Treatise of Physics

An evidence-based account of theoretical, mathematical, and experimental physics

Nicolas Gonzalez

Written with the assistance of Claude (Opus 5, Fable 5);
see *A note on method* in the preface.

August 25, 2026

Preface

To be human is to change. Physics is the description of change; even the acquisition of knowledge is itself a change. And yet beneath all of it Nature keeps a set of regularities — rules — that allow us to reproduce everything empirical that she is telling us. That set of rules, taken whole, is the entire framework of Theoretical, Mathematical, and Experimental Physics.

This treatise attempts a single, connected account of physics restricted to what has actually been observed in Nature. Its three threads — theoretical, mathematical, and experimental physics — are kept deliberately interwoven: every theoretical claim is anchored to the experiments that force it, every experiment is presented in full phenomenological detail, and every observed phenomenon carries its mathematical derivation, either in place or in Appendix A.

Speculative programs without experimental support — string theory, supersymmetry, supergravity, and their relatives — are outside the scope of this work, however mathematically compelling. High-energy physics up to the Standard Model, being among the best-tested science ever produced, is treated in full. The frontier of what is observed but not understood is stated honestly in the closing part.

Two working rules govern the text throughout. First, SI units are used axiomatically in every branch. Second, the mathematics is developed in the generality it deserves — in particular, geometric and group-theoretic chapters are written for general dimension $D = p + q$ with signature (p, q) — while the physics is always instantiated in the observed $3 + 1$ dimensions.

Derivations not yet written are never silently omitted: they are marked as pending, and collected in the List of Pending Derivations.

A note on method

Much of this text was drafted with the assistance of large language models — Anthropic's Claude, in the Opus 5 and Fable 5 versions. The division of labour was this: I set the scope, the editorial rules and the structure, directed the work chapter by chapter, and decided what the book would and would not claim; the models translated and ported the older Spanish sources, drafted prose and derivations to those rules, and wrote most of the checking apparatus described below. They are named on the title page as what they were, an instrument. They are not authors, and could not be: an author answers for the work, and a language model cannot. Every error here is mine.

That apparatus is worth describing, because it is what the honesty of a book this size actually rests on. The sources are checked mechanically before every commit: that each citation key resolves to a real entry, that each entry's DOI resolves to the work claimed for it, that no equation silently loses a unit, that nothing reaches the reader as raw markup, and that the logical skeleton of the book — which statement rests on which, which derivation discharges which claim — is a machine-readable database whose two representations are proved equivalent. Where a result is quoted rather than proved, a remark headed *What is*

quoted here says so and names the missing theorem.

None of that is peer review, and it should not be mistaken for it. A derivation drafted by a model and checked by a program can still be wrong in ways neither notices; a citation can resolve perfectly and support nothing. What the machinery buys is that the book's gaps are visible rather than hidden — the pending derivations are listed, the quoted inputs are named, and the claims that no experiment tests are marked as such. A reader who distrusts a particular argument is meant to be able to find its foundations and check them. That is the standard this book sets for itself, and the reason to prefer a stated gap to a confident sentence.

Editorial Conventions

The four rules below govern how this treatise is written, read once here and applied silently everywhere after. They are rules of editorial *policy*, not of notation; for what the symbols themselves mean, see the single canonical reference, Appendix E.

Units

> **Axiom (SI).** All physical quantities in this treatise are expressed in the International System of Units (SI), in every branch of physics.

Numerical values are typeset with their units, e.g. the standard acceleration of free fall $g_0 = 9.806\,65\,\mathrm{m/s^2}$. Natural or Gaussian units appear only inside clearly flagged remarks, always with the conversion back to SI stated. The SI base units and the defining constants of the 2019 revision are collected in Appendix B.

Proofs of observed phenomena

Every observed phenomenon stated in this treatise (environment *Phenomenon*) carries its mathematical derivation: short derivations follow the statement directly; long derivations live in Appendix A and are linked from the statement. A derivation not yet written is marked visibly as pending — never omitted.

Experiments

Experiments are presented in dedicated boxes with a fixed structure: apparatus, procedure, observations and data (with uncertainties, in SI), interpretation, and primary references. The List of Experiments in the front matter indexes all of them.

Notation

Every symbol, macro, and index convention used in this treatise — vector notation, the differential d, index-letter conventions for general dimension $D = p + q$, the full macro-to-meaning dictionary — is collected in one place, Appendix E, rather than restated piecemeal through the front matter.

Contents

Preface iii

Editorial Conventions v

List of Experiments xcv

List of Pending Derivations xcix

I Foundations 1

1 Epistemology and the Scientific Method 3
- 1.1 What physics is . . . 3
- 1.2 Physics, mathematics, and philosophy . . . 3
- 1.3 Why this treatise follows analytic philosophy . . . 4
  - 1.3.1 The choice cannot be avoided . . . 4
  - 1.3.2 What the commitment amounts to . . . 4
  - 1.3.3 Why this tradition and not another . . . 5
  - 1.3.4 What the alternatives would cost . . . 6
  - 1.3.5 The price of the choice . . . 6
  - 1.3.6 How the choice shows on the page . . . 7
- 1.4 The scientific method as used in this treatise . . . 7
- 1.5 Primitive notions . . . 8
- 1.6 How this treatise is organized . . . 9

2 Measurement, SI Units, and the Theory of Errors 11
- 2.1 Measurement . . . 11
- 2.2 The International System of Units . . . 11
  - 2.2.1 Dimensional analysis . . . 11
- 2.3 The theory of errors . . . 12
  - 2.3.1 Repeated measurements . . . 12
  - 2.3.2 Propagation of uncertainty . . . 13
  - 2.3.3 Combining and fitting data . . . 14
  - 2.3.4 Reporting . . . 14
- 2.4 A worked example: timing a pendulum . . . 14

II Mathematical Methods 15

3 Logic, Sets, and Maps 17
- 3.1 Logic . . . 17
  - 3.1.1 Propositions . . . 17

3.1.2 Logical operators . . . 18
3.1.3 Implications and logical equivalences . . . 21
3.1.4 Predicates and quantifiers . . . 23
3.1.5 Methods of proof . . . 24
3.2 Sets . . . 25
3.2.1 Sets . . . 25
3.2.2 Numbers . . . 26
3.2.3 Operations with sets . . . 28
3.2.4 Cardinality . . . 30
3.3 Maps . . . 31
3.3.1 Maps . . . 31
3.4 Equivalence classes . . . 34
3.4.1 Equivalence relations . . . 34
3.4.2 Quotient space . . . 35
3.5 The infinite . . . 36
3.5.1 Equipotence and countability . . . 36
3.5.2 The diagonal argument . . . 37
3.5.3 The continuum hypothesis . . . 38
3.6 The crisis of foundations . . . 39
3.6.1 Russell's paradox . . . 39
3.6.2 The Zermelo–Fraenkel axioms . . . 40
3.6.3 The axiom of choice . . . 41
3.6.4 Three answers to the crisis . . . 41
3.7 The analytical perspective . . . 42
3.7.1 The linguistic turn . . . 42
3.7.2 Logical positivism and its collapse . . . 43
3.7.3 The position this treatise takes . . . 44
3.8 Formal systems and incompleteness . . . 44
3.8.1 Formal systems . . . 45
3.8.2 Completeness — what logic does deliver . . . 46
3.8.3 Arithmetization . . . 46
3.8.4 The first incompleteness theorem . . . 47
3.8.5 The second incompleteness theorem . . . 48
3.8.6 What the theorems do not say . . . 48
3.9 Computability and decidability . . . 49
3.9.1 Turing machines . . . 50
3.9.2 The halting problem . . . 51
3.9.3 The Entscheidungsproblem answered . . . 52
3.9.4 Where decision procedures do exist . . . 52
3.9.5 Undecidability inside physics . . . 53
3.10 Can mathematics be built without contradictions? . . . 54
3.10.1 What Hilbert asked for . . . 54
3.10.2 What survived: relative consistency . . . 54
3.10.3 The honest answer . . . 55
3.10.4 How much of this reaches physics . . . 55

4 Algebraic Structures **57**
4.1 Binary operations and algebraic structures . . . 57
4.1.1 Binary operations . . . 57
4.1.2 Maps between algebraic structures . . . 60
4.1.3 Kernel and image . . . 61
4.2 Simple algebraic structures . . . 62

4.2.1 Semigroups, monoids, and groups . . . 62
4.2.2 Basic properties of groups . . . 64
4.3 Composite algebraic structures . . . 65
4.3.1 Axioms for composite algebraic structures . . . 65
4.3.2 Composite algebraic structures . . . 66
4.4 Group theory: further definitions . . . 68
4.4.1 Groups . . . 68
4.4.2 Cayley tables . . . 69
4.5 Basic groups . . . 70
4.5.1 The first finite groups . . . 70
4.5.2 The group $\mathbb{Z}$ . . . 71
4.6 Functions on groups . . . 74
4.6.1 Homomorphisms . . . 75
4.6.2 Properties of homomorphisms . . . 76
4.7 The symmetric group . . . 77
4.7.1 The symmetric group . . . 77
4.7.2 The permutation group . . . 78
4.8 Decompositions of a group . . . 81
4.8.1 Conjugacy classes . . . 81
4.8.2 Cosets . . . 83
4.8.3 The quotient group . . . 85
4.9 Direct products . . . 87
4.9.1 Direct product . . . 87
4.9.2 Semidirect product . . . 90

5 Linear Algebra and Representation Theory 95
5.1 Preliminary formulas from elementary algebra . . . 95
5.1.1 Factorials and the binomial theorem . . . 95
5.1.2 Trigonometric formulas . . . 95
5.1.3 The sine and cosine theorems . . . 96
5.1.4 Cramer's rule . . . 96
5.1.5 Derivations of the elementary identities . . . 96
5.1.6 The derivative of a determinant . . . 100
5.2 Vector spaces . . . 101
5.2.1 Vector spaces . . . 101
5.2.2 Bases . . . 104
5.3 Inner product, norm, and metric . . . 106
5.3.1 Inner product . . . 106
5.3.2 Orthogonal bases . . . 111
5.3.3 Vector algebra in three dimensions . . . 113
5.4 Linear transformations . . . 116
5.4.1 Linear transformations . . . 116
5.4.2 The vector space of linear transformations . . . 121
5.4.3 Morphisms of vector spaces . . . 124
5.5 The dual space . . . 127
5.5.1 The dual space . . . 127
5.5.2 Pullback and pushforward . . . 131
5.6 Eigenvectors, eigenvalues, and decompositions . . . 133
5.6.1 Eigenvectors and eigenvalues . . . 133
5.6.2 Decomposition of vector spaces . . . 147
5.6.3 Bilinear transformations . . . 151
5.7 Algebras . . . 157

5.7.1 Algebras . . . 157
5.7.2 The algebra of linear transformations . . . 158
5.7.3 Lie algebras . . . 159
5.7.4 Derivations . . . 160
5.8 Representations of groups . . . 161
5.8.1 The classical groups . . . 161
5.8.2 Representations . . . 168
5.8.3 Irreducible representations . . . 169
5.9 Representations of finite groups . . . 171
5.9.1 Complete reducibility of a unitary representation . . . 171
5.9.2 Every representation of a finite group is unitary . . . 173
5.9.3 The regular representation . . . 176
5.9.4 Direct-sum and direct-product representations . . . 178
5.10 Properties of irreducible representations . . . 178
5.10.1 Fundamental theorems of representations . . . 178
5.10.2 Orthonormality relations . . . 181
5.10.3 Decomposition into irreducible representations . . . 183
5.11 Representations of algebras . . . 187
5.11.1 Basic properties of the representations . . . 187

6 Topological and Metric Spaces 193
6.1 Topological spaces . . . 193
6.1.1 Basic definitions . . . 193
6.1.2 Continuity . . . 194
6.1.3 Compactness . . . 194
6.1.4 Connectedness . . . 195
6.1.5 The winding number . . . 196
6.2 Metric spaces . . . 198
6.2.1 The metric . . . 198
6.2.2 Convergence and completeness . . . 199
6.2.3 Continuity in metric terms . . . 199
6.2.4 Compactness in metric spaces . . . 200
6.3 The topology of $\mathbb{R}^N$ . . . 201
6.3.1 Vectors in $\mathbb{R}^N$ . . . 201

7 Real Analysis 203
7.1 The real numbers . . . 203
7.2 Sequences . . . 203
7.3 Real functions . . . 204
7.4 Limits . . . 205
7.4.1 One-sided limits . . . 206
7.4.2 Infinite limits and limits at infinity . . . 206
7.5 Continuity . . . 207
7.6 The derivative . . . 208
7.6.1 Geometric interpretation of the derivative . . . 208
7.6.2 The derivative function . . . 208
7.6.3 Rules of differentiation . . . 213
7.6.4 The mean value theorems . . . 214
7.7 The Riemann integral . . . 216
7.8 The fundamental theorem of calculus . . . 216
7.9 Series . . . 217
7.9.1 The numbers e and π . . . 219

7.10 Multivariable calculus . . . 225
7.10.1 Planes in $\mathbb{R}^3$ . . . 225
7.10.2 Differentiable functions . . . 225
7.10.3 Taylor expansion in several variables . . . 228
7.10.4 Differentiation under the integral sign . . . 230
7.10.5 The implicit function theorem . . . 232
7.11 Vector analysis . . . 236
7.11.1 The nabla operator . . . 236
7.11.2 Identities of the nabla calculus . . . 236
7.11.3 The integral theorems . . . 239
7.11.4 Conservative fields . . . 243
7.11.5 The Helmholtz decomposition . . . 244
7.12 Sets of measure zero and Hausdorff dimension . . . 245
7.13 Functional calculus . . . 247
7.14 The Dirac delta . . . 248

8 Complex Analysis 249
8.1 The complex field . . . 249
8.2 Holomorphic functions . . . 250
8.3 Contour integrals . . . 251
8.4 Cauchy's integral theorem . . . 251
8.5 The Cauchy integral formula . . . 253
8.6 Laurent series and singularities . . . 255
8.7 The residue theorem . . . 255

9 Ordinary Differential Equations and Sturm–Liouville Theory 257
9.1 Ordinary differential equations . . . 257
9.1.1 Order, linearity, homogeneity . . . 258
9.1.2 Existence and uniqueness . . . 258
9.1.3 The solution space and the Wronskian . . . 262
9.1.4 Elementary first-order equations . . . 263
9.2 Linear systems with constant coefficients . . . 264
9.2.1 The matrix exponential . . . 264
9.2.2 The shape of the solutions . . . 266
9.2.3 Planar systems: classification of the equilibrium . . . 268
9.2.4 Equilibria of a nonlinear system . . . 269
9.2.5 Planar flows and closed orbits . . . 270
9.3 Linear systems with periodic coefficients . . . 271
9.3.1 The monodromy matrix . . . 271
9.3.2 Hill's equation and the tongues of instability . . . 273
9.3.3 The Mathieu equation . . . 274
9.4 Orthonormal systems of functions . . . 275
9.4.1 Orthonormality . . . 275
9.4.2 Expansion in series of orthonormal functions . . . 275
9.4.3 Expansion of a product . . . 276
9.4.4 Parseval relation . . . 277
9.5 The Sturm–Liouville operator . . . 278
9.5.1 Self-adjointness and orthogonality . . . 278
9.6 The semiclassical approximation . . . 279
9.6.1 The equation and its Sturm–Liouville data . . . 280
9.6.2 The expansion in powers of $\hbar$ . . . 280
9.6.3 Where the approximation is valid, and where it is not . . . 283

9.6.4 Connection across a simple turning point . . . 284
9.6.5 Turning points of higher order . . . 287
9.6.6 Barrier penetration and the Gamow factor . . . 288
9.6.7 Bohr–Sommerfeld quantisation and the Maslov index . . . 290
9.6.8 Worked example: a charged particle against a Coulomb barrier . . . 291
9.6.9 The radial equation . . . 292
9.6.10 The Langer substitution . . . 293
9.6.11 Boundary layers and matched asymptotic expansions . . . 294
9.7 The one-dimensional Helmholtz equation . . . 296
9.7.1 Sturm–Liouville data . . . 296
9.7.2 Homogeneous solution . . . 296
9.7.3 Orthonormality . . . 297
9.7.4 Fourier series . . . 297
9.7.5 The Fourier transform . . . 298
9.8 The Bessel equation . . . 299
9.8.1 Sturm–Liouville data . . . 299
9.8.2 General solution . . . 299
9.8.3 Recurrence relations . . . 300
9.8.4 Orthogonality . . . 301
9.8.5 Generating function . . . 301
9.8.6 The Fourier–Bessel series . . . 302
9.8.7 The zeros . . . 302
9.8.8 The modified Bessel functions . . . 304
9.9 The Legendre equation and spherical harmonics . . . 305
9.9.1 The Legendre equation . . . 305
9.9.2 The associated Legendre equation . . . 308
9.9.3 Spherical harmonics . . . 311
9.10 The Hermite equation . . . 312
9.10.1 Sturm–Liouville data . . . 312
9.10.2 Solutions and generating function . . . 312
9.10.3 Orthogonality and Hermite series . . . 313
9.11 The Laguerre equation . . . 314
9.11.1 Sturm–Liouville data . . . 314
9.11.2 Solutions and generating function . . . 314
9.11.3 Orthogonality and Laguerre series . . . 315
9.12 Elliptic integrals and the Jacobi elliptic functions . . . 316
9.12.1 The three normal forms . . . 316
9.12.2 The Jacobi elliptic functions . . . 318
9.12.3 Quadratures that reduce to elliptic integrals . . . 319
9.13 The Laplace transform . . . 321
9.14 Summary of the classical families . . . 322

10 Partial Differential Equations **323**
10.1 Definitions . . . 323
10.1.1 Classification and characteristics . . . 324
10.1.2 Well-posedness in the sense of Hadamard . . . 330
10.1.3 The equations of physics . . . 331
10.2 From partial to ordinary differential equations . . . 332
10.2.1 Separable coordinate systems . . . 333
10.2.2 Eigenfunction expansions . . . 334
10.2.3 Integral-transform methods . . . 335
10.3 Green's functions . . . 336

10.3.1 The Green's function of a boundary-value problem . . . 336
10.3.2 Causal Green's functions and Duhamel's principle . . . 338
10.4 The canonical equations and their fundamental solutions . . . 339
10.4.1 The wave equation . . . 339
10.4.2 The heat equation . . . 342
10.4.3 The Laplace and Poisson equations . . . 343
10.4.4 The Helmholtz decomposition . . . 344
10.5 Maximum principles . . . 345
10.5.1 Elliptic maximum principles . . . 345
10.5.2 Parabolic maximum principles . . . 347
10.6 Weak solutions . . . 347
10.6.1 Distributional solutions and Sobolev spaces . . . 347
10.6.2 Variational formulation . . . 348
10.6.3 Shocks and conservation laws . . . 350

11 Probability and Statistics **355**
11.1 Discrete statistics . . . 355
11.1.1 Definitions . . . 355
11.2 Continuous statistics . . . 361
11.2.1 Definitions . . . 361
11.2.2 Measure spaces, null sets and "almost every" . . . 364
11.2.3 Invariant measures and the ergodic theorem . . . 366
11.3 Characteristic functions and the central limit theorem . . . 367
11.3.1 Characteristic functions . . . 367
11.3.2 The central limit theorem . . . 369
11.4 Estimation . . . 371
11.4.1 The likelihood function . . . 373
11.4.2 Score and Fisher information . . . 374
11.4.3 The Cramér–Rao bound . . . 376
11.4.4 Consistency and asymptotic normality . . . 378
11.4.5 Nuisance parameters and the profile likelihood . . . 381
11.5 Hypothesis testing . . . 382
11.5.1 Tests, size, power and p-values . . . 382
11.5.2 The Neyman–Pearson lemma . . . 383
11.5.3 The chi-squared distribution . . . 384
11.5.4 The likelihood-ratio statistic and Wilks' theorem . . . 386
11.5.5 Significance . . . 389
11.6 The look-elsewhere effect . . . 390
11.6.1 Local and global p-values . . . 390
11.6.2 Upcrossings . . . 391
11.6.3 A worked example: a bump search over a mass window . . . 394

12 Hilbert Spaces **397**
12.1 Definitions . . . 397
12.2 Orthonormal bases and separability . . . 400
12.2.1 Orthogonality and projections . . . 400
12.2.2 Gram–Schmidt orthogonalization . . . 403
12.2.3 Fourier expansion, Bessel and Parseval . . . 404
12.3 Bounded operators and duality . . . 406
12.3.1 Bounded operators and their adjoints . . . 406
12.3.2 The Riesz representation theorem . . . 410
12.4 Self-adjointness and the spectral theorem . . . 411

12.4.1 Spectrum and resolvent . . . 411
12.4.2 The spectral theorem . . . 414
12.4.3 Stone's theorem . . . 416
12.5 Unbounded operators and domains . . . 418
12.5.1 Closed operators and domains . . . 418
12.5.2 Symmetric versus self-adjoint: deficiency indices . . . 420
12.6 Direct sums, tensor products and rigged spaces . . . 423
12.6.1 Direct sums and orthogonal decompositions . . . 423
12.6.2 Tensor products . . . 425
12.6.3 Rigged Hilbert spaces . . . 427
12.7 The canonical commutation relations in Weyl form . . . 429

13 Differentiable Manifolds, Tensors, and Curvature 433
13.1 Index conventions and the symmetry ladder . . . 433
13.2 Tensors under orthogonal transformations . . . 434
13.2.1 Orthogonal coordinate transformations . . . 434
13.2.2 Algebra of Cartesian tensors . . . 438
13.3 The theory of curves and surfaces in $\mathbb{R}^3$ . . . 441
13.3.1 Theory of curves . . . 441
13.3.2 Theory of surfaces . . . 445
13.3.3 Curvatures of a surface . . . 451
13.4 Differentiable manifolds . . . 462
13.4.1 Charts, atlases, and differentiable structure . . . 462
13.4.2 Euclidean space . . . 474
13.4.3 Curves in Euclidean space . . . 477
13.5 Tensor analysis on manifolds . . . 483
13.5.1 Vectors, covectors, and tensors on a manifold . . . 483
13.5.2 Tensors under general coordinate transformations . . . 488
13.5.3 Tensor densities . . . 492
13.6 Differential forms: k-forms and k-vectors . . . 495
13.6.1 k-forms and k-vectors . . . 495
13.6.2 The basis of the cotangent space . . . 496
13.6.3 The wedge product . . . 497
13.6.4 The exterior derivative . . . 498
13.6.5 Closed and exact forms . . . 499
13.6.6 The volume form . . . 502
13.6.7 The Hodge dual . . . 502
13.7 The metric tensor . . . 503
13.7.1 Isometries . . . 504
13.8 Vielbein and local frames . . . 504
13.9 Flows, the Lie derivative, and Killing vectors . . . 505
13.9.1 Distributions and the Frobenius theorem . . . 508
13.9.2 Killing vectors . . . 512
13.10 Connections, parallel transport, and torsion . . . 514
13.10.1 The spin connection . . . 516
13.11 Curvature . . . 516
13.12 The Cartan formalism . . . 517
13.13 Maximally symmetric spaces and the symmetry ladder . . . 519
13.13.1 The conformal group . . . 522
13.13.2 The observed case . . . 523

14 Lie Groups, Lie Algebras, and Fibre Bundles **525**
14.1 Lie groups . . . 525
14.1.1 Definition and generators . . . 525
14.2 Representations of Lie groups . . . 534
14.2.1 The group SO(2, ℝ) . . . 534
14.2.2 The groups SO(3), SU(2) and SU(3) . . . 537
14.3 The Lie algebra associated with a Lie group . . . 552
14.3.1 Translations in D dimensions . . . 552
14.3.2 The algebra $\mathfrak{so}(p,q)$ . . . 552
14.4 Central extensions . . . 563
14.4.1 Extensions and cocycles . . . 563
14.4.2 The instances that occur in physics . . . 565
14.5 Contraction of Lie algebras . . . 566
14.5.1 The Inönü–Wigner limit . . . 566
14.5.2 The classification of kinematical algebras . . . 569
14.6 Fibre bundles . . . 571
14.6.1 Bundles, sections and principal bundles . . . 572
14.6.2 Connection and curvature . . . 573
14.7 Invariant polynomials . . . 575
14.8 The Chern–Weil theorem and transgression forms . . . 578
14.8.1 Chern–Simons forms . . . 581
14.8.2 Characteristic classes . . . 582
14.8.3 The transgression of an $\mathfrak{so}(p,q)$ connection . . . 583

15 Expansions of Lie Algebras **585**
15.1 The tensor product construction . . . 586
15.2 Ideals, the radical, and the Killing form . . . 587
15.3 The Takiff tower . . . 590
15.4 Gradings and resonant subalgebras . . . 591
15.4.1 The gradings of the anti-de Sitter algebra . . . 592
15.5 Contraction as a degeneration of A . . . 593
15.5.1 The curvature parameter lives in A . . . 594
15.5.2 The (anti-)de Sitter radius and the cosmological constant . . . 595
15.6 Central charges as the top level of an expansion . . . 597
15.7 The Bacry–Lévy-Leblond cube as a lattice of algebras . . . 601
15.7.1 The kinematical generators are a Klein grading . . . 601
15.7.2 Resonance for a graded group . . . 601
15.7.3 The algebra with three structure constants . . . 602
15.7.4 The cube . . . 602
15.8 Invariant forms and why the extension is needed . . . 605
15.9 Relation to the published expansion methods . . . 609
15.10 The construction in 3+1 dimensions . . . 610
15.11 Positive functionals and the Gram form . . . 611

16 Calculus of Variations **615**
16.1 The classical problems . . . 615
16.1.1 The brachistochrone . . . 615
16.1.2 The catenary and the isoperimetric problem . . . 616
16.1.3 Minimal surfaces and soap films . . . 618
16.2 Functionals and the first variation . . . 619
16.2.1 Function spaces and functionals . . . 619
16.2.2 The fundamental lemma . . . 621

16.2.3 The Euler–Lagrange equation . . . 622
16.2.4 First integrals and natural boundary conditions . . . 624
16.2.5 Several functions, several variables, higher derivatives . . . 626
16.3 Constrained variation . . . 628
16.3.1 Lagrange multipliers . . . 628
16.3.2 Isoperimetric constraints . . . 629
16.3.3 Holonomic and nonholonomic constraints . . . 630
16.4 The second variation . . . 631
16.4.1 The Legendre condition . . . 631
16.4.2 The Jacobi equation and conjugate points . . . 632
16.4.3 Strong variations and the Weierstrass condition . . . 634
16.4.4 Fields of extremals and sufficiency . . . 635
16.5 Direct methods and existence . . . 638
16.5.1 The Dirichlet principle and its failure . . . 638
16.5.2 Tonelli's existence theory . . . 639
16.5.3 The Rayleigh–Ritz method . . . 641
16.5.4 The Plateau problem . . . 644
16.6 Symmetry and conservation: Noether's theorems . . . 644
16.6.1 Noether's first theorem . . . 645
16.6.2 Noether's second theorem . . . 648
16.7 The variational principles of physics . . . 649
16.7.1 Hamilton's principle . . . 649
16.7.2 Maupertuis' principle . . . 650
16.7.3 Fermat's principle . . . 651
16.7.4 Field actions . . . 653

17 Fourier Analysis and Integral Transforms 655
17.1 Fourier series . . . 655
17.1.1 Trigonometric expansion of periodic functions . . . 655
17.1.2 Convergence: from Dirichlet to Carleson . . . 658
17.1.3 The Gibbs phenomenon and summability . . . 665
17.1.4 Series in several variables . . . 667
17.2 The Fourier transform . . . 668
17.2.1 Definition and inversion . . . 668
17.2.2 Parseval and Plancherel . . . 671
17.2.3 Tempered distributions . . . 673
17.2.4 The uncertainty principle . . . 675
17.2.5 Poisson summation . . . 676
17.3 Convolution and correlation . . . 677
17.3.1 The convolution theorem . . . 677
17.3.2 Correlation and the Wiener–Khinchin theorem . . . 679
17.4 The Laplace transform . . . 682
17.4.1 Definition, convergence and inversion . . . 682
17.4.2 Initial-value problems and the operational calculus . . . 685
17.5 Sampling and the discrete transform . . . 687
17.5.1 The sampling theorem . . . 687
17.5.2 The discrete Fourier transform and the FFT . . . 689
17.6 Other integral transforms . . . 691
17.6.1 Hankel and Mellin transforms . . . 691
17.6.2 Multiplicative convolution and the moments . . . 693
17.6.3 The Radon transform . . . 695

III Classical Mechanics **699**

18 Kinematics **701**
18.1 Foundations . . . 701
18.1.1 Fundamental concepts . . . 701
18.1.2 Frames of reference . . . 703
18.1.3 Kinematics of a particle . . . 703
18.1.4 The itinerary equation for constant acceleration . . . 705
18.2 Galilean relativity . . . 706
18.2.1 Newton's postulates of absolute space and time . . . 706
18.2.2 Galileo's principle of relativity . . . 707
18.2.3 Galilean transformations . . . 708
18.2.4 Transformation of velocities and accelerations . . . 710
18.3 Circular motion . . . 711
18.3.1 The angular description . . . 712
18.3.2 Uniform circular motion . . . 712
18.3.3 Non-uniform circular motion . . . 713
18.3.4 The intrinsic decomposition of the acceleration . . . 714
18.3.5 Magnitudes . . . 714

19 Newtonian Dynamics **717**
19.1 Inertia and mass . . . 717
19.2 Newton's laws and the Newtonian formulation: translation . . . 718
19.2.1 Linear momentum and Newton's first law . . . 718
19.2.2 Force and Newton's second law . . . 719
19.3 Newton's laws and the Newtonian formulation: rotation . . . 722
19.3.1 Angular momentum . . . 722
19.3.2 Torque . . . 724
19.4 Energy . . . 726
19.4.1 Kinetic energy . . . 726
19.4.2 Potential energy and conservative forces . . . 726
19.4.3 Energy . . . 727
19.4.4 The law of conservation of energy . . . 727
19.4.5 Work . . . 728
19.4.6 The work–energy theorem . . . 730
19.5 Systems of particles . . . 731
19.5.1 Inertia and mass of a system . . . 731
19.5.2 Internal dynamics of the system . . . 734
19.5.3 Translational Newton laws for systems of particles . . . 736
19.5.4 Rotational Newton laws for systems of particles . . . 738
19.5.5 Conservation laws . . . 742

20 Experiment: Free Fall and Projectile Motion **747**
20.1 Historical context: the Aristotelian expectation and its collapse . . . 747
20.2 The observed phenomenon and its derivation . . . 748
20.3 Galileo's Inclined Planes (1604–1638) . . . 750
20.3.1 Apparatus . . . 750
20.3.2 Procedure . . . 750
20.3.3 Observations and data . . . 750
20.3.4 Interpretation . . . 751
20.3.5 Primary references . . . 751
20.4 Absolute Ballistic Gravimetry (1995–present) . . . 751

20.4.1 Apparatus . . . 751
20.4.2 Procedure . . . 751
20.4.3 Observations and data . . . 752
20.4.4 Interpretation . . . 752
20.4.5 Primary references . . . 752
20.5 Deviations and the domain of validity . . . 752
20.6 From Galileo's ratios to the equivalence principle . . . 757

21 Lagrangian Mechanics **759**
21.1 Generalized coordinates . . . 759
21.1.1 Constraints . . . 759
21.1.2 Degrees of freedom . . . 760
21.1.3 Generalized coordinates . . . 760
21.1.4 Constraint forces and active forces . . . 761
21.2 Virtual work . . . 761
21.2.1 Virtual displacements . . . 761
21.2.2 The constraint-force hypothesis . . . 761
21.2.3 Virtual work and its principle . . . 762
21.3 d'Alembert's principle . . . 763
21.4 The Euler–Lagrange equations . . . 763
21.4.1 Lagrange equations of the first kind . . . 763
21.4.2 Lagrange multipliers as constraint forces . . . 764
21.4.3 Lagrange equations of the second kind . . . 767
21.4.4 The Lagrangian . . . 769
21.4.5 Cyclic coordinates . . . 770
21.4.6 The energy of a Lagrangian system . . . 770
21.5 Hamilton's principle . . . 773
21.5.1 The action . . . 773
21.5.2 Hamilton's principle . . . 773
21.6 Nonholonomic constraints . . . 774
21.6.1 Pfaffian constraints . . . 775
21.6.2 The d'Alembert–Chetaev prescription . . . 775
21.6.3 Why the vakonomic prescription is not the observed one . . . 776
21.7 General coordinate transformations . . . 778
21.7.1 Point transformations . . . 778
21.7.2 Covariance of the Euler–Lagrange equations . . . 779
21.7.3 The kinetic metric and the geodesic form . . . 780
21.7.4 Inertial forces are connection coefficients . . . 781
21.8 The universality of general covariance . . . 782
21.8.1 Flat spacetime in arbitrary coordinates . . . 782
21.8.2 What general relativity adds . . . 784

22 Hamiltonian Mechanics **789**
22.1 From the Lagrangian to the Hamiltonian . . . 789
22.1.1 The Legendre transformation of the Lagrangian . . . 789
22.1.2 Hamilton's equations . . . 790
22.1.3 The Hamiltonian and the energy . . . 792
22.2 Canonical transformations . . . 793
22.2.1 Definition and transformation conditions . . . 793
22.2.2 Generating functions . . . 794
22.2.3 Derivation of the transformation conditions . . . 795
22.3 Invariant integrals . . . 797

22.3.1 The Lagrange bracket . . . 798
22.3.2 The Poisson bracket . . . 800
22.4 Motion in phase space . . . 802
22.5 Symplectic formulation . . . 804
22.5.1 Symplectic form of the transformation condition . . . 805
22.6 Beyond the canonical formalism . . . 805
22.6.1 Hamilton–Jacobi theory . . . 805
22.6.2 Symplectic geometry . . . 805
22.6.3 Symmetries . . . 806

23 Hamilton–Jacobi Theory and the Optical–Mechanical Analogy 807
23.1 The Hamilton–Jacobi equation . . . 807
23.1.1 The transformation to constant coordinates . . . 807
23.1.2 Jacobi's theorem . . . 808
23.1.3 Time-independent systems and the characteristic function . . . 810
23.2 Complete integrals and separation of variables . . . 811
23.2.1 Conditions for separability . . . 812
23.2.2 Worked separations . . . 815
23.3 Action–angle variables and integrable systems . . . 817
23.3.1 Adiabatic invariance . . . 819
23.3.2 The old quantum theory . . . 821
23.4 The optical–mechanical analogy . . . 821
23.4.1 The eikonal limit of wave optics . . . 821
23.4.2 The dictionary . . . 823
23.4.3 Where the analogy is not an analogy . . . 824
23.5 From Hamilton–Jacobi to the Schrödinger equation . . . 825
23.5.1 The semiclassical expansion . . . 826
23.5.2 Caustics . . . 827

24 Symplectic Geometry of Phase Space 833
24.1 Symplectic vector spaces . . . 833
24.2 Symplectic manifolds . . . 835
24.3 Hamiltonian vector fields and the flow . . . 837
24.4 Invariants of the flow . . . 840
24.4.1 Integral invariants . . . 841
24.4.2 Symplectic capacity and non-squeezing . . . 843
24.5 Poisson manifolds . . . 847
24.6 Symmetry, momentum maps, and reduction . . . 849
24.7 Geometric quantization . . . 852
24.7.1 Prequantization . . . 852
24.7.2 Polarization . . . 853
24.7.3 Bohr–Sommerfeld from holonomy . . . 854
24.8 What the symplectic view buys . . . 855

25 The Poisson Algebra and the Canonical Bridge to Quantum Mechanics 857
25.1 The Poisson algebra of observables . . . 857
25.2 Generators, symmetries, and constants of motion . . . 859
25.2.1 Mass as a central charge . . . 863
25.2.2 The generators as a bridge . . . 864
25.3 The first-order formalism . . . 865
25.3.1 The Faddeev–Jackiw form . . . 866
25.4 Canonical quantization: the correspondence rule . . . 868

25.5 The obstruction: what the correspondence cannot be . . . 872
25.5.1 Ordering . . . 874
25.5.2 Deformation quantization and the Moyal bracket . . . 877
25.5.3 Phase-space quantum mechanics . . . 877
25.6 What carries over to field theory . . . 881

26 Constrained Hamiltonian Systems: the Dirac–Bergmann Formalism 883
26.1 Singular Lagrangians and primary constraints . . . 884
26.2 The Dirac–Bergmann consistency algorithm . . . 885
26.3 First- and second-class constraints . . . 887
26.4 The Dirac bracket . . . 889
26.5 Gauge fixing and the reduced phase space . . . 892
26.6 Worked systems . . . 894
26.6.1 The relativistic free particle . . . 895
26.6.2 The electromagnetic field . . . 896
26.6.3 The massive vector field . . . 897
26.6.4 Yang–Mills fields . . . 898
26.6.5 General relativity . . . 900
26.7 Quantizing a constrained system . . . 903
26.7.1 Reduced-phase-space quantization . . . 904
26.7.2 Weak imposition: the Gupta–Bleuler construction . . . 905
26.7.3 BRST . . . 906
26.8 Constraints in field theory . . . 907

27 Central Forces and Statics 909
27.1 Statics . . . 909
27.1.1 The centre of gravity . . . 911
27.1.2 Equilibrium as a stationary potential energy . . . 912
27.2 Central forces . . . 913
27.2.1 Areal velocity . . . 914
27.2.2 Conservation laws of central motion . . . 915
27.2.3 The Binet equation . . . 916
27.3 The equivalent one-dimensional problem . . . 917
27.3.1 Energy of the system . . . 917
27.3.2 Effective potential . . . 918
27.3.3 Turning points . . . 918
27.3.4 The radial quadrature . . . 919
27.4 The two-body problem and Kepler orbits . . . 920
27.4.1 Reduction to one body . . . 920
27.4.2 The orbit of an inverse-square attraction . . . 921
27.4.3 Kepler's third law . . . 923
27.4.4 The Laplace–Runge–Lenz vector . . . 925
27.4.5 Closure, and what its failure measures . . . 926
27.5 The virial theorem . . . 927
27.6 Where the claims of this chapter are tested . . . 928

28 Oscillations and Mechanical Waves 929
28.1 The harmonic oscillator . . . 930
28.1.1 Free oscillations . . . 930
28.1.2 The pendulum . . . 931
28.1.3 Universality of the harmonic approximation . . . 933
28.2 Damped and driven oscillations . . . 934

28.2.1 Linear damping . . . 934
28.2.2 The driven oscillator and resonance . . . 936
28.2.3 Parametric resonance . . . 938
28.3 Coupled oscillators and normal modes . . . 940
28.3.1 Two coupled oscillators and beats . . . 940
28.3.2 Small oscillations about equilibrium . . . 941
28.3.3 The loaded string and the continuum limit . . . 944
28.4 The vibrating string and the wave equation . . . 945
28.4.1 d'Alembert's equation and its general solution . . . 945
28.4.2 Superposition and the vibrating-string controversy . . . 946
28.4.3 Standing waves . . . 947
28.4.4 Fourier decomposition . . . 948
28.5 Sound . . . 949
28.5.1 The acoustic wave equation . . . 949
28.5.2 Acoustic phenomena . . . 951
28.6 Dispersion . . . 953
28.6.1 Phase and group velocity . . . 953
28.6.2 Energy transport and impedance . . . 954
28.7 Anharmonicity: a first look . . . 956
28.7.1 The anharmonic oscillator . . . 956
28.7.2 Secular terms and the Lindstedt–Poincaré method . . . 959

29 Rigid Bodies and Rotating Frames 961
29.1 Rigid-body kinematics . . . 961
29.1.1 Angular velocity and Euler's theorem . . . 961
29.1.2 Euler angles . . . 963
29.2 The inertia tensor . . . 964
29.2.1 Definition and principal axes . . . 964
29.2.2 Computing moments of inertia . . . 966
29.3 Euler's equations and free rotation . . . 968
29.3.1 Euler's equations . . . 968
29.3.2 The free symmetric top . . . 969
29.3.3 Stability of rotation about the principal axes . . . 971
29.3.4 Free nutation of the Earth: the Chandler wobble . . . 972
29.4 The heavy top . . . 974
29.4.1 The Lagrange top . . . 974
29.4.2 Integrable cases and the Kovalevskaya top . . . 978
29.4.3 Gyroscopes . . . 979
29.4.4 Precession of the equinoxes and astronomical nutation . . . 981
29.5 Motion in rotating frames . . . 984
29.5.1 The transport theorem and the fictitious forces . . . 984
29.5.2 Absolute rotation: Newton's bucket and Mach's critique . . . 984
29.5.3 Terrestrial evidence of the Earth's rotation . . . 985
29.5.4 The Foucault pendulum . . . 988

30 Continuum Mechanics and Elasticity 991
30.1 The continuum hypothesis . . . 991
30.2 Deformation: the strain tensor . . . 993
30.2.1 Displacement and linear strain . . . 993
30.2.2 Finite strain and compatibility . . . 994
30.3 Stress: the Cauchy tensor . . . 997
30.3.1 Traction and the existence of the stress tensor . . . 997

30.3.2 Momentum balance . . . 998
30.4 Linear elasticity: Hooke's law and the moduli . . . 999
30.4.1 Hooke's law . . . 999
30.4.2 The elastic moduli and their measurement . . . 1001
30.4.3 Anisotropic elasticity . . . 1005
30.5 Elastodynamics and elastic waves . . . 1006
30.5.1 The Navier–Cauchy equations . . . 1006
30.5.2 P and S waves . . . 1007
30.5.3 Surface waves and the seismological evidence . . . 1009
30.6 Rods, beams and plates . . . 1011
30.6.1 The elastica and Euler buckling . . . 1012
30.6.2 Euler–Bernoulli beam theory . . . 1014
30.6.3 Torsion . . . 1015
30.6.4 Plates and shells . . . 1016
30.7 Static problems and contact . . . 1018
30.8 The limits of elasticity . . . 1020

31 Fluid Dynamics **1025**
31.1 Kinematics of flow . . . 1025
31.1.1 The continuum hypothesis and the two descriptions . . . 1025
31.1.2 The continuity equation . . . 1028
31.2 Hydrostatics . . . 1029
31.3 Ideal flow: the Euler equations . . . 1031
31.3.1 The Euler equations . . . 1031
31.3.2 Bernoulli's theorem . . . 1032
31.3.3 Potential flow and the d'Alembert paradox . . . 1034
31.4 Viscous flow: the Navier–Stokes equations . . . 1035
31.4.1 Viscosity . . . 1035
31.4.2 The Navier–Stokes equations . . . 1036
31.4.3 Exact solutions: Couette and Poiseuille flow . . . 1038
31.4.4 Stokes flow . . . 1040
31.4.5 Existence and smoothness: an open problem . . . 1041
31.5 Dynamical similarity . . . 1042
31.5.1 The Reynolds number . . . 1042
31.5.2 Drag laws . . . 1045
31.6 Vorticity . . . 1046
31.6.1 Vortex lines and Helmholtz's theorems . . . 1046
31.6.2 Kelvin's circulation theorem . . . 1048
31.7 Lift and the boundary layer . . . 1049
31.7.1 The Kutta–Joukowski theorem . . . 1049
31.7.2 The Prandtl boundary layer . . . 1051
31.8 Instability and transition . . . 1052
31.8.1 Shear and buoyancy instabilities . . . 1052
31.8.2 Rayleigh–Bénard convection and Taylor–Couette flow . . . 1055
31.9 Turbulence . . . 1057
31.9.1 The Reynolds decomposition and the closure problem . . . 1057
31.9.2 The energy cascade and Kolmogorov's 1941 theory . . . 1058
31.10 Compressible flow . . . 1060
31.10.1 Sound . . . 1060
31.10.2 Shock waves . . . 1061
31.11 Surface waves and solitary waves . . . 1062

32 Nonlinear Dynamics and Chaos **1065**
32.1 Flows in phase space . . . 1066
32.1.1 Dynamical systems, flows and maps . . . 1066
32.1.2 Fixed points and Lyapunov stability . . . 1067
32.1.3 Bifurcations . . . 1070
32.2 Limit cycles . . . 1072
32.2.1 The Poincaré–Bendixson theorem . . . 1072
32.2.2 Relaxation oscillations: van der Pol . . . 1073
32.3 Poincaré sections and discrete dynamics . . . 1076
32.3.1 The return map . . . 1076
32.3.2 The Smale horseshoe and symbolic dynamics . . . 1077
32.4 Hamiltonian chaos . . . 1078
32.4.1 The three-body problem . . . 1078
32.4.2 The KAM theorem . . . 1079
32.4.3 Resonance overlap and the standard map . . . 1080
32.4.4 Numerical discovery: Hénon–Heiles . . . 1082
32.4.5 Chaos in the solar system . . . 1083
32.5 Strange attractors . . . 1086
32.5.1 The Lorenz system . . . 1086
32.5.2 Strange attractors and the Ruelle–Takens proposal . . . 1088
32.5.3 Sensitive dependence and Lyapunov exponents . . . 1089
32.5.4 Fractal dimension . . . 1092
32.6 Universal routes to chaos . . . 1093
32.6.1 The logistic map . . . 1093
32.6.2 Feigenbaum universality . . . 1094
32.6.3 Quasiperiodicity and intermittency . . . 1097
32.7 Experimental chaos . . . 1099
32.7.1 Rayleigh–Bénard convection . . . 1099
32.7.2 Taylor–Couette flow . . . 1100
32.7.3 Circuits and driven oscillators . . . 1101

33 Experiment: The Cavendish Torsion Balance **1103**
33.1 Historical context and the prediction under test . . . 1103
33.1.1 Universal gravitation without a constant . . . 1104
33.1.2 The Schiehallion experiment . . . 1105
33.1.3 Michell's apparatus . . . 1106
33.2 Torsion pendulum theory . . . 1106
33.2.1 The torsion pendulum . . . 1106
33.2.2 From deflection to attraction . . . 1107
33.2.3 Systematics of the method . . . 1109
33.3 Apparatus . . . 1110
33.4 Procedure . . . 1110
33.5 Observations and data . . . 1112
33.6 Interpretation . . . 1113
33.7 Modern determinations and their discrepancies . . . 1115
33.7.1 Classic redeterminations . . . 1115
33.7.2 Modern torsion-balance measurements . . . 1115
33.7.3 Beyond torsion . . . 1116
33.7.4 The discrepancy, honestly stated . . . 1116
33.8 Primary references . . . 1119

34 Experiment: The Pendulum **1121**
34.1 Galileo and Huygens: the isochronous pendulum (1602–1673) 1121
34.1.1 Apparatus 1121
34.1.2 Procedure 1122
34.1.3 Observations and data 1122
34.1.4 Interpretation 1124
34.1.5 Primary references 1125
34.2 Coulomb: the torsion pendulum (1784) 1125
34.2.1 Apparatus 1125
34.2.2 Procedure 1125
34.2.3 Observations and data 1125
34.2.4 Interpretation 1128
34.2.5 Primary references 1130
34.3 Kater: the reversible pendulum (1818) 1130
34.3.1 Apparatus 1130
34.3.2 Procedure 1131
34.3.3 Observations and data 1131
34.3.4 Interpretation 1133
34.3.5 Primary references 1134
34.4 The seconds pendulum and the metre 1134
34.5 Foucault: the pendulum at the Panthéon (1851) 1135
34.5.1 Apparatus 1135
34.5.2 Procedure 1135
34.5.3 Observations and data 1135
34.5.4 Interpretation 1140
34.5.5 Primary references 1140
34.6 The pendulum as the first precision instrument 1140

35 Experiment: Waves and Acoustics **1143**
35.1 Mersenne: the laws of the vibrating string (1636) 1143
35.1.1 Apparatus 1143
35.1.2 Procedure 1144
35.1.3 Observations and data 1144
35.1.4 Interpretation 1145
35.1.5 Primary references 1145
35.2 Melde: standing waves on a driven string (1860) 1145
35.2.1 Apparatus 1145
35.2.2 Procedure 1145
35.2.3 Observations and data 1146
35.2.4 Interpretation 1147
35.2.5 Primary references 1148
35.3 Chladni: the nodal figures of a vibrating plate (1787) 1148
35.3.1 Apparatus 1148
35.3.2 Procedure 1148
35.3.3 Observations and data 1148
35.3.4 Interpretation 1151
35.3.5 Primary references 1151
35.4 Newton, Laplace and the speed of sound (1687–1822) 1151
35.4.1 Apparatus 1151
35.4.2 Procedure 1151
35.4.3 Observations and data 1152
35.4.4 Interpretation 1153

35.4.5 Primary references . . . 1154
35.5 Kundt's tube: sound speeds from dust figures (1866) . . . 1154
35.5.1 Apparatus . . . 1154
35.5.2 Procedure . . . 1154
35.5.3 Observations and data . . . 1154
35.5.4 Interpretation . . . 1155
35.5.5 Primary references . . . 1155
35.6 Buys Ballot: the Doppler effect on a moving train (1845) . . . 1156
35.6.1 Apparatus . . . 1156
35.6.2 Procedure . . . 1156
35.6.3 Observations and data . . . 1156
35.6.4 Interpretation . . . 1157
35.6.5 Primary references . . . 1157
35.7 Summary of the evidence . . . 1158

36 Experiment: Fluid Flow and Turbulence 1159
36.1 Efflux from a vessel: Torricelli's law (1738) . . . 1159
36.1.1 Apparatus . . . 1159
36.1.2 Procedure . . . 1160
36.1.3 Observations and data . . . 1160
36.1.4 Interpretation . . . 1163
36.1.5 Primary references . . . 1163
36.2 Poiseuille: flow in fine capillaries (1840) . . . 1163
36.2.1 Apparatus . . . 1163
36.2.2 Procedure . . . 1164
36.2.3 Observations and data . . . 1164
36.2.4 Interpretation . . . 1165
36.2.5 Primary references . . . 1165
36.3 Reynolds: the dye filament and the critical number (1883) . . . 1165
36.3.1 Apparatus . . . 1166
36.3.2 Procedure . . . 1166
36.3.3 Observations and data . . . 1166
36.3.4 Interpretation . . . 1170
36.3.5 Primary references . . . 1170
36.4 Bénard: cellular convection (1900) . . . 1170
36.4.1 Apparatus . . . 1171
36.4.2 Procedure . . . 1171
36.4.3 Observations and data . . . 1171
36.4.4 Interpretation . . . 1173
36.4.5 Primary references . . . 1173
36.5 Prandtl: the boundary layer made visible (1904) . . . 1174
36.5.1 Apparatus . . . 1174
36.5.2 Procedure . . . 1174
36.5.3 Observations and data . . . 1174
36.5.4 Interpretation . . . 1177
36.5.5 Primary references . . . 1178
36.6 What the record establishes, and what it does not . . . 1178

IV Special Relativity **1181**

37 Experiments: Light, the Aether, and Time **1183**
37.1 Foundations . . . 1183
37.1.1 The vacuum and the aether . . . 1183
37.2 The finite speed of light . . . 1184
37.3 Michelson and Morley: the search for the aether wind (1887) . . . 1184
37.3.1 Apparatus . . . 1184
37.3.2 Procedure . . . 1184
37.3.3 Observations and data . . . 1184
37.3.4 Interpretation . . . 1185
37.3.5 Primary references . . . 1185
37.4 Fizeau: light dragged by moving water (1851) . . . 1185
37.4.1 Apparatus . . . 1185
37.4.2 Procedure . . . 1186
37.4.3 Observations and data . . . 1186
37.4.4 Interpretation . . . 1186
37.4.5 Primary references . . . 1186
37.5 Cavity resonators: the modern Michelson–Morley (1979–2015) . . . 1186
37.5.1 Apparatus . . . 1186
37.5.2 Procedure . . . 1186
37.5.3 Observations and data . . . 1186
37.5.4 Interpretation . . . 1187
37.5.5 Primary references . . . 1187
37.6 The muon lifetime . . . 1187
37.7 What the experiments establish . . . 1187

38 Lorentz Transformations **1189**
38.1 The postulates . . . 1189
38.2 Lorentz boosts . . . 1190
38.2.1 The one-dimensional boost . . . 1190
38.2.2 Minkowski space . . . 1191
38.2.3 The general boost . . . 1194
38.3 Spacetime diagrams . . . 1194
38.4 Consequences of the postulates . . . 1196
38.4.1 Relativity of simultaneity . . . 1196
38.4.2 Time dilation . . . 1197
38.4.3 Length contraction . . . 1198
38.5 Composition of velocities . . . 1199
38.5.1 Rapidity . . . 1200
38.6 General Lorentz transformations . . . 1201
38.6.1 Boosts composed with rotations . . . 1201
38.6.2 Directions normal to the relative velocity . . . 1203

39 Minkowski Space and Its Symmetries **1205**
39.1 Conserved quantities . . . 1205
39.2 The Lorentz group . . . 1205
39.2.1 Foundations . . . 1205
39.2.2 The proper Lorentz group . . . 1207
39.2.3 The improper Lorentz group . . . 1207
39.2.4 The orthochronous Lorentz group . . . 1207
39.2.5 The antichronous Lorentz group . . . 1208

39.2.6 The four components . . . 1208
39.2.7 The algebra of the proper Lorentz group . . . 1209
39.3 The Poincaré group . . . 1210
39.3.1 Foundations . . . 1210
39.3.2 The algebra of the Poincaré group . . . 1211

40 Relativistic Dynamics **1213**
40.1 Four-vectors . . . 1213
40.2 Four-momentum and mass–energy . . . 1214
40.3 The relativistic force law . . . 1215
40.4 Massless particles . . . 1215
40.5 Collisions . . . 1216
40.6 The twin paradox . . . 1217
40.7 Covariant electrodynamics in one line . . . 1217

41 Experiment: Time Dilation and Relativistic Kinematics **1219**
41.1 Ives and Stilwell: the moving atomic clock (1938) . . . 1219
41.1.1 Apparatus . . . 1219
41.1.2 Procedure . . . 1219
41.1.3 Observations and data . . . 1219
41.1.4 Interpretation . . . 1220
41.1.5 Primary references . . . 1220
41.2 Rossi and Hall: muon survival with altitude (1941) . . . 1220
41.2.1 Apparatus . . . 1220
41.2.2 Procedure . . . 1220
41.2.3 Observations and data . . . 1221
41.2.4 Interpretation . . . 1221
41.2.5 Primary references . . . 1221
41.3 Frisch and Smith: Mount Washington to sea level (1963) . . . 1221
41.3.1 Apparatus . . . 1221
41.3.2 Procedure . . . 1221
41.3.3 Observations and data . . . 1222
41.3.4 Interpretation . . . 1222
41.3.5 Primary references . . . 1222
41.4 The CERN muon storage ring (1977) . . . 1222
41.4.1 Apparatus . . . 1222
41.4.2 Procedure . . . 1222
41.4.3 Observations and data . . . 1222
41.4.4 Interpretation . . . 1223
41.4.5 Primary references . . . 1223
41.5 Hafele and Keating: caesium clocks flown around the world (1972) . . . 1223
41.5.1 Apparatus . . . 1223
41.5.2 Procedure . . . 1223
41.5.3 Observations and data . . . 1223
41.5.4 Interpretation . . . 1223
41.5.5 Primary references . . . 1224
41.6 Bertozzi: the speed limit (1964) . . . 1224
41.6.1 Apparatus . . . 1224
41.6.2 Procedure . . . 1224
41.6.3 Observations and data . . . 1224
41.6.4 Interpretation . . . 1224
41.6.5 Primary references . . . 1224

41.7 Summary of the evidence . . . 1224

V General Relativity and Cosmology 1227

42 The Equivalence Principle and Classical Tests 1229
42.1 The universality of free fall . . . 1229
42.1.1 From Galileo to Newton . . . 1229
42.1.2 The Eötvös torsion balance . . . 1230
42.1.3 Modern limits . . . 1230
42.2 The equivalence principle in three strengths . . . 1231
42.2.1 Weak, Einstein and strong forms . . . 1231
42.2.2 Lunar laser ranging . . . 1232
42.3 Gravitational redshift . . . 1234
42.3.1 The prediction . . . 1234
42.3.2 Pound–Rebka and Pound–Snider . . . 1234
42.3.3 Clocks on rockets and in orbit . . . 1235
42.4 The perihelion of Mercury . . . 1236
42.4.1 Le Verrier's anomaly . . . 1236
42.4.2 Einstein's 1915 computation . . . 1237
42.5 The deflection of light . . . 1237
42.5.1 Half and full values . . . 1237
42.5.2 The 1919 eclipse expeditions . . . 1238
42.5.3 Radio interferometry and gravitational lensing . . . 1239
42.6 The Shapiro time delay . . . 1239
42.6.1 Radar ranging to the inner planets . . . 1239
42.6.2 Viking and Cassini . . . 1239
42.7 Relativity in the Global Positioning System . . . 1240
42.8 Beyond the classical tests . . . 1241

43 Geometric Formulation of Gravity 1243
43.1 Geodesics as the trajectories of free fall . . . 1243
43.2 Curvature and tidal forces . . . 1244
43.3 The Einstein–Hilbert action . . . 1244

44 The Einstein Field Equations 1247
44.1 The route to the field equations . . . 1247
44.1.1 Requirements on a relativistic theory of gravity . . . 1247
44.1.2 The November 1915 papers . . . 1248
44.1.3 The Hilbert action . . . 1249
44.2 The stress–energy tensor . . . 1253
44.2.1 Matter models . . . 1253
44.2.2 Conservation and Noether's theorem . . . 1255
44.3 The Bianchi identities . . . 1256
44.4 The Newtonian correspondence . . . 1257
44.4.1 Geodesics and the potential . . . 1257
44.4.2 Fixing the coupling constant . . . 1258
44.5 The linearized theory . . . 1258
44.6 The cosmological constant . . . 1259
44.7 Lovelock's uniqueness theorem . . . 1260
44.8 Energy conditions . . . 1261
44.9 The initial-value problem . . . 1263

44.9.1 Constraints and evolution . . . 1263
44.9.2 Well-posedness . . . 1265
44.10 Exact solutions: an overview . . . 1266

45 Schwarzschild Geometry and Black Holes 1269
45.1 The Schwarzschild solution . . . 1269
45.1.1 Derivation and the 1916 paper . . . 1269
45.1.2 Birkhoff's theorem . . . 1271
45.2 Geodesics and orbits . . . 1272
45.2.1 The effective potential . . . 1272
45.2.2 Circular orbits, the ISCO and the photon sphere . . . 1274
45.2.3 Null geodesics: deflection and delay . . . 1275
45.3 The horizon . . . 1277
45.3.1 Coordinate and curvature singularities . . . 1277
45.3.2 The Kruskal–Szekeres extension . . . 1278
45.4 Gravitational collapse . . . 1279
45.4.1 Oppenheimer–Snyder collapse . . . 1279
45.4.2 The singularity theorems . . . 1281
45.4.3 Cosmic censorship . . . 1282
45.5 Rotating and charged black holes . . . 1282
45.5.1 The Kerr solution . . . 1282
45.5.2 Charge: Reissner–Nordström and Kerr–Newman . . . 1285
45.5.3 The no-hair theorems . . . 1286
45.6 Black-hole mechanics . . . 1287
45.6.1 The four laws . . . 1287
45.6.2 Hawking radiation: prediction without observation . . . 1289
45.7 The shadow . . . 1291

46 Gravitational-Wave Theory 1293
46.1 Linearized gravity and the wave equation . . . 1293
46.2 Transverse-traceless gauge and the two polarizations . . . 1293
46.3 The quadrupole formula . . . 1294
46.4 Beyond linear order . . . 1295

47 Experiment: Gravitational Waves 1297
47.1 The Hulse–Taylor Binary Pulsar (PSR B1913+16) (1974–2016) . . . 1297
47.1.1 Apparatus . . . 1297
47.1.2 Procedure . . . 1297
47.1.3 Observations and data . . . 1297
47.1.4 Interpretation . . . 1298
47.1.5 Primary references . . . 1299
47.2 Direct detection: the instrument . . . 1299
47.3 GW150914 (2015) . . . 1299
47.3.1 Apparatus . . . 1299
47.3.2 Procedure . . . 1300
47.3.3 Observations and data . . . 1300
47.3.4 Interpretation . . . 1300
47.3.5 Primary references . . . 1300
47.4 The multimessenger outlook . . . 1301

48 Evidence-Based Cosmology **1303**
48.1 The expanding universe . . . 1303
48.1.1 Nebular redshifts . . . 1303
48.1.2 The distance ladder . . . 1304
48.1.3 The Hubble–Lemaître law . . . 1304
48.2 The FLRW geometry . . . 1305
48.2.1 Homogeneity, isotropy and the metric . . . 1305
48.2.2 The Friedmann equations . . . 1307
48.2.3 Distances and horizons . . . 1308
48.3 The thermal history . . . 1310
48.3.1 Radiation, matter and Λ eras . . . 1310
48.3.2 Big-Bang nucleosynthesis . . . 1311
48.3.3 The prediction of the relic radiation . . . 1313
48.4 The cosmic microwave background . . . 1313
48.4.1 Discovery . . . 1313
48.4.2 Anisotropies . . . 1314
48.5 Large-scale structure . . . 1315
48.5.1 Growth of perturbations . . . 1315
48.5.2 Baryon acoustic oscillations . . . 1317
48.5.3 Evidence for dark matter . . . 1318
48.6 The accelerating universe . . . 1319
48.6.1 Type Ia supernovae . . . 1319
48.6.2 The cosmological constant as dark energy . . . 1320
48.7 The ΛCDM parametrization . . . 1321
48.7.1 Six numbers . . . 1321
48.7.2 The Hubble tension . . . 1321
48.8 Inflation: the evidence status . . . 1323
48.8.1 The proposal . . . 1323
48.8.2 Measured and unmeasured . . . 1324

49 Compact Stars and Relativistic Astrophysics **1325**
49.1 White dwarfs . . . 1325
49.1.1 The observational puzzle: Sirius B . . . 1325
49.1.2 Electron degeneracy pressure . . . 1326
49.1.3 The Chandrasekhar mass limit . . . 1329
49.1.4 White-dwarf cooling and ages . . . 1330
49.2 Neutron stars . . . 1331
49.2.1 The Baade–Zwicky hypothesis . . . 1331
49.2.2 The Tolman–Oppenheimer–Volkoff equations . . . 1332
49.2.3 Birth: core collapse and supernova neutrinos . . . 1334
49.2.4 The dense-matter equation of state . . . 1335
49.3 Pulsars . . . 1336
49.3.1 Discovery and identification . . . 1336
49.3.2 The pulsar population . . . 1338
49.3.3 Binary pulsars: precision gravity . . . 1338
49.4 Maximum masses and radii . . . 1339
49.4.1 Bounds on the maximum mass . . . 1339
49.4.2 Radius measurements: X-ray pulse-profile modelling . . . 1340
49.4.3 Tidal deformability: GW170817 . . . 1340
49.5 The black-hole endpoint . . . 1342

50 Experiment: Black-Hole Observations **1343**
50.1 Historical context and the prediction under test . . . 1343
50.1.1 From invisible companions to relativistic collapse . . . 1343
50.1.2 What general relativity predicts . . . 1344
50.2 Apparatus . . . 1344
50.2.1 X-ray observatories . . . 1344
50.2.2 Adaptive optics and interferometry on the Galactic Centre . . . 1345
50.2.3 The Event Horizon Telescope . . . 1345
50.3 Procedure . . . 1346
50.4 Observations and data . . . 1346
50.4.1 Cygnus X-1 and the X-ray binaries . . . 1346
50.4.2 Sgr A*: stellar orbits . . . 1348
50.4.3 Event-horizon-scale imaging . . . 1350
50.4.4 Black-hole masses from gravitational waves . . . 1352
50.5 Interpretation . . . 1352
50.6 Modern repetitions and precision . . . 1354
50.7 Primary references . . . 1355

51 Experiment: The Cosmic Microwave Background **1357**
51.1 Historical context and the prediction under test . . . 1357
51.1.1 The hot-big-bang prediction . . . 1357
51.1.2 Precursors . . . 1358
51.2 Apparatus . . . 1358
51.2.1 Radiometers and the Holmdel horn . . . 1358
51.2.2 Satellites: COBE, WMAP, Planck . . . 1359
51.2.3 Balloons and ground stations . . . 1359
51.3 Procedure . . . 1360
51.4 Observations and data . . . 1361
51.4.1 Discovery . . . 1361
51.4.2 The blackbody spectrum . . . 1362
51.4.3 The dipole . . . 1363
51.4.4 Primary anisotropies . . . 1364
51.4.5 The acoustic peaks . . . 1365
51.4.6 Polarization . . . 1366
51.4.7 Secondary anisotropies . . . 1367
51.5 Interpretation . . . 1368
51.6 Modern repetitions and precision . . . 1369
51.7 Primary references . . . 1370

52 Stellar Structure and Nucleosynthesis **1371**
52.1 The equations of stellar structure . . . 1371
52.1.1 Hydrostatic equilibrium and timescales . . . 1371
52.1.2 Polytropes and the Lane–Emden equation . . . 1373
52.1.3 Energy transport: radiation and convection . . . 1375
52.1.4 The mass–luminosity relation and the Eddington limit . . . 1377
52.2 Nuclear energy generation . . . 1379
52.2.1 Tunnelling and the Gamow peak . . . 1379
52.2.2 The pp chains . . . 1380
52.2.3 The CNO cycle . . . 1381
52.2.4 Helium burning and beyond . . . 1382
52.3 The Hertzsprung–Russell diagram and stellar evolution . . . 1384
52.3.1 The diagram . . . 1384

52.3.2 Main-sequence lifetimes . . . 1384
52.3.3 Before and after the main sequence . . . 1386
52.4 Stellar nucleosynthesis . . . 1387
52.4.1 The abundance record . . . 1387
52.4.2 B^2FH and the eight processes . . . 1387
52.4.3 Direct evidence . . . 1389
52.5 Solar neutrinos . . . 1390
52.5.1 The chlorine experiment and the solar-neutrino problem . . . 1390
52.5.2 Resolution: flavour transformation . . . 1391
52.5.3 Complete spectroscopy of solar fusion . . . 1391
52.6 Helioseismology . . . 1392
52.6.1 The five-minute oscillations . . . 1392
52.6.2 The seismic Sun . . . 1393

53 Experiment: The Classical Tests of General Relativity 1395
53.1 Le Verrier and the perihelion of Mercury (1859) . . . 1395
53.1.1 Apparatus . . . 1395
53.1.2 Procedure . . . 1396
53.1.3 Observations and data . . . 1396
53.1.4 Interpretation . . . 1397
53.1.5 Primary references . . . 1398
53.2 Dyson, Eddington and Davidson: the eclipse of 1919 . . . 1398
53.2.1 Apparatus . . . 1398
53.2.2 Procedure . . . 1398
53.2.3 Observations and data . . . 1399
53.2.4 Interpretation . . . 1400
53.2.5 Primary references . . . 1400
53.3 Pound, Rebka and Snider: redshift in a tower (1960–1965) . . . 1400
53.3.1 Apparatus . . . 1401
53.3.2 Procedure . . . 1401
53.3.3 Observations and data . . . 1401
53.3.4 Interpretation . . . 1402
53.3.5 Primary references . . . 1402
53.4 Shapiro: the radar time delay (1964–2003) . . . 1402
53.4.1 Apparatus . . . 1403
53.4.2 Procedure . . . 1403
53.4.3 Observations and data . . . 1403
53.4.4 Interpretation . . . 1404
53.4.5 Primary references . . . 1404
53.5 The Eötvös lineage and MICROSCOPE (1922–2022) . . . 1405
53.5.1 Apparatus . . . 1405
53.5.2 Procedure . . . 1405
53.5.3 Observations and data . . . 1406
53.5.4 Interpretation . . . 1407
53.5.5 Primary references . . . 1407
53.6 Lunar laser ranging and Gravity Probe B (1969–2011) . . . 1408
53.6.1 Apparatus . . . 1408
53.6.2 Procedure . . . 1408
53.6.3 Observations and data . . . 1408
53.6.4 Interpretation . . . 1410
53.6.5 Primary references . . . 1410
53.7 Summary of the evidence . . . 1410

VI Classical Thermodynamics **1413**

54 Classical Thermodynamics **1415**
54.1 Systems, equilibrium and temperature . . . 1415
54.1.1 State variables and equilibrium . . . 1415
54.1.2 The zeroth law and empirical temperature . . . 1415
54.1.3 Thermometry and the kelvin . . . 1416
54.2 Equations of state . . . 1416
54.2.1 The gas laws . . . 1416
54.2.2 The ideal-gas equation of state . . . 1416
54.3 The first law . . . 1417
54.3.1 Work, heat and internal energy . . . 1417
54.3.2 Heat capacities . . . 1418
54.3.3 Processes and cycles . . . 1418
54.4 The second law . . . 1418
54.4.1 Carnot's theorem . . . 1418
54.4.2 The Clausius and Kelvin statements . . . 1418
54.4.3 Thermodynamic temperature and the Carnot efficiency . . . 1419
54.4.4 Entropy and the Clausius inequality . . . 1419
54.4.5 Carathéodory's axiomatics . . . 1420
54.5 Thermodynamic potentials . . . 1420
54.5.1 Legendre transforms and the four potentials . . . 1420
54.5.2 Maxwell relations . . . 1420
54.5.3 Extremum principles, stability and the chemical potential . . . 1420
54.6 Phase equilibria . . . 1420
54.6.1 Coexistence and the Clausius–Clapeyron equation . . . 1420
54.6.2 The Gibbs phase rule . . . 1420
54.6.3 Real gases: van der Waals and the critical point . . . 1420
54.6.4 The Joule–Thomson effect . . . 1421
54.7 The third law . . . 1421
54.8 The bridge to the molecular theory . . . 1422

55 Kinetic Theory of Gases **1423**
55.1 The kinetic hypothesis . . . 1423
55.1.1 From Bernoulli to Clausius . . . 1423
55.1.2 Pressure from molecular impacts . . . 1423
55.1.3 The mean free path . . . 1423
55.2 The Maxwell velocity distribution . . . 1424
55.2.1 Maxwell's derivation . . . 1424
55.2.2 Molecular-beam verification . . . 1424
55.3 Equipartition and specific heats . . . 1425
55.3.1 The equipartition theorem . . . 1425
55.3.2 Specific heats and the classical failure . . . 1425
55.4 The Boltzmann equation . . . 1426
55.4.1 The transport equation . . . 1426
55.4.2 The H-theorem . . . 1426
55.4.3 Reversibility and recurrence . . . 1426
55.4.4 The Maxwell–Boltzmann distribution . . . 1426
55.5 Transport phenomena . . . 1426
55.5.1 Viscosity . . . 1426
55.5.2 Heat conduction . . . 1427
55.5.3 Diffusion . . . 1427

55.5.4 Chapman–Enskog and the road to hydrodynamics . . . 1427
55.6 Effusion and rarefied gases . . . 1428
55.6.1 Graham's law of effusion . . . 1428
55.6.2 The Knudsen regime . . . 1428
55.7 Molecular reality and Avogadro's constant . . . 1428
55.7.1 Loschmidt's count . . . 1428
55.7.2 From estimates to an exact constant . . . 1429

56 Experiment: The Mechanical Equivalent of Heat 1431
56.1 Historical context: heat as substance or motion . . . 1431
56.1.1 Rumford's cannon-boring observations . . . 1431
56.1.2 Mayer's inference . . . 1432
56.1.3 Joule's programme, 1843–1849 . . . 1432
56.2 The claim under test . . . 1432
56.3 Apparatus . . . 1433
56.4 Procedure . . . 1433
56.5 Observations . . . 1433
56.6 Interpretation . . . 1434
56.7 The Joule–Thomson companion experiments . . . 1434
56.8 Modern repetitions and precision . . . 1435

57 Experiment: The Kinetic Theory Tested 1437
57.1 Stern: the direct measurement of molecular speeds (1920) . . . 1437
57.1.1 Apparatus . . . 1437
57.1.2 Procedure . . . 1438
57.1.3 Observations and data . . . 1438
57.1.4 Interpretation . . . 1438
57.1.5 Primary references . . . 1438
57.2 Miller and Kusch: the shape of the distribution (1955) . . . 1439
57.2.1 Apparatus . . . 1439
57.2.2 Procedure . . . 1439
57.2.3 Observations and data . . . 1439
57.2.4 Interpretation . . . 1440
57.2.5 Primary references . . . 1440
57.3 Maxwell: the viscosity of air and its density (1866) . . . 1441
57.3.1 Apparatus . . . 1441
57.3.2 Procedure . . . 1441
57.3.3 Observations and data . . . 1441
57.3.4 Interpretation . . . 1441
57.3.5 Primary references . . . 1442
57.4 Graham: diffusion, effusion and the transport coefficients (1833–1846) . . . 1443
57.4.1 Apparatus . . . 1443
57.4.2 Procedure . . . 1443
57.4.3 Observations and data . . . 1443
57.4.4 Interpretation . . . 1443
57.4.5 Primary references . . . 1444
57.5 Loschmidt: the size of an air molecule (1865) . . . 1445
57.5.1 Apparatus . . . 1445
57.5.2 Procedure . . . 1445
57.5.3 Observations and data . . . 1445
57.5.4 Interpretation . . . 1445
57.5.5 Primary references . . . 1446

57.6 Joule and Thomson: the porous plug and the forces between molecules (1852) 1446
57.6.1 Apparatus . . . 1446
57.6.2 Procedure . . . 1447
57.6.3 Observations and data . . . 1447
57.6.4 Interpretation . . . 1447
57.6.5 Primary references . . . 1447
57.7 Summary of the evidence . . . 1448
57.8 Giauque: adiabatic demagnetization and the third law (1933) . . . 1448
57.8.1 Apparatus . . . 1450
57.8.2 Procedure . . . 1450
57.8.3 Observations and data . . . 1450
57.8.4 Interpretation . . . 1451
57.8.5 Primary references . . . 1452

VII Classical Field Theory and Electrodynamics 1455

58 Electrostatics 1457
58.1 The electrostatic field . . . 1457
58.1.1 Fundamentals . . . 1457
58.1.2 Geometry of the field . . . 1460
58.2 Methods for obtaining the potential . . . 1461
58.2.1 Boundary conditions . . . 1461
58.2.2 The method of images . . . 1462
58.2.3 The Poisson equation . . . 1464
58.2.4 Green functions . . . 1465
58.2.5 Multipole expansion . . . 1465
58.3 Electrostatic potential energy . . . 1465
58.4 Charge distributions in external fields . . . 1465
58.4.1 Potential energy of a charge distribution in an external field . . . 1465
58.4.2 Force on a charge distribution in an external field . . . 1465
58.4.3 Torque on a charge distribution in an external field . . . 1465
58.4.4 The ideal dipole . . . 1465
58.5 Electrostatics in macroscopic media . . . 1466

59 Magnetostatics 1467
59.1 Steady currents . . . 1467
59.1.1 Current density and charge conservation . . . 1467
59.1.2 The SI unit of current . . . 1467
59.2 The discovery of electromagnetism . . . 1467
59.2.1 Magnetism before 1820 . . . 1467
59.2.2 Ørsted's experiment . . . 1468
59.2.3 The Biot–Savart law . . . 1469
59.2.4 Ampère's electrodynamics . . . 1469
59.3 The field equations of magnetostatics . . . 1470
59.3.1 Gauss's law for the magnetic field . . . 1470
59.3.2 Ampère's circuital law . . . 1470
59.3.3 Boundary conditions . . . 1470
59.4 The vector potential . . . 1470
59.4.1 Existence and gauge freedom . . . 1470
59.4.2 The Biot–Savart law as the general solution . . . 1471
59.4.3 Multipole expansion and the magnetic dipole . . . 1471

59.4.4 Force and torque on a dipole . . . 1471
59.5 Magnetostatics in matter . . . 1471
59.5.1 Magnetization and bound currents . . . 1471
59.5.2 The auxiliary field $\boldsymbol{H}$ . . . 1472
59.5.3 Permanent magnets and boundary-value problems . . . 1472
59.6 Magnetic monopoles: the empirical status . . . 1472
59.6.1 The Dirac quantization argument . . . 1472
59.6.2 Searches and the null record . . . 1472

60 The Maxwell Equations 1473
60.1 Electromagnetic induction . . . 1473
60.1.1 Faraday's discovery . . . 1473
60.1.2 Lenz's law . . . 1474
60.1.3 The Maxwell–Faraday equation . . . 1474
60.2 The displacement current and the complete system . . . 1474
60.2.1 The inconsistency of the static circuital law . . . 1474
60.2.2 Maxwell's dynamical theory . . . 1474
60.2.3 The Heaviside vector form . . . 1475
60.3 Potentials and gauge freedom . . . 1475
60.3.1 Scalar and vector potentials . . . 1475
60.3.2 Gauge transformations . . . 1475
60.4 Conservation laws . . . 1475
60.4.1 Charge conservation . . . 1475
60.4.2 Poynting's theorem . . . 1476
60.4.3 The stress tensor and field momentum . . . 1476
60.5 The electromagnetic wave prediction . . . 1477
60.5.1 The wave equation and the speed c . . . 1477
60.5.2 Hertz's confirmation . . . 1477
60.6 Structure of the theory . . . 1478
60.6.1 Relativity of the field decomposition . . . 1478
60.6.2 Duality and the absence of magnetic charge . . . 1478
60.6.3 Limits and the domain of validity . . . 1478

61 Electromagnetic Waves and Optics 1479
61.1 Plane electromagnetic waves . . . 1479
61.1.1 Monochromatic plane waves . . . 1479
61.1.2 Energy and momentum transport . . . 1479
61.2 The speed of light . . . 1479
61.2.1 Astronomical determinations . . . 1479
61.2.2 Terrestrial measurements . . . 1480
61.2.3 From measured constant to defined constant . . . 1481
61.3 Reflection and refraction . . . 1481
61.3.1 Snell's law and Fermat's principle . . . 1481
61.3.2 The Fresnel equations . . . 1482
61.4 Polarization . . . 1482
61.4.1 Discovery and phenomenology . . . 1482
61.4.2 Mathematical description . . . 1483
61.4.3 Birefringence and wave plates . . . 1483
61.5 Interference . . . 1483
61.5.1 Young's double slit . . . 1483
61.5.2 Thin films and interferometers . . . 1484
61.5.3 Coherence . . . 1484

61.6 Diffraction . . . 1484
61.6.1 Fresnel diffraction and the Arago spot . . . 1484
61.6.2 Fraunhofer diffraction and gratings . . . 1484
61.6.3 Kirchhoff's scalar theory . . . 1485
61.7 Geometrical optics . . . 1485
61.7.1 The eikonal limit . . . 1485
61.7.2 Images and instruments . . . 1485
61.8 Dispersion . . . 1485
61.8.1 The prism and the solar spectrum . . . 1485
61.8.2 The rainbow . . . 1486
61.8.3 Phase velocity, group velocity, and media . . . 1486

62 Electrodynamics in Matter 1487
62.1 From microscopic to macroscopic fields . . . 1487
62.1.1 Bound charges and the polarization field . . . 1487
62.1.2 Bound currents and the magnetization field . . . 1487
62.1.3 Constitutive relations and boundary conditions . . . 1487
62.2 Dielectrics . . . 1488
62.2.1 Polarizability and local fields . . . 1488
62.2.2 Polar molecules and relaxation . . . 1488
62.2.3 Piezo- and ferroelectricity . . . 1488
62.3 Conduction . . . 1488
62.3.1 Ohm's law . . . 1488
62.3.2 The Drude model and the Hall effect . . . 1489
62.3.3 The skin effect . . . 1489
62.4 Magnetic materials . . . 1489
62.4.1 Dia-, para- and ferromagnetism . . . 1489
62.4.2 Langevin theory and the Curie–Weiss law . . . 1490
62.4.3 The Bohr–van Leeuwen theorem . . . 1491
62.4.4 Superconductors: deferred . . . 1491
62.5 Dispersion and causality . . . 1491
62.5.1 The Lorentz oscillator model . . . 1491
62.5.2 The Kramers–Kronig relations . . . 1492
62.5.3 Signal velocity and the front . . . 1492
62.6 Free-carrier response and plasmas . . . 1492
62.6.1 Plasma frequency and transparency . . . 1492
62.7 Magneto- and electro-optics . . . 1493

63 Experiments: Electromagnetism 1495
63.1 The predictions under test . . . 1495
63.2 Coulomb: the torsion balance (1785) . . . 1495
63.2.1 Apparatus and procedure . . . 1495
63.2.2 Observations . . . 1496
63.2.3 Interpretation . . . 1496
63.3 Cavendish: the concentric spheres . . . 1496
63.4 Galvanic circuits: Ohm and Kirchhoff . . . 1496
63.5 Ørsted: the current and the needle (1820) . . . 1497
63.6 Ampère: the force between currents (1820) . . . 1497
63.7 Faraday: electromagnetic induction (1831–1832) . . . 1497
63.7.1 Apparatus and procedure . . . 1497
63.7.2 Observations . . . 1497
63.7.3 Interpretation . . . 1498

63.8 Weber and Kohlrausch: the ratio of the units (1856) . . . 1498
63.8.1 Convection currents: Rowland (1876) . . . 1498
63.9 Hertz: electromagnetic waves (1888) . . . 1499
63.9.1 Apparatus and procedure . . . 1499
63.9.2 Observations . . . 1499
63.9.3 Interpretation . . . 1500
63.10 Millikan: the quantization of charge (1911–1913) . . . 1500
63.10.1 Apparatus and procedure . . . 1500
63.10.2 Observations . . . 1500
63.10.3 Interpretation . . . 1501
63.11 Modern repetitions and precision . . . 1501
63.11.1 The Coulomb exponent . . . 1501
63.11.2 Photon-mass bounds . . . 1501
63.11.3 Charge quantization today . . . 1501

64 Generalized Classical Field Theory 1503
64.1 Fields as dynamical systems . . . 1503
64.1.1 From many degrees of freedom to a continuum . . . 1503
64.1.2 The action functional and the Lagrangian density . . . 1504
64.1.3 Functional derivatives . . . 1504
64.2 The Euler–Lagrange field equations . . . 1504
64.2.1 Variation, boundary terms and natural conditions . . . 1504
64.2.2 The real scalar field: the Klein–Gordon equation . . . 1504
64.2.3 The complex scalar field and internal symmetry . . . 1504
64.3 Symmetries and conservation laws . . . 1504
64.3.1 Noether's first theorem . . . 1504
64.3.2 Poincaré invariance and the ten charges . . . 1505
64.3.3 Internal symmetry and electric charge . . . 1505
64.3.4 Noether's second theorem and local symmetries . . . 1505
64.4 The energy–momentum tensor . . . 1505
64.4.1 The canonical tensor and its defects . . . 1505
64.4.2 The Belinfante–Rosenfeld improvement . . . 1505
64.4.3 The Hilbert tensor and the gravitational source . . . 1505
64.4.4 What is actually measured . . . 1506
64.5 The electromagnetic field as a Lagrangian system . . . 1506
64.5.1 The free-field Lagrangian . . . 1506
64.5.2 Minimal coupling and the gauge principle . . . 1506
64.5.3 Gauge invariance and charge conservation . . . 1506
64.5.4 The electromagnetic energy–momentum tensor . . . 1506
64.5.5 The non-abelian generalization . . . 1506
64.6 Massive vector fields: the Proca theory . . . 1507
64.6.1 The Proca equation . . . 1507
64.6.2 The photon mass as a measured quantity . . . 1507
64.7 Hamiltonian field theory . . . 1507
64.7.1 Momentum density and the Hamiltonian . . . 1507
64.7.2 Poisson brackets for fields . . . 1507
64.7.3 Singular Lagrangians and constraints . . . 1507
64.7.4 Covariant Hamiltonian formulations . . . 1507
64.8 From classical fields to quanta . . . 1508

65 Radiation and Scattering of Electromagnetic Waves **1509**
65.1 Fields of a charge in arbitrary motion . . . 1509
65.1.1 Retarded potentials . . . 1509
65.1.2 The Liénard–Wiechert potentials . . . 1510
65.1.3 The velocity and acceleration fields . . . 1510
65.2 Radiated power . . . 1510
65.2.1 The Larmor formula . . . 1510
65.2.2 The relativistic generalization . . . 1511
65.2.3 Angular distribution and beaming . . . 1511
65.3 Multipole radiation . . . 1511
65.3.1 Electric dipole radiation . . . 1511
65.3.2 Magnetic dipole and electric quadrupole . . . 1512
65.3.3 The multipole expansion of the radiation field . . . 1512
65.4 Antennas . . . 1512
65.4.1 Radiation resistance . . . 1512
65.4.2 Arrays, directivity and reciprocity . . . 1512
65.5 Radiation reaction . . . 1512
65.5.1 The Abraham–Lorentz force . . . 1512
65.5.2 The Lorentz–Abraham–Dirac equation . . . 1512
65.5.3 Runaways, pre-acceleration and the domain of validity . . . 1513
65.5.4 The Landau–Lifshitz reduction of order . . . 1513
65.5.5 Radiation reaction observed . . . 1513
65.6 Scattering . . . 1513
65.6.1 Thomson scattering . . . 1513
65.6.2 Rayleigh scattering and the colour of the sky . . . 1514
65.6.3 Mie scattering . . . 1515
65.6.4 Where the classical account fails: Compton scattering . . . 1515
65.7 Radiation from relativistic charges and in matter . . . 1516
65.7.1 Synchrotron radiation . . . 1516
65.7.2 Bremsstrahlung . . . 1516
65.7.3 Cherenkov radiation . . . 1516
65.7.4 Transition radiation . . . 1517

66 Plasmas and Magnetohydrodynamics **1519**
66.1 The plasma state . . . 1519
66.1.1 Ionization equilibrium . . . 1519
66.1.2 Debye screening and the plasma parameter . . . 1519
66.1.3 Quasi-neutrality and the plasma frequency . . . 1520
66.1.4 Where plasmas are found . . . 1520
66.2 Waves in a plasma . . . 1521
66.2.1 Langmuir waves . . . 1521
66.2.2 Electromagnetic waves and the cutoff . . . 1522
66.2.3 Magnetized waves and ion-acoustic waves . . . 1522
66.3 Kinetic theory and Landau damping . . . 1522
66.3.1 The Vlasov equation . . . 1522
66.3.2 Landau damping . . . 1523
66.3.3 The observation . . . 1523
66.3.4 Collisions and transport . . . 1523
66.4 Single-particle motion . . . 1524
66.4.1 Gyration and the guiding centre . . . 1524
66.4.2 Drifts . . . 1524
66.4.3 Adiabatic invariants and magnetic mirrors . . . 1524

66.4.4 Fermi acceleration . . . 1525
66.5 Magnetohydrodynamics . . . 1525
66.5.1 The MHD equations and their validity . . . 1525
66.5.2 Flux freezing and the magnetic Reynolds number . . . 1525
66.5.3 Alfvén and magnetosonic waves . . . 1525
66.5.4 Equilibria and the pinch . . . 1526
66.6 Magnetic reconnection . . . 1526
66.6.1 The Sweet–Parker layer . . . 1526
66.6.2 Petschek and fast reconnection . . . 1526
66.6.3 Reconnection observed . . . 1526
66.7 Instabilities and turbulence . . . 1527
66.7.1 Ideal MHD instabilities . . . 1527
66.7.2 Microinstabilities and turbulence . . . 1527
66.8 Fusion and confinement . . . 1527
66.8.1 The Lawson criterion . . . 1527
66.8.2 Magnetic confinement . . . 1528
66.8.3 Inertial confinement and ignition . . . 1528
66.9 Space and astrophysical plasmas . . . 1528
66.9.1 The solar wind . . . 1528
66.9.2 The magnetosphere . . . 1529
66.9.3 The dynamo . . . 1529

67 Experiment: Wave Optics 1531
67.1 The predictions under test . . . 1531
67.1.1 Two accounts of light . . . 1531
67.1.2 What separates them . . . 1532
67.2 Young: the double slit (1801–1804) . . . 1532
67.2.1 Apparatus and procedure . . . 1532
67.2.2 Observations . . . 1532
67.2.3 Interpretation . . . 1533
67.3 Fresnel and the Poisson–Arago spot (1818–1819) . . . 1533
67.3.1 Apparatus and procedure . . . 1533
67.3.2 Observations . . . 1533
67.3.3 Interpretation . . . 1533
67.4 The Fresnel–Arago laws and transversality . . . 1533
67.4.1 Apparatus and procedure . . . 1533
67.4.2 Observations . . . 1534
67.4.3 Interpretation . . . 1534
67.5 Diffraction gratings and wavelength metrology . . . 1535
67.5.1 Fraunhofer's gratings . . . 1535
67.5.2 Rowland's ruling engine and concave grating . . . 1535
67.5.3 Resolving power measured . . . 1535
67.5.4 The same experiment with X-rays . . . 1536
67.6 Interferometry . . . 1536
67.6.1 The Michelson interferometer . . . 1536
67.6.2 The metre in wavelengths of light . . . 1536
67.6.3 Light in a moving medium . . . 1537
67.6.4 Rotation: the Sagnac effect . . . 1537
67.7 Stellar interferometry . . . 1538
67.7.1 Michelson's beam interferometer . . . 1538
67.7.2 Intensity interferometry . . . 1538
67.8 Gabor: holography and the recording of phase (1948, 1962) . . . 1538

67.8.1 Apparatus . . . 1539
67.8.2 Procedure . . . 1539
67.8.3 Observations and data . . . 1540
67.8.4 Interpretation . . . 1544
67.8.5 Primary references . . . 1544
67.9 Modern repetitions and precision . . . 1545
67.9.1 One photon at a time . . . 1545
67.9.2 Matter-wave double slits . . . 1545
67.9.3 Beyond the diffraction limit . . . 1545
67.9.4 Interferometers as detectors . . . 1545

VIII The Transition to Quantum Physics 1547

68 Black-Body Radiation and Planck's Hypothesis 1549
68.1 Cavity radiation and Kirchhoff's theorem . . . 1549
68.1.1 Emissive and absorptive power . . . 1549
68.1.2 The cavity as a thermodynamic system . . . 1550
68.2 What classical thermodynamics could settle . . . 1550
68.2.1 The Stefan–Boltzmann law . . . 1550
68.2.2 Wien's displacement law . . . 1551
68.2.3 Wien's distribution law . . . 1552
68.3 The classical field theory and its failure . . . 1552
68.3.1 Counting the modes of a cavity . . . 1552
68.3.2 The Rayleigh–Jeans law . . . 1552
68.3.3 The ultraviolet catastrophe . . . 1552
68.4 The measurements that decided the matter . . . 1553
68.4.1 Realizing a black body . . . 1553
68.4.2 Lummer–Pringsheim and the failure of Wien's law . . . 1553
68.4.3 Rubens–Kurlbaum in the far infrared . . . 1553
68.5 Planck's law . . . 1553
68.5.1 The interpolation of October 1900 . . . 1553
68.5.2 The combinatorial derivation of December 1900 . . . 1554
68.5.3 The 1901 Annalen paper . . . 1555
68.5.4 Limits, moments and consistency . . . 1555
68.6 What exactly was quantized . . . 1555
68.6.1 Energy elements . . . 1555
68.6.2 The reception, 1900–1911 . . . 1555
68.7 Out of the cavity: specific heats . . . 1555
68.7.1 The Einstein solid . . . 1555
68.7.2 Debye's improvement and the phonon . . . 1556
68.8 The statistics behind the law . . . 1556
68.8.1 Bose's derivation . . . 1556
68.8.2 Detailed balance and the Einstein coefficients . . . 1556
68.9 Black bodies in nature . . . 1556
68.9.1 The cosmic microwave background . . . 1556
68.9.2 Stellar spectra and radiation thermometry . . . 1557

69 The Photon: Photoelectric and Compton Effects **1559**
69.1 The photoelectric effect . . . 1559
69.1.1 Discovery: Hertz and Hallwachs . . . 1559
69.1.2 Lenard's measurements and the classical impasse . . . 1560
69.2 The light-quantum hypothesis . . . 1560
69.2.1 Einstein's heuristic viewpoint of 1905 . . . 1560
69.2.2 The photoelectric equation and the work function . . . 1561
69.2.3 Millikan's verification . . . 1561
69.2.4 What the photoelectric effect does not prove . . . 1561
69.3 Momentum and angular momentum of light . . . 1562
69.3.1 Radiation pressure . . . 1562
69.3.2 Angular momentum of light . . . 1562
69.4 Compton scattering . . . 1563
69.4.1 The 1923 experiment . . . 1563
69.4.2 Kinematics and the Compton wavelength . . . 1564
69.4.3 Debye's independent derivation . . . 1564
69.4.4 The Klein–Nishina cross-section . . . 1564
69.5 Coincidence: the last classical alternative . . . 1564
69.5.1 The Bohr–Kramers–Slater proposal . . . 1564
69.5.2 Bothe–Geiger coincidences . . . 1565
69.5.3 Compton–Simon cloud-chamber pairs . . . 1565
69.6 The photon as a countable object . . . 1565
69.6.1 Anticorrelation at a beam splitter . . . 1565
69.6.2 Counting photons . . . 1566
69.7 Creation and annihilation of quanta . . . 1566
69.7.1 Pair production and annihilation . . . 1566
69.7.2 The photon in quantum electrodynamics . . . 1566

70 Atomic Models and Spectra **1567**
70.1 Spectra before any model . . . 1567
70.1.1 The Fraunhofer lines . . . 1567
70.1.2 Spectral analysis . . . 1568
70.1.3 Balmer's formula . . . 1568
70.1.4 The Rydberg formula and the combination principle . . . 1568
70.2 The constituents of the atom . . . 1569
70.2.1 The electron . . . 1569
70.2.2 Thomson's model and its failure . . . 1569
70.3 The nuclear atom . . . 1569
70.3.1 Large-angle alpha scattering . . . 1569
70.3.2 Rutherford's scattering formula . . . 1570
70.3.3 Why the classical nuclear atom cannot exist . . . 1570
70.4 The Bohr model . . . 1571
70.4.1 The 1913 trilogy . . . 1571
70.4.2 What the model gets right . . . 1571
70.4.3 Reduced mass and the isotope shift . . . 1572
70.4.4 Where the model fails . . . 1573
70.5 The old quantum theory . . . 1573
70.5.1 Sommerfeld's elliptic orbits and the fine structure . . . 1573
70.5.2 The correspondence principle . . . 1573
70.5.3 Space quantization . . . 1574
70.6 X-ray spectra and the atomic number . . . 1574
70.6.1 X-rays and their wavelengths . . . 1574

70.6.2 Moseley's law . . . 1574
70.7 Atoms in external fields . . . 1575
70.7.1 The Zeeman effect . . . 1575
70.7.2 The anomalous Zeeman effect . . . 1575
70.7.3 The Stark effect . . . 1576
70.8 Levels made visible . . . 1576

71 Matter Waves **1577**
71.1 The de Broglie hypothesis . . . 1577
71.1.1 The relation . . . 1577
71.1.2 Phase waves and the Bohr condition . . . 1578
71.1.3 The optical–mechanical analogy completed . . . 1578
71.2 Electron diffraction . . . 1578
71.2.1 Elsasser's prediction . . . 1578
71.2.2 Davisson and Germer . . . 1578
71.2.3 Thomson's transmission experiment . . . 1579
71.3 Interference of atoms and molecules . . . 1579
71.3.1 Helium beams: Estermann and Stern . . . 1579
71.3.2 The electron double slit . . . 1580
71.3.3 Building the pattern one particle at a time . . . 1580
71.4 Neutron interferometry . . . 1580
71.4.1 The perfect-crystal interferometer . . . 1580
71.4.2 Gravitationally induced phase . . . 1581
71.4.3 Rotation by a full turn changes the sign . . . 1581
71.4.4 Bound states of a neutron in the Earth's field . . . 1582
71.5 How large can an interfering object be? . . . 1582
71.5.1 Fullerenes . . . 1582
71.5.2 The present limits . . . 1583
71.6 Wave packets . . . 1583
71.6.1 Superposition and localization . . . 1583
71.6.2 Phase velocity, group velocity and spreading . . . 1583
71.6.3 Bandwidth and the uncertainty relation . . . 1584
71.7 From a wavelength to a wave equation . . . 1584
71.7.1 What the hypothesis does not supply . . . 1584
71.7.2 Matter waves as instruments . . . 1584

72 Experiment: Franck–Hertz **1585**
72.1 Historical context and the prediction under test . . . 1585
72.1.1 What the Bohr postulates predict for a collision . . . 1585
72.1.2 What was known in 1913 . . . 1586
72.2 Apparatus . . . 1586
72.2.1 The tube . . . 1586
72.2.2 Electrical arrangement and the systematics it controls . . . 1586
72.3 Procedure . . . 1586
72.4 Observations . . . 1586
72.4.1 The current–voltage curve . . . 1586
72.4.2 The emission line at 253.7 nanometres . . . 1587
72.4.3 Uncertainties and the dataset . . . 1588
72.5 Interpretation . . . 1588
72.5.1 Discrete energy transfer . . . 1588
72.5.2 What the authors concluded, and why it was wrong . . . 1588
72.5.3 Bohr's correction and its experimental confirmation . . . 1589

72.6 Modern repetitions and precision . . . 1589
72.6.1 The neon variant . . . 1589
72.6.2 Fine structure of the mercury curve . . . 1589
72.6.3 Present-day level energies . . . 1590
72.7 What the experiment settles . . . 1590

73 Experiment: The Blackbody Spectrum Measured 1591
73.1 Stefan: the fourth power, from Tyndall's data (1879) . . . 1591
73.1.1 Apparatus . . . 1592
73.1.2 Procedure . . . 1592
73.1.3 Observations and data . . . 1592
73.1.4 Interpretation . . . 1593
73.1.5 Primary references . . . 1593
73.2 Lummer and Kurlbaum: a black body one can build (1898) . . . 1593
73.2.1 Apparatus . . . 1593
73.2.2 Procedure . . . 1594
73.2.3 Observations and data . . . 1594
73.2.4 Interpretation . . . 1595
73.2.5 Primary references . . . 1595
73.3 Lummer and Pringsheim: Wien's law breaks (1899) . . . 1595
73.3.1 Apparatus . . . 1595
73.3.2 Procedure . . . 1595
73.3.3 Observations and data . . . 1596
73.3.4 Interpretation . . . 1597
73.3.5 Primary references . . . 1597
73.4 Rubens and Kurlbaum: residual rays in the far infrared (1900) . . . 1597
73.4.1 Apparatus . . . 1597
73.4.2 Procedure . . . 1597
73.4.3 Observations and data . . . 1598
73.4.4 Interpretation . . . 1599
73.4.5 Primary references . . . 1599
73.5 Radiometric realization of the kelvin (2019) . . . 1599
73.5.1 Apparatus . . . 1600
73.5.2 Procedure . . . 1600
73.5.3 Observations and data . . . 1600
73.5.4 Interpretation . . . 1601
73.5.5 Primary references . . . 1601
73.6 What the measurements settle . . . 1601

74 Experiment: The Photoelectric Effect and Compton Scattering 1603
74.1 Hertz and Hallwachs: ultraviolet light discharges a metal (1887–1888) . . . 1603
74.1.1 Apparatus . . . 1604
74.1.2 Procedure . . . 1604
74.1.3 Observations and data . . . 1604
74.1.4 Interpretation . . . 1605
74.1.5 Primary references . . . 1605
74.2 Lenard: the energy of the electrons ignores the intensity (1902) . . . 1605
74.2.1 Apparatus . . . 1605
74.2.2 Procedure . . . 1606
74.2.3 Observations and data . . . 1606
74.2.4 Interpretation . . . 1607
74.2.5 Primary references . . . 1607

74.3 Millikan: the slope of the line is h/e (1916) . . . 1607
74.3.1 Apparatus . . . 1608
74.3.2 Procedure . . . 1608
74.3.3 Observations and data . . . 1608
74.3.4 Interpretation . . . 1610
74.3.5 Primary references . . . 1610
74.4 Compton: the wavelength of scattered X-rays changes (1923) . . . 1610
74.4.1 Apparatus . . . 1610
74.4.2 Procedure . . . 1610
74.4.3 Observations and data . . . 1611
74.4.4 Interpretation . . . 1612
74.4.5 Primary references . . . 1612
74.5 Bothe and Geiger: the two products appear together (1925) . . . 1612
74.5.1 Apparatus . . . 1612
74.5.2 Procedure . . . 1613
74.5.3 Observations and data . . . 1613
74.5.4 Interpretation . . . 1614
74.5.5 Primary references . . . 1615
74.6 Grangier, Roger and Aspect: a single quantum is not split (1986) . . . 1615
74.6.1 Apparatus . . . 1615
74.6.2 Procedure . . . 1615
74.6.3 Observations and data . . . 1615
74.6.4 Interpretation . . . 1617
74.6.5 Primary references . . . 1617
74.7 What the measurements settle . . . 1617

75 Experiment: Matter Waves Observed 1619
75.1 Davisson and Germer: electron diffraction by a crystal (1927) . . . 1620
75.1.1 Apparatus . . . 1620
75.1.2 Procedure . . . 1620
75.1.3 Observations and data . . . 1620
75.1.4 Interpretation . . . 1621
75.1.5 Primary references . . . 1621
75.2 Thomson and Reid: transmission rings from thin films (1927) . . . 1622
75.2.1 Apparatus . . . 1622
75.2.2 Procedure . . . 1622
75.2.3 Observations and data . . . 1622
75.2.4 Interpretation . . . 1623
75.2.5 Primary references . . . 1623
75.3 Estermann and Stern: helium and hydrogen diffracted (1930) . . . 1623
75.3.1 Apparatus . . . 1623
75.3.2 Procedure . . . 1624
75.3.3 Observations and data . . . 1624
75.3.4 Interpretation . . . 1625
75.3.5 Primary references . . . 1625
75.4 The electron double slit, one electron at a time (1961–1989) . . . 1625
75.4.1 Apparatus . . . 1625
75.4.2 Procedure . . . 1626
75.4.3 Observations and data . . . 1626
75.4.4 Interpretation . . . 1627
75.4.5 Primary references . . . 1627
75.5 Colella, Overhauser and Werner: gravity shifts a phase (1975) . . . 1627

75.5.1 Apparatus . . . 1627
75.5.2 Procedure . . . 1628
75.5.3 Observations and data . . . 1628
75.5.4 Interpretation . . . 1629
75.5.5 Primary references . . . 1629
75.6 Molecule interferometry and the edge of coherence (1999–2019) . . . 1629
75.6.1 Apparatus . . . 1630
75.6.2 Procedure . . . 1630
75.6.3 Observations and data . . . 1630
75.6.4 Interpretation . . . 1631
75.6.5 Primary references . . . 1632
75.7 What the measurements settle . . . 1632

76 Experiment: X-ray Spectra and the Charge of the Nucleus 1635
76.1 Röntgen: a new kind of rays (1895) . . . 1635
76.1.1 Apparatus . . . 1636
76.1.2 Procedure . . . 1636
76.1.3 Observations and data . . . 1636
76.1.4 Interpretation . . . 1637
76.1.5 Primary references . . . 1638
76.2 Barkla: characteristic radiation and the transversality of X-rays (1905–1911) 1638
76.2.1 Apparatus . . . 1638
76.2.2 Procedure . . . 1638
76.2.3 Observations and data . . . 1638
76.2.4 Interpretation . . . 1639
76.2.5 Primary references . . . 1640
76.3 Friedrich, Knipping and von Laue: diffraction by a crystal (1912) . . . 1640
76.3.1 Apparatus . . . 1640
76.3.2 Procedure . . . 1640
76.3.3 Observations and data . . . 1640
76.3.4 Interpretation . . . 1641
76.3.5 Primary references . . . 1641
76.4 The Braggs: the reflection law and the X-ray spectrometer (1913) . . . 1641
76.4.1 Apparatus . . . 1641
76.4.2 Procedure . . . 1642
76.4.3 Observations and data . . . 1642
76.4.4 Interpretation . . . 1643
76.4.5 Primary references . . . 1643
76.5 Moseley: the high-frequency spectra of the elements (1913–1914) . . . 1643
76.5.1 Apparatus . . . 1643
76.5.2 Procedure . . . 1643
76.5.3 Observations and data . . . 1643
76.5.4 Interpretation . . . 1644
76.5.5 Primary references . . . 1645
76.6 Siegbahn: precision X-ray spectroscopy (1916–1925) . . . 1645
76.6.1 Apparatus . . . 1645
76.6.2 Procedure . . . 1645
76.6.3 Observations and data . . . 1645
76.6.4 Interpretation . . . 1646
76.6.5 Primary references . . . 1646
76.7 What the X-ray spectra established . . . 1646

IX Quantum Mechanics **1649**

77 The Postulates of Quantum Mechanics **1651**
77.1 The conventional postulates . . . 1651
77.1.1 The state postulate . . . 1651
77.1.2 The observable postulate . . . 1651
77.1.3 The measurement postulate . . . 1652
77.1.4 The probability postulate: the Born rule . . . 1652
77.1.5 The collapse postulate . . . 1652
77.1.6 The time-evolution postulate . . . 1652
77.1.7 The composite-system postulate . . . 1653
77.2 Density-matrix postulates . . . 1653
77.3 On Planck's constant . . . 1653
77.4 Pictures of time evolution . . . 1653
77.4.1 The Schrödinger picture . . . 1653
77.4.2 The Heisenberg picture . . . 1653
77.5 Canonical quantization rules . . . 1653
77.5.1 From the Poisson bracket to the commutator . . . 1653
77.5.2 Linear momentum as a differential operator . . . 1653
77.6 The Heisenberg uncertainty principle . . . 1654
77.6.1 Mean value of an observable . . . 1654
77.6.2 Uncertainty . . . 1654
77.6.3 The generalized uncertainty principle . . . 1654

78 Elementary Quantum Systems **1655**
78.1 The free particle . . . 1655
78.1.1 Plane waves and the continuum . . . 1655
78.1.2 Wave packets, group velocity and spreading . . . 1655
78.1.3 Ehrenfest's theorem . . . 1655
78.2 Potential wells . . . 1656
78.2.1 The infinite square well . . . 1656
78.2.2 The finite square well . . . 1656
78.2.3 The delta potential . . . 1656
78.2.4 The linear potential: the quantum bouncer . . . 1657
78.3 Tunnelling . . . 1657
78.3.1 Steps, barriers and the tunnel effect . . . 1657
78.3.2 Alpha decay . . . 1658
78.3.3 Field emission and the scanning tunnelling microscope . . . 1658
78.4 The harmonic oscillator . . . 1659
78.4.1 Analytic solution . . . 1659
78.4.2 Ladder operators . . . 1659
78.4.3 Zero-point energy . . . 1659
78.4.4 Coherent states . . . 1660
78.5 Periodic potentials: a preview . . . 1660

79 Angular Momentum and Spin **1661**
79.1 The angular-momentum algebra . . . 1661
79.1.1 Rotations and their generators . . . 1661
79.1.2 The spectrum from the algebra alone . . . 1662
79.1.3 Representation matrices and Euler angles . . . 1662
79.1.4 SU(2), SO(3) and the double cover . . . 1662
79.2 Orbital angular momentum . . . 1663

79.2.1 Spherical harmonics . . . 1663
79.2.2 Central potentials and the radial reduction . . . 1663
79.2.3 Rotational spectra of molecules . . . 1663
79.2.4 Parity and selection rules . . . 1664
79.3 Spin . . . 1664
79.3.1 The evidence for an intrinsic angular momentum . . . 1664
79.3.2 Pauli matrices and spinors . . . 1665
79.3.3 Magnetic moments and the g factor . . . 1665
79.3.4 Thomas precession . . . 1665
79.3.5 Spin beyond one half . . . 1666
79.4 Addition of angular momenta . . . 1666
79.4.1 The coupling problem and the Clebsch–Gordan series . . . 1666
79.4.2 Clebsch–Gordan coefficients . . . 1667
79.4.3 Two spins one half: singlet and triplet . . . 1667
79.4.4 $3j$, $6j$ and $9j$ symbols . . . 1667
79.5 Tensor operators and the Wigner–Eckart theorem . . . 1667
79.5.1 Spherical tensor operators . . . 1667
79.5.2 The Wigner–Eckart theorem . . . 1667
79.5.3 What the theorem buys . . . 1667
79.6 Spin dynamics and magnetic resonance . . . 1668
79.6.1 Larmor precession . . . 1668
79.6.2 Rabi's resonance method . . . 1668
79.6.3 Nuclear magnetic resonance . . . 1669
79.6.4 Spin as the elementary two-state system . . . 1669

80 Experiment: Stern–Gerlach 1671
80.1 Historical context, 1921–1922 . . . 1671
80.2 Stern–Gerlach (1921–1922) . . . 1672
80.2.1 Apparatus . . . 1672
80.2.2 Procedure . . . 1673
80.2.3 Observations and data . . . 1673
80.2.4 Interpretation . . . 1674
80.2.5 Primary references . . . 1674
80.3 The phenomenon: space quantization . . . 1674
80.4 Sequential Stern–Gerlach experiments . . . 1675
80.5 Legacy . . . 1675

81 The Hydrogen Atom 1677
81.1 The spectrum before the theory . . . 1677
81.1.1 Balmer's formula and the Rydberg constant . . . 1677
81.1.2 What the old quantum theory got right . . . 1679
81.2 The exact solution . . . 1679
81.2.1 Two bodies, one Coulomb potential . . . 1679
81.2.2 The radial equation and its quantization . . . 1679
81.2.3 Eigenfunctions, orbitals and expectation values . . . 1680
81.2.4 Levels, degeneracy and scales . . . 1680
81.3 Degeneracy and the hidden symmetry . . . 1680
81.3.1 The Runge–Lenz vector and the algebraic solution . . . 1680
81.3.2 SO(4) and Fock's construction . . . 1680
81.4 Fine and hyperfine structure . . . 1681
81.4.1 Relativistic corrections and spin–orbit coupling . . . 1681
81.4.2 The Lamb shift . . . 1681

81.4.3 Hyperfine structure . . . 1681
81.4.4 Hydrogen in external fields . . . 1682
81.5 The 21-centimetre line . . . 1682
81.5.1 A prediction made before the instrument existed . . . 1682
81.5.2 Detection . . . 1682
81.5.3 What the line measures . . . 1682
81.6 Precision hydrogen and the constants . . . 1683
81.6.1 The Rydberg constant and optical frequency metrology . . . 1683
81.6.2 The proton radius . . . 1683
81.6.3 Antihydrogen and the tests it carries . . . 1684
81.7 Hydrogen-like systems . . . 1684
81.7.1 One electron, other cores . . . 1684
81.7.2 Exotic atoms . . . 1684

82 Approximation Methods 1685
82.1 Stationary perturbation theory . . . 1685
82.1.1 The Rayleigh–Schrödinger series . . . 1685
82.1.2 Degenerate and near-degenerate levels . . . 1686
82.1.3 Does the series converge? Usually not . . . 1686
82.2 Perturbation theory at work . . . 1686
82.2.1 The Stark effect . . . 1686
82.2.2 The Zeeman effect . . . 1686
82.2.3 Fine structure, hyperfine structure and isotope shifts . . . 1686
82.2.4 The Hellmann–Feynman and virial theorems . . . 1687
82.3 The variational method . . . 1687
82.3.1 The Rayleigh–Ritz principle . . . 1687
82.3.2 Helium . . . 1687
82.3.3 Self-consistent fields . . . 1687
82.4 The semiclassical approximation . . . 1687
82.4.1 The WKB expansion . . . 1687
82.4.2 Turning points and connection formulae . . . 1688
82.4.3 Semiclassical quantization . . . 1688
82.4.4 Barrier penetration . . . 1688
82.5 Time-dependent perturbation theory . . . 1688
82.5.1 Variation of constants and the interaction picture . . . 1688
82.5.2 Fermi's golden rule . . . 1688
82.5.3 Periodic perturbations and the two-level check . . . 1688
82.5.4 The Born approximation . . . 1689
82.6 Adiabatic and sudden limits . . . 1689
82.6.1 The adiabatic theorem . . . 1689
82.6.2 Berry's geometric phase . . . 1689
82.6.3 Landau–Zener transitions . . . 1689
82.6.4 The sudden approximation . . . 1689
82.6.5 Born–Oppenheimer separation . . . 1690

83 Identical Particles 1691
83.1 Indistinguishability and the exchange operator . . . 1691
83.1.1 What indistinguishability asserts . . . 1691
83.1.2 The permutation group on the Hilbert space . . . 1692
83.1.3 The exchange operator and its eigenvalues . . . 1692
83.1.4 The symmetrization postulate . . . 1692
83.2 Bosons and fermions . . . 1692

83.2.1 Bose–Einstein statistics . . . 1692
83.2.2 Fermi–Dirac statistics and the exclusion principle . . . 1692
83.2.3 Parastatistics, and why nature does not use it . . . 1693
83.2.4 Exchange in an effectively two-dimensional system . . . 1693
83.3 Constructing many-particle states . . . 1693
83.3.1 Slater determinants and permanents . . . 1693
83.3.2 Occupation numbers and second quantization . . . 1693
83.3.3 The self-consistent field . . . 1694
83.4 The exchange interaction . . . 1694
83.4.1 Exchange energy has no classical analogue . . . 1694
83.4.2 Helium: ortho and para . . . 1694
83.4.3 From exchange to magnetic order . . . 1695
83.5 The spin–statistics connection . . . 1695
83.5.1 Statement of the theorem . . . 1695
83.5.2 Experimental tests of the symmetrization postulate . . . 1695
83.6 Two-particle interference . . . 1696
83.6.1 The Hong–Ou–Mandel dip . . . 1696
83.6.2 Intensity correlations and the fermionic counterpart . . . 1697
83.7 Consequences for bulk matter . . . 1698
83.7.1 Stability of matter . . . 1698
83.7.2 Degeneracy pressure . . . 1698

84 Entanglement and Bell Tests 1699
84.1 Entangled states . . . 1699
84.1.1 Tensor products and separability . . . 1699
84.1.2 The Schmidt decomposition . . . 1700
84.1.3 Reduced states, purification and steering . . . 1700
84.2 The EPR argument and the replies . . . 1700
84.2.1 Einstein, Podolsky and Rosen . . . 1700
84.2.2 Verschränkung, the cat, and the amplification argument . . . 1700
84.2.3 Bohm's spin version . . . 1700
84.3 Bell's theorem . . . 1701
84.3.1 Local hidden variables and the 1964 inequality . . . 1701
84.3.2 CHSH: the form that is actually measured . . . 1701
84.3.3 The quantum bound . . . 1702
84.3.4 GHZ: contradiction without inequalities . . . 1703
84.3.5 What a violation excludes — and what it does not . . . 1703
84.4 What entanglement may not do . . . 1703
84.4.1 No-signalling . . . 1703
84.4.2 No-cloning . . . 1704
84.4.3 Monogamy . . . 1704
84.5 Quantifying entanglement . . . 1705
84.5.1 Separability criteria . . . 1705
84.5.2 Entropy of entanglement . . . 1705
84.5.3 Mixed-state measures . . . 1705
84.6 Entanglement as a resource . . . 1705
84.6.1 Teleportation and entanglement swapping . . . 1705
84.6.2 Device-independent protocols . . . 1706
84.7 The experimental record . . . 1706

85 Interpretations (Evidence-Anchored) 1707
85.1 The measurement problem . . . 1707
85.1.1 The Born rule . . . 1707
85.1.2 The von Neumann chain . . . 1708
85.1.3 Three problems, stated separately . . . 1708
85.2 What experiment has already settled . . . 1708
85.2.1 Gleason's theorem . . . 1708
85.2.2 Contextuality . . . 1708
85.2.3 Bell and Leggett–Garg constraints . . . 1708
85.2.4 PBR and the status of the quantum state . . . 1709
85.2.5 Decoherence . . . 1709
85.3 The interpretive programmes . . . 1709
85.3.1 Copenhagen and complementarity . . . 1709
85.3.2 Pilot-wave theory . . . 1709
85.3.3 Everett's relative states . . . 1709
85.3.4 Consistent histories . . . 1710
85.3.5 Epistemic and relational accounts . . . 1710
85.4 Dynamical collapse: interpretation made testable . . . 1710
85.4.1 GRW and continuous spontaneous localization . . . 1710
85.4.2 Gravitationally induced collapse . . . 1710
85.4.3 The experimental bounds . . . 1711
85.5 Adjacent experiments that constrain the discussion . . . 1711
85.5.1 Delayed choice and quantum erasure . . . 1711
85.5.2 Weak measurement . . . 1711
85.6 The ledger . . . 1711

86 Scattering Theory 1713
86.1 What a scattering experiment measures . . . 1713
86.1.1 Cross sections . . . 1713
86.1.2 Kinematics, frames and flux . . . 1713
86.1.3 The scattering amplitude . . . 1714
86.2 The stationary formulation . . . 1714
86.2.1 Asymptotic conditions and scattering states . . . 1714
86.2.2 Green functions and the Lippmann–Schwinger equation . . . 1714
86.2.3 The Born series and the Born approximation . . . 1714
86.3 Partial waves and phase shifts . . . 1714
86.3.1 The partial-wave expansion . . . 1714
86.3.2 Unitarity and the optical theorem . . . 1715
86.3.3 Low energy: scattering length and effective range . . . 1715
86.3.4 Levinson's theorem . . . 1716
86.4 The S-matrix . . . 1716
86.4.1 The S operator and its symmetries . . . 1716
86.4.2 Analyticity, poles and dispersion relations . . . 1716
86.5 Resonances . . . 1716
86.5.1 The Breit–Wigner form . . . 1716
86.5.2 Time delay . . . 1717
86.5.3 Fano profiles and Feshbach resonances . . . 1717
86.6 Coulomb scattering . . . 1717
86.6.1 Rutherford, classically and quantum mechanically . . . 1717
86.6.2 Identical particles: Mott scattering . . . 1719
86.7 Observed effects the formalism must reproduce . . . 1719
86.7.1 The Ramsauer–Townsend effect . . . 1719

86.7.2 Threshold behaviour . . . 1720
86.7.3 High energy: the eikonal and Glauber regime . . . 1721

87 Open Quantum Systems and Decoherence 1723
87.1 States of an open system . . . 1723
87.1.1 The density operator . . . 1723
87.1.2 Reduced states and the partial trace . . . 1724
87.1.3 Proper and improper mixtures . . . 1724
87.1.4 Entropy and mixedness . . . 1724
87.2 Quantum channels . . . 1724
87.2.1 Complete positivity . . . 1724
87.2.2 The Kraus representation . . . 1724
87.2.3 The standard channels . . . 1724
87.2.4 Process tomography . . . 1725
87.3 Master equations . . . 1725
87.3.1 Projection-operator derivations . . . 1725
87.3.2 The Lindblad–Gorini–Kossakowski–Sudarshan generator . . . 1725
87.3.3 Quantum Brownian motion . . . 1725
87.3.4 Unravellings and quantum trajectories . . . 1725
87.3.5 Beyond Markov . . . 1726
87.4 Decoherence . . . 1726
87.4.1 Environment-induced superselection . . . 1726
87.4.2 Pointer states . . . 1726
87.4.3 Decoherence rates . . . 1727
87.4.4 Quantum Darwinism . . . 1727
87.4.5 Thermalization of an isolated system . . . 1727
87.4.6 What decoherence does not settle . . . 1728
87.5 Decoherence observed . . . 1728
87.5.1 Cavity quantum electrodynamics . . . 1728
87.5.2 Engineered reservoirs in ion traps . . . 1729
87.5.3 Matter-wave interferometry . . . 1729
87.5.4 Solid-state qubits . . . 1730
87.6 The quantum Zeno effect . . . 1730
87.6.1 The Misra–Sudarshan theorem . . . 1730
87.6.2 Observations . . . 1730
87.7 Fighting decoherence . . . 1731
87.7.1 Decoherence-free subspaces . . . 1731
87.7.2 Dynamical decoupling . . . 1731
87.7.3 Quantum error correction . . . 1732

88 Atoms and Molecules 1733
88.1 The many-electron problem . . . 1733
88.1.1 The Coulomb Hamiltonian . . . 1733
88.1.2 Antisymmetry and Slater determinants . . . 1733
88.1.3 Helium: exchange and the ortho–para splitting . . . 1734
88.2 Self-consistent field theory . . . 1734
88.2.1 The Hartree method . . . 1734
88.2.2 Hartree–Fock . . . 1735
88.2.3 Correlation energy . . . 1735
88.2.4 Density functional theory . . . 1735
88.3 The Thomas–Fermi model . . . 1735
88.4 The periodic table . . . 1736

88.4.1 The building-up principle . . . 1736
88.4.2 The Madelung rule and its exceptions . . . 1736
88.4.3 Periodic trends and their evidence . . . 1737
88.5 Angular-momentum coupling . . . 1737
88.5.1 LS coupling and atomic terms . . . 1737
88.5.2 Hund's rules . . . 1737
88.5.3 jj coupling and the intermediate regime . . . 1737
88.5.4 Atoms in external fields . . . 1737
88.5.5 Hyperfine structure . . . 1737
88.6 The Born–Oppenheimer separation . . . 1738
88.6.1 The adiabatic approximation . . . 1738
88.6.2 Where the surface fails . . . 1738
88.7 The chemical bond . . . 1739
88.7.1 The one-electron bond . . . 1739
88.7.2 Valence bond theory . . . 1739
88.7.3 Molecular orbitals and LCAO . . . 1739
88.7.4 What the two pictures agree on . . . 1740
88.8 Molecular spectra . . . 1740
88.8.1 Rotational spectra . . . 1740
88.8.2 Vibrational spectra . . . 1741
88.8.3 Electronic transitions . . . 1741
88.8.4 Raman scattering . . . 1741
88.9 Intermolecular forces . . . 1742
88.9.1 London dispersion . . . 1742
88.9.2 Retardation and the Casimir–Polder force . . . 1743
88.9.3 Weakly bound dimers . . . 1743

89 Experiment: Bell Tests 1745
89.1 Historical context and the prediction under test . . . 1745
89.1.1 From EPR to a testable inequality . . . 1745
89.1.2 CHSH: the inequality that is measured . . . 1746
89.1.3 The three loopholes . . . 1746
89.1.4 The first tests . . . 1746
89.2 Apparatus . . . 1747
89.2.1 Atomic cascade sources . . . 1747
89.2.2 Parametric down-conversion . . . 1747
89.2.3 Detectors . . . 1747
89.2.4 Entangled matter systems . . . 1747
89.2.5 Setting generators . . . 1747
89.3 Procedure . . . 1748
89.4 Observations . . . 1748
89.5 Interpretation . . . 1749
89.5.1 What the violations exclude . . . 1749
89.5.2 What remains open . . . 1749
89.5.3 Why a violation is not fragile . . . 1750
89.5.4 The 2022 Nobel citation . . . 1750
89.6 Modern repetitions and precision . . . 1750
89.6.1 Closing the locality loophole . . . 1750
89.6.2 Closing the detection loophole . . . 1751
89.6.3 The loophole-free experiments . . . 1751
89.6.4 Pushing back the setting choice . . . 1751
89.6.5 Other platforms and longer baselines . . . 1751

89.6.6 From test to instrument . . . 1752

90 Experiment: Precision Spectroscopy and Atomic Clocks 1753
90.1 Historical context and the prediction under test . . . 1753
90.1.1 The degeneracy the Dirac equation predicts . . . 1753
90.1.2 The Lamb shift . . . 1754
90.1.3 What a frequency measures . . . 1754
90.2 Apparatus . . . 1755
90.2.1 Resonance methods . . . 1755
90.2.2 Doppler-free laser spectroscopy . . . 1755
90.2.3 Cooling and trapping . . . 1756
90.2.4 Stable lasers and optical cavities . . . 1756
90.2.5 The frequency comb . . . 1756
90.3 Procedure . . . 1757
90.4 Observations . . . 1758
90.5 Interpretation . . . 1758
90.5.1 Tests of quantum electrodynamics . . . 1758
90.5.2 The proton-radius puzzle . . . 1758
90.5.3 Constants, and whether they are constant . . . 1758
90.5.4 Relativity measured with clocks . . . 1759
90.6 Modern repetitions and precision . . . 1760
90.6.1 Optical clocks . . . 1760
90.6.2 Redefining the second . . . 1760
90.6.3 Two-electron and exotic atoms . . . 1760
90.6.4 Antihydrogen . . . 1760
90.6.5 Toward a nuclear clock . . . 1760

91 Experiment: Quantum Tunnelling Observed 1763
91.1 The prediction under test . . . 1763
91.2 Alpha decay and the Geiger–Nuttall relation (1911, 1928) . . . 1764
91.2.1 Apparatus . . . 1764
91.2.2 Procedure . . . 1764
91.2.3 Observations and data . . . 1765
91.2.4 Interpretation . . . 1767
91.2.5 Primary references . . . 1767
91.3 Fowler and Nordheim: cold field emission from metals (1928) . . . 1767
91.3.1 Apparatus . . . 1767
91.3.2 Procedure . . . 1768
91.3.3 Observations and data . . . 1768
91.3.4 Interpretation . . . 1769
91.3.5 Primary references . . . 1769
91.4 Esaki: the tunnel diode and negative differential resistance (1958) . . . 1769
91.4.1 Apparatus . . . 1769
91.4.2 Procedure . . . 1769
91.4.3 Observations and data . . . 1770
91.4.4 Interpretation . . . 1771
91.4.5 Primary references . . . 1771
91.5 Giaever: tunnelling spectroscopy of the superconducting gap (1960) . . . 1771
91.5.1 Apparatus . . . 1771
91.5.2 Procedure . . . 1772
91.5.3 Observations and data . . . 1772
91.5.4 Interpretation . . . 1772

91.5.5 Primary references . . . 1773
91.6 Binnig and Rohrer: the scanning tunnelling microscope (1982) . . . 1773
91.6.1 Apparatus . . . 1773
91.6.2 Procedure . . . 1773
91.6.3 Observations and data . . . 1774
91.6.4 Interpretation . . . 1774
91.6.5 Primary references . . . 1775
91.7 Summary of the evidence . . . 1775

X Relativistic Quantum Mechanics 1777

92 The Klein–Gordon Equation 1779
92.1 The equation . . . 1779
92.1.1 Schrödinger's discarded first attempt . . . 1779
92.1.2 Klein, Gordon, and the equation itself . . . 1780
92.1.3 The mass shell and the Poincaré Casimir . . . 1780
92.1.4 Lorentz covariance and the scalar transformation law . . . 1780
92.2 Negative energies and the indefinite density . . . 1780
92.2.1 Both branches of the dispersion relation . . . 1781
92.2.2 The conserved current and why its density changes sign . . . 1781
92.2.3 The two problems are one problem . . . 1781
92.3 Minimal coupling and the charged scalar . . . 1781
92.3.1 The covariant derivative and gauge invariance . . . 1781
92.3.2 The current as charge, not probability . . . 1782
92.3.3 The Compton effect after Gordon . . . 1782
92.3.4 Exact solutions in external fields . . . 1782
92.4 The Feshbach–Villars form and the nonrelativistic limit . . . 1783
92.4.1 The two-component representation . . . 1783
92.4.2 Recovery of the Schrödinger equation . . . 1783
92.4.3 Relativistic corrections . . . 1783
92.5 Propagation, causality and the propagator . . . 1783
92.5.1 Green functions of the Klein–Gordon operator . . . 1783
92.5.2 The Feynman prescription . . . 1783
92.5.3 Localization and its limits . . . 1783
92.5.4 The Yukawa potential . . . 1784
92.6 Pauli–Weisskopf: the equation is a field . . . 1784
92.6.1 Quantization of the scalar field . . . 1784
92.6.2 Bose statistics and microcausality . . . 1784
92.6.3 What survives of the one-particle reading . . . 1785
92.7 The observed spin-0 sector . . . 1785
92.7.1 Pions and the light pseudoscalars . . . 1785
92.7.2 Pionic and kaonic atoms . . . 1785
92.7.3 The Higgs boson . . . 1786

93 The Dirac Equation 1787
93.1 Construction of the equation . . . 1787
93.1.1 Dirac's requirement: first order in time . . . 1787
93.1.2 The Clifford algebra of the gamma matrices . . . 1788
93.1.3 Representations and the choice of basis . . . 1788
93.1.4 The conserved positive-definite current . . . 1788
93.2 Lorentz covariance and the spinor . . . 1788

93.2.1 The spinor transformation law . . . 1788
93.2.2 Bilinear covariants . . . 1789
93.2.3 Plane-wave solutions and spin projectors . . . 1789
93.3 Spin as a consequence, not an assumption . . . 1789
93.3.1 Orbital angular momentum is not conserved . . . 1789
93.3.2 The gyromagnetic ratio $g = 2$ . . . 1790
93.3.3 The Pauli equation as the nonrelativistic limit . . . 1790
93.4 The hydrogen atom solved exactly . . . 1791
93.4.1 The Dirac–Coulomb spectrum . . . 1791
93.4.2 Fine structure: spin–orbit, Darwin and mass terms . . . 1791
93.4.3 Where the exact spectrum fails: the Lamb shift . . . 1791
93.5 Negative energies, holes and the positron . . . 1792
93.5.1 Hole theory . . . 1792
93.5.2 The prediction of the anti-electron . . . 1792
93.5.3 The Stueckelberg–Feynman reading . . . 1793
93.5.4 Charge conjugation and the Majorana alternative . . . 1793
93.6 Phenomena of the one-particle theory . . . 1793
93.6.1 Zitterbewegung . . . 1793
93.6.2 The Klein paradox . . . 1793
93.6.3 Thomas precession and spin transport . . . 1794
93.7 The systematic nonrelativistic expansion . . . 1794
93.7.1 The Foldy–Wouthuysen transformation . . . 1794
93.7.2 The corrections order by order . . . 1794
93.8 The massless limit and the boundary of the theory . . . 1794
93.8.1 Weyl spinors and chirality . . . 1794
93.8.2 Where relativistic quantum mechanics ends . . . 1794

94 Particles as Poincaré Representations 1795
94.1 Symmetries act unitarily: Wigner's theorem . . . 1795
94.2 The unitary generators of the Poincaré group . . . 1796
94.3 The universal cover $\mathrm{SL}(2,\mathbb{C})$ . . . 1798
94.4 The invariants: mass and spin . . . 1799
94.5 Orbits, little groups, and induced representations . . . 1800
94.6 Massive particles: spin . . . 1803
94.7 Massless particles: helicity . . . 1804
94.8 The particle dictionary . . . 1805

95 Experiment: The Positron 1807
95.1 Historical context and the prediction under test . . . 1807
95.1.1 Dirac's hole theory as a falsifiable prediction . . . 1807
95.1.2 Cosmic rays as the available source . . . 1808
95.1.3 What else the track could have been . . . 1808
95.2 Apparatus . . . 1808
95.2.1 The cloud chamber . . . 1808
95.2.2 The magnet . . . 1808
95.2.3 The lead plate . . . 1808
95.3 Procedure . . . 1808
95.3.1 Expansion, triggering and photography . . . 1808
95.3.2 Track reconstruction and momentum from curvature . . . 1809
95.4 Observations . . . 1809
95.4.1 The photograph of 2 August 1932 . . . 1809
95.4.2 The tabulated quantities . . . 1810

95.4.3 Uncertainties and systematics . . . 1810
95.5 Interpretation . . . 1810
95.5.1 Identification with Dirac's anti-electron . . . 1810
95.5.2 Independent confirmation and pair production . . . 1810
95.5.3 Annihilation radiation . . . 1811
95.5.4 The muon: the same apparatus, a different particle . . . 1811
95.6 Modern repetitions and precision . . . 1812
95.6.1 Positronium . . . 1812
95.6.2 The antiproton and antihydrogen . . . 1812
95.6.3 Precision comparisons of positron and electron . . . 1812
95.6.4 Applied annihilation: positron emission tomography . . . 1813

96 The Weyl Equation and Neutrinos 1815
96.1 The two-component equation . . . 1815
96.1.1 Weyl's equation . . . 1815
96.1.2 Two-component spinors and $\mathrm{SL}(2,\mathbb{C})$ . . . 1815
96.1.3 Chirality and helicity . . . 1816
96.1.4 Relation to the Dirac equation . . . 1816
96.1.5 Discrete symmetries of a Weyl field . . . 1816
96.2 The neutrino hypothesis . . . 1816
96.2.1 The continuous beta spectrum . . . 1816
96.2.2 Pauli's 1930 letter . . . 1817
96.2.3 Fermi's theory of beta decay . . . 1817
96.2.4 Detection: Reines and Cowan . . . 1817
96.2.5 How many neutrinos . . . 1817
96.3 Parity and the reinstatement of the equation . . . 1818
96.3.1 Parity as an assumed symmetry . . . 1818
96.3.2 The measurements . . . 1818
96.3.3 Maximal violation . . . 1818
96.4 The two-component theory of the neutrino . . . 1818
96.4.1 Landau, Lee–Yang and Salam . . . 1818
96.4.2 The measured helicity . . . 1819
96.4.3 The V–A current . . . 1819
96.5 Chirality in the Standard Model . . . 1820
96.5.1 Left-handed doublets, right-handed singlets . . . 1820
96.5.2 The chiral anomaly . . . 1820
96.5.3 Chiral fermions on a lattice . . . 1820
96.6 Mass, and where the massless picture fails . . . 1820
96.6.1 Dirac and Majorana mass terms . . . 1820
96.6.2 Oscillations as the evidence for mass . . . 1820
96.6.3 Absolute mass bounds . . . 1821
96.6.4 Where the mass could come from . . . 1822
96.7 The Majorana question . . . 1822
96.7.1 Majorana's symmetric theory . . . 1822
96.7.2 Neutrinoless double beta decay . . . 1822
96.7.3 Null results and present status . . . 1822
96.8 Weyl fermions as an observed phenomenon . . . 1823

97 Higher-Spin Wave Equations **1825**
97.1 The classification problem . . . 1825
97.1.1 From representations to field equations . . . 1825
97.1.2 What a wave equation must deliver . . . 1826
97.2 Spin 1: the Proca field . . . 1826
97.2.1 The Proca equation . . . 1826
97.2.2 The Duffin–Kemmer form . . . 1826
97.2.3 Stueckelberg's restoration of gauge invariance . . . 1826
97.2.4 Massive vectors in Nature . . . 1826
97.3 Spin 3/2: the Rarita–Schwinger field . . . 1826
97.3.1 The equation and its constraints . . . 1826
97.3.2 The Velo–Zwanziger causality problem . . . 1827
97.3.3 The $\Delta(1232)$ resonance . . . 1827
97.4 Spin 2: the Fierz–Pauli field . . . 1827
97.4.1 The Fierz–Pauli Lagrangian . . . 1827
97.4.2 The vDVZ discontinuity . . . 1827
97.4.3 The Boulware–Deser ghost and the Vainshtein radius . . . 1827
97.4.4 The massless graviton and its evidence . . . 1827
97.5 General spin . . . 1828
97.5.1 Massive fields of arbitrary spin . . . 1828
97.5.2 Massless fields of arbitrary spin . . . 1828
97.5.3 Coupling obstructions . . . 1828
97.6 No-go theorems . . . 1828
97.6.1 Soft theorems and Weinberg's low-energy argument . . . 1828
97.6.2 Weinberg–Witten . . . 1828
97.6.3 Coleman–Mandula . . . 1828
97.7 Experimental status . . . 1829
97.7.1 The observed spectrum . . . 1829
97.7.2 Programmes without evidence . . . 1829

98 Experiment: Neutrino Helicity **1831**
98.1 Historical context and the prediction under test . . . 1831
98.1.1 From parity violation to the two-component neutrino . . . 1831
98.1.2 Why direct neutrino polarimetry is impossible . . . 1831
98.1.3 The chosen decay scheme . . . 1832
98.2 Apparatus . . . 1833
98.2.1 The source . . . 1833
98.2.2 The analysing magnet . . . 1833
98.2.3 Resonant scatterer and detector . . . 1833
98.3 Procedure . . . 1833
98.3.1 Selecting the recoil-aligned photons . . . 1833
98.3.2 Field reversal and counting . . . 1833
98.4 Observations . . . 1833
98.4.1 The counting asymmetry . . . 1833
98.4.2 Systematic checks . . . 1834
98.5 Interpretation . . . 1834
98.5.1 The neutrino is left-handed . . . 1834
98.5.2 Consequences for V–A . . . 1835
98.5.3 What the result does not say . . . 1835
98.6 Modern repetitions and precision . . . 1835
98.6.1 Chirality tests in leptonic decays . . . 1835
98.6.2 Parity violation in electron scattering and atoms . . . 1835

98.6.3 Limits on right-handed currents . . . 1836

XI Quantum Field Theory and the Standard Model 1837

99 Canonical Quantization of Fields 1839
99.1 Why fields . . . 1840
99.1.1 Failure of the single-particle relativistic theories . . . 1840
99.1.2 Field as a mechanical system with continuously many degrees of freedom 1843
99.1.3 The historical sequence . . . 1845
99.2 The free real scalar field . . . 1845
99.2.1 Canonical commutation relations . . . 1845
99.2.2 Mode expansion, creation and annihilation operators . . . 1848
99.2.3 Fock space . . . 1851
99.2.4 Normal ordering and the vacuum energy . . . 1852
99.3 The complex scalar field and antiparticles . . . 1853
99.3.1 Global phase symmetry and conserved charge . . . 1853
99.3.2 Two species of quantum: particle and antiparticle . . . 1855
99.3.3 Charge conjugation . . . 1856
99.4 Noether charges and the Poincaré algebra . . . 1857
99.4.1 Noether's theorem in field form . . . 1857
99.4.2 The Poincaré algebra and the labels of a particle . . . 1858
99.5 Fermions and gauge fields . . . 1859
99.5.1 The Dirac field and anticommutation . . . 1859
99.5.2 The electromagnetic field and gauge redundancy . . . 1862
99.5.3 Massive vector fields and the general spin case . . . 1864
99.6 Propagators . . . 1865
99.6.1 Time ordering and the Feynman propagator . . . 1865
99.6.2 The family of Green functions . . . 1868
99.6.3 Wick's theorem . . . 1870
99.7 Causality and spin–statistics . . . 1871
99.7.1 Microcausality . . . 1871
99.7.2 The spin–statistics theorem . . . 1873
99.7.3 Measurability of the field operators . . . 1875
99.8 Vacuum energy made observable: the Casimir effect . . . 1876
99.8.1 The prediction . . . 1876
99.8.2 Measurements . . . 1878
99.9 Limits of the canonical approach . . . 1879
99.9.1 Haag's theorem . . . 1879
99.9.2 What comes next . . . 1880

100 Quantum Electrodynamics and Renormalization 1883
100.1 The QED Lagrangian . . . 1883
100.1.1 Conventions, and the SI dimension of everything . . . 1883
100.1.2 Local phase invariance and minimal coupling . . . 1885
100.1.3 Gauge fixing and the photon propagator . . . 1887
100.1.4 Discrete symmetries and Furry's theorem . . . 1888
100.2 Perturbation theory and its three formalisms . . . 1889
100.2.1 Tomonaga–Schwinger covariant formulation . . . 1889
100.2.2 Feynman's space-time approach and the diagrams . . . 1890
100.2.3 Dyson's S-matrix and the equivalence of the formalisms . . . 1892
100.3 Tree-level QED . . . 1893

100.3.1 Elementary processes . . . 1893
100.3.2 From amplitudes to observables . . . 1895
100.4 One loop: where the infinities are . . . 1898
100.4.1 Vacuum polarization . . . 1900
100.4.2 Electron self-energy and mass renormalization . . . 1903
100.4.3 The vertex correction and the anomalous magnetic moment . . . 1905
100.4.4 Regularization . . . 1907
100.5 Renormalization . . . 1908
100.5.1 Counterterms and renormalization constants . . . 1908
100.5.2 The Ward–Takahashi identity . . . 1909
100.5.3 All orders: the BPHZ theorem . . . 1910
100.5.4 Scheme dependence and the meaning of a renormalized parameter . . . 1911
100.6 Infrared structure . . . 1912
100.6.1 Soft photons and the Bloch–Nordsieck mechanism . . . 1912
100.6.2 Mass singularities . . . 1914
100.7 The running coupling . . . 1914
100.7.1 Gell-Mann–Low and the beta function . . . 1914
100.7.2 Measured running of α . . . 1916
100.7.3 The Landau pole, stated honestly . . . 1916
100.8 Precision tests . . . 1917
100.8.1 The Lamb shift . . . 1917
100.8.2 The electron anomalous magnetic moment . . . 1919
100.8.3 Light-by-light scattering . . . 1920
100.8.4 Bound-state QED and the fine-structure constant . . . 1922

101 Weak Interactions **1925**
101.1 The beta-decay crisis . . . 1925
101.1.1 The continuous spectrum . . . 1925
101.1.2 Pauli's neutrino . . . 1927
101.1.3 Detection . . . 1928
101.2 Fermi's theory . . . 1930
101.2.1 The four-fermion interaction . . . 1930
101.2.2 Allowed transitions: Fermi and Gamow–Teller . . . 1933
101.2.3 The Kurie plot and the neutrino mass . . . 1936
101.3 Parity violation . . . 1938
101.3.1 The tau–theta puzzle and the Lee–Yang question . . . 1938
101.3.2 Observation . . . 1939
101.3.3 Neutrino helicity . . . 1941
101.4 The $V - A$ interaction . . . 1944
101.4.1 Chirality and the left-handed current . . . 1944
101.4.2 Conserved vector current and universality . . . 1946
101.4.3 Muon decay and the Fermi constant . . . 1949
101.5 Strangeness-changing currents . . . 1955
101.5.1 The strangeness deficit and the Cabibbo rotation . . . 1955
101.5.2 Selection rules . . . 1956
101.5.3 The GIM mechanism . . . 1957
101.5.4 Three generations: the CKM matrix . . . 1958
101.6 Neutral currents and the intermediate bosons . . . 1960
101.6.1 The intermediate vector boson hypothesis . . . 1960
101.6.2 Gargamelle and the weak neutral current . . . 1963
101.6.3 Discovery of the W and Z . . . 1965
101.6.4 Precision electroweak measurements . . . 1966

101.7 Why Fermi theory had to be replaced . . . 1968
101.7.1 Unitarity violation . . . 1968
101.7.2 Non-renormalizability and the effective-theory reading . . . 1969

102 Quantum Chromodynamics 1973
102.1 Hadron spectroscopy and the quark hypothesis . . . 1973
102.1.1 The eightfold way . . . 1973
102.1.2 The Ω^- prediction and its confirmation . . . 1975
102.1.3 Quarks . . . 1976
102.2 Colour . . . 1978
102.2.1 The statistics paradox . . . 1978
102.2.2 Counting the colours . . . 1979
102.3 The QCD Lagrangian . . . 1981
102.3.1 Yang–Mills theory . . . 1981
102.3.2 Gauge fixing, ghosts and the Slavnov–Taylor identities . . . 1983
102.3.3 The Lagrangian and its parameters . . . 1985
102.4 Asymptotic freedom . . . 1987
102.4.1 The sign of the beta function . . . 1987
102.4.2 The measured running of α_s . . . 1991
102.5 Deep inelastic scattering and the parton model . . . 1993
102.5.1 Bjorken scaling and partons . . . 1993
102.5.2 Scaling violation and the DGLAP equations . . . 1995
102.5.3 Factorization and parton distributions . . . 1998
102.6 Jets and the gluon . . . 1999
102.6.1 Two-jet events and the quark spin . . . 1999
102.6.2 Three-jet events: the gluon . . . 2000
102.7 Confinement . . . 2001
102.7.1 Lattice gauge theory . . . 2001
102.7.2 The static potential and hadron masses . . . 2002
102.7.3 No free quark has ever been observed . . . 2004
102.8 Chiral symmetry and its breaking . . . 2005
102.8.1 The chiral limit . . . 2005
102.8.2 Dynamical symmetry breaking . . . 2006
102.8.3 The pion as a pseudo-Goldstone boson . . . 2007
102.9 QCD under extreme conditions . . . 2008
102.10 The strong CP problem . . . 2010

103 Flavour Physics and Neutrinos 2013
103.1 The discovery of new quantum numbers . . . 2014
103.1.1 V particles . . . 2014
103.1.2 Strangeness . . . 2015
103.2 Quark mixing . . . 2016
103.2.1 The Cabibbo angle . . . 2016
103.2.2 GIM and the prediction of charm . . . 2017
103.2.3 The November revolution . . . 2019
103.3 Three generations . . . 2020
103.3.1 Kobayashi and Maskawa . . . 2020
103.3.2 The tau lepton . . . 2021
103.3.3 The bottom quark . . . 2022
103.3.4 The top quark . . . 2023
103.3.5 How many generations . . . 2024
103.4 The CKM matrix . . . 2025

103.4.1 Structure and parametrization . . . 2025
103.4.2 Measured magnitudes . . . 2028
103.4.3 Unitarity triangle . . . 2028
103.5 Neutral meson mixing and CP violation . . . 2030
103.5.1 Neutral kaon mixing . . . 2030
103.5.2 Discovery of CP violation . . . 2032
103.5.3 B mixing and the B factories . . . 2033
103.5.4 Charm and the present picture . . . 2036
103.5.5 What CP violation cannot yet explain . . . 2036
103.6 Lepton flavour and neutrino mixing . . . 2037
103.6.1 The PMNS matrix . . . 2037
103.6.2 Oscillation in vacuum . . . 2039
103.6.3 Matter effects . . . 2042
103.6.4 The evidence . . . 2045
103.6.5 Measured parameters and what is still unknown . . . 2046
103.7 Neutrino mass . . . 2047
103.7.1 Direct kinematic limits . . . 2047
103.7.2 Neutrinoless double beta decay . . . 2048
103.7.3 Cosmological bounds . . . 2050
103.7.4 Dirac or Majorana: an open question . . . 2051
103.8 Open questions in flavour . . . 2052

104 Electroweak Unification and the Higgs Boson 2055
104.1 Conventions carried into this chapter . . . 2055
104.2 From Fermi's theory to a gauge theory . . . 2056
104.2.1 The four-fermion interaction and its breakdown . . . 2056
104.2.2 Yang–Mills fields and the mass problem . . . 2058
104.2.3 Glashow's $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ . . . 2060
104.3 Spontaneous symmetry breaking . . . 2061
104.3.1 The superconducting analogy . . . 2061
104.3.2 Goldstone's theorem . . . 2062
104.3.3 The Anderson–Higgs mechanism . . . 2064
104.3.4 The 1964 papers . . . 2065
104.4 The Glashow–Weinberg–Salam model . . . 2066
104.4.1 The electroweak Lagrangian . . . 2066
104.4.2 Mixing and the weak mixing angle . . . 2070
104.4.3 Fermion masses and the Yukawa couplings . . . 2072
104.4.4 Custodial symmetry and the ρ parameter . . . 2074
104.5 Renormalizability . . . 2077
104.5.1 't Hooft's proofs . . . 2077
104.5.2 Dimensional regularization and gauge cancellations . . . 2080
104.5.3 Anomaly cancellation . . . 2081
104.6 Experimental confirmation . . . 2083
104.6.1 Weak neutral currents . . . 2083
104.6.2 Discovery of the W and Z bosons . . . 2084
104.6.3 LEP, SLD and the Z resonance . . . 2085
104.6.4 Precision fits and the W mass . . . 2087
104.7 The Higgs boson . . . 2087
104.7.1 Search and discovery . . . 2087
104.7.2 Quantum numbers and couplings . . . 2089
104.7.3 The potential and the self-coupling . . . 2091
104.8 What the model does not explain . . . 2092

104.8.1 Vacuum stability . . . 2092
104.8.2 Open questions, stated honestly . . . 2093

105 Discrete Symmetries and CPT 2095
105.1 The three discrete operations . . . 2096
105.1.1 Parity . . . 2096
105.1.2 Charge conjugation . . . 2100
105.1.3 Time reversal . . . 2102
105.2 Parity violation . . . 2106
105.2.1 The tau–theta puzzle and the Lee–Yang question . . . 2106
105.2.2 The experimental collapse of 1957 . . . 2107
105.2.3 Parity violation as a precision tool . . . 2110
105.3 CP violation . . . 2111
105.3.1 The neutral kaon system . . . 2111
105.3.2 The 1964 discovery . . . 2116
105.3.3 The CKM phase . . . 2118
105.3.4 CP violation beyond the kaon . . . 2120
105.4 Direct tests of time reversal . . . 2121
105.4.1 The CPLEAR asymmetry . . . 2121
105.4.2 Time reversal in the B system . . . 2122
105.4.3 Electric dipole moments as T tests . . . 2123
105.5 The CPT theorem . . . 2125
105.5.1 Statement and hypotheses . . . 2125
105.5.2 Consequences and loopholes . . . 2129
105.5.3 Experimental tests . . . 2131
105.6 The baryon asymmetry . . . 2132
105.6.1 Sakharov's conditions . . . 2132
105.6.2 Why the Standard Model fails . . . 2134
105.7 The strong CP problem . . . 2135
105.7.1 The theta term . . . 2135
105.7.2 The experimental bound and its status . . . 2137

106 Path-Integral Quantization 2139
106.1 The path integral in quantum mechanics . . . 2140
106.1.1 Dirac's Lagrangian formulation . . . 2140
106.1.2 Feynman's sum over histories . . . 2141
106.1.3 Time slicing, the measure and the classical limit . . . 2145
106.2 Euclidean formulation . . . 2148
106.2.1 Wick rotation . . . 2148
106.2.2 The Feynman–Kac formula . . . 2149
106.2.3 Reconstruction of the real-time theory . . . 2152
106.3 Functional quantization of fields . . . 2153
106.3.1 Generating functionals . . . 2153
106.3.2 Free fields and the perturbative expansion . . . 2155
106.3.3 Fermions and Grassmann integration . . . 2159
106.3.4 The effective action . . . 2162
106.3.5 The LSZ reduction formula . . . 2165
106.4 Gauge fields . . . 2166
106.4.1 Faddeev–Popov ghosts . . . 2166
106.4.2 BRST symmetry . . . 2169
106.4.3 Slavnov–Taylor identities . . . 2172
106.4.4 The Gribov ambiguity . . . 2173

106.5 The lattice . . . 2174
106.5.1 Wilson's lattice gauge theory . . . 2174
106.5.2 Monte Carlo evidence . . . 2178
106.6 Topology of the gauge vacuum . . . 2180
106.6.1 Instantons . . . 2180
106.6.2 Theta vacua . . . 2182
106.7 Anomalies . . . 2184
106.7.1 The Adler–Bell–Jackiw anomaly . . . 2184
106.7.2 Neutral pion decay as the measurement . . . 2186
106.7.3 Fujikawa's derivation from the measure . . . 2188
106.7.4 Anomaly cancellation . . . 2190

107 Nuclear Forces and Nuclear Structure 2193
107.1 Discovery of the nucleus . . . 2194
107.1.1 Radioactivity . . . 2194
107.1.2 Rutherford scattering . . . 2195
107.1.3 The neutron and the constitution of nuclei . . . 2197
107.2 The nucleon–nucleon interaction . . . 2199
107.2.1 Yukawa's exchange mechanism . . . 2199
107.2.2 The pion . . . 2201
107.2.3 The deuteron . . . 2203
107.2.4 Phenomenological and effective-field-theory potentials . . . 2206
107.3 Nuclear masses and binding . . . 2208
107.3.1 The binding-energy curve . . . 2208
107.3.2 The semi-empirical mass formula . . . 2209
107.3.3 Radii, form factors and the charge distribution . . . 2213
107.4 Nuclear models . . . 2215
107.4.1 The liquid-drop model and the compound nucleus . . . 2216
107.4.2 Magic numbers and the shell model . . . 2217
107.4.3 The collective model . . . 2221
107.4.4 Where the models fail . . . 2222
107.5 Radioactive decay . . . 2223
107.5.1 Alpha decay . . . 2223
107.5.2 Beta decay . . . 2225
107.5.3 Gamma decay and the Mössbauer effect . . . 2228
107.6 Fission and fusion . . . 2231
107.6.1 Fission . . . 2231
107.6.2 Fusion . . . 2234
107.6.3 The heaviest nuclei . . . 2237

108 The Renormalization Group 2239
108.1 Conventions carried into this chapter . . . 2239
108.2 Origins in field theory . . . 2241
108.2.1 The arbitrariness of the subtraction point . . . 2241
108.2.2 The Callan–Symanzik equation . . . 2243
108.2.3 Renormalization schemes and their arbitrariness . . . 2245
108.3 Regularization . . . 2247
108.3.1 Dimensional regularization . . . 2247
108.3.2 Other regulators and the counterterm expansion . . . 2249
108.4 The Wilsonian picture . . . 2251
108.4.1 Block spins and Kadanoff scaling . . . 2251
108.4.2 Integrating out momentum shells . . . 2253

108.4.3 Fixed points, relevant and irrelevant operators . . . 2256
108.4.4 Universality . . . 2257
108.5 Running couplings in the Standard Model . . . 2259
108.5.1 Asymptotic freedom . . . 2260
108.5.2 The measured running . . . 2261
108.5.3 Landau poles and triviality . . . 2263
108.6 Effective field theory . . . 2265
108.6.1 Decoupling . . . 2265
108.6.2 Phenomenological Lagrangians and power counting . . . 2267
108.6.3 The Standard Model as an effective theory . . . 2269
108.7 The renormalization group beyond perturbation theory . . . 2273
108.7.1 Scale invariance and its breaking . . . 2273
108.7.2 Monotonicity of the flow . . . 2275

109 Axiomatic Quantum Field Theory 2277
109.1 Why axioms . . . 2277
109.1.1 Fields as operator-valued distributions . . . 2277
109.1.2 Haag's theorem . . . 2279
109.1.3 What axiomatization is for . . . 2280
109.2 The Wightman axioms . . . 2281
109.2.1 Statement of the axioms . . . 2281
109.2.2 Wightman functions and reconstruction . . . 2283
109.2.3 Analyticity and the Wightman–Hall domain . . . 2286
109.2.4 The Reeh–Schlieder theorem . . . 2288
109.2.5 Borchers classes . . . 2290
109.3 Theorems from the axioms . . . 2290
109.3.1 The PCT theorem . . . 2290
109.3.2 Spin and statistics . . . 2293
109.3.3 The Källén–Lehmann representation . . . 2296
109.3.4 Asymptotic states and LSZ . . . 2298
109.4 Algebraic quantum field theory . . . 2300
109.4.1 Haag–Kastler nets . . . 2300
109.4.2 States, GNS and superselection . . . 2301
109.4.3 Modular theory . . . 2302
109.5 The Euclidean route . . . 2304
109.5.1 Osterwalder–Schrader axioms . . . 2304
109.5.2 Reflection positivity and the lattice . . . 2306
109.6 The operator product expansion . . . 2307
109.7 Constructive results and what is missing . . . 2309
109.7.1 Models below four dimensions . . . 2309
109.7.2 Triviality of the four-dimensional scalar . . . 2309
109.7.3 The Yang–Mills mass gap . . . 2311
109.7.4 What axiomatics does and does not secure . . . 2311

110 Experiment: Parity Violation 2313
110.1 Historical context and the prediction under test . . . 2313
110.1.1 Parity as a conserved quantum number . . . 2313
110.1.2 The tau–theta puzzle . . . 2314
110.1.3 Lee and Yang: parity untested in the weak interaction . . . 2315
110.1.4 The observable and what a measurement must deliver . . . 2315
110.2 Apparatus . . . 2317
110.2.1 The polarized cobalt-60 source . . . 2317

110.2.2 Adiabatic demagnetization and the cryostat . . . 2318
110.2.3 Beta and gamma detection . . . 2320
110.3 Procedure . . . 2321
110.3.1 The run cycle . . . 2321
110.3.2 Interleaving, and why it is not optional . . . 2322
110.3.3 The controls . . . 2322
110.4 Observations and data . . . 2323
110.4.1 Gamma anisotropy as the polarization thermometer . . . 2323
110.4.2 The beta asymmetry and its warm-up decay . . . 2323
110.4.3 Numerical results . . . 2324
110.4.4 Systematic checks . . . 2326
110.5 Interpretation . . . 2326
110.5.1 Maximal violation and the V–A structure . . . 2326
110.5.2 Charge conjugation, CP and CPT . . . 2327
110.5.3 Neutrino helicity . . . 2327
110.6 Immediate confirmation: the pion–muon–electron chain . . . 2329
110.6.1 Garwin, Lederman and Weinrich . . . 2329
110.6.2 Friedman and Telegdi . . . 2331
110.7 Modern repetitions and precision . . . 2332
110.7.1 Neutral currents and the weak mixing angle . . . 2332
110.7.2 Atomic parity violation . . . 2332
110.7.3 Polarized electron scattering . . . 2334
110.7.4 The weak charge of the proton . . . 2336
110.7.5 Consistency and what remains open . . . 2338
110.8 Primary references . . . 2339

111 Experiment: CP Violation 2341
111.1 Historical context and the prediction under test . . . 2342
111.1.1 The neutral kaon system . . . 2342
111.1.2 Regeneration . . . 2343
111.1.3 CP as the symmetry that was supposed to survive . . . 2344
111.1.4 The observable and what a measurement must deliver . . . 2345
111.2 Apparatus . . . 2345
111.2.1 The Brookhaven beam and the decay volume . . . 2345
111.2.2 The two-arm spectrometer . . . 2346
111.3 Procedure . . . 2349
111.3.1 Trigger and event recording . . . 2349
111.3.2 Reconstruction in the mass-angle plane . . . 2349
111.3.3 Backgrounds and the control samples . . . 2350
111.4 Observations and data . . . 2350
111.4.1 The forward peak at the kaon mass . . . 2350
111.4.2 The branching ratio and the parameter epsilon . . . 2351
111.4.3 Charge asymmetry in semileptonic decay . . . 2353
111.4.4 The measured CP-violating observables . . . 2353
111.5 Interpretation . . . 2353
111.5.1 Indirect versus direct violation . . . 2353
111.5.2 The CKM matrix and the Jarlskog invariant . . . 2355
111.6 Direct CP violation in the kaon system . . . 2356
111.6.1 NA31 and the first evidence . . . 2356
111.6.2 NA48 and KTeV . . . 2356
111.6.3 The Standard Model prediction and what it is worth . . . 2358
111.7 CP violation beyond the kaons . . . 2359

111.7.1 The B factories . . . 2359
111.7.2 Neutral B_s mesons at the LHC . . . 2361
111.7.3 Charm . . . 2362
111.7.4 Direct T violation . . . 2363
111.7.5 The unitarity triangle . . . 2364
111.8 Cosmological relevance and open questions . . . 2365
111.8.1 The Sakharov conditions . . . 2365
111.8.2 Insufficiency of Standard Model CP violation . . . 2366
111.8.3 Electric dipole moments and the strong CP problem . . . 2367
111.9 Primary references . . . 2369

112 Experiment: The Electron Anomalous Magnetic Moment 2371
112.1 Historical context and the prediction under test . . . 2372
112.1.1 Dirac's $g = 2$ . . . 2372
112.1.2 The hyperfine anomaly and the Kusch–Foley measurement . . . 2373
112.1.3 Schwinger's $\alpha/2\pi$ . . . 2374
112.1.4 What a measurement must deliver . . . 2375
112.2 The theoretical prediction . . . 2376
112.2.1 The QED series through tenth order . . . 2376
112.2.2 Hadronic and electroweak contributions . . . 2377
112.2.3 The fine-structure constant as an input . . . 2379
112.3 Apparatus . . . 2380
112.3.1 Free-electron precession experiments . . . 2380
112.3.2 The Penning trap . . . 2382
112.3.3 The cylindrical trap and the one-electron quantum cyclotron . . . 2385
112.4 Procedure . . . 2387
112.5 Observations and data . . . 2390
112.5.1 The measured electron anomaly . . . 2390
112.5.2 Positrons and the CPT test . . . 2391
112.6 The muon anomaly . . . 2392
112.6.1 CERN and Brookhaven . . . 2392
112.6.2 Fermilab E989 . . . 2393
112.6.3 Hadronic vacuum polarization: lattice versus data-driven . . . 2394
112.6.4 Status . . . 2396
112.7 Interpretation . . . 2396
112.7.1 Direction one: theory plus α against the measured a_e . . . 2396
112.7.2 Direction two: the measurement plus theory against α . . . 2397
112.7.3 What the agreement bounds . . . 2398
112.7.4 A closing accounting . . . 2399
112.8 Primary references . . . 2399

113 Experiment: The Higgs Boson Discovery 2401
113.1 Historical context and the prediction under test . . . 2402
113.1.1 The Brout–Englert–Higgs mechanism . . . 2402
113.1.2 The one free parameter . . . 2403
113.1.3 Indirect constraints . . . 2404
113.1.4 LEP and Tevatron searches . . . 2406
113.2 Apparatus . . . 2407
113.2.1 The Large Hadron Collider . . . 2407
113.2.2 ATLAS . . . 2409
113.2.3 CMS . . . 2411
113.3 Procedure . . . 2412

113.3.1 Production and decay channels . . . 2412
113.3.2 Event selection and background estimation . . . 2415
113.3.3 Blinding and the statistical procedure . . . 2417
113.4 Observations and data . . . 2418
113.4.1 The diphoton channel . . . 2418
113.4.2 The four-lepton channel . . . 2419
113.4.3 Combined significance, 4 July 2012 . . . 2419
113.4.4 The measured numbers . . . 2420
113.5 Interpretation . . . 2421
113.5.1 The look-elsewhere effect . . . 2421
113.5.2 The mass . . . 2426
113.5.3 Spin and parity . . . 2426
113.5.4 Couplings against mass . . . 2428
113.5.5 The total width . . . 2429
113.5.6 What the discovery does and does not establish . . . 2431
113.6 What is not yet measured . . . 2431
113.6.1 The self-coupling and the shape of the potential . . . 2431
113.6.2 First-generation couplings and rare decays . . . 2432
113.6.3 Vacuum stability . . . 2433
113.7 Primary references . . . 2434

114 Experiment: Neutrino Oscillations 2437
114.1 Historical context and the prediction under test . . . 2438
114.1.1 Neutrinos: detection and flavour . . . 2438
114.1.2 Mixing and oscillation . . . 2439
114.1.3 The standard solar model prediction . . . 2440
114.1.4 The observable and what a measurement must deliver . . . 2441
114.2 Apparatus . . . 2442
114.2.1 Radiochemical detectors . . . 2442
114.2.2 The Homestake chlorine tank . . . 2443
114.2.3 The gallium detectors . . . 2443
114.2.4 Water Cherenkov detectors . . . 2444
114.2.5 Heavy water and the neutral-current channel . . . 2445
114.2.6 Liquid scintillator detectors . . . 2447
114.2.7 Accelerator beams . . . 2448
114.2.8 What all of them have in common . . . 2449
114.3 Procedure . . . 2450
114.3.1 The radiochemical run cycle . . . 2450
114.3.2 Reconstructing a Cherenkov event . . . 2451
114.3.3 Separating the three Sudbury channels . . . 2451
114.3.4 The delayed coincidence and the reactor flux model . . . 2452
114.3.5 The near-to-far ratio . . . 2452
114.3.6 Calibration, blindness and the treatment of systematics . . . 2453
114.4 Observations and data . . . 2454
114.4.1 The solar neutrino deficit . . . 2454
114.4.2 The directional and real-time solar measurements . . . 2455
114.4.3 Atmospheric zenith-angle dependence . . . 2456
114.4.4 Sudbury and the flavour-blind flux . . . 2457
114.4.5 KamLAND and reactor antineutrinos . . . 2459
114.4.6 The third mixing angle . . . 2460
114.4.7 Accelerator beams: disappearance and appearance . . . 2461
114.4.8 The global fit . . . 2462

114.5 Interpretation . . . 2463
114.5.1 What the data force . . . 2463
114.5.2 Matter effects: the MSW mechanism . . . 2463
114.5.3 The PMNS matrix and the measured parameters . . . 2465
114.5.4 What oscillation cannot say . . . 2465
114.6 Modern repetitions and precision . . . 2466
114.6.1 Borexino and the solar fusion chains . . . 2466
114.6.2 Supernova and astrophysical neutrinos . . . 2467
114.6.3 Absolute mass: kinematics, cosmology, double beta decay . . . 2467
114.6.4 Anomalies and sterile searches . . . 2468
114.7 Open questions . . . 2469
114.7.1 The mass ordering . . . 2469
114.7.2 The CP phase . . . 2470
114.7.3 Dirac or Majorana . . . 2470
114.8 Primary references . . . 2470

115 Cosmic Rays and Astroparticle Physics 2473
115.1 Discovery of the cosmic radiation . . . 2474
115.1.1 The penetrating radiation . . . 2474
115.1.2 Naming and the extraterrestrial origin . . . 2476
115.1.3 The latitude effect . . . 2476
115.1.4 The east–west asymmetry . . . 2478
115.1.5 Solar modulation and geomagnetic transients . . . 2479
115.2 Cosmic rays as the first particle accelerator . . . 2480
115.2.1 Track detectors . . . 2480
115.2.2 The positron . . . 2480
115.2.3 The muon . . . 2481
115.2.4 The pion . . . 2483
115.2.5 Strange particles . . . 2484
115.3 The primary spectrum and composition . . . 2485
115.3.1 The all-particle spectrum . . . 2485
115.3.2 The knee . . . 2486
115.3.3 The ankle and the highest energies . . . 2487
115.3.4 Elemental and leptonic composition . . . 2488
115.4 Extensive air showers . . . 2489
115.4.1 Auger's coincidence experiment . . . 2489
115.4.2 Cascade development . . . 2490
115.4.3 Detection techniques . . . 2493
115.4.4 The Cherenkov technique . . . 2494
115.4.5 Depth of maximum and mass composition . . . 2495
115.4.6 Energy reconstruction and its uncertainty . . . 2496
115.5 Acceleration and origin . . . 2498
115.5.1 The supernova-remnant hypothesis . . . 2498
115.5.2 Fermi acceleration . . . 2499
115.5.3 Diffusive shock acceleration . . . 2500
115.5.4 The Hillas criterion and candidate sources . . . 2503
115.5.5 Galactic propagation and confinement . . . 2504
115.6 The end of the spectrum . . . 2505
115.6.1 The Greisen–Zatsepin–Kuz'min prediction . . . 2505
115.6.2 The observed suppression . . . 2507
115.6.3 Anisotropy at the highest energies . . . 2508
115.7 Neutral messengers . . . 2509

115.7.1 Gamma-ray astronomy . . . 2509
115.7.2 Atmospheric neutrinos . . . 2509
115.7.3 Supernova 1987A . . . 2511
115.7.4 High-energy astrophysical neutrinos . . . 2513
115.7.5 Multimessenger astronomy . . . 2514
115.8 Cosmic rays as a physics laboratory . . . 2515
115.8.1 Centre-of-mass energies beyond the colliders . . . 2515
115.8.2 Tests of Lorentz invariance . . . 2516
115.8.3 Muon time dilation as a standing test . . . 2516
115.9 What is settled and what is not . . . 2517

116 Experiment: Deep Inelastic Scattering 2519
116.1 Historical context and the prediction under test . . . 2520
116.1.1 Elastic scattering and the proton form factors . . . 2520
116.1.2 The quark hypothesis . . . 2521
116.1.3 Bjorken scaling . . . 2522
116.1.4 The parton model . . . 2522
116.1.5 The Callan–Gross relation . . . 2523
116.2 Apparatus . . . 2524
116.2.1 The Stanford linear accelerator . . . 2524
116.2.2 The End Station A spectrometers . . . 2524
116.2.3 Targets, monitoring and normalisation . . . 2526
116.3 Procedure . . . 2527
116.3.1 Kinematics of the measurement . . . 2527
116.3.2 The measurement programme . . . 2528
116.3.3 Radiative corrections and the systematic budget . . . 2529
116.4 Observations and data . . . 2530
116.4.1 The weak fall-off of the inelastic cross section . . . 2530
116.4.2 Scaling of the structure functions . . . 2531
116.4.3 The longitudinal-to-transverse ratio . . . 2531
116.4.4 Quark charges from the structure functions . . . 2532
116.4.5 The measured numbers . . . 2533
116.5 Interpretation . . . 2533
116.5.1 Point-like constituents . . . 2533
116.5.2 The missing momentum fraction . . . 2537
116.5.3 Neutrino deep inelastic scattering . . . 2539
116.5.4 The 1990 Nobel Prize . . . 2542
116.6 Scaling violations and the confirmation of QCD . . . 2542
116.6.1 Asymptotic freedom . . . 2542
116.6.2 The evolution equations . . . 2543
116.6.3 Muon-beam measurements . . . 2544
116.6.4 HERA . . . 2544
116.6.5 Parton distributions today . . . 2546
116.7 Extensions and modern precision . . . 2547
116.7.1 The EMC effect . . . 2547
116.7.2 The proton spin . . . 2547
116.7.3 The Drell–Yan process and the sea . . . 2549
116.7.4 Continuing programmes . . . 2549
116.8 Primary references . . . 2550

117 Experiment: Artificial Transmutation and the Instruments of Nuclear Physics **2553**
117.1 Conventions and constants . . . 2553
117.2 Rutherford: the first artificial transmutation (1919) . . . 2554
117.2.1 Apparatus . . . 2554
117.2.2 Procedure . . . 2555
117.2.3 Observations and data . . . 2555
117.2.4 Interpretation . . . 2557
117.2.5 Primary references . . . 2559
117.3 Cockcroft and Walton: lithium split by accelerated protons (1932) . . . 2559
117.3.1 Apparatus . . . 2560
117.3.2 Procedure . . . 2560
117.3.3 Observations and data . . . 2561
117.3.4 Interpretation . . . 2562
117.3.5 Primary references . . . 2566
117.4 Lawrence and Livingston: the cyclotron (1932) . . . 2566
117.4.1 Apparatus . . . 2566
117.4.2 Procedure . . . 2567
117.4.3 Observations and data . . . 2567
117.4.4 Interpretation . . . 2567
117.4.5 Primary references . . . 2570
117.5 Wilson: the expansion cloud chamber (1912) . . . 2570
117.5.1 Apparatus . . . 2570
117.5.2 Procedure . . . 2571
117.5.3 Observations and data . . . 2571
117.5.4 Interpretation . . . 2573
117.5.5 Primary references . . . 2575
117.6 Glaser: the bubble chamber (1952) . . . 2576
117.6.1 Apparatus . . . 2576
117.6.2 Procedure . . . 2576
117.6.3 Observations and data . . . 2576
117.6.4 Interpretation . . . 2578
117.6.5 Primary references . . . 2579
117.7 Charpak: the multiwire proportional chamber (1968) . . . 2580
117.7.1 Apparatus . . . 2580
117.7.2 Procedure . . . 2580
117.7.3 Observations and data . . . 2580
117.7.4 Interpretation . . . 2581
117.7.5 Primary references . . . 2584
117.8 Summary of the evidence . . . 2585

XII Statistical Mechanics **2587**

118 Statistical Mechanics **2589**
118.1 From mechanics to thermodynamics . . . 2589
118.1.1 Why a statistical description . . . 2589
118.1.2 Phase space and Liouville's theorem . . . 2589
118.1.3 Macrostates, microstates and coarse graining . . . 2590
118.2 Entropy and the Boltzmann principle . . . 2590
118.2.1 The combinatorial definition . . . 2590
118.2.2 The relation between entropy and multiplicity . . . 2590

118.2.3 The H-theorem and the reversibility objections . . . 2590
118.2.4 The Gibbs entropy . . . 2590
118.3 Ensembles . . . 2590
118.3.1 The microcanonical ensemble . . . 2590
118.3.2 The canonical ensemble . . . 2591
118.3.3 The grand canonical ensemble . . . 2591
118.3.4 Equivalence of ensembles . . . 2591
118.3.5 Partition functions in practice . . . 2591
118.4 The classical ideal gas . . . 2591
118.4.1 The Maxwell–Boltzmann distribution . . . 2591
118.4.2 Equipartition and its failure . . . 2591
118.4.3 The Gibbs paradox . . . 2592
118.4.4 The Sackur–Tetrode entropy . . . 2592
118.5 The thermodynamic limit . . . 2592
118.5.1 Existence and extensivity . . . 2592
118.5.2 Analyticity and its failure . . . 2592
118.5.3 Long-range forces and non-extensive systems . . . 2592
118.6 The ergodic hypothesis: status . . . 2592
118.6.1 What the hypothesis is asked to do . . . 2592
118.6.2 The ergodic theorems . . . 2593
118.6.3 Systems proved ergodic . . . 2593
118.6.4 Honest status of the foundations . . . 2593
118.6.5 Thermalisation in isolated quantum systems . . . 2593
118.7 Fluctuations . . . 2593
118.7.1 Energy and particle-number fluctuations . . . 2593
118.7.2 Critical opalescence . . . 2593
118.7.3 Thermal noise . . . 2593
118.8 Entropy, information and inference . . . 2594
118.8.1 The Shannon functional . . . 2594
118.8.2 Maximum entropy as inference . . . 2594
118.8.3 Information and thermodynamic cost . . . 2594
118.8.4 Negative absolute temperature . . . 2594
118.8.5 Beyond equilibrium . . . 2594

119 Experiment: Brownian Motion and Avogadro's Number 2595
119.1 Historical context and the prediction under test . . . 2595
119.1.1 Brown's observation . . . 2595
119.1.2 The atomic hypothesis under attack . . . 2596
119.1.3 Einstein's prediction . . . 2596
119.1.4 Smoluchowski and Sutherland . . . 2596
119.1.5 The Langevin equation . . . 2596
119.2 Apparatus . . . 2597
119.2.1 The gamboge and mastic emulsions . . . 2597
119.2.2 Microscope, cell and recording . . . 2597
119.2.3 Calibration of the grain radius . . . 2597
119.3 Procedure . . . 2597
119.3.1 Sedimentation equilibrium . . . 2597
119.3.2 Displacement statistics . . . 2598
119.3.3 Rotational Brownian motion . . . 2599
119.4 Observations . . . 2599
119.4.1 Measured displacements and distributions . . . 2599
119.4.2 Values of Avogadro's constant . . . 2599

119.4.3 Uncertainty budget . . . 2599
119.5 Interpretation . . . 2600
119.5.1 The fluctuation–dissipation content . . . 2600
119.5.2 Why this settled the atomic hypothesis . . . 2600
119.5.3 The second law as a statistical statement . . . 2600
119.5.4 Recognition . . . 2600
119.6 Mathematical aftermath . . . 2600
119.6.1 The Wiener process . . . 2600
119.6.2 Ornstein–Uhlenbeck and the velocity correlation . . . 2600
119.6.3 The Fokker–Planck description . . . 2600
119.7 Modern repetitions and precision . . . 2601
119.7.1 Kappler's torsion balance . . . 2601
119.7.2 Optical tweezers . . . 2601
119.7.3 The instantaneous velocity . . . 2601
119.7.4 Avogadro's constant today . . . 2602

120 Quantum Statistics 2603
120.1 Indistinguishability . . . 2603
120.1.1 Counting states for identical particles . . . 2603
120.1.2 The exclusion principle . . . 2603
120.1.3 The spin–statistics theorem . . . 2604
120.1.4 Composite particles . . . 2604
120.1.5 Experimental limits on violations . . . 2604
120.2 The two distributions . . . 2604
120.2.1 Bose's derivation of the Planck law . . . 2604
120.2.2 The Bose–Einstein distribution . . . 2605
120.2.3 Fermi–Dirac statistics . . . 2605
120.2.4 The classical limit . . . 2605
120.2.5 Grand canonical treatment . . . 2605
120.3 The ideal Bose gas . . . 2605
120.3.1 Einstein's condensation . . . 2605
120.3.2 Critical temperature and condensate fraction . . . 2605
120.3.3 Superfluid helium-4 . . . 2605
120.3.4 The weakly interacting Bose gas . . . 2606
120.3.5 Dilute atomic condensates . . . 2606
120.4 The ideal Fermi gas . . . 2606
120.4.1 Fermi energy and degeneracy pressure . . . 2606
120.4.2 The Sommerfeld expansion . . . 2606
120.4.3 Pauli paramagnetism . . . 2607
120.4.4 Landau diamagnetism and quantum oscillations . . . 2608
120.4.5 Degenerate Fermi gases in the laboratory . . . 2608
120.5 Degenerate matter in astrophysics . . . 2608
120.5.1 White dwarfs . . . 2608
120.5.2 Neutron stars . . . 2609
120.6 Correlations as direct evidence . . . 2609
120.6.1 Photon bunching . . . 2609
120.6.2 Bunching and antibunching compared . . . 2609
120.6.3 Exchange effects in scattering . . . 2609
120.7 Beyond the two statistics . . . 2610
120.7.1 Why three space dimensions admit only two . . . 2610
120.7.2 Quasiparticle statistics in planar systems . . . 2610

121 Experiment: Bose–Einstein Condensation **2611**
121.1 The prediction under test . . . 2611
121.1.1 Bose statistics and Einstein's condensation . . . 2611
121.1.2 The condensation criterion . . . 2612
121.1.3 Seventy years of near misses . . . 2612
121.2 Apparatus . . . 2612
121.2.1 Laser cooling . . . 2612
121.2.2 Sub-Doppler cooling . . . 2612
121.2.3 Magnetic traps . . . 2613
121.2.4 Evaporative cooling . . . 2613
121.2.5 Imaging and thermometry . . . 2613
121.3 Procedure . . . 2613
121.3.1 The JILA rubidium-87 sequence . . . 2613
121.3.2 The MIT sodium-23 sequence . . . 2613
121.3.3 Lithium-7 and attractive interactions . . . 2613
121.4 Observations . . . 2614
121.4.1 The bimodal velocity distribution . . . 2614
121.4.2 Condensate fraction and transition temperature . . . 2614
121.4.3 Collective excitations . . . 2615
121.4.4 Interference of two condensates . . . 2615
121.5 Interpretation . . . 2616
121.5.1 The order parameter and the Gross–Pitaevskii equation . . . 2616
121.5.2 Bogoliubov excitations and superfluidity . . . 2616
121.5.3 What was and was not shown . . . 2616
121.6 Repetitions, extensions and precision . . . 2616
121.6.1 Other species, and hydrogen . . . 2616
121.6.2 Vortices and quantized circulation . . . 2616
121.6.3 Optical lattices . . . 2617
121.6.4 Degenerate Fermi gases and the BEC–BCS crossover . . . 2617
121.6.5 Recognition . . . 2617

122 Quantum Optics and the Photon **2619**
122.1 The quantized radiation field . . . 2619
122.1.1 Modes, Fock states and the vacuum . . . 2619
122.1.2 Einstein A and B coefficients . . . 2619
122.1.3 Is the photon necessary? . . . 2620
122.2 Photon statistics . . . 2620
122.2.1 Photodetection and photocount distributions . . . 2620
122.2.2 Poissonian, super- and sub-Poissonian light . . . 2620
122.2.3 The Hanbury Brown–Twiss experiment . . . 2620
122.3 Coherence . . . 2620
122.3.1 Correlation functions and the Glauber hierarchy . . . 2620
122.3.2 Coherent states . . . 2621
122.3.3 The optical equivalence theorem . . . 2621
122.3.4 Nonclassical light . . . 2621
122.4 Antibunching and single photons . . . 2621
122.4.1 Resonance fluorescence and photon antibunching . . . 2621
122.4.2 Single-photon sources . . . 2621
122.4.3 Two-photon interference . . . 2621
122.5 Squeezed light . . . 2622
122.5.1 Quadratures and the uncertainty ellipse . . . 2622
122.5.2 Generation and detection . . . 2622

122.5.3 Squeezed light in gravitational-wave detectors . . . 2622
122.6 Cavity quantum electrodynamics . . . 2622
122.6.1 The Jaynes–Cummings model . . . 2622
122.6.2 Modified spontaneous emission . . . 2622
122.6.3 Strong coupling and vacuum Rabi splitting . . . 2622
122.6.4 Photon counting without absorption . . . 2623
122.7 The laser . . . 2623
122.7.1 Stimulated emission, inversion and the maser . . . 2623
122.7.2 The optical maser and its realization . . . 2623
122.7.3 Laser linewidth and coherence . . . 2623
122.8 Zewail: femtosecond probing of the transition state (1987) . . . 2623
122.8.1 Apparatus . . . 2624
122.8.2 Procedure . . . 2624
122.8.3 Observations and data . . . 2624
122.8.4 Interpretation . . . 2625
122.8.5 Primary references . . . 2626
122.9 Which-path information and the quantum eraser . . . 2626
122.9.1 Complementarity and the duality relation . . . 2626
122.9.2 The quantum eraser . . . 2626

123 Phase Transitions and Critical Phenomena 2627
123.1 Phases and phase diagrams . . . 2627
123.1.1 Coexistence, triple and critical points . . . 2627
123.1.2 Latent heat and the Clausius–Clapeyron relation . . . 2628
123.1.3 The Ehrenfest classification and its limits . . . 2628
123.2 The liquid–gas transition . . . 2628
123.2.1 Andrews and the critical point . . . 2628
123.2.2 The van der Waals equation and the Maxwell construction . . . 2629
123.2.3 Critical opalescence . . . 2629
123.3 The Ising model . . . 2630
123.3.1 Lenz, Ising and the one-dimensional failure . . . 2630
123.3.2 Existence of a transition in two dimensions . . . 2630
123.3.3 Onsager's exact solution . . . 2630
123.3.4 Lattice gas and the Yang–Lee theory . . . 2630
123.4 Mean-field theory and order parameters . . . 2631
123.4.1 The Weiss molecular field . . . 2631
123.4.2 Landau's expansion of the free energy . . . 2631
123.4.3 Symmetry breaking and the order parameter . . . 2631
123.4.4 Where mean field fails . . . 2631
123.5 Critical exponents, scaling and universality . . . 2631
123.5.1 The exponents and the inequalities they obey . . . 2631
123.5.2 Corresponding states and universality . . . 2632
123.5.3 Scaling hypotheses . . . 2633
123.5.4 Correlation functions and the correlation length . . . 2633
123.5.5 The renormalization group and measured exponents . . . 2633
123.6 Lower critical dimension and topological order . . . 2634
123.6.1 The Mermin–Wagner–Hohenberg theorem . . . 2634
123.6.2 The Berezinskii–Kosterlitz–Thouless transition . . . 2634
123.6.3 Observation in helium films . . . 2634
123.7 Liquid crystals . . . 2634
123.7.1 Discovery and the mesomorphic states . . . 2634
123.7.2 The nematic transition . . . 2635

123.7.3 Defects, elasticity and textures . . . 2635

124 Nonequilibrium Thermodynamics and Transport 2637
124.1 The kinetic route . . . 2637
124.1.1 The Boltzmann equation and the H-theorem . . . 2637
124.1.2 Chapman–Enskog and the transport coefficients . . . 2638
124.1.3 The phenomenological transport laws . . . 2638
124.2 Linear irreversible thermodynamics . . . 2638
124.2.1 Entropy production and thermodynamic forces . . . 2638
124.2.2 Onsager reciprocal relations . . . 2638
124.2.3 Reciprocity with a magnetic field . . . 2638
124.2.4 Steady states and minimum entropy production . . . 2638
124.3 Thermoelectricity . . . 2639
124.3.1 The Seebeck, Peltier and Thomson effects . . . 2639
124.3.2 The Thomson relations as a test of reciprocity . . . 2639
124.3.3 Conduction in metals and the figure of merit . . . 2639
124.4 Brownian motion and stochastic dynamics . . . 2639
124.4.1 The Einstein–Smoluchowski relation . . . 2639
124.4.2 The Langevin equation . . . 2639
124.4.3 The Fokker–Planck equation . . . 2639
124.4.4 Master equations . . . 2640
124.5 Linear response and fluctuation–dissipation . . . 2640
124.5.1 Johnson–Nyquist noise . . . 2640
124.5.2 The fluctuation–dissipation theorem . . . 2640
124.5.3 Kubo formulae and Green–Kubo relations . . . 2640
124.5.4 Projection operators and memory kernels . . . 2640
124.6 Fluctuation theorems . . . 2640
124.6.1 The steady-state fluctuation theorem . . . 2640
124.6.2 Work relations far from equilibrium . . . 2641
124.6.3 Experimental verification . . . 2641
124.6.4 Stochastic thermodynamics and information . . . 2641

125 Phonons and Lattice Dynamics 2643
125.1 The crystal lattice . . . 2643
125.1.1 Bravais lattices and the crystallographic restriction . . . 2643
125.1.2 Point groups, space groups and the basis . . . 2643
125.1.3 The reciprocal lattice and Brillouin zones . . . 2644
125.1.4 Quasicrystals: the exception that fixed the rule . . . 2644
125.2 Experiment: the lattice made visible . . . 2644
125.2.1 Laue diffraction of X-rays . . . 2644
125.2.2 Bragg's law and structure determination . . . 2644
125.2.3 Structure factor and systematic absences . . . 2645
125.3 Heat capacity of solids . . . 2645
125.3.1 The Dulong–Petit law and its failure . . . 2645
125.3.2 The Einstein model . . . 2646
125.3.3 The Debye model and the cubic law . . . 2646
125.3.4 Calorimetric tests at low temperature . . . 2646
125.4 Lattice dynamics . . . 2647
125.4.1 The Born–von Kármán theory . . . 2647
125.4.2 Acoustic and optical branches . . . 2647
125.4.3 Quantization: the phonon . . . 2647
125.4.4 Polar crystals and the Lyddane–Sachs–Teller relation . . . 2648

125.5 Measured dispersion relations . . . 2648
125.5.1 Inelastic neutron scattering . . . 2648
125.5.2 Inelastic X-ray and light scattering . . . 2649
125.5.3 The Debye–Waller factor . . . 2649
125.6 Anharmonicity . . . 2649
125.6.1 Thermal expansion and the Grüneisen parameter . . . 2649
125.6.2 Phonon–phonon scattering and umklapp . . . 2650
125.6.3 Thermal conductivity of insulators . . . 2650

126 Electrons in Solids: Band Theory 2651
126.1 The free-electron metal . . . 2651
126.1.1 Drude's kinetic model . . . 2651
126.1.2 The Wiedemann–Franz law . . . 2652
126.1.3 What the classical electron gas cannot do . . . 2653
126.1.4 The Sommerfeld model . . . 2654
126.2 Electrons in a periodic potential . . . 2654
126.2.1 Bloch's theorem . . . 2654
126.2.2 Reciprocal lattice and Brillouin zones . . . 2654
126.2.3 Nearly free electrons . . . 2654
126.2.4 Tight binding and Wannier functions . . . 2654
126.2.5 The Kronig–Penney model . . . 2655
126.3 Band structure and the classification of solids . . . 2655
126.3.1 Filling and Wilson's criterion . . . 2655
126.3.2 Semimetals, semiconductors and the gap scale . . . 2655
126.3.3 Computing band structures . . . 2656
126.3.4 Density functional theory . . . 2656
126.4 Semiclassical dynamics of Bloch electrons . . . 2656
126.4.1 Group velocity and effective mass . . . 2656
126.4.2 Holes . . . 2656
126.4.3 Bloch oscillations and Zener breakdown . . . 2656
126.4.4 Magnetotransport and the Hall coefficient . . . 2656
126.5 The Fermi surface and how it is measured . . . 2657
126.5.1 Landau levels and quantum oscillations . . . 2657
126.5.2 The de Haas–van Alphen effect . . . 2657
126.5.3 Shubnikov–de Haas and cyclotron resonance . . . 2657
126.5.4 Photoemission . . . 2658
126.6 Transport and scattering . . . 2658
126.6.1 Residual resistivity and Matthiessen's rule . . . 2658
126.6.2 Phonon scattering and the Bloch–Grüneisen law . . . 2658
126.7 Where the one-electron picture fails . . . 2659
126.7.1 Fermi-liquid theory . . . 2659
126.7.2 Mott insulators and the Hubbard model . . . 2659
126.7.3 Disorder and localization . . . 2659

127 Semiconductors 2661
127.1 What a semiconductor is . . . 2661
127.1.1 Wilson's band picture . . . 2661
127.1.2 The band structures of silicon, germanium and gallium arsenide . . . 2662
127.1.3 Measuring the gap optically . . . 2662
127.2 Carriers and their statistics . . . 2663
127.2.1 Intrinsic conduction . . . 2663
127.2.2 Doping and extrinsic conduction . . . 2663

127.2.3 The Hall effect as a carrier measurement . . . 2664
127.2.4 Drift, diffusion and mobility . . . 2664
127.2.5 Recombination and carrier lifetime . . . 2664
127.3 Junctions and rectification . . . 2664
127.3.1 Metal–semiconductor contacts . . . 2664
127.3.2 The p–n junction . . . 2665
127.3.3 The Haynes–Shockley experiment . . . 2665
127.3.4 Tunnelling and the Esaki diode . . . 2666
127.4 Devices . . . 2666
127.4.1 The transistor . . . 2666
127.4.2 The MOSFET and the inversion layer . . . 2667
127.4.3 The integrated circuit . . . 2667
127.5 Light and semiconductors . . . 2667
127.5.1 Photoconductivity and the solar cell . . . 2667
127.5.2 Light-emitting diodes . . . 2667
127.5.3 The semiconductor laser and heterostructures . . . 2668
127.5.4 Excitons . . . 2668
127.6 Low-dimensional semiconductors . . . 2668
127.6.1 Quantum wells and superlattices . . . 2668
127.6.2 The two-dimensional electron gas . . . 2668

128 Experiment: The Quantum Hall Effect 2669
128.1 Historical context and the prediction under test . . . 2669
128.1.1 The classical Hall effect . . . 2669
128.1.2 Two-dimensional electron gases . . . 2670
128.1.3 Landau quantization and what was expected . . . 2670
128.2 Apparatus . . . 2670
128.2.1 The silicon MOSFET Hall bar . . . 2670
128.2.2 Heterostructure samples . . . 2670
128.2.3 Cryogenics, magnets and electrical metrology . . . 2670
128.3 Procedure . . . 2670
128.4 Observations . . . 2671
128.4.1 The integer plateaux . . . 2671
128.4.2 Exactness and universality . . . 2672
128.4.3 The fractional plateaux . . . 2672
128.5 Interpretation . . . 2673
128.5.1 Laughlin's gauge argument . . . 2673
128.5.2 Disorder, localization and plateau width . . . 2673
128.5.3 Edge states . . . 2673
128.5.4 Topology: TKNN and the Chern number . . . 2673
128.5.5 The Laughlin state and fractional charge . . . 2673
128.6 Later measurements and precision . . . 2674
128.6.1 Fractional charge and anyon statistics measured . . . 2674
128.6.2 Composite fermions . . . 2674
128.6.3 Graphene and the relativistic quantum Hall effect . . . 2674
128.6.4 Quantum spin Hall and topological insulators . . . 2675
128.6.5 The quantum anomalous Hall effect . . . 2675
128.7 Metrology: the ohm and the SI . . . 2676

129 Magnetism in Matter **2677**
129.1 The magnetic response of matter . . . 2677
129.1.1 Diamagnetism . . . 2677
129.1.2 Paramagnetism and the Curie law . . . 2678
129.1.3 The Bohr–van Leeuwen theorem . . . 2679
129.1.4 Magnetism of the conduction electrons . . . 2679
129.2 Where the moments come from . . . 2680
129.2.1 Local moments and the effective moment . . . 2680
129.2.2 The gyromagnetic ratio measured . . . 2680
129.3 Ferromagnetism . . . 2681
129.3.1 Weiss's molecular field . . . 2681
129.3.2 Exchange . . . 2682
129.3.3 The Heisenberg model . . . 2682
129.3.4 Itinerant ferromagnetism . . . 2682
129.4 Antiferromagnetism and ferrimagnetism . . . 2682
129.4.1 Néel's theory . . . 2682
129.4.2 Neutron diffraction as the decisive evidence . . . 2682
129.4.3 Superexchange and RKKY coupling . . . 2683
129.5 Domains and hysteresis . . . 2683
129.5.1 Domain structure and the Barkhausen effect . . . 2683
129.5.2 Imaging domains . . . 2683
129.5.3 Domain walls and anisotropy . . . 2684
129.6 Spin waves . . . 2684
129.6.1 Magnons and the three-halves law . . . 2684
129.6.2 Magnons measured by neutron scattering . . . 2685
129.7 Magnetic phase transitions . . . 2685
129.8 Magnetic resonance . . . 2685
129.8.1 Electron spin resonance . . . 2685
129.8.2 Nuclear magnetic resonance . . . 2685
129.9 Magnetotransport and spintronics . . . 2685
129.9.1 Giant magnetoresistance . . . 2685
129.9.2 Tunnelling magnetoresistance and spin transport . . . 2686

130 Superconductivity and Superfluidity **2687**
130.1 Discovery and phenomenology . . . 2687
130.1.1 Vanishing resistance . . . 2687
130.1.2 The Meissner–Ochsenfeld effect . . . 2688
130.1.3 Critical temperature, field and current . . . 2689
130.1.4 The two-fluid picture . . . 2689
130.2 Phenomenological electrodynamics . . . 2689
130.2.1 The London equations . . . 2689
130.2.2 Non-local response and the coherence length . . . 2689
130.2.3 Ginzburg–Landau theory . . . 2689
130.2.4 Type II superconductors and the Abrikosov lattice . . . 2690
130.3 Microscopic theory . . . 2690
130.3.1 The isotope effect . . . 2690
130.3.2 The energy gap . . . 2690
130.3.3 Cooper pairing . . . 2691
130.3.4 The BCS ground state . . . 2691
130.3.5 Broken gauge symmetry . . . 2691
130.4 Macroscopic quantum coherence . . . 2691
130.4.1 Off-diagonal long-range order . . . 2691

130.4.2 Flux quantization . . . 2692
130.4.3 The Josephson effects . . . 2692
130.4.4 SQUIDs and quantum metrology . . . 2693
130.5 Superconductors beyond the BCS regime . . . 2693
130.5.1 The cuprates . . . 2693
130.5.2 Pairing symmetry in the cuprates . . . 2693
130.5.3 Other families . . . 2693
130.5.4 What remains unexplained . . . 2694
130.6 Superfluid helium-4 . . . 2694
130.6.1 The lambda transition and the discovery of superflow . . . 2694
130.6.2 Bose–Einstein condensation as the mechanism . . . 2694
130.6.3 The two-fluid model . . . 2695
130.6.4 The excitation spectrum and rotons . . . 2695
130.6.5 Quantized circulation and vortices . . . 2695
130.7 Superfluid helium-3 . . . 2696
130.7.1 Discovery of the superfluid phases . . . 2696
130.7.2 The order parameter and its consequences . . . 2696
130.8 Where the phenomena reappear . . . 2696

131 Experiment: Superconductivity 2697
131.1 Historical context and the claim under test . . . 2697
131.1.1 The race to liquefy the permanent gases . . . 2697
131.1.2 Three rival predictions for the low-temperature resistance . . . 2697
131.1.3 Mercury as the test material . . . 2698
131.2 Apparatus . . . 2698
131.2.1 The Leiden cryostat and the helium cycle . . . 2698
131.2.2 Resistance measurement . . . 2698
131.2.3 Magnetometry for flux expulsion . . . 2698
131.2.4 Tunnel junctions . . . 2698
131.3 Procedure . . . 2698
131.3.1 The 1911 mercury run . . . 2698
131.3.2 Field-cooled and zero-field-cooled protocols . . . 2699
131.3.3 Isotope separation and the mercury series . . . 2699
131.4 Observations . . . 2699
131.4.1 The mercury transition . . . 2699
131.4.2 Bounds on the residual resistivity . . . 2700
131.4.3 Critical temperatures and critical fields . . . 2701
131.4.4 The isotope exponent . . . 2701
131.4.5 The gap . . . 2701
131.4.6 The flux quantum . . . 2702
131.4.7 Josephson currents and Shapiro steps . . . 2702
131.5 Interpretation . . . 2703
131.5.1 What each measurement establishes . . . 2703
131.5.2 What the 1986 discovery did not settle . . . 2703
131.6 Modern repetitions and precision . . . 2704
131.6.1 Metrology: the Josephson volt . . . 2704
131.6.2 SQUID magnetometry and the London moment . . . 2704
131.6.3 High-pressure hydrides and the standard of proof . . . 2704
131.6.4 Datasets . . . 2704

132 Experiment: Neutron Scattering and the Structure of Solids **2705**
132.1 Why the neutron, and not the X-ray . . . 2705
132.2 Shull and Smart: antiferromagnetic order in MnO (1949) . . . 2706
132.2.1 Apparatus . . . 2707
132.2.2 Procedure . . . 2707
132.2.3 Observations and data . . . 2707
132.2.4 Interpretation . . . 2708
132.2.5 Primary references . . . 2708
132.3 Brockhouse and Stewart: phonons in aluminium (1955) . . . 2708
132.3.1 Apparatus . . . 2708
132.3.2 Procedure . . . 2709
132.3.3 Observations and data . . . 2709
132.3.4 Interpretation . . . 2710
132.3.5 Primary references . . . 2710
132.4 Brockhouse: spin waves in magnetite (1957) . . . 2710
132.4.1 Apparatus . . . 2710
132.4.2 Procedure . . . 2710
132.4.3 Observations and data . . . 2710
132.4.4 Interpretation . . . 2712
132.4.5 Primary references . . . 2712
132.5 Brockhouse and Iyengar: the normal modes of germanium (1958) . . . 2712
132.5.1 Apparatus . . . 2712
132.5.2 Procedure . . . 2712
132.5.3 Observations and data . . . 2712
132.5.4 Interpretation . . . 2713
132.5.5 Primary references . . . 2713
132.6 The Debye–Waller factor, measured on the same patterns . . . 2713
132.7 What the neutron settled . . . 2715
132.8 Hodgkin, Hauptman and Karle, and Klug: the phase problem (1934–1969) . . . 2716
132.8.1 Apparatus . . . 2716
132.8.2 Procedure . . . 2717
132.8.3 Observations and data . . . 2717
132.8.4 Interpretation . . . 2721
132.8.5 Primary references . . . 2722

XIII Frontiers and Open Problems **2723**

133 What We Observe but Do Not Understand **2725**
133.1 The mass and mixing puzzles . . . 2725
133.2 The hierarchy problem . . . 2725
133.3 Dark matter . . . 2726
133.4 Dark energy and the Hubble tension . . . 2726
133.5 The baryon asymmetry . . . 2726
133.6 Confinement and the Yang–Mills mass gap . . . 2726
133.7 What lies outside this treatise . . . 2727

134 The Dark Sector: Evidence Without Explanation **2729**
134.1 The dynamical evidence . . . 2729
134.1.1 Cluster velocity dispersions . . . 2729
134.1.2 Galaxy rotation curves . . . 2729
134.1.3 Neutral-hydrogen curves beyond the optical disc . . . 2730

134.1.4 Galaxy stability and halo masses . . . 2730
134.1.5 X-ray hydrostatic cluster masses . . . 2730
134.2 Gravitational lensing . . . 2730
134.2.1 Strong lensing . . . 2730
134.2.2 Weak lensing and shear surveys . . . 2730
134.2.3 The Bullet Cluster . . . 2730
134.2.4 Microlensing and the compact-object bound . . . 2731
134.3 Cosmological evidence . . . 2731
134.3.1 The microwave-background budget . . . 2731
134.3.2 Nucleosynthesis and the baryon count . . . 2731
134.3.3 Baryon acoustic oscillations . . . 2731
134.3.4 Structure formation . . . 2731
134.4 Candidates and the searches that found nothing . . . 2731
134.4.1 What a candidate must satisfy . . . 2731
134.4.2 Direct detection: null results . . . 2732
134.4.3 The DAMA claim and its non-confirmation . . . 2732
134.4.4 Axions . . . 2732
134.4.5 Collider and indirect searches . . . 2732
134.4.6 Primordial black holes . . . 2732
134.5 Modified dynamics as the rival hypothesis . . . 2732
134.5.1 MOND and the radial-acceleration relation . . . 2732
134.5.2 Where modified dynamics fails . . . 2733
134.6 Dark energy . . . 2733
134.6.1 The supernova evidence for acceleration . . . 2733
134.6.2 The cosmological constant and its problem . . . 2733
134.6.3 The equation of state . . . 2733
134.7 Tensions and the honest status . . . 2733
134.7.1 The Hubble tension . . . 2733
134.7.2 The S_8 tension . . . 2734
134.7.3 What is and is not known . . . 2734

135 Quantum Gravity: The Honest Status 2735
135.1 Why the question arises . . . 2735
135.1.1 Two frameworks with incompatible foundations . . . 2735
135.1.2 The Planck scale . . . 2736
135.1.3 Must gravity be quantized at all? . . . 2736
135.2 What is established . . . 2736
135.2.1 Gravity as an effective field theory . . . 2736
135.2.2 Perturbative non-renormalizability . . . 2736
135.2.3 Quantum field theory in curved spacetime . . . 2736
135.2.4 Black-hole thermodynamics as theory . . . 2736
135.2.5 The information problem . . . 2737
135.3 What has actually been measured . . . 2737
135.3.1 Quantum matter in a classical gravitational field . . . 2737
135.3.2 Newton's law at short range . . . 2737
135.3.3 Graviton and photon propagation speed . . . 2737
135.3.4 Lorentz invariance at the highest energies . . . 2737
135.3.5 Classical black-hole predictions confirmed . . . 2738
135.4 What has not been measured . . . 2738
135.4.1 Hawking radiation . . . 2738
135.4.2 Analogue systems are analogues . . . 2738
135.4.3 Gravitational decoherence and collapse models . . . 2738

135.4.4 Planck-scale interferometry . . . 2738
135.5 Proposed experiments . . . 2738
135.5.1 Tabletop gravitational entanglement . . . 2738
135.5.2 Other proposals . . . 2739
135.6 Programmes without evidence . . . 2739
135.6.1 The editorial position . . . 2739
135.6.2 String theory . . . 2739
135.6.3 Supersymmetry and the collider nulls . . . 2739
135.6.4 Loop quantum gravity . . . 2739
135.6.5 Asymptotic safety . . . 2739
135.7 Status . . . 2740

136 Experiment: Dark-Matter Searches and Their Null Results 2741
136.1 What is being searched for, and what a null result bounds . . . 2742
136.1.1 The halo the searches assume . . . 2742
136.1.2 Elastic recoil kinematics . . . 2742
136.1.3 How to read an exclusion curve . . . 2742
136.2 LUX-ZEPLIN: the largest xenon target (2023) . . . 2743
136.2.1 Apparatus . . . 2743
136.2.2 Procedure . . . 2743
136.2.3 Observations and data . . . 2743
136.2.4 Interpretation . . . 2744
136.2.5 Primary references . . . 2745
136.3 XENONnT: the same signature in an independent detector (2023) . . . 2745
136.3.1 Apparatus . . . 2745
136.3.2 Procedure . . . 2745
136.3.3 Observations and data . . . 2745
136.3.4 Interpretation . . . 2746
136.3.5 Primary references . . . 2746
136.4 Cryogenic bolometers: the low-mass frontier . . . 2746
136.5 DAMA/LIBRA and the annual-modulation claim (2008–2018) . . . 2747
136.5.1 Apparatus . . . 2747
136.5.2 Procedure . . . 2747
136.5.3 Observations and data . . . 2748
136.5.4 Interpretation . . . 2749
136.5.5 Primary references . . . 2749
136.6 ADMX: the axion haloscope (2018–2020) . . . 2749
136.6.1 Apparatus . . . 2749
136.6.2 Procedure . . . 2749
136.6.3 Observations and data . . . 2749
136.6.4 Interpretation . . . 2750
136.6.5 Primary references . . . 2750
136.7 ATLAS: missing transverse momentum at the Large Hadron Collider (2021) 2750
136.7.1 Apparatus . . . 2750
136.7.2 Procedure . . . 2751
136.7.3 Observations and data . . . 2751
136.7.4 Interpretation . . . 2752
136.7.5 Primary references . . . 2752
136.8 Fermi-LAT: gamma rays from dwarf spheroidal galaxies (2009–) . . . 2752
136.8.1 Apparatus . . . 2752
136.8.2 Procedure . . . 2752
136.8.3 Observations and data . . . 2752

136.8.4 Interpretation . . . 2753
136.8.5 Primary references . . . 2754
136.9 The state of the search . . . 2754

Appendices **2755**

A Long Proofs **2759**
A.1 The Quadrupole Formula for Gravitational Radiation . . . 2759
A.1.1 Setup: linearized geometry . . . 2760
A.1.2 Gauge freedom and the Lorenz gauge . . . 2761
A.1.3 Vacuum solutions: plane waves and the transverse–traceless gauge . . . 2762
A.1.4 Physical effect: geodesic deviation and the strain . . . 2763
A.1.5 Energy carried by the wave: the Isaacson tensor . . . 2764
A.1.6 Generation: the quadrupole formula . . . 2765
A.1.7 Application: the circular compact binary . . . 2767
A.2 Oppenheimer–Snyder Collapse: the Interior Matching . . . 2771
A.2.1 The interior: a contracting closed dust universe . . . 2771
A.2.2 The boundary, and the matching conditions . . . 2772
A.2.3 The global picture . . . 2773
A.3 The Cantor–Schröder–Bernstein Theorem . . . 2773
A.4 The Completeness Theorem . . . 2775
A.4.1 Soundness . . . 2775
A.4.2 Maximal consistent sets . . . 2776
A.4.3 Henkin witnesses . . . 2777
A.4.4 The term model . . . 2778
A.4.5 The truth lemma . . . 2778
A.4.6 Completeness, and compactness . . . 2779
A.5 Arithmetization and the Incompleteness Theorems . . . 2780
A.5.1 Gödel numbering . . . 2780
A.5.2 Representability . . . 2781
A.5.3 The diagonal lemma . . . 2785
A.5.4 Rosser's form of the first theorem . . . 2786
A.5.5 The derivability conditions and the second theorem . . . 2786
A.5.6 Church's theorem . . . 2787
A.6 The Universal Turing Machine . . . 2788
A.6.1 Coding a machine as a string . . . 2788
A.6.2 Extra tapes buy no extra power . . . 2789
A.6.3 Construction of U . . . 2789
A.7 Completeness of the Real Numbers . . . 2791
A.7.1 Statement . . . 2791
A.7.2 The construction: Dedekind cuts . . . 2791
A.7.3 Order and the least-upper-bound property . . . 2792
A.7.4 The field operations . . . 2792
A.7.5 Consequences . . . 2793
A.8 Equivalence of the Einstein–Hilbert Action in Metric and Vielbein Variables 2794
A.8.1 Statement . . . 2794
A.8.2 A contraction lemma for the Levi-Civita symbol . . . 2794
A.8.3 Proof of the theorem . . . 2795
A.9 Wigner's Theorem on Quantum Symmetries . . . 2797
A.9.1 Orthonormal bases map to orthonormal bases . . . 2797
A.9.2 Phase alignment along a basis . . . 2797

A.9.3 The induced scalar map . . . 2798
A.9.4 Proof of the theorem . . . 2798
A.10 The Classification of Kinematical Algebras . . . 2799
A.10.1 The generators and the hypotheses . . . 2800
A.10.2 Step one: the hypotheses reduce the algebra to five constants . . . 2800
A.10.3 Step two: the Jacobi identity eliminates two of the five . . . 2801
A.10.4 Step three: rescaling reduces the constants to their signs . . . 2802
A.10.5 Step four: the enumeration . . . 2802
A.10.6 Two corollaries the enumeration makes visible . . . 2803
A.11 Invariant Bilinear Forms on an Expanded Algebra . . . 2804
A.12 Darboux's Theorem: Local Canonical Coordinates . . . 2807
A.12.1 Statement . . . 2808
A.12.2 The pullback of a k-form . . . 2808
A.12.3 Flows of time-dependent vector fields . . . 2810
A.12.4 The local converse of the Poincaré lemma . . . 2813
A.12.5 Moser's homotopy argument . . . 2814
A.12.6 Reading the result . . . 2816
A.13 The Korn–Lichtenstein Theorem and the Elliptic Canonical Form . . . 2817
A.13.1 Statement . . . 2817
A.13.2 From the canonical form to a Beltrami equation . . . 2818
A.13.3 Normalizing the Beltrami coefficient . . . 2820
A.13.4 The Cauchy and Beurling transforms . . . 2821
A.13.5 The construction and the Jacobian . . . 2823
A.14 The Cauchy–Kovalevskaya Theorem . . . 2825
A.14.1 Formal series, the recursion, and the reduction . . . 2825
A.14.2 Majorants . . . 2827
A.14.3 The majorant problem has a closed-form solution . . . 2828
A.14.4 Proof of the theorem . . . 2829
A.15 The Stäckel Conditions and the Eleven Separable Systems . . . 2831
A.15.1 Orthogonal coordinates and the Laplacian . . . 2832
A.15.2 What separation means, and what it forces . . . 2833
A.15.3 The three systems the treatise uses . . . 2835
A.15.4 The generic system: confocal quadrics . . . 2836
A.15.5 The enumeration . . . 2839
A.16 The Mean-Value Property Characterizes Harmonic Functions . . . 2841
A.16.1 Statement . . . 2842
A.16.2 Differentiation under the integral sign . . . 2842
A.16.3 A radial mollifier . . . 2843
A.16.4 The mollification reproduces the function . . . 2845
A.16.5 Proof of the theorem . . . 2846
A.17 Kruzhkov's Doubling of Variables and the Entropy Solution . . . 2847
A.17.1 Statement . . . 2848
A.17.2 The inputs this treatise does not build . . . 2848
A.17.3 From the admissible entropies to Kruzhkov's family . . . 2849
A.17.4 Doubling the variables . . . 2851
A.17.5 The limit and the cone estimate . . . 2852
A.17.6 Existence by vanishing viscosity . . . 2853
A.18 Lévy's Continuity Theorem . . . 2856
A.18.1 Statement . . . 2856
A.18.2 The two facts imported from Lebesgue theory . . . 2856
A.18.3 The Dirichlet integral . . . 2857

A.18.4 Helly's selection theorem . . . 2858
A.18.5 The inversion formula . . . 2859
A.18.6 Tightness from continuity at the origin . . . 2860
A.18.7 Proof of the theorem . . . 2861
A.19 The Lindeberg–Feller Central Limit Theorem . . . 2862
A.19.1 Statement and notation . . . 2863
A.19.2 Lindeberg's condition implies uniform asymptotic negligibility . . . 2863
A.19.3 The uniform second-order expansion . . . 2864
A.19.4 Two product estimates . . . 2864
A.19.5 Proof of the theorem . . . 2865
A.20 The Berry–Esseen Inequality . . . 2867
A.20.1 Statement . . . 2867
A.20.2 The Fejér kernel . . . 2867
A.20.3 Esseen's smoothing inequality . . . 2868
A.20.4 The characteristic function of the standardised sum . . . 2870
A.20.5 Integration, and the constant . . . 2872
A.21 Wilks' Theorem in the Multi-Parameter Case . . . 2873
A.21.1 Setting and notation . . . 2873
A.21.2 What is quoted . . . 2874
A.21.3 The uniform quadratic expansion . . . 2875
A.21.4 Localisation of both estimators . . . 2876
A.21.5 Replacing the surface by its tangent space . . . 2877
A.21.6 The chi-squared limit . . . 2878
A.22 Rice's Formula for the Expected Number of Upcrossings . . . 2879
A.22.1 Hypotheses and statement . . . 2880
A.22.2 The joint law of the field and its derivative . . . 2881
A.22.3 Kac's counting integral . . . 2881
A.22.4 Taking the expectation . . . 2882
A.22.5 Evaluation . . . 2883
A.23 The Hilbert–Schmidt Theorem for Compact Self-Adjoint Operators . . . 2885
A.23.1 Statement . . . 2885
A.23.2 Compactness in sequential form . . . 2886
A.23.3 The supremum is attained: the first eigenvector . . . 2886
A.23.4 The recursion . . . 2887
A.23.5 Passage to the limit . . . 2888
A.24 The Spectral Theorem for a Bounded Self-Adjoint Operator . . . 2889
A.24.1 Statement . . . 2890
A.24.2 The polynomial calculus . . . 2890
A.24.3 The continuous calculus . . . 2891
A.24.4 The spectral measures . . . 2892
A.24.5 The Borel calculus . . . 2892
A.24.6 The projection-valued measure . . . 2893
A.24.7 The multiplication form . . . 2895
A.25 Stone's Theorem on One-Parameter Unitary Groups . . . 2897
A.25.1 Statement . . . 2897
A.25.2 Vector-valued integrals . . . 2897
A.25.3 The domain is dense . . . 2898
A.25.4 The generator is closed and self-adjoint . . . 2899
A.25.5 The converse: exponentiating a self-adjoint operator . . . 2900
A.26 Von Neumann's Criterion for Self-Adjoint Extensions . . . 2904
A.26.1 Statement . . . 2904

A.26.2 The basic identity . . . 2904
A.26.3 The Cayley transform . . . 2905
A.26.4 From an isometry back to an operator . . . 2906
A.26.5 Proof of the criterion . . . 2907
A.26.6 The two worked cases . . . 2908
A.27 The Nuclear Spectral Theorem of Gelfand and Maurin . . . 2910
A.27.1 Hypotheses and statement . . . 2910
A.27.2 The direct integral . . . 2911
A.27.3 What nuclearity buys . . . 2911
A.27.4 Generalized eigenvectors, and completeness . . . 2913
A.28 The Implicit Function Theorem . . . 2914
A.28.1 Statement . . . 2915
A.28.2 The contraction . . . 2915
A.28.3 Existence and uniqueness of the implicit map . . . 2916
A.28.4 Continuity, differentiability and the derivative formula . . . 2917
A.28.5 The two consequences the treatise uses . . . 2918
A.28.6 Reading the result . . . 2920
A.29 The General Stokes Theorem for Differential Forms . . . 2920
A.29.1 Manifolds with boundary . . . 2921
A.29.2 Orientation and the integral of an n-form . . . 2921
A.29.3 Bump functions and partitions of unity . . . 2923
A.29.4 Statement and the half-space computation . . . 2924
A.29.5 Proof of the theorem . . . 2926
A.29.6 The instance stated in the chapter . . . 2926
A.29.7 What is quoted, and what is not . . . 2928
A.30 The Poincaré–Birkhoff–Witt Theorem . . . 2928
A.30.1 Statement . . . 2929
A.30.2 The filtration, and spanning . . . 2929
A.30.3 The action on the polynomial algebra . . . 2930
A.30.4 Independence, and proof of the theorem . . . 2932
A.30.5 Consequences: the symbol and the symmetrization map . . . 2932
A.31 Cartan's Criterion for Semisimplicity . . . 2934
A.31.1 Statement . . . 2934
A.31.2 Solvability, and the language of the proof . . . 2934
A.31.3 The easy half: an abelian ideal lies in the radical of the Killing form . . . 2935
A.31.4 The Jordan decomposition . . . 2936
A.31.5 Engel's theorem . . . 2936
A.31.6 The trace lemma . . . 2937
A.31.7 Proof of the criterion . . . 2939
A.31.8 Consequences, and the case of $\mathfrak{so}(p,q)$ . . . 2939
A.32 Racah's Theorem on the Number of Casimir Operators . . . 2941
A.32.1 Statement . . . 2941
A.32.2 The centre of the enveloping algebra and the invariant polynomials . . . 2942
A.32.3 The two quoted theorems . . . 2943
A.32.4 The count . . . 2944
A.32.5 The count checked against the chapter's algebras . . . 2945
A.33 Covering Spaces and the Fundamental Group of the Rotation Group . . . 2946
A.33.1 Statement . . . 2946
A.33.2 Two lifting lemmas . . . 2947
A.33.3 The monodromy theorem . . . 2949
A.33.4 Covering homomorphisms and deck transformations . . . 2950

A.33.5 The rotation group . . . 2950
A.34 The Structure Constants of $\mathfrak{su}(3)$ . . . 2951
A.34.1 One trace for both arrays . . . 2952
A.34.2 Which triples can be nonzero at all . . . 2952
A.34.3 Family (1): three diagonal indices . . . 2954
A.34.4 Family (2): one diagonal index and one off-diagonal pair . . . 2954
A.34.5 Family (3): one index from each pair . . . 2955
A.34.6 The tables . . . 2955
A.34.7 A Jacobi check . . . 2956
A.35 Simplicity of $\mathfrak{su}(3)$ . . . 2957
A.35.1 Statement . . . 2957
A.35.2 Complexification . . . 2957
A.35.3 The root decomposition . . . 2958
A.35.4 Walking round the hexagon . . . 2959
A.35.5 Proof of the theorem . . . 2959
A.36 From the Little Algebra to the Labels of a Massive Representation . . . 2960
A.36.1 Conventions, and what is assumed . . . 2961
A.36.2 Statement . . . 2961
A.36.3 The geometry of the mass shell . . . 2962
A.36.4 The induced representation . . . 2963
A.36.5 The classification, and the count . . . 2965
A.36.6 The observed case . . . 2966
A.37 Whitehead's Lemmas and the Rigidity of Semisimple Algebras . . . 2967
A.37.1 The Chevalley–Eilenberg complex . . . 2967
A.37.2 The Casimir operator of a representation . . . 2969
A.37.3 Averaging a cocycle against the Casimir . . . 2970
A.37.4 Complete reducibility . . . 2971
A.37.5 Trivial coefficients . . . 2972
A.37.6 The Whitehead lemmas . . . 2973
A.38 The Second Cohomology of the Galilei Algebra . . . 2975
A.38.1 The algebra, and the two spaces . . . 2975
A.38.2 Three tensor lemmas . . . 2976
A.38.3 Solving the cocycle condition . . . 2977
A.38.4 The coboundaries, and the quotient . . . 2978
A.39 Central Extensions of Three Infinite-Dimensional Algebras . . . 2979
A.39.1 The Virasoro cocycle . . . 2980
A.39.2 The Kac–Moody cocycle . . . 2982
A.39.3 The Schwinger term . . . 2985
A.40 The Strengthened Jacobi Condition and Conjugate Points . . . 2988
A.40.1 Statement . . . 2988
A.40.2 The accessory equation as a first-order system . . . 2989
A.40.3 Continuous dependence on the initial point . . . 2989
A.40.4 The Sturm separation theorem . . . 2991
A.40.5 Proof of the equivalence . . . 2992
A.40.6 The displaced initial point . . . 2993
A.41 Tonelli's Existence Theorem for an Integrand Depending on the Function . . . 2994
A.41.1 Statement . . . 2995
A.41.2 The Sobolev space in one dimension . . . 2995
A.41.3 The Arzelà–Ascoli theorem and the compact embedding . . . 2997
A.41.4 Weak lower semicontinuity with y present . . . 2998
A.41.5 Proof of the theorem . . . 3000

A.42 Completeness of the Sturm–Liouville Eigenfunctions in the Energy Norm . 3001
A.42.1 Statement . . . 3002
A.42.2 The energy space . . . 3002
A.42.3 The Green function and the inverse operator . . . 3003
A.42.4 Compactness and the spectral decomposition . . . 3006
A.42.5 From the weighted norm to the energy norm . . . 3007
A.43 The Plateau Problem: the Architecture of Douglas' Proof . . . 3009
A.43.1 Setting and statement . . . 3009
A.43.2 The Dirichlet integral dominates the area . . . 3010
A.43.3 The Douglas functional . . . 3011
A.43.4 The conformal group and the three-point normalisation . . . 3012
A.43.5 Compactness: the Courant–Lebesgue estimate . . . 3014
A.43.6 The two quoted inputs, and the conclusion . . . 3015
A.44 Change of Variables in a Multiple Integral . . . 3018
A.44.1 What the derivation uses . . . 3018
A.44.2 Sets of zero content . . . 3019
A.44.3 Iterated integration in $\mathbb{R}^N$ . . . 3020
A.44.4 Two reductions . . . 3021
A.44.5 Permutations and primitive maps . . . 3022
A.44.6 The local factorisation . . . 3023
A.44.7 From compact support to the theorem . . . 3024
A.44.8 Reading the result . . . 3026
A.45 The Helmholtz Decomposition . . . 3026
A.45.1 What the derivation uses . . . 3027
A.45.2 The kernel is integrable, and uniformly so . . . 3027
A.45.3 Differentiating under the integral sign . . . 3028
A.45.4 Poisson's equation . . . 3029
A.45.5 Existence of the decomposition . . . 3030
A.45.6 Decay at infinity . . . 3030
A.45.7 Uniqueness . . . 3031
A.45.8 Reading the result . . . 3031
A.46 The Quotient Manifold Theorem . . . 3032
A.46.1 What the derivation uses, and what it must state itself . . . 3033
A.46.2 The orbit space is Hausdorff and second countable . . . 3034
A.46.3 Orbits are embedded submanifolds . . . 3034
A.46.4 The slice . . . 3035
A.46.5 The charts and the smooth structure . . . 3036
A.46.6 The universal property and uniqueness . . . 3037
A.46.7 Reading the result . . . 3038
A.47 The Liouville–Arnold Theorem . . . 3039
A.47.1 Setting and statement . . . 3039
A.47.2 The commuting fields and the level set . . . 3040
A.47.3 The translation action and the torus . . . 3040
A.47.4 The action variables . . . 3042
A.47.5 The angles, and the proof of the theorem . . . 3043
A.48 Marsden–Weinstein Reduction . . . 3044
A.48.1 Setting and statement . . . 3045
A.48.2 Two facts of symplectic linear algebra . . . 3046
A.48.3 The level set is a manifold . . . 3046
A.48.4 The degenerate directions of the restricted form . . . 3047
A.48.5 Descent of the form . . . 3048

A.48.6 Descent of the dynamics . . . 3049
A.48.7 The example the chapter names . . . 3049
A.48.8 What is quoted here . . . 3050
A.49 The Stone–von Neumann Theorem . . . 3051
A.49.1 Statement, conventions and units . . . 3051
A.49.2 The Weyl operators and their composition law . . . 3052
A.49.3 Two analytic tools . . . 3053
A.49.4 The Gaussian average . . . 3054
A.49.5 The vacuum vector . . . 3056
A.49.6 The intertwining unitary . . . 3057
A.49.7 The Schrödinger system . . . 3058
A.49.8 What is quoted here . . . 3059
A.50 Hudson's Theorem: the Pure States of Non-Negative Wigner Function . . . 3060
A.50.1 Statement and conventions . . . 3060
A.50.2 Gaussian integrals . . . 3061
A.50.3 The easy direction . . . 3062
A.50.4 Coherent states, and where the Wigner function cannot vanish . . . 3063
A.50.5 An entire function without zeros . . . 3064
A.50.6 What normalisability forces . . . 3066
A.50.7 Recovering the wavefunction . . . 3067
A.50.8 Scope, and what is quoted . . . 3068
A.51 The Jacobi Identity for the Dirac Bracket . . . 3069
A.51.1 Notation, and the four identities consumed . . . 3069
A.51.2 Route A: the direct expansion . . . 3070
A.51.3 Route B: the bracket of the induced symplectic form . . . 3072
A.52 The Gauss–Codazzi Form of the Einstein–Hilbert Lagrangian . . . 3073
A.52.1 The kit: normal, inverse metric, projector, determinant . . . 3074
A.52.2 The extrinsic curvature, and the dictionary . . . 3075
A.52.3 The Gauss equation . . . 3077
A.52.4 The Codazzi equation . . . 3078
A.52.5 The twice-normal contraction . . . 3078
A.52.6 Assembling the Lagrangian . . . 3079
A.52.7 The boundary term . . . 3081
A.53 The Hypersurface-Deformation Algebra . . . 3082
A.53.1 Conventions: weights, smearing functions, and the bracket . . . 3082
A.53.2 The momentum constraint generates the spatial drag . . . 3083
A.53.3 The first two brackets . . . 3084
A.53.4 The variation of the intrinsic curvature . . . 3085
A.53.5 The third bracket . . . 3085
A.53.6 The geometric reading . . . 3087
A.54 Existence and Uniqueness of a Nilpotent BRST Charge . . . 3088
A.54.1 The extended phase space and its two gradings . . . 3088
A.54.2 The Koszul–Tate differential . . . 3089
A.54.3 The recursion . . . 3091
A.54.4 Existence . . . 3092
A.54.5 Termination at the displayed terms . . . 3093
A.54.6 Uniqueness . . . 3093
A.55 Bertrand's Theorem . . . 3094
A.55.1 The orbit equation and its circular solutions . . . 3095
A.55.2 First stage: closure near a circle forces a power law . . . 3096
A.55.3 Second stage: the third-order solvability condition . . . 3097

A.55.4 The two laws, and that they really do close . . . 3099
A.56 Isotropic and Cubic Elastic Tensors . . . 3100
A.56.1 What is proved here, and what is proved elsewhere . . . 3101
A.56.2 The rotation group of the cube leaves four constants . . . 3101
A.56.3 From twenty-one constants to three, and then to two . . . 3103
A.57 Reduction of Three-Dimensional Elasticity to the Kirchhoff Plate Equation 3105
A.57.1 Kinematics: the Kirchhoff hypothesis . . . 3105
A.57.2 Plane stress, not plane strain . . . 3106
A.57.3 Integrating the energy across the thickness . . . 3107
A.57.4 The field equation . . . 3109
A.57.5 The free edge: three conditions are one too many . . . 3110
A.58 Hertz's Solution for the Elastic Half-Space Under an Axisymmetric Pressure 3112
A.58.1 What is quoted here . . . 3112
A.58.2 Superposition and the contact conditions . . . 3113
A.58.3 The elliptic pressure and its displacement . . . 3114
A.58.4 Matching, and the coefficient $\frac{4}{3}$ . . . 3115
A.58.5 Checks and consequences . . . 3117
A.59 D'Alembert's Paradox for a Body of Arbitrary Shape . . . 3118
A.59.1 Statement and hypotheses . . . 3118
A.59.2 The far field carries no source . . . 3118
A.59.3 The momentum theorem on a large sphere . . . 3120
A.59.4 Every surviving term integrates to zero . . . 3120
A.60 The Coefficient 6π in Stokes' Drag Law . . . 3121
A.60.1 The stream function and the equation it obeys . . . 3122
A.60.2 Solution for the sphere . . . 3122
A.60.3 The pressure, and a check . . . 3123
A.60.4 The traction and its integral . . . 3124
A.61 The Kármán Spacing Ratio of the Vortex Street . . . 3125
A.61.1 One summation formula does all the work . . . 3126
A.61.2 The street and its rigid translation . . . 3127
A.61.3 Linearization . . . 3127
A.61.4 The growth rates . . . 3129
A.61.5 Only one spacing survives . . . 3130
A.62 The Joukowski Map, the Kutta Condition and the 2π Lift Slope . . . 3131
A.62.1 Flow past a circle with circulation . . . 3132
A.62.2 The Joukowski map and the cusped edge . . . 3132
A.62.3 The Kutta condition and the lift . . . 3133
A.63 The Kármán–Howarth Relation and the Four-Fifths Law . . . 3135
A.63.1 Setting . . . 3135
A.63.2 Isotropic tensor functions of one vector . . . 3136
A.63.3 The pressure drops out . . . 3137
A.63.4 The two-point energy equation . . . 3138
A.63.5 Structure functions . . . 3139
A.63.6 The inertial range and the four-fifths law . . . 3140
A.64 The Korteweg–de Vries Equation from the Free-Surface Problem . . . 3142
A.64.1 The exact free-surface problem . . . 3142
A.64.2 Two small parameters, and a choice . . . 3142
A.64.3 Eliminating the depth . . . 3143
A.64.4 Unidirectional reduction . . . 3144
A.64.5 Two checks . . . 3145
A.65 The Lorenz System from Boussinesq Convection . . . 3146

A.65.1 The layer, in SI units . . . 3147
A.65.2 Rolls, the stream function, and the elimination of the pressure . . . 3148
A.65.3 The scaling, and the two dimensionless groups . . . 3149
A.65.4 Linear stability and the critical Rayleigh number . . . 3150
A.65.5 The Galerkin truncation . . . 3151
A.65.6 Reading the parameters back . . . 3153
A.66 Rayleigh Acoustic Streaming Above a Vibrating Surface . . . 3154
A.66.1 What is quoted here . . . 3154
A.66.2 Why the effect cannot appear before second order . . . 3155
A.66.3 The first-order Stokes layer . . . 3155
A.66.4 The second-order mean flow and the slip velocity . . . 3156
A.66.5 The direction of the circulation over a plate . . . 3159
A.66.6 The same mechanism in Kundt's tube . . . 3160
A.66.7 What is not derived here . . . 3161
A.67 The Free-Streamline Jet: Kirchhoff's Contraction Coefficient . . . 3161
A.67.1 What is quoted here . . . 3161
A.67.2 Formulation . . . 3162
A.67.3 The hodograph region is a half-strip . . . 3163
A.67.4 Two maps onto the same half plane . . . 3164
A.67.5 Integrating back along the free streamline . . . 3165
A.67.6 What the theorem does not cover . . . 3166
A.68 Prandtl's Boundary-Layer Equations and the Blasius Similarity Solution . . . 3167
A.68.1 What is quoted here . . . 3167
A.68.2 Scaling, and why the inviscid limit is singular . . . 3168
A.68.3 The inner expansion: Prandtl's equations . . . 3169
A.68.4 Matching hands over the pressure . . . 3170
A.68.5 The flat plate: similarity and the Blasius equation . . . 3170
A.68.6 The three constants . . . 3171
A.68.7 What these equations cannot describe . . . 3173

B Physical Constants and SI Units 3175
B.1 Mathematical constants . . . 3175
B.2 The seven defining constants of the SI . . . 3176
B.3 CODATA 2018 derived and measured constants . . . 3176
B.4 Planck units . . . 3176
B.5 SI base units . . . 3177

C The Free Parameters of Physics 3179
C.1 Electroweak and Higgs . . . 3179
C.2 Quantum Chromodynamics . . . 3181
C.3 Charged Lepton Masses and Quark Mixing (CKM) . . . 3185
C.4 Neutrino Oscillations and the PMNS Matrix . . . 3187
C.5 Lambda-CDM Cosmological Parameters . . . 3190
C.6 Gravitational Coupling, Planck Units, and the Hierarchy Problem . . . 3193

D The Dimensionless Constants 3197
D.1 Why only dimensionless numbers carry content . . . 3197
D.1.1 The dimension map . . . 3197
D.1.2 How many dimensionful constants are fundamental? . . . 3199
D.2 The systematic construction . . . 3199
D.2.1 Planck units and the mass ratios . . . 3199
D.2.2 The couplings . . . 3200

D.2.3 The full catalogue . . . 3200
D.3 Universality: the numbers a stranger could check . . . 3201
D.3.1 Why α and not a metre . . . 3201
D.3.2 The message that needs no dictionary . . . 3201
D.3.3 A primer built from nothing . . . 3202
D.3.4 Is it actually constant? . . . 3205
D.4 Relations to mathematical constants . . . 3205
D.4.1 When a mathematical constant legitimately appears . . . 3206
D.4.2 Eddington, and the failure mode . . . 3206
D.4.3 The Koide relation . . . 3206
D.4.4 Fine-tuning and the anthropic move . . . 3207
D.5 Summary . . . 3208

E Notation **3209**
E.1 General conventions . . . 3209
E.2 Index conventions for general dimension . . . 3209
E.3 Macro-to-meaning dictionary . . . 3210

Bibliography **3213**

List of Experiments

20.3 Galileo's Inclined Planes (1604–1638) . . . 750
20.4 Absolute Ballistic Gravimetry (1995–present) . . . 751

33 The Cavendish Torsion Balance (1798) 1103
34.1 Galileo and Huygens: isochronism (1602–1673) . . . 1121
34.2 Coulomb: the torsion pendulum (1784) . . . 1125
34.3 Kater: the reversible pendulum (1818) . . . 1130
34.5 Foucault: the pendulum at the Panthéon (1851) . . . 1135
35.1 Mersenne: the laws of the vibrating string (1636) . . . 1143
35.2 Melde: standing waves on a driven string (1860) . . . 1145
35.3 Chladni: the nodal figures of a vibrating plate (1787) . . . 1148
35.4 Newton, Laplace and the speed of sound (1687–1822) . . . 1151
35.5 Kundt's tube: sound speeds from dust figures (1866) . . . 1154
35.6 Buys Ballot: the Doppler effect on a moving train (1845) . . . 1156
36.1 Efflux from a vessel: Torricelli's law (1738) . . . 1159
36.2 Poiseuille: flow in fine capillaries (1840) . . . 1163
36.3 Reynolds: the dye filament and the critical number (1883) . . . 1165
36.4 Bénard: cellular convection (1900) . . . 1170
36.5 Prandtl: the boundary layer made visible (1904) . . . 1174
37.3 Michelson and Morley: the search for the aether wind (1887) . . . 1184
37.4 Fizeau: light dragged by moving water (1851) . . . 1185
37.5 Cavity resonators: the modern Michelson–Morley (1979–2015) . . . 1186
41.1 Ives and Stilwell: the moving atomic clock (1938) . . . 1219
41.2 Rossi and Hall: muon survival with altitude (1941) . . . 1220
41.3 Frisch and Smith: Mount Washington to sea level (1963) . . . 1221
41.4 The CERN muon storage ring (1977) . . . 1222
41.5 Hafele and Keating: caesium clocks flown around the world (1972) . . . 1223
41.6 Bertozzi: the speed limit (1964) . . . 1224
47.1 The Hulse–Taylor Binary Pulsar (PSR B1913+16) (1974–2016) . . . 1297
47.3 GW150914 (2015) . . . 1299

50 Black-Hole Observations (1972–2022) 1343

51 The Cosmic Microwave Background 1357
53.1 Le Verrier and the perihelion of Mercury (1859) . . . 1395
53.2 Dyson, Eddington and Davidson: the eclipse of 1919 (1919) . . . 1398
53.3 Pound, Rebka and Snider: redshift in a tower (1960–1965) . . . 1400
53.4 Shapiro: the radar time delay (1964–2003) . . . 1402
53.5 The Eötvös lineage and MICROSCOPE (1922–2022) . . . 1405
53.6 Lunar laser ranging and Gravity Probe B (1969–2011) . . . 1408

56 The Mechanical Equivalent of Heat **1431**
57.1 Stern: the direct measurement of molecular speeds (1920) . . . 1437
57.2 Miller and Kusch: the shape of the speed distribution (1955) . . . 1439
57.3 Maxwell: the viscosity of air is independent of its density (1866) . . . 1441
57.4 Graham: the diffusion and transpiration of gases (1833–1846) . . . 1443
57.5 Loschmidt: the size of an air molecule (1865) . . . 1445
57.6 Joule and Thomson: the porous plug (1852) . . . 1446
57.8 Giauque and MacDougall: cooling below one kelvin by adiabatic demagnetization (1933) . . . 1448

63 Electromagnetism **1495**

67 Wave Optics **1531**
67.8 Gabor: holography and the recording of phase (1948, 1962) . . . 1538

72 Franck–Hertz **1585**
73.1 Stefan: the fourth power from Tyndall's data (1879) . . . 1591
73.2 Lummer and Kurlbaum: a realizable black body (1898) . . . 1593
73.3 Lummer and Pringsheim: the failure of Wien's law (1899) . . . 1595
73.4 Rubens and Kurlbaum: the far-infrared radiance (1900) . . . 1597
73.5 Radiometric realization of the kelvin (2019) . . . 1599
74.1 Hertz and Hallwachs: ultraviolet light discharges a metal (1887–1888) . . . 1603
74.2 Lenard: the electron energy is independent of the intensity (1902) . . . 1605
74.3 Millikan: a direct photoelectric determination of Planck's constant (1916) . . . 1607
74.4 Compton: the wavelength shift of scattered X-rays (1923) . . . 1610
74.5 Bothe and Geiger: coincidence of scattered quantum and recoil electron (1925) 1612
74.6 Grangier, Roger and Aspect: single-photon anticorrelation at a beamsplitter (1986) . . . 1615
75.1 Davisson and Germer: electron diffraction by a nickel crystal (1927) . . . 1620
75.2 Thomson and Reid: electron transmission rings (1927) . . . 1622
75.3 Estermann and Stern: diffraction of neutral molecular beams (1930) . . . 1623
75.4 The electron double slit and its single-electron build-up (1961–1989) . . . 1625
75.5 Colella, Overhauser and Werner: gravitationally induced quantum interference (1975) . . . 1627
75.6 Molecule interferometry, from fullerenes to macromolecules (1999–2019) . . . 1629
76.1 Röntgen: a new kind of rays (1895) . . . 1635
76.2 Barkla: characteristic radiation and the transversality of X-rays (1905–1911) 1638
76.3 Friedrich, Knipping and von Laue: diffraction by a crystal (1912) . . . 1640
76.4 The Braggs: the reflection law and the X-ray spectrometer (1913) . . . 1641
76.5 Moseley: the high-frequency spectra of the elements (1913–1914) . . . 1643
76.6 Siegbahn: precision X-ray spectroscopy (1916–1925) . . . 1645
80.2 Stern–Gerlach (1921–1922) . . . 1672

89 Bell Tests **1745**

90 Precision Spectroscopy and Atomic Clocks **1753**
91.2 Alpha decay and the Geiger–Nuttall relation (1911, 1928) . . . 1764
91.3 Fowler and Nordheim: cold field emission from metals (1928) . . . 1767
91.4 Esaki: the tunnel diode and negative differential resistance (1958) . . . 1769
91.5 Giaever: tunnelling spectroscopy of the superconducting gap (1960) . . . 1771
91.6 Binnig and Rohrer: the scanning tunnelling microscope (1982) . . . 1773

95 The Positron **1807**

98 Neutrino Helicity **1831**

110 Parity Violation **2313**

111 CP Violation **2341**

112 The Electron Anomalous Magnetic Moment (2023) **2371**

113 The Higgs Boson Discovery (2012) **2401**

114 Neutrino Oscillations **2437**

116 Deep Inelastic Scattering **2519**
117.2 Rutherford: the first artificial transmutation (1919) . . . 2554
117.3 Cockcroft and Walton: lithium split by accelerated protons (1932) . . . 2559
117.4 Lawrence and Livingston: the cyclotron (1932) . . . 2566
117.5 Wilson: the expansion cloud chamber (1912) . . . 2570
117.6 Glaser: the bubble chamber (1952) . . . 2576
117.7 Charpak: the multiwire proportional chamber (1968) . . . 2580

119 Brownian Motion and Avogadro's Number **2595**

121 Bose–Einstein Condensation **2611**
122.8 Zewail: femtosecond probing of the transition state (1987) . . . 2623

128 The Quantum Hall Effect **2669**

131 Superconductivity **2697**
132.2 Shull and Smart: antiferromagnetic order in MnO (1949) . . . 2706
132.3 Brockhouse and Stewart: phonons in aluminium (1955) . . . 2708
132.4 Brockhouse: spin waves in magnetite (1957) . . . 2710
132.5 Brockhouse and Iyengar: normal modes of germanium (1958) . . . 2712
132.8 Hodgkin, Hauptman and Karle, and Klug: the phase problem (1934–1969) . . . 2716
136.2 LUX-ZEPLIN: dark-matter search in seven tonnes of liquid xenon (2023) . . . 2743
136.3 XENONnT: nuclear-recoil dark-matter search at the Gran Sasso laboratory (2023) . . . 2745
136.5 DAMA/LIBRA: annual modulation in sodium iodide, and the same-target replications (2008–2018) . . . 2747
136.6 ADMX: microwave-cavity search for halo axions (2018–2020) . . . 2749
136.7 ATLAS: jet plus missing transverse momentum search in proton–proton collisions (2021) . . . 2750
136.8 Fermi Large Area Telescope: search for annihilation gamma rays from dwarf spheroidal galaxies (2009–) . . . 2752

List of Pending Derivations

1 Classical Thermodynamics: all derivations of this chapter are pending. . . . 1415

2 Kinetic Theory of Gases: all derivations of this chapter are pending. 1423

3 The H-theorem: that the binary-collision integral of the Boltzmann transport equation makes the functional $H = \int f \ln f \, \mathrm{d}^3 v$ non-increasing in time, with stationarity exactly on the Maxwell distribution; together with an explicit statement of the molecular-chaos assumption on which the monotonicity rests, and of what the reversibility and recurrence objections do and do not refute. 1426

4 Experiment: The Mechanical Equivalent of Heat: all derivations of this chapter are pending. 1431

5 The Chapman–Enskog transport coefficients: that the first-order solution of the Boltzmann transport equation about local equilibrium fixes the numerical coefficients left undetermined by the mean-free-path argument, and in particular gives the ratio of thermal conductivity to viscosity times specific heat as $5/2$ for a monatomic gas, against the value 1 of the elementary estimate. . . 1445

6 The residual internal field of a paramagnetic salt: that the dipolar sum over the lattice of magnetic ions, together with the crystal-field splitting of the ground multiplet, fixes the ordering temperature and hence the floor under a single demagnetization stage, so that the cooling ratio is a computable property of the salt rather than a fitted parameter. 1454

7 The spectroscopic entropy of a gas: that the translational, rotational, vibrational and electronic partition functions, evaluated from band-spectrum constants and molecular masses, give the absolute molar entropy against which the calorimetric integration is compared, together with the symmetry number and the nuclear-spin terms that cancel between the two routes. . . 1454

8 Gradient identity Equation (58.12): direct differentiation of $|\boldsymbol{x} - \boldsymbol{x}'|^{-1}$ (asserted without proof in the source). 1459

9 Distributional Laplacian of $|\boldsymbol{x} - \boldsymbol{x}'|^{-1}$, Equation (58.17) (asserted without proof in the source). 1459

10 Continuity and interface conditions of the electrostatic field and the ideal-conductor boundary conditions, including Equation (58.25): the source reserves these headings without content, while its worked problems use the results. 1462

11 Ideal-dipole potential Equation (58.30), field Equation (58.31) and force Equation (58.32): to be derived from the multipole expansion of Section 58.2.5 and from the energy of a distribution in an external field, both of which the source reserves without content. 1465

12 Magnetostatics: all derivations of this chapter are pending. 1467

13 Inseparability of magnetic poles: derive, from the Amperian representation of a magnetized body by bound currents, that every subdivision of such a body again carries a complete circulating current and is therefore again a dipole, and connect the statement to the vanishing divergence of the magnetic field. 1468

14 Monopole searches: derive the flux quantum that an induction detector would register for the passage of one Dirac charge through a superconducting loop, and the conversion of a null exposure into a flux limit per unit area, time and solid angle. The Dirac quantization argument itself requires the quantum mechanics of Part IX and is stated, not derived, here. 1472

15 The Maxwell Equations: all derivations of this chapter are pending. 1473

16 The flux rule: derive it from the Maxwell–Faraday equation together with the Lorentz force acting on the charge carriers of a moving circuit, using the transport theorem for a surface whose boundary moves with the circuit, and show that the two physically distinct mechanisms — an induced electric field in the transformer case, a magnetic force on carried charge in the motional case — combine into the single total rate of change of the flux. 1473

17 Angular momentum of the electromagnetic field: derive the angular momentum density from the field momentum density, and show that for a circularly polarized beam of angular frequency ω the ratio of angular momentum to energy is $1/\omega$, so that reversing the handedness transfers twice that amount per unit energy. 1477

18 Electromagnetic Waves and Optics: all derivations of this chapter are pending. 1479

19 Brewster's angle: derive the Fresnel amplitude coefficients for the two polarizations from the interface conditions on the tangential fields, and show that the coefficient for light polarized in the plane of incidence has a simple zero exactly where the reflected and refracted directions are perpendicular, which by the sine law is at the angle whose tangent is the refractive index. 1482

20 The axial bright point: derive it from the Huygens–Fresnel superposition of secondary wavelets over the plane of the obstacle, showing that every point of the disc's rim is equidistant from the axial field point, so that the wavelets from the rim arrive exactly in phase whatever the radius of the disc, and evaluate the resulting on-axis amplitude for an ideal circular obstacle. . . . 1484

21 Dispersion: derive the frequency dependence of the refractive index from the response of bound charges to the field of the wave, obtaining normal dispersion away from resonance and anomalous dispersion with absorption near it, and recover the empirical Sellmeier form as its low-loss limit. The model belongs to the treatment of electrodynamics in matter, and the discreteness of the solar lines themselves is not classical at all and awaits the atomic chapters. 1485

22 Electrodynamics in Matter: all derivations of this chapter are pending. . . . 1487

23 The magnetic response of matter: derive the Bohr–van Leeuwen theorem, that a classical system of charges in thermal equilibrium has identically zero magnetization because the magnetic field enters the classical Hamiltonian only through a shift of the canonical momenta, which is absorbed by a translation of the integration variables in the partition function. The theorem states that the phenomenon above has no classical explanation at all, so every classical model in this section is a parametrization and not a derivation, and the real account is the quantum one of the magnetism chapter. 1490

24 Ferromagnetic hysteresis: derive the energy dissipated per unit volume per cycle as the area enclosed by the loop traced in the plane of applied field against magnetization, and identify precisely which hypothesis behind a linear constitutive relation the loop violates. The microscopic origin — domains, exchange coupling and magnetocrystalline anisotropy — is quantum and belongs to the magnetism chapter. 1490

25 Absorption and dispersion: derive the Kramers–Kronig relations, in which the real and imaginary parts of the permittivity are each the Hilbert transform of the other. The only physical input is causality — no response before the stimulus — which makes the response function analytic in the upper half of the complex frequency plane, whence a contour integral gives the pair of dispersion relations and the sum rules that follow from the high-frequency limit. The argument needs the complex analysis of the corresponding methods chapter, and it is worth noting that it uses no model of the medium at all. . 1492

26 The Faraday effect: derive the difference between the refractive indices of the two circular polarizations for a medium of bound charges in a magnetic field along the ray, by solving the driven oscillator equation with the magnetic force included; then show that a linear polarization, resolved into two circular components of equal amplitude, emerges rotated by half the accumulated phase difference, which gives the Verdet constant. The non-reciprocity should be derived too, and it follows from the field being an axial vector that does not reverse when the propagation direction does. 1493

27 Experiments: Electromagnetism: all derivations of this chapter are pending. 1495

28 The Coulomb exponent: derive the interior field of a uniformly charged spherical shell for a force law with exponent $2 + \delta$, show that it vanishes identically only when δ is zero, and give the conversion of a measured null voltage between concentric conductors into a bound on δ. Then relate δ to a photon rest mass through the Yukawa form that a massive vector field gives in place of the Coulomb potential. 1496

29 The ratio of the units: show that the electrostatic and electromagnetic definitions of the unit of charge differ by a factor with the dimensions of a speed, equal in SI terms to the reciprocal square root of the product of the electric and magnetic constants, and hence that this measurement determines exactly the combination that reappears as the propagation speed in the vacuum wave equation. The error budget of the original determination, and its agreement with the contemporary optical values, are to be reconstructed here. 1498

30 Generalized Classical Field Theory: all derivations of this chapter are pending. 1503

31 Radiation and Scattering of Electromagnetic Waves: all derivations of this chapter are pending. 1509

32 The Larmor formula: derive the fields of a point charge in arbitrary motion from the retarded potentials, separate the velocity field, which falls as the inverse square of the distance and carries no energy to infinity, from the acceleration field, which falls as the inverse first power and is proportional to the acceleration, and integrate the Poynting flux of the acceleration field over a sphere of large radius. 1510

33 Synchrotron radiation: derive the emission of a charge on a circular orbit at arbitrary speed from the acceleration field and the relativistic generalization of the Larmor formula, obtaining the forward collimation into a cone whose half-angle is the reciprocal of the Lorentz factor, the critical frequency proportional to the cube of that factor divided by the orbit radius, the power-law spectrum below it, the energy radiated per turn, and the degree and sense of the polarization. 1516

34 Plasmas and Magnetohydrodynamics: all derivations of this chapter are pending. 1519

35 Landau damping: solve the linearized Vlasov–Poisson initial-value problem by Laplace transform, analytically continue the dielectric function below the real axis to obtain the correct prescription for the singularity at the phase velocity, and derive the damping rate proportional to the slope of the distribution function there. Then give the resonant-particle reading — particles slightly slower than the wave gain energy and slightly faster ones lose it, and a Maxwellian has more of the former — and explain why the process is not thermodynamically irreversible, the information being carried into fine structure in velocity space and recoverable as a plasma-wave echo. 1523

36 Alfvén waves: linearize the ideal magnetohydrodynamic equations about a uniform field and a uniform density, retaining the tension part of the Lorentz force, and obtain the transverse mode propagating along the field at the speed above, together with the fast and slow magnetosonic modes and their anisotropic phase speeds. The tension reading should be derived as well — the magnetic force decomposed into a pressure and a tension along the field lines, whence the wave speed is the tension divided by the mass per unit length in exactly the form of a wave on a string. 1526

37 Magnetic reconnection: derive the Sweet–Parker current sheet from mass and flux conservation across a resistive layer, obtaining a reconnection rate scaling as the inverse square root of the magnetic Reynolds number, and show by substituting coronal values that this rate is orders of magnitude too slow for a flare. Then derive Petschek's slow-shock geometry with its much weaker logarithmic dependence, and state honestly which of the modern resolutions — plasmoid instability of thin sheets, or the collisionless Hall and pressure-tensor terms in Ohm's law — the measurements support. 1527

38 The solar wind: derive Parker's transonic solution. Integrate the steady, spherically symmetric momentum equation for an isothermal corona in the Sun's gravitational field, exhibit the critical point at which the flow speed equals the sound speed, show that of the family of solutions only the transonic branch matches both a subsonic base and a vanishing pressure at large distance, and obtain the asymptotic speed. The contrasting hydrostatic solution must be shown to leave a finite pressure at infinity, which is the argument that rules it out. 1528

39 Experiment: Wave Optics: all derivations of this chapter are pending. . . . 1531

40 The diffraction limit: derive the chromatic resolving power of a grating from the angular width of an order — the wavelength divided by the order times the number of rulings — using the N-aperture sum and the convention that two lines are resolved when the maximum of one falls on the first zero of the other. Then derive the imaging bound from the requirement that at least two diffracted orders from a periodic object be collected by the objective, which gives the smallest resolvable period as the wavelength divided by twice the numerical aperture, and state the hypotheses — linearity in the illumination, far-field collection — on which the bound depends. 1536

41 Diffraction efficiency of a hologram: what fraction of the reconstructing beam is delivered into the reconstruction, derived for a thin absorption record and a thin phase record by expanding the transmittance of a sinusoidal grating, and for a thick phase record by the coupled-wave treatment in which two waves exchange energy under the Bragg condition. What has to be established is that only a volume record can send almost all of the incident light into a single order — which is what makes a white-light reflection hologram bright enough to look at — while the thin records are bounded well below that. . 1543

42 Single-quantum interference: the derivation is quantum and belongs to the later parts. What must be shown here is negative and can be made precise — that no classical field theory reproduces the combination of a single localized detection with a fringe pattern of full visibility, since the second-order correlation at zero delay is bounded below by unity for any classical field, while the measured value for a heralded single photon falls well below it. The wave account of this chapter must therefore be read as the high-occupancy limit of something else. 1545

43 Black-Body Radiation and Planck's Hypothesis: all derivations of this chapter are pending. 1549

44 Wien's adiabatic argument: a slow compression of a perfectly reflecting cavity leaves the entropy unchanged while shifting every mode frequency, which forces the universal function of Kirchhoff's theorem into the scaling form given by ν cubed times a function of ν/T alone; and the maximisation of the corresponding wavelength density, which turns that scaling form into a fixed product of peak wavelength and temperature. 1552

45 The low-frequency limit of the cavity spectral density: that each mode carries the equipartition energy given by Boltzmann's constant times the temperature, so that the spectral density becomes proportional to the square of the frequency times the temperature, and the radiance at fixed long wavelength is therefore linear in T. 1553

46 The quantum theory of lattice heat capacities: the mean energy of a quantised oscillator applied to the vibrational degrees of freedom of a crystal, which gives the exponential freeze-out of a single Einstein frequency; and the replacement of that single frequency by the elastic-continuum mode spectrum with a cut-off preserving the total mode count, which produces the observed cubic low-temperature law. 1556

47 Why the expanding universe leaves a Planck spectrum Planckian: thermal equilibrium of matter and radiation before recombination fixes the form, and the redshift of every mode by a common factor rescales the spectral density into another Planck spectrum with the temperature inversely proportional to the scale factor, so that no spectral distortion is generated by the expansion itself. The limits on distortions then become limits on energy release after equilibrium was lost. 1557

48 The Photon: Photoelectric and Compton Effects: all derivations of this chapter are pending. 1559

49 The classical accumulation time: the energy an electromagnetic wave of a stated irradiance delivers per second to a target of atomic cross-section, compared with the work function of the metal, giving the delay a wave theory requires before the first electron can be emitted; and the demonstration that no redistribution of energy within the classical field shortens it, which is what makes item 4 above a contradiction rather than an estimate. 1560

50 The angular momentum of a circularly polarized wave: the flux of angular momentum computed from the Maxwell stress tensor for a beam of finite transverse extent, its ratio to the energy flux, and the torque this predicts on a half-wave plate that reverses the handedness; and the identification of that ratio with one unit of the reduced Planck constant per quantum, which is what makes the result a statement about the spin of the photon rather than about the field alone. 1563

51 The angular correlation implied by the collision kinematics: the recoil-electron direction as a function of the scattering angle of the quantum, obtained from the same conservation laws that give the wavelength shift; the rate of accidental coincidences to be subtracted from the observed rate, given the counting rates and the resolving time; and the contrasting prediction of statistical-only conservation, under which the two directions are independent and the coincidence rate is the accidental rate alone. 1565

52 The two bounds that make the measurement decisive: the classical inequality on the normalised joint detection rate, from the Cauchy–Schwarz inequality applied to the intensity correlations of any positive classical intensity distribution; and the quantum prediction that a one-quantum state incident on a beam splitter gives a vanishing joint rate, together with the correction for the imperfect heralding and the accidental coincidences of a real source. . . 1566

53 Atomic Models and Spectra: all derivations of this chapter are pending. . . 1567

54 Two things must be derived and only the second is classical. That a bound system radiates at a discrete set of frequencies at all follows only from the quantisation of its internal energies, and is taken up in the quantum-mechanics part; that emission and absorption occur at the same wavelengths is the detailed-balance content of Kirchhoff's radiation theorem, and follows from the equilibrium argument of the black-body chapter applied frequency by frequency. 1568

55 The term structure: that there exists a set of stationary energies of the atom whose differences, divided by Planck's constant, are exactly the observed frequencies. The stationary-state postulate assumed in the Bohr section below turns the combination principle into a law rather than deriving it; the derivation proper is the eigenvalue problem for the atomic Hamiltonian, solved for hydrogen in the quantum-mechanics part. 1569

56 The Rutherford cross-section: the relation between impact parameter and scattering angle for a repulsive inverse-square force, read off the Kepler hyperbola of the classical two-body problem, and the Jacobian that converts it into a differential cross-section; the summation over foil thickness that gives the counting rate actually measured; and the remark, which must be made honestly, that the same expression is also the exact Born approximation for a Coulomb potential, so that a mechanics now known to be wrong gave the right answer here for a reason no one could have known in 1911. 1570

57 Stability of the atomic ground state: the existence of a lowest energy level, which no classical bound system in an inverse-square potential possesses, and the bound relating the localisation of the electron to its kinetic energy which fixes the size and energy of that level. The derivation belongs to the quantum-mechanics part, where the hydrogen atom is solved; what can be established here is only the negative statement of the remark below. 1571

58 Two derivations of different standing. The normal triplet follows classically, from the Larmor precession that a magnetic field imposes on a bound charge: the three normal modes of an isotropic oscillator in a field, their frequencies, and the circular and linear polarizations of the light each of them emits. The anomalous patterns do not follow at all without the electron spin and the coupling of spin to orbital angular momentum, and their treatment belongs to the angular-momentum chapter of the quantum-mechanics part; what must be recorded here is that the empirical splitting factor was fitted, not derived. 1575

59 Matter Waves: all derivations of this chapter are pending. 1577

60 Diffraction of a thermal beam: the grating condition for a two-dimensional surface lattice, and the shape of the observed maxima when the incident beam is not monochromatic but carries the Maxwellian velocity distribution of the oven, each velocity class contributing its own de Broglie wavelength. What must be shown is that the position of each maximum tracks the most probable momentum in the predicted way as the source temperature is varied, which is the quantitative content of the experiment. 1579

61 The accumulation statistics: that the arrival positions are independent samples drawn from the squared modulus of the single-particle amplitude, so that the histogram of many arrivals converges on the intensity pattern of the wave while no individual arrival carries any fringe at all. What must be derived is the convergence itself and its rate, and what must be stated is that the probabilistic reading of the amplitude is an additional postulate and not a consequence of the de Broglie relation. 1580

62 The gravitationally induced phase: the local wavenumber of a neutron in a linear potential, obtained from the free-particle dispersion relation with the gravitational potential energy subtracted from the total energy; the phase accumulated along each arm as the line integral of that wavenumber; and the reduction of the difference to the stated product of the squared mass, the acceleration of free fall, the enclosed area and the wavelength. The systematic corrections that must accompany it — bending of the crystal under its own weight as the device is rotated, and the Sagnac phase from the Earth's rotation — belong with it. 1581

63 The rotation operator for a spin-half state: that the generator of rotations has half-integer eigenvalues, so that a rotation by a full turn acts as minus the identity rather than the identity, and that a spin-half state therefore carries a representation of the covering group and not of the rotation group. The interferometric argument that turns this global sign into a measurable relative phase must accompany it, together with the relation between the applied magnetic field, the transit time and the precession angle. 1582

64 The eigenvalue problem for a particle in a linear potential above a hard wall: the equation reduces to Airy's equation, the admissible solutions are the decaying Airy function shifted so as to vanish at the mirror, and the eigenvalues are therefore fixed by its zeros. The classical turning-point heights follow from the eigenvalues, and it is those heights that the absorber measures; the step positions predicted this way are to be compared with the observed ones. 1582

65 Two things, of which only the first is optics. The near-field Talbot–Lau fringe condition and its scaling with mass and wavelength, which is what fixes the grating periods and the source coherence a given molecular mass demands. And the decoherence criterion: how much which-path information the emission or scattering of a single photon of given wavelength carries, and hence the residual gas pressure and internal temperature at which the fringes must disappear — the quantitative statement of what these experiments do and do not bound. 1583

66 Experiment: Franck–Hertz: all derivations and data analysis of this chapter are pending. 1585

67 Why the glow layers sit where they do: an electron starting from rest at the cathode reaches the excitation energy after falling through a fixed potential difference, so the luminous shells mark equal potential intervals, and their migration towards the cathode as the voltage is raised follows from the field distribution in the tube. The dependence of the effective spacing on current and geometry requires in addition the branching among the accessible neon levels and the collision cross-section of each. 1589

68 The effective spacing: the excitation cross-sections of the competing mercury levels as functions of electron energy, and the weighted mean interval they produce in a tube of given vapour density and geometry; and the mechanism of the drift with maximum number, in which electrons surviving several accelerating stages sample a different mixture of levels from those excited near the first drop. 1590

69 Experiment: The Blackbody Spectrum Measured: what is pending is the tabulation of the measured radiance against wavelength and temperature for each experiment, in SI with its uncertainty budget, together with the residuals against the Wien and the Planck distributions. The phenomena stated below each carry their derivation inline. 1591

70 Experiment: The Photoelectric Effect and Compton Scattering: what is pending is the tabulation of the measured numbers for each experiment, in SI with its uncertainty budget — stopping potential against frequency for each emitting metal, scattered wavelength against scattering angle, coincidence and singles counts against resolving time, and the gated detection probabilities of the beamsplitter experiment. The phenomena stated below carry their derivations inline except where a pending derivation is marked against them individually. 1603

71 The operational definition of the stopping potential. The measured current does not fall to zero at a sharp potential but approaches it along a tail, because the emitted electrons have a distribution of energies rather than a single maximum, because electrons originate at a range of depths below the surface, and because the collector is not equipotential. What is pending is the derivation of the shape of that current–voltage curve near its foot from the energy distribution of the photoelectrons, and the demonstration that the extrapolation Millikan used recovers the true intercept without bias — the point being that the linearity of the plotted line is a statement about a quantity defined by an extrapolation, and that extrapolation must be shown not to have manufactured it. 1609

72 The angular correlation measured by the cloud-chamber experiment. What is pending is the direction of the recoil electron as a function of the scattering angle of the quantum, obtained from the same conservation laws that give the wavelength shift, together with the angular spread introduced by the initial motion of the struck electron and by multiple scattering of the recoil track in the chamber gas — the latter being what sets the accuracy with which the predicted correlation can be confirmed, and hence what the measurement is really comparing. 1614

73 Experiment: Matter Waves Observed: what is pending is the apparatus and procedure of each measurement and the tabulation of its data — scattered intensity against angle and accelerating voltage, ring radii against voltage, diffraction angle against source temperature, fringe visibility against count, neutron intensity against tilt angle, and molecular fringe contrast against internal temperature and residual gas — each in SI with its uncertainty budget. The phenomena stated below each carry their derivation inline. . . 1619

74 The quantitative decoherence rates: the thermal photon emission spectrum of a hot molecule and the resulting product of which-path factors as a function of internal temperature, giving the predicted visibility curve; and the collisional case, in which the localization length follows from the momentum transfer distribution of the residual gas at a given pressure and temperature, giving the predicted decay of visibility with pressure. Both are stated qualitatively in the derivation above from the single-photon overlap; what must be supplied is the master equation whose decoherence rate they are, together with the statement of which of its assumptions the data actually test. 1631

75 X-ray spectra and the charge of the nucleus: the classical account of the emission itself, namely the bremsstrahlung continuum radiated by cathode-ray electrons decelerated in the anticathode, and its short-wavelength cut-off. . 1637

76 Characteristic X-radiation: the quantum account, in which the primary beam ejects an electron from an inner shell and the radiation is emitted when an outer electron fills the vacancy, so that the excitation threshold is the binding energy of the shell and the monotone hardening along the table is the growth of that binding energy with nuclear charge. 1639

77 The X-ray term system: the derivation of the K, L and M level multiplicities from the quantum numbers of the many-electron atom, with the L and M splittings as the combined effect of screening and of spin-orbit coupling, and the identification of an absorption edge with a one-electron binding energy. 1646

78 From Poisson brackets to commutators: statement of the canonical quantization rule and derivation of the canonical commutation relations between conjugate coordinates, which are used throughout the following chapters. . 1653

79 Justification of the coordinate-representation rule $p_i = -\mathrm{i}\hbar\,\partial_i$ from the canonical commutation relations; the source asserts the rule without derivation. . 1653

80 Statement and derivation of the generalized uncertainty relation bounding the product of the uncertainties of two non-commuting observables by the mean value of their commutator. 1654

81 Elementary Quantum Systems: all derivations of this chapter are pending. . 1655

82 Resonant transmission by an attractive well: the condition under which the amplitudes reflected from its two edges cancel, making it perfectly transparent, and the three-dimensional form of the same statement as an s-wave phase shift passing through a half turn at the Ramsauer minimum 1658

83 The Gamow factor: semiclassical penetration of the Coulomb barrier by a particle whose energy lies below its top, the exponential of the integrated imaginary momentum across the forbidden region, and the linear dependence of the logarithm of the decay constant on the inverse square root of the alpha energy that follows from it 1658

84 Angular Momentum and Spin: all derivations of this chapter are pending. . 1661

85 Landé's g factor from the projection theorem: that the matrix elements of any vector operator inside one multiplet are proportional to those of the total angular momentum, so that the observed splitting unit becomes a closed formula in the three quantum numbers of the level, and that fitting it to the data requires the electron's own gyromagnetic factor to be twice the orbital one 1665
86 Force on a magnetic dipole in an inhomogeneous field 1673
87 Quantum derivation of the two-valued splitting from spin-1/2 (to Chapter 79) 1674
88 Sequential SG probabilities from the measurement postulates 1675
89 The Hydrogen Atom: all derivations of this chapter are pending. 1677
90 The bound-state eigenvalues of the Coulomb problem: the radial equation for the potential that falls as the inverse first power of the distance, the termination of its series solution which quantizes the energy as minus a constant over the square of an integer, and the identification of that constant with a combination of the electron mass, the elementary charge and Planck's constant 1678
91 The ℓ-independence of the Coulomb spectrum: the conserved Runge–Lenz vector peculiar to the inverse-square-law force, the closure of its commutators with the angular momentum on the algebra of SO(4) for bound states, and the identification of the levels with irreducible representations of that group whose dimension is the square of the principal quantum number 1680
92 The radiative level shift: the interaction of the bound electron with the fluctuating quantized electromagnetic field, the difference between the bound and free self-energies which is what remains finite after regularization, and the vacuum-polarization contribution of opposite sign that accompanies it . 1681
93 The ground-state hyperfine splitting of hydrogen: the Fermi contact interaction between the electron and nuclear magnetic moments, its expectation value in the ground state, which is proportional to the probability density of the electron at the origin, and the resulting separation of the singlet from the triplet 1682
94 Approximation Methods: all derivations of this chapter are pending. 1685
95 Identical Particles: all derivations of this chapter are pending. 1691
96 The Ramberg–Snow bound: the energy displacement of the anomalous X-ray line produced by the extra screening of a doubly occupied inner shell, the number of fresh electrons that a given charge delivers into the target lattice, and the conversion of a null count and its background into an upper limit on the violation probability 1696
97 The two-particle correlation of a chaotic source: the sum over the two indistinguishable histories by which one particle from each of two source points reaches the two detectors, the factor of two it produces in the zero-delay coincidence rate for symmetric states, and its replacement by zero for antisymmetric ones, together with the loss of the effect once the path difference exceeds the coherence length 1697
98 The degenerate electron gas: its zero-temperature pressure obtained from the exclusion principle and the filling of momentum space up to the Fermi momentum, the change of the exponent relating pressure to density once that momentum exceeds the electron mass times the speed of light, and the resulting limiting mass, which is of the order of the cube of the Planck mass divided by the square of the nucleon mass 1698
99 Entanglement and Bell Tests: all derivations of this chapter are pending. . . 1699
100 Interpretations (Evidence-Anchored): all derivations of this chapter are pending. 1707
101 Scattering Theory: all derivations of this chapter are pending. 1713

102 Open Quantum Systems and Decoherence: all derivations of this chapter are pending. . . . 1723
103 Eigenstate thermalization: that the expectation of a local observable in an individual many-body energy eigenstate is a smooth function of the energy density alone, so that any superposition confined to a narrow energy window relaxes to the microcanonical value of every local observable; and that the entanglement entropy of a small subregion then grows to the thermodynamic entropy that subregion would carry at the corresponding temperature . . . 1727
104 Atoms and Molecules: all derivations of this chapter are pending. . . . 1733
105 The Heitler–London calculation: that the symmetrized and the antisymmetrized products of two hydrogenic ground-state orbitals have expectation energies differing by twice an exchange integral; that this integral has the sign, over the relevant range of internuclear distance, which puts the singlet below the separated atoms and the triplet above them; and the size of the discrepancy that remains between so simple a variational estimate and the measured dissociation energy . . . 1739
106 Experiment: Bell Tests: all derivations of this chapter are pending. . . . 1745
107 The detection-efficiency threshold: the least heralding efficiency at which the observed coincidence and singles rates cannot be reproduced by any local model free to decide which pairs go undetected, worked out both for the CHSH inequality with a maximally entangled state and for the Eberhard inequality with a partially entangled one, the latter threshold being the lower and being the reason the photonic experiments deliberately used a non-maximally entangled state . . . 1751
108 Experiment: Precision Spectroscopy and Atomic Clocks: all derivations of this chapter are pending. . . . 1753
109 Magnitude of the interband tunnel current: the transmission of a carrier across the forbidden gap of a uniformly biased semiconductor, the imaginary wavevector inside the gap obtained by continuing the two-band dispersion relation, and the resulting exponential dependence of the peak current on the gap, the effective mass and the junction field . . . 1770
110 The tunnel current between a normal metal and a superconductor: the elastic-tunnelling current as the convolution of the two densities of states with the difference of the two Fermi functions [1198], the BCS quasiparticle density of states with its square-root edge at the gap, and the consequence that the differential conductance at low temperature is proportional to that density of states evaluated at the bias energy . . . 1772
111 The tunnel current as a probe of the local density of states: the transfer-Hamiltonian expression for the current between two weakly coupled electrodes in terms of the overlap of their surface wavefunctions [1198], and its reduction, for a tip modelled by a single spherically symmetric orbital, to the local density of states of the sample at the position of the tip and at the bias energy [1765] . . . 1774
112 The Klein–Gordon Equation: all derivations of this chapter are pending. . . . 1779
113 Determination of the spin and parity of the boson near 125 GeV from the angular distributions of its four-lepton and diphoton decays, and the electroweak argument that requires an elementary scalar field in the first place. The experiment is the Higgs discovery chapter of Part XII and the mechanism is the electroweak chapter of the same part. . . . 1786
114 The Dirac Equation: all derivations of this chapter are pending. . . . 1787

115 Separation of the Dirac equation in a Coulomb potential and solution of the resulting radial system in closed form, together with its expansion in powers of the nuclear charge times the fine-structure constant, which must reproduce the Bohr levels at leading order and the fine-structure intervals at the next. 1791

116 The radiative level shift between the two lowest hydrogen states of equal total angular momentum. Deriving it requires the quantized electromagnetic field, the electron self-energy and the vacuum polarization, and a renormalization prescription, so it lies outside relativistic quantum mechanics altogether and belongs to the quantum-electrodynamics chapter of Part XII. 1792

117 Bargmann's theorem for the Poincaré group: the proof computes the second cohomology of the Lie algebra $\mathfrak{iso}(3,1)$ and shows every local phase cocycle is a coboundary, then lifts to the simply connected cover; see [1029]. The cohomological machinery is not yet developed in this treatise. 1796

118 Completeness of the classification: that the representation defined by Equations (94.21) and (94.22) is irreducible precisely when D is, and that *every* irreducible unitary representation of the cover with $\hat{P}$-spectrum on the given orbit is equivalent to an induced one, is Mackey's imprimitivity theorem, specialized by Wigner to this group [849]. The systematic imprimitivity machinery is not yet developed in this treatise. 1803

119 Experiment: The Positron: all derivations of this chapter are pending. . . . 1807

120 The identification of the penetrating cosmic-ray component as a heavy lepton rather than as the quantum of the nuclear force: its anomalously weak absorption in nuclei against the strong coupling a nuclear-force carrier must have, and the three-body shape of its decay electron spectrum. The muon's own phenomenology is taken up in the cosmic-ray chapter of Part XII. . . . 1812

121 The CPT theorem — that any local, Lorentz-invariant quantum field theory with a Hermitian Hamiltonian and the normal spin–statistics connection is invariant under the combined operation of charge conjugation, parity and time reversal — together with the equalities of mass, lifetime and magnetic moment that follow from it, and the reasons a null result here bounds rather than confirms. 1812

122 The Weyl Equation and Neutrinos: all derivations of this chapter are pending. 1815

123 Extraction of the invisible width of the Z from the measured total and visible widths at the resonance peak, and the Standard Model expression for the partial width into one neutrino species, whose ratio is the species count. The measurement requires the electroweak theory of Part XII to convert a width into a number. 1818

124 The shape of the beta spectrum near its endpoint as a function of the neutrino mass, and the reason the observable is the incoherent sum of squared mixing amplitudes times squared masses rather than any single mass; and, separately, the free-streaming suppression of the matter power spectrum by massive neutrinos, whose derivation belongs to the cosmology part and whose model dependence must be stated with the bound. 1821

125 The two-neutrino and neutrinoless double beta decay rates: the phase-space and nuclear-matrix-element factorization, the effective Majorana mass that the neutrinoless rate measures, and the argument that observing the decay implies a Majorana mass term whatever mechanism drives it. The nuclear matrix elements are the dominant systematic and must be presented as such. 1822

126 Higher-Spin Wave Equations: all derivations of this chapter are pending. . . 1825

127 Experiment: Neutrino Helicity: all derivations of this chapter are pending. . 1831

128 The Michel spectrum: the most general Lorentz-invariant four-fermion matrix element for muon decay, its reduction to four measurable parameters in the decay distribution, and the values those parameters take when the current is the purely left-handed one. Deriving it also shows which combinations of couplings the experiment cannot separate, which is what makes the bounds on right-handed admixtures the shape they are. 1835

129 The parity-violating asymmetry in the scattering of longitudinally polarized electrons, as the interference between the photon-exchange and neutral-current amplitudes, and its expression in terms of the weak mixing angle; together with the analogous expression for the parity-non-conserving amplitude in a heavy atom, where the nuclear weak charge and the atomic structure factorize. Both belong to the electroweak chapter of Part XII. 1836

130 The general left–right parametrization of the charged current, the observables in beta and muon decay that are first order in the right-handed admixture as against those that are second order, and the translation of a null asymmetry into a bound on the mixing angle and on the mass of an additional charged vector boson. The distinction between the two orders is what determines how the sensitivity scales, and it must be stated with the bounds. 1836

131 The indefinite-metric mode algebra of covariant quantization, displayed just above: that the gauge-fixing term makes all four components of the potential dynamical, that the covariant equal-time bracket $[A_\mu, \pi^\nu]$ then delivers $-\eta_{\lambda\lambda'}$ where the Coulomb-gauge algebra has $\delta_{\lambda\lambda'}$, and hence that the timelike mode has negative norm. It needs a basis of four polarization vectors orthonormal in the Minkowski metric, which this chapter does not set up, so it belongs in the long-proof appendix. 1863

132 Consistency of the Gupta–Bleuler subsidiary condition displayed just above: that the subspace it selects carries a non-negative norm, that timelike and longitudinal quanta occur in it only in zero-norm combinations, and that those combinations contribute nothing to any observable, so that the counting reduces to the two transverse polarizations. It runs through the mode decomposition of the gauge-fixed field and belongs in the long-proof appendix, next to the indefinite-metric algebra above. 1864

133 The spin–statistics theorem for arbitrary spin, by analytic continuation of the Wightman functions to the Jost points: this is a long argument and belongs in the long-proof appendix. The two cases relevant to the Standard Model — spin 0 and spin 1/2 — are proved in full in the text below. 1873

134 Haag's theorem: the proof runs through the equality of all equal-time vacuum expectation values under the unitary, the reconstruction of the theory from them, and the identification of the mass from the two-point function. It is a page or two of distribution-theoretic argument and belongs in the long-proof appendix; the statement and its consequences are used here, and the axiomatic setting in which the hypotheses are stated is the axiomatic quantum field theory chapter. 1879

135 Klein–Nishina cross-section: the spin-summed trace computation of the two-diagram Compton amplitude leading to the squared matrix element and, after the rest-frame kinematics, to the angular distribution; the same computation gives the total cross section as a function of the ratio of photon energy to electron rest energy. It is a page of Dirac-algebra manipulation of the kind carried out in full for the vertex correction below, and belongs in Appendix A. 1894

136 Tree-level cross-sections: the spin-summed trace computations for Møller, Bhabha, muon-pair and two-photon annihilation, and the Born-approximation Bethe–Heitler bremsstrahlung cross-section together with the definition of the radiation length that follows from it. Each is an application of the trace technology established in the next subsection and belongs in Appendix A. . 1895

137 The field-strength renormalization as a residue, and its bound: the Källén–Lehmann spectral representation of the exact fermion two-point function, obtained by inserting a complete set of eigenstates of the four-momentum and using the spectrum condition; the positivity of the spectral density, which follows because each weight is a squared modulus of a vacuum-to-state matrix element; the isolated one-particle delta whose weight is the field-strength constant, identifying it with the residue of the exact propagator at its pole; and the sum rule from the canonical equal-time anticommutator, which normalizes the total spectral weight to one and hence bounds the constant between zero and one, the deficit being the weight carried by the multiparticle continuum. The subsection of the axiomatic chapter named for this representation is at present a reserved heading, so the derivation is owed there or in Appendix A. 1904

138 The Bogoliubov–Parasiuk–Hepp–Zimmermann theorem: the recursive subtraction, the combinatorial proof that the forest formula solves the recursion, and Hepp's sector decomposition establishing absolute convergence of the subtracted integrand, together with the demonstration that overlapping divergences require no counterterm beyond those of the local Lagrangian. This is a long combinatorial argument and belongs in Appendix A; the statement is used here only to license the claim that the four constants suffice at every order. 1911

139 Bloch–Nordsieck cancellation: the explicit infrared part of the one-loop form factor with a photon mass regulator, the integral of the eikonal emission rate up to the detector resolution, the demonstration that the coefficient functions of the two logarithms are equal and opposite, and the exponentiation of the soft series into the Sudakov form factor. The structure and its physical reading are established above; what is owed is the coefficient function, and it belongs in Appendix A. 1913

140 Kinoshita–Lee–Nauenberg theorem: the classification of the singular regions of a Feynman integral in the massless limit — soft and collinear — and the proof that the sum over the degenerate subspace of initial and final states is free of them, together with the counting that shows why an average over degenerate initial states is needed and not only a sum over final ones. It belongs in Appendix A. 1914

141 Hadronic vacuum polarization from data: the once-subtracted dispersion relation for the renormalized polarization function in the cut plane, with the subtraction fixed by the renormalization condition at zero momentum transfer and the convergence of the integral argued from the behaviour of the function at infinity; the optical theorem identifying the discontinuity across the timelike cut with the measured ratio of the annihilation cross section into hadrons to that into muon pairs; and the resulting integral over that measured ratio, which is the number quoted above as the hadronic half of the shift in the coupling at the mass of the Z boson. Only the analyticity input is proved in the treatise, in the Fourier and integral-transforms chapter; the dispersion integral itself is owed, and belongs in Appendix A. 1916

142 Lamb shift, covariant completion: the relativistic one-loop self-energy of a bound electron, which supplies the constant accompanying Bethe's logarithm and fixes the cutoff-independence of the result; the binding corrections organized as an expansion in $Z\alpha$; and the finite-nuclear-size term through which the proton charge radius enters the comparison. The non-relativistic leading term and the vacuum-polarization term are derived in full above; this box covers the remainder, and it belongs in Appendix A. 1919

143 Light-by-light scattering: the explicit four-photon amplitude from a closed charged-fermion loop, the demonstration by gauge invariance that the six box diagrams combine into a finite answer despite a superficial degree of divergence of zero, the expansion of the one-loop effective action of a constant field yielding the coefficients in the Euler–Heisenberg Lagrangian, and the phase-space integration giving the low-energy cross-section. It belongs in Appendix A. 1921

144 Bound-state QED: the annihilation rates of para- and ortho-positronium quoted above, the ground-state hyperfine splitting of positronium and of muonium, and the systematic expansion of a two-body bound-state energy in powers of α and $\ln \alpha$ that organizes them. The selection rule forbidding a two-photon decay of a $J = 1$ state is cited rather than derived here and is proved in the discrete-symmetry chapter. This box belongs in Appendix A. 1923

145 Muon decay, the parts not derived here: the phase-space function $f(z)$ of the electron-mass correction, obtained by repeating the phase-space integral of this section with $p_e^2 = m_e^2 c^2$ retained; the one-loop electromagnetic correction $[25/4 - \pi^2]$, which is a QED calculation of the kind carried out in the QED chapter but applied to a contact operator; and the spin-dependent decay distribution $x^2[(3 - 2x) + P_\mu \cos\theta\,(2x - 1)]$ obtained by keeping the muon spin projector in place of the average taken in the trace lemma. All three are Appendix A material and all three are quoted with their numerical values in the derivation above. 1954

146 The $[\Delta I = 1/2]$ enhancement in nonleptonic strange decays: why the isospin-one-half amplitude dominates the isospin-three-halves one by a factor near twenty, when the current–current operator produces both with comparable coefficients. The explanation combines short-distance operator mixing under the renormalization group with long-distance hadronic matrix elements, so the derivation belongs with the strong-interaction and renormalization-group material rather than here; the measured ratio is quoted with its source above. 1957

147 The hadronic neutral-current ratio: the prediction for the ratio of neutral-current to charged-current rates on an isoscalar nuclear target, which is what the Gargamelle hadronic sample actually measured. It requires the quark-parton content of the nucleon and the quark neutral-current couplings derived in this section, and therefore belongs with the deep-inelastic material rather than here; the leptonic ratio, which needs nothing beyond this chapter, is derived above. 1964

148 Renormalizability of Yang–Mills theory. The proof combines the power-counting theorem, which for a gauge theory in four dimensions gives superficial degree of divergence four minus the number of external lines weighted by their dimension exactly as in the abelian case, with the Slavnov–Taylor identities, which show that the divergent structures allowed by power counting are constrained to be gauge invariant and therefore already present in the Lagrangian. The regularization must respect the gauge symmetry, and dimensional regularization does. This is a long argument of the same shape as the BPHZ theorem of the electromagnetic chapter and belongs beside it in Appendix A. 1984

149 Collinear factorization to all orders. What is available inline is the leading-order statement, which is the impulse approximation together with the observation that the collinear logarithms are universal because they come from the region where the emitted parton is nearly on shell and therefore long-lived compared with the hard collision. The all-orders proof requires showing that the leading regions of the loop momenta are exhausted by hard, collinear and soft subgraphs, that the soft gluons cancel or are absorbed into eikonal lines, and that what remains is the same operator matrix element in every process. That analysis, and the counterexamples at subleading power, belong in Appendix A. 1998

150 The exact order-α_s three-jet distribution. What is derived inline is the soft and collinear structure, which is where the rate is concentrated and which already excludes a scalar gluon. The full matrix element requires the spin-summed trace over the two interfering diagrams in which the gluon is emitted from the quark and from the antiquark, together with the three-body phase space in the Dalitz variables; and the corresponding statement for a scalar emitter, needed to make the exclusion quantitative rather than qualitative. Both are trace computations of the kind carried out for the electromagnetic processes and belong beside them in Appendix A. 2001

151 Confinement: no analytic derivation from the continuum Lagrangian exists, and none should be implied. What is available is Wilson's area law for the loop expectation value, exact in the strong-coupling lattice expansion, together with the numerical demonstration that the lattice theory joins continuously onto the asymptotically free continuum. Closing the gap between those two statements is an open problem, not an exercise. 2005

152 The short-distance two-W box amplitude for the kaon mass difference. What is carried out inline is the GIM structure of the internal sum, the dimensional check of every factor, and the numerical evaluation; the loop integral itself — the box function in its limit of internal mass small against the W mass, together with the reduction of the four-quark operator to the decay constant and the bag parameter — is not derived here and belongs in Appendix A. . 2019

153 The phase-space function $f(x) = 1 - 8x + 8x^3 - x^4 - 12x^2 \ln x$ for a purely leptonic charged-current decay with a *massive* daughter lepton. The muon-decay rate inherited from the weak-interactions chapter is derived there only for a massless daughter, where $f(0) = 1$; the massive case is the one used above, and the three-body phase-space integral that produces it belongs in Appendix A. 2022

154 The tree-level partial width of the Z into a fermion pair. The formula is stated above with its dimensional check and then used to count the neutrino species; its derivation from the neutral-current Lagrangian — the squared matrix element summed over final and averaged over initial polarizations, times the two-body phase space — is not carried out here and belongs in Appendix A. 2025

155 The loop function $S_0(x)$ and the B-meson box amplitude for the mass difference. Both are stated above and then used twice to weigh the top quark, so the derivation is load-bearing: the box integral for an internal mass comparable with the W mass, whose value is S_0, is not carried out here and belongs in Appendix A. 2034

156 The wave-packet coherence length for neutrino oscillation, including its factor $4\sqrt{2}$. What is given inline is the group-velocity separation, which fixes the scaling with the packet width, the energy and the squared-mass splitting; the Gaussian overlap integral that fixes the numerical coefficient, and with it the precise meaning of the packet width, is not carried out here and belongs in Appendix A. 2040

157 The one-loop coefficient in the top contribution to the rho parameter: the W and Z vacuum-polarization integrals with a t–b doublet in the loop, regulated dimensionally, whose difference at zero momentum transfer gives the factor $3/8\sqrt{2}\pi^2$ quoted from Veltman's paper above; and the accompanying logarithmic dependence on the scalar mass, which is what the screening theorem asserts. Belongs in Appendix A. 2077

158 The loop-induced scalar vertices: the amplitudes for $h \to \gamma\gamma$ and $h \to gg$, from the W and top triangles, which have no tree-level counterpart and which supply the discovery channel and the dominant production mechanism (gluon fusion) respectively; together with the resulting production cross-sections at the LHC. These require the full one-loop tensor reduction and belong in Appendix A. Also owed are the three-body off-shell rates $h \to WW^*$ and $h \to ZZ^*$, whose branching fractions the table quotes without derivation. What is derived above is the two-body tree-level fermionic width formula and the tau and muon entries computed from it; every other entry of the branching table quotes the standard values. 2089

159 Vacuum stability: the coupled integration of the one-loop renormalization-group equations for the quartic coupling, the top Yukawa and the three gauge couplings from the electroweak scale upward, which is what locates the zero of the quartic near 1×10^{10} GeV; and the bounce action governing the tunnelling rate out of the electroweak minimum. The beta function and its sign are derived in the text above; the integration and the bounce belong in Appendix A. 2093

160 The axiomatic CPT theorem: Jost's proof requires the Wightman axioms, the analyticity of the Wightman functions in the forward tube, the extended tube and the edge-of-the-wedge theorem. None of that machinery is developed here; the axiomatic-QFT chapter reserves it, and the proof belongs in Appendix A alongside it. What is proved in this chapter is the Lagrangian-level theorem, which is weaker in that it assumes a Lagrangian, a field content and a perturbative expansion. 2127

161 The spin–statistics theorem for arbitrary spin: the free-field computation sketched here settles spin 0 and spin 1/2, but the general statement requires the analyticity of the two-point Wightman function and the same extended-tube machinery as the axiomatic CPT proof. It belongs in Appendix A next to it, and the identical-particle chapter of Part IX states the result it needs. 2130

162 The triangle anomaly and the Chern–Simons current: the derivation that the baryon current's divergence is proportional to the topological density, whether by the triangle diagram or by the non-invariance of the path-integral measure, is not carried out in this chapter. It belongs to the path-integral chapter of this part, which reserves the anomaly, and to Appendix A. What is used here is only the integrated statement, that a change of Chern-Simons number by one unit changes baryon number by the number of generations. 2134

163 Osterwalder–Schrader reconstruction: the derivation above establishes the easy half in quantum mechanics — that reflection positivity is the positivity of the reconstructed inner product — and states the field-theoretic theorem. The construction of the Wightman functions from the Schwinger functions by analytic continuation in the time arguments, and the verification of the spectrum condition and locality, are not carried out here; they belong in Appendix A, alongside the axiomatic material of the axiomatic quantum field theory chapter of this part. 2153

164 The Gribov horizon and Singer's obstruction: the derivation above establishes the criterion — a copy exists exactly when the Faddeev–Popov operator acquires a normalizable zero mode — and states that configurations reaching it exist. Gribov's explicit spherically symmetric family, the proof that the Gribov region is convex and bounded, and Singer's theorem that the bundle of gauge fields over the orbit space admits no continuous global section are stated here and not proved; all three belong in Appendix A. 2174

165 The Nielsen–Ninomiya theorem: the fermion doubling of the naive operator is derived above, but the no-go theorem itself — that no local, chirally symmetric, doubler-free operator exists, proved by counting the zeros of the energy–momentum relation as a topological invariant of the Brillouin torus — is stated and not proved here. It belongs in Appendix A. 2178

166 The instanton fluctuation determinant: the Bogomolny bound, the integrality of the winding number and the exponential weight are derived above, but the explicit self-dual solution of Belavin, Polyakov, Schwartz and Tyupkin and the one-loop determinant about it, which supply the measure over collective coordinates quoted in the dilute-gas remark, are not computed here. They belong in Appendix A. 2182

167 The chiral matching for the pion two-photon coupling: the phase space, the polarization sum and the colour counting of the neutral-pion width are derived above, but the step from the anomalous divergence of the axial isospin current to the effective coupling $g_{\pi\gamma\gamma}$ — the partial conservation of the axial current, the pion pole dominance and the normalization by the decay constant — is quoted rather than derived. It belongs in Appendix A, alongside the chiral material of the QCD chapter of this part. 2188

168 The effective-range expansion itself: the proof that $k \cot \delta_0$ is an even analytic function of the wave number at low energy, with the scattering length and the effective range as its first two coefficients. It is the premise of the proposition above; its derivation belongs to the low-energy section of the scattering-theory chapter, where it is not yet written. 2204

169 Chiral power counting: the derivation of the counting formula for the exponent, by assembling the powers of Q from each propagator, each loop integration and each vertex of the effective Lagrangian, and the demonstration that every bracket in the sum is non-negative for a Lagrangian consistent with chiral symmetry. It belongs in Appendix A, since it needs the full vertex enumeration and runs to several pages; the statement and its consequences are used above without it. 2208

170 The Breit–Wigner formula: the derivation of the single-level resonance cross-section used above, with its partial widths and the statistical spin factor g, from the pole structure of the S matrix. It is taken here as a premise; its derivation belongs to the resonance section of the scattering-theory chapter, where it is not yet written. 2217

171 The Rayleigh plane-wave expansion: the expansion of a plane wave in Legendre polynomials and spherical Bessel functions, with the small-argument behaviour $j_L(x) \sim x^L/(2L+1)!!$ used above. Both are premises of the forbiddenness ladder; the expansion belongs to the chapter on Fourier and integral transforms, where it is not yet derived. 2227

172 The multipole normalization: the coefficient $8\pi(L+1)/L[(2L+1)!!]^2$ of the electric multipole rate is quoted, not derived. Its derivation — the multipole expansion of the quantized radiation field, the normalization of the vector spherical harmonics and Fermi's golden rule — belongs in Appendix A; the power counting in kR and the dimensional structure are derived in the text. 2229

173 The Debye–Waller factor: the derivation of the zero-phonon probability $f = \exp(-k^2 \langle x^2 \rangle)$ for a harmonic lattice, and of the Debye-model zero-temperature mean square displacement $\langle x^2 \rangle = 3\hbar^2/(4Mk_B\Theta_D)$. Both are premises of the proposition above; they belong to the chapter on phonons and lattices, where they are not yet derived. 2231

174 The fission barrier: the deformation energy surface of the charged liquid drop beyond second order, through the saddle point to the scission configuration, giving the barrier height as a function of the fissility x and its collapse to zero as x approaches one. It belongs in Appendix A, since it requires the multipole expansion of the electrostatic energy of an arbitrary axially symmetric shape. The second-order result and the critical fissility are derived in the text. . . . 2233

175 The shell correction: the derivation that a low single-particle level density near a closed shell adds several MeV of binding relative to the smooth liquid-drop average and so restores a fission barrier — the Strutinsky shell-correction method. It belongs in Appendix A; the phenomenon above uses the statement without it. 2237

176 The $O(N)$ combinatorial factors of the one-loop shell integration: the explicit index contractions of $(\phi \cdot \phi)^2$ that produce $(N+2)/6$ in the mass renormalization and $(N+8)/6$ in the vertex renormalization, for general N. The $N = 1$ case is counted in the derivation above and both general factors reduce to it; the general count belongs in Appendix A. 2255

177 The two-loop coefficient $b_1 = 102 - \frac{38}{3} n_f$ of the SU(3) beta function: the two-loop vacuum-polarization, vertex and self-energy graphs, their renormalization, and the cancellation that leaves a scheme-independent result. It is quoted from the Particle Data Group here and it fixes the whole two-loop column of the table above, so it is a number the book uses and does not yet derive; the derivation belongs in Appendix A. 2262

178 Appelquist–Carazzone to all orders: the proof that in a mass-dependent scheme the effects of a heavy field on light-field observables are exhausted, at every order in the loop expansion and for a heavy field of any spin whose mass is an independent parameter, by a finite renormalization of the light theory's parameters plus corrections of positive power in the ratio of the probe energy to the heavy mass. What is derived in the text is the one-loop case for a heavy Dirac fermion in quantum electrodynamics, including the quadratic suppression rate; the general theorem belongs in Appendix A. . . 2266

179 Weinberg's generality claim: the demonstration that a Lagrangian containing every term allowed by the assumed symmetries and built from the light fields generates the most general S-matrix consistent with analyticity, perturbative unitarity and cluster decomposition, that is, that no further S-matrix satisfying those constraints exists which the construction fails to reach. The predictivity half of the theorem is derived in the text above; the generality half belongs in Appendix A. 2268

180 The trace anomaly, complete: the operator equation for the trace of the improved energy–momentum tensor in a non-abelian gauge theory, with the numerical coefficient of the beta-function term obtained by explicit regularization rather than by the scaling argument given in the text, together with the demonstration that the result is independent of the regulator used and that the improvement terms do not contribute. The structure — proportionality to the beta function times the operator conjugate to the coupling, and vanishing of the trace exactly at a zero of the beta function — is derived above; the coefficient is not, and the full operator derivation belongs in Appendix A. . 2274

181 The four-dimensional a-theorem: the construction of the dilaton effective action for a spontaneously broken conformal symmetry, the identification of the difference of the anomaly coefficients with the coefficient of the four-derivative dilaton self-interaction, and the positivity of that coefficient from the analyticity and unitarity of forward dilaton scattering, by an argument of the same shape as an unsubtracted dispersion relation. The statement is used here only qualitatively, to say that the flow has a direction; the proof is several pages and belongs in Appendix A. 2275

182 The Bargmann–Hall–Wightman extension theorem: that the holomorphic Wightman function on the forward tube extends single-valuedly to the extended tube and is covariant under the complex Lorentz group; together with Jost's characterization of the real points of that extended tube as the totally spacelike configurations, stated below. These are the substantial analytic imports of the chapter and belong in Appendix A, together with the Paley–Wiener–Schwartz theorem and the edge-of-the-wedge theorem they use. Everything the chapter derives from them is derived explicitly in the text, and the non-emptiness of the set of real points is derived below. 2287

183 Borchers' theorem: that two relatively local Wightman fields generate the same asymptotic states and hence the same S matrix. The argument runs through the Haag–Ruelle construction and belongs in Appendix A next to it. The consequence used in this book, that a field redefinition leaves scattering amplitudes unchanged, is stated with its mechanism in the text. 2290

184 The analytic continuation of the finite-dimensional representation matrix $S([\Lambda])$ along the path joining the identity to the total inversion in the complex Lorentz group, which is what supplies the factor $(-1)^{2s}$ used in the half-integer case. The path itself and the scalar case are derived in full in the text; the index bookkeeping for arbitrary spin follows Burgoyne and belongs in Appendix A with the Bargmann–Hall–Wightman theorem it rests on. . . . 2295

185 The Jost–Schroer theorem: a Wightman field whose two-point function is that of a free field of mass m is a free field. The argument runs through the analyticity of the truncated Wightman functions in the extended tube and the vanishing of the truncated two-point function's support off the mass shell; it belongs in Appendix A beside the other analyticity results of this chapter. 2298

186 The Haag–Ruelle convergence estimate: that the vectors of the displayed limit form a Cauchy family, with the rate supplied by cluster decomposition together with the stationary-phase decay of a wave packet built on an isolated mass shell. The construction and its mechanism are given in the text; the estimate is a page of analysis and belongs in Appendix A. 2299

187 The Doplicher–Haag–Roberts reconstruction: that localizable superselection sectors of a local net form the representation category of a compact group, with Bose or Fermi statistics forced in $[3+1]$ dimensions, and that a field algebra carrying that group exists. The statement, its inputs and its two limitations are given in the text; the proof is long and belongs in Appendix A. 2302

188 Modular theory: the Tomita–Takesaki theorem, that the polar decomposition of the Tomita operator yields a modular automorphism group preserving the algebra and a conjugation exchanging it with its commutant; and the Bisognano–Wichmann identification of that group, for a wedge in a Wightman theory, with the boosts preserving the wedge. Both are stated with their hypotheses and used only through the KMS property, from which the acceleration temperature is derived in full in the text. The two proofs are long and belong in Appendix A. 2304

189 Triviality of the four-dimensional scalar: the correlation-inequality argument bounding the connected four-point function of the lattice model above dimension four, and the tree-diagram bound that settles the marginal four-dimensional case. Both are results in classical probability rather than in field theory and are quoted here with their citations; they belong in Appendix A. What is derived in the text is the consequence, the scale at which the isolated scalar coupling diverges. 2309

190 Doubly differential muon-decay rate from the $V-A$ four-fermion matrix element, i.e. the angular factor $[(3-2x)+P_{\mu}(2x-1)\cos\theta]$ that is quoted rather than computed here; the energy spectrum alone is parametrized in the weak-interactions chapter, and the integration of the angular factor is carried out in this chapter. Belongs in the long-proofs appendix 2331

191 Atomic many-body calculation converting the measured caesium amplitude into a nuclear weak charge: the relativistic coupled-cluster evaluation of the valence electron density at the nucleus and its error budget. Belongs in the long-proofs appendix; the mixing of opposite-parity states and the Z-cubed scaling are derived in this chapter 2334

192 Coefficients of the deep-inelastic parity-violating asymmetry on an isoscalar target: the combinations $1 - \frac{20}{9}\sin^2\theta_W$ and $1 - 4\sin^2\theta_W$ follow from the electron and quark neutral-current couplings C_{1q}, C_{2q} contracted with the quark-parton structure functions of a deuteron, and are quoted rather than derived here. Belongs in the long-proofs appendix; the couplings themselves are written down in this chapter for the proton case, and the deep-inelastic kinematics live in the deep-inelastic chapter 2336

193 One-loop electroweak corrections to the weak charge of the proton: the ρ factor, the charge-radius term, and the WW, ZZ and γZ box graphs that together shift the tree value $1 - 4\sin^2\theta_W = 0.047$ to the Standard Model prediction near 0.071. They are a third of the predicted value, so the comparison with experiment does not exist without them. Belongs in the long-proofs appendix 2337

194 The short-distance box-diagram computation of the $[\Delta S = 2]$ mixing amplitude, giving the mass difference and epsilon in terms of the CKM elements and the quark masses — the step that converts the measured epsilon into a constraint on the unitarity triangle. It is an appendix-scale calculation requiring the loop functions of the box with internal charm and top, and it belongs in Appendix A rather than in an experiment chapter. 2359

195 The U-spin relation between the direct CP asymmetries of $[B^0 \to K^+\pi^-]$ and $[B_s^0 \to K^-\pi^+]$ is stated here in outline only. Owed: the explicit tree-plus-penguin decomposition, the action of the $d \leftrightarrow s$ interchange on each amplitude, and the width weighting that turns the amplitude statement into the ratio -1, together with the size of the U-spin-breaking corrections. This belongs in Appendix A. No key in the bibliography covers the U-spin literature, so the claim carries no citation. 2362

196 The three-loop, chirally suppressed Cabibbo–Kobayashi–Maskawa contribution to the neutron electric dipole moment is quoted here as an order of magnitude and is not derived. Owed: the argument that the one- and two-loop quark diagrams cancel by the Glashow–Iliopoulos–Maiani mechanism because the rephasing invariant needs all six quark masses, and the resulting estimate of order $3 \times 10^{-53}\,\mathrm{C\,m}$. This belongs in Appendix A. The bibliography carries no key for the electric-dipole calculation, so the estimate is uncited. 2368

197 The invariance theorem for a tilted and elliptical Penning trap: the general Brown–Gabrielse result, that the sum of squares of the three measured mode frequencies equals the squared free-space cyclotron frequency independently of the tilt angle and the ellipticity parameter. Belongs in Appendix A beside the ideal-trap identity proved in this chapter. 2383

198 The dispersive representation of the hadronic vacuum-polarization contribution to a lepton anomaly: analyticity of the photon self-energy, the optical theorem relating its discontinuity to the total hadronic cross section, and the derivation of the kernel $[K(s)]$ from the one-loop vertex with a dressed photon propagator. Belongs in Appendix A. 2395

199 The coefficient of the scalar-mass logarithm in the radiative correction Δr, written parametrically in this subsection, together with the constant term that accompanies it — which is not a refinement at the measured mass but comparable to the logarithm itself. Obtaining both requires the one-loop W and Z self-energies with the scalar circulating, renormalized on shell. The electroweak chapter already records the same logarithm as owed, in the paragraph following its screening-theorem discussion; the two belong together in Appendix A. Nothing in this chapter is evaluated numerically from the coefficient. 2404

200 The Standard Model production cross sections quoted in this section. Obtaining them requires the partonic amplitudes — above all the top-quark triangle that drives gluon fusion, which has no tree-level counterpart — convolved with the measured parton distributions and corrected to next-to-next-to-leading order in the strong coupling, where the corrections are large (a K factor near two for gluon fusion). The loop amplitude and the resulting cross sections are already recorded as owed in the electroweak chapter's pending derivation of the loop-induced scalar vertices; the convolution and the perturbative corrections belong with them in Appendix A. What is derived in this chapter from those inputs is everything downstream of them: the produced event counts, the channel comparison, the mass resolution and the significance. . 2413

201 The azimuthal distributions that separate 0^+ from 0^- and from the spin-2 hypotheses in $H \to ZZ^* \to 4\ell$. What is owed is the decay amplitude for each hypothesis — the scalar contraction of the two boson polarizations, the Levi-Civita contraction that a pseudoscalar is restricted to, and the tensor couplings — projected onto the two lepton pairs and integrated to give the distribution in the angle between their decay planes, exhibiting the $\cos\Phi$ term the scalar produces, its absence for the pseudoscalar, and the $\cos 2\Phi$ term of opposite sign that the pseudoscalar carries. It belongs in Appendix A beside the vector-boson vertices of the electroweak chapter, whose loop-induced counterparts are already recorded as owed there. 2427

202 The double-production cross section $\sigma(gg \to HH)$ as a function of κ_λ: the interference of the top-loop triangle carrying the trilinear vertex with the top-loop box that does not, giving the quadratic whose minimum locates the least sensitive value of κ_λ, and the normalization that puts the rate three orders of magnitude below single production. The location of the minimum and the absolute rate are stated in this subsection on the authority of the experimental compilations and are not derived here; the two loop amplitudes belong in Appendix A with the single-production ones already recorded as owed. The published constraint on κ_λ quoted below is a measurement and needs no derivation. 2432

203 The Störmer theory itself: the first integral of the motion of a charge in a magnetic dipole field, the partition of directions at each point into an allowed and a forbidden cone that follows from it, and the integration supplying the coefficient one quarter in the vertical cutoff together with the full azimuthal form used in the east–west subsection below. Belongs in Appendix A; both formulae are quoted here from the source cited beside them, and only their consequences are derived 2477

204 The Bethe–Heitler bremsstrahlung cross section, the resulting exponential energy-loss law, and the radiation length of air and of lead used throughout this chapter. Belongs in Appendix A or in the bremsstrahlung section of the radiation and scattering chapter, which currently reserves the heading . . . 2482

205 The Frank–Tamm intensity formula in SI, obtained from the retarded fields of a uniformly moving charge in a dispersive medium by Fourier transform in space and time. Belongs in Appendix A; this chapter quotes the result and uses only its numerical consequences 2494

206 The non-resonant streaming instability by which the accelerated particles amplify the magnetic field at the shock, and the saturation level that fixes $[B]$ in the maximum-energy estimate. Belongs in Appendix A; the estimate in this chapter takes the amplified field from the X-ray filament observations instead of deriving it 2503

207 The photopion energy-loss length as a proper integral of the measured $p\gamma$ cross section over the Planck spectrum and over collision angle, replacing the order-of-magnitude estimate given here. Belongs in Appendix A 2507

208 The perturbative correction to the Gross–Llewellyn Smith sum rule. The parton-model value of 3 is derived in the text above; what is still owed is the calculation of the coefficient function that multiplies it by $1 - \alpha_s/\pi - \ldots$, order by order, and the demonstration that the same coefficient function is the one appearing in the Bjorken sum rule for polarized scattering. That is what makes a measurement of this integral a competitive determination of the strong coupling, and it belongs in Appendix A alongside the other sum rules. 2542

209 The nuclear modification of parton distributions, stated as the open problem it is: no derivation from quantum chromodynamics currently reproduces the observed dependence on the momentum fraction. The candidate mechanisms — nucleon swelling, pion-cloud enhancement, multi-quark clusters, and a dependence on the local nuclear density rather than on the mass number — each account for part of the shape and none for all of it, and they are not mutually exclusive. What is owed in Appendix A is a statement of what each predicts and of which measurements discriminate among them. 2547

210 The Ellis–Jaffe sum rule, the theoretical number the EMC measurement contradicts. What is owed is the derivation of the expectation for the first moment of the polarized proton structure function from the octet axial charges: the isovector combination fixed by the neutron beta-decay ratio, the octet combination fixed by the hyperon semileptonic decays under the flavour symmetry of the baryon octet, and the assumption of an unpolarized strange sea that sets the singlet piece equal to the octet one — together with the leading perturbative correction, which lowers the parton-model value. Only that chain shows why the expected number is near 0.19 and which of its assumptions the measurement actually falsifies. It belongs in Appendix A beside the other sum rules; this book's bibliography carries no entry for the original Ellis–Jaffe paper. 2548

211 The decomposition of the proton spin into quark spin, gluon spin and orbital contributions, and the reason the split is ambiguous. What is owed is the demonstration that the gauge-invariant decomposition of the angular momentum of a gauge theory into a "spin" and an "orbital" part is not unique — different decompositions differ by terms that vanish for a free field and do not vanish here — so that a measured ΔG has meaning only relative to a stated convention and factorization scheme. It belongs in Appendix A with the operator definitions written out. 2548

212 The thermal-spike calculation of bubble nucleation: the temperature field around an ionising track from the heat equation with a line source, the fraction of the deposited energy thermalised within the critical radius before the bubble can grow, and the resulting threshold in $[\mathrm{d}E/\mathrm{d}x]$ as a function of the superheat. It belongs in Appendix A; the derivation above establishes the barrier and the mechanism but treats the energy delivery as instantaneous and perfectly local. 2579

213 Statistical Mechanics: all derivations of this chapter are pending. 2589

214 Experiment: Brownian Motion and Avogadro's Number: all derivations of this chapter are pending. 2595

215 Molecular origin of the agitation: the unbalanced momentum delivered by molecular impacts on a grain of radius a, and the reason the relative size of the resulting fluctuation grows as the grain shrinks 2596

216 The rotational displacement law for a sphere, in which the mean square angular displacement grows linearly in time with a coefficient carrying the inverse cube of the grain radius rather than its first power — the different power being what makes agreement between the two routes an independent check and not a repetition 2599

217 The velocity autocorrelation of the Ornstein–Uhlenbeck process obtained from the Langevin equation, its short-time ballistic limit, and the crossover to the diffusive law at the momentum relaxation time; and separately the hydrodynamic-memory kernel that replaces it for a sphere in a dense fluid . 2601

218 Quantum Statistics: all derivations of this chapter are pending. 2603

219 Exclusion derived rather than postulated: the symmetrisation of the state of identical particles, the vanishing of a totally antisymmetric state with two coincident single-particle labels, and the spin–statistics theorem that assigns antisymmetry to half-integer spin 2604

220 Saturation of the excited-state population of the ideal Bose gas: that the sum over excited modes of the Bose occupation is bounded above at fixed temperature and volume when the chemical potential reaches the ground-state energy, so that any further particles must enter the ground state; the resulting critical temperature, condensate fraction and heat-capacity cusp in three dimensions 2606

221 The ground-state pressure of the ideal Fermi gas, its persistence at zero temperature, the softening of the equation of state when the Fermi momentum becomes relativistic, and the resulting limiting mass for a self-gravitating degenerate sphere 2608

222 The two-particle detection amplitude for identical particles from a chaotic source, the interference of the direct and exchanged terms, and the resulting second-order correlation of two at zero separation for bosons against zero for fermions, with the source coherence length as the correlation scale 2609

223 Experiment: Bose–Einstein Condensation: all derivations of this chapter are pending. 2611

224 The polarization-gradient (Sisyphus) mechanism: the spatially modulated light shifts of the ground-state sublattice in a standing wave of varying polarization, the optical pumping that preferentially transfers an atom from a hill to the bottom of the next well, the resulting friction coefficient, and the recoil energy as the residual floor 2612

225 Interference of two overlapping expanding condensates: the fringe spacing set by the relative momentum of the two sources, and the emergence of a definite relative phase from repeated measurement on a state of fixed total atom number, in which no relative phase exists before the first detection . . 2615
226 The Bose–Hubbard description of atoms in a deep optical lattice, the tunnelling and on-site interaction parameters expressed through the lattice depth and the scattering length, and the mean-field estimate of the ratio at which the compressible superfluid gives way to the gapped, integer-filled insulator 2617
227 Quantum Optics and the Photon: all derivations of this chapter are pending. 2619
228 The shape of the pump–probe signal for two fragments separating on a repulsive potential energy surface: the classical clocking relation that maps probe detuning onto internuclear separation, the wave-packet computation of the fluorescence yield as a function of delay, and the extraction of the free-fragment appearance time from the saturating on-resonance curve . . . 2625
229 The two-state model of alkali-halide dissociation: the diabatic covalent and ionic curves and the adiabatic well their crossing produces, the Landau–Zener probability of remaining on the adiabatic branch at each pass, and the resulting periodic recurrences with a geometrically decaying envelope and a step-wise product signal 2625
230 Phase Transitions and Critical Phenomena: all derivations of this chapter are pending. 2627
231 The scattered intensity as the Fourier transform of the density–density correlation function; the Ornstein–Zernike form with its Lorentzian structure factor and correlation length; and the fluctuation–response identity that ties the forward scattering to the isothermal compressibility, so that a diverging compressibility and a diverging correlation length are one statement 2630
232 Universality as the basin of attraction of a fixed point of the coarse-graining flow: that the couplings distinguishing one microscopic system from another are irrelevant in the technical sense, so that the singular behaviour is governed by the fixed point alone; and the accompanying calculation of the exponents, developed in the renormalization-group chapter 2633
233 The superfluid transition as the three-dimensional XY universality class, and the calculation of its specific-heat exponent to a precision comparable with the measurement, by renormalization-group and high-order perturbative methods; together with the hyperscaling relation that ties the negative sign of the exponent to the correlation-length exponent in three dimensions . . . 2633
234 The orientational order parameter as a traceless symmetric second-rank tensor, forced by the equivalence of the director and its negative; the cubic invariant that this symmetry permits and that therefore makes the isotropic–nematic transition weakly first order, with a small latent heat and a jump in the order parameter; and the purely entropic hard-rod mechanism that produces the same transition with no attraction at all 2634
235 Nonequilibrium Thermodynamics and Transport: all derivations of this chapter are pending. 2637
236 Phonons and Lattice Dynamics: all derivations of this chapter are pending. 2643
237 The crystallographic restriction, proved from the integrality of the matrix of a lattice rotation in the lattice basis and hence of its trace, which admits only two-, three-, four- and six-fold axes; and the demonstration that a non-periodic but long-range-ordered point set can nevertheless have a Fourier transform consisting of sharp discrete peaks, so that sharp diffraction does not entail a periodic lattice 2644

238 The Grüneisen relation: that a phonon frequency depends on volume, quantified by the logarithmic derivative of frequency with respect to volume; the resulting thermal pressure of the phonon gas; and the equality of that pressure gradient with the elastic restoring stress, which yields the expansion coefficient as the product of the compressibility, the heat capacity and the mean logarithmic derivative, together with the sign reversal produced by transverse modes that stiffen under expansion 2649

239 Why a harmonic crystal has infinite thermal conductivity, and why anharmonicity alone is not enough: that three-phonon processes conserving the total crystal momentum cannot degrade a heat current, so that a finite thermal resistance requires umklapp processes in which the summed wavevector leaves the Brillouin zone and is returned by a reciprocal-lattice vector; the exponential freezing-out of those processes at low temperature; and the kinetic-theory expression for the conductivity in terms of heat capacity, group velocity and mean free path, whose crossover to boundary-limited and isotope-limited scattering produces the observed maximum 2650

240 Electrons in Solids: Band Theory: all derivations of this chapter are pending. 2651

241 The dynamics of a nearly filled band described through its vacancies: that the total current of a filled band vanishes, so that the current of an almost-filled band equals minus the current of the missing states; that a wave packet near a band maximum has negative band curvature and therefore responds as a particle of positive charge and positive mass; and the resulting positive Hall coefficient, with the multi-band and open-orbit corrections that govern its magnitude 2653

242 Bloch's theorem for a periodic potential and the counting that follows from it: that the allowed wavevectors in one Brillouin zone number the primitive cells of the crystal, so that each band holds exactly two electrons per cell; that a crystal whose bands are either completely full or completely empty has no states available at infinitesimal excitation energy and therefore carries no current at zero temperature; and the consequent identification of a semiconductor as an insulator whose gap is comparable with the thermal energy 2655

243 Semiclassical quantization of a closed cyclotron orbit, giving orbits of quantized area in reciprocal space evenly spaced in the magnetic field; the resulting oscillation of the free energy as successive quantized tubes cross the Fermi surface, periodic in the inverse field with a frequency proportional to the extremal cross-sectional area; and the Lifshitz–Kosevich amplitude with its temperature and scattering damping factors 2657

244 That a Bloch state is a stationary state of the perfect periodic potential and therefore propagates without attenuation, so that resistance requires a departure from periodicity; the Boltzmann equation in the relaxation-time approximation and the additivity of scattering rates from independent channels that yields Matthiessen's rule, together with the conditions under which the additivity fails; and the electron–phonon calculation giving the linear high-temperature and the steep low-temperature resistivity 2658

245 The Hubbard Hamiltonian, in which a hopping amplitude competes with an on-site repulsion, and the localization of one electron per site when the repulsion exceeds the bandwidth — a transition to an insulating state that occurs at fixed, partial band filling and therefore has no description in band theory at all; noting that no controlled solution of this model in three dimensions exists, so the derivation available is a limiting-case argument and not a theory 2659

246 Semiconductors: all derivations of this chapter are pending. 2661

247 The hydrogenic impurity level: a donor treated as a charge screened by the dielectric constant of the host and binding an electron of the band effective mass, giving a binding energy reduced from the hydrogenic value by the effective-mass ratio and by the square of the relative permittivity, hence tens of meV; and the charge-neutrality condition combined with the law of mass action, which yields the freeze-out, saturation and intrinsic regimes as limits of one expression 2664

248 The p–n junction in equilibrium: the diffusion of carriers across the metallurgical boundary, the depletion region and built-in potential that arrest it, and the equality of drift and diffusion currents that fixes that potential through the law of mass action; then, under bias, the Boltzmann factor by which the barrier lowering multiplies the minority-carrier density at the depletion edge, the diffusion of those excess carriers into the neutral regions, and the resulting saturation current — together with the recombination, high-injection and series-resistance terms by which real diodes depart from the ideal law . . . 2665

249 Transistor action from the junction physics already in hand: injection of minority carriers across a forward-biased emitter junction, their diffusion across a base thin compared with the diffusion length so that few recombine, and their collection by the reverse-biased collector junction; the resulting current gain as the fraction surviving the base, and the power gain as the product of that fraction with the ratio of the collector to the emitter impedance — the transfer of current from a low-resistance to a high-resistance circuit that gives the device its name 2666

250 Radiative recombination of injected minority carriers across a direct gap: the emission spectrum as the joint density of states weighted by the carrier occupations, giving a peak just above the gap with a width of order the thermal energy; the selection rule that suppresses the process in an indirect-gap material, where a phonon must participate, and the competing non-radiative Auger and trap-assisted channels that fix the internal quantum efficiency 2668

251 Experiment: The Quantum Hall Effect: all derivations of this chapter are pending. 2669

252 Laughlin's gauge argument: that adiabatically threading one flux quantum through a Hall cylinder is a gauge transformation and must map the spectrum to itself, so that its only effect is to transfer an integer number of electrons from one edge to the other; the resulting Hall conductance as that integer times the square of the elementary charge over Planck's constant, with no reference to the Hamiltonian; and the requirement of a mobility gap at the Fermi level that makes the process adiabatic — together with the topological restatement in which the integer is the Chern number of the occupied states, which cannot change under any smooth deformation of the sample 2672

253 Laughlin's correlated wavefunction for filling one over an odd integer, an incompressible liquid rather than a filled single-particle band; the plasma analogy that fixes its density and its excitation gap; the antisymmetry requirement that forces the exponent to be odd; and the hierarchy or composite-fermion construction that accounts for the remaining observed fractions 2673

254 The quasihole of the correlated ground state and the determination of its charge from the deficit of density produced by adiabatically threading one flux quantum, together with the plasma analogy that makes the deficit exact; and the Schottky formula for the noise of a Poisson stream of independent carriers, whose proportionality to charge times current is what converts a noise measurement into a charge measurement 2674

255 Landau quantization of a linear rather than quadratic dispersion, which gives level energies proportional to the square root of the product of level index and field and places one level exactly at zero energy; the shared occupation of that zero level by electrons and holes, which shifts the filling sequence by one half; the fourfold spin and valley degeneracy of the observed steps; and the Berry phase of the cyclotron orbit that appears as the offset in the Shubnikov–de Haas phase 2675

256 That a quantized Hall conductance requires only a gapped band structure with broken time-reversal symmetry and a non-zero topological invariant of the occupied bands, not a uniform magnetic field; the lattice model exhibiting this with zero net flux through the unit cell; and the mechanism by which magnetic dopants in a strong-spin–orbit thin film open the required gap in the surface states 2675

257 Magnetism in Matter: all derivations of this chapter are pending. 2677

258 Diamagnetism as the induced response of the bound electrons: the Larmor precession an applied field imposes on an atomic charge distribution, the induced moment opposing the field by Lenz's rule, and the resulting susceptibility proportional to the mean square radius of the electronic charge distribution and to the number of electrons per unit volume, with no temperature dependence because no level populations change 2678

259 The two-sublattice molecular-field theory, giving the Néel temperature, the susceptibility maximum at it and the negative Curie–Weiss intercept above it; and the magnetic neutron cross section, in which the neutron's own moment couples to the ordered electronic moments so that the scattering amplitude alternates between sublattices, producing reflections at the reciprocal-lattice vectors of the doubled magnetic cell that are absent from the nuclear pattern 2683

260 Superconductivity and Superfluidity: all derivations of this chapter are pending. 2687

261 Why the normal state cannot produce this: the temperature-independent residual resistivity that impurity scattering guarantees in any metal, so that a discontinuous drop to zero requires a new state rather than a limit of the old one; and the conversion of the persistent-current decay bound into a resistivity bound through the inductance of the loop 2688

262 Why an electron–phonon-mediated pairing energy carries the phonon frequency as its prefactor, so that the transition temperature inherits the ionic mass dependence; and the corrections — from the screened Coulomb repulsion between the electrons, and from strong coupling — that reduce the exponent below one half without removing the effect 2690

263 The gap as the energy to break one pair, obtained from the self-consistent gap equation of the pairing theory; the quasiparticle dispersion, whose minimum is the gap and whose density of states diverges at the gap edge, giving the observed tunnelling threshold and infrared edge; and the exponential activation of the heat capacity, together with the derivation of the measured ratio of twice the gap to the transition temperature as a parameter-free consequence of weak coupling 2691

264 Two weakly coupled condensates described by their amplitudes and phases: the coupled equations of motion for the pair, in which the current across the barrier is proportional to the sine of the phase difference and the phase difference evolves at a rate proportional to the electrochemical potential difference; the resulting zero-voltage supercurrent and the alternating current at the Josephson frequency; and the spatial variation of the phase difference across a junction in a field, whose integral gives the diffraction-like dependence of the critical current on the enclosed flux 2693

265 The two-fluid description in which the liquid behaves as an interpenetrating mixture of a normal component carrying all the entropy and all the viscosity and a superfluid component carrying neither, and the reconciliation of the capillary and rotating-body measurements that it provides; the thermomechanical effect as the statement that a temperature difference and a pressure difference are equivalent driving forces for the superfluid component; and Landau's criterion, in which an excitation spectrum with no low-lying free-particle branch forbids dissipation below a critical velocity — which is why an ideal Bose gas would condense without being a superfluid 2694

266 Experiment: Superconductivity: all derivations and data tables of this chapter are pending. 2697

267 The inference chain from the raw run: how a four-terminal measurement removes lead and contact resistance so that the reading belongs to the specimen; how the galvanometer sensitivity and the measuring current together convert into the quoted upper limit on the resistance; and the control observations that exclude the obvious alternative, a short circuit developing on cooling, by their dependence on measuring current and on applied magnetic field 2700

268 The inference from the measured isotope exponent: that a lattice frequency scales as the inverse square root of the ionic mass with the force constants, which are electronic, held fixed, so that an exponent of one half identifies the phonon energy as the prefactor of the pairing scale; and the corrections from the screened Coulomb repulsion, which enter the exponent and depress it below one half without altering the identification 2701

269 The coherence factors: that the matrix element for a process between quasiparticle states is a sum of two amplitudes whose relative sign depends on the behaviour of the perturbation under time reversal, so that the two combinations cancel and reinforce respectively; and the resulting opposite temperature dependences of ultrasonic attenuation and nuclear spin relaxation just below the transition, against the divergent density of states at the gap edge that both share 2702

270 What can and cannot be derived here: the demonstration that flux expulsion on field cooling requires a bulk, thermodynamic superconducting phase and excludes a percolating filamentary path; and the reading of the half-integer flux quantum trapped at a tricrystal ring as a sign change of the order parameter between directions. The mechanism itself is not derived anywhere in this treatise and is recorded as an open problem 2704

271 The one-phonon scattering cross section itself: the factor that weights each peak by the square of the projection of the polarization vector on the momentum transfer, by the Bose population of the mode, and by the Debye–Waller factor, so that the measured intensities constrain the eigenvectors of the dynamical matrix and not only its eigenvalues 2709

272 The mode sum for the mean square displacement, giving $2W$ in terms of the measured dispersion relation, its evaluation in the Debye model with the zero-point term separated from the thermal one, and the resulting temperature dependence of the attenuation at fixed momentum transfer 2714

273 The probability distribution of the triplet phase sum: the random-atom model in which the atomic positions are independent and uniform over the cell, the central-limit argument that makes the joint distribution of three normalized structure factors Gaussian, and the conditional distribution of the phase sum that follows from it, giving the hyperbolic tangent in the centrosymmetric case and the von Mises form in general 2720

274 The contrast transfer function of a defocused electron microscope: the weak-phase-object approximation, the phase shift introduced by defocus and by spherical aberration, and the resulting oscillating transfer function whose zeros must be filled in by combining images taken at different defocus . . . 2722

275 The Dark Sector: all derivations of this chapter are pending. 2729

276 Quantum Gravity: The Honest Status: all derivations of this chapter are pending. 2735

277 Experiment: Dark-Matter Searches: the standard halo model and the nuclear form factor that together convert a measured recoil spectrum into a bound on a cross-section, and the statistical construction of a 90 per cent confidence exclusion contour in the presence of an estimated background, are pending. The kinematic and rate arguments of this chapter are derived inline. 2741

278 The form factor and the recoil spectrum: derivation of the Helm form factor as the Fourier transform of a uniform nuclear density convolved with a surface term, its first zero in relation to the recoil energies of interest for xenon, and the resulting distortion of the differential rate away from the pure exponential that the maximum-recoil relation alone would give; together with the detector response function that converts a recoil energy into an expected number of scintillation photons and ionization electrons, which is what actually sets the threshold 2744

279 The quantitative incompatibility: the calculation that converts a modulation amplitude measured in sodium iodide into an allowed region in the mass–cross-section plane, including the quenching factors for sodium and for iodine recoils, which are measured quantities with their own uncertainties and which set the recoil-energy scale of the claimed region; and the comparison of that region against the exclusion curves of the xenon experiments, together with an enumeration of the non-standard halo and interaction assumptions — streams, inelastic scattering, isospin-violating couplings — that have been proposed to reconcile them and the reasons each is now disfavoured 2748

280 The thermal relic cross-section: solution of the Boltzmann equation for a species annihilating in an expanding universe, freeze-out when the annihilation rate falls below the expansion rate, and the resulting inverse proportionality of the relic abundance to the annihilation cross-section, which is what fixes the benchmark value quoted above from the abundance measured in the dark sector chapter; and the reason the required value falls near a weak-scale cross-section, which is the coincidence that motivated a generation of searches 2753

281 Digits and irrationality of π: an explicit error bound pinning $\pi = 3.141592653\ldots$, and Niven's proof that π is irrational. The latter needs $\sin\pi = 0$ and $\cos\pi = -1$, hence the addition formulas, which the real-analysis chapter does not yet establish — it defines π as twice the first positive zero of the cosine series and proves it is the circle constant, but computes no decimal of it. 3175

Part I

Foundations

Chapter 1

Epistemology and the Scientific Method

1.1 What physics is

Physics is the attempt to describe, with the least possible number of independent assumptions, everything that can be observed in Nature, and to predict what will be observed next. Three commitments separate it from every neighbouring discipline. First, its statements are *quantitative*: they concern numbers obtained by measurement, and are meaningless without a statement of how those numbers were obtained and how much they may be trusted (Chapter 2). Second, its statements are *testable*: a proposition belongs to physics only if some conceivable observation could refute it [1173]. Third, its statements are *connected*: isolated regularities are catalogued, but the goal is always the smallest set of principles from which the catalogue follows deductively.

This treatise takes the second commitment as a hard rule of scope. Theories of great mathematical beauty whose predictions have not been observed — string theory, supersymmetry, supergravity, grand unification — are not treated here, however influential they may be as research programs; they are named only once, in Chapter 133, precisely in order to say honestly that the evidence for them is currently absent. What *is* treated is everything for which Nature has already voted: from the parabola of a thrown stone (Chapter 20) to the strain of 10^{-21} that two colliding black holes imprinted on an interferometer (Chapter 47).

1.2 Physics, mathematics, and philosophy

Mathematics supplies the language in which physical principles are stated and their consequences extracted. The relationship is asymmetric: mathematics owes physics nothing, while physics without mathematics cannot even formulate a prediction. Why abstract structures invented for their own sake — complex numbers, Riemannian geometry, group representations — repeatedly turn out to be exactly what Nature uses is a genuine puzzle, famously called the *unreasonable effectiveness of mathematics* [1195]. This treatise responds to that effectiveness practically: Part II develops the mathematics in the generality it deserves — often in arbitrary dimension $D = p + q$ and signature (p, q) — while the physical parts instantiate it in the observed $3 + 1$ dimensions.

Philosophy, by contrast, supplies the questions physics cannot answer from within: what counts as an explanation, whether unobservable entities exist, how theory choice works when data underdetermine it [1233]. We borrow from it sparingly and explicitly — and from one tradition in particular, for reasons set out in Section 1.3. Where an interpretive question has no observable consequence — as with the interpretations of quantum mechanics — this

treatise states the competing readings and the experiments that bound them (Chapter 85), and goes no further.

1.3 Why this treatise follows analytic philosophy

The preceding section borrowed from philosophy; this one says from which philosophy, and why. The borrowing is not incidental. The rule that governs what may appear in this book — only physics for which Nature has already voted — is not itself a result of physics, since no experiment establishes it. It is a philosophical position, and a treatise that acts on one owes its reader the argument rather than the posture.

1.3.1 The choice cannot be avoided

Three questions must be settled before the first equation is written. What counts as evidence? What counts as having explained something? Which apparently physical questions are questions at all? Physics does not answer them from within: an experiment can decide between two theories, but no experiment can decide whether a proposition is the kind of thing an experiment could decide.

Every treatise answers all three regardless — in its selection of material, in what it is willing to call a derivation, and in the questions it declines to discuss. The only real choice is whether the answers are stated or merely enacted. Enacted, they are usually inconsistent: the same textbook will define temperature by the procedure that measures it on one page and speak of the wavefunction as a thing on the next, without noticing that the two paragraphs answer to incompatible theories of meaning. This treatise states its answers instead, and takes them from the analytic tradition — the line running from Frege's logic [207] through the Vienna Circle [624] to the modern philosophy of science.

1.3.2 What the commitment amounts to

"Analytic" here names a method, not a creed, and the method reduces to five working rules.

1. **Logical form, not grammatical form.** A well-formed sentence need not pose a well-formed question. "What is the temperature of this single molecule?" is impeccable English and has no answer, because temperature is a property of an ensemble in equilibrium and of nothing else (Chapter 55). Before a question is answered it is worth asking what would count as an answer.

2. **To explain is to derive.** A phenomenon is explained when it is exhibited as a consequence of premises that are stated in full [915]. That is editorial rule 2 of this treatise in one sentence, and it is why an unproved claim here is marked as pending rather than covered by a plausible-sounding paragraph: the premise list is either complete or it is visibly not.

3. **Content through procedure.** A quantity that no procedure could measure carries no empirical content [565]. Hence the fixed structure of every experiment box — apparatus, procedure, data with uncertainties — and hence the axiomatic status of the SI (Chapter 2). The strong form of this rule, which *identifies* a quantity with one procedure, is too strong and was abandoned: thermodynamic temperature is fixed by the agreement of gas thermometry, radiation pyrometry and Johnson-noise thermometry, not by any one of them (Chapter 54). What survives is the demand that at least one procedure exist and that all of them agree.

4. **Convention separated from content.** Parts of any theory are decisions rather than discoveries [309], and the stipulations that tie a formalism to measurements — Reichenbach's coordinative definitions [619] — must be declared as such. Which coordinates, which gauge, which metric signature, which unit system: none of these is measured. The empirical content is what stays invariant when they are varied. This is why the SI is imposed by axiom and labelled a convention, and why Part II is written for signature (p, q) instead of quietly building $3 + 1$ into the mathematics.

5. **Triage before answer.** Every question is first sorted: empirical, to be settled by measurement; conventional, to be settled by stipulation and then recorded; or defective, to be repaired or dropped. Delivering that verdict is a result, not an evasion — "the aether's rest frame" was disposed of in the third category (Chapter 37).

1.3.3 Why this tradition and not another

Three reasons of principle, and a fourth of practice.

It was built out of this material. Analytic philosophy is the only school formed in direct response to the crises of physics and mathematics between 1880 and 1930. Mach's critique of absolute space [225] is part of the prehistory of general relativity; Einstein's operational dissection of simultaneity is conceptual analysis performed by a physicist (Chapter 38), and he stated the methodology explicitly [473]: "insofar as the propositions of mathematics refer to reality, they are not certain; and insofar as they are certain, they do not refer to reality." Frege's new logic, and the paradox that struck it, opened the foundational crisis (Section 3.6); the Vienna Circle assembled around relativity and the new quantum theory [624]; Reichenbach's philosophy of space and time [619] is also a technical treatment of the conventionality of simultaneity. The working vocabulary of a careful physicist — operational definition, auxiliary hypothesis, underdetermination, model error as distinct from statistical error — is that tradition's vocabulary. Adopting it costs no translation.

Its results occasionally become physics. The traffic runs both ways, and this is the strongest single argument for the choice. Einstein, Podolsky and Rosen raised a philosophical objection to the completeness of quantum mechanics [753]; Bell converted the objection into an inequality between measurable correlations [1271] (Chapter 84); experiment then decided it, against local hidden variables (Chapter 89). A tradition that can promote a metaphysical dispute into a laboratory result has demonstrated the only credential this treatise recognizes.

It audits itself to destruction. Its central doctrine was proposed, refuted and abandoned in public. The verification principle fails its own test; Duhem showed that no hypothesis meets experience alone [325] and Quine generalized the holism until the analytic–synthetic distinction went with it [987]; Popper replaced verifiability by falsifiability [1173]; Kuhn supplied the historical objection [1233]; Lakatos replaced the single falsifiable theory by the progressive or degenerating research programme [1428]; and from inside mathematics Gödel ended the dream of complete formalization [681]. The full sequence is set out in Section 3.7, with what remains of it in Section 3.7.3. What matters here is the pattern: a book whose own rule is that no assertion stands without its derivation cannot consistently adopt a philosophy that exempts itself from that treatment.

It yields the decisions this book must make. Beyond provenance and self-audit, the tradition supplies standards for choices that recur on every page: whether a subject is admissible (testability); whether an explanation is finished (derivation from stated premises);

how a superseded theory relates to its successor — as a limit case with a stated domain of validity rather than as an error, so that Newtonian mechanics survives inside relativity and quantum mechanics [1214] (Chapters 40 and 77); and how a theory relates to its mathematics, namely as a class of mathematical models together with a rule mapping some of their features onto measurements [1193]. That last one is literally the architecture of this book: Part II builds the structures, and the physical parts supply the interpretation and the measured numbers.

1.3.4 What the alternatives would cost

Four alternatives deserve naming. None is dismissed as foolish; each is declined for a stated reason.

1. **A priori metaphysics** — determining the structure of Nature by reason in advance of measurement. Its track record in physics is the argument against it: Aristotle's natural places, Descartes' vortices, Kant's Euclidean space and absolute simultaneity as preconditions of any possible experience, all refuted by measurement (Part IV). Where a priori reasoning has succeeded in physics it did so as symmetry argument (Chapter 21) — and symmetry claims are empirical, as parity discovered when an experiment found it violated (Chapter 110). Accepting this alternative would license precisely what editorial rule 1 excludes.

2. **Phenomenological and hermeneutic traditions.** Their questions are real ones: how a scientific object is constituted in experience, how a practice is situated in its history. But they yield no procedure for deciding whether a given claim belongs in a physics book, and no account of when a derivation is complete. They are not false; they are the wrong instrument for this particular job.

3. **History- and sociology-first accounts** [1233, 1544]. As description these are substantially right, and the five-station cycle of Section 1.4 is an idealization that says so. Feyerabend's demonstration that no proposed rule of method has gone unviolated in some successful episode is accepted here without reserve. The reply is that this treatise legislates presentation, not discovery: how a result was found may be as anarchic as the record shows, but what is asserted in a book still has to come with its evidence, its premises and an admission of what remains unproven.

4. **Instrumentalism and constructive empiricism** [1652]. This is not a rival method but an answer, within the same tradition, to a question the treatise deliberately leaves open: whether the unobservable structures of a successful theory are real or are devices for organizing observations. Where that dispute acquires observable consequences it is treated as experimental (Chapter 89); where it does not, the competing readings are reported and the matter is left there (Chapter 85).

1.3.5 The price of the choice

The tradition's failures are part of the reason to trust it, and they are inherited along with the method.

1. **No sharp line of demarcation.** Holism guarantees that any prediction can be rescued by adjusting an auxiliary hypothesis, so testability is a policy applied case by case with the reasons given, not a decision procedure. The dark sector is treated as physics because it has independent quantitative signatures in several channels (Chapter 134); programmes with no such signature are named once, to record that absence honestly (Chapters 133 and 135).

2. **No theory-free observation.** Every number in this book was produced by an apparatus modelled with theory, so the observational base is not neutral ground (Chapter 2). The response is disclosure — apparatus, procedure, calibration and uncertainty budget in every experiment box — not the pretence of raw data.

3. **No complete formalization.** Axiomatization is a tool with a proven ceiling [681]. It is used where it earns its keep (Chapters 77 and 109) and not imposed elsewhere for appearances.

4. **No theory of discovery.** Logical reconstruction is not history: derivations here are ordered for validity, not chronology, and Planck's radiation law was a successful fit years before anyone could say why it worked (Chapter 68). Where the order of discovery carries the lesson, it is recorded in the experiment chapters rather than smuggled into the proofs.

1.3.6 How the choice shows on the page

The commitment is not decorative; it is the reason this book has the shape it has. The numbers below are those of the five rules in Section 1.3.2.

- *Phenomenon* boxes state what was observed, with its uncertainty — rule 3.
- *Experiment* boxes carry a fixed field structure, so that the procedure giving each term its content is on the page and not in the reader's imagination — rule 3, made procedural.
- Definitions and axioms are declared before use, and conventions are labelled conventions rather than presented as findings — rule 4.
- Every phenomenon carries its derivation, or a visible *pending derivation* marker collected in a front-matter list — rule 2, and the bookkeeping that holism demands.
- Sources are cited so that a reader can retrieve them and check the claim independently; where a source cannot be retrieved, that fact is recorded rather than hidden — rule 2 applied to the evidence base.
- Where no observation discriminates between competing readings, the treatise stops (Chapter 85) — rule 5.

Remark 1.1. The commitment is methodological, not partisan. No physical result in this book depends on it: the Lorentz transformations, the mass of the Higgs boson and the CMB power spectrum are what they are under any philosophy. What depends on it is everything editorial — what was included, what was allowed to count as a derivation, what was left out and for which reason — and that is exactly the part a reader cannot check against Nature, and must therefore be told.

1.4 The scientific method as used in this treatise

The working cycle of physics, as this treatise practices it, has five stations.

1. **Phenomenon.** Something is observed, reproducibly, with stated apparatus and stated uncertainty. In the text, such facts are boxed in the *Phenomenon* environment.

2. **Model.** A mathematical structure is proposed whose solutions mimic the phenomenon. Its assumptions are listed explicitly.

3. **Derivation.** The phenomenon is *proven* to follow from the model. This treatise's first editorial rule is that this step is never skipped: every Phenomenon carries its derivation inline, in Appendix A, or — when not yet written — as a visibly marked pending derivation.

4. **Prediction.** The model is pushed beyond the data that motivated it, into statements that could fail.

5. **Experiment.** The prediction is confronted with Nature, in an experiment reported with apparatus, procedure, data, and uncertainty — the fixed structure of every *Experiment* box in this book. A failed confrontation sends the cycle back to station 2 with the failure as new input.

Galileo's inclined planes [4] already exhibit the full cycle, and Newton's *Principia* [14] is its first complete monument: observed regularities (Kepler's laws), a model (three laws of motion and universal gravitation), derivations, and predictions reaching to the tides. The cycle has not changed since; only the precision has, from Galileo's pulse beats to the 10^{-18} m displacement sensitivity of a gravitational-wave interferometer.

Remark 1.2. The method never certifies a theory as true; it certifies survival under attempted refutation [1173]. Newtonian gravity survived for two centuries before the perihelion of Mercury and the bending of starlight retired it to the status of a limit case (Chapter 42). Every theory in this treatise, including the best-tested ones, is to be read with that clause attached.

1.5 Primitive notions

Every deductive structure must start from undefined terms. Physics is no exception; the following notions are taken as primitives, sharpened but never fully defined by the theories built on them.

Definition 1.3 (System)**.** A *system* is the part of Nature under study, conceptually separated from the rest (its *environment*) by a stated boundary. The separation is a choice of the physicist, and a fallible one: much of thermodynamics and of quantum measurement theory is the study of what that boundary costs.

Definition 1.4 (State)**.** The *state* of a system is the minimal set of data that, together with the dynamical laws, determines the outcome statistics of every measurement that can be performed on it. What data suffice is itself a physical discovery: a point in phase space classically (Chapter 22), a ray in Hilbert space quantum mechanically (Chapter 77).

Definition 1.5 (Space and time)**.** *Space* is the catalogue of possible coincidences of systems; *time* the ordering of events along a system's history. Both enter pre-relativistic physics as an absolute stage (Chapter 18) and are demoted to dynamical, observer-dependent structure by relativity (Parts IV and V).

Definition 1.6 (Particle and field)**.** A *particle* is a system idealized as carrying its full state at a single spatial point; a *field* is a system whose state assigns data to every point of space. The deepest lesson of twentieth-century physics is that the second notion is the fundamental one, the first emerging as its quantized excitations (Part XI).

Definition 1.7 (Law and theory)**.** A *law* is a compact statement of an observed regularity, valid in a stated domain. A *theory* is a deductive structure from which many laws follow, together with the specification of its own domain of validity. Neither term implies certainty; both are graded by the evidence recorded in their experiment boxes.

1.6 How this treatise is organized

Three threads are interwoven throughout. *Mathematical physics* (Part II) develops the language once, in general dimension, so that no later part need pause for it. *Theoretical physics* (Parts III to XIII) builds the models and carries out the derivations. *Experimental physics* lives inside the theory parts as dedicated experiment chapters placed immediately after the theory each experiment tests, and is indexed globally by the List of Experiments in the front matter. Two conventions bind all three threads: SI units are used axiomatically everywhere (Chapter 2), and no observed phenomenon is left without its derivation or an explicit admission that the derivation is pending.

Remark 1.8 (The logical structure is itself a checked object). The chains this organization promises — axiom or postulate to definition, theorem and equation, to observed phenomenon, to the experiment that tests it, to the source that reports it — are not left to the prose. They are maintained alongside the text as a knowledge base in two presentations at once: a relational one (entity tables of the objects just named, with primary and foreign keys, plus junction tables for the many-to-many relations) and a property multigraph (every object a node; every non-empty foreign-key cell and every junction row an edge of a declared type). The two are one dataset presented twice: objects and nodes are in bijection, so are relationship instances and edges, either presentation rebuilds the other, and every consistency rule is computed in both languages — as relational algebra and as graph traversal — with the verdicts required to agree. That correspondence is developed in a manuscript in preparation [2433], an unpublished self-citation the reader cannot yet retrieve; nothing here depends on it, because the equivalence just stated is verified mechanically over this book's own data on every build, and the verification is published with the book's validation report.

Chapter 2

Measurement, SI Units, and the Theory of Errors

2.1 Measurement

A *measurement* is a physical interaction between a system and an apparatus whose outcome is a number (or a set of numbers) together with a unit and a statement of uncertainty. All three components are mandatory: a number without a unit does not identify a quantity, and a number without an uncertainty does not identify how strongly it constrains a theory [1945, 2129]. The quantity a measurement intends to determine is called the *measurand*.

Definition 2.1 (Physical quantity)**.** A *physical quantity* is a property of a system that can be compared quantitatively with a reference property of the same kind, the *unit*. The result of the comparison is the *numerical value*, so that every quantity Q decomposes as $Q = \{Q\}\,[Q]$ with $\{Q\}$ a real number and $[Q]$ the unit.

2.2 The International System of Units

Axiom 2.2 (SI)**.** All physical quantities in this treatise are expressed in the International System of Units (SI), in every branch of physics.

Since the 2019 revision, the SI is defined not by artefacts but by fixing the exact numerical values of seven defining constants — the caesium hyperfine frequency $\Delta\nu_{\mathrm{Cs}}$, the speed of light c, the Planck constant h, the elementary charge e, the Boltzmann constant k, the Avogadro constant N_A, and the luminous efficacy K_{cd} — from which the seven base units (second, metre, kilogram, ampere, kelvin, mole, candela) are derived [2333]. The complete tables, with the CODATA 2018 values of the remaining measured constants, are collected in Appendix B [2392].

Two practical consequences of Axiom 2.2 run through the whole treatise. First, equations keep their constants: c, G, $\hbar$ and k are never set to one, so every formula can be evaluated against an experiment without unit bookkeeping. Where a natural-unit shortcut is genuinely illuminating it appears inside a flagged remark with the conversion stated. Second, every reported number carries its unit through the typography itself, e.g. $g_0 = 9.806\,65\,\mathrm{m/s^2}$ (exact, by convention [2333]).

2.2.1 Dimensional analysis

Every physical quantity has a *dimension*, a formal product of powers of the base dimensions $\mathsf{T}, \mathsf{L}, \mathsf{M}, \mathsf{I}, \Theta, \mathsf{N}, \mathsf{J}$. An equation between quantities is admissible only if both sides carry the

same dimension; this trivial requirement is unreasonably powerful, because it constrains the possible form of an answer before any dynamics is solved.

Theorem 2.3 (Buckingham π theorem)**.** *Let a physical relation connect n quantities $q_1, \dots, q_n$ whose dimensions involve r independent base dimensions. Then the relation can be rewritten as a relation among $n - r$ independent dimensionless combinations $\pi_1, \dots, \pi_{n-r}$ of the q_i.*

Rests on Axiom 2.2.

Derivation. *Derives Theorem 2.3.*

Write the dimension of each q_i as an integer (or rational) exponent vector in the r-dimensional space of base dimensions. The map sending a monomial $q_1^{a_1} \cdots q_n^{a_n}$ to its dimension is linear in the exponents a_i, with matrix M of shape $r \times n$ and rank $r' \le r$. Dimensionless monomials are precisely the kernel of M, which by the rank–nullity theorem (Chapter 5) has dimension $n - r'$. Choosing a basis $\pi_1, \dots, \pi_{n-r'}$ of the kernel, any dimensionally consistent relation among the q_i, being invariant under independent rescalings of the base units, can depend on the q_i only through kernel monomials, hence is a relation among the π_j. With the customary assumption that the r base dimensions actually occur independently, $r' = r$. □

Example 2.4. For a pendulum of length ℓ, mass m, in gravity g, with period T: the four quantities involve $\mathsf{T}, \mathsf{L}, \mathsf{M}$, so there is $4 - 3 = 1$ dimensionless group, $\pi = T\sqrt{g/\ell}$. Dimensional analysis alone forces $T = C\sqrt{\ell/g}$ with C a pure number — mass cannot appear. Dynamics is needed only for $C = 2\pi$ (Chapter 28).

2.3 The theory of errors

No measurement returns the measurand exactly. The deviations are of two kinds, treated by two different disciplines.

Definition 2.5 (Random and systematic errors)**.** A *random error* varies unpredictably between repetitions of the measurement under nominally identical conditions; it is treated statistically, and shrinks under repetition. A *systematic error* biases every repetition the same way (a miscalibrated scale, a clock that runs slow); it is immune to repetition and must be hunted by varying the apparatus and the method. The corresponding quantitative statements are called *type A* and *type B* uncertainty evaluations [2129].

2.3.1 Repeated measurements

Let $x_1, \dots, x_N$ be independent repetitions of a measurement of a quantity whose true value is μ, each modelled as a random variable with mean μ and variance σ^2 (the probabilistic machinery is developed in Chapter 11).

Definition 2.6 (Sample mean and variance)**.**

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i\,, \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2\,. \tag{2.1}$$

Theorem 2.7 (The mean improves as $1/\sqrt{N}$)**.** *$\bar{x}$ is an unbiased estimator of μ, and its standard deviation is $\sigma/\sqrt{N}$.*

Rests on Definition 2.6.

Derivation. *Derives Theorem 2.7.*

Linearity of the expectation gives $\langle \bar{x} \rangle = \frac{1}{N} \sum_i \langle x_i \rangle = \mu$. For the variance, since the x_i are independent, variances add:

$$\operatorname{var} \bar{x} = \frac{1}{N^2} \sum_{i=1}^{N} \operatorname{var} x_i = \frac{\sigma^2}{N} , \tag{2.2}$$

whence the standard deviation of the mean is $\sigma/\sqrt{N}$. An analogous computation shows $\langle s^2 \rangle = \sigma^2$, which is the reason for the $N - 1$ (Bessel) denominator. □

The standard uncertainty quoted for a type-A evaluation is therefore $u = s/\sqrt{N}$. The ubiquity of the Gaussian distribution for random errors is explained by the central limit theorem, proved as Theorem 11.46: a sum of many small independent disturbances tends to a Gaussian whatever the individual disturbances look like. The form that applies here is the Lindeberg–Feller one (Theorem 11.48), since the disturbances acting on a measurement — a thermal drift, a vibration, a quantisation step, a reading error — are independent but certainly not identically distributed.

Two hypotheses of that theorem are worth keeping in view, because a measurement can violate either. The contributions must have finite variance: a heavy-tailed error source need not average down at all (Remark 11.47). And none may dominate — Theorem 11.48 implies that every individual source contributes a vanishing fraction of the total variance, so a single dominant source defeats its hypothesis. The converse does not hold, and the theorem is not a licence to argue backwards: failing the condition does not by itself make the total non-Gaussian, and satisfying every variance share is not enough on its own.

The rate of approach is $O(M^{-1/2})$ in the number M of contributing sources, by the Berry–Esseen bound Equation (11.40) — which additionally assumes a finite third absolute moment. That bound is uniform in the deviation, so it constrains the *absolute* error of the Gaussian approximation equally everywhere; in the tails, where the probability being approximated is itself tiny, the *relative* error is correspondingly larger. Since a coverage factor $k = 3$ is a claim about a tail probability, it deserves more caution than the same approximation applied near the centre.

2.3.2 Propagation of uncertainty

Derived quantities inherit uncertainty from their inputs.

Theorem 2.8 (First-order propagation)**.** *Let $f(x_1, \dots, x_n)$ be differentiable, and let the x_i be independent measurements with standard uncertainties u_i small enough that f is approximately linear over their spread. Then the standard uncertainty of $y = f(x_1, \dots, x_n)$ is*

$$u_y^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 u_i^2 . \tag{2.3}$$

Derivation. *Derives Theorem 2.8.*

Expand f to first order about the means, $y - \langle y \rangle \approx \sum_i \frac{\partial f}{\partial x_i}(x_i - \langle x_i \rangle)$. Squaring and taking expectations, the cross terms $\langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$ vanish for $i \neq j$ by independence, leaving exactly Equation (2.3). For correlated inputs the covariances reappear as $2 \sum_{i<j} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \operatorname{cov}(x_i, x_j)$ [2129]. □

Corollary 2.9 (Quotients and powers)**.** *For $y = x_1^{a_1} \cdots x_n^{a_n}$, relative uncertainties combine as $(u_y/y)^2 = \sum_i a_i^2 \, (u_i/x_i)^2$.*
Rests on Theorem 2.8.

Derivation. *Derives Corollary 2.9.*

Apply Theorem 2.8 to $\ln y = \sum_i a_i \ln x_i$, for which $\partial \ln y / \partial \ln x_i = a_i$. □

2.3.3 Combining and fitting data

Theorem 2.10 (Weighted mean)**.** *Given independent measurements x_i of the same quantity with uncertainties u_i, the linear combination of minimal variance is*

$$\hat{x} = \frac{\sum_i x_i/u_i^2}{\sum_i 1/u_i^2}\,, \qquad u_{\hat{x}}^2 = \left(\sum_i \frac{1}{u_i^2}\right)^{-1}. \tag{2.4}$$

Rests on Theorem 2.7.

Derivation. *Derives Theorem 2.10.*

Seek $\hat{x} = \sum_i w_i x_i$ with $\sum_i w_i = 1$ (unbiasedness) and $\operatorname{var}\hat{x} = \sum_i w_i^2 u_i^2$ minimal. With a Lagrange multiplier λ (Chapter 21 uses the same device for constraints), $\partial_{w_i}\left[\sum_j w_j^2 u_j^2 - \lambda\left(\sum_j w_j - 1\right)\right] = 0$ gives $w_i = \lambda/(2u_i^2)$; the constraint fixes λ, yielding Equation (2.4), and substitution gives the stated variance. □

Theorem 2.11 (Least-squares line)**.** *Given points (x_i, y_i), $i = 1, \dots, N$, with equal uncertainties on the y_i and negligible uncertainty on the x_i, the line $y = a + bx$ minimizing $\chi^2 = \sum_i (y_i - a - bx_i)^2$ has*

$$b = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}\,, \qquad a = \bar{y} - b\,\bar{x}\,, \tag{2.5}$$

where overbars denote sample means.

Rests on Definition 2.6.

Derivation. *Derives Theorem 2.11.*

$\partial_a \chi^2 = -2\sum_i (y_i - a - bx_i) = 0$ gives $\bar{y} = a + b\bar{x}$; $\partial_b \chi^2 = -2\sum_i x_i(y_i - a - bx_i) = 0$ gives $\overline{xy} = a\bar{x} + b\overline{x^2}$. Solving the two linear equations yields Equation (2.5). The second-derivative matrix is positive definite (its determinant is proportional to the variance of the x_i), so the extremum is the minimum. □

2.3.4 Reporting

A result is reported as $y = (\hat{y} \pm u)\,[\,\text{unit}\,]$, the uncertainty with at most two significant figures and the value rounded to match. Where higher confidence is wanted, an expanded uncertainty $U = k\,u$ is quoted with the *coverage factor* k stated ($k = 2$ covers about 95 % for a Gaussian) [2129]. When an experiment box in this treatise quotes asymmetric bounds (as the gravitational-wave observations of Chapter 47 do), they are the bounds of the quoted posterior interval of the discovery analysis.

2.4 A worked example: timing a pendulum

Suppose a pendulum of length $\ell = (0.9500 \pm 0.0005)\,\text{m}$ swings with small amplitude, and 50 periods are timed five times with a stopwatch reading to 0.01 s: $T_{50} = 97.81\,\text{s}$, $97.95\,\text{s}$, $97.73\,\text{s}$, $98.02\,\text{s}$, $97.89\,\text{s}$. The sample mean is $\bar{T}_{50} = 97.88\,\text{s}$ with $s = 0.11\,\text{s}$, so $u_{\bar{T}_{50}} = s/\sqrt{5} = 0.05\,\text{s}$, and the period is $T = (1.9576 \pm 0.0010)\,\text{s}$. With $T = 2\pi\sqrt{\ell/g}$ (derived in Chapter 28), $g = 4\pi^2\ell/T^2$; Corollary 2.9 gives

$$\frac{u_g}{g} = \sqrt{\left(\frac{u_\ell}{\ell}\right)^2 + \left(2\,\frac{u_T}{T}\right)^2} \approx 1.2 \times 10^{-3}\,, \tag{2.6}$$

so $g = (9.79 \pm 0.01)\,\text{m/s}^2$ — consistent with the local value and eight orders of magnitude less precise than the ballistic gravimetry of Section 20.4, which is the same physics with the error budget engineered down [1906, 1945].

Part II

Mathematical Methods

Chapter 3

Logic, Sets, and Maps

Propositional logic, sets, maps, and equivalence classes form the raw language in which the rest of this treatise is written. This chapter fixes that language: what we mean by a definition, an axiom, and a theorem; how propositions combine under logical operators; what sets, numbers, and maps are; and how an equivalence relation partitions a set into classes.

The chapter then turns the language on itself. Comparing infinite sets uncovers a strict hierarchy of infinities (Section 3.5); unrestricted set-building collapses in Russell's paradox, and the Zermelo–Fraenkel axioms repair it (Section 3.6); Hilbert's demand that mathematics certify its own consistency, completeness, and decidability is answered — negatively — by Gödel's incompleteness theorems (Section 3.8) and by Turing's analysis of computation (Section 3.9). We close by asking, with the full twentieth-century record on the table, whether mathematics can be built free of contradiction, and what the honest answer means for a book that intends to derive physics from evidence and proof (Section 3.10). A physics treatise owes its readers this chapter: every derivation in the book is a formal proof in the sense made precise here, and the limits of that method are part of its description.

3.1 Logic

3.1.1 Propositions

Definitions

The definition of the word *definition* is redundant in an obvious sense. We could say instead: the accepted meaning of *definition* is... — but then, what is the definition of *accepted meaning*? Better still: what is *something*? And beyond that: why does something exist? Whenever we try to reach a primordial question we run into problems, circular loops, and paradoxes. To avoid entering into philosophical discussion (though such discussion is never wasted), we simply adopt the dictionary meaning [2236]: a *definition* is a declarative sentence that states with clarity and exactness the generic and specific characteristics of something, concrete or abstract.

In mathematical notation, when we wish to record a definition — that is, when an equality holds not as the result of other equalities but by fiat — we use the symbol :=. For example, let $\Diamond$ be an object. We define $\blacklozenge$ by $\blacklozenge := \Diamond^{-1}$, and we define $\Diamond\Diamond^{-1}$ by $\Diamond\Diamond^{-1} := \circlearrowleft$. Every time the object $\blacklozenge$ appears we may replace it by $\Diamond^{-1}$, since by definition they are exactly the same thing, and likewise we may replace $\Diamond\Diamond^{-1}$ by $\circlearrowleft$. For instance,

$$\Diamond\blacklozenge = \Diamond\Diamond^{-1} = \circlearrowleft \ .$$

A definition need not be a fundamental truth. However, for an idea built from definitions — together with the observations made upon them — to be consistent, a definition must be

irrevocable within its context (at least, this book demands it). We cannot build an idea on definitions that we later change, much less on new definitions that contradict the earlier ones. In what follows, the definitions we give will not be dictionary entries but will be tailored to the studies we undertake.

Generalization and specification of a definition

It is often necessary to extend a definition so that it covers more concepts; when this happens one usually renames the definition to fit the new context. As a toy example, define first a *hydroliving* object as an object that lives on Earth and contains at least ten hydrogen atoms. We ourselves are then hydroliving objects. Now extend the definition to an object living anywhere in the Milky Way that contains at least ten hydrogen atoms, which we call *locally hydroliving*: we have generalized the definition, and we are locally hydroliving. We may also *sharpen* the parameters of a definition, cutting down the candidates that satisfy it: call *submarine hydroliving* a hydroliving object that is born, lives, and reproduces in the sea.

We call *generalization* and *specification* of a definition, respectively, the definitions obtained by enlarging or reducing the set of candidates satisfying the original one. Both notions are relative to the aspect of the definition being modified; for example, one may generalize and specify simultaneously by defining a *locally hydrolivoid* object as one living in the Milky Way but containing at least 1 000 000 hydrogen atoms.

Definition 3.1 (Logical proposition)**.** A *logical proposition* p (or simply a proposition) is a declarative sentence that can be only one of two ways: true or false. The quality of a proposition of being true or false is its *truth value*, which we denote by T if true and F if false.

For example, let $p =$ "Albert Einstein was born in 1879". The truth value of p is T, since that is indeed what happened. Let $q =$ "Paul Adrien Maurice Dirac was born in 1790". The truth value of q is F, because Dirac was born in 1902.

Definition 3.2 (Axiom)**.** An *axiom* is a proposition assumed irrevocably true.

An axiom, accompanied by the pertinent definitions, allows one to build a self-consistent idea (not necessarily consistent with other ideas). The word *axiom* is used in mathematics; in physics one uses, depending on the importance of the proposition in its context, the words *law*, *principle*, or *postulate*. As an example, let $\ltimes$ be given and define $\rtimes := -\ltimes$. Separately, enunciate the axiom $-(-\ltimes) = \ltimes$. We conclude a new, consistent statement: $-\rtimes = \ltimes$.

Definition 3.3 (Theorem)**.** Within a context of axioms and definitions, a *theorem* is a proposition whose truth value has been demonstrated.

In the preceding example it is a theorem that $-\rtimes = \ltimes$. In the literature it is not customary to call every demonstrated result a theorem: depending on the relevance of the result one speaks of a *theorem*, a *lemma*, a *corollary*, or even a *proposition* — the last not to be confused with a logical proposition.

Definition 3.4 (Theory)**.** A *theory* is an agglomeration of axioms, definitions, theorems, and interpretations of results that model the behaviour of a concrete or abstract entity, and whose good correspondence between model and results is vouched for by us as human beings, through our senses and available technology.

3.1.2 Logical operators

Given a proposition, we can form new propositions from it. For example, let $p =$ "It is raining outside". The truth value of p is not settled — until we look outside we

do not know whether it is raining — but from p we may consider the proposition q = "It is not raining outside". Moreover, given the proposition r = "I put on a coat", from p and r we can form s = "If it is raining outside, then I put on a coat".

Definition 3.5 (Logical operator)**.** A *logical operator* is an entity that takes one or more propositions and generates a new proposition.

Definition 3.6 (Truth table)**.** The truth value of a proposition generated by a logical operator is a function of the truth values of the propositions that generate it. The table of all possible truth values of the generated proposition is its *truth table*.

In the trivial case, the truth table of a single proposition p is

p
T
F

(3.1)

Definition 3.7 (Negation)**.** Let p be a proposition. The *negation* operator $\neg$ takes p into the proposition

$$\neg p = \text{"It is not true that } p\text{"}\,, \tag{3.2}$$

whose truth table is

p	$\neg p$
T	F
F	T

(3.3)

For example, let p = "Today is Monday". Its negation is

$$\neg p = \text{"It is not true that today is Monday"}\,,$$

which we read more naturally as "Today is not Monday". According to the table, if it is true that today is Monday, then it is false that today is not Monday; conversely, if it is false that today is Monday, it is true that today is not Monday.

Definition 3.8 (Conjunction)**.** Let p and q be propositions. The *conjunction* operator $\wedge$ takes p and q into the proposition

$$p \wedge q = p \text{ and } q\,, \tag{3.4}$$

whose truth table is

p	q	$p \wedge q$
T	T	T
T	F	F
F	T	F
F	F	F

(3.5)

For example, let p = "I am in Asia" and q = "I am in Africa". Then $p \wedge q$ = "I am in Asia and I am in Africa". Common sense tells us it is impossible to be in two places at once, i.e. $p \wedge q$ is false. According to the table this may be because it is false that I am in Asia, false that I am in Africa, or false that I am in either place.

Definition 3.9 (Disjunction)**.** Let p and q be propositions. The *disjunction* operator $\vee$ takes p and q into the proposition

$$p \vee q = p \text{ or } q\,, \tag{3.6}$$

whose truth table is

p	q	$p \vee q$
T	T	T
T	F	T
F	T	T
F	F	F

(3.7)

For example, let $p =$ "I will arrive at 8:00" and $q =$ "I will arrive at 8:05". Then $p \vee q =$ "I will arrive at 8:00 or at 8:05".

Definition 3.10 (Exclusive disjunction)**.** The *exclusive disjunction* operator $\veebar$ generates the proposition $p \veebar q$ with truth table

p	q	$p \veebar q$
T	T	F
T	F	T
F	T	T
F	F	F

(3.8)

Definition 3.11 (Conditional)**.** Let p and q be propositions. The *conditional* operator $\rightarrow$ takes p and q into the proposition

$$p \rightarrow q = \text{"If } p\text{, then } q\text{"}\,, \tag{3.9}$$

whose truth table is

p	q	$p \rightarrow q$
T	T	T
T	F	F
F	T	T
F	F	T

(3.10)

For example, let $p =$ "Today is Sunday" and $q =$ "Today I will rest". Then $p \rightarrow q =$ "If today is Sunday, then I will rest". From the table: if today really is Sunday but it is false that I will rest, then it is indeed false that "if today is Sunday, then I will rest". On the other hand, if today is not Sunday but I do rest anyway, the conditional stands. Even if today is neither Sunday nor a day I rest, the proposition is true, because my not resting on a non-Sunday takes nothing away from my resting when Sunday does come.

Definition 3.12 (Biconditional)**.** Let p and q be propositions. The *biconditional* operator $\leftrightarrow$ takes p and q into the proposition

$$p \leftrightarrow q = p \text{ if and only if } q\,, \tag{3.11}$$

whose truth table is

p	q	$p \leftrightarrow q$
T	T	T
T	F	F
F	T	F
F	F	T

(3.12)

For example, let $p =$ "Wake me up" and $q =$ "It is 10:00". Then $p \leftrightarrow q =$ "Wake me up if and only if it is 10:00".

Definition 3.13 (Logical expression)**.** A *logical expression* is a proposition composed of propositions and logical operators.

For example, given propositions p and q, a logical expression is

$$s = \neg p \wedge (p \to q)\,. \tag{3.13}$$

The expression s can take various truth values depending on the truth values of p and q. Let us build its truth table. First, from Equation (3.3),

p	$\neg p$
T	F
T	F
F	T
F	T

From Equation (3.10),

p	q	$p \to q$
T	T	T
T	F	F
F	T	T
F	F	T

Then, from Equation (3.5),

$\neg p$	$p \to q$	$\neg p \wedge (p \to q)$
F	T	F
F	F	F
T	T	T
T	T	T

3.1.3 Implications and logical equivalences

Definition 3.14 (Tautology)**.** A logical expression is a *tautology* if all of its possible truth values are T.

Definition 3.15 (Contradiction)**.** A logical expression is a *contradiction* if all of its possible truth values are F.

Definition 3.16 (Contingency)**.** A logical expression is a *contingency* if among its possible truth values we find both truths and falsehoods.

All the logical operators above generate propositions that are contingencies, as their truth tables show. Indeed, any single proposition is a contingency, since it can be true or false. A more elaborate example is the expression defined in Equation (3.13).

Definition 3.17 (Logical equivalence)**.** Two logical expressions s and t built from the same propositions are *logically equivalent*, written $s \equiv t$, if and only if the expression $s \leftrightarrow t$ is a tautology — equivalently, if and only if s and t have identical truth tables.

Rests on Definitions 3.12 and 3.14.

Remark 3.18 (Object language and metalanguage). The distinction between $\leftrightarrow$ and $\equiv$ is not pedantry, and it will matter a great deal in Section 3.8. The connective $\leftrightarrow$ belongs to the language we are studying: it builds one proposition out of two. The symbol $\equiv$ belongs to the language in which we *talk about* that language, and asserts a fact about truth tables. The first is called the *object language*, the second the *metalanguage*. The same split separates $\to$ from $\Longrightarrow$ and, later, provability from truth. Confusing the two levels is the single most common way to misread Gödel's theorems.

Proposition 3.19 (Conditional as disjunction)**.** *For any propositions p and q,*

$$(p \to q) \equiv (\neg p \vee q)\,. \tag{3.14}$$

Rests on Definitions 3.7, 3.9, 3.11 and 3.17.

Proof. *Derives Proposition 3.19.*

Both expressions are built from the same two propositions, so both truth tables have the same four rows. Reading $\neg p$ from Equation (3.3), $p \to q$ from Equation (3.10) and $\neg p \vee q$ from the disjunction table,

p	q	$\neg p$	$p \to q$	$\neg p \vee q$
T	T	F	T	T
T	F	F	F	F
F	T	T	T	T
F	F	T	T	T

The fourth and fifth columns agree row by row, so $(p \to q) \leftrightarrow (\neg p \vee q)$ takes the value T in every row and is a tautology. □

Proposition 3.20 (De Morgan's laws)**.** *For any propositions p and q,*

$$\neg(p \wedge q) \equiv \neg p \vee \neg q\,, \tag{3.15}$$

$$\neg(p \vee q) \equiv \neg p \wedge \neg q\,. \tag{3.16}$$

Rests on Definitions 3.7 to 3.9 and 3.17.

Proof. *Derives Proposition 3.20.*

Tabulate all four rows at once:

p	q	$\neg(p \wedge q)$	$\neg p \vee \neg q$	$\neg(p \vee q)$	$\neg p \wedge \neg q$
T	T	F	F	F	F
T	F	T	T	F	F
F	T	T	T	F	F
F	F	T	T	T	T

Columns three and four agree, and so do columns five and six. □

Proposition 3.21 (Contraposition)**.** *For any propositions p and q,*

$$(p \to q) \equiv (\neg q \to \neg p)\,. \tag{3.17}$$

Rests on Proposition 3.19 and Equation (3.3).

Proof. *Derives Proposition 3.21.*

By Proposition 3.19 applied to the right-hand side, $\neg q \to \neg p$ is equivalent to $\neg(\neg q) \vee \neg p$. Double negation gives $\neg(\neg q) \equiv q$ — immediate from Equation (3.3), read twice — so the right-hand side is equivalent to $q \vee \neg p$. Disjunction is symmetric in its two arguments, as its truth table shows, hence $q \vee \neg p \equiv \neg p \vee q$, which by Proposition 3.19 again is $p \to q$. □

The contrapositive $\neg q \to \neg p$ must not be confused with the *converse* $q \to p$, which is *not* equivalent to $p \to q$: rows two and three of Equation (3.10) differ under the exchange of p and q. Asserting the converse from the conditional is the fallacy of *affirming the consequent*, and it is a standing hazard in experimental physics: a theory p predicting an observation q is not confirmed as true by observing q, since some rival theory may predict the same q. What the observation of $\neg q$ does establish — by Equation (3.17) — is $\neg p$. This asymmetry between confirmation and refutation is the logical skeleton of falsificationism [1173] and of Remark 1.2.

Proposition 3.22 (Algebra of propositions)**.** *For any propositions p, q, r the following equivalences hold:*

$$\neg(\neg p) \equiv p\,, \tag{3.18}$$

$$p \wedge q \equiv q \wedge p\,, \qquad p \vee q \equiv q \vee p\,, \tag{3.19}$$

$$(p \wedge q) \wedge r \equiv p \wedge (q \wedge r)\,, \qquad (p \vee q) \vee r \equiv p \vee (q \vee r)\,, \tag{3.20}$$

$$p \wedge (q \vee r) \equiv (p \wedge q) \vee (p \wedge r)\,, \tag{3.21}$$

$$p \vee (q \wedge r) \equiv (p \vee q) \wedge (p \vee r)\,, \tag{3.22}$$

$$p \wedge p \equiv p\,, \qquad p \vee p \equiv p\,, \tag{3.23}$$

$$p \wedge (p \vee q) \equiv p\,, \qquad p \vee (p \wedge q) \equiv p\,, \tag{3.24}$$

$$p \vee \neg p \equiv T\,, \tag{3.25}$$

$$p \wedge \neg p \equiv F\,, \tag{3.26}$$

$$(p \leftrightarrow q) \equiv (p \to q) \wedge (q \to p)\,. \tag{3.27}$$

Rests on Definitions 3.6 and 3.17.

Proof. *Derives Proposition 3.22.*

Each is a finite truth-table check: the expressions involve at most three propositions, hence at most $2^3 = 8$ rows, and the two sides are compared row by row exactly as in Proposition 3.20. We display only the least mechanical one, the first distributive law Equation (3.21):

p	q	r	$q \vee r$	$p \wedge (q \vee r)$	$(p \wedge q) \vee (p \wedge r)$
T	T	T	T	T	T
T	T	F	T	T	T
T	F	T	T	T	T
T	F	F	F	F	F
F	T	T	T	F	F
F	T	F	T	F	F
F	F	T	T	F	F
F	F	F	F	F	F

The last two columns coincide. □

Equation (3.25) is the *law of the excluded middle*: a proposition is true or its negation is, with no third option. It is a theorem of the two-valued truth tables of Definition 3.1 precisely because we built those tables with two values. Whether it should be assumed when reasoning about infinite collections — where no finite inspection can decide which disjunct holds — is the point at which intuitionism parts company with classical mathematics (Section 3.6.4). This treatise reasons classically, and says so.

3.1.4 Predicates and quantifiers

Propositional logic is too coarse for mathematics or physics. The sentence "x is greater than 3" has no truth value until x is supplied; only then does it become a proposition. Such an incomplete sentence is a predicate, and completing it is what quantifiers do.

Definition 3.23 (Predicate)**.** A *predicate* P on a set X is a rule assigning to each $x \in X$ a proposition $P(x)$. The set of elements for which the proposition is true,

$$\{x \in X \mid P(x)\}\,, \tag{3.28}$$

is the *truth set* of P. Predicates of several variables, $P(x, y)$, $P(x, y, z), \ldots$, are defined in the same way on products of sets (Definition 3.38).

Rests on Definitions 3.1 and 3.38.

Definition 3.24 (Quantifiers)**.** Let P be a predicate on a set X. The *universal quantifier* $\forall$ and the *existential quantifier* $\exists$ build the propositions

$$\forall x \in X,\ P(x) = \text{“}P(x) \text{ is true for every } x \text{ in } X\text{”}, \tag{3.29}$$

$$\exists x \in X \mid P(x) = \text{“}P(x) \text{ is true for at least one } x \text{ in } X\text{”}. \tag{3.30}$$

We write $\exists!\, x \in X \mid P(x)$ — “there exists a unique x” — for the conjunction of $\exists x \in X \mid P(x)$ with the uniqueness clause $\forall x, x' \in X,\ (P(x) \wedge P(x')) \to x = x'$.

Rests on Definition 3.23.

Over a finite set $X = \{x_1, \dots, x_n\}$ the quantifiers are abbreviations for finite connectives,

$$\forall x \in X,\ P(x) \ \equiv\ P(x_1) \wedge \cdots \wedge P(x_n), \qquad \exists x \in X \mid P(x) \ \equiv\ P(x_1) \vee \cdots \vee P(x_n), \tag{3.31}$$

and this is the way to remember their negations: De Morgan's laws (Proposition 3.20) pass to the limit of infinite X, where the finite reading is no longer available but the rule survives,

$$\neg\,(\forall x \in X,\ P(x)) \equiv \exists x \in X \mid \neg P(x), \tag{3.32}$$

$$\neg\,(\exists x \in X \mid P(x)) \equiv \forall x \in X,\ \neg P(x). \tag{3.33}$$

To refute a universal claim one exhibits a single counterexample; to refute an existential claim one must argue about every element. This is why “no experiment has ever violated the second law” is a far weaker statement than “the second law holds”, and why the search for a violating experiment is nonetheless the right way to test it.

Remark 3.25 (The order of quantifiers is not free)*.* Quantifiers of the same kind commute; quantifiers of different kinds do not. The proposition

$$\forall \varepsilon > 0,\ \exists \delta > 0 \mid \cdots$$

allows δ to depend on ε, while

$$\exists \delta > 0 \mid \forall \varepsilon > 0,\ \cdots$$

demands one δ serving all ε at once, and is strictly stronger. The whole difference between continuity and uniform continuity, or between pointwise and uniform convergence, lives in that exchange (Chapter 7); so does the difference between a bound that holds for each energy and a bound that holds uniformly in energy.

3.1.5 Methods of proof

Every result asserted in this treatise carries a derivation, by editorial rule. The derivations use a small and closed repertoire of argument forms, each of which is a tautology of Proposition 3.22 put to work.

1. **Direct proof.** To establish $p \to q$, assume p and derive q by a chain of previously established implications. The underlying tautology is transitivity of the conditional, $((p \to r) \wedge (r \to q)) \to (p \to q)$, together with *modus ponens*, $(p \wedge (p \to q)) \to q$.

2. **Contraposition.** To establish $p \to q$, assume $\neg q$ and derive $\neg p$. Legitimate by Proposition 3.21. The proof of injectivity in Definition 3.45 uses exactly this move.

3. **Contradiction** (*reductio ad absurdum*). To establish p, assume $\neg p$ and derive a contradiction $q \wedge \neg q$, which is false by Equation (3.26); since a false conclusion cannot follow from a true assumption, $\neg p$ is false and, by Equation (3.18), p holds.

4. **Cases.** If $p_1 \vee \cdots \vee p_n$ is a tautology and each $p_i \to q$ has been proved, then q holds. The two-case version rests on Equation (3.25).

5. **Induction.** To establish $\forall n \in \mathbb{N},\ P(n)$, prove $P(1)$ and prove $\forall n \in \mathbb{N},\ P(n) \to P(n+1)$. This is not a tautology of propositional logic but an axiom about $\mathbb{N}$ — Axiom 3.34 below — and it is the one place where the structure of the natural numbers enters a proof.

6. **Counterexample.** To refute $\forall x \in X,\ P(x)$, exhibit one x with $\neg P(x)$, by Equation (3.32). One suffices; no number of confirming instances does the opposite job.

Example 3.26 (Irrationality of the square root of two)**.** Proof by contradiction, in its classical instance. Suppose $\sqrt{2} \in \mathbb{Q}$. Then there are $n, m \in \mathbb{Z}$ with $m \neq 0$ such that $\sqrt{2} = n/m$, and we may take the fraction in lowest terms, i.e. with n and m not both even. Squaring and clearing the denominator,

$$n^2 = 2m^2\,, \tag{3.34}$$

so n^2 is even. An odd $n = 2k+1$ has $n^2 = 2(2k^2+2k)+1$ odd, so by contraposition n is even: $n = 2j$ for some $j \in \mathbb{Z}$. Substituting into Equation (3.34) gives $4j^2 = 2m^2$, that is $m^2 = 2j^2$, and the same argument makes m even. Then n and m are both even, contradicting the choice of lowest terms. Hence $\sqrt{2} \notin \mathbb{Q}$, and the irrational numbers of Section 3.2.2 are not an optional embellishment of $\mathbb{Q}$ but a necessity forced by the diagonal of a unit square.

3.2 Sets

Set theory descends through Cantor, Frege, Russell, and Zermelo–Fraenkel. In what follows we shall not include all the formalism that the full theory demands; we give only the definitions, axioms, and theorems that allow us to understand the many aspects of mathematical physics whose properties we can link with nature.

3.2.1 Sets

Definition 3.27 (Element)**.** We consider an *element* x to be an object that satisfies some collection of logical propositions.

Definition 3.28 (Set)**.** A *set* is an abstract grouping of elements that satisfy the same logical propositions.

For example, the vowels form the set X given by

$$\begin{aligned} X &= \{x \mid x \text{ is a vowel}\} \\ &= \{a, e, i, o, u\}\ . \end{aligned}$$

Notation 3.29 (Membership). That x is an element of the set X is written $x \in X$; that it is not, $x \notin X$. Membership is the sole primitive relation of set theory: every other notion in this section is defined from it, from $\in$ and from the logical apparatus of Section 3.1. The two displays above illustrate the two ways of presenting a set — by *extension*, listing its elements, and by *comprehension*, giving a predicate (Definition 3.23) whose truth set it is.

Definition 3.30 (Subset)**.** Let X and Y be sets. We say X is a *subset* of Y, written $X \subset Y$, if and only if every element of X is an element of Y:

$$X \subset Y \iff \forall x \in X,\ x \in Y\,. \tag{3.35}$$

In particular, every set is a subset of itself. If $X \subset Y$ and $X \neq Y$ we call X a *proper* subset of Y.

Two sets are equal when they have the same elements, and nothing else about them counts: a set is determined by its extension alone. This is the *principle of extensionality*, and in the axiomatic treatment of Section 3.6.2 it is promoted to an axiom. It is what licenses the double-inclusion argument Equation (3.47) used throughout this book, and it is the reason $\{a, e, i, o, u\}$ and $\{e, a, u, i, o, a\}$ are the same set: order and repetition are features of the notation, not of the object.

Definition 3.31 (Empty set)**.** The *empty set* is the set with no elements, denoted $\varnothing$.

Every set has the empty set as a subset: the condition Equation (3.35) quantifies over the elements of $\varnothing$, of which there are none, so it holds vacuously — an instance of rows three and four of Equation (3.10). Extensionality makes the empty set unique, since any two sets with no elements have the same elements. That so much can be built from nothing at all is the first sign that the naive notion of set carries more weight than it looks able to bear; Section 3.6 shows where it breaks and how the Zermelo–Fraenkel axioms repair it.

Definition 3.32 (Trivial subsets)**.** Given a set X, it always has as subsets itself and the empty set. When it matters to refer to this fact, we call these the *trivial* subsets of X.

3.2.2 Numbers

Number theory is an essential branch of mathematics that studies the properties of numbers. We shall not define *number*; instead we explain a way of understanding the familiar number sets.

Natural numbers $\mathbb{N}$

The set known as the *natural numbers* is a set of numbers with which we can identify, for example, a quantity of whole objects. Concretely: take the proposition "a farmer owns cows", assumed true. We can then assert that the farmer has one, two, three, or more cows. The farmer cannot have infinitely many cows, for they would not fit in the world, and cannot have none, for that would be a contradiction. We write

$$\mathbb{N} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, \ldots\} \,. \tag{3.36}$$

That description says what $\mathbb{N}$ is *for*, not what it *is*. The axiomatic characterization is Peano's [245], and it needs one undefined operation: the successor, "the next one after".

Axiom 3.33 (Peano axioms)**.** The natural numbers form a set $\mathbb{N}$ carrying a distinguished element $1 \in \mathbb{N}$ and a map $S : \mathbb{N} \longrightarrow \mathbb{N}$, called the *successor*, such that

1. 1 is not a successor: $\forall n \in \mathbb{N},\ S(n) \neq 1$;
2. distinct numbers have distinct successors, i.e. S is injective (Definition 3.45):

$$S(n) = S(m) \implies n = m \,.$$

The two clauses forbid the successor chain from closing back on itself, but by themselves they do not forbid stray elements sitting outside the chain $1, S(1), S(S(1)), \ldots$ altogether. Excluding those is the work of the fifth and deepest of Peano's axioms, which is also the licence for the induction proofs of Section 3.1.5.

Axiom 3.34 (Principle of induction)**.** Let P be a predicate on $\mathbb{N}$. If $P(1)$ holds and $\forall n \in \mathbb{N},\ P(n) \to P(S(n))$ holds, then $\forall n \in \mathbb{N},\ P(n)$. Equivalently, in the language of Definition 3.30: the only subset of $\mathbb{N}$ containing 1 and closed under S is $\mathbb{N}$ itself.

Addition and multiplication are then *defined* by recursion on the successor — $n+1 := S(n)$, $n+S(m) := S(n+m)$, $n \cdot 1 := n$, $n \cdot S(m) := n \cdot m + n$ — and every familiar arithmetic law is a theorem proved by induction. This matters beyond arithmetic: the axioms above, written in first-order logic with induction restricted to predicates expressible in that language, constitute *Peano arithmetic* (Definition 3.83), the theory whose incompleteness Gödel proved.

Integers $\mathbb{Z}$

The set known as the *integers* consists of the natural numbers together with a new element, the *number zero*, denoted 0, which quantifies the existence of nothing: given the proposition "a farmer owns no sheep", we say the farmer has zero sheep.

The set also contains, for each natural number, a dual element manifesting its opposite. Consider a sport played by two teams, where a team gains a point by stealing a banner from the other and loses a point when a banner is stolen from it. If a team has two banners stolen at the start, what is its score? Clearly it has gained nothing, but we intuit it holds something dual to the number 2 — something like "minus 2". Writing these dual numbers with the minus sign $-$ on the left, the set of integers is

$$\mathbb{Z} = \{\ldots, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, \ldots\} \ . \tag{3.37}$$

Several subsets of $\mathbb{Z}$ are heavily used in mathematics and physics. Among them:

1. *Even numbers*: the set

$$\{2n \mid n \in \mathbb{Z}\} = \{\ldots, -4, -2, 0, 2, 4, \ldots\} \ , \tag{3.38}$$

   that is, the set of all multiples of 2 (positive and negative), clearly a subset of $\mathbb{Z}$.

2. *Odd numbers*: the set

$$\{2n+1 \mid n \in \mathbb{Z}\} = \{\ldots, -5, -3, -1, 1, 3, 5, \ldots\} \ , \tag{3.39}$$

   that is, the set of all multiples of 2, plus or minus 1.

3. *Prime numbers*: the natural numbers that are multiples only of 1 and of themselves. For example, 4 is a multiple of 1, 2, and 4, so it is not prime; 5 is a multiple only of 1 and 5, so it is prime. By convention 1 is not considered prime, since the primes satisfy many properties in number theory that 1 does not. The set of primes is denoted $\mathbb{P}$; its first ten elements are

$$\mathbb{P} = \{2, 3, 5, 7, 11, 13, 17, 19, 23, 29, \ldots\} \ . \tag{3.40}$$

   Note that the only even prime is 2.

Rational numbers $\mathbb{Q}$

The set known as the *rational numbers* is a set of numbers with which we can identify portions. If we have a watermelon and eat half of it, we say we have eaten the fraction "one half" of the watermelon, written $\frac{1}{2}$. We could likewise have eaten $\frac{1}{3}$ of it, or even $\frac{2}{2}$, which corresponds to eating the whole watermelon. In a fraction, the upper number is the *numerator* and the lower one the *denominator*. These new numbers, together with the integers, form the set of rational numbers

$$\mathbb{Q} = \left\{ \frac{n}{m} \mid n, m \in \mathbb{Z},\ m \neq 0 \right\} \ , \tag{3.41}$$

where zero is excluded as a denominator because it leads to complications: there is no practical or canonical way to define how many times zero fits into an arbitrary number.

Irrational numbers $\mathbb{R} \setminus \mathbb{Q}$

Geometry produces new numbers whose infinite decimal expansions follow no known general formula. A classic example is the number π, the ratio of the perimeter of a circle to its diameter,

$$\pi = \frac{P_{\bigcirc}}{D_{\bigcirc}} = 3.141\,592\,653\,589\,793\,238\,46\ldots\,. \tag{3.42}$$

Another example is the length of the diagonal of a unit square, the Pythagorean number $\sqrt{2}$. These new numbers are called the *irrational numbers*. That they exist at all — that the rationals do not already exhaust the number line — is not an observation but a theorem, proved for $\sqrt{2}$ in Example 3.26. That they are in fact the overwhelming majority of the real numbers is a much later and much stranger result, Theorem 3.69.

Real numbers $\mathbb{R}$

The rationals together with the irrationals form the *real numbers* $\mathbb{R}$. We have no general symbolic rule for them, but graphically we can represent them by a straight line built out of them, as in Figure 3.1: each $x \in \mathbb{R}$ is a point of the line.

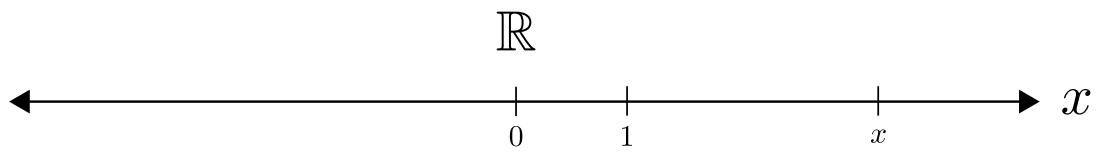


Figure 3.1: The real line: a graphical representation of the set $\mathbb{R}$.

The number sets assemble into the hierarchy

$$\mathbb{C} \supset \begin{cases} \mathbb{R} \supset \begin{cases} \mathbb{Q} \supset \mathbb{Z} \supset \mathbb{N} \\ \mathbb{R} \setminus \mathbb{Q} \end{cases} \\ \mathbb{I} \end{cases} \tag{3.43}$$

where $\mathbb{I}$ denotes the imaginary numbers and $\mathbb{C}$ the complex numbers, both studied in detail in Chapter 8.

3.2.3 Operations with sets

The logical operators of Section 3.1.2 induce operations on sets, because a set is the truth set of a predicate and the operators combine predicates. Each definition below pairs one connective with one symbol, and this pairing is the whole content of the subject.

Definition 3.35 (Union, intersection, difference)**.** Let X and Y be sets. We define

$$X \cup Y := \{x \mid x \in X \ \vee \ x \in Y\}\,, \tag{3.44}$$
$$X \cap Y := \{x \mid x \in X \ \wedge \ x \in Y\}\,, \tag{3.45}$$
$$X \setminus Y := \{x \mid x \in X \ \wedge \ x \notin Y\}\,, \tag{3.46}$$

called respectively the *union*, the *intersection* and the *difference* of X and Y. When all sets under discussion are subsets of one ambient set U, the *complement* of X in U is $X^{\mathrm{c}} := U \setminus X$.

Rests on Definitions 3.7 to 3.9.

The complement is defined relative to a stated ambient set and never absolutely. There is no set of all things to take complements in: assuming one is exactly the mistake that produces Russell's paradox (Theorem 3.75).

Definition 3.36 (Disjoint sets)**.** Two sets X and Y are *disjoint* if and only if $X \cap Y = \varnothing$.

Two properties will be used repeatedly in this book. First, mutual inclusion implies equality — this is extensionality in working form:

$$(X \subset Y) \ \wedge \ (Y \subset X) \implies X = Y \,. \tag{3.47}$$

Second, intersecting with the empty set yields the empty set:

$$X \cap \varnothing = \varnothing \,. \tag{3.48}$$

Proposition 3.37 (De Morgan's laws for sets)**.** *Let X, Y be subsets of an ambient set U. Then*

$$(X \cup Y)^{\mathrm{c}} = X^{\mathrm{c}} \cap Y^{\mathrm{c}} \,, \tag{3.49}$$

$$(X \cap Y)^{\mathrm{c}} = X^{\mathrm{c}} \cup Y^{\mathrm{c}} \,. \tag{3.50}$$

Rests on Definition 3.35 and Equations (3.15), (3.16) and (3.47).

Proof. *Derives Proposition 3.37.*

Let $u \in U$. By Definition 3.35, $u \in (X \cup Y)^{\mathrm{c}}$ means $\neg(u \in X \vee u \in Y)$, which by Equation (3.16) is equivalent to $(u \notin X) \wedge (u \notin Y)$, i.e. to $u \in X^{\mathrm{c}} \cap Y^{\mathrm{c}}$. Every step is an equivalence, so each set is contained in the other and Equation (3.47) gives equality. The second law follows the same way from Equation (3.15). □

That proof is the template for the whole algebra of sets: commutativity, associativity, distributivity and absorption of $\cup$ and $\cap$ are the set-theoretic shadows of Equations (3.19) to (3.22) and (3.24), obtained by the same translation and proved by the same two lines. We shall use them without further comment.

For families of more than two sets, indexed by an arbitrary set I — finite, countable, or neither — we write

$$\bigcup_{i \in I} X_i := \{x \mid \exists\, i \in I \mid x \in X_i\} \,, \qquad \bigcap_{i \in I} X_i := \{x \mid \forall\, i \in I, \ x \in X_i\} \,. \tag{3.51}$$

The pairing of $\cup$ with $\exists$ and of $\cap$ with $\forall$ is the same one as before, now in its quantified form; the negation rules Equations (3.32) and (3.33) turn into De Morgan's laws for arbitrary families.

Definition 3.38 (Ordered pair and Cartesian product)**.** The *ordered pair* (x, y) is an object satisfying

$$(x, y) = (x', y') \iff (x = x') \wedge (y = y') \,. \tag{3.52}$$

The *Cartesian product* of the sets X and Y is the set of ordered pairs

$$X \times Y := \{(x, y) \mid x \in X, \ y \in Y\} \,, \tag{3.53}$$

and analogously for finitely many factors; we write X^n for the product of n copies of X.

Remark 3.39 (Pairs are sets too). Equation (3.52) states the property an ordered pair must have, not what it is. If ordered pairs are to be sets like everything else, the standard implementation is Kuratowski's [477],

$$(x, y) := \{\{x\}, \{x, y\}\} \,, \tag{3.54}$$

from which Equation (3.52) is a short exercise in extensionality. The implementation is a convenience, not a discovery: nothing in mathematics or physics depends on *which* coding is chosen, only on the fact that *some* coding satisfies Equation (3.52). The same remark will apply, with far greater force, to Gödel's coding of formulas as numbers (Section 3.8.3).

Cartesian products are the reason this section belongs in a physics book. Configuration space is a product of position ranges, phase space a product of configuration and momentum spaces (Chapter 22), spacetime a product of a time axis with a spatial manifold in every chart, and a relation — including every equivalence relation of Section 3.4 and every map of Section 3.3 — is formally a subset of a product.

3.2.4 Cardinality

Definition 3.40 (Cardinal)**.** The *cardinal* of a set X, written $|X|$, is the number of elements it contains. A set is *finite* if and only if its cardinal is 0 or lies in $\mathbb{N}$; otherwise it is *infinite*.

For example, the cardinal of the set of vowels is $|\{a, e, i, o, u\}| = 5$, and $|\varnothing| = 0$. The cardinal of the set of naturals greater than or equal to 6 and less than 15 is $|\{6, 7, 8, 9, 10, 11, 12, 13, 14\}| = 9$. Other sets are infinite: $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and even the set of reals between 0 and 1.

Writing $|\mathbb{R}| = \infty$, as is often done informally, throws away the most interesting fact in this chapter. Both $\mathbb{N}$ and $\mathbb{R}$ are infinite, yet we intuit that $\mathbb{R}$ is somehow "more" infinite — and that intuition is correct, provable, and the beginning of a whole hierarchy (Section 3.5). Until the tools for comparing infinite sets are in place we confine the arithmetic of cardinals to finite sets.

Definition 3.41 (Power set)**.** Let X be a set. The *power set* of X is the set of all possible subsets of X, denoted $\mathcal{P}(X)$.

Rests on Definition 3.30.

For example, if $X = \{a, b, c\}$, then

$$\mathcal{P}(X) = \{\varnothing, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, X\} .$$

Proposition 3.42 (Cardinal of the power set)**.** *For every finite set X,*

$$|\mathcal{P}(X)| = 2^{|X|} . \tag{3.55}$$

Rests on Definitions 3.40 and 3.41, Axiom 3.34, and Equations (3.25) and (3.26).

Proof. *Derives Proposition 3.42.*

By induction (Axiom 3.34) on $n = |X|$, the statement $P(n)$ being "every set of cardinal n has exactly 2^n subsets".

Base. If $n = 0$ then $X = \varnothing$, whose only subset is $\varnothing$ itself, so $\mathcal{P}(\varnothing) = \{\varnothing\}$ and $|\mathcal{P}(X)| = 1 = 2^0$.

Step. Assume $P(n)$ and let $|X| = n + 1$. Since $n + 1 \geq 1$ the set is non-empty; pick $a \in X$ and put $Y := X \setminus \{a\}$, so $|Y| = n$. Split the subsets of X according to whether they contain a:

$$\mathcal{A} := \{A \in \mathcal{P}(X) \mid a \notin A\} , \qquad \mathcal{B} := \{A \in \mathcal{P}(X) \mid a \in A\} .$$

These are disjoint by Equation (3.26) and exhaust $\mathcal{P}(X)$ by Equation (3.25). Now $\mathcal{A}$ is exactly $\mathcal{P}(Y)$, since a subset of X omitting a is a subset of Y and conversely; hence $|\mathcal{A}| = 2^n$ by the induction hypothesis. And the map $B \longmapsto B \setminus \{a\}$ sends $\mathcal{B}$ into $\mathcal{P}(Y)$ with inverse $A \longmapsto A \cup \{a\}$, so it is a bijection (Definition 3.47) and $|\mathcal{B}| = 2^n$ as well. Adding the two disjoint families,

$$|\mathcal{P}(X)| = 2^n + 2^n = 2^{n+1} ,$$

which is $P(n + 1)$. By induction $P(n)$ holds for every n. □

The exponential notation is not a coincidence of counting. Sending a subset $A \subset X$ to its indicator function — the map taking the value 1 on A and 0 elsewhere — is a bijection between $\mathcal{P}(X)$ and the set $\mathcal{F}(X, \{0, 1\})$ of Definition 3.52, for finite and infinite X alike. This is why the power set is also written 2^X, and why Theorem 3.71 below can be read as the statement that there are always more functions on a set than points in it.

3.3 Maps

3.3.1 Maps

Consider two non-empty sets X and Y. In the everyday sense we can always relate some element of X with some element of Y by means of an arbitrary rule: some elements of X may be related to several elements of Y, others to none. The relations that turn out to be most useful are the more disciplined ones, in which *each* element of X is related to *one and only one* element of Y: we demand both the existence of a related element in Y and its uniqueness.

Definition 3.43 (Map)**.** A *map* (also *function* or *application*) f from X to Y is an assignment rule relating the elements of X with elements of Y that satisfies the axiom of existence and uniqueness:

$$\forall\, x \in X,\ \exists!\ f(x) \in Y \mid \text{to } x \text{ we assign } f(x)\,, \tag{3.56}$$

written in notation $x \longmapsto f(x)$. To make the related sets explicit we write $f : X \longrightarrow Y$, and we call X the *domain* of f and Y the *codomain* of f.

Rests on Definitions 3.24 and 3.28.

We may picture a map as a "machine" to which we feed an element of the domain, and which transforms it into an element of the codomain. For example, the map f assigning to each natural number its square root is

$$\begin{array}{rccc} f: & \mathbb{N} & \longrightarrow & \mathbb{R} \\ & n & \longmapsto & \sqrt{n}\,. \end{array}$$

Definition 3.44 (Image)**.** Let $f : X \longrightarrow Y$ be a map. The *image* of f is

$$f(X) = \{y \in Y \mid y = f(x),\ x \in X\} \subset Y\,. \tag{3.57}$$

Not every element of the codomain need be the result of mapping some element of the domain. Note that the image of a map is defined canonically: a map brings its image with it, uniquely determined, which is not the case for the codomain. One must not confuse this with two distinct maps sharing the same assignment formula: the image of

$$\begin{array}{rccc} f: & \mathbb{N} & \longrightarrow & \mathbb{R} \\ & n & \longmapsto & \sqrt{n} \end{array}$$

is different from the image of

$$\begin{array}{rccc} g: & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & \sqrt{x}\,. \end{array}$$

Definition 3.45 (Injective map)**.** A map $f : X \longrightarrow Y$ is *injective* if and only if

$$x \neq x' \implies f(x) \neq f(x')\,. \tag{3.58}$$

Injectivity means that two distinct elements of the domain are mapped to distinct elements of the codomain. The definition is equivalent to: $f : X \longrightarrow Y$ is injective if and only if

$$f(x) = f(x') \implies x = x'\,, \tag{3.59}$$

since the two implications are contrapositives of one another.

Definition 3.46 (Surjective map)**.** A map $f : X \longrightarrow Y$ is *surjective* if and only if

$$f(X) = Y\,, \tag{3.60}$$

which is equivalent to saying that

$$\forall\, y \in Y,\ \exists\, x \in X \mid y = f(x)\,. \tag{3.61}$$

Definition 3.47 (Bijective map)**.** A map $f : X \longrightarrow Y$ is *bijective*, or a *bijection*, if and only if

$$\forall\, y \in Y,\ \exists!\, x \in X \mid y = f(x)\,. \tag{3.62}$$

Bijectivity sharpens surjectivity by demanding uniqueness in the existence. A more practical characterization is the following.

Proposition 3.48 (Characterization of bijectivity)**.** *A map $f : X \longrightarrow Y$ is bijective if and only if it is injective and surjective.*

Rests on Definitions 3.45 to 3.47.

Proof. *Derives Proposition 3.48.*

Both directions unpack Definition 3.47, whose content is that each $y \in Y$ has an antecedent and has only one.

($\Longrightarrow$) Let f be bijective. The existential half of $\exists!\, x \in X \mid y = f(x)$ says that every $y \in Y$ is attained, i.e. $f(X) = Y$, which is surjectivity. For injectivity, suppose $f(x) = f(x')$ and call this common value y. Then x and x' are both antecedents of y, and the uniqueness half forces $x = x'$.

($\Longleftarrow$) Let f be injective and surjective, and let $y \in Y$. Surjectivity supplies at least one $x \in X$ with $f(x) = y$. If $x' \in X$ also satisfies $f(x') = y$, then $f(x) = f(x')$ and injectivity gives $x = x'$. So the antecedent exists and is unique, which is $\exists!\, x \in X \mid y = f(x)$. □

Definition 3.49 (Composition map)**.** Let $f : X \longrightarrow Y$ and $g : Y \longrightarrow Z$ be maps. The *composition of g with f* is the map

$$\begin{aligned} g \circ f : X &\longrightarrow Z \\ x &\longmapsto g(f(x))\,. \end{aligned}$$

Definition 3.50 (Identity map)**.** The *identity map* on X is

$$\begin{aligned} \mathrm{id}_X : X &\longrightarrow X \\ x &\longmapsto x\,. \end{aligned}$$

Definition 3.51 (Inverse map)**.** Let $f : X \longrightarrow Y$ be a bijection. The *inverse map* $f^{-1} : Y \longrightarrow X$ is the map assigning to each $y \in Y$ the unique $x \in X$ with $f(x) = y$; it satisfies $f^{-1} \circ f = \mathrm{id}_X$ and $f \circ f^{-1} = \mathrm{id}_Y$.

Rests on Definitions 3.47, 3.49 and 3.50.

Definition 3.52 (Set of maps)**.** We define $\mathcal{F}(X, Y)$ as the set of all maps of the form $f : X \longrightarrow Y$.

Definition 3.53 (Preimage)**.** Let $f : X \longrightarrow Y$ be a map and $B \subset Y$. The *preimage* of B under f is

$$f^{-1}(B) := \{x \in X \mid f(x) \in B\} \subset X\,. \tag{3.63}$$

The notation is a deliberate abuse: $f^{-1}(B)$ is defined for every map, bijective or not, and is a set, whereas the inverse map of Definition 3.51 exists only for bijections and returns an element. The two agree, up to the identification of a point with the singleton containing it, exactly when f is a bijection.

Proposition 3.54 (Preimages respect the set operations)**.** *Let $f : X \longrightarrow Y$ and let $B, C \subset Y$. Then*

$$f^{-1}(B \cup C) = f^{-1}(B) \cup f^{-1}(C)\,, \tag{3.64}$$

$$f^{-1}(B \cap C) = f^{-1}(B) \cap f^{-1}(C)\,, \tag{3.65}$$

$$f^{-1}(B^{\mathrm{c}}) = f^{-1}(B)^{\mathrm{c}}\,. \tag{3.66}$$

Images respect unions, $f(A \cup A') = f(A) \cup f(A')$, but in general only $f(A \cap A') \subset f(A) \cap f(A')$.

Rests on Definitions 3.35 and 3.53 and Equation (3.47).

Proof. *Derives Proposition 3.54.*

For the first identity, $x \in f^{-1}(B \cup C)$ says $f(x) \in B \vee f(x) \in C$, which is $x \in f^{-1}(B) \vee x \in f^{-1}(C)$; the second and third replace $\vee$ by $\wedge$ and by $\neg$ in the same way. Each step is an equivalence of propositions, so Equation (3.47) applies. For images the corresponding step is not an equivalence: $y \in f(A \cap A')$ requires *one* x lying in both A and A', whereas $y \in f(A) \cap f(A')$ allows two different antecedents, one in each. A non-injective map exhibits the gap — take $f(x) = x^2$ on $\mathbb{R}$ with $A = \{1\}$ and $A' = \{-1\}$, for which $f(A \cap A') = \varnothing$ while $f(A) \cap f(A') = \{1\}$. □

This asymmetry is why preimages, not images, are the right notion in topology (continuity is "preimages of open sets are open", Chapter 6) and in measure theory (a random variable is measurable when preimages of intervals are measurable, Chapter 11).

Definition 3.55 (Restriction)**.** Let $f : X \longrightarrow Y$ and $A \subset X$. The *restriction* of f to A is the map $f|_A : A \longrightarrow Y$ with $f|_A(a) = f(a)$ for every $a \in A$.

Proposition 3.56 (Composition is associative)**.** *Let $f : X \longrightarrow Y$, $g : Y \longrightarrow Z$ and $h : Z \longrightarrow W$. Then*

$$h \circ (g \circ f) = (h \circ g) \circ f \,, \tag{3.67}$$

and $f \circ \mathrm{id}_X = f = \mathrm{id}_Y \circ f$.

Rests on Definitions 3.49 and 3.50.

Proof. *Derives Proposition 3.56.*

Two maps with the same domain and codomain are equal when they agree at every point. For $x \in X$, both sides of Equation (3.67) unfold to $h(g(f(x)))$ by Definition 3.49 applied twice; the bracketing only records the order in which the unfolding was performed, and the result does not depend on it. The identity laws are immediate from Definition 3.50. □

Associativity with a two-sided identity is precisely the axiom system of a monoid, and the bijections of a set onto itself — which by Proposition 3.57 are closed under composition and by Definition 3.51 under inversion — form a group, the symmetric group of the set. That is the bridge from this chapter to Chapter 4, and through it to every symmetry argument in the physical parts of this treatise.

Proposition 3.57 (Composition of bijective maps)**.** *Let $f : X \longrightarrow Y$ and $g : Y \longrightarrow Z$ be bijections. Then $g \circ f : X \longrightarrow Z$ is a bijection, and*

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1} \,. \tag{3.68}$$

Rests on Propositions 3.48 and 3.56 and Definition 3.51.

Proof. *Derives Proposition 3.57.*

By Proposition 3.48 it suffices to prove that $g \circ f$ is injective and surjective.

Injective. Suppose $(g \circ f)(x) = (g \circ f)(x')$, that is $g(f(x)) = g(f(x'))$. Since g is injective, $f(x) = f(x')$; since f is injective, $x = x'$.

Surjective. Let $z \in Z$. Since g is surjective there is $y \in Y$ with $g(y) = z$, and since f is surjective there is $x \in X$ with $f(x) = y$. Then $(g \circ f)(x) = g(f(x)) = g(y) = z$.

For Equation (3.68), compose and use associativity (Proposition 3.56):

$$(f^{-1} \circ g^{-1}) \circ (g \circ f) = f^{-1} \circ (g^{-1} \circ g) \circ f = f^{-1} \circ \mathrm{id}_Y \circ f = f^{-1} \circ f = \mathrm{id}_X \,,$$

and symmetrically $(g \circ f) \circ (f^{-1} \circ g^{-1}) = \mathrm{id}_Z$. A two-sided inverse of a bijection is unique — if u and v both invert w then $u = u \circ \mathrm{id} = u \circ (w \circ v) = (u \circ w) \circ v = \mathrm{id} \circ v = v$ — so $f^{-1} \circ g^{-1}$ is *the* inverse of $g \circ f$. □

3.4 Equivalence classes

3.4.1 Equivalence relations

Definition 3.58 (Equivalence relation)**.** Consider a set $X = \{x, y, z, \ldots\}$. A relation $\sim$ on X is an *equivalence relation* if and only if it satisfies

1. Reflexivity:
$$\forall x \in X,\ x \sim x\,. \tag{3.69}$$
2. Symmetry:
$$\forall x, y \in X,\ x \sim y \implies y \sim x\,. \tag{3.70}$$
3. Transitivity:
$$\forall x, y, z \in X,\ x \sim y \ \wedge\ y \sim z \implies x \sim z\,. \tag{3.71}$$

Two elements $x, y \in X$ with $x \sim y$ are said to be *equivalent*.

Definition 3.59 (Equivalence class)**.** Consider a set X with an equivalence relation $\sim$, and let $a \in X$. The set of all elements of X equivalent to a,
$$[a] = \{x \in X \mid x \sim a\}\,, \tag{3.72}$$
is the *equivalence class*[1] of a, and a is called the *representative* of the class $[a]$.
Rests on Definition 3.58.

By reflexivity, $a \in [a]$. Moreover, let $b \in [a]$; then $b \sim a$ and, by symmetry, $a \sim b$. Now let $x \in [a]$; then $x \sim a$ and, by transitivity, $x \sim b$. Since x is arbitrary in $[a]$, every element of $[a]$ is equivalent to b, and we conclude that any element of $[a]$ may serve as its representative.

Two examples:

- On the set $\mathbb{R}$, define the relation
$$x \sim y \iff \exists\, n \in \mathbb{Z} \mid y = x + 2\pi n, \qquad x, y \in \mathbb{R}\,,$$
which is indeed an equivalence relation:
  1. $x \sim x$, since there exists $0 \in \mathbb{Z}$ such that
  $$x = x + 2\pi \cdot 0\,.$$
  2. Suppose $x \sim y$. Then there exists $n \in \mathbb{Z}$ such that
  $$\begin{aligned} & y = x + 2\pi n \\ \implies\ & x = y - 2\pi n \\ \implies\ & x = y + 2\pi m, \quad m \in \mathbb{Z}\,, \end{aligned}$$
  hence $y \sim x$ and symmetry holds.
  3. Suppose $x \sim y$ and $y \sim z$. Then there exist $n, m \in \mathbb{Z}$ such that
  $$y = x + 2\pi n, \quad z = y + 2\pi m\,.$$
  Adding the equations gives
  $$z = x + 2\pi(n + m), \quad n + m \in \mathbb{Z}\,,$$
  hence $x \sim z$ and transitivity holds.

[1] Equivalence classes are colloquially called simply *classes*.

One says that x and y are *congruent modulo* 2π; in particular 0 and 2π are congruent, i.e. they belong to the same equivalence class. If you know the trigonometric functions, this is surely something you had always intuited.

- In a non-leap year of the Gregorian calendar, define the relation

$$x \sim y \iff x \text{ has their birthday on the same day as } y\,,$$

which is an equivalence relation, very direct to verify. There are clearly 365 equivalence classes.

3.4.2 Quotient space

Theorem 3.60 (Equivalence classes partition the set)**.** *Let $\sim$ be an equivalence relation on a set X. Two equivalence classes are either disjoint or identical: for any classes $[a]$ and $[b]$,*

$$[a] \cap [b] \neq \varnothing \implies [a] = [b]\,, \tag{3.73}$$

or, equivalently, in view of the logical property Equation (3.14),

$$[a] \cap [b] = \varnothing \quad \vee \quad [a] = [b]\,. \tag{3.74}$$

The relation $\sim$ therefore divides X into disjoint classes: it partitions *the set.*
Rests on Definitions 3.36, 3.58 and 3.59 and Equation (3.47).

Proof. *Derives Theorem 3.60.*
Suppose the classes $[a]$ and $[b]$ are not disjoint,

$$[a] \cap [b] \neq \varnothing\,.$$

We must prove $[a] = [b]$: if they share at least one element, they are the same class. Recall that by definition

$$[a] = \{x \in X \mid x \sim a\}\,, \quad [b] = \{x \in X \mid x \sim b\}\ .$$

By reflexivity $[a] \neq \varnothing$, since $a \sim a$, and analogously $[b] \neq \varnothing$; there is thus no obstruction to considering $[a] \cap [b] \neq \varnothing$ (we make this remark because of property Equation (3.48) of set theory).

Since $[a] \cap [b] \neq \varnothing$, there exists $c \in [a] \cap [b]$ satisfying $c \sim a$ and $c \sim b$. By symmetry $a \sim c$, and finally by transitivity $a \sim b$.

Let $x \in [a]$; then $x \sim a$. Since also $a \sim b$, transitivity gives $x \sim b$, which means $x \in [b]$ for all $x \in [a]$, i.e. $[a] \subset [b]$. Analogously, starting from $y \in [b]$ we conclude $[b] \subset [a]$. By double inclusion, Equation (3.47), we conclude $[a] = [b]$. □

Definition 3.61 (Quotient space)**.** Consider a set X and an equivalence relation $\sim$ among its elements. The *quotient space* is

$$\frac{X}{\sim} = \{[y] \mid y \in X\}\ , \tag{3.75}$$

that is, the quotient space is the set of all equivalence classes given by $\sim$.
Rests on Definition 3.59.

Quotienting is how physics disposes of descriptive surplus. Angles are real numbers modulo 2π, the example above. States in quantum mechanics are vectors of a Hilbert space modulo nonzero complex scaling, which is why the state space is the projective space $\mathcal{P}(\mathcal{H})$ and why the global phase of a wavefunction is unobservable (Chapter 77). Gauge fields related by a gauge transformation are the same physical configuration, and the physical configuration space is the quotient by the gauge group (Chapter 64). In each case the equivalence relation encodes exactly what no measurement can distinguish, and Theorem 3.60 guarantees that this notion of "same physical situation" is consistent: the classes never partly overlap.

3.5 The infinite

Cardinality was defined in Section 3.2.4 by counting, which works only for finite sets. Cantor's insight was that comparison does not require counting: two shepherds can agree that they have equally many sheep by pairing the animals off, without either of them knowing how to count. The pairing is a bijection, and it makes sense whether the flocks are finite or not [188].

3.5.1 Equipotence and countability

Definition 3.62 (Equipotence)**.** Two sets X and Y are *equipotent*, written $X \approx Y$, if and only if there exists a bijection $f : X \longrightarrow Y$. We then say they have the same *cardinal*, $|X| = |Y|$.
Rests on Definitions 3.40 and 3.47.

Proposition 3.63. *Equipotence is reflexive, symmetric and transitive.*
Rests on Definitions 3.50, 3.51 and 3.62 and Proposition 3.57.

Proof. *Derives Proposition 3.63.*
Reflexivity: id_X is a bijection of X onto itself. Symmetry: if $f : X \longrightarrow Y$ is a bijection then so is $f^{-1} : Y \longrightarrow X$, by Definition 3.51. Transitivity: if $f : X \longrightarrow Y$ and $g : Y \longrightarrow Z$ are bijections then so is $g \circ f$, by Proposition 3.57. □

Remark 3.64. Proposition 3.63 lists the three clauses of Definition 3.58, but it stops short of calling $\approx$ an equivalence relation, and the omission is deliberate. An equivalence relation lives on a *set*, and there is no set of all sets to put this one on (Section 3.6). The classes of $\approx$ are therefore proper classes, not elements of a quotient set. Cardinals are recovered inside ZFC by choosing a canonical representative of each class rather than by forming the quotient.

Definition 3.65 (Countable set)**.** A set is *countably infinite* if and only if it is equipotent to $\mathbb{N}$, and *at most countable* (or simply *countable*) if it is finite or countably infinite. A set that is neither is *uncountable*. The cardinal of $\mathbb{N}$ is denoted $\aleph_0$.
Rests on Definition 3.62 and Axiom 3.33.

Being countably infinite means precisely that the elements can be arranged in an infinite list $x_1, x_2, x_3, \ldots$ in which every element appears exactly once: the list *is* the bijection. The natural numbers are trivially countable, via $n \longmapsto n$. Three less obvious examples follow, and each is a warning that infinite sets do not behave like finite ones.

Example 3.66 (The integers are countable)**.** Although $\mathbb{N} \subset \mathbb{Z}$ properly — $\mathbb{Z}$ contains 0 and all the negatives besides — the two sets are equipotent. Interleave the signs:

$$f : \mathbb{N} \longrightarrow \mathbb{Z}\,, \qquad f(1) = 0\,, \qquad f(2k) = k\,, \qquad f(2k+1) = -k \quad (k \in \mathbb{N})\,, \tag{3.76}$$

which lists $\mathbb{Z}$ as $0, 1, -1, 2, -2, 3, -3, \ldots$. Every integer occurs, and occurs once, so f is a bijection and $|\mathbb{Z}| = \aleph_0$.

That a proper subset can be equipotent to the whole is impossible for finite sets and characteristic of infinite ones — indeed Dedekind took it as the *definition* of infinite [237]. It is the content of the parable of Hilbert's hotel: a hotel with rooms $1, 2, 3, \ldots$, all occupied, still accommodates a new guest by moving the occupant of room n to room $n + 1$.

Proposition 3.67 (Countability of the plane of naturals)**.** $\mathbb{N} \times \mathbb{N}$ *is countably infinite.*
Rests on Definitions 3.38, 3.47 and 3.65.

Proof. *Derives Proposition 3.67.*

Enumerate the pairs along the finite antidiagonals $m+n=2,3,4,\ldots$, each of which contains finitely many pairs, taking them in order of increasing m. Explicitly, the Cantor pairing map

$$\pi : \mathbb{N}\times\mathbb{N} \longrightarrow \mathbb{N}\,, \qquad \pi(m,n) = \frac{(m+n-2)(m+n-1)}{2} + m \tag{3.77}$$

sends the antidiagonal $m+n=d$ onto the block of integers $\{T_{d-2}+1,\ldots,T_{d-2}+(d-1)\}$, where $T_k = k(k+1)/2$ is the k-th triangular number and $d-1$ is the number of pairs on that antidiagonal. Consecutive blocks abut without gap or overlap, since $T_{d-2}+(d-1)=T_{d-1}$, and within a block m runs over $1,\ldots,d-1$ injectively. Hence π is a bijection. □

Corollary 3.68 (The rationals are countable)**.** $|\mathbb{Q}| = \aleph_0$.
Rests on Proposition 3.67, Definition 3.65, and Equation (3.76).

Proof. *Derives Corollary 3.68.*

Each positive rational is n/m with $n,m\in\mathbb{N}$, so $(n,m)\longmapsto n/m$ maps $\mathbb{N}\times\mathbb{N}$ onto the positive rationals. A surjection from a countable set makes the target at most countable: list the source by Proposition 3.67 and strike out every term whose value has already appeared, leaving a list of the positive rationals without repetition. That list is infinite because $\mathbb{N}$ embeds in it, so the positive rationals are countably infinite. Interleaving with the negatives and 0 as in Equation (3.76) gives $|\mathbb{Q}| = \aleph_0$. □

So the rationals, dense though they are — between any two of them lies another — are no more numerous than the integers. At this point one might guess that all infinite sets are equipotent, and that "infinite" admits no useful gradation. Cantor showed otherwise.

3.5.2 The diagonal argument

Theorem 3.69 (Uncountability of the continuum)**.** *The set $\mathbb{R}$ is uncountable. More precisely, no map $f:\mathbb{N}\longrightarrow[0,1)$ is surjective [249].*
Rests on Definitions 3.46 and 3.65.

Proof. *Derives Theorem 3.69.*

Let $f:\mathbb{N}\longrightarrow[0,1)$ be any map. Each value has a decimal expansion

$$f(n) = 0.a_{n1}a_{n2}a_{n3}\ldots\,, \qquad a_{nk}\in\{0,1,\ldots,9\}\,,$$

and where a number admits two expansions — one ending in all 0s, one in all 9s — we fix the first, so that the digits a_{nk} are well defined functions of n and k.

Now build a number that differs from every listed one, using the diagonal digits a_{nn}. Define

$$b_n := \begin{cases} 5 & \text{if } a_{nn}\neq 5\,,\\ 4 & \text{if } a_{nn}=5\,, \end{cases} \tag{3.78}$$

and set $b := 0.b_1b_2b_3\ldots$. Since every digit of b is 4 or 5, we have $b\in[0,1)$; moreover b's expansion neither terminates in all 0s nor in all 9s, so it is the *only* expansion of b and the digits of b are exactly the b_n.

Suppose $b=f(n)$ for some $n\in\mathbb{N}$. Comparing the n-th digits of the two (unique, on our convention) expansions gives $b_n = a_{nn}$, contradicting Equation (3.78), which was rigged so that $b_n\neq a_{nn}$. Hence $b\neq f(n)$ for every n, and f is not surjective.

Since no map $\mathbb{N}\longrightarrow[0,1)$ is surjective, none is bijective, so $[0,1)$ is uncountable; and $[0,1)\subset\mathbb{R}$, so $\mathbb{R}$ is uncountable too — an at most countable set has no uncountable subset. □

Corollary 3.70. *The irrational numbers are uncountable, and in this precise sense almost every real number is irrational.*
Rests on Theorem 3.69 and Corollary 3.68.

Proof. *Derives Corollary 3.70.*
If $\mathbb{R} \setminus \mathbb{Q}$ were at most countable then $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$ would be the union of two at most countable sets, hence at most countable — interleave the two lists — contradicting Theorem 3.69. □

The pattern of the proof, not its arithmetic, is what matters. A list is given; a new object is built by disagreeing with the n-th listed object in its n-th feature; the new object cannot be on the list. This *diagonal argument* reappears at every turn in the rest of the chapter: in Cantor's theorem immediately below, in Russell's paradox (Theorem 3.75), in Gödel's undecidable sentence (Theorem 3.87), and in Turing's halting problem (Theorem 3.96). They are four faces of one idea.

Theorem 3.71 (Cantor)**.** *For every set X there is no surjection $f : X \longrightarrow \mathcal{P}(X)$. Consequently $|X| < |\mathcal{P}(X)|$, and there is no largest cardinal.*
Rests on Definitions 3.41 and 3.46, Axiom 3.76, and Equation (3.26).

Proof. *Derives Theorem 3.71.*
Let $f : X \longrightarrow \mathcal{P}(X)$ be any map. Each $x \in X$ is thereby assigned a subset $f(x) \subset X$, and it makes sense to ask whether x belongs to the subset assigned to it. Collect the elements for which the answer is no:

$$D := \{x \in X \mid x \notin f(x)\} \ . \tag{3.79}$$

This is a legitimate subset of X — it is carved out of the existing set X by a predicate, as the axiom of separation requires (Axiom 3.76) — so $D \in \mathcal{P}(X)$.

Suppose $D = f(d)$ for some $d \in X$. Ask whether $d \in D$. By the definition Equation (3.79), $d \in D$ holds if and only if $d \notin f(d)$; but $f(d) = D$, so $d \in D$ holds if and only if $d \notin D$. That is $p \leftrightarrow \neg p$, a contradiction by Equation (3.26). Hence D is in the image of no element and f is not surjective.

Since $x \longmapsto \{x\}$ is an injection $X \longrightarrow \mathcal{P}(X)$ while no surjection exists, the inequality is strict. Applying the theorem to $\mathcal{P}(X)$, then to $\mathcal{P}(\mathcal{P}(X))$, and so on produces an unending strictly increasing chain of cardinals. □

Theorem 3.71 generalizes Proposition 3.42 from finite sets to all sets, and it strengthens Theorem 3.69: taking $X = \mathbb{N}$ gives an uncountable set directly, and one can show $|\mathcal{P}(\mathbb{N})| = |\mathbb{R}|$, the cardinal of the continuum, written $\mathfrak{c} = 2^{\aleph_0}$. Comparison of cardinals is made rigorous by the following theorem, whose proof is long enough to belong in the appendix.

Theorem 3.72 (Cantor–Schröder–Bernstein)**.** *If there exist injections $f : X \longrightarrow Y$ and $g : Y \longrightarrow X$, then there exists a bijection between X and Y. Hence $|X| \leq |Y|$ and $|Y| \leq |X|$ imply $|X| = |Y|$.*
Rests on Definitions 3.45, 3.47 and 3.62.

Full derivation: Appendix A.3 (p. 2773).
Derives Theorem 3.72.

3.5.3 The continuum hypothesis

Cantor found $\aleph_0 = |\mathbb{N}|$ and $\mathfrak{c} = |\mathbb{R}|$ with $\aleph_0 < \mathfrak{c}$, and could not settle the obvious next question: is there anything strictly in between?

Axiom 3.73 (Continuum hypothesis)**.** There is no set X with $\aleph_0 < |X| < \mathfrak{c}$; equivalently, every uncountable subset of $\mathbb{R}$ is equipotent to $\mathbb{R}$, i.e. $\mathfrak{c} = \aleph_1$, the least uncountable cardinal.

Hilbert placed the question first on his list of twenty-three problems for the new century [291], and its resolution took two thirds of that century and came in two halves, neither of them an answer in the sense Hilbert expected.

1. Gödel showed in 1938 that if ZF is consistent, then so is ZFC together with the continuum hypothesis [815]. He did it by constructing the universe L of *constructible sets*, an inner model of ZF in which both the axiom of choice and Axiom 3.73 come out true. So Axiom 3.73 cannot be refuted from the axioms.

2. Cohen showed in 1963 that if ZF is consistent, then so is ZFC together with the *negation* of the continuum hypothesis [1249]. His method, *forcing*, adjoins to a model new sets constructed by finite approximations, in such a way that the cardinal $\mathfrak{c}$ can be pushed above $\aleph_1$ while every axiom of ZFC survives. So Axiom 3.73 cannot be proved from the axioms either.

Together: the continuum hypothesis is *independent* of ZFC. It is not that we do not yet know whether it is true; it is that the axioms we use for mathematics do not decide it, as the parallel postulate is not decided by the other axioms of geometry. One may adjoin it or its negation and get consistent mathematics either way.

Remark 3.74 (Does the physicist care?). Directly, no — and it is worth being clear why not. Physical predictions are statements about measured quantities with finite precision, and the mathematics that generates them lives, almost without exception, in separable spaces: the Hilbert spaces of quantum mechanics have countable orthonormal bases (Chapter 12), fields are described by functions determined by their values on a countable dense set, and every experimental record is a finite list of rationals with uncertainties. No measurement can distinguish a world in which Axiom 3.73 holds from one in which it fails, because nothing observable depends on the cardinal arithmetic strictly between $\aleph_0$ and $\mathfrak{c}$.

Indirectly, it matters a great deal, and this is the reason the section is here. The continuum hypothesis is the first proof that a perfectly meaningful mathematical question can be immune to the axioms used to pose it. Physics inherits mathematics wholesale; it is entitled to know that its inherited language has questions it cannot answer, and where the boundary runs. The rest of this chapter maps that boundary.

3.6 The crisis of foundations

Definition 3.28 says a set is a grouping of elements satisfying the same logical propositions. Taken at face value this is the principle of *unrestricted comprehension*: every predicate P determines a set $\{x \mid P(x)\}$. It is the natural reading; it is, in the guise of Basic Law V on value-ranges, what Frege built his reduction of arithmetic to logic upon [257]; and it is false.

3.6.1 Russell's paradox

Theorem 3.75 (Russell, 1901)**.** *Unrestricted comprehension is inconsistent.*
Rests on Definitions 3.7, 3.12 and 3.28.

Proof. *Derives Theorem 3.75.*

Sets can be elements of sets, so for any set x the sentence $x \in x$ is a well formed proposition and $x \notin x$ is its negation. Apply unrestricted comprehension to the predicate $P(x) =$ "$x \notin x$" and let

$$R := \{x \mid x \notin x\} \,. \tag{3.80}$$

Ask whether $R \in R$. By Equation (3.80), an object belongs to R if and only if it does not belong to itself; applied to R itself this reads

$$R \in R \iff R \notin R \,, \tag{3.81}$$

which is $p \leftrightarrow \neg p$. Both truth values of p make Equation (3.81) false, while comprehension asserts it. The principle therefore yields a contradiction. □

Russell communicated the argument to Frege in June 1902, while the second volume of Frege's *Grundgesetze* was at the printer; Frege appended a note beginning "Hardly anything more unwelcome can befall a scientific writer than that one of the foundations of his edifice be shaken after the work is finished". The exchange is reprinted in [1349].

Compare Equation (3.80) with the set D of Theorem 3.71. They are the same construction: Russell's paradox is Cantor's diagonal argument applied to the map that sends each set to itself. Cantor's theorem is a theorem because D is carved out of an existing set X; Russell's is a paradox because R is carved out of nothing at all. That distinction is the entire repair.

3.6.2 The Zermelo–Fraenkel axioms

Zermelo's response [350], completed by Fraenkel and Skolem [488], was to abandon comprehension as a general principle and to legislate instead exactly which set formations are permitted. Sets are then not *defined* at all: like "point" in geometry or "event" in relativity, "set" and "$\in$" are primitive notions (Section 1.5), constrained only by axioms.

Axiom 3.76 (Zermelo–Fraenkel with choice)**.** The theory ZFC is written in first-order logic with the single non-logical symbol $\in$, and consists of:

1. **Extensionality.** Sets with the same elements are equal.
2. **Pairing.** For all a, b there is a set $\{a, b\}$.
3. **Union.** For every set A there is a set $\bigcup A$ whose elements are the elements of the elements of A.
4. **Power set.** For every set A there is a set $\mathcal{P}(A)$ whose elements are the subsets of A.
5. **Infinity.** There is a set containing $\varnothing$ and closed under $x \longmapsto x \cup \{x\}$ — an infinite set exists, and this is where $\mathbb{N}$ comes from.
6. **Separation** (an axiom for each predicate P). For every set A, the collection $\{x \in A \mid P(x)\}$ is a set.
7. **Replacement** (an axiom for each definable function F). The image $\{F(x) \mid x \in A\}$ of a set is a set.
8. **Foundation.** Every non-empty set has an element disjoint from it; equivalently, there is no infinite descending chain $\cdots \in x_2 \in x_1 \in x_0$.
9. **Choice.** For every set of non-empty sets there is a function selecting one element from each.

Axioms 1–8 constitute ZF; adjoining 9 gives ZFC. Items 6 and 7 are *schemata*: infinitely many axioms, one per predicate, since first-order logic cannot quantify over predicates.

Separation is the repair. It permits $\{x \in A \mid P(x)\}$ but never $\{x \mid P(x)\}$: a set may only be sieved out of a set already known to exist. Russell's R is unobtainable, since no prior set is available to sieve. What survives of his argument is a theorem.

Corollary 3.77 (There is no set of all sets)**.** *For every set A there is a set that is not an element of A.*

Rests on Axiom 3.76.

Proof. *Derives Corollary 3.77.*

Let A be a set. By separation, $B := \{x \in A \mid x \notin x\}$ is a set. Suppose $B \in A$. Then $B \in B$ holds if and only if $B \in A$ and $B \notin B$, and since $B \in A$ was assumed, this reduces to $B \in B \iff B \notin B$, a contradiction. Hence $B \notin A$, so A fails to contain the set B. $\square$

The universe of sets is thus not itself a set — it is a *proper class*, and Remark 3.64 was an instance of the same fact. Foundation goes further and rules out $x \in x$ outright, so that the sets form a cumulative hierarchy built up from $\varnothing$ by iterated power sets. Nothing in physics ever reaches beyond the first few dozen levels of that hierarchy.

3.6.3 The axiom of choice

The ninth axiom is different in kind from the other eight. Each of those gives an explicit recipe for the set it asserts; choice asserts that a selection exists without providing any rule for making it. For finitely many non-empty sets the selection can be made one at a time and choice is a theorem; for infinitely many it is a genuine assumption.

Axiom 3.78 (Choice)**.** Let $\{X_i\}_{i \in I}$ be a family of non-empty sets indexed by any set I. Then there exists a map

$$c : I \longrightarrow \bigcup_{i \in I} X_i \quad \text{with} \quad c(i) \in X_i \ \text{ for every } i \in I\,. \tag{3.82}$$

Choice is equivalent, over ZF, to a list of statements that look nothing like it and are used constantly: Zermelo's well-ordering theorem (every set admits a well ordering), Zorn's lemma (a partially ordered set in which every chain has an upper bound has a maximal element), the statement that every vector space has a basis (Chapter 5), Tychonoff's theorem on products of compact spaces (Chapter 6), and the claim that every surjection has a right inverse.

Its cost is that it produces objects no construction could reach. Vitali's non-measurable subset of $\mathbb{R}$ and the Banach–Tarski decomposition — a solid ball cut into five pieces and reassembled, by rigid motions alone, into two balls of the original size — both follow from choice, and both are immune to physical interpretation precisely because the pieces have no volume. The lesson for physics is not that choice is false but that measure-theoretic hygiene is mandatory: probability is defined on a σ-algebra of measurable sets, and the pathological sets are simply not in it (Chapter 11).

Remark 3.79 (How much choice does physics use?)*.* Far less than the full axiom. The Hilbert spaces of quantum mechanics are separable, and constructing a countable orthonormal basis from a countable dense set by Gram–Schmidt needs only *dependent choice*, a weak fragment consistent with ZF plus "every set of reals is measurable". Where full choice does enter — the Hahn–Banach theorem, the existence of a basis for a non-separable space, the spectral theorem in its most general form (Chapter 12) — the applications physics makes of the result can invariably be recovered constructively in the separable case. No experimental prediction has ever depended on the axiom of choice.

3.6.4 Three answers to the crisis

The paradoxes forced a question that mathematics had never needed to ask: what are its statements *about*, and what makes a proof valid? Three programmes answered differently, and their disagreement is the direct ancestor of the results in the remainder of this chapter.

1. **Logicism** (Frege, Russell, Whitehead). Mathematics is a branch of logic: numbers are definable in purely logical terms and arithmetic truths are logical truths. *Principia Mathematica* [365] carried the reduction out in three volumes: the cardinal groundwork

for $1 + 1 = 2$ is proposition *54.43, midway through the first volume, and the arithmetical proposition itself is *110.643, on page 86 of the second. The paradoxes were blocked with a ramified theory of types forbidding a collection to contain itself. The programme foundered on its own repairs: the axioms of infinity and of reducibility, needed to make the development work, are not plausibly *logical* truths, so the reduction relocated the problem rather than dissolving it.

2. **Formalism** (Hilbert). Mathematics is the manipulation of finite strings of symbols according to stated rules; questions of what the symbols denote are set aside. Infinite objects are "ideal elements", admissible so long as reasoning with them never yields a false statement about finite ones [546]. Hilbert's programme is the demand that this be *proved*, by finitary means, once and for all. It is the subject of Sections 3.8 and 3.10.

3. **Intuitionism** (Brouwer). Mathematics is a mental construction, and a mathematical object exists only when it has been constructed [401]. On this view $\forall x\, P(x)$ asserts a method producing a proof of $P(x)$ for each x, and $\exists x\, P(x)$ asserts a construction exhibiting a witness. The law of the excluded middle Equation (3.25) then fails for infinite domains: to assert $p \vee \neg p$ over $\mathbb{N}$ is to claim a decision procedure one may not possess. Proof by contradiction survives for refutations but not for existence claims — an intuitionist rejects a proof that some x exists which merely shows that the non-existence of x is absurd.

This treatise reasons classically, uses ZFC without apology, and states the fact rather than assuming it. But the intuitionist objection has an afterlife the physicist should recognize: a constructive existence proof yields an algorithm, and a classical one need not. When a physical theory guarantees that a solution, a ground state, or an equilibrium *exists*, only the constructive version tells anyone how to compute it — and Section 3.9 shows that the gap between the two is not always bridgeable, even in principle. Bishop's constructive redevelopment of analysis [1341] shows how much of the working mathematics of physics can be had on those terms, which is more than the classical mathematician expects and less than the physicist would like.

3.7 The analytical perspective

The same decades that produced the paradoxes produced a movement in philosophy that took the new logic as its instrument: analytical philosophy. Its founding conviction was that philosophical problems are to be attacked by analysing the logical structure of the language in which they are posed — and its central case study was physics. This treatise's first editorial rule, that only physics with observational evidence is in scope, is a philosophical commitment, not a mathematical one. It deserves an argument rather than an assertion, and the argument was worked out here.

3.7.1 The linguistic turn

Frege's *Begriffsschrift* [207] introduced quantifiers, variables and a rigorous notion of proof, replacing two millennia of syllogistic; Section 3.1.4 is his legacy. His *Grundlagen* [228] then asked what a number *is*, and answered by analysing the sentences numbers occur in rather than by introspecting about quantity. That reversal — study the statement, not the thing — is the linguistic turn.

Frege's distinction between the *sense* of an expression and its *reference* is the part a physicist uses daily. "The morning star" and "the evening star" refer to the same object,

Venus, but differ in sense, which is why their identity was a discovery and not a triviality. The pattern recurs throughout this book. Inertial mass and gravitational mass have entirely different senses — one is the resistance to acceleration in Chapter 19, the other the coupling to a gravitational field — and their identity is an empirical finding, tested to parts in 10^{15} and elevated to the equivalence principle (Chapter 42). Energy and mass, inertia and curvature, temperature and mean kinetic energy: each is an identity of reference across a difference of sense, each was an experimental result, and none is a definition.

Russell's theory of descriptions supplied the second tool. The sentence "the present King of France is bald" is not about a bald man; analysed, it asserts that exactly one thing kings France and that it is bald — a perfectly meaningful claim, and false, with no ghostly monarch required. The grammatical form of a sentence can thus misrepresent its logical form. This is precisely what happened to the aether: "the frame in which the aether is at rest" is a definite description whose presupposition fails, and special relativity did not find that frame moving oddly, it denied that the description picks anything out (Chapter 37). A physicist who keeps the distinction in view is protected against a whole family of pseudo-questions — "what is the wavefunction really made of", "where was the particle before measurement" — which may or may not have answers, but which certainly do not have them merely because the grammar of the question is well formed.

3.7.2 Logical positivism and its collapse

Wittgenstein's *Tractatus* [496] pushed the programme to its limit: propositions picture facts, the sense of a proposition is the state of affairs it depicts, and what cannot be depicted cannot be said. Its closing proposition — whereof one cannot speak, thereof one must be silent — was read by the Vienna Circle as a mandate. Their manifesto [624] proposed the *verification principle*: the meaning of a statement is the method of its verification, and a statement admitting no verification is not false but meaningless. Metaphysics was to be eliminated, not refuted.

For a while physics appeared to vindicate them, because physics had just done it. Einstein's 1905 analysis of simultaneity is the operationalist move in its purest form: instead of asking what simultaneity *is*, he asked what procedure establishes it, found that the procedure is frame-dependent, and derived the Lorentz transformations from that (Chapter 38). Heisenberg's 1925 matrix mechanics was motivated by the resolve to build a theory from observable transition frequencies and intensities alone, discarding unobservable electron orbits (Chapter 70). Two revolutions in twenty years, both achieved by discarding a quantity no experiment could reach.

The principle nonetheless failed, in three distinct ways, and the failures are as instructive as the successes.

1. **Self-refutation.** The verification principle is not itself verifiable by any observation, so by its own criterion it is meaningless. No repair commanded agreement.

2. **Holism.** Quine argued that no individual statement can be confronted with experience on its own [987]: a prediction follows only from a theory together with auxiliary hypotheses about apparatus, background and approximation, so a failed prediction refutes the conjunction and never says which conjunct to abandon. The analytic–synthetic distinction on which the positivists relied — truths of meaning versus truths of fact — does not survive the analysis either. Every anomalous result in this book illustrates the point: the anomalous perihelion of Mercury could in principle have been an unseen planet, and deciding otherwise took decades and independent evidence.

3. **Verification is the wrong relation.** A universal law has infinitely many instances and can never be verified by finitely many observations, but a single counterinstance

refutes it. Popper therefore replaced verifiability by *falsifiability* as the criterion of scientific status [1173]. This is nothing but Equation (3.17) taken seriously, as Section 3.1.3 already noted.

Kuhn added the historical objection [1233]: observations are theory-laden, the auxiliary hypotheses are supplied by the reigning framework, and scientists abandon a framework not when a refutation arrives but when a better framework does.

3.7.3 The position this treatise takes

The programme did not survive, but its residue is the working philosophy of modern physics, and it is worth stating explicitly what this book has adopted from the wreckage.

1. **Demarcation by testability, not by verifiability.** Editorial rule 1 admits a subject when observation could in principle decide against it. That is why the Standard Model is in and string theory, supersymmetry and speculative quantum gravity are named only in Chapter 135 and only to record that the evidence is absent.

2. **Operational meaning of quantities.** A physical quantity is individuated by the procedure that measures it and the unit that reports it, which is why the SI is treated as axiomatic here (Chapter 2) and why every experiment in this book carries apparatus, procedure and uncertainty.

3. **Holism, taken as a bookkeeping obligation.** Since no prediction confronts data alone, the auxiliary assumptions must be visible. That is the real function of editorial rule 2: a derivation is the complete list of what a prediction depends on, and a *pending derivation* is an admission that the list is not yet complete.

4. **No final truth.** Survival under attempted refutation is what the method certifies, never truth (Remark 1.2).

What this treatise does *not* do is settle the question the positivists thought they had dissolved: whether the unobservable structures of a successful theory — quantum states, gauge fields, spacetime curvature — are real, or are instruments for organizing observations. Where an interpretive dispute has observable consequences, it is an experimental question and is treated as one (Chapter 89); where it has none, this book reports the competing readings and stops (Chapter 85). That restraint is itself the analytical inheritance: not every well formed question has an answer, and one of the uses of logic is to tell which.

There is a last, sharper lesson, and the following sections prove it rather than argue it. The Vienna Circle's ambition was a language for science in which every meaningful question could be settled by inspection of the world plus computation. Within four years of the manifesto, Gödel had shown that no such language exists even for arithmetic.

3.8 Formal systems and incompleteness

Hilbert's answer to the paradoxes (Section 3.6.4) was to make mathematical reasoning itself a mathematical object. Strip the meaning from a proof and what remains is a finite sequence of finite strings, each obtained from earlier ones by stated rules. Such an object can be studied with the same rigour as a triangle, and the study of it is *metamathematics*.

3.8.1 Formal systems

Definition 3.80 (Formal system)**.** A *formal system* T consists of

1. a finite or countable *alphabet* of symbols;
2. a set of *formulas*: those finite strings over the alphabet that a purely syntactic rule declares well formed;
3. a set of *axioms*, distinguished formulas;
4. a finite set of *rules of inference*, each taking finitely many formulas to a formula.

A *derivation* of φ in T is a finite sequence of formulas ending in φ, each member of which is an axiom or follows from earlier members by a rule. We write $T \vdash \varphi$ — "T proves φ" — when a derivation exists, and $T \nvdash \varphi$ when none does.

Everything in Definition 3.80 is finite and mechanical. Whether a given string is a formula, whether a given formula is an axiom, whether a given sequence is a derivation: each is settled by inspection, without understanding what the symbols mean. This is the whole point. A formal system is a game with tokens, and its trustworthiness rests on the fact that a clerk who understands nothing can check a proof.

Definition 3.81 (Effective axiomatization)**.** A formal system is *effectively axiomatized* if there is a mechanical procedure that, given a formula, decides whether it is an axiom.
Rests on Definition 3.80.

"Mechanical procedure" is made precise in Section 3.9; until then the intuitive reading suffices. The condition excludes a cheap evasion of the theorems below — taking as axioms *all* true statements of arithmetic gives a complete theory, but no clerk could ever recognize its axioms, and it is therefore no use to anyone.

Definition 3.82 (Consistency, completeness, soundness)**.** Let T be a formal system in a language with negation.

1. T is *consistent* if there is no formula φ with both $T \vdash \varphi$ and $T \vdash \neg\varphi$.
2. T is *negation-complete* if for every sentence φ of its language, $T \vdash \varphi$ or $T \vdash \neg\varphi$: the system decides every question expressible in it.
3. T is *sound* with respect to an interpretation if every provable sentence is true under that interpretation.
4. T is *decidable* if there is a mechanical procedure that, given a sentence, determines whether $T \vdash \varphi$.

Rests on Definitions 3.7 and 3.80.

Consistency is indispensable, because an inconsistent system proves everything: from φ and $\neg\varphi$ the tautology $(\varphi \wedge \neg\varphi) \to \psi$ — vacuously true since its antecedent is false by Equation (3.26) — yields any ψ whatever by modus ponens. A theory that proves everything distinguishes nothing and says nothing.

Definition 3.83 (Peano arithmetic)**.** *Peano arithmetic*, denoted PA, is the first-order theory in the language $\{0, S, +, \cdot\}$ whose axioms are: S is injective and takes the value 0 nowhere; the recursion equations

$$n + 0 = n\,, \quad n + S(m) = S(n + m)\,, \qquad n \cdot 0 = 0\,, \quad n \cdot S(m) = n \cdot m + n\,, \tag{3.83}$$

which are those of Section 3.2.2 rebased at 0, as first-order arithmetic conventionally is; and the induction schema: for each formula φ of the language, the axiom

$$(\varphi(0) \wedge \forall n\,(\varphi(n) \to \varphi(Sn))) \to \forall n\,\varphi(n)\,. \tag{3.84}$$

Rests on Definition 3.80 and Axioms 3.33 and 3.34.

Equation (3.84) is weaker than Axiom 3.34. The axiom of Section 3.2.2 quantifies over *all* predicates on $\mathbb{N}$, of which there are uncountably many by Theorem 3.71; Equation (3.84) provides one axiom for each of the countably many formulas of a first-order language. That gap between second-order and first-order induction is exactly where incompleteness will enter.

3.8.2 Completeness — what logic does deliver

Before the negative results, the positive one, and it is easy to confuse them because both are called "completeness".

Theorem 3.84 (Gödel's completeness theorem, 1930)**.** *A sentence of first-order logic is derivable from a set of premises if and only if it is true in every structure satisfying those premises:*

$$T \vDash \varphi \iff T \vdash \varphi\,. \tag{3.85}$$

Equivalently, a consistent first-order theory has a model [654].

Rests on Definitions 3.80 and 3.82.

Full derivation: Appendix A.4 (p. 2775).

Derives Theorem 3.84.

This says that the proof calculus of first-order logic misses nothing: semantic consequence and syntactic derivability coincide. Note carefully what it is *about* — the logic, not any particular theory. It says that whatever follows from the axioms of a theory can be proved from them. It does not say the axioms settle every question, and the theorem of the next subsection says they do not.

Remark 3.85 (Two senses of "complete"). Theorem 3.84 concerns *semantic* completeness of a logic: everything true in all models is provable. Definition 3.82 defines *negation* completeness of a theory: every sentence or its negation is provable. Gödel proved the first in 1930 and refuted the second, for arithmetic, in 1931. Both results are his, they point in opposite directions, and mixing them up is the commonest error made about this subject.

3.8.3 Arithmetization

Gödel's decisive move was to make arithmetic talk about itself. Assign to every symbol a number, to every finite string of symbols the number coding its sequence of symbols, and to every finite sequence of strings — that is, every derivation — a number coding it. Any injective coding with mechanically computable encoding and decoding will do; the choice matters as little as the choice of Kuratowski pairs in Remark 3.39. Write $\ulcorner\varphi\urcorner$ for the numeral denoting the code of the formula φ: its *Gödel number*.

Syntactic properties become arithmetical properties under this coding. "n codes a well formed formula", "m codes a derivation of the formula coded by n" — these are conditions on numbers, decidable by finite computation, and for an effectively axiomatized theory they can be expressed by formulas of the language of arithmetic and proved of the right numbers inside PA. In particular there is a formula $\mathrm{Prov}_T(x)$ of one free variable, the *provability predicate*, such that

$$\mathrm{Prov}_T(x) \text{ expresses "}\exists m \text{ such that } m \text{ codes a } T\text{-derivation of the formula coded by } x\text{"}\,. \tag{3.86}$$

The construction is elementary but long; it is carried out in Appendix A.5.

Lemma 3.86 (Diagonal lemma)**.** *Let T be a theory in which the elementary syntactic functions are representable, and let $\psi(x)$ be any formula with one free variable. Then there is a sentence σ with*

$$T \vdash \ \sigma \leftrightarrow \psi(\ulcorner \sigma \urcorner)\,. \tag{3.87}$$

Rests on Definitions 3.80 and 3.83.

Full derivation: Appendix A.5 (p. 2780).

Derives Lemma 3.86.

The sentence σ asserts of its own Gödel number that it has the property ψ; informally, σ says "I am ψ". This is the diagonal argument of Section 3.5.2 once more, now applied to formulas indexed by their own codes, and it is the only sleight of hand in the entire subject. Nothing self-referential is smuggled in: σ is an ordinary arithmetical sentence about numbers, which happens to be true exactly when a certain number — its own code — has a certain computable property.

3.8.4 The first incompleteness theorem

Theorem 3.87 (Gödel, 1931)**.** *Let T be an effectively axiomatized, consistent theory that proves a modest amount of arithmetic — enough to represent the computable functions, which PA and ZFC amply do. Then there is a sentence G_T of the language of T such that*

$$T \nvdash G_T \qquad \text{and} \qquad T \nvdash \neg G_T\,. \tag{3.88}$$

T is therefore not negation-complete. Moreover G_T is true of the natural numbers [681, 786].

Rests on Lemma 3.86, Equation (3.86), and Definitions 3.81 and 3.82.

Proof. *Derives Theorem 3.87.*

Apply Lemma 3.86 to the formula $\psi(x) = \neg\mathrm{Prov}_T(x)$. This yields a sentence G_T with

$$T \vdash \ G_T \leftrightarrow \neg\mathrm{Prov}_T(\ulcorner G_T \urcorner)\,, \tag{3.89}$$

a sentence that, read through the coding, says "I am not provable in T".

T does not prove G_T. Suppose $T \vdash G_T$. Then some finite derivation of G_T exists; let m be its code. The statement "m codes a derivation of G_T" is a true statement about particular numbers, verified by a finite computation, and a theory representing the computable functions proves every such true statement. Hence $T \vdash \mathrm{Prov}_T(\ulcorner G_T \urcorner)$, and by Equation (3.89), $T \vdash \neg G_T$. So T proves both G_T and $\neg G_T$, contradicting consistency. Therefore $T \nvdash G_T$.

G_T is true. What the previous paragraph established is that no derivation of G_T exists. That is precisely what G_T asserts, so G_T is a true sentence about the natural numbers.

T does not prove $\neg G_T$. If T is sound — if it proves only truths — this is immediate: $\neg G_T$ is false, so T cannot prove it. Soundness is a stronger hypothesis than consistency, and Gödel replaced it by the syntactic condition of ω-consistency; Rosser later removed even that, at the cost of a subtler sentence, so that bare consistency suffices [786]. Both refinements are carried out in Appendix A.5. □

Note what the theorem does not permit: patching. Adjoin G_T to T as a new axiom and the resulting theory $T' = T + G_T$ is still effectively axiomatized, still consistent, still arithmetically adequate — so the theorem applies again and produces $G_{T'}$, undecided by T'. The gap cannot be closed by adding axioms, one at a time or infinitely many, so long as a clerk can recognize the axioms.

3.8.5 The second incompleteness theorem

Consistency is itself an arithmetical statement: T is consistent exactly when some particular sentence has no derivation, say

$$\mathrm{Con}_T := \neg \mathrm{Prov}_T(\ulcorner 0 = 1 \urcorner)\,. \tag{3.90}$$

Theorem 3.88 (Gödel's second incompleteness theorem)**.** *Let T be effectively axiomatized and contain, or interpret, PA. This is more than Theorem 3.87 demands — representing the computable functions does not by itself make the derivability conditions below available — and PA and ZFC both qualify. If T is consistent, then*

$$T \nvdash \mathrm{Con}_T\,. \tag{3.91}$$

No such theory can prove its own consistency [681].
Rests on Theorem 3.87, Equation (3.90), and Definition 3.83.

Derivation. *Derives Theorem 3.88.*

The first half of the proof of Theorem 3.87 is an argument, from the consistency of T, to the unprovability of G_T — that is, an argument for $\mathrm{Con}_T \to G_T$. The content of the second theorem is that this argument can itself be formalized inside T, giving

$$T \vdash\ \mathrm{Con}_T \to G_T\,. \tag{3.92}$$

Granting Equation (3.92), suppose $T \vdash \mathrm{Con}_T$. Modus ponens would give $T \vdash G_T$, which Theorem 3.87 forbids for consistent T. Hence $T \nvdash \mathrm{Con}_T$.

Establishing Equation (3.92) requires that Prov_T behave inside T the way provability behaves outside it. The sufficient conditions, isolated by Hilbert and Bernays [839] and streamlined by Löb [1074], are the three *derivability conditions*:

$$T \vdash \varphi \implies T \vdash \mathrm{Prov}_T(\ulcorner \varphi \urcorner)\,, \tag{3.93}$$

$$T \vdash \mathrm{Prov}_T(\ulcorner \varphi \to \chi \urcorner) \to (\mathrm{Prov}_T(\ulcorner \varphi \urcorner) \to \mathrm{Prov}_T(\ulcorner \chi \urcorner))\,, \tag{3.94}$$

$$T \vdash \mathrm{Prov}_T(\ulcorner \varphi \urcorner) \to \mathrm{Prov}_T(\ulcorner \mathrm{Prov}_T(\ulcorner \varphi \urcorner) \urcorner)\,. \tag{3.95}$$

Verifying them for a specific coding is laborious rather than deep. The verification and the derivation of Equation (3.92) from them are given in Appendix A.5. □

Remark 3.89 (Löb's theorem). The derivability conditions have a striking consequence: if $T \vdash \mathrm{Prov}_T(\ulcorner \varphi \urcorner) \to \varphi$, then $T \vdash \varphi$ [1074]. A theory can affirm "if this is provable then it is true" only of things it already proves. Taking φ to be $0 = 1$ recovers Theorem 3.88. Whatever confidence we have that provable arithmetic statements are true, arithmetic cannot supply it from within.

3.8.6 What the theorems do not say

Few results have been abused more, so the boundaries deserve stating as plainly as the theorems.

1. **They do not exhibit an unknowable truth.** We know G_T is true — Theorem 3.87 proves it, in a metatheory. The sentence is unprovable *in* T, and provable in any reasonable stronger system; PA's Gödel sentence is a theorem of ZFC. Incompleteness is always relative to a fixed system.

2. **They do not apply to every formal system.** A theory must be effectively axiomatized, consistent, and strong enough to represent computation. Drop the last and complete decidable theories abound: the theory of dense linear orders, Presburger arithmetic — the naturals with addition but not multiplication [666] — and, most strikingly, Tarski's result that the first-order theory of the real closed fields, and with it all of elementary algebra and Euclidean geometry, is complete and decidable [992]. Multiplication together with addition on the integers is what tips a theory over the edge.

3. **They do not show mathematics is inconsistent, or unreliable.** They show a specific ambition — self-certification — is unattainable. See Section 3.10.

4. **They do not settle whether minds exceed machines.** Arguments that a human "sees" the truth of G_T while a machine cannot assume that the human's own reasoning is a consistent formal system and that we know which one; neither premise is available. The question is open and these theorems do not close it.

5. **They do not by themselves make physics incomplete.** A physical theory is not a formal system in the required sense, and its content is fixed by measurement rather than derivation. Genuine undecidability results in physics exist, and they come from computability (Section 3.9), not from Gödel numbering.

Remark 3.90 (Independent statements that are not contrived)*.* The sentence G_T is built for the purpose and of no independent interest, which for forty years left the impression that incompleteness was a curiosity confined to self-reference. Paris and Harrington closed that escape [1608]: a natural finitary strengthening of Ramsey's combinatorial theorem, whose statement mentions nothing about provability, is true and unprovable in PA. Goodstein's theorem [874], about certain rapidly growing integer sequences that nevertheless terminate, is another: Kirby and Paris showed it too escapes PA [1690]. Ordinary mathematics does reach past the boundary.

Theorem 3.91 (Tarski's undefinability of truth)**.** *Let T be a consistent theory as above. There is no formula* $\mathrm{True}(x)$ *of the language of arithmetic such that $\varphi \leftrightarrow \mathrm{True}(\ulcorner\varphi\urcorner)$ holds for every sentence φ [761].*

Rests on Lemma 3.86 and Definition 3.82.

Proof. *Derives Theorem 3.91.*

Suppose such a formula existed. Apply Lemma 3.86 to $\neg\mathrm{True}(x)$ to obtain a sentence λ with $\lambda \leftrightarrow \neg\mathrm{True}(\ulcorner\lambda\urcorner)$. Combining with the assumed property of True at $\varphi = \lambda$ gives $\lambda \leftrightarrow \neg\lambda$, a contradiction. □

The sentence λ says "I am false": it is the liar paradox, made into a theorem instead of a contradiction by locating the fault correctly. Truth for a language is definable, but only in a richer metalanguage — exactly the stratification anticipated in Remark 3.18. Provability, by contrast, *is* expressible in the object language, which is why Equation (3.86) exists. That asymmetry between truth and provability is the substance of incompleteness.

3.9 Computability and decidability

Two of Hilbert's demands survived Theorem 3.87: even if a theory cannot prove every truth, one might still hope for a mechanical procedure that decides, of any given sentence, whether it is provable. Hilbert and Ackermann posed this as *das Entscheidungsproblem* — the decision problem — and called it the principal problem of mathematical logic [606]. Answering it required saying what a mechanical procedure *is*, a question nobody had needed

to make precise. Three answers arrived within a year of one another, and they turned out to define the same class.

3.9.1 Turing machines

Turing's analysis began not with mathematics but with a description of a human computer at work: a person with a pencil, an unlimited supply of ruled paper, a finite number of states of mind, and a finite book of rules, attending to one square at a time [803]. Idealizing this gives the following.

Definition 3.92 (Turing machine)**.** A *Turing machine* is a tuple $M = (Q, \Gamma, \sqcup, \Sigma, \delta, q_0)$ where

1. Q is a finite set of *states*;
2. Γ is a finite *tape alphabet* containing a distinguished *blank* symbol $\sqcup$;
3. $\Sigma \subset \Gamma \setminus \{\sqcup\}$ is the *input alphabet*;
4. $q_0 \in Q$ is the *initial state*;
5. the *transition function* δ is a partial map

$$\delta : Q \times \Gamma \longrightarrow Q \times \Gamma \times \{L, R\} \ . \tag{3.96}$$

The machine acts on a tape of cells unbounded in both directions, each holding one symbol of Γ, all but finitely many blank, with a head positioned over one cell. A *configuration* is the triple consisting of the tape contents, the head position and the current state. One *step* reads the scanned symbol a in state q, and if $\delta(q, a) = (q', a', D)$ is defined, writes a' in place of a, moves the head one cell in the direction D and enters state q'. If $\delta(q, a)$ is undefined the machine *halts*. The machine *computes* the partial function sending an input string, written on an otherwise blank tape with the head at its first symbol, to the tape contents at halting; where it never halts, the function is undefined.

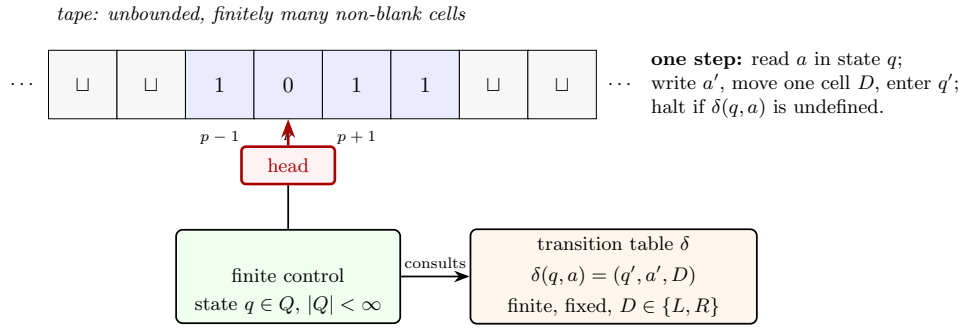


Figure 3.2: A Turing machine (Definition 3.92). The tape is unbounded and holds finitely many non-blank symbols; the head reads and writes one cell at a time; the finite control is in one of finitely many states and consults a fixed finite table δ. Everything unbounded is on the tape, and everything the machine *is* — the state set and the table — is finite. That separation is what makes the machine codeable as a number and hence a possible input to another machine.

Nothing about this device is electronic, and nothing about it is fast. Its significance is that it is manifestly mechanical and just as manifestly unrestricted in the memory it may use — and that the machine itself, being finite, can be written down as a finite string and so coded by a number $\ulcorner M \urcorner$, exactly as formulas were coded in Section 3.8.3.

Definition 3.93 (Computable function, decidable set)**.** A partial function on strings is *computable* if some Turing machine computes it. A set is *decidable* if its indicator function is computable — some machine halts on every input and reports correctly whether the input belongs to the set. A set is *semi-decidable* if some machine halts exactly on its members.
Rests on Definition 3.92.

Definition 3.81 can now be read literally: a theory is effectively axiomatized when its set of axioms is decidable. And a theory that is effectively axiomatized has a semi-decidable set of theorems — enumerate all finite derivations in order and watch for one ending in the target sentence — which is why Prov_T of Equation (3.86) exists.

Axiom 3.94 (Church–Turing thesis)**.** Every function computable by any mechanical procedure whatsoever is computable by a Turing machine.

This is not a theorem and cannot be one, since "mechanical procedure" on the left is informal. It is an empirical and conceptual claim, and its support is the convergence of independent attempts: Church's λ-definable functions [773], Kleene's general recursive functions [778], Post's combinatory processes [784] and Turing's machines all define the same class, and every model of computation proposed since — register machines, cellular automata, the physical computers on our desks, quantum computers — has been proved no more powerful in what it can compute at all, whatever the differences in efficiency.

Theorem 3.95 (Universal machine)**.** *There exists a Turing machine U which, on input the pair $(\ulcorner M \urcorner, w)$, simulates M on w: it halts precisely when M halts on w, with the same output.*
Rests on Definitions 3.92 and 3.93.

Full derivation: Appendix A.6 (p. 2788).
Derives Theorem 3.95.

The universal machine is why a computer is a general-purpose device rather than a special-purpose one, and it is the technological content of [803]. For the present chapter its importance is that programs are data: a machine can be fed to a machine, including to itself.

3.9.2 The halting problem

Theorem 3.96 (Turing)**.** *There is no Turing machine which, given $(\ulcorner M \urcorner, w)$, always halts and correctly reports whether M halts on input w. The halting problem is undecidable [803].*
Rests on Definitions 3.92 and 3.93.

Proof. *Derives Theorem 3.96.*

Suppose such a machine H existed, halting on every input and returning *yes* when M halts on w and *no* otherwise.

Construct from it a machine D which, on input a single machine code $\ulcorner M \urcorner$, runs H on the pair $(\ulcorner M \urcorner, \ulcorner M \urcorner)$ — asking whether M halts when fed its own code — and then does the opposite of the answer:

$$D(\ulcorner M \urcorner) = \begin{cases} \text{loop forever} & \text{if } H \text{ answers yes}, \\ \text{halt} & \text{if } H \text{ answers no}. \end{cases} \tag{3.97}$$

D is an ordinary machine, built from H by adding a copying step and a deliberate infinite loop, so it has a code $\ulcorner D \urcorner$.

Run D on $\ulcorner D \urcorner$. If D halts on $\ulcorner D \urcorner$, then H answered *yes*, and Equation (3.97) made D loop forever — so it does not halt. If D does not halt on $\ulcorner D \urcorner$, then H answered *no*, and Equation (3.97) made D halt. Either branch contradicts itself: D halts on its own code if and only if it does not. Since D was built from H by unimpeachable steps, the fault lies with the assumption that H exists. □

That is the diagonal argument for the fourth time, and the family resemblance is now unmistakable: replace "does not contain itself" by "does not halt on its own code" and Russell becomes Turing; replace it by "is not provable" and Russell becomes Gödel.

Theorem 3.97 (Rice)**.** *Let $\mathcal{P}$ be any set of partial computable functions that is neither empty nor all of them. Then the set of machine codes $\{\ulcorner M \urcorner \mid M \text{ computes a function in } \mathcal{P}\}$ is undecidable.*

Rests on Theorems 3.95 and 3.96 and Definition 3.93.

Proof. *Derives Theorem 3.97.*

Write $\varnothing$ for the nowhere-defined function, computed by a machine that immediately loops. Replacing $\mathcal{P}$ by its complement if necessary — which changes decidability not at all — assume $\varnothing \notin \mathcal{P}$, and pick $g \in \mathcal{P}$, computed by some machine G.

Given an arbitrary pair $(\ulcorner M \urcorner, w)$, build the machine N which, on any input x, first simulates M on w using Theorem 3.95, and only if that simulation halts goes on to simulate G on x and return its output. If M halts on w, then N computes $g \in \mathcal{P}$; if M does not halt on w, then N halts on no input and computes $\varnothing \notin \mathcal{P}$. The code $\ulcorner N \urcorner$ is obtained from $(\ulcorner M \urcorner, w)$ by a computable transformation.

A decision procedure for $\mathcal{P}$ would therefore decide, of any $(\ulcorner M \urcorner, w)$, whether M halts on w, contradicting Theorem 3.96. □

Rice's theorem is the sweeping form of the bad news: *no* non-trivial question about what a program computes — as opposed to how it is written — is answerable by a general algorithm. Not "does it ever output zero", not "does it compute the same function as this other program", not "does it terminate on all inputs".

3.9.3 The Entscheidungsproblem answered

Theorem 3.98 (Church–Turing)**.** *The set of valid sentences of first-order logic is undecidable: no mechanical procedure determines, of an arbitrary first-order sentence, whether it is provable [772, 803].*

Rests on Theorem 3.96 and Definitions 3.80 and 3.93.

Full derivation: Appendix A.5 (p. 2780).

Derives Theorem 3.98.

The proof encodes the behaviour of a Turing machine into a first-order sentence which is valid if and only if the machine halts, reducing Theorem 3.96 to the decision problem. Hilbert's principal problem of mathematical logic thus has a negative answer, and the two negative answers fit together: by Theorem 3.84 the valid sentences are exactly the provable ones, so they are semi-decidable — search all derivations — but by Theorem 3.98 not decidable. A search that will succeed if there is anything to find, but that gives no sign of when to stop, is the best any general procedure can do.

The corresponding statement for arithmetic is sharper still: the set of true sentences of arithmetic is not even semi-decidable, which is Theorem 3.91 in computational dress.

3.9.4 Where decision procedures do exist

The negative results are about *general* procedures over *all* sentences of a sufficiently rich language. Restrict the language and algorithms reappear, sometimes spectacularly.

1. **Propositional logic** is decidable: build the truth table, as in Proposition 3.22. The cost is 2^n rows for n propositions, and whether an essentially faster method exists is the P versus NP problem, still open.

2. **Presburger arithmetic** — the naturals with addition and order but *without* multiplication — is consistent, complete and decidable [666]. Multiplication is what makes arithmetic able to encode syntax, and hence what makes it incomplete.

3. **Real closed fields.** Tarski proved that the first-order theory of $(\mathbb{R}, +, \cdot, <)$ admits elimination of quantifiers and is therefore complete and decidable [992]. Every statement of elementary algebra and of Euclidean and analytic geometry expressible with real variables, polynomials, equalities, inequalities and quantifiers can be settled by algorithm. This is the technical foundation of computer algebra and of automated geometry; it also means that the real-number algebra a physicist actually manipulates sits on the decidable side of the line. The decision procedure is doubly exponential and rarely practical, but it exists.

The boundary between these islands and the undecidable ocean is sharp and somewhat surprising: the reals with addition and multiplication are decidable, the integers with addition are decidable, but the integers with addition *and* multiplication are not.

3.9.5 Undecidability inside physics

Physical questions can be phrased so as to encode computations, and when they are, they inherit undecidability. Three results, in increasing order of physical relevance.

1. **Hilbert's tenth problem.** There is no algorithm deciding whether an arbitrary polynomial equation in several variables with integer coefficients has an integer solution. Davis, Putnam and Robinson reduced the question to the arithmetic definability of exponential growth, and Matiyasevich completed the proof in 1970 by showing the Fibonacci sequence is Diophantine [1429, 1496]. Diophantine constraints are not exotic in physics — they govern commensurability of orbital periods, lattice matching conditions and anomaly-cancellation conditions on integer charges.

2. **Tilings and ground states.** Whether a finite set of tile shapes tiles the plane is undecidable [1326], and a tiling problem can be rewritten as the problem of finding the ground-state configuration of a classical lattice model with finite-range interactions. The ground-state energy density of a general such model is therefore not computable by any algorithm uniform in the model.

3. **The spectral gap.** Cubitt, Pérez-García and Wolf constructed a family of translationally invariant, nearest-neighbour Hamiltonians on a two-dimensional lattice of spins, depending on a parameter, such that determining whether the model is gapped or gapless in the thermodynamic limit is undecidable [2251]. The construction embeds a universal Turing machine in the low-energy physics: the model is gapped exactly when the encoded machine fails to halt.

Remark 3.99 (How to read these results honestly). They do not say that Nature computes something uncomputable, and they do not say any particular laboratory question is unanswerable. Every one of them concerns an *infinite family* of idealized models and asks for a single algorithm covering the whole family. A specific finite system has a finite-dimensional Hamiltonian whose gap is computable to any desired accuracy by diagonalization; what fails is the existence of a general criterion — a finite set of rules, a closed-form classification — valid for all models at once. The thermodynamic limit, not quantum mechanics, is where the undecidability enters, since it is the limit that supplies the unbounded tape.

The practical content is a warning about ambitions rather than about measurements. Programmes seeking a complete classification of quantum phases, or a general algorithm to determine whether a lattice model is gapped — the Yang–Mills mass gap being the

celebrated instance (Chapter 133) — are seeking something that provably does not exist in full generality. Progress must come by restricting the class of models, exactly as Section 3.9.4 restricts the language. That is a genuine, and genuinely useful, piece of physical knowledge.

3.10 Can mathematics be built without contradictions?

The question that opened the crisis can now be asked precisely, and answered precisely — though not in the form anyone wanted.

3.10.1 What Hilbert asked for

Hilbert's programme had a definite shape [546]. Mathematics was to be split into a *finitary* part, concerning concrete finite configurations of symbols, whose meaning and reliability nobody doubts, and an *ideal* part — infinite sets, arbitrary real numbers, the whole Cantorian apparatus — introduced because it makes the theory smooth, in the way that points at infinity make projective geometry smooth. The programme demanded a proof, *carried out by finitary means alone*, that the ideal part never yields a false finitary statement. Consistency was the crux: an inconsistent system proves everything, so a consistency proof would license the entire ideal apparatus at one stroke, and the mathematician could then use infinite sets with a clear conscience while regarding them as convenient fictions.

Hilbert put the case in an address broadcast from Königsberg on 8 September 1930, ending with the words carved on his gravestone: we must know, we shall know [658]. The previous day, at a conference in the same city, Gödel had mentioned to a roundtable that he could construct true but unprovable arithmetical statements. Almost nobody present grasped what they had heard.

1. Theorem 3.87 killed the hope that a single effectively axiomatized system could settle every arithmetical question.

2. Theorem 3.88 killed the programme proper: since finitary reasoning is certainly formalizable within PA, and PA cannot prove its own consistency, no finitary proof of the consistency of PA — let alone of analysis or set theory — can exist.

3. Theorem 3.98 killed the remaining hope of a mechanical decision procedure.

3.10.2 What survived: relative consistency

What did *not* die is the mathematical study of consistency, which became one of the most productive branches of logic. The trick is to stop demanding absolute certification and to measure systems against each other instead.

Gentzen proved in 1936 that Peano arithmetic is consistent [776]. There is no contradiction with Theorem 3.88, because his proof is not carried out inside PA: it uses finitary reasoning *plus* one transfinite ingredient, induction along the ordinal

$$\varepsilon_0 = \sup\left\{\omega,\ \omega^{\omega},\ \omega^{\omega^{\omega}},\ \dots\right\}, \tag{3.98}$$

the first ordinal not reachable from ω by finitely many exponentiations. Every proper initial segment of ε_0 admits an induction principle provable in PA; ε_0 itself does not, and that single step beyond is exactly what a consistency proof costs. The number ε_0 is thereby a measurement of the strength of PA, and assigning such *proof-theoretic ordinals* to theories is now a discipline in its own right.

The same relativization organizes set theory. None of the following results asserts consistency outright; each transfers it:

$$\mathrm{Con}(\mathrm{ZF}) \implies \mathrm{Con}(\mathrm{ZFC})\,, \tag{3.99}$$

$$\mathrm{Con}(\mathrm{ZF}) \implies \mathrm{Con}(\mathrm{ZFC} + \mathrm{CH})\,, \tag{3.100}$$

$$\mathrm{Con}(\mathrm{ZF}) \implies \mathrm{Con}(\mathrm{ZFC} + \neg\mathrm{CH})\,, \tag{3.101}$$

$$\mathrm{Con}(\mathrm{ZFC}) \implies \mathrm{Con}(\mathrm{PA})\,. \tag{3.102}$$

The first two are Gödel's, by the constructible universe [815]; the third is Cohen's, by forcing [1249]. Equation (3.102) runs one way only: ZFC proves the consistency of PA — it constructs a model, namely $\mathbb{N}$ — and by Theorem 3.88 PA cannot return the favour, nor can ZFC certify itself. Stronger axioms of infinity extend the ladder upward indefinitely, each rung provably consistent relative to the ones above and never to itself.

3.10.3 The honest answer

Can mathematics be constructed without contradictions? The precise answers are these.

1. **Almost certainly yes** — but this cannot be proved from within, and never will be. Any proof of the consistency of a system strong enough for ordinary mathematics must use assumptions strictly stronger than the system it certifies, whose own consistency is then open. The regress does not terminate. Absolute self-certification is not merely undone; it is impossible.

2. **Consistency is not an unknowable, though.** It is a mathematical statement with a definite truth value, provable in stronger systems, refutable in principle by a single explicit derivation of $0 = 1$, and studied quantitatively by ordinal analysis. What is unavailable is a proof using nothing but means we are already certain of.

Remark 3.100 (Consistency has the epistemic status of a physical law). The parallel is exact, and it is the reason this section belongs in a physics treatise. The statement "ZFC is consistent" is *falsifiable*: a derivation of a contradiction would refute it decisively, and could be checked by any competent clerk. It is *unverifiable*: no finite body of successful practice establishes it, just as no finite run of confirming experiments establishes a universal law (Section 3.7.2). And it is *well corroborated*: a century of intensive, adversarial use by thousands of mathematicians, an enormous body of consequences, and a family of natural inner models have all failed to produce a contradiction, exactly as two centuries of mechanics failed to refute Newton.

We are therefore in the same position with respect to the foundations of mathematics as with respect to the foundations of physics: holding a well-tested conjecture, aware that it is a conjecture, and getting on with the work (Remark 1.2). That is not a comfortable position, but it is an honest one, and it is the only one available.

3.10.4 How much of this reaches physics

Very little, and the reason is worth stating precisely rather than comfortingly.

The heavy axioms of ZFC — power set applied without limit, replacement, full choice — are what generate the exotic cardinals and the independence phenomena. Physics uses almost none of that. The programme of *reverse mathematics* determines, for each classical theorem, exactly which axioms are needed to prove it [2149], and the answer for the mathematics of physics is consistently: far less than ZFC. The convergence of sequences, the intermediate value theorem, the existence and uniqueness of solutions to ordinary differential equations (Chapter 9), the Riemann and Lebesgue integrals, the spectral theorem for separable Hilbert

spaces (Chapter 12) — these live in weak subsystems of second-order arithmetic whose consistency strength is at or below that of PA itself. Physics never leaves the lower reaches of the hierarchy, and its mathematics is certified by Gentzen's theorem, not left waiting on the consistency of ZFC.

Suppose, nonetheless, that a contradiction in ZFC were found tomorrow. What would happen to physics? Essentially nothing. Every measured value in this book would still be what the apparatus reported, since a measurement is a physical event and not a theorem. The derivations would be re-examined and, in overwhelming likelihood, re-derived in a repaired system, because they never used the offending strength in the first place — exactly as the discovery that unrestricted comprehension is inconsistent (Theorem 3.75) invalidated not one line of Newtonian mechanics or one entry in a laboratory notebook. Foundations are the plumbing of mathematics; when plumbing fails it is replaced, and the building stands.

Remark 3.101 (The position of this treatise)*.* This book reasons classically, works in ZFC, assumes its consistency, and now has said so explicitly rather than silently. Every derivation it contains is a proof in the sense of Definition 3.80 — informal in presentation, formalizable in principle — and the editorial rule that no phenomenon may be stated without its derivation is therefore a commitment whose meaning and whose limits are both fixed by this chapter. The limits are real: there are arithmetical truths no derivation in a fixed system will reach (Theorem 3.87), questions no algorithm will settle (Theorem 3.96), and physical model classes whose gross qualitative behaviour admits no general criterion (Remark 3.99).

None of that licenses vagueness, and it is worth being clear why. What the theorems remove is the dream of a complete, self-certifying, mechanically decidable body of knowledge. What they leave entirely intact is the obligation to prove what one asserts, to state what one assumes, and to report what one measured. The rest of this treatise does exactly that.

Chapter 4

Algebraic Structures

Algebra begins when a bare set is equipped with a rule for combining its elements. This chapter builds that idea up axiom by axiom: binary operations and closure first, then associativity, neutral and inverse elements, until the simple structures (semigroups, monoids, groups) and the composite ones (rings, fields, modules, vector spaces, algebras) emerge as successive strengthenings of the same scheme. The second half of the chapter studies groups in detail — the structure that physics uses to express symmetry — including their basic examples, the maps that preserve them, the symmetric group, the decompositions of a group into conjugacy classes and cosets, the quotient group, and the direct product.

4.1 Binary operations and algebraic structures

4.1.1 Binary operations

Consider two sets A and B. We may ask whether, starting from A and B, it is possible to define a set containing all the possible combinations of two elements of A and of B respectively. For example, let $A = \{a, b\}$ and $B = \{\alpha, \beta\}$. Is the collection of the pairings (a, α), (a, β), (b, α) and (b, β) then a set? The answer is yes, owing to the arbitrariness in the definition of a set (Definition 3.28). According to the historical account, René Descartes was the first to think of ordering pairs of numbers in this way.

Definition 4.1 (Cartesian product)**.** The *Cartesian product* of two sets A and B, denoted $A \times B$, is the set containing as elements all the possible pairs of combinations of elements of A and B. The elements of $A \times B$ are called 2-*tuples*. The generalization is immediate: the elements of the Cartesian product $A_1 \times \ldots \times A_n$ are called n-*tuples*.

Rests on Definition 3.28.

As an example, note that it is also possible to construct the Cartesian product of a set with itself,

$$\{a, b\} \times \{a, b\} = \{(a, a), (a, b), (b, a), (b, b)\} ,$$

where it must be noted that $(a, b) \neq (b, a)$, since they are distinct combinations.

Definition 4.2 (Binary operation)**.** Consider the sets A and X. A *binary operation* $*$ is a map of the form

$$\begin{array}{rccc} * : & A \times A & \longrightarrow & X \\ & (a, b) & \longmapsto & a * b \, . \end{array}$$

Rests on Definitions 3.43 and 4.1.

Definition 4.3 (Algebraic structure)**.** Consider a set A and an arbitrary set X, together with a binary operation $* : A \times A \longrightarrow X$. The set A together with the operation $*$ is called an *algebraic structure*, and is denoted $(A, *)$.

Rests on Definition 4.2.

For example, the set $\{0,1\} \subset \mathbb{Z}$ together with ordinary addition,

$$\begin{array}{rccc} + : & \{0,1\} \times \{0,1\} & \longrightarrow & \{0,1,2\} \\ & (a,b) & \longmapsto & a+b\,, \end{array}$$

forms an algebraic structure whose binary operation is made explicit by the table

$+$	0	1
0	0	1
1	1	2

The closure axiom

Consider an algebraic structure $(A, *)$. If the binary operation has not previously been fixed explicitly, in general $a * b$ need not be an element of A. The current theories of physics, however, rest on a mathematical formalism in which sets lacking such an internal mechanism are not of interest, and we therefore present the formalism only for algebraic structures that possess one.

Definition 4.4 (Internal binary operation; magma)**.** A binary operation $* : A \times A \longrightarrow X$ is an *internal binary operation* (IBO) if and only if $X = A$. An algebraic structure $(A, *)$ whose operation $*$ is an IBO is said to be *closed*, to be a *magma*, or to have *closure* with respect to $*$. Closure is expressed formally by the axiom

$$\forall\ a, b \in A,\ \ a * b \in A\,. \tag{4.1}$$

Rests on Definitions 4.2 and 4.3.

Remark 4.5. When no ambiguity is possible or relevant, we shall refer to a closed algebraic structure simply as an algebraic structure, and we shall denote an algebraic structure $(A, *)$ simply by A, the name of its underlying set.

The associativity axiom

Just as happens with the real numbers, we can formalize the fact that the product of one number with the product of two others equals the product of the first two with the last.

Definition 4.6 (Associativity)**.** An algebraic structure $(A, *)$ satisfies the property of *associativity* if and only if

$$\forall\ a, b, c \in A,\ \ a * (b * c) = (a * b) * c\,. \tag{4.2}$$

Rests on Definition 4.3.

Remark 4.7. The associativity axiom is genuinely important because of the following property. Consider the equation $a = b * c$. With associativity available we may operate on the whole equation with a further quantity d; that is, $a * d = (b * c) * d$, or equivalently $a * d = b * (c * d)$. If we did not have that property we would be left with something like $a * d = b * c * d$, an expression devoid of meaning, since we have not defined $*$ as a "ternary operation".

The commutativity axiom

For the real numbers, the order of the factors does not alter the product. We can formalize this for algebraic structures through the following axiom.

Definition 4.8 (Commutativity; abelian structure)**.** An IBO $*$ is *commutative* — equivalently, an algebraic structure $(A, *)$ is *abelian* with respect to $*$ — if and only if the following axiom holds:

$$\forall\ a, b \in A,\ \ a * b = b * a\,. \tag{4.3}$$

Rests on Definition 4.4.

The neutral-element axiom and uniqueness

In order to construct more sophisticated structures, we may demand the existence of an element e_A whose operation with any element is irrelevant.

Definition 4.9 (Neutral element)**.** An element e_A of an algebraic structure $(A, *)$ is a *neutral element* if and only if it satisfies the axiom

$$\forall\, a \in A,\ \exists\, e_A \in A \mid a * e_A = a \ \wedge\ e_A * a = a\,. \tag{4.4}$$

Rests on Definition 4.3.

We must observe that the axiom does not assert commutativity among all the elements of the set: commutation is demanded only for the neutral element. Many of the structures we shall define make use of this axiom, which will let us uncover properties usable throughout mathematical physics. On occasion, when we wish to specify the symbol employed for the neutral element, we shall denote the algebraic structure by $(A, *, e_A)$.

Proposition 4.10 (Uniqueness of the neutral element)**.** *The neutral element of an algebraic structure is unique.*

Rests on Definition 4.9.

Proof. *Derives Proposition 4.10.*

Consider two neutral elements e_A and f_A. By the axiom we have $e_A = e_A * f_A$, and furthermore

$$\begin{aligned} f_A &= f_A * e_A \\ &= e_A * f_A \qquad\qquad \text{by Equation (4.4)}\,, \end{aligned}$$

hence $e_A = f_A$. □

The inverse-element axiom

Let $(A, *, e_A)$ be an algebraic structure equipped with the neutral-element axiom. A further axiom we may impose is that to each element a there be assigned a dual element a^{-1} such that, when the two are operated, the neutral element results.

Definition 4.11 (Inverse element)**.** An algebraic structure $(A, *, e_A)$ satisfies the *inverse-element axiom* if and only if

$$\forall\, a \in A,\ \exists\, a^{-1} \in A \mid a * a^{-1} = e_A \ \wedge\ a^{-1} * a = e_A\,, \tag{4.5}$$

and we call a^{-1} an *inverse element* of a.

Rests on Definition 4.9.

Substructures

Consider a closed algebraic structure $(A, *)$, and let $S \subseteq A$ be a non-empty subset. The IBO $*$ induces a new binary operation $*_S : S \times S \longrightarrow A$. Then $(S, *_S)$ is an algebraic structure, but not necessarily a closed one. If, however, we restrict the codomain, the operation $*'_S : S \times S \longrightarrow S$ is by definition an IBO, and $(S, *'_S)$ is a closed algebraic structure. When no ambiguity is possible we shall simply write $*'_S = *$.

Definition 4.12 (Substructure)**.** $(S, *)$ is a *substructure* of $(A, *)$ if and only if

1. $S \neq \varnothing$;
2. $\forall\, a, b \in S,\ a * b \in S$.

We further introduce the notation

$$(S, *) \text{ substructure of } (A, *) \quad \overset{\text{not.}}{\Longleftrightarrow} \quad S \leq A\,.$$

Rests on Definition 4.4.

4.1.2 Maps between algebraic structures

Product of maps

Consider the algebraic structures $(A, *_A)$ and $(B, *_B)$. As we know (Definition 3.52), $\mathcal{F}(A, B)$ is the set of all maps from A to B. Let us define the IBO

$$\begin{array}{rccccccc} *: & \mathcal{F}(A,B)\times\mathcal{F}(A,B) & \longrightarrow & \mathcal{F}(A,B) & & & & \\ & (f,g) & \longmapsto & f*g: & A & \longrightarrow & B \\ & & & & a & \longmapsto & (f*g)(a) = f(a) *_B g(a)\,. \end{array}$$

Given the maps f and g, we have thereby defined a new map $f * g \in \mathcal{F}(A, B)$, which we shall call the *product of maps*.

Neutral map

Let $(A, *_A)$ be an algebraic structure and $(B, *_B, e_B)$ an algebraic structure satisfying the neutral-element axiom. Consider a map of the form

$$\begin{array}{rccc} \Theta: & A & \longrightarrow & B \\ & a & \longmapsto & e_B\,, \end{array}$$

that is, a map assigning to every element of A the neutral element of B. We shall call this map the *neutral map*, *neutral function*, or *neutral operator*. Now consider a map $f : A \longrightarrow B$. We see that

$$\begin{aligned} (f * \Theta)(a) &= f(a) *_B \Theta(a) \\ &= f(a) *_B e_B \\ &= f(a)\,, \end{aligned}$$

so, since a is arbitrary, we simply have $f * \Theta = f$, and analogously $\Theta * f = f$.

Opposite map

The previous two headings supplied $\mathcal{F}(A, B)$ with an operation and with a neutral element for it. One ingredient is missing before the set of maps is a group in its own right, and it is the pointwise inverse.

Let $(A, *_A)$ be an algebraic structure and let $(B, *_B, e_B)$ satisfy both the neutral-element axiom and the inverse-element axiom Equation (4.5). Given $f : A \longrightarrow B$, define

$$\begin{array}{rccc} f^{\mathrm{op}}: & A & \longrightarrow & B \\ & a & \longmapsto & (f(a))^{-1}\,, \end{array} \tag{4.6}$$

which we call the *opposite map* of f: it inverts the value of f at each point separately, and does nothing else.

Notation 4.13 (The opposite map is not the inverse map). The superscript is op, and not -1, on purpose. The symbol f^{-1} is already spoken for twice over in Chapter 3: it is the inverse map of a bijection (Definition 3.51), which exists only when f is bijective and satisfies $f^{-1} \circ f = \mathrm{id}_A$; and it is the preimage (Definition 3.53), which returns a subset. The opposite map Equation (4.6) is neither. It is defined for *every* f, bijective or not, it involves no composition, and the inversion happens in the codomain B rather than between A and B. The three notions coincide in no interesting case, and conflating them is the commonest slip in this part of the subject.

Rests on Equation (4.6) and Definitions 3.51 and 3.53.

For f^{op} to be a map at all, the value $(f(a))^{-1}$ must be determined by $f(a)$ — the existence and uniqueness demanded by Definition 3.43. Existence is the inverse-element axiom. Uniqueness is *not* automatic: Equation (4.5) asserts that an inverse exists, not that it is unique, and the proof of uniqueness (Proposition 4.21) needs associativity. We therefore assume throughout this heading that $(B, *_B)$ is associative — in practice that B is a group, Definition 4.20 — and note that without it Equation (4.6) selects one inverse among possibly several and is a definition only after such a selection has been made.

Granting that, the opposite map does for $\mathcal{F}(A, B)$ exactly what a^{-1} does for B:

$$\begin{aligned}(f * f^{\mathrm{op}})(a) &= f(a) *_B f^{\mathrm{op}}(a) \\ &= f(a) *_B (f(a))^{-1} \\ &= e_B \\ &= \Theta(a)\,,\end{aligned}$$

and since $a \in A$ was arbitrary, $f * f^{\mathrm{op}} = \Theta$. The same computation with the factors exchanged gives $f^{\mathrm{op}} * f = \Theta$.

Proposition 4.14 (Maps into a group form a group)**.** *Let A be a non-empty set and $(B, *_B, e_B)$ a group. Then $(\mathcal{F}(A, B), *)$ is a group, with neutral element the neutral map Θ and with the inverse of f given by f^{op}. If B is abelian, so is $\mathcal{F}(A, B)$.*

Rests on Definition 4.20, Equation (4.6), and Proposition 4.21.

Proof. *Derives Proposition 4.14.*

Each group axiom is checked pointwise, which is available because two maps are equal exactly when they agree at every point.

Closure. $f * g$ was defined as a map $A \longrightarrow B$, so it lies in $\mathcal{F}(A, B)$.

Associativity. For every $a \in A$,

$$((f * g) * h)(a) = (f(a) *_B g(a)) *_B h(a) = f(a) *_B (g(a) *_B h(a)) = (f * (g * h))(a)\,,$$

the middle step being associativity in B.

Neutral element. $f * \Theta = \Theta * f = f$ was shown under the previous heading.

Inverses. $f * f^{\mathrm{op}} = f^{\mathrm{op}} * f = \Theta$ was shown above, and f^{op} is well defined because inverses in a group are unique by Proposition 4.21.

Commutativity. If B is abelian then $(f * g)(a) = f(a) *_B g(a) = g(a) *_B f(a) = (g * f)(a)$ for every a, so $f * g = g * f$. □

This small proposition is used more often in physics than its size suggests. It is the reason that the B-valued functions on any set inherit B's algebraic structure without further argument: real-valued functions on a region form an abelian group under pointwise addition, and with the scalar multiplication of Chapter 5 a vector space — which is what makes function spaces the natural home of fields (Chapter 12). When B is a Lie group rather than $\mathbb{R}$, the same pointwise construction produces the group of gauge transformations, the maps from spacetime into the gauge group, whose structure is the subject of Chapter 14 and whose physical content occupies Chapter 64. Nothing in the construction requires A to carry any structure at all; only the codomain does the work.

4.1.3 Kernel and image

Definition 4.15 (Kernel)**.** Let $(A, *_A)$ be an algebraic structure and $(B, *_B, e_B)$ an algebraic structure satisfying the neutral-element axiom, and consider a map $f : A \longrightarrow B$. The *kernel* of f is the set

$$\ker f = \{a \in A \mid f(a) = e_B\}\,. \tag{4.7}$$

Rests on Definition 4.9.

Definition 4.16 (Image)**.** Let $(A, *_A)$ and $(B, *_B)$ be algebraic structures and $f : A \longrightarrow B$ a map. The *image* of f is the set

$$\operatorname{im} f = \{b \in B \mid b = f(a),\ a \in A\}\ . \tag{4.8}$$

As a set this is Definition 3.44, the image of a map; it is restated here only to fix the notation $\operatorname{im} f$ used throughout this chapter. Nothing about the structures enters the definition, and nothing is claimed by it — that the image of a homomorphism is itself a structure of the same kind is a separate statement and needs its own proof.

Rests on Definitions 3.43 and 3.44.

Proposition 4.17 (Trivial kernel of a homomorphism)**.** *Consider two algebraic structures satisfying the neutral-element axiom, $(A, *_A, e_A)$ and $(B, *_B, e_B)$, and let $f : A \longrightarrow B$ be a homomorphism. Then*

$$\ker f = \{e_A\} \iff f \text{ is injective}\,. \tag{4.9}$$

Rests on Definitions 4.11, 4.15 and 4.39.

Proof. *Derives Proposition 4.17.*

($\Rightarrow$) We must prove that if $\ker f = \{e_A\}$ then f is injective. Indeed, since $e_A \in \ker f$,

$$f(e_A) = e_B\,. \tag{4.10}$$

On the other hand, let $a_1, a_2 \in A$ and suppose that $f(a_1) = f(a_2)$. Operating on both sides with $*_B\, f(a_2^{-1})$,

$$\begin{aligned} f(a_1) *_B f(a_2^{-1}) &= f(a_2) *_B f(a_2^{-1}) \\ f(a_1 *_A a_2^{-1}) &= f(a_2 *_A a_2^{-1}) \\ &= f(e_A) \\ &= e_B\,, \end{aligned}$$

so $a_1 *_A a_2^{-1} \in \ker f$; hence $a_1 *_A a_2^{-1} = e_A$, and therefore $a_1 = a_2$. This proves injectivity.

($\Leftarrow$) Conversely, suppose f is injective. Because f is a homomorphism and $e_A = e_A *_A e_A$,

$$f(e_A) = f(e_A *_A e_A) = f(e_A) *_B f(e_A)\,,$$

and operating on the right with $*_B\ (f(e_A))^{-1}$ leaves $e_B = f(e_A)$ — the computation stated on its own account in Proposition 4.42. Hence $e_A \in \ker f$, that is, $\{e_A\} \subseteq \ker f$. Conversely, let $a \in \ker f$. Then $f(a) = e_B = f(e_A)$, and injectivity gives $a = e_A$; thus $\ker f \subseteq \{e_A\}$. The two inclusions together give $\ker f = \{e_A\}$. □

4.2 Simple algebraic structures

The conditions defining an algebraic structure form a purely abstract set of axioms. The axioms we now impose are chosen because the resulting structures are relevant to physics: they possess properties that we can link with nature.

4.2.1 Semigroups, monoids, and groups

Definition 4.18 (Semigroup)**.** An algebraic structure $(A, *)$ is a *semigroup* if and only if

1. $\forall\ a, b \in A, \quad a * b \in A$;
2. $\forall\ a, b, c \in A, \quad a * (b * c) = (a * b) * c$;

that is, $(A, *)$ is a magma whose elements moreover satisfy associativity under $*$.
Rests on Definitions 4.4 and 4.6.

For example, the set of natural numbers together with addition as internal binary operation forms a semigroup with infinitely many elements, the closure and associativity properties being clearly satisfied; this semigroup is moreover abelian. Another example is the abelian semigroup formed by the set $\{0, 1\} \subset \mathbb{Z}$ together with the usual multiplication, whose internal operation is summarized in the table

$\cdot$	0	1
0	0	0
1	0	1

(4.11)

Here closure is clear, and associativity follows[1] from the associativity of $\mathbb{Z}$.

Definition 4.19 (Monoid)**.** An algebraic structure $(A, *)$ is a *monoid* if and only if

1. $\forall\ a, b \in A, \quad a * b \in A$;
2. $\forall\ a, b, c \in A, \quad a * (b * c) = (a * b) * c$;
3. $\forall\ a \in A,\ \exists\ e_A \in A \mid a * e_A = a \quad \wedge \quad e_A * a = a$;

that is, $(A, *)$ is a semigroup possessing in addition a neutral element (which, as we have seen, is unique).
Rests on Definitions 4.9 and 4.18.

Returning to the earlier example, $(\mathbb{N}, +)$ is not a monoid, since that set contains no element serving as a neutral element for addition. However, if we adjoin the number zero to the naturals, that very element fulfils the property with respect to addition, so $(\mathbb{N} \cup \{0\}, +, 0)$ is a monoid. Another example, easy to check, is the monoid $(\mathbb{N}, \cdot, 1)$. Finally, the example of Equation (4.11) is furthermore an abelian monoid.

Definition 4.20 (Group)**.** An algebraic structure $(A, *)$ is a *group* if and only if

1. $\forall\ a, b \in A, \quad a * b \in A$;
2. $\forall\ a, b, c \in A, \quad a * (b * c) = (a * b) * c$;
3. $\forall\ a \in A,\ \exists\ e_A \in A \mid a * e_A = a \quad \wedge \quad e_A * a = a$;
4. $\forall\ a \in A,\ \exists\ a^{-1} \in A \mid a * a^{-1} = e_A \quad \wedge \quad a^{-1} * a = e_A$;

that is, $(A, *)$ is a monoid in which every element of the set moreover possesses an *inverse element* (and since associativity holds by axiom, the inverse of each element is unique).
Rests on Definitions 4.11 and 4.19.

As an example, upon adjoining the negative numbers to $(\mathbb{N} \cup \{0\}, +, 0)$ we obtain the group $(\mathbb{Z}, +, 0)$: every element $k \in \mathbb{Z}$ has its inverse element $-k$, giving for each of them $k + (-k) = -k + k = 0$. Another well-known example is the set $\{1, -1\} \subset \mathbb{Z}$ together with ordinary multiplication, whose internal operation is described by the table

$\cdot$	1	-1
1	1	-1
-1	-1	1

(4.12)

Here closure and associativity are clear. The (as we know, unique) neutral element is the number 1, which is thus an inverse element for itself, since $1 \cdot 1 = 1$. An inverse element for -1 is -1 itself, since $(-1) \cdot (-1) = 1$.

[1] The reader is invited to carry out the proof of associativity explicitly. To do so one must verify the eight possibilities $(0 \cdot 0) \cdot 0 \stackrel{?}{=} 0 \cdot (0 \cdot 0)$, $(0 \cdot 0) \cdot 1 \stackrel{?}{=} 0 \cdot (0 \cdot 1)$, etc.

4.2.2 Basic properties of groups

Since in a group we have associativity at our disposal, the following important properties valid in any group can be proved. Strictly speaking they hold in any structure with neutral element, inverses, and associative elements; but as we shall not encounter structures that fail to close, we state them here as properties of a group.

Proposition 4.21 (Uniqueness of the inverse element)**.** *Given an element of a group A, its inverse is unique.*

Rests on Definition 4.20.

Proof. *Derives Proposition 4.21.*

Let $a \in A$ and let a^{-1}, $\tilde{a}^{-1}$ be two inverse elements of a. By the axiom, $a * a^{-1} = e_A$ and $a * \tilde{a}^{-1} = e_A$, so that, operating on the left with $a^{-1}*$,

$$\begin{aligned} a * a^{-1} &= a * \tilde{a}^{-1} \\ a^{-1} * (a * a^{-1}) &= a^{-1} * (a * \tilde{a}^{-1}) \\ (a^{-1} * a) * a^{-1} &= (a^{-1} * a) * \tilde{a}^{-1} \\ e_A * a^{-1} &= e_A * \tilde{a}^{-1}\,, \end{aligned}$$

and thus $a^{-1} = \tilde{a}^{-1}$. □

Proposition 4.22 (Inverse of the neutral element)**.** *The inverse of the neutral element is the neutral element itself.*

Rests on Definition 4.20 and Proposition 4.21.

Proof. *Derives Proposition 4.22.*

By the neutral-element axiom we have $e_A * e_A = e_A$. Operating on both sides with $* e_A^{-1}$ we obtain $e_A * (e_A * e_A^{-1}) = e_A^{-1}$, that is, $e_A = e_A^{-1}$. Since the inverse is unique, we conclude that the inverse of the neutral element is itself. □

Proposition 4.23 (Neutral element equal to a product)**.** *In a group,*

$$e_A = a * b \iff a = b^{-1} \tag{4.13}$$

and analogously

$$e_A = a * b \iff b = a^{-1}\,. \tag{4.14}$$

Rests on Definition 4.20.

Proof. *Derives Proposition 4.23.*

Starting from $a * b = e_A$ and operating on the right with $* b^{-1}$,

$$\begin{aligned} a * b &= e_A \\ a * (b * b^{-1}) &= e_A * b^{-1} \\ a * e_A &= b^{-1} \\ a &= b^{-1}\,, \end{aligned}$$

and the converse and the second equivalence follow analogously. □

Proposition 4.24 (Involution of the inverse)**.** *The inverse of the inverse of an element is the element itself:*

$$\left(a^{-1}\right)^{-1} = a\,. \tag{4.15}$$

Rests on Proposition 4.23.

Proof. *Derives Proposition 4.24.*

In the context of the preceding proposition, from Equation (4.14) we have

$$b^{-1} = \left(a^{-1}\right)^{-1} ,$$

and substituting into Equation (4.13) we obtain $\left(a^{-1}\right)^{-1} = a$. □

Proposition 4.25 (Inverse of a product)**.** *Let $a, b \in G$. The inverse of the product of a and b is*

$$(a * b)^{-1} = b^{-1} * a^{-1} . \tag{4.16}$$

Rests on Definition 4.20 and Proposition 4.21.

Proof. *Derives Proposition 4.25.*

We have

$$\begin{aligned} (a * b) * (b^{-1} * a^{-1}) &= a * (b * b^{-1}) * a^{-1} \\ &= a * e_A * a^{-1} \\ &= e_A , \end{aligned}$$

and analogously on the other side. Hence, since the inverse is unique, $b^{-1} * a^{-1}$ is the inverse of $a * b$. □

4.3 Composite algebraic structures

4.3.1 Axioms for composite algebraic structures

Distributivity axioms

Let $(A, +)$ and $(V, \bar{+})$ be two algebraic structures, and consider further a map $\circ$ of the form

$$\begin{array}{rccc} \circ : & A \times V & \longrightarrow & V \\ & (a, v) & \longmapsto & a \circ v . \end{array}$$

Definition 4.26 (Distributivity)**.** We say that $\circ$ is *distributive* with respect to $\bar{+}$ if and only if

1. $\forall\ a \in A,\ \forall\ u, v \in V,\ a \circ (u \bar{+} v) = a \circ u \bar{+} a \circ v$;
2. $\forall\ a, b \in A,\ \forall\ v \in V,\ (a + b) \circ v = a \circ v \bar{+} b \circ v$.

Rests on Definition 4.3.

That the structure A acts from the left is only a convention, or the traditional usage; nothing prevents the creation of structures with distributivity from the left with respect to $\circ$ together with distributivity from the right with respect to $\circ$ (or even another binary map). In this book we shall use only left distributivity, and we shall mention it explicitly when absolutely necessary. Note also that if the structures $(A, +)$ and $(V, \bar{+})$ are identical — that is, $A = V$ and there is an isomorphism between $(A, +)$ and $(A, \bar{+})$ — then $\circ$ becomes an internal binary operation; in this context, if the magma $(A, \circ)$ is abelian, the two axioms above are completely equivalent.

Action axioms

Let $(A, *, e)$ be an algebraic structure with neutral element and X an arbitrary set. Consider the map $\bullet$ of the form

$$\begin{array}{rccc} \bullet: & A \times X & \longrightarrow & X \\ & (a, x) & \longmapsto & a \bullet x\,. \end{array}$$

Definition 4.27 (Action)**.** We say that $\bullet$ is an *action* of the structure A on the set X if and only if $\bullet$ satisfies

1. $\forall\ x \in X,\ e \bullet x = x$;
2. $\forall\ x \in X,\ \forall\ a, b \in A,\ (a * b) \bullet x = a \bullet (b \bullet x)$.

Rests on Definition 4.9.

4.3.2 Composite algebraic structures

Definition 4.28 (Ring)**.** Consider an abelian group $(A, +, 0_A)$ and a semigroup $(A, \circ)$. We say that $[(A, +, 0_A), \circ, (A, \circ)] \overset{\text{not.}}{=} (A, +, \circ)$ is a *ring* if and only if

1. $\forall\ a, b, c \in A, \quad a \circ (b + c) = a \circ b + a \circ c$;
2. $\forall\ a, b, c \in A, \quad (a + b) \circ c = a \circ c + b \circ c$;

that is, if and only if the property of *distributivity* of $\circ$ with respect to $+$ holds.
Rests on Definitions 4.8, 4.18, 4.20 and 4.26.

For example, in $\mathbb{Z}$ multiplication distributes with respect to addition, so $(\mathbb{Z}, +, \cdot)$ is a ring. Further:

- Upon imposing commutativity of the IBO $\circ$ — that is, if $(A, \circ)$ becomes an abelian semigroup — we say that $(A, +, \circ)$ is a *commutative ring*. $(\mathbb{Z}, +, \cdot)$ is a commutative ring.
- Imposing the existence of a neutral element 1_A for the semigroup — that is, if $(A, \circ, 1_A)$ becomes a monoid — we say that $(A, +, \circ)$ is a *ring with unity*. Since in $\mathbb{Z}$ the number 1 is a neutral element for multiplication, $(\mathbb{Z}, +, \cdot)$ is a (commutative) ring with unity. Another example is the real numbers: $(\mathbb{R}, +, 0)$ is an abelian group, and $(\mathbb{R}, \cdot, 1)$ is an abelian monoid. As is well known, multiplication distributes with respect to addition, and consequently $(\mathbb{R}, +, \cdot)$ is a (commutative) ring with unity.

Definition 4.29 (Module)**.** Consider a ring with unity $[(A, +, 0_A), \circ, (A, \circ, 1_A)] = (A, +, \circ)$ and an abelian group $(V, \bar{+}, 0_V)$. Let there be, in addition, the binary map

$$\begin{array}{rccc} \cdot: & A \times V & \longrightarrow & V \\ & (a, u) & \longmapsto & au\,. \end{array}$$

We say that $[(A, +, \circ), \cdot, (V, \bar{+})]$ is a *module* V *over* A if and only if the binary map $\cdot$ distributes with respect to $\bar{+}$ and, in addition, $\cdot$ is an action of the monoid $(A, \circ, 1_A)$ on the group $(V, \bar{+}, 0_V)$ regarded as a set. That is, if and only if

1. $\forall\ a \in A,\ \forall\ u, v \in V, \quad a(u \bar{+} v) = au \bar{+} av$;
2. $\forall\ a, b \in A,\ \forall\ u \in V, \quad (a + b)u = au \bar{+} bu$;
3. $\forall\ a, b \in A,\ \forall\ u \in V, \quad a(bu) = (a \circ b)u$;
4. $\forall\ u \in V, \quad 1_A u = u$.

Rests on Definitions 4.20 and 4.26 to 4.28.

Definition 4.30 (Field)**.** Consider a ring with unity $(A, +, \circ)$, that is, such that $(A, +, 0_A)$ is an abelian group and $(A, \circ, 1_A)$ is a monoid. We say that $[(A, +, 0_A), \circ, (A \setminus \{0_A\}, \circ, 1_A)] \overset{\text{not.}}{=} (A, +, \circ)$ is a *field* if for every element of the monoid $(A, \circ, 1_A)$ there exists an inverse element, except for the neutral element 0_A of $(A, +)$; that is, if in addition

1. $\forall\ a \in A \setminus \{0_A\},\ \exists\ a^{-1} \in A \mid a \circ a^{-1} = 1_A\ \wedge\ a^{-1} \circ a = 1_A$,

which means that the monoid $(A, \circ)$ almost becomes a group — as it would, completely, if an inverse existed for 0_A.

Rests on Definitions 4.11, 4.19 and 4.28.

- If the monoid $(A, \circ)$ is moreover abelian, we say that $(A, +, \circ)$ is a *commutative field*.

Definition 4.31 (Vector space)**.** Consider a module $[(A, +, \circ), \cdot, (V, \bar{+})]$. We say that $[(A, +, \circ), \cdot, (V, \bar{+})]$ is a *vector space V over A* if and only if the ring with unity $(A, +, \circ)$ is a commutative field; that is, if and only if the following further axioms hold:

1. $\forall\ a \in A \setminus \{0_A\},\ \exists\ a^{-1} \mid a \circ a^{-1} = 1_A\ \wedge\ a^{-1} \circ a = 1_A$;
2. $\forall\ a, b \in A,\ a \circ b = b \circ a$.

Rests on Definitions 4.8, 4.29 and 4.30.

As a characterization, a vector space is thus a commutative field (arising from a ring with unity in which inverses now exist except for 0_A) together with a distributive action of the field's monoid-with-inverses on an abelian group.

Vector spaces are very important in physics. In classical physics they manifest an abstract arena serving to model nature at each point of the Universe; in the context of quantum physics, they serve to model the possible states of a system. Even if a system is confined to some smaller set of states, we can model that smaller system as a vector-space substructure.

Definition 4.32 (Algebra)**.** Consider a vector space $[(A, +, \circ), \cdot, (V, \bar{+})]$, together with the IBO

$$\begin{array}{rccc} \times : & V \times V & \longrightarrow & V \\ & (u, v) & \longmapsto & u \times v\,. \end{array}$$

We say that $[(A, +, \circ), \cdot, (V, \bar{+}, \times)]$ is an *algebra V over A* if and only if the following axioms hold:

1. $\forall\ u, v, w \in V,\ u \times (v \bar{+} w) = u \times v \bar{+} u \times w$;
2. $\forall\ u, v, w \in V,\ (u \bar{+} v) \times w = u \times w \bar{+} v \times w$;
3. $\forall\ a \in A,\ \forall\ u, v \in V,\ a(u \times v) = (au) \times v = u \times (av)$.

Rests on Definitions 4.4, 4.26 and 4.31.

Thus the algebra $[(A, +, \circ), \cdot, (V, \bar{+}, \times)]$ is such that

- $(A, +, \circ)$ is a ring with unity;
- $(V, \bar{+})$ is an abelian group;
- $(V, \times)$ is a magma;
- A acts on V through the operation $\cdot$;
- various distributivity properties hold among the operations, as specified above.

Let us examine the properties that may be added, giving rise to remarkable kinds of algebras.

- If $(V, \bar{+}, \times)$ is a ring — that is, if $(V, \times)$ is moreover a semigroup — we say that $[(A, +, \circ), \cdot, (V, \bar{+}, \times)]$ is an *associative algebra*. There holds

  1. $\forall\ u, v, w \in V,\ u \times (v \times w) = (u \times v) \times w$.

- If $(V, \bar{+}, \times)$ is a ring with unity — that is, if $(V, \times)$ is moreover a monoid — we say that $[(A, +, \circ), \cdot, (V, \bar{+}, \times)]$ is a *unital associative algebra*. There holds

  1. $\forall\ u \in V,\ \exists\ \mathbb{1}_V \in V \mid \mathbb{1}_V \times u = u\ \wedge\ u \times \mathbb{1}_V = u$.

- If the axioms

  1. $\forall\ u, v \in V,\ u \times v = -v \times u$,
  2. $\forall\ u, v, w \in V,\ u \times (v \times w) \bar{+} v \times (w \times u) \bar{+} w \times (u \times v) = 0_V$

  hold, we say that $[(A, +, \circ), \cdot, (V, \bar{+}, \times)]$ is a *Lie algebra*.

Algebras are a most interesting algebraic structure for physics, since — like groups — they provide an elegant mathematical machinery for the symmetries of a physical system, classical or quantum.

4.4 Group theory: further definitions

Group theory is an immense branch of mathematics that studies these structures in detail. Because of the axioms imposed, groups are potentially applicable to situations in geometry and in nature that involve symmetry. There is a large literature on groups containing a finite number of elements. Such groups display a complexity in their structure that is tied directly to their cardinality and to the prime decomposition of the cardinal. There is a certain beauty in this: arithmetic appears unavoidably within the theory of algebraic structures, even though the two theories study different things. To organize these finite groups there are classification theorems for abelian groups, and for the non-abelian ones there are the Sylow theorems. A problem to solve may be, for example: find all abelian groups of cardinal 8.

4.4.1 Groups

Let us discuss the definition of a group here in more detail. Consider a set G, in general of finite or infinite cardinality, together with a binary operation $*$. We say that $(G, *)$ is a group if and only if it satisfies the following axioms.

1. Closure:
$$\forall\ a, b \in G,\ a * b \in G\,. \tag{4.17}$$
This property is exactly equivalent to having taken the binary operation $*$ to be an IBO, or the structure $(G, *)$ to be a magma.

2. Associativity:
$$\forall\ a, b, c \in G,\ a * (b * c) = (a * b) * c\,. \tag{4.18}$$

3. Existence of a neutral element:
$$\forall\ a \in G,\ \exists\ e \in G \mid a * e = a \quad \wedge \quad e * a = a\,. \tag{4.19}$$

4. Existence of inverse elements:

$$\forall\ a \in G,\ \exists\ a^{-1} \in G \mid a * a^{-1} = e \quad \wedge \quad a^{-1} * a = e\,. \tag{4.20}$$

Let us examine some examples whose underlying sets have infinite cardinality. The structure $(\mathbb{R}^*, \cdot)$ is a group whose neutral element, in multiplicative notation, is $1 \in \mathbb{R}$. On the other hand we can verify that $(\mathbb{Q}^*, \div)$, where

$$\begin{array}{rccc} \div : & \mathbb{Q}^* \times \mathbb{Q}^* & \longrightarrow & \mathbb{Q}^* \\ & (a, b) & \longmapsto & \dfrac{a}{b}\,, \end{array}$$

is not a group, since the associativity axiom fails: indeed,

$$\frac{\frac{a}{b}}{c} \neq \frac{a}{\frac{b}{c}}\,.$$

To denote a group we shall use the notations $(G, *)$ and $G = \{e, a, b, c, \ldots\}$, where e is the neutral element, though we may also use the index notation $G = \{g_i\}_{i \in I \subset \mathbb{N}}$. When no ambiguity is possible we shall denote the operation $\cdot$ by simple juxtaposition.

Definition 4.33 (Subgroup)**.** Consider a group $(G, *)$ and a subset $H \subseteq G$. We say that $(H, *)$ is a *subgroup* of $(G, *)$ if and only if

1. $\forall\ a, b \in H,\ a * b \in H$;
2. $\forall\ a \in H,\ \exists\ e_H \in H \mid a * e_H = a\ \wedge\ e_H * a = a$ — which is completely equivalent to saying that the neutral element of $(G, *)$ lies in H, since a neutral element of G, if it exists, is unique;
3. $\forall\ a \in H,\ \exists\ a^{-1} \in H \mid a * a^{-1} = e\ \wedge\ a^{-1} * a = e$ — which is equivalent to saying that the inverse elements of the elements of H belong to H.

Rests on Definition 4.20.

Since all the elements of the group $(G, *)$ satisfy associativity, we may dispense with that condition in asserting that a subgroup is indeed a group. In particular a subgroup is a substructure, for H is closed by definition. We therefore introduce the notation

$$(H, *) \text{ subgroup of } (G, *) \quad \overset{\text{not.}}{\iff} \quad H \leq G\,.$$

Given a group $G = \{e, a, \ldots\}$ it is clear that $\{e\} \leq G$ and that $G \leq G$; these subgroups are known as the *trivial subgroups*.

4.4.2 Cayley tables

Definition 4.34 (Cayley table)**.** The *Cayley table* of a finite group is the resource we shall use to symbolize how the elements of the group operate. Let $(\{e, a, b, c, \ldots\}, \cdot)$ be a group. Its Cayley table is given by

$\cdot$	e	a	b	c	$\ldots$
e	e	a	b	c	$\ldots$
a	a	a^2	ab	ac	$\ldots$
b	b	ba	b^2	bc	$\ldots$
c	c	ca	cb	c^2	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Rests on Definition 4.20.

The Cayley table of a group stores the whole "machinery" of the group: it contains all the information needed to describe the group's internal product explicitly. Two important observations about Cayley tables are the following.

- **The elements in any one row or column of a Cayley table cannot repeat.** Indeed, suppose that within one column we had $ab = x$ and also $ac = x$. This says $ab = ac$, and since we are in a group, a has an inverse. Multiplying on the left by the inverse of a we obtain $b = c$, which is a contradiction, since the elements of the group as a set must all be distinct.

- **For an abelian group the Cayley table is symmetric** with respect to the diagonal e, a^2, b^2, $c^2, \ldots$ — as can be seen trivially from the general Cayley table.

4.5 Basic groups

4.5.1 The first finite groups

The trivial group

Consider the algebraic structure $(\{e\}, *)$, where e satisfies the neutral-element axiom. Trivially this set satisfies the four group axioms; we call it the *trivial group*. Since every group contains a neutral element (by definition), given a group $(G, \cdot, e)$ we can always affirm that the set formed by the neutral element of G is a subgroup of G, that is, $\{e\} \leq G$. The Cayley table of the trivial group is

$*$	e
e	e

For example, consider $0 \in \mathbb{Z}$ together with the IBO $+$. We have

$+$	0
0	0

that is, the structure $(\{0\}, +)$ is a trivial group. This does not happen for an arbitrary IBO defined on $\mathbb{Z}$: for example, $(\{0\}, \cdot)$ is not a group, since 0 has no multiplicative inverse, so the inverse axiom fails.

The group of two elements

Consider the algebraic structure $(\{e, a\}, *)$, where e satisfies the neutral-element axiom. Let us see how the elements of $\{e, a\}$ must operate for the group axioms to be satisfied. Clearly we have

$*$	e	a
e	e	a
a	a	$a * a$

But what must the result of $a * a$ be? That is, supposing that $\{e, a, a * a\}$ is a group, let us deduce the conditions a must satisfy. The options are:

1. the option $a * a = a$ is not valid, since we know the elements of a row or column of a Cayley table cannot repeat;

2. on the other hand, to respect closure, $a * a = e$ is the only other option for $\{e, a\}$ to be a group of cardinal 2.

Thus the Cayley table of a group of two elements is

$*$	e	a
e	e	a
a	a	e

We have therefore proved that this is the only manner, or configuration, in which a set of two elements can form a group. For example, consider the elements $1, -1 \in \mathbb{Z}$ together with the IBO $\cdot$. We have

$\cdot$	1	-1
1	1	-1
-1	-1	1

that is, the structure $(\{1,-1\},\cdot)$ is a group of two elements. By contrast, $(\{1,-1\},+)$ is not a group, closure being clearly violated since in this structure $1+1=2$. Indeed, one may regard the question "how much is one plus one?" as a trick question, to be answered with the question "in which algebraic structure?".

The group of three elements

In this case we have

$*$	e	a	b
e	e	a	b
a	a	$a*a$	$a*b$
b	b	$b*a$	$b*b$

Since, however, the elements of a row or column cannot repeat, we conclude that $a*b=e$, and in the same way $b*a=e$. Filling in the remaining entries by the same logic we find $a*a=b$ and $b*b=a$, so the general Cayley table for a group of three elements is

$*$	e	a	b
e	e	a	b
a	a	b	e
b	b	e	a

which is clearly unique. As an example, the set of the three cube roots of unity together with complex multiplication forms a group, with the Cayley table of a group of three elements:

$\cdot$	1	$-\frac{1}{2}+\mathrm{i}\frac{\sqrt{3}}{2}$	$-\frac{1}{2}-\mathrm{i}\frac{\sqrt{3}}{2}$
1	1	$-\frac{1}{2}+\mathrm{i}\frac{\sqrt{3}}{2}$	$-\frac{1}{2}-\mathrm{i}\frac{\sqrt{3}}{2}$
$-\frac{1}{2}+\mathrm{i}\frac{\sqrt{3}}{2}$	$-\frac{1}{2}+\mathrm{i}\frac{\sqrt{3}}{2}$	$-\frac{1}{2}-\mathrm{i}\frac{\sqrt{3}}{2}$	1
$-\frac{1}{2}-\mathrm{i}\frac{\sqrt{3}}{2}$	$-\frac{1}{2}-\mathrm{i}\frac{\sqrt{3}}{2}$	1	$-\frac{1}{2}+\mathrm{i}\frac{\sqrt{3}}{2}$

4.5.2 The group $\mathbb{Z}$

In the discussion of the definition of a group we mentioned that $(\mathbb{Z},+,0)$ is a group. This group is very fruitful for explaining the ideas of subgroup and of quotient spaces, and we therefore study it in detail.

The subgroup $n\mathbb{Z}$

Let $n \in \mathbb{Z}$, and define the set $n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}$; that is, $n\mathbb{Z}$ consists of all the multiples of n within $\mathbb{Z}$. For example,

$$3\mathbb{Z} = \{\ldots,-9,-6,-3,0,3,6,9,\ldots\}, \quad \mathbb{Z} = \{\ldots,-3,-2,-1,0,1,2,3,\ldots\}\,.$$

Proposition 4.35. *$n\mathbb{Z}$ is a subgroup of $\mathbb{Z}$.*
Rests on Definition 4.33.

Proof. *Derives Proposition 4.35.*
We check the requirements:

1. $n\mathbb{Z} \neq \varnothing$, since $0 \in n\mathbb{Z}$ (take $k = 0$).

2. Trivially the neutral element 0 lies in $n\mathbb{Z}$, since the number zero is a multiple of every integer.

3. Finally, for every element of $n\mathbb{Z}$ there exists an inverse: if $a \in n\mathbb{Z}$ is a multiple of n, so is $-a$. Indeed, we may write $a = nl$ with $l \in \mathbb{Z}$; then, since distributivity holds in $\mathbb{Z}$,

$$\begin{aligned} a + (-a) &= nl + (-nl) \\ &= nl + n(-l) \\ &= n(l - l) \\ &= 0\,, \end{aligned}$$

so the inverse axiom is satisfied.

□

Congruence modulo n

As human beings we do not count the hours from the beginning of the Universe. As a reference for indicating clock time we speak of k hours past midnight: a day is completed when 24 hours have elapsed since midnight, and hour number 25 we begin to call hour 1 again. We shall build definitions that say that 25 is mathematically equivalent to 1 (though not equal), every 24 hours.

Let $k, l \in \mathbb{Z}$, and define the relation

$$k \sim_n l \iff k - l \in n\mathbb{Z}\,. \tag{4.21}$$

Proposition 4.36. *The relation Equation (4.21) is an equivalence relation.*
Rests on Definitions 3.58 and 4.20.

Proof. *Derives Proposition 4.36.*
We verify the three properties of Definition 3.58:

1. $k \sim_n k$, since $k - k = 0 \in n\mathbb{Z}$; reflexivity therefore holds.

2. If $k \sim_n l$, we have

$$\begin{aligned} k - l \in n\mathbb{Z} &\iff k - l = np,\ p \in \mathbb{Z}, \\ &\iff l - k = n(-p),\ p \in \mathbb{Z}, \\ &\iff l - k = nq,\ q \in \mathbb{Z}\,, \end{aligned}$$

because $\mathbb{Z}$ is a group, so there exists $-p = q \in \mathbb{Z}$. Thus

$$k \sim_n l \implies l \sim_n k\,,$$

and symmetry holds.

3. If $k \sim_n l$ and $l \sim_n p$, we have

$$k - l = nq,\ q \in \mathbb{Z},$$
$$l - p = nr,\ r \in \mathbb{Z}\,.$$

Adding the equations member by member gives $k - p = n(q + r)$, and by closure $q + r = s \in \mathbb{Z}$; hence $k - p \in n\mathbb{Z}$. Thus

$$k \sim_n l \ \wedge\ l \sim_n p \implies k \sim_n p\,,$$

and transitivity holds.

□

We have proved that the relation defined is an equivalence relation. We shall say that two elements equivalent under this relation are *congruent modulo* n. In the earlier example, 25 is congruent to 1 modulo 24.

The quotient space $\mathbb{Z}/n\mathbb{Z}$

As we have seen, congruence modulo n is an equivalence relation. Let $k \in \mathbb{Z}$. We denote by $\overline{k}$ the class of k, given by

$$\overline{k} = \{l \in \mathbb{Z} \mid l \sim_n k\} = \{l \in \mathbb{Z} \mid l - k \in n\mathbb{Z}\} = \{\text{multiples of } n, \text{ plus } k\}\,.$$

As we know, the quotient space is the set of all the classes. Let us analyse the possible classes making up our quotient space:

- $\overline{0}$ consists of the multiples of n;
- $\overline{1}$ consists of the multiples of n, plus 1;
- $\overline{n}$ consists of the multiples of n, plus n — that is, simply the multiples of n. Thus

$$\overline{0} = \overline{n}\,. \tag{4.22}$$

- $\overline{n+1}$ consists of the multiples of n, plus $n + 1$ — that is, simply the multiples of n, plus 1.

Hence the elements of our quotient space are the classes (none repeated)

$$\frac{\mathbb{Z}}{n\mathbb{Z}} = \{\overline{0}, \ldots, \overline{n-1}\}\ . \tag{4.23}$$

The set $\mathbb{Z}/n\mathbb{Z}$ as an abelian group

Define the internal binary operation

$$\begin{aligned} \overline{+}: \ \frac{\mathbb{Z}}{n\mathbb{Z}} \times \frac{\mathbb{Z}}{n\mathbb{Z}} &\longrightarrow \frac{\mathbb{Z}}{n\mathbb{Z}} \\ (\overline{k}, \overline{l}) &\longmapsto \overline{k} \,\overline{+}\, \overline{l} = \overline{k+l}\,. \end{aligned}$$

Proposition 4.37. $\left(\frac{\mathbb{Z}}{n\mathbb{Z}}, \overline{+}, \overline{0}\right)$ *is an abelian group.*
Rests on Definitions 4.8 and 4.20 and Proposition 4.36.

Proof. *Derives Proposition 4.37.*
To avoid overloading the notation we write $\overline{+}$ simply as $+$. We have:

1. There is closure by construction: the class $\overline{k+l} = \overline{k} + \overline{l}$ always lies in the quotient space, since the class of $k+l$ can always be written as the class of n plus something, and finally as the class of that something. For example, for $n = 5$,
$$\overline{3} + \overline{8} = \overline{3+8} = \overline{11} = \overline{5+6} = \overline{6} = \overline{5+1} = \overline{1} \in \frac{\mathbb{Z}}{5\mathbb{Z}}\,.$$
2. There is associativity, since
$$\begin{aligned}(\overline{k} + \overline{l}) + \overline{m} &= \overline{k+l} + \overline{m} \\ &= \overline{(k+l)+m} \\ &= \overline{k+(l+m)} \\ &= \overline{k} + \overline{l+m} \\ &= \overline{k} + (\overline{l} + \overline{m})\,.\end{aligned}$$
3. The neutral element $\overline{0}$ exists: indeed, $\overline{k} + \overline{0} = \overline{k+0} = \overline{k}$.
4. An inverse element $\overline{-k}$ exists for each element $\overline{k}$ of the set: indeed, $\overline{k} + \overline{-k} = \overline{k-k} = \overline{0}$. For example, for $n = 5$ we have
$$\frac{\mathbb{Z}}{5\mathbb{Z}} = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}\}\ ,$$
and the inverse element of $\overline{1}$ is $\overline{4}$, that is, $\overline{-1} = \overline{4}$ — not by imposition, but by the definition of the class: the multiples of 5, minus 1, are the same numbers (the same class) as the multiples of 5, plus 4.
5. The IBO $\overline{+}$ is commutative, owing to the commutativity of $\mathbb{Z}$. Indeed,
$$\begin{aligned}\overline{k} + \overline{l} &= \overline{k+l} \\ &= \overline{l+k} \\ &= \overline{l} + \overline{k}\,.\end{aligned}$$

□

In conclusion, $\left(\frac{\mathbb{Z}}{n\mathbb{Z}}, \overline{+}, \overline{0}\right)$ is an abelian group, distinct from the group $(\mathbb{Z}, +, 0)$.

4.6 Functions on groups

The group of maps into a group

The construction this section needs has already been carried out in the generality of arbitrary algebraic structures: the pointwise product of Section 4.1.2, the neutral map Θ, and the opposite map Equation (4.6) make $\mathcal{F}(A, B)$ a group whenever B is one, by Proposition 4.14. Taking A to be a group G and B a group H costs nothing extra and adds nothing new: G's own operation plays no part, because the pointwise construction never operates on the argument.

What is genuinely new here is that the maps of interest between groups are not all of $\mathcal{F}(G, H)$ but the homomorphisms, and those are a different matter.

Proposition 4.38 (When the homomorphisms form a group)**.** *Let G and H be groups and let* $\mathrm{Hom}(G, H)$ *denote the set of group homomorphisms from G to H. If H is abelian, then* $\mathrm{Hom}(G, H)$ *is a subgroup of $(\mathcal{F}(G, H), *)$, and it is abelian. If H is not abelian, then* $\mathrm{Hom}(H, H)$ *is not even closed under $*$.*

Rests on Definitions 4.33 and 4.39 and Propositions 4.14 and 4.25.

Proof. *Derives Proposition 4.38.*

Let $f, g \in \mathrm{Hom}(G,H)$ and $a, b \in G$. Expanding both sides,

$$\begin{aligned}(f*g)(a *_G b) &= f(a *_G b) *_H g(a *_G b)\\ &= f(a) *_H f(b) *_H g(a) *_H g(b)\,,\\ (f*g)(a) *_H (f*g)(b) &= f(a) *_H g(a) *_H f(b) *_H g(b)\,.\end{aligned}$$

The two agree for all a, b precisely when $f(b) *_H g(a) = g(a) *_H f(b)$, that is, when the images of f and g commute with one another.

If H is abelian this holds automatically, so $f * g$ is again a homomorphism. The neutral map Θ is a homomorphism, since $\Theta(a *_G b) = e_H = e_H *_H e_H = \Theta(a) *_H \Theta(b)$; and f^{op} is one because $f^{\mathrm{op}}(a *_G b) = (f(a) *_H f(b))^{-1} = f(b)^{-1} *_H f(a)^{-1} = f^{\mathrm{op}}(a) *_H f^{\mathrm{op}}(b)$, the last step using commutativity again. So $\mathrm{Hom}(G,H)$ contains the neutral element and is closed under the operation and under opposites, which is what it is to be a subgroup; it is abelian because $\mathcal{F}(G,H)$ is, by Proposition 4.14.

If H is not abelian, pick $a, b \in H$ with $a *_H b \neq b *_H a$ and take $f = g = \mathrm{id}_H$, which is certainly a homomorphism. The displayed condition then reads $b *_H a = a *_H b$, which fails, so $\mathrm{id}_H * \mathrm{id}_H$ is not a homomorphism and $\mathrm{Hom}(H,H)$ is not closed. □

The proposition is the reason the theory of group homomorphisms does not reduce to the theory of maps. For abelian targets one may add homomorphisms and stay inside the class — which is what makes $\mathrm{Hom}(G,\mathbb{R})$ and its relatives vector spaces, and ultimately what makes cohomology groups groups. For non-abelian targets the pointwise product leaves the class immediately, and the structures that survive are built from composition instead. The remainder of this section develops the notions that do work in both cases.

4.6.1 Homomorphisms

Definition 4.39 (Group homomorphism)**.** Consider the groups $(G_1, *_1)$ and $(G_2, *_2)$, together with a map $f : G_1 \longrightarrow G_2$. We say that f is a *group homomorphism* from G_1 to G_2 if and only if $f(a_1 *_1 b_1) = f(a_1) *_2 f(b_1)$ — that is, if between the two distinct groups f preserves the multiplication law.

Rests on Definitions 3.43 and 4.20.

Definition 4.40 (Group isomorphism)**.** Consider the groups $(G_1, *_1)$ and $(G_2, *_2)$ and let $f : G_1 \longrightarrow G_2$ be a group homomorphism. We say that f is an *isomorphism* from G_1 to G_2 if and only if the homomorphism is a bijection; we then say that G_1 and G_2 are *isomorphic*.

Rests on Definitions 3.47 and 4.39.

Isomorphism between groups is of great importance, because it is the concept expressing that two groups operate in exactly the same way even though their elements are distinct. What totally determines a group is the operation among its elements — its Cayley table — and what an isomorphism does is precisely to preserve the operation. For example, consider the set $\{1, \mathrm{i}, -1, -\mathrm{i}\}$, where $\mathrm{i} = \sqrt{-1} \in \mathbb{C}$. It is easy to verify that this set, together with the usual multiplication, forms an abelian group whose Cayley table is

$\cdot$	1	i	-1	$-\mathrm{i}$
1	1	i	-1	$-\mathrm{i}$
i	i	-1	$-\mathrm{i}$	1
-1	-1	$-\mathrm{i}$	1	i
$-\mathrm{i}$	$-\mathrm{i}$	1	i	-1

On the other hand, it is easy to compute the Cayley table of the (abelian) group $\frac{\mathbb{Z}}{4\mathbb{Z}}$:

$\overline{+}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$
$\overline{0}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$
$\overline{1}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{0}$
$\overline{2}$	$\overline{2}$	$\overline{3}$	$\overline{0}$	$\overline{1}$
$\overline{3}$	$\overline{3}$	$\overline{0}$	$\overline{1}$	$\overline{2}$

Comparing the two tables, we see that these groups are essentially the same thing: they are exactly the same up to the symbols we have used to designate the elements of the sets. This may seem magical, but so it is: taking two sets with no relation whatsoever and adjoining the group axioms, an intrinsic operational structure appears when the sets are viewed as algebraic structures. The situation has a precise mathematical underpinning: these groups are isomorphic. This is the true meaning of isomorphisms. The explicit form of the isomorphism between $\{1, \mathrm{i}, -1, -\mathrm{i}\}$ and $\frac{\mathbb{Z}}{4\mathbb{Z}}$ is given by $f(x) = \overline{-(2\mathrm{i}/\pi)\ln x}$ with $x \in \{1, \mathrm{i}, -1, -\mathrm{i}\}$, resulting in

$$\begin{array}{ccc} & f & \\ 1 & \longmapsto & \overline{0} \\ \mathrm{i} & \longmapsto & \overline{1} \\ -1 & \longmapsto & \overline{2} \\ -\mathrm{i} & \longmapsto & \overline{3}\,. \end{array}$$

One verifies, moreover, the preservation of the product law:

$$\begin{aligned} f(x \cdot y) &= \overline{-(2\mathrm{i}/\pi)\ln(x \cdot y)} \\ &= \overline{-(2\mathrm{i}/\pi)(\ln x + \ln y)} \\ &= \overline{-(2\mathrm{i}/\pi)\ln x} \,\overline{+}\, \overline{-(2\mathrm{i}/\pi)\ln y} \\ &= f(x) \,\overline{+}\, f(y)\,. \end{aligned}$$

Definition 4.41 (Group automorphism)**.** Let $f : G_1 \longrightarrow G_2$ be a group isomorphism. We say that f is an *automorphism* if and only if $G_1 = G_2$ — that is, the homomorphism f is moreover a bijection sending elements of a group to the group itself.

Rests on Definition 4.40.

Since every group is isomorphic to itself, we can say that for every group there exists a trivial automorphism sending each element to itself; that is, the bijection

$$\begin{array}{rccc} f: & G & \longrightarrow & G \\ & e & \longmapsto & e \\ & a & \longmapsto & a \\ & b & \longmapsto & b \\ & & \vdots & \end{array}$$

is a (trivial, but canonical) example of an automorphism.

4.6.2 Properties of homomorphisms

Proposition 4.42 (Image of the neutral element)**.** *Under a group homomorphism, the image of the neutral element is the neutral element of the codomain:*

$$f(e_1) = e_2\,. \tag{4.24}$$

Rests on Definition 4.39.

Proof. *Derives Proposition 4.42.*

For the neutral element of the domain we have $f(e_1) = f(e_1 *_1 e_1)$, and since f is a homomorphism, operating with $*_2 (f(e_1))^{-1}$,

$$\begin{aligned} f(e_1) &= f(e_1) *_2 f(e_1) \\ f(e_1) *_2 (f(e_1))^{-1} &= f(e_1) \\ e_2 &= f(e_1)\,. \end{aligned}$$

□

Proposition 4.43 (Image of the inverse element)**.** *Analogously to the preceding case, the image of an inverse element of the domain is the inverse of the image in the codomain:*

$$f(a_1^{-1}) = (f(a_1))^{-1}\,. \tag{4.25}$$

Rests on Definition 4.39 and Proposition 4.42.

Proof. *Derives Proposition 4.43.*

Indeed, operating on the left with $(f(a_1))^{-1} *_2$,

$$\begin{aligned} f(e_1) &= f(a_1 *_1 a_1^{-1}) \\ e_2 &= f(a_1) *_2 f(a_1^{-1}) && \text{by Equation (4.24)} \\ (f(a_1))^{-1} &= f(a_1^{-1})\,. \end{aligned}$$

□

4.7 The symmetric group

4.7.1 The symmetric group

Definition 4.44 (Symmetric group)**.** Consider a set X, and let S_X be the set of all the possible bijections from X to X, that is,

$$S_X = \{f \mid f : X \longrightarrow X \text{ is a bijection}\}\,. \tag{4.26}$$

Rests on Definition 3.47.

Let us recall the following properties (Section 3.3):

1. the composition of two bijective maps is a bijective map;
2. the composition of maps is associative — in particular, so is the composition of bijective maps;
3. there exists the identity map, whose composition with any map returns that same map — in particular, this works for any bijective map;
4. every bijective map possesses an inverse map.

From this we may conclude that $(S_X, \circ)$ is a group under composition of maps, which we shall call the *symmetric group*.

4.7.2 The permutation group

Definition 4.45 (Permutation group)**.** We define the *permutation group* as a symmetric group S_G in which G is a finite group. Thus

$$S_G = \{f \mid f : G \longrightarrow G \text{ is a bijection, } |G| \in \mathbb{N}\} \,. \tag{4.27}$$

Rests on Definitions 4.20 and 4.44.

Consider $G = \{g_i\}_{i=1}^{n}$. As discussed in the definition of the symmetric group, the elements of S_G are bijections. Since G here has a finite number of elements, a possible bijection is, for example, for $n = 5$,

$$\begin{array}{ccc} & f & \\ g_1 & \longmapsto & g_4 \\ g_2 & \longmapsto & g_3 \\ g_3 & \longmapsto & g_2 \\ g_4 & \longmapsto & g_5 \\ g_5 & \longmapsto & g_1 \end{array}$$

that is, in some manner f permutes the elements. We can symbolize this permutation as

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 3 & 2 & 5 & 1 \end{pmatrix} .$$

As an alternative notation for permutations we introduce

$$\sigma_a(g_i) = a g_i \overset{\text{not.}}{=} g_{a_i} \,, \tag{4.28}$$

where the index a_i denotes the new place of the element. For example, one may have $a g_2 = g_1$ (that is, $a_2 = 1$).

Proposition 4.46 (Cardinal of the permutation group)**.** $|S_G| = |G|!$.

Rests on Definition 4.45.

Proof. *Derives Proposition 4.46.*

Write $n = |G|$ and $G = \{g_1, \ldots, g_n\}$. For $0 \le k \le n$ let A_k denote the set of *injective* maps from $\{g_1, \ldots, g_k\}$ into G, so that A_0 has exactly one element, the empty map. We claim that

$$|A_k| = n(n-1)\cdots(n-k+1) \,, \qquad 0 \le k \le n \,, \tag{4.29}$$

the right-hand side being read as 1 when $k = 0$.

The claim holds for $k = 0$. Assume it for some $k < n$ and consider the restriction map $A_{k+1} \longrightarrow A_k$ that forgets the value at g_{k+1}. An injective $u \in A_k$ takes k pairwise distinct values, so the injective extensions of u to $\{g_1, \ldots, g_{k+1}\}$ are exactly the maps sending g_{k+1} to one of the elements of $G \setminus \{u(g_1), \ldots, u(g_k)\}$, a set of $n - k$ elements. Every $u \in A_k$ therefore has exactly $n - k$ preimages under the restriction map, and every element of A_{k+1} has exactly one image, whence

$$|A_{k+1}| = (n-k)\ |A_k| = n(n-1)\cdots(n-k) \,,$$

which is Equation (4.29) for $k+1$. By induction it holds for every $k \le n$; in particular

$$|A_n| = n(n-1)\cdots 2 \cdot 1 = n! \,.$$

It remains to identify A_n with S_G. An element of A_n is an injective map $f : G \longrightarrow G$; its image has exactly n elements, and the only subset of the n-element set G with n elements is G itself, so f is also surjective and hence a bijection. Conversely every bijection of G is injective. Thus $A_n = S_G$ and $|S_G| = n! = |G|!$. □

The permutation group of order 3!

The smallest symmetric group that is not abelian is the one built on a three-element set. Take $X = \{1, 2, 3\}$ and write $S_3 = S_X$; by Proposition 4.46 it has $3! = 6$ elements. It is small enough to be written out completely, and large enough that everything the rest of this chapter defines — subgroups, conjugacy, normality, quotients — can be read off its table.

Notation 4.47 (Cycle notation). Let $i_1, \dots, i_r$ be distinct elements of a finite set X. The *cycle* $(i_1\, i_2\, \dots\, i_r)$ is the bijection of X sending $i_1 \longmapsto i_2$, $i_2 \longmapsto i_3$, …, $i_r \longmapsto i_1$ and leaving every other element of X fixed; a cycle of length two is a *transposition*. A permutation is also written in two-line form, the second line listing the images of the entries of the first. Products are compositions of maps, so fg means $f \circ g$: the right-hand factor acts first.

Rests on Definitions 3.47 and 4.44.

In this notation the six elements of S_3 are

$$\begin{aligned} e &= \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \qquad r = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} = (1\,2\,3), \quad r^2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} = (1\,3\,2), \\ t_1 &= \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} = (2\,3), \quad t_2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} = (1\,3), \quad t_3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} = (1\,2). \end{aligned} \tag{4.30}$$

Three of them are transpositions, two are three-cycles, and $r^3 = e$, so $r^{-1} = r^2$ and $t_i^{-1} = t_i$ for each i. Composing the maps of Equation (4.30) pair by pair gives the Cayley table (Definition 4.34) shown in Table 4.1.

Table 4.1: The Cayley table of S_3 in the notation of Equation (4.30). The entry in row f and column g is the composite $f \circ g$, in which g acts first. Every row and every column is a permutation of the six elements, as Section 4.4.2 requires; the table is *not* symmetric about its diagonal — $r\,t_1 = t_3$ while $t_1 r = t_2$ — which is the whole content of Proposition 4.48. The upper-left 3×3 block is closed, and is the group of three elements of Section 4.5.1.

$\circ$	e	r	r^2	t_1	t_2	t_3
e	e	r	r^2	t_1	t_2	t_3
r	r	r^2	e	t_3	t_1	t_2
r^2	r^2	e	r	t_2	t_3	t_1
t_1	t_1	t_2	t_3	e	r	r^2
t_2	t_2	t_3	t_1	r^2	e	r
t_3	t_3	t_1	t_2	r	r^2	e

Proposition 4.48 (S_3 is a smallest non-abelian group)**.** *Every group with fewer than six elements is abelian, and S_3 has six elements and is not abelian.*

Rests on Definitions 4.8 and 4.20 and Proposition 4.46.

Proof. *Derives Proposition 4.48.*

That S_3 has six elements is Proposition 4.46, and Table 4.1 exhibits $r\,t_1 = t_3 \neq t_2 = t_1 r$, so S_3 is not abelian.

For the first assertion, let G be a group that is not abelian and choose $x, y \in G$ with $xy \neq yx$. We show that G has at least six elements. First, the five elements

$$e, \quad x, \quad y, \quad xy, \quad yx$$

are pairwise distinct. Indeed $x \neq e$ and $y \neq e$, since the neutral element commutes with everything, and $x \neq y$, since every element commutes with itself. If $xy = e$ then $y = x^{-1}$

by Proposition 4.23, and then $yx = e = xy$, excluded; the same argument rules out $yx = e$. Cancelling x on the left in $xy = x$, or y on the right in $xy = y$, would give $y = e$ or $x = e$, and likewise for yx; finally $xy \neq yx$ by choice.

It remains to produce a sixth element. Consider x^2. It cannot equal x (that gives $x = e$), nor xy (that gives $x = y$), nor yx (cancelling x on the right gives $x = y$), nor y — for $x^2 = y$ would make $xy = x^3 = yx$. So either x^2 lies outside the five elements listed, and G has at least six, or $x^2 = e$. Applying the same argument to y, we may assume $x^2 = y^2 = e$, so that $x^{-1} = x$ and $y^{-1} = y$. Then xyx is the sixth element: $xyx = e$ gives $yx = x^{-1} = x$ and so $y = e$; $xyx = x$ gives $yx = e$ and so $y = x^{-1} = x$; $xyx = y$ gives $xy = yx^{-1} = yx$; $xyx = xy$ gives $x = e$; and $xyx = yx$ gives $xy = y$ and so $x = e$ — every case being excluded. Hence $|G| \geq 6$. □

Proposition 4.49 (The subgroups of S_3)**.** *S_3 has exactly six subgroups: the two trivial ones $\{e\}$ and S_3; the three two-element subgroups $\{e, t_i\}$, $i = 1, 2, 3$; and the three-element subgroup*

$$A_3 = \left\{e, r, r^2\right\} . \tag{4.31}$$

Rests on Definitions 4.33 and 4.45.

Proof. *Derives Proposition 4.49.*

Each of the six listed sets is a subgroup: $\{e\}$ and S_3 trivially, $\{e, t_i\}$ because $t_i\, t_i = e$, and A_3 because the upper-left 3×3 block of Table 4.1 takes values in A_3 and contains e, with $r^{-1} = r^2$.

Conversely let $H \leq S_3$ be a subgroup and suppose $H \neq \{e\}$. If H contains r or r^2 then, being closed and containing inverses, it contains $\{e, r, r^2\} = A_3$. If in addition H contains some t_i, then it contains the three products $r\, t_i$, $r^2 t_i$ and t_i, which by Table 4.1 are the three distinct transpositions; hence $H = S_3$. So a subgroup containing a three-cycle is A_3 or S_3.

Suppose instead that H contains no three-cycle. Then H consists of e and transpositions. It cannot contain two distinct transpositions: the product of two distinct transpositions is, by Table 4.1, a three-cycle, which would then lie in H. Hence $H = \{e, t_i\}$ for a single i. The list is therefore complete. □

Remark 4.50. Anticipating Definition 4.56, A_3 is normal in S_3: conjugation by an element of A_3 maps A_3 into itself because A_3 is a subgroup, and Table 4.1 gives $t_1 r\, t_1^{-1} = t_1 r\, t_1 = r^2$, with the same outcome for t_2 and t_3. The two-element subgroups are not normal: the same table gives $r\, t_1 r^{-1} = r\, t_1 r^2 = t_2$, which is outside $\{e, t_1\}$. The three of them are therefore conjugate to one another in the sense of Definition 4.55. Moreover S_3 is the smallest group possessing a subgroup that is not normal: in an abelian group $aHa^{-1} = \{aha^{-1} \mid h \in H\} = \{h \mid h \in H\} = H$ for every subgroup H and every a, so all subgroups are normal, and by Proposition 4.48 every group of fewer than six elements is abelian.

The elements of S_3 are not merely relabellings: they are the symmetries of a figure.

Proposition 4.51 (S_3 and the equilateral triangle)**.** *Let $\Delta \subset \mathbb{R}^2$ be an equilateral triangle with vertices labelled $1, 2, 3$, and let T be the set of isometries of the plane that map Δ onto itself, a group under composition. The map sending each element of T to the permutation it induces on the vertex labels is an isomorphism $T \longrightarrow S_3$.*

Rests on Definition 4.40 and Equation (4.30).

Proof. *Derives Proposition 4.51.*

An isometry mapping Δ onto itself maps the vertex set onto itself, since a vertex is characterized metrically as a point of Δ at maximal distance from the centroid; call $\pi(u) \in S_3$ the resulting permutation of the labels. Composition of isometries induces composition of the permutations, so π is a homomorphism (Definition 4.39).

π is injective. If $\pi(u) = e$ then u fixes the three vertices, and these are not collinear; since a point of the plane is determined by its distances to three non-collinear points, and u preserves all distances, u fixes every point of the plane. So $\ker \pi = \{\mathrm{id}\}$, and Proposition 4.17 gives injectivity.

π is surjective. The rotation through $2\pi/3$ about the centroid of Δ carries Δ onto itself and induces the three-cycle r of Equation (4.30); its square induces r^2. The reflection in the median through the vertex i carries Δ onto itself and induces the transposition t_i. Together with the identity these six isometries realize all six elements of S_3, so π is onto and hence an isomorphism. □

The six isometries just exhibited are all of T, since π is a bijection. This is the first instance in this chapter of the pattern that makes group theory the language of symmetry in physics: an abstract group defined by a multiplication table turns out to *be* the symmetry group of a concrete object, and the two descriptions carry exactly the same information.

Theorem 4.52 (Cayley's theorem)**.** *Every finite group of order n is isomorphic to a subgroup of order n of the permutation group S_n.*
Rests on Definitions 4.33, 4.40 and 4.45.

Proof. *Derives Theorem 4.52.*
Let $(G, \cdot, e)$ be a group with $|G| = n$. For each $a \in G$ define the *left translation*

$$\begin{array}{rccc} p_a : & G & \longrightarrow & G \\ & g & \longmapsto & ag\,. \end{array} \tag{4.32}$$

Each p_a is a bijection. It is a map into G by closure. It is injective: if $ag = ag'$, then operating on the left with a^{-1} and using associativity gives $g = g'$. It is surjective: given $h \in G$, the element $a^{-1}h$ lies in G and $p_a(a^{-1}h) = a(a^{-1}h) = h$. By Proposition 3.48 it is therefore a bijection, so $p_a \in S_G$; its inverse map is $p_{a^{-1}}$.

The assignment $a \longmapsto p_a$ is a homomorphism. For $a, b, g \in G$,

$$p_{ab}(g) = (ab)g = a(bg) = p_a\left(p_b(g)\right) = \left(p_a \circ p_b\right)(g)\,,$$

and since g was arbitrary, $p_{ab} = p_a \circ p_b$. Writing $P : G \longrightarrow S_G$ for the map $a \longmapsto p_a$, this says precisely that P is a group homomorphism into $(S_G, \circ)$ (Definition 4.39).

P is injective. If p_a is the identity permutation then in particular $ae = e$, so $a = e$. Hence $\ker P = \{e\}$, and Proposition 4.17 gives injectivity.

The image is a subgroup of order n. The set $P(G) = \{p_a \mid a \in G\}$ contains $p_e = \mathrm{id}_G$, is closed under composition because $p_a \circ p_b = p_{ab}$, and contains the inverse $p_a^{-1} = p_{a^{-1}}$ of each of its elements; it is therefore a subgroup of S_G (Definition 4.33). Since P is injective, $P(G)$ has exactly n elements, and $P : G \longrightarrow P(G)$ is a bijective homomorphism, that is, an isomorphism (Definition 4.40).

Finally, any bijection $\varphi : G \longrightarrow \{1, \ldots, n\}$ induces an isomorphism $S_G \longrightarrow S_n$ by $f \longmapsto \varphi \circ f \circ \varphi^{-1}$ — a bijection with inverse $h \longmapsto \varphi^{-1} \circ h \circ \varphi$, and a homomorphism because the two inner factors $\varphi^{-1}\varphi$ cancel in a product. It carries $P(G)$ to a subgroup of S_n of order n isomorphic to G. □

4.8 Decompositions of a group

4.8.1 Conjugacy classes

Consider a group G. Define the relation

$$a \sim b \iff \exists\, p \in G \mid b = pap^{-1}, \quad a, b \in G\,. \tag{4.33}$$

Proposition 4.53. *The relation Equation (4.33) is an equivalence relation.*
Rests on Definitions 3.58 and 4.20 and Proposition 4.25.

Proof. *Derives Proposition 4.53.*
We verify the three properties:

1. Reflexivity holds, since
$$a = eae = eae^{-1},$$
that is, thanks to the existence of the neutral element, $a \sim a$.

2. If $a \sim b$, it is because there exists $p \in G$ such that $b = pap^{-1}$. Multiplying on the right by p and then on the left by p^{-1},
$$\begin{aligned} b &= pap^{-1} \\ bp &= pa \\ p^{-1}bp &= a\,. \end{aligned}$$
Hence, by the existence of inverse elements in G, there exists $q = p^{-1}$ such that $a = qbq^{-1}$, so $b \sim a$; symmetry therefore holds.

3. Supposing $a \sim b$ and $b \sim c$, there exist $p, q \in G$ such that
$$b = pap^{-1}, \quad c = qbq^{-1},$$
and substituting the first into the second,
$$\begin{aligned} c &= q(pap^{-1})q^{-1} \\ &= (qp)a(p^{-1}q^{-1}) \\ &= (qp)a(qp)^{-1} && \text{by Equation (4.16)}\,. \end{aligned}$$
Hence by closure there exists $r = qp \in G$ such that $c = rar^{-1}$, so $a \sim c$, and transitivity holds.

□

Definition 4.54 (Conjugacy classes)**.** We call *conjugacy classes* the equivalence classes given by the relation Equation (4.33). We say that two elements of a conjugacy class are *conjugate*; that is, a and b are conjugate if and only if $a \sim b$. We have
$$[a] = \{x \in G \mid x \sim a\} = \{pap^{-1} \mid p \in G\}\,.$$
Rests on Proposition 4.53 and Definition 3.59.

Definition 4.55 (Conjugate subgroups)**.** Consider the group G and $H, K \leq G$. We say that K and H are *conjugate subgroups* if and only if
$$h \sim k, \quad \forall\, h \in H,\ \forall\, k \in K\,,$$
that is,
$$\begin{aligned} K &= \{k \mid k \in K\} \\ &= \left\{php^{-1} \mid h \in H,\, p \in G\right\}, \end{aligned}$$
which we shall symbolize simply as $K = pHp^{-1}$, $p \in G$.
Rests on Definitions 4.33 and 4.54.

Definition 4.56 (Normal subgroup)**.** Let G be a group and $H \leq G$. We say that H is a *normal* (or *invariant*) subgroup in G if and only if it is conjugate to itself by all the elements of the group, that is, if $H = pHp^{-1}$ for all $p \in G$. We introduce the notation

$$H \text{ normal in } G \quad \overset{\text{not.}}{\iff} \quad H \trianglelefteq G\,.$$

Rests on Definitions 4.33 and 4.55.

Given a group G, it is easy to see that the trivial subgroups are also normal subgroups in G; that is, $\{e\} \trianglelefteq G$ and $G \trianglelefteq G$. It is worth mentioning that in the group-theory literature there are many ways of defining what a normal subgroup is; one may define it as one wishes, provided the other definitions are mentioned as possible characterizations.

4.8.2 Cosets

Consider the group $(G, \cdot, e)$ and $H \leq G$. Define the relation

$$a \sim_H b \iff \exists\, h \in H \mid b = ah, \quad a, b \in G\,. \tag{4.34}$$

Proposition 4.57. *The relation Equation (4.34) is an equivalence relation.*
Rests on Definitions 3.58 and 4.33.

Proof. *Derives Proposition 4.57.*
We verify the three properties:

1. Reflexivity holds because H is a subgroup of G, so the neutral element exists in H and

$$a = ae\,,$$

whence $a \sim_H a$.

2. Suppose $a \sim_H b$. Then there exists $h \in H$ such that $b = ah$. Since H is a subgroup of G (and in essence a group), h^{-1} exists. Operating on the right with $\cdot\, h^{-1}$,

$$\begin{aligned} b &= ah \\ bh^{-1} &= a\,. \end{aligned}$$

Then there exists $k = h^{-1} \in H$ such that $a = bk$; hence $b \sim_H a$ and symmetry holds.

3. If $a \sim_H b$ and $b \sim_H c$, it is because there exist $h, k \in H$ such that

$$b = ah, \quad c = bk\,.$$

Substituting the first into the second,

$$\begin{aligned} c &= (ah)k \\ &= a(hk) \\ &= al\,, \end{aligned}$$

and thus, since H is a subgroup, by closure there exists $l = hk \in H$; we conclude that $a \sim_H c$, so transitivity holds.

□

Definition 4.58 (Cosets)**.** We call *left lateral class* or *left coset* each equivalence class given by the relation Equation (4.34). We have

$$\begin{aligned}[a] &= \{x \in G \mid x \sim_H a\} \\ &= \{x \in G \mid x = ah,\ h \in H\} \\ &= \{ah \mid h \in H\}\,,\end{aligned}$$

and, in view of the fact that h ranges over H, we shall symbolize this class as

$$[a] \overset{\text{not.}}{=} aH\,.$$

Rests on Proposition 4.57 and Definition 3.59.

Analogously we can define Ha as a *right lateral class* or *right coset* of G. The study of left and right cosets is entirely parallel, so we shall refer only to left cosets, which we call simply *lateral classes* or *cosets*.

Finally, we must mention that it is trivial to verify that

$$eH = H, \quad He = H\,. \tag{4.35}$$

Compatibility between left and right cosets

Consider the group G and $H \leq G$. A natural question we may ask is: under which conditions is a left coset equal to a right coset? That is, under which conditions does $aH = Ha$ hold? We may suspect a possible condition (perhaps necessary and sufficient) by looking at the definition of a normal subgroup. Does

$$H = aHa^{-1},\ \forall\, a \in G \quad \overset{?}{\iff} \quad aH = Ha, \quad \forall\ a \in G$$

hold?

Indeed it does. We must prove the two implications.

⇒) Supposing $H \trianglelefteq G$, we have $H = aHa^{-1}$ for all a. Then the elements of H can be written as

$$h' = aha^{-1}\,,$$

whence

$$\begin{aligned}Ha &= \{h'a \mid h' \in H\} \\ &= \left\{(aha^{-1})a \mid h \in H\right\} \\ &= \{ah \mid h \in H\} \\ &= aH\,.\end{aligned}$$

⇐) Recall that

$$aH = \{ah \mid h \in H\}\,, \quad Ha = \left\{h'a \mid h' \in H\right\}\,,$$

and supposing $aH = Ha$ — that is, that these sets have the same elements — we can always write an element of aH as

$$ah = h'a\,,$$

whence

$$\begin{aligned}aHa^{-1} &= \left\{aha^{-1} \mid h \in H\right\} \\ &= \left\{(h'a)a^{-1} \mid h' \in H\right\} \\ &= \{h' \mid h' \in H\} \\ &= H,\ \forall\, a\,,\end{aligned}$$

and thus $H \trianglelefteq G$.

In conclusion, we have found a characterization of the normality of a subgroup:

$$\boxed{H \trianglelefteq G \iff aH = Ha,\ \forall\, a \in G} \tag{4.36}$$

From Equation (4.36) a further characterization of the normality of a subgroup can be seen:

$$\boxed{H \trianglelefteq G \iff \forall\, a \in G,\ \forall\, h \in H,\ \exists\, h' \in H \mid ah = h'a} \tag{4.37}$$

From Equation (4.36) (writing the equality of the two cosets elementwise).

4.8.3 The quotient group

Product of cosets

Consider a group $G = \{e, a, b, \ldots\}$ and a subgroup $H \leq G$. When we introduced the coset equivalence relation we did not analyse the quotient space. That quotient space is the set of all the cosets,

$$G/\!\sim_H = \{H, aH, bH, \ldots\} \overset{\text{not.}}{=} G/H\,.$$

We define the IBO *product of cosets* as

$$\begin{array}{rccc} * : & G/H \times G/H & \longrightarrow & G/H \\ & (aH, bH) & \longmapsto & aH * bH = abH\,. \end{array}$$

This product provides yet another characterization of the normality of a subgroup:

$$\boxed{H \trianglelefteq G \iff \forall\, a, b \in G,\ aH * bH = abH}\,. \tag{4.38}$$

The prescription $aH * bH = abH$ needs a word of care before it can be called an operation at all. A coset is a set, and a is merely one of its elements; the recipe reads that element off, multiplies, and forms a new coset. Nothing so far guarantees that a different choice of representatives returns the same answer — and in fact that is exactly what normality is.

Proposition 4.59 (Well-definedness of the coset product)**.** *Let G be a group and $H \leq G$. The prescription $aH * bH = abH$ is independent of the representatives a and b chosen, and so defines an IBO on G/H, if and only if $H \trianglelefteq G$. This is Equation (4.38).*

Rests on Definitions 4.56 and 4.58 and Equation (4.35).

Proof. *Derives Proposition 4.59.*

Throughout we use the fact that the cosets are the equivalence classes of Equation (4.34) (Proposition 4.57), so that

$$x \in yH \iff xH = yH\,. \tag{4.39}$$

($\Leftarrow$) Let $H \trianglelefteq G$ and let $a' \in aH$, $b' \in bH$, say $a' = ah_1$ and $b' = bh_2$ with $h_1, h_2 \in H$. Normality applied with $p = b^{-1}$ gives $H = b^{-1}Hb$, so $b^{-1}h_1 b \in H$; call this element h_3, so that $h_1 b = bh_3$. Then

$$\begin{aligned} a'b' &= (ah_1)(bh_2) \\ &= a(h_1 b)h_2 \\ &= a(bh_3)h_2 \\ &= (ab)(h_3 h_2)\,, \end{aligned}$$

and $h_3 h_2 \in H$ by closure, so $a'b' \in abH$. By Equation (4.39), $a'b'H = abH$: the answer does not depend on the representatives, and $*$ is a well-defined IBO on G/H.

($\Rightarrow$) Suppose the prescription is independent of representatives, and let $a \in G$ and $h \in H$ be arbitrary. Since $ah \in aH$ and $a^{-1} \in a^{-1}H$, evaluating $aH * a^{-1}H$ on the representatives ah and a^{-1} must give the same coset as evaluating it on a and a^{-1}, that is,

$$(ah)a^{-1}H = aa^{-1}H = eH = H\,,$$

the last step by Equation (4.35). By Equation (4.39) this says $aha^{-1} \in H$. As h ranged over H, we have $aHa^{-1} \subseteq H$ for every $a \in G$; applying the same inclusion with a^{-1} in place of a gives $a^{-1}Ha \subseteq H$, which upon multiplying by a on the left and a^{-1} on the right reads $H \subseteq aHa^{-1}$. The two inclusions give $H = aHa^{-1}$ for every $a \in G$, which is Definition 4.56. □

The quotient group

The quotient space G/H by itself has no further relevance for our study; but if we add the condition that H be normal in G, then G/H, together with the product of cosets, is a group.

Theorem 4.60 (Quotient group)**.** *Let $H \trianglelefteq G$. Then $(G/H, *)$ is a group, known as the* quotient group *of G over H.*

Rests on Definition 4.56 and Equations (4.35) and (4.38).

Proof. *Derives Theorem 4.60.*

We verify the four axioms:

1. Taking H normal in G, by Equation (4.38) we see the following: by the closure of G, ab also belongs to G; hence

   $$aH * bH = abH \in G/H\,,$$

   so the closure property holds for $(G/H, *)$.

2. The associativity of $(G/H, *)$ is a consequence of the associativity of G. Indeed,

   $$\begin{aligned}(aH * bH) * cH &= abH * cH\\ &= (ab)cH\\ &= a(bc)H\\ &= aH * bcH\\ &= aH * (bH * cH)\,.\end{aligned}$$

3. The existence of the neutral element is due to the existence of the neutral element e in G. From Equation (4.35) we see that $eH = H$. Hence, for an arbitrary element of G/H, say aH, there exists $H \in G/H$ such that

   $$H * aH = eH * aH = eaH = aH, \quad aH * H = aH * eH = aeH = aH\,.$$

4. It remains to prove the existence of inverse elements. Let $a \in G$. Since G is a group, a^{-1} exists. Now, $a^{-1}H$ is the inverse of aH, since

   $$aH * a^{-1}H = aa^{-1}H = H, \quad a^{-1}H * aH = a^{-1}aH = H\,.$$

□

4.9 Direct products

4.9.1 Direct product

Definition 4.61 (Direct product of groups)**.** Let $(G_1, *_1, e_1)$ and $(G_2, *_2, e_2)$ be groups. We define the *direct product* of the groups as the Cartesian product of the sets G_1 and G_2 together with the binary operation

$$\begin{array}{rccc} * : & (G_1 \times G_2) \times (G_1 \times G_2) & \longrightarrow & G_1 \times G_2 \\ & ((a_1, a_2), (b_1, b_2)) & \longmapsto & (a_1 *_1 b_1, a_2 *_2 b_2)\,. \end{array}$$

Rests on Definitions 4.1 and 4.20.

Proposition 4.62. $(G_1 \times G_2, *, (e_1, e_2))$ *is a group.*
Rests on Definitions 4.20 and 4.61.

Proof. *Derives Proposition 4.62.*
We verify the four axioms:

1. The binary operation $*$ is an IBO, since $a_1 *_1 b_1 \in G_1$ and $a_2 *_2 b_2 \in G_2$ by closure. Hence always $(a_1 *_1 b_1, a_2 *_2 b_2) \in G_1 \times G_2$, so there is closure with respect to $*$.

2. We have

$$\begin{aligned} ((a_1, a_2) * (b_1, b_2)) * (c_1, c_2) &= (a_1 *_1 b_1, a_2 *_2 b_2) * (c_1, c_2) \\ &= ((a_1 *_1 b_1) *_1 c_1, (a_2 *_2 b_2) *_2 c_2) \\ &= (a_1 *_1 (b_1 *_1 c_1), a_2 *_2 (b_2 *_2 c_2)) \\ &= (a_1, a_2) * (b_1 *_1 c_1, b_2 *_2 c_2) \\ &= (a_1, a_2) * ((b_1, b_2) * (c_1, c_2))\,, \end{aligned}$$

   so associativity holds.

3. We can also see that the neutral element $(e_1, e_2) \in G_1 \times G_2$ exists, since

$$\begin{aligned} (a_1, a_2) * (e_1, e_2) &= (a_1 *_1 e_1, a_2 *_2 e_2) = (a_1, a_2)\,, \\ (e_1, e_2) * (a_1, a_2) &= (e_1 *_1 a_1, e_2 *_2 a_2) = (a_1, a_2)\,. \end{aligned}$$

4. Finally, given $(a_1, a_2) \in G_1 \times G_2$, its inverse $(a_1, a_2)^{-1} = (a_1^{-1}, a_2^{-1})$ exists. Indeed,

$$(a_1, a_2) * (a_1, a_2)^{-1} = (a_1, a_2) * (a_1^{-1}, a_2^{-1}) = (a_1 *_1 a_1^{-1}, a_2 *_2 a_2^{-1}) = (e_1, e_2)\,,$$

   and analogously $(a_1, a_2)^{-1} * (a_1, a_2) = (e_1, e_2)$.

□

Definition 4.63 (Product of subgroups)**.** Consider a group G and $H, K \leq G$. We define the *product of subgroups* HK as the set

$$HK = \{hk \mid h \in H,\ k \in K\}\,. \tag{4.40}$$

Rests on Definition 4.33.

Proposition 4.64 (Direct product of abelian groups)**.** *The direct product of two abelian groups is an abelian group.*
Rests on Definitions 4.8 and 4.61.

Proof. *Derives Proposition 4.64.*

Let $(G_1, *_1)$ and $(G_2, *_2)$ be abelian. That $G_1 \times G_2$ is a group is Proposition 4.62; commutativity is inherited componentwise, since for all $a_1, b_1 \in G_1$ and $a_2, b_2 \in G_2$

$$\begin{aligned}(a_1, a_2) * (b_1, b_2) &= (a_1 *_1 b_1, a_2 *_2 b_2) \\ &= (b_1 *_1 a_1, b_2 *_2 a_2) \\ &= (b_1, b_2) * (a_1, a_2)\,,\end{aligned}$$

the middle step applying Equation (4.3) in each factor separately. □

Example 4.65 (The Cayley table of $\mathbb{Z}_2 \times \mathbb{Z}_4$)**.** Write $\mathbb{Z}_n$ for the additive group $\mathbb{Z}/n\mathbb{Z}$ of Proposition 4.37 and denote the class $\overline{k}$ simply by k. Both factors are abelian, so by Proposition 4.64 the eight-element group $\mathbb{Z}_2 \times \mathbb{Z}_4$ is abelian, and its Cayley table (Table 4.2) is symmetric about its diagonal. The operation is addition in each slot separately,

$$(i, j) * (k, l) = \left(\overline{i+k} \text{ in } \mathbb{Z}_2,\ \overline{j+l} \text{ in } \mathbb{Z}_4\right)\,, \tag{4.41}$$

so the neutral element is $(0, 0)$ and the inverse of (i, j) is $(-i, -j)$ read in the respective factors.

Note that adding any element of this group to itself four times returns $(0, 0)$, since $4i$ is a multiple of 2 and $4j$ a multiple of 4. No single element therefore runs through all eight, so $\mathbb{Z}_2 \times \mathbb{Z}_4$ is not isomorphic to $\mathbb{Z}_8$ — two abelian groups of the same order need not be isomorphic, which is why the classification of finite abelian groups mentioned in Section 4.4 has content.

Rests on Propositions 4.37 and 4.64 and Definition 4.34.

Table 4.2: The Cayley table of the direct product $\mathbb{Z}_2 \times \mathbb{Z}_4$, whose operation is Equation (4.41). The first slot is read modulo 2 and the second modulo 4. The table is symmetric about its diagonal, as Proposition 4.64 requires, and no row or column repeats an entry, as Section 4.4.2 requires of any group.

$*$	$(0,0)$	$(0,1)$	$(0,2)$	$(0,3)$	$(1,0)$	$(1,1)$	$(1,2)$	$(1,3)$
$(0,0)$	$(0,0)$	$(0,1)$	$(0,2)$	$(0,3)$	$(1,0)$	$(1,1)$	$(1,2)$	$(1,3)$
$(0,1)$	$(0,1)$	$(0,2)$	$(0,3)$	$(0,0)$	$(1,1)$	$(1,2)$	$(1,3)$	$(1,0)$
$(0,2)$	$(0,2)$	$(0,3)$	$(0,0)$	$(0,1)$	$(1,2)$	$(1,3)$	$(1,0)$	$(1,1)$
$(0,3)$	$(0,3)$	$(0,0)$	$(0,1)$	$(0,2)$	$(1,3)$	$(1,0)$	$(1,1)$	$(1,2)$
$(1,0)$	$(1,0)$	$(1,1)$	$(1,2)$	$(1,3)$	$(0,0)$	$(0,1)$	$(0,2)$	$(0,3)$
$(1,1)$	$(1,1)$	$(1,2)$	$(1,3)$	$(1,0)$	$(0,1)$	$(0,2)$	$(0,3)$	$(0,0)$
$(1,2)$	$(1,2)$	$(1,3)$	$(1,0)$	$(1,1)$	$(0,2)$	$(0,3)$	$(0,0)$	$(0,1)$
$(1,3)$	$(1,3)$	$(1,0)$	$(1,1)$	$(1,2)$	$(0,3)$	$(0,0)$	$(0,1)$	$(0,2)$

Proposition 4.66 (Characterization of the direct product)**.** *Given a group G with subgroups $H, K \leq G$, the map*

$$\begin{array}{rccc}\varphi: & H \times K & \longrightarrow & G \\ & (h, k) & \longmapsto & hk\,,\end{array} \tag{4.42}$$

is an isomorphism — in which case G is called the internal direct product *of the subgroups H and K — if and only if*

1. *$G = HK$;*
2. *$H \cap K = \{e\}$;*

3. $H \trianglelefteq G$ *and* $K \trianglelefteq G$.

Rests on Definitions 4.40, 4.56, 4.61 and 4.63.

Proof. *Derives Proposition 4.66.*

Throughout, $H \times K$ carries the operation of Definition 4.61, whose neutral element is (e, e).

($\Leftarrow$) Assume (1), (2) and (3). We first show that every element of H commutes with every element of K. Let $h \in H$, $k \in K$ and put

$$c = hkh^{-1}k^{-1}\,.$$

Read as $c = (hkh^{-1})\,k^{-1}$, the normality of K gives $hkh^{-1} \in K$, so $c \in K$; read as $c = h\,(kh^{-1}k^{-1})$, the normality of H gives $kh^{-1}k^{-1} \in H$, so $c \in H$. Hence $c \in H \cap K = \{e\}$ by (2), that is,

$$hk = kh\,, \qquad \forall\ h \in H,\ \forall\ k \in K\,. \tag{4.43}$$

Now φ is a homomorphism: for $h, h' \in H$ and $k, k' \in K$,

$$\begin{aligned} \varphi\left((h,k) * (h',k')\right) &= \varphi\left(hh', kk'\right) \\ &= hh'kk' \\ &= hkh'k' && \text{by Equation (4.43)} \\ &= \varphi(h,k)\,\varphi(h',k')\,. \end{aligned}$$

It is surjective by (1), since $G = HK$ is exactly the set of products hk. It is injective because its kernel is trivial: if $\varphi(h,k) = e$ then $h = k^{-1}$, and $k^{-1} \in K$, so $h \in H \cap K = \{e\}$; thus $h = e$ and then $k = e$, that is, $\ker\varphi = \{(e,e)\}$, and Proposition 4.17 applies. A bijective homomorphism is an isomorphism (Definition 4.40).

($\Rightarrow$) Assume φ is an isomorphism. Surjectivity says that every $g \in G$ can be written $g = hk$ with $h \in H$ and $k \in K$, which is $G = HK$: this is (1). For (2), let $x \in H \cap K$. Then $x^{-1} \in K$, and

$$\varphi\left(x, x^{-1}\right) = xx^{-1} = e = \varphi(e,e)\,,$$

so injectivity forces $(x, x^{-1}) = (e,e)$ and $x = e$.

For (3), note that $H \times \{e\}$ is a normal subgroup of $H \times K$, since

$$(h',k') * (h,e) * (h',k')^{-1} = \left(h'h(h')^{-1},\ k'e(k')^{-1}\right) = \left(h'h(h')^{-1}, e\right) \in H \times \{e\}\ ,$$

and that $\varphi\left(H \times \{e\}\right) = H$, because $\varphi(h,e) = he = h$. Let $g \in G$ and write $g = \varphi(u)$ with $u \in H \times K$, which is possible by surjectivity. Since φ is a homomorphism, $\varphi(u)^{-1} = \varphi\left(u^{-1}\right)$ (Proposition 4.43), and therefore

$$\begin{aligned} gHg^{-1} &= \varphi(u)\,\varphi\left(H \times \{e\}\right)\,\varphi\left(u^{-1}\right) \\ &= \varphi\left(u * (H \times \{e\}) * u^{-1}\right) \\ &= \varphi\left(H \times \{e\}\right) \\ &= H\,, \end{aligned}$$

so $H \trianglelefteq G$. The same argument with $\{e\} \times K$ gives $K \trianglelefteq G$. □

Remark 4.67. It is essential that the proposition names the map Equation (4.42). The weaker reading — that G be isomorphic to $H \times K$ by *some* isomorphism or other — does not imply (1)–(3), and the difference is not a technicality. Take $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ and let $H = K = \{(0,0),(1,0)\} \leq G$. Each of H and K is a two-element group, hence isomorphic to $\mathbb{Z}_2$ by the uniqueness of the two-element Cayley table established in Section 4.5.1, so $H \times K$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2 = G$; yet $HK = H \neq G$ and $H \cap K = H \neq \{e\}$. An abstract isomorphism knows nothing about how H and K sit inside G, and it is exactly that placement which conditions (1)–(3) describe.

4.9.2 Semidirect product

The direct product is symmetric in its two factors: Proposition 4.66 demands that *both* H and K be normal in G, and Equation (4.43) then forces every element of one to commute with every element of the other. That is a strong condition, and most of the groups physics actually uses fail it. In the group of rigid motions of space, a rotation and a translation do not commute; the translations form a normal subgroup, and the rotations do not. What is needed is a product in which one factor is normal and the other merely *acts* on it. That is the semidirect product, and it is the shape of the Euclidean and Poincaré groups alike.

Semidirect product of groups; characterization of the semidirect product

The action of one group on another is by automorphisms, so we first note that the automorphisms of a group form a group in their own right.

Proposition 4.68 (The automorphism group)**.** *Let N be a group. The set* $\mathrm{Aut}(N)$ *of all automorphisms of N is a subgroup of the symmetric group S_N, and in particular a group under composition.*

Rests on Definitions 4.33, 4.41 and 4.44.

Proof. *Derives Proposition 4.68.*

An automorphism of N is in particular a bijection of N onto itself, so $\mathrm{Aut}(N) \subseteq S_N$. The identity map is an automorphism, so the neutral element of S_N lies in $\mathrm{Aut}(N)$. If $\alpha, \beta \in \mathrm{Aut}(N)$ then for all $n_1, n_2 \in N$

$$(\alpha \circ \beta)(n_1 n_2) = \alpha\left(\beta(n_1)\beta(n_2)\right) = (\alpha \circ \beta)(n_1)\,(\alpha \circ \beta)(n_2)\,,$$

so $\alpha \circ \beta$ is a homomorphism, and it is a bijection as a composite of bijections. Finally, since α is a homomorphism,

$$\alpha\left(\alpha^{-1}(n_1)\,\alpha^{-1}(n_2)\right) = n_1 n_2\,,$$

and applying α^{-1} to both sides gives $\alpha^{-1}(n_1)\,\alpha^{-1}(n_2) = \alpha^{-1}(n_1 n_2)$, so $\alpha^{-1} \in \mathrm{Aut}(N)$. The three conditions of Definition 4.33 are met. □

Definition 4.69 (External semidirect product)**.** Let N and H be groups and let $\varphi : H \longrightarrow \mathrm{Aut}(N)$ be a group homomorphism, written $h \longmapsto \varphi_h$. The *external semidirect product* $N \rtimes_\varphi H$ is the Cartesian product $N \times H$ equipped with the internal binary operation

$$(n_1, h_1) * (n_2, h_2) = (n_1\,\varphi_{h_1}(n_2),\, h_1 h_2)\,. \tag{4.44}$$

Rests on Proposition 4.68 and Definitions 4.1 and 4.39.

The first slot is not multiplied blindly: the second factor of the left element is allowed to act on the first slot of the right element before the two are combined. Setting $\varphi_h = \mathrm{id}_N$ throughout removes the action and returns Definition 4.61, as Proposition 4.73 makes precise.

Proposition 4.70 (The semidirect product is a group)**.** *$(N \rtimes_\varphi H, *)$ is a group, with neutral element (e_N, e_H) and with*

$$(n, h)^{-1} = \left(\varphi_{h^{-1}}\left(n^{-1}\right),\, h^{-1}\right)\,. \tag{4.45}$$

Rests on Definitions 4.20 and 4.69.

Proof. *Derives Proposition 4.70.*

We use twice over that φ is a homomorphism into $\mathrm{Aut}(N)$: each φ_h is an automorphism of N, and $\varphi_{h_1h_2} = \varphi_{h_1} \circ \varphi_{h_2}$. In particular $\varphi_{e_H} = \mathrm{id}_N$ and $\varphi_h(e_N) = e_N$, by Proposition 4.42 applied to φ and to φ_h respectively.

Closure. $\varphi_{h_1}(n_2) \in N$ and $h_1h_2 \in H$, so the right-hand side of Equation (4.44) lies in $N \times H$.

Associativity. On the one hand,

$$\begin{aligned}((n_1, h_1) * (n_2, h_2)) * (n_3, h_3) &= (n_1\varphi_{h_1}(n_2),\, h_1h_2) * (n_3, h_3) \\ &= (n_1\varphi_{h_1}(n_2)\,\varphi_{h_1h_2}(n_3),\, h_1h_2h_3)\ ,\end{aligned}$$

and on the other,

$$\begin{aligned}(n_1, h_1) * ((n_2, h_2) * (n_3, h_3)) &= (n_1, h_1) * (n_2\varphi_{h_2}(n_3),\, h_2h_3) \\ &= (n_1\varphi_{h_1}(n_2\varphi_{h_2}(n_3))\,,\, h_1h_2h_3) \\ &= (n_1\varphi_{h_1}(n_2)\,\varphi_{h_1}(\varphi_{h_2}(n_3))\,,\, h_1h_2h_3) \\ &= (n_1\varphi_{h_1}(n_2)\,\varphi_{h_1h_2}(n_3),\, h_1h_2h_3)\ ,\end{aligned}$$

the third step because φ_{h_1} is a homomorphism of N and the fourth because φ is a homomorphism into $\mathrm{Aut}(N)$. The two agree.

Neutral element. Using $\varphi_h(e_N) = e_N$ and $\varphi_{e_H} = \mathrm{id}_N$,

$$(n, h) * (e_N, e_H) = (n\varphi_h(e_N), h) = (n, h)\,, \qquad (e_N, e_H) * (n, h) = (\varphi_{e_H}(n), h) = (n, h)\,.$$

Inverses. With the candidate Equation (4.45),

$$\begin{aligned}(n, h) * \left(\varphi_{h^{-1}}\left(n^{-1}\right), h^{-1}\right) &= \left(n\,\varphi_h\left(\varphi_{h^{-1}}\left(n^{-1}\right)\right),\, hh^{-1}\right) \\ &= \left(n\,\varphi_{e_H}\left(n^{-1}\right),\, e_H\right) \\ &= (e_N, e_H)\,,\end{aligned}$$

and in the other order

$$\begin{aligned}\left(\varphi_{h^{-1}}\left(n^{-1}\right), h^{-1}\right) * (n, h) &= \left(\varphi_{h^{-1}}\left(n^{-1}\right)\,\varphi_{h^{-1}}(n),\, h^{-1}h\right) \\ &= \left(\varphi_{h^{-1}}\left(n^{-1}n\right),\, e_H\right) \\ &= (e_N, e_H)\,.\end{aligned}$$

All four axioms of Definition 4.20 hold. □

Proposition 4.71 (The two factors inside the semidirect product)**.** *Write $G = N \rtimes_\varphi H$, $\widetilde{N} = N \times \{e_H\}$ and $\widetilde{H} = \{e_N\} \times H$. Then $\widetilde{N} \trianglelefteq G$, $\widetilde{H} \leq G$, $\widetilde{N} \cap \widetilde{H} = \{(e_N, e_H)\}$ and $G = \widetilde{N}\widetilde{H}$; moreover $n \longmapsto (n, e_H)$ and $h \longmapsto (e_N, h)$ are isomorphisms of N onto $\widetilde{N}$ and of H onto $\widetilde{H}$, and conjugation inside G reproduces φ:*

$$(e_N, h) * (n, e_H) * (e_N, h)^{-1} = (\varphi_h(n), e_H)\ . \tag{4.46}$$

Rests on Proposition 4.70 and Definitions 4.56 and 4.63.

Proof. *Derives Proposition 4.71.*

By Equation (4.44), $(n_1, e_H) * (n_2, e_H) = (n_1\varphi_{e_H}(n_2), e_H) = (n_1n_2, e_H)$ and $(e_N, h_1) * (e_N, h_2) = (\varphi_{h_1}(e_N), h_1h_2) = (e_N, h_1h_2)$, so both maps are homomorphisms; each is evidently bijective onto its image, hence an isomorphism (Definition 4.40), and each image contains the neutral element and the inverses of its elements, so both are subgroups. Their intersection

is $\{(e_N, e_H)\}$ because a pair lying in both has $h = e_H$ and $n = e_N$, and $(n, e_H) * (e_N, h) = (n\varphi_{e_H}(e_N), h) = (n, h)$ shows $G = \widetilde{N}\widetilde{H}$.

For normality, let $(m, k) \in G$ and $(n, e_H) \in \widetilde{N}$. Then

$$\begin{aligned}(m,k) * (n,e_H) * (m,k)^{-1} &= (m\,\varphi_k(n), k) * \left(\varphi_{k^{-1}}\left(m^{-1}\right), k^{-1}\right) \\ &= \left(m\,\varphi_k(n)\,\varphi_k\left(\varphi_{k^{-1}}\left(m^{-1}\right)\right), e_H\right) \\ &= \left(m\,\varphi_k(n)\,m^{-1}, e_H\right),\end{aligned}$$

which lies in $\widetilde{N}$; hence $\widetilde{N} \trianglelefteq G$. Finally, $(e_N, h)^{-1} = (\varphi_{h^{-1}}(e_N), h^{-1}) = (e_N, h^{-1})$, so

$$(e_N, h) * (n, e_H) * (e_N, h^{-1}) = (\varphi_h(n), h) * (e_N, h^{-1}) = (\varphi_h(n)\varphi_h(e_N), e_H)\ ,$$

which is Equation (4.46). □

The construction is therefore faithful to its data, and the converse holds too: whenever a group contains a normal subgroup and a complement to it, it *is* a semidirect product. Note that the complement is *not* required to be normal — that is the whole difference from Proposition 4.66.

Theorem 4.72 (Internal characterization of the semidirect product)**.** *Let G be a group, $N \trianglelefteq G$ and $H \leq G$ with $G = NH$ and $N \cap H = \{e\}$. Then*

$$\varphi_h(n) = hnh^{-1} \tag{4.47}$$

defines a homomorphism $\varphi : H \longrightarrow \mathrm{Aut}(N)$, and

$$\begin{array}{rccc}\psi: & N \rtimes_\varphi H & \longrightarrow & G \\ & (n,h) & \longmapsto & nh\end{array}$$

is an isomorphism. Conversely, by Proposition 4.71, every external semidirect product is of this form.

Rests on Definitions 4.56, 4.63 and 4.69.

Proof. *Derives Theorem 4.72.*

φ_h is an automorphism of N. Since $N \trianglelefteq G$ and $h \in G$, we have $hnh^{-1} \in N$, so φ_h maps N into N; it is a homomorphism because

$$\varphi_h(n_1n_2) = hn_1n_2h^{-1} = \left(hn_1h^{-1}\right)\left(hn_2h^{-1}\right) = \varphi_h(n_1)\varphi_h(n_2)\,,$$

and $\varphi_{h^{-1}}$ is a two-sided inverse for it, so it is bijective.

φ is a homomorphism. For $h_1, h_2 \in H$ and $n \in N$,

$$\varphi_{h_1h_2}(n) = h_1h_2nh_2^{-1}h_1^{-1} = \varphi_{h_1}(\varphi_{h_2}(n))\ ,$$

using Equation (4.16) for $(h_1h_2)^{-1}$; hence $\varphi_{h_1h_2} = \varphi_{h_1} \circ \varphi_{h_2}$.

ψ is a homomorphism. Using Equation (4.44) and then Equation (4.47),

$$\begin{aligned}\psi\left((n_1,h_1) * (n_2,h_2)\right) &= \psi\left(n_1\varphi_{h_1}(n_2),\, h_1h_2\right) \\ &= n_1\left(h_1n_2h_1^{-1}\right)h_1h_2 \\ &= (n_1h_1)(n_2h_2) \\ &= \psi(n_1,h_1)\,\psi(n_2,h_2)\,.\end{aligned}$$

ψ is bijective. It is surjective because $G = NH$ is precisely the set of products nh. Its kernel is trivial: if $nh = e$ then $n = h^{-1}$, which lies in $N \cap H = \{e\}$, so $n = e$ and then $h = e$. By Proposition 4.17, ψ is injective, and a bijective homomorphism is an isomorphism. □

Proposition 4.73 (When a semidirect product is direct)**.** *Let $G = N \rtimes_\varphi H$ and let $\widetilde{N}$, $\widetilde{H}$ be as in Proposition 4.71. The following are equivalent:*

1. *$\widetilde{H} \trianglelefteq G$;*
2. *$\varphi_h = \mathrm{id}_N$ for every $h \in H$;*
3. *the operation Equation (4.44) is the operation of Definition 4.61, so that $N \rtimes_\varphi H$ is the direct product $N \times H$.*

Rests on Proposition 4.71 and Definitions 4.56 and 4.61.

Proof. *Derives Proposition 4.73.*

(2) $\implies$ (3): substituting $\varphi_{h_1} = \mathrm{id}_N$ into Equation (4.44) gives $(n_1, h_1) * (n_2, h_2) = (n_1 n_2, h_1 h_2)$, which is Definition 4.61.

(3) $\implies$ (1): in the direct product,

$$(n', h') * (e_N, h) * (n', h')^{-1} = \left(n' e_N n'^{-1},\, h' h (h')^{-1}\right) = \left(e_N,\, h' h (h')^{-1}\right) \in \widetilde{H}\,,$$

so $\widetilde{H} \trianglelefteq G$.

(1) $\implies$ (2): let $h \in H$ and $n \in N$. Since $(n, e_H)^{-1} = \left(\varphi_{e_H}(n^{-1}), e_H\right) = (n^{-1}, e_H)$,

$$\begin{aligned}(n, e_H) * (e_N, h) * (n, e_H)^{-1} &= (n\varphi_{e_H}(e_N),\, h) * \left(n^{-1}, e_H\right) \\ &= \left(n\,\varphi_h\left(n^{-1}\right),\, h\right)\,.\end{aligned}$$

If $\widetilde{H}$ is normal this must lie in $\{e_N\} \times H$, that is, $n\,\varphi_h(n^{-1}) = e_N$, or $\varphi_h(n^{-1}) = n^{-1}$. As n runs over N so does n^{-1} (Proposition 4.24), hence $\varphi_h = \mathrm{id}_N$ for every h. □

Example 4.74 (The Euclidean group)**.** Let $N = (\mathbb{R}^3, +)$, the translations of Euclidean space, and let $H = \mathrm{SO}(3)$, the rotations about a fixed origin. A rotation R is a linear bijection of $\mathbb{R}^3$, so $\varphi_R(\boldsymbol{a}) = R\boldsymbol{a}$ is an automorphism of the *additive* group $\mathbb{R}^3$, and $\varphi_{R_1 R_2} = \varphi_{R_1} \circ \varphi_{R_2}$ because matrix multiplication is associative. The semidirect product

$$\mathrm{SE}(3) = \mathbb{R}^3 \rtimes_\varphi \mathrm{SO}(3) \tag{4.48}$$

therefore exists, and by Equation (4.44) its operation is

$$(\boldsymbol{a}_1, R_1) * (\boldsymbol{a}_2, R_2) = (\boldsymbol{a}_1 + R_1\boldsymbol{a}_2,\, R_1 R_2)\,, \tag{4.49}$$

which is exactly the composition of the rigid motions $\boldsymbol{x} \longmapsto R\boldsymbol{x} + \boldsymbol{a}$ of $\mathbb{R}^3$: applying $(\boldsymbol{a}_2, R_2)$ and then $(\boldsymbol{a}_1, R_1)$ sends $\boldsymbol{x}$ to $R_1 R_2 \boldsymbol{x} + R_1\boldsymbol{a}_2 + \boldsymbol{a}_1$. The same construction with $\mathrm{O}(3)$ in place of $\mathrm{SO}(3)$ adjoins the reflections and gives the full group of distance-preserving maps of $\mathbb{R}^3$, the Euclidean group proper,

$$\mathrm{E}(3) = \mathrm{ISO}(3) = \mathbb{R}^3 \rtimes_\varphi \mathrm{O}(3)\,, \tag{4.50}$$

of which $\mathrm{SE}(3)$ is the orientation-preserving subgroup of index two; the isometries of the equilateral triangle in Proposition 4.51 are the planar analogue. The naming follows the convention Equation (4.51) uses below and Definition 39.10 fixes: an inhomogeneous group $\mathrm{ISO}(p, q)$ carries the *full* orthogonal group of its form, reflections included, and the prefix S marks the orientation-preserving subgroup.

The product is genuinely semidirect and not direct. Conjugating a rotation by a translation gives, by the computation in Proposition 4.73,

$$(\boldsymbol{a}, \mathbb{1}) * (\boldsymbol{0}, R) * (\boldsymbol{a}, \mathbb{1})^{-1} = (\boldsymbol{a} - R\boldsymbol{a},\, R)\,,$$

whose translation part vanishes only when $R\boldsymbol{a} = \boldsymbol{a}$. The rotations therefore do not form a normal subgroup — a rotation about the origin, viewed from a translated origin, is a rotation about a different point — while the translations do, by Proposition 4.71. This is the algebraic content of the familiar fact that rotating and then translating is not the same as translating and then rotating.

Rests on Definition 4.69 and Proposition 4.73.

Example 4.75 (The Poincaré group)**.** The same construction with the translations of Minkowski space and the Lorentz group in place of the rotations gives the symmetry group of special relativity. Take $N = (\mathbb{R}^4, +)$ and for H the Lorentz group L of Definition 39.1, whose elements are invertible linear maps and therefore automorphisms of the additive group $\mathbb{R}^4$, with $\varphi_\Lambda(a) = \Lambda a$. Then

$$\mathrm{ISO}(3,1) = \mathbb{R}^4 \rtimes_\varphi L\,, \tag{4.51}$$

and Equation (4.44) reads

$$(a_2, \Lambda_2) * (a_1, \Lambda_1) = (a_2 + \Lambda_2 a_1,\ \Lambda_2\Lambda_1)\ ,$$

which is the composition law Equation (39.17) of the Poincaré transformations, defined and used in Chapter 39; restricting L to its proper subgroup $\mathrm{SO}(3,1)$ gives the proper Poincaré group in the same way. That the translations are normal and the Lorentz transformations are not is precisely why Equation (39.17) is not the componentwise product of a direct product, by Proposition 4.73. The infinitesimal counterpart of Equation (4.51) is the semidirect *sum* Equation (14.99) of Chapter 14, in which the translation generators span an abelian ideal on which the Lorentz generators act.

Rests on Definition 4.69 and Proposition 4.71.

Chapter 5

Linear Algebra and Representation Theory

The vector space is the algebraic structure on which almost all of the mathematics of physics is built: classical states, quantum states, fields, and the linearized neighbourhoods of curved spaces are all modelled on vector spaces. This chapter develops the theory of vector spaces and of the linear transformations between them, the dual space, and the algebras — in particular Lie algebras — that linear transformations generate. It then turns to representation theory: the systematic study of how an abstract group can act by linear transformations on a vector space. The chapter closes with the fundamental properties of irreducible representations of finite groups, the working machinery behind the symmetry analysis used throughout the treatise. A short section of elementary algebraic formulas, collected from a companion summary sheet of the source, opens the chapter for later reference.

5.1 Preliminary formulas from elementary algebra

The following classical identities are collected here for reference; they are used freely in computations throughout the treatise.

5.1.1 Factorials and the binomial theorem

The *double factorial* abbreviates the product of the odd or of the even numbers:

$$1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n+1) = (2n+1)!!\,, \tag{5.1}$$

$$2 \cdot 4 \cdot 6 \cdot \ldots \cdot (2n) = (2n)!! = 2^n n!\,. \tag{5.2}$$

For all $a, b \in \mathbb{R}$ and all $n \in \mathbb{N}$, the *binomial theorem* states that

$$(a+b)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!}\, a^{n-k} b^k\,. \tag{5.3}$$

5.1.2 Trigonometric formulas

- **Cofunctions.**

$$\begin{aligned} \sin\left(\frac{\pi}{2} - x\right) &= \cos x\,, \\ \tan\left(\frac{\pi}{2} - x\right) &= \cot x\,, \\ \sec\left(\frac{\pi}{2} - x\right) &= \csc x\,. \end{aligned} \tag{5.4}$$

- **Angle addition.**

$$\sin(x \pm y) = \sin x \cos y \pm \cos x \sin y\,, \tag{5.5}$$
$$\cos(x \pm y) = \cos x \cos y \mp \sin x \sin y\,. \tag{5.6}$$

- **Double angle.**

$$\sin 2x = 2\cos x \sin x\,, \tag{5.7}$$
$$\cos 2x = \cos^2 x - \sin^2 x\,. \tag{5.8}$$

- **Power reduction.**

$$\sin^2 x = \frac{1-\cos 2x}{2}\,, \tag{5.9}$$
$$\cos^2 x = \frac{1+\cos 2x}{2}\,. \tag{5.10}$$

- **Sums of functions.**

$$\sin x + \sin y = 2\sin\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right), \tag{5.11}$$
$$\cos x + \cos y = 2\cos\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right). \tag{5.12}$$

- **Inverse trigonometric functions.**

$$\begin{aligned}\sin(\arccos x) &= \cos(\arcsin x) = \sqrt{1-x^2}\,,\\ \sin(\arctan x) &= \frac{x}{\sqrt{x^2+1}}\,,\\ \cos(\arctan x) &= \frac{1}{\sqrt{x^2+1}}\,.\end{aligned} \tag{5.13}$$

5.1.3 The sine and cosine theorems

In a triangle with sides a, b, c and opposite angles α, β, γ, one has the *law of sines*

$$\frac{\sin\alpha}{a} = \frac{\sin\beta}{b} \tag{5.14}$$

and the *law of cosines*

$$c^2 = a^2 + b^2 - 2ab\cos\gamma\,. \tag{5.15}$$

5.1.4 Cramer's rule

In a system of linear equations with matrix representation $AX = B$, with $X = (x_1, \ldots, x_n)$ and $\det A \neq 0$, one has

$$X = \left(\frac{D_1}{\det A}, \ldots, \frac{D_n}{\det A}\right), \tag{5.16}$$

where D_i is the determinant of the matrix obtained by replacing column i of the square matrix A by the column matrix B.

5.1.5 Derivations of the elementary identities

The summary sheet these identities were collected from states them without proof. They are elementary, but the treatise derives what it states, so the derivations follow here in the order the formulas were listed. Nothing is used beyond the field axioms of $\mathbb{R}$ and $\mathbb{C}$, the functional equation $\mathrm{e}^{z+w} = \mathrm{e}^z\mathrm{e}^w$ of the exponential together with Euler's formula, the inner product of Section 5.3 on the physical space $\mathbb{R}^3$ together with the one geometric fact that the area of a triangle is half its base times the corresponding altitude, and the Leibniz expansion of the determinant, which is written out where it is needed.

Double factorials. Equation (5.1) is a definition: the symbol $(2n+1)!!$ abbreviates the product of the odd numbers up to $2n+1$. Equation (5.2) is not a definition, and follows by extracting one factor 2 from each of the n even factors:

$$\begin{aligned}(2n)!! &= \prod_{k=1}^{n} (2k) \\ &= \left(\prod_{k=1}^{n} 2\right)\left(\prod_{k=1}^{n} k\right) \\ &= 2^n\, n!\,.\end{aligned}$$

The two double factorials are tied to the ordinary one by

$$(2n+1)! = (2n+1)!!\,(2n)!!\,,$$

since the factors of $(2n+1)!$ split without remainder into the odd ones and the even ones.

The binomial theorem. Write $\binom{n}{k} = n!/(k!\,(n-k)!)$ for $0 \le k \le n$, and adopt the convention $\binom{n}{k} = 0$ outside that range. The proof of Equation (5.3) is by induction on n, and it rests on Pascal's rule

$$\binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}, \quad 1 \le k \le n\,, \tag{5.17}$$

which is obtained by putting the two fractions over a common denominator:

$$\begin{aligned}\binom{n}{k-1} + \binom{n}{k} &= \frac{n!}{(k-1)!\,(n-k+1)!} + \frac{n!}{k!\,(n-k)!} \\ &= \frac{n!\,k}{k!\,(n-k+1)!} + \frac{n!\,(n-k+1)}{k!\,(n-k+1)!} \\ &= \frac{n!\,(n+1)}{k!\,(n+1-k)!} \\ &= \binom{n+1}{k}.\end{aligned}$$

For $n = 0$ both sides of Equation (5.3) equal 1. Assume the identity for n. Then

$$\begin{aligned}(a+b)^{n+1} &= (a+b)\sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^k \\ &= \sum_{k=0}^{n} \binom{n}{k} a^{n+1-k}b^k + \sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^{k+1} \\ &= \sum_{k=0}^{n} \binom{n}{k} a^{n+1-k}b^k + \sum_{k=1}^{n+1} \binom{n}{k-1} a^{n+1-k}b^k \\ &= \sum_{k=0}^{n+1} \left[\binom{n}{k} + \binom{n}{k-1}\right] a^{n+1-k}b^k \\ &= \sum_{k=0}^{n+1} \binom{n+1}{k} a^{n+1-k}b^k\,,\end{aligned}$$

where the third line relabels $k+1 \mapsto k$ in the second sum, the fourth uses the vanishing convention to extend both sums to $0 \le k \le n+1$, and the last invokes Equation (5.17). Only commutativity of the product $ab = ba$ was used, so the identity holds in any commutative ring, and in particular for $a, b \in \mathbb{C}$.

Angle addition, and the identities that follow from it. All the trigonometric formulas of Section 5.1.2 descend from Equations (5.5) and (5.6), which in turn descend from Euler's formula $\mathrm{e}^{\mathrm{i}x} = \cos x + \mathrm{i}\sin x$ and the functional equation of the exponential. Indeed,

$$\begin{aligned}\cos(x+y) + \mathrm{i}\sin(x+y) &= \mathrm{e}^{\mathrm{i}(x+y)} \\ &= \mathrm{e}^{\mathrm{i}x}\mathrm{e}^{\mathrm{i}y} \\ &= (\cos x + \mathrm{i}\sin x)(\cos y + \mathrm{i}\sin y) \\ &= (\cos x\cos y - \sin x\sin y) \\ &\quad + \mathrm{i}(\sin x\cos y + \cos x\sin y)\,,\end{aligned}$$

and, x and y being real, the real and imaginary parts may be equated separately, which is the upper sign of Equations (5.5) and (5.6). The lower sign follows on replacing y by $-y$ and using $\mathrm{e}^{-\mathrm{i}y} = (\mathrm{e}^{\mathrm{i}y})^*$, that is, $\cos(-y) = \cos y$ and $\sin(-y) = -\sin y$.

The remaining identities are specializations.

- *Cofunctions* Equation (5.4). Evaluate Equations (5.5) and (5.6) with the lower sign on the pair of angles $\pi/2$ and x, using $\sin(\pi/2) = 1$ and $\cos(\pi/2) = 0$:

$$\begin{aligned}\sin\left(\frac{\pi}{2} - x\right) &= \sin\frac{\pi}{2}\cos x - \cos\frac{\pi}{2}\sin x = \cos x\,, \\ \cos\left(\frac{\pi}{2} - x\right) &= \cos\frac{\pi}{2}\cos x + \sin\frac{\pi}{2}\sin x = \sin x\,.\end{aligned}$$

  Dividing the second into the first gives $\tan(\pi/2 - x) = \cos x/\sin x = \cot x$, and taking the reciprocal of the second gives $\sec(\pi/2 - x) = 1/\sin x = \csc x$.

- *Double angle* Equations (5.7) and (5.8). Put $y = x$ in the addition formulas with the upper sign.

- *Power reduction* Equations (5.9) and (5.10). Combine $\cos 2x = \cos^2 x - \sin^2 x$ with $\cos^2 x + \sin^2 x = 1$ — itself $\left|\mathrm{e}^{\mathrm{i}x}\right|^2 = 1$ — to eliminate first $\cos^2 x$ and then $\sin^2 x$:

$$\cos 2x = 1 - 2\sin^2 x\,, \qquad \cos 2x = 2\cos^2 x - 1\,,$$

  and solve each for the square.

- *Sums of functions* Equations (5.11) and (5.12). Set $s = (x+y)/2$ and $t = (x-y)/2$, so that $x = s + t$ and $y = s - t$. Adding the two addition formulas for $\sin(s \pm t)$ makes the terms in $\cos s\sin t$ cancel, and adding those for $\cos(s \pm t)$ makes the terms in $\sin s\sin t$ cancel:

$$\begin{aligned}\sin x + \sin y &= \sin(s+t) + \sin(s-t) = 2\sin s\cos t\,, \\ \cos x + \cos y &= \cos(s+t) + \cos(s-t) = 2\cos s\cos t\,.\end{aligned}$$

- *Inverse functions* Equation (5.13). Let $\theta = \arccos x \in [0, \pi]$, where $\sin\theta \geq 0$; then $\cos\theta = x$ and $\sin\theta = +\sqrt{1-x^2}$. Let $\varphi = \arcsin x \in [-\pi/2, \pi/2]$, where $\cos\varphi \geq 0$; then $\cos\varphi = +\sqrt{1-x^2}$. This is the first line. For the other two let $\psi = \arctan x \in (-\pi/2, \pi/2)$, where $\cos\psi > 0$; dividing $\cos^2\psi + \sin^2\psi = 1$ by $\cos^2\psi$ gives $1 + x^2 = 1/\cos^2\psi$, whence $\cos\psi = 1/\sqrt{x^2+1}$ and $\sin\psi = x\cos\psi = x/\sqrt{x^2+1}$.

The angle at a vertex. Both theorems of Section 5.1.3 speak of the angle at a vertex of a triangle, and they must mean the same thing by it or their proofs are not proofs of statements about one figure. The notion is therefore fixed once, here, and both derivations are read off it. The triangle is placed in the physical space $\mathbb{R}^3$, carrying the inner product of Section 5.3 — a $3+0$ instantiation, the triangle lying in any plane of it — and the angle between two non-zero vectors u, v is *defined* by

$$\langle u|v\rangle = \|u\|\,\|v\|\cos\theta\,, \qquad \theta \in [0,\pi]\,. \tag{5.18}$$

This determines θ uniquely: the Cauchy–Schwarz inequality Equation (5.43) confines $\langle u|v\rangle\,/(\|u\|\,\|v\|)$ to $[-1,1]$, and the cosine is a bijection from $[0,\pi]$ onto $[-1,1]$. The angle of a triangle at a vertex is the angle so defined between the two sides issuing from that vertex.

The elementary construction the law of sines uses — dropping a perpendicular — measures that same angle, and saying so is what makes the two derivations below derivations about the same α, β, γ. Let u and v issue from a common vertex, let θ be the angle Equation (5.18) between them, and take the line carrying v as the base. The foot of the perpendicular dropped from the tip of u is $\left(\langle v|u\rangle\,/\,\|v\|^2\right)v$, since subtracting it from u leaves a vector orthogonal to v; the altitude is the length of what remains, and

$$\begin{aligned}
\left\| u - \frac{\langle v|u\rangle}{\|v\|^2}\,v \right\|^2 &= \|u\|^2 - \frac{\langle u|v\rangle^2}{\|v\|^2} \\
&= \|u\|^2\left(1-\cos^2\theta\right) && \text{by Equation (5.18)} \\
&= \|u\|^2\sin^2\theta\,,
\end{aligned}$$

the inner product being real and symmetric here. Hence the altitude has length $\|u\|\sin\theta$, the sine being non-negative on $[0,\pi]$. No case distinction between an acute and an obtuse angle is needed: the computation never asked where the foot of the perpendicular falls relative to the endpoints of the base.

The law of sines. Let the triangle have area S, which is half of any base times the corresponding altitude. Take b as the base. The altitude is then dropped from the vertex at which a and c meet, and by the computation just made it has length $c\sin\alpha$ — the side c times the sine of the angle α that c makes with the base b at the vertex the two share — so $S = \frac{1}{2}bc\sin\alpha$. Taking c and then a as the base gives in the same way, cycling $a \to b \to c \to a$ and $\alpha \to \beta \to \gamma \to \alpha$, $S = \frac{1}{2}ca\sin\beta = \frac{1}{2}ab\sin\gamma$. The pairing is worth checking rather than remembering: each expression is the product of the two sides that *meet* at a vertex with the sine of the angle at that same vertex, and the base has to be one of the two factors. That is what forbids pairing the base c with $\frac{1}{2}ab\sin\gamma$ — an expression in which c does not appear at all. Dividing the three equal quantities by $\frac{1}{2}abc$,

$$\frac{\sin\alpha}{a} = \frac{\sin\beta}{b} = \frac{\sin\gamma}{c} = \frac{2S}{abc}\,,$$

of which Equation (5.14) is the first equality.

The law of cosines. The same setting serves, and no second notion of angle is introduced. Place the vertex carrying the angle γ at the origin of $\mathbb{R}^3$ and let $\boldsymbol{u}, \boldsymbol{v}$ be the two sides issuing from it, so that $\|\boldsymbol{u}\| = a$, $\|\boldsymbol{v}\| = b$ and the opposite side is $\boldsymbol{u}-\boldsymbol{v}$, of length c. The angle between them is γ in the sense of Equation (5.18), that is $\langle \boldsymbol{u}|\boldsymbol{v}\rangle = ab\cos\gamma$. Then

$$\begin{aligned}
c^2 &= \langle \boldsymbol{u}-\boldsymbol{v}|\boldsymbol{u}-\boldsymbol{v}\rangle \\
&= \langle \boldsymbol{u}|\boldsymbol{u}\rangle + \langle \boldsymbol{v}|\boldsymbol{v}\rangle - 2\,\langle \boldsymbol{u}|\boldsymbol{v}\rangle \\
&= a^2 + b^2 - 2ab\cos\gamma\,,
\end{aligned}$$

which is Equation (5.15). The middle step uses the distributivity Equations (5.35) and (5.37) and the symmetry of the real inner product Equation (5.33). Note that $\gamma = \pi/2$ returns Pythagoras' theorem.

Cramer's rule. Let A be a square matrix of order n over a field of characteristic zero, with entries $A^i{}_j$ (i the row, j the column) and columns $A_1, \dots, A_n$. The determinant is the Leibniz sum over the permutations σ of n symbols,

$$\det A = \sum_{\sigma} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} A^{\sigma(j)}_{\ j} \,, \tag{5.19}$$

from which the two properties we need follow at once.

- det *is linear in each column separately.* Every term of Equation (5.19) contains exactly one entry from each column, to the first power.
- det *changes sign when two columns are exchanged.* Exchanging columns p and q replaces $\prod_j A^{\sigma(j)}_{\ j}$ by $\prod_j A^{\sigma(\tau(j))}_{\ j}$ with τ the transposition $(p\,q)$; as σ runs over all permutations so does $\sigma\tau$, and $\operatorname{sgn}(\sigma\tau) = -\operatorname{sgn}(\sigma)$. Consequently a matrix with two equal columns has det $= -$ det, hence det $= 0$.

The system $AX = B$ says precisely that B is the combination of the columns of A with the unknowns as coefficients,

$$B = \sum_{j=1}^{n} x_j A_j \,.$$

Substituting this for the i-th column of A and using linearity in that column,

$$\begin{aligned} D_i &= \det\left(A_1, \dots, A_{i-1}, B, A_{i+1}, \dots, A_n\right) \\ &= \sum_{j=1}^{n} x_j \, \det\left(A_1, \dots, A_{i-1}, A_j, A_{i+1}, \dots, A_n\right) \\ &= x_i \det A \,, \end{aligned}$$

because for $j \neq i$ the determinant on the second line has the column A_j twice and therefore vanishes. With $\det A \neq 0$ we may divide, obtaining $x_i = D_i / \det A$, which is Equation (5.16).

5.1.6 The derivative of a determinant

One consequence of the Leibniz expansion is worth stating on its own, because it is what governs how a volume element changes along a motion.

Proposition 5.1 (Jacobi's formula, column form)**.** *Let $A(t)$ be a square array of order n whose entries are differentiable functions of a real parameter t. Then* $\det A$ *is differentiable and*

$$\frac{\mathrm{d}}{\mathrm{d}t} \det A = \sum_{j=1}^{n} \det\left(A_1, \dots, \frac{\mathrm{d}A_j}{\mathrm{d}t}, \dots, A_n\right) \tag{5.20}$$

in terms of the columns A_j of A. If moreover $A(t)$ is invertible then

$$\frac{\mathrm{d}}{\mathrm{d}t} \det A = (\det A)\, \operatorname{tr}\left(A^{-1} \frac{\mathrm{d}A}{\mathrm{d}t}\right) \,. \tag{5.21}$$

Rests on Equation (5.19) and Propositions 5.47 and 7.30.

Proof. *Derives Proposition 5.1.*

By Equation (5.19), $\det A$ is a sum of products of n entries, each entry from a different column. Differentiating a product of n differentiable functions — Proposition 7.30 extended to n factors by induction — replaces it by the sum of the n products in which one factor has been differentiated. Collecting, for each j, the terms in which the differentiated factor came from column j gives exactly the Leibniz expansion of the determinant with that column replaced by its derivative, which is Equation (5.20).

Now let A be invertible, so that its columns are a basis of the coordinate space (Proposition 5.47), and expand each $\frac{\mathrm{d}A_j}{\mathrm{d}t}$ in that basis. The array of coefficients is $B = A^{-1}\frac{\mathrm{d}A}{\mathrm{d}t}$, since $AB = \frac{\mathrm{d}A}{\mathrm{d}t}$ says column by column that $\frac{\mathrm{d}A_j}{\mathrm{d}t} = \sum_i B_{ij}A_i$. Substituting this into the j-th term of Equation (5.20) and using linearity in the j-th column,

$$\det\left(A_1, \ldots, \frac{\mathrm{d}A_j}{\mathrm{d}t}, \ldots, A_n\right) = \sum_{i=1}^{n} B_{ij} \det\left(A_1, \ldots, A_i, \ldots, A_n\right) = B_{jj} \det A\,,$$

because the determinant on the right has the column A_i standing in the j-th place and therefore repeats a column, hence vanishes, unless $i = j$. Summing over j produces $\operatorname{tr} B$ and Equation (5.21). □

5.2 Vector spaces

5.2.1 Vector spaces

Let us discuss the definition of a vector space in more detail. As we have seen (Definition 4.31), the composite structure $[(\mathbb{K}, +, \cdot), \cdot, (\mathbb{V}, +)]$ is a vector space if and only if the following hold.

1. $\mathbb{K}$ is a commutative field.
2. $\mathbb{V}$ is closed with respect to $+$.
3. $\mathbb{V}$ is an abelian group with respect to $+$:
   a) $\forall\, u, v, w \in \mathbb{V},\ u + (v + w) = (u + v) + w$;
   b) $\forall\, u \in \mathbb{V},\ \exists\ 0_{\mathbb{V}} \in \mathbb{V} \mid u + 0_{\mathbb{V}} = u\ \wedge\ 0_{\mathbb{V}} + u = u$;
   c) $\forall\, u \in \mathbb{V},\ \exists\ (-u) \in \mathbb{V} \mid u + (-u) = 0_{\mathbb{V}}\ \wedge\ (-u) + u = 0_{\mathbb{V}}$;
   d) $\forall\, u, v \in \mathbb{V},\ u + v = v + u$.
4. The binary operation $\cdot : \mathbb{K} \times \mathbb{V} \longrightarrow \mathbb{V}$ relates $\mathbb{K}$ and $\mathbb{V}$ in the following manner:
   e) $\forall\, a, b \in \mathbb{K},\ \forall\, u \in \mathbb{V},\ a(bu) = (ab)u$;
   f) $\forall\, u \in \mathbb{V},\ 1u = u$;

   that is, the operation $\cdot$ is an action of $(\mathbb{K}, \cdot)$ on $\mathbb{V}$. Finally, the following distributivity properties hold:
   g) $\forall\, a \in \mathbb{K},\ \forall\, u, v \in \mathbb{V},\ a(u + v) = au + av$;
   h) $\forall\, a, b \in \mathbb{K},\ \forall\, u \in \mathbb{V},\ (a + b)u = au + bu$.

Given a vector space, the following properties hold:

$$\forall\, v \in \mathbb{V},\ 0 \cdot v = 0_{\mathbb{V}} \tag{5.22}$$

and

$$\forall\, a \in \mathbb{K},\ a \cdot 0_{\mathbb{V}} = 0_{\mathbb{V}}\,. \tag{5.23}$$

Derivation. *Derives Equation (5.22).*

Both are consequences of distributivity together with the existence of additive inverses in $(\mathbb{V}, +)$; note that the two zeros involved are different objects, $0 \in \mathbb{K}$ and $0_{\mathbb{V}} \in \mathbb{V}$, and that the content of the statements is exactly that they are compatible.

For Equation (5.22), apply distributivity h) with $a = b = 0$, using $0 + 0 = 0$ in the field:

$$0 \cdot v = (0 + 0) \cdot v = 0 \cdot v + 0 \cdot v \,.$$

By axiom c) the vector $0 \cdot v$ possesses an additive inverse $-(0 \cdot v)$. Adding it on the left of both sides and using associativity a), the inverse axiom c) and the neutral element b),

$$\begin{aligned} -(0 \cdot v) + 0 \cdot v &= -(0 \cdot v) + (0 \cdot v + 0 \cdot v) \\ 0_{\mathbb{V}} &= (-(0 \cdot v) + 0 \cdot v) + 0 \cdot v \\ &= 0_{\mathbb{V}} + 0 \cdot v \\ &= 0 \cdot v \,. \end{aligned}$$

For Equation (5.23) the argument is the same one run on the other distributive law. By axiom b), $0_{\mathbb{V}} = 0_{\mathbb{V}} + 0_{\mathbb{V}}$, so distributivity g) gives

$$a \cdot 0_{\mathbb{V}} = a \cdot (0_{\mathbb{V}} + 0_{\mathbb{V}}) = a \cdot 0_{\mathbb{V}} + a \cdot 0_{\mathbb{V}} \,,$$

and adding $-(a \cdot 0_{\mathbb{V}})$ on the left of both sides yields $0_{\mathbb{V}} = a \cdot 0_{\mathbb{V}}$. □

Two further consequences are worth recording, since they are used silently later: $(-1) \cdot v = -v$, because $v + (-1) \cdot v = (1 + (-1)) \cdot v = 0 \cdot v = 0_{\mathbb{V}}$ by axiom f) and Equation (5.22); and, conversely, $a \cdot v = 0_{\mathbb{V}}$ with $a \neq 0$ forces $v = 0_{\mathbb{V}}$, since multiplying by $a^{-1} \in \mathbb{K}$ — which exists because $\mathbb{K}$ is a field — gives $v = 1 \cdot v = (a^{-1}a) \cdot v = a^{-1} \cdot 0_{\mathbb{V}} = 0_{\mathbb{V}}$ by axioms e), f) and Equation (5.23).

When no ambiguity is possible we shall refer to the vector space as $\mathbb{V}$. When it is necessary to specify the field, we denote it $\mathbb{V}(\mathbb{K})$. If $\mathbb{K} = \mathbb{R}$ we say that $\mathbb{V}$ is a *real vector space*; if $\mathbb{K} = \mathbb{C}$, that $\mathbb{V}$ is a *complex vector space*. From now on we shall consider only the cases $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$.

Definition 5.2 (Scalar)**.** Consider a vector space $\mathbb{V}(\mathbb{K})$. An element belonging to the field is called a *scalar*. We name *zero* the neutral element of $(\mathbb{K}, +)$ and *unity* or *one* the neutral element of $(\mathbb{K}, \cdot)$.

Rests on Definition 4.31.

Definition 5.3 (Vector)**.** Consider a vector space $\mathbb{V}$. An element belonging to the vector space is called a *vector*, and the neutral element $0_{\mathbb{V}}$ is called the *neutral vector*.

Rests on Definition 4.31.

Definition 5.4 (Linear combination)**.** Consider a set of scalars $\{a^i\}_{i=1}^N$ in $\mathbb{K}$ and a set of vectors $W = \{v_i\}_{i=1}^N$ in $\mathbb{V}$. The quantity

$$v = \sum_{i=1}^{N} a^i v_i$$

is a vector, by the closure axioms. This vector is called a *linear combination* of the elements of W.

Rests on Definition 4.31.

Remark 5.5. The concept of linear combination unifies the concepts of multiplication by a scalar and of vector addition: if we wish to prove a property, or state an axiom, both for a vector of the form av and for one of the form $u + v$, it is sometimes simpler to consider a linear combination of u and v, that is, to carry out the proof for a vector of the form $au + bv$. This is at the discretion of the writer.

Definition 5.6 (Vector subspace)**.** Consider a vector space $\mathbb{V}$ and $\mathbb{S} \subset \mathbb{V}$. We say that $\mathbb{S}$ is a *vector subspace* if and only if every linear combination of vectors in $\mathbb{S}$ is a vector in $\mathbb{S}$, that is, if

1. $\mathbb{S} \neq \varnothing$;
2. $\forall\ a, b \in \mathbb{K},\ \forall\ u, v \in \mathbb{S},\ au + bv \in \mathbb{S}$.

Rests on Definition 5.4.

The foregoing says that a vector subspace is a substructure. A vector subspace is evidently a vector space in itself, since:

- $\mathbb{K}$ continues to be a commutative field;
- we have specified the closure of $\mathbb{S}$ with respect to $+$;
- the axioms listed in the definition of a vector space hold for all the elements of $\mathbb{V}$, and in particular for those of $\mathbb{S}$.

Basic vector spaces: real numbers; complex numbers; matrices; functions

Four families of examples carry very nearly every vector space this treatise uses, and in each of them the axioms of Section 5.2.1 are verified in the same way: the operations are defined entry by entry, or point by point, so that each axiom reduces to the corresponding axiom of the field $\mathbb{K}$. The verification is therefore written out once, for the first example, and only its outcome is recorded for the others.

Example 5.7 (The coordinate spaces $\mathbb{R}^n$ and $\mathbb{C}^n$)**.** Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ and let $\mathbb{K}^n$ be the set of ordered n-tuples $v = (v^1, \dots, v^n)$ of elements of $\mathbb{K}$, with the componentwise operations

$$(u + v)^i = u^i + v^i\,, \qquad (av)^i = a\,v^i\,, \qquad i = 1, \dots, n\,. \tag{5.24}$$

Closure is immediate, since $\mathbb{K}$ is closed under its own operations. Associativity and commutativity of $+$ hold because they hold in $\mathbb{K}$ in each slot separately; the neutral vector is $0_{\mathbb{V}} = (0, \dots, 0)$ and the opposite of v is $(-v^1, \dots, -v^n)$; and the four axioms e) to h) relating the field to the space are the associativity, the unit, and the two distributive laws of $\mathbb{K}$, read one component at a time. The n vectors

$$e_i = (0, \dots, 0, \underset{i}{1}, 0, \dots, 0)\,, \qquad i = 1, \dots, n\,, \tag{5.25}$$

form a basis in the sense of Definition 5.14: the identity $v = \sum_i v^i e_i$ exhibits an arbitrary v as a linear combination of them, and $\sum_i a^i e_i = 0_{\mathbb{V}}$ reads $a^i = 0$ componentwise, which is Equation (5.29). Hence $\dim \mathbb{K}^n = n$ over $\mathbb{K}$. Two degenerate cases deserve their names: $\mathbb{R}^1 = \mathbb{R}$ is a real vector space of dimension one, and $\mathbb{C}$ is a complex vector space of dimension one but a *real* vector space of dimension two, with basis $\{1, \mathrm{i}\}$ — the distinction insisted on in Remark 5.53. Physical space is the case $n = 3$ over $\mathbb{R}$ and the spacetime of Chapter 39 the case $n = 4$ over $\mathbb{R}$.

Rests on Definitions 4.31 and 5.4.

Example 5.8 (The matrices of a given shape)**.** Let $\mathcal{M}_{m\times n}[\mathbb{K}]$ be the set of rectangular arrays $A = \left(A^i{}_j\right)$ of m rows and n columns with entries in $\mathbb{K}$, added and scaled entry by entry,

$$(A + B)^i{}_j = A^i{}_j + B^i{}_j\,, \qquad (aA)^i{}_j = a\,A^i{}_j\,.$$

This is Example 5.7 with the single index i replaced by the pair (i, j), so every axiom holds for the same reason, the neutral vector being the array all of whose entries are 0. A

basis is furnished by the mn arrays $E_{(kl)}$ carrying 1 in the entry (k,l) and 0 everywhere else: $A = \sum_{k,l} A^k{}_l E_{(kl)}$, and a vanishing combination of the $E_{(kl)}$ has every coefficient zero because it has every entry zero. Hence

$$\dim \mathcal{M}_{m\times n}[\mathbb{K}] = mn\,. \tag{5.26}$$

Note what is *not* used here: the matrix product plays no part, and the square arrays acquire the further structure of an algebra (Proposition 5.121) only once it is added.

Rests on Definitions 4.31 and 5.14.

Example 5.9 (Functions on a set)**.** Let X be a non-empty set and let $\mathcal{F}(X,\mathbb{R})$ be the set of all maps $f : X \longrightarrow \mathbb{R}$, with the *pointwise* operations

$$(f+g)(x) = f(x) + g(x)\,, \qquad (af)(x) = a\,f(x)\,, \qquad \forall\, x \in X\,. \tag{5.27}$$

Once more each axiom is an axiom of $\mathbb{R}$ evaluated at one point at a time; the neutral vector is the function constant equal to 0 and the opposite of f is $-f$. Many of the spaces met later in the treatise are subspaces of this one in the sense of Definition 5.6 — the polynomial functions of degree at most N, the continuous functions on an interval, the solutions of a homogeneous linear differential equation — and each is a subspace for the same reason: the defining condition is preserved by linear combinations.

The dimension separates two regimes. If X is finite with N elements $x_1,\dots,x_N$, then $f \longmapsto (f(x_1),\dots,f(x_N))$ is a bijection onto $\mathbb{R}^N$ respecting Equation (5.27), so $\mathcal{F}(X,\mathbb{R})$ is a space of dimension N and nothing new. If X is infinite — $X = \mathbb{R}$, say — the space is infinite-dimensional in the sense of Definition 5.15: the monomial functions $p_k : x \longmapsto x^k$, $k = 0,1,2,\dots$, are linearly independent, because a vanishing linear combination $\sum_{k=0}^{K} a^k x^k = 0$ for *every* real x is a polynomial with infinitely many roots and hence has all its coefficients zero. No finite set can therefore generate $\mathcal{F}(\mathbb{R},\mathbb{R})$.

Rests on Definitions 4.31 and 5.6.

Remark 5.10 (Why the last example forces a later chapter)*.* Example 5.9 is the first space in this chapter whose dimension is not a natural number, and the difference is not one of degree. Everything proved from Definition 5.11 onwards rests on *finite* linear combinations (Remark 5.12), whereas the expansions physics actually writes in function spaces — a Fourier series, a power series, an expansion of a quantum state in energy eigenstates — are infinite sums, and an infinite sum means nothing at all until a notion of convergence is supplied. Supplying it takes a norm (Definition 5.18) and then completeness with respect to the associated metric Equation (5.45); that is exactly the equipment of a Hilbert space, and it is why Chapter 12 exists as a separate chapter rather than as a paragraph of this one. The state space of Postulate 77.1 is a function space of this kind.

5.2.2 Bases

Definition 5.11 (Subspace generated by a set of vectors)**.** Consider a non-empty set of vectors $W \subset \mathbb{V}$. The set of all possible linear combinations of the elements of W forms a vector subspace of $\mathbb{V}$. We denote this subspace by $\langle W\rangle$ and say that it is the *subspace generated* by W.

Rests on Definitions 5.4 and 5.6.

Derivation. *Derives Definition 5.11.*

Write out what the set is:

$$\langle W\rangle = \left\{ \sum_{i=1}^{N} a^i w_i \;\middle|\; N \in \mathbb{N},\ a^i \in \mathbb{K},\ w_i \in W \right\},$$

the linear combinations being *finite* — N is arbitrary but each element of $\langle W\rangle$ uses only finitely many vectors of W. We verify the two conditions of Definition 5.6.

1. $\langle W\rangle \neq \varnothing$. By hypothesis $W \neq \varnothing$, so pick $w \in W$; then $w = 1 \cdot w \in \langle W\rangle$ by axiom f), the case $N = 1$, $a^1 = 1$.

2. Closure under linear combination. Let $u, v \in \langle W\rangle$ and $c, d \in \mathbb{K}$. By definition there are finite families
$$u = \sum_{i=1}^{N} a^i w_i\,, \qquad v = \sum_{j=1}^{M} b^j w'_j\,,$$
with all $w_i, w'_j \in W$. Then
$$\begin{aligned} cu + dv &= c\sum_{i=1}^{N} a^i w_i + d\sum_{j=1}^{M} b^j w'_j \\ &= \sum_{i=1}^{N} c\left(a^i w_i\right) + \sum_{j=1}^{M} d\left(b^j w'_j\right) && \text{by } g) \\ &= \sum_{i=1}^{N} \left(ca^i\right) w_i + \sum_{j=1}^{M} \left(db^j\right) w'_j && \text{by } e)\,, \end{aligned}$$
which is again a finite linear combination of elements of W, with coefficients $ca^i, db^j \in \mathbb{K}$ by closure of the field. Hence $cu + dv \in \langle W\rangle$.

So $\langle W\rangle$ is a vector subspace. Note in passing that $0_{\mathbb{V}} = 0 \cdot w \in \langle W\rangle$ by Equation (5.22), as it must be for a subspace. □

Remark 5.12 ($\langle W\rangle$ is the smallest such subspace). If $\mathbb{S}$ is any vector subspace with $W \subset \mathbb{S}$, then $\langle W\rangle \subset \mathbb{S}$: applying condition 2 of Definition 5.6 repeatedly to the elements $w_1, \ldots, w_N \in W \subset \mathbb{S}$ shows that every finite linear combination of them lies in $\mathbb{S}$. Thus $\langle W\rangle$ is the intersection of all subspaces containing W, which is the usual characterization of the generated subspace. The restriction to *finite* combinations is not a technicality that could be relaxed: an infinite sum presupposes a notion of convergence, which a bare vector space does not carry — it needs the norm of Definition 5.18 and, beyond that, completeness. This is exactly the difference between the algebraic (Hamel) basis of Definition 5.14 and the topological bases used in infinite dimensions.

Definition 5.13 (Linear independence)**.** A set of vectors $\{v_i\}_{i=1}^{N}$ is *linearly independent* (LI) if none of them can be written as a linear combination of the remaining ones, that is, if and only if
$$\forall\ \left\{a^i\right\}_{i=1}^{N}, \quad \sum_{i=1}^{N} a^i v_i = v_j \implies a^k = 0_{\mathbb{K}},\ \forall j,\ \forall k\,, \tag{5.28}$$
where $\left\{a^i\right\}_{i=1}^{N} \subset \mathbb{K}$ is an arbitrary set of an arbitrary (but fixed) number N of scalars in $\mathbb{K}$.
Rests on Definition 5.4.

It is important to mention that Equation (5.28) is equivalent to
$$\forall\ \left\{a^i\right\}_{i=1}^{N+1}, \quad \sum_{i=1}^{N+1} a^i v_i = 0_{\mathbb{V}} \implies a^k = 0_{\mathbb{K}},\ \forall k\,, \tag{5.29}$$
although, since N is arbitrary, we may leave the characterization of linear independence as:

1. $\forall\ \{a^i\}_{i=1}^{N+1}, \quad \sum_{i=1}^{N+1} a^i v_i = 0_{\mathbb{V}} \implies a^k = 0_{\mathbb{K}},\ \forall k.$

Definition 5.14 (Basis)**.** Let $B \subset \mathbb{V}$ be a subset of vectors. We say that B is a *basis* of $\mathbb{V}$ if and only if

1. $\langle B \rangle = \mathbb{V}$;
2. B is LI.

Rests on Definitions 5.11 and 5.13.

The basis of a vector space is not unique, since:

- given a basis, we may multiply each of its elements by a scalar, and the new set will still generate the whole space;
- there may exist another set of LI vectors generating the whole space.

Definition 5.15 (Dimension)**.** Let B be a basis of $\mathbb{V}$, with $|B| = n$. We say that $\mathbb{V}$ is a *finite-dimensional* vector space of *dimension* n if and only if $n \in \mathbb{N}$. We say that $\mathbb{V}$ is an *infinite-dimensional* vector space if and only if $n = \infty$.
Rests on Definition 5.14.

When it is prudent to specify the dimension of an n-dimensional vector space we shall denote it $\mathbb{V}(n, \mathbb{K})$, or simply $\mathbb{V}_n$.

Definition 5.16 (Components of a vector)**.** Let $\mathbb{V} = \langle \{e_i\}_{i=1} \rangle$ be a vector space. Since $\mathbb{V}$ is generated by the basis $\{e_i\}_{i=1}$, we can write any vector $v \in \mathbb{V}$ as a linear combination of the elements of that basis, that is,

$$v = \sum_{i=1} v^i e_i \,. \tag{5.30}$$

We say that the quantities $v^i \in \mathbb{K}$ are the *components* of the vector v with respect to the basis $\{e_i\}_{i=1}$.
Rests on Definition 5.14.

It is important to mention that the components of a vector are always *basis-dependent*: if we choose another basis generating $\mathbb{V}$, the components will be different.

5.3 Inner product, norm, and metric

5.3.1 Inner product

Definition 5.17 (Inner product)**.** Consider the map

$$\begin{array}{rccc} \langle \,|\, \rangle : & \mathbb{V} \times \mathbb{V} & \longrightarrow & \mathbb{K} \\ & (u, v) & \longmapsto & \langle u|v \rangle \,. \end{array}$$

We say that $\langle \,|\, \rangle$ defines an *inner product* on $\mathbb{V}$ if and only if it satisfies the following properties:

1. Identity of indiscernibles:
$$\langle v|v \rangle = 0 \iff v = 0_{\mathbb{V}} \,. \tag{5.31}$$
2. Positive definiteness:
$$\forall\, v \in \mathbb{V} \setminus \{0_{\mathbb{V}}\} \,,\ \langle v|v \rangle > 0 \,. \tag{5.32}$$
3. Hermiticity:
$$\forall\, u, v \in \mathbb{V},\ \langle u|v \rangle^* = \langle v|u \rangle \,. \tag{5.33}$$
4. Product by a scalar:
$$\forall\, u, v \in \mathbb{V},\ \forall\, a \in \mathbb{K},\ \langle u|av \rangle = a \, \langle u|v \rangle \,. \tag{5.34}$$

5. Distributivity:
$$\forall\, u, v, w \in \mathbb{V},\ \langle u|v+w\rangle = \langle u|v\rangle + \langle u|w\rangle\ . \tag{5.35}$$

Rests on Definition 4.31.

Given an inner product, the following properties hold as direct consequences of the axioms.

- From Equation (5.33) we see that
$$\begin{aligned}\langle au|v\rangle &= \langle v|au\rangle^* \\ &= (a\,\langle v|u\rangle)^* && \text{by Equation (5.34)} \\ &= a^*\,\langle v|u\rangle^*\ ,\end{aligned}$$
so that
$$\forall\, u, v \in \mathbb{V},\ \forall\, a \in \mathbb{K},\ \langle au|v\rangle = a^*\,\langle u|v\rangle\ . \tag{5.36}$$
- Again from Equation (5.33) we see that
$$\begin{aligned}\langle u+v|w\rangle &= \langle w|u+v\rangle^* \\ &= (\langle w|u\rangle + \langle w|v\rangle)^* && \text{by Equation (5.35)} \\ &= \langle w|u\rangle^* + \langle w|v\rangle^*\ ,\end{aligned}$$
hence
$$\forall\, u, v, w \in \mathbb{V},\ \langle u+v|w\rangle = \langle u|w\rangle + \langle v|w\rangle\ . \tag{5.37}$$
- From Equation (5.34) and Equation (5.22) we see that
$$\begin{aligned}\langle u|0_{\mathbb{V}}\rangle &= \langle u|0\cdot v\rangle \\ &= 0\,\langle u|v\rangle\ ,\end{aligned}$$
hence
$$\forall\, u \in \mathbb{V},\ \langle u|0_{\mathbb{V}}\rangle = 0\,. \tag{5.38}$$
- The inner product of a vector with itself is always a real number. Indeed, from hermiticity we have $\langle v|v\rangle = \langle v|v\rangle^*$, and, as we know from the complex numbers, every number equal to its conjugate is real. Hence
$$\forall\, v \in \mathbb{V},\ \langle v|v\rangle \in \mathbb{R}\,. \tag{5.39}$$

On occasion we shall call a vector space equipped with an inner product a *pre-Hilbert space*.

Definition 5.18 (Norm)**.** Consider the map
$$\begin{array}{rccc} \|\ \|: & \mathbb{V} & \longrightarrow & \mathbb{R} \\ & v & \longmapsto & \|v\|\ . \end{array}$$
We say that the map $\|\ \|$ defines a *norm* on $\mathbb{V}$ if and only if the following properties hold:

1. Identity of indiscernibles:
$$\|v\| = 0 \iff v = 0_{\mathbb{V}}\,. \tag{5.40}$$
2. Product by a scalar:
$$\forall\, v \in \mathbb{V},\ \forall\, a \in \mathbb{K},\ \|av\| = |a|\,\|v\|\ . \tag{5.41}$$
3. Triangle inequality:
$$\forall\, u, v \in \mathbb{V},\ \|u+v\| \leq \|u\| + \|v\|\ . \tag{5.42}$$

Rests on Definition 4.31.

We say that a vector space equipped with a norm is a *normed space*.

The Schwarz inequality

The single inequality that the whole of the theory of inner-product spaces rests on is the following. It is used at once to prove that Equation (5.44) really defines a norm, and again in Chapter 6 for the metric axioms of $\mathbb{R}^N$.

Proposition 5.19 (Cauchy–Schwarz inequality)**.** *Let $\mathbb{V}(\mathbb{K})$, with $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, be a vector space equipped with an inner product in the sense of Definition 5.17. Then*

$$|\langle u|v\rangle|^2 \le \langle u|u\rangle\,\langle v|v\rangle\ , \qquad \forall\, u, v \in \mathbb{V}\,, \tag{5.43}$$

with equality if and only if u and v are linearly dependent.
Rests on Definitions 5.13 and 5.17.

Proof. *Derives Proposition 5.19.*

If $v = 0_{\mathbb{V}}$ both sides vanish, by Equation (5.38) on the left and Equation (5.31) on the right, and u and $0_{\mathbb{V}}$ are linearly dependent; so assume $v \neq 0_{\mathbb{V}}$, whence $\langle v|v\rangle > 0$ by Equation (5.32) and we may divide by it. Put

$$\lambda = \frac{\langle v|u\rangle}{\langle v|v\rangle} \in \mathbb{K}\,, \qquad w = u - \lambda v\,,$$

that is, subtract from u its component along v. Expanding $\langle w|w\rangle$ by distributivity Equations (5.35) and (5.37) and the two scalar rules Equations (5.34) and (5.36),

$$\begin{aligned}\langle w|w\rangle &= \langle u - \lambda v|u - \lambda v\rangle \\ &= \langle u|u\rangle - \lambda\,\langle u|v\rangle - \lambda^*\,\langle v|u\rangle + \lambda^*\lambda\,\langle v|v\rangle\ .\end{aligned}$$

Now $\langle v|u\rangle = \langle u|v\rangle^*$ by hermiticity Equation (5.33), and $\langle v|v\rangle$ is a positive real by Equation (5.39), so each of the last three terms is $|\langle u|v\rangle|^2 / \langle v|v\rangle$, with signs $-, -, +$:

$$\begin{aligned}\langle w|w\rangle &= \langle u|u\rangle - \frac{|\langle u|v\rangle|^2}{\langle v|v\rangle} - \frac{|\langle u|v\rangle|^2}{\langle v|v\rangle} + \frac{|\langle u|v\rangle|^2}{\langle v|v\rangle} \\ &= \langle u|u\rangle - \frac{|\langle u|v\rangle|^2}{\langle v|v\rangle}\,.\end{aligned}$$

By positive definiteness Equation (5.32) together with Equation (5.31) we have $\langle w|w\rangle \ge 0$, and multiplying through by $\langle v|v\rangle > 0$ gives Equation (5.43).

For the equality case: the chain above shows that equality holds if and only if $\langle w|w\rangle = 0$, that is — by Equation (5.31) — if and only if $u = \lambda v$, and a pair $\{u, v\}$ with $v \neq 0_{\mathbb{V}}$ is linearly dependent precisely when u is a multiple of v. □

Remark 5.20 (What the proof does *not* need). Only the five axioms of Definition 5.17 were used. Neither finite-dimensionality, nor a basis, nor completeness of $\mathbb{V}$ entered anywhere, so Proposition 5.19 holds verbatim in the infinite-dimensional pre-Hilbert spaces of quantum mechanics (Chapter 12). What positive definiteness buys is the division by $\langle v|v\rangle$; on a space carrying an indefinite form — the Minkowski form of Chapter 39 is the case physics meets first — the inequality fails, and in fact reverses on the timelike vectors.

Norm associated with an inner product

Consider a vector space $\mathbb{V}$. Given an inner product $\langle\ |\ \rangle$ on $\mathbb{V}$, we can automatically define a map that satisfies the axioms of a norm, namely

$$\|v\| = \sqrt{\langle v|v\rangle}\,. \tag{5.44}$$

Let us prove that the map defined in Equation (5.44) is a norm.

1. For the identity of indiscernibles:

   ⇒) We must prove that if
   $$\|v\| = \sqrt{\langle v|v\rangle} = 0\,,$$
   then $v = 0_{\mathbb{V}}$. Indeed, squaring the above equation we see that $\langle v|v\rangle = 0$; hence, by axiom Equation (5.31), necessarily $v = 0_{\mathbb{V}}$.

   ⇐) We must now prove that if $v = 0_{\mathbb{V}}$ then $\|v\| = 0$. Since $v = 0_{\mathbb{V}}$, by axiom Equation (5.31) necessarily $\langle v|v\rangle = 0$; substituting into Equation (5.44),
   $$\|v\| = \sqrt{0} = 0\,.$$

2. We must prove that $\|av\| = |a|\,\|v\|$. Indeed,
   $$\begin{aligned}\|av\| &= \sqrt{\langle av|av\rangle}\\ &= \sqrt{a^*a\,\langle v|v\rangle} && \text{by Equations (5.34) and (5.36)}\\ &= \sqrt{|a|^2\,\langle v|v\rangle}\\ &= |a|\,\sqrt{\langle v|v\rangle}\\ &= |a|\,\|v\|\ .\end{aligned}$$

3. It remains to prove that this norm satisfies the triangle inequality. This is the one step that is not immediate from the axioms of the inner product, and it is where Proposition 5.19 is needed. Since both sides of Equation (5.42) are non-negative reals, it is equivalent — and easier — to compare their squares. Expanding by distributivity Equations (5.35) and (5.37),
   $$\begin{aligned}\|u+v\|^2 &= \langle u+v|u+v\rangle\\ &= \langle u|u\rangle + \langle u|v\rangle + \langle v|u\rangle + \langle v|v\rangle\\ &= \|u\|^2 + \langle u|v\rangle + \langle u|v\rangle^* + \|v\|^2 && \text{by Equation (5.33)}\\ &= \|u\|^2 + 2\,\mathrm{Re}\,\langle u|v\rangle + \|v\|^2\ ,\end{aligned}$$
   the cross terms having combined into twice a real part because a complex number added to its conjugate is twice its real part. Now $\mathrm{Re}\,z \le |z|$ for any $z \in \mathbb{C}$, and Equation (5.43) bounds $|\langle u|v\rangle|$ by $\|u\|\,\|v\|$, so
   $$\begin{aligned}\|u+v\|^2 &\le \|u\|^2 + 2\,|\langle u|v\rangle| + \|v\|^2\\ &\le \|u\|^2 + 2\,\|u\|\,\|v\| + \|v\|^2 && \text{by Equation (5.43)}\\ &= (\|u\| + \|v\|)^2\ .\end{aligned}$$
   Taking the (non-negative) square root of both sides preserves the inequality, and Equation (5.42) follows. Equality demands equality in both steps: $\langle u|v\rangle$ real and non-negative, and u, v linearly dependent — that is, u and v must be non-negative multiples of one another, the vectors "pointing the same way", which is the geometric content of the triangle inequality degenerating into an equality.

Metric associated with a norm

Consider a vector space $\mathbb{V}$. Given a norm $\|\ \|$ defined on $\mathbb{V}$, we can define a map that satisfies the axioms of a metric (Definition 6.24), namely
$$d(u,v) = \|u - v\|\ . \tag{5.45}$$

Let us prove that the map defined by Equation (5.45) really satisfies the three axioms of Definition 6.24, in their order.

1. *Positivity and the identity of indiscernibles* Equation (6.7). The first half is not an axiom of Definition 5.18 and has to be derived: for any $w \in \mathbb{V}$,

$$\begin{aligned} 0 &= \|0_{\mathbb{V}}\| && \text{by Equation (5.40)} \\ &= \|w + (-w)\| \\ &\leq \|w\| + \|-w\| && \text{by Equation (5.42)} \\ &= \|w\| + |-1|\,\|w\| && \text{by Equation (5.41)} \\ &= 2\,\|w\|\,, \end{aligned}$$

so $\|w\| \geq 0$, and with $w = u - v$ this is $d(u, v) \geq 0$. The second half is Equation (5.40) read on $w = u - v$:

$$d(u, v) = 0 \iff \|u - v\| = 0 \iff u - v = 0_{\mathbb{V}} \iff u = v\,,$$

the last step because $\mathbb{V}$ is a group under addition. In particular $d(v, v) = \|0_{\mathbb{V}}\| = 0$.

2. *Symmetry* Equation (6.8). We have

$$\begin{aligned} d(u, v) &= \|u - v\| \\ &= \|-(v - u)\| \\ &= |-1|\,\|v - u\| && \text{by Equation (5.41)} \\ &= \|v - u\| \\ &= d(v, u)\,, \end{aligned}$$

so symmetry holds.

3. *The triangle inequality* Equation (6.9). Finally, we must prove that the triangle inequality holds for the metric we have defined. We have

$$\begin{aligned} d(u, v) &= \|u - v\| \\ &= \|u - w + w - v\| \\ &\leq \|u - w\| + \|w - v\| && \text{by Equation (5.42)} \\ &= d(u, w) + d(w, v)\,, \end{aligned}$$

so the triangle inequality is indeed satisfied.

Remark 5.21 (On the inner product, the norm, and the metric). Note the following. Given a vector space equipped with an inner product, we can always define a norm associated with that inner product, and therefore we can always define a metric associated with that norm. This is a direct consequence of the axioms we have laid down for those maps. Even so, one could define an inner product, a norm, and a metric separately, that is, not associated with one another.

The axioms defined for an inner product, a norm, and a metric are motivated when these maps are indeed associated. For example, the inner product must be positive definite because otherwise the associated norm (defined in Equation (5.44)) could be imaginary, and the triangle inequality Equation (5.42) would then be meaningless, since imaginary numbers are neither smaller nor greater than one another.

The norm defined in Equation (5.44) is consistent because, by property Equation (5.39), the inner product of a vector with itself is always a real number; hence the square root in Equation (5.44) is always real as well.

5.3.2 Orthogonal bases

Definition 5.22 (Orthogonal vectors)**.** Consider a vector space equipped with an inner product. We say that two vectors $u, v \in \mathbb{V}$ are *orthogonal* if and only if their inner product vanishes, that is, if

$$\langle u|v\rangle = 0\,. \tag{5.46}$$

Rests on Definition 5.17.

Definition 5.23 (Unit vector)**.** Consider a vector space equipped with a norm. We say that a vector $v \in \mathbb{V}$ is *normalized*, or is a *unit vector*, if and only if its norm is one, that is, if $\|v\| = 1$.

Rests on Definition 5.18.

Definition 5.24 (Orthogonal basis)**.** Consider a vector space equipped with an inner product, generated by the basis $\{e_i\}_{i=1}$. We say that the basis is *orthogonal* if and only if all its elements are mutually orthogonal, that is, if

$$\langle e_i|e_j\rangle = a\,\delta_{ij},\ \forall\, i, j\,. \tag{5.47}$$

Rests on Definitions 5.14 and 5.22.

Definition 5.25 (Orthonormal basis)**.** Consider a vector space equipped with an inner product and a norm (not necessarily the associated one). We say that the basis is *orthonormal* if and only if it is orthogonal and, in addition, all its elements are unit vectors, that is, if

$$\|e_i\| = 1,\ \forall\, i\,. \tag{5.48}$$

Rests on Definitions 5.23 and 5.24.

If we consider the norm associated with the inner product, defined in Equation (5.44), it is clear that the orthonormality condition is given by the expression

$$\langle e_i|e_j\rangle = \delta_{ij},\ \forall\, i, j\,. \tag{5.49}$$

The Gram–Schmidt orthogonalization process

An orthonormal basis is so much more convenient than a general one that it is worth knowing that every finite-dimensional inner-product space has one. The construction is due to Gram [223] and Schmidt [338], and it is completely explicit: it turns a basis into an orthonormal basis one vector at a time, at each step subtracting from the next vector its components along the vectors already orthonormalized.

Proposition 5.26 (Gram–Schmidt)**.** *Let $\mathbb{V}$ be a vector space with an inner product and let $\{v_1, \ldots, v_n\}$ be a linearly independent set. Then there exists an orthonormal set $\{e_1, \ldots, e_n\}$ with*

$$\langle\{e_1, \ldots, e_k\}\rangle = \langle\{v_1, \ldots, v_k\}\rangle\,, \qquad k = 1, \ldots, n\,. \tag{5.50}$$

In particular every finite-dimensional inner-product space, and every finite-dimensional subspace of an inner-product space, possesses an orthonormal basis.

Rests on Definitions 5.13, 5.17 and 5.25 and Equation (5.44).

Proof. *Derives Proposition 5.26.*

By induction on k. For $k = 1$: $v_1 \neq 0_{\mathbb{V}}$ by linear independence, so $\|v_1\| > 0$ by Equations (5.32) and (5.44), and $e_1 = v_1/\|v_1\|$ is a unit vector generating the same line, by Equation (5.41).

Suppose $e_1, \dots, e_{k-1}$ have been constructed, orthonormal and satisfying Equation (5.50). Put

$$\tilde{e}_k = v_k - \sum_{i=1}^{k-1} \langle e_i | v_k \rangle \, e_i \,, \qquad e_k = \frac{\tilde{e}_k}{\|\tilde{e}_k\|} \,. \tag{5.51}$$

The subtraction is legitimate — $\tilde{e}_k \neq 0_{\mathbb{V}}$ — because otherwise v_k would lie in $\langle \{e_1, \dots, e_{k-1}\} \rangle = \langle \{v_1, \dots, v_{k-1}\} \rangle$, contradicting linear independence. For $j < k$,

$$\begin{aligned}
\langle e_j | \tilde{e}_k \rangle &= \langle e_j | v_k \rangle - \sum_{i=1}^{k-1} \langle e_i | v_k \rangle \, \langle e_j | e_i \rangle && \text{by Equations (5.34) and (5.35)} \\
&= \langle e_j | v_k \rangle - \sum_{i=1}^{k-1} \langle e_i | v_k \rangle \, \delta_{ji} && \text{by Equation (5.49)} \\
&= \langle e_j | v_k \rangle - \langle e_j | v_k \rangle \\
&= 0 \,,
\end{aligned}$$

so e_k is orthogonal to all its predecessors and, by construction, a unit vector. Equation (5.50) holds at step k because Equation (5.51) expresses e_k in terms of $v_k, e_1, \dots, e_{k-1}$ and, read backwards, expresses v_k in terms of $e_k, \dots, e_1$; each generated subspace therefore contains the other. □

Corollary 5.27 (Orthogonal decomposition)**.** *Let $\mathbb{S} \leq \mathbb{V}$ be a finite-dimensional subspace of an inner-product space and let*

$$\mathbb{S}^{\perp} = \{ w \in \mathbb{V} \mid \langle s | w \rangle = 0, \ \forall \, s \in \mathbb{S} \} \tag{5.52}$$

be its orthogonal complement*. Then $\mathbb{S}^{\perp}$ is a subspace and*

$$\mathbb{V} = \mathbb{S} \oplus \mathbb{S}^{\perp} \,. \tag{5.53}$$

Rests on Proposition 5.26 and Definitions 5.6 and 5.17.

Proof. *Derives Corollary 5.27.*

$\mathbb{S}^{\perp}$ is a subspace because $\langle s | a w_1 + b w_2 \rangle = a \langle s | w_1 \rangle + b \langle s | w_2 \rangle = 0$ by Equations (5.34) and (5.35), and it is non-empty ($0_{\mathbb{V}} \in \mathbb{S}^{\perp}$ by Equation (5.38)). Let $\{e_1, \dots, e_m\}$ be an orthonormal basis of $\mathbb{S}$, which exists by Proposition 5.26, and for $v \in \mathbb{V}$ define

$$Pv = \sum_{i=1}^{m} \langle e_i | v \rangle \, e_i \in \mathbb{S} \,.$$

The computation just performed in Proposition 5.26 shows $\langle e_j | v - Pv \rangle = 0$ for every j, hence $v - Pv \in \mathbb{S}^{\perp}$, so $v = Pv + (v - Pv)$ exhibits $\mathbb{V}$ as a sum. The sum is direct: if $s \in \mathbb{S} \cap \mathbb{S}^{\perp}$ then $\langle s | s \rangle = 0$, so $s = 0_{\mathbb{V}}$ by Equation (5.31). □

Remark 5.28 (Where the finiteness is needed)*.* The recursion Equation (5.51) is finite at every step, so Proposition 5.26 applies verbatim to a countable linearly independent family and produces a countable orthonormal family. What does *not* survive to infinite dimension is Equation (5.53) for an arbitrary subspace: the projection P is then an infinite sum, which needs convergence, and the decomposition holds for *closed* subspaces of a *complete* inner-product space — a Hilbert space, Definition 12.2. The projection theorem that says so is not proved in this treatise: it is reserved for Section 12.2.1, where it belongs. Every use made of Corollary 5.27 in this chapter is in finite dimension, where no such caveat arises.

The Gram matrix

Gram–Schmidt answers whether an orthonormal basis exists. The companion question — whether a given family is independent at all — has an answer in the same terms, and it is the form in which independence is usually tested in practice, because it replaces a statement about vanishing linear combinations by the definiteness of a single matrix assembled from inner products.

Proposition 5.29 (Gram criterion)**.** *Let* $\mathbb{V}$ *be a real vector space with an inner product and let* $v_1, \dots, v_m \in \mathbb{V}$*. Their* Gram matrix

$$G_{kl} = \langle v_k | v_l \rangle \,, \qquad k, l = 1, \dots, m \,, \tag{5.54}$$

is symmetric and positive semi-definite, and the following three conditions are equivalent: the family $\{v_1, \dots, v_m\}$ *is linearly independent;* G *is positive definite;* $\det G \neq 0$*.*

Rests on Definitions 5.13 and 5.17 and Proposition 5.47.

Proof. *Derives Proposition 5.29.*

Symmetry is Equation (5.33), the conjugation being the identity on $\mathbb{R}$. For $c = (c^1, \dots, c^m) \in \mathbb{R}^m$ put $w = \sum_{k=1}^m c^k v_k$; expanding by Equations (5.34) to (5.37),

$$\sum_{k,l=1}^{m} c^k G_{kl} c^l = \langle w | w \rangle \geq 0 \tag{5.55}$$

by Equation (5.32), which is positive semi-definiteness; and by Equation (5.31) the value vanishes exactly when $w = 0_{\mathbb{V}}$.

If the family is independent, $w = 0_{\mathbb{V}}$ forces $c = 0$, so Equation (5.55) is strictly positive for every $c \neq 0$ and G is positive definite. If it is dependent, some $c \neq 0$ gives $w = 0_{\mathbb{V}}$ and the quadratic form vanishes there, so G is not positive definite. For the third condition: if G is positive definite and $Gc = 0$, then Equation (5.55) vanishes at c, hence $c = 0$, so the array is injective and $\det G \neq 0$ by Proposition 5.47 and Equation (5.81). Conversely, if G is not positive definite there is a $c \neq 0$ with $\langle w | w \rangle = 0$, hence $w = 0_{\mathbb{V}}$, hence $\sum_l G_{kl} c^l = \langle v_k | w \rangle = 0$ for every k; so $Gc = 0$ with $c \neq 0$ and $\det G = 0$. □

5.3.3 Vector algebra in three dimensions

Physical space is $\mathbb{R}^3$ carrying the inner product of Definition 5.17 — the $3 + 0$ instantiation of Example 5.114. Its vector algebra is used in every physical part of this treatise, and the handful of identities that carry all of it are collected and proved here. Throughout this subsection the basis is orthonormal and fixed, components are written with subscripts, and a repeated index is summed over $1, 2, 3$.

Definition 5.30 (Levi–Civita symbol; cross product)**.** The *Levi–Civita symbol* is the array of 27 numbers

$$\varepsilon_{ijk} = \begin{cases} +1 & \text{if } (i,j,k) \text{ is an even permutation of } (1,2,3)\,, \\ -1 & \text{if it is an odd one}\,, \\ 0 & \text{if two of } i,j,k \text{ are equal}\,. \end{cases} \tag{5.56}$$

It changes sign under the exchange of any two of its indices and is therefore unchanged under a cyclic permutation of them. The *cross product* of $u, v \in \mathbb{R}^3$ is the element of $\mathbb{R}^3$ with components

$$(u \times v)_i = \varepsilon_{ijk} u_j v_k \,. \tag{5.57}$$

Rests on Definitions 4.44, 5.17 and 5.25.

Lemma 5.31 (Determinant through the Levi–Civita symbol)**.** *For every real* 3×3 *array* a *with entries* a_{ij}*,*

$$\varepsilon_{ijk}\, a_{il} a_{jm} a_{kn} = \varepsilon_{lmn} \det a = \varepsilon_{ijk}\, a_{li} a_{mj} a_{nk}\,, \tag{5.58}$$

and in particular $\det a^{\mathsf{T}} = \det a$*. Consequently, for a real array* R *with* $R^{\mathsf{T}} R = \mathbb{1}$ *— an orthogonal array in the sense of Equation* (5.161) *—*

$$R_{ia} R_{jb} R_{kc}\, \varepsilon_{abc} = (\det R)\, \varepsilon_{ijk}\,, \tag{5.59}$$

so that the symbol is unchanged when $\det R = 1$ *and merely changes sign when* $\det R = -1$*.*
Rests on Definition 5.30 and Equation (5.19)*.*

Proof. *Derives Lemma 5.31.*

Write $L_{lmn} = \varepsilon_{ijk} a_{il} a_{jm} a_{kn}$. Exchanging two of l, m, n exchanges two of the three factors; relabelling the summed indices to put the factors back in order exchanges two indices of ε_{ijk} and so costs a sign. Hence L is totally antisymmetric in (l, m, n): it vanishes whenever two of them are equal, and on the six remaining index triples it is $L_{lmn} = \varepsilon_{lmn} L_{123}$. Finally

$$L_{123} = \varepsilon_{ijk} a_{i1} a_{j2} a_{k3} = \sum_{\sigma} \operatorname{sgn}(\sigma)\, a_{\sigma(1)1} a_{\sigma(2)2} a_{\sigma(3)3} = \det a$$

by Equation (5.19), the surviving terms of the sum over i, j, k being exactly those in which (i, j, k) is a permutation of $(1, 2, 3)$, with ε_{ijk} its sign.

That $\det a^{\mathsf{T}} = \det a$ is Equation (5.19) reindexed by $\sigma \longmapsto \sigma^{-1}$, a bijection of the permutations which preserves the sign and turns $\prod_j a_{\sigma(j)j}$ into $\prod_j a_{j\sigma(j)}$. The second equality of Equation (5.58) is therefore the first one applied to a^{T}. Equation (5.59) is that second equality with $a = R$, after renaming the summed indices i, j, k to a, b, c and the free indices l, m, n to i, j, k; and $\det R = \pm 1$ because $(\det R)^2 = \det\left(R^{\mathsf{T}} R\right) = 1$. □

Lemma 5.32 (Contraction of two Levi–Civita symbols)**.** *For all* $j, k, l, m \in \{1, 2, 3\}$*,*

$$\sum_{i=1}^{3} \varepsilon_{ijk} \varepsilon_{ilm} = \delta_{jl} \delta_{km} - \delta_{jm} \delta_{kl}\,. \tag{5.60}$$

Rests on Definition 5.30.

Proof. *Derives Lemma 5.32.*

If $j = k$ both sides vanish: the left because $\varepsilon_{ijk} = 0$, the right because it becomes $\delta_{jl}\delta_{jm} - \delta_{jm}\delta_{jl}$. The same argument disposes of $l = m$. So let $j \neq k$ and $l \neq m$.

A term of the sum is non-zero only if $i \notin \{j, k\}$ and $i \notin \{l, m\}$. Since $\{j, k\}$ has two elements and i ranges over three values, exactly one i meets the first condition, so the sum has at most one non-zero term. Both $\{j, k\}$ and $\{l, m\}$ are two-element subsets of $\{1, 2, 3\}$; if they are different they share exactly one element, and the value missing from $\{j, k\}$ then lies in $\{l, m\}$, so the surviving term is killed by its second factor and the sum vanishes. The right-hand side vanishes too, since (l, m) is then neither (j, k) nor (k, j) and neither product of deltas can be 1.

There remain the two cases $\{l, m\} = \{j, k\}$. If $(l, m) = (j, k)$ the surviving term is $(\varepsilon_{ijk})^2 = 1$ and the right side is $\delta_{jj}\delta_{kk} - \delta_{jk}\delta_{kj} = 1 - 0 = 1$. If $(l, m) = (k, j)$ the term is $\varepsilon_{ijk}\varepsilon_{ikj} = -1$ and the right side is $\delta_{jk}\delta_{kj} - \delta_{jj}\delta_{kk} = -1$. □

Proposition 5.33 (The identities of the vector algebra of $\mathbb{R}^3$)**.** *For all* $u, v, w \in \mathbb{R}^3$*:*

1. Antisymmetry and orthogonality. $u \times v = -v \times u$*, and* $\langle u | u \times v \rangle = \langle v | u \times v \rangle = 0$*.*

2. Scalar triple product.

$$\langle u|v\times w\rangle = \varepsilon_{ijk}u_i v_j w_k = \det\begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix}, \tag{5.61}$$

which is therefore unchanged by a cyclic permutation of u, v, w and changes sign under the exchange of any two of them.

3. Expansion of a double cross product.

$$u\times(v\times w) = v\,\langle u|w\rangle - w\,\langle u|v\rangle\ . \tag{5.62}$$

4. Lagrange's identity.

$$\|u\times v\|^2 = \|u\|^2\,\|v\|^2 - \langle u|v\rangle^2\ , \tag{5.63}$$

so that $\|u\times v\| = \|u\|\,\|v\|\sin\theta$ with θ the angle Equation (5.18) between u and v; that is, the length of $u\times v$ is the area of the parallelogram they span.

5. Behaviour under an orthogonal change of basis. *If $R^{\mathsf{T}}R = \mathbb{1}$ then*

$$(Ru)\times(Rv) = (\det R)\,R\,(u\times v)\ . \tag{5.64}$$

Rests on Lemmas 5.31 and 5.32 and Equation (5.18).

Proof. *Derives Proposition 5.33.*

1. $(v\times u)_i = \varepsilon_{ijk}v_j u_k = \varepsilon_{ikj}v_k u_j = -\varepsilon_{ijk}u_j v_k$, the middle step being a relabelling of the two summed indices and the last the antisymmetry of Equation (5.56). For the orthogonality, $\langle u|u\times v\rangle = \varepsilon_{ijk}u_i u_j v_k$ vanishes because $u_i u_j$ is unchanged and ε_{ijk} changes sign when i and j are exchanged, so the sum equals minus itself; and $\langle v|u\times v\rangle$ vanishes by the same argument on j and k after using part 1's antisymmetry.

2. The first equality is Equation (5.57) paired with u. For the second, apply the second equality of Equation (5.58) at $(l,m,n) = (1,2,3)$ to the array a whose rows are u, v, w: it gives $\varepsilon_{ijk}a_{1i}a_{2j}a_{3k} = \varepsilon_{123}\det a = \det a$. The stated symmetries are those of ε_{ijk}.

3. Using Equation (5.57) twice, then the cyclic symmetry $\varepsilon_{ijk} = \varepsilon_{kij}$ and Equation (5.60),

$$\begin{aligned}[u\times(v\times w)]_i &= \varepsilon_{ijk}u_j\varepsilon_{klm}v_l w_m \\ &= (\varepsilon_{kij}\varepsilon_{klm})\,u_j v_l w_m \\ &= (\delta_{il}\delta_{jm} - \delta_{im}\delta_{jl})\,u_j v_l w_m \\ &= v_i\,u_j w_j - w_i\,u_j v_j\ ,\end{aligned}$$

which is Equation (5.62).

4. By Equation (5.57) and Equation (5.60),

$$\begin{aligned}\|u\times v\|^2 &= (\varepsilon_{ijk}u_j v_k)\,(\varepsilon_{ilm}u_l v_m) \\ &= (\delta_{jl}\delta_{km} - \delta_{jm}\delta_{kl})\,u_j v_k u_l v_m \\ &= u_j u_j\,v_k v_k - (u_j v_j)^2\ ,\end{aligned}$$

which is Equation (5.63). Substituting Equation (5.18) turns the right-hand side into $\|u\|^2\,\|v\|^2\,(1-\cos^2\theta)$, and $\sin\theta \geq 0$ on $[0,\pi]$; the parallelogram spanned by u and v has base $\|v\|$ and altitude $\|u\|\sin\theta$, the altitude computed in Section 5.1.5.

5. Multiply the first equality of Equation (5.58), taken with $a = R$, by R_{dl} and sum over l. On the left the two factors carrying l contract to $\sum_l R_{il}R_{dl} = \delta_{id}$, which is $R^{\mathsf{T}}R = \mathbb{1}$ read

for RR^{T} — the two are the same statement in finite dimension, by Proposition 5.47. There results

$$\varepsilon_{djk} R_{jm} R_{kn} = (\det R)\, R_{dl}\, \varepsilon_{lmn}\,.$$

Contracting this with $u_m v_n$ and using Equation (5.57) twice gives $[(Ru)\times(Rv)]_d = (\det R)\, R_{dl}\, (u\times v)_l$, which is Equation (5.64). □

Remark 5.34 (Why the cross product is not a vector). Equation (5.64) is the whole content of the word *pseudo-vector*. Under a rotation, $\det R = 1$, the cross product transforms exactly as a vector does; under a reflection, $\det R = -1$, it acquires an extra sign, so it is not a vector in the sense of a quantity whose components transform by R under *every* orthogonal change of basis. Equivalently, Equation (5.57) depends on which of the two orientations of $\mathbb{R}^3$ was fixed with the basis. Chapter 13 recovers the same object intrinsically, as the dual Equation (13.224) of the antisymmetric part of a tensor product, which is what makes its transformation law manifest; the identities above are its algebra, and they are established here once for all the physical parts that use them.

5.4 Linear transformations

5.4.1 Linear transformations

Definition 5.35 (Linear transformation)**.** Consider two vector spaces $\mathbb{V}$ and $\mathbb{W}$, together with a map $T : \mathbb{V} \longrightarrow \mathbb{W}$. We say that the map T is a *linear map, application, operator, or transformation* (LT) if and only if

1. $\forall\ u, v \in \mathbb{V},\ T(u+v) = T(u) + T(v)$;
2. $\forall\ v \in \mathbb{V},\ \forall\, a \in \mathbb{K},\ T(av) = aT(v)$.

Rests on Definition 4.31.

Clearly we can unify the above axioms into

1. $\forall\ u, v \in \mathbb{V},\ \forall\, a, b \in \mathbb{K},\ T(au+bv) = aT(u) + bT(v)$.

Let us denote by $\mathcal{L}(\mathbb{V},\mathbb{W})$ the set of all linear transformations from $\mathbb{V}$ to $\mathbb{W}$. Obviously

$$\mathcal{L}(\mathbb{V},\mathbb{W}) \subset \mathcal{F}(\mathbb{V},\mathbb{W})\,.$$

The fundamental theorem of linear algebra

The name is traditionally given to two statements about one linear transformation: that its domain splits into the part it annihilates and a part it maps faithfully onto its image — the rank–nullity relation — and that the four subspaces attached to T and to its adjoint fit together in two orthogonal pairs. Both are proved below. A preliminary is needed first, and it is one the chapter has so far used without stating: that Definition 5.15 is unambiguous, because all bases of a finite-dimensional space have the same number of elements.

Lemma 5.36 (Exchange and completion)**.** *Let $\mathbb{V} = \langle\{w_1,\dots,w_m\}\rangle$ and let $\{v_1,\dots,v_k\} \subset \mathbb{V}$ be linearly independent. Then $k \le m$, and $\{v_1,\dots,v_k\}$ can be completed to a basis of $\mathbb{V}$ by adjoining vectors taken from $\{w_1,\dots,w_m\}$. Consequently any two bases of a finite-dimensional vector space have the same number of elements, and Definition 5.15 is well posed.*

Rests on Definitions 5.11, 5.13 and 5.14.

Proof. *Derives Lemma 5.36.*

Exchange. We show by induction on $j = 0, 1, \ldots, k$ that, after a renumbering of the w's,

$$\langle \{v_1, \ldots, v_j, w_{j+1}, \ldots, w_m\} \rangle = \mathbb{V}\,, \tag{5.65}$$

and in particular that $j \leq m$. For $j = 0$ this is the hypothesis. Assume Equation (5.65) for $j-1$, with $j \leq k$. Then v_j is a finite linear combination

$$v_j = \sum_{i<j} a^i v_i + \sum_{l \geq j} b^l w_l\,.$$

Not every b^l can vanish: otherwise v_j would be a linear combination of $v_1, \ldots, v_{j-1}$, which Definition 5.13 forbids. In particular the second sum is not empty, so $j \leq m$. Renumber the remaining w's so that $b^j \neq 0$ and solve:

$$w_j = \frac{1}{b^j}\left(v_j - \sum_{i<j} a^i v_i - \sum_{l>j} b^l w_l\right)\,,$$

which places w_j in $\langle \{v_1, \ldots, v_j, w_{j+1}, \ldots, w_m\} \rangle$. That set therefore generates everything the set of the previous step generated, namely $\mathbb{V}$, which is Equation (5.65) at j. Taking $j = k$ gives $k \leq m$.

Equal cardinality. If B and B' are two bases, each is independent and each generates, so the exchange applied twice gives $|B| \leq |B'|$ and $|B'| \leq |B|$.

Completion. Take $j = k$ in Equation (5.65): the set $G = \{v_1, \ldots, v_k, w_{k+1}, \ldots, w_m\}$ generates $\mathbb{V}$. If it is not linearly independent, some vanishing linear combination of its elements has a non-zero coefficient; that coefficient cannot belong only to the v's, which are independent, so it may be taken to belong to some w_l. Solving for that w_l expresses it through the others, so deleting it leaves a set that still generates $\mathbb{V}$. Repeating terminates, G being finite, and leaves a generating, linearly independent set containing every v_i — a basis in the sense of Definition 5.14. □

Definition 5.37 (Kernel, image, nullity, rank)**.** Let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be a linear transformation. Its *kernel* and *image* are

$$\ker T = \{v \in \mathbb{V} \mid T(v) = 0_{\mathbb{W}}\}\,, \tag{5.66}$$

$$\operatorname{im} T = \{w \in \mathbb{W} \mid \exists\, v \in \mathbb{V},\ w = T(v)\}\,, \tag{5.67}$$

subspaces of $\mathbb{V}$ and of $\mathbb{W}$ respectively. The *nullity* of T is $\dim \ker T$ and its *rank* is $\dim \operatorname{im} T$.

Rests on Definitions 5.6 and 5.35.

Derivation. *Derives Definition 5.37.*

That the two sets are subspaces is the content asserted, and it is one line each. Both are non-empty, containing $0_{\mathbb{V}}$ and $0_{\mathbb{W}}$ respectively, since $T(0_{\mathbb{V}}) = T(0 \cdot 0_{\mathbb{V}}) = 0\,T(0_{\mathbb{V}}) = 0_{\mathbb{W}}$ by Equation (5.22). If $u, v \in \ker T$ and $a, b \in \mathbb{K}$ then $T(au + bv) = aT(u) + bT(v) = 0_{\mathbb{W}}$; and if $w = T(u)$, $z = T(v)$ lie in $\operatorname{im} T$ then $aw + bz = T(au + bv) \in \operatorname{im} T$. Both verify Definition 5.6. □

Theorem 5.38 (Rank–nullity)**.** *Let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be a linear transformation with $\dim \mathbb{V} = n < \infty$. Then*

$$\dim \mathbb{V} = \dim \ker T + \dim \operatorname{im} T\,. \tag{5.68}$$

Rests on Definitions 5.14 and 5.37 and Lemma 5.36.

Proof. *Derives Theorem 5.38.*

$\ker T$ is a subspace of a space of dimension n, so by Lemma 5.36 it has a finite basis $\{u_1, \dots, u_k\}$ with $k \leq n$, and that independent set completes to a basis $\{u_1, \dots, u_k, x_1, \dots, x_{n-k}\}$ of $\mathbb{V}$. We claim that $\{T(x_1), \dots, T(x_{n-k})\}$ is a basis of $\operatorname{im} T$.

It generates. Any element of $\operatorname{im} T$ is $T(v)$ for some $v = \sum_i a^i u_i + \sum_j b^j x_j$, and linearity together with $T(u_i) = 0_{\mathbb{W}}$ gives $T(v) = \sum_j b^j T(x_j)$.

It is independent. Suppose $\sum_j c^j T(x_j) = 0_{\mathbb{W}}$. Then $T\left(\sum_j c^j x_j\right) = 0_{\mathbb{W}}$, so $\sum_j c^j x_j \in \ker T$ and therefore $\sum_j c^j x_j = \sum_i d^i u_i$ for some scalars d^i. That is a vanishing linear combination of the completed basis, so every coefficient vanishes by Equation (5.29); in particular every $c^j = 0$.

Hence $\dim \operatorname{im} T = n - k = n - \dim \ker T$, which is Equation (5.68). □

The adjoint and the four fundamental subspaces

The second half of the theorem needs an inner product on each of the two spaces, and with it the transformation that mirrors T on the other side of the bracket.

Definition 5.39 (Adjoint)**.** Let $\mathbb{V}$ and $\mathbb{W}$ be finite-dimensional inner-product spaces over $\mathbb{K}$ and let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be linear. A map $T^\dagger : \mathbb{W} \longrightarrow \mathbb{V}$ is an *adjoint* of T if

$$\left\langle T^\dagger w \middle| v \right\rangle = \langle w | T v \rangle \,, \qquad \forall\, v \in \mathbb{V},\ \forall\, w \in \mathbb{W}\,, \tag{5.69}$$

the left-hand bracket being that of $\mathbb{V}$ and the right-hand one that of $\mathbb{W}$. An operator $T : \mathbb{V} \longrightarrow \mathbb{V}$ with $T^\dagger = T$ is called *self-adjoint*, or *hermitian*.

Rests on Definitions 5.17 and 5.35.

Proposition 5.40 (The adjoint exists, is unique, and is linear)**.** *Under the hypotheses of Definition 5.39 the adjoint exists, is unique, and is itself a linear transformation. In orthonormal bases of $\mathbb{V}$ and $\mathbb{W}$ its matrix is the conjugate transpose of the matrix of T,*

$$\left[T^\dagger\right] = [T]^\dagger \,, \qquad \left(T^\dagger\right)^i{}_j = \left(T^j{}_i\right)^* \,, \tag{5.70}$$

and $\left(T^\dagger\right)^\dagger = T$.

Rests on Definition 5.39, Proposition 5.26, and Equation (5.74).

Proof. *Derives Proposition 5.40.*

Uniqueness. If S and S' both satisfy Equation (5.69), then $\langle Sw - S'w | v \rangle = 0$ for every v by Equation (5.37); taking $v = Sw - S'w$ and using Equation (5.31) gives $Sw = S'w$ for every w.

Existence. By Proposition 5.26 choose orthonormal bases $\{e_i\}$ of $\mathbb{V}$ and $\{f_j\}$ of $\mathbb{W}$, so that Equation (5.49) holds in each. Define $T^\dagger$ on the basis of $\mathbb{W}$ by

$$T^\dagger f_j = \sum_i \left(T^j{}_i\right)^* e_i \,,$$

and extend by linearity, which is legitimate by the uniqueness of components used in Proposition 5.43. The map so defined is linear by construction, so both sides of Equation (5.69) are antilinear in w and linear in v, and it suffices to check the identity on basis vectors:

$$\begin{aligned}
\left\langle T^\dagger f_j \middle| e_k \right\rangle &= \sum_i \left(\left(T^j{}_i\right)^*\right)^* \langle e_i | e_k \rangle && \text{by Equation (5.36)} \\
&= T^j{}_k \,, \\
\langle f_j | T e_k \rangle &= \sum_l T^l{}_k \, \langle f_j | f_l \rangle && \text{by Equations (5.34) and (5.74)} \\
&= T^j{}_k \,,
\end{aligned}$$

so the two agree. This also establishes Equation (5.70).

Linearity is forced, not assumed. Whatever map satisfies Equation (5.69) is linear, so nothing was lost by building a linear one: for $a, b \in \mathbb{K}$ and $w, z \in \mathbb{W}$,

$$\begin{aligned}\left\langle T^\dagger\,(aw+bz)\middle|v\right\rangle &= \langle aw+bz|Tv\rangle \\ &= a^*\,\langle w|Tv\rangle + b^*\,\langle z|Tv\rangle \qquad \text{by Equations (5.36) and (5.37)} \\ &= \left\langle aT^\dagger w + bT^\dagger z\middle|v\right\rangle\,,\end{aligned}$$

for every v, and the uniqueness argument above forces the two arguments to coincide.

Involutivity. For all v, w, $\left\langle \left(T^\dagger\right)^\dagger v\middle|w\right\rangle = \left\langle v\middle|T^\dagger w\right\rangle = \left\langle T^\dagger w\middle|v\right\rangle^* = \langle w|Tv\rangle^* = \langle Tv|w\rangle$, using Equation (5.33) twice, so $\left(T^\dagger\right)^\dagger = T$. □

Theorem 5.41 (The four fundamental subspaces)**.** *Let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be a linear transformation between finite-dimensional inner-product spaces. Then*

$$(\operatorname{im} T)^\perp = \ker T^\dagger\,, \tag{5.71}$$

$$(\ker T)^\perp = \operatorname{im} T^\dagger\,, \tag{5.72}$$

so that each space splits orthogonally,

$$\mathbb{W} = \operatorname{im} T \oplus \ker T^\dagger\,, \qquad \mathbb{V} = \ker T \oplus \operatorname{im} T^\dagger\,, \tag{5.73}$$

and T and $T^\dagger$ have the same rank.

Rests on Definition 5.39, Theorem 5.38, and Corollary 5.27.

Proof. *Derives Theorem 5.41.*

A preliminary: $\left(\mathbb{S}^\perp\right)^\perp = \mathbb{S}$ for a subspace $\mathbb{S}$ of a finite-dimensional inner-product space. The inclusion $\mathbb{S} \subset \left(\mathbb{S}^\perp\right)^\perp$ is immediate from Equation (5.52) and the hermiticity Equation (5.33). Conversely let $v \in \left(\mathbb{S}^\perp\right)^\perp$ and split $v = s + w$ with $s \in \mathbb{S}$, $w \in \mathbb{S}^\perp$, which Corollary 5.27 permits. Then $w = v - s \in \left(\mathbb{S}^\perp\right)^\perp$, being a difference of two of its elements, while also $w \in \mathbb{S}^\perp$; hence $\langle w|w\rangle = 0$ and $w = 0_\mathbb{V}$ by Equation (5.31), so $v = s \in \mathbb{S}$.

Equation (5.71). By Equation (5.52), $w \in (\operatorname{im} T)^\perp$ means $\langle Tv|w\rangle = 0$ for every $v \in \mathbb{V}$. Now

$$\langle Tv|w\rangle = \langle w|Tv\rangle^* = \left\langle T^\dagger w\middle|v\right\rangle^* = \left\langle v\middle|T^\dagger w\right\rangle\,,$$

using Equation (5.33) and Equation (5.69), so the condition reads $\left\langle v\middle|T^\dagger w\right\rangle = 0$ for every v; taking $v = T^\dagger w$ and using Equation (5.31) this holds exactly when $T^\dagger w = 0_\mathbb{V}$, that is, when $w \in \ker T^\dagger$.

Equation (5.72). Apply Equation (5.71) to $T^\dagger$ in place of T, which is legitimate because $\left(T^\dagger\right)^\dagger = T$ by Proposition 5.40: it gives $\left(\operatorname{im} T^\dagger\right)^\perp = \ker T$. Taking the orthogonal complement of both sides and using the preliminary yields $\operatorname{im} T^\dagger = (\ker T)^\perp$.

The splittings and the ranks. Equation (5.73) is Corollary 5.27 applied to $\operatorname{im} T \leq \mathbb{W}$ and to $\ker T \leq \mathbb{V}$, rewritten through Equations (5.71) and (5.72). Finally, the second splitting gives $\dim \mathbb{V} = \dim \ker T + \dim \operatorname{im} T^\dagger$, whereas Equation (5.68) gives $\dim \mathbb{V} = \dim \ker T + \dim \operatorname{im} T$; subtracting, $\dim \operatorname{im} T = \dim \operatorname{im} T^\dagger$. □

Remark 5.42 (Two daggers, one star, and what each of them needs). Three things are worth separating, because the notation invites confusing them.

- The *adjoint* $T^\dagger : \mathbb{W} \longrightarrow \mathbb{V}$ of Definition 5.39 needs an inner product on each space, and changes if either is changed. The *pullback* $T^* : \mathbb{W}^* \longrightarrow \mathbb{V}^*$ of Section 5.5.2 needs no structure at all, and acts between the dual spaces rather than between the spaces themselves. They are different maps between different pairs of spaces, and the correspondence between them is exactly the antilinear identification $\mathbb{V} \simeq \mathbb{V}^*$ that an inner product supplies through Proposition 5.59.

- Over $\mathbb{R}$ the conjugation in Equation (5.70) is the identity, so in an orthonormal basis $\left[T^\dagger\right] = [T]^\mathsf{T}$ and self-adjoint means symmetric. Over $\mathbb{C}$ it means hermitian, in the sense of Definition 5.111. The requirement that the basis be *orthonormal* cannot be dropped: in a general basis the matrix of the adjoint involves the matrix of the inner product as well.

- Equation (5.73) is the whole of the solvability theory of a linear system in one line: $Tv = w$ has a solution if and only if $w \perp \ker T^\dagger$, and the solution is unique up to the addition of an arbitrary element of $\ker T$. The equality of ranks that closes the proof is the classical statement that the row rank and the column rank of a matrix agree.

Matrix representation associated with a linear transformation

Let $\mathbb{V}$ and $\mathbb{W}$ be vector spaces, generated by the bases $B_1 = \{e_i\}_{i=1}^{\dim \mathbb{V}}$ and $B_2 = \{f_i\}_{i=1}^{\dim \mathbb{W}}$, and consider a linear transformation $T : \mathbb{V} \longrightarrow \mathbb{W}$. The image $T(e_i)$ of the i-th vector of the basis of $\mathbb{V}$ is a new vector belonging to $\mathbb{W}$, so we can express it as a linear combination of the elements of the basis of $\mathbb{W}$ — the expansion exists because $\langle B_2 \rangle = \mathbb{W}$, which is the first clause of Definition 5.14, and its coefficients are unique because B_2 is linearly independent, which is the second clause and is the step written out in Proposition 5.43 below. That is,

$$T(e_i) = \sum_{j=1}^{\dim \mathbb{W}} T^j{}_i f_j \,. \tag{5.74}$$

We denote by $[T]$ the matrix with elements $T^j{}_i$, with $i = 1, \ldots, \dim \mathbb{V}$ and $j = 1, \ldots, \dim \mathbb{W}$. We say that the matrix $[T]$ is the *matrix representation* associated with the linear transformation T. Note that the coefficients of $[T]$ are basis-dependent: they depend on the choice of the bases of $\mathbb{V}$ and of $\mathbb{W}$. For this reason, when it is necessary to specify the bases, we denote the matrix representation of $T : \langle B_1 \rangle \longrightarrow \langle B_2 \rangle$ by $[T]_{B_1}^{B_2}$.

Proposition 5.43. *Given an LT and the respective bases, its matrix representation is unique.*

Rests on Definitions 5.14 and 5.35 and Equations (5.29) and (5.74).

Derivation. *Derives Proposition 5.43.*

The claim has two halves, and it is worth separating them: the array $[T]$ is *well defined* (its entries do not depend on how $T(e_i)$ happens to be written), and it *determines* T completely.

The entries are unique. Suppose the images of the basis vectors admitted two expansions,

$$T(e_i) = \sum_{j=1}^{\dim \mathbb{W}} T^j{}_i f_j = \sum_{j=1}^{\dim \mathbb{W}} S^j{}_i f_j \,.$$

Subtracting, and collecting terms by h),

$$\sum_{j=1}^{\dim \mathbb{W}} \left(T^j{}_i - S^j{}_i\right) f_j = 0_{\mathbb{W}} \,.$$

But $B_2 = \{f_j\}$ is a basis and therefore linearly independent, so by Equation (5.29) every coefficient of such a vanishing combination is zero:

$$T^j{}_i = S^j{}_i\,, \qquad \forall\, i = 1, \ldots, \dim \mathbb{V}\,,\ \forall\, j = 1, \ldots, \dim \mathbb{W}\,.$$

The array determines the transformation. Conversely, let v be any vector of $\mathbb{V}$, with components v^i in B_1 (Definition 5.16). Linearity gives

$$\begin{aligned} T(v) &= T\left(\sum_{i=1}^{\dim \mathbb{V}} v^i e_i\right) \\ &= \sum_{i=1}^{\dim \mathbb{V}} v^i T(e_i) \\ &= \sum_{i=1}^{\dim \mathbb{V}} \sum_{j=1}^{\dim \mathbb{W}} T^j{}_i v^i f_j \qquad\qquad \text{by Equation (5.74)}\,, \end{aligned}$$

so the components of $T(v)$ in B_2 are $(T(v))^j = \sum_i T^j{}_i v^i$, the familiar matrix-times-column product. Two linear transformations with the same matrix therefore agree on every vector, and are the same transformation.

The two halves together say that, once the bases are fixed, $T \mapsto [T]$ is a bijection between $\mathcal{L}(\mathbb{V}, \mathbb{W})$ and the $\dim \mathbb{W} \times \dim \mathbb{V}$ arrays over $\mathbb{K}$. What is *not* claimed is independence of the bases: change B_1 or B_2 and the same T acquires a different matrix, which is the remark made after Equation (5.74) and, later, the reason a representation is characterized by its character (Equation (5.181)) rather than by its matrices. □

Definition 5.44 (Functional)**.** A *functional* is a linear transformation of the form $T : \mathbb{V} \longrightarrow \mathbb{K}$.

Rests on Definition 5.35.

5.4.2 The vector space of linear transformations

Null operator

Consider the map

$$\begin{array}{rccc} \Theta: & \mathbb{V} & \longrightarrow & \mathbb{W} \\ & v & \longmapsto & \Theta(v) = 0_{\mathbb{W}}\,. \end{array}$$

Let us see that Θ is a linear operator. We have

$$\begin{aligned} \Theta(au + bv) &= 0_{\mathbb{W}} \\ &= 0_{\mathbb{W}} + 0_{\mathbb{W}} \\ &= a0_{\mathbb{W}} + b0_{\mathbb{W}} \qquad\qquad \text{by Equation (5.23)} \\ &= a\Theta(u) + b\Theta(v)\,, \end{aligned}$$

so Θ is a linear operator, which we shall call the *null operator*. Let us see, further, the following. Let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be a linear transformation. We have

$$\begin{aligned} (T + \Theta)(v) &= T(v) + \Theta(v) \\ &= T(v) + 0_{\mathbb{W}} \\ &= T(v)\,, \end{aligned}$$

hence $T + \Theta = T$, and analogously $\Theta + T = T$. Thus Θ satisfies the neutral-element axiom:

$$\forall\, T \in \mathcal{L}(\mathbb{V}, \mathbb{W}),\ \exists\, \Theta \in \mathcal{L}(\mathbb{V}, \mathbb{W}) \mid T + \Theta = T\ \wedge\ \Theta + T = T\,. \tag{5.75}$$

Opposite operator

Let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be an LT. Define $-T$ as the map

$$\begin{array}{rccc} -T: & \mathbb{V} & \longrightarrow & \mathbb{W} \\ & v & \longmapsto & (-T)(v) = -T(v)\,. \end{array}$$

Let us see that $-T$ is a linear operator. We have

$$\begin{aligned} (-T)(au+bv) &= -T(au+bv) \\ &= -(aT(u)+bT(v)) \\ &= -aT(u) - bT(v) \\ &= a(-T)(u) + b(-T)(v)\,, \end{aligned}$$

so $-T$ is indeed a linear operator, which we shall call the *opposite operator* of T. Note, moreover, that

$$\begin{aligned} (T+(-T))(v) &= T(v) + (-T)(v) \\ &= T(v) - T(v) \\ &= 0_{\mathbb{W}} \\ &= \Theta(v)\,, \end{aligned}$$

so $T+(-T) = \Theta$, and analogously $-T+T = \Theta$. Thus the inverse-element axiom is satisfied:

$$\forall\ T \in \mathcal{L}(\mathbb{V},\mathbb{W}),\ \exists\ (-T) \in \mathcal{L}(\mathbb{V},\mathbb{W}) \mid T+(-T) = \Theta\ \wedge\ -T+T = \Theta\,. \tag{5.76}$$

Linear combination of linear transformations

Consider the linear transformations $T_1 : \mathbb{V} \longrightarrow \mathbb{W}$ and $T_2 : \mathbb{V} \longrightarrow \mathbb{W}$. Let us see that a linear combination of these LT's is also an LT. Indeed,

$$\begin{aligned} (aT_1+bT_2)(cu+dv) &= aT_1(cu+dv) + bT_2(cu+dv) \\ &= a(T_1(cu)+T_1(dv)) + b(T_2(cu)+T_2(dv)) \\ &= c(aT_1(u)+bT_2(u)) + d(aT_1(v)+bT_2(v)) \\ &= c(aT_1+bT_2)(u) + d(aT_1+bT_2)(v)\,. \end{aligned}$$

The vector space of linear transformations

As we have seen, the set $\mathcal{L}(\mathbb{V},\mathbb{W})$ satisfies the neutral-element and inverse-element axioms, and moreover there is closure: the linear combination of two LT's is an element of $\mathcal{L}(\mathbb{V},\mathbb{W})$. It is then evident that the axioms of a vector space are satisfied; that is, $[(\mathbb{K},+,\cdot),\cdot,(\mathcal{L}(\mathbb{V},\mathbb{W}),+)]$ is a vector space.

Inverse operator

The opposite operator of the preceding subsubsection undoes the *addition* of T; the operator defined here undoes its *action*. The two are unrelated, and the source's placement of them side by side is the only reason for saying so.

Definition 5.45 (Inverse of a linear transformation)**.** Let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be a linear transformation. A map $S : \mathbb{W} \longrightarrow \mathbb{V}$ is an *inverse* of T if

$$S \circ T = \mathrm{id}_{\mathbb{V}}\,, \qquad T \circ S = \mathrm{id}_{\mathbb{W}}\,. \tag{5.77}$$

When such an S exists we call T *invertible* and write $S = T^{-1}$; the notation is justified by the uniqueness proved in Proposition 5.46.

Rests on Definition 5.35.

Proposition 5.46 (Existence, uniqueness, and linearity of the inverse)**.** *Let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be a linear transformation.*

1. *T is invertible if and only if it is a bijection.*
2. *The inverse, when it exists, is unique.*
3. *The inverse is automatically a linear transformation; hence an invertible linear transformation is precisely an isomorphism in the sense of Definition 5.51.*

Rests on Definitions 5.35 and 5.45.

Proof. *Derives Proposition 5.46.*

1. If S satisfies Equation (5.77) then T is injective, since $T(u) = T(v)$ gives $u = S(T(u)) = S(T(v)) = v$, and surjective, since $w = T(S(w))$ for every $w \in \mathbb{W}$. Conversely, a bijection has a set-theoretic inverse, and that inverse satisfies Equation (5.77) by construction.

2. If S and S' are both inverses then

$$S = S \circ \mathrm{id}_{\mathbb{W}} = S \circ (T \circ S') = (S \circ T) \circ S' = \mathrm{id}_{\mathbb{V}} \circ S' = S' ,$$

composition of maps being associative.

3. Let $w, z \in \mathbb{W}$ and $a, b \in \mathbb{K}$, and put $v = T^{-1}(w)$, $u = T^{-1}(z)$. Linearity of T gives $T(av + bu) = aT(v) + bT(u) = aw + bz$, and applying T^{-1} to both sides,

$$T^{-1}(aw + bz) = av + bu = aT^{-1}(w) + bT^{-1}(z) .$$

So T^{-1} satisfies Definition 5.35. With part 1 this identifies the invertible linear transformations with the isomorphisms. □

In finite dimension the three conditions of Proposition 5.46 collapse into one, and this is one of the very few places where a statement true of finite-dimensional spaces is flatly false in general.

Proposition 5.47 (Injective, surjective, invertible)**.** *Let $\mathbb{V}$ and $\mathbb{W}$ be vector spaces over the same field with $\dim \mathbb{V} = \dim \mathbb{W} = n < \infty$, and let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be linear. The following are equivalent: T is injective; T is surjective; T is invertible. In particular an endomorphism of a finite-dimensional space is invertible as soon as it is injective.*

Rests on Theorem 5.38, Proposition 5.46, and Definition 5.37.

Proof. *Derives Proposition 5.47.*

T is injective if and only if $\ker T = \{0_{\mathbb{V}}\}$: one direction is immediate, and for the other $T(u) = T(v)$ gives $T(u - v) = 0_{\mathbb{W}}$, hence $u = v$. By Equation (5.68) this is equivalent to $\dim \operatorname{im} T = n$, and a subspace of $\mathbb{W}$ of dimension $n = \dim \mathbb{W}$ is the whole of $\mathbb{W}$ — a basis of it is an independent set of n vectors in $\mathbb{W}$, which Lemma 5.36 completes to a basis of $\mathbb{W}$ by adjoining $n - n = 0$ further vectors. So injectivity is equivalent to surjectivity, and each is therefore equivalent to bijectivity, which is invertibility by Proposition 5.46. □

Remark 5.48 (The hypothesis of finite dimension is doing all the work). On $\mathcal{F}(\mathbb{N}, \mathbb{R})$ of Example 5.9 — the sequences of real numbers — the shift $T(a^1, a^2, \ldots) = (0, a^1, a^2, \ldots)$ is linear and injective and is not surjective, while the shift in the other direction is linear and surjective and is not injective. Both are one-sided inverses of one another and neither is an inverse in the sense of Equation (5.77). This is the reason the operator theory of Chapter 12 must distinguish the spectrum of an operator from its set of eigenvalues, a distinction that simply does not arise here.

Matrix of the inverse

Proposition 5.49. *Let $\mathbb{V}, \mathbb{W}$ be of the same finite dimension n with fixed bases B_1, B_2, and let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be invertible. Then $[T]$ is an invertible array and*

$$\left[T^{-1}\right]_{B_2}^{B_1} = \left([T]_{B_1}^{B_2}\right)^{-1} . \tag{5.78}$$

Moreover, for invertible arrays A and B of order n the product is invertible and

$$(AB)^{-1} = B^{-1}A^{-1} . \tag{5.79}$$

Rests on Propositions 5.43 and 5.46 and Equation (5.74).

Proof. Derives Proposition 5.49.

First, the matrix of a composition is the product of the matrices. Write $B_1 = \{e_i\}$, $B_2 = \{f_j\}$, and let $S : \mathbb{W} \longrightarrow \mathbb{V}$ be any linear transformation, so that Equation (5.74) reads $T(e_i) = \sum_j T^j{}_i f_j$ and $S(f_j) = \sum_k S^k{}_j e_k$. Then

$$\begin{aligned} (S \circ T)(e_i) &= S\left(\sum_j T^j{}_i f_j\right) \\ &= \sum_j T^j{}_i \, S(f_j) && \text{by linearity of } S \\ &= \sum_k \left(\sum_j S^k{}_j T^j{}_i\right) e_k , \end{aligned}$$

so that, by the uniqueness of Proposition 5.43, $[S \circ T]_{B_1}^{B_1} = [S]_{B_2}^{B_1} [T]_{B_1}^{B_2}$; and the matrix of $\mathrm{id}_{\mathbb{V}}$ in the basis B_1 twice over is $\mathbb{1}$, since $\mathrm{id}_{\mathbb{V}}(e_i) = e_i$. Applying this to both halves of Equation (5.77),

$$\left[T^{-1}\right][T] = \mathbb{1} , \qquad [T]\left[T^{-1}\right] = \mathbb{1} ,$$

which is Equation (5.78). For Equation (5.79), multiply out:

$$\begin{aligned} (AB)\left(B^{-1}A^{-1}\right) &= A\left(BB^{-1}\right)A^{-1} = AA^{-1} = \mathbb{1} , \\ \left(B^{-1}A^{-1}\right)(AB) &= B^{-1}\left(A^{-1}A\right)B = B^{-1}B = \mathbb{1} , \end{aligned}$$

so $B^{-1}A^{-1}$ is *an* inverse of AB, and the uniqueness argument of Proposition 5.46 — which used nothing but associativity — makes it *the* inverse. The order reverses because the factor applied last must be undone first; the same reversal appears for the transpose in Equation (5.161) and for the pullback in Equation (5.98). □

The determinant criterion that decides invertibility in practice, $\det[T] \neq 0$, is recorded with the general linear group in Equation (5.81).

5.4.3 Morphisms of vector spaces

Remark 5.50 (Homomorphism). An LT is by definition a group homomorphism: from the first axiom of an LT we see that T establishes a homomorphism between the groups $(\mathbb{V}, +)$ and $(\mathbb{W}, +)$.

Definition 5.51 (Isomorphism)**.** We say that a linear transformation is an *isomorphism* if it is in addition a bijection. If an isomorphism exists between two vector spaces, we say that they are *isomorphic*.

Rests on Definition 5.35.

Proposition 5.52. *Two vector spaces* over the same field $\mathbb{K}$ *and of the same finite dimension* n *are isomorphic.*

Rests on Definitions 5.14, 5.15 and 5.51 and Proposition 5.43.

Proof. *Derives Proposition 5.52.*

Let $\mathbb{V}$ and $\mathbb{W}$ be the two spaces, with bases $B_1 = \{e_i\}_{i=1}^n$ and $B_2 = \{f_i\}_{i=1}^n$; these have the same cardinality n by hypothesis, which is what allows the two to be listed with a common index. Define $T : \mathbb{V} \longrightarrow \mathbb{W}$ on the basis by

$$T(e_i) = f_i\,, \qquad i = 1, \ldots, n\,,$$

and extend to all of $\mathbb{V}$ by

$$T\left(\sum_{i=1}^{n} v^i e_i\right) = \sum_{i=1}^{n} v^i f_i\,. \tag{5.80}$$

Equation (5.80) is a legitimate definition precisely because the components v^i of a vector in a basis are unique — the argument just given for Proposition 5.43 — so no vector is assigned two images.

T is linear. If $u = \sum_i u^i e_i$ and $v = \sum_i v^i e_i$ then $au + bv = \sum_i (au^i + bv^i) e_i$, so

$$\begin{aligned} T(au + bv) &= \sum_{i=1}^{n} \left(au^i + bv^i\right) f_i \\ &= a \sum_{i=1}^{n} u^i f_i + b \sum_{i=1}^{n} v^i f_i \\ &= aT(u) + bT(v)\,. \end{aligned}$$

T is injective. Suppose $T(v) = 0_{\mathbb{W}}$. Then $\sum_i v^i f_i = 0_{\mathbb{W}}$, and B_2 being linearly independent, Equation (5.29) forces $v^i = 0$ for every i, whence $v = 0_{\mathbb{V}}$. Since T is linear, $T(u) = T(v)$ implies $T(u - v) = 0_{\mathbb{W}}$ and hence $u = v$.

T is surjective. Any $w \in \mathbb{W}$ can be written $w = \sum_i w^i f_i$, because B_2 generates $\mathbb{W}$; then $w = T\left(\sum_i w^i e_i\right)$.

Thus T is a bijective linear transformation, that is, an isomorphism in the sense of Definition 5.51. □

Remark 5.53 (The hypotheses are not decoration). Two of them are doing work, and the source states neither.

- *The same field.* Dimension is meaningless without saying over what. The space $\mathbb{C}$ regarded as a complex vector space has dimension 1; regarded as a real vector space it has dimension 2 and is isomorphic to $\mathbb{R}^2$, not to $\mathbb{R}$. A complex space of dimension n is a real space of dimension $2n$.

- *Finiteness.* For infinite-dimensional spaces the statement survives in the form "two spaces whose bases have the same cardinality are isomorphic", with the same proof read as a bijection $B_1 \longrightarrow B_2$ rather than as a common index; but the existence of a basis at all is then no longer constructive — that every vector space has a basis is equivalent to the axiom of choice (Chapter 3) — and the resulting isomorphism cannot be written down. In every physical application in this book the relevant spaces are either finite-dimensional or Hilbert spaces (Definition 12.2), where the topological notion of basis — a countable orthonormal set whose closed span is the whole space, the notion reserved for Section 12.2 — replaces this one and behaves far better.

The converse is immediate and worth stating: an isomorphism carries a basis to a basis (it preserves linear independence, being injective, and generation, being surjective), so isomorphic spaces have equal dimension. Dimension is therefore a *complete* invariant of a vector space over a fixed field — which is why so little of physics is contained in the vector space itself, and so much of it in the extra structure (inner product, algebra, group action) laid on top.

Definition 5.54 (Endomorphism)**.** An *endomorphism* is a linear transformation of the form $T : \mathbb{V} \longrightarrow \mathbb{V}$.

Rests on Definition 5.35.

Definition 5.55 (Automorphism)**.** An endomorphism is an *automorphism* if it is in addition an isomorphism.

Rests on Definitions 5.51 and 5.54.

Now consider the vector space $\mathbb{V}$. The set of all automorphisms of the form $T : \mathbb{V} \longrightarrow \mathbb{V}$, together with the composition of linear transformations, forms a group. Indeed, closure holds because automorphisms are bijections and, as we know, the composition of bijections is a bijection; associativity, the existence of the neutral element, and the existence of inverse elements are easy to see, and are likewise consequences of our considering bijections. We call this group the *general linear group*, and symbolize it as

$$\left(\{T \mid T : \mathbb{V} \longrightarrow \mathbb{V} \text{ is a bijection}\}, \circ\right) \overset{\text{not.}}{=} \mathrm{GL}(\mathbb{V}) .$$

In the case where the vector space considered is of the form $\mathbb{V} = \mathbb{K}^n$, with $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, we denote the general linear group by $\mathrm{GL}(n, \mathbb{K})$.

As we have seen, we can treat linear transformations as matrices. Let $\mathcal{M}_{n\times n}[\mathbb{K}]$ denote the set of *all* square arrays of order n over $\mathbb{K}$. That is the meaning the symbol carries throughout this chapter, and in particular in Proposition 5.121, where $T \longmapsto [T]$ is a bijection of the *whole* of $\mathcal{L}(\mathbb{V})$ onto it. The general linear group is then not that set, but its subset of invertible elements:

$$\mathrm{GL}(n, \mathbb{K}) \simeq \{M \in \mathcal{M}_{n\times n}[\mathbb{K}] \mid \det M \neq 0\} , \tag{5.81}$$

where the isomorphism is the bijection associating with each linear transformation T its matrix representation $[T]$ (for we have seen that, given an LT and a basis, its matrix representation is unique). The restriction on the right is not cosmetic. The full set of arrays is not a group under multiplication at all — the zero array has no inverse — so an isomorphism of groups could not have it on the right-hand side; only the invertible arrays, which are exactly the images of the automorphisms, form one. Recall, moreover, that

- $T \circ T^{-1} = \mathrm{id}_{\mathrm{GL}(n,\mathbb{K})}$;
- $[T][T]^{-1} = \mathbb{1}$.

Note that this isomorphism is not canonical, in the following sense: the matrix representation of a linear transformation is not unique, being basis-dependent.

There exists a very important subgroup of the general linear group. If we consider the set of all automorphisms of the form $T : \mathbb{K}^n \longrightarrow \mathbb{K}^n$ with $\det[T] = 1$, there is closure under this condition. Indeed, let A and B be two automorphisms belonging to this set, so $\det[A] = 1$ and $\det[B] = 1$. By closure of $\mathrm{GL}(n, \mathbb{K})$ there must exist C such that $A \circ B = C$, where we know that $[A][B] = [C]$. Thus, taking determinants,

$$\begin{aligned} [A][B] &= [C] \\ \det\left([A][B]\right) &= \det[C] \\ \det[A]\det[B] &= \det[C] \\ 1 &= \det[C] , \end{aligned}$$

so $\det[C] = 1$. That is, there is a smaller group satisfying the properties of the general linear group, for which $\det[T] = 1$. We call this subgroup the *special linear group* and symbolize it $\mathrm{SL}(n, \mathbb{K})$; thus

$$\mathrm{SL}(n, \mathbb{K}) = \{T \in \mathrm{GL}(n, \mathbb{K}) \mid \det[T] = 1\} \,. \tag{5.82}$$

5.5 The dual space

5.5.1 The dual space

Definition 5.56 (Dual space)**.** Consider a vector space $\mathbb{V}(\mathbb{K})$. We define the *dual space* of $\mathbb{V}$ as the vector space

$$\mathcal{L}(\mathbb{V}, \mathbb{K}) \overset{\text{not.}}{=} \mathbb{V}^* \,, \tag{5.83}$$

that is, as the set of all functionals of the form $f : \mathbb{V} \longrightarrow \mathbb{K}$.

Rests on Definition 5.44.

Vectors as matrices

Consider the vector space $\mathbb{V} = \langle \{e_i\}_{i=1} \rangle$. Define the functional g by

$$g(e_i) = g_i \,, \tag{5.84}$$

where $\{g_i\}_{i=1}$ is a set of scalars. Then, for an arbitrary vector with components v^i in the mentioned basis,

$$\begin{aligned} g(v) &= g\left(\sum_{i=1}^{\dim \mathbb{V}} v^i e_i\right) \\ &= \sum_{i=1}^{\dim \mathbb{V}} v^i g(e_i) \\ &= \sum_{i=1}^{\dim \mathbb{V}} v^i g_i && \text{by Equation (5.84)} \\ &= (g_1 \ \ldots)_{1\times \dim \mathbb{V}} \begin{pmatrix} v^1 \\ \vdots \end{pmatrix}_{\dim \mathbb{V}\times 1} \,. \end{aligned} \tag{5.85}$$

The preceding result suggests that we can represent the vector $v \in \mathbb{V}$ and the vector $g \in \mathbb{V}^*$ as a column matrix and a row matrix respectively; that is,

$$v \sim \begin{pmatrix} v^1 \\ \vdots \end{pmatrix}_{\dim \mathbb{V}\times 1} \,, \quad g \sim (g_1 \ \ldots)_{1\times \dim \mathbb{V}} \,.$$

Basis of the dual space

In the preceding context, consider the functionals

$$\begin{aligned} f^1 &\sim (1\ 0\ 0\ \ldots\ 0\ \ldots) \,, \\ f^2 &\sim (0\ 1\ 0\ \ldots\ 0\ \ldots) \,, \\ &\vdots \end{aligned}$$

From Equation (5.84) and Equation (5.85) we see that

$$f^i(e_j) = \delta^i{}_j \,. \tag{5.86}$$

Now, for an arbitrary functional $g \in \mathbb{V}^*$ we have

$$\begin{aligned}
g(v) &= g\left(\sum_{i=1}^{\dim \mathbb{V}} v^i e_i\right) \\
&= \sum_{i=1}^{\dim \mathbb{V}} v^i g(e_i) \\
&= \sum_{i=1}^{\dim \mathbb{V}} v^i g_i && \text{by Equation (5.84)} \\
&= \sum_{i,j=1}^{\dim \mathbb{V}} v^i \delta^j{}_i g_j \\
&= \sum_{i,j=1}^{\dim \mathbb{V}} v^i f^j(e_i) g_j && \text{by Equation (5.86)} \\
&= \sum_{j=1}^{\dim \mathbb{V}} f^j \left(\sum_i v^i e_i\right) g_j \\
&= \sum_{j=1}^{\dim \mathbb{V}} f^j(v) g_j \\
&= \left(\sum_{j=1}^{\dim \mathbb{V}} g_j f^j\right)(v)\,,
\end{aligned}$$

hence

$$g = \sum_{j=1}^{\dim \mathbb{V}} g_j f^j\,. \tag{5.87}$$

This result tells us that any functional in $\mathbb{V}^*$ can be expressed as a linear combination of the functionals f^j. Hence $\{f^j\}_{j=1}^{\dim \mathbb{V}}$ is a basis for the space $\mathbb{V}^*$. As we see, the dimension of the dual space $\mathbb{V}^*$ is the same as that of $\mathbb{V}$. We thus conclude the important result: $\mathbb{V}$ *and* $\mathbb{V}^*$ *are isomorphic.*

That a vector space and its dual are isomorphic is a very important result: it means that a space and its dual are in reality the same structure. Given an isomorphism $\Phi : \mathbb{V} \longrightarrow \mathbb{V}^*$, we say that $\Phi(v)$ is the *dual element* of v, and vice versa.

The annihilator of a subspace

Proposition 5.57 (Annihilator of a subspace)**.** *Let* $\mathbb{W}$ *be a subspace of a vector space* $\mathbb{V}$ *with* $\dim \mathbb{V} = n < \infty$*, and let*

$$\mathbb{W}^0 = \{g \in \mathbb{V}^* \mid g(w) = 0 \quad \forall\, w \in \mathbb{W}\} \tag{5.88}$$

be its annihilator*. Then* $\mathbb{W}^0$ *is a subspace of* $\mathbb{V}^*$ *and*

$$\dim \mathbb{W}^0 = n - \dim \mathbb{W}\,. \tag{5.89}$$

Rests on Definitions 5.6 and 5.56 and Lemma 5.36.

Proof. *Derives Proposition 5.57.*

$\mathbb{W}^0$ is defined by conditions linear in g, so it is a subspace. Let $\{e_1, \ldots, e_m\}$ be a basis of $\mathbb{W}$, $m = \dim \mathbb{W}$, completed to a basis $\{e_1, \ldots, e_n\}$ of $\mathbb{V}$ by Lemma 5.36, and let $\{f^1, \ldots, f^n\}$ be the dual basis Equation (5.86). By Equation (5.87) every $g \in \mathbb{V}^*$ is $g = \sum_j g_j f^j$ with $g_j = g(e_j)$. Since $\mathbb{W}$ is generated by $e_1, \ldots, e_m$, the condition $g(w) = 0$ for all $w \in \mathbb{W}$ is

equivalent to $g_1 = \ldots = g_m = 0$, that is, to $g \in \langle\{f^{m+1}, \ldots, f^n\}\rangle$. Those $n-m$ functionals are independent, being part of a basis, so they are a basis of $\mathbb{W}^0$ and Equation (5.89) follows. □

Corollary 5.58 (A functional that annihilates a set of constraints)**.** *Let $f^1, \ldots, f^m \in \mathbb{V}^*$ be linearly independent,* $\dim \mathbb{V} = n < \infty$*, and put $\mathbb{W} = \bigcap_{\alpha=1}^m \ker f^\alpha$. Then* $\dim \mathbb{W} = n - m$*, and every functional vanishing on $\mathbb{W}$ is a linear combination $\sum_\alpha \mu_\alpha f^\alpha$ with uniquely determined coefficients μ_α.*

Rests on Proposition 5.57, Theorem 5.38, and Definition 5.13.

Proof. *Derives Corollary 5.58.*

Let $A : \mathbb{V} \longrightarrow \mathbb{K}^m$ send v to $(f^1(v), \ldots, f^m(v))$; it is linear and its kernel is $\mathbb{W}$. It is surjective: otherwise its image is a proper subspace $\mathbb{S}$ of $\mathbb{K}^m$, and Proposition 5.57 applied in $\mathbb{K}^m$ gives $\dim \mathbb{S}^0 = m - \dim \mathbb{S} \geq 1$, so some non-zero $(c_1, \ldots, c_m)$ satisfies $\sum_\alpha c_\alpha f^\alpha(v) = 0$ for every v — a vanishing non-trivial combination, contradicting independence. Hence Equation (5.68) gives $\dim \mathbb{W} = n - m$.

By Equation (5.89), $\dim \mathbb{W}^0 = n - (n - m) = m$. Each f^α vanishes on $\mathbb{W}$ and so lies in $\mathbb{W}^0$; the m of them are independent, hence a basis of $\mathbb{W}^0$, which is the assertion, the coefficients being the unique components in that basis. □

Inner product established by an isomorphism between a space and its dual

Proposition 5.59. *Let $\mathbb{V}$ be a finite-dimensional vector space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, and let $\Phi : \mathbb{V} \longrightarrow \mathbb{V}^*$ be a bijection that is linear if $\mathbb{K} = \mathbb{R}$ and* antilinear *if $\mathbb{K} = \mathbb{C}$, that is,*

$$\Phi(au + bv) = a^* \Phi(u) + b^* \Phi(v)\,. \tag{5.90}$$

Consider the map defined by

$$\langle u|v\rangle = (\Phi(u))(v)\,. \tag{5.91}$$

Then:

1. *for* any *such Φ, Equation (5.91) satisfies the linearity axioms Equations (5.34) and (5.35) and their conjugate partners Equations (5.36) and (5.37), and it is* non-degenerate*: $\langle u|v\rangle = 0$ for all v forces $u = 0_{\mathbb{V}}$, and $\langle u|v\rangle = 0$ for all u forces $v = 0_{\mathbb{V}}$;*
2. *the remaining axioms Equations (5.31) to (5.33) hold if and only if Φ is in addition* hermitian*,*
$$(\Phi(u))(v) = ((\Phi(v))(u))^*\,, \qquad \forall\, u, v \in \mathbb{V}\,, \tag{5.92}$$
and positive*,*
$$(\Phi(v))(v) > 0\,, \qquad \forall\, v \in \mathbb{V} \setminus \{0_{\mathbb{V}}\}\,. \tag{5.93}$$

Conversely, every inner product on $\mathbb{V}$ arises from exactly one Φ satisfying Equations (5.90), (5.92) and (5.93).

Rests on Definitions 5.17 and 5.56 and Equations (5.86) and (5.87).

Proof. *Derives Proposition 5.59.*

Part 1. $\Phi(u)$ is by construction an element of $\mathbb{V}^* = \mathcal{L}(\mathbb{V}, \mathbb{K})$, that is, a *linear* functional on $\mathbb{V}$. Hence, for $a, b \in \mathbb{K}$ and $v, w \in \mathbb{V}$,

$$\langle u|av + bw\rangle = (\Phi(u))(av + bw) = a(\Phi(u))(v) + b(\Phi(u))(w) = a\,\langle u|v\rangle + b\,\langle u|w\rangle\,,$$

which is Equation (5.34) (take $b = 0$) and Equation (5.35) (take $a = b = 1$). In the first slot the behaviour is that of Φ: by Equation (5.90),

$$\langle au + bv|w\rangle = (\Phi(au + bv))(w) = a^*(\Phi(u))(w) + b^*(\Phi(v))(w) = a^*\,\langle u|w\rangle + b^*\,\langle v|w\rangle\,,$$

which is Equation (5.36) and Equation (5.37). (For $\mathbb{K} = \mathbb{R}$ the conjugation is the identity and antilinearity is linearity, so the same two lines cover both fields.)

For non-degeneracy: if $\langle u|v\rangle = 0$ for every v then $\Phi(u)$ is the zero functional, and Φ being injective — with $\Phi(0_{\mathbb{V}}) = 0$ by Equation (5.90) — gives $u = 0_{\mathbb{V}}$. If instead $\langle u|v\rangle = 0$ for every u, then $f(v) = 0$ for every $f \in \mathbb{V}^*$, because Φ is surjective; taking f to be the dual-basis functional f^1 of Equation (5.86) built from a basis whose first element is v — possible if $v \neq 0_{\mathbb{V}}$ — would give $f^1(v) = 1 \neq 0$, a contradiction, so $v = 0_{\mathbb{V}}$.

Part 2. The three remaining axioms are, one by one, restatements of the two extra conditions.

- Hermiticity Equation (5.33) reads $\langle u|v\rangle^* = \langle v|u\rangle$, which written out through Equation (5.91) is exactly Equation (5.92).
- Positive definiteness Equation (5.32) reads $\langle v|v\rangle = (\Phi(v))(v) > 0$ for $v \neq 0_{\mathbb{V}}$, which is Equation (5.93).
- Identity of indiscernibles Equation (5.31): the direction ($\Leftarrow$) holds for any Φ, since $\Phi(0_{\mathbb{V}}) = 0$ and the zero functional sends every vector, in particular $0_{\mathbb{V}}$, to 0. The direction ($\Rightarrow$) is the contrapositive of Equation (5.93): if $v \neq 0_{\mathbb{V}}$ then $\langle v|v\rangle > 0 \neq 0$.

Converse. Given an inner product on $\mathbb{V}$, define $\Phi(u) = \langle u|\ \rangle$, the functional $v \mapsto \langle u|v\rangle$. It is a functional by Equations (5.34) and (5.35), it is antilinear in u by Equations (5.36) and (5.37), and it is injective, for $\Phi(u) = 0$ gives in particular $\langle u|u\rangle = 0$ and hence $u = 0_{\mathbb{V}}$ by Equation (5.31). An injective map that is $\mathbb{R}$-linear between two real vector spaces of the same finite dimension — and $\mathbb{V}$, $\mathbb{V}^*$ have the same dimension over $\mathbb{K}$, hence the same dimension over $\mathbb{R}$ — is surjective, so Φ is a bijection of the required kind, and it manifestly reproduces the given inner product through Equation (5.91). Uniqueness is immediate: two such maps agreeing in Equation (5.91) agree as functionals on every v, hence are equal. □

Remark 5.60 (Why an arbitrary isomorphism will not do). The source states Proposition 5.59 for an arbitrary isomorphism $\Phi : \mathbb{V} \longrightarrow \mathbb{V}^*$, on the strength of the isomorphism established after Equation (5.87). That is false, and it fails in two separate ways which are worth keeping apart.

- *Even over $\mathbb{R}$, most isomorphisms give no inner product.* Take $\mathbb{V} = \mathbb{R}^2$ with basis $\{e_1, e_2\}$ and the dual basis $\{f^1, f^2\}$ of Equation (5.86), and let Φ be the linear isomorphism with $\Phi(e_1) = f^2$ and $\Phi(e_2) = f^1$. Then $\langle e_1|e_1\rangle = f^2(e_1) = 0$ although $e_1 \neq 0_{\mathbb{V}}$, so both Equation (5.31) and Equation (5.32) fail. The choice $\Phi(e_1) = f^1$, $\Phi(e_2) = -f^2$ fails differently, giving $\langle e_2|e_2\rangle = -1$: it is a perfectly good non-degenerate symmetric form, of signature $(1, 1)$ rather than $(2, 0)$, and the reader will meet exactly this distinction again in Section 5.6.3. What Equation (5.91) produces from a bare isomorphism is a non-degenerate form, no more.
- *Over $\mathbb{C}$ no linear isomorphism can work at all.* Suppose Φ were $\mathbb{C}$-linear and Equation (5.91) an inner product. Linearity of Φ would give $\langle au|v\rangle = a\,\langle u|v\rangle$, whereas the axioms give $\langle au|v\rangle = a^*\,\langle u|v\rangle$ by Equation (5.36). Choosing $a = \mathrm{i}$ and $u = v$ with $\langle u|u\rangle = 1$, which Equation (5.32) permits, yields $\mathrm{i} = -\mathrm{i}$: a contradiction. Hence the map must be antilinear, as Equation (5.90) requires.

The upshot is the one physics uses constantly. The isomorphism $\mathbb{V} \simeq \mathbb{V}^*$ of Equation (5.87) is *not canonical* — it was built from a chosen basis, and a different basis gives a different isomorphism. An inner product is precisely the extra datum that singles one out, and the singled-out map is the antilinear correspondence $|v\rangle \longmapsto \langle v|$ of Dirac notation: the bra is the functional $\langle v|\ \rangle$, and it is conjugate-linear in the ket, which is why a phase $\mathrm{e}^{\mathrm{i}\theta}$ on a ket appears as $\mathrm{e}^{-\mathrm{i}\theta}$ on the corresponding bra. In infinite dimension the surjectivity step of the

converse is no longer a dimension count and becomes a theorem requiring completeness — the Riesz representation theorem [336], which this chapter uses nowhere and which belongs to Section 12.3.2, reserved for it there.

5.5.2 Pullback and pushforward

Pullback

Consider a linear transformation of the form $T : \mathbb{V} \longrightarrow \mathbb{W}$ and the linear functional $f : \mathbb{W} \longrightarrow \mathbb{K}$. Clearly $f \in \mathbb{W}^*$. We see moreover that $f \circ T : \mathbb{V} \longrightarrow \mathbb{K}$, and it is for this reason that $f \circ T \in \mathbb{V}^*$. We may ask: given a linear map T taking us from $\mathbb{V}$ to $\mathbb{W}$, what map can we define that brings us back from $\mathbb{W}^*$ to $\mathbb{V}^*$, in terms of T?

The answer is simple. We define the *pullback* of T under f as the map that takes f and assigns to it $f \circ T \in \mathbb{V}^*$. We symbolize the pullback of T as T^*. The idea is clearly illustrated in the diagram

$$\begin{array}{ccccc} \mathbb{V} & \xrightarrow{T} & \mathbb{W} & \xrightarrow{f} & \mathbb{K} \\ \mathbb{V}^* & \xleftarrow{T^*} & \mathbb{W}^* & & \\ f \circ T & \longleftarrow & f & & \end{array}$$

Pushforward

The pullback moves functionals backwards along T. The construction is completed by naming what moves forwards, and the two names together are what fix the index bookkeeping used for the rest of the treatise.

Notation 5.61 (The dual pairing). For $\omega \in \mathbb{V}^*$ and $v \in \mathbb{V}$ we write

$$\langle \omega, v \rangle = \omega(v) \in \mathbb{K} \tag{5.94}$$

and call it the *dual pairing*. It is linear in each argument separately — in v because ω is a functional, in ω because the vector-space operations on $\mathbb{V}^*$ are defined pointwise — so it is a bilinear map in the sense of Definition 5.106. It is *not* an inner product: it does not pair a space with itself, it involves no conjugation, and it needs no structure on $\mathbb{V}$ beyond the vector-space axioms.

Definition 5.62 (Pushforward)**.** Let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be a linear transformation. The *pushforward* along T is the map

$$T_* : \mathbb{V} \longrightarrow \mathbb{W}\,, \qquad T_* v = T(v)\,. \tag{5.95}$$

It is T itself. The notation exists to record a direction, not a new object: a vector is carried *forwards* by T, from $\mathbb{V}$ to $\mathbb{W}$, whereas a functional is carried *backwards* by the pullback T^* of Section 5.5.2, from $\mathbb{W}^*$ to $\mathbb{V}^*$.

Rests on Definitions 5.35 and 5.56.

Proposition 5.63 (Adjointness of pullback and pushforward)**.** *Let $T : \mathbb{V} \longrightarrow \mathbb{W}$ be a linear transformation. Then T^* is a linear transformation $\mathbb{W}^* \longrightarrow \mathbb{V}^*$, and it is characterized by the identity*

$$\langle T^* \omega, v \rangle = \langle \omega, T_* v \rangle\,, \qquad \forall\, \omega \in \mathbb{W}^*,\ \forall\, v \in \mathbb{V}\,. \tag{5.96}$$

The two constructions compose in opposite orders: for $T : \mathbb{V} \longrightarrow \mathbb{W}$ and $S : \mathbb{W} \longrightarrow \mathbb{Z}$,

$$(S \circ T)_* = S_* \circ T_*\,, \tag{5.97}$$

$$(S \circ T)^* = T^* \circ S^*\,. \tag{5.98}$$

Finally, in a basis $\{e_i\}$ of $\mathbb{V}$ and a basis $\{f_a\}$ of $\mathbb{W}$ with their dual bases $\{\varepsilon^i\}$ and $\{\phi^a\}$ built as in Equation (5.86), the matrix of T^* is the transpose of that of T:

$$T^*\phi^a = \sum_i T^a{}_i\, \varepsilon^i \,. \tag{5.99}$$

Rests on Definitions 5.56 and 5.62 and Notation 5.61.

Proof. *Derives Proposition 5.63.*

Linearity of T^*. By definition $T^*\omega = \omega \circ T$, which is a functional on $\mathbb{V}$ as noted in Section 5.5.2. For $a, b \in \mathbb{K}$ and $\omega, \eta \in \mathbb{W}^*$, evaluation at an arbitrary $v \in \mathbb{V}$ gives

$$\left(T^*(a\omega + b\eta)\right)(v) = (a\omega + b\eta)\,(Tv) = a\,\omega(Tv) + b\,\eta(Tv) = \left(aT^*\omega + bT^*\eta\right)(v)\,,$$

the middle step being the pointwise definition of the operations on $\mathbb{W}^*$. Since v was arbitrary the two functionals are equal.

Equation (5.96). Both sides are the same number, written twice:

$$\langle T^*\omega, v\rangle = (\omega \circ T)\,(v) = \omega\,(T(v)) = \langle \omega, T_* v\rangle \ .$$

It characterizes T^* because a functional is determined by its values: if $\langle \Phi\omega, v\rangle = \langle \omega, T_* v\rangle$ for every v then $\Phi\omega$ and $T^*\omega$ agree at every point of $\mathbb{V}$.

The two composition laws. Equation (5.97) is the definition of composition. For Equation (5.98), evaluate on $\omega \in \mathbb{Z}^*$:

$$(S \circ T)^*\,\omega = \omega \circ (S \circ T) = (\omega \circ S) \circ T = T^*\,(S^*\omega) \ ,$$

associativity of composition doing the whole of the work. The same conclusion is reached from Equation (5.96) alone: $\langle (S\circ T)^*\omega, v\rangle = \langle \omega, S_* T_* v\rangle = \langle S^*\omega, T_* v\rangle = \langle T^* S^*\omega, v\rangle$ for all v.

Equation (5.99). Expand $T^*\phi^a$ in the dual basis of $\mathbb{V}^*$ using Equation (5.87), whose coefficients are the values on the basis of $\mathbb{V}$:

$$\begin{aligned}
\left(T^*\phi^a\right)(e_i) &= \phi^a\,(T(e_i)) \\
&= \phi^a\left(\sum_b T^b{}_i f_b\right) && \text{by Equation (5.74)} \\
&= \sum_b T^b{}_i\, \delta^a{}_b && \text{by Equation (5.86)} \\
&= T^a{}_i \,,
\end{aligned}$$

which is Equation (5.99). Note the index positions: the row index of $[T]$ has become the label of the pulled-back functional and the column index the label of the dual basis vector, which is what "transpose" means here. □

Remark 5.64 (Contravariant and covariant, settled once). The pair (T_*, T^*) is what fixes the index conventions used throughout the treatise, and the rule is best read off a change of basis. Let $e'_i = \sum_j A^j{}_i e_j$ with $[A]$ invertible, as in Equation (5.139). Comparing $v = \sum_i v'^i e'_i = \sum_j v^j e_j$ gives $v^j = \sum_i A^j{}_i v'^i$, that is

$$v' = [A]^{-1}\, v \,, \qquad g'_i = \sum_j A^j{}_i g_j \,, \tag{5.100}$$

the second equality being $g'_i = g(e'_i)$ read through Equation (5.84). The components of a *vector* therefore transform with the *inverse* of the basis change — they are *contravariant*, and carry an upper index — while the components of a *functional* transform with the change

itself — they are *covariant*, and carry a lower index. Nothing but Equation (5.96) is at work: the pairing $\langle g, v\rangle = \sum_i g_i v^i$ is a number, and a number cannot depend on the basis, so the two sets of components must transform by mutually inverse rules.

One asymmetry is worth stating plainly, because it is the reason the two names are not interchangeable. A vector can always be pushed forward and a functional can always be pulled back, for *any* linear T. The reverse operations do not exist in general: pushing a functional forward would require $(T^{-1})^*$, hence an invertible T. This is exactly the constraint that reappears on a manifold, where the same construction is performed pointwise on each tangent space — the pushforward acting on tangent vectors and the pullback on covectors (Definition 13.83) — and where the map is accordingly required to be a diffeomorphism before a vector *field* can be transported at all. See Chapter 13; the linear algebra done here is the whole of the content, repeated at every point.

5.6 Eigenvectors, eigenvalues, and decompositions

5.6.1 Eigenvectors and eigenvalues

Eigenvectors and eigenvalues of a linear transformation; characteristic polynomial; diagonalization

An operator mixes the directions of a space; the directions it does *not* mix are the ones along which its action is a single number, and they are what the rest of this section is about.

Definition 5.65 (Eigenvector, eigenvalue, eigenspace)**.** Let $T : \mathbb{V} \longrightarrow \mathbb{V}$ be an endomorphism of a vector space over $\mathbb{K}$. A vector $v \in \mathbb{V}$ is an *eigenvector* of T with *eigenvalue* $\lambda \in \mathbb{K}$ if

$$T(v) = \lambda v\,, \qquad v \neq 0_{\mathbb{V}}\,. \tag{5.101}$$

The set of all λ for which Equation (5.101) has a solution is the *spectrum* of T, and for each such λ the *eigenspace*

$$\mathbb{E}_\lambda = \ker\left(T - \lambda\mathbb{1}\right) = \{v \in \mathbb{V} \mid T(v) = \lambda v\} \tag{5.102}$$

is a subspace, by Definition 5.37; it consists of the eigenvectors for that λ together with $0_{\mathbb{V}}$, which the exclusion in Equation (5.101) keeps out of the list of eigenvectors and which the kernel necessarily contains.

Rests on Definitions 5.37 and 5.54.

The exclusion $v \neq 0_{\mathbb{V}}$ is not fussiness: without it every scalar would be an eigenvalue of every operator, and the notion would carry no information at all.

Definition 5.66 (Characteristic polynomial)**.** Let $\dim \mathbb{V} = n < \infty$ and let T be an endomorphism of $\mathbb{V}$ with matrix $[T]$ in some basis. The *characteristic polynomial* of T is

$$p_T(\mu) = \det\left([T] - \mu\mathbb{1}\right)\,. \tag{5.103}$$

Rests on Definition 5.54 and Equations (5.19) and (5.74).

Derivation. *Derives Definition 5.66.*

Two things have to be checked before Equation (5.103) defines anything: that it is a polynomial of degree n, and that it does not depend on the basis in which $[T]$ was written.

Degree. In the Leibniz sum Equation (5.19) for $\det\left([T] - \mu\mathbb{1}\right)$ the identity permutation contributes $\prod_{i=1}^{n}\left(T^i{}_i - \mu\right)$, a polynomial of degree n with leading term $(-1)^n\mu^n$. Any other permutation moves at least two indices, so at least two of its n factors are off-diagonal entries of $[T]$, carrying no μ; such a term has degree at most $n-2$ and cannot cancel the leading one. Hence $\deg p_T = n$ exactly.

Basis-independence. A change of basis conjugates the array. Let $e'_i = \sum_j A^j{}_i e_j$ with $[A]$ invertible, and let $[T']$ be the matrix of T in the primed basis. Expanding $T(e'_i)$ in the two ways Equation (5.74) allows,

$$T(e'_i) = \sum_j A^j{}_i T(e_j) = \sum_k \left(\sum_j T^k{}_j A^j{}_i \right) e_k \,,$$
$$T(e'_i) = \sum_l \left(T'\right)^l{}_i \, e'_l = \sum_k \left(\sum_l A^k{}_l \left(T'\right)^l{}_i \right) e_k \,,$$

and equating coefficients, which Proposition 5.43 permits, gives $[T]\,[A] = [A]\,[T']$, that is $[T'] = [A]^{-1}\,[T]\,[A]$ — the similarity law recorded again, in its algebra setting, in Remark 5.122. Since $[A]^{-1}\,\mathbb{1}\,[A] = \mathbb{1}$,

$$\begin{aligned} \det\left([A]^{-1}\,[T]\,[A] - \mu\mathbb{1}\right) &= \det\left([A]^{-1}\left([T] - \mu\mathbb{1}\right)[A]\right) \\ &= \det\left([A]^{-1}\right) \det\left([T] - \mu\mathbb{1}\right) \det[A] \\ &= \det\left([T] - \mu\mathbb{1}\right) \,, \end{aligned}$$

the determinant of a product being the product of the determinants and $\det\left([A]^{-1}\right) \det[A] = \det\mathbb{1} = 1$ by Equation (5.78). So p_T is an attribute of the operator, not of the array. □

Theorem 5.67 (The eigenvalues are the roots of the characteristic polynomial)**.** *Let $\dim\mathbb{V} = n < \infty$ and let T be an endomorphism of $\mathbb{V}$. A scalar $\lambda \in \mathbb{K}$ is an eigenvalue of T if and only if $p_T(\lambda) = 0$.*

Rests on Definitions 5.65 and 5.66 and Proposition 5.47.

Proof. *Derives Theorem 5.67.*

The chain is one of equivalences, each already available. Equation (5.101) has a solution $v \neq 0_{\mathbb{V}}$ exactly when $\ker(T - \lambda\mathbb{1}) \neq \{0_{\mathbb{V}}\}$, that is, exactly when the endomorphism $T - \lambda\mathbb{1}$ fails to be injective. By Proposition 5.47 — and here the finite dimension is indispensable — that is exactly when $T - \lambda\mathbb{1}$ fails to be invertible, which by Equation (5.81) is exactly when $\det([T] - \lambda\mathbb{1}) = 0$. The last expression is $p_T(\lambda)$. □

Corollary 5.68 (Existence of an eigenvalue over $\mathbb{C}$)**.** *Every endomorphism of a complex vector space $\mathbb{V}$ with $0 < \dim\mathbb{V} < \infty$ has at least one eigenvalue, and hence at least one eigenvector. Over $\mathbb{R}$ this fails: a rotation of $\mathbb{R}^2$ by an angle that is not a multiple of π has no eigenvector, its characteristic polynomial $\mu^2 - 2\mu\cos\theta + 1$ having no real root.*

Rests on Theorems 5.67 and 8.19.

Proof. *Derives Corollary 5.68.*

By Definition 5.66 the polynomial p_T has degree $n = \dim\mathbb{V} \geq 1$, so by the fundamental theorem of algebra (Theorem 8.19) it has a root $\lambda \in \mathbb{C}$, and Theorem 5.67 makes that root an eigenvalue. For the rotation, the matrix in the standard basis is $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$, whose characteristic polynomial is $(\cos\theta - \mu)^2 + \sin^2\theta = \mu^2 - 2\mu\cos\theta + 1$ with discriminant $4\left(\cos^2\theta - 1\right) = -4\sin^2\theta$, negative unless $\sin\theta = 0$. □

This corollary is the only part of the present subsubsection that the rest of the chapter actually consumes: it is the step with which Schur's first lemma opens (Theorem 5.153), and the two hypotheses it carries — a complex field and a finite non-zero dimension — are inherited by everything that lemma supports, as Remark 5.154 records.

Definition 5.69 (Algebraic and geometric multiplicity)**.** Let λ be an eigenvalue of T. Its *algebraic multiplicity* is the multiplicity of λ as a root of p_T, that is, the largest m with $(\lambda - \mu)^m$ dividing $p_T(\mu)$; its *geometric multiplicity* is $\dim \mathbb{E}_\lambda$.

Rests on Definitions 5.65 and 5.66.

Proposition 5.70. *For every eigenvalue, $1 \leq$ geometric multiplicity $\leq$ algebraic multiplicity, and the second inequality can be strict.*

Rests on Definition 5.69, Lemma 5.36, and Equation (5.19).

Proof. *Derives Proposition 5.70.*

The first inequality restates that an eigenvalue has an eigenvector, so $\mathbb{E}_\lambda \neq \{0_\mathbb{V}\}$.

For the second, write $g = \dim \mathbb{E}_\lambda$, take a basis $\{u_1, \dots, u_g\}$ of $\mathbb{E}_\lambda$ and complete it by Lemma 5.36 to a basis $\{u_1, \dots, u_g, x_1, \dots, x_{n-g}\}$ of $\mathbb{V}$. Since $T(u_i) = \lambda u_i$, the first g columns of $[T]$ in this basis carry λ on the diagonal and nothing below row g:

$$[T] = \begin{pmatrix} \lambda \mathbb{1}_g & B \\ 0 & C \end{pmatrix} . \tag{5.104}$$

A block array of this shape has $\det = \det(\lambda \mathbb{1}_g) \det C$, as the Leibniz sum Equation (5.19) shows: a permutation σ contributes a non-zero term only if $\sigma(j) \leq g$ for every $j \leq g$, because the entries below the first g rows in those columns vanish; being injective, such a σ maps $\{1, \dots, g\}$ onto itself and therefore maps the remaining indices onto themselves as well, so it splits as a pair (σ_1, σ_2) with $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma_1) \operatorname{sgn}(\sigma_2)$ and the sum factorizes into the two Leibniz sums. Applying this to $[T] - \mu \mathbb{1}$, which has the same shape,

$$p_T(\mu) = (\lambda - \mu)^g \det(C - \mu \mathbb{1}) ,$$

so $(\lambda - \mu)^g$ divides p_T and the algebraic multiplicity is at least g.

That the inequality can be strict is shown by $[T] = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ on $\mathbb{C}^2$, the shear that reappears in Remark 5.147 as a representation of $\mathbb{Z}$: here $p_T(\mu) = (1 - \mu)^2$, so $\lambda = 1$ has algebraic multiplicity 2, while $[T] - \mathbb{1} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ has a one-dimensional kernel, spanned by $(1, 0)^\mathsf{T}$, so the geometric multiplicity is 1. □

Proposition 5.71 (Eigenvectors for distinct eigenvalues are independent)**.** *Let $v_1, \dots, v_m$ be eigenvectors of T with pairwise distinct eigenvalues $\lambda_1, \dots, \lambda_m$. Then $\{v_1, \dots, v_m\}$ is linearly independent.*

Rests on Definitions 5.13 and 5.65.

Proof. *Derives Proposition 5.71.*

By induction on m. For $m = 1$ the statement is that a single non-zero vector is independent, which Equation (5.29) gives at once, since $a^1 v_1 = 0_\mathbb{V}$ with $v_1 \neq 0_\mathbb{V}$ forces $a^1 = 0$ (the remark after Equation (5.23)).

Assume the claim for $m - 1$ vectors and let $\sum_{i=1}^m a^i v_i = 0_\mathbb{V}$. Apply the linear operator $T - \lambda_m \mathbb{1}$ to both sides. Each term gives $a^i (T(v_i) - \lambda_m v_i) = a^i (\lambda_i - \lambda_m) v_i$, and the last term vanishes, so

$$\sum_{i=1}^{m-1} a^i (\lambda_i - \lambda_m) v_i = 0_\mathbb{V} .$$

By the induction hypothesis every coefficient vanishes, and $\lambda_i \neq \lambda_m$ for $i < m$ then forces $a^i = 0$. The original relation reduces to $a^m v_m = 0_\mathbb{V}$ with $v_m \neq 0_\mathbb{V}$, whence $a^m = 0$ as well. □

Definition 5.72 (Diagonalizable operator)**.** An endomorphism T of a finite-dimensional space is *diagonalizable* if there is a basis of $\mathbb{V}$ in which $[T] = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$.

Rests on Definition 5.54 and Equation (5.74).

Theorem 5.73 (Criterion for diagonalizability)**.** *An endomorphism T of a space of finite dimension n is diagonalizable if and only if $\mathbb{V}$ possesses a basis consisting of eigenvectors of T. In particular T is diagonalizable whenever p_T has n distinct roots in $\mathbb{K}$.*

Rests on Definitions 5.65 and 5.72 and Proposition 5.71.

Proof. *Derives Theorem 5.73.*

Read Equation (5.74) in a basis $\{e_i\}$: the array $[T]$ is diagonal, with entries λ_i, precisely when $T(e_i) = \lambda_i e_i$ for every i — that is, precisely when every basis vector is an eigenvector. Both directions of the first claim are that one observation.

For the second, let p_T have n distinct roots $\lambda_1, \dots, \lambda_n$ in $\mathbb{K}$. Each is an eigenvalue by Theorem 5.67, so we may choose an eigenvector v_i for each; by Proposition 5.71 the set $\{v_1, \dots, v_n\}$ is linearly independent, and n independent vectors in a space of dimension n are a basis — by Lemma 5.36 they complete to a basis by the adjunction of $n - n = 0$ further vectors. □

Remark 5.74 (Not every operator is diagonalizable). The shear of Proposition 5.70 is not: a basis of eigenvectors would need two independent vectors in a one-dimensional eigenspace. This is the generic obstruction, and Proposition 5.70 localizes it exactly — T is diagonalizable if and only if the geometric multiplicities sum to n, which over an algebraically closed field means that each of them equals its algebraic partner. What replaces diagonalization in general is the Jordan normal form, which this treatise does not need and does not prove: every operator it diagonalizes in earnest is normal — self-adjoint or unitary — and for those Theorem 5.75 and its extension Corollary 5.78 settle the question outright. The shear is not normal, which is the structural reason it resists.

The spectral theorem

Physics almost never diagonalizes an arbitrary operator. It diagonalizes self-adjoint ones (Definition 5.39), and for those the answer is as strong as it could be: they are always diagonalizable, their eigenvalues are always real, and the diagonalizing basis can always be taken orthonormal.

Theorem 5.75 (Spectral theorem for a self-adjoint operator)**.** *Let $\mathbb{V}$ be a complex vector space with an inner product and $0 < \dim \mathbb{V} = n < \infty$, and let T be self-adjoint, $T^\dagger = T$. Then*

1. *every eigenvalue of T is real;*
2. *eigenvectors belonging to distinct eigenvalues are orthogonal;*
3. *$\mathbb{V}$ has an orthonormal basis $\{e_1, \dots, e_n\}$ of eigenvectors of T, in which*

$$[T] = \mathrm{diag}\,(\lambda_1, \dots, \lambda_n)\,, \qquad \lambda_i \in \mathbb{R}\,, \tag{5.105}$$

and $\mathbb{V}$ is the orthogonal direct sum of the eigenspaces.

Rests on Definition 5.39, Corollary 5.68, and Theorem 5.38.

Proof. *Derives Theorem 5.75.*

1. Let $T(v) = \lambda v$ with $v \neq 0_{\mathbb{V}}$. Then

$$\begin{aligned} \lambda \langle v|v\rangle &= \langle v|Tv\rangle && \text{by Equation (5.34)} \\ &= \left\langle T^\dagger v \middle| v \right\rangle && \text{by Equation (5.69)} \\ &= \langle Tv|v\rangle = \langle \lambda v|v\rangle \\ &= \lambda^* \langle v|v\rangle && \text{by Equation (5.36)}\,, \end{aligned}$$

and $\langle v|v\rangle > 0$ by Equation (5.32), so $\lambda = \lambda^*$.

2. Let $T(u) = \lambda u$ and $T(v) = \mu v$ with $\lambda \neq \mu$, both real by part 1. Then $\langle u|Tv\rangle = \mu \langle u|v\rangle$, while $\langle u|Tv\rangle = \left\langle T^\dagger u \middle| v \right\rangle = \langle Tu|v\rangle = \lambda^* \langle u|v\rangle = \lambda \langle u|v\rangle$. Subtracting, $(\mu - \lambda)\langle u|v\rangle = 0$, and $\mu \neq \lambda$ gives $\langle u|v\rangle = 0$.

3. By induction on n, the case $n = 0$ being vacuous. Let $n \geq 1$. By Corollary 5.68 — available because $\mathbb{K} = \mathbb{C}$ and $0 < n < \infty$ — there is an eigenvalue λ_1 with an eigenvector, which we normalize to a unit vector e_1 using Equation (5.41). Put $\mathbb{S} = \langle\{e_1\}\rangle$, so that Equation (5.52) gives $\mathbb{S}^\perp = \{w \in \mathbb{V} \mid \langle e_1|w\rangle = 0\}$.

The complement is T-invariant. For $w \in \mathbb{S}^\perp$,

$$\langle e_1|Tw\rangle = \left\langle T^\dagger e_1 \middle| w \right\rangle = \langle Te_1|w\rangle = \langle \lambda_1 e_1|w\rangle = \lambda_1^* \langle e_1|w\rangle = 0\,,$$

so $T(w) \in \mathbb{S}^\perp$.

Its dimension is $n - 1$. The map $w \longmapsto \langle e_1|w\rangle$ is a functional on $\mathbb{V}$ by Equations (5.34) and (5.35), whose kernel is $\mathbb{S}^\perp$ and whose image is the whole of $\mathbb{C}$, since it sends e_1 to 1. Equation (5.68) gives $\dim \mathbb{S}^\perp = n - 1$.

The restriction $T|_{\mathbb{S}^\perp}$ is an endomorphism of $\mathbb{S}^\perp$, and it is self-adjoint for the restricted inner product, since Equation (5.69) holds in particular for vectors lying in the subspace. The induction hypothesis therefore supplies an orthonormal basis $\{e_2, \ldots, e_n\}$ of $\mathbb{S}^\perp$ consisting of eigenvectors. Adjoining e_1, which is a unit vector orthogonal to every element of $\mathbb{S}^\perp$, gives an orthonormal set of n eigenvectors; it is a basis by Lemma 5.36, an orthonormal set being independent (a vanishing combination paired with e_j returns its j-th coefficient). Equation (5.105) is then Theorem 5.73, and the eigenvalues are real by part 1.

Finally, grouping the basis by eigenvalue writes $\mathbb{V}$ as a sum of subspaces $\mathbb{V}_\lambda \subseteq \mathbb{E}_\lambda$, mutually orthogonal by part 2. The inclusion is an equality: an eigenvector v with eigenvalue λ expands as $v = \sum_i \langle e_i|v\rangle\, e_i$, and pairing $T(v) = \lambda v$ with e_i gives $\lambda_i \langle e_i|v\rangle = \lambda \langle e_i|v\rangle$, so the coefficients with $\lambda_i \neq \lambda$ vanish and $v \in \mathbb{V}_\lambda$. □

Theorem 5.76 (Simultaneous diagonalization of commuting self-adjoint operators)**.** *Let A and B be self-adjoint operators on a complex inner-product space $\mathbb{V}$ with $0 < \dim \mathbb{V} < \infty$. There exists an orthonormal basis of $\mathbb{V}$ whose elements are eigenvectors of A and of B simultaneously if and only if*

$$[A, B] = AB - BA = \Theta\,. \tag{5.106}$$

Rests on Theorem 5.75, Definition 5.39, and Equation (5.102).

Proof. *Derives Theorem 5.76.*

(⇒) The easy direction. Let $\{e_i\}$ be a common eigenbasis, $Ae_i = \alpha_i e_i$ and $Be_i = \beta_i e_i$. Then

$$(AB)\, e_i = \beta_i\, Ae_i = \alpha_i \beta_i e_i = \alpha_i\, Be_i = (BA)\, e_i\,,$$

the scalars commuting in $\mathbb{C}$; two linear operators agreeing on a basis agree everywhere, so $[A, B] = \Theta$.

(⇐) Assume Equation (5.106). By Theorem 5.75 applied to A, the space is the orthogonal direct sum of the eigenspaces $\mathbb{E}_\lambda$ of A, over the finitely many distinct eigenvalues λ of A.

Each eigenspace of A is B-invariant. Let $v \in \mathbb{E}_\lambda$. Then

$$A\,(Bv) = B\,(Av) = B\,(\lambda v) = \lambda\,(Bv)\ ,$$

the first equality being Equation (5.106), so $Bv \in \mathbb{E}_\lambda$ by Equation (5.102). Note that this step, and only this step, uses the commutation.

The restriction $B|_{\mathbb{E}_\lambda}$ is therefore an endomorphism of $\mathbb{E}_\lambda$, and it is self-adjoint for the restricted inner product. Theorem 5.75, applied to it on the space $\mathbb{E}_\lambda$, supplies an orthonormal basis of $\mathbb{E}_\lambda$ consisting of eigenvectors of B; and every vector of $\mathbb{E}_\lambda$ is already an eigenvector of A with eigenvalue λ. Collecting these bases over all λ gives an orthonormal set — vectors from different eigenspaces are orthogonal by Theorem 5.75, part 2 — of $\sum_\lambda \dim \mathbb{E}_\lambda = \dim \mathbb{V}$ vectors, each an eigenvector of both operators. It is a basis for the reason given at the end of the proof of Theorem 5.75. □

Remark 5.77 (This is the theorem of compatible observables). Theorem 5.76 is the piece of linear algebra on which the quantum-mechanical notion of compatibility rests. An observable is represented by a self-adjoint operator (Postulate 77.2) and a measurement returns one of its eigenvalues (Postulate 77.3), so two observables can be assigned definite values in one and the same state exactly when the two operators share an eigenvector — and can be assigned definite values in *every* state of a complete set exactly when they share a whole eigenbasis, which by the theorem happens exactly when they commute. The familiar statement that non-commuting observables cannot be simultaneously sharp is the contrapositive of the direction proved first, the easy one: a common eigenbasis forces the two operators to commute, so operators that fail to commute have none. Note that the eigenvalues are the *measured* quantities and therefore carry SI units — joules for an energy, J s for an angular momentum — while the operators' being self-adjoint is what makes those numbers real, by part 1 of Theorem 5.75.

The other operator physics diagonalizes constantly is the unitary one, and self-adjointness plainly fails for it. Splitting an operator into a self-adjoint and an anti-self-adjoint part reduces that case to the two theorems just proved, at the cost of nothing new.

Corollary 5.78 (Spectral theorem for a normal operator)**.** *Let $\mathbb{V}$ be a complex vector space with an inner product and $0 < \dim \mathbb{V} < \infty$, and let A be an endomorphism of $\mathbb{V}$. Then $\mathbb{V}$ possesses an orthonormal basis of eigenvectors of A if and only if A is* normal*,*

$$A^\dagger A = A A^\dagger\,. \tag{5.107}$$

In particular every unitary operator, $U^\dagger U = U U^\dagger = \mathbb{1}$ (Equation (5.171)), is normal, hence orthonormally diagonalizable, and its eigenvalues all have modulus 1*.*

Rests on Theorem 5.76 and Proposition 5.40.

Proof. *Derives Corollary 5.78.*

Necessity. In an orthonormal eigenbasis $[A]$ is diagonal, so $\left[A^\dagger\right] = [A]^\dagger$ is the array of the conjugate eigenvalues by Equation (5.70), and two diagonal arrays commute.

Sufficiency. Put

$$H = \frac{1}{2}\left(A + A^\dagger\right)\ , \qquad K = \frac{1}{2\mathrm{i}}\left(A - A^\dagger\right)\ ,$$

so that $A = H + \mathrm{i}K$ and $A^\dagger = H - \mathrm{i}K$ by inspection. Both are self-adjoint: in an orthonormal basis Equation (5.70) makes $[H]^\dagger = \left([A]^\dagger + [A]\right)/2 = [H]$, while conjugating the factor $1/(2\mathrm{i})$ changes its sign and reversing the difference changes it back, so $[K]^\dagger = [K]$ as well. Multiplying out,

$$A^\dagger A = H^2 + K^2 + \mathrm{i}\,[H, K]\ , \qquad A A^\dagger = H^2 + K^2 - \mathrm{i}\,[H, K]\ ,$$

so Equation (5.107) holds precisely when $[H, K] = \Theta$. Theorem 5.76 then supplies an orthonormal basis $\{e_i\}$ with $He_i = h_i e_i$ and $Ke_i = k_i e_i$, the eigenvalues real by part 1 of Theorem 5.75, whence $Ae_i = (h_i + \mathrm{i}k_i)\, e_i$: the basis diagonalizes A, with eigenvalues sweeping the complex plane rather than the real line.

Unitary operators. $U^\dagger U = UU^\dagger = \mathbb{1}$ is Equation (5.107) outright. If $Ue = \lambda e$ with $e \neq 0_{\mathbb{V}}$, then Equations (5.34), (5.36) and (5.69) give $\langle e|e\rangle = \left\langle U^\dagger U e \middle| e \right\rangle = \langle Ue|Ue\rangle = |\lambda|^2 \langle e|e\rangle$, and $\langle e|e\rangle > 0$ by Equation (5.32) leaves $|\lambda| = 1$. □

Remark 5.79 (What fails in infinite dimension). Every step above used $\dim \mathbb{V} < \infty$, and the failure in infinite dimension is not a technicality but a change of subject.

- *An operator may have no eigenvalue at all.* On the space of square-integrable functions on an interval, multiplication by the coordinate, $(Qf)(x) = xf(x)$, is self-adjoint and bounded, yet $Qf = \lambda f$ forces f to vanish wherever $x \neq \lambda$, hence to be the zero element. Corollary 5.68 is unavailable because p_T does not exist: there is no determinant.

- *The spectrum must therefore be redefined*, as the set of λ for which $T - \lambda\mathbb{1}$ fails to be invertible rather than the set for which it fails to be injective — two conditions that Proposition 5.47 makes identical here and that Remark 5.48 has already shown to part company.

- *The diagonal sum is replaced by an integral.* The finite-dimensional statement Equation (5.105) becomes a spectral decomposition against a projection-valued measure, which is the form the theorem takes for the operators of quantum mechanics, including the unbounded ones whose domains must be specified before the word "self-adjoint" even means anything [706]. All of this, the spectral theorem in its Hilbert-space form and the specification of domains alike, is the business of Chapter 12.

What survives unchanged is parts 1 and 2 of Theorem 5.75: their proofs used only Equation (5.69) and the inner-product axioms, never a basis or a dimension.

The real symmetric case

Theorem 5.75 is a statement about a complex space, and it has to be: its proof reaches for Corollary 5.68, which holds over $\mathbb{C}$ and fails over $\mathbb{R}$. The operators of mechanics, however, are real symmetric arrays — a matrix of masses, a matrix of stiffnesses, a tensor of inertia — and their eigenvectors have to be real if they are to be displacements or axes. What follows is the real form of the theorem, obtained by borrowing from the complex one only the one thing $\mathbb{R}$ cannot supply, namely that the roots exist.

Theorem 5.80 (Spectral theorem for a real symmetric operator)**.** *Let $\mathbb{V}$ be a* real *vector space with an inner product and $0 < \dim \mathbb{V} = n < \infty$, and let S be self-adjoint, $\langle Su|v\rangle = \langle u|Sv\rangle$ for all $u, v \in \mathbb{V}$. Then the characteristic polynomial of S has n real roots counted with multiplicity, and $\mathbb{V}$ possesses an orthonormal basis $\{e_1, \ldots, e_n\}$ of eigenvectors of S. Equivalently: for every real symmetric array A of order n there is a real orthogonal R, $R^{\mathsf{T}}R = \mathbb{1}$, with*

$$R^{\mathsf{T}} A R = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)\,, \qquad \lambda_r \in \mathbb{R}\,. \tag{5.108}$$

Rests on Theorems 5.67 and 5.75 and Proposition 5.26.

Proof. *Derives Theorem 5.80.*

The two statements are one. Fix an orthonormal basis of $\mathbb{V}$, which exists by Proposition 5.26, and let A be the matrix of S in it. Over $\mathbb{R}$ the conjugation in Equation (5.70) is the identity, so self-adjointness reads $A^{\mathsf{T}} = A$; and a basis of eigenvectors written as

the columns of R turns orthonormality into $R^{\mathsf{T}}R = \mathbb{1}$ and the eigenvalue equations into Equation (5.108).

The roots are real. Read the same array A as the matrix of an operator on $\mathbb{C}^n$ with the standard inner product $\langle u|v\rangle = \sum_i (u^i)^* v^i$. Its adjoint matrix is the conjugate transpose Equation (5.70), which for a real symmetric A is A itself, so the operator is self-adjoint and part 1 of Theorem 5.75 makes every one of its eigenvalues real. The characteristic polynomial is the same polynomial in both readings, being computed from the same array by Equation (5.19); over $\mathbb{C}$ it has n roots with multiplicity by Theorem 8.19, and each of them is an eigenvalue by Theorem 5.67, hence real.

Each root has a real eigenvector. Let λ be one of them. It is a real scalar and a root of the characteristic polynomial of S, so Theorem 5.67 — applied now over $\mathbb{K} = \mathbb{R}$, where it holds verbatim — makes it an eigenvalue of S: there is a $v \in \mathbb{V} \setminus \{0_{\mathbb{V}}\}$ with $Sv = \lambda v$. This is the step at which the complexification is discharged, and it is the only one that needed it.

Induction on n. Normalize $e_1 = v/\|v\|$ and let $\mathbb{S}^{\perp}$ be the orthogonal complement Equation (5.52) of $\langle\{e_1\}\rangle$. The computation made in part 3 of the proof of Theorem 5.75 used only Equation (5.69) and the inner-product axioms, never the complex field, so it holds here word for word: $\mathbb{S}^{\perp}$ is S-invariant, $\dim \mathbb{S}^{\perp} = n - 1$ by Equation (5.68), and the restriction of S to it is self-adjoint for the restricted inner product. The induction hypothesis supplies an orthonormal basis $\{e_2, \dots, e_n\}$ of $\mathbb{S}^{\perp}$ made of eigenvectors, and adjoining e_1 completes one for $\mathbb{V}$. □

Remark 5.81 (Principal axes). Equation (5.108) is the statement a physicist uses under the name *principal axes*: a real symmetric array — of inertia, of strain, of stress, of electric polarizability — is diagonal in some orthonormal frame, and that frame is obtained from any other by a rotation or a reflection, never by a shear. The eigenvalues are then physically measured numbers and carry SI units: $\mathrm{kg\,m^2}$ for the principal moments of inertia, Pa for the principal stresses. Note what distinguishes this from Theorem 5.112, which also brings a real symmetric form to a diagonal one: there the change of basis is an arbitrary congruence and only the *signs* survive it, here it is orthogonal and the eigenvalues themselves survive. Both are true of the same array and they answer different questions.

Theorem 5.82 (Simultaneous diagonalization of a definite pencil)**.** *Let M and K be real symmetric arrays of order n with M positive definite. Then there is an invertible real array U with*

$$U^{\mathsf{T}}MU = \mathbb{1}\,, \qquad U^{\mathsf{T}}KU = \operatorname{diag}(\lambda_1, \dots, \lambda_n)\,, \qquad \lambda_r \in \mathbb{R}\,. \tag{5.109}$$

Equivalently, the generalized eigenvalue problem

$$Ku = \lambda Mu \tag{5.110}$$

has exactly n real roots $\lambda_1, \dots, \lambda_n$ counted with multiplicity, and $\mathbb{R}^n$ has a basis $\{u_1, \dots, u_n\}$ of solutions which is orthonormal for the inner product $\langle u|v\rangle_M = u^{\mathsf{T}}Mv$. The numbers of positive, negative and vanishing λ_r are the triple (p, q, r) that Theorem 5.112 attaches to K, and are therefore the same for every U achieving Equation (5.109).

Rests on Theorems 5.80 and 5.112 and Proposition 5.26.

Proof. *Derives Theorem 5.82.*

M supplies an inner product. Put $\langle u|v\rangle_M = u^{\mathsf{T}}Mv$ on $\mathbb{R}^n$. It is linear in each argument by inspection, symmetric because $M^{\mathsf{T}} = M$, and positive definite by hypothesis, so it satisfies Equations (5.31) to (5.35) and is an inner product in the sense of Definition 5.17, the conjugation being trivial over $\mathbb{R}$.

A first change of basis makes M the identity. Proposition 5.26, applied to the standard basis of $\mathbb{R}^n$ in this inner product, produces a basis $\{w_1, \dots, w_n\}$ with $\langle w_k|w_l\rangle_M = \delta_{kl}$. Let W be the array whose columns are the w_k: it is invertible, its columns being a basis, and the orthonormality relations are exactly $W^{\mathsf{T}}MW = \mathbb{1}$.

A second, orthogonal, change of basis diagonalizes K. The array $\tilde{K} = W^{\mathsf{T}} K W$ is real and symmetric, since $\left(W^{\mathsf{T}} K W\right)^{\mathsf{T}} = W^{\mathsf{T}} K^{\mathsf{T}} W = \tilde{K}$. By Theorem 5.80 there is a real orthogonal R with $R^{\mathsf{T}} \tilde{K} R = \operatorname{diag}(\lambda_r)$ and every λ_r real. Put $U = WR$, invertible as a product of invertible arrays. Then

$$U^{\mathsf{T}} M U = R^{\mathsf{T}} \left(W^{\mathsf{T}} M W\right) R = R^{\mathsf{T}} R = \mathbb{1}\,, \qquad U^{\mathsf{T}} K U = R^{\mathsf{T}} \tilde{K} R = \operatorname{diag}(\lambda_r)\,,$$

which is Equation (5.109).

The second statement is the first one read column by column. Write u_r for the r-th column of U. The first relation of Equation (5.109) says $u_k^{\mathsf{T}} M u_l = \delta_{kl}$, the asserted M-orthonormality; and with $\Lambda = \operatorname{diag}(\lambda_r)$ the two relations combine into $U^{\mathsf{T}} K U = \Lambda = U^{\mathsf{T}} M U \Lambda$, so that $U^{\mathsf{T}}(KU - MU\Lambda) = 0$ and, U^{T} being invertible, $KU = MU\Lambda$ — whose r-th column is Equation (5.110) for u_r with $\lambda = \lambda_r$. That these are all the roots follows from

$$\det(K - \lambda M) = (\det U)^{-2} \det\left(U^{\mathsf{T}} K U - \lambda\, U^{\mathsf{T}} M U\right) = (\det U)^{-2} \prod_{r=1}^{n} (\lambda_r - \lambda)\,,$$

a polynomial of degree n whose roots with multiplicity are the λ_r.

Signature. Equation (5.109) exhibits $\operatorname{diag}(\lambda_r)$ as congruent to K in the sense of Equation (5.139). Rescaling each column of U by $|\lambda_r|^{-1/2}$ wherever $\lambda_r \neq 0$ is a further congruence and brings the array to Equation (5.142) with p, q and r the counts of positive, negative and vanishing λ_r; Theorem 5.112 says those counts are the same for every basis reaching that form. □

Remark 5.83 (What a definite pencil is). The hypothesis that carries Theorem 5.82 is the definiteness of *one* of the two forms, not of both: K is allowed any signature, and the sign of λ_r records it. Neither $M^{-1}K$ nor KM^{-1} is symmetric in general, which is why the theorem is not a corollary of Theorem 5.80 applied to one of them; what is true, and is what the proof exploits, is that $M^{-1}K$ is self-adjoint for the inner product M itself defines. The statement fails outright without definiteness — for $M = \operatorname{diag}(1, -1)$ and $K = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ the roots of Equation (5.110) are $\pm\mathrm{i}$ — so the positive definiteness is not a convenience. In mechanics it is supplied by the kinetic energy: M is a matrix of masses, in kg, and $u^{\mathsf{T}} M u$ is twice a kinetic energy per unit squared velocity, which is positive for every motion that is not rest.

Proposition 5.84 (Rayleigh quotient of a definite pencil)**.** *In the notation of Theorem 5.82, order the roots $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ and define the* Rayleigh quotient

$$\mathcal{R}(u) = \frac{u^{\mathsf{T}} K u}{u^{\mathsf{T}} M u}\,, \qquad u \in \mathbb{R}^n \setminus \{0\}\,. \tag{5.111}$$

Then

$$\lambda_1 \leq \mathcal{R}(u) \leq \lambda_n\,, \tag{5.112}$$

and $\mathcal{R}(u) = \lambda_1$ holds exactly when u solves Equation (5.110) with $\lambda = \lambda_1$, while $\mathcal{R}(u) = \lambda_n$ holds exactly when it solves it with $\lambda = \lambda_n$. In particular λ_1 is the minimum and λ_n the maximum of $\mathcal{R}$, both attained.

Rests on Theorem 5.82.

Proof. *Derives Proposition 5.84.*

The denominator never vanishes, M being positive definite. Expand u in the M-orthonormal basis of Theorem 5.82, $u = \sum_r c^r u_r$, not all c^r zero. Then $u^{\mathsf{T}} M u = \sum_r (c^r)^2$ and $u^{\mathsf{T}} K u = \sum_r \lambda_r (c^r)^2$, so

$$\mathcal{R}(u) = \frac{\sum_r \lambda_r (c^r)^2}{\sum_r (c^r)^2}\,, \tag{5.113}$$

a weighted average of the roots with non-negative weights not all zero, which lies between the smallest and the largest of them — that is Equation (5.112). Subtracting,

$$\mathcal{R}(u) - \lambda_1 = \frac{\sum_r (\lambda_r - \lambda_1)\,(c^r)^2}{\sum_r (c^r)^2}\,,$$

a sum of non-negative terms, which vanishes exactly when $c^r = 0$ for every r with $\lambda_r > \lambda_1$ — that is, exactly when u lies in the span of the u_r belonging to λ_1, which by Equation (5.110) is exactly the solution set for $\lambda = \lambda_1$. The maximum is the same computation with the signs reversed, and both are attained, at u_1 and at u_n. □

Corollary 5.85 (Rayleigh–Ritz)**.** *Every trial vector bounds the lowest root from above, $\lambda_1 \le \mathcal{R}(u)$, and the bound is second-order accurate: if $u = u_1 + w$ with w M-orthogonal to u_1 and $w^{\mathsf{T}} M w = \varepsilon^2$, then*

$$0 \le \mathcal{R}(u) - \lambda_1 \le (\lambda_n - \lambda_1)\,\frac{\varepsilon^2}{1+\varepsilon^2}\,. \tag{5.114}$$

A first-order error in the trial vector therefore costs only a second-order error in the estimate, which is what makes the method worth using.

Rests on Proposition 5.84 and Theorem 5.82.

Proof. *Derives Corollary 5.85.*

The inequality $\lambda_1 \le \mathcal{R}(u)$ is Equation (5.112). For the estimate, write $w = \sum_{r\ge 2} c^r u_r$ — legitimate because w is M-orthogonal to u_1 — so that $\varepsilon^2 = \sum_{r\ge 2} (c^r)^2$ and, by Equation (5.113) with $c^1 = 1$,

$$\mathcal{R}(u) - \lambda_1 = \frac{\sum_{r\ge 2} (\lambda_r - \lambda_1)\,(c^r)^2}{1+\varepsilon^2} \le (\lambda_n - \lambda_1)\,\frac{\varepsilon^2}{1+\varepsilon^2}\,,$$

each factor $\lambda_r - \lambda_1$ being at most $\lambda_n - \lambda_1$. □

Proposition 5.86 (Polar decomposition)**.** *Every invertible real array A of order n factors as*

$$A = OP\,, \qquad O^{\mathsf{T}} O = \mathbb{1}\,, \qquad P^{\mathsf{T}} = P \text{ positive definite}\,, \tag{5.115}$$

and the two factors are uniquely determined by A.

Rests on Theorem 5.80 and Proposition 5.47.

Proof. *Derives Proposition 5.86.*

Existence. $B = A^{\mathsf{T}} A$ is symmetric, and for $x \neq 0$, $x^{\mathsf{T}} B x = (Ax)^{\mathsf{T}} (Ax) > 0$, since A is invertible and therefore injective (Proposition 5.47), so $Ax \neq 0$. By Theorem 5.80, $B = R\,\mathrm{diag}\,(\mu_r)\,R^{\mathsf{T}}$ with R orthogonal, and each $\mu_r > 0$, being $x^{\mathsf{T}} B x$ for the corresponding unit eigenvector. Put

$$P = R\,\mathrm{diag}\,(\sqrt{\mu_r})\,R^{\mathsf{T}}\,,$$

symmetric by inspection, positive definite because its eigenvalues $\sqrt{\mu_r}$ are positive, and satisfying $P^2 = B$ and hence invertible. Set $O = AP^{-1}$; then, P^{-1} being symmetric,

$$O^{\mathsf{T}} O = P^{-1} A^{\mathsf{T}} A P^{-1} = P^{-1} P^2 P^{-1} = \mathbb{1}\,.$$

Uniqueness. Suppose $A = OP$ as in Equation (5.115). Then $A^{\mathsf{T}} A = P\,O^{\mathsf{T}} O\,P = P^2$, so P is *a* positive definite symmetric square root of the array B fixed by A. There is only one. Indeed such a P commutes with $B = P^2$, so it maps each eigenspace $\mathbb{E}_\mu$ of B into itself: $Bx = \mu x$ gives $B(Px) = P(Bx) = \mu\,Px$. The restriction of P to $\mathbb{E}_\mu$ is symmetric, so by Theorem 5.80 that subspace has a basis of eigenvectors of P, whose eigenvalues ν satisfy $\nu^2 = \mu$ and $\nu > 0$, hence $\nu = \sqrt{\mu}$: P acts on $\mathbb{E}_\mu$ as multiplication by $\sqrt{\mu}$. Since the eigenspaces of B span $\mathbb{R}^n$ by Theorem 5.80, P is determined by B, and then $O = AP^{-1}$ is determined too. □

Corollary 5.87 (The operator norm of a real array)**.** *For a real array L of order n put $\|L\| = \max\{\|Lx\| \mid x \in \mathbb{R}^n,\ \|x\| = 1\}$, the maximum being attained. Then*

$$\|L\|^2 = \lambda_{\max}\left(L^{\mathsf{T}}L\right) = \lambda_{\max}\left(LL^{\mathsf{T}}\right), \tag{5.116}$$

the largest eigenvalue of either of those two symmetric positive semi-definite arrays.
Rests on Proposition 5.84 and Theorem 5.80.

Proof. *Derives Corollary 5.87.*
$\|Lx\|^2 = x^{\mathsf{T}}\left(L^{\mathsf{T}}L\right)x$, and $L^{\mathsf{T}}L$ is symmetric with $x^{\mathsf{T}}L^{\mathsf{T}}Lx = \|Lx\|^2 \geq 0$, hence positive semi-definite. Taking $M = \mathbb{1}$ in Proposition 5.84, which is legitimate because the identity is positive definite, the quotient $\|Lx\|^2 / \|x\|^2$ has maximum $\lambda_{\max}\left(L^{\mathsf{T}}L\right)$, attained at a corresponding eigenvector; restricting to $\|x\| = 1$ gives the first equality of Equation (5.116) and the attainment.

For the second, let $L^{\mathsf{T}}Lv = \lambda v$ with $v \neq 0$ and $\lambda \neq 0$. Then $Lv \neq 0$, since $L^{\mathsf{T}}Lv \neq 0$, and $LL^{\mathsf{T}}(Lv) = L\left(L^{\mathsf{T}}Lv\right) = \lambda Lv$: every non-zero eigenvalue of $L^{\mathsf{T}}L$ is one of LL^{T}, and by symmetry of the argument the converse holds too. If instead $\lambda_{\max}\left(L^{\mathsf{T}}L\right) = 0$ then, all its eigenvalues being non-negative and at most 0, $L^{\mathsf{T}}L$ has every eigenvalue zero, so $\|Lx\|^2 = 0$ for every x by Equation (5.108), so $L = 0$ and both sides vanish. □

Proposition 5.88 (The eigenvalues do not control the norm)**.** *Let A be a real array of order $n \geq 2$ and $\|A\|$ its operator norm.*

1. *If $Av = \lambda v$ with $v \in \mathbb{R}^n \setminus \{0\}$ and $\lambda \in \mathbb{R}$, then $|\lambda| \leq \|A\|$.*
2. *If A is symmetric, then $\|A\| = \max_r |\lambda_r|$ over its eigenvalues.*
3. *Without symmetry the second statement fails, and fails unboundedly: for*

$$A = \begin{pmatrix} \lambda & c \\ 0 & \lambda \end{pmatrix} \tag{5.117}$$

— extended to order n by zeros — every eigenvalue has modulus $|\lambda|$ or is zero, while $\|A\| \geq |c|$, and c is free.

Rests on Corollary 5.87 and Theorem 5.80.

Proof. *Derives Proposition 5.88.*
1. Normalize v; then $|\lambda| = \|\lambda v\| = \|Av\| \leq \|A\|$ by the definition of the operator norm.
2. For symmetric A, Equation (5.108) gives $A^{\mathsf{T}}A = A^2 = R\,\mathrm{diag}\left(\lambda_r^2\right)R^{\mathsf{T}}$, so the eigenvalues of $A^{\mathsf{T}}A$ are the λ_r^2 and Equation (5.116) gives $\|A\|^2 = \max_r \lambda_r^2$.
3. The characteristic polynomial of Equation (5.117) is $(\lambda - \mu)^2$, so λ is its only eigenvalue, whatever c; the padding contributes only the eigenvalue 0. Applying A to the second standard basis vector gives the vector with components (c, λ), of norm at least $|c|$, so $\|A\| \geq |c|$. □

Remark 5.89 (What this costs elsewhere). Proposition 5.88 is the linear-algebraic root of a fact that surprises whoever meets it first in a physical problem: an array every one of whose eigenvalues is small can still stretch some vector by a large factor. The eigenvectors of a non-symmetric array are in general not orthogonal, so a superposition of them can be much shorter than its individual terms, and cancellations of that kind can be undone by the array. The dynamical consequence — that a linear system every one of whose modes decays can nevertheless amplify a disturbance by a large factor before decaying — is a statement about the exponential of A rather than about A, and this chapter does not make it; what belongs here is the statement that the spectrum is not the norm.

Quadrics and the conic sections

Proposition 5.90 (Principal axes of a real quadratic form)**.** *Let $q(x) = x^{\mathsf{T}} A x$ on $\mathbb{R}^n$ with A real and symmetric. Then there is an orthogonal change of coordinates $x = Ry$ in which*

$$q = \sum_{r=1}^{n} \lambda_r \, (y^r)^2 \,, \tag{5.118}$$

the λ_r being the eigenvalues of A and the columns of R — the principal axes *of q — an orthonormal basis of eigenvectors. The level set $\{q = 1\}$ is accordingly an ellipsoid when every $\lambda_r > 0$; when the signs differ it is a hyperboloid, when some λ_r vanishes a cylinder over one of these, and when no λ_r is positive it is empty.*

Rests on Theorems 5.80 and 5.112 and Definition 5.113.

Proof. *Derives Proposition 5.90.*

Substituting $x = Ry$ gives $q = y^{\mathsf{T}} \left(R^{\mathsf{T}} A R \right) y$, and Equation (5.108) makes the array in the middle diagonal with the eigenvalues along it, which is Equation (5.118). The level set is then read off Equation (5.118) term by term, and the pattern of signs is the signature of Theorem 5.112 and Definition 5.113, hence independent of every choice made along the way. □

Proposition 5.91 (The focal polar equation of a conic)**.** *Let $p > 0$ and $e \geq 0$, and let $C \subset \mathbb{R}^2$ be the set of points whose polar coordinates $x = r\cos\varphi$, $y = r\sin\varphi$ with $r > 0$ satisfy*

$$r = \frac{p}{1 + e\cos\varphi} \,. \tag{5.119}$$

Then C is exactly the set of points of the quadric

$$\left(1 - e^2\right) x^2 + y^2 + 2pe\, x - p^2 = 0 \tag{5.120}$$

obeying $ex < p$; and for $e > 0$ it is the locus of points whose distance to the origin is e times their distance to the line $x = p/e$. The quadratic part of Equation (5.120) is already in the principal axes of Proposition 5.90, with eigenvalues $1 - e^2$ and 1, so that

1. *for $0 \leq e < 1$, C is an ellipse with semi-axes*

$$a = \frac{p}{1 - e^2} \,, \qquad b = \frac{p}{\sqrt{1 - e^2}} \,, \tag{5.121}$$

   centred at $(-ae, 0)$: the origin is one of its two foci, the foci lie $2ae$ apart, and the area enclosed is πab;

2. *for $e = 1$, C is the parabola $y^2 = p^2 - 2px$, whose focus is the origin;*

3. *for $e > 1$, C is the branch $ex < p$ of a hyperbola, again with the origin at a focus.*

In every case $p = r(\pi/2)$, the semi-latus rectum.

Rests on Proposition 5.90, Equation (5.10), and Corollary 7.44.

Proof. *Derives Proposition 5.91.*

The two descriptions agree. Multiplying Equation (5.119) out and using $r\cos\varphi = x$ gives

$$r = p - ex \,, \tag{5.122}$$

which requires $ex < p$ because $r > 0$; squaring it and substituting $r^2 = x^2 + y^2$ produces Equation (5.120). Conversely let a point satisfy Equation (5.120) with $ex < p$. It is not the origin, which would give $-p^2 = 0$, so $r > 0$; and Equation (5.120) says $r^2 = (p - ex)^2$ with $p - ex > 0$, so Equation (5.122) holds and, divided by r, returns Equation (5.119). For $e > 0$

the distance from the point to the line $x = p/e$ is $p/e - x$, positive by the same inequality, and Equation (5.122) says exactly $r = e\,(p/e - x)$. Putting $\varphi = \pi/2$ in Equation (5.119) gives $r = p$.

$0 \le e < 1$. Completing the square in Equation (5.120),

$$\left(1 - e^2\right)\left(x + \frac{pe}{1 - e^2}\right)^2 + y^2 = p^2 + \frac{p^2e^2}{1 - e^2} = \frac{p^2}{1 - e^2}\,,$$

and dividing by the right-hand side gives

$$\frac{(x + ae)^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{5.123}$$

with a and b as in Equation (5.121), since $pe/(1 - e^2) = ae$. This is an ellipse centred at $(-ae, 0)$, and

$$a^2 - b^2 = \frac{p^2}{(1 - e^2)^2} - \frac{p^2}{1 - e^2} = \frac{p^2e^2}{(1 - e^2)^2} = (ae)^2\,,$$

so the two foci, at distance $\sqrt{a^2 - b^2} = ae$ from the centre along the major axis, are the origin and $(-2ae, 0)$. The side condition $ex < p$ is automatic here: Equation (5.123) confines x to $x \le a(1 - e) = p/(1 + e)$, and $ep/(1 + e) < p$.

The area. For each $u = x + ae$ in $[-a, a]$, Equation (5.123) gives a vertical segment of length $2b\sqrt{1 - u^2/a^2}$, so the area enclosed is, substituting first $u = at$ and then $t = \sin\theta$ by Corollary 7.44,

$$\begin{aligned}
\int_{-a}^{a} 2b\sqrt{1 - \frac{u^2}{a^2}}\,\mathrm{d}u &= 2ab\int_{-1}^{1}\sqrt{1 - t^2}\,\mathrm{d}t \\
&= 2ab\int_{-\pi/2}^{\pi/2}\cos^2\theta\,\mathrm{d}\theta \\
&= ab\int_{-\pi/2}^{\pi/2}(1 + \cos 2\theta)\,\mathrm{d}\theta \qquad \text{by Equation (5.10)} \\
&= \pi ab\,,
\end{aligned}$$

the integral of $\cos 2\theta$ vanishing over a full period of it.

$e = 1$ and $e > 1$. For $e = 1$ the quadratic part loses its x^2 term and Equation (5.120) becomes $y^2 = p^2 - 2px$, a parabola with vertex $(p/2, 0)$ opening towards decreasing x; written as $y^2 = -2p\,(x - p/2)$ it has focal length $p/2$, so its focus is at $(p/2 - p/2, 0)$, the origin. For $e > 1$ the eigenvalues $1 - e^2$ and 1 have opposite signs, so by Proposition 5.90 the quadric is a hyperbola; the side condition $ex < p$ selects the branch containing the point $\varphi = \pi/2$, since the other branch lies where $p - ex < 0$. □

Remark 5.92 (Why a physics chapter needs this). Equation (5.119) is the form in which an orbit under an inverse-square force emerges from its equation of motion, with p and e built out of the angular momentum and the energy. Without Proposition 5.91 the sentence "the orbit is an ellipse with the force centre at a focus" would be a name attached to a formula; with it, it is a theorem about a locus, and the semi-axes Equation (5.121) and the area πab are what turn the formula into the periods that are actually measured. The three cases are distinguished by the signature of the quadratic part, so the trichotomy ellipse–parabola–hyperbola is a statement about Theorem 5.112 and not about the coordinates chosen.

Generalized eigenspaces

Theorem 5.73 tells us when an operator has an eigenbasis. When it has none, the question is what it has instead, and the answer over $\mathbb{C}$ is that the space still splits into subspaces on each of which the operator is a scalar plus something nilpotent. That is all the structure the physical parts of this treatise ever use, and it is proved here in full.

Lemma 5.93 (Fitting splitting)**.** *Let T be an endomorphism of a vector space $\mathbb{V}$ over $\mathbb{K}$ with $\dim \mathbb{V} = n < \infty$, let $\lambda \in \mathbb{K}$, and write $N = T - \lambda \mathbb{1}$. Then*

$$\mathbb{V} = \ker\left(N^n\right) \oplus \operatorname{im}\left(N^n\right) , \tag{5.124}$$

both summands are T-invariant, and N is nilpotent on the first and invertible on the second.
Rests on Theorem 5.38, Definition 5.37, and Proposition 5.47.

Proof. *Derives Lemma 5.93.*

The kernels stabilize by step n. The subspaces $\ker\left(N^k\right)$ increase with k. If two consecutive ones agree, $\ker\left(N^k\right) = \ker\left(N^{k+1}\right)$, the chain is constant from there on: $N^{k+2}v = 0$ gives $N^{k+1}\left(Nv\right) = 0$, so $Nv \in \ker\left(N^{k+1}\right) = \ker\left(N^k\right)$ and $N^{k+1}v = 0$. Every strict increase costs at least one dimension and the dimensions are bounded by n, so a repetition has occurred by $k = n$ and $\ker\left(N^n\right) = \ker\left(N^{2n}\right)$.

The sum is direct and fills the space. Let v lie in both summands, $v = N^n w$ with $N^n v = 0$. Then $N^{2n} w = 0$, so $w \in \ker\left(N^{2n}\right) = \ker\left(N^n\right)$ and $v = N^n w = 0_{\mathbb{V}}$. The two dimensions add up to n by Equation (5.68), so Equation (5.124) holds.

Invariance and the two behaviours. T commutes with N and hence with N^n, so it carries kernel and image of N^n into themselves. On $\ker\left(N^n\right)$ the power N^n vanishes by definition, so N is nilpotent there. On $\operatorname{im}\left(N^n\right)$ the map N is injective — if v lies there and $Nv = 0$, then v lies in $\ker N$, a subspace of $\ker\left(N^n\right)$, so v lies in both summands and vanishes — and therefore invertible on it, by Proposition 5.47. □

Theorem 5.94 (Primary decomposition)**.** *Let T be an endomorphism of a complex vector space $\mathbb{V}$ with $0 < \dim \mathbb{V} = n < \infty$, and let $\lambda_1, \ldots, \lambda_k$ be its distinct eigenvalues. Write*

$$\mathbb{G}_i = \ker\left(\left(T - \lambda_i \mathbb{1}\right)^n\right) \tag{5.125}$$

for the generalized eigenspace *of λ_i. Then*

$$\mathbb{V} = \mathbb{G}_1 \oplus \ldots \oplus \mathbb{G}_k , \tag{5.126}$$

every $\mathbb{G}_i$ is T-invariant and contains the eigenspace $\mathbb{E}_{\lambda_i}$, and $T - \lambda_i \mathbb{1}$ is nilpotent on $\mathbb{G}_i$.
Rests on Lemma 5.93, Corollary 5.68, and Equation (5.102).

Proof. *Derives Theorem 5.94.*

By induction on n. Corollary 5.68 supplies an eigenvalue λ_1, and Lemma 5.93 applied to it splits $\mathbb{V} = \mathbb{G}_1 \oplus \mathbb{W}$ with $\mathbb{W} = \operatorname{im}\left(N^n\right)$, $N = T - \lambda_1 \mathbb{1}$. The first summand contains an eigenvector, so it is not trivial and $\dim \mathbb{W} < n$. If $\mathbb{W}$ is trivial, then $k = 1$ and there is nothing left to prove; otherwise T restricts to an endomorphism of $\mathbb{W}$ and the induction hypothesis decomposes $\mathbb{W}$ into the generalized eigenspaces, computed inside $\mathbb{W}$, of the distinct eigenvalues of that restriction. Two identifications complete the argument, and both rest on one observation:

$T - \mu \mathbb{1}$ is invertible on $\mathbb{G}_1$ for every $\mu \neq \lambda_1$. On that subspace $T - \mu \mathbb{1} = N + (\lambda_1 - \mu)\,\mathbb{1}$ is a nilpotent operator plus a non-zero scalar, and such a sum is invertible: with $c = \lambda_1 - \mu$ and $N^n = \Theta$ there, $(N + c\mathbb{1})^{-1} = c^{-1} \sum_{j=0}^{n-1} (-N/c)^j$, a finite sum, as multiplying out and telescoping shows.

The eigenvalues of T are λ_1 together with those of the restriction. An eigenvector of the restriction is one of T. Conversely let $Tv = \mu v$ with $v \neq 0_{\mathbb{V}}$ and $\mu \neq \lambda_1$, and split $v = g + w$ along Equation (5.124); both summands being T-invariant, $Tg = \mu g$, so $(T - \mu\mathbb{1})\, g = 0_{\mathbb{V}}$ and the observation forces $g = 0_{\mathbb{V}}$: the eigenvector lies in $\mathbb{W}$.

The generalized eigenspaces agree. Fix $\mu \neq \lambda_1$ and let $v \in \ker\left((T - \mu\mathbb{1})^n\right)$. Split $v = g + w$ along Equation (5.124); both summands are T-invariant, so $(T - \mu\mathbb{1})^n\, g = 0_{\mathbb{V}}$, and the observation — which makes that operator invertible on $\mathbb{G}_1$, a power of an invertible operator being invertible — forces $g = 0_{\mathbb{V}}$. Hence $\ker\left((T - \mu\mathbb{1})^n\right)$ lies inside $\mathbb{W}$. Computed inside $\mathbb{W}$ it is the kernel of the $\dim \mathbb{W}$-th power, which by the first paragraph of the proof of Lemma 5.93 is the same subspace as the kernel of the n-th power. So Equation (5.126) is the induction hypothesis with $\mathbb{G}_1$ adjoined.

Finally $\mathbb{E}_{\lambda_i} \subseteq \mathbb{G}_i$ because a vector killed by $T - \lambda_i\mathbb{1}$ is killed by its powers (Equation (5.102)); invariance is Lemma 5.93; and the nilpotency is Equation (5.125) itself. □

Corollary 5.95 (Semisimple and nilpotent parts)**.** *With the notation of Theorem 5.94, let S be the operator acting on $\mathbb{G}_i$ as multiplication by λ_i, and put $N = T - S$. Then S is diagonalizable, N satisfies $N^n = \Theta$, and $[S, N] = \Theta$. Consequently, in any basis adapted to Equation (5.126), the matrix of T is block diagonal, the i-th block being $\lambda_i\mathbb{1}$ plus a nilpotent array.*

Rests on Theorem 5.94 and Definition 5.72.

Proof. *Derives Corollary 5.95.*

Any basis of $\mathbb{G}_i$ consists of eigenvectors of S with eigenvalue λ_i, so collecting such bases over i diagonalizes S, by Equation (5.126). On $\mathbb{G}_i$ the operator N is $T - \lambda_i\mathbb{1}$, nilpotent with $N^n = \Theta$ there by Equation (5.125); since the subspaces exhaust $\mathbb{V}$, $N^n = \Theta$ throughout. And on $\mathbb{G}_i$ the operator S is a scalar, which commutes with everything that preserves the subspace, so $[S, N] = \Theta$ on each summand and hence everywhere. □

Remark 5.96 (What is and is not claimed here). Corollary 5.95 is the whole of the structure this treatise uses. It is weaker than the classical *normal form*, which refines the nilpotent part of each block into a direct sum of single chains of prescribed lengths; that refinement is a finer classification of nilpotent operators, nothing in this treatise needs it, and it is therefore not carried here — an omission recorded rather than a result assumed. Note also what fails outside $\mathbb{C}$: Theorem 5.94 rests on Corollary 5.68, so over $\mathbb{R}$ the decomposition can be empty of content, exactly as diagonalizability can, and the honest statement for a real array is the one obtained by regarding it as complex. The use the physical parts make of the corollary — that a linear system of differential equations with constant coefficients has solutions of the form $t^j \mathrm{e}^{\lambda t}$ — is a statement about differential equations rather than about arrays, and this chapter does not make it.

5.6.2 Decomposition of vector spaces

Direct sum; direct product

Definition 5.97 (External direct sum)**.** Let $\mathbb{V}$ and $\mathbb{W}$ be vector spaces over the same field $\mathbb{K}$. Their *direct sum* $\mathbb{V} \oplus \mathbb{W}$ is the Cartesian product $\mathbb{V} \times \mathbb{W}$ with the operations performed in each slot separately,

$$(v, w) + (v', w') = (v + v',\, w + w') \;, \qquad a\,(v, w) = (av,\, aw) \;. \tag{5.127}$$

Rests on Definitions 4.31 and 5.4.

That this is a vector space is Example 5.7 once more, with the two slots playing the role of the two components: the neutral vector is $(0_{\mathbb{V}}, 0_{\mathbb{W}})$ and every axiom is verified slot by slot.

Proposition 5.98 (Dimensions add)**.** *If $\mathbb{V}$ and $\mathbb{W}$ are finite-dimensional over the same field, then*

$$\dim(\mathbb{V} \oplus \mathbb{W}) = \dim \mathbb{V} + \dim \mathbb{W}. \tag{5.128}$$

Rests on Definitions 5.14 and 5.97 and Lemma 5.36.

Proof. *Derives Proposition 5.98.*

Let $\{e_i\}_{i=1}^{p}$ and $\{f_j\}_{j=1}^{q}$ be bases of $\mathbb{V}$ and $\mathbb{W}$. The $p+q$ vectors $(e_i, 0_{\mathbb{W}})$ and $(0_{\mathbb{V}}, f_j)$ generate: any (v, w) equals $\sum_i v^i (e_i, 0_{\mathbb{W}}) + \sum_j w^j (0_{\mathbb{V}}, f_j)$ by Equation (5.127). They are independent: a vanishing combination has first slot $\sum_i a^i e_i = 0_{\mathbb{V}}$ and second slot $\sum_j b^j f_j = 0_{\mathbb{W}}$, and each forces its own coefficients to vanish by Equation (5.29). So they form a basis, and Lemma 5.36 makes the count $p+q$ the dimension. □

The construction is also met the other way round, inside a space that is already given: one asks whether $\mathbb{V}$ is assembled from two of its own subspaces.

Definition 5.99 (Sum of subspaces; internal direct sum)**.** Let $\mathbb{U}, \mathbb{W} \leq \mathbb{V}$ be subspaces. Their *sum* is

$$\mathbb{U} + \mathbb{W} = \{u + w \mid u \in \mathbb{U},\ w \in \mathbb{W}\}, \tag{5.129}$$

a subspace of $\mathbb{V}$, since a linear combination of two such elements is again one, the $\mathbb{U}$-parts and the $\mathbb{W}$-parts combining separately. We say that $\mathbb{V}$ is the *internal direct sum* of $\mathbb{U}$ and $\mathbb{W}$, written $\mathbb{V} = \mathbb{U} \oplus \mathbb{W}$, if every $v \in \mathbb{V}$ has exactly one decomposition $v = u + w$ with $u \in \mathbb{U}$ and $w \in \mathbb{W}$; $\mathbb{W}$ is then called a *complement* of $\mathbb{U}$ in $\mathbb{V}$.

Rests on Definitions 5.6 and 5.97.

Proposition 5.100 (Criterion for a direct sum)**.** *For subspaces $\mathbb{U}, \mathbb{W} \leq \mathbb{V}$ the following are equivalent:*

1. *$\mathbb{V} = \mathbb{U} \oplus \mathbb{W}$ in the sense of Definition 5.99;*
2. *$\mathbb{U} + \mathbb{W} = \mathbb{V}$ and $\mathbb{U} \cap \mathbb{W} = \{0_{\mathbb{V}}\}$.*

When they hold, and $\mathbb{V}$ is finite-dimensional, the map $(u, w) \longmapsto u + w$ is an isomorphism from the external direct sum onto $\mathbb{V}$, so that

$$\dim \mathbb{V} = \dim \mathbb{U} + \dim \mathbb{W}. \tag{5.130}$$

Rests on Definitions 5.51 and 5.99 and Proposition 5.98.

Proof. *Derives Proposition 5.100.*

(1) ⇒ (2): existence of a decomposition for every v is $\mathbb{U} + \mathbb{W} = \mathbb{V}$. If $x \in \mathbb{U} \cap \mathbb{W}$ then $x = x + 0_{\mathbb{V}} = 0_{\mathbb{V}} + x$ are two decompositions of the same vector, the first with $\mathbb{U}$-part x and the second with $\mathbb{U}$-part $0_{\mathbb{V}}$; uniqueness forces $x = 0_{\mathbb{V}}$.

(2) ⇒ (1): existence is again the first half. For uniqueness, let $u + w = u' + w'$ with $u, u' \in \mathbb{U}$ and $w, w' \in \mathbb{W}$. Then $u - u' = w' - w$ lies in $\mathbb{U}$ and in $\mathbb{W}$, hence is $0_{\mathbb{V}}$, so $u = u'$ and $w = w'$.

The isomorphism. The map $\Sigma : (u, w) \longmapsto u + w$ is linear by Equation (5.127), surjective by the first half of (2), and injective because $\Sigma(u, w) = 0_{\mathbb{V}}$ gives $u = -w \in \mathbb{U} \cap \mathbb{W} = \{0_{\mathbb{V}}\}$. It is therefore an isomorphism (Definition 5.51), and isomorphic spaces have equal dimension (Remark 5.53), so Equation (5.130) follows from Equation (5.128). □

Proposition 5.101 (Grassmann's formula)**.** *For subspaces $\mathbb{U}, \mathbb{W}$ of a finite-dimensional space,*

$$\dim(\mathbb{U} + \mathbb{W}) = \dim \mathbb{U} + \dim \mathbb{W} - \dim(\mathbb{U} \cap \mathbb{W}). \tag{5.131}$$

Rests on Definitions 5.14 and 5.99 and Lemma 5.36.

Proof. *Derives Proposition 5.101.*

$\mathbb{U} \cap \mathbb{W}$ is a subspace, being defined by membership of two subspaces at once. Let $\{z_1, \dots, z_r\}$ be a basis of it and complete it, by Lemma 5.36, to a basis $\{z_1, \dots, z_r, u_1, \dots, u_p\}$ of $\mathbb{U}$ and to a basis $\{z_1, \dots, z_r, w_1, \dots, w_q\}$ of $\mathbb{W}$. We claim that $\{z_1, \dots, z_r, u_1, \dots, u_p, w_1, \dots, w_q\}$ is a basis of $\mathbb{U} + \mathbb{W}$, which gives Equation (5.131) at once, the count being $r + p + q = (r + p) + (r + q) - r$.

It generates. Every element of $\mathbb{U} + \mathbb{W}$ is a sum of an element of $\mathbb{U}$ and an element of $\mathbb{W}$, each of which expands in its own basis.

It is independent. Suppose

$$\sum_k c^k z_k + \sum_i a^i u_i + \sum_j b^j w_j = 0_{\mathbb{V}} \,.$$

Then $x = \sum_j b^j w_j = -\sum_k c^k z_k - \sum_i a^i u_i$ lies in $\mathbb{W}$, by the left-hand expression, and in $\mathbb{U}$, by the right-hand one; so $x \in \mathbb{U} \cap \mathbb{W}$ and $x = \sum_k d^k z_k$ for some scalars. Comparing the two expressions for x in the basis of $\mathbb{W}$, $\sum_j b^j w_j - \sum_k d^k z_k = 0_{\mathbb{V}}$ forces every $b^j = 0$ by Equation (5.29). The original relation then reduces to a vanishing combination of the basis of $\mathbb{U}$, so the remaining coefficients vanish too. □

Remark 5.102 (Direct sum, tensor product, and one name used for both). The heading this subsubsection carries names two constructions, and the subsubsection that follows it names a third thing with the second of those names. The reader is owed a plain statement of which object is which, because the two behave oppositely in the one respect that matters most.

- The *direct sum* $\mathbb{V} \oplus \mathbb{W}$ is the object defined above. Its elements are *pairs*, its dimension is the *sum* Equation (5.128), and the operators built on it are block diagonal — which is what Equations (5.193) and (5.217) record for representations.

- The *tensor product* $\mathbb{V} \otimes \mathbb{W}$ is the object the next subsubsection builds with the symbol $\otimes$, and calls a "direct product" following the older physics usage that the source inherits. Its elements are not pairs, its basis is the set of *products* $e_i \otimes f_j$ of the two bases, and its dimension is accordingly the *product*

$$\dim(\mathbb{V} \otimes \mathbb{W}) = \dim \mathbb{V} \, \dim \mathbb{W} \,, \tag{5.132}$$

  which is the count of the pairs (i, j) indexing Equation (5.133). It is the space on which a bilinear map becomes linear, as noted after Definition 5.106.

The distinction is not a matter of names. A system built from two others by a direct sum has a state space whose dimension grows additively; a composite quantum system has a state space whose dimension grows multiplicatively, because its states are built on the tensor product (Postulate 77.6), and that single fact is why the dimension of the state space of N particles grows exponentially in N — the starting point of Chapter 83. Where the older literature writes "direct product" for $\otimes$, this treatise keeps the two words apart and uses the symbols to disambiguate.

Direct product of linear transformations

Because an LT itself belongs to a vector space, we can define the direct product of operators.

Consider the linear transformations $A \in \mathcal{L}^1(\mathbb{V}^1)$ and $B \in \mathcal{L}^2(\mathbb{V}^2)$, and the direct-product space $\mathbb{W} = \mathcal{L}^1(\mathbb{V}^1) \otimes \mathcal{L}^2(\mathbb{V}^2)$. Consider further the bases $\{\hat{e}^1_i\}$ and $\{\hat{e}^2_i\}$ of $\mathbb{V}^1$ and $\mathbb{V}^2$. The basis of $\mathbb{W}$ is given by

$$\{\hat{w}_k\} = \{\hat{e}^1_i \otimes \hat{e}^2_j\}, \qquad k = (i, j) \,.$$

Then, to compute the components of the matrix associated with $C = A \otimes B$ in the chosen bases, we have

$$A \otimes B \left| \hat{e}^1_i \otimes \hat{e}^2_j \right\rangle = A^{i'}{}_i B^{j'}{}_j \left| \hat{e}^1_{i'} \otimes \hat{e}^2_{j'} \right\rangle$$

and, in addition,

$$C \left| \hat{w}_k \right\rangle = C^{k'}_k \left| \hat{w}_{k'} \right\rangle .$$

Therefore

$$C^{k'}_k = A^{i'}{}_i B^{j'}{}_j, \qquad k = (i,j),\ k' = (i',j') . \tag{5.133}$$

Union of vector spaces

The intersection of two subspaces is a subspace, as Proposition 5.101 used without comment: a linear combination of vectors lying in both lies in both. The union is a different matter, and the answer is as negative as it can be.

Proposition 5.103 (The union of two subspaces)**.** *Let $\mathbb{U}, \mathbb{W} \leq \mathbb{V}$ be subspaces. Then $\mathbb{U} \cup \mathbb{W}$ is a subspace of $\mathbb{V}$ if and only if $\mathbb{U} \subset \mathbb{W}$ or $\mathbb{W} \subset \mathbb{U}$.*

Rests on Definitions 5.4 and 5.6.

Proof. *Derives Proposition 5.103.*

($\Leftarrow$) If one contains the other the union *is* the larger of the two, which is a subspace by hypothesis.

($\Rightarrow$) Suppose neither containment holds, so that there are vectors $u \in \mathbb{U} \setminus \mathbb{W}$ and $w \in \mathbb{W} \setminus \mathbb{U}$. If $\mathbb{U} \cup \mathbb{W}$ were a subspace it would contain $u + w$, which must then lie in $\mathbb{U}$ or in $\mathbb{W}$. Both are impossible:

- if $u + w \in \mathbb{U}$ then $w = (u + w) - u \in \mathbb{U}$, since $\mathbb{U}$ is closed under linear combinations, contradicting the choice of w;
- if $u + w \in \mathbb{W}$ then $u = (u + w) - w \in \mathbb{W}$, contradicting the choice of u.

Hence $\mathbb{U} \cup \mathbb{W}$ is not a subspace. □

Corollary 5.104 (The sum is the right operation)**.** *For subspaces $\mathbb{U}, \mathbb{W} \leq \mathbb{V}$,*

$$\langle \mathbb{U} \cup \mathbb{W} \rangle = \mathbb{U} + \mathbb{W} , \tag{5.134}$$

so the sum Equation (5.129) is the smallest subspace containing both.

Rests on Proposition 5.103 and Definitions 5.11 and 5.99.

Proof. *Derives Corollary 5.104.*

$\mathbb{U} + \mathbb{W}$ is a subspace containing $\mathbb{U} \cup \mathbb{W}$ — take $w = 0_{\mathbb{V}}$ or $u = 0_{\mathbb{V}}$ in Equation (5.129) — so it contains the smallest such subspace, which is $\langle \mathbb{U} \cup \mathbb{W} \rangle$ by Remark 5.12. Conversely every $u + w$ is a linear combination of two elements of $\mathbb{U} \cup \mathbb{W}$ and therefore lies in $\langle \mathbb{U} \cup \mathbb{W} \rangle$ by Definition 5.11. The two inclusions give Equation (5.134). □

Example 5.105 (Two lines in physical space)**.** Take $\mathbb{V} = \mathbb{R}^3$, a $3 + 0$ instantiation, and let $\mathbb{U}$ and $\mathbb{W}$ be two distinct lines through the origin, say the spans of e_1 and of e_2 in the notation of Equation (5.25). Their union is the pair of coordinate axes, which fails to contain $e_1 + e_2$ and is therefore not a subspace; the smallest subspace containing both is the plane $\mathbb{U} + \mathbb{W} = \langle \{e_1, e_2\} \rangle$, of dimension $1 + 1 - 0 = 2$ by Equation (5.131). The moral is worth keeping: set union is not an operation of linear algebra, and every construction in this chapter that looks as though it combines two subspaces — the decompositions Equation (5.53) and Equation (5.73), the direct sums of representations Equation (5.184) — combines them by sum, not by union.

Rests on Proposition 5.103 and Corollary 5.104.

5.6.3 Bilinear transformations

A linear transformation consumes one vector. The maps that consume two — the inner product of Definition 5.17 is the example already in hand, the metric tensor of Chapter 13 the one physics needs most — are linear in each argument separately, and they behave differently enough from linear transformations under a change of basis that they deserve their own treatment.

Definition 5.106 (Bilinear map)**.** Let $\mathbb{V}, \mathbb{W}, \mathbb{Z}$ be vector spaces over the same field $\mathbb{K}$. A map $B : \mathbb{V} \times \mathbb{W} \longrightarrow \mathbb{Z}$ is *bilinear* if it is linear in each argument when the other is held fixed:

$$B(au + bv, w) = aB(u, w) + bB(v, w)\,, \tag{5.135}$$

$$B(u, aw + bz) = aB(u, w) + bB(u, z)\,, \tag{5.136}$$

for all $a, b \in \mathbb{K}$ and all vectors for which the expressions are defined. If $\mathbb{Z} = \mathbb{K}$ we call B a *bilinear form*; if in addition $\mathbb{W} = \mathbb{V}$ we speak of a bilinear form *on* $\mathbb{V}$.

Rests on Definition 5.35.

A bilinear map is *not* a linear map on the product space: the product $\mathbb{V} \times \mathbb{W}$ is itself a vector space, and a linear map on it would satisfy $B(au, aw) = aB(u, w)$, whereas Equations (5.135) and (5.136) give $B(au, aw) = a^2 B(u, w)$. Bilinearity is instead linearity on the tensor product $\mathbb{V} \otimes \mathbb{W}$ met in Equation (5.133): that is exactly what the tensor product is built to arrange.

Matrix representation of a bilinear form

Let $\mathbb{V}$ have the basis $\{e_i\}_{i=1}^n$ and let B be a bilinear form on $\mathbb{V}$. Define

$$B_{ij} = B(e_i, e_j)\,, \tag{5.137}$$

and call $[B]$ the matrix with these entries — note that *both* indices are down, which is not the placement carried by the matrix of a linear transformation in Equation (5.74), and the difference is about to matter. For arbitrary $u = \sum_i u^i e_i$ and $v = \sum_j v^j e_j$, applying Equation (5.135) and then Equation (5.136),

$$B(u, v) = \sum_{i,j=1}^{n} u^i v^j B_{ij} = \begin{pmatrix} u^1 & \dots \end{pmatrix} [B] \begin{pmatrix} v^1 \\ \vdots \end{pmatrix}\,, \tag{5.138}$$

so the n^2 numbers B_{ij} determine B completely; and they are unique, by the uniqueness of components in a basis used already in Proposition 5.43. Conversely any array defines a bilinear form through Equation (5.138).

Change of basis: congruence, not similarity

Let $\{e'_i\}$ be a second basis, related to the first by

$$e'_i = \sum_{j=1}^{n} A^j{}_i e_j\,,$$

with $[A]$ invertible. Then

$$\begin{aligned} B'_{ik} &= B(e'_i, e'_k) \\ &= \sum_{j,l=1}^{n} A^j{}_i A^l{}_k B(e_j, e_l) && \text{by Equations (5.135) and (5.136)} \\ &= \sum_{j,l=1}^{n} \left(A^{\mathsf{T}}\right)_i{}^j B_{jl} A^l{}_k\,, \end{aligned} \tag{5.139}$$

that is, $[B'] = [A]^{\mathsf{T}}\,[B]\,[A]$. Two matrices related in this way are called *congruent*. Compare Remark 5.122: the matrix of a linear *transformation* changes by the *similarity* $[A]^{-1}\,[T]\,[A]$. The two laws agree only when $[A]^{\mathsf{T}} = [A]^{-1}$, that is, for orthogonal changes of basis — which is the reason a physicist who works only with orthonormal bases can go a long way without noticing that a metric and an operator are different kinds of object, and the reason the distinction reappears the moment curvilinear coordinates are introduced in Chapter 13.

Two immediate consequences of Equation (5.139): the trace of $[B]$ is *not* basis-independent (unlike the character of a representation, Equation (5.181)), while

$$\det\,[B'] = (\det\,[A])^2\,\det\,[B]$$

shows that whether $\det\,[B]$ vanishes, and — over $\mathbb{R}$ — its sign, are basis-independent.

Symmetry, non-degeneracy, and the map into the dual

Definition 5.107 (Symmetric and antisymmetric forms)**.** A bilinear form on $\mathbb{V}$ is *symmetric* if $B(u,v) = B(v,u)$ for all u, v, and *antisymmetric* (or *alternating*) if $B(u,v) = -B(v,u)$. In terms of Equation (5.137) these read $[B]^{\mathsf{T}} = [B]$ and $[B]^{\mathsf{T}} = -\,[B]$, and both conditions are preserved by Equation (5.139).

Rests on Definition 5.106 and Equation (5.137).

Proposition 5.108 (Symmetric–antisymmetric splitting)**.** *Let $\mathbb{V}$ be a vector space over a field of characteristic other than 2 — $\mathbb{R}$ and $\mathbb{C}$ are such fields. Every bilinear form B on $\mathbb{V}$ is the sum of a symmetric and an antisymmetric form, namely*

$$B = B_{\mathrm{s}} + B_{\mathrm{a}}\,, \qquad B_{\mathrm{s}}(u,v) = \tfrac{1}{2}\,(B(u,v) + B(v,u))\,, \quad B_{\mathrm{a}}(u,v) = \tfrac{1}{2}\,(B(u,v) - B(v,u))\,, \tag{5.140}$$

and this decomposition is unique.

Rests on Definitions 5.106 and 5.107.

Derivation. *Derives Proposition 5.108.*

Existence. Both B_{s} and B_{a} are bilinear, being linear combinations of the bilinear maps $(u,v) \longmapsto B(u,v)$ and $(u,v) \longmapsto B(v,u)$; exchanging u and v in Equation (5.140) leaves the first unchanged and reverses the sign of the second, so they are respectively symmetric and antisymmetric in the sense of Definition 5.107; and adding them returns B. The halving is where the characteristic hypothesis is spent here — $\frac{1}{2}$ must exist in the field — and it is spent again, in the other direction, in the next paragraph.

Uniqueness. Suppose $B = S + A = S' + A'$ with S, S' symmetric and A, A' antisymmetric, and put $C = S - S' = A' - A$. Then C is symmetric, being a difference of symmetric forms, and antisymmetric, being a difference of antisymmetric ones, so for all u, v

$$C(u,v) = C(v,u) = -C(u,v)\,, \qquad \text{hence} \qquad 2C(u,v) = 0\,.$$

Since $2 \neq 0$ in the field, $C = 0$, that is $S = S'$ and $A = A'$. In matrix terms Equation (5.140) is the familiar $[B] = \frac{1}{2}\left([B] + [B]^{\mathsf{T}}\right) + \frac{1}{2}\left([B] - [B]^{\mathsf{T}}\right)$, and the uniqueness argument says that the only array equal to both its own transpose and minus it is the zero array. □

Physics uses both halves and keeps them apart: the symmetric one is a metric (Chapter 13), the antisymmetric one a symplectic form or a field strength.

Definition 5.109 (Non-degenerate form)**.** A bilinear form B on $\mathbb{V}$ is *non-degenerate* if $B(u,v) = 0$ for all $v \in \mathbb{V}$ implies $u = 0_{\mathbb{V}}$, and $B(u,v) = 0$ for all $u \in \mathbb{V}$ implies $v = 0_{\mathbb{V}}$.

Rests on Definition 5.106.

Proposition 5.110. *Let* $\dim \mathbb{V} = n < \infty$ *and let* B *be a bilinear form on* $\mathbb{V}$*. Define*

$$\begin{array}{rccc} \flat_B : & \mathbb{V} & \longrightarrow & \mathbb{V}^* \\ & u & \longmapsto & B(u,\ ) \,. \end{array} \tag{5.141}$$

Then $\flat_B$ *is a linear transformation, and the following are equivalent:*

1. B *is non-degenerate;*
2. $\flat_B$ *is an isomorphism;*
3. $\det[B] \neq 0$ *in any (hence every) basis.*

Rests on Definitions 5.56, 5.106 and 5.109 and Equations (5.86) and (5.87).

Proof. *Derives Proposition 5.110.*

$\flat_B(u)$ is a functional by Equation (5.136), and $u \longmapsto \flat_B(u)$ is linear by Equation (5.135), so $\flat_B \in \mathcal{L}(\mathbb{V}, \mathbb{V}^*)$.

(1) $\Rightarrow$ (2): the first half of Definition 5.109 says exactly that $\ker \flat_B = \{0_{\mathbb{V}}\}$, so $\flat_B$ is injective; and an injective linear map between spaces of the same finite dimension — $\dim \mathbb{V}^* = \dim \mathbb{V} = n$ was established after Equation (5.87) — is surjective.

(2) $\Rightarrow$ (3): in the basis $\{e_i\}$ and its dual basis $\{f^j\}$ of Equation (5.86),

$$\flat_B(e_i) = B(e_i,\ ) = \sum_{j=1}^{n} B_{ij} f^j \,,$$

because both sides agree on every e_k: the left gives B_{ik} and the right gives $\sum_j B_{ij}\delta^j{}_k = B_{ik}$. So the matrix of $\flat_B$ in these bases is $[B]^{\mathsf{T}}$; a linear map between spaces of equal finite dimension is bijective exactly when its matrix is invertible, that is, when its determinant does not vanish, and $\det\left([B]^{\mathsf{T}}\right) = \det[B]$.

(3) $\Rightarrow$ (1): if $B(u, v) = 0$ for all v then in components $\sum_i u^i B_{ij} = 0$ for every j, a homogeneous linear system whose matrix is invertible, so $u^i = 0$ for all i; the other half is the same argument on the other index. Independence of the basis follows from $\det[B'] = (\det[A])^2 \det[B]$ with $\det[A] \neq 0$. □

Proposition 5.110 is the precise sense in which a non-degenerate form "identifies a space with its dual", and it is the algebraic content of raising and lowering indices with a metric. It also locates exactly what Proposition 5.59 adds: an inner product is a non-degenerate form with two further properties, hermiticity and positivity, and it is those — not non-degeneracy — that fail in the counterexamples of Remark 5.60.

Sesquilinear and hermitian forms

Over $\mathbb{C}$ the useful notion is not quite bilinearity.

Definition 5.111 (Sesquilinear form)**.** A map $H : \mathbb{V} \times \mathbb{V} \longrightarrow \mathbb{C}$ is *sesquilinear* — "one and a half times linear" — if it is antilinear in the first argument and linear in the second:

$$H(au + bv, w) = a^* H(u, w) + b^* H(v, w) \,, \qquad H(u, aw + bz) = aH(u, w) + bH(u, z) \,.$$

It is *hermitian* if $H(u, v) = H(v, u)^*$ for all u, v, which in components means $[H]^\dagger = [H]$, and *positive definite* if $H(v, v) > 0$ for every $v \neq 0_{\mathbb{V}}$.

Rests on Definition 4.31 and Equation (5.137).

Comparing with Definition 5.17: *an inner product on a complex vector space is precisely a positive-definite hermitian sesquilinear form*, and on a real vector space precisely a positive-definite symmetric bilinear form. Everything proved in Section 5.3 is therefore a statement about a very particular corner of the present section, and the Schwarz inequality Equation (5.43) is exactly what positive definiteness buys and what an indefinite form loses. Note also that $H(v,v)$ is automatically real for a hermitian form, by the same one-line argument as Equation (5.39), so the inequality $H(v,v) > 0$ is meaningful.

The change-of-basis law for a sesquilinear form carries a conjugate: repeating the computation of Equation (5.139) with the first slot antilinear gives $[H'] = [A]^\dagger\,[H]\,[A]$, which is why the unitary group $\mathrm{U}(n,\mathbb{C})$ of Equation (5.171) — and not the orthogonal group — is the group preserving a complex inner product.

Sylvester's law of inertia

For a real symmetric form the congruence Equation (5.139) can always be used to reach a normal form, and what survives the choice of basis is a pair of integers. This is what makes the "signature (p,q)" spoken of throughout the treatise a well-defined attribute of a form rather than of a basis.

Theorem 5.112 (Sylvester's law of inertia)**.** *Let B be a symmetric bilinear form on a real vector space $\mathbb{V}$ of finite dimension n. Then there is a basis in which*

$$[B] = \mathrm{diag}(\underbrace{1,\dots,1}_{p},\underbrace{-1,\dots,-1}_{q},\underbrace{0,\dots,0}_{r})\,, \qquad p+q+r=n\,, \tag{5.142}$$

and the triple (p,q,r) is the same for every such basis.

Rests on Definitions 5.106 and 5.107 and Equation (5.139).

Proof. *Derives Theorem 5.112.*

Existence. By induction on n. If $B=0$ then $p=q=0$, $r=n$ and any basis will do. Otherwise there is a vector v with $B(v,v)\neq 0$: were $B(u,u)=0$ for every u, polarization would give

$$B(u,w) = \tfrac{1}{2}\left(B(u+w,u+w) - B(u,u) - B(w,w)\right) = 0$$

for all u,w, using symmetry, and B would vanish identically. Set $e_1 = v/\sqrt{|B(v,v)|}$, so that $B(e_1,e_1)=\pm 1$, and let

$$\mathbb{V}_1 = \{u\in\mathbb{V} \mid B(e_1,u)=0\}\,,$$

which is a subspace, being defined by a linear condition. Every $u\in\mathbb{V}$ decomposes as

$$u = \frac{B(e_1,u)}{B(e_1,e_1)}e_1 + \left(u - \frac{B(e_1,u)}{B(e_1,e_1)}e_1\right)\,,$$

whose second term lies in $\mathbb{V}_1$, as substituting it into $B(e_1,\ )$ shows; and $\langle\{e_1\}\rangle\cap\mathbb{V}_1 = \{0_\mathbb{V}\}$, since $B(e_1,\lambda e_1)=\lambda B(e_1,e_1)=0$ forces $\lambda=0$. Hence $\mathbb{V}=\langle\{e_1\}\rangle\oplus\mathbb{V}_1$ and $\dim\mathbb{V}_1 = n-1$. Applying the induction hypothesis to the restriction of B to $\mathbb{V}_1$ — which is again symmetric — and reordering the resulting basis gives Equation (5.142).

Uniqueness. r is basis-independent because $n-r$ is the rank of $[B]$, and congruent matrices have equal rank ($[A]$ being invertible). For p, suppose two bases give (p,q,r) and (p',q',r) with, say, $p>p'$. Let $\mathbb{P}$ be the span of the p basis vectors of the first basis on which B is $+1$: on $\mathbb{P}\setminus\{0_\mathbb{V}\}$ the form is strictly positive, since $B\left(\sum_{i\le p}a^i e_i, \sum_{j\le p}a^j e_j\right) = \sum_{i\le p}(a^i)^2 > 0$. Let $\mathbb{M}'$ be the span of the $q'+r = n-p'$ vectors of the second basis on which B is -1 or 0: on $\mathbb{M}'$ the form satisfies $B(u,u)\le 0$, by the same computation with the signs of Equation (5.142). Now

$$\dim\mathbb{P} + \dim\mathbb{M}' = p + (n-p') > n\,,$$

so $\mathbb{P}\cap\mathbb{M}'$ contains a non-zero vector u — two subspaces whose dimensions sum to more than n cannot meet only in $0_{\mathbb{V}}$, since $\dim(\mathbb{P}+\mathbb{M}')\le n$. That vector would satisfy $B(u,u)>0$ and $B(u,u)\le 0$ at once. Hence $p\le p'$, and by symmetry $p'\le p$; with r fixed, $q=n-p-r$ is determined too. □

Definition 5.113 (Signature)**.** The pair (p,q) of Theorem 5.112 is the *signature* of the symmetric form. A non-degenerate form has $r=0$ and $p+q=n$; it is *positive definite*, that is, an inner product, exactly when $q=r=0$.

Rests on Theorem 5.112 and Definition 5.109.

Example 5.114 (The two forms physics uses)**.** Both instantiations below are the observed ones, $3+0$ and $3+1$; the treatise takes no other (editorial rule on dimensions).

- *Euclidean space,* $3+0$. On $\mathbb{R}^3$ with the standard basis, $[B]=\operatorname{diag}(1,1,1)$: signature $(3,0)$, non-degenerate and positive definite, hence an inner product, and the source of the length $\|\boldsymbol{x}\|=\sqrt{x^2+y^2+z^2}$ in metres. The subgroup of $\mathrm{GL}(3,\mathbb{R})$ preserving it is $\mathrm{O}(3,\mathbb{R})$ of Equation (5.161), the rotations and reflections.

- *Minkowski spacetime,* $3+1$. On $\mathbb{R}^4$ with coordinates $(x^0,x^1,x^2,x^3)=(ct,x,y,z)$ — all four carrying the SI unit metre, which is exactly what the factor c is there for —

  $$[\eta]=\operatorname{diag}(-1,1,1,1)\,,$$

  of signature $(3,1)$: non-degenerate, and *not* positive definite, so it is not an inner product and Proposition 5.19 does not apply to it. The quantity $\eta(u,u)$ carries the SI unit m^2. The subgroup of $\mathrm{GL}(4,\mathbb{R})$ preserving it is the Lorentz group $\mathrm{O}(3,1)$ of Chapter 39. Theorem 5.112 is what guarantees that no change of basis whatever can turn this form into the Euclidean one: the signature is an invariant, so the distinction between a timelike and a spacelike direction is not a coordinate artefact.

Proposition 5.115 (Normal form of a non-degenerate antisymmetric form)**.** *Let ω be a non-degenerate antisymmetric bilinear form on a real vector space $\mathbb{V}$ of finite dimension n. Then $n=2m$ is even and $\mathbb{V}$ has a basis $\{e_1,\dots,e_m,f_1,\dots,f_m\}$ in which*

$$\omega(e_i,e_j)=\omega(f_i,f_j)=0\,,\qquad \omega(e_i,f_j)=\delta_{ij}\,,\qquad i,j=1,\dots,m\,, \tag{5.143}$$

that is, in which the matrix of the form is the block array

$$[\omega]=\begin{pmatrix}0 & \mathbb{1}_m\\ -\mathbb{1}_m & 0\end{pmatrix}. \tag{5.144}$$

Consequently any two non-degenerate antisymmetric forms on real spaces of the same even dimension are congruent in the sense of Equation (5.139): there is no invariant here answering to the signature of Theorem 5.112.

Rests on Definitions 5.106, 5.107 and 5.109 and Equation (5.139).

Derivation. Derives Proposition 5.115.

Antisymmetry first gives $\omega(u,u)=-\omega(u,u)$, hence $\omega(u,u)=0$ for every u, the field being $\mathbb{R}$. The construction is then an induction on n — a *symplectic* Gram–Schmidt, in which vectors are paired off instead of being normalized one at a time.

If $\mathbb{V}=\{0_{\mathbb{V}}\}$ there is nothing to prove and $m=0$. Otherwise choose any $e_1\neq 0_{\mathbb{V}}$. Non-degeneracy (Definition 5.109) supplies a w with $\omega(e_1,w)\neq 0$, and

$$f_1=\frac{w}{\omega(e_1,w)}\qquad\text{gives}\qquad \omega(e_1,f_1)=1\,.$$

These two are linearly independent: were $f_1 = \lambda e_1$ we should have $\omega(e_1, f_1) = \lambda\,\omega(e_1, e_1) = 0$. Put $W = \langle\{e_1, f_1\}\rangle$ and

$$W^\omega = \{v \in \mathbb{V} \mid \omega(v, e_1) = \omega(v, f_1) = 0\}\ ,$$

a subspace, being the intersection of the kernels of two functionals.

The two split $\mathbb{V}$. Given $v \in \mathbb{V}$, set

$$v' = v - \omega(v, f_1)\, e_1 + \omega(v, e_1)\, f_1\,.$$

Using $\omega(e_1, e_1) = \omega(f_1, f_1) = 0$ and $\omega(f_1, e_1) = -1$,

$$\begin{aligned}\omega(v', e_1) &= \omega(v, e_1) + \omega(v, e_1)\,\omega(f_1, e_1) = 0\,,\\ \omega(v', f_1) &= \omega(v, f_1) - \omega(v, f_1)\,\omega(e_1, f_1) = 0\,,\end{aligned}$$

so $v' \in W^\omega$ and $v = (\omega(v, f_1)e_1 - \omega(v, e_1)f_1) + v' \in W + W^\omega$. The sum is direct: if $v = ae_1 + bf_1$ lies in W^ω then $0 = \omega(v, f_1) = a$ and $0 = \omega(v, e_1) = -b$. Hence $\mathbb{V} = W \oplus W^\omega$ and $\dim W^\omega = n - 2$.

The restriction is again non-degenerate. Let $v \in W^\omega$ satisfy $\omega(v, u) = 0$ for every $u \in W^\omega$. It also annihilates e_1 and f_1, hence the whole of W; by the direct sum $\omega(v,\ )$ vanishes on $\mathbb{V}$, and $v = 0_{\mathbb{V}}$ by non-degeneracy of ω.

The induction hypothesis applied to W^ω therefore makes $n - 2$ even — so n is even, $n = 2m$ — and supplies vectors $e_2, \dots, e_m, f_2, \dots, f_m$ of W^ω satisfying Equation (5.143) among themselves. All products of e_1 or f_1 with one of them vanish, since those vectors lie in W^ω, so the whole collection satisfies Equation (5.143) and, being a basis of W together with a basis of W^ω, is a basis of $\mathbb{V}$. Listing the e_i first and the f_i second turns Equation (5.143) into Equation (5.144). That two such forms are congruent now follows because each is carried to Equation (5.144) by a change of basis, and Equation (5.139) composes. □

Remark 5.116 (Antisymmetric forms and the symplectic group). Proposition 5.115 is the antisymmetric counterpart of Sylvester's law, and the contrast is sharp: a symmetric form keeps its signature, an antisymmetric one keeps nothing but its dimension. That even dimension is forced can also be seen without the construction, by taking determinants in $[\omega]^{\mathsf{T}} = -[\omega]$:

$$\det[\omega] = \det\left([\omega]^{\mathsf{T}}\right) = \det(-[\omega]) = (-1)^n \det[\omega]\ ,$$

which for odd n forces $\det[\omega] = 0$, excluded by non-degeneracy (Proposition 5.110). The subgroup of $\mathrm{GL}(2m, \mathbb{R})$ preserving ω is the symplectic group $\mathrm{Sp}(2m, \mathbb{R})$, constructed with the other classical groups in Definition 5.130; every one of its elements has determinant $+1$ (Theorem 5.133), so the antisymmetric case admits no analogue of the split between $\mathrm{O}(n, \mathbb{K})$ and $\mathrm{SO}(n, \mathbb{K})$.

The absence of a signature is a statement of linear algebra and is proved above. Its celebrated geometric counterpart — that a symplectic *manifold* has no local invariants at all, where a Riemannian one has curvature — is Darboux's theorem, a statement about a form varying from point to point rather than about one form on one vector space; it does not follow from the proposition, which only says that the forms at each single point look alike. What the proposition supplies is precisely that pointwise input: at each single point the form can be brought to Equation (5.144), and the work left to Darboux's theorem is to show that the bases achieving it at neighbouring points can be chosen to fit together into a single chart. That is where a condition with no counterpart in linear algebra enters — the form must be closed — and it is proved in the appendix, in the language of the differential forms of Chapter 13.

Full derivation: Appendix A.12 (p. 2807).

Derives Remark 5.116.

The instance physics meets is the phase space of a mechanical system with N degrees of freedom, of dimension $2N$, on which ω pairs a coordinate against its conjugate momentum; in SI the value $\omega(u, v)$ then carries the unit of action, J s, the same unit as $\hbar$ — which is the first hint of why the quantization of Chapter 77 is expressed in units of $\hbar$.

5.7 Algebras

5.7.1 Algebras

Let us discuss the definition of an algebra here in more detail. The axioms that the elements of an algebra must satisfy are many, but they are summarized as follows. Let $((\mathbb{K}, +, \cdot), \cdot, (A, +))$ be a vector space, and consider the IBO

$$\begin{array}{rccc} [\ ,\ ]: & A \times A & \longrightarrow & A \\ & (X, Y) & \longmapsto & [X, Y]\ . \end{array}$$

Definition 5.117 (Algebra)**.** The space A together with the IBO $[\ ,\ ]$ is an *algebra* if and only if the following linearity axioms hold:

1. Left distributivity:

$$\forall\ X, Y, Z \in A, \quad [X + Y, Z] = [X, Z] + [Y, Z]\ . \tag{5.145}$$

2. Right distributivity:

$$\forall\ X, Y, Z \in A, \quad [X, Y + Z] = [X, Y] + [X, Z]\ . \tag{5.146}$$

3. Product by scalars:

$$\forall\ X, Y \in A,\ \forall\, \alpha, \beta \in \mathbb{K}, \quad [\alpha X, \beta Y] = \alpha\beta\, [X, Y]\ . \tag{5.147}$$

Rests on Definitions 4.4 and 4.31.

When no ambiguity is possible we shall refer to an algebra simply as A, unless it is necessary to specify the field and the internal binary operations accompanying it.

Definition 5.118 (Linear hull)**.** Consider $M, N \subset A$. We define the set $[M, N]$ as

$$[M, N] = \left\{ \sum_i \alpha_i\, [X_i, Y_i] \mid X_i \in M,\ Y_i \in N,\ \alpha_i \in \mathbb{K} \right\}\ . \tag{5.148}$$

Rests on Definition 5.117.

Definition 5.119 (Subalgebra)**.** Consider an algebra A and $N \subset A$. We say that N is a *subalgebra* of A if and only if all possible linear combinations of brackets of vectors in N remain in N, that is, if

$$\left\{ \sum_i \alpha_i\, [X_i, Y_i] \mid X_i, Y_i \in N \right\} \subset N\,,$$

hence

$$N \text{ is a subalgebra of } A \iff [N, N] \subset N\,.$$

Rests on Definitions 5.117 and 5.118.

Definition 5.120 (Ideal)**.** We say that a subset N of an algebra A is an *ideal* of A if and only if

$$[N, A] \subset N\,.$$

Rests on Definitions 5.117 and 5.118.

Structure constants

Let $A = \langle \{e_i\}^n \rangle$ be an algebra and $X, Y \in A$. We can expand X and Y as linear combinations of the elements of the chosen basis. We have

$$\begin{aligned}[X,Y] &= \left[X^i e_i, Y^j e_j\right] \\ &= X^i Y^j \left[e_i, e_j\right] \qquad \text{by Equation (5.147)}\,. \end{aligned} \tag{5.149}$$

Now, since $[\ ,\ ]$ is an IBO — that is, closure holds — we can write $[e_i, e_j]$ as a linear combination of the elements of the basis. That is,

$$[e_i, e_j] = C^k_{\ ij} e_k\,, \tag{5.150}$$

where the quantities $C^k_{\ ij} \in \mathbb{K}$ are called the *structure constants* of the algebra. It should be mentioned that the structure constants are basis-dependent.

On the basis of the structure constants we can compute the coefficients of the linear-combination expansion of an element $[X, Y]$ of the algebra. Substituting Equation (5.150) into Equation (5.149),

$$[X,Y] = X^i Y^j C^k_{\ ij} e_k = [X,Y]^k\, e_k\,,$$

hence

$$[X,Y]^k = X^i Y^j C^k_{\ ij}\,. \tag{5.151}$$

From Equations (5.149) and (5.150) (expand both arguments in the basis and replace the bracket of basis elements by the structure constants).

5.7.2 The algebra of linear transformations

Proposition 5.121. *Let $\mathbb{V}$ be a vector space of finite dimension n with a fixed basis $B = \{e_i\}_{i=1}^n$, and let $\mathcal{L}(\mathbb{V}) = \mathcal{L}(\mathbb{V}, \mathbb{V})$ carry the commutator bracket $[S,T] = S \circ T - T \circ S$. Then the map $T \longmapsto [T]$ that assigns to an endomorphism its matrix representation in B satisfies*

$$[\alpha S + \beta T] = \alpha\,[S] + \beta\,[T]\ , \tag{5.152}$$

$$[S \circ T] = [S]\,[T]\ , \tag{5.153}$$

$$[[S,T]] = [[S]\,, [T]]\ , \tag{5.154}$$

and is a bijection onto $\mathcal{M}_{n\times n}[\mathbb{K}]$. Consequently any subalgebra of $\mathcal{L}(\mathbb{V})$ and its image in $\mathcal{M}_{n\times n}[\mathbb{K}]$ have the same structure constants: the matrices satisfy the same algebra as the transformations they represent.

Rests on Definitions 5.35 and 5.117, Proposition 5.43, and Equations (5.74) and (5.150).

Derivation. *Derives Proposition 5.121.*

Every identity is read off the definition Equation (5.74) of the matrix representation, here with $\mathbb{W} = \mathbb{V}$ and $B_2 = B_1 = B$, so that $T(e_i) = \sum_j T^j_{\ i} e_j$.

Linearity, Equation (5.152). By the definition of the linear combination of two LT's,

$$\begin{aligned}(\alpha S + \beta T)(e_i) &= \alpha S(e_i) + \beta T(e_i) \\ &= \sum_{j=1}^{n} \left(\alpha S^j_{\ i} + \beta T^j_{\ i}\right) e_j\,, \end{aligned}$$

and by the uniqueness established in Proposition 5.43 the coefficients displayed are *the* matrix entries of $\alpha S + \beta T$.

Composition, Equation (5.153). Apply the two transformations in turn:

$$\begin{aligned}(S\circ T)(e_i) &= S\left(T(e_i)\right)\\ &= S\left(\sum_{j=1}^{n} T^{j}{}_{i}\, e_j\right)\\ &= \sum_{j=1}^{n} T^{j}{}_{i}\, S(e_j) && \text{by linearity of } S\\ &= \sum_{k=1}^{n}\left(\sum_{j=1}^{n} S^{k}{}_{j} T^{j}{}_{i}\right) e_k\,,\end{aligned}$$

so that $(S\circ T)^{k}{}_{i} = S^{k}{}_{j} T^{j}{}_{i}$ in the summation convention — precisely the row-by-column rule for the matrix product $[S]\,[T]$. Note that the order is preserved: the transformation applied first stands to the right in both notations.

Bracket, Equation (5.154). Combine the two:

$$\begin{aligned}[[S,T]] &= [S\circ T - T\circ S]\\ &= [S\circ T] - [T\circ S] && \text{by Equation (5.152)}\\ &= [S]\,[T] - [T]\,[S] && \text{by Equation (5.153)}\\ &= [[S]\,,[T]]\ .\end{aligned}$$

Bijectivity is Proposition 5.43: distinct endomorphisms have distinct matrices, and every array arises from the endomorphism it defines through Equation (5.74).

Structure constants. Let $A\subset\mathcal{L}(\mathbb{V})$ be a subalgebra with basis $\{X_a\}$ and structure constants defined by Equation (5.150), $[X_a,X_b] = C^{c}{}_{ab} X_c$. Taking matrix representations of both sides and using Equations (5.152) and (5.154),

$$[[X_a]\,,[X_b]] = C^{c}{}_{ab}\,[X_c]\ ,$$

with the *same* coefficients $C^{c}{}_{ab}$. Moreover the $[X_a]$ are linearly independent, since $T\longmapsto[T]$ is injective and linear, so they are a basis of the image and the two algebras have the same structure constants in corresponding bases. Any identity following from the axioms — the antisymmetry Equation (5.157) and the Jacobi identity Equation (5.158) of a Lie algebra, for instance — therefore holds for the matrices exactly when it holds for the operators. □

Remark 5.122 (The basis is fixed, and that matters). Equation (5.153) is false if S and T are referred to different bases: it is the *same* B appearing three times that lets the inner index j be summed. Changing B replaces every $[T]$ by $[S_0]\,[T]\,[S_0]^{-1}$ for one fixed invertible $[S_0]$, and all three identities are invariant under that simultaneous conjugation — which is why the structure constants are carried along unchanged, and why Definition 5.140 declares two representations related in this way to be equivalent.

5.7.3 Lie algebras

Definition 5.123 (Lie algebra)**.** An algebra $((\mathbb{K},+,\cdot)\,,\cdot,(L,+,[\ ,\ ]))$ in the sense of Definition 5.117 — equivalently, of Definition 4.32, whose axioms are the bilinearity of the bracket and nothing more — is a *Lie algebra* if and only if the following two further axioms are satisfied:

1. Antisymmetry:
$$\forall\ X,Y\in L,\quad [X,Y] = -\,[Y,X]\ . \tag{5.155}$$

2. Jacobi identity:

$$\forall\ X, Y, Z \in L, \quad [X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0_L\,. \tag{5.156}$$

Rests on Definition 5.117.

In the context of Lie algebras we shall name the IBO [,] the *Lie bracket*.

Properties of the structure constants

As a direct consequence of the axioms of a Lie algebra, there are two important properties satisfied by the structure constants. Let $L = \langle \{e_i\}^n \rangle$ be a Lie algebra.

1. Antisymmetry. From axiom Equation (5.155),

$$\begin{aligned} [e_i, e_j] &= -\,[e_j, e_i] \\ C^k{}_{ij} e_k &= -C^k{}_{ji} e_k \qquad\qquad \text{by Equation (5.150)}\,, \end{aligned}$$

hence

$$C^k{}_{ij} = -C^k{}_{ji}\,. \tag{5.157}$$

From Equations (5.150) and (5.155) (antisymmetry of the Lie bracket, written on a pair of basis elements).

2. Jacobi identity. From axiom Equation (5.156),

$$\begin{aligned} [e_i, [e_j, e_k]] + [e_j, [e_k, e_i]] + [e_k, [e_i, e_j]] &= 0_L \\ \left[e_i, C^l{}_{jk} e_l\right] + \left[e_j, C^l{}_{ki} e_l\right] + \left[e_k, C^l{}_{ij} e_l\right] &= 0_L \\ C^l{}_{jk}\,[e_i, e_l] + C^l{}_{ki}\,[e_j, e_l] + C^l{}_{ij}\,[e_k, e_l] &= 0_L \\ C^l{}_{jk} C^m{}_{il} e_m + C^l{}_{ki} C^m{}_{jl} e_m + C^l{}_{ij} C^m{}_{kl} e_m &= 0_L \\ (C^l{}_{jk} C^m{}_{il} + C^l{}_{ki} C^m{}_{jl} + C^l{}_{ij} C^m{}_{kl}) e_m &= 0_L\,, \end{aligned}$$

so, writing 0_L as

$$0_L = \sum_{m=1}^{\dim L} 0 \cdot e_m, \quad 0 \in \mathbb{K}\,,$$

we obtain

$$C^l{}_{jk} C^m{}_{il} + C^l{}_{ki} C^m{}_{jl} + C^l{}_{ij} C^m{}_{kl} = 0\,. \tag{5.158}$$

From Equations (5.150) and (5.156) (the Jacobi identity on basis elements, with the inner and outer brackets expanded in turn).

5.7.4 Derivations

Definition 5.124 (Derivation)**.** Consider an algebra L and let $D : L \longrightarrow L$ be a map. We say that D is a *derivation* if it satisfies the following properties:

1. Linearity:

$$D(\alpha X + \beta Y) = \alpha DX + \beta DY\,,$$

that is, D is a linear transformation.

2. Leibniz rule:

$$D\,[X, Y] = [DX, Y] + [X, DY]\,.$$

Rests on Definitions 5.35 and 5.117.

The algebra of derivations

Derivations satisfy the following properties. Let D_1 and D_2 be derivations.

1. *Linear combination of derivations.* The linear combination of derivations is also a derivation. Indeed, the linear combination of linear transformations is again a linear transformation; moreover, the Leibniz rule holds:

$$\begin{aligned}(\alpha D_1 + \beta D_2)\,[X,Y] &= \alpha D_1\,[X,Y] + \beta D_2\,[X,Y] \\ &= \alpha([D_1X,Y] + [X,D_1Y]) + \beta([D_2X,Y] + [X,D_2Y]) \\ &= [\alpha D_1X,Y] + [\beta D_2X,Y] + [X,\alpha D_1Y] + [X,\beta D_2Y] \\ &= [(\alpha D_1 + \beta D_2)X,Y] + [X,(\alpha D_1 + \beta D_2)Y]\ .\end{aligned}$$

2. *Commutator of two derivations.* The derivation of a derivation is, in general, not a derivation. Indeed, applying D_2 to the Leibniz rule for D_1,

$$\begin{aligned}D_1\,[X,Y] &= [D_1X,Y] + [X,D_1Y] \\ D_2D_1\,[X,Y] &= [D_2D_1X,Y] + [D_1X,D_2Y] + [D_2X,D_1Y] + [X,D_2D_1Y] \\ &= [D_2D_1X,Y] + [X,D_2D_1Y] + \underbrace{[D_1X,D_2Y] + [D_2X,D_1Y]}_{\text{in general } \neq 0_L}\ . \end{aligned} \tag{5.159}$$

   Analogously to Equation (5.159) we have

$$D_1D_2\,[X,Y] = [D_1D_2X,Y] + [X,D_1D_2Y] + [D_2X,D_1Y] + [D_1X,D_2Y]\ . \tag{5.160}$$

   Now, subtracting Equation (5.159) from Equation (5.160),

$$\begin{aligned}(D_1D_2 - D_2D_1)\,[X,Y] &= [D_1D_2X,Y] + [X,D_1D_2Y] - [D_2D_1X,Y] - [X,D_2D_1Y] \\ &= [(D_1D_2 - D_2D_1)X,Y] + [X,(D_1D_2 - D_2D_1)Y]\ ,\end{aligned}$$

   hence

$$[D_1,D_2]\,[X,Y] = [[D_1,D_2]\,X,Y] + [X,[D_1,D_2]\,Y]\ ,$$

   so the commutator of two derivations is again a derivation.

On the basis of the two properties above we may conclude that the set of all derivations definable on L, together with the binary operation (internal, thanks to the closure proved in the second property) defined by $[D_1,D_2] = D_1D_2 - D_2D_1$, forms an algebra, which we shall call the *algebra of derivations* and denote L_A.

5.8 Representations of groups

5.8.1 The classical groups

The orthogonal group

There is an important subgroup of the general linear group $\mathrm{GL}(n,\mathbb{K})$ obtained by adding the property that every automorphism T satisfy $[T][T]^{\mathsf{T}} = \mathbb{1}$. Indeed, let A and B lie in this set, so that $[A][A]^{\mathsf{T}} = \mathbb{1}$ and $[B][B]^{\mathsf{T}} = \mathbb{1}$. By closure of $\mathrm{GL}(n,\mathbb{K})$ there exists C such that $AB = C$. We must prove that $[C][C]^{\mathsf{T}} = \mathbb{1}$. We have

$$\begin{aligned}[C][C]^{\mathsf{T}} &= [A][B]([A][B])^{\mathsf{T}} \\ &= [A][B][B]^{\mathsf{T}}[A]^{\mathsf{T}} \\ &= [A]\,\mathbb{1}\,[A]^{\mathsf{T}} \\ &= \mathbb{1}\,,\end{aligned}$$

so this set possesses closure. We call this subgroup the *orthogonal group* and symbolize it $\mathrm{O}(n,\mathbb{K})$; thus

$$\mathrm{O}(n,\mathbb{K}) = \left\{ T \in \mathrm{GL}(n,\mathbb{K}) \mid [T][T]^{\mathsf{T}} = \mathbb{1} \right\} . \tag{5.161}$$

Note that, taking determinants,

$$\begin{aligned} [T][T]^{\mathsf{T}} &= \mathbb{1} \\ \det\left([T][T]^{\mathsf{T}} \right) &= 1 \\ \det[T]\, \det[T]^{\mathsf{T}} &= 1 \\ \det[T]\, \det[T] &= 1 \\ (\det[T])^2 &= 1 , \end{aligned}$$

so that

$$\det[T] = \pm 1 .$$

As we saw for the special linear group, the condition $\det[T] = 1$ possesses closure, which in general does not happen for the condition $\det[T] = -1$. We therefore define the *special orthogonal group* as

$$\mathrm{SO}(n,\mathbb{K}) = \{ T \in \mathrm{O}(n,\mathbb{K}) \mid \det[T] = 1 \} . \tag{5.162}$$

Proposition 5.125 (An orthogonal transformation is an isometry)**.** *Let $\mathbb{R}^n$ carry the standard inner product $\langle u|v\rangle = u^{\mathsf{T}} v$. A real array T of order n lies in $\mathrm{O}(n,\mathbb{R})$ if and only if*

$$\langle Tu|Tv\rangle = \langle u|v\rangle \qquad \text{for all } u, v \in \mathbb{R}^n , \tag{5.163}$$

and then in particular $\|Tu\| = \|u\|$: an orthogonal transformation preserves lengths, and with them the angle Equation (5.18) between any two non-zero vectors.

Rests on Definition 5.17, Proposition 5.47, and Equation (5.44).

Proof. *Derives Proposition 5.125.*

Equation (5.161) says $TT^{\mathsf{T}} = \mathbb{1}$; in finite dimension a one-sided inverse is the inverse (Proposition 5.47), so $T^{\mathsf{T}}T = \mathbb{1}$ as well, and the two conditions are interchangeable. Given them, $\langle Tu|Tv\rangle = u^{\mathsf{T}}T^{\mathsf{T}}Tv = u^{\mathsf{T}}v$, which is Equation (5.163). Conversely, taking u and v to be the standard basis vectors in Equation (5.163) reads off $\left(T^{\mathsf{T}}T\right)_{ij} = \delta_{ij}$. The norm statement is Equation (5.44) with $u = v$, and the angle statement then follows because both the numerator and the denominator of Equation (5.18) are unchanged. □

Isotropic Cartesian tensors

An array of numbers carrying N indices, each running over the three directions of physical space, is a *Cartesian tensor of rank N* when its entries in a rotated orthonormal frame are obtained from those in the old one by contracting each index with the rotation. Almost every constitutive law of continuum physics is written with such an array, and almost every material to which those laws are first applied is *isotropic*: it has no preferred direction, so the array must be the same in every frame. That single requirement determines the array almost completely, and the following is the statement that does the work.

Definition 5.126 (Isotropic Cartesian tensor)**.** An array $T_{i_1 \ldots i_N}$ of 3^N real numbers is an *isotropic Cartesian tensor of rank N* if

$$T_{i_1 \ldots i_N} = R_{i_1 j_1} \cdots R_{i_N j_N} T_{j_1 \ldots j_N} \qquad \text{for every } R \in \mathrm{SO}(3) , \tag{5.164}$$

repeated indices being summed over $1, 2, 3$ as in Section 5.3.3.

Rests on Equation (5.162) and Definition 5.30.

Lemma 5.127 (Parity constraint)**.** *Let T be isotropic of rank N. A component $T_{i_1\dots i_N}$ vanishes unless, for each value $p \in \{1,2,3\}$, the number m_p of indices equal to p has the same parity as N.*

Rests on Definition 5.126 and Equation (5.162).

Proof. *Derives Lemma 5.127.*

Fix p and let $R^{(p)}$ be the diagonal array with $+1$ in the p-th place and -1 in the other two: it satisfies $R^{\mathsf{T}}R = \mathbb{1}$ and has determinant $(+1)(-1)(-1) = +1$, so it lies in SO(3) by Equation (5.162). Being diagonal, $R^{(p)}_{ij} = s_i\delta_{ij}$ with $s_p = 1$ and $s_q = -1$ for $q \neq p$, so Equation (5.164) collapses to

$$T_{i_1\dots i_N} = s_{i_1}\cdots s_{i_N}\, T_{i_1\dots i_N} = (-1)^{N-m_p}\, T_{i_1\dots i_N}$$

with no summation, $N - m_p$ being the number of indices different from p. If $N - m_p$ is odd the component equals its own negative. □

Theorem 5.128 (Isotropic Cartesian tensors of rank at most four)**.** *Let T be an isotropic Cartesian tensor. Then, with $a, b, c \in \mathbb{R}$ arbitrary,*

$$N = 1: \quad T_i = 0\,, \tag{5.165}$$
$$N = 2: \quad T_{ij} = a\,\delta_{ij}\,, \tag{5.166}$$
$$N = 3: \quad T_{ijk} = a\,\varepsilon_{ijk}\,, \tag{5.167}$$
$$N = 4: \quad T_{ijkl} = a\,\delta_{ij}\delta_{kl} + b\,\delta_{ik}\delta_{jl} + c\,\delta_{il}\delta_{jk}\,, \tag{5.168}$$

and conversely every array of these forms is isotropic. A rank-four isotropic tensor therefore carries three free numbers, not $3^4 = 81$.

Rests on Lemmas 5.31 and 5.127 and Proposition 5.125.

Proof. *Derives Theorem 5.128.*

The converse first. δ_{ij} is isotropic, because $R_{ia}R_{jb}\delta_{ab} = R_{ia}R_{ja} = \delta_{ij}$ is $RR^{\mathsf{T}} = \mathbb{1}$ (Proposition 5.125); products of Kronecker deltas are then isotropic index group by index group; and ε_{ijk} is isotropic by Equation (5.59), since $\det R = 1$ on SO(3).

Two further elements of SO(3) do the rest of the work. The *cyclic* array C, with $C_{21} = C_{32} = C_{13} = 1$ and every other entry zero, is orthogonal and has determinant $+1$, a 3-cycle being an even permutation; substituted in Equation (5.164) it gives

$$T_{i_1\dots i_N} = T_{\sigma(i_1)\dots\sigma(i_N)}\,, \qquad \sigma: 1 \mapsto 3 \mapsto 2 \mapsto 1\,, \tag{5.169}$$

that is, the components are unchanged when every index *value* is relabelled cyclically. The *quarter turn* Q about the third axis, with $Q_{21} = 1$, $Q_{12} = -1$, $Q_{33} = 1$, is likewise in SO(3) and gives

$$T_{i_1\dots i_N} = (-1)^{m_1}\, T_{\tau(i_1)\dots\tau(i_N)}\,, \qquad \tau: 1 \leftrightarrow 2\,,\ 3 \mapsto 3\,, \tag{5.170}$$

m_1 being the number of indices equal to 1.

Rank one. Lemma 5.127 asks every m_p to be odd while $m_1 + m_2 + m_3 = 1$, which is impossible; every component vanishes.

Rank two. Every m_p must be even and they sum to 2, so one of them is 2 and the others vanish: $T_{ij} = 0$ unless $i = j$. By Equation (5.169), $T_{11} = T_{33} = T_{22}$, and Equation (5.166) holds with a that common value.

Rank three. Every m_p must be odd and they sum to 3, so each equals 1: only components with three distinct indices survive. Equation (5.169) gives $T_{123} = T_{312} = T_{231}$ and $T_{213} = T_{132} = T_{321}$, and Equation (5.170) applied to $(1,2,3)$, where $m_1 = 1$, gives $T_{123} = -T_{213}$. The six surviving components are therefore $\pm T_{123}$ according to the parity of the index arrangement, which is Equation (5.167) with $a = T_{123}$.

Rank four. Every multiplicity must be even and they sum to 4, so they are either $(4,0,0)$ or $(2,2,0)$ in some order: the surviving components are the three T_{pppp} and, for each ordered pair of distinct values $p \neq q$, the three arrangements T_{ppqq}, T_{pqpq}, T_{pqqp}. Equation (5.169) makes $T_{1111} = T_{3333} = T_{2222}$, a common value α, and carries the ordered pair (p,q) to $(\sigma(p),\sigma(q))$, whose orbits are $\{(1,2),(3,1),(2,3)\}$ and $\{(2,1),(1,3),(3,2)\}$; Equation (5.170) applied to $(1,1,2,2)$, $(1,2,1,2)$ and $(1,2,2,1)$ — each with $m_1 = 2$, hence with sign $+1$ — gives $T_{1122} = T_{2211}$, $T_{1212} = T_{2121}$ and $T_{1221} = T_{2112}$, joining the two orbits to each other. Hence the three numbers

$$a = T_{ppqq}\,, \qquad b = T_{pqpq}\,, \qquad c = T_{pqqp}$$

do not depend on which ordered pair $p \neq q$ is used, and Equation (5.168) reproduces every surviving component except possibly T_{pppp}, where it returns $a+b+c$ instead of α.

The last relation. Let $R(\theta)$ be the rotation by θ about the third axis, $R_{11} = R_{22} = \cos\theta$, $R_{21} = -R_{12} = \sin\theta$, $R_{33} = 1$. Only the values 1 and 2 occur in R_{1j}, so Equation (5.164) for the component T_{1111} involves only components whose indices lie in $\{1,2\}$, each carrying a factor $\cos\theta$ for an index 1 and $-\sin\theta$ for an index 2. Those with an odd number of 2's vanish by Lemma 5.127, and the rest give, writing $\gamma = \cos\theta$ and $s = \sin\theta$,

$$\alpha = \left(\gamma^4 + s^4\right)\alpha + 2\gamma^2 s^2\,(a+b+c)\,,$$

the six mixed components $T_{1122}, T_{1212}, T_{1221}, T_{2112}, T_{2121}, T_{2211}$ each carrying the factor $\gamma^2 s^2$ and summing to $2(a+b+c)$. Since $\gamma^4 + s^4 = 1 - 2\gamma^2 s^2$, this reads $2\gamma^2 s^2\,(a+b+c-\alpha) = 0$, and $\theta = \pi/4$ makes $\gamma^2 s^2 = 1/4$, so $\alpha = a+b+c$. □

Remark 5.129 (What the theorem buys, and what it does not). Equation (5.168) is the step at which the constitutive law of an isotropic elastic solid collapses from twenty-one independent constants to two, and at which the viscous stress of an isotropic Newtonian fluid collapses from eighty-one components to two viscosities: in both cases the array is symmetric in its first and last pairs of indices, which identifies two of the three constants of Equation (5.168) with each other and leaves two. The theorem says nothing about anisotropic materials — a crystal has a preferred set of directions and its tensor is invariant only under the finite point group of the lattice, so the count is a different and larger one — and it is a statement about SO(3) alone: adding invariance under reflections would in addition kill the rank-three case, since ε_{ijk} changes sign there by Equation (5.59).

The unitary group

Another very important subgroup of the general linear group is a group analogous to $\mathrm{O}(n,\mathbb{K})$, except that the added condition is $[T][T]^\dagger = \mathbb{1}$; we call it the *unitary group* and symbolize it $\mathrm{U}(n,\mathbb{K})$. Again, analogously, we can define the *special unitary group* by adding the condition $\det[T] = 1$. We have

$$\mathrm{U}(n,\mathbb{K}) = \left\{T \in \mathrm{GL}(n,\mathbb{K}) \mid [T][T]^\dagger = \mathbb{1}\right\}, \tag{5.171}$$

$$\mathrm{SU}(n,\mathbb{K}) = \{T \in \mathrm{U}(n,\mathbb{K}) \mid \det[T] = 1\}\,. \tag{5.172}$$

The symplectic group

The orthogonal and unitary groups are the invariance groups of a *symmetric* form. The third classical family is the invariance group of a non-degenerate *antisymmetric* one, which Proposition 5.115 has already reduced to a single normal form.

Definition 5.130 (The symplectic group)**.** Let ω be the non-degenerate antisymmetric bilinear form on $\mathbb{R}^{2n}$ whose matrix in the basis of Equation (5.143) is

$$J = [\omega] = \begin{pmatrix} 0 & \mathbb{1}_n \\ -\mathbb{1}_n & 0 \end{pmatrix}, \qquad (J)^{\mathsf{T}} = -J\,, \quad J^2 = -\mathbb{1}_{2n}\,, \quad J^{-1} = -J\,. \tag{5.173}$$

The *symplectic group* is the set of linear transformations of $\mathbb{R}^{2n}$ preserving ω,

$$\mathrm{Sp}(2n,\mathbb{R}) = \left\{ M \in \mathcal{M}_{2n\times 2n}\left[\mathbb{R}\right] \;\middle|\; (M)^{\mathsf{T}} J\, M = J \right\}, \tag{5.174}$$

the condition being $\omega(Mu, Mv) = \omega(u, v)$ for all u, v, written in components through Equation (5.138).

Rests on Proposition 5.115, Definition 5.109, and Equation (5.138).

Proposition 5.131 ($\mathrm{Sp}(2n,\mathbb{R})$ is a subgroup of $\mathrm{GL}(2n,\mathbb{R})$)**.** *Every M satisfying Equation (5.174) is invertible, and the set of such M is closed under multiplication and under inversion; it is therefore a subgroup of* $\mathrm{GL}(2n,\mathbb{R})$.

Rests on Definition 5.130, Equation (5.81), and Proposition 5.46.

Proof. *Derives Proposition 5.131.*

Invertibility. Taking determinants in Equation (5.174) and using $\det\left((M)^{\mathsf{T}}\right) = \det M$, which is immediate from the Leibniz expansion Equation (5.19) on relabelling $\sigma \longmapsto \sigma^{-1}$,

$$(\det M)^2 \det J = \det J\,,$$

and $\det J \neq 0$ because J is invertible by Equation (5.173). Hence $(\det M)^2 = 1$, so $\det M = \pm 1 \neq 0$ and $M \in \mathrm{GL}(2n,\mathbb{R})$ by Equation (5.81).

Closure. If M and N both satisfy Equation (5.174) then

$$\begin{aligned} (MN)^{\mathsf{T}} J\,(MN) &= (N)^{\mathsf{T}} \left((M)^{\mathsf{T}} J\, M\right) N \\ &= (N)^{\mathsf{T}} J\, N \\ &= J\,, \end{aligned}$$

the first line using $(MN)^{\mathsf{T}} = (N)^{\mathsf{T}}(M)^{\mathsf{T}}$.

Inverses. Transposing $MM^{-1} = M^{-1}M = \mathbb{1}_{2n}$ gives $(M^{-1})^{\mathsf{T}}(M)^{\mathsf{T}} = (M)^{\mathsf{T}}(M^{-1})^{\mathsf{T}} = \mathbb{1}_{2n}$, so $(M^{-1})^{\mathsf{T}}$ is the inverse of $(M)^{\mathsf{T}}$ — the uniqueness of Proposition 5.46. Multiplying $(M)^{\mathsf{T}} J\, M = J$ on the left by $(M^{-1})^{\mathsf{T}}$ and on the right by M^{-1} therefore yields

$$J = (M^{-1})^{\mathsf{T}} J\, M^{-1}\,,$$

so M^{-1} is symplectic as well. Associativity and the identity $\mathbb{1}_{2n}$, which trivially satisfies Equation (5.174), are inherited from $\mathrm{GL}(2n,\mathbb{R})$. □

The determinant is in fact never -1. Proving that requires one genuine piece of multilinear algebra, which is worth isolating because it is the characterization of the determinant that makes the argument work.

Lemma 5.132 (The alternating top form is unique up to scale)**.** *Let $\mathbb{V}$ be a real vector space of dimension N with basis $\{b_1, \ldots, b_N\}$, and let $\Omega : \mathbb{V}^N \longrightarrow \mathbb{R}$ be N-linear (linear in each argument with the others held fixed) and* alternating*, meaning that it changes sign whenever two of its arguments are exchanged. Then, with $v_j = \sum_i v^i{}_j b_i$,*

$$\Omega\left(v_1, \ldots, v_N\right) = \Omega\left(b_1, \ldots, b_N\right) \det\left[v^i{}_j\right], \tag{5.175}$$

and consequently, for every linear $M : \mathbb{V} \longrightarrow \mathbb{V}$,

$$\Omega\left(Mv_1, \ldots, Mv_N\right) = \det M\; \Omega\left(v_1, \ldots, v_N\right)\,. \tag{5.176}$$

Rests on Equation (5.19) and Definitions 5.14 and 5.106.

Proof. *Derives Lemma 5.132.*

First, Ω vanishes whenever two of its arguments are equal: exchanging those two leaves it unchanged and, by hypothesis, reverses its sign, so $\Omega = -\Omega$ and $2\Omega = 0$ in $\mathbb{R}$.

Expanding each argument by N-linearity,

$$\Omega\left(v_1, \ldots, v_N\right) = \sum_{i_1, \ldots, i_N} v^{i_1}{}_1 \cdots v^{i_N}{}_N \, \Omega\left(b_{i_1}, \ldots, b_{i_N}\right) ,$$

and every term in which two indices coincide vanishes by the previous paragraph. The surviving tuples are exactly those of the form $i_j = \sigma(j)$ for a permutation σ of N symbols; writing σ as a product of transpositions and applying the sign change once per transposition gives $\Omega\left(b_{\sigma(1)}, \ldots, b_{\sigma(N)}\right) = \operatorname{sgn}(\sigma)\, \Omega\left(b_1, \ldots, b_N\right)$. Hence

$$\Omega\left(v_1, \ldots, v_N\right) = \Omega\left(b_1, \ldots, b_N\right) \sum_{\sigma} \operatorname{sgn}(\sigma) \prod_{j=1}^{N} v^{\sigma(j)}_{\ j} ,$$

whose sum is the Leibniz expansion Equation (5.19) of $\det \left[v^i{}_j\right]$. This is Equation (5.175).

For Equation (5.176), note that the components of Mv_j are the entries of the array $[M]\left[v^i{}_j\right]$, since a matrix acts on a column of components by Proposition 5.43. Applying Equation (5.175) twice and using the multiplicativity of the determinant,

$$\begin{aligned} \Omega\left(Mv_1, \ldots, Mv_N\right) &= \Omega\left(b_1, \ldots, b_N\right) \det\left([M]\left[v^i{}_j\right]\right) \\ &= \det M \; \Omega\left(b_1, \ldots, b_N\right) \det\left[v^i{}_j\right] \\ &= \det M \; \Omega\left(v_1, \ldots, v_N\right) . \end{aligned}$$

□

Theorem 5.133 (A symplectic transformation has determinant +1)**.** *Every* $M \in \mathrm{Sp}(2n, \mathbb{R})$ *satisfies* $\det M = 1$. *Hence* $\mathrm{Sp}(2n, \mathbb{R}) \leq \mathrm{SL}(2n, \mathbb{R})$, *and for* $n = 1$ *the two coincide.*

Rests on Definition 5.130, Lemma 5.132, and Proposition 5.115.

Proof. *Derives Theorem 5.133.*

Build from ω the $2n$-argument function

$$\Omega\left(v_1, \ldots, v_{2n}\right) = \frac{1}{2^n n!} \sum_{\sigma} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} \omega\left(v_{\sigma(2i-1)}, v_{\sigma(2i)}\right) , \tag{5.177}$$

the sum running over all permutations σ of $2n$ symbols.

Ω *is* $2n$*-linear.* In each term every argument occurs in exactly one of the n factors and occurs there once, and ω is linear in each of its slots.

Ω *is alternating.* Let τ be the transposition exchanging two of the argument positions. Replacing the arguments by their τ-permutation and substituting $\sigma' = \tau\sigma$ in the sum — which merely reindexes it, $\sigma \longmapsto \tau\sigma$ being a bijection of the permutations onto themselves — gives back Equation (5.177) with $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\tau)\operatorname{sgn}(\sigma') = -\operatorname{sgn}(\sigma')$, that is, with the opposite sign.

Ω *does not vanish identically.* Evaluate it on the canonical basis of Equation (5.143), taken in the interleaved order $(e_1, f_1, e_2, f_2, \ldots, e_n, f_n)$. A term survives only if each factor pairs some e_i with the corresponding f_i, that is, only if the n unordered pairs $\{\sigma(2i-1), \sigma(2i)\}$ are exactly the n pairs $\{2i-1, 2i\}$ of argument positions; there are $n!$ ways to order those pairs among the factors and 2 ways to order each pair internally, so $2^n n!$ terms survive. Each contributes $+1$: reversing one pair internally multiplies its factor $\omega(e_i, f_i) = 1$ by -1, by antisymmetry, and multiplies $\operatorname{sgn}(\sigma)$ by -1 as well, while permuting whole pairs is an

even permutation, being a product of transpositions of two-element blocks. Dividing by $2^n n!$,

$$\Omega\,(e_1, f_1, \ldots, e_n, f_n) = 1\,.$$

Conclusion. If M is symplectic then $\omega(Mu, Mv) = \omega(u, v)$, so every factor in Equation (5.177) is unchanged when each argument is replaced by its image, and $\Omega(Mv_1, \ldots, Mv_{2n}) = \Omega(v_1, \ldots, v_{2n})$. But Equation (5.176) says the left-hand side equals $\det M\,\Omega(v_1, \ldots, v_{2n})$. Evaluating both on the canonical basis, where $\Omega = 1$, gives $\det M = 1$.

For $n = 1$ a direct computation settles the converse inclusion: for any real 2×2 array M one has $(M)^{\mathsf{T}} J\, M = \det M\, J$, so Equation (5.174) reduces to $\det M = 1$ and $\mathrm{Sp}(2, \mathbb{R}) = \mathrm{SL}(2, \mathbb{R})$. □

Proposition 5.134 (Dimension of the symplectic group)**.** *The set*

$$\mathfrak{sp}(2n, \mathbb{R}) = \Big\{ X \in \mathcal{M}_{2n\times 2n}\,[\mathbb{R}] \;\Big|\; (X)^{\mathsf{T}} J + J\,X = 0 \Big\} \tag{5.178}$$

is a Lie subalgebra of the matrix algebra of Proposition 5.121, and

$$\dim \mathfrak{sp}(2n, \mathbb{R}) = n\,(2n + 1)\ , \tag{5.179}$$

which is also the dimension of $\mathrm{Sp}(2n, \mathbb{R})$ *as a Lie group.*
Rests on Definitions 5.123 and 5.130 and Proposition 5.121.

Proof. *Derives Proposition 5.134.*

It is a subspace and a subalgebra. The defining condition Equation (5.178) is linear in X, so the set is a subspace. For the bracket, let $X, Y \in \mathfrak{sp}(2n, \mathbb{R})$, so that $(X)^{\mathsf{T}} J = -JX$ and $(Y)^{\mathsf{T}} J = -JY$. Then

$$\begin{aligned} ([X, Y])^{\mathsf{T}} J &= \Big((Y)^{\mathsf{T}} (X)^{\mathsf{T}} - (X)^{\mathsf{T}} (Y)^{\mathsf{T}} \Big)\, J \\ &= (Y)^{\mathsf{T}}\,(-JX) - (X)^{\mathsf{T}}\,(-JY) \\ &= -\,(-JY)\,X + (-JX)\,Y \\ &= -J\,(XY - YX) = -J\,[X, Y]\ , \end{aligned}$$

so $[X, Y] \in \mathfrak{sp}(2n, \mathbb{R})$, and the two Lie-algebra axioms Equations (5.155) and (5.156) hold because the commutator satisfies them on the whole matrix algebra.

The dimension. Put $S = J\,X$, so that $X = J^{-1} S = -J\,S$ by Equation (5.173); the assignment $X \longmapsto S$ is a linear bijection of $\mathcal{M}_{2n\times 2n}\,[\mathbb{R}]$ onto itself. Using $(J^{-1})^{\mathsf{T}} = (-J)^{\mathsf{T}} = J$ and $J^2 = -\mathbb{1}_{2n}$,

$$(X)^{\mathsf{T}} J = (J^{-1} S)^{\mathsf{T}} J = (S)^{\mathsf{T}} (J^{-1})^{\mathsf{T}} J = (S)^{\mathsf{T}} J^2 = -(S)^{\mathsf{T}}\,,$$

so the condition $(X)^{\mathsf{T}} J + JX = 0$ becomes $-(S)^{\mathsf{T}} + S = 0$: precisely the statement that S is symmetric. The symmetric arrays of order $2n$ form a subspace of dimension $2n(2n+1)/2 = n(2n+1)$, a basis being given by the arrays with 1 in a diagonal entry and by those with 1 in a pair of mirror-image off-diagonal entries. This is Equation (5.179).

That the group has the same dimension is the general correspondence between a Lie group and its algebra, established in Chapter 14; the check $n = 1$ is reassuring, for $n(2n + 1) = 3$ agrees with the three parameters of $\mathrm{SL}(2, \mathbb{R})$ — four entries subject to one determinant condition — as Theorem 5.133 requires. □

Remark 5.135 (Where the symplectic group is used). The group is the exact linear counterpart of a structure physics meets in its first serious reformulation of mechanics. The phase space of a system with N degrees of freedom carries a non-degenerate antisymmetric form pairing

each coordinate with its conjugate momentum (Remark 5.116), so $\omega(u, v)$ has the SI unit of action, J s; the linear transformations preserving it are exactly the linear *canonical* transformations, which is why Hamilton's equations take the same shape in every canonical coordinate system (Chapter 22). The passage from the fixed form on one vector space to a form varying smoothly over a manifold, and the theorem of Darboux that says nothing is lost in the passage, belong to Chapter 24. Note also the contrast with Equation (5.161): for the orthogonal group the determinant condition $\det M = 1$ is an extra requirement, singling out SO inside O, whereas for the symplectic group Theorem 5.133 makes it automatic — there is no "special symplectic group" to define.

5.8.2 Representations

Definition 5.136 (Representation of a group)**.** Let G be a group and X a set. We define a *representation* of G on X as a group homomorphism

$$\begin{array}{rccc} \rho: & G & \longrightarrow & S_X \\ & a & \longmapsto & \rho_a : X \longrightarrow X \,. \end{array}$$

Rests on Definitions 4.20, 4.39 and 4.44.

From now on we shall consider representations in which the set X is a vector space, that is, the case

$$S_X = \mathrm{GL}(\mathbb{V}) \,,$$

so the representations we shall study are of the form

$$\begin{array}{rccccccc} U: & G & \longrightarrow & \mathrm{GL}(\mathbb{V}) & & & & \\ & a & \longmapsto & U(a): & \mathbb{V} & \longrightarrow & \mathbb{V} \\ & & & & |v\rangle & \longmapsto & U(a)\,(|v\rangle) \overset{\text{not.}}{=} U(a)\,|v\rangle \ , \end{array}$$

where $U(a)$ is a linear transformation by definition.

Definition 5.137 (Dimension of a representation)**.** We refer to the dimension of a representation $U : G \longrightarrow \mathrm{GL}(\mathbb{V})$ as the dimension of $\mathbb{V}$.

Rests on Definitions 5.15 and 5.136.

Definition 5.138 (Faithful representation)**.** We say that a representation U is *faithful* if and only if U is an isomorphism. We say that a representation is *degenerate* if it is not faithful.

Rests on Definitions 4.40 and 5.136.

Definition 5.139 (Unitary representation)**.** Let $U : G \longrightarrow \mathrm{GL}\,(\mathbb{V})$ be a representation. We say that U is *unitary* if and only if

1. the vector space $\mathbb{V}$ is equipped with an inner product;

2. the linear transformation $U(a)$ is unitary for every a.

Rests on Definitions 5.17 and 5.136 and Equation (5.171)*.*

From now on we shall consider only unitary representations, unless we specify a more general case.

Representation matrix

Consider a linear space $\mathbb{V} = \left\langle \{|\hat{e}_i\rangle\}_i^{\dim \mathbb{V}} \right\rangle$ and a representation U. We can write

$$U(a)\,|\hat{e}_i\rangle = U^j{}_i(a)\,|\hat{e}_j\rangle\ ,$$

that is, the effect of a linear transformation on an element of the basis can be written as a linear combination of the elements of the basis — for it is, indeed, a basis.

We can associate the matrix $\left[U^j{}_i(a)\right]$ with the coefficients of this linear combination. For a more compact notation we symbolize this simply as

$$\left[U^j{}_i(a)\right] \overset{\text{not.}}{=} D(a)\,,$$

which is the *matrix representation of the representation* U with respect to a; equivalently, $D(a)$ is the matrix representation of the linear transformation $U(a)$.

Let us see that the matrix representation of a representation is also a homomorphism. We have

$$\begin{aligned} U(a)U(b)\,|\hat{e}_i\rangle &= U(a)U^j{}_i(b)\,|\hat{e}_j\rangle \\ U(ab)\,|\hat{e}_i\rangle &= U^k{}_j(a)U^j{}_i(b)\,|\hat{e}_k\rangle \\ U^k{}_i(ab)\,|\hat{e}_k\rangle &= U^k{}_j(a)U^j{}_i(b)\,|\hat{e}_k\rangle\ , \end{aligned}$$

from which we conclude that

$$U^k{}_i(ab) = U^k{}_j(a)U^j{}_i(b)\,,$$

or, in matrix notation,

$$D(ab) = D(a)D(b)\,. \tag{5.180}$$

Definition 5.140 (Equivalent representations)**.** Consider the representations U and U'. We say that these representations are *equivalent* if and only if there exists an operator S such that

$$U'(a) = SU(a)S^{-1}\,,$$

which is a similarity transformation; consequently, the set of all equivalent representations of a group forms an equivalence class.

Rests on Definition 5.136.

Definition 5.141 (Character)**.** We define the *character* of a representation U as

$$\chi(a) = \operatorname{tr}\left(D(a)\right). \tag{5.181}$$

Rests on Definition 5.136.

As we know, the trace is invariant under similarity transformations, so the character of a representation is invariant under similarity transformations. In other words, an equivalence class of representations is characterized by a character. The character of a representation is thus useful because it is basis-independent.

5.8.3 Irreducible representations

Definition 5.142 (Invariant subspace)**.** Let $\mathbb{V}$ be a vector space and $\mathbb{S} \leq \mathbb{V}$ a subspace. We say that the subspace $\mathbb{S}$ is *invariant* with respect to the representation U if and only if

$$U(a)\,|v\rangle \in \mathbb{S}, \quad \forall\ |v\rangle \in \mathbb{S}\ \wedge\ \forall\, a \in G\,,$$

that is, the vectors of the subspace do not leave this set when operated on by $U(a)$.

Rests on Definitions 5.6 and 5.136.

To visualize this more easily, we may conceive an invariant subspace as a set of trapped vectors, which do not escape from $\mathbb{S}$ when operated on by the linear transformation $U(a)$.

Given a representation there are always two trivial invariant subspaces: the null space and $\mathbb{V}$ itself.

Definition 5.143 (Irreducible representation)**.** Let $U : G \longrightarrow \mathrm{GL}(\mathbb{V})$ be a representation. We say that U is *irreducible* if and only if the only invariant subspaces with respect to it are the trivial ones, namely the null space and $\mathbb{V}$.

We say that a representation is *reducible* if and only if it is not irreducible, that is, if there exists a (non-trivial) invariant subspace with respect to U.

Rests on Definition 5.142.

Definition 5.144 (Totally reducible representation)**.** Consider an invariant subspace $\mathbb{S}$ with respect to U. We say that the representation U is *totally reducible* if the orthogonal complement of $\mathbb{S}$ is also invariant with respect to U.

Rests on Definition 5.142 and Equation (5.52)*.*

Before stating the proposition, fix what "unitary" says about the operator rather than about its matrix. By Definition 5.139 and Equation (5.171), U is unitary when $\mathbb{V}$ carries an inner product and $D(a)^\dagger D(a) = \mathbb{1}$ for every a. In an orthonormal basis (Equation (5.49)) that condition is equivalent to the statement that $U(a)$ *preserves the inner product*, since

$$\begin{aligned}\langle U(a)e_i | U(a)e_j\rangle &= \sum_{k,l} \left(D^k{}_i(a)\right)^* D^l{}_j(a)\,\langle e_k|e_l\rangle \qquad \text{by Equations (5.34) and (5.36)}\\ &= \sum_{k} \left(D^k{}_i(a)\right)^* D^k{}_j(a)\\ &= \left(D(a)^\dagger D(a)\right)^i{}_j\,, \end{aligned} \tag{5.182}$$

which equals $\delta^i{}_j = \langle e_i|e_j\rangle$ for all i, j exactly when $D(a)^\dagger D(a) = \mathbb{1}$; the general case follows by expanding arbitrary u, v in the basis. We shall use the invariant form

$$\langle U(a)u | U(a)v\rangle = \langle u|v\rangle\,, \qquad \forall\, u, v \in \mathbb{V},\ \forall\, a \in G\,. \tag{5.183}$$

Lemma 5.145 (Invariance of the orthogonal complement)**.** *Let $U : G \longrightarrow \mathrm{GL}(\mathbb{V})$ be a unitary representation of a group G on a vector space $\mathbb{V}$ with an inner product, and let $\mathbb{S} \leq \mathbb{V}$ be an invariant subspace (Definition 5.142). Then the orthogonal complement $\mathbb{S}^\perp$ of Equation* (5.52) *is invariant as well.*

Rests on Definitions 5.139 and 5.142 and Equations (4.25)*,* (5.52) *and* (5.183)*.*

Proof. *Derives Lemma 5.145.*

Let $w \in \mathbb{S}^\perp$ and $a \in G$; we must show $U(a)w \in \mathbb{S}^\perp$, that is, $\langle s|U(a)w\rangle = 0$ for every $s \in \mathbb{S}$. Because U is a group homomorphism into $\mathrm{GL}(\mathbb{V})$, the operator $U(a)$ is invertible with $U(a)^{-1} = U(a^{-1})$ by Equation (4.25), so we may write $s = U(a)U(a^{-1})s$ and use Equation (5.183):

$$\begin{aligned}\langle s|U(a)w\rangle &= \Big\langle U(a)U(a^{-1})s \Big| U(a)w\Big\rangle\\ &= \Big\langle U(a^{-1})s \Big| w\Big\rangle \qquad \text{by Equation (5.183)}\\ &= 0\,, \end{aligned}$$

the last step because $a^{-1} \in G$ — a group contains the inverse of each of its elements — so $U(a^{-1})s \in \mathbb{S}$ by the invariance of $\mathbb{S}$, and w is orthogonal to every vector of $\mathbb{S}$. As s was arbitrary, $U(a)w \in \mathbb{S}^\perp$. □

Three hypotheses were used and no others: that U is unitary, that G is a group, and that U is a homomorphism. Neither the dimension of $\mathbb{V}$ nor any property of G beyond its being a group entered, so the lemma holds for infinite groups and in infinite dimension.

Proposition 5.146. *Let U be a unitary representation. If U is not totally reducible, then it is not reducible.*

Rests on Lemma 5.145 and Definitions 5.143 and 5.144.

Proof. *Derives Proposition 5.146.*

We prove the contrapositive: if U is reducible then it is totally reducible. Suppose U is reducible, so that by Definition 5.143 there exists a non-trivial invariant subspace $\mathbb{S}$. By Lemma 5.145 the orthogonal complement $\mathbb{S}^{\perp}$ is invariant too, and that is exactly the condition Definition 5.144 imposes for U to be totally reducible.

Hence no unitary representation can be reducible without being totally reducible, and the stated implication follows. □

Remark 5.147 (The proposition is the contrapositive of a stronger one). Read in the direction just proved, the statement says something the source does not spell out: for a *unitary* representation the case "reducible but not totally reducible" is empty. It is not empty for representations in general. The standard witness lives already in $\mathrm{GL}(2,\mathbb{C})$: the map

$$U(n) = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}, \qquad n \in \mathbb{Z},$$

is a representation of the additive group $\mathbb{Z}$, since $U(n)U(m) = U(n+m)$. The subspace spanned by $(1,0)^{\mathsf{T}}$ is invariant, so the representation is reducible; but *no* complementary subspace is invariant, because any vector with non-zero second component is moved by $U(1)$ out of every line through it other than the first axis. The representation is therefore reducible and not totally reducible — and, consistently with Proposition 5.146, it is not unitary: $U(n)$ does not preserve the standard inner product, as $\left\|U(n)(0,1)^{\mathsf{T}}\right\|^2 = 1 + n^2$ grows without bound. This is the exact point at which Proposition 5.150 does its work: for a *finite* group no such example exists, because every representation of a finite group can be made unitary.

5.9 Representations of finite groups

The two results of this section are the ones the rest of the chapter — and, through it, a good deal of the physics later in the treatise — actually stands on. The source states both and leaves their derivations in draft form; they are given here in full, with the hypotheses each of them really needs made explicit, because those hypotheses are what decides where the results may be used and where they may not.

5.9.1 Complete reducibility of a unitary representation

Proposition 5.148. *Let $U : G \longrightarrow \mathrm{GL}(\mathbb{V})$ be a unitary representation of a group G on a* finite-dimensional *vector space $\mathbb{V}$ equipped with an inner product. Then:*

1. *if U is reducible it is totally reducible — for every invariant subspace $\mathbb{S}$ the orthogonal complement $\mathbb{S}^{\perp}$ is invariant as well, and $\mathbb{V} = \mathbb{S} \oplus \mathbb{S}^{\perp}$;*
2. *consequently $\mathbb{V}$ is the orthogonal direct sum of finitely many invariant subspaces on each of which U acts irreducibly,*

$$\mathbb{V} = \bigoplus_{\mu} \mathbb{V}^{\mu}, \qquad U(a) = \bigoplus_{\mu} U^{\mu}(a), \tag{5.184}$$

with $U^{\mu}(a) = U(a)|_{\mathbb{V}^{\mu}}$ a unitary irreducible representation of G on $\mathbb{V}^{\mu}$.

Rests on Lemma 5.145, Corollary 5.27, and Definitions 5.143 and 5.144.

Proof. *Derives Proposition 5.148.*

Part 1. That $\mathbb{S}^{\perp}$ is invariant is Lemma 5.145, which used no finiteness at all. That the two subspaces exhaust $\mathbb{V}$,

$$\mathbb{V} = \mathbb{S} \oplus \mathbb{S}^{\perp},$$

is Corollary 5.27, and *this* is where the finite dimension enters: it is what guarantees the orthonormal basis of $\mathbb{S}$ from which the projection was built. Every vector of $\mathbb{V}$ therefore splits uniquely as $v = s + w$ with $s \in \mathbb{S}$ and $w \in \mathbb{S}^{\perp}$, and since $U(a)s \in \mathbb{S}$ and $U(a)w \in \mathbb{S}^{\perp}$, the operator $U(a)$ acts on the two summands separately:

$$U(a)v = U(a)|_{\mathbb{S}}\, s + U(a)|_{\mathbb{S}^{\perp}}\, w\,.$$

In a basis adapted to the splitting the matrix $D(a)$ is block diagonal for every a simultaneously, which is what *totally reducible* means operationally.

Each restriction is itself a representation: it is linear; it is invertible on its subspace, because $U(a^{-1})$ maps that subspace into itself and $U(a)U(a^{-1}) = U(e) = \mathrm{id}$; and it inherits $U(ab) = U(a)U(b)$ from U. It is unitary, because Equation (5.183) holds in particular for vectors of the subspace.

Part 2. Induction on $n = \dim \mathbb{V}$. If $n = 1$ the only subspaces are the trivial ones, so U is irreducible by Definition 5.143 and the decomposition has one summand. Let $n > 1$ and suppose the claim proved for all dimensions smaller than n. If U is irreducible we are again done. If not, there is a non-trivial invariant subspace $\mathbb{S}$, and by Part 1

$$\mathbb{V} = \mathbb{S} \oplus \mathbb{S}^{\perp}, \qquad 1 \le \dim \mathbb{S} < n\,, \quad 1 \le \dim \mathbb{S}^{\perp} < n\,,$$

the strict inequalities holding because $\mathbb{S}$ is neither the null space nor the whole of $\mathbb{V}$. The restrictions of U to $\mathbb{S}$ and to $\mathbb{S}^{\perp}$ are unitary representations on spaces of strictly smaller dimension, so the induction hypothesis applies to each and decomposes it into irreducible summands; collecting the two lists gives Equation (5.184). The recursion terminates after at most n steps because the dimension drops by at least one each time, and the summands are mutually orthogonal by construction. □

Remark 5.149 (What each hypothesis is for). Three are in play and they are used in three different places.

- *An inner product* is needed to have an orthogonal complement at all. Without one, "totally reducible" in the sense of Definition 5.144 is not even a statement; the basis-free version of the result is that every invariant subspace has *some* invariant complement, and the inner product is what supplies it canonically.

- *Unitarity* is what makes that complement invariant (Lemma 5.145). Drop it and the conclusion is false: the representation of $\mathbb{Z}$ in Remark 5.147 is reducible with no invariant complement whatsoever.

- *Finite dimension* is used twice, for the splitting Equation (5.53) and for the termination of the induction. In infinite dimension the first survives for closed subspaces of a Hilbert space (Definition 12.2) — by the projection theorem reserved for Section 12.2.1 — but the second does not, and a unitary representation there need not be a direct sum of irreducibles at all — the representation of the translation group $(\mathbb{R}, +)$ on $L^2(\mathbb{R})$ by $(U(a)f)(x) = f(x - a)$ has no irreducible subrepresentation whatever, and its decomposition is a direct *integral* over the characters $\mathrm{e}^{\mathrm{i}kx}$, which is the Fourier transform of Chapter 17 in disguise.

5.9.2 Every representation of a finite group is unitary

The preceding proposition is worth little on its own, since it presumes what is usually not given: that the representation one happens to be handed is unitary. The following result removes the presumption for finite groups, and it does so constructively — it exhibits the inner product that makes the representation unitary. The device is the averaging or "unitarian" trick, which replaces a quantity by its average over the group and thereby makes it group-invariant by construction; it is the discrete ancestor of every projection onto a symmetric configuration used later in the treatise.

Proposition 5.150 (Averaging trick)**.** *Let $G = \{g_i\}_{i=1}^{n_G}$ be a* finite *group of order n_G and let $U : G \longrightarrow \mathrm{GL}(\mathbb{V})$ be a representation on a finite-dimensional vector space $\mathbb{V}$ over $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, equipped with an inner product $\langle\ |\ \rangle$. Define*

$$\langle u, v\rangle_G = \frac{1}{n_G}\sum_{i=1}^{n_G} \langle U(g_i)u | U(g_i)v\rangle \ . \tag{5.185}$$

Then $\langle\ ,\ \rangle_G$ is an inner product on $\mathbb{V}$; it is G-invariant,

$$\langle U(a)u, U(a)v\rangle_G = \langle u, v\rangle_G \ , \qquad \forall\, a \in G\,, \tag{5.186}$$

so that U is unitary with respect to it; and there exists $S \in \mathrm{GL}(\mathbb{V})$ such that

$$U'(a) = SU(a)S^{-1} \tag{5.187}$$

is unitary with respect to the original inner product. Hence U is equivalent, in the sense of Definition 5.140, to a unitary representation.

Rests on Definitions 5.17, 5.139 and 5.140, Proposition 5.26, and Equation (5.183).

Proof. *Derives Proposition 5.150.*

Step 1: Equation (5.185) is an inner product. Each summand is the original inner product evaluated on the images of u and v, so the five axioms transfer term by term.

- *Linearity in the second argument* Equations (5.34) and (5.35): $U(g_i)$ is linear, so $U(g_i)(av + bw) = aU(g_i)v + bU(g_i)w$, and Equations (5.34) and (5.35) applied inside each summand give the same for the average. The conjugate rules Equations (5.36) and (5.37) follow in the same way.

- *Hermiticity* Equation (5.33): each summand satisfies

  $$\langle U(g_i)u | U(g_i)v\rangle^* = \langle U(g_i)v | U(g_i)u\rangle \ ,$$

  and $1/n_G$ is real, so the sum inherits the property.

- *Positive definiteness* Equation (5.32): let $u \neq 0_{\mathbb{V}}$. Since $U(g_i) \in \mathrm{GL}(\mathbb{V})$ is invertible, $U(g_i)u \neq 0_{\mathbb{V}}$, so every summand $\langle U(g_i)u | U(g_i)u\rangle$ is strictly positive by Equation (5.32); a sum of n_G strictly positive reals divided by n_G is strictly positive. Equation (5.31) follows, the direction ($\Leftarrow$) from Equation (5.38).

Note where the finiteness of G has been used: the sum has finitely many terms, and n_G is invertible in $\mathbb{K}$.

Step 2: invariance, Equation (5.186). Using the homomorphism property $U(g_i)U(a) = U(g_i a)$,

$$\begin{aligned}\langle U(a)u, U(a)v\rangle_G &= \frac{1}{n_G}\sum_{i=1}^{n_G}\langle U(g_i)U(a)u|U(g_i)U(a)v\rangle \\ &= \frac{1}{n_G}\sum_{i=1}^{n_G}\langle U(g_i a)u|U(g_i a)v\rangle \\ &= \frac{1}{n_G}\sum_{j=1}^{n_G}\langle U(g_j)u|U(g_j)v\rangle \\ &= \langle u, v\rangle_G \ ,\end{aligned}$$

the third line because right multiplication by a is a bijection of G onto itself: it is injective, since $g_i a = g_j a$ implies $g_i = g_j$ on multiplying by a^{-1}, and therefore surjective on a finite set, so $\{g_i a\}$ is a relabelling of $\{g_j\}$ and the summands are the same numbers in another order. *This is the whole of the trick*: an average over the group cannot tell which element of the group it is looking at. With Equation (5.186) in hand, U satisfies Equation (5.183) for the product $\langle\ ,\ \rangle_G$, that is, it is a unitary representation on the inner-product space $(\mathbb{V}, \langle\ ,\ \rangle_G)$.

Step 3: the equivalence Equation (5.187). What remains is bookkeeping: to say "unitary" with respect to the original product rather than the new one. By Proposition 5.26 applied twice, choose a basis $\{e_i\}_{i=1}^n$ orthonormal for $\langle\ |\ \rangle$ and a basis $\{\tilde{e}_i\}_{i=1}^n$ orthonormal for $\langle\ ,\ \rangle_G$, and let S be the linear map determined by $S\tilde{e}_i = e_i$, which is invertible because it carries a basis to a basis. For $u = \sum_i u^i \tilde{e}_i$ and $v = \sum_i v^i \tilde{e}_i$, orthonormality of each basis for its own product gives

$$\langle Su|Sv\rangle = \sum_{i=1}^{n}\left(u^i\right)^* v^i = \langle u, v\rangle_G \ ,$$

that is, S carries the averaged product to the original one. Therefore, for all $x, y \in \mathbb{V}$,

$$\begin{aligned}\langle U'(a)x|U'(a)y\rangle &= \left\langle SU(a)S^{-1}x\middle|SU(a)S^{-1}y\right\rangle \\ &= \left\langle U(a)S^{-1}x, U(a)S^{-1}y\right\rangle_G \\ &= \left\langle S^{-1}x, S^{-1}y\right\rangle_G \qquad \text{by Equation (5.186)} \\ &= \langle x|y\rangle \ ,\end{aligned}$$

so U' satisfies Equation (5.183) for $\langle\ |\ \rangle$ and is a unitary representation, equivalent to U by construction. □

Remark 5.151 (Where the finiteness is really needed, and what replaces it). Finiteness of G entered in exactly two places, both in Step 1–2: the sum Equation (5.185) must converge, and n_G must be invertible. Neither is a statement about $\mathbb{V}$, and that is what makes the trick portable.

- *Compact continuous groups.* Replace the normalized sum $n_G^{-1}\sum_i$ by the integral $\int_G \mathrm{d}\mu(g)$ against the normalized bi-invariant *Haar measure* of the group, which exists and is unique on a compact group. Step 2 goes through verbatim, since invariance of the measure under right translation is precisely what replaced the rearrangement of the finite sum; positivity in Step 1 survives because the integrand is continuous and strictly positive. Every *continuous* finite-dimensional representation of a compact group — and $\mathrm{SO}(3,\mathbb{R})$, $\mathrm{SU}(2)$ and $\mathrm{U}(1)$, the groups on which most of the physics in this treatise turns, are compact — is therefore equivalent to a unitary one and, by Proposition 5.148, completely reducible. The continuity hypothesis is not decoration and cannot be

dropped: the whole argument consists of integrating $g \longmapsto \langle U(g)u | U(g)v \rangle$ against $\mathrm{d}\mu$, and a representation that is not at least measurable has nothing there to integrate. Continuity is the hypothesis every physical example satisfies and the one under which the statement is normally quoted; measurability suffices, and for these groups the two coincide in practice. This treatise takes the existence and uniqueness of the normalized bi-invariant measure on a compact group as an imported theorem of topological group theory, due to Haar [721]; it is not constructed anywhere in this book, and the natural home for it would be the Lie-group material of Chapter 14, which at the time of writing does not carry it either.

- *Non-compact groups.* There the argument fails, and not for want of ingenuity: a non-compact group carries no finite invariant measure, so there is nothing to average with. The failure is physical as well as technical. The Lorentz group is non-compact — its boosts form a non-compact one-parameter family — and its *non-trivial* finite-dimensional representations, the ones that classify fields in Chapter 39, are *not* unitary; its non-trivial unitary representations, the ones that classify particle states, are all infinite-dimensional [849]. The qualifier is needed on both halves and for the same reason: the trivial representation, which sends every group element to the identity operator on a one-dimensional space, is finite-dimensional and unitary at once, and would falsify either statement read without it. The exception is trivial in the literal sense — it is the one representation that records nothing whatever about the group — but a statement a counterexample falsifies is a false statement, so the qualifier stands. Complete reducibility must then be argued case by case, and the sharp division between the two kinds of representation of the Poincaré group is a direct consequence of this one missing average.

- *The arithmetic caveat.* Dividing by n_G presupposes that $n_G \neq 0$ in the field. Over $\mathbb{R}$ or $\mathbb{C}$ — the only fields this chapter admits — that is automatic, but over a field whose characteristic divides the order of the group the trick collapses, and with it complete reducibility. This is the hypothesis of Maschke's theorem, and it is not idle: the modular representation theory that results is a genuinely different subject.

Remark 5.152 (Why these two propositions are cited from the Standard Model). Taken together, Proposition 5.150 and Proposition 5.148 say that a symmetry acting on a finite-dimensional space may always be taken to act by inner-product-preserving operators, and that the space then splits cleanly into irreducible pieces. The flavour sector of the Standard Model is where a physicist meets the consequence in its bluntest form.

The three known generations of quarks span a three-dimensional complex space — an *internal* space, indexed by generation, with no relation to the three spatial dimensions of Chapter 39 — and there are two natural bases for it: the states of definite mass and the states that couple to the charged weak current. The matrix V relating the two bases is the Cabibbo–Kobayashi–Maskawa matrix [1248, 1505]; the corresponding matrix in the lepton sector, relating neutrino mass states to flavour states, is the Pontecorvo–Maki–Nakagawa–Sakata matrix [1159, 1237]. Both are dimensionless, as any change of basis must be: their entries are pure numbers, and no SI unit attaches to them.

Both bases are orthonormal for the same inner product — which is available precisely because the symmetry can be realized unitarily — so the matrix carrying one to the other is unitary. Unitarity of a matrix is two families of relations, not one — its rows are orthonormal and so are its columns — and they say different physical things, so which is which has to be kept straight. That single word carries the whole experimental programme:

- unitarity of the quark matrix gives, with i, j labelling the up-type rows and k, l the

down-type columns, the row relations

$$\sum_{k} V_{ik} V_{jk}^{*} = \delta_{ij}$$

and the column relations $\sum_i V_{ik}^{*} V_{il} = \delta_{kl}$. In either family the off-diagonal case is a sum of three complex numbers equal to zero, so the three must close into a triangle in the complex plane — six triangles in all, one for each pair of rows and one for each pair of columns. The one that is actually measured, and that the phrase *the unitarity triangle* refers to, comes from the down and the bottom column. It is customarily written in the conjugated form

$$V_{ud} V_{ub}^{*} + V_{cd} V_{cb}^{*} + V_{td} V_{tb}^{*} = 0\,,$$

the complex conjugate of the column relation just displayed with $k = d$ and $l = b$, and zero exactly when that one is. Its three sides are of comparable length, which is what makes its angles measurable in CP-violation experiments (Chapter 111); a failure to close would be evidence for a fourth generation or for physics outside the model, and is reported as such;

- unitarity of the leptonic matrix likewise gives two relations, and they are not interchangeable. The *row* relation

$$\sum_{i} |U_{\alpha i}|^{2} = 1$$

holds the flavour index α fixed and sums over the mass states i: it says that the state produced together with a charged lepton of flavour α is a superposition of the three mass states whose probabilities sum to one. The *column* relation $\sum_{\alpha} |U_{\alpha i}|^{2} = 1$ holds the mass index i fixed instead, and says the different thing that the mass state i has unit total flavour content. The statement that a neutrino produced in one flavour is detected in some flavour with total probability one is $\sum_{\beta} P(\nu_\alpha \to \nu_\beta) = 1$, and it consumes both: the column relation collapses the interference between mass states when the sum over the detected flavour β is done, and the row relation is what makes the surviving $\sum_i |U_{\alpha i}|^{2}$ equal to one (Chapter 114). Without unitarity the oscillation formula would not conserve probability.

The physics of both matrices belongs to Chapter 103, which at the time of writing carries its headings, its citations and its editorial specification but not yet the derivations. What this chapter contributes is only the licence to write the word “unitary” — but the two propositions above are where that licence comes from, and neither is free: each fails outside its hypotheses, as Remark 5.151 and Remark 5.149 record.

5.9.3 The regular representation

Recall from Equation (4.28) that upon multiplying the i-th element of a group by an element a of the group, it takes the place a_i; that is, a permutation occurs. We can write

$$\begin{aligned} a g_i &= g_{a_i} \\ &= g_j \, (\Delta_a)^{j}{}_{i}\,, \end{aligned}$$

where

$$(\Delta_a)^{j}{}_{i} = \begin{cases} 1, & j = a_i \\ 0, & j \neq a_i \end{cases} \tag{5.188}$$

from which we see that

$$(\Delta_a)^{j}{}_{i} = \delta^{j}{}_{a_i}\,. \tag{5.189}$$

Consider then elements $a, b, c \in G$ such that $c = ab$. In this permutation notation we can write

$$\begin{aligned} abg_i &= ag_{b_i} \\ &= ag_j \left(\Delta_b\right)^j{}_i \\ &= g_{a_j} \left(\Delta_b\right)^j{}_i \\ &= g_k \left(\Delta_a\right)^k{}_j \left(\Delta_b\right)^j{}_i \,, \end{aligned}$$

and, in addition,

$$\begin{aligned} cg_i &= g_{c_i} \\ &= g_k \left(\Delta_c\right)^k{}_i \,, \end{aligned}$$

and since $ab = c$ we have

$$\left(\Delta_a\right)^k{}_j \left(\Delta_b\right)^j{}_i = \left(\Delta_c\right)^k{}_i \,,$$

or, in operator notation,

$$\Delta_a \Delta_b = \Delta_c \,.$$

We have therefore found a homomorphism $a \overset{U}{\longmapsto} U(a) = \Delta_a$, since

$$\begin{aligned} U(a)U(b) &= U(c) \\ &= U(ab) \,. \end{aligned}$$

This homomorphism is called the *regular representation* of G.

The coefficients of Δ_a are determined by the following cases.

1. If $a = e$, we have

$$\begin{aligned} eg_i &= g_{e_i} \\ g_i &= g_{e_i} \,. \end{aligned}$$

Since the neutral element does not permute the others, $i = e_i$, so that

$$\left(\Delta_e\right)^j{}_i = \delta^j{}_i \,, \tag{5.190}$$

and we can compute the character by setting $j = i$, obtaining

$$\chi^R_e = n_G \,, \tag{5.191}$$

where n_G is the order of the group.

2. If $a \neq e$, we have the usual case, Equation (5.189). But upon setting $j = i$, since a_i is a permutation of i, the two indices are always distinct (for $a \neq e$); hence $\delta^i{}_{a_i} = 0$, and thus

$$\chi^R_a = 0 \,. \tag{5.192}$$

For example, for the group C_2 we have

$$\Delta_e = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$\begin{aligned} \Delta_a &= \begin{pmatrix} \delta^1{}_{a_1} & \delta^1{}_{a_2} \\ \delta^2{}_{a_1} & \delta^2{}_{a_2} \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} . \end{aligned}$$

5.9.4 Direct-sum and direct-product representations

Direct-sum representation

Consider the irreducible representations $U^\mu : G \longrightarrow \mathrm{GL}(\mathbb{V}^\mu)$, $\mu \in \mathbb{N}$. Their invariant subspaces are the $\mathbb{V}^\mu$ themselves, since they are irreducible. Let $\mathbb{V}$ be the direct sum of these spaces, that is,

$$\mathbb{V} = \bigoplus_\mu \mathbb{V}^\mu ,$$

and consider further a reducible representation U for which $\mathbb{V}$ is an invariant subspace, and which coincides with $U^\mu(a)$ on $\mathbb{V}^\mu$.

We have seen that the representation of a finite group is equivalent to a unitary representation (Proposition 5.150), and that if a representation is unitary and reducible then it is completely reducible (Proposition 5.148).

In this context we can decompose the representation U into the irreducible representations U^μ. In this procedure, equivalent representations may appear a_μ times each. That is,

$$U(a) = \underbrace{U^1(a) + \ldots + U^1(a)}_{a_1 \text{ times}} + \ldots + \underbrace{U^\mu(a) + \ldots + U^\mu(a)}_{a_\mu \text{ times}} + \ldots$$

or, in compact form,

$$U(a) = \sum_\mu a_\mu U^\mu(a) . \tag{5.193}$$

$U(a)$ is called the *direct-sum representation* of the irreducible representations $U^\mu(a)$.

Direct-product representation

Consider the representations defined by $U^\mu : G \longrightarrow \mathrm{GL}(\mathbb{V}^\mu)$, $\mu \in \mathbb{N}$. We define the *direct-product representation* as

$$\left[U^{\mu\times\nu}(a)\right] = \left[U^\mu(a)\right] \otimes \left[U^\nu(a)\right] . \tag{5.194}$$

Let us compute the character of the direct-product representation. We have

$$\begin{aligned} \chi^{\mu\times\nu}(a) &= \mathrm{tr}\left[U^{\mu\times\nu}(a)\right] \\ &= \left[U^{\mu\times\nu}(a)\right]^k{}_k \\ &= \left[U^\mu(a)\right]^i{}_i \left[U^\nu(a)\right]^j{}_j \qquad \text{by Equation (5.133)} \\ &= \chi^\mu(a)\,\chi^\nu(a) . \end{aligned}$$

Thus

$$\chi^{\mu\times\nu} = \chi^\mu \chi^\nu . \tag{5.195}$$

5.10 Properties of irreducible representations

5.10.1 Fundamental theorems of representations

Theorem 5.153 (Schur's first lemma)**.** *Let $\mathbb{V}$ be a vector space over $\mathbb{C}$ with $0 < \dim \mathbb{V} < \infty$, let $U : G \longrightarrow \mathrm{GL}(\mathbb{V})$ be an irreducible representation (Definition 5.143), and let $A : \mathbb{V} \longrightarrow \mathbb{V}$ be a linear operator commuting with every operator of the representation, $[U(a), A] = 0$ for every $a \in G$. Then A is proportional to the identity operator:*

$$A = \lambda \mathbb{1} , \qquad \lambda \in \mathbb{C} . \tag{5.196}$$

Rests on Definitions 5.35, 5.142 and 5.143.

Proof. *Derives Theorem 5.153.*

Both hypotheses on $\mathbb{V}$ are used at once, and only once: because $\dim \mathbb{V}$ is finite and non-zero, A has, in any basis, a characteristic polynomial $\det([A] - \lambda \mathbb{1})$ of degree $\dim \mathbb{V} \geq 1$, and because $\mathbb{C}$ is algebraically closed that polynomial has a root λ. For that λ the operator $A - \lambda\mathbb{1}$ is not invertible, so its kernel

$$\mathbb{E} = \ker(A - \lambda\mathbb{1}) = \{v \in \mathbb{V} \mid Av = \lambda v\}$$

is a subspace different from the null space: an eigenvector exists, which is precisely what a real or an infinite-dimensional space would not guarantee.

$\mathbb{E}$ is invariant under the representation. Indeed, for $v \in \mathbb{E}$ and any $a \in G$,

$$\begin{aligned} (A - \lambda\mathbb{1})\, U(a)v &= U(a)\,(A - \lambda\mathbb{1})\, v && \text{by } [U(a), A] = 0 \\ &= U(a)\, 0_{\mathbb{V}} \\ &= 0_{\mathbb{V}}\,, \end{aligned}$$

so $U(a)v \in \mathbb{E}$.

Irreducibility (Definition 5.143) leaves an invariant subspace only the two trivial possibilities, and $\mathbb{E}$ is not the null space; hence $\mathbb{E} = \mathbb{V}$. That is $Av = \lambda v$ for every $v \in \mathbb{V}$, which is Equation (5.196). Note that nothing was assumed about A beyond linearity — in particular it was not assumed diagonalizable, and no eigenbasis was used, only one eigenvalue. □

Remark 5.154 (The hypotheses are not decoration). Drop the algebraic closure of the field and the conclusion is false, not merely unproved. Take $G = \mathrm{SO}(2)$ acting on $\mathbb{V} = \mathbb{R}^2$ by rotations. That representation is irreducible over $\mathbb{R}$, since a rotation by an angle that is not a multiple of π leaves no line fixed; yet the group is abelian, so every $U(\theta_0)$ commutes with every $U(\theta)$ and is an intertwining operator, and no rotation other than $U(0) = \mathbb{1}$ and $U(\pi) = -\mathbb{1}$ is a multiple of the identity. The same failure in its algebra form, with the same matrix, is recorded in Remark 5.160. The moral is worth stating plainly because it is easy to violate at a distance: a chapter that invokes this lemma about a *real* representation — and the representations that physics writes down are frequently real — is invoking a statement that does not hold. The finite dimension is equally load-bearing: on an infinite-dimensional space an operator need have no eigenvalue at all, which is why the unitary representation theory of non-compact groups (Remark 5.151) is a different subject rather than a longer version of this one.

Theorem 5.155 (Schur's second lemma)**.** *Consider the irreducible representations $U : G \longrightarrow \mathrm{GL}(\mathbb{V})$ and $U' : G \longrightarrow \mathrm{GL}(\mathbb{V}')$, and an operator $A : \mathbb{V}' \longrightarrow \mathbb{V}$ such that $AU'(a) = U(a)A$ for all a, with $A \neq 0$. Then $\mathbb{V}$ and $\mathbb{V}'$ are isomorphic and, moreover, U and U' are equivalent.*
Rests on Definitions 5.51, 5.140, 5.142 and 5.143.

Proof. *Derives Theorem 5.155.*

We proceed in two steps.

- First note that the image and the kernel of the operator A are the subspaces

$$\begin{aligned} \mathrm{im}(A) &= \{|v\rangle \in \mathbb{V} \mid |v\rangle = A\,|v'\rangle\} \leq \mathbb{V}\,, \\ \ker(A) &= \{|v'\rangle \in \mathbb{V}' \mid A\,|v'\rangle = |0\rangle\,,\ 0 \in \mathbb{V}\} \leq \mathbb{V}'\,. \end{aligned}$$

  The image of A is an invariant subspace with respect to U, since

$$\begin{aligned} U(a)\,|v\rangle &= U(a)A\,|v'\rangle \\ &= AU'(a)\,|v'\rangle\,, \end{aligned}$$

and since $U'(a)\,|v'\rangle \in \mathbb{V}'$, we have $AU'(a)\,|v'\rangle \in \mathrm{im}(A)$, and hence $U(a)\,|v\rangle \in \mathrm{im}(A)$. As U is irreducible, $\mathrm{im}(A)$ is a trivial invariant subspace: the image of A would be the null space if and only if $A = 0$; otherwise one may have $\mathrm{im}(A) = \mathbb{V}$.

On the other hand, the kernel of A is an invariant subspace with respect to U'. This is because

$$\begin{aligned} AU'(a)\,|v'\rangle &= U(a)A\,|v'\rangle \\ &= 0 \in \mathbb{V}\,, \end{aligned}$$

so $U'(a)\,|v'\rangle \in \ker(A)$. Since we are assuming that U' is irreducible, the kernel of A is a trivial invariant subspace: again, the kernel of A would be the whole space $\mathbb{V}'$ if and only if $A = 0$; otherwise the kernel of A is the null space.

From these observations we may summarize: either $A = 0$, or the kernel of A is the null space and the image of A is the whole codomain, which tells us that A is a bijection establishing an isomorphism between the vector spaces $\mathbb{V}'$ and $\mathbb{V}$.

- Moreover, since A is bijective its inverse exists. Therefore, operating on the right with A^{-1},

$$\begin{aligned} AU'(a) &= U(a)A \\ AU'(a)A^{-1} &= U(a)\,, \end{aligned}$$

so $U(a)$ and $U'(a)$ are equivalent.

□

Proposition 5.156 (Irreducible representations of an abelian group)**.** *The irreducible representations of an abelian group on a finite-dimensional complex vector space are one-dimensional. The hypotheses are inherited from Theorem 5.153 and cannot be dropped: over* $\mathbb{R}$ *the rotations* $\mathrm{SO}(2)$ *form an abelian group with a two-dimensional irreducible representation, as Remark 5.154 records.*

Rests on Theorem 5.153 and Definition 5.143.

Proof. *Derives Proposition 5.156.*

Consider an irreducible representation $U : G \longrightarrow \mathrm{GL}(\mathbb{V})$ of an abelian group G, with $\mathbb{V}$ complex and $0 < \dim \mathbb{V} < \infty$. Let $a, b \in G$. Applying U to $ab = ba$,

$$\begin{aligned} U(ab) &= U(ba) \\ U(a)U(b) &= U(b)U(a)\,. \end{aligned}$$

By Schur's first lemma (Theorem 5.153), applied to the operator $U(a)$, which commutes with every $U(b)$,

$$U(a) = \lambda_a \mathbb{1}\,,$$

so that

$$U(a)\,|v\rangle = \lambda_a \mathbb{1}\,|v\rangle$$

and thus

$$U(a)\,|v\rangle = \lambda_a\,|v\rangle\,, \quad \forall\ |v\rangle \in \mathbb{V}\,.$$

This does not mean that all the vectors of $\mathbb{V}$ are eigenvectors, but rather that the irreducible representations of an abelian group are one-dimensional. □

5.10.2 Orthonormality relations

Orthonormality of irreducible inequivalent representations

Consider an irreducible representation $U : G \longrightarrow \mathrm{GL}(\mathbb{V})$ of a finite group $G = \{g_i\}_{i=1}^{n_G}$ and an arbitrary operator X. Define

$$A(X) = \sum_{i=1}^{n_G} U(g_i) X U(g_i^{-1}) \,.$$

Let us compute the following:

$$\begin{aligned}
U(a)A &= \sum_{i=1}^{n_G} U(a)U(g_i)XU(g_i^{-1}) \\
&= \sum_{i=1}^{n_G} U(a)U(g_i)XU(g_i^{-1}) \Big(U(a^{-1})U(a)\Big) \\
&= \sum_{i=1}^{n_G} U(ag_i)XU(g_i^{-1}a^{-1})U(a) \\
&= \sum_{i=1}^{n_G} U(ag_i)XU((ag_i)^{-1})U(a) \\
&= \sum_{i=1}^{n_G} U(g_{a_i})XU(g_{a_i}^{-1})U(a) \qquad \text{by Equation (4.28)} \,.
\end{aligned}$$

Since it makes no difference whether we sum over i or over a_i — it is a permutation — we may relabel the indices, so that

$$U(a)A = \underbrace{\sum_{i=1}^{n_G} U(g_i)XU(g_i^{-1})}_{A}\, U(a) \,,$$

whence $[U(a), A] = 0$.

On this basis we may use Schur's first lemma (as we are assuming U irreducible), which tells us that

$$A(X) = \lambda(X)\, \mathbb{1} \,. \tag{5.197}$$

Since X is arbitrary, we may choose it as follows:

$$X = \begin{pmatrix}
 & & l & & & \\
0 & \dots & 0 & \dots & 0 & \\
\vdots & \ddots & \vdots & & & \\
0 & \dots & 1 & \dots & 0 & m \\
\vdots & & \vdots & \ddots & & \\
0 & & 0 & & 0 &
\end{pmatrix} ,$$

that is, such that only $X^l{}_m = 1$, all other entries being zero. Identify the coefficients $\lambda(X)$ as $\lambda(X) = \lambda^m_l$. With this, writing the definition of A in index notation and using the result Equation (5.197),

$$\begin{aligned}
\lambda^m_l \delta^j_k &= \sum_{i=1}^{n_G} (D(g_i))^j{}_l \, X^l{}_m \left(D(g_i^{-1})\right)^m{}_k \\
&= \sum_{i=1}^{n_G} (D(g_i))^j{}_l \left(D(g_i^{-1})\right)^m{}_k \,.
\end{aligned} \tag{5.198}$$

Evaluating at $k = j$ and summing,

$$\begin{aligned}\lambda^m{}_l\,\delta^j{}_j &= \sum_{i=1}^{n_G} (D(g_i))^j{}_l \left(D(g_i^{-1})\right)^m{}_j \\ \lambda^m{}_l\,n_U &= \sum_{i=1}^{n_G} \left(D(g_i^{-1})D(g_i)\right)^m{}_l \\ &= \sum_{i=1}^{n_G} (D(e))^m{}_l\ ,\end{aligned}$$

where n_U is the dimension of the representation U (equal, by its definition, to the order of the operator A). Since D is a homomorphism, from Equation (4.24) we see that $U(e) = \mathrm{id}_{\mathrm{GL}(\mathbb{V})}$, that is, $D(e) = \mathbb{1}$. Thus

$$\begin{aligned}\lambda^m{}_l\,n_U &= \sum_{i=1}^{n_G} \delta^m{}_l \\ &= n_G\,\delta^m{}_l\,,\end{aligned}$$

so the coefficients $\lambda^m{}_l$ are given by

$$\lambda^m{}_l = \frac{n_G}{n_U}\,\delta^m{}_l\,. \tag{5.199}$$

Substituting Equation (5.199) into Equation (5.198) we obtain

$$\sum_{i=1}^{n_G} (D(g_i))^j{}_l \left(D(g_i^{-1})\right)^m{}_k = \frac{n_G}{n_U}\,\delta^m{}_l\,\delta^j{}_k\,. \tag{5.200}$$

From Equations (5.198) and (5.199) (substitute the coefficients fixed by the trace back into the intertwining identity).

Now, had we constructed A distinguishing the irreducible representations,

$$A(X) = \sum_{i=1}^{n_G} U^\mu(g_i) X U_\nu(g_i^{-1})\,,$$

then, by a procedure analogous to the preceding one, we would arrive at

$$U^\mu(a)A = AU_\nu(a)\,,$$

and we must therefore place ourselves in the following cases:

1. if $\mu = \nu$ we recover the preceding case;
2. if $\mu \neq \nu$ we may use Schur's second lemma (Theorem 5.155): then either the representations $U^\mu(a)$ and $U_\nu(a)$ are equivalent, or $A = 0$ and the representations are inequivalent.

Consider, then, inequivalent representations (irreducible by hypothesis). We can synthesize the above in Equation (5.200) as

$$\frac{n_\mu}{n_G} \sum_{i=1}^{n_G} (D^\mu(g_i))^j{}_l \left(D_\nu(g_i^{-1})\right)^m{}_k = \delta^m{}_l\,\delta^j{}_k\,\delta^\mu{}_\nu\,, \tag{5.201}$$

where $\delta^\mu{}_\nu$ appears in order to annihilate A in the case $\mu \neq \nu$ (the representations being taken inequivalent), and where in the case $\mu = \nu$ one has $n_{U^\mu} = n_{U^\nu}$, so we make the replacement $n_U \longrightarrow n_\mu$.

We must note that the indices μ and ν only label the Kronecker delta, so $\mu = \nu$ does not indicate summation.

Equation (5.201) expresses the *orthonormality of the irreducible, inequivalent representations* a group may have.

Orthonormality of characters of irreducible unitary representations

As we have seen, the character of a representation actually identifies a whole class of equivalent representations, so in this context we need not insist on inequivalent representations.

Characters describe the intrinsic structure of a representation better, because they are basis-independent; not so the representations themselves, since the components of $D(a)$ depend on the chosen basis.

Consider then the irreducible representations U^μ. Since the matrix representation of a representation is also a homomorphism, from Equation (4.25) and Equation (5.201) we have

$$\frac{n_\mu}{n_G}\sum_{i=1}^{n_G}(D^\mu(g_i))^j{}_l\left((D_\nu(g_i))^{-1}\right)^m{}_k=\delta^m_{\ l}\,\delta^j{}_k\,\delta^\mu{}_\nu\,.$$

If we further take the representations to be unitary,

$$\frac{n_\mu}{n_G}\sum_{i=1}^{n_G}(D^\mu(g_i))^j{}_l\left((D_\nu(g_i))^{\dagger}\right)^m{}_k=\delta^m_{\ l}\,\delta^j{}_k\,\delta^\mu{}_\nu\,.$$

Evaluating at $l=j$ and $m=k$ and summing over both,

$$\frac{n_\mu}{n_G}\sum_{i=1}^{n_G}(D^\mu(g_i))^j{}_j\left((D_\nu(g_i))^{\dagger}\right)^k{}_k=\delta^k{}_j\,\delta^j{}_k\,\delta^\mu{}_\nu$$

$$\frac{n_\mu}{n_G}\sum_{i=1}^{n_G}\chi^\mu(g_i)\chi^*_\nu(g_i)=n_\mu\,\delta^\mu{}_\nu\,,$$

and the factor n_μ cancels on both sides ($\delta^k{}_j\delta^j{}_k=\delta^j{}_j=n_\mu$), leaving

$$\frac{1}{n_G}\sum_{i=1}^{n_G}\chi^\mu(g_i)\chi^*_\nu(g_i)=\delta^\mu{}_\nu\,.$$

Grouping into n_C conjugacy classes and denoting by n_s the number of elements of the s-th class, we convert the sum over each element of the group into a sum over each class. Thus

$$\frac{1}{n_G}\sum_{s=1}^{n_C}n_s\,\chi^\mu_s\chi^*_{s\nu}=\delta^\mu{}_\nu\,,\tag{5.202}$$

where χ^μ_s denotes the character of the μ-th irreducible representation on the s-th class.

5.10.3 Decomposition into irreducible representations

Coefficients of the direct-sum representation

By means of the orthonormality relations we can compute the coefficients of the direct-sum representation appearing in Equation (5.193).

Evaluating the trace on the s-th class of the group,

$$\chi_s=\sum_\mu a_\mu\chi^\mu_s\,.\tag{5.203}$$

Multiplying the equation by $n_s\chi^*_{s\nu}$ and summing over s for all the classes,

$$\begin{aligned}\sum_{s=1}^{n_C}n_s\,\chi_s\chi^*_{s\nu}&=\sum_\mu a_\mu\sum_{s=1}^{n_C}n_s\,\chi^\mu_s\chi^*_{s\nu}\\&=\sum_\mu a_\mu\,n_G\,\delta^\mu{}_\nu&&\text{by Equation (5.202)}\\&=n_G\,a_\nu\,,\end{aligned}$$

so that

$$a_\mu = \frac{1}{n_G} \sum_{s=1}^{n_C} n_s \, \chi_s \chi^*_{s\mu} \, . \tag{5.204}$$

From Equations (5.202) and (5.203) (weight the class decomposition of the character by $n_s \chi^*_{s\nu}$ and sum over the classes).

Completeness relation

We can compute the number of times an irreducible representation appears in the regular representation. From Equation (5.204),

$$a^R_\mu = \frac{1}{n_G} \sum_{s=1}^{n_C} n_s \, \chi^R_s \chi^*_{s\mu} \, ,$$

but from Equation (5.191) and Equation (5.192) we see that the only contribution comes from the class of the neutral element. That is,

$$\begin{aligned} a^R_\mu &= \frac{1}{n_G} \, n_e \, \chi^R_e \, \chi^*_{e\mu} \\ &= n_e \, \chi^*_{e\mu} \qquad\qquad \text{by Equation (5.191)} \, , \end{aligned}$$

and we know, moreover, that the class of the neutral element contains only the neutral element, that is, $n_e = 1$. On the other hand,

$$\chi^*_{e\mu} = \operatorname{tr} \left((D^\mu(e))^\dagger \right) \, ,$$

but since the matrix representation of a representation is also a homomorphism, $D^\mu(e) = \mathbb{1}$, so

$$\chi_{e\mu} = n_\mu, \quad \chi^*_{e\mu} = n_\mu \, , \tag{5.205}$$

and there results

$$a^R_\mu = n_\mu \, . \tag{5.206}$$

From Equations (5.191), (5.192), (5.204) and (5.205) (only the class of the neutral element contributes to the multiplicity).

Now, writing the regular representation as a direct sum of irreducible representations, we have — in particular for the neutral element of G —

$$\begin{aligned} \Delta_e &= \sum_\mu a^R_\mu U^\mu(e) \\ \implies \chi^R_e &= \sum_\mu a^R_\mu \chi_{e\mu} \\ &= \sum_\mu n^2_\mu \qquad\qquad \text{by Equations (5.205) and (5.206)} \, . \end{aligned}$$

Hence, from Equation (5.191) we arrive at

$$n_G = \sum_\mu n^2_\mu \, . \tag{5.207}$$

From Equations (5.191), (5.205) and (5.206) (evaluate the direct-sum decomposition of the regular representation at the neutral element).

Equation (5.207) is a completeness relation for the irreducible representations.

Irreducibility criterion

To evaluate whether a given representation is irreducible, we can find a condition that the character of a representation must satisfy with respect to the equivalence classes of the group.

From Equation (5.203) we have, taking the complex conjugate,

$$\chi_s = \sum_\mu a_\mu \chi_s^\mu \,,$$
$$\chi_s^* = \sum_\nu a_\nu \chi_{s\nu}^* \,.$$

Multiplying the two equations member by member, weighting by n_s, and summing over s for all the equivalence classes,

$$\begin{aligned}
\sum_{s=1}^{n_C} n_s \, \chi_s \chi_s^* &= \sum_\mu a_\mu \sum_\nu a_\nu \sum_{s=1}^{n_C} n_s \, \chi_s^\mu \chi_{s\nu}^* \\
\sum_{s=1}^{n_C} n_s \, |\chi_s|^2 &= \sum_\mu a_\mu \sum_\nu a_\nu \, n_G \, \delta^\mu{}_\nu \qquad \text{by Equation (5.202)} \\
&= n_G \sum_\mu a_\mu^2 \,,
\end{aligned}$$

so that

$$\sum_{s=1}^{n_C} n_s \, |\chi_s|^2 = n_G \sum_\mu a_\mu^2 \,. \tag{5.208}$$

From Equations (5.202) and (5.203) (multiply the class decomposition by its conjugate and sum over the classes).

Now, if we take $U(a)$ to be an irreducible representation, then its direct-sum decomposition (see Equation (5.193)) holds with $a_\mu = 1$ for a single μ, the others being 0. Thus

$$\sum_{s=1}^{n_C} n_s \, |\chi_s|^2 = n_G \tag{5.209}$$

is the condition that an irreducible representation must verify.

Clebsch–Gordan coefficients

The direct-product representation Equation (5.194) is in general reducible even when both factors are irreducible: the product of two characters is not a character of an irreducible representation. The apparatus of this section decomposes it. From Equation (5.195) the character of $U^{\mu\times\nu}$ on the s-th class is $\chi_s^\mu \chi_s^\nu$, so Equation (5.204) gives the multiplicity with which the irreducible representation U^λ occurs in it:

$$a_{\mu\nu}^\lambda = \frac{1}{n_G} \sum_{s=1}^{n_C} n_s \, \chi_s^\mu \chi_s^\nu \chi_{s\lambda}^* \,. \tag{5.210}$$

From Equations (5.195) and (5.204) (the character of the product representation is the product of the characters, read into the multiplicity formula).

These non-negative integers are the *Clebsch–Gordan series* of the group; they answer the question "which irreducible representations appear, and how often". Equation (5.184) then reads, for the product space,

$$\mathbb{V}^\mu \otimes \mathbb{V}^\nu = \bigoplus_\lambda \bigoplus_{\alpha=1}^{a_{\mu\nu}^\lambda} \mathbb{V}^{\lambda,\alpha} \,, \tag{5.211}$$

the second index α distinguishing the $a^{\lambda}_{\mu\nu}$ copies of one and the same $\mathbb{V}^{\lambda}$.

What Equation (5.211) does *not* give is the vectors. There are now two natural bases of the same space:

- the *product*, or uncoupled, basis $\{|\mu i\rangle \otimes |\nu j\rangle\}$, built from bases of the two factors, in which the group acts by Equation (5.194) — the natural basis if one thinks of two independent systems;
- the *coupled* basis $\{|\lambda\alpha; k\rangle\}$ adapted to Equation (5.211), in which the matrix of every $U^{\mu\times\nu}(a)$ is block diagonal with irreducible blocks — the natural basis if one thinks of the composite system as carrying a symmetry of its own.

The *Clebsch–Gordan coefficients* are the entries of the change of basis between them:

$$|\lambda\alpha; k\rangle = \sum_{i,j} \langle \mu i; \nu j | \lambda\alpha k\rangle \; |\mu i\rangle \otimes |\nu j\rangle \; . \tag{5.212}$$

They are pure numbers, carrying no SI unit whatever — a change of basis cannot — however they are used in physics.

Two properties follow from the general theory and require no further computation. First, if the representations are unitary (Proposition 5.150 makes this no restriction for a finite group) and both bases are orthonormal, then the matrix of coefficients is *unitary*, being the matrix of an operator carrying one orthonormal basis to another — the argument used in Step 3 of Proposition 5.150. Its unitarity is precisely the pair of relations

$$\sum_{i,j} \langle \mu i; \nu j | \lambda\alpha k\rangle^{*} \langle \mu i; \nu j | \lambda'\alpha' k'\rangle = \delta^{\lambda}{}_{\lambda'} \delta_{\alpha\alpha'} \delta_{kk'} \; , \tag{5.213}$$

$$\sum_{\lambda,\alpha,k} \langle \mu i; \nu j | \lambda\alpha k\rangle \langle \mu i'; \nu j' | \lambda\alpha k\rangle^{*} = \delta_{ii'} \delta_{jj'} \; , \tag{5.214}$$

orthonormality of the coupled basis and completeness of the product one. Second, the coefficients are *not* determined by the group alone. Equation (5.211) fixes the subspaces only up to an arbitrary unitary mixing within each multiplicity label α and an overall phase on each irreducible block — the freedom left by Schur's first lemma (Theorem 5.153), which pins an intertwining operator down only to a multiple of the identity. Tables of Clebsch–Gordan coefficients therefore always come attached to a phase convention; the one universally used in physics is that of Condon and Shortley [751].

For the groups this treatise actually computes with, the coefficients are tabulated and closed forms exist in the literature: for SU(2), whose Clebsch–Gordan series is the rule for adding angular momenta, Racah's closed expression [867] and the recoupling coefficients that follow it [873]. Where that development sits in this book has to be stated carefully, because the two halves of it are in different states. Which irreducible representations occur, and with what multiplicity, is proved for SU(2) in Theorem 14.47, where the group is continuous and its irreducible representations have been classified. The coefficients *themselves*, and with them the physics — the addition of angular momenta and the Wigner–Eckart theorem — belong to Section 79.4.2, which at the time of writing carries its heading and its editorial specification but not the construction; that chapter reserves it explicitly, and nothing in the present chapter depends on it. What is established here is only what holds for any finite group: the multiplicities Equation (5.210) are computable from characters alone, and the coefficients Equation (5.212) exist and form a unitary matrix. The systematic use of both in physics goes back to Wigner [687].

5.11 Representations of algebras

Everything so far has represented a *group*: a set closed under a product, with inverses. An algebra is a different object — a vector space with a bracket — and representing it means something correspondingly different. The distinction is not pedantry: it is the reason angular momenta add while rotations multiply, and getting it backwards produces expressions that are not representations of anything.

5.11.1 Basic properties of the representations

Definition 5.157 (Representation of an algebra)**.** Let A be an algebra over $\mathbb{K}$ with bracket $[\,,\,]$ (Definition 5.117) and let $\mathbb{V}$ be a vector space over $\mathbb{K}$. A *representation* of A on $\mathbb{V}$ is a *linear* map

$$\rho : A \longrightarrow \mathcal{L}(\mathbb{V})$$

that carries the bracket of A into the corresponding bracket of $\mathcal{L}(\mathbb{V})$:

$$\rho\left([X,Y]\right) = [\rho(X),\rho(Y)] = \rho(X)\rho(Y) - \rho(Y)\rho(X)\,, \qquad \forall\, X, Y \in A\,, \tag{5.215}$$

the commutator being the bracket $\mathcal{L}(\mathbb{V})$ carries by Proposition 5.121. (For an *associative* algebra, whose product is written as a juxtaposition rather than a bracket, the requirement is instead $\rho(XY) = \rho(X)\rho(Y)$; the Lie case Equation (5.215) is the one this treatise uses, and the two agree once an associative algebra is given its commutator bracket.) The *dimension* of ρ is $\dim \mathbb{V}$.

Rests on Definition 5.117 and Proposition 5.121.

How this differs from a group representation

The two definitions look similar and behave differently, in ways worth listing because each has a physical consequence.

1. *A representation of an algebra is linear in its argument; a representation of a group is multiplicative.* It makes sense to write $\rho(\alpha X + \beta Y) = \alpha\rho(X) + \beta\rho(Y)$, and it makes no sense at all to add two group elements. This is the formal expression of the fact that an algebra element is an infinitesimal generator and a group element a finite transformation.

2. *The image lies in $\mathcal{L}(\mathbb{V})$, not in* $\mathrm{GL}(\mathbb{V})$. Linearity forces $\rho(0_A) = \Theta$, the null operator of Equation (5.75), which is not invertible. More generally $\rho(X)$ may be singular for $X \neq 0_A$. A group representation, by contrast, takes values in the invertible operators only, since $U(a)U(a^{-1}) = U(e) = \mathrm{id}$.

3. *The bracket, not the product, is preserved.* A group representation preserves a product, Equation (5.180); an algebra representation preserves a commutator, Equation (5.215). This is what makes the structure constants the invariant content of an algebra representation, as Proposition 5.121 already showed for the defining one.

Everything the two notions *do* share is the vocabulary built on the underlying vector space, and it transfers unchanged.

Definition 5.158 (Faithful, equivalent, invariant)**.** A representation ρ of A on $\mathbb{V}$ is *faithful* if ρ is injective. Two representations ρ, ρ' on the same space are *equivalent* if there is $S \in \mathrm{GL}(\mathbb{V})$ with $\rho'(X) = S\rho(X)S^{-1}$ for every $X \in A$. A subspace $\mathbb{S} \leq \mathbb{V}$ is *invariant* if $\rho(X)\mathbb{S} \subset \mathbb{S}$ for every $X \in A$, and ρ is *irreducible* if the only invariant subspaces are the trivial ones.

Rests on Definitions 5.6 and 5.157.

Proposition 5.159 (The kernel is an ideal)**.** *Let ρ be a representation of the algebra A. Then*

$$\ker\rho = \{X \in A \mid \rho(X) = \Theta\}$$

is an ideal of A in the sense of Definition 5.120. Consequently, if A is simple *— its only ideals being $\{0_A\}$ and A — then every representation of A is either faithful or identically zero.*

Rests on Definitions 5.120, 5.157 and 5.158 and Equation (5.215).

Proof. *Derives Proposition 5.159.*

$\ker\rho$ is a subspace because ρ is linear. Let $X \in \ker\rho$ and $Y \in A$. Then, by Equation (5.215),

$$\begin{aligned}\rho([X,Y]) &= [\rho(X),\rho(Y)]\\ &= [\Theta,\rho(Y)]\\ &= \Theta\rho(Y) - \rho(Y)\Theta\\ &= \Theta\,,\end{aligned}$$

so $[X,Y] \in \ker\rho$; taking linear combinations, $[\ker\rho, A] \subset \ker\rho$, which is Definition 5.120. If A is simple the ideal $\ker\rho$ is either $\{0_A\}$, and ρ is injective, that is, faithful, or the whole of A, and ρ sends everything to Θ. □

This is the algebra counterpart of Definition 5.138 and it is sharper: for a group, the analogous statement is that the kernel of a representation is a normal subgroup, and a simple group likewise admits only faithful or trivial representations. Both are the reason the simple Lie algebras are classified before their representations are: the classification of the algebras is what makes the representation theory finite work.

The induced module structure

A representation is the same thing as an action, and writing it that way makes several statements shorter. Given ρ, define

$$X \cdot v = \rho(X)v\,, \qquad X \in A,\ v \in \mathbb{V}\,. \tag{5.216}$$

Then $\mathbb{V}$ becomes an *A-module*: the action is bilinear — linear in v because $\rho(X)$ is an LT, linear in X because ρ is linear — and it is compatible with the bracket in the sense that

$$[X,Y]\cdot v = X\cdot(Y\cdot v) - Y\cdot(X\cdot v)\ ,$$

which is Equation (5.215) read on a vector. In this language an invariant subspace is a *submodule*, an irreducible representation is a *simple module*, and the equivalence of Definition 5.158 is an isomorphism of modules. Nothing is added by the change of words except economy; it is used here only to say that the two Schur lemmas transfer.

Remark 5.160 (Schur's lemmas hold verbatim). The proofs of Theorems 5.153 and 5.155 use exactly three things: that the operators of the representation are linear, that the intertwining operator commutes with all of them, and that the invariant subspaces of an irreducible representation are trivial. None of the three mentions the group product, so both proofs apply word for word to a representation of an algebra — with one caveat that must be stated, since it is the only place the group was doing work: the proof of Theorem 5.155 concluded that A is a bijection from $\mathrm{im}(A) = \mathbb{V}$ and $\ker(A) = \{0\}$, which remains valid, whereas Theorem 5.153 carries the hypotheses $\mathbb{K} = \mathbb{C}$ and $0 < \dim\mathbb{V} < \infty$ — needed for

the eigenvalue its proof starts from to exist at all — and they must be carried over with it. Over $\mathbb{R}$ the conclusion genuinely fails: the algebra $\mathfrak{so}(2)$ acting on $\mathbb{R}^2$ by

$$\rho(X) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

is irreducible — a rotation generator leaves no real line fixed — yet $\rho(X)$ commutes with itself and is not a multiple of the identity.

Direct sums and tensor products

Let ρ_1 and ρ_2 be representations of the same algebra A on $\mathbb{V}_1$ and $\mathbb{V}_2$.

Direct sum. On $\mathbb{V}_1 \oplus \mathbb{V}_2$ define

$$(\rho_1 \oplus \rho_2)(X) = \begin{pmatrix} \rho_1(X) & 0 \\ 0 & \rho_2(X) \end{pmatrix} . \tag{5.217}$$

It is a representation: linearity in X is block by block, and the commutator of two block-diagonal matrices is block-diagonal with the blockwise commutators, so Equation (5.215) holds in each block. Traces add, $\operatorname{tr}(\rho_1 \oplus \rho_2)(X) = \operatorname{tr}\rho_1(X) + \operatorname{tr}\rho_2(X)$, exactly as characters did in Equation (5.193). This much is identical to the group case.

Tensor product. Here the two theories part company, and the difference is the single most consequential fact in this section. For groups the product representation is the product of the matrices, Equation (5.194). For algebras that rule is *wrong*: the map $X \mapsto \rho_1(X) \otimes \rho_2(X)$ is not even linear in X. The correct construction is

$$(\rho_1 \otimes \rho_2)(X) = \rho_1(X) \otimes \mathbb{1} + \mathbb{1} \otimes \rho_2(X) . \tag{5.218}$$

Derivation. *Derives Equation (5.218).*

Linearity in X is manifest, both terms being linear. For Equation (5.215), write $\sigma(X)$ for the right-hand side of Equation (5.218) and use that operators acting on different factors commute, $[P \otimes \mathbb{1}, \mathbb{1} \otimes Q] = 0$, together with $(P \otimes \mathbb{1})(P' \otimes \mathbb{1}) = PP' \otimes \mathbb{1}$:

$$\begin{aligned} [\sigma(X), \sigma(Y)] &= [\rho_1(X) \otimes \mathbb{1}, \rho_1(Y) \otimes \mathbb{1}] + [\mathbb{1} \otimes \rho_2(X), \mathbb{1} \otimes \rho_2(Y)] \\ &= [\rho_1(X), \rho_1(Y)] \otimes \mathbb{1} + \mathbb{1} \otimes [\rho_2(X), \rho_2(Y)] \\ &= \rho_1([X,Y]) \otimes \mathbb{1} + \mathbb{1} \otimes \rho_2([X,Y]) \qquad \text{by Equation (5.215)} \\ &= \sigma([X,Y]) , \end{aligned}$$

the two cross terms of the first line having cancelled in pairs. □

Remark 5.161 (Why generators add and group elements multiply). Equation (5.218) is not an arbitrary convention: it is forced by Equation (5.194) for the corresponding group. If a one-parameter family of group elements is written $U(t) = \exp(t\rho(X))$, then the product representation satisfies

$$U_1(t) \otimes U_2(t) = \left(\mathrm{e}^{t\rho_1(X)}\right) \otimes \left(\mathrm{e}^{t\rho_2(X)}\right) = \exp\left[t\left(\rho_1(X) \otimes \mathbb{1} + \mathbb{1} \otimes \rho_2(X)\right)\right],$$

the last step being legitimate because the two exponents commute. Differentiating at $t = 0$ recovers Equation (5.218): the multiplicative law of the group becomes an *additive* law for the generators. This is the algebraic content of the statement that angular momenta, and every other additive quantum number, add when systems are combined — the total generator of a composite system is the sum of the generators of its parts, each acting on its own factor. The exponential map that makes the step precise belongs to Chapter 14.

The adjoint representation

Every algebra represents itself, on its own underlying vector space, and the matrices of that representation are its structure constants. This is the content of the source's marginal note at this heading.

Proposition 5.162 (Adjoint representation)**.** *Let L be a Lie algebra (Definition 5.123) and define, for each $X \in L$, the map*

$$\begin{array}{rccc} \mathrm{ad}_X : & L & \longrightarrow & L \\ & Y & \longmapsto & \mathrm{ad}_X Y = [X, Y] \,. \end{array} \tag{5.219}$$

Then $X \longmapsto \mathrm{ad}_X$ is a representation of L on L itself, of dimension $\dim L$, and in a basis $\{e_i\}$ its matrices are

$$(\mathrm{ad}_{e_i})^k{}_j = C^k{}_{ij} \,. \tag{5.220}$$

Its kernel is the centre *of L, the set of elements commuting with everything.*

Rests on Definitions 5.123 and 5.157, Equation (5.150), and Proposition 5.43.

Proof. *Derives Proposition 5.162.*

Each ad_X is a linear transformation of L by right distributivity Equation (5.146) and the scalar rule Equation (5.147), so $\mathrm{ad}_X \in \mathcal{L}(L)$; and $X \longmapsto \mathrm{ad}_X$ is itself linear by left distributivity Equation (5.145). What must be verified is Equation (5.215), namely $\mathrm{ad}_{[X,Y]} = [\mathrm{ad}_X, \mathrm{ad}_Y]$. Applying both sides to an arbitrary Z,

$$\begin{aligned} [\mathrm{ad}_X, \mathrm{ad}_Y]\, Z &= \mathrm{ad}_X \,\mathrm{ad}_Y\, Z - \mathrm{ad}_Y \,\mathrm{ad}_X\, Z \\ &= [X, [Y, Z]] - [Y, [X, Z]] \\ &= [X, [Y, Z]] + [Y, [Z, X]] && \text{by Equation (5.155)} \\ &= -\,[Z, [X, Y]] && \text{by Equation (5.156)} \\ &= [[X, Y]\,, Z] && \text{by Equation (5.155)} \\ &= \mathrm{ad}_{[X,Y]}\, Z \,, \end{aligned}$$

so the adjoint map is a representation — and note what the derivation shows: *the Jacobi identity is exactly the statement that* ad *is a representation*, neither more nor less.

For the matrices, apply Equation (5.219) to a basis vector and compare with Equation (5.74) and the definition Equation (5.150) of the structure constants:

$$\mathrm{ad}_{e_i}\, e_j = [e_i, e_j] = C^k{}_{ij} e_k \,,$$

whose coefficients are, by Proposition 5.43, the matrix entries claimed in Equation (5.220).

Finally $\mathrm{ad}_X = \Theta$ means $[X, Y] = 0_L$ for every Y, which is the definition of the centre; by Proposition 5.159 the centre is an ideal, as it must be. □

Corollary 5.163. *The adjoint representation of a Lie algebra with trivial centre — in particular of a simple Lie algebra, which is by definition non-abelian and has no proper non-zero ideal — is faithful. Every such algebra is therefore isomorphic to an algebra of $\dim L \times \dim L$ matrices under the commutator.*

Rests on Propositions 5.121, 5.159 and 5.162 and Definition 5.158.

Proof. *Derives Corollary 5.163.*

Immediate from Proposition 5.162: the kernel is the centre, and a representation is faithful exactly when its kernel is trivial. For a simple L the centre is an ideal, hence $\{0_L\}$ or L; the second would make L abelian and is excluded. The matrices are those of Equation (5.220), and Proposition 5.121 guarantees they satisfy the same algebra. □

Back to unitary group representations

The two theories are joined by the exponential map, and the join is where the SI units of the treatise re-enter what has so far been pure algebra. Let G be a Lie group with algebra L, and let U be a representation of G that is unitary in the sense of Equation (5.183). Writing a one-parameter subgroup as $U(t) = \exp(t\rho(X))$, unitarity for every t requires

$$\rho(X)^\dagger = -\rho(X)\,, \tag{5.221}$$

that is, the algebra is represented by *anti-hermitian* operators: differentiate $U(t)^\dagger U(t) = \mathbb{1}$ at $t = 0$. Physics prefers hermitian operators, because those are the ones with real spectra and hence the observable ones, so one writes $\rho(X) = -\mathrm{i}\Lambda/\hbar$ with $\Lambda^\dagger = \Lambda$ — and this is where the units are fixed.

Remark 5.164 (The generator of time translations, in SI). Take G to be the group of time translations of the $3+1$ spacetime of Chapter 39, acting on the state space of a quantum system. Its unitary representation is

$$U(t) = \exp\left(-\frac{\mathrm{i}}{\hbar}\mathcal{H}t\right),$$

with $\mathcal{H}$ the Hamiltonian, of SI unit the joule, t in seconds, and $\hbar = h/2\pi = 1.054\,571\,817 \times 10^{-34}\,\mathrm{J\,s}$ the reduced Planck constant — h being exact by the SI definition of the kilogram, so that $\hbar$ is exact as a multiple of $1/2\pi$ and the decimal shown is rounded (Chapter 77 and the constants appendix). The exponent $\mathcal{H}t/\hbar$ is then a pure number, as the argument of an exponential must be, and the algebra element is represented by $\rho(X) = -\mathrm{i}\mathcal{H}/\hbar$, whose SI unit is $1/\mathrm{s}$ — an inverse time, which is what a generator of time translations has to be. The same bookkeeping applies to every generator in this treatise: an angular-momentum operator is $\rho(X) = -\mathrm{i}J/\hbar$ with J in $\mathrm{J\,s}$, so that $\rho(X)$ is dimensionless, rotation angles being dimensionless too. The habit of setting $\hbar = 1$, universal in the representation-theory literature, is what hides this; the treatise does not adopt it (editorial rule 3).

Chapter 6

Topological and Metric Spaces

Topology isolates the minimal structure a set must carry for the notions of nearness, continuity, and limit to make sense at all, before any notion of distance is introduced. Metric spaces then recover distance as an additional layer of structure, and the topology of $\mathbb{R}^N$ is the concrete arena in which the analysis of the neighbouring chapters takes place.

6.1 Topological spaces

6.1.1 Basic definitions

Definition 6.1 (Topological space)**.** Consider a set X, and consider further a family $T = \{U_i\}_{i\in I}$ of subsets $U_i \subset X$. We say that (X, T) is a *topological space* with *topology* T if and only if it satisfies the following axioms:

1. $\varnothing \in T$ and $X \in T$;
2. the union of any collection of sets in T belongs to T;
3. the intersection of any finite collection of sets in T belongs to T.

Rests on Definitions 3.30, 3.31 and 3.35 and Equation (3.51)*.*

Definition 6.2 (Open set)**.** A set that belongs, as an element, to the topology of a space is called an *open set*.
Rests on Definition 6.1.

Definition 6.3 (Closed set)**.** A set is *closed* in a topological space if and only if its complement in that space is an open set.
Rests on Definitions 3.35 and 6.2.

Definition 6.4 (Neighbourhood)**.** A *neighbourhood* of a point $x \in X$ is any open set containing x.
Rests on Definition 6.2.

Definition 6.5 (Open cover)**.** A family $\{U_i\}_{i\in I}$ of open sets in X is an *open cover* of X if and only if the union of all of them is the whole space, that is,

$$\bigcup_{i\in I} U_i = X\,. \tag{6.1}$$

An open cover of a subset $K \subseteq X$ is a family of open sets whose union contains K; a *subcover* is a subfamily that still covers.
Rests on Definition 6.2 and Equation (3.51)*.*

6.1.2 Continuity

Continuity is the property that topology is built to express: a map between topological spaces is continuous when it respects the open-set structure of the two spaces, as the following definition makes precise.

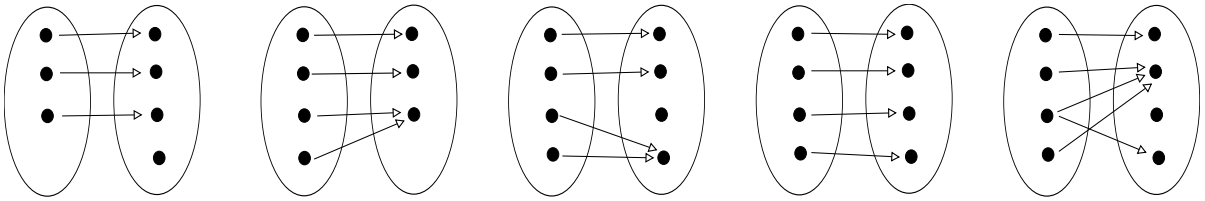


Figure 6.1: Schematic representation of maps between sets: a map assigns to each element of its domain exactly one element of its codomain. Continuity constrains how such maps interact with the topologies of the two spaces.

Definition 6.6 (Continuous map)**.** Consider two topological spaces X_1 and X_2 with topologies T_1 and T_2, and consider further a map $f : X_1 \longrightarrow X_2$. We say that f is *continuous* if and only if the inverse image of every open set in X_2 is an open set in X_1; that is,

$$\forall\, U \in T_2, \quad f^{-1}(U) \in T_1\,. \tag{6.2}$$

Rests on Definitions 3.53, 6.1 and 6.2.

Definition 6.7 (Homeomorphism)**.** A *homeomorphism* is a bijective continuous map $f : X_1 \longrightarrow X_2$ whose inverse f^{-1} is also continuous. Two spaces related by a homeomorphism are *homeomorphic*: they carry identical topological structure, and every property expressed purely in terms of open sets (compactness, connectedness, ...) transfers between them. Topology is the study of what homeomorphisms preserve.

Rests on Definitions 3.47, 3.51 and 6.6.

Remark 6.8. Bijectivity plus continuity does *not* suffice: the map $t \longmapsto (\cos t, \sin t)$ from $[0, 2\pi)$ onto the circle is continuous and bijective, but its inverse is discontinuous at the seam — the interval and the circle are not homeomorphic (removing a point disconnects one but not the other; see Section 6.1.4).

6.1.3 Compactness

Definition 6.9 (Compact set)**.** A subset K of a topological space is *compact* if and only if every open cover of K admits a finite subcover.

Rests on Definition 6.5.

Proposition 6.10 (Continuous images of compact sets)**.** *If $f : X_1 \longrightarrow X_2$ is continuous and $K \subseteq X_1$ is compact, then $f(K)$ is compact.*

Rests on Definitions 6.6 and 6.9.

Proof. *Derives Proposition 6.10.*

Let $\{U_i\}$ be an open cover of $f(K)$. The sets $f^{-1}(U_i)$ are open (Definition 6.6) and cover K, so finitely many $f^{-1}(U_{i_1}), \dots, f^{-1}(U_{i_n})$ cover K; then $U_{i_1}, \dots, U_{i_n}$ cover $f(K)$. □

The archetypal compact sets are the closed bounded intervals of the real line:

Theorem 6.11 (Heine–Borel on $\mathbb{R}$)**.** *The interval $[a, b] \subset \mathbb{R}$ is compact (in the topology of Section 6.3).*

Rests on Definition 6.9 and Axiom 7.1.

Proof. *Derives Theorem 6.11.*

Let $\{U_i\}$ be an open cover of $[a, b]$ and define

$$S = \{x \in [a, b] \mid [a, x] \text{ is covered by finitely many } U_i\} .$$

S is nonempty (a lies in some U_i) and bounded above by b, so $c = \sup S$ exists by the completeness of $\mathbb{R}$ (Axiom 7.1, proven in Appendix A.7). Some U_{i_0} contains c, and with it an interval $(c - \varepsilon, c + \varepsilon)$. Since $c - \varepsilon$ is not an upper bound of S, some $x \in S$ exceeds it; the finite subcover of $[a, x]$ together with U_{i_0} covers $[a, \min(c + \varepsilon/2, b)]$. If $c < b$ this puts points beyond c into S, contradicting $c = \sup S$; hence $c = b$, and the same finite family covers $[a, b]$. □

In $\mathbb{R}^N$ the theorem extends to the characterization that gives compactness its concrete meaning in analysis. Throughout, $B_r(\boldsymbol{x})$ denotes the Euclidean ball of Definition 6.26 for the metric Equation (6.12).

Theorem 6.12 (Heine–Borel in $\mathbb{R}^N$)**.** *A subset $K \subseteq \mathbb{R}^N$ is compact, in the topology of Section 6.3, if and only if it is closed and bounded.*

Rests on Definitions 6.3 and 6.9 and Axiom 7.1.

Proof. *Derives Theorem 6.12.*

Necessity. Let K be compact. The balls $B_n(\mathbf{0})$, $n \in \mathbb{N}$, form an open cover of K; a finite subcover is contained in the largest of its members, so K is bounded. For closedness, take $\boldsymbol{y} \notin K$ and set $r_{\boldsymbol{x}} = |\boldsymbol{x} - \boldsymbol{y}|/2 > 0$ for each $\boldsymbol{x} \in K$. The balls $B_{r_{\boldsymbol{x}}}(\boldsymbol{x})$ cover K, so finitely many of them, centred at $\boldsymbol{x}_1, \dots, \boldsymbol{x}_p$, already do; abbreviate $r_j = r_{\boldsymbol{x}_j}$ and put $\rho = \min_{j \le p} r_j > 0$. No point $\boldsymbol{z}$ can lie in both $B_\rho(\boldsymbol{y})$ and $B_{r_j}(\boldsymbol{x}_j)$, since that would give

$$2r_j = |\boldsymbol{x}_j - \boldsymbol{y}| \le |\boldsymbol{x}_j - \boldsymbol{z}| + |\boldsymbol{z} - \boldsymbol{y}| < r_j + \rho \le 2r_j .$$

Hence $B_\rho(\boldsymbol{y})$ misses K, and the complement of K is open.

Sufficiency. Consider first the closed box $Q_0 = [-R, R]^N$, and suppose some open cover $\{U_i\}$ of it admitted no finite subcover. Halving every edge splits Q_0 into 2^N closed boxes whose union is Q_0; were each of them covered by finitely many U_i, so would Q_0 be, so at least one of them — call it Q_1 — is not. Iterating produces nested closed boxes $Q_0 \supseteq Q_1 \supseteq \cdots$, none of them finitely covered, the k-th of edge $2R/2^k$ and hence of diameter $\sqrt{N}\, 2R/2^k$. In each coordinate the left endpoints of the Q_k form a nondecreasing bounded sequence and the right endpoints a nonincreasing one, the two differing by $2R/2^k$; by completeness (Axiom 7.1) each converges, and to the same number, which therefore lies in every one of the nested coordinate intervals. The point $\boldsymbol{x}$ with those N coordinates thus belongs to every Q_k. Some U_{i_0} contains $\boldsymbol{x}$, and with it a ball $B_\varepsilon(\boldsymbol{x})$; choosing k with $\sqrt{N}\, 2R/2^k < \varepsilon$ gives $Q_k \subseteq B_\varepsilon(\boldsymbol{x}) \subseteq U_{i_0}$, a subcover of Q_k by a single set — contradicting the choice of Q_k. So Q_0 is compact.

Let now K be closed and bounded, so that $K \subseteq Q_0$ for some R, and let $\{U_i\}$ be an open cover of K. Since K is closed, $\mathbb{R}^N \setminus K$ is open (Definition 6.3), so $\{U_i\}$ together with $\mathbb{R}^N \setminus K$ is an open cover of Q_0. Extract a finite subcover and discard $\mathbb{R}^N \setminus K$ from it: what remains is a finite subfamily of $\{U_i\}$, and it still covers K, because the discarded set contains no point of K. □

6.1.4 Connectedness

Definition 6.13 (Connected space)**.** We say that a topological space is *connected* if and only if it cannot be expressed as a union of disjoint nonempty open sets.

Rests on Definitions 3.36 and 6.2.

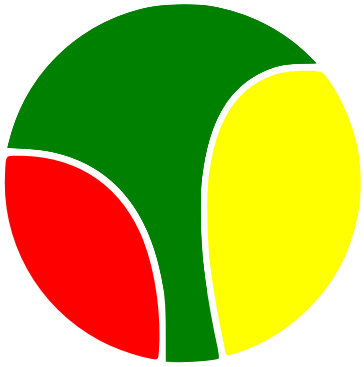


Figure 6.2: A set decomposed into pairwise disjoint subsets whose union is the whole set. A topological space that admits such a decomposition into disjoint nonempty open sets is disconnected.

Theorem 6.14 (Intervals are connected)**.** *Every interval of* $\mathbb{R}$ *is connected.*
Rests on Definition 6.13 and Axiom 7.1.

Proof. *Derives Theorem 6.14.*
Suppose $I = U \cup V$ with U, V open in I, disjoint and nonempty; pick $u \in U$, $v \in V$, say $u < v$. Let $c = \sup\{x \in U \mid x < v\}$, which exists by completeness and satisfies $u \le c \le v$, so $c \in I$. If $c \in U$, openness of U puts an interval around c inside U; since $c < v$ (as $v \in V$), points of U above c would contradict the definition of c. If $c \in V$, openness of V puts $(c - \varepsilon, c] \subseteq V$, making $c - \varepsilon$ an upper bound of the set defining c — again a contradiction. So c lies in neither, contradicting $I = U \cup V$. □

Definition 6.15 (Path-connected space)**.** X is *path-connected* iff any two points $x, y \in X$ are joined by a continuous map $\gamma : [0,1] \longrightarrow X$ with $\gamma(0) = x$, $\gamma(1) = y$.
Rests on Definition 6.6.

Proposition 6.16. *A path-connected space is connected.*
Rests on Definitions 6.13 and 6.15 and Theorem 6.14.

Proof. *Derives Proposition 6.16.*
If $X = U \cup V$ were a disconnection, join $x \in U$ to $y \in V$ by a path γ. Then $\gamma^{-1}(U)$ and $\gamma^{-1}(V)$ disconnect $[0,1]$: both are open by continuity, disjoint, nonempty (0 and 1), and cover. This contradicts Theorem 6.14. □

Definition 6.17 (Simply connected space)**.** A path-connected space X is *simply connected* iff every loop is contractible: for every continuous $\gamma : [0,1] \longrightarrow X$ with $\gamma(0) = \gamma(1) = x_0$ there is a continuous homotopy $H : [0,1] \times [0,1] \longrightarrow X$ with $H(s,0) = \gamma(s)$, $H(s,1) = x_0$, and $H(0,t) = H(1,t) = x_0$ for all s, t.
Rests on Definitions 6.6 and 6.15.

Example 6.18. $\mathbb{R}^N$, and any convex subset of it, is simply connected: the straight-line homotopy $H(s,t) = (1-t)\,\gamma(s) + t\,x_0$ stays inside a convex set and contracts any loop. The punctured plane $\mathbb{R}^2 \setminus \{0\}$ is not: a loop encircling the origin cannot be contracted without crossing it, as Proposition 6.23 proves — the fact underlying the residue theorem of Chapter 8 and property (3) of conservative fields in Chapter 7.
Rests on Definition 6.17.

6.1.5 The winding number

That a loop around a puncture cannot be contracted is a theorem, not a picture, and what proves it is an integer attached to every loop: the number of turns the loop makes about the puncture. Identify $\mathbb{R}^2$ with $\mathbb{C}$ (Chapter 8), so that a loop avoiding the origin is a continuous

map into $\mathbb{C} \setminus \{0\}$. The integer is built from a continuous choice of angle along the loop, and the whole argument rests on the fact that such a choice exists and is unique up to one additive constant.

Lemma 6.19 (Continuous argument along a path)**.** *Let $\gamma : [0,1] \longrightarrow \mathbb{C} \setminus \{0\}$ be continuous. There exists a continuous $\theta : [0,1] \longrightarrow \mathbb{R}$ with*

$$\gamma(t) = |\gamma(t)|\, \mathrm{e}^{\mathrm{i}\theta(t)}\,, \qquad t \in [0,1]\,, \tag{6.3}$$

and any two such functions differ by a constant lying in $2\pi\mathbb{Z}$.

Rests on Theorems 7.23 to 7.25.

Proof. *Derives Lemma 6.19.*

Existence. The function $t \longmapsto |\gamma(t)|$ is continuous and strictly positive on $[0,1]$, so it attains a minimum $m > 0$ (Theorem 7.24); and γ is uniformly continuous, because its real and imaginary parts are (Theorem 7.25). Uniform continuity supplies a $\delta > 0$ with $|\gamma(t) - \gamma(s)| < m$ whenever $|t - s| < \delta$; choose any partition $0 = t_0 < t_1 < \cdots < t_M = 1$ of mesh smaller than δ, so that $|\gamma(t) - \gamma(t_{j-1})| < m$ for every $t \in [t_{j-1}, t_j]$ and every j. On such an interval the quotient $w(t) = \gamma(t)/\gamma(t_{j-1})$ satisfies

$$|w(t) - 1| = \frac{|\gamma(t) - \gamma(t_{j-1})|}{|\gamma(t_{j-1})|} < \frac{m}{|\gamma(t_{j-1})|} \leq 1\,,$$

so $w(t)$ lies in the open disc of radius 1 centred at 1, hence in the half-plane $\Re w > 0$. On that half-plane the function

$$A(w) = \arctan \frac{\Im w}{\Re w} \tag{6.4}$$

is continuous, takes values in $(-\pi/2, \pi/2)$, vanishes at $w = 1$, and satisfies $w = |w|\, \mathrm{e}^{\mathrm{i}A(w)}$. Fix $\theta(0)$ with $\gamma(0) = |\gamma(0)|\, \mathrm{e}^{\mathrm{i}\theta(0)}$ and define recursively

$$\theta(t) = \theta(t_{j-1}) + A\left(\frac{\gamma(t)}{\gamma(t_{j-1})}\right), \qquad t \in [t_{j-1}, t_j]\,.$$

Each piece is continuous and, since $A(1) = 0$, the pieces agree at the nodes, so θ is continuous on $[0,1]$. Multiplying $\gamma(t_{j-1}) = |\gamma(t_{j-1})|\, \mathrm{e}^{\mathrm{i}\theta(t_{j-1})}$ by the polar form of the quotient gives Equation (6.3), one interval at a time.

Uniqueness up to $2\pi\mathbb{Z}$. If θ_1 and θ_2 both satisfy Equation (6.3), then $\mathrm{e}^{\mathrm{i}(\theta_1 - \theta_2)} = 1$, so $\varphi = (\theta_1 - \theta_2)/2\pi$ is continuous with values in $\mathbb{Z}$. Were it to take two distinct values, the intermediate value theorem (Theorem 7.23) would make it take some value strictly between two consecutive integers as well, which is not in $\mathbb{Z}$; hence φ is constant. □

Definition 6.20 (Winding number)**.** Let $\gamma : [0,1] \longrightarrow \mathbb{C} \setminus \{0\}$ be a continuous *loop*, $\gamma(1) = \gamma(0)$, and let θ be a continuous argument for it (Lemma 6.19). Since $\mathrm{e}^{\mathrm{i}\theta(1)} = \mathrm{e}^{\mathrm{i}\theta(0)}$, the number

$$n(\gamma) = \frac{\theta(1) - \theta(0)}{2\pi} \tag{6.5}$$

is an integer; by the uniqueness clause of Lemma 6.19 it does not depend on which continuous argument was chosen. It is the *winding number* of γ about the origin.

Rests on Lemma 6.19.

Example 6.21. The unit circle traversed once counterclockwise, $\gamma(s) = \mathrm{e}^{2\pi\mathrm{i}s}$, admits the continuous argument $\theta(s) = 2\pi s$, so $n(\gamma) = 1$. A constant loop admits a constant argument, so its winding number is 0.

Rests on Definition 6.20.

Lemma 6.22 (Nearby loops wind alike)**.** *Let $\gamma, \sigma : [0,1] \longrightarrow \mathbb{C} \setminus \{0\}$ be loops with*

$$|\sigma(s) - \gamma(s)| < |\gamma(s)| \,, \qquad s \in [0,1] \,. \tag{6.6}$$

Then $n(\sigma) = n(\gamma)$.
Rests on Definition 6.20 and Lemma 6.19.

Proof. *Derives Lemma 6.22.*
The quotient $w = \sigma/\gamma$ is a continuous loop, and Equation (6.6) says $|w(s) - 1| < 1$; in particular w never vanishes and takes its values in the half-plane $\Re w > 0$, where the function A of Equation (6.4) provides the continuous argument $s \longmapsto A(w(s))$. Hence $n(w) = \big(A(w(1)) - A(w(0))\big)/2\pi = 0$, because $w(1) = w(0)$. If θ is a continuous argument for γ, then $\theta + A(w)$ is one for $\sigma = w\,\gamma$, so $n(\sigma) = n(\gamma) + n(w) = n(\gamma)$. □

Proposition 6.23 (The punctured plane is not simply connected)**.** *The loop $\gamma(s) = \mathrm{e}^{2\pi \mathrm{i} s}$ in $\mathbb{R}^2 \setminus \{0\} = \mathbb{C} \setminus \{0\}$ admits no contraction in the sense of Definition 6.17. The punctured plane is therefore path-connected but not simply connected.*
Rests on Definition 6.17, Lemma 6.22, and Theorem 6.11.

Proof. *Derives Proposition 6.23.*
Path-connectedness is immediate: any two points of the punctured plane are joined by a circular arc followed by a radial segment, and neither meets the origin. Suppose then that $H : [0,1] \times [0,1] \longrightarrow \mathbb{C} \setminus \{0\}$ were a contraction of γ, that is, a continuous map with $H(s,0) = \gamma(s)$, $H(s,1) = 1$ and $H(0,t) = H(1,t) = 1$. For each fixed t the map $\gamma_t(s) = H(s,t)$ is then a loop in $\mathbb{C} \setminus \{0\}$, and we may set $n(t) = n(\gamma_t)$.

The function n is locally constant. Fix t_0 and let $m = \min_s |H(s,t_0)| > 0$, the minimum being attained by Theorem 7.24. For each $s_0 \in [0,1]$, continuity of H at (s_0,t_0) supplies a $\delta(s_0) > 0$ such that $|H(s,t) - H(s_0,t_0)| < m/2$ whenever $|s - s_0| < \delta(s_0)$ and $|t - t_0| < \delta(s_0)$. The open intervals $(s_0 - \delta(s_0)/2,\ s_0 + \delta(s_0)/2)$ cover $[0,1]$, which is compact (Theorem 6.11), so finitely many of them suffice, say those centred at $s_1, \ldots, s_k$; put $\delta = \min_j \delta(s_j)/2 > 0$. Let $|t - t_0| < \delta$ and let $s \in [0,1]$. Then $|s - s_j| < \delta(s_j)/2$ for some j, and $|t - t_0| < \delta \le \delta(s_j)$; the defining property of $\delta(s_j)$ applies both at t and at t_0, so $H(s,t)$ and $H(s,t_0)$ both lie within $m/2$ of $H(s_j,t_0)$ and

$$|H(s,t) - H(s,t_0)| \le |H(s,t) - H(s_j,t_0)| + |H(s_j,t_0) - H(s,t_0)| < m \le |H(s,t_0)| \,.$$

Lemma 6.22 then gives $n(t) = n(t_0)$ for every t with $|t - t_0| < \delta$.

Conclusion. A locally constant function is continuous, so $n : [0,1] \longrightarrow \mathbb{Z}$ is continuous and integer-valued, and the intermediate value theorem (Theorem 7.23) forces it to be constant, exactly as in Lemma 6.19. But $n(0) = 1$, since $\gamma_0 = \gamma$, and $n(1) = 0$, since γ_1 is the constant loop (Example 6.21). No such H can exist. □

The circle's winding number of 1 has an analytic counterpart in Chapter 8: the integral $\oint_C \mathrm{d}z/(z-a) = 2\pi\mathrm{i}$ of Lemma 8.11 measures the same single turn of the exponential by integration rather than by lifting.

6.2 Metric spaces

6.2.1 The metric

Definition 6.24 (Metric)**.** Let X be a set and let $d : X \times X \longrightarrow \mathbb{R}$ be a map. We say that d defines a *metric* on X if and only if it satisfies the following axioms:

1. positivity and identity of indiscernibles,

$$\forall\, x, y \in X, \quad d(x,y) \ge 0 \,, \qquad d(x,y) = 0 \iff x = y \,, \tag{6.7}$$

2. symmetry,
$$\forall\, x, y \in X, \quad d(x,y) = d(y,x)\,, \tag{6.8}$$

3. the triangle inequality,
$$\forall\, x, y, z \in X, \quad d(x,y) \le d(x,z) + d(z,y)\,. \tag{6.9}$$

Rests on Definitions 3.38 and 3.43.

Remark 6.25. These axioms bear a notable resemblance to the axioms of an equivalence relation: reflexivity, symmetry, and a triangle-shaped substitute for transitivity.

Definition 6.26 (Open ball; metric topology)**.** The *open ball* of centre $x_0 \in X$ and radius $r > 0$ is
$$B_r(x_0) = \{x \in X \mid d(x, x_0) < r\}\,. \tag{6.10}$$
The *metric topology* of (X, d) declares $U \subseteq X$ open iff every point of U has a ball around it inside U. The axioms of Definition 6.1 are verified directly: arbitrary unions inherit the property pointwise, and for a finite intersection the smallest of the finitely many radii works. Every metric space is thus a topological space.

Rests on Definitions 6.1, 6.2 and 6.24.

6.2.2 Convergence and completeness

Definition 6.27 (Convergence; Cauchy sequence; completeness)**.** A sequence (x_n) in (X, d) *converges* to x iff $d(x_n, x) \longrightarrow 0$; it is *Cauchy* iff $\forall\, \varepsilon > 0\ \exists\, N \mid m, n \ge N \implies d(x_m, x_n) < \varepsilon$. The space is *complete* iff every Cauchy sequence converges in it.

Rests on Definition 6.24.

Every convergent sequence is Cauchy (triangle inequality through the limit); the converse is exactly what completeness demands. The real line is complete (Corollary A.47, Appendix A.7); $\mathbb{Q}$ with the same metric is not, which is the entire reason $\mathbb{R}$ was constructed. Completeness of the function spaces built on this notion is what makes Hilbert spaces possible (Chapter 12).

6.2.3 Continuity in metric terms

Proposition 6.28 (ε–δ characterization)**.** *For metric spaces (X_1, d_1), (X_2, d_2), a map $f : X_1 \longrightarrow X_2$ is continuous at every point in the sense of Definition 6.6 if and only if*
$$\forall\, x_0,\ \forall\, \varepsilon > 0,\ \exists\, \delta > 0 \mid d_1(x, x_0) < \delta \implies d_2(f(x), f(x_0)) < \varepsilon\,. \tag{6.11}$$

Rests on Definitions 6.6 and 6.26.

Proof. *Derives Proposition 6.28.*

($\Rightarrow$) $B_\varepsilon(f(x_0))$ is open, so $f^{-1}(B_\varepsilon(f(x_0)))$ is open and contains x_0, hence contains a ball $B_\delta(x_0)$ — which is Equation (6.11). ($\Leftarrow$) Let U be open in X_2 and $x_0 \in f^{-1}(U)$. Some $B_\varepsilon(f(x_0)) \subseteq U$; the δ provided by Equation (6.11) gives $B_\delta(x_0) \subseteq f^{-1}(U)$, so $f^{-1}(U)$ is open. □

This recovers, for $X_1 = X_2 = \mathbb{R}$ with $d(x,y) = |x - y|$, the definition of continuity used throughout Chapter 7 — the two chapters describe one notion at two levels of generality.

6.2.4 Compactness in metric spaces

In a metric space — and there only — compactness admits a second, sequential description, which is the form in which analysis normally uses it.

Definition 6.29 (Sequential compactness)**.** A metric space (X, d) is *sequentially compact* if and only if every sequence in X has a subsequence converging to a point of X. A subset is sequentially compact if and only if it is so as a metric space in its own right, with the restricted distance.

Rests on Definitions 6.24 and 6.27.

Lemma 6.30 (Lebesgue number)**.** *Let (X, d) be sequentially compact and let $\{U_i\}_{i\in I}$ be an open cover of X. There is a $\delta > 0$ — a* Lebesgue number *of the cover — such that every ball $B_\delta(x)$, $x \in X$, is contained in a single member of the cover.*

Rests on Definitions 6.5, 6.26 and 6.29.

Proof. *Derives Lemma 6.30.*

If no such δ existed, then for every $n \in \mathbb{N}$ there would be a point $x_n \in X$ with $B_{1/n}(x_n)$ contained in no member of the cover. By sequential compactness some subsequence $x_{n_k} \longrightarrow x \in X$, and x lies in some U_{i_0}, which contains a ball $B_\varepsilon(x)$. Choose k so large that $d(x_{n_k}, x) < \varepsilon/2$ and $1/n_k < \varepsilon/2$. Every $y \in B_{1/n_k}(x_{n_k})$ then satisfies

$$d(y, x) \le d(y, x_{n_k}) + d(x_{n_k}, x) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\,,$$

so $B_{1/n_k}(x_{n_k}) \subseteq B_\varepsilon(x) \subseteq U_{i_0}$ — contradicting the choice of x_{n_k}. □

Theorem 6.31 (Compactness and sequential compactness)**.** *A metric space (X, d) is compact if and only if it is sequentially compact.*

Rests on Definitions 6.9 and 6.29 and Lemma 6.30.

Proof. *Derives Theorem 6.31.*

Compact $\Longrightarrow$ sequentially compact. Let (x_n) be a sequence with no convergent subsequence. Then no point $x \in X$ is the limit of a subsequence, and consequently each x admits a radius $r_x > 0$ such that $B_{r_x}(x)$ contains x_n for only finitely many indices n: otherwise every ball $B_{1/k}(x)$ would contain terms of arbitrarily large index, and picking $n_1 < n_2 < \cdots$ with $x_{n_k} \in B_{1/k}(x)$ would produce a subsequence converging to x. The balls $B_{r_x}(x)$ form an open cover of X, so finitely many of them cover X; their union then contains x_n for only finitely many indices, although it is all of X and so contains every term of the sequence. Since the index set is infinite, this is absurd.

Sequentially compact $\Longrightarrow$ compact. Let $\{U_i\}$ be an open cover of X and let $\delta > 0$ be a Lebesgue number for it (Lemma 6.30). We claim first that finitely many balls of radius δ cover X. If not, choose $x_1 \in X$ and, having chosen $x_1, \dots, x_n$, a point x_{n+1} outside $B_\delta(x_1) \cup \cdots \cup B_\delta(x_n)$; the resulting sequence satisfies $d(x_m, x_n) \ge \delta$ for all $m \ne n$, so none of its subsequences is Cauchy and none converges — contradicting sequential compactness. So $X = B_\delta(y_1) \cup \cdots \cup B_\delta(y_p)$ for suitable points $y_1, \dots, y_p$, and by the defining property of δ each $B_\delta(y_q)$ lies inside some U_{i_q}. The finitely many sets $U_{i_1}, \dots, U_{i_p}$ cover X. □

Remark 6.32. On a closed bounded subset of $\mathbb{R}$ the sequential form is exactly the Bolzano–Weierstrass theorem (Theorem 7.7), which Theorem 6.31 converts into the covering statement of Theorem 6.11. Applying Bolzano–Weierstrass in each coordinate in turn gives the sequential compactness of a closed bounded subset of $\mathbb{R}^N$, and the same conversion then reproduces Theorem 6.12 — whose direct proof above is independent of this equivalence, and so may be read as a second, covering-side derivation of it. The restriction to metric spaces is essential rather than technical: for general topological spaces neither implication holds.

6.3 The topology of $\mathbb{R}^N$

The real numbers admit no general symbolic rule that generates them all, but graphically they can be represented by a straight line built out of them: the real line, on which every $x \in \mathbb{R}$ has its place.

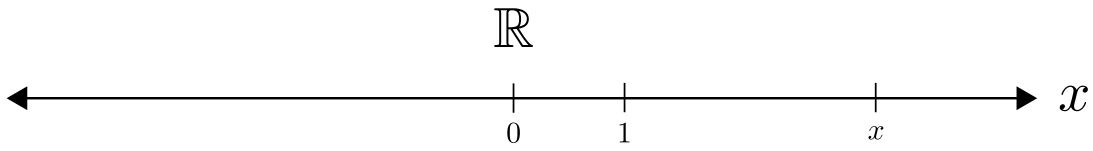


Figure 6.3: The real line: the set $\mathbb{R}$ represented as a straight line, with the integers 0 and 1 fixing origin and unit and an arbitrary real number x marked.

6.3.1 Vectors in $\mathbb{R}^N$

The set $\mathbb{R}^N$ of ordered N-tuples $\boldsymbol{x} = (x_1, \ldots, x_N)$ carries three compatible structures at once: the vector-space structure of Chapter 5 (componentwise addition and scaling), the *Euclidean metric*

$$d(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}| = \left(\sum_{i=1}^{N} (x_i - y_i)^2 \right)^{1/2}, \tag{6.12}$$

whose metric axioms follow from the Cauchy–Schwarz inequality (Chapter 5), and the metric topology this induces (Definition 6.26). The identification of the N-tuple with the geometric vector is precisely the statement that these structures cohabit: the topology sees nearness, the vector space sees displacement, and the metric ties them together. Physics happens on this triple structure — configuration spaces in Part III, and, once the metric is allowed to vary from point to point and take general signature (p, q), the manifolds of Chapter 13.

Convergence in $\mathbb{R}^N$ is componentwise (each coordinate is squeezed by the Euclidean distance), so the completeness of $\mathbb{R}$ (Appendix A.7) lifts coordinatewise: $\mathbb{R}^N$ is a complete metric space.

Chapter 7

Real Analysis

Real analysis makes precise the two operations on which the whole of physics rests: the limit, from which continuity and the derivative follow, and the integral, which undoes the derivative and sums continuous distributions. Everything in this chapter descends from a single property of the real numbers — completeness — which is therefore stated first, as an axiom here and as a theorem in Appendix A.7, where the real numbers are constructed and the property is proven. It is, by design, the deepest proof of Parts I and II.

7.1 The real numbers

The rational numbers $\mathbb{Q}$ form an ordered field (Chapter 4), but a gapped one: no rational squares to 2. Analysis begins by filling the gaps.

Axiom 7.1 (Completeness of $\mathbb{R}$)**.** $\mathbb{R}$ is an ordered field containing $\mathbb{Q}$, in which every nonempty subset S that is bounded above has a least upper bound $\sup S \in \mathbb{R}$.

Full derivation: Appendix A.7 (p. 2791) constructs $\mathbb{R}$ from $\mathbb{Q}$ by Dedekind cuts and proves Axiom 7.1 as a theorem, together with three consequences used constantly below: the *Archimedean property* (Corollary A.45: for $x > 0$ and any y some $nx > y$), the *density of* $\mathbb{Q}$ (between any two reals lies a rational), and *monotone convergence* (Corollary A.46: a bounded nondecreasing sequence converges to its supremum). By symmetry, bounded nonempty sets have infima and bounded nonincreasing sequences converge to them.

Definition 7.2 (Absolute value)**.** For $x \in \mathbb{R}$, $|x| = x$ if $x \geq 0$ and $|x| = -x$ otherwise.

Proposition 7.3 (Triangle inequality)**.** $|x + y| \leq |x| + |y|$ *for all* $x, y \in \mathbb{R}$.
Rests on Definition 7.2.

Proof. *Derives Proposition 7.3.*
From $-|x| \leq x \leq |x|$ and the same for y, adding gives $-(|x| + |y|) \leq x + y \leq |x| + |y|$, which is the claim. □

7.2 Sequences

Definition 7.4 (Convergence)**.** A sequence $(a_n)_{n \in \mathbb{N}}$ in $\mathbb{R}$ *converges* to $L \in \mathbb{R}$, written $a_n \longrightarrow L$ or $\lim_{n \to \infty} a_n = L$, iff

$$\forall\, \varepsilon > 0,\ \exists\, N \in \mathbb{N} \mid n \geq N \implies |a_n - L| < \varepsilon\,. \tag{7.1}$$

Rests on Definition 7.2.

Proposition 7.5 (Uniqueness of the limit)**.** *A sequence has at most one limit.*
Rests on Definition 7.4 and Proposition 7.3.

Proof. *Derives Proposition 7.5.*

If $a_n \to L$ and $a_n \to L'$ with $L \neq L'$, take $\varepsilon = |L - L'|/2 > 0$. For n beyond both thresholds, $|L - L'| \leq |L - a_n| + |a_n - L'| < 2\varepsilon = |L - L'|$, a contradiction. □

Proposition 7.6 (Algebra of limits)**.** *If $a_n \to A$ and $b_n \to B$, then $a_n + b_n \to A + B$, $a_n b_n \to AB$, and, if $B \neq 0$, $a_n / b_n \to A/B$.*

Rests on Definition 7.4 and Proposition 7.3.

Proof. *Derives Proposition 7.6.*

Sum: given ε, beyond the larger of the two thresholds for $\varepsilon/2$, $|(a_n + b_n) - (A + B)| \leq |a_n - A| + |b_n - B| < \varepsilon$. *Product:* convergent sequences are bounded (beyond N_1 for $\varepsilon = 1$ all terms lie within 1 of the limit, and finitely many remain), say $|a_n| \leq M$; then

$$|a_n b_n - AB| = |a_n(b_n - B) + B(a_n - A)| \leq M\,|b_n - B| + |B|\,|a_n - A|\,,$$

and each term is made smaller than $\varepsilon/2$ beyond a suitable threshold. *Quotient:* it suffices to prove $1/b_n \to 1/B$. Beyond some N_0, $|b_n| > |B|/2$ (take $\varepsilon = |B|/2$); then $|1/b_n - 1/B| = |B - b_n|/(|b_n|\,|B|) \leq 2\,|b_n - B|\,/\,|B|^2$, which is controlled as before. □

Theorem 7.7 (Bolzano–Weierstrass)**.** *Every bounded sequence in $\mathbb{R}$ has a convergent subsequence.*

Rests on Corollaries A.45 and A.46.

Proof. *Derives Theorem 7.7.*

Let $a_n \in [A_0, B_0]$ for all n. Bisect the interval: at least one half contains a_n for infinitely many indices n; call it $[A_1, B_1]$ and pick n_1 with $a_{n_1} \in [A_1, B_1]$. Iterating, we obtain nested intervals $[A_k, B_k]$ of length $(B_0 - A_0)/2^k$, each containing infinitely many terms, and indices $n_1 < n_2 < \cdots$ with $a_{n_k} \in [A_k, B_k]$. The sequence (A_k) is nondecreasing and bounded above by B_0, so by monotone convergence (Corollary A.46) $A_k \to L$; since $0 \leq B_k - A_k = (B_0 - A_0)/2^k \to 0$ (Archimedean property), $B_k \to L$ as well, and $A_k \leq a_{n_k} \leq B_k$ squeezes $a_{n_k} \to L$. □

Theorem 7.8 (Cauchy criterion)**.** *A sequence in $\mathbb{R}$ converges if and only if it is a Cauchy sequence: $\forall\, \varepsilon > 0\ \exists N \mid m, n \geq N \implies |a_m - a_n| < \varepsilon$.*

Rests on Definition 7.4 and Corollary A.47.

Proof. *Derives Theorem 7.8.*

Necessity: if $a_n \to L$, then beyond the threshold for $\varepsilon/2$, $|a_m - a_n| \leq |a_m - L| + |L - a_n| < \varepsilon$. Sufficiency is Corollary A.47, proven in Appendix A.7 from monotone convergence. □

The Cauchy criterion is the working form of completeness: it certifies convergence without knowing the limit, which is how every series and every integral below will be shown to exist.

7.3 Real functions

Definition 7.9 (Real function)**.** A *real function* is a map $f : A \subseteq \mathbb{R} \longrightarrow \mathbb{R}$ (Chapter 3); A is its *domain* and $f(A) = \{f(x) \mid x \in A\}$ its *image*.

Rests on Definition 3.43.

Definition 7.10 (Constant function)**.** Consider a real number $C \in \mathbb{R}$. We define the *constant function* as

$$\begin{aligned} f : A \subseteq \mathbb{R} &\longrightarrow B \subseteq \mathbb{R} \\ x &\longmapsto f(x) = C\,. \end{aligned}$$

Rests on Definition 7.9.

Definition 7.11 (Sum of functions)**.** Consider two real functions f and g. We define the *sum of functions* as the function

$$\begin{aligned} f+g : A \subseteq \mathbb{R} &\longrightarrow B \subseteq \mathbb{R} \\ x &\longmapsto (f+g)(x) = f(x) + g(x)\,. \end{aligned}$$

Rests on Definition 7.9.

Definition 7.12 (Product of functions)**.** Consider two real functions f and g. We define the *product of functions* as the function

$$\begin{aligned} fg : A \subseteq \mathbb{R} &\longrightarrow B \subseteq \mathbb{R} \\ x &\longmapsto (fg)(x) = f(x)g(x)\,. \end{aligned}$$

Rests on Definition 7.9.

Definition 7.13 (Composite function)**.** Consider two real functions f and g. We define the *composition of f with g* as the function

$$\begin{aligned} f \circ g : A \subseteq \mathbb{R} &\longrightarrow B \subseteq \mathbb{R} \\ x &\longmapsto (f \circ g)(x) = f(g(x))\,. \end{aligned}$$

Rests on Definition 7.9.

Definition 7.14 (Inverse function)**.** Consider a bijective real function $f : A \longrightarrow B$, $x \longmapsto f(x) = y$. The *inverse function* $f^{-1} : B \longrightarrow A$ assigns to each $y \in B$ the unique $x \in A$ with $f(x) = y$, so that $f^{-1} \circ f = \mathrm{id}_A$ and $f \circ f^{-1} = \mathrm{id}_B$. A strictly monotonic function is injective, and is therefore invertible onto its image.
Rests on Definitions 3.47 and 7.13.

Definition 7.15 (Absolute-value function)**.** $|\cdot| : \mathbb{R} \longrightarrow \mathbb{R}$, $x \longmapsto |x|$, with $|x|$ as in Definition 7.2.
Rests on Definitions 7.2 and 7.9.

7.4 Limits

Definition 7.16 (Limit)**.** Consider a real function $f : x \longmapsto f(x)$. We say that the *limit* of f, as x tends to x_0, exists and equals L if and only if

$$\forall\, \varepsilon > 0,\ \exists\, \delta(\varepsilon) \in \mathbb{R}^+ \mid 0 < |x - x_0| < \delta \implies |f(x) - L| < \varepsilon\,. \tag{7.2}$$

If it exists, we denote this limit by

$$\lim_{x \longrightarrow x_0} f(x) = L\,. \tag{7.3}$$

Rests on Definitions 7.2 and 7.9.

Remark 7.17. Note that the definition requires

$$|x - x_0| > 0 \quad \implies \quad x \neq x_0\,, \tag{7.4}$$

that is, the point x_0 itself is excluded: the limit probes the values of f near x_0, never at x_0.
Rests on Definition 7.16.

The algebra of limits of sequences (Proposition 7.6) transfers verbatim to limits of functions, by the same ε–δ estimates; we use it freely.

7.4.1 One-sided limits

Consider a real function $f : (a, b) \longrightarrow B \subseteq \mathbb{R}$. From the definition of the limit we have the condition

$$0 < |x - x_0| < \delta \,.$$

- If we choose $x_0 = a$, then, since the interval is open,

$$x > a \quad \Longrightarrow \quad 0 < x - a < \delta \,.$$

We say that the *limit of f from the right, as x tends to a, is L* if and only if

$$\lim_{x \longrightarrow a} f(x) = L, \quad x > a \,, \tag{7.5}$$

which we denote by

$$\lim_{x \longrightarrow a^+} f(x) = L \,. \tag{7.6}$$

- If we choose $x_0 = b$, then, since the interval is open,

$$\begin{aligned} x < b \implies & 0 < -(x - b) < \delta \\ \iff & 0 > x - b > -\delta \\ \iff & -\delta < x - b < 0 \,. \end{aligned}$$

We say that the *limit of f from the left, as x tends to b, is L* if and only if

$$\lim_{x \longrightarrow b} f(x) = L, \quad x < b \,, \tag{7.7}$$

which we denote by

$$\lim_{x \longrightarrow b^-} f(x) = L \,. \tag{7.8}$$

Proposition 7.18 (Two-sided limit from one-sided limits)**.** *Let x_0 be interior to the domain of f. Then $\lim_{x \to x_0} f(x) = L$ if and only if both one-sided limits exist and*

$$\lim_{x \longrightarrow x_0^-} f(x) = \lim_{x \longrightarrow x_0^+} f(x) = L \,. \tag{7.9}$$

Rests on Definition 7.16 and Equations (7.5) and (7.7).

Proof. *Derives Proposition 7.18.*

($\Rightarrow$) The two-sided δ works for each side separately. ($\Leftarrow$) Given ε, let δ_- and δ_+ be the one-sided thresholds and $\delta = \min(\delta_-, \delta_+)$. Any x with $0 < |x - x_0| < \delta$ lies on one of the two sides and satisfies the corresponding estimate, so $|f(x) - L| < \varepsilon$. □

7.4.2 Infinite limits and limits at infinity

Definition 7.19 (Infinite limits)**.** $\lim_{x \to a} f(x) = \infty$ iff for every $M \in \mathbb{R}$ there is $\delta > 0$ with $0 < |x - a| < \delta \implies f(x) > M$; and $L = \lim_{x \to \infty} f(x)$ iff for every $\varepsilon > 0$ there is $R \in \mathbb{R}$ with $x > R \implies |f(x) - L| < \varepsilon$. The variants with $-\infty$ are defined by the obvious sign changes.

Rests on Definitions 7.2 and 7.16.

An *asymptote* of f is a line approached by its graph in one of these senses: vertical at a when $\lim_{x \to a^{\pm}} f(x) = \pm\infty$, horizontal at height L when $\lim_{x \to \pm\infty} f(x) = L$, and oblique $y = mx + c$ when $\lim_{x \to \pm\infty} [f(x) - mx - c] = 0$.

7.5 Continuity

Definition 7.20 (Continuity at a point)**.** Consider a function $f : A \subseteq \mathbb{R} \longrightarrow B \subseteq \mathbb{R}$ and a point $x_0 \in A$. We say that f is *continuous* at x_0 if and only if

$$\exists \lim_{x \longrightarrow x_0} f(x) = f(x_0)\,. \tag{7.10}$$

f is *continuous on* A iff it is continuous at every point of A; the set of such functions is denoted $C^0(A)$. It is *piecewise continuous* on $[a, b]$ iff it is continuous there except at finitely many points, at each of which both one-sided limits exist.

Rests on Definition 7.16 and Equations (7.5) and (7.7).

Remark 7.21. Two ways of failing this condition are distinguished:

- if the limit exists but differs from $f(x_0)$, we say that f has a *removable discontinuity* at x_0;
- if the limit does not exist, we say that f has an *essential discontinuity* at x_0.

Rests on Definitions 7.16 and 7.20.

Proposition 7.22 (Sequential characterization)**.** *f is continuous at x_0 if and only if for every sequence $x_n \longrightarrow x_0$ in its domain, $f(x_n) \longrightarrow f(x_0)$.*

Rests on Definitions 7.4 and 7.20.

Proof. *Derives Proposition 7.22.*

($\Rightarrow$) Given ε, choose δ from continuity, then N so that $n \geq N \implies |x_n - x_0| < \delta$; then $|f(x_n) - f(x_0)| < \varepsilon$. ($\Leftarrow$) If f is not continuous at x_0, some ε_0 admits no δ: for each n there is x_n with $|x_n - x_0| < 1/n$ but $|f(x_n) - f(x_0)| \geq \varepsilon_0$. Then $x_n \to x_0$ while $f(x_n) \not\to f(x_0)$. □

Sums, products, quotients (where defined) and compositions of continuous functions are continuous, by Proposition 7.22 and Proposition 7.6.

Theorem 7.23 (Intermediate value theorem)**.** *Let $f \in C^0([a, b])$ with $f(a) < 0 < f(b)$. Then there is $c \in (a, b)$ with $f(c) = 0$.*

Rests on Axiom 7.1 and Definition 7.20.

Proof. *Derives Theorem 7.23.*

Let $S = \{x \in [a, b] \mid f(x) < 0\}$; S is nonempty ($a \in S$) and bounded above by b, so $c = \sup S$ exists (Axiom 7.1). We claim $f(c) = 0$. If $f(c) < 0$, then $c < b$, and continuity provides $\delta > 0$ with $f < 0$ on $(c - \delta, c + \delta) \cap [a, b]$ (take $\varepsilon = -f(c)$); points just above c would then lie in S, contradicting that c bounds S. If $f(c) > 0$, continuity gives $f > 0$ on $(c - \delta, c]$ for some $\delta > 0$; then $c - \delta$ is already an upper bound of S, contradicting leastness. Hence $f(c) = 0$, and $c \neq a, b$ because $f(a) < 0 < f(b)$. □

Theorem 7.24 (Extreme value theorem)**.** *A continuous function on a closed bounded interval $[a, b]$ is bounded and attains its supremum and infimum.*

Rests on Axiom 7.1, Theorem 7.7, and Proposition 7.22.

Proof. *Derives Theorem 7.24.*

Bounded: if not, pick $x_n \in [a, b]$ with $|f(x_n)| > n$. By Bolzano–Weierstrass (Theorem 7.7) a subsequence $x_{n_k} \to c \in [a, b]$; continuity forces $f(x_{n_k}) \to f(c)$, but $|f(x_{n_k})| > n_k \to \infty$ — contradiction. *Attained:* let $M = \sup f([a, b])$, which now exists; pick x_n with $f(x_n) > M - 1/n$. A convergent subsequence $x_{n_k} \to c$ gives $f(c) = \lim f(x_{n_k}) = M$. The infimum is the supremum of $-f$. □

Theorem 7.25 (Heine–Cantor: uniform continuity)**.** *A continuous function on $[a,b]$ is uniformly continuous: for every $\varepsilon > 0$ there is a single $\delta > 0$ such that $|x-y| < \delta \implies |f(x) - f(y)| < \varepsilon$ for* all *$x, y \in [a,b]$.*

Rests on Theorem 7.7 and Proposition 7.22.

Proof. *Derives Theorem 7.25.*

If not, some ε_0 admits points x_n, y_n with $|x_n - y_n| < 1/n$ but $|f(x_n) - f(y_n)| \geq \varepsilon_0$. Extract $x_{n_k} \to c$ (Theorem 7.7); then $y_{n_k} \to c$ too, and continuity gives $f(x_{n_k}) - f(y_{n_k}) \to f(c) - f(c) = 0$, contradicting $|f(x_{n_k}) - f(y_{n_k})| \geq \varepsilon_0$. □

7.6 The derivative

Definition 7.26 (Derivative of a function at a point)**.** Consider a real function

$$\begin{array}{rccc} f: & (a,b) & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & f(x) \end{array}$$

and a point $x_0 \in (a,b)$. We define the *derivative of f at x_0* as the real number given by

$$f'(x_0) = \lim_{x \longrightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}\,. \tag{7.11}$$

Rests on Definitions 7.9 and 7.16.

We may introduce

$$h = x - x_0\,, \tag{7.12}$$

in terms of which the derivative reads

$$f'(x_0) = \lim_{h \longrightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}\,. \tag{7.13}$$

Proposition 7.27 (Differentiable implies continuous)**.** *If $f'(x_0)$ exists, f is continuous at x_0.*

Rests on Definitions 7.20 and 7.26 and Proposition 7.6.

Proof. *Derives Proposition 7.27.*

$f(x) - f(x_0) = \dfrac{f(x) - f(x_0)}{x - x_0}\,(x - x_0) \longrightarrow f'(x_0) \cdot 0 = 0$ as $x \to x_0$, by the algebra of limits. □

7.6.1 Geometric interpretation of the derivative

The construction underlying Equation (7.11) is geometric: the difference quotient is the slope of the secant line through the points $(x_0, f(x_0))$ and $(x, f(x))$ of the graph, and the derivative is the slope of the tangent line obtained in the limit $x \longrightarrow x_0$, as the sequence of figures Figures 7.1 to 7.4 illustrates. It is for this reason that we must take only open intervals: the limit must be able to approach the point from within the domain.

7.6.2 The derivative function

From the definition in Equation (7.11) we may take $x_0 = u$ to be an arbitrary point of the interval (a,b) at which the derivative is evaluated, and thereby obtain a general expression for the derivative at any point of (a,b). We define the *derivative function* of f as the real function that assigns to each point of the domain of f the derivative of f at that point, which we denote by f'. Making the change of variable $h = x - u$,

$$f'(x) = \lim_{h \longrightarrow 0} \frac{f(x + h) - f(x)}{h}\,. \tag{7.14}$$

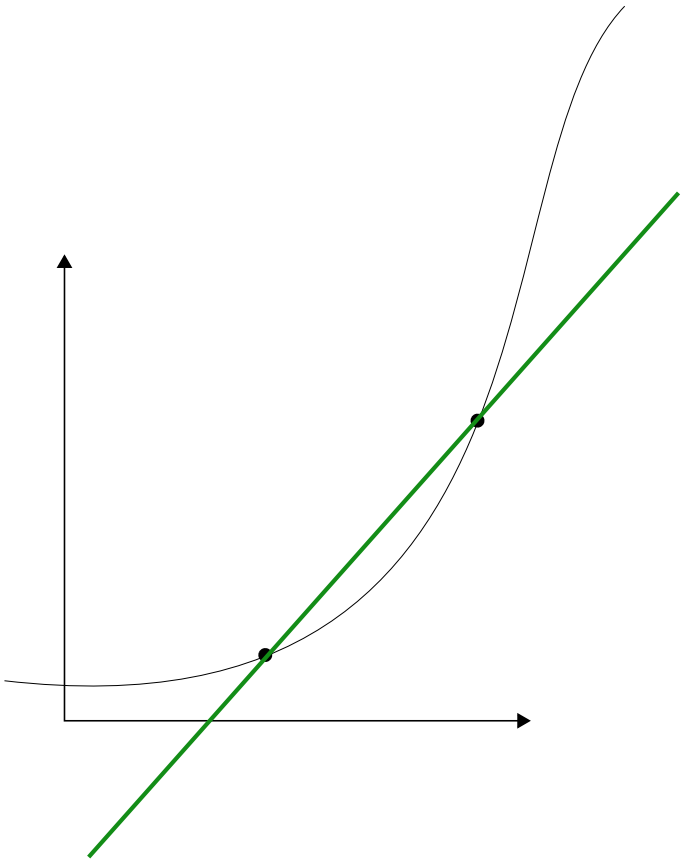


Figure 7.1: The secant-to-tangent construction of the derivative (panel 1 of 4).

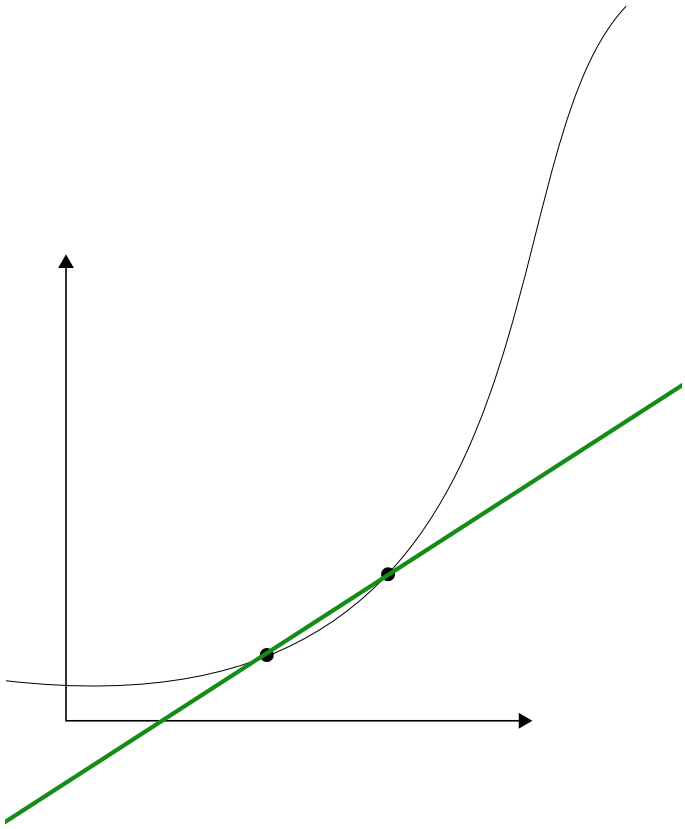


Figure 7.2: The secant-to-tangent construction of the derivative (panel 2 of 4).

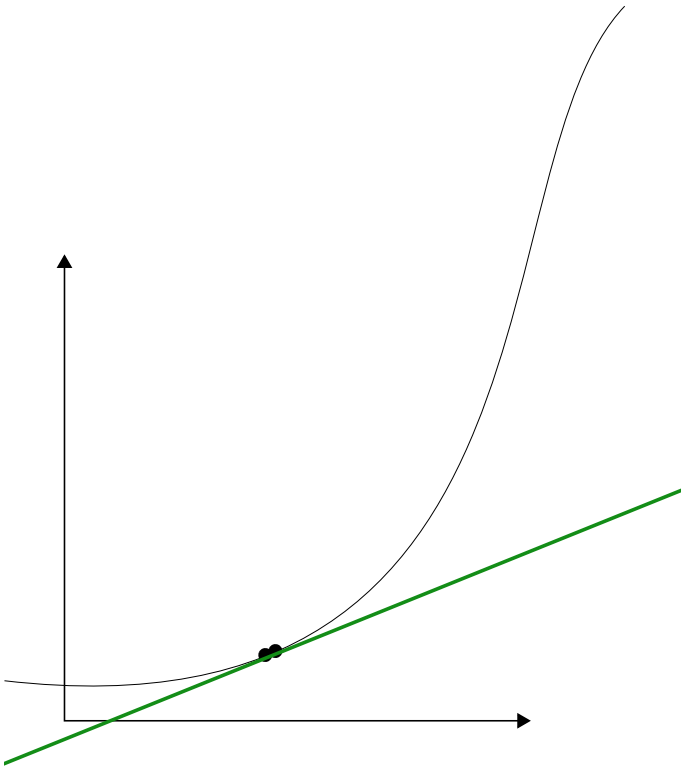


Figure 7.3: The secant-to-tangent construction of the derivative (panel 3 of 4).

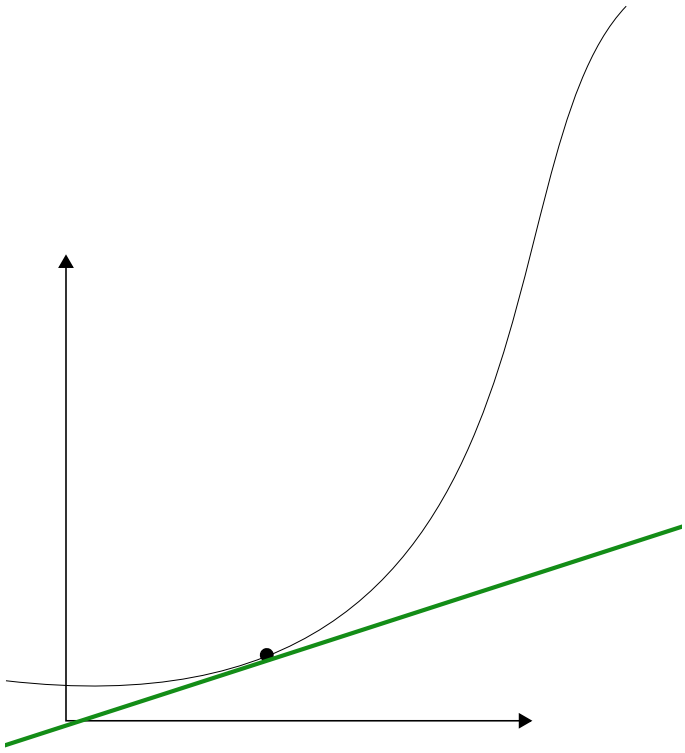


Figure 7.4: The secant-to-tangent construction of the derivative (panel 4 of 4).

7.6.3 Rules of differentiation

Proposition 7.28 (Derivative of a constant function)**.** *Let $f : x \longmapsto C$ be a constant real function. Then*

$$f = C \quad \Longrightarrow \quad f' = 0\,. \tag{7.15}$$

Rests on Definition 7.10 and Equation (7.14).

Derivation. Derives Proposition 7.28.

From Equation (7.14) we have

$$\begin{aligned} f'(x) &= \lim_{h \longrightarrow 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h \longrightarrow 0} \frac{C - C}{h} = 0\,. \end{aligned}$$

□

Proposition 7.29 (Linearity)**.** $(\lambda f)' = \lambda f'$ *and* $(f+g)' = f' + g'$ *for* $\lambda \in \mathbb{R}$.
Rests on Definition 7.11, Equation (7.14), and Proposition 7.6.

Proof. Derives Proposition 7.29.

Both difference quotients split termwise, and the algebra of limits (Proposition 7.6) distributes the limit over the pieces. □

Proposition 7.30 (Leibniz rule)**.** *Let f and g be real functions. Then*

$$(fg)' = f'g + fg'\,. \tag{7.16}$$

Rests on Definition 7.12, Equation (7.14), and Propositions 7.6 and 7.27.

Derivation. Derives Proposition 7.30.

We have

$$\begin{aligned} (fg)'(x) &= \lim_{h \longrightarrow 0} \frac{(fg)(x+h) - (fg)(x)}{h} \\ &= \lim_{h \longrightarrow 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h} \\ &= \lim_{h \longrightarrow 0} \frac{f(x+h)g(x+h) - f(x)g(x) + f(x)g(x+h) - f(x)g(x+h)}{h} \\ &= \lim_{h \longrightarrow 0} \frac{f(x)\,[g(x+h) - g(x)] + g(x+h)\,[f(x+h) - f(x)]}{h} \\ &= f(x) \lim_{h \longrightarrow 0} \frac{g(x+h) - g(x)}{h} + \lim_{h \longrightarrow 0} g(x+h) \lim_{h \longrightarrow 0} \frac{f(x+h) - f(x)}{h} \\ &= f(x)g'(x) + f'(x)g(x)\,, \end{aligned}$$

where the last step uses Equation (7.14) and the continuity of g (Proposition 7.27). □

Proposition 7.31 (Chain rule)**.** *Let f be differentiable at x and g differentiable at $f(x)$. Then*

$$(g \circ f)'(x) = g'(f(x))\, f'(x)\,. \tag{7.17}$$

Rests on Definition 7.13 and Propositions 7.6 and 7.27.

Derivation. Derives Proposition 7.31.

Define, for y in the domain of g and $u_0 = f(x)$,

$$\varphi(y) = \begin{cases} \dfrac{g(y) - g(u_0)}{y - u_0}\,, & y \neq u_0\,, \\[2ex] g'(u_0)\,, & y = u_0\,. \end{cases}$$

Differentiability of g at u_0 says precisely that φ is continuous at u_0, and by construction $g(y) - g(u_0) = \varphi(y)\,(y - u_0)$ for *all* y — including $y = u_0$, which is the case the naive quotient argument cannot handle. Substituting $y = f(x+h)$,

$$\frac{g(f(x+h)) - g(f(x))}{h} = \varphi(f(x+h))\,\frac{f(x+h) - f(x)}{h}\,.$$

As $h \to 0$: $f(x+h) \to f(x)$ by Proposition 7.27, so $\varphi(f(x+h)) \to \varphi(f(x)) = g'(f(x))$ by continuity of φ, while the second factor tends to $f'(x)$. The algebra of limits gives Equation (7.17). □

Proposition 7.32 (Derivative of the inverse function)**.** *Let f be strictly monotonic and continuous on an interval, and differentiable at x_0 with $f'(x_0) \neq 0$. Then f^{-1} is differentiable at $y_0 = f(x_0)$ and*

$$(f^{-1})'(y_0) = \frac{1}{f'(x_0)}\,. \tag{7.18}$$

Rests on Definition 7.14, Theorem 7.23, and Proposition 7.6.

Proof. *Derives Proposition 7.32.*

For $y \neq y_0$ write $x = f^{-1}(y)$; continuity and strict monotonicity make f^{-1} continuous (its one-sided limits exist by monotonicity and cannot jump without leaving a gap in the image, forbidden by Theorem 7.23), so $y \to y_0$ forces $x \to x_0$. Then

$$\frac{f^{-1}(y) - f^{-1}(y_0)}{y - y_0} = \frac{x - x_0}{f(x) - f(x_0)} \longrightarrow \frac{1}{f'(x_0)}\,.$$

□

7.6.4 The mean value theorems

Lemma 7.33 (Fermat: interior extremum)**.** *If f attains a local maximum or minimum at an interior point c and $f'(c)$ exists, then $f'(c) = 0$.*

Rests on Definition 7.26 and Proposition 7.18.

Proof. *Derives Lemma 7.33.*

Say the extremum is a maximum. For $h > 0$ small, $[f(c+h) - f(c)]\,/h \leq 0$, so the right limit gives $f'(c) \leq 0$; for $h < 0$, the quotient is ≥ 0, so $f'(c) \geq 0$. Hence $f'(c) = 0$. □

Theorem 7.34 (Rolle)**.** *If $f \in C^0([a,b])$ is differentiable on (a,b) and $f(a) = f(b)$, then $f'(c) = 0$ for some $c \in (a,b)$.*

Rests on Theorem 7.24 and Lemma 7.33.

Proof. *Derives Theorem 7.34.*

By the extreme value theorem (Theorem 7.24) f attains a maximum and a minimum on $[a,b]$. If both are attained at the endpoints, f is constant (max = min) and any interior c works. Otherwise one extremum is interior, and Lemma 7.33 finishes. □

Theorem 7.35 (Mean value theorem)**.** *If $f \in C^0([a,b])$ is differentiable on (a,b), there is $c \in (a,b)$ with*

$$f(b) - f(a) = f'(c)\,(b - a)\,. \tag{7.19}$$

Rests on Theorem 7.34 and Proposition 7.29.

Proof. *Derives Theorem 7.35.*

Apply Rolle to $\phi(x) = f(x) - f(a) - \dfrac{f(b) - f(a)}{b - a}\,(x - a)$, which is continuous on $[a,b]$, differentiable inside, and vanishes at both endpoints; $\phi'(c) = 0$ is Equation (7.19). □

Corollary 7.36. *Let f be continuous on an interval and differentiable at its interior points. If $f' = 0$ there, f is constant; if $f' > 0$ there, f is strictly increasing — on the whole interval, endpoints included.*

Rests on Theorem 7.35.

Proof. *Derives Corollary 7.36.*

For $x < y$ in the interval, Theorem 7.35 on $[x, y]$ gives $f(y) - f(x) = f'(c)(y - x)$ with $c \in (x, y)$ — an interior point — and the right side vanishes in the first case and is positive in the second. □

Theorem 7.37 (Cauchy mean value theorem and l'Hôpital's rule)**.** *Let $f, g \in C^0([a, b])$ be differentiable on (a, b) with $g' \neq 0$ there. Then there is $c \in (a, b)$ with $[f(b) - f(a)]\, g'(c) = [g(b) - g(a)]\, f'(c)$. Consequently, if $f(x_0) = g(x_0) = 0$ and $\lim_{x \to x_0} f'(x)/g'(x) = L$ exists, then*

$$\lim_{x \longrightarrow x_0} \frac{f(x)}{g(x)} = L\,. \tag{7.20}$$

Rests on Theorem 7.34 and Definition 7.16.

Proof. *Derives Theorem 7.37.*

Rolle applied to $\psi(x) = [f(b) - f(a)]\, g(x) - [g(b) - g(a)]\, f(x)$ (which takes the value $f(b)g(a) - g(b)f(a)$ at both ends) gives the first claim. For the second, take $x \neq x_0$ and apply it on the interval with endpoints x_0, x: since $f(x_0) = g(x_0) = 0$,

$$\frac{f(x)}{g(x)} = \frac{f'(c_x)}{g'(c_x)} \quad \text{for some } c_x \text{ between } x_0 \text{ and } x\,.$$

As $x \to x_0$, also $c_x \to x_0$, and the right-hand side tends to L. □

Theorem 7.38 (Taylor's theorem with Lagrange remainder)**.** *Let f be $n+1$ times differentiable on an interval containing x_0 and x. Then*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}\,(x - x_0)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!}\,(x - x_0)^{n+1} \tag{7.21}$$

for some ξ between x_0 and x.

Rests on Theorem 7.34 and Proposition 7.30.

Proof. *Derives Theorem 7.38.*

Fix x and let M be defined by $f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + M\,(x - x_0)^{n+1}$. Consider

$$g(t) = f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(t)}{k!}\,(x - t)^k - M\,(x - t)^{n+1}\,.$$

Then $g(x) = 0$ and $g(x_0) = 0$ by the choice of M, so Rolle gives ξ between x_0 and x with $g'(\xi) = 0$. Differentiating the sum in t telescopes: all terms cancel pairwise except the last, leaving

$$g'(t) = -\frac{f^{(n+1)}(t)}{n!}\,(x - t)^n + (n+1)M\,(x - t)^n\,.$$

Setting $t = \xi$ and dividing by $(x - \xi)^n \neq 0$ yields $M = f^{(n+1)}(\xi)/(n+1)!$, which is Equation (7.21). □

7.7 The Riemann integral

Definition 7.39 (Darboux sums and the definite integral)**.** Let $f : [a,b] \longrightarrow \mathbb{R}$ be bounded and $P : a = x_0 < x_1 < \cdots < x_n = b$ a partition. With $m_i = \inf_{[x_{i-1},x_i]} f$ and $M_i = \sup_{[x_{i-1},x_i]} f$, the *lower* and *upper sums* are

$$L(f,P) = \sum_{i=1}^{n} m_i \,(x_i - x_{i-1})\,, \qquad U(f,P) = \sum_{i=1}^{n} M_i \,(x_i - x_{i-1})\,. \tag{7.22}$$

f is *Riemann integrable* iff $\sup_P L(f,P) = \inf_P U(f,P)$, and the common value is the *definite integral* $\int_a^b f(x)\,\mathrm{d}x$. (Both extrema exist by Axiom 7.1: every lower sum is $\le$ every upper sum, since refining a partition raises L and lowers U.)

Rests on Axiom 7.1.

Theorem 7.40 (Continuous functions are integrable)**.** *$f \in C^0([a,b])$ is Riemann integrable.*

Rests on Definition 7.39 and Theorems 7.24 and 7.25.

Proof. *Derives Theorem 7.40.*

Given $\varepsilon > 0$, Heine–Cantor (Theorem 7.25) provides δ with $|f(x) - f(y)| < \varepsilon/(b-a)$ whenever $|x - y| < \delta$. For any partition with all steps $x_i - x_{i-1} < \delta$, the extreme value theorem gives points where m_i, M_i are attained, at distance $< \delta$, so $M_i - m_i < \varepsilon/(b-a)$ and

$$U(f,P) - L(f,P) = \sum_i (M_i - m_i)(x_i - x_{i-1}) < \frac{\varepsilon}{b-a}\,(b-a) = \varepsilon\,.$$

Upper and lower extrema squeezed within every ε coincide. □

The integral is linear and monotone in f, and additive over subintervals, $\int_a^b = \int_a^c + \int_c^b$ — each property follows directly from the corresponding property of the Darboux sums.

Definition 7.41 (Antiderivative)**.** We define the *antiderivative* of a function f as the function F such that

$$F' = f\,, \tag{7.23}$$

and we denote it by

$$F = \int f\,. \tag{7.24}$$

By Corollary 7.36, two antiderivatives of the same f on an interval differ by a constant.

Rests on Definition 7.26 and Corollary 7.36.

7.8 The fundamental theorem of calculus

Theorem 7.42 (Fundamental theorem of calculus, I)**.** *Let $f \in C^0([a,b])$ and define $F(x) = \int_a^x f(t)\,\mathrm{d}t$. Then F is differentiable on (a,b) and*

$$F'(x) = f(x)\,. \tag{7.25}$$

Rests on Theorems 7.24 and 7.40 and Definition 7.26.

Proof. *Derives Theorem 7.42.*

For $h > 0$ (the case $h < 0$ is symmetric), additivity gives

$$\frac{F(x+h) - F(x)}{h} = \frac{1}{h}\int_x^{x+h} f(t)\,\mathrm{d}t\,.$$

On $[x, x+h]$ let m_h, M_h be the (attained) minimum and maximum of f; monotonicity of the integral squeezes $m_h \le \frac{1}{h}\int_x^{x+h} f \le M_h$. As $h \to 0$, continuity of f at x drives $m_h, M_h \to f(x)$, and the difference quotient converges to $f(x)$. □

Theorem 7.43 (Fundamental theorem of calculus, II)**.** *If $f \in C^0([a,b])$ and G is any antiderivative of f, then*

$$\int_a^b f(x)\,\mathrm{d}x = G(b) - G(a)\,. \tag{7.26}$$

Rests on Theorem 7.42 and Definition 7.41.

Proof. *Derives Theorem 7.43.*

$F(x) = \int_a^x f$ is an antiderivative by Theorem 7.42, so $G = F + C$ for a constant C (Definition 7.41). Then $G(b) - G(a) = F(b) - F(a) = \int_a^b f - 0$. □

Corollary 7.44 (Substitution and integration by parts)**.** *For C^1 functions on the relevant intervals,*

$$\int_a^b f(g(t))\,g'(t)\,\mathrm{d}t = \int_{g(a)}^{g(b)} f(x)\,\mathrm{d}x\,, \tag{7.27}$$

$$\int_a^b f'(x)\,g(x)\,\mathrm{d}x = [f(x)g(x)]_a^b - \int_a^b f(x)\,g'(x)\,\mathrm{d}x\,. \tag{7.28}$$

Rests on Theorem 7.43 and Propositions 7.30 and 7.31.

Proof. *Derives Corollary 7.44.*

Both sides of Equation (7.27) equal $F(g(b)) - F(g(a))$ for an antiderivative F of f, by Theorem 7.43 and the chain rule $(F \circ g)' = f(g)\,g'$ (Proposition 7.31). Equation Equation (7.28) is Theorem 7.43 applied to the Leibniz rule $(fg)' = f'g + fg'$ (Proposition 7.30). □

7.9 Series

Definition 7.45 (Series)**.** Given a sequence (a_n), the *series* $\sum_{n=0}^{\infty} a_n$ is the limit of its partial sums $s_N = \sum_{n=0}^{N} a_n$, when it exists. It *converges absolutely* iff $\sum |a_n|$ converges.

Rests on Definitions 7.2 and 7.4.

Proposition 7.46 (Geometric series)**.** *For $|q| < 1$,* $\displaystyle\sum_{n=0}^{\infty} q^n = \frac{1}{1-q}$.

Rests on Definition 7.45 and Corollary A.46.

Proof. *Derives Proposition 7.46.*

$(1-q)\,s_N = 1 - q^{N+1}$, so $s_N = (1-q^{N+1})/(1-q)$. Since $|q| < 1$, $q^{N+1} \to 0$ (its absolute value is decreasing and bounded below, so it converges by monotone convergence, and the limit ℓ satisfies $\ell = |q|\,\ell$, forcing $\ell = 0$). □

Proposition 7.47 (Comparison; absolute convergence)**.** *If $0 \le a_n \le b_n$ and $\sum b_n$ converges, then $\sum a_n$ converges. Moreover an absolutely convergent series converges.*

Rests on Definition 7.45, Corollary A.46, and Theorem 7.8.

Proof. *Derives Proposition 7.47.*

The partial sums of $\sum a_n$ are nondecreasing and bounded above by $\sum b_n$; monotone convergence (Corollary A.46) concludes. For the second claim, apply the Cauchy criterion (Theorem 7.8): for $M > N$, $|s_M - s_N| = \left|\sum_{N+1}^{M} a_n\right| \le \sum_{N+1}^{M} |a_n|$, and the right side is the Cauchy difference of the convergent series $\sum |a_n|$. □

Proposition 7.48 (Ratio test)**.** *If $\limsup_n |a_{n+1}/a_n| = r < 1$, then $\sum a_n$ converges absolutely; if $|a_{n+1}/a_n| \ge 1$ eventually, it diverges.*

Rests on Propositions 7.46 and 7.47 and Theorem 7.8.

Proof. *Derives Proposition 7.48.*

Pick q with $r < q < 1$: beyond some N, $|a_{n+1}| \le q\,|a_n|$, so $|a_{N+k}| \le q^k\,|a_N|$ and comparison with the geometric series (Propositions 7.46 and 7.47) gives absolute convergence. In the divergent case $|a_n|$ is eventually nondecreasing and nonzero, so $a_n \not\to 0$, while convergence of a series forces its terms to 0 (Cauchy criterion with $M = N+1$). □

Proposition 7.49 (Cauchy product)**.** *Let $\sum_{n=0}^{\infty} a_n = A$ and $\sum_{n=0}^{\infty} b_n = B$ converge absolutely. Then their* Cauchy product

$$\sum_{n=0}^{\infty} c_n\,, \qquad c_n = \sum_{k=0}^{n} a_k\, b_{n-k}\,, \tag{7.29}$$

converges absolutely, with sum AB.

Rests on Definition 7.45, Corollary A.46, and Proposition 7.6.

Proof. *Derives Proposition 7.49.*

Write $A^* = \sum_n |a_n|$ and $B^* = \sum_n |b_n|$. For any N,

$$\sum_{n=0}^{N} |c_n| \le \sum_{n=0}^{N}\sum_{k=0}^{n} |a_k|\;|b_{n-k}| \le \left(\sum_{k=0}^{N} |a_k|\right)\left(\sum_{j=0}^{N} |b_j|\right) \le A^*B^*\,,$$

since every product $|a_k|\,|b_j|$ with $k+j \le N$ appears exactly once in the middle double sum and all products are nonnegative. The partial sums of $\sum |c_n|$ are nondecreasing and bounded, hence convergent (Corollary A.46): the Cauchy product converges absolutely. For its value, compare $C_N = \sum_{n=0}^{N} c_n$ with the product of partial sums $A_N B_N$: expanding both, $A_N B_N - C_N$ consists exactly of the products $a_j b_k$ with $j,k \le N$ but $j+k > N$. Every such pair has $j > N/2$ or $k > N/2$ (double counting only enlarges the bound), so

$$|A_N B_N - C_N| \le B^* \sum_{j>N/2} |a_j| + A^* \sum_{k>N/2} |b_k| \longrightarrow 0\,,$$

because each tail is the difference between A^* (respectively B^*) and one of its own partial sums, and those partial sums converge to it. Since $A_N B_N \to AB$ by the algebra of limits (Proposition 7.6), also $C_N \to AB$. □

Theorem 7.50 (Power series; radius of convergence)**.** *For a power series $\sum_n c_n (x-x_0)^n$ there is $R \in [0,\infty]$ such that the series converges absolutely for $|x-x_0| < R$ and diverges for $|x-x_0| > R$.*

Rests on Axiom 7.1 and Propositions 7.46 and 7.47.

Proof. *Derives Theorem 7.50.*

Let $S = \{\rho \ge 0 \mid (|c_n|\,\rho^n) \text{ is bounded}\}$ and $R = \sup S$ (allowing ∞). If $|x-x_0| < R$, pick $\rho \in S$ with $|x-x_0| < \rho$ and a bound $|c_n|\,\rho^n \le M$; then $|c_n (x-x_0)^n| \le M\,(|x-x_0|/\rho)^n$, a convergent geometric majorant, so comparison gives absolute convergence. If $|x-x_0| > R$, the terms $c_n(x-x_0)^n$ are unbounded (else $|x-x_0| \in S$), so they cannot tend to 0 and the series diverges. □

Theorem 7.51 (Termwise differentiation)**.** *Let $f(x) = \sum_{n=0}^{\infty} c_n (x-x_0)^n$ have radius of convergence $R > 0$. Then the derived series $\sum_{n\ge 1} n\,c_n\,(x-x_0)^{n-1}$ has the same radius R, and on $|x-x_0| < R$ the function f is differentiable with*

$$f'(x) = \sum_{n=1}^{\infty} n\,c_n\,(x-x_0)^{n-1}\,. \tag{7.30}$$

Iterating, f has derivatives of every order inside its radius; in particular it is continuous there.

Rests on Theorem 7.50 and Propositions 7.6, 7.27, 7.47 and 7.48.

Proof. *Derives Theorem 7.51.*

Translate so that $x_0 = 0$. *Same radius.* Let $|x| < R$. As in the proof of Theorem 7.50, pick ρ in the set S of that proof with $|x| < \rho$ (possible because $R = \sup S > |x|$) and a bound $|c_n|\,\rho^n \le M$; then

$$n\,|c_n|\,|x|^{n-1} \le \frac{M}{\rho}\,n\left(\frac{|x|}{\rho}\right)^{n-1},$$

and $\sum_n n\,t^{n-1}$ converges for $0 < t < 1$ by the ratio test (Proposition 7.48: the ratio of consecutive terms is $\frac{n+1}{n}\,t \to t < 1$), and trivially at $t = 0$, where only the $n = 1$ term survives. Comparison (Proposition 7.47) gives absolute convergence of the derived series whenever $|x| < R$. Conversely $|c_n x^n| \le |x| \cdot n\,|c_n|\,|x|^{n-1}$ for $n \ge 1$, so wherever the derived series converges absolutely the original does too: the two radii coincide. Applying the statement to the derived series itself, $\sum_{n\ge2} n(n-1)\,c_n\,x^{n-2}$ also has radius R.

The difference quotient. Fix $|x| < R$, choose r with $|x| < r < R$, and let $y \ne x$ with $|y| \le r$. Subtracting the two absolutely convergent series term by term and dividing by $y - x$ (Proposition 7.6 on partial sums),

$$\frac{f(y) - f(x)}{y - x} - \sum_{n=1}^{\infty} n\,c_n\,x^{n-1} = \sum_{n=1}^{\infty} c_n\left[\frac{y^n - x^n}{y - x} - n\,x^{n-1}\right].$$

The factorization $y^n - x^n = (y - x)\sum_{k=0}^{n-1} y^k x^{n-1-k}$ turns the bracket into $\sum_{k=0}^{n-1}\left(y^k - x^k\right)x^{n-1-k}$, and the same factorization once more bounds $\left|y^k - x^k\right| \le k\,r^{k-1}\,|y - x|$, so

$$\left|\frac{y^n - x^n}{y - x} - n\,x^{n-1}\right| \le |y - x|\sum_{k=0}^{n-1} k\,r^{k-1}\,r^{n-1-k} = |y - x|\,\frac{n(n-1)}{2}\,r^{n-2}.$$

Multiplying by $|c_n|$ and summing,

$$\left|\frac{f(y) - f(x)}{y - x} - \sum_{n=1}^{\infty} n\,c_n\,x^{n-1}\right| \le \frac{|y - x|}{2}\sum_{n=2}^{\infty} n(n-1)\,|c_n|\,r^{n-2} = \frac{K}{2}\,|y - x|,$$

with $K < \infty$ because $r < R$ (the twice-derived series above). Letting $y \to x$, the difference quotient converges to $\sum_n n\,c_n\,x^{n-1}$, which is Equation (7.30). Differentiability implies continuity (Proposition 7.27), and iterating on the derived series — again a power series of radius R — yields derivatives of every order. □

Termwise *integration* is the mirror statement: the antiderivative series $\sum_n c_n(x - x_0)^{n+1}/(n+1)$ has the same radius (the same comparison, run backwards) and differentiates termwise back to f by Theorem 7.51. Both are used freely in Chapter 8, where power series become the *definition* of analyticity.

7.9.1 The numbers e and π

Two real numbers recur in every branch of this treatise, from the oscillator of Chapter 28 to the phase factors of Part IX. Both are defined here, rigorously, from the machinery just built — not quoted as decimal approximations.

Proposition 7.52 (Alternating series test)**.** *Let $(b_n)_{n\ge0}$ be a nonincreasing sequence of nonnegative reals with $b_n \to 0$. Then $\sum_{n=0}^{\infty}(-1)^n b_n$ converges, and its sum S satisfies $S_1 \le S \le S_0 = b_0$, where S_N denotes the N-th partial sum.*

Rests on Definition 7.45 and Corollary A.46.

Proof. *Derives Proposition 7.52.*

The even partial sums are nonincreasing, $S_{2k+2} - S_{2k} = -b_{2k+1} + b_{2k+2} \leq 0$, and bounded below by S_1 (since $S_{2k+1} \leq S_{2k+2} \leq S_{2k}$, shown by the same computation applied once more); the odd partial sums are nondecreasing and bounded above by S_0 by the mirror argument. Both are monotone and bounded, hence convergent (Corollary A.46), to limits S_{e} and S_{o} respectively; since $S_{2k} - S_{2k+1} = b_{2k+1} \to 0$, $S_{\mathrm{e}} = S_{\mathrm{o}} =: S$. Monotonicity of each subsequence gives $S_1 \leq S \leq S_0$. □

Figure 7.5 shows the mechanism at work: each pair of consecutive partial sums traps the limit in an interval of width $b_{N+1} \to 0$, which is what makes the test also an *error bound* — the property Lemma 7.60 will exploit.

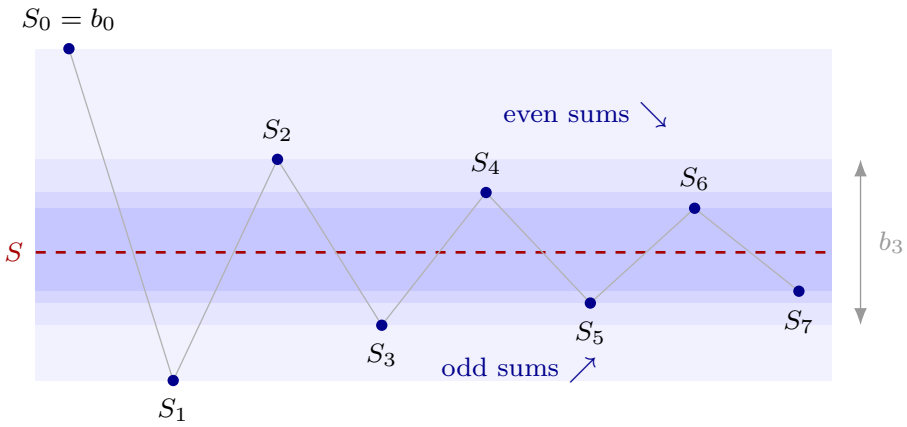


Figure 7.5: The alternating series test (Proposition 7.52) on $\sum_n (-1)^n b_n$ with $b_n = 1/(n+1)$. Even partial sums descend, odd partial sums ascend, and every consecutive pair brackets the sum S in a nested interval of width b_{N+1}.

Euler's number

Definition 7.53 (e)**.**

$$\mathrm{e} = \sum_{n=0}^{\infty} \frac{1}{n!}\,. \tag{7.31}$$

Rests on Definition 7.45 and Proposition 7.48.

The series converges absolutely by the ratio test (Proposition 7.48: the ratio of consecutive terms is $1/(n+1) \to 0$).

Convergence by itself yields no digit. What turns Equation (7.31) into a *computational* definition is that its truncation error admits a closed bound, and one sharp enough that a dozen terms already settle nine decimals.

Lemma 7.54 (Truncation error)**.** *Let $s_n = \sum_{k=0}^{n} 1/k!$ denote the n-th partial sum of Equation (7.31). Then for every integer $n \geq 1$,*

$$\frac{1}{(n+1)!} \;\leq\; \mathrm{e} - s_n \;<\; \frac{1}{n \cdot n!}\,. \tag{7.32}$$

Rests on Equation (7.31) and Proposition 7.46.

Proof. *Derives Lemma 7.54.*

The tail $\mathrm{e} - s_n = \sum_{k=n+1}^{\infty} 1/k!$ has positive terms, so it is at least its first term $1/(n+1)!$; that is the left inequality. For the right one, put $k = n+1+j$ with $j \geq 0$ and factor out $1/(n+1)!$:

$$\mathrm{e} - s_n = \frac{1}{(n+1)!} \sum_{j=0}^{\infty} \prod_{i=2}^{j+1} \frac{1}{n+i}\,, \tag{7.33}$$

the empty product at $j = 0$ being 1. Each of the j factors obeys $n+i \geq n+2 > n+1$, so the product is at most $(n+1)^{-j}$, strictly so once $j \geq 1$. The comparison series is geometric with ratio $1/(n+1) \in (0,1)$, hence summable (Proposition 7.46), and

$$\mathrm{e} - s_n < \frac{1}{(n+1)!} \sum_{j=0}^{\infty} \frac{1}{(n+1)^j} = \frac{1}{(n+1)!} \cdot \frac{n+1}{n} = \frac{1}{n \cdot n!}\,.$$

□

The two bounds trap e between two *rationals* computable in finitely many integer operations, and the width of the trap collapses faster than any geometric rate. This is what produces digits.

Corollary 7.55 (The decimal expansion)**.**

$$\mathrm{e} = 2.718\,281\,828\ldots\,. \tag{7.34}$$

Rests on Lemma 7.54.

Derivation. *Derives Corollary 7.55.*

Take $n = 12$ in Lemma 7.54. Over the common denominator $12! = 479\,001\,600$ the partial sum is a ratio of integers obtained by finitely many additions, with no limit process anywhere:

$$s_{12} = \frac{1}{12!} \sum_{k=0}^{12} \frac{12!}{k!} = \frac{1\,302\,061\,345}{479\,001\,600} = \frac{260\,412\,269}{95\,800\,320}\,. \tag{7.35}$$

The error bound is $1/(12 \cdot 12!) = 1/5\,748\,019\,200$, itself smaller than 1.74×10^{-10}. Carrying out the two divisions to sixteen decimals — rounding the lower endpoint down and the upper endpoint up, so that the enclosure is certain — Equation (7.32) gives

$$2.718\,281\,828\,286\,168\,5 \;<\; \mathrm{e} \;<\; 2.718\,281\,828\,460\,141\,6\,. \tag{7.36}$$

The endpoints agree in their first nine decimals, so every real number between them — e among them — begins 2.718281828. □

Nothing about $n = 12$ is special: raising n narrows Equation (7.36) at will, $n = 15$ already fixing eleven decimals. What the argument does *not* do is terminate. No finite decimal is ever reached, and the next result explains why not.

Proposition 7.56 (Irrationality)**.** $\mathrm{e} \notin \mathbb{Q}$. *Consequently its decimal expansion neither terminates nor becomes periodic, and no finite decimal equals* e*: Equation* (7.34) *records the opening digits of an expansion that never ends.*

Rests on Definition 7.53 and Lemma 7.54.

Proof. *Derives Proposition 7.56.*

Suppose $\mathrm{e} = p/q$ with p, q positive integers, and set $n = \max\{q, 2\}$. Consider

$$x = n!\,(\mathrm{e} - s_n) = n!\,\mathrm{e} - \sum_{k=0}^{n} \frac{n!}{k!}\,.$$

Every $n!/k!$ with $k \le n$ is an integer, so the sum is an integer; and $n!\,\mathrm{e} = n!\,p/q$ is an integer because $q \le n$ makes q a divisor of $n!$. Hence $x \in \mathbb{Z}$. Multiplying Equation (7.32) by $n! > 0$, however,

$$0 < \frac{1}{n+1} \le x < \frac{1}{n} \le \frac{1}{2}\,,$$

so x is an integer lying strictly between 0 and 1. No such integer exists, and the assumption fails. □

Proposition 7.57 (The limit definition)**.**

$$\mathrm{e} = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n . \tag{7.37}$$

Rests on Definition 7.53, Equation (5.3), and Proposition 7.6.

Proof. *Derives Proposition 7.57.*

By the binomial theorem (Equation (5.3)),

$$\left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^{n} \binom{n}{k} \frac{1}{n^k} = \sum_{k=0}^{n} \frac{1}{k!} \prod_{j=0}^{k-1} \left(1 - \frac{j}{n}\right) , \tag{7.38}$$

using $\binom{n}{k}/n^k = \frac{1}{k!}\prod_{j=0}^{k-1}(n-j)/n = \frac{1}{k!}\prod_{j=0}^{k-1}(1-j/n)$. Each factor $1 - j/n \in (0,1]$, so every term of Equation (7.38) is at most $1/k!$; summing,

$$\left(1 + \frac{1}{n}\right)^n \le \sum_{k=0}^{n} \frac{1}{k!} \le \mathrm{e}\,,$$

the last step because a partial sum of a convergent series of positive terms is bounded by the full sum. This gives $\limsup_n (1+1/n)^n \le \mathrm{e}$. Conversely, fix N; for $n \ge N$, truncating Equation (7.38) at $k = N$ (dropping the remaining nonnegative terms) gives

$$\left(1 + \frac{1}{n}\right)^n \ge \sum_{k=0}^{N} \frac{1}{k!} \prod_{j=0}^{k-1} \left(1 - \frac{j}{n}\right) .$$

For each fixed $k \le N$, the product $\to 1$ as $n \to \infty$ (a finite product of terms each $\to 1$, by Proposition 7.6), so the right side $\to \sum_{k=0}^{N} 1/k!$ as $n \to \infty$. Hence $\liminf_n (1+1/n)^n \ge \sum_{k=0}^{N} 1/k!$ for every N, and letting $N \to \infty$, $\liminf_n (1+1/n)^n \ge \mathrm{e}$. The limsup and liminf bounds coincide, so the limit exists and equals e. □

Remark 7.58 (An identity, not a recipe). Proposition 7.57 is often taken as the definition of e, but it is useless for computing one. Taking logarithms and expanding, $n\ln(1+1/n) = 1 - \frac{1}{2n} + O(n^{-2})$, so

$$\left(1 + \frac{1}{n}\right)^n = \mathrm{e}\left(1 - \frac{1}{2n} + O(n^{-2})\right) , \qquad \mathrm{e} - \left(1 + \frac{1}{n}\right)^n \sim \frac{\mathrm{e}}{2n}\,. \tag{7.39}$$

The error falls off like $1/n$: one further decimal per tenfold increase in n. Matching the nine decimals that Equation (7.35) obtains from twelve terms would take $n > 2.7 \times 10^9$, and the eleven decimals available at $n = 15$ would take $n > 2.7 \times 10^{11}$. The factorial series converges faster than any geometric rate and the limit slower than one; that asymmetry, not tradition, is why Equation (7.31) is the definition here and Equation (7.37) a theorem about it.

Rests on Proposition 7.57 and Lemma 7.54.

The real exponential $\exp(x) = \sum_{n=0}^{\infty} x^n/n!$ (Theorem 7.50: radius ∞, ratio test) satisfies $\exp(1) = \mathrm{e}$ and extends to the complex exponential e^z of Definition 8.3 without changing a single coefficient.

The number π

The trigonometric functions are defined, on $\mathbb{R}$, by the same series used in Definition 8.3 restricted to real argument:

$$\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}\,, \qquad \sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}\,, \tag{7.40}$$

both with infinite radius of convergence (ratio test, as for e^x).

Lemma 7.59 (Derivatives; the Pythagorean identity)**.** *For every $x \in \mathbb{R}$,*

$$\sin' x = \cos x\,, \qquad \cos' x = -\sin x\,, \qquad \sin^2 x + \cos^2 x = 1\,. \tag{7.41}$$

Rests on Equations (5.3) and (7.40), Theorem 7.51, and Proposition 7.49.

Proof. *Derives Lemma 7.59.*

Derivatives. Termwise differentiation (Theorem 7.51) of Equation (7.40) gives

$$\sin' x = \sum_{n=0}^{\infty} \frac{(-1)^n (2n+1)\, x^{2n}}{(2n+1)!} = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} = \cos x\,,$$

and, re-indexing $n = m+1$ in the derived series of cos,

$$\cos' x = \sum_{n=1}^{\infty} \frac{(-1)^n (2n)\, x^{2n-1}}{(2n)!} = \sum_{m=0}^{\infty} \frac{(-1)^{m+1} x^{2m+1}}{(2m+1)!} = -\sin x\,.$$

Pythagorean identity. Both series converge absolutely at every x, so each square is a Cauchy product (Proposition 7.49). Squaring the cosine series, the terms $a_j = (-1)^j x^{2j}/(2j)!$ collect at fixed $j+k=n$ into

$$\cos^2 x = \sum_{n=0}^{\infty} (-1)^n x^{2n} \sum_{j=0}^{n} \frac{1}{(2j)!\,(2n-2j)!} = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} \sum_{j=0}^{n} \binom{2n}{2j}\,,$$

multiplying and dividing by $(2n)!$ in the last step. The same computation squares the sine series (terms $(-1)^j x^{2j+1}/(2j+1)!$, product powers x^{2n+2}, then re-indexed so the power reads x^{2n}):

$$\sin^2 x = \sum_{n=1}^{\infty} \frac{(-1)^{n-1} x^{2n}}{(2n)!} \sum_{j=0}^{n-1} \binom{2n}{2j+1}\,.$$

Adding the two, the constant term is 1 and the coefficient of x^{2n} for $n \geq 1$ is

$$\frac{(-1)^n}{(2n)!}\left[\sum_{j=0}^{n} \binom{2n}{2j} - \sum_{j=0}^{n-1} \binom{2n}{2j+1}\right] = \frac{(-1)^n}{(2n)!}\,(1-1)^{2n} = 0\,,$$

by the binomial theorem (Equation (5.3)): the two brackets are the even and the odd halves of $\sum_{i=0}^{2n} (-1)^i \binom{2n}{i}$. Hence $\sin^2 x + \cos^2 x = 1$. (Once the derivatives are known the identity also follows in one line — $(\sin^2 + \cos^2)' = 2\sin\cos - 2\cos\sin = 0$, so the function is constant (Corollary 7.36) with value 1 at $x=0$; the Cauchy-product computation above is that statement read coefficient by coefficient.) □

Lemma 7.60 ($\cos 2 < 0$)**.** *Rests on Equation (7.40) and Proposition 7.52.*

Proof. *Derives Lemma 7.60.*

At $x = 2$ the terms of Equation (7.40) have magnitude $t_n = 2^{2n}/(2n)!$; for $n \geq 1$,

$$\frac{t_{n+1}}{t_n} = \frac{4}{(2n+1)(2n+2)} \leq \frac{4}{3 \cdot 4} = \frac{1}{3},$$

so $(t_n)_{n\geq 1}$ is nonincreasing (and $t_n \to 0$). By Proposition 7.52 applied to the tail starting at $n = 2$, the sum $S = \sum_{n=2}^{\infty}(-1)^n t_n = t_2 - t_3 + t_4 - \cdots$ satisfies $0 \leq S \leq t_2 = 16/24 = 2/3$. Since $\cos 2 = t_0 - t_1 + S = 1 - 2 + S = -1 + S$,

$$\cos 2 \in \left[-1, -\tfrac{1}{3}\right],$$

in particular $\cos 2 < 0$. □

Definition 7.61 (π)**.** By Lemma 7.60, $\cos 0 = 1 > 0$, and cos is continuous (Theorem 7.51: a power series is differentiable, hence continuous, inside its radius — here infinite). The intermediate value theorem (Theorem 7.23) applied to cos on $[0, 2]$ gives at least one zero (Figure 7.6); let

$$z = \inf\{x > 0 \mid \cos x = 0\},$$

a nonempty set bounded below by 0 and above by 2, so z exists by completeness (Axiom 7.1) and $\cos z = 0$ by continuity (a limit of a sequence of zeros approaching the infimum, or directly: cos > 0 on $[0, z)$ by minimality of z, and $\cos z \leq 0$ by continuity from the right, forcing $\cos z = 0$). We define

$$\pi = 2z. \tag{7.42}$$

Rests on Lemma 7.60, Theorems 7.23 and 7.51, and Axiom 7.1.

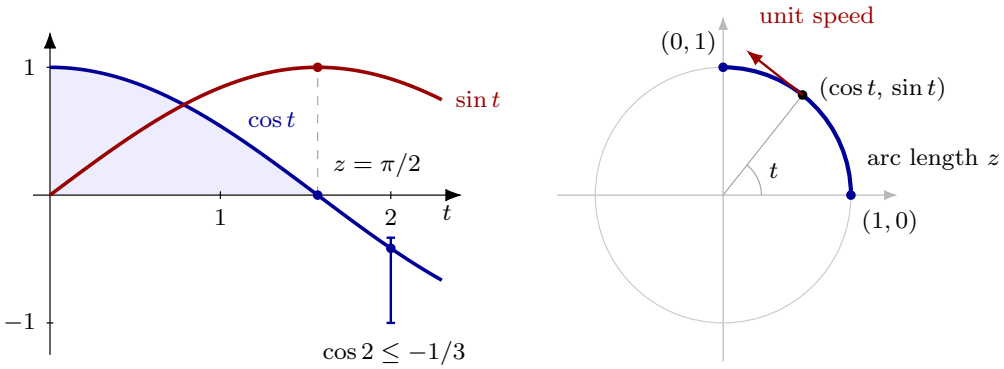


Figure 7.6: Left: the series Equation (7.40) pin $\cos 0 = 1$ and $\cos 2 \leq -\frac{1}{3}$ (Lemma 7.60), so the first positive zero $z = \pi/2$ of cos exists by the intermediate value theorem (Definition 7.61). Right: on $[0, z]$ the path $t \mapsto (\cos t, \sin t)$ traverses a quarter of the unit circle once, at unit speed, so the arc has length exactly z (Proposition 7.62).

The *arc length* of a continuously differentiable path $\gamma : [a, b] \longrightarrow \mathbb{R}^2$ is $\int_a^b |\gamma'(t)|\, \mathrm{d}t$, the integral of its speed — the notion taken up systematically in Definition 13.11. Only two of its immediate properties are needed here: reflections preserve it, since they leave the speed $|\gamma'|$ unchanged, and the circumference of a circle is the sum of the lengths of arcs covering it with only endpoints in common.

Proposition 7.62 (π as the circle constant)**.** *The arc length of the unit circle is* 2π*.*

Rests on Definitions 7.61 and 13.11, Lemma 7.59, Corollary 7.36, and Theorem 7.23.

Proof. *Derives Proposition 7.62.*

On $[0, z)$, $\cos > 0$ by minimality of z in Definition 7.61, and $\sin' = \cos$ (Lemma 7.59), so sin is strictly increasing on $[0, z]$ (Corollary 7.36); from $\sin 0 = 0$ it is positive on $(0, z]$, and $\sin^2 z = 1 - \cos^2 z = 1$ (Lemma 7.59 again, with $\cos z = 0$) forces $\sin z = 1$. The path $t \mapsto (\cos t, \sin t)$, $t \in [0, z]$, therefore lies on the unit circle by the Pythagorean identity, starts at $(1, 0)$, ends at $(0, 1)$, and traverses the first-quadrant arc exactly once: sin is a continuous strictly increasing bijection of $[0, z]$ onto $[0, 1]$ (Theorem 7.23), and on that arc the height $\sin t$ determines the point, since $\cos t \geq 0$. The speed of the path is

$$\sqrt{\cos'^2 t + \sin'^2 t} = \sqrt{\sin^2 t + \cos^2 t} = 1$$

(both parts of Lemma 7.59), so its arc length is $\int_0^z 1 \, \mathrm{d}t = z$. The reflections $(x, y) \mapsto (-x, y)$ and $(x, y) \mapsto (x, -y)$ map the unit circle to itself and preserve arc length, so the four quadrantal arcs have the same length z, and the full circumference is $4z = 2\pi$. □

Scaling by $r > 0$ multiplies every arc length by r (the scaled path $r\gamma$ has speed $r\,|\gamma'|$, pulling the factor r out of the integral), so a circle of radius r has circumference $2\pi r$: *the circumference of any circle is π times its diameter* (Figure 7.7).

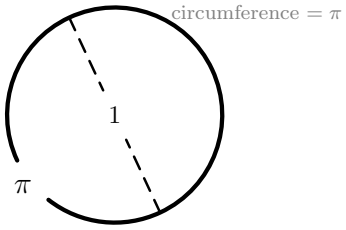


Figure 7.7: The circle constant, reduced to its emblem: a circle of diameter 1 has circumference π (Proposition 7.62). The number so drawn — twice the first positive zero of cos (Definition 7.61) — is the same π in every formula of this treatise.

Remark 7.63 (π elsewhere in this treatise). The connection between Definition 7.61 (an analytic definition, the first positive zero of cos) and Euler's identity $\mathrm{e}^{\mathrm{i}\pi} = -1$ is made in Proposition 8.4: π is exactly half the period of $\theta \mapsto \mathrm{e}^{\mathrm{i}\theta}$ (Lemma 8.11), the same constant defined here through cos alone. Every subsequent appearance of π in this treatise — the $2\pi\mathrm{i}$ of the residue theorem, the phase-space volumes of statistical mechanics, the angles of Part III — is this one number.

7.10 Multivariable calculus

7.10.1 Planes in $\mathbb{R}^3$

Definition 7.64 (Plane)**.** Let $\boldsymbol{n} = (a, b, c) \neq \boldsymbol{0}$, $\boldsymbol{p}_0 = (x_0, y_0, z_0)$, and $\boldsymbol{p} = (x, y, z)$. A *plane* is the set of points $\boldsymbol{p}$ satisfying

$$\boldsymbol{n} \cdot \boldsymbol{p} = \boldsymbol{n} \cdot \boldsymbol{p}_0 \,, \tag{7.43}$$

where $\boldsymbol{n} \cdot \boldsymbol{p}_0$ is a real constant.

Rests on Definition 5.17.

7.10.2 Differentiable functions

Let $\boldsymbol{F} : A \subseteq \mathbb{R}^m \longrightarrow \mathbb{R}^n$ be a function, with $\boldsymbol{x} = (x_1, \ldots, x_m)$, $\boldsymbol{F} = (F_1(\boldsymbol{x}), \ldots, F_n(\boldsymbol{x}))$, and A an open set.

Definition 7.65 (Partial derivative; gradient)**.** The *k-th partial derivative* of a scalar function $f : A \subseteq \mathbb{R}^m \longrightarrow \mathbb{R}$ at $\boldsymbol{a} \in A$ is the ordinary derivative of the one-variable function obtained by freezing every coordinate but the k-th,

$$\frac{\partial f}{\partial x_k}(\boldsymbol{a}) = \partial_k f(\boldsymbol{a}) = \lim_{t \longrightarrow 0} \frac{f(\boldsymbol{a} + t\,\boldsymbol{e}_k) - f(\boldsymbol{a})}{t}\,, \tag{7.44}$$

where $\boldsymbol{e}_k$ is the k-th vector of the canonical basis of $\mathbb{R}^m$. The *gradient* $\nabla f(\boldsymbol{a}) = (\partial_1 f(\boldsymbol{a}), \ldots, \partial_m f(\boldsymbol{a}))$ collects the m of them.

Rests on Definitions 5.14 and 7.26.

Definition 7.66 (Functions of class C^1)**.** We say that $\boldsymbol{F}$ is C^1 on A if and only if the m first-order partial derivatives of each of its n components have domain A and are continuous on A.

Rests on Definitions 7.20 and 7.65.

Definition 7.67 (Differentiability at a point)**.** $\boldsymbol{F}$ is *differentiable* at $\boldsymbol{a} \in A$ if and only if there is a linear map $L : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ such that

$$\lim_{\boldsymbol{h} \longrightarrow \boldsymbol{0}} \frac{|\boldsymbol{F}(\boldsymbol{a} + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{a}) - L(\boldsymbol{h})|}{|\boldsymbol{h}|} = 0\,, \tag{7.45}$$

the bars denoting Euclidean length (Equation (6.12)). Such an L is unique: if L and L' both satisfy Equation (7.45), subtracting gives $|(L - L')(\boldsymbol{h})| \,/\, |\boldsymbol{h}| \longrightarrow 0$, and taking $\boldsymbol{h} = t\,\boldsymbol{u}$ with $|\boldsymbol{u}| = 1$ and $t \longrightarrow 0$ leaves $(L - L')(\boldsymbol{u}) = \boldsymbol{0}$ for every unit vector, so $L = L'$. The same substitution also identifies L: taking $\boldsymbol{h} = t\,\boldsymbol{e}_k$ in Equation (7.45) and letting $t \longrightarrow 0$ gives $(\boldsymbol{F}(\boldsymbol{a} + t\,\boldsymbol{e}_k) - \boldsymbol{F}(\boldsymbol{a}))\,/t \longrightarrow L(\boldsymbol{e}_k)$, so each partial derivative $\partial_k F_j(\boldsymbol{a})$ exists (Definition 7.65) and $L(\boldsymbol{e}_k)$ is the vector of the n of them. The matrix of L in the canonical bases is therefore $(\partial_k F_j(\boldsymbol{a}))$, whose j-th row is the gradient $\nabla F_j(\boldsymbol{a})$. Note what this does and does not say: differentiability at a single point already forces the partial derivatives to exist and to assemble into the differential, with nothing assumed about their continuity — but the converse fails, as Remark 7.69 shows.

Rests on Definitions 5.35 and 7.16 and Equation (6.12).

Theorem 7.68 (C^1 implies differentiable)**.** *Let $A \subseteq \mathbb{R}^m$ be open and let $\boldsymbol{F} : A \longrightarrow \mathbb{R}^n$ be of class C^1. Then $\boldsymbol{F}$ is differentiable at every $\boldsymbol{a} \in A$, and its differential is the linear map*

$$L(\boldsymbol{h}) = (\nabla F_1(\boldsymbol{a}) \cdot \boldsymbol{h}, \ldots, \nabla F_n(\boldsymbol{a}) \cdot \boldsymbol{h})\,, \tag{7.46}$$

that is, the differential of Definition 7.70 below.

Rests on Definitions 7.66 and 7.67 and Theorem 7.35.

Proof. *Derives Theorem 7.68.*

Reduction to one component. With L as in Equation (7.46), the j-th component of $\boldsymbol{F}(\boldsymbol{a} + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{a}) - L(\boldsymbol{h})$ is $R_j(\boldsymbol{h}) = F_j(\boldsymbol{a} + \boldsymbol{h}) - F_j(\boldsymbol{a}) - \nabla F_j(\boldsymbol{a}) \cdot \boldsymbol{h}$, and the Euclidean length of a vector is at most the sum of the absolute values of its components. It therefore suffices to prove that $|R_j(\boldsymbol{h})| \,/\, |\boldsymbol{h}| \longrightarrow 0$ for a single scalar C^1 function $f = F_j$.

A staircase of coordinate steps. Since A is open, fix $r > 0$ with $B_r(\boldsymbol{a}) \subseteq A$ (Definition 6.26) and let $|\boldsymbol{h}| < r$. Put $\boldsymbol{p}_0 = \boldsymbol{a}$ and

$$\boldsymbol{p}_k = \boldsymbol{a} + \sum_{i=1}^{k} h_i\,\boldsymbol{e}_i\,, \qquad k = 1, \ldots, m\,,$$

so that $\boldsymbol{p}_m = \boldsymbol{a} + \boldsymbol{h}$. Every point of the segment joining $\boldsymbol{p}_{k-1}$ to $\boldsymbol{p}_k$ has the form $\boldsymbol{a} + \sum_{i<k} h_i \boldsymbol{e}_i + t\,h_k \boldsymbol{e}_k$ with $t \in [0, 1]$, and its distance to $\boldsymbol{a}$ is at most $|\boldsymbol{h}| < r$: the whole staircase lies inside $B_r(\boldsymbol{a})$. Telescoping,

$$f(\boldsymbol{a} + \boldsymbol{h}) - f(\boldsymbol{a}) = \sum_{k=1}^{m} [f(\boldsymbol{p}_k) - f(\boldsymbol{p}_{k-1})]\,. \tag{7.47}$$

The mean value theorem on each step. The k-th bracket moves only the k-th coordinate, so it is the increment over $[0,1]$ of the one-variable function $g_k(t) = f(\boldsymbol{p}_{k-1} + t\,h_k \boldsymbol{e}_k)$. Assume $h_k \neq 0$. The difference quotient of g_k at t with increment s becomes, on substituting $u = s\,h_k$, exactly h_k times the quotient appearing in Equation (7.44), so $g_k'(t) = h_k\,\partial_k f(\boldsymbol{p}_{k-1} + t\,h_k \boldsymbol{e}_k)$ wherever the partial derivative $\partial_k f$ exists — that is, throughout $B_r(\boldsymbol{a})$. Both endpoints of the segment lie at distance $|\boldsymbol{h}| < r$ from $\boldsymbol{a}$, hence in the interior of the ball, so g_k is defined and differentiable on an open interval containing $[0,1]$ and is in particular continuous there (Proposition 7.27). By Theorem 7.35 there is $\theta_k \in (0,1)$ with

$$f(\boldsymbol{p}_k) - f(\boldsymbol{p}_{k-1}) = h_k\,\partial_k f(\boldsymbol{\xi}_k)\,, \qquad \boldsymbol{\xi}_k = \boldsymbol{p}_{k-1} + \theta_k h_k \boldsymbol{e}_k\,, \tag{7.48}$$

and $|\boldsymbol{\xi}_k - \boldsymbol{a}| \leq |\boldsymbol{h}|$. If $h_k = 0$ both sides of Equation (7.48) vanish and $\boldsymbol{\xi}_k = \boldsymbol{p}_{k-1}$ serves.

The estimate. Subtracting $\nabla f(\boldsymbol{a}) \cdot \boldsymbol{h} = \sum_k h_k\,\partial_k f(\boldsymbol{a})$ from Equation (7.47) and inserting Equation (7.48),

$$f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a}) - \nabla f(\boldsymbol{a}) \cdot \boldsymbol{h} = \sum_{k=1}^{m} h_k \left[\partial_k f(\boldsymbol{\xi}_k) - \partial_k f(\boldsymbol{a})\right]\,,$$

so that, since $|h_k| \leq |\boldsymbol{h}|$ for every k,

$$\frac{|f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a}) - \nabla f(\boldsymbol{a}) \cdot \boldsymbol{h}|}{|\boldsymbol{h}|} \leq \sum_{k=1}^{m} |\partial_k f(\boldsymbol{\xi}_k) - \partial_k f(\boldsymbol{a})|\,. \tag{7.49}$$

Given $\varepsilon > 0$, continuity of the m partial derivatives at $\boldsymbol{a}$ (Definition 7.66) provides $\delta \in (0, r)$ such that $|\boldsymbol{y} - \boldsymbol{a}| < \delta$ implies $|\partial_k f(\boldsymbol{y}) - \partial_k f(\boldsymbol{a})| < \varepsilon/m$ for every k. For $|\boldsymbol{h}| < \delta$ each $\boldsymbol{\xi}_k$ obeys $|\boldsymbol{\xi}_k - \boldsymbol{a}| \leq |\boldsymbol{h}| < \delta$, so the right-hand side of Equation (7.49) is smaller than ε. This is Equation (7.45). □

Remark 7.69 (Partial derivatives alone do not suffice). Continuity of the partial derivatives is doing real work in Theorem 7.68: their bare existence does not even force continuity of the function, let alone differentiability. The standard witness on $\mathbb{R}^2$ is

$$f(x,y) = \frac{xy}{x^2+y^2} \quad \text{for } (x,y) \neq (0,0)\,, \qquad f(0,0) = 0\,. \tag{7.50}$$

Both partial derivatives exist at the origin and vanish there, because f vanishes identically on each axis; yet $f(x,x) = 1/2$ for every $x \neq 0$, so f has no limit at the origin and is discontinuous there. The partial derivatives of this f are accordingly not continuous at the origin, and the hypothesis of Theorem 7.68 fails.

Rests on Theorem 7.68 and Definition 7.20.

Definition 7.70 (Differential of a function)**.** We say that the linear map

$$\begin{aligned} \mathrm{d}\boldsymbol{F}(\boldsymbol{a}) : \mathbb{R}^m &\longrightarrow \mathbb{R}^n \\ \boldsymbol{x} &\longmapsto \mathrm{d}\boldsymbol{F}(\boldsymbol{a})(\boldsymbol{x}) = (\nabla F_1(\boldsymbol{a}) \cdot \boldsymbol{x}, \ldots, \nabla F_n(\boldsymbol{a}) \cdot \boldsymbol{x}) \end{aligned}$$

is the *differential* of $\boldsymbol{F}$ at $\boldsymbol{a} \in A$ if and only if $\boldsymbol{F}$ is differentiable at $\boldsymbol{a}$. The two prescriptions agree whenever they apply at all: if $\boldsymbol{F}$ is differentiable at $\boldsymbol{a}$ then the unique linear map of Definition 7.67 has exactly the matrix written here, as identified there — differentiability alone, with no continuity assumed of the partial derivatives. Theorem 7.68 adds that such a map exists whenever $\boldsymbol{F}$ is C^1.

Rests on Definitions 7.65 and 7.67.

Remark 7.71. If $n = 1$, one has

$$\mathrm{d}F_1(\boldsymbol{a}) = \nabla F_1(\boldsymbol{a}) \cdot \boldsymbol{x}\,. \tag{7.51}$$

Rests on Definition 7.70.

Proposition 7.72 (Chain rule in several variables)**.** *Let $A \subseteq \mathbb{R}^m$ and $B \subseteq \mathbb{R}^p$ be open, let $\boldsymbol{G} : A \longrightarrow B$ be differentiable at $\boldsymbol{a} \in A$, and let $\boldsymbol{F} : B \longrightarrow \mathbb{R}^n$ be differentiable at $\boldsymbol{b} = \boldsymbol{G}(\boldsymbol{a})$. Then $\boldsymbol{F} \circ \boldsymbol{G}$ is differentiable at $\boldsymbol{a}$, with differential the composite $\mathrm{d}\boldsymbol{F}(\boldsymbol{b}) \circ \mathrm{d}\boldsymbol{G}(\boldsymbol{a})$; in components,*

$$\partial_j \left(F_i \circ \boldsymbol{G}\right)(\boldsymbol{a}) = \sum_{l=1}^{p} \partial_l F_i(\boldsymbol{b})\, \partial_j G_l(\boldsymbol{a})\,. \tag{7.52}$$

Rests on Definitions 7.13, 7.67 and 7.70.

Proof. *Derives Proposition 7.72.*

Write $L = \mathrm{d}\boldsymbol{G}(\boldsymbol{a})$ and $M = \mathrm{d}\boldsymbol{F}(\boldsymbol{b})$. A linear map between Euclidean spaces is bounded: expanding in the canonical bases and applying the Cauchy–Schwarz inequality (Chapter 5) row by row produces constants C_L, C_M with $|L\boldsymbol{v}| \le C_L\,|\boldsymbol{v}|$ and $|M\boldsymbol{w}| \le C_M\,|\boldsymbol{w}|$. By Definition 7.67,

$$\begin{aligned} \boldsymbol{G}(\boldsymbol{a}+\boldsymbol{h}) &= \boldsymbol{b} + L\boldsymbol{h} + \boldsymbol{\rho}(\boldsymbol{h})\,, \\ \boldsymbol{F}(\boldsymbol{b}+\boldsymbol{k}) &= \boldsymbol{F}(\boldsymbol{b}) + M\boldsymbol{k} + \boldsymbol{\sigma}(\boldsymbol{k})\,, \end{aligned}$$

where $|\boldsymbol{\rho}(\boldsymbol{h})| \,/\, |\boldsymbol{h}| \longrightarrow 0$, and where $|\boldsymbol{\sigma}(\boldsymbol{k})| = \varepsilon(\boldsymbol{k})\ |\boldsymbol{k}|$ with $\varepsilon(\boldsymbol{k}) \longrightarrow 0$ as $\boldsymbol{k} \longrightarrow \boldsymbol{0}$; setting $\varepsilon(\boldsymbol{0}) = 0$ is consistent, since $\boldsymbol{\sigma}(\boldsymbol{0}) = \boldsymbol{0}$. Put $\boldsymbol{k}(\boldsymbol{h}) = L\boldsymbol{h} + \boldsymbol{\rho}(\boldsymbol{h})$, so that

$$\boldsymbol{F}(\boldsymbol{G}(\boldsymbol{a}+\boldsymbol{h})) = \boldsymbol{F}(\boldsymbol{b}) + ML\,\boldsymbol{h} + M\boldsymbol{\rho}(\boldsymbol{h}) + \boldsymbol{\sigma}(\boldsymbol{k}(\boldsymbol{h}))\,.$$

The first remainder obeys $|M\boldsymbol{\rho}(\boldsymbol{h})| \le C_M\,|\boldsymbol{\rho}(\boldsymbol{h})|$ and is therefore $o(|\boldsymbol{h}|)$. For the second, taking $|\boldsymbol{h}|$ small enough that $|\boldsymbol{\rho}(\boldsymbol{h})| \le |\boldsymbol{h}|$ gives $|\boldsymbol{k}(\boldsymbol{h})| \le (C_L + 1)\,|\boldsymbol{h}|$, whence

$$|\boldsymbol{\sigma}(\boldsymbol{k}(\boldsymbol{h}))| = \varepsilon(\boldsymbol{k}(\boldsymbol{h}))\ |\boldsymbol{k}(\boldsymbol{h})| \le (C_L+1)\,\varepsilon(\boldsymbol{k}(\boldsymbol{h}))\ |\boldsymbol{h}|\,,$$

and $\boldsymbol{k}(\boldsymbol{h}) \longrightarrow \boldsymbol{0}$ with $\boldsymbol{h}$, so this remainder is $o(|\boldsymbol{h}|)$ as well. Hence ML satisfies Equation (7.45) for $\boldsymbol{F} \circ \boldsymbol{G}$ and is its differential, by the uniqueness established in Definition 7.67. Reading off the matrix of ML in the canonical bases, the rows of L and of M being the gradients of the components — which mere differentiability already secures, by the identification in Definition 7.67 — is Equation (7.52). □

Proposition 7.73 (Clairaut–Schwarz)**.** *If $f : A \subseteq \mathbb{R}^2 \longrightarrow \mathbb{R}$ has continuous second partial derivatives, then $\partial_x \partial_y f = \partial_y \partial_x f$.*

Rests on Theorem 7.35 and Definition 7.20.

Proof. *Derives Proposition 7.73.*

Fix (x, y) and consider the second difference $\Delta(h,k) = f(x+h, y+k) - f(x+h, y) - f(x, y+k) + f(x, y)$. Applying the mean value theorem (Theorem 7.35) in x to $u(t) = f(t, y+k) - f(t, y)$ gives $\Delta = [\partial_x f(\xi, y+k) - \partial_x f(\xi, y)]\,h$ for some ξ, and a second application in y gives $\Delta = \partial_y \partial_x f(\xi, \eta)\,hk$. Symmetrically, $\Delta = \partial_x \partial_y f(\xi', \eta')\,hk$. Dividing by hk and letting $(h,k) \to 0$, continuity of the second partials forces $\partial_y \partial_x f(x,y) = \partial_x \partial_y f(x,y)$. □

7.10.3 Taylor expansion in several variables

Theorem 7.74 (Taylor's theorem in several variables)**.** *Let $A \subseteq \mathbb{R}^m$ be open, let $f : A \longrightarrow \mathbb{R}$ have continuous second partial derivatives on A, let $\boldsymbol{a} \in A$, and let $\boldsymbol{h} \in \mathbb{R}^m$ be small enough that the whole segment $\{\boldsymbol{a} + t\boldsymbol{h} \mid t \in [0,1]\}$ lies in A. Write*

$$H_{ij}(\boldsymbol{x}) = \partial_i \partial_j f(\boldsymbol{x}) \tag{7.53}$$

for the Hessian *of f, a symmetric matrix by Proposition 7.73. Then there is $\theta \in (0,1)$ with*

$$f(\boldsymbol{a}+\boldsymbol{h}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a}) \cdot \boldsymbol{h} + \frac{1}{2} \sum_{i,j=1}^{m} H_{ij}(\boldsymbol{a} + \theta\boldsymbol{h})\, h_i h_j\,, \tag{7.54}$$

and consequently, as $\boldsymbol{h} \longrightarrow \mathbf{0}$*,*

$$f(\boldsymbol{a}+\boldsymbol{h}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a}) \cdot \boldsymbol{h} + \frac{1}{2}\, \boldsymbol{h}^{\mathsf{T}} H(\boldsymbol{a})\, \boldsymbol{h} + o\Big(|\boldsymbol{h}|^2\Big)\,. \tag{7.55}$$

Rests on Theorem 7.38 and Propositions 7.72 and 7.73.

Proof. *Derives Theorem 7.74.*

Since A is open and the segment is compact, there is $\varepsilon > 0$ with $\boldsymbol{a} + t\boldsymbol{h} \in A$ for every $t \in (-\varepsilon, 1+\varepsilon)$; put $g(t) = f(\boldsymbol{a} + t\boldsymbol{h})$ there. The map $t \longmapsto \boldsymbol{a} + t\boldsymbol{h}$ is differentiable with differential $\boldsymbol{h}$, so the chain rule Equation (7.52) gives

$$g'(t) = \sum_{i=1}^{m} \partial_i f(\boldsymbol{a} + t\boldsymbol{h})\, h_i\,, \qquad g''(t) = \sum_{i,j=1}^{m} \partial_j \partial_i f(\boldsymbol{a} + t\boldsymbol{h})\, h_i h_j\,, \tag{7.56}$$

the second by applying the same rule to each $\partial_i f$, which is C^1 by hypothesis. Hence g is twice differentiable on an interval containing $[0,1]$, and Theorem 7.38 at order $n = 1$ supplies $\theta \in (0,1)$ with $g(1) = g(0) + g'(0) + \frac{1}{2} g''(\theta)$, which is Equation (7.54).

For Equation (7.55), subtract the two quadratic terms: the difference is

$$\frac{1}{2} \sum_{i,j=1}^{m} \big[H_{ij}(\boldsymbol{a} + \theta\boldsymbol{h}) - H_{ij}(\boldsymbol{a})\big]\, h_i h_j\,,$$

whose modulus is at most $\frac{1}{2}\,|\boldsymbol{h}|^2 \sum_{i,j} |H_{ij}(\boldsymbol{a} + \theta\boldsymbol{h}) - H_{ij}(\boldsymbol{a})|$, since $|h_i| \le |\boldsymbol{h}|$ for every i. As $\boldsymbol{h} \longrightarrow \mathbf{0}$ the point $\boldsymbol{a} + \theta\boldsymbol{h}$ tends to $\boldsymbol{a}$ whatever value $\theta \in (0,1)$ takes, so continuity of the second partial derivatives sends that sum to zero. The difference is therefore $o(|\boldsymbol{h}|^2)$. □

Remark 7.75 (What the Hessian is for). Equation (7.55) is the statement behind every expansion of a potential about an equilibrium: at a stationary point $\nabla f(\boldsymbol{a}) = \mathbf{0}$ the leading behaviour is the quadratic form $\boldsymbol{h}^{\mathsf{T}} H(\boldsymbol{a}) \boldsymbol{h}$, whose signature (Theorem 5.112) decides whether the point is a minimum, a maximum or a saddle, and whose symmetry — Proposition 7.73, not a modelling assumption — is what allows it to be diagonalized by an orthogonal change of axes.

Rests on Theorems 5.112 and 7.74.

Proposition 7.76 (Euler's theorem on homogeneous functions)**.** *Let* $U \subseteq \mathbb{R}^n$ *be open and a* cone*, meaning that* $\mu\boldsymbol{u} \in U$ *whenever* $\boldsymbol{u} \in U$ *and* $\mu > 0$*, and let* $f : U \longrightarrow \mathbb{R}$ *be differentiable. Then* f *is* homogeneous of degree k*,*

$$f(\mu\boldsymbol{u}) = \mu^k f(\boldsymbol{u})\,, \qquad \mu > 0\,, \quad \boldsymbol{u} \in U\,, \tag{7.57}$$

if and only if

$$\sum_{i=1}^{n} u^i\, \frac{\partial f}{\partial u^i}(\boldsymbol{u}) = \boldsymbol{u} \cdot \nabla f(\boldsymbol{u}) = k\, f(\boldsymbol{u}) \tag{7.58}$$

at every $\boldsymbol{u} \in U$*.*

Rests on Propositions 7.30 and 7.72 and Corollary 7.36.

Proof. *Derives Proposition 7.76.*

Fix $\boldsymbol{u} \in U$. Because U is a cone, $\mu \longmapsto f(\mu\boldsymbol{u})$ is defined for every $\mu > 0$, and the chain rule Equation (7.52) makes it differentiable with

$$\frac{\mathrm{d}}{\mathrm{d}\mu} f(\mu\boldsymbol{u}) = \sum_{i=1}^{n} u^i\, \partial_i f(\mu\boldsymbol{u}) = \frac{1}{\mu}\, (\mu\boldsymbol{u}) \cdot \nabla f(\mu\boldsymbol{u})\,. \tag{7.59}$$

If Equation (7.57) holds, differentiate it in μ: the left-hand side is Equation (7.59) and the right-hand side is $k\mu^{k-1} f(\boldsymbol{u})$. Setting $\mu = 1$ leaves Equation (7.58).

Conversely, assume Equation (7.58) throughout U and put $\psi(\mu) = \mu^{-k} f(\mu \boldsymbol{u})$ for $\mu > 0$. By the product rule (Proposition 7.30) and Equation (7.59),

$$\psi'(\mu) = \mu^{-k-1} \left[(\mu \boldsymbol{u}) \cdot \nabla f(\mu \boldsymbol{u}) - k\, f(\mu \boldsymbol{u}) \right] = 0 \,,$$

the bracket vanishing because Equation (7.58) holds at the point $\mu \boldsymbol{u} \in U$. A function with vanishing derivative on an interval is constant (Corollary 7.36), so $\psi(\mu) = \psi(1) = f(\boldsymbol{u})$ for every $\mu > 0$, which is Equation (7.57). □

7.10.4 Differentiation under the integral sign

An integral over a fixed region, of an integrand carrying a parameter, is a function of that parameter, and the question when it may be differentiated by differentiating the integrand is answered by the mean value theorem and uniform continuity alone. The statement is placed here, beside the chain rule, because that is what it is used with; the multiple integral it is written for is Definition 7.93 below, and the one-dimensional case is contained in it as $N = 1$.

Theorem 7.77 (Leibniz integral rule)**.** *Let $D \subset \mathbb{R}^N$ be a compact region that decomposes into finitely many simple regions (Definition 7.95), of volume $V = \int_D 1$; let $I \subseteq \mathbb{R}$ be an open interval; and let $f : D \times I \longrightarrow \mathbb{R}$ be continuous with continuous partial derivative $\partial f / \partial t$ on $D \times I$. Then $F(t) = \int_D f(\boldsymbol{x}, t)\, \mathrm{d}V$ is differentiable on I and*

$$\frac{\mathrm{d}F}{\mathrm{d}t}(t) = \int_D \frac{\partial f}{\partial t}(\boldsymbol{x}, t)\, \mathrm{d}V \,, \qquad t \in I \,, \tag{7.60}$$

the right-hand side being itself continuous in t.

Rests on Definition 7.93 and Theorems 7.25 and 7.35.

Proof. *Derives Theorem 7.77.*

Fix $t \in I$ and choose $\eta > 0$ with $J = [t - \eta, t + \eta] \subset I$. The set $D \times J$ is closed and bounded in $\mathbb{R}^{N+1}$, hence compact (Theorem 6.12), so $\partial f / \partial t$ is uniformly continuous on it (Theorem 7.25). Both $f(\cdot, s)$ and $\partial_t f(\cdot, s)$ are continuous on D, hence integrable there (Remark 7.96); write $G(t) = \int_D \partial_t f(\boldsymbol{x}, t)\, \mathrm{d}V$.

Let $0 < |h| \leq \eta$. For each fixed $\boldsymbol{x} \in D$ the one-variable function $s \longmapsto f(\boldsymbol{x}, s)$ is differentiable on J, so the mean value theorem (Theorem 7.35) supplies $\theta = \theta(\boldsymbol{x}, h) \in (0, 1)$ with

$$\frac{f(\boldsymbol{x}, t + h) - f(\boldsymbol{x}, t)}{h} = \partial_t f\left(\boldsymbol{x}, t + \theta h\right) . \tag{7.61}$$

Subtracting $\partial_t f(\boldsymbol{x}, t)$, integrating over D and bounding the integral of a function by the volume times its supremum,

$$\left| \frac{F(t + h) - F(t)}{h} - G(t) \right| \leq V \sup_{\boldsymbol{x} \in D} \left| \partial_t f\left(\boldsymbol{x}, t + \theta h\right) - \partial_t f(\boldsymbol{x}, t) \right| . \tag{7.62}$$

Given $\varepsilon > 0$, uniform continuity provides $\delta \in (0, \eta]$ such that $|s - s'| < \delta$ implies $|\partial_t f(\boldsymbol{x}, s) - \partial_t f(\boldsymbol{x}, s')| < \varepsilon / (V + 1)$ for every $\boldsymbol{x} \in D$. Since $|(t + \theta h) - t| \leq |h|$, taking $|h| < \delta$ makes the right-hand side of Equation (7.62) smaller than ε; this is Equation (7.60). The same bound with $t + \theta h$ replaced by an arbitrary $t' \in J$ gives $|G(t) - G(t')| \leq V \sup_D |\partial_t f(\cdot, t) - \partial_t f(\cdot, t')|$, so G is continuous. □

Corollary 7.78 (Variable limits of integration)**.** *Let f and $\partial f / \partial t$ be continuous on $[c, d] \times I$ and let $\alpha, \beta : I \longrightarrow [c, d]$ be of class C^1. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\alpha(t)}^{\beta(t)} f(x, t)\, \mathrm{d}x = f\left(\beta(t), t\right) \beta'(t) - f\left(\alpha(t), t\right) \alpha'(t) + \int_{\alpha(t)}^{\beta(t)} \frac{\partial f}{\partial t}(x, t)\, \mathrm{d}x \,. \tag{7.63}$$

Rests on Theorems 7.42 and 7.77 and Proposition 7.72.

Proof. *Derives Corollary 7.78.*

Put $\Phi(u,v,t) = \int_u^v f(x,t)\,\mathrm{d}x$ on $[c,d]\times[c,d]\times I$. The fundamental theorem of calculus (Theorem 7.42), together with the sign rule for exchanging the limits of integration, gives $\partial_v\Phi = f(v,t)$ and $\partial_u\Phi = -f(u,t)$; Theorem 7.77 applied on the interval with endpoints u and v gives $\partial_t\Phi = \int_u^v \partial_t f(x,t)\,\mathrm{d}x$. All three are continuous in (u,v,t): the first two because f is, the third by the last clause of Theorem 7.77 together with the estimate $\left|\int_u^v g - \int_{u'}^{v'} g\right| \le (|u-u'|+|v-v'|)\sup|g|$ applied to $g = \partial_t f$. Hence Φ is C^1 and therefore differentiable (Theorem 7.68), and the chain rule Equation (7.52) applied to $t \longmapsto \Phi\left(\alpha(t),\beta(t),t\right)$ is Equation (7.63). □

The derivative of a determinant has already appeared, in Proposition 5.1 of Chapter 5, differentiated a column at a time — the form wanted when the columns are the objects that move. The same derivative resolved into the n^2 individual entries is the form wanted when the entries are, and it is what the change-of-variables and Liouville arguments below use. The two are the same identity: expand each differentiated column of the column form along that column and its cofactors, which contain no entry of it, and the column sum becomes the entry sum term for term.

Proposition 7.79 (Jacobi's formula, cofactor form)**.** *Let $I \subseteq \mathbb{R}$ be an open interval and let $A : I \longrightarrow \mathbb{R}^{n\times n}$ be a differentiable matrix-valued function, with entries $A^i{}_j(t)$. Then $t \longmapsto \det A(t)$ is differentiable and*

$$\frac{\mathrm{d}}{\mathrm{d}t}\det A = \sum_{i,j=1}^{n} C^i{}_j\,\frac{\mathrm{d}A^i{}_j}{\mathrm{d}t}\,, \qquad C^i{}_j = \frac{\partial \det A}{\partial A^i{}_j}\,, \tag{7.64}$$

$C^i{}_j$ being the cofactor *of the entry $A^i{}_j$. Wherever $A(t)$ is invertible this reads*

$$\frac{\mathrm{d}}{\mathrm{d}t}\det A(t) = \det A(t)\ \mathrm{tr}\Big(A(t)^{-1}\frac{\mathrm{d}A}{\mathrm{d}t}(t)\Big)\,, \tag{7.65}$$

and where moreover $\det A > 0$, $\mathrm{d}\left(\ln\det A\right)/\mathrm{d}t = \mathrm{tr}\left(A^{-1}\mathrm{d}A/\mathrm{d}t\right)$.

Rests on Propositions 5.1, 7.30 and 7.31 and Equation (5.19).

Proof. *Derives Proposition 7.79.*

The cofactor form. By Equation (5.19) the determinant is a polynomial in the n^2 entries, and each of its terms carries exactly one entry from each column. Differentiating that sum of products by the product rule (Proposition 7.30),

$$\frac{\mathrm{d}}{\mathrm{d}t}\det A = \sum_{\sigma}\mathrm{sgn}(\sigma)\sum_{k=1}^{n}\frac{\mathrm{d}A^{\sigma(k)}{}_k}{\mathrm{d}t}\prod_{j\neq k}A^{\sigma(j)}{}_j\,.$$

Collecting the terms that multiply one fixed $\mathrm{d}A^i{}_j/\mathrm{d}t$ leaves precisely the coefficient of $A^i{}_j$ in Equation (5.19), that is $\partial\det A/\partial A^i{}_j = C^i{}_j$, which is Equation (7.64). Note that $C^i{}_j$ contains no entry of the j-th column at all.

The trace form. Since det is linear in the j-th column and $C^i{}_j$ is free of that column, expanding along it gives $\sum_i A^i{}_j C^i{}_j = \det A$ for each j; and for $k \neq j$ the sum $\sum_i A^i{}_k C^i{}_j$ is that same expansion for the matrix whose j-th column has been replaced by its k-th, a matrix with two equal columns, whose determinant therefore vanishes. These are the two properties of Equation (5.19) on which Equation (5.16) also rests, and together they say

$$\sum_{i=1}^{n} A^i{}_k\, C^i{}_j = \delta_{jk}\det A\,. \tag{7.66}$$

If $\det A \neq 0$, Equation (7.66) states that the matrix with entries $C^i{}_j/\det A$, read with its indices exchanged, is a left inverse of A and hence its inverse: $(A^{-1})^j{}_i = C^i{}_j/\det A$. Substituting in Equation (7.64),

$$\frac{\mathrm{d}}{\mathrm{d}t}\det A = \det A \sum_{i,j} \left(A^{-1}\right)^j{}_i \frac{\mathrm{d}A^i{}_j}{\mathrm{d}t} = \det A \ \mathrm{tr}\left(A^{-1}\frac{\mathrm{d}A}{\mathrm{d}t}\right) ,$$

because the j-th diagonal entry of $A^{-1}\mathrm{d}A/\mathrm{d}t$ is $\sum_i (A^{-1})^j{}_i \mathrm{d}A^i{}_j/\mathrm{d}t$. This is Equation (7.65), and the logarithmic form follows from it by the chain rule (Proposition 7.31). □

7.10.5 The implicit function theorem

Theorem 7.80 (Implicit function theorem)**.** *Let $A \subseteq \mathbb{R}^m \times \mathbb{R}^n$ be open, let $\boldsymbol{F} : A \longrightarrow \mathbb{R}^n$ be of class C^k with $k \geq 1$, write the points of A as $(\boldsymbol{x}, \boldsymbol{y})$ with $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$, and abbreviate $(D_{\boldsymbol{y}}\boldsymbol{F})_{ij} = \partial F_i/\partial y_j$ and $(D_{\boldsymbol{x}}\boldsymbol{F})_{ij} = \partial F_i/\partial x_j$. Let $(\boldsymbol{a}, \boldsymbol{b}) \in A$ satisfy*

$$\boldsymbol{F}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{0} , \qquad \det\left[D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a}, \boldsymbol{b})\right] \neq 0 . \tag{7.67}$$

Then there are open sets $W \ni \boldsymbol{a}$ in $\mathbb{R}^m$ and $V \ni \boldsymbol{b}$ in $\mathbb{R}^n$ with $W \times V \subseteq A$, and a unique map $\boldsymbol{h} : W \longrightarrow V$ whose graph is the zero set,

$$\{(\boldsymbol{x}, \boldsymbol{y}) \in W \times V \mid \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}\} = \{(\boldsymbol{x}, \boldsymbol{h}(\boldsymbol{x})) \mid \boldsymbol{x} \in W\} , \tag{7.68}$$

with $\boldsymbol{h}(\boldsymbol{a}) = \boldsymbol{b}$. That map is of class C^k, and

$$D\boldsymbol{h}(\boldsymbol{x}) = -\left[D_{\boldsymbol{y}}\boldsymbol{F}\right]^{-1} D_{\boldsymbol{x}}\boldsymbol{F} , \tag{7.69}$$

both blocks being evaluated at $(\boldsymbol{x}, \boldsymbol{h}(\boldsymbol{x}))$.

Rests on Theorem 7.68, Proposition 7.72, and Definition 7.67.

Full derivation: Appendix A.28 (p. 2914).

Derives Theorem 7.80.

Corollary 7.81 (Inverse function theorem)**.** *Let $U \subseteq \mathbb{R}^n$ be open, let $\boldsymbol{f} : U \longrightarrow \mathbb{R}^n$ be of class C^k with $k \geq 1$, and let $\boldsymbol{\alpha} \in U$ satisfy $\det\left[D\boldsymbol{f}(\boldsymbol{\alpha})\right] \neq 0$. Then $\boldsymbol{f}$ maps some open $U_0 \ni \boldsymbol{\alpha}$ bijectively onto an open $V_0 \ni \boldsymbol{f}(\boldsymbol{\alpha})$; the inverse $\boldsymbol{g} : V_0 \longrightarrow U_0$ is of class C^k; and*

$$D\boldsymbol{g}(\boldsymbol{v}) = \left[D\boldsymbol{f}\left(\boldsymbol{g}(\boldsymbol{v})\right)\right]^{-1} . \tag{7.70}$$

A C^k map whose derivative is everywhere invertible is therefore a local C^k diffeomorphism*: a bijection between open sets with C^k inverse.*

Rests on Theorem 7.80, Definition 7.66, and Proposition 7.72.

Full derivation: Appendix A.28 (p. 2914).

Derives Corollary 7.81.

Remark 7.82 (Where the proof is, and why). Both statements are proved in Appendix A.28, as Theorem A.286 and Corollary A.292, from exactly the material of Section 7.10.2 and nothing else: no fixed-point theorem is imported, the Newton iteration being run explicitly and its convergence read off the Cauchy criterion (Theorem 7.8). The proof runs to several pages and would unbalance this chapter, which is why it sits in the appendix. The *statements* belong here, because this is where several-variable differentiability is developed and because what uses them — the constant-rank theorem and the multiplier rule below, the regular value theorem of Chapter 13, the inversion of a Legendre transformation in Chapter 16 — reaches for them as calculus, not as geometry.

Rests on Theorem 7.80 and Corollary 7.81.

Theorem 7.83 (Constant rank)**.** *Let $U \subseteq \mathbb{R}^N$ be open, let $\boldsymbol{F} : U \longrightarrow \mathbb{R}^{N'}$ be of class C^k with $k \geq 1$, and suppose the derivative $D\boldsymbol{F}$ has one and the same rank r at every point of U. Then about each $\boldsymbol{p} \in U$ and about its image there are C^k diffeomorphisms (Corollary 7.81) φ, of a neighbourhood of $\boldsymbol{p}$ onto a neighbourhood of the origin of $\mathbb{R}^N$, and χ, of a neighbourhood of $\boldsymbol{F}(\boldsymbol{p})$ onto a neighbourhood of the origin of $\mathbb{R}^{N'}$, in which $\boldsymbol{F}$ reads*

$$\chi \circ \boldsymbol{F} \circ \varphi^{-1}\left(v^1, \dots, v^N\right) = \left(v^1, \dots, v^r, 0, \dots, 0\right) . \tag{7.71}$$

In particular the image of a small enough neighbourhood of $\boldsymbol{p}$ is, in the coordinates χ, the intersection of an r-dimensional coordinate plane with an open set, cut out by the vanishing of $N' - r$ coordinate functions with independent differentials.

Rests on Corollaries 7.36 and 7.81 and Proposition 7.72.

Proof. *Derives Theorem 7.83.*

Replacing $\boldsymbol{F}$ by $\boldsymbol{F} - \boldsymbol{F}(\boldsymbol{p})$ — a translation of the target, which changes neither $D\boldsymbol{F}$ nor its rank — we may assume $\boldsymbol{F}(\boldsymbol{p}) = \boldsymbol{0}$.

Straightening the source. Since $D\boldsymbol{F}(\boldsymbol{p})$ has rank r, some $r \times r$ minor of it is non-zero; relabelling the coordinates of $\mathbb{R}^N$ and of $\mathbb{R}^{N'}$, which are linear changes and hence C^∞, we may take it to be the minor built from the first r rows and the first r columns. Define

$$\varphi(\boldsymbol{u}) = \left(F^1(\boldsymbol{u}), \dots, F^r(\boldsymbol{u}), u^{r+1} - p^{r+1}, \dots, u^N - p^N\right) . \tag{7.72}$$

Its derivative at $\boldsymbol{p}$ is block triangular, carrying that invertible $r \times r$ minor in the upper left and the identity in the lower right, so $\det D\varphi(\boldsymbol{p}) \neq 0$ and Corollary 7.81 makes φ a C^k diffeomorphism of a neighbourhood of $\boldsymbol{p}$ onto a neighbourhood of $\varphi(\boldsymbol{p}) = \boldsymbol{0}$. Shrink the latter to an open box $B = B' \times B''$ with $B' \subseteq \mathbb{R}^r$ and $B'' \subseteq \mathbb{R}^{N-r}$, and shrink the former to its preimage.

The tail depends only on the head. Put $\boldsymbol{G} = \boldsymbol{F} \circ \varphi^{-1}$ on B, of class C^k. By construction the first r components of $\boldsymbol{G}$ are the first r coordinates, so

$$\boldsymbol{G}\left(\boldsymbol{v}', \boldsymbol{v}''\right) = \left(\boldsymbol{v}', \, \boldsymbol{g}\left(\boldsymbol{v}', \boldsymbol{v}''\right)\right) , \qquad \boldsymbol{g} : B \longrightarrow \mathbb{R}^{N'-r} , \tag{7.73}$$

and the derivative has the block form

$$D\boldsymbol{G} = \begin{pmatrix} \mathbb{1}_r & 0 \\ D_{\boldsymbol{v}'}\boldsymbol{g} & D_{\boldsymbol{v}''}\boldsymbol{g} \end{pmatrix} . \tag{7.74}$$

The chain rule Equation (7.52) gives $D\boldsymbol{G} = D\boldsymbol{F} \cdot D\varphi^{-1}$ with $D\varphi^{-1}$ invertible, so $D\boldsymbol{G}$ has rank r at every point of B. Its first r columns are already independent by Equation (7.74), so every column of $D_{\boldsymbol{v}''}\boldsymbol{g}$ must be a combination of them; but those columns vanish in the first r rows, so each is a combination with zero coefficients, that is $D_{\boldsymbol{v}''}\boldsymbol{g} \equiv 0$ on B. Because B'' is a box, any two of its points are joined inside B by a chain of coordinate steps, along each of which $\boldsymbol{g}$ is constant (Corollary 7.36); hence $\boldsymbol{g}$ does not depend on $\boldsymbol{v}''$, and we write $\boldsymbol{g} = \boldsymbol{g}(\boldsymbol{v}')$.

Straightening the target. On $B' \times \mathbb{R}^{N'-r}$ define

$$\chi\left(\boldsymbol{w}', \boldsymbol{w}''\right) = \left(\boldsymbol{w}', \;\; \boldsymbol{w}'' - \boldsymbol{g}(\boldsymbol{w}')\right) . \tag{7.75}$$

It is C^k, and it is a bijection of $B' \times \mathbb{R}^{N'-r}$ onto itself with the C^k inverse $(\boldsymbol{w}', \boldsymbol{w}'') \longmapsto (\boldsymbol{w}', \boldsymbol{w}'' + \boldsymbol{g}(\boldsymbol{w}'))$, hence a C^k diffeomorphism. Composing with Equation (7.73),

$$\chi\left(\boldsymbol{G}\left(\boldsymbol{v}', \boldsymbol{v}''\right)\right) = \left(\boldsymbol{v}', \, \boldsymbol{g}(\boldsymbol{v}') - \boldsymbol{g}(\boldsymbol{v}')\right) = \left(\boldsymbol{v}', \boldsymbol{0}\right) ,$$

which is Equation (7.71). The final clause is that statement read backwards: in the coordinates χ the image is where the last $N' - r$ coordinates vanish. □

Lemma 7.84 (Functions vanishing on a regular zero set)**.** *Let $A \subseteq \mathbb{R}^n$ be open, let $\phi_1, \dots, \phi_J : A \longrightarrow \mathbb{R}$ be smooth with linearly independent gradients at every point of their common zero set*

$$\Sigma = \{\boldsymbol{x} \in A \mid \phi_1(\boldsymbol{x}) = \cdots = \phi_J(\boldsymbol{x}) = 0\} \,, \tag{7.76}$$

and let f be smooth on A and vanish identically on Σ. Then every point of Σ has a neighbourhood on which

$$f = \sum_{j=1}^{J} c^j \, \phi_j \tag{7.77}$$

for smooth functions $c^1, \dots, c^J$.

Rests on Corollary 7.81 and Theorems 7.43 and 7.77.

Proof. *Derives Lemma 7.84.*

Adapted coordinates. Fix $\boldsymbol{p} \in \Sigma$. The $J \times n$ matrix of the gradients $\nabla\phi_j(\boldsymbol{p})$ has rank J, so J of its columns are independent; relabelling the coordinates, take them to be the first J. Let $i_1 < \cdots < i_{n-J}$ be the remaining indices and put

$$\boldsymbol{\Psi}(\boldsymbol{x}) = \Big(\phi_1(\boldsymbol{x}), \dots, \phi_J(\boldsymbol{x}),\, x^{i_1} - p^{i_1}, \dots, x^{i_{n-J}} - p^{i_{n-J}}\Big) \,. \tag{7.78}$$

Expanding $\det D\boldsymbol{\Psi}(\boldsymbol{p})$ along its last $n-J$ rows, each of which is a vector of the canonical basis, leaves — up to sign — exactly the non-zero $J \times J$ minor just chosen, so $\boldsymbol{\Psi}$ is a smooth diffeomorphism of a neighbourhood of $\boldsymbol{p}$ onto a neighbourhood of $\boldsymbol{\Psi}(\boldsymbol{p}) = \boldsymbol{0}$ (Corollary 7.81), which we shrink to an open box $B = B' \times B''$, $B' \subseteq \mathbb{R}^J$, $B'' \subseteq \mathbb{R}^{n-J}$. Write $\boldsymbol{y} = (\boldsymbol{y}', \boldsymbol{y}'') = \boldsymbol{\Psi}(\boldsymbol{x})$ and $\tilde{f} = f \circ \boldsymbol{\Psi}^{-1}$, smooth on B. By construction $\boldsymbol{y}' = (\phi_1, \dots, \phi_J)$, so Σ meets the chart in $\{\boldsymbol{y}' = \boldsymbol{0}\}$ and $\tilde{f}$ vanishes there.

Integrating along the segment. For $\boldsymbol{y} \in B$ the segment from $(\boldsymbol{0}, \boldsymbol{y}'')$ to $(\boldsymbol{y}', \boldsymbol{y}'')$ lies in the box B, so the fundamental theorem of calculus (Theorem 7.43) and the chain rule Equation (7.52) give

$$\tilde{f}\,(\boldsymbol{y}', \boldsymbol{y}'') = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s} \tilde{f}\,(s\boldsymbol{y}', \boldsymbol{y}'')\,\mathrm{d}s = \sum_{j=1}^{J} y'^{\,j}\, \tilde{c}_j(\boldsymbol{y}) \,, \quad \tilde{c}_j(\boldsymbol{y}) = \int_0^1 \frac{\partial \tilde{f}}{\partial y'^{\,j}}\,(s\boldsymbol{y}', \boldsymbol{y}'')\,\mathrm{d}s \,, \tag{7.79}$$

where the integral computes $\tilde{f}(\boldsymbol{y}', \boldsymbol{y}'') - \tilde{f}(\boldsymbol{0}, \boldsymbol{y}'')$ and the subtracted term is zero because $\tilde{f}$ vanishes on $\{\boldsymbol{y}' = \boldsymbol{0}\}$. Each $\tilde{c}_j$ is smooth: its integrand and every partial derivative of the integrand are continuous on $B \times [0,1]$, so Theorem 7.77 may be applied to it as often as one pleases. Setting $c^j = \tilde{c}_j \circ \boldsymbol{\Psi}$ and recalling $y'^{\,j} = \phi_j(\boldsymbol{x})$ turns Equation (7.79) into Equation (7.77). □

Proposition 7.85 (Lagrange multipliers in finitely many variables)**.** *Let $A \subseteq \mathbb{R}^n$ be open, let f and $g_1, \dots, g_J$ be of class C^1 on A with $J < n$, and let $\boldsymbol{x}_\star \in A$ satisfy $g_\alpha(\boldsymbol{x}_\star) = 0$ for every α and have linearly independent gradients $\nabla g_\alpha(\boldsymbol{x}_\star)$. If the restriction of f to the constraint set $\{\boldsymbol{x} \in A \mid g_1(\boldsymbol{x}) = \cdots = g_J(\boldsymbol{x}) = 0\}$ has a local extremum at $\boldsymbol{x}_\star$, then there are unique numbers $\lambda^1, \dots, \lambda^J$ with*

$$\nabla f(\boldsymbol{x}_\star) = \sum_{\alpha=1}^{J} \lambda^\alpha \, \nabla g_\alpha(\boldsymbol{x}_\star) \,. \tag{7.80}$$

Rests on Theorem 7.80, Lemma 7.33, and Proposition 7.72.

Proof. *Derives Proposition 7.85.*

Write $\boldsymbol{G} = (g_1, \dots, g_J)$. Its derivative at $\boldsymbol{x}_\star$ has rank J, so J of its columns are independent; relabelling the coordinates, split $\boldsymbol{x} = (\boldsymbol{x}', \boldsymbol{y})$ with $\boldsymbol{y} \in \mathbb{R}^J$ carrying those columns, so that $\det\,[D_{\boldsymbol{y}}\boldsymbol{G}(\boldsymbol{x}_\star)] \neq 0$. By Theorem 7.80 there are neighbourhoods $W \ni \boldsymbol{x}'_\star$

and $V \ni \boldsymbol{y}_\star$ and a C^1 map $\boldsymbol{h} : W \longrightarrow V$ whose graph is the constraint set inside $W \times V$, with the derivative Equation (7.69).

The function $\Xi(\boldsymbol{x}') = f\,(\boldsymbol{x}', \boldsymbol{h}(\boldsymbol{x}'))$ therefore has an ordinary, unconstrained local extremum at $\boldsymbol{x}'_\star$; applying Lemma 7.33 to each one-variable function $t \longmapsto \Xi(\boldsymbol{x}'_\star + t\boldsymbol{e}_i)$ gives $\nabla\Xi(\boldsymbol{x}'_\star) = \boldsymbol{0}$, and the chain rule Equation (7.52) turns that into

$$D_{\boldsymbol{x}'} f + D_{\boldsymbol{y}} f \cdot D\boldsymbol{h} = 0\,, \qquad D\boldsymbol{h} = -\,[D_{\boldsymbol{y}}\boldsymbol{G}]^{-1}\, D_{\boldsymbol{x}'}\boldsymbol{G}\,, \tag{7.81}$$

everything evaluated at $\boldsymbol{x}_\star$, the second equality being Equation (7.69). Now *define* the row vector $\lambda = D_{\boldsymbol{y}} f\,[D_{\boldsymbol{y}}\boldsymbol{G}]^{-1}$, which exists and is unique because $D_{\boldsymbol{y}}\boldsymbol{G}$ is invertible; unwinding, λ is the unique solution of $\partial f/\partial y_j = \sum_\alpha \lambda^\alpha \partial g_\alpha/\partial y_j$, which is Equation (7.80) in the $\boldsymbol{y}$ block. Substituting it into Equation (7.81),

$$D_{\boldsymbol{x}'} f = -D_{\boldsymbol{y}} f \cdot D\boldsymbol{h} = \lambda\, D_{\boldsymbol{x}'}\boldsymbol{G}\,,$$

which is Equation (7.80) in the $\boldsymbol{x}'$ block. Uniqueness over all n components follows from the linear independence of the $\nabla g_\alpha(\boldsymbol{x}_\star)$: two solutions would differ by a vanishing combination of them. □

Remark 7.86 (The multiplier rule and its functional counterpart). Proposition 7.85 is the finite-dimensional statement from which the functional multiplier rule of Chapter 16 takes its name and its shape: the gradient of the objective lies in the span of the gradients of the constraints, so that the constrained problem is the unconstrained problem for $f - \lambda^\alpha g_\alpha$. What the variational version must supply for itself is the infinite-dimensional replacement of the implicit function theorem used here; the algebra of the multipliers is the same.

Rests on Proposition 7.85.

Definition 7.87 (Envelope of a family)**.** Let $\Omega \subseteq \mathbb{R}^n$ be open, let $I \subseteq \mathbb{R}$ be an open interval and let $f : \Omega \times I \longrightarrow \mathbb{R}$ be of class C^2. The member of the family labelled by $a \in I$ is the level set $S_a = \{\boldsymbol{x} \in \Omega \mid f(\boldsymbol{x}, a) = 0\}$, and the *envelope* of the family is

$$E = \left\{ \boldsymbol{x} \in \Omega \mid \exists\, a \in I :\ f(\boldsymbol{x}, a) = 0 \ \text{ and } \ \frac{\partial f}{\partial a}(\boldsymbol{x}, a) = 0 \right\}. \tag{7.82}$$

Rests on Definitions 7.65 and 7.66.

Proposition 7.88 (The envelope touches every member it meets)**.** *Let $\boldsymbol{x}_0 \in E$ with witness a_0, and suppose that $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, a_0) \neq \boldsymbol{0}$ and that $\partial^2 f/\partial a^2(\boldsymbol{x}_0, a_0) \neq 0$. Then on a neighbourhood of $\boldsymbol{x}_0$ the witness is a C^1 function $a(\boldsymbol{x})$, the set E is contained in the zero set of $\Psi(\boldsymbol{x}) = f\,(\boldsymbol{x}, a(\boldsymbol{x}))$, and*

$$\nabla\Psi(\boldsymbol{x}_0) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, a_0) \neq \boldsymbol{0}\,. \tag{7.83}$$

Hence E and S_{a_0} have the same tangent plane at $\boldsymbol{x}_0$: the envelope is tangent to the member through each of its points.

Rests on Definition 7.87, Theorem 7.80, and Proposition 7.72.

Proof. *Derives Proposition 7.88.*

Apply Theorem 7.80 to the scalar equation $\partial f/\partial a(\boldsymbol{x}, a) = 0$ at $(\boldsymbol{x}_0, a_0)$, where the derivative with respect to a is $\partial^2 f/\partial a^2 \neq 0$: it returns a neighbourhood of $\boldsymbol{x}_0$ and a C^1 function $a(\boldsymbol{x})$ with $a(\boldsymbol{x}_0) = a_0$ and

$$\frac{\partial f}{\partial a}\,(\boldsymbol{x}, a(\boldsymbol{x})) = 0 \tag{7.84}$$

identically there, and no other root of $\partial f/\partial a$ nearby — so every point of E in that neighbourhood has $a(\boldsymbol{x})$ as its witness and satisfies $\Psi(\boldsymbol{x}) = 0$. By the chain rule Equation (7.52),

$$\nabla\Psi(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} f\,(\boldsymbol{x}, a(\boldsymbol{x})) + \frac{\partial f}{\partial a}\,(\boldsymbol{x}, a(\boldsymbol{x}))\,\nabla a(\boldsymbol{x})\,,$$

whose second term vanishes identically by Equation (7.84); at $\boldsymbol{x}_0$ this is Equation (7.83). That same vector is the gradient at $\boldsymbol{x}_0$ of $\boldsymbol{x} \longmapsto f(\boldsymbol{x}, a_0)$, the defining function of S_{a_0}. Two level sets through one point whose defining functions have the same non-zero gradient there have the same tangent plane (Definition 7.64), which is the assertion. □

7.11 Vector analysis

7.11.1 The nabla operator

Let $\psi : \mathbb{R}^3 \longrightarrow \mathbb{R}$ and $\boldsymbol{A} : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$.

- In Cartesian coordinates,

$$\nabla\psi = \hat{\boldsymbol{x}}\,\frac{\partial\psi}{\partial x} + \hat{\boldsymbol{y}}\,\frac{\partial\psi}{\partial y} + \hat{\boldsymbol{z}}\,\frac{\partial\psi}{\partial z}\,, \tag{7.85}$$

$$\nabla\cdot\boldsymbol{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}\,, \tag{7.86}$$

$$\nabla\times\boldsymbol{A} = \hat{\boldsymbol{x}}\left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right) + \hat{\boldsymbol{y}}\left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right) + \hat{\boldsymbol{z}}\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right), \tag{7.87}$$

$$\nabla^2\psi = \nabla\cdot\nabla\psi = \frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2}\,. \tag{7.88}$$

- In cylindrical coordinates (ρ, φ, z),

$$\nabla\psi = \hat{\boldsymbol{\rho}}\,\frac{\partial\psi}{\partial\rho} + \hat{\boldsymbol{\varphi}}\,\frac{1}{\rho}\frac{\partial\psi}{\partial\varphi} + \hat{\boldsymbol{z}}\,\frac{\partial\psi}{\partial z}\,, \tag{7.89}$$

$$\nabla\cdot\boldsymbol{A} = \frac{1}{\rho}\frac{\partial(\rho A_\rho)}{\partial\rho} + \frac{1}{\rho}\frac{\partial A_\varphi}{\partial\varphi} + \frac{\partial A_z}{\partial z}\,, \tag{7.90}$$

$$\nabla^2\psi = \frac{1}{\rho}\,\partial_\rho(\rho\,\partial_\rho\psi) + \frac{1}{\rho^2}\,\partial_\varphi^2\psi + \partial_z^2\psi\,. \tag{7.91}$$

- In spherical coordinates (r, θ, φ) (Figure 7.8),

$$\nabla\psi = \hat{\boldsymbol{r}}\,\frac{\partial\psi}{\partial r} + \hat{\boldsymbol{\theta}}\,\frac{1}{r}\frac{\partial\psi}{\partial\theta} + \hat{\boldsymbol{\varphi}}\,\frac{1}{r\sin\theta}\frac{\partial\psi}{\partial\varphi}\,, \tag{7.92}$$

$$\nabla\cdot\boldsymbol{A} = \frac{1}{r^2}\frac{\partial(r^2 A_r)}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial(\sin\theta\, A_\theta)}{\partial\theta} + \frac{1}{r\sin\theta}\frac{\partial A_\varphi}{\partial\varphi}\,, \tag{7.93}$$

$$\nabla^2\psi = \frac{1}{r^2}\,\partial_r\Big(r^2\,\partial_r\psi\Big) + \frac{1}{r^2\sin\theta}\,\partial_\theta(\sin\theta\,\partial_\theta\psi) + \frac{1}{r^2\sin^2\theta}\,\partial_\varphi^2\psi\,. \tag{7.94}$$

  The curl expressions in curvilinear coordinates follow from the general covariant machinery of Chapter 13.

7.11.2 Identities of the nabla calculus

Every identity below is an equality of Cartesian components, and each is proved by writing both sides in index notation. Indices i, j, k, l, m run over $1, 2, 3$; a repeated index is summed; $\partial_i = \partial/\partial x^i$; δ_{ij} is the Kronecker delta, equal to 1 when $i = j$ and 0 otherwise; and ϵ_{ijk} is the *Levi-Civita symbol*, the totally antisymmetric array with $\epsilon_{123} = +1$. In this notation Equation (7.87) and the cross product of two vectors of $\mathbb{R}^3$ read

$$(\nabla\times\boldsymbol{A})_i = \epsilon_{ijk}\,\partial_j A_k\,, \qquad (\boldsymbol{A}\times\boldsymbol{B})_i = \epsilon_{ijk} A_j B_k\,. \tag{7.95}$$

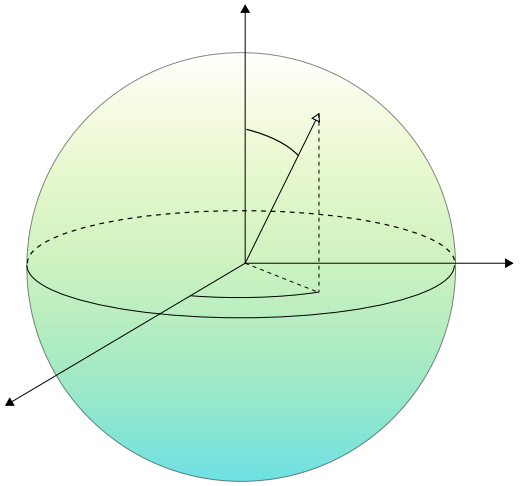


Figure 7.8: Spherical coordinates of a point in $\mathbb{R}^3$.

Lemma 7.89 (Contraction of two Levi-Civita symbols)**.** *For all $j,k,l,m \in \{1,2,3\}$,*

$$\epsilon_{ijk}\,\epsilon_{ilm} = \delta_{jl}\delta_{km} - \delta_{jm}\delta_{kl}\,, \tag{7.96}$$

the repeated index i being summed over $1,2,3$.
Rests on Equation (7.95).

Proof. *Derives Lemma 7.89.*

Both sides change sign under $j \leftrightarrow k$ and under $l \leftrightarrow m$, and both vanish when $j = k$ or $l = m$; so only the cases with $j \neq k$ and $l \neq m$ need checking. On the left the sum over i has at most one surviving term, the one whose i differs from both j and k, and it survives only if i also differs from l and m — that is, only if $\{l,m\}$ is the same pair as $\{j,k\}$. If $(l,m) = (j,k)$ that term is $(\epsilon_{ijk})^2 = 1$ and the right-hand side is $1\cdot 1 - 0 = 1$; if $(l,m) = (k,j)$ the term is $\epsilon_{ijk}\epsilon_{ikj} = -1$ and the right-hand side is $0 - 1\cdot 1 = -1$; and if $\{l,m\} \neq \{j,k\}$ the left-hand side vanishes, while on the right $\delta_{jl}\delta_{km} = 1$ would force $(l,m) = (j,k)$ and $\delta_{jm}\delta_{kl} = 1$ would force $(l,m) = (k,j)$, so both products vanish too. This exhausts the cases. □

Proposition 7.90 (Product identities of the nabla calculus)**.** *Let ψ be a C^1 scalar field and let $\boldsymbol{A}, \boldsymbol{B}$ be C^1 vector fields on an open subset of $\mathbb{R}^3$. Then*

$$\nabla\cdot(\psi\boldsymbol{A}) = \psi\,\nabla\cdot\boldsymbol{A} + \boldsymbol{A}\cdot\nabla\psi\,, \tag{7.97}$$

$$\nabla\times(\psi\boldsymbol{A}) = \psi\,\nabla\times\boldsymbol{A} + \nabla\psi\times\boldsymbol{A}\,, \tag{7.98}$$

$$\nabla\cdot(\boldsymbol{A}\times\boldsymbol{B}) = \boldsymbol{B}\cdot(\nabla\times\boldsymbol{A}) - \boldsymbol{A}\cdot(\nabla\times\boldsymbol{B})\,, \tag{7.99}$$

$$\nabla\times(\boldsymbol{A}\times\boldsymbol{B}) = (\boldsymbol{B}\cdot\nabla)\,\boldsymbol{A} - (\boldsymbol{A}\cdot\nabla)\,\boldsymbol{B} + \boldsymbol{A}\,(\nabla\cdot\boldsymbol{B}) - \boldsymbol{B}\,(\nabla\cdot\boldsymbol{A})\,, \tag{7.100}$$

$$\nabla\,(\boldsymbol{A}\cdot\boldsymbol{B}) = (\boldsymbol{A}\cdot\nabla)\,\boldsymbol{B} + (\boldsymbol{B}\cdot\nabla)\,\boldsymbol{A} + \boldsymbol{A}\times(\nabla\times\boldsymbol{B}) + \boldsymbol{B}\times(\nabla\times\boldsymbol{A})\,, \tag{7.101}$$

where $(\boldsymbol{A}\cdot\nabla)\,\boldsymbol{B}$ denotes the vector with components $A_j\,\partial_j B_i$. Setting $\boldsymbol{B} = \boldsymbol{A}$ in Equation (7.101) gives the identity

$$(\boldsymbol{A}\cdot\nabla)\,\boldsymbol{A} = \nabla\left(\frac{|\boldsymbol{A}|^2}{2}\right) - \boldsymbol{A}\times(\nabla\times\boldsymbol{A})\,, \tag{7.102}$$

which splits an advective derivative into a gradient and a term orthogonal to $\boldsymbol{A}$.
Rests on Lemma 7.89, Equation (7.95), and Proposition 7.30.

Proof. *Derives Proposition 7.90.*

Throughout, the product rule Proposition 7.30 is applied to each Cartesian component and Equation (7.96) is used to contract the two symbols that appear whenever two cross products meet.

Equations (7.97) *and* (7.98). Directly, $\partial_i(\psi A_i) = \psi\,\partial_i A_i + A_i\,\partial_i\psi$, which is the first; and $\epsilon_{ijk}\partial_j(\psi A_k) = \psi\,\epsilon_{ijk}\partial_j A_k + \epsilon_{ijk}(\partial_j\psi)A_k$, whose second term is $(\nabla\psi\times\boldsymbol{A})_i$ by Equation (7.95).

Equation (7.99). $\partial_i\left(\epsilon_{ijk}A_jB_k\right) = \epsilon_{ijk}\left(\partial_i A_j\right)B_k + \epsilon_{ijk}A_j\left(\partial_i B_k\right)$. In the first term relabel $\epsilon_{ijk} = \epsilon_{kij}$, so that it is $B_k\left(\nabla\times\boldsymbol{A}\right)_k$; in the second, $\epsilon_{ijk} = -\epsilon_{jik}$, so that it is $-A_j\left(\nabla\times\boldsymbol{B}\right)_j$.

Equation (7.100). Using Equation (7.95) twice and then Equation (7.96) with the summed index moved to the front of both symbols,

$$\begin{aligned}\left[\nabla\times(\boldsymbol{A}\times\boldsymbol{B})\right]_i &= \epsilon_{ijk}\,\partial_j\left(\epsilon_{klm}A_lB_m\right) = \epsilon_{kij}\epsilon_{klm}\,\partial_j\left(A_lB_m\right)\\ &= \left(\delta_{il}\delta_{jm}-\delta_{im}\delta_{jl}\right)\partial_j\left(A_lB_m\right) = \partial_m\left(A_iB_m\right) - \partial_l\left(A_lB_i\right)\,,\end{aligned}$$

and expanding the two derivatives by the product rule gives $A_i\,\partial_m B_m + B_m\,\partial_m A_i - B_i\,\partial_l A_l - A_l\,\partial_l B_i$, which is Equation (7.100) component by component.

Equation (7.101). By the same two steps,

$$\left[\boldsymbol{A}\times(\nabla\times\boldsymbol{B})\right]_i = \epsilon_{ijk}A_j\,\epsilon_{klm}\,\partial_l B_m = A_j\,\partial_i B_j - A_j\,\partial_j B_i\,,$$

and interchanging $\boldsymbol{A}$ and $\boldsymbol{B}$, $\left[\boldsymbol{B}\times(\nabla\times\boldsymbol{A})\right]_i = B_j\,\partial_i A_j - B_j\,\partial_j A_i$. Adding the two and using $\partial_i\left(A_jB_j\right) = A_j\,\partial_i B_j + B_j\,\partial_i A_j$ rearranges to Equation (7.101). Equation (7.102) is the case $\boldsymbol{B} = \boldsymbol{A}$, in which the two advective terms coincide and $\partial_i(A_jA_j) = \partial_i\left|\boldsymbol{A}\right|^2$, so that the whole identity may be divided by two. □

Proposition 7.91 (Second-order identities of the nabla calculus)**.** *Let ψ be a C^2 scalar field and let $\boldsymbol{A}$ be a C^2 vector field on an open subset of $\mathbb{R}^3$. Then*

$$\nabla\times\nabla\psi = \boldsymbol{0}\,, \tag{7.103}$$

$$\nabla\cdot(\nabla\times\boldsymbol{A}) = 0\,, \tag{7.104}$$

$$\nabla\times(\nabla\times\boldsymbol{A}) = \nabla\left(\nabla\cdot\boldsymbol{A}\right) - \nabla^2\boldsymbol{A}\,, \tag{7.105}$$

where $\nabla^2\boldsymbol{A}$ is the vector whose Cartesian components are the Laplacians Equation (7.88) of the Cartesian components of $\boldsymbol{A}$. The first of the three is Equation (7.128) of Proposition 7.104 below, recorded here so that the three second-order identities stand together.

Rests on Lemma 7.89, Proposition 7.73, and Equation (7.88).

Proof. *Derives Proposition 7.91.*

The first two are the contraction of an object antisymmetric in two indices with one symmetric in them. By Proposition 7.73 the second partial derivatives of a C^2 field commute, so $\partial_j\partial_k\psi$ is symmetric in (j,k) while ϵ_{ijk} is antisymmetric, and $(\nabla\times\nabla\psi)_i = \epsilon_{ijk}\partial_j\partial_k\psi$ vanishes: interchanging the summation labels j and k turns the sum into minus itself. The same argument on $\nabla\cdot(\nabla\times\boldsymbol{A}) = \epsilon_{ijk}\,\partial_i\partial_j A_k$ gives Equation (7.104). For the third, by Equation (7.95) and then Equation (7.96),

$$\begin{aligned}\left[\nabla\times(\nabla\times\boldsymbol{A})\right]_i &= \epsilon_{ijk}\,\partial_j\left(\epsilon_{klm}\,\partial_l A_m\right) = \left(\delta_{il}\delta_{jm}-\delta_{im}\delta_{jl}\right)\partial_j\partial_l A_m\\ &= \partial_j\partial_i A_j - \partial_j\partial_j A_i = \partial_i\left(\nabla\cdot\boldsymbol{A}\right) - \nabla^2 A_i\,,\end{aligned}$$

the last step again by Proposition 7.73, and this is Equation (7.105). □

Remark 7.92 (The Laplacian of a vector field is a Cartesian notion). Equation (7.105) is the *definition* of ∇^2 on a vector field in any coordinate system other than the Cartesian one: the componentwise Laplacian of Equation (7.91) or Equation (7.94) applied to A_ρ, A_θ and the rest is *not* $\nabla^2\boldsymbol{A}$, because those basis vectors themselves vary from point to point. The combination $\nabla(\nabla\cdot\boldsymbol{A}) - \nabla\times(\nabla\times\boldsymbol{A})$ is built from operations that are defined without reference to a basis and is therefore the right object to carry over; the covariant reading of the same statement is in Chapter 13.

Rests on Proposition 7.91 and Equation (7.94).

7.11.3 The integral theorems

The three classical integral theorems trade a derivative integrated over a region for the field itself integrated over the boundary of that region. They are the multivariable descendants of the fundamental theorem of calculus (Theorem 7.43), and each derivation below reduces to it. We first fix the integrals that appear in them.

Definition 7.93 (Multiple integral)**.** Let $R = [a_1, b_1] \times \cdots \times [a_N, b_N] \subset \mathbb{R}^N$ be a closed box and $f : R \longrightarrow \mathbb{R}$ bounded. Partition each edge; the lower and upper sums of Equation (7.22) are then formed with the infimum and the supremum of f on each sub-box weighted by the volume of that sub-box, and f is *integrable* on R, with integral $\int_R f$, when the supremum of the lower sums and the infimum of the upper sums coincide. For a bounded $D \subseteq R$, $\int_D f$ means $\int_R \tilde{f}$, where $\tilde{f}$ agrees with f on D and vanishes off it. We write $\iint_D$ and $\iiint_V$ for $N = 2$ and $N = 3$, and $\mathrm{d}V = \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$.

Rests on Definition 7.39 and Axiom 7.1.

Definition 7.94 (Line and surface integrals)**.** Let γ be a piecewise-C^1 path $\boldsymbol{r} : [\alpha, \beta] \longrightarrow \mathbb{R}^3$ and let $\boldsymbol{A}$ be continuous on its image. The *line integral* of $\boldsymbol{A}$ along γ is

$$\int_\gamma \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{l} = \int_\alpha^\beta \boldsymbol{A}(\boldsymbol{r}(t)) \cdot \boldsymbol{r}'(t)\,\mathrm{d}t\,, \tag{7.106}$$

and in the plane, with $\boldsymbol{r}(t) = (x(t), y(t))$,

$$\int_\gamma (P\,\mathrm{d}x + Q\,\mathrm{d}y) = \int_\alpha^\beta \left[P(\boldsymbol{r}(t))\,x'(t) + Q(\boldsymbol{r}(t))\,y'(t)\right]\mathrm{d}t\,. \tag{7.107}$$

Let S be a surface carried by a C^1 parametrization $\boldsymbol{r} : D \subseteq \mathbb{R}^2 \longrightarrow \mathbb{R}^3$, injective on the interior of D and with $\partial_u \boldsymbol{r} \times \partial_v \boldsymbol{r} \neq \boldsymbol{0}$ there. Its *oriented surface element*, and the *flux* of $\boldsymbol{A}$ through S, are

$$\mathrm{d}\boldsymbol{S} = (\partial_u \boldsymbol{r} \times \partial_v \boldsymbol{r})\,\mathrm{d}u\,\mathrm{d}v\,, \tag{7.108}$$

$$\iint_S \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{S} = \iint_D \boldsymbol{A}(\boldsymbol{r}(u,v)) \cdot (\partial_u \boldsymbol{r} \times \partial_v \boldsymbol{r})\,\mathrm{d}u\,\mathrm{d}v\,. \tag{7.109}$$

Reversing the orientation of γ or of S reverses the sign.

Rests on Definitions 7.66 and 7.93.

Definition 7.95 (Simple regions)**.** A compact set $D \subset \mathbb{R}^2$ is *vertically simple* if

$$D = \left\{(x, y) \in \mathbb{R}^2 \mid a \le x \le b,\ \varphi_1(x) \le y \le \varphi_2(x)\right\} \tag{7.110}$$

for piecewise-C^1 functions $\varphi_1 \le \varphi_2$ on $[a, b]$, and *horizontally simple* if it admits the same description with the two coordinates exchanged, between $\psi_1 \le \psi_2$ on $[c, d]$. It is *simple* if it is both. A compact $V \subset \mathbb{R}^3$ is *simple* if, for each of the three coordinate directions, it is the region between the graphs of two piecewise-C^1 functions over a simple plane region — for the z direction,

$$V = \left\{(x, y, z) \in \mathbb{R}^3 \mid (x, y) \in D_z,\ \chi_1(x, y) \le z \le \chi_2(x, y)\right\}. \tag{7.111}$$

Rests on Definitions 6.9 and 7.66.

Remark 7.96 (What the derivations below take as given)*.* Three properties of the multiple integral are used repeatedly. A function continuous on a simple region is integrable there; the integral is additive over a decomposition of the region into finitely many pieces meeting

only along their boundaries; and on a vertically simple region the integral of a continuous f reduces to the iterated integral

$$\iint_D f \,\mathrm{d}x\,\mathrm{d}y = \int_a^b \left[\int_{\varphi_1(x)}^{\varphi_2(x)} f(x,y)\,\mathrm{d}y \right] \mathrm{d}x\,, \tag{7.112}$$

with the analogous statements on a horizontally simple region and in $\mathbb{R}^3$. All three follow from the Darboux construction by the argument of Theorem 7.40 once one knows that the graph of a continuous φ carries no area: given $\varepsilon > 0$, uniform continuity (Theorem 7.25) splits $[a,b]$ into n subintervals on each of which φ varies by less than ε, and the n rectangles so obtained cover the graph with total area at most $(b-a)\varepsilon$. Hence the boundary of a simple region contributes nothing to any upper–lower gap. These statements are inputs to what follows, not consequences of it.

Rests on Definition 7.93 and Theorems 7.25 and 7.40.

The one further property of the multiple integral that the physical parts use constantly is the several-variable substitution rule. In one variable it is Corollary 7.44; in N variables the derivative of the substitution is a matrix, and what replaces it is the modulus of the determinant.

Theorem 7.97 (Change of variables in a multiple integral)**.** *Let $U, V \subseteq \mathbb{R}^N$ be open and bounded and let $\boldsymbol{\Phi} : U \longrightarrow V$ be a bijection of class C^1 with $\det D\boldsymbol{\Phi}(\boldsymbol{u}) \neq 0$ at every $\boldsymbol{u} \in U$ — equivalently, by Corollary 7.81, a C^1 bijection whose inverse is C^1. Let f be continuous and bounded on V. Then both integrals below exist in the sense of Definition 7.93 and are equal:*

$$\int_V f(\boldsymbol{x})\,\mathrm{d}^N x = \int_U f\left(\boldsymbol{\Phi}(\boldsymbol{u})\right) \left|\det D\boldsymbol{\Phi}(\boldsymbol{u})\right| \,\mathrm{d}^N u\,. \tag{7.113}$$

The same identity holds when $\boldsymbol{\Phi}$ fails to be injective, or its determinant fails to be nonzero, on a set contained in finitely many graphs of continuous functions, since such a set contributes nothing to a Darboux sum (Remark 7.96).

Rests on Definition 7.93, Corollary 7.81, and Remark 7.96.

Full derivation: Appendix A.44 (p. 3018).

Derives Theorem 7.97.

Example 7.98 (The two Jacobians this treatise uses)**.** In the plane, the polar map $\boldsymbol{\Phi}(\rho,\varphi) = (\rho\cos\varphi, \rho\sin\varphi)$ has

$$D\boldsymbol{\Phi} = \begin{pmatrix} \cos\varphi & -\rho\sin\varphi \\ \sin\varphi & \rho\cos\varphi \end{pmatrix}, \qquad \det D\boldsymbol{\Phi} = \rho\,, \tag{7.114}$$

so $\mathrm{d}x\,\mathrm{d}y = \rho\,\mathrm{d}\rho\,\mathrm{d}\varphi$; in space, the spherical map of Figure 7.8, $(r,\theta,\varphi) \longmapsto (r\sin\theta\cos\varphi, r\sin\theta\sin\varphi, r\cos\theta)$, has $\det D\boldsymbol{\Phi} = r^2\sin\theta$ on expanding the 3×3 determinant along its last row, so $\mathrm{d}V = r^2\sin\theta\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\varphi$. Each map is a diffeomorphism away from its coordinate degeneracies — $\rho = 0$ in the plane, $r = 0$ or $\sin\theta = 0$ in space — and those lie in finitely many graphs of continuous functions, so the last clause of Theorem 7.97 covers them. These are the volume elements used for the Gaussian integral of Chapter 11 and for every radial integration in the physical parts.

Rests on Theorem 7.97 and Equation (7.92).

Theorem 7.99 (Green)**.** *Let $D \subset \mathbb{R}^2$ be a region bounded by a piecewise-C^1 simple closed curve ∂D, traversed counterclockwise, and let P, Q be C^1 on $D \cup \partial D$. Then*

$$\oint_{\partial D} (P\,\mathrm{d}x + Q\,\mathrm{d}y) = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y\,. \tag{7.115}$$

Rests on Theorem 7.43, Definition 7.95, and Remark 7.96.

Proof. *Derives Theorem 7.99.*

We derive the identity for a simple region (Definition 7.95) and then for any region that can be cut into finitely many of them; what that leaves uncovered is stated in Remark 7.102.

Step 1: the P term on a vertically simple region. Let D be as in Equation (7.110). Since $\partial_y P$ is continuous, Equation (7.112) and the fundamental theorem of calculus (Theorem 7.43), applied in y at fixed x, give

$$\iint_D \frac{\partial P}{\partial y}\,\mathrm{d}x\,\mathrm{d}y = \int_a^b \left[P(x,\varphi_2(x)) - P(x,\varphi_1(x))\right]\mathrm{d}x\,. \tag{7.116}$$

Traversed counterclockwise, ∂D consists of the lower graph $x \longmapsto (x,\varphi_1(x))$ run from $x=a$ to $x=b$, the vertical segment at $x=b$, the upper graph $x \longmapsto (x,\varphi_2(x))$ run back from $x=b$ to $x=a$, and the vertical segment at $x=a$. On the two vertical pieces x is constant, so $x'(t)=0$ and they contribute nothing to Equation (7.107). Hence, using Equation (7.116) in the last step,

$$\oint_{\partial D} P\,\mathrm{d}x = \int_a^b P(x,\varphi_1(x))\,\mathrm{d}x + \int_b^a P(x,\varphi_2(x))\,\mathrm{d}x = -\iint_D \frac{\partial P}{\partial y}\,\mathrm{d}x\,\mathrm{d}y\,. \tag{7.117}$$

Step 2: the Q term on a horizontally simple region. With $D = \{(x,y) \mid c \le y \le d,\ \psi_1(y) \le x \le \psi_2(y)\}$ the same two moves give

$$\iint_D \frac{\partial Q}{\partial x}\,\mathrm{d}x\,\mathrm{d}y = \int_c^d \left[Q(\psi_2(y),y) - Q(\psi_1(y),y)\right]\mathrm{d}y\,.$$

The counterclockwise boundary now runs up the right graph $y \longmapsto (\psi_2(y),y)$ from $y=c$ to $y=d$, which is the sense that keeps the interior on the left, and down the left graph $y \longmapsto (\psi_1(y),y)$ from $y=d$ to $y=c$; the two horizontal pieces have $y'(t)=0$ and contribute nothing to $\oint Q\,\mathrm{d}y$. Therefore

$$\oint_{\partial D} Q\,\mathrm{d}y = \iint_D \frac{\partial Q}{\partial x}\,\mathrm{d}x\,\mathrm{d}y\,. \tag{7.118}$$

Step 3: simple regions. On a region that is both vertically and horizontally simple, both steps apply, and adding Equation (7.117) to Equation (7.118) is exactly Equation (7.115).

Step 4: decomposition. Let $D = \bigcup_{k=1}^{N} D_k$ with every D_k simple and any two of them meeting at most along parts of their boundaries. Summing Equation (7.115) over k, the area integrals add to $\iint_D (\partial_x Q - \partial_y P)$ by additivity (Remark 7.96). On the boundary side, an arc interior to D belongs to exactly two of the ∂D_k, whose regions lie on opposite sides of it, so it is traversed once in each direction; the two line integrals then cancel, since reversing the parametrization reverses the sign of Equation (7.107). What survives is the integral over the arcs making up ∂D, traversed counterclockwise. □

Theorem 7.100 (Stokes)**.** *For an oriented piecewise-smooth surface S with boundary curve ∂S (oriented compatibly) and $\boldsymbol{A} \in C^1$,*

$$\oint_{\partial S} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{l} = \iint_S (\nabla\times\boldsymbol{A})\cdot\mathrm{d}\boldsymbol{S}\,. \tag{7.119}$$

Rests on Theorem 7.99 and Propositions 7.72 and 7.73.

Proof. *Derives Theorem 7.100.*

It suffices to treat a surface that is a graph over a plane region to which Theorem 7.99 applies: a piecewise-smooth oriented surface can be cut into finitely many such pieces — after a permutation of the coordinate axes for each — and the line integrals along the interior cuts cancel in pairs exactly as in Step 4 of the derivation of Theorem 7.99.

So let $S = \{(x, y, g(x, y)) \in \mathbb{R}^3 \mid (x, y) \in D\}$ with g of class C^2 on a neighbourhood of D, oriented by the upward normal. The parametrization $\boldsymbol{r}(x, y) = (x, y, g(x, y))$ has

$$\partial_x \boldsymbol{r} \times \partial_y \boldsymbol{r} = \left(-\frac{\partial g}{\partial x},\ -\frac{\partial g}{\partial y},\ 1 \right), \tag{7.120}$$

whose third component is positive, so Equation (7.108) with this parametrization is the chosen orientation; and ∂S is the lift of ∂D, traversed so that ∂D runs counterclockwise.

The line integral. Parametrize ∂D by $t \longmapsto (x(t), y(t))$. Its lift has $z(t) = g(x(t), y(t))$, so the chain rule (Proposition 7.72) gives $z' = \partial_x g\, x' + \partial_y g\, y'$. Writing $\tilde{A}_i(x, y) = A_i(x, y, g(x, y))$ and collecting the coefficients of x' and y' in Equation (7.106),

$$\oint_{\partial S} \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{l} = \oint_{\partial D} (P\,\mathrm{d}x + Q\,\mathrm{d}y), \qquad P = \tilde{A}_x + \tilde{A}_z \frac{\partial g}{\partial x}, \quad Q = \tilde{A}_y + \tilde{A}_z \frac{\partial g}{\partial y}. \tag{7.121}$$

Both P and Q are C^1 on D, because $\boldsymbol{A}$ is C^1 and g is C^2.

Green's theorem. Apply Theorem 7.99 to Equation (7.121). The chain rule (Equation (7.52), applied to the composition of $(x, y) \longmapsto (x, y, g(x, y))$ with A_i) gives $\partial_x \tilde{A}_i = \partial_x A_i + \partial_z A_i\, \partial_x g$ and $\partial_y \tilde{A}_i = \partial_y A_i + \partial_z A_i\, \partial_y g$, every partial derivative of A_i being evaluated on S, so that

$$\begin{aligned}
\frac{\partial Q}{\partial x} &= \partial_x A_y + \partial_z A_y \frac{\partial g}{\partial x} + \left(\partial_x A_z + \partial_z A_z \frac{\partial g}{\partial x} \right) \frac{\partial g}{\partial y} + \tilde{A}_z \frac{\partial^2 g}{\partial x\, \partial y}, \\
\frac{\partial P}{\partial y} &= \partial_y A_x + \partial_z A_x \frac{\partial g}{\partial y} + \left(\partial_y A_z + \partial_z A_z \frac{\partial g}{\partial y} \right) \frac{\partial g}{\partial x} + \tilde{A}_z \frac{\partial^2 g}{\partial y\, \partial x}.
\end{aligned}$$

On subtraction the two second-derivative terms cancel by Clairaut–Schwarz (Proposition 7.73), which is the one place where g is required to be C^2, and so do the two terms carrying $\partial_z A_z\, \partial_x g\, \partial_y g$. What is left is

$$\begin{aligned}
\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} &= (\partial_x A_y - \partial_y A_x) + \frac{\partial g}{\partial x} (\partial_z A_y - \partial_y A_z) + \frac{\partial g}{\partial y} (\partial_x A_z - \partial_z A_x) \\
&= (\nabla \times \boldsymbol{A}) \cdot \left(-\frac{\partial g}{\partial x},\ -\frac{\partial g}{\partial y},\ 1 \right),
\end{aligned} \tag{7.122}$$

the second line being just Equation (7.87) read component by component. Integrating Equation (7.122) over D and comparing with Equation (7.109) and Equation (7.120),

$$\oint_{\partial S} \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{l} = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \mathrm{d}x\, \mathrm{d}y = \iint_S (\nabla \times \boldsymbol{A}) \cdot \mathrm{d}\boldsymbol{S},$$

which is Equation (7.119). □

Theorem 7.101 (Gauss)**.** *For a bounded region $V \subset \mathbb{R}^3$ with piecewise-smooth boundary ∂V, outward oriented, and $\boldsymbol{A} \in C^1$,*

$$\oint_{\partial V} \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{S} = \iiint_V \nabla \cdot \boldsymbol{A}\, \mathrm{d}V. \tag{7.123}$$

Rests on Theorem 7.43, Definition 7.95, and Remark 7.96.

Proof. *Derives Theorem 7.101.*

Again we derive the identity for a simple region (Definition 7.95) and extend it by decomposition.

The z component. Let V be as in Equation (7.111). Because $\partial_z A_z$ is continuous, the three-dimensional form of Equation (7.112) and the fundamental theorem of calculus (Theorem 7.43), applied in z at fixed (x, y), give

$$\iiint_V \frac{\partial A_z}{\partial z}\, \mathrm{d}V = \iint_{D_z} [A_z(x, y, \chi_2(x, y)) - A_z(x, y, \chi_1(x, y))]\, \mathrm{d}x\, \mathrm{d}y. \tag{7.124}$$

The boundary ∂V splits into the upper cap $z = \chi_2$, the lower cap $z = \chi_1$, and a lateral part on which the outward normal is horizontal — possibly empty, as for a ball, whose two caps meet along the equator. On the upper cap the parametrization $\boldsymbol{r}(x, y) = (x, y, \chi_2(x, y))$ has $\partial_x \boldsymbol{r} \times \partial_y \boldsymbol{r} = (-\partial_x \chi_2, -\partial_y \chi_2, 1)$, whose third component is positive; this is therefore the outward orientation, and Equation (7.109) gives $\iint A_z\, \hat{\boldsymbol{z}} \cdot \mathrm{d}\boldsymbol{S} = \iint_{D_z} A_z(x, y, \chi_2)\, \mathrm{d}x\, \mathrm{d}y$. On the lower cap the outward normal points downwards, so $\mathrm{d}\boldsymbol{S} = (\partial_x \chi_1, \partial_y \chi_1, -1)\, \mathrm{d}x\, \mathrm{d}y$ and the same computation returns $-\iint_{D_z} A_z(x, y, \chi_1)\, \mathrm{d}x\, \mathrm{d}y$. On the lateral part $\hat{\boldsymbol{z}} \cdot \mathrm{d}\boldsymbol{S} = 0$. Adding the three contributions and comparing with Equation (7.124),

$$\oint_{\partial V} A_z\, \hat{\boldsymbol{z}} \cdot \mathrm{d}\boldsymbol{S} = \iiint_V \frac{\partial A_z}{\partial z}\, \mathrm{d}V\,. \tag{7.125}$$

The other two components. Because V is simple in all three coordinate directions, the identical argument with the axes permuted gives the analogues of Equation (7.125) for $A_x \hat{\boldsymbol{x}}$ and $A_y \hat{\boldsymbol{y}}$. Adding the three identities, and using $\boldsymbol{A} = A_x \hat{\boldsymbol{x}} + A_y \hat{\boldsymbol{y}} + A_z \hat{\boldsymbol{z}}$ together with Equation (7.86), yields Equation (7.123).

Decomposition. If $V = \bigcup_k V_k$ with every V_k simple and any two meeting at most along parts of their boundaries, summing the identity over k adds the volume integrals. On the surface side an interior face belongs to exactly two of the ∂V_k, which assign it opposite outward normals, so those two fluxes cancel; the flux through ∂V is what remains. □

Remark 7.102 (Scope of these three derivations). Three limitations should be stated exactly. First, the derivations cover every region and every surface that decomposes into finitely many simple pieces (Definition 7.95) — which is every region and surface this treatise integrates over — but not an arbitrary region bounded by a piecewise-C^1 curve; the passage to that generality is an exhaustion argument, not carried out here. Second, the derivation of Theorem 7.100 asks the surface to be C^2, one degree smoother than the statement, because it applies Clairaut–Schwarz to the height function; the C^1 case follows by approximating the surface, again not done here. Third, the flux Equation (7.109) was defined through a parametrization, and its independence of that choice rests on the change-of-variables formula Theorem 7.97: two parametrizations of one surface differ by a C^1 diffeomorphism of their parameter domains, and Equation (7.113) converts one integral into the other, the Jacobian factor cancelling the corresponding factor in $\partial_u \boldsymbol{r} \times \partial_v \boldsymbol{r}$.

The three theorems are instances of a single statement, the general Stokes theorem $\int_M \mathrm{d}\omega = \int_{\partial M} \omega$ for a differential form ω on an oriented manifold with boundary: Theorem 7.99 is the case of a 1-form on a plane region, Theorem 7.100 that of a 1-form on a surface in $\mathbb{R}^3$, and Theorem 7.101 that of a 2-form on a solid. The language for that formulation is built in Chapter 13, where the antisymmetric-tensor form Equation (13.231) is stated — but its derivation is still owed there, which is why the classical theorems are proved directly above rather than deduced from it.

7.11.4 Conservative fields

Definition 7.103 (Conservative field)**.** For a function $\boldsymbol{F} : A \subseteq \mathbb{R}^3 \longrightarrow \mathbb{R}^3$, with A an open connected set, we say that $\boldsymbol{F}$ is *conservative* on A if and only if

$$\exists\, f : A \subseteq \mathbb{R}^3 \longrightarrow \mathbb{R} \mid \boldsymbol{F} = \nabla f\,. \tag{7.126}$$

Rests on Equation (7.85).

Proposition 7.104 (Properties of conservative fields)**.** *Let A be open and connected.*

1. *If $\boldsymbol{F} = \nabla f$ on A and $L \subseteq A$ is a closed piecewise-C^1 loop, then*

$$\oint_L \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{l} = 0\,. \tag{7.127}$$

2. *If $\boldsymbol{F} \in C^1(A)$ is conservative, then*

$$\nabla \times \boldsymbol{F} = \boldsymbol{0}\,. \tag{7.128}$$

3. *If $\boldsymbol{F} \in C^1(A)$, A is moreover simply connected (Chapter 6), and $\nabla \times \boldsymbol{F} = \boldsymbol{0}$, then $\boldsymbol{F}$ is conservative.*

Rests on Definition 7.103, Proposition 7.72, and Theorem 7.100.

Proof. *Derives Proposition 7.104.*

(1) Parametrize the loop by $\boldsymbol{r}(t)$, $t \in [0,1]$, with $\boldsymbol{r}(0) = \boldsymbol{r}(1)$. By the chain rule (Proposition 7.72), $\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{r}(t)) = \nabla f(\boldsymbol{r}(t)) \cdot \boldsymbol{r}'(t)$, so

$$\oint_L \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{l} = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{r}(t))\,\mathrm{d}t = f(\boldsymbol{r}(1)) - f(\boldsymbol{r}(0)) = 0$$

by the fundamental theorem of calculus (Theorem 7.43). (2) Componentwise, e.g. $(\nabla \times \nabla f)_z = \partial_x \partial_y f - \partial_y \partial_x f = 0$ by Clairaut–Schwarz (Proposition 7.73); the hypothesis $\boldsymbol{F} \in C^1$ makes f twice continuously differentiable. (3) Fix $\boldsymbol{a} \in A$ and define $f(\boldsymbol{p}) = \int_\gamma \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{l}$ along any path $\gamma \subset A$ from $\boldsymbol{a}$ to $\boldsymbol{p}$. Well-definedness is the point: two paths bound (by simple connectedness) a surface swept by a homotopy inside A, and Stokes' theorem (Theorem 7.100) with $\nabla \times \boldsymbol{F} = \boldsymbol{0}$ makes the two line integrals equal. Differentiating along a coordinate step then gives $\partial_i f = F_i$, i.e. $\nabla f = \boldsymbol{F}$. □

7.11.5 The Helmholtz decomposition

Proposition 7.104 says when a field is a gradient; the companion statement says that an arbitrary field is a gradient plus a curl. It is the capstone of this section, and it is what licenses the routine physical move of splitting a vector field into an irrotational and a solenoidal part and treating the two separately.

Theorem 7.105 (Helmholtz decomposition)**.** *Let $\boldsymbol{u} : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ be smooth with compact support. Then there are a smooth scalar field ϕ and a smooth vector field $\boldsymbol{\Psi}$ on $\mathbb{R}^3$ with*

$$\boldsymbol{u} = \nabla\phi + \nabla \times \boldsymbol{\Psi}\,, \qquad \nabla \cdot \boldsymbol{\Psi} = 0\,, \tag{7.129}$$

namely

$$\boldsymbol{w}(\boldsymbol{x}) = -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{\boldsymbol{u}(\boldsymbol{y})}{|\boldsymbol{x} - \boldsymbol{y}|}\,\mathrm{d}^3 y\,, \qquad \phi = \nabla \cdot \boldsymbol{w}\,, \qquad \boldsymbol{\Psi} = -\nabla \times \boldsymbol{w}\,. \tag{7.130}$$

The first summand is curl-free and the second divergence-free (Proposition 7.91), and the two are unique: if $\nabla\phi_1 + \nabla \times \boldsymbol{\Psi}_1 = \nabla\phi_2 + \nabla \times \boldsymbol{\Psi}_2$ with all four fields tending to zero at infinity, then $\nabla\phi_1 = \nabla\phi_2$ and $\nabla \times \boldsymbol{\Psi}_1 = \nabla \times \boldsymbol{\Psi}_2$.

Rests on Propositions 7.91 and 10.67 and Theorem 7.77.

Full derivation: Appendix A.45 (p. 3026).

Derives Theorem 7.105.

The mechanism is worth stating even though the estimates are not. Each Cartesian component of $\boldsymbol{w}$ is the Newtonian potential of the corresponding component of $\boldsymbol{u}$, so $\nabla^2 \boldsymbol{w} = \boldsymbol{u}$ by Proposition 10.67; and Equation (7.105), read backwards, turns that Laplacian into $\nabla(\nabla \cdot \boldsymbol{w}) - \nabla \times (\nabla \times \boldsymbol{w})$, which is Equation (7.129) with the potentials Equation (7.130). That $\nabla \cdot \boldsymbol{\Psi} = 0$ is Equation (7.104). Uniqueness is the observation that the difference of two decompositions is a field that is at once a gradient and a curl, hence by Equations (7.103) and (7.104) both curl-free and divergence-free, hence harmonic in each component by Equation (7.105); a harmonic function on $\mathbb{R}^3$ that tends to zero at infinity vanishes, by

the maximum principle Theorem 10.75 applied on balls of growing radius. What the appendix supplies is the analytic work the display above skips: that the singular integral Equation (7.130) may be differentiated under the integral sign as often as one likes — Theorem 7.77 after the singularity has been excised, with the excised part estimated uniformly — and that the result decays.

7.12 Sets of measure zero and Hausdorff dimension

The Darboux construction already asks which sets are small enough to be ignored: the graph of a continuous function carries no area (Remark 7.96), which is why the boundary of a simple region contributes nothing to an integral. This section makes "small" precise in the two senses the physical parts use — a set of measure zero, and the dimension of a set that is small in no whole number of dimensions. Neither needs the Lebesgue integral, and this treatise does not construct it; what is defined here is the covering notion alone, and every statement below is proved from it.

Definition 7.106 (Set of measure zero)**.** A set $E \subseteq \mathbb{R}^N$ is *null*, or *of measure zero*, if for every $\varepsilon > 0$ there is a countable family of boxes $R_1, R_2, \ldots$ (Definition 7.93) with

$$E \subseteq \bigcup_{k \geq 1} R_k \,, \qquad \sum_{k \geq 1} \operatorname{vol} R_k < \varepsilon \,, \tag{7.131}$$

where $\operatorname{vol} R$ is the product of the edge lengths of R.

Rests on Definitions 3.65 and 7.93.

Lemma 7.107 (A countable union of null sets is null)**.** *If $E_1, E_2, \ldots$ are null subsets of $\mathbb{R}^N$ then so is $\bigcup_{j \geq 1} E_j$. In particular every countable subset of $\mathbb{R}^N$ is null, and the rational points $\mathbb{Q}^N$ are a null set.*

Rests on Definition 7.106 and Propositions 3.67 and 7.46.

Proof. *Derives Lemma 7.107.*

Let $\varepsilon > 0$. For each j choose, by Definition 7.106, a countable family of boxes covering E_j of total volume less than $\varepsilon 2^{-j}$. The union of these families is a countable family — a countable union of countable families is countable, by Proposition 3.67 — it covers $\bigcup_j E_j$, and its total volume is at most $\sum_{j \geq 1} \varepsilon 2^{-j} = \varepsilon$ by the geometric series (Proposition 7.46). A single point is null, being contained in a cube of arbitrarily small volume, so a countable set is a countable union of null sets and the first part applies; and $\mathbb{Q}^N$ is countable by Corollary 3.68 and Proposition 3.67. □

Definition 7.108 (Hausdorff measure)**.** The *diameter* of a non-empty $U \subseteq \mathbb{R}^N$ is $\operatorname{diam} U = \sup \{ |\boldsymbol{p} - \boldsymbol{q}| \mid \boldsymbol{p}, \boldsymbol{q} \in U \}$. For $E \subseteq \mathbb{R}^N$, $s \geq 0$ and $\delta > 0$ let

$$\mathcal{H}^s_\delta(E) = \inf \sum_{k \geq 1} (\operatorname{diam} U_k)^s \,, \tag{7.132}$$

the infimum taken over all countable families $\{U_k\}$ that cover E and satisfy $\operatorname{diam} U_k \leq \delta$, with the convention that an empty set of competitors gives $+\infty$. Fewer families compete as δ decreases, so $\mathcal{H}^s_\delta(E)$ does not decrease, and the *s-dimensional Hausdorff measure*

$$\mathcal{H}^s(E) = \lim_{\delta \longrightarrow 0^+} \mathcal{H}^s_\delta(E) = \sup_{\delta > 0} \mathcal{H}^s_\delta(E) \in [0, +\infty] \tag{7.133}$$

exists. The normalising constants that make $\mathcal{H}^N$ agree with the ordinary volume on $\mathbb{R}^N$ are omitted; nothing below needs them.

Rests on Definitions 6.24 and 7.106 and Axiom 7.1.

Proposition 7.109 (The critical exponent, and the Hausdorff dimension)**.** *For every $E \subseteq \mathbb{R}^N$ there is exactly one $d = \dim_{\mathrm{H}} E \in [0, N]$, called the* Hausdorff dimension *of E, with*

$$\mathcal{H}^s(E) = \begin{cases} +\infty\,, & 0 \le s < d\,, \\ 0\,, & s > d\,. \end{cases} \tag{7.134}$$

Moreover $\mathcal{H}^N(E) = 0$ implies that E is null in the sense of Definition 7.106; in particular every set of Hausdorff dimension smaller than N is null.

Rests on Definitions 7.106 and 7.108 and Lemma 7.107.

Proof. *Derives Proposition 7.109.*

One threshold. Let $0 \le s < t$ and suppose $\mathcal{H}^s(E) < \infty$. For any family covering E with diameters at most δ,

$$\sum_k (\operatorname{diam} U_k)^t \le \delta^{\,t-s} \sum_k (\operatorname{diam} U_k)^s\,, \tag{7.135}$$

so $\mathcal{H}^t_\delta(E) \le \delta^{\,t-s}\mathcal{H}^s_\delta(E) \le \delta^{\,t-s}\mathcal{H}^s(E)$, which tends to 0 as $\delta \longrightarrow 0^+$: hence $\mathcal{H}^t(E) = 0$ for every $t > s$. Consequently the set $Z = \{s \ge 0 \mid \mathcal{H}^s(E) = 0\}$ contains, with any of its points, every larger one, and contains every s at which $\mathcal{H}^s(E)$ is merely finite together with everything above it. Put $d = \inf Z$, which exists by Axiom 7.1 once Z is known to be non-empty — the next paragraph shows $(N, \infty) \subseteq Z$. If $s > d$ there is $t \in Z$ with $t < s$, so $\mathcal{H}^s(E) = 0$. If $s < d$ then $\mathcal{H}^s(E)$ cannot be finite: were it finite, every $t > s$ would lie in Z and $\inf Z$ would be at most $s < d$. So $\mathcal{H}^s(E) = +\infty$, which is Equation (7.134); and those two clauses determine d as $\inf Z$, so it is unique.

The bound $d \le N$. It is enough to show $\mathcal{H}^s(E) = 0$ for every $s > N$. First, $\mathcal{H}^s_\delta$ is countably subadditive — covers of the pieces concatenate into a cover of the union — so $\mathcal{H}^s_\delta\left(\bigcup_j E_j\right) \le \sum_j \mathcal{H}^s_\delta(E_j) \le \sum_j \mathcal{H}^s(E_j)$, and letting $\delta \longrightarrow 0^+$ shows $\mathcal{H}^s$ is countably subadditive too. Since $\mathbb{R}^N$ is the union of countably many unit lattice cubes, it is enough to treat E inside one such cube Q. Split Q into n^N closed subcubes of edge $1/n$, hence of diameter $\sqrt{N}/n$: they cover E, and

$$n^N \left(\frac{\sqrt{N}}{n}\right)^s = N^{s/2}\, n^{\,N-s} \longrightarrow 0 \qquad (n \longrightarrow \infty)$$

whenever $s > N$, while their diameters tend to 0. Hence $\mathcal{H}^s_\delta(E) = 0$ for every $\delta > 0$, so $\mathcal{H}^s(E) = 0$.

Null sets. Suppose $\mathcal{H}^N(E) = 0$ and let $\varepsilon > 0$. Choose a family $\{U_k\}$ covering E with $\sum_k (\operatorname{diam} U_k)^N < \varepsilon$. A set of diameter t has each of its N coordinate projections contained in an interval of length t, hence lies in a closed cube of edge t and volume t^N. Replacing each U_k by such a cube covers E by countably many boxes of total volume less than ε, which is Equation (7.131). If $\dim_{\mathrm{H}} E < N$ then $\mathcal{H}^N(E) = 0$ by Equation (7.134), so E is null. □

Definition 7.110 (Box-counting dimension)**.** For a bounded non-empty $E \subseteq \mathbb{R}^N$ let $N_\delta(E)$ be the least number of sets of diameter at most δ needed to cover E, a finite number because E is bounded. The *upper* and *lower box-counting dimensions* of E are

$$\overline{\dim}_{\mathrm{B}} E = \limsup_{\delta \longrightarrow 0^+} \frac{\ln N_\delta(E)}{\ln(1/\delta)}\,, \qquad \underline{\dim}_{\mathrm{B}} E = \liminf_{\delta \longrightarrow 0^+} \frac{\ln N_\delta(E)}{\ln(1/\delta)}\,, \tag{7.136}$$

and where the two agree their common value is the *box-counting dimension* $\dim_{\mathrm{B}} E$. It is this quantity, not the Hausdorff dimension, that a numerical box-count measures.

Rests on Definitions 6.9 and 7.108.

Proposition 7.111 (The Hausdorff dimension never exceeds the box dimension)**.** *For every bounded non-empty $E \subseteq \mathbb{R}^N$,*

$$\dim_{\mathrm{H}} E \le \underline{\dim}_{\mathrm{B}} E \le \overline{\dim}_{\mathrm{B}} E \,, \tag{7.137}$$

and the first inequality can be strict: the rationals of $[0,1]$ have Hausdorff dimension 0 and box-counting dimension 1.

Rests on Definition 7.110, Proposition 7.109, and Lemma 7.107.

Proof. *Derives Proposition 7.111.*

The second inequality is the definition of a lower and an upper limit. For the first, let $s > \underline{\dim}_{\mathrm{B}} E$ and pick s' with $\underline{\dim}_{\mathrm{B}} E < s' < s$. By Equation (7.136) there is a sequence $\delta_1 > \delta_2 > \cdots \longrightarrow 0$ along which $\ln N_{\delta_j}(E)/\ln(1/\delta_j) < s'$, that is $N_{\delta_j}(E) < \delta_j^{-s'}$. Covering E by that many sets of diameter at most δ_j,

$$\mathcal{H}^s_{\delta_j}(E) \le N_{\delta_j}(E)\, \delta_j^{\,s} < \delta_j^{\,s-s'} \longrightarrow 0 \,,$$

and since $\mathcal{H}^s(E)$ is the limit of $\mathcal{H}^s_\delta(E)$ along any sequence $\delta \longrightarrow 0^+$ (Equation (7.133)), it follows that $\mathcal{H}^s(E) = 0$ and hence $\dim_{\mathrm{H}} E \le s$ by Equation (7.134). Letting s decrease to $\underline{\dim}_{\mathrm{B}} E$ gives Equation (7.137).

For the example, let $E = \mathbb{Q} \cap [0,1]$. It is countable, so for $s > 0$ and $\varepsilon \in (0, \delta)$ covering its k-th point by an interval of length $\varepsilon 2^{-k}$ gives $\mathcal{H}^s_\delta(E) \le \varepsilon^s \sum_{k \ge 1} 2^{-ks} = \varepsilon^s/(2^s - 1)$, which tends to 0 with ε; hence $\mathcal{H}^s(E) = 0$ for every $s > 0$ and $\dim_{\mathrm{H}} E = 0$ (consistently with Lemma 7.107, which already makes E null). But any cover of E by sets of diameter at most δ may be replaced by the cover by their closures, of the same diameters, which then covers the closure $[0,1]$ of E; so $N_\delta(E) \ge 1/\delta$, and Equation (7.136) gives $\underline{\dim}_{\mathrm{B}} E \ge 1$, while $E \subseteq [0,1]$ forces $\overline{\dim}_{\mathrm{B}} E \le 1$. □

Remark 7.112 (What is developed here, and what is not). Exactly three notions are constructed above — the null set, the Hausdorff measures $\mathcal{H}^s$, and the two counting dimensions — and nothing else is assumed. The Lebesgue integral, the σ-algebra of measurable sets and the countable additivity of a measure defined on it are not built in this treatise, and no statement in it rests on them: where an argument needs "almost every point" it needs Definition 7.106 and Lemma 7.107, and where it needs a fractal dimension it needs Definition 7.108 or Definition 7.110. The integral used throughout remains the Darboux–Riemann integral of Definition 7.39 and Definition 7.93.

Rests on Definitions 7.106, 7.108 and 7.110.

7.13 Functional calculus

A *functional* assigns a number to a whole function, the archetype being the action integral $S[q] = \int_{t_1}^{t_2} \mathcal{L}(q(t), \dot q(t), t)\, \mathrm{d}t$ of Chapter 21. The *functional derivative* $\delta S/\delta q(t)$ is defined by the first-order response

$$S[q + \varepsilon\eta] - S[q] = \varepsilon \int \frac{\delta S}{\delta q(t)}\, \eta(t)\, \mathrm{d}t + O(\varepsilon^2) \tag{7.138}$$

for arbitrary smooth variations η vanishing at the endpoints. Its calculus — the Euler–Lagrange equation, constrained variations, second variations — is developed where it is first needed, in Chapter 21, with the general field-theoretic case in Chapter 64.

7.14 The Dirac delta

No function $\delta : \mathbb{R} \to \mathbb{R}$ satisfies $\int f(x)\,\delta(x)\,\mathrm{d}x = f(0)$ for all continuous f; the Dirac delta is instead a *distribution*: the linear functional $\delta : f \longmapsto f(0)$ on smooth test functions of compact support. The integral notation

$$\int_{-\infty}^{\infty} f(x)\,\delta(x - x_0)\,\mathrm{d}x = f(x_0) \tag{7.139}$$

is shorthand for its action, the *sifting property*, which is the definition rather than a theorem.

Proposition 7.113 (Delta as a limit; elementary properties)**.** *Let* $\delta_\sigma(x) = \dfrac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2\sigma^2}$. *Then for every bounded continuous* f*,*

$$\lim_{\sigma \to 0^+} \int_{-\infty}^{\infty} f(x)\,\delta_\sigma(x - x_0)\,\mathrm{d}x = f(x_0)\,, \tag{7.140}$$

and, as distributions, $\delta(ax) = \delta(x)/\,|a|$ *for* $a \neq 0$*, and* $x\,\delta(x) = 0$*.*
Rests on Equation (7.139)*, Corollary 7.44, and Definition 7.20.*

Proof. *Derives Proposition 7.113.*

Each δ_σ is positive with unit integral (the Gaussian integral; computed in Chapter 11), so

$$\left| \int f\,\delta_\sigma(x - x_0)\,\mathrm{d}x - f(x_0) \right| \leq \int |f(x) - f(x_0)|\;\delta_\sigma(x - x_0)\,\mathrm{d}x\,.$$

Split the domain at $|x - x_0| = \eta$: on the inner part continuity makes $|f - f(x_0)| < \varepsilon$ and the weight integrates to at most 1; on the outer part $|f - f(x_0)| \leq 2\sup|f|$ while the Gaussian tail mass tends to 0 as $\sigma \to 0$ for fixed η. Both contributions are eventually below ε. The scaling rule follows by substituting $u = ax$ in $\int f(x)\,\delta(ax)\,\mathrm{d}x$ (Corollary 7.44), which gives $f(0)/\,|a|$ (the sign of a flips the orientation, whence the absolute value); and $\int f(x)\,x\,\delta(x)\,\mathrm{d}x = 0 \cdot f(0) = 0$ for every test function, which is the meaning of $x\,\delta(x) = 0$. □

The three-dimensional delta $\delta^3(\boldsymbol{r}) = \delta(x)\delta(y)\delta(z)$ represents point sources; its central identity for field theory,

$$\nabla^2 \frac{1}{|\boldsymbol{r}|} = -4\pi\,\delta^3(\boldsymbol{r})\,, \tag{7.141}$$

is derived where it is used, in electrostatics (Chapter 58).

Chapter 8

Complex Analysis

Complex analysis is the calculus of functions differentiable in the complex sense — a condition so much stronger than real differentiability that it forces infinite smoothness, rigidity, and the exact integral formulas carrying the factor $2\pi\mathrm{i}$ on which physics relies: contour representations of Green functions (Chapter 10), dispersion relations, and the spectral calculus of operators (Chapter 12). Everything rests on the real analysis of Chapter 7, hence ultimately on the completeness of $\mathbb{R}$ (Appendix A.7).

8.1 The complex field

Definition 8.1 (Complex numbers)**.** $\mathbb{C}$ is the set $\mathbb{R}^2$ equipped with componentwise addition and the multiplication

$$(x_1, y_1)\,(x_2, y_2) = (x_1x_2 - y_1y_2,\; x_1y_2 + y_1x_2)\,. \tag{8.1}$$

Writing $z = x + \mathrm{i}y$ with $\mathrm{i} = (0,1)$, so that $\mathrm{i}^2 = -1$, the maps $z \mapsto \Re z = x$ and $z \mapsto \Im z = y$ extract the real and imaginary parts, $\bar z = x - \mathrm{i}y$ is the *conjugate*, and $|z| = \sqrt{x^2+y^2} = \sqrt{z\bar z}$ the *modulus*.

Proposition 8.2 ($\mathbb{C}$ is a field)**.** *$(\mathbb{C}, +, \cdot)$ is a field, with $z^{-1} = \bar z / |z|^2$ for $z \neq 0$; the modulus is multiplicative, $|zw| = |z|\,|w|$, and satisfies the triangle inequality $|z+w| \le |z| + |w|$.*
Rests on Equations (6.12) and (8.1).

Proof. *Derives Proposition 8.2.*
The field axioms are direct computations with Equation (8.1); associativity and distributivity reduce to real arithmetic, and $z\,\bar z/|z|^2 = |z|^2/|z|^2 = 1$ exhibits the inverse. Multiplicativity: $|zw|^2 = zw\,\overline{zw} = z\bar z\, w\bar w = |z|^2\,|w|^2$, using $\overline{zw} = \bar z\,\bar w$ (a computation). Triangle inequality: with the Euclidean identification this is the triangle inequality of $\mathbb{R}^2$ (Equation (6.12)); algebraically, $|z+w|^2 = |z|^2 + |w|^2 + 2\Re(z\bar w) \le |z|^2 + |w|^2 + 2\,|z|\,|w|$ since $\Re u \le |u|$. □

$\mathbb{C}$ inherits the metric $d(z,w) = |z-w|$ of the Euclidean plane, hence its topology (Chapter 6) and its completeness ($\mathbb{R}^2$ is complete, Section 6.3). Unlike $\mathbb{R}$, the field $\mathbb{C}$ carries no compatible order — $\mathrm{i}^2 = -1$ would violate the sign rules of an ordered field — and in exchange it will turn out to be algebraically closed (Theorem 8.19).

Definition 8.3 (Exponential, sine, cosine)**.** For $z \in \mathbb{C}$,

$$\exp z = \mathrm{e}^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}\,, \qquad \cos z = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!}\,, \qquad \sin z = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!}\,. \tag{8.2}$$

All three converge absolutely for every z by the ratio test (Proposition 7.48), so their radius of convergence is infinite (Theorem 7.50).
Rests on Proposition 7.48 and Theorem 7.50.

Proposition 8.4 (Euler's formula)**.** *For $\theta \in \mathbb{R}$,*

$$\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta\,, \tag{8.3}$$

and every $z \neq 0$ has the polar form $z = |z|\,\mathrm{e}^{\mathrm{i}\arg z}$, with $\arg z$ defined modulo 2π.

Rests on Equation (8.2) and Proposition 7.47.

Proof. *Derives Proposition 8.4.*

Insert $z = \mathrm{i}\theta$ into Equation (8.2) and split even and odd powers, using $\mathrm{i}^{2n} = (-1)^n$ and $\mathrm{i}^{2n+1} = (-1)^n\mathrm{i}$: absolute convergence permits the rearrangement (Proposition 7.47), and the two sub-series are exactly $\cos\theta$ and $\mathrm{i}\sin\theta$. The polar form is plane trigonometry: $(x, y) = |z|\,(\cos\theta, \sin\theta)$ for the angle θ of the point. □

The number 2π enters complex analysis here, as the period of $\theta \mapsto \mathrm{e}^{\mathrm{i}\theta}$; every $2\pi\mathrm{i}$ in the theorems below descends from one full turn of this exponential.

8.2 Holomorphic functions

Definition 8.5 (Complex derivative)**.** Let $\Omega \subseteq \mathbb{C}$ be open and $f : \Omega \longrightarrow \mathbb{C}$. f is *complex differentiable* at z_0 iff

$$f'(z_0) = \lim_{h \longrightarrow 0} \frac{f(z_0 + h) - f(z_0)}{h} \tag{8.4}$$

exists, with $h \in \mathbb{C}$ approaching 0 from *every* direction of the plane. f is *holomorphic on* Ω iff it is complex differentiable at every point of Ω.

The rules of differentiation — linearity, Leibniz, chain, quotient — hold with the proofs of Section 7.6.3 verbatim, since those proofs used only field arithmetic and limits. Polynomials are holomorphic; so is e^z term by term inside its radius (Section 7.9), with $(\mathrm{e}^z)' = \mathrm{e}^z$.

Theorem 8.6 (Cauchy–Riemann equations)**.** *Write $f(x + \mathrm{i}y) = u(x, y) + \mathrm{i}v(x, y)$. If f is complex differentiable at $z_0 = x_0 + \mathrm{i}y_0$, then the partial derivatives exist there and satisfy*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}\,, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}\,. \tag{8.5}$$

Conversely, if u, v are C^1 near z_0 and satisfy Equation (8.5) at z_0, then f is complex differentiable at z_0, with $f' = \partial_x u + \mathrm{i}\,\partial_x v$.

Rests on Equation (8.4) and Definition 7.66.

Proof. *Derives Theorem 8.6.*

Necessity. Take the limit Equation (8.4) first along real $h = t$: $f'(z_0) = \partial_x u + \mathrm{i}\,\partial_x v$. Then along imaginary $h = \mathrm{i}t$:

$$f'(z_0) = \lim_{t \to 0} \frac{f(z_0 + \mathrm{i}t) - f(z_0)}{\mathrm{i}t} = \frac{1}{\mathrm{i}}\left(\frac{\partial u}{\partial y} + \mathrm{i}\,\frac{\partial v}{\partial y}\right) = \frac{\partial v}{\partial y} - \mathrm{i}\,\frac{\partial u}{\partial y}\,.$$

Equating real and imaginary parts of the two expressions gives Equation (8.5).

Sufficiency. C^1 regularity makes (u, v) differentiable as a map $\mathbb{R}^2 \to \mathbb{R}^2$ (Definition 7.66), so

$$f(z_0 + h) - f(z_0) = (\partial_x u + \mathrm{i}\,\partial_x v)\,\eta + (\partial_y u + \mathrm{i}\,\partial_y v)\,\xi + o(|h|)\,,$$

with $h = \eta + \mathrm{i}\xi$. Substituting Equation (8.5) turns the second coefficient into $\mathrm{i}\,(\partial_x u + \mathrm{i}\,\partial_x v)$, so the two terms combine into $(\partial_x u + \mathrm{i}\,\partial_x v)\,(\eta + \mathrm{i}\xi) = (\partial_x u + \mathrm{i}\,\partial_x v)\,h$; dividing by h and letting $h \to 0$ kills the $o(|h|)/h$ remainder. □

Example 8.7. $f(z) = \bar{z}$ has $u = x$, $v = -y$, so $\partial_x u = 1 \neq -1 = \partial_y v$: nowhere complex differentiable, despite being real-smooth. Complex differentiability is a genuine constraint, not a notational variant — it demands that the Jacobian of (u, v) be a rotation-scaling, i.e. that f preserve angles where $f' \neq 0$.

Rests on Theorem 8.6.

8.3 Contour integrals

Definition 8.8 (Contour integral)**.** A *contour* γ is a piecewise-C^1 path $z : [a,b] \longrightarrow \mathbb{C}$. For continuous f,

$$\int_\gamma f(z)\,\mathrm{d}z = \int_a^b f(z(t))\,z'(t)\,\mathrm{d}t\,, \tag{8.6}$$

the right side being a pair of real integrals (Section 7.7). The integral is independent of the parametrization (substitution rule, Corollary 7.44) and changes sign under orientation reversal. A contour is *closed* iff $z(b) = z(a)$, written $\oint_\gamma$.

Rests on Corollary 7.44.

Lemma 8.9 (ML estimate)**.** *If $|f| \le M$ on γ and $L = \int_a^b |z'(t)|\,\mathrm{d}t$ is the length of γ, then $\left|\int_\gamma f\,\mathrm{d}z\right| \le ML$.*

Rests on Equation (8.6).

Proof. *Derives Lemma 8.9.*

Write $\int_\gamma f\,\mathrm{d}z = R\,\mathrm{e}^{\mathrm{i}\varphi}$ with $R \ge 0$. Then $R = \Re \int_a^b \mathrm{e}^{-\mathrm{i}\varphi} f(z(t))\,z'(t)\,\mathrm{d}t \le \int_a^b |f(z(t))|\,|z'(t)|\,\mathrm{d}t \le ML$, using $\Re u \le |u|$ and monotonicity of the real integral. □

Proposition 8.10 (Fundamental theorem for contours)**.** *If $f = F'$ for a holomorphic F on Ω and $\gamma \subset \Omega$ runs from z_1 to z_2, then $\int_\gamma f\,\mathrm{d}z = F(z_2) - F(z_1)$; in particular the integral of f over every closed contour vanishes.*

Rests on Equation (8.6) and Theorem 7.43.

Proof. *Derives Proposition 8.10.*

$\frac{\mathrm{d}}{\mathrm{d}t}F(z(t)) = F'(z(t))\,z'(t)$ by the chain rule, so Equation (8.6) is $\int_a^b \frac{\mathrm{d}}{\mathrm{d}t}F(z(t))\,\mathrm{d}t = F(z(b)) - F(z(a))$ by the fundamental theorem of calculus (Theorem 7.43), applied to real and imaginary parts. □

Lemma 8.11 (The fundamental $2\pi\mathrm{i}$)**.** *Let C_r be the circle $|z-a| = r$ traversed once counterclockwise. Then for $n \in \mathbb{Z}$,*

$$\oint_{C_r} (z-a)^n\,\mathrm{d}z = \begin{cases} 2\pi\mathrm{i}\,, & n = -1\,, \\ 0\,, & n \ne -1\,. \end{cases} \tag{8.7}$$

Rests on Equation (8.6) and Proposition 8.4.

Proof. *Derives Lemma 8.11.*

Parametrize $z = a + r\mathrm{e}^{\mathrm{i}\theta}$, $\theta \in [0, 2\pi]$, $\mathrm{d}z = \mathrm{i}r\mathrm{e}^{\mathrm{i}\theta}\mathrm{d}\theta$:

$$\oint_{C_r} (z-a)^n\,\mathrm{d}z = \mathrm{i}r^{\,n+1} \int_0^{2\pi} \mathrm{e}^{\mathrm{i}(n+1)\theta}\,\mathrm{d}\theta\,.$$

For $n \ne -1$ the integrand is the derivative of $\mathrm{e}^{\mathrm{i}(n+1)\theta}/(\mathrm{i}(n+1))$, periodic over $[0, 2\pi]$, so the integral vanishes; for $n = -1$ it is $\mathrm{i}\int_0^{2\pi} \mathrm{d}\theta = 2\pi\mathrm{i}$. □

Every $2\pi\mathrm{i}$ below is this lemma: one full turn of Euler's exponential, surviving only in the term $(z-a)^{-1}$ — precisely the power with no single-valued antiderivative.

8.4 Cauchy's integral theorem

Theorem 8.12 (Cauchy)**.** *Let f be holomorphic with continuous derivative on an open set containing a region D and its piecewise-C^1 boundary ∂D (counterclockwise). Then*

$$\oint_{\partial D} f(z)\,\mathrm{d}z = 0\,. \tag{8.8}$$

Rests on Theorem 7.99 and Equation (8.5).

Proof. *Derives Theorem 8.12.*

Separate real and imaginary parts with $f = u + \mathrm{i}v$ and $\mathrm{d}z = \mathrm{d}x + \mathrm{i}\,\mathrm{d}y$:

$$\oint_{\partial D} f\,\mathrm{d}z = \oint_{\partial D} (u\,\mathrm{d}x - v\,\mathrm{d}y) + \mathrm{i}\oint_{\partial D} (v\,\mathrm{d}x + u\,\mathrm{d}y)\ .$$

Green's theorem (Theorem 7.99) converts the two circulations to

$$\iint_D \left(-\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\mathrm{d}x\,\mathrm{d}y + \mathrm{i}\iint_D \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right)\mathrm{d}x\,\mathrm{d}y\,,$$

and both integrands vanish identically by the Cauchy–Riemann equations Equation (8.5). □

Continuity of f' need not be assumed. Goursat's quadrisection argument reaches the same conclusion on a triangle from bare complex differentiability, and pays for the weaker hypothesis by giving up Green's theorem in favour of an estimate that shrinks the contour.

Theorem 8.13 (Goursat)**.** *Let $\Omega \subseteq \mathbb{C}$ be open and let f be complex differentiable at every point of Ω, with no assumption whatever on the continuity of f'. Then for every closed triangle $\Delta \subset \Omega$, with boundary $\partial\Delta$ traversed once,*

$$\oint_{\partial\Delta} f(z)\,\mathrm{d}z = 0\,. \tag{8.9}$$

Rests on Equation (8.4), Proposition 8.10, and Lemma 8.9.

Proof. *Derives Theorem 8.13.*

Write $I(\Delta) = \oint_{\partial\Delta} f\,\mathrm{d}z$ and let L and d be the perimeter and the diameter of Δ.

Quadrisection. Joining the midpoints of the three sides divides Δ into four closed triangles $\Delta^{(1)}, \dots, \Delta^{(4)}$, each similar to Δ with ratio $1/2$, hence of perimeter $L/2$ and diameter $d/2$. Traversing the four boundaries coherently, every interior edge is covered twice in opposite senses and cancels, so

$$I(\Delta) = \sum_{j=1}^{4} I(\Delta^{(j)})\,,$$

and at least one of the four obeys $\left|I(\Delta^{(j)})\right| \geq |I(\Delta)|\,/4$; call it Δ_1. Iterating the construction produces nested closed triangles $\Delta = \Delta_0 \supseteq \Delta_1 \supseteq \cdots$ with

$$|I(\Delta)| \leq 4^n\ |I(\Delta_n)|\,, \qquad L_n = 2^{-n}L\,, \qquad d_n = 2^{-n}d\,. \tag{8.10}$$

The limit point. Pick $z_n \in \Delta_n$. For $m \geq n$ both z_m and z_n lie in Δ_n, so $|z_m - z_n| \leq d_n \longrightarrow 0$ and the sequence is Cauchy; by completeness of $\mathbb{C}$ (Section 6.3) it converges to some z_0, which belongs to every Δ_n, each Δ_n being closed and containing all later terms. In particular $z_0 \in \Omega$.

The estimate. Complex differentiability at z_0 (Equation (8.4)) means that

$$f(z) = f(z_0) + f'(z_0)\,(z - z_0) + \eta(z)\,(z - z_0)\,, \tag{8.11}$$

with $\eta(z) \longrightarrow 0$ as $z \longrightarrow z_0$ and $\eta(z_0) = 0$. The affine part of Equation (8.11) is the derivative of the polynomial $f(z_0)\,z + \frac{1}{2}f'(z_0)\,(z - z_0)^2$, so its integral over the closed contour $\partial\Delta_n$ vanishes (Proposition 8.10), leaving

$$I(\Delta_n) = \oint_{\partial\Delta_n} \eta(z)\,(z - z_0)\,\mathrm{d}z\,.$$

Given $\varepsilon > 0$, choose $\rho > 0$ with $|\eta(z)| \le \varepsilon$ whenever $|z - z_0| < \rho$, and take n large enough that $d_n < \rho$. Since $z_0 \in \Delta_n$, every z on $\partial\Delta_n$ satisfies $|z - z_0| \le d_n$, so the integrand is bounded by $\varepsilon\, d_n$ and the ML estimate (Lemma 8.9) gives

$$|I(\Delta_n)| \le \varepsilon\, d_n\, L_n = \varepsilon\, d\, L\, 4^{-n}\,.$$

Combined with Equation (8.10) this yields $|I(\Delta)| \le \varepsilon\, d\, L$ for every $\varepsilon > 0$, whence $I(\Delta) = 0$. □

Remark 8.14 (Goursat). Theorem 8.13 removes the continuity hypothesis on f' from Theorem 8.12. The triangle suffices: in a convex domain the function $F(z) = \int_{z_*}^{z} f(w)\,\mathrm{d}w$, integrated along the straight segment from a fixed z_*, is well defined, and Equation (8.9) applied to the triangle $z_*, z, z+h$ gives $F(z+h) - F(z) = \int_z^{z+h} f(w)\,\mathrm{d}w$, whence $F' = f$ by continuity of f; Proposition 8.10 then returns Equation (8.8) for every closed contour in the domain, and a general region is handled by triangulation. Once the integral formula Equation (8.12) is available it shows that f' is automatically continuous — indeed infinitely differentiable (Theorem 8.17) — so the hypothesis discarded here was never a restriction, and the circle closes.

Corollary 8.15 (Deformation of contours)**.** *If f is holomorphic on a region containing two closed contours γ_1, γ_2 and the annular region between them, then $\oint_{\gamma_1} f\,\mathrm{d}z = \oint_{\gamma_2} f\,\mathrm{d}z$. In a simply connected domain (Definition 6.17), the integral of a holomorphic function over any closed contour vanishes.*

Rests on Theorem 8.12 and Definition 6.17.

Proof. *Derives Corollary 8.15.*

Join the two contours by a cut; the boundary of the cut annulus is $\gamma_1 - \gamma_2$ plus the cut traversed both ways, whose contributions cancel. Cauchy's theorem on the annulus gives $\oint_{\gamma_1} - \oint_{\gamma_2} = 0$. In a simply connected domain every closed contour bounds, and Theorem 8.12 applies directly. □

8.5 The Cauchy integral formula

Theorem 8.16 (Cauchy integral formula)**.** *Let f be holomorphic on an open set containing the disc $|z - a| \le r$ and its boundary circle C (counterclockwise). Then for every z_0 inside C,*

$$f(z_0) = \frac{1}{2\pi\mathrm{i}} \oint_C \frac{f(z)}{z - z_0}\,\mathrm{d}z\,. \tag{8.12}$$

Rests on Corollary 8.15 and Lemmas 8.9 and 8.11.

Proof. *Derives Theorem 8.16.*

The integrand is holomorphic between C and any small circle $C_\varepsilon : |z - z_0| = \varepsilon$, so by deformation (Corollary 8.15)

$$\oint_C \frac{f(z)}{z - z_0}\,\mathrm{d}z = \oint_{C_\varepsilon} \frac{f(z)}{z - z_0}\,\mathrm{d}z = f(z_0) \oint_{C_\varepsilon} \frac{\mathrm{d}z}{z - z_0} + \oint_{C_\varepsilon} \frac{f(z) - f(z_0)}{z - z_0}\,\mathrm{d}z\,.$$

The first integral is $2\pi\mathrm{i}$ by Lemma 8.11. In the second, continuity of f at z_0 bounds the integrand by $\varepsilon^{-1} \sup_{C_\varepsilon} |f - f(z_0)|$, so the ML estimate (Lemma 8.9) bounds the integral by $2\pi \sup_{C_\varepsilon} |f - f(z_0)| \longrightarrow 0$ as $\varepsilon \to 0$. Since the left side does not depend on ε, it equals $2\pi\mathrm{i} f(z_0)$. □

A holomorphic function is thus determined inside a circle by its boundary values — the first sign of the rigidity of holomorphy.

Theorem 8.17 (Derivatives of all orders; Cauchy estimates)**.** *Under the hypotheses of Theorem 8.16, f has complex derivatives of every order inside C, given by*

$$f^{(n)}(z_0) = \frac{n!}{2\pi\mathrm{i}} \oint_C \frac{f(z)}{(z-z_0)^{n+1}}\,\mathrm{d}z\,, \tag{8.13}$$

and if $|f| \le M$ on C and z_0 is its centre, $\left|f^{(n)}(z_0)\right| \le n!\,M/r^n$.

Rests on Theorem 8.16 and Lemma 8.9.

Proof. *Derives Theorem 8.17.*

For $n = 1$, form the difference quotient of Equation (8.12) at z_0 and $z_0 + h$:

$$\frac{f(z_0+h) - f(z_0)}{h} = \frac{1}{2\pi\mathrm{i}} \oint_C \frac{f(z)}{(z-z_0-h)(z-z_0)}\,\mathrm{d}z\,,$$

an algebraic identity of the two kernels. As $h \to 0$ the integrand converges uniformly on C to $f(z)/(z-z_0)^2$ (the denominators are bounded away from zero for z_0 interior and $|h|$ small), so the ML estimate carries the limit inside, giving Equation (8.13) with $n = 1$. Induction repeats the argument on each successive formula. The estimate follows from Lemma 8.9 with $|z - z_0| = r$ and $L = 2\pi r$. □

Theorem 8.18 (Liouville)**.** *A function holomorphic and bounded on all of $\mathbb{C}$ is constant.*

Rests on Theorem 8.17 and Proposition 8.10.

Proof. *Derives Theorem 8.18.*

If $|f| \le M$ everywhere, the Cauchy estimate with $n = 1$ on a circle of radius r about any z_0 gives $|f'(z_0)| \le M/r \longrightarrow 0$ as $r \to \infty$. So $f' \equiv 0$, and (integrating along any path, Proposition 8.10) f is constant. □

Theorem 8.19 (Fundamental theorem of algebra)**.** *Every non-constant polynomial p with complex coefficients has a root in $\mathbb{C}$.*

Rests on Theorem 8.18.

Proof. *Derives Theorem 8.19.*

If p had no root, $1/p$ would be holomorphic on $\mathbb{C}$ (quotient rule). For large $|z|$ the leading term dominates, $|p(z)| \ge \frac{1}{2}\,|c_n|\,|z|^n \longrightarrow \infty$, so $1/p \to 0$ at infinity and, being continuous on the (compact, Theorem 6.11) complement of the large-$|z|$ region, is bounded on $\mathbb{C}$. Liouville (Theorem 8.18) forces $1/p$ constant — contradicting non-constancy of p. Dividing out roots inductively, p factors completely: $\mathbb{C}$ is algebraically closed. □

Theorem 8.20 (Taylor expansion)**.** *A function holomorphic on the open disc $|z - a| < R$ equals its Taylor series there:*

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}\,(z-a)^n\,, \qquad |z-a| < R\,. \tag{8.14}$$

Rests on Equations (8.12) and (8.13), Proposition 7.46, and Lemma 8.9.

Proof. *Derives Theorem 8.20.*

Fix z with $|z - a| = \rho < r < R$ and integrate over $C : |w - a| = r$. Expand the Cauchy kernel geometrically (Proposition 7.46), uniformly on C since $|(z-a)/(w-a)| = \rho/r < 1$:

$$\frac{1}{w-z} = \frac{1}{(w-a)\left(1 - \frac{z-a}{w-a}\right)} = \sum_{n=0}^{\infty} \frac{(z-a)^n}{(w-a)^{n+1}}\,.$$

Insert into Equation (8.12) and exchange sum and integral (uniform convergence, ML estimate); the coefficient integrals are Equation (8.13) evaluated at a, giving Equation (8.14). □

Holomorphy and analyticity thus coincide on $\mathbb{C}$ — the deep justification for defining the elementary functions by their series in Definition 8.3.

8.6 Laurent series and singularities

Theorem 8.21 (Laurent expansion)**.** *Let f be holomorphic on the annulus $r < |z-a| < R$. Then, uniquely,*

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z-a)^n, \qquad c_n = \frac{1}{2\pi \mathrm{i}} \oint_C \frac{f(w)}{(w-a)^{n+1}} \, \mathrm{d}w, \tag{8.15}$$

with C any counterclockwise circle in the annulus.

Rests on Theorems 8.16 and 8.20, Corollary 8.15, and Lemma 8.11.

Proof. *Derives Theorem 8.21.*

For z in the annulus choose radii $r < r_1 < |z-a| < r_2 < R$. The cut-annulus argument of Corollary 8.15 applied to the Cauchy formula gives

$$f(z) = \frac{1}{2\pi \mathrm{i}} \oint_{|w-a|=r_2} \frac{f(w)}{w-z} \, \mathrm{d}w - \frac{1}{2\pi \mathrm{i}} \oint_{|w-a|=r_1} \frac{f(w)}{w-z} \, \mathrm{d}w .$$

On the outer circle expand the kernel as in Theorem 8.20 (powers $(z-a)^n$, $n \geq 0$); on the inner circle the roles reverse, $|(w-a)/(z-a)| < 1$, giving $-1/(w-z) = \sum_{m \geq 0} (w-a)^m/(z-a)^{m+1}$, i.e. the negative powers. Deforming both coefficient circles to the common C (Corollary 8.15) yields Equation (8.15). Uniqueness: integrating $\sum c_n (z-a)^n$ against $(z-a)^{-m-1}$ term by term over C kills every term except $n = m$ by Lemma 8.11, returning c_m. □

Definition 8.22 (Isolated singularities; residue)**.** If f is holomorphic on a punctured disc $0 < |z-a| < R$, the point a is an *isolated singularity*, classified by the Laurent series Equation (8.15) on the punctured disc:

- *removable* iff $c_n = 0$ for all $n < 0$ (then f extends holomorphically through a);
- a *pole of order m* iff $c_{-m} \neq 0$ and $c_n = 0$ for $n < -m$;
- *essential* iff infinitely many negative coefficients are nonzero.

The coefficient c_{-1} is the *residue* of f at a, written $\operatorname{Res}_{z=a} f$ — the sole survivor of termwise integration (Lemma 8.11).

Rests on Theorem 8.21 and Lemma 8.11.

Proposition 8.23 (Residue at a simple pole)**.** *If $f = g/h$ with g, h holomorphic at a, $g(a) \neq 0$, $h(a) = 0$, $h'(a) \neq 0$, then*

$$\operatorname{Res}_{z=a} \frac{g}{h} = \frac{g(a)}{h'(a)} . \tag{8.16}$$

Rests on Theorem 8.20 and Definition 8.22.

Proof. *Derives Proposition 8.23.*

$h(z) = h'(a)(z-a)\,[1 + O(z-a)]$ by its Taylor expansion (Theorem 8.20), so $f(z) = \dfrac{g(a)}{h'(a)} (z-a)^{-1} +$ (holomorphic part); the coefficient of $(z-a)^{-1}$ is read off. □

8.7 The residue theorem

Theorem 8.24 (Residue theorem)**.** *Let f be holomorphic on an open set containing a region D and its boundary ∂D (counterclockwise), except for finitely many isolated singularities $a_1, \ldots, a_k$ in the interior of D. Then*

$$\oint_{\partial D} f(z) \, \mathrm{d}z = 2\pi \mathrm{i} \sum_{j=1}^{k} \operatorname{Res}_{z=a_j} f . \tag{8.17}$$

Rests on Corollary 8.15, Theorem 8.21, Lemma 8.11, and Definition 8.22.

Proof. *Derives Theorem 8.24.*

Excise from D a small disc $|z-a_j|\le\varepsilon$ around each singularity. On the excised region f is holomorphic, so Cauchy's theorem in the cut-boundary form (Corollary 8.15) gives

$$\oint_{\partial D} f\,\mathrm{d}z=\sum_{j=1}^{k}\oint_{C_{\varepsilon,j}} f\,\mathrm{d}z\,,$$

with $C_{\varepsilon,j}$ the small circles, counterclockwise. On each, insert the Laurent series of f about a_j (Theorem 8.21), which converges uniformly on $C_{\varepsilon,j}$, and integrate term by term: by Lemma 8.11 every power dies except $(z-a_j)^{-1}$, which contributes $2\pi\mathrm{i}\,\operatorname{Res}_{z=a_j} f$. □

Example 8.25 (A real integral by residues)**.** For the archetype

$$\int_{-\infty}^{\infty}\frac{\mathrm{d}x}{1+x^2}=\pi\,,\tag{8.18}$$

close the contour with the upper semicircle S_R of radius R. The integrand $1/(1+z^2)$ has simple poles at $z=\pm\mathrm{i}$; only $z=\mathrm{i}$ lies inside, with residue $1/(2\mathrm{i})$ by Equation (8.16) (here $h'=2z$). The residue theorem gives $\oint=2\pi\mathrm{i}\cdot\frac{1}{2\mathrm{i}}=\pi$, while on S_R the ML estimate (Lemma 8.9) bounds the contribution by $\pi R/(R^2-1)\longrightarrow 0$. The real line's share is therefore π — in agreement with the elementary antiderivative arctan, but obtained by a method that survives when no elementary antiderivative exists. This closing of contours is the engine behind the Green-function and propagator computations of Chapter 10 and Part XI.

Rests on Theorem 8.24, Equation (8.16), and Lemma 8.9.

With the residue theorem the chapter's arc is complete: one full turn of Euler's exponential (Lemma 8.11) has become an exact bookkeeping device relating global contour data to local expansion coefficients. The deeper function theory — analytic continuation, the argument principle, conformal mapping — is developed where physics first demands it.

Chapter 9

Ordinary Differential Equations and Sturm–Liouville Theory

The linear boundary-value problems of physics — the vibrating string, the temperature of a bar, the electrostatic potential inside a cavity, the stationary states of a quantum particle — all reduce, once the partial differential equation is separated into ordinary ones (Chapter 10), to a single template: a linear second-order ordinary differential equation containing an eigenvalue parameter and a weight function. This is the Sturm–Liouville problem. Its power lies in what all its instances share: the solutions belonging to distinct eigenvalues are orthogonal with respect to the weight, and the full solution set is a basis in which arbitrary admissible functions can be expanded. The present chapter sets up the language of orthonormal function systems, defines the Sturm–Liouville operator, derives the semiclassical approximation that serves every potential for which no closed-form solution exists, and then works through the classical families — Helmholtz (Fourier), Bessel and its modified companion, Legendre and its associated equation with the spherical harmonics, Hermite, Laguerre and Mathieu — recording for each the identification of the Sturm–Liouville data, the solutions, and the orthogonality relations given in the source.

Around that core sit the parts of the general theory of ordinary differential equations that the physical parts of this treatise consume constantly and that no other chapter owns: existence, uniqueness and continuous dependence; the elementary first-order equations solvable by an integrating factor or by separation; linear systems with constant coefficients, the matrix exponential, and the reading of stability off the eigenvalues; Floquet's theorem for periodic coefficients, which is what makes parametric resonance a statement rather than an estimate; and the elliptic integrals and Jacobi elliptic functions, which are what a conservative system with one degree of freedom reduces to when the small-amplitude approximation is dropped.

9.1 Ordinary differential equations

An *ordinary differential equation* (ODE) relates a function of a single real variable to its derivatives; every equation treated in this chapter is an ODE that is linear and of second order. The source outline opens the subject with a section of general definitions — order, linearity, homogeneity, the structure of solution spaces — but that section contains only its heading, so the present section supplies it. None of what follows is special to the Sturm–Liouville problem. It is what entitles every closed form recorded later in this chapter to be called *the* general solution rather than merely *a* pair of solutions: a linear homogeneous equation of order n has a solution set of dimension exactly n, so two solutions of a second-order equation with non-vanishing Wronskian already exhaust it.

9.1.1 Order, linearity, homogeneity

Definition 9.1 (Ordinary differential equation)**.** Let $I \subseteq \mathbb{R}$ be an open interval. An *ordinary differential equation of order* n on I is a relation

$$F\left(x,\ y(x),\ \frac{\mathrm{d}y}{\mathrm{d}x}(x),\ \ldots,\ \frac{\mathrm{d}^n y}{\mathrm{d}x^n}(x)\right) = 0 \tag{9.1}$$

between a function y of the single variable $x \in I$ and its derivatives up to the n-th, the n-th being present. A *solution* is a function y, n times differentiable on I, satisfying Equation (9.1) at every point of I.

Definition 9.2 (Linear equation; homogeneity)**.** Equation (9.1) is *linear* when it can be written

$$L[y] = b\,, \qquad L[y](x) = \sum_{i=0}^{n} a_i(x)\,\frac{\mathrm{d}^i y}{\mathrm{d}x^i}(x)\,, \tag{9.2}$$

with the coefficients $a_0, \ldots, a_n$ and the inhomogeneity b continuous on I and a_n nowhere zero there. It is *homogeneous* when b vanishes identically. The map L is linear: $L[\alpha u + \beta v] = \alpha L[u] + \beta L[v]$ for all real constants α, β.

Rests on Definition 9.1.

Remark 9.3 (Normal form and first-order systems). Dividing Equation (9.2) by a_n — legitimate because a_n does not vanish — solves the equation for its highest derivative. Any equation so solved, linear or not, becomes a first-order system for the vector $\boldsymbol{Y} = (y, y', \ldots, y^{(n-1)})$ of the function and its first $n-1$ derivatives,

$$\frac{\mathrm{d}\boldsymbol{Y}}{\mathrm{d}x} = \boldsymbol{f}\big(x, \boldsymbol{Y}(x)\big)\,, \qquad \boldsymbol{f} = \big(Y_2,\ Y_3,\ \ldots,\ Y_n,\ f(x, Y_1, \ldots, Y_n)\big)\,, \tag{9.3}$$

where $y^{(n)} = f\big(x, y, y', \ldots, y^{(n-1)}\big)$ is the equation in normal form. Solutions of the system and of the scalar equation correspond one to one, so every statement below is made for Equation (9.3) and read back for the scalar equation.

9.1.2 Existence and uniqueness

Definition 9.4 (Lipschitz condition)**.** Let $R = \{(x, \boldsymbol{Y}) : |x - x_0| \le a\,,\ \|\boldsymbol{Y} - \boldsymbol{Y}_0\| \le b\}$. A map $\boldsymbol{f} : R \longrightarrow \mathbb{R}^n$ is *Lipschitz in its second argument*, with constant L, if

$$\|\boldsymbol{f}(x, \boldsymbol{U}) - \boldsymbol{f}(x, \boldsymbol{V})\| \le L\,\|\boldsymbol{U} - \boldsymbol{V}\| \tag{9.4}$$

for every $(x, \boldsymbol{U})$ and $(x, \boldsymbol{V})$ in R.

Remark 9.5 (Continuity is not enough). A continuously differentiable $\boldsymbol{f}$ satisfies Equation (9.4) on the compact box R, with L the maximum there of the norm of its derivative with respect to $\boldsymbol{Y}$, by the mean value theorem Theorem 7.35. Mere continuity does not suffice for uniqueness: $\mathrm{d}y/\mathrm{d}x = 3\,y^{2/3}$ with $y(0) = 0$ is solved both by $y \equiv 0$ and by $y = x^3$, and $3y^{2/3}$ is continuous everywhere but Lipschitz nowhere near $y = 0$.

Lemma 9.6 (Weierstrass M-test; uniform limits are continuous)**.** *Let $S \subseteq \mathbb{R}$ and let $\boldsymbol{u}_j : S \longrightarrow \mathbb{R}^n$, $j \ge 0$, satisfy $\|\boldsymbol{u}_j(x)\| \le M_j$ for every $x \in S$, with $\sum_{j\ge 0} M_j$ convergent. Then $\sum_{j\ge 0} \boldsymbol{u}_j(x)$ converges absolutely at every $x \in S$, and its partial sums converge to the sum $\boldsymbol{U}$* uniformly on S*: for every $\varepsilon > 0$ there is an N, depending on ε alone and not on x, such that*

$$\left\| \boldsymbol{U}(x) - \sum_{j=0}^{k} \boldsymbol{u}_j(x) \right\| < \varepsilon \qquad \text{for all } k \ge N \text{ and all } x \in S\,. \tag{9.5}$$

If in addition every $\boldsymbol{u}_j$ is continuous on S, then so is $\boldsymbol{U}$.

Rests on Proposition 7.47 and Definition 7.20.

Derivation. *Derives Lemma 9.6.*

Fix $x \in S$. Each component obeys $\left|u_j^i(x)\right| \le \|\boldsymbol{u}_j(x)\| \le M_j$, so $\sum_j \left|u_j^i(x)\right|$ converges by comparison with $\sum_j M_j$ and the component series converges, both by Proposition 7.47. Call the limit $\boldsymbol{U}(x)$.

For the uniformity, subtract the k-th partial sum: for $k \ge N$,

$$\left\| \boldsymbol{U}(x) - \sum_{j=0}^{k} \boldsymbol{u}_j(x) \right\| = \left\| \sum_{j>k} \boldsymbol{u}_j(x) \right\| \le \sum_{j>k} M_j \le \sum_{j>N} M_j \,,$$

and the right-hand side is the tail of a convergent series of numbers, hence below any prescribed ε once N is large enough. It contains no x, which is exactly Equation (9.5).

Suppose now every $\boldsymbol{u}_j$ is continuous and fix $x_0 \in S$ and $\varepsilon > 0$. Choose N with $\sum_{j>N} M_j < \varepsilon/3$ and write $\boldsymbol{S}_N = \sum_{j=0}^{N} \boldsymbol{u}_j$, a finite sum of continuous maps and therefore continuous (Definition 7.20): there is a $\delta > 0$ with $\|\boldsymbol{S}_N(x) - \boldsymbol{S}_N(x_0)\| < \varepsilon/3$ for every $x \in S$ with $|x - x_0| < \delta$. For such an x,

$$\|\boldsymbol{U}(x) - \boldsymbol{U}(x_0)\| \le \|\boldsymbol{U}(x) - \boldsymbol{S}_N(x)\| + \|\boldsymbol{S}_N(x) - \boldsymbol{S}_N(x_0)\| + \|\boldsymbol{S}_N(x_0) - \boldsymbol{U}(x_0)\| < \varepsilon \,,$$

the outer two terms being bounded by $\sum_{j>N} M_j$ by the estimate just proved. Hence $\boldsymbol{U}$ is continuous at x_0. □

Remark 9.7 (Why the numerical comparison test is not enough). Proposition 7.47 is a statement about series of *numbers*: it decides convergence at each fixed x separately and says nothing about how fast. What Lemma 9.6 adds is the single N that serves every x at once, and it is that uniformity — not the pointwise convergence — which carries continuity across to the limit and licenses the exchange of a limit with an integral. Both halves are stated and proved here because Theorem 9.8 needs them in exactly this form; the same two facts are what the Fourier arguments of Chapter 17 invoke under the name of the M-test.

Theorem 9.8 (Picard–Lindelöf)**.** *Let $\boldsymbol{f}$ be continuous on the box R of Definition 9.4, bounded there by $\|\boldsymbol{f}\| \le M$, and Lipschitz in its second argument with constant L. Then the initial value problem*

$$\frac{\mathrm{d}\boldsymbol{Y}}{\mathrm{d}x} = \boldsymbol{f}(x, \boldsymbol{Y}(x)) \,, \qquad \boldsymbol{Y}(x_0) = \boldsymbol{Y}_0 \,, \tag{9.6}$$

has exactly one solution on $|x - x_0| \le h$, where $h = \min\{a,\ b/M\}$.

Rests on Definition 9.4, Equation (9.3), and Lemma 9.6.

Derivation. *Derives Theorem 9.8.*

A continuous $\boldsymbol{Y}$ solves Equation (9.6) if and only if it solves the integral equation

$$\boldsymbol{Y}(x) = \boldsymbol{Y}_0 + \int_{x_0}^{x} \boldsymbol{f}(s, \boldsymbol{Y}(s))\, \mathrm{d}s \,, \tag{9.7}$$

by the fundamental theorem of calculus (Theorems 7.42 and 7.43): differentiating Equation (9.7) returns Equation (9.6) and evaluating it at x_0 returns the initial value, while integrating Equation (9.6) from x_0 to x returns Equation (9.7).

Define the *Picard iterates* by $\boldsymbol{Y}_0(x) = \boldsymbol{Y}_0$ and

$$\boldsymbol{Y}_{k+1}(x) = \boldsymbol{Y}_0 + \int_{x_0}^{x} \boldsymbol{f}(s, \boldsymbol{Y}_k(s))\, \mathrm{d}s \,. \tag{9.8}$$

Every iterate stays inside the box on $|x - x_0| \le h$: if $\|\boldsymbol{Y}_k(s) - \boldsymbol{Y}_0\| \le b$ there, the integrand of Equation (9.8) is defined and bounded by M, so $\|\boldsymbol{Y}_{k+1}(x) - \boldsymbol{Y}_0\| \le M\,|x - x_0| \le Mh \le b$; and the hypothesis holds trivially for $k = 0$.

The successive differences obey

$$\|\boldsymbol{Y}_{k+1}(x)-\boldsymbol{Y}_k(x)\| \le \frac{M\,L^k\,|x-x_0|^{k+1}}{(k+1)!}\,, \qquad |x-x_0|\le h\,. \tag{9.9}$$

For $k=0$ this is the bound just used. Assuming it at $k-1$ and applying Equation (9.4),

$$\begin{aligned}\|\boldsymbol{Y}_{k+1}(x)-\boldsymbol{Y}_k(x)\| &\le \left|\int_{x_0}^{x} \|\boldsymbol{f}(s,\boldsymbol{Y}_k(s))-\boldsymbol{f}(s,\boldsymbol{Y}_{k-1}(s))\|\,\mathrm{d}s\right| \\ &\le L\left|\int_{x_0}^{x} \|\boldsymbol{Y}_k(s)-\boldsymbol{Y}_{k-1}(s)\|\,\mathrm{d}s\right| \le L\left|\int_{x_0}^{x} \frac{M\,L^{k-1}\,|s-x_0|^{k}}{k!}\,\mathrm{d}s\right| \\ &= \frac{M\,L^k\,|x-x_0|^{k+1}}{(k+1)!}\,,\end{aligned}$$

which is Equation (9.9).

Write $\boldsymbol{Y}_k=\boldsymbol{Y}_0+\sum_{j=0}^{k-1}(\boldsymbol{Y}_{j+1}-\boldsymbol{Y}_j)$. By Equation (9.9) the terms of that series are bounded in supremum norm on $|x-x_0|\le h$ by $ML^jh^{j+1}/(j+1)!$, whose sum is $M(\mathrm{e}^{Lh}-1)/L$, a finite number; these are the constants M_j of Lemma 9.6, so the series converges absolutely and uniformly on $|x-x_0|\le h$. Its sum $\boldsymbol{Y}$ is therefore continuous, by the last clause of the same lemma, and obeys $\|\boldsymbol{Y}-\boldsymbol{Y}_0\|\le b$. Passing to the limit under the integral sign in Equation (9.8) is legitimate because Equation (9.4) gives $\|\boldsymbol{f}(s,\boldsymbol{Y}_k)-\boldsymbol{f}(s,\boldsymbol{Y})\|\le L\,\|\boldsymbol{Y}_k-\boldsymbol{Y}\|$, which tends to zero uniformly by Equation (9.5), while the error committed on the integral is at most h times that supremum. Hence $\boldsymbol{Y}$ solves Equation (9.7).

For uniqueness let $\boldsymbol{Y}$ and $\tilde{\boldsymbol{Y}}$ both solve Equation (9.7) on $|x-x_0|\le h$ and put $\delta=\sup\left\|\boldsymbol{Y}-\tilde{\boldsymbol{Y}}\right\|$, finite because both stay in the box. Subtracting the two integral equations and using Equation (9.4) gives $\left\|\boldsymbol{Y}(x)-\tilde{\boldsymbol{Y}}(x)\right\|\le L\,\delta\,|x-x_0|$; feeding that bound back into the same inequality gives $L^2\delta\,|x-x_0|^2/2$, and after k rounds

$$\left\|\boldsymbol{Y}(x)-\tilde{\boldsymbol{Y}}(x)\right\| \le \frac{\delta\,(Lh)^k}{k!}\,,$$

which tends to zero as $k\to\infty$ because the exponential series Equation (7.31) converges. Hence $\delta=0$. □

Corollary 9.9 (Linear equations: existence on the whole interval)**.** *Let A be a continuous matrix-valued function and $\boldsymbol{g}$ a continuous vector-valued function on a compact interval J containing x_0. Then $\mathrm{d}\boldsymbol{Y}/\mathrm{d}x=A(x)\,\boldsymbol{Y}+\boldsymbol{g}(x)$ with $\boldsymbol{Y}(x_0)=\boldsymbol{Y}_0$ has exactly one solution, and it is defined on the whole of J. In particular the scalar linear equation Equation (9.2) has exactly one solution on I for each choice of the initial data $y(x_0)$, $y'(x_0)$, …, $y^{(n-1)}(x_0)$.*

Rests on Theorem 9.8, Remark 9.3, and Definition 9.2.

Derivation. Derives Corollary 9.9.

Here $\boldsymbol{f}(x,\boldsymbol{Y})=A(x)\boldsymbol{Y}+\boldsymbol{g}(x)$ is defined for every $\boldsymbol{Y}$, so the box of Definition 9.4 may be taken to be $J\times\mathbb{R}^n$ and Equation (9.4) holds globally with $L=\max_J\|A\|$, the maximum existing because J is compact and A continuous (Theorem 7.24). The construction in the proof of Theorem 9.8 then runs on all of J without the restriction $h\le b/M$, which was needed only to keep the iterates inside a bounded box: the estimate Equation (9.9) holds on J with $M=\max_J\|\boldsymbol{f}(\cdot,\boldsymbol{Y}_0)\|$, and the comparison series is again the exponential one. Uniqueness is unchanged. The scalar statement follows through Equation (9.3), whose coefficients $-a_i/a_n$ and inhomogeneity b/a_n are continuous precisely because a_n does not vanish. □

Theorem 9.8 says that the solution through a point exists and is unique. It says nothing about how two solutions through *nearby* points separate, and that quantitative question is the one every stability argument asks. The answer is a single inequality, and it is worth isolating because it is used far more often than it is named.

Lemma 9.10 (Grönwall's inequality)**.** *Let u be continuous and non-negative on $[x_0, x_1]$ and suppose that*

$$u(x) \le \alpha + \beta \int_{x_0}^{x} u(s)\,\mathrm{d}s \qquad \text{for every } x \in [x_0, x_1]\,, \tag{9.10}$$

with constants $\alpha \ge 0$ and $\beta > 0$. Then

$$u(x) \le \alpha\, \mathrm{e}^{\beta\,(x - x_0)} \qquad \text{for every } x \in [x_0, x_1]\,. \tag{9.11}$$

Rests on Theorem 7.42 and Corollary 7.36.

Derivation. *Derives Lemma 9.10.*

Write $U(x) = \alpha + \beta \int_{x_0}^{x} u(s)\,\mathrm{d}s$. Since u is continuous, U is differentiable with $U' = \beta u$ (Theorem 7.42), and $U(x_0) = \alpha$. The hypothesis Equation (9.10) is exactly $u \le U$, so $U' = \beta u \le \beta U$ and

$$\frac{\mathrm{d}}{\mathrm{d}x}\Big[\mathrm{e}^{-\beta(x-x_0)}\,U(x)\Big] = \mathrm{e}^{-\beta(x-x_0)}\big[U'(x) - \beta\,U(x)\big] \le 0\,.$$

A function whose derivative is nowhere positive does not increase (Corollary 7.36), so $\mathrm{e}^{-\beta(x-x_0)}U(x) \le U(x_0) = \alpha$ throughout, and $u(x) \le U(x) \le \alpha\,\mathrm{e}^{\beta(x-x_0)}$. □

Proposition 9.11 (Continuous dependence on the initial data)**.** *Let $\boldsymbol{f}$ satisfy the hypotheses of Theorem 9.8, with Lipschitz constant L, and let $\boldsymbol{Y}$ and $\tilde{\boldsymbol{Y}}$ solve $\mathrm{d}\boldsymbol{Y}/\mathrm{d}x = \boldsymbol{f}(x, \boldsymbol{Y})$ on $[x_0, x_1]$ with initial values $\boldsymbol{Y}_0$ and $\tilde{\boldsymbol{Y}}_0$, both trajectories remaining in the box R. Then*

$$\big\|\boldsymbol{Y}(x) - \tilde{\boldsymbol{Y}}(x)\big\| \le \big\|\boldsymbol{Y}_0 - \tilde{\boldsymbol{Y}}_0\big\|\ \mathrm{e}^{L\,(x-x_0)}\,, \qquad x \in [x_0, x_1]\,. \tag{9.12}$$

In particular the value of the solution at x depends continuously on its initial value, and two solutions that agree at one point agree throughout.

Rests on Theorem 9.8, Lemma 9.10, and Equation (9.4).

Derivation. *Derives Proposition 9.11.*

Both solutions satisfy the integral equation Equation (9.7) with their own initial values. Subtracting and using Equation (9.4),

$$\big\|\boldsymbol{Y}(x) - \tilde{\boldsymbol{Y}}(x)\big\| \le \big\|\boldsymbol{Y}_0 - \tilde{\boldsymbol{Y}}_0\big\| + L \int_{x_0}^{x} \big\|\boldsymbol{Y}(s) - \tilde{\boldsymbol{Y}}(s)\big\|\,\mathrm{d}s\,.$$

This is Equation (9.10) for the continuous non-negative function $u = \big\|\boldsymbol{Y} - \tilde{\boldsymbol{Y}}\big\|$ with $\alpha = \big\|\boldsymbol{Y}_0 - \tilde{\boldsymbol{Y}}_0\big\|$ and $\beta = L$, so Lemma 9.10 gives Equation (9.12). Taking $\alpha = 0$ recovers the uniqueness half of Theorem 9.8. □

Remark 9.12 (What the exponential in the bound is, and is not). Equation (9.12) is an upper bound and nothing more. It licenses the statement that neighbouring solutions cannot separate faster than e^{Lx} — which is what makes a numerical trajectory meaningful over a finite time, and what turns a linearised decay estimate into a genuine statement about the nonlinear flow in Proposition 9.31. It does *not* assert that any pair of solutions actually separates at that rate. A system in which some pairs do separate exponentially, and in which the rate is a property of the flow rather than of a Lipschitz constant, is a different subject altogether; the exponents that measure it are defined in Chapter 32, and Grönwall's inequality is the tool that shows the linear estimate there survives the neglected quadratic terms.

9.1.3 The solution space and the Wronskian

Theorem 9.13 (Dimension of the solution space)**.** *The solutions on I of the homogeneous linear equation $L[y] = 0$ of Equation (9.2) form a real vector space of dimension exactly n. The solutions of the inhomogeneous equation $L[y] = b$ form the affine set $y_p + \ker L$, with y_p any one solution of it.*

Rests on Definition 9.2 and Corollary 9.9.

Derivation. *Derives Theorem 9.13.*

Linearity of L makes $\ker L$ a vector space and makes the difference of two solutions of $L[y] = b$ a solution of $L[y] = 0$, which is the second statement. Fix $x_0 \in I$ and consider the evaluation map

$$E : \ker L \longrightarrow \mathbb{R}^n , \qquad E(y) = \big(y(x_0),\, y'(x_0),\, \ldots,\, y^{(n-1)}(x_0)\big) ,$$

which is linear. It is surjective because Corollary 9.9 produces a solution for every prescribed initial vector, and injective because that solution is unique: a y in $\ker L$ with $E(y) = \mathbf{0}$ agrees with the zero solution. An isomorphism onto $\mathbb{R}^n$ forces $\dim \ker L = n$. □

Definition 9.14 (Wronskian)**.** The *Wronskian* of n functions $y_1, \ldots, y_n$, each $n-1$ times differentiable on I, is the determinant

$$W(x) = \det \begin{pmatrix} y_1 & y_2 & \cdots & y_n \\ y_1' & y_2' & \cdots & y_n' \\ \vdots & \vdots & & \vdots \\ y_1^{(n-1)} & y_2^{(n-1)} & \cdots & y_n^{(n-1)} \end{pmatrix} (x) . \tag{9.13}$$

Proposition 9.15 (Abel's identity)**.** *If $y_1, \ldots, y_n$ all solve $L[y] = 0$, their Wronskian satisfies*

$$\frac{\mathrm{d}W}{\mathrm{d}x}(x) = -\frac{a_{n-1}(x)}{a_n(x)}\, W(x) , \qquad W(x) = W(x_0)\, \exp\left[-\int_{x_0}^{x} \frac{a_{n-1}(s)}{a_n(s)}\, \mathrm{d}s \right] . \tag{9.14}$$

In particular W either vanishes identically on I or vanishes nowhere.

Rests on Definitions 9.2 and 9.14.

Derivation. *Derives Proposition 9.15.*

A determinant is multilinear in its rows, so $\mathrm{d}W/\mathrm{d}x$ is a sum of n determinants, the i-th obtained by differentiating the i-th row of Equation (9.13). For $i < n$ the differentiated row becomes equal to the row below it, so that determinant has two equal rows and vanishes. Only the last term survives, and in it the bottom row is $(y_1^{(n)}, \ldots, y_n^{(n)})$. Each solution obeys $y_j^{(n)} = -\sum_{i<n} (a_i/a_n)\, y_j^{(i)}$, so that row is a linear combination of the rows of Equation (9.13); subtracting the multiples of the rows carrying $y^{(i)}$ for $i \le n-2$, which changes no determinant, leaves $-a_{n-1}/a_n$ times the row $(y_1^{(n-1)}, \ldots, y_n^{(n-1)})$, that is $-(a_{n-1}/a_n)\, W$. The resulting first-order linear equation integrates to the stated exponential, and an exponential never vanishes, so $W(x)$ and $W(x_0)$ vanish together. □

Proposition 9.16 (Wronskian test for independence)**.** *Let $y_1, \ldots, y_n$ solve $L[y] = 0$ on I. They are linearly independent if and only if $W(x) \neq 0$ at one — equivalently at every — point of I, and in that case every solution of $L[y] = 0$ is a unique linear combination of them.*

Rests on Theorem 9.13 and Proposition 9.15.

Derivation. *Derives Proposition 9.16.*

If $\sum_j c_j y_j$ vanishes identically with the c_j not all zero, differentiating $n-1$ times shows that the columns of Equation (9.13) are linearly dependent at every x, so W vanishes

identically. Conversely, if $W(x_0) = 0$ there is a non-zero vector $\boldsymbol{c}$ annihilating the columns at x_0, and $y = \sum_j c_j y_j$ is then a solution with $E(y) = \mathbf{0}$ in the notation of the proof of Theorem 9.13; by uniqueness y vanishes identically, so the y_j are dependent. Proposition 9.15 makes "at one point" and "at every point" the same statement. Finally n independent solutions inside an n-dimensional space (Theorem 9.13) are a basis of it. □

Remark 9.17 (What this buys the rest of the chapter). Every equation treated below is linear, homogeneous and of second order, so its solution space is two-dimensional and exhibiting two solutions with non-vanishing Wronskian settles the general solution completely. That is the status of Equation (9.107), Equation (9.126), Equation (9.147) and Equation (9.158). The Sturm–Liouville equation Equation (9.55) is in normal form wherever p does not vanish; the endpoints at which p *does* vanish — $x = 0$ for Bessel, $x = \pm 1$ for Legendre and its associated equation — are exactly the singular points at which Corollary 9.9 does not apply, and they are also the points at which one of the two solutions ceases to be bounded.

9.1.4 Elementary first-order equations

Two first-order equations can be integrated outright, and between them they cover almost every first-order equation the physical parts of this treatise meet: a decay law, the charging of a capacitor, a relaxation towards equilibrium, a body falling against linear or quadratic drag. Theorem 9.8 already guarantees that each has exactly one solution; what follows produces that solution in closed form, so a later chapter can quote a formula instead of exhibiting an answer and verifying it.

Proposition 9.18 (The linear equation: integrating factor)**.** *Let p and q be continuous on an interval I and let $x_0 \in I$. Put $P(x) = \int_{x_0}^{x} p(s)\,\mathrm{d}s$. Then the solutions on I of*

$$\frac{\mathrm{d}y}{\mathrm{d}x}(x) + p(x)\,y(x) = q(x) \tag{9.15}$$

are exactly the functions

$$y(x) = \mathrm{e}^{-P(x)}\left[\,C + \int_{x_0}^{x} q(s)\,\mathrm{e}^{P(s)}\,\mathrm{d}s\right], \qquad C \in \mathbb{R}\,, \tag{9.16}$$

one for each constant $C = y(x_0)$. The factor e^{P}, which turns the left-hand side of Equation (9.15) into an exact derivative, is the integrating factor.

Rests on Corollary 9.9 and Theorems 7.42 and 7.43.

Derivation. *Derives Proposition 9.18.*

P is differentiable with $P' = p$ (Theorem 7.42), so for any differentiable y,

$$\frac{\mathrm{d}}{\mathrm{d}x}\Big[\mathrm{e}^{P(x)}y(x)\Big] = \mathrm{e}^{P(x)}\left[\frac{\mathrm{d}y}{\mathrm{d}x}(x) + p(x)\,y(x)\right], \tag{9.17}$$

and e^{P} never vanishes. Hence y solves Equation (9.15) if and only if $(\mathrm{e}^{P}y)' = q\,\mathrm{e}^{P}$, which by Theorem 7.43 holds if and only if $\mathrm{e}^{P(x)}y(x) - \mathrm{e}^{P(x_0)}y(x_0) = \int_{x_0}^{x} q\,\mathrm{e}^{P}$; since $P(x_0) = 0$ this is Equation (9.16) with $C = y(x_0)$. Every C occurs and no solution is missed, because Corollary 9.9 gives exactly one solution on I for each initial value. □

Proposition 9.19 (Separation of variables)**.** *Let g be continuous on an interval I and h continuous and nowhere zero on an interval J, let $x_0 \in I$ and let y_0 be an interior point of J. Then, near x_0, the initial-value problem*

$$\frac{\mathrm{d}y}{\mathrm{d}x}(x) = g(x)\,h(y(x))\,, \qquad y(x_0) = y_0\,, \tag{9.18}$$

has exactly one solution with values in J, and it is obtained by inverting the quadrature

$$\int_{y_0}^{y} \frac{\mathrm{d}v}{h(v)} = \int_{x_0}^{x} g(s)\,\mathrm{d}s\,. \tag{9.19}$$

For the autonomous equation $\mathrm{d}y/\mathrm{d}x = F(y)$ with F continuous and non-vanishing this reads $x - x_0 = \int_{y_0}^{y} \mathrm{d}v/F(v)$: the independent variable is recovered from the dependent one by a single integral, and inverting that integral is the whole of the solution.

Rests on Proposition 7.32, Theorem 7.42, and Corollary 7.36.

Derivation. Derives Proposition 9.19.

Set $H(y) = \int_{y_0}^{y} \mathrm{d}v/h(v)$ on J and $G(x) = \int_{x_0}^{x} g(s)\,\mathrm{d}s$ on I; both are differentiable, with $H' = 1/h$ and $G' = g$ (Theorem 7.42). Since h is continuous and nowhere zero on the interval J it has one sign there (Theorem 7.23), so H' has one sign and H is strictly monotone (Corollary 7.36); it therefore maps J bijectively onto an interval $H(J)$ containing $0 = H(y_0)$, and its inverse is differentiable with $(H^{-1})' = 1/H' \circ H^{-1} = h \circ H^{-1}$ (Proposition 7.32).

Let $I_0 \subseteq I$ be a neighbourhood of x_0 on which $G(x) \in H(J)$; such a neighbourhood exists because G is continuous and $G(x_0) = H(y_0) = 0$ is an interior point of $H(J)$, H being a strictly monotone continuous map and y_0 interior to J. On I_0 define $y = H^{-1} \circ G$. It takes the value $H^{-1}(0) = y_0$ at x_0, and by the chain rule (Proposition 7.31)

$$\frac{\mathrm{d}y}{\mathrm{d}x}(x) = \big(H^{-1}\big)'\big(G(x)\big)\,G'(x) = h\big(y(x)\big)\,g(x)\,,$$

so it solves Equation (9.18). Conversely, if y is any solution with values in J, then $\mathrm{d}\big[H(y(x)) - G(x)\big]/\mathrm{d}x = y'/h(y) - g = 0$, so $H(y(x)) - G(x)$ is constant (Corollary 7.36) and equal to its value 0 at x_0; that is Equation (9.19), and H being injective determines y uniquely. □

Remark 9.20 (Where the hypotheses bite). Both hypotheses on h are doing work. If h vanishes at y_* then $y \equiv y_*$ is a constant solution, the integral $\int \mathrm{d}v/h$ diverges there or ceases to be defined, and uniqueness may fail — Remark 9.5 is exactly this situation with $h(y) = 3y^{2/3}$. Physically the zeros of h are the equilibria, and the divergence of the quadrature at one of them is the statement that the equilibrium is reached only asymptotically: a body falling against drag approaches its terminal speed and never attains it, and the integral $\int^{v_t} \mathrm{d}v/(1 - v/v_t)$ diverging is that fact in one line.

9.2 Linear systems with constant coefficients

The theory so far is scalar. A great deal of physics is not: Euler's equations for a rotating body, a chain of coupled oscillators, a chemical or nuclear decay chain and every linearisation of a nonlinear flow are first-order systems $\mathrm{d}\boldsymbol{Y}/\mathrm{d}x = A\boldsymbol{Y}$ with A a constant matrix. Corollary 9.9 already says such a system has exactly one solution for each initial vector; this section produces that solution, describes the shape of every solution, and reads the stability of the equilibrium $\boldsymbol{Y} = \boldsymbol{0}$ off the eigenvalues of A. Nothing here is deeper than the series for the exponential and the residue theorem of Chapter 8, but the conclusions are what every stability argument in the physical parts appeals to.

9.2.1 The matrix exponential

Definition 9.21 (Matrix exponential)**.** Let A be an $n \times n$ matrix with real or complex entries and let $\|\cdot\|$ be the matrix norm induced by a norm on the underlying vector space (Definition 5.18),

$$\|A\| = \sup_{\|\boldsymbol{v}\|=1} \|A\boldsymbol{v}\|\,, \qquad \text{so that} \qquad \|A\boldsymbol{v}\| \le \|A\|\,\|\boldsymbol{v}\| \quad \text{and} \quad \|AB\| \le \|A\|\,\|B\|\,. \tag{9.20}$$

The *matrix exponential* is

$$\mathrm{e}^{A} = \sum_{k=0}^{\infty} \frac{A^k}{k!}\,, \qquad A^0 = \mathbb{1}\,. \tag{9.21}$$

Rests on Definition 5.18 and Proposition 5.121.

Proposition 9.22 (The exponential and its derivative)**.** *The series Equation (9.21) converges absolutely for every A, and uniformly in x on every bounded interval when A is replaced by xA. The map $x \mapsto \mathrm{e}^{xA}$ is differentiable with*

$$\frac{\mathrm{d}}{\mathrm{d}x}\,\mathrm{e}^{xA} = A\,\mathrm{e}^{xA} = \mathrm{e}^{xA}A\,, \qquad \mathrm{e}^{0\cdot A} = \mathbb{1}\,, \tag{9.22}$$

and $\mathrm{e}^{(x+y)A} = \mathrm{e}^{xA}\mathrm{e}^{yA}$; in particular e^{xA} is invertible, with inverse e^{-xA}. If $AB = BA$ then $\mathrm{e}^{A+B} = \mathrm{e}^{A}\mathrm{e}^{B}$.

Rests on Definition 9.21, Lemma 9.6, and Theorem 7.51.

Derivation. Derives Proposition 9.22.

Submultiplicativity gives $\left\|x^k A^k/k!\right\| \le |x|^k \,\|A\|^k/k! \le R^k \,\|A\|^k/k!$ on $|x| \le R$, and those bounds sum to $\mathrm{e}^{R\|A\|}$; by Lemma 9.6 the series converges absolutely and uniformly there, entry by entry. Each entry is a power series in x with infinite radius of convergence, so it may be differentiated term by term (Theorem 7.51), and $\sum_{k\ge1} k x^{k-1}A^k/k! = A\sum_{k\ge0} x^k A^k/k!$, the factor A coming out on either side since it commutes with its own powers. That is Equation (9.22).

For the addition law with $AB = BA$, expand the Cauchy product of the two absolutely convergent series (Proposition 7.49) and collect the terms of total degree m: $\sum_{j+k=m} A^j B^k/(j!\,k!) = (A+B)^m/m!$, the binomial theorem being available precisely because A and B commute. Taking $B = yA$, which commutes with A, gives $\mathrm{e}^{(x+y)A} = \mathrm{e}^{xA}\mathrm{e}^{yA}$, and $y = -x$ identifies the inverse. □

Theorem 9.23 (Solution of a constant-coefficient system)**.** *Let A be a constant $n \times n$ matrix. The initial-value problem*

$$\frac{\mathrm{d}\boldsymbol{Y}}{\mathrm{d}x}(x) = A\,\boldsymbol{Y}(x)\,, \qquad \boldsymbol{Y}(x_0) = \boldsymbol{Y}_0\,, \tag{9.23}$$

has exactly one solution, defined for every $x \in \mathbb{R}$, namely

$$\boldsymbol{Y}(x) = \mathrm{e}^{(x-x_0)A}\,\boldsymbol{Y}_0\,. \tag{9.24}$$

The inhomogeneous system $\boldsymbol{Y}' = A\boldsymbol{Y} + \boldsymbol{g}(x)$ with $\boldsymbol{g}$ continuous has the unique solution

$$\boldsymbol{Y}(x) = \mathrm{e}^{(x-x_0)A}\boldsymbol{Y}_0 + \int_{x_0}^{x} \mathrm{e}^{(x-s)A}\,\boldsymbol{g}(s)\,\mathrm{d}s\,. \tag{9.25}$$

Rests on Proposition 9.22 and Corollary 9.9.

Derivation. Derives Theorem 9.23.

Equation (9.22) makes Equation (9.24) a solution with the right initial value, and Corollary 9.9 says there is no other and that it lives on the whole line. For the inhomogeneous system, multiply by the integrating factor $\mathrm{e}^{-(x-x_0)A}$ — the matrix analogue of Equation (9.17), legitimate because e^{-xA} commutes with A — to get $\left(\mathrm{e}^{-(x-x_0)A}\boldsymbol{Y}\right)' = \mathrm{e}^{-(x-x_0)A}\boldsymbol{g}$, and integrate (Theorem 7.43). Multiplying back by $\mathrm{e}^{(x-x_0)A}$ and using the addition law gives Equation (9.25). □

Remark 9.24 (The same exponential, seen twice). Equation (9.22) says that $x \mapsto \mathrm{e}^{xA}$ is a one-parameter group of invertible matrices with generator A. Proposition 14.3 proves the converse — every differentiable one-parameter group of linear maps is the exponential of its generator — and is the entry point to the Lie theory of Chapter 14. The two statements are the same computation read in opposite directions, and it is worth seeing that the flow of a linear autonomous system and the exponential map of a matrix group are not two constructions but one.

9.2.2 The shape of the solutions

The exponential is a closed form but not yet an answer: one still wants to know what the solutions *look like*. They are combinations of $x^k \mathrm{e}^{\lambda x}$, and the eigenvalues of A supply the λ. The diagonalizable case is immediate; the general case is proved below without any appeal to a normal form for A, by reading the exponential off the resolvent with the residue theorem.

Corollary 9.25 (Normal modes of a diagonalizable system)**.** *Suppose A has n linearly independent eigenvectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ with eigenvalues $\lambda_1, \dots, \lambda_n$. Then every solution of Equation (9.23) is*

$$\boldsymbol{Y}(x) = \sum_{r=1}^{n} c_r \, \mathrm{e}^{\lambda_r (x - x_0)} \, \boldsymbol{v}_r \,, \tag{9.26}$$

the constants c_r being the components of $\boldsymbol{Y}_0$ in the eigenbasis. If A is real and $\lambda = \alpha + \mathrm{i}\beta$ is a non-real eigenvalue with eigenvector $\boldsymbol{v} = \boldsymbol{a} + \mathrm{i}\boldsymbol{b}$, then $\bar{\lambda}$ is an eigenvalue with eigenvector $\bar{\boldsymbol{v}}$, and the two real solutions carried by the pair are

$$\mathrm{e}^{\alpha x}\big[\boldsymbol{a}\cos\beta x - \boldsymbol{b}\sin\beta x\big] \,, \qquad \mathrm{e}^{\alpha x}\big[\boldsymbol{a}\sin\beta x + \boldsymbol{b}\cos\beta x\big] \,. \tag{9.27}$$

Rests on Theorem 9.23 and Definitions 5.14 and 5.65.

Derivation. *Derives Corollary 9.25.*

For an eigenvector, $A^k \boldsymbol{v}_r = \lambda_r^k \boldsymbol{v}_r$, so term-by-term application of Equation (9.21) gives $\mathrm{e}^{xA}\boldsymbol{v}_r = \mathrm{e}^{\lambda_r x}\boldsymbol{v}_r$. Expanding $\boldsymbol{Y}_0 = \sum_r c_r \boldsymbol{v}_r$ in the eigenbasis — a basis by hypothesis — and applying Equation (9.24) gives Equation (9.26). For the real form, conjugating $A\boldsymbol{v} = \lambda\boldsymbol{v}$ with A real gives $A\bar{\boldsymbol{v}} = \bar{\lambda}\bar{\boldsymbol{v}}$; the real and imaginary parts of the complex solution $\mathrm{e}^{\lambda x}\boldsymbol{v}$ are then themselves solutions, because A is real and the equation is linear, and expanding $\mathrm{e}^{(\alpha+\mathrm{i}\beta)x}(\boldsymbol{a}+\mathrm{i}\boldsymbol{b})$ with Euler's formula (Proposition 8.4) exhibits them as Equation (9.27). □

Theorem 9.26 (Structure of the solutions)**.** *Let A be an $n \times n$ complex matrix, let $\lambda_1, \dots, \lambda_p$ be the distinct roots of its characteristic polynomial $\chi(\lambda) = \det(\lambda\mathbb{1} - A)$, and let m_j be the multiplicity of λ_j as a root. Then every entry of e^{xA} has the form*

$$\sum_{j=1}^{p} \pi_j(x)\, \mathrm{e}^{\lambda_j x} \,, \qquad \deg \pi_j \le m_j - 1 \,, \tag{9.28}$$

and the same is true of every component of every solution of Equation (9.23). Every entry of the N-th power A^N likewise has the form $\sum_j \varpi_j(N)\, \lambda_j^N$ with $\deg \varpi_j \le m_j - 1$.

Rests on Theorems 5.67, 8.24 and 9.23.

Derivation. *Derives Theorem 9.26.*

The resolvent. For $|\lambda| > \|A\|$ the series $\sum_{k\ge 0} A^k \lambda^{-k-1}$ converges absolutely and uniformly, by Lemma 9.6 with $M_k = \|A\|^k\, |\lambda|^{-k-1}$ and the geometric series Proposition 7.46; multiplying it by $\lambda\mathbb{1} - A$ telescopes to $\mathbb{1}$, so its sum is the resolvent

$$R(\lambda) = (\lambda\mathbb{1} - A)^{-1} = \sum_{k=0}^{\infty} \frac{A^k}{\lambda^{k+1}} \,, \qquad |\lambda| > \|A\| \,. \tag{9.29}$$

Solving $(\lambda \mathbb{1} - A)\boldsymbol{X} = \boldsymbol{e}_j$ by Cramer's rule Equation (5.16) expresses the (i,j) entry of $R(\lambda)$ as a quotient $D_{ij}(\lambda)/\chi(\lambda)$ in which D_{ij} is the determinant of a matrix whose entries are affine in λ, hence a polynomial of degree at most $n-1$. So each entry of R is a rational function of λ whose poles lie among the roots of χ, that of λ_j being of order at most m_j.

The contour integral. Let Γ be the positively oriented circle $|\lambda| = \rho$ with $\rho > \|A\|$, so that Equation (9.29) holds on Γ and every root of χ lies inside (an eigenvalue λ_j has $|\lambda_j| \le \|A\|$, since an eigenvector would otherwise be lengthened by A beyond the bound its norm allows). Multiplying Equation (9.29) by $\mathrm{e}^{\lambda x}$, which is bounded on Γ, preserves the uniform convergence, so the integral may be taken term by term; and $(2\pi\mathrm{i})^{-1}\oint_\Gamma \mathrm{e}^{\lambda x}\lambda^{-k-1}\mathrm{d}\lambda = x^k/k!$, the k-th Taylor coefficient of $\mathrm{e}^{\lambda x}$ at $\lambda = 0$ (Theorem 8.17). Summing,

$$\mathrm{e}^{xA} = \frac{1}{2\pi\mathrm{i}} \oint_\Gamma \mathrm{e}^{\lambda x}\, R(\lambda)\, \mathrm{d}\lambda\,. \tag{9.30}$$

The residues. By the residue theorem Theorem 8.24 the right-hand side of Equation (9.30) is the sum over j of the residues of $\mathrm{e}^{\lambda x}R(\lambda)$ at λ_j. Near λ_j the entry $(\lambda-\lambda_j)^{m_j}R_{ik}(\lambda)$ is holomorphic and has a Taylor expansion $\sum_{l\ge 0} c_l(\lambda-\lambda_j)^l$ (Theorem 8.20), while $\mathrm{e}^{\lambda x} = \mathrm{e}^{\lambda_j x}\sum_{i\ge 0} x^i(\lambda-\lambda_j)^i/i!$. The residue is the coefficient of $(\lambda-\lambda_j)^{-1}$ in the product (Theorem 8.21 and Definition 8.22), that is

$$\mathrm{e}^{\lambda_j x}\sum_{i=0}^{m_j-1} c_{m_j-1-i}\,\frac{x^i}{i!}\,,$$

a polynomial in x of degree at most m_j-1 times $\mathrm{e}^{\lambda_j x}$. Summing over j gives Equation (9.28), and Equation (9.24) carries it to every solution. Replacing $\mathrm{e}^{\lambda x}$ throughout by λ^N — also holomorphic, with $(2\pi\mathrm{i})^{-1}\oint \lambda^{N-k-1}\mathrm{d}\lambda = \delta_{kN}$ — gives A^N in place of e^{xA} and the stated form of its entries. □

Corollary 9.27 (Linear stability)**.** *Write $\sigma(A) = \max_j \Re\lambda_j$. For every $\sigma > \sigma(A)$ there is a constant C with*

$$\left\|\mathrm{e}^{xA}\right\| \le C\,\mathrm{e}^{\sigma x}\,, \qquad x \ge 0\,. \tag{9.31}$$

Consequently every solution of Equation (9.23) tends to $\mathbf{0}$ as $x \to +\infty$ if and only if every eigenvalue of A has strictly negative real part, and some solution is unbounded if some eigenvalue has strictly positive real part. If in addition A has n independent eigenvectors, every solution stays bounded on $x \ge 0$ whenever no eigenvalue has positive real part. Similarly $A^N \to 0$ as $N \to \infty$ if and only if every eigenvalue satisfies $|\lambda_j| < 1$, and the powers A^N stay bounded whenever every $|\lambda_j| \le 1$ and A has n independent eigenvectors.

Rests on Theorem 9.26 and Corollary 9.25.

Derivation. Derives Corollary 9.27.

By Equation (9.28) each entry of e^{xA} is a finite sum of terms $x^k\mathrm{e}^{\lambda_j x}$, whose modulus is $x^k\mathrm{e}^{x\Re\lambda_j}$. For $\sigma > \sigma(A)$ the quotient $x^k\mathrm{e}^{x\Re\lambda_j}/\mathrm{e}^{\sigma x} = x^k\mathrm{e}^{-(\sigma-\Re\lambda_j)x}$ tends to zero as $x \to +\infty$ and is continuous on $[0,\infty)$, hence bounded there; summing the finitely many bounds gives Equation (9.31). If every $\Re\lambda_j < 0$, choose $\sigma < 0$ and every solution decays exponentially. Conversely an eigenvector for λ furnishes the solution $\mathrm{e}^{\lambda x}\boldsymbol{v}$ of Corollary 9.25, whose norm is $\mathrm{e}^{x\Re\lambda}\,\|\boldsymbol{v}\|$: it does not tend to zero when $\Re\lambda \ge 0$ and is unbounded when $\Re\lambda > 0$. When A has an eigenbasis, Equation (9.26) writes every solution as a fixed combination of the $\mathrm{e}^{\lambda_r x}\boldsymbol{v}_r$, each of norm $\mathrm{e}^{x\Re\lambda_r}\,\|\boldsymbol{v}_r\| \le \|\boldsymbol{v}_r\|$ for $x \ge 0$ when $\Re\lambda_r \le 0$; the triangle inequality bounds the sum. The statements about A^N are the same two arguments on the second half of Theorem 9.26, with $|\lambda_j|^N$ in place of $\mathrm{e}^{x\Re\lambda_j}$ and, in the diagonalizable case, with $A^N\boldsymbol{v}_r = \lambda_r^N\boldsymbol{v}_r$. □

Proposition 9.28 (The scalar constant-coefficient equation)**.** *Let $a_0, \dots, a_n$ be constants with $a_n \neq 0$, let*

$$P(\lambda) = \sum_{i=0}^{n} a_i \lambda^i \tag{9.32}$$

be the characteristic polynomial *of $L[y] = \sum_i a_i \,\mathrm{d}^i y/\mathrm{d}x^i$, and let $\lambda_1, \dots, \lambda_p$ be its distinct roots, with multiplicities $m_1, \dots, m_p$. Then the n functions*

$$x^{\kappa}\, \mathrm{e}^{\lambda_j x}, \qquad 1 \leq j \leq p, \quad 0 \leq \kappa \leq m_j - 1, \tag{9.33}$$

are a basis of the solution space of $L[y] = 0$. In particular, for $n = 2$ the general solution is $c_1 \mathrm{e}^{\lambda_1 x} + c_2 \mathrm{e}^{\lambda_2 x}$ when the two roots differ and $(c_1 + c_2 x)\mathrm{e}^{\lambda x}$ when they coincide; and if the a_i are real and the roots are $\alpha \pm \mathrm{i}\beta$ with $\beta \neq 0$, the real solutions are spanned by $\mathrm{e}^{\alpha x}\cos \beta x$ and $\mathrm{e}^{\alpha x}\sin \beta x$.

Rests on Theorems 8.19 and 9.13 and Proposition 7.30.

Derivation. Derives Proposition 9.28.

Write D for $\mathrm{d}/\mathrm{d}x$. Polynomials in D with constant coefficients multiply exactly as polynomials in λ do, because D commutes with itself and with constants, so the factorization $P(\lambda) = a_n \prod_j (\lambda - \lambda_j)^{m_j}$ furnished by the fundamental theorem of algebra (Theorem 8.19) gives the operator identity $L = a_n \prod_j (D - \lambda_j)^{m_j}$, in which the factors may be applied in any order.

A single factor acts on the proposed solutions by

$$(D - \lambda)\big[x^{\kappa}\mathrm{e}^{\lambda x}\big] = \kappa\, x^{\kappa - 1}\mathrm{e}^{\lambda x}$$

(Proposition 7.30), so $(D - \lambda)^m$ annihilates $x^{\kappa}\mathrm{e}^{\lambda x}$ for every $\kappa \leq m - 1$. Applying the factors of L in an order that puts $(D - \lambda_j)^{m_j}$ last shows every function Equation (9.33) solves $L[y] = 0$.

They are independent. Suppose $\sum_j p_j(x)\mathrm{e}^{\lambda_j x}$ vanishes identically, with p_j polynomials; we show every p_j is zero, by induction on the number p of distinct exponents. For $p = 1$, an exponential never vanishes, so p_1 has infinitely many roots and is the zero polynomial. For $p > 1$, multiply by $\mathrm{e}^{-\lambda_p x}$ and differentiate $\deg p_p + 1$ times. The last term dies; each other term $p_j \mathrm{e}^{\mu_j x}$, with $\mu_j = \lambda_j - \lambda_p \neq 0$, becomes $q_j \mathrm{e}^{\mu_j x}$ with $\deg q_j = \deg p_j$, since each differentiation multiplies the leading coefficient by $\mu_j \neq 0$. The inductive hypothesis makes every q_j, hence every p_j with $j < p$, zero; and then p_p is zero by the case $p = 1$.

There are $\sum_j m_j = n$ of them, and the solution space has dimension n (Theorem 9.13), so they are a basis. The real form of a conjugate pair is obtained as in Corollary 9.25, by taking the real and imaginary parts of $\mathrm{e}^{(\alpha + \mathrm{i}\beta)x}$ (Proposition 8.4); they are independent because their Wronskian is $\beta\, \mathrm{e}^{2\alpha x} \neq 0$ (Proposition 9.16). □

9.2.3 Planar systems: classification of the equilibrium

Proposition 9.29 (The planar classification)**.** *Let A be a real 2×2 matrix with $\tau = \operatorname{tr} A$, $\Delta = \det A$ and discriminant $D = \tau^2 - 4\Delta$, so that the eigenvalues are*

$$\lambda_{\pm} = \tfrac{1}{2}\big(\tau \pm \sqrt{D}\big)\,. \tag{9.34}$$

The origin, the only equilibrium of $\boldsymbol{Y}' = A\boldsymbol{Y}$ when $\Delta \neq 0$, is classified by (τ, Δ) exactly as in Table 9.1. In particular it is asymptotically stable if and only if $\tau < 0$ and $\Delta > 0$, and unstable whenever $\Delta < 0$ or $\tau > 0$.

Rests on Corollaries 9.25 and 9.27.

Derivation. Derives Proposition 9.29.

The characteristic polynomial of a 2×2 matrix is $\lambda^2 - \tau\lambda + \Delta$ (Theorem 5.67), whose roots are Equation (9.34); they multiply to Δ and add to τ. If $D > 0$ the roots are real

and distinct: they have opposite signs when $\Delta < 0$ (a saddle, one growing and one decaying direction) and the common sign of τ when $\Delta > 0$ (a node). If $D < 0$ they are a conjugate pair $\tau/2 \pm \mathrm{i}\beta$ with $\beta = \sqrt{-D}/2 \neq 0$, and Equation (9.27) shows every solution to be a rotation of period $2\pi/\beta$ scaled by $\mathrm{e}^{\tau x/2}$: a focus spiralling in for $\tau < 0$ and out for $\tau > 0$, and for $\tau = 0$ a centre, every orbit of which is closed and periodic. If $D = 0$ the eigenvalue $\tau/2$ is double and Equation (9.28) permits a factor of x, present exactly when A is not a multiple of the identity: a degenerate node, or a star node when it is. The stability statements are Corollary 9.27 read through $\Re\lambda_{\pm}$, both of which are negative precisely when their sum τ is negative and their product Δ positive. □

Table 9.1: Classification of the equilibrium of a planar linear system $\boldsymbol{Y}' = A\boldsymbol{Y}$ by $\tau = \operatorname{tr} A$, $\Delta = \det A$ and $D = \tau^2 - 4\Delta$, as proved in Proposition 9.29. The parabola $D = 0$ separates the nodes from the foci and the axis $\Delta = 0$ from the saddles; the half-line $\tau = 0$, $\Delta > 0$ carries the centres.

Condition	Eigenvalues	Equilibrium
$\Delta < 0$	real, opposite signs	saddle (unstable)
$\Delta > 0$, $D > 0$, $\tau < 0$	real, both negative	stable node
$\Delta > 0$, $D > 0$, $\tau > 0$	real, both positive	unstable node
$\Delta > 0$, $D < 0$, $\tau < 0$	complex, $\Re < 0$	stable focus
$\Delta > 0$, $D < 0$, $\tau > 0$	complex, $\Re > 0$	unstable focus
$\Delta > 0$, $\tau = 0$	purely imaginary	centre
$\Delta > 0$, $D = 0$	real, equal	degenerate or star node
$\Delta = 0$	one eigenvalue zero	a line of equilibria

9.2.4 Equilibria of a nonlinear system

A nonlinear autonomous system $\mathrm{d}\boldsymbol{Y}/\mathrm{d}x = \boldsymbol{F}(\boldsymbol{Y})$ with $\boldsymbol{F}(\boldsymbol{Y}_*) = \boldsymbol{0}$ is approximated near its equilibrium by the linear system carried by the derivative matrix $A = D\boldsymbol{F}(\boldsymbol{Y}_*)$. Two questions follow, and they have different answers. Whether the *stability* of the equilibrium is decided by A is settled below, and affirmatively when no eigenvalue sits on the imaginary axis. Whether the whole *phase portrait* is decided by A is a deeper statement, and is the one theorem of this chapter that is quoted rather than proved.

Definition 9.30 (Hyperbolic equilibrium)**.** An equilibrium $\boldsymbol{Y}_*$ of $\mathrm{d}\boldsymbol{Y}/\mathrm{d}x = \boldsymbol{F}(\boldsymbol{Y})$, $\boldsymbol{F}$ of class C^1, is *hyperbolic* when no eigenvalue of $A = D\boldsymbol{F}(\boldsymbol{Y}_*)$ has zero real part.

Rests on Definitions 5.65 and 7.66.

Proposition 9.31 (Lyapunov's first method: asymptotic stability)**.** *Let $\boldsymbol{F}$ be of class C^1 near an equilibrium $\boldsymbol{Y}_*$ and suppose every eigenvalue of $A = D\boldsymbol{F}(\boldsymbol{Y}_*)$ has $\Re\lambda < 0$. Then $\boldsymbol{Y}_*$ is asymptotically stable: there are $\varepsilon > 0$, $C > 0$ and $\mu > 0$ such that every solution starting within ε of $\boldsymbol{Y}_*$ satisfies*

$$\|\boldsymbol{Y}(x) - \boldsymbol{Y}_*\| \leq C\,\mathrm{e}^{-\mu x}\,\|\boldsymbol{Y}(0) - \boldsymbol{Y}_*\|\,, \qquad x \geq 0\,. \tag{9.35}$$

Rests on Corollary 9.27, Lemma 9.10, and Theorem 9.23.

Derivation. *Derives Proposition 9.31.*

Put $\boldsymbol{Z} = \boldsymbol{Y} - \boldsymbol{Y}_*$ and $\boldsymbol{N}(\boldsymbol{Z}) = \boldsymbol{F}(\boldsymbol{Y}_* + \boldsymbol{Z}) - A\boldsymbol{Z}$, so that $\boldsymbol{Z}' = A\boldsymbol{Z} + \boldsymbol{N}(\boldsymbol{Z})$. Because $\boldsymbol{F}$ is C^1 and A is its derivative at the equilibrium, $\|\boldsymbol{N}(\boldsymbol{Z})\| / \|\boldsymbol{Z}\| \to 0$ as $\boldsymbol{Z} \to \boldsymbol{0}$ (Theorem 7.68): given any $\eta > 0$ there is a $\delta > 0$ with $\|\boldsymbol{N}(\boldsymbol{Z})\| \leq \eta\,\|\boldsymbol{Z}\|$ for $\|\boldsymbol{Z}\| \leq \delta$.

Choose σ with $\sigma(A) < \sigma < 0$ and let C be the constant of Equation (9.31). Fix $\eta = -\sigma/(2C)$ and the corresponding δ, and treat $\boldsymbol{N}$ as an inhomogeneity in Equation (9.25): as long as the solution stays within δ,

$$\|\boldsymbol{Z}(x)\| \le C\mathrm{e}^{\sigma x}\,\|\boldsymbol{Z}(0)\| + \int_0^x C\,\mathrm{e}^{\sigma(x-s)}\,\eta\,\|\boldsymbol{Z}(s)\|\,\mathrm{d}s\,.$$

Multiplying by $\mathrm{e}^{-\sigma x}$ and writing $u(x) = \mathrm{e}^{-\sigma x}\,\|\boldsymbol{Z}(x)\|$ turns this into $u(x) \le C\,\|\boldsymbol{Z}(0)\| + C\eta\int_0^x u$, so Lemma 9.10 gives $u(x) \le C\,\|\boldsymbol{Z}(0)\|\,\mathrm{e}^{C\eta x}$, that is

$$\|\boldsymbol{Z}(x)\| \le C\;\|\boldsymbol{Z}(0)\|\;\mathrm{e}^{(\sigma+C\eta)x} = C\;\|\boldsymbol{Z}(0)\|\;\mathrm{e}^{\sigma x/2}\,,$$

by the choice of η. Taking $\varepsilon = \delta/C$ makes the right-hand side at most δ for every $x \ge 0$, so the solution never leaves the ball on which the estimate was assumed and the bound holds for all $x \ge 0$. This is Equation (9.35) with $\mu = -\sigma/2 > 0$. □

Theorem 9.32 (Hartman–Grobman; quoted)**.** *Let $\boldsymbol{F}$ be of class C^1 near a hyperbolic equilibrium $\boldsymbol{Y}_*$ and let $A = D\boldsymbol{F}(\boldsymbol{Y}_*)$. Then there are neighbourhoods U of $\boldsymbol{Y}_*$ and V of $\boldsymbol{0}$ and a homeomorphism $\varphi : U \longrightarrow V$ carrying the trajectories of $\boldsymbol{Y}' = \boldsymbol{F}(\boldsymbol{Y})$ in U onto the trajectories of $\boldsymbol{Z}' = A\boldsymbol{Z}$ in V, preserving the direction and the parametrization of time [1183]. The nonlinear flow near a hyperbolic equilibrium is therefore a continuous distortion of its linearization, and in particular the equilibrium is unstable whenever some eigenvalue of A has positive real part.*

Rests on Definitions 6.7 and 9.30 and Theorem 9.23.

Remark 9.33 (What is quoted here). Theorem 9.32 is the only statement of this chapter about ordinary differential equations that carries no derivation, and the omission is a decision rather than a debt. Its proof constructs the conjugating map as the fixed point of a contraction on a space of bounded continuous maps, after splitting the linearization into its contracting and expanding parts and cutting the nonlinearity off outside a small ball; the technique is that of Theorem 9.8 applied in a function space, but the bookkeeping is a chapter's worth and it serves this one statement.

Three things about the statement are worth fixing in the reader's mind, because each is a place where more is often claimed than is true. The conjugacy is a *homeomorphism* and in general not differentiable, so it preserves the qualitative phase portrait — which trajectories go where — and not the eigenvalues or any other number computed from the linearization. Hyperbolicity is essential: Proposition 9.29 shows a centre, where $\Re\lambda = 0$, and an arbitrarily small nonlinear perturbation can turn a centre into a stable or an unstable focus, so no conjugacy can exist there. And the theorem is local; it says nothing about the flow away from the equilibrium.

What the rest of this treatise actually uses is weaker and is proved here. The decay statement is Proposition 9.31, obtained from Equation (9.31) and Grönwall's inequality with no conjugacy anywhere; only the instability half of Theorem 9.32 is genuinely borrowed.

9.2.5 Planar flows and closed orbits

Theorem 9.34 (Poincaré–Bendixson; quoted)**.** *Let $\boldsymbol{F}$ be of class C^1 on an open subset of $\mathbb{R}^2$ and let K be a compact subset of it containing no equilibrium of $\boldsymbol{Y}' = \boldsymbol{F}(\boldsymbol{Y})$. If a trajectory enters K and remains in K for all later times, then it is a closed orbit or it approaches one [220, 298]. A bounded planar trajectory that does not converge to an equilibrium therefore has periodic behaviour as its only possible fate.*

Rests on Definitions 6.9 and 9.1 and Theorem 9.8.

Remark 9.35 (What is quoted here, and why it is a statement about the plane). The proof rests on two facts about the plane and on nothing else: a simple closed curve separates

the plane into an inside and an outside (the Jordan curve theorem, which this treatise does not prove — Chapter 6 carries the winding number, Definition 6.20, which is the tool such a proof uses, but stops short of the separation theorem), and successive crossings of a short segment transverse to the flow are therefore monotone along that segment, because a trajectory arc together with the chord joining its ends bounds a region the flow cannot re-enter. It is the dimension that does the work, and the theorem is false in three dimensions: a bounded trajectory there may wander forever without approaching any closed orbit. That failure is not a technicality but the opening of a subject, and it is where Chapter 32 begins.

Proposition 9.36 (Bendixson–Dulac negative criterion)**.** *Let $\boldsymbol{F} = (F_1, F_2)$ be of class C^1 on a simply connected open $\Omega \subseteq \mathbb{R}^2$ and let φ be a positive C^1 function on Ω. If*

$$\frac{\partial}{\partial y_1}(\varphi F_1) + \frac{\partial}{\partial y_2}(\varphi F_2) \tag{9.36}$$

has one strict sign throughout Ω, then $\boldsymbol{Y}' = \boldsymbol{F}(\boldsymbol{Y})$ has no closed orbit lying in Ω.
Rests on Theorem 7.99 and Definition 6.17.

Derivation. *Derives Proposition 9.36.*

Suppose a closed orbit γ lay in Ω, and let S be the region it bounds, contained in Ω because Ω is simply connected. Green's theorem Theorem 7.99 equates the integral of Equation (9.36) over S with the circulation $\oint_\gamma \varphi\,(F_1\,\mathrm{d}y_2 - F_2\,\mathrm{d}y_1)$. On the orbit $\mathrm{d}y_1 = F_1\,\mathrm{d}x$ and $\mathrm{d}y_2 = F_2\,\mathrm{d}x$, so the integrand is $\varphi\,(F_1F_2 - F_2F_1)\,\mathrm{d}x = 0$ and the circulation vanishes. But the area integral of a function of one strict sign over a region of positive area cannot vanish — a contradiction. □

9.3 Linear systems with periodic coefficients

A linear system whose coefficients vary periodically in the independent variable is not solvable in closed form, and yet its stability is decided completely by a single matrix. That is Floquet's theorem, and it is the exact statement behind every parametric-resonance calculation: a swing pumped by shortening the rope at twice its natural frequency, a mass on a vibrating support, a charged particle in the radio-frequency field of an ion trap, the tension modulation of a plucked string. The physical parts of this treatise derive the leading resonance band by a slowly-varying-amplitude argument; what follows is the theorem that argument approximates, and it is what entitles anyone to speak of a *tongue* structure at all.

9.3.1 The monodromy matrix

Definition 9.37 (Fundamental matrix)**.** Let A be continuous on $\mathbb{R}$ with values in the $n \times n$ matrices. The *fundamental matrix* of $\boldsymbol{Y}' = A(x)\boldsymbol{Y}$ based at x_0 is the matrix-valued solution Φ of

$$\frac{\mathrm{d}\Phi}{\mathrm{d}x}(x) = A(x)\,\Phi(x)\,, \qquad \Phi(x_0) = \mathbb{1}\,, \tag{9.37}$$

that is, the matrix whose j-th column is the solution with initial value $\boldsymbol{e}_j$. Every solution is $\boldsymbol{Y}(x) = \Phi(x)\boldsymbol{Y}(x_0)$.
Rests on Corollary 9.9 and Theorem 9.13.

Proposition 9.38 (Liouville's formula)**.** *The fundamental matrix satisfies*

$$\det \Phi(x) = \exp\left[\int_{x_0}^{x} \operatorname{tr} A(s)\,\mathrm{d}s\right], \tag{9.38}$$

and is therefore invertible at every x.
Rests on Definition 9.37 and Proposition 9.15.

Derivation. *Derives Proposition 9.38.*

This is the computation of Proposition 9.15 carried out on the columns of Φ instead of on a Wronskian. A determinant is multilinear in its rows, so $\mathrm{d}(\det\Phi)/\mathrm{d}x$ is the sum over i of the determinant of Φ with its i-th row differentiated. By Equation (9.37) that row is $\sum_k A_{ik}\Phi_{k\cdot}$, a combination of the rows of Φ; subtracting the multiples of the other rows, which changes no determinant, leaves $A_{ii}\det\Phi$. Summing over i gives $\mathrm{d}(\det\Phi)/\mathrm{d}x = (\operatorname{tr}A)\det\Phi$, a scalar linear equation whose solution with $\det\Phi(x_0)=1$ is Equation (9.38) by Proposition 9.18. An exponential never vanishes. □

Theorem 9.39 (Floquet)**.** *Let A be continuous and T-periodic, $A(x+T)=A(x)$, and let Φ be the fundamental matrix based at $x_0=0$. Write $M=\Phi(T)$, the* monodromy matrix*, and let $\rho_1,\dots,\rho_n$ be its eigenvalues, the* characteristic multipliers*. Then*

1. *$\Phi(x+T)=\Phi(x)\,M$ for every x, and hence $\Phi(x+NT)=\Phi(x)M^N$ for every integer $N\geq 0$;*
2. *M is invertible, with $\det M=\prod_i\rho_i=\exp\bigl[\int_0^T\operatorname{tr}A\bigr]$, so no multiplier is zero;*
3. *to each multiplier ρ belongs a solution with $\boldsymbol{Y}(x+T)=\rho\,\boldsymbol{Y}(x)$; writing $\rho=\mathrm{e}^{\mu T}$ for a* characteristic exponent *μ, that solution is*

$$\boldsymbol{Y}(x)=\mathrm{e}^{\mu x}\,\boldsymbol{P}(x)\,,\qquad \boldsymbol{P}(x+T)=\boldsymbol{P}(x)\,, \tag{9.39}$$

   an exponential times a T-periodic vector function;
4. *every solution tends to $\mathbf{0}$ as $x\to+\infty$ if and only if $|\rho_i|<1$ for every i, and some solution is unbounded if $|\rho_i|>1$ for some i.*

Rests on Definition 9.37, Proposition 9.38, and Corollary 9.27.

Derivation. *Derives Theorem 9.39.*

(i) Put $\Psi(x)=\Phi(x+T)$. Then $\Psi'=A(x+T)\Psi=A(x)\Psi$ by periodicity, so Ψ solves Equation (9.37) column by column; and $\Phi(x)M$ solves it too, being a fixed linear combination of solutions. The two agree at $x=0$, where both equal M, so they agree everywhere by the uniqueness in Corollary 9.9. Iterating gives the statement for NT.

(ii) is Equation (9.38) at $x=T$ together with the fact that the determinant is the product of the eigenvalues (Theorem 5.67, comparing the constant terms of the characteristic polynomial).

(iii) Let $M\boldsymbol{w}=\rho\boldsymbol{w}$ with $\boldsymbol{w}\neq\mathbf{0}$ and put $\boldsymbol{Y}(x)=\Phi(x)\boldsymbol{w}$. By (i), $\boldsymbol{Y}(x+T)=\Phi(x)M\boldsymbol{w}=\rho\,\boldsymbol{Y}(x)$. Since $\rho\neq 0$ by (ii), the complex exponential takes the value ρ (Proposition 8.4: write $\rho=|\rho|\,\mathrm{e}^{\mathrm{i}\theta}$ and take $\mu T=\ln|\rho|+\mathrm{i}\theta$). Setting $\boldsymbol{P}(x)=\mathrm{e}^{-\mu x}\boldsymbol{Y}(x)$ gives $\boldsymbol{P}(x+T)=\mathrm{e}^{-\mu T}\mathrm{e}^{-\mu x}\rho\,\boldsymbol{Y}(x)=\boldsymbol{P}(x)$, which is Equation (9.39).

(iv) By (i), $\boldsymbol{Y}(x+NT)=\Phi(x)M^N\boldsymbol{Y}(0)$. The matrix Φ is continuous, hence bounded on the compact interval $[0,T]$ (Theorem 7.24), and $\Phi(x)^{-1}$ is bounded there too by Equation (9.38). So the solution tends to zero, or is bounded, exactly as the sequence $M^N\boldsymbol{Y}(0)$ does, and the statements about the powers A^N in Corollary 9.27 settle that by the moduli of the multipliers: $M^N\to 0$ when every $|\rho_i|<1$, while an eigenvector for a multiplier of modulus greater than one gives $M^N\boldsymbol{w}=\rho^N\boldsymbol{w}$, unbounded. □

Remark 9.40 (The normal form, and the one place a logarithm is needed)*.* Floquet's theorem is often stated as the single assertion $\Phi(x)=P(x)\,\mathrm{e}^{xB}$ with P a T-periodic invertible matrix function and B constant. That form follows from Theorem 9.39(iii) whenever M has n independent eigenvectors: collect them as the columns of V, put $B=V\operatorname{diag}(\mu_1,\dots,\mu_n)V^{-1}$ and $P(x)=\Phi(x)\mathrm{e}^{-xB}$, which is periodic because each column is by (iii). In general it requires a logarithm of the matrix M itself, which exists because M is invertible but whose

construction needs a normal form for M that Chapter 5 does not develop. Nothing below uses the aggregate form: every statement this treatise makes about a periodically forced linear system is a statement about the multipliers, and Theorem 9.39(iii) and (iv) supply those without any logarithm of a matrix. A further caution: B and P are generally complex even for a real system, because a negative real multiplier has no real logarithm; the standard repair is to work with period $2T$, under which the multipliers are squared and become positive.

9.3.2 Hill's equation and the tongues of instability

Corollary 9.41 (Hill's equation)**.** *Let Q be continuous and T-periodic and consider*

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2}(x) + Q(x)\, y(x) = 0\,, \tag{9.40}$$

written as a first-order system for (y, y'). Then $\operatorname{tr} A = 0$, so $\det M = 1$ and the two multipliers satisfy

$$\rho^2 - (\operatorname{tr} M)\,\rho + 1 = 0\,, \qquad \rho_1\rho_2 = 1\,, \qquad \rho_1 + \rho_2 = \operatorname{tr} M\,. \tag{9.41}$$

Consequently, with $\Delta_{\mathrm{H}} = \operatorname{tr} M$ the discriminant *of Equation (9.40):*

1. *if $|\Delta_{\mathrm{H}}| < 2$ the multipliers are $\mathrm{e}^{\pm \mathrm{i}\theta}$ with $2\cos\theta = \Delta_{\mathrm{H}}$, distinct and of unit modulus, and every solution is bounded on $\mathbb{R}$: the equation is* stable*;*
2. *if $|\Delta_{\mathrm{H}}| > 2$ the multipliers are real, reciprocal and distinct, one of modulus greater than one, so some solution grows like $|\rho|^{x/T}$ — exponentially, at the rate $\mu = T^{-1}\ln|\rho|$: the equation is* unstable*;*
3. *if $|\Delta_{\mathrm{H}}| = 2$ the multipliers are both $+1$ or both -1, and Equation (9.40) has a solution of period T or of period $2T$ respectively.*

Rests on Theorem 9.39, Proposition 9.38, and Corollary 9.27.

Derivation. *Derives Corollary 9.41.*

In the system for $\boldsymbol{Y} = (y, y')$ the matrix is $A = \begin{pmatrix} 0 & 1 \\ -Q & 0 \end{pmatrix}$, of zero trace, so Equation (9.38) gives $\det M = 1$ and the characteristic polynomial of the 2×2 matrix M is Equation (9.41) (Proposition 9.29, the same identity read for M). Its roots are $\rho_{\pm} = \left[\Delta_{\mathrm{H}} \pm \sqrt{\Delta_{\mathrm{H}}^2 - 4}\,\right]/2$.

If $|\Delta_{\mathrm{H}}| < 2$ the discriminant is negative, the roots are complex conjugates of product 1, hence of modulus 1, and distinct; M then has two independent eigenvectors (Proposition 5.71), so the powers M^N stay bounded by the last clause of Corollary 9.27, and by the argument of Theorem 9.39(iv) every solution is bounded. If $|\Delta_{\mathrm{H}}| > 2$ the roots are real and distinct with product 1, so one has modulus greater than 1 and Theorem 9.39(iv) gives an unbounded solution; by Equation (9.39) it is $\mathrm{e}^{\mu x}$ times a periodic function, with $\mathrm{e}^{\mu T} = \rho$. If $|\Delta_{\mathrm{H}}| = 2$ the double root is $\rho = \Delta_{\mathrm{H}}/2 = \pm 1$, and the eigenvector of M belonging to it furnishes, by Theorem 9.39(iii), a solution with $y(x + T) = \pm y(x)$. □

Remark 9.42 (Why the unstable set is a family of tongues)*.* Let Q depend on two parameters — a mean value a and a modulation depth h — continuously. Then M, and with it $\Delta_{\mathrm{H}}(a, h)$, depends continuously on (a, h), because the solution of an initial-value problem does (Proposition 9.11, applied to the system with parameters carried along). Corollary 9.41 therefore partitions the parameter plane into the open stable set $|\Delta_{\mathrm{H}}| < 2$, the open unstable set $|\Delta_{\mathrm{H}}| > 2$, and the level curves $\Delta_{\mathrm{H}} = \pm 2$ that separate them. On those curves a periodic

or antiperiodic solution exists, by Corollary 9.41(iii), which is what makes them locatable in practice: one looks for solutions of period T and $2T$ rather than computing M. The unstable set is the union of the regions bounded by those curves, and at $h = 0$ it degenerates — as Proposition 9.43 shows for the Mathieu equation — to a discrete set of points on the a axis, from which the regions open out as h grows. That is the origin of the name.

9.3.3 The Mathieu equation

The canonical Hill equation, and the one that the separation of the Helmholtz operator in elliptic-cylinder coordinates produces [173], is

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2}(x) + \left[a - 2h\cos 2x\right] y(x) = 0\,. \tag{9.42}$$

Its Sturm–Liouville data, in the convention of Definition 9.54, are

$$p(x) = 1\,, \qquad q(x) = -2h\cos 2x\,, \qquad \lambda^2 = a\,, \qquad r(x) = 1\,, \tag{9.43}$$

so the spectral parameter is a and the modulation depth h is a parameter of the operator; much of the literature writes q for what is here called h, which would collide with the Sturm–Liouville coefficient $q(x)$. The coefficient has period π in x, so Theorem 9.39 applies with $T = \pi$. The solutions that are periodic — which by Corollary 9.41(iii) exist exactly on the boundaries of the unstable regions — are the *Mathieu functions*; the remaining solutions are, by Equation (9.39), an exponential times a periodic function, and it is the sign of the exponent that the experiment sees.

Proposition 9.43 (The unmodulated line, and the feet of the tongues)**.** *On the line $h = 0$, Equation (9.42) is stable in the sense of Corollary 9.41 at every $a > 0$ with $\sqrt{a}$ not an integer, and unstable at every $a \le 0$. Consequently the closure of the unstable set meets the ray $a > 0$ exactly at the points*

$$a = n^2\,, \qquad n = 1, 2, 3, \ldots\,, \tag{9.44}$$

which are the feet from which the instability tongues open out as $|h|$ grows; the whole ray $a \le 0$ is unstable already at $h = 0$ and is the $n = 0$ region.

Rests on Corollary 9.41, Theorem 9.23, and Equation (9.42).

Derivation. *Derives Proposition 9.43.*

At $h = 0$ the equation is $y'' + ay = 0$ with constant coefficients, whose fundamental matrix over one period $T = \pi$ is computed from Theorem 9.23. For $a > 0$ the solutions are $\cos(\sqrt{a}\,x)$ and $\sin(\sqrt{a}\,x)/\sqrt{a}$, so

$$M = \begin{pmatrix} \cos(\pi\sqrt{a}) & \sin(\pi\sqrt{a})/\sqrt{a} \\ -\sqrt{a}\,\sin(\pi\sqrt{a}) & \cos(\pi\sqrt{a}) \end{pmatrix}, \qquad \Delta_{\mathrm{H}} = 2\cos(\pi\sqrt{a})\,.$$

Hence $|\Delta_{\mathrm{H}}| < 2$, the stable case of Corollary 9.41, at every $a > 0$ except where $\cos(\pi\sqrt{a}) = \pm 1$, that is where $\pi\sqrt{a} = n\pi$, that is at $a = n^2$ with $n \ge 1$; there $|\Delta_{\mathrm{H}}| = 2$ and the second solution grows linearly, so those points belong to the closure of the unstable set. For $a = 0$ the solutions are 1 and x, giving $\Delta_{\mathrm{H}} = 2$ with an unbounded solution, and for $a < 0$ they are real exponentials with $\Delta_{\mathrm{H}} = 2\cosh(\pi\sqrt{-a}) > 2$, the unstable case outright. So on the line $h = 0$ the unstable set is the ray $a \le 0$ together with the isolated points $a = n^2$, $n \ge 1$, which is the statement. □

Remark 9.44 (Reading the feet of the tongues as a resonance condition). An oscillator of natural angular frequency ω_0 whose restoring coefficient is modulated at angular frequency γ obeys $\ddot{u} + \omega_0^2[1 + \epsilon\cos\gamma t]u = 0$. Substituting $x = \gamma t/2$ turns it into Equation (9.42) with

$$a = \frac{4\omega_0^2}{\gamma^2}\,, \qquad 2h = -\frac{4\omega_0^2\epsilon}{\gamma^2}\,, \tag{9.45}$$

so Equation (9.44) becomes $\gamma = 2\omega_0/n$ with $n \geq 1$: the modulation destabilises the oscillator when it is driven at twice the natural frequency, or at an integer submultiple of that. The $n = 1$ tongue is the widest and is the one observed; the others narrow rapidly with n, which is the quantitative statement that a slowly-varying-amplitude argument cannot make and this theorem can. Chapter 28 derives the width of the first tongue and the damping threshold, and Chapter 35 measures the effect on a string.

9.4 Orthonormal systems of functions

9.4.1 Orthonormality

Definition 9.45 (Orthonormal set of functions)**.** Let $\{\phi_i\}$ be a collection of functions

$$\begin{aligned}\phi_i : I \subset \mathbb{R} &\longrightarrow \mathbb{C}\\ x &\longmapsto \phi_i(x)\,.\end{aligned}$$

The functions ϕ_i form an *orthonormal set of functions* if and only if

$$\int_I \overline{\phi_i(x)}\,\phi_j(x)\,\mathrm{d}x = \delta_{ij}\,. \tag{9.46}$$

Remark 9.46. For real-valued functions the conjugation in Equation (9.46) is immaterial and the relation reduces to $\int_I \phi_i\phi_j\,\mathrm{d}x = \delta_{ij}$, which is the form in which the Legendre and Bessel relations below are stated.

9.4.2 Expansion in series of orthonormal functions

Definition 9.47 (Orthonormal series expansion)**.** The *expansion of a function f in a series of orthonormal functions* $\{\phi_i\}$ is

$$f(x) = \sum_{i=1}^{\infty} c_i\,\phi_i(x)\,. \tag{9.47}$$

Rests on Definition 9.45.

Proposition 9.48 (Expansion coefficients)**.** *The coefficients of the expansion Equation (9.47) are*

$$c_i = \int_I f(x)\,\overline{\phi_i(x)}\,\mathrm{d}x\,. \tag{9.48}$$

Rests on Definitions 9.45 and 9.47.

Derivation. Derives Proposition 9.48.

Multiply Equation (9.47) by $\overline{\phi_j(x)}$ and integrate over I:

$$\begin{aligned}f(x) &= \sum_{i=1}^{\infty} c_i\,\phi_i(x)\,,\\ f(x)\,\overline{\phi_j(x)} &= \sum_{i=1}^{\infty} c_i\,\phi_i(x)\,\overline{\phi_j(x)}\,,\\ \int_I f(x)\,\overline{\phi_j(x)}\,\mathrm{d}x &= \sum_{i=1}^{\infty} c_i \int_I \overline{\phi_j(x)}\,\phi_i(x)\,\mathrm{d}x = \sum_{i=1}^{\infty} c_i\,\delta_{ji} = c_j\,,\end{aligned}$$

where the orthonormality relation Equation (9.46) collapsed the sum. □

9.4.3 Expansion of a product

Consider two functions f_1 and f_2, each expanded in the same orthonormal set, together with the expansion of their pointwise product:

$$f_1(x) = \sum_{i=1}^{\infty} c_i^{(1)}\, \phi_i(x)\,,$$
$$f_2(x) = \sum_{i=1}^{\infty} c_i^{(2)}\, \phi_i(x)\,,$$
$$f_1(x)\, f_2(x) = \sum_{i=1}^{\infty} c_i^{(12)}\, \phi_i(x)\,,$$

where $c_i^{(1)}$ and $c_i^{(2)}$ are the expansion coefficients of f_1 and f_2, and $c_i^{(12)}$ those of the product $f_1 f_2$.

Definition 9.49 (Structure constants of an orthonormal set)**.** The *structure constants* of the orthonormal set $\{\phi_i\}$ are its triple overlap integrals

$$T_{ijk} = \int_I \overline{\phi_i(x)}\, \phi_j(x)\, \phi_k(x)\, \mathrm{d}x\,. \tag{9.49}$$

Rests on Definition 9.45.

Proposition 9.50 (Coefficients of a product)**.** *Whenever the three expansions above are valid and the double series below converges absolutely, the coefficients of the product are the values of one fixed bilinear form on the two coefficient sequences,*

$$c_i^{(12)} = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} T_{ijk}\, c_j^{(1)}\, c_k^{(2)}\,, \tag{9.50}$$

with T_{ijk} the structure constants Equation (9.49).
Rests on Proposition 9.48 and Definition 9.49.

Derivation. *Derives Proposition 9.50.*
Applying Equation (9.48) to the product and substituting the two series,

$$\begin{aligned} c_i^{(12)} &= \int_I f_1(x)\, f_2(x)\, \overline{\phi_i(x)}\, \mathrm{d}x \\ &= \int_I \left(\sum_{j=1}^{\infty} c_j^{(1)} \phi_j(x) \right) \left(\sum_{k=1}^{\infty} c_k^{(2)} \phi_k(x) \right) \overline{\phi_i(x)}\, \mathrm{d}x \\ &= \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} c_j^{(1)} c_k^{(2)} \int_I \overline{\phi_i(x)}\, \phi_j(x)\, \phi_k(x)\, \mathrm{d}x\,, \end{aligned}$$

and the remaining integral is T_{ijk} by Equation (9.49). □

Remark 9.51 (This is the closed form). No further reduction is available for a general orthonormal set, and none should be looked for: the numbers T_{ijk} are independent data attached to the set — they measure exactly how far it is from being closed under multiplication — and Equation (9.50) is the statement that pointwise multiplication, read in coefficients, is the bilinear map they define. What makes a particular set useful is that its structure constants are sparse. The complex exponentials are the extreme case: a single term of the double sum survives and the bilinear map degenerates into a convolution.

Proposition 9.52 (Product coefficients for the Fourier system)**.** *Take the complex exponentials $\phi_n(x) = \mathrm{e}^{\mathrm{i}nx}$ on $[-\pi, \pi]$ in the normalisation of Equations (9.110) and (9.111), that is with every integral of Section 9.4 — and in particular the triple overlap Equation (9.49) — read against the normalised measure $\mathrm{d}x/(2\pi)$ of Remark 9.91, which is what makes them an orthonormal set at all. Their structure constants are then $T_{njk} = \delta_{n,\,j+k}$ and Equation (9.50) becomes the discrete convolution*

$$c^{(12)}_n = \sum_{m=-\infty}^{\infty} c^{(1)}_m \, c^{(2)}_{n-m} \,, \tag{9.51}$$

absolutely convergent whenever both coefficient sequences are absolutely summable.

Rests on Proposition 9.50 and Definition 9.49.

Derivation. *Derives Proposition 9.52.*

Every integral carries the measure $\mathrm{d}x/(2\pi)$ (Remark 9.91), which is also what puts the factor into Equation (9.111), so the triple overlap to be computed is

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}nx}\,\mathrm{e}^{\mathrm{i}jx}\,\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}x = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}(j+k-n)x}\,\mathrm{d}x = \delta_{n,\,j+k}\,,$$

by Equation (9.109) applied to the integer $j + k - n$. Inserting this in Equation (9.50) collapses the double sum to the single sum over $j = m$, $k = n - m$, which is Equation (9.51). Absolute convergence, and with it the right to rearrange the double series into that single one, is the Cauchy product of two absolutely convergent series (Proposition 7.49).

The other reading of Remark 9.91 gives the same answer, as it must. For the set $\phi_n = \mathrm{e}^{\mathrm{i}nx}/\sqrt{2\pi}$ with plain $\mathrm{d}x$ the same integral yields $T_{njk} = \delta_{n,\,j+k}/\sqrt{2\pi}$ while every coefficient is multiplied by $\sqrt{2\pi}$, and the three factors cancel in Equation (9.50), leaving Equation (9.51) unchanged. □

9.4.4 Parseval relation

Let f be a function whose expansion Equation (9.47) in the orthonormal set $\{\phi_i\}$ converges to f in the mean square. The *Parseval relation* then equates the norm of the function to the norm of its coefficient sequence,

$$\int_I |f(x)|^2 \,\mathrm{d}x = \sum_{i=1}^{\infty} |c_i|^2 \,. \tag{9.52}$$

Derivation. *Derives Equation (9.52).*

Mean-square convergence permits the series to be integrated term by term against its own conjugate:

$$\begin{aligned}\int_I |f(x)|^2 \,\mathrm{d}x &= \int_I \left(\sum_{i=1}^{\infty} c_i\,\phi_i(x)\right)\overline{\left(\sum_{j=1}^{\infty} c_j\,\phi_j(x)\right)}\,\mathrm{d}x \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} c_i\,\overline{c_j}\int_I \overline{\phi_j(x)}\,\phi_i(x)\,\mathrm{d}x = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} c_i\,\overline{c_j}\,\delta_{ji} = \sum_{i=1}^{\infty}|c_i|^2 \,,\end{aligned}$$

the orthonormality relation Equation (9.46) collapsing the double sum to its diagonal. □

Remark 9.53. The hypothesis is not decoration: it is exactly the statement that the orthonormal set is *complete* for f. For an arbitrary orthonormal set only an inequality holds, and it costs one line: expanding the non-negative quantity $\int_I \big|f - \sum_{i\le N} c_i\phi_i\big|^2\mathrm{d}x =$

$\int_I |f|^2\,\mathrm{d}x - \sum_{i\le N} |c_i|^2 \ge 0$, where the cross terms collapse by orthonormality and the definition $c_i = \int_I \bar{\phi}_i f\,\mathrm{d}x$, and letting $N \to \infty$ gives

$$\sum_i |c_i|^2 \le \int_I |f|^2\,\mathrm{d}x\,, \tag{9.53}$$

Bessel's inequality, with equality for every square integrable f being the definition of completeness. (Proposition 17.4 is this same argument carried out for the trigonometric system, where it also identifies the partial sum as the least-squares approximant; it does not by itself cover an arbitrary orthonormal set, which is why the general statement is given here.) The two systems used in this chapter are complete, and each is proved so: Theorem 17.34 for the trigonometric system of Section 9.7, and Theorem 17.36 for the Fourier integral of Section 9.7.5. The statement in the abstract, for an orthonormal basis of an arbitrary Hilbert space, belongs to Chapter 12.

9.5 The Sturm–Liouville operator

Definition 9.54 (Sturm–Liouville differential operator)**.** Let y be a function

$$\begin{aligned} y : \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto y(x)\,. \end{aligned}$$

The *Sturm–Liouville differential operator* acting on y is defined as

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[p(x)\,\frac{\mathrm{d}y}{\mathrm{d}x}(x)\right] + \left[q(x) + \lambda^2\, r(x)\right] y(x)\,, \tag{9.54}$$

where p, q and r are given real functions and λ^2 is a parameter.

Remark 9.55. Setting the expression Equation (9.54) equal to zero yields the *Sturm–Liouville equation*. Wherever $r > 0$ it may be rewritten as

$$-\frac{1}{r(x)}\left\{\frac{\mathrm{d}}{\mathrm{d}x}\left[p(x)\,\frac{\mathrm{d}y}{\mathrm{d}x}(x)\right] + q(x)\,y(x)\right\} = \lambda^2\, y(x)\,, \tag{9.55}$$

an eigenvalue problem with eigenvalue λ^2 and *weight function* r. Each classical family of special functions treated below is obtained by one specific choice of the data (p, q, λ^2, r); the identifications are collected in Table 9.2.

9.5.1 Self-adjointness and orthogonality

The whole force of the Sturm–Liouville form is one identity. Because the second-order part of Equation (9.54) is written as the derivative of $p\,\mathrm{d}y/\mathrm{d}x$ and not as $p\,\mathrm{d}^2y/\mathrm{d}x^2$, two solutions can be played off against each other under the integral sign, and every orthogonality relation proved in this chapter follows from that single manoeuvre rather than from a separate trick per family.

Lemma 9.56 (Lagrange identity)**.** *Write* $L[y] = \frac{\mathrm{d}}{\mathrm{d}x}\left[p\,\frac{\mathrm{d}y}{\mathrm{d}x}\right] + q\,y$ *for the part of Equation* (9.54) *that does not carry the spectral parameter. For any twice differentiable* u *and* v*,*

$$u\,L[v] - v\,L[u] = \frac{\mathrm{d}}{\mathrm{d}x}\left[p\left(u\,\frac{\mathrm{d}v}{\mathrm{d}x} - v\,\frac{\mathrm{d}u}{\mathrm{d}x}\right)\right]\,. \tag{9.56}$$

Rests on Definition 9.54.

Derivation. *Derives Lemma 9.56.*

The terms carrying q cancel, and by the Leibniz rule Equation (7.16),

$$u\frac{\mathrm{d}}{\mathrm{d}x}\left[p\frac{\mathrm{d}v}{\mathrm{d}x}\right]-v\frac{\mathrm{d}}{\mathrm{d}x}\left[p\frac{\mathrm{d}u}{\mathrm{d}x}\right]=\frac{\mathrm{d}}{\mathrm{d}x}\left[u\,p\frac{\mathrm{d}v}{\mathrm{d}x}\right]-p\frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x}-\frac{\mathrm{d}}{\mathrm{d}x}\left[v\,p\frac{\mathrm{d}u}{\mathrm{d}x}\right]+p\frac{\mathrm{d}v}{\mathrm{d}x}\frac{\mathrm{d}u}{\mathrm{d}x},$$

in which the two products of first derivatives cancel, leaving Equation (9.56). □

Theorem 9.57 (Orthogonality of Sturm–Liouville eigenfunctions)**.** *Let u and v be real solutions of the Sturm–Liouville equation Equation (9.55) on $[a,b]$ belonging to different eigenvalues, $\lambda_u^2\neq\lambda_v^2$, and suppose the boundary term vanishes,*

$$\left[p\left(u\frac{\mathrm{d}v}{\mathrm{d}x}-v\frac{\mathrm{d}u}{\mathrm{d}x}\right)\right]_a^b=0\,. \tag{9.57}$$

Then u and v are orthogonal with respect to the weight r,

$$\int_a^b u(x)\,v(x)\,r(x)\,\mathrm{d}x=0\,. \tag{9.58}$$

Rests on Lemma 9.56 and Equation (9.55).

Derivation. *Derives Theorem 9.57.*

By hypothesis $L[u]=-\lambda_u^2\,r\,u$ and $L[v]=-\lambda_v^2\,r\,v$, so that $u\,L[v]-v\,L[u]=(\lambda_u^2-\lambda_v^2)\,r\,u\,v$. Integrating Equation (9.56) over $[a,b]$ therefore gives

$$(\lambda_u^2-\lambda_v^2)\int_a^b u\,v\,r\,\mathrm{d}x=\left[p\left(u\frac{\mathrm{d}v}{\mathrm{d}x}-v\frac{\mathrm{d}u}{\mathrm{d}x}\right)\right]_a^b,$$

which is zero by Equation (9.57). Dividing by the non-zero factor $\lambda_u^2-\lambda_v^2$ gives Equation (9.58). □

Remark 9.58 (The three ways the boundary term dies). Equation (9.57) is satisfied in each of the situations met below, and they are worth naming because they are the only ones used. *Separated boundary conditions*: u and v each satisfy $\alpha\,y(a)+\beta\,\frac{\mathrm{d}y}{\mathrm{d}x}(a)=0$ with the same pair (α,β), and similarly at b; the bracket then vanishes at each endpoint separately, because $(u,\mathrm{d}u/\mathrm{d}x)$ and $(v,\mathrm{d}v/\mathrm{d}x)$ are proportional there. This is the Fourier–Bessel problem of Section 9.8.4, where the condition is $y(a)=0$. *A vanishing coefficient function*: $p(a)=0$ or $p(b)=0$ kills the bracket at that endpoint with no condition on y beyond boundedness of y and $\mathrm{d}y/\mathrm{d}x$ there, which is what happens at $x=0$ for Bessel and at $x=\pm1$ for Legendre and its associated equation. *Decay at infinity*: on an unbounded interval it suffices that $p\,y\,\mathrm{d}y/\mathrm{d}x$ tend to zero, which the Gaussian and exponential weights of Sections 9.10 and 9.11 guarantee for polynomial solutions. The second mechanism is the one that requires care: p vanishing is not by itself enough if the solutions are unbounded there, and Remark 9.112 is a worked instance in which the bracket survives.

9.6 The semiclassical approximation

The families catalogued in the sections that follow solve the Sturm–Liouville equation exactly, one closed form for each choice of the data (p,q,λ^2,r). Almost no potential met in a laboratory is one of them. What survives when the exact solution does not exist is an asymptotic one: whenever the coefficient multiplying the unknown in a second-order linear equation is large, the equation has a universal approximate solution, an oscillation whose

local wavelength and local amplitude are both read off that coefficient. Jeffreys obtained it for the general second-order equation [526], and Wentzel, Kramers and Brillouin obtained it independently of him and of each other, within months, for the wave equation of quantum mechanics [535, 549, 563] — which is why the method is called WKB by most authors and JWKB by the careful ones. This section derives the approximation, the formulae that carry a solution across the points where it fails, the tunnelling exponent that follows from them — the Gamow factor [600, 603] — and the quantisation condition of the old quantum theory [395, 441], which the same machinery delivers with its half-integer correction attached rather than postulated.

9.6.1 The equation and its Sturm–Liouville data

The equation treated is the stationary wave equation of a particle of mass m moving on a line in a potential V [557],

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2}(x) + V(x)\,\psi(x) = E\,\psi(x)\,, \tag{9.59}$$

with x in m, m in kg, V and E in J, and $\hbar = 1.054\,571\,817 \times 10^{-34}$ J s exactly (Appendix B). In the template of Definition 9.54 this is a Sturm–Liouville equation with coefficient function 1, potential term $-2mV(x)/\hbar^2$, weight 1 and eigenvalue $\lambda^2 = 2mE/\hbar^2$. The potential term and the eigenvalue both carry the SI dimension of an inverse squared length, as they must, being added to a second derivative with respect to x; the coefficient function and the weight are dimensionless.

Dividing Equation (9.59) by $-\hbar^2/(2m)$ puts it in the form used throughout this section,

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2}(x) + \frac{P(x)^2}{\hbar^2}\,\psi(x) = 0\,, \qquad P(x) = \sqrt{2m\,[E - V(x)]}\,, \tag{9.60}$$

where $P(x)$, in kg m/s, is the momentum a classical particle of energy E would have at x. It is real on the *classically allowed* set $\{x : V(x) < E\}$ and imaginary on the *classically forbidden* set $\{x : V(x) > E\}$, where the real quantity

$$\kappa(x) = \sqrt{2m\,[V(x) - E]} = |P(x)| \tag{9.61}$$

is used instead. A point a with $V(a) = E$, where the two sets meet, is a *turning point*: classically the particle stops there and reverses.

Remark 9.59 (Notation). Throughout this section P denotes the local momentum Equation (9.60) and not the coefficient function called p in Definition 9.54, which for this equation is the constant 1 and plays no further role. The capital is a deliberate departure from the usual lower-case p of the semiclassical literature, adopted here only to keep the two meanings apart inside one chapter.

9.6.2 The expansion in powers of $\hbar$

Write the solution as a single exponential,

$$\psi(x) = \exp\left[\frac{\mathrm{i}\,S(x)}{\hbar}\right], \tag{9.62}$$

with S carrying the SI dimension of an action, J s, so that the exponent is dimensionless. Nothing is approximated yet: any nowhere-vanishing ψ can be written this way with a complex S.

Proposition 9.60 (Exact Riccati form)**.** *The function ψ of Equation (9.62) solves Equation (9.60) if and only if*

$$\left(\frac{\mathrm{d}S}{\mathrm{d}x}\right)^2 - \mathrm{i}\hbar\,\frac{\mathrm{d}^2 S}{\mathrm{d}x^2} = P(x)^2\,. \tag{9.63}$$

Rests on Equations (9.60) and (9.62).

Derivation. *Derives Proposition 9.60.*

Differentiating Equation (9.62) twice,

$$\frac{\mathrm{d}\psi}{\mathrm{d}x} = \frac{\mathrm{i}}{\hbar}\frac{\mathrm{d}S}{\mathrm{d}x}\,\psi\,, \qquad \frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} = \left[\frac{\mathrm{i}}{\hbar}\frac{\mathrm{d}^2 S}{\mathrm{d}x^2} - \frac{1}{\hbar^2}\left(\frac{\mathrm{d}S}{\mathrm{d}x}\right)^2\right]\psi\,.$$

Substituting into Equation (9.60) and cancelling the nowhere-vanishing factor ψ leaves

$$\frac{\mathrm{i}}{\hbar}\frac{\mathrm{d}^2 S}{\mathrm{d}x^2} - \frac{1}{\hbar^2}\left(\frac{\mathrm{d}S}{\mathrm{d}x}\right)^2 + \frac{P^2}{\hbar^2} = 0\,,$$

and multiplication by $-\hbar^2$ gives Equation (9.63). □

The substitution has traded a linear equation of second order for a non-linear one of first order in $\mathrm{d}S/\mathrm{d}x$ — a Riccati equation — and that is a bad bargain except for one feature: $\hbar$ now appears in exactly one place, multiplying the highest derivative. The equation is therefore a singular perturbation problem in $\hbar$, and its natural treatment is a power series in that constant,

$$S(x) = S_0(x) + \hbar\,S_1(x) + \hbar^2 S_2(x) + \cdots\,, \tag{9.64}$$

in which S_0 has the dimension of an action, S_1 is dimensionless, S_2 has the dimension of a reciprocal action, and so on. Equation (9.64) is the one place in this treatise where a physical constant is treated as a formal expansion parameter; the whole structure of the method is the ordering of Equation (9.63) by powers of $\hbar$, so a convention that gave that constant the numerical value one would not simplify the argument but abolish it.

Proposition 9.61 (Eikonal and transport equations)**.** *Substituting Equation (9.64) into Equation (9.63) and collecting powers of $\hbar$ gives, at the first three orders,*

$$\left(\frac{\mathrm{d}S_0}{\mathrm{d}x}\right)^2 = P(x)^2 \qquad \textit{(eikonal)}\,, \tag{9.65}$$

$$2\,\frac{\mathrm{d}S_0}{\mathrm{d}x}\,\frac{\mathrm{d}S_1}{\mathrm{d}x} = \mathrm{i}\,\frac{\mathrm{d}^2 S_0}{\mathrm{d}x^2} \qquad \textit{(transport)}\,, \tag{9.66}$$

$$2\,\frac{\mathrm{d}S_0}{\mathrm{d}x}\,\frac{\mathrm{d}S_2}{\mathrm{d}x} + \left(\frac{\mathrm{d}S_1}{\mathrm{d}x}\right)^2 = \mathrm{i}\,\frac{\mathrm{d}^2 S_1}{\mathrm{d}x^2} \qquad \textit{(first correction)}\,. \tag{9.67}$$

Their solutions are

$$S_0(x) = \pm\int_{x_0}^{x} P(x')\,\mathrm{d}x'\,, \qquad S_1(x) = \frac{\mathrm{i}}{2}\,\ln P(x) + \textit{const}\,. \tag{9.68}$$

Rests on Proposition 9.60 and Equation (9.64).

Derivation. *Derives Proposition 9.61.*

Insert Equation (9.64) in Equation (9.63). The squared term contributes $(S_0')^2 + 2\hbar S_0' S_1' + \hbar^2\left[(S_1')^2 + 2S_0' S_2'\right] + \cdots$, and the second term contributes $-\mathrm{i}\hbar S_0'' - \mathrm{i}\hbar^2 S_1'' - \cdots$; the right-hand side P^2 is of order $\hbar^0$. Equating coefficients of $\hbar^0$, $\hbar^1$ and $\hbar^2$ gives Equations (9.65) to (9.67). The eikonal equation is solved by quadrature, with the two signs corresponding to the two directions of classical motion, which is the first equality of Equation (9.68). Dividing Equation (9.66) by $2S_0'$ turns it into

$$\frac{\mathrm{d}S_1}{\mathrm{d}x} = \frac{\mathrm{i}}{2}\,\frac{S_0''}{S_0'} = \frac{\mathrm{i}}{2}\,\frac{\mathrm{d}}{\mathrm{d}x}\ln\left|\frac{\mathrm{d}S_0}{\mathrm{d}x}\right|\,,$$

and $|S_0'| = P$ by Equation (9.65), so integrating gives the second equality. □

Corollary 9.62 (The semiclassical wave)**.** *Truncating Equation* (9.64) *after the term of order ħ, the general solution of Equation* (9.60) *in a classically allowed region free of turning points is*

$$\psi(x) \approx \frac{C_+}{\sqrt{P(x)}} \exp\left[\frac{\mathrm{i}}{\hbar}\int_{x_0}^{x} P\,\mathrm{d}x'\right] + \frac{C_-}{\sqrt{P(x)}} \exp\left[-\frac{\mathrm{i}}{\hbar}\int_{x_0}^{x} P\,\mathrm{d}x'\right], \tag{9.69}$$

and in a classically forbidden region, with κ of Equation (9.61)*,*

$$\psi(x) \approx \frac{D_+}{\sqrt{\kappa(x)}} \exp\left[\frac{1}{\hbar}\int_{x_0}^{x} \kappa\,\mathrm{d}x'\right] + \frac{D_-}{\sqrt{\kappa(x)}} \exp\left[-\frac{1}{\hbar}\int_{x_0}^{x} \kappa\,\mathrm{d}x'\right]. \tag{9.70}$$

Rests on Proposition 9.61 and Equations (9.61) *and* (9.62)*.*

Derivation. *Derives Corollary 9.62.*

By Equation (9.62) and Equation (9.68),

$$\psi = \exp\left[\frac{\mathrm{i}S_0}{\hbar}\right] \exp\left[\mathrm{i}S_1\right] = \exp\left[\pm\frac{\mathrm{i}}{\hbar}\int^{x} P\,\mathrm{d}x'\right] \exp\left[-\tfrac{1}{2}\ln P\right],$$

because $\mathrm{i}S_1 = \mathrm{i}\cdot\frac{\mathrm{i}}{2}\ln P = -\frac{1}{2}\ln P$, and the second factor is $P^{-1/2}$. Linear superposition of the two signs gives Equation (9.69). In a forbidden region $P = \mathrm{i}\kappa$, the two exponents become real, and absorbing the constant phase of $P^{-1/2}$ into the coefficients gives Equation (9.70). □

The amplitude factor is not decoration. It is the exact content of the transport equation, and it says something physical.

Proposition 9.63 (The transport equation is flux conservation)**.** *For the single travelling wave* $\psi = C_+ P^{-1/2}\exp\left[(\mathrm{i}/\hbar)\int^x P\,\mathrm{d}x'\right]$ *of Equation* (9.69)*, the probability current*

$$j = \frac{\hbar}{m}\,\Im\left[\overline{\psi(x)}\,\frac{\mathrm{d}\psi}{\mathrm{d}x}(x)\right] = \frac{|C_+|^2}{m} \tag{9.71}$$

is independent of x.

Rests on Corollary 9.62.

Derivation. *Derives Proposition 9.63.*

Write $\psi = C_+ P^{-1/2}\mathrm{e}^{\mathrm{i}\theta}$ with $\theta(x) = \hbar^{-1}\int^x P\,\mathrm{d}x'$, so that $\mathrm{d}\theta/\mathrm{d}x = P/\hbar$. Then

$$\overline{\psi}\,\frac{\mathrm{d}\psi}{\mathrm{d}x} = |C_+|^2\left[-\frac{1}{2}\,\frac{1}{P^2}\,\frac{\mathrm{d}P}{\mathrm{d}x} + \frac{\mathrm{i}}{\hbar}\right],$$

the first term inside the bracket being real because P is real in the allowed region. Taking the imaginary part and multiplying by $\hbar/m$ gives Equation (9.71). Had the amplitude been anything other than $P^{-1/2}$ the bracket would have carried an x-dependent imaginary part and the current would not have been constant, which for a stationary state on a line is impossible. □

If ψ is normalised so that $|\psi|^2$ is a probability per unit length, in 1/m, then $|C_+|^2$ is in kg/s and j in 1/s: a probability per unit time, as a current must be. Since the classical speed is $v = P/m$, Equation (9.71) reads $|\psi|^2 v =$ constant — the particle is most likely to be found where it moves slowest, which is the classical statement the approximation reproduces.

9.6.3 Where the approximation is valid, and where it is not

Definition 9.64 (Reduced de Broglie wavelength)**.** For the local momentum of Equation (9.60),

$$\bar{\lambda}(x) = \frac{\hbar}{P(x)}\,, \tag{9.72}$$

in m: the de Broglie wavelength [521] of the local motion, divided by 2π.

Rests on Equation (9.60).

Proposition 9.65 (Validity criterion)**.** *The term of order $\hbar$ in Equation (9.64) is a small correction to the leading term, in the sense that $|\hbar\,\mathrm{d}S_1/\mathrm{d}x| \ll |\mathrm{d}S_0/\mathrm{d}x|$, if and only if*

$$\left|\frac{\mathrm{d}\bar{\lambda}}{\mathrm{d}x}(x)\right| \ll 1\,, \qquad \textit{equivalently} \qquad \frac{\hbar\, m\,|V'(x)|}{P(x)^3} \ll 1\,. \tag{9.73}$$

Both expressions are dimensionless.

Rests on Proposition 9.61, Equation (9.64), and Definition 9.64.

Derivation. *Derives Proposition 9.65.*

By Equation (9.68), $\hbar\,S_1' = \mathrm{i}\hbar P'/(2P)$ and $S_0' = \pm P$, so

$$\left|\frac{\hbar\,S_1'}{S_0'}\right| = \frac{\hbar\,|P'|}{2P^2} = \frac{1}{2}\left|\frac{\mathrm{d}\bar{\lambda}}{\mathrm{d}x}\right|\,,$$

the last equality because $\mathrm{d}\bar{\lambda}/\mathrm{d}x = -\hbar P'/P^2$ from Equation (9.72). Differentiating $P^2 = 2m(E-V)$ gives $2PP' = -2mV'$, hence $P' = -mV'/P$ and $\hbar\,|P'|/P^2 = \hbar m\,|V'|/P^3$. □

The first form of Equation (9.73) is the statement usually quoted: the local wavelength must change by much less than itself over a distance of one wavelength, so that the wave sees an essentially constant potential over each oscillation. The second form makes the same point in terms of the potential, and shows that the criterion is not a smallness condition on V' alone — a gentle slope still defeats the approximation where P is small.

Remark 9.66 (Failure at a turning point). At a turning point a the momentum vanishes, so $\bar{\lambda}$ diverges and Equation (9.73) fails however smooth V may be. The failure is written on the face of the solution: the amplitudes $P^{-1/2}$ and $\kappa^{-1/2}$ in Equations (9.69) and (9.70) become infinite there, whereas the true solution of Equation (9.59) is perfectly finite. Quantitatively, near a simple turning point — one at which $F = V'(a) \neq 0$ — one has $P(x)^2 \approx 2mF(a-x)$, so the second form of Equation (9.73) behaves as $|a-x|^{-3/2}$ and reaches unity at a distance of order

$$\ell = \left(\frac{\hbar^2}{2m\,|F|}\right)^{1/3}\,, \tag{9.74}$$

in m. Outside a neighbourhood of a few ℓ the semiclassical forms are good; inside it they are worthless, and something else must be put there. That is the business of the next subsection.

Remark 9.67 (The series is asymptotic, not convergent). Equation (9.64) is an asymptotic expansion. Carrying it to high order does not converge to the exact solution and generally makes matters worse beyond an optimal truncation, in the same way as the perturbation series of the anharmonic oscillator [1392]. All statements below therefore use the first two terms only, and the accuracy claimed is the one certified by Proposition 9.65.

9.6.4 Connection across a simple turning point

Let a be a simple turning point, $V(a) = E$ with $F = V'(a) \neq 0$, and take $F > 0$, so that $x < a$ is classically allowed and $x > a$ is forbidden. On the scale ℓ of Equation (9.74) the potential may be replaced by its tangent, $V(x) \approx E + F\,(x - a)$, and Equation (9.60) becomes

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} = \frac{2mF}{\hbar^2}\,(x - a)\,\psi\,. \tag{9.75}$$

From Equation (9.60) (linearise V about the turning point).

Introducing the dimensionless variable $z = (x - a)/\ell$ turns Equation (9.75) into

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}z^2} = z\,\psi\,, \tag{9.76}$$

the *Airy equation*, whose solutions Airy introduced to describe the intensity of light near a caustic — the optical turning point, and the same mathematics [103]. Its two standard solutions are fixed by a contour integral, and it is their behaviour far from the turning point that the connection formulae are made of.

Definition 9.68 (Airy functions)**.** For $z \in \mathbb{C}$,

$$\mathrm{Ai}(z) = \frac{1}{2\pi\mathrm{i}} \int_C \exp\left[\frac{t^3}{3} - z\,t\right] \mathrm{d}t\,, \tag{9.77}$$

where C comes in from infinity along the direction $\arg t = -\pi/3$ and returns to infinity along $\arg t = +\pi/3$, and

$$\mathrm{Bi}(z) = \mathrm{e}^{\mathrm{i}\pi/6}\,\mathrm{Ai}\big(z\,\mathrm{e}^{2\pi\mathrm{i}/3}\big) + \mathrm{e}^{-\mathrm{i}\pi/6}\,\mathrm{Ai}\big(z\,\mathrm{e}^{-2\pi\mathrm{i}/3}\big)\,. \tag{9.78}$$

Proposition 9.69 (The integral solves the Airy equation)**.** *The functions of Definition 9.68 are entire, both solve Equation (9.76), both are real on the real axis, and they are linearly independent.*

Rests on Definition 9.68 and Equation (9.76).

Derivation. *Derives Proposition 9.69.*

The integrand is entire in t and decays superexponentially as $|t| \to \infty$ in each of the three open sectors where $\Re(t^3) < 0$, namely those centred on $\arg t = \pi/3$, $\arg t = \pi$ and $\arg t = -\pi/3$. Both ends of C lie in such sectors, so Equation (9.77) converges and may be differentiated under the integral sign. Since

$$\frac{\mathrm{d}}{\mathrm{d}t} \exp\left[\frac{t^3}{3} - z\,t\right] = \left(t^2 - z\right) \exp\left[\frac{t^3}{3} - z\,t\right],$$

one has

$$\frac{\mathrm{d}^2\mathrm{Ai}}{\mathrm{d}z^2}(z) - z\,\mathrm{Ai}(z) = \frac{1}{2\pi\mathrm{i}} \int_C \left(t^2 - z\right) \exp\left[\frac{t^3}{3} - z\,t\right] \mathrm{d}t = \frac{1}{2\pi\mathrm{i}} \left[\exp\left(\frac{t^3}{3} - z\,t\right)\right]_C = 0\,,$$

the bracket vanishing because both ends of C lie in decay sectors. If $\omega^3 = 1$ then $w(z) = \mathrm{Ai}(\omega z)$ satisfies $\mathrm{d}^2w/\mathrm{d}z^2 = \omega^2\,\mathrm{Ai}''(\omega z) = \omega^3\,z\,\mathrm{Ai}(\omega z) = z\,w$, so each term of Equation (9.78) solves the equation and so does their sum. Reflection: conjugating Equation (9.77) maps C onto itself with reversed orientation and sends the integrand to $\exp[t^3/3 - \bar{z}t]$, while $\overline{1/(2\pi\mathrm{i})} = -1/(2\pi\mathrm{i})$; the two sign changes cancel and $\overline{\mathrm{Ai}(z)} = \mathrm{Ai}(\bar{z})$. Hence Ai is real on the real axis, and for real z the two terms of Equation (9.78) are complex conjugates of each other, so Bi is real there too. Independence follows from Theorem 9.70 below: one solution decays and the other grows as $z \to +\infty$, so neither is a multiple of the other. □

Theorem 9.70 (Asymptotics of the Airy functions)**.** *As the real variable z tends to $+\infty$,*

$$\mathrm{Ai}(z) \sim \frac{1}{2\sqrt{\pi}\, z^{1/4}}\, \mathrm{e}^{-\frac{2}{3}z^{3/2}}, \qquad \mathrm{Ai}(-z) \sim \frac{1}{\sqrt{\pi}\, z^{1/4}} \sin\left(\tfrac{2}{3}z^{3/2} + \tfrac{\pi}{4}\right), \tag{9.79}$$

$$\mathrm{Bi}(z) \sim \frac{1}{\sqrt{\pi}\, z^{1/4}}\, \mathrm{e}^{+\frac{2}{3}z^{3/2}}, \qquad \mathrm{Bi}(-z) \sim \frac{1}{\sqrt{\pi}\, z^{1/4}} \cos\left(\tfrac{2}{3}z^{3/2} + \tfrac{\pi}{4}\right). \tag{9.80}$$

The first of these holds in the whole sector $|\arg z| < \pi$, with the principal branches of $z^{3/2}$ and $z^{1/4}$.

Rests on Definition 9.68 and Proposition 9.69.

Derivation. *Derives Theorem 9.70.*

Write the exponent of Equation (9.77) as $\varphi(t) = t^3/3 - z\,t$, so that $\varphi'(t) = t^2 - z$ and $\varphi''(t) = 2t$; the saddle points are $t = \pm\sqrt{z}$. By Cauchy's theorem and the deformation corollary (Theorem 8.12 and Corollary 8.15) the contour may be moved at will, the integrand being entire, provided its ends stay in the decay sectors. The method of steepest descent chooses the deformation on which $\Im\varphi$ is constant: the integrand then does not oscillate, its modulus has a sharp maximum at the saddle, and the whole contribution comes from a neighbourhood of it whose width shrinks as $|\varphi''|^{-1/2}$.

Positive argument. Let $z = x > 0$. The two saddles are at $t = \pm\sqrt{x}$, with $\varphi(\sqrt{x}) = -\frac{2}{3}x^{3/2}$ and $\varphi(-\sqrt{x}) = +\frac{2}{3}x^{3/2}$. The ends of C lie in the two sectors adjacent to the positive real axis, so C is deformed through the saddle at $t = +\sqrt{x}$ alone. There $\varphi'' = 2\sqrt{x}$ is real and positive, so putting $t = \sqrt{x} + \mathrm{i}u$ with u real makes $\frac{1}{2}\varphi''\,(t-\sqrt{x})^2 = -\sqrt{x}\,u^2$: the vertical line through the saddle is the path of steepest descent. On it $\mathrm{d}t = \mathrm{i}\,\mathrm{d}u$, and

$$\mathrm{Ai}(x) \approx \frac{1}{2\pi\mathrm{i}}\, \mathrm{e}^{-\frac{2}{3}x^{3/2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\sqrt{x}\,u^2}\, \mathrm{i}\,\mathrm{d}u = \frac{\mathrm{e}^{-\frac{2}{3}x^{3/2}}}{2\pi} \sqrt{\frac{\pi}{\sqrt{x}}} = \frac{\mathrm{e}^{-\frac{2}{3}x^{3/2}}}{2\sqrt{\pi}\, x^{1/4}},$$

the first member of Equation (9.79). Nothing in this computation used z real: the same deformation is available, and the same result holds, for every z with $|\arg z| < \pi$.

Negative argument. Let $z = -x$ with $x > 0$, so that $\varphi(t) = t^3/3 + x\,t$ and the saddles sit at $t = \pm\mathrm{i}\sqrt{x}$ with $\varphi(\pm\mathrm{i}\sqrt{x}) = \pm\mathrm{i}\,\frac{2}{3}x^{3/2}$, purely imaginary: neither dominates the other, and both must be kept. The bookkeeping is topological. Call V_1, V_2, V_3 the decay sectors around $\arg t = \pi/3$, π and $-\pi/3$; the contour runs from V_3 to V_1. At a saddle s the steepest descent direction $\mathrm{e}^{\mathrm{i}\theta}$ is fixed by requiring $\frac{1}{2}\varphi''(s)\,\mathrm{e}^{2\mathrm{i}\theta}$ to be real and negative. At $s = -\mathrm{i}\sqrt{x}$, where $\varphi'' = 2\sqrt{x}\,\mathrm{e}^{-\mathrm{i}\pi/2}$, this gives $\theta = -\pi/4$ or $\theta = 3\pi/4$, so that descent path joins V_3 to V_2; at $s = +\mathrm{i}\sqrt{x}$, where $\varphi'' = 2\sqrt{x}\,\mathrm{e}^{\mathrm{i}\pi/2}$, it gives $\theta = \pi/4$ or $\theta = 5\pi/4$, joining V_2 to V_1. Their concatenation runs from V_3 to V_1 and is therefore a legitimate deformation of C. Each saddle contributes $\mathrm{e}^{\varphi(s)}\,\mathrm{e}^{\mathrm{i}\theta}\sqrt{\pi/\sqrt{x}}$ with θ the direction in which it is traversed — $3\pi/4$ at the lower saddle and $\pi/4$ at the upper one. Writing $\zeta = \frac{2}{3}x^{3/2}$,

$$\begin{aligned}\mathrm{Ai}(-x) &\approx \frac{1}{2\pi\mathrm{i}}\sqrt{\frac{\pi}{\sqrt{x}}}\left[\mathrm{e}^{-\mathrm{i}\zeta}\,\mathrm{e}^{3\mathrm{i}\pi/4} + \mathrm{e}^{\mathrm{i}\zeta}\,\mathrm{e}^{\mathrm{i}\pi/4}\right] \\ &= \frac{1}{2\pi\mathrm{i}}\sqrt{\frac{\pi}{\sqrt{x}}}\,\mathrm{e}^{\mathrm{i}\pi/2}\left[\mathrm{e}^{\mathrm{i}(\pi/4-\zeta)} + \mathrm{e}^{-\mathrm{i}(\pi/4-\zeta)}\right] = \frac{\cos\left(\zeta - \pi/4\right)}{\sqrt{\pi}\, x^{1/4}},\end{aligned}$$

and $\cos(\zeta - \pi/4) = \sin(\zeta + \pi/4)$, which is the second member of Equation (9.79).

The second solution. Equation (9.78) reduces both members of Equation (9.80) to the result already obtained. For $z = x > 0$ the two arguments are $x\,\mathrm{e}^{\pm 2\pi\mathrm{i}/3}$, inside the sector $|\arg z| < \pi$ where the first expansion holds; there $\left(x\mathrm{e}^{\pm 2\pi\mathrm{i}/3}\right)^{3/2} = x^{3/2}\mathrm{e}^{\pm\mathrm{i}\pi} = -x^{3/2}$ and $\left(x\mathrm{e}^{\pm 2\pi\mathrm{i}/3}\right)^{1/4} = x^{1/4}\mathrm{e}^{\pm\mathrm{i}\pi/6}$, so each term of Equation (9.78) equals $\mathrm{e}^{\pm\mathrm{i}\pi/6}\,\mathrm{e}^{+\frac{2}{3}x^{3/2}} \,/\, (2\sqrt{\pi}\, x^{1/4}\,\mathrm{e}^{\pm\mathrm{i}\pi/6})$: the phases cancel, the two terms are equal, and their sum is the first

member of Equation (9.80). For $z=-x$ the two arguments are $x\,\mathrm{e}^{\mp\mathrm{i}\pi/3}$, again inside the sector, and now $\left(x\mathrm{e}^{\mp\mathrm{i}\pi/3}\right)^{3/2}=\mp\mathrm{i}\,x^{3/2}$ and $\left(x\mathrm{e}^{\mp\mathrm{i}\pi/3}\right)^{1/4}=x^{1/4}\mathrm{e}^{\mp\mathrm{i}\pi/12}$, so the two terms are $\mathrm{e}^{\pm\mathrm{i}\pi/4}\,\mathrm{e}^{\pm\mathrm{i}\zeta}\,/\,\left(2\sqrt{\pi}\,x^{1/4}\right)$ and their sum is $\cos(\zeta+\pi/4)\,/\,\left(\sqrt{\pi}\,x^{1/4}\right)$, the second member of Equation (9.80). □

Remark 9.71 (What "asymptotic" means here). Each result above is the leading term of a full asymptotic expansion, and the expansion is obtained from the same contour. On the steepest-descent path the exponent is $\varphi(s)+\frac{1}{2}\varphi''(s)(t-s)^2+\frac{1}{6}\varphi'''(s)(t-s)^3$ *exactly*, because φ is a cubic and $\varphi'''=2$; the quadratic term is the Gaussian already used, and expanding the cubic term and integrating the Gaussian moments one by one produces a series in inverse powers of $\zeta=\frac{2}{3}z^{3/2}$, whose first correction to Ai is the relative factor $-5/(72\zeta)$. The series does not converge — it is asymptotic in the sense of Remark 9.67 — but truncating it after any fixed number of terms commits an error of the order of the first term omitted, uniformly on $|\arg z|\le\pi-\delta$ for each fixed $\delta>0$. Only the leading term is used in this chapter, and the connection formulae inherit exactly its accuracy: they are correct to relative order ζ^{-1}, which is small precisely where the semiclassical forms they connect are themselves good.

Theorem 9.72 (Connection formulae)**.** *With a a simple turning point and $F=V'(a)>0$, the semiclassical solutions on the two sides of a are related by*

$$\frac{C}{\sqrt{\kappa(x)}}\exp\left[-\frac{1}{\hbar}\int_a^x\kappa\,\mathrm{d}x'\right]\leftrightarrow\frac{2C}{\sqrt{P(x)}}\sin\left[\frac{1}{\hbar}\int_x^a P\,\mathrm{d}x'+\frac{\pi}{4}\right],\tag{9.81}$$

$$\frac{D}{\sqrt{\kappa(x)}}\exp\left[+\frac{1}{\hbar}\int_a^x\kappa\,\mathrm{d}x'\right]\leftrightarrow\frac{D}{\sqrt{P(x)}}\cos\left[\frac{1}{\hbar}\int_x^a P\,\mathrm{d}x'+\frac{\pi}{4}\right],\tag{9.82}$$

the left-hand member holding for $x>a$ and the right-hand member for $x<a$. If instead $F<0$, so that the allowed region lies to the right of a, the same formulae hold with the roles of the two sides interchanged and every integral taken outward from a.

Rests on Corollary 9.62 and Equations (9.74), (9.76), (9.79) and (9.80).

Derivation. *Derives Theorem 9.72.*

Work with the linearised equation, for which the exact solution is a combination of $\mathrm{Ai}(z)$ and $\mathrm{Bi}(z)$, and evaluate the two sides of Equation (9.81) on it. For $x>a$, that is $z>0$, Equation (9.61) gives $\kappa(x)=\sqrt{2mF(x-a)}=\sqrt{2mF\ell}\,z^{1/2}$, hence

$$\frac{1}{\hbar}\int_a^x\kappa\,\mathrm{d}x'=\frac{\ell^{3/2}\sqrt{2mF}}{\hbar}\int_0^z\sqrt{z'}\,\mathrm{d}z'=\frac{2}{3}\,z^{3/2},$$

because $\ell^{3/2}=\hbar/\sqrt{2mF}$ by Equation (9.74). So on the forbidden side the decaying semiclassical form is $C\,(2mF\ell)^{-1/4}z^{-1/4}\mathrm{e}^{-\frac{2}{3}z^{3/2}}$, which by Equation (9.79) is $2\sqrt{\pi}\,C\,(2mF\ell)^{-1/4}$ times $\mathrm{Ai}(z)$. For $x<a$ put $z=-u$ with $u>0$; then $P(x)=\sqrt{2mF\ell}\,u^{1/2}$ and, by the same computation, $\hbar^{-1}\int_x^a P\,\mathrm{d}x'=\frac{2}{3}u^{3/2}$. The second relation of Equation (9.79) therefore gives

$$2\sqrt{\pi}\,C\,(2mF\ell)^{-1/4}\,\mathrm{Ai}(-u)\sim 2C\,(2mF\ell)^{-1/4}u^{-1/4}\sin\left(\tfrac{2}{3}u^{3/2}+\tfrac{\pi}{4}\right)=\frac{2C}{\sqrt{P}}\sin\left[\frac{1}{\hbar}\int_x^a P\,\mathrm{d}x'+\frac{\pi}{4}\right],$$

which is Equation (9.81). The factor 2 is nothing but the ratio of the two prefactors in Equation (9.79), and the phase $\pi/4$ is the constant carried across by the same pair. Repeating the computation with Bi in place of Ai, whose two prefactors in Equation (9.80) are equal, gives Equation (9.82) with no factor of 2. The case $F<0$ is the reflection $x\mapsto 2a-x$, which leaves Equation (9.60) invariant. □

Remark 9.73 (The formulae are not symmetric in practice). Equation (9.81) may be read reliably only from left to right, and Equation (9.82) only from right to left. The reason is that an exponentially small admixture of the growing solution is invisible beside the

decaying one on the forbidden side, yet dominates once traced back to the allowed side; the asymptotic representation of a solution therefore changes discontinuously as the turning point is crossed in the complex plane. This is the Stokes phenomenon, and ignoring it is the commonest error in applications of the method. Every connection made below is therefore labelled with its direction: the quantisation condition of Section 9.6.7 uses only the safe one, and the single delicate step in the tunnelling derivation is flagged where it occurs and its cost accounted for in Remark 9.77.

9.6.5 Turning points of higher order

The linearisation Equation (9.75) presumed $F = V'(a) \neq 0$. Where the potential meets the energy level tangentially that derivative vanishes too and the Airy problem is empty, so let $\mu \geq 1$ be the order of contact,

$$V(x) - E \approx F_\mu \, (x-a)^\mu \,, \qquad F_\mu = \frac{V^{(\mu)}(a)}{\mu!} \neq 0 \,, \tag{9.83}$$

F_μ carrying the SI unit of an energy divided by the μ-th power of a length, so that F_1 is the force F of Section 9.6.4 in J/m. The odd and even cases are genuinely different, and only the odd one is a turning point at all.

Proposition 9.74 (Connection at a turning point of odd order)**.** *Let μ be odd and $F_\mu > 0$ in Equation (9.83), so that $x < a$ is classically allowed and $x > a$ forbidden. Then the decaying solution on the forbidden side connects to*

$$\frac{C}{\sqrt{\kappa(x)}} \exp\left[-\frac{1}{\hbar}\int_a^x \kappa \,\mathrm{d}x'\right] \quad \leftrightarrow \quad \frac{C}{\sin\left[\pi/\left(2\mu+4\right)\right]} \frac{1}{\sqrt{P(x)}} \sin\left[\frac{1}{\hbar}\int_x^a P \,\mathrm{d}x' + \frac{\pi}{4}\right] . \tag{9.84}$$

The phase $\pi/4$ is the same as in Equation (9.81) for every odd μ; only the amplitude changes, and at $\mu = 1$ the factor is $1/\sin(\pi/6) = 2$, which recovers Theorem 9.72.

Rests on Theorem 9.72 and Equation (9.83).

Derivation. *Derives Proposition 9.74.*

Put $s = x - a$ and $k = \sqrt{2mF_\mu}/\hbar$, a positive constant. Near a, Equation (9.60) reads

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}s^2} = k^2 \, s^\mu \, \psi \,,$$

and for $s > 0$ the substitution $\zeta = 2k\, s^q/(\mu+2)$ with $q = (\mu+2)/2$, together with $\psi = \sqrt{s}\,\chi(\zeta)$, turns it by a direct computation into the modified Bessel equation of order $\nu = 1/(\mu+2)$. A basis of local solutions is therefore $\sqrt{s}\, I_{\pm\nu}(\zeta)$, and the decaying combination is $\sqrt{s}\,[I_{-\nu}(\zeta) - I_\nu(\zeta)]$, proportional to $\sqrt{s}\, K_\nu(\zeta)$. Note that $\zeta = \hbar^{-1}\int_a^x \kappa \,\mathrm{d}x'$ exactly for the local potential, since $\kappa = \hbar k\, s^{\mu/2}$ there.

Continue to the allowed side by rotating $s = u\,\mathrm{e}^{\mathrm{i}\pi}$ with $u = a - x > 0$. Then $\zeta = w\,\mathrm{e}^{\mathrm{i}\pi q}$ with $w = 2k\,u^q/(\mu+2) = \hbar^{-1}\int_x^a P \,\mathrm{d}x'$. In the power series

$$I_\sigma(\zeta) = \sum_{j\geq 0} \frac{1}{j!\,\Gamma(j+\sigma+1)} \left(\frac{\zeta}{2}\right)^{2j+\sigma}$$

the rotation multiplies the j-th term by $\mathrm{e}^{2\pi\mathrm{i}qj} = \mathrm{e}^{\mathrm{i}\pi(\mu+2)j} = (-1)^j$, because $\mu+2$ is odd; and an alternating sign is exactly what turns the modified Bessel series into the ordinary one. Hence $I_\sigma(w\,\mathrm{e}^{\mathrm{i}\pi q}) = \mathrm{e}^{\mathrm{i}\pi q\sigma}\, J_\sigma(w)$. With $q\nu = \tfrac{1}{2}$ and $\sqrt{s} = \mathrm{e}^{\mathrm{i}\pi/2}\sqrt{u}$ the two basis solutions continue to $\mathrm{e}^{\mathrm{i}\pi/2}\mathrm{e}^{\pm\mathrm{i}\pi/2}\sqrt{u}\, J_{\pm\nu}(w)$, that is to $-\sqrt{u}\, J_\nu(w)$ and $+\sqrt{u}\, J_{-\nu}(w)$, so the decaying solution continues to $\sqrt{u}\,[J_{-\nu}(w) + J_\nu(w)]$. This much is exact.

It remains to read off the two amplitudes, and for that one classical fact is imported: the large-argument behaviour of the Bessel functions, $K_\nu(\zeta) \sim \sqrt{\pi/(2\zeta)}\ \mathrm{e}^{-\zeta}$ and $J_\sigma(w) \sim \sqrt{2/(\pi w)}\cos(w - \sigma\pi/2 - \pi/4)$ [1023]. The prefactors $\sqrt{s/\zeta}$ and $\sqrt{u/w}$ are the same function of $|s|$ with the same constant in front, and each reproduces the semiclassical amplitude, $\kappa^{-1/2}$ on one side and $P^{-1/2}$ on the other, so only the pure numbers matter. On the forbidden side $I_{-\nu} - I_\nu = (2\sin\nu\pi/\pi)\,K_\nu$ leaves the number $(2\sin\nu\pi/\pi)\sqrt{\pi/2} = \sqrt{2/\pi}\,\sin\nu\pi$. On the allowed side

$$J_\nu(w) + J_{-\nu}(w) \sim \sqrt{\frac{2}{\pi w}}\,2\cos\left(\frac{\nu\pi}{2}\right)\cos\left(w - \frac{\pi}{4}\right),$$

the two phase shifts $\mp\nu\pi/2$ averaging away and leaving the universal $-\pi/4$. The ratio of the two numbers is

$$\frac{2\cos(\nu\pi/2)\sqrt{2/\pi}}{\sqrt{2/\pi}\,\sin\nu\pi} = \frac{2\cos(\nu\pi/2)}{2\sin(\nu\pi/2)\cos(\nu\pi/2)} = \frac{1}{\sin(\nu\pi/2)},$$

which with $\nu = 1/(\mu+2)$ is the factor in Equation (9.84); and $\cos(w-\pi/4) = \sin(w+\pi/4)$ completes it. □

Remark 9.75 (Even order is tangency, not a turning point). For even μ the difference $V - E$ does not change sign at a, so a is not a turning point at all: the motion is either forbidden on both sides of it or allowed on both, and there are no two regimes to connect. The rotation argument above fails at exactly that point — $\mu+2$ is then even, $\mathrm{e}^{2\pi\mathrm{i}qj} = +1$, and the modified Bessel series is not turned into the ordinary one. The instance that matters physically is $\mu = 2$ with $F_2 < 0$: the top of a barrier sitting exactly at the energy of the incident particle. There the local equation is the parabolic-cylinder equation rather than Equation (9.76), and the transmission coefficient passes smoothly through the barrier top instead of jumping. That regime is outside the linear connection theory of this section, and Remark 9.77 has already recorded that $\gamma \lesssim 1$ is precisely where the formulae derived here are not entitled to be used.

The other place where the semiclassical treatment fails for a reason that is not a simple turning point is the origin of the radial equation, where the centrifugal term diverges. That case has its own repair, and it is derived in Section 9.6.10.

9.6.6 Barrier penetration and the Gamow factor

Theorem 9.76 (Semiclassical transmission through a barrier)**.** *Let $V > E$ on an interval (a,b) and $V < E$ outside it, with simple turning points at a and b. Write*

$$\gamma = \frac{1}{\hbar}\int_a^b \sqrt{2m\,[V(x) - E]}\,\mathrm{d}x\,, \tag{9.85}$$

a dimensionless number, the integrand being in kg m/s *and $\hbar$ in* J s*. Then the transmission coefficient — the ratio of transmitted to incident probability current — is*

$$T = \left(\mathrm{e}^{\gamma} + \tfrac{1}{4}\mathrm{e}^{-\gamma}\right)^{-2}, \tag{9.86}$$

which for a thick barrier, $\gamma \gg 1$, reduces to

$$T \approx \mathrm{e}^{-2\gamma} = \exp\left[-\frac{2}{\hbar}\int_a^b \sqrt{2m\,[V(x) - E]}\,\mathrm{d}x\right], \tag{9.87}$$

the correction being of relative size $\mathrm{e}^{-2\gamma}/2$. The exponent 2γ is the Gamow factor.

Rests on Corollary 9.62, Theorem 9.72, and Proposition 9.63.

Derivation. *Derives Theorem 9.76.*

Beyond the barrier, for $x > b$, impose a purely outgoing wave and write it with the phase measured from b,

$$\psi(x) = \frac{\mathcal{A}}{\sqrt{P(x)}} \exp\left\{ \mathrm{i} \left[\frac{1}{\hbar} \int_b^x P \, \mathrm{d}x' + \frac{\pi}{4} \right] \right\} = \frac{\mathcal{A}}{\sqrt{P}} \left[\cos\Theta + \mathrm{i} \sin\Theta \right] ,$$

with $\Theta(x) = \hbar^{-1} \int_b^x P \, \mathrm{d}x' + \pi/4$. At b the forbidden region lies to the left, so Theorem 9.72 is used in its reflected form. The direction has to be said carefully, because "growing towards b" and "growing away from b" name opposite solutions: measured *into* the barrier, the way Equations (9.81) and (9.82) measure it, the $\cos\Theta$ part comes from the exponential growing away from b and the $\sin\Theta$ part from the one decaying away from it, with the coefficients read off Equation (9.82) and Equation (9.81) respectively. Hence inside the barrier

$$\psi(x) = \frac{\mathcal{A}}{\sqrt{\kappa}} \exp\left[+\frac{1}{\hbar} \int_x^b \kappa \, \mathrm{d}x' \right] + \frac{\mathrm{i}\mathcal{A}/2}{\sqrt{\kappa}} \exp\left[-\frac{1}{\hbar} \int_x^b \kappa \, \mathrm{d}x' \right] .$$

The first of these two readings is in the safe direction of Remark 9.73; the second is not, and it is the sole source of the order-unity ambiguity recorded in Remark 9.77 — it is the term that carries the coefficient $\frac{1}{4}$ into Equation (9.86), and it does not affect the exponent.

Now re-express the exponents about the other turning point using $\int_x^b \kappa = \hbar\gamma - \int_a^x \kappa$, which turns the first term into $\mathcal{A}\mathrm{e}^{\gamma}$ times the exponential decaying away from a and the second into $(\mathrm{i}\mathcal{A}/2)\mathrm{e}^{-\gamma}$ times the one growing away from a. At a the allowed region lies to the left, which is exactly the configuration of Theorem 9.72, so with $\Xi(x) = \hbar^{-1} \int_x^a P \, \mathrm{d}x' + \pi/4$,

$$\psi(x) = \frac{1}{\sqrt{P}} \left[2\mathcal{A}\mathrm{e}^{\gamma} \sin\Xi + \frac{\mathrm{i}\mathcal{A}}{2} \mathrm{e}^{-\gamma} \cos\Xi \right] , \qquad x < a \, .$$

Writing $\sin\Xi = (\mathrm{e}^{\mathrm{i}\Xi} - \mathrm{e}^{-\mathrm{i}\Xi})/(2\mathrm{i})$ and $\cos\Xi = (\mathrm{e}^{\mathrm{i}\Xi} + \mathrm{e}^{-\mathrm{i}\Xi})/2$ and collecting,

$$\psi(x) = \frac{\mathrm{i}\mathcal{A}}{\sqrt{P}} \left[\left(\mathrm{e}^{\gamma} + \tfrac{1}{4}\mathrm{e}^{-\gamma} \right) \mathrm{e}^{-\mathrm{i}\Xi} - \left(\mathrm{e}^{\gamma} - \tfrac{1}{4}\mathrm{e}^{-\gamma} \right) \mathrm{e}^{\mathrm{i}\Xi} \right] .$$

Since $\mathrm{d}\Xi/\mathrm{d}x = -P/\hbar$, the term in $\mathrm{e}^{-\mathrm{i}\Xi}$ advances in phase with increasing x and is the incident wave, while $\mathrm{e}^{\mathrm{i}\Xi}$ is the reflected one. The amplitude $P^{-1/2}$ has already made $|\cdot|^2$ proportional to the current, by Proposition 9.63, so the transmission coefficient is the ratio of the squared coefficients of the transmitted and incident waves,

$$T = \frac{|\mathcal{A}|^2}{|\mathcal{A}|^2 \left(\mathrm{e}^{\gamma} + \frac{1}{4}\mathrm{e}^{-\gamma} \right)^2} ,$$

which is Equation (9.86). Expanding, $T = \mathrm{e}^{-2\gamma}\left(1 + \frac{1}{4}\mathrm{e}^{-2\gamma}\right)^{-2} = \mathrm{e}^{-2\gamma}\left[1 - \frac{1}{2}\mathrm{e}^{-2\gamma} + \cdots\right]$. □

Remark 9.77 (How much of the prefactor is real). Only the exponential in Equation (9.87) is a reliable output of the method. The order-unity factor multiplying it depends on how the two turning points are treated, and Equation (9.86) is not exact: the linear connection formulae assume the two turning points are far apart on the scale ℓ of Equation (9.74), and a uniform treatment that does not gives $T = \left(1 + \mathrm{e}^{2\gamma}\right)^{-1}$ instead. The place where the two disagree is worth being exact about, because it is the opposite of what one might guess. Both tend to $\mathrm{e}^{-2\gamma}$ with the *same* unit prefactor as γ grows, and they are numerically indistinguishable in the regime a barrier problem usually lives in: their ratio is 1.009 at $\gamma = 2$, and 1 to ten significant figures at the $\gamma = 33.6$ of Example 9.81. They part company only near the top of the barrier, $\gamma \lesssim 1$, where the ratio reaches 1.28 at $\gamma = 0$ — and that is precisely the regime in which the two turning points are no longer far apart on the scale ℓ, so the linear connection formulae were never entitled to be used there. The exponent is the whole physical content, and it is the part the method actually determines.

The exponential is the result Gamow, and independently Gurney and Condon, put to work on the escape of a charged particle from a nucleus [600, 603, 625]; the same expression, with a triangular barrier, is Fowler and Nordheim's account of electron emission from a cold metal in a strong field [599]. What this section supplies is the mathematics: a formula for the exponent, its derivation, and an honest statement of what it does and does not determine.

9.6.7 Bohr–Sommerfeld quantisation and the Maslov index

Theorem 9.78 (Semiclassical quantisation of a well)**.** *Let $V < E$ on an interval (a, b) and $V > E$ outside it, with simple turning points at a and b. A solution of Equation (9.59) that decays in both forbidden regions exists only for those energies satisfying*

$$\frac{1}{\hbar}\int_a^b P(x)\,\mathrm{d}x = \left(n+\frac{1}{2}\right)\pi\,, \qquad n = 0, 1, 2, \dots\,. \tag{9.88}$$

Equivalently, in terms of the action accumulated over one full classical period, with $h = 2\pi\hbar$ the Planck constant,

$$\oint P\,\mathrm{d}x = 2\int_a^b P(x)\,\mathrm{d}x = \left(n+\frac{\mu}{4}\right)h\,, \qquad \mu = 2\,. \tag{9.89}$$

Rests on Equation (9.59), Theorem 9.72, and Corollary 9.62.

Derivation. Derives Theorem 9.78.

Decay in the left forbidden region, $x < a$, is the configuration of Theorem 9.72 reflected, and connects to

$$\psi(x) = \frac{2C}{\sqrt{P}}\sin\left[\frac{1}{\hbar}\int_a^x P\,\mathrm{d}x' + \frac{\pi}{4}\right]$$

inside the well. Decay in the right forbidden region, $x > b$, connects by Equation (9.81) to

$$\psi(x) = \frac{2C'}{\sqrt{P}}\sin\left[\frac{1}{\hbar}\int_x^b P\,\mathrm{d}x' + \frac{\pi}{4}\right]\,.$$

These are two expressions for the same function, so they must agree up to sign for every x in (a, b). Write $\theta(x) = \hbar^{-1}\int_a^x P\,\mathrm{d}x'$ and $\Phi = \hbar^{-1}\int_a^b P\,\mathrm{d}x'$, so that $\hbar^{-1}\int_x^b P\,\mathrm{d}x' = \Phi - \theta$, and expand

$$\sin\left(\Phi - \theta + \tfrac{\pi}{4}\right) = \sin\left(\Phi + \tfrac{\pi}{2}\right)\cos\left(\theta + \tfrac{\pi}{4}\right) - \cos\left(\Phi + \tfrac{\pi}{2}\right)\sin\left(\theta + \tfrac{\pi}{4}\right)\,.$$

For this to be a constant multiple of $\sin(\theta + \pi/4)$ for all θ, the coefficient of the independent function $\cos(\theta + \pi/4)$ must vanish, that is $\cos\Phi = 0$, whence $\Phi = (n + \frac{1}{2})\pi$ with n a non-negative integer, which is Equation (9.88). The remaining coefficient is then $-\cos(\Phi + \pi/2) = \sin\Phi = (-1)^n$, so $C' = (-1)^n C$ and the two expressions do agree. Doubling and multiplying by $\hbar$ gives $\oint P\,\mathrm{d}x = (2n+1)\pi\hbar = (n + \frac{1}{2})h$, which is Equation (9.89). □

Remark 9.79 (Where the one half comes from)*.* The old quantum theory quantised the phase integral as $\oint P\,\mathrm{d}x = nh$ [441], generalising Bohr's quantisation of the angular momentum of a circular orbit [395], and had no account of the additional $h/2$. Theorem 9.78 shows that it is not an empirical patch but an accounting of turning points: each of the two simple turning points contributed a phase $\pi/4$ through Equation (9.81), and $2 \times \pi/4 = \pi/2$ is exactly the half-integer. The integer μ of Equation (9.89) counts those contributions — one per turning point for a one-dimensional well, so $\mu = 2$ — and is the *Maslov index*. Keller identified the same index as the count of caustics traversed by the classical trajectory, and thereby extended the condition to systems of more than one degree of freedom, where the quantisation applies to each independent circuit of an invariant torus [1150]. A turning point at an infinitely steep wall, where the wave function must simply vanish, contributes $\pi/2$ rather than $\pi/4$ and so counts twice in μ.

Example 9.80 (The harmonic oscillator)**.** For $V(x) = \frac{1}{2}m\omega^2 x^2$, with m in kg and ω in rad/s, the turning points of energy E are at $x = \pm x_E$ with $x_E = \sqrt{2E/(m\omega^2)}$, and

$$\oint P\,\mathrm{d}x = 2\int_{-x_E}^{x_E} \sqrt{2mE - m^2\omega^2 x^2}\,\mathrm{d}x = \frac{2\pi E}{\omega}\,, \tag{9.90}$$

the integral being the area of an ellipse with semi-axes x_E and $\sqrt{2mE}$ in the phase plane. Equation (9.89) then gives $2\pi E/\omega = (n + \frac{1}{2})\,2\pi\hbar$, that is

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega\,, \qquad n = 0, 1, 2, \ldots\,, \tag{9.91}$$

in J. These are the exact eigenvalues of Equation (9.59) for this potential, zero-point energy included — an accident of the harmonic case, not a general property of the approximation, but a reassuring one. For a vibrational mode of angular frequency $\omega = 5.0 \times 10^{14}\,\mathrm{rad/s}$, typical of a diatomic molecule, the spacing is $\hbar\omega = 5.3 \times 10^{-20}\,\mathrm{J} = 0.33\,\mathrm{eV}$.

Rests on Theorem 9.78.

9.6.8 Worked example: a charged particle against a Coulomb barrier

Example 9.81 (Gamow factor for an alpha particle, in SI)**.** Take a particle of charge ze and mass m inside a spherical region of radius R whose boundary it must cross against the electrostatic repulsion of a residual charge Ze. Outside R the potential energy is

$$V(r) = \frac{zZe^2}{4\pi\varepsilon_0\, r}\,, \qquad r > R\,, \tag{9.92}$$

in J, with $e = 1.602\,176\,634 \times 10^{-19}\,\mathrm{C}$ exactly and $\varepsilon_0 = 8.854\,187\,812\,8 \times 10^{-12}\,\mathrm{F/m}$ [2392]. Take the alpha particle, $z = 2$ and $m_\alpha = 6.644\,657\,335\,7 \times 10^{-27}\,\mathrm{kg}$ [2392], with $Z = 82$, $R = 9.0\,\mathrm{fm}$ and a kinetic energy at infinity of

$$E = 5.00\,\mathrm{MeV} = 8.0109 \times 10^{-13}\,\mathrm{J}\,.$$

The charge factor is $zZe^2/(4\pi\varepsilon_0) = 3.7836 \times 10^{-26}\,\mathrm{J\,m}$, so the outer turning point, where $V(b) = E$, sits at

$$b = \frac{zZe^2}{4\pi\varepsilon_0 E} = 4.7231 \times 10^{-14}\,\mathrm{m} = 47.23\,\mathrm{fm}\,,$$

five times the radius of the source region. Only b is a turning point here: at $r = R$ the model potential is cut off discontinuously by the nuclear surface, and $V(R) = 26.24\,\mathrm{MeV}$ is far above E (Example 9.82), so the momentum does not vanish there and no connection formula can be applied at R. That puts the configuration *outside* the hypothesis of Theorem 9.76, which asks for simple turning points at both ends of the barrier and performs one connection at each: here the inner end is a hard wall, off which the incident amplitude is reflected rather than matched to an Airy function. What is taken from that theorem is therefore its exponent alone. The barrier integral Equation (9.85) is taken from the edge of the source region to the single turning point b, while the order-unity factor multiplying $\mathrm{e}^{-2\gamma}$ — which Equation (9.86) determines only in the two-turning-point case, and there only to the accuracy Remark 9.77 records — is not fixed by anything proved here. Since $V(r) - E = E\,(b/r - 1)$ by Equation (9.92), the exponent Equation (9.85) taken between R and b is $\hbar^{-1}\sqrt{2mE}\int_R^b \sqrt{b/r - 1}\,\mathrm{d}r$, and the substitution $r = b\cos^2\vartheta$ evaluates it in closed form: with $\xi = R/b = 0.1906$,

$$\gamma = \frac{\sqrt{2mE}\,b}{\hbar}\left[\arccos\sqrt{\xi} - \sqrt{\xi\,(1-\xi)}\right]. \tag{9.93}$$

Numerically $\sqrt{2mE} = 1.0318 \times 10^{-19}\,\mathrm{kg\,m/s}$, $\arccos\sqrt{\xi} = 1.1191$, $\sqrt{\xi(1-\xi)} = 0.3927$, and the bracket is 0.7263, so

$$\gamma = \frac{1.0318 \times 10^{-19}\,\mathrm{kg\,m/s} \times 4.7231 \times 10^{-14}\,\mathrm{m} \times 0.7263}{1.054\,571\,817 \times 10^{-34}\,\mathrm{J\,s}} = 33.56\,,$$

dimensionless as required, since kg m/s × m divided by J s is a pure number. The Gamow factor is $2\gamma = 67.13$, and the exponential of Equation (9.87) gives

$$T \approx \mathrm{e}^{-67.13} = 7.0 \times 10^{-30}\,, \tag{9.94}$$

about one crossing in 1.4×10^{29} attempts — an order-of-magnitude statement, the prefactor being undetermined here for the reason given above. Two checks are worth recording. First, letting $R \to 0$ replaces the bracket of Equation (9.93) by $\pi/2$ and gives $2\gamma \to 2\pi\eta$ with $\eta = zZe^2/(4\pi\varepsilon_0\hbar v)$ and $v = \sqrt{2E/m} = 1.553 \times 10^7\,\mathrm{m/s}$; here $\eta = 23.11$ and $2\pi\eta = 145.2$, more than twice the actual exponent, so the finite size of the source region is not a detail. Second, with $c = 299\,792\,458\,\mathrm{m/s}$ exactly, $v/c = 0.052$, which justifies using the non-relativistic Equation (9.59) at all.

Rests on Theorem 9.76.

Example 9.82 (Checking the validity criterion for the barrier above)**.** Return to the data of Example 9.81. At the inner edge $r = R$ the potential energy is $V(R) = 26.24\,\mathrm{MeV}$, so $V(R) - E = 21.24\,\mathrm{MeV}$, $\kappa(R) = 2.1266 \times 10^{-19}\,\mathrm{kg\,m/s}$ and $\bar{\lambda}(R) = \hbar/\kappa(R) = 0.496\,\mathrm{fm}$. With $|V'(R)| = zZe^2/(4\pi\varepsilon_0 R^2) = 467\,\mathrm{J/m}$, the criterion Equation (9.73) evaluates to

$$\frac{\hbar\, m_\alpha\, |V'(R)|}{\kappa(R)^3} = 0.034\,,$$

comfortably small. It grows without bound on approach to b, where $|V'(b)| = 16.96\,\mathrm{J/m}$ gives an Airy scale Equation (9.74) of $\ell = 3.67\,\mathrm{fm}$: the semiclassical forms are worthless over a few femtometres at the outer turning point, of order a tenth of the barrier width $b - R = 38\,\mathrm{fm}$. That region is precisely the one handled by Theorem 9.72 and not by Equation (9.70). Finally, the free momentum $\sqrt{2mE}$ corresponds to $\bar{\lambda} = 1.02\,\mathrm{fm}$ outside the barrier, small against that same width, which is the qualitative statement the criterion makes quantitative.

Rests on Proposition 9.65 and Example 9.81.

9.6.9 The radial equation

In three-dimensional space with a potential depending only on the radius, separation of variables (Chapter 10) writes a stationary state as $\psi = u(r)\,Y(\theta,\varphi)/r$, with Y a spherical harmonic of degree n (Equation (9.163)). The azimuthal index of Y does not enter the radial equation, which is why it is not displayed here beside the mass m; the radial factor u obeys

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2u}{\mathrm{d}r^2}(r) + \left[V(r) + \frac{\hbar^2\, n(n+1)}{2m\, r^2}\right] u(r) = E\,u(r)\,, \qquad r > 0\,, \tag{9.95}$$

with $u(0) = 0$. This is Equation (9.59) on the half-line with the centrifugal term added to the potential, so everything derived above applies verbatim once V is replaced by the bracket. For $n = 0$ the bracket is V itself, which is the case used in Example 9.81. For $n \geq 1$ the replacement is legitimate away from the origin but not at it: the centrifugal term diverges as r^{-2}, the criterion Equation (9.73) fails there however smooth V is, and the naive substitution gives the wrong energies. The repair, due to Langer, is to replace $n(n+1)$ by $\left(n + \tfrac{1}{2}\right)^2$ before applying the approximation, and it is derived next.

9.6.10 The Langer substitution

Theorem 9.83 (Langer transformation of the radial equation)**.** *Fix a reference length r_0 in* m *and let u solve the radial equation Equation (9.95) on $r > 0$. Under the change of independent and dependent variable*

$$r = r_0\,\mathrm{e}^{\varrho}\,, \qquad u(r) = \mathrm{e}^{\varrho/2}\,w(\varrho)\,, \qquad \varrho \in \mathbb{R}\,, \tag{9.96}$$

the radial equation becomes, exactly and with no approximation,

$$\frac{\mathrm{d}^2 w}{\mathrm{d}\varrho^2}(\varrho) + \frac{Q(\varrho)^2}{\hbar^2}\,w(\varrho) = 0\,, \qquad Q^2 = 2m\,[E - V(r)]\,r^2 - \hbar^2\left(n+\tfrac{1}{2}\right)^2\,, \tag{9.97}$$

a problem of the form Equation (9.60) on the whole line, with Q in J s *and ϱ dimensionless. Its semiclassical phase integral, transported back to the radial variable, is*

$$\frac{1}{\hbar}\int Q\,\mathrm{d}\varrho = \frac{1}{\hbar}\int \sqrt{2m\,[E - V(r)] - \frac{\hbar^2\left(n+\tfrac{1}{2}\right)^2}{r^2}}\;\mathrm{d}r\,, \tag{9.98}$$

which is the radial phase integral with $n(n+1)$ replaced by $\left(n+\tfrac{1}{2}\right)^2$.

Rests on Equations (9.60) and (9.95).

Derivation. *Derives Theorem 9.83.*

From Equation (9.96), $\mathrm{d}/\mathrm{d}r = r^{-1}\mathrm{d}/\mathrm{d}\varrho$ with $r = r_0\mathrm{e}^{\varrho}$, so

$$\frac{\mathrm{d}u}{\mathrm{d}r} = \frac{1}{r_0}\,\mathrm{e}^{-\varrho/2}\left(\frac{\mathrm{d}w}{\mathrm{d}\varrho} + \frac{w}{2}\right)\,, \qquad \frac{\mathrm{d}^2 u}{\mathrm{d}r^2} = \frac{1}{r_0^2}\,\mathrm{e}^{-3\varrho/2}\left(\frac{\mathrm{d}^2 w}{\mathrm{d}\varrho^2} - \frac{w}{4}\right)\,,$$

the cross terms in $\mathrm{d}w/\mathrm{d}\varrho$ cancelling in the second derivative, which is what the factor $\mathrm{e}^{\varrho/2}$ in Equation (9.96) was chosen to arrange. Substituting into Equation (9.95) written as

$$\frac{\mathrm{d}^2 u}{\mathrm{d}r^2} + \left[\frac{2m\,(E-V)}{\hbar^2} - \frac{n(n+1)}{r^2}\right]u = 0$$

and multiplying through by $r_0^2\,\mathrm{e}^{3\varrho/2}$ gives

$$\frac{\mathrm{d}^2 w}{\mathrm{d}\varrho^2} - \frac{w}{4} + \frac{2m\,(E-V)}{\hbar^2}\,r^2\,w - n(n+1)\,w = 0\,,$$

because $r_0^2\mathrm{e}^{2\varrho} = r^2$. The two constant terms combine as $n(n+1) + \tfrac{1}{4} = \left(n+\tfrac{1}{2}\right)^2$, which is Equation (9.97). For Equation (9.98), $\mathrm{d}\varrho = \mathrm{d}r/r$, so

$$Q\,\mathrm{d}\varrho = \sqrt{2m\,(E-V)\,r^2 - \hbar^2\left(n+\tfrac{1}{2}\right)^2}\;\frac{\mathrm{d}r}{r} = \sqrt{2m\,(E-V) - \frac{\hbar^2\left(n+\tfrac{1}{2}\right)^2}{r^2}}\;\mathrm{d}r\,.$$

Both terms of Q^2 carry the square of J s — the first because $2m\,(E-V)\,r^2$ is a mass times an energy times an area, the second manifestly — so $Q/\hbar$ is dimensionless, as an exponent per unit of ϱ must be; the arbitrary r_0 has cancelled out of both Equation (9.97) and Equation (9.98), as it must, since shifting it only translates ϱ. □

Remark 9.84 (Why the replacement is not optional). Three things are settled at once by Theorem 9.83. First, the half-line $r > 0$, whose endpoint at the origin is singular, becomes the whole line, on which Q^2 tends to the finite constant $-\hbar^2\left(n+\tfrac{1}{2}\right)^2$ as $\varrho \to -\infty$: the coefficient no longer diverges, the criterion Equation (9.73) is satisfied there, and

the semiclassical solution $w \sim \exp\left[\pm\left(n+\tfrac{1}{2}\right)\varrho\right]$ reproduces through Equation (9.96) the exact behaviours $u \sim r^{n+1}$ and $u \sim r^{-n}$ of Equation (9.95) at the origin — which the untransformed approximation does not, since with $n(n+1)$ in place of $\left(n+\tfrac{1}{2}\right)^2$ the same computation returns $u \sim r^{1/2\pm\sqrt{n(n+1)}}$, a pair of irrational exponents that agree with the exact ones only in the limit of large n. Second, the transformation is exact, so nothing has been assumed: the substitution $n(n+1) \to \left(n+\tfrac{1}{2}\right)^2$ is not a patch applied to the answer but the statement that the approximation had been made in the wrong variable. Third, the turning points of the transformed problem are simple for a potential of the usual kind, so Theorem 9.72 applies to it unchanged and Equation (9.88) may be used with the phase integral Equation (9.98). The Coulomb potential is the standard check [1023]: with the corrected coefficient the quantisation condition returns the exact hydrogen levels, and with $n(n+1)$ it does not.

Rests on Theorem 9.83.

Remark 9.85 (The $\hbar = 1$ convention of the literature). Most treatments of this material set $\hbar = 1$ and write the semiclassical exponent as $\int P\,\mathrm{d}x$ and the Gamow factor as $2\int \kappa\,\mathrm{d}x$. This treatise does not, for the reason given after Equation (9.64): $\hbar$ is the expansion parameter, and a convention that hides it hides the structure of the method. A reader coming from such a text restores the SI form of every expression above by the single rule that each exponent must be dimensionless: divide any integral of a momentum over a length by $\hbar$, in J s, and any energy by $\hbar$ before comparing it with an angular frequency. Momenta are then in kg m/s, lengths in m and energies in J, and no formula in this section needs any other change.

9.6.11 Boundary layers and matched asymptotic expansions

The semiclassical expansion is one half of the theory of a small parameter multiplying the highest derivative. It is the half in which the reduced problem is oscillatory and the small parameter sets a wavelength. The other half is the one in which the reduced problem is of lower order and *cannot meet all the boundary conditions*: the neglected derivative then reasserts itself in a thin region, a *boundary layer*, whose thickness is set by the small parameter and inside which the solution turns rapidly to meet the condition the outer solution had to abandon. The technique for handling that case is the same in outline — expand, and match — and is set out here because the physical parts need it and because a reader who has followed Section 9.6 has already done the harder half.

Definition 9.86 (Outer and inner expansions; matching)**.** Consider a problem containing a small parameter $\varepsilon > 0$ on a fixed interval. An *outer expansion* is a series in ε solving the problem with ε set to zero at leading order, valid where the solution varies on the scale of the interval. An *inner expansion* is a series solving the problem rewritten in a stretched variable $\xi = (x - x_c)/\delta(\varepsilon)$ about a point x_c, with $\delta \to 0$; the choice of δ for which the rewritten equation retains its highest derivative in a non-trivial balance is the *distinguished limit*. The two are joined by the *matching condition*

$$\lim_{\xi\to\infty} y_{\mathrm{in}}(\xi) = \lim_{x\to x_c} y_{\mathrm{out}}(x)\,, \tag{9.99}$$

which says that the inner solution far out and the outer solution close in describe the same thing, and the *composite* approximation is

$$y_{\mathrm{c}}(x) = y_{\mathrm{out}}(x) + y_{\mathrm{in}}\!\left(\frac{x-x_c}{\delta}\right) - (\text{their common limit})\,, \tag{9.100}$$

the subtraction removing the part they both already carry.

Rests on Definitions 7.16 and 9.1.

Proposition 9.87 (The distinguished limit, and which end carries the layer)**.** *Let a and b be continuous on $[0,1]$ with $a(x) \geq a_0 > 0$, and consider*

$$\varepsilon \frac{\mathrm{d}^2 y}{\mathrm{d}x^2}(x) + a(x) \frac{\mathrm{d}y}{\mathrm{d}x}(x) + b(x)\, y(x) = 0\,, \qquad y(0) = A\,, \quad y(1) = B\,. \tag{9.101}$$

Then (i) *the only stretching $x = x_c + \delta\xi$ under which the second-derivative term is retained in a balance with a term at least as large as every neglected one is $\delta = \varepsilon$;* (ii) *at that scaling the leading inner equation is $\mathrm{d}^2Y/\mathrm{d}\xi^2 + a(x_c)\,\mathrm{d}Y/\mathrm{d}\xi = 0$, whose solutions are constants and $\mathrm{e}^{-a(x_c)\xi}$; and* (iii) *that exponential decays away from x_c into the interval only at $x_c = 0$. The layer therefore sits at the left endpoint, and the outer solution — the solution of the first-order reduced equation $a\,y' + b\,y = 0$ — must carry the condition at $x = 1$.*

Rests on Equation (9.101), Theorem 9.23, and Proposition 9.18.

Derivation. Derives Proposition 9.87.

Under $x = x_c + \delta\xi$ the three terms of Equation (9.101) carry the factors ε/δ^2, $1/\delta$ and 1. Suppose the first is retained. If it balances the third, $\varepsilon/\delta^2 \sim 1$ and $\delta \sim \sqrt{\varepsilon}$, whereupon the second term is of order $\varepsilon^{-1/2}$ and dominates both retained terms — so the balance is inconsistent. If it balances the second, $\varepsilon/\delta^2 \sim 1/\delta$, that is $\delta \sim \varepsilon$, and the third term, of order 1, is smaller by a factor ε than the two retained ones: consistent. That is (i), and substituting $\delta = \varepsilon$ and $a(x_c + \varepsilon\xi) \to a(x_c)$ gives (ii), whose solutions are read off Theorem 9.23 for the constant-coefficient equation — or directly from Proposition 9.18 applied to $\mathrm{d}Y/\mathrm{d}\xi$.

For (iii): at $x_c = 0$ the interval is $\xi \geq 0$ and $\mathrm{e}^{-a(0)\xi} \to 0$ as $\xi \to \infty$, so the inner solution settles to a constant and Equation (9.99) can be imposed. At $x_c = 1$ the interval is $\xi \leq 0$ and $\mathrm{e}^{-a(1)\xi} \to \infty$ as $\xi \to -\infty$: the inner solution grows without bound where it should match, so no matching is possible and there is no layer there. The reduced equation is of first order and therefore accepts exactly one boundary condition, which must then be the one at $x = 1$. □

Proposition 9.88 (The model problem, and the error of the composite)**.** *Let $A, B \in \mathbb{R}$ and consider Equation (9.101) with $a \equiv 1$, $b \equiv 1$: $\varepsilon y'' + y' + y = 0$, $y(0) = A$, $y(1) = B$. Its outer solution is $y_{\mathrm{out}}(x) = B\,\mathrm{e}^{1-x}$, its inner solution is $y_{\mathrm{in}}(\xi) = (A - B\mathrm{e})\mathrm{e}^{-\xi} + B\mathrm{e}$, and the composite Equation (9.100) is*

$$y_{\mathrm{c}}(x) = B\,\mathrm{e}^{1-x} + (A - B\mathrm{e})\,\mathrm{e}^{-x/\varepsilon}\,. \tag{9.102}$$

There is a constant C, depending only on $|A|$ and $|B|$, with

$$\sup_{0 \leq x \leq 1} |y(x) - y_{\mathrm{c}}(x)| \leq C\,\varepsilon \tag{9.103}$$

for every $\varepsilon \in (0, 1/8]$. The approximation is therefore uniform: it is correct both inside the layer and outside it.

Rests on Propositions 9.28 and 9.87 and Theorem 7.38.

Derivation. Derives Proposition 9.88.

The equation has constant coefficients, so by Proposition 9.28 its solutions are combinations of $\mathrm{e}^{\lambda_\pm x}$ with $\lambda_\pm = \left[-1 \pm \sqrt{1 - 4\varepsilon}\right]/(2\varepsilon)$, real and distinct for $\varepsilon < 1/4$. Taylor's theorem (Theorem 7.38) applied to $\sqrt{1-u}$ at $u = 4\varepsilon$ gives $\left|\sqrt{1-4\varepsilon} - (1 - 2\varepsilon)\right| \leq c_0\varepsilon^2$ on $(0, 1/8]$, whence

$$\lambda_+ = -1 + r\,, \quad |r| \leq c_0\varepsilon/2\,, \qquad \lambda_- = -\frac{1}{\varepsilon} + s\,, \quad |s| \leq 1 + c_0\varepsilon/2 \leq 2\,. \tag{9.104}$$

In particular $\lambda_- \leq -1/(2\varepsilon)$, so $\eta = \mathrm{e}^{\lambda_-} \leq \mathrm{e}^{-1/(2\varepsilon)}$ is smaller than any power of ε.

Write $y = C_+\mathrm{e}^{\lambda_+ x} + C_-\mathrm{e}^{\lambda_- x}$. The boundary conditions give $C_+ + C_- = A$ and $C_+\mathrm{e}^{\lambda_+} + C_-\eta = B$, so $C_+ = (B - A\eta)/(\mathrm{e}^{\lambda_+} - \eta)$. Since $\mathrm{e}^{\lambda_+} = \mathrm{e}^{-1}\mathrm{e}^{r} = \mathrm{e}^{-1}(1 + O(\varepsilon))$ by

Equation (9.104) and η is negligible beside it, $C_+ = B\mathrm{e} + O(\varepsilon)$ and hence $C_- = A - B\mathrm{e} + O(\varepsilon)$, every implied constant depending only on $|A|$, $|B|$ and c_0.

Now compare term by term. For the outer term, on $0 \le x \le 1$,

$$\left| C_+ \mathrm{e}^{\lambda_+ x} - B\mathrm{e}^{1-x} \right| \le |C_+ - B\mathrm{e}|\ \mathrm{e}^{|\lambda_+|} + B\mathrm{e}\ \mathrm{e}^{-x}\,|\mathrm{e}^{rx} - 1| \le C_1 \varepsilon\,,$$

using $|\mathrm{e}^{rx} - 1| \le |r|\, x\mathrm{e}^{|r|}$ and Equation (9.104). For the layer term, writing $\mathrm{e}^{\lambda_- x} = \mathrm{e}^{-x/\varepsilon}\mathrm{e}^{sx}$ and using $\mathrm{e}^{\lambda_- x} \le 1$ for $x \ge 0$,

$$\left| C_- \mathrm{e}^{\lambda_- x} - (A - B\mathrm{e})\mathrm{e}^{-x/\varepsilon} \right| \le |C_- - (A - B\mathrm{e})| + |A - B\mathrm{e}|\ \mathrm{e}^{-x/\varepsilon}\,|\mathrm{e}^{sx} - 1|\,.$$

The first term is $O(\varepsilon)$. In the second, $|\mathrm{e}^{sx} - 1| \le |s|\, x\mathrm{e}^{|s|} \le 2\mathrm{e}^2 x$ on $[0,1]$, and $x\,\mathrm{e}^{-x/\varepsilon}$ attains its maximum ε/e at $x = \varepsilon$; so the second term is $O(\varepsilon)$ too. Adding the two comparisons gives Equation (9.103). □

Remark 9.89 (Two singular perturbations, and where each belongs). Section 9.6 and this subsection treat the same syntactic situation — a small parameter on the highest derivative — and produce different objects, because the sign of the coefficient of the undifferentiated term decides whether the neglected derivative buys oscillation or a layer. Where the reduced problem is oscillatory the rapid variation is spread over the whole interval and the expansion is in the phase, which is Corollary 9.62; where the reduced problem is of lower order the rapid variation is confined to a region of thickness ε and the expansion is the two-region one of Definition 9.86. The turning point of Theorem 9.72 is the same matching problem in the first setting: an inner solution — there the Airy function of Definition 9.68 — joined to two outer ones by their common asymptotics.

Nothing above is specific to one dimension or to linear equations. The reduction of the equations of a viscous fluid to Prandtl's boundary-layer equations [315] is Proposition 9.87 carried out on a partial differential equation, with the Reynolds number in place of $1/\varepsilon$: the inviscid outer flow cannot satisfy the no-slip condition at a wall, the viscous term is restored in a layer whose thickness the same balance fixes, and Equation (9.99) is what hands the outer pressure gradient to the inner problem. Chapter 36 carries that computation and its measurement.

9.7 The one-dimensional Helmholtz equation

9.7.1 Sturm–Liouville data

The functions p, q and r of the operator are

$$p(x) = 1\,, \qquad q(x) = 0\,, \qquad r(x) = 1\,, \tag{9.105}$$

with λ^2 the free eigenvalue parameter.

9.7.2 Homogeneous solution

The homogeneous equation becomes

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2}(x) + \lambda^2\, y(x) = 0\,, \tag{9.106}$$

whose solution is

$$y(x) = A\,\mathrm{e}^{\mathrm{i}\lambda x} + B\,\mathrm{e}^{-\mathrm{i}\lambda x}\,. \tag{9.107}$$

Remark 9.90. Using Euler's formula, $\mathrm{e}^{\pm \mathrm{i}t} = \cos t \pm \mathrm{i}\sin t$, the solution Equation (9.107) may equivalently be written in the real form

$$y(x) = A'\cos(\lambda x) + B'\sin(\lambda x)\,. \tag{9.108}$$

9.7.3 Orthonormality

On the interval $[-\pi, \pi]$, and for integer values $\lambda = n \in \mathbb{Z}$, the solution functions satisfy

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}(m-n)x} \, \mathrm{d}x = \delta_{mn} \, . \tag{9.109}$$

Equation (9.109) is Lemma 17.2, proved there by direct integration: the integrand is $\mathrm{e}^{\mathrm{i}(m-n)x}$, which equals 1 when $m = n$ and, when $m \neq n$, has an antiderivative that returns to its starting value over a full period, so the integral vanishes.

Remark 9.91 (Which measure makes the exponentials orthonormal). The factor $1/(2\pi)$ in Equation (9.109) is not cosmetic, and it is worth placing once. Against the plain $\mathrm{d}x$ of Equation (9.46) the bare exponentials are merely *orthogonal*: $\int_{-\pi}^{\pi} \overline{\mathrm{e}^{\mathrm{i}nx}} \, \mathrm{e}^{\mathrm{i}nx} \, \mathrm{d}x = 2\pi$, not 1, so $\{\mathrm{e}^{\mathrm{i}nx}\}$ is not an orthonormal set in the sense of Definition 9.45. Two readings repair this, and they are one statement written twice: either keep $\mathrm{d}x$ and take the normalised functions $\phi_n(x) = \mathrm{e}^{\mathrm{i}nx}/\sqrt{2\pi}$, or keep the bare exponentials $\phi_n(x) = \mathrm{e}^{\mathrm{i}nx}$ and read every integral of Section 9.4 — Equation (9.46), the coefficient formula Equation (9.48) and the triple overlap Equation (9.49) alike — against the normalised measure $\mathrm{d}x/(2\pi)$. The Fourier sections of this chapter take the second reading throughout, which is exactly what puts the $1/(2\pi)$ into Equation (9.111) and the $1/L$ into Equation (9.113). The choice cancels out of every expansion, since the factor a coefficient gains is the one its basis function loses; what must not happen is the two readings being mixed inside one calculation.

9.7.4 Fourier series

Expanding a function on $[-\pi, \pi]$ in the complex exponentials — orthonormal in the second sense of Remark 9.91 — gives the *Fourier series*

$$f(x) = \sum_{n=-\infty}^{\infty} c_n \, \mathrm{e}^{\mathrm{i}nx} \, , \tag{9.110}$$

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \, \mathrm{e}^{-\mathrm{i}nx} \, \mathrm{d}x \, , \tag{9.111}$$

an instance of the general coefficient formula Equation (9.48).

Under a *change of interval* to an arbitrary interval $[a, b]$ of length $L = b - a$, the series and its coefficients become

$$f(x) = \sum_{n=-\infty}^{\infty} f_n \, \mathrm{e}^{\mathrm{i}k_n x} \, , \qquad k_n = \frac{2n\pi}{L} \, , \tag{9.112}$$

$$f_n = \frac{1}{L} \int_{a}^{b} f(x) \, \mathrm{e}^{-\mathrm{i}k_n x} \, \mathrm{d}x \, . \tag{9.113}$$

Trigonometric form

On the symmetric interval $[-L, L]$ the cosine and sine functions satisfy the orthonormality relations

$$\frac{1}{L} \int_{-L}^{L} \cos\left(\frac{m\pi x}{L}\right) \cos\left(\frac{n\pi x}{L}\right) \mathrm{d}x = \delta_{mn} \, , \tag{9.114}$$

$$\frac{1}{L} \int_{-L}^{L} \sin\left(\frac{m\pi x}{L}\right) \sin\left(\frac{n\pi x}{L}\right) \mathrm{d}x = \delta_{mn} \, , \tag{9.115}$$

$$\int_{-L}^{L} \cos\left(\frac{m\pi x}{L}\right) \sin\left(\frac{n\pi x}{L}\right) \mathrm{d}x = 0 \, , \tag{9.116}$$

with $m, n \geq 1$, and the Fourier series takes the real form

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right] , \tag{9.117}$$

$$a_n = \frac{1}{L} \int_{-L}^{L} f(x) \cos \frac{n\pi x}{L} \, \mathrm{d}x \, , \tag{9.118}$$

$$b_n = \frac{1}{L} \int_{-L}^{L} f(x) \sin \frac{n\pi x}{L} \, \mathrm{d}x \, . \tag{9.119}$$

The relations Equations (9.114) to (9.116) are proved at Equation (17.9), where the product formulas of trigonometry reduce each of them to Lemma 17.2: the identity $\cos A \cos B = \frac{1}{2}[\cos(A-B) + \cos(A+B)]$ turns the first into δ_{mn}, since the second index coincidence $m = -n$ is impossible for positive m and n; the sine relation is the same identity with the opposite sign; and the mixed integral vanishes as that of an odd function over a symmetric interval. The coefficient formulas Equations (9.118) and (9.119) are Equation (17.7) written with $T = 2L$.

9.7.5 The Fourier transform

Consider a function that can be expanded in the Fourier series Equation (9.112). If $L \to +\infty$, the sum over the discrete wavenumbers k_n passes into a Riemann integral, and one obtains

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(k) \, \mathrm{e}^{\mathrm{i}kx} \, \mathrm{d}k \, , \tag{9.120}$$

where $g(k)$, called the *Fourier transform* of $f(x)$, is defined by

$$g(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \, \mathrm{e}^{-\mathrm{i}kx} \, \mathrm{d}x \, . \tag{9.121}$$

The Riemann-sum argument sketched above is carried out at Equation (17.33), in the same normalisation: the coefficients of the series on $[-L, L]$ are $\sqrt{2\pi}$ times the transform sampled at the harmonics k_n, divided by the interval length, and the mesh of the sum is $\Delta k = 2\pi/L \to 0$. That argument is heuristic, because the summand depends on L as well as on the mesh; the inversion formula Equation (9.120) is established properly, by Gaussian regularisation, in Theorem 17.32, and the injectivity it implies is Corollary 17.33.

Properties

Transform of derivatives. If f is n times differentiable, $f^{(n)}$ is absolutely integrable, and every derivative below the top one vanishes at infinity,

$$\lim_{x \to \pm\infty} f^{(j)}(x) = 0 \qquad \text{for } j = 0, 1, \ldots, n-1 \, ,$$

then the transform g_n of the n-th derivative $f^{(n)}$ satisfies

$$g_n(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f^{(n)}(x) \, \mathrm{e}^{-\mathrm{i}kx} \, \mathrm{d}x = (\mathrm{i}k)^n \, g(k) \, . \tag{9.122}$$

The case $n = 1$ is Equation (17.38) of Proposition 17.30, proved there by a single integration by parts: the boundary term $[f \, \mathrm{e}^{-\mathrm{i}kx}]_{-\infty}^{\infty}$ vanishes because f tends to zero, and what is left is $\mathrm{i}k$ times the transform of f. The general n follows by induction, applying that result to $f^{(j)}$ in turn — which is exactly why the hypothesis above is imposed on *every* derivative below the top one and not on f alone: each step of the induction consumes one boundary term. This is the property that turns a constant-coefficient differential operator into a polynomial and is the reason the transform is a method at all.

Convolution. The convolution of two functions and the theorem that its transform is the product of theirs are Definition 17.50 and Theorem 17.53: in the normalisation of Equation (9.121) the transform of $f * g$ is $\sqrt{2\pi}$ times the product of the transforms, and dually the transform of a product is $(2\pi)^{-1/2}$ times the convolution of the transforms. The derivation is a single application of Fubini's theorem to the double integral, justified by the bound Equation (17.64).

Parseval relation. The transform preserves the norm,

$$\int_{-\infty}^{\infty} |f(x)|^2 \, \mathrm{d}x = \int_{-\infty}^{\infty} |g(k)|^2 \, \mathrm{d}k \,. \tag{9.123}$$

Equation (9.123) is Plancherel's theorem, Theorem 17.36: the transform of Equation (9.121) is an isometry on the smooth rapidly decreasing functions, by Theorem 17.32 and an exchange of integrals, and extends uniquely to a unitary map of the square-integrable functions onto themselves. The statement for the Fourier *series* — the completeness of the harmonics, and hence Equation (9.52) for that system — is Theorem 17.34.

9.8 The Bessel equation

9.8.1 Sturm–Liouville data

The identification of the operator data is

$$p(x) = x \,, \qquad q(x) = x \,, \qquad \lambda^2 = -\nu^2 \,, \qquad r(x) = \frac{1}{x} \,. \tag{9.124}$$

9.8.2 General solution

The homogeneous equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[x \, \frac{\mathrm{d}y}{\mathrm{d}x}(x)\right] + \left[x - \frac{\nu^2}{x}\right] y(x) = 0 \,, \tag{9.125}$$

whose solution is

$$y(x) = A \, J_\nu(x) + B \, N_\nu(x) \,, \tag{9.126}$$

with J_ν the Bessel functions of the first kind and N_ν (also written Y_ν) the Neumann functions, the Bessel functions of the second kind [1976].

Remark 9.92. Expanding the derivative in Equation (9.125) and multiplying by x puts the equation into its familiar form $x^2 y'' + x y' + (x^2 - \nu^2)\, y = 0$.

Definition 9.93 (Bessel functions of the first kind)**.** For $\nu \geq 0$ and $x > 0$,

$$J_\nu(x) = \sum_{j=0}^{\infty} \frac{(-1)^j}{j! \, \Gamma(j + \nu + 1)} \left(\frac{x}{2}\right)^{2j+\nu} \,, \tag{9.127}$$

a series convergent for every x by the ratio test (Proposition 7.48). For non-integer ν the same formula with ν replaced by $-\nu$ defines a second, independent solution $J_{-\nu}$. At integer $\nu = n$ the reciprocal gamma function vanishes at the non-positive integers, so the first n terms of that series drop out and $J_{-n} = (-1)^n J_n$ is not independent; the second solution is then the Neumann function $N_\nu = \big[J_\nu \cos \nu\pi - J_{-\nu}\big] / \sin \nu\pi$, read at integer order as the limit of that expression.

Proposition 9.94 (The series solves the Bessel equation)**.** *J_ν of Equation (9.127) satisfies Equation (9.125).*

Rests on Definition 9.93 and Equation (9.125).

Derivation. *Derives Proposition 9.94.*

Use the expanded form of Remark 9.92 and substitute Equation (9.127) term by term, which is legitimate inside the disc of convergence (Theorem 7.51). Write $J_\nu = \sum_j c_j x^{2j+\nu}$ with $c_j = (-1)^j \big[j!\,\Gamma(j+\nu+1)\,2^{2j+\nu}\big]^{-1}$. The operator $x^2\mathrm{d}^2/\mathrm{d}x^2 + x\,\mathrm{d}/\mathrm{d}x - \nu^2$ sends $x^{2j+\nu}$ to $\big[(2j+\nu)^2 - \nu^2\big]x^{2j+\nu} = 4j(j+\nu)\,x^{2j+\nu}$, while the remaining term $x^2 J_\nu$ shifts the index by one. The coefficient of $x^{2j+\nu}$ in the whole expression is therefore $4j(j+\nu)\,c_j + c_{j-1}$, and this vanishes: by $j!\,\Gamma(j+\nu+1) = j\,(j-1)!\,(j+\nu)\,\Gamma(j+\nu)$ and $2^{2j+\nu} = 4\cdot 2^{2(j-1)+\nu}$,

$$4j(j+\nu)\,c_j = \frac{4\,(-1)^j}{(j-1)!\,\Gamma(j+\nu)\,2^{2j+\nu}} = -c_{j-1}\,.$$

The term $j=0$ has no partner and vanishes on its own, because $4j(j+\nu)=0$ there. □

9.8.3 Recurrence relations

Proposition 9.95 (The two ladder identities)**.** *For every order ν and every $x>0$,*

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[x^{-\nu}\,J_\nu(x)\right] = -x^{-\nu}\,J_{\nu+1}(x)\,, \tag{9.128}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[x^{\nu}\,J_\nu(x)\right] = x^{\nu}\,J_{\nu-1}(x)\,. \tag{9.129}$$

Rests on Definition 9.93.

Derivation. *Derives Proposition 9.95.*

Both are term-by-term computations on Equation (9.127). For Equation (9.128),

$$x^{-\nu}J_\nu(x) = \sum_{j\ge 0} \frac{(-1)^j\,x^{2j}}{j!\,\Gamma(j+\nu+1)\,2^{2j+\nu}}\,,$$

whose derivative is $\sum_{j\ge1}(-1)^j\,2j\,x^{2j-1}\big[j!\,\Gamma(j+\nu+1)2^{2j+\nu}\big]^{-1}$; putting $j=i+1$ and cancelling the factor $2(i+1)$ against $(i+1)!$ turns this into $-\sum_{i\ge0}(-1)^i x^{2i+1}\big[i!\,\Gamma(i+\nu+2)\,2^{2i+\nu+1}\big]^{-1}$, which is $-x^{-\nu}J_{\nu+1}(x)$ by Equation (9.127) at order $\nu+1$. For Equation (9.129), the derivative of $x^\nu J_\nu = \sum_j (-1)^j x^{2j+2\nu}\big[j!\,\Gamma(j+\nu+1)2^{2j+\nu}\big]^{-1}$ carries the factor $2j+2\nu = 2(j+\nu)$, which cancels against $\Gamma(j+\nu+1) = (j+\nu)\Gamma(j+\nu)$ and lowers the power of two by one, giving $\sum_j (-1)^j x^{2j+2\nu-1}\big[j!\,\Gamma(j+\nu)2^{2j+\nu-1}\big]^{-1} = x^\nu J_{\nu-1}(x)$. □

Corollary 9.96 (Recurrence relations)**.** *For every order ν,*

$$J_{\nu-1}(x) + J_{\nu+1}(x) = \frac{2\nu}{x}\,J_\nu(x)\,, \tag{9.130}$$

$$J_{\nu-1}(x) - J_{\nu+1}(x) = 2\,\frac{\mathrm{d}J_\nu}{\mathrm{d}x}(x)\,. \tag{9.131}$$

In particular $\frac{\mathrm{d}J_\nu}{\mathrm{d}x} = -J_{\nu+1}$ at every zero of J_ν.

Rests on Proposition 9.95.

Derivation. *Derives Corollary 9.96.*

Carrying out the differentiations on the left of Equations (9.128) and (9.129) and dividing by $x^{-\nu}$ and x^ν respectively gives

$$\frac{\mathrm{d}J_\nu}{\mathrm{d}x} - \frac{\nu}{x}J_\nu = -J_{\nu+1}\,, \qquad \frac{\mathrm{d}J_\nu}{\mathrm{d}x} + \frac{\nu}{x}J_\nu = J_{\nu-1}\,.$$

Subtracting gives Equation (9.130) and adding gives Equation (9.131). At a zero of J_ν the first of the two displayed identities reduces to $\mathrm{d}J_\nu/\mathrm{d}x = -J_{\nu+1}$. □

9.8.4 Orthogonality

Remark 9.97 (Which eigenvalue problem the orthogonality belongs to). Equation (9.124) reads Equation (9.125) as a Sturm–Liouville equation in which the order ν carries the spectral parameter. The orthogonality relation below belongs to the other reading of the same equation, the one the applications use: fix ν, and let $y(x) = J_\nu(\alpha x/a)$ on $[0, a]$. Rescaling Equation (9.125) gives

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[x\,\frac{\mathrm{d}y}{\mathrm{d}x}(x)\right] + \left[-\frac{\nu^2}{x} + \frac{\alpha^2}{a^2}\,x\right] y(x) = 0\,, \tag{9.132}$$

which is Equation (9.54) with $p(x) = x$, $q(x) = -\nu^2/x$, weight $r(x) = x$ and eigenvalue $\lambda^2 = \alpha^2/a^2$. It is that spectral parameter, and not ν, which distinguishes the functions appearing in Equation (9.133).

Theorem 9.98 (Orthogonality of the Bessel functions)**.** *Let $\nu \geq 0$ and let $\alpha_{\nu m}$ denote the m-th positive zero of J_ν. On the interval $[0, a]$ the rescaled Bessel functions are orthogonal with weight x:*

$$\int_0^a J_\nu\left(\frac{\alpha_{\nu m}x}{a}\right) J_\nu\left(\frac{\alpha_{\nu n}x}{a}\right) x\,\mathrm{d}x = \frac{a^2}{2}\left[J_{\nu+1}(\alpha_{\nu m})\right]^2 \delta_{mn}\,. \tag{9.133}$$

Rests on Theorem 9.57, Corollary 9.96, and Equation (9.132).

Derivation. Derives Theorem 9.98.

Take $u(x) = J_\nu(\lambda x)$ and $v(x) = J_\nu(\mu x)$ with $\lambda, \mu > 0$, both solutions of Equation (9.132) with $p(x) = x$, weight x and eigenvalues λ^2 and μ^2. Integrating the Lagrange identity Equation (9.56) over $[0, a]$ gives

$$\left(\lambda^2 - \mu^2\right) \int_0^a J_\nu(\lambda x)\, J_\nu(\mu x)\, x\,\mathrm{d}x = a\left[\mu\, J_\nu(\lambda a)\, J'_\nu(\mu a) - \lambda\, J_\nu(\mu a)\, J'_\nu(\lambda a)\right]\,, \tag{9.134}$$

the prime denoting the derivative of J_ν with respect to its own argument. The contribution at $x = 0$ vanishes because $p(x) = x$ does while J_ν behaves there as x^ν by Equation (9.127) — the second mechanism of Remark 9.58. Taking $\lambda = \alpha_{\nu m}/a$ and $\mu = \alpha_{\nu n}/a$ with $m \neq n$ makes both Bessel values on the right vanish while the factor on the left does not, which is the off-diagonal statement; it is Theorem 9.57 with the boundary condition $y(a) = 0$, the first mechanism of Remark 9.58.

For the diagonal, fix $\lambda = \alpha_{\nu m}/a$, keep μ free and use $J_\nu(\lambda a) = 0$ in Equation (9.134):

$$\int_0^a J_\nu(\lambda x)\, J_\nu(\mu x)\, x\,\mathrm{d}x = \frac{-a\,\lambda\, J_\nu(\mu a)\, J'_\nu(\lambda a)}{\lambda^2 - \mu^2}\,.$$

Let $\mu \to \lambda$. The numerator vanishes like $J_\nu(\mu a) = (\mu - \lambda)\, a\, J'_\nu(\lambda a) + \ldots$ and the denominator like $-(\mu - \lambda)\, 2\lambda$, so the quotient tends to $\frac{1}{2}a^2\left[J'_\nu(\lambda a)\right]^2$. Finally $J'_\nu(\alpha_{\nu m}) = -J_{\nu+1}(\alpha_{\nu m})$ by Corollary 9.96, and squaring removes the sign. □

9.8.5 Generating function

For integer order the Bessel functions of the first kind are generated by

$$g(x, t) = \mathrm{e}^{\frac{x}{2}\left(t - \frac{1}{t}\right)} = \sum_{n=-\infty}^{\infty} J_n(x)\, t^n\,. \tag{9.135}$$

Proposition 9.99 (Generating function)**.** *Equation (9.135) holds for every x and every complex $t \neq 0$.*

Rests on Definition 9.93.

Derivation. *Derives Proposition 9.99.*

Both factors of $g(x,t) = \mathrm{e}^{xt/2}\,\mathrm{e}^{-x/(2t)}$ are entire in their arguments, so the two exponential series may be multiplied out and the resulting double series rearranged absolutely (Proposition 7.49):

$$\begin{aligned}\mathrm{e}^{\frac{x}{2}t}\,\mathrm{e}^{-\frac{x}{2t}} &= \left[\sum_{j=0}^{\infty}\frac{1}{j!}\left(\frac{x}{2}\right)^{j}t^{j}\right]\left[\sum_{k=0}^{\infty}\frac{(-1)^{k}}{k!}\left(\frac{x}{2}\right)^{k}t^{-k}\right]\\ &= \sum_{n=-\infty}^{\infty}t^{n}\sum_{k}\frac{(-1)^{k}}{(k+n)!\,k!}\left(\frac{x}{2}\right)^{2k+n},\end{aligned}$$

the outer index being $n = j-k$ and the inner sum running over those $k \geq 0$ with $k+n \geq 0$. Comparing with Equation (9.127) at integer order, where $\Gamma(k+n+1) = (k+n)!$ and the reciprocal factorial is zero whenever $k+n<0$, the inner sum is exactly $J_n(x)$, which is Equation (9.135). □

9.8.6 The Fourier–Bessel series

Proposition 9.100 (Fourier–Bessel coefficients)**.** *Let $\nu \geq 0$, and let f on $[0,a]$ admit an expansion*

$$f(x) = \sum_{m=1}^{\infty} c_m\,J_\nu\left(\frac{\alpha_{\nu m}\,x}{a}\right), \tag{9.136}$$

convergent in the mean with respect to the weight x. Then

$$c_m = \frac{2}{a^2\left[J_{\nu+1}(\alpha_{\nu m})\right]^2}\int_0^a f(x)\,J_\nu\left(\frac{\alpha_{\nu m}\,x}{a}\right)x\,\mathrm{d}x\,. \tag{9.137}$$

Rests on Theorem 9.98 and Definition 9.47.

Derivation. *Derives Proposition 9.100.*

Multiply Equation (9.136) by $J_\nu(\alpha_{\nu n}x/a)\,x$ and integrate over $[0,a]$. Mean convergence permits the exchange of sum and integral, and Equation (9.133) leaves the single term $m=n$ carrying the constant $a^2\left[J_{\nu+1}(\alpha_{\nu n})\right]^2/2$. This is the argument of Proposition 9.48 run for a set that is orthogonal but not normalised, the normalising constant being divided out at the end instead of absorbed into the functions. □

Remark 9.101 (Completeness is a separate theorem). That an expansion Equation (9.136) exists at all — that the system $\{J_\nu(\alpha_{\nu m}x/a)\}_{m\geq 1}$ is complete in the square-integrable functions on $[0,a]$ with weight x — is a statement about the Sturm–Liouville problem of Remark 9.97, not about Bessel functions in particular, and it is not proved here; the general statement belongs to Chapter 12 [1222]. What Proposition 9.100 does establish is that if the expansion exists then it is this one — the same division of labour as in Remark 9.53.

9.8.7 The zeros

The orthogonality relation of Theorem 9.98 and the Fourier–Bessel series of Section 9.8.6 are indexed by the positive zeros of J_ν, and every eigenfrequency of a circular membrane or a cylindrical cavity is one of them. Their existence, their simplicity and their spacing are therefore not decoration.

Proposition 9.102 (Liouville normal form of the Bessel equation)**.** *On $x>0$ the substitution $y(x) = x^{-1/2}u(x)$ turns Equation (9.125) into*

$$\frac{\mathrm{d}^2u}{\mathrm{d}x^2}(x) + \left[1 - \frac{\nu^2-\frac{1}{4}}{x^2}\right]u(x) = 0\,. \tag{9.138}$$

The first-derivative term is gone, and for large x the equation is a vanishing perturbation of $u'' + u = 0$.

Rests on Equation (9.125) and Remark 9.92.

Derivation. *Derives Proposition 9.102.*

Use the expanded form $x^2y'' + xy' + (x^2 - \nu^2)y = 0$ of Remark 9.92. With $y = x^{-1/2}u$,

$$y' = -\tfrac{1}{2}x^{-3/2}u + x^{-1/2}u' , \qquad y'' = \tfrac{3}{4}x^{-5/2}u - x^{-3/2}u' + x^{-1/2}u'' ,$$

so $x^2y'' = \tfrac{3}{4}x^{-1/2}u - x^{1/2}u' + x^{3/2}u''$ and $xy' = -\tfrac{1}{2}x^{-1/2}u + x^{1/2}u'$. The terms in u' cancel, and adding $(x^2 - \nu^2)x^{-1/2}u$ leaves

$$x^{3/2}u'' + x^{3/2}u + \left(\tfrac{3}{4} - \tfrac{1}{2} - \nu^2\right) x^{-1/2}u = 0 ,$$

which is Equation (9.138) after division by $x^{3/2}$, since $\tfrac{3}{4} - \tfrac{1}{2} = \tfrac{1}{4}$. □

Proposition 9.103 (The positive zeros and their spacing)**.** *Let $\nu \geq 0$. Every zero of J_ν in $x > 0$ is simple, and there are infinitely many of them; writing $j_{\nu,n}$ for the n-th in increasing order,*

$$j_{\nu,n} = \left(n + \frac{\nu}{2} - \frac{1}{4}\right)\pi + O\left(\frac{1}{n}\right) , \qquad n \to \infty . \tag{9.139}$$

Successive zeros are therefore asymptotically π apart, whatever the order.

Rests on Proposition 9.102 and Corollaries 9.9 and 9.96.

Derivation. *Derives Proposition 9.103.*

Simplicity. On $x > 0$ the Bessel equation is regular — its leading coefficient x^2 does not vanish — so Corollary 9.9 applies: a solution with $y(x_*) = y'(x_*) = 0$ at some $x_* > 0$ is identically zero. Since J_ν is not, no positive zero of it can be a zero of its derivative. (Equivalently, by Corollary 9.96, $J_\nu' = -J_{\nu+1}$ at a zero of J_ν, and the two functions have no common positive zero.)

Location. Proposition 9.102 writes $J_\nu(x) = x^{-1/2}u(x)$ with u solving an equation that tends to $u'' + u = 0$: the solutions oscillate with asymptotic half-period π, and the classical large-argument form quoted in Section 9.6.5 from [1023],

$$J_\nu(x) = \sqrt{\frac{2}{\pi x}}\left[\cos\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) + O\left(\frac{1}{x}\right)\right] , \qquad x \to \infty , \tag{9.140}$$

makes the phase explicit. The cosine vanishes where $x - \nu\pi/2 - \pi/4 = \pi/2 + (n-1)\pi$, that is at $x = (n + \nu/2 - \tfrac{1}{4})\pi$, and it crosses each of those points with slope of modulus 1. A perturbation of size $O(1/x)$ of a function crossing zero with unit slope displaces the crossing by $O(1/x)$, and the amplitude $\sqrt{2/\pi x}$, being non-zero, cannot create or destroy a crossing; so for large n there is exactly one zero of J_ν within $O(1/n)$ of each of those points, which is Equation (9.139), and there are infinitely many of them. □

Remark 9.104 (What is imported, and how much rests on it)*.* Equation (9.140) is not proved in this treatise; it is the same imported fact that Section 9.6.5 uses to fix the connection coefficients at a turning point of higher order, and it is flagged there for the same reason. Only the constant $-\pi/4$ in the phase depends on it. The π spacing — which is what converts a count of nodal circles into a frequency, and is all that Chapter 35 needs for the square law of a circular plate — follows from Equation (9.138) alone, since the coefficient there tends to 1 and the oscillation of $u'' + u = 0$ has half-period exactly π.

9.8.8 The modified Bessel functions

Replacing x by $\mathrm{i}x$ in Equation (9.125) reverses the sign of the term that makes the solutions oscillate, and produces the equation obeyed by the radial part of a field that decays rather than propagates — the evanescent interior of a waveguide below cutoff, the screened potential around a charge, the edge-localised part of the deflection of a vibrating plate. This treatise has already used its solutions, in Section 9.6.5; they are defined here.

Definition 9.105 (Modified Bessel functions)**.** The *modified Bessel equation* of order ν is

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[x\,\frac{\mathrm{d}y}{\mathrm{d}x}(x)\right]-\left[x+\frac{\nu^2}{x}\right]y(x)=0\,, \tag{9.141}$$

equivalently $x^2y''+xy'-(x^2+\nu^2)y=0$. Its Sturm–Liouville data in the convention of Definition 9.54 are $p(x)=x$, $q(x)=-x$, $\lambda^2=-\nu^2$, $r(x)=1/x$. The *modified Bessel function of the first kind* is

$$I_\nu(x)=\sum_{j=0}^{\infty}\frac{1}{j!\,\Gamma(j+\nu+1)}\left(\frac{x}{2}\right)^{2j+\nu}\,,\qquad x>0\,, \tag{9.142}$$

the series of Equation (9.127) with the alternating sign removed, so that $I_\nu(x)=\mathrm{i}^{-\nu}J_\nu(\mathrm{i}x)$. The *modified Bessel function of the second kind* is

$$K_\nu(x)=\frac{\pi}{2}\,\frac{I_{-\nu}(x)-I_\nu(x)}{\sin\nu\pi}\,, \tag{9.143}$$

read at integer order as the limit of that expression, exactly as N_ν is in Definition 9.93.

Rests on Definitions 9.54 and 9.93.

Proposition 9.106 (The modified series and what it does not do)**.** *I_ν of Equation (9.142) converges for every x and solves Equation (9.141). For $\nu\geq 0$ it is positive and strictly increasing on $x>0$, with $I_\nu(0^+)=0$ for $\nu>0$ and $I_0(0)=1$; in particular it has no positive zero, so it carries no oscillation and no eigenvalue condition. Together with K_ν it spans the solution space of Equation (9.141).*

Rests on Definition 9.105, Proposition 9.94, and Theorem 9.13.

Derivation. *Derives Proposition 9.106.*

Convergence is the ratio test (Proposition 7.48) exactly as for Equation (9.127), the moduli of the terms being the same. For the equation, run the computation of Proposition 9.94 with $c_j=\left[j!\,\Gamma(j+\nu+1)2^{2j+\nu}\right]^{-1}$ and with the sign of the term x^2y reversed: the coefficient of $x^{2j+\nu}$ becomes $4j(j+\nu)c_j-c_{j-1}$, and the identity $j!\,\Gamma(j+\nu+1)=j\,(j-1)!\,(j+\nu)\,\Gamma(j+\nu)$ used there gives $4j(j+\nu)c_j=c_{j-1}$, so it vanishes; the term $j=0$ vanishes on its own.

Every coefficient in Equation (9.142) is strictly positive for $\nu\geq 0$, and so is every term for $x>0$; the same is true of the term-by-term derivative, legitimate inside the radius of convergence (Theorem 7.51). Hence $I_\nu>0$ and $I_\nu'>0$ on $x>0$, which forbids a zero there. The leading power is $(x/2)^\nu/\Gamma(\nu+1)$, giving the stated behaviour at the origin. Finally I_ν and K_ν are independent — one is bounded at infinity and the other is not, by Remark 9.107 — so by Theorem 9.13 they span the two-dimensional solution space. □

Remark 9.107 (Growth, and where it is imported from)*.* The large-argument behaviour

$$I_\nu(x)\sim\frac{\mathrm{e}^x}{\sqrt{2\pi x}}\,,\qquad K_\nu(x)\sim\sqrt{\frac{\pi}{2x}}\,\mathrm{e}^{-x}\,,\qquad x\to\infty\,, \tag{9.144}$$

is quoted from [1023]; the second half of it is the fact already imported in Section 9.6.5, so nothing new is being taken on credit. The qualitative half of Equation (9.144) — that I_ν

grows without bound and K_ν decays — does not depend on the import: I_ν is a positive increasing function whose terms already exceed those of $e^x/2^\nu$ term by term for ν an integer, and K_ν is the solution that Proposition 9.106 leaves over. This is what licenses the reading used in the physical parts: in a superposition $A\,J_m(kr) + B\,I_m(kr)$ on a disc, the second term is negligible except within a few $1/k$ of the rim, so the interior nodal pattern is that of J_m alone.

9.9 The Legendre equation and spherical harmonics

9.9.1 The Legendre equation

Sturm–Liouville data

The identification of the operator data is

$$p(x) = 1 - x^2\,, \qquad q(x) = n(n+1)\,, \qquad \lambda^2 = 0\,, \qquad r(x) = 1\,. \tag{9.145}$$

This is the source's identification and it is the one reproduced in Table 9.2, but with $\lambda^2 = 0$ it is not an eigenvalue problem: nothing in it distinguishes one degree from another. Theorem 9.111 therefore reads the same equation the other way round, with $q = 0$ and $\lambda^2 = n(n+1)$, exactly as Remark 9.97 does for the Bessel equation.

Homogeneous solution

The homogeneous equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[(1-x^2)\,\frac{\mathrm{d}y}{\mathrm{d}x}(x)\right] + n(n+1)\,y(x) = 0\,, \tag{9.146}$$

whose solution is

$$y(x) = A\,P_n(x) + B\,Q_n(x)\,, \tag{9.147}$$

with P_n the Legendre polynomials and Q_n the Legendre functions of the second kind.

The Legendre polynomials

Definition 9.108 (Rodrigues formula)**.** The *Legendre polynomial* of degree n is

$$P_n(x) = \frac{1}{2^n\,n!}\,\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(x^2-1\right)^n\,, \qquad n = 0, 1, 2, \dots \tag{9.148}$$

Proposition 9.109 (Rodrigues' polynomials solve the equation)**.** *P_n of Equation (9.148) is a polynomial of degree n with leading coefficient $(2n)!/\left[2^n (n!)^2\right]$, it satisfies $P_n(1) = 1$, and it solves the Legendre equation Equation (9.146).*

Rests on Definition 9.108 and Equation (9.146).

Derivation. *Derives Proposition 9.109.*

Write $v = \left(x^2-1\right)^n$, a polynomial of degree $2n$ whose n-th derivative has degree n; the coefficient of x^n comes from $\mathrm{d}^n x^{2n}/\mathrm{d}x^n = [(2n)!/n!]\,x^n$, which after the division by $2^n n!$ gives the stated leading coefficient. For the value at 1, factor $v = (x-1)^n (x+1)^n$: in the Leibniz expansion of $\mathrm{d}^n v/\mathrm{d}x^n$ every term retains at least one factor $x-1$ except the one in which all n derivatives fall on $(x-1)^n$, so at $x = 1$ only $n!\,(x+1)^n = n!\,2^n$ survives, and division by $2^n n!$ gives $P_n(1) = 1$.

For the equation, start from the identity $\left(x^2-1\right)\frac{\mathrm{d}v}{\mathrm{d}x} = 2n\,x\,v$, which is immediate from the definition of v, and differentiate it $n+1$ times by the Leibniz rule Equation (7.16). On the left only three terms survive, because x^2-1 is quadratic:

$$\left(x^2-1\right)v^{(n+2)} + (n+1)\,2x\,v^{(n+1)} + \frac{(n+1)n}{2}\,2\,v^{(n)}\,,$$

and on the right only two, because x is linear: $2n\,x\,v^{(n+1)} + 2n(n+1)\,v^{(n)}$. Writing $u = v^{(n)}$ and collecting,

$$\left(x^2 - 1\right)\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} + 2x\,\frac{\mathrm{d}u}{\mathrm{d}x} - n(n+1)\,u = 0\,,$$

which on multiplication by -1 is $\left[(1-x^2)u'\right]' + n(n+1)u = 0$, that is Equation (9.146). Since P_n is a constant multiple of u, it solves the same linear equation. □

Lemma 9.110 (The polynomial solution is unique up to scale)**.** *Every polynomial solution of Equation (9.146) is a constant multiple of P_n.*

Rests on Proposition 9.109.

Derivation. *Derives Lemma 9.110.*

Let $\Lambda[y] = \left[(1-x^2)y'\right]'$. On monomials

$$\Lambda\left[x^k\right] = \frac{\mathrm{d}}{\mathrm{d}x}\left[k\,x^{k-1} - k\,x^{k+1}\right] = -k(k+1)\,x^k + k(k-1)\,x^{k-2}\,,$$

so Λ maps the space of polynomials of degree at most N into itself and is triangular in the monomial basis ordered by degree, with diagonal entries $-k(k+1)$ for $k = 0, \dots, N$. Those entries are pairwise distinct, so each eigenvalue of Λ on that space is simple and its eigenspace is one-dimensional. A polynomial solution of Equation (9.146) of degree at most N is an eigenvector for the eigenvalue $-n(n+1)$, and P_n is one such by Proposition 9.109; taking N at least as large as the degree of the given solution identifies the two up to a constant. □

Orthogonality

Theorem 9.111 (Orthogonality of the Legendre polynomials)**.** *The Legendre polynomials are orthogonal on $[-1, 1]$:*

$$\int_{-1}^{1} P_m(x)\,P_n(x)\,\mathrm{d}x = \frac{2}{2n+1}\,\delta_{mn}\,. \tag{9.149}$$

Rests on Theorem 9.57 and Proposition 9.109.

Derivation. *Derives Theorem 9.111.*

For the off-diagonal terms read Equation (9.146) as the eigenvalue problem it is: $p(x) = 1 - x^2$, $q = 0$, weight $r = 1$ and eigenvalue $\lambda^2 = n(n+1)$, distinct for distinct non-negative degrees. The coefficient function p vanishes at both endpoints while P_m, P_n and their derivatives are polynomials and hence bounded there, so Equation (9.57) holds by the second mechanism of Remark 9.58 and Theorem 9.57 gives the vanishing of the integral for $m \neq n$.

For the diagonal use Equation (9.148) and integrate by parts n times (Equation (7.28)). At each step the integrated term carries a factor $\mathrm{d}^j(x^2-1)^n/\mathrm{d}x^j$ with $j < n$, which retains a factor (x^2-1) and so vanishes at $x = \pm 1$. Hence

$$\int_{-1}^{1} P_n^2\,\mathrm{d}x = \frac{(-1)^n}{(2^n n!)^2}\int_{-1}^{1}\left(x^2-1\right)^n\frac{\mathrm{d}^{2n}}{\mathrm{d}x^{2n}}\left(x^2-1\right)^n\mathrm{d}x = \frac{(2n)!}{(2^n n!)^2}\int_{-1}^{1}\left(1-x^2\right)^n\mathrm{d}x\,,$$

since the $2n$-th derivative of a monic-leading polynomial of degree $2n$ is the constant $(2n)!$ and $(-1)^n(x^2-1)^n = (1-x^2)^n$. Write $I_n = \int_{-1}^{1}(1-x^2)^n\,\mathrm{d}x$. Integration by parts gives

$$I_n = \left[x\left(1-x^2\right)^n\right]_{-1}^{1} + 2n\int_{-1}^{1} x^2\left(1-x^2\right)^{n-1}\mathrm{d}x = 2n\left(I_{n-1} - I_n\right)\,,$$

the boundary term vanishing and $x^2 = 1 - (1 - x^2)$ splitting the integral; hence $I_n = [2n/(2n+1)]I_{n-1}$, and with $I_0 = 2$,

$$I_n = 2\prod_{k=1}^{n} \frac{2k}{2k+1} = \frac{2^{2n+1}\,(n!)^2}{(2n+1)!}\,,$$

the closed form following from $\prod_{k\le n} 2k = 2^n n!$ and $\prod_{k\le n}(2k+1) = (2n+1)!/\,(2^n n!)$. Multiplying,

$$\int_{-1}^{1} P_n^2\,\mathrm{d}x = \frac{(2n)!}{2^{2n}\,(n!)^2}\cdot\frac{2^{2n+1}\,(n!)^2}{(2n+1)!} = \frac{2\,(2n)!}{(2n+1)!} = \frac{2}{2n+1}\,.$$

□

Remark 9.112 (There is no such relation for the Q_n). The source reserves headings for an orthogonality relation of the second-kind solutions Q_n, and for one of the general solution Equation (9.147). Neither exists, and the reason is the warning at the end of Remark 9.58: p vanishing at an endpoint is not enough when the solutions do not stay bounded there. The Q_n diverge logarithmically at $x = \pm 1$ — $Q_0(x) = \frac{1}{2}\ln[(1+x)/(1-x)]$ and $Q_n = P_nQ_0 -$ (a polynomial) — and a short computation with $u = Q_0$, $v = Q_2$ shows that the surviving part of the bracket is $\frac{3}{2}\,p\,\mathrm{d}Q_0/\mathrm{d}x = \frac{3}{2}$ at $x \to 1$ and $-\frac{3}{2}$ at $x \to -1$, so Equation (9.57) equals 3 and not 0. The identity in the proof of Theorem 9.57 then returns $\int_{-1}^{1} Q_0Q_2\,\mathrm{d}x = -\frac{1}{2}$ rather than zero. Orthogonality on $[-1,1]$ is a property of the solutions that remain bounded at the two singular endpoints, and that requirement is exactly what selects the P_n and makes the Legendre problem a self-adjoint one.

Generating function

Theorem 9.113 (Generating function)**.** *For $|x| \le 1$ and $|t| < 1$ the Legendre polynomials are generated by*

$$g(x,t) = \frac{1}{\sqrt{1-2xt+t^2}} = \sum_{n=0}^{\infty} P_n(x)\,t^n\,. \tag{9.150}$$

Rests on Lemma 9.110 and Proposition 9.109.

Derivation. *Derives Theorem 9.113.*

Write $W = 1 - 2xt + t^2$, so that $g = W^{-1/2}$. For $|x| \le 1$ the roots of W in t are $\mathrm{e}^{\pm\mathrm{i}\theta}$ with $x = \cos\theta$, both of modulus one, so g is holomorphic in t on $|t| < 1$ and has there a Taylor expansion $g = \sum_{n\ge 0} p_n(x)\,t^n$ (Theorem 8.20). Expanding $[1 - t(2x - t)]^{-1/2}$ by the binomial series and collecting powers of t shows each p_n to be a polynomial in x of degree at most n.

Next, g satisfies the identity

$$\frac{\partial}{\partial x}\left[\left(1-x^2\right)\frac{\partial g}{\partial x}\right] + \frac{\partial}{\partial t}\left[t^2\,\frac{\partial g}{\partial t}\right] = 0\,.$$

Indeed $\partial g/\partial x = t\,W^{-3/2}$ and $\partial g/\partial t = (x-t)\,W^{-3/2}$, so the first bracket differentiates to $-2xt\,W^{-3/2} + 3\,(1-x^2)\,t^2W^{-5/2}$ and the second to $2t(x-t)W^{-3/2} - t^2W^{-3/2} + 3t^2(x-t)^2W^{-5/2}$. The terms in $W^{-3/2}$ add to $-3t^2W^{-3/2}$, and those in $W^{-5/2}$ add to $3t^2[(1-x^2) + (x-t)^2]W^{-5/2} = 3t^2W\cdot W^{-5/2}$, which cancels the first.

Substituting the Taylor series in that identity and equating coefficients of t^n, using $\partial_t[t^2\partial_t g] = \sum_n n(n+1)p_nt^n$, gives

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\left(1-x^2\right)\frac{\mathrm{d}p_n}{\mathrm{d}x}\right] + n(n+1)\,p_n = 0\,,$$

which is Equation (9.146). So p_n is a polynomial solution of the Legendre equation of that degree, hence a constant multiple of P_n by Lemma 9.110. The constant is 1 because $g(1,t) = (1-t)^{-1} = \sum_n t^n$ gives $p_n(1) = 1$, and $P_n(1) = 1$ by Proposition 9.109. □

Legendre series and recurrence relations

Proposition 9.114 (Recurrence relations)**.** *For $n \geq 1$,*

$$(n+1)\,P_{n+1}(x) = (2n+1)\,x\,P_n(x) - n\,P_{n-1}(x)\,, \tag{9.151}$$

$$\frac{\mathrm{d}P_{n+1}}{\mathrm{d}x}(x) - \frac{\mathrm{d}P_{n-1}}{\mathrm{d}x}(x) = (2n+1)\,P_n(x)\,, \tag{9.152}$$

with $P_0(x) = 1$ and $P_1(x) = x$.

Rests on Theorem 9.113.

Derivation. *Derives Proposition 9.114.*

Differentiating Equation (9.150) with respect to t gives $W\,\partial g/\partial t = (x-t)\,g$, and with respect to x gives $W\,\partial g/\partial x = t\,g$. Substituting the series and reading off the coefficient of t^n turns the first into

$$(n+1)P_{n+1} - 2n\,x\,P_n + (n-1)P_{n-1} = x\,P_n - P_{n-1}\,,$$

which is Equation (9.151), and the second into

$$\frac{\mathrm{d}P_n}{\mathrm{d}x} - 2x\,\frac{\mathrm{d}P_{n-1}}{\mathrm{d}x} + \frac{\mathrm{d}P_{n-2}}{\mathrm{d}x} = P_{n-1}\,. \tag{9.153}$$

Differentiating Equation (9.151) gives $(n+1)P'_{n+1} = (2n+1)P_n + (2n+1)xP'_n - nP'_{n-1}$, and Equation (9.153) at index $n+1$ gives $2xP'_n = P'_{n+1} + P'_{n-1} - P_n$. Eliminating xP'_n between the two leaves $\frac{1}{2}P'_{n+1} = \frac{1}{2}(2n+1)P_n + \frac{1}{2}P'_{n-1}$, which is Equation (9.152). The initial values follow from $g(x,t) = 1 + xt + O\,(t^2)$. □

Proposition 9.115 (Legendre series)**.** *Let f on $[-1,1]$ admit an expansion*

$$f(x) = \sum_{n=0}^{\infty} a_n\,P_n(x) \tag{9.154}$$

convergent in the mean. Then

$$a_n = \frac{2n+1}{2}\int_{-1}^{1} f(x)\,P_n(x)\,\mathrm{d}x\,. \tag{9.155}$$

Rests on Theorem 9.111 and Definition 9.47.

Derivation. *Derives Proposition 9.115.*

Multiply Equation (9.154) by P_m and integrate over $[-1,1]$; mean convergence permits the exchange of sum and integral, and Equation (9.149) leaves the single term $n = m$ with the factor $2/(2m+1)$, which inverts to Equation (9.155). That an expansion exists for every square-integrable f — the completeness of the system — is again a Sturm–Liouville theorem and not proved here, exactly as in Remark 9.101. □

9.9.2 The associated Legendre equation

Sturm–Liouville data

The identification of the operator data is

$$p(x) = 1 - x^2\,, \qquad q(x) = n(n+1)\,, \qquad \lambda^2 = -m^2\,, \qquad r(x) = \frac{1}{1-x^2}\,. \tag{9.156}$$

General solution

The homogeneous equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[(1-x^2)\,\frac{\mathrm{d}y}{\mathrm{d}x}(x)\right]+\left[n(n+1)-\frac{m^2}{1-x^2}\right]y(x)=0\,, \tag{9.157}$$

whose solution is

$$y(x)=A\,P_n^m(x)+B\,Q_n^m(x)\,, \tag{9.158}$$

with P_n^m the associated Legendre functions and Q_n^m the associated Legendre functions of the second kind.

Construction from the Legendre polynomials

Definition 9.116 (Associated Legendre functions)**.** For integers $0\le m\le n$ and $|x|\le 1$,

$$P_n^m(x)=\left(1-x^2\right)^{m/2}\,\frac{\mathrm{d}^m P_n}{\mathrm{d}x^m}(x)\,, \tag{9.159}$$

so that $P_n^0=P_n$ and $P_n^m=0$ for $m>n$, the m-th derivative of a polynomial of degree n then vanishing.

Rests on Definition 9.108.

Lemma 9.117 (The differentiated Legendre equation)**.** *If y solves Equation (9.146) then $u=\mathrm{d}^m y/\mathrm{d}x^m$ solves*

$$\left(1-x^2\right)\frac{\mathrm{d}^2u}{\mathrm{d}x^2}-2(m+1)\,x\,\frac{\mathrm{d}u}{\mathrm{d}x}+\left[n(n+1)-m(m+1)\right]u=0\,. \tag{9.160}$$

Rests on Equation (9.146).

Derivation. *Derives Lemma 9.117.*

Differentiate $(1-x^2)\,y''-2xy'+n(n+1)y=0$ exactly m times by the Leibniz rule Equation (7.16). The first term contributes $(1-x^2)\,y^{(m+2)}-2mx\,y^{(m+1)}-m(m-1)y^{(m)}$, the second $-2x\,y^{(m+1)}-2m\,y^{(m)}$, and the third $n(n+1)y^{(m)}$; only three and two terms survive respectively because $1-x^2$ is quadratic and x linear. Collecting, the coefficient of $y^{(m+1)}$ is $-2(m+1)x$ and that of $y^{(m)}$ is $n(n+1)-m(m-1)-2m=n(n+1)-m(m+1)$. □

Proposition 9.118 (The construction solves the associated equation)**.** *P_n^m of Equation (9.159) solves Equation (9.157).*

Rests on Definition 9.116 and Lemma 9.117.

Derivation. *Derives Proposition 9.118.*

Put $\varpi=(1-x^2)^{m/2}$ and $y=\varpi\,u$ with $u=\mathrm{d}^m P_n/\mathrm{d}x^m$, so that $\varpi'/\varpi=-mx/\left(1-x^2\right)$ and

$$\frac{\varpi''}{\varpi}=-\frac{m}{1-x^2}+\frac{(m^2-2m)\,x^2}{(1-x^2)^2}\,.$$

Substituting $y=\varpi u$ into the associated equation Equation (9.157), written out as $(1-x^2)\,y''-2xy'+\left[n(n+1)-m^2/(1-x^2)\right]y=0$, and dividing by ϖ leaves an equation for u whose coefficient of $\mathrm{d}u/\mathrm{d}x$ is $2\,(1-x^2)\,\varpi'/\varpi-2x=-2(m+1)x$ and whose coefficient of u is

$$\left(1-x^2\right)\frac{\varpi''}{\varpi}-2x\,\frac{\varpi'}{\varpi}+n(n+1)-\frac{m^2}{1-x^2}=-m+n(n+1)+\frac{m^2\,(x^2-1)}{1-x^2}\,,$$

that is $n(n+1)-m(m+1)$. The result is exactly Equation (9.160), which u satisfies by Lemma 9.117. □

Orthogonality

Theorem 9.119 (Orthogonality at fixed order)**.** *At fixed order m, the associated Legendre functions are orthogonal on $[-1,1]$:*

$$\int_{-1}^{1} P_n^m(x)\, P_{n'}^m(x)\,\mathrm{d}x = \frac{2}{2n+1}\,\frac{(n+m)!}{(n-m)!}\,\delta_{nn'}\,. \tag{9.161}$$

Rests on Theorems 9.57 and 9.111 and Proposition 9.118.

Derivation. *Derives Theorem 9.119.*

For $n \neq n'$ read Equation (9.157) at fixed m as the eigenvalue problem with $p(x) = 1-x^2$, $q(x) = -m^2/(1-x^2)$, weight $r = 1$ and eigenvalue $\lambda^2 = n(n+1)$. The functions P_n^m carry the factor $(1-x^2)^{m/2}$, so for $m \geq 1$ both they and p times their derivatives vanish at $x = \pm 1$, while for $m = 0$ the argument is that of Theorem 9.111; either way Equation (9.57) holds and Theorem 9.57 applies.

For the diagonal put $I_m = \int_{-1}^{1} (1-x^2)^m \left[u_m(x)\right]^2 \mathrm{d}x$ with $u_m = \mathrm{d}^m P_n/\mathrm{d}x^m$, which is the integral to be computed because $(P_n^m)^2 = (1-x^2)^m\, u_m^2$. For $m \geq 1$ write $u_m = \mathrm{d}u_{m-1}/\mathrm{d}x$ and integrate by parts (Equation (7.28)); the boundary term carries the factor $(1-x^2)^m$ and vanishes, so

$$I_m = -\int_{-1}^{1} u_{m-1}\,\frac{\mathrm{d}}{\mathrm{d}x}\left[\left(1-x^2\right)^m \frac{\mathrm{d}u_{m-1}}{\mathrm{d}x}\right]\mathrm{d}x = -\int_{-1}^{1} \left(1-x^2\right)^{m-1} u_{m-1}\,\Theta\,\mathrm{d}x\,,$$

where $\Theta = (1-x^2)\,\mathrm{d}^2u_{m-1}/\mathrm{d}x^2 - 2m\,x\,\mathrm{d}u_{m-1}/\mathrm{d}x$. By Equation (9.160) at order $m-1$ that combination is $-\left[n(n+1) - m(m-1)\right]u_{m-1}$, so

$$I_m = \left[n(n+1) - m(m-1)\right] I_{m-1} = (n+m)(n-m+1)\, I_{m-1}\,.$$

Iterating from $I_0 = 2/(2n+1)$, which is Theorem 9.111, gives

$$I_m = \frac{2}{2n+1}\prod_{k=1}^{m}(n+k)(n-k+1) = \frac{2}{2n+1}\cdot\frac{(n+m)!}{n!}\cdot\frac{n!}{(n-m)!}\,,$$

which is Equation (9.161). □

Recurrence relations

Proposition 9.120 (Recurrence in the degree at fixed order)**.** *For $1 \leq m \leq n$,*

$$(n-m+1)\,P_{n+1}^m(x) = (2n+1)\,x\,P_n^m(x) - (n+m)\,P_{n-1}^m(x)\,. \tag{9.162}$$

Rests on Proposition 9.114 and Definition 9.116.

Derivation. *Derives Proposition 9.120.*

Write $D = \mathrm{d}/\mathrm{d}x$. Applying D^m to Equation (9.151) by the Leibniz rule Equation (7.16), in which the factor x contributes only two terms,

$$(n+1)\,D^m P_{n+1} = (2n+1)\left[x\,D^m P_n + m\,D^{m-1}P_n\right] - n\,D^m P_{n-1}\,,$$

while applying D^{m-1} to Equation (9.152) gives $D^m P_{n+1} - D^m P_{n-1} = (2n+1)\,D^{m-1}P_n$. Substituting the second into the first eliminates $D^{m-1}P_n$ and leaves

$$(n-m+1)\,D^m P_{n+1} = (2n+1)\,x\,D^m P_n - (n+m)\,D^m P_{n-1}\,,$$

and multiplying through by $(1-x^2)^{m/2}$ turns each $D^m P_k$ into P_k^m by Equation (9.159). □

9.9.3 Spherical harmonics

The associated Legendre functions in $\cos\theta$, combined with the complex exponentials of Section 9.7 in the azimuthal angle φ, form the *spherical harmonics*

$$Y_n^m(\theta,\varphi) = \sqrt{\frac{2n+1}{4\pi}\,\frac{(n-m)!}{(n+m)!}}\;P_n^m(\cos\theta)\,\mathrm{e}^{\mathrm{i}m\varphi}\,, \tag{9.163}$$

which are orthonormal on the unit sphere [1976],

$$\int_0^{2\pi}\mathrm{d}\varphi\int_0^{\pi}\overline{Y_{n'}^{m'}(\theta,\varphi)}\,Y_n^m(\theta,\varphi)\,\sin\theta\,\mathrm{d}\theta = \delta_{nn'}\,\delta_{mm'}\,. \tag{9.164}$$

Proposition 9.121 (Where the construction comes from)**.** *Separating the angular part of the Laplace operator on the unit sphere as $Y(\theta,\varphi)=\Theta(\theta)\,\Phi(\varphi)$ forces $\Phi(\varphi)=\mathrm{e}^{\mathrm{i}m\varphi}$ with m an integer, and $\Theta(\theta)=P_n^m(\cos\theta)$ with n a non-negative integer and $|m|\le n$. Every other choice of the two separation constants gives a solution that is either multivalued in φ or unbounded at one of the poles.*

Rests on Equations (9.107) and (9.157).

Derivation. Derives Proposition 9.121.

The angular eigenvalue equation is

$$\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left[\sin\theta\,\frac{\partial Y}{\partial\theta}\right]+\frac{1}{\sin^2\theta}\,\frac{\partial^2 Y}{\partial\varphi^2} = -\Lambda\,Y\,.$$

Substituting the product and dividing by $\Theta\Phi/\sin^2\theta$ separates it: the φ part is $\mathrm{d}^2\Phi/\mathrm{d}\varphi^2 = -m^2\Phi$, whose solutions are the exponentials Equation (9.107), and single-valuedness under $\varphi\mapsto\varphi+2\pi$ restricts m to the integers. The θ part, in the variable $x=\cos\theta$ — for which $\mathrm{d}/\mathrm{d}\theta = -\sin\theta\,\mathrm{d}/\mathrm{d}x$ and $\sin^2\theta = 1-x^2$ — is exactly the associated Legendre equation Equation (9.157) with $n(n+1)$ replaced by Λ. Boundedness at the two poles $x=\pm1$, the singular endpoints of that equation in the sense of Remark 9.17, selects the solutions P_n^m: for $\Lambda=n(n+1)$ with n a non-negative integer and $|m|\le n$ these are bounded by Equation (9.159), and the second solution Q_n^m diverges there, as does every solution for any other value of Λ. □

Theorem 9.122 (Normalisation and orthonormality)**.** *With the constant displayed in Equation (9.163) the functions Y_n^m satisfy Equation (9.164), and that constant is the only positive one that does so.*

Rests on Theorem 9.119 and Equation (9.109).

Derivation. Derives Theorem 9.122.

The integral factorises. The azimuthal factor is

$$\int_0^{2\pi}\mathrm{e}^{\mathrm{i}(m-m')\varphi}\,\mathrm{d}\varphi = 2\pi\,\delta_{mm'}\,,$$

which is Equation (9.109) on an interval of one period, so every term with $m\neq m'$ vanishes whatever the polar factor does. At $m=m'$ the polar factor is, substituting $x=\cos\theta$ and $\mathrm{d}x=-\sin\theta\,\mathrm{d}\theta$,

$$\int_0^{\pi}P_{n'}^m(\cos\theta)\,P_n^m(\cos\theta)\,\sin\theta\,\mathrm{d}\theta = \int_{-1}^{1}P_{n'}^m(x)\,P_n^m(x)\,\mathrm{d}x = \frac{2}{2n+1}\,\frac{(n+m)!}{(n-m)!}\,\delta_{nn'}$$

by Theorem 9.119. The product of the two factors is $2\pi\cdot 2(n+m)!/\big[(2n+1)(n-m)!\big]$ times $\delta_{nn'}\delta_{mm'}$, and the squared modulus of the constant in Equation (9.163) is the reciprocal of that number, which gives Equation (9.164). Since the constant enters squared and is required positive, it is unique. □

Remark 9.123 (Negative order, and completeness). Equation (9.159) defines P_n^m for $m \geq 0$ only, and Equation (9.163) is extended to $-n \leq m < 0$ by

$$P_n^{-m}(x) = (-1)^m \, \frac{(n-m)!}{(n+m)!} \, P_n^m(x) \,, \qquad 0 \leq m \leq n \,, \tag{9.165}$$

a convention — the Condon–Shortley one — chosen so that Equation (9.163) then satisfies $\overline{Y_n^m} = (-1)^m \, Y_n^{-m}$ and so that Theorem 9.122 continues to hold with the same formula, as substituting Equation (9.165) into the two factorials shows. That the Y_n^m are moreover *complete* in the square-integrable functions on the sphere — so that every such function has an expansion in them — is not proved here: it is the statement that the angular Laplace operator has no other eigenfunctions, which belongs with the general spectral theory of Chapter 12 [1222]. The addition theorem, which expresses $P_n(\cos\gamma)$ for the angle γ between two directions as a sum of products of spherical harmonics of those two directions, rests on the rotation behaviour of the space spanned by the Y_n^m at fixed n and is not derived here either; it is used in this treatise only through [1976].

9.10 The Hermite equation

The source announces the Hermite family with the same systematic headings as the families above — Sturm–Liouville data p, q and r, general solution, recurrence relations, orthonormality, and Hermite series — but every one of those subsections is empty, so they are supplied here. This family governs the stationary states of the quantum harmonic oscillator, where the Hermite polynomials appear multiplied by a Gaussian weight [1593, 2313].

9.10.1 Sturm–Liouville data

The Hermite equation is

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2}(x) - 2x \, \frac{\mathrm{d}y}{\mathrm{d}x}(x) + 2n \, y(x) = 0 \,, \qquad x \in \mathbb{R} \,. \tag{9.166}$$

Multiplying it by the integrating factor e^{-x^2} puts it in the form Equation (9.54), since $\frac{\mathrm{d}}{\mathrm{d}x}\left[\mathrm{e}^{-x^2} \mathrm{d}y/\mathrm{d}x\right] = \mathrm{e}^{-x^2}\left[y'' - 2xy'\right]$, with data

$$p(x) = \mathrm{e}^{-x^2} \,, \qquad q(x) = 0 \,, \qquad \lambda^2 = 2n \,, \qquad r(x) = \mathrm{e}^{-x^2} \,, \tag{9.167}$$

on the whole line. The weight is the Gaussian, and the interval is unbounded, so the third mechanism of Remark 9.58 is the one that will apply.

9.10.2 Solutions and generating function

Definition 9.124 (Hermite polynomials)**.**

$$H_n(x) = (-1)^n \, \mathrm{e}^{x^2} \, \frac{\mathrm{d}^n}{\mathrm{d}x^n} \, \mathrm{e}^{-x^2} \,, \qquad n = 0, 1, 2, \ldots \tag{9.168}$$

Proposition 9.125 (Generating function and first properties)**.** *The polynomials Equation (9.168) are generated by*

$$\mathrm{e}^{2xt - t^2} = \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} \, t^n \,, \tag{9.169}$$

and H_n is a polynomial of degree n with leading coefficient 2^n and parity $(-1)^n$.
Rests on Definition 9.124.

Derivation. *Derives Proposition 9.125.*

Complete the square: $2xt-t^2 = x^2-(x-t)^2$, so the left of Equation (9.169) is $\mathrm{e}^{x^2}\,\mathrm{e}^{-(x-t)^2}$. The second factor is a function of $x-t$ alone, so $\partial/\partial t$ acting on it equals $-\partial/\partial x$; hence its n-th t-derivative at $t=0$ is $(-1)^n \mathrm{d}^n\mathrm{e}^{-x^2}/\mathrm{d}x^n$, and the Taylor expansion in t of a function entire in t (Theorem 8.20) gives

$$\mathrm{e}^{x^2}\mathrm{e}^{-(x-t)^2} = \sum_{n\geq 0} \frac{t^n}{n!}\,(-1)^n\,\mathrm{e}^{x^2}\,\frac{\mathrm{d}^n\mathrm{e}^{-x^2}}{\mathrm{d}x^n}\,,$$

which is Equation (9.169) with Equation (9.168). Differentiating Equation (9.168) once more, and using $\mathrm{d}\mathrm{e}^{-x^2}/\mathrm{d}x = -2x\,\mathrm{e}^{-x^2}$, gives $H_{n+1} = 2x\,H_n - \mathrm{d}H_n/\mathrm{d}x$; with $H_0 = 1$ this shows by induction that H_n is a polynomial of degree n whose leading coefficient doubles at each step, hence equals 2^n. The parity follows from $\mathrm{e}^{2(-x)t-t^2} = \mathrm{e}^{2x(-t)-(-t)^2}$, which sends $H_n(-x)$ to $(-1)^n H_n(x)$. □

Proposition 9.126 (Recurrence relations and the equation)**.** *For $n \geq 1$,*

$$\frac{\mathrm{d}H_n}{\mathrm{d}x}(x) = 2n\,H_{n-1}(x)\,, \tag{9.170}$$

$$H_{n+1}(x) = 2x\,H_n(x) - 2n\,H_{n-1}(x)\,, \tag{9.171}$$

and H_n solves the Hermite equation Equation (9.166).

Rests on Proposition 9.125 and Equation (9.166).

Derivation. *Derives Proposition 9.126.*

Differentiating Equation (9.169) with respect to x gives $2t\,\mathrm{e}^{2xt-t^2} = \sum_n (\mathrm{d}H_n/\mathrm{d}x)t^n/n!$, and comparing coefficients of t^n gives $\mathrm{d}H_n/\mathrm{d}x \;/\; n! = 2H_{n-1}/(n-1)!$, which is Equation (9.170). Combining it with $H_{n+1} = 2xH_n - \mathrm{d}H_n/\mathrm{d}x$, established in the proof of Proposition 9.125, gives Equation (9.171). For the equation, differentiate that same relation: $\mathrm{d}H_{n+1}/\mathrm{d}x = 2H_n + 2x\,\mathrm{d}H_n/\mathrm{d}x - \mathrm{d}^2H_n/\mathrm{d}x^2$, and use Equation (9.170) on the left, which turns it into $2(n+1)H_n$. Cancelling $2H_n$ from both sides leaves $\mathrm{d}^2H_n/\mathrm{d}x^2 - 2x\,\mathrm{d}H_n/\mathrm{d}x + 2n\,H_n = 0$. □

9.10.3 Orthogonality and Hermite series

Theorem 9.127 (Orthogonality)**.**

$$\int_{-\infty}^{\infty} H_m(x)\,H_n(x)\,\mathrm{e}^{-x^2}\,\mathrm{d}x = 2^n\,n!\,\sqrt{\pi}\,\delta_{mn}\,. \tag{9.172}$$

Rests on Theorem 9.57, Proposition 9.126, and Equation (9.167).

Derivation. *Derives Theorem 9.127.*

For $m \neq n$ apply Theorem 9.57 with the data Equation (9.167): the eigenvalues $2m$ and $2n$ differ, and the boundary term Equation (9.57) vanishes because $p = \mathrm{e}^{-x^2}$ multiplies polynomials and e^{-x^2} beats every polynomial at both ends of the line — the third mechanism of Remark 9.58.

For $m = n$ insert Equation (9.168) for one factor and integrate by parts n times (Equation (7.28)), every boundary term vanishing for the same reason:

$$\int_{-\infty}^{\infty} H_n^2\,\mathrm{e}^{-x^2}\,\mathrm{d}x = (-1)^n \int_{-\infty}^{\infty} H_n(x)\,\frac{\mathrm{d}^n\mathrm{e}^{-x^2}}{\mathrm{d}x^n}\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{\mathrm{d}^n H_n}{\mathrm{d}x^n}\,\mathrm{e}^{-x^2}\,\mathrm{d}x\,.$$

By Proposition 9.125 the n-th derivative of H_n is the constant $2^n n!$, and the Gaussian integral $\int_{-\infty}^{\infty}\mathrm{e}^{-x^2}\mathrm{d}x = \sqrt{\pi}$ — its square being $\int\int \mathrm{e}^{-x^2-y^2}\mathrm{d}x\,\mathrm{d}y = \pi$ in plane polar coordinates — completes Equation (9.172). □

Proposition 9.128 (Hermite series)**.** *Let f admit an expansion $f(x) = \sum_{n\geq 0} a_n H_n(x)$ convergent in the mean with respect to the weight e^{-x^2}. Then*

$$a_n = \frac{1}{2^n\, n!\, \sqrt{\pi}} \int_{-\infty}^{\infty} f(x)\, H_n(x)\, \mathrm{e}^{-x^2}\, \mathrm{d}x\,. \tag{9.173}$$

Rests on Theorem 9.127 and Definition 9.47.

Derivation. *Derives Proposition 9.128.*

Multiply by $H_m(x)\mathrm{e}^{-x^2}$, integrate over the line, exchange sum and integral by mean convergence, and use Equation (9.172); the surviving term is $n = m$ and carries the constant $2^m m!\sqrt{\pi}$. Completeness of the system carries the same caveat as Remark 9.101. □

Remark 9.129 (The quantum oscillator, in SI). For the potential $V(x) = \frac{1}{2}m_0\omega^2 x^2$ of Example 9.80, with m_0 in kg and ω in rad/s, the substitution $\xi = x/\ell_0$ with $\ell_0 = \sqrt{\hbar/(m_0\omega)}$ in m turns Equation (9.59) into $\mathrm{d}^2\psi/\mathrm{d}\xi^2 + [2E/(\hbar\omega) - \xi^2]\psi = 0$, and the substitution $\psi = \mathrm{e}^{-\xi^2/2}H(\xi)$ turns that into Equation (9.166) for H with $2n = 2E/(\hbar\omega) - 1$. Square integrability forces the series solution to terminate, that is n to be a non-negative integer, which returns the levels Equation (9.91) $E_n = (n + \frac{1}{2})\hbar\omega$ in J — there exactly, not semiclassically. The mass is written m_0 here only because m is the order of the associated Legendre functions elsewhere in this chapter.

9.11 The Laguerre equation

The source likewise reserves the full set of systematic headings for the Laguerre family — Sturm–Liouville data p, q and r, general solution, recurrence relations, orthonormality, and Laguerre series — all of them empty, and they are supplied here. The associated Laguerre polynomials furnish the radial eigenfunctions of the hydrogen atom [1593, 2313].

9.11.1 Sturm–Liouville data

The Laguerre equation is

$$x\,\frac{\mathrm{d}^2y}{\mathrm{d}x^2}(x) + (1-x)\,\frac{\mathrm{d}y}{\mathrm{d}x}(x) + n\,y(x) = 0\,, \qquad x > 0\,. \tag{9.174}$$

Multiplying by e^{-x} puts it in the form Equation (9.54), since $\frac{\mathrm{d}}{\mathrm{d}x}[x\mathrm{e}^{-x}\mathrm{d}y/\mathrm{d}x] = \mathrm{e}^{-x}[xy'' + (1-x)y']$, with data

$$p(x) = x\,\mathrm{e}^{-x}\,, \qquad q(x) = 0\,, \qquad \lambda^2 = n\,, \qquad r(x) = \mathrm{e}^{-x}\,, \tag{9.175}$$

on the half-line. Here p vanishes at the origin and decays at infinity, so the boundary term dies at both ends at once.

9.11.2 Solutions and generating function

Definition 9.130 (Laguerre polynomials)**.**

$$L_n(x) = \frac{\mathrm{e}^x}{n!}\,\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(x^n\,\mathrm{e}^{-x}\right)\,, \qquad n = 0, 1, 2, \ldots \tag{9.176}$$

Proposition 9.131 (Explicit form and generating function)**.** *The Rodrigues formula Equation (9.176) evaluates to*

$$L_n(x) = \sum_{j=0}^{n} \binom{n}{j}\,\frac{(-1)^j}{j!}\,x^j\,, \tag{9.177}$$

a polynomial of degree n with leading coefficient $(-1)^n/n!$ and $L_n(0) = 1$, and these polynomials are generated by

$$\frac{1}{1-t}\,\exp\left[-\frac{x\,t}{1-t}\right] = \sum_{n=0}^{\infty} L_n(x)\,t^n\,, \qquad |t|<1\,. \tag{9.178}$$

Rests on Definition 9.130.

Derivation. Derives Proposition 9.131.

By the Leibniz rule Equation (7.16),

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(x^n \mathrm{e}^{-x}\right) = \sum_{k=0}^{n} \binom{n}{k} \frac{n!}{(n-k)!}\, x^{n-k}\,(-1)^{n-k}\,\mathrm{e}^{-x}\,,$$

and dividing by $n!\,\mathrm{e}^{-x}$ and putting $j = n-k$ gives Equation (9.177); the term $j=n$ gives the leading coefficient and the term $j=0$ gives $L_n(0)=1$. For the generating function expand the exponential and then each factor $(1-t)^{-1-j}$ by the binomial series,

$$\frac{1}{1-t}\exp\left[-\frac{xt}{1-t}\right] = \sum_{j\ge 0}\frac{(-x)^j}{j!}\,\frac{t^j}{(1-t)^{j+1}} = \sum_{j\ge 0}\frac{(-x)^j}{j!}\,t^j \sum_{i\ge 0}\binom{i+j}{j} t^i\,,$$

both series being absolutely convergent for $|t|<1$, so that they may be rearranged (Proposition 7.49). The coefficient of t^n collects $i=n-j$ and is $\sum_{j\le n}\binom{n}{j}(-x)^j/j!$, which is Equation (9.177). □

Proposition 9.132 (The equation and the recurrence)**.** *L_n solves Equation (9.174), and for $n \ge 1$*

$$(n+1)\,L_{n+1}(x) = (2n+1-x)\,L_n(x) - n\,L_{n-1}(x)\,. \tag{9.179}$$

Rests on Proposition 9.131 and Equation (9.174).

Derivation. Derives Proposition 9.132.

Write $v = x^n\mathrm{e}^{-x}$, which satisfies $x\,\mathrm{d}v/\mathrm{d}x = (n-x)\,v$. Differentiating that identity $n+1$ times by the Leibniz rule Equation (7.16), each side producing only two terms because x and $n-x$ are linear, gives for $u = v^{(n)}$

$$x\,\frac{\mathrm{d}^2u}{\mathrm{d}x^2} + (n+1)\,\frac{\mathrm{d}u}{\mathrm{d}x} = (n-x)\,\frac{\mathrm{d}u}{\mathrm{d}x} - (n+1)\,u\,,$$

that is $x\,u'' + (1+x)\,u' + (n+1)\,u = 0$. Now $u = n!\,\mathrm{e}^{-x}L_n$ by Equation (9.176), so $u' = n!\,\mathrm{e}^{-x}\,(L_n' - L_n)$ and $u'' = n!\,\mathrm{e}^{-x}\,(L_n'' - 2L_n' + L_n)$; substituting and cancelling $n!\mathrm{e}^{-x}$,

$$x\,L_n'' + L_n'\,(-2x+1+x) + L_n\,(x-1-x+n+1) = x\,L_n'' + (1-x)L_n' + n\,L_n = 0\,,$$

which is Equation (9.174). For the recurrence, write G for the left of Equation (9.178); then $\ln G = -\ln(1-t) - xt/(1-t)$ gives $\partial G/\partial t = G\big[(1-t)^{-1} - x(1-t)^{-2}\big]$, that is $(1-t)^2\,\partial G/\partial t = (1-t-x)\,G$. Substituting the series and reading off the coefficient of t^n gives $(n+1)L_{n+1} - 2n\,L_n + (n-1)L_{n-1} = L_n - L_{n-1} - x\,L_n$, which rearranges to Equation (9.179). □

9.11.3 Orthogonality and Laguerre series

Theorem 9.133 (Orthonormality)**.** *The Laguerre polynomials are orthonormal on the half-line with weight e^{-x}:*

$$\int_0^{\infty} L_m(x)\,L_n(x)\,\mathrm{e}^{-x}\,\mathrm{d}x = \delta_{mn}\,. \tag{9.180}$$

Rests on Theorem 9.57, Proposition 9.131, and Equation (9.175).

Derivation. *Derives Theorem 9.133.*

For $m \neq n$ apply Theorem 9.57 with the data Equation (9.175): the eigenvalues m and n differ, and $p(x) = x\mathrm{e}^{-x}$ vanishes at the origin and decays at infinity while the solutions are polynomials, so Equation (9.57) holds.

For $m = n$ use Equation (9.176) on one factor and integrate by parts n times (Equation (7.28)). Every boundary term vanishes: at infinity because of e^{-x}, and at the origin because each derivative $\mathrm{d}^k(x^n\mathrm{e}^{-x})/\mathrm{d}x^k$ with $k < n$ still carries a positive power of x. Hence

$$\int_0^\infty L_n^2\,\mathrm{e}^{-x}\,\mathrm{d}x = \frac{1}{n!}\int_0^\infty L_n(x)\,\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(x^n\mathrm{e}^{-x}\right)\mathrm{d}x = \frac{(-1)^n}{n!}\int_0^\infty \frac{\mathrm{d}^n L_n}{\mathrm{d}x^n}\,x^n\,\mathrm{e}^{-x}\,\mathrm{d}x\,.$$

By Equation (9.177) the n-th derivative of L_n is the constant $(-1)^n$, and $\int_0^\infty x^n\mathrm{e}^{-x}\mathrm{d}x = \Gamma(n+1) = n!$, so the whole expression is $[(-1)^n]^2\,n!/n! = 1$. □

Proposition 9.134 (Laguerre series)**.** *Let f on $(0,\infty)$ admit an expansion $f(x) = \sum_{n\geq 0} a_n L_n(x)$ convergent in the mean with respect to the weight e^{-x}. Then $a_n = \int_0^\infty f(x)\,L_n(x)\,\mathrm{e}^{-x}\,\mathrm{d}x$.*

Rests on Theorem 9.133 and Definition 9.47.

Derivation. *Derives Proposition 9.134.*

This is Proposition 9.48 verbatim: the system $\{L_n\}$ is already orthonormal for the weight e^{-x} by Equation (9.180), so no normalising constant appears. Completeness carries the caveat of Remark 9.101. □

Remark 9.135 (The associated polynomials and the hydrogen atom). The radial eigenfunctions of the hydrogen atom involve not L_n but the *associated* Laguerre polynomials $L_n^k = (-1)^k\,\mathrm{d}^k L_{n+k}/\mathrm{d}x^k$, which stand to Definition 9.130 as the associated Legendre functions of Definition 9.116 stand to the Legendre polynomials. They solve $x\,y'' + (k+1-x)\,y' + n\,y = 0$, which is Equation (9.174) with $1-x$ replaced by $k+1-x$ — a Sturm–Liouville problem with $p(x) = x^{k+1}\mathrm{e}^{-x}$ and weight $x^k\mathrm{e}^{-x}$ on the same half-line, so that the boundary term dies at both ends exactly as it does here. Their orthogonality, recurrences and generating function are established by the arguments used above, and they are used in Chapter 81 [1593, 2313].

9.12 Elliptic integrals and the Jacobi elliptic functions

Every family treated so far arose from separating a linear partial differential equation. The family of this section arises differently and is needed just as often: whenever a one-degree-of-freedom conservative system is integrated by energy conservation, the result is a quadrature $t = \int \mathrm{d}u/\sqrt{f(u)}$, and when f is a cubic or a quartic that quadrature is an elliptic integral. The pendulum at finite amplitude, the nutation of a heavy symmetric top, the free rotation of a body with three unequal moments of inertia and the shape of an elastic strip under end thrust are all of this kind, and each stops at a quadrature the reader cannot invert without the functions defined here. They are due to Legendre, who reduced every such integral to three normal forms, and to Jacobi and Abel, who saw that the fruitful object is the *inverse* of the integral.

9.12.1 The three normal forms

Lemma 9.136 (Wallis integrals)**.** *For every integer $n \geq 0$,*

$$\int_0^{\pi/2} \sin^{2n}\theta\,\mathrm{d}\theta = \frac{\pi}{2}\,\frac{(2n)!}{4^n\,(n!)^2}\,. \tag{9.181}$$

Rests on Corollary 7.44 and Lemma 7.59.

Derivation. *Derives Lemma 9.136.*

Write W_n for the integral. Integrating by parts (Corollary 7.44) with $u = \sin^{2n-1}\theta$ and $\mathrm{d}v = \sin\theta\,\mathrm{d}\theta$, the boundary term $\left[-\sin^{2n-1}\theta\cos\theta\right]_0^{\pi/2}$ vanishes at both ends for $n \geq 1$, leaving

$$W_n = (2n-1)\int_0^{\pi/2} \sin^{2n-2}\theta\,\cos^2\theta\,\mathrm{d}\theta = (2n-1)(W_{n-1} - W_n)\,,$$

by $\cos^2 = 1 - \sin^2$ (Lemma 7.59). Hence $2n\,W_n = (2n-1)W_{n-1}$, and with $W_0 = \pi/2$ induction gives

$$W_n = \frac{\pi}{2}\prod_{j=1}^{n}\frac{2j-1}{2j} = \frac{\pi}{2}\,\frac{(2n)!}{4^n(n!)^2}\,,$$

the last step because $\prod_{j\leq n}(2j-1) = (2n)!/(2^n n!)$ and $\prod_{j\leq n}(2j) = 2^n n!$. □

Definition 9.137 (Elliptic integrals of the three kinds)**.** Let $0 \leq k < 1$ be the *modulus* and $\varphi \in \mathbb{R}$ the *amplitude*. The incomplete elliptic integrals of the first, second and third kind are

$$F(\varphi,k) = \int_0^{\varphi} \frac{\mathrm{d}\theta}{\sqrt{1-k^2\sin^2\theta}}\,, \tag{9.182}$$

$$E(\varphi,k) = \int_0^{\varphi} \sqrt{1-k^2\sin^2\theta}\,\mathrm{d}\theta\,, \tag{9.183}$$

$$\Pi(n;\varphi,k) = \int_0^{\varphi} \frac{\mathrm{d}\theta}{(1-n\sin^2\theta)\sqrt{1-k^2\sin^2\theta}}\,, \tag{9.184}$$

with n the *characteristic*. The *complete* integrals are their values at $\varphi = \pi/2$,

$$K(k) = F\!\left(\frac{\pi}{2},k\right), \qquad E(k) = E\!\left(\frac{\pi}{2},k\right). \tag{9.185}$$

Rests on Definition 7.93 and Theorem 7.40.

Remark 9.138 (Modulus, parameter, and a trap in the literature). Three conventions are in use and they are not distinguished by the symbol K. This treatise writes $K(k)$ with the *modulus* k, so that k^2 appears in Equation (9.182). Much of the mathematical literature, and most computer-algebra systems, write $K(m)$ with the *parameter* $m = k^2$; a third convention uses the *modular angle* α with $k = \sin\alpha$. The three disagree by a squaring, so a formula copied across conventions without checking is wrong by an amount that vanishes at small amplitude and is therefore not caught by the obvious limit. Where a result of this treatise is quoted elsewhere — the pendulum period Equation (9.199) is the case that matters — the argument is stated as $\sin(\theta_0/2)$, a modulus.

Proposition 9.139 (Series for the complete integral of the first kind)**.** *For $0 \leq k < 1$,*

$$K(k) = \frac{\pi}{2}\sum_{n=0}^{\infty}\left[\frac{(2n)!}{4^n(n!)^2}\right]^2 k^{2n} = \frac{\pi}{2}\left[1 + \frac{k^2}{4} + \frac{9k^4}{64} + \frac{25k^6}{256} + \cdots\right], \tag{9.186}$$

the series converging for every k with $|k| < 1$. Moreover K is strictly increasing on $[0,1)$ and $K(k) \to +\infty$ as $k \to 1^-$.

Rests on Definition 9.137 and Lemmas 9.6 and 9.136.

Derivation. *Derives Proposition 9.139.*

The binomial series $(1-u)^{-1/2} = \sum_{n\geq 0}\left[(2n)!/(4^n(n!)^2)\right]u^n$ converges for $|u| < 1$ (Theorem 7.50, Proposition 7.48: the ratio of successive coefficients is $(2n+1)/(2n+2) \to 1$). Put $u = k^2\sin^2\theta$, so that $0 \leq u \leq k^2 < 1$ uniformly in θ. The n-th term is bounded on $[0,\pi/2]$ by $M_n = \left[(2n)!/(4^n(n!)^2)\right]k^{2n}$, and $\sum_n M_n = (1-k^2)^{-1/2}$ is finite, so by Lemma 9.6 the series converges uniformly on $[0,\pi/2]$ and may be integrated term by term. Each

term contributes $[(2n)!/(4^n(n!)^2)]k^{2n}W_n$ with W_n the Wallis integral Equation (9.181), and substituting Equation (9.181) produces the square in Equation (9.186). The first coefficients are 1, $(1/2)^2 = 1/4$, $(3/8)^2 = 9/64$ and $(5/16)^2 = 25/256$.

Monotonicity is immediate: the integrand of Equation (9.182) increases with k at every $\theta \in (0, \pi/2]$. For the divergence, use $1 - k^2\sin^2\theta = (1-k^2) + k^2\cos^2\theta \le c^2 + \cos^2\theta$ with $c = \sqrt{1-k^2}$, and substitute $w = \pi/2 - \theta$, so $\cos\theta = \sin w \le w$:

$$K(k) \ge \int_0^{\pi/2} \frac{\mathrm{d}w}{\sqrt{c^2 + w^2}} = \ln\left[\frac{\pi}{2c} + \sqrt{1 + \frac{\pi^2}{4c^2}}\right],$$

which tends to $+\infty$ as $c \to 0^+$, that is as $k \to 1^-$. □

9.12.2 The Jacobi elliptic functions

Definition 9.140 (Amplitude and the Jacobi elliptic functions)**.** Fix $0 \le k < 1$. The integrand of Equation (9.182) is positive and π-periodic in θ, so $\varphi \mapsto F(\varphi, k)$ is a strictly increasing bijection of $\mathbb{R}$ onto $\mathbb{R}$ satisfying $F(\varphi + \pi, k) = F(\varphi, k) + 2K(k)$. Its inverse is the *amplitude* $\operatorname{am}(u, k)$, and the *Jacobi elliptic functions* are

$$\operatorname{sn}(u,k) = \sin\operatorname{am}(u,k), \quad \operatorname{cn}(u,k) = \cos\operatorname{am}(u,k), \quad \operatorname{dn}(u,k) = \sqrt{1 - k^2\operatorname{sn}^2(u,k)}. \tag{9.187}$$

Rests on Definition 9.137 and Proposition 7.32.

Proposition 9.141 (First properties)**.** *With $K = K(k)$ and the argument k suppressed,*

$$\operatorname{sn}^2 u + \operatorname{cn}^2 u = 1, \qquad \operatorname{dn}^2 u + k^2\operatorname{sn}^2 u = 1, \tag{9.188}$$

$$\frac{\mathrm{d}}{\mathrm{d}u}\operatorname{sn} u = \operatorname{cn} u\,\operatorname{dn} u, \qquad \frac{\mathrm{d}}{\mathrm{d}u}\operatorname{cn} u = -\operatorname{sn} u\,\operatorname{dn} u, \tag{9.189}$$

and $\mathrm{d}(\operatorname{dn} u)/\mathrm{d}u = -k^2\operatorname{sn} u\,\operatorname{cn} u$. *Furthermore* sn *is odd and* cn, dn *are even;* $\operatorname{sn} 0 = 0$, $\operatorname{cn} 0 = \operatorname{dn} 0 = 1$, $\operatorname{sn} K = 1$*; and*

$$\operatorname{sn}(u+2K) = -\operatorname{sn} u, \quad \operatorname{cn}(u+2K) = -\operatorname{cn} u, \quad \operatorname{dn}(u+2K) = \operatorname{dn} u, \tag{9.190}$$

so sn *and* cn *have period* $4K$ *and* dn *has period* $2K$*. At* $k = 0$ *they reduce to* $\sin u$, $\cos u$ *and* 1.

Rests on Definition 9.140, Proposition 7.32, and Lemma 7.59.

Derivation. *Derives Proposition 9.141.*

The first identity in Equation (9.188) is the Pythagorean identity (Lemma 7.59) applied to am, and the second is Equation (9.187) squared. Since $\partial F/\partial\varphi = (1 - k^2\sin^2\varphi)^{-1/2}$ is positive and continuous, Proposition 7.32 gives

$$\frac{\mathrm{d}}{\mathrm{d}u}\operatorname{am}(u,k) = \sqrt{1 - k^2\sin^2\operatorname{am}(u,k)} = \operatorname{dn}(u,k), \tag{9.191}$$

and the chain rule (Proposition 7.31) turns Equation (9.187) into Equation (9.189). Differentiating $\operatorname{dn}^2 = 1 - k^2\operatorname{sn}^2$ and dividing by $2\operatorname{dn}$ gives the third derivative, $-k^2\operatorname{sn}\operatorname{cn}$.

F is odd in φ because its integrand is even, so am is odd and the parities follow. The values at $u = 0$ are those at $\varphi = 0$, and $\operatorname{am}(K) = \pi/2$ by Equation (9.185), giving $\operatorname{sn} K = 1$. Finally $F(\varphi + \pi) = F(\varphi) + 2K$ inverts to $\operatorname{am}(u + 2K) = \operatorname{am}(u) + \pi$, and Equation (9.190) is what sin, cos and $\sin^2$ do under a shift by π. At $k = 0$, Equation (9.182) gives $F(\varphi, 0) = \varphi$ and hence $\operatorname{am}(u, 0) = u$. □

Proposition 9.142 (The differential equation of the sine amplitude)**.** *The function* $y(u) = \operatorname{sn}(u,k)$ *satisfies*

$$\left(\frac{\mathrm{d}y}{\mathrm{d}u}\right)^2 = (1-y^2)(1-k^2y^2)\,, \tag{9.192}$$

and therefore the second-order equation

$$\frac{\mathrm{d}^2y}{\mathrm{d}u^2} = -(1+k^2)\,y + 2k^2\,y^3\,, \tag{9.193}$$

with $y(0)=0$ *and* $y'(0)=1$*. Conversely Equation* (9.193) *with those initial data has* $\operatorname{sn}(\cdot,k)$ *as its unique solution.*

Rests on Proposition 9.141 and Theorem 9.8.

Derivation. *Derives Proposition 9.142.*

Squaring Equation (9.189) and eliminating cn and dn by Equation (9.188) gives Equation (9.192). Differentiating $\operatorname{sn}' = \operatorname{cn}\operatorname{dn}$ once more and using the three derivative formulas,

$$\operatorname{sn}'' = -\operatorname{sn}\operatorname{dn}^2 - k^2\operatorname{sn}\operatorname{cn}^2 = -\operatorname{sn}(1-k^2\operatorname{sn}^2) - k^2\operatorname{sn}(1-\operatorname{sn}^2)\,,$$

which is Equation (9.193). The initial values are those of Proposition 9.141, with $\operatorname{sn}'(0) = \operatorname{cn}0\,\operatorname{dn}0 = 1$. The right-hand side of Equation (9.193) is a polynomial and therefore locally Lipschitz, so Theorem 9.8 makes the solution unique. Note that Equation (9.192), whose right-hand side has a square root at the turning points, does *not* by itself determine the solution there; the second-order form does. □

9.12.3 Quadratures that reduce to elliptic integrals

Proposition 9.143 (A cubic quadrature between two turning points)**.** *Let* $f(v) = c\,(v-e_1)(v-e_2)(v-e_3)$ *with* $c>0$ *and* $e_1<e_2<e_3$*, so that* $f>0$ *on* (e_1,e_2)*. Put*

$$k^2 = \frac{e_2-e_1}{e_3-e_1} \in (0,1)\,, \qquad \omega = \tfrac{1}{2}\sqrt{c\,(e_3-e_1)}\,. \tag{9.194}$$

Then for $u \in [e_1,e_2]$*,*

$$\int_{e_1}^{u} \frac{\mathrm{d}v}{\sqrt{f(v)}} = \frac{1}{\omega}\,F(\varphi,k)\,, \qquad \sin^2\varphi = \frac{u-e_1}{e_2-e_1}\,, \tag{9.195}$$

so the integral is finite at both endpoints, its value over the whole interval is $K(k)/\omega$*, and the function* $t \mapsto u(t)$ *inverting it is*

$$u(t) = e_1 + (e_2-e_1)\,\operatorname{sn}^2(\omega t,k)\,, \tag{9.196}$$

which oscillates between e_1 *and* e_2 *with period* $2K(k)/\omega$ *and satisfies* $(\mathrm{d}u/\mathrm{d}t)^2 = f(u)$*.*

Rests on Definition 9.140, Proposition 9.141, and Corollary 7.44.

Derivation. *Derives Proposition 9.143.*

Substitute $v = e_1 + (e_2-e_1)\sin^2\theta$ (Corollary 7.44), which increases from e_1 to e_2 as θ runs over $[0,\pi/2]$. Then

$$v-e_1 = (e_2-e_1)\sin^2\theta\,, \qquad v-e_2 = -(e_2-e_1)\cos^2\theta\,,$$

and, using Equation (9.194), $v-e_3 = -(e_3-e_1)\,[1-k^2\sin^2\theta]$. Hence

$$f(v) = c\,(e_2-e_1)^2(e_3-e_1)\,\sin^2\theta\,\cos^2\theta\,[1-k^2\sin^2\theta]\,,$$

whose square root is $2\omega(e_2-e_1)\sin\theta\cos\theta\sqrt{1-k^2\sin^2\theta}$ on $[0,\pi/2]$, every factor being non-negative there. Since $\mathrm{d}v = 2(e_2-e_1)\sin\theta\cos\theta\,\mathrm{d}\theta$, everything cancels except

$$\frac{\mathrm{d}v}{\sqrt{f(v)}} = \frac{1}{\omega}\,\frac{\mathrm{d}\theta}{\sqrt{1-k^2\sin^2\theta}}\,,$$

and integrating from 0 to φ gives Equation (9.195), finite at both ends because the integrand of Equation (9.182) is bounded for $k<1$. Taking $\varphi=\pi/2$ gives $K(k)/\omega$.

Inverting Equation (9.195) means $\varphi=\mathrm{am}(\omega t,k)$, hence $\sin^2\varphi=\mathrm{sn}^2(\omega t,k)$, which is Equation (9.196). Differentiating it with Equation (9.189), $\mathrm{d}u/\mathrm{d}t = 2\omega(e_2-e_1)\,\mathrm{sn}\,\mathrm{cn}\,\mathrm{dn}$, whose square is $4\omega^2(e_2-e_1)^2\,\mathrm{sn}^2\,\mathrm{cn}^2\,\mathrm{dn}^2$; and the same substitutions that produced $f(v)$ above, now with $\sin\theta=\mathrm{sn}$, give $f(u)=c(e_2-e_1)^2(e_3-e_1)\,\mathrm{sn}^2\,\mathrm{cn}^2\,\mathrm{dn}^2$. The two agree precisely when $4\omega^2=c(e_3-e_1)$, which is Equation (9.194). Finally sn^2 has period $2K$ by Equation (9.190), so u has period $2K/\omega$. □

Proposition 9.144 (The pendulum equation at finite amplitude)**.** *Let $\omega_0>0$ and let θ solve*

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}(t)+\omega_0^2\,\sin\theta(t)=0 \tag{9.197}$$

with $\theta(0)=0$ and $\mathrm{d}\theta/\mathrm{d}t(0)=2\omega_0\sin(\theta_0/2)$ for some $\theta_0\in(0,\pi)$. Then the motion is periodic between $-\theta_0$ and $+\theta_0$, with

$$\sin\frac{\theta(t)}{2} = k\;\mathrm{sn}(\omega_0 t,k)\,,\qquad k=\sin\frac{\theta_0}{2}\,, \tag{9.198}$$

and its period is

$$T(\theta_0)=\frac{4}{\omega_0}\,K\!\left(\sin\frac{\theta_0}{2}\right)=T_0\,\frac{2}{\pi}\,K\!\left(\sin\frac{\theta_0}{2}\right),\qquad T_0=\frac{2\pi}{\omega_0}\,, \tag{9.199}$$

which for small amplitude expands as

$$T(\theta_0)=T_0\left[1+\frac{\theta_0^2}{16}+\frac{11\,\theta_0^4}{3072}+O\!\left(\theta_0^6\right)\right]. \tag{9.200}$$

Rests on Proposition 9.139, Definition 9.140, and Corollary 7.44.

Derivation. *Derives Proposition 9.144.*

Multiplying Equation (9.197) by $\mathrm{d}\theta/\mathrm{d}t$ and integrating once gives the first integral $\tfrac12\dot\theta^2-\omega_0^2\cos\theta=\text{const}$. The stated initial data fix the constant so that $\dot\theta$ vanishes at $\theta=\pm\theta_0$, and with $\cos\theta=1-2\sin^2(\theta/2)$,

$$\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2=4\omega_0^2\left[k^2-\sin^2\frac{\theta}{2}\right],\qquad k=\sin\frac{\theta_0}{2}\,. \tag{9.201}$$

The right-hand side is negative beyond $|\theta|=\theta_0$, so the motion is confined there and reverses at the two turning points. Equation (9.197) is unchanged under $t\mapsto -t$, and also under $\theta\mapsto-\theta$ because the sine is odd, so the four quarters of the motion are congruent and the period is four times the time from $\theta=0$ to $\theta=\theta_0$.

On that quarter $\dot\theta>0$. Substitute $\sin(\theta/2)=k\sin\varphi$, which carries $\theta:0\to\theta_0$ into $\varphi:0\to\pi/2$ and gives $\tfrac12\cos(\theta/2)\,\mathrm{d}\theta=k\cos\varphi\,\mathrm{d}\varphi$ with $\cos(\theta/2)=\sqrt{1-k^2\sin^2\varphi}$, while Equation (9.201) gives $\dot\theta=2\omega_0 k\cos\varphi$. Hence

$$\mathrm{d}t=\frac{\mathrm{d}\theta}{\dot\theta}=\frac{2k\cos\varphi\,\mathrm{d}\varphi}{\sqrt{1-k^2\sin^2\varphi}}\,\frac{1}{2\omega_0 k\cos\varphi}=\frac{1}{\omega_0}\,\frac{\mathrm{d}\varphi}{\sqrt{1-k^2\sin^2\varphi}}\,,$$

so $\omega_0 t = F(\varphi, k)$, that is $\varphi = \operatorname{am}(\omega_0 t, k)$, which is Equation (9.198); and $\varphi = \pi/2$ at the turning point gives $T/4 = K(k)/\omega_0$, which is Equation (9.199).

For Equation (9.200), expand $k = \sin(\theta_0/2)$ by Taylor's theorem (Theorem 7.38): $k^2 = \theta_0^2/4 - \theta_0^4/48 + O(\theta_0^6)$ and $k^4 = \theta_0^4/16 + O(\theta_0^6)$. Substituting into Equation (9.186) truncated at k^4,

$$\frac{2}{\pi}K = 1 + \frac{k^2}{4} + \frac{9k^4}{64} = 1 + \frac{\theta_0^2}{16} - \frac{\theta_0^4}{192} + \frac{9\theta_0^4}{1024} + O\left(\theta_0^6\right) ,$$

and $-1/192 + 9/1024 = (-16 + 27)/3072 = 11/3072$. □

Remark 9.145 (What this section is for). Proposition 9.143 is the general shape of the answer whenever a conservative system with one degree of freedom is reduced to a quadrature whose radicand is a cubic with three real roots and the motion runs between the lower two. That is the case of the heavy symmetric top, whose nutation angle satisfies exactly such an equation in the variable $\cos\theta$, so its nutation is Equation (9.196) and its nutation period is the complete integral $2K(k)/\omega$; the same reduction applies to the free rotation of a body with three distinct moments of inertia. The quartic case is reached from the cubic by the substitution that sends one root to infinity, and the case of two real and two complex roots by a different linear fractional substitution; those reductions are not carried out here, and no statement of this treatise depends on them. Proposition 9.144 is the second standard instance and is used directly by Chapter 28 and Chapter 34: Equation (9.199) is what a pendulum clock's rate depends on, Equation (9.200) is the correction a swinging amplitude introduces, and the two together are the quantitative content of the isochronism experiment. Note that Equation (9.200) is an expansion of an exact formula, not an independent approximation: the exact period is available at every amplitude below π, and the series is quoted only because it is what an error budget uses.

9.13 The Laplace transform

Beside the Fourier transform of Section 9.7.5, the source reserves a section for the Laplace transform — the integral transform adapted to initial-value problems on the half-line — but leaves it empty. The material is developed in full in Section 17.4, and this section records only what it is for and where each piece of it is proved.

The Fourier transform requires a function to decay at both ends of the line. A physical initial-value problem supplies neither condition: the system is switched on at $t = 0$ and its response may grow. Restricting the integral to $t > 0$ and inserting the convergence factor e^{-st} with $\Re s$ large enough repairs both defects at once, and gives the one-sided Laplace transform of Definition 17.65. Three facts make it a method rather than a definition. It converges on a half-plane $\Re s > \sigma_c$ and nowhere to the left of it (Theorem 17.66), and it is holomorphic there (Proposition 17.67); on each vertical line it is literally a Fourier transform, of the causal signal damped by $\mathrm{e}^{-\sigma t}$ (Proposition 17.68), which is what reduces its inversion to Theorem 17.32 and produces the Bromwich contour integral (Theorem 17.69), evaluable by residues (Corollary 17.71). And it converts differentiation into multiplication while swallowing the initial data, which enter the transformed equation as extra polynomial terms rather than as conditions imposed afterwards (Equation (17.87)).

That last rule is what makes the transform the natural tool for the constant-coefficient equations of Section 9.1: a linear ordinary differential equation becomes an algebraic one, its solution is a rational function of s, and the inverse transform is read off the poles by Heaviside's expansion theorem (Corollary 17.76). The location of those poles is the stability of the system (Proposition 17.77), and the requirement that the response be causal, applied to the same half-plane of holomorphy, yields the Kramers–Kronig dispersion relations (Theorem 17.78).

9.14 Summary of the classical families

Table 9.2 collects the Sturm–Liouville data (p, q, λ^2, r) of every family treated above, together with the pair of independent solutions. Three of the rows have to be read with Remark 9.97 in mind: the Bessel, Legendre and associated Legendre equations each admit two Sturm–Liouville readings. In the first the order or the degree sits in q — the source's identification, reproduced in the table — and in the second a scaling parameter or the degree carries the spectral parameter instead. It is the second reading to which the orthogonality relations of Theorems 9.98, 9.111 and 9.119 belong, since orthogonality of eigenfunctions needs distinct eigenvalues to separate them. The Legendre row is where the swap is bluntest, and it is the reason the table alone cannot be read as an eigenvalue problem: it carries $q = n(n+1)$ and $\lambda^2 = 0$, so Equation (9.55) has no spectral parameter left, whereas the proof of Theorem 9.111 reads the same equation with $q = 0$ and $\lambda^2 = n(n+1)$, distinct for distinct degrees.

Two rows of Table 9.2 are of a different character and are included because later parts of the book need them by name. The modified Bessel row (Definition 9.105) is the ordinary Bessel row with the sign of q reversed, which is exactly the difference between oscillation and monotone growth, and its solutions therefore carry no zeros and no eigenvalue condition (Proposition 9.106). The Mathieu row Equation (9.43) is the only entry whose coefficient q is periodic rather than polynomial or exponential: its eigenvalue problem is not a boundary-value problem on an interval but the stability question of Corollary 9.41, and its named solutions ce_n and se_n are the periodic solutions that exist only on the boundaries of the instability regions (Proposition 9.43). Finally, the elliptic integrals and Jacobi functions of Section 9.12 appear in no row at all, and deliberately: they do not solve a linear Sturm–Liouville equation but invert a quadrature, and Equation (9.193) — the equation sn does satisfy — is nonlinear.

Table 9.2: Sturm–Liouville data of the classical equations, in the convention of Definition 9.54: the operator is $\frac{\mathrm{d}}{\mathrm{d}x}\big[p\,y'\big] + \big[q + \lambda^2 r\big]y$.

Equation	$p(x)$	$q(x)$	λ^2	$r(x)$	Solutions
Helmholtz (1D)	1	0	λ^2	1	$\mathrm{e}^{\pm \mathrm{i}\lambda x}$
Bessel	x	x	$-\nu^2$	$1/x$	$J_\nu,\ N_\nu$
Modified Bessel	x	$-x$	$-\nu^2$	$1/x$	$I_\nu,\ K_\nu$
Legendre	$1-x^2$	$n(n+1)$	0	1	$P_n,\ Q_n$
Associated Legendre	$1-x^2$	$n(n+1)$	$-m^2$	$\dfrac{1}{1-x^2}$	$P_n^m,\ Q_n^m$
Hermite	e^{-x^2}	0	$2n$	e^{-x^2}	H_n
Laguerre	$x\,\mathrm{e}^{-x}$	0	n	e^{-x}	L_n
Mathieu	1	$-2h\cos 2x$	a	1	$\mathrm{ce}_n,\ \mathrm{se}_n$

Chapter 10

Partial Differential Equations

A *partial differential equation* (PDE) relates a function of several variables to its partial derivatives. The fundamental equations of physics are of this kind: the wave equation, the heat equation, the Laplace and Poisson equations, and the Schrödinger equation are all linear second-order PDEs, and their boundary- and initial-value problems occupy the rest of this treatise.

Three facts organise the whole subject and are established below. The first is that a second-order equation carries an invariant — the inertia of its principal coefficient matrix — which sorts the equations of physics into three families with entirely different behaviour, and which no change of coordinates can alter. The second is that each family admits its own kind of side condition and no other: asking an elliptic equation for initial data, or a parabolic one for a solution running backwards in time, is not merely difficult but ill posed in Hadamard's precise sense. The third is that a linear problem is solved once the response to a point source is known, which is what a Green's function is.

The general theory followed here — classification, characteristics, well-posedness and the canonical equations — is that of Courant–Hilbert [1222]; the operator-by-operator catalogue of separable coordinates and Green's functions is Morse–Feshbach [1023].

Throughout, Ω denotes a nonempty open subset of $\mathbb{R}^N$, repeated indices are summed from 1 to N, and $\partial_i = \partial/\partial x^i$. The Laplacian is written $\nabla^2 = \partial_i \partial_i$ in Cartesian coordinates, and the wave operator of speed c is

$$\Box = \frac{1}{c^2}\,\partial_t^2 - \nabla^2 \,. \tag{10.1}$$

10.1 Definitions

Definition 10.1 (Partial differential equation; order)**.** Let $u : \Omega \longrightarrow \mathbb{R}$ be the unknown. For a multi-index $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{N}^N$ write $|\alpha| = \alpha_1 + \cdots + \alpha_N$ and

$$\partial^\alpha = \partial_1^{\alpha_1} \cdots \partial_N^{\alpha_N} \,, \qquad \xi^\alpha = \xi_1^{\alpha_1} \cdots \xi_N^{\alpha_N} \,. \tag{10.2}$$

A *partial differential equation of order* m on Ω is a relation

$$F\big(x,\; u(x),\; \{\partial^\alpha u(x)\}_{1 \le |\alpha| \le m}\big) = 0 \,, \qquad x \in \Omega \,, \tag{10.3}$$

in which F depends nontrivially on at least one derivative of order exactly m. A *classical solution* is a function $u \in C^m(\Omega)$ satisfying Equation (10.3) at every point of Ω.

Rests on Definitions 6.2 and 7.66.

Definition 10.2 (Linear, semilinear, quasilinear)**.** The equation Equation (10.3) is

1. *linear* if it reads $\sum_{|\alpha| \le m} a_\alpha(x)\, \partial^\alpha u = f(x)$, and *homogeneous* if in addition $f \equiv 0$;

2. *semilinear* if the terms of order m are linear with coefficients depending on x alone, the remaining dependence being arbitrary;

3. *quasilinear* if it is linear in the derivatives of order m with coefficients that may depend on x, on u and on derivatives of order below m;

4. *fully nonlinear* otherwise.

Rests on Definition 10.1.

Linearity is what makes superposition available, and superposition is what every method of this chapter rests on: if u_1 and u_2 solve a homogeneous linear equation, so does $\lambda_1 u_1 + \lambda_2 u_2$, because a linear differential operator is a linear map on the vector space $C^m(\Omega)$ (Chapter 5). The last section of the chapter shows what breaks when it is unavailable.

Definition 10.3 (Principal part and principal symbol)**.** For a linear operator $L = \sum_{|\alpha|\le m} a_\alpha(x)\,\partial^\alpha$ the *principal part* is $\sum_{|\alpha|=m} a_\alpha \partial^\alpha$ and the *principal symbol* is the homogeneous polynomial

$$\sigma_L(x,\xi) = \sum_{|\alpha|=m} a_\alpha(x)\,\xi^\alpha\,, \qquad \xi \in \mathbb{R}^N\,. \tag{10.4}$$

Rests on Definition 10.2 and Equation (10.2).

The symbol is obtained by the substitution $\partial^\alpha \longmapsto \xi^\alpha$, which is exactly what the Fourier transform performs up to a factor i per derivative (Equation (17.38)): for constant coefficients, $\widehat{Lu}(\xi) = \sigma_L(\mathrm{i}\xi)\,\hat{u}(\xi)$ modulo the lower-order terms. Everything in this section is therefore a statement about a polynomial.

Definition 10.4 (The second-order operator)**.** The general linear second-order operator on Ω is

$$Lu = a^{ij}(x)\,\partial_i\partial_j u + b^i(x)\,\partial_i u + c(x)\,u\,, \tag{10.5}$$

with real coefficients and $a^{ij} = a^{ji}$. Its principal symbol is the quadratic form $\sigma_L(x,\xi) = a^{ij}(x)\,\xi_i\xi_j$, and $A(x)$ denotes the symmetric matrix $\big(a^{ij}(x)\big)$.

Rests on Definition 10.3 and Proposition 7.73.

Remark 10.5. The symmetry $a^{ij} = a^{ji}$ costs nothing. For $u \in C^2$ one has $\partial_i\partial_j u = \partial_j\partial_i u$ (Proposition 7.73), so only the symmetric part $\frac{1}{2}(a^{ij} + a^{ji})$ of the coefficient array acts on the solution; the antisymmetric part is invisible and is discarded. This is why the classification below is a statement about a *symmetric* matrix, and hence about its inertia.

10.1.1 Classification and characteristics

Definition 10.6 (Type of a second-order operator)**.** Let (n_+, n_-, n_0) be the inertia of $A(x)$ — the numbers of positive, negative and zero eigenvalues, so that $n_+ + n_- + n_0 = N$. At the point x the operator L of Equation (10.5) is

1. *elliptic* if $n_0 = 0$ and $n_+ n_- = 0$, that is if $A(x)$ is definite;

2. *hyperbolic* if $n_0 = 0$ and $\min(n_+, n_-) = 1$;

3. *ultrahyperbolic* if $n_0 = 0$ and $\min(n_+, n_-) \ge 2$;

4. *parabolic* if $n_0 \ge 1$, that is if $A(x)$ is singular.

L is elliptic (hyperbolic, …) *on* Ω if it is so at every point of Ω.

Rests on Definition 10.4 and Theorem 5.112.

Remark 10.7. Case (iv) as stated is only degeneracy of the principal part; the heat operator $\partial_t - D\nabla^2$ is the case that matters, in which $n_0 = 1$ exactly and the first-order part supplies a derivative in the one direction the principal part cannot see. An equation with $n_0 \geq 1$ and no such first-order term — $\partial_x^2 u = 0$ in two variables — is degenerate rather than parabolic. Ultrahyperbolic operators do not occur in this treatise: they need at least two time-like directions, and the observed world has one.

The classification would be worthless if it depended on the coordinates used to write the equation down. It does not, and the reason is Sylvester's law of inertia.

Theorem 10.8 (The type is a coordinate invariant)**.** *Let $y = y(x)$ be a C^2 change of coordinates on Ω with invertible Jacobian $J^a{}_i = \partial y^a / \partial x^i$, and let $\tilde{A}$ be the principal coefficient matrix of the transformed operator. Then*

$$\tilde{A} = J\, A\, (J)^{\mathsf{T}}, \tag{10.6}$$

and the inertia (n_+, n_-, n_0) of A is unchanged. The type of a second-order operator at a point is therefore an attribute of the equation, not of the chart in which it is written.

Rests on Definition 10.6, Theorem 5.112, and Proposition 7.31.

Proof. *Derives Theorem 10.8.*

Write $u(x) = \tilde{u}(y(x))$ and let $\tilde{\partial}_a = \partial/\partial y^a$. The chain rule (Proposition 7.31) gives $\partial_i u = J^a{}_i\, \tilde{\partial}_a \tilde{u}$ and, differentiating once more,

$$\partial_i \partial_j u = J^a{}_i J^b{}_j\, \tilde{\partial}_a \tilde{\partial}_b \tilde{u} + (\partial_i \partial_j y^a)\, \tilde{\partial}_a \tilde{u}\,. \tag{10.7}$$

The second term carries only one derivative of $\tilde{u}$, so it joins the first-order part and cannot affect the principal coefficients. Contracting the first term with a^{ij} gives $\tilde{a}^{ab} = J^a{}_i\, a^{ij}\, J^b{}_j$, which is Equation (10.6). Since J is invertible, this is a congruence of symmetric matrices, and by Sylvester's law of inertia (Theorem 5.112) congruent real symmetric matrices have the same triple (n_+, n_-, n_0); Equation (5.142) is the normal form both reach. □

Example 10.9. In $\mathbb{R}^4$ with coordinates (ct, x^1, x^2, x^3) the wave operator Equation (10.1) has $A = \operatorname{diag}(1, -1, -1, -1)$, inertia $(1, 3, 0)$: hyperbolic. The Laplacian ∇^2 on $\mathbb{R}^3$ has $A = \operatorname{diag}(1, 1, 1)$, inertia $(3, 0, 0)$: elliptic. The heat operator $\partial_t - D\nabla^2$ on $\mathbb{R} \times \mathbb{R}^3$ has $A = \operatorname{diag}(0, -D, -D, -D)$, inertia $(0, 3, 1)$: parabolic, with the single null direction the time direction, exactly as in Remark 10.7.

Rests on Definition 10.6 and Equation (10.1).

Definition 10.10 (Characteristic surface)**.** Let $\phi \in C^1(\Omega)$ with $\nabla\phi \neq 0$, and let $S = \{x \in \Omega \mid \phi(x) = 0\}$. The hypersurface S is *characteristic* for L at $x \in S$ if and only if

$$\sigma_L\big(x, \nabla\phi(x)\big) = a^{ij}(x)\, \partial_i \phi(x)\, \partial_j \phi(x) = 0\,. \tag{10.8}$$

Rests on Definitions 10.3 and 10.4.

Proposition 10.11 (What a characteristic surface is)**.** *Let S be a C^2 hypersurface and let Cauchy data — the values of u and of its normal derivative on S — be prescribed. Every second-order derivative of u on S is then determined by the data together with the equation $Lu = f$ if and only if S is non-characteristic at the point in question.*

Rests on Definition 10.10 and Theorem 10.8.

Proof. *Derives Proposition 10.11.*

Choose coordinates $y = (y^1, \ldots, y^N)$ near the point with $y^N = \phi$, so that S is the slice $\{y^N = 0\}$; this is admissible because $\nabla\phi \neq 0$. The Cauchy data are $\tilde{u}$ and $\tilde{\partial}_N \tilde{u}$ on the slice. Differentiating those two functions *along* the slice determines every derivative $\tilde{\partial}_a \tilde{\partial}_b \tilde{u}$ with

$a, b < N$ and every $\tilde{\partial}_a \tilde{\partial}_N \tilde{u}$ with $a < N$. The only second derivative not so obtained is $\tilde{\partial}_N^2 \tilde{u}$, and the equation reads

$$\tilde{a}^{NN}\, \tilde{\partial}_N^2 \tilde{u} = f - (\text{determined terms})\,, \tag{10.9}$$

because every other principal term carries at least one index below N. By Equation (10.6), $\tilde{a}^{NN} = a^{ij} \partial_i \phi\, \partial_j \phi = \sigma_L(x, \nabla\phi)$. Hence Equation (10.9) determines $\tilde{\partial}_N^2 \tilde{u}$ if and only if that symbol is nonzero, which by Equation (10.8) is exactly non-characteristic. □

Corollary 10.12 (Characteristics of the three canonical operators)**.** *An elliptic operator has no real characteristic surface. The characteristics of* $\Box$ *are the surfaces with* $|\nabla\phi|^2 = c^{-2}(\partial_t \phi)^2$*, that is the light cones* $|\boldsymbol{x} - \boldsymbol{x}_0| = c\,|t - t_0|$*. The characteristics of* $\partial_t - D\nabla^2$ *are the slices* $t =$ *constant.*

Rests on Proposition 10.11 and Equation (10.1).

Proof. *Derives Corollary 10.12.*

Ellipticity means A definite, so $a^{ij}\xi_i\xi_j = 0$ forces $\xi = 0$, which $\nabla\phi \neq 0$ excludes. For $\Box$ the symbol is $c^{-2}(\partial_t\phi)^2 - |\nabla\phi|^2$, and the stated cones are the level sets of $\phi = c(t - t_0) - |\boldsymbol{x} - \boldsymbol{x}_0|$, which annihilate it. For the heat operator the principal symbol is $-D\,|\boldsymbol{\xi}|^2$, with no dependence on the time component of the conormal; it vanishes precisely when $\nabla_{\boldsymbol{x}}\phi = 0$, i.e. when ϕ is a function of t alone. □

That the slices of constant time are characteristic for the heat operator is not a technicality: it is the reason a parabolic problem takes one initial condition and not two, and it will reappear as the one-sided character of diffusion in Section 10.5.2.

Proposition 10.13 (Canonical form in two variables)**.** *Let* $Lu = a\,u_{xx} + 2b\,u_{xy} + c\,u_{yy} + \cdots$ *have* C^1 *coefficients near a point where* $(a, b, c) \neq (0, 0, 0)$*, and set* $\Delta = b^2 - ac$*. There are new coordinates* (ξ, η) *in which the principal part becomes*

$$u_{\xi\eta} \quad (\Delta > 0)\,, \qquad u_{\eta\eta} \quad (\Delta = 0)\,, \qquad u_{\xi\xi} + u_{\eta\eta} \quad (\Delta < 0)\,, \tag{10.10}$$

up to a nonvanishing factor. The three cases are respectively hyperbolic, parabolic and elliptic in the sense of Definition 10.6.

Rests on Theorem 10.8 and Definition 10.10.

Proof. *Derives Proposition 10.13.*

Here the principal coefficient matrix is

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \qquad \det A = ac - b^2 = -\Delta\,, \tag{10.11}$$

so the sign of Δ is the sign pattern of the eigenvalues of A, which is Definition 10.6 for $N = 2$: two eigenvalues of opposite sign when $\Delta > 0$, one zero eigenvalue when $\Delta = 0$, and two of the same sign when $\Delta < 0$. Under $\xi = \xi(x, y)$, $\eta = \eta(x, y)$ with Jacobian $J_0 = \xi_x\eta_y - \xi_y\eta_x \neq 0$, Equation (10.6) gives

$$\tilde{a} = a\xi_x^2 + 2b\xi_x\xi_y + c\xi_y^2\,, \qquad \tilde{c} = a\eta_x^2 + 2b\eta_x\eta_y + c\eta_y^2\,, \tag{10.12}$$

with $\tilde{b}$ the associated polarisation, and taking determinants of Equation (10.6) yields $\tilde{b}^2 - \tilde{a}\tilde{c} = J_0^2\,\Delta$.

Now $\tilde{a}$ vanishes exactly when $\xi =$ constant is a characteristic, by Equation (10.8). Suppose first that $a \neq 0$. On such a curve $\xi_y \neq 0$, since $\xi_y = 0$ would reduce Equation (10.12) to $a\xi_x^2 = 0$ and force $\nabla\xi = 0$, which Definition 10.10 excludes; dividing Equation (10.12)

by ξ_y^2 and writing $m = \mathrm{d}y/\mathrm{d}x = -\xi_x/\xi_y$ for the slope of such a curve, the condition is the quadratic

$$a\,m^2 - 2b\,m + c = 0\,, \qquad m = \frac{b \pm \sqrt{\Delta}}{a} \quad (a \neq 0)\,. \tag{10.13}$$

If $a = 0$ the closed form is empty and the families must be read off Equation (10.12) directly, which then factorises as $\tilde{a} = \xi_y(2b\,\xi_x + c\,\xi_y)$: the vertical family $x =$ constant, of infinite slope and so invisible to m, is always characteristic, and if $b \neq 0$ a second family of slope $m = c/(2b)$ accompanies it — two families, as $\Delta = b^2 > 0$ demands. If $a = b = 0$ then $\Delta = 0$, only the vertical family survives, and the principal part is already $c\,u_{yy}$ with $c \neq 0$, which is the parabolic normal form. The archetype of the first of these degenerate cases is $Lu = 2u_{xy}$, whose two characteristic families are the coordinate lines themselves and which is already in hyperbolic canonical form. With the vertical family admitted, the count of real families in each of the three cases is the one the sign of Δ dictates, and the constructions below apply verbatim.

If $\Delta > 0$ there are two real families; take ξ and η to be first integrals of the two, so $\tilde{a} = \tilde{c} = 0$ and $\tilde{b}^2 = J_0^2\Delta > 0$, leaving $2\tilde{b}\,u_{\xi\eta}$ as the principal part. If $\Delta = 0$ the two families coincide; take ξ to be a first integral of it and η any independent function, so $\tilde{a} = 0$ and then $\tilde{b}^2 = \tilde{a}\tilde{c} = 0$, leaving $\tilde{c}\,u_{\eta\eta}$ with $\tilde{c} \neq 0$: otherwise the whole transformed principal matrix would vanish, and by Equation (10.6) with J invertible so would A, contradicting $(a, b, c) \neq (0, 0, 0)$. If $\Delta < 0$ then $ac > b^2 \geq 0$, so $a \neq 0$ and the roots of Equation (10.13) are a complex-conjugate pair. Assume for this branch that the coefficients are real analytic. Complexifying y and integrating $\mathrm{d}y/\mathrm{d}x = (b + \mathrm{i}\sqrt{-\Delta})/a$ then supplies a complex first integral ζ with $a\zeta_x^2 + 2b\zeta_x\zeta_y + c\zeta_y^2 = 0$ and with $\Re\zeta$, $\Im\zeta$ functionally independent; setting $\xi = \Re\zeta$, $\eta = \Im\zeta$ turns that one complex condition into the real pair $\tilde{a} = \tilde{c}$, $\tilde{b} = 0$, which is $\tilde{a}\,(u_{\xi\xi} + u_{\eta\eta})$. Remark 10.14 records how the same branch is settled at C^1. □

Remark 10.14 (Regularity in the elliptic branch). The hyperbolic and parabolic branches of the proof are complete at C^1: their characteristic equations are real, and Theorem 9.8 integrates them under exactly that hypothesis. The elliptic branch is not. A complex first integral is not delivered by any real ordinary differential equation theorem, because the object sought is a function ζ of two real variables solving a Beltrami equation with non-vanishing Jacobian; complexifying the independent variable and integrating the characteristic equation constructs it when the coefficients are real analytic, and that is the hypothesis under which the classical argument of Courant and Hilbert is stated [1222]. For merely C^1 — hence locally Hölder — coefficients the existence of ζ is a separate and much harder statement, the Korn–Lichtenstein theorem on the local solvability of the Beltrami equation. The conclusion of Proposition 10.13 does hold at C^1, and the appendix supplies it: the reduction of the canonical form to a Beltrami equation, and the construction of its solution, are carried out there from a single quoted singular-integral estimate. Two things should be carried away from that section. The map (ξ, η) it produces is $C^{1,\alpha}$ and not C^2, so at merely Hölder coefficients the canonical form is a statement about the *principal part* alone — which is all Equation (10.6) ever needed, since that congruence involves the Jacobian and no second derivative; and one more derivative on the coefficients buys a $C^{2,\alpha}$ map and the classical reading of Equation (10.10), first-order remainder included.

Full derivation: Appendix A.13 (p. 2817).
Derives Proposition 10.13.

Proposition 10.15 (Method of characteristics, first order)**.** *Let u be a C^1 solution of the quasilinear equation*

$$a(x, y, u)\,\partial_x u + b(x, y, u)\,\partial_y u = c(x, y, u)\,. \tag{10.14}$$

Then along every solution of the characteristic system $\dot{x} = a$, $\dot{y} = b$, $\dot{u} = c$ the graph of u is swept out; the equation is thus reduced to a system of ordinary differential equations.

Rests on Definition 10.2 and Proposition 7.31.

Proof. *Derives Proposition 10.15.*

Let $s \longmapsto (x(s), y(s))$ solve $\dot{x} = a(x, y, u(x, y))$, $\dot{y} = b(x, y, u(x, y))$ and put $U(s) = u(x(s), y(s))$. The chain rule (Proposition 7.31) gives $\dot{U} = \partial_x u\,\dot{x} + \partial_y u\,\dot{y} = a\,\partial_x u + b\,\partial_y u = c$, which is the third equation. The converse is the statement that the surface swept by the characteristic curves through a non-characteristic initial curve is a solution surface, since its tangent plane contains the vector (a, b, c) at every point. □

The equation of a mechanical principal function is not of the quasilinear form Equation (10.14): the Hamiltonian is quadratic in the derivatives. What covers that case is Cauchy's method, in which the derivatives are carried along the curve as unknowns of their own, so that a curve in $\mathbb{R}^n$ is replaced by a *strip* — a point, the value there, and the tangent plane there.

Proposition 10.16 (Cauchy's characteristic strips)**.** *Let $F(\boldsymbol{x}, u, \boldsymbol{p})$ be of class C^2 on an open subset of $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$ and let u be a C^2 solution of the fully nonlinear first-order equation*

$$F(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x})) = 0\,. \tag{10.15}$$

Fix $\boldsymbol{x}_0$ and let $s \longmapsto (\boldsymbol{x}(s), U(s), \boldsymbol{p}(s))$ solve the characteristic strip system

$$\dot{x}^i = \partial_{p_i} F\,, \qquad \dot{U} = p_j\,\partial_{p_j} F\,, \qquad \dot{p}_i = -\partial_{x^i} F - p_i\,\partial_u F\,, \tag{10.16}$$

with data $\boldsymbol{x}(0) = \boldsymbol{x}_0$, $U(0) = u(\boldsymbol{x}_0)$ and $\boldsymbol{p}(0) = \nabla u(\boldsymbol{x}_0)$. Then $U(s) = u(\boldsymbol{x}(s))$ and $\boldsymbol{p}(s) = \nabla u(\boldsymbol{x}(s))$ for as long as the solution exists, and F keeps the value zero along the strip. The system Equation (10.16) is closed*: its right-hand sides involve only $(\boldsymbol{x}, U, \boldsymbol{p})$, so it can be integrated without knowing u beforehand, and that is what makes it a method of solution rather than an identity satisfied by a solution already found.*

Rests on Propositions 7.73 and 10.15 and Theorem 9.8.

Proof. *Derives Proposition 10.16.*

Let $s \longmapsto \boldsymbol{x}(s)$ solve $\dot{x}^i = \partial_{p_i} F(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x}))$ with $\boldsymbol{x}(0) = \boldsymbol{x}_0$; the right-hand side is C^1 in $\boldsymbol{x}$, so Theorem 9.8 supplies a unique solution on some interval about $s = 0$. Put $U(s) = u(\boldsymbol{x}(s))$ and $p_i(s) = \partial_i u(\boldsymbol{x}(s))$. The chain rule (Proposition 7.72) gives $\dot{U} = \partial_j u\,\dot{x}^j = p_j\,\partial_{p_j} F$, which is the second equation, and $\dot{p}_i = \partial_j \partial_i u\,\dot{x}^j = \partial_i \partial_j u\,\partial_{p_j} F$, the order of the two derivatives being immaterial by Proposition 7.73. Differentiating Equation (10.15) with respect to x^i,

$$\partial_{x^i} F + \partial_u F\,\partial_i u + \partial_{p_j} F\,\partial_i \partial_j u = 0\,, \tag{10.17}$$

so that $\partial_i \partial_j u\,\partial_{p_j} F = -\partial_{x^i} F - p_i\,\partial_u F$, which is the third equation. The triple $(\boldsymbol{x}, U, \boldsymbol{p})$ therefore solves the closed system Equation (10.16) with the stated data; that system has a C^1 right-hand side because F is C^2, so Theorem 9.8 applied to it says the triple is the only solution with those data, which is the claim. Finally, along any solution of Equation (10.16),

$$\frac{\mathrm{d}}{\mathrm{d}s} F = \partial_{x^i} F\,\dot{x}^i + \partial_u F\,\dot{U} + \partial_{p_i} F\,\dot{p}_i = \partial_{x^i} F\,\partial_{p_i} F + \partial_u F\,p_j\,\partial_{p_j} F - \partial_{p_i} F\,(\partial_{x^i} F + p_i\,\partial_u F) = 0\,, \tag{10.18}$$

so F retains its initial value, which is zero. □

Remark 10.17 (Two special cases). For $F = a^i(\boldsymbol{x}, u)\,p_i - c(\boldsymbol{x}, u)$ the first two equations of Equation (10.16) read $\dot{x}^i = a^i$ and $\dot{U} = c$, and the third is not needed: that is Proposition 10.15 again, in n variables. For $F = p_0 + H(\boldsymbol{x}, \boldsymbol{p})$ — the shape the Hamilton–Jacobi

equation of Chapter 23 takes when the time is written as one further coordinate x^0, so that $p_0 = \partial S/\partial t$ — one has $\partial_u F = 0$ and $\partial_{p_0} F = 1$, so s may be identified with t and the remaining equations of Equation (10.16) are $\dot{x}^i = \partial H/\partial p_i$ and $\dot{p}_i = -\partial H/\partial x^i$: Hamilton's equations themselves. The characteristics of the Hamilton–Jacobi equation are the trajectories of the mechanical system, and this is the calculation that establishes it.

A complete integral is a family of solutions large enough to generate the rest by a purely algebraic operation. The operation is the formation of an envelope, and the one thing that has to be proved about it is that it does not leave the equation.

Definition 10.18 (Complete integral; envelope)**.** Let $u = \varphi(\boldsymbol{x}; \boldsymbol{a})$ be a family of C^2 solutions of Equation (10.15), one for each $\boldsymbol{a}$ in an open set $A \subseteq \mathbb{R}^n$, with φ of class C^2 in both arguments jointly. The family is a *complete integral* when the $n \times (n+1)$ matrix whose rows are $\left(\partial\varphi/\partial a_k,\ \partial^2\varphi/\partial a_k \partial x^1, \dots, \partial^2\varphi/\partial a_k \partial x^n\right)$ has rank n everywhere, so that distinct parameters give genuinely distinct solutions. Given a C^1 map $\boldsymbol{x} \longmapsto \boldsymbol{a}(\boldsymbol{x})$ into A with

$$\frac{\partial \varphi}{\partial a_k}\big(\boldsymbol{x}; \boldsymbol{a}(\boldsymbol{x})\big) = 0\,, \qquad k = 1, \dots, n\,, \tag{10.19}$$

the function $v(\boldsymbol{x}) = \varphi(\boldsymbol{x}; \boldsymbol{a}(\boldsymbol{x}))$ is an *envelope* of the family: at each point it touches one member of the family and has the same tangent plane there.

Rests on Equation (10.15) *and Proposition 7.72.*

Proposition 10.19 (An envelope of solutions is a solution)**.** *Every envelope in the sense of Definition 10.18 solves Equation* (10.15)*.*

Rests on Definition 10.18 and Proposition 7.72.

Proof. *Derives Proposition 10.19.*

By the chain rule (Proposition 7.72),

$$\partial_i v(\boldsymbol{x}) = \frac{\partial \varphi}{\partial x^i} + \frac{\partial \varphi}{\partial a_k}\,\frac{\partial a_k}{\partial x^i} = \frac{\partial \varphi}{\partial x^i}\big(\boldsymbol{x}; \boldsymbol{a}(\boldsymbol{x})\big)\,, \tag{10.20}$$

the second group of terms vanishing by Equation (10.19). Hence the triple $(\boldsymbol{x}, v(\boldsymbol{x}), \nabla v(\boldsymbol{x}))$ is exactly the triple $(\boldsymbol{x}, \varphi, \nabla_{\boldsymbol{x}}\varphi)$ formed from the single member $\varphi(\cdot\,; \boldsymbol{a}(\boldsymbol{x}))$ of the family, and F vanishes on it because that member is a solution. □

Remark 10.20 (What the envelope construction is for)*.* Proposition 10.19 is what makes a complete integral worth having. Restricting the parameters by $n-1$ arbitrary relations and enveloping over what is left produces, from the n-parameter family, a solution containing an arbitrary function — the *general integral* of the equation; enveloping over all n parameters at once produces the *singular* solution, which belongs to no member of the family and is where the family's members touch one another. In Chapter 23 this is the sense in which a caustic is the envelope of a family of solutions, and Equation (10.19) is the condition that chapter writes as the vanishing of a Jacobian. The construction and its use for first-order equations are Cauchy's; the classical account, with the strip system Equation (10.16) and the envelope theorem side by side, is [1222], chapter II.

Remark 10.21*.* Proposition 10.16 is the mechanism behind the Hamilton–Jacobi equation of Chapter 23: there the first-order PDE for the principal function is integrated along the characteristic strips, and those strips are, by Remark 10.17, the trajectories of the mechanical system carried together with their momenta. Its quasilinear predecessor Proposition 10.15 does not cover that equation, and the distinction matters: a Hamiltonian is quadratic in $\partial S/\partial q^a$, so the derivatives do not enter linearly. The quasilinear construction reappears below as the mechanism of shock formation (Section 10.6.3), where the characteristics cross.

Theorem 10.22 (Cauchy–Kovalevskaya)**.** *If the coefficients and the data of a normal Cauchy problem are real analytic near a point, and the initial surface is non-characteristic there, then the problem has a solution analytic in a neighbourhood of that point, and it is unique among analytic solutions.*

Rests on Proposition 10.11 and Definition 10.2.

Full derivation: Appendix A.14 (p. 2825).

Derives Theorem 10.22.

Remark 10.23. The theorem is due to Kovalevskaya [190], and its hypotheses cannot be relaxed casually. Analyticity is essential: Lewy exhibited a linear equation of first order with polynomial coefficients and a smooth right-hand side possessing no solution whatever, in any neighbourhood of any point [1122]. Nor does the theorem say anything useful about physics even when it applies, because analytic data cannot be prescribed independently in disjoint regions: the Cauchy problem it solves for the Laplace equation is the ill-posed one of Section 10.1.2. Existence and well-posedness are different questions, and only the second is physical.

10.1.2 Well-posedness in the sense of Hadamard

Definition 10.24 (Well-posed problem)**.** A problem consisting of a differential equation together with side conditions, with data d in a normed space $\mathcal{D}$ and solutions u in a normed space $\mathcal{U}$, is *well posed in the sense of Hadamard* if and only if

1. a solution exists for every admissible d;
2. it is unique;
3. it depends continuously on the data: $d_n \to d$ in $\mathcal{D}$ implies $u_n \to u$ in $\mathcal{U}$.

A problem failing any one of the three is *ill posed*.

Rests on Definitions 6.24 and 10.1.

Hadamard introduced the three conditions together, and insisted on the third [305]; the systematic development is in his Yale lectures [502]. The third condition is the one that carries the physics. Data are measured, hence known only to within an experimental uncertainty, and a problem that fails (iii) converts an arbitrarily small uncertainty into an arbitrarily large one: its solution is not a prediction. The following is the standard example, and it is the heat equation run backwards.

Proposition 10.25 (The backward heat problem is ill posed)**.** *Consider $\partial_t u = D\,\partial_x^2 u$ on $(0,\pi)\times(0,T)$ with $u(0,t) = u(\pi,t) = 0$, and let the datum be the final profile $u(\cdot,T) = g$, the unknown being the initial profile $u(\cdot,0)$. With the supremum norm on both, this problem violates condition (iii) of Definition 10.24.*

Rests on Definitions 10.6 and 10.24.

Proof. *Derives Proposition 10.25.*

For $n \in \mathbb{N}$ put

$$u_n(x,t) = \frac{1}{n}\,\mathrm{e}^{Dn^2(T-t)}\,\sin(nx)\,. \tag{10.21}$$

Then $\partial_t u_n = -Dn^2 u_n$ and $D\partial_x^2 u_n = -Dn^2 u_n$, so each u_n solves the equation and vanishes at $x = 0, \pi$. Its datum is $g_n = u_n(\cdot,T) = n^{-1}\sin(nx)$, with $\|g_n\|_\infty = 1/n \longrightarrow 0$, so g_n tends to the zero datum, whose solution is $u \equiv 0$. But $\|u_n(\cdot,0)\|_\infty = n^{-1}\mathrm{e}^{Dn^2T} \longrightarrow \infty$. Data differing by ε therefore admit initial profiles differing by any prescribed amount, and no continuity estimate can hold. □

Remark 10.26 (Hadamard's example). The same phenomenon forbids Cauchy data for an elliptic equation. Take $\partial_x^2 u + \partial_y^2 u = 0$ on the half-plane $y > 0$ with $u(x,0) = 0$ and $\partial_y u(x,0) = n^{-1}\sin(nx)$. The solution is $u = n^{-2}\sin(nx)\sinh(ny)$, and while the Cauchy data tend uniformly to zero, $\sup_x |u(x,y)| = n^{-2}\sinh(ny)$ diverges for every fixed $y > 0$. An elliptic equation accepts data on a closed boundary — Dirichlet or Neumann — and nothing else; a hyperbolic equation accepts Cauchy data on a non-characteristic surface; a parabolic equation accepts one initial condition and boundary conditions, forward in time only. The match between type and side condition is not a convention but a theorem about which problems are well posed.

Remark 10.27 (Where ill-posedness is unavoidable). Ill-posed problems cannot simply be avoided, because some of them are the questions one wants to ask: reconstructing a source from its field, deconvolving an instrumental response from a measured spectrum, or inferring a past state from a present one. The Fourier analysis of the same difficulty is Remark 17.55: the forward map annihilates high frequencies, so its inverse amplifies them without bound, and the practical answer is regularisation — solving a nearby well-posed problem and quantifying the difference.

10.1.3 The equations of physics

The equations this chapter serves are five, and between them they cover almost every field equation of the treatise.

The *wave equation*

$$\Box u = \frac{1}{c^2}\,\partial_t^2 u - \nabla^2 u = 0\,, \tag{10.22}$$

with c a speed in m/s, is hyperbolic (Example 10.9). It was written down by d'Alembert for the vibrating string [25] and returns for sound (Chapter 28), for light (Chapter 61) and for gravitational radiation (Chapter 46).

The *heat* or *diffusion equation*

$$\partial_t u - D\,\nabla^2 u = 0\,, \tag{10.23}$$

with diffusivity D in $\mathrm{m^2/s}$, is parabolic. It is Fourier's [72], and it governs conduction, molecular diffusion (Chapter 119) and the transport coefficients of Chapter 124.

The *Laplace* and *Poisson equations*

$$\nabla^2 u = 0\,, \qquad \nabla^2 u = -\varrho\,, \tag{10.24}$$

are elliptic. The homogeneous form is the potential equation of Laplace's celestial mechanics [45]; Poisson added the source term valid inside the distribution [57]. They govern electrostatics (Chapter 58) and Newtonian gravity.

The *Helmholtz equation*

$$\nabla^2 \psi + k^2 \psi = 0\,, \tag{10.25}$$

with k a wavenumber in 1/m, is elliptic and is the common time-reduced form of the first two: separating the time dependence out of Equation (10.22) or Equation (10.23) leaves Equation (10.25) for the spatial factor, as Section 10.2 shows. Its separable coordinate systems are catalogued by Morse and Feshbach [1023].

The *Schrödinger equation*

$$\mathrm{i}\hbar\,\partial_t \psi = -\frac{\hbar^2}{2m}\,\nabla^2 \psi + V(\boldsymbol{x})\,\psi\,, \tag{10.26}$$

with $\hbar$ in J s and m in kg, is the equation of Chapter 77 [557]. Its principal coefficients are imaginary, so the real trichotomy of Definition 10.6 does not apply to it directly: it has the shape of a diffusion equation with imaginary diffusivity $\mathrm{i}\hbar/2m$, and behaves like neither, propagating wave packets that spread without damping. Its relativistic counterparts in Chapter 92 are genuinely hyperbolic.

10.2 From partial to ordinary differential equations

The link between this chapter and the preceding one is the method of *separation of variables*: writing the unknown as a product of single-variable factors decomposes a linear PDE, in coordinates adapted to the boundary, into ordinary differential equations of Sturm–Liouville type. It is by this route that the classical families of Chapter 9 enter physics: the one-dimensional Helmholtz equation and its Fourier system arise on intervals (Section 9.7), the Bessel functions in cylindrical problems (Section 9.8), and the Legendre functions and spherical harmonics in spherical ones (Section 9.9) [1976].

The whole method rests on one elementary observation.

Lemma 10.28 (The separation constant)**.** *Let I_1, I_2 be intervals and let $F : I_1 \to \mathbb{R}$, $G : I_2 \to \mathbb{R}$ satisfy $F(s) = G(t)$ for every $(s,t) \in I_1 \times I_2$. Then F and G are one and the same constant.*

Rests on Definition 10.2.

Proof. *Derives Lemma 10.28.*

Fix $t_0 \in I_2$. Then $F(s) = G(t_0)$ for every $s \in I_1$, so F is the constant $G(t_0)$; symmetrically G is the constant $F(s_0)$, and the two constants agree. □

Proposition 10.29 (Time separation reduces the canonical equations to Helmholtz)**.** *Let $u(\boldsymbol{x}, t) = T(t)\,\psi(\boldsymbol{x})$ with T, ψ not identically zero. Then u solves Equation (10.23) if and only if there is a constant λ with*

$$\nabla^2\psi + \lambda\,\psi = 0\,, \qquad \frac{\mathrm{d}T}{\mathrm{d}t} = -D\lambda\,T\,, \tag{10.27}$$

and it solves Equation (10.22) if and only if

$$\nabla^2\psi + \lambda\,\psi = 0\,, \qquad \frac{\mathrm{d}^2T}{\mathrm{d}t^2} = -c^2\lambda\,T\,. \tag{10.28}$$

In both cases the spatial factor obeys the Helmholtz equation Equation (10.25) with $k^2 = \lambda$.

Rests on Lemma 10.28 and Equation (10.25).

Proof. *Derives Proposition 10.29.*

Substituting $u = T\psi$ into Equation (10.23) and dividing by $DT\psi$ where it is nonzero gives

$$\frac{1}{D}\,\frac{T'(t)}{T(t)} = \frac{\nabla^2\psi(\boldsymbol{x})}{\psi(\boldsymbol{x})}\,. \tag{10.29}$$

The left side depends on t alone and the right on $\boldsymbol{x}$ alone, so by Lemma 10.28 both equal a constant, written $-\lambda$; that is Equation (10.27). The wave case is identical with T'/D replaced by T''/c^2. Conversely a pair satisfying Equation (10.27) reassembles into a solution by direct substitution, and likewise for Equation (10.28). □

Remark 10.30 (What the ansatz is, and what it can miss). The product form is not an assumption about the solution sought. It is a device for manufacturing a supply of exact solutions, each of which is then a legitimate solution of the linear homogeneous equation; the solution of the actual problem is assembled from them by superposition, and the assembly is what has to be justified. Two things can go wrong. The separated solutions may fail to span the data — the question of completeness, settled for a regular Sturm–Liouville problem in Chapter 9 and in the abstract in Section 12.2; and the boundary may fail to be a coordinate surface of any system in which the operator separates, in which case the method does not start at all. Superposition also requires linearity, so nothing in this section survives into Section 10.6.3.

10.2.1 Separable coordinate systems

The three systems this treatise uses are worked out here; the general classification follows.

Example 10.31 (Cartesian separation)**.** Put $\psi = X(x)Y(y)Z(z)$ in Equation (10.25) and divide by ψ:

$$\frac{X''}{X} + \frac{Y''}{Y} + \frac{Z''}{Z} + k^2 = 0\,. \tag{10.30}$$

Each quotient depends on one variable alone, so each is a constant by Lemma 10.28: $X'' = -\alpha^2 X$, $Y'' = -\beta^2 Y$, $Z'' = -\gamma^2 Z$ with $\alpha^2 + \beta^2 + \gamma^2 = k^2$. Each factor obeys the one-dimensional Helmholtz equation Equation (9.106), whose eigenfunctions on an interval are the Fourier system (Section 9.7.4).

Rests on Lemma 10.28 and Equation (9.106).

Example 10.32 (Cylindrical separation)**.** In circular cylindrical coordinates (ρ, φ, z),

$$\nabla^2 = \frac{1}{\rho}\,\partial_\rho(\rho\,\partial_\rho) + \frac{1}{\rho^2}\,\partial_\varphi^2 + \partial_z^2\,. \tag{10.31}$$

With $\psi = R(\rho)\Phi(\varphi)Z(z)$, single-valuedness forces $\Phi'' = -m^2\Phi$ with $m \in \mathbb{Z}$, and $Z'' = -\gamma^2 Z$; the radial factor then obeys

$$\frac{1}{\rho}\frac{\mathrm{d}}{\mathrm{d}\rho}\left(\rho\,\frac{\mathrm{d}R}{\mathrm{d}\rho}\right) + \left(k^2 - \gamma^2 - \frac{m^2}{\rho^2}\right) R = 0\,, \tag{10.32}$$

which is Bessel's equation Equation (9.125) of order m in the variable $\sqrt{k^2-\gamma^2}\,\rho$.

Rests on Lemma 10.28 and Equation (9.125).

Example 10.33 (Spherical separation)**.** In spherical coordinates (r, θ, φ),

$$\nabla^2 = \frac{1}{r^2}\,\partial_r(r^2\,\partial_r) + \frac{1}{r^2}\left[\frac{1}{\sin\theta}\,\partial_\theta(\sin\theta\,\partial_\theta) + \frac{1}{\sin^2\theta}\,\partial_\varphi^2\right]. \tag{10.33}$$

With $\psi = R(r)\,Y(\theta,\varphi)$ the bracket acting on Y must be a constant multiple of Y; regularity on the whole sphere fixes that constant to be $-l(l+1)$ with $l \in \mathbb{N}$, and Y is a spherical harmonic Equation (9.163), its polar factor obeying the associated Legendre equation Equation (9.157) in $\cos\theta$. The radial factor obeys

$$\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}R}{\mathrm{d}r}\right) + \left(k^2 - \frac{l(l+1)}{r^2}\right) R = 0\,, \tag{10.34}$$

the spherical Bessel equation, which the substitution $R = r^{-1/2}\,W$ turns into Equation (9.125) of order $l + \frac{1}{2}$. For $k = 0$ this is the Laplace equation and Equation (10.34) is solved by r^l and r^{-l-1}: the multipole expansion of Chapter 58, and, with the Coulomb potential restored, the radial problem of Chapter 81.

Rests on Lemma 10.28 and Equations (9.157) and (9.163).

Theorem 10.34 (Separable systems for the Helmholtz operator)**.** *In three dimensions the Helmholtz equation Equation (10.25) separates in exactly eleven orthogonal coordinate systems: Cartesian, circular cylindrical, elliptic cylindrical, parabolic cylindrical, spherical, prolate spheroidal, oblate spheroidal, parabolic, conical, ellipsoidal and paraboloidal. Separation holds if and only if the metric coefficients of the system satisfy the Stäckel conditions, which state that the reciprocals of the scale factors are expressible through a matrix whose rows depend each on one coordinate alone.*

Rests on Equation (10.25) and Lemma 10.28.

Full derivation: Appendix A.15 (p. 2831).
Derives Theorem 10.34.

The Stäckel conditions are derived there — both their necessity and their sufficiency — by substituting the product ansatz into the Laplacian of an orthogonal system and demanding that the three separation constants enter the separated equations linearly; the ellipsoidal system, of which the three examples above are degenerate limits, is verified against them outright, and the eleven systems are obtained as the degenerations of one confocal quadric family. Only the completeness of that list — that no orthogonal system outside the confocal scheme separates — is quoted, from the catalogue of Morse and Feshbach [1023], whose tables give the separated equations system by system.

Remark 10.35. The Laplace equation, being the case $k = 0$, separates in strictly more systems than the Helmholtz equation — bispherical and toroidal coordinates among them — because it also admits *R-separation*, in which the product ansatz is preceded by division by a fixed function of position. A Stäckel matrix of the same shape governs the separability of the Hamilton–Jacobi equation of Chapter 23, which is not a coincidence: both ask when a single operator falls apart along the coordinate directions. The two statements are nevertheless different, the Hamilton–Jacobi equation being of first order and nonlinear, and it is Theorem 23.17 that carries that case; Remark 23.19 sets the two side by side and explains why the eleven systems above coincide with the separable systems of a free particle. In practice the coordinate system is chosen by the boundary, not by the operator: one separates in the system whose surfaces the boundary is made of, and the worked boundary-value problems appear in Chapter 58 and Chapter 81.

10.2.2 Eigenfunction expansions

Separation supplies a family of solutions indexed by the separation constants; the data are matched by superposing them, which is possible exactly when the separated eigenfunctions are complete.

Theorem 10.36 (Solution of the heat problem on an interval)**.** *Let f be continuous on $[0, L]$ with $f(0) = f(L) = 0$ and let*

$$c_n = \frac{2}{L}\int_0^L f(x)\,\sin\!\left(\frac{n\pi x}{L}\right)\mathrm{d}x\,. \tag{10.35}$$

Then

$$u(x,t) = \sum_{n=1}^{\infty} c_n\,\mathrm{e}^{-Dn^2\pi^2 t/L^2}\,\sin\!\left(\frac{n\pi x}{L}\right) \tag{10.36}$$

converges for $t > 0$ together with all its derivatives, solves Equation (10.23) on $(0, L)\times(0,\infty)$ with $u(0,t) = u(L,t) = 0$, and satisfies $u(\cdot,t) \to f$ in the mean square as $t \to 0^+$.
Rests on Proposition 10.29 and Remark 9.53.

Proof. *Derives Theorem 10.36.*

Each summand is a separated solution: by Proposition 10.29 with $\lambda = n^2\pi^2/L^2$ the spatial factor $\sin(n\pi x/L)$ solves the one-dimensional Helmholtz equation with the stated boundary values and the temporal factor solves $T' = -D\lambda T$. Bessel's inequality Equation (9.53) bounds the coefficients, $|c_n| \le C$ uniformly in n; for $t \ge t_0 > 0$ the factor $\mathrm{e}^{-Dn^2\pi^2 t/L^2}$ then dominates any fixed power of n, so the series and every series obtained from it by differentiating term by term in x or t converge absolutely and uniformly on $[0, L]\times[t_0,\infty)$. Term-by-term differentiation is therefore legitimate there and each term satisfies the equation, so u does. Since $t_0 > 0$ was arbitrary, this holds for every $t > 0$. Finally the partial sums of Equation (10.36) at $t = 0$ are the partial sums of the sine series of f, which converge to f in the mean square by completeness of the Sturm–Liouville eigenfunctions (Remark 9.53);

the exponential factors are bounded by 1 and tend to 1 for each fixed n, whence the stated limit. □

Remark 10.37 (Smoothing). The proof shows more than it was asked for. The initial datum was merely continuous, yet for every $t > 0$ the solution is infinitely differentiable, because the factor $\mathrm{e}^{-Dn^2\pi^2 t/L^2}$ kills the high harmonics faster than differentiation can excite them. Diffusion destroys the information carried by the fine structure of the datum, and that destruction is exactly the obstruction met in Proposition 10.25.

Example 10.38 (The vibrating string)**.** The same separation applied to Equation (10.22) on $[0, L]$ with fixed ends gives, by Equation (10.28),

$$u(x,t) = \sum_{n=1}^{\infty} \left[a_n \cos\!\left(\frac{n\pi c t}{L}\right) + b_n \sin\!\left(\frac{n\pi c t}{L}\right) \right] \sin\!\left(\frac{n\pi x}{L}\right), \tag{10.37}$$

with a_n and $n\pi c\, b_n/L$ the sine coefficients of the initial displacement and velocity. Here the exponentials of Theorem 10.36 are replaced by oscillations of the same magnitude for all n, so nothing is damped and no smoothing occurs: the wave equation remembers. That the general motion is a superposition of such modes is Daniel Bernoulli's claim [28], contested at the time and vindicated only by Fourier's theory of the coefficients [72]; the physical reading, in which the n-th term is a normal mode, is developed in Chapter 28.

Rests on Proposition 10.29 and Equation (10.28).

10.2.3 Integral-transform methods

On an unbounded domain the eigenfunctions are not discrete and the sum of Section 10.2.2 becomes an integral. The machinery is that of Chapter 17, and the principle is one line: a transform that converts differentiation into multiplication converts a PDE in $N+1$ variables into a problem in one.

Proposition 10.39 (Fourier reduction on the line)**.** *Let $P(\partial_x)$ be a constant-coefficient operator in x and consider $\partial_t u = P(\partial_x)u$ for $x \in \mathbb{R}$, $t > 0$, with $u(\cdot, 0) = f$ absolutely integrable. Then the transform $\hat{u}(\xi, t)$ of u in x satisfies the ordinary differential equation*

$$\partial_t \hat{u}(\xi,t) = P(\mathrm{i}\xi)\,\hat{u}(\xi,t), \qquad \hat{u}(\xi,0) = \hat{f}(\xi), \tag{10.38}$$

one equation for each ξ, whose solution is $\hat{u} = \hat{f}\,\mathrm{e}^{P(\mathrm{i}\xi)t}$.

Rests on Equation (17.38) *and Definition 17.29.*

Proof. *Derives Proposition 10.39.*

Each x-derivative contributes a factor $\mathrm{i}\xi$ under the transform (Equation (17.38)), so $\widehat{P(\partial_x)u} = P(\mathrm{i}\xi)\hat{u}$; the t-derivative passes through the x-integral. What remains is a linear first-order ordinary differential equation in t with ξ as a parameter, and Equation (10.38) is its statement. The inversion Equation (17.34) returns u. □

Remark 10.40 (Laplace on the half-line). When the variable runs over a half-line — as time does, a system being switched on at $t = 0$ — the Laplace transform of Definition 17.65 is the right tool, because its derivative rule Equation (17.87) does not merely multiply by s but absorbs the initial values into the transformed equation as polynomial terms. The initial conditions are thus imposed automatically rather than fitted afterwards, and the solution is recovered from the Bromwich integral Theorem 17.69 by residues. The choice of transform is therefore dictated by the domain and the side conditions, not by taste: Fourier for a whole line with decay at both ends, Laplace for a half-line with data at its origin, a series for a bounded interval.

Remark 10.41 (Dispersion relations). Substituting the plane wave $u = \mathrm{e}^{\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{x}-\omega t)}$ into a constant-coefficient linear equation reduces it to an algebraic relation between ω and $\boldsymbol{k}$ — the dispersion relation, which is the symbol Equation (10.4) evaluated on the wave. For Equation (10.22) it is $\omega = \pm c\,|\boldsymbol{k}|$: all wavelengths travel at the same speed and a signal keeps its shape. For Equation (10.23) it is $\omega = -\mathrm{i}D\,|\boldsymbol{k}|^2$, purely imaginary, so nothing propagates and everything decays, the finer structure faster. For Equation (10.26) with $V = 0$ it is $\hbar\omega = \hbar^2\,|\boldsymbol{k}|^2\,/2m$, real but nonlinear in $|\boldsymbol{k}|$: propagation with spreading. The three types of Definition 10.6 are visible in these three answers, and the physical consequences are drawn in Chapter 28 and Chapter 71.

10.3 Green's functions

For an inhomogeneous linear problem, the *Green's function* — the response of the differential operator to a point source — converts the equation into an integral representation of the solution. Green introduced both the function and the reciprocity it obeys, for the potential equation, in a privately printed essay [83].

Definition 10.42 (Green's function)**.** Let L be a linear differential operator on Ω together with homogeneous boundary conditions. A *Green's function* for the pair is a function $G(\boldsymbol{x}, \boldsymbol{x}')$ satisfying, in the sense of distributions in $\boldsymbol{x}$ for each fixed $\boldsymbol{x}' \in \Omega$,

$$L\,G(\boldsymbol{x}, \boldsymbol{x}') = \delta(\boldsymbol{x} - \boldsymbol{x}')\,, \tag{10.39}$$

and the boundary conditions in the variable $\boldsymbol{x}$.

Rests on Definitions 10.4 and 17.40.

Proposition 10.43 (Representation of the solution)**.** *If G is a Green's function for L with the given homogeneous boundary conditions and f is such that the integral converges, then*

$$u(\boldsymbol{x}) = \int_\Omega G(\boldsymbol{x}, \boldsymbol{x}')\, f(\boldsymbol{x}')\, \mathrm{d}^3x' \tag{10.40}$$

solves $Lu = f$ with those boundary conditions.

Rests on Definitions 10.2 and 10.42.

Proof. *Derives Proposition 10.43.*

L acts on $\boldsymbol{x}$ and the integration is over $\boldsymbol{x}'$, so L may be taken under the integral sign: $Lu(\boldsymbol{x}) = \int_\Omega [LG(\boldsymbol{x}, \boldsymbol{x}')] f(\boldsymbol{x}')\mathrm{d}^3x' = \int_\Omega \delta(\boldsymbol{x} - \boldsymbol{x}') f(\boldsymbol{x}')\mathrm{d}^3x' = f(\boldsymbol{x})$ by Equation (10.39) and the defining property of the Dirac distribution (Definition 17.38). Each $G(\cdot, \boldsymbol{x}')$ satisfies the homogeneous boundary conditions, and so therefore does the superposition. □

The content of the method is thus entirely in the construction of G, to which the next two subsections are devoted. Note what has been gained: one function of two points, computed once, solves the problem for every source.

10.3.1 The Green's function of a boundary-value problem

Lemma 10.44 (Green's identities)**.** *Let $V \subset \mathbb{R}^3$ be a bounded region with piecewise-smooth boundary and let $u, v \in C^2(\overline{V})$, $p \in C^1$. Then*

$$\int_V \left[v\,\nabla\cdot(p\nabla u) - u\,\nabla\cdot(p\nabla v)\right] \mathrm{d}^3x = \oint_{\partial V} p\,(v\,\partial_\nu u - u\,\partial_\nu v)\,\mathrm{d}S\,, \tag{10.41}$$

where ∂_ν is the outward normal derivative.

Rests on Theorem 7.101 and Definition 7.66.

Proof. *Derives Lemma 10.44.*

Apply the divergence theorem Equation (7.123) to the vector field $\boldsymbol{A} = p\,(v\nabla u - u\nabla v)$. Its divergence is $v\,\nabla\cdot(p\nabla u) - u\,\nabla\cdot(p\nabla v)$, the two terms $p\nabla v\cdot\nabla u$ produced by the product rule cancelling; and $\boldsymbol{A}\cdot\mathrm{d}\boldsymbol{S} = p(v\partial_\nu u - u\partial_\nu v)\mathrm{d}S$. □

Theorem 10.45 (Symmetry of the Green's function)**.** *Let $Lu = \nabla\cdot(p\nabla u) + q\,u$ with p, q real, and let the boundary conditions be homogeneous of the form $\alpha u + \beta\,\partial_\nu u = 0$ on ∂V with real α, β not both zero. Then the Green's function is symmetric,*

$$G(\boldsymbol{x}_1, \boldsymbol{x}_2) = G(\boldsymbol{x}_2, \boldsymbol{x}_1)\,. \tag{10.42}$$

Rests on Lemma 10.44 and Definition 10.42.

Proof. *Derives Theorem 10.45.*

Put $u(\boldsymbol{x}) = G(\boldsymbol{x}, \boldsymbol{x}_1)$ and $v(\boldsymbol{x}) = G(\boldsymbol{x}, \boldsymbol{x}_2)$ in Equation (10.41). The terms in q are symmetric in u and v and cancel, so the left-hand side is

$$\int_V \left[v\,\delta(\boldsymbol{x} - \boldsymbol{x}_1) - u\,\delta(\boldsymbol{x} - \boldsymbol{x}_2)\right]\mathrm{d}^3x = G(\boldsymbol{x}_1, \boldsymbol{x}_2) - G(\boldsymbol{x}_2, \boldsymbol{x}_1)\,. \tag{10.43}$$

On the boundary both u and v satisfy $\alpha w + \beta\partial_\nu w = 0$. If $\beta \neq 0$ then $\partial_\nu w = -(\alpha/\beta)w$ for both, and $v\partial_\nu u - u\partial_\nu v = -(\alpha/\beta)(vu - uv) = 0$; if $\beta = 0$ then $u = v = 0$ there and the same combination vanishes. So the right-hand side of Equation (10.41) is zero, and Equation (10.43) gives Equation (10.42). □

Remark 10.46. The hypothesis is exactly that L is formally self-adjoint and that the boundary conditions make it symmetric on its domain — the Sturm–Liouville form Equation (9.54) of Chapter 9, in one dimension. The symmetry of G is therefore not an accident of the potential problem but the kernel-level shadow of self-adjointness, and it is the statement Green called reciprocity: the field at $\boldsymbol{x}_1$ due to a unit source at $\boldsymbol{x}_2$ equals the field at $\boldsymbol{x}_2$ due to a unit source at $\boldsymbol{x}_1$. The operator-theoretic setting, in which a self-adjoint operator is diagonalised by the spectral theorem, is Chapter 12.

Proposition 10.47 (Bilinear expansion)**.** *Let $\{\psi_n\}$ be a complete orthonormal system of eigenfunctions of L on V satisfying the boundary conditions, with $L\psi_n = \lambda_n\psi_n$ and $\lambda_n \neq \mu$ for all n. Then the Green's function of $L - \mu$ is*

$$G_\mu(\boldsymbol{x}, \boldsymbol{x}') = \sum_n \frac{\psi_n(\boldsymbol{x})\,\overline{\psi_n(\boldsymbol{x}')}}{\lambda_n - \mu}\,. \tag{10.44}$$

Rests on Definition 10.42 and Remark 9.53.

Proof. *Derives Proposition 10.47.*

Apply $L - \mu$ term by term in $\boldsymbol{x}$: the n-th term becomes $\psi_n(\boldsymbol{x})\overline{\psi_n(\boldsymbol{x}')}$, and completeness of the system is the closure relation $\sum_n \psi_n(\boldsymbol{x})\overline{\psi_n(\boldsymbol{x}')} = \delta(\boldsymbol{x} - \boldsymbol{x}')$ (Remark 9.53), which is Equation (10.39) for $L - \mu$. Each term satisfies the boundary conditions, hence so does the sum. □

Remark 10.48. Equation (10.44) displays the Green's function as a meromorphic function of the parameter μ whose poles are exactly the eigenvalues and whose residues are the eigenprojections. This is the resolvent operator of Chapter 12 written as a kernel, and it is why the analytic structure of a response function encodes a spectrum — the observation that drives the contour techniques of Chapter 8 and, in Chapter 65, the identification of resonances with poles.

Example 10.49 (The method of images)**.** For $L = \nabla^2$ on the half-space $z > 0$ with $G = 0$ on the plane $z = 0$, write $\boldsymbol{x}'^{*}$ for the mirror image of $\boldsymbol{x}'$ in that plane. Then

$$G(\boldsymbol{x}, \boldsymbol{x}') = -\frac{1}{4\pi}\left(\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} - \frac{1}{|\boldsymbol{x} - \boldsymbol{x}'^{*}|}\right). \tag{10.45}$$

The second term is harmonic throughout $z > 0$, since its singularity lies outside; so Equation (10.39) holds by Proposition 10.67, proved below, which identifies $-1/4\pi r$ as the fundamental solution of the Laplacian. On $z = 0$ the two distances are equal and G vanishes. The construction succeeds whenever the boundary is a plane or a sphere, and fails otherwise; the electrostatic reading is in Chapter 58 [1976], and the catalogue of Green's functions for the standard operators and domains is Morse–Feshbach [1023].

Rests on Proposition 10.67 and Definition 10.42.

10.3.2 Causal Green's functions and Duhamel's principle

Definition 10.50 (Retarded and advanced Green's functions)**.** The *retarded* and *advanced* Green's functions of the wave operator Equation (10.1) in three space dimensions are

$$G_{\pm}(\boldsymbol{x}, t) = \frac{\delta\left(t \mp |\boldsymbol{x}|/c\right)}{4\pi\,|\boldsymbol{x}|}, \tag{10.46}$$

each satisfying $\Box G_{\pm} = \delta(\boldsymbol{x})\,\delta(t)$, with G_+ supported on the future light cone of the source and G_- on its past cone.

Rests on Definition 10.42 and Equation (10.1).

Both solve the same equation, so the equation does not choose between them; the choice is a physical postulate, that an effect follows its cause. Selecting G_+ turns Equation (10.40) into the retarded potentials of Chapter 60 and the radiation fields of Chapter 65: the field at $(\boldsymbol{x}, t)$ is the source evaluated at the earlier time $t - |\boldsymbol{x} - \boldsymbol{x}'|/c$, integrated over space.

Remark 10.51 (The causal choice as a contour prescription). Transforming Equation (10.39) for $\Box$ in space and time gives $(|\boldsymbol{k}|^2 - \omega^2/c^2)\hat{G} = 1$, whose right-hand side is finite while the left vanishes on the light cone $\omega = \pm c\,|\boldsymbol{k}|$: the inverse transform is an integral through two poles on the real ω axis and is undefined until they are avoided. Which way to close the contour is fixed by the synthesis kernel of Equation (17.34), which in the convention of Definition 17.29 is $\mathrm{e}^{+\mathrm{i}\omega t}$: on a large semicircle at $\Im\omega = \tau$ its modulus is $\mathrm{e}^{-\tau t}$, so the contour must be closed in the upper half-plane for $t > 0$ and in the lower for $t < 0$. Displacing the poles by $\omega \to \omega - \mathrm{i}\varepsilon$ in the denominator moves them to $\omega = \pm c\,|\boldsymbol{k}| + \mathrm{i}\varepsilon$, both in the upper half-plane, so closing downward encloses nothing and gives zero for $t < 0$ (Corollary 8.15) while closing upward gives the retarded function for $t > 0$; the opposite displacement gives the advanced one. Causality is thus one of the two possible pole prescriptions, and the Sokhotski–Plemelj formula Equation (17.54) is the statement that the difference between them is a delta function on the light cone. The residue calculus that evaluates either is Theorem 8.24.

The sign of ε is convention-dependent and is quoted the other way round in much of the physics literature, which synthesises with $\mathrm{e}^{-\mathrm{i}\omega t}$ — the kernel of the plane-wave ansatz $\mathrm{e}^{\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)}$ of Remark 10.41 — rather than with the $\mathrm{e}^{+\mathrm{i}\omega t}$ fixed for this treatise in Definition 17.29. Under that reading the half-planes are exchanged and the retarded prescription is $\omega \to \omega + \mathrm{i}\varepsilon$. What is convention-independent, and is the content of the remark, is that the retarded function is the one whose poles lie on the side of the real axis that the closure for $t > 0$ encircles.

Theorem 10.52 (Duhamel's principle)**.** *Let $\mathcal{A}$ be a linear operator acting on the space variables only. Suppose that for each $s \geq 0$ the function $v(\cdot, t; s)$ solves the homogeneous*

problem $\partial_t v = \mathcal{A} v$ *for* $t > s$ *with* $v(\cdot, s; s) = f(\cdot, s)$. *Then*

$$u(\boldsymbol{x}, t) = \int_0^t v(\boldsymbol{x}, t; s)\, \mathrm{d}s \tag{10.47}$$

solves the driven problem $\partial_t u = \mathcal{A} u + f$ *with* $u(\cdot, 0) = 0$.

Rests on Definition 10.2 and Theorem 7.42.

Proof. *Derives Theorem 10.52.*

Differentiating Equation (10.47) in t, the upper limit contributes the integrand evaluated at $s = t$ and the integrand contributes its own t-derivative (Theorem 7.42):

$$\partial_t u(\boldsymbol{x}, t) = v(\boldsymbol{x}, t; t) + \int_0^t \partial_t v(\boldsymbol{x}, t; s)\, \mathrm{d}s = f(\boldsymbol{x}, t) + \int_0^t \mathcal{A} v(\boldsymbol{x}, t; s)\, \mathrm{d}s\,, \tag{10.48}$$

using the initial condition of v in the first term and its equation in the second. Since $\mathcal{A}$ does not act on s, the remaining integral is $\mathcal{A} u$. That $u(\cdot, 0) = 0$ is immediate from Equation (10.47). □

Corollary 10.53 (Duhamel for the wave equation)**.** *If* $w(\cdot, t; s)$ *solves* $\partial_t^2 w = c^2 \nabla^2 w$ *for* $t > s$ *with* $w(\cdot, s; s) = 0$ *and* $\partial_t w(\cdot, s; s) = f(\cdot, s)$, *then* $u = \int_0^t w(\boldsymbol{x}, t; s) \mathrm{d}s$ *solves* $\partial_t^2 u - c^2 \nabla^2 u = f$ *with vanishing initial data.*

Rests on Theorem 10.52 and Equation (10.22).

Proof. *Derives Corollary 10.53.*

The first derivative is $\partial_t u = w(\boldsymbol{x}, t; t) + \int_0^t \partial_t w\, \mathrm{d}s = \int_0^t \partial_t w\, \mathrm{d}s$, since w vanishes on the diagonal. Differentiating once more, $\partial_t^2 u = \partial_t w(\boldsymbol{x}, t; t) + \int_0^t \partial_t^2 w\, \mathrm{d}s = f + c^2 \nabla^2 u$, by the second initial condition and the equation for w. □

Remark 10.54. Duhamel's principle is superposition in the time variable: a distributed source is a continuum of impulses, and the response is the integral of the impulse responses. It is the same statement as Proposition 10.43 with the causal Green's function inserted, and in the language of Chapter 17 it is the convolution Equation (17.66) of the source with the impulse response of a time-invariant linear system.

10.4 The canonical equations and their fundamental solutions

10.4.1 The wave equation

Theorem 10.55 (d'Alembert's formula)**.** *Let* $f \in C^2(\mathbb{R})$ *and* $g \in C^1(\mathbb{R})$. *The unique* C^2 *solution of* $\partial_t^2 u = c^2 \partial_x^2 u$ *on* $\mathbb{R} \times \mathbb{R}$ *with* $u(x, 0) = f(x)$ *and* $\partial_t u(x, 0) = g(x)$ *is*

$$u(x, t) = \frac{f(x - ct) + f(x + ct)}{2} + \frac{1}{2c} \int_{x - ct}^{x + ct} g(y)\, \mathrm{d}y\,. \tag{10.49}$$

Rests on Proposition 10.13 and Theorem 7.43.

Proof. *Derives Theorem 10.55.*

Introduce the characteristic coordinates $\xi = x - ct$, $\eta = x + ct$ of Proposition 10.13. Then $\partial_x = \partial_\xi + \partial_\eta$ and $\partial_t = c\,(\partial_\eta - \partial_\xi)$, so

$$\partial_t^2 - c^2 \partial_x^2 = c^2\,(\partial_\eta - \partial_\xi)^2 - c^2\,(\partial_\xi + \partial_\eta)^2 = -4c^2\, \partial_\xi \partial_\eta\,. \tag{10.50}$$

The equation is therefore $\partial_\xi \partial_\eta u = 0$. Integrating in η shows $\partial_\xi u$ is a function of ξ alone, and integrating in ξ (Theorem 7.43) gives $u = F(\xi) + G(\eta)$ with $F, G \in C^2$; conversely every such u solves the equation, so this is the general solution.

Imposing the data at $t = 0$, where $\xi = \eta = x$, gives $F(x) + G(x) = f(x)$ and $c\,(G'(x) - F'(x)) = g(x)$. Integrating the second from a fixed x_0, $G(x) - F(x) = c^{-1}\int_{x_0}^{x} g + A$ for a constant A. Solving the pair,

$$F(x) = \frac{f(x)}{2} - \frac{1}{2c}\int_{x_0}^{x} g - \frac{A}{2}\,, \qquad G(x) = \frac{f(x)}{2} + \frac{1}{2c}\int_{x_0}^{x} g + \frac{A}{2}\,, \tag{10.51}$$

and $u = F(x - ct) + G(x + ct)$ is Equation (10.49), the constant A cancelling and the two partial integrals combining into one over $[x - ct, x + ct]$. Uniqueness follows because every solution has the form $F(\xi) + G(\eta)$ and the data determine F and G up to the constant that cancels. □

Corollary 10.56 (Domain of dependence and influence)**.** *The value $u(x_0, t_0)$ with $t_0 > 0$ depends on the data only through their restriction to the interval $[x_0 - ct_0,\ x_0 + ct_0]$. Conversely the data near a point x_1 influence the solution only inside the forward cone $|x - x_1| \le c\,t$. In particular no signal travels faster than c.*

Rests on Theorem 10.55.

Proof. *Derives Corollary 10.56.*

Read Equation (10.49): f enters only at the two endpoints of the interval and g only through its integral over the interval. The converse statement is the contrapositive, since x_1 lies in the interval of dependence of (x, t) if and only if $|x - x_1| \le ct$. □

The same conclusion holds in three space dimensions, and is proved there without any formula, by an energy estimate that will also give uniqueness.

Proposition 10.57 (Energy in a backward cone)**.** *Let $u \in C^2$ solve $\partial_t^2 u = c^2 \nabla^2 u$ on $\mathbb{R}^3 \times (0, \infty)$, fix $(\boldsymbol{x}_0, t_0)$ and set*

$$e(t) = \frac{1}{2}\int_{B(\boldsymbol{x}_0,\, c(t_0 - t))} \Big[(\partial_t u)^2 + c^2\,|\nabla u|^2\Big]\,\mathrm{d}^3x\,, \qquad 0 \le t < t_0\,. \tag{10.52}$$

Then e is non-increasing. Consequently, if u and $\partial_t u$ vanish at $t = 0$ on the ball $B(\boldsymbol{x}_0, ct_0)$, then u vanishes throughout the backward cone, and solutions of the Cauchy problem are unique.

Rests on Theorem 7.101 and Equation (10.22).

Proof. *Derives Proposition 10.57.*

Write $R(t) = c(t_0 - t)$ and differentiate Equation (10.52). The integrand's own t-derivative gives, on using the equation and the divergence theorem Equation (7.123) on $\nabla \cdot (\partial_t u\,\nabla u)$,

$$\int_B \Big[\partial_t u\,\partial_t^2 u + c^2 \nabla u \cdot \nabla \partial_t u\Big]\,\mathrm{d}^3x = c^2 \oint_{\partial B} \partial_t u\,\partial_\nu u\,\mathrm{d}S\,, \tag{10.53}$$

because the remaining volume term is $\int \partial_t u\,(\partial_t^2 u - c^2\nabla^2 u) = 0$. The shrinking of the ball, $\dot R = -c$, contributes $-\frac{c}{2}\oint_{\partial B}\Big[(\partial_t u)^2 + c^2\,|\nabla u|^2\Big]\,\mathrm{d}S$. Now for real numbers $a = \partial_t u$ and $b = |\nabla u|$ one has $c^2 ab \le \frac{c}{2}\,(a^2 + c^2 b^2)$, since the difference is $\frac{c}{2}(a - cb)^2 \ge 0$; and $|\partial_\nu u| \le |\nabla u|$. The surface term of Equation (10.53) is therefore dominated by the negative one, so $\dot e(t) \le 0$.

If the data vanish on $B(\boldsymbol{x}_0, ct_0)$ then $e(0) = 0$, and since $e \ge 0$ and e is non-increasing, $e \equiv 0$: both $\partial_t u$ and ∇u vanish in the cone, so u is constant there, and the constant is zero. Uniqueness follows by applying this to the difference of two solutions, whose data vanish everywhere. □

Lemma 10.58 (Darboux's equation for spherical means)**.** *For $h \in C^2(\mathbb{R}^3)$ define the spherical mean*

$$M_h(\boldsymbol{x}, r) = \frac{1}{4\pi} \int_{S^2} h(\boldsymbol{x} + r\boldsymbol{n}) \, \mathrm{d}\Omega \,, \tag{10.54}$$

the integral running over the unit sphere with total solid angle 4π. Then $M_h(\boldsymbol{x}, r) \to h(\boldsymbol{x})$ as $r \to 0$ and

$$\partial_r M_h = \frac{1}{4\pi r^2} \int_{B_r(\boldsymbol{x})} \nabla^2 h \, \mathrm{d}^3 y \,, \qquad \left(\partial_r^2 + \frac{2}{r}\partial_r\right) M_h = M_{\nabla^2 h} = \nabla_{\boldsymbol{x}}^2 M_h \,. \tag{10.55}$$

Rests on Theorem 7.101 and Definition 7.66.

Proof. *Derives Lemma 10.58.*

Differentiating Equation (10.54) under the integral sign, $\partial_r M_h = \frac{1}{4\pi} \int_{S^2} \boldsymbol{n} \cdot \nabla h(\boldsymbol{x} + r\boldsymbol{n}) \mathrm{d}\Omega$, which on rewriting the solid angle as a surface element, $\mathrm{d}S = r^2 \mathrm{d}\Omega$, is $\frac{1}{4\pi r^2} \oint_{\partial B_r} \partial_\nu h \, \mathrm{d}S$; the divergence theorem Equation (7.123) turns this into the first half of Equation (10.55). Multiplying by r^2 and differentiating again,

$$\partial_r \left(r^2 \partial_r M_h\right) = \frac{1}{4\pi} \partial_r \int_{B_r} \nabla^2 h \, \mathrm{d}^3 y = \frac{1}{4\pi} \oint_{\partial B_r} \nabla^2 h \, \mathrm{d}S = r^2 \, M_{\nabla^2 h} \,, \tag{10.56}$$

which is the second half after dividing by r^2. The last equality in Equation (10.55) holds because $\nabla_{\boldsymbol{x}}^2$ commutes with the average over $\boldsymbol{n}$. The limit $r \to 0$ is the continuity of h. □

Theorem 10.59 (Kirchhoff's formula)**.** *Let $f \in C^3(\mathbb{R}^3)$ and $g \in C^2(\mathbb{R}^3)$. The solution of $\partial_t^2 u = c^2 \nabla^2 u$ with $u(\cdot, 0) = f$ and $\partial_t u(\cdot, 0) = g$ is*

$$u(\boldsymbol{x}, t) = \partial_t \left[t \, M_f(\boldsymbol{x}, ct)\right] + t \, M_g(\boldsymbol{x}, ct) \,, \tag{10.57}$$

with M the spherical mean Equation (10.54).

Rests on Lemma 10.58 and Theorem 10.55.

Proof. *Derives Theorem 10.59.*

Fix $\boldsymbol{x}$ and write $M(r, t) = M_{u(\cdot,t)}(\boldsymbol{x}, r)$. Averaging the equation over the sphere of radius r and using Equation (10.55),

$$\partial_t^2 M = c^2 M_{\nabla^2 u} = c^2 \left(\partial_r^2 + \frac{2}{r}\partial_r\right) M \,. \tag{10.58}$$

Set $W(r, t) = r \, M(r, t)$. Then $\partial_r^2 W = r \left(\partial_r^2 M + \frac{2}{r}\partial_r M\right)$, so Equation (10.58) becomes the one-dimensional wave equation $\partial_t^2 W = c^2 \partial_r^2 W$ on $r > 0$, with $W(0, t) = 0$ and data $W(r, 0) = r \, M_f(\boldsymbol{x}, r)$, $\partial_t W(r, 0) = r \, M_g(\boldsymbol{x}, r)$. Extending both data as odd functions of r — which is consistent with the boundary condition at the origin — d'Alembert's formula Equation (10.49) applies. Writing $\tilde{F}(s) = s \, M_f(\boldsymbol{x}, s)$ and $\tilde{G}(s) = s \, M_g(\boldsymbol{x}, s)$ for the odd extensions,

$$W(r, t) = \frac{\tilde{F}(r + ct) - \tilde{F}(ct - r)}{2} + \frac{1}{2c} \int_{ct-r}^{ct+r} \tilde{G}(s) \, \mathrm{d}s \,, \tag{10.59}$$

for $0 < r < ct$, the oddness having been used to replace $\tilde{F}(r - ct)$ by $-\tilde{F}(ct - r)$. Now $u(\boldsymbol{x}, t) = \lim_{r \to 0^+} M(r, t) = \lim_{r \to 0^+} W(r, t)/r$. The first term of Equation (10.59) divided by r tends to $\tilde{F}'(ct)$ by the definition of the derivative, and the second to $\tilde{G}(ct)/c$ by continuity of $\tilde{G}$. Hence $u = \tilde{F}'(ct) + t \, M_g(\boldsymbol{x}, ct)$, and since $\partial_t \left[t \, M_f(\boldsymbol{x}, ct)\right] = M_f(ct) + ct \, \partial_s M_f(s)\big|_{s=ct} = \tilde{F}'(ct)$, this is Equation (10.57). □

Corollary 10.60 (Huygens' principle in three space dimensions)**.** *The value of u at $(\boldsymbol{x}, t)$ depends on the initial data only through their values on the sphere $|\boldsymbol{y} - \boldsymbol{x}| = ct$, and not at all on their values inside it. A signal emitted at one instant from one point is therefore received at one instant, with no trailing wake.*

Rests on Theorem 10.59 and Corollary 10.12.

Proof. *Derives Corollary 10.60.*

Every term of Equation (10.57) is an average over the sphere of radius ct, or a t-derivative of such an average, and no integration over the interior occurs. □

Remark 10.61. Huygens stated the sharp propagation of light as a construction of wave fronts [15]; Kirchhoff turned it into the formula Equation (10.57) [224], and Hadamard identified the property as special — it holds for the wave operator in three space dimensions and fails in general [502]. It is what makes hearing and vision possible: sound arrives when it arrives, and the acoustic and optical consequences are drawn in Chapter 28 and Chapter 61. The support of the solution on the characteristic cone of Corollary 10.12 is the same statement in the language of the classification.

10.4.2 The heat equation

Theorem 10.62 (The heat kernel)**.** *Let f be bounded and continuous on $\mathbb{R}$. Then*

$$K(x,t) = \frac{1}{\sqrt{4\pi D t}}\,\exp\!\left(-\frac{x^2}{4Dt}\right), \qquad t>0\,, \tag{10.60}$$

satisfies $\int_{\mathbb{R}} K(x,t)\mathrm{d}x = 1$ for every $t>0$, and $u = K(\cdot,t) * f$ solves Equation (10.23) on $\mathbb{R}\times(0,\infty)$ with $u(\cdot,t)\to f$ pointwise as $t\to 0^{+}$. In three dimensions the kernel is the product of three such factors, $K_3(\boldsymbol{x},t) = (4\pi Dt)^{-3/2}\exp\left(-|\boldsymbol{x}|^2/4Dt\right)$.

Rests on Proposition 10.39, Equation (17.42), and Theorem 17.53.

Proof. *Derives Theorem 10.62.*

By Proposition 10.39 with $P(\partial_x) = D\partial_x^2$, the transform of the solution is

$$\hat{u}(\xi,t) = \hat{f}(\xi)\,\mathrm{e}^{-D\xi^2 t}\,. \tag{10.61}$$

The convolution theorem Equation (17.66) says that a product of transforms is $(2\pi)^{-1/2}$ times the transform of the convolution, so $u = K * f$ with $\hat{K}(\xi) = (2\pi)^{-1/2}\mathrm{e}^{-Dt\xi^2}$. The Gaussian pair Equation (17.42) with $a = \sqrt{2Dt}$ reads $\widehat{\mathrm{e}^{-x^2/(4Dt)}}(\xi) = \sqrt{2Dt}\,\mathrm{e}^{-Dt\xi^2}$, whence $K = (2\pi)^{-1/2}(2Dt)^{-1/2}\mathrm{e}^{-x^2/(4Dt)}$, which is Equation (10.60). Its integral is $\hat{K}(0)\sqrt{2\pi} = 1$. That u solves the equation may now be checked directly: K is smooth for $t>0$ and differentiation under the integral sign is justified by the Gaussian decay, while $\partial_t K = D\partial_x^2 K$ is a computation from Equation (10.60). For the initial values, $K(\cdot,t)$ is a nonnegative kernel of unit mass concentrating on the origin, so $(K*f)(x)\to f(x)$ wherever f is continuous. In three dimensions the transformed solution is $\hat{f}\,\mathrm{e}^{-D|\boldsymbol{\xi}|^2 t}$, which factorises over the three coordinates, and so does the kernel. □

Proposition 10.63 (Infinite propagation speed)**.** *Let $f\geq 0$ be continuous, not identically zero, and supported in a bounded set. Then the solution of Equation (10.23) given by Theorem 10.62 satisfies $u(x,t)>0$ for every $x\in\mathbb{R}$ and every $t>0$.*

Rests on Theorem 10.62.

Proof. *Derives Proposition 10.63.*

$u(x,t) = \int_{\mathbb{R}} K(x-y,t)f(y)\mathrm{d}y$ with $K>0$ everywhere for $t>0$ by Equation (10.60), and $f\geq 0$ is strictly positive on some interval. The integrand is therefore nonnegative and strictly positive on a set of positive measure. □

Remark 10.64 (What the infinite speed means). The disturbance is felt everywhere immediately, in flat contradiction with Corollary 10.56 for the wave equation. The contradiction is with relativity, not with observation: at distance x and time t the response is suppressed by $\exp(-x^2/4Dt)$, which for any laboratory value of D is unmeasurably small outside the diffusion length. The heat equation is a macroscopic limit of a kinetic description with a finite mean free path (Chapter 124), and the limit is where the infinite speed comes from.

Remark 10.65 (The spreading law). The kernel Equation (10.60) is a Gaussian of variance $2Dt$, so a point release spreads as

$$\left\langle x^2 \right\rangle = \int_{\mathbb{R}} x^2 K(x,t)\,\mathrm{d}x = 2Dt\,, \tag{10.62}$$

and the diffusion length grows as $\sqrt{t}$, not as t: a random walk, not a flight. With D in m^2/s the dimensions are those of Equation (10.62) as written. This is Einstein's relation for Brownian motion, whose molecular origin and measurement are Chapter 119. Time enters Equation (10.60) only through Dt and only for $t > 0$: the equation is not invariant under $t \to -t$, and running it backwards is the ill-posed problem of Proposition 10.25.

10.4.3 The Laplace and Poisson equations

Definition 10.66 (Harmonic function)**.** A function $u \in C^2(\Omega)$ is *harmonic* on Ω if and only if $\nabla^2 u = 0$ there; it is *subharmonic* if $\nabla^2 u \geq 0$ and *superharmonic* if $\nabla^2 u \leq 0$.

Rests on Equation (10.24) and Definition 7.66.

Proposition 10.67 (The Newtonian potential is the fundamental solution)**.** *In $\mathbb{R}^3$ the function*

$$\Phi(\boldsymbol{x}) = -\frac{1}{4\pi\,|\boldsymbol{x}|} \tag{10.63}$$

is harmonic away from the origin and satisfies $\nabla^2\Phi = \delta(\boldsymbol{x})$ in the sense of distributions.

Rests on Lemma 10.44 and Definition 17.38.

Proof. *Derives Proposition 10.67.*

For $r = |\boldsymbol{x}| > 0$ the radial part of Equation (10.33) gives $\nabla^2 r^{-1} = r^{-2}\partial_r\left(r^2\partial_r r^{-1}\right) = r^{-2}\partial_r(-1) = 0$, so Φ is harmonic there. Let φ be a test function (Definition 17.38). Then

$$\int_{\mathbb{R}^3} \Phi\,\nabla^2\varphi\,\mathrm{d}^3x = \lim_{\varepsilon\to 0^+} \int_{r>\varepsilon} \Phi\,\nabla^2\varphi\,\mathrm{d}^3x\,, \tag{10.64}$$

the integral converging absolutely because Φ is integrable near the origin. On the region $r > \varepsilon$ apply Equation (10.41) with $p = 1$, $u = \varphi$, $v = \Phi$; the term $\varphi\nabla^2\Phi$ vanishes there, and only the inner boundary contributes, with outward normal $-\boldsymbol{n}$ pointing towards the origin. The two surface terms are

$$\frac{1}{4\pi\varepsilon}\oint_{r=\varepsilon} \partial_r\varphi\,\mathrm{d}S = O(\varepsilon)\,, \qquad \frac{1}{4\pi\varepsilon^2}\oint_{r=\varepsilon} \varphi\,\mathrm{d}S \longrightarrow \varphi(\boldsymbol{0})\,, \tag{10.65}$$

the first because the sphere has area $4\pi\varepsilon^2$ and $\nabla\varphi$ is bounded, the second because the factor is exactly the reciprocal of that area and φ is continuous. Hence Equation (10.64) equals $\varphi(\boldsymbol{0})$, which by Equation (17.50) says $\nabla^2\Phi = \delta$. □

Remark 10.68. Combining Proposition 10.67 with Proposition 10.43, the solution of $\nabla^2 u = -\varrho$ on all of $\mathbb{R}^3$ decaying at infinity is $u(\boldsymbol{x}) = \frac{1}{4\pi}\int \varrho(\boldsymbol{x}')\,|\boldsymbol{x}-\boldsymbol{x}'|^{-1}\,\mathrm{d}^3x'$: the superposition of $1/r$ potentials that Laplace and Poisson wrote down [45], [57], and the electrostatic potential of a charge density once ϱ is ρ/ε_0 (Chapter 58).

Theorem 10.69 (Mean-value property)**.** *Let u be harmonic on Ω and let $\overline{B(\boldsymbol{x},r)} \subset \Omega$. Then*

$$u(\boldsymbol{x}) = \frac{1}{4\pi}\int_{S^2} u(\boldsymbol{x}+r\boldsymbol{n})\,\mathrm{d}\Omega = \frac{3}{4\pi r^3}\int_{B(\boldsymbol{x},r)} u(\boldsymbol{y})\,\mathrm{d}^3y\,. \tag{10.66}$$

Conversely a function of class C^2 with the first of these properties for every admissible ball is harmonic.

Rests on Lemma 10.58 and Definition 10.66.

Proof. *Derives Theorem 10.69.*

By Equation (10.55), $\partial_r M_u(\boldsymbol{x}, r)$ is $(4\pi r^2)^{-1}$ times the integral of $\nabla^2 u$ over the ball, which vanishes. So $M_u(\boldsymbol{x}, \cdot)$ is constant on $(0, r]$, and its limit as the radius tends to zero is $u(\boldsymbol{x})$: that is the first equality. Multiplying it by $4\pi s^2$ and integrating over $s \in [0, r]$ converts the spherical average into the solid one, since $\int_0^r 4\pi s^2 \mathrm{d}s = \frac{4}{3}\pi r^3$: that is the second. For the converse, if $\nabla^2 u(\boldsymbol{x}_0) > 0$ then by continuity $\nabla^2 u > 0$ on a small ball, and Equation (10.55) makes $\partial_r M_u > 0$ there, so M_u is not constant — contradicting the mean-value property; the case $\nabla^2 u < 0$ is symmetric. Both $\nabla^2 u(\boldsymbol{x}_0)$ and the appeal to Equation (10.55) use the hypothesis $u \in C^2$, which is why the converse is stated for that class. □

Remark 10.70 (The converse for merely continuous functions). The converse in Theorem 10.69 is in fact true for a function only assumed continuous, but not by the argument above: that argument decides the sign of a Laplacian whose existence is what is to be proved. The route taken in the appendix first establishes smoothness — averaging a continuous function with the mean-value property against a radial bump reproduces the function itself, exactly and on every compact subset, so it is automatically infinitely differentiable — and only then runs the sign argument. That smoothing step needs a mollifier, which is built there from scratch, together with the one calculus lemma it consumes; nothing is quoted, and the whole construction stays inside the Riemann integral of Chapter 7, every integrand being continuous on a compact region. Everything the chapter subsequently uses, including Theorem 10.77, needs only the C^2 form; what the continuous form adds is that the mean-value property may be taken as the *definition* of harmonicity, and with it the stability of the class under locally uniform limits.

Full derivation: Appendix A.16 (p. 2841).
Derives Theorem 10.69.

Proposition 10.71 (Green's representation formula)**.** *Let V be a bounded region with piecewise-smooth boundary and let $u \in C^2(\overline{V})$. Then for $\boldsymbol{x}$ in the interior,*

$$u(\boldsymbol{x}) = \int_V \Phi(\boldsymbol{x} - \boldsymbol{y})\, \nabla^2 u(\boldsymbol{y})\, \mathrm{d}^3 y + \oint_{\partial V} \left[u\, \partial_\nu \Phi - \Phi\, \partial_\nu u\right] \mathrm{d}S_y\,, \tag{10.67}$$

with Φ the Newtonian potential Equation (10.63) and the normal derivatives taken in $\boldsymbol{y}$.
Rests on Proposition 10.67 and Lemma 10.44.

Proof. *Derives Proposition 10.71.*

Apply Equation (10.41) with $p = 1$ on the region V punctured by a ball of radius ε about $\boldsymbol{x}$, where $\Phi(\cdot - \boldsymbol{x})$ is harmonic. The inner boundary contributes, in the limit $\varepsilon \to 0$, exactly the two terms evaluated in Equation (10.65), the first vanishing and the second producing $-u(\boldsymbol{x})$ with the orientation of the punctured region; rearranging gives Equation (10.67). □

Remark 10.72 (From the representation to the Green's function). Equation (10.67) is not yet a solution of the Dirichlet problem, because it uses both u and $\partial_\nu u$ on the boundary while the problem prescribes only the first. Adding to Φ a harmonic correction that cancels it on ∂V produces the Green's function of Definition 10.42 and removes the unwanted term — Equation (10.45) is the case where the correction can be written down. That such a correction always exists is the solvability of the Dirichlet problem, argued by Riemann from the existence of a minimiser of the energy $\int |\nabla v|^2$, criticised by Weierstrass for assuming the minimum attained, and rehabilitated by Hilbert [314]. The variational side of the story is Chapter 16, and the modern form of the argument is Section 10.6.2.

10.4.4 The Helmholtz decomposition

The Newtonian potential solves Poisson's equation one scalar at a time. Applied to each Cartesian component of a vector field it yields a structural statement about vector fields

themselves, and it is the statement most often wanted in physics: that a field splits into a part carrying all of its divergence and a part carrying all of its curl. That statement is Theorem 7.105, and this chapter is where its engine lives: each Cartesian component of the vector potential Equation (7.130) is a Newtonian potential, so $\nabla^2 \boldsymbol{w} = \boldsymbol{u}$ by Proposition 10.67, and the identity $\nabla \times (\nabla \times \boldsymbol{w}) = \nabla(\nabla \cdot \boldsymbol{w}) - \nabla^2 \boldsymbol{w}$ read backwards turns that Laplacian into a gradient plus a curl. The analysis the step needs — convergence of the singular integral, differentiation under it, and the decay that forces uniqueness — is Appendix A.45.

Remark 10.73 (What compact support is doing, and the rate that replaces it). Theorem 7.105 is stated for a field that is *smooth* and of *compact* support, and both words are load-bearing: smoothness is inherited by the potentials, which the theorem asserts to be smooth as well, and compact support is what makes the singular integral converge and lets it be differentiated under the sign. Compact support is nonetheless a convenience rather than a necessity, and Remark A.527 says exactly how much of it is needed: the same argument runs for a smooth field decaying faster than $|\boldsymbol{x}|^{-2}$ whose first derivatives decay faster than $|\boldsymbol{x}|^{-3}$, that is, for some $\varepsilon > 0$,

$$|\boldsymbol{u}(\boldsymbol{x})| + |\boldsymbol{x}| \max_{i,j} |\partial_j u_i(\boldsymbol{x})| = O\big(|\boldsymbol{x}|^{-2-\varepsilon}\big)\,, \qquad |\boldsymbol{x}| \longrightarrow \infty\,. \tag{10.68}$$

What that relaxation does *not* buy is a drop in differentiability: the proof obtains the potentials from Proposition A.521, which returns C^k potentials for a C^k field, so a merely C^2 field would give a C^1 vector potential and not the smooth one the theorem states. Some condition at infinity is indispensable, and the failure is immediate rather than technical — the constant field $\hat{\boldsymbol{z}}$ is both a gradient and a curl, so the split is not unique for it. That case is worked in the same appendix remark.

Rests on Theorem 7.105 and Proposition 10.67.

Remark 10.74 (Where the decomposition is used, and what it costs). Theorem 7.105 is the reason a wave in an elastic solid splits into one longitudinal and two transverse polarizations (Chapter 30): the irrotational part and the solenoidal part obey wave equations with different speeds, so a general disturbance separates into a pressure wave and a shear wave. It is also what lies behind the Coulomb gauge of Chapter 60, where $\boldsymbol{A}$ is chosen divergence-free, and behind the split of a fluid velocity field into a potential flow and a vortical part (Chapter 31). Some hypothesis at infinity is not decoration: without one the uniqueness fails outright, since a constant field is both a gradient and a curl, and on a bounded region the statement acquires boundary terms and a third, harmonic, summand. Both the compact support of Theorem 7.105 and the rate Equation (10.68) are convenient sufficient conditions, and both are comfortably met by the fields the physical chapters decompose, which are compactly supported or exponentially localised. Sharper hypotheses exist and are not needed here.

Rests on Theorem 7.105.

10.5 Maximum principles

10.5.1 Elliptic maximum principles

Theorem 10.75 (Weak maximum principle)**.** *Let $\Omega \subset \mathbb{R}^3$ be bounded and open and let $u \in C^2(\Omega)$ be continuous on $\overline{\Omega}$ with $\nabla^2 u \geq 0$. Then*

$$\max_{\overline{\Omega}} u = \max_{\partial\Omega} u\,. \tag{10.69}$$

The same holds for the minimum of a superharmonic function, and a harmonic function attains both its extremes on the boundary.

Rests on Definition 10.66 and Theorem 6.11.

Proof. *Derives Theorem 10.75.*

Suppose first that $\nabla^2 u > 0$ strictly. If u attained an interior maximum at $\boldsymbol{x}_0$, then each second pure derivative $\partial_i^2 u(\boldsymbol{x}_0)$ would be non-positive, since $x^i \longmapsto u$ has an interior maximum along each coordinate line; summing, $\nabla^2 u(\boldsymbol{x}_0) \le 0$, a contradiction. So the maximum is attained on the boundary, which is where it must be, both sets being compact and u continuous.

In general set $u_\varepsilon = u + \varepsilon\,|\boldsymbol{x}|^2$ for $\varepsilon > 0$. Then $\nabla^2 u_\varepsilon = \nabla^2 u + 6\varepsilon > 0$, so the strict case applies: $\max_{\overline{\Omega}} u_\varepsilon = \max_{\partial\Omega} u_\varepsilon$. Since Ω is bounded there is R with $|\boldsymbol{x}| \le R$ on $\overline{\Omega}$, and therefore

$$\max_{\overline{\Omega}} u \le \max_{\overline{\Omega}} u_\varepsilon = \max_{\partial\Omega} u_\varepsilon \le \max_{\partial\Omega} u + \varepsilon R^2 \,. \tag{10.70}$$

Letting $\varepsilon \to 0$ gives $\le$ in Equation (10.69), and the reverse inequality is trivial. The minimum statement follows on replacing u by $-u$. □

Corollary 10.76 (Uniqueness and stability for the Dirichlet problem)**.** *Let Ω be bounded and let u_1, u_2 solve $\nabla^2 u_i = f$ in Ω with boundary values g_i. Then*

$$\max_{\overline{\Omega}} |u_1 - u_2| = \max_{\partial\Omega} |g_1 - g_2| \,. \tag{10.71}$$

In particular the solution is unique, and the Dirichlet problem is well posed in the sense of Definition 10.24 with respect to the supremum norm.

Rests on Theorem 10.75 and Definition 10.24.

Proof. *Derives Corollary 10.76.*

$w = u_1 - u_2$ is harmonic, so by Equation (10.69) applied to w and to $-w$ its maximum and its minimum are attained on $\partial\Omega$; hence $\max|w|$ is attained there and equals $\max_{\partial\Omega}|g_1 - g_2|$. With $g_1 = g_2$ this gives $w \equiv 0$. □

Theorem 10.77 (Strong maximum principle for harmonic functions)**.** *Let Ω be open and connected and let u be harmonic on Ω. If u attains its supremum at some point of Ω, then u is constant.*

Rests on Theorem 10.69 and Definition 6.13.

Proof. *Derives Theorem 10.77.*

Let $M = \sup_\Omega u$ and $E = \{\boldsymbol{x} \in \Omega \mid u(\boldsymbol{x}) = M\}$, nonempty by hypothesis and closed in Ω by continuity. Let $\boldsymbol{x}_0 \in E$ and let $\overline{B(\boldsymbol{x}_0, r)} \subset \Omega$. By the solid mean-value property Equation (10.66), $M = u(\boldsymbol{x}_0)$ is the average of u over the ball, and $u \le M$ there; a continuous function no greater than its own average must equal it everywhere on the ball, since otherwise $u < M$ on a nonempty open subset and the average would be strictly less than M. So $B(\boldsymbol{x}_0, r) \subseteq E$ and E is open. Being nonempty, open and closed in the connected set Ω (Definition 6.13), $E = \Omega$. □

Remark 10.78 (Hopf's extension). Hopf proved the strong principle for general second-order elliptic operators, without the mean-value property, by a barrier construction [578]. The boundary-point lemma that usually travels with it — at a boundary point where the maximum is attained, the outward normal derivative is strictly positive unless the solution is constant — is not in that paper, whose content is the strong principle alone; Hopf published the boundary lemma a quarter of a century later [1002], and Oleinik obtained it independently the same year, in a paper this bibliography does not carry (*Matematicheskii Sbornik* 30(72) (1952) 695–702). The physical reading of both results is that a static temperature or potential field has no interior hot spot: an extreme in the interior requires a source there, and where there is no source the field is controlled entirely by its boundary values. It is also why Equation (10.66) is sometimes taken as the definition of harmonicity — legitimately, by the converse of Theorem 10.69 for twice-differentiable functions and, with the smoothing step quoted in Remark 10.70, for merely continuous ones.

10.5.2 Parabolic maximum principles

Definition 10.79 (Parabolic boundary)**.** For $\Omega \subset \mathbb{R}^3$ bounded and open and $T > 0$ put $\Omega_T = \Omega \times (0, T]$. The *parabolic boundary* is

$$\Gamma_T = \left(\overline{\Omega} \times \{0\}\right) \cup (\partial\Omega \times [0, T]) \,, \tag{10.72}$$

the initial slice together with the lateral wall. The final slice $\Omega \times \{T\}$ belongs to Ω_T and not to Γ_T.

Rests on Definitions 6.3 and 10.6.

The asymmetry is the content: a parabolic problem is posed on the parabolic boundary and predicts the final slice, never the reverse.

Theorem 10.80 (Parabolic maximum principle)**.** *Let u be continuous on $\overline{\Omega_T}$, with $\partial_t u$ and the second space derivatives continuous in Ω_T, and suppose $\partial_t u - D\nabla^2 u \le 0$ there. Then*

$$\max_{\overline{\Omega_T}} u = \max_{\Gamma_T} u \,. \tag{10.73}$$

Rests on Definition 10.79 and Theorem 10.75.

Proof. *Derives Theorem 10.80.*

Assume first the strict inequality $\partial_t u - D\nabla^2 u < 0$ in Ω_T, and suppose the maximum over the compact set $\overline{\Omega_T}$ were attained at a point $(\boldsymbol{x}_0, t_0) \in \Omega_T$, so $\boldsymbol{x}_0$ interior and $0 < t_0 \le T$. Spatially the point is an interior maximum, so $\nabla^2 u(\boldsymbol{x}_0, t_0) \le 0$ by the argument of Theorem 10.75. In time, if $t_0 < T$ then $\partial_t u = 0$ there, while if $t_0 = T$ then $\partial_t u \ge 0$, because u cannot decrease on approach to a maximum at the end of the interval. Either way $\partial_t u - D\nabla^2 u \ge 0$, contradicting the strict inequality.

For the general case put $u_\varepsilon = u - \varepsilon t$ with $\varepsilon > 0$. Then $\partial_t u_\varepsilon - D\nabla^2 u_\varepsilon = (\partial_t u - D\nabla^2 u) - \varepsilon < 0$, so $\max_{\overline{\Omega_T}} u_\varepsilon = \max_{\Gamma_T} u_\varepsilon$, whence $\max_{\overline{\Omega_T}} u \le \max_{\Gamma_T} u + \varepsilon T$. Let $\varepsilon \to 0$. □

Corollary 10.81 (Uniqueness for the heat equation)**.** *The initial-boundary-value problem $\partial_t u = D\nabla^2 u + f$ in Ω_T, with u prescribed on Γ_T, has at most one solution, and two solutions whose data differ by at most ε on Γ_T differ by at most ε everywhere.*

Rests on Theorem 10.80 and Definition 10.24.

Proof. *Derives Corollary 10.81.*

The difference w of two solutions satisfies $\partial_t w - D\nabla^2 w = 0$, so Equation (10.73) applies to w and to $-w$, giving $\max_{\overline{\Omega_T}} |w| = \max_{\Gamma_T} |w|$. □

Remark 10.82. Theorem 10.80 says that the extremes of a temperature field can only shrink: no interior point ever becomes hotter than the hottest point of the initial datum or of the wall. Together with Remark 10.37 this is the analytic face of the irreversibility of diffusion — the equation has a preferred direction of time built into its type, and the statistical account of why the world's diffusion equations run that way rather than the other is Chapter 118. The analogy should not be pressed into an identification: the maximum principle is a theorem about a parabolic PDE, while entropy growth is a statement about ensembles.

10.6 Weak solutions

10.6.1 Distributional solutions and Sobolev spaces

Every construction above has assumed a classical solution, differentiable as many times as the equation demands. Three things force that requirement to be relaxed: the fundamental solutions themselves are not twice differentiable at the source; the data of a physical problem may be discontinuous; and, as Section 10.6.3 shows, a perfectly smooth initial condition may produce a solution that ceases to be differentiable in finite time. The remedy is to move the derivatives off the solution and onto a test function.

Definition 10.83 (Weak solution)**.** Let $L = \sum_{|\alpha|\le m} a_\alpha \partial^\alpha$ have smooth coefficients and let its *formal adjoint* be

$$L^*\varphi = \sum_{|\alpha|\le m} (-1)^{|\alpha|}\, \partial^\alpha (a_\alpha \varphi)\,. \tag{10.74}$$

A locally integrable u is a *weak solution* of $Lu = f$ on Ω if and only if

$$\int_\Omega u\, L^*\varphi \,\mathrm{d}^N x = \int_\Omega f\,\varphi \,\mathrm{d}^N x \tag{10.75}$$

for every smooth φ with compact support in Ω.

Rests on Definitions 10.3 and 17.38.

Proposition 10.84 (Consistency)**.** *Every classical solution is a weak solution. Conversely a weak solution of class C^m is a classical solution.*

Rests on Definition 10.83 and Equation (7.28).

Proof. *Derives Proposition 10.84.*

For $u \in C^m$, integrating by parts $|\alpha|$ times (Equation (7.28), applied in each variable) moves each ∂^α from u onto $a_\alpha\varphi$ with a sign $(-1)^{|\alpha|}$; the boundary terms vanish because φ has compact support. Hence $\int uL^*\varphi = \int (Lu)\varphi$, and Equation (10.75) says $\int (Lu - f)\varphi = 0$ for all test φ. It remains to show that a continuous h with $\int h\varphi = 0$ for all such φ vanishes. If $h(\boldsymbol{x}_0) > 0$, continuity gives a ball on which $h > h(\boldsymbol{x}_0)/2$; choosing $\varphi \ge 0$ smooth, supported in that ball and not identically zero makes $\int h\varphi > 0$, a contradiction, and $h(\boldsymbol{x}_0) < 0$ is symmetric. This proves both directions at once. □

Definition 10.85 (Sobolev space)**.** For $\Omega \subseteq \mathbb{R}^N$ open, $H^1(\Omega)$ is the space of square-integrable u whose first-order distributional derivatives $\partial_i u$ are again square-integrable, with the norm

$$\|u\|_{H^1}^2 = \int_\Omega \left(|u|^2 + |\nabla u|^2 \right) \mathrm{d}^N x\,, \tag{10.76}$$

and $H^1_0(\Omega)$ is the closure in this norm of the smooth functions of compact support — the functions "vanishing on $\partial\Omega$" in the only sense available.

Rests on Definitions 10.83 and 12.2.

Remark 10.86. H^1 is complete, hence a Hilbert space (Chapter 12); the generalised derivatives are Sobolev's, introduced precisely to handle the Cauchy problem for hyperbolic equations [787], and the systematic theory in which they and the Dirac distribution are the same kind of object is Schwartz's [971], developed in this treatise at Section 17.2.3. In that language every Green's function of Section 10.3 is a genuine solution of its equation: Φ of Equation (10.63) has no second derivative at the origin as a function, yet Proposition 10.67 exhibits $\nabla^2\Phi$ as the Dirac distribution, and $G_\pm$ of Equation (10.46) is not a function at all. A fundamental solution is a distribution or it is nothing.

10.6.2 Variational formulation

Definition 10.87 (Weak form of an elliptic problem)**.** For p, q bounded and measurable on a bounded Ω, define on $H^1_0(\Omega)$ the bilinear form and the linear functional

$$a(u,v) = \int_\Omega \left(p\,\nabla u \cdot \nabla v + q\,u\,v \right) \mathrm{d}^N x\,, \qquad \ell(v) = \int_\Omega f\,v\,\mathrm{d}^N x\,. \tag{10.77}$$

A *weak solution* of $-\nabla\cdot(p\nabla u) + qu = f$ with vanishing boundary values is a $u \in H^1_0(\Omega)$ with $a(u,v) = \ell(v)$ for every $v \in H^1_0(\Omega)$.

Rests on Definitions 10.83 and 10.85.

The form is obtained from the equation by multiplying by v and integrating by parts once: the divergence theorem Equation (7.123) moves one derivative from u to v and produces a boundary term $\oint p\, v\, \partial_\nu u$, which vanishes because v vanishes on the boundary. Only one derivative of u now appears, which is why H^1 is the natural space.

Theorem 10.88 (Lax–Milgram)**.** *Let H be a real Hilbert space and $a : H \times H \to \mathbb{R}$ bilinear with*

$$|a(u,v)| \le C\, \|u\|\, \|v\|\,, \qquad a(u,u) \ge \alpha\, \|u\|^2\,, \tag{10.78}$$

for constants C and $\alpha > 0$, and let ℓ be a bounded linear functional on H. Then there is exactly one $u \in H$ with $a(u,v) = \ell(v)$ for every $v \in H$, and $\|u\| \le \|\ell\|\,/\alpha$.

Rests on Definitions 10.87 and 12.2.

Proof. *Derives Theorem 10.88.*

For fixed u the map $v \longmapsto a(u,v)$ is a bounded linear functional, so by the Riesz representation theorem (Chapter 12) there is a unique $Au \in H$ with $a(u,v) = \langle Au, v\rangle$ for all v; uniqueness makes A linear, and the first hypothesis of Equation (10.78) gives $\|Au\| \le C\,\|u\|$.

Coercivity gives $\alpha\,\|u\|^2 \le a(u,u) = \langle Au, u\rangle \le \|Au\|\,\|u\|$, hence

$$\|Au\| \ge \alpha\, \|u\|\,, \qquad u \in H\,. \tag{10.79}$$

So A is injective. Its range is closed: if $Au_n \to w$ then Equation (10.79) applied to $u_n - u_m$ shows (u_n) is Cauchy, so $u_n \to u$ by completeness, and $Au = w$ by continuity. The range is also dense: if $w \perp A(H)$ then in particular $\langle Aw, w\rangle = 0$, so $\alpha\,\|w\|^2 \le a(w,w) = 0$ and $w = 0$. A closed dense subspace is the whole space, so A is a bijection.

Riesz again gives F with $\ell(v) = \langle F, v\rangle$, and $u = A^{-1}F$ is the unique solution. Finally $\alpha\,\|u\|^2 \le a(u,u) = \ell(u) \le \|\ell\|\,\|u\|$. □

Corollary 10.89 (Existence for the elliptic problem)**.** *If $p \ge p_0 > 0$ and $q \ge q_0 > 0$ almost everywhere and f is square-integrable, the problem of Definition 10.87 has exactly one weak solution $u \in H^1_0(\Omega)$, and $\|u\|_{H^1} \le \|f\| \,/\, \min(p_0, q_0)$.*

Rests on Theorem 10.88 and Definition 10.87.

Proof. *Derives Corollary 10.89.*

Boundedness of a follows from the boundedness of p and q and the Cauchy–Schwarz inequality; coercivity is immediate, $a(u,u) \ge \min(p_0,q_0)\,\|u\|^2_{H^1}$ by Equation (10.76); and ℓ is bounded because $|\ell(v)| \le \|f\|\,\|v\| \le \|f\|\,\|v\|_{H^1}$. Theorem 10.88 applies with $\alpha = \min(p_0, q_0)$. □

Remark 10.90. The hypothesis $q \ge q_0 > 0$ was used only to control the term $\int u^2$ in Equation (10.76). For the Dirichlet problem proper, $q = 0$, coercivity instead follows from the Poincaré inequality — on a bounded Ω the H^1_0 norm is equivalent to $\|\nabla u\|$ alone — and the conclusion is unchanged.

Proposition 10.91 (Dirichlet principle)**.** *Let a be symmetric in addition to the hypotheses of Theorem 10.88, and set $J(v) = \tfrac12 a(v,v) - \ell(v)$. Then u is the weak solution if and only if u minimises J over H.*

Rests on Theorem 10.88 and Definition 10.87.

Proof. *Derives Proposition 10.91.*

For $v \in H$ and $t \in \mathbb{R}$, bilinearity and symmetry give

$$J(u + tv) = J(u) + t\,[a(u,v) - \ell(v)] + \frac{t^2}{2}\, a(v,v)\,. \tag{10.80}$$

If $a(u,v) = \ell(v)$ for all v, the linear term vanishes and $a(v,v) \ge \alpha\,\|v\|^2 \ge 0$, so $J(u+tv) \ge J(u)$ with equality only for $v = 0$: u is the strict minimiser. Conversely if u minimises then the derivative of Equation (10.80) at $t = 0$ vanishes for every v, which is $a(u,v) = \ell(v)$. □

Remark 10.92. The three conclusions of Theorem 10.88 are exactly Hadamard's three conditions (Definition 10.24): existence, uniqueness, and the energy estimate $\|u\| \le \|\ell\|/\alpha$, which is continuous dependence because the map $\ell \longmapsto u$ is linear. Coercivity is thus the hypothesis that carries well-posedness, and the Euler–Lagrange correspondence of Proposition 10.91 is the one developed in general in Chapter 16. The functional-analytic ingredients — completeness, the Riesz theorem, orthogonality — are Chapter 12.

10.6.3 Shocks and conservation laws

Weak solutions were introduced above to accommodate irregular data. For nonlinear equations they are unavoidable for a stronger reason: the classical solution can destroy itself.

Definition 10.93 (Scalar conservation law)**.** Given a flux $F \in C^1(\mathbb{R})$, the associated *conservation law* is

$$\partial_t u + \partial_x F(u) = 0\,, \tag{10.81}$$

equivalently $\partial_t u + F'(u)\,\partial_x u = 0$ for smooth u. The case $F(u) = \tfrac{1}{2}u^2$, that is $\partial_t u + u\,\partial_x u = 0$, is *Burgers' equation*.

Rests on Definition 10.2 and Proposition 10.15.

The name is the content: integrating Equation (10.81) over an interval shows that $\int u\,\mathrm{d}x$ changes only through the flux at the two ends, so u is the density of a conserved quantity.

Proposition 10.94 (Gradient catastrophe)**.** *Let $u_0 \in C^1(\mathbb{R})$ and let u solve Burgers' equation with $u(\cdot,0) = u_0$. Then u is constant along each straight characteristic $x = x_0 + u_0(x_0)\,t$, and*

$$\partial_x u(x_0 + u_0(x_0)t,\ t) = \frac{u_0'(x_0)}{1 + u_0'(x_0)\,t}\,. \tag{10.82}$$

If $u_0' < 0$ somewhere, no C^1 solution exists beyond

$$t_* = \frac{-1}{\min_{x_0 \in \mathbb{R}} u_0'(x_0)}\,. \tag{10.83}$$

Rests on Proposition 10.15 and Definition 10.93.

Proof. *Derives Proposition 10.94.*

By Proposition 10.15 with $a = 1$ in the time variable, $b = u$ and $c = 0$, the solution is constant along the curves $\mathrm{d}x/\mathrm{d}t = u$; being constant there, those curves have constant slope and are the stated straight lines. Hence $u\,(x_0 + u_0(x_0)t,\, t) = u_0(x_0)$. Differentiating this identity with respect to x_0,

$$\partial_x u \cdot \big(1 + u_0'(x_0)\,t\big) = u_0'(x_0)\,, \tag{10.84}$$

which is Equation (10.82). Where $u_0'(x_0) < 0$ the denominator vanishes at $t = -1/u_0'(x_0)$, and the earliest such time is Equation (10.83); as $t \to t_*^-$ the slope diverges. Equivalently, two characteristics with u_0 decreasing between them meet at that time, and the classical solution would have to take two values at once. □

Definition 10.95 (Weak solution of a conservation law)**.** A bounded measurable u on $\mathbb{R} \times [0,\infty)$ is a weak solution of Equation (10.81) with datum u_0 if and only if

$$\int_0^\infty \!\! \int_{\mathbb{R}} \big[u\,\partial_t \varphi + F(u)\,\partial_x \varphi\big]\,\mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}} u_0(x)\,\varphi(x,0)\,\mathrm{d}x = 0 \tag{10.85}$$

for every smooth φ of compact support in $\mathbb{R} \times [0,\infty)$.

Rests on Definitions 10.83 and 10.93.

Theorem 10.96 (Rankine–Hugoniot condition)**.** *Let u be a weak solution which is C^1 on either side of a C^1 curve $x = s(t)$ and has one-sided limits $u_\pm$ across it. Then at every point of the curve*

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{F(u_+) - F(u_-)}{u_+ - u_-}\,. \tag{10.86}$$

Rests on Definition 10.95 and Theorem 7.99.

Proof. *Derives Theorem 10.96.*

Let φ be supported in a small disc B meeting the curve and no other singularity, and away from $t = 0$. Split B into $B_- = B \cap \{x < s(t)\}$ and $B_+ = B \cap \{x > s(t)\}$. On B_-,

$$\int_{B_-} \left[u\,\partial_t\varphi + F(u)\partial_x\varphi\right] = \int_{B_-} \left[\partial_t(u\varphi) + \partial_x(F(u)\varphi)\right] - \int_{B_-} \varphi\left[\partial_t u + \partial_x F(u)\right]\,, \tag{10.87}$$

and the last integral vanishes because u is a classical solution there. The first is a divergence in the plane (t, x), so by the divergence form of Green's theorem (Theorem 7.99) it equals the flux of $(u\varphi,\ F(u)\varphi)$ through ∂B_-; only the part of the boundary lying on the curve contributes, since φ vanishes on ∂B. There the outward normal of B_- is proportional to $(-\dot s,\,1)$ and the arclength element is $\sqrt{1+\dot s^2}\,\mathrm{d}t$, so the contribution is $\int \left[F(u_-) - \dot s\,u_-\right]\varphi\,\mathrm{d}t$. The region B_+ gives the same expression with u_+ and the opposite sign. Adding, and using Equation (10.85),

$$\int \left\{\left[F(u_-) - F(u_+)\right] - \dot s\left[u_- - u_+\right]\right\}\varphi\,\mathrm{d}t = 0 \tag{10.88}$$

for every such φ, and the bracket, being continuous, must vanish identically; that is Equation (10.86). □

Example 10.97 (Burgers shock)**.** For $F(u) = u^2/2$, Equation (10.86) gives

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{u_+^2/2 - u_-^2/2}{u_+ - u_-} = \frac{u_- + u_+}{2}\,, \tag{10.89}$$

the shock travels at the arithmetic mean of the states it separates. Riemann obtained the same relation for finite-amplitude sound waves and identified the steepening that produces it [156]; the jump conditions across a shock in a compressible fluid are Rankine's [178] and Hugoniot's [242], and the gas-dynamical form is developed in Chapter 31.

Rests on Theorem 10.96 and Definition 10.93.

Example 10.98 (Weak solutions are not unique)**.** Take Burgers' equation with $u_0 = 0$ for $x < 0$ and $u_0 = 1$ for $x > 0$. Both of

$$u_{\mathrm{r}} = \begin{cases} 0\,, & x \le 0\,,\\ x/t\,, & 0 < x < t\,,\\ 1\,, & x \ge t\,, \end{cases} \qquad\qquad u_{\mathrm{s}} = \begin{cases} 0\,, & x < t/2\,,\\ 1\,, & x > t/2\,, \end{cases} \tag{10.90}$$

are weak solutions in the sense of Definition 10.95: the first is continuous and solves the equation classically wherever it is smooth, and the second satisfies Equation (10.89) with $u_- = 0$, $u_+ = 1$. The weak formulation alone therefore does not determine the evolution.

Rests on Theorem 10.96 and Definition 10.95.

Definition 10.99 (Entropy condition)**.** A discontinuity of a weak solution of Equation (10.81) is *admissible* if and only if

$$F'(u_-) \ge \frac{\mathrm{d}s}{\mathrm{d}t} \ge F'(u_+)\,, \tag{10.91}$$

that is, if the characteristics on both sides run *into* the discontinuity as time increases. For Burgers' equation this reads $u_- > u_+$.

Rests on Theorem 10.96 and Definition 10.95.

Remark 10.100. In Example 10.98 the jump of u_{s} has $u_- = 0 < 1 = u_+$ and is inadmissible, so the entropy condition rejects it and keeps the rarefaction u_{r}. The criterion is not an extra postulate but a limit: adding a viscous term $\varepsilon\,\partial_x^2 u$ to Equation (10.81) makes the problem parabolic and its solution unique, and as $\varepsilon \to 0^+$ the solutions converge to the weak solution satisfying Equation (10.91) and to no other [1222]. The name records that for the equations of gas dynamics the admissible jumps are exactly those across which the entropy of a fluid element does not decrease, which is why a compression shock exists in nature and a rarefaction shock does not (Chapter 31).

Definition 10.101 (Entropy pair; entropy inequality)**.** An *entropy pair* for Equation (10.81) is a pair (η, q) with η of class C^2 and convex and q of class C^1 with $q' = \eta' F'$. A weak solution u in the sense of Definition 10.95 is an *entropy solution* if and only if

$$\int_0^\infty \!\! \int_{\mathbb{R}} \left[\eta(u)\,\partial_t \varphi + q(u)\,\partial_x \varphi\right] \mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}} \eta(u_0(x))\,\varphi(x,0)\,\mathrm{d}x \geq 0 \tag{10.92}$$

for every entropy pair and every non-negative smooth φ of compact support in $\mathbb{R} \times [0,\infty)$; briefly, $\partial_t \eta(u) + \partial_x q(u) \leq 0$ in the sense of distributions.

Rests on Definitions 10.95 and 10.99.

Equation (10.91) is a statement about the one-sided limits across a curve of discontinuity, so it presupposes that those limits exist. A bounded measurable weak solution need not be piecewise C^1 and its set of discontinuities need not be a union of curves at all, so for such a solution the jump condition selects nothing; Equation (10.92) is meaningful for every weak solution and says the same thing wherever both do.

Proposition 10.102 (Jump form of the entropy inequality)**.** *Let F be convex and let u be a weak solution which is C^1 on either side of a C^1 curve $x = s(t)$, with one-sided limits $u_\pm$. Then Equation (10.92) holds for every test function supported near the curve and away from $t = 0$ if and only if*

$$q(u_-) - q(u_+) \geq \frac{\mathrm{d}s}{\mathrm{d}t}\,\left[\eta(u_-) - \eta(u_+)\right] \tag{10.93}$$

for every entropy pair; and for a strictly convex flux that is in turn equivalent to $u_- \geq u_+$, hence to Equation (10.91).

Rests on Definition 10.101 and Theorem 10.96.

Proof. *Derives Proposition 10.102.*

Off the curve u is a classical solution, so there $\partial_t \eta(u) + \partial_x q(u) = \eta'(u)\,[\partial_t u + F'(u)\partial_x u] = 0$. Repeating the computation of Theorem 10.96 with $(\eta(u), q(u))$ in place of $(u, F(u))$ therefore reduces the left-hand side of Equation (10.92) to $\int \{q(u_-) - q(u_+) - \dot{s}\,[\eta(u_-) - \eta(u_+)]\}\,\varphi\,\mathrm{d}t$, and requiring that to be non-negative for every non-negative φ is Equation (10.93).

For the second equivalence write $\sigma = \mathrm{d}s/\mathrm{d}t$ and

$$G(v) = F(v) - F(u_-) - \sigma\,(v - u_-)\ , \tag{10.94}$$

the difference between the flux and the chord joining the two states. It is convex and vanishes at $v = u_-$ and, by Equation (10.86), at $v = u_+$, so $G \leq 0$ between them. Since $q' = \eta' F'$ one has $\mathrm{d}\big[q(v) - \sigma\eta(v)\big]/\mathrm{d}v = \eta'(v)\,G'(v)$, and integrating that by parts between the two states, the boundary term vanishing with G,

$$q(u_-) - q(u_+) - \sigma\,[\eta(u_-) - \eta(u_+)] = \int_{u_-}^{u_+} \eta''(v)\,G(v)\,\mathrm{d}v\,. \tag{10.95}$$

With $\eta'' \geq 0$ and $G \leq 0$ the integrand is non-positive on the interval between the two states, so the integral in Equation (10.95), whose limits run backwards precisely when $u_- > u_+$, is

non-negative exactly in that case; and for strictly convex F one has $G < 0$ strictly between the states, so a strictly convex η makes the inequality fail whenever $u_- < u_+$. Finally F' is non-decreasing and, by Equation (10.86), σ is a mean value of F' between the two states, so Equation (10.91) is likewise equivalent to $u_- \geq u_+$. □

Theorem 10.103 (Uniqueness in the entropy class)**.** *For a convex flux F and bounded measurable datum u_0, the conservation law Equation (10.81) has exactly one entropy solution in the sense of Definition 10.101, and it depends continuously on u_0 in the mean. Where that solution is C^1 on either side of a C^1 curve of discontinuity, its jumps satisfy Equation (10.91) by Proposition 10.102.*

Rests on Definitions 10.95 and 10.101.

Full derivation: Appendix A.17 (p. 2847).

Derives Theorem 10.103.

The uniqueness half is Kruzhkov's doubling of variables, carried out there in full: the entropy inequality Equation (10.92) is first extended, by approximation from the C^2 family admitted in Definition 10.101, to the entropies $\eta(u) = |u - k|$; each of two solutions is then tested against the other, in two sets of variables at once, and the L^1 contraction is read off. Convexity of F turns out not to be needed for that half. The existence half is the vanishing-viscosity limit described above, whose entropy estimate is derived there and whose two compactness inputs are named. What the appendix does *not* build is the Lebesgue theory the subject is written in — a bounded measurable weak solution is not an object of Chapter 7 — and its section on quoted inputs says so item by item.

Remark 10.104. With Theorem 10.103 the pattern of the chapter closes on itself. A nonlinear equation loses its classical solution in finite time (Proposition 10.94); weak solutions restore existence but cost uniqueness (Example 10.98); and an admissibility criterion drawn from the physics restores uniqueness and with it Hadamard's three conditions (Definition 10.24). Which class of solutions to admit is thus not a technical preference but part of the statement of the problem, and it is settled by asking what the equation is a model of.

Chapter 11

Probability and Statistics

Probability spaces, distributions, and estimators form the mathematical backbone of the theory of measurement error developed in Chapter 2 and of the statistical mechanics of Chapter 118. The source manuscript reserves a part for this material but leaves it unwritten: it fixes only the division of the subject into a discrete and a continuous theory. That division is retained in the first two sections below, whose content — the descriptive statistics of a distribution and of a sample — is authored rather than ported, the source having left both headings empty. So are the remaining four sections: the central limit theorem, which Part I deferred to here, and then the theory of inference built on it — estimation, hypothesis testing, and the correction that a search over a range of an unknown parameter forces on any significance quoted at a point. Those last three exist because the collider search chapters of Part XI test a hypothesis and quote a significance, and a number quoted without the mathematics that defines it is not evidence.

11.1 Discrete statistics

A discrete model is one whose possible outcomes can be listed. Everything in it is a sum, and every identity below is an identity between absolutely convergent series, so nothing is needed here beyond the theory of series of Chapter 7. The continuous case of Section 11.2 replaces those sums by integrals and inherits the algebra unchanged. Together the two supply the vocabulary — mean, variance, standard deviation, moments, median, quantile, covariance, correlation — that the rest of this chapter and the error theory of Chapter 2 are written in.

11.1.1 Definitions

Definition 11.1 (Discrete probability space)**.** A *discrete probability space* is a pair $(\Omega, \Pr)$ in which the *sample space* Ω is a finite or countably infinite set, whose elements are the possible *outcomes*, and $\Pr$ assigns to each $\omega \in \Omega$ a weight $\Pr(\{\omega\}) \geq 0$ with

$$\sum_{\omega \in \Omega} \Pr(\{\omega\}) = 1 \,. \tag{11.1}$$

An *event* is a subset $A \subseteq \Omega$, and its probability is $\Pr(A) = \sum_{\omega \in A} \Pr(\{\omega\})$. A probability is a pure number and carries no SI unit.

Rests on Definition 7.45.

The series in Equation (11.1) has non-negative terms, so it converges to the same value in every order and a sum over an unordered countable set is well defined; the same remark licenses every rearrangement made below. The general setting, in which Ω need not be countable and the events form a σ-algebra, is not needed anywhere in this chapter and is set out in [1894].

Proposition 11.2 (Elementary rules)**.** *Let $A, B \subseteq \Omega$ be events and $A_1, A_2, \ldots$ pairwise disjoint events. Then*

1. $\Pr(\Omega \setminus A) = 1 - \Pr(A)$*, and* $A \subseteq B$ *implies* $\Pr(A) \leq \Pr(B)$*;*
2. $\Pr(\varnothing) = 0$*,* $\Pr(\Omega) = 1$ *and* $0 \leq \Pr(A) \leq 1$*;*
3. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$*;*
4. $\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$*.*

Rests on Definition 11.1.

Derivation. *Derives Proposition 11.2.*

Every claim is a statement about a series of non-negative terms indexed by a countable set, and such a series may be split and regrouped at will. (i) Splitting the sum Equation (11.1) over A and its complement gives the first identity; splitting the sum over B over A and $B \setminus A$ gives $\Pr(B) = \Pr(A) + \Pr(B \setminus A) \geq \Pr(A)$. (ii) The empty sum is 0 and the sum over Ω is 1 by Equation (11.1); the bounds then follow from monotonicity, $\Pr(\varnothing) \leq \Pr(A) \leq \Pr(\Omega)$. (iii) Split $A \cup B$ into $A \setminus B$, $B \setminus A$ and $A \cap B$: the left side is the sum of the three, and the right side counts $A \cap B$ twice and subtracts it once. (iv) Each outcome of the union lies in exactly one A_i, so grouping the terms of the sum over the union by the index i reproduces the right-hand side. □

Definition 11.3 (Discrete random variable)**.** A *discrete random variable* is a map $X : \Omega \to \mathbb{R}$ taking at most countably many values. Its *probability mass function* is

$$p_X(x) = \Pr(X = x) = \Pr(\{\omega \in \Omega : X(\omega) = x\}) \,, \tag{11.2}$$

a non-negative function summing to 1 over the values X attains. If the outcome carries a physical quantity — a count, a length in m, an energy in J — then X carries that SI unit and p_X does not.

Rests on Definition 11.1.

Definition 11.4 (Expectation)**.** The *expectation* of a discrete random variable is

$$\langle X \rangle = \sum_{\omega \in \Omega} X(\omega)\, \Pr(\{\omega\}) \,, \tag{11.3}$$

defined whenever $\sum_\omega |X(\omega)| \Pr(\{\omega\}) < \infty$, in which case X is called *integrable*. The expectation carries the SI unit of X.

Rests on Definitions 7.45 and 11.3.

Proposition 11.5 (Transfer, linearity, monotonicity)**.** *Let X and Y be integrable random variables on one discrete probability space and let $a, b, c \in \mathbb{R}$. Then*

1. *(transfer) for every $g : \mathbb{R} \to \mathbb{R}$ with $\sum_x |g(x)|\, p_X(x) < \infty$, $\langle g(X) \rangle = \sum_x g(x)\, p_X(x)$; in particular $\langle X \rangle = \sum_x x\, p_X(x)$;*
2. *(linearity) $aX + bY$ is integrable and $\langle aX + bY \rangle = a \langle X \rangle + b \langle Y \rangle$, and $\langle c \rangle = c$;*
3. *(monotonicity) if $X \leq Y$ at every outcome then $\langle X \rangle \leq \langle Y \rangle$, and $|\langle X \rangle| \leq \langle |X| \rangle$.*

Rests on Definition 11.4 and Proposition 11.2.

Derivation. *Derives Proposition 11.5.*

(i) Group the outcomes according to the value of X: the terms of $\sum_\omega g(X(\omega)) \Pr(\{\omega\})$ with $X(\omega) = x$ contribute $g(x) \sum_{\omega : X(\omega) = x} \Pr(\{\omega\}) = g(x) p_X(x)$, and the regrouping is legitimate because the series converges absolutely by hypothesis. (ii) Termwise, $(aX + bY)(\omega) \Pr(\{\omega\}) = aX(\omega) \Pr(\{\omega\}) + bY(\omega) \Pr(\{\omega\})$; both series on the right converge absolutely, so their sum may be split, and $|aX + bY| \leq |a|\,|X| + |b|\,|Y|$ gives the integrability. For a constant the sum is c times Equation (11.1). (iii) Termwise again, since each weight is non-negative; and $\pm X \leq |X|$ gives $\pm \langle X \rangle \leq \langle |X| \rangle$. □

Those three properties are the whole toolkit. Every identity in the rest of this section is a formal consequence of them, which is why the continuous theory of Section 11.2 can take them over wholesale.

Definition 11.6 (Moments, variance, standard deviation)**.** Let $\langle |X|^n \rangle < \infty$. The *nth raw moment* of X is $\langle X^n \rangle$ and its *nth central moment* is $\mu_n = \langle (X-\mu)^n \rangle$, where $\mu = \langle X \rangle$ is the *mean*. The *variance* and the *standard deviation* are

$$\operatorname{var} X = \mu_2 = \left\langle (X-\mu)^2 \right\rangle , \qquad \sigma = +\sqrt{\operatorname{var} X} , \tag{11.4}$$

and the *skewness* and *excess kurtosis* are $\gamma_1 = \mu_3/\sigma^3$ and $\gamma_2 = \mu_4/\sigma^4 - 3$.

Rests on Proposition 11.5.

If X carries the SI unit u then μ and σ carry u, the variance carries u^2, the nth moment carries u^n, and γ_1 and γ_2 are pure numbers — which is what makes the last two comparable between a length distribution and an energy distribution. A finite second moment implies a finite first one, since $|x| \le \tfrac{1}{2}(1+x^2)$; the same inequality is used again below to make a covariance exist. The subtracted 3 in γ_2 is a convention fixed by Proposition 11.29, which gives every Gaussian $\mu_4 = 3\sigma^4$ and hence $\gamma_2 = 0$.

Proposition 11.7 (Variance identity)**.** *If* $\langle X^2 \rangle < \infty$ *then*

$$\operatorname{var} X = \left\langle X^2 \right\rangle - \langle X \rangle^2 , \tag{11.5}$$

$\operatorname{var}(aX+b) = a^2 \operatorname{var} X$ *for constants* a *and* b*, and* $\operatorname{var} X = 0$ *if and only if* $X = \mu$ *with probability one.*

Rests on Definition 11.6 and Proposition 11.5.

Derivation. *Derives Proposition 11.7.*

Expand the square and use linearity twice: $\langle X^2 - 2\mu X + \mu^2 \rangle = \langle X^2 \rangle - 2\mu \langle X \rangle + \mu^2 = \langle X^2 \rangle - \mu^2$, since $\langle X \rangle = \mu$. For the second claim, $\langle aX+b \rangle = a\mu + b$, so $(aX+b) - \langle aX+b \rangle = a(X-\mu)$ and squaring produces the factor a^2. For the third, $\operatorname{var} X = \sum_\omega (X(\omega)-\mu)^2 \Pr(\{\omega\})$ is a series of non-negative terms, so it vanishes exactly when every term does, that is when $X(\omega) = \mu$ at every outcome of non-zero probability. □

Equation (11.5) is what allows a variance to be accumulated in one pass over the data, keeping only $\sum X_i$ and $\sum X_i^2$. It is also a trap in finite-precision arithmetic: when $\sigma \ll |\mu|$ the two terms on its right are nearly equal and their difference loses most of its significant digits, so a numerical implementation subtracts the mean first and uses Equation (11.4) directly.

Definition 11.8 (Quantiles and median)**.** For $\alpha \in (0,1)$, an *α-quantile* of X is any $q \in \mathbb{R}$ with

$$\Pr(X \le q) \ge \alpha , \qquad \Pr(X \ge q) \ge 1-\alpha . \tag{11.6}$$

A *median* is a $\tfrac{1}{2}$-quantile. A quantile carries the SI unit of X.

Rests on Definition 11.3.

Remark 11.9 (Quantiles exist and need not be unique). Write $F(x) = \Pr(X \le x)$. By countable additivity (Proposition 11.2(iv)) $F(x_n) \to F(q)$ whenever $x_n \downarrow q$ and $F(x_n) \to \Pr(X < q)$ whenever $x_n \uparrow q$; also $F \to 1$ at $+\infty$ and $F \to 0$ at $-\infty$. The set $\{x : F(x) \ge \alpha\}$ is therefore a non-empty half-line $[q, \infty)$ with q finite, and that q satisfies both requirements of Equation (11.6): the first by construction, the second because $F(x) < \alpha$ for $x < q$ forces $\Pr(X < q) \le \alpha$. Uniqueness fails in general: for X taking the values 0 and 1 with probability $\tfrac{1}{2}$ each, every point of $[0,1]$ is a median.

Proposition 11.10 (What the mean and the median minimise)**.** *Let X be integrable and let m be a median of X. Then*

$$\langle |X - m| \rangle \;\le\; \langle |X - c| \rangle \qquad \text{for every } c \in \mathbb{R}\,, \tag{11.7}$$

and if in addition $\langle X^2 \rangle < \infty$ then

$$\left\langle (X-c)^2 \right\rangle = \operatorname{var} X + (\mu - c)^2 \;\ge\; \operatorname{var} X\,, \tag{11.8}$$

with equality exactly at $c = \mu$.

Rests on Definition 11.8 and Proposition 11.7.

Derivation. *Derives Proposition 11.10.*

For Equation (11.8) write $X - c = (X - \mu) + (\mu - c)$ and expand; the cross term is $2(\mu - c)\,\langle X - \mu \rangle = 0$.

For Equation (11.7) take first $c > m$ and put $d = c - m > 0$. Then

$$|x - c| - |x - m| \;\ge\; d\left(1 - 2\,\mathbf{1}_{\{x > m\}}\right) \qquad \text{for every } x \in \mathbb{R}\,. \tag{11.9}$$

Indeed, for $x \le m < c$ both sides equal d, because $|x - c| - |x - m| = (c - x) - (m - x) = d$; and for $x > m$ the right-hand side is $-d$, while the left-hand side is at least $-\,|c - m| = -d$ by the triangle inequality. Put $x = X$ in Equation (11.9) and take expectations, using monotonicity and linearity (Proposition 11.5):

$$\langle |X - c| \rangle - \langle |X - m| \rangle \;\ge\; d\,(1 - 2\Pr(X > m)) \;\ge\; 0\,,$$

since $\Pr(X > m) = 1 - \Pr(X \le m) \le \tfrac{1}{2}$ by Equation (11.6) at $\alpha = \tfrac{1}{2}$. For $c < m$ the same argument with $d = m - c$ and the indicator of $\{x < m\}$ applies, using $\Pr(X < m) = 1 - \Pr(X \ge m) \le \tfrac{1}{2}$. □

So the mean is the least-squares measure of location and the median the least-absolute-deviation one. The difference between them is robustness: a single grossly wrong reading displaces the mean without bound, and displaces the median by at most the gap to the next value in the sample.

Definition 11.11 (Covariance and correlation)**.** For X and Y with finite second moments,

$$\operatorname{cov}(X, Y) = \langle (X - \langle X \rangle)\,(Y - \langle Y \rangle) \rangle\,, \tag{11.10}$$

which exists because $|uv| \le \tfrac{1}{2}\,(u^2 + v^2)$ makes the product integrable. When both variances are non-zero, the *correlation coefficient* is

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var} X}\,\sqrt{\operatorname{var} Y}}\,. \tag{11.11}$$

A covariance carries the product of the SI units of X and Y; a correlation coefficient is a pure number.

Rests on Definition 11.6 and Proposition 11.5.

Proposition 11.12 (Bilinearity)**.** *Covariance is symmetric and bilinear:* $\operatorname{cov}(X, X) = \operatorname{var} X$, $\operatorname{cov}(X, c) = 0$ *for a constant* c, $\operatorname{cov}(aX + bY, Z) = a \operatorname{cov}(X, Z) + b \operatorname{cov}(Y, Z)$, *and* $\operatorname{cov}(X, Y) = \langle XY \rangle - \langle X \rangle \langle Y \rangle$. *Consequently, for constants* $a_1, \dots, a_n$,

$$\operatorname{var}\left(\sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i^2 \operatorname{var} X_i + 2 \sum_{i<j} a_i a_j \operatorname{cov}(X_i, X_j)\,. \tag{11.12}$$

Rests on Definition 11.11 and Proposition 11.5.

Derivation. *Derives Proposition 11.12.*

The first two claims are Equation (11.4) and the vanishing of $c - \langle c \rangle$. For bilinearity, the centred variable of $aX + bY$ is $a\,(X - \langle X \rangle) + b\,(Y - \langle Y \rangle)$ by linearity of the expectation, and multiplying by $Z - \langle Z \rangle$ and taking expectations splits the result by linearity again; symmetry then gives linearity in the second slot. Expanding Equation (11.10) gives $\langle XY \rangle - \langle X \rangle \langle Y \rangle - \langle Y \rangle \langle X \rangle + \langle X \rangle \langle Y \rangle = \langle XY \rangle - \langle X \rangle \langle Y \rangle$. Finally Equation (11.12) is bilinearity applied to $\operatorname{cov}\left(\sum_i a_i X_i, \sum_j a_j X_j\right)$, the diagonal terms giving the variances and the off-diagonal ones pairing up by symmetry. □

Definition 11.13 (Independence)**.** Discrete random variables $X_1, \ldots, X_n$ are *independent* when

$$\Pr(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p_{X_i}(x_i) \tag{11.13}$$

for every choice of values $x_1, \ldots, x_n$. An infinite family is independent when every finite subfamily is.

Rests on Definition 11.3.

Proposition 11.14 (Independence implies zero covariance; the converse fails)**.** *If X and Y are independent with finite second moments then $\langle XY \rangle = \langle X \rangle \langle Y \rangle$ and $\operatorname{cov}(X, Y) = 0$. The converse is false.*

Rests on Definition 11.13 and Proposition 11.12.

Derivation. *Derives Proposition 11.14.*

Grouping the outcomes by the value of the pair (X, Y), exactly as in Proposition 11.5(i), gives $\langle XY \rangle = \sum_{x,y} xy \Pr(X = x, Y = y)$, and the double series converges absolutely because $\sum_{x,y} |x|\,|y|\, p_X(x) p_Y(y) = \langle |X| \rangle \langle |Y| \rangle < \infty$. Absolute convergence permits summing in either order and splitting the product of the two series (Proposition 7.49), so Equation (11.13) gives $\langle XY \rangle = \left(\sum_x x\, p_X(x)\right)\left(\sum_y y\, p_Y(y)\right) = \langle X \rangle \langle Y \rangle$, and Proposition 11.12 makes the covariance vanish. For the converse, let X take the values $-1, 0, 1$ with probability $\frac{1}{3}$ each and put $Y = X^2$. Then $\langle X \rangle = 0$ and $\langle XY \rangle = \langle X^3 \rangle = 0$, so the covariance vanishes; but $\Pr(X = 0, Y = 0) = \frac{1}{3}$ while $p_X(0)\, p_Y(0) = \frac{1}{3} \cdot \frac{1}{3}$, so Equation (11.13) fails and Y is in fact a function of X. □

Zero covariance is a statement about second moments alone, and the example shows how little it excludes. This is exactly why Proposition 11.43(iii) needs full independence rather than mere uncorrelatedness: $\mathrm{e}^{\mathrm{i}tX}$ is a nonlinear function of X.

Corollary 11.15 (Bounds on the correlation coefficient)**.** $|\rho(X, Y)| \le 1$*, with $\rho = \pm 1$ if and only if $Y = aX + b$ with probability one for constants $a \neq 0$ and b; the sign of ρ is then the sign of a.*

Rests on Definition 11.11 and Lemma 11.63.

Derivation. *Derives Corollary 11.15.*

Divide Equation (11.50) by the positive number $\operatorname{var} X \cdot \operatorname{var} Y$ and take square roots. Equality holds exactly under the condition recorded in Lemma 11.63: there are constants $(a_0, c_0) \neq (0, 0)$ with $a_0\,(X - \langle X \rangle) + c_0\,(Y - \langle Y \rangle) = 0$ with probability one. Neither coefficient can vanish here, because $a_0 = 0$ would force $\operatorname{var} Y = 0$ and $c_0 = 0$ would force $\operatorname{var} X = 0$. Hence $Y - \langle Y \rangle = k\,(X - \langle X \rangle)$ with $k = -a_0/c_0 \neq 0$, which is the stated affine relation with $b = \langle Y \rangle - k \langle X \rangle$. For such a Y, Proposition 11.12 gives $\operatorname{cov}(X, Y) = k \operatorname{var} X$ and $\operatorname{var} Y = k^2 \operatorname{var} X$, so $\rho = k/|k| = \operatorname{sgn} k$. □

Lemma 11.63 is stated and proved in Section 11.4.3, where it is needed a second time; its derivation uses nothing from this section, so borrowing it here is not circular.

Definition 11.16 (Sample mean, variance and correlation)**.** Let $X_1, \dots, X_n$ be independent with a common distribution and $n \geq 2$. The *sample mean*, *sample variance* and *sample standard deviation* are

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2\,, \qquad s_n = +\sqrt{s_n^2}\,, \tag{11.14}$$

and for pairs (X_i, Y_i) the *sample covariance* and *sample correlation coefficient* are

$$c_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)\left(Y_i - \bar{Y}_n\right)\,, \qquad r_n = \frac{c_n}{s_{X,n}\, s_{Y,n}}\,. \tag{11.15}$$

These are functions of the data alone: they are *statistics*, estimating the μ, σ^2, σ, $\operatorname{cov}(X,Y)$ and ρ of the distribution that produced them.

Rests on Definitions 11.6, 11.11 and 11.13.

Proposition 11.17 (Sampling identities)**.** *With the notation of Equation* (11.14) *and a common mean μ and variance $\sigma^2 < \infty$,*

$$\left\langle \bar{X}_n \right\rangle = \mu\,, \qquad \operatorname{var} \bar{X}_n = \frac{\sigma^2}{n}\,, \qquad \left\langle s_n^2 \right\rangle = \sigma^2\,, \tag{11.16}$$

while $\langle s_n \rangle \leq \sigma$, with equality only if s_n is constant with probability one.

Rests on Definition 11.16 and Propositions 11.12 and 11.14.

Derivation. *Derives Proposition 11.17.*

The first identity is linearity. The second is Equation (11.12) with $a_i = 1/n$: the covariances vanish by Proposition 11.14, leaving $n \cdot n^{-2}\sigma^2$. For the third, write $X_i - \bar{X}_n = (X_i - \mu) - \left(\bar{X}_n - \mu\right)$ and sum the squares; since $\sum_i (X_i - \mu) = n\left(\bar{X}_n - \mu\right)$, the cross term is $-2n\left(\bar{X}_n - \mu\right)^2$ and

$$\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n\left(\bar{X}_n - \mu\right)^2\,. \tag{11.17}$$

Taking expectations gives $n\sigma^2 - n \cdot \sigma^2/n = (n-1)\sigma^2$, and dividing by $n-1$ gives $\langle s_n^2 \rangle = \sigma^2$. Finally $0 \leq \operatorname{var} s_n = \langle s_n^2 \rangle - \langle s_n \rangle^2 = \sigma^2 - \langle s_n \rangle^2$ by Equation (11.5), so $\langle s_n \rangle \leq \sigma$, with equality exactly when $\operatorname{var} s_n$ vanishes. □

The divisor $n-1$ is the entire content of the third identity: the deviations in Equation (11.14) are taken from $\bar{X}_n$ rather than from the unknown μ, and $\bar{X}_n$ is itself pulled towards the data, which is what Equation (11.17) measures. The last clause is a warning worth stating plainly: correcting the variance does not correct the standard deviation. Whatever the distribution, s_n underestimates σ on average, and no divisor can fix that for every distribution at once, since the deficit $\sigma - \langle s_n \rangle$ depends on the sampling distribution of s_n itself.

Example 11.18 (Three standard discrete distributions)**.** *Bernoulli*: $X \in \{0,1\}$ with $\Pr(X=1) = p$. Then $\langle X \rangle = p$ and $\langle X^2 \rangle = p$, so $\operatorname{var} X = p(1-p)$ by Equation (11.5), largest at $p = \frac{1}{2}$.

Binomial: $S = \sum_{i=1}^{n} X_i$ with the X_i independent Bernoulli variables, so that $\Pr(S = k) = \binom{n}{k} p^k (1-p)^{n-k}$. Linearity gives $\langle S \rangle = np$ and Equation (11.12) gives $\operatorname{var} S = np(1-p)$, neither calculation touching the mass function.

Poisson: $p(k) = \nu^k \mathrm{e}^{-\nu}/k!$ for $k = 0, 1, 2, \dots$ and $\nu > 0$, which sums to $\mathrm{e}^{-\nu}\mathrm{e}^{\nu} = 1$. Cancelling one factor from $k/k! = 1/(k-1)!$ gives $\langle X \rangle = \nu \sum_{j \geq 0} \nu^j \mathrm{e}^{-\nu}/j! = \nu$, and the same manoeuvre twice gives $\langle X(X-1) \rangle = \nu^2$, so $\langle X^2 \rangle = \nu^2 + \nu$ and $\operatorname{var} X = \nu$. Mean and variance coincide: that is the signature of a counting process, and the reason a count of N events is quoted with an uncertainty $\sqrt{N}$.

Rests on Definition 11.6 and Propositions 11.12 and 11.14.

Example 11.19 (A counting measurement)**.** A detector is read out five times over equal live times $t_{\text{live}} = 60\,\text{s}$ and records 41, 52, 47, 44 and 46 counts. Counts are pure numbers; only the derived rate carries a unit. The sample mean is $\bar{n}_5 = 230/5 = 46.0$, the deviations are $-5, 6, 1, -2, 0$, their squares sum to 66, and Equation (11.14) gives $s_5^2 = 66/4 = 16.5$ and $s_5 = 4.06$. The standard deviation of the mean is $s_5/\sqrt{5} = 1.82$ by Equation (11.16). Sorting the readings as $41, 44, 46, 47, 52$ shows the median to be 46, close to the mean here because nothing is grossly discrepant.

Dividing by the live time turns the estimate into a rate,

$$\hat{\Gamma} = \frac{\bar{n}_5}{t_{\text{live}}} = 0.767/\text{s}\,, \qquad \frac{s_5}{\sqrt{5}\,t_{\text{live}}} = 0.030/\text{s}\,, \tag{11.18}$$

the division by a constant scaling the standard deviation by that same constant, as Proposition 11.7 requires. Had the five readings come from a Poisson law, Example 11.18 would predict $\sigma = \sqrt{\nu} \simeq 6.8$ for a single reading against the $s_5 = 4.06$ observed — a difference that five readings cannot resolve, and exactly the kind of question Section 11.5 is built to answer.

Rests on Definition 11.16, Proposition 11.17, and Example 11.18.

11.2 Continuous statistics

A continuously distributed variable takes each individual value with probability zero, so the mass function of Equation (11.2) is useless and a density takes its place. The sums of Section 11.1 become integrals, and because every identity proved there used only linearity, monotonicity and positivity of the expectation, the algebra survives the change intact. What has to be redone is only what depended on rearranging a series.

11.2.1 Definitions

Definition 11.20 (Density)**.** A real random variable X is *continuously distributed* with *probability density* f when $f \geq 0$ is Riemann integrable on every bounded interval, the improper integral

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \tag{11.19}$$

converges, and $\Pr(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$ for all $a < b$. If X carries the SI unit u, then f carries u^{-1}: only the product $f(x)\,\mathrm{d}x$ is a probability, and only a probability is a pure number.

Rests on Definition 7.39.

Definition 11.21 (Distribution function)**.** The *distribution function* of a real random variable X is $F_X(x) = \Pr(X \leq x)$, whether or not X has a density. When it has one, $F_X(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$. In general F_X is only non-decreasing, and it may jump: for a variable taking the single value c with probability one it steps from 0 to 1 at c. The proposition below is the density case, and continuity is what that case adds.

Rests on Definition 11.20.

Proposition 11.22 (Properties of the distribution function)**.** *If X has a density f, then F_X is non-decreasing and continuous, tends to 0 at $-\infty$ and to 1 at $+\infty$, satisfies $\Pr(a < X \leq b) = F_X(b) - F_X(a)$ and $\Pr(X = x) = 0$ for every x, and has $F_X'(x) = f(x)$ at every point where f is continuous.*

Rests on Definition 11.21 and Theorem 7.42.

Derivation. Derives Proposition 11.22.

Monotonicity and the difference formula follow from additivity of the integral over subintervals together with $f \geq 0$; the two limits are Equation (11.19). A Riemann integrable

function is bounded on each bounded interval, say by M near x, so $|F_X(x+h) - F_X(x)| \le M\,|h|$ and F_X is continuous — indeed locally Lipschitz. Hence $\Pr(X = x) \le F_X(x+\varepsilon) - F_X(x-\varepsilon) \to 0$. The last claim is the first fundamental theorem of calculus (Theorem 7.42), which needs f continuous only at the point in question. □

Definition 11.23 (Expectation, moments and quantiles in the continuous case)**.** For $g : \mathbb{R} \to \mathbb{R}$ with $\int |g(x)|\, f(x)\,\mathrm{d}x$ convergent,

$$\langle g(X) \rangle = \int_{-\infty}^{\infty} g(x)\, f(x)\,\mathrm{d}x\,, \tag{11.20}$$

and the mean, moments, variance, standard deviation, skewness, excess kurtosis, covariance and correlation coefficient are defined by the very formulas of Definitions 11.6 and 11.11, read with this expectation. Quantiles are defined by Equation (11.6); by Proposition 11.22 the two requirements there collapse to $F_X(q) = \alpha$.

Rests on Definitions 11.4 and 11.20.

Proposition 11.24 (The algebra of the discrete case transfers)**.** *The expectation Equation (11.20) is linear and monotone and satisfies $|\langle g(X)\rangle| \le \langle |g(X)|\rangle$. Consequently Propositions 11.7, 11.10, 11.12 and 11.17 and Corollary 11.15 hold verbatim for continuously distributed variables.*

Rests on Definition 11.23 and Proposition 11.5.

Derivation. *Derives Proposition 11.24.*

Linearity and monotonicity of the integral are the elementary properties of the Darboux construction recorded with Definition 7.39, and applying monotonicity to $\pm g f \le |g|\, f$ gives the third property. Reading the proofs of the five results named, each uses the expectation only through those three properties together with pointwise algebraic identities and the pointwise inequality Equation (11.9); none of them rearranges a series. The same holds of Lemma 11.63, whose proof needs only that $\left\langle \left(s\tilde{V} + \tilde{W} \right)^2 \right\rangle \ge 0$, and which is in any case stated for arbitrary random variables of finite variance. One dependence is not of that kind and has to be replaced rather than reread: Proposition 11.17 also uses the vanishing of the covariance of two independent variables, which here is Proposition 11.26 below and not Proposition 11.14. □

What does *not* transfer for free is the product rule for independent variables, since its discrete proof rearranged a double series. It is restated and reproved here.

Definition 11.25 (Joint density and independence)**.** A pair (X, Y) has *joint density* $f_{XY} \ge 0$ when $\Pr(a \le X \le b,\ c \le Y \le d) = \int_a^b \int_c^d f_{XY}(x,y)\,\mathrm{d}y\,\mathrm{d}x$ for all $a < b$ and $c < d$. The two variables are *independent* when $f_{XY}(x,y) = f_X(x) f_Y(y)$ for all x and y.

Rests on Definitions 11.13 and 11.20.

Proposition 11.26 (Product rule)**.** *If X and Y are independent in the sense of Definition 11.25 and both are integrable, then $\langle XY \rangle = \langle X \rangle \langle Y \rangle$ and $\operatorname{cov}(X,Y) = 0$.*

Rests on Definition 11.25 and Proposition 11.24.

Derivation. *Derives Proposition 11.26.*

In the iterated integral $\langle XY \rangle = \int \left(\int xy\, f_X(x) f_Y(y)\,\mathrm{d}y \right) \mathrm{d}x$ the inner integrand is $x f_X(x)$ times a function of y alone, so the inner integral equals $x f_X(x) \int y f_Y(y)\,\mathrm{d}y$ — a constant multiple of $x f_X(x)$, the constant being $\langle Y \rangle$, which is finite by hypothesis. Integrating in x gives $\langle X \rangle \langle Y \rangle$. No interchange of integrations is used: the factorisation is of the integrand, not of the order of integration. Proposition 11.12, which transfers by Proposition 11.24, then makes the covariance vanish. □

Proposition 11.27 (Probability integral transform)**.** *Let $F = F_X$ be strictly increasing on an interval carrying all the probability. Then $U = F(X)$ is uniformly distributed on $[0,1]$, that is $\Pr(U \le \alpha) = \alpha$ for $\alpha \in (0,1)$; conversely, if U is uniformly distributed on $[0,1]$ then $F^{-1}(U)$ has distribution function F.*

Rests on Proposition 11.22 and Definition 11.23.

Derivation. Derives Proposition 11.27.

F is continuous by Proposition 11.22 and strictly increasing by hypothesis, so it has a continuous strictly increasing inverse on $(0,1)$ and $\{F(X) \le \alpha\} = \{X \le F^{-1}(\alpha)\}$. Hence $\Pr(U \le \alpha) = F(F^{-1}(\alpha)) = \alpha$. Conversely $\Pr(F^{-1}(U) \le x) = \Pr(U \le F(x)) = F(x)$. □

The transform is what makes quantiles computable — the α-quantile is $F^{-1}(\alpha)$ — and it is the same statement, read through the survival function instead of the distribution function, as the uniformity of a p-value in Proposition 11.76.

Definition 11.28 (Gaussian distribution)**.** The *Gaussian* (or normal) distribution $\mathcal{N}(\mu, \sigma^2)$ with $\sigma > 0$ has density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) . \tag{11.21}$$

The case $\mu = 0$, $\sigma = 1$ is the *standard* Gaussian, with density φ and distribution function Φ. Here μ and σ carry the SI unit of X, f carries its reciprocal, and the exponent is a pure number.

Rests on Definition 11.20.

Proposition 11.29 (Gaussian moments)**.** *Equation (11.21) is a density; its mean is μ and its variance σ^2; its central moments are $\mu_{2n+1} = 0$ and $\mu_{2n} = (2n-1)!!\,\sigma^{2n}$; and its skewness and excess kurtosis both vanish.*

Rests on Definitions 11.23 and 11.28.

Derivation. Derives Proposition 11.29.

The substitution $u = (x-\mu)/\sigma$ (Equation (7.27)) reduces every statement to the standard case, with $\mu_n = \sigma^n \langle U^n \rangle$ for U standard. For the normalisation put $I = \int_{-\infty}^{\infty} \mathrm{e}^{-u^2/2}\,\mathrm{d}u$ and evaluate I^2 as a double integral in polar coordinates, whose Jacobian is r:

$$I^2 = \int_0^{2\pi}\int_0^{\infty} \mathrm{e}^{-r^2/2}\, r\,\mathrm{d}r\,\mathrm{d}\theta = 2\pi\,, \tag{11.22}$$

so $I = \sqrt{2\pi}$ and φ integrates to 1. This is the one step of the derivation that uses the change-of-variables formula for a double integral.

The odd moments vanish because $u^{2n+1}\varphi(u)$ is odd and absolutely integrable. For the even ones use $\varphi'(u) = -u\varphi(u)$ and integrate by parts (Equation (7.28)):

$$\left\langle U^{2n} \right\rangle = -\int_{-\infty}^{\infty} u^{2n-1}\varphi'(u)\,\mathrm{d}u = (2n-1)\int_{-\infty}^{\infty} u^{2n-2}\varphi(u)\,\mathrm{d}u = (2n-1)\left\langle U^{2n-2} \right\rangle , \tag{11.23}$$

the boundary term vanishing because $u^{2n-1}\varphi(u) \to 0$ at both ends. With $\langle U^0 \rangle = 1$ this gives $\langle U^{2n} \rangle = (2n-1)!!$, hence $\langle U^2 \rangle = 1$ and $\langle U^4 \rangle = 3$. So the mean is μ, the variance is σ^2, $\gamma_1 = 0$ and $\gamma_2 = 3 - 3 = 0$, which is the convention fixed in Definition 11.6. □

Example 11.30 (Exponential lifetimes)**.** A decay with constant λ, carried in 1/s, has density $f(t) = \lambda \mathrm{e}^{-\lambda t}$ for $t \ge 0$ and $f(t) = 0$ for $t < 0$. It integrates to 1, and integrating by parts n times (Equation (7.28)) gives $\langle T^n \rangle = n!/\lambda^n$. Hence the mean lifetime and the standard deviation are both $\tau = 1/\lambda$, in s, and

$$\operatorname{var} T = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \tau^2\,, \qquad \mu_3 = \frac{6}{\lambda^3} - \frac{6}{\lambda^3} + \frac{3}{\lambda^3} - \frac{1}{\lambda^3} = 2\tau^3\,, \qquad \gamma_1 = 2\,. \tag{11.24}$$

The distribution function is $F(t) = 1 - \mathrm{e}^{-\lambda t}$, so by Definition 11.23 the median is the half-life $t_{1/2} = \tau \ln 2 = 0.693\,\tau$, strictly less than the mean — the quantitative form of the positive skewness, and a reminder that for an asymmetric law the two measures of location of Proposition 11.10 answer different questions.

Rests on Definition 11.23 and Propositions 11.10 and 11.24.

Example 11.31 (A length measurement)**.** Six readings of a gauge block on a comparator give, in m, 0.025 001 3, 0.025 000 9, 0.025 001 5, 0.025 001 1, 0.025 001 0 and 0.025 001 4. Working in units of 1×10^{-7} m above 0.025 m, the readings are 13, 9, 15, 11, 10, 14, with sample mean 12 and deviations $1, -3, 3, -1, -2, 2$ whose squares sum to 28. By Equation (11.14),

$$\bar{\ell}_6 = 2.500\,120 \times 10^{-2}\,\mathrm{m}\,, \qquad s_6^2 = 5.6 \times 10^{-14}\,\mathrm{m}^2\,, \qquad s_6 = 2.4 \times 10^{-7}\,\mathrm{m}\,, \tag{11.25}$$

and the standard deviation of the mean is $s_6/\sqrt{6} = 9.7 \times 10^{-8}$ m. The units track the definitions exactly: a variance in m^2, a standard deviation in m, and a skewness or a correlation coefficient that would be a pure number. The ordered readings are $9, 10, 11, 13, 14, 15$, so any point of $[11, 13]$ is a median by Equation (11.6); the customary choice is the midpoint, here 12, which coincides with the mean because this sample happens to be symmetric.

Rests on Definitions 11.8 and 11.16 and Proposition 11.17.

The limit theorems that join the two theories are the subject of the next section. The binomial variable of Example 11.18 is a sum of independent identically distributed Bernoulli variables, so Theorem 11.46 applies to it verbatim: a discrete law, suitably centred and scaled, is described in the limit by the continuous Gaussian density Equation (11.21). That is the passage from the first of these two sections to the second, and it is proved rather than asserted in Section 11.3.2.

Remark 11.32 (Neither case is the general one). A distribution need be neither discrete nor continuous. The asymptotic law of the discovery statistic in Equation (11.86) is a half-and-half mixture of a point mass at zero and a chi-squared density, and is of neither kind: its distribution function jumps at the origin and grows continuously above it. The theory that covers every case at once replaces both the sum Equation (11.3) and the integral Equation (11.20) by an integral against a measure [1894]. Every identity proved in these two sections survives that generalisation unchanged, because each rests only on the linearity, monotonicity and positivity of the expectation — which is also why they could be proved once and transferred once, rather than twice.

11.2.2 Measure spaces, null sets and “almost every”

Both sections above are built on sums and on Riemann integrals, and that is enough for every distribution they treat. It is not enough for one thing the rest of the treatise says repeatedly: that a statement holds *for almost every* point, or that some exceptional set has *zero volume*. That phrase has no meaning until sets are assigned sizes, and the assignment has to survive countable operations or the phrase is useless — the whole force of “almost every” is that countably many exceptions may be discarded at once. The definitions below supply exactly that much and nothing more. They are used in Chapter 24 for the recurrence theorem, in Chapter 32 for the sets on which a Lyapunov exponent is defined, and in Chapter 12 wherever L^2 appears.

Definition 11.33 (Sigma-algebra)**.** A *σ-algebra* on a set Ω is a family $\mathcal{F}$ of subsets of Ω such that

1. $\Omega \in \mathcal{F}$;
2. $A \in \mathcal{F}$ implies $\Omega \setminus A \in \mathcal{F}$;

3. $A_1, A_2, \dots \in \mathcal{F}$ implies $\bigcup_{n\geq 1} A_n \in \mathcal{F}$.

Its members are the *measurable sets*, and $(\Omega, \mathcal{F})$ is a *measurable space*. Complements and countable unions give countable intersections as well, by de Morgan's laws (Proposition 3.37), so $\mathcal{F}$ is closed under every countable Boolean operation. The *Borel* σ-algebra of a topological space is the smallest σ-algebra containing the open sets — smallest because an arbitrary intersection of σ-algebras is one.

Rests on Definitions 6.1 and 11.1.

Definition 11.34 (Measure; measure space)**.** A *measure* on $(\Omega, \mathcal{F})$ is a map $\mu : \mathcal{F} \longrightarrow [0, \infty]$ with $\mu(\varnothing) = 0$ which is *countably additive*: for pairwise disjoint $A_1, A_2, \dots \in \mathcal{F}$,

$$\mu\Big(\bigcup_{n\geq 1} A_n\Big) = \sum_{n\geq 1} \mu(A_n)\,. \tag{11.26}$$

The triple $(\Omega, \mathcal{F}, \mu)$ is a *measure space*, and a *probability space* when $\mu(\Omega) = 1$, in which case μ is written Pr. A measure carries the SI unit of the quantity it measures — m^3 for a volume in space, the product of the units of the phase-space coordinates for the Liouville measure of Chapter 24 — and a probability is a pure number.

Rests on Definitions 11.1 and 11.33.

Definition 11.35 (Null set; almost everywhere)**.** A set $N \in \mathcal{F}$ is *null* (or *of measure zero*) if $\mu(N) = 0$. A statement holds *almost everywhere*, or for *almost every* point, if the set where it fails is contained in a null set. In a probability space one says *almost surely*.

Rests on Definition 11.34.

Proposition 11.36 (Monotonicity, countable subadditivity, and null sets)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then*

1. *$A \subseteq B$ implies $\mu(A) \leq \mu(B)$;*
2. *$\mu\left(\bigcup_n A_n\right) \leq \sum_n \mu(A_n)$ for any countable family, disjoint or not;*
3. *a countable union of null sets is null;*
4. *consequently, countably many statements each holding almost everywhere hold simultaneously almost everywhere.*

Rests on Definitions 11.34 and 11.35.

Derivation. *Derives Proposition 11.36.*

(i) Write $B = A \cup (B \setminus A)$, a disjoint union of measurable sets; Equation (11.26) with all later terms empty gives $\mu(B) = \mu(A) + \mu(B \setminus A) \geq \mu(A)$, the last step because a measure is non-negative.

(ii) Disjointify: put $B_1 = A_1$ and $B_n = A_n \setminus (A_1 \cup \dots \cup A_{n-1})$, which lie in $\mathcal{F}$ by Definition 11.33, are pairwise disjoint, have the same union as the A_n, and satisfy $B_n \subseteq A_n$. Then $\mu\left(\bigcup_n A_n\right) = \sum_n \mu(B_n) \leq \sum_n \mu(A_n)$ by Equation (11.26) and (i).

(iii) If every $\mu(A_n) = 0$ then the right-hand side of (ii) is a series of zeros, so the union has measure at most zero, hence exactly zero.

(iv) Apply (iii) to the union of the countably many exceptional sets: outside it, every one of the statements holds. □

Remark 11.37 (What is quoted here, and what is not). The definitions above are the whole of what this treatise needs from measure theory, and everything asserted about them is proved in Proposition 11.36. What is *not* proved anywhere here is that interesting measures exist: that there is a measure on the Borel sets of $\mathbb{R}^n$ assigning to a box the product of its edge lengths — Lebesgue measure — is a construction (Carathéodory's extension of a

premeasure from an algebra to the σ-algebra it generates), and it is quoted, from [1894], wherever it is used; so is the Lebesgue integral built on it, which enters this book only through the completeness of L^2 in Theorem 12.12 and through the convergence theorems listed in Remark 12.1. The distinction is worth keeping sharp. A statement of the form "the exceptional set has measure zero" needs only Definition 11.35 and Proposition 11.36 to be meaningful and to be combined with others; it needs the construction only when a particular set is to be shown null by computing something. The recurrence theorem of Chapter 24 is of the first kind, which is why it can be stated honestly here.

11.2.3 Invariant measures and the ergodic theorem

A dynamical system supplies a measure space with a map of it into itself, and the question that then arises — whether a long time average along one trajectory equals an average over the space — is the question every use of statistical mechanics and of Lyapunov exponents in this book turns on. It is a theorem of probability, not of mechanics, and it belongs here.

Definition 11.38 (Measure-preserving map; invariant measure; ergodicity)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space with $\mu(\Omega) < \infty$. A measurable map $T : \Omega \longrightarrow \Omega$ *preserves* μ, equivalently μ is *invariant* under T, if

$$\mu(T^{-1}(A)) = \mu(A)\,, \qquad \forall\, A \in \mathcal{F}\,. \tag{11.27}$$

A set A is *invariant* if $T^{-1}(A) = A$, and the system (T, μ) is *ergodic* if every invariant set is null or has null complement — that is, if the space cannot be split into two parts of positive measure that the motion never mixes. The same definitions apply to a one-parameter family $\{T_t\}_{t\in\mathbb{R}}$, a *measure-preserving flow*, with Equation (11.27) required of every T_t.

Rests on Definitions 11.34 and 11.35.

Theorem 11.39 (Birkhoff's pointwise ergodic theorem)**.** *Let T preserve a finite measure μ on $(\Omega, \mathcal{F})$ and let f be μ-integrable. Then the time average*

$$\bar{f}(\omega) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n\omega) \tag{11.28}$$

exists for μ-almost every ω, is T-invariant, and satisfies $\int \bar{f}\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$. If (T, μ) is ergodic then $\bar{f}$ is almost everywhere equal to the constant $\mu(\Omega)^{-1}\int f\,\mathrm{d}\mu$: the time average along almost every trajectory equals the space average. The same statements hold for a measure-preserving flow, with the time average $T^{-1}\int_0^T f(T_t\omega)\,\mathrm{d}t$.

Rests on Definition 11.38 and Proposition 11.36.

Remark 11.40 (What is quoted here). Theorem 11.39 is quoted, not proved. It is Birkhoff's theorem [674], and its proof rests on a maximal inequality for the partial sums together with the dominated convergence theorem of Lebesgue integration — exactly the theory Remark 11.37 declares this treatise does not develop. Von Neumann's mean ergodic theorem [705], proved a little earlier, gives convergence in L^2 rather than pointwise and is a statement about the unitary operator $f \mapsto f \circ T$ on that space; it is the weaker conclusion and the easier proof, and it is not enough for a statement about individual trajectories. Nothing in this chapter is derived from either. The theorem is stated here so that Chapter 32, which reports measured Lyapunov exponents, and Chapter 118, which replaces a time average by an ensemble average, can name the hypothesis they need — ergodicity of a specified invariant measure — instead of leaving it implicit. The multiplicative extension that governs the whole Lyapunov spectrum, rather than a single average, is Oseledets' [1378]; it is stated and used, with the same declaration, where the exponents are defined.

Remark 11.41 (Ergodicity is a hypothesis, not a consequence). Theorem 11.39 says what follows *if* the system is ergodic; it says nothing about whether any given system is. That is a question about the system, and for mechanical systems it is usually open and sometimes answered in the negative: an integrable system is never ergodic on an energy surface, because each invariant torus of Chapter 23 is an invariant set of measure zero whose union of positive-measure families fills the surface, and the surviving tori of the KAM theorem in Chapter 32 do the same for a nearly integrable one. The honest form of the statistical-mechanical postulate is therefore an assumption about the measure, not a theorem about the dynamics.

11.3 Characteristic functions and the central limit theorem

The theory of measurement error in Chapter 2 rests on one theorem: that a quantity assembled from many small independent contributions is Gaussian, whatever the contributions individually look like. That is what licenses quoting a standard uncertainty at all, and it is proved here.

The proof goes through the Fourier transform of a distribution, which turns the convolution that describes a sum of independent variables into a product. Throughout, $\langle \cdot \rangle$ denotes expectation, as in Chapter 2.

11.3.1 Characteristic functions

Definition 11.42 (Characteristic function)**.** The *characteristic function* of a real random variable X is

$$\varphi_X(t) = \left\langle \mathrm{e}^{\mathrm{i}tX} \right\rangle , \qquad t \in \mathbb{R} . \tag{11.29}$$

Unlike the moment generating function $\left\langle \mathrm{e}^{tX} \right\rangle$, this exists for every distribution without exception: $\left|\mathrm{e}^{\mathrm{i}tX}\right| = 1$, so the expectation is an integral of a bounded function and always converges. That unconditional existence is why the argument below needs no assumption beyond a finite variance.

Proposition 11.43 (Elementary properties)**.** *For any random variables X, Y and constants $a, b \in \mathbb{R}$:*

1. $\varphi_X(0) = 1$ *and* $|\varphi_X(t)| \le 1$ *for all* t*;*
2. $\varphi_{aX+b}(t) = \mathrm{e}^{\mathrm{i}bt}\, \varphi_X(at)$*;*
3. *if X and Y are independent, then* $\varphi_{X+Y} = \varphi_X\, \varphi_Y$*.*

Rests on Definition 11.42.

Derivation. *Derives Proposition 11.43.*
(i) $\varphi_X(0) = \langle 1 \rangle = 1$, and $|\varphi_X(t)| = \left|\left\langle \mathrm{e}^{\mathrm{i}tX} \right\rangle\right| \le \left\langle \left|\mathrm{e}^{\mathrm{i}tX}\right| \right\rangle = 1$. (ii) $\left\langle \mathrm{e}^{\mathrm{i}t(aX+b)} \right\rangle = \mathrm{e}^{\mathrm{i}bt} \left\langle \mathrm{e}^{\mathrm{i}(ta)X} \right\rangle$. (iii) $\mathrm{e}^{\mathrm{i}t(X+Y)} = \mathrm{e}^{\mathrm{i}tX}\mathrm{e}^{\mathrm{i}tY}$, and if X and Y are independent then $\langle f(X)g(Y)\rangle = \langle f(X)\rangle\, \langle g(Y)\rangle$ for all bounded measurable f, g — which is what independence means, and is strictly stronger than being uncorrelated: the latter is the case $f = g = \mathrm{id}$ alone and would not suffice here, since $\mathrm{e}^{\mathrm{i}tX}$ is a nonlinear function of X. Both factors are bounded, so both expectations exist. □

Property (iii) is the whole point: addition of independent variables, which is a convolution of distributions, becomes multiplication of characteristic functions — and no density need exist for that to hold.

Lemma 11.44 (Second-order expansion)**.** *If* $\langle X\rangle = 0$ *and* $\langle X^2\rangle = \sigma^2 < \infty$*, then*

$$\varphi_X(t) = 1 - \tfrac{1}{2}\sigma^2 t^2 + o(t^2) \qquad (t \to 0)\,. \tag{11.30}$$

Rests on Definition 11.42.

Derivation. *Derives Lemma 11.44.*

Writing the remainder of $\mathrm{e}^{\mathrm{i}x}$ in integral form by integrating by parts n times gives, for real x,

$$\left|\mathrm{e}^{\mathrm{i}x} - \sum_{k=0}^{n} \frac{(\mathrm{i}x)^k}{k!}\right| \le \min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2\,|x|^n}{n!}\right). \tag{11.31}$$

Bounding the integrand of that remainder gives the first estimate; the second follows instead from the triangle inequality applied to the order-$(n-1)$ remainder together with the last term of the series. The minimum matters because only the second branch survives when no moment beyond the nth exists. Taking $n = 2$ and $x = tX$,

$$\left|\mathrm{e}^{\mathrm{i}tX} - 1 - \mathrm{i}tX + \tfrac{1}{2}t^2X^2\right| \le \min\left(\tfrac{1}{6}\,|t|^3\,|X|^3\,,\; t^2X^2\right).$$

Take expectations, and use $|\langle Z\rangle| \le \langle |Z|\rangle$. Since $\langle X\rangle = 0$ the left side becomes $\left|\varphi_X(t) - 1 + \tfrac{1}{2}t^2\sigma^2\right|$, so

$$\frac{\left|\varphi_X(t) - 1 + \tfrac{1}{2}\sigma^2 t^2\right|}{t^2} \le \left\langle \min\left(\tfrac{1}{6}\,|t|\,|X|^3\,,\; X^2\right)\right\rangle.$$

The right-hand side tends to 0 as $t \to 0$. The usual reason given is dominated convergence, which this treatise cannot yet invoke: Chapter 7 develops the Riemann integral only, and the Lebesgue theory the dominated-convergence theorem belongs to is not built anywhere in the book. An elementary truncation does the same work. Fix $K > 0$ and split the expectation at $|X| = K$: where $|X| \le K$ the first branch of the minimum bounds the integrand by $\tfrac{1}{6}\,|t|\,K^3$, and where $|X| > K$ the second branch bounds it by $X^2\mathbf{1}_{\{|X|>K\}}$, so

$$\left\langle \min\left(\tfrac{1}{6}\,|t|\,|X|^3\,,\; X^2\right)\right\rangle \;\le\; \tfrac{1}{6}\,|t|\,K^3 + \left\langle X^2\,\mathbf{1}_{\{|X|>K\}}\right\rangle. \tag{11.32}$$

From Equation (11.31) (take $n = 2$ and $x = tX$, then split the expectation at $|X| = K$ and use one branch of the minimum on each piece).

Given $\varepsilon > 0$, choose K so large that the second term is below $\varepsilon/2$ — possible precisely because $\langle X^2\rangle$ is finite, a finite integral being one whose tail contribution vanishes [1894] — and then take $|t| < 3\varepsilon/K^3$, which puts the first term below $\varepsilon/2$ as well. That is exactly Equation (11.30). □

The lemma is where the finite variance enters, and it is the only place it is needed. Note what it does *not* require: no third moment, no density, no symmetry.

Theorem 11.45 (Lévy's continuity theorem)**.** *Let* $X_1, X_2, \dots$ *be random variables with characteristic functions* φ_n*. If* $\varphi_n(t) \to \varphi(t)$ *for every* $t \in \mathbb{R}$ *and* φ *is continuous at* $t = 0$*, then* φ *is the characteristic function of a random variable* X *and* $X_n \to X$ *in distribution.*

Rests on Definition 11.42.

This is the bridge from pointwise convergence of transforms back to convergence of the distributions themselves, and it is the deepest analytic statement this section rests on: the proof needs Helly's selection theorem, the Fourier inversion formula and a tightness estimate that is where continuity of the limit at the origin does its work [1446, 1894]. It is carried out in full in the appendix, so Theorem 11.45 is derived rather than quoted; what the derivation does assume, and all it assumes, are two theorems of Lebesgue theory, named there.

Full derivation: Appendix A.18 (p. 2856).

Derives Theorem 11.45.

11.3.2 The central limit theorem

Theorem 11.46 (Central limit theorem, Lindeberg–Lévy)**.** *Let $X_1, X_2, \dots$ be independent, identically distributed random variables with finite mean μ and finite variance $\sigma^2 \in (0, \infty)$. Then*

$$Z_n = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow[n\to\infty]{\text{in distribution}} \mathcal{N}(0,1)\,, \tag{11.33}$$

that is, $\Pr(Z_n \le z) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-u^2/2}\,\mathrm{d}u$ *for every* $z \in \mathbb{R}$.

Rests on Proposition 11.43, Lemma 11.44, and Theorem 11.45.

Derivation. *Derives Theorem 11.46.*

Put $Y_i = (X_i - \mu)/\sigma$, so the Y_i are independent and identically distributed with $\langle Y_i \rangle = 0$ and $\langle Y_i^2 \rangle = 1$, and $Z_n = n^{-1/2} \sum_i Y_i$. Write φ for the common characteristic function of the Y_i.

By Proposition 11.43(ii) and (iii), the latter extended to n factors by induction,

$$\varphi_{Z_n}(t) = \left[\varphi\!\left(\frac{t}{\sqrt{n}} \right) \right]^n . \tag{11.34}$$

Fix t. By Lemma 11.44 applied at argument $t/\sqrt{n}$,

$$\varphi\!\left(\frac{t}{\sqrt{n}} \right) = 1 - \frac{t^2}{2n} + r_n(t)\,, \qquad n\, r_n(t) \longrightarrow 0\,. \tag{11.35}$$

From Equation (11.30) (the second-order expansion evaluated at argument $t/\sqrt{n}$).

The remainder $r_n(t)$ is defined by that equation; the convergence is the lemma rescaled. Writing $\varphi(s) = 1 - s^2/2 + R(s)$ with $R(s)/s^2 \to 0$, we have $r_n(t) = R(t/\sqrt{n})$ and hence $n\, r_n(t) = t^2\, R(s)/s^2$ at $s = t/\sqrt{n}$, which tends to 0 for each fixed t because $s \to 0$. Now compare the two products using the elementary inequality

$$\left| \prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i \right| \le \sum_{i=1}^{n} |a_i - b_i|\,, \qquad |a_i| \le 1,\ |b_i| \le 1\,, \tag{11.36}$$

which follows by telescoping $\prod a_i - \prod b_i = \sum_i \left(\prod_{j<i} a_j \right) (a_i - b_i) \left(\prod_{j>i} b_j \right)$ and bounding each partial product by 1. Take $a_i = \varphi(t/\sqrt{n})$, whose modulus is at most 1 by Proposition 11.43(i), and $b_i = 1 - t^2/(2n)$, which lies in $[0,1]$ once $n \ge t^2/2$. Then

$$\left| \varphi_{Z_n}(t) - \left(1 - \frac{t^2}{2n} \right)^n \right| \le n\, |r_n(t)| \longrightarrow 0\,.$$

Since $(1 - t^2/(2n))^n \to \mathrm{e}^{-t^2/2}$, it follows that

$$\varphi_{Z_n}(t) \longrightarrow \mathrm{e}^{-t^2/2} \qquad \text{for every } t \in \mathbb{R}\,. \tag{11.37}$$

From Equations (11.34) to (11.36) (compare the two products factor by factor and let n grow).

The limit is the characteristic function of the standard Gaussian. That is quickest to see without contour shifting: writing $g(t) = \int \mathrm{e}^{\mathrm{i}tu} \mathrm{e}^{-u^2/2} \mathrm{d}u/\sqrt{2\pi}$ and differentiating under the integral sign, an integration by parts gives $g'(t) = -t\, g(t)$ with $g(0) = 1$, whence $g(t) = \mathrm{e}^{-t^2/2}$. It is continuous at $t = 0$. Theorem 11.45 therefore gives $Z_n \to \mathcal{N}(0,1)$ in distribution. □

Remark 11.47 (What the theorem does and does not say). Three qualifications matter for the use made of this result in Chapter 2.

First, the convergence is *in distribution* and says nothing about any individual sample. Second, it is a statement about the centre: the approximation is worst in the tails, which is

exactly where a coverage factor $k = 3$ is being asked to do work. Third — and this is the hypothesis that actually fails in practice — a finite variance is required. Errors drawn from a heavy-tailed law need not converge to anything Gaussian: the Cauchy distribution has no finite variance — indeed no finite mean — and the average of n Cauchy samples has the *same* distribution as a single one, so averaging never narrows it at all and Theorem 2.7's $\sigma/\sqrt{N}$ has nothing to describe.

Theorem 11.48 (Lindeberg–Feller)**.** *For each n let $X_{n,1}, \dots, X_{n,k_n}$ be independent with $\langle X_{n,k} \rangle = 0$ and $s_n^2 = \sum_k \operatorname{var} X_{n,k}$. If for every $\varepsilon > 0$*

$$\frac{1}{s_n^2} \sum_{k=1}^{k_n} \left\langle X_{n,k}^2 \, \mathbf{1}_{\{|X_{n,k}| > \varepsilon s_n\}} \right\rangle \longrightarrow 0 \,, \tag{11.38}$$

then $s_n^{-1} \sum_k X_{n,k} \to \mathcal{N}(0,1)$ in distribution.

Theorem 11.46 is the special case of identically distributed summands. The general form is the one the error theory actually needs, because the disturbances that perturb a measurement are *not* identically distributed: they are a thermal drift, a vibration, a quantisation step, a reading error.

Condition Equation (11.38) is *sufficient*, and the implication it does carry is the one wanted here: it forces every individual variance share to vanish. Indeed, splitting the expectation at $|X_{n,k}| = \varepsilon s_n$ gives, for any $\varepsilon > 0$,

$$\frac{\operatorname{var} X_{n,k}}{s_n^2} \le \varepsilon^2 + \frac{1}{s_n^2} \sum_j \left\langle X_{n,j}^2 \, \mathbf{1}_{\{|X_{n,j}| > \varepsilon s_n\}} \right\rangle \qquad \text{for every } \varepsilon > 0 \,, \tag{11.39}$$

so Equation (11.38) implies $\max_k \operatorname{var} X_{n,k}/s_n^2 \to 0$, which is *uniform asymptotic negligibility* — no single source carries a fixed fraction of the total variance. A dominant source therefore violates the condition.

The converse fails, and it is worth being precise about how, because the temptation is to read Equation (11.38) as a characterisation of "many small contributions". It is strictly stronger than negligibility: it constrains tail mass, not merely variance shares. Summands equal to $\pm\sqrt{n}$ with probability $1/(2n)$ each and 0 otherwise have variance share $1/n \to 0$, yet the Lindeberg sum equals 1 for every n and the limit is compound Poisson, not Gaussian. Nor does failure of the condition force a non-Gaussian limit: adjoining one $\mathcal{N}(0,n)$ summand to n independent $\mathcal{N}(0,1)$ summands breaks Equation (11.38) while leaving the normalised sum exactly $\mathcal{N}(0,1)$. Feller's companion theorem closes the loop only within the negligible class: if $\max_k \operatorname{var} X_{n,k}/s_n^2 \to 0$ *and* the normalised sum is asymptotically Gaussian, then Equation (11.38) holds [1446].

Lindeberg introduced the condition in 1922 [493]; the method of that paper — replacing the summands by Gaussians one at a time and controlling the accumulated error — also yields an elementary proof of Theorem 11.46 that avoids characteristic functions altogether. A sufficient condition that is easier to check in practice is Lyapunov's: $s_n^{-(2+\delta)} \sum_k \left\langle |X_{n,k}|^{2+\delta} \right\rangle \to 0$ for some $\delta > 0$ [1446, 1894]; it implies Equation (11.38) in a line, and that line is in the appendix. The characteristic-function argument that establishes Theorem 11.46 generalises to Theorem 11.48, with Lemma 11.44 replaced by an expansion uniform in the summand index and controlled by Equation (11.38); that generalisation is carried out there in full, and Equation (11.39) above is one of its lemmas.

Full derivation: Appendix A.19 (p. 2862).
Derives Theorem 11.48.

Remark 11.49 (How fast). The theorem is a limit, and an uncertainty budget needs a rate. The Berry–Esseen inequality supplies one: if $\rho = \left\langle |X - \mu|^3 \right\rangle < \infty$, then

$$\sup_{z \in \mathbb{R}} \left| \Pr(Z_n \le z) - \Phi(z) \right| \le \frac{C\rho}{\sigma^3 \sqrt{n}} \,, \tag{11.40}$$

with Φ the standard normal distribution function and C an absolute constant, shown finite by Berry [856] and, independently and in the same period, by Esseen. It is proved in the appendix along Esseen's route — a smoothing inequality, a bound on the characteristic function of Z_n over $|t| \lesssim \sqrt{n}\,\sigma^3/\rho$, and an integration — which delivers the explicit admissible value $C = 7.23$ [1446]. Determining the *smallest* admissible constant is a separate literature and is not attempted there; the sharpest published value for identically distributed summands is $C \leq 0.4690$, which is due to Shevtsova alone [2238].

Two things follow, and neither is comfortable. The bound is *uniform in* z: it limits the absolute error of the Gaussian approximation equally at the centre and in the tails, so where the probability being approximated is itself small the *relative* error is large — which is exactly the regime a coverage factor $k = 3$ operates in. And the $n^{-1/2}$ is slow. Since $\rho/\sigma^3 \geq 1$ always, the right-hand side of Equation (11.40) is at least $0.4690/\sqrt{n}$, which for a dozen contributions guarantees only about 0.14 in Kolmogorov distance; several hundred are needed before the inequality itself certifies a few per cent. Actual convergence is usually far better than the worst case the bound protects against — but "usually" is not a quantity an uncertainty budget can carry, and the bound is what can be asserted.

Full derivation: Appendix A.20 (p. 2867).
Derives Equation (11.40).

11.4 Estimation

The central limit theorem says what a sum of many small contributions looks like. Inference runs the other way: given the numbers a measurement actually produced, what can be said about the parameter that produced them, and how well? This section builds the likelihood function, the estimator that maximises it, the exact lower bound on how well *any* estimator can do, and the asymptotic law the maximum-likelihood estimator obeys. The last of these is what the next two sections need, and the profile likelihood constructed at the end of this one is the object a collider search actually evaluates.

Two elementary tools are used repeatedly and the earlier sections of this chapter do not yet supply them, so they are derived here.

Lemma 11.50 (Markov and Chebyshev inequalities)**.** *Let $Y \geq 0$ be a random variable and $a > 0$. Then $\Pr(Y \geq a) \leq \langle Y \rangle / a$. Consequently, if X has finite mean μ and finite variance σ^2, then for every $a > 0$*

$$\Pr(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}\,. \tag{11.41}$$

Derivation. *Derives Lemma 11.50.*

Since $Y \geq 0$, $Y \geq Y\,\mathbf{1}_{\{Y \geq a\}} \geq a\,\mathbf{1}_{\{Y \geq a\}}$ pointwise; taking expectations gives $\langle Y \rangle \geq a \Pr(Y \geq a)$. Apply this to $Y = (X - \mu)^2$ with the level a^2: the event $|X - \mu| \geq a$ is the event $Y \geq a^2$, and $\langle Y \rangle = \sigma^2$. □

Corollary 11.51 (Weak law of large numbers)**.** *If $X_1, X_2, \ldots$ are independent and identically distributed with mean μ and finite variance σ^2, then the sample mean $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$ converges to μ in probability: $\Pr(\left|\bar{X}_n - \mu\right| \geq a) \to 0$ for every $a > 0$.*

Rests on Lemma 11.50.

Derivation. *Derives Corollary 11.51.*

$\left\langle \bar{X}_n \right\rangle = \mu$ and, by independence, $\operatorname{var} \bar{X}_n = \sigma^2/n$. Then Equation (11.41) gives $\Pr(\left|\bar{X}_n - \mu\right| \geq a) \leq \sigma^2/(na^2) \to 0$. □

One further result is needed, a statement about modes of convergence rather than about any model here. It is elementary and is derived in full.

Lemma 11.52 (Slutsky)**.** *If $A_n \to A$ in distribution and $B_n \to b$ in probability with b a constant, then $A_n + B_n \to A + b$ and $A_n B_n \to bA$ in distribution. If moreover $b \neq 0$, then $A_n / B_n \to A/b$ in distribution.*

Rests on Definition 11.21.

Derivation of the sum and product clauses. *Derives Lemma 11.52.*

Write F for the distribution function of A and $F_n(y) = \Pr(A_n \le y)$, both in the general sense of Definition 11.21: no density is assumed here, and F may jump — as it does for the constant limits met below — which is exactly why the continuity points have to be watched. Convergence in distribution means $F_n(y) \to F(y)$ at every continuity point y of F. Two remarks are used repeatedly. First, F is non-decreasing, so its discontinuities are at most countable — the open intervals $(F(y^-), F(y^+))$ belonging to distinct jumps are disjoint and each contains a rational — and continuity points are therefore dense. Second, $F(y) \to 1$ as $y \to +\infty$ and $F(y) \to 0$ as $y \to -\infty$.

The sum. Let x be a continuity point of the distribution function of $A + b$, that is, let F be continuous at $x - b$, and let $\varepsilon > 0$. If $A_n + B_n \le x$ and $|B_n - b| \le \varepsilon$ then $A_n \le x - B_n \le x - b + \varepsilon$; conversely, if $A_n \le x - b - \varepsilon$ and $|B_n - b| \le \varepsilon$ then $A_n + B_n \le x$. Passing to probabilities,

$$F_n(x - b - \varepsilon) - \delta_n \;\le\; \Pr(A_n + B_n \le x) \;\le\; F_n(x - b + \varepsilon) + \delta_n \,, \tag{11.42}$$

with $\delta_n = \Pr(|B_n - b| > \varepsilon)$, which tends to 0 because $B_n \to b$ in probability. Restrict ε to the values — all but countably many, by the first remark — for which both $x - b \pm \varepsilon$ are continuity points of F. Taking the lower and upper limits in Equation (11.42) gives

$$F(x - b - \varepsilon) \;\le\; \liminf_n \Pr(A_n + B_n \le x) \;\le\; \limsup_n \Pr(A_n + B_n \le x) \;\le\; F(x - b + \varepsilon) \,,$$

and letting ε decrease to zero through admissible values makes both ends tend to $F(x - b)$, since F is continuous at $x - b$. Hence $\Pr(A_n + B_n \le x) \to F(x - b) = \Pr(A + b \le x)$, which is the claim.

Boundedness in probability. Given $\delta > 0$, choose $K > 0$ with $\pm K$ both continuity points of F, $F(K) > 1 - \delta/2$ and $F(-K) < \delta/2$; the second remark makes this possible and the first makes it possible at continuity points. Since $\Pr(|A_n| > K) \le (1 - F_n(K)) + F_n(-K)$,

$$\limsup_n \Pr(|A_n| > K) \;\le\; 1 - F(K) + F(-K) \;<\; \delta \,. \tag{11.43}$$

The product. Decompose $A_n B_n = b\, A_n + A_n\,(B_n - b)$ and treat the two pieces separately. For the second, fix $\varepsilon > 0$ and $\delta > 0$ and take K as in Equation (11.43). Since $|A_n (B_n - b)| > \varepsilon$ forces either $|A_n| > K$ or $|B_n - b| > \varepsilon / K$,

$$\Pr(|A_n\,(B_n - b)| > \varepsilon) \;\le\; \Pr(|A_n| > K) + \Pr(|B_n - b| > \varepsilon/K) \;,$$

whose second term tends to 0; so the upper limit is at most δ, and δ was arbitrary. Thus $A_n (B_n - b) \to 0$ in probability.

For the first piece, $b\, A_n \to b\, A$ in distribution. This is trivial for $b = 0$, both sides being the constant 0. For $b > 0$ the events $\{bA_n \le x\}$ and $\{A_n \le x/b\}$ coincide, and x is a continuity point of the law of bA exactly when x/b is one of F. For $b < 0$ they are $\{A_n \ge x/b\}$, so $\Pr(bA_n \le x) = 1 - \Pr(A_n < x/b)$; and at a continuity point y of F one has $F_n(y') \le \Pr(A_n < y) \le F_n(y)$ for every continuity point $y' < y$, so the limits squeeze $\Pr(A_n < y)$ between $F(y')$ and $F(y)$, and $F(y') \to F(y)$ as y' increases to y. Hence $\Pr(bA_n \le x) \to 1 - F(x/b) = \Pr(bA \le x)$ at every such x.

Applying the sum clause already proved, with bA_n in place of A_n and the constant 0 in place of b, gives $A_n B_n \to bA$ in distribution. □

The quotient is used below and is not a special case of the other two, so it is worth reducing it to them explicitly rather than leaving the reader to supply the step.

Derivation of the quotient clause from the product clause. *Derives Lemma 11.52.*

Let $b \neq 0$ and $\varepsilon > 0$. The map $y \mapsto 1/y$ is continuous at b, so there is a $\delta \in (0, |b|/2)$ such that $|y - b| < \delta$ implies both $y \neq 0$ and $|1/y - 1/b| < \varepsilon$. Hence

$$\Pr(|1/B_n - 1/b| \geq \varepsilon) \;\leq\; \Pr(|B_n - b| \geq \delta) \;\longrightarrow\; 0\,, \tag{11.44}$$

the same bound also showing that B_n is nonzero with probability tending to one, so the quotient is defined on an event of probability approaching 1. Thus $1/B_n \to 1/b$ in probability — this is the continuous-mapping theorem, in the only case this chapter needs, and its proof is the two lines just given. Applying the product clause to A_n and $1/B_n$ gives $A_n/B_n \to A/b$ in distribution. □

All three clauses are therefore derived here, and Lemma 11.52 joins the list of results this chapter proves rather than imports. Four results that are stated in this chapter are proved not here but in the appendix, because their derivations are long: Theorem 11.45, Theorem 11.48 and the Berry–Esseen inequality Equation (11.40) of Section 11.3, and Rice's formula Theorem 11.102. What remains genuinely imported is smaller and is worth naming exactly: a handful of theorems of Lebesgue theory — dominated convergence, Fubini–Tonelli — which Chapter 7 does not build and which the appendix names at each use; the weak law of large numbers in the form that assumes only integrability of the summand, which Section 11.4.4 needs wherever Corollary 11.51, resting as it does on Chebyshev's inequality, has no finite variance to work with; and, for the multi-parameter form of Theorem 11.88, the Cramér–Wold device and the continuous-mapping theorem in $\mathbb{R}^k$.

11.4.1 The likelihood function

Definition 11.53 (Statistical model)**.** A *statistical model* for data taking values in a set $\mathcal{X}$ is a family $\{f(\,\cdot\,;\theta) : \theta \in \Theta\}$ of probability densities (or, for discrete data, probability mass functions) on $\mathcal{X}$, indexed by a parameter θ ranging over a set $\Theta \subseteq \mathbb{R}^k$.

The parameter is what the experiment is trying to learn; the model is the physics, written as a statement about what data each possible value of the parameter would produce. Nothing in the definition is dimensionful, but every component of θ in a physical application carries an SI dimension — a mass in kg, a rate in 1/s, a cross-section in m^2 — and the estimators below inherit it. What is dimensionless is everything built from a *ratio* of likelihoods: the test statistics of Section 11.5, the p-values and significances read off them, and the trials factor, global p-value and global significance of Section 11.6. The dimensionful quantities in that last section are the scanned range and the resolution that fix the correction, not the correction itself.

The likelihood function, and with it most of the vocabulary of this section — parameter, statistic, likelihood, sufficiency, efficiency, and the information of Definition 11.60 — is Fisher's, introduced in a single paper [487].

Definition 11.54 (Likelihood)**.** Given observed data $x \in \mathcal{X}$, the *likelihood function* is $L(\theta) = f(x;\theta)$, regarded as a function of θ with x held fixed at its observed value. For an independent sample $x = (x_1, \ldots, x_n)$ from a common density,

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)\,, \qquad \ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta)\,. \tag{11.45}$$

Rests on Definition 11.53.

Remark 11.55 (What a likelihood is not). L is not a probability density in θ. The integral $\int_\Theta L(\theta)\,\mathrm{d}\theta$ is not 1, need not be finite, and changes if θ is reparametrised — it has no meaning. Nor is L determined by the data: multiplying f by any factor that depends on x alone leaves the model unchanged and rescales L. What survives both objections is a *ratio* of likelihoods at two parameter values, and that is why every construction below is built from ratios and never from L on its own.

Definition 11.56 (Maximum-likelihood estimator)**.** A *maximum-likelihood estimator* (MLE) is any $\hat\theta = \hat\theta(x)$ with $L(\hat\theta) = \sup_{\theta\in\Theta} L(\theta)$.
Rests on Definition 11.54.

Proposition 11.57 (Invariance)**.** *Let $g : \Theta \to \Theta^*$ be any map and define the* induced likelihood *of $\tau \in \Theta^*$ by*

$$L^*(\tau) = \sup\{L(\theta) : \theta \in \Theta,\ g(\theta) = \tau\}\ . \tag{11.46}$$

If $\hat\theta$ maximises L, then $g(\hat\theta)$ maximises L^*.
Rests on Definition 11.56.

Derivation. *Derives Proposition 11.57.*
For every τ the supremum in Equation (11.46) is taken over a subset of Θ, so $L^*(\tau) \le \sup_\Theta L = L(\hat\theta)$. On the other hand $\hat\theta$ itself lies in the fibre over $\tau = g(\hat\theta)$, so $L^*(g(\hat\theta)) \ge L(\hat\theta)$. The two inequalities force $L^*(g(\hat\theta)) = L(\hat\theta) = \max_\tau L^*(\tau)$. □

Invariance is the property that makes maximum likelihood usable in physics at all: an experiment that estimates a decay rate has thereby estimated the lifetime, with no separate calculation and no ambiguity about which parametrisation was "the right one". It is also, taken with g a projection, exactly the construction of the profile likelihood in Section 11.4.5.

Example 11.58 (Counting experiment)**.** A detector records n events in a fixed exposure, the events being independent and the expected number $\nu > 0$. Then $f(n;\nu) = \nu^n \mathrm{e}^{-\nu}/n!$, so $\ell(\nu) = n\ln\nu - \nu - \ln n!$ and $\frac{\mathrm{d}\ell}{\mathrm{d}\nu} = n/\nu - 1$, which vanishes at $\hat\nu = n$. The second derivative $-n/\nu^2$ is negative there, so $\hat\nu = n$: the estimate of the expected count is the observed count. By Proposition 11.57 the estimate of $\sqrt{\nu}$ is $\sqrt{n}$, and if the exposure is a live time t_{live} then the estimate of the rate $\Gamma = \nu/t_{\mathrm{live}}$ is n/t_{live}, carried in 1/s.
Rests on Definition 11.56 and Proposition 11.57.

11.4.2 Score and Fisher information

Everything asymptotic below needs the model to be smooth in θ and the data to be informative about it. Those requirements are made explicit once, here, and referred to by name afterwards. Let θ_0 denote the true value.

Definition 11.59 (Regularity conditions)**.** A model satisfies the *regularity conditions* at θ_0 when

(R1) *common support*: the set $\{x : f(x;\theta) > 0\}$ does not depend on θ;

(R2) *identifiability*: distinct parameter values give distinct distributions, $\theta \neq \theta' \Rightarrow f(\,\cdot\,;\theta) \neq f(\,\cdot\,;\theta')$;

(R3) *interior point*: θ_0 lies in the interior of Θ;

(R4) *differentiability under the integral*: $\theta \mapsto f(x;\theta)$ is three times continuously differentiable near θ_0, and the first two derivatives may be taken under $\int \mathrm{d}x$;

(R5) *non-degenerate information*: the quantity $I_1(\theta_0)$ of Definition 11.60 is finite and nonzero, and there are a neighbourhood of θ_0 and an integrable $M(x)$ with $|\partial^3 \ln f(x;\theta)/\partial\theta^3| \le M(x)$ throughout it.

Rests on Definitions 11.53 and 11.60.

Each of the five fails somewhere in physics, and the failures are not pathologies of the mathematics: (R1) fails for a uniform density on $[0,\theta]$, (R2) for a signal mass when the signal strength is zero, (R3) for a rate or a squared mass constrained to be non-negative and tested at zero. The last two are exactly the cases Section 11.5.4 has to treat.

Both objects of the next definition, and the word "information" attached to the second of them, are Fisher's [487].

Definition 11.60 (Score and Fisher information)**.** For a one-parameter model the *score* is

$$U(\theta;x) = \frac{\partial \ln f(x;\theta)}{\partial \theta}\,, \tag{11.47}$$

and the *Fisher information* carried by one observation is $I_1(\theta) = \langle U(\theta;X)^2 \rangle$, the expectation taken under $f(\,\cdot\,;\theta)$. For a k-parameter model the information is the $k \times k$ matrix with entries $[I_1(\theta)]_{ab} = \left\langle \frac{\partial \ln f}{\partial \theta^a} \frac{\partial \ln f}{\partial \theta^b} \right\rangle$.

Rests on Definition 11.53.

Lemma 11.61 (Score identities)**.** *Under (R1) and (R4), $\langle U(\theta;X) \rangle = 0$, so $I_1(\theta) = \operatorname{var} U(\theta;X)$, and*

$$I_1(\theta) = -\left\langle \frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2} \right\rangle . \tag{11.48}$$

For an independent sample of size n the total information is $I_n(\theta) = n\, I_1(\theta)$.

Rests on Definitions 11.59 and 11.60.

Derivation. *Derives Lemma 11.61.*

Because the support does not move with θ, the domain of the integral below is fixed, and (R4) licenses differentiating under it:

$$\langle U \rangle = \int \frac{1}{f}\frac{\partial f}{\partial \theta}\, f \,\mathrm{d}x = \frac{\partial}{\partial \theta}\int f \,\mathrm{d}x = \frac{\partial 1}{\partial \theta} = 0\,.$$

Differentiate that identity once more with respect to θ, again under the integral:

$$0 = \frac{\partial}{\partial \theta}\int \left(\frac{\partial \ln f}{\partial \theta}\right) f \,\mathrm{d}x = \int \left(\frac{\partial^2 \ln f}{\partial \theta^2}\right) f \,\mathrm{d}x + \int \left(\frac{\partial \ln f}{\partial \theta}\right)\left(\frac{\partial f}{\partial \theta}\right) \mathrm{d}x\,.$$

In the second integral write $\partial f/\partial\theta = (\partial \ln f/\partial\theta) f$; it becomes $\langle U^2 \rangle = I_1$. Rearranging gives Equation (11.48). Finally, for an independent sample $\ell = \sum_i \ln f(x_i;\theta)$, so the total score is a sum of n independent copies of U, each of mean zero; variances of independent variables add, giving $I_n = n I_1$. □

Information is additive in the data and that is the whole reason a measurement improves with running time: doubling the exposure doubles I_n and, by the bound of the next subsection, halves the smallest attainable variance.

Example 11.62 (Information in a counting experiment)**.** For the Poisson model of Example 11.58, $U(\nu;n) = n/\nu - 1$ and $\partial^2\ell/\partial\nu^2 = -n/\nu^2$, so Equation (11.48) gives $I(\nu) = \langle n \rangle / \nu^2 = 1/\nu$. The information is therefore largest where the expected count is smallest, but the conclusion to draw from that is the opposite of the obvious one, because

the bound it supplies is a bound on an *absolute* variance. By the Cramér–Rao inequality of the next subsection, Equation (11.53),

$$\operatorname{var}\hat{\nu} \;\geq\; \frac{1}{I(\nu)} = \nu\,, \qquad \text{so} \qquad \frac{\sqrt{\operatorname{var}\hat{\nu}}}{\nu} \;\geq\; \frac{1}{\sqrt{\nu}}\,. \tag{11.49}$$

A rare process is thus measured *well* in absolute terms — the standard deviation $\sqrt{\nu}$ shrinks as the process gets rarer — and *badly* in fractional terms, the fractional uncertainty $\nu^{-1/2}$ growing without bound as $\nu \to 0$. Ten expected events are known to ± 3.2 events, which is about 32 per cent; 1×10^6 expected events are known to $\pm 1 \times 10^3$ events, which is 0.1 per cent. The scale-free comparison is the information about $\ln \nu$, which by the same calculation is $I(\ln \nu) = \nu^2 I(\nu) = \nu$: it is *smallest* for a rare process, and that is the honest statement of why a rare process needs a long exposure.

Rests on Example 11.58, Lemma 11.61, and Theorem 11.64.

11.4.3 The Cramér–Rao bound

The subject of this subsection is the information inequality [870, 882], universally called the Cramér–Rao bound; Remark 11.66 records why that name and the priority disagree. It rests on one algebraic fact about second moments, which the treatise states and proves here rather than importing: Chapter 12, where a reader would look for the Cauchy–Schwarz inequality, reserves its headings and proves nothing, and Chapter 5 likewise leaves the Schwarz inequality a reserved heading with no content.

Lemma 11.63 (Cauchy–Schwarz for random variables)**.** *Let V and W be random variables of finite variance. Then*

$$\operatorname{cov}(V,W)^2 \;\leq\; \operatorname{var} V \cdot \operatorname{var} W\,, \tag{11.50}$$

with equality if and only if there are constants $(a,c) \neq (0,0)$ with $a\,(V - \langle V\rangle) + c\,(W - \langle W\rangle) = 0$ with probability one.

Rests on Definitions 11.6 and 11.11.

Derivation. *Derives Lemma 11.63.*

Write $\tilde{V} = V - \langle V\rangle$ and $\tilde{W} = W - \langle W\rangle$, both of mean zero and finite second moment. Their product is integrable, since $\left|\tilde{V}\tilde{W}\right| \leq \frac{1}{2}(\tilde{V}^2 + \tilde{W}^2)$, so $\operatorname{cov}(V,W) = \left\langle \tilde{V}\tilde{W} \right\rangle$ is a finite number and

$$\phi(s) = \left\langle \left(s\tilde{V} + \tilde{W}\right)^2 \right\rangle = s^2 \operatorname{var} V + 2s \operatorname{cov}(V,W) + \operatorname{var} W \;\geq\; 0 \qquad \text{for every } s \in \mathbb{R}\,. \tag{11.51}$$

If $\operatorname{var} V = 0$ then $\tilde{V} = 0$ almost surely, both sides of Equation (11.50) vanish, and the stated equality condition holds with $(a,c) = (1,0)$. Otherwise ϕ is a genuine quadratic that is nowhere negative, so its discriminant is non-positive: $4\operatorname{cov}(V,W)^2 - 4 \operatorname{var} V \operatorname{var} W \leq 0$, which is Equation (11.50). Equality holds exactly when the discriminant vanishes, i.e. when ϕ has a real root s_0; and $\phi(s_0) = 0$ says that the non-negative variable $\left(s_0\tilde{V} + \tilde{W}\right)^2$ has zero mean, hence vanishes almost surely, which is the stated condition with $(a,c) = (s_0, 1)$. □

Theorem 11.64 (Cramér–Rao inequality)**.** *Let the model satisfy (R1) and (R4), let the information be non-degenerate,*

$$0 < I(\theta) < \infty\,, \tag{11.52}$$

and let $T(X)$ be an estimator with finite variance and mean $\langle T\rangle = \theta + b(\theta)$, where the bias *$b$ is differentiable and $\partial/\partial\theta$ may be taken under $\int T f \,\mathrm{d}x$. Then*

$$\operatorname{var} T \;\geq\; \frac{\left(1 + b'(\theta)\right)^2}{I(\theta)}\,. \tag{11.53}$$

In particular an unbiased estimator has $\operatorname{var} T \geq 1/I(\theta)$*, which for an independent sample of size* n *is* $1/(nI_1(\theta))$*.*
Rests on Definitions 11.59 and 11.60 and Lemmas 11.61 and 11.63.

Both halves of Equation (11.52) are load-bearing and neither is cosmetic: finiteness is what lets the Cauchy–Schwarz step be applied to the score at all, since $\operatorname{var} U = I(\theta)$ must be a number, and non-vanishing is what lets Equation (11.53) divide by it. Where the regularity conditions of Definition 11.59 are in force, Equation (11.52) is the first half of their clause (R5) and nothing new is being assumed; the theorem is stated without (R2), (R3) and (R5) because it needs none of the rest of them.

Derivation. *Derives Theorem 11.64.*
By Lemma 11.61 the score has mean zero, so $\operatorname{cov}(T, U) = \langle TU \rangle - \langle T \rangle \langle U \rangle = \langle TU \rangle$. Evaluate that expectation by the same manoeuvre as before, writing $Uf = \partial f / \partial \theta$:

$$\langle TU \rangle = \int T \, \frac{\partial f}{\partial \theta} \, \mathrm{d}x = \frac{\partial}{\partial \theta} \int T f \, \mathrm{d}x = \frac{\partial}{\partial \theta} \left(\theta + b(\theta) \right) = 1 + b'(\theta) \, . \tag{11.54}$$

Now apply Lemma 11.63 to T and U, whose variances are finite by hypothesis and by Equation (11.52) respectively:

$$\left(1 + b'(\theta) \right)^2 = \operatorname{cov}(T, U)^2 \leq \operatorname{var} T \cdot \operatorname{var} U = \operatorname{var} T \cdot I(\theta) \, ,$$

which is Equation (11.53). □

Corollary 11.65 (When the bound is attained)**.** *Equality holds in Equation (11.53) for a given* θ *if and only if* $U(\theta; x) = c(\theta) \left(T(x) - \langle T \rangle \right)$ *with probability one, for some* $c(\theta)$*. If this holds for all* θ *in an interval then, integrating in* θ*,*

$$\ln f(x; \theta) = A(\theta) \, T(x) + B(\theta) + C(x) \, , \tag{11.55}$$

a one-parameter exponential family with T *as its natural sufficient statistic.*
Rests on Theorem 11.64 and Lemma 11.63.

Derivation. *Derives Corollary 11.65.*
By Lemma 11.63 equality holds exactly when the two centred variables are proportional with probability one; since $\operatorname{var} U = I(\theta) > 0$ the proportionality can be solved for the score, giving $U = c(\theta)(T - \langle T \rangle)$ almost surely. Given that for every θ, integrate $\partial \ln f / \partial \theta = c(\theta)(T(x) - \langle T \rangle_\theta)$ with respect to θ: the x-dependence enters only through the factor $T(x)$ multiplying $\int c \, \mathrm{d}\theta$, and the remaining terms are functions of θ alone plus a constant of integration $C(x)$. □

So the bound is achievable at finite sample size only inside the exponential family; the Poisson model of Example 11.62 is such a case, and indeed $\operatorname{var} \hat{\nu} = \operatorname{var} n = \nu = 1/I(\nu)$, saturating Equation (11.53) exactly. Outside that family the bound is still true and still useful, because Theorem 11.67 shows the MLE attains it asymptotically.

Remark 11.66 (Attribution: the name and the priority disagree)*.* The inequality was obtained independently, within a few years of one another, by Fréchet, Darmois, Rao and Cramér; the name that stuck records only two of the four, and neither of those two was first. The earliest published statement is Fréchet's, in 1943 [870]; Rao's independent derivation, which is the one that introduced the information-geometric reading now standard, appeared in 1945 [882]; Cramér's is later still, in his 1946 textbook. This treatise keeps the customary name "Cramér–Rao" because that is what a reader will meet everywhere else, and cites the two papers it actually has — [870, 882] — rather than the book, for which it carries no bibliography entry at all. A reader tracing the priority should start from Fréchet, not from either of the names in the title. The statement above is in any case derived in full and stands on its derivation.

11.4.4 Consistency and asymptotic normality

Theorem 11.67 (Asymptotics of the MLE)**.** *Let $X_1, \dots, X_n$ be independent and identically distributed with density $f(\,\cdot\,;\theta_0)$, and let the regularity conditions (R1)–(R5) hold at θ_0. Then there is a sequence $\hat\theta_n$ of solutions of the score equation $\ell_n'(\theta) = 0$ such that*

1. *$\hat\theta_n \to \theta_0$ in probability (*consistency*); and*
2. *$\sqrt{n}\left(\hat\theta_n - \theta_0\right) \to \mathcal{N}(0,\, I_1(\theta_0)^{-1})$ in distribution (*asymptotic normality*).*

The same holds in k parameters with $I_1(\theta_0)^{-1}$ the inverse of the information matrix.
Rests on Definitions 11.56 and 11.59, Lemmas 11.52 and 11.61, Theorem 11.46, and Corollary 11.51.

Derivation of (ii), given (i). *Derives Theorem 11.67.*

Write $\ell_n(\theta) = \sum_i \ln f(X_i;\theta)$ and expand its derivative about θ_0 with a Lagrange remainder, which (R4) permits:

$$0 = \ell_n'(\hat\theta_n) = \ell_n'(\theta_0) + \left(\hat\theta_n - \theta_0\right)\ell_n''(\theta_0) + \tfrac{1}{2}\left(\hat\theta_n - \theta_0\right)^2 \ell_n'''(\bar\theta_n)\,, \tag{11.56}$$

with $\bar\theta_n$ between θ_0 and $\hat\theta_n$. Solve for the deviation and multiply by $\sqrt{n}$:

$$\sqrt{n}\left(\hat\theta_n - \theta_0\right) = \frac{n^{-1/2}\,\ell_n'(\theta_0)}{-n^{-1}\ell_n''(\theta_0) - \tfrac{1}{2}n^{-1}\left(\hat\theta_n - \theta_0\right)\ell_n'''(\bar\theta_n)}\,. \tag{11.57}$$

Treat the three pieces separately.

The numerator is $n^{-1/2}\sum_i U(\theta_0;X_i)$, a normalised sum of independent identically distributed variables of mean zero (Lemma 11.61) and variance $I_1(\theta_0)$, finite and nonzero by (R5). Theorem 11.46 therefore gives $n^{-1/2}\ell_n'(\theta_0) \to \mathcal{N}(0, I_1(\theta_0))$ in distribution.

The first term of the denominator is $-n^{-1}\sum_i \partial^2 \ln f(X_i;\theta_0)/\partial\theta^2$, an average of independent identically distributed variables whose common mean is $I_1(\theta_0)$ by Equation (11.48) — an identity that presupposes the summand integrable, and asserts nothing about its second moment. Nothing in (R1)–(R5) bounds that second moment, so Corollary 11.51, which is the Chebyshev form and needs a finite variance, does not apply here; the weak law in the form that needs only integrability [1894] gives convergence to $I_1(\theta_0)$ in probability.

The second term of the denominator is bounded in absolute value by $\tfrac{1}{2}\left|\hat\theta_n - \theta_0\right| \cdot n^{-1}\sum_i M(X_i)$, using the domination in (R5). The average converges in probability to $\langle M\rangle < \infty$, and $\left|\hat\theta_n - \theta_0\right| \to 0$ in probability by (i); the product therefore tends to zero in probability.

The two denominator terms therefore converge in probability, by the sum clause of Lemma 11.52, to the constant $I_1(\theta_0)$, which is nonzero by (R5). Equation (11.57) is a *quotient*, not a sum or a product, so the passage to the limit is the quotient clause of Lemma 11.52 — available exactly because $I_1(\theta_0) \neq 0$, and resting on the convergence $1/B_n \to 1/I_1(\theta_0)$ in probability established at Equation (11.44). The ratio therefore converges in distribution to $\mathcal{N}(0, I_1(\theta_0))/I_1(\theta_0) = \mathcal{N}(0, I_1(\theta_0)^{-1})$, since scaling a Gaussian by $1/I_1$ scales its variance by $1/I_1^2$. □

The variance $1/(nI_1)$ is precisely the Cramér–Rao bound Equation (11.53) for an unbiased estimator: the MLE is *asymptotically efficient* in Fisher's sense [487], wasting none of the information in the data in the limit of many observations. That is the whole case for maximum likelihood, and it is an asymptotic case only — at finite n the MLE can be biased, and Example 11.58 is the exception rather than the rule in being exactly unbiased.

The consistency statement (i) has a short heart and a long technical remainder, and it is worth separating them. The heart is an inequality of Jensen's type, used nowhere else in this treatise and therefore proved here rather than quoted.

Lemma 11.68 (Jensen's inequality for the logarithm)**.** *Let $Y > 0$ almost surely with $\langle Y \rangle < \infty$. Then*

$$\langle \ln Y \rangle \;\le\; \ln \langle Y \rangle \,, \tag{11.58}$$

with equality if and only if $Y = \langle Y \rangle$ with probability one. (The left side may be $-\infty$, in which case the inequality is trivially true.)

Rests on Theorem 7.35.

Derivation. *Derives Lemma 11.68.*

Put $a = \langle Y \rangle$, which is finite and strictly positive because $Y > 0$. The logarithm has second derivative $-1/y^2 < 0$ on $(0, \infty)$, so it is strictly concave there and its graph lies below the tangent at a:

$$\ln y \;\le\; \ln a + \frac{y - a}{a} \qquad \text{for every } y > 0 \,, \tag{11.59}$$

with equality only at $y = a$. To see Equation (11.59) without appealing to a general theory of convexity, apply the mean value theorem (Theorem 7.35) to $h(y) = \ln y - \ln a - (y - a)/a$: $h(a) = 0$ and $h'(y) = 1/y - 1/a$, which is positive for $y < a$ and negative for $y > a$, so h increases up to a and decreases after it, attaining its strict maximum 0 there. Now put $y = Y$ in Equation (11.59) and take expectations, which is legitimate because the right-hand side is integrable:

$$\langle \ln Y \rangle \;\le\; \ln a + \frac{\langle Y \rangle - a}{a} = \ln a = \ln \langle Y \rangle \,.$$

Equality forces the non-positive variable $\ln Y - \ln a - (Y - a)/a$ to have zero mean, hence to vanish almost surely, which by the strictness in Equation (11.59) means $Y = a$ almost surely. □

Apply Lemma 11.68 to the likelihood ratio $Y = f(X;\theta)/f(X;\theta_0)$, which is positive almost surely because (R1) makes the support common to all θ, and whose mean is $\int f(x;\theta)\,\mathrm{d}x = 1$. For any $\theta \neq \theta_0$,

$$\left\langle \ln \frac{f(X;\theta)}{f(X;\theta_0)} \right\rangle \le \ln \left\langle \frac{f(X;\theta)}{f(X;\theta_0)} \right\rangle = \ln \int f(x;\theta)\,\mathrm{d}x = 0 \,, \tag{11.60}$$

with equality only if the ratio is constant — and, its mean being 1, only if that constant is 1, i.e. only if the two distributions coincide, which (R2) excludes. So the expected log-likelihood per observation is strictly maximised at the true value, and by Corollary 11.51 the observed $n^{-1}[\ell_n(\theta) - \ell_n(\theta_0)]$ converges to that negative number for each fixed θ. Passing from "for each fixed θ" to a statement about the *global* maximiser of the likelihood needs that convergence to be uniform on Θ, or a compactness argument in its place; that step is Wald's, under conditions weaker than the ones assumed here [950], and it is not carried out in this treatise.

Clause (i) of Theorem 11.67 claims less than that, and the less is enough for clause (ii): it asserts a consistent sequence of *solutions of the score equation*, which the expansion already in hand delivers.

Derivation of (i). *Derives Theorem 11.67.*

Take $k = 1$; the modification for k parameters is given at the end. Write $I = I_1(\theta_0)$, which is finite and strictly positive by (R5), and let $a_0 > 0$ be small enough that $[\theta_0 - a_0, \theta_0 + a_0]$ lies both inside Θ, which (R3) permits, and inside the neighbourhood on which (R5) supplies the dominating function M. For $0 < a \le a_0$ and $\theta = \theta_0 \pm a$, expand with a Lagrange remainder, as (R4) allows:

$$\frac{\ell_n(\theta) - \ell_n(\theta_0)}{n} = (\theta - \theta_0)\,\frac{\ell_n'(\theta_0)}{n} + \frac{(\theta - \theta_0)^2}{2}\,\frac{\ell_n''(\theta_0)}{n} + \frac{(\theta - \theta_0)^3}{6}\,\frac{\ell_n'''(\bar{\theta})}{n} \,, \tag{11.61}$$

with $\bar\theta$ between θ_0 and θ.

Each of the three coefficients is an average of independent identically distributed terms. The first, $n^{-1}\sum_i U(\theta_0;X_i)$, converges in probability to 0, the score having mean zero by Lemma 11.61; the second converges in probability to $-I$ by Equation (11.48); and the third is bounded in absolute value by $n^{-1}\sum_i M(X_i)$, which converges in probability to $\bar M = \langle M\rangle < \infty$. Corollary 11.51 gives the first of these, the score having variance $I_1(\theta_0)$ by (R5). It gives neither of the other two: (R5) bounds the variance of $\partial^2 \ln f/\partial\theta^2$ no more than it bounds that of M, which it assumes integrable and nothing more, and Equation (11.48) says only that the second derivative has a mean. Both are averages of integrable summands, so for both the weak law in the form that needs only integrability is used instead [1894].

Now shrink the radius, requiring in addition

$$0 < a < \frac{I}{\bar M + 1}\,, \tag{11.62}$$

and let $E_n(a)$ be the event on which all three of

$$\left|\frac{\ell_n'(\theta_0)}{n}\right| < \frac{aI}{16}\,, \qquad \left|\frac{\ell_n''(\theta_0)}{n} + I\right| < \frac{I}{4}\,, \qquad \frac{1}{n}\sum_{i=1}^{n} M(X_i) < \bar M + 1$$

hold. By the previous paragraph $\Pr(E_n(a)) \to 1$. On $E_n(a)$, and for either sign of $\theta - \theta_0 = \pm a$, the three terms of Equation (11.61) are bounded by $a^2 I/16$, by $-a^2 I/2 + a^2 I/8$ and, using Equation (11.62), by $a^3(\bar M + 1)/6 < a^2 I/6$ respectively, so that

$$\frac{\ell_n(\theta_0 \pm a) - \ell_n(\theta_0)}{n} \;<\; a^2 I\left(\frac{1}{16} - \frac{1}{2} + \frac{1}{8} + \frac{1}{6}\right) \;<\; 0\,. \tag{11.63}$$

From Equations (11.61) and (11.62) (bound the three terms on the event $E_n(a)$ and add).

So on $E_n(a)$ the log-likelihood is strictly smaller at both endpoints of $[\theta_0 - a, \theta_0 + a]$ than at its midpoint. It is continuous there, so by the extreme value theorem (Theorem 7.24) it attains a maximum on that interval; the maximum is at neither endpoint, hence at an interior point, where the derivative of a differentiable function vanishes. The score equation $\ell_n'(\theta) = 0$ therefore has a solution in $(\theta_0 - a, \theta_0 + a)$.

Define $\hat\theta_n$ to be a solution of the score equation in $[\theta_0 - a_0, \theta_0 + a_0]$ nearest to θ_0 when one exists — the solution set is closed, being the zero set of a continuous function, so a nearest point is attained — and any fixed point of Θ otherwise. For every a satisfying Equation (11.62), $E_n(a)$ implies $\left|\hat\theta_n - \theta_0\right| < a$, so $\Pr(\left|\hat\theta_n - \theta_0\right| < a) \to 1$; and since the event grows with a, the same holds for every $a > 0$. That is convergence in probability.

In k parameters, replace the interval by the closed ball of radius a about θ_0 and its endpoints by the sphere, expanding along $\theta = \theta_0 + a\boldsymbol{e}$ with $\boldsymbol{e}$ a unit vector. The linear term is bounded by a times the length of $n^{-1}\nabla\ell_n$; the quadratic term is $-\frac{1}{2}a^2\,\boldsymbol{e}^{\mathsf T} I \boldsymbol{e}$ up to the same error; and the cubic term is bounded as before, with M dominating every third partial derivative. Read in k parameters, (R5) makes I positive definite, and $\boldsymbol{e} \mapsto \boldsymbol{e}^{\mathsf T} I \boldsymbol{e}$ is continuous and strictly positive on the unit sphere, which is closed and bounded, so it has a strictly positive minimum λ (Theorem 7.24); putting λ in place of I throughout gives Equation (11.63) at every point of the sphere, and the argument closes as before with the gradient in place of the derivative. □

Remark 11.69 (In what sense the sequence solves the score equation). The sequence just constructed solves $\ell_n'(\hat\theta_n) = 0$ on the event $E_n(a_0)$, whose probability tends to one, and is defined arbitrarily on the complement. That is the standard reading of Theorem 11.67(i), and the arbitrariness is immaterial to both clauses: each is a statement about a limit in probability or in distribution, and neither is affected by altering the sequence on events whose probability tends to zero. What is *not* claimed, here or anywhere in this chapter,

is that the root so chosen is the global maximiser of the likelihood; that is the content of Wald's theorem [950], and a likelihood with several local maxima can have a consistent root far from its highest peak at any finite n.

11.4.5 Nuisance parameters and the profile likelihood

A real measurement almost never has a one-dimensional parameter. A search for a signal must simultaneously determine the size of the background, the detector efficiency and the energy scale, none of which is of interest and all of which are unknown. Write

$$\theta = (\mu, \boldsymbol{\nu}) , \qquad \mu \in \mathbb{R} , \quad \boldsymbol{\nu} \in \mathbb{R}^{m} , \tag{11.64}$$

with μ the *parameter of interest* and $\boldsymbol{\nu}$ the *nuisance parameters*.

Definition 11.70 (Profile likelihood)**.** The *profile likelihood* is

$$L_{\mathrm{p}}(\mu) = L\Big(\mu, \hat{\hat{\boldsymbol{\nu}}}(\mu)\Big) , \qquad \hat{\hat{\boldsymbol{\nu}}}(\mu) = \arg\max_{\boldsymbol{\nu}} L(\mu, \boldsymbol{\nu}) , \tag{11.65}$$

the double hat marking a maximisation carried out at fixed μ. The *profile likelihood ratio* is $\lambda(\mu) = L_{\mathrm{p}}(\mu)/L(\hat{\mu}, \hat{\boldsymbol{\nu}})$, where $(\hat{\mu}, \hat{\boldsymbol{\nu}})$ is the unconstrained MLE.
Rests on Definitions 11.54 and 11.56.

Proposition 11.71 (Elementary properties)**.** $0 < \lambda(\mu) \leq 1$ *for every* μ*, with* $\lambda(\hat{\mu}) = 1$*; and* L_{p} *is maximised at* $\hat{\mu}$*.*
Rests on Definition 11.70 and Proposition 11.57.

Derivation. *Derives Proposition 11.71.*

L_{p} is the induced likelihood Equation (11.46) for the map $g(\mu, \boldsymbol{\nu}) = \mu$, so Proposition 11.57 applies verbatim: g of the unconstrained maximiser, namely $\hat{\mu}$, maximises L_{p}, and the maximum value is $L(\hat{\mu}, \hat{\boldsymbol{\nu}})$. The ratio is therefore at most 1 and equals 1 at $\hat{\mu}$; it is positive because $L_{\mathrm{p}}(\mu) \geq L(\mu, \boldsymbol{\nu}) > 0$ for any admissible $\boldsymbol{\nu}$. □

Proposition 11.72 (Nuisance parameters cost information)**.** *Partition the information matrix of the full model as*

$$I = \begin{pmatrix} I_{\mu\mu} & I_{\mu\nu} \\ I_{\nu\mu} & I_{\nu\nu} \end{pmatrix} , \qquad I_{\nu\mu} = I_{\mu\nu}^{\mathsf{T}} , \tag{11.66}$$

with $I_{\nu\nu}$ *positive definite. Then the asymptotic variance of* $\hat{\mu}$ *is* $(n I_{\mathrm{p}})^{-1}$ *with*

$$I_{\mathrm{p}} = I_{\mu\mu} - I_{\mu\nu} I_{\nu\nu}^{-1} I_{\nu\mu} \;\leq\; I_{\mu\mu} , \tag{11.67}$$

with equality if and only if $I_{\mu\nu} = 0$*.*
Rests on Definition 11.60 and Theorem 11.67.

Derivation. *Derives Proposition 11.72.*

By Theorem 11.67 the asymptotic covariance of $\sqrt{n}(\hat{\theta} - \theta_0)$ is I^{-1}, so the asymptotic variance of $\sqrt{n}(\hat{\mu} - \mu_0)$ is the $\mu\mu$ entry of I^{-1}. To identify it, solve $I\,(x, \boldsymbol{y})^{\mathsf{T}} = (1, \boldsymbol{0})^{\mathsf{T}}$: the lower block gives $I_{\nu\mu} x + I_{\nu\nu} \boldsymbol{y} = \boldsymbol{0}$, hence $\boldsymbol{y} = -I_{\nu\nu}^{-1} I_{\nu\mu} x$, and substituting in the upper block gives $\left(I_{\mu\mu} - I_{\mu\nu} I_{\nu\nu}^{-1} I_{\nu\mu}\right) x = 1$. By construction $x = \left(I^{-1}\right)_{\mu\mu}$, so $\left(I^{-1}\right)_{\mu\mu} = I_{\mathrm{p}}^{-1}$. Finally $I_{\mu\nu} I_{\nu\nu}^{-1} I_{\nu\mu} \geq 0$ because $I_{\nu\nu}^{-1}$ is positive definite, and it vanishes exactly when $I_{\mu\nu} = 0$. □

The content of Equation (11.67) is that not knowing the background costs precision on the signal, and costs nothing only when the two are *information-orthogonal*. Designing an analysis so that $I_{\mu\nu}$ is small — a control region that fixes the background without touching the signal region — is the practical form of that statement.

Remark 11.73 (The profile is not a likelihood). L_{p} is not the likelihood of any probability model for the data: there is generally no density $h(x;\mu)$ with $h(x;\mu) = L_{\mathrm{p}}(\mu)$ for all x, because the maximising $\hat{\boldsymbol{\nu}}(\mu)$ depends on x as well as on μ. The identities of Lemma 11.61 therefore do not hold for it, and its curvature can understate the true uncertainty at finite sample size. What survives is the asymptotic statement proved in Section 11.5.4, which is the only property the search analyses of Part XI use [2172].

11.5 Hypothesis testing

Estimation asks what value a parameter has. Testing asks a coarser and often more consequential question: is a proposed value compatible with the data at all? The apparatus is small — a critical region, its size, its power — and one theorem, due to Neyman and Pearson, says which critical region to use. Everything a modern search does is a corollary of that theorem plus the asymptotic distribution derived in Section 11.5.4.

11.5.1 Tests, size, power and p-values

Definition 11.74 (Hypotheses and tests)**.** A *hypothesis* is a subset $\Theta_0 \subseteq \Theta$; it is *simple* if Θ_0 is a single point and *composite* otherwise. A *test* is a measurable function $\psi : \mathcal{X} \to [0,1]$ giving the probability with which the null hypothesis $H_0 : \theta \in \Theta_0$ is rejected when x is observed; a test is *non-randomised* when ψ is the indicator of a *critical region* $W \subseteq \mathcal{X}$. Its *size* is

$$\alpha = \sup_{\theta \in \Theta_0} \langle \psi \rangle_\theta \,, \tag{11.68}$$

and its *power* against an alternative $\theta \in \Theta_1$ is $\beta(\theta) = \langle \psi \rangle_\theta$.

Rests on Definition 11.53.

Size is the probability of rejecting a true null — the rate of false discovery claims, fixed in advance by the experimenter. Power is the probability of rejecting a false one, and is not free: it is whatever the data and the chosen critical region deliver.

Definition 11.75 (p-value)**.** Let $t(X)$ be a test statistic, large values of which disfavour a simple H_0. The *p-value* of an observation t_{obs} is

$$p = \Pr(t(X) \geq t_{\mathrm{obs}} \mid H_0) \,. \tag{11.69}$$

Rests on Definition 11.74.

Proposition 11.76 (Uniformity)**.** *If $t(X)$ has a continuous, strictly decreasing survival function $G(s) = \Pr(t \geq s \mid H_0)$ on its support, then $p = G(t(X))$ is uniformly distributed on $[0,1]$ under H_0. Consequently the test that rejects when $p \leq \alpha$ has size exactly α.*

Rests on Definitions 11.74 and 11.75.

Derivation. *Derives Proposition 11.76.*

For $\alpha \in (0,1)$, G is invertible on its range, and G is decreasing, so $\{G(t) \leq \alpha\} = \{t \geq G^{-1}(\alpha)\}$. Hence $\Pr(p \leq \alpha \mid H_0) = \Pr(t \geq G^{-1}(\alpha) \mid H_0) = G(G^{-1}(\alpha)) = \alpha$, which is the distribution function of the uniform law on $[0,1]$. □

That uniformity is what makes a p-value comparable across experiments: it is a calibrated quantity, and a threshold on it means the same thing whatever the statistic and whatever the model. It is also the source of the most persistent misreading in the subject, which is worth stating without diplomacy.

Remark 11.77 (A p-value is not the probability that the hypothesis is true). The p-value is a probability computed *assuming H_0 holds*. It is $\Pr(\text{data at least this extreme} \mid H_0)$. It is not $\Pr(H_0 \mid \text{data})$, and the two are different numbers that answer different questions. Nothing in Equation (11.69) even mentions the alternative hypothesis or the plausibility of H_0 before the data arrived, and both enter the second quantity:

$$\Pr(H_0 \mid x) = \frac{f(x;\theta_0)\,\pi_0}{f(x;\theta_0)\,\pi_0 + \int_{\Theta_1} f(x;\theta)\,\pi(\theta)\,\mathrm{d}\theta}\,, \tag{11.70}$$

with π_0 the prior probability of H_0 and π the prior density over the alternative. The gap is not small. Take an observation with $q = 25$, which Section 11.5.5 will show corresponds to $p = 2.87 \times 10^{-7}$ — a "five sigma" result.

The comparison needs the direction of one inequality stated, or it proves nothing. What Equation (11.80) supplies is the *maximised* likelihood ratio $\mathrm{e}^{q/2}$, whereas Equation (11.70) contains the likelihood *averaged* over the prior. Since π is a probability density, integrating to 1,

$$\int_{\Theta_1} f(x;\theta)\,\pi(\theta)\,\mathrm{d}\theta \;\le\; \sup_{\theta\in\Theta_1} f(x;\theta) \;=\; f(x;\hat\theta) \;=\; \mathrm{e}^{q/2} f(x;\theta_0)\,, \tag{11.71}$$

the last two equalities holding whenever the unconstrained maximiser lies in Θ_1, which is the case here because $q > 0$. So replacing the average by the maximum can only *inflate* the odds against H_0: whatever the prior, the posterior odds on the alternative are at most $\mathrm{e}^{q/2}(1-\pi_0)/\pi_0$, and $\Pr(H_0 \mid x)$ is at least $\left[1 + \mathrm{e}^{q/2}(1-\pi_0)/\pi_0\right]^{-1}$.

With $\mathrm{e}^{q/2} = 2.68 \times 10^5$ and prior odds of 1 to 1000 against the alternative, the posterior odds are therefore at most $2.68 \times 10^5 \times 1 \times 10^{-3} = 2.68 \times 10^2$, so $\Pr(H_0 \mid x) \ge 1/(1 + 2.68 \times 10^2) = 3.7 \times 10^{-3}$. The inequality runs the helpful way: the true posterior probability of the null is *at least* four orders of magnitude larger than the p-value — larger by a factor of at least 1.3×10^4 — and using a real prior density in place of its maximum can only widen the gap. The two numbers were never the same quantity.

Two corollaries follow. A small p-value is evidence against H_0 only in the presence of an alternative that predicts the observation better; with no such alternative it is a sign that something in the model is wrong, and the something is at least as likely to be the background estimate as the physics. And a p-value cannot be converted into a probability that a discovery is real without supplying a prior, which is a judgement and must be declared as one.

11.5.2 The Neyman–Pearson lemma

The lemma below is Neyman and Pearson's, from the joint paper of 1933 that created the whole framework of size, power and critical regions used above [725].

Theorem 11.78 (Neyman–Pearson). *Let H_0 and H_1 be simple, with densities f_0 and f_1. Fix $k \ge 0$ and let*

$$\psi^*(x) = \begin{cases} 1 & f_1(x) > k f_0(x)\,, \\ \gamma(x) & f_1(x) = k f_0(x)\,, \\ 0 & f_1(x) < k f_0(x)\,, \end{cases} \tag{11.72}$$

for any γ taking values in $[0,1]$, and write $\alpha = \langle\psi^*\rangle_0$ for its size. Then every test ψ with $\langle\psi\rangle_0 \le \alpha$ satisfies $\langle\psi\rangle_1 \le \langle\psi^*\rangle_1$: no test of size at most α is more powerful.

Rests on Definition 11.74.

Derivation. *Derives Theorem 11.78.*

Consider the function

$$D(x) = \left(\psi^*(x) - \psi(x)\right)\left(f_1(x) - k f_0(x)\right)\,.$$

It is non-negative everywhere. Where $f_1 > k f_0$ the second factor is positive and $\psi^* = 1 \geq \psi$, so the first is non-negative; where $f_1 < k f_0$ the second factor is negative and $\psi^* = 0 \leq \psi$, so the first is non-positive; where $f_1 = k f_0$ the second factor vanishes. Integrating,

$$0 \leq \int D \, \mathrm{d}x = \left(\langle \psi^* \rangle_1 - \langle \psi \rangle_1 \right) - k \left(\langle \psi^* \rangle_0 - \langle \psi \rangle_0 \right) ,$$

so that $\langle \psi^* \rangle_1 - \langle \psi \rangle_1 \geq k \, (\alpha - \langle \psi \rangle_0)$. The right-hand side is non-negative because $k \geq 0$ and $\langle \psi \rangle_0 \leq \alpha$ by hypothesis. □

The lemma answers the design question completely for a simple-versus-simple test: order the sample space by the likelihood ratio f_1/f_0 and reject the largest values. It is the reason every construction that follows is built from a likelihood ratio rather than from anything else. For composite hypotheses no uniformly most powerful test exists in general, and the statistic of Section 11.5.4 — the ratio of maximised likelihoods — is the standard substitute: not optimal by any theorem, but reducing to the Neyman–Pearson statistic when both hypotheses are simple.

Remark 11.79 (Attribution). The lemma is Neyman and Pearson's, from their joint work of the early 1930s, and the reference is [725]. Two traps attend that citation and are recorded here because both have been fallen into. The Pearson is Egon, not his father Karl, and not either of the two other Pearsons this bibliography carries, who work on silicon and on convection cells. And Crossref stores the title with the issue's article number "IX." prefixed to it; that numeral is not part of the title and is dropped in the bibliography entry.

11.5.3 The chi-squared distribution

Definition 11.80 (Chi-squared)**.** Let $Z_1, \ldots, Z_k$ be independent standard Gaussian variables. The distribution of $\sum_{i=1}^{k} Z_i^2$ is the *chi-squared distribution with k degrees of freedom*, written χ^2_k.

Proposition 11.81 (Density, mean and variance)**.** *The χ^2_k distribution has density*

$$f_k(x) = \frac{1}{2^{k/2}\,\Gamma(k/2)} \, x^{k/2-1} \, \mathrm{e}^{-x/2} , \qquad x > 0 , \tag{11.73}$$

mean k and variance $2k$.

Rests on Definitions 11.42 and 11.80 and Proposition 11.43.

Derivation. *Derives Proposition 11.81.*

Work with the Laplace transform $M_k(s) = \left\langle \mathrm{e}^{-sX} \right\rangle$, which for a distribution on $(0, \infty)$ exists for all $s \geq 0$. For one degree of freedom, completing the exponent,

$$M_1(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-su^2} \mathrm{e}^{-u^2/2} \, \mathrm{d}u = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-(1+2s)u^2/2} \, \mathrm{d}u = (1+2s)^{-1/2} , \tag{11.74}$$

using $\int \mathrm{e}^{-au^2/2} \mathrm{d}u = \sqrt{2\pi/a}$. The k squares are independent, so their transforms multiply exactly as characteristic functions do in Proposition 11.43(iii), giving $M_k(s) = (1+2s)^{-k/2}$.

Now transform the claimed density. Substituting $y = (\tfrac{1}{2} + s)x$,

$$\int_0^{\infty} \frac{x^{k/2-1} \mathrm{e}^{-x/2}}{2^{k/2}\Gamma(k/2)} \mathrm{e}^{-sx} \, \mathrm{d}x = \frac{\left(\tfrac{1}{2}+s\right)^{-k/2}}{2^{k/2}\Gamma(k/2)} \int_0^{\infty} y^{k/2-1} \mathrm{e}^{-y} \, \mathrm{d}y = \frac{\left(\tfrac{1}{2}+s\right)^{-k/2}}{2^{k/2}} = (1+2s)^{-k/2} ,$$

the middle integral being $\Gamma(k/2)$ by definition. The two distributions have the same Laplace transform on $s > 0$; since both are supported on $[0, \infty)$ their transforms extend analytically to $\operatorname{Re} s > -\tfrac{1}{2}$ and agree there, in particular on the imaginary axis, where they are the

characteristic functions of Definition 11.42. Uniqueness of the characteristic function [1446, 1894] then forces the distributions to coincide, which is Equation (11.73).

For the moments, differentiate M_k at $s = 0$: $M_k'(s) = -k(1+2s)^{-k/2-1}$ and $M_k''(s) = k(k+2)(1+2s)^{-k/2-2}$, so $\langle X \rangle = -M_k'(0) = k$ and $\langle X^2 \rangle = M_k''(0) = k(k+2)$, whence $\operatorname{var} X = k(k+2) - k^2 = 2k$. □

Corollary 11.82 (The one degree of freedom tail)**.** *For $z \geq 0$,*

$$\Pr\!\left(\chi^2_1 \geq z^2\right) = 2\,(1 - \Phi(z))\ . \tag{11.75}$$

Rests on Definition 11.80.

Derivation. *Derives Corollary 11.82.*

$\chi^2_1 = Z^2$ with Z standard Gaussian, and $Z^2 \geq z^2$ exactly when $|Z| \geq z$, an event of probability $\Pr(Z \geq z) + \Pr(Z \leq -z) = 2(1 - \Phi(z))$ by symmetry. □

That factor of two is small, obvious, and responsible for more misquoted significances than any other step in the subject. It is tracked explicitly in Section 11.5.5.

Lemma 11.83 (Gaussian quadratic forms)**.** *If $\boldsymbol{Y} \in \mathbb{R}^k$ is Gaussian with mean zero and positive-definite covariance Σ, then $\boldsymbol{Y}^{\mathsf{T}}\Sigma^{-1}\boldsymbol{Y} \sim \chi^2_k$. More generally, if $\boldsymbol{V} \in \mathbb{R}^k$ is Gaussian with mean zero and covariance $\mathbb{1}$ and P is the orthogonal projector onto a subspace of dimension r, then $\|P\boldsymbol{V}\|^2 \sim \chi^2_r$.*

Rests on Definition 11.80.

Derivation. *Derives Lemma 11.83.*

By the spectral theorem for real symmetric matrices — diagonalise Σ by an orthogonal matrix and take the positive square roots of its eigenvalues, which are positive because Σ is positive definite — Σ has a symmetric positive-definite square root $\Sigma^{1/2}$. Put $\boldsymbol{V} = \Sigma^{-1/2}\boldsymbol{Y}$; a linear image of a Gaussian vector is Gaussian, with covariance $\Sigma^{-1/2}\Sigma\,\Sigma^{-1/2} = \mathbb{1}$, so the components of $\boldsymbol{V}$ are independent standard Gaussians. Then $\boldsymbol{Y}^{\mathsf{T}}\Sigma^{-1}\boldsymbol{Y} = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \sum_i V_i^2$, which is χ^2_k by Definition 11.80. For the second statement choose an orthonormal basis $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$ of the image of P; the components $\boldsymbol{u}_j \cdot \boldsymbol{V}$ are jointly Gaussian with $\langle (\boldsymbol{u}_i \cdot \boldsymbol{V})(\boldsymbol{u}_j \cdot \boldsymbol{V}) \rangle = \boldsymbol{u}_i \cdot \boldsymbol{u}_j = \delta_{ij}$, hence independent standard Gaussians, and $\|P\boldsymbol{V}\|^2 = \sum_{j=1}^{r} (\boldsymbol{u}_j \cdot \boldsymbol{V})^2$. □

The projector form of Lemma 11.83 has one consequence that metrology uses constantly and that is worth having as a labelled statement rather than rederived at each site.

Corollary 11.84 (Residuals about a weighted mean; the Birge ratio)**.** *Let $x_1, \dots, x_n$ be independent Gaussian measurements of one quantity μ, with known standard uncertainties u_i, and let*

$$\hat{x} = \frac{\sum_i x_i/u_i^2}{\sum_i 1/u_i^2} \tag{11.76}$$

be their inverse-variance weighted mean (Theorem 2.10). Then

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \hat{x})^2}{u_i^2} \sim \chi^2_{n-1}\,, \tag{11.77}$$

whatever the value of μ. In particular $\langle \chi^2 \rangle = n - 1$ with variance $2(n-1)$, and the Birge ratio *$R = \sqrt{\chi^2/(n-1)}$ is the factor by which every u_i must be multiplied to make the set internally consistent, in the sense that the χ^2 recomputed with uncertainties Ru_i equals $n-1$ exactly.*

Rests on Lemma 11.83, Proposition 11.81, and Theorem 2.10.

Derivation. *Derives Corollary 11.84.*

Put $V_i = (x_i - \mu)/u_i$, so that $\boldsymbol{V} \in \mathbb{R}^n$ is Gaussian with mean zero and covariance $\mathbb{1}$. Write $S = \sum_k u_k^{-2}$ and let $\boldsymbol{m}$ be the vector with components $m_i = u_i^{-1}/\sqrt{S}$, which is a unit vector because $\sum_i m_i^2 = S^{-1}\sum_i u_i^{-2} = 1$. From Equation (11.76),

$$\hat{x} - \mu = \frac{1}{S}\sum_j \frac{x_j - \mu}{u_j^2} = \frac{1}{S}\sum_j \frac{V_j}{u_j} = \frac{1}{\sqrt{S}}\,\boldsymbol{m}\cdot\boldsymbol{V}\,, \tag{11.78}$$

and therefore

$$\frac{x_i - \hat{x}}{u_i} = V_i - \frac{\hat{x} - \mu}{u_i} = V_i - m_i\,(\boldsymbol{m}\cdot\boldsymbol{V}) = \big(P\boldsymbol{V}\big)_i\,, \tag{11.79}$$

where $P = \mathbb{1} - \boldsymbol{m}\boldsymbol{m}^{\mathsf{T}}$ is the orthogonal projector onto the hyperplane orthogonal to $\boldsymbol{m}$, of dimension $n-1$. Hence $\chi^2 = \|P\boldsymbol{V}\|^2$, which is χ^2_{n-1} by the second statement of Lemma 11.83; μ has cancelled, as Equation (11.79) shows. The mean and variance are Proposition 11.81 with $k = n-1$, and the rescaling claim is immediate because χ^2 is homogeneous of degree -2 in the uncertainties. □

Remark 11.85 (What the ratio does and does not say). A Birge ratio near unity is consistency, not correctness: it says the scatter of the measurements matches the uncertainties they quote, and it is blind to a bias common to all of them. A ratio well above unity says at least one quoted uncertainty is too small, and gives no information about which. Expanding every uncertainty by R is therefore a confession rather than a correction, and it is what the CODATA adjustments do where inputs disagree [2431]; the worked case is the six modern determinations of the gravitational constant in Chapter 33, where $R \approx 4$. The count of degrees of freedom is $n-1$ and not n because one linear combination of the residuals is forced to vanish by the definition of $\hat{x}$ — that is exactly the content of the projector in Equation (11.79).

Remark 11.86 (What the last derivation imports). The finite-dimensional spectral theorem is proved in this treatise for a self-adjoint operator on a *complex* inner-product space (Theorem 5.75 in Section 5.6, whose infinite-dimensional counterpart is Theorem 12.59); what is used above is its real orthogonal form, for a real symmetric matrix, which is a standard corollary and is nowhere separately stated. It is quoted here, in the same explicit way as in Chapter 15, and the reader should treat it as an assumption of this lemma rather than as something established earlier in the book. What is actually needed is weaker than the full theorem: only the existence of *some* invertible A with $\Sigma = AA^{\mathsf{T}}$, since $\boldsymbol{V} = A^{-1}\boldsymbol{Y}$ then has covariance $\mathbb{1}$ and $\boldsymbol{Y}^{\mathsf{T}}\Sigma^{-1}\boldsymbol{Y} = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V}$; the symmetric square root is one such A, and Gram–Schmidt applied to the inner product defined by Σ^{-1} supplies a triangular one. The same weaker statement is all that Theorem 11.88 uses when it writes $I^{1/2}$.

11.5.4 The likelihood-ratio statistic and Wilks' theorem

Definition 11.87 (Likelihood-ratio test statistic)**.** For a null hypothesis $H_0 : \theta \in \Theta_0$ inside a model with parameter set Θ, the *likelihood-ratio test statistic* is

$$q = -2\ln\frac{\sup_{\theta\in\Theta_0} L(\theta)}{\sup_{\theta\in\Theta} L(\theta)} = -2\left[\ell(\hat{\theta}_0) - \ell(\hat{\theta})\right] \;\geq\; 0\,, \tag{11.80}$$

where $\hat{\theta}_0$ is the constrained and $\hat{\theta}$ the unconstrained maximum-likelihood estimator. For a simple null $\Theta_0 = \{\theta_0\}$ this is $q(\theta_0) = -2\ln\big[L(\theta_0)/L(\hat{\theta})\big]$; for a parameter of interest with nuisance parameters it is $q(\mu) = -2\ln\lambda(\mu)$ with λ the profile likelihood ratio of Definition 11.70.

Rests on Definitions 11.54, 11.56 and 11.70.

Non-negativity is immediate: the constrained supremum is over a subset, so the ratio is at most 1 and its logarithm at most zero. The statistic is also a pure number — the SI dimensions of θ cancel in the ratio — which is what allows a single tabulated distribution to serve every measurement.

The asymptotic law of that statistic is Wilks' theorem [825]; its hypotheses are stated in full below because the two ways they fail are exactly the two corrections a collider search has to apply.

Theorem 11.88 (Wilks)**.** *Let the model satisfy (R1)–(R5) at the true value θ_0, let the sample be independent and identically distributed of size n, let $\Theta \subseteq \mathbb{R}^k$ and let Θ_0 be, near θ_0, a smooth submanifold of dimension $k - r$ — equivalently, let H_0 impose r smooth, functionally independent constraints. If θ_0 lies in Θ_0 and in the interior of Θ, then under H_0*

$$q \;\longrightarrow\; \chi^2_r \qquad \text{in distribution as } n \to \infty\,, \tag{11.81}$$

where r is the number of parameters constrained by the null.

Rests on Definitions 11.59, 11.80 and 11.87, Theorem 11.67, and Lemmas 11.52 and 11.83.

Derivation, one parameter and a simple null. *Derives Theorem 11.88.*

Take $k = r = 1$ and $\Theta_0 = \{\theta_0\}$. Expand ℓ_n about the unconstrained MLE $\hat\theta_n$, where ℓ_n' vanishes:

$$\ell_n(\theta_0) = \ell_n(\hat\theta_n) + \tfrac{1}{2}\ell_n''(\hat\theta_n)\left(\theta_0 - \hat\theta_n\right)^2 + \tfrac{1}{6}\ell_n'''(\bar\theta_n)\left(\theta_0 - \hat\theta_n\right)^3\,, \tag{11.82}$$

with $\bar\theta_n$ between the two. Multiply by -2:

$$q = \left[\sqrt{n}\left(\hat\theta_n - \theta_0\right)\right]^2 \left(\frac{-\ell_n''(\hat\theta_n)}{n}\right) - \frac{1}{3}\ell_n'''(\bar\theta_n)\left(\theta_0 - \hat\theta_n\right)^3\,. \tag{11.83}$$

From Equation (11.82) (multiply by -2 and rescale the deviation by $\sqrt{n}$).

The first bracket converges in distribution to $\mathcal{N}(0, I_1(\theta_0)^{-1})$ by Theorem 11.67, and is therefore bounded in probability. The second factor converges in probability to $I_1(\theta_0)$: it is an average of independent identically distributed terms evaluated at $\hat\theta_n$ rather than at θ_0, and the difference is controlled by the domination in (R5) together with the consistency of $\hat\theta_n$. The last term is bounded in absolute value by $\tfrac{1}{3}\left(n^{-1}\textstyle\sum_i M(X_i)\right) n\left|\hat\theta_n - \theta_0\right|^3$, and $n\left|\hat\theta_n - \theta_0\right|^3 = n^{-1/2}\left[\sqrt{n}\left|\hat\theta_n - \theta_0\right|\right]^3 \to 0$ in probability. By Lemma 11.52,

$$q \;\longrightarrow\; \left[\sqrt{I_1(\theta_0)}\,W\right]^2\,, \qquad W \sim \mathcal{N}\!\left(0, I_1(\theta_0)^{-1}\right)\,,$$

and $\sqrt{I_1}\,W$ is standard Gaussian, so its square is χ^2_1 by Definition 11.80. □

Derivation, k parameters and r constraints. *Derives Theorem 11.88.*

The same expansion in k dimensions gives, for any θ near $\hat\theta_n$,

$$Q(\theta) \equiv -2\left[\ell_n(\theta) - \ell_n(\hat\theta_n)\right] = n\left(\theta - \hat\theta_n\right)^{\mathsf T}\left[I_1(\theta_0) + o_P(1)\right]\left(\theta - \hat\theta_n\right) + o_P(1)\,, \tag{11.84}$$

the Hessian term being handled exactly as above and the cubic remainder again vanishing. Since $\hat\theta_0$ maximises ℓ_n over Θ_0, it minimises Q there, and $q = Q(\hat\theta_0) = \min_{\theta\in\Theta_0} Q(\theta)$. Near θ_0 the submanifold Θ_0 is its tangent space to the accuracy the expansion retains, $\Theta_0 \simeq \theta_0 + E$ with E that tangent space, of dimension $\dim E = k - r$. Writing $\boldsymbol{W} = \sqrt{n}\left(\hat\theta_n - \theta_0\right)$ and $I = I_1(\theta_0)$,

$$q \;\longrightarrow\; \min_{\boldsymbol{e}\in E} (\boldsymbol{W} - \boldsymbol{e})^{\mathsf T}\, I\, (\boldsymbol{W} - \boldsymbol{e})\,. \tag{11.85}$$

Substitute $\boldsymbol{V} = I^{1/2}\boldsymbol{W}$ and $S = I^{1/2}E$; the right side is $\min_{\boldsymbol{s}\in S}\|\boldsymbol{V} - \boldsymbol{s}\|^2 = \|P\boldsymbol{V}\|^2$ with P the orthogonal projector onto $S^\perp$, of dimension r. By Theorem 11.67 $\boldsymbol{W} \to \mathcal{N}(0, I^{-1})$, so $\boldsymbol{V}$ is asymptotically Gaussian with covariance $I^{1/2}I^{-1}I^{1/2} = \mathbb{1}$, and Lemma 11.83 gives $\|P\boldsymbol{V}\|^2 \sim \chi^2_r$. □

Two steps of that derivation are argued here rather than proved: the remainder in the k-dimensional Taylor expansion must be shown uniform on a neighbourhood of the true parameter shrinking at rate $n^{-1/2}$, and both maximum-likelihood estimators must be shown to lie in that neighbourhood with probability tending to one, which is what licenses replacing the constraint surface by its tangent space at Equation (11.85). Both are supplied in the appendix, from the same regularity conditions (R1)–(R5) assumed here.

Full derivation: Appendix A.21 (p. 2873).

Derives Theorem 11.88.

Corollary 11.89 (Degrees of freedom with nuisance parameters)**.** *For a single parameter of interest μ with any number m of nuisance parameters, the profile statistic $q(\mu_0) = -2\ln\lambda(\mu_0)$ of Definition 11.70 is asymptotically χ^2_1 under $H_0 : \mu = \mu_0$, whatever m is.*

Rests on Definition 11.70 and Theorem 11.88.

Derivation. *Derives Corollary 11.89.*

The null constrains one coordinate and leaves the other m free, so Θ_0 is a submanifold of codimension $r = 1$ and Theorem 11.88 applies with $r = 1$. The nuisance parameters affect the *variance* of $\hat\mu$, through the Schur complement Equation (11.67), but not the number of degrees of freedom. □

This is the result a collider search relies on: however many nuisance parameters describe the background and the detector, the significance is read off a one-degree-of-freedom distribution [2172]. It is worth saying exactly when it stops being true.

Remark 11.90 (Failure I: a parameter on the boundary). Condition (R3) requires θ_0 to be interior. A signal strength constrained to $\mu \geq 0$ and tested at $\mu = 0$ violates it, and the distribution changes. In the quadratic regime let $\hat\mu_{\mathrm{u}}$ be the unconstrained maximiser, asymptotically Gaussian with mean zero and variance σ^2 under H_0; the maximiser over $\mu \geq 0$ is $\hat\mu = \max(\hat\mu_{\mathrm{u}}, 0)$, because the quadratic log-likelihood opens downward and, with its vertex at a negative abscissa, is decreasing throughout $[0, \infty)$. Then

$$q = \begin{cases} 0 & \hat\mu_{\mathrm{u}} \leq 0\,, \\ \hat\mu_{\mathrm{u}}^2/\sigma^2 & \hat\mu_{\mathrm{u}} > 0\,, \end{cases} \qquad \text{so} \qquad q \;\longrightarrow\; \tfrac{1}{2}\delta_0 + \tfrac{1}{2}\chi^2_1\,, \tag{11.86}$$

a half-and-half mixture of a point mass at zero and a chi-squared law, since $\hat\mu_{\mathrm{u}} \leq 0$ has asymptotic probability $\frac{1}{2}$ by symmetry. Hence $\Pr(q \geq u) = \frac{1}{2}\Pr(\chi^2_1 \geq u)$ for $u > 0$: using the uncorrected χ^2_1 tail overstates the p-value by exactly a factor of two, and therefore understates the significance. The boundary correction and the half-and-half mixture are Chernoff's [1033]; the same mixture is used deliberately in Section 11.5.5, and the convention a search actually quotes is fixed in [2172].

Rests on Definition 11.59 and Corollary 11.89.

Remark 11.91 (Failure II: a parameter unidentified under the null). Condition (R2) can fail on Θ_0 itself. Consider a model $f(x; \mu, m)$ in which μ is the strength of a signal and m its mass. At $\mu = 0$ the density does not depend on m at all: every value of m gives the same distribution, the parameter is unidentified, the information matrix is singular in the m direction, and the expansion behind Theorem 11.88 has nothing to expand in. The statistic $q(m)$ is then not a single number but a *random field* over the mass range, and the quantity actually reported — $\sup_m q(m)$ — has no chi-squared distribution and is not covered by any theorem above. This is the look-elsewhere problem, and Section 11.6 treats it. The standard analysis of the unidentified-parameter case is R. B. Davies' [1594], extended in his sequel of ten years later [1791]; the two papers share a title and neither supersedes the other. He is not the E. B. Davies of the master-equation literature cited elsewhere in this treatise.

11.5.5 Significance

Definition 11.92 (Significance)**.** The *significance* Z corresponding to a p-value is defined by

$$p = 1 - \Phi(Z)\,, \qquad \text{equivalently} \qquad Z = \Phi^{-1}(1-p)\,, \tag{11.87}$$

with Φ the standard Gaussian distribution function: Z is the number of standard deviations beyond which a *one-sided* Gaussian tail carries probability p.

Rests on Definition 11.75.

Z is a pure number and carries no unit. The thresholds conventionally called evidence and discovery in particle physics are $Z = 3$ and $Z = 5$; the corresponding tail probabilities, computed directly from Equation (11.87), are $p = 1.35 \times 10^{-3}$ and $p = 2.87 \times 10^{-7}$. The statistical practice of the field, these conventions among it, is reviewed in [2427], and the profile-likelihood asymptotics that turn a measured statistic into one of these numbers are collected in [2172].

Proposition 11.93 ($Z = \sqrt{q}$)**.** *Let μ be a signal strength with $H_0 : \mu = 0$, and define the* one-sided discovery statistic

$$\tilde{q}_0 = \begin{cases} -2\ln\left[L(0)/L(\hat{\mu})\right] & \hat{\mu} \geq 0\,, \\ 0 & \hat{\mu} < 0\,, \end{cases} \tag{11.88}$$

which refuses to count a deficit as evidence for a signal. Then under H_0, asymptotically, $\tilde{q}_0 \sim \frac{1}{2}\delta_0 + \frac{1}{2}\chi^2_1$, and for an observed value $q > 0$

$$p = \Pr(\tilde{q}_0 \geq q \mid H_0) = \tfrac{1}{2}\Pr\!\left(\chi^2_1 \geq q\right) = 1 - \Phi(\sqrt{q})\ , \qquad \textit{so} \qquad Z = \sqrt{q}\,. \tag{11.89}$$

Rests on Definition 11.92, Corollaries 11.82 and 11.89, and Equation (11.86)*.*

Derivation. *Derives Proposition 11.93.*

The mixture is the one derived in Equation (11.86): writing $q_0 = \hat{\mu}^2/\sigma^2$ for the two-sided statistic, which is asymptotically χ^2_1 by Corollary 11.89, the event $\tilde{q}_0 \geq q$ with $q > 0$ is the event $\{\hat{\mu} > 0\} \cap \{q_0 \geq q\}$. The sign of $\hat{\mu}$ and the magnitude $|\hat{\mu}|$ are independent for a Gaussian of mean zero, and the sign is positive with probability $\frac{1}{2}$, so $\Pr(\tilde{q}_0 \geq q) = \frac{1}{2}\Pr(\chi^2_1 \geq q)$. Insert Equation (11.75) with $z = \sqrt{q}$: the right-hand side is $\frac{1}{2} \cdot 2\,(1 - \Phi(\sqrt{q})) = 1 - \Phi(\sqrt{q})$. Comparing with Equation (11.87) gives $Z = \sqrt{q}$. □

The identity $Z = \sqrt{q}$ is the reason a collider search can quote a significance without simulating anything, and it is the convention fixed, together with the asymptotic distributions it rests on, in [2172]. The same reference supplies the companion device for the *expected* significance of a planned measurement: a single artificial "Asimov" data set, in which every observable is set to its expectation under the assumed signal, returns the median q of an ensemble of simulated experiments directly, so that the sensitivity of a search can be quoted without generating the ensemble at all. The treatise uses only the observed-data statement above; the Asimov construction is named here because the searches of Part XI quote sensitivities computed with it.

Remark 11.94 (The factor of two). The identity $Z = \sqrt{q}$ is a property of the one-sided convention and of nothing else. Used with the two-sided statistic q_0, which counts a downward fluctuation as evidence against $\mu = 0$ just as readily as an upward one, the correct p-value is $\Pr(\chi^2_1 \geq q) = 2\,(1 - \Phi(\sqrt{q}))$, exactly twice as large, and the significance is $Z = \Phi^{-1}(1 - 2\,(1 - \Phi(\sqrt{q})))$, which is not $\sqrt{q}$. At $q = 25$ the two conventions give $p = 2.87 \times 10^{-7}$ with $Z = 5.00$, and $p = 5.73 \times 10^{-7}$ with $Z = 4.86$. The difference of 0.14 in Z is small enough to look like a rounding disagreement and large enough to move a result across the conventional discovery threshold. State which convention is in force, once, and do not mix them.

Remark 11.95 (Exclusion is the same machinery reversed). An upper limit uses a statistic that is one-sided the other way — large values of μ disfavoured — and the same asymptotic distributions apply [2172]. The additional difficulty there is that a downward background fluctuation can exclude a signal the experiment had no sensitivity to; the standard remedy is to report the ratio of the two tail probabilities rather than the numerator alone [2036]. That construction belongs with the searches that use it and is not developed here.

11.6 The look-elsewhere effect

Everything in Section 11.5 assumes the hypothesis was fixed before the data were examined. A search does not work that way. It scans a range of an unknown parameter — a mass, a lifetime, a frequency — computes a test statistic at every point of the range, and reports the largest value found. The probability that *some* point of a wide range fluctuates upward is far greater than the probability that a nominated point does, and the difference can be one or two units of significance. This section makes the correction quantitative.

11.6.1 Local and global p-values

Definition 11.96 (Local, global, trials factor)**.** Let $M = [m_1, m_2]$ be the scanned range and $\tilde{q}(m)$ the one-sided statistic Equation (11.88) evaluated with the signal parameter fixed at m. Let $u = \max_{m\in M} \tilde{q}(m)$ be the observed maximum. The *local p-value* is the tail probability at a point fixed in advance,

$$p_{\text{loc}} = \Pr(\tilde{q}(m_{\text{fix}}) \geq u \mid H_0) = 1 - \Phi(\sqrt{u}) \,, \tag{11.90}$$

by Equation (11.89); the *global p-value* is

$$p_{\text{glob}} = \Pr\Big(\max_{m\in M} \tilde{q}(m) \geq u \mid H_0\Big) \,, \tag{11.91}$$

and the *trials factor* is $N_{\text{tr}} = p_{\text{glob}}/p_{\text{loc}}$.

Rests on Definition 11.75 and Equations (11.88) *and* (11.89).

Proposition 11.97 (The trials factor is at least one)**.** $p_{\text{glob}} \geq p_{\text{loc}}$*, so* $N_{\text{tr}} \geq 1$.

Rests on Definition 11.96.

Derivation. *Derives Proposition 11.97.*

For any fixed $m_{\text{fix}} \in M$ the event $\{\tilde{q}(m_{\text{fix}}) \geq u\}$ is contained in the event $\{\max_{m\in M} \tilde{q}(m) \geq u\}$; a probability is monotone under inclusion. □

Proposition 11.98 (Union bound and the naive estimate)**.** *Suppose the maximum over M is attained, to the accuracy required, at one of N points $m_1, \dots, m_N$. Then*

$$p_{\text{glob}} \leq \sum_{j=1}^{N} \Pr(\tilde{q}(m_j) \geq u) = N\, p_{\text{loc}} \,, \qquad \text{so} \qquad N_{\text{tr}} \leq N \,. \tag{11.92}$$

If in addition the N tests were independent, then

$$p_{\text{glob}} = 1 - (1 - p_{\text{loc}})^N = N p_{\text{loc}} - \binom{N}{2} p_{\text{loc}}^2 + \cdots \simeq N p_{\text{loc}} \quad \text{when } N p_{\text{loc}} \ll 1 \,. \tag{11.93}$$

Rests on Definition 11.96.

Derivation. *Derives Proposition 11.98.*

Write $A_j = \{\tilde{q}(m_j) \geq u\}$. The first claim is Boole's inequality $\Pr(\bigcup_j A_j) \leq \sum_j \Pr(A_j)$, which follows by induction from $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \leq \Pr(A) + \Pr(B)$, together with the fact that each $\Pr(A_j)$ equals p_{loc} because the statistic has the same null distribution at every m. If the A_j are independent then so are their complements, so $\Pr(\bigcap_j A_j^c) = (1 - p_{\mathrm{loc}})^N$; expanding by the binomial theorem gives the series. □

The customary choice is $N \simeq \Delta m / \delta m$, the width of the scanned window divided by the experimental resolution, on the reasoning that two hypotheses separated by much less than the resolution are tested by essentially the same data and are not independent trials.

Remark 11.99 (What is and is not proved here). Equation (11.92) is a theorem and Equation (11.93) is not. The union bound holds unconditionally but is weak, and it becomes vacuous as the grid is refined — nothing stops N from being taken larger, and the bound then exceeds 1. The independence assumption is false by construction: $\tilde{q}(m)$ and $\tilde{q}(m')$ are computed from the same events and are strongly correlated when $|m - m'|$ is of order the resolution. The number N is therefore a judgement about a correlation length, not a quantity the model supplies, and the estimate it gives can err in either direction. The next subsection replaces the judgement by a measurable quantity.

11.6.2 Upcrossings

The honest object is the whole function $m \mapsto \tilde{q}(m)$, a random field on M whose null distribution at each point is known but whose maximum is not. What controls the maximum is how often the field crosses a high level on its way up.

Definition 11.100 (Upcrossings)**.** For a continuous path $\tilde{q}(\,\cdot\,)$ on M and a level u, an *upcrossing* is a point $m \in (m_1, m_2)$ at which the path passes from below u to above it. Write N_u for the number of upcrossings in M, a non-negative integer-valued random variable.

Rests on Definition 11.96.

The bound below is Davies' [1594, 1791]: it is the result that replaces Wilks' theorem when a parameter — here the mass — is present only under the alternative, and it is what the trials-factor practice of a search is built on [2158].

Theorem 11.101 (Davies' bound)**.** *If the paths of $\tilde{q}$ are continuous, then for every level u*

$$\Pr\Big(\max_{m \in M} \tilde{q}(m) \geq u\Big) \leq \Pr(\tilde{q}(m_1) \geq u) + \langle N_u \rangle \,, \tag{11.94}$$

that is, $p_{\mathrm{glob}} \leq p_{\mathrm{loc}} + \langle N_u \rangle$.

Rests on Definition 11.100, Lemma 11.50, and Theorem 7.23.

Derivation. *Derives Theorem 11.101.*

Suppose $\max_{m \in M} \tilde{q}(m) \geq u$ and that the left endpoint is below the level, $\tilde{q}(m_1) < u$. The path is continuous and reaches a value at least u somewhere in M, so by the intermediate value theorem (Theorem 7.23) it attains the value u at some interior point, having come from below; that is an upcrossing, so $N_u \geq 1$. Hence

$$\Big\{\max_{m \in M} \tilde{q}(m) \geq u\Big\} \subseteq \{\tilde{q}(m_1) \geq u\} \cup \{N_u \geq 1\} \,,$$

and subadditivity gives $\Pr(\max \geq u) \leq \Pr(\tilde{q}(m_1) \geq u) + \Pr(N_u \geq 1)$. Finally N_u is non-negative, so Markov's inequality (Lemma 11.50) with $a = 1$ gives $\Pr(N_u \geq 1) \leq \langle N_u \rangle$. □

The bound is exact mathematics and it is also practical, because it replaces an unknown — the distribution of a maximum in the far tail, which no feasible number of simulated experiments can sample — by a mean, which can be estimated at a *low* level where upcrossings are common. What makes that replacement useful is the way $\langle N_u \rangle$ depends on the level.

The level dependence below is Gross and Vitells' [2158], and it descends from Rice's formula for the expected number of level crossings of a smooth Gaussian process. The memoir on random noise is usually cited whole, but the crossing count is specifically section 3.3 of its second instalment [883]; the first [877] carries the shot effect and the power spectra, and not the formula used here. That formula is derived in the appendix — by a Kac counting integral whose expectation factorises, because a field of constant variance is independent of its own derivative at each point — and everything built on it here follows from it.

Theorem 11.102 (Rice's formula)**.** *Let X be a Gaussian process on $M = [m_1, m_2]$ with zero mean, unit variance at every point and continuously differentiable sample paths, and write $\sigma_1(m)^2 = \operatorname{var} X'(m)$, assumed finite and continuous in m. Then the expected number of upcrossings of a level $z \in \mathbb{R}$ by X on M is*

$$\langle N_z(X) \rangle = \frac{\mathrm{e}^{-z^2/2}}{2\pi} \int_{m_1}^{m_2} \sigma_1(m)\, \mathrm{d}m \,. \tag{11.95}$$

Rests on Definitions 11.28 and 11.100.

Two features of Equation (11.95) are worth naming, because they are what the next theorem lives on. Constancy of the variance forces $\operatorname{cov}(X(m), X'(m)) = \frac{1}{2}\,\mathrm{d}\operatorname{var} X(m)/\mathrm{d}m = 0$, so field and derivative are independent at each fixed m; the level therefore enters only through the Gaussian density of $X(m)$ at z, the expected positive part of the derivative, $\langle X'(m)^+ \rangle = \sigma_1(m)/\sqrt{2\pi}$, being the same at every level. That is the whole reason the right-hand side factorises into a function of z times a function of the field. For a stationary field σ_1 is a constant $\sqrt{\lambda_2}$ and Equation (11.95) reduces to Rice's original expression $\langle N_z \rangle = (L/2\pi)\sqrt{\lambda_2}\,\mathrm{e}^{-z^2/2}$, with L the length of M.

The appendix carries the derivation, and states there what it must assume: the smoothness hypotheses above in the sharper form the proof uses, the absence of tangential crossings, and the two Lebesgue-theory statements — Tonelli's theorem and dominated convergence — that license the interchange of the expectation with the counting integral and with the limit. Everything the chapter builds on the formula, including the exponential level dependence and the trials factor, is derived in full.

Full derivation: Appendix A.22 (p. 2879).

Derives Theorem 11.102.

Theorem 11.103 (Level dependence of the mean upcrossing count)**.** *Let the one-sided statistic Equation (11.88) be, asymptotically, $\tilde{q}(m) = X(m)^2\, \mathbf{1}_{\{X(m)>0\}}$ for a Gaussian field X satisfying the hypotheses of Theorem 11.102, indexed by the single scanned parameter m. Then for levels $u, u_0 > 0$ in the range where that asymptotic description holds,*

$$\langle N_u \rangle = \langle N_{u_0} \rangle\, \mathrm{e}^{-(u-u_0)/2} \,. \tag{11.96}$$

Rests on Definition 11.100 and Theorem 11.102.

Derivation. *Derives Theorem 11.103.*

Fix $u > 0$. An upcrossing of the level u by $\tilde{q}$ is exactly an upcrossing of the level $\sqrt{u}$ by X. Where $X \le 0$ the statistic takes the value 0, which is below u, so no crossing of u occurs there; and on the open set where $X > 0$ one has $\tilde{q} = X^2$, a strictly increasing function of X, so $\tilde{q}$ passes from below u to above it precisely where X passes from below $\sqrt{u}$ to above it.

The two counts therefore agree path by path, $N_u = N_{\sqrt{u}}(X)$, and taking expectations and applying Equation (11.95) at $z = \sqrt{u}$,

$$\langle N_u \rangle = \frac{\mathrm{e}^{-u/2}}{2\pi} \int_{m_1}^{m_2} \sigma_1(m)\,\mathrm{d}m\,, \tag{11.97}$$

where the integral does not depend on u. Writing the same identity at u_0 and dividing cancels the integral and leaves Equation (11.96). □

Remark 11.104 (More than one scanned parameter). Gross and Vitells extend the result to a scan over d parameters, where the right-hand side of Equation (11.96) acquires a factor $(u/u_0)^{(d-1)/2}$ [2158]; the upcrossing count is there replaced by the expected Euler characteristic of the set on which the field exceeds the level, of which it is the case $d = 1$. That extension is reported, not derived here, and nothing in this treatise uses it: every application below is a scan over a single mass.

The procedure this licenses is the one searches actually use: simulate a manageable number of background-only experiments, count upcrossings of a *low* threshold u_0 — $u_0 = 1$, i.e. a one-sigma level, is a common choice — average to get $\langle N_{u_0} \rangle$, and extrapolate with Equation (11.96) to the observed level. Nothing has to be simulated at 5×10^{-7} probability.

Corollary 11.105 (Trials factor at high significance)**.** *With $u = Z_{\mathrm{loc}}^2$ and Z_{loc} large,*

$$N_{\mathrm{tr}} \;\le\; 1 + \frac{\langle N_{u_0} \rangle\, \mathrm{e}^{(u_0-u)/2}}{1 - \Phi(\sqrt{u})} \;\simeq\; 1 + \sqrt{2\pi}\,\langle N_{u_0} \rangle\, \mathrm{e}^{u_0/2}\, Z_{\mathrm{loc}}\,, \tag{11.98}$$

so the trials factor grows linearly with the local significance.
Rests on Definition 11.96 and Theorems 11.101 and 11.103.

Derivation. *Derives Corollary 11.105.*

Divide Equation (11.94) by p_{loc} and use Equations (11.90) and (11.96) for the first form. For the second, the Gaussian tail satisfies

$$\frac{z}{z^2+1}\,\frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}} \;\le\; 1 - \Phi(z) \;\le\; \frac{1}{z}\,\frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}}\,, \qquad z > 0\,, \tag{11.99}$$

so that $1 - \Phi(z) = \mathrm{e}^{-z^2/2}\,(1 + O(z^{-2}))\,/\,\big(z\sqrt{2\pi}\big)$. To see Equation (11.99), write $J = \int_z^\infty \mathrm{e}^{-t^2/2}\mathrm{d}t$ and integrate by parts with $\mathrm{e}^{-t^2/2} = t^{-1} \cdot t\,\mathrm{e}^{-t^2/2}$:

$$J = \frac{\mathrm{e}^{-z^2/2}}{z} - \int_z^\infty \frac{\mathrm{e}^{-t^2/2}}{t^2}\,\mathrm{d}t\,.$$

The subtracted integral is positive, giving the upper bound, and is at most $z^{-2}J$, giving $J(1 + z^{-2}) \ge \mathrm{e}^{-z^2/2}/z$ and hence the lower bound. Now put $z = \sqrt{u} = Z_{\mathrm{loc}}$ in the first form of Equation (11.98): the exponentials $\mathrm{e}^{-u/2}$ and $\mathrm{e}^{-z^2/2}$ cancel and what remains is $\langle N_{u_0} \rangle\, \mathrm{e}^{u_0/2}\, Z_{\mathrm{loc}}\sqrt{2\pi}$. □

Remark 11.106 (The factor of two, once more). The same relation appears in the literature with $\sqrt{\pi/2}$ in place of $\sqrt{2\pi}$. Both are right, in their own conventions. There the local p-value is the two-sided tail $\Pr(\chi^2_1 \ge u) = 2\,(1 - \Phi(\sqrt{u}))$ and N_u counts upcrossings of the two-sided statistic $q(m)$, which are twice as numerous because the underlying Gaussian field crosses $+\sqrt{u}$ and $-\sqrt{u}$ with equal frequency. Both the numerator and the denominator of N_{tr} are therefore doubled, and the trials factor — a ratio — is the same number either way, as it must be.

What is fatal is to mix them — but the damage is *not* exactly a factor of two, because the Davies bound Equation (11.94) carries an additive 1 that no convention doubles. Write

$x = \langle N_u \rangle / p_{\mathrm{loc}}$ in the one-sided convention, so that the number a search quotes as its trials factor is $N_{\mathrm{tr}} = 1 + x$, the right-hand side of Equation (11.98). A two-sided upcrossing count divided by a one-sided local p-value doubles x alone and returns

$$1{+}2x = 2\,(1+x){-}1 = 2N_{\mathrm{tr}}{-}1\,, \qquad \text{too large by the factor } \frac{2N_{\mathrm{tr}}-1}{N_{\mathrm{tr}}} = 2{-}\frac{1}{N_{\mathrm{tr}}}\,. \tag{11.100}$$

What the mistake doubles is the *excess* of the trials factor over one, not the trials factor itself. The overstatement is therefore strictly less than a factor of two, and reaches two only asymptotically, for $N_{\mathrm{tr}} \gg 1$. At the $N_{\mathrm{tr}} \simeq 52$ of Example 11.107 the mixed number is about 104, too large by 1.98 — close enough to two that the loose statement passes unnoticed. In the opposite regime it is nearly harmless: a genuine $N_{\mathrm{tr}} = 1.05$ is reported as 1.10, an overstatement of five per cent, not of a factor of two. The mistake is worth avoiding for its own sake; "exactly twice" is the wrong way to describe it.

11.6.3 A worked example: a bump search over a mass window

Example 11.107 (Degrading a local five sigma)**.** A search scans for a narrow resonance in an invariant-mass spectrum over the window

$$m \in \left[110\,\mathrm{GeV}/c^2,\ 150\,\mathrm{GeV}/c^2\right], \qquad \Delta m = 40\,\mathrm{GeV}/c^2 = 7.13\times10^{-26}\,\mathrm{kg}\,, \tag{11.101}$$

with a mass resolution $\delta m = 1.5\,\mathrm{GeV}/c^2 = 2.67\times10^{-27}\,\mathrm{kg}$. The statistic $\tilde{q}(m)$ of Equation (11.88) is evaluated across the window with the background normalisation and shape treated as nuisance parameters, so Corollary 11.89 applies at each fixed m. Its largest value is $u = 25.0$, at some mass inside the window.

Local significance. By Equation (11.89), $Z_{\mathrm{loc}} = \sqrt{u} = 5.00$ and $p_{\mathrm{loc}} = 1 - \Phi(5) = 2.87\times10^{-7}$. This is the significance the excess would have if the mass had been nominated in advance, and it is not the significance of the search.

Naive trials factor. Taking $N = \Delta m/\delta m = 40/1.5 \simeq 27$ resolution elements and treating them as independent, Equation (11.93) gives $p_{\mathrm{glob}} \simeq 27\times2.87\times10^{-7} = 7.7\times10^{-6}$, whence $Z_{\mathrm{glob}} = \Phi^{-1}(1-p_{\mathrm{glob}}) = 4.32$.

Upcrossings. Suppose background-only simulations at the low threshold $u_0 = 1.0$ yield an average of $\langle N_{u_0}\rangle = 2.4$ upcrossings per experiment across the window. Then Equation (11.96) extrapolates to

$$\langle N_{25}\rangle = 2.4\,\mathrm{e}^{-(25-1)/2} = 2.4\,\mathrm{e}^{-12} = 1.47\times10^{-5}\,, \tag{11.102}$$

and Equation (11.94) gives

$$p_{\mathrm{glob}} \le 2.87\times10^{-7} + 1.47\times10^{-5} = 1.50\times10^{-5}\,, \qquad N_{\mathrm{tr}} = 52\,, \qquad Z_{\mathrm{glob}} = 4.17\,. \tag{11.103}$$

The closed form Equation (11.98) gives $1 + \sqrt{2\pi}\times2.4\times\mathrm{e}^{0.5}\times5.00 = 51$, agreeing with the direct evaluation to about three per cent — the difference being the $O(z^{-2})$ term dropped from the Gaussian tail.

Rests on Corollary 11.89, Propositions 11.93 and 11.98, and Theorems 11.101 and 11.103.

Remark 11.108 (Reading the example). Three things in Table 11.1 are worth separating. First, the degradation is real and large in probability — a factor of 52 — and mild in significance, $5.00 \to 4.17$; that compression is the exponential in Equation (11.87) at work and is why an excess that survives the correction at all tends to survive it comfortably. Second, the naive estimate is not conservative: it happened to be optimistic by a factor of about two here, and which way it errs depends on whether the resolution over-counts or under-counts the correlation length of the field. Third, Equation (11.103) is an upper bound on p_{glob}, so $Z_{\mathrm{glob}} = 4.17$ is a *lower* bound on the global significance — the honest direction for a discovery claim.

Table 11.1: The bump search of Example 11.107, a local $q = 25.0$ found somewhere in a mass window $40\,\mathrm{GeV}/c^2$ wide with a $1.5\,\mathrm{GeV}/c^2$ resolution. The trials factor and both p-values are pure numbers; the window and the resolution that generate them are not. The naive estimate is optimistic here by a factor of about two, and understates the correction by 0.15 in significance.

Treatment	N_{tr}	p_{glob}	Z_{glob}
Mass nominated in advance	1	2.87×10^{-7}	5.00
Independent resolution elements, $N = 27$	27	7.7×10^{-6}	4.32
Expected upcrossings, Equation (11.94)	52	1.50×10^{-5}	4.17

Nothing in the arithmetic depends on the mass window being the one chosen here, and nothing in it is specific to any particular resonance. What the example fixes is the discipline: a significance quoted for a search must state the range searched, and a significance quoted without it is a local number wearing a global name.

Chapter 12

Hilbert Spaces

The vector-space machinery of Chapter 5 was developed largely without regard to dimension, but the passage to infinite dimension is not innocent: limits enter, and with them the topological and metric notions of Chapter 6. A Hilbert space is the structure in which the algebra and the topology cooperate — an inner-product space whose associated metric is complete — and it is the arena in which the postulates of quantum mechanics will be formulated in Chapter 77.

The subject grew out of Hilbert's spectral theory of integral equations [388], was axiomatised by von Neumann for exactly that quantum-mechanical purpose [706], and its standard modern exposition is Reed–Simon [1481]; those three are the references of record for this chapter.

Three things are new here, and each is a place where the finite-dimensional intuition of Chapter 5 fails. A subspace need not be closed, so it need not admit an orthogonal complement; an operator need not be defined on the whole space, and if it is symmetric and everywhere defined it is forced to be bounded; and a self-adjoint operator need have no eigenvector at all, so the decomposition of a state into eigenstates has to be replaced by an integral against a projection-valued measure. The chapter is organised around repairing those three failures.

Remark 12.1 (Analytical inputs quoted, not proved). Three classical results are used below and are *not* derived in this treatise; each is flagged where it is used and all three are in [1481]. (i) The convergence theorems of Lebesgue integration — monotone convergence and Fatou's lemma — which enter only through the completeness of L^2 (Theorem 12.12); the treatise develops no measure theory, and Chapter 11 makes the same declaration. (ii) The closed-graph theorem for Banach spaces, used once, in Theorem 12.74; it rests on the Baire category theorem, which is not proved here either. (iii) The Weierstrass approximation theorem, used once, to pin down the functional calculus in Proposition 12.61. Everything else in this chapter is proved from Chapter 5, Chapter 6, Chapter 7 and Chapter 8.

The long derivations this chapter sends to Appendix A quote two further results — the Riesz–Markov representation theorem, in Appendix A.24, and the Hilbert–Schmidt embedding property of a nuclear space, in Appendix A.27 — together with the Radon–Nikodym theorem and the dominated convergence theorem, which belong to the same measure theory as (i). Each is stated as a displayed theorem where it is used, and each of those four appendix sections closes with a remark titled *What is quoted here* naming its imports exactly.

12.1 Definitions

Definition 12.2 (Hilbert space)**.** Consider a pre-Hilbert space $\mathcal{H}$ — a vector space equipped with an inner product — of infinite dimension, together with the norm associated with that

inner product and the metric associated with that norm. We say that $\mathcal{H}$ is a *Hilbert space* if and only if $\mathcal{H}$ is complete, that is, if and only if every Cauchy sequence in $\mathcal{H}$ converges.

Rests on Definitions 5.17 and 6.27 and Equation (5.44).

Nothing in Definition 12.2 is new except the word *complete*. The inner product is the map $\langle\ |\ \rangle$ of Definition 5.17, antilinear in its first slot and linear in its second (Equations (5.34) and (5.36)); the norm is $\|v\| = \sqrt{\langle v|v\rangle}$ of Equation (5.44); the metric is $d(u,v) = \|u - v\|$ of Equation (5.45); and completeness is Definition 6.27 read in that metric. The infinite dimension is not a hypothesis of any theorem below — a finite-dimensional inner-product space is automatically complete — but it is the case of interest, and we shall say *Hilbert space* for any complete inner-product space and note the finite-dimensional degeneracies where they matter.

Notation 12.3 (Scalars and adjoints). Throughout, $\mathcal{H}$ is a complex Hilbert space unless the real case is named; $\mathbb{K} = \mathbb{C}$, and a^* is the complex conjugate of $a \in \mathbb{C}$. Vectors are written x, y, z or, where the quantum-mechanical reading is intended, $|\psi\rangle$. The dagger $A^\dagger$ denotes the adjoint of Theorem 12.38 throughout the treatise, never the transpose, which is written A^{T} (Chapter 5).

Three consequences of the axioms are used on nearly every page below: the Cauchy–Schwarz inequality, the continuity of the inner product that follows from it, and the parallelogram law, which is what completeness will be levered against in Section 12.2.1.

Proposition 12.4 (Cauchy–Schwarz and continuity of the inner product)**.** *Let $\mathcal{H}$ be an inner-product space. Then*

$$|\langle u|v\rangle| \le \|u\|\ \|v\|\ , \qquad \forall\, u, v \in \mathcal{H}\,, \tag{12.1}$$

and the inner product is jointly continuous: if $u_n \to u$ and $v_n \to v$ in norm, then $\langle u_n|v_n\rangle \to \langle u|v\rangle$ in $\mathbb{C}$.

Rests on Proposition 5.19 and Definition 12.2.

Proof. *Derives Proposition 12.4.*

Equation (12.1) is Equation (5.43) with both sides put under a square root through Equation (5.44). That inequality is available here without any further argument: Remark 5.20 records that its proof used only the five axioms of Definition 5.17 — neither a basis, nor finite dimension, nor completeness — so it holds verbatim in the present setting.

For continuity, a convergent sequence is bounded, say $\|v_n\| \le C$ for all n. Then, adding and subtracting $\langle u|v_n\rangle$ and using Equations (5.35) and (5.37),

$$|\langle u_n|v_n\rangle - \langle u|v\rangle| \le |\langle u_n - u|v_n\rangle| + |\langle u|v_n - v\rangle| \le C\ \|u_n - u\| + \|u\|\ \|v_n - v\|\ ,$$

by Equation (12.1) applied to each term, and both summands tend to zero. □

Corollary 12.5 (Continuity of the norm and of orthogonality)**.** *The norm is continuous, $|\|u\| - \|v\|| \le \|u - v\|$; and if $x_n \to x$ with $\langle y|x_n\rangle = 0$ for all n, then $\langle y|x\rangle = 0$. Orthogonality to a fixed vector is therefore a closed condition.*

Rests on Proposition 12.4 and Definition 5.22.

Proof. *Derives Corollary 12.5.*

The first is the triangle inequality Equation (5.42) used twice, $\|u\| \le \|u - v\| + \|v\|$ and symmetrically. The second is Proposition 12.4 with the constant sequence y in the first slot. □

Proposition 12.6 (Parallelogram law and polarization)**.** *For all u, v in an inner-product space,*

$$\|u + v\|^2 + \|u - v\|^2 = 2\,\|u\|^2 + 2\,\|v\|^2\ , \tag{12.2}$$

and the inner product is recovered from the norm alone by

$$\langle u|v\rangle = \frac{1}{4}\sum_{k=0}^{3} \mathrm{i}^{-k}\left\|u+\mathrm{i}^{k}v\right\|^{2}. \tag{12.3}$$

Rests on Definition 5.17 and Equation (5.44).

Proof. *Derives Proposition 12.6.*

Expanding by Equations (5.35) and (5.37) and the scalar rules Equations (5.34) and (5.36), for any $\lambda \in \mathbb{C}$,

$$\|u+\lambda v\|^{2} = \|u\|^{2} + |\lambda|^{2}\|v\|^{2} + \lambda\langle u|v\rangle + \lambda^{*}\langle v|u\rangle . \tag{12.4}$$

Taking $\lambda = +1$ and $\lambda = -1$ and adding cancels the two cross terms and gives Equation (12.2).

For Equation (12.3), put $\lambda = \mathrm{i}^{k}$ in Equation (12.4), multiply by i^{-k} and sum over $k = 0, 1, 2, 3$. The first two terms carry $\sum_k \mathrm{i}^{-k} = 0$ and vanish. The third carries $\sum_k \mathrm{i}^{-k}\mathrm{i}^{k} = 4$ and leaves $4\langle u|v\rangle$. The fourth carries $\sum_k \mathrm{i}^{-k}\mathrm{i}^{-k} = \sum_k (-1)^{k} = 0$ and vanishes. □

Remark 12.7 (Which norms come from an inner product). Equation (12.2) is a genuine test, not a mere identity: a norm satisfies it if and only if it is the norm associated with an inner product, and Equation (12.3) then constructs that inner product from the norm. The supremum norm on continuous functions fails it — take u, v two bumps with disjoint supports and equal heights, so that all four norms in Equation (12.2) equal that height and the identity reads $2 = 4$ — which is why the space of continuous functions with the supremum norm is a Banach space and not a Hilbert space. The whole of the geometry below, orthogonality and projection included, is unavailable there.

Proposition 12.8 (Absolutely convergent series test)**.** *A normed space is complete if and only if every series $\sum_n x_n$ with $\sum_n \|x_n\| < \infty$ converges in it.*

Rests on Definitions 5.18 and 6.27.

Proof. *Derives Proposition 12.8.*

($\Rightarrow$) The partial sums $S_N = \sum_{n\le N} x_n$ satisfy $\|S_M - S_N\| \le \sum_{n=N+1}^{M} \|x_n\|$ for $M > N$, which is a tail of a convergent series of non-negative reals and so tends to zero; (S_N) is Cauchy and converges by completeness.

($\Leftarrow$) Let (y_n) be Cauchy. Choose $n_1 < n_2 < \cdots$ with $\|y_m - y_{n_k}\| < 2^{-k}$ for all $m \ge n_k$. Then $x_1 = y_{n_1}$ and $x_{k+1} = y_{n_{k+1}} - y_{n_k}$ have $\sum_k \|x_k\| \le \|y_{n_1}\| + \sum_k 2^{-k} < \infty$, so by hypothesis the telescoping series converges, that is, $y_{n_k} \to y$ for some y. A Cauchy sequence with a convergent subsequence converges to the same limit, by the triangle inequality. □

Two examples carry the whole subject, and Theorem 12.33 will show that they are one example seen twice.

Example 12.9 (The sequence space ℓ^2)**.** Let

$$\ell^{2} = \left\{ x = (x_1, x_2, \dots) \in \mathbb{C}^{\mathbb{N}} \;\middle|\; \sum_{n=1}^{\infty} |x_n|^{2} < \infty \right\}, \qquad \langle x|y\rangle = \sum_{n=1}^{\infty} x_n^{*} y_n . \tag{12.5}$$

This is a Hilbert space. The series defining $\langle x|y\rangle$ converges absolutely, because $|x_n^{*}y_n| \le \frac{1}{2}(|x_n|^{2} + |y_n|^{2})$; the five axioms of Definition 5.17 are inherited termwise from $\mathbb{C}$; and ℓ^2 is a vector space because the same bound applied to $|x_n + y_n|^{2} \le 2|x_n|^{2} + 2|y_n|^{2}$ keeps the sum finite.

Rests on Definitions 5.17 and 12.2.

Proposition 12.10 (ℓ^2 is complete)**.** *The space Equation (12.5) is complete.*

Rests on Example 12.9 and Axiom 7.1.

Proof. *Derives Proposition 12.10.*

Let $(x^{(m)})_m$ be Cauchy in ℓ^2. For each fixed n, $\left|x_n^{(m)} - x_n^{(p)}\right| \le \left\|x^{(m)} - x^{(p)}\right\|$, so $(x_n^{(m)})_m$ is Cauchy in $\mathbb{C}$ and converges to some x_n by the completeness of $\mathbb{R}$ (Theorem 7.8, applied to real and imaginary parts). Given $\varepsilon > 0$ choose M with $\left\|x^{(m)} - x^{(p)}\right\| < \varepsilon$ for $m, p \ge M$. For every N and every $m, p \ge M$,

$$\sum_{n=1}^{N} \left|x_n^{(m)} - x_n^{(p)}\right|^2 \le \varepsilon^2 ,$$

a finite sum, in which we may let $p \to \infty$ termwise to get $\sum_{n=1}^{N} \left|x_n^{(m)} - x_n\right|^2 \le \varepsilon^2$; since N is arbitrary, $\left\|x^{(m)} - x\right\| \le \varepsilon$. In particular $x = x^{(M)} - (x^{(M)} - x)$ lies in ℓ^2, and $x^{(m)} \to x$. □

Example 12.11 (The function space L^2)**.** Let $\Omega \subseteq \mathbb{R}^N$ be measurable and let $L^2(\Omega)$ be the set of measurable $f : \Omega \longrightarrow \mathbb{C}$ with $\int_\Omega |f|^2 < \infty$, two functions being identified when they agree outside a set of measure zero, with

$$\langle f | g \rangle = \int_\Omega f(x)^* \, g(x) \, \mathrm{d}^N x \, . \tag{12.6}$$

The identification is not a convenience but a necessity: without it $\langle f|f\rangle = 0$ would not force $f = 0$, and Equation (5.31) would fail. An element of L^2 is therefore an equivalence class, and a statement about "the value $f(x_0)$" is meaningless in L^2 — a point that returns in Section 12.6.3, where the delta functional is shown not to be an element of L^2 at all.

Rests on Definitions 5.17 and 12.2.

Theorem 12.12 (Riesz–Fischer)**.** *$L^2(\Omega)$ is complete, and is therefore a Hilbert space.*

Rests on Example 12.11 and Proposition 12.8.

Proof. *Derives Theorem 12.12.*

By Proposition 12.8 it suffices to sum an absolutely convergent series: let $\sum_k \|f_k\| = S < \infty$ and put $g_K = \sum_{k \le K} |f_k|$, an increasing sequence of non-negative measurable functions with $\|g_K\| \le S$ by the triangle inequality. By monotone convergence (Remark 12.1) $g_K \uparrow g$ pointwise with $\int |g|^2 \le S^2$, so g is finite outside a null set. Where g is finite the series $\sum_k f_k(x)$ converges absolutely in $\mathbb{C}$; call its sum $f(x)$, defined almost everywhere, and note $|f| \le g$, so $f \in L^2$. Finally $\left|f - \sum_{k \le K} f_k\right|^2 \le g^2$ pointwise and tends to zero almost everywhere, so Fatou's lemma applied to $g^2 - \left|f - \sum_{k \le K} f_k\right|^2$ gives $\left\|f - \sum_{k \le K} f_k\right\| \longrightarrow 0$. □

The theorem is the joint work of Riesz [336] and Fischer [335], published within weeks of one another in 1907, and it is what made L^2 usable: before it, "convergence in the mean" had no limit to converge to. Theorem 12.33 will show that ℓ^2 and L^2 are not two examples but one, which is the mathematical content of the equivalence of matrix and wave mechanics.

12.2 Orthonormal bases and separability

12.2.1 Orthogonality and projections

In finite dimension every subspace splits off an orthogonal complement, $\mathbb{V} = \mathbb{S} \oplus \mathbb{S}^\perp$ (Corollary 5.27), and the proof simply orthonormalises a basis of $\mathbb{S}$. In infinite dimension that argument stops at the first step, because $\mathbb{S}$ need have no finite basis, and the conclusion itself becomes false for a general subspace: Remark 5.28 flagged the gap and reserved its repair for this section. The repair rests on completeness and on the parallelogram law, and it begins with a purely metric statement about convex sets.

Definition 12.13 (Convex set)**.** A subset C of a vector space is *convex* if $tx + (1-t)y \in C$ whenever $x, y \in C$ and $t \in [0,1]$. Every subspace, and every translate of a subspace, is convex.

Rests on Definition 5.6.

Theorem 12.14 (Closest point in a closed convex set)**.** *Let $C \subseteq \mathcal{H}$ be non-empty, closed and convex, and let $x \in \mathcal{H}$. Then there is exactly one $y \in C$ with*

$$\|x - y\| = d\,, \qquad d = \inf_{z \in C} \|x - z\|\ . \tag{12.7}$$

Rests on Definitions 12.2 and 12.13 and Proposition 12.6.

Proof. *Derives Theorem 12.14.*

Apply Equation (12.2) to the vectors $u = x - z_1$ and $v = x - z_2$, for $z_1, z_2 \in C$. Then $u + v = 2x - (z_1 + z_2)$ and $u - v = z_2 - z_1$, so

$$\|z_1 - z_2\|^2 = 2\,\|x - z_1\|^2 + 2\,\|x - z_2\|^2 - 4\left\|x - \tfrac{1}{2}(z_1 + z_2)\right\|^2\ . \tag{12.8}$$

The midpoint $\frac{1}{2}(z_1 + z_2)$ lies in C by convexity, so the last norm is at least d and

$$\|z_1 - z_2\|^2 \le 2\,\|x - z_1\|^2 + 2\,\|x - z_2\|^2 - 4d^2\,.$$

Take a minimising sequence (z_n) in C with $\|x - z_n\| \to d$: the right-hand side, with $z_1 = z_n$ and $z_2 = z_m$, tends to $2d^2 + 2d^2 - 4d^2 = 0$ as $n, m \to \infty$, so (z_n) is Cauchy and converges, by Definition 12.2, to some y, which lies in C because C is closed. Continuity of the norm (Corollary 12.5) gives $\|x - y\| = d$.

Uniqueness is the same estimate: if y and y' both realise the infimum, then $\|y - y'\|^2 \le 2d^2 + 2d^2 - 4d^2 = 0$. □

Remark 12.15 (Where each hypothesis is used). All three hypotheses are needed and each fails somewhere. Without closedness the infimum need not be attained (take C an open ball). Without convexity the minimiser need not be unique (take C a sphere and x its centre). Without completeness the Cauchy sequence has nowhere to go: on the pre-Hilbert space of continuous functions in $L^2[-1,1]$, the closed convex set of functions with $\int_{-1}^{0} f = \int_{0}^{1} f$ has no closest point to a suitable x. It is completeness that makes Theorem 12.14 a theorem about Hilbert spaces rather than about inner-product spaces, and everything in this section inherits that.

Definition 12.16 (Orthogonal complement)**.** For any subset $S \subseteq \mathcal{H}$,

$$S^{\perp} = \{\, y \in \mathcal{H} \mid \langle x | y \rangle = 0 \ \ \forall\, x \in S \,\}\ . \tag{12.9}$$

Rests on Definition 5.22 and Equation (5.52).

Proposition 12.17 (The complement is always a closed subspace)**.** *For any $S \subseteq \mathcal{H}$, $S^{\perp}$ is a closed subspace, $S \subseteq S^{\perp\perp}$, and $S^{\perp} = \overline{\langle S \rangle}^{\perp}$, where $\overline{\langle S \rangle}$ is the closure of the set of finite linear combinations of elements of S. Moreover $S \cap S^{\perp} \subseteq \{0\}$.*

Rests on Definition 12.16 and Corollary 12.5.

Proof. *Derives Proposition 12.17.*

$S^{\perp}$ is a subspace by linearity of the bracket in its second slot, Equations (5.34) and (5.35), and it is closed as an intersection over $x \in S$ of the sets $\{y \mid \langle x|y\rangle = 0\}$, each closed by Corollary 12.5. If $y \in S^{\perp}$ then y is orthogonal to every finite combination of elements of S, again by linearity, and hence to every limit of such combinations, again by Corollary 12.5; the reverse inclusion is immediate from $S \subseteq \overline{\langle S \rangle}$. That $S \subseteq S^{\perp\perp}$ is a restatement of the definition read the other way round, using hermiticity Equation (5.33). Finally $x \in S \cap S^{\perp}$ gives $\langle x|x\rangle = 0$, so $x = 0$ by Equation (5.31). □

Theorem 12.18 (Projection theorem)**.** *Let $M \subseteq \mathcal{H}$ be a* closed *subspace. Then*

$$\mathcal{H} = M \oplus M^{\perp}\,, \tag{12.10}$$

that is, every $x \in \mathcal{H}$ has a unique decomposition $x = x_{\parallel} + x_{\perp}$ with $x_{\parallel} \in M$ and $x_{\perp} \in M^{\perp}$; moreover $x_{\parallel}$ is the closest point of M to x, and $\|x\|^2 = \left\|x_{\parallel}\right\|^2 + \|x_{\perp}\|^2$.

Rests on Theorem 12.14 and Proposition 12.17.

Proof. *Derives Theorem 12.18.*

A subspace is convex and M is closed and contains 0, so Theorem 12.14 supplies a unique closest point $x_{\parallel} \in M$; put $x_{\perp} = x - x_{\parallel}$.

We claim $x_{\perp} \in M^{\perp}$. Let $m \in M$ with $m \neq 0$ and let $t \in \mathbb{C}$. Since $x_{\parallel} + tm \in M$, minimality gives $\|x_{\perp} - tm\|^2 \geq \|x_{\perp}\|^2$, while Equation (12.4) with $\lambda = -t$ reads

$$\|x_{\perp} - tm\|^2 = \|x_{\perp}\|^2 + |t|^2\,\|m\|^2 - t\,\langle x_{\perp}|m\rangle - t^*\,\langle m|x_{\perp}\rangle\ . \tag{12.11}$$

Choose $t = \langle m|x_{\perp}\rangle\,/\,\|m\|^2$. Using $\langle x_{\perp}|m\rangle = \langle m|x_{\perp}\rangle^*$ (Equation (5.33)), each of the last three terms of Equation (12.11) equals $|\langle m|x_{\perp}\rangle|^2\,/\,\|m\|^2$ with signs $+, -, -$, so

$$\|x_{\perp} - tm\|^2 = \|x_{\perp}\|^2 - \frac{|\langle m|x_{\perp}\rangle|^2}{\|m\|^2}\,.$$

Comparison with the minimality inequality forces $\langle m|x_{\perp}\rangle = 0$, and m was arbitrary.

Uniqueness: if $x = m + w = m' + w'$ with $m, m' \in M$ and $w, w' \in M^{\perp}$, then $m - m' = w' - w$ lies in $M \cap M^{\perp} = \{0\}$ by Proposition 12.17. The norm identity is Equation (12.4) with $\lambda = 1$ and the cross terms vanishing — the Pythagorean theorem. □

Corollary 12.19 (Double complement; the density criterion)**.** *For a closed subspace M, $M^{\perp\perp} = M$. For an arbitrary subset S, $S^{\perp\perp} = \overline{\langle S\rangle}$. In particular a subspace $D \subseteq \mathcal{H}$ is dense if and only if $D^{\perp} = \{0\}$.*

Rests on Theorem 12.18 and Proposition 12.17.

Proof. *Derives Corollary 12.19.*

$M \subseteq M^{\perp\perp}$ always (Proposition 12.17). Conversely let $x \in M^{\perp\perp}$ and split $x = x_{\parallel} + x_{\perp}$ by Theorem 12.18. Then $x_{\perp} = x - x_{\parallel} \in M^{\perp\perp}$, since $M^{\perp\perp}$ is a subspace containing M; but also $x_{\perp} \in M^{\perp}$, and $M^{\perp} \cap M^{\perp\perp} = \{0\}$. Hence $x = x_{\parallel} \in M$.

For general S, apply this to the closed subspace $M = \overline{\langle S\rangle}$ and use $S^{\perp} = M^{\perp}$ from Proposition 12.17. The density criterion is the case $S = D$: D is dense exactly when $\overline{\langle D\rangle} = \mathcal{H}$, i.e. when $D^{\perp\perp} = \mathcal{H}$, i.e. when $D^{\perp} = \{0\}$. □

The density criterion of Corollary 12.19 is used more often below than the projection theorem itself: it converts every question of the form "is this family total?" into the computation of an orthogonal complement, and it is the standard way to test whether a domain is dense (Section 12.5.1).

Definition 12.20 (Orthogonal projection operator)**.** For M a closed subspace, the map $P_M : \mathcal{H} \longrightarrow \mathcal{H}$, $P_M x = x_{\parallel}$, given by Theorem 12.18, is the *orthogonal projection* onto M.

Rests on Theorem 12.18.

Proposition 12.21 (Characterization of orthogonal projections)**.** *$P = P_M$ is linear and satisfies*

$$P^2 = P\,, \qquad \langle Px|y\rangle = \langle x|Py\rangle \quad \forall\, x, y \in \mathcal{H}\,, \tag{12.12}$$

with $\|Px\| \leq \|x\|$, image M and kernel $M^{\perp}$. Conversely, if $P : \mathcal{H} \longrightarrow \mathcal{H}$ is linear and satisfies Equation (12.12), then its image is a closed subspace M and $P = P_M$.

Rests on Definition 12.20 and Theorem 12.18.

Proof. *Derives Proposition 12.21.*

Linearity. If $x = m + w$ and $x' = m' + w'$ are the decompositions of Theorem 12.18, then $ax + bx' = (am + bm') + (aw + bw')$ is a decomposition of $ax + bx'$ of the required kind, and by uniqueness it is the one; so $P(ax + bx') = aPx + bPx'$.

Idempotence and norm. $Px \in M$, and a vector of M is its own closest point, so $P^2 = P$. The norm identity of Theorem 12.18 gives $\|Px\| \le \|x\|$. The image is M and $Px = 0$ exactly when $x = x_\perp \in M^\perp$.

Symmetry. Write $x = Px + (x - Px)$ and $y = Py + (y - Py)$ with the second term of each in $M^\perp$. Then $\langle Px|y\rangle = \langle Px|Py\rangle$ and $\langle x|Py\rangle = \langle Px|Py\rangle$, because in each case the omitted term pairs a vector of M with a vector of $M^\perp$.

Converse. Let $M = \operatorname{im} P$. If $x \in M$, say $x = Pz$, then $Px = P^2 z = Pz = x$, so $M = \{x \mid Px = x\}$. This set is closed, and the argument needs no boundedness of P: if $x_n \in M$ and $x_n \to x$, then for every y the symmetry in Equation (12.12) gives $\langle x_n|Py - y\rangle = \langle Px_n|y\rangle - \langle x_n|y\rangle = 0$, whence $\langle x|Py - y\rangle = 0$ by Proposition 12.4, and reading the symmetry backwards, $\langle Px - x|y\rangle = 0$ for all y, so $Px = x$. For any x and any $m \in M$, $\langle m|x - Px\rangle = \langle Pm|x - Px\rangle = \langle m|Px - P^2 x\rangle = 0$, using Equation (12.12) twice; so $x - Px \in M^\perp$ and Px is the M-component of x. □

Remark 12.22 (Yes–no questions). Equation (12.12) is the whole algebraic content of an orthogonal projection, and it is what makes projections the carriers of propositions in quantum mechanics: $P^2 = P$ says that asking the same question twice adds nothing, and the symmetry says the question has real answers. Two projections commute if and only if the corresponding subspaces decompose compatibly, and non-commuting projections are the formal seat of complementarity. The second identity of Equation (12.12) is exactly self-adjointness, $P^\dagger = P$, once the adjoint is available (Theorem 12.38); it is written out here because the adjoint is defined two sections later and nothing above depends on it. The use made of all this is in Chapter 77 [706].

12.2.2 Gram–Schmidt orthogonalization

Proposition 12.23 (Gram–Schmidt in a Hilbert space). *Let $\{v_1, v_2, \dots\}$ be a finite or countable linearly independent family in an inner-product space. Then there is an orthonormal family $\{e_1, e_2, \dots\}$ with*

$$\langle\{e_1, \dots, e_k\}\rangle = \langle\{v_1, \dots, v_k\}\rangle\,, \qquad k = 1, 2, \dots\,, \tag{12.13}$$

and consequently $\overline{\langle\{e_k\}\rangle} = \overline{\langle\{v_k\}\rangle}$.

Rests on Proposition 5.26 and Definition 5.25.

Proof. *Derives Proposition 12.23.*

The recursion Equation (5.51) is finite at every step, so Proposition 5.26 constructs $e_1, \dots, e_k$ for each k and the construction at step k does not disturb its predecessors; Remark 5.28 already recorded that this much survives to a countable family. Only the last clause is new. By Equation (12.13) the two increasing unions $\bigcup_k \langle\{e_1, \dots, e_k\}\rangle$ and $\bigcup_k \langle\{v_1, \dots, v_k\}\rangle$ coincide, and each is exactly the set of finite linear combinations of the corresponding family; equal sets have equal closures. □

Corollary 12.24 (Separable spaces have countable orthonormal families). *Let $\mathcal{H}$ be an inner-product space containing a countable dense subset. Then there is a finite or countable orthonormal family $\{e_k\}$ with $\overline{\langle\{e_k\}\rangle} = \mathcal{H}$.*

Rests on Proposition 12.23 and Definition 6.27.

Proof. *Derives Corollary 12.24.*

Let $\{w_1, w_2, \dots\}$ be dense. Discard w_n whenever it lies in $\langle\{w_1, \dots, w_{n-1}\}\rangle$; the surviving subfamily $\{v_k\}$ is linearly independent and spans the same set of finite linear combinations, so $\overline{\langle\{v_k\}\rangle} \supseteq \{w_n\} = \mathcal{H}$. Now apply Proposition 12.23. □

Example 12.25 (Orthogonal polynomials)**.** Take $\mathcal{H} = L^2([-1,1])$ and apply Proposition 12.23 to the monomials $v_k(x) = x^{k-1}$, which are linearly independent because a non-trivial polynomial has finitely many roots and so cannot vanish almost everywhere. The output is, up to normalisation, the sequence of Legendre polynomials; running the same recursion in L^2 of a weighted measure $w(x)\,\mathrm{d}x$ produces the Hermite, Laguerre and Chebyshev families according to the weight. That these are the eigenfunctions of the corresponding second-order differential operators is the content of Sturm–Liouville theory (Chapter 9); that the resulting family is *total*, so that the expansion of Section 12.2.3 converges to the function expanded, is a separate fact about each weight and is not supplied by Gram–Schmidt.
Rests on Proposition 12.23 and Example 12.11.

The construction is due to Gram [223], who introduced it for least-squares approximation by series of functions, and to Schmidt [338], who put it in the form used here in the course of his work on integral equations.

12.2.3 Fourier expansion, Bessel and Parseval

Definition 12.26 (Orthonormal system; Fourier coefficients)**.** A finite or countable family $\{e_n\}$ in $\mathcal{H}$ with $\langle e_n | e_m \rangle = \delta_{nm}$ is an *orthonormal system*. For $x \in \mathcal{H}$ the scalars

$$c_n = \langle e_n | x \rangle \tag{12.14}$$

are the *Fourier coefficients* of x in that system.
Rests on Definition 5.25 and Equation (5.49).

Proposition 12.27 (Best approximation and Bessel's inequality)**.** *Let $\{e_n\}$ be an orthonormal system and $x \in \mathcal{H}$. For every N the partial sum $S_N = \sum_{n \le N} c_n e_n$ is the orthogonal projection of x onto $\langle \{e_1, \dots, e_N\} \rangle$, hence the unique closest point of that subspace to x, and*

$$\sum_{n} |\langle e_n | x \rangle|^2 \le \|x\|^2 \,. \tag{12.15}$$

Rests on Definition 12.26 and Theorem 12.18.

Proof. *Derives Proposition 12.27.*
Expanding with Equations (5.35) and (5.37) and using orthonormality,

$$\|x - S_N\|^2 = \|x\|^2 - \sum_{n \le N} |c_n|^2 \,, \tag{12.16}$$

since each of the three quantities $\langle x | S_N \rangle$, $\langle S_N | x \rangle$ and $\|S_N\|^2$ equals $\sum_{n \le N} |c_n|^2$ — for the first, $\langle x | S_N \rangle = \sum_{n \le N} c_n \langle x | e_n \rangle = \sum_{n \le N} c_n c_n^*$ by Equation (5.33), and the other two are the same computation — so that the two cross terms together cancel one copy of $\|S_N\|^2$. The same computation gives $\langle e_m | x - S_N \rangle = 0$ for $m \le N$, so $x - S_N$ is orthogonal to the finite-dimensional — hence closed — subspace $\langle \{e_1, \dots, e_N\} \rangle$, and S_N is its orthogonal projection of x by the uniqueness in Theorem 12.18. The left side of Equation (12.16) is non-negative, so $\sum_{n \le N} |c_n|^2 \le \|x\|^2$ for every N; a bounded increasing sequence of partial sums converges, by Axiom 7.1, and its limit obeys the same bound. □

Bessel's inequality says that the coefficients of any vector are square summable. The converse question — which square-summable coefficient sequences arise — has the best possible answer, and it is completeness that supplies it.

Proposition 12.28 (Convergence criterion for orthogonal series)**.** *Let $\{e_n\}$ be an orthonormal system in a Hilbert space and let $c_n \in \mathbb{C}$. The series $\sum_n c_n e_n$ converges in $\mathcal{H}$ if and only if $\sum_n |c_n|^2 < \infty$, and then $\|\sum_n c_n e_n\|^2 = \sum_n |c_n|^2$ and $\langle e_m | \sum_n c_n e_n \rangle = c_m$.*
Rests on Definitions 12.2 and 12.26.

Proof. *Derives Proposition 12.28.*

By orthonormality and the Pythagorean identity, for $M > N$,

$$\left\| \sum_{n=N+1}^{M} c_n e_n \right\|^2 = \sum_{n=N+1}^{M} |c_n|^2 \,. \tag{12.17}$$

So the partial sums of $\sum_n c_n e_n$ form a Cauchy sequence in $\mathcal{H}$ if and only if the partial sums of $\sum_n |c_n|^2$ form a Cauchy sequence in $\mathbb{R}$. By Definition 12.2 and Theorem 7.8 respectively, each is equivalent to the convergence of the corresponding series. The norm identity follows by letting $N = 0$ and $M \to \infty$ in Equation (12.17) and using Corollary 12.5; the coefficient identity follows from Proposition 12.4, which allows $\langle e_m | \cdot \rangle$ to be taken through the limit. □

Definition 12.29 (Orthonormal basis)**.** An orthonormal system $\{e_n\}$ in $\mathcal{H}$ is an *orthonormal basis*, or a *complete* orthonormal system, if it is maximal: the only vector orthogonal to every e_n is 0.

Rests on Definitions 12.16 and 12.26.

Theorem 12.30 (Completeness, expansion, Parseval)**.** *Let $\{e_n\}$ be an orthonormal system in a Hilbert space $\mathcal{H}$. The following are equivalent.*

1. *$\{e_n\}$ is maximal, i.e. an orthonormal basis (Definition 12.29);*
2. *the finite linear combinations of the e_n are dense in $\mathcal{H}$;*
3. *every $x \in \mathcal{H}$ is the sum of its Fourier series,*

$$x = \sum_n \langle e_n | x \rangle \; e_n \,, \tag{12.18}$$

   the series converging in the norm of $\mathcal{H}$;

4. *Parseval's identity holds for every $x \in \mathcal{H}$,*

$$\|x\|^2 = \sum_n |\langle e_n | x \rangle|^2 \,. \tag{12.19}$$

Any of these implies the polarized form

$$\langle x | y \rangle = \sum_n \langle x | e_n \rangle \, \langle e_n | y \rangle \,, \qquad \forall\, x, y \in \mathcal{H} \,. \tag{12.20}$$

Rests on Propositions 12.27 and 12.28 and Corollary 12.19.

Proof. *Derives Theorem 12.30.*

(1) ⇒ (3). By Equation (12.15) the coefficients $c_n = \langle e_n | x \rangle$ are square summable, so by Proposition 12.28 the series $\sum_n c_n e_n$ converges to some $y \in \mathcal{H}$ with $\langle e_m | y \rangle = c_m = \langle e_m | x \rangle$ for every m. Then $x - y$ is orthogonal to every e_m, hence $x = y$ by maximality.

(3) ⇒ (2). The partial sums of Equation (12.18) are finite linear combinations converging to x.

(2) ⇒ (1). If x is orthogonal to every e_n then $x \in \{e_n\}^\perp = \overline{\langle \{e_n\} \rangle}^{\perp} = \mathcal{H}^\perp = \{0\}$, by Proposition 12.17 and density.

(3) ⇒ (4). Take norms in Equation (12.18) using the norm identity of Proposition 12.28.

(4) ⇒ (1). If $x \perp e_n$ for every n, then Equation (12.19) gives $\|x\|^2 = 0$.

Finally Equation (12.20) follows from Equation (12.19) by polarization: apply Equation (12.3) to both sides, each of the four terms of which is an instance of Equation (12.19). Alternatively, substitute Equation (12.18) for y and take $\langle x | \cdot \rangle$ through the sum by Proposition 12.4. □

Remark 12.31 (What "basis" has come to mean). An orthonormal basis in the sense of Definition 12.29 is *not* a basis in the algebraic sense of Definition 5.14: the expansion Equation (12.18) is an infinite sum, a limit in the topology, and not a finite linear combination. The two notions coincide only in finite dimension. An infinite-dimensional Hilbert space does have algebraic bases, but they are uncountable and useless; the topological notion is the one physics means, and Chapter 5 said as much when it reserved this section. The identity Equation (12.20), written in Dirac notation as $\sum_n |e_n\rangle\langle e_n| = \mathbb{1}$, is the completeness relation used constantly from Chapter 77 onwards.

Definition 12.32 (Separable Hilbert space)**.** $\mathcal{H}$ is *separable* if it contains a countable dense subset.

Rests on Definitions 6.27 and 12.2.

Theorem 12.33 (Every separable Hilbert space is ℓ^2)**.** *Let $\mathcal{H}$ be a separable Hilbert space of infinite dimension. Then there is a linear bijection $U : \mathcal{H} \longrightarrow \ell^2$ preserving the inner product, $\langle Ux|Uy\rangle_{\ell^2} = \langle x|y\rangle$. Any two separable infinite-dimensional Hilbert spaces are therefore isomorphic; up to isomorphism there is only one.*

Rests on Corollary 12.24, Theorem 12.30, and Proposition 12.10.

Proof. *Derives Theorem 12.33.*

By Corollary 12.24 there is a countable orthonormal family $\{e_n\}$ whose finite combinations are dense, which is condition (2) of Theorem 12.30; the family is infinite because $\mathcal{H}$ is. Define $Ux = (\langle e_n|x\rangle)_n$. It lands in ℓ^2 by Equation (12.15), is linear because the bracket is linear in its second slot, and is isometric by Equation (12.19); an isometric linear map is injective. It is surjective because a sequence $(c_n) \in \ell^2$ is, by Proposition 12.28, the coefficient sequence of the convergent sum $\sum_n c_n e_n$. Preservation of the inner product is Equation (12.20), or equivalently isometry plus Equation (12.3). Composing the map for one space with the inverse of the map for another gives the last claim. □

Remark 12.34 (Why wave mechanics and matrix mechanics agree). Theorem 12.33 is a mathematical triviality with a large physical consequence. Heisenberg's mechanics is written on ℓ^2 — infinite matrices acting on square-summable coefficient sequences — and Schrödinger's on L^2, and the two were presented as rival theories. They are formulations on the same space: L^2 is separable, so Theorem 12.33 produces a bracket-preserving bijection with ℓ^2, and the map is explicitly the assignment of Fourier coefficients in any orthonormal basis, Equation (12.14). This is the observation von Neumann made the organising principle of his treatise [706], and it is why Chapter 77 may state the postulates on an abstract $\mathcal{H}$ without choosing a representation. The name attached to Equation (12.19) is Parseval's [51], who wrote the identity for trigonometric series more than a century before there was a space to write it on; the general form is inseparable from the Riesz–Fischer theorem [335, 336], of which it is the abstract shadow, and the trigonometric case is worked out in Chapter 17.

12.3 Bounded operators and duality

12.3.1 Bounded operators and their adjoints

In finite dimension every linear map is continuous, so Chapter 5 never had to distinguish the algebra of linear transformations from the analysis of continuous ones. Here they part company, and the first order of business is to show that for linear maps the two words mean the same thing.

Definition 12.35 (Bounded operator; operator norm)**.** A linear map $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ is *bounded* if there is $C \geq 0$ with $\|Ax\| \leq C\,\|x\|$ for all x. The least such C,

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\| \,, \tag{12.21}$$

is the *operator norm*. The bounded operators from $\mathcal{H}$ to itself form a set written $\mathcal{B}(\mathcal{H})$.
Rests on Definitions 5.18 and 5.35.

Proposition 12.36 (Boundedness is continuity)**.** *For a linear $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ the following are equivalent: (i) A is continuous at 0; (ii) A is continuous everywhere; (iii) A is bounded.*
Rests on Definition 12.35 and Proposition 6.28.

Proof. *Derives Proposition 12.36.*

(iii) $\Rightarrow$ (ii): $\|Ax - Ax_0\| = \|A(x - x_0)\| \leq \|A\| \, \|x - x_0\|$, so $\delta = \varepsilon/(1 + \|A\|)$ serves in Equation (6.11). (ii) $\Rightarrow$ (i) is a special case. (i) $\Rightarrow$ (iii): continuity at 0 with $\varepsilon = 1$ gives a $\delta > 0$ with $\|Ax\| \leq 1$ whenever $\|x\| \leq \delta$. For $x \neq 0$ apply this to $\delta x / \|x\|$, which has norm δ: $\|Ax\| \leq \|x\| / \delta$, so A is bounded with $\|A\| \leq 1/\delta$. □

Proposition 12.37 ($\mathcal{B}(\mathcal{H})$ is a Banach algebra)**.** *Equation (12.21) is a norm on $\mathcal{B}(\mathcal{H})$; it is submultiplicative,*

$$\|AB\| \leq \|A\| \, \|B\| \,, \tag{12.22}$$

the identity $\mathbb{1}$ has norm 1, and $\mathcal{B}(\mathcal{H})$ is complete.
Rests on Definition 12.35 and Proposition 12.8.

Proof. *Derives Proposition 12.37.*

The norm axioms Equations (5.40) to (5.42) follow pointwise from those of $\mathcal{H}$ and the definition of a supremum; Equation (12.22) from $\|ABx\| \leq \|A\| \, \|Bx\| \leq \|A\| \, \|B\| \, \|x\|$.

For completeness let (A_n) be Cauchy in $\mathcal{B}(\mathcal{H})$. For each x, $\|A_n x - A_m x\| \leq \|A_n - A_m\| \, \|x\|$, so $(A_n x)$ is Cauchy in $\mathcal{H}$ and converges; call the limit Ax. The map A is linear, being a pointwise limit of linear maps (Proposition 12.4 is not even needed: addition and scalar multiplication are continuous). Given $\varepsilon > 0$ pick N with $\|A_n - A_m\| \leq \varepsilon$ for $n, m \geq N$; then $\|A_n x - A_m x\| \leq \varepsilon \, \|x\|$, and letting $m \to \infty$ with x fixed gives $\|A_n x - Ax\| \leq \varepsilon \, \|x\|$. Hence $A - A_N$ is bounded, so A is, and $\|A_n - A\| \leq \varepsilon$ for $n \geq N$. □

Theorem 12.38 (Existence and uniqueness of the adjoint)**.** *For every $A \in \mathcal{B}(\mathcal{H})$ there is exactly one $A^\dagger \in \mathcal{B}(\mathcal{H})$ with*

$$\left\langle A^\dagger x \middle| y \right\rangle = \langle x | Ay \rangle \,, \qquad \forall \, x, y \in \mathcal{H} \,, \tag{12.23}$$

and $\left\| A^\dagger \right\| = \|A\|$.
Rests on Definition 5.39, Theorem 12.46, and Proposition 12.37.

Proof. *Derives Theorem 12.38.*

Fix x. The map $y \longmapsto \langle x | Ay \rangle$ is linear, by Equations (5.34) and (5.35) and linearity of A, and bounded, $|\langle x | Ay \rangle| \leq \|x\| \, \|A\| \, \|y\|$ by Equation (12.1). By the Riesz representation theorem (Theorem 12.46) there is a unique vector, call it $A^\dagger x$, with $\left\langle A^\dagger x \middle| y \right\rangle = \langle x | Ay \rangle$ for all y. No circularity is involved: Theorem 12.46 is proved in Section 12.3.2 from Theorem 12.18 alone and uses nothing from the present subsection. Uniqueness of $A^\dagger x$ is uniqueness in Riesz's theorem; linearity of $x \longmapsto A^\dagger x$ follows because $\left\langle A^\dagger(ax_1 + bx_2) - aA^\dagger x_1 - bA^\dagger x_2 \middle| y \right\rangle = 0$ for every y, using Equations (5.36) and (5.37), so the vector in the first slot is 0 by Equation (5.31).

For the norm, Equation (12.23) with $y = A^\dagger x$ gives

$$\left\| A^\dagger x \right\|^2 = \left\langle A^\dagger x \middle| A^\dagger x \right\rangle = \left\langle x \middle| AA^\dagger x \right\rangle \leq \|x\| \, \|A\| \, \left\| A^\dagger x \right\| \,,$$

whence $\left\| A^\dagger x \right\| \leq \|A\| \, \|x\|$ and $\left\| A^\dagger \right\| \leq \|A\|$. Reading Equation (12.23) backwards shows $(A^\dagger)^\dagger = A$, and the same estimate then gives the reverse inequality. □

Equation (12.23) is Equation (5.69) verbatim; what has changed is the proof. In finite dimension Proposition 5.40 built $A^\dagger$ from an orthonormal basis and read off the conjugate transpose. Here there is no finite basis to write a matrix in, and the existence of $A^\dagger$ is a theorem of analysis, resting on completeness through the projection theorem.

Proposition 12.39 (Algebra of the adjoint; the C^* identity)**.** *For $A, B \in \mathcal{B}(\mathcal{H})$ and $a \in \mathbb{C}$:*

$$(A+B)^\dagger = A^\dagger + B^\dagger\,, \quad (aA)^\dagger = a^* A^\dagger\,, \quad (AB)^\dagger = B^\dagger A^\dagger\,, \quad (A^\dagger)^\dagger = A\,, \tag{12.24}$$

and

$$\left\| A^\dagger A \right\| = \|A\|^2\,. \tag{12.25}$$

Moreover $\ker A^\dagger = (\operatorname{im} A)^\perp$.

Rests on Theorem 12.38 and Proposition 12.37.

Proof. *Derives Proposition 12.39.*

Each identity in Equation (12.24) is verified by checking that both sides satisfy Equation (12.23) for the operator in question and invoking uniqueness; for the product, $\left\langle B^\dagger A^\dagger x \middle| y \right\rangle = \left\langle A^\dagger x \middle| By \right\rangle = \langle x | ABy \rangle$.

For Equation (12.25), one inequality is Equation (12.22) together with $\left\|A^\dagger\right\| = \|A\|$. For the other, for any x with $\|x\| = 1$,

$$\|Ax\|^2 = \langle Ax | Ax \rangle = \left\langle A^\dagger A x \middle| x \right\rangle \le \left\| A^\dagger A x \right\| \, \|x\| \le \left\| A^\dagger A \right\|\,,$$

using Equation (12.23) and Equation (12.1); taking the supremum over such x gives $\|A\|^2 \le \left\|A^\dagger A\right\|$.

Finally $x \in \ker A^\dagger$ means $\left\langle A^\dagger x \middle| y \right\rangle = 0$ for all y, that is, $\langle x | Ay \rangle = 0$ for all y, that is, $x \in (\operatorname{im} A)^\perp$. □

Remark 12.40 (Why $\mathcal{B}(\mathcal{H})$ is the model C^*-algebra). A Banach algebra with an involution obeying Equation (12.24) and the norm identity Equation (12.25) is called a C^*-algebra, and Propositions 12.37 and 12.39 say that $\mathcal{B}(\mathcal{H})$ is one. The identity is far stronger than the inequality $\left\|A^\dagger A\right\| \le \|A\| \left\|A^\dagger\right\|$ that any Banach $*$-algebra has: it ties the norm to the algebra so tightly that a C^*-algebra admits exactly one norm making it one, so the analytic structure is determined by the algebraic. That rigidity is what the algebraic formulations of quantum theory (Chapter 109) trade on. This treatise does not develop C^*-algebras; Equation (12.25) is proved here because it is used directly — it is what makes $\|A^2\| = \|A\|^2$ for self-adjoint A, and hence what controls the functional calculus of Section 12.4.2.

Definition 12.41 (The operator classes)**.** Let $A \in \mathcal{B}(\mathcal{H})$. It is

1. *self-adjoint* (or hermitian) if $A^\dagger = A$;
2. *unitary* if $A^\dagger A = A A^\dagger = \mathbb{1}$;
3. *normal* if $A^\dagger A = A A^\dagger$;
4. *positive*, written $A \ge 0$, if $\langle x | Ax \rangle \ge 0$ for every x;
5. a *projection* if $A^2 = A = A^\dagger$ (Proposition 12.21);
6. *compact* if the image under A of the unit ball $\{x \mid \|x\| \le 1\}$ has compact closure (Definition 6.9).

Rests on Theorem 12.38, Definition 6.9, and Proposition 12.21.

Proposition 12.42 (Elementary consequences)**.** *(i) Self-adjoint and unitary operators are normal, and a projection is self-adjoint and positive. (ii) A unitary U preserves the inner product, $\langle Ux|Uy\rangle = \langle x|y\rangle$, and $\|U\| = 1$ when $\mathcal{H} \neq \{0\}$. (iii) On a complex Hilbert space, A is self-adjoint if and only if $\langle x|Ax\rangle \in \mathbb{R}$ for every x; in particular a positive operator is self-adjoint. (iv) $A^\dagger A$ is positive for every A.*

Rests on Definition 12.41 and Proposition 12.39.

Proof. *Derives Proposition 12.42.*

(i) and (iv) are immediate from the definitions, (iv) because $\left\langle x\middle|A^\dagger Ax\right\rangle = \|Ax\|^2 \geq 0$. (ii) $\langle Ux|Uy\rangle = \left\langle U^\dagger Ux\middle|y\right\rangle = \langle x|y\rangle$, and $y = x$ gives $\|Ux\| = \|x\|$.

(iii) If $A^\dagger = A$ then $\langle x|Ax\rangle = \langle Ax|x\rangle = \langle x|Ax\rangle^*$ by Equation (5.33), so the number is real. Conversely suppose $\langle x|Ax\rangle$ is real for every x, and put $B = A - A^\dagger$, so that $\langle x|Bx\rangle = \langle x|Ax\rangle - \langle Ax|x\rangle = 0$ for every x. Replacing x by $x + y$ and by $x + \mathrm{i}y$ and expanding as in Equation (12.4) gives $\langle x|By\rangle + \langle y|Bx\rangle = 0$ and $\langle x|By\rangle - \langle y|Bx\rangle = 0$, whence $\langle x|By\rangle = 0$ for all x, y and $B = 0$. (The complex field is essential: on a real Hilbert space a rotation by a quarter turn in a plane has $\langle x|Ax\rangle = 0$ identically and is not self-adjoint.) □

Proposition 12.43 (Norm of a self-adjoint operator)**.** *If $A \in \mathcal{B}(\mathcal{H})$ is self-adjoint then*

$$\|A\| = \sup_{\|x\|=1} |\langle x|Ax\rangle| \,. \tag{12.26}$$

In particular a self-adjoint operator with $\langle x|Ax\rangle = 0$ for all x is 0.

Rests on Propositions 12.39 and 12.42.

Proof. *Derives Proposition 12.43.*

Write M for the supremum. That $M \leq \|A\|$ is Equation (12.1). For the converse, note first that $|\langle z|Az\rangle| \leq M\|z\|^2$ for every z, by homogeneity. Since A is self-adjoint, $\langle y|Ax\rangle = \langle Ay|x\rangle = \langle x|Ay\rangle^*$, so

$$\langle x+y|A(x+y)\rangle - \langle x-y|A(x-y)\rangle = 2\langle x|Ay\rangle + 2\langle y|Ax\rangle = 4\,\mathrm{Re}\,\langle x|Ay\rangle \,.$$

Bounding each term on the left and using Equation (12.2),

$$4\,|\mathrm{Re}\,\langle x|Ay\rangle| \leq M\left(\|x+y\|^2 + \|x-y\|^2\right) = 2M\left(\|x\|^2 + \|y\|^2\right) .$$

Take $\|x\| = \|y\| = 1$, so $|\mathrm{Re}\,\langle x|Ay\rangle| \leq M$; replacing x by $\mathrm{e}^{\mathrm{i}\theta}x$ for a suitable θ makes $\langle x|Ay\rangle$ real and non-negative, giving $|\langle x|Ay\rangle| \leq M$. Now take the supremum over unit x: if $Ay \neq 0$ the choice $x = Ay/\|Ay\|$ gives $|\langle x|Ay\rangle| = \|Ay\|$, so $\|Ay\| \leq M$, and if $Ay = 0$ that bound is trivial. Hence $\|Ay\| \leq M$ for every unit y, i.e. $\|A\| \leq M$. □

Theorem 12.44 (Hilbert–Schmidt: compact self-adjoint operators)**.** *Let $A \in \mathcal{B}(\mathcal{H})$ be compact and self-adjoint on a Hilbert space $\mathcal{H} \neq \{0\}$. Then $\mathcal{H}$ has an orthonormal basis of eigenvectors of A: there are real $\lambda_1, \lambda_2, \ldots$ with $\lambda_n \to 0$ and an orthonormal system $\{e_n\}$ with*

$$Ax = \sum_n \lambda_n \langle e_n|x\rangle\, e_n \,, \qquad \forall\, x \in \mathcal{H} \,, \tag{12.27}$$

the series converging in norm. Every non-zero eigenvalue has finite-dimensional eigenspace, and $\|A\| = \max_n |\lambda_n|$.

Rests on Definition 12.41, Proposition 12.43, and Theorem 12.30.

The derivation is carried out in Appendix A.23. The route is: the supremum in the numerical-radius formula Equation (12.26) is attained on a compact operator, which produces a first eigenvector; its orthogonal complement is invariant, and the argument recurses; the eigenvalues tend to zero because an orthonormal sequence of eigenvectors with eigenvalues

bounded away from 0 has an image with no convergent subsequence. The attainment step is the one that needs compactness, and the appendix carries it out without the weak topology, which the standard treatment [1481] uses: self-adjointness alone forces a maximising sequence to satisfy $Ax_n - \lambda x_n \to 0$ in norm, after which the sequential form of compactness delivers a genuine eigenvector.

Full derivation: Appendix A.23 (p. 2885).

Derives Theorem 12.44.

Equation (12.27) is the exact infinite-dimensional analogue of the finite-dimensional spectral theorem Equation (5.105), and it is the only case in which the analogy is that close — the eigenvectors really do form a basis. It is also the historical starting point: this is the theorem Hilbert proved for the symmetric integral kernels of his spectral theory [388] and that Schmidt recast in the language used here [338]. The kernels concerned,

$$(Kf)(x) = \int_\Omega K(x,y)\, f(y)\, \mathrm{d}^N y\,, \qquad \int_\Omega \int_\Omega |K(x,y)|^2 \, \mathrm{d}^N x\, \mathrm{d}^N y < \infty\,, \tag{12.28}$$

define compact operators on $L^2(\Omega)$, and K is self-adjoint exactly when $K(y,x) = K(x,y)^*$. The Green functions of Chapter 10 are of this form, which is why a boundary-value problem has a discrete spectrum of modes while the free operators of Section 12.4.1 need not.

12.3.2 The Riesz representation theorem

Definition 12.45 (Continuous linear functional; the dual)**.** A *continuous linear functional* on $\mathcal{H}$ is a bounded linear map $f : \mathcal{H} \longrightarrow \mathbb{C}$. The set of them is the *topological dual* $\mathcal{H}^*$, normed by Equation (12.21).

Rests on Definitions 5.44 and 12.35.

Theorem 12.46 (Riesz representation)**.** *For every $f \in \mathcal{H}^*$ there is exactly one $z \in \mathcal{H}$ with*

$$f(x) = \langle z | x \rangle\,, \qquad \forall\, x \in \mathcal{H}\,, \tag{12.29}$$

and $\|f\| = \|z\|$.

Rests on Theorem 12.18 and Definition 12.45.

Proof. *Derives Theorem 12.46.*

If $f = 0$ take $z = 0$; uniqueness is proved below. Otherwise let $N = \ker f$, a subspace, closed because f is continuous and $N = f^{-1}(\{0\})$ is the preimage of a closed set (Definition 6.6), and proper because $f \neq 0$. By Theorem 12.18 applied to N, and by $N^\perp \neq \{0\}$ — which holds because otherwise $\mathcal{H} = N \oplus \{0\} = N$ — we may choose $u \in N^\perp$ with $\|u\| = 1$; note $f(u) \neq 0$, since $u \notin N$.

For arbitrary x, the vector $f(x)\,u - f(u)\,x$ is annihilated by f, by linearity, so it lies in N and is orthogonal to u:

$$0 = \langle u | f(x)u - f(u)x \rangle = f(x)\,\langle u|u\rangle - f(u)\,\langle u|x\rangle = f(x) - f(u)\,\langle u|x\rangle\,.$$

Hence $f(x) = f(u)\,\langle u|x\rangle = \langle f(u)^* u | x\rangle$ by Equation (5.36), and $z = f(u)^* u$ represents f.

Uniqueness. If $\langle z|x\rangle = \langle z'|x\rangle$ for all x, take $x = z - z'$ and use Equation (5.31). *Norm.* $|f(x)| \le \|z\|\,\|x\|$ by Equation (12.1), so $\|f\| \le \|z\|$; and $f(z) = \|z\|^2$ gives $\|f\| \ge \|z\|$. □

Corollary 12.47 ($\mathcal{H}$ is its own dual, antilinearly)**.** *The map $J : \mathcal{H} \longrightarrow \mathcal{H}^*$, $Jz = \langle z|\cdot\rangle$, is a bijection, is isometric, and is* antilinear*: $J(az_1 + bz_2) = a^* Jz_1 + b^* Jz_2$.*

Rests on Theorem 12.46 and Definition 12.45.

Proof. *Derives Corollary 12.47.*

Surjectivity and injectivity are the existence and uniqueness in Theorem 12.46, isometry is its last clause, and antilinearity is Equation (5.36). □

Remark 12.48 (What the bra–ket notation is). Corollary 12.47 is the whole content of Dirac's notation [1140]. The edition cited is the fourth, and deliberately so: the bra–ket split is not in the first edition of the *Principles* of 1930 [646], which is the source of the delta-function calculus used later in this chapter and of nothing else here. Dirac introduced the split in a note of 1939 [832], and brought it into the book at the third edition of 1947. Writing the vector z as $|z\rangle$ and the functional Jz as $\langle z|$, the value of the functional on a vector is $\langle z|x\rangle$ — the bracket split into a bra and a ket, which is where the words come from. The antilinearity of J is the rule that scalars are conjugated on passing the vertical bar, $\langle az| = a^* \langle z|$, and it is not a convention but a theorem: it is forced by Equation (5.36). In finite dimension the same identification was made in Proposition 5.59, where it needed only a basis; the point of Theorem 12.46 is that it survives without one.

The identification is special to Hilbert spaces and is exactly what fails elsewhere. The dual of a Banach space is in general a different space, and even for a pre-Hilbert space that is not complete the theorem is false — completeness enters through Theorem 12.18. It fails again, in the opposite direction, when one insists on writing $|x\rangle$ and $|p\rangle$ for position and momentum "eigenstates": those are functionals on a smaller space than $\mathcal{H}$ and correspond to no vector of $\mathcal{H}$ at all. That is the subject of Section 12.6.3, and it is the one place in this treatise where the bra–ket notation has to be read with care.

12.4 Self-adjointness and the spectral theorem

12.4.1 Spectrum and resolvent

Definition 12.49 (Resolvent set; spectrum)**.** Let $A \in \mathcal{B}(\mathcal{H})$. The *resolvent set* is

$$\rho(A) = \{\lambda \in \mathbb{C} \mid A - \lambda \mathbb{1} \text{ has an inverse in } \mathcal{B}(\mathcal{H})\} \,, \tag{12.30}$$

the map $R(\lambda) = (A - \lambda\mathbb{1})^{-1}$ on $\rho(A)$ is the *resolvent*, and the *spectrum* is $\sigma(A) = \mathbb{C} \setminus \rho(A)$.
Rests on Definitions 5.45 and 12.35.

Definition 12.50 (Point, continuous and residual spectrum)**.** For $\lambda \in \sigma(A)$ write $T = A - \lambda\mathbb{1}$. Then λ belongs to

1. the *point spectrum* $\sigma_p(A)$ if T is not injective; λ is then an *eigenvalue* and $\ker T$ the eigenspace;
2. the *continuous spectrum* $\sigma_c(A)$ if T is injective, $\operatorname{im} T$ is dense, and the inverse map T^{-1}, defined on $\operatorname{im} T$, is unbounded;
3. the *residual spectrum* $\sigma_r(A)$ if T is injective and $\operatorname{im} T$ is not dense.

Rests on Definition 12.49 and Corollary 12.19.

Proposition 12.51 (The three cases are exclusive and exhaustive)**.** *$\sigma(A)$ is the disjoint union of $\sigma_p(A)$, $\sigma_c(A)$ and $\sigma_r(A)$.*
Rests on Definitions 12.2 and 12.50.

Proof. *Derives Proposition 12.51.*

Exclusivity is clear from the statements. For exhaustiveness, suppose λ belongs to none of the three: then T is injective, $\operatorname{im} T$ is dense, and T^{-1} is bounded on $\operatorname{im} T$, say $\|T^{-1}y\| \le c\,\|y\|$. Then $\operatorname{im} T$ is closed: if $Tx_n \to y$, the sequence (Tx_n) is Cauchy, so $\|x_n - x_m\| \le c\,\|Tx_n - Tx_m\|$ shows (x_n) is Cauchy, and $x_n \to x$ for some x by Definition 12.2, with $Tx = y$ because T is continuous. A dense closed subspace is the whole space, so $\operatorname{im} T = \mathcal{H}$ and $T^{-1} \in \mathcal{B}(\mathcal{H})$, i.e. $\lambda \in \rho(A)$. □

In finite dimension injectivity and surjectivity coincide (Proposition 5.47) and only σ_p survives; Remark 5.48 already recorded that this is the identity that infinite dimension destroys.

Proposition 12.52 (Neumann series; the spectrum is bounded)**.** *If $|\lambda| > \|A\|$ then $\lambda \in \rho(A)$ and*

$$R(\lambda) = -\sum_{n=0}^{\infty} \frac{A^n}{\lambda^{n+1}}\,, \qquad \|R(\lambda)\| \le \frac{1}{|\lambda| - \|A\|}\,. \tag{12.31}$$

Hence $\sigma(A)$ is contained in the closed disc of radius $\|A\|$ about the origin, and $\|R(\lambda)\| \to 0$ as $|\lambda| \to \infty$.

Rests on Proposition 12.37 and Definition 12.49.

Proof. *Derives Proposition 12.52.*

By Equation (12.22), $\|A^n/\lambda^{n+1}\| \le \|A\|^n / |\lambda|^{n+1}$, a geometric series of ratio $\|A\| / |\lambda| < 1$ (Proposition 7.46), so the series converges absolutely and hence in $\mathcal{B}(\mathcal{H})$, by Propositions 12.8 and 12.37; its sum S obeys the stated norm bound. Multiplying the partial sums by $A - \lambda\mathbb{1}$ telescopes to $-\mathbb{1} + A^{N+1}/\lambda^{N+1}$, whose second term tends to 0; by continuity of multiplication in a Banach algebra, $(A - \lambda\mathbb{1})S = S(A - \lambda\mathbb{1}) = -\mathbb{1}$. Hence $A - \lambda\mathbb{1}$ is invertible with $R(\lambda) = (A - \lambda\mathbb{1})^{-1} = -S$, which is Equation (12.31), and $\|R(\lambda)\| = \|S\|$ obeys the bound already established. □

Proposition 12.53 (The resolvent set is open, the resolvent analytic)**.** *$\rho(A)$ is open, and for $x, y \in \mathcal{H}$ the scalar function $\lambda \longmapsto \langle x|R(\lambda)y\rangle$ is holomorphic on $\rho(A)$.*

Rests on Proposition 12.52 and Definition 8.5.

Proof. *Derives Proposition 12.53.*

Let $\lambda_0 \in \rho(A)$ and write $A - \lambda\mathbb{1} = (A - \lambda_0\mathbb{1})\big[\mathbb{1} - (\lambda - \lambda_0)R(\lambda_0)\big]$. For $|\lambda - \lambda_0| < 1/\|R(\lambda_0)\|$ the bracket is invertible by the same geometric-series argument as in Proposition 12.52, so $\lambda \in \rho(A)$ and

$$R(\lambda) = \sum_{n=0}^{\infty} (\lambda - \lambda_0)^n R(\lambda_0)^{n+1}\,. \tag{12.32}$$

Pairing Equation (12.32) with x and y and using Proposition 12.4 to take the bracket through the sum exhibits $\langle x|R(\lambda)y\rangle$ as a convergent power series about λ_0, hence holomorphic there (Theorem 7.50, and Chapter 8 for the complex statement). □

Theorem 12.54 (The spectrum is compact and non-empty)**.** *For $A \in \mathcal{B}(\mathcal{H})$ with $\mathcal{H} \neq \{0\}$, the set $\sigma(A)$ is compact and non-empty.*

Rests on Propositions 12.52 and 12.53 and Theorem 8.18.

Proof. *Derives Theorem 12.54.*

$\sigma(A)$ is closed because $\rho(A)$ is open (Proposition 12.53) and bounded by Proposition 12.52; a closed bounded subset of $\mathbb{C}$, identified with $\mathbb{R}^2$ as a metric space, is compact by the Heine–Borel theorem in $\mathbb{R}^N$ (Section 6.1.3, where the one-dimensional case Theorem 6.11 is proved and the N-dimensional one is recorded as owed).

Suppose $\sigma(A) = \varnothing$. Then for every x, y the function $g(\lambda) = \langle x|R(\lambda)y\rangle$ is entire, by Proposition 12.53, and by the norm bound in Equation (12.31) together with Equation (12.1) it satisfies $|g(\lambda)| \le \|x\|\,\|y\| / (|\lambda| - \|A\|)$ for $|\lambda| > \|A\|$, so g is bounded on $\mathbb{C}$ and tends to 0 at infinity. By Liouville's theorem (Theorem 8.18) g is constant, and the constant is 0. Since x and y were arbitrary, $R(\lambda) = 0$ for every λ — impossible, because $R(\lambda)(A - \lambda\mathbb{1}) = \mathbb{1} \neq 0$ on $\mathcal{H} \neq \{0\}$. □

Theorem 12.55 (The spectrum of a self-adjoint operator is real)**.** *Let $A \in \mathcal{B}(\mathcal{H})$ be self-adjoint. Then $\sigma(A) \subseteq \mathbb{R}$, the residual spectrum is empty, eigenvalues are real, and eigenvectors belonging to distinct eigenvalues are orthogonal.*

Rests on Definition 12.41, Theorem 12.54, and Corollary 12.19.

Proof. *Derives Theorem 12.55.*

Let $\lambda = a + \mathrm{i}b$ with $a, b \in \mathbb{R}$ and $b \neq 0$, and put $B = A - a\mathbb{1}$, which is self-adjoint. Since $\langle x|Bx\rangle$ is real (Proposition 12.42), expanding as in Equation (12.4) with $\lambda = -\mathrm{i}b$ makes the two cross terms $-\mathrm{i}b\,\langle Bx|x\rangle$ and $+\mathrm{i}b\,\langle x|Bx\rangle$ cancel, leaving

$$\|(A - \lambda\mathbb{1})x\|^2 = \|Bx\|^2 + b^2\,\|x\|^2 \geq b^2\,\|x\|^2 \,. \tag{12.33}$$

So $A - \lambda\mathbb{1}$ is injective with $\|(A - \lambda\mathbb{1})^{-1}\| \leq 1/|b|$ on its image, and that image is dense: by Equation (12.24) the adjoint of $A - \lambda\mathbb{1}$ is $A - \lambda^*\mathbb{1}$, so

$$\big(\mathrm{im}(A - \lambda\mathbb{1})\big)^{\perp} = \ker(A - \lambda^*\mathbb{1}) = \{0\} \,,$$

by Proposition 12.39 and by Equation (12.33) applied to λ^*, which has the same $|b|$; density then follows from Corollary 12.19. So $A - \lambda\mathbb{1}$ is injective with dense image and bounded inverse on it, which by the argument of Proposition 12.51 puts λ in $\rho(A)$. Hence $\sigma(A) \subseteq \mathbb{R}$.

The same computation for real λ shows $\big(\mathrm{im}(A - \lambda\mathbb{1})\big)^{\perp} = \ker(A - \lambda\mathbb{1})$, so if $A - \lambda\mathbb{1}$ is injective its image is dense: the residual spectrum is empty. If $Ax = \lambda x$ with $x \neq 0$ then $\lambda\,\|x\|^2 = \langle x|Ax\rangle$ is real; and if $Ax = \lambda x$, $Ay = \mu y$ with $\lambda \neq \mu$ real, then $\lambda\,\langle x|y\rangle = \langle Ax|y\rangle = \langle x|Ay\rangle = \mu\,\langle x|y\rangle$, forcing $\langle x|y\rangle = 0$. □

Example 12.56 (Multiplication by the coordinate: spectrum without eigenvectors)**.** On $\mathcal{H} = L^2([0,1])$ define

$$(Qf)(t) = t\,f(t) \,. \tag{12.34}$$

Then Q is bounded and self-adjoint with $\|Q\| = 1$, and

$$\sigma(Q) = \sigma_c(Q) = [0,1] \,, \qquad \sigma_p(Q) = \varnothing \,. \tag{12.35}$$

The operator has a full interval of spectrum and not one eigenvector.

Rests on Definition 12.50, Theorem 12.55, and Example 12.11.

Proof. *Derives Example 12.56.*

Bounded and self-adjoint. $|tf(t)| \leq |f(t)|$ on $[0,1]$ gives $\|Qf\| \leq \|f\|$; and since t is real, $\langle f|Qg\rangle = \int_0^1 f(t)^* t\,g(t)\,\mathrm{d}t = \langle Qf|g\rangle$, which is Equation (12.23).

No eigenvectors. If $Qf = \lambda f$ then $(t - \lambda)f(t) = 0$ for almost every t, so f vanishes almost everywhere on $[0,1] \setminus \{\lambda\}$, a set whose complement in $[0,1]$ has measure zero; hence $f = 0$ as an element of L^2 (Example 12.11). This is exactly where the identification of functions agreeing almost everywhere does its work.

$\sigma(Q) \subseteq [0,1]$. For $\lambda \notin [0,1]$ the function $t \longmapsto (t - \lambda)^{-1}$ is continuous on the compact set $[0,1]$, hence bounded there by Theorem 7.24, so multiplication by it is a bounded operator and is a two-sided inverse of $Q - \lambda\mathbb{1}$.

$[0,1] \subseteq \sigma(Q)$. Fix $\lambda \in [0,1]$ and let I_n be the intersection of $[0,1]$ with the interval of half-width $1/n$ about λ, of length $\ell_n \geq 1/n$. Put $f_n = \ell_n^{-1/2}$ on I_n and 0 elsewhere, so $\|f_n\| = 1$ while

$$\|(Q - \lambda\mathbb{1})f_n\|^2 = \frac{1}{\ell_n}\int_{I_n} |t - \lambda|^2\,\mathrm{d}t \leq \frac{1}{n^2} \,.$$

If $Q - \lambda\mathbb{1}$ had a bounded inverse R we would get $1 = \|f_n\| \leq \|R\|\,\|(Q - \lambda)f_n\| \leq \|R\|/n$ for every n, which is false. Hence $\lambda \in \sigma(Q)$, and $\|Q\| = 1$ follows from Proposition 12.52 and $1 \in \sigma(Q)$. By Theorem 12.55 the residual spectrum is empty and by the previous paragraph the point spectrum is, so all of $[0,1]$ is continuous spectrum. □

Remark 12.57 (What the example costs). Example 12.56 is the operator Remark 5.79 promised, and it settles the shape of the rest of the chapter. Every statement of Theorem 5.75 that mentions eigenvectors is simply false for Q: there is no orthonormal basis of eigenvectors,

no diagonal matrix Equation (5.105), no decomposition of $\mathcal{H}$ into eigenspaces. What survives is the *real spectrum*, and what replaces the eigenbasis is an integral over that spectrum against a projection-valued measure — for Q itself, the measure that assigns to a Borel set $\Omega \subseteq [0,1]$ the projection "multiply by the indicator of Ω". Physically, the difference is the difference between a bound state, which is a genuine eigenvector with a normalisable wavefunction, and a scattering state, which is not; the two coexist in the hydrogen spectrum of Chapter 81 and are the whole subject of Chapter 86. It is also why a position "eigenstate" $|x_0\rangle$ cannot be a vector of L^2 — Section 12.6.3 says what it is instead.

12.4.2 The spectral theorem

Definition 12.58 (Projection-valued measure)**.** A *projection-valued measure* on $\mathbb{R}$, acting on $\mathcal{H}$, is a map E assigning to every Borel set $\Omega \subseteq \mathbb{R}$ a projection $E(\Omega) \in \mathcal{B}(\mathcal{H})$ (Definition 12.41) such that

1. $E(\varnothing) = 0$ and $E(\mathbb{R}) = \mathbb{1}$;
2. $E(\Omega_1 \cap \Omega_2) = E(\Omega_1)E(\Omega_2)$;
3. for pairwise disjoint $\Omega_1, \Omega_2, \ldots$ and every $x \in \mathcal{H}$,

$$E\Big(\bigcup_n \Omega_n\Big)x = \sum_n E(\Omega_n)\,x\,, \tag{12.36}$$

the series converging in the norm of $\mathcal{H}$.

For $x, y \in \mathcal{H}$ the complex measure $\Omega \longmapsto \langle x|E(\Omega)y\rangle$ is written $\langle x|\mathrm{d}E(\lambda)y\rangle$; taking $y = x$ gives a positive measure of total mass $\|x\|^2$.

Rests on Definition 12.41 and Proposition 12.21.

The last remark is the reason the definition is shaped this way: for a unit vector x, $\Omega \longmapsto \langle x|E(\Omega)x\rangle = \|E(\Omega)x\|^2$ is a probability measure on $\mathbb{R}$. That is already the Born rule in embryo, before any physics has been assumed (Chapter 77).

Theorem 12.59 (Spectral theorem for a bounded self-adjoint operator)**.** *Let $A \in \mathcal{B}(\mathcal{H})$ be self-adjoint. Then there is exactly one projection-valued measure E on $\mathbb{R}$, supported on $\sigma(A)$, such that*

$$\langle x|Ay\rangle = \int_{\sigma(A)} \lambda\, \langle x|\mathrm{d}E(\lambda)y\rangle\,, \qquad \forall\, x, y \in \mathcal{H}\,, \tag{12.37}$$

written $A = \int_{\sigma(A)} \lambda\,\mathrm{d}E(\lambda)$. Every bounded operator commuting with A commutes with every $E(\Omega)$.

Equivalently, in the multiplication form: if $\mathcal{H}$ is separable there is a measure space (M,μ) with $\mu(M) < \infty$, a bounded real measurable function a on M, and a unitary $U : \mathcal{H} \longrightarrow L^2(M,\mu)$ with

$$\big(UAU^{-1}f\big)(m) = a(m)\,f(m)\,, \qquad \forall\, f \in L^2(M,\mu)\,. \tag{12.38}$$

Every bounded self-adjoint operator is a multiplication operator in disguise.

Rests on Definition 12.58, Theorem 12.55, and Proposition 12.43.

Both halves are derived in Appendix A.24. The route is the standard one: the continuous functional calculus is built first, from the fact that for a self-adjoint operator the norm of a polynomial in it equals the supremum of that polynomial on the spectrum; a complex measure is then obtained for each pair of vectors by the Riesz–Markov representation theorem; the calculus is extended to bounded Borel functions, and the projections $E(\Omega)$ are the values it takes on indicator functions. The multiplication form follows by decomposing $\mathcal{H}$ into cyclic subspaces, on each of which the calculus itself provides the unitary. Two

classical theorems are used there and are not proved in this treatise — Riesz–Markov, which is measure theory, and Stone–Weierstrass, of which only the Weierstrass approximation theorem already quoted in Remark 12.1 is needed; the appendix states each precisely where it is used and collects both in a remark, and nothing else is assumed. The original is von Neumann's memoir [664]; the modern account followed there is [1481], chapters VII and VIII.
Full derivation: Appendix A.24 (p. 2889).

Derives Theorem 12.59.

The uniqueness of the functional calculus is worth isolating from that argument, and it costs only a paragraph, so it is proved here: it is what makes "the" functions of an observable well defined — without it the expression $f(A)$ would depend on how it was constructed.

Definition 12.60 (Functional calculus)**.** Given E as in Theorem 12.59 and a bounded Borel function $f : \sigma(A) \longrightarrow \mathbb{C}$, the operator $f(A)$ is defined by

$$\langle x | f(A)\, y \rangle = \int_{\sigma(A)} f(\lambda)\ \langle x | \mathrm{d}E(\lambda) y \rangle\ . \tag{12.39}$$

Rests on Theorem 12.59 and Definition 12.58.

Proposition 12.61 (Uniqueness of the continuous functional calculus)**.** *Let $A \in \mathcal{B}(\mathcal{H})$ be self-adjoint and let Φ_1, Φ_2 be maps from the continuous functions on $\sigma(A)$ to $\mathcal{B}(\mathcal{H})$ such that each is linear and multiplicative, sends the constant function 1 to $\mathbb{1}$ and the function $\lambda \longmapsto \lambda$ to A, and is continuous for the supremum norm on the domain and the operator norm on the target. Then $\Phi_1 = \Phi_2$.*

Rests on Definition 12.60 and Proposition 12.37.

Proof. *Derives Proposition 12.61.*

Linearity and multiplicativity, together with the two normalisations, force $\Phi_i(p) = p(A)$ for every polynomial p, where $p(A)$ is the algebraic expression: the monomial λ^n goes to A^n by multiplicativity and induction, and linear combinations follow. So Φ_1 and Φ_2 agree on polynomials. By the Weierstrass approximation theorem (Remark 12.1) the polynomials are dense, in the supremum norm, in the continuous functions on the compact set $\sigma(A)$ (Theorem 12.54). Two continuous maps agreeing on a dense subset of a metric space agree everywhere. □

Remark 12.62 (The finite-dimensional case is the whole picture)*.* In finite dimension Theorem 12.59 is Theorem 5.75, already proved, and the translation is worth carrying out once because it fixes the meaning of every symbol above. Let $\dim \mathcal{H} = n < \infty$ and let A be self-adjoint, with distinct eigenvalues $\lambda_1, \dots, \lambda_r$ and P_i the orthogonal projection onto the eigenspace of λ_i. Then Theorem 5.75 says

$$A = \sum_{i=1}^{r} \lambda_i P_i\,, \qquad \sum_{i=1}^{r} P_i = \mathbb{1}\,, \qquad P_i P_j = \delta_{ij} P_i\,, \tag{12.40}$$

and the projection-valued measure is $E(\Omega) = \sum_{\lambda_i \in \Omega} P_i$, which manifestly satisfies the three conditions of Definition 12.58; the integral Equation (12.37) is then a finite sum and reproduces Equation (12.40). The functional calculus Equation (12.39) becomes $f(A) = \sum_i f(\lambda_i) P_i$, which is the familiar rule that a function of a diagonalizable operator acts by that function on each eigenvalue. What the general theorem does is exactly to replace the finite sum over eigenvalues by an integral over the spectrum — nothing more, and nothing less. Example 12.56 shows why nothing less will do.

Remark 12.63 (Reading the spectral theorem physically)*.* Three of the quantum postulates of Chapter 77 are statements about Theorem 12.59. That an observable is represented by a self-adjoint operator is the demand that the theorem apply at all; that the possible values

of a measurement are the points of $\sigma(A)$ is the support statement; and that the probability of finding a value in Ω is $\|E(\Omega)\psi\|^2$ is the positive measure of Definition 12.58, whose total mass is 1 for a normalised state. The expectation value is $\langle\psi|A\psi\rangle$, which is Equation (12.37) read as the mean of that probability measure. Functions of observables — $\mathcal{H}^2$, $\mathrm{e}^{-\mathrm{i}\mathcal{H}t/\hbar}$, the projectors of a filtering measurement — are Equation (12.39), and Proposition 12.61 is what allows the treatise to speak of *the* function of an operator. This is the content von Neumann extracted from the formalism [664, 706], and it is the reason the postulates can be stated without approximation or interpretation of the mathematics.

12.4.3 Stone's theorem

Definition 12.64 (Strongly continuous one-parameter unitary group)**.** A family $\{U(t)\}_{t\in\mathbb{R}}$ of unitary operators on $\mathcal{H}$ is a *strongly continuous one-parameter unitary group* if

$$U(t+s) = U(t)\,U(s)\,, \qquad U(0) = \mathbb{1}\,, \tag{12.41}$$

and for every $x \in \mathcal{H}$ the map $t \longmapsto U(t)x$ is continuous from $\mathbb{R}$ to $\mathcal{H}$. The parameter t carries the SI unit s throughout this section.

Rests on Definitions 6.6 and 12.41.

Continuity is demanded of $t \longmapsto U(t)x$ for each fixed x (*strong* continuity), not of $t \longmapsto U(t)$ in the operator norm, which would be a much stronger and, for the groups physics needs, a false requirement — norm continuity forces the generator to be bounded, and Theorem 12.75 will show that the generators of interest are not.

Proposition 12.65 (Exponential of a bounded self-adjoint operator)**.** *Let $H \in \mathcal{B}(\mathcal{H})$ be self-adjoint, with the SI dimension of energy,* J*. Then the series*

$$U(t) = \exp\left(-\frac{\mathrm{i}Ht}{\hbar}\right) = \sum_{n=0}^{\infty} \frac{1}{n!}\left(-\frac{\mathrm{i}t}{\hbar}\right)^n H^n \tag{12.42}$$

converges in $\mathcal{B}(\mathcal{H})$ for every t and defines a strongly continuous one-parameter unitary group with

$$\mathrm{i}\hbar\,\frac{\mathrm{d}}{\mathrm{d}t}U(t)\,x = H\,U(t)\,x\,, \qquad \forall\, x \in \mathcal{H}\,, \tag{12.43}$$

the derivative existing in the norm of $\mathcal{H}$.

Rests on Propositions 12.37 and 12.39 and Definition 12.64.

Proof. *Derives Proposition 12.65.*

Since $\hbar$ has the SI unit J s, the combination $Ht/\hbar$ is dimensionless and Equation (12.42) is a sum of operators, as it must be. By Equation (12.22) the nth term has norm at most $(|t|\,\|H\|/\hbar)^n/n!$, so the series converges absolutely and hence in $\mathcal{B}(\mathcal{H})$, by Propositions 12.8 and 12.37.

The group law Equation (12.41) follows from the Cauchy product of the two absolutely convergent series (Proposition 7.49) and the binomial theorem, which is legitimate because the two exponents commute, both being polynomials in the single operator H. Taking adjoints term by term — legitimate by Equation (12.24) and continuity of the adjoint, which is isometric — and using $H^\dagger = H$ gives $U(t)^\dagger = \exp(+\mathrm{i}Ht/\hbar) = U(-t)$, which by the group law is $U(t)^{-1}$; so each $U(t)$ is unitary.

For Equation (12.43), the difference quotient

$$\frac{U(t+s)-U(t)}{s} + \frac{\mathrm{i}}{\hbar}H\,U(t) = \left[\frac{U(s)-\mathbb{1}}{s} + \frac{\mathrm{i}}{\hbar}H\right]U(t)$$

has norm at most $\|(U(s)-\mathbb{1})/s + \mathrm{i}H/\hbar\|$, and the bracket is the tail of Equation (12.42) divided by s, of norm at most $|s|\,\|H\|^2\,\hbar^{-2}\exp(|s|\,\|H\|/\hbar)$, which tends to 0. Multiplying by $\mathrm{i}\hbar$ gives Equation (12.43), in the operator norm and hence also strongly. □

Theorem 12.66 (Stone)**.** *The map $H \longmapsto U(t) = \exp(-\mathrm{i}Ht/\hbar)$ is a bijection between the self-adjoint operators on $\mathcal{H}$ — bounded or not, in the sense of Section 12.5.2 — and the strongly continuous one-parameter unitary groups on $\mathcal{H}$. Given the group, the operator is recovered as*

$$H\,x = \mathrm{i}\hbar \lim_{t\to 0} \frac{U(t)\,x - x}{t}\,, \tag{12.44}$$

on the domain $D(H)$ of vectors for which the limit exists, and $D(H)$ is dense with H self-adjoint on it.

Rests on Definition 12.64, Proposition 12.65, and Theorem 12.59.

Half of Theorem 12.66 was proved in Proposition 12.65 for bounded H, and one further piece — everything that does not require the unbounded functional calculus — is proved next; the remainder is derived in Appendix A.25.

Proposition 12.67 (The generator is symmetric, and generates the motion)**.** *Let U be a strongly continuous one-parameter unitary group and let $D(H)$ and H be as in Equation (12.44). Then $D(H)$ is a subspace, $U(t)D(H) = D(H)$ for every t, and*

$$\langle x|Hy\rangle = \langle Hx|y\rangle\,, \qquad \forall\, x, y \in D(H)\,. \tag{12.45}$$

Moreover, for $x \in D(H)$ the curve $t \longmapsto U(t)x$ is differentiable and

$$\mathrm{i}\hbar\,\frac{\mathrm{d}}{\mathrm{d}t}\big(U(t)x\big) = H\,U(t)\,x\,. \tag{12.46}$$

Rests on Definition 12.64 and Theorem 12.66.

Proof. *Derives Proposition 12.67.*

$D(H)$ is a subspace because the limit in Equation (12.44) is linear in x where it exists. If $x \in D(H)$ then, by Equation (12.41), $\big(U(t)U(s)x - U(t)x\big)/s = U(t)\big(U(s)x - x\big)/s$, and $U(t)$ is isometric, so the limit as $s \to 0$ exists and equals $-\mathrm{i}\hbar^{-1}U(t)Hx$; thus $U(t)x \in D(H)$ with $HU(t)x = U(t)Hx$, which is Equation (12.46), and applying the same with $-t$ gives $U(t)D(H) = D(H)$.

For Equation (12.45), let $x, y \in D(H)$ and let $t \neq 0$. Since $U(t)^\dagger = U(-t)$, and writing $s = -t$,

$$\left\langle x \middle| \frac{U(t)y - y}{t} \right\rangle = \left\langle \frac{U(-t)x - x}{t} \middle| y \right\rangle = -\left\langle \frac{U(s)x - x}{s} \middle| y \right\rangle . \tag{12.47}$$

Now $s \to 0$ exactly when $t \to 0$, and by Equation (12.44) both difference quotients converge: $\big(U(t)y - y\big)/t \to -\mathrm{i}Hy/\hbar$ and $\big(U(s)x - x\big)/s \to -\mathrm{i}Hx/\hbar$. Taking the limit in Equation (12.47), with the bracket carried through the limit by Proposition 12.4 and the antilinearity Equation (5.36) applied on the right,

$$-\frac{\mathrm{i}}{\hbar}\,\langle x|Hy\rangle = -\left(-\frac{\mathrm{i}}{\hbar}\right)^{*}\langle Hx|y\rangle = -\frac{\mathrm{i}}{\hbar}\,\langle Hx|y\rangle\,,$$

which is Equation (12.45). □

The two hard halves — that the domain defined by the limit in Equation (12.44) is dense, and that the operator on it is self-adjoint rather than merely symmetric; and the converse, that every self-adjoint operator, bounded or not, exponentiates to a group — are derived in Appendix A.25. The first uses a vector-valued integral to produce enough vectors in the domain by smoothing, and then an elementary differential-equation argument to kill both deficiency subspaces; the second needs the spectral theorem for *unbounded* self-adjoint operators, which the appendix obtains from Theorem 12.59 through the Cayley transform rather than quoting. Stone's paper [707] and [1481], chapter VIII, carry both.

Full derivation: Appendix A.25 (p. 2897).

Derives Theorem 12.66.

Remark 12.68 (Time evolution, and why $\hbar$ is written). Equation (12.46) is the Schrödinger equation before any Hamiltonian has been chosen: written for a state vector it is

$$\mathrm{i}\hbar\,\frac{\mathrm{d}}{\mathrm{d}t}\,|\psi(t)\rangle = \mathcal{H}\,|\psi(t)\rangle\ , \qquad |\psi(t)\rangle = \exp\left(-\frac{\mathrm{i}\mathcal{H}t}{\hbar}\right)|\psi(0)\rangle\ , \tag{12.48}$$

and Chapter 77 adopts it as a postulate. Stone's theorem [707] is what makes the two halves of Equation (12.48) equivalent, and it explains the otherwise unmotivated demand that the Hamiltonian be self-adjoint and not merely symmetric: symmetry alone gives neither a unitary group nor a conserved norm, so probability would not be conserved. That is not a technicality — Section 12.5.2 exhibits a symmetric momentum operator with no self-adjoint extension at all.

The factor $\hbar$ is carried explicitly here and everywhere in this treatise (CLAUDE.md rule 3). Its role is dimensional and is worth stating once: $\mathcal{H}$ has the unit J and t the unit s, so $\mathcal{H}t$ has the unit J s, which is the unit of $\hbar$; the argument of the exponential is therefore a pure number, as the series Equation (12.42) requires. The generator of the group is thus not the Hamiltonian but $\mathcal{H}/\hbar$, an angular frequency in 1/s — which is the mathematical form of the Planck–Einstein relation, and the reason a stationary state of energy E oscillates at angular frequency $E/\hbar$. The convention $\hbar = 1$, common in the literature, sets a unit of action to 1 and is not used here.

12.5 Unbounded operators and domains

12.5.1 Closed operators and domains

Definition 12.69 (Operator with a domain)**.** An *operator* on $\mathcal{H}$ is a pair $(A, D(A))$ with $D(A) \subseteq \mathcal{H}$ a subspace, the *domain*, and $A : D(A) \longrightarrow \mathcal{H}$ linear. It is *densely defined* if $D(A)$ is dense. Two operators are equal only if their domains agree; $A \subseteq B$ means $D(A) \subseteq D(B)$ and $Bx = Ax$ on $D(A)$, and B is then an *extension* of A.

Rests on Definition 5.35 and Corollary 12.19.

The insistence that the domain be part of the operator is not bookkeeping. Section 12.5.2 exhibits one differential expression, on one interval, that is several different operators according to the domain chosen, with different spectra and different physics; and Theorem 12.75 shows that the operators of quantum mechanics have no choice in the matter.

Definition 12.70 (Graph; closed and closable operators)**.** The *graph* of $(A, D(A))$ is

$$\Gamma(A) = \{(x, Ax) \mid x \in D(A)\} \subseteq \mathcal{H} \oplus \mathcal{H}\,, \tag{12.49}$$

where $\mathcal{H} \oplus \mathcal{H}$ carries the inner product $\langle (x_1, y_1)|(x_2, y_2)\rangle = \langle x_1|x_2\rangle + \langle y_1|y_2\rangle$. The operator is *closed* if $\Gamma(A)$ is a closed subspace, and *closable* if the closure of $\Gamma(A)$ is itself the graph of an operator, called the *closure* $\overline{A}$.

Rests on Definitions 6.3 and 12.69.

Unwinding Equation (12.49): A is closed if and only if, whenever $x_n \in D(A)$ with $x_n \to x$ *and* $Ax_n \to y$, it follows that $x \in D(A)$ and $Ax = y$. This is strictly weaker than continuity, which would require the second hypothesis to follow from the first. The closure of the graph fails to be a graph exactly when it contains some $(0, y)$ with $y \neq 0$, which is the obstruction to closability.

Definition 12.71 (Adjoint of a densely defined operator)**.** Let $(A, D(A))$ be densely defined. Put

$$D(A^\dagger) = \{y \in \mathcal{H} \mid \exists\, z \in \mathcal{H} : \langle z|x\rangle = \langle y|Ax\rangle \quad \forall\, x \in D(A)\}\ , \tag{12.50}$$

and $A^\dagger y = z$ on it. The vector z is unique because $D(A)$ is dense: if $\langle z - z'|x\rangle = 0$ for every x in a dense set then $z = z'$, by Corollary 12.19.

Rests on Definition 12.69, Theorem 12.38, and Corollary 12.19.

Definition 12.72 (Symmetric; self-adjoint)**.** A densely defined $(A, D(A))$ is *symmetric* if $\langle x|Ay\rangle = \langle Ax|y\rangle$ for all $x, y \in D(A)$, equivalently $A \subseteq A^\dagger$; it is *self-adjoint* if $A = A^\dagger$, that is, if in addition $D(A^\dagger) = D(A)$.

Rests on Definitions 12.69 and 12.71.

Proposition 12.73 (The adjoint is always closed)**.** *For densely defined A, the operator $A^\dagger$ is closed. A self-adjoint operator is therefore closed, and a symmetric operator is closable, with $\overline{A} \subseteq A^\dagger$.*

Rests on Definitions 12.70 and 12.71.

Proof. *Derives Proposition 12.73.*

Let $y_n \in D(A^\dagger)$ with $y_n \to y$ and $A^\dagger y_n \to z$. For every $x \in D(A)$, $\left\langle A^\dagger y_n \middle| x \right\rangle = \langle y_n|Ax\rangle$; letting $n \to \infty$ and using Proposition 12.4 on both sides gives $\langle z|x\rangle = \langle y|Ax\rangle$. By Equation (12.50) this says $y \in D(A^\dagger)$ with $A^\dagger y = z$, which is closedness. If A is symmetric then $A \subseteq A^\dagger$ with $A^\dagger$ closed, so the closure of $\Gamma(A)$ lies inside $\Gamma(A^\dagger)$ and is therefore a graph. □

Theorem 12.74 (Hellinger–Toeplitz)**.** *Let A be defined on all of $\mathcal{H}$ and symmetric, that is $\langle x|Ay\rangle = \langle Ax|y\rangle$ for all $x, y \in \mathcal{H}$. Then A is bounded.*

Rests on Definition 12.72, Proposition 12.36, and Remark 12.1.

Proof. *Derives Theorem 12.74.*

First, A is closed. Let $x_n \to x$ and $Ax_n \to y$. For every $z \in \mathcal{H}$,

$$\langle z|y\rangle = \lim_n \langle z|Ax_n\rangle = \lim_n \langle Az|x_n\rangle = \langle Az|x\rangle = \langle z|Ax\rangle \ ,$$

using symmetry twice and Proposition 12.4 for the limits. Hence $\langle z|y - Ax\rangle = 0$ for all z, so $y = Ax$ by Equation (5.31): the graph is closed. Since A is everywhere defined and closed between Banach spaces, the closed-graph theorem (Remark 12.1) makes it bounded. □

Theorem 12.75 (The canonical commutation relation admits no bounded solution)**.** *Let $\mathcal{H} \neq \{0\}$ and let $Q, P \in \mathcal{B}(\mathcal{H})$. Then*

$$[Q, P] = QP - PQ = \mathrm{i}\hbar\, \mathbb{1} \tag{12.51}$$

is impossible.

Rests on Proposition 12.37 and Definition 12.35.

Proof. *Derives Theorem 12.75.*

Assume Equation (12.51). By induction,

$$QP^n - P^n Q = \mathrm{i}\hbar\, n\, P^{n-1} \,, \qquad n \geq 1 \,, \tag{12.52}$$

the step being $QP^n - P^n Q = (QP - PQ)\, P^{n-1} + P\,(QP^{n-1} - P^{n-1}Q)$, which by the inductive hypothesis is $\mathrm{i}\hbar P^{n-1} + \mathrm{i}\hbar(n-1)P^{n-1}$.

No power of P vanishes. For suppose $m \geq 1$ is least with $P^m = 0$. Then Equation (12.52) at $n = m$ reads $0 = \mathrm{i}\hbar\, m\, P^{m-1}$, so $P^{m-1} = 0$; minimality forces $m - 1 = 0$, i.e. $P^0 = \mathbb{1} = 0$, which is false on $\mathcal{H} \neq \{0\}$.

So $\|P^{n-1}\| > 0$ for every n, and taking norms in Equation (12.52) with Equation (12.22),

$$\hbar\, n \left\|P^{n-1}\right\| = \|QP^n - P^n Q\| \leq 2\,\|Q\|\,\|P^n\| \leq 2\,\|Q\|\,\|P\|\,\left\|P^{n-1}\right\| \,,$$

whence $\hbar\, n \leq 2\,\|Q\|\,\|P\|$ for every $n \in \mathbb{N}$ — absurd. □

Corollary 12.76 (Position and momentum are unbounded, and cannot be everywhere defined)**.** *Any pair of operators satisfying Equation (12.51) on a dense domain must be unbounded, and if they are symmetric they cannot be defined on all of $\mathcal{H}$.*

Rests on Theorems 12.74 and 12.75.

Proof. *Derives Corollary 12.76.*

Unboundedness is Theorem 12.75. If such an operator were symmetric and defined on all of $\mathcal{H}$ it would be bounded by Theorem 12.74, a contradiction. □

Remark 12.77 (Why quantum mechanics needs domains). Corollary 12.76 is the reason this section exists. The canonical relation Equation (12.51) is not optional — it is the content of canonical quantisation (Chapter 99) and the source of the uncertainty relation — and it forbids position and momentum from being tame operators. They are unbounded, hence discontinuous, hence not defined on every state; every statement about them carries a domain, and Chapter 77 must say on which vectors an observable acts. The point was von Neumann's, and it is the technical core of the axiomatisation [664, 706].

A second consequence is that a domain is a piece of *physics*, not a technicality to be suppressed. For a differential operator the domain encodes the boundary conditions, so choosing it is choosing the physical situation — a particle in a box, on a ring, or on a half-line — and the eigenvalue problem that results is the Sturm–Liouville problem of Chapter 9, whose "self-adjoint boundary conditions" are exactly the choices that make the operator self-adjoint in the sense of Definition 12.72. The next subsection works one such family out completely.

12.5.2 Symmetric versus self-adjoint: deficiency indices

Definition 12.78 (Essential self-adjointness)**.** A symmetric operator A is *essentially self-adjoint* if its closure $\overline{A}$ (Proposition 12.73) is self-adjoint; equivalently, if A has exactly one self-adjoint extension. A *core* for a self-adjoint B is a subspace $D \subseteq D(B)$ on which B is essentially self-adjoint.

Rests on Definition 12.72 and Proposition 12.73.

The three notions — symmetric, essentially self-adjoint, self-adjoint — are genuinely distinct, and the distinction is measured by two integers.

Definition 12.79 (Deficiency subspaces and indices)**.** Let A be a closed symmetric operator and let $\mu > 0$ be a fixed number carrying the same SI dimension as A, so that $A \mp \mathrm{i}\mu$ is dimensionally consistent. The *deficiency subspaces* are

$$K_{\pm} = \ker\left(A^{\dagger} \mp \mathrm{i}\mu\right) = \left(\mathrm{im}(A \pm \mathrm{i}\mu)\right)^{\perp}, \tag{12.53}$$

and the *deficiency indices* are $n_{\pm} = \dim K_{\pm}$. They do not depend on the choice of μ.

Rests on Definition 12.71, Proposition 12.39, and Corollary 12.19.

The second equality in Equation (12.53) is Proposition 12.39 carried over to the unbounded case, where it reads $\ker A^{\dagger} = (\mathrm{im}\, A)^{\perp}$ with the same one-line proof; the independence of μ is part of von Neumann's analysis. Writing μ rather than 1 is not pedantry: for a momentum operator μ is a momentum in $\mathrm{kg\,m/s}$, and $A^{\dagger} \mp \mathrm{i}$ would be the difference of a momentum and a pure number.

Theorem 12.80 (von Neumann's criterion)**.** *Let A be a closed symmetric operator with deficiency indices (n_+, n_-).*

1. *A is self-adjoint if and only if $n_+ = n_- = 0$;*
2. *A has self-adjoint extensions if and only if $n_+ = n_-$, and they are then in bijective correspondence with the unitary maps $K_+ \longrightarrow K_-$ — a family parametrised by the group $\mathrm{U}(n)$ when $n_+ = n_- = n < \infty$;*
3. *if $n_+ \neq n_-$, A has no self-adjoint extension at all.*

Rests on Definitions 12.72, 12.78 and 12.79.

The criterion is proved in Appendix A.26. The proof passes to the Cayley transform, which converts a closed symmetric operator into an isometry between the orthogonal complements of the deficiency subspaces and a self-adjoint operator into a unitary; the self-adjoint extensions of the operator then correspond to the unitary extensions of that isometry, which exist exactly when the two deficiency subspaces have the same dimension. The appendix also derives the explicit description of each extension, and the independence of the indices from μ asserted in Definition 12.79. Von Neumann's memoir of 1930 [664] is the original.

Full derivation: Appendix A.26 (p. 2904).

Derives Theorem 12.80.

The deficiency-index theory and von Neumann's extension theorem are in the second Reed–Simon volume [1559], chapter X; the first volume [1481], which is what the rest of this chapter cites, does not carry them. The two examples that follow are the ones the criterion was made for, and they differ only in the interval.

Example 12.81 (Momentum on a finite interval: a circle of self-adjoint momenta)**.** Let $L > 0$ be a length and $\mathcal{H} = L^2([0, L])$, and let

$$P = -\mathrm{i}\hbar \, \frac{\mathrm{d}}{\mathrm{d}x} \, , \tag{12.54}$$

with the initial domain D_0 of continuously differentiable functions vanishing at both endpoints, which is dense. Then P is symmetric on D_0 but not self-adjoint; its deficiency indices are $(n_+, n_-) = (1, 1)$; and it has a one-parameter family of self-adjoint extensions P_θ, $\theta \in [0, 2\pi)$, with domains

$$D_\theta = \left\{ \psi \;\middle|\; \psi \text{ absolutely continuous},\ \psi' \in \mathcal{H},\ \psi(L) = \mathrm{e}^{\mathrm{i}\theta}\psi(0) \right\} , \tag{12.55}$$

whose spectra are the discrete sets

$$\sigma(P_\theta) = \left\{ p_n = \frac{\hbar\,(\theta + 2\pi n)}{L} \;\middle|\; n \in \mathbb{Z} \right\} , \tag{12.56}$$

with normalised eigenfunctions $\psi_n(x) = L^{-1/2}\mathrm{e}^{\mathrm{i}p_n x/\hbar}$.

Rests on Definition 12.79, Theorem 12.80, and Example 12.11.

Proof. *Derives Example 12.81.*

Symmetry. For $\varphi, \psi \in D_0$, integration by parts (Theorem 7.43) gives

$$\langle \varphi | P\psi \rangle - \langle P\varphi | \psi \rangle = -\mathrm{i}\hbar \Big[\varphi(x)^* \psi(x) \Big]_0^L = -\mathrm{i}\hbar \left(\varphi(L)^*\psi(L) - \varphi(0)^*\psi(0) \right) , \tag{12.57}$$

which vanishes because both functions vanish at both endpoints.

The adjoint. A computation with Equation (12.50), using the du Bois-Reymond lemma to identify the weak derivative, shows that $P^\dagger$ is the same differential expression on the *maximal* domain — absolutely continuous ψ with $\psi' \in \mathcal{H}$ and *no* boundary condition [1481]. That is already the failure of self-adjointness: $D(P^\dagger)$ is far larger than D_0.

Deficiency indices. Solve $P^\dagger\psi = \pm\mathrm{i}\mu\psi$ with $\mu > 0$ a momentum. The equation $-\mathrm{i}\hbar\psi' = \mathrm{i}\mu\psi$ gives $\psi(x) = C\mathrm{e}^{-\mu x/\hbar}$, and $-\mathrm{i}\hbar\psi' = -\mathrm{i}\mu\psi$ gives $\psi(x) = C\mathrm{e}^{+\mu x/\hbar}$. Both are continuous on the compact interval $[0, L]$, hence bounded, hence in $L^2([0, L])$; each solution space is one-dimensional, so $n_+ = n_- = 1$ and Theorem 12.80 predicts a family of extensions parametrised by $\mathrm{U}(1)$ — a circle.

The extensions. On D_θ the boundary term Equation (12.57) vanishes, since $\varphi(L)^*\psi(L) = \mathrm{e}^{-\mathrm{i}\theta}\mathrm{e}^{\mathrm{i}\theta}\varphi(0)^*\psi(0) = \varphi(0)^*\psi(0)$, so each P_θ is symmetric. That each is exactly self-adjoint,

rather than merely symmetric, and that the D_θ exhaust the possibilities — a boundary condition making Equation (12.57) vanish for all φ, ψ in the domain must relate $\psi(L)$ to $\psi(0)$ by a phase — are proved in Corollary A.274, which also computes which unitary map of Theorem 12.80 carries which value of θ.

Spectrum. $P_\theta\psi = p\psi$ has the solutions $\psi(x) = C\mathrm{e}^{\mathrm{i}px/\hbar}$, which lie in D_θ exactly when $\mathrm{e}^{\mathrm{i}pL/\hbar} = \mathrm{e}^{\mathrm{i}\theta}$, that is, when $pL/\hbar = \theta + 2\pi n$ for an integer n — which is Equation (12.56). These eigenfunctions are orthonormal after the stated normalisation and, being the trigonometric system of Chapter 17 up to the phase θ, form an orthonormal basis; so the spectrum is pure point. □

Example 12.82 (Momentum on the half-line: no self-adjoint extension)**.** Let $\mathcal{H} = L^2([0,\infty))$ and let P be Equation (12.54) on the dense domain of continuously differentiable functions of compact support in $(0,\infty)$. Then P is symmetric, its deficiency indices are $(n_+, n_-) = (1, 0)$, and by Theorem 12.80 it has *no* self-adjoint extension. There is no momentum observable for a particle confined to a half-line.

Rests on Example 12.81, Theorem 12.80, and Definition 12.79.

Proof. *Derives Example 12.82.*

Symmetry is Equation (12.57) again, with both boundary terms zero because the functions have compact support in the open half-line. The adjoint is once more the maximal operator, with no boundary condition [1481]. Solving $P^\dagger\psi = \mathrm{i}\mu\psi$ gives $\psi(x) = C\mathrm{e}^{-\mu x/\hbar}$, which is square integrable on $[0,\infty)$ with $\|\psi\|^2 = |C|^2\,\hbar/(2\mu)$, so $n_+ = 1$. Solving $P^\dagger\psi = -\mathrm{i}\mu\psi$ gives $\psi(x) = C\mathrm{e}^{+\mu x/\hbar}$, whose square integral diverges unless $C = 0$, so $n_- = 0$. The indices differ, and case (3) of Theorem 12.80 applies. □

Remark 12.83 (Three intervals, three fates)*.* The same expression Equation (12.54) behaves in three different ways according to the interval, and nothing but the domain distinguishes the cases.

- On the whole line, P on the smooth rapidly decreasing functions is essentially self-adjoint, with purely continuous spectrum $\mathbb{R}$ and no eigenvectors — the Fourier transform of Chapter 17 turns it into multiplication by $\hbar k$, which is Example 12.56 on an unbounded interval.

- On a finite interval it is not self-adjoint but has a circle of self-adjoint extensions, each with a discrete spectrum Equation (12.56). The parameter θ is physical: $\theta = 0$ is the particle on a ring, and a non-zero θ is the ring threaded by a magnetic flux, the phase being the Aharonov–Bohm phase. The mathematics does not choose; the physical situation does.

- On the half-line it has no self-adjoint extension. The reason is visible in the proof: the translations generated by P push probability off the end of the half-line and back on again, which no unitary group can do. By Theorem 12.66, no self-adjoint generator, no unitary group; and no unitary group, no conservation of probability.

The corresponding question for the *Hamiltonian* — whether $-\hbar^2\nabla^2/(2m) + V$ is essentially self-adjoint on the smooth compactly supported functions — is the question of whether the Schrödinger equation has a unique unitary dynamics, and it is answered, potential by potential, in Chapters 81 and 82 [1335, 1559].

12.6 Direct sums, tensor products and rigged spaces

12.6.1 Direct sums and orthogonal decompositions

Definition 12.84 (External direct sum)**.** Let $\mathcal{H}_1, \mathcal{H}_2, \dots$ be Hilbert spaces. Their *direct sum* is

$$\bigoplus_n \mathcal{H}_n = \left\{ (x_1, x_2, \dots) \;\middle|\; x_n \in \mathcal{H}_n,\ \sum_n \|x_n\|^2 < \infty \right\}, \tag{12.58}$$

with componentwise operations and $\langle x|y\rangle = \sum_n \langle x_n|y_n\rangle$.

Rests on Definition 5.97 and Example 12.9.

Proposition 12.85 (The direct sum is a Hilbert space)**.** *Equation (12.58) is a Hilbert space, and the maps $J_n : \mathcal{H}_n \longrightarrow \bigoplus_m \mathcal{H}_m$ placing a vector in the nth slot are isometric with mutually orthogonal images.*

Rests on Definition 12.84 and Proposition 12.10.

Proof. *Derives Proposition 12.85.*

The inner product is well defined and the space is a vector space by the same two estimates as in Example 12.9, applied to the numbers $\|x_n\|$; the axioms of Definition 5.17 hold slotwise. Completeness is the argument of Proposition 12.10 verbatim, with the modulus $|x_n|$ replaced by the norm $\|x_n\|$ and the completeness of $\mathbb{C}$ replaced by that of $\mathcal{H}_n$: each slot converges, the finite truncations give a uniform bound, and the limit lies in the sum. The last claim is immediate, distinct slots pairing to zero. Note that ℓ^2 is the special case $\mathcal{H}_n = \mathbb{C}$ for every n. □

Definition 12.86 (Internal orthogonal decomposition)**.** Closed subspaces $M_1, M_2, \dots$ of $\mathcal{H}$ are said to *decompose* $\mathcal{H}$, written $\mathcal{H} = \bigoplus_n M_n$, if they are pairwise orthogonal and the only vector orthogonal to all of them is 0.

Rests on Definitions 12.16 and 12.29.

Proposition 12.87 (Expansion in an orthogonal decomposition)**.** *If $\mathcal{H} = \bigoplus_n M_n$ with orthogonal projections $P_n = P_{M_n}$, then $P_n P_m = \delta_{nm} P_n$ and, for every $x \in \mathcal{H}$,*

$$x = \sum_n P_n x\,, \qquad \|x\|^2 = \sum_n \|P_n x\|^2\,, \tag{12.59}$$

the series converging in norm. The map $x \longmapsto (P_1 x, P_2 x, \dots)$ is an inner-product-preserving bijection onto $\bigoplus_n M_n$ in the sense of Definition 12.84.

Rests on Definition 12.86, Theorem 12.30, and Proposition 12.85.

Proof. *Derives Proposition 12.87.*

Orthogonality of the subspaces gives $\operatorname{im} P_m \subseteq M_m \subseteq M_n^{\perp} = \ker P_n$ for $m \neq n$, which is $P_n P_m = 0$; and $P_n^2 = P_n$ is Proposition 12.21.

Choose an orthonormal basis of each M_n — possible in the separable case by Corollary 12.24, and in general by the same argument applied to a maximal orthonormal family — and let $\{e_k\}$ be the union of all of them, which is orthonormal because the M_n are pairwise orthogonal. This family is maximal: a vector orthogonal to every e_k is orthogonal to every M_n, hence 0 by Definition 12.86. So Theorem 12.30 applies, and grouping the terms of Equation (12.18) by the subspace they belong to — legitimate because the partial sums converge absolutely in the sense of Proposition 12.28, whose criterion is insensitive to order — gives Equation (12.59), the group of index n summing to $P_n x$ by Proposition 12.27. The norm identity is Equation (12.19) grouped the same way, and it says precisely that the displayed map lands in Equation (12.58) and is isometric; surjectivity follows from Proposition 12.28 applied slotwise. □

Definition 12.88 (Reducing subspace)**.** A closed subspace M *reduces* $A \in \mathcal{B}(\mathcal{H})$ if $A(M) \subseteq M$ and $A(M^{\perp}) \subseteq M^{\perp}$.
Rests on Definitions 12.20 and 12.35.

Proposition 12.89 (Reduction is commutation)**.** *A closed subspace M reduces $A \in \mathcal{B}(\mathcal{H})$ if and only if $P_M A = A P_M$.*
Rests on Definition 12.88 and Proposition 12.21.

Proof. *Derives Proposition 12.89.*
Write $P = P_M$. ($\Leftarrow$) If $x \in M$ then $Ax = APx = PAx \in M$; if $x \in M^{\perp}$ then $Px = 0$, so $PAx = APx = 0$ and $Ax \in M^{\perp}$ by Proposition 12.21.
($\Rightarrow$) For arbitrary x split $x = Px + (\mathbb{1} - P)x$. The first term lies in M, so $APx \in M$ and $PAPx = APx$; the second lies in $M^{\perp}$, so $A(\mathbb{1} - P)x \in M^{\perp}$ and $PA(\mathbb{1} - P)x = 0$. Adding, $PAx = PAPx + PA(\mathbb{1} - P)x = APx$. □

The extreme case — no proper reducing subspace at all — is what representation theory calls irreducibility, and Proposition 12.89 converts it into a statement about commuting operators that can actually be used.

Definition 12.90 (Self-adjoint family; commutant; irreducibility)**.** A set $\mathcal{S} \subseteq \mathcal{B}(\mathcal{H})$ is *self-adjoint* if $A \in \mathcal{S}$ implies $A^{\dagger} \in \mathcal{S}$. Its *commutant* is

$$\mathcal{S}' = \{B \in \mathcal{B}(\mathcal{H}) \;\mid\; BA = AB \;\; \forall\, A \in \mathcal{S}\}\,, \tag{12.60}$$

and $\mathcal{S}$ *acts irreducibly* on $\mathcal{H}$ if the only closed subspaces reducing every member of $\mathcal{S}$ are $\{0\}$ and $\mathcal{H}$.
Rests on Definitions 12.41 and 12.88.

Theorem 12.91 (Schur's lemma, commutant form)**.** *Let $\mathcal{S} \subseteq \mathcal{B}(\mathcal{H})$ be self-adjoint. Then $\mathcal{S}$ acts irreducibly if and only if $\mathcal{S}' = \mathbb{C}\,\mathbb{1}$: every bounded operator commuting with every member of an irreducible self-adjoint family is a multiple of the identity, and conversely.*
Rests on Definition 12.90, Proposition 12.89, and Theorem 12.59.

Proof. *Derives Theorem 12.91.*
($\Rightarrow$) Let $B \in \mathcal{S}'$. Taking adjoints in $BA = AB$ and using Equation (12.24) gives $A^{\dagger}B^{\dagger} = B^{\dagger}A^{\dagger}$ for every $A \in \mathcal{S}$; since $\mathcal{S}$ is self-adjoint this says $B^{\dagger} \in \mathcal{S}'$, and hence the self-adjoint operators $\frac{1}{2}(B + B^{\dagger})$ and $\frac{1}{2\mathrm{i}}(B - B^{\dagger})$ lie in $\mathcal{S}'$ too. As B is their combination $B_1 + \mathrm{i}B_2$, it is enough to treat a self-adjoint $B \in \mathcal{S}'$.

Let E be the projection-valued measure of B (Theorem 12.59). Every $A \in \mathcal{S}$ is a bounded operator commuting with B, so by the last clause of that theorem it commutes with every $E(\Omega)$; by Proposition 12.89 the range of $E(\Omega)$ therefore reduces every member of $\mathcal{S}$, and irreducibility forces $E(\Omega) \in \{0, \mathbb{1}\}$ for every Borel Ω.

The spectrum $\sigma(B)$ is compact and non-empty (Theorem 12.54) and E is supported on it, so $\lambda \longmapsto E((-\infty, \lambda])$ passes from 0 to $\mathbb{1}$ within a bounded range and $c = \inf\{\lambda \in \mathbb{R} \mid E((-\infty, \lambda]) = \mathbb{1}\}$ is finite. The sets $(-\infty, c - 1/n]$ increase to $(-\infty, c)$ and each carries $E = 0$; the sets $(c + 1/n, \infty)$ increase to (c, ∞) and each carries $E = \mathbb{1} - E((-\infty, c + 1/n]) = 0$. Writing an increasing union as the disjoint union of its successive differences and applying Equation (12.36) gives $E((-\infty, c)) = 0$ and $E((c, \infty)) = 0$, whence $E(\{c\}) = \mathbb{1}$. Then Equation (12.37) reads $\langle x|By\rangle = c\,\langle x|y\rangle$ for all x, y, that is $B = c\,\mathbb{1}$, with c real because B is self-adjoint.

($\Leftarrow$) If a closed subspace M reduces every member of $\mathcal{S}$ then $P_M \in \mathcal{S}'$ by Proposition 12.89, so $P_M = c\,\mathbb{1}$; squaring, $c^2 = c$, so $c \in \{0, 1\}$ and M is $\{0\}$ or $\mathcal{H}$. □

Remark 12.92 (The finite-dimensional lemma, and the unbounded case). Theorem 12.91 is the Hilbert-space form of Theorem 5.155, the second lemma of Schur for finite-dimensional group representations; the hypothesis that has to be added in infinite dimension is that $\mathcal{S}$

be closed under adjoints, without which the argument above cannot pass to the self-adjoint parts of B and the conclusion is false. The self-adjointness costs nothing in practice, because the families that arise are families of unitaries or of self-adjoint operators. That matters for the way the lemma is used on canonical pairs: position and momentum are unbounded (Corollary 12.76) and so are not members of $\mathcal{B}(\mathcal{H})$ at all, but the unitary groups they generate are, and it is to those that Theorem 12.91 is applied in Section 12.7.

Remark 12.93 (Where decompositions come from)*.* Three sources of orthogonal decompositions matter later, and all three are Proposition 12.87 with a different supply of subspaces. The spectral projections $E(\Omega)$ of Theorem 12.59 decompose $\mathcal{H}$ over any Borel partition of the spectrum, and by Proposition 12.89 every such piece reduces the operator — the abstract form of "diagonalise, then work block by block". The eigenspaces of the total angular momentum decompose the state space of a rotationally invariant system, which is how Chapter 79 reduces a problem to one value of ℓ at a time, and the same decomposition organises the hydrogen spectrum of Chapter 81. And a superselection rule is the statement that a decomposition is respected by *every* observable, so that no superposition across the summands is ever prepared or detected; Chapter 85 takes that up.

12.6.2 Tensor products

The direct sum adds dimensions; the tensor product multiplies them (Remark 5.102), and it is the tensor product that describes a composite system. The algebraic construction and its universal property are those of Chapter 5; what is new here is the inner product and the completion.

Definition 12.94 (Tensor product of Hilbert spaces)**.** Let $\mathcal{H}_1, \mathcal{H}_2$ be Hilbert spaces and let $\mathcal{H}_1 \odot \mathcal{H}_2$ be their algebraic tensor product, the space of finite sums $\sum_i u_i \otimes v_i$ characterised by the universal property that every bilinear map out of $\mathcal{H}_1 \times \mathcal{H}_2$ factors uniquely through it. On it define

$$\langle u_1 \otimes v_1 | u_2 \otimes v_2 \rangle = \langle u_1 | u_2 \rangle \, \langle v_1 | v_2 \rangle \, , \tag{12.61}$$

extended to finite sums by sesquilinearity. The *tensor product* $\mathcal{H}_1 \otimes \mathcal{H}_2$ is the completion of $\mathcal{H}_1 \odot \mathcal{H}_2$ in the associated norm.

Rests on Definitions 5.17, 5.106 and 12.2.

Proposition 12.95 (The tensor inner product is well defined and positive definite)**.** *Equation* (12.61) *extends to a genuine inner product on $\mathcal{H}_1 \odot \mathcal{H}_2$, and if $\{e_i\}$ and $\{f_j\}$ are orthonormal bases of $\mathcal{H}_1$ and $\mathcal{H}_2$ then $\{e_i \otimes f_j\}$ is an orthonormal basis of $\mathcal{H}_1 \otimes \mathcal{H}_2$.*

Rests on Definition 12.94, Theorem 12.30, and Proposition 12.23.

Proof. *Derives Proposition 12.95.*

Well defined. The right-hand side of Equation (12.61) is bilinear in (u_2, v_2) and antibilinear in (u_1, v_1), so by the universal property it factors through the algebraic tensor product in each argument separately; the resulting form on $\mathcal{H}_1 \odot \mathcal{H}_2$ does not depend on the representation of an element as a finite sum.

Positive definite. Let $w = \sum_{i=1}^{r} u_i \otimes v_i$ be non-zero. Applying Proposition 12.23 to the v_i and re-expanding, we may assume the v_i orthonormal, in which case Equation (12.61) gives $\langle w | w \rangle = \sum_i \|u_i\|^2 \geq 0$, with equality only if every $u_i = 0$, i.e. only if $w = 0$. The remaining axioms of Definition 5.17 are inherited from the two factors.

Basis. The family $\{e_i \otimes f_j\}$ is orthonormal by Equation (12.61). It is maximal: its closed span contains every $u \otimes v$, because expanding u and v by Equation (12.18) and using continuity of Equation (12.61) in each slot gives $u \otimes v = \sum_{i,j} \langle e_i | u \rangle \, \langle f_j | v \rangle \, e_i \otimes f_j$; hence it contains $\mathcal{H}_1 \odot \mathcal{H}_2$, which is dense in the completion. Now apply criterion (2) of Theorem 12.30. □

Proposition 12.96 (Operators on a tensor product)**.** *For $A_1 \in \mathcal{B}(\mathcal{H}_1)$ and $A_2 \in \mathcal{B}(\mathcal{H}_2)$ there is a unique $A_1 \otimes A_2 \in \mathcal{B}(\mathcal{H}_1 \otimes \mathcal{H}_2)$ with $(A_1 \otimes A_2)(u \otimes v) = A_1 u \otimes A_2 v$, and*

$$\|A_1 \otimes A_2\| = \|A_1\|\,\|A_2\|\,, \qquad (A_1 \otimes A_2)^\dagger = A_1^\dagger \otimes A_2^\dagger\,. \tag{12.62}$$

Rests on Propositions 12.39 and 12.95.

Proof. *Derives Proposition 12.96.*

Uniqueness is density of $\mathcal{H}_1 \odot \mathcal{H}_2$ and continuity. For existence and the norm, factor $A_1 \otimes A_2 = (A_1 \otimes \mathbb{1})(\mathbb{1} \otimes A_2)$ and bound each factor separately. Expanding $w = \sum_j u_j \otimes f_j$ in the orthonormal basis $\{f_j\}$ of $\mathcal{H}_2$ — possible by Proposition 12.95 — gives $\|w\|^2 = \sum_j \|u_j\|^2$ and

$$\|(A_1 \otimes \mathbb{1})w\|^2 = \sum_j \|A_1 u_j\|^2 \le \|A_1\|^2 \sum_j \|u_j\|^2 = \|A_1\|^2\,\|w\|^2\,,$$

so $\|A_1 \otimes \mathbb{1}\| \le \|A_1\|$, and symmetrically for the other factor; hence $\|A_1 \otimes A_2\| \le \|A_1\|\,\|A_2\|$, which also shows the map is bounded on the dense subspace and so extends. The reverse inequality follows by testing on product vectors, $\|A_1 u \otimes A_2 v\| = \|A_1 u\|\,\|A_2 v\|$, and taking suprema. The adjoint identity is Equation (12.61) checked on product vectors, together with uniqueness in Theorem 12.38. □

Remark 12.97 (The universal property does not survive completion)*.* It is worth stating what Definition 12.94 does and does not claim. The universal property — every bilinear map factors uniquely — belongs to the *algebraic* tensor product $\mathcal{H}_1 \odot \mathcal{H}_2$, and it is what makes Equation (12.61) well defined. The completion $\mathcal{H}_1 \otimes \mathcal{H}_2$ does *not* inherit it for arbitrary bounded bilinear maps: a bilinear B with $\|B(u,v)\| \le C\,\|u\|\,\|v\|$ need not extend continuously to the completion, because the Hilbert norm on $\mathcal{H}_1 \odot \mathcal{H}_2$ is not the largest such norm. The Hilbert tensor product is one of several completions of the same algebraic object, singled out by Equation (12.61) and by nothing else, and it is singled out for a physical reason: it is the one whose orthonormal basis is the set of products of basis vectors (Proposition 12.95), which is the statement that a joint state is specified by an amplitude for each pair of outcomes.

Example 12.98 (Two particles in three-dimensional space)**.** For a single particle in ordinary space the state space is $L^2(\mathbb{R}^3)$, and for two distinguishable particles it is

$$L^2(\mathbb{R}^3) \otimes L^2(\mathbb{R}^3) \cong L^2(\mathbb{R}^6)\,, \tag{12.63}$$

the isomorphism carrying $f \otimes g$ to the function $(x_1, x_2) \longmapsto f(x_1)\,g(x_2)$. It is an isomorphism because the products $f(x_1)g(x_2)$ span a dense subspace of $L^2(\mathbb{R}^6)$ and because Equation (12.61) matches the Fubini factorisation of the integral over $\mathbb{R}^6$ [1481]. Note that the configuration space of the pair is $\mathbb{R}^6$, not $\mathbb{R}^3$: a two-particle wavefunction is a function of six coordinates and is not a field on space. That is a statement about the formalism with observable consequences, and it is where the non-classical correlations begin.

Rests on Definition 12.94 and Example 12.11.

Definition 12.99 (Product and entangled vectors)**.** A vector $w \in \mathcal{H}_1 \otimes \mathcal{H}_2$ is a *product vector* if $w = u \otimes v$ for some u, v, and is *entangled* otherwise.

Rests on Definition 12.94.

Example 12.100 (Entangled vectors exist)**.** Let $\mathcal{H}_1 = \mathcal{H}_2 = \mathbb{C}^2$ with orthonormal bases $\{e_0, e_1\}$ and $\{f_0, f_1\}$. The unit vector

$$w = \frac{1}{\sqrt{2}}\,(e_0 \otimes f_0 + e_1 \otimes f_1) \tag{12.64}$$

is entangled.

Rests on Definition 12.99 and Proposition 12.95.

Proof. *Derives Example 12.100.*

Suppose $w = (ae_0 + be_1) \otimes (cf_0 + df_1)$. Expanding in the basis $\{e_i \otimes f_j\}$, which is a basis by Proposition 12.95, and comparing coefficients with Equation (12.64) gives

$$ac = bd = \frac{1}{\sqrt{2}}\,, \qquad ad = bc = 0\,.$$

From $ad = 0$ either $a = 0$, contradicting $ac = 1/\sqrt{2}$, or $d = 0$, contradicting $bd = 1/\sqrt{2}$. □

Remark 12.101 (Composite systems). Definition 12.94 is the mathematics behind the composition postulate: the state space of a system assembled from two others is the tensor product of their state spaces, an assumption already explicit in von Neumann [706] and stated as a postulate in Chapter 77. Example 12.100 is then not a curiosity but a forced consequence: the tensor product contains vectors that are not products, so a composite system has states in which neither part has a state of its own. Whether nature actually realises them is a question for experiment, and it has been settled affirmatively; the argument that turns Equation (12.64) into a testable inequality is Chapter 84, and the measurements are in Chapter 89. For identical rather than merely distinguishable particles the state space is not the whole tensor product but a symmetric or antisymmetric subspace of it, which is a reducing subspace in the sense of Definition 12.88; Chapter 83 takes that up.

12.6.3 Rigged Hilbert spaces

Two objects used constantly in quantum mechanics are not elements of $\mathcal{H}$: the plane wave, which is not square integrable, and the delta "function", which is not a function. Neither is an approximation to be apologised for; both are legitimate objects, and the structure they live in is a triple of spaces rather than a single one.

Proposition 12.102 (Neither plane waves nor deltas are in L^2)**.** *(i) The function $x \longmapsto \mathrm{e}^{\mathrm{i}px/\hbar}$ is not in $L^2(\mathbb{R})$ for any real p. (ii) There is no $g \in L^2(\mathbb{R})$ with $\langle g|\varphi\rangle = \varphi(0)$ for every continuous, compactly supported φ.*

Rests on Example 12.11 and Proposition 12.4.

Proof. *Derives Proposition 12.102.*

(i) The integrand $\left|\mathrm{e}^{\mathrm{i}px/\hbar}\right|^2 = 1$, so the integral over $\mathbb{R}$ diverges.

(ii) Suppose such a g existed. Let φ_n be the continuous function equal to 1 at 0, vanishing outside $[-1/n, 1/n]$, and linear in between, so $\varphi_n(0) = 1$ while $\|\varphi_n\|^2 = \int |\varphi_n|^2 \le 2/n$. Then Equation (12.1) gives $1 = |\langle g|\varphi_n\rangle| \le \|g\|\,\sqrt{2/n}$ for every n, which is false for large n. □

Definition 12.103 (Gelfand triple)**.** A *Gelfand triple*, or *rigged Hilbert space*, is a chain

$$\Phi \subseteq \mathcal{H} \subseteq \Phi'\,, \tag{12.65}$$

in which Φ is a dense subspace of $\mathcal{H}$ carrying its own, strictly finer, topology under which it is a nuclear space, and Φ' is the space of continuous *linear* functionals on Φ for that topology. The middle inclusion is the Riesz identification of Corollary 12.47: each $x \in \mathcal{H}$ acts on Φ by $\varphi \longmapsto \langle x|\varphi\rangle$, which is linear in φ by Equation (5.34) and therefore is an element of Φ', and that action determines x because Φ is dense. The identification $x \longmapsto \langle x|\cdot\rangle$ is itself antilinear in x, exactly as Corollary 12.47 says; it is the functionals, not the embedding, that are linear.

Rests on Corollaries 12.19 and 12.47 and Definition 12.45.

Example 12.104 (The Schwartz triple)**.** The standard triple of quantum mechanics on the line is

$$\mathcal{S}(\mathbb{R}) \subseteq L^2(\mathbb{R}) \subseteq \mathcal{S}'(\mathbb{R})\,, \tag{12.66}$$

where $\mathcal{S}(\mathbb{R})$ is the Schwartz space of smooth functions all of whose derivatives decay faster than any power, and $\mathcal{S}'(\mathbb{R})$ is the space of tempered distributions [971]. Both objects of Proposition 12.102 live in the third space: the delta is the functional $\varphi \longmapsto \varphi(0)$, which is continuous for the topology of $\mathcal{S}$ though not for the L^2 norm, and the plane wave is the functional

$$F_p(\varphi) = \frac{1}{\sqrt{2\pi\hbar}} \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}px/\hbar}\, \varphi(x)\, \mathrm{d}x\,, \tag{12.67}$$

which is the Fourier transform of Chapter 17 evaluated at the wavenumber $p/\hbar$.

Rests on Definition 12.103, Proposition 12.102, and Example 12.11.

Definition 12.105 (Generalized eigenvector)**.** Let A map Φ into Φ continuously. A functional $F \in \Phi'$ is a *generalized eigenvector* of A with eigenvalue λ if

$$F(A\varphi) = \lambda\, F(\varphi)\,, \qquad \forall\, \varphi \in \Phi\,. \tag{12.68}$$

Rests on Definitions 12.50 and 12.103.

Proposition 12.106 (The plane wave is a generalized momentum eigenvector)**.** *On the triple Equation (12.66), with $P = -\mathrm{i}\hbar\,\mathrm{d}/\mathrm{d}x$ mapping $\mathcal{S}(\mathbb{R})$ into itself, the functional Equation (12.67) satisfies $F_p(P\varphi) = p\,F_p(\varphi)$ for every $\varphi \in \mathcal{S}(\mathbb{R})$, and p ranges over all of $\mathbb{R}$.*

Rests on Definition 12.105, Example 12.104, and Equation (12.54).

Proof. *Derives Proposition 12.106.*

Integrate by parts in Equation (12.67); the boundary terms vanish because φ and all its derivatives decay faster than any power. Writing $C = (2\pi\hbar)^{-1/2}$,

$$F_p(P\varphi) = C \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}px/\hbar}\, \bigl(-\mathrm{i}\hbar\varphi'(x)\bigr)\, \mathrm{d}x = C \int_{\mathbb{R}} \mathrm{i}\hbar\, \frac{\mathrm{d}}{\mathrm{d}x}\Bigl(\mathrm{e}^{-\mathrm{i}px/\hbar}\Bigr)\, \varphi(x)\, \mathrm{d}x\,,$$

and $\mathrm{i}\hbar\,\mathrm{d}(\mathrm{e}^{-\mathrm{i}px/\hbar})/\mathrm{d}x = \mathrm{i}\hbar\,(-\mathrm{i}p/\hbar)\,\mathrm{e}^{-\mathrm{i}px/\hbar} = p\,\mathrm{e}^{-\mathrm{i}px/\hbar}$, which gives $p\,F_p(\varphi)$. □

Theorem 12.107 (Nuclear spectral theorem)**.** *Let $\Phi \subseteq \mathcal{H} \subseteq \Phi'$ be a Gelfand triple with Φ nuclear, and let A be a self-adjoint operator on $\mathcal{H}$ mapping Φ continuously into itself. Then A possesses a* complete *set of generalized eigenvectors: there is a measure μ on $\sigma(A)$ and a family $\{F_\lambda\}$ in Φ', with F_λ a generalized eigenvector of eigenvalue λ for μ-almost every λ, such that*

$$\langle \varphi | \psi \rangle = \int_{\sigma(A)} F_\lambda(\varphi)^*\, F_\lambda(\psi)\, \mathrm{d}\mu(\lambda)\,, \qquad \forall\, \varphi, \psi \in \Phi\,. \tag{12.69}$$

Rests on Definitions 12.103 and 12.105 and Theorem 12.59.

The proof is in Appendix A.27: it builds the direct-integral decomposition supplied by the spectral theorem Theorem 12.59, and then uses nuclearity of the test space to realise the fibre maps as continuous functionals, so that almost every fibre furnishes a generalized eigenvector. This is the one long derivation of the chapter that rests on a theory this treatise does not develop — the theory of nuclear spaces, for which Gelfand and Vilenkin, volume 4, is the reference of record [1276]. The appendix isolates exactly what is assumed, in one displayed statement: that a nuclear test space embeds in $\mathcal{H}$ through some Hilbert–Schmidt map. Everything else is derived, and two things the statement above suppresses are made explicit there: $\mathcal{H}$ and Φ are assumed separable, and when the spectrum of A has multiplicity greater than one the integrand of Equation (12.69) carries a sum over a multiplicity index — the display as written is the simple-spectrum case, which is the case of Proposition 12.106 and of every use made of the theorem in this treatise.

Full derivation: Appendix A.27 (p. 2910).

Derives Theorem 12.107.

Remark 12.108 (The working rule). Equation (12.69) is the completeness relation that physics writes as $\int |\lambda\rangle\langle\lambda| \, \mathrm{d}\mu(\lambda) = \mathbb{1}$, and Theorem 12.107 is what licenses it. The rule this treatise follows is therefore precise, and it is worth stating once for all the later chapters that use it.

- A ket $|\lambda\rangle$ belonging to a point of continuous spectrum denotes an element of Φ', not of $\mathcal{H}$. It is not a state, it cannot be normalised, and $\langle\lambda|\lambda\rangle$ is meaningless. Proposition 12.102 is the proof that no other reading is available.
- An expression $\langle\lambda|\varphi\rangle$ is legitimate whenever $\varphi \in \Phi$ — it is the value $F_\lambda(\varphi)$ — and the "normalisation" $\langle\lambda|\lambda'\rangle = \delta(\lambda - \lambda')$ is shorthand for Equation (12.69), an identity between functionals and never between numbers.
- A physical state is always a vector of $\mathcal{H}$. A wave packet built by superposing generalized eigenvectors against a square-integrable amplitude is such a vector; a single plane wave is not, which is why every scattering calculation in Chapter 86 that is done with plane waves must be read as a statement about packets.

The triple also explains, retrospectively, why Dirac's formalism worked before it had a foundation [646]: he was computing in Φ' with functionals that the nuclear spectral theorem later showed to exist. The rigorous setting is due to Gelfand and his collaborators [1276], built on Schwartz's theory of distributions [971]. It is a completion of the formalism, not a correction to it, and no result of this chapter is disturbed by it: the spectral theorem Theorem 12.59 remains the statement that carries the physics, and Equation (12.69) is a convenient way of writing its conclusion in the notation physicists actually use.

12.7 The canonical commutation relations in Weyl form

Theorem 12.75 shows that no pair of bounded operators satisfies $[\hat{q}, \hat{p}] = \mathrm{i}\hbar\mathbb{1}$, so any realisation of that relation involves unbounded operators, each with a domain that is not the whole space (Corollary 12.76). This is not a technicality to be waved through. Two pairs can satisfy the relation on every vector where both sides are defined and still be inequivalent, because the relation says nothing about what happens off the common domain: momentum on a half-line has no self-adjoint realisation at all (Example 12.82), and momentum on an interval has a whole circle of them (Example 12.81), and in neither case does the formal commutator notice. The repair is to exponentiate, replacing the unbounded generators by the unitary groups they generate — which by Theorem 12.66 carry exactly the same information — and to impose the commutation relation on those.

Definition 12.109 (Weyl system)**.** A *Weyl system* of one degree of freedom on $\mathcal{H}$ is a pair of strongly continuous one-parameter unitary groups (Definition 12.64)

$$U(\alpha) = \exp\left(\frac{\mathrm{i}\alpha Q}{\hbar}\right), \qquad V(\beta) = \exp\left(-\frac{\mathrm{i}\beta P}{\hbar}\right), \tag{12.70}$$

with Q and P the self-adjoint generators supplied by Theorem 12.66, satisfying the *Weyl relation*

$$U(\alpha)\,V(\beta) = \mathrm{e}^{\mathrm{i}\alpha\beta/\hbar}\,V(\beta)\,U(\alpha), \qquad \forall\,\alpha,\beta \in \mathbb{R}. \tag{12.71}$$

Here Q carries the SI unit m and P the unit kg m/s, so that the parameters carry kg m/s and m respectively and both $\alpha Q/\hbar$ and $\alpha\beta/\hbar$ are pure numbers. For f degrees of freedom the parameters become vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^f$, the exponents become $\boldsymbol{\alpha}\cdot\boldsymbol{Q}$ and $\boldsymbol{\beta}\cdot\boldsymbol{P}$, the phase in Equation (12.71) becomes $\mathrm{e}^{\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}/\hbar}$, and the groups belonging to different degrees of freedom are required to commute.

Rests on Definition 12.64 and Theorems 12.66 and 12.75.

Example 12.110 (The Schrödinger system)**.** On $\mathcal{H} = L^2(\mathbb{R})$ put

$$\big(U(\alpha)\psi\big)(x) = \mathrm{e}^{\mathrm{i}\alpha x/\hbar}\,\psi(x)\,, \qquad \big(V(\beta)\psi\big)(x) = \psi(x-\beta)\,. \tag{12.72}$$

Both are unitary, both are groups, and both are strongly continuous — the first by dominated convergence, the second because translation is continuous in the mean. Their generators are multiplication by x and $-\mathrm{i}\hbar\,\mathrm{d}/\mathrm{d}x$, on the domains where the corresponding limits Equation (12.44) exist. Applying the two definitions in the two orders, $(U(\alpha)V(\beta)\psi)(x) = \mathrm{e}^{\mathrm{i}\alpha x/\hbar}\psi(x-\beta)$ while $(V(\beta)U(\alpha)\psi)(x) = \mathrm{e}^{\mathrm{i}\alpha(x-\beta)/\hbar}\psi(x-\beta)$, so Equation (12.71) holds. This system is also irreducible in the sense of Definition 12.90 — a fact taken here from [1481], theorem VIII.14 and its preamble, not proved: the step that does the work is that a bounded operator commuting with multiplication by every bounded measurable function is itself such a multiplication, which is the maximal abelian property of $L^\infty(\mathbb{R})$ acting on $L^2(\mathbb{R})$, and only then does commuting with every translation force the multiplier to be constant.

Rests on Definition 12.109 and Example 12.11.

Proposition 12.111 (The Weyl relation is a covariance statement)**.** *In any Weyl system,*

$$V(\beta)^\dagger\, Q\, V(\beta) = Q + \beta\,\mathbb{1}\,, \qquad U(\alpha)^\dagger\, P\, U(\alpha) = P - \alpha\,\mathbb{1}\,, \tag{12.73}$$

the domains being $V(-\beta)D(Q)$ and $U(-\alpha)D(P)$ respectively. Conversely either identity in Equation (12.73) implies Equation (12.71).

Rests on Definition 12.109, Theorem 12.66, and Proposition 12.67.

Proof. *Derives Proposition 12.111.*

Multiply Equation (12.71) on the left by $V(\beta)^\dagger = V(-\beta)$:

$$V(\beta)^\dagger\, U(\alpha)\, V(\beta) = \mathrm{e}^{\mathrm{i}\alpha\beta/\hbar}\, U(\alpha)\,. \tag{12.74}$$

For fixed β the left-hand side is, as a function of α, a strongly continuous one-parameter unitary group: the group law and the continuity are inherited from U because conjugation by a fixed unitary is multiplicative and isometric. Its generator is $V(\beta)^\dagger Q V(\beta)$ on the domain $V(-\beta)D(Q)$, because

$$\frac{V(\beta)^\dagger U(\alpha) V(\beta)x - x}{\alpha} = V(\beta)^\dagger\,\frac{U(\alpha)\big(V(\beta)x\big) - V(\beta)x}{\alpha}\,, \tag{12.75}$$

and $V(\beta)^\dagger$ is isometric, so the limit on the left exists exactly when the limit Equation (12.44) defining Q exists at the vector $V(\beta)x$, and is then its image under $V(\beta)^\dagger$. The right-hand side of Equation (12.74) is likewise such a group, and its generator is $Q + \beta\mathbb{1}$ on $D(Q)$: the summand $\beta\mathbb{1}$ is bounded and self-adjoint, and by Proposition 12.65 it contributes exactly the constant phase $\mathrm{e}^{\mathrm{i}\alpha\beta/\hbar}$. Two groups that are equal have equal generators, by the bijection of Theorem 12.66; that is the first identity of Equation (12.73), and the second follows in the same way from Equation (12.71) rewritten as $U(\alpha)^\dagger V(\beta) U(\alpha) = \mathrm{e}^{-\mathrm{i}\alpha\beta/\hbar} V(\beta)$. Conversely, exponentiating either identity of Equation (12.73) through the bijection of Theorem 12.66 returns Equation (12.74), hence Equation (12.71). □

Corollary 12.112 (Commutator of the momentum with a function of the position)**.** *Let (U, V) be a Weyl system and let*

$$g(\lambda) = \int_{\mathbb{R}} m(\alpha)\,\mathrm{e}^{\mathrm{i}\alpha\lambda/\hbar}\,\mathrm{d}\alpha\,, \qquad \int_{\mathbb{R}} (1+|\alpha|)\,|m(\alpha)|\,\mathrm{d}\alpha < \infty\,, \tag{12.76}$$

and set $g(Q) = \int_{\mathbb{R}} m(\alpha)\,U(\alpha)\,\mathrm{d}\alpha$, a bounded operator defined by the convergent vector-valued integral. Then for every $x \in D(P)$ one has $g(Q)x \in D(P)$ and

$$[P, g(Q)]\; x = -\mathrm{i}\hbar\, g'(Q)\, x\,, \tag{12.77}$$

where $g'(Q) = \int_{\mathbb{R}} (\mathrm{i}\alpha/\hbar)\, m(\alpha)\, U(\alpha)\,\mathrm{d}\alpha$ is the operator built the same way from g'.

Rests on Proposition 12.111, Theorem 12.66, and Definition 12.109.

Proof. *Derives Corollary 12.112.*

By Equation (12.74) and linearity of the integral,

$$V(\beta)^{\dagger}\, g(Q)\, V(\beta) = \int_{\mathbb{R}} m(\alpha)\, \mathrm{e}^{\mathrm{i}\alpha\beta/\hbar}\, U(\alpha)\, \mathrm{d}\alpha\,, \tag{12.78}$$

which is the operator built the same way from the translate $\lambda \longmapsto g(\lambda+\beta)$. Subtracting $g(Q)$, dividing by β and comparing with $g'(Q)$, the operator norm of the difference is at most

$$\int_{\mathbb{R}} |m(\alpha)| \left| \frac{\mathrm{e}^{\mathrm{i}\alpha\beta/\hbar}-1}{\beta} - \frac{\mathrm{i}\alpha}{\hbar} \right| \mathrm{d}\alpha\,, \tag{12.79}$$

because every $U(\alpha)$ has norm 1. The integrand tends to zero pointwise as $\beta \to 0$ and is bounded by $2\,|\alpha|\,|m(\alpha)|/\hbar$, using $\left|\mathrm{e}^{\mathrm{i}\theta}-1\right| \le |\theta|$; that bound is integrable by the hypothesis in Equation (12.76), so dominated convergence (Remark 12.1) sends Equation (12.79) to zero. The right-hand side of Equation (12.78) is therefore differentiable in β at $\beta = 0$ in the operator norm, with derivative $g'(Q)$.

Now fix $x \in D(P)$ and write, for $\beta \neq 0$,

$$\frac{V(-\beta)\, g(Q)\, V(\beta)x - g(Q)x}{\beta} = V(-\beta)\, g(Q)\, \frac{V(\beta)x - x}{\beta} + \frac{V(-\beta)\, g(Q)x - g(Q)x}{\beta}\,. \tag{12.80}$$

The left-hand side converges to $g'(Q)x$ by the previous paragraph. In the first term on the right, $(V(\beta)x - x)/\beta \to -\mathrm{i}Px/\hbar$ because $x \in D(P)$ and V has generator P (Equation (12.44)), $g(Q)$ is bounded, and $V(-\beta) \to \mathbb{1}$ strongly, so that term converges to $-\mathrm{i}g(Q)Px/\hbar$. The second term therefore converges as well, and by Equation (12.44) applied to the group $V(-\cdot)$ its convergence *is* the statement that $g(Q)x \in D(P)$, the limit being $+\mathrm{i}P\, g(Q)x/\hbar$. Collecting,

$$g'(Q)x = \frac{\mathrm{i}}{\hbar}\Big(P\, g(Q) - g(Q)\, P\Big)x = \frac{\mathrm{i}}{\hbar}\,[P, g(Q)]\, x\,,$$

which is Equation (12.77). □

Remark 12.113 (Scope of the commutator identity). Equation (12.77) is the identity every derivation of Ehrenfest's theorem uses, and the reason for proving it in the class Equation (12.76) rather than for a general function is that this class needs nothing but the Weyl relation: the operator $g(Q)$ is built from the group itself, so no functional calculus for the unbounded Q has to be invoked and no question of domains arises beyond the one settled in the proof. The class contains every bounded potential whose Fourier transform decays fast enough — a Gaussian well, a smoothly screened Coulomb potential, any C^2 function of compact support — and for a polynomial the identity is elementary and is obtained directly from $[P, Q] = -\mathrm{i}\hbar$ by induction. For a general bounded Borel g the identity still holds, with $g(Q)$ read through the functional calculus of Definition 12.60 extended to unbounded operators in Appendix A.25; that extension is not carried out here, and where a chapter needs it the fact is flagged.

Theorem 12.114 (Stone–von Neumann)**.** *Let f be finite and let (U, V) be a Weyl system of f degrees of freedom, in the sense of Definition 12.109, on a separable Hilbert space $\mathcal{H}$, acting irreducibly (Definition 12.90). Then there is a unitary $T : \mathcal{H} \longrightarrow L^2(\mathbb{R}^f)$ carrying it to the Schrödinger system Equation (12.72), and T is unique up to a phase. Without the irreducibility assumption, every such Weyl system on a separable space is a direct sum (Definition 12.84) of copies of the Schrödinger system.*

Rests on Definitions 12.90 and 12.109 and Theorem 12.91.

Remark 12.115 (What is quoted here). Theorem 12.114 is quoted, not proved. The argument builds from the Weyl operators a weighted average $\Pi = \iint w(\alpha,\beta)\, U(\alpha)V(\beta)\, \mathrm{d}\alpha\, \mathrm{d}\beta$ with

a Gaussian weight w chosen so that the phase Equation (12.71) reproduces itself on multiplication, shows by that relation alone that $\Pi^2 = \Pi = \Pi^\dagger$ and that Π is not zero, deduces from irreducibility and Theorem 12.91 that its range is one-dimensional, and then shows that the Weyl operators applied to a unit vector of that range generate $\mathcal{H}$ and reproduce the Schrödinger action — whence the unitary. It is functional analysis about the Weyl algebra, and the treatise takes the statement from [1481], theorem VIII.14, where the proof is given in full; the uniqueness question was von Neumann's, raised by Weyl's formulation of the commutation relations in exponentiated form. What is proved here is everything Equation (12.71) yields directly — the covariance Equation (12.73), the commutator Equation (12.77), and the Schur step Theorem 12.91 on which the quoted argument turns — so exactly one thing is owed, and it is named.

Remark 12.116 (Why the theorem is stated for finitely many degrees of freedom)*.* The hypothesis that f be finite is essential and its failure is physics rather than pathology. A field has infinitely many degrees of freedom, the theorem then fails, and inequivalent irreducible representations of the commutation relations exist in profusion; the choice among them becomes a physical question, and it is the origin of superselection sectors, of inequivalent vacua, and of the theorems of Chapter 109. Within its hypotheses the theorem is what makes "the" Schrödinger representation a legitimate phrase: the quantum kinematics of a system with finitely many canonical pairs is fixed, up to unitary equivalence, by the commutation relations alone, which is the statement Chapter 25 needs when it asks what a quantization can and cannot be.

Chapter 13

Differentiable Manifolds, Tensors, and Curvature

The differential geometry of curves and surfaces, and its generalisation to differentiable manifolds of arbitrary dimension, provides the language in which modern physics describes configuration spaces, spacetimes, and gauge fields. This chapter builds that language in the order in which the source manuscript develops it: tensors under orthogonal transformations in Euclidean space; the classical theory of curves and surfaces in $\mathbb{R}^3$, where the metric, the Christoffel symbols, and the Riemann symbols first appear in concrete form; the definition of a differentiable manifold through charts and atlases; tensor analysis on manifolds and under general coordinate transformations; differential forms; the metric tensor; the covariant derivative and affine connections; and the curvature tensors. The chapters on general relativity build directly on this material [1509, 1746].

Notation 13.1. Repeated indices are summed. In the sections on Cartesian tensors and on curve and surface theory all indices are written as subscripts and summation runs over repeated subscripts, as is customary for Cartesian tensors; from the section on differentiable manifolds onwards, coordinate indices are written as superscripts (x^i) and the summation convention pairs one upper with one lower index. Throughout this chapter, the *type* of a tensor is written (r, s) (r contravariant and s covariant slots); the pair (p, q) is reserved for the signature of a metric, with $D = p + q$ the dimension. Where a derivation manipulates both sides of an equation, the operation applied is displayed after the equation being manipulated, e.g. $A = B, \quad \boldsymbol{x}_l \cdot (\ )$ means "take the scalar product of both sides with $\boldsymbol{x}_l$"; this compact notation is inherited from the source manuscript.

13.1 Index conventions and the symmetry ladder

This chapter, and every chapter that builds on it, works in general dimension $D = p + q$ with metric signature (p, q) (project rule, Chapter 1); the observed case is the instantiation $p + q = 3 + 1$ of Parts IV and V. Two families of transformations act on everything we construct: *diffeomorphisms* (general coordinate transformations of the manifold) and *local frame transformations* (rotations of the orthonormal frames introduced in Section 13.8). Moreover, the maximally symmetric geometries of Section 13.13 are most transparently described by embedding the $(p+q)$-dimensional space in flat spaces of one and two dimensions more. We therefore fix, once and for all, the covariant index notation for tensors and connections defined in $(p + q)$-, $(p + q + 1)$- and $(p + q + 2)$-dimensional spaces. The following types of letters are used for indices of (pseudo-)tensors and connections under

diffeomorphisms:

$$\mu, \nu, \ldots = 1, \ldots, p+q\,,$$
$$m, n, \ldots = 1, \ldots, p+q+1\,,$$
$$M, N, \ldots = 1, \ldots, p+q+2\,,$$

while for the indices of (pseudo-)tensors and connections under *local Lorentz, (anti-)de Sitter and conformal transformations* we use

$$a, b, \ldots = 1, \ldots, p+q\,,$$
$$A, B, \ldots = 1, \ldots, p+q+1\,,$$
$$I, J, \ldots = 1, \ldots, p+q+2\,,$$

respectively.

The groups these indices transform under form a ladder that recurs throughout the treatise; we fix its names and symbols in Table 13.1.

Name	Symbol	Group
Poincaré	P_{p+q}	$\mathrm{ISO}(p,q)$
Lorentz	L_{p+q}	$\mathrm{SO}(p,q)$
translations	T_{p+q}	$\mathrm{T}(p,q)$ (abelian)
anti-de Sitter	AdS_{p+q}	$\mathrm{SO}(p,q+1)$
de Sitter	dS_{p+q}	$\mathrm{SO}(p+1,q)$
Conformal	C_{p+q}	$\mathrm{SO}(p+1,q+1)$

Table 13.1: The symmetry ladder of signature (p,q): the groups acting on flat space (Poincaré and its Lorentz and translation subgroups), on the two maximally symmetric curved spaces (anti-de Sitter and de Sitter, realized linearly in $p+q+1$ dimensions), and on the conformal structure of flat space (realized linearly in $p+q+2$ dimensions). See Section 13.13 for the constructions and Chapter 14 for the group theory.

The frame metric of signature (p,q) is fixed as

$$\eta_{ab} = \mathrm{diag}(\underbrace{+1, \ldots, +1}_{p}, \underbrace{-1, \ldots, -1}_{q})\,, \tag{13.1}$$

and likewise η_{AB} and η_{IJ} denote the flat metrics of the signatures indicated by the corresponding group in Table 13.1 — $(p,q+1)$ for AdS_{p+q}, $(p+1,q)$ for dS_{p+q}, and $(p+1,q+1)$ for C_{p+q}. In the relativistic chapters of the physical parts, where $p+q=3+1$ and the signature convention is $(-,+,+,+)$ (Part IV), the dictionary is $(p,q)=(3,1)$ with the roles of the signs read from context; nothing in this chapter depends on the overall sign choice.

13.2 Tensors under orthogonal transformations

Before tensors are defined on a general manifold, it is instructive to meet them in their simplest habitat: Euclidean space referred to orthogonal coordinate systems, where all indices can be written as subscripts and the transformation matrices are constant.

13.2.1 Orthogonal coordinate transformations

Linear transformations

Definition 13.2 (Linear coordinate transformation)**.** We say that a general coordinate transformation is *linear* if and only if it can be written as

$$x'_i = \Lambda_{ij} x_j\,, \tag{13.2}$$

with Λ_{ij} the *transformation matrix*, with constant coefficients.

Differentiating Equation (13.2) we see that

$$\Lambda_{ij} = \frac{\partial x'_i}{\partial x_j} \,. \tag{13.3}$$

From Equations (13.2) and (13.3) we have

$$\begin{aligned} x'_i &= \Lambda_{ij} x_j \\ &= \frac{\partial x'_i}{\partial x_j} x_j \,, \qquad \frac{\partial x_k}{\partial x'_i}(\ ) \\ x'_i \frac{\partial x_k}{\partial x'_i} &= \delta_{jk} x_j = x_k \,. \end{aligned}$$

Renaming indices, we obtain the inverse transformation of the coordinates,

$$x_j = \frac{\partial x_j}{\partial x'_i} x'_i \,. \tag{13.4}$$

We denote

$$\left(\Lambda^{-1}\right)_{ij} = \frac{\partial x_j}{\partial x'_i} \,, \tag{13.5}$$

which is justified because

$$\begin{aligned} \left(\Lambda^{-1}\right)_{ik} \Lambda_{ij} &= \frac{\partial x_k}{\partial x'_i} \frac{\partial x'_i}{\partial x_j} \\ &= \delta_{jk} \,, \end{aligned}$$

where $\left(\Lambda^{-1}\right)_{ik}$ is the *inverse transformation matrix*, with constant coefficients. Therefore, from Equation (13.4),

$$x_j = \left(\Lambda^{-1}\right)_{ij} x'_i \,. \tag{13.6}$$

Orthogonal coordinate systems

Definition 13.3 (Orthogonal coordinate system)**.** Consider an n-dimensional vector space generated by the basis $\{\hat{e}_i\}_{i=1}^n$. We say that the given basis forms an *orthogonal coordinate system* if and only if

$$\hat{e}_i \cdot \hat{e}_j = \delta_{ij} \,. \tag{13.7}$$

We denote the orthogonal coordinate system by K.

Definition 13.4 (Orthogonal coordinate transformation)**.** Consider two orthogonal coordinate systems K and K' with respective bases $\{\hat{e}_i\}_{i=1}^n$ and $\{\hat{e}'_i\}_{i=1}^n$. We call *orthogonal coordinate transformation* the coordinate transformation between the two systems,

$$x'_i = a_{ij} x_j \,, \tag{13.8}$$

where a_{ij} is the *transformation matrix*, with constant coefficients.

Rests on Definition 13.3.

Owing to Equation (13.8), the unit vectors of the new basis are related to the old ones by

$$\hat{e}'_i = a_{ij} \hat{e}_j \,, \tag{13.9}$$

and since the new coordinate system is also orthogonal,

$$\hat{e}'_i \cdot \hat{e}'_j = \delta_{ij} \,. \tag{13.10}$$

Geometric interpretation of the transformation matrix

From Equations (13.7) and (13.9) we see that

$$\begin{aligned}\hat{e}'_i \cdot \hat{e}_j &= (a_{ik}\hat{e}_k) \cdot \hat{e}_j \\ &= a_{ik}\,(\hat{e}_k \cdot \hat{e}_j) \\ &= a_{ik}\delta_{jk} \\ &= a_{ij}\,. \end{aligned} \tag{13.11}$$

On the other hand, let θ_{ij} be the angles between the unit vectors $\hat{e}_j$ and $\hat{e}'_i$. Then

$$\begin{aligned}\hat{e}'_i \cdot \hat{e}_j &= |\hat{e}'_i|\,|\hat{e}_j| \cos\theta_{ij} \\ &= \cos\theta_{ij}\,, \end{aligned}$$

so that, by Equation (13.11),

$$a_{ij} = \cos\theta_{ij}\,. \tag{13.12}$$

The transformation matrix therefore contains the cosines of the angles formed between the unit vectors $\hat{e}_j$ and $\hat{e}'_i$.

Orthogonality condition

Consider two orthogonal coordinate systems K and K' related by an orthogonal transformation of the form Equation (13.8). Since K' is orthogonal, from Equations (13.7), (13.9) and (13.10),

$$\begin{aligned}\delta_{ij} &= \hat{e}'_i \cdot \hat{e}'_j \\ &= a_{il}\hat{e}_l \cdot a_{jm}\hat{e}_m \\ &= a_{il}a_{jm}\delta_{lm}\,. \end{aligned}$$

Hence the *orthogonality condition* reads

$$\delta_{ij} = a_{il}a_{jl}\,. \tag{13.13}$$

In matrix notation this corresponds to

$$AA^{\mathsf{T}} = I\,. \tag{13.14}$$

Taking the determinant of both sides,

$$\begin{aligned}\det(A)\det(A^{\mathsf{T}}) &= 1 \\ (\det(A))^2 &= 1 \\ \det(A) &= \pm 1\,. \end{aligned} \tag{13.15}$$

Since $\det(A) \neq 0$, the inverse A^{-1} exists, with

$$AA^{-1} = I\,. \tag{13.16}$$

Comparing Equation (13.14) with Equation (13.16) we see that for orthogonal transformations

$$A^{\mathsf{T}} = A^{-1}\,. \tag{13.17}$$

Since Equation (13.16) may equally be written $A^{-1}A = I$, from Equation (13.17) we also have $A^{\mathsf{T}}A = I$; in index notation the orthogonality condition can therefore also be written

$$\delta_{ij} = a_{li}a_{lj}\,. \tag{13.18}$$

Inverse orthogonal transformation

From Equations (13.8) and (13.18),

$$\begin{aligned} x'_i &= a_{ik}x_k\,, \qquad a_{ij}(\ ) \\ a_{ij}x'_i &= \delta_{jk}x_k = x_j\,. \end{aligned}$$

Thus the *inverse orthogonal coordinate transformation* is given by

$$x_j = a_{ij}x'_i\,. \tag{13.19}$$

Finally, differentiating Equation (13.8) with respect to x_j and Equation (13.19) with respect to x'_i, we see that

$$a_{ij} = \frac{\partial x'_i}{\partial x_j} = \frac{\partial x_j}{\partial x'_i}\,. \tag{13.20}$$

Cartesian tensors

Definition 13.5 (Tensor under orthogonal transformations)**.** We say that the n^r quantities $T_{i_1\cdots i_r}$ are the components of a *tensor of rank r* if and only if under an orthogonal coordinate transformation their components satisfy

$$T'_{i_1\cdots i_r} = a_{i_1j_1}\cdots a_{i_rj_r}T_{j_1\cdots j_r}\,. \tag{13.21}$$

Rests on Definition 13.4.

In particular, for $r = 2$,

$$T'_{ij} = a_{il}a_{jk}T_{lk}\,. \tag{13.22}$$

To find the inverse relation, note from Equations (13.18) and (13.21) that

$$\begin{aligned} T'_{i_1\cdots i_r} &= a_{i_1k_1}\cdots a_{i_rk_r}T_{k_1\cdots k_r}\,, \qquad a_{i_1j_1}\cdots a_{i_rj_r}(\ ) \\ a_{i_1j_1}\cdots a_{i_rj_r}T'_{i_1\cdots i_r} &= \delta_{j_1k_1}\cdots\delta_{j_rk_r}T_{k_1\cdots k_r}\,, \end{aligned}$$

so that

$$T_{j_1\cdots j_r} = a_{i_1j_1}\cdots a_{i_rj_r}T'_{i_1\cdots i_r}\,. \tag{13.23}$$

Notable Cartesian tensors

We call *vector* a tensor of rank 1.

From the orthogonality condition Equation (13.13) we may write

$$\delta_{ij} = a_{il}a_{jk}\delta_{lk}\,.$$

Moreover, by its definition the Kronecker delta takes the same values in every orthogonal coordinate system, so we may write

$$\delta'_{ij} = a_{il}a_{jk}\delta_{lk}\,. \tag{13.24}$$

Comparing with Equation (13.22), the Kronecker delta satisfies the transformation law of a rank-2 tensor. Therefore *the Kronecker delta is an invariant tensor under orthogonal transformations.*

13.2.2 Algebra of Cartesian tensors

Addition

Consider the quantities $A_{i_1\cdots i_r}$ and $B_{i_1\cdots i_r}$, components of tensors under orthogonal transformations, both of rank r. Define the quantities

$$C_{i_1\cdots i_r} = A_{i_1\cdots i_r} + B_{i_1\cdots i_r} .$$

Then, by Equation (13.21),

$$\begin{aligned} C'_{i_1\cdots i_r} &= A'_{i_1\cdots i_r} + B'_{i_1\cdots i_r} \\ &= a_{i_1 j_1}\cdots a_{i_r j_r} A_{j_1\cdots j_r} + a_{i_1 j_1}\cdots a_{i_r j_r} B_{j_1\cdots j_r} \\ &= a_{i_1 j_1}\cdots a_{i_r j_r} \left(A_{j_1\cdots j_r} + B_{j_1\cdots j_r}\right) \\ &= a_{i_1 j_1}\cdots a_{i_r j_r} C_{j_1\cdots j_r} . \end{aligned}$$

Therefore the addition of tensors forms the components of a new tensor under orthogonal transformations.

Multiplication by a scalar

Consider the quantities $A_{i_1\cdots i_r}$, components of a rank-r tensor under orthogonal transformations, and a scalar λ. Define

$$C_{i_1\cdots i_r} = \lambda A_{i_1\cdots i_r} .$$

Then

$$\begin{aligned} C'_{i_1\cdots i_r} &= \lambda A'_{i_1\cdots i_r} \\ &= \lambda\, a_{i_1 j_1}\cdots a_{i_r j_r} A_{j_1\cdots j_r} \\ &= a_{i_1 j_1}\cdots a_{i_r j_r} \lambda A_{j_1\cdots j_r} \\ &= a_{i_1 j_1}\cdots a_{i_r j_r} C_{j_1\cdots j_r} . \end{aligned}$$

Therefore the multiplication of a tensor by a scalar forms the components of a new tensor under orthogonal transformations.

Product

Consider the quantities $A_{i_1\cdots i_r}$ and $B_{i_1\cdots i_s}$, components of tensors under orthogonal transformations, of ranks r and s respectively. Define

$$C_{i_1\cdots i_{r+s}} = A_{i_1\cdots i_r} B_{i_{r+1}\cdots i_{r+s}} . \tag{13.25}$$

Then, by Equation (13.21),

$$\begin{aligned} C'_{i_1\cdots i_{r+s}} &= A'_{i_1\cdots i_r} B'_{i_{r+1}\cdots i_{r+s}} \\ &= \left(a_{i_1 j_1}\cdots a_{i_r j_r} A_{j_1\cdots j_r}\right)\left(a_{i_{r+1} j_{r+1}}\cdots a_{i_{r+s} j_{r+s}} B_{j_{r+1}\cdots j_{r+s}}\right) \\ &= a_{i_1 j_1}\cdots a_{i_{r+s} j_{r+s}} A_{j_1\cdots j_r} B_{j_{r+1}\cdots j_{r+s}} \\ &= a_{i_1 j_1}\cdots a_{i_{r+s} j_{r+s}} C_{j_1\cdots j_{r+s}} . \end{aligned}$$

Therefore the product of tensors forms the components of a new tensor under orthogonal transformations.

Permutation of indices

Consider the quantities $A_{i_1\cdots i_r}$, components of a rank-r tensor under orthogonal transformations. Define the symbol

$$C_{i_1\cdots i_{s-1}i_s i_{s+1}\cdots i_{t-1}i_t i_{t+1}\cdots i_r} = A_{i_1\cdots i_{s-1}i_t i_{s+1}\cdots i_{t-1}i_s i_{t+1}\cdots i_r}\,, \tag{13.26}$$

in which the s-th and t-th indices of $A_{i_1\cdots i_r}$ are permuted. Then, by Equation (13.21) and reordering the constant factors a_{ij},

$$\begin{aligned} C'_{i_1\cdots i_s\cdots i_t\cdots i_r} &= A'_{i_1\cdots i_t\cdots i_s\cdots i_r} \\ &= a_{i_1 j_1}\cdots a_{i_t j_t}\cdots a_{i_s j_s}\cdots a_{i_r j_r} A_{j_1\cdots j_t\cdots j_s\cdots j_r} \\ &= a_{i_1 j_1}\cdots a_{i_s j_s}\cdots a_{i_t j_t}\cdots a_{i_r j_r} C_{j_1\cdots j_s\cdots j_t\cdots j_r}\,. \end{aligned}$$

Thus the permutation of indices of a tensor defines a new tensor under orthogonal transformations.

Contraction of indices

Consider the quantities $A_{i_1\cdots i_r}$, components of a rank-r tensor under orthogonal transformations. Define the n^{r-2} quantities

$$C_{i_1\cdots i_{r-2}} = A_{i_1\cdots i\cdots i\cdots i_{r-2}}\,. \tag{13.27}$$

Then, by Equations (13.18) and (13.21),

$$\begin{aligned} C'_{i_1\cdots i_{r-2}} &= A'_{i_1\cdots i\cdots i\cdots i_{r-2}} \\ &= a_{i_1 j_1}\cdots a_{ij}\cdots a_{il}\cdots a_{i_{r-2}j_{r-2}} A_{j_1\cdots j\cdots l\cdots j_{r-2}} \\ &= a_{i_1 j_1}\cdots a_{i_{r-2}j_{r-2}}\,\delta_{jl} A_{j_1\cdots j\cdots l\cdots j_{r-2}} \\ &= a_{i_1 j_1}\cdots a_{i_{r-2}j_{r-2}} A_{j_1\cdots j\cdots j\cdots j_{r-2}} \\ &= a_{i_1 j_1}\cdots a_{i_{r-2}j_{r-2}} C_{j_1\cdots j_{r-2}}\,. \end{aligned}$$

Thus the contraction of two indices of a tensor defines a new tensor, of rank $r-2$, under orthogonal transformations.

Quotient law

Consider the quantities $A_{i_1\cdots i_s j_1\cdots j_r}$, and consider further the quantities $B_{j_1\cdots j_r}$, components of a rank-r tensor under orthogonal transformations. Let $C_{i_1\cdots i_s}$ be given by

$$A_{i_1\cdots i_s j_1\cdots j_r} B_{j_1\cdots j_r} = C_{i_1\cdots i_s}\,. \tag{13.28}$$

With respect to another orthogonal coordinate system,

$$A'_{i_1\cdots i_s j_1\cdots j_r} B'_{j_1\cdots j_r} = C'_{i_1\cdots i_s}\,.$$

Supposing that the $C_{i_1\cdots i_s}$ are the components of a rank-s tensor under orthogonal transformations, from Equation (13.21) we have

$$\begin{aligned} A'_{i_1\cdots i_s j_1\cdots j_r}\, a_{j_1 k_1}\cdots a_{j_r k_r} B_{k_1\cdots k_r} &= a_{i_1 l_1}\cdots a_{i_s l_s} C_{l_1\cdots l_s} \\ \text{[Equation (13.28)]} \quad &= a_{i_1 l_1}\cdots a_{i_s l_s} A_{l_1\cdots l_s m_1\cdots m_r} B_{m_1\cdots m_r} \\ &= a_{i_1 l_1}\cdots a_{i_s l_s} A_{l_1\cdots l_s k_1\cdots k_r} B_{k_1\cdots k_r}\,. \end{aligned}$$

Since the $B_{k_1\cdots k_r}$ are arbitrary, applying $a_{m_1 k_1}\cdots a_{m_r k_r}$ to both sides of the resulting relation and using Equation (13.13),

$$\begin{aligned} \delta_{j_1 m_1}\cdots\delta_{j_r m_r} A'_{i_1\cdots i_s j_1\cdots j_r} &= a_{i_1 l_1}\cdots a_{i_s l_s} a_{m_1 k_1}\cdots a_{m_r k_r} A_{l_1\cdots l_s k_1\cdots k_r} \\ A'_{i_1\cdots i_s m_1\cdots m_r} &= a_{i_1 l_1}\cdots a_{i_s l_s} a_{m_1 k_1}\cdots a_{m_r k_r} A_{l_1\cdots l_s k_1\cdots k_r}\,. \end{aligned}$$

Hence the quantities $A_{i_1\cdots i_s j_1\cdots j_r}$ satisfying Equation (13.28) are the components of a tensor under orthogonal transformations. This result is called the *quotient law*.

The Levi-Civita symbol

Beside the Kronecker delta, three dimensions carry a second array with the same components in every frame. Define ε_{ijk}, with $i,j,k \in \{1,2,3\}$, to be totally antisymmetric with $\varepsilon_{123} = +1$; equivalently $\varepsilon_{ijk} = \operatorname{sgn}\sigma$ when $(i,j,k) = (\sigma(1),\sigma(2),\sigma(3))$ for a permutation σ, and zero when two indices coincide. (The general n-index symbol reappears in Section 13.5.3 as a density of weight -1 under arbitrary coordinate changes; under *orthogonal* changes the weight factor is $\det a = \pm 1$, which is the content of Equation (13.31) below.)

Proposition 13.6 (Levi-Civita identities in three dimensions)**.** *For $n = 3$, with ε_{ijk} as just defined,*

$$\varepsilon_{ijk}\varepsilon_{ilm} = \delta_{jl}\delta_{km} - \delta_{jm}\delta_{kl}\,, \tag{13.29}$$

$$\varepsilon_{ijk}\varepsilon_{ijm} = 2\,\delta_{km}\,, \qquad \varepsilon_{ijk}\varepsilon_{ijk} = 6\,, \tag{13.30}$$

and, for any 3×3 array a_{ij},

$$\varepsilon_{ijk}\,a_{il}a_{jm}a_{kn} = \varepsilon_{lmn}\,\det a\,. \tag{13.31}$$

In particular ε_{ijk} is an isotropic tensor under proper orthogonal transformations, and changes sign under improper ones.

Rests on Definition 13.5 and Equation (5.19).

Derivation. *Derives Proposition 13.6.*

Equation (13.29). Both sides vanish when $j = k$ or when $l = m$: on the left ε_{ijk} or ε_{ilm} is then zero, and on the right, putting $j = k$, the two products coincide and cancel (similarly for $l = m$). So let $j \neq k$ and $l \neq m$. A nonzero ε_{ijk} forces $\{i,j,k\} = \{1,2,3\}$, hence fixes i; the same for ε_{ilm}. The sum on the left is therefore nonzero only if the two constraints select the same i, i.e. only if $\{l,m\} = \{j,k\}$, and then exactly one term survives. If $(l,m) = (j,k)$ that term is $\varepsilon_{ijk}^2 = 1$, while the right side is $1 - \delta_{jk}\delta_{kj} = 1$; if $(l,m) = (k,j)$ it is $\varepsilon_{ijk}\varepsilon_{ikj} = -1$, and the right side is $\delta_{jk}\delta_{kj} - 1 = -1$. In every remaining case $\{l,m\} \neq \{j,k\}$ and both sides are zero.

Equation (13.30). Contract Equation (13.29) with δ_{jl}: the right side gives $\delta_{jj}\delta_{km} - \delta_{jm}\delta_{kj} = 3\delta_{km} - \delta_{km} = 2\delta_{km}$. Contracting once more with δ_{km} gives $2\delta_{kk} = 6$.

Equation (13.31). The left side is totally antisymmetric in (l,m,n): exchanging two of them exchanges two of the factors a, which may be undone by exchanging the corresponding two summation indices of ε_{ijk}, at the cost of one sign. A totally antisymmetric array of three three-valued indices is a multiple of ε_{lmn}, the multiple being its value at $(l,m,n) = (1,2,3)$. That value is $\varepsilon_{ijk}a_{i1}a_{j2}a_{k3}$, which is the Leibniz expansion Equation (5.19) of $\det a$ read by columns: the nonzero terms are indexed by the permutations σ with $(i,j,k) = (\sigma(1),\sigma(2),\sigma(3))$, and $\varepsilon_{\sigma(1)\sigma(2)\sigma(3)} = \operatorname{sgn}\sigma$.

The last sentence. Equation (13.31) holds for every 3×3 array, so it may be applied to the transpose of a, giving $a_{li}a_{mj}a_{nk}\varepsilon_{ijk} = \varepsilon_{lmn}\det a^{\mathsf{T}} = \varepsilon_{lmn}\det a$ by Equation (13.14). The left side is the rank-3 transformation law Equation (13.21) applied to ε; the right side is ε itself when $\det a = +1$ and its negative when $\det a = -1$, the two cases of Equation (13.15). □

Isotropic tensors

The algebra above says how tensors combine. The question left open is which tensors are singled out by the group itself — those whose components are the same in every frame. That question is invariant theory for the rotation group rather than tensor algebra, and it is settled once and for all in Theorem 5.128. In *three* dimensions — and the dimension is essential, as Remark 13.7 records — a Cartesian tensor (Definition 13.5) whose components

are unchanged by every proper rotation vanishes at rank one, is a multiple of δ_{ij} at rank two, a multiple of ε_{ijk} at rank three, and a combination

$$T_{ijkl} = a\,\delta_{ij}\delta_{kl} + b\,\delta_{ik}\delta_{jl} + c\,\delta_{il}\delta_{jk} \tag{13.32}$$

of the three products of Kronecker deltas at rank four — three free numbers in place of $3^4 = 81$. The invariance of δ_{ij} and of ε_{ijk} under proper rotations, which is what makes the converse half true, is Equations (13.24) and (13.31) of this section.

Remark 13.7 (What the classification is used for). A linear constitutive law between two rank-2 Cartesian tensors, $\sigma_{ij} = A_{ijkl}\,e_{kl}$, has 81 coefficients in three dimensions. If the medium is isotropic — no direction distinguished — then A_{ijkl} is an isotropic tensor, and Equation (13.32) reduces the 81 to three, or to two once e_{kl} is symmetric, since then only the combination $b+c$ acts. That single step is what makes linear elasticity a two-constant theory (Lemma 30.26) and the Newtonian stress of a fluid a two-viscosity one. The classification is a statement about SO(3) specifically and does not survive a change of dimension: it is not only the rank-3 case that is three-dimensional. In two dimensions ε_{ij} is SO(2)-invariant, so an isotropic rank-2 tensor is $a\,\delta_{ij} + b\,\varepsilon_{ij}$ rather than a multiple of δ_{ij}; in four, ε_{ijkl} is SO(4)-invariant and is not a combination of Kronecker products, so a rank-4 isotropic tensor carries a fourth independent coefficient. Every step of the proof in Chapter 5 uses $n = 3$ — the parity argument needs $\det \operatorname{diag}(1,-1,-1) = +1$, and the relabelling argument needs a 3-cycle to be even.

Rests on Theorem 5.128 and Definition 13.5.

13.3 The theory of curves and surfaces in $\mathbb{R}^3$

13.3.1 Theory of curves

Curves in $\mathbb{R}^3$

Definition 13.8 (Curve)**.** A *curve* is a differentiable map

$$\begin{aligned} \boldsymbol{x} : I_\lambda \subset \mathbb{R} &\longrightarrow E \subset \mathbb{R}^3 \\ \lambda &\longmapsto \boldsymbol{x}(\lambda)\,, \end{aligned}$$

with I_λ an open interval (so as to guarantee differentiability). The variable λ parametrises, or labels, the curve.

Definition 13.9 (Tangent vector to a curve)**.** The *tangent vector to the curve*, with parameter t, is defined as

$$\boldsymbol{T} = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\,. \tag{13.33}$$

Rests on Definition 13.8.

Definition 13.10 (Admissible change of parameter)**.** An *admissible change of parameter* is a function α such that

$$t = \alpha(\theta)\,, \quad \theta \in I_\theta\,, \quad \text{with} \quad \alpha \in C^1(I_\theta) \ \wedge\ \alpha'(\theta) \neq 0 \quad \forall\,\theta \in I_\theta\,. \tag{13.34}$$

Definition 13.11 (Arc length)**.** The *arc length of the curve from t_0 to t* is given by

$$s(t) = \int_{t_0}^{t} \left|\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\right| \mathrm{d}t\,. \tag{13.35}$$

Rests on Definitions 13.8 and 13.9.

From this we see that the inverse function of s is an admissible change of parameter for t. Moreover, differentiating both sides of Equation (13.35) with respect to t,

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \left|\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\right| , \tag{13.36}$$

and therefore, from Equation (13.33),

$$\begin{aligned} \boldsymbol{T} &= \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} \\ &= \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}t} \\ &= \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s}\left|\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\right| \\ \implies \quad \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s} &= \frac{\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}}{\left|\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\right|} . \end{aligned} \tag{13.37}$$

Definition 13.12 (Unit tangent vector)**.** We define the *unit tangent vector* as

$$\hat{T} = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s} . \tag{13.38}$$

Rests on Definitions 13.9 and 13.11.

From Equation (13.37) we see that $\hat{T}$ is indeed a unit vector.

Definition 13.13 (Normal curvature vector of a curve)**.** We define the *normal curvature vector of the curve* as

$$\boldsymbol{k} = \frac{\mathrm{d}^2\boldsymbol{x}}{\mathrm{d}s^2} = \frac{\mathrm{d}\hat{T}}{\mathrm{d}s} . \tag{13.39}$$

Rests on Definition 13.12.

The derivative of any unit vector $\hat{e}$ lies in the plane orthogonal to $\hat{e}$, for

$$\begin{aligned} & |\hat{e}| = 1 \\ \iff \quad & \hat{e}\cdot\hat{e} = 1 , \qquad \frac{\mathrm{d}}{\mathrm{d}s}(\,) \\ \implies \quad & \frac{\mathrm{d}\hat{e}}{\mathrm{d}s}\cdot\hat{e} = 0 \\ \implies \quad & \frac{\mathrm{d}\hat{e}}{\mathrm{d}s} \perp \hat{e} , \end{aligned} \tag{13.40}$$

and therefore the vector $\boldsymbol{k}$ lies in the plane orthogonal to $\hat{T}$.

We define the *curvature function* κ as

$$\kappa(s) = |\boldsymbol{k}| . \tag{13.41}$$

Definition 13.14 (Principal unit normal vector)**.** We define the *principal unit normal vector* as

$$\hat{N} = \frac{\boldsymbol{k}}{\kappa} . \tag{13.42}$$

Rests on Definition 13.13.

We thus find a relation between the derivative of the unit tangent vector and the unit normal,

$$\frac{\mathrm{d}\hat{T}}{\mathrm{d}s} = \kappa\hat{N} . \tag{13.43}$$

Definition 13.15 (Unit binormal vector)**.** We define the *unit binormal vector* as

$$\hat{B} = \hat{T} \times \hat{N} \,. \tag{13.44}$$

Rests on Definitions 13.12 and 13.14.

We thereby obtain a system of three orthonormal vectors called the *moving trihedron* (the Frenet frame).

Torsion

Differentiating Equation (13.44) with respect to s,

$$\frac{\mathrm{d}\hat{B}}{\mathrm{d}s} = \frac{\mathrm{d}\hat{T}}{\mathrm{d}s} \times \hat{N} + \hat{T} \times \frac{\mathrm{d}\hat{N}}{\mathrm{d}s} \,. \tag{13.45}$$

From Equation (13.43) the first term of this expression vanishes. From Equation (13.40) the vector $\frac{\mathrm{d}\hat{N}}{\mathrm{d}s}$ lies in the plane spanned by $\hat{T}$ and $\hat{B}$, so we may write $\frac{\mathrm{d}\hat{N}}{\mathrm{d}s}$ as a linear combination of $\hat{T}$ and $\hat{B}$. Thus

$$\begin{aligned} \frac{\mathrm{d}\hat{B}}{\mathrm{d}s} &= \hat{T} \times \left(\alpha(s)\hat{T} + \tau(s)\hat{B}\right) \\ &= \tau\, \hat{T} \times \hat{B} \\ \\ \implies \qquad \frac{\mathrm{d}\hat{B}}{\mathrm{d}s} &= -\tau \hat{N} \,, \end{aligned} \tag{13.46}$$

where the function τ is called the *torsion of the curve.*

From Equation (13.44) we see that

$$\begin{aligned} \hat{N} &= \hat{B} \times \hat{T} \,, \qquad \frac{\mathrm{d}}{\mathrm{d}s}(\,) \\ \frac{\mathrm{d}\hat{N}}{\mathrm{d}s} &= \frac{\mathrm{d}\hat{B}}{\mathrm{d}s} \times \hat{T} + \hat{B} \times \frac{\mathrm{d}\hat{T}}{\mathrm{d}s} \,, \end{aligned}$$

and by Equations (13.43) and (13.46),

$$\begin{aligned} \frac{\mathrm{d}\hat{N}}{\mathrm{d}s} &= -\tau \hat{N} \times \hat{T} + \kappa \hat{B} \times \hat{N} \\ \\ \implies \qquad \frac{\mathrm{d}\hat{N}}{\mathrm{d}s} &= \tau \hat{B} - \kappa \hat{T} \,. \end{aligned} \tag{13.47}$$

The Frenet–Serret equations

Equations Equations (13.43), (13.46) and (13.47) are the *Frenet–Serret equations*, which relate the vectors of the moving trihedron to their derivatives:

$$\begin{aligned} \frac{\mathrm{d}\hat{T}}{\mathrm{d}s} &= \kappa \hat{N} \\ \frac{\mathrm{d}\hat{N}}{\mathrm{d}s} &= \tau \hat{B} - \kappa \hat{T} \\ \frac{\mathrm{d}\hat{B}}{\mathrm{d}s} &= -\tau \hat{N} \,, \end{aligned}$$

or, in matrix form,

$$\frac{\mathrm{d}}{\mathrm{d}s} \begin{pmatrix} \hat{T} \\ \hat{N} \\ \hat{B} \end{pmatrix} = \begin{pmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{pmatrix} \begin{pmatrix} \hat{T} \\ \hat{N} \\ \hat{B} \end{pmatrix} \,. \tag{13.48}$$

The fundamental theorem of the theory of curves

The Frenet–Serret system determines the moving trihedron from the two functions κ and τ, and the trihedron determines the curve. The precise statement is the following.

Theorem 13.16 (Fundamental theorem of the theory of curves)**.** *Let $I =]a, b[\subset \mathbb{R}$ and let $\kappa, \tau : I \longrightarrow \mathbb{R}$ be continuous with $\kappa(s) > 0$ for every $s \in I$. Then there exists a curve $\boldsymbol{x} : I \longrightarrow \mathbb{R}^3$, parametrised by arc length, whose curvature function is κ and whose torsion is τ. It is unique up to a rigid motion of $\mathbb{R}^3$: any two such curves are related by $\boldsymbol{x} \longmapsto A\boldsymbol{x} + \boldsymbol{c}$ with $A \in \mathrm{SO}(3)$ and $\boldsymbol{c} \in \mathbb{R}^3$ constant.*

Rests on Equations (13.43)*,* (13.46) *and* (13.47)*.*

Proof. *Derives Theorem 13.16.*

Existence. Read the Frenet–Serret equations as a linear system of ordinary differential equations for the nine components of the triple $(\hat{T}, \hat{N}, \hat{B})$,

$$\frac{\mathrm{d}}{\mathrm{d}s}\begin{pmatrix}\hat{T}\\ \hat{N}\\ \hat{B}\end{pmatrix} = \Omega(s)\begin{pmatrix}\hat{T}\\ \hat{N}\\ \hat{B}\end{pmatrix}, \qquad \Omega = \begin{pmatrix}0 & \kappa & 0\\ -\kappa & 0 & \tau\\ 0 & -\tau & 0\end{pmatrix}. \tag{13.49}$$

The coefficients are continuous on I, so by the existence and uniqueness theorem for linear systems (Chapter 9) the system has exactly one solution on the whole of I for each choice of initial data at some $s_0 \in I$. Choose as initial data any positively oriented orthonormal triple $(\hat{T}_0, \hat{N}_0, \hat{B}_0)$.

The solution remains orthonormal. Write $e_1 = \hat{T}$, $e_2 = \hat{N}$, $e_3 = \hat{B}$ and consider the Gram matrix $G_{AB} = e_A \cdot e_B$. From Equation (13.49),

$$\frac{\mathrm{d}G_{AB}}{\mathrm{d}s} = \Omega_{AC}G_{CB} + \Omega_{BC}G_{AC}\,, \tag{13.50}$$

a linear system for G. Because Ω is antisymmetric, the constant matrix $G_{AB} = \delta_{AB}$ solves Equation (13.50); it agrees with the initial data at s_0, and by uniqueness of solutions it is *the* solution. Hence the triple is orthonormal for every $s \in I$, and $\det(e_1, e_2, e_3)$, a continuous function taking values in $\{+1, -1\}$ on the interval I, stays equal to its initial value $+1$: the triple stays positively oriented.

Now define

$$\boldsymbol{x}(s) = \boldsymbol{x}_0 + \int_{s_0}^{s} \hat{T}(\sigma)\,\mathrm{d}\sigma\,. \tag{13.51}$$

Then $\mathrm{d}\boldsymbol{x}/\mathrm{d}s = \hat{T}$ is a unit vector, so s is the arc-length parameter of $\boldsymbol{x}$ by Equation (13.36); the first row of Equation (13.49) gives $\mathrm{d}\hat{T}/\mathrm{d}s = \kappa\hat{N}$ with $\hat{N}$ a unit vector and $\kappa > 0$, so the curvature function of $\boldsymbol{x}$ is κ and its principal unit normal is $\hat{N}$ (Equations (13.39) and (13.42)); orthonormality and orientation give $\hat{B} = \hat{T} \times \hat{N}$, so $\hat{B}$ is the binormal Equation (13.44); and the third row gives $\mathrm{d}\hat{B}/\mathrm{d}s = -\tau\hat{N}$, so the torsion of $\boldsymbol{x}$ is τ by Equation (13.46).

Uniqueness. A rigid motion $\boldsymbol{x} \longmapsto A\boldsymbol{x} + \boldsymbol{c}$ with $A \in \mathrm{SO}(3)$ preserves arc length, scalar products and cross products, hence carries the trihedron of a curve to the trihedron of the image curve and leaves κ and τ unchanged. Conversely, let $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ both realize the given κ and τ. Their trihedra at s_0 are two positively oriented orthonormal bases of $\mathbb{R}^3$, so there is a unique $A \in \mathrm{SO}(3)$ carrying the second to the first; let $\boldsymbol{c} = \boldsymbol{x}(s_0) - A\tilde{\boldsymbol{x}}(s_0)$. The curve $A\tilde{\boldsymbol{x}} + \boldsymbol{c}$ then has the same κ and τ as $\boldsymbol{x}$ and the same trihedron at s_0; since both trihedra solve Equation (13.49) with the same coefficients and the same data at s_0, they coincide on all of I. In particular the two unit tangents coincide, and the two curves, having the same value at s_0, coincide by Equation (13.51). □

Remark 13.17. The hypothesis $\kappa > 0$ cannot be dropped: where κ vanishes the principal normal Equation (13.42) is undefined, the trihedron is not determined by the curve, and a straight segment may be joined smoothly to curved pieces in inequivalent ways. The theorem is the geometric statement that κ and τ are a *complete* set of invariants of a curve in $\mathbb{R}^3$ under the group of rigid motions — the pattern that recurs throughout this chapter, where a geometric object is classified by the scalars that survive a group of transformations.

13.3.2 Theory of surfaces

Surfaces in $\mathbb{R}^3$

Definition 13.18 (Surface)**.** A *surface* is a differentiable map

$$\begin{aligned} \boldsymbol{x} : I_{u_1} \times I_{u_2} = A \subset \mathbb{R}^2 &\longrightarrow E \subset \mathbb{R}^3 \\ (u_1, u_2) &\longmapsto \boldsymbol{x}(u_1, u_2)\,, \end{aligned}$$

with I_{u_1}, I_{u_2}, and A open sets.

Definition 13.19 (Parameter curves)**.** We call *curve of parameter* u_1 and *curve of parameter* u_2 the curves

$$\boldsymbol{x} = \boldsymbol{x}(u_1, a)\,, \qquad \boldsymbol{x} = \boldsymbol{x}(b, u_2)\,, \tag{13.52}$$

respectively, with a and b constants.

Rests on Definitions 13.8 and 13.18.

From the definition we see that these curves lie on the surface.

The *tangent vectors* to the curves of parameter u_1 and u_2 are given by

$$\frac{\partial \boldsymbol{x}}{\partial u_1} \overset{\text{not.}}{=} \boldsymbol{x}_1\,, \qquad \frac{\partial \boldsymbol{x}}{\partial u_2} \overset{\text{not.}}{=} \boldsymbol{x}_2\,, \tag{13.53}$$

respectively. These vectors lie in the tangent plane of the surface.

Definition 13.20 (Unit normal vector of a surface)**.** We define the *unit normal vector of the surface* as

$$\hat{n} = \frac{\boldsymbol{x}_1 \times \boldsymbol{x}_2}{|\boldsymbol{x}_1 \times \boldsymbol{x}_2|}\,. \tag{13.54}$$

Rests on Definition 13.19.

First fundamental form

Consider the differential

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{x}_i \mathrm{d}u_i = \boldsymbol{x}_1 \mathrm{d}u_1 + \boldsymbol{x}_2 \mathrm{d}u_2\,. \tag{13.55}$$

This differential is written as a linear combination of the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and therefore also lies in the tangent plane of the surface. Consider a curve on the surface, given by

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{x}\left(u_1(s), u_2(s)\right) \\ &= \boldsymbol{x}(s)\,, \end{aligned} \tag{13.56}$$

where s is the arc-length parameter of the curve. From Equation (13.36) we have

$$\mathrm{d}s = |\mathrm{d}\boldsymbol{x}|\,, \tag{13.57}$$

where $\mathrm{d}s$ is called the *arc element*. Thus

$$\mathrm{d}s^2 = \mathrm{d}\boldsymbol{x} \cdot \mathrm{d}\boldsymbol{x}\,. \tag{13.58}$$

Definition 13.21 (First fundamental form)**.** We call *first fundamental form* the square of the arc element of the curve given by Equation (13.56), that is, the quantity $\mathrm{d}s^2$.

Rests on Definitions 13.11 and 13.18.

From Equations (13.55) and (13.58) we see that

$$\mathrm{d}s^2 = \boldsymbol{x}_i \cdot \boldsymbol{x}_j \, \mathrm{d}u_i \mathrm{d}u_j \,. \tag{13.59}$$

Definition 13.22 (First fundamental coefficients)**.** We define the *first fundamental coefficients* as

$$g_{ij} = \boldsymbol{x}_i \cdot \boldsymbol{x}_j \,. \tag{13.60}$$

Rests on Definitions 13.19 and 13.21.

The first fundamental form is therefore given by

$$\mathrm{d}s^2 = g_{ij} \, \mathrm{d}u_i \mathrm{d}u_j \,. \tag{13.61}$$

From the definition we see that the g_{ij} are the components of a symmetric tensor of rank 2, which is called the *metric tensor*, or simply the *metric*. The *determinant of the metric tensor* is given by

$$g = g_{11} g_{22} - g_{12}^2 \,. \tag{13.62}$$

Definition 13.23 (Inverse metric tensor)**.** We define the *inverse metric tensor* g^{ij} as the quantities that satisfy

$$g_{li} \, g^{ij} = \delta_l{}^j \,. \tag{13.63}$$

Rests on Definition 13.22.

We can compute these quantities using matrix notation: writing

$$[g_{li}][g^{ij}] = I \tag{13.64}$$

with

$$[g_{li}] = \begin{pmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{pmatrix} , \qquad [g^{ij}] = \begin{pmatrix} g^{11} & g^{12} \\ g^{12} & g^{22} \end{pmatrix} , \tag{13.65}$$

we see from Equation (13.64) that $[g^{ij}] = [g_{li}]^{-1}$, and therefore, from the algebra of matrices,

$$\begin{pmatrix} g^{11} & g^{12} \\ g^{12} & g^{22} \end{pmatrix} = \begin{pmatrix} \dfrac{g_{22}}{g} & -\dfrac{g_{12}}{g} \\ -\dfrac{g_{12}}{g} & \dfrac{g_{11}}{g} \end{pmatrix} . \tag{13.66}$$

Length of a curve and area of a surface

Consider the curve on a surface $\boldsymbol{x}$ given by Equation (13.56). Since the surface carries a first fundamental form, we can compute the *length of the whole curve* in terms of the first fundamental coefficients. As $s = \int \mathrm{d}s$, the length of the whole curve is given by

$$L = \int_{I_s} \sqrt{g_{ij} \, \mathrm{d}u_i \mathrm{d}u_j} \,. \tag{13.67}$$

Now define an angle γ between the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Since

$$\boldsymbol{x}_1 \cdot \boldsymbol{x}_2 = |\boldsymbol{x}_1| \, |\boldsymbol{x}_2| \cos\gamma \,, \tag{13.68}$$

from Equation (13.60) we have

$$\cos\gamma = \frac{g_{12}}{\sqrt{g_{11} g_{22}}} \,. \tag{13.69}$$

On the other hand, the area of an infinitesimal parallelogram of the surface is given by

$$\begin{aligned} \mathrm{d}S &= |\boldsymbol{x}_1 \times \boldsymbol{x}_2| \, \mathrm{d}u_1 \mathrm{d}u_2 \\ &= |\boldsymbol{x}_1| \, |\boldsymbol{x}_2| \sin\gamma \, \mathrm{d}u_1 \mathrm{d}u_2 \\ &= \sqrt{g_{11} g_{22}} \, \sin\gamma \, \mathrm{d}u_1 \mathrm{d}u_2 \,, \end{aligned}$$

and from Equation (13.69),

$$\sin\gamma = \sqrt{1 - \frac{g_{12}^2}{g_{11} g_{22}}} \,. \tag{13.70}$$

Thus we see that

$$\mathrm{d}S = \sqrt{g} \, \mathrm{d}u_1 \mathrm{d}u_2 \,. \tag{13.71}$$

Integrating, the *area of a surface* is

$$S = \int_{I_{u_1}} \int_{I_{u_2}} \sqrt{g} \, \mathrm{d}u_1 \mathrm{d}u_2 \,. \tag{13.72}$$

Second fundamental form

The differential of any unit vector $\hat{e}$ lies in the plane orthogonal to $\hat{e}$, for

$$\begin{aligned} & |\hat{e}| = 1 \\ \iff \quad & \hat{e} \cdot \hat{e} = 1 \,, \qquad \mathrm{d}(\,) \\ \implies \quad & \mathrm{d}\hat{e} \cdot \hat{e} = 0 \\ \implies \quad & \mathrm{d}\hat{e} \perp \hat{e} \,, \end{aligned} \tag{13.73}$$

so the vector $\mathrm{d}\hat{n}$ lies in the plane orthogonal to $\hat{n}$. By the chain rule, the differential $\mathrm{d}\hat{n}$ is given by

$$\mathrm{d}\hat{n} = \frac{\partial \hat{n}}{\partial u_i} \mathrm{d}u_i \overset{\text{not.}}{=} \hat{n}_i \, \mathrm{d}u_i \,. \tag{13.74}$$

Definition 13.24 (Second fundamental form)**.** We define the *second fundamental form* as the quantity

$$\mathrm{I\!I} = -\mathrm{d}\boldsymbol{x} \cdot \mathrm{d}\hat{n} \,. \tag{13.75}$$

Rests on Definition 13.20.

From Equations (13.55) and (13.74) we see that

$$\mathrm{I\!I} = -\boldsymbol{x}_i \cdot \hat{n}_j \, \mathrm{d}u_i \mathrm{d}u_j \,. \tag{13.76}$$

Definition 13.25 (Second fundamental coefficients)**.** We define the *second fundamental coefficients* as

$$b_{ij} = -\boldsymbol{x}_i \cdot \hat{n}_j \,. \tag{13.77}$$

Rests on Definitions 13.20 and 13.24.

Thus the second fundamental form is given by

$$\mathrm{I\!I} = b_{ij} \, \mathrm{d}u_i \mathrm{d}u_j \,. \tag{13.78}$$

From their definition, the second fundamental coefficients are the components of a symmetric tensor of rank 2. The *determinant of the matrix formed by the second fundamental coefficients* is given by

$$b = b_{11} b_{22} - b_{12}^2 \,. \tag{13.79}$$

Alternative expression. Note that the unit normal $\hat{n}$ is orthogonal to each of the partial tangent vectors, $\hat{n} \cdot \boldsymbol{x}_i = 0$. Hence

$$\begin{aligned} \hat{n} \cdot \boldsymbol{x}_i &= 0\,, \qquad \frac{\partial}{\partial u_j}(\,) \\ \hat{n}_j \cdot \boldsymbol{x}_i + \hat{n} \cdot \boldsymbol{x}_{ij} &= 0 \\ \text{[Equation (13.77)]} \quad b_{ij} &= \hat{n} \cdot \frac{\partial^2 \boldsymbol{x}}{\partial u_j \partial u_i}\,, \end{aligned} \tag{13.80}$$

and therefore, from Equation (13.78), the second fundamental form is given by

$$\mathrm{I\!I} = \hat{n} \cdot \mathrm{d}^2 \boldsymbol{x}\,. \tag{13.81}$$

Geometric interpretation. Recall from linear algebra that the projection $\boldsymbol{P}$ of a vector $\boldsymbol{A}$ onto another vector $\boldsymbol{B}$ is given by

$$\boldsymbol{P} = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{|\boldsymbol{B}|^2}\,\boldsymbol{B}\,. \tag{13.82}$$

Consider now the vector that, with respect to a tangent plane of the surface, indicates the small advance of the surface in a neighbourhood of the point of tangency. This vector is $\boldsymbol{x}(u_1 + \mathrm{d}u_1, u_2 + \mathrm{d}u_2) - \boldsymbol{x}(u_1, u_2)$; thus the distance d from the surface to its tangent plane, in a neighbourhood of the point of tangency, equals the projection of this vector onto the unit normal $\hat{n}$. From Equation (13.82),

$$d = [\boldsymbol{x}(u_1 + \mathrm{d}u_1, u_2 + \mathrm{d}u_2) - \boldsymbol{x}(u_1, u_2)] \cdot \hat{n}\,.$$

Making a multivariable Taylor expansion of $\boldsymbol{x}(u_1 + \mathrm{d}u_1, u_2 + \mathrm{d}u_2)$ about (u_1, u_2),

$$\begin{aligned} d &\approx \left[\boldsymbol{x}(u_1, u_2) + \frac{\partial \boldsymbol{x}}{\partial u_i}\mathrm{d}u_i + \frac{1}{2}\frac{\partial^2 \boldsymbol{x}}{\partial u_j \partial u_i}\mathrm{d}u_i \mathrm{d}u_j - \boldsymbol{x}(u_1, u_2)\right] \cdot \hat{n} \\ \text{[Equation (13.54)]} \quad &= \frac{1}{2}\,\mathrm{d}^2 \boldsymbol{x} \cdot \hat{n} \\ &= \frac{1}{2}\,\mathrm{I\!I}\,. \end{aligned}$$

Therefore the second fundamental form is, approximately, *twice the distance from the surface to its tangent plane in a neighbourhood of the point of tangency.*

The Gauss equations and the Christoffel symbols

Since at each point of the surface we have the three linearly independent vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\hat{n}$, we can write any vector of $\mathbb{R}^3$ as a linear combination of them. In particular, we can express the partial derivative of $\boldsymbol{x}_i$ with respect to u_j as

$$\boldsymbol{x}_{ij} = \Gamma_{ij}^{\,k} \boldsymbol{x}_k + \alpha_{ij} \hat{n}\,,$$

where $\Gamma_{ij}^{\,k}$ and α_{ij} are the coefficients of the linear combination. Taking the scalar product with $\hat{n}$ and using Equation (13.80),

$$b_{ij} = \alpha_{ij}\,.$$

Thus we obtain the *Gauss equations*,

$$\boldsymbol{x}_{ij} = \Gamma_{ij}^{\,k} \boldsymbol{x}_k + b_{ij} \hat{n}\,. \tag{13.83}$$

From Equation (13.83), taking the scalar product with $\boldsymbol{x}_l$ and using Equations (13.54) and (13.60),

$$\boldsymbol{x}_{ij} \cdot \boldsymbol{x}_l = \Gamma_{ij}^{\,k} g_{kl}\,.$$

Definition 13.26 (Christoffel symbols)**.** We define the *Christoffel symbol of the first kind* as the quantities

$$\Gamma_{ijk} = \boldsymbol{x}_{ij} \cdot \boldsymbol{x}_k \,, \tag{13.84}$$

and the *Christoffel symbol of the second kind* as the quantities

$$\Gamma_{ij}^{\ k} = \Gamma_{ijl}\, g^{lk} \,. \tag{13.85}$$

Rests on Equation (13.83) *and Definition 13.23.*

Note from their definition that the Christoffel symbols are symmetric in their first two indices: considering that $\boldsymbol{x}$ is a map of class C^2, by Schwarz's theorem

$$\boldsymbol{x}_{ij} = \boldsymbol{x}_{ji} \,. \tag{13.86}$$

Alternative expression for the Christoffel symbols. From the definition of the first fundamental coefficients, Equation (13.60),

$$\begin{aligned} \frac{\partial g_{ij}}{\partial u_k} &\overset{\text{not.}}{=} \partial_k g_{ij} = \boldsymbol{x}_{ik} \cdot \boldsymbol{x}_j + \boldsymbol{x}_{jk} \cdot \boldsymbol{x}_i \\ \text{[Equation (13.84)]} \quad &= \Gamma_{ikj} + \Gamma_{jki} \,. \end{aligned}$$

In the same way,

$$\begin{aligned} \partial_i g_{jk} &= \Gamma_{jik} + \Gamma_{kij} \,, \\ \partial_j g_{ki} &= \Gamma_{kji} + \Gamma_{ijk} \,. \end{aligned}$$

Then, since the Christoffel symbols are symmetric in their first two indices,

$$\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij} = 2\Gamma_{ijk} \,,$$

and therefore

$$\Gamma_{ijk} = \frac{1}{2} \left(\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij} \right) . \tag{13.87}$$

From Equation (13.86), multiplying Equation (13.87) by g^{kl} and renaming indices,

$$\Gamma_{ij}^{\ k} = \frac{1}{2} g^{kl} \left(\partial_i g_{jl} + \partial_j g_{li} - \partial_l g_{ij} \right) . \tag{13.88}$$

Note from Equation (13.88) that setting $j = k$ gives

$$\begin{aligned} \Gamma_{ij}^{\ j} &= \frac{1}{2} g^{jl} \left(\partial_i g_{jl} + \partial_j g_{li} - \partial_l g_{ij} \right) \\ &= \frac{1}{2} g^{jl} \partial_i g_{jl} + \frac{1}{2} \left(g^{jl} \partial_j g_{li} - g^{jl} \partial_l g_{ij} \right) \\ &= \frac{1}{2} g^{jl} \partial_i g_{jl} + \frac{1}{2} \left(g^{lj} \partial_l g_{ji} - g^{jl} \partial_l g_{ij} \right) , \end{aligned}$$

where the last two terms cancel after renaming the summed indices; therefore

$$\Gamma_{ij}^{\ j} = \frac{1}{2} g^{jl} \partial_i g_{jl} \,. \tag{13.89}$$

The Weingarten equations

By the same justification as for the Gauss equations, we can write the partial derivative of $\hat{n}$ with respect to u_i as a linear combination of the vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\hat{n}$. Moreover, in analogy with Equations (13.40) and (13.73), the partial derivative of any unit vector $\hat{e}$ with respect to u_i lies in the plane orthogonal to $\hat{e}$:

$$\begin{aligned} & |\hat{e}| = 1 \\ \iff \quad & \hat{e} \cdot \hat{e} = 1\,, \qquad \partial_i(\,) \\ \implies \quad & \hat{e}_i \cdot \hat{e} = 0 \\ \implies \quad & \hat{e}_i \perp \hat{e}\,. \end{aligned} \tag{13.90}$$

Then $\hat{n}_i$ lies in the plane orthogonal to $\hat{n}$, so we may write

$$\hat{n}_i = \beta_i{}^j \boldsymbol{x}_j\,,$$

where the quantities $\beta_i{}^j$ are the factors of the linear combination. To compute them, take the scalar product with $\boldsymbol{x}_k$ and use Equations (13.60) and (13.77):

$$\begin{aligned} -b_{ik} &= \beta_i{}^j g_{jk} \\ b_{ik} &= -\beta_{ik} \\ b_i{}^k &= -\beta_i{}^k\,. \end{aligned}$$

Finally, we obtain the *Weingarten equations*,

$$\hat{n}_i = -b_i{}^j \boldsymbol{x}_j\,. \tag{13.91}$$

The Gauss–Codazzi equations

Starting from the Gauss equations,

$$\begin{aligned} \boldsymbol{x}_{ij} &= \Gamma_{ij}^{\ l} \boldsymbol{x}_l + b_{ij}\hat{n}\,, \qquad \partial_k(\,) \\ \boldsymbol{x}_{ijk} &= \partial_k\left(\Gamma_{ij}^{\ l}\right)\boldsymbol{x}_l + \Gamma_{ij}^{\ l}\boldsymbol{x}_{lk} + \partial_k\left(b_{ij}\right)\hat{n} + b_{ij}\hat{n}_k \\ \text{[Equations (13.83) and (13.91)]} \quad &= \partial_k\left(\Gamma_{ij}^{\ l}\right)\boldsymbol{x}_l + \Gamma_{ij}^{\ l}\Gamma_{lk}^{\ m}\boldsymbol{x}_m + \Gamma_{ij}^{\ l} b_{lk}\hat{n} + \partial_k\left(b_{ij}\right)\hat{n} - b_{ij}b_k{}^l\boldsymbol{x}_l \\ &= \left[\partial_k\left(\Gamma_{ij}^{\ l}\right) + \Gamma_{ij}^{\ m}\Gamma_{mk}^{\ l} - b_{ij}b_k{}^l\right]\boldsymbol{x}_l + \left[\Gamma_{ij}^{\ l}b_{lk} + \partial_k\left(b_{ij}\right)\right]\hat{n}\,. \end{aligned}$$

Renaming indices, we have

$$\begin{aligned} \boldsymbol{x}_{ijk} &= \left[\partial_k\left(\Gamma_{ij}^{\ l}\right) + \Gamma_{ij}^{\ m}\Gamma_{mk}^{\ l} - b_{ij}b_k{}^l\right]\boldsymbol{x}_l + \left[\Gamma_{ij}^{\ l}b_{lk} + \partial_k\left(b_{ij}\right)\right]\hat{n}\,, \\ \boldsymbol{x}_{ikj} &= \left[\partial_j\left(\Gamma_{ik}^{\ l}\right) + \Gamma_{ik}^{\ m}\Gamma_{mj}^{\ l} - b_{ik}b_j{}^l\right]\boldsymbol{x}_l + \left[\Gamma_{ik}^{\ l}b_{lj} + \partial_j\left(b_{ik}\right)\right]\hat{n}\,. \end{aligned}$$

By the compatibility conditions, $\boldsymbol{x}_{ijk} - \boldsymbol{x}_{ikj} = \boldsymbol{0}$; subtracting the previous equations,

$$\begin{aligned} \Big[\partial_k\left(\Gamma_{ij}^{\ l}\right) + \Gamma_{ij}^{\ m}\Gamma_{mk}^{\ l} - b_{ij}b_k{}^l - \partial_j\left(\Gamma_{ik}^{\ l}\right) - \Gamma_{ik}^{\ m}\Gamma_{mj}^{\ l} + b_{ik}b_j{}^l\Big]\boldsymbol{x}_l \\ + \left[\Gamma_{ij}^{\ l}b_{lk} + \partial_k\left(b_{ij}\right) - \Gamma_{ik}^{\ l}b_{lj} - \partial_j\left(b_{ik}\right)\right]\hat{n} = \boldsymbol{0}\,. \end{aligned}$$

Since $\boldsymbol{x}_l$ and $\hat{n}$ are linearly independent, their factors vanish separately. From this we see that

$$b_{ik}b_j{}^l - b_{ij}b_k{}^l = \partial_j\left(\Gamma_{ik}^{\ l}\right) - \partial_k\left(\Gamma_{ij}^{\ l}\right) + \Gamma_{ik}^{\ m}\Gamma_{mj}^{\ l} - \Gamma_{ij}^{\ m}\Gamma_{mk}^{\ l}\,, \tag{13.92}$$

$$\partial_j\left(b_{ik}\right) - \partial_k\left(b_{ij}\right) = \Gamma_{ij}^{\ l}b_{lk} - \Gamma_{ik}^{\ l}b_{lj}\,, \tag{13.93}$$

which are the *Gauss–Codazzi equations*. These equations relate the metric to the second fundamental coefficients.

The Riemann symbols

Definition 13.27 (Riemann symbols)**.** We define the *Riemann symbol of the first kind* as the quantities

$$R^l{}_{ijk} = b_{ik} b_j{}^l - b_{ij} b_k{}^l \,. \tag{13.94}$$

From Equation (13.92) we see that

$$R^l{}_{ijk} = \partial_j \left(\Gamma_{ik}{}^l\right) - \partial_k \left(\Gamma_{ij}{}^l\right) + \Gamma_{ik}{}^m \Gamma_{mj}{}^l - \Gamma_{ij}{}^m \Gamma_{mk}{}^l \,. \tag{13.95}$$

We define the *Riemann symbol of the second kind* as the quantities

$$R_{lijk} = R^m_{ijk}\, g_{ml} = b_{ik} b_{jl} - b_{ij} b_{kl} \,. \tag{13.96}$$

Rests on Definitions 13.25 and 13.26 and Equation (13.92).

13.3.3 Curvatures of a surface

Normal curvature

Let $\boldsymbol{x} = \boldsymbol{x}(u_1(t), u_2(t))$ be a curve whose points lie on the surface $\boldsymbol{x} = \boldsymbol{x}(u_1, u_2)$, and consider the normal curvature vector $\boldsymbol{k}$ of that curve (see Equation (13.39)).

Definition 13.28 (Normal curvature vector of a surface)**.** We define the *normal curvature vector of the surface* as the projection (see Equation (13.82)) of the curve's normal curvature vector $\boldsymbol{k}$ onto the unit normal $\hat{n}$, that is,

$$\boldsymbol{k}_n = (\boldsymbol{k} \cdot \hat{n})\, \hat{n} \,. \tag{13.97}$$

Rests on Definitions 13.13 and 13.20.

We call *normal curvature* the quantity

$$|\boldsymbol{k}_n| = k_n = \boldsymbol{k} \cdot \hat{n} \,. \tag{13.98}$$

Let us derive an alternative expression for the normal curvature. We have, using Equations (13.37) and (13.39),

$$\begin{aligned} k_n &= \boldsymbol{k} \cdot \hat{n} \\ &= \frac{\frac{\mathrm{d}\hat{T}}{\mathrm{d}t} \cdot \hat{n}}{\left|\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\right|} \,. \end{aligned} \tag{13.99}$$

Since the tangent vector $\hat{T}$ of the curve is orthogonal to the surface normal $\hat{n}$,

$$\hat{T} \cdot \hat{n} = 0 \,, \qquad \frac{\mathrm{d}}{\mathrm{d}t}(\,)$$

$$\frac{\mathrm{d}\hat{T}}{\mathrm{d}t} \cdot \hat{n} + \hat{T} \cdot \frac{\mathrm{d}\hat{n}}{\mathrm{d}t} = 0 \,.$$

Therefore Equation (13.99) becomes, using Equations (13.37) and (13.38),

$$\begin{aligned} k_n &= -\frac{\hat{T} \cdot \frac{\mathrm{d}\hat{n}}{\mathrm{d}t}}{\left|\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\right|} \\ &= -\frac{\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} \cdot \frac{\mathrm{d}\hat{n}}{\mathrm{d}t}}{\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} \cdot \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}} \\ &= -\frac{\mathrm{d}\boldsymbol{x} \cdot \mathrm{d}\hat{n}}{\mathrm{d}\boldsymbol{x} \cdot \mathrm{d}\boldsymbol{x}} \,. \end{aligned}$$

Finally, we arrive at

$$k_n = \frac{b_{ij}\, \mathrm{d}u_i \mathrm{d}u_j}{g_{ij}\, \mathrm{d}u_i \mathrm{d}u_j} \,. \tag{13.100}$$

Principal curvatures and Gaussian curvature

From Equation (13.100) we see that

$$k_n = \frac{b_{11}\mathrm{d}u_1^2 + 2b_{12}\mathrm{d}u_1\mathrm{d}u_2 + b_{22}\mathrm{d}u_2^2}{g_{11}\mathrm{d}u_1^2 + 2g_{12}\mathrm{d}u_1\mathrm{d}u_2 + g_{22}\mathrm{d}u_2^2}\,. \tag{13.101}$$

Introducing the auxiliary variable

$$\lambda = \frac{\mathrm{d}u_2}{\mathrm{d}u_1}\,, \tag{13.102}$$

we have

$$k_n = \frac{b_{11} + 2b_{12}\lambda + b_{22}\lambda^2}{g_{11} + 2g_{12}\lambda + g_{22}\lambda^2}\,. \tag{13.103}$$

We now seek an extremal value of k_n, denoted k_{np}, which we call a *principal curvature*. That is,

$$\begin{aligned} &\frac{\mathrm{d}k_n}{\mathrm{d}\lambda} \stackrel{!}{=} 0 \\ \Longrightarrow \quad &(b_{12} + b_{22}\lambda)\left(g_{11} + 2g_{12}\lambda + g_{22}\lambda^2\right) = (g_{12} + g_{22}\lambda)\left(b_{11} + 2b_{12}\lambda + b_{22}\lambda^2\right)\,. \end{aligned} \tag{13.104}$$

This relation holds for a value of λ that makes k_n extremal. Substituting it back into Equation (13.103),

$$k_{np} = \frac{b_{12} + b_{22}\lambda}{g_{12} + g_{22}\lambda}\,. \tag{13.105}$$

Note that Equation (13.104) can be rewritten in the form

$$\begin{aligned} (b_{12} + b_{22}\lambda)\,(g_{11} + g_{12}\lambda + (g_{12} + g_{22}\lambda)\,\lambda) &= (g_{12} + g_{22}\lambda)\,(b_{11} + b_{12}\lambda + (b_{12} + b_{22}\lambda)\,\lambda) \\ \frac{g_{11} + g_{12}\lambda}{g_{12} + g_{22}\lambda} &= \frac{b_{11} + b_{12}\lambda}{b_{12} + b_{22}\lambda}\,, \end{aligned}$$

and substituting this result into Equation (13.105) yields a second expression for the principal curvature,

$$k_{np} = \frac{b_{11} + b_{12}\lambda}{g_{11} + g_{12}\lambda}\,. \tag{13.106}$$

Returning to the original variables, Equations (13.105) and (13.106) give the system of equations

$$\begin{aligned} (b_{12} - g_{12}k_{np})\,\mathrm{d}u_1 + (b_{22} - g_{22}k_{np})\,\mathrm{d}u_2 &= 0\,, \\ (b_{11} - g_{11}k_{np})\,\mathrm{d}u_1 + (b_{12} - g_{12}k_{np})\,\mathrm{d}u_2 &= 0\,. \end{aligned}$$

Being a homogeneous system of two equations in the two variables $\mathrm{d}u_1$ and $\mathrm{d}u_2$, it has a non-trivial solution if and only if the determinant of the system vanishes:

$$\begin{aligned} (b_{12} - g_{12}k_{np})^2 - (b_{11} - g_{11}k_{np})\,(b_{22} - g_{22}k_{np}) &= 0 \\ g\,k_{np}^2 - (g_{11}b_{22} + g_{22}b_{11} - 2g_{12}b_{12})\,k_{np} + b &= 0\,. \end{aligned}$$

We thus obtain a quadratic equation for k_{np} in terms of the first and second fundamental coefficients. Recall from the algebra of polynomials that for a quadratic equation $ax^2+bx+c = 0$ the product of its roots x_1 and x_2 is $x_1x_2 = c/a$.

Definition 13.29 (Gaussian curvature)**.** We define the *Gaussian curvature* as the product of the two principal curvatures,

$$K = k_{np,1}\,k_{np,2} = \frac{b}{g}\,. \tag{13.107}$$

Rests on Definitions 13.22, 13.25 and 13.28.

Geometric interpretation of the Gaussian curvature

The sign of K classifies the local shape of the surface at a point: where $K > 0$ the two principal curvatures share a sign and the surface curves away from its tangent plane on every side (elliptic point, as on a sphere or an ellipsoid); where $K < 0$ they have opposite signs and the surface crosses its tangent plane (a saddle, as on a hyperboloid or the inner rim of a torus); where $K = 0$ at least one principal curvature vanishes (a cylinder, a cone, or a plane). Figure 13.1 illustrates the flat and cylindrical cases, both with $K = 0$.

Theorema Egregium

The definition $K = b/g$ (Equation (13.107)) is built from *both* fundamental forms: g_{ij}, intrinsic to the surface, and b_{ij}, which records how the surface bends inside the ambient space. Gauss's celebrated theorem is that the second form is, in the end, dispensable.

Theorem 13.30 (Theorema Egregium; Gauss, 1827)**.** *The Gaussian curvature K of a surface is determined by its first fundamental form g_{ij} and derivatives of g_{ij} alone; it does not depend on the second fundamental form b_{ij}, i.e. not on how the surface is embedded in $\mathbb{R}^3$.*

Rests on Definitions 13.27 and 13.29 and Equations (13.88) and (13.95).

Proof. *Derives Theorem 13.30.*

Specializing the Riemann symbol of the second kind (Definition 13.27) to $D = 2$ at $(l, i, j, k) = (1, 2, 1, 2)$, and using $b_{21} = b_{12}$,

$$R_{1212} = b_{22}b_{11} - b_{21}b_{12} = b_{11}b_{22} - b_{12}^2 = \det b_{ij} = b\,, \tag{13.108}$$

so that $K = b/g = R_{1212}/g$. Now trace the definition of R_{1212} backwards through what has already been established: by Equation (13.95), $R^l{}_{ijk}$ — and hence, lowering with g_{lm}, R_{lijk} — is built entirely from the Christoffel symbols $\Gamma_{ij}^{\,k}$ and their first derivatives; and by Equation (13.88), $\Gamma_{ij}^{\,k}$ is itself built entirely from g_{ij} and its first derivatives. No occurrence of b_{ij} survives this substitution: R_{1212}, and therefore

$$K = \frac{R_{1212}}{g}\,, \tag{13.109}$$

is a function of g_{ij} and its derivatives up to second order alone. This proves the theorem; K is an *intrinsic* invariant of the surface. □

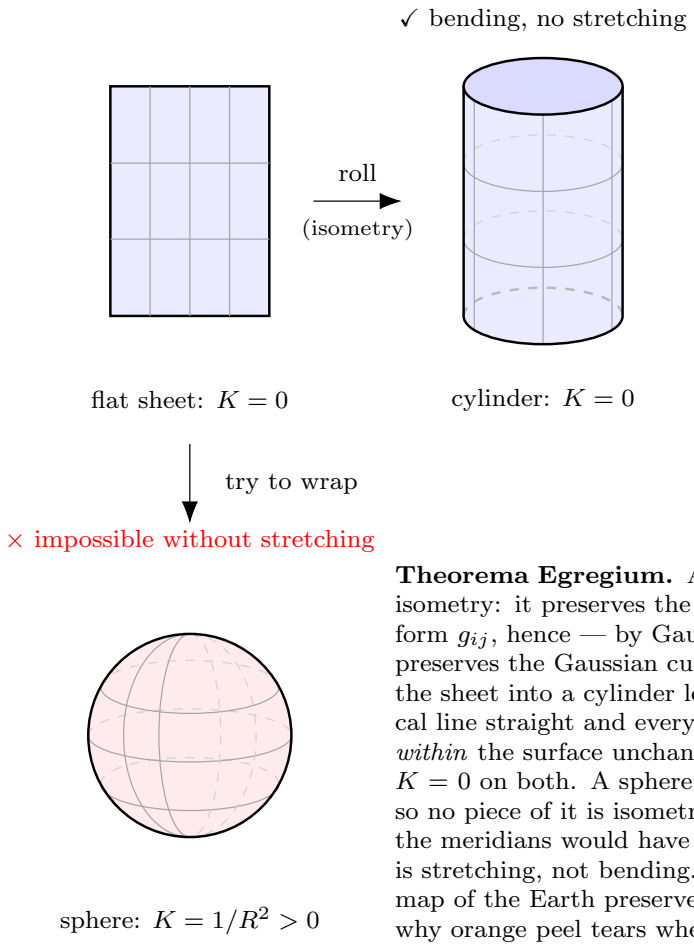
Corollary 13.31 (Bending invariance)**.** *If $f : S \longrightarrow S'$ is a local isometry between two surfaces (Definition 13.120 specialized to $D = 2$), then $K_{S'}(f(P)) = K_S(P)$ for every $P \in S$.*

Rests on Theorem 13.30 and Definition 13.120.

Proof. *Derives Corollary 13.31.*

An isometry pulls back the metric to itself, $f^*g_{S'} = g_S$ (Equation (13.262)). Since K is, by Theorem 13.30, a fixed differential expression in the metric components alone, $K_S(P) = K_{S'}(f(P))$ follows by evaluating that same expression on both sides of the pullback. □

A flat sheet of paper bent, without stretching or tearing, into a cylinder realizes exactly this corollary: the bending is a local isometry (distances measured *along* the surface do not change), and indeed both surfaces have $K = 0$ (Figure 13.1). The same sheet cannot be bent into any piece of a sphere of radius R, whose curvature $K = 1/R^2 \neq 0$ (Example 13.33 below) can never match the flat sheet's $K = 0$ — this is the everyday fact that maps of the round Earth cannot be flat without distortion, and the reason orange peel tears when flattened.



Theorema Egregium. A bending is a local isometry: it preserves the first fundamental form g_{ij}, hence — by Gauss's theorem — it preserves the Gaussian curvature K. Rolling the sheet into a cylinder leaves every vertical line straight and every distance measured *within* the surface unchanged, and indeed $K = 0$ on both. A sphere has $K = 1/R^2 \neq 0$, so no piece of it is isometric to a flat sheet: the meridians would have to converge, which is stretching, not bending. This is why no flat map of the Earth preserves all distances, and why orange peel tears when flattened.

Figure 13.1: Isometric bending preserves K (flat sheet to cylinder, both $K = 0$); no bending of a flat sheet reaches the sphere, whose $K = 1/R^2 > 0$ obstructs it by Theorem 13.30.

Remark 13.32 (Restatement via the Cartan formalism). The proof above reuses the nineteenth-century route already built into Equations (13.88) and (13.95). The Cartan formalism of Section 13.12, developed later in this chapter, makes the same content more transparent and supplies, as a byproduct, an explicit formula for K.

Fix any local orthogonal coordinates (u_1, u_2) on the surface, so that $g_{11} = E(u_1, u_2)$, $g_{22} = G(u_1, u_2)$, $g_{12} = 0$ (such coordinates always exist locally — e.g. the curvature lines of Equation (13.104) away from umbilic points). An orthonormal coframe is then $e^1 = \sqrt{E}\,\mathrm{d}u_1$, $e^2 = \sqrt{G}\,\mathrm{d}u_2$, and the torsion-free structure equation Equation (13.311) in $D = 2$ reads $\mathrm{d}e^1 = -\omega^1{}_2 \wedge e^2$, $\mathrm{d}e^2 = \omega^1{}_2 \wedge e^1$, a linear system for the single connection one-form $\omega^1{}_2$ that involves *only* $\mathrm{d}e^1, \mathrm{d}e^2$, hence only E, G. Solving,

$$\omega^1{}_2 = \frac{\partial_2 E}{2\sqrt{EG}}\,\mathrm{d}u_1 - \frac{\partial_1 G}{2\sqrt{EG}}\,\mathrm{d}u_2\,, \tag{13.110}$$

which is verified by substitution: with $\mathrm{d}e^1 = -\dfrac{\partial_2 E}{2\sqrt{E}}\,\mathrm{d}u_1 \wedge \mathrm{d}u_2$ and $\mathrm{d}e^2 = \dfrac{\partial_1 G}{2\sqrt{G}}\,\mathrm{d}u_1 \wedge \mathrm{d}u_2$, one checks $-\omega^1{}_2 \wedge e^2 = \mathrm{d}e^1$ and $\omega^1{}_2 \wedge e^1 = \mathrm{d}e^2$ directly. In $D = 2$ the Lorentz algebra $\mathfrak{so}(2)$ is abelian, so the quadratic term of Equation (13.312) vanishes identically and $R^1{}_2 = \mathrm{d}\omega^1{}_2$; defining K through $R^1{}_2 = K\, e^1 \wedge e^2$ (consistent with Equation (13.107), checked below) and differentiating Equation (13.110) gives the classical *Liouville formula*,

$$K = -\frac{1}{2\sqrt{EG}}\left[\partial_1\!\left(\frac{\partial_1 G}{\sqrt{EG}}\right) + \partial_2\!\left(\frac{\partial_2 E}{\sqrt{EG}}\right)\right], \tag{13.111}$$

manifestly a functional of E, G and their derivatives alone [1575]. Not one component of $\omega^1{}_2$, hence not one term of Equation (13.111), involves the embedding: the Cartan route makes the Theorema Egregium visible *by construction* rather than by cancellation, since b_{ij} never entered the computation.

Example 13.33 (Sphere of radius R)**.** In spherical coordinates $(u_1, u_2) = (\theta, \phi)$, $E = R^2$, $G = R^2 \sin^2\theta$, so $\partial_2 E = 0$ and $\partial_1 G = 2R^2 \sin\theta\cos\theta$. Then $\partial_1 G/\sqrt{EG} = 2\cos\theta$, whose θ-derivative is $-2\sin\theta$, and Equation (13.111) gives

$$K = -\frac{1}{2R^2\sin\theta}\,(-2\sin\theta) = \frac{1}{R^2}\,,$$

confirming $K = 1/R^2$ used above and matching the direct computation $K = k_{np,1}k_{np,2} = (1/R)(1/R)$ from Definition 13.29, since $b_{ij} = \pm g_{ij}/R$ gives $b = \det b_{ij} = g/R^2$ regardless of the sign convention for the normal.

Remark 13.34 (Consistency of the two computations of $D = 2$ curvature). Definition 13.153 already recorded, for a general torsion-free connection, that $R = 2K$ in $D = 2$ (with R the Ricci scalar). The identity $\Omega = \epsilon_{12}R^{12} + \epsilon_{21}R^{21} = 2R^1{}_2 = 2K\, e^1 \wedge e^2 = 2K\sqrt{g}\,\mathrm{d}u_1 \wedge \mathrm{d}u_2 = R\sqrt{g}\,\mathrm{d}u_1 \wedge \mathrm{d}u_2$ (using $R = 2K$ in the last step), i.e. $\epsilon_{ab}R^{ab} = R\sqrt{g}\,\mathrm{d}^2u$ with *no* extra numerical factor, is exactly the $D = 2$ special case ($(D-2)! = 0! = 1$) of the Einstein–Hilbert equivalence proven in Appendix A.8 — an independent check of the general-D combinatorial factor derived there.

Geodesic curvature

Definition 13.35 (Projection curve)**.** We define the *projection curve* as the curve projected onto the tangent plane at some point of a surface.

Definition 13.36 (Geodesic curvature vector)**.** Under the same conditions used to define the normal curvature (see Equation (13.97)), we define the *geodesic curvature vector* as the normal curvature vector of the projection curve.

Rests on Definitions 13.13, 13.28 and 13.35.

We may take the curve to pass through the origin of an orthogonal system; the curve projected onto the tangent plane, called $\boldsymbol{x}_g$, is then given by

$$\boldsymbol{x}_g = \left(\boldsymbol{x}\cdot\hat{T}'\right)\hat{T}' + \left(\boldsymbol{x}\cdot\hat{N}'\right)\hat{N}', \tag{13.112}$$

where $\hat{T}'$ and $\hat{N}'$ span the tangent plane at the point of tangency. Differentiating with respect to the arc-length parameter of the curve on the surface,

$$\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s} = \left(\hat{T}\cdot\hat{T}'\right)\hat{T}' + \left(\hat{T}\cdot\hat{N}'\right)\hat{N}', \tag{13.113}$$

$$\frac{\mathrm{d}^2\boldsymbol{x}_g}{\mathrm{d}s^2} = \left(\boldsymbol{k}\cdot\hat{T}'\right)\hat{T}' + \left(\boldsymbol{k}\cdot\hat{N}'\right)\hat{N}'. \tag{13.114}$$

On the other hand, from Equation (13.39), the normal curvature vector of the projected curve is given by

$$\boldsymbol{k}_g = \frac{\mathrm{d}\hat{T}_g}{\mathrm{d}s_g},$$

where $\hat{T}_g$ is the unit tangent vector defined in Equation (13.38) and s_g is the arc-length parameter defined in Equation (13.35), both for the projection curve. By the chain rule we may write

$$\boldsymbol{k}_g = \frac{\mathrm{d}\hat{T}_g}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}s_g}.$$

From Equation (13.37) we see that

$$\frac{\mathrm{d}s_g}{\mathrm{d}s} = \left|\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\right|,$$

so that

$$\boldsymbol{k}_g = \frac{\frac{\mathrm{d}\hat{T}_g}{\mathrm{d}s}}{\left|\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\right|}. \tag{13.115}$$

Furthermore, from the definition of the unit tangent vector (Equation (13.38)),

$$\begin{aligned}\hat{T}_g &= \frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s_g}\\ &= \frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}s_g}.\end{aligned} \tag{13.116}$$

Again by Equation (13.37), substituting in Equation (13.116) and differentiating,

$$\begin{aligned}\hat{T}_g &= \frac{\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}}{\left|\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\right|}, \qquad \frac{\mathrm{d}}{\mathrm{d}s}(\,)\\ \frac{\mathrm{d}\hat{T}_g}{\mathrm{d}s} &= \frac{\frac{\mathrm{d}^2\boldsymbol{x}_g}{\mathrm{d}s^2}\left|\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\right| - \frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\left(\frac{\mathrm{d}^2\boldsymbol{x}_g}{\mathrm{d}s^2}\cdot\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\right)}{\left|\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\right|^2}\\ \text{[Equation (13.115)]}\quad \boldsymbol{k}_g &= \frac{\frac{\mathrm{d}^2\boldsymbol{x}_g}{\mathrm{d}s^2}\left|\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\right| - \frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\left(\frac{\mathrm{d}^2\boldsymbol{x}_g}{\mathrm{d}s^2}\cdot\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\right)}{\left|\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\right|^3}.\end{aligned} \tag{13.117}$$

Finally, since in this case we are considering that at the point of tangency $\hat{T} = \hat{T}'$, from Equations (13.113) and (13.114),

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s} &= \hat{T}' + \left(\hat{T}'\cdot\hat{N}'\right)\hat{N}',\\ \frac{\mathrm{d}^2\boldsymbol{x}_g}{\mathrm{d}s^2} &= \left(\boldsymbol{k}\cdot\hat{T}\right)\hat{T} + \left(\boldsymbol{k}\cdot\hat{N}'\right)\hat{N}',\end{aligned}$$

and from Equations (13.40) and (13.42),

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s} &= \hat{T}' \,, \\ \frac{\mathrm{d}^2\boldsymbol{x}_g}{\mathrm{d}s^2} &= \left(\boldsymbol{k}\cdot\hat{N}'\right)\hat{N}' \\ \implies \quad \left|\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s}\right| &= 1 \,, \\ \implies \quad \frac{\mathrm{d}^2\boldsymbol{x}_g}{\mathrm{d}s^2}\cdot\frac{\mathrm{d}\boldsymbol{x}_g}{\mathrm{d}s} &= 0 \,. \end{aligned}$$

Hence Equation (13.117) reduces to

$$\boldsymbol{k}_g = \left(\boldsymbol{k}\cdot\hat{N}'\right)\hat{N}' \,. \tag{13.118}$$

Considering Equation (13.97), we see that

$$\boldsymbol{k} = \boldsymbol{k}_g + \boldsymbol{k}_n \,. \tag{13.119}$$

Alternative expression for the geodesic curvature vector. Consider a curve on a surface $\boldsymbol{x}$ parametrised by arc length (see Equation (13.56)). From the definition of the normal curvature vector of the curve (Equation (13.39)),

$$\begin{aligned} \boldsymbol{k} &= \frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s}\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}s}\left(\boldsymbol{x}_i\frac{\mathrm{d}u_i}{\mathrm{d}s}\right) \\ &= \boldsymbol{x}_{ij}\frac{\mathrm{d}u_i\mathrm{d}u_j}{\mathrm{d}s^2} + \boldsymbol{x}_i\frac{\mathrm{d}^2u_i}{\mathrm{d}s^2} \\ \text{[Equation (13.83)]} \quad &= \Gamma_{ij}{}^k\frac{\mathrm{d}u_i\mathrm{d}u_j}{\mathrm{d}s^2}\boldsymbol{x}_k + \frac{b_{ij}\mathrm{d}u_i\mathrm{d}u_j}{\mathrm{d}s^2}\hat{n} + \boldsymbol{x}_k\frac{\mathrm{d}^2u_k}{\mathrm{d}s^2} \\ \text{[Equation (13.100)]} \quad &= \left(\Gamma_{ij}{}^k\frac{\mathrm{d}u_i\mathrm{d}u_j}{\mathrm{d}s^2} + \frac{\mathrm{d}^2u_k}{\mathrm{d}s^2}\right)\boldsymbol{x}_k + \boldsymbol{k}_n \,. \end{aligned}$$

Comparing with Equation (13.119), we see that

$$\boldsymbol{k}_g = \left(\Gamma_{ij}{}^k\frac{\mathrm{d}u_i\mathrm{d}u_j}{\mathrm{d}s^2} + \frac{\mathrm{d}^2u_k}{\mathrm{d}s^2}\right)\boldsymbol{x}_k \,. \tag{13.120}$$

Geodesic lines

Consider two points on a surface, and a curve on the surface containing them. Among all curves satisfying this, there must be one of least arc length, as given by Equation (13.67). Intuitively, such a curve possesses a projection curve with vanishing geodesic curvature.

Definition 13.37 (Geodesic lines)**.** We define the *geodesic lines* as the curves whose projection curve has vanishing geodesic curvature, that is, the curves satisfying

$$\Gamma_{ij}{}^k\frac{\mathrm{d}u_i\mathrm{d}u_j}{\mathrm{d}s^2} + \frac{\mathrm{d}^2u_k}{\mathrm{d}s^2} = 0 \,, \tag{13.121}$$

with the initial conditions

$$u_i(s=0) = u_i^0 \quad \wedge \quad \frac{\mathrm{d}u_i}{\mathrm{d}s}(s=0) = \dot{u}_i^0 \,. \tag{13.122}$$

Rests on Definitions 13.26 and 13.36 and Equation (13.120).

By the calculus of variations it can be proved that Equation (13.121) results from extremising the arc length given by Equation (13.67), as the following proposition shows.

Proposition 13.38 (Geodesics as stationary curves of the arc length)**.** *Let $u_i(s)$ be a regular curve on a surface, parametrised by arc length. It is a stationary point of the arc-length functional Equation (13.67) among curves with the same endpoints if and only if it satisfies the geodesic equation Equation (13.121).*

Rests on Equations (13.67), (13.87) and (13.121).

Proof. *Derives Proposition 13.38.*

Write the functional Equation (13.67) for a curve $u_i(t)$ joining two fixed points as

$$L[u] = \int_{t_0}^{t_1} F(u, \dot{u})\, \mathrm{d}t\,, \qquad F(u, \dot{u}) = \sqrt{g_{ij}(u)\, \dot{u}_i \dot{u}_j}\,, \qquad \dot{u}_i = \frac{\mathrm{d}u_i}{\mathrm{d}t}\,. \tag{13.123}$$

For a regular curve $F > 0$, so F is differentiable in its arguments and the Euler–Lagrange equations of Chapter 16,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial F}{\partial \dot{u}_k}\right) - \frac{\partial F}{\partial u_k} = 0\,, \tag{13.124}$$

are the condition for stationarity under variations vanishing at the endpoints. The functional is invariant under reparametrisation (Definition 13.10), so its stationary points come in reparametrisation classes and we may work with the representative parametrised by arc length, for which $F \equiv 1$ and $t = s$. With

$$\frac{\partial F}{\partial \dot{u}_k} = \frac{g_{kj}\dot{u}_j}{F}\,, \qquad \frac{\partial F}{\partial u_k} = \frac{1}{2F}\, \partial_k g_{ij}\, \dot{u}_i \dot{u}_j\,,$$

setting $F = 1$ turns Equation (13.124) into

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}s}\left(g_{kj}\dot{u}_j\right) - \frac{1}{2}\, \partial_k g_{ij}\, \dot{u}_i \dot{u}_j \\ &= g_{kj}\ddot{u}_j + \partial_i g_{kj}\, \dot{u}_i \dot{u}_j - \frac{1}{2}\, \partial_k g_{ij}\, \dot{u}_i \dot{u}_j\,. \end{aligned}$$

The product $\dot{u}_i \dot{u}_j$ is symmetric in i and j, so only the symmetric part of $\partial_i g_{kj}$ in those indices contributes and we may replace it by $\frac{1}{2}(\partial_i g_{kj} + \partial_j g_{ki})$, giving

$$g_{kj}\ddot{u}_j + \frac{1}{2}\left(\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}\right) \dot{u}_i \dot{u}_j = 0\,, \tag{13.125}$$

where the symmetry $g_{kj} = g_{jk}$ was used. The bracket is exactly the Christoffel symbol of the first kind Equation (13.87), Γ_{ijk}, so Equation (13.125) reads $g_{kj}\ddot{u}_j + \Gamma_{ijk}\dot{u}_i\dot{u}_j = 0$. Contracting with the inverse metric g^{kl} and using $\Gamma_{ij}^{\ \ l} = \Gamma_{ijk}g^{kl}$ (Definition 13.26) gives

$$\frac{\mathrm{d}^2 u_l}{\mathrm{d}s^2} + \Gamma_{ij}^{\ \ l}\, \frac{\mathrm{d}u_i \mathrm{d}u_j}{\mathrm{d}s^2} = 0\,,$$

which is Equation (13.121). Every step is an equivalence, so the converse holds as well. □

Remark 13.39. Stationarity is necessary but not sufficient for minimality. On the sphere the geodesics are the great circles, and the longer of the two arcs joining two non-antipodal points satisfies Equation (13.121) while plainly failing to minimise the length. What is true is local: a sufficiently short piece of a geodesic is the shortest curve between its endpoints. On a pseudo-Riemannian manifold Equation (13.121) becomes the autoparallel equation Equation (13.297) of the Levi-Civita connection, but the minimising statement does *not* carry over verbatim. The line element changes sign there, so the functional has to be

written $\int \sqrt{|\mathrm{d}s^2|}$ (Remark 13.151), and the conclusion then depends on the causal character of the curve: a short enough spacelike geodesic still minimises the arc length, whereas a short enough timelike geodesic *maximises* the proper time between its endpoints — the twin-paradox inequality of Part IV — and on a null curve the functional vanishes identically, so every null curve is stationary for trivial reasons and the null geodesics are singled out by Equation (13.297) rather than by any extremal property of the length. See Remark 13.151, where the distinction between the shortest and the straightest curve is what torsion measures.

Geodesic parallelism

Consider a curve $\boldsymbol{x} = \boldsymbol{x}(u_1(t), u_2(t))$ on the surface $\boldsymbol{x} = \boldsymbol{x}(u_1, u_2)$. Assign to the curve a vector field $\boldsymbol{A}$ of the form

$$\boldsymbol{A}(t) = v_i(t)\, \boldsymbol{x}_i \,. \tag{13.126}$$

At each point of the curve, $\boldsymbol{A}$ is written as a linear combination of the partial tangent vectors of the surface, and therefore lies in the tangent plane. From vector analysis, the vectors $\boldsymbol{A}$ are parallel for every value of t if and only if

$$\boldsymbol{A}(t) \parallel \boldsymbol{A}(t + \mathrm{d}t) \quad \Longleftrightarrow \quad \frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}t} = \boldsymbol{0} \quad \forall\, t \,. \tag{13.127}$$

Analogously, we say that the vectors $\boldsymbol{A}$ are *geodesically parallel* for every value of t if and only if

$$\boldsymbol{A}(t) \parallel_g \boldsymbol{A}(t + \mathrm{d}t) \quad \Longleftrightarrow \quad \boldsymbol{x}_l \cdot \frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}t} = 0 \quad \forall\, t \ \wedge\ \forall\, l \,. \tag{13.128}$$

From Equation (13.126),

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}t} &= \frac{\mathrm{d}v_i}{\mathrm{d}t}\boldsymbol{x}_i + v_i \boldsymbol{x}_{ij}\frac{\mathrm{d}u_j}{\mathrm{d}t} \\
\text{[Equation (13.83)]} \quad &= \frac{\mathrm{d}v_i}{\mathrm{d}t}\boldsymbol{x}_i + v_i \left(\Gamma_{ij}^{\ k}\boldsymbol{x}_k + b_{ij}\hat{n}\right)\frac{\mathrm{d}u_j}{\mathrm{d}t} \\
&= \frac{\mathrm{d}v_i}{\mathrm{d}t}\boldsymbol{x}_i + \Gamma_{ij}^{\ k} v_i \frac{\mathrm{d}u_j}{\mathrm{d}t}\boldsymbol{x}_k + b_{ij} v_i \frac{\mathrm{d}u_j}{\mathrm{d}t}\hat{n} \,, \qquad \boldsymbol{x}_l \cdot (\ ) \\
\text{[Equations (13.54) and (13.60)]} \quad \boldsymbol{x}_l \cdot \frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}t} &= \frac{\mathrm{d}v_i}{\mathrm{d}t} g_{il} + \Gamma_{ij}^{\ k} g_{kl} v_i \frac{\mathrm{d}u_j}{\mathrm{d}t} \\
&= \frac{\mathrm{d}v_l}{\mathrm{d}t} + \Gamma_{ijl}\, v_i \frac{\mathrm{d}u_j}{\mathrm{d}t} \,.
\end{aligned}$$

Therefore the field $\boldsymbol{A}$ is geodesically parallel if and only if

$$\frac{\mathrm{d}v_l}{\mathrm{d}t} + \Gamma_{ijl}\, v_i \frac{\mathrm{d}u_j}{\mathrm{d}t} = 0 \,, \qquad g^{lk}(\ )$$

whence the condition of geodesic parallelism is

$$\frac{\mathrm{d}v^k}{\mathrm{d}t} + \Gamma_{ij}^{\ k} \frac{\mathrm{d}u_j}{\mathrm{d}t} v_i = 0 \,. \tag{13.129}$$

Holonomy and the local Gauss–Bonnet theorem

Transport a vector geodesically parallel all the way around a closed curve on a surface and it comes back to the point it started from — but not, in general, to the vector it started as. The failure is the *holonomy* of the curve, and the theorem of this subsection identifies it: it is the total Gaussian curvature enclosed. That is the local form of the Gauss–Bonnet theorem, and it is what makes curvature an observable rather than a bookkeeping device, since the angle it predicts can be measured by carrying a gyroscope, or a swinging pendulum, around a circuit.

Proposition 13.40 (Parallel transport is an isometry of the tangent plane)**.** *If $\boldsymbol{A}$ and $\boldsymbol{B}$ are geodesically parallel along a curve on a surface, then $\boldsymbol{A}\cdot\boldsymbol{B}$ is constant along it. In particular lengths and angles are preserved, so transport around a closed curve acts on the tangent plane as a rotation.*

Rests on Equations (13.126) *and* (13.129) *and Definition 13.21.*

Derivation. Derives Proposition 13.40.

Both fields are tangent to the surface, and geodesic parallelism says that $\mathrm{d}\boldsymbol{A}/\mathrm{d}t$ has vanishing tangential part, i.e. $\boldsymbol{x}_l\cdot\mathrm{d}\boldsymbol{A}/\mathrm{d}t=0$ for every l; since $\boldsymbol{B}=v_i\boldsymbol{x}_i$ is a combination of the $\boldsymbol{x}_l$, $\boldsymbol{B}\cdot\mathrm{d}\boldsymbol{A}/\mathrm{d}t=0$, and likewise with the roles exchanged. Hence $\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{A}\cdot\boldsymbol{B})=\boldsymbol{B}\cdot\frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}t}+\boldsymbol{A}\cdot\frac{\mathrm{d}\boldsymbol{B}}{\mathrm{d}t}=0$. A linear map of a two-dimensional Euclidean plane preserving the inner product is a rotation or a reflection, and transport is a rotation because it is reached continuously from the identity at $t=0$. □

Theorem 13.41 (Holonomy equals the enclosed curvature; local Gauss–Bonnet)**.** *Let (u_1,u_2) be orthogonal coordinates on a surface, with $g_{11}=E>0$, $g_{22}=G>0$, $g_{12}=0$, and let γ be a piecewise smooth simple closed curve, positively oriented in the (u_1,u_2) plane, bounding a region D whose closure lies in the coordinate patch. Write ϑ for the angle a geodesically parallel field makes with the orthonormal frame*

$$\hat{e}_1=\frac{\boldsymbol{x}_1}{\sqrt{E}}\,,\qquad \hat{e}_2=\frac{\boldsymbol{x}_2}{\sqrt{G}}\,. \tag{13.130}$$

Then the holonomy of γ — the total change of ϑ on one circuit — is

$$\Delta\vartheta=\int_D K\,\mathrm{d}A\,,\qquad \mathrm{d}A=\sqrt{EG}\,\mathrm{d}u_1\mathrm{d}u_2\,, \tag{13.131}$$

with K the Gaussian curvature Equation (13.107)*. Reversing the orientation of γ reverses the sign.*

Rests on Proposition 13.40, Equation (13.111)*, and Theorem 7.99.*

Proof. Derives Theorem 13.41.

The angle obeys a first-order equation. By Proposition 13.40 a parallel field has constant length, so we may write $\boldsymbol{A}=|\boldsymbol{A}|\,(\cos\vartheta\,\hat{e}_1+\sin\vartheta\,\hat{e}_2)$. Put $\hat{e}_\perp=-\sin\vartheta\,\hat{e}_1+\cos\vartheta\,\hat{e}_2$, a unit tangent vector orthogonal to $\boldsymbol{A}$. Since $\hat{e}_1\cdot\hat{e}_1=1$ the tangential part of $\mathrm{d}\hat{e}_1/\mathrm{d}t$ is a multiple of $\hat{e}_2$; call the coefficient $\varpi=\hat{e}_2\cdot\mathrm{d}\hat{e}_1/\mathrm{d}t$, and then $\hat{e}_1\cdot\mathrm{d}\hat{e}_2/\mathrm{d}t=-\varpi$ because $\hat{e}_1\cdot\hat{e}_2=0$. Taking the scalar product of $\mathrm{d}\boldsymbol{A}/\mathrm{d}t$ with $\hat{e}_\perp$ and dividing by $|\boldsymbol{A}|$,

$$\frac{\hat{e}_\perp}{|\boldsymbol{A}|}\cdot\frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}t}=\frac{\mathrm{d}\vartheta}{\mathrm{d}t}+\cos\vartheta\left(\hat{e}_\perp\cdot\frac{\mathrm{d}\hat{e}_1}{\mathrm{d}t}\right)+\sin\vartheta\left(\hat{e}_\perp\cdot\frac{\mathrm{d}\hat{e}_2}{\mathrm{d}t}\right)=\frac{\mathrm{d}\vartheta}{\mathrm{d}t}+\varpi\,,$$

the last step using $\hat{e}_\perp\cdot\mathrm{d}\hat{e}_1/\mathrm{d}t=\varpi\cos\vartheta$ and $\hat{e}_\perp\cdot\mathrm{d}\hat{e}_2/\mathrm{d}t=\varpi\sin\vartheta$ together with $\cos^2+\sin^2=1$. The left-hand side vanishes for a parallel field, because $\hat{e}_\perp$ is tangent, so

$$\frac{\mathrm{d}\vartheta}{\mathrm{d}t}=-\varpi\,. \tag{13.132}$$

ϖ is a one-form built from E and G alone. Using $\boldsymbol{x}_1\cdot\boldsymbol{x}_1=E$, $\boldsymbol{x}_2\cdot\boldsymbol{x}_2=G$, $\boldsymbol{x}_1\cdot\boldsymbol{x}_2=0$,

$$\boldsymbol{x}_2\cdot\boldsymbol{x}_{11}=\partial_1(\boldsymbol{x}_2\cdot\boldsymbol{x}_1)-\boldsymbol{x}_{12}\cdot\boldsymbol{x}_1=-\tfrac{1}{2}\partial_2E\,,\qquad \boldsymbol{x}_2\cdot\boldsymbol{x}_{12}=\tfrac{1}{2}\partial_1G\,,$$

so that, since the term coming from differentiating $E^{-1/2}$ is proportional to $\boldsymbol{x}_1$ and dies against $\hat{e}_2$,

$$\varpi\,\mathrm{d}t=\omega_1\mathrm{d}u_1+\omega_2\mathrm{d}u_2\,,\qquad \omega_1=-\frac{\partial_2E}{2\sqrt{EG}}\,,\qquad \omega_2=\frac{\partial_1G}{2\sqrt{EG}}\,. \tag{13.133}$$

This is minus the connection one-form $\omega^1{}_2$ of Equation (13.110), computed here again directly so that nothing is owed to the later Cartan formalism.

Green's theorem. Integrating Equation (13.132) once around γ and applying Theorem 7.99 in the (u_1, u_2) plane,

$$\Delta\vartheta = -\oint_\gamma (\omega_1 \mathrm{d}u_1 + \omega_2 \mathrm{d}u_2) = -\int_D (\partial_1\omega_2 - \partial_2\omega_1)\,\mathrm{d}u_1\mathrm{d}u_2\,.$$

By Equation (13.133),

$$\partial_1\omega_2 - \partial_2\omega_1 = \frac{1}{2}\left[\partial_1\left(\frac{\partial_1 G}{\sqrt{EG}}\right) + \partial_2\left(\frac{\partial_2 E}{\sqrt{EG}}\right)\right] = -K\sqrt{EG}\,,$$

the last step being the Liouville formula Equation (13.111). Substituting gives Equation (13.131). Reversing the orientation of γ reverses the sign of the line integral and hence of $\Delta\vartheta$. □

Example 13.42 (The sphere, the solid angle, and the pole)**.** On the sphere of radius R in the coordinates $(u_1, u_2) = (\theta, \phi)$ of Example 13.33, $E = R^2$ and $G = R^2\sin^2\theta$, so $\partial_2 E = 0$, $\partial_1 G = 2R^2\sin\theta\cos\theta$ and $\sqrt{EG} = R^2\sin\theta$. Equation (13.133) then gives the strikingly simple

$$\omega_1 \mathrm{d}\theta + \omega_2 \mathrm{d}\phi = \cos\theta\,\mathrm{d}\phi\,, \tag{13.134}$$

so a parallel field carried once round the latitude circle $\theta = \theta_0$, in the direction of increasing ϕ, turns relative to the frame Equation (13.130) by

$$\Delta\vartheta\,(\theta_0) = -\oint \cos\theta_0\,\mathrm{d}\phi = -2\pi\cos\theta_0\,. \tag{13.135}$$

The theorem itself is checked on a curve that does satisfy its hypotheses — the boundary of the coordinate rectangle $\theta \in [\theta_1, \theta_2]$, $\phi \in [\phi_1, \phi_2]$, whose closure lies in the patch. Only the two edges of constant θ contribute to $\oint \cos\theta\,\mathrm{d}\phi$, and with the positive orientation they give $(\phi_2 - \phi_1)(\cos\theta_2 - \cos\theta_1)$, so $\Delta\vartheta = (\phi_2 - \phi_1)(\cos\theta_1 - \cos\theta_2)$; and since $K = 1/R^2$ with $\mathrm{d}A = R^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$, the right-hand side of Equation (13.131) is $(\phi_2 - \phi_1)\int_{\theta_1}^{\theta_2}\sin\theta\,\mathrm{d}\theta$, the same number.

Comparing Equation (13.135) at θ_0 and at a small ε gives, by the same direct computation,

$$\Delta\vartheta\,(\theta_0) - \Delta\vartheta\,(\varepsilon) = 2\pi\,(\cos\varepsilon - \cos\theta_0) = \int_{\varepsilon\le\theta\le\theta_0} K\,\mathrm{d}A\,,$$

the total curvature of the annulus between the two latitudes.

The cap $\theta \le \theta_0$ itself is *not* covered by this chart: the pole $\theta = 0$ is a coordinate singularity, where the frame Equation (13.130) is undefined, and the hypothesis of the theorem fails. The bookkeeping is exact and worth stating, because it is the whole content of the Foucault deficit. Letting $\varepsilon \to 0$ above, the total curvature of the cap is

$$\int_{\mathrm{cap}} K\,\mathrm{d}A = 2\pi\,(1 - \cos\theta_0)\,, \tag{13.136}$$

which is precisely the solid angle the cap subtends at the centre, while Equation (13.135) gives $\Delta\vartheta = -2\pi\cos\theta_0$: the two differ by exactly 2π, the limit of $-\Delta\vartheta(\varepsilon)$, which is the turn the singular frame itself makes on one circuit of the pole. So it is the *deficit from a full turn*, $2\pi - 2\pi\cos\theta_0$, and not the frame-relative rotation, that is the geometric invariant of the region enclosed. Written in terms of the latitude $\lambda = \pi/2 - \theta_0$, that deficit is $2\pi\,(1 - \sin\lambda)$.

Rests on Theorem 13.41 and Example 13.33.

13.4 Differentiable manifolds

The definitions given below serve to introduce the concept of a differentiable manifold, which results from generalising the idea of curves and surfaces. The definition is built up from the concept of a topological space (Chapter 6), which we progressively equip with the pertinent structure so as to obtain an object rich in properties that can be linked with nature. One of the key concepts is that of a coordinate system, which is in intimate correspondence with reference frames: the tool of theoretical physics that lets us fix a point in the Universe with respect to which a physical system is being evaluated.

13.4.1 Charts, atlases, and differentiable structure

Coordinate systems

Consider a topological space. In the context of this chapter we refer to the elements of a topological space as *points*.

Let $M \subset \mathbb{R}^N$ be a topological space and $U \subset \mathbb{R}^n$ with $n \leq N$. Consider a point $P \in M$ and $x = (x^1, \ldots, x^n) \in U$, and consider further the injective map $\phi : x \longmapsto \phi(x) = P$.

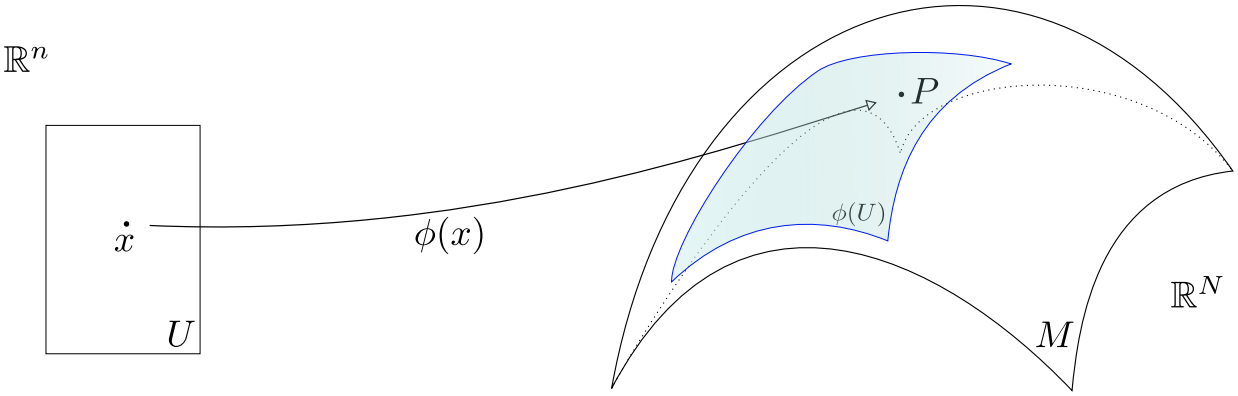


Figure 13.2: A coordinate system: the map ϕ carries a point $x = (x^1, \ldots, x^n)$ of the parameter domain $U \subset \mathbb{R}^n$ to a point $P = \phi(x)$ of the topological space M.

Intuitively, ϕ must respect the topology of M in order to describe its points correctly through coordinates. To this end we must demand that points of M be close if and only if their coordinates are. The idea is to require

$$P_l \longrightarrow P \quad \overset{!}{\Longleftrightarrow} \quad x_l \longrightarrow x\,,$$

where x_l is a sequence of points in U and $P_l = \phi(x_l)$ a sequence of points in M. Demanding "$\Leftarrow$" we meet the continuity of ϕ, and demanding "$\Rightarrow$" we meet the continuity of ϕ^{-1}. We further wish a one-to-one relation between the points of U and (at least some of, since in general more than one set like U is needed to describe all the points of M) the points of M, which suggests that ϕ must be bijective. In conclusion, the topology is respected if and only if ϕ is a homeomorphism. By the definition of continuity, demanding "$\Leftarrow$" the set $\phi^{-1}(\phi(U)) = U$ must be open, and demanding "$\Rightarrow$" the set $(\phi^{-1})^{-1}(U) = \phi(U)$ must be open. The geometric interpretation is that (by the definition of an open set) we can place an open ball centred at P and one centred at x,

from which we see that the points neighbouring x should give us information about the points neighbouring P (by the definition of a neighbourhood), which happens if and only if the map ϕ is smooth.

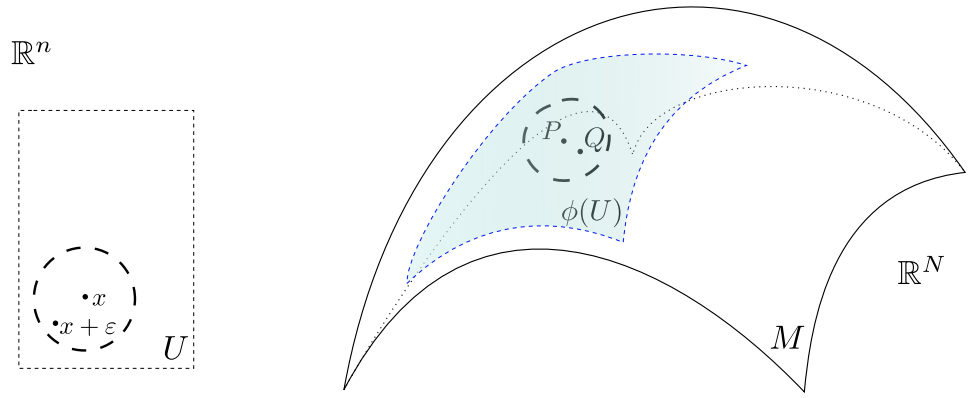


Figure 13.3: Open balls centred at $x \in U$ and at $P = \phi(x) \in M$: the neighbours of x carry information about the neighbours of P precisely when ϕ is a homeomorphism.

Definition 13.43 (Coordinate system)**.** We say that the pair (ϕ, U) is an *n-dimensional coordinate system* (CS), or *local chart*, if and only if

1. the map ϕ is a homeomorphism, which implies that U is open in $\mathbb{R}^n$;
2. the map ϕ is smooth.

Coordinates

As we have seen, given a coordinate system there is a one-to-one relation between the points of U and some points of M. The variables $(x^1, \ldots, x^n)$ corresponding to a point $P \in M$ receive the name of *coordinates* of the point P with respect to the coordinate system (ϕ, U). To symbolise (and not forget) the exact point of M that we wish to describe by means of (ϕ, U), we denote the coordinates of a point $P \in M$ as $(x^1, \ldots, x^n) \overset{\text{not.}}{=} x^i(P)$ with $i = 1, \ldots, n$. Thus, when we want to refer to the point $x \in U$ we write x or $(x^1, \ldots, x^n)$; when we want to refer explicitly to the point P through its coordinates, we write $x^i(P)$; and for the coordinates of arbitrary points of M (within the codomain of ϕ) we use x^i. In summary,

$$\underbrace{x^i(P)}_{\text{coordinates}} = \underbrace{(x^1, \ldots, x^n)}_{\text{point in } U}, \qquad \text{for } P = \phi(x)\,. \tag{13.137}$$

Even though we have insisted on the difference between the quantities in U and the coordinates of the point, numerically they are the same, which is useful in practice; in the theory, however, it is necessary to be rigorous with the notation. It is very important to note that the coordinates of a point depend on the subset U of $\mathbb{R}^n$ we have chosen. Note that $\phi : U \longrightarrow \phi(U)$ is bijective, so this one-to-one relation characterises the points of $\phi(U)$ by their coordinates in $\mathbb{R}^n$. Note also that a correct notation for the point P in terms of its coordinates in U would be $P = \phi(x) = \phi(x^1, \ldots, x^n)$, but not $P = \phi(x^i)$ nor $P = \phi(x^1(P), \ldots, x^n(P))$.

Atlases

Consider a family of n-dimensional coordinate systems $\{(\phi_a, U_a)\}_{a \in I}$ that cover the whole space M, that is,

$$M \subset \bigcup_{a \in I} \phi_a(U_a)\,,$$

such that $U_a \cap U_b \neq \varnothing$ for $a \neq b$. Let the *overlap regions* be

$$\Omega_{ab} = \phi_a(U_a) \cap \phi_b(U_b)\,.$$

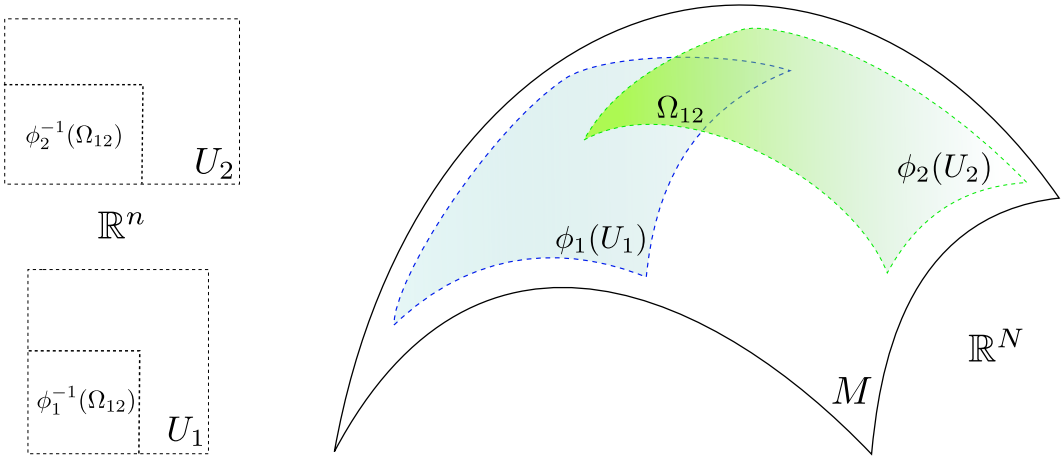


Figure 13.4: Two charts (ϕ_a, U_a) and (ϕ_b, U_b) covering overlapping portions of M; the overlap region is $\Omega_{ab} = \phi_a(U_a) \cap \phi_b(U_b)$.

The sets $\phi_a^{-1}(\Omega_{ab})$ and $\phi_b^{-1}(\Omega_{ab})$ are open because ϕ_a and ϕ_b are homeomorphisms. We further define the *transition maps* as

$$\phi_{ba} = \phi_b^{-1} \circ \phi_a : \phi_a^{-1}(\Omega_{ab}) \longrightarrow \phi_b^{-1}(\Omega_{ab}) \,.$$

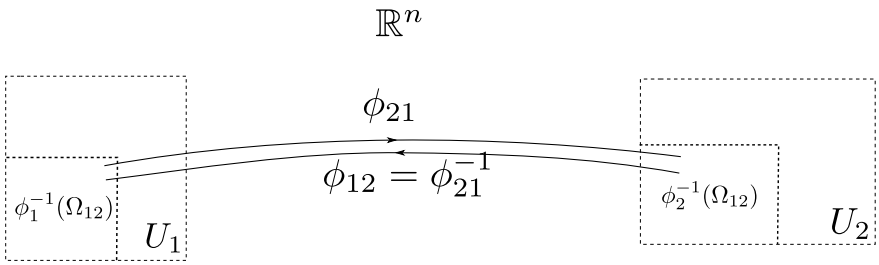


Figure 13.5: The transition map $\phi_{ba} = \phi_b^{-1} \circ \phi_a$ between the coordinate domains of two overlapping charts.

The transition maps are continuous, being compositions of continuous maps, and since the sets $\phi_a^{-1}(\Omega_{ab})$ and $\phi_b^{-1}(\Omega_{ab})$ are open, the transition maps are moreover homeomorphisms. A point P may be described by $P = \phi_a(x_a^1, \ldots, x_a^n)$ or by $P = \phi_b(x_b^1, \ldots, x_b^n)$. This means we can perform a change of coordinate system of the form

$$(x_b^1, \ldots, x_b^n) = \phi_{ba}(x_a^1, \ldots, x_a^n) \,,$$

which respects the topology of M because the ϕ_{ba} are homeomorphisms. For the information about the points neighbouring P to be well defined after changing the coordinate system, we must demand that the transition maps be smooth.

Definition 13.44 (Atlas)**.** We say that a family of coordinate systems $\{(\phi_a, U_a)\}_{a \in I}$ is an *n-dimensional atlas* if and only if

1. the coordinate systems cover the whole space M, that is,

$$M \subset \bigcup_{a \in I} \phi_a(U_a) \,,$$

2. the coordinate systems are smoothly related[1].

Rests on Definition 13.43.

[1]These smoothly related changes of coordinate system are exclusively within the atlas. This concept must not be confused with the general coordinate transformations studied further below.

Differentiable maps on a topological space

Consider now a topological space M equipped with an atlas $\{(\phi_a, U_a)\}_{a\in I}$, and let

$$f : M \longrightarrow \mathbb{R}$$

be a map. A key question is: under what conditions is f differentiable? Since we know the usual definition of differentiability at a point $x \in \mathbb{R}^n$, we wish to extend it to the points of a topological space by virtue of the given atlas. Of course, we know that a map that is not continuous is not differentiable, so at the very least we must demand that f be continuous. Choose the coordinate system (ϕ_a, U_a) and define the map $\bar{f}_a = f \circ \phi_a$.

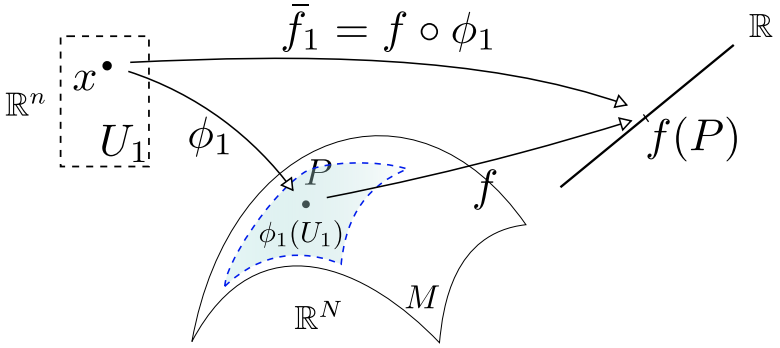


Figure 13.6: A map $f : M \to \mathbb{R}$ read through a chart: the composition $\bar{f}_a = f \circ \phi_a$ is an ordinary map between subsets of $\mathbb{R}^n$ and $\mathbb{R}$, on which differentiability makes sense.

The map $\bar{f}_a$ is continuous, being a composition of continuous maps; moreover, the atlas allows us to define f on a neighbourhood of P, so that ϕ_a is well defined on a neighbourhood of $(x_a^1, \dots, x_a^n)$, since ϕ_a respects the topology (again a consequence of working with an atlas). Thus our map $\bar{f}_a$ is in a position to be differentiable.

Definition 13.45 (Differentiable map on a topological space)**.** We say that f is differentiable at $P = \phi_a(x)$ if and only if

1. f is continuous;
2. $\bar{f}_a = f \circ \phi_a$ is differentiable.

Rests on Definitions 13.43 and 13.44.

But what would happen if we changed to a coordinate system (ϕ_b, U_b) and $\bar{f}_b = f \circ \phi_b$ were not differentiable at P? This does not occur, since

$$\begin{aligned}
\bar{f}_b &= f \circ \phi_b \\
&= f \circ \mathrm{id} \circ \phi_b \\
&= f \circ \left(\phi_a \circ \phi_a^{-1}\right) \circ \phi_b \\
&= (f \circ \phi_a) \circ \left(\phi_a^{-1} \circ \phi_b\right) \\
&= \bar{f}_a \circ \phi_{ab}\,,
\end{aligned}$$

and thus, since ϕ_{ab} is smooth, it is at least of class C^1 and hence differentiable; and since $\bar{f}_a$ is differentiable, $\bar{f}_b$ is differentiable, being a composition of differentiable maps. The atlas has again entered the game: it is the smoothness condition in the atlas that allows us to define a differentiable map on M.

Differentiable structure

Consider a topological space now equipped with the atlases $\{(\phi_a, U_a)\}_{a\in I}$ and $\{(\psi_b, V_b)\}_{b\in J}$. The relation

$$\{(\phi_a, U_a)\} \cup \{(\psi_b, V_b)\} \text{ is an atlas} \iff \{(\phi_a, U_a)\} \wedge \{(\psi_b, V_b)\} \text{ are atlases}$$

is an equivalence relation.

Proposition 13.46 (Compatibility of atlases)**.** *On the set of n-dimensional atlases of a topological space M, the relation*

$$\mathcal{A} \sim \mathcal{B} \quad \iff \quad \mathcal{A} \cup \mathcal{B} \text{ is an atlas} \tag{13.138}$$

is an equivalence relation.

Rests on Definitions 13.43 and 13.44.

Proof. *Derives Proposition 13.46.*

First unpack Equation (13.138). The union of two atlases always satisfies the covering condition of Definition 13.44, since $\mathcal{A}$ alone already covers M; and the transition maps between two charts of $\mathcal{A}$, or between two charts of $\mathcal{B}$, are smooth because each family is an atlas. So $\mathcal{A} \sim \mathcal{B}$ holds if and only if every *crossed* transition map $\psi^{-1} \circ \phi$, between a chart $(\phi, U) \in \mathcal{A}$ and a chart $(\psi, V) \in \mathcal{B}$ with overlapping images, is smooth.

Reflexivity. $\mathcal{A} \cup \mathcal{A} = \mathcal{A}$, which is an atlas by hypothesis.

Symmetry. The union is a symmetric operation, so the condition Equation (13.138) is symmetric as it stands. (Concretely: a crossed transition map in one direction is the inverse of the one in the other, $\psi^{-1} \circ \phi = (\phi^{-1} \circ \psi)^{-1}$, and the atlas condition demands smoothness of both.)

Transitivity. Let $\mathcal{A} \sim \mathcal{B}$ and $\mathcal{B} \sim \mathcal{C}$, and take charts $(\phi, U) \in \mathcal{A}$ and $(\chi, W) \in \mathcal{C}$ whose images overlap. Let $P \in \phi(U) \cap \chi(W)$. Since $\mathcal{B}$ is an atlas it covers M, so there is a chart $(\psi, V) \in \mathcal{B}$ with $P \in \psi(V)$. The set

$$\phi^{-1}\left(\phi(U) \cap \psi(V) \cap \chi(W)\right)$$

is open — it is the preimage under the homeomorphism ϕ of an intersection of three sets open in M — and it contains $\phi^{-1}(P)$. On it,

$$\chi^{-1} \circ \phi = \left(\chi^{-1} \circ \psi\right) \circ \left(\psi^{-1} \circ \phi\right) , \tag{13.139}$$

a composition of two smooth maps by the hypotheses $\mathcal{A} \sim \mathcal{B}$ and $\mathcal{B} \sim \mathcal{C}$, hence smooth. Smoothness is a local property and P was an arbitrary point of the overlap, so $\chi^{-1} \circ \phi$ is smooth on the whole of $\phi^{-1}(\phi(U) \cap \chi(W))$; the same argument with the roles of ϕ and χ exchanged gives the smoothness of $\phi^{-1} \circ \chi$. Every crossed transition map of $\mathcal{A} \cup \mathcal{C}$ is therefore smooth, and $\mathcal{A} \cup \mathcal{C}$ is an atlas, that is $\mathcal{A} \sim \mathcal{C}$.

Note where each hypothesis was used: reflexivity and symmetry are formal, while transitivity rests on the covering property of the intermediate atlas $\mathcal{B}$ — without a chart of $\mathcal{B}$ around P the factorisation Equation (13.139) would not exist. □

Definition 13.47 (Differentiable structure)**.** We say that an equivalence class of atlases is a *differentiable structure*. Two atlases belonging to the same differentiable structure are called *compatible*.

Rests on Definition 13.44.

Note the following: all the coordinate systems of two compatible atlases are smoothly related, because their union is an atlas (and the coordinate systems of an atlas are smoothly related).

Differentiable manifold

Definition 13.48 (Differentiable manifold)**.** We say that a topological space M is an *n-dimensional differentiable manifold* if and only if M is equipped with an (n-dimensional) differentiable structure. When it is necessary to specify the dimension of the manifold, we denote it M_n.

Rests on Definition 13.47.

It is important to discuss the following. The dimension of the manifold refers to the number of coordinates we need to describe completely an arbitrary point of it. The definitions have been constructed in such a way that it is not possible to have an n-dimensional manifold in which some points are described by fewer than n coordinates.

Diffeomorphism

The notion of a differentiable map developed above compares a manifold with $\mathbb{R}$; nothing prevents us from comparing two manifolds with each other.

Definition 13.49 (Smooth map between manifolds)**.** Let M_m and N_n be differentiable manifolds. We say that a map $f : M_m \longrightarrow N_n$ is *smooth* if and only if it is continuous and, for every coordinate system (ϕ_a, U_a) of M_m and every coordinate system (ψ_b, V_b) of N_n, the *coordinate representation*

$$\psi_b^{-1} \circ f \circ \phi_a \,, \tag{13.140}$$

which maps a subset of $\mathbb{R}^m$ into $\mathbb{R}^n$, is smooth wherever it is defined.

Rests on Definitions 13.44, 13.45 and 13.48.

That the condition need be checked in one pair of charts only is the computation already performed for $\bar{f}_a$ and $\bar{f}_b$ above: changing either chart composes Equation (13.140) with a transition map, which is smooth because we are working inside atlases. Definition 13.45 is the special case $N_n = E_1$ read in the Cartesian chart.

Definition 13.50 (Diffeomorphism)**.** A *diffeomorphism* is a smooth homeomorphism $f : M_m \longrightarrow N_n$ whose inverse f^{-1} is also smooth. Two manifolds related by a diffeomorphism are said to be *diffeomorphic*.

Rests on Definitions 6.7 and 13.49.

Remark 13.51. Diffeomorphic manifolds are, for every purpose of this chapter, the same manifold. Applying the chain rule to $f^{-1} \circ f = \mathrm{id}$ shows that the Jacobian of Equation (13.140) is invertible at every point, so $m = n$; and then $(f \circ \phi_a, U_a)$ is a coordinate system of N_n whenever (ϕ_a, U_a) is one of M_n, so f carries atlases to atlases and the differentiable structure of M_n (Definition 13.47) to that of N_n. Every statement expressible through charts, smooth maps and tangent spaces alone therefore transfers. Diffeomorphism is to differential geometry what homeomorphism (Definition 6.7) is to topology.

Smoothness of the inverse must be demanded, for it does not follow: the map $f : \mathbb{R} \longrightarrow \mathbb{R}$ with $f(x) = x^3$ is a smooth homeomorphism, but $f^{-1}(y) = y^{1/3}$ is not differentiable at the origin, so f is not a diffeomorphism — precisely as continuity of a bijection fails to give a homeomorphism (Remark 6.8).

The differential of a smooth map

Definition 13.52 (Differential; pushforward)**.** Let $f : M_m \longrightarrow N_n$ be smooth and let $P \in M_m$. The *differential* of f at P is the map

$$f_{*P} : T_M(P) \longrightarrow T_N(f(P))$$

that assigns to the vector tangent at P to a curve $\lambda \longmapsto x(\lambda)$ (Definition 13.81) the vector tangent at $f(P)$ to the image curve $\lambda \longmapsto f(x(\lambda))$. Writing $y^\alpha = f^\alpha(x^1, \dots, x^m)$ for the coordinate representation Equation (13.140), the chain rule gives, in the coordinate bases $\{\partial/\partial x^i\}$ and $\{\partial/\partial y^\alpha\}$,

$$(f_{*P}V)^\alpha = \left.\frac{\partial f^\alpha}{\partial x^i}\right|_P V^i \,, \tag{13.141}$$

so f_{*P} is linear and is represented by the Jacobian matrix of f at P. It is the pushforward used below in the discussion of pullbacks (Section 13.5.1).

Rests on Definitions 13.49 and 13.81.

Immersion, embedding, and submanifold

Definition 13.53 (Immersion, submersion, embedding)**.** Let $f : M_m \longrightarrow N_n$ be smooth. We say that f is

1. an *immersion* if and only if f_{*P} is injective for every $P \in M_m$, which requires $m \le n$;
2. a *submersion* if and only if f_{*P} is surjective for every $P \in M_m$, which requires $m \ge n$;
3. an *embedding* if and only if it is an injective immersion which is, in addition, a homeomorphism onto its image $f(M_m)$, that image carrying the topology it inherits as a subset of N_n.

Rests on Definitions 6.7 and 13.52.

Definition 13.54 (Embedded submanifold)**.** A subset $S \subseteq N_n$ is an *embedded submanifold of dimension* m if and only if S, equipped with the topology inherited from N_n, admits an m-dimensional differentiable structure for which the inclusion $\iota : S \longrightarrow N_n$ is an embedding. The integer $n - m$ is the *codimension* of S in N_n.

Rests on Definitions 13.48 and 13.53.

The image of an embedding is an embedded submanifold: transporting the differentiable structure of M_m through the homeomorphism $f : M_m \longrightarrow f(M_m)$ produces an atlas on $f(M_m)$ whose charts are the $f \circ \phi_a$, and with respect to it f becomes a diffeomorphism onto its image. The third condition in Definition 13.53 is not implied by the first two, as the following example shows.

Example 13.55 (An injective immersion that is not an embedding)**.** Consider the *figure-eight* curve in E_3,

$$\gamma :]-\pi, \pi[\longrightarrow \mathbb{R}^3 \,, \qquad x^i(\gamma(\lambda)) = (\sin 2\lambda,\ \sin\lambda,\ 0) \,. \tag{13.142}$$

It is an immersion: $\frac{\mathrm{d}x^i}{\mathrm{d}\lambda} = (2\cos 2\lambda, \cos\lambda, 0)$ never vanishes, because $\cos\lambda = 0$ forces $\lambda = \pm\pi/2$ and there $2\cos 2\lambda = -2$. It is injective: let $\lambda_1, \lambda_2 \in]-\pi, \pi[$ have $\gamma(\lambda_1) = \gamma(\lambda_2)$. Equality of the second components, $\sin\lambda_1 = \sin\lambda_2$, leaves besides $\lambda_2 = \lambda_1$ exactly two possibilities inside the interval, $\lambda_2 = \pi - \lambda_1$ (which lies in $]-\pi, \pi[$ only for $\lambda_1 > 0$) and $\lambda_2 = -\pi - \lambda_1$ (only for $\lambda_1 < 0$); both branches must be checked, since the second is the one available on the negative half of the interval. On either of them $\sin 2\lambda_2 = -\sin 2\lambda_1$, because sin has period 2π and $2\lambda_2 = \pm 2\pi - 2\lambda_1$. Equality of the first components then gives $\sin 2\lambda_1 = 0$, that is $\lambda_1 \in \{0, \pm\pi/2\}$; and $\lambda_1 = 0$ sends both branches out of the interval, while $\lambda_1 = \pi/2$ and $\lambda_1 = -\pi/2$ send the available branch back to $\lambda_2 = \lambda_1$. Hence $\lambda_1 = \lambda_2$ in every case.

It is not an embedding. As $\lambda \to \pm\pi$ both components tend to 0, so the image accumulates at $\gamma(0) = (0, 0, 0)$: every neighbourhood of $\gamma(0)$ in the subspace topology of the image contains points $\gamma(\lambda)$ with λ arbitrarily close to $\pm\pi$, whose parameters are *not* close to 0.

Hence γ^{-1} is discontinuous at $\gamma(0)$, and the image, which is a self-touching figure eight, is not an embedded submanifold of E_3: no neighbourhood of the origin in it is homeomorphic to an interval.

Rests on Definitions 13.52 and 13.53.

Product of manifolds

Proposition 13.56 (Product manifold)**.** *Let M_m and N_n be differentiable manifolds with atlases $\{(\phi_a, U_a)\}_{a\in I}$ and $\{(\psi_b, V_b)\}_{b\in J}$. Then the topological product $M_m \times N_n$, carrying the product topology, is a differentiable manifold with the* product atlas

$$\{(\phi_a \times \psi_b,\ U_a \times V_b)\}_{(a,b)\in I\times J}\ , \qquad (\phi_a \times \psi_b)\,(x,y) = (\phi_a(x), \psi_b(y))\ , \tag{13.143}$$

and

$$\dim\,(M_m \times N_n) = \dim M_m + \dim N_n = m + n\,. \tag{13.144}$$

Rests on Definitions 13.43, 13.44 and 13.48.

Proof. *Derives Proposition 13.56.*

Each $U_a \times V_b$ is open in $\mathbb{R}^m \times \mathbb{R}^n = \mathbb{R}^{m+n}$, and $\phi_a \times \psi_b$ is a homeomorphism onto its image for the product topology, being a product of homeomorphisms; it is smooth because each factor is, so Definition 13.43 is met and each pair is an $(m+n)$-dimensional coordinate system. The images cover, since $\phi_a(U_a)$ cover M_m and $\psi_b(V_b)$ cover N_n. Finally, the transition map between two such charts is

$$(\phi_{a'} \times \psi_{b'})^{-1} \circ (\phi_a \times \psi_b) = (\phi_{a'a},\ \psi_{b'b})\ ,$$

acting on the two blocks of coordinates separately; it is smooth because the transition maps of the two given atlases are. Both conditions of Definition 13.44 hold, so Equation (13.143) is an atlas, and the number of coordinates needed to locate a point, $(x^1, \ldots, x^m, y^1, \ldots, y^n)$, is $m+n$, which is Equation (13.144). □

Example 13.57 (The cylinder)**.** Take $M = S_1$, the circle of radius R with the two-chart atlas of Section 13.4.3, and $N = \mathbb{R}$ with the single Cartesian chart. The index sets of Equation (13.143) then have two and one element, so the product atlas has $2 \times 1 = 2$ charts, both of the form $(\phi, z) \longmapsto (R\cos\phi, R\sin\phi, z)$ on the corresponding angular range, and it makes $S_1 \times \mathbb{R}$ a two-dimensional manifold, the cylinder of radius R in E_3. Its dimension, $1 + 1 = 2$, is Equation (13.144); it is the surface whose vanishing Gaussian curvature (Section 13.3.3) makes it *locally* isometric to the plane. Globally it is not: a flat strip is simply connected and the cylinder is not, so no isometry of the two exists — the cylinder is a strip with its two edges identified, not a subset of the plane.

Rests on Proposition 13.56.

Hypersurfaces

Definition 13.58 (Hypersurface)**.** A *hypersurface* of N_n is an embedded submanifold of codimension one, that is, of dimension $n - 1$.

Rests on Definition 13.54.

Hypersurfaces are produced in practice by one scalar equation, and the theorem that licenses this is the following.

Theorem 13.59 (Regular value theorem)**.** *Let $F : N_n \longrightarrow \mathbb{R}$ be smooth and let $c \in \mathbb{R}$ be a* regular value *of F: the differential F_{*P} is nonzero at every P with $F(P) = c$, that is, in any chart around such a P at least one of the $\partial F/\partial x^i$ is nonzero there. Then the level set*

$$S = F^{-1}(c) = \{P \in N_n \mid F(P) = c\}\ , \tag{13.145}$$

if non-empty, is a hypersurface of N_n.

Rests on Definitions 13.49, 13.52 and 13.58.

Proof. *Derives Theorem 13.59.*

Let $P \in S$ and choose a chart of N_n around P with coordinates $(x^1, \dots, x^n)$. Relabelling the coordinates we may assume $\partial F/\partial x^n \neq 0$ at P. By the implicit function theorem of multivariable calculus there are then an open set $W \subset \mathbb{R}^{n-1}$ containing $(x^1, \dots, x^{n-1})(P)$, an open interval I containing $x^n(P)$, and a smooth function $h : W \longrightarrow I$ such that, inside $W \times I$,

$$F(x^1, \dots, x^n) = c \quad \Longleftrightarrow \quad x^n = h(x^1, \dots, x^{n-1})\,. \tag{13.146}$$

Thus S meets the coordinate box $W \times I$ in the graph of h, and

$$\chi : W \longrightarrow S\,, \qquad \chi(u^1, \dots, u^{n-1}) = \Big(u^1, \dots, u^{n-1}, h(u)\Big)$$

is smooth and injective; its inverse is the restriction to S of the projection onto the first $n-1$ coordinates, which is continuous, so χ is a homeomorphism onto an open subset of S in the subspace topology. The family of all such χ, one around each point of S, covers S. Two of them are smoothly related, because the transition map is a composition of one such χ with a projection and with a transition map of N_n, all smooth. Hence Definition 13.44 is met and S is an $(n-1)$-dimensional manifold whose inclusion into N_n reads, in these charts, $u \longmapsto (u, h(u))$: an injective immersion — its Jacobian $(\mathbb{1}, \partial h/\partial u)$ has rank $n-1$ — and a homeomorphism onto its image by construction. By Definitions 13.53 and 13.58, S is a hypersurface. □

Remark 13.60. The single input of the proof above is the implicit function theorem. Chapter 7 develops differentiability in several variables but stops short of stating it, so it is stated and proved in the appendix — by an explicit contraction, in the manner of the Picard iteration of Theorem 9.8, so that no fixed-point theorem is imported. The form used here is the scalar one: where $\partial F/\partial x^n \neq 0$, the equation $F = c$ determines x^n as a function of the remaining coordinates, of the same differentiability class as F, which is exactly the assertion Equation (13.146) appealed to above.

Full derivation: Appendix A.28 (p. 2914).
Derives Equation (13.146).

Example 13.61 (The 2-sphere in $\mathbb{R}^3$)**.** Take $N_3 = E_3$ with Cartesian coordinates (x, y, z) and

$$F(x, y, z) = x^2 + y^2 + z^2\,, \qquad c = R^2 > 0\,.$$

On the level set the differential $\partial F/\partial x^i = 2(x, y, z)$ is nonzero, since the origin does not satisfy $F = R^2$; hence R^2 is a regular value and the sphere

$$S_2 = \Big\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = R^2\Big\} \tag{13.147}$$

is a hypersurface of E_3, of dimension 2 and codimension 1. The charts furnished by Theorem 13.59 are here explicit: the six open hemispheres, on each of which one coordinate is a smooth function of the other two, for instance $z = +\sqrt{R^2 - x^2 - y^2}$ on the northern one. No single chart suffices, by the compactness argument used above for the circle: S_2 is closed and bounded, hence compact (Definition 6.9), while a chart domain must be open in $\mathbb{R}^2$.

The induced metric (Definition 13.119) of S_2 is the first fundamental form of Section 13.3.2; in the spherical parametrisation it is $\mathrm{d}s^2 = R^2\mathrm{d}\theta^2 + R^2 \sin^2\theta\, \mathrm{d}\phi^2$, whose Gaussian curvature was computed to be $K = 1/R^2$ in Example 13.33.

Rests on Theorem 13.59 and Definition 13.58.

Higher codimension and maps of constant rank

Theorem 13.59 cuts out a submanifold with *one* equation. Physics almost never presents one equation: a constraint surface in phase space, the level set of a momentum map, the shell of a conserved four-vector are all defined by several. The extension is not automatic — "at least one partial derivative is nonzero" has to become a rank condition — and the general statement is the following.

Theorem 13.62 (Regular value theorem in codimension k)**.** *Let $F : N_n \longrightarrow \mathbb{R}^k$ be smooth with $1 \le k \le n$, and let $c \in \mathbb{R}^k$ be a* regular value*: at every P with $F(P) = c$ the differential $F_{*P} : T_P N_n \longrightarrow \mathbb{R}^k$ is surjective, equivalently the k differentials $\mathrm{d}F^1, \dots, \mathrm{d}F^k$ are linearly independent at P. Then*

$$S = F^{-1}(c) = \{P \in N_n \mid F(P) = c\} \,, \tag{13.148}$$

if non-empty, is an embedded submanifold of N_n of dimension $n - k$, and at each of its points

$$T_P S = \ker F_{*P} \,. \tag{13.149}$$

For $k = 1$ this is Theorem 13.59.

Rests on Theorem A.286 and Definitions 13.52 and 13.54.

Derivation. Derives Theorem 13.62.

Let $P \in S$ and take a chart around P with coordinates $(x^1, \dots, x^n)$, in which F becomes a smooth map of an open set of $\mathbb{R}^n$ into $\mathbb{R}^k$. Surjectivity of F_{*P} says that the $k \times n$ Jacobian $\partial F^i / \partial x^j$ has rank k at P, so some k of its columns are independent; relabelling the coordinates, take them to be the last k, and split $x = (x', x'')$ with $x' \in \mathbb{R}^{n-k}$ and $x'' \in \mathbb{R}^k$. Then $\det [D_{x''} F(P)] \neq 0$, which is the hypothesis Equation (A.523) of Theorem A.286 for the map $F - c$. That theorem supplies open sets $W \subseteq \mathbb{R}^{n-k}$ and $V \subseteq \mathbb{R}^k$ around the two blocks of P and a smooth $h : W \longrightarrow V$ with

$$\{(x', x'') \in W \times V \mid F(x', x'') = c\} = \{(x', h(x')) \mid x' \in W\} \,. \tag{13.150}$$

From here the argument of Theorem 13.59 runs verbatim with $n - 1$ replaced by $n - k$: S meets the coordinate box $W \times V$ in the graph of h; the map $\chi(u) = (u, h(u))$ is smooth, injective, and a homeomorphism onto its image because its inverse is the restriction of the projection onto the first $n - k$ coordinates; two such charts are smoothly related; and the inclusion has Jacobian $(\mathbb{1}, \partial h / \partial u)$ of rank $n - k$, so it is an injective immersion and a homeomorphism onto its image. By Definitions 13.53 and 13.54, S is an embedded submanifold of dimension $n - k$.

For Equation (13.149): if γ is a curve in S through P then $F \circ \gamma \equiv c$, so $F_{*P} \dot\gamma = 0$ and $T_P S \subseteq \ker F_{*P}$. Both spaces have dimension $n - k$ — the first by the paragraph above, the second by the rank–nullity relation Equation (5.68) applied to the surjection F_{*P} — so they coincide. □

The rank of F_* need not be maximal, and when it is merely *constant* the map still has a normal form. This is the statement that turns "the image of a degenerate Legendre map" or "the set swept out by a family of solutions" into a submanifold with a definite dimension, and it is proved from the same implicit function theorem.

Theorem 13.63 (Constant rank theorem)**.** *Let $U \subseteq \mathbb{R}^N$ be open and let $f : U \longrightarrow \mathbb{R}^{N'}$ be of class C^k, $k \ge 1$, with*

$$\operatorname{rank} Df(x) = \rho \qquad \text{for every } x \in U \,. \tag{13.151}$$

Then about each $p \in U$ and its image $f(p)$ there are C^k changes of coordinates — diffeomorphisms of a neighbourhood of p in $\mathbb{R}^N$ and of a neighbourhood of $f(p)$ in $\mathbb{R}^{N'}$ — in which f reads

$$\left(u^1, \dots, u^N\right) \longmapsto \left(u^1, \dots, u^\rho, 0, \dots, 0\right) . \tag{13.152}$$

Rests on Corollary A.292, Theorem A.286, and Definition 7.66.

Derivation. Derives Theorem 13.63.

Translating source and target we may take $p = 0$ and $f(0) = 0$. Since $Df(0)$ has rank ρ, some $\rho \times \rho$ minor of it is nonsingular; permuting the coordinates of $\mathbb{R}^N$ and of $\mathbb{R}^{N'}$ — themselves linear diffeomorphisms — we may assume it is the leading one. Split $x = (x', x'') \in \mathbb{R}^\rho \times \mathbb{R}^{N-\rho}$ and $y = (y', y'') \in \mathbb{R}^\rho \times \mathbb{R}^{N'-\rho}$, and write $f = (f', f'')$ accordingly, so that $\det [D_{x'} f'(0)] \neq 0$.

Straightening the source. Put

$$\varphi(x', x'') = \left(f'(x', x''),\, x''\right) , \tag{13.153}$$

a C^k map of U into $\mathbb{R}^N$ whose Jacobian at the origin is the block matrix with rows $(D_{x'} f', D_{x''} f')$ and $(0, \mathbb{1})$, of determinant $\det [D_{x'} f'(0)] \neq 0$. By the inverse function theorem (Corollary A.292) φ restricts to a C^k diffeomorphism of a neighbourhood of 0 onto an open set of $\mathbb{R}^N$, inside which we choose a product of open balls $W = W' \times W''$ containing 0; write $\psi = \varphi^{-1}$ on W. By construction the first ρ components of φ are f', so

$$(f \circ \psi)(u', u'') = \left(u',\, g(u', u'')\right) , \qquad (u', u'') \in W , \tag{13.154}$$

with $g = f'' \circ \psi$ of class C^k.

The rank forces g to forget u''. Composing with a diffeomorphism does not change the rank of a differential, by the chain rule Equation (7.52) and the invertibility of $D\psi$; so $D(f \circ \psi)$ has rank ρ throughout W. Its block form, read off Equation (13.154), has rows $(\mathbb{1}, 0)$ and $(D_{u'} g, D_{u''} g)$. The first ρ rows are already independent, so rank ρ forces every row of $D_{u''} g$ to vanish: $D_{u''} g \equiv 0$ on W. Since W'' is a ball, hence connected, $g(u', u'') = g(u')$ depends on u' alone.

Straightening the target. On the open set $W' \times \mathbb{R}^{N'-\rho}$ define

$$\chi(y', y'') = \left(y',\, y'' - g(y')\right) , \tag{13.155}$$

a C^k diffeomorphism onto $W' \times \mathbb{R}^{N'-\rho}$, its inverse being $(z', z'') \mapsto (z', z'' + g(z'))$. It is defined on the image of Equation (13.154), whose first block lies in W', and

$$(\chi \circ f \circ \psi)(u', u'') = \left(u',\, g(u') - g(u')\right) = (u', 0) ,$$

which is Equation (13.152). □

Corollary 13.64 (The image of a constant-rank map, locally)**.** *Under the hypotheses of Theorem 13.63, every $p \in U$ has a neighbourhood U_0 such that $f(U_0)$ is an embedded submanifold of $\mathbb{R}^{N'}$ of dimension ρ, cut out near each of its points by exactly $N' - \rho$ functions with linearly independent differentials.*

Rests on Theorems 13.62 and 13.63 and Definition 13.54.

Derivation. Derives Corollary 13.64.

In the coordinates of Theorem 13.63, take $U_0 = \psi(W)$. Its image is $\{(u', 0) \mid u' \in W'\}$, which is the level set of the last $N' - \rho$ coordinate functions $z^{\rho+1}, \dots, z^{N'}$ at the value 0; those are $N' - \rho$ functions whose differentials $\mathrm{d}z^{\rho+1}, \dots, \mathrm{d}z^{N'}$ are independent everywhere. By Theorem 13.62 the level set is an embedded submanifold of dimension $N' - (N' - \rho) = \rho$, and transporting it back by the target diffeomorphism χ^{-1}, itself a diffeomorphism onto an open set, preserves both statements. □

Remark 13.65 (Two words that mean different things here). In this chapter *rank* names the number of indices of a tensor (Definition 13.5). In Theorems 13.62 and 13.63 it means the rank of a linear map (Definition 5.37), namely the dimension of the image of the differential. The two usages are standard and both are kept; which is meant is fixed by the object carrying the word. Note also that Theorem 13.63 is stated in $\mathbb{R}^N$ because that is where the implicit function theorem lives; it transfers to a smooth map $M_m \longrightarrow N_n$ of constant rank by reading it in charts, since a change of chart is exactly one of the coordinate changes the theorem already allows.

Group actions and quotient manifolds

The other standard way a manifold arises is not by cutting a bigger one with equations but by identifying its points: the configuration space of a rigid body is the set of frames modulo nothing, but the space of *shapes* is a set of configurations modulo rotations, and a reduced phase space is a level set modulo a symmetry group. The question is when the set of orbits is again a manifold, and the answer needs two hypotheses that are easy to state and easy to forget.

Definition 13.66 (Smooth action; free; proper; orbit)**.** Let G be a Lie group (Chapter 14) acting on a manifold M_n by a smooth map $G \times M_n \longrightarrow M_n$, $(g, P) \mapsto g \cdot P$, with $e \cdot P = P$ and $g \cdot (h \cdot P) = (gh) \cdot P$. The *orbit* of P is $G \cdot P = \{g \cdot P \mid g \in G\}$, and the *orbit space* M_n/G is the set of orbits with the quotient topology. The action is

1. *free* iff $g \cdot P = P$ for some P forces $g = e$;
2. *proper* iff the map $G \times M_n \longrightarrow M_n \times M_n$, $(g, P) \mapsto (g \cdot P, P)$, is proper, i.e. the preimage of every compact set is compact.

Rests on Definitions 6.9, 13.48 and 13.49.

Theorem 13.67 (Quotient manifold theorem)**.** *Let a Lie group G of dimension d act smoothly, freely and properly on M_n. Then the orbit space M_n/G carries exactly one smooth structure, of dimension $n - d$, for which the projection*

$$\pi : M_n \longrightarrow M_n/G\,, \qquad \pi(P) = G \cdot P\,, \tag{13.156}$$

is a smooth submersion (Definition 13.53); with it, M_n/G is Hausdorff and second countable. Every orbit is an embedded submanifold of M_n diffeomorphic to G, and

$$\dim\,(M_n/G) = \dim M_n - \dim G = n - d\,. \tag{13.157}$$

A map $f : M_n/G \longrightarrow N$ is smooth if and only if $f \circ \pi$ is.
Rests on Definitions 13.54 and 13.66 and Theorem 13.63.

Full derivation: Appendix A.46 (p. 3032).
Derives Theorem 13.67.

Example 13.68 (Why each hypothesis is there)**.** Let $G = \mathbb{R}$ act on $M = \mathbb{R}^2$ by $t \cdot (x, y) = (x + t,\, y)$. The action is free, and it is proper because $(t, x, y) \mapsto ((x + t, y), (x, y))$ has continuous inverse $((x', y'), (x, y)) \mapsto (x' - x, x, y)$ on its image, so the preimage of a compact set is a closed subset of a compact one. The orbits are the horizontal lines, and Equation (13.157) gives $\dim(M/G) = 2 - 1 = 1$: the orbit space is the y-axis, as it should be.

Freeness is not decoration: let $\mathrm{SO}(2)$ act on $\mathbb{R}^2$ by rotation. The action is proper — the group is compact — but the origin is fixed, so it is not free, and the orbit space is the half-line $[0, \infty[$, which is not a manifold at its endpoint. Properness is not decoration either:

let $\mathbb{R}$ act on the torus $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ by $t \cdot [x, y] = [x + t,\ y + \kappa t]$ with κ irrational. This action is free, but every orbit is dense in T^2, because the successive returns of the line to a fixed meridian occur at the points $\{n\kappa \bmod 1\}$, which are distinct for distinct n and therefore accumulate. Hence the quotient topology on the orbit space is indiscrete: two distinct orbits have no disjoint neighbourhoods and the space is not Hausdorff, hence not a manifold. The failure is exactly the failure of properness, the group $\mathbb{R}$ escaping to infinity while its points stay in a compact set.

Rests on Theorem 13.67 and Definition 13.66.

13.4.2 Euclidean space

The Cartesian coordinate system

Consider the topological space $M = \mathbb{R}^n$. Note that the pair $(\mathrm{id}_{\mathbb{R}^n}, \mathbb{R}^n)$ defines a coordinate system, for $\mathrm{id}_{\mathbb{R}^n}$ is trivially a smooth homeomorphism: it is the bijective map that associates with each point $x \in \mathbb{R}^n$ the point $x \in \mathbb{R}^n$ itself.

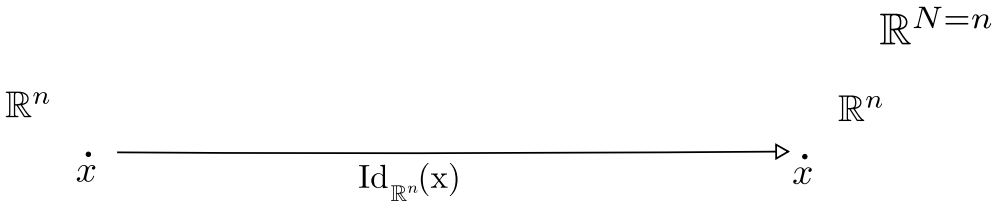


Figure 13.7: The identity chart on $\mathbb{R}^n$: each point is its own coordinate label.

We say that $(\mathrm{id}_{\mathbb{R}^n}, \mathbb{R}^n)$ is a *Cartesian* coordinate system of $\mathbb{R}^n$. In this coordinate system the coordinates of a point $P \in \mathbb{R}^n$ are $x^i(P)$, which are called the *Cartesian coordinates* of the point P (with respect to the coordinate system $(\mathrm{id}_{\mathbb{R}^n}, \mathbb{R}^n)$).

Euclidean space E_n

In the preceding context, a single Cartesian coordinate system suffices to cover all of $\mathbb{R}^n$, so $\{(\mathrm{id}_{\mathbb{R}^n}, \mathbb{R}^n)\}$ defines an n-dimensional atlas and, trivially, a differentiable structure. We conclude that $\mathbb{R}^n$ is an n-dimensional differentiable manifold, which we denote E_n and call *Euclidean space* (or the Euclidean manifold); we also say that spaces of this type possess a *Euclidean geometry*. The fact that a single coordinate system describes the whole Euclidean space leads us to consider describing a manifold, such as a curve or a two-dimensional surface, as a submanifold of E_n.

We say that the *origin* of E_n is the point $O = \mathrm{id}_{\mathbb{R}^n}(0, \ldots, 0)$ (with respect to the coordinate system $(\mathrm{id}_{\mathbb{R}^n}, \mathbb{R}^n)$); therefore $x^i(O) = 0$ for all i.

Pictorial representation

We represent pictorially a Cartesian coordinate system and a point $P = \mathrm{id}_{\mathbb{R}^n}(x) \in E_n$ of coordinates $x^i(P)$ as a diagram of perpendicular axes[2]. For the case $n = 2$ it has been historically common to denote the coordinates of a generic point, x^i, as (x, y):

where E_2 is called *flat two-dimensional space*. For the case $n = 3$ it has been historically common to denote the coordinates of a generic point, x^i, as (x, y, z):

where E_3 receives the name *flat three-dimensional space*.

[2]That the axes are perpendicular has a mathematical background studied further below. This representation is due to René Descartes.

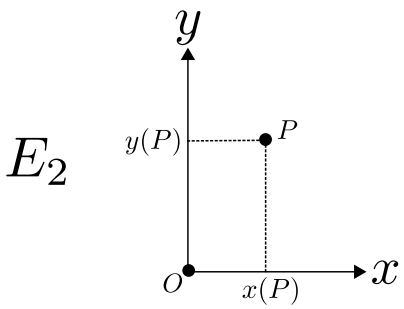


Figure 13.8: The plane E_2 with perpendicular Cartesian axes (x, y).

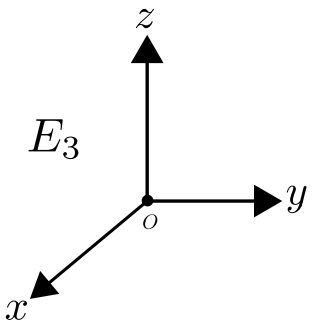


Figure 13.9: The space E_3 with Cartesian axes (x, y, z).

For the case $n = 2$, the Euclidean space E_2 works very well for describing plane geometric figures, for example a square: given a square in the plane, we can specify each of its points by its coordinates. The same occurs with solid geometric figures in the space E_3.

Vectors in E_n

Definition 13.69 (Vector in E_n)**.** Let $P, Q \in E_n$ be two points. We define a *vector* in E_n[3] as a straight-line segment joining P to Q, endowed with the orientation from P towards Q; we symbolise it $\overrightarrow{PQ}$ and represent it pictorially as an arrow pointing from P to Q.

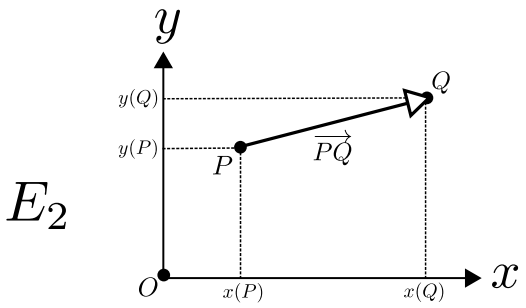


Figure 13.10: A vector in E_n: an oriented straight-line segment $\overrightarrow{PQ}$.

Definition 13.70 (Position vector)**.** We define the *position vector* of a point P with respect to the origin O as the vector $\overrightarrow{OP}$. We introduce the notation $\overrightarrow{OP} \overset{\text{not.}}{=} \boldsymbol{x}_P$.

To refer to the position vector of arbitrary points of E_n we make no reference to the point, denoting it simply $\boldsymbol{x}$. The position vector of the origin is denoted $\mathbf{0}$.

[3] A vector in E_n must not be confused with a vector in the context of vector spaces.

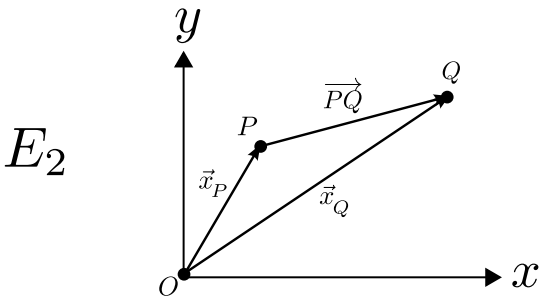


Figure 13.11: The position vector $\boldsymbol{x}_P = \overrightarrow{OP}$ of a point P with respect to the origin O.

Cartesian coordinates as components of the position vector

The fact that a single coordinate system describes the whole Euclidean space provides us with the bijective maps $\boldsymbol{x} : P \longmapsto \boldsymbol{x}_P$ and $x^i : P \longmapsto x^i(P)$; that is, the map assigning to each point of E_n its position vector (with respect to O), and the map assigning to each point of E_n its coordinates. A point of E_n is determined by its position vector or by its Cartesian coordinates, so we may establish a relation between the two concepts.

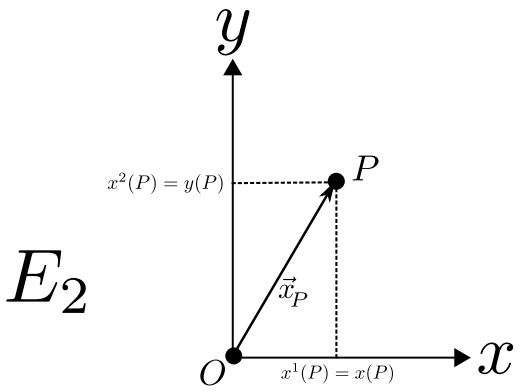


Figure 13.12: The Cartesian coordinates of P are the components of its position vector: $(\boldsymbol{x}_P)^i = x^i(P)$.

We say that the *Cartesian components* of the position vector of a point P with respect to O are the Cartesian coordinates of P (with respect to the coordinate system $(\mathrm{id}_{\mathbb{R}^n}, \mathbb{R}^n)$), which we symbolise as

$$(\boldsymbol{x}_P)^i = x^i(P)\,, \quad i = 1, \ldots, n \tag{13.158}$$

(for a generic point we write $(\boldsymbol{x})^i = x^i$). Note that this relation is one-to-one: given a coordinate system and an origin, the components of the position vector of a point are unique. With respect to the maps just mentioned we have, moreover, $O \longmapsto \mathbf{0}$ and $O \longmapsto (0, \ldots, 0)$ respectively. In this sense, giving a coordinate system amounts to giving an origin (and vice versa) with respect to which we can locate points in E_n.

E_n as a vector space

As we know, the set of all points of $\mathbb{R}^n$ forms an n-dimensional vector space. Moreover, each of these points gives the Cartesian components of the position vector of the corresponding point of Euclidean space E_n. We may therefore picture the vector space $\mathbb{R}^n$ as follows:

Since the relation between the position vector of P (with respect to a given origin) and the coordinates of P (with respect to a given coordinate system) is one-to-one, from now on, given a Cartesian coordinate system (and hence an origin), we refer to the points of E_n by

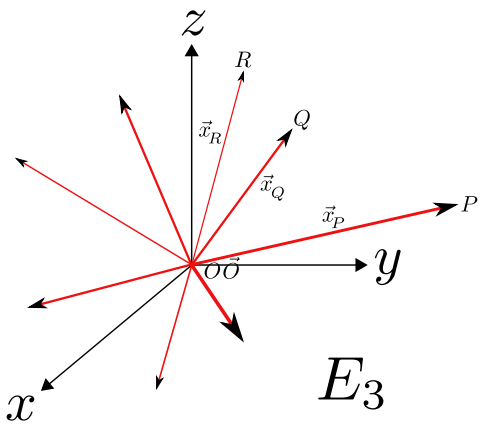


Figure 13.13: $\mathbb{R}^n$ pictured as the space of position vectors of E_n: each point doubles as the arrow from the origin to it.

means of position vectors. This is only a notation to make the equations more transparent. Since we can now conceive the manifold E_n as a vector space — that is, position vectors as vectors of a vector space — we can define the *Euclidean (or standard) inner product* on E_n as

$$\boldsymbol{x}_P \cdot \boldsymbol{x}_Q = \sum_{i=1}^{n} x^i(P)\, x^i(Q)\,, \tag{13.159}$$

which clearly satisfies the axioms of an inner product. We further define the *Euclidean (or standard) norm* of E_n as the norm associated with the Euclidean inner product,

$$|\boldsymbol{x}_P| = \sqrt{\boldsymbol{x}_P \cdot \boldsymbol{x}_P}\,, \tag{13.160}$$

and, lastly, the *Euclidean (or standard) metric* of E_n as the metric associated with the Euclidean norm,

$$d\,(\boldsymbol{x}_P, \boldsymbol{x}_Q) = |\boldsymbol{x}_P - \boldsymbol{x}_Q|\ . \tag{13.161}$$

Displacements

Consider two points P and Q in Euclidean space, with position vectors $\boldsymbol{x}_P$ and $\boldsymbol{x}_Q$ with respect to a given origin. In this context it is common to symbolise the vector $\overrightarrow{PQ}$ as

$$\overrightarrow{PQ} \overset{\text{not.}}{=} \delta\boldsymbol{x}_{PQ}\,, \tag{13.162}$$

or, for two arbitrary points, simply as $\delta\boldsymbol{x}$, which is called the *displacement vector*. It serves to denote the change of position between two points.

In the preceding context, one then analyses the case in which the Euclidean distance between these points is infinitesimally small.

13.4.3 Curves in Euclidean space

Differentiable curves

Definition 13.71 (Differentiable curve)**.** We say that a *differentiable curve*[4] is a one-dimensional differentiable submanifold of E_n, which we denote $\mathcal{C}$.

Rests on Definition 13.48.

[4]More general, topologically founded definitions of a curve exist; they will not be needed in our studies.

We therefore need only one coordinate to describe a curve locally — we say locally because in general more than one coordinate system is needed to describe it globally, though each of them involves a single coordinate. Thus the possible coordinate systems needed to describe the curve are of the form

$$(\gamma,]a,b[\,)\,, \quad \text{where} \quad \begin{array}{rccl} \gamma: &]a,b[& \longrightarrow & \mathcal{C}\subset E_n \\ & \lambda & \longmapsto & P=\gamma(\lambda)\,. \end{array} \tag{13.163}$$

The variable λ is a coordinate associated with the coordinate system $(\gamma,]a,b[\,)$, and $\gamma(\lambda_P)$, with $\lambda_P \in]a,b[$, is a point of the curve. For a generic point of the curve we may identify $\gamma(\lambda)$ with its position vector $\boldsymbol{x}_{\gamma(\lambda)}$. The components of the position vector are therefore the quantities

$$\left(\boldsymbol{x}_{\gamma(\lambda)}\right)^i = x^i(\gamma(\lambda))\,, \tag{13.164}$$

which are the (Cartesian, since the curve lives in E_n) coordinates of the point $\gamma(\lambda) \in \mathcal{C} \subset E_n$. The coordinate λ parametrises the curve: for each value of $\lambda \in \mathbb{R}$ we have a position vector $\boldsymbol{x} = \boldsymbol{x}(\lambda)$ characterising a point of the curve.

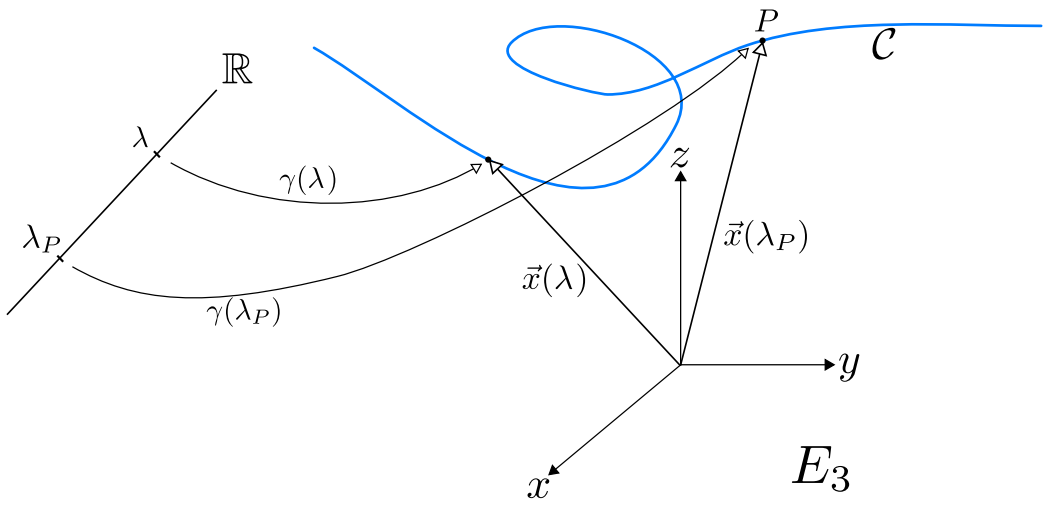


Figure 13.14: A curve in E_n described by a single chart $(\gamma,]a,b[\,)$: the coordinate λ labels the points $\gamma(\lambda)$, whose position vectors are $\boldsymbol{x}(\lambda)$.

It is necessary to clarify the difference between the two statements: λ is the coordinate of the point $\gamma(\lambda) \in \mathcal{C}$, while the $x^i(\gamma(\lambda))$ are the Cartesian coordinates of the point $\gamma(\lambda) \in E_n$. Now, as mentioned before, in general it is not possible to describe a curve completely with a single coordinate system (we need an atlas). The problem of needing more than one coordinate system is not transcendental for calculations (though it does bring topological consequences to be studied in theoretical physics). Let us illustrate with a well-known and much-cited example: the circle of radius R as a hypersurface of E_2, $S_1 = \{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 = R^2\}$. We wish to describe by coordinate systems all points of E_2 satisfying $x^2+y^2=R^2$. A possible candidate is $(\gamma,]0, 2\pi[\,)$, where

$$\begin{array}{rccl} \gamma: &]0,2\pi[& \longrightarrow & \mathcal{C} \\ & \phi & \longmapsto & \gamma(\phi)\,, \end{array}$$

with the Cartesian coordinates of $\gamma(\phi)$ given by

$$x^i(\gamma(\phi)) = (R\cos\phi, R\sin\phi)\ . \tag{13.165}$$

This coordinate system describes almost all of the desired region — almost, because the point $(R,0)$ is not included. If we put the interval $[0,2\pi[$ or $]0,2\pi]$ in the domain of γ we

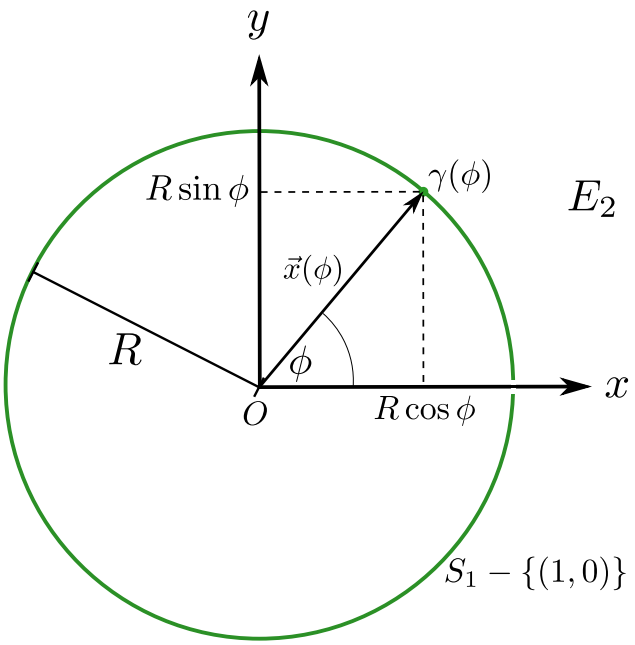


Figure 13.15: The circle $S_1 \subset E_2$ described by the angular chart $\phi \mapsto (R\cos\phi, R\sin\phi)$ on $]0, 2\pi[$; the point $(R, 0)$ escapes this single chart.

would include the point, but in the definition of a differentiable manifold we worked only with open domains, to make sure we respect the topology (the maps of a coordinate system must be homeomorphisms). We could contemplate an interval $]0, 2\pi + \delta[$ so as to include the mentioned point, but then the map γ would no longer be bijective, which as we know is a crucial property of γ as a homeomorphism, meant to preserve the topology. How can we fix this problem? Is there a coordinate system that describes all the points of the circle? The answer is no, and as we shall see this is related to the fact that the circle is a compact subset of E_2. If a single coordinate system (γ, I) described every point of the circle, the homeomorphism $\gamma : I \longrightarrow S_1$ would force I to be compact — a contradiction, since I must necessarily be open.

We define the coordinate systems $(\gamma_1,]-\delta, \pi + \delta[\,)$ and $(\gamma_2,]\pi - \delta, 2\pi + \delta[\,)$ with

$$\begin{array}{rccc} \gamma_1 : &]-\delta, \pi+\delta[& \longrightarrow & E_2 \\ & \phi & \longmapsto & \gamma_1(\phi)\,, \end{array} \qquad \begin{array}{rccc} \gamma_2 : &]\pi-\delta, 2\pi+\delta[& \longrightarrow & E_2 \\ & \phi & \longmapsto & \gamma_2(\phi)\,, \end{array}$$

where $x^i(\gamma_1(\phi)) = (R\cos\phi, R\sin\phi)$ with ϕ varying in $]-\delta, \pi + \delta[$, and $x^i(\gamma_2(\phi)) = (R\cos\phi, R\sin\phi)$ with ϕ varying in $]\pi - \delta, 2\pi + \delta[$. We mention that we have used the same variable ϕ for both coordinate systems because it is a dummy variable and there is no ambiguity (it merely happens that in both coordinate systems the curve behaves locally in the same way). The two coordinate systems cover the whole circle. Thus $\{(\gamma_1,]-\delta, \pi + \delta[\,), (\gamma_2,]\pi - \delta, 2\pi + \delta[\,)\}$ defines an atlas on S_1, which clearly possesses a one-dimensional differentiable structure. This means that S_1 is a hypersurface of E_2.

An example of a non-compact curve that can be described by a single coordinate system is the straight-line segment between two points $P, Q \in E_n$, denoted L_{PQ}. The straight line joining the two points is completely described by the coordinate system $(\gamma,]0, 1[\,)$ with

$$\begin{array}{rccc} \gamma : &]0,1[& \longrightarrow & E_n \\ & \lambda & \longmapsto & \gamma(\lambda)\,, \end{array}$$

where $x^i(\gamma(\lambda)) = (1 - \lambda)x^i(P) + \lambda x^i(Q)$ with λ varying in $]0, 1[$. Visually, the curve "grows" from P to Q as λ grows in $]0, 1[$. Note further that $x^i(\gamma(\lambda \to 0)) = x^i(P)$ and $x^i(\gamma(\lambda \to 1)) = x^i(Q)$.

Remark 13.72. In the example of the circle of radius R we used for both coordinate systems the same parametrisation of the points of the curve, $(R\cos\phi, R\sin\phi)$; this does not mean we are describing the curve locally with only one coordinate system — it only means that the circle behaves locally in the same way in both coordinate systems, which is why we used the same variable. On the other hand, whenever we describe a curve locally from a Euclidean space by means of a coordinate system $(\gamma,]a,b[\,)$, we shall refer to the curve simply as the map $x^i = x^i(\lambda)$, unless it is necessary to make explicit the interval over which λ varies together with the homeomorphism of that coordinate system. In this case, unlike for the circle, we give a parameter to specify that we are describing the curve locally. The stated notation is warranted by the fact that the coordinates of E_n used to describe the curve depend solely on the coordinate used to describe the curve locally.

Admissible change of parameter

Given a curve $\mathcal{C}$, described by $\boldsymbol{x} = \boldsymbol{x}(\lambda)$, we can perform coordinate changes in our coordinate systems of the form seen in Equation (13.163). A general coordinate transformation may take the form $\bar\lambda = \bar\lambda(\lambda)$, and, as we know, the map

$$\begin{array}{rccc} \bar\lambda: &]a,b[& \longrightarrow &]c,d[\\ & \lambda & \longmapsto & \bar\lambda(\lambda) \end{array}$$

must be a smooth function of λ to guarantee that we possess a differentiable structure. The new coordinate system is then of the form $(\bar\gamma,]c,d[\,)$ where

$$\begin{array}{rccc} \bar\gamma: &]c,d[& \longrightarrow & \mathcal{C} \\ & \bar\lambda & \longmapsto & \bar\gamma(\bar\lambda)\,, \end{array}$$

with $\bar\gamma = \gamma \circ \bar\lambda$.

Definition 13.73 (Admissible change of parameter)**.** In the context of the theory of curves, we call the coordinate transformation $\bar\lambda = \bar\lambda(\lambda)$ an *admissible change of parameter* if and only if

1. $\bar\lambda$ is a function of class C^1 $(]a,b[)$, so as to guarantee the differentiable structure;
2. $\forall\,\lambda \in]a,b[$, $\frac{\mathrm{d}\bar\lambda}{\mathrm{d}\lambda}(\lambda) \neq 0$. This ensures the function has no extrema in $]a,b[$, hence possesses an inverse on the given interval, so that $\lambda = \lambda(\bar\lambda)$ is also a coordinate transformation.

It is important to point out that an admissible change of parameter (a general coordinate transformation of the single coordinate describing the curve locally) must not be confused with a general coordinate transformation of the Euclidean space in which the curve is embedded. There is, however, no problem in performing a general coordinate transformation of the Euclidean space in order to describe the curve. As an example of an admissible change of parameter, let the straight-line curve L_{PQ} with $P, Q \in E_2$ be described by

$$x^i(\lambda) = (1-\lambda)x^i(P) + \lambda x^i(Q) = ((1-\lambda)x(P) + \lambda x(Q),\, (1-\lambda)y(P) + \lambda y(Q))\,,$$

with λ varying in $]0,1[$. We may perform the change of parameter $\bar\lambda(\lambda) = 1-\lambda$, which is clearly admissible, with $\bar\lambda$ still varying in $]0,1[$. The straight line is now described by

$$x^i(\bar\lambda) = \bar\lambda x^i(P) + (1-\bar\lambda)x^i(Q) = (1-\bar\lambda)x^i(Q) + \bar\lambda x^i(P)\,.$$

Visually, this change of parameter makes the curve "grow" from Q to P instead.

Tangent vector

Consider a differentiable curve in E_n described locally by $x^i = x^i(\lambda)$, $i = 1, \dots, n$, with respect to $(\gamma,]a, b[\,)$, or equivalently with respect to the basis $\{\hat{x}_i\}$ of E_n as a vector space. We define the *tangent vector* to the curve at the point $P = \gamma(\lambda_P)$ as the vector

$$\boldsymbol{V}_P = \left.\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\right|_{\lambda=\lambda_P} \hat{x}_i \,, \tag{13.166}$$

that is, as the variation of the coordinates of the curve as the parameter varies.

Arc element and length of a curve

Section 13.3.1 built the theory of curves *extrinsically*: the curve was a map into $\mathbb{R}^3$, and its tangent, normal and binormal were vectors of the ambient Euclidean space, differentiated componentwise. The same content can be stated using nothing but the metric of the space the curve lives in and the covariant derivative it determines — the *intrinsic* formulation, which is the one that survives when the ambient space is removed. Throughout this stretch (M, g) is a Riemannian manifold of dimension D (Definition 13.117 with $q = 0$), with coordinates x^μ and Levi-Civita connection $\mathring{\Gamma}$ (Theorem 13.150); the reader who prefers a concrete setting may read $M = E_3$ in arbitrary curvilinear coordinates, in which case $g_{\mu\nu}$ is the Euclidean metric written in those coordinates.

The *arc element* is the line element Equation (13.259) of the displacement along the curve,

$$\mathrm{d}s^2 = g_{\mu\nu}\, \mathrm{d}x^\mu \mathrm{d}x^\nu \,, \tag{13.167}$$

which for a surface in E_3 is exactly the first fundamental form Equation (13.61), the metric g_{ij} there being the induced metric Equation (13.261).

Definition 13.74 (Length of a curve)**.** Let $x^\mu(\lambda)$, $\lambda \in]a, b[$, be a differentiable curve in (M, g). Its *length* is

$$L = \int_a^b \sqrt{g_{\mu\nu} \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda} \frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda}}\; \mathrm{d}\lambda = \int_a^b \left\| \frac{\mathrm{d}x}{\mathrm{d}\lambda} \right\| \mathrm{d}\lambda \,, \tag{13.168}$$

the integrand being the metric norm of the tangent vector Equation (13.177). The curve is *regular* if that norm never vanishes.

Rests on Definitions 13.81, 13.117 and 13.118.

Proposition 13.75 (Invariance of the length)**.** *The length Equation (13.168) is unchanged by a general coordinate transformation of M and by an admissible change of parameter (Definition 13.73).*

Rests on Definitions 13.73, 13.74 and 13.88.

Proof. *Derives Proposition 13.75.*

Under a general coordinate transformation the integrand is a scalar: it is built by contracting $g_{\mu\nu}$, a $(0,2)$ tensor (Definition 13.117), with two copies of the contravariant vector $\mathrm{d}x^\mu/\mathrm{d}\lambda$ (Equation (13.199)), and the Jacobians cancel in pairs exactly as in Equation (13.206). Under a change of parameter $\lambda = \lambda(\bar{\lambda})$ the chain rule gives $\mathrm{d}x^\mu/\mathrm{d}\bar{\lambda} = (\mathrm{d}\lambda/\mathrm{d}\bar{\lambda})\, \mathrm{d}x^\mu/\mathrm{d}\lambda$, so the integrand is multiplied by $\left|\mathrm{d}\lambda/\mathrm{d}\bar{\lambda}\right|$, which is exactly the factor produced by $\mathrm{d}\lambda = (\mathrm{d}\lambda/\mathrm{d}\bar{\lambda})\mathrm{d}\bar{\lambda}$ in the measure, the derivative having a fixed sign by Definition 13.73. □

Definition 13.76 (Arclength parametrisation)**.** For a regular curve, the function

$$s(\lambda) = \int_a^\lambda \sqrt{g_{\mu\nu} \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda'} \frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda'}}\; \mathrm{d}\lambda' \tag{13.169}$$

has $\mathrm{d}s/\mathrm{d}\lambda > 0$ everywhere and is therefore an admissible change of parameter. In the parameter s the tangent $u^\mu = \mathrm{d}x^\mu/\mathrm{d}s$ is a *unit* vector,

$$g_{\mu\nu} u^\mu u^\nu = 1 \,, \tag{13.170}$$

and we call s the *arclength parameter* and u^μ the *unit tangent*, the intrinsic counterpart of Definition 13.12.

Rests on Definitions 13.73 and 13.74.

Curvature vector, binormal, and torsion

Definition 13.77 (Curvature vector)**.** The *curvature vector* of a curve parametrised by arclength is the covariant derivative of its unit tangent along itself,

$$k^\mu = \frac{\mathrm{D}u^\mu}{\mathrm{d}s} \equiv u^\nu \mathring{\nabla}_\nu u^\mu = \frac{\mathrm{d}^2 x^\mu}{\mathrm{d}s^2} + \mathring{\Gamma}^\mu{}_{\nu\rho} \frac{\mathrm{d}x^\nu}{\mathrm{d}s} \frac{\mathrm{d}x^\rho}{\mathrm{d}s} \,, \tag{13.171}$$

and the *curvature function* is $\kappa = \sqrt{g_{\mu\nu} k^\mu k^\nu}$.

Rests on Definitions 13.76 and 13.145 and Theorem 13.150.

In E_3 referred to Cartesian coordinates the Christoffel symbols vanish and Equation (13.171) collapses to $\mathrm{d}^2\boldsymbol{x}/\mathrm{d}s^2$, which is the extrinsic curvature vector Equation (13.39); comparison with Equations (13.121) and (13.297) shows that the curves with $k^\mu = 0$ are precisely the geodesics.

Proposition 13.78 (The curvature vector is orthogonal to the tangent)**.** $g_{\mu\nu} u^\mu k^\nu = 0$.

Rests on Definitions 13.77 and 13.149 and Equation (13.170).

Proof. *Derives Proposition 13.78.*

Differentiate Equation (13.170) along the curve. The Levi-Civita connection is metric compatible (Definition 13.149), so $\mathring{\nabla}_\lambda g_{\mu\nu} = 0$ and

$$0 = \frac{\mathrm{D}}{\mathrm{d}s} \left(g_{\mu\nu} u^\mu u^\nu \right) = 2\, g_{\mu\nu} u^\mu \frac{\mathrm{D}u^\nu}{\mathrm{d}s} = 2\, g_{\mu\nu} u^\mu k^\nu \,.$$

This is the intrinsic form of Equation (13.40), the ambient differentiation there being replaced by covariant differentiation here. □

Where $\kappa > 0$ we may therefore define the *unit principal normal* $N^\mu = k^\mu/\kappa$, orthogonal to u^μ, in parallel with Definition 13.14. For $D = 3$ — the case treated extrinsically in Section 13.3.1, and the observed spatial dimension — the orthogonal complement of $\langle\{u, N\}\rangle$ in each tangent space is one-dimensional, and we fix the *unit binormal* B^μ as the unique unit vector completing (u, N, B) to a positively oriented orthonormal frame; equivalently, with the Levi-Civita tensor Equation (13.257),

$$B^\mu = g^{\mu\sigma}\, \eta_{\sigma\nu\rho}\, u^\nu N^\rho \,, \tag{13.172}$$

which reduces to the cross product Equation (13.44) in Cartesian coordinates.

Theorem 13.79 (Frenet–Serret equations, covariant form)**.** *Let $D = 3$ and let a curve in (M, g) be parametrised by arclength with $\kappa > 0$. Then there is a function τ, the* torsion*, such that the frame (u, N, B) obeys*

$$\frac{\mathrm{D}u^\mu}{\mathrm{d}s} = \kappa\, N^\mu \,, \tag{13.173}$$

$$\frac{\mathrm{D}N^\mu}{\mathrm{d}s} = \tau\, B^\mu - \kappa\, u^\mu \,, \tag{13.174}$$

$$\frac{\mathrm{D}B^\mu}{\mathrm{d}s} = -\tau\, N^\mu \,. \tag{13.175}$$

In E_3 referred to Cartesian coordinates these are Equations (13.43), (13.46) and (13.47).

Rests on Proposition 13.78, Definition 13.149, and Equation (13.172).

Proof. *Derives Theorem 13.79.*

Write $e_1 = u$, $e_2 = N$, $e_3 = B$ and label the frame with capital indices $A, C \in \{1, 2, 3\}$, so that $g_{\mu\nu} e^\mu_A e^\nu_C = \delta_{AC}$ along the whole curve. Expand the covariant derivative of each frame vector in the frame itself, which is legitimate because the frame is a basis of the tangent space at each point of the curve:

$$\frac{\mathrm{D}e^\mu_C}{\mathrm{d}s} = \omega_{AC}\, e^\mu_A\,, \qquad \omega_{AC} = g_{\mu\nu}\, e^\mu_A\, \frac{\mathrm{D}e^\nu_C}{\mathrm{d}s}\,. \tag{13.176}$$

Differentiating $g_{\mu\nu} e^\mu_A e^\nu_C = \delta_{AC}$ along the curve and using metric compatibility (Definition 13.149) gives $\omega_{AC} + \omega_{CA} = 0$: the matrix ω is antisymmetric, so it has exactly three independent entries. They are fixed one by one. By Definition 13.77 and the definition of N, $\mathrm{D}u^\mu/\mathrm{d}s = \kappa N^\mu$, which is Equation (13.173) and reads $\omega_{21} = \kappa$, $\omega_{11} = \omega_{31} = 0$. Antisymmetry then gives $\omega_{12} = -\kappa$ and $\omega_{13} = 0$. The single entry left undetermined is ω_{32}, which we *define* to be the torsion τ; antisymmetry gives $\omega_{23} = -\tau$. Substituting into Equation (13.176) for $C = 2$ and $C = 3$ yields Equations (13.174) and (13.175). In Cartesian coordinates on E_3 the Christoffel symbols vanish, $\mathrm{D}/\mathrm{d}s$ becomes $\mathrm{d}/\mathrm{d}s$, and the three equations are those obtained extrinsically in Section 13.3.1. □

Remark 13.80. The proof above is shorter than the extrinsic one of Section 13.3.1 because it uses no property of the ambient space: the antisymmetry of ω is metric compatibility and nothing else, and it is what makes the coefficient matrix of the Frenet–Serret system antisymmetric in *any* Riemannian manifold. The same argument in dimension D produces $D-1$ curvature functions and a frame of D vectors. What does *not* carry over unchanged is the fundamental theorem of the theory of curves — Theorem 13.16, that κ and τ determine the curve up to a rigid motion — proved in Section 13.3.1 for E_3. On a general Riemannian manifold prescribed κ and τ still determine the curve uniquely once an initial point and an initial orthonormal frame are fixed, and for as far as the solution extends: the three Frenet equations above, together with $\mathrm{d}x^\mu/\mathrm{d}s = u^\mu$, are a first-order system in which the frame enters linearly. Uniqueness *up to isometry* is a stronger statement, and it needs the isometries of Section 13.7.1 to act transitively on orthonormal frames, so that two solutions with different initial data can be carried onto one another. That is a genuine hypothesis and not a technicality: a generic metric admits no isometry but the identity, and then two curves with the same κ and τ started at different points are simply two different curves. Enough isometries are available exactly on the maximally symmetric spaces of Theorem 13.160, whose curvature is constant, E_3 being the case $\epsilon = 0$.

13.5 Tensor analysis on manifolds

13.5.1 Vectors, covectors, and tensors on a manifold

The definition of a vector on a manifold

The definition of a vector in $\mathbb{R}^N$ is an oriented straight-line segment joining two points. This tells us that a vector as we know it is a *bilocal* object. Would this definition remain consistent if the space no longer had Euclidean geometry — that is, on a general manifold? Consider an n-dimensional manifold $M \subset \mathbb{R}^N$. The Euclidean metric of $\mathbb{R}^N$ allows us to define straight lines joining points on the manifold, but such straight lines are clearly not contained in M.

Along which route, then, should the line between A and B be traced? There is no possible route unless we define beforehand a metric on the space M_n, for without this concept we cannot have a concept of straight line. For this reason we cannot define a vector on the manifold as a bilocal object: we must develop a formalism to understand a vector as a *local* object.

Let $(\gamma,]a, b[\,)$ be a coordinate system describing a curve contained in M. The curve is characterised by the position vectors $\boldsymbol{x}_{\gamma(\lambda)}$ with λ varying in $]a, b[$. Denote by λ_0 the value of the parameter at which the point of the curve is $P = \gamma(\lambda_0) \in M$.

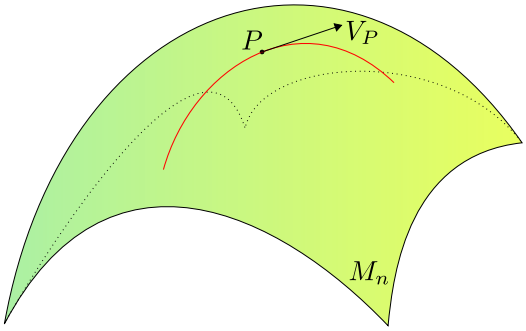


Figure 13.16: A curve through $P = \gamma(\lambda_0)$ on a manifold M: its tangent vector at P does not lie in M but in a tangent plane touching M at P.

We say that the vector $V_P \in \mathbb{R}^N$ is a *tangent vector* to M at P if and only if it is tangent to the curve at P, that is, if its components with respect to the coordinate system $(\gamma,]a, b[\,)$ are

$$V^i(P) = \left.\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\right|_{\lambda_0} . \tag{13.177}$$

The tangent vector V_P must not be viewed as a displacement lying in M, but as lying in some kind of tangent surface coinciding with M at P. Since M is equipped with an atlas, we can define a differentiable map of the form $f : M \longrightarrow \mathbb{R}$. Then $f \circ x : \mathbb{R} \longrightarrow \mathbb{R}$ is differentiable, being a composition of differentiable maps, and

$$\left.\frac{\mathrm{d}}{\mathrm{d}\lambda}(f \circ x)\right|_{\lambda_0} = \lim_{\lambda \to \lambda_0} \frac{f(x(\lambda)) - f(x(\lambda_0))}{\lambda - \lambda_0} . \tag{13.178}$$

But which map f do we choose? It does not matter. What matters is to see how the curve generates changes in f. Indeed, $f \circ x : \mathbb{R} \longrightarrow \mathbb{R}$ allows us to evaluate the changes in f; these changes are given by the directional derivative (Equation (13.178)) in the direction of the curve, that is, in the direction of the vector V_P.

We may conclude that the operator generating the changes of f in the direction of V_P is $\left.\frac{\mathrm{d}}{\mathrm{d}\lambda}\right|_{\lambda_0}$. By the chain rule and Equation (13.177),

$$\left.\frac{\mathrm{d}}{\mathrm{d}\lambda}\right|_{\lambda_0} = V_P^i \, \partial_i \, .$$

Therefore $\left.\frac{\mathrm{d}}{\mathrm{d}\lambda}\right|_{\lambda_0}$ can be written as a linear combination of the basis $\{\partial_i\}$, with components exactly V_P^i.

Definition 13.81 (Vector on a manifold)**.** We define a *vector* on M_n at P as the operator

$$V_P = V^i(P) \, \frac{\partial}{\partial x^i(P)} = \left.\frac{\mathrm{d}}{\mathrm{d}\lambda}\right|_{\lambda_0} . \tag{13.179}$$

Rests on Definitions 13.45, 13.48 and 13.71.

Definition 13.82 (Vector field)**.** We say that a *vector field* is an assignment rule that associates a vector with each point of a manifold.

Rests on Definition 13.81.

We must then symbolise a vector field without reference to the point of the manifold, since we are considering all of them. In the context of the definition of a vector, we symbolise a vector field V as

$$V = V^i \partial_i = \frac{\mathrm{d}}{\mathrm{d}\lambda} \,. \tag{13.180}$$

When no reference to the point is made, it is understood that we are speaking of a field.

Tangent space and coordinate basis

Owing to the linearity of the derivative, and since it has the properties of addition and multiplication by a scalar, the set of all vectors tangent to M at P forms an n-dimensional vector space, which we call the *tangent space* and denote $T_M(P)$ (for an arbitrary point, T_M). This may seem inconceivable, but so it is: at each point of the manifold we can define a vector space.

As discussed in the definition of a vector on a manifold, we can write a vector V on M as a linear combination of the vectors ∂_i,

$$V = V^i \partial_i \,. \tag{13.181}$$

This means that the set $\{\partial_i\}_{i=1}^n$ forms a basis of the tangent space, which we call the *coordinate (or holonomic) basis*. We therefore have

$$T_M = \langle \{\partial_i\} \rangle \,, \quad i = 1, \ldots, n \,. \tag{13.182}$$

The vectors of the coordinate basis are tangent to the coordinate lines.

For example, the coordinate-basis vectors of E_2 in Cartesian coordinates are ∂_x and ∂_y, while in polar coordinates they are ∂_r and ∂_ϕ. The relation between the Cartesian and polar coordinates describing E_2 is

$$x = r\cos\phi \,, \quad y = r\sin\phi \,,$$

so that

$$\begin{aligned} \partial_r &= \frac{\partial x}{\partial r}\frac{\partial}{\partial x} + \frac{\partial y}{\partial r}\frac{\partial}{\partial y} \\ &= \cos\phi \, \frac{\partial}{\partial x} + \sin\phi \, \frac{\partial}{\partial y} \,, \end{aligned}$$

and likewise

$$\begin{aligned} \partial_\phi &= \frac{\partial x}{\partial \phi}\frac{\partial}{\partial x} + \frac{\partial y}{\partial \phi}\frac{\partial}{\partial y} \\ &= -r\sin\phi \, \frac{\partial}{\partial x} + r\cos\phi \, \frac{\partial}{\partial y} \,, \end{aligned}$$

whence

$$\partial_r = \cos\phi \, \partial_x + \sin\phi \, \partial_y \,, \qquad \partial_\phi = -r\sin\phi \, \partial_x + r\cos\phi \, \partial_y \,. \tag{13.183}$$

Covectors and the cotangent space

Definition 13.83 (Covector)**.** Let $P \in M$. We define a *covector* at P as a linear functional

$$\begin{array}{rccc} \psi_P : & T_M(P) & \longrightarrow & \mathbb{R} \\ & V_P & \longmapsto & \psi_P(V_P) \,. \end{array}$$

Rests on Definition 13.81.

From the theory of vector spaces we know that the set of all linear functionals defined on a vector space forms a vector space, called the dual. Thus the set of all covectors defined at a point $P \in M$ forms the vector space

$$\mathcal{L}\left(T_M(P), \mathbb{R}\right) = T_M^*(P)\,, \tag{13.184}$$

which we call the *cotangent space* at P. As shown below (Equation (13.239)), the differentials $\mathrm{d}x^i$ furnish the basis dual to the coordinate basis, so that

$$T_M^*(P) = \left\langle \left\{ \mathrm{d}x^i \right\} \right\rangle\,, \quad i = 1, \ldots, n\,. \tag{13.185}$$

Tensors

Definition 13.84 (Tensor)**.** Let $P \in M$, and consider $T_M(P) = \langle\{e_i\}\rangle$ and $T_M^*(P) = \langle\{\omega^i\}\rangle$. We define a *tensor* of type (r, s) at P as a multilinear map

$$T : \underbrace{T_M^*(P) \times \cdots \times T_M^*(P)}_{r \text{ times}} \times \underbrace{T_M(P) \times \cdots \times T_M(P)}_{s \text{ times}} \longrightarrow \mathbb{R}\,,$$

whose components are its values on the basis elements,

$$T\left(\omega^{i_1}, \ldots, \omega^{i_r}, e_{j_1}, \ldots, e_{j_s}\right) = T^{i_1 \ldots i_r}{}_{j_1 \ldots j_s}\,.$$

Rests on Definitions 13.81 and 13.83.

We say that a tensor is *contravariant* if it is of type $(r, 0)$ with $r \in \mathbb{N}$, and *covariant* if it is of type $(0, s)$ with $s \in \mathbb{N}$. For example, a vector is a contravariant tensor of type $(1, 0)$ and a covector is a covariant tensor of type $(0, 1)$.

Pullback and pushforward

If we wished to compare two vectors assigned at distinct points of a differentiable manifold, we could not simply subtract them, since they belong to different vector spaces. We must provide a mechanism that maps one of the vectors into the space of the vector with which we wish to compare it.

Consider a diffeomorphism $f : M \longrightarrow M$. For a point P, the map f induces the *pushforward* $f_* : T_M(P) \longrightarrow T_M(f(P))$. This amounts to taking a tangent vector at P and locating it at $f(P)$.

One also has the *pullback* $f^* : T_M^*(f(P)) \longrightarrow T_M^*(P)$, which to each covector at $f(P)$, say $\omega_{f(P)}$, assigns by definition the covector $f^*(\omega_{f(P)}) = \omega_{f(P)} \circ f_*$ at P. Covectors are linear functionals acting on vectors of the tangent space: for a tangent vector V_P at P,

$$\begin{aligned}\left(f^*(\omega_{f(P)})\right)(V_P) &= \left(\omega_{f(P)} \circ f_*\right)(V_P) \\ &= \omega_{f(P)}\left(f_*(V_P)\right) \\ &= \omega_{f(P)}\left(V_{f(P)}\right)\,.\end{aligned}$$

Vector fields on basic manifolds

Example 13.85 (Euclidean space E_3)**.** The Cartesian chart of Section 13.4.2 is global, so the coordinate frame

$$\{\partial_x, \partial_y, \partial_z\} \tag{13.186}$$

consists of three vector fields defined at every point, and every vector field on E_3 is $V = V^x\partial_x + V^y\partial_y + V^z\partial_z$ with three functions as components. Spherical coordinates,

$$x = r\sin\theta\cos\phi\,, \qquad y = r\sin\theta\sin\phi\,, \qquad z = r\cos\theta\,, \tag{13.187}$$

give a second chart, and the two frames are related by the chain rule, which is the transformation law of a coordinate basis, $\bar{\partial}_i = (\partial x^j/\partial \bar{x}^i)\partial_j$:

$$\partial_r = \sin\theta\cos\phi\,\partial_x + \sin\theta\sin\phi\,\partial_y + \cos\theta\,\partial_z\,, \tag{13.188}$$

$$\partial_\theta = r\cos\theta\cos\phi\,\partial_x + r\cos\theta\sin\phi\,\partial_y - r\sin\theta\,\partial_z\,, \tag{13.189}$$

$$\partial_\phi = -r\sin\theta\sin\phi\,\partial_x + r\sin\theta\cos\phi\,\partial_y\,. \tag{13.190}$$

Two features are worth recording. First, the spherical chart is not global: it degenerates at $r = 0$ and along the polar axis $\sin\theta = 0$, where ∂_ϕ vanishes and ϕ is undefined — the same obstruction met for the circle in Section 13.4.3. Second, the spherical coordinate frame is not orthonormal: Equations (13.188) to (13.190) give $|\partial_r| = 1$, $|\partial_\theta| = r$ and $|\partial_\phi| = r\sin\theta$ in the Euclidean metric Equation (13.159). The normalized frame $\hat{e}_r = \partial_r$, $\hat{e}_\theta = \partial_\theta/r$, $\hat{e}_\phi = \partial_\phi/(r\sin\theta)$ familiar from vector calculus is therefore *not* a coordinate frame: it is an anholonomic frame, with nonvanishing structure functions in the sense of Equation (13.266), and this is the elementary origin of the vielbein of Section 13.8.

Rests on Definition 13.82 and Equation (13.199).

Example 13.86 (A curve)**.** Let $\mathcal{C}$ be a differentiable curve described locally by $(\gamma,]a, b[\,)$ as in Equation (13.163). Its tangent space at every point is one-dimensional, spanned by the single coordinate-basis vector $\mathrm{d}/\mathrm{d}\lambda$, so *every* vector field on $\mathcal{C}$ has the form

$$V = f(\lambda)\,\frac{\mathrm{d}}{\mathrm{d}\lambda}\,, \tag{13.191}$$

with one component function f. Read in the ambient E_n through Equation (13.166), the field $\mathrm{d}/\mathrm{d}\lambda$ is the tangent vector $(\mathrm{d}x^i/\mathrm{d}\lambda)\,\hat{x}_i$ of the curve. Under an admissible change of parameter (Definition 13.73) the chain rule gives $\mathrm{d}/\mathrm{d}\lambda = \left(\mathrm{d}\bar{\lambda}/\mathrm{d}\lambda\right)\mathrm{d}/\mathrm{d}\bar{\lambda}$, the one-dimensional instance of Equation (13.199). Since the coordinate field $\mathrm{d}/\mathrm{d}\lambda$ never vanishes, a curve always carries a nowhere-vanishing vector field; the next example shows that this is a privilege of low dimension and not a general fact.

Rests on Definitions 13.71 and 13.82.

Example 13.87 (The 2-sphere)**.** Let $S_2 \subset E_3$ be the sphere of radius R (Example 13.61) and consider the three rotation generators of E_3,

$$L_1 = y\,\partial_z - z\,\partial_y\,, \qquad L_2 = z\,\partial_x - x\,\partial_z\,, \qquad L_3 = x\,\partial_y - y\,\partial_x\,. \tag{13.192}$$

Each is tangent to S_2, because each annihilates the defining function $F = x^2 + y^2 + z^2$ of Equation (13.147): for instance $L_3 F = x(2y) - y(2x) = 0$. A vector that annihilates F is tangent to the level set, since F is constant along any curve inside it; hence the L_i restrict to vector fields on S_2. In the spherical chart Equation (13.187) with $r = R$ held fixed, so that (θ, ϕ) are coordinates on the sphere, comparison with Equations (13.189) and (13.190) gives

$$L_1 = -\sin\phi\,\partial_\theta - \cot\theta\cos\phi\,\partial_\phi\,, \tag{13.193}$$

$$L_2 = \cos\phi\,\partial_\theta - \cot\theta\sin\phi\,\partial_\phi\,, \tag{13.194}$$

$$L_3 = \partial_\phi\,. \tag{13.195}$$

The last is immediate from Equation (13.190), which reads $\partial_\phi = -y\,\partial_x + x\,\partial_y = L_3$; the other two follow by the same substitution. Computing the Lie brackets Equation (13.273) of the fields Equation (13.192) directly gives

$$[L_i, L_j] = -\varepsilon_{ijk} L_k\,, \tag{13.196}$$

so the three fields span a three-dimensional Lie algebra; it is isomorphic to $\mathfrak{so}(3)$, the sign being absorbed by $L_i \to -L_i$. These are the Killing vectors of the round sphere

(Section 13.9.2), and they saturate the bound $D(D+1)/2 = 3$ of Proposition 13.140 at $D = 2$: the round S_2 is maximally symmetric.

Each of the three fields vanishes somewhere: $L_3 = \partial_\phi$ vanishes at the two poles $\sin\theta = 0$, and L_1, L_2 likewise vanish at the two points of S_2 on their own axis. This is not an accident of the choice. *No continuous vector field on S_2 is nowhere vanishing* — the "hairy ball" theorem. It follows from the Poincaré index theorem, which states that on a compact orientable surface the indices of the isolated zeros of a vector field sum to the Euler characteristic of the surface [1575]; for S_2 that characteristic is 2, so the sum of the indices cannot be zero and zeros cannot be absent. Two consequences are worth naming: S_2 admits no global frame of the kind Equation (13.186) — it is not parallelizable, unlike E_n and unlike the curve of Example 13.86 — and consequently no vielbein (Section 13.8) on S_2 can be defined by a single chart-independent formula valid everywhere. Charts are not a convenience here but a necessity.

Rests on Definitions 13.81 and 13.82 and Example 13.61.

13.5.2 Tensors under general coordinate transformations

General coordinate transformations

We call a *coordinate system* K a set of functions $\{x^i\}_{i=1}^n$ such that the location of a point P is given by $x^i(P)$.

Definition 13.88 (General coordinate transformation)**.** We call *general coordinate transformation* (GCT) a transformation relating the coordinates x^i of a coordinate system K to the coordinates $\bar{x}^i$ of a coordinate system $\bar{K}$, that is, a transformation of the form

$$\bar{x}^i = \bar{x}^i(x^j)\,, \qquad x^i = x^i(\bar{x}^j)\,. \tag{13.197}$$

For practical purposes we assume that these n functions of n variables are of class C^2.

In the context of coordinate systems it is necessary to extend the definition of the *Kronecker delta* to every coordinate system: we define the *Kronecker delta symbol* as the quantities

$$\delta^i{}_j = \begin{cases} 1\,, & i = j \\ 0\,, & i \neq j \end{cases} \tag{13.198}$$

in every coordinate system.

A GCT is *linear* if and only if it can be written as $\bar{x}^i = \Lambda^i{}_j x^j$ with a constant matrix $\Lambda^i{}_j$ (compare Definition 13.2).

Transformation law of a vector

Since a vector field is the coordinate-independent object $V = V^i\partial_i$, its components in two coordinate systems obey

$$V^i\partial_i = \bar{V}^i\bar{\partial}_i\,, \qquad (\bar{x}^j)$$
$$V^i\frac{\partial\bar{x}^j}{\partial x^i} = \bar{V}^i\delta^j{}_i\,,$$

where both sides have been applied to the coordinate function $\bar{x}^j$; hence

$$\bar{V}^i = \frac{\partial\bar{x}^i}{\partial x^j}V^j\,. \tag{13.199}$$

Contravariant, covariant, and mixed tensors

Definition 13.89 (Contravariant and covariant tensors)**.** We define the *components of a contravariant tensor of rank* r as quantities that change coordinate system under the transformation

$$\bar{A}^{i_1\cdots i_r} = \frac{\partial \bar{x}^{i_1}}{\partial x^{j_1}} \cdots \frac{\partial \bar{x}^{i_r}}{\partial x^{j_r}} A^{j_1\cdots j_r} \,, \tag{13.200}$$

placing the indices in the upper position to denote that these indices enumerate the components of a contravariant tensor. Analogously, we define the *components of a covariant tensor of rank* s as quantities that change coordinate system under the transformation

$$\bar{B}_{i_1\cdots i_s} = \frac{\partial x^{j_1}}{\partial \bar{x}^{i_1}} \cdots \frac{\partial x^{j_s}}{\partial \bar{x}^{i_s}} B_{j_1\cdots j_s} \,, \tag{13.201}$$

placing the indices in the lower position.

Rests on Definition 13.88.

Definition 13.90 (Mixed tensor)**.** We define the components of a *mixed tensor of order* $r+s$ as quantities that change coordinate system under the transformation

$$\bar{T}^{i_1\cdots i_r}{}_{l_1\cdots l_s} = \frac{\partial \bar{x}^{i_1}}{\partial x^{j_1}} \cdots \frac{\partial \bar{x}^{i_r}}{\partial x^{j_r}} \frac{\partial x^{k_1}}{\partial \bar{x}^{l_1}} \cdots \frac{\partial x^{k_s}}{\partial \bar{x}^{l_s}} T^{j_1\cdots j_r}{}_{k_1\cdots k_s} \,, \tag{13.202}$$

which we call a tensor of type (r,s) and order $r+s$.

Rests on Definitions 13.88 and 13.89.

We call *contravariant vector* a tensor of type $(1,0)$; from Equation (13.200), its components with respect to another coordinate system are given by

$$\bar{T}^i = \frac{\partial \bar{x}^i}{\partial x^j} T^j \,, \qquad T^j = \frac{\partial x^j}{\partial \bar{x}^i} \bar{T}^i \,. \tag{13.203}$$

We call *covariant vector* a tensor of type $(0,1)$; from Equation (13.201),

$$\bar{T}_i = \frac{\partial x^j}{\partial \bar{x}^i} T_j \,, \qquad T_j = \frac{\partial \bar{x}^i}{\partial x^j} \bar{T}_i \,. \tag{13.204}$$

The displacement vector

For a GCT one cannot say that the coordinates themselves transform as a vector would: under a GCT the coordinates are neither contravariant nor covariant. Differentiating the relations Equation (13.197), however, we see that

$$\mathrm{d}\bar{x}^i = \frac{\partial \bar{x}^i}{\partial x^j} \mathrm{d}x^j \,, \qquad \mathrm{d}x^i = \frac{\partial x^i}{\partial \bar{x}^j} \mathrm{d}\bar{x}^j \,, \tag{13.205}$$

so the differential of the coordinates transforms as a contravariant vector, which we call the *displacement vector*.

Scalars and invariance

The notion of the rank of a tensor under general transformations is the number of indices it carries, and the number of times the Jacobian matrix of the transformation must be applied to carry out a change of coordinates. If we apply this matrix zero times, we obtain the same quantities in both coordinate systems, since we are not transforming them at all. We may extend the definition to *tensors of rank* 0 by defining these *invariant* quantities: quantities that depend on position with respect to a coordinate system but take the same value in every coordinate system. That is,

$$\bar{T}(\bar{x}) = T(x) \,. \tag{13.206}$$

More generally, we call *invariant tensor* a tensor satisfying

$$\bar{T}_{i_1\cdots i_r}(\bar{x}) = T_{i_1\cdots i_r}(x)\,. \tag{13.207}$$

We call *scalar* an invariant tensor of rank 0.

Scalar product

Consider the components of a contravariant vector T^i and of a covariant vector U_l. With respect to $\bar{K}$,

$$\bar{T}^i = \frac{\partial \bar{x}^i}{\partial x^j} T^j\,, \qquad \bar{U}_l = \frac{\partial x^k}{\partial \bar{x}^l} U_k\,.$$

Consider the quantity

$$\begin{aligned}
\bar{T}^i \bar{U}_i &= \frac{\partial \bar{x}^i}{\partial x^j} T^j \frac{\partial x^k}{\partial \bar{x}^i} U_k \\
&= \frac{\partial \bar{x}^i}{\partial x^j} \frac{\partial x^k}{\partial \bar{x}^i}\, T^j U_k \\
&= \frac{\partial x^k}{\partial x^j} T^j U_k \\
&= \delta^k{}_j T^j U_k \\
&= T^j U_j\,,
\end{aligned}$$

from which we see that this type of product is a scalar under GCTs. We call this multiplication of vectors the *scalar product*.

The Kronecker delta as a mixed tensor

From its definition, the Kronecker delta is an invariant tensor. Observe that

$$\begin{aligned}
\bar{\delta}^i{}_l &= \frac{\partial \bar{x}^i}{\partial \bar{x}^l} \\
&= \frac{\partial \bar{x}^i}{\partial x^j} \frac{\partial x^j}{\partial \bar{x}^l} \\
&= \frac{\partial \bar{x}^i}{\partial x^j} \frac{\partial x^k}{\partial \bar{x}^l}\, \delta^j{}_k\,.
\end{aligned}$$

Thus, from Equation (13.202), the Kronecker delta satisfies the definition of a tensor of type $(1,1)$.

Symmetry and antisymmetry

We say that a tensor with components $T^{i_1\cdots i_r}{}_{l_1\cdots l_s}$ is a *tensor symmetric with respect to the indices* i_p and i_q if and only if

$$T^{i_1\cdots i_p\cdots i_q\cdots i_r}{}_{l_1\cdots l_s} = T^{i_1\cdots i_q\cdots i_p\cdots i_r}{}_{l_1\cdots l_s}\,, \tag{13.208}$$

or a *tensor symmetric with respect to the indices* l_p and l_q if and only if

$$T^{i_1\cdots i_r}{}_{l_1\cdots l_p\cdots l_q\cdots l_s} = T^{i_1\cdots i_r}{}_{l_1\cdots l_q\cdots l_p\cdots l_s}\,. \tag{13.209}$$

The quantities so defined are tensors under GCTs, by the permutation of indices. Analogously, the tensor is *antisymmetric* with respect to a pair of upper indices, or a pair of lower indices, if the interchange of the pair produces a minus sign; again the quantities so defined are tensors under GCTs.

Consider now a tensor of type $(1,1)$, and suppose that in some coordinate system

$$T^i{}_j = T^j{}_i \,, \tag{13.210}$$

that is, quantities symmetric with respect to one contravariant and one covariant index. Then

$$\begin{aligned} \bar{T}^i{}_j &= \frac{\partial \bar{x}^i}{\partial x^l}\frac{\partial x^k}{\partial \bar{x}^j} T^l{}_k \\ \text{[Equation (13.210)]} \quad &= \frac{\partial \bar{x}^i}{\partial x^l}\frac{\partial x^k}{\partial \bar{x}^j} T^k_l \,, \end{aligned}$$

from which we see that this type of symmetry does not generate tensors under GCTs. Symmetry or antisymmetry properties must therefore always be taken with respect to two contravariant or two covariant indices; for this reason we display them below only for covariant indices.

The symmetry or antisymmetry of a rank-2 tensor reads

$$S_{ij} = S_{ji} \,, \qquad A_{ij} = -A_{ji} \,, \tag{13.211}$$

which in matrix notation is equivalent to

$$\mathbb{S} = \mathbb{S}^{\mathsf{T}} \,, \qquad \mathbb{A} = -\mathbb{A}^{\mathsf{T}} \,. \tag{13.212}$$

Symmetric and antisymmetric parts

Definition 13.91 (Symmetric and antisymmetric parts)**.** We define the *symmetric part* of a rank-r tensor $T_{i_1\cdots i_r}$ with respect to the indices i_s and i_t as the quantities

$$S_{i_1\cdots(i_s|\cdots|i_t)\cdots i_r} \overset{\text{not.}}{=} \frac{1}{2}\left(T_{i_1\cdots i_s\cdots i_t\cdots i_r} + T_{i_1\cdots i_t\cdots i_s\cdots i_r}\right) \,, \tag{13.213}$$

and analogously the *antisymmetric part*, with respect to the indices i_s and i_t, as the quantities

$$A_{i_1\cdots[i_s|\cdots|i_t]\cdots i_r} \overset{\text{not.}}{=} \frac{1}{2}\left(T_{i_1\cdots i_s\cdots i_t\cdots i_r} - T_{i_1\cdots i_t\cdots i_s\cdots i_r}\right) \,. \tag{13.214}$$

Adding Equations (13.213) and (13.214) we see that

$$T_{i_1\cdots i_r} = S_{i_1\cdots(i_s|\cdots|i_t)\cdots i_r} + A_{i_1\cdots[i_s|\cdots|i_t]\cdots i_r} \,, \tag{13.215}$$

so a rank-r tensor can be decomposed into its symmetric and its antisymmetric part. For a tensor of order 2,

$$S_{ij} = \frac{1}{2}\left(T_{ij} + T_{ji}\right) \,, \qquad A_{ij} = \frac{1}{2}\left(T_{ij} - T_{ji}\right) \,, \qquad T_{ij} = S_{ij} + A_{ij} \,. \tag{13.216}$$

The vector-product tensor

We define the *vector-product tensor* as

$$C_{ij} = A_i B_j - A_j B_i \,. \tag{13.217}$$

From the algebra of tensors, the quantities C_{ij} are the components of a tensor of type $(0,2)$ under GCTs. Moreover, from Equation (13.217),

$$\begin{aligned} C_{ji} &= A_j B_i - A_i B_j \\ &= -\left(A_i B_j - A_j B_i\right) \,, \end{aligned}$$

so that

$$C_{ij} = -C_{ji} \,, \tag{13.218}$$

that is, the vector-product tensor is an antisymmetric tensor of order 2.

Gradient of a scalar

Consider Equation (13.206) and apply the partial derivative with respect to the coordinate $\bar{x}^i$:

$$\begin{aligned} \bar{T} &= T\,, \qquad \frac{\partial}{\partial \bar{x}^i}(\,) \\ \frac{\partial \bar{T}}{\partial \bar{x}^i} &= \frac{\partial T}{\partial \bar{x}^i} \\ &= \frac{\partial T}{\partial x^j}\frac{\partial x^j}{\partial \bar{x}^i}\,, \end{aligned}$$

from which we see that

$$\frac{\partial \bar{T}}{\partial \bar{x}^i} = \frac{\partial x^j}{\partial \bar{x}^i}\frac{\partial T}{\partial x^j}\,, \tag{13.219}$$

that is, the *gradient of a scalar* is a covariant vector under GCTs.

A Jacobian identity

From the chain rule,

$$\begin{aligned} \frac{\partial \bar{x}^i}{\partial \bar{x}^j} &= \delta^i{}_j \\ \frac{\partial \bar{x}^i}{\partial x^k}\frac{\partial x^k}{\partial \bar{x}^j} &= \delta^i{}_j\,, \qquad \det(\,) \\ \left|\frac{\partial \bar{x}}{\partial x}\,\frac{\partial x}{\partial \bar{x}}\right| &= 1\,, \end{aligned} \tag{13.220}$$

where $|\partial\bar{x}/\partial x|$ denotes the Jacobian determinant of the transformation.

13.5.3 Tensor densities

The quantities met so far transform under a GCT by factors of $\partial\bar{x}/\partial x$ alone. A second family of objects, indispensable as soon as one integrates over a manifold, transforms by those factors *and* by a power of the Jacobian determinant.

Definition 13.92 (Tensor density)**.** Write

$$J = \det\left(\frac{\partial \bar{x}^i}{\partial x^j}\right) \tag{13.221}$$

for the Jacobian determinant of the GCT Equation (13.197). Quantities $\mathfrak{T}^{i_1\ldots i_r}{}_{j_1\ldots j_s}$ are the components of a *tensor density of weight* $w \in \mathbb{Z}$, of type (r, s), if under a GCT they transform as

$$\bar{\mathfrak{T}}^{i_1\ldots i_r}{}_{j_1\ldots j_s} = J^{-w}\,\frac{\partial \bar{x}^{i_1}}{\partial x^{k_1}}\cdots\frac{\partial \bar{x}^{i_r}}{\partial x^{k_r}}\,\frac{\partial x^{l_1}}{\partial \bar{x}^{j_1}}\cdots\frac{\partial x^{l_s}}{\partial \bar{x}^{j_s}}\,\mathfrak{T}^{k_1\ldots k_r}{}_{l_1\ldots l_s}\,. \tag{13.222}$$

A density of weight 0 is an ordinary tensor (Equations (13.200) and (13.201)); a density of weight 0 and type $(0, 0)$ is a scalar Equation (13.206). A quantity transforming by $|J|^{-w}$ rather than J^{-w} is a *relative* tensor density of weight w, and the two notions differ only for orientation-reversing transformations.

Rests on Definitions 13.88 and 13.89.

Proposition 13.93 (Algebra of densities)**.** *Densities of the same weight and type form a vector space under pointwise addition; the product of a density of weight w_1 with one of weight w_2 is a density of weight $w_1 + w_2$; contraction of an upper with a lower index preserves the weight; and the quotient of two densities of equal weight, where defined, is an ordinary tensor.*

Rests on Definition 13.92.

Derivation. Derives Proposition 13.93.

Each clause is read off Equation (13.222). Linearity is immediate, since the same factor J^{-w} multiplies both summands. For the product, the Jacobian factors multiply as $J^{-w_1}J^{-w_2} = J^{-(w_1+w_2)}$ while the partial-derivative factors are those of the tensor product. For a contraction, one factor $\partial\bar{x}^i/\partial x^k$ meets one factor $\partial x^l/\partial\bar{x}^i$ and the two collapse to $\delta^l{}_k$ by Equation (13.198), leaving J^{-w} untouched. For the quotient, the factors J^{-w} cancel between numerator and denominator, so the result transforms with no Jacobian factor at all. □

Example 13.94 (The two densities this chapter uses)**.** The Levi-Civita symbol $\varepsilon_{i_1\dots i_n}$, defined to take the values $0, \pm 1$ in *every* coordinate system, is a covariant density of weight -1: expanding the determinant by the Leibniz formula gives

$$\frac{\partial x^{k_1}}{\partial\bar{x}^{i_1}}\cdots\frac{\partial x^{k_n}}{\partial\bar{x}^{i_n}}\,\varepsilon_{k_1\dots k_n} = J^{-1}\,\varepsilon_{i_1\dots i_n}\,, \tag{13.223}$$

which is Equation (13.222) with $w = -1$. Correspondingly $\sqrt{|\det g|}$ is a scalar density of weight $+1$, because $\bar{g}_{ij} = (\partial x^k/\partial\bar{x}^i)(\partial x^l/\partial\bar{x}^j)g_{kl}$ gives $\det\bar{g} = J^{-2}\det g$. Their product $\sqrt{|\det g|}\,\varepsilon_{i_1\dots n}$ has weight 0 and is therefore an honest tensor — the volume form of Section 13.6.7 — and $\sqrt{|\det g|}\,\mathrm{d}^n x$ is invariant, which is what makes an integral over a manifold well defined.

Rests on Definition 13.92 and Proposition 13.93.

The sources develop none of this in general and treat only the following three-dimensional instances.

The pseudo-vector cross product

Consider the vector-product tensor (Equation (13.217)). Since this tensor is totally antisymmetric, we can define the associated dual pseudo-tensor, in three dimensions, as

$$C_i = \frac{1}{2}\varepsilon_{ijk}C_{jk}\,, \tag{13.224}$$

which we call the *pseudo-vector cross product*, or *cross product of two vectors*. From Equation (13.217),

$$\begin{aligned} C_i &= \frac{1}{2}\varepsilon_{ijk}\,(A_jB_k - A_kB_j) \\ &= \frac{1}{2}\varepsilon_{ijk}A_jB_k - \frac{1}{2}\varepsilon_{ijk}A_kB_j \\ &= \frac{1}{2}\varepsilon_{ijk}A_jB_k + \frac{1}{2}\varepsilon_{ikj}A_kB_j \\ &= \frac{1}{2}\varepsilon_{ijk}A_jB_k + \frac{1}{2}\varepsilon_{ijk}A_jB_k\,, \end{aligned}$$

so that

$$C_i = \varepsilon_{ijk}A_jB_k\,. \tag{13.225}$$

The source calls Equation (13.225) a pseudo-vector and stops there. The name records a property, and the property has to be established; the three results below do that, and supply at the same time the elementary identities of the three-dimensional cross product, which the sources use without stating and which the physical parts of this treatise cref constantly.

Proposition 13.95 (The elementary cross-product identities)**.** *For vectors $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ of E_3, with $(\boldsymbol{A}\times\boldsymbol{B})_i = \varepsilon_{ijk}A_jB_k$ as in Equation (13.225),*

$$\boldsymbol{A}\cdot(\boldsymbol{B}\times\boldsymbol{C}) = \boldsymbol{B}\cdot(\boldsymbol{C}\times\boldsymbol{A}) = \boldsymbol{C}\cdot(\boldsymbol{A}\times\boldsymbol{B}) = \varepsilon_{ijk}A_iB_jC_k\,, \tag{13.226}$$

$$\boldsymbol{A}\times(\boldsymbol{B}\times\boldsymbol{C}) = \boldsymbol{B}\,(\boldsymbol{A}\cdot\boldsymbol{C}) - \boldsymbol{C}\,(\boldsymbol{A}\cdot\boldsymbol{B})\,, \tag{13.227}$$

$$|\boldsymbol{A}\times\boldsymbol{B}|^2 = |\boldsymbol{A}|^2\,|\boldsymbol{B}|^2 - (\boldsymbol{A}\cdot\boldsymbol{B})^2\,. \tag{13.228}$$

Rests on Proposition 13.6 and Equation (13.225).

Derivation. *Derives Proposition 13.95.*

Equation (13.226): all three scalars equal $\varepsilon_{ijk}A_iB_jC_k$ after renaming the summation indices, because $\varepsilon_{ijk} = \varepsilon_{jki} = \varepsilon_{kij}$ — a cyclic permutation of three symbols is even.

Equation (13.227): by Equation (13.225) and Equation (13.29),

$$\begin{aligned}[\boldsymbol{A}\times(\boldsymbol{B}\times\boldsymbol{C})]_i &= \varepsilon_{ijk}A_j\,\varepsilon_{klm}B_lC_m = \varepsilon_{kij}\varepsilon_{klm}A_jB_lC_m\\ &= (\delta_{il}\delta_{jm} - \delta_{im}\delta_{jl})\,A_jB_lC_m = B_i\,(A_jC_j) - C_i\,(A_jB_j)\,.\end{aligned}$$

Equation (13.228): with the same contraction,

$$|\boldsymbol{A}\times\boldsymbol{B}|^2 = \varepsilon_{ijk}A_jB_k\,\varepsilon_{ilm}A_lB_m = (\delta_{jl}\delta_{km} - \delta_{jm}\delta_{kl})\,A_jB_kA_lB_m\,,$$

which is $A_jA_jB_kB_k - (A_jB_j)^2$. □

Proposition 13.96 (The cross product is an axial vector)**.** *Under an orthogonal coordinate transformation Equation (13.8) with matrix a_{ij}, the quantities Equation (13.225) obey*

$$C'_i = (\det a)\,a_{ij}\,C_j\,. \tag{13.229}$$

They therefore transform as the components of a Cartesian vector (Equation (13.21) with $r = 1$) under a proper *orthogonal transformation,* $\det a = +1$*, and pick up an extra minus sign under an improper one,* $\det a = -1$*: the cross product is a* pseudo-vector*, or* axial vector*, and not a vector.*

Rests on Proposition 13.6, Definition 13.5, and Equation (13.13).

Derivation. *Derives Proposition 13.96.*

Write Equation (13.31) with the index names (i, j, k) on the symbol and (s, l, m) on the free slots, $\varepsilon_{ijk}a_{is}a_{jl}a_{km} = \varepsilon_{slm}\det a$, and contract both sides with a_{ts}, summing over s. On the left the orthogonality condition Equation (13.13) gives $a_{is}a_{ts} = \delta_{it}$, so

$$\varepsilon_{tjk}\,a_{jl}a_{km} = (\det a)\,a_{ts}\,\varepsilon_{slm}\,. \tag{13.230}$$

Now C_{jk} of Equation (13.217) is an ordinary rank-2 tensor, so $C'_{jk} = a_{jl}a_{km}C_{lm}$ by Equation (13.22), and Equation (13.224) with Equation (13.230) gives

$$C'_i = \tfrac{1}{2}\varepsilon_{ijk}C'_{jk} = \tfrac{1}{2}\varepsilon_{ijk}a_{jl}a_{km}C_{lm} = (\det a)\,a_{is}\,\tfrac{1}{2}\varepsilon_{slm}C_{lm} = (\det a)\,a_{is}C_s\,,$$

which is Equation (13.229). The factor $\det a = \pm 1$ of Equation (13.15) is what distinguishes the two cases. □

Stokes' theorem for an antisymmetric tensor field

For an antisymmetric tensor field H_{ij} with associated vector H_i,

$$\frac{1}{2}\int_S H_{ij}\,\mathrm{d}f^{ij} = \oint_{\partial S} H_i\,\mathrm{d}x^i\,, \tag{13.231}$$

where $\mathrm{d}f^{ij}$ is the antisymmetric surface element of S and ∂S its boundary. The source states the relation without derivation; it is the flat three-dimensional instance of the general Stokes theorem for differential forms, and both are proved in the appendix.

Full derivation: Appendix A.29 (p. 2920).

Derives Equation (13.231).

Remark 13.97 (What the statement assumes). The relation between H_{ij} and H_i is left implicit in the source and is worth making explicit, because the identity is false under the other natural reading. It holds when H_{ij} is the *rotor tensor* of H_i,

$$H_{ij} = \partial_i H_j - \partial_j H_i \,, \tag{13.232}$$

and *not* when H_i is the dual pseudo-vector $\frac{1}{2}\varepsilon_{ijk}H_{jk}$ of Equation (13.224): with the latter reading the left-hand side of Equation (13.231) is the flux of H_i through S, which has no reason to equal its circulation around ∂S. With Equation (13.232), and writing the surface element as $\mathrm{d}f^{ij} = \varepsilon^{ijk}\mathrm{d}f_k$, the left-hand side of Equation (13.231) becomes $\frac{1}{2}\varepsilon^{ijk}(\partial_i H_j - \partial_j H_i)\mathrm{d}f_k = (\nabla \times \boldsymbol{H}) \cdot \mathrm{d}\boldsymbol{f}$, so in three dimensions Equation (13.231) is the classical Stokes theorem (Theorem 7.100) transcribed into tensor notation. In the language of Section 13.6 it is the statement $\int_S \mathrm{d}H = \oint_{\partial S} H$ for the one-form $H = H_i \mathrm{d}x^i$, whose exterior derivative has components Equation (13.232). In three dimensions nothing here is owed to the manifold machinery: the classical statement is derived outright in Theorem 7.100 from Green's theorem, with the scope of that derivation set out in Remark 7.102. The general form $\int_M \mathrm{d}\omega = \oint_{\partial M} \omega$ on a compact oriented manifold with boundary, of which Equation (13.231) is the flat three-dimensional instance, is proved in the appendix, and Equation (13.231) is deduced there from it under exactly the hypothesis Equation (13.232) made explicit above. The two derivations are independent — the appendix does not use Theorem 7.100 and Chapter 7 does not use the appendix — so they check each other rather than one resting on the other.

13.6 Differential forms: k-forms and k-vectors

13.6.1 k-forms and k-vectors

Definition 13.98 (k-form)**.** We say that a *k-form* is a totally antisymmetric tensor of type $(0, k)$. In particular, a covector is a 1-form.

Rests on Definitions 13.83 and 13.84.

Definition 13.99 (k-vector)**.** We say that a *k-vector* is a totally antisymmetric tensor of type $(k, 0)$. In particular, a vector is a 1-vector.

Rests on Definitions 13.81 and 13.84.

The set of all k-forms at P forms a vector space, which we denote $\Lambda^k_{M_n}(P)$. Of course,

$$\Lambda^k_{M_n}(P) \subset \underbrace{T^*_{M_n}(P) \otimes \cdots \otimes T^*_{M_n}(P)}_{k \text{ times}} \,. \tag{13.233}$$

Analogously, the set of all k-vectors forms a vector space, denoted ${}^*\Lambda^k_{M_n}(P)$, with

$${}^*\Lambda^k_{M_n}(P) \subset \underbrace{T_{M_n}(P) \otimes \cdots \otimes T_{M_n}(P)}_{k \text{ times}} \,. \tag{13.234}$$

Both assertions, and the dimension of the two spaces, follow from a single count.

Proposition 13.100 (The space of k-forms)**.** *Let M_n be an n-dimensional differentiable manifold and $P \in M_n$. The set $\Lambda^k_{M_n}(P)$ of k-forms at P is a vector subspace of the space of tensors of type $(0, k)$ at P, and*

$$\dim\left(\Lambda^k_{M_n}(P)\right) = \binom{n}{k} = \frac{n!}{(n-k)!\,k!} \quad \text{for } 0 \le k \le n \,, \tag{13.235}$$

while $\Lambda^k_{M_n}(P) = \{0\}$ for $k > n$. The same holds for the space ${}^*\Lambda^k_{M_n}(P)$ of k-vectors, with upper indices throughout.

Rests on Definitions 13.83, 13.84 and 13.98.

Proof. *Derives Proposition 13.100.*

The tensors of type $(0,k)$ at P form a vector space under pointwise addition and scaling of their components (Definition 13.84). Total antisymmetry is the requirement

$$\psi_{i_1\dots i_s\dots i_t\dots i_k} + \psi_{i_1\dots i_t\dots i_s\dots i_k} = 0 \qquad \text{for every pair } s \neq t\,, \tag{13.236}$$

a family of *linear* homogeneous conditions on the components; its solution set is therefore a linear subspace, which is the first assertion. Equivalently: if ψ and φ are totally antisymmetric so is $\alpha\psi + \beta\varphi$, because each condition Equation (13.236) is preserved by linear combination.

For the dimension, fix the coordinate cobasis $\{\mathrm{d}x^i\}$ of $T^*_{M_n}(P)$ (Equation (13.239)) and read a k-form through its components $\psi_{i_1\dots i_k}$. Total antisymmetry has two consequences. First, any component with a repeated index vanishes: exchanging the two equal slots leaves the component unchanged and, by Equation (13.236), reverses its sign. Second, any component whose indices are distinct is determined by the one whose indices are arranged in increasing order, since a permutation σ of the slots multiplies the component by $\operatorname{sgn}\sigma$:

$$\psi_{i_{\sigma(1)}\dots i_{\sigma(k)}} = \operatorname{sgn}(\sigma)\,\psi_{i_1\dots i_k}\,. \tag{13.237}$$

Hence the linear map sending ψ to the family of its *ordered* components $\psi_{i_1\dots i_k}$ with $i_1 < \dots < i_k$ is injective. It is also surjective: given arbitrary numbers assigned to the ordered index sets, Equation (13.237) extends them to a full family of components, and the extension is consistent because sgn is a homomorphism, so the result is a totally antisymmetric tensor with the prescribed ordered components. The map is therefore an isomorphism, and $\dim \Lambda^k_{M_n}(P)$ equals the number of strictly increasing k-tuples drawn from $\{1,\dots,n\}$, that is, the number of k-element subsets of an n-element set: $\binom{n}{k}$, which is Equation (13.235). If $k > n$ no such tuple exists — any k indices taken from n values must repeat one — so every component vanishes and the space is trivial. The argument used nothing but the antisymmetry of the components, so it applies verbatim to k-vectors. □

Remark 13.101. Once the wedge product is available (Definition 13.102), the isomorphism just constructed is realized by an explicit basis: the $\binom{n}{k}$ products $\mathrm{d}x^{i_1}\wedge\dots\wedge\mathrm{d}x^{i_k}$ with $i_1 < \dots < i_k$. Equation (13.235) is the dimension count used in Section 13.6.7 to pair k-forms with $(n-k)$-forms.

13.6.2 The basis of the cotangent space

Let $P \in M_n$, with M_n a differentiable manifold, and let $f : M_n \longrightarrow \mathbb{R}$ be a differentiable map. Consider the map

$$\begin{aligned} \mathrm{d}f &: T_{M_n}(P) \longrightarrow \mathbb{R} \\ V_P = \left.\frac{\mathrm{d}}{\mathrm{d}\lambda}\right|_{\lambda_0} &\longmapsto \mathrm{d}f(V_P) = \frac{\mathrm{d}f}{\mathrm{d}\lambda}(P)\,. \end{aligned}$$

We claim that $\mathrm{d}f(V_P) = V_P(f)$. Indeed, $\mathrm{d}f$ is defined as

$$\begin{aligned} \mathrm{d}f &= \frac{\partial f}{\partial x^i}\mathrm{d}x^i \\ \mathrm{d}f &= \frac{\partial f}{\partial x^i}\frac{\mathrm{d}x^i}{\mathrm{d}\lambda}\mathrm{d}\lambda \\ \frac{\mathrm{d}f}{\mathrm{d}\lambda}(P) &= \frac{\partial f}{\partial x^i}\frac{\mathrm{d}x^i}{\mathrm{d}\lambda} \\ \mathrm{d}f(V_P) &= V_P^i\,\partial_i f\,, \end{aligned}$$

and, recalling the discussion of the definition of a vector, $\{\partial_i\}$ is a basis of the tangent space at P. Thus

$$\mathrm{d}f(V_P) = V_P(f)\,. \tag{13.238}$$

From Equation (13.238), taking the vectors $\partial_i \in T_{M_n}(P)$ and $f = x^j$, we see that

$$\mathrm{d}x^j(\partial_i) = \delta^j{}_i\,, \tag{13.239}$$

so $\{\mathrm{d}x^i\}$ is the basis dual to $\{\partial_i\}$, that is,

$$T^*_{M_n}(P) = \left\langle \left\{ \mathrm{d}x^i \right\}_{i=1}^n \right\rangle\,,$$

and in particular the $\mathrm{d}x^i$ are covectors, that is, 1-forms. From this we see that for the set of all tensors of type $(0,k)$,

$$\underbrace{T^*_{M_n}(P) \otimes \cdots \otimes T^*_{M_n}(P)}_{k \text{ times}} = \left\langle \left\{ \mathrm{d}x^{i_1} \otimes \cdots \otimes \mathrm{d}x^{i_k} \right\} \right\rangle\,. \tag{13.240}$$

13.6.3 The wedge product

Wedge product and change of basis

Observe the following. The space of all 2-forms, $\Lambda^2_{M_n}(P)$, can be generated by the basis $\{\mathrm{d}x^i \otimes \mathrm{d}x^j\}$, but we must be careful: the direct product of two 2-forms is not necessarily a 4-form, since the result is a tensor of type $(0,4)$ that is not necessarily antisymmetric; that is, it can happen that

$$\psi \otimes \varphi \notin \Lambda^4_{M_n}(P)\,, \qquad \psi, \varphi \in \Lambda^2_{M_n}(P)\,.$$

For closure — referring to the antisymmetry property — we must define a new type of product. As we now show, this can be achieved by a change of basis. In the basis already discussed for the cotangent space,

$$\begin{aligned}
\psi &= \psi_{ij}\,\mathrm{d}x^i \otimes \mathrm{d}x^j \\
&= \frac{1}{2}\left(\psi_{ij} + \psi_{ij}\right)\mathrm{d}x^i \otimes \mathrm{d}x^j \\
&= \frac{1}{2}\left(\psi_{ij} - \psi_{ji}\right)\mathrm{d}x^i \otimes \mathrm{d}x^j \\
&= \frac{1}{2}\left(\psi_{ij}\,\mathrm{d}x^i \otimes \mathrm{d}x^j - \psi_{ji}\,\mathrm{d}x^i \otimes \mathrm{d}x^j\right) \\
&= \frac{1}{2}\left(\psi_{ij}\,\mathrm{d}x^i \otimes \mathrm{d}x^j - \psi_{ij}\,\mathrm{d}x^j \otimes \mathrm{d}x^i\right) \\
&= \frac{1}{2}\psi_{ij}\left(\mathrm{d}x^i \otimes \mathrm{d}x^j - \mathrm{d}x^j \otimes \mathrm{d}x^i\right)\,,
\end{aligned}$$

where the antisymmetry of ψ_{ij} was used in the second step.

Definition 13.102 (Wedge product)**.** As we have seen, the $\mathrm{d}x^i$ are 1-forms. We define the *wedge product* as the binary operation on the basis elements

$$\begin{aligned}
\wedge : \Lambda^1_{M_n}(P) \times \Lambda^1_{M_n}(P) &\longrightarrow \Lambda^2_{M_n}(P) \\
(\mathrm{d}x^i, \mathrm{d}x^j) &\longmapsto \mathrm{d}x^i \wedge \mathrm{d}x^j = \mathrm{d}x^i \otimes \mathrm{d}x^j - \mathrm{d}x^j \otimes \mathrm{d}x^i\,.
\end{aligned}$$

Rests on Definition 13.98 and Equation (13.239).

Therefore

$$\psi = \frac{1}{2}\psi_{ij}\,\mathrm{d}x^i \wedge \mathrm{d}x^j\,.$$

In short, we have performed the change of basis

$$\left\{ \mathrm{d}x^i \otimes \mathrm{d}x^j \right\}_{i,j}^n \;\longrightarrow\; \left\{ \mathrm{d}x^i \wedge \mathrm{d}x^j \right\}_{i,j}^n\,.$$

Antisymmetry

From the definition of the wedge product it is immediate that

$$\mathrm{d}x^i \wedge \mathrm{d}x^j = -\mathrm{d}x^j \wedge \mathrm{d}x^i \,, \tag{13.241}$$

and clearly, taking $j = i$ (no sum),

$$\mathrm{d}x^i \wedge \mathrm{d}x^i = 0 \,. \tag{13.242}$$

Powers of a k-form

We introduce the notation

$$\underbrace{\psi \wedge \cdots \wedge \psi}_{m \text{ times}} = \psi^m \,,$$

which we call the m-th *power* of a k-form, with respect to the wedge product.

Commutation rule

Having defined the wedge product explicitly, we can analyse its commutativity. Let ψ be a k-form and φ an l-form. In the basis discussed above,

$$\begin{aligned}
\psi \wedge \varphi &= \left(\frac{1}{k!}\psi_{i_1\ldots i_k}\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}\right) \wedge \left(\frac{1}{l!}\varphi_{j_1\ldots j_l}\mathrm{d}x^{j_1} \wedge \cdots \wedge \mathrm{d}x^{j_l}\right) \\
&= \frac{1}{k!\,l!}\psi_{i_1\ldots i_k}\varphi_{j_1\ldots j_l}\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k} \wedge \mathrm{d}x^{j_1} \wedge \cdots \wedge \mathrm{d}x^{j_l} \\
\text{[Equation (13.241)]} \quad &= -\frac{1}{k!\,l!}\psi_{i_1\ldots i_k}\varphi_{j_1\ldots j_l}\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_{k-1}} \wedge \mathrm{d}x^{j_1} \wedge \mathrm{d}x^{i_k} \wedge \mathrm{d}x^{j_2} \wedge \cdots \wedge \mathrm{d}x^{j_l} \\
&= (-1)^k \frac{1}{k!\,l!}\psi_{i_1\ldots i_k}\varphi_{j_1\ldots j_l}\,\mathrm{d}x^{j_1} \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k} \wedge \cdots \wedge \mathrm{d}x^{j_l} \\
&= \left((-1)^k\right)^l \frac{1}{k!\,l!}\psi_{i_1\ldots i_k}\varphi_{j_1\ldots j_l}\,\mathrm{d}x^{j_1} \wedge \cdots \wedge \mathrm{d}x^{j_l} \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k} \\
&= (-1)^{kl}\left(\frac{1}{l!}\varphi_{j_1\ldots j_l}\mathrm{d}x^{j_1} \wedge \cdots \wedge \mathrm{d}x^{j_l}\right) \wedge \left(\frac{1}{k!}\psi_{i_1\ldots i_k}\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}\right) .
\end{aligned}$$

We thus obtain the commutation rule

$$\psi \wedge \varphi = (-1)^{kl}\,\varphi \wedge \psi \,. \tag{13.243}$$

Various properties follow from this relation by simple arithmetic. For example, consider the case $\varphi = \psi$ with k odd. Since the square of an odd number is odd,

$$\begin{aligned}
\psi^2 &= (-1)^{k^2}\psi^2 \\
&= -\psi^2 \,,
\end{aligned}$$

hence

$$\psi^2 = 0 \,. \tag{13.244}$$

13.6.4 The exterior derivative

Definition 13.103 (Exterior derivative)**.** We define the *exterior derivative* as the map

$$\begin{aligned}
\mathrm{d} : \Lambda^k_{M_n}(P) &\longrightarrow \Lambda^{k+1}_{M_n}(P) \\
\psi &\longmapsto \mathrm{d}\psi = \frac{1}{k!}\,\partial_j\psi_{i_1\ldots i_k}\,\mathrm{d}x^j \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k} \,.
\end{aligned}$$

Rests on Definitions 13.98 and 13.102.

It is clear that the exterior derivative is a linear operator, owing to the linearity of the ordinary derivative ∂_j.

Nilpotency

Let ψ be a k-form. Then

$$\mathrm{d}(\mathrm{d}\psi) = \frac{1}{k!}\,\partial_i\partial_j\psi_{i_1\dots i_k}\,\mathrm{d}x^i\wedge\mathrm{d}x^j\wedge\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}\,.$$

Now, if $\psi_{i_1\dots i_k}$ is of sufficiently high differentiability class, $\partial_i\partial_j$ is symmetric in i and j, while $\mathrm{d}x^i\wedge\mathrm{d}x^j$ is antisymmetric in those indices; hence their contraction vanishes. Thus

$$\mathrm{d}(\mathrm{d}\psi) = 0\,. \tag{13.245}$$

Leibniz rule

Let ψ be a k-form and φ an l-form. Then

$$\begin{aligned}
\mathrm{d}(\psi\wedge\varphi) &= \mathrm{d}\left(\left(\frac{1}{k!}\psi_{i_1\dots i_k}\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}\right)\wedge\left(\frac{1}{l!}\varphi_{j_1\dots j_l}\mathrm{d}x^{j_1}\wedge\cdots\wedge\mathrm{d}x^{j_l}\right)\right)\\
&= \frac{1}{k!\,l!}\,\partial_j\left(\psi_{i_1\dots i_k}\varphi_{j_1\dots j_l}\right)\mathrm{d}x^j\wedge\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}\wedge\mathrm{d}x^{j_1}\wedge\cdots\wedge\mathrm{d}x^{j_l}\\
&= \frac{1}{k!\,l!}\left(\partial_j(\psi_{i_1\dots i_k})\varphi_{j_1\dots j_l}+\psi_{i_1\dots i_k}\partial_j\varphi_{j_1\dots j_l}\right)\mathrm{d}x^j\wedge\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}\wedge\mathrm{d}x^{j_1}\wedge\cdots\wedge\mathrm{d}x^{j_l}\\
&= \mathrm{d}\psi\wedge\varphi+\frac{1}{k!\,l!}\,\psi_{i_1\dots i_k}\partial_j\varphi_{j_1\dots j_l}\,\mathrm{d}x^j\wedge\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}\wedge\mathrm{d}x^{j_1}\wedge\cdots\wedge\mathrm{d}x^{j_l}\\
&= \mathrm{d}\psi\wedge\varphi+\frac{(-1)^k}{k!\,l!}\,\psi_{i_1\dots i_k}\partial_j\varphi_{j_1\dots j_l}\,\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}\wedge\mathrm{d}x^j\wedge\mathrm{d}x^{j_1}\wedge\cdots\wedge\mathrm{d}x^{j_l}\,,
\end{aligned}$$

where in the last step the 1-form $\mathrm{d}x^j$ was carried through the k factors $\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_k}$ at the cost of a factor $(-1)^k$, by Equation (13.241). We conclude the identity

$$\mathrm{d}(\psi\wedge\varphi) = \mathrm{d}\psi\wedge\varphi+(-1)^k\,\psi\wedge\mathrm{d}\varphi\,, \tag{13.246}$$

known as the *Leibniz rule* for differential forms.

Remark 13.104. All the properties of the exterior derivative mentioned above are consequences of the properties of the ordinary derivative. This is because this special derivative is only a definition designed to deal with differential forms; the essential object is the derivative we all know.

13.6.5 Closed and exact forms

Definition 13.105 (Closed form)**.** Let ψ be a k-form. We say that ψ is *closed* if and only if

$$\mathrm{d}\psi = 0\,.$$

Rests on Definitions 13.98 and 13.103.

Definition 13.106 (Exact form)**.** Let $\psi\in\Lambda^k_{M_n}(P)$ be a k-form. We say that ψ is *exact* if and only if there exists a $(k-1)$-form $\varphi\in\Lambda^{k-1}_{M_n}(P)$ such that

$$\psi = \mathrm{d}\varphi\,.$$

Rests on Definitions 13.98 and 13.103.

Lemma 13.107 (Poincaré lemma)**.** *If a k-form is exact, then it is closed.*
Rests on Definitions 13.105 and 13.106 and Equation (13.245).

Derivation. Derives Lemma 13.107.

Immediate from the nilpotency of the exterior derivative, Equation (13.245): if $\psi=\mathrm{d}\varphi$ then $\mathrm{d}\psi=\mathrm{d}(\mathrm{d}\varphi)=0$. □

The converse is not automatic, and its validity depends on the topology of the domain. Two statements settle the matter for everything this treatise needs: on a star-shaped domain the converse holds outright, and in general it fails, with a completely explicit counterexample.

Definition 13.108 (Star-shaped domain)**.** An open set $U \subseteq \mathbb{R}^n$ is *star-shaped* about $x_0 \in U$ if and only if the segment from x_0 to x lies in U for every $x \in U$. A ball is star-shaped about its centre, and every point of every open set has a star-shaped neighbourhood.

Rests on Definition 6.2 and Equation (6.12).

Theorem 13.109 (Converse of the Poincaré lemma on a star-shaped domain)**.** *Let $U \subseteq \mathbb{R}^n$ be open and star-shaped about the origin, and let ψ be a closed k-form on U with $k \geq 1$. Then ψ is exact: $\psi = \mathrm{d}\varphi$ with φ the $(k-1)$-form*

$$\varphi_{i_2 \ldots i_k}(x) = \int_0^1 t^{k-1}\, x^j\, \psi_{j\, i_2 \ldots i_k}(tx)\, \mathrm{d}t\,. \tag{13.247}$$

Consequently a closed form is locally *exact on any manifold, each point having a chart domain that may be taken star-shaped.*

Rests on Definitions 13.105, 13.106 and 13.108.

Proof. *Derives Theorem 13.109.*

Write $h\psi$ for the $(k-1)$-form Equation (13.247); it is defined because $tx \in U$ for $t \in [0,1]$, and it is totally antisymmetric in $i_2, \ldots, i_k$ because ψ is. We prove the *homotopy identity*

$$\mathrm{d}\,(h\psi) + h\,(\mathrm{d}\psi) = \psi\,, \tag{13.248}$$

from which the theorem follows by putting $\mathrm{d}\psi = 0$.

In the convention of Definition 13.103 the components of the exterior derivative of an l-form χ are $(\mathrm{d}\chi)_{i_1 \ldots i_{l+1}} = (l+1)\, \partial_{[i_1} \chi_{i_2 \ldots i_{l+1}]}$, the brackets denoting total antisymmetrization. Differentiating Equation (13.247) under the integral sign, the derivative falls either on the explicit factor x^j or on the argument tx:

$$[\mathrm{d}\,(h\psi)]_{i_1 \ldots i_k} = k \int_0^1 t^{k-1} \psi_{i_1 \ldots i_k}(tx)\, \mathrm{d}t + k \int_0^1 t^k x^j \left(\partial_{[i_1} \psi_{|j|\, i_2 \ldots i_k]}\right)(tx)\, \mathrm{d}t\,, \tag{13.249}$$

the first term arising from $\delta^j_{[i_1} \psi_{|j| i_2 \ldots i_k]} = \psi_{[i_1 i_2 \ldots i_k]} = \psi_{i_1 \ldots i_k}$, and the vertical bars excluding j from the antisymmetrization. For the second term of Equation (13.248),

$$[h\,(\mathrm{d}\psi)]_{i_1 \ldots i_k} = (k+1) \int_0^1 t^k x^j \left(\partial_{[j} \psi_{i_1 \ldots i_k]}\right)(tx)\, \mathrm{d}t\,. \tag{13.250}$$

Separating the first slot from a total antisymmetrization over $k+1$ indices gives the algebraic identity

$$(k+1)\, \partial_{[j} \psi_{i_1 \ldots i_k]} = \partial_j \psi_{i_1 \ldots i_k} - k\, \partial_{[i_1} \psi_{|j|\, i_2 \ldots i_k]}\,, \tag{13.251}$$

which for $k = 1$ reads $\partial_j \psi_i - \partial_i \psi_j = \partial_j \psi_i - \partial_i \psi_j$ and in general is the statement that antisymmetrizing over $k+1$ slots may be performed by holding the first slot and antisymmetrizing the k transpositions that move j into it. Adding Equations (13.249) and (13.250) through Equation (13.251), the two terms carrying $\partial_{[i_1} \psi_{|j| \ldots]}$ cancel and

$$[\mathrm{d}(h\psi) + h(\mathrm{d}\psi)]_{i_1 \ldots i_k} = \int_0^1 \left[k\, t^{k-1} \psi_{i_1 \ldots i_k}(tx) + t^k x^j \left(\partial_j \psi_{i_1 \ldots i_k}\right)(tx)\right] \mathrm{d}t\,.$$

By the chain rule the bracket is exactly $\frac{\mathrm{d}}{\mathrm{d}t}\left[t^k \psi_{i_1 \ldots i_k}(tx)\right]$, so the integral is $\left[t^k \psi_{i_1 \ldots i_k}(tx)\right]_0^1 = \psi_{i_1 \ldots i_k}(x)$, the lower limit vanishing because $k \geq 1$. That is Equation (13.248). □

Example 13.110 (Closed but not exact: the angle form)**.** On $U = \mathbb{R}^2 \setminus \{0\}$, which is not star-shaped, put

$$\alpha = \frac{-y\,\mathrm{d}x + x\,\mathrm{d}y}{x^2 + y^2}\,. \tag{13.252}$$

It is closed: with $\alpha_1 = -y/(x^2+y^2)$ and $\alpha_2 = x/(x^2+y^2)$, both $\partial_1\alpha_2$ and $\partial_2\alpha_1$ equal $(y^2-x^2)/(x^2+y^2)^2$, so $\mathrm{d}\alpha = 0$. It is not exact: on the unit circle $x = \cos s$, $y = \sin s$ one computes $\oint \alpha = \int_0^{2\pi} \mathrm{d}s = 2\pi$, whereas an exact form $\mathrm{d}f$ integrates to zero around any closed curve, since $\oint \mathrm{d}f = \int_0^{2\pi} \frac{\mathrm{d}}{\mathrm{d}s} f\,(\gamma(s))\,\mathrm{d}s = f(\gamma(2\pi)) - f(\gamma(0)) = 0$ by the fundamental theorem of calculus (Theorem 7.43). So the converse of Lemma 13.107 genuinely fails, and what fails is topological: α is exact on any star-shaped subset of U — there it is d of a branch of the polar angle — but no branch is single-valued on the whole punctured plane.

Rests on Theorem 13.109 and Definition 13.106.

The symmetric-tensor analogue of Theorem 13.109 is the result that decides which strain fields of a continuum are realisable by an actual displacement, and it follows from the antisymmetric one by applying it twice.

Definition 13.111 (Linearised curvature of a symmetric field)**.** For a C^2 symmetric field $e_{ij} = e_{ji}$ on an open set of $\mathbb{R}^n$, put

$$\mathcal{R}_{jlkm}(e) = \partial_l\partial_k e_{jm} + \partial_j\partial_m e_{lk} - \partial_l\partial_m e_{jk} - \partial_j\partial_k e_{lm}\,. \tag{13.253}$$

It is antisymmetric under $j \leftrightarrow l$ and under $k \leftrightarrow m$ and symmetric under the exchange of the two pairs — the index symmetries of the curvature tensor (Proposition 13.154), which is what it is: the linearisation of Equation (13.95) about a flat metric in the perturbation e. The equations $\mathcal{R}(e) = 0$ are the *Saint-Venant compatibility conditions*.

Rests on Definition 13.91 and Proposition 13.154.

Theorem 13.112 (Symmetric analogue of the converse Poincaré lemma)**.** *Let $U \subseteq \mathbb{R}^n$ be open and star-shaped and let e_{ij} be a symmetric field of class C^2 on U. Then*

$$\mathcal{R}_{jlkm}(e) = 0 \quad \text{on } U \quad\Longleftrightarrow\quad e_{jk} = \tfrac{1}{2}\,(\partial_j u_k + \partial_k u_j) \quad \text{for some } C^3 \text{ field } u\,. \tag{13.254}$$

The field u is unique up to $u_j \mapsto u_j + a_j + \Omega_{jk}x^k$ with a constant and Ω a constant antisymmetric array — an infinitesimal rigid motion.

Rests on Theorem 13.109, Definition 13.111, and Proposition 7.73.

Proof. *Derives Theorem 13.112.*

Necessity. Substituting $e_{jm} = \frac{1}{2}\,(\partial_j u_m + \partial_m u_j)$ into Equation (13.253) and commuting partial derivatives freely (Proposition 7.73), the eight third derivatives cancel in four pairs: $\partial_l\partial_k\partial_j u_m$ against $\partial_j\partial_k\partial_l u_m$, $\partial_l\partial_k\partial_m u_j$ against $\partial_l\partial_m\partial_k u_j$, $\partial_j\partial_m\partial_l u_k$ against $\partial_l\partial_m\partial_j u_k$, and $\partial_j\partial_m\partial_k u_l$ against $\partial_j\partial_k\partial_m u_l$.

Sufficiency, first application. For each fixed pair (j,l) consider the one-form on U

$$\alpha^{(jl)} = (\partial_l e_{jk} - \partial_j e_{lk})\,\mathrm{d}x^k\,, \tag{13.255}$$

antisymmetric in (j,l). Its exterior derivative has components $\partial_m\,(\partial_l e_{jk} - \partial_j e_{lk}) - \partial_k\,(\partial_l e_{jm} - \partial_j e_{lm}) = -\mathcal{R}_{jlkm}(e)$, which vanishes by hypothesis. By Theorem 13.109 there are functions w_{jl} with $\partial_k w_{jl} = \partial_l e_{jk} - \partial_j e_{lk}$; since the right-hand side changes sign under $j \leftrightarrow l$, the function $w_{jl} + w_{lj}$ has vanishing gradient and is therefore a constant on the connected set U (Corollary 7.36), which may be subtracted, so w may be taken antisymmetric.

Second application. For each fixed j consider $\beta^{(j)} = (e_{jk} + w_{jk})\,\mathrm{d}x^k$. Its exterior derivative has components

$$\begin{aligned}\partial_m\,(e_{jk} + w_{jk}) - \partial_k\,(e_{jm} + w_{jm}) &= \partial_m e_{jk} - \partial_k e_{jm} + (\partial_k e_{jm} - \partial_j e_{km}) - (\partial_m e_{jk} - \partial_j e_{mk}) \\ &= \partial_j\,(e_{mk} - e_{km}) = 0\end{aligned}$$

by the symmetry of e. So $\beta^{(j)} = \mathrm{d}u_j$ for some function u_j, again by Theorem 13.109; that is $\partial_k u_j = e_{jk} + w_{jk}$, and symmetrizing in (j,k) kills w and returns $e_{jk} = \tfrac{1}{2}(\partial_j u_k + \partial_k u_j)$. The field u is C^3 because its first derivatives are C^2.

Uniqueness. If two solutions differ by u with $\partial_j u_k + \partial_k u_j = 0$, then

$$\partial_k \partial_l u_j = -\partial_k \partial_j u_l = -\partial_j \partial_k u_l = \partial_j \partial_l u_k = \partial_l \partial_j u_k = -\partial_l \partial_k u_j = -\partial_k \partial_l u_j \,,$$

so all second derivatives vanish and u is affine, $u_j = a_j + \Omega_{jk} x^k$; the condition then forces $\Omega_{jk} + \Omega_{kj} = 0$. □

Proposition 13.113 (Product of a closed and an exact form)**.** *If ψ is a closed form and φ is an exact form, then $\psi \wedge \varphi$ is a closed form.*

Rests on Definitions 13.105 and 13.106, Lemma 13.107, and Equation (13.246)*.*

Derivation. *Derives Proposition 13.113.*

Let ψ be a closed k-form and φ an exact l-form. By the Poincaré lemma, φ is moreover closed; then by the Leibniz rule

$$\mathrm{d}(\psi \wedge \varphi) = \mathrm{d}\psi \wedge \varphi + (-1)^k \, \psi \wedge \mathrm{d}\varphi$$

vanishes. □

13.6.6 The volume form

Definition 13.114 (Volume form)**.** We define the *volume form* as

$$\eta = \sqrt{g}\; \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \,, \tag{13.256}$$

where g is the determinant of the metric tensor (Section 13.7).

Rests on Definitions 13.102 and 13.117.

η is in fact an n-form, since

$$\begin{aligned} \eta &= \sqrt{g}\; \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \\ &= \frac{\sqrt{g}}{n!}\, \delta^{1\ldots n}_{i_1 \ldots i_n}\, \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_n} \\ &= \frac{1}{n!} \sqrt{g}\; \varepsilon_{i_1 \ldots i_n}\, \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_n} \,, \end{aligned}$$

where $\delta^{1\ldots n}_{i_1\ldots i_n}$ is the generalised Kronecker delta and $\varepsilon_{i_1\ldots i_n}$ the Levi-Civita symbol. Hence η is an n-form with components

$$\eta_{i_1 \ldots i_n} = \sqrt{g}\; \varepsilon_{i_1 \ldots i_n} \,. \tag{13.257}$$

Remark 13.115. For a metric of signature (p,q) with $D = p+q$ the determinant g carries the sign $(-1)^q$, and the volume form is written with $\sqrt{|g|}$ in place of $\sqrt{g}$; the source treats the Riemannian case $q = 0$.

13.6.7 The Hodge dual

Consider a differentiable manifold M_n. Recall that at each point P of the manifold there exist a tangent space $T_{M_n}(P)$ and a cotangent space $T^*_{M_n}(P)$, which are n-dimensional vector spaces, and that the set of all k-forms is a vector space $\Lambda^k_{M_n}(P)$ with (Proposition 13.100)

$$\dim\left(\Lambda^k_{M_n}(P)\right) = \frac{n!}{(n-k)!\,k!} \,,$$

and likewise for the set of all k-vectors, ${}^*\Lambda^k_{M_n}(P)$. Owing to the fact that

$$\binom{n}{k} = \binom{n}{n-k},$$

we can associate four $\binom{n}{k}$-dimensional vector spaces with each point of the manifold: the space of k-forms, the space of $(n-k)$-forms, the space of k-vectors, and the space of $(n-k)$-vectors. Recalling that two vector spaces of the same dimension are isomorphic, the four spaces mentioned are isomorphic; that is, we can guarantee that isomorphisms exist among them.

Which are the isomorphisms that move us between these spaces? The answer to this question is not in itself essential: we define the maps and prove their quality as isomorphisms when necessary.

Definition 13.116 (Hodge dual)**.** Let T be a k-vector with components $T^{i_1\dots i_k}$. We define the *Hodge operator* (*duality operator*, or *star operator*) as

$$\begin{aligned} *\,:\, {}^*\Lambda^k_{M_n}(P) &\longrightarrow \Lambda^{n-k}_{M_n}(P) \\ T &\longmapsto {}^*T = \frac{1}{(n-k)!}\underbrace{\frac{1}{k!}\,\eta_{i_1\dots i_k i_{k+1}\dots i_n} T^{i_1\dots i_k}}_{{}^*T_{i_{k+1}\dots i_n}}\,\mathrm{d}x^{i_{k+1}}\wedge\cdots\wedge\mathrm{d}x^{i_n}\,. \end{aligned}$$

Rests on Definitions 13.98, 13.99 and 13.114.

We see that the change of components is given by

$${}^*T_{i_{k+1}\dots i_n} = \frac{1}{k!}\,\eta_{i_1\dots i_n} T^{i_1\dots i_k}\,. \tag{13.258}$$

13.7 The metric tensor

A manifold, as defined so far, knows nearness but not distance: the charts of Section 13.4.1 let us differentiate, but nothing yet measures lengths, areas, volumes, or angles. As the source manuscript asks: the ambient $\mathbb{R}^N$ of an embedded manifold carries the Euclidean metric, but can that law serve an arbitrary manifold? It cannot — a vector on a manifold lives in the tangent space at its point (Section 13.5.1), and each point carries its own. The remedy is a field that equips every tangent space with its own scalar product.

Definition 13.117 (Metric tensor of signature (p,q))**.** A *metric* on a D-dimensional manifold M, $D = p+q$, is a $(0,2)$ tensor field $g = g_{\mu\nu}\,\mathrm{d}x^\mu \otimes \mathrm{d}x^\nu$ that is symmetric, $g_{\mu\nu} = g_{\nu\mu}$, and nondegenerate, $\det g_{\mu\nu} \neq 0$, and whose matrix at every point has p positive and q negative eigenvalues (*signature* (p,q); the counts are point-independent on a connected manifold by continuity of the eigenvalues). The inverse metric $g^{\mu\nu}$ is defined by $g^{\mu\lambda}g_{\lambda\nu} = \delta^\mu{}_\nu$. The pair (M,g) is a *pseudo-Riemannian manifold* (Riemannian iff $q=0$).

Rests on Definitions 13.48 and 13.84.

Definition 13.118 (Line element)**.** The *invariant line element* of a displacement $\mathrm{d}x^\mu$ is

$$\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu \mathrm{d}x^\nu\,. \tag{13.259}$$

It is a scalar under general coordinate transformations (Section 13.5.2), because the transformation Jacobians of $g_{\mu\nu}$ and of the two $\mathrm{d}x$'s cancel. For $q \neq 0$ the line element is not positive definite: displacements split into those with $\mathrm{d}s^2 > 0$, $\mathrm{d}s^2 = 0$, and $\mathrm{d}s^2 < 0$, the geometric germ of the causal structure of Part IV.

Rests on Definition 13.117.

The metric and its inverse convert between the tangent and cotangent spaces — they *lower and raise indices*, $V_\mu = g_{\mu\nu}V^\nu$, $\omega^\mu = g^{\mu\nu}\omega_\nu$ — and define the invariant volume element

$$\mathrm{d}V = \sqrt{|g|}\,\mathrm{d}^D x\,, \qquad g \equiv \det g_{\mu\nu}\,, \tag{13.260}$$

whose invariance follows from the Jacobian transformation of $\det g$ (Section 13.5.3: $\sqrt{|g|}$ is a scalar density of weight one).

Definition 13.119 (Induced metric)**.** Let $\Sigma \subset M$ be a submanifold described by the embedding $x^\mu(\xi^i)$. The *induced metric* on Σ is the pullback of g,

$$h_{ij} = g_{\mu\nu}\,\frac{\partial x^\mu}{\partial \xi^i}\,\frac{\partial x^\nu}{\partial \xi^j}\,, \tag{13.261}$$

generalizing the first fundamental form of surface theory (Section 13.3.2), which is exactly Equation (13.261) for a surface in Euclidean $\mathbb{R}^3$.

Rests on Definition 13.117.

13.7.1 Isometries

Definition 13.120 (Isometry)**.** A diffeomorphism $f : M \longrightarrow M$ is an *isometry* iff the metric at P has the same value as the metric at $f(P)$, compared by pulling back:

$$f^* g_{f(P)} = g_P\,. \tag{13.262}$$

Rests on Definition 13.117.

Proposition 13.121 (The isometry group)**.** *The set of all isometries of (M, g) forms a group under composition, denoted* $\mathrm{Isom}(M, g)$.

Rests on Definition 13.120.

Proof. *Derives Proposition 13.121.*

Closure: for isometries f and h, the pullback of a composition is the composition of pullbacks in reverse order (Section 13.5.1), so

$$\begin{aligned}(h \circ f)^* g_{(h\circ f)(P)} &= f^*\left(h^* g_{h(f(P))}\right)\\ &= f^* g_{f(P)}\\ &= g_P\,.\end{aligned} \tag{13.263}$$

Associativity is inherited from composition of maps. The identity id_M is an isometry, $\mathrm{id}_M^*\, g_P = g_P$. Inverses: an isometry is a diffeomorphism, so f^{-1} exists, and applying $(f^{-1})^*$ to both sides of Equation (13.262) at the point $f^{-1}(Q)$ gives $(f^{-1})^* g_{f^{-1}(Q)} = g_Q$, i.e. f^{-1} is an isometry. □

The infinitesimal version of Equation (13.262) — the Killing equation — is derived in Section 13.9.2, and the manifolds whose isometry groups are as large as possible are classified in Section 13.13: they are precisely the flat, anti-de Sitter, and de Sitter geometries of Table 13.1.

13.8 Vielbein and local frames

The coordinate basis ∂_μ diagonalizes differentiation, not the metric. For the physics of Part V — and for spinors, which transform under no representation of the diffeomorphism group — one needs at each point a basis in which the metric is the flat η_{ab} of Equation (13.1). Indices $a, b, \ldots$ follow the conventions of Section 13.1.

Definition 13.122 (Vielbein)**.** A *vielbein* (frame field) is a set of D one-forms $e^a = e^a{}_\mu \, \mathrm{d}x^\mu$ such that

$$g_{\mu\nu} = \eta_{ab} \, e^a{}_\mu \, e^b{}_\nu \, . \tag{13.264}$$

The dual frame vectors $e_a = e_a{}^\mu \partial_\mu$ satisfy $e^a{}_\mu e_b{}^\mu = \delta^a_b$ and $e^a{}_\mu e_a{}^\nu = \delta_\mu{}^\nu$; frame indices are raised and lowered with η_{ab}, coordinate indices with $g_{\mu\nu}$, and the vielbein converts one type into the other, $V^a = e^a{}_\mu V^\mu$. Taking determinants of Equation (13.264), $e \equiv \det e^a{}_\mu = \sqrt{|g|}$ (up to orientation), so the volume element Equation (13.260) is $e \, \mathrm{d}^D x$.

Rests on Definition 13.117 and Equation (13.1).

Proposition 13.123 (Local Lorentz freedom)**.** *The vielbein of a given metric is unique up to*

$$e^a{}_\mu(x) \longrightarrow \Lambda^a{}_b(x) \, e^b{}_\mu(x) \, , \qquad \Lambda^a{}_c \, \Lambda^b{}_d \, \eta_{ab} = \eta_{cd} \, , \tag{13.265}$$

a local *(point-dependent) transformation with values in the Lorentz group* $\mathrm{L}_{p+q} = \mathrm{SO}(p,q)$ *of Table 13.1.*

Rests on Definition 13.122.

Proof. *Derives Proposition 13.123.*

If e and $\tilde{e}$ both satisfy Equation (13.264), then $\Lambda^a{}_b \equiv \tilde{e}^a{}_\mu e_b{}^\mu$ maps one to the other, and substituting into Equation (13.264) for both sides forces $\eta_{ab}\Lambda^a{}_c\Lambda^b{}_d = \eta_{cd}$, which is the defining condition of $\mathrm{O}(p,q)$; continuity and choice of orientation restrict to $\mathrm{SO}(p,q)$. Conversely any such $\Lambda(x)$ preserves Equation (13.264). □

Geometry in the frame formulation therefore has *two* covariances, acting on different indices per Section 13.1: diffeomorphisms on $\mu, \nu, \ldots$ and local Lorentz transformations on $a, b, \ldots$ — the structure that makes gravity resemble a gauge theory (Section 13.12).

Unlike the coordinate basis, the frame e_a is in general *anholonomic*: its Lie brackets (Section 13.9) do not vanish,

$$[e_a, e_b] = C^c{}_{ab} \, e_c \, , \qquad C^c{}_{ab} = e_a{}^\mu e_b{}^\nu \left(\partial_\mu e^c{}_\nu - \partial_\nu e^c{}_\mu \right) \, , \tag{13.266}$$

and the $C^c{}_{ab}$ measure the failure of the frame to come from coordinates.

13.9 Flows, the Lie derivative, and Killing vectors

A vector field $\boldsymbol{\xi} = \xi^\mu \partial_\mu$ generates a flow: the one-parameter family of diffeomorphisms ϕ_t obtained by sliding every point along the integral curves $\frac{\mathrm{d}x^\mu}{\mathrm{d}t} = \xi^\mu(x)$. The *Lie derivative* measures the rate of change of a tensor field as it is dragged by this flow, comparing the dragged tensor with the actual one:

$$\mathcal{L}_\xi T = \lim_{t \to 0} \frac{\phi_t^* \, T_{\phi_t(P)} - T_P}{t} \, . \tag{13.267}$$

It requires no metric and no connection — only the differentiable structure.

That opening sentence asserts something, and the assertion is a theorem: that a smooth vector field *has* a flow, that the flow is unique, that it is smooth in the point as well as in the parameter, and that each ϕ_t is a diffeomorphism. Everything in this section, and the Frobenius theorem of Section 13.9.1 below, rests on it.

Definition 13.124 (Integral curve; complete vector field)**.** An *integral curve* of a smooth vector field X on M_n is a smooth curve γ with $\dot{\gamma}(t) = X\left(\gamma(t)\right)$ for every t in its domain. X is *complete* if and only if through every point there is an integral curve defined on all of $\mathbb{R}$.

Rests on Definitions 13.71 and 13.82.

Theorem 13.125 (Existence, uniqueness and smoothness of the flow)**.** *Let X be a smooth vector field on M_n and $P \in M_n$. Then there are an open set $U \ni P$, a number $\varepsilon > 0$ and a smooth map*

$$\phi :]-\varepsilon, \varepsilon[\times U \longrightarrow M_n , \qquad \frac{\mathrm{d}}{\mathrm{d}t}\phi_t(x) = X\left(\phi_t(x)\right) , \qquad \phi_0 = \mathrm{id}_U , \tag{13.268}$$

such that each ϕ_t is a diffeomorphism of U onto an open subset of M_n. Two integral curves of X that agree at one parameter value agree on the intersection of their domains; consequently

$$\phi_{t+s} = \phi_t \circ \phi_s \tag{13.269}$$

wherever both sides are defined, and $\phi_t^{-1} = \phi_{-t}$. If X is complete, ϕ is defined on all of $\mathbb{R} \times M_n$ and $t \mapsto \phi_t$ is a one-parameter group of diffeomorphisms of M_n.

Rests on Definition 13.124 and Theorems 9.8 and A.74.

Proof. *Derives Theorem 13.125.*

Work in a chart around P, in which X becomes a smooth map $X : V \longrightarrow \mathbb{R}^n$ on an open $V \subseteq \mathbb{R}^n$ and P becomes x_0. Smoothness makes X bounded and Lipschitz on any closed box inside V (Theorem 7.24 and Theorem 7.35 applied componentwise), so Theorem 9.8 gives a unique integral curve c with $c(0) = x_0$ on some interval, and we fix $T > 0$ with $c\left([-T, T]\right) \subset V$.

Reduction to a field that vanishes at a point. For $s \in [0,1]$ and z in the open set $O = \{z \in \mathbb{R}^n \mid z + c(Ts) \in V \text{ for all } s \in [0,1]\}$, which contains 0 because $c([0,T])$ is compact and V open, put

$$Z_s(z) = T\left[X\left(z + c(Ts)\right) - X\left(c(Ts)\right)\right] . \tag{13.270}$$

This is smooth in (s, z) and $Z_s(0) = 0$ for every s, which is exactly the hypothesis of Theorem A.74. That theorem supplies a ball $B \ni 0$ and a smooth $\zeta : [0,1] \times B \longrightarrow O$ with $\partial_s \zeta_s(z) = Z_s\left(\zeta_s(z)\right)$, $\zeta_0 = \mathrm{id}$, each ζ_s a diffeomorphism onto an open set. Define, for $x \in U := x_0 + B$ and $t \in [0, T]$,

$$\phi_t(x) = \zeta_{t/T}(x - x_0) + c(t) . \tag{13.271}$$

Then $\phi_0(x) = (x - x_0) + x_0 = x$, and by Equation (13.270)

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(x) = \frac{1}{T}\, Z_{t/T}\left(\zeta_{t/T}(x - x_0)\right) + \dot{c}(t) = X\left(\phi_t(x)\right) - X\left(c(t)\right) + X\left(c(t)\right) ,$$

which is Equation (13.268). Each ϕ_t is a diffeomorphism onto an open set, being $\zeta_{t/T}$ composed with two translations, and ϕ is smooth in (t, x) jointly because ζ and c are. Negative t is covered by running the same construction for $-X$, whose integral curves are those of X traversed backwards; taking ε smaller than both times obtained gives Equation (13.268) on $]-\varepsilon, \varepsilon[\times U$.

Uniqueness and the group law. Two integral curves agreeing at t_1 satisfy the same initial value problem there, so they agree near t_1 by Theorem 9.8; the set where they agree is therefore open, and it is closed by continuity, so it is the whole of the (connected) intersection of their domains. Now fix s and x: both $t \mapsto \phi_{t+s}(x)$ and $t \mapsto \phi_t\left(\phi_s(x)\right)$ are integral curves of X taking the value $\phi_s(x)$ at $t = 0$, whence Equation (13.269); putting $s = -t$ gives $\phi_t \circ \phi_{-t} = \mathrm{id}$. Completeness removes the restriction on the interval, and Equation (13.269) then says that $t \mapsto \phi_t$ is a homomorphism of $(\mathbb{R}, +)$ into the diffeomorphisms of M_n. □

Remark 13.126 (What the proof imports, and what it does not). Only two things: the Picard–Lindelöf theorem of Chapter 9, for existence and uniqueness, and the flow theorem proved in the appendix for Moser's trick (Theorem A.74), for smooth dependence on the initial point. The latter is stated there for a time-dependent field vanishing at a point,

which looks like a special case and is not: subtracting the value of X along the integral curve through P, as Equation (13.270) does, turns *any* smooth field into one of that kind, at the cost of making it time-dependent — which the appendix theorem already allows. No fixed-point theorem and no compactness argument beyond Theorem 7.24 enters.

Proposition 13.127 (Component formulas)**.** *For a scalar, a vector, a covector, and a* $(0,2)$ *tensor,*

$$\mathcal{L}_\xi f = \xi^\mu \partial_\mu f \,, \tag{13.272}$$

$$(\mathcal{L}_\xi V)^\mu = \xi^\nu \partial_\nu V^\mu - V^\nu \partial_\nu \xi^\mu = [\xi, V]^\mu \,, \tag{13.273}$$

$$(\mathcal{L}_\xi \omega)_\mu = \xi^\nu \partial_\nu \omega_\mu + \omega_\nu \partial_\mu \xi^\nu \,, \tag{13.274}$$

$$(\mathcal{L}_\xi g)_{\mu\nu} = \xi^\lambda \partial_\lambda g_{\mu\nu} + g_{\lambda\nu} \partial_\mu \xi^\lambda + g_{\mu\lambda} \partial_\nu \xi^\lambda \,, \tag{13.275}$$

with one drag term per index in general. All partial derivatives may be replaced by covariant derivatives of any torsion-free connection: the connection terms cancel pairwise.

Rests on Equation (13.267) *and Definitions 13.89 and 13.90.*

Proof. *Derives Proposition 13.127.*

Expand ϕ_t to first order, $x'^\mu = x^\mu + t\,\xi^\mu$, in the transformation law of each tensor type (Section 13.5.2), and differentiate at $t = 0$: the Jacobian $\partial x'^\mu / \partial x^\nu = \delta^\mu{}_\nu + t\,\partial_\nu \xi^\mu$ contributes the drag terms with the signs shown (one $+$ per lower index, one $-$ per upper index), and the argument shift contributes the transport term $\xi \cdot \partial$. The cancellation of connection terms is a direct check using the symmetry $\Gamma^\lambda{}_{\mu\nu} = \Gamma^\lambda{}_{\nu\mu}$ of a torsion-free connection. □

On differential forms (Section 13.6) the Lie derivative has a purely exterior expression:

Proposition 13.128 (Cartan's magic formula)**.** *On any* k*-form* α*,*

$$\mathcal{L}_\xi \alpha = \mathrm{d}\,(i_\xi \alpha) + i_\xi\,(\mathrm{d}\alpha) \,, \tag{13.276}$$

where i_ξ *is the interior product,* $(i_\xi \alpha)_{\mu_2 \ldots \mu_k} = \xi^{\mu_1} \alpha_{\mu_1 \mu_2 \ldots \mu_k}$*.*

Rests on Proposition 13.127, Definition 13.103, and Equations (13.245) *and* (13.246)*.*

Proof. *Derives Proposition 13.128.*

On functions, $i_\xi f \equiv 0$ and $i_\xi \mathrm{d}f = \xi^\mu \partial_\mu f = \mathcal{L}_\xi f$. On the coordinate differentials, $\mathcal{L}_\xi \mathrm{d}x^\mu = \mathrm{d}\xi^\mu = \mathrm{d}(i_\xi \mathrm{d}x^\mu) + i_\xi \mathrm{d}(\mathrm{d}x^\mu)$, using $\mathrm{d}^2 = 0$ (Section 13.6.4). Both sides of Equation (13.276) are derivations of the exterior algebra (they obey the graded Leibniz rule — for the right side this follows from the graded Leibniz rules of d and i_ξ), and a derivation is fixed by its action on functions and their differentials, which generate all forms locally. The two sides agree there, hence everywhere. □

The three operators $\mathcal{L}$, d and i close on one another, and Equation (13.276) is only the first of the identities they satisfy. The second is the one that lets a computation move an interior product past a Lie derivative, which is what makes many arguments with forms one line long instead of a component calculation.

Proposition 13.129 (Commutation of Lie derivative and interior product)**.** *For any smooth vector fields* X, Y *and any* k*-form* α*,*

$$\mathcal{L}_X\, i_Y \alpha - i_Y\, \mathcal{L}_X \alpha = i_{[X,Y]} \alpha \,. \tag{13.277}$$

Rests on Propositions 13.127 and 13.128 and Definition 13.98.

Proof. *Derives Proposition 13.129.*

Both sides are $(k-1)$-forms; compare components. Write $\beta = i_Y\alpha$, so $\beta_{\mu_2\dots\mu_k} = Y^\nu\alpha_{\nu\mu_2\dots\mu_k}$ by the convention of Proposition 13.128, and use the general component formula of Proposition 13.127 — transport term plus one drag term per lower index:

$$(\mathcal{L}_X\beta)_{\mu_2\dots\mu_k} = X^\lambda\partial_\lambda\left(Y^\nu\alpha_{\nu\mu_2\dots\mu_k}\right) + \sum_{i=2}^{k} Y^\nu\alpha_{\nu\dots\lambda\dots\mu_k}\,\partial_{\mu_i}X^\lambda\,,$$

$$(i_Y\mathcal{L}_X\alpha)_{\mu_2\dots\mu_k} = Y^\nu\left[X^\lambda\partial_\lambda\alpha_{\nu\mu_2\dots\mu_k} + \alpha_{\lambda\mu_2\dots\mu_k}\partial_\nu X^\lambda + \sum_{i=2}^{k}\alpha_{\nu\dots\lambda\dots\mu_k}\,\partial_{\mu_i}X^\lambda\right],$$

where in each sum λ sits in the i-th slot. Subtracting, the transport terms of α and all the drag terms carried by $\mu_2,\dots,\mu_k$ cancel, leaving

$$\left(X^\lambda\partial_\lambda Y^\nu - Y^\lambda\partial_\lambda X^\nu\right)\alpha_{\nu\mu_2\dots\mu_k} = [X,Y]^\nu\,\alpha_{\nu\mu_2\dots\mu_k}\,,$$

by Equation (13.273), and that is $\left(i_{[X,Y]}\alpha\right)_{\mu_2\dots\mu_k}$. □

13.9.1 Distributions and the Frobenius theorem

A single vector field, by Theorem 13.125, threads the manifold with curves: through every point passes one integral curve, and the curves fill the manifold. The question of this subsection is what happens when a k-dimensional field of tangent planes is prescribed instead of a field of tangent lines. Does a k-dimensional surface pass through each point with the prescribed tangent plane everywhere? For $k=1$ the answer is always yes; for $k\geq 2$ it is yes precisely when an algebraic condition on brackets holds, and that is the Frobenius theorem. It is the theorem behind the symplectic foliation of a Poisson manifold, behind the counting of gauge degrees of freedom generated by first-class constraints, and behind the very meaning of the word *nonholonomic*: a constraint is nonholonomic exactly when the distribution it defines fails the condition below.

Definition 13.130 (Distribution; involutive; integrable)**.** A *distribution of rank k* on M_n is an assignment $P\mapsto D_P\subseteq T_PM_n$ of a k-dimensional subspace to each point which is *smooth*: every point has a neighbourhood on which there are smooth vector fields $X_1,\dots,X_k$ spanning D_Q at each Q of it. A vector field X *belongs to* D iff $X_Q\in D_Q$ everywhere. The distribution is

1. *involutive* iff $[X,Y]$ belongs to D whenever X and Y do;
2. *integrable* iff through each $P\in M_n$ there passes an *integral manifold*: an embedded k-dimensional submanifold $S\ni P$ (Definition 13.54) with $T_QS = D_Q$ for every $Q\in S$.

Rests on Definitions 13.54 and 13.82 and Equation (13.273).

Proposition 13.131 (Commuting fields have commuting flows)**.** *Let X and Y be smooth vector fields with flows ϕ_t and η_s. If $[X,Y]=0$ then $(\phi_t)_*Y = Y$ and*

$$\phi_t\circ\eta_s = \eta_s\circ\phi_t \tag{13.278}$$

wherever both sides are defined.

Rests on Theorem 13.125 and Equations (13.267) and (13.273).

Proof. *Derives Proposition 13.131.*

Fix x and put $W(t) = (\phi_{-t})_*\,Y_{\phi_t(x)}\in T_xM_n$, a smooth curve in one fixed vector space. By the group law Equation (13.269), for any t_0,

$$W(t_0+h) = (\phi_{-t_0})_*\left[(\phi_{-h})_*\,Y_{\phi_h(y)}\right],\qquad y=\phi_{t_0}(x)\,,$$

so differentiating at $h = 0$ and using the definition Equation (13.267) of the Lie derivative together with Equation (13.273), $\frac{\mathrm{d}W}{\mathrm{d}t}(t_0) = (\phi_{-t_0})_* (\mathcal{L}_X Y)_y = (\phi_{-t_0})_* [X, Y]_y = 0$. Hence $W(t) = W(0) = Y_x$, i.e. $(\phi_t)_* Y_x = Y_{\phi_t(x)}$ for every t and x.

A diffeomorphism that carries Y to Y carries integral curves of Y to integral curves of Y: if $\dot\gamma = Y \circ \gamma$ then $\frac{\mathrm{d}}{\mathrm{d}s}(\phi_t \circ \gamma) = (\phi_t)_* Y_\gamma = Y_{\phi_t(\gamma)}$. So $s \mapsto \phi_t(\eta_s(x))$ is the integral curve of Y through $\phi_t(x)$, which is $s \mapsto \eta_s(\phi_t(x))$ by the uniqueness clause of Theorem 13.125. That is Equation (13.278). □

Proposition 13.132 (Simultaneous straightening of commuting fields)**.** *Let $Y_1, \dots, Y_k$ be smooth vector fields on M_n, pairwise commuting, $[Y_a, Y_b] = 0$, and linearly independent at P. Then there is a chart $(u^1, \dots, u^n)$ around P in which*

$$Y_a = \frac{\partial}{\partial u^a}, \qquad a = 1, \dots, k. \tag{13.279}$$

Rests on Proposition 13.131, Theorem 13.125, and Corollary A.292.

Proof. *Derives Proposition 13.132.*

Take any chart around P sending P to the origin of $\mathbb{R}^n$, and follow it by the linear change of coordinates that sends $Y_1(P), \dots, Y_k(P)$ — independent by hypothesis, hence extendable to a basis — to the first k coordinate vectors. Write ϕ^a for the flow of Y_a (Theorem 13.125) and, for u in a small neighbourhood of 0 in $\mathbb{R}^n$, set

$$\Psi\left(u^1, \dots, u^n\right) = \left(\phi^1_{u^1} \circ \cdots \circ \phi^k_{u^k}\right)\left(0, \dots, 0, u^{k+1}, \dots, u^n\right), \tag{13.280}$$

the argument on the right being the point of the slice $u^1 = \cdots = u^k = 0$. This is defined and smooth near 0, each flow being smooth in both arguments.

Its differential at 0 is invertible: all parameters vanish there, so $\partial\Psi/\partial u^a\big|_0 = Y_a(P)$ for $a \le k$, while $\partial\Psi/\partial u^j\big|_0 = \partial/\partial x^j\big|_P$ for $j > k$, and these n vectors are a basis by the choice of the initial chart. By the inverse function theorem (Corollary A.292) Ψ is a diffeomorphism of a neighbourhood of 0 onto a neighbourhood of P, and we take Ψ^{-1} as the chart.

Finally, fix a. Because the flows commute (Proposition 13.131), the composition in Equation (13.280) may be reordered to put ϕ^a outermost, and then $\partial\Psi/\partial u^a$ is the velocity of the integral curve of Y_a through $\Psi(u)$, that is $Y_a(\Psi(u))$. Hence $\Psi_*(\partial/\partial u^a) = Y_a$, which is Equation (13.279). □

Theorem 13.133 (Frobenius)**.** *A smooth distribution D of constant rank k on M_n is integrable if and only if it is involutive. When it is, every point has a chart $(u^1, \dots, u^n)$ in which*

$$D = \operatorname{span}\left\{\frac{\partial}{\partial u^1}, \dots, \frac{\partial}{\partial u^k}\right\}, \tag{13.281}$$

so that the slices $u^{k+1} = const, \dots, u^n = const$ are integral manifolds: the chart domain is foliated by them, and every connected integral manifold inside it lies in one slice.

Rests on Definition 13.130, Proposition 13.132, and Equation (13.273).

Proof. *Derives Theorem 13.133.*

Integrable implies involutive. Let X, Y belong to D and let $P \in M_n$ lie on an integral manifold S. Both fields are tangent to S, hence restrict to vector fields on S; the bracket Equation (13.273) of two fields tangent to an embedded submanifold is again tangent to it, since in a chart adapted to S both fields have vanishing components transverse to S and Equation (13.273) differentiates those components only along directions inside S. So $[X, Y]_P \in T_P S = D_P$.

Involutive implies integrable. Fix P and a local spanning frame $X_1, \dots, X_k$ of D (Definition 13.130). Choose a chart around P with coordinates $(x^1, \dots, x^n)$ such that D_P is transverse to the span of $\partial/\partial x^{k+1}, \dots, \partial/\partial x^n$; equivalently, writing $X_a = X_a{}^i \partial/\partial x^i$, the

$k \times k$ block $A_a{}^b = X_a{}^b$ with $b \le k$ is invertible at P, hence on a neighbourhood. Replace the frame by

$$Y_a = \left(A^{-1}\right)_a{}^b X_b = \frac{\partial}{\partial x^a} + \sum_{j>k} Y_a{}^j \frac{\partial}{\partial x^j}\,, \tag{13.282}$$

still a frame for D, since the change is by an invertible matrix of functions.

Now compute $[Y_a, Y_b]$ from Equation (13.273). The components of Y_a along $\partial/\partial x^c$ with $c \le k$ are the constants δ_a^c, so every derivative of them vanishes and the bracket has *no* component along $\partial/\partial x^1, \dots, \partial/\partial x^k$:

$$[Y_a, Y_b] = \sum_{j>k} \left(Y_a\left(Y_b{}^j\right) - Y_b\left(Y_a{}^j\right)\right) \frac{\partial}{\partial x^j}\,. \tag{13.283}$$

By involutivity the left side belongs to D, so it is a combination $\lambda^a Y_a$; comparing the components along $\partial/\partial x^c$, $c \le k$, which are λ^c on the right and zero on the left by Equation (13.283), gives $\lambda^c = 0$ and hence

$$[Y_a, Y_b] = 0\,. \tag{13.284}$$

The frame $Y_1, \dots, Y_k$ is therefore commuting and pointwise independent, and Proposition 13.132 supplies a chart in which $Y_a = \partial/\partial u^a$. Since the Y_a span D, this is Equation (13.281), and the slices $u^{k+1}, \dots, u^n = \text{const}$ are embedded k-dimensional submanifolds whose tangent space at each point is exactly D: integral manifolds through every point of the chart domain.

For the last clause, let S be a connected integral manifold inside the chart domain. On S each function u^j with $j > k$ has vanishing differential, because $T_Q S = D_Q$ is spanned by the $\partial/\partial u^a$ with $a \le k$; a function with vanishing differential on a connected manifold is constant (Corollary 7.36 applied along paths in a chart). So S lies in one slice. □

The dual formulation is the one that appears whenever the planes are prescribed by equations rather than by spanning fields — a mechanical constraint written as a set of one-forms annihilating the admissible velocities is exactly that situation.

Corollary 13.134 (Frobenius for a Pfaffian system)**.** *Let D have rank k and let $\theta^1, \dots, \theta^{n-k}$ be pointwise independent one-forms whose common kernel is D, so that a vector v lies in D if and only if $\theta^\alpha(v) = 0$ for every α. Then D is integrable if and only if the ideal generated by the θ^α is closed under exterior differentiation, i.e. iff there are one-forms $\eta^\alpha{}_\beta$ with*

$$\mathrm{d}\theta^\alpha = \sum_\beta \eta^\alpha{}_\beta \wedge \theta^\beta\,, \tag{13.285}$$

equivalently iff $\mathrm{d}\theta^\alpha \wedge \theta^1 \wedge \cdots \wedge \theta^{n-k} = 0$ for every α.

Rests on Theorem 13.133 and Definitions 13.103 and 13.130.

Proof. *Derives Corollary 13.134.*

For a one-form $\theta = \theta_\mu \mathrm{d}x^\mu$, the components of $\mathrm{d}\theta$ are $\partial_\mu\theta_\nu - \partial_\nu\theta_\mu$ (Definition 13.103), and for any vector fields X, Y

$$\left(\partial_\mu\theta_\nu - \partial_\nu\theta_\mu\right) X^\mu Y^\nu = X(\theta(Y)) - Y(\theta(X)) - \theta([X,Y])\,, \tag{13.286}$$

as expanding the three terms on the right with Equation (13.273) shows: the terms carrying a derivative of X or of Y cancel in pairs and only the derivatives of θ survive. If X and Y belong to D, the first two terms on the right of Equation (13.286) vanish, so

$$\left(\mathrm{d}\theta^\alpha\right)_{\mu\nu} X^\mu Y^\nu = -\,\theta^\alpha([X,Y])\,. \tag{13.287}$$

Involutivity says the right-hand side vanishes for every α and every X, Y in D — because the θ^α detect exactly membership of D — and that is the statement that each $\mathrm{d}\theta^\alpha$ annihilates

$D \times D$. Completing the θ^α to a coframe by one-forms σ^a dual to a frame of D, a two-form annihilating $D \times D$ has no $\sigma^a \wedge \sigma^b$ term, i.e. every term carries at least one factor θ^β, which is Equation (13.285); and a two-form of that shape wedges to zero against $\theta^1 \wedge \cdots \wedge \theta^{n-k}$, while conversely a $\sigma^a \wedge \sigma^b$ term would survive that wedge. The three conditions are therefore equivalent, and by Theorem 13.133 each is equivalent to integrability. □

Remark 13.135 (Constant rank is a hypothesis, not a formality). Theorem 13.133 assumes the rank of D is the same at every point, and the hypothesis is doing work: Equation (13.282) needs a local frame of fixed size, and Equation (13.281) asserts a fixed number of slice coordinates. Distributions whose rank jumps do occur — a bivector field may be invertible on the complement of a point and vanish at it, so that the distribution it spans has rank 2 almost everywhere and rank 0 there — and at such a point the conclusion above says nothing whatever, not even that a leaf of some smaller dimension passes through. This treatise establishes no generalisation covering that case, and nothing in it rests on one; the limitation is recorded here so that an appeal to Theorem 13.133 elsewhere in the book can be checked against its hypothesis rather than against its name.

Lemma 13.136 (Discrete subgroups of $\mathbb{R}^f$)**.** *Let Γ be a subgroup of $\left(\mathbb{R}^f, +\right)$ which is* discrete*, i.e. $\Gamma \cap B_r(0) = \{0\}$ for some $r > 0$. Then there are $m \leq f$ linearly independent vectors $w_1, \ldots, w_m \in \Gamma$ with*

$$\Gamma = \mathbb{Z}w_1 \oplus \cdots \oplus \mathbb{Z}w_m \,, \tag{13.288}$$

and m is the dimension of the linear span of Γ.

Rests on Definitions 5.13 and 6.9.

Proof. *Derives Lemma 13.136.*

First, a bounded set meets Γ in finitely many points: two distinct elements of Γ differ by a nonzero element of Γ, hence by at least r, so a ball of radius R can contain at most finitely many of them.

Induct on $m = \dim \operatorname{span} \Gamma$. For $m = 0$, $\Gamma = \{0\}$. Let $m \geq 1$ and choose linearly independent $v_1, \ldots, v_m \in \Gamma$ spanning that span. Put $W = \operatorname{span}\{v_1, \ldots, v_{m-1}\}$. The subgroup $\Gamma' = \Gamma \cap W$ is discrete with span exactly W, so by the induction hypothesis $\Gamma' = \mathbb{Z}w_1 \oplus \cdots \oplus \mathbb{Z}w_{m-1}$ with the w_i linearly independent, hence a basis of W.

Every $x \in \Gamma$ decomposes uniquely as $x = \sum_{i<m} a_i w_i + t\, v_m$ with a_i, t real. Let

$$S = \{x \in \Gamma \mid 0 \leq a_i < 1 \text{ for all } i < m, \quad 0 < t \leq 1\} \,,$$

a bounded set, hence finite, and non-empty: subtracting from v_m suitable integer multiples of the w_i — which are in Γ — brings its a_i into $[0, 1[$ without changing $t = 1$. Choose $w_m \in S$ with $t = t_0$ minimal.

Let $x \in \Gamma$ have v_m-coefficient t, put $n = \lfloor t/t_0 \rfloor$ and consider $x - n\, w_m \in \Gamma$, whose v_m-coefficient is $t - nt_0 \in [0, t_0[$. If it were positive, subtracting integer multiples of the w_i would produce an element of S with a v_m-coefficient smaller than t_0, contradicting minimality. So $t = nt_0$ and $x - n\, w_m \in \Gamma \cap W = \Gamma'$. Hence $\Gamma = \Gamma' + \mathbb{Z}w_m$, the sum is direct because $w_m \notin W$, and $w_1, \ldots, w_m$ are linearly independent. Finally $m \leq f$, being the number of independent vectors in $\mathbb{R}^f$. □

Theorem 13.137 (Commuting complete fields on a compact manifold)**.** *Let M_f be a compact connected manifold of dimension f carrying f complete, pairwise commuting vector fields $Y_1, \ldots, Y_f$ that are linearly independent at every point. Then the map*

$$\Theta : \mathbb{R}^f \times M_f \longrightarrow M_f \,, \qquad \Theta\,(t, x) = \left(\phi^1_{t^1} \circ \cdots \circ \phi^f_{t^f}\right)(x) \tag{13.289}$$

is a smooth action of the abelian group $\left(\mathbb{R}^f, +\right)$ which is transitive, and M_f is diffeomorphic to the torus $T^f = \mathbb{R}^f / \mathbb{Z}^f$.

Rests on Propositions 13.131 and 13.132 and Theorem 13.125.

Proof. *Derives Theorem 13.137.*

Each ϕ^a is defined for all parameters by completeness (Definition 13.124 and Theorem 13.125) and the f flows commute in pairs (Proposition 13.131), so Equation (13.289) does not depend on the order of composition and $\Theta(t, \Theta(s, x)) = \Theta(t + s, x)$: it is a smooth action of $\mathbb{R}^f$.

The orbits are open, hence there is one. Fix x and apply Proposition 13.132 with $k = f = n$: near x there is a chart in which $Y_a = \partial/\partial u^a$, and in it $\Theta(\cdot, x)$ is $t \mapsto u = t$ near $t = 0$. So the orbit of x contains a neighbourhood of x; being a union of such, every orbit is open. The orbits partition M_f, and M_f is connected, so there is exactly one: the action is transitive. The same computation shows that $\Theta(\cdot, x)$ is a local diffeomorphism at every t, by the group law.

The stabiliser is a lattice. Let $\Gamma = \left\{ t \in \mathbb{R}^f \mid \Theta(t, x) = x \right\}$, a subgroup of $\mathbb{R}^f$; since $\Theta(\cdot, x)$ is injective near 0, Γ meets some neighbourhood of 0 only in 0, so Γ is discrete, and by transitivity $\Theta(\cdot, x)$ descends to a bijection $\mathbb{R}^f/\Gamma \longrightarrow M_f$, a diffeomorphism because $\Theta(\cdot, x)$ is a local diffeomorphism everywhere. By Lemma 13.136, $\Gamma = \mathbb{Z}w_1 \oplus \cdots \oplus \mathbb{Z}w_m$ with $w_1, \ldots, w_m$ linearly independent; completing them to a basis of $\mathbb{R}^f$ and using that basis as coordinates identifies $\mathbb{R}^f/\Gamma$ with $T^m \times \mathbb{R}^{f-m}$. Compactness of M_f forces $m = f$, since $\mathbb{R}^{f-m}$ is non-compact for $m < f$, and the same change of basis carries Γ to $\mathbb{Z}^f$. Hence $M_f \cong \mathbb{R}^f/\mathbb{Z}^f = T^f$. □

Remark 13.138 (Where this is used). Theorem 13.137 is the differential-topological half of the Liouville–Arnold theorem of Hamiltonian mechanics: an integrable system with f degrees of freedom supplies, on a compact connected level set of its f commuting conserved quantities, exactly f commuting independent vector fields, and the theorem then says the level set is a torus. The symplectic half — that the resulting angle coordinates are canonical — belongs to the mechanics part and is not a statement about manifolds alone.

13.9.2 Killing vectors

Definition 13.139 (Killing vector)**.** $\boldsymbol{\xi}$ is a *Killing vector* of (M, g) iff its flow consists of isometries, i.e. iff $\mathcal{L}_\xi g = 0$.

Rests on Definition 13.120 and Equation (13.267).

Proposition 13.140 (Killing's equation)**.** *With ∇ the Levi-Civita connection of g (Theorem 13.150),*

$$\mathcal{L}_\xi g = 0 \quad \Longleftrightarrow \quad \nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu = 0\,. \tag{13.290}$$

A D-dimensional manifold admits at most $D(D+1)/2$ linearly independent Killing vectors.

Rests on Definitions 13.139 and 13.149, Proposition 13.127, and Theorems 13.150 and 13.152.

Proof. *Derives Proposition 13.140.*

Replace ∂ by ∇ in Equation (13.275) (Proposition 13.127) and use metricity $\nabla_\lambda g_{\mu\nu} = 0$: the transport term drops and the drag terms become $\nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu$. For the count: differentiating Equation (13.290) and commuting derivatives (Section 13.11) yields $\nabla_\mu \nabla_\nu \xi_\lambda = R^\rho{}_{\mu\nu\lambda}\, \xi_\rho$ (torsion-free), so a Killing vector is determined everywhere by the D values $\xi_\mu(P)$ and the $D(D-1)/2$ values $\nabla_{[\mu}\xi_{\nu]}(P)$ at a single point: $D + D(D-1)/2 = D(D+1)/2$. □

A Killing vector is worth having because it gives a conserved quantity, and the conservation law — not the symmetry — is what a physical chapter actually uses. It is one line from Killing's equation.

Proposition 13.141 (The invariant of a Killing vector along a geodesic)**.** *Let $\boldsymbol{\xi}$ be a Killing vector of (M, g) and let γ be an affinely parametrized geodesic of the Levi-Civita connection, with tangent $u^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}t}$ and $u^\nu \nabla_\nu u^\mu = 0$. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\xi_\mu u^\mu\right) = 0\,, \tag{13.291}$$

that is, $\xi_\mu u^\mu$ is constant along γ.
Rests on Proposition 13.140 and Definitions 13.146 and 13.149.

Proof. *Derives Proposition 13.141.*
The derivative along the curve of a scalar is $u^\nu\nabla_\nu$ of it, so

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\xi_\mu u^\mu\right) = u^\nu u^\mu \nabla_\nu \xi_\mu + \xi_\mu\, u^\nu \nabla_\nu u^\mu \,.$$

The second term vanishes by the geodesic equation (Equation (13.297) with $\Gamma = \mathring{\Gamma}$). In the first, $u^\nu u^\mu$ is symmetric under $\mu \leftrightarrow \nu$, so only the symmetric part of $\nabla_\nu \xi_\mu$ contributes:

$$u^\nu u^\mu \nabla_\nu \xi_\mu = \tfrac{1}{2} u^\nu u^\mu \left(\nabla_\nu \xi_\mu + \nabla_\mu \xi_\nu\right) \,,$$

and the bracket vanishes by Killing's equation Equation (13.290). □

Not every geodesic invariant comes from an isometry. The object that produces the rest is the symmetric-tensor generalisation of a Killing vector.

Definition 13.142 (Killing tensor)**.** A *Killing tensor of rank r* on (M, g) is a totally symmetric tensor field $K_{\mu_1\ldots\mu_r}$ satisfying

$$\nabla_{(\lambda} K_{\mu_1\ldots\mu_r)} = 0 \,, \tag{13.292}$$

the round brackets denoting here the *total* symmetrization over the $r+1$ indices they enclose — the average over all $(r+1)!$ orderings of them — which extends to any number of indices the two-index convention Equation (13.213) of Definition 13.91. For $r = 1$, Equation (13.292) is Killing's equation Equation (13.290).
Rests on Definitions 13.91, 13.139 and 13.149.

Theorem 13.143 (Killing tensors and geodesic invariants)**.** *Let K be a Killing tensor of rank r and γ an affinely parametrized geodesic with tangent u^μ. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(K_{\mu_1\ldots\mu_r}\, u^{\mu_1}\cdots u^{\mu_r}\right) = 0 \,. \tag{13.293}$$

Conversely, a function on the tangent bundle that is a homogeneous polynomial of degree r in u and constant along every geodesic has totally symmetric coefficients obeying Equation (13.292).
Rests on Definitions 13.142 and 13.146 and Proposition 13.141.

Proof. *Derives Theorem 13.143.*
Differentiating along the curve and using $u^\nu\nabla_\nu u^\mu = 0$ for each of the r factors,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(K_{\mu_1\ldots\mu_r} u^{\mu_1}\cdots u^{\mu_r}\right) = \left(\nabla_\lambda K_{\mu_1\ldots\mu_r}\right) u^\lambda u^{\mu_1}\cdots u^{\mu_r} \,.$$

The product $u^\lambda u^{\mu_1}\cdots u^{\mu_r}$ is totally symmetric in its $r+1$ indices, so contracting it with any tensor picks out that tensor's totally symmetric part; the right-hand side is therefore $\left(\nabla_{(\lambda} K_{\mu_1\ldots\mu_r)}\right) u^\lambda u^{\mu_1}\cdots u^{\mu_r}$, which vanishes by Equation (13.292).

For the converse, the same computation shows that the derivative along every geodesic is $\left(\nabla_{(\lambda} K_{\mu_1\ldots\mu_r)}\right) u^\lambda u^{\mu_1}\cdots u^{\mu_r}$, and through a given point a geodesic may be started with any tangent u (Equation (13.297) is a second-order equation with arbitrary initial position and velocity, solvable by Theorem 9.8). A totally symmetric tensor whose contraction with $u^{\otimes(r+1)}$ vanishes for every u is zero, by polarization: the contraction is a homogeneous polynomial in the components of u whose coefficients are exactly the independent components of the symmetric tensor. □

Example 13.144 (The two Killing tensors every metric carries)**.** The metric itself is a Killing tensor of rank 2: metricity (Definition 13.149) gives $\nabla_\lambda g_{\mu\nu} = 0$, so Equation (13.292) holds trivially, and the invariant Equation (13.293) is $g_{\mu\nu}u^\mu u^\nu$ — the statement that a geodesic keeps the norm of its tangent, i.e. that timelike, null and spacelike geodesics stay so. Symmetrized products of Killing vectors are Killing tensors too: if ξ and χ are Killing vectors then $K_{\mu\nu} = \xi_{(\mu}\chi_{\nu)}$ obeys Equation (13.292), since Equation (13.290) makes each $\nabla_{(\lambda}\xi_{\mu)}$ vanish. A Killing tensor is called *irreducible* when it is not a combination of the metric and such products; an irreducible one gives a constant of the motion that no isometry accounts for, which is why the bound $D(D+1)/2$ of Proposition 13.140 does not bound the number of geodesic invariants.

Rests on Definitions 13.142 and 13.149 and Proposition 13.140.

Metrics that saturate this bound are *maximally symmetric* (Section 13.13). Relaxing isometry to angle preservation defines the conformal Killing vectors, $\mathcal{L}_\xi g = 2\sigma g$ with $\sigma = \nabla_\mu \xi^\mu / D$; for flat space of signature (p, q) with $D > 2$ their maximal number is $(D+1)(D+2)/2$, the dimension of the conformal group $\mathrm{C}_{p+q} = \mathrm{SO}(p+1, q+1)$ (Section 13.13.1).

13.10 Connections, parallel transport, and torsion

In Euclidean geometry a homogeneous (constant) vector field is one whose derivative with respect to position vanishes; its vectors are parallel and of equal length. How does this notion extend to an arbitrary manifold? The partial derivative fails: differentiating the vector transformation law $V'^\mu = (\partial x'^\mu/\partial x^\nu)V^\nu$ gives

$$\partial'_\nu V'^\mu = \frac{\partial x^\rho}{\partial x'^\nu}\frac{\partial x'^\mu}{\partial x^\sigma}\,\partial_\rho V^\sigma + \frac{\partial x^\rho}{\partial x'^\nu}\frac{\partial^2 x'^\mu}{\partial x^\rho \partial x^\sigma}\,V^\sigma\,, \tag{13.294}$$

whose second, inhomogeneous term spoils tensoriality whenever the coordinate change is nonlinear. The covariant derivative is the repair, and it is the concept on which the whole of theoretical physics leans.

Definition 13.145 (Affine connection)**.** An *affine connection* ∇ assigns to vector fields the $(1,1)$-tensor-valued derivative

$$\nabla_\mu V^\lambda = \partial_\mu V^\lambda + \Gamma^\lambda{}_{\mu\nu} V^\nu\,, \tag{13.295}$$

extended to covectors by $\nabla_\mu \omega_\nu = \partial_\mu \omega_\nu - \Gamma^\lambda{}_{\mu\nu}\omega_\lambda$ (so that $\nabla(V^\mu\omega_\mu)$ obeys the Leibniz rule and reduces to ∂), and to arbitrary (r, s) tensors with one $+\Gamma$ term per upper and one $-\Gamma$ term per lower index. The connection coefficients are *not* tensor components: requiring Equation (13.295) to transform as a tensor forces

$$\Gamma'^\lambda{}_{\mu\nu} = \frac{\partial x'^\lambda}{\partial x^\rho}\frac{\partial x^\sigma}{\partial x'^\mu}\frac{\partial x^\tau}{\partial x'^\nu}\,\Gamma^\rho{}_{\sigma\tau} + \frac{\partial x'^\lambda}{\partial x^\rho}\frac{\partial^2 x^\rho}{\partial x'^\mu \partial x'^\nu}\,, \tag{13.296}$$

the inhomogeneous term cancelling that of Equation (13.294).

Rests on Definitions 13.82 and 13.90.

Definition 13.146 (Parallel transport and autoparallels)**.** A tensor T is *parallelly transported* along a curve with tangent $u^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}t}$ iff $u^\mu \nabla_\mu T = 0$. A curve is an *autoparallel* iff it parallelly transports its own tangent,

$$\frac{\mathrm{d}^2 x^\lambda}{\mathrm{d}t^2} + \Gamma^\lambda{}_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}t}\frac{\mathrm{d}x^\nu}{\mathrm{d}t} = 0\,, \tag{13.297}$$

generalizing the geodesics of surface theory (Section 13.3.3). Only the symmetric part $\Gamma^\lambda{}_{(\mu\nu)}$ enters Equation (13.297).

Rests on Definition 13.145.

Definition 13.147 (Torsion)**.** The *torsion* of a connection is

$$T^{\lambda}{}_{\mu\nu} = \Gamma^{\lambda}{}_{\mu\nu} - \Gamma^{\lambda}{}_{\nu\mu}\,. \tag{13.298}$$

Rests on Definition 13.145.

Proposition 13.148. *$T^{\lambda}{}_{\mu\nu}$ is a tensor, although $\Gamma^{\lambda}{}_{\mu\nu}$ is not.*
Rests on Definitions 13.90 and 13.147, Equation (13.296), and Proposition 7.73.

Proof. *Derives Proposition 13.148.*
The inhomogeneous term of Equation (13.296) is symmetric in $\mu\nu$ (mixed second partials commute, Proposition 7.73), so it cancels in the antisymmetrization Equation (13.298), leaving the homogeneous tensor law. □

Geometrically, torsion measures the failure of infinitesimal parallelograms to close: transporting $\mathrm{d}x_{(1)}$ along $\mathrm{d}x_{(2)}$ and vice versa, the two endpoints differ by $T^{\lambda}{}_{\mu\nu}\,\mathrm{d}x^{\mu}_{(1)}\mathrm{d}x^{\nu}_{(2)}$. In Section 13.12 it becomes one of the two field strengths of the geometry.

Definition 13.149 (Metric compatibility)**.** A connection is *metric compatible* iff $\nabla_{\lambda} g_{\mu\nu} = 0$: parallel transport preserves lengths and angles.
Rests on Definitions 13.117, 13.145 and 13.146.

Theorem 13.150 (Levi-Civita connection and contorsion)**.** *On a pseudo-Riemannian manifold there is exactly one metric-compatible connection for each prescribed torsion. Lowering the first index, $\Gamma_{\lambda\mu\nu} = g_{\lambda\rho}\Gamma^{\rho}{}_{\mu\nu}$,*

$$\Gamma_{\lambda\mu\nu} = \underbrace{\tfrac{1}{2}\left(\partial_{\mu} g_{\nu\lambda} + \partial_{\nu} g_{\lambda\mu} - \partial_{\lambda} g_{\mu\nu}\right)}_{\mathring{\Gamma}_{\lambda\mu\nu}} + \underbrace{\tfrac{1}{2}\left(T_{\lambda\mu\nu} + T_{\mu\lambda\nu} + T_{\nu\lambda\mu}\right)}_{K_{\lambda\mu\nu}}, \tag{13.299}$$

with $T_{\lambda\mu\nu} = g_{\lambda\rho}T^{\rho}{}_{\mu\nu}$. The torsion-free case $K = 0$ is the Levi-Civita connection $\mathring{\Gamma}$ *(the Christoffel symbols, generalizing Section 13.3.2); the* contorsion $K_{\lambda\mu\nu}$ *is a tensor, antisymmetric in its outer pair, $K_{\lambda\mu\nu} = -K_{\nu\mu\lambda}$.*
Rests on Definitions 13.117, 13.145, 13.147 and 13.149.

Proof. *Derives Theorem 13.150.*
Write metricity three times with cyclically permuted indices,

$$\begin{aligned} \partial_{\lambda} g_{\mu\nu} &= \Gamma_{\mu\lambda\nu} + \Gamma_{\nu\lambda\mu}\,, \\ \partial_{\mu} g_{\nu\lambda} &= \Gamma_{\nu\mu\lambda} + \Gamma_{\lambda\mu\nu}\,, \\ \partial_{\nu} g_{\lambda\mu} &= \Gamma_{\lambda\nu\mu} + \Gamma_{\mu\nu\lambda}\,, \end{aligned} \tag{13.300}$$

and form the combination (second) + (third) − (first). Grouping with the torsion Equation (13.298), $\Gamma_{\lambda\mu\nu} + \Gamma_{\lambda\nu\mu} = 2\Gamma_{\lambda\mu\nu} - T_{\lambda\mu\nu}$, $\Gamma_{\nu\mu\lambda} - \Gamma_{\nu\lambda\mu} = T_{\nu\mu\lambda}$, and $\Gamma_{\mu\nu\lambda} - \Gamma_{\mu\lambda\nu} = T_{\mu\nu\lambda}$, so that

$$\partial_{\mu} g_{\nu\lambda} + \partial_{\nu} g_{\lambda\mu} - \partial_{\lambda} g_{\mu\nu} = 2\Gamma_{\lambda\mu\nu} - T_{\lambda\mu\nu} + T_{\nu\mu\lambda} + T_{\mu\nu\lambda}\,. \tag{13.301}$$

Solving for $\Gamma_{\lambda\mu\nu}$ and using the antisymmetry $T_{\nu\mu\lambda} = -T_{\nu\lambda\mu}$, $T_{\mu\nu\lambda} = -T_{\mu\lambda\nu}$ to reorder gives Equation (13.299). Uniqueness is manifest: the metric determines $\mathring{\Gamma}$, the prescribed torsion determines K. The outer antisymmetry of K is a two-line check using $T_{\lambda\mu\nu} = -T_{\lambda\nu\mu}$, and it is precisely the condition that the contorsion contribution preserve metricity. □

Remark 13.151 (Geodesics versus autoparallels). Curves that extremize the length functional $\int \sqrt{|\mathrm{d}s^2|}$ obey Equation (13.297) with $\mathring{\Gamma}$ (the variational computation of Chapter 21 applied to the line element). Since $K^{\lambda}{}_{(\mu\nu)} \neq 0$ in general, autoparallels of a torsionful connection differ from geodesics of its metric: straightest and shortest part ways.

13.10.1 The spin connection

Frame-indexed fields need their own connection: local Lorentz transformations Equation (13.265) are point-dependent, so $\partial_\mu V^a$ is not covariant. The *spin connection* $\omega_\mu{}^a{}_b$ enters as

$$\nabla_\mu V^a = \partial_\mu V^a + \omega_\mu{}^a{}_b V^b \,, \tag{13.302}$$

and the requirement that frame and coordinate descriptions agree — the *vielbein postulate*, total covariant constancy of the vielbein — ties it to Γ:

$$\nabla_\mu e^a{}_\nu = \partial_\mu e^a{}_\nu + \omega_\mu{}^a{}_b\, e^b{}_\nu - \Gamma^\lambda{}_{\mu\nu}\, e^a{}_\lambda = 0 \,. \tag{13.303}$$

Metricity of Γ is then equivalent to $\omega_{\mu ab} = -\omega_{\mu ba}$: the spin connection takes values in the Lorentz algebra $\mathfrak{so}(p,q)$ (Chapter 14). Under a local Lorentz transformation it transforms inhomogeneously,

$$\omega_\mu \longrightarrow \Lambda\, \omega_\mu\, \Lambda^{-1} - (\partial_\mu \Lambda)\, \Lambda^{-1} \,, \tag{13.304}$$

exactly like a gauge potential — the observation the Cartan formalism turns into machinery (Section 13.12).

13.11 Curvature

In the theory of curves a curve is characterized by its curvatures, built from first and second derivatives (Section 13.3.1); the curvature of a manifold must likewise live in the first and second derivatives of the metric, and vanish for Euclidean space. The correct object measures the failure of parallel transport around a closed loop, i.e. the non-commutativity of covariant derivatives.

Theorem 13.152 (Riemann tensor; Ricci identity with torsion)**.** *For any affine connection and any vector field,*

$$[\nabla_\mu, \nabla_\nu]\, V^\lambda = R^\lambda{}_{\rho\mu\nu}\, V^\rho - T^\rho{}_{\mu\nu}\, \nabla_\rho V^\lambda \,, \tag{13.305}$$

with

$$R^\lambda{}_{\rho\mu\nu} = \partial_\mu \Gamma^\lambda{}_{\nu\rho} - \partial_\nu \Gamma^\lambda{}_{\mu\rho} + \Gamma^\lambda{}_{\mu\sigma}\Gamma^\sigma{}_{\nu\rho} - \Gamma^\lambda{}_{\nu\sigma}\Gamma^\sigma{}_{\mu\rho} \,. \tag{13.306}$$

$R^\lambda{}_{\rho\mu\nu}$ *is a tensor: the* Riemann curvature tensor.

Rests on Definitions 13.145 and 13.147 and Proposition 7.73.

Proof. *Derives Theorem 13.152.*

Expand $\nabla_\mu \nabla_\nu V^\lambda = \partial_\mu(\nabla_\nu V^\lambda) + \Gamma^\lambda{}_{\mu\sigma}\nabla_\nu V^\sigma - \Gamma^\rho{}_{\mu\nu}\nabla_\rho V^\lambda$ (the middle slot of $\nabla_\nu V^\lambda$ is a covector index) and antisymmetrize in $\mu\nu$. The $\partial\partial V$ terms cancel by equality of mixed partials; the terms with one Γ and one ∂V cancel except for those coming from the last piece, which assemble into $-T^\rho{}_{\mu\nu}\nabla_\rho V^\lambda$ by Equation (13.298); the remaining terms contain no derivatives of V and define Equation (13.306). Tensoriality follows because the left side of Equation (13.305) and the torsion term are tensors for arbitrary V. □

Definition 13.153 (Contractions)**.** The *Ricci tensor* and *curvature scalar* are

$$R_{\mu\nu} = R^\lambda{}_{\mu\lambda\nu} \,, \qquad R = g^{\mu\nu} R_{\mu\nu} \,. \tag{13.307}$$

For a surface ($D = 2$, torsion-free) $R = 2K$ with K the Gaussian curvature of Section 13.3.3, closing the circle with the classical theory.

Rests on Theorem 13.152 and Definition 13.117.

Proposition 13.154 (Symmetries of the curvature)**.** *Always:* $R^{\lambda}{}_{\rho\mu\nu} = -R^{\lambda}{}_{\rho\nu\mu}$. *With metricity:* $R_{\lambda\rho\mu\nu} = -R_{\rho\lambda\mu\nu}$. *For the Levi-Civita connection additionally*

$$R_{\lambda\rho\mu\nu} = R_{\mu\nu\lambda\rho}\,, \qquad R^{\lambda}{}_{[\rho\mu\nu]} = 0\,, \qquad \nabla_{[\sigma}R^{\lambda}{}_{|\rho|\mu\nu]} = 0\,, \tag{13.308}$$

(first and second Bianchi identities). With torsion the last two acquire torsion terms, displayed in form language in Proposition 13.157.

Rests on Theorems 13.150 and 13.152, Definition 13.149, and Proposition 13.157.

Proof. *Derives Proposition 13.154.*

The first is manifest from Equation (13.306). The second: apply Equation (13.305) to the scalar $g_{\lambda\rho}V^{\lambda}W^{\rho}$ and use metricity. The torsion-free identities follow from Equation (13.306) with $\Gamma^{\lambda}{}_{[\mu\nu]} = 0$ by direct cyclic summation — carried out most efficiently in the Cartan formalism below (Proposition 13.157 with $T^{a} = 0$); the pair-exchange symmetry is an algebraic consequence of the other three. □

Proposition 13.155 (Geodesic deviation; Jacobi equation)**.** *Let* $x^{\lambda}(\tau, s)$ *be a smooth one-parameter family of geodesics of a torsion-free connection, with tangent* $u^{\lambda} = \partial x^{\lambda}/\partial\tau$ *and connecting field* $\xi^{\lambda} = \partial x^{\lambda}/\partial s$. *Then* ξ *obeys the* Jacobi equation

$$\frac{\mathrm{D}^2\xi^{\lambda}}{\mathrm{d}\tau^2} = R^{\lambda}{}_{\rho\mu\nu}\,u^{\rho}\,u^{\mu}\,\xi^{\nu}\,, \tag{13.309}$$

with $\mathrm{D}/\mathrm{d}\tau = u^{\mu}\nabla_{\mu}$ *the covariant derivative along the curve.*

Rests on Theorem 13.152 and Definition 13.146.

Proof. *Derives Proposition 13.155.*

The coordinate fields commute: $[u,\xi]^{\lambda} = u^{\mu}\partial_{\mu}\xi^{\lambda} - \xi^{\mu}\partial_{\mu}u^{\lambda} = \partial_{\tau}\partial_{s}x^{\lambda} - \partial_{s}\partial_{\tau}x^{\lambda} = 0$. For any connection, $\nabla_{u}\xi^{\lambda} - \nabla_{\xi}u^{\lambda} = [u,\xi]^{\lambda} + T^{\lambda}{}_{\mu\nu}u^{\mu}\xi^{\nu}$ by Equation (13.298), so with $T = 0$ the two mixed derivatives agree, $\nabla_{u}\xi = \nabla_{\xi}u$. Hence

$$\frac{\mathrm{D}^2\xi^{\lambda}}{\mathrm{d}\tau^2} = \nabla_{u}\nabla_{u}\xi^{\lambda} = \nabla_{u}\nabla_{\xi}u^{\lambda} = \nabla_{\xi}\nabla_{u}u^{\lambda} + R^{\lambda}{}_{\rho\mu\nu}\,u^{\rho}u^{\mu}\xi^{\nu}\,,$$

the exchange being the Ricci identity Equation (13.305) applied to the pair (u,ξ): the commutator term $\nabla_{[u,\xi]}$ vanishes because $[u,\xi] = 0$, and the torsion term because $T = 0$. The first term on the right vanishes because every member of the family is a geodesic — $\nabla_{u}u = 0$ identically in s, and ∇_{ξ} differentiates that identity. □

13.12 The Cartan formalism

The results of Sections 13.8, 13.10 and 13.11 compress into two lines when written in the language of differential forms (Section 13.6), with all coordinate indices hidden. The dynamical variables are the vielbein one-forms e^{a} and the $\mathfrak{so}(p,q)$-valued *connection one-form*

$$\omega^{a}{}_{b} = \omega_{\mu}{}^{a}{}_{b}\,\mathrm{d}x^{\mu}\,, \qquad \omega^{ab} = -\omega^{ba}\,, \tag{13.310}$$

and the *Lorentz-covariant exterior derivative* acts on frame-indexed forms as $\mathrm{D}\alpha^{a} = \mathrm{d}\alpha^{a} + \omega^{a}{}_{b}\wedge\alpha^{b}$ (one ω term per frame index).

Theorem 13.156 (Cartan structure equations)**.** *The torsion and curvature of Definition 13.147 and Theorem 13.152 are the two-forms*

$$T^{a} = \mathrm{d}e^{a} + \omega^{a}{}_{b}\wedge e^{b} = \mathrm{D}e^{a} = \tfrac{1}{2}\,T^{a}{}_{\mu\nu}\,\mathrm{d}x^{\mu}\wedge\mathrm{d}x^{\nu}\,, \tag{13.311}$$

$$R^{a}{}_{b} = \mathrm{d}\omega^{a}{}_{b} + \omega^{a}{}_{c}\wedge\omega^{c}{}_{b} = \tfrac{1}{2}\,R^{a}{}_{b\mu\nu}\,\mathrm{d}x^{\mu}\wedge\mathrm{d}x^{\nu}\,, \tag{13.312}$$

with $T^{a}{}_{\mu\nu} = e^{a}{}_{\lambda}T^{\lambda}{}_{\mu\nu}$ *and* $R^{a}{}_{b\mu\nu} = e^{a}{}_{\lambda}\,e_{b}{}^{\rho}\,R^{\lambda}{}_{\rho\mu\nu}$.

Rests on Definitions 13.102, 13.103, 13.122 and 13.147, Equation (13.303), and Theorem 13.152.

Proof. *Derives Theorem 13.156.*

For Equation (13.311): by the vielbein postulate Equation (13.303), $\partial_\mu e^a{}_\nu + \omega_\mu{}^a{}_b e^b{}_\nu = \Gamma^\lambda{}_{\mu\nu} e^a{}_\lambda$; antisymmetrizing in $\mu\nu$, the left side gives the components of $\mathrm{d}e^a + \omega^a{}_b \wedge e^b$ and the right side gives $\frac{1}{2}(\Gamma^\lambda{}_{\mu\nu} - \Gamma^\lambda{}_{\nu\mu}) e^a{}_\lambda = \frac{1}{2} T^\lambda{}_{\mu\nu} e^a{}_\lambda$. For Equation (13.312): solve Equation (13.303) for $\omega_\mu{}^a{}_b$ in terms of Γ and e, insert into $\mathrm{d}\omega + \omega \wedge \omega$, and the inhomogeneous ∂e contributions cancel, leaving exactly the components Equation (13.306) converted with two vielbeins. (Equivalently: $\mathrm{D}^2 V^a = R^a{}_b V^b$ for any zero-form V^a, by a two-line computation, which identifies Equation (13.312) as the frame version of the Ricci identity Equation (13.305).) □

Proposition 13.157 (Bianchi identities)**.** *Applying* d *to the structure equations,*

$$\mathrm{D}T^a = R^a{}_b \wedge e^b \,, \tag{13.313}$$

$$\mathrm{D}R^a{}_b = 0 \,. \tag{13.314}$$

In the torsion-free case, Equation (13.313) *reduces to the first Bianchi identity* $R^a{}_b \wedge e^b = 0$*, i.e.* $R^\lambda{}_{[\rho\mu\nu]} = 0$*, and Equation* (13.314) *is the second identity of Equation* (13.308)*.*

Rests on Theorem 13.156 and Equations (13.245) *and* (13.246)*.*

Proof. *Derives Proposition 13.157.*

Differentiate Equation (13.311) with $\mathrm{d}^2 = 0$ (Section 13.6.4):

$$\begin{aligned} \mathrm{d}T^a &= \mathrm{d}\omega^a{}_b \wedge e^b - \omega^a{}_b \wedge \mathrm{d}e^b \\ &= (R^a{}_b - \omega^a{}_c \wedge \omega^c{}_b) \wedge e^b - \omega^a{}_b \wedge \left(T^b - \omega^b{}_c \wedge e^c\right) \\ &= R^a{}_b \wedge e^b - \omega^a{}_b \wedge T^b \,, \end{aligned} \tag{13.315}$$

which is Equation (13.313) after moving the $\omega \wedge T$ term to the left as part of $\mathrm{D}T^a$. Differentiating Equation (13.312) likewise, $\mathrm{d}R^a{}_b = \mathrm{d}\omega^a{}_c \wedge \omega^c{}_b - \omega^a{}_c \wedge \mathrm{d}\omega^c{}_b = R^a{}_c \wedge \omega^c{}_b - \omega^a{}_c \wedge R^c{}_b$ (the $\omega \wedge \omega \wedge \omega$ terms cancel pairwise), which is $\mathrm{D}R^a{}_b = 0$. □

Proposition 13.158 (The connection determined by vielbein and torsion)**.** *Given* e^a *and a prescribed torsion two-form* T^a*, the first structure equation Equation* (13.311) *determines* $\omega^{ab} = -\omega^{ba}$ *uniquely; for* $T^a = 0$*,* $\omega = \omega(e)$ *is the frame form of the Levi-Civita connection, with the contorsion one-form* $K^a{}_b$ *(*$T^a = K^a{}_b \wedge e^b$*) supplying the general case, in exact parallel with Equation* (13.299)*.*

Rests on Theorems 13.150 and 13.156 and Equation (13.266)*.*

Proof. *Derives Proposition 13.158.*

Counting: $\omega_\mu{}^{ab}$ has $D \cdot D(D-1)/2$ independent components, exactly the number of components of the two-form system Equation (13.311); the explicit solution repeats the cyclic permutation trick of Theorem 13.150 on the anholonomy coefficients Equation (13.266),

$$\omega_{\mu ab} = \tfrac{1}{2}\, e_\mu{}^c \,(C_{abc} - C_{bca} - C_{cab}) + K_{\mu ab} \quad \text{with } C_{abc} = \eta_{cd}\, C^d{}_{ab} \,, \tag{13.316}$$

whose verification is the same computation as in Theorem 13.150 transcribed to frame indices. □

Remark 13.159 (Gravity as gauge geometry). Equations Equations (13.304), (13.311) and (13.312) exhibit the frame bundle of (M, g) as a gauge structure with group $\mathrm{L}_{p+q} = \mathrm{SO}(p, q)$: ω is the gauge potential, R its field strength, and e a Lorentz-vector-valued one-form whose field strength is the torsion. The fibre-bundle formulation is taken up in Chapter 14; the physical instantiation is general relativity (Part V), where the Einstein–Hilbert action becomes

$$S_{\mathrm{EH}} = \frac{1}{2\kappa\,(D-2)!} \int \epsilon_{a_1 \ldots a_D}\, R^{a_1 a_2} \wedge e^{a_3} \wedge \cdots \wedge e^{a_D} = \frac{1}{2\kappa} \int \mathrm{d}^D x \, \sqrt{|g|}\, R \,, \tag{13.317}$$

the equality of the two integrands being proven in Appendix A.8 and used in Chapter 43.

13.13 Maximally symmetric spaces and the symmetry ladder

The spaces admitting the full complement of $D(D+1)/2$ Killing vectors (Proposition 13.140) are characterized by a curvature built from the metric alone, with one constant:

$$R_{\lambda\rho\mu\nu} = \frac{\epsilon}{\ell^2}\left(g_{\lambda\mu}g_{\rho\nu} - g_{\lambda\nu}g_{\rho\mu}\right), \qquad \epsilon \in \{+1,\, 0,\, -1\}\,, \tag{13.318}$$

with $T^a = 0$ and ℓ a length (for $\epsilon = 0$ the curvature vanishes and ℓ drops out). In form language, $R^{ab} = (\epsilon/\ell^2)\, e^a \wedge e^b$.

Theorem 13.160 (Maximal symmetry forces constant curvature)**.** *Let (M, g) be a connected pseudo-Riemannian manifold of dimension $D \geq 2$ with the Levi-Civita connection, and suppose it admits the maximal number $D(D+1)/2$ of linearly independent Killing vectors. Then its curvature has the form Equation (13.318), with*

$$\frac{\epsilon}{\ell^2} = \frac{R}{D(D-1)}\,, \tag{13.319}$$

R being the constant curvature scalar of Equation (13.307).

Rests on Propositions 13.140 and 13.154 and Definition 13.153.

Proof. *Derives Theorem 13.160.*

Step 1: the data a Killing vector may carry at a point. The proof of Proposition 13.140 showed that a Killing vector is determined throughout M by the D numbers $\xi_\mu(P)$ and the $D(D-1)/2$ numbers $\nabla_{[\mu}\xi_{\nu]}(P)$ at one point P; that is, the linear map sending ξ to this data is injective. By hypothesis its domain has dimension $D(D+1)/2$, which is exactly the dimension of its target, so the map is a bijection: for *every* vector v_μ and every antisymmetric $A_{\mu\nu}$ there is a Killing vector with $\xi_\mu(P) = v_\mu$ and $\nabla_\mu\xi_\nu(P) = A_{\mu\nu}$, the latter being antisymmetric already by Killing's equation Equation (13.290).

Step 2: isotropy of the curvature. The flow of a Killing vector consists of isometries (Definition 13.139), and the Riemann tensor is built from the metric by a rule that commutes with diffeomorphisms, so $\phi_t^* R[g] = R[\phi_t^* g] = R[g]$; differentiating at $t = 0$ gives $\mathcal{L}_\xi R = 0$. Writing the Lie derivative of a $(0,4)$ tensor with covariant derivatives, which is legitimate for a torsion-free connection (Proposition 13.127),

$$\begin{aligned} 0 = \xi^\sigma\nabla_\sigma R_{\lambda\rho\mu\nu} + (\nabla_\lambda\xi^\sigma)\, R_{\sigma\rho\mu\nu} + (\nabla_\rho\xi^\sigma)\, R_{\lambda\sigma\mu\nu} \\ + (\nabla_\mu\xi^\sigma)\, R_{\lambda\rho\sigma\nu} + (\nabla_\nu\xi^\sigma)\, R_{\lambda\rho\mu\sigma}\,. \end{aligned} \tag{13.320}$$

Evaluate at P on a Killing vector with $\xi(P) = 0$ and $\nabla_\mu\xi_\nu(P) = A_{\mu\nu}$, which Step 1 provides for arbitrary antisymmetric A. The transport term drops and

$$A_\lambda{}^\sigma R_{\sigma\rho\mu\nu} + A_\rho{}^\sigma R_{\lambda\sigma\mu\nu} + A_\mu{}^\sigma R_{\lambda\rho\sigma\nu} + A_\nu{}^\sigma R_{\lambda\rho\mu\sigma} = 0 \tag{13.321}$$

holds at P for every antisymmetric A.

Step 3: the algebra. Fix two indices α, β and take $A_{\sigma\tau} = g_{\sigma\alpha}g_{\tau\beta} - g_{\sigma\beta}g_{\tau\alpha}$, so that $A_\lambda{}^\sigma = g_{\lambda\alpha}\delta^\sigma{}_\beta - g_{\lambda\beta}\delta^\sigma{}_\alpha$. Then Equation (13.321) becomes

$$\begin{aligned} g_{\lambda\alpha}R_{\beta\rho\mu\nu} - g_{\lambda\beta}R_{\alpha\rho\mu\nu} + g_{\rho\alpha}R_{\lambda\beta\mu\nu} - g_{\rho\beta}R_{\lambda\alpha\mu\nu} \\ + g_{\mu\alpha}R_{\lambda\rho\beta\nu} - g_{\mu\beta}R_{\lambda\rho\alpha\nu} + g_{\nu\alpha}R_{\lambda\rho\mu\beta} - g_{\nu\beta}R_{\lambda\rho\mu\alpha} = 0\,. \end{aligned} \tag{13.322}$$

Contract Equation (13.322) with $g^{\lambda\alpha}$. The first term gives $D\, R_{\beta\rho\mu\nu}$; the second and the third give $-R_{\beta\rho\mu\nu}$ each, the third after using $R_{\rho\beta\mu\nu} = -R_{\beta\rho\mu\nu}$ (Proposition 13.154); the fourth vanishes by the same antisymmetry; the sixth and eighth produce Ricci tensors

(Definition 13.153), $-g_{\mu\beta}R_{\rho\nu}$ and $+g_{\nu\beta}R_{\rho\mu}$; the fifth and seventh survive as $R_{\mu\rho\beta\nu}$ and $R_{\nu\rho\mu\beta}$. Hence

$$(D-2)R_{\beta\rho\mu\nu} + R_{\mu\rho\beta\nu} + R_{\nu\rho\mu\beta} = g_{\mu\beta}R_{\rho\nu} - g_{\nu\beta}R_{\rho\mu}\,. \tag{13.323}$$

Use the pair-exchange symmetry of Equation (13.308) to write $R_{\mu\rho\beta\nu} = R_{\beta\nu\mu\rho}$ and $R_{\nu\rho\mu\beta} = -R_{\beta\mu\nu\rho}$, both with β in the first slot, and then the first Bianchi identity $R_{\beta[\rho\mu\nu]} = 0$ in the form $R_{\beta\nu\mu\rho} = R_{\beta\rho\mu\nu} + R_{\beta\mu\nu\rho}$. The last two terms of the left side of Equation (13.323) collapse to $R_{\beta\rho\mu\nu}$, leaving

$$(D-1)R_{\lambda\rho\mu\nu} = g_{\lambda\mu}R_{\rho\nu} - g_{\lambda\nu}R_{\rho\mu}\,, \tag{13.324}$$

after renaming β to λ.

The left side of Equation (13.324) is antisymmetric under $\lambda \leftrightarrow \rho$, so the right side must be too:

$$g_{\lambda\mu}R_{\rho\nu} - g_{\lambda\nu}R_{\rho\mu} = -g_{\rho\mu}R_{\lambda\nu} + g_{\rho\nu}R_{\lambda\mu}\,.$$

Contracting this with $g^{\lambda\mu}$ gives $(D-1)R_{\rho\nu} = -R_{\rho\nu} + g_{\rho\nu}R$, that is

$$R_{\rho\nu} = \frac{R}{D}\,g_{\rho\nu}\,. \tag{13.325}$$

Substituting Equation (13.325) into Equation (13.324),

$$R_{\lambda\rho\mu\nu} = \frac{R}{D(D-1)}\left(g_{\lambda\mu}g_{\rho\nu} - g_{\lambda\nu}g_{\rho\mu}\right)\,, \tag{13.326}$$

which is the tensor structure of Equation (13.318) at the point P, and P was arbitrary.

Step 4: the coefficient is constant. By Step 1 there is, at each point P and for each vector v, a Killing vector with $\xi(P) = v$. Since R is a scalar built from the metric and the flow of ξ consists of isometries, $\mathcal{L}_\xi R = \xi^\sigma \partial_\sigma R = 0$ (Equation (13.272)), so $v^\sigma \partial_\sigma R = 0$ at P for every v, whence $\partial_\sigma R = 0$ there. As P was arbitrary and M is connected, R is constant. Writing that constant as Equation (13.319) — with $\epsilon = 0$ when $R = 0$, and otherwise $\epsilon = \pm 1$ the sign of R and $\ell = \left(|R|\,/D(D-1)\right)^{-1/2}$ the associated length — turns Equation (13.326) into Equation (13.318). □

Remark 13.161. Step 4 did not need the second Bianchi identity, only homogeneity. Where a space is isotropic about every point but the full Killing count is not assumed, Equation (13.326) still holds with a function $K(x)$ in place of the constant, and inserting it into the second Bianchi identity of Equation (13.308) and contracting twice yields $(D-1)(D-2)\nabla_\sigma K = 0$: for $D > 2$ pointwise isotropy alone already forces K to be constant. This is Schur's theorem, and it is the reason the three cases below exhaust the possibilities.

The three cases realize the groups of Table 13.1:

Flat space ($\epsilon = 0$). The manifold is $\mathbb{R}^{p,q}$ with metric $\eta_{\mu\nu}$; the Killing vectors are the D translations ∂_μ (group T_{p+q}, abelian) and the $D(D-1)/2$ pseudo-rotations $x_\mu\partial_\nu - x_\nu\partial_\mu$ (group $\mathrm{L}_{p+q} = \mathrm{SO}(p,q)$), closing into the Poincaré group $\mathrm{P}_{p+q} = \mathrm{ISO}(p,q)$, the semidirect product of the two.

Anti-de Sitter space ($\epsilon = -1$). AdS_{p+q} is the quadric

$$\eta_{AB}\,X^A X^B = -\ell^2 \tag{13.327}$$

in the flat $(p+q+1)$-dimensional space of signature $(p, q+1)$, with embedding coordinates X^m and frame metric η_{AB} per the conventions of Section 13.1. The induced metric (Definition 13.119) satisfies Equation (13.318) with $\epsilon = -1$, by Proposition 13.162 below. Every linear transformation of X^m preserving η_{AB} preserves the quadric and its induced metric, so $\mathrm{Isom}(\mathrm{AdS}_{p+q}) \supseteq \mathrm{SO}(p, q+1)$; its dimension $(D+1)D/2$ saturates the Killing bound, so equality holds (connected component).

de Sitter space ($\epsilon = +1$). Identically, dS_{p+q} is the quadric $\eta_{AB}X^A X^B = +\ell^2$ in signature $(p+1, q)$, with isometry group $\mathrm{SO}(p+1, q)$.

The curvature of the two quadrics is computed once, for both signs at once, by the Gauss equation of Section 13.3.2 transcribed to a pseudo-Riemannian hypersurface of a flat space.

Proposition 13.162 (Curvature induced on the quadric)**.** *Let η_{AB} be the flat metric of a $(D+1)$-dimensional space with Cartesian coordinates X^A, let $\sigma = \pm 1$, and let Σ be the quadric*

$$\eta_{AB}\, X^A X^B = \sigma\, \ell^2 \,, \tag{13.328}$$

described by an embedding $X^A(\xi^i)$, $i = 1, \dots, D$. Then the metric h_{ij} induced on Σ (Definition 13.119) has curvature

$$R_{ijkl} = \frac{\sigma}{\ell^2}\left(h_{ik}h_{jl} - h_{il}h_{jk}\right) , \tag{13.329}$$

that is, Equation (13.318) with $\epsilon = \sigma$.

Rests on Definition 13.119, Equation (13.83), and Theorem 13.152.

Proof. *Derives Proposition 13.162.*

Write $e_i^A = \partial X^A/\partial\xi^i$ for the tangent vectors, so that $h_{ij} = \eta_{AB}e_i^A e_j^B$ by Equation (13.261). Differentiating Equation (13.328) along Σ gives $\eta_{AB}X^A e_i^B = 0$: the position vector is normal to the quadric. Hence

$$N^A = \frac{X^A}{\ell} \,, \qquad \eta_{AB}N^A N^B = \sigma \,, \qquad \frac{\partial N^A}{\partial \xi^i} = \frac{e_i^A}{\ell} \,, \tag{13.330}$$

the last relation because X^A is itself the embedding.

At each point the $D+1$ vectors $\left\{e_i^A, N^A\right\}$ are a basis of the ambient space, so we may expand, exactly as in the Gauss equations Equation (13.83),

$$\frac{\partial e_j^A}{\partial \xi^i} = \Gamma^k{}_{ij}\, e_k^A + \sigma\, b_{ij}\, N^A \,, \qquad b_{ij} = \eta_{AB}N^A \frac{\partial e_j^B}{\partial \xi^i} \,, \tag{13.331}$$

the factor σ being what makes the stated formula for b_{ij} follow from Equation (13.330) on contracting with $\eta_{AB}N^B$; contracting instead with $\eta_{AB}e_k^B$ identifies $\Gamma^k{}_{ij}$ as the Levi-Civita connection of h_{ij}, by the same computation that produced Equation (13.87) from Equation (13.83). Differentiating $\eta_{AB}N^A e_j^B = 0$ with respect to ξ^i and using Equation (13.330),

$$b_{ij} = -\eta_{AB}\left(\frac{\partial N^A}{\partial \xi^i}\right) e_j^B = -\frac{h_{ij}}{\ell} \,. \tag{13.332}$$

The second fundamental form of the quadric is thus proportional to its first: every direction is a curvature direction, which is the analytic form of the statement that the quadric looks the same from every side.

The ambient space is flat and referred to Cartesian coordinates, so $\partial_i\partial_j e_k^A$ is symmetric in i and j. Differentiating Equation (13.331) once more and antisymmetrizing in i, j, the component along e_m^A must vanish:

$$\partial_i\Gamma^m{}_{jk} - \partial_j\Gamma^m{}_{ik} + \Gamma^m{}_{il}\Gamma^l{}_{jk} - \Gamma^m{}_{jl}\Gamma^l{}_{ik} + \frac{\sigma}{\ell}\left(b_{jk}\delta^m_{\ i} - b_{ik}\delta^m_{\ j}\right) = 0 \,,$$

the last bracket arising from the terms $\sigma b_{jk}\partial_i N^A$ through Equation (13.330). The first four terms are precisely $R^m{}_{kij}$ of the induced metric, in the convention Equation (13.306). Inserting Equation (13.332),

$$R^m{}_{kij} = -\frac{\sigma}{\ell}\left(b_{jk}\delta^m_{\ i} - b_{ik}\delta^m_{\ j}\right) = \frac{\sigma}{\ell^2}\left(h_{jk}\delta^m_{\ i} - h_{ik}\delta^m_{\ j}\right) \,,$$

and lowering the first index with h_{lm} and renaming gives Equation (13.329).

For anti-de Sitter space Equation (13.327) has $\sigma = -1$, so $\epsilon = -1$; for de Sitter space $\sigma = +1$, so $\epsilon = +1$. In both cases the induced metric is nondegenerate of the signature stated, because the normal direction Equation (13.330) carries the extra sign σ of the ambient metric and is removed from it. □

In the limit $\ell \to \infty$ both quadrics flatten and their isometry groups contract to P_{p+q} — the İnönü–Wigner contraction, treated algebraically in Chapter 14.

13.13.1 The conformal group

Relaxing isometry to angle preservation, $f^*g = \Omega^2(x)\, g$, enlarges the symmetry. For flat space of signature (p, q), $D = p + q > 2$, the conformal Killing vectors are

$$\begin{aligned} \text{translations}: &\quad \xi^\mu = a^\mu && (D)\,, \\ \text{pseudo-rotations}: &\quad \xi^\mu = \lambda^\mu{}_\nu x^\nu,\ \lambda_{\mu\nu} = -\lambda_{\nu\mu} && (D(D-1)/2)\,, \\ \text{dilatation}: &\quad \xi^\mu = x^\mu && (1)\,, \\ \text{special conformal}: &\quad \xi^\mu = 2(b\cdot x)x^\mu - x^2 b^\mu && (D)\,, \end{aligned} \tag{13.333}$$

for a total of $(D+1)(D+2)/2$ generators. That this list is complete is the content of the following proposition.

Proposition 13.163 (Integration of the conformal Killing equation)**.** *Let $\eta_{\mu\nu}$ be the flat metric of signature (p, q) in Cartesian coordinates, $D = p + q > 2$. Every solution of the conformal Killing equation*

$$\partial_\mu \xi_\nu + \partial_\nu \xi_\mu = 2\sigma\, \eta_{\mu\nu}\,, \qquad \sigma = \frac{1}{D}\,\partial_\lambda \xi^\lambda\,, \tag{13.334}$$

is a polynomial of degree at most two in x, and the solution space is spanned by the vector fields Equation (13.333); its dimension is $(D+1)(D+2)/2$.

Rests on Equations (13.275) and (13.333) and Definition 13.139.

Proof. *Derives Proposition 13.163.*

Equation (13.334) is $\mathcal{L}_\xi \eta = 2\sigma\eta$ written out with Equation (13.275), the trace of which fixes σ as stated.

Step 1: second derivatives in terms of σ. Put $E_{\mu\nu} = \partial_\mu\xi_\nu + \partial_\nu\xi_\mu - 2\sigma\eta_{\mu\nu}$, so that $E_{\mu\nu} = 0$. Form the combination $\partial_\rho E_{\mu\nu} + \partial_\mu E_{\nu\rho} - \partial_\nu E_{\rho\mu} = 0$. Six of the eight derivative terms cancel in pairs because partial derivatives commute, and what survives is

$$\partial_\mu\partial_\rho\xi_\nu = \eta_{\mu\nu}\partial_\rho\sigma + \eta_{\nu\rho}\partial_\mu\sigma - \eta_{\rho\mu}\partial_\nu\sigma\,. \tag{13.335}$$

Step 2: σ is affine. Apply ∂^ν to Equation (13.335). On the left this gives $\partial_\mu\partial_\rho\,(\partial^\nu\xi_\nu) = D\,\partial_\mu\partial_\rho\sigma$, and on the right $2\partial_\mu\partial_\rho\sigma - \eta_{\rho\mu}\Box\sigma$ with $\Box = \partial^\nu\partial_\nu$, whence

$$(D-2)\,\partial_\mu\partial_\rho\sigma = -\eta_{\mu\rho}\,\Box\sigma\,. \tag{13.336}$$

Tracing Equation (13.336) gives $(D-2)\Box\sigma = -D\,\Box\sigma$, that is $\Box\sigma = 0$ for $D \neq 1$; and then Equation (13.336) itself gives $\partial_\mu\partial_\rho\sigma = 0$, since $D > 2$. Hence σ is an affine function, which we write

$$\sigma(x) = \lambda + 2\,b_\mu x^\mu\,, \tag{13.337}$$

with λ and b_μ constant, the factor 2 chosen for later convenience. This is the only place where $D > 2$ is used, and it is where the argument genuinely fails in two dimensions, whose conformal algebra is infinite-dimensional.

Step 3: ξ is quadratic. With $\partial\partial\sigma = 0$ the right-hand side of Equation (13.335) is constant, so all third derivatives of ξ vanish and its Taylor expansion about the origin terminates:

$$\xi_\nu(x) = a_\nu + b_{\nu\mu}x^\mu + \tfrac{1}{2}\left(\partial_\mu\partial_\rho\xi_\nu\right)x^\mu x^\rho\,, \qquad a_\nu = \xi_\nu(0)\,, \quad b_{\nu\mu} = \partial_\mu\xi_\nu(0)\,. \tag{13.338}$$

Step 4: the three orders. Each order of Equation (13.338) is constrained separately, because Equation (13.334) must hold identically in x.

1. *Constant part.* a_ν is unconstrained: D parameters, the translations $\xi^\mu = a^\mu$.
2. *Linear part.* Here Equation (13.334) reads $b_{\nu\mu} + b_{\mu\nu} = 2\lambda\eta_{\mu\nu}$, so $b_{\nu\mu} = \lambda\eta_{\nu\mu} + \lambda_{\nu\mu}$ with $\lambda_{\nu\mu} = -\lambda_{\mu\nu}$ arbitrary: one parameter for the dilatation $\xi^\mu = \lambda\, x^\mu$ and $D(D-1)/2$ for the pseudo-rotations $\xi^\mu = \lambda^\mu{}_\nu x^\nu$.
3. *Quadratic part.* By Equations (13.335) and (13.337), $\partial_\mu\partial_\rho\xi_\nu = 2\left(\eta_{\mu\nu}b_\rho + \eta_{\nu\rho}b_\mu - \eta_{\rho\mu}b_\nu\right)$, so the last term of Equation (13.338) is

$$\left(\eta_{\mu\nu}b_\rho + \eta_{\nu\rho}b_\mu - \eta_{\rho\mu}b_\nu\right)x^\mu x^\rho = 2\,(b\cdot x)\,x_\nu - x^2 b_\nu\,,$$

   which is the special conformal transformation of Equation (13.333): D parameters b_μ.

Conversely each of the four families satisfies Equation (13.334) by direct substitution — for the last, with $\sigma = 2\,b\cdot x$ — so the solution space is exactly their span. The parameters are independent, being the coefficients of distinct powers of x together with the irreducible pieces of $b_{\nu\mu}$, and they number

$$D + 1 + \frac{D(D-1)}{2} + D = \frac{(D+1)(D+2)}{2}\,, \tag{13.339}$$

which is the dimension of $\mathrm{SO}(p+1, q+1)$ quoted below. □

These close into the conformal group $\mathrm{C}_{p+q} = \mathrm{SO}(p+1, q+1)$ of Table 13.1, realized *linearly* on the $(p+q+2)$-dimensional flat space of signature $(p+1, q+1)$: flat (p,q)-space is identified with rays of the null cone

$$\eta_{MN}\,X^M X^N = 0\,, \tag{13.340}$$

(projective light-cone construction), on which $\mathrm{SO}(p+1, q+1)$ acts by its defining representation, indices M, N and I, J per Section 13.1. The dimension matches: $\dim \mathrm{SO}(p+1, q+1) = (D+1)(D+2)/2$.

13.13.2 The observed case

Physics instantiates this chapter at $p+q = 3+1$. The kinematics of Part IV is the flat row of Table 13.1; general relativity (Part V) is the torsion-free, Levi-Civita case of Sections 13.10 to 13.12, tested by the experiments recorded there; and the accelerating expansion of the Universe (Chapter 48) makes the de Sitter row observationally relevant. Torsion, by contrast, is retained in this chapter as mathematics: no experiment to date requires $T^a \neq 0$, and the physical parts of this treatise accordingly set it to zero — the scope rule of Chapter 1 applied to geometry itself.

Chapter 14

Lie Groups, Lie Algebras, and Fibre Bundles

A continuous symmetry of a physical system is described by a group whose elements depend smoothly on a set of real parameters: a Lie group. This chapter introduces Lie groups and their generators, builds the universal enveloping algebra and the Casimir operators that label an irreducible representation, and works out the rotation group $\mathrm{SO}(2,\mathbb{R})$ as the fundamental example. It then constructs the three compact groups on which the Standard Model rests — $\mathrm{SO}(3,\mathbb{R})$, its double cover $\mathrm{SU}(2)$, and $\mathrm{SU}(3)$ — together with their irreducible representations, and the Lie algebras $\mathfrak{t}(D)$ and $\mathfrak{so}(p,q)$ with the algebras built from them: Lorentz, Poincaré, (anti-)de Sitter and conformal, in general dimension $D = p + q$ and signature (p, q). The two Casimir invariants of the Poincaré algebra, which are mass and spin, are obtained in the observed $3 + 1$ dimensions. The closing sections develop invariant polynomials on a principal fibre bundle and prove the Chern–Weil theorem, from which the transgression and Chern–Simons forms follow.

Notation 14.1. Throughout this chapter $D = p + q$ denotes the dimension of the underlying pseudo-Euclidean space; capital Latin indices $A, B, C, \ldots$ run from 1 to D; and $\eta_{AB} = \mathrm{diag}(\underbrace{+1,\ldots,+1}_{p},\underbrace{-1,\ldots,-1}_{q})$ is the flat metric of signature (p, q), used to raise and lower indices, $x_A = \eta_{AB}\, x^B$. The upright symbol D denotes the covariant exterior derivative and must not be confused with the italic dimension D.

14.1 Lie groups

14.1.1 Definition and generators

Lie group

Definition 14.2 (Lie group)**.** Consider an n-dimensional differentiable manifold G. To denote the points of the manifold we introduce the differentiable operator g, with $g(\theta^a) \in G$, where the θ^a, $a = 1, \ldots, M$, are real parameters. We say that G is a *Lie group* if and only if

1. the algebraic structure $(G, \circ)$ is a group;
2. the operator g is of class C^∞ in the parameters θ^a.

Thus, in order to characterise a continuous-group transformation, the transformation must be a function of the old coordinates and of the parameters θ^a; that is, continuous-group transformations take the form

$$\bar{x}^i = f^i(\theta^a, x^i)\,. \tag{14.1}$$

Generator of a Lie group

Consider an infinitesimal continuous-group transformation in a Lie group G,

$$\bar{x}^i = f^i(\delta\theta^a, x^i)\,.$$

Since with no variation of the parameters the coordinates do not change, we may write

$$x^i = f^i(0, x^i)\,,$$

so the infinitesimal variation of the coordinates is given by

$$\begin{aligned}\delta\bar{x}^i &= \bar{x}^i - x^i \\ &= f^i(\delta\theta^a, x^i) - f^i(0, x^i) \\ &\approx f^i(0, x^i) + \left.\frac{\partial f^i}{\partial\theta^a}\right|_{\theta^a=0} \delta\theta^a - f^i(0, x^i) \\ &= \left.\frac{\partial f^i}{\partial\theta^a}\right|_{\theta^a=0} \delta\theta^a\,. \end{aligned} \tag{14.2}$$

Consider now the variation of a field $F = F(x)$ defined on G between two infinitesimally close points P and Q:

$$\begin{aligned}\delta F &= F(Q) - F(P) \\ &= F(x + \delta x) - F(x) \\ &\approx F(x) + \frac{\partial F}{\partial x^i}\,\delta x^i - F(x) \\ &= \frac{\partial F}{\partial x^i}\,\delta x^i \\ \text{[Equation (14.2)]} \quad &= \frac{\partial F}{\partial x^i}\,\left.\frac{\partial f^i}{\partial\theta^a}\right|_{\theta^a=0} \delta\theta^a \\ &= \left(\left.\frac{\partial f^i}{\partial\theta^a}\right|_{\theta^a=0} \partial_i\right) F\,\delta\theta^a\,. \end{aligned}$$

We define the *generators of the Lie group* as the operators

$$J_a = \left.\frac{\partial f^i}{\partial\theta^a}\right|_{\theta^a=0} \partial_i\,, \tag{14.3}$$

so that

$$\delta F = J_a F\,\delta\theta^a\,. \tag{14.4}$$

From Equations (14.2) and (14.3) (insert the infinitesimal coordinate variation into the variation of the field and read off the operator acting on it).

The generators of a Lie group owe their name to the fact that they generate the variation of a field, as Equation (14.4) shows. They are mathematical objects that can always be assigned to a Lie group, and they are of fundamental importance in physics for the study of the symmetries of a system, as the physical parts of this treatise will show at length.

Geometric interpretation of the Lie bracket

In the preceding setting, let us examine what happens when we apply the variations along different paths leading from P to Q. Let θ_1^a and θ_2^a be the two sets of parameters used to follow the two distinct paths. From Equation (14.4) we have

$$\delta_1 F = J_a F\,\delta\theta_1^a\,, \qquad \delta_2 F = J_b F\,\delta\theta_2^b\,.$$

Then

$$\delta_2\delta_1 F = J_b J_a F\, \delta\theta_1^a \delta\theta_2^b\,, \qquad \delta_1\delta_2 F = J_a J_b F\, \delta\theta_2^b \delta\theta_1^a\,,$$

so that the variation of the variations becomes

$$\begin{aligned}\delta_1\delta_2 F - \delta_2\delta_1 F &= J_a J_b F\, \delta\theta_2^b \delta\theta_1^a - J_b J_a F\, \delta\theta_1^a \delta\theta_2^b \\ &= [J_a, J_b]\, F\, \delta\theta_1^a \delta\theta_2^b\,.\end{aligned}$$

Hence, for the quantity $\delta_1\delta_2 F - \delta_2\delta_1 F$ to be, indeed, a variation of F leading from P to Q — that is, of the form Equation (14.4) — the operator $[J_a, J_b]$ must be a linear combination of the generators:

$$[J_a, J_b] = C^c{}_{ab}\, J_c\,. \tag{14.5}$$

From Equation (14.4) (demanding that the commutator of two variations be itself a variation of the same form).

Exponentiation of the generators

A finite transformation is obtained by exponentiating the generators:

$$g(\theta^a) = \mathrm{e}^{J_a \theta^a}\,. \tag{14.6}$$

The source asserts Equation (14.6) without derivation. Two statements are hidden in it and they are of different kinds. That a *one-parameter* family of transformations closing under composition is an exponential is a theorem, proved next. That a *many*-parameter family can be written as the single exponential Equation (14.6) is a statement about the choice of parameters, and is discussed in Remark 14.4.

Proposition 14.3 (A one-parameter subgroup is an exponential)**.** *Let $\mathbb{V}$ be a finite-dimensional real or complex vector space and let $t \mapsto g(t)$ be a differentiable curve of linear operators on $\mathbb{V}$ satisfying*

$$g(s+t) = g(s)\, g(t)\,, \qquad g(0) = \mathbb{1}\,. \tag{14.7}$$

Then, with $J = \left.\frac{\mathrm{d}g}{\mathrm{d}t}\right|_{t=0}$,

$$\boxed{g(t) = \mathrm{e}^{tJ} = \sum_{n=0}^{\infty} \frac{t^n}{n!}\, J^n}\,, \tag{14.8}$$

the series converging absolutely for every t. In particular $g(t)$ is invertible, with $g(t)^{-1} = g(-t)$.

Rests on Definition 14.2 and Equation (14.3).

Proof. *Derives Proposition 14.3.*

Fix any submultiplicative norm on the operators of $\mathbb{V}$, so that $\|J^n\| \le \|J\|^n$. Then $\sum_n |t|^n \|J\|^n / n! = \mathrm{e}^{|t|\|J\|} < \infty$, so the series Equation (14.8) converges absolutely in the finite-dimensional — hence complete — space of operators, term by term differentiation is legitimate, and $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{tJ} = J\mathrm{e}^{tJ} = \mathrm{e}^{tJ} J$.

Differentiate Equation (14.7) with respect to s at $s = 0$, holding t fixed. The left-hand side gives $\frac{\mathrm{d}g}{\mathrm{d}t}(t)$ and the right-hand side gives $g(t)\left.\frac{\mathrm{d}g}{\mathrm{d}s}\right|_{s=0} = g(t)J$, so

$$\frac{\mathrm{d}g}{\mathrm{d}t} = g(t)\, J\,, \qquad g(0) = \mathbb{1}\,. \tag{14.9}$$

Put $h(t) = g(t)\mathrm{e}^{-tJ}$. Since J commutes with e^{-tJ},

$$\frac{\mathrm{d}h}{\mathrm{d}t} = g(t)J\mathrm{e}^{-tJ} - g(t)J\mathrm{e}^{-tJ} = 0\,,$$

so h is constant and equal to $h(0) = \mathbb{1}$. Multiplying on the right by e^{tJ} gives Equation (14.8). Finally $g(t)g(-t) = g(0) = \mathbb{1}$ by Equation (14.7). □

Remark 14.4 (What the exponential formula asserts, and what it does not). For several parameters, $\mathrm{e}^{\theta^a J_a}$ is a well-defined operator for every choice of the θ^a, and for fixed θ^a the curve $t \mapsto \mathrm{e}^{t\theta^a J_a}$ is a one-parameter subgroup by Proposition 14.3. The map $\theta^a \mapsto \mathrm{e}^{\theta^a J_a}$ is smooth, and its differential at $\theta^a = 0$ sends θ^a to $\theta^a J_a$, which is injective because the generators are linearly independent; by the inverse function theorem the map is therefore a diffeomorphism of a neighbourhood of the origin in parameter space onto a neighbourhood of the identity in G. Equation (14.6) is the statement that the parameters θ^a of Definition 14.2 have been chosen to be those coordinates — the *canonical coordinates of the first kind*. Three things it does *not* say. It does not say that parameters compose additively: unless the generators commute, $\mathrm{e}^{X}\mathrm{e}^{Y} \neq \mathrm{e}^{X+Y}$, the leading correction to the exponent being $\tfrac{1}{2}[X,Y]$. It does not say that every element of a connected Lie group is a single exponential; what is true, and is all that Corollary 14.18 uses, is that the exponentials cover a neighbourhood of the identity and that a connected group is generated by any such neighbourhood. And when the generators are realized as the differential operators Equation (14.3) rather than as matrices, $\mathrm{e}^{\theta^a J_a}$ is to be read as the flow of the vector field $\theta^a J_a$, which exists locally by the existence and uniqueness theorem for ordinary differential equations and obeys Equation (14.7) for that reason.

Alternative expression for the generators

From Equation (14.6) we see that, differentiating with respect to θ^a,

$$\begin{aligned}\frac{\mathrm{d}g}{\mathrm{d}\theta^a} &= \mathrm{e}^{J_b\theta^b}\,\frac{\mathrm{d}}{\mathrm{d}\theta^a}\left(J_c\theta^c\right)\\ &= \mathrm{e}^{J_b\theta^b} J_c\,\delta^c{}_a\\ &= \mathrm{e}^{J_b\theta^b} J_a\,,\end{aligned}$$

and evaluating at $\theta^a = 0$ we obtain

$$J_a = \left.\frac{\mathrm{d}g}{\mathrm{d}\theta^a}\right|_{\theta^a=0}. \tag{14.10}$$

From Equation (14.6) (differentiate the exponential with respect to a parameter and evaluate at the identity).

The universal enveloping algebra

The bracket Equation (14.5) is not an associative product. The juxtaposition J_aJ_b has no meaning in the Lie algebra itself; it acquires one only inside a representation, where the generators are operators and juxtaposition is composition. Yet the quantities that label the states of a physical system — the squared angular momentum, the squared four-momentum — are polynomials of exactly that kind. Before they can be discussed once and for all, rather than representation by representation, they must be given a home.

Definition 14.5 (Universal enveloping algebra)**.** Let $\mathfrak{g}$ be a Lie algebra over $\mathbb{R}$ with basis $\{T_a\}$, $a = 1, \dots, \dim\mathfrak{g}$, and structure constants fixed as in Equation (14.5),

$$[T_a, T_b] = C^c{}_{ab}\,T_c\,. \tag{14.11}$$

Its *universal enveloping algebra* $U(\mathfrak{g})$ is the associative algebra with unit generated by the symbols T_a subject to the relations

$$\boxed{T_aT_b - T_bT_a = C^c{}_{ab}\,T_c}\,, \tag{14.12}$$

equivalently the quotient of the tensor algebra of the vector space $\mathfrak{g}$ by the two-sided ideal generated by all elements $X \otimes Y - Y \otimes X - [X,Y]$ with $X, Y \in \mathfrak{g}$.

Rests on Equation (14.5).

Remark 14.6 (What $U(\mathfrak{g})$ is for). Every linear representation D of $\mathfrak{g}$ on a vector space $\mathbb{V}$ — that is, every linear map with $D([X,Y]) = D(X)D(Y) - D(Y)D(X)$ — extends uniquely to a homomorphism of associative algebras $U(\mathfrak{g}) \to \mathrm{End}(\mathbb{V})$, because Equation (14.12) is precisely the relation the operators $D(T_a)$ already satisfy. This is the sense in which $U(\mathfrak{g})$ is *universal*, and it is what makes a statement proved once in $U(\mathfrak{g})$ a statement about every representation at once.

Theorem 14.7 (Poincaré–Birkhoff–Witt)**.** *The ordered monomials $T_1^{n_1} T_2^{n_2} \cdots T_M^{n_M}$, with $M = \dim \mathfrak{g}$ and $n_a \in \mathbb{N} \cup \{0\}$, form a basis of $U(\mathfrak{g})$ as a vector space. In particular $\mathfrak{g}$ embeds in $U(\mathfrak{g})$, and no relation holds among the generators beyond those forced by Equation (14.12).*

Rests on Definition 14.5.

That the ordered monomials *span* $U(\mathfrak{g})$ is immediate from the defining relations, which allow any product to be reordered at the cost of terms of lower degree; the content of the theorem is their linear *independence*, and it is what guarantees that the Casimir element built below is not accidentally the zero element of the enveloping algebra.

Full derivation: Appendix A.30 (p. 2928).

Derives Theorem 14.7.

Casimir operators

Definition 14.8 (Casimir element)**.** A *Casimir element*, or *Casimir operator*, of a Lie algebra $\mathfrak{g}$ is an element $C \in U(\mathfrak{g})$ commuting with every generator,

$$\boxed{[C, T_a] = 0\,, \qquad a = 1, \ldots, \dim \mathfrak{g}}\,, \tag{14.13}$$

equivalently an element of the centre $Z\left(U(\mathfrak{g})\right)$.

Rests on Definition 14.5.

A Casimir element is in general *not* an element of $\mathfrak{g}$: for a semisimple algebra the centre of $\mathfrak{g}$ itself is zero, and the invariants appear only at quadratic and higher degree in the generators. This is the whole reason for Definition 14.5.

Notation 14.9 (Casimirs against structure constants). Casimir elements are written $C_2, C_3, \ldots$, the single numerical subscript recording the degree of the polynomial in the generators. The structure constants always carry three algebra indices, ${C^c}_{ab}$, so no ambiguity arises. Note also that Section 14.7 writes the same structure constants with the index pair first, $[\boldsymbol{T}_A, \boldsymbol{T}_B] = {C_{AB}}^C \boldsymbol{T}_C$; only the index placement differs, and ${C_{ab}}^c = {C^c}_{ab}$.

The Killing form and the quadratic Casimir

Every Lie algebra carries a canonical symmetric bilinear form, built from the structure constants alone, and it is this form that produces the Casimir element every physicist meets first.

Definition 14.10 (Killing form)**.** The *Killing form* of $\mathfrak{g}$ is

$$\boxed{\kappa(X,Y) = \mathrm{tr}\left(\mathrm{ad}_X\, \mathrm{ad}_Y\right)}\,, \qquad \mathrm{ad}_X\, Y = [X,Y]\,, \tag{14.14}$$

a symmetric bilinear form on $\mathfrak{g}$. Since $(\mathrm{ad}_{T_a})^c{}_b = {C^c}_{ab}$ is the matrix of ad_{T_a} in the basis $\{T_a\}$, its components are

$$\kappa_{ab} = \kappa(T_a, T_b) = {C^c}_{ad}\, {C^d}_{bc}\,. \tag{14.15}$$

Rests on Equation (14.5).

Lemma 14.11 (Invariance of the Killing form)**.** *For all $X, Y, Z \in \mathfrak{g}$,*

$$\kappa\left([X,Y],Z\right)+\kappa\left(Y,[X,Z]\right)=0\,. \tag{14.16}$$

Rests on Definition 14.10.

Proof. *Derives Lemma 14.11.*

The adjoint map is itself a representation, $\mathrm{ad}_{[X,Y]} = [\mathrm{ad}_X, \mathrm{ad}_Y]$, which is the Jacobi identity rewritten. Hence

$$\begin{aligned}\kappa\left([X,Y],Z\right)&=\mathrm{tr}\left([\mathrm{ad}_X,\mathrm{ad}_Y]\,\mathrm{ad}_Z\right)=\mathrm{tr}\left(\mathrm{ad}_X\,\mathrm{ad}_Y\,\mathrm{ad}_Z\right)-\mathrm{tr}\left(\mathrm{ad}_Y\,\mathrm{ad}_X\,\mathrm{ad}_Z\right)\,,\\ \kappa\left(Y,[X,Z]\right)&=\mathrm{tr}\left(\mathrm{ad}_Y\,[\mathrm{ad}_X,\mathrm{ad}_Z]\right)=\mathrm{tr}\left(\mathrm{ad}_Y\,\mathrm{ad}_X\,\mathrm{ad}_Z\right)-\mathrm{tr}\left(\mathrm{ad}_Y\,\mathrm{ad}_Z\,\mathrm{ad}_X\right)\,.\end{aligned}$$

Adding the two lines, the terms $\mp\,\mathrm{tr}(\mathrm{ad}_Y\,\mathrm{ad}_X\,\mathrm{ad}_Z)$ cancel against each other, and the survivors $\mathrm{tr}(\mathrm{ad}_X\,\mathrm{ad}_Y\,\mathrm{ad}_Z)$ and $-\,\mathrm{tr}(\mathrm{ad}_Y\,\mathrm{ad}_Z\,\mathrm{ad}_X)$ cancel by cyclicity of the trace. □

Corollary 14.12 (Total antisymmetry of the structure constants)**.** *Lower the upper index of the structure constants with the Killing form,*

$$C_{abc}:=C^{d}{}_{ab}\,\kappa_{dc}=\kappa\left([T_a,T_b],T_c\right)\,. \tag{14.17}$$

Then C_{abc} is antisymmetric under the exchange of any *two of its three indices.*

Rests on Lemma 14.11 and Definition 14.10.

Proof. *Derives Corollary 14.12.*

Antisymmetry in the first two indices is the antisymmetry of the bracket. For the last two, put $X = T_a$, $Y = T_b$, $Z = T_c$ in Equation (14.16) and use the symmetry of κ:

$$C_{abc}=\kappa\left([T_a,T_b],T_c\right)=-\kappa\left(T_b,[T_a,T_c]\right)=-\kappa\left([T_a,T_c],T_b\right)=-C_{acb}\,.$$

The transpositions (ab) and (bc) generate the whole symmetric group on three letters, so C_{abc} changes sign under every transposition. □

Theorem 14.13 (Cartan's criterion)**.** *A finite-dimensional Lie algebra over a field of characteristic zero is semisimple — it contains no nonzero abelian ideal — if and only if its Killing form is nondegenerate.*

Rests on Definition 14.10.

One direction — that a nonzero abelian ideal lies in the radical of the Killing form — is a short index computation; the converse, that a degenerate Killing form produces a solvable ideal, goes through Cartan's criterion for solvability and the Jordan decomposition of an endomorphism, and is what makes the theorem long. It is used here to know when the Killing form may be inverted, and again in Whitehead's lemmas in the section on central extensions.

Full derivation: Appendix A.31 (p. 2934).

Derives Theorem 14.13.

When κ_{ab} is nondegenerate we write κ^{ab} for its inverse, $\kappa^{ab}\kappa_{bc} = \delta^{a}{}_{c}$, and use it freely to raise algebra indices.

Definition 14.14 (Invariant symmetric tensor)**.** A totally symmetric array $k^{a_1\cdots a_r}$ is *invariant* if, for every c,

$$\boxed{\sum_{l=1}^{r} C^{a_l}{}_{bc}\,k^{a_1\cdots a_{l-1}\,b\,a_{l+1}\cdots a_r}=0}\,. \tag{14.18}$$

Rests on Equation (14.5).

Remark 14.15 (This is the invariance already proved). Condition Equation (14.18) is the infinitesimal invariance of Definition 14.109 with the indices raised: it is Equation (14.148), established as part 2 of Theorem 14.110, transported to the contravariant position by κ^{ab}. On a semisimple algebra, where the Killing form may be inverted, the two notions therefore coincide, and every G-invariant polynomial of Section 14.7 supplies one of these tensors. Nothing new has to be proved to use them.

Theorem 14.16 (Invariant tensors give Casimir operators). *Let $k^{a_1\cdots a_r}$ be a totally symmetric invariant tensor in the sense of Definition 14.14. Then*

$$\boxed{C_r = k^{a_1\cdots a_r}\, T_{a_1} T_{a_2} \cdots T_{a_r} \in U(\mathfrak{g})} \tag{14.19}$$

is a Casimir element.

Rests on Definitions 14.5, 14.8 and 14.14.

Proof. *Derives Theorem 14.16.*

In the associative algebra $U(\mathfrak{g})$ the commutator with a fixed element obeys the Leibniz rule, $[XY, Z] = X\,[Y, Z] + [X, Z]\,Y$, so

$$\begin{aligned}[C_r, T_c] &= k^{a_1\cdots a_r} \sum_{l=1}^{r} T_{a_1} \cdots T_{a_{l-1}}\, [T_{a_l}, T_c]\, T_{a_{l+1}} \cdots T_{a_r} \\ &= \sum_{l=1}^{r} k^{a_1\cdots a_r}\, C^{b}{}_{a_l c}\, T_{a_1} \cdots T_{a_{l-1}}\, T_b\, T_{a_{l+1}} \cdots T_{a_r}\,.\end{aligned}$$

In the l-th term rename the two summation indices a_l and b into each other. This is a relabelling of dummy indices and moves no operator past another, so the ordered string becomes literally $T_{a_1} T_{a_2} \cdots T_{a_r}$, the *same* string for every value of l, while the coefficient becomes $C^{a_l}{}_{bc}\, k^{a_1\cdots a_{l-1}\, b\, a_{l+1}\cdots a_r}$. The sum over l may therefore be taken inside:

$$[C_r, T_c] = \left(\sum_{l=1}^{r} C^{a_l}{}_{bc}\, k^{a_1\cdots a_{l-1}\, b\, a_{l+1}\cdots a_r} \right) T_{a_1} T_{a_2} \cdots T_{a_r}\,,$$

and the parenthesis vanishes by Equation (14.18). □

Corollary 14.17 (The quadratic Casimir). *Let $\mathfrak{g}$ be semisimple, so that κ_{ab} is invertible by Theorem 14.13. Then*

$$\boxed{C_2 = \kappa^{ab}\, T_a T_b} \tag{14.20}$$

commutes with every generator of $\mathfrak{g}$.

Rests on Theorems 14.13 and 14.16 and Corollary 14.12.

Proof. *Derives Corollary 14.17.*

The statement is Theorem 14.16 for $r = 2$ and $k^{ab} = \kappa^{ab}$, whose invariance is Corollary 14.12 with both free indices raised. It is worth seeing the two-index case done by hand, because it is the whole mechanism in three lines. By the Leibniz rule,

$$[C_2, T_c] = \kappa^{ab}\,(T_a\,[T_b, T_c] + [T_a, T_c]\,T_b) = \kappa^{ab} C^{d}{}_{bc}\, T_a T_d + \kappa^{ab} C^{d}{}_{ac}\, T_d T_b\,.$$

Fix c and set

$$M^{ad} := \kappa^{ab}\, C^{d}{}_{bc} = \kappa^{ab}\kappa^{de}\, C_{bce}\,,$$

using Equation (14.17). Exchanging the two free indices and then renaming the dummies $b \leftrightarrow e$,

$$M^{da} = \kappa^{db}\kappa^{ae}\, C_{bce} = \kappa^{de}\kappa^{ab}\, C_{ecb} = -\kappa^{ab}\kappa^{de}\, C_{bce} = -M^{ad}\,,$$

the last step by Corollary 14.12, since C_{ecb} differs from C_{bce} by the transposition of the first and third indices. The first term above is $M^{ad}T_aT_d$; the second is $\kappa^{ab}C^{d}{}_{ac}T_dT_b = M^{bd}T_dT_b$ by the symmetry of κ, which is $M^{ad}T_dT_a$ after renaming $b \to a$. Hence

$$[C_2, T_c] = M^{ad}\,(T_aT_d + T_dT_a) = 0\,,$$

an antisymmetric array contracted with a symmetric one. □

Corollary 14.18 (A Casimir acts as a number on an irreducible representation)**.** *Let G be a connected Lie group with Lie algebra $\mathfrak{g}$, let U be a finite-dimensional irreducible representation of G on a complex vector space $\mathbb{V}$, let D be the induced representation of $\mathfrak{g}$ as in Equation* (14.6)*, and let C be a Casimir element. Then*

$$\boxed{D(C) = c\,\mathbb{1}\,, \qquad c \in \mathbb{C}}\,. \tag{14.21}$$

Rests on Definition 14.8, Remark 14.6, Equation (14.6)*, and Theorem 5.153.*

Proof. *Derives Corollary 14.18.*

By Remark 14.6 the representation D extends to $U(\mathfrak{g})$, so $D(C)$ is a well-defined operator on $\mathbb{V}$ and $[D(C), D(T_a)] = D\,([C, T_a]) = 0$ for every a by Equation (14.13). An operator commuting with $D(T_a)$ commutes with every power of $D(T_a)$, hence with the exponential $\mathrm{e}^{\theta^a D(T_a)} = U\,(g(\theta^a))$ of Equation (14.6), that is, with $U(g)$ for every g in a neighbourhood of the identity. A connected topological group is generated by any neighbourhood of its identity, so $D(C)$ commutes with $U(g)$ for every $g \in G$. Schur's first lemma (Theorem 5.153), applied to the irreducible representation U, gives $D(C) = c\,\mathbb{1}$. □

Remark 14.19 (A Casimir eigenvalue labels; one Casimir may not separate). Equation (14.21) is what makes a Casimir useful. The number c takes the same value on every vector of the representation and is unchanged by any change of basis, so it is an invariant *label* of the equivalence class of the irreducible representation, and two representations with different values of c cannot be equivalent. The converse does not follow, and it is worth being precise about it rather than repeating the usual slogan. A second imprecision has to be removed first: there is no such thing as *the* quadratic Casimir, because any nonzero multiple of a Casimir is again one and its eigenvalue scales with it, so a numerical eigenvalue means something only once the normalisation is stated. For $\mathfrak{su}(2)$, in the normalisation $\boldsymbol{J}^2 = J_1^2 + J_2^2 + J_3^2$ of Equation (14.49) — which by Lemma 14.41 is twice the Killing-form Casimir $C_2 = \kappa^{ab}J_aJ_b$ of Equation (14.20) — the quadratic Casimir does separate the classes, because Theorem 14.43 gives its eigenvalue as $j(j+1)$ and $j \mapsto j(j+1)$ is injective for $j \geq 0$. For $\mathfrak{su}(3)$ it does not: the fundamental and the antifundamental representations of Proposition 14.55 share the value of C_2 and are told apart only by a second, cubic invariant. How many invariants are needed in general is the subject of what follows.

Rank, and how many Casimir operators there are

Definition 14.20 (Cartan subalgebra and rank)**.** A *Cartan subalgebra* $\mathfrak{h} \subset \mathfrak{g}$ of a semisimple Lie algebra is a maximal abelian subalgebra whose elements all act diagonalizably in the adjoint representation. Its dimension is the same for every such subalgebra and is called the *rank* ℓ of $\mathfrak{g}$.

Theorem 14.21 (Racah)**.** *For a finite-dimensional semisimple Lie algebra of rank ℓ, the centre $Z\,(U(\mathfrak{g}))$ is a polynomial algebra in ℓ algebraically independent Casimir elements. A complete set of invariant labels for its finite-dimensional irreducible representations therefore consists of ℓ numbers, and of no fewer.*

Rests on Definitions 14.8 and 14.20.

The derivation identifies the centre of the enveloping algebra with the adjoint-invariant polynomials on the algebra, by the symmetrization map, and then those with the Weyl-invariant polynomials on a Cartan subalgebra; it reduces the theorem to two results of the structure theory of semisimple Lie algebras — the Chevalley restriction theorem and Chevalley's theorem on finite reflection groups — which this treatise quotes rather than proves, and the appendix says exactly where. The systematic use of such invariants as representation labels in spectroscopy is Racah's.

Full derivation: Appendix A.32 (p. 2941).

Derives Theorem 14.21.

The rank is read off the algebra without difficulty in the cases this treatise needs. For $\mathfrak{so}(3) \cong \mathfrak{su}(2)$ any single generator spans a maximal abelian subalgebra, so $\ell = 1$ and there is exactly one Casimir: it is the $\boldsymbol{J}^2$ constructed in Definition 14.40. For $\mathfrak{su}(3)$ the diagonal traceless matrices give $\ell = 2$, hence two Casimirs, one quadratic and one cubic (Proposition 14.54). The remaining family this treatise needs is $\mathfrak{so}(p,q)$, and because the count of Lorentz invariants rests on it the statement is proved rather than quoted.

Proposition 14.22 (The rank of $\mathfrak{so}(p,q)$)**.** *Let $D = p + q \geq 3$. Whatever the signature,*

$$\boxed{\ell\left(\mathfrak{so}(p,q)\right) = \left\lfloor \frac{D}{2} \right\rfloor}. \tag{14.22}$$

In particular the Lorentz algebra $\mathfrak{so}(3,1)$ of the observed spacetime has rank 2, and therefore, by Theorem 14.21, exactly two Casimir operators.

Rests on Definition 14.20, Equation (14.90), and Theorem 14.21.

Proof. *Derives Proposition 14.22.*

Write $r = \lfloor D/2 \rfloor$, work in the defining representation on $\mathbb{C}^D$, and let η be diagonal as in Notation 14.1, so that $X \in \mathfrak{so}(p,q)$ means

$$\eta(Xu, v) + \eta(u, Xv) = 0 \qquad \text{for all } u, v\,. \tag{14.23}$$

Diagonalizability in the defining representation is what is used throughout; that it agrees with the diagonalizability in the adjoint representation asked for by Definition 14.20 is standard structure theory, quoted here as Theorem 14.13 is. One direction is immediate — in a basis diagonalizing X the map $Y \mapsto [X, Y]$ is diagonal on the matrix units, with eigenvalue the difference of two eigenvalues of X — and the other is that the semisimple and nilpotent parts of an element of a semisimple linear Lie algebra again lie in it, so that an element with diagonalizable adjoint action has central, hence zero, nilpotent part.

A commuting set of size r. Take

$$H_k = J_{(2k-1)(2k)}\,, \qquad k = 1, \ldots, r\,, \tag{14.24}$$

one generator for each of the r disjoint index pairs $\{1,2\}$, $\{3,4\}$, and so on. No two of them share an index, so every η on the right of Equation (14.90) carries two distinct indices and vanishes, η being diagonal; hence $[H_k, H_l] = 0$. In the defining representation J_{AB} is the matrix $(J_{AB})^C{}_D = \delta^C_A \eta_{BD} - \delta^C_B \eta_{AD}$, the identification already made for $\mathfrak{so}(3)$ at Equation (14.32): it acts inside the plane of its own two coordinates and annihilates the others, and on that plane it is either the rotation generator Equation (14.27), with eigenvalues $\pm\mathrm{i}$, or a boost generator, with eigenvalues ± 1; in both cases it is diagonalizable, and so is any linear combination of the H_k, the blocks being independent. The span of the H_k is therefore an abelian subalgebra of diagonalizable elements, of dimension r.

No larger one exists. Let $\mathfrak{h}$ be any abelian subalgebra of $\mathfrak{so}(p,q)$ all of whose elements are diagonalizable, and let $\mathfrak{h}_{\mathbb{C}}$ be its complex span inside the complex $D \times D$ matrices. Since the elements of $\mathfrak{h}$ are real matrices, a real basis of $\mathfrak{h}$ stays independent over $\mathbb{C}$ —

the real and imaginary parts of a vanishing complex combination vanish separately — so $\dim_{\mathbb{C}} \mathfrak{h}_{\mathbb{C}} = \dim_{\mathbb{R}} \mathfrak{h}$. Commuting diagonalizable operators are simultaneously diagonalizable, and every complex combination of operators diagonal in one basis is diagonal in that basis, so there are a basis $\{v_1, \dots, v_D\}$ of $\mathbb{C}^D$ and complex-linear functionals λ_i on $\mathfrak{h}_{\mathbb{C}}$ with $X v_i = \lambda_i(X) v_i$. Condition Equation (14.23) is complex-bilinear and therefore holds on $\mathfrak{h}_{\mathbb{C}}$; applied to $u = v_i$, $v = v_j$ it gives

$$(\lambda_i(X) + \lambda_j(X))\, \eta(v_i, v_j) = 0 \qquad \text{for all } X \in \mathfrak{h}_{\mathbb{C}}\,.$$

Collect the basis vectors by their functional: for each α occurring among the λ_i, let V_α be the span of those v_i with $\lambda_i = \alpha$. The display says $\eta(V_\alpha, V_\beta) = 0$ unless $\beta = -\alpha$, so nondegeneracy of η forces $-\alpha$ to occur whenever α does, and forces the pairing of V_α with $V_{-\alpha}$ to be nondegenerate, whence $\dim V_\alpha = \dim V_{-\alpha}$. The occurring functionals thus consist of k pairs $\pm\alpha_1, \dots, \pm\alpha_k$ with $\alpha_i \neq 0$, possibly together with 0; each pair accounts for at least two basis vectors, so $2k \leq D$ and $k \leq r$. Finally the occurring functionals span the dual of $\mathfrak{h}_{\mathbb{C}}$, since an X annihilated by all of them acts as zero on every v_i and is therefore zero, and they span the same space as $\alpha_1, \dots, \alpha_k$. Hence $\dim_{\mathbb{R}} \mathfrak{h} = \dim_{\mathbb{C}} \mathfrak{h}_{\mathbb{C}} \leq k \leq r$.

The subalgebra Equation (14.24) is therefore maximal, and every maximal one has the same dimension r. □

Remark 14.23 (Racah's theorem does not apply to every algebra that matters). Theorem 14.21 is a statement about *semisimple* algebras, and the Poincaré algebra Equation (14.99) is not semisimple: it is a semidirect sum with an abelian ideal, so its Killing form is degenerate and Theorem 14.13 excludes it. Its invariants must be found by hand, which is done in Theorem 14.65, and the count that comes out — two, in the observed $3+1$ dimensions — is not predicted by Racah's theorem but merely happens to agree with the rank of its Lorentz subalgebra. Conflating the two is an easy error and would give the wrong answer for the Galilei algebra of Example 14.80, whose invariants include the central mass.

Remark 14.24 (Casimir operators and SI units). Everything in this subsection is dimensionless. The generators of Equation (14.3) are differential operators built from coordinates and derivatives with respect to them, the structure constants are pure numbers, and a Casimir element is a polynomial in the generators, so its eigenvalue on an irreducible representation is a pure number too. A physical observable is obtained by multiplying a generator by the constant that carries its dimension, and for the rotation generators that constant is the reduced Planck constant $\hbar = h/2\pi$, with $h = 6.626\,070\,15 \times 10^{-34}\,\mathrm{J\,s}$ exactly, by the definition of the kilogram [2333]. The observable angular momentum is $\hbar \boldsymbol{J}$, of SI unit J s, and the observable Casimir is $\hbar^2 \boldsymbol{J}^2$, of SI unit $\mathrm{J^2\,s^2}$. It is worth saying this once and plainly, because the literature almost never does: the familiar statement "the eigenvalue is $j(j+1)$" is a statement about a dimensionless generator, and the measured quantity is $\hbar^2 j(j+1)$. The same distinction returns for the Poincaré algebra in Proposition 14.60, where the first invariant is m^2c^2 and not m^2.

14.2 Representations of Lie groups

14.2.1 The group $\mathrm{SO}(2, \mathbb{R})$

Discussion of the definition

Consider the vector space $\mathbb{R}^2 = \mathrm{span}\,\{|e_x\rangle, |e_y\rangle\}$ and let $R(\phi)$ be the operator that takes the basis vector $|e_i\rangle$ to the rotated vector $R(\phi)\,|e_i\rangle$ [2027]. Elementary trigonometry on the rotated basis gives

$$\begin{aligned} R(\phi)\,|e_x\rangle &= \cos\phi\;|e_x\rangle + \sin\phi\;|e_y\rangle\;, \\ R(\phi)\,|e_y\rangle &= -\sin\phi\;|e_x\rangle + \cos\phi\;|e_y\rangle\;, \end{aligned}$$

or, compactly,

$$R(\phi)\,|e_i\rangle = (D(\phi))^j{}_i\,|e_j\rangle\ ,$$

with

$$D(\phi) = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}\ . \tag{14.25}$$

Now consider an arbitrary vector $|x\rangle \in \mathbb{R}^2$. In the basis above, $|x\rangle = x^i\,|e_i\rangle$. As we have seen, the operator $R(\phi)$ rotates a vector when acting upon it:

$$\begin{aligned} |\bar{x}\rangle &= R(\phi)\,|x\rangle\ , \\ \bar{x}^j\,|e_j\rangle &= R(\phi)\,x^i\,|e_i\rangle = x^i\,(D(\phi))^j{}_i\,|e_j\rangle\ , \end{aligned}$$

so that

$$\bar{x}^j = (D(\phi))^j{}_i\,x^i\,.$$

But does the rotation of the vector preserve its norm? This seems obvious; let us see what happens if we impose it. We require

$$\begin{aligned} \||x\rangle\| &\stackrel{!}{=} \||\bar{x}\rangle\|\ , \\ x^i x^i &= \bar{x}^j \bar{x}^j \\ &= (D(\phi))^j{}_k\,(D(\phi))^j{}_l\,x^k x^l \\ &= \left(D(\phi)^{\mathsf{T}}\right)^k{}_j\,(D(\phi))^j{}_l\,x^k x^l\,, \end{aligned}$$

whence

$$\left(D(\phi)^{\mathsf{T}}\right)^k{}_j\,(D(\phi))^j{}_l = \delta^k{}_l\,,$$

or, in matrix notation,

$$D(\phi)^{\mathsf{T}} D(\phi) = \mathbb{1}\,. \tag{14.26}$$

In conclusion, we have found the condition for the operator $R(\phi)$ to be an automorphism belonging to the group $\mathrm{O}(2,\mathbb{R})$. Moreover,

$$\det D(\phi) = \cos^2\phi + \sin^2\phi = 1\,,$$

and therefore $R(\phi) \in \mathrm{SO}(2,\mathbb{R})$.

With this we can conclude the following fact: the set of all two-dimensional rotations forms a group, which is isomorphic to the group $\mathrm{SO}(2,\mathbb{R})$. Furthermore, one can see geometrically that rotating a vector by an angle ϕ_1 and then by an angle ϕ_2 is exactly the same as doing it the other way around — first by ϕ_2 and then by ϕ_1 — which means that $\mathrm{SO}(2,\mathbb{R})$ is an abelian group.

Generator

We have

$$\begin{aligned} D(\phi) &= \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \\ &= \begin{pmatrix} \cos\phi & 0 \\ 0 & \cos\phi \end{pmatrix} + \begin{pmatrix} 0 & -\sin\phi \\ \sin\phi & 0 \end{pmatrix} \\ &= \cos\phi \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \sin\phi \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\ . \end{aligned}$$

Defining the matrix

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} , \tag{14.27}$$

we may write

$$D(\phi) = \mathbb{1} \cos\phi + J \sin\phi \, .$$

Note that

$$J^2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -\mathbb{1} \, ,$$

and thus

$$J^3 = -J \, , \qquad J^4 = \mathbb{1} \, , \qquad J^5 = J \, ,$$

so that, for $n \in \mathbb{N}$,

$$J^{2n} = (-1)^n \, \mathbb{1} \, , \qquad J^{2n+1} = (-1)^n \, J \, .$$

On the other hand, recalling the Taylor series of the cosine and the sine,

$$\cos\phi = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} \, \phi^{2n} \, , \qquad \sin\phi = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} \, \phi^{2n+1} \, ,$$

we obtain

$$\begin{aligned} D(\phi) &= \sum_{n=0}^{\infty} \frac{(-1)^n \, \mathbb{1}}{(2n)!} \, \phi^{2n} + \sum_{n=0}^{\infty} \frac{(-1)^n \, J}{(2n+1)!} \, \phi^{2n+1} \\ &= \sum_{n=0}^{\infty} \frac{J^{2n}}{(2n)!} \, \phi^{2n} + \sum_{n=0}^{\infty} \frac{J^{2n+1}}{(2n+1)!} \, \phi^{2n+1} \\ &= \sum_{n=0}^{\infty} \frac{J^n}{n!} \, \phi^n \, , \end{aligned}$$

whence

$$D(\phi) = \mathrm{e}^{J\phi} \, , \tag{14.28}$$

so J is the generator of the group $\mathrm{SO}(2, \mathbb{R})$.

Irreducible representations

We know that the group $\mathrm{SO}(2, \mathbb{R})$ is abelian, and we have seen that the irreducible representations of an abelian group are necessarily one-dimensional — Proposition 5.156, a consequence of the first Schur lemma (Theorem 5.153). Which, then, are the possible irreducible representations? Let $|\lambda\rangle$ be an eigenvector of J,

$$J \, |\lambda\rangle = \lambda \, |\lambda\rangle \ ,$$

so that, by Equation (14.28),

$$\begin{aligned} R(\phi) \, |\lambda\rangle &= \mathrm{e}^{J\phi} \, |\lambda\rangle \\ &= \mathrm{e}^{\lambda\phi} \, |\lambda\rangle \ . \end{aligned}$$

But since one must have

$$R(\phi) \, |x\rangle = R(\phi + 2\pi) \, |x\rangle \ ,$$

it follows that

$$\begin{aligned} \mathrm{e}^{\lambda\phi} \, |\lambda\rangle &= \mathrm{e}^{\lambda(\phi + 2\pi)} \, |\lambda\rangle \\ &= \mathrm{e}^{\lambda\phi} \, \mathrm{e}^{2\pi\lambda} \, |\lambda\rangle \implies \mathrm{e}^{2\pi\lambda} = 1 \, , \end{aligned}$$

so that

$$\lambda = \mathrm{i}m\,, \qquad m \in \mathbb{Z}\,.$$

Therefore, the irreducible representations of $\mathrm{SO}(2,\mathbb{R})$ are

$$R^m(\phi) = \mathrm{e}^{\mathrm{i}m\phi}\,, \qquad m \in \mathbb{Z}\,. \tag{14.29}$$

From Equation (14.28) (diagonalize the generator and impose 2π-periodicity of the rotation).

14.2.2 The groups $\mathrm{SO}(3)$, $\mathrm{SU}(2)$ and $\mathrm{SU}(3)$

Three compact groups carry almost all of the group theory the physical parts of this treatise need. The rotation group $\mathrm{SO}(3,\mathbb{R})$ is the symmetry of space itself. The special unitary group $\mathrm{SU}(2)$ is its double cover, and the existence of that cover — not any postulate about particles — is why half-integer angular momentum exists. The group $\mathrm{SU}(3)$ is the first case here in which the rank exceeds one, so that two Casimir operators are required by Theorem 14.21 and the weight diagram becomes two-dimensional. The gauge group of the Standard Model is $\mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1)$, and this subsection supplies the mathematics that Chapters 102 and 104 will use; the physics, and every claim about what the groups act on, is left to them. The quantum-mechanical treatment of angular momentum, which starts from the algebra derived here, is Chapter 79 [1593, 2287].

Remark 14.25 ($\mathrm{SU}(3)$ is an internal symmetry). This treatise instantiates every spacetime construction in the observed $3+1$ dimensions. The 3 of $\mathrm{SU}(3)$ is not a spacetime dimension: it is the dimension of an internal complex vector space attached to each point, on which the group acts without touching the coordinates at all. Nothing in this subsection is a statement about a three-dimensional spacetime.

The rotation group $\mathrm{SO}(3,\mathbb{R})$

Definition 14.26 (The rotation group)**.**

$$\mathrm{SO}(3,\mathbb{R}) = \left\{ R \in \mathrm{GL}(3,\mathbb{R}) \mid R^{\mathsf{T}}R = \mathbb{1},\ \det R = 1 \right\}. \tag{14.30}$$

This is the case $p=3$, $q=0$ of Section 14.3.2: the defining condition $R^{\mathsf{T}}\eta R = \eta$ of the pseudo-orthogonal group reduces to $R^{\mathsf{T}}R = \mathbb{1}$ when $\eta = \mathbb{1}$, and $\det R = 1$ selects the component containing the identity. Accordingly Equation (14.89) gives

$$\dim \mathfrak{so}(3) = \frac{3\cdot 2}{2} = 3\,. \tag{14.31}$$

From Equation (14.89) (the case $p=3$, $q=0$, so that $D=3$).

Proposition 14.27 (Generators of $\mathfrak{so}(3)$)**.** *The Lie algebra $\mathfrak{so}(3)$ is the space of real antisymmetric 3×3 matrices. A basis of it is*

$$(L_k)_{ij} = -\epsilon_{kij}\,, \tag{14.32}$$

that is

$$L_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad L_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{14.33}$$

and it satisfies

$$[L_i, L_j] = \epsilon_{ijk}\, L_k\,. \tag{14.34}$$

Rests on Definition 14.26.

Proof. *Derives Proposition 14.27.*

Write $R = \mathbb{1} + \varepsilon X + O(\varepsilon^2)$ in Equation (14.30). To first order $R^{\mathsf{T}} R = \mathbb{1}$ gives $X^{\mathsf{T}} + X = 0$, and $\det R = 1$ gives $\operatorname{tr} X = 0$, which is implied by antisymmetry; conversely exp of an antisymmetric matrix is orthogonal with unit determinant. The real antisymmetric 3×3 matrices form a three-dimensional space, matching Equation (14.31), and the L_k of Equation (14.32) are three independent members of it. Note that L_3 is the generator J of $\mathrm{SO}(2,\mathbb{R})$ found in Equation (14.27), embedded in the 1–2 block.

For the bracket, compute the (m,n) entry using $(L_i)_{mk} = -\epsilon_{imk}$:

$$\begin{aligned} \left([L_i, L_j]\right)_{mn} &= \epsilon_{imk}\epsilon_{jkn} - \epsilon_{jmk}\epsilon_{ikn} \\ &= (\delta_{in}\delta_{mj} - \delta_{ij}\delta_{mn}) - (\delta_{jn}\delta_{mi} - \delta_{ij}\delta_{mn}) \\ &= \delta_{in}\delta_{jm} - \delta_{im}\delta_{jn}\,, \end{aligned}$$

where the contraction $\epsilon_{kim}\epsilon_{knj} = \delta_{in}\delta_{mj} - \delta_{ij}\delta_{mn}$ was used twice after a cyclic relabelling. On the other hand

$$\left(\epsilon_{ijk} L_k\right)_{mn} = -\epsilon_{ijk}\epsilon_{kmn} = -\left(\delta_{im}\delta_{jn} - \delta_{in}\delta_{jm}\right)\,,$$

which is the same expression. □

For representation theory it is convenient to use Hermitian generators, because Hermitian operators are the ones that can represent observables and because the resulting brackets are those of the quantum theory.

Definition 14.28 (Hermitian rotation generators)**.**

$$\boxed{J_k = \mathrm{i} L_k\,, \qquad [J_i, J_j] = \mathrm{i}\,\epsilon_{ijk} J_k\,, \qquad R(\hat{n},\phi) = \exp\left(-\mathrm{i}\,\phi\,\hat{n}\cdot\boldsymbol{J}\right)}\,. \tag{14.35}$$

Each J_k is Hermitian: L_k is real and antisymmetric, so $(\mathrm{i}L_k)^\dagger = -\mathrm{i}L_k^{\mathsf{T}} = \mathrm{i}L_k$. The bracket follows from Equation (14.34) by $[J_i, J_j] = -[L_i, L_j] = -\epsilon_{ijk}L_k = \mathrm{i}\epsilon_{ijk}J_k$.

Rests on Proposition 14.27.

Remark 14.29 (Where $\hbar$ enters). Equation (14.35) is a statement about dimensionless matrices. The physical angular momentum operator is $\hat{\boldsymbol{\mathcal{J}}} = \hbar \boldsymbol{J}$, of SI unit J s, and multiplying Equation (14.35) by $\hbar^2$ turns it into $\left[\hat{\mathcal{J}}_i, \hat{\mathcal{J}}_j\right] = \mathrm{i}\hbar\,\epsilon_{ijk}\hat{\mathcal{J}}_k$, which is the commutator measured in the laboratory and the one used throughout Chapters 79 and 94 (cf. Equation (94.8)). The rotation operator is correspondingly $\exp\left(-\mathrm{i}\,\phi\,\hat{n}\cdot\hat{\boldsymbol{\mathcal{J}}}/\hbar\right)$, the ratio in the exponent being dimensionless as it must be. See Remark 14.24.

Proposition 14.30 (Rodrigues formula; the exponential map is onto)**.** *Write $N = \hat{n}\cdot\boldsymbol{L}$ for a unit vector $\hat{n}$. Then $N\boldsymbol{x} = \hat{n}\times\boldsymbol{x}$, $N^3 = -N$, and*

$$\exp(\phi N) = \mathbb{1} + \sin\phi\; N + (1 - \cos\phi)\, N^2\,, \tag{14.36}$$

so that

$$R(\hat{n},\phi)\,\boldsymbol{x} = \boldsymbol{x}\cos\phi + \hat{n}\,(\hat{n}\cdot\boldsymbol{x})\,(1-\cos\phi) + (\hat{n}\times\boldsymbol{x})\sin\phi\,. \tag{14.37}$$

Every element of $\mathrm{SO}(3,\mathbb{R})$ is of this form with $\phi \in [0,\pi]$, so the exponential map $\mathfrak{so}(3) \to \mathrm{SO}(3,\mathbb{R})$ is surjective.

Rests on Proposition 14.27 and Equations (14.25) and (14.28).

Proof. *Derives Proposition 14.30.*

From Equation (14.32), $(N\boldsymbol{x})_i = -\epsilon_{kij} n_k x_j = \epsilon_{ikj} n_k x_j = (\hat{n}\times\boldsymbol{x})_i$. Hence $N^2\boldsymbol{x} = \hat{n}\times(\hat{n}\times\boldsymbol{x}) = \hat{n}(\hat{n}\cdot\boldsymbol{x}) - \boldsymbol{x}$ and $N^3\boldsymbol{x} = \hat{n}\times(\hat{n}(\hat{n}\cdot\boldsymbol{x}) - \boldsymbol{x}) = -\hat{n}\times\boldsymbol{x} = -N\boldsymbol{x}$. Consequently

$N^{2k+1} = (-1)^k N$ and $N^{2k+2} = (-1)^k N^2$ for $k \in \mathbb{N} \cup \{0\}$, and splitting the exponential series into its odd and even parts, exactly as was done for $\mathrm{SO}(2,\mathbb{R})$ at Equation (14.28),

$$\exp(\phi N) = \mathbb{1} + \left(\sum_{k=0}^{\infty} \frac{(-1)^k \phi^{2k+1}}{(2k+1)!} \right) N + \left(\sum_{k=0}^{\infty} \frac{(-1)^k \phi^{2k+2}}{(2k+2)!} \right) N^2 ,$$

whose two coefficients are $\sin\phi$ and $1 - \cos\phi$. Substituting $N^2 \boldsymbol{x}$ gives Equation (14.37).

Surjectivity is Euler's theorem on rotations: $R \in \mathrm{SO}(3,\mathbb{R})$ is a real 3×3 matrix, so its characteristic polynomial is a real cubic and has a real root; the roots have modulus 1 because R is orthogonal, and their product is $\det R = 1$, so the real root is $+1$ and there is a unit vector $\hat{n}$ with $R\hat{n} = \hat{n}$. In an orthonormal basis adapted to $\hat{n}$, R acts as the identity on $\hat{n}$ and as an element of $\mathrm{SO}(2,\mathbb{R})$ on the orthogonal plane, which is Equation (14.25) for some angle ϕ; replacing $\hat{n}$ by $-\hat{n}$ if necessary brings ϕ into $[0,\pi]$. □

Proposition 14.31 ($\mathrm{SO}(3,\mathbb{R})$ is not simply connected)**.** *The map sending $\boldsymbol{\phi} = \phi\hat{n}$ with $\phi \in [0,\pi]$ to $R(\hat{n},\phi)$ identifies $\mathrm{SO}(3,\mathbb{R})$ with the closed ball of radius π in $\mathbb{R}^3$ with antipodal boundary points identified, that is with the real projective space $\mathbb{R}P^3$. It is compact, connected, and* not *simply connected in the sense of Definition 6.17: the loop traced by $R(\hat{n},\phi)$ as ϕ runs from 0 to 2π cannot be contracted to a point, while the same loop traversed twice can.*

Rests on Proposition 14.30, Definition 6.17, and Corollary 14.38.

Derivation. *Derives Proposition 14.31.*

That the map is onto is Proposition 14.30, and it is injective on the open ball because $R(\hat{n},\phi)$ determines its fixed axis and its rotation angle; on the boundary sphere $\phi = \pi$ the only failure of injectivity is $R(\hat{n},\pi) = R(-\hat{n},\pi)$, immediate from Equation (14.37) since $\cos\pi = -1$ and $\sin\pi = 0$ leave only terms even in $\hat{n}$. A continuous bijection from a compact space to a Hausdorff space is a homeomorphism, so $\mathrm{SO}(3,\mathbb{R}) \cong \mathbb{R}P^3$; connectedness and compactness follow. The statement about loops is Corollary 14.38 below, where it is read off the double cover. □

The group SU(2)

Definition 14.32 (The special unitary group in two dimensions)**.**

$$\mathrm{SU}(2) = \left\{ U \in \mathrm{GL}(2,\mathbb{C}) \mid U^\dagger U = \mathbb{1},\ \det U = 1 \right\} . \tag{14.38}$$

Proposition 14.33 ($\mathrm{SU}(2)$ is the three-sphere)**.** *Every $U \in \mathrm{SU}(2)$ has the form*

$$U = \begin{pmatrix} \alpha & \beta \\ -\bar{\beta} & \bar{\alpha} \end{pmatrix} , \qquad |\alpha|^2 + |\beta|^2 = 1 , \tag{14.39}$$

and $(\alpha,\beta) \mapsto U$ is a homeomorphism from the unit sphere $S^3 \subset \mathbb{R}^4 \cong \mathbb{C}^2$ onto $\mathrm{SU}(2)$. Hence $\mathrm{SU}(2)$ is compact, connected and simply connected.

Rests on Definition 14.32.

Proof. *Derives Proposition 14.33.*

Write

$$U = \begin{pmatrix} a & b \\ c & d \end{pmatrix} .$$

With $\det U = 1$ the inverse is $U^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$, while unitarity says $U^{-1} = U^\dagger = \begin{pmatrix} \bar{a} & \bar{c} \\ \bar{b} & \bar{d} \end{pmatrix}$. Comparing entries, $d = \bar{a}$ and $c = -\bar{b}$, which is Equation (14.39) with $\alpha = a$, $\beta = b$; the determinant condition then reads $|\alpha|^2 + |\beta|^2 = 1$. Conversely every such matrix is unitary with unit determinant. The correspondence and its inverse are given by polynomials in the real and imaginary parts, so both are continuous. Finally S^3 is connected and simply connected by Lemma 14.34. □

The last step is the only one that is not immediate, and since the distinction between $\mathrm{SU}(2)$ and $\mathrm{SO}(3,\mathbb{R})$ turns entirely on it — and with that distinction, the existence of half-integer spin — it is proved rather than quoted. Its natural home is the topology chapter, whose Definition 6.17 it uses.

Lemma 14.34 (The n-sphere is simply connected for $n \geq 2$)**.** *For $n \geq 2$ the unit sphere $S^n = \{y \in \mathbb{R}^{n+1} \mid \|y\| = 1\}$ is path-connected and simply connected.*
Rests on Definition 6.17, Example 6.18, and Theorem 7.25.

Proof. *Derives Lemma 14.34.*
For $p, q \in S^n$ with $q \neq -p$ define the normalized chord

$$\sigma_{p,q}(u) = \frac{(1-u)p + uq}{\|(1-u)p + uq\|}\,, \qquad 0 \leq u \leq 1\,. \tag{14.40}$$

It is well defined, because $\|(1-u)p+uq\|^2 = (1-u)^2 + u^2 + 2u(1-u)\,p\cdot q$ can vanish only if $u = 1/2$ and $p \cdot q = -1$, that is only if $q = -p$; it is jointly continuous in p, q and u; it runs from p to q; and its image lies in the plane spanned by p and q. Path-connectedness follows at once: two points are joined by Equation (14.40) unless they are antipodal, and an antipodal pair is joined through any third point.

Every loop is homotopic to a piecewise-circular one. Let $\gamma : [0,1] \longrightarrow S^n$ be continuous with $\gamma(0) = \gamma(1)$. Each of the $n+1$ components of γ is uniformly continuous by Theorem 7.25, whose proof applies verbatim to a map into $\mathbb{R}^{n+1}$, so there is an N with $\|\gamma(s) - \gamma(s')\| < 1$ whenever $|s - s'| \leq 1/N$. Put $p_k = \gamma(k/N)$. Consecutive p_k are not antipodal, their distance being less than $1 < 2$, so

$$\tilde\gamma(s) = \sigma_{p_k,p_{k+1}}\,(Ns - k)\;, \qquad \frac{k}{N} \leq s \leq \frac{k+1}{N}\,,$$

is a well-defined continuous loop with $\tilde\gamma(0) = p_0 = \gamma(0)$ and $\tilde\gamma(1) = p_N = \gamma(1)$. For s in the k-th subinterval both $\|\gamma(s) - p_k\| < 1$ and $\|\gamma(s) - p_{k+1}\| < 1$, and for unit vectors $\|a-b\|^2 = 2 - 2\,a\cdot b$, so $\gamma(s)\cdot p_k > 1/2$ and $\gamma(s)\cdot p_{k+1} > 1/2$. Since $\tilde\gamma(s)$ is a combination of p_k and p_{k+1} with nonnegative coefficients divided by a norm at most 1,

$$\gamma(s)\cdot\tilde\gamma(s) = \frac{(1-u)\,\gamma(s)\cdot p_k + u\,\gamma(s)\cdot p_{k+1}}{\|(1-u)p_k + up_{k+1}\|} \geq \tfrac{1}{2}\,, \qquad u = Ns - k\,.$$

Hence $(1-t)\gamma(s) + t\tilde\gamma(s)$ never vanishes, its squared norm being at least $(1-t)^2 + t^2$, and

$$H(s,t) = \frac{(1-t)\gamma(s) + t\,\tilde\gamma(s)}{\|(1-t)\gamma(s) + t\,\tilde\gamma(s)\|}$$

is a homotopy from γ to $\tilde\gamma$ fixing the base point.

A piecewise-circular loop misses a point. The image of $\tilde\gamma$ lies in the union of the N sets $C_k = S^n \cap \Pi_k$, where Π_k is a plane of dimension at most two containing p_k and p_{k+1}; enlarging it if necessary, take $\dim \Pi_k = 2$. Because $n+1 \geq 3$, the planes $\Pi' = \operatorname{span}\{e_1, \cos\alpha\, e_2 + \sin\alpha\, e_3\}$ are pairwise distinct for $\alpha \in [0,\pi)$ and infinite in number, so some Π' differs from all N of the Π_k. Two distinct two-dimensional planes meet in a subspace of dimension at most one, so $S^n \cap \Pi' \cap \Pi_k$ has at most two points for each k; the circle $S^n \cap \Pi'$ is infinite, and therefore contains a point x lying on no C_k.

The complement of a point is $\mathbb{R}^n$. Stereographic projection from x,

$$\pi(y) = \frac{y - (y\cdot x)\,x}{1 - y\cdot x}\,, \tag{14.41}$$

maps $S^n \setminus \{x\}$ into the n-dimensional space $x^\perp \cong \mathbb{R}^n$ and is continuous there. So is $w \mapsto \big(2w + (\|w\|^2 - 1)x\big)/(\|w\|^2 + 1)$, which lands on S^n — its squared norm is

$\left(4\|w\|^2 + (\|w\|^2-1)^2\right)/(\|w\|^2+1)^2 = 1$ — and a direct substitution shows the two maps are mutually inverse. Hence $S^n \setminus \{x\}$ is homeomorphic to $\mathbb{R}^n$, which is simply connected by Example 6.18. The loop $\tilde{\gamma}$ therefore contracts to its base point inside $S^n \setminus \{x\}$, hence inside S^n; composing that homotopy with H contracts γ. □

Proposition 14.35 (Generators of $\mathfrak{su}(2)$)**.** *The Lie algebra $\mathfrak{su}(2)$ is the space of traceless anti-Hermitian 2×2 matrices, of real dimension 3. Writing its elements as $-\mathrm{i}\theta^k T_k$ with θ^k real, a basis of Hermitian generators is*

$$\boxed{T_k = \tfrac{1}{2}\sigma_k}, \qquad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{14.42}$$

the Pauli matrices [587]. They obey

$$\sigma_i\sigma_j = \delta_{ij}\mathbb{1} + \mathrm{i}\,\epsilon_{ijk}\sigma_k\,, \tag{14.43}$$

and therefore

$$\boxed{[T_i, T_j] = \mathrm{i}\,\epsilon_{ijk}T_k}, \qquad \operatorname{tr}(T_iT_j) = \tfrac{1}{2}\delta_{ij}\,. \tag{14.44}$$

Rests on Definition 14.32.

Proof. *Derives Proposition 14.35.*

Setting $U = \mathbb{1} + \varepsilon X + O(\varepsilon^2)$ in Equation (14.38) gives $X^\dagger + X = 0$ and $\operatorname{tr} X = 0$. A traceless anti-Hermitian 2×2 matrix is $\begin{pmatrix} \mathrm{i}u & v+\mathrm{i}w \\ -v+\mathrm{i}w & -\mathrm{i}u \end{pmatrix}$ with u, v, w real, a three-dimensional real space, and the three matrices $-\mathrm{i}\sigma_k/2$ span it. Equation (14.43) is checked directly: $\sigma_i^2 = \mathbb{1}$ for each i by inspection, and the three products $\sigma_1\sigma_2 = \mathrm{i}\sigma_3$, $\sigma_2\sigma_3 = \mathrm{i}\sigma_1$, $\sigma_3\sigma_1 = \mathrm{i}\sigma_2$ are likewise immediate, the remaining cases following from $\sigma_j\sigma_i = 2\delta_{ij}\mathbb{1} - \sigma_i\sigma_j$. Antisymmetrizing Equation (14.43) gives $[\sigma_i, \sigma_j] = 2\mathrm{i}\epsilon_{ijk}\sigma_k$, that is Equation (14.44); taking the trace of Equation (14.43) and using $\operatorname{tr}\sigma_k = 0$ gives $\operatorname{tr}(\sigma_i\sigma_j) = 2\delta_{ij}$. □

Equation (14.44) is *the same algebra* as Equation (14.35). The two groups therefore have isomorphic Lie algebras, $\mathfrak{su}(2) \cong \mathfrak{so}(3)$, and cannot be distinguished by any local construction. They are nevertheless different groups, and the next result says exactly how.

Proposition 14.36 (Closed form of the exponential)**.** *For a unit vector $\hat{n}$,*

$$U(\hat{n}, \psi) = \exp\left(-\tfrac{\mathrm{i}}{2}\psi\,\hat{n}\cdot\boldsymbol{\sigma}\right) = \mathbb{1}\cos\frac{\psi}{2} - \mathrm{i}\,(\hat{n}\cdot\boldsymbol{\sigma})\sin\frac{\psi}{2}\,, \tag{14.45}$$

and in particular $U(\hat{n}, 2\pi) = -\mathbb{1}$ and $U(\hat{n}, 4\pi) = \mathbb{1}$.

Rests on Proposition 14.35 and Equation (14.28).

Proof. *Derives Proposition 14.36.*

By Equation (14.43), $(\hat{n}\cdot\boldsymbol{\sigma})^2 = n_in_j\sigma_i\sigma_j = n_in_j\delta_{ij}\mathbb{1} = \mathbb{1}$, the ϵ term dropping because n_in_j is symmetric. The even and odd powers of $\hat{n}\cdot\boldsymbol{\sigma}$ therefore alternate exactly as J did in Equation (14.27), and the same resummation that produced Equation (14.28) produces Equation (14.45). Setting $\psi = 2\pi$ gives $\cos\pi = -1$, $\sin\pi = 0$. □

The double cover $\mathrm{SU}(2) \to \mathrm{SO}(3,\mathbb{R})$

Theorem 14.37 ($\mathrm{SU}(2)$ is a two-to-one cover of $\mathrm{SO}(3,\mathbb{R})$)**.** *Identify $\boldsymbol{x} \in \mathbb{R}^3$ with the traceless Hermitian matrix $X = \boldsymbol{x}\cdot\boldsymbol{\sigma}$, and define*

$$\boxed{\Phi(U):\ \boldsymbol{x}\cdot\boldsymbol{\sigma} \longmapsto U\,(\boldsymbol{x}\cdot\boldsymbol{\sigma})\,U^\dagger}. \tag{14.46}$$

Then Φ *is a surjective group homomorphism* $\mathrm{SU}(2) \to \mathrm{SO}(3,\mathbb{R})$ *with*

$$\ker \Phi = \{\mathbb{1}, -\mathbb{1}\} \cong \mathbb{Z}_2\,, \tag{14.47}$$

and explicitly $\Phi\left(U(\hat{n},\phi)\right) = R(\hat{n},\phi)$.

Rests on Propositions 14.30, 14.33, 14.35 and 14.36.

Proof. *Derives Theorem 14.37.*

We check the four claims in turn.

$\Phi(U)$ *is a well-defined real linear map on* $\mathbb{R}^3$. The matrices $\boldsymbol{x}\cdot\boldsymbol{\sigma}$ are exactly the traceless Hermitian 2×2 matrices, since $\{\sigma_1,\sigma_2,\sigma_3\}$ is a real basis of that space. If X is traceless Hermitian then so is $UXU^\dagger$: Hermiticity because $\left(UXU^\dagger\right)^\dagger = UX^\dagger U^\dagger = UXU^\dagger$, and tracelessness by cyclicity together with $U^\dagger U = \mathbb{1}$. So $\Phi(U)$ maps $\mathbb{R}^3$ to $\mathbb{R}^3$ linearly.

$\Phi(U)$ *preserves the Euclidean norm.* From Equation (14.43), $\det(\boldsymbol{x}\cdot\boldsymbol{\sigma}) = -\boldsymbol{x}\cdot\boldsymbol{x}$, as one sees by writing out $\begin{pmatrix} x_3 & x_1 - \mathrm{i}x_2 \\ x_1 + \mathrm{i}x_2 & -x_3 \end{pmatrix}$. Since $\det U = 1$, $\det\left(UXU^\dagger\right) = \det X$, so the norm of $\boldsymbol{x}$ is unchanged and $\Phi(U) \in \mathrm{O}(3,\mathbb{R})$.

The image lies in $\mathrm{SO}(3,\mathbb{R})$*, and* Φ *is a homomorphism.* $\Phi(U_1U_2) = \Phi(U_1)\Phi(U_2)$ is immediate from Equation (14.46). The map $U \mapsto \det\Phi(U) \in \{+1,-1\}$ is continuous on the connected space $\mathrm{SU}(2)$ (Proposition 14.33) and equals $+1$ at $U = \mathbb{1}$, hence equals $+1$ everywhere.

The kernel. Suppose $UXU^\dagger = X$ for every traceless Hermitian X, i.e. U commutes with $\sigma_1,\sigma_2,\sigma_3$. Writing $U = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, commuting with σ_3 forces $b = c = 0$, and commuting with σ_1 then forces $a = d$. So $U = a\mathbb{1}$, and $\det U = a^2 = 1$ gives $a = \pm 1$. Both values do lie in the kernel, giving Equation (14.47).

Surjectivity, and the explicit form. It suffices to compute $\Phi\left(U(\hat{n},\phi)\right)$ and find the rotation Equation (14.37), since those exhaust $\mathrm{SO}(3,\mathbb{R})$ by Proposition 14.30. Abbreviate $\mathrm{c} = \cos(\phi/2)$, $\mathrm{s} = \sin(\phi/2)$ and $n = \hat{n}\cdot\boldsymbol{\sigma}$, $X = \boldsymbol{x}\cdot\boldsymbol{\sigma}$. Using Equation (14.45),

$$\begin{aligned} UXU^\dagger &= (\mathrm{c}\,\mathbb{1} - \mathrm{i}\mathrm{s}\,n)\,X\,(\mathrm{c}\,\mathbb{1} + \mathrm{i}\mathrm{s}\,n) \\ &= \mathrm{c}^2 X + \mathrm{i}\,\mathrm{cs}\,[X,n] + \mathrm{s}^2\,nXn\,. \end{aligned}$$

Now Equation (14.43) gives $[X,n] = x_i n_j\,[\sigma_i,\sigma_j] = 2\mathrm{i}\,\epsilon_{ijk}x_i n_j \sigma_k = -2\mathrm{i}\,(\hat{n}\times\boldsymbol{x})\cdot\boldsymbol{\sigma}$, and, using $nX + Xn = 2\,(\hat{n}\cdot\boldsymbol{x})\,\mathbb{1}$ from the symmetric part of the same identity together with $n^2 = \mathbb{1}$,

$$nXn = \left(2\,(\hat{n}\cdot\boldsymbol{x})\,\mathbb{1} - Xn\right)n = 2\,(\hat{n}\cdot\boldsymbol{x})\,n - X\,.$$

Collecting the three pieces and using $\mathrm{c}^2 - \mathrm{s}^2 = \cos\phi$, $2\mathrm{cs} = \sin\phi$, $2\mathrm{s}^2 = 1 - \cos\phi$,

$$UXU^\dagger = \left[\boldsymbol{x}\cos\phi + \hat{n}\,(\hat{n}\cdot\boldsymbol{x})\,(1-\cos\phi) + (\hat{n}\times\boldsymbol{x})\sin\phi\right]\cdot\boldsymbol{\sigma}\,,$$

which is exactly Equation (14.37). □

Corollary 14.38 (SU(2) is the universal cover)**.** $\mathrm{SO}(3,\mathbb{R}) \cong \mathrm{SU}(2)/\{\mathbb{1},-\mathbb{1}\}$*, and since* $\mathrm{SU}(2)$ *is simply connected (Proposition 14.33) it is the universal covering group of* $\mathrm{SO}(3,\mathbb{R})$*, with*

$$\pi_1\left(\mathrm{SO}(3,\mathbb{R})\right) \cong \mathbb{Z}_2\,. \tag{14.48}$$

Concretely: the path $\phi \mapsto U(\hat{n},\phi)$*,* $0 \le \phi \le 2\pi$*, joins* $\mathbb{1}$ *to* $-\mathbb{1}$ *in* $\mathrm{SU}(2)$ *and projects to a closed loop in* $\mathrm{SO}(3,\mathbb{R})$ *that is not contractible, while the path with* $0 \le \phi \le 4\pi$ *is a closed loop upstairs and projects to a contractible one.*

Rests on Theorem 14.37 and Proposition 14.33.

Everything in the corollary except Equation (14.48) — that the quotient is $\mathrm{SO}(3,\mathbb{R})$, that the kernel has two elements, that the two paths are as described — is proved above from the material of this chapter. The remaining step, from a two-sheeted covering by a simply connected space to the fundamental group of the base, is covering-space theory: the homotopy lifting property, unique path lifting, and the resulting isomorphism between the fundamental group of the base and the deck-transformation group, here the kernel of the covering homomorphism.

Full derivation: Appendix A.33 (p. 2946).

Derives Corollary 14.38.

Remark 14.39 (This is why half-integer spin exists). A quantum state is a ray, so a symmetry need only be represented up to a phase: what a connected symmetry group must furnish is a projective unitary representation, and by Bargmann's analysis these are the ordinary unitary representations of the universal covering group [1029]. For rotations that covering group is $\mathrm{SU}(2)$, not $\mathrm{SO}(3,\mathbb{R})$. Theorem 14.43 below will show that an irreducible representation of $\mathfrak{su}(2)$ carries a label j with $2j \in \mathbb{N} \cup \{0\}$ and Proposition 14.44 that every such j occurs, and Proposition 14.45 will show that exactly the integer ones descend to $\mathrm{SO}(3,\mathbb{R})$ while the half-integer ones do not. The half-integer representations are therefore not an extra postulate: they are what the double cover makes available, and the electron uses them [562, 587]. The same argument in the relativistic setting replaces $\mathrm{SU}(2)$ by $\mathrm{SL}(2,\mathbb{C})$ and is carried out in Proposition 94.10.

The irreducible representations of $\mathfrak{su}(2)$

The construction that follows is the ladder-operator argument, and it is worth noting how little goes into it: only the brackets Equation (14.44), the existence of the Casimir guaranteed by Corollary 14.17, and finite dimensionality [687].

Definition 14.40 (Ladder operators and the Casimir)**.** In the complexified algebra put

$$J_{\pm} = J_1 \pm \mathrm{i} J_2 \,, \qquad \boldsymbol{J}^2 = J_1^2 + J_2^2 + J_3^2 \,. \tag{14.49}$$

Rests on Equation (14.44).

Lemma 14.41 (Ladder algebra)**.**

$$[J_3, J_{\pm}] = \pm J_{\pm} \,, \qquad [J_+, J_-] = 2J_3 \,, \tag{14.50}$$

and

$$\boldsymbol{J}^2 = J_- J_+ + J_3\,(J_3+1) = J_+ J_- + J_3\,(J_3-1) \ . \tag{14.51}$$

Moreover $\boldsymbol{J}^2$ is the quadratic Casimir of Corollary 14.17 up to normalisation: the Killing form of $\mathfrak{su}(2)$ in the Hermitian basis is $\kappa_{ab} = 2\delta_{ab}$, so $C_2 = \kappa^{ab} J_a J_b = \tfrac{1}{2}\boldsymbol{J}^2$.

Rests on Definition 14.40, Equations (14.15) and (14.44), and Corollary 14.17.

Proof. *Derives Lemma 14.41.*

From Equation (14.44), $[J_3, J_1 \pm \mathrm{i}J_2] = \mathrm{i}J_2 \pm \mathrm{i}(-\mathrm{i}J_1) = \pm(J_1 \pm \mathrm{i}J_2)$, and $[J_+, J_-] = -2\mathrm{i}\,[J_1, J_2] = 2J_3$. Next

$$J_- J_+ = (J_1 - \mathrm{i}J_2)\,(J_1 + \mathrm{i}J_2) = J_1^2 + J_2^2 + \mathrm{i}\,[J_1, J_2] = \boldsymbol{J}^2 - J_3^2 - J_3 \,,$$

which rearranges to the first form of Equation (14.51); the second follows by exchanging the roles of J_+ and J_-, which flips the sign of the $[J_1, J_2]$ term. For the Killing form use Equation (14.15) with $C^c{}_{ab} = \mathrm{i}\epsilon_{abc}$ from Equation (14.44): $\kappa_{ab} = (\mathrm{i}\epsilon_{adc})\,(\mathrm{i}\epsilon_{bcd}) = -\epsilon_{adc}\epsilon_{bcd} = \epsilon_{adc}\epsilon_{bdc} = 2\delta_{ab}$. □

Remark 14.42 (The compact real form). The value $\kappa_{ab} = +2\delta_{ab}$ is computed in the basis of *Hermitian* generators J_a, which spans the complexification. On the real algebra $\mathfrak{su}(2)$ itself, whose elements are the anti-Hermitian $-\mathrm{i}\theta^a J_a = \theta^a L_a$, the same computation gives $\kappa_{ab} = -2\delta_{ab}$: the Killing form of a compact real form is negative definite. Nondegeneracy, which is all that Corollary 14.17 requires, holds either way.

Theorem 14.43 (The irreducible representations of $\mathfrak{su}(2)$)**.** *Let D be a finite-dimensional irreducible representation of $\mathfrak{su}(2)$ on a complex inner-product space, unitary in the sense that the J_a are represented by Hermitian operators. Then there is a number j with*

$$\boxed{2j \in \mathbb{N} \cup \{0\}} \tag{14.52}$$

such that the representation has dimension $2j+1$, an orthonormal basis $\{|j,m\rangle\}$ with $m = -j, -j+1, \dots, j-1, j$, and

$$\boxed{J_3\,|j,m\rangle = m\,|j,m\rangle\ , \qquad \boldsymbol{J}^2\,|j,m\rangle = j(j+1)\,|j,m\rangle}\,, \tag{14.53}$$

$$J_\pm\,|j,m\rangle = \sqrt{j(j+1) - m(m\pm 1)}\ \,|j,m\pm 1\rangle\ . \tag{14.54}$$

Rests on Lemma 14.41 and Equation (14.44).

Proof. *Derives Theorem 14.43.*

J_3 is Hermitian on a finite-dimensional space, so it has an eigenvector; let μ be an eigenvalue of largest real part, with eigenvector $|\mu\rangle$. From Equation (14.50), $J_3\,(J_\pm\,|\mu\rangle) = (\mu \pm 1)\,J_\pm\,|\mu\rangle$, so $J_\pm$ raises and lowers the J_3 eigenvalue by one. Maximality of μ forces $J_+\,|\mu\rangle = 0$. Write $j := \mu$ and $|j,j\rangle := |\mu\rangle$ normalised. Then by Equation (14.51),

$$\boldsymbol{J}^2\,|j,j\rangle = (J_-J_+ + J_3(J_3+1))\,|j,j\rangle = j(j+1)\,|j,j\rangle\ ,$$

and since $\boldsymbol{J}^2$ commutes with J_3 and with $J_\pm$ its eigenspaces are invariant under the whole algebra, so by irreducibility the value $j(j+1)$ is taken on the entire representation. (This is Corollary 14.18 again, argued here directly so that nothing about the group is needed.)

Define $|j,m\rangle$ recursively by lowering. Because J_+ and J_- are adjoint to each other when the J_a are Hermitian,

$$\|J_-\,|j,m\rangle\|^2 = \langle j,m|\,J_+J_-\,|j,m\rangle = j(j+1) - m(m-1)\,,$$

using the second form of Equation (14.51); the same computation with J_+ and the first form gives $\|J_+\,|j,m\rangle\|^2 = j(j+1) - m(m+1)$, which is Equation (14.54) once phases are fixed by choosing the square roots positive. Since the space is finite-dimensional the chain of eigenvalues $j, j-1, j-2, \dots$ must terminate, which by the norm formula happens exactly at the value $m_{\min}$ with $j(j+1) - m_{\min}(m_{\min}-1) = 0$, that is $m_{\min} \in \{-j,\ j+1\}$; the second is excluded because $m_{\min} \le j$ and would need $j+1 \le j$. Hence $m_{\min} = -j$, and the chain runs from j down to $-j$ in unit steps, so $j - (-j) = 2j$ is a non-negative integer. *This is the whole reason half-integer angular momentum is possible: nothing in the algebra forbids the chain from having an even number of steps.* The span of the $|j,m\rangle$ is invariant under J_3 and $J_\pm$, hence under the whole algebra, and is nonzero, so by irreducibility it is the whole space, of dimension $2j+1$. □

Theorem 14.43 is a statement of *necessity*: an irreducible representation, if there is one, looks like that. It does not produce a single representation, and everything below — the descent to $\mathrm{SO}(3,\mathbb{R})$, the Clebsch–Gordan series, the spin label of a Poincaré representation — needs to know that the list is not empty. The converse is the shorter half, because the theorem has already written down the matrix elements a representation would have to have; it remains to check that they define one.

Proposition 14.44 (Existence: every allowed j is realized)**.** *For each j with $2j \in \mathbb{N} \cup \{0\}$ there exists a $(2j+1)$-dimensional irreducible representation of $\mathfrak{su}(2)$ by Hermitian operators, and its matrix elements are Equations (14.53) and (14.54). Together with Theorem 14.43 this puts the equivalence classes of finite-dimensional irreducible representations of $\mathfrak{su}(2)$ in bijection with the numbers $j = 0, \frac{1}{2}, 1, \frac{3}{2}, \dots$*

Rests on Theorem 14.43 and Lemma 14.41.

Proof. *Derives Proposition 14.44.*

Let $\mathbb{V}_j$ be the complex inner-product space with orthonormal basis $\{|j,m\rangle\}$, $m = -j, -j+1, \dots, j$, of dimension $2j+1$, and define operators on it by

$$J_3 \,|j,m\rangle = m \,|j,m\rangle \;, \qquad J_\pm \,|j,m\rangle = c_\pm(m) \,|j,m\pm 1\rangle \;, \tag{14.55}$$

$$c_\pm(m) = \sqrt{j(j+1) - m(m\pm 1)}\,, \tag{14.56}$$

with the convention that $|j,m\rangle$ denotes the zero vector for $|m| > j$. The convention is consistent because $j(j+1) - m(m+1) = (j-m)(j+m+1)$ and $j(j+1) - m(m-1) = (j+m)(j-m+1)$, which are non-negative for $-j \le m \le j$ and vanish exactly at $c_+(j) = 0$ and $c_-(-j) = 0$: the operators do not attempt to leave the basis. Put $J_1 = \frac{1}{2}(J_+ + J_-)$ and $J_2 = \frac{1}{2\mathrm{i}}(J_+ - J_-)$.

The brackets hold. $J_3 J_\pm \,|j,m\rangle = (m \pm 1) c_\pm(m) \,|j,m\pm 1\rangle$ while $J_\pm J_3 \,|j,m\rangle = m\, c_\pm(m) \,|j,m\pm 1\rangle$, so $[J_3, J_\pm] = \pm J_\pm$. Since $c_+(m-1) = c_-(m)$ by Equation (14.56), $J_+ J_- \,|j,m\rangle = c_-(m)^2 \,|j,m\rangle$ and $J_- J_+ \,|j,m\rangle = c_+(m)^2 \,|j,m\rangle$, whence

$$[J_+, J_-] \,|j,m\rangle = \big([j(j+1) - m(m-1)] - [j(j+1) - m(m+1)]\big) \,|j,m\rangle = 2m \,|j,m\rangle \;,$$

which is $2J_3 \,|j,m\rangle$. That is Equation (14.50), equivalent to Equation (14.44) by the computation in Lemma 14.41 run backwards.

The generators are Hermitian. The basis is orthonormal and the $c_\pm$ are real, so the only nonzero matrix elements of the ladder operators are $\langle j, m+1| \, J_+ \,|j,m\rangle = c_+(m)$ and $\langle j,m| \, J_- \,|j,m+1\rangle = c_-(m+1) = c_+(m)$. Hence $J_- = J_+^\dagger$, and J_1, J_2, J_3 are Hermitian.

It is irreducible. Let $\mathbb{W} \neq 0$ be an invariant subspace. It is invariant under the Hermitian J_3, whose eigenvalues on $\mathbb{V}_j$ are the $2j+1$ distinct numbers m, so $\mathbb{W}$ is spanned by those $|j,m\rangle$ it contains and contains at least one of them. From any one of them the ladder operators reach every other, their coefficients $c_\pm(m)$ vanishing only at the two ends of the chain. So $\mathbb{W} = \mathbb{V}_j$.

Finally Equation (14.51) gives $\boldsymbol{J}^2 \,|j,m\rangle = \big(c_+(m)^2 + m(m+1)\big) \,|j,m\rangle = j(j+1) \,|j,m\rangle$, so Equations (14.55) and (14.56) are Equations (14.53) and (14.54). □

Proposition 14.45 (Which representations descend to $\mathrm{SO}(3,\mathbb{R})$)**.** *Let $D^{(j)}$ be the representation of $\mathrm{SU}(2)$ obtained by exponentiating the representation of $\mathfrak{su}(2)$ constructed in Proposition 14.44. Then*

$$D^{(j)}(-\mathbb{1}) = (-1)^{2j}\,\mathbb{1}\,, \tag{14.57}$$

so $D^{(j)}$ descends to a genuine representation of $\mathrm{SO}(3,\mathbb{R}) = \mathrm{SU}(2)/\{\pm\mathbb{1}\}$ if and only if j is an integer. For half-integer j the map $R \mapsto D^{(j)}(U)$ is defined only up to the sign ambiguity in the choice of U above R: a projective representation, in the sense of Remark 14.39.

Rests on Propositions 14.36 and 14.44 and Corollary 14.38.

Proof. *Derives Proposition 14.45.*

By Proposition 14.36, $-\mathbb{1} = U(\hat{e}_3, 2\pi) = \exp(-2\pi\mathrm{i}\, T_3)$, so $D^{(j)}(-\mathbb{1}) = \exp(-2\pi\mathrm{i} J_3)$ in the representation, and on $|j,m\rangle$ this is $\mathrm{e}^{-2\pi\mathrm{i}m}$ by Equation (14.53). The values of m are $j, j-1, \dots, -j$: all integers when j is an integer, all in $\mathbb{Z} + \frac{1}{2}$ when j is half-odd-integer. In the first case $\mathrm{e}^{-2\pi\mathrm{i}m} = 1$ for every m, in the second $\mathrm{e}^{-2\pi\mathrm{i}m} = -1$ for every m; either way the operator is $(-1)^{2j}\mathbb{1}$. A representation of $\mathrm{SU}(2)$ factors through the quotient by $\{\pm\mathbb{1}\}$ precisely when it is trivial on that subgroup, which by Equation (14.57) happens exactly for integer j. □

Remark 14.46 (The integer case is what one already knew). For $j = 1$ the representation of Theorem 14.43 is three-dimensional and is equivalent to the defining representation of $\mathrm{SO}(3,\mathbb{R})$ on $\mathbb{R}^3$ complexified — both have $\boldsymbol{J}^2 = 2$, as one checks directly from Equation (14.35), and both are irreducible of dimension three. For general integer j the representation is the one carried by the spherical harmonics of degree j, of dimension $2j+1$. In SI terms (Remark 14.29) the observable statement is

$$\boxed{\hat{\mathcal{J}}^2 = \hbar^2\, j(j+1)}\,, \qquad \hat{\mathcal{J}}_3 = \hbar m\,, \tag{14.58}$$

with $\hbar^2 j(j+1)$ carrying the SI unit $\mathrm{J}^2\,\mathrm{s}^2$ and $\hbar m$ the unit $\mathrm{J}\,\mathrm{s}$.

The Clebsch–Gordan decomposition

Theorem 14.47 (Clebsch–Gordan series)**.** *Let V_{j_1} and V_{j_2} be the irreducible representations of $\mathfrak{su}(2)$ of Theorem 14.43. Under the total generators $\boldsymbol{J} = \boldsymbol{J}^{(1)} \otimes \mathbb{1} + \mathbb{1} \otimes \boldsymbol{J}^{(2)}$,*

$$\boxed{V_{j_1} \otimes V_{j_2} = \bigoplus_{J=|j_1-j_2|}^{j_1+j_2} V_J}\,, \tag{14.59}$$

each V_J occurring exactly once, and the dimensions check:

$$(2j_1+1)\,(2j_2+1) = \sum_{J=|j_1-j_2|}^{j_1+j_2} (2J+1)\,. \tag{14.60}$$

Rests on Theorem 14.43 and Equation (14.44).

Proof. *Derives Theorem 14.47.*

That $\boldsymbol{J}$ satisfies Equation (14.44) is immediate, the two factors commuting. Take $j_1 \geq j_2$ without loss of generality. In the product basis $|j_1 m_1\rangle \otimes |j_2 m_2\rangle$ the operator J_3 is diagonal with eigenvalue $m = m_1 + m_2$, so the multiplicity $n(m)$ of the eigenvalue m is the number of pairs (m_1, m_2) with $m_1 + m_2 = m$, $|m_i| \leq j_i$. Counting them,

$$n(m) = \begin{cases} j_1 + j_2 - |m| + 1\,, & j_1 - j_2 \leq |m| \leq j_1 + j_2\,, \\ 2j_2 + 1\,, & |m| \leq j_1 - j_2\,, \\ 0\,, & |m| > j_1 + j_2\,. \end{cases} \tag{14.61}$$

Now the tensor product is a finite-dimensional representation, hence a direct sum of irreducibles V_J with multiplicities $a_J \geq 0$; by Equation (14.53) each V_J contributes exactly 1 to $n(m)$ for every $|m| \leq J$ and nothing otherwise, so $n(m) = \sum_{J \geq |m|} a_J$. Subtracting consecutive values,

$$a_J = n(J) - n(J+1)\,,$$

and Equation (14.61) makes this 1 for $|j_1 - j_2| \leq J \leq j_1 + j_2$ and 0 otherwise, which is Equation (14.59). For Equation (14.60) write $J = j_1 - j_2 + k$ with $k = 0, \dots, 2j_2$:

$$\begin{aligned} \sum_{k=0}^{2j_2} (2(j_1 - j_2 + k) + 1) &= (2j_2+1)\,(2j_1 - 2j_2 + 1) + 2\,\frac{2j_2\,(2j_2+1)}{2} \\ &= (2j_2+1)\,(2j_1+1)\,. \end{aligned}$$

□

Remark 14.48 (What is and is not proved here). Theorem 14.47 settles *which* irreducible representations occur and with what multiplicity, and that is all the group theory of the following parts requires of it. The unitary matrix relating the product basis to the coupled basis — the Clebsch–Gordan coefficients themselves, together with the $3j$, $6j$ and $9j$ symbols built from them — is a separate construction and belongs where it is used. This treatise reserves it for Sections 79.4.2 and 79.4.4 of Chapter 79, which at the time of writing carry their headings and their editorial specification but not yet the construction; nothing in the present chapter depends on it. The sources are Wigner's book for the coefficients themselves, of which the $3j$ symbol is his symmetrized form [687]; Racah's second paper for the closed form of the general coefficient and for the recoupling coefficient now written $6j$ [867]; and his third for the recoupling of four angular momenta, the modern $9j$ symbol [873]. The character-theoretic route to the same decomposition, using the product rule Equation (5.195), is available from Chapter 5; the weight count used above was preferred because it needs nothing beyond Equation (14.53).

The group SU(3)

Definition 14.49 (The special unitary group in three dimensions)**.**

$$\mathrm{SU}(3) = \left\{ U \in \mathrm{GL}(3,\mathbb{C}) \mid U^{\dagger}U = \mathbb{1},\ \det U = 1 \right\}, \tag{14.62}$$

whose Lie algebra $\mathfrak{su}(3)$ is the space of traceless anti-Hermitian 3×3 matrices.

Proposition 14.50 (Dimension and rank)**.** $\dim \mathfrak{su}(3) = 8$ *and the rank of* $\mathfrak{su}(3)$ *in the sense of Definition 14.20 is* $\ell = 2$*.*

Rests on Definitions 14.20 and 14.49.

Proof. *Derives Proposition 14.50.*

Linearizing Equation (14.62) as in Proposition 14.35 gives $X^{\dagger} + X = 0$ and $\operatorname{tr} X = 0$. The anti-Hermitian 3×3 matrices form a real vector space of dimension 9 — three real diagonal entries times i, and three complex entries above the diagonal — and the trace condition removes one, leaving 8. For the rank, a maximal abelian subalgebra consisting of diagonalizable elements may be taken to be the traceless diagonal anti-Hermitian matrices, of which there are two independent, and no larger commuting set of diagonalizable elements exists because a set of commuting Hermitian matrices is simultaneously diagonalizable and the diagonal traceless ones already exhaust that possibility. □

Definition 14.51 (Gell-Mann basis)**.** The standard basis of Hermitian generators is $T_a = \frac{1}{2}\lambda_a$, $a = 1, \dots, 8$, with [1205, 2427]

$$\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_2 = \begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{14.63}$$

$$\lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \lambda_5 = \begin{pmatrix} 0 & 0 & -\mathrm{i} \\ 0 & 0 & 0 \\ \mathrm{i} & 0 & 0 \end{pmatrix}, \quad \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \tag{14.64}$$

$$\lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathrm{i} \\ 0 & \mathrm{i} & 0 \end{pmatrix}, \quad \lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}. \tag{14.65}$$

They are Hermitian and traceless and satisfy the normalisation

$$\operatorname{tr}(\lambda_a \lambda_b) = 2\delta_{ab}, \qquad \text{i.e.} \qquad \operatorname{tr}(T_a T_b) = \tfrac{1}{2}\delta_{ab}, \tag{14.66}$$

the same normalisation as Equation (14.44).

Rests on Definition 14.49.

Proposition 14.52 (Structure constants and the symmetric tensor)**.** *There are real arrays f_{abc}, totally antisymmetric, and d_{abc}, totally symmetric, with*

$$\boxed{[T_a, T_b] = \mathrm{i} f_{abc} T_c\,, \qquad \{T_a, T_b\} = \tfrac{1}{3}\delta_{ab}\mathbb{1} + d_{abc} T_c}\,, \tag{14.67}$$

recovered from the generators by

$$f_{abc} = -2\mathrm{i}\ \mathrm{tr}\left([T_a, T_b]\, T_c\right)\,, \qquad d_{abc} = 2\ \mathrm{tr}\left(\{T_a, T_b\}\, T_c\right)\,. \tag{14.68}$$

Their independent nonvanishing values are

$$f_{123} = 1\,, \quad f_{147} = f_{246} = f_{257} = f_{345} = \tfrac{1}{2}\,, \quad f_{156} = f_{367} = -\tfrac{1}{2}\,, \quad f_{458} = f_{678} = \tfrac{\sqrt{3}}{2}\,, \tag{14.69}$$

$$\begin{aligned} &d_{118} = d_{228} = d_{338} = \tfrac{1}{\sqrt{3}}\,, && d_{888} = -\tfrac{1}{\sqrt{3}}\,, \\ &d_{448} = d_{558} = d_{668} = d_{778} = -\tfrac{1}{2\sqrt{3}}\,, && d_{146} = d_{157} = d_{256} = d_{344} = d_{355} = \tfrac{1}{2}\,, \\ &d_{247} = d_{366} = d_{377} = -\tfrac{1}{2}\,. \end{aligned} \tag{14.70}$$

Rests on Definition 14.51.

Proof. *Derives Proposition 14.52.*

The product $T_a T_b$ is a 3×3 matrix, and $\{\mathbb{1}, T_1, \ldots, T_8\}$ is a basis of the complex 3×3 matrices, since the T_c span the traceless ones. Expanding $T_a T_b$ in it and separating the antisymmetric and symmetric parts in (a, b) gives Equation (14.67), the coefficient of $\mathbb{1}$ in the symmetric part being $\frac{1}{3}\,\mathrm{tr}\,\{T_a, T_b\} = \frac{1}{3}\delta_{ab}$ by Equation (14.66). Multiplying either relation by T_c and taking the trace with Equation (14.66), and using $\mathrm{tr}\, T_c = 0$, gives Equation (14.68). Total symmetry of d and total antisymmetry of f then follow from Equation (14.68) and the cyclicity of the trace: f is antisymmetric in (a, b) by construction and

$$\mathrm{tr}\left([T_a, T_b]\, T_c\right) = \mathrm{tr}\left(T_a T_b T_c\right) - \mathrm{tr}\left(T_b T_a T_c\right) = \mathrm{tr}\left(T_a\, [T_b, T_c]\right)$$

is invariant under cyclic permutation of (a, b, c), which together with antisymmetry in the first pair gives total antisymmetry; the same argument with the anticommutator gives total symmetry of d. The tables Equations (14.69) and (14.70) are then a finite computation from Equations (14.63) to (14.65); two entries are done here as specimens. For f_{458}: a direct multiplication gives $[\lambda_4, \lambda_5] = \mathrm{diag}(2\mathrm{i}, 0, -2\mathrm{i})$, so

$$f_{458} = -\frac{\mathrm{i}}{4}\,\mathrm{tr}\left([\lambda_4, \lambda_5]\,\lambda_8\right) = -\frac{\mathrm{i}}{4}\cdot\frac{1}{\sqrt{3}}\,(2\mathrm{i} + 0 + 4\mathrm{i}) = \frac{\sqrt{3}}{2}\,.$$

For d_{247}: $\{\lambda_2, \lambda_4\} = -\lambda_7$ by direct multiplication, so $d_{247} = \frac{1}{4}\,\mathrm{tr}\,(-\lambda_7\lambda_7) = -\frac{1}{4}\cdot 2 = -\frac{1}{2}$. □

The remaining entries are three-matrix traces of the same kind: nine independent nonvanishing components of f and sixteen of d, twenty-five in all, every other nonzero component being fixed by the total antisymmetry of f and the total symmetry of d. They are computed and tabulated in full in the appendix.

Full derivation: Appendix A.34 (p. 2951).

Derives Proposition 14.52.

Proposition 14.53 (Killing form of $\mathfrak{su}(3)$)**.** *In the Hermitian basis of Definition 14.51,*

$$\kappa_{ab} = f_{acd} f_{bcd} = 3\,\delta_{ab}\,, \tag{14.71}$$

which is nondegenerate — as Theorem 14.13 demands of $\mathfrak{su}(3)$, which is simple and hence semisimple — so that $\kappa^{ab} = \frac{1}{3}\delta^{ab}$.

Rests on Proposition 14.52, Definition 14.51, Equation (14.15), and Theorem 14.13.

Proof. *Derives Proposition 14.53.*

With $C^c{}_{ab} = \mathrm{i} f_{abc}$ from Equation (14.67), Equation (14.15) gives $\kappa_{ab} = (\mathrm{i} f_{adc})(\mathrm{i} f_{bcd}) = -f_{adc} f_{bcd} = f_{adc} f_{bdc} = f_{acd} f_{bcd}$ after renaming.

The array is evaluated here directly. The shorter route — both κ_{ab} and $\operatorname{tr}(T_a T_b)$ are symmetric bilinear forms invariant under the adjoint action, and on a simple algebra such a form is unique up to a scale, so it is enough to evaluate one entry — would have to obtain that uniqueness from Schur's lemma applied to the adjoint representation, and Theorem 5.153 is stated for a representation on a *complex* space, whereas the adjoint representation of $\mathfrak{su}(3)$ acts on the eight-dimensional *real* algebra. Over $\mathbb{R}$ the conclusion of that lemma is false in general, so its hypotheses would first have to be recovered by complexifying and by quoting the irreducibility of the complexified adjoint — which is no cheaper than the computation that follows and rests on the same quoted simplicity.

Because $[T_a, T_c] = \mathrm{i} f_{acd} T_d$ and $\operatorname{tr}(T_d T_e) = \frac{1}{2}\delta_{de}$ by Equation (14.66),

$$\sum_c \operatorname{tr}\left([T_a, T_c]\,[T_b, T_c]\right) = -f_{acd} f_{bce}\ \operatorname{tr}\left(T_d T_e\right) = -\tfrac{1}{2} f_{acd} f_{bcd} = -\tfrac{1}{2}\kappa_{ab}\,. \tag{14.72}$$

Expanding the two commutators gives four traces. Summed over c they collapse to two, by cyclicity of the trace together with $\sum_c T_c T_c \propto \mathbb{1}$, which Equation (14.74) below establishes:

$$\sum_c \operatorname{tr}\left([T_a, T_c]\,[T_b, T_c]\right) = 2\sum_c \operatorname{tr}\left(T_a T_c T_b T_c\right) - 2\sum_c \operatorname{tr}\left(T_a T_b T_c T_c\right)\,. \tag{14.73}$$

Both sums follow from the completeness of the basis $\{\mathbb{1}, T_1, \dots, T_8\}$ of the complex 3×3 matrices already used in Proposition 14.52. Expanding an arbitrary M in that basis, with the coefficients read off by Equation (14.66) and $\operatorname{tr} T_a = 0$, gives $M = \frac{1}{3}(\operatorname{tr} M)\,\mathbb{1} + 2\,\operatorname{tr}(M T_a)\,T_a$, and taking for M the matrix whose only nonzero entry is a 1 in position (j, k) turns that into

$$\boxed{\sum_{a=1}^{8} (T_a)_{il}\,(T_a)_{kj} = \tfrac{1}{2}\left(\delta_{ij}\delta_{lk} - \tfrac{1}{3}\delta_{il}\delta_{jk}\right)}\,. \tag{14.74}$$

Setting $l = k$ and summing gives $\sum_c T_c T_c = \frac{1}{2}\left(3 - \frac{1}{3}\right)\mathbb{1} = \frac{4}{3}\mathbb{1}$, so that $\sum_c \operatorname{tr}(T_a T_b T_c T_c) = \frac{4}{3}\operatorname{tr}(T_a T_b) = \frac{2}{3}\delta_{ab}$; and contracting Equation (14.74) into $\operatorname{tr}(T_a T_c T_b T_c) = (T_a)_{ij}\,(T_c)_{jk}\,(T_b)_{kl}\,(T_c)_{li}$ gives

$$\sum_c \operatorname{tr}\left(T_a T_c T_b T_c\right) = \tfrac{1}{2}\left(\operatorname{tr} T_a\ \operatorname{tr} T_b - \tfrac{1}{3}\operatorname{tr}\left(T_a T_b\right)\right) = -\tfrac{1}{12}\delta_{ab}\,.$$

Substituting both into Equation (14.73),

$$\sum_c \operatorname{tr}\left([T_a, T_c]\,[T_b, T_c]\right) = -\tfrac{1}{6}\delta_{ab} - \tfrac{4}{3}\delta_{ab} = -\tfrac{3}{2}\delta_{ab}\,,$$

and Equation (14.72) gives $\kappa_{ab} = 3\delta_{ab}$.

As a check on one entry, take $a = b = 3$ and read the nonzero f_{3cd} from Equation (14.69) — $f_{312} = 1$, $f_{345} = \frac{1}{2}$, $f_{367} = -\frac{1}{2}$, each contributing twice because both index orders are summed:

$$\kappa_{33} = f_{3cd} f_{3cd} = 2\,(1)^2 + 2\left(\tfrac{1}{2}\right)^2 + 2\left(\tfrac{1}{2}\right)^2 = 3\,.$$

□

The simplicity of $\mathfrak{su}(3)$ is used below, where the adjoint representation is asserted to be irreducible — that is, where the traceless part of the endomorphisms of the fundamental is identified with the eight-dimensional representation. The argument decomposes a putative ideal into eigenspaces of the two-dimensional Cartan subalgebra and then uses the fact that

the six roots of the algebra — the hexagon exhibited below — are connected, every root being reachable from every other by adding roots, so that a single root vector in the ideal drags the whole algebra in with it.

Full derivation: Appendix A.35 (p. 2957).

Derives Theorem 14.57.

Proposition 14.54 (The two Casimir operators of $\mathfrak{su}(3)$)**.**

$$\boxed{C_2 = \sum_{a=1}^{8} T_a T_a \,, \qquad C_3 = d_{abc}\, T_a T_b T_c} \tag{14.75}$$

are Casimir elements of $U(\mathfrak{su}(3))$, and by Theorem 14.21 together with Proposition 14.50 they generate all of them: $\ell = 2$, so there are exactly two.

Rests on Corollary 14.17, Theorems 14.16, 14.21 and 14.110, and Propositions 14.50, 14.52 and 14.53.

Proof. *Derives Proposition 14.54.*

For C_2 this is Corollary 14.17 with $\kappa^{ab} = \frac{1}{3}\delta^{ab}$ from Equation (14.71), up to the harmless factor of 3. For C_3, apply Theorem 14.16 with $r = 3$ and $k^{abc} = d_{abc}$: the tensor is totally symmetric by Proposition 14.52, and its invariance in the sense of Equation (14.18) is Equation (14.148) for the G-invariant polynomial $\langle T_a, T_b, T_c\rangle = \frac{1}{2} d_{abc}$ — symmetric and invariant because $\operatorname{tr}(\{T_a, T_b\} T_c)$ is a trace of a product of generators and therefore unchanged under $T_a \mapsto g T_a g^{-1}$, which is Definition 14.109. Written out, the invariance reads

$$f_{ead}\, d_{dbc} + f_{ebd}\, d_{adc} + f_{ecd}\, d_{abd} = 0 \,, \tag{14.76}$$

and it is exactly this identity that the proof of Theorem 14.16 consumes. □

Proposition 14.55 (The fundamental, the antifundamental and the adjoint)**.** SU(3) *has three representations of immediate use.*

1. *The* fundamental **3***: the defining action on $\mathbb{C}^3$, generated by the T_a themselves. Its quadratic Casimir is*
$$C_2(\mathbf{3}) = \tfrac{4}{3} \,. \tag{14.77}$$
2. *The* antifundamental $\bar{\mathbf{3}}$*: the action on the complex-conjugate space, generated by $-T_a^*$. It has the same C_2 and is* not *equivalent to* **3***.*
3. *The* adjoint **8***: the action on $\mathfrak{su}(3)$ itself, with generators $\left(T_a^{\mathrm{ad}}\right)_{bc} = -\mathrm{i} f_{abc}$, of dimension 8 and quadratic Casimir*
$$C_2(\mathbf{8}) = 3 \,. \tag{14.78}$$

Rests on Definition 14.51, Corollary 14.18, and Proposition 14.53.

Proof. *Derives Proposition 14.55.*

For Equation (14.77): C_2 is central, so by Corollary 14.18 it acts as a number c on the irreducible **3**; taking the trace, $3c = \sum_a \operatorname{tr}(T_a T_a) = 8 \cdot \frac{1}{2} = 4$ by Equation (14.66), so $c = 4/3$. For Equation (14.78): the same argument with $\operatorname{tr}\left(T_a^{\mathrm{ad}} T_b^{\mathrm{ad}}\right) = -f_{acd} f_{bdc} = f_{acd} f_{bcd} = 3\delta_{ab}$ from Equation (14.71) gives $8c = \sum_a 3 = 24$, so $c = 3$.

That $\bar{\mathbf{3}}$ is a representation is the observation that complex-conjugating $[T_a, T_b] = \mathrm{i} f_{abc} T_c$ and multiplying by -1 reproduces the same brackets. That it is inequivalent to **3** is read off the weights. Take the Cartan subalgebra spanned by T_3 and T_8, which are diagonal; the *weights* of a representation are the pairs of simultaneous eigenvalues of (T_3, T_8), and being

eigenvalues they are unchanged by any equivalence. For $\mathbf{3}$ they are, from Equations (14.63) and (14.65),

$$\left(\tfrac{1}{2}, \tfrac{1}{2\sqrt{3}}\right), \quad \left(-\tfrac{1}{2}, \tfrac{1}{2\sqrt{3}}\right), \quad \left(0, -\tfrac{1}{\sqrt{3}}\right), \tag{14.79}$$

summing to zero as tracelessness demands, and for $\bar{\mathbf{3}}$ they are the three negatives of these. The two sets are different — an upward-pointing triangle and a downward-pointing one — so no equivalence can carry one representation to the other. □

Remark 14.56 (Roots, and why $\mathbf{3}$ and $\bar{\mathbf{3}}$ differ here but not for SU(2)). The *roots* are the weights of the adjoint representation, that is the eigenvalues of ad_{T_3} and ad_{T_8}. They are the differences of the weights Equation (14.79),

$$\pm(1,0), \qquad \pm\left(\tfrac{1}{2}, \tfrac{\sqrt{3}}{2}\right), \qquad \pm\left(-\tfrac{1}{2}, \tfrac{\sqrt{3}}{2}\right), \tag{14.80}$$

six vectors of unit length at 60 degrees to one another, forming a regular hexagon, together with two zeros for the Cartan directions themselves: $6+2=8$ weights for an eight-dimensional representation, as required. For $\mathfrak{su}(2)$ the rank is one, the two roots are ± 1, and the weight set of any representation is symmetric about zero, which is why there the fundamental and its conjugate *are* equivalent. The asymmetry of the SU(3) weight diagram under $w \mapsto -w$ is precisely the statement that its fundamental is complex.

Theorem 14.57 ($\mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{1} \oplus \mathbf{8}$).

$$\boxed{\mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{1} \oplus \mathbf{8}}, \qquad 3 \times 3 = 1 + 8\,. \tag{14.81}$$

Rests on Proposition 14.55.

Proof. *Derives Theorem 14.57.*

Identify $\mathbb{C}^3 \otimes (\mathbb{C}^3)^*$ with the space of complex 3×3 matrices $\mathrm{End}(\mathbb{C}^3)$; the action induced by $\mathbf{3}$ on the first factor and $\bar{\mathbf{3}}$ on the second is $M \mapsto UMU^\dagger$. The trace is invariant under it, so $\mathrm{End}(\mathbb{C}^3)$ splits as an SU(3)-invariant direct sum

$$\mathrm{End}\left(\mathbb{C}^3\right) = \mathbb{C}\,\mathbb{1} \;\oplus\; \{M \mid \mathrm{tr}\, M = 0\}\,,$$

of dimensions 1 and 8. The first summand is the trivial representation $\mathbf{1}$. The second is the complexification of $\mathfrak{su}(3)$ — a traceless complex matrix splits uniquely into Hermitian and anti-Hermitian traceless parts — carrying the action $M \mapsto UMU^\dagger$, which is the adjoint action; it is irreducible because an invariant subspace would be an ideal of a simple Lie algebra. Hence it is $\mathbf{8}$.

The weights confirm the count independently: the weights of the tensor product are the sums of a weight of $\mathbf{3}$ and a weight of $\bar{\mathbf{3}}$, that is all differences $w_i - w_j$ of the three weights Equation (14.79). The three coincident cases $i = j$ give the weight zero three times, and the six cases $i \neq j$ give exactly the six roots Equation (14.80). That is one zero for $\mathbf{1}$ plus the six roots and two zeros of $\mathbf{8}$, and nothing left over. □

Remark 14.58 (What is deliberately not said here). This subsection has supplied group theory and nothing else. Which physical states, if any, carry the $\mathbf{3}$; whether the $\mathbf{8}$ is realized by anything observable; what the eight generators have to do with the strong interaction — none of that is a mathematical question, and all of it is settled, where it can be settled, by measurement in Chapters 102 and 104. The historical route ran the other way: the group was proposed as an organizing symmetry of the observed hadron spectrum [1205, 1216] before the mathematics above was put to the use it now has [1907]. The order in this treatise is the opposite one on purpose — the mathematics is established first and then confronted with evidence, so that agreement counts for something.

14.3 The Lie algebra associated with a Lie group

14.3.1 Translations in D dimensions

The algebra of the group of translations in D dimensions, which we denote by $\mathfrak{t}(D)$, is generated by

$$\boxed{P_A = \partial_A\,, \quad A = 1, \dots, D} \tag{14.82}$$

so that

$$\boxed{\dim \mathfrak{t}(D) = D}\,. \tag{14.83}$$

Let f be a function of class C^2 on $\mathbb{R}^D$. Then

$$\begin{aligned} [P_A, P_B]\, f &= [\partial_A, \partial_B]\, f \\ &= \partial_A \partial_B f - \partial_B \partial_A f \\ (f \text{ of class } C^2) \quad &= 0\,, \end{aligned}$$

whence

$$\boxed{[P_A, P_B] = 0}\,. \tag{14.84}$$

14.3.2 The algebra $\mathfrak{so}(p, q)$

Generators and commutation relations

Consider $\mathbb{R}^{p+q}$ endowed with the flat metric η_{AB} of signature (p, q) of Notation 14.1, and in it the unit quadric (a pseudo-sphere) defined by the equation

$$\left(x^1\right)^2 + \cdots + \left(x^p\right)^2 - \left(x^{p+1}\right)^2 - \cdots - \left(x^{p+q}\right)^2 = 1\,. \tag{14.85}$$

This space carries the Killing vector fields (cf. Chapter 13 and [1746])

$$\boxed{J_{AB} = x_A \partial_B - x_B \partial_A\,, \quad A, B = 1, \dots, p+q}\,. \tag{14.86}$$

That the fields Equation (14.86) are Killing fields of η_{AB}, and that they generate the isometries of the quadric Equation (14.85), is asserted by the source without proof. The source introduces them through the quadric; they are in fact a statement about the flat space itself, and are proved here in that form, the quadric following as a corollary.

Proposition 14.59 (The Killing fields of a flat pseudo-Euclidean space)**.** *Let $\mathbb{R}^{p+q}$ carry the flat metric η_{AB} of Notation 14.1, $D = p + q$. Then:*

1. *every J_{AB} of Equation (14.86) is a Killing field;*
2. *conversely, every Killing field of η_{AB} is a real linear combination of the D translations P_A of Equation (14.82) and the $D(D-1)/2$ fields J_{AB} with $A < B$, and those $D(D+1)/2$ fields are linearly independent. Flat space is therefore maximally symmetric in the sense of Proposition 13.140, and its Killing algebra is the inhomogeneous algebra $\mathfrak{t}(D) \mathbin{\bar{\oplus}} \mathfrak{so}(p, q)$, which for the Lorentzian signature $q = 1$ is the Poincaré algebra Equation (14.99);*
3. *J_{AB} annihilates the quadratic form $x^C x_C$, so its flow preserves each quadric $x^C x_C = \text{const}$, in particular Equation (14.85), and restricts on it to a Killing field of the induced metric.*

Rests on Equation (14.86), Proposition 13.140, and Definition 13.139.

Proof. *Derives Proposition 14.59.*

In the Cartesian coordinates x^A the metric components are constant, so the Christoffel symbols vanish and Killing's equation Equation (13.290) reads

$$\partial_C \xi_D + \partial_D \xi_C = 0\,. \tag{14.87}$$

(i) The field J_{AB} has components $\xi^C = x_A \delta^C_B - x_B \delta^C_A$, hence $\xi_D = x_A \eta_{BD} - x_B \eta_{AD}$, and with $\partial_C x_A = \eta_{AC}$,

$$\partial_C \xi_D = \eta_{AC}\eta_{BD} - \eta_{BC}\eta_{AD}\,,$$

which is antisymmetric under $C \leftrightarrow D$. So Equation (14.87) holds. That the translations satisfy it is immediate, their components being constant.

(ii) Let ξ satisfy Equation (14.87) and set $T_{CDE} = \partial_C \partial_D \xi_E$. It is symmetric in its first two indices, because partial derivatives commute, and antisymmetric in its last two by Equation (14.87). Those two symmetries are incompatible unless T vanishes:

$$T_{CDE} = T_{DCE} = -T_{DEC} = -T_{EDC} = T_{ECD} = T_{CED} = -T_{CDE}\,,$$

each step being one of the two symmetries applied once. Hence $\partial_C \partial_D \xi_E = 0$, so each ξ_E is a polynomial of degree at most one,

$$\xi_E = a_E + \omega_{EC}\, x^C\,, \qquad a_E = \xi_E(0)\,, \quad \omega_{EC} = \partial_C \xi_E\,, \tag{14.88}$$

with constant coefficients, and Equation (14.87) says exactly that $\omega_{EC} = -\omega_{CE}$. Comparing with the components computed in (i), the constant part is $a^E P_E$ and the linear part is $-\frac{1}{2}\omega^{AB} J_{AB}$, a combination of the J_{AB}. Conversely ξ determines a_E and ω_{EC} through Equation (14.88), so distinct pairs (a, ω) give distinct fields and the $D + D(D-1)/2 = D(D+1)/2$ listed fields are independent. That number is the maximum allowed by Proposition 13.140. The brackets of these fields are Equations (14.84), (14.90) and (14.97), computed below; they are those of the semidirect sum $\mathfrak{t}(D) \mathbin{\bar{\oplus}} \mathfrak{so}(p,q)$, which at $q = 1$ is Equation (14.99).

(iii) Directly,

$$J_{AB}\left(x^C x_C\right) = (x_A \partial_B - x_B \partial_A)\, x^C x_C = 2 x_A x_B - 2 x_B x_A = 0\,,$$

so J_{AB} is tangent to every level set of $x^C x_C$ and its flow maps each such quadric to itself. That flow consists of isometries of η by (i), and an isometry of the ambient metric preserving a submanifold restricts to an isometry of the induced metric; by Definition 13.139 the restriction of J_{AB} is therefore a Killing field of the quadric. □

Clearly $J_{AB} = -J_{BA}$, and therefore

$$\boxed{\dim \mathfrak{so}(p,q) = \frac{(p+q)(p+q-1)}{2} = \frac{D(D-1)}{2}}\,. \tag{14.89}$$

Let us compute the commutation relations among the symmetry generators J_{AB}. Let f be a function of class C^2 on $\mathbb{R}^{p+q}$. We have

$$\begin{aligned}
[J_{AB}, J_{CD}]\, f &= [x_A\partial_B - x_B\partial_A, x_C\partial_D - x_D\partial_C]\, f \\
&= [x_A\partial_B, x_C\partial_D]\, f - [x_A\partial_B, x_D\partial_C]\, f - [x_B\partial_A, x_C\partial_D]\, f + [x_B\partial_A, x_D\partial_C]\, f\,,
\end{aligned}$$

where

$$\begin{aligned}
[x_A\partial_B, x_C\partial_D]\, f &= x_A\partial_B\,(x_C\partial_D f) - x_C\partial_D\,(x_A\partial_B f) \\
&= x_A\eta_{BC}\partial_D f + x_A x_C \partial_B\partial_D f - x_C\eta_{DA}\partial_B f - x_C x_A \partial_D\partial_B f \\
(f \text{ of class } C^2) \quad &= \eta_{BC}\, x_A \partial_D f - \eta_{DA}\, x_C \partial_B f\,,
\end{aligned}$$

and where we have used that $\partial_A x_B = \eta_{BC}\,\partial_A x^C = \eta_{BC}\,\delta^C_A = \eta_{AB}$. In this manner we obtain

$$\begin{aligned}[J_{AB}, J_{CD}] &= \eta_{BC}\,x_A\partial_D - \eta_{DA}\,x_C\partial_B - \eta_{BD}\,x_A\partial_C + \eta_{CA}\,x_D\partial_B \\ &\quad - \eta_{AC}\,x_B\partial_D + \eta_{DB}\,x_C\partial_A + \eta_{AD}\,x_B\partial_C - \eta_{CB}\,x_D\partial_A \\ &= \eta_{AC}\,(x_D\partial_B - x_B\partial_D) + \eta_{BD}\,(x_C\partial_A - x_A\partial_C) \\ &\quad + \eta_{AD}\,(x_B\partial_C - x_C\partial_B) + \eta_{BC}\,(x_A\partial_D - x_D\partial_A)\ ,\end{aligned}$$

so the commutation relations of the algebra $\mathfrak{so}(p,q)$ are given by

$$\boxed{[J_{AB}, J_{CD}] = \eta_{AC}J_{DB} + \eta_{BD}J_{CA} + \eta_{AD}J_{BC} + \eta_{BC}J_{AD}}\,. \tag{14.90}$$

Structure constants

To express the commutation relations Equation (14.90) in the usual form $[J_{AB}, J_{CD}] = f_{AB,CD}{}^{EF}J_{EF}$, the symmetries of the indices must be taken into account. A mistake one may commit is to simply attach the corresponding Kronecker deltas, writing

$$f_{AB,CD}{}^{EF} = \eta_{AC}\delta^E_D\delta^F_B + \eta_{BD}\delta^E_C\delta^F_A + \eta_{AD}\delta^E_B\delta^F_C + \eta_{BC}\delta^E_A\delta^F_D\,, \tag{14.91}$$

or

$$f_{AB,CD}{}^{EF} = \eta_{AD}\delta^E_B\delta^F_C + \eta_{BC}\delta^E_A\delta^F_D - \eta_{AC}\delta^E_B\delta^F_D - \eta_{BD}\delta^E_A\delta^F_C\,. \tag{14.92}$$

However, upon substituting $J_{AB} = -J_{BA}$ into $[J_{AB}, J_{CD}] = f_{AB,CD}{}^{EF}J_{EF}$ it is necessary that $f_{AB,CD}{}^{EF} = -f_{BA,CD}{}^{EF}$, which is clearly satisfied neither by the quantities Equation (14.91) nor by Equation (14.92). In the same way, one must have $f_{AB,CD}{}^{EF} = -f_{AB,DC}{}^{EF}$. The only thing that can be said about the upper indices of f is that f must not be symmetric under their exchange, since the contraction of symmetric upper indices with the antisymmetric indices of J_{EF} would force $[J_{AB}, J_{CD}] = 0$; moreover, this naive choice of f does not satisfy the antisymmetry in the lower indices either. Let us see what happens if we antisymmetrize in the indices E and F:

$$\begin{aligned}[J_{AB}, J_{CD}] &= \eta_{AC}J_{DB} + \eta_{BD}J_{CA} + \eta_{AD}J_{BC} + \eta_{BC}J_{AD} \\ &= \eta_{AC}\delta^E_D\delta^F_B J_{EF} + \eta_{BD}\delta^E_C\delta^F_A J_{EF} + \eta_{AD}\delta^E_B\delta^F_C J_{EF} + \eta_{BC}\delta^E_A\delta^F_D J_{EF} \\ &= \frac{1}{2}\eta_{AC}\left(\delta^E_D\delta^F_B + \delta^E_D\delta^F_B\right)J_{EF} + \frac{1}{2}\eta_{BD}\left(\delta^E_C\delta^F_A + \delta^E_C\delta^F_A\right)J_{EF} \\ &\quad + \frac{1}{2}\eta_{AD}\left(\delta^E_B\delta^F_C + \delta^E_B\delta^F_C\right)J_{EF} + \frac{1}{2}\eta_{BC}\left(\delta^E_A\delta^F_D + \delta^E_A\delta^F_D\right)J_{EF}\,,\end{aligned}$$

but, using the antisymmetry of J_{EF} and relabelling the dummy indices in the second copy of each term,

$$\begin{aligned}\left(\delta^E_D\delta^F_B + \delta^E_D\delta^F_B\right)J_{EF} &= \delta^E_D\delta^F_B J_{EF} + \delta^E_D\delta^F_B J_{EF} \\ &= \delta^E_D\delta^F_B J_{EF} - \delta^F_B\delta^E_D J_{FE} \\ &= \delta^E_D\delta^F_B J_{EF} - \delta^E_B\delta^F_D J_{EF} \\ &= \left(\delta^E_D\delta^F_B - \delta^E_B\delta^F_D\right)J_{EF}\,,\end{aligned}$$

whence

$$\begin{aligned}[J_{AB}, J_{CD}] = \frac{1}{2}\Big\{&\eta_{AC}\left(\delta^E_D\delta^F_B - \delta^E_B\delta^F_D\right) + \eta_{BD}\left(\delta^E_C\delta^F_A - \delta^E_A\delta^F_C\right) \\ &+ \eta_{AD}\left(\delta^E_B\delta^F_C - \delta^E_C\delta^F_B\right) + \eta_{BC}\left(\delta^E_A\delta^F_D - \delta^E_D\delta^F_A\right)\Big\}J_{EF}\,.\end{aligned}$$

In this case we have

$$f_{AB,CD}{}^{EF} = \frac{1}{2}\Big\{\eta_{AC}\left(\delta^E_D\delta^F_B - \delta^E_B\delta^F_D\right) + \eta_{BD}\left(\delta^E_C\delta^F_A - \delta^E_A\delta^F_C\right) + \eta_{AD}\left(\delta^E_B\delta^F_C - \delta^E_C\delta^F_B\right) + \eta_{BC}\left(\delta^E_A\delta^F_D - \delta^E_D\delta^F_A\right)\Big\},$$

for which we must verify the aforementioned symmetries. In the first place,

$$\begin{aligned} f_{CD,AB}{}^{EF} &= \frac{1}{2}\Big\{\eta_{CA}\left(\delta^E_B\delta^F_D - \delta^E_D\delta^F_B\right) + \eta_{DB}\left(\delta^E_A\delta^F_C - \delta^E_C\delta^F_A\right) \\ &\quad + \eta_{CB}\left(\delta^E_D\delta^F_A - \delta^E_A\delta^F_D\right) + \eta_{DA}\left(\delta^E_C\delta^F_B - \delta^E_B\delta^F_C\right)\Big\} \\ &= \frac{1}{2}\Big\{\eta_{AC}\left(\delta^E_B\delta^F_D - \delta^E_D\delta^F_B\right) + \eta_{BD}\left(\delta^E_A\delta^F_C - \delta^E_C\delta^F_A\right) \\ &\quad + \eta_{BC}\left(\delta^E_D\delta^F_A - \delta^E_A\delta^F_D\right) + \eta_{AD}\left(\delta^E_C\delta^F_B - \delta^E_B\delta^F_C\right)\Big\} \\ &= -f_{AB,CD}{}^{EF}, \end{aligned}$$

so we do have the antisymmetry under exchange of the index pairs. Moreover,

$$\begin{aligned} f_{BA,CD}{}^{EF} &= \frac{1}{2}\Big\{\eta_{BC}\left(\delta^E_D\delta^F_A - \delta^E_A\delta^F_D\right) + \eta_{AD}\left(\delta^E_C\delta^F_B - \delta^E_B\delta^F_C\right) \\ &\quad + \eta_{BD}\left(\delta^E_A\delta^F_C - \delta^E_C\delta^F_A\right) + \eta_{AC}\left(\delta^E_B\delta^F_D - \delta^E_D\delta^F_B\right)\Big\} \\ &= -f_{AB,CD}{}^{EF}, \end{aligned}$$

so that $f_{AB,DC}{}^{EF} = -f_{DC,AB}{}^{EF} = f_{CD,AB}{}^{EF} = -f_{AB,CD}{}^{EF}$; and additionally

$$\begin{aligned} f_{AB,CD}{}^{FE} &= \frac{1}{2}\Big\{\eta_{AC}\left(\delta^F_D\delta^E_B - \delta^F_B\delta^E_D\right) + \eta_{BD}\left(\delta^F_C\delta^E_A - \delta^F_A\delta^E_C\right) \\ &\quad + \eta_{AD}\left(\delta^F_B\delta^E_C - \delta^F_C\delta^E_B\right) + \eta_{BC}\left(\delta^F_A\delta^E_D - \delta^F_D\delta^E_A\right)\Big\} \\ &= \frac{1}{2}\Big\{\eta_{AC}\left(\delta^E_B\delta^F_D - \delta^E_D\delta^F_B\right) + \eta_{BD}\left(\delta^E_A\delta^F_C - \delta^E_C\delta^F_A\right) \\ &\quad + \eta_{AD}\left(\delta^E_C\delta^F_B - \delta^E_B\delta^F_C\right) + \eta_{BC}\left(\delta^E_D\delta^F_A - \delta^E_A\delta^F_D\right)\Big\} \\ &= -f_{AB,CD}{}^{EF}. \end{aligned}$$

In summary, the commutation relations of the algebra $\mathfrak{so}(p,q)$ are given by Equation (14.90) or, equivalently, by

$$[J_{AB}, J_{CD}] = f_{AB,CD}{}^{EF} J_{EF}, \tag{14.93}$$

where the $f_{AB,CD}{}^{EF}$ are the structure constants, given by

$$f_{AB,CD}{}^{EF} = \frac{1}{2}\Big\{\eta_{AC}\left(\delta^E_D\delta^F_B - \delta^E_B\delta^F_D\right) + \eta_{BD}\left(\delta^E_C\delta^F_A - \delta^E_A\delta^F_C\right) + \eta_{AD}\left(\delta^E_B\delta^F_C - \delta^E_C\delta^F_B\right) + \eta_{BC}\left(\delta^E_A\delta^F_D - \delta^E_D\delta^F_A\right)\Big\}, \tag{14.94}$$

and which satisfy the following antisymmetry relations:

$$f_{AB,CD}{}^{EF} = -f_{CD,AB}{}^{EF} = -f_{BA,CD}{}^{EF} = -f_{AB,DC}{}^{EF} = -f_{AB,CD}{}^{FE}. \tag{14.95}$$

The Lorentz algebra in D dimensions

The Lorentz algebra in D spacetime dimensions arises as the case of Lorentzian signature, $\mathfrak{so}(D-1,1)$, of the family $\mathfrak{so}(p,q)$. At the observed $D=4$ it is the algebra $\mathfrak{so}(3,1)$ of [1746], whose treatment is four-dimensional throughout; the general-D statement here is the mathematics of Section 14.3.2 and is not taken from that source. From Equation (14.89),

$$\dim \mathfrak{so}(D-1,1) = \frac{D(D-1)}{2}\,. \tag{14.96}$$

From Equation (14.89) (Lorentzian signature, $p=D-1$ and $q=1$).

The Poincaré algebra in D dimensions

The Poincaré algebra is the algebra generated by the translations P_A together with the pseudo-rotations J_{AB}. Let us compute the commutation relations between J_{AB} and P_C. Let f be a function of class C^2 on $\mathbb{R}^D$. We have

$$\begin{aligned}
[J_{AB},P_C]\,f &= [x_A\partial_B - x_B\partial_A,\partial_C]\,f\\
&= [x_A\partial_B,\partial_C]\,f - [x_B\partial_A,\partial_C]\,f\\
&= [x_A,\partial_C]\,\partial_B f + x_A\,[\partial_B,\partial_C]\,f - [x_B,\partial_C]\,\partial_A f - x_B\,[\partial_A,\partial_C]\,f\\
(f \text{ of class } C^2)\quad &= [x_A,\partial_C]\,\partial_B f - [x_B,\partial_C]\,\partial_A f\\
&= x_A\partial_C\partial_B f - \partial_C\,(x_A\partial_B f) - x_B\partial_C\partial_A f + \partial_C\,(x_B\partial_A f)\\
&= x_A\partial_C\partial_B f - \partial_C x_A\,\partial_B f - x_A\partial_C\partial_B f - x_B\partial_C\partial_A f + \partial_C x_B\,\partial_A f + x_B\partial_C\partial_A f\\
&= \partial_C x_B\,\partial_A f - \partial_C x_A\,\partial_B f\\
&= \eta_{CB}\,\partial_A f - \eta_{CA}\,\partial_B f\,,
\end{aligned}$$

so that

$$\boxed{[J_{AB},P_C] = \eta_{BC}P_A - \eta_{AC}P_B}\,, \tag{14.97}$$

or equivalently

$$\boxed{[J_{AB},P_C] = f_{AB,C}{}^{E}\,P_E}\,, \qquad \text{with} \qquad \boxed{f_{AB,C}{}^{E} = \eta_{BC}\,\delta^E_A - \eta_{AC}\,\delta^E_B}\,. \tag{14.98}$$

Since J_{AB} and P_A do not commute, the algebra generated by $\{P_A\}\cup\{J_{AB}\}$ is the semidirect sum

$$\boxed{\mathfrak{iso}(D-1,1) = \mathfrak{t}(D)\,\bar{\oplus}\,\mathfrak{so}(D-1,1)}\,, \tag{14.99}$$

whose dimension is

$$\dim \mathfrak{iso}(D-1,1) = D + \frac{D(D-1)}{2}\,,$$

that is,

$$\dim \mathfrak{iso}(D-1,1) = \frac{D(D+1)}{2}\,. \tag{14.100}$$

From Equations (14.83) and (14.96) (add the dimensions of the two summands of the semidirect sum).

The Casimir operators of the Poincaré algebra

The Poincaré algebra is not semisimple — Equation (14.99) is a semidirect sum with the abelian ideal $\mathfrak{t}(D)$ — so Theorem 14.21 does not apply and its invariants must be found by hand (Remark 14.23). Two of them are found here. They matter out of all proportion to the difficulty of the computation, because they are what a particle *is*: an irreducible representation of this algebra is labelled by their values, and those two numbers are the mass and the spin [849].

Proposition 14.60 (The quadratic invariant)**.** $P^2 = \eta^{AB} P_A P_B$ *is a Casimir element of* $U(\mathfrak{iso}(D-1,1))$ *in every dimension* D*.*

Rests on Equations (14.84) *and* (14.97) *and Definition 14.8.*

Proof. *Derives Proposition 14.60.*

$[P^2, P_C] = 0$ is immediate from Equation (14.84). For the pseudo-rotations, use the Leibniz rule and Equation (14.97) in the form $\left[J_{AB}, P^C\right] = \delta^C_B P_A - \delta^C_A P_B$:

$$\begin{aligned}\left[J_{AB}, P_C P^C\right] &= \left[J_{AB}, P_C\right] P^C + P_C \left[J_{AB}, P^C\right] \\ &= \left(\eta_{BC} P_A - \eta_{AC} P_B\right) P^C + P_C \left(\delta^C_B P_A - \delta^C_A P_B\right) \\ &= \left(P_A P_B - P_B P_A\right) + \left(P_B P_A - P_A P_B\right) = 0\,,\end{aligned}$$

each parenthesis vanishing separately by Equation (14.84). □

The second invariant is specific to the observed four-dimensional spacetime, because it is built from the Levi-Civita symbol, which carries exactly D indices and pairs with the two-index J and the one-index P only when $D = 4$. Writing it down requires a decision about index order that the rest of the chapter had no occasion to make, and the decision is recorded before anything is computed.

Notation 14.61 (Signature and index order in this subsubsection)*.* From here to the end of this subsubsection, and here alone, $D = 4$ and the *physical* convention of Chapter 94 is adopted: the timelike direction is written x^0 and listed *first*, capital indices run from 0 to 3, and

$$\eta_{AB} = \mathrm{diag}(+1,-1,-1,-1)\,, \qquad \epsilon_{0123} = +1\,. \tag{14.101}$$

This is the *reverse* ordering of Notation 14.1, which lists the p positive directions first and therefore, at $(p,q) = (3,1)$, writes the timelike direction last, as x^4 with $\eta_{44} = -1$. The dictionary between the two is the relabelling $x^4 \leftrightarrow x^0$ of the timelike coordinate together with an overall change of sign of the metric, $\eta \mapsto -\eta$.

Nothing in the algebra changes under that dictionary. Lowering an index with $-\eta$ sends $x_A \mapsto -x_A$ and hence $J_{AB} \mapsto -J_{AB}$, while $P_A = \partial_A$ is untouched; both sides of Equation (14.90) are then even and both sides of Equation (14.97) odd, so each relation holds verbatim in either convention. What *does* change sign is every scalar built by contracting with the inverse metric. $P^2 = \eta^{AB} P_A P_B$ carries one inverse metric and flips. The vector W_A of Equation (14.103) below does *not*: it carries three raised indices, contributing $(-1)^3$, and one factor of J, contributing a fourth sign, so it comes out unchanged. But $W^2 = \eta^{AB} W_A W_B$ again carries one inverse metric, and so flips as well. The two invariants change sign *together*, so their ratio — which is what carries the spin — is the same in both conventions, and the results of this subsubsection, read in the ordering of Notation 14.1, would appear with both signs reversed: $\hat{P}^2 = -m^2c^2$ and $\hat{W}^2 = +m^2c^2\hbar^2 s(s+1)$. The convention adopted here is the one under which they agree with Theorem 94.18 and Equation (94.25), which is the form in which the rest of this treatise uses them.

Lemma 14.62 (The Levi-Civita symbol is invariant)**.** *For every antisymmetric* $\omega_{AB} = -\omega_{BA}$ *in four dimensions,*

$$\omega^E{}_A \epsilon_{EBCD} + \omega^E{}_B \epsilon_{AECD} + \omega^E{}_C \epsilon_{ABED} + \omega^E{}_D \epsilon_{ABCE} = 0\,. \tag{14.102}$$

Proof. *Derives Lemma 14.62.*

The left-hand side is totally antisymmetric in $ABCD$: exchanging any two of the four free indices swaps two of the terms with each other and reverses the sign of the remaining two through the antisymmetry of ϵ. A totally antisymmetric four-index array in four dimensions is a multiple of ϵ_{ABCD}, and contracting the left-hand side with ϵ^{ABCD} shows that the multiple is proportional to $\omega^A{}_A = \eta^{AB}\omega_{BA} = 0$, a symmetric array contracted with an antisymmetric one. □

Definition 14.63 (Pauli–Lubanski vector)**.**

$$\boxed{W_A = \tfrac{1}{2}\,\epsilon_{ABCD}\,J^{BC}P^D}\,. \tag{14.103}$$

Rests on Equations (14.82) *and* (14.86)*.*

Lemma 14.64 (Properties of W)**.** *1.* $[W_A, P_E] = 0$*;*

2. $W_A P^A = P^A W_A = 0$*;*

3. W *transforms as a vector,* $[J_{AB}, W_C] = \eta_{BC}W_A - \eta_{AC}W_B$*.*
Rests on Definition 14.63, Equations (14.84)*,* (14.90) *and* (14.97)*, and Lemma 14.62.*

Proof. *Derives Lemma 14.64.*
(i) By Equation (14.84) only the J factor contributes, and Equation (14.97) gives

$$[W_A, P_E] = \tfrac{1}{2}\epsilon_{ABCD}\left(\delta^C_E P^B - \delta^B_E P^C\right)P^D = \tfrac{1}{2}\left(\epsilon_{ABED}P^B P^D - \epsilon_{AECD}P^C P^D\right) = 0\,,$$

since in each term ϵ antisymmetrizes a pair of indices carried by the symmetric product $P\,P$ of commuting translations.

(ii) $W_A P^A = \tfrac{1}{2}\epsilon_{ABCD}J^{BC}P^D P^A$ vanishes for the same reason, the pair (A, D) now being the symmetric one; the reversed order follows from (i).

(iii) Apply the Leibniz rule to Equation (14.103) using Equation (14.90) for the J factor and Equation (14.97) for the P factor. Both brackets are of the form "η times a generator with one index replaced", that is, exactly the action of an infinitesimal pseudo-rotation on each index of the tensors J^{BC} and P^D. The three contributions are therefore the four terms of Equation (14.102) with ω the antisymmetrized index pair (AB), less the one term in which the free index C of W is rotated; by Lemma 14.62 the first four cancel and the survivor is

$$[J_{AB}, W_C] = \eta_{BC}W_A - \eta_{AC}W_B\,.$$

In words: W is a contraction of the invariant tensor ϵ with two tensors, so it is a tensor with the one uncontracted index it displays. □

Theorem 14.65 (The two Casimir operators of the Poincaré algebra in $3+1$ dimensions)**.** *In* $D = 4$*,*

$$\boxed{P^2 = P_A P^A\,, \qquad W^2 = W_A W^A} \tag{14.104}$$

commute with every generator P_C *and* J_{AB} *of* $\mathfrak{iso}(3,1)$*, and they are functionally independent.*
Rests on Proposition 14.60, Lemma 14.64, and Definition 14.63.

Proof. *Derives Theorem 14.65.*
P^2 is Proposition 14.60. For W^2: $[W^2, P_E] = 0$ follows from Lemma 14.64(i) by the Leibniz rule, and $[W^2, J_{AB}] = 0$ is the computation already done for P^2 in Proposition 14.60 with P replaced by W, which is legitimate because Lemma 14.64(iii) says W carries its index the same way P does; the square of a vector is a scalar. Independence: W^2 contains the generators J^{BC}, which P^2 does not, so neither is a function of the other. □

Remark 14.66 (What these two numbers are, in SI units). The generators above are the differential operators of Equations (14.82) and (14.86), and — unlike the rotation generators of Remark 14.24 — they do not all carry the same dimension. The pseudo-rotation generator $J_{AB} = x_A\partial_B - x_B\partial_A$ is dimensionless, being a coordinate times a derivative with respect to a coordinate; the translation generator $P_A = \partial_A$ is not, and carries the SI dimension of inverse length, of unit $1/\mathrm{m}$. In a quantum theory the observable attached to a generator is $\hbar$ times it, and the dimensions then come out right in both cases. With $P_A = \partial_A$ from Equation (14.82) the momentum operator is $\hat{P}_A = -\mathrm{i}\hbar\,P_A$, of SI unit $\mathrm{kg\,m/s}$, and with J_{AB}

from Equation (14.86) the angular-momentum operator is $\hat{J}_{AB} = -\mathrm{i}\hbar\, J_{AB}$, of unit J s; both are Hermitian because the generators are anti-Hermitian, and the normalisation is that of Equation (94.3). By Corollary 14.18 each invariant is then a number on an irreducible representation, and the numbers are

$$\hat{P}^2 = m^2c^2\,, \qquad \hat{W}^2 = -m^2c^2\,\hbar^2\, s(s+1)\,, \tag{14.105}$$

defining the *mass* m, of SI unit kg, and the *spin* s, dimensionless, with $2s \in \mathbb{N} \cup \{0\}$ by the $\mathfrak{su}(2)$ analysis of Theorem 14.43 applied to the subalgebra that leaves a timelike momentum fixed.

Note the factors, and note in particular that the $\hbar^2$ of the second is not simply one power of $\hbar$ per generator. Each generator does bring one: $\hat{P}^2 = -\hbar^2 P^2$ carries two, and $\hat{W}_A = -\hbar^2 W_A$ carries two, so that $\hat{W}^2 = \hbar^4 W^2$ carries four. What survives in the eigenvalues is the ratio,

$$\frac{\hat{W}^2}{\hat{P}^2} = -\hbar^2\, s(s+1)\,, \tag{14.106}$$

and the $\hbar^2$ left standing there is exactly the $\hbar^2$ of the angular-momentum Casimir of Remark 14.24. The mechanism is visible in the rest frame of a timelike momentum, where $\hat{P}^D = mc\,\delta^D_0$: the definition Equation (14.103) collapses to $\hat{W}_0 = 0$ and $\hat{W}_i = -mc\,\hat{\mathcal{J}}_i$, with $\hat{\mathcal{J}}_i = \frac{1}{2}\epsilon_{ijk}\hat{J}^{jk}$ the angular momentum of the subalgebra that fixes it, of unit J s, so that $\hat{W}^2 = -m^2c^2\,\hat{\boldsymbol{\mathcal{J}}}^2$ and the remaining factor is Equation (14.58) with $j = s$. One factor m^2c^2 comes from the momentum, one factor $\hbar^2 s(s+1)$ from the spin, and that is the whole content of the second line of Equation (14.105). For the same reason the first invariant is m^2c^2 and not m^2: $\hat{P}_A$ is a momentum, and the square of a momentum has the unit $\mathrm{kg^2\,m^2/s^2}$.

The classification of the representations themselves — massive, massless, the little groups, and why a massless particle has helicity rather than spin — is carried out in Chapter 94, whose Theorems 94.18 and 94.29 are Equation (14.105) in their physical setting [849, 1912]. What is established here is only the mathematical fact that the two invariants exist.

Remark 14.67 (Why this is $3+1$ and not general D). Proposition 14.60 holds for every D. The Pauli–Lubanski construction does not: Equation (14.103) contracts a D-index Levi-Civita symbol with one J and one P, which balances only when $D = 4$. In general dimension the higher invariants of $\mathfrak{iso}(D-1,1)$ are built instead from the antisymmetrized products of J with P, and how many there are is fixed by the subalgebra that leaves a timelike momentum invariant. That subalgebra is $\mathfrak{so}(D-1)$, of rank $\lfloor (D-1)/2 \rfloor$ by Lemma 14.68, and by Theorem 14.21 its representations carry exactly that many labels; together with the mass they make

$$1 + \left\lfloor \frac{D-1}{2} \right\rfloor = \left\lceil \frac{D}{2} \right\rceil \tag{14.107}$$

labels for a massive irreducible representation — two at $D = 4$, namely P^2 and W^2, and three at both $D = 5$ and $D = 6$. This is *not* $\lfloor D/2 \rfloor$: the two expressions agree in even dimension and differ in odd. This treatise instantiates the observed case, and the observed case is $D = 4$.

The algebraic half of that count is proved here; what is not proved is identified immediately after it.

Lemma 14.68 (The algebra fixing a timelike momentum)**.** *Let $D \geq 4$ and let $p \neq 0$ be a timelike vector of $\mathbb{R}^{D-1,1}$, that is $\eta(p,p) < 0$ in the ordering of Notation 14.1. The elements of $\mathfrak{so}(D-1,1)$ annihilating p form a subalgebra isomorphic to $\mathfrak{so}(D-1)$, of rank $\lfloor (D-1)/2 \rfloor$, and*

$$1 + \left\lfloor \frac{D-1}{2} \right\rfloor = \left\lceil \frac{D}{2} \right\rceil . \tag{14.108}$$

Rests on Equation (14.23) *and Proposition 14.22.*

Proof. *Derives Lemma 14.68.*

Let $\mathfrak{h} = \{X \in \mathfrak{so}(D-1,1) \mid Xp = 0\}$; it is a subalgebra, being the kernel of the linear map $X \mapsto Xp$ and closed under the bracket because $[X,Y]\,p = 0$ whenever $Xp = Yp = 0$. For $X \in \mathfrak{h}$ and any u, the defining condition Equation (14.23) gives $\eta(Xu,p) = -\eta(u,Xp) = 0$, so X maps the whole space into $p^\perp$ and kills p. Since $\eta(p,p) \neq 0$ the space splits as $\mathbb{R}p \oplus p^\perp$ with η nondegenerate on each summand, and X is determined by its restriction to $p^\perp$, which by Equation (14.23) again is an element of the pseudo-orthogonal algebra of η restricted to $p^\perp$. Conversely any such element, extended by $Xp = 0$, satisfies Equation (14.23) on all of $\mathbb{R}^{D-1,1}$: both sides vanish when either argument is p. Removing one negative direction from the signature $(D-1,1)$ leaves $\eta|_{p^\perp}$ positive definite in $D-1$ dimensions, so $\mathfrak{h} \cong \mathfrak{so}(D-1)$. Its rank is $\lfloor (D-1)/2 \rfloor$ by Proposition 14.22, applicable because $D-1 \geq 3$.

For Equation (14.108): if $D = 2k$ then $\lfloor (D-1)/2 \rfloor = k-1$ and $\lceil D/2 \rceil = k$; if $D = 2k+1$ then $\lfloor (D-1)/2 \rfloor = k$ and $\lceil D/2 \rceil = k+1$. In both cases the two sides differ by one. □

What the lemma supplies is the algebra: the little algebra and its rank. Turning a rank into a count of labels is representation theory — that a massive irreducible representation of the inhomogeneous algebra is determined by the mass together with an irreducible representation of the subalgebra fixing a timelike momentum, and that no further invariant of the enveloping algebra escapes those labels. That is the method of induced representations. Only the case $D = 4$, proved in full above, is used anywhere in this book; the general statement is derived in the appendix.

Full derivation: Appendix A.36 (p. 2960).

Derives Equation (14.107).

The de Sitter algebras in D dimensions

The anti-de Sitter (AdS) and de Sitter (dS) algebras in D spacetime dimensions are

$$\text{AdS}: \ \mathfrak{so}(D-1,2)\,, \qquad \text{dS}: \ \mathfrak{so}(D,1)\,,$$

with, by Equation (14.89),

$$\dim \mathfrak{so}(D-1,2) = \frac{D(D+1)}{2}\,, \tag{14.109}$$

$$\dim \mathfrak{so}(D,1) = \frac{D(D+1)}{2}\,. \tag{14.110}$$

The dS and AdS algebras therefore have the same dimension as the Poincaré algebra Equation (14.100). That coincidence of dimensions is not an accident, and the reason is that all three algebras are the Killing algebras of maximally symmetric D-dimensional spacetimes. The identification of the two written above is proved next; the geometry of the two spaces, and the curvature that distinguishes them, belong to Section 13.13.

Proposition 14.69 (The de Sitter algebras are isometry algebras)**.** *Let $D \geq 3$ and $\ell > 0$. In the flat space of signature $(D-1,2)$ let AdS_D be the quadric*

$$\eta_{MN} X^M X^N = -\ell^2\,, \tag{14.111}$$

and in the flat space of signature $(D,1)$ let dS_D be the quadric

$$\eta_{MN} X^M X^N = +\ell^2\,. \tag{14.112}$$

Each is a D-dimensional submanifold whose induced metric has the Lorentzian signature $(D-1,1)$, and the restrictions of the ambient fields J_{MN} of Equation (14.86) *are linearly independent Killing fields of it spanning $\mathfrak{so}(D-1,2)$ and $\mathfrak{so}(D,1)$ respectively. These exhaust the Killing fields: the two algebras are the full isometry algebras.*

Rests on Propositions 13.140 and 14.59 and Equation (14.89).

Proof. *Derives Proposition 14.69.*

Write $f(X) = \eta_{MN} X^M X^N$, so that $\mathrm{d}f = 2\eta_{MN} X^M \mathrm{d}X^N$ is nonzero at every point of either quadric; each is therefore a regular level set, a submanifold of dimension D, with tangent space at X the orthogonal complement $X^\perp$. Since $f(X) \neq 0$ there, the ambient space splits as $\mathbb{R}X \oplus X^\perp$ with η nondegenerate on each summand, so the signature of the induced metric is that of the ambient space less one direction: less a negative one for Equation (14.111) in signature $(D-1,2)$, and less a positive one for Equation (14.112) in signature $(D,1)$. Both give $(D-1,1)$.

By Proposition 14.59(iii) each J_{MN} is tangent to the quadric and restricts to a Killing field of the induced metric. The restrictions are still independent. Indeed a vanishing combination $\omega^{MN} J_{MN}$ has ambient components proportional to $\omega^M{}_N X^N$, so it vanishes on the quadric only if the matrix $\omega^M{}_N$ annihilates every point of the quadric, hence the whole span of the quadric. That span is everything. In an indefinite signature the vectors of negative norm already span: a negative basis direction has negative norm, and a positive basis direction e is the half-sum of $e + 2u$ and $e - 2u$ for any negative unit u, both of norm -3. Every vector of negative norm is a real multiple of a point of Equation (14.111), since rescaling by λ multiplies the norm by λ^2. Interchanging the two signs throughout gives the same conclusion for Equation (14.112). Hence $\omega = 0$.

The count closes the argument. By Equation (14.89) with $p + q = D + 1$, both algebras have dimension $D(D+1)/2$, which is the largest number of independent Killing fields a D-dimensional manifold can carry by Proposition 13.140. No further Killing field can be independent of them, so the isometry algebra is exactly what has been exhibited. □

The conformal algebra in D dimensions

The conformal algebra in D spacetime dimensions is

$$\mathcal{C}_D = \mathfrak{so}(D,2)\,, \tag{14.113}$$

with

$$\dim \mathcal{C}_D = \frac{(D+1)(D+2)}{2}\,. \tag{14.114}$$

From Equation (14.89) (signature $(D,2)$, so the underlying space has dimension $D+2$).

Equation (14.113) is an assertion, and what it asserts is that an algebra defined by a *conformal* condition on D-dimensional flat space is the *isometry* algebra of a flat space with two more dimensions. The conformal Killing fields themselves are written down in Equation (13.333); what remains, and is done here because it is pure Lie algebra, is to compute their brackets and exhibit the isomorphism. The dilatation generator is written Δ throughout, the letter D being reserved for the dimension by Notation 14.1.

Proposition 14.70 (The conformal algebra of flat space)**.** *Let $D = p + q \geq 3$ and, on $\mathbb{R}^{p,q}$ with the flat metric of Notation 14.1, put*

$$\boxed{P_A = \partial_A\,, \quad J_{AB} = x_A \partial_B - x_B \partial_A\,, \quad \Delta = x^C \partial_C\,, \quad K_A = 2x_A \Delta - x^2 P_A}\,, \tag{14.115}$$

with $x^2 = x^C x_C$. These $(D+1)(D+2)/2$ vector fields are the conformal Killing fields Equation (13.333) of the flat metric; they close into a Lie algebra, with brackets Equations (14.84), (14.90) and (14.97) together with

$$\begin{aligned} &[\Delta, P_A] = -P_A\,, \qquad [\Delta, K_A] = +K_A\,, \qquad\qquad [\Delta, J_{AB}] = 0\,, \\ &[K_A, K_B] = 0\,, \qquad [J_{AB}, K_C] = \eta_{BC} K_A - \eta_{AC} K_B\,, \quad [P_A, K_B] = 2\eta_{AB}\Delta - 2J_{AB}\,. \end{aligned} \tag{14.116}$$

That algebra is isomorphic to $\mathfrak{so}(p+1, q+1)$. For Minkowski space, $(p,q) = (D-1,1)$, it is $\mathfrak{so}(D,2)$, which is Equation (14.113).

Rests on Equations (13.333), (14.90) and (14.97).

Proof. *Derives Proposition 14.70.*

The fields are conformal Killing fields. In Cartesian coordinates the conformal Killing equation reads $\partial_B\xi_C + \partial_C\xi_B = \frac{2}{D}\left(\partial^E\xi_E\right)\eta_{BC}$. For P_A and J_{AB} both sides vanish, by Proposition 14.59. For Δ, $\xi_C = x_C$ gives $\partial_B\xi_C + \partial_C\xi_B = 2\eta_{BC}$ and $\partial^E\xi_E = D$, and the equation holds. For K_A, $\xi_C = 2x_Ax_C - x^2\eta_{AC}$ gives

$$\partial_B\xi_C = 2\eta_{AB}x_C + 2x_A\eta_{BC} - 2x_B\eta_{AC}\,,$$

whose symmetric part is $4x_A\eta_{BC}$ and whose trace is $\partial^E\xi_E = 2x_A + 2Dx_A - 2x_A = 2Dx_A$; again the equation holds. That these are *all* of them is Equation (13.333), and the count is $D + D(D-1)/2 + 1 + D = (D+1)(D+2)/2$.

The brackets. Every computation below uses only the derivation rule $[X, fY] = X(f)\,Y + f\,[X,Y]$ for a function f, the values

$$P_A(x_B) = \eta_{AB}\,,\quad \Delta(x_A) = x_A\,,\quad \Delta(x^2) = 2x^2\,,\quad J_{AB}(x_C) = x_A\eta_{BC} - x_B\eta_{AC}\,,\quad J_{AB}(x^2) = 0\,,$$

the last from Proposition 14.59(iii), and the identity $x_AP_B - x_BP_A = J_{AB}$.

For $[\Delta, P_A]$: $\left[x^C\partial_C, \partial_A\right] = -\left(\partial_Ax^C\right)\partial_C = -P_A$. For $[\Delta, J_{AB}]$, the same computation on each of the two terms of J_{AB} gives $x_AP_B - x_AP_B$ and so zero. For $[\Delta, K_A]$,

$$\left[\Delta, 2x_A\Delta - x^2P_A\right] = 2x_A\Delta - 2x^2P_A - x^2\,[\Delta, P_A] = 2x_A\Delta - x^2P_A = K_A\,.$$

For $[P_A, K_B]$, using $[P_A, \Delta] = P_A$ and $P_A(x^2) = 2x_A$,

$$[P_A, K_B] = 2\eta_{AB}\Delta + 2x_BP_A - 2x_AP_B = 2\eta_{AB}\Delta - 2J_{AB}\,.$$

For $[J_{AB}, K_C]$, using $[J_{AB}, \Delta] = 0$, $J_{AB}(x^2) = 0$ and Equation (14.97),

$$\begin{aligned}[J_{AB}, K_C] &= 2\,(x_A\eta_{BC} - x_B\eta_{AC})\,\Delta - x^2\,(\eta_{BC}P_A - \eta_{AC}P_B)\\ &= \eta_{BC}\left(2x_A\Delta - x^2P_A\right) - \eta_{AC}\left(2x_B\Delta - x^2P_B\right) = \eta_{BC}K_A - \eta_{AC}K_B\,.\end{aligned}$$

Finally, with $K_A(x_B) = 2x_Ax_B - x^2\eta_{AB}$, $K_A(x^2) = 4x_Ax^2 - 2x_Ax^2 = 2x_Ax^2$, $[K_A, \Delta] = -K_A$ and $[K_A, P_B] = -2\eta_{AB}\Delta - 2J_{AB}$,

$$\begin{aligned}[K_A, K_B] &= 2\left(2x_Ax_B - x^2\eta_{AB}\right)\Delta - 2x_BK_A - 2x_Ax^2P_B + x^2\,(2\eta_{AB}\Delta + 2J_{AB})\\ &= 4x_Ax_B\Delta - 2x_B\left(2x_A\Delta - x^2P_A\right) - 2x^2x_AP_B + 2x^2J_{AB}\\ &= 2x^2\,(x_BP_A - x_AP_B) + 2x^2J_{AB} = 0\,.\end{aligned}$$

The isomorphism. Let the indices M, N run over $1, \ldots, D+2$, extend η by

$$\eta_{D+1,D+1} = -1\,,\qquad \eta_{D+2,D+2} = +1\,, \tag{14.117}$$

all other new components vanishing, and set

$$\boxed{J_{A,D+1} = \tfrac{1}{2}\,(K_A - P_A)\,,\quad J_{A,D+2} = \tfrac{1}{2}\,(K_A + P_A)\,,\quad J_{D+1,D+2} = \Delta}\,, \tag{14.118}$$

the remaining J_{MN} being the J_{AB} already defined and $J_{NM} = -J_{MN}$. This is a linear bijection onto the span, since P_A and K_A are recovered as $J_{A,D+2} \mp J_{A,D+1}$. It remains to check that Equation (14.90) holds for every choice of the four indices among $1, \ldots, D+2$. With four indices in $1, \ldots, D$ it is Equation (14.90) itself. With exactly one index equal to $D+1$ or $D+2$ it is the bracket $[J_{AB}, K_C]$ of Equation (14.116) together with Equation (14.97), both of which say that K and P carry their index as vectors. The case $[J_{AB}, J_{D+1,D+2}] =$

$[J_{AB},\Delta]=0$ agrees with Equation (14.90), every term of which carries a vanishing η. There remain the pairs built from Equation (14.118). Using $[K_A,P_B]=-2\eta_{AB}\Delta-2J_{AB}$ and $[P_A,K_B]=2\eta_{AB}\Delta-2J_{AB}$,

$$\begin{aligned}[J_{A,D+1},J_{B,D+1}]&=\tfrac{1}{4}\,[K_A-P_A,K_B-P_B]=J_{AB}\,,\\ [J_{A,D+2},J_{B,D+2}]&=\tfrac{1}{4}\,[K_A+P_A,K_B+P_B]=-J_{AB}\,,\\ [J_{A,D+1},J_{B,D+2}]&=\tfrac{1}{4}\,[K_A-P_A,K_B+P_B]=-\eta_{AB}\Delta\,,\end{aligned}$$

while Equation (14.90) demands, for these three, $\eta_{D+1,D+1}J_{BA}=J_{AB}$, $\eta_{D+2,D+2}J_{BA}=-J_{AB}$ and $\eta_{AB}J_{D+2,D+1}=-\eta_{AB}\Delta$, which are the same three results. Likewise

$$[J_{D+1,D+2},J_{A,D+1}]=\Big[\Delta,\tfrac{1}{2}\,(K_A-P_A)\Big]=J_{A,D+2}\,,\qquad [J_{D+1,D+2},J_{A,D+2}]=J_{A,D+1}\,,$$

against the demands $\eta_{D+1,D+1}J_{D+2,A}=J_{A,D+2}$ and $\eta_{D+2,D+2}J_{A,D+1}=J_{A,D+1}$ of Equation (14.90). All cases agree, so the fields Equation (14.115) satisfy the brackets of $\mathfrak{so}$ of the extended metric Equation (14.117). That metric adds one positive and one negative direction, so its signature is $(p+1,q+1)$ after the coordinates are reordered to list the positive directions first. □

Remark 14.71 (Two spaces of two extra dimensions, and neither is physical). Propositions 14.69 and 14.70 both realize a D-dimensional symmetry linearly on a flat space of more dimensions — one more for the de Sitter quadrics, two more for the conformal algebra. The extra directions are a device for linearizing an action, not a claim about the world: the manifold in both cases remains D-dimensional, and this treatise instantiates $D=4$. Nothing here asserts the existence of extra dimensions, which have no evidence base and are excluded by the scope rule of Chapter 1. The geometric side of the same statement, including the projective null-cone realization of the conformal action, is Section 13.13.1.

14.4 Central extensions

An algebra of generators need not close on itself. It may close only on itself together with an extra element that commutes with everything and generates no transformation—a *central charge*. This is not a pathology: two of the most important constants in physics, the mass in nonrelativistic mechanics and Planck's constant in the canonical commutation relations, appear in exactly this position, and neither can be removed by any redefinition of the generators. This section says what the obstruction is and how it is classified.

14.4.1 Extensions and cocycles

Definition 14.72 (Central extension)**.** A *central extension* of a Lie algebra $\mathfrak{g}$ by $\mathbb{R}$ is a Lie algebra $\tilde{\mathfrak{g}}$ fitting into the exact sequence

$$0\longrightarrow\mathbb{R}\longrightarrow\tilde{\mathfrak{g}}\longrightarrow\mathfrak{g}\longrightarrow 0\,,\tag{14.119}$$

with the image of $\mathbb{R}$ contained in the centre of $\tilde{\mathfrak{g}}$. Choosing a linear splitting $\xi\mapsto\tilde{\xi}$ of the projection, the bracket of $\tilde{\mathfrak{g}}$ takes the form

$$\boxed{\big[\tilde{\xi},\tilde{\eta}\big]=\widetilde{[\xi,\eta]}+c(\xi,\eta)\,Z}\,,\tag{14.120}$$

where Z is the central generator and $c:\mathfrak{g}\times\mathfrak{g}\to\mathbb{R}$ is a bilinear map.

Proposition 14.73 (The extension datum is a 2-cocycle)**.** *The map c of Equation (14.120) is antisymmetric and satisfies*

$$\boxed{c\left([\xi,\eta],\zeta\right)+c\left([\eta,\zeta],\xi\right)+c\left([\zeta,\xi],\eta\right)=0}, \tag{14.121}$$

for all $\xi,\eta,\zeta\in\mathfrak{g}$. A bilinear map with these two properties is called a 2-cocycle *on $\mathfrak{g}$.*
Rests on Definition 14.72.

Proof. *Derives Proposition 14.73.*

Antisymmetry of c follows from antisymmetry of the bracket in Equation (14.120). Imposing the Jacobi identity on $\tilde{\mathfrak{g}}$ for the three elements $\tilde{\xi},\tilde{\eta},\tilde{\zeta}$, the terms not involving Z cancel because the Jacobi identity already holds in $\mathfrak{g}$, and Z is central so it contributes nothing to a nested bracket; what survives is the coefficient of Z, which is exactly Equation (14.121). □

Definition 14.74 (Coboundary and triviality)**.** A cocycle is a *coboundary* if there is a linear map $b:\mathfrak{g}\to\mathbb{R}$ with

$$c(\xi,\eta)=b\left([\xi,\eta]\right)\,. \tag{14.122}$$

An extension whose cocycle is a coboundary is called *trivial*: the shift $\tilde{\xi}\mapsto\tilde{\xi}+b(\xi)Z$ removes c from Equation (14.120) entirely, so that $\tilde{\mathfrak{g}}\cong\mathfrak{g}\oplus\mathbb{R}$ as a direct sum of Lie algebras and the central charge decouples.
Rests on Proposition 14.73.

Definition 14.75 (The classifying group)**.** The 2-cocycles of Proposition 14.73 form a real vector space and the coboundaries of Definition 14.74 a subspace of it. The quotient is the second Chevalley–Eilenberg cohomology group $H^2(\mathfrak{g},\mathbb{R})$. Two central extensions of $\mathfrak{g}$ by $\mathbb{R}$ are called *equivalent* when there is an isomorphism of Lie algebras between them commuting with the inclusion of $\mathbb{R}$ and with the projection onto $\mathfrak{g}$ in Equation (14.119).
Rests on Proposition 14.73 and Definition 14.74.

That H^2 deserves to be called the classifying group is a theorem, not part of the definition, and it is proved next.

Proposition 14.76 (H^2 classifies the central extensions)**.** *The assignment Equation (14.120) of a cocycle to an extension induces a bijection between $H^2(\mathfrak{g},\mathbb{R})$ and the set of central extensions of $\mathfrak{g}$ by $\mathbb{R}$ up to equivalence, the zero class corresponding to the trivial extension. In particular*

$$\boxed{H^2(\mathfrak{g},\mathbb{R})=0\iff\textit{every central extension of }\mathfrak{g}\textit{ is trivial}}. \tag{14.123}$$

Rests on Definitions 14.72, 14.74 and 14.75 and Proposition 14.73.

Proof. *Derives Proposition 14.76.*

Write ∂b for the coboundary of Equation (14.122), $(\partial b)(\xi,\eta)=b\left([\xi,\eta]\right)$. Three things have to be checked.

The class does not depend on the splitting. Two linear splittings of the projection differ by a linear map $\mathfrak{g}\to\mathbb{R}$: if $\xi\mapsto\tilde{\xi}$ and $\xi\mapsto\tilde{\xi}'$ both project to ξ, their difference lies in the kernel of the projection, which is the line spanned by Z, so $\tilde{\xi}'=\tilde{\xi}+b(\xi)Z$ with b linear. Since Z is central, $\left[\tilde{\xi}',\tilde{\eta}'\right]=\left[\tilde{\xi},\tilde{\eta}\right]$, while the first term of Equation (14.120) shifts by $b\left([\xi,\eta]\right)Z$. Hence $c'=c-\partial b$ and the two cocycles define the same class. An isomorphism of extensions in the sense of Definition 14.75 is the identity on $\mathbb{R}$ and induces the identity on $\mathfrak{g}$, so it acts on a splitting exactly as a change of splitting does; equivalent extensions therefore have equal classes.

Every class is realized. Given a cocycle c, put $\tilde{\mathfrak{g}}=\mathfrak{g}\oplus\mathbb{R}Z$ as a vector space, with Z central and the bracket Equation (14.120). It is antisymmetric because c is, and it satisfies

the Jacobi identity: the $\mathfrak{g}$-component of the Jacobi expression is the Jacobi identity of $\mathfrak{g}$, and the coefficient of Z is Equation (14.121). The obvious inclusion and projection make it a central extension, whose cocycle for the obvious splitting is c again.

Cohomologous cocycles give equivalent extensions, and only those. A map between the extensions built from c and from c' that commutes with the inclusion and the projection is the identity on $\mathbb{R}Z$ and induces the identity on $\mathfrak{g}$, so it sends $\tilde{\xi} \mapsto \tilde{\xi} + \beta(\xi)Z$ for some linear $\beta : \mathfrak{g} \to \mathbb{R}$ and is automatically bijective. Requiring it to respect the two brackets and reading off the coefficient of Z gives $c' = c + \partial\beta$. So the two extensions are equivalent precisely when $c' - c$ is a coboundary, that is, precisely when c and c' define the same class. The trivial extension is the one with $c = 0$, which is Definition 14.74, and it represents the zero class. □

Proposition 14.77 (Semisimple algebras admit no nontrivial extension)**.** *If $\mathfrak{g}$ is a finite-dimensional semisimple Lie algebra over a field of characteristic zero, then $H^1(\mathfrak{g}, \mathbb{R}) = H^2(\mathfrak{g}, \mathbb{R}) = 0$. In particular $\mathfrak{so}(p, q)$, and hence the Lorentz algebra $\mathfrak{so}(D-1, 1)$ of Equation (14.90), admits no nontrivial central extension.*

Rests on Definition 14.75 and Theorem 14.13.

These are Whitehead's first and second lemmas. Both are proved by averaging a cochain against the Casimir element of the algebra, which Theorem 14.13 makes invertible on any module with no trivial summand; the corollary for the Lorentz algebra then follows from its semisimplicity, which is itself the nondegeneracy of the Killing form.

Full derivation: Appendix A.37 (p. 2967).

Derives Proposition 14.77.

Remark 14.78 (Where the obstruction can live). Proposition 14.77 localises the interesting cases. A semisimple algebra is rigid in this respect, so nontrivial central charges occur only where the algebra is *not* semisimple: in abelian algebras, in the inhomogeneous algebras built as semidirect sums such as Equation (14.99), in algebras obtained by the contractions of Section 14.5, and in infinite-dimensional algebras. All four cases occur in physics.

14.4.2 The instances that occur in physics

Example 14.79 (The Heisenberg algebra)**.** Take $\mathfrak{g} = \mathbb{R}^{2f}$, abelian, spanned by the phase-space translations. Every antisymmetric bilinear form on an abelian algebra is a cocycle, since Equation (14.121) is vacuous, and no coboundary is available because $[\xi, \eta] = 0$ makes the right-hand side of Equation (14.122) vanish identically. Hence $H^2(\mathbb{R}^{2f}, \mathbb{R}) = \Lambda^2(\mathbb{R}^{2f})^*$, which is as far from zero as possible. The extension defined by the symplectic form is the *Heisenberg algebra*

$$[X_i, P_j] = \delta_{ij} Z\,, \qquad [X_i, X_j] = [P_i, P_j] = [Z, \cdot] = 0\,, \tag{14.124}$$

and the canonical commutation relations of quantum mechanics are Equation (14.124) with $Z = \mathrm{i}\hbar$. Planck's constant is a central charge.

Rests on Proposition 14.73 and Definition 14.74.

Example 14.80 (The Galilei algebra)**.** The ten-generator Galilei algebra of three space dimensions carries the cocycle $c(K_i, P_j) = \delta_{ij}$, whose central charge is the mass, and H^2 of that algebra is one-dimensional, so up to a scale this is the only central extension it admits. The classical mechanics of a free particle realizes the extension and not the Galilei algebra itself; the computation, and the proof that this cocycle is not a coboundary, are carried out in Proposition 25.14. That proposition establishes only that the class is nonzero; that it exhausts H^2 is the stronger statement, and it needs the cocycle condition solved on the whole ten-dimensional algebra and quotiented by the coboundaries.

Rests on Definition 14.75 and Proposition 25.14.

Full derivation: Appendix A.38 (p. 2975).
Derives Example 14.80.

Example 14.81 (Infinite-dimensional cases)**.** The Witt algebra of vector fields on the circle has a one-dimensional H^2, whose extension is the Virasoro algebra with central charge c; loop algebras extend to affine Kac–Moody algebras with the level as central charge; and the equal-time current-algebra commutators of a field theory acquire Schwinger terms of the same nature.

All three are computations rather than citations: the Virasoro cocycle and its uniqueness up to coboundary, the Kac–Moody cocycle built from the Killing form and the residue pairing, and the Schwinger term as the statement that a formally central extension survives regularization.
Full derivation: Appendix A.39 (p. 2979).
Derives Example 14.81.

Remark 14.82 (Central charges are measured, not postulated)*.* The four examples above are not on the same evidential footing, and the distinction matters under the scope rule of this treatise. Planck's constant and the mass are measured quantities appearing as central charges of algebras whose representations describe observed systems in the observed $3+1$ dimensions; nothing stands between the algebra and the instrument.

The Virasoro central charge is not on that footing, and this treatise records no measurement of it. Where it is called observable, the claim runs through the universal critical behaviour of a laboratory sample — a three-dimensional piece of matter, in the regime treated in Chapter 123, whose long-wavelength behaviour is that of an effectively two-dimensional system when one direction is frozen out by geometry or by temperature. What is measured there are critical exponents and amplitude ratios; the step from those numbers to a value of c is an identification made inside a conformal field theory, not a reading of an instrument, and this book does not make it. The system is $3+1$-dimensional and the effective description is not a spacetime.

The Schwinger terms are on a third footing again. They are the equal-time-commutator face of the anomalies [1388, 1391], and it is the anomaly that carries the measurable consequence — the cleanest case being the two-photon decay of the neutral pion, whose rate has been measured directly [2361]. That the Schwinger term and the anomaly are two faces of one obstruction is asserted here and not proved; what can be said without qualification is that the decay rate is measured and that an anomaly-free theory predicts the wrong one.

What is *not* evidence in any of these cases is the existence of an algebra as such: writing down a consistent extension is a mathematical act, and whether nature realizes it is a separate question, settled only by experiment.

14.5 Contraction of Lie algebras

14.5.1 The Inönü–Wigner limit

A symmetry algebra is rarely exact. More often it is the limit of another one as some parameter—a velocity ratio, a curvature radius—is taken to an extreme, and the limit is singular: the generators must be rescaled as it is taken, or the brackets diverge. Inönü and Wigner made the procedure precise [1021].

Definition 14.83 (Contraction)**.** Let $\mathfrak{g}$ be a Lie algebra with basis $\{T_\alpha\}$ and let $U(\varepsilon)$ be a family of invertible linear maps, singular as $\varepsilon \to 0$. If the limit

$$\boxed{[T_\alpha, T_\beta]_0 = \lim_{\varepsilon\to 0} U(\varepsilon)^{-1}\,[U(\varepsilon)T_\alpha, U(\varepsilon)T_\beta]} \tag{14.125}$$

exists for all α, β, it defines a Lie algebra $\mathfrak{g}_0$ of the same dimension, called a *contraction* of $\mathfrak{g}$.

Proposition 14.84 (Properties of a contraction)**.** *Let $\mathfrak{g}_0$ be a contraction of $\mathfrak{g}$ in the sense of Definition 14.83, and fix a basis $\{T_\alpha\}$, in which $U(\varepsilon)$ has matrix U and $\mathfrak{g}$ has structure constants ${C^\gamma}_{\alpha\beta}$ and Killing form κ. Then:*

1. *$[\cdot,\cdot]_0$ is a Lie bracket, and $\dim \mathfrak{g}_0 = \dim \mathfrak{g}$;*
2. *the derivation algebras obey $\dim \mathrm{Der}(\mathfrak{g}_0) \geq \dim \mathrm{Der}(\mathfrak{g})$, so that a contraction is* not reversible*: whenever the inequality is strict, the two algebras are not isomorphic and $\mathfrak{g}$ is not a contraction of $\mathfrak{g}_0$;*
3. *the Killing form of $\mathfrak{g}_0$ is $\kappa_0 = \lim_{\varepsilon\to 0} \kappa_\varepsilon$ with $\kappa_\varepsilon(X,Y) = \kappa\left(U(\varepsilon)X, U(\varepsilon)Y\right)$, so that $\det \kappa_0 = \lim_{\varepsilon\to 0} (\det U)^2 \det \kappa$. If $\det U(\varepsilon) \to 0$ and $\mathfrak{g}$ is semisimple, then $\mathfrak{g}_0$ is not;*
4. *a contraction can create the cohomology that a central charge needs. Every Lie algebra of dimension n contracts to the abelian algebra of the same dimension, whose second cohomology is the whole space $\Lambda^2 (\mathbb{R}^n)^*$ of antisymmetric bilinear forms by the argument of Example 14.79, while H^2 vanishes for a semisimple $\mathfrak{g}$ by Proposition 14.77.*

Rests on Definitions 14.10 and 14.83 and Proposition 14.77.

Proof. *Derives Proposition 14.84.*

Throughout, write $[X,Y]_\varepsilon = U(\varepsilon)^{-1}\,[U(\varepsilon)X, U(\varepsilon)Y]$ for the bracket at finite ε. It is the pullback of the bracket of $\mathfrak{g}$ along the linear isomorphism $U(\varepsilon)$, so $(\mathfrak{g}, [\cdot,\cdot]_\varepsilon)$ is a Lie algebra isomorphic to $\mathfrak{g}$ for every $\varepsilon \neq 0$, and its structure constants ${C^\gamma}_{\alpha\beta}(\varepsilon)$ converge, by hypothesis, to those of $\mathfrak{g}_0$.

(i) Bilinearity and antisymmetry pass to the limit because they are linear conditions on the structure constants. The Jacobi identity is the system of quadratic equations ${C^\delta}_{\alpha\beta}(\varepsilon){C^\sigma}_{\delta\gamma}(\varepsilon) + {C^\delta}_{\beta\gamma}(\varepsilon){C^\sigma}_{\delta\alpha}(\varepsilon) + {C^\delta}_{\gamma\alpha}(\varepsilon){C^\sigma}_{\delta\beta}(\varepsilon) = 0$, valid for every $\varepsilon \neq 0$; a polynomial is continuous, so the same equations hold in the limit. The underlying vector space never changed, whence the dimension.

(ii) A derivation is a solution of the linear system $d\,([T_\alpha, T_\beta]) = [dT_\alpha, T_\beta] + [T_\alpha, dT_\beta]$ in the unknown matrix d, and the coefficients of that system are linear in the structure constants. Write $L(\varepsilon)$ for its matrix. Since Der of isomorphic algebras are conjugate, $\dim \ker L(\varepsilon) = \dim \mathrm{Der}(\mathfrak{g})$ for all $\varepsilon \neq 0$. Rank is lower semicontinuous: if $L(0)$ carries a nonvanishing minor of size r, that same minor is a continuous function of ε and so is nonvanishing for small ε, whence $\operatorname{rank} L(0) \leq \operatorname{rank} L(\varepsilon)$ there, and the kernels satisfy the reverse inequality. That is the claim. If it is strict, then $\mathfrak{g}_0$ is not isomorphic to $\mathfrak{g}$, and a contraction the other way would give $\dim \mathrm{Der}(\mathfrak{g}) \geq \dim \mathrm{Der}(\mathfrak{g}_0)$ and hence equality, which is excluded.

(iii) From $\mathrm{ad}^\varepsilon_X = U(\varepsilon)^{-1}\,\mathrm{ad}_{U(\varepsilon)X}\,U(\varepsilon)$ and cyclicity of the trace,

$$\kappa_\varepsilon(X,Y) = \operatorname{tr}\left(\mathrm{ad}^\varepsilon_X\,\mathrm{ad}^\varepsilon_Y\right) = \operatorname{tr}\left(\mathrm{ad}_{U(\varepsilon)X}\,\mathrm{ad}_{U(\varepsilon)Y}\right) = \kappa\left(U(\varepsilon)X, U(\varepsilon)Y\right)\,,$$

which in the fixed basis is the matrix $(U)^{\mathsf{T}}\kappa\,U$, of determinant $(\det U)^2 \det\kappa$. The Killing form is a quadratic polynomial in the structure constants by Equation (14.15), so κ_0 is the limit of κ_ε and $\det\kappa_0$ the limit of the determinants. If $\det U \to 0$ then $\det\kappa_0 = 0$, the Killing form of $\mathfrak{g}_0$ is degenerate, and Theorem 14.13 denies it semisimplicity.

(iv) Take $U(\varepsilon) = \varepsilon\,\mathbb{1}$. Then $[X,Y]_\varepsilon = \varepsilon^{-1}\,[\varepsilon X, \varepsilon Y] = \varepsilon\,[X,Y] \to 0$, so every algebra contracts to the abelian one of the same dimension; and $\det U = \varepsilon^n \to 0$, consistently with (iii). The second cohomology of an abelian algebra is computed in Example 14.79 and is the whole space of antisymmetric bilinear forms, nonzero as soon as $n \geq 2$. □

Remark 14.85 (Why (iv) is only an existence statement). Part (iv) proves that a contraction *can* create a nonvanishing H^2, by exhibiting the extreme case. It does not say that any

particular contraction does so, and no general theorem does: which classes survive a given limit has to be computed limit by limit. What the physical cases have in common is the structure of Lemma 14.86 — an abelian ideal appears where there was none — and it is on an abelian ideal that a cocycle has the most room, by the same computation as in Example 14.79.

Lemma 14.86 (The Inönü–Wigner contraction along a subalgebra)**.** *Let $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ as a vector space, let π be the projection onto $\mathfrak{m}$ along $\mathfrak{h}$, and let $U(\varepsilon)$ be the identity on $\mathfrak{h}$ and multiplication by ε on $\mathfrak{m}$. The limit Equation (14.125) exists if and only if $\mathfrak{h}$ is a subalgebra, and then*

$$\boxed{[X,Y]_0 = [X,Y]\,, \quad [X,Z]_0 = \pi\,[X,Z]\,, \quad [Z,Z']_0 = 0}\,, \tag{14.126}$$

for $X, Y \in \mathfrak{h}$ and $Z, Z' \in \mathfrak{m}$. In particular $\mathfrak{m}$ is an abelian ideal of $\mathfrak{g}_0$ and $\mathfrak{g}_0/\mathfrak{m} \cong \mathfrak{h}$.
Rests on Definition 14.83.

Proof. *Derives Lemma 14.86.*
For $X, Y \in \mathfrak{h}$, $U(\varepsilon)$ acts as the identity and $U(\varepsilon)^{-1}$ multiplies the $\mathfrak{m}$-part by ε^{-1}, so $[X,Y]_\varepsilon = (1-\pi)\,[X,Y] + \varepsilon^{-1}\pi\,[X,Y]$, which converges if and only if $\pi\,[X,Y] = 0$, that is if and only if $[\mathfrak{h},\mathfrak{h}] \subseteq \mathfrak{h}$; the limit is then $[X,Y]$. For $X \in \mathfrak{h}$ and $Z \in \mathfrak{m}$,

$$[X,Z]_\varepsilon = U(\varepsilon)^{-1}\,[X,\varepsilon Z] = \varepsilon\,(1-\pi)\,[X,Z] + \pi\,[X,Z] \longrightarrow \pi\,[X,Z]\,,$$

and for $Z, Z' \in \mathfrak{m}$, $[Z,Z']_\varepsilon = \varepsilon^2\,(1-\pi)\,[Z,Z'] + \varepsilon\pi\,[Z,Z'] \to 0$. Both limits exist unconditionally, so the condition on $\mathfrak{h}$ is the whole condition. Equation (14.126) then puts $[\mathfrak{g}_0,\mathfrak{m}]_0 \subseteq \mathfrak{m}$ and $[\mathfrak{m},\mathfrak{m}]_0 = 0$, and the quotient bracket is that of $\mathfrak{h}$. □

Lemma 14.87 (A rescaled Casimir stays central)**.** *Let $C \in U(\mathfrak{g})$ be a Casimir element in the sense of Definition 14.8, and for $\varepsilon \neq 0$ let $C_\varepsilon = \lambda(\varepsilon)\,\hat{U}(\varepsilon)^{-1}C$, where $\hat{U}(\varepsilon)$ is the isomorphism of enveloping algebras induced by $U(\varepsilon)$ and $\lambda(\varepsilon) \neq 0$ is a scalar. If the coefficients of C_ε in the ordered monomials of Theorem 14.7 converge as $\varepsilon \to 0$, their limit defines an element $C_0 \in U(\mathfrak{g}_0)$, and C_0 is a Casimir element of $\mathfrak{g}_0$.*
Rests on Definitions 14.8 and 14.83 and Theorem 14.7.

Proof. *Derives Lemma 14.87.*
Because $U(\varepsilon)$ is an isomorphism of Lie algebras from $(\mathfrak{g}, [\cdot,\cdot]_\varepsilon)$ to $\mathfrak{g}$, it extends to an isomorphism $\hat{U}(\varepsilon)$ of the corresponding enveloping algebras, which carries the centre to the centre. Hence $[C_\varepsilon, T_\alpha]_\varepsilon = 0$ for every α and every $\varepsilon \neq 0$, the scalar $\lambda(\varepsilon)$ being irrelevant to that.

Reduce both sides to the ordered basis of Theorem 14.7. Every reordering step uses the relations Equation (14.12) for the bracket $[\cdot,\cdot]_\varepsilon$, so the coefficients of $[C_\varepsilon, T_\alpha]_\varepsilon$ in that basis are polynomials in the coefficients of C_ε and in the structure constants ${C^\gamma}_{\alpha\beta}(\varepsilon)$, with the same polynomials for every ε — the degree of C_ε being fixed. Both arguments converge, the polynomials are continuous, and the value is identically zero; therefore $[C_0, T_\alpha]_0 = 0$ for every α, which is Equation (14.13) for $\mathfrak{g}_0$. □

Remark 14.88 (The Galilei mass, precisely). Lemma 14.87 is the mechanism by which an invariant of the parent algebra reappears as a central quantity of the contracted one, and the nonrelativistic limit is its standard instance — but the instance has to be stated carefully, because the plain contraction of the Poincaré algebra does *not* produce a central charge. Take $D = 4$ in the ordering of Notation 14.1, so that $x^4 = ct$ and the generator of time translation is $H = c\,P_4 = \partial_t$, of SI unit 1/s, and put $G_i = J_{i4}/c$ for the rescaled boost, which tends to the Galilean $t\,\partial_i$. Then Equations (14.90) and (14.97) give

$$[G_i, P_j] = -\frac{\delta_{ij}}{c^2}\,H\,, \qquad [G_i, H] = -P_i\,, \qquad [G_i, G_j] = \frac{1}{c^2}J_{ij}\,, \tag{14.127}$$

and the limit $c \to \infty$ is the Galilei algebra with $[G_i, P_j] = 0$ and no central charge at all. The mass appears when the same limit is taken in the *trivially extended* algebra $\mathfrak{iso}(3,1) \oplus \mathbb{R}M$, with M central, after the c-dependent redefinition $H' = H - c^2 M$: the first bracket of Equation (14.127) then reads $[G_i, P_j] = -\delta_{ij}\,(H'/c^2 + M)$ and tends to $-\delta_{ij}M$. The extension that survives is nontrivial, which is the content of Proposition 25.14; this treatise proves that nontriviality directly, from the Poisson brackets of a free particle, rather than through the limit. Remark 25.15 is to be read in the sense set out here.

Example 14.89 (Poincaré from de Sitter, Galilei from Poincaré)**.** Two contractions are used throughout this treatise.

1. *Vanishing curvature.* Rescaling the translation generators of $\mathfrak{so}(D-1,2)$ or $\mathfrak{so}(D,1)$ by the inverse radius and letting the radius diverge contracts either de Sitter algebra to the Poincaré algebra Equation (14.99). This is the statement that spacetime looks flat on scales small compared with the curvature radius set by the cosmological constant.

2. *Vanishing velocity ratio.* Rescaling the boosts of the Poincaré algebra by $1/c$ and letting $c \to \infty$ contracts it to the Galilei algebra. The Casimir $p_\mu p^\mu = m^2 c^2$ that labelled a Poincaré representation becomes, in the limit, the central charge of Example 14.80.

14.5.2 The classification of kinematical algebras

A *kinematical* algebra is one generated by the transformations any spacetime symmetry must contain: rotations J_i, boosts K_i, space translations P_i, and a time translation H—ten generators in four spacetime dimensions. Bacry and Lévy-Leblond asked which Lie algebras these can form, and answered the question completely [1364].

Theorem 14.90 (Bacry–Lévy-Leblond)**.** *Assume that*

1. *space is isotropic, so that the J_i generate an $\mathfrak{so}(3)$ under which K_i and P_i are vectors and H a scalar;*

2. *parity and time reversal are automorphisms of the algebra; and*

3. *each boost generates a noncompact one-parameter subgroup.*

Then there are exactly eleven such algebras, up to a rescaling of the generators, and they are organized by three constants (γ, μ, ν) appearing in

$$\boxed{[K_i, P_j] = \gamma\,\delta_{ij} H\,, \qquad [K_i, H] = \mu\, P_i\,, \qquad [P_i, H] = \nu\, K_i}\,, \tag{14.128}$$

the two remaining brackets being fixed by the Jacobi identity to

$$[K_i, K_j] = -\gamma\mu\,\epsilon_{ijk} J_k\,, \qquad [P_i, P_j] = \gamma\nu\,\epsilon_{ijk} J_k\,. \tag{14.129}$$

Setting each of γ, μ, ν to zero is an Inönü–Wigner contraction, so the eight possible vanishing patterns form a cube of contractions; the eleven algebras are those eight cases, three of which split into two according to a sign.

Rests on Equation (14.34) and Definition 14.83.

Full derivation: Appendix A.10 (p. 2799).

Derives Theorem 14.90.

Remark 14.91 (Eleven kinematics, not eleven isomorphism classes)*.* The theorem must not be stated as "eleven algebras up to isomorphism". Exchanging $K_i \leftrightarrow P_i$ is an isomorphism of Lie algebras carrying Poincaré to para-Poincaré and Galilei to para-Galilei (Corollary A.60), but it is not an equivalence of kinematics, because it does not commute with the time reversal of hypothesis (2) and because it exchanges the physical roles of a boost and a translation. Two entries of Table 14.1 are therefore abstractly the same algebra as another entry while describing different physics.

Table 14.1: The eleven kinematical algebras of Theorem 14.90, what each describes, and whether this treatise records a measurement in its regime. The third column gives the vanishing pattern of the three constants of Equation (14.128), $*$ marking a nonzero one; the eight patterns are the vertices of the contraction cube, and the three cases carrying two entries differ only by a sign. The last column is a verdict about the *evidence base*, not about the mathematics — all eleven are consistent Lie algebras — and the evidence for the three realized cases is cited in Remark 14.92.

Algebra	Sym.	(γ, μ, ν)	Regime it describes	Realized in nature?
de Sitter	dS	$(*, *, *)$	relativistic, positive curvature	yes, asymptotically: the measured Λ is positive
anti-de Sitter	AdS	$(*, *, *)$	relativistic, negative curvature	no: the measured Λ has the opposite sign
Poincaré	P	$(*, *, 0)$	relativistic, flat	yes: the symmetry of Part IV, tested directly
Galilei	G	$(0, *, 0)$	nonrelativistic, flat	yes, as the $c \to \infty$ limit used throughout Part III
Newton–Hooke	NH_+	$(0, *, *)$	nonrelativistic, positive curvature	only as an effective description of Λ in local dynamics
anti-Newton–Hooke	NH_-	$(0, *, *)$	nonrelativistic, negative curvature	no, for the same reason as AdS
Carroll	C	$(*, 0, 0)$	ultrarelativistic, $c \to 0$, flat	no laboratory regime; it is the intrinsic symmetry of a null hypersurface
para-Poincaré	P'	$(*, 0, *)$	ultrarelativistic, curved	no
para-Galilei	G'	$(0, 0, *)$	mixed limit	no
inhomogeneous SO(4)	E'	$(*, 0, *)$	Euclidean signature	not a spacetime kinematics at all
static	S	$(0, 0, 0)$	no relative motion whatever	degenerate limit only

Remark 14.92 (What the table is claiming). Three of the eleven are realized in the strong sense that measurements have been made in their regime and agree with them. Poincaré is the subject of Part IV and is tested to the precision recorded in Chapters 37 and 41. Galilei is its low-velocity limit, which is what Part III uses throughout. De Sitter qualifies in a weaker and more careful sense: the measured cosmological constant is positive — the supernova evidence for an accelerating expansion [1962, 1981], with the value fixed by the cosmic microwave background [2348] — so the asymptotic symmetry of the observed universe is de Sitter rather than anti-de Sitter or Poincaré — but this is an inference from the expansion history of Chapter 48, not a symmetry test. The remaining eight are mathematically consistent and, at present, physically unoccupied.

The Carroll algebra is the interesting borderline case. It has no laboratory regime, being the limit $c \to 0$, yet it is the intrinsic symmetry of a null hypersurface, and null hypersurfaces are observed — the horizons of Chapter 50. That is a statement about where an algebra appears in a description, which is not the same as a measurement of the algebra, and the table's last column is deliberately worded to keep the two apart.

Remark 14.93 (The cube is an algebra, not only a picture). The eight vanishing patterns of (γ, μ, ν) were called the vertices of a cube of contractions, which so far is a way of drawing the theorem rather than a statement about it. It is more than that. Section 15.7 of the next chapter shows that the four generator species J_i, K_i, H, P_i — the boosts are written K_i here and in Theorem 14.90, and G_i there — are the four components of a $\mathbb{Z}_2 \times \mathbb{Z}_2$-grading of the algebra, that the three constants (γ, μ, ν) are the three structure constants of a single four-dimensional commutative associative algebra A carrying the same grading, and that the eight vertices are the eight ways those three constants can vanish — so each edge of the cube is a degeneration of A and nothing else. The three rescaling-invariant signs of that algebra are the squares of its three graded generators; hypothesis (3) above, that the boosts be noncompact, is exactly the condition that one of those three squares be nonnegative; and counting the vertices against the signs that survive it returns $8 + 3 = 11$. The two accounts are complementary rather than redundant: this chapter's, completed in Appendix A.10, proves that the list is *exhaustive*, which no construction can do; the next chapter's says what the list *is* — one algebra and its degenerations.

Remark 14.94 (Extensions of the classification, and the scope rule). Theorem 14.90 is not the end of the subject. Dropping the discrete automorphisms, or working in general dimension, enlarges the list; extending the algebras—by further generators or by central charges—produces infinite families, and the construction of gauge theories from them is an active programme. Supersymmetric extensions are the other standard enlargement, and this treatise names them only to say what its scope rule requires: no supersymmetric partner of any known particle has been observed, so supersymmetry has no evidence base and nothing built on it belongs in this book. A recent example builds *Maxwell* extensions of the kinematical algebras by a semigroup expansion method that generalizes contraction, and uses the resulting invariant tensors to construct three-dimensional Chern–Simons gravity theories [2432]. These constructions are mathematics, and this treatise treats them as such: gravity in three spacetime dimensions has no propagating degrees of freedom and no observational status, and the paper claims none. The Chern–Simons machinery they use is developed in Section 14.8.1 of this chapter, so the mathematical apparatus is present in the book; what is absent, and is stated here rather than left implicit, is any measurement that selects such an extension. Section 133.7 records the general policy.

14.6 Fibre bundles

The invariant-polynomial machinery of the following sections is formulated on a *principal fibre bundle*, and so is every gauge theory in Part XI. This section supplies the definitions

and proves the one identity those constructions consume, the Bianchi identity. Nothing beyond that is developed: the classification of bundles, characteristic numbers as topological invariants, and holonomy are not needed anywhere in this treatise and are not treated.

14.6.1 Bundles, sections and principal bundles

Definition 14.95 (Fibre bundle)**.** A *fibre bundle* consists of differentiable manifolds E (the *total space*), M (the *base*) and F (the *typical fibre*), a Lie group G (the *structure group*) acting effectively on F on the left, and a smooth surjection $\pi : E \longrightarrow M$ (the *projection*), such that M carries an open cover $\{U_i\}$ and there are diffeomorphisms

$$\varphi_i : \pi^{-1}(U_i) \longrightarrow U_i \times F\,, \qquad \mathrm{pr}_1 \circ \varphi_i = \pi\,, \tag{14.130}$$

called *local trivializations*, whose overlaps act on the fibre through G:

$$\varphi_i \circ \varphi_j^{-1}(x, f) = (x, t_{ij}(x) f)\ , \qquad t_{ij} : U_i \cap U_j \longrightarrow G \tag{14.131}$$

smooth. The t_{ij} are the *transition functions*; composing Equation (14.131) on triple overlaps gives the *cocycle conditions*

$$t_{ii} = e\,, \qquad t_{ij} t_{ji} = e\,, \qquad t_{ij} t_{jk} = t_{ik}\,. \tag{14.132}$$

The bundle is *trivial* if a single trivialization with $U_1 = M$ exists, in which case $E \cong M \times F$.

Definition 14.96 (Section)**.** A *section* of E over an open $U \subseteq M$ is a smooth map $s : U \longrightarrow E$ with $\pi \circ s = \mathrm{id}_U$; it is *global* if $U = M$. Local sections always exist — take $s(x) = \varphi_i^{-1}(x, f_0)$ inside a trivializing neighbourhood — and a global section need not.
Rests on Definition 14.95.

Definition 14.97 (Principal bundle)**.** A *principal G-bundle* $P(M, G)$ is a fibre bundle whose typical fibre is the group G itself, acting on itself by left translation, together with the right action of G on P defined in each trivialization by

$$\varphi_i(p) = (x, h) \quad \longmapsto \quad \varphi_i(p \cdot g) = (x, hg)\ . \tag{14.133}$$

The action is well defined — right translation commutes with the left translation by t_{ij} of Equation (14.131), so Equation (14.133) does not depend on the trivialization — it is free, and its orbits are exactly the fibres of π.
Rests on Definition 14.95.

Proposition 14.98 (A principal bundle is trivial if and only if it has a global section)**.** *$P(M, G)$ is trivial if and only if it admits a global section.*
Rests on Definitions 14.96 and 14.97.

Proof. *Derives Proposition 14.98.*
If $P \cong M \times G$, then $x \mapsto (x, e)$ is a global section. Conversely let s be one and define

$$\Phi : M \times G \longrightarrow P\,, \qquad \Phi(x, g) = s(x) \cdot g\,.$$

It is smooth, satisfies $\pi \circ \Phi = \mathrm{pr}_1$, and is bijective: surjective because G acts transitively on each fibre, so every p with $\pi(p) = x$ is $s(x)g$ for some g; injective because the action is free. Its inverse is smooth, as one sees in any local trivialization φ_i: writing $\varphi_i(s(x)) = (x, \sigma_i(x))$ with σ_i smooth, $\varphi_i \circ \Phi(x, g) = (x, \sigma_i(x) g)$, whose inverse $(x, h) \mapsto (x, \sigma_i(x)^{-1} h)$ is smooth because inversion and multiplication in G are. Hence Φ is a diffeomorphism commuting with the right action, and P is trivial. □

Definition 14.99 (Associated bundle)**.** Let $P(M,G)$ be principal and let G act on a manifold F on the left. The *associated bundle* $P\times_G F$ is the quotient of $P\times F$ by the equivalence

$$(p,f)\sim\left(p\cdot g,\ g^{-1}f\right),\qquad g\in G\,, \tag{14.134}$$

with projection $[p,f]\mapsto\pi(p)$. It is a fibre bundle with typical fibre F and the same transition functions Equation (14.131). When F is a vector space and the action is a linear representation, $P\times_G F$ is a *vector bundle*, and a matter field is a section of it.

Rests on Definitions 14.95 and 14.97.

Remark 14.100 (Where the nontriviality is felt). The definitions above would be idle if every bundle were trivial, since a trivial bundle reduces every construction to a global choice of gauge. Two places in this treatise turn on their not being. A gauge-fixing prescription, such as the Faddeev–Popov construction of Section 106.4, is a section of a principal bundle over the space of gauge orbits, and Proposition 14.98 is the statement quoted in support of Theorem 106.65: for a non-abelian structure group that bundle is not trivial, so it admits no global section — no global section, no global gauge fixing, and the Faddeev–Popov prescription is a local construction used globally [1626]. And the connection of Definition 14.103 below is what makes the phase acquired around a closed path a property of the path rather than of the coordinates, the effect predicted in [1164], first measured by Chambers with an iron whisker behind an electron biprism [1178] and put beyond the leakage-field objection by Tonomura and collaborators, who enclosed the flux in a superconducting shield [1784]. The non-abelian case, in which the structure group is not U(1), was introduced for the isotopic-spin symmetry [1054].

14.6.2 Connection and curvature

Notation 14.101 (Algebra-valued forms and their bracket). A $\mathfrak{g}$-valued p-form on a manifold is $\boldsymbol{\alpha}=\alpha^a\boldsymbol{T}_a$ with α^a ordinary p-forms and $\{\boldsymbol{T}_a\}$ a basis of $\mathfrak{g}$. For a p-form $\boldsymbol{\alpha}$ and a q-form $\boldsymbol{\beta}$ put

$$[\boldsymbol{\alpha},\boldsymbol{\beta}]=\alpha^a\wedge\beta^b\,[\boldsymbol{T}_a,\boldsymbol{T}_b]\,, \tag{14.135}$$

a $\mathfrak{g}$-valued $(p+q)$-form, and $\mathrm{d}\boldsymbol{\alpha}=\mathrm{d}\alpha^a\,\boldsymbol{T}_a$. Expanding the wedge product of two 1-forms on a pair of vectors gives

$$[\boldsymbol{\alpha},\boldsymbol{\beta}]\,(V,W)=[\boldsymbol{\alpha}(V),\boldsymbol{\beta}(W)]-[\boldsymbol{\alpha}(W),\boldsymbol{\beta}(V)]\,. \tag{14.136}$$

Lemma 14.102 (Two identities for a $\mathfrak{g}$-valued 1-form)**.** *For every $\mathfrak{g}$-valued 1-form $\boldsymbol{\alpha}$,*

$$\mathrm{d}\,[\boldsymbol{\alpha},\boldsymbol{\alpha}]=2\,[\mathrm{d}\boldsymbol{\alpha},\boldsymbol{\alpha}]\,,\qquad [\boldsymbol{\alpha},[\boldsymbol{\alpha},\boldsymbol{\alpha}]]=0\,. \tag{14.137}$$

Rests on Equations (14.5) and (14.135).

Proof. *Derives Lemma 14.102.*

For the first, the graded Leibniz rule gives $\mathrm{d}\left(\alpha^a\wedge\alpha^b\right)=\mathrm{d}\alpha^a\wedge\alpha^b-\alpha^a\wedge\mathrm{d}\alpha^b$, and in the second term a 1-form and a 2-form commute, so $-\alpha^a\wedge\mathrm{d}\alpha^b\,[\boldsymbol{T}_a,\boldsymbol{T}_b]=-\mathrm{d}\alpha^b\wedge\alpha^a\,[\boldsymbol{T}_a,\boldsymbol{T}_b]=+\mathrm{d}\alpha^a\wedge\alpha^b\,[\boldsymbol{T}_a,\boldsymbol{T}_b]$ after renaming $a\leftrightarrow b$ and using the antisymmetry of the bracket; the two terms are equal. For the second,

$$[\boldsymbol{\alpha},[\boldsymbol{\alpha},\boldsymbol{\alpha}]]=\alpha^c\wedge\alpha^a\wedge\alpha^b\,C^d{}_{ab}C^e{}_{cd}\,\boldsymbol{T}_e\,,$$

and the 3-form $\alpha^c\wedge\alpha^a\wedge\alpha^b$ is totally antisymmetric in (c,a,b), so only the totally antisymmetric part of the coefficient survives. That part is the cyclic sum $C^d{}_{ab}C^e{}_{cd}+C^d{}_{bc}C^e{}_{ad}+C^d{}_{ca}C^e{}_{bd}$, which vanishes: it is the Jacobi identity for $\mathfrak{g}$ written in the structure constants of Equation (14.5). □

Definition 14.103 (Connection 1-form)**.** Let $P(M,G)$ be a principal bundle. For $X \in \mathfrak{g}$ the *fundamental vector field* $X^\sharp$ is the generator of the flow $p \mapsto p \cdot \mathrm{e}^{tX}$; it is tangent to the fibres, and since the action is free and transitive on each fibre, $X \mapsto X^\sharp_p$ is a linear isomorphism of $\mathfrak{g}$ onto the *vertical* subspace $V_p = \ker \mathrm{d}\pi_p$. A *connection* on P is a $\mathfrak{g}$-valued 1-form $\boldsymbol{A}$ on P with

$$\boxed{\boldsymbol{A}\left(X^\sharp\right) = X \quad \text{for all } X \in \mathfrak{g}\,, \qquad R_g^* \boldsymbol{A} = g^{-1}\boldsymbol{A}g \quad \text{for all } g \in G}\,, \tag{14.138}$$

where R_g is the right action Equation (14.133) and the second condition is written for a matrix group, the adjoint action in general. The *horizontal* subspace is $H_p = \ker \boldsymbol{A}_p$.

Rests on Definition 14.97 and Notation 14.101.

Proposition 14.104 (A connection splits the tangent space)**.** *$T_pP = V_p \oplus H_p$ for every $p \in P$, and the projections onto the two summands are $v \mapsto (\boldsymbol{A}(v))^\sharp$ and $v \mapsto hv = v - (\boldsymbol{A}(v))^\sharp$.*

Rests on Definition 14.103.

Proof. *Derives Proposition 14.104.*

For $v \in T_pP$ write $v = (\boldsymbol{A}(v))^\sharp + hv$ with hv defined by that equation. The first term is vertical by construction, and $\boldsymbol{A}(hv) = \boldsymbol{A}(v) - \boldsymbol{A}\left((\boldsymbol{A}(v))^\sharp\right) = \boldsymbol{A}(v) - \boldsymbol{A}(v) = 0$ by the first condition of Equation (14.138), so $hv \in H_p$. The sum is direct: a vector in $V_p \cap H_p$ is $X^\sharp$ for some X, and $0 = \boldsymbol{A}(X^\sharp) = X$ forces it to vanish. □

Definition 14.105 (Curvature 2-form)**.** The *curvature* of the connection $\boldsymbol{A}$ is the $\mathfrak{g}$-valued 2-form

$$\boxed{\boldsymbol{F} = \mathrm{d}\boldsymbol{A} + \frac{1}{2}\,[\boldsymbol{A},\boldsymbol{A}]}\,, \tag{14.139}$$

and the *covariant exterior derivative* of a $\mathfrak{g}$-valued form is $\mathrm{D} = \mathrm{d} + [\boldsymbol{A},\ \cdot\ ]$, the upright D being distinguished from the dimension D as in Notation 14.1.

Rests on Definition 14.103 and Notation 14.101.

Proposition 14.106 (The curvature is horizontal, and measures non-integrability)**.** *For all $v, w \in T_pP$,*

$$\boldsymbol{F}(v,w) = \mathrm{d}\boldsymbol{A}\,(hv, hw)\ . \tag{14.140}$$

In particular $\boldsymbol{F}$ vanishes whenever either argument is vertical, and for horizontal vector fields V, W,

$$\boldsymbol{F}(V,W) = -\boldsymbol{A}\,([V,W])\ , \tag{14.141}$$

so that $\boldsymbol{F} = 0$ if and only if the horizontal distribution is closed under the Lie bracket.

Rests on Definition 14.105 and Proposition 14.104.

Proof. *Derives Proposition 14.106.*

By Equation (14.136), $\frac{1}{2}\,[\boldsymbol{A},\boldsymbol{A}]\,(v,w) = [\boldsymbol{A}(v),\boldsymbol{A}(w)]$. By bilinearity it is enough to check Equation (14.140) when each argument is either horizontal or vertical, using Proposition 14.104, and the invariant formula $\mathrm{d}\boldsymbol{A}(V,W) = V\boldsymbol{A}(W) - W\boldsymbol{A}(V) - \boldsymbol{A}\,([V,W])$ for extensions to vector fields. That formula is the coordinate identity $\mathrm{d}\alpha(V,W) = (\partial_\mu\alpha_\nu - \partial_\nu\alpha_\mu)\,V^\mu W^\nu$ of Section 13.6.4, rearranged: the derivatives of the components of V and W cancel against the bracket term.

Both horizontal. Then $\boldsymbol{A}(v) = \boldsymbol{A}(w) = 0$, the bracket term vanishes and $hv = v$, $hw = w$, so both sides are $\mathrm{d}\boldsymbol{A}(v,w)$; extending to horizontal fields and using the invariant formula gives Equation (14.141).

Both vertical, $v = X^\sharp$, $w = Y^\sharp$. The bracket term is $[X,Y]$. For the fundamental fields $\left[X^\sharp, Y^\sharp\right] = [X,Y]^\sharp$, because $X \mapsto X^\sharp$ is the differential of a right action, and $\boldsymbol{A}(X^\sharp) = X$ is constant along P, so the invariant formula gives $\mathrm{d}\boldsymbol{A}(X^\sharp, Y^\sharp) = -\,[X,Y]$. The two cancel, and the right-hand side of Equation (14.140) vanishes because $hv = hw = 0$.

Mixed, v horizontal and $w = Y^\sharp$. The bracket term vanishes, since $\boldsymbol{A}(v) = 0$. Extend v to the horizontal lift V of a vector field on M, which is invariant under every R_g by the second condition of Equation (14.138); then $\left[Y^\sharp, V\right] = 0$, being the derivative along the flow of $Y^\sharp$ — that is, along the action of e^{tY} — of an invariant field. The invariant formula leaves $V\boldsymbol{A}(Y^\sharp) = V(Y) = 0$. Both sides vanish, the right-hand one because $hw = 0$. □

Theorem 14.107 (Bianchi identity)**.** *Every connection satisfies*

$$\boxed{\mathrm{D}\boldsymbol{F} = \mathrm{d}\boldsymbol{F} + [\boldsymbol{A}, \boldsymbol{F}] = 0}\,. \tag{14.142}$$

Rests on Definition 14.105 and Lemma 14.102.

Proof. *Derives Theorem 14.107.*

Differentiate Equation (14.139). Since the exterior derivative squares to zero, and by the first identity of Equation (14.137),

$$\mathrm{d}\boldsymbol{F} = \tfrac{1}{2}\,\mathrm{d}\,[\boldsymbol{A}, \boldsymbol{A}] = [\mathrm{d}\boldsymbol{A}, \boldsymbol{A}]\ .$$

On the other hand, by the second identity of Equation (14.137),

$$[\boldsymbol{A}, \boldsymbol{F}] = [\boldsymbol{A}, \mathrm{d}\boldsymbol{A}] + \tfrac{1}{2}\,[\boldsymbol{A}, [\boldsymbol{A}, \boldsymbol{A}]] = [\boldsymbol{A}, \mathrm{d}\boldsymbol{A}]\ .$$

Finally $[\boldsymbol{A}, \mathrm{d}\boldsymbol{A}] = A^a \wedge \mathrm{d}A^b\,[\boldsymbol{T}_a, \boldsymbol{T}_b] = \mathrm{d}A^b \wedge A^a\,[\boldsymbol{T}_a, \boldsymbol{T}_b] = -\,[\mathrm{d}\boldsymbol{A}, \boldsymbol{A}]$, moving the 1-form past the 2-form at no cost in sign and then renaming $a \leftrightarrow b$. The two contributions cancel. □

Remark 14.108 (The same objects on the base)*.* A local section $s : U \longrightarrow P$ (Definition 14.96) pulls the connection and its curvature back to $\mathfrak{g}$-valued forms $s^*\boldsymbol{A}$ and $s^*\boldsymbol{F}$ on the patch $U \subseteq M$; the first is what a physicist calls the gauge potential and the second the field strength. Pullback commutes with d and with the pointwise bracket Equation (14.135), so Equations (14.139) and (14.142) hold verbatim for the pulled-back forms. They depend on the section chosen, which is why a gauge potential is not a tensor on M, and by Proposition 14.98 a single section covering all of M exists only when the bundle is trivial. The sections that follow work on P, where no choice has to be made.

14.7 Invariant polynomials

Consider a principal bundle $P(M, G)$, where M is a D-dimensional differentiable manifold and G a Lie group. We denote the Lie algebra associated with G by $\mathfrak{g} = \mathrm{span}\,\{\boldsymbol{T}_A\}$, with $A = 1, \ldots, \dim \mathfrak{g}$. The possible linear combinations of the $\boldsymbol{T}_A$ form a vector space, which, when no ambiguity is possible, we shall also denote by $\mathfrak{g}$. Let, moreover, $[\ ] : G \to \mathrm{GL}(\mathfrak{g})$ be the linear representation of the Lie group G defined by $[g] : \boldsymbol{B} \mapsto [g]\boldsymbol{B} = g\boldsymbol{B}g^{-1}$, with $g \in G$ and $\boldsymbol{B} = B^A\boldsymbol{T}_A \in \mathfrak{g}$. Finally, let $S_r(\mathfrak{g}, \mathbb{R})$ be the space of real symmetric r-linear functionals, and let $P \in S_r(\mathfrak{g}, \mathbb{R})$, i.e. a mapping of the form

$$P : \underbrace{\mathfrak{g} \otimes \cdots \otimes \mathfrak{g}}_{r \text{ times}} \longrightarrow \mathbb{R} \tag{14.143}$$

satisfying the following axioms:

1. the mapping is an r-linear transformation, i.e.

$$P(\boldsymbol{B}_1, \ldots, a\boldsymbol{B} + b\boldsymbol{C}, \ldots, \boldsymbol{B}_r) = a\,P(\boldsymbol{B}_1, \ldots, \boldsymbol{B}, \ldots, \boldsymbol{B}_r) + b\,P(\boldsymbol{B}_1, \ldots, \boldsymbol{C}, \ldots, \boldsymbol{B}_r)\,, \tag{14.144}$$

2. the mapping is totally symmetric:

$$P(\boldsymbol{B}_1,\dots,\boldsymbol{B}_i,\dots,\boldsymbol{B}_j,\dots,\boldsymbol{B}_r) = P(\boldsymbol{B}_1,\dots,\boldsymbol{B}_j,\dots,\boldsymbol{B}_i,\dots,\boldsymbol{B}_r)\,, \quad \forall\, i,j\,. \tag{14.145}$$

Of all these totally symmetric polynomials, we shall concentrate on those that are invariant under the action induced by the representation [ ] of the Lie group G, which is enshrined in the following definition.

Definition 14.109 (G-invariant polynomial)**.** We say that $P \in S_r(\mathfrak{g},\mathbb{R})$ is a *G-invariant polynomial* if

$$\forall\, g \in G, \quad \forall\, \boldsymbol{B}_l \in \mathfrak{g},\ l = 1,\dots,r, \quad P\left([g]\boldsymbol{B}_1,\dots,[g]\boldsymbol{B}_r\right) = P\left(\boldsymbol{B}_1,\dots,\boldsymbol{B}_r\right)\,,$$

where $[g]\boldsymbol{B} = g\boldsymbol{B}g^{-1}$. We introduce, moreover, the notation $P\left(\boldsymbol{B}_1,\dots,\boldsymbol{B}_r\right) = \langle \boldsymbol{B}_1,\dots,\boldsymbol{B}_r\rangle$. Throughout this chapter the angle brackets denote this multilinear functional and nothing else; a linear span is written span. The brackets appear once before this definition, in the proof of Proposition 14.54, and there they already carry the meaning fixed here.

Rests on Equations (14.144) *and* (14.145)*.*

The polynomial extends to $\mathfrak{g}$-valued differential forms by multilinearity: if $\boldsymbol{B}_l = B_l^{A_l}\boldsymbol{T}_{A_l}$ with $B_l^{A_l}$ ordinary p_l-forms, then $\langle \boldsymbol{B}_1,\dots,\boldsymbol{B}_r\rangle = B_1^{A_1}\cdots B_r^{A_r}\,\langle \boldsymbol{T}_{A_1},\dots,\boldsymbol{T}_{A_r}\rangle$, where the product of the component forms is the wedge product.

Theorem 14.110. *Let $\langle \boldsymbol{B}_1,\dots,\boldsymbol{B}_r\rangle$ be a G-invariant polynomial evaluated on $\mathfrak{g}$-valued differential forms of degrees $p_1,\dots,p_r$ defined on the principal bundle $P(M,G)$.*

1. *One has*

$$\mathrm{d}\,\langle \boldsymbol{B}_1,\dots,\boldsymbol{B}_r\rangle = \sum_{l=1}^{r}(-1)^{p_1+\cdots+p_{l-1}}\,\langle \boldsymbol{B}_1,\dots,\boldsymbol{B}_{l-1},\mathrm{d}\boldsymbol{B}_l,\boldsymbol{B}_{l+1},\dots,\boldsymbol{B}_r\rangle\,, \tag{14.146}$$

$$\mathrm{D}\,\langle \boldsymbol{B}_1,\dots,\boldsymbol{B}_r\rangle = \sum_{l=1}^{r}(-1)^{p_1+\cdots+p_{l-1}}\,\langle \boldsymbol{B}_1,\dots,\boldsymbol{B}_{l-1},\mathrm{D}\boldsymbol{B}_l,\boldsymbol{B}_{l+1},\dots,\boldsymbol{B}_r\rangle\,, \tag{14.147}$$

   where d *is the exterior derivative and* $\mathrm{D} = \mathrm{d} + [\boldsymbol{A},\ \cdot\ ]$ *is the covariant exterior derivative, with $\boldsymbol{A}$ a connection defined on $P(M,G)$.*

2. *Let $\{\boldsymbol{T}_A\}$ be a basis of the Lie algebra $\mathfrak{g}$ associated with the Lie group G. One has*

$$\sum_{l=1}^{r} C_{AA_l}{}^{B}\left\langle \boldsymbol{T}_{A_1},\dots,\boldsymbol{T}_{A_{l-1}},\boldsymbol{T}_B,\boldsymbol{T}_{A_{l+1}},\dots,\boldsymbol{T}_{A_r}\right\rangle = 0\,. \tag{14.148}$$

3. *Let, moreover, $\boldsymbol{B} = B^A\boldsymbol{T}_A$ be a $\mathfrak{g}$-valued p-form. Then*

$$\sum_{l=1}^{r}(-1)^{p(p_1+\cdots+p_{l-1})}\,\langle \boldsymbol{B}_1,\dots,\boldsymbol{B}_{l-1},[\boldsymbol{B},\boldsymbol{B}_l],\boldsymbol{B}_{l+1},\dots,\boldsymbol{B}_r\rangle = 0\,. \tag{14.149}$$

4. *One has*

$$\mathrm{D}\,\langle \boldsymbol{B}_1,\dots,\boldsymbol{B}_r\rangle = \mathrm{d}\,\langle \boldsymbol{B}_1,\dots,\boldsymbol{B}_r\rangle\,. \tag{14.150}$$

Rests on Definition 14.109 and Equation (14.5)*.*

Proof. *Derives Theorem 14.110.*

We prove the four statements in turn.

1. Expanding $\boldsymbol{B}_1, \dots, \boldsymbol{B}_r$ in the basis $\{\boldsymbol{T}_A\}$ of the Lie algebra $\mathfrak{g}$ of the group G and using the linearity of the polynomial, we have $\mathrm{d}\,\langle \boldsymbol{B}_1, \dots, \boldsymbol{B}_r\rangle = \mathrm{d}\left(B_1^{A_1}\cdots B_r^{A_r}\right)\langle \boldsymbol{T}_{A_1}, \dots, \boldsymbol{T}_{A_r}\rangle$. By the graded Leibniz rule for the exterior derivative,

$$\begin{aligned}\mathrm{d}\left(B_1^{A_1}\cdots B_r^{A_r}\right) &= \mathrm{d}B_1^{A_1}\, B_2^{A_2}\cdots B_r^{A_r} + (-1)^{p_1} B_1^{A_1}\,\mathrm{d}B_2^{A_2}\, B_3^{A_3}\cdots B_r^{A_r}\\ &\quad + \cdots + (-1)^{p_1+\cdots+p_{r-1}} B_1^{A_1}\cdots B_{r-1}^{A_{r-1}}\,\mathrm{d}B_r^{A_r}\\ &= \sum_{l=1}^{r}(-1)^{p_1+\cdots+p_{l-1}} B_1^{A_1}\cdots B_{l-1}^{A_{l-1}}\,\mathrm{d}B_l^{A_l}\, B_{l+1}^{A_{l+1}}\cdots B_r^{A_r}\,,\end{aligned}$$

and hence

$$\begin{aligned}&\mathrm{d}\,\langle \boldsymbol{B}_1, \dots, \boldsymbol{B}_r\rangle\\ &= \sum_{l=1}^{r}(-1)^{p_1+\cdots+p_{l-1}} B_1^{A_1}\cdots B_{l-1}^{A_{l-1}}\,\mathrm{d}B_l^{A_l}\, B_{l+1}^{A_{l+1}}\cdots B_r^{A_r}\,\langle \boldsymbol{T}_{A_1}, \dots, \boldsymbol{T}_{A_r}\rangle\\ &= \sum_{l=1}^{r}(-1)^{p_1+\cdots+p_{l-1}}\,\langle \boldsymbol{B}_1, \dots, \boldsymbol{B}_{l-1}, \mathrm{d}\boldsymbol{B}_l, \boldsymbol{B}_{l+1}, \dots, \boldsymbol{B}_r\rangle\,.\end{aligned}$$

The expression Equation (14.147) for the covariant exterior derivative is proved by an entirely analogous procedure, yielding the claim.

2. From the definition of a G-invariant polynomial we have

$$\langle \boldsymbol{T}_{A_1}, \dots, \boldsymbol{T}_{A_r}\rangle = \langle [g]\boldsymbol{T}_{A_1}, \dots, [g]\boldsymbol{T}_{A_r}\rangle = \left\langle g\boldsymbol{T}_{A_1}g^{-1}, \dots, g\boldsymbol{T}_{A_r}g^{-1}\right\rangle\,, \tag{14.151}$$

where $g = \mathrm{e}^{t^A\boldsymbol{T}_A}$ is an element of the Lie group G. Furthermore, since the generators $\boldsymbol{T}_A$ do not depend on the parameters of the infinitesimal transformations, neither does the polynomial $\langle \boldsymbol{T}_{A_1}, \dots, \boldsymbol{T}_{A_r}\rangle$; therefore $\frac{\mathrm{d}}{\mathrm{d}t^A}\langle \boldsymbol{T}_{A_1}, \dots, \boldsymbol{T}_{A_r}\rangle\Big|_{t^A=0} = 0$, and from Equation (14.151) we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t^A}\left\langle g\boldsymbol{T}_{A_1}g^{-1}, \dots, g\boldsymbol{T}_{A_r}g^{-1}\right\rangle\Big|_{t^A=0} = 0\,. \tag{14.152}$$

From part 1 — analogously to the procedure used to prove Equation (14.146) — one obtains

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t^A}\left\langle g\boldsymbol{T}_{A_1}g^{-1}, \dots, g\boldsymbol{T}_{A_r}g^{-1}\right\rangle\\ &= \sum_{l=1}^{r}\left\langle g\boldsymbol{T}_{A_1}g^{-1}, \dots, g\boldsymbol{T}_{A_{l-1}}g^{-1}, \frac{\mathrm{d}}{\mathrm{d}t^A}\left(g\boldsymbol{T}_{A_l}g^{-1}\right), g\boldsymbol{T}_{A_{l+1}}g^{-1}, \dots, g\boldsymbol{T}_{A_r}g^{-1}\right\rangle\,,\end{aligned} \tag{14.153}$$

but $g\boldsymbol{T}_Ag^{-1}\big|_{t^A=0} = \mathrm{e}^{t^B\boldsymbol{T}_B}\boldsymbol{T}_A\,\mathrm{e}^{-t^C\boldsymbol{T}_C}\Big|_{t^A=0} = \boldsymbol{T}_A$, and moreover

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t^A}\left(g\boldsymbol{T}_Bg^{-1}\right)\Big|_{t^A=0} &= \left(\frac{\mathrm{d}}{\mathrm{d}t^A}\left(\mathrm{e}^{t^C\boldsymbol{T}_C}\right)\boldsymbol{T}_B\,\mathrm{e}^{-t^E\boldsymbol{T}_E} + \mathrm{e}^{t^C\boldsymbol{T}_C}\boldsymbol{T}_B\,\frac{\mathrm{d}}{\mathrm{d}t^A}\left(\mathrm{e}^{-t^E\boldsymbol{T}_E}\right)\right)\Big|_{t^A=0}\\ &= \left(\mathrm{e}^{t^C\boldsymbol{T}_C}\boldsymbol{T}_A\boldsymbol{T}_B\,\mathrm{e}^{-t^E\boldsymbol{T}_E} - \mathrm{e}^{t^C\boldsymbol{T}_C}\boldsymbol{T}_B\,\mathrm{e}^{-t^E\boldsymbol{T}_E}\boldsymbol{T}_A\right)\Big|_{t^A=0}\\ &= [\boldsymbol{T}_A, \boldsymbol{T}_B]\\ &= C_{AB}{}^C\,\boldsymbol{T}_C\,,\end{aligned}$$

so that we finally conclude from Equation (14.152) and Equation (14.153) that the claim holds.

3. In the basis $\{\boldsymbol{T}_A\}$ we have

$$\begin{aligned}
&\langle \boldsymbol{B}_1, \dots, \boldsymbol{B}_{l-1}, [\boldsymbol{B}, \boldsymbol{B}_l], \boldsymbol{B}_{l+1}, \dots, \boldsymbol{B}_r \rangle \\
&= B_1^{A_1} \cdots B_{l-1}^{A_{l-1}} B^A B_l^{A_l} B_{l+1}^{A_{l+1}} \cdots B_r^{A_r} \left\langle \boldsymbol{T}_{A_1}, \dots, \boldsymbol{T}_{A_{l-1}}, [\boldsymbol{T}_A, \boldsymbol{T}_{A_l}], \boldsymbol{T}_{A_{l+1}}, \dots, \boldsymbol{T}_{A_r} \right\rangle \\
&= (-1)^{p(p_1+\cdots+p_{l-1})} B^A B_1^{A_1} \cdots B_{l-1}^{A_{l-1}} B_l^{A_l} B_{l+1}^{A_{l+1}} \cdots B_r^{A_r} \\
&\quad \times \left\langle \boldsymbol{T}_{A_1}, \dots, \boldsymbol{T}_{A_{l-1}}, C_{AA_l}{}^B \boldsymbol{T}_B, \boldsymbol{T}_{A_{l+1}}, \dots, \boldsymbol{T}_{A_r} \right\rangle ,
\end{aligned}$$

where the sign arises from commuting the p-form components B^A past the forms $B_1^{A_1}, \dots, B_{l-1}^{A_{l-1}}$ of total degree $p_1 + \cdots + p_{l-1}$. Hence

$$\begin{aligned}
&(-1)^{p(p_1+\cdots+p_{l-1})} \langle \boldsymbol{B}_1, \dots, \boldsymbol{B}_{l-1}, [\boldsymbol{B}, \boldsymbol{B}_l], \boldsymbol{B}_{l+1}, \dots, \boldsymbol{B}_r \rangle \\
&= B^A B_1^{A_1} \cdots B_r^{A_r} C_{AA_l}{}^B \left\langle \boldsymbol{T}_{A_1}, \dots, \boldsymbol{T}_{A_{l-1}}, \boldsymbol{T}_B, \boldsymbol{T}_{A_{l+1}}, \dots, \boldsymbol{T}_{A_r} \right\rangle ,
\end{aligned}$$

so that, summing over all values of $l = 1, \dots, r$, part 2 yields the desired result.

4. Since $\mathrm{D} = \mathrm{d} + [\boldsymbol{A}, \ \cdot\ ]$, from part 1 we see that

$$\mathrm{D} \langle \boldsymbol{B}_1, \dots, \boldsymbol{B}_r \rangle = \mathrm{d} \langle \boldsymbol{B}_1, \dots, \boldsymbol{B}_r \rangle + \sum_{l=1}^{r} (-1)^{p_1+\cdots+p_{l-1}} \langle \boldsymbol{B}_1, \dots, \boldsymbol{B}_{l-1}, [\boldsymbol{A}, \boldsymbol{B}_l], \boldsymbol{B}_{l+1}, \dots, \boldsymbol{B}_r \rangle .$$

From part 3 with $\boldsymbol{B} = \boldsymbol{A}$ and $p = 1$ — since the connection is a 1-form — the second term vanishes.

□

14.8 The Chern–Weil theorem and transgression forms

Theorem 14.111 (Chern–Weil)**.** *Let $P(M, G)$ be a principal bundle equipped with a connection 1-form $\boldsymbol{A}$ of curvature $\boldsymbol{F} = \mathrm{d}\boldsymbol{A} + \frac{1}{2}[\boldsymbol{A}, \boldsymbol{A}]$, and let P be a G-invariant polynomial of degree r. Then:*

1. *$P(\boldsymbol{F}, \dots, \boldsymbol{F}) = \langle \boldsymbol{F}, \dots, \boldsymbol{F} \rangle$ is a closed form, i.e. $\mathrm{d} \langle \boldsymbol{F}, \dots, \boldsymbol{F} \rangle = 0$.*
2. *If $\boldsymbol{F}$ and $\bar{\boldsymbol{F}}$ are the curvature 2-forms corresponding to the connections $\boldsymbol{A}$ and $\bar{\boldsymbol{A}}$ respectively, both defined on $P(M, G)$, then $\langle \boldsymbol{F}, \dots, \boldsymbol{F} \rangle - \left\langle \bar{\boldsymbol{F}}, \dots, \bar{\boldsymbol{F}} \right\rangle$ is an exact form, i.e.*

$$\langle \boldsymbol{F}, \dots, \boldsymbol{F} \rangle - \left\langle \bar{\boldsymbol{F}}, \dots, \bar{\boldsymbol{F}} \right\rangle = \mathrm{d}Q^{(2r-1)}(\boldsymbol{A}, \bar{\boldsymbol{A}}) , \tag{14.154}$$

where the $(2r-1)$-form $Q^{(2r-1)}(\boldsymbol{A}, \bar{\boldsymbol{A}})$ is called the transgression form *of the polynomial P and is explicitly given by*

$$Q^{(2r-1)}(\boldsymbol{A}, \bar{\boldsymbol{A}}) = r \int_0^1 \left\langle \boldsymbol{A} - \bar{\boldsymbol{A}}, \boldsymbol{F}_t, \dots, \boldsymbol{F}_t \right\rangle \mathrm{d}t , \tag{14.155}$$

where $\boldsymbol{F}_t = \mathrm{d}\boldsymbol{A}_t + \frac{1}{2}[\boldsymbol{A}_t, \boldsymbol{A}_t]$ is the interpolating field strength *corresponding to the* interpolating connection *$\boldsymbol{A}_t$ defined as*

$$\boldsymbol{A}_t(\boldsymbol{A}, \bar{\boldsymbol{A}}) = \bar{\boldsymbol{A}} + t \left(\boldsymbol{A} - \bar{\boldsymbol{A}} \right) , \qquad 0 \le t \le 1 . \tag{14.156}$$

Rests on Theorem 14.110 and Equations (14.139) and (14.142).

Proof. *Derives Theorem 14.111.*

Theorem 14.110 is the principal tool.

1. Consider part 1 of Theorem 14.110, in its covariant form Equation (14.147), for $\boldsymbol{B}_1, \dots, \boldsymbol{B}_r = \boldsymbol{F}$, so that $p_1, \dots, p_r = 2$. We obtain

$$\begin{aligned} \mathrm{D}\left\langle \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle &= \sum_{l=1}^{r} (-1)^{2+\dots+2} \left\langle \boldsymbol{F}, \dots, \boldsymbol{F}, \mathrm{D}\boldsymbol{F}, \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle \\ &= \sum_{l=1}^{r} \left\langle \boldsymbol{F}, \dots, \boldsymbol{F}, \mathrm{D}\boldsymbol{F}, \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle \\ &= r \left\langle \boldsymbol{F}, \dots, \boldsymbol{F}, \mathrm{D}\boldsymbol{F}, \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle . \end{aligned} \tag{14.157}$$

   From the Bianchi identity we have $\mathrm{D}\boldsymbol{F} = \mathrm{d}\boldsymbol{F} + [\boldsymbol{A}, \boldsymbol{F}] = 0$, and $\langle \boldsymbol{F}, \dots, \boldsymbol{F}, 0, \boldsymbol{F}, \dots, \boldsymbol{F} \rangle = 0$ as a direct consequence of multilinearity:

$$\begin{aligned} \left\langle \boldsymbol{F}, \dots, \boldsymbol{F}, 0, \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle &= \left\langle \boldsymbol{F}, \dots, \boldsymbol{F}, \boldsymbol{B} - \boldsymbol{B}, \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle \\ &= \left\langle \boldsymbol{F}, \dots, \boldsymbol{F}, \boldsymbol{B}, \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle + \left\langle \boldsymbol{F}, \dots, \boldsymbol{F}, -\boldsymbol{B}, \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle \\ &= \left\langle \boldsymbol{F}, \dots, \boldsymbol{F}, \boldsymbol{B}, \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle - \left\langle \boldsymbol{F}, \dots, \boldsymbol{F}, \boldsymbol{B}, \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle \\ &= 0 . \end{aligned}$$

   Together with part 4 of Theorem 14.110, which gives $\mathrm{d}\left\langle \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle = \mathrm{D}\left\langle \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle = 0$, this proves the statement.

2. Write $\boldsymbol{\theta} = \boldsymbol{A} - \bar{\boldsymbol{A}}$, so that $\boldsymbol{A}_0 = \bar{\boldsymbol{A}}$ and $\boldsymbol{A}_1 = \bar{\boldsymbol{A}} + \boldsymbol{\theta} = \boldsymbol{A}$, and denote by $\bar{\mathrm{D}} = \mathrm{d} + \left[\bar{\boldsymbol{A}}, \ \cdot\ \right]$ the covariant exterior derivative of the connection $\bar{\boldsymbol{A}}$. The field strength corresponding to $\boldsymbol{A}_t$ is given by $\boldsymbol{F}_t = \mathrm{d}\boldsymbol{A}_t + \frac{1}{2}[\boldsymbol{A}_t, \boldsymbol{A}_t]$, so that $\boldsymbol{F}_0 = \bar{\boldsymbol{F}}$ and $\boldsymbol{F}_1 = \boldsymbol{F}$. We have

$$\left\langle \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle - \left\langle \bar{\boldsymbol{F}}, \dots, \bar{\boldsymbol{F}} \right\rangle = \left\langle \boldsymbol{F}_1, \dots, \boldsymbol{F}_1 \right\rangle - \left\langle \boldsymbol{F}_0, \dots, \boldsymbol{F}_0 \right\rangle = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} \left\langle \boldsymbol{F}_t, \dots, \boldsymbol{F}_t \right\rangle \mathrm{d}t .$$

   Analogously to the procedure used to obtain Equation (14.146), one finds

$$\left\langle \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle - \left\langle \bar{\boldsymbol{F}}, \dots, \bar{\boldsymbol{F}} \right\rangle = \int_0^1 \sum_{l=1}^{r} \left\langle \boldsymbol{F}_t, \dots, \boldsymbol{F}_t, \frac{\mathrm{d}\boldsymbol{F}_t}{\mathrm{d}t}, \boldsymbol{F}_t, \dots, \boldsymbol{F}_t \right\rangle \mathrm{d}t .$$

   Since the summand does not depend on the index l, and since a G-invariant polynomial is by definition totally symmetric, we obtain r equal terms, arriving at

$$\left\langle \boldsymbol{F}, \dots, \boldsymbol{F} \right\rangle - \left\langle \bar{\boldsymbol{F}}, \dots, \bar{\boldsymbol{F}} \right\rangle = r \int_0^1 \left\langle \frac{\mathrm{d}\boldsymbol{F}_t}{\mathrm{d}t}, \boldsymbol{F}_t, \dots, \boldsymbol{F}_t \right\rangle \mathrm{d}t .$$

   To compute the derivative of the interpolating field strength $\boldsymbol{F}_t$ with respect to t, observe that

$$\begin{aligned} \boldsymbol{F}_t &= \mathrm{d}\boldsymbol{A}_t + \frac{1}{2}[\boldsymbol{A}_t, \boldsymbol{A}_t] \\ &= \mathrm{d}\bar{\boldsymbol{A}} + t\,\mathrm{d}\boldsymbol{\theta} + \frac{1}{2}\left[\bar{\boldsymbol{A}} + t\boldsymbol{\theta}, \bar{\boldsymbol{A}} + t\boldsymbol{\theta}\right] \\ &= \mathrm{d}\bar{\boldsymbol{A}} + t\,\mathrm{d}\boldsymbol{\theta} + \frac{1}{2}\left[\bar{\boldsymbol{A}}, \bar{\boldsymbol{A}}\right] + \frac{t}{2}\left[\bar{\boldsymbol{A}}, \boldsymbol{\theta}\right] + \frac{t}{2}\left[\boldsymbol{\theta}, \bar{\boldsymbol{A}}\right] + \frac{t^2}{2}[\boldsymbol{\theta}, \boldsymbol{\theta}] \\ &= \mathrm{d}\bar{\boldsymbol{A}} + \frac{1}{2}\left[\bar{\boldsymbol{A}}, \bar{\boldsymbol{A}}\right] + t\,\mathrm{d}\boldsymbol{\theta} + t\left[\bar{\boldsymbol{A}}, \boldsymbol{\theta}\right] + \frac{t^2}{2}[\boldsymbol{\theta}, \boldsymbol{\theta}] , \end{aligned}$$

   that is,

$$\boldsymbol{F}_t = \bar{\boldsymbol{F}} + t\,\bar{\mathrm{D}}\boldsymbol{\theta} + \frac{t^2}{2}[\boldsymbol{\theta}, \boldsymbol{\theta}] , \tag{14.158}$$

so that $\frac{\mathrm{d}\boldsymbol{F}_t}{\mathrm{d}t} = \bar{\mathrm{D}}\boldsymbol{\theta} + t\,[\boldsymbol{\theta},\boldsymbol{\theta}]$, and thus

$$\langle \boldsymbol{F},\dots,\boldsymbol{F}\rangle - \left\langle \bar{\boldsymbol{F}},\dots,\bar{\boldsymbol{F}}\right\rangle = r\int_0^1 \left\langle \boldsymbol{F}_t,\dots,\boldsymbol{F}_t,\bar{\mathrm{D}}\boldsymbol{\theta} + t\,[\boldsymbol{\theta},\boldsymbol{\theta}],\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\right\rangle \mathrm{d}t\,.$$

Since the polynomial is totally symmetric we may write

$$\langle \boldsymbol{F},\dots,\boldsymbol{F}\rangle - \left\langle \bar{\boldsymbol{F}},\dots,\bar{\boldsymbol{F}}\right\rangle = r\int_0^1 \left\langle \bar{\mathrm{D}}\boldsymbol{\theta} + t\,[\boldsymbol{\theta},\boldsymbol{\theta}],\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\right\rangle \mathrm{d}t\,. \tag{14.159}$$

On the other hand, applying part 1 of Theorem 14.110 — for the connection $\bar{\boldsymbol{A}}$ — with $\boldsymbol{B}_1 = \boldsymbol{\theta}$ and $\boldsymbol{B}_2,\dots,\boldsymbol{B}_r = \boldsymbol{F}_t$, so that $p_1 = 1$ and $p_2,\dots,p_r = 2$, together with part 4, we observe that

$$\begin{aligned}
\mathrm{d}\,\langle \boldsymbol{\theta},\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\rangle &= \bar{\mathrm{D}}\,\langle \boldsymbol{\theta},\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\rangle \\
&= \left\langle \bar{\mathrm{D}}\boldsymbol{\theta},\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\right\rangle + \sum_{l=2}^{r}(-1)^{1+2+\cdots+2}\left\langle \boldsymbol{\theta},\boldsymbol{F}_t,\dots,\boldsymbol{F}_t,\bar{\mathrm{D}}\boldsymbol{F}_t,\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\right\rangle \\
&= \left\langle \bar{\mathrm{D}}\boldsymbol{\theta},\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\right\rangle - (r-1)\left\langle \boldsymbol{\theta},\bar{\mathrm{D}}\boldsymbol{F}_t,\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\right\rangle\,.
\end{aligned} \tag{14.160}$$

Recall that the difference of two connections is a tensor, and that a connection plus a tensor defines a new connection. In this way $\boldsymbol{A}_t$ determines a connection on the bundle, to which corresponds a covariant derivative $\mathrm{D}_{\boldsymbol{A}_t} = \mathrm{d} + [\boldsymbol{A}_t,\ \cdot\ ]$. For the field strength $\boldsymbol{F}_t$ we have the Bianchi identity $\mathrm{D}_{\boldsymbol{A}_t}\boldsymbol{F}_t = 0$, which implies

$$\begin{aligned}
\mathrm{d}\boldsymbol{F}_t + [\boldsymbol{A}_t,\boldsymbol{F}_t] &= 0\,,\\
\mathrm{d}\boldsymbol{F}_t + \left[\bar{\boldsymbol{A}} + t\boldsymbol{\theta},\boldsymbol{F}_t\right] &= 0\,,\\
\mathrm{d}\boldsymbol{F}_t + \left[\bar{\boldsymbol{A}},\boldsymbol{F}_t\right] + t\,[\boldsymbol{\theta},\boldsymbol{F}_t] &= 0\,,\\
\Longrightarrow\ \bar{\mathrm{D}}\boldsymbol{F}_t &= -t\,[\boldsymbol{\theta},\boldsymbol{F}_t]\ ,
\end{aligned}$$

so that

$$\mathrm{d}\,\langle \boldsymbol{\theta},\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\rangle = \left\langle \bar{\mathrm{D}}\boldsymbol{\theta},\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\right\rangle + t(r-1)\,\langle \boldsymbol{\theta},[\boldsymbol{\theta},\boldsymbol{F}_t],\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\rangle\ . \tag{14.161}$$

Finally, applying part 3 of Theorem 14.110 with $\boldsymbol{B} = \boldsymbol{\theta}$, $\boldsymbol{B}_1 = \boldsymbol{\theta}$ and $\boldsymbol{B}_2,\dots,\boldsymbol{B}_r = \boldsymbol{F}_t$, so that $p = 1$, $p_1 = 1$ and $p_2,\dots,p_r = 2$, we obtain

$$\begin{aligned}
\langle [\boldsymbol{\theta},\boldsymbol{\theta}],\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\rangle + \sum_{l=2}^{r}(-1)^{1\cdot(1+2+\cdots+2)}\,\langle \boldsymbol{\theta},\boldsymbol{F}_t,\dots,\boldsymbol{F}_t,[\boldsymbol{\theta},\boldsymbol{F}_t],\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\rangle &= 0\,,\\
\Longrightarrow\ \langle [\boldsymbol{\theta},\boldsymbol{\theta}],\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\rangle - (r-1)\,\langle \boldsymbol{\theta},[\boldsymbol{\theta},\boldsymbol{F}_t],\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\rangle &= 0\,,
\end{aligned}$$

from which we see that Equation (14.161) becomes $\mathrm{d}\,\langle \boldsymbol{\theta},\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\rangle = \left\langle \bar{\mathrm{D}}\boldsymbol{\theta} + t\,[\boldsymbol{\theta},\boldsymbol{\theta}],\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\right\rangle$. Substituting this result into Equation (14.159), we obtain

$$\begin{aligned}
\langle \boldsymbol{F},\dots,\boldsymbol{F}\rangle - \left\langle \bar{\boldsymbol{F}},\dots,\bar{\boldsymbol{F}}\right\rangle &= r\int_0^1 \mathrm{d}\,\langle \boldsymbol{\theta},\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\rangle\ \mathrm{d}t\\
&= \mathrm{d}\left(r\int_0^1 \left\langle \boldsymbol{A}-\bar{\boldsymbol{A}},\boldsymbol{F}_t,\dots,\boldsymbol{F}_t\right\rangle \mathrm{d}t\right)\\
&= \mathrm{d}Q^{(2r-1)}(\boldsymbol{A},\bar{\boldsymbol{A}})\,.
\end{aligned}$$

□

14.8.1 Chern–Simons forms

Definition 14.112 (Chern–Simons form)**.** Let $P(\boldsymbol{F}) = \langle \boldsymbol{F}, \dots, \boldsymbol{F} \rangle$ be a G-invariant polynomial of degree r. By the Chern–Weil theorem we know that $\langle \boldsymbol{F}, \dots, \boldsymbol{F} \rangle$ is a closed form, and it can therefore be written locally as $\langle \boldsymbol{F}, \dots, \boldsymbol{F} \rangle = \mathrm{d}Q^{(2r-1)}(\boldsymbol{A})$. The $(2r-1)$-form $Q^{(2r-1)}(\boldsymbol{A})$ is called the *Chern–Simons form* corresponding to the polynomial P.

Rests on Theorem 14.111 and Definition 14.109.

By the Chern–Weil theorem we have

$$\mathrm{d}Q^{(2r-1)}(\boldsymbol{A}) - \mathrm{d}Q^{(2r-1)}(\bar{\boldsymbol{A}}) = \mathrm{d}\left(r \int_0^1 \mathrm{d}t \left\langle \boldsymbol{A} - \bar{\boldsymbol{A}}, \boldsymbol{F}_t, \dots, \boldsymbol{F}_t \right\rangle \right) ,$$

so that for $\bar{\boldsymbol{A}} = 0$ one has $Q^{(2r-1)}(\bar{\boldsymbol{A}}) = 0$ and thus

$$Q^{(2r-1)}(\boldsymbol{A}) = r \int_0^1 \mathrm{d}t \, \langle \boldsymbol{A}, \boldsymbol{F}_t, \dots, \boldsymbol{F}_t \rangle \, . \tag{14.162}$$

However, an important warning must be made: since the coordinate transformation law of a connection is inhomogeneous, it is impossible to choose a connection that vanishes globally. Note, moreover, that locally a transgression form with $\bar{\boldsymbol{A}} = 0$ corresponds to a Chern–Simons form (which is, of course, defined locally):

$$Q^{(2r-1)}(\boldsymbol{A}, 0) = Q^{(2r-1)}(\boldsymbol{A}) \, . \tag{14.163}$$

When $\bar{\boldsymbol{A}} = 0$, the interpolating connection is given by $\boldsymbol{A}_t = t\boldsymbol{A}$, with $0 \leq t \leq 1$, so that $\boldsymbol{A}_0 = 0$ and $\boldsymbol{A}_1 = \boldsymbol{A}$. The interpolating field strength corresponding to $\boldsymbol{A}_t$ is given by

$$\begin{aligned} \boldsymbol{F}_t &= \mathrm{d}\boldsymbol{A}_t + \frac{1}{2} [\boldsymbol{A}_t, \boldsymbol{A}_t] \\ &= t \, \mathrm{d}\boldsymbol{A} + \frac{t^2}{2} [\boldsymbol{A}, \boldsymbol{A}] \\ &= t \left(\boldsymbol{F} - \frac{1}{2} [\boldsymbol{A}, \boldsymbol{A}] \right) + \frac{t^2}{2} [\boldsymbol{A}, \boldsymbol{A}] \\ &= t\boldsymbol{F} + \frac{1}{2} t(t-1) [\boldsymbol{A}, \boldsymbol{A}] \, . \end{aligned}$$

From Equation (14.155) we then find that the sought Chern–Simons form is

$$Q^{(2r-1)}(\boldsymbol{A}) = r \int_0^1 \mathrm{d}t \, t^{r-1} \left\langle \boldsymbol{A}, \boldsymbol{F} + \frac{1}{2}(t-1) [\boldsymbol{A}, \boldsymbol{A}], \dots, \boldsymbol{F} + \frac{1}{2}(t-1) [\boldsymbol{A}, \boldsymbol{A}] \right\rangle . \tag{14.164}$$

The source manuscript stops at Equation (14.164). The three things it was evidently going to do next — evaluate the Chern–Simons forms in the degrees that occur, name the invariant polynomials whose Chern–Weil forms are the characteristic classes, and apply the construction to an $\mathrm{SO}(p, q)$ connection — are done here, and the chapter closes with them.

Proposition 14.113 (The Chern–Simons forms of degree 3 and 5)**.** *For $r = 2$ and $r = 3$, Equation (14.164) evaluates to*

$$Q^{(3)}(\boldsymbol{A}) = \langle \boldsymbol{A}, \boldsymbol{F} \rangle - \tfrac{1}{6} \langle \boldsymbol{A}, [\boldsymbol{A}, \boldsymbol{A}] \rangle \, , \tag{14.165}$$

$$Q^{(5)}(\boldsymbol{A}) = \langle \boldsymbol{A}, \boldsymbol{F}, \boldsymbol{F} \rangle - \tfrac{1}{4} \langle \boldsymbol{A}, \boldsymbol{F}, [\boldsymbol{A}, \boldsymbol{A}] \rangle + \tfrac{1}{40} \langle \boldsymbol{A}, [\boldsymbol{A}, \boldsymbol{A}], [\boldsymbol{A}, \boldsymbol{A}] \rangle \, . \tag{14.166}$$

When the invariant polynomial is the trace of a product of matrices in some representation, $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \mathrm{tr}(\boldsymbol{XY})$, Equation (14.165) takes the familiar form

$$\boxed{Q^{(3)}(\boldsymbol{A}) = \mathrm{tr} \left(\boldsymbol{A} \wedge \mathrm{d}\boldsymbol{A} + \tfrac{2}{3} \boldsymbol{A} \wedge \boldsymbol{A} \wedge \boldsymbol{A} \right)} \, . \tag{14.167}$$

Rests on Equation (14.164) and Definitions 14.109 and 14.112.

Proof. *Derives Proposition 14.113.*

Abbreviate $\boldsymbol{G}_t = \boldsymbol{F} + \frac{1}{2}(t-1)\,[\boldsymbol{A},\boldsymbol{A}]$, which is the entry appearing in every slot but the first of Equation (14.164). For $r=2$ there is one such slot, and linearity gives

$$Q^{(3)} = 2\int_0^1 \mathrm{d}t\, t\,\langle \boldsymbol{A}, \boldsymbol{G}_t\rangle = 2\,\langle \boldsymbol{A},\boldsymbol{F}\rangle \int_0^1 t\,\mathrm{d}t + \langle \boldsymbol{A},[\boldsymbol{A},\boldsymbol{A}]\rangle \int_0^1 t(t-1)\,\mathrm{d}t\,,$$

and the two integrals are $1/2$ and $1/3 - 1/2 = -1/6$, which is Equation (14.165). For $r=3$ there are two such slots, and by total symmetry

$$\langle \boldsymbol{A},\boldsymbol{G}_t,\boldsymbol{G}_t\rangle = \langle \boldsymbol{A},\boldsymbol{F},\boldsymbol{F}\rangle + (t-1)\,\langle \boldsymbol{A},\boldsymbol{F},[\boldsymbol{A},\boldsymbol{A}]\rangle + \tfrac{1}{4}(t-1)^2\,\langle \boldsymbol{A},[\boldsymbol{A},\boldsymbol{A}],[\boldsymbol{A},\boldsymbol{A}]\rangle\,,$$

so that the three coefficients of Equation (14.166) are $3\int_0^1 t^2\mathrm{d}t = 1$, $3\int_0^1 t^2(t-1)\mathrm{d}t = 3\left(\frac{1}{4}-\frac{1}{3}\right) = -\frac{1}{4}$ and $\frac{3}{4}\int_0^1 t^2(t-1)^2\mathrm{d}t = \frac{3}{4}\left(\frac{1}{5}-\frac{1}{2}+\frac{1}{3}\right) = \frac{1}{40}$.

For Equation (14.167), matrix-valued 1-forms obey $[\boldsymbol{A},\boldsymbol{A}] = 2\,\boldsymbol{A}\wedge\boldsymbol{A}$ by Equation (14.135), so $\langle \boldsymbol{A},\boldsymbol{F}\rangle = \operatorname{tr}(\boldsymbol{A}\wedge\mathrm{d}\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{A}\wedge\boldsymbol{A}\wedge\boldsymbol{A})$ and $\frac{1}{6}\langle \boldsymbol{A},[\boldsymbol{A},\boldsymbol{A}]\rangle = \frac{1}{3}\operatorname{tr}(\boldsymbol{A}\wedge\boldsymbol{A}\wedge\boldsymbol{A})$. □

14.8.2 Characteristic classes

The Chern–Weil forms of Theorem 14.111 live on the total space P. They are forms on the base as well, and that is what makes them invariants of M rather than of a choice of gauge.

Proposition 14.114 (Chern–Weil forms descend to the base)**.** *Let $\boldsymbol{A}$ be a connection on $P(M,G)$ with curvature $\boldsymbol{F}$, and let $\langle\,\cdot\,\rangle$ be a G-invariant polynomial of degree r. Then, for a matrix group,*

$$R_g^*\boldsymbol{F} = g^{-1}\boldsymbol{F}g\,, \tag{14.168}$$

and $\langle \boldsymbol{F},\dots,\boldsymbol{F}\rangle$ is the pullback $\pi^*\rho$ of a unique closed $2r$-form ρ on M.

Rests on Theorem 14.111, Proposition 14.106, and Definition 14.109.

Proof. *Derives Proposition 14.114.*

Pullback commutes with d and with the bracket Equation (14.135), and g is a fixed group element, so

$$R_g^*\boldsymbol{F} = \mathrm{d}\left(g^{-1}\boldsymbol{A}g\right) + \tfrac{1}{2}\left[g^{-1}\boldsymbol{A}g, g^{-1}\boldsymbol{A}g\right] = g^{-1}\left(\mathrm{d}\boldsymbol{A} + \tfrac{1}{2}\,[\boldsymbol{A},\boldsymbol{A}]\right)g\,,$$

using Equation (14.138) in the first step. Hence $R_g^*\,\langle \boldsymbol{F},\dots,\boldsymbol{F}\rangle = \langle g^{-1}\boldsymbol{F}g,\dots,g^{-1}\boldsymbol{F}g\rangle = \langle \boldsymbol{F},\dots,\boldsymbol{F}\rangle$ by Definition 14.109: the form is invariant. It is also horizontal, being a wedge of the component forms of $\boldsymbol{F}$, each of which vanishes as soon as one argument is vertical by Proposition 14.106.

An invariant horizontal form descends. Define $\rho_x(v_1,\dots,v_{2r}) = \langle \boldsymbol{F},\dots,\boldsymbol{F}\rangle_p(\tilde v_1,\dots,\tilde v_{2r})$ for any $p\in\pi^{-1}(x)$ and any $\tilde v_i$ with $\mathrm{d}\pi\,\tilde v_i = v_i$. Two lifts at the same p differ by vertical vectors, which horizontality kills, so the value does not depend on the choice; two points of the same fibre differ by some g, and invariance makes the two values agree. Smoothness follows by computing ρ on a local section, $\rho|_U = s^*\,\langle \boldsymbol{F},\dots,\boldsymbol{F}\rangle$. Uniqueness holds because $\mathrm{d}\pi$ is onto. Finally $\mathrm{d}\rho$ pulls back to $\mathrm{d}\,\langle \boldsymbol{F},\dots,\boldsymbol{F}\rangle = 0$ by Theorem 14.111, and π^* is injective, so ρ is closed. □

Definition 14.115 (Characteristic forms)**.** Let G act by matrices in a representation of dimension N, and define the polynomials σ_r on $\mathfrak{g}$ by

$$\det(\mathbb{1} + \lambda\boldsymbol{X}) = \sum_{r=0}^{N} \lambda^r\,\sigma_r(\boldsymbol{X})\,, \tag{14.169}$$

and, for $\dim = 2m$ and $\boldsymbol{X}$ antisymmetric, the Pfaffian

$$\mathrm{Pf}(\boldsymbol{X}) = \frac{1}{2^m m!}\,\epsilon^{a_1\cdots a_{2m}} X_{a_1a_2}\cdots X_{a_{2m-1}a_{2m}}\,. \tag{14.170}$$

By Lemma 14.116 these are G-invariant polynomials, so Theorem 14.111 and Proposition 14.114 attach to each of them a closed form on M whose de Rham class does not depend on the connection. The *Chern* forms are those of σ_r for G unitary, evaluated at $\boldsymbol{X} = \mathrm{i}\boldsymbol{F}/2\pi$; the *Pontryagin* forms are those of σ_r for G orthogonal, evaluated at $\boldsymbol{X} = \boldsymbol{F}/2\pi$, only the even r surviving; and the *Euler* form is that of the Pfaffian, for $G = \mathrm{SO}(2m)$.

Rests on Definition 14.109, Theorem 14.111, and Proposition 14.114.

Lemma 14.116 (Invariance of the characteristic coefficients)**.** *The σ_r of Equation (14.169) are G-invariant polynomials of degree r in the sense of Definition 14.109, and σ_r vanishes identically on antisymmetric matrices for odd r. The Pfaffian Equation (14.170) obeys $\mathrm{Pf}\left(g\boldsymbol{X}g^{\mathsf{T}}\right) = \det(g)\,\mathrm{Pf}(\boldsymbol{X})$ for orthogonal g, so it is invariant under $\mathrm{SO}(2m)$ and changes sign under an orientation-reversing orthogonal transformation; the Euler form therefore belongs to an oriented bundle.*

Rests on Definition 14.109.

Proof. *Derives Lemma 14.116.*

Invariance is $\det\left(\mathbb{1} + \lambda\, g\boldsymbol{X}g^{-1}\right) = \det\left(g\left(\mathbb{1} + \lambda\boldsymbol{X}\right)g^{-1}\right) = \det\left(\mathbb{1} + \lambda\boldsymbol{X}\right)$, true coefficient by coefficient in λ; each σ_r is a homogeneous polynomial of degree r in the entries of $\boldsymbol{X}$, and the associated symmetric multilinear form is its polarization. For antisymmetric $\boldsymbol{X}$,

$$\det\left(\mathbb{1} + \lambda\boldsymbol{X}\right) = \det\left(\left(\mathbb{1} + \lambda\boldsymbol{X}\right)^{\mathsf{T}}\right) = \det\left(\mathbb{1} - \lambda\boldsymbol{X}\right)\,,$$

so the polynomial is even in λ and the odd σ_r vanish. For the Pfaffian, an orthogonal g has $g\boldsymbol{X}g^{-1} = g\boldsymbol{X}g^{\mathsf{T}}$, each factor in Equation (14.170) carries two of the g's, and

$$\epsilon^{b_1\cdots b_{2m}}\, g^{a_1}{}_{b_1}\cdots g^{a_{2m}}{}_{b_{2m}} = \det(g)\,\epsilon^{a_1\cdots a_{2m}}\,,$$

which is the definition of the determinant. The stated transformation follows, and $\det g = \pm 1$ for an orthogonal matrix. □

14.8.3 The transgression of an $\mathfrak{so}(p,q)$ connection

Proposition 14.117 (The quadratic invariant and its Chern–Simons form)**.** *Let $\boldsymbol{\omega} = \frac{1}{2}\omega^{AB}J_{AB}$ be a connection whose structure group is $\mathrm{SO}(p,q)$, taken in the defining representation of Equation (14.23), and let $\boldsymbol{R} = \frac{1}{2}R^{AB}J_{AB}$ be its curvature. With the invariant polynomial $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \mathrm{tr}(\boldsymbol{X}\boldsymbol{Y})$,*

$$\langle \boldsymbol{R}, \boldsymbol{R}\rangle = -R^{AB}\wedge R_{AB} = -8\pi^2\,\sigma_2\left(\boldsymbol{R}/2\pi\right)\,, \tag{14.171}$$

that is, $-8\pi^2$ times the first Pontryagin form $\sigma_2(\boldsymbol{R}/2\pi)$ of Definition 14.115, the invariant $\langle \boldsymbol{R}, \boldsymbol{R}\rangle$ itself carrying none of the 2π normalization of the characteristic forms. Its Chern–Simons 3-form is

$$\boxed{Q^{(3)}(\boldsymbol{\omega}) = -\omega^{AB}\wedge \mathrm{d}\omega_{AB} + \tfrac{2}{3}\,\omega^{A}{}_{B}\wedge\omega^{B}{}_{C}\wedge\omega^{C}{}_{A}}\,, \tag{14.172}$$

with $\mathrm{d}Q^{(3)}(\boldsymbol{\omega}) = \langle \boldsymbol{R}, \boldsymbol{R}\rangle$.

Rests on Proposition 14.113, Equation (14.86), and Definition 14.109.

Proof. *Derives Proposition 14.117.*

In the defining representation the generators are the matrices $(J_{AB})^{M}{}_{N} = \delta^M_A\eta_{BN} - \delta^M_B\eta_{AN}$ already used in Proposition 14.22, so

$$\mathrm{tr}\left(J_{AB}J_{CD}\right) = \left(\delta^M_A\eta_{BN} - \delta^M_B\eta_{AN}\right)\left(\delta^N_C\eta_{DM} - \delta^N_D\eta_{CM}\right) = 2\left(\eta_{AD}\eta_{BC} - \eta_{AC}\eta_{BD}\right)\,,$$

which is symmetric under the exchange of the index pairs and invariant, being a trace of a product of generators. Contracting,

$$\langle \boldsymbol{R}, \boldsymbol{R} \rangle = \tfrac{1}{4} R^{AB} \wedge R^{CD} \, \mathrm{tr}\,(J_{AB} J_{CD}) = \tfrac{1}{2} \left(R^{AB} \wedge R_{BA} - R^{AB} \wedge R_{AB} \right) = -R^{AB} \wedge R_{AB} \,,$$

the two 2-forms commuting. For the second equality in Equation (14.171): the elements of $\mathfrak{so}(p,q)$ are traceless, so $\sigma_1(\boldsymbol{X}) = \mathrm{tr}\,\boldsymbol{X}$ vanishes, and expanding Equation (14.169) to second order gives $\sigma_2(\boldsymbol{X}) = \frac{1}{2}\left[(\mathrm{tr}\,\boldsymbol{X})^2 - \mathrm{tr}(\boldsymbol{X}\boldsymbol{X})\right] = -\frac{1}{2}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{X})$; at $\boldsymbol{X} = \boldsymbol{R}/2\pi$, the value prescribed for the Pontryagin forms in Definition 14.115, this is $-\langle \boldsymbol{R}, \boldsymbol{R} \rangle / 8\pi^2$. The same substitution in the matrix components gives $(\boldsymbol{\omega})^M{}_N = \omega^M{}_N$, so Equation (14.167) reads $Q^{(3)} = \omega^M{}_N \wedge \mathrm{d}\omega^N{}_M + \frac{2}{3}\omega^M{}_N \wedge \omega^N{}_P \wedge \omega^P{}_M$, and lowering the indices in the first term turns $\omega^M{}_N \wedge \mathrm{d}\omega^N{}_M$ into $-\omega^{AB} \wedge \mathrm{d}\omega_{AB}$. That $\mathrm{d}Q^{(3)}$ is $\langle \boldsymbol{R}, \boldsymbol{R} \rangle$ is Definition 14.112. □

Remark 14.118 (Where this ends, and why). Equation (14.172) is where the source manuscript was heading and is the natural end of the chapter, because it is the last construction this treatise uses. The physical use of the invariant polynomials is made in Part XI, where the obstruction to a classical symmetry surviving quantization is an anomaly (Theorem 106.94); the measured consequence quoted in this chapter is the two-photon decay rate of the neutral pion, with its evidence, in Remark 14.82. The gauge connection of Definition 14.103 is the object every chapter of Part XI works with.

Three directions are deliberately not taken. The classification of bundles over a given base, and the reading of characteristic classes as integers attached to a topology, is algebraic topology and nothing in this book requires it. Index theorems, which relate those integers to the spectra of the operators of Chapter 12, are likewise outside what the evidence under discussion needs. And Chern–Simons *actions* — as opposed to Chern–Simons forms — are field theories in three spacetime dimensions, which the scope rule of Chapter 1 keeps out of this treatise: Remark 14.94 records the position. The mathematics is here because the anomaly and the gauge principle need it, and stops where they stop.

Chapter 15

Expansions of Lie Algebras

Chapter 14 left two mechanisms side by side without saying how they are related. Section 14.4 classified the central charges an algebra may carry by the cohomology group $H^2(\mathfrak{g}, \mathbb{R})$; Section 14.5 described how one algebra degenerates into another in a physical limit, and Theorem 14.90 listed the eleven algebras the degeneration produces from the kinematical hypotheses. What connects them is an observation, not a theorem: contraction *creates* the cohomology that makes a central charge possible (Proposition 14.84), and the mass of nonrelativistic mechanics is the standard example (Remark 25.15). The observation is correct but structurally unsatisfying, because it leaves the central charge looking like an accident of a limit rather than a part of the same construction.

This chapter supplies the missing structure. The whole family — the de Sitter and anti-de Sitter algebras, the Poincaré algebra, the kinematical algebras of Table 14.1, their central extensions, and the towers above them — is obtained from a *single* Lie algebra $\mathfrak{g}$ by tensoring it with a commutative associative algebra A. Changing A changes the resulting Lie algebra; degenerating A performs the contraction; enlarging A by one level produces the central extension. One datum, three operations that were previously three separate constructions.

Notation 15.1 (Standing conventions of this chapter). Throughout, $\mathfrak{g}$ is a finite-dimensional real Lie algebra and A is a finite-dimensional *commutative associative unital* $\mathbb{R}$-algebra, with unit 1_A and Jacobson radical $\mathrm{rad}(A)$. The symbol $\otimes$ means $\otimes_{\mathbb{R}}$, and $\langle v_1, \dots, v_n \rangle$ denotes the linear span of the listed generators. The spacetime dimension D is left general in the general statements; Section 15.10 instantiates every one of them at the observed $D = 4$, and each place where a claim holds *only* in $D = 3$ is marked as such where it occurs.

Remark 15.2 (Status of the material). The construction of this chapter, the classification of Theorem 15.13, and the identification of central charges with the top level of a truncation (Theorem 15.34 and Corollary 15.36) follow an unpublished manuscript in preparation by one of the authors of this treatise and a collaborator [2434]. That is stated here rather than buried in a citation: the reader cannot obtain the manuscript, so nothing in this chapter may rest on its authority. Every statement is therefore proved in full below or in Appendix A.11, from the definitions given here; where an argument is standard commutative algebra it is marked as such and is not used for any claim about a specific algebra. The published antecedents — the Maurer–Cartan expansion [2044] and the semigroup expansion [2096] — are recovered as special cases in Section 15.9, and a reader who wants a citable source for those special cases should use them.

15.1 The tensor product construction

Definition 15.3 (Tensor product Lie algebra)**.** Let $\mathfrak{g}$ be a Lie algebra and A a commutative associative unital algebra. The *tensor product Lie algebra* is the vector space

$$\boxed{\mathfrak{g}_A := A \otimes \mathfrak{g}} \tag{15.1}$$

equipped with the bracket determined by

$$\boxed{[a \otimes X, b \otimes Y] := (ab) \otimes [X, Y]} \tag{15.2}$$

for $a, b \in A$ and $X, Y \in \mathfrak{g}$, extended bilinearly.

Theorem 15.4 (The construction produces a Lie algebra)**.** *Equation* (15.2) *is well defined on $A \otimes \mathfrak{g}$ and makes $\mathfrak{g}_A$ a Lie algebra of dimension* $\dim \mathfrak{g}_A = (\dim A)(\dim \mathfrak{g})$*. In a basis $\{e_\alpha\}$ of A with $e_\alpha e_\beta = \mu_{\alpha\beta}{}^{\gamma} e_\gamma$ and a basis $\{T_a\}$ of $\mathfrak{g}$ with $[T_a, T_b] = C_{ab}{}^{c} T_c$, the structure constants* factorise*:*

$$\boxed{[e_\alpha \otimes T_a, e_\beta \otimes T_b] = \mu_{\alpha\beta}{}^{\gamma}\, C_{ab}{}^{c}\, e_\gamma \otimes T_c}\,. \tag{15.3}$$

Rests on Definition 15.3.

Derivation. *Derives Theorem 15.4.*

Well-definedness. The assignment $(a, X, b, Y) \mapsto (ab) \otimes [X, Y]$ is linear in each of its four arguments, because multiplication in A and the bracket of $\mathfrak{g}$ are bilinear. A multilinear map factors uniquely through the tensor product in each pair of arguments, so it descends to a bilinear map $\mathfrak{g}_A \times \mathfrak{g}_A \to \mathfrak{g}_A$, which is Equation (15.2). Substituting basis elements gives Equation (15.3), and the dimension count is that of a tensor product of vector spaces.

Antisymmetry. Using commutativity of A and antisymmetry of the bracket of $\mathfrak{g}$,

$$[a \otimes X, b \otimes Y] = (ab) \otimes [X, Y] = (ba) \otimes (-[Y, X]) = -[b \otimes Y, a \otimes X]\,.$$

Both hypotheses are used, and only these: associativity plays no part here.

Jacobi identity. Applying Equation (15.2) twice,

$$[a \otimes X, [b \otimes Y, c \otimes Z]] = (a(bc)) \otimes [X, [Y, Z]]\,. \tag{15.4}$$

Because A is associative and commutative, the three products obtained by cycling $(a, X) \to (b, Y) \to (c, Z) \to (a, X)$ are equal,

$$a(bc) = b(ca) = c(ab) =: abc\,,$$

so the common factor comes out of the cyclic sum of Equation (15.4):

$$\sum_{\text{cyclic}} [a \otimes X, [b \otimes Y, c \otimes Z]] = (abc) \otimes \Big([X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]]\Big) = 0\,,$$

the parenthesis vanishing by the Jacobi identity of $\mathfrak{g}$. Since elementary tensors span $\mathfrak{g}_A$ and both sides are trilinear, the identity holds throughout $\mathfrak{g}_A$. □

Remark 15.5 (What the factorisation buys). Equation (15.3) is the point of the construction. Every structure constant of $\mathfrak{g}_A$ is a product of one structure constant of A and one of $\mathfrak{g}$. Consequently any question about $\mathfrak{g}_A$ that is linear in the structure constants — semisimplicity, the Killing form, the invariant tensors, the effect of a limit — separates into a question about A and a question about $\mathfrak{g}$. Keeping $\mathfrak{g}$ fixed and varying A generates the families studied below.

Proposition 15.6 (Functoriality)**.** *Let $f : A \to A'$ be a homomorphism of commutative associative unital algebras. Then*

$$f \otimes \mathrm{id} : \mathfrak{g}_A \longrightarrow \mathfrak{g}_{A'}\,, \qquad a \otimes X \longmapsto f(a) \otimes X\,, \tag{15.5}$$

is a homomorphism of Lie algebras, with kernel $\ker(f) \otimes \mathfrak{g}$, surjective whenever f is.
Rests on Definition 15.3 and Theorem 15.4.

Derivation. *Derives Proposition 15.6.*
On elementary tensors,

$$(f \otimes \mathrm{id})\,[a \otimes X, b \otimes Y] = f(ab) \otimes [X,Y] = (f(a)f(b)) \otimes [X,Y] = [f(a) \otimes X, f(b) \otimes Y]\,,$$

using that f is multiplicative; both sides are bilinear, so the identity extends. The kernel is $\ker(f) \otimes \mathfrak{g}$ because tensoring the exact sequence $0 \to \ker f \to A \to \operatorname{im} f \to 0$ with a vector space over a field is exact; surjectivity is inherited for the same reason. □

Proposition 15.6 is the engine of the rest of the chapter: every relation between two algebras A and A' becomes a relation between the Lie algebras $\mathfrak{g}_A$ and $\mathfrak{g}_{A'}$, at no extra cost.

15.2 Ideals, the radical, and the Killing form

Proposition 15.7 (Ideals of A become ideals of $\mathfrak{g}_A$)**.** *If $I \subseteq A$ is an ideal of A, then $I \otimes \mathfrak{g}$ is a Lie ideal of $\mathfrak{g}_A$. In particular $\mathfrak{g}_A$ can be simple only if A is a field.*
Rests on Definition 15.3 and Theorem 15.4.

Derivation. *Derives Proposition 15.7.*
For $u \in I$, $b \in A$ and $X, Y \in \mathfrak{g}$, $[u \otimes X, b \otimes Y] = (ub) \otimes [X,Y]$ with $ub \in I$, because an ideal absorbs multiplication. Bilinearity extends the containment, so $[I \otimes \mathfrak{g}, \mathfrak{g}_A] \subseteq I \otimes \mathfrak{g}$. If A is not a field it has a proper nonzero ideal I, and then $I \otimes \mathfrak{g}$ is a proper nonzero ideal of $\mathfrak{g}_A$. □

Proposition 15.8 (The radical of A is solvable in $\mathfrak{g}_A$)**.** *Let $J = \mathrm{rad}(A)$, nilpotent with $J^{N+1} = 0$. Then $J \otimes \mathfrak{g}$ is a* solvable *ideal of $\mathfrak{g}_A$. If moreover $\mathfrak{g}$ is perfect, $[\mathfrak{g},\mathfrak{g}] = \mathfrak{g}$, the k-th derived algebra of $J \otimes \mathfrak{g}$ is exactly $J^{2^k} \otimes \mathfrak{g}$.*
Rests on Proposition 15.7 and Definition 15.3.

Derivation. *Derives Proposition 15.8.*
Write $\mathfrak{s} = J \otimes \mathfrak{g}$, an ideal by Proposition 15.7. By Equation (15.2),

$$[J^m \otimes \mathfrak{g}, J^m \otimes \mathfrak{g}] \subseteq J^{2m} \otimes [\mathfrak{g},\mathfrak{g}] \subseteq J^{2m} \otimes \mathfrak{g}\,, \tag{15.6}$$

so by induction $\mathfrak{s}^{(k)} \subseteq J^{2^k} \otimes \mathfrak{g}$. Choosing k with $2^k > N$ gives $J^{2^k} = 0$ and hence $\mathfrak{s}^{(k)} = 0$: $\mathfrak{s}$ is solvable. When $\mathfrak{g}$ is perfect the second inclusion in Equation (15.6) is an equality, and the first is too, because J^{2m} is spanned by products uv with $u, v \in J^m$; the induction then gives equality at every step. □

The Killing form of $\mathfrak{g}_A$ factorises as cleanly as the structure constants, and that turns the two propositions above into a criterion.

Definition 15.9 (Trace form of an algebra)**.** For $x \in A$ let $L_x : A \to A$ be multiplication by x. The *trace functional* and the *trace form* of A are

$$\tau_A(x) := \mathrm{tr}(L_x)\,, \qquad (x,y)_A := \tau_A(xy)\,. \tag{15.7}$$

Theorem 15.10 (Factorisation of the Killing form)**.** *Let $\kappa_{\mathfrak{g}}$ be the Killing form of $\mathfrak{g}$ and κ_A that of $\mathfrak{g}_A$. Then*

$$\boxed{\kappa_A\left(a \otimes X,\ b \otimes Y\right) = \tau_A(ab)\ \kappa_{\mathfrak{g}}(X, Y)}, \tag{15.8}$$

and κ_A is nondegenerate if and only if both the trace form of A and the Killing form of $\mathfrak{g}$ are nondegenerate.

Rests on Definitions 15.3 and 15.9.

Derivation. *Derives Theorem 15.10.*

By Equation (15.2), the adjoint action of $a \otimes X$ sends $c \otimes Z$ to $(ac) \otimes [X, Z]$, that is

$$\mathrm{ad}_{a\otimes X} = L_a \otimes \mathrm{ad}_X \tag{15.9}$$

as an operator on $A \otimes \mathfrak{g}$. Composing two of them,

$$\mathrm{ad}_{a\otimes X}\,\mathrm{ad}_{b\otimes Y} = (L_aL_b) \otimes (\mathrm{ad}_X\,\mathrm{ad}_Y) = L_{ab} \otimes (\mathrm{ad}_X\,\mathrm{ad}_Y)\ ,$$

where $L_aL_b = L_{ab}$ is associativity of A. The trace of a tensor product of operators is the product of the traces, which gives Equation (15.8).

For the second statement, Equation (15.8) exhibits κ_A as the tensor product of the bilinear forms $(\cdot,\cdot)_A$ on A and $\kappa_{\mathfrak{g}}$ on $\mathfrak{g}$. A tensor product of two bilinear forms is nondegenerate exactly when both factors are: in bases its Gram matrix is the Kronecker product of the two Gram matrices, of determinant $(\det G_A)^{\dim \mathfrak{g}}(\det G_{\mathfrak{g}})^{\dim A}$. □

Corollary 15.11 (Semisimplicity criterion)**.** *$\mathfrak{g}_A$ is semisimple if and only if $\mathfrak{g}$ is semisimple and the trace form of A is nondegenerate. For a finite-dimensional commutative $\mathbb{R}$-algebra the latter holds exactly when $A \cong \mathbb{R}^r \times \mathbb{C}^s$.*

Rests on Theorem 15.10 and Lemma 15.12.

Derivation. *Derives Corollary 15.11.*

Cartan's criterion states that a finite-dimensional Lie algebra over a field of characteristic zero is semisimple if and only if its Killing form is nondegenerate; apply it to both sides of Theorem 15.10. For the last sentence: if $x \in \mathrm{rad}(A)$ is nilpotent then L_{xy} is nilpotent for every y, so $\tau_A(xy) = 0$ and x lies in the radical of the trace form; conversely $\mathbb{R}^r \times \mathbb{C}^s$ has trace form $\mathrm{diag}(1, \dots, 1)$ on the real factors and, on each complex factor in the basis $\{1, \mathrm{i}\}$, the Gram matrix

$$\begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix},$$

both nondegenerate. The identification of the radical-free case with $\mathbb{R}^r \times \mathbb{C}^s$ is Lemma 15.12. □

Lemma 15.12 (Artin decomposition)**.** *Every finite-dimensional commutative associative unital $\mathbb{R}$-algebra decomposes uniquely as a finite direct product*

$$A \cong A_1 \times \cdots \times A_k \tag{15.10}$$

of local *algebras A_i, that is, algebras with a unique maximal ideal $\mathfrak{m}_i$, which is nilpotent. The decomposition is implemented by a partition $1_A = \pi_1 + \cdots + \pi_k$ into orthogonal idempotents. When every $\mathfrak{m}_i = 0$ the factors are fields, hence copies of $\mathbb{R}$ or $\mathbb{C}$, and $A \cong \mathbb{R}^r \times \mathbb{C}^s$.*

Rests on Notation 15.1 and Theorem 8.19.

Derivation. *Derives Lemma 15.12.*

The argument is elementary commutative algebra and uses nothing else from this chapter. Write $n = \dim A < \infty$ and assume $A \neq 0$.

Step 1: the nilpotents form a nilpotent ideal. Let $N \subseteq A$ be the set of nilpotent elements. If $u^r = v^t = 0$ then $(u+v)^{r+t-1} = 0$, since every term of the binomial expansion — which is

available because A is commutative — carries either u^r or v^t; and $(cu)^r = c^r u^r = 0$ for every $c \in A$. So N is an ideal. Pick a basis $u_1, \dots, u_d$ of N, necessarily consisting of nilpotents, with $u_i^{r_i} = 0$, and set $M := 1 + \sum_i (r_i - 1)$. Then N^M is spanned by products of M of the u_i, and in any such product some u_i occurs at least r_i times; hence

$$N^M = 0\,. \tag{15.11}$$

Step 2: either A is local, or it splits. Fix $a \in A$ and let $p \in \mathbb{R}[X]$ be its minimal polynomial, the monic generator of the kernel of the evaluation $\mathbb{R}[X] \to A$, $X \mapsto a$; that kernel is nonzero because $1, a, \dots, a^n$ are $n+1$ vectors in an n-dimensional space. Factor $p = p_1^{e_1} \cdots p_s^{e_s}$ into powers of distinct monic irreducibles of $\mathbb{R}[X]$, and write $q_i := p/p_i^{e_i}$.

Suppose $s \geq 2$. The q_i have no common irreducible factor, so there are $f_i \in \mathbb{R}[X]$ with $\sum_i f_i q_i = 1$; put $\pi_i := (f_i q_i)(a)$. Then $\sum_i \pi_i = 1_A$; for $i \neq j$ the product $f_i q_i f_j q_j$ is divisible by p, so $\pi_i \pi_j = 0$, and multiplying $\sum_j \pi_j = 1_A$ by π_i gives $\pi_i^2 = \pi_i$. No π_i vanishes: $f_i q_i \equiv 1$ modulo $p_i^{e_i}$, because every other summand of $\sum_j f_j q_j = 1$ is divisible by $p_i^{e_i}$; if $\pi_i = 0$ then p, hence $p_i^{e_i}$, would divide $f_i q_i$, forcing $p_i^{e_i}$ to divide 1. So the π_i are nonzero orthogonal idempotents summing to 1_A, the sum $A = \bigoplus_i \pi_i A$ is direct (multiply a vanishing sum by π_j), and $\pi_i A \cdot \pi_j A = 0$ for $i \neq j$: A is the product of the algebras $\pi_i A$, each unital with unit π_i and each of dimension strictly less than n.

Suppose instead that $s = 1$ for *every* $a \in A$, so that each minimal polynomial is $p_1^{e_1}$ with p_1 irreducible. If $p_1 = X$ then $a^{e_1} = 0$ and $a \in N$. Otherwise $p_1(0) \neq 0$ — an irreducible real polynomial with vanishing constant term is a multiple of X — so p has constant term $c_0 = p_1(0)^{e_1} \neq 0$; writing $p(X) = Xg(X) + c_0$ and evaluating at a gives $a \cdot (-g(a)/c_0) = 1_A$, so a is invertible. Thus every element is nilpotent or invertible. A nilpotent is not invertible, since $a^r = 0$ and $ab = 1_A$ would give $1_A = (ab)^r = a^r b^r = 0$; hence N is exactly the set of non-units. An ideal that contains every non-unit and consists of non-units is the unique maximal ideal, so A is local with $\mathfrak{m} = N$, nilpotent by Equation (15.11).

Step 3: induction. If $\dim A = 1$ then $A = \mathbb{R}\,1_A$ is a field, hence local with $\mathfrak{m} = 0$. If $\dim A > 1$, Step 2 either makes A local or splits it into factors of strictly smaller dimension, to which the inductive hypothesis applies. An idempotent of a factor is an idempotent of A, and idempotents belonging to different factors are orthogonal, so the partitions of the units compose into a single partition $1_A = \pi_1 + \cdots + \pi_k$ with $A_i = \pi_i A$ local.

Step 4: uniqueness. A local algebra has no idempotents but 0 and its unit: if $e^2 = e$ then e is a unit, whence $e = 1$ after multiplying $e^2 = e$ by e^{-1}, or e is nilpotent, whence $e = e^m = 0$. Let $1_A = \sum_i \pi_i = \sum_j \rho_j$ be the unit partitions of two decompositions into local factors. Each $\pi_i \rho_j$ is an idempotent lying in $\pi_i A$ and in $\rho_j A$, so it is 0 or π_i, and 0 or ρ_j. Since $\sum_j \pi_i \rho_j = \pi_i \neq 0$, at least one j gives $\pi_i \rho_j = \pi_i = \rho_j$; a second such j' is impossible, because then $\pi_i = \pi_i \rho_j$ multiplied by $\rho_{j'}$ would give $\pi_i \rho_{j'} = \pi_i \rho_j \rho_{j'} = 0$. The two partitions therefore coincide up to order, and with them the factors.

Step 5: the radical-free case. Suppose $\mathfrak{m}_i = 0$. Then the non-units of A_i are 0 alone, so A_i is a field K, finite dimensional over $\mathbb{R}\,1_K \cong \mathbb{R}$. For $b \in K$ the minimal polynomial p_b is irreducible: a factorisation $p_b = gh$ into nonconstant factors would give $g(b)h(b) = 0$ in a field, hence $g(b) = 0$ or $h(b) = 0$, contradicting minimality. By the fundamental theorem of algebra (Theorem 8.19) a real polynomial splits over $\mathbb{C}$ into linear factors, and its non-real roots occur in conjugate pairs; pairing those factors gives a factorisation into real factors of degree 1 and 2, so an irreducible real polynomial has degree at most 2. If $K \neq \mathbb{R}\,1_K$, choose $b \in K$ outside $\mathbb{R}\,1_K$: its minimal polynomial is an irreducible quadratic, and $\mathbb{R}[b] \cong \mathbb{R}[X]/(p_b) \cong \mathbb{C}$. Every $c \in K$ then has a minimal polynomial over that copy of $\mathbb{C}$ which is irreducible in $\mathbb{C}[X]$, hence linear by Theorem 8.19, so $c \in \mathbb{R}[b]$ and $K \cong \mathbb{C}$. Otherwise $K \cong \mathbb{R}$. Collecting the factors gives $A \cong \mathbb{R}^r \times \mathbb{C}^s$. □

Theorem 15.13 (Classification of $\mathfrak{g}_A$ by the algebra)**.** *Let $\mathfrak{g}$ be semisimple and A a finite-dimensional commutative associative unital $\mathbb{R}$-algebra with Artin decomposition Equation (15.10). Then*

$$\mathfrak{g}_A \cong \bigoplus_{i=1}^{k} \mathfrak{g}_{A_i} \qquad \text{as Lie algebras,} \tag{15.12}$$

and the summands fall into two branches.

1. Semisimple branch. *If A_i is a field then $\mathfrak{g}_{A_i} \cong \mathfrak{g}$ when $A_i \cong \mathbb{R}$, and $\mathfrak{g}_{A_i} \cong \mathbb{C} \otimes_{\mathbb{R}} \mathfrak{g}$, the complexification regarded as a real Lie algebra, when $A_i \cong \mathbb{C}$. Both are semisimple.*

2. Local branch. *If A_i has nonzero maximal ideal $\mathfrak{m}_i$ then $\mathfrak{g}_{A_i}$ is not semisimple, its solvable radical is exactly $\mathfrak{m}_i \otimes \mathfrak{g}$, and the quotient by that radical is $\mathfrak{g}_{A_i/\mathfrak{m}_i}$, semisimple by branch (1).*

The isomorphism class of $\mathfrak{g}_A$ is therefore determined by those of A and $\mathfrak{g}$.
Rests on Definition 15.3, Lemma 15.12, Propositions 15.6 and 15.8, and Corollary 15.11.

Derivation. *Derives Theorem 15.13.*

The splitting. With the orthogonal idempotents of Lemma 15.12, $\pi_i \pi_j = \delta_{ij} \pi_i$, so Equation (15.2) gives

$$[\pi_i a \otimes X, \pi_j b \otimes Y] = (\pi_i \pi_j ab) \otimes [X, Y] = \delta_{ij} \, (\pi_i ab) \otimes [X, Y] \ ,$$

so the subspaces $\pi_i A \otimes \mathfrak{g} = \mathfrak{g}_{A_i}$ are mutually commuting ideals whose sum is $\mathfrak{g}_A$, which is Equation (15.12).

Branch (1). If $A_i \cong \mathbb{R}$ then $A_i \otimes \mathfrak{g} \cong \mathfrak{g}$ with the same bracket, since $1 \cdot 1 = 1$. If $A_i \cong \mathbb{C}$ then $A_i \otimes_{\mathbb{R}} \mathfrak{g}$ is by definition the complexification of $\mathfrak{g}$, a real Lie algebra of twice the dimension. Both are semisimple by Corollary 15.11, the trace forms of $\mathbb{R}$ and $\mathbb{C}$ over $\mathbb{R}$ being nondegenerate as computed there.

Branch (2). By Proposition 15.8, $\mathfrak{m}_i \otimes \mathfrak{g}$ is a nonzero solvable ideal, so $\mathfrak{g}_{A_i}$ is not semisimple. Applying Proposition 15.6 to the quotient map $A_i \to A_i/\mathfrak{m}_i$ gives a surjection $\mathfrak{g}_{A_i} \to \mathfrak{g}_{A_i/\mathfrak{m}_i}$ with kernel $\mathfrak{m}_i \otimes \mathfrak{g}$. Since A_i is local, $A_i/\mathfrak{m}_i$ is a field, so the image is semisimple by branch (1) and its own radical is zero. Any solvable ideal of $\mathfrak{g}_{A_i}$ therefore maps into that zero radical, hence lies in $\mathfrak{m}_i \otimes \mathfrak{g}$; together with the solvability of $\mathfrak{m}_i \otimes \mathfrak{g}$ itself this identifies the radical exactly. □

15.3 The Takiff tower

The single most important family of algebras A is the truncated polynomial ring, because truncation is what a physical limit does to an expansion parameter.

Definition 15.14 (Truncated polynomial algebra and Takiff algebra)**.** For $N \geq 0$ set

$$A_N := \mathbb{R}[\varepsilon]/(\varepsilon^{N+1}) \, , \tag{15.13}$$

the commutative algebra with basis $\left\{1, \varepsilon, \ldots, \varepsilon^N\right\}$ and $\varepsilon^i \varepsilon^j = \varepsilon^{i+j}$ for $i + j \leq N$, zero otherwise. The associated Lie algebra

$$\mathfrak{g}\,\langle N \rangle := A_N \otimes \mathfrak{g} \tag{15.14}$$

is the *N-th Takiff algebra* of $\mathfrak{g}$ [1457]. Its elements are written $X^{(i)} := \varepsilon^i \otimes X$, the *level-$i$* copy of X.
Rests on Definition 15.3.

Proposition 15.15 (Takiff brackets)**.** *In* $\mathfrak{g}\langle N\rangle$*,*

$$\boxed{\left[X^{(i)},Y^{(j)}\right]=\begin{cases}[X,Y]^{(i+j)}\,, & i+j\le N\,,\\ 0\,, & i+j>N\,,\end{cases}} \tag{15.15}$$

so $\dim\mathfrak{g}\langle N\rangle=(N+1)\dim\mathfrak{g}$*, level* 0 *is a subalgebra isomorphic to* $\mathfrak{g}$*, and* $\bigoplus_{i\ge1}\varepsilon^i\otimes\mathfrak{g}$ *is a nilpotent ideal.*

Rests on Definitions 15.3 and 15.14 and Proposition 15.8.

Derivation. *Derives Proposition 15.15.*

The bracket is Equation (15.2) with $a=\varepsilon^i$, $b=\varepsilon^j$ and the multiplication table of A_N. Level 0 is $1\otimes\mathfrak{g}$, closed because $1\cdot1=1$, and $X\mapsto X^{(0)}$ is an isomorphism onto it. The maximal ideal $\mathfrak{m}=(\varepsilon)$ of A_N satisfies $\mathfrak{m}^{N+1}=0$, so $\mathfrak{m}\otimes\mathfrak{g}$ is a nilpotent, in particular solvable, ideal by Proposition 15.8. □

Example 15.16 (The Maxwell tower)**.** Take any $\mathfrak{g}$ and write $J:=X^{(0)}$, $P:=X^{(1)}$, $Z:=X^{(2)}$, $B:=X^{(3)}$ for the successive levels of a generator X. Then Equation (15.15) gives

$$[P,P']=[X,X']^{(2)}\,,\qquad [P,Z']=[X,X']^{(3)}\,,\qquad [Z,Z']=[X,X']^{(4)}\,, \tag{15.16}$$

each vanishing when its level exceeds N. At $N=1$ the level-1 sector is abelian; at $N=2$ one has $[P,P']=Z$ but $[P,Z']=0$, which for $\mathfrak{g}$ the Lorentz algebra is the *Maxwell algebra* [1414, 1483], in which Z_{ab} is the algebraic image of a constant electromagnetic field strength; at $N=3$ the bracket $[P,Z']=B$ reopens and the algebra is the one called $\mathfrak{B}$ in the gauge-gravity literature. The tower continues for every N.

Rests on Proposition 15.15.

Remark 15.17 (An expansion is not a contraction)*.* Proposition 14.84 recorded that a contraction preserves the dimension of the algebra. The passage $\mathfrak{g}\to\mathfrak{g}\langle N\rangle$ does not: it multiplies the dimension by $N+1$. The two operations are therefore different, and the word *expansion* is used for the second. What Section 15.5 shows is that a contraction is a degeneration *of* A *at fixed dimension of* A, whereas an expansion enlarges A. Confusing them is the commonest error in this subject.

15.4 Gradings and resonant subalgebras

The Takiff algebra puts a full copy of $\mathfrak{g}$ at every level. Physically one wants less: the boosts at one level and the rotations at another. The mechanism that permits a consistent choice is a grading.

Definition 15.18 ($\mathbb{Z}_2$-grading, symmetric coset)**.** A $\mathbb{Z}_2$*-grading* of $\mathfrak{g}$ is a vector-space decomposition $\mathfrak{g}=V_0\oplus V_1$ with

$$\boxed{[V_0,V_0]\subseteq V_0\,,\qquad [V_0,V_1]\subseteq V_1\,,\qquad [V_1,V_1]\subseteq V_0}\,. \tag{15.17}$$

The pair (V_0,V_1) is then a *symmetric coset* decomposition, V_0 being a subalgebra.

Lemma 15.19 (A grading is an involution, and the Killing form respects it)**.** *Let* $\mathfrak{g}=V_0\oplus V_1$ *be a* $\mathbb{Z}_2$*-grading and let* θ *act as* $+1$ *on* V_0 *and* -1 *on* V_1*. Then* θ *is an involutive automorphism of* $\mathfrak{g}$*, and*

$$\kappa_{\mathfrak{g}}(V_0,V_1)=0\,. \tag{15.18}$$

If $\kappa_{\mathfrak{g}}$ *is nondegenerate on* $\mathfrak{g}$*, its restrictions to* V_0 *and to* V_1 *are nondegenerate.*

Rests on Definition 15.18.

Derivation. *Derives Lemma 15.19.*

For $X \in V_i$, $Y \in V_j$ one has $[X,Y] \in V_{i+j}$ by Equation (15.17), so $\theta\,[X,Y] = (-1)^{i+j}\,[X,Y] = [(-1)^i X, (-1)^j Y] = [\theta X, \theta Y]$: θ is an automorphism, and $\theta^2 = \mathrm{id}$. The Killing form is invariant under every automorphism, so for $X \in V_0$, $Y \in V_1$,

$$\kappa_{\mathfrak{g}}(X,Y) = \kappa_{\mathfrak{g}}(\theta X, \theta Y) = \kappa_{\mathfrak{g}}(X,-Y) = -\kappa_{\mathfrak{g}}(X,Y)\,,$$

which forces Equation (15.18) in characteristic zero. If $X \in V_0$ satisfies $\kappa_{\mathfrak{g}}(X,V_0) = 0$ then by Equation (15.18) also $\kappa_{\mathfrak{g}}(X,V_1) = 0$, so X lies in the radical of $\kappa_{\mathfrak{g}}$ and vanishes; likewise for V_1. □

Definition 15.20 (Resonant subalgebra)**.** Let σ denote a $\mathbb{Z}_2$-grading $\mathfrak{g} = V_0 \oplus V_1$ and let $[k] \in \{0,1\}$ be the parity of k. The *resonant subspace* of $\mathfrak{g}\langle N\rangle$ is

$$\boxed{\mathfrak{R}_N^\sigma := \bigoplus_{k=0}^{N} \varepsilon^k \otimes V_{[k]}}\,, \tag{15.19}$$

that is: even levels carry V_0, odd levels carry V_1.

Rests on Definitions 15.14 and 15.18.

Proposition 15.21 (Resonance)**.** *$\mathfrak{R}_N^\sigma$ is a Lie subalgebra of $\mathfrak{g}\langle N\rangle$, of dimension*

$$\dim \mathfrak{R}_N^\sigma = \left\lceil \tfrac{N+1}{2} \right\rceil \dim V_0 + \left\lfloor \tfrac{N+1}{2} \right\rfloor \dim V_1\,. \tag{15.20}$$

Rests on Definitions 15.18 and 15.20 and Proposition 15.15.

Derivation. *Derives Proposition 15.21.*

Take $X \in V_{[i]}$ and $Y \in V_{[j]}$, so $\varepsilon^i \otimes X$ and $\varepsilon^j \otimes Y$ lie in $\mathfrak{R}_N^\sigma$. By Equation (15.15) their bracket is $\varepsilon^{i+j} \otimes [X,Y]$ when $i+j \le N$ and zero otherwise; the zero case is trivially in the subspace. In the surviving case Equation (15.17) gives $[X,Y] \in V_{[i]+[j]}$, and $[i]+[j] = [i+j]$ because parity is additive, so the bracket lies in $\varepsilon^{i+j} \otimes V_{[i+j]} \subseteq \mathfrak{R}_N^\sigma$. The dimension is the number of even levels in $\{0,\dots,N\}$ times $\dim V_0$ plus the number of odd ones times $\dim V_1$. □

15.4.1 The gradings of the anti-de Sitter algebra

Everything below is applied to one algebra, in every dimension.

Notation 15.22 (Generators of $\mathfrak{so}(D-1,2)$). Let η_{AB} have signature $(D-1,2)$ on $\mathbb{R}^{D+1}$, with indices $A,B \in \{0,1,\dots,D-1,D\}$; the index 0 is timelike, the indices $i,j \in \{1,\dots,D-1\}$ are spacelike, and the extra index D is timelike, $\eta_{DD} = -1$. The generators $J_{AB} = -J_{BA}$ satisfy Equation (14.90). Split them as

$$\boxed{J_{ij}\ \text{(rotations)}\,,\quad G_i := J_{0i}\ \text{(boosts)}\,,\quad H := J_{0D}\ \text{(time translation)}\,,\quad P_i := J_{iD}}\,. \tag{15.21}$$

This is the generator set of Theorem 14.90, with $K_i = G_i$ there.

Rests on Equation (14.90).

Proposition 15.23 (Brackets in the kinematical splitting)**.** *With Notation 15.22 and $\eta_{ij} = \delta_{ij}$, $\eta_{00} = -1$, $\eta_{DD} = \varsigma$ where $\varsigma = -1$ for $\mathfrak{so}(D-1,2)$ and $\varsigma = +1$ for $\mathfrak{so}(D,1)$,*

$$[G_i,G_j] = J_{ij}\,, \qquad [G_i,P_j] = \delta_{ij} H\,, \qquad [G_i,H] = P_i\,, \tag{15.22}$$

$$[P_i,P_j] = -\varsigma J_{ij}\,, \qquad [P_i,H] = \varsigma G_i\,, \qquad [J_{ij},H] = 0\,, \tag{15.23}$$

with J_{ij} acting on G_k and P_k as on a vector index. In the parametrisation Equation (14.128) of Theorem 14.90 this is $(\gamma,\mu,\nu) = (1,1,\varsigma)$: all three nonzero, the anti-de Sitter row of Table 14.1 for $\varsigma = -1$ and the de Sitter row for $\varsigma = +1$.

Rests on Notation 15.22 and Equations (14.90) and (14.128).

Derivation. *Derives Proposition 15.23.*

Each bracket is Equation (14.90) with the indices substituted. For instance

$$[P_i, P_j] = [J_{iD}, J_{jD}] = \eta_{ij} J_{DD} + \eta_{DD} J_{ji} + \eta_{iD} J_{Dj} + \eta_{Dj} J_{iD} = \eta_{DD} J_{ji} = -\varsigma J_{ij} \,,$$

using $J_{DD} = 0$ and $\eta_{iD} = \eta_{jD} = 0$; and

$$[G_i, P_j] = [J_{0i}, J_{jD}] = \eta_{0j} J_{Di} + \eta_{iD} J_{j0} + \eta_{0D} J_{ij} + \eta_{ij} J_{0D} = \delta_{ij} H \,.$$

The remaining four are the same computation. The identification of (γ, μ, ν) is read off by comparing with Equation (14.128), whose second and third relations are $[G_i, H] = \mu P_i$ and $[P_i, H] = \nu G_i$. □

Definition 15.24 (The four kinematical gradings)**.** Four splittings of $\{J_{ij}, G_i, H, P_i\}$ are relevant:

$$\begin{aligned}
\sigma_{\mathrm{st}} &: \; V_0 = \langle J_{ij}, G_i \rangle \,, \quad V_1 = \langle H, P_i \rangle \,, && \text{space–time,} \\
\sigma_{\mathrm{ss}} &: \; V_0 = \langle J_{ij}, H \rangle \,, \quad V_1 = \langle G_i, P_i \rangle \,, && \text{speed–space,} \\
\sigma_{\mathrm{sT}} &: \; V_0 = \langle J_{ij}, P_i \rangle \,, \quad V_1 = \langle G_i, H \rangle \,, && \text{speed–time,} \\
\sigma_{\mathrm{gen}} &: \; V_0 = \langle J_{ij} \rangle \,, \quad\;\; V_1 = \langle G_i, H, P_i \rangle \,, && \text{general.}
\end{aligned} \tag{15.24}$$

Rests on Notation 15.22.

Proposition 15.25 (Which of the four are symmetric cosets)**.** *The first three of Equation (15.24) satisfy Equation (15.17); the fourth does not.*

Rests on Definitions 15.18 and 15.24 and Proposition 15.23.

Derivation. *Derives Proposition 15.25.*

Read Equations (15.22) and (15.23). For σ_{st}, the set $V_0 = \langle J_{ij}, G_i \rangle$ is the Lorentz subalgebra, so $[V_0, V_0] \subseteq V_0$; $[G_i, H] = P_i$ and $[J, P] \sim P$ give $[V_0, V_1] \subseteq V_1$; and $[P_i, P_j] = -\varsigma J_{ij}$, $[P_i, H] = \varsigma G_i$ give $[V_1, V_1] \subseteq V_0$.

For σ_{ss}: $[J, J] \sim J$ and $[J_{ij}, H] = 0$ close V_0; $[J, G] \sim G$, $[J, P] \sim P$, $[H, G_i] = -P_i$ and $[H, P_i] = -\varsigma G_i$ keep $[V_0, V_1]$ in V_1; and $[G_i, G_j] = J_{ij}$, $[G_i, P_j] = \delta_{ij} H$, $[P_i, P_j] = -\varsigma J_{ij}$ all land in V_0.

For σ_{sT}: $[P_i, P_j] = -\varsigma J_{ij}$ closes V_0; $[P_i, H] = \varsigma G_i$ and $[P_j, G_i] = -\delta_{ij} H$ send $[V_0, V_1]$ into V_1; and $[G_i, G_j] = J_{ij}$, $[G_i, H] = P_i$ land in V_0.

For σ_{gen} the third condition fails at once: $[G_i, P_j] = \delta_{ij} H \in V_1$, not V_0. □

15.5 Contraction as a degeneration of A

Proposition 15.26 (The contraction along a grading)**.** *Let $\mathfrak{g} = V_0 \oplus V_1$ be a $\mathbb{Z}_2$-grading and let $U(\lambda)$ act as the identity on V_0 and as multiplication by λ on V_1. Then the Inönü–Wigner limit Equation (14.125) exists, and the contracted algebra $\mathfrak{g}_0$ is isomorphic to the resonant subalgebra at $N = 1$,*

$$\boxed{\mathfrak{g}_0 \cong \mathfrak{R}_1^{\sigma} = (1 \otimes V_0) \oplus (\varepsilon \otimes V_1) \subseteq \mathfrak{g}\,\langle 1 \rangle} \,. \tag{15.25}$$

Rests on Definitions 15.18 and 15.20, Equation (14.125), and Proposition 15.15.

Derivation. *Derives Proposition 15.26.*

Compute the three cases of Equation (14.125) using Equation (15.17). For $X, X' \in V_0$ the map $U(\lambda)$ is the identity on the arguments and, since $[X, X'] \in V_0$, on the result too, so $U^{-1}[UX, UX'] = [X, X']$, independent of λ. For $X \in V_0$, $Y \in V_1$ we have $[X, Y] \in V_1$, so

$$U(\lambda)^{-1}\,[U(\lambda)X, U(\lambda)Y] = \lambda^{-1}\,[X, \lambda Y] = [X, Y] \,,$$

again independent of λ. For $Y, Y' \in V_1$ we have $[Y, Y'] \in V_0$, on which U^{-1} acts trivially, so

$$U(\lambda)^{-1}\,[U(\lambda)Y, U(\lambda)Y'] = \lambda^2\,[Y, Y'] \xrightarrow{\lambda \to 0} 0 \,.$$

The limit exists and $\mathfrak{g}_0$ has the brackets of $\mathfrak{g}$ except that $[V_1, V_1] = 0$.

Define $\varphi : \mathfrak{g}_0 \to \mathfrak{R}_1^\sigma$ by $\varphi(X) = 1 \otimes X$ for $X \in V_0$ and $\varphi(Y) = \varepsilon \otimes Y$ for $Y \in V_1$, extended linearly; it is a linear bijection, both spaces having dimension $\dim V_0 + \dim V_1$. It is a homomorphism by Equation (15.15): the V_0–V_0 bracket lands at level 0 with value in V_0, the V_0–V_1 bracket at level 1 with value in V_1, and the V_1–V_1 bracket at level $2 > 1$, hence zero — matching the three cases just computed. □

Corollary 15.27 (The flat limit)**.** *Applying Proposition 15.26 to $\mathfrak{so}(D-1,2)$ with the space–time grading σ_{st} gives the Poincaré algebra $\mathfrak{iso}(D-1,1)$ of Equation (14.99); applying it to $\mathfrak{iso}(D-1,1)$ with the speed–space grading σ_{ss} gives the Galilei algebra. These are the two contractions of Example 14.89.*

Rests on Propositions 15.23 and 15.26, Definition 15.24, and Equation (14.99).

Derivation. *Derives Corollary 15.27.*

For σ_{st}, Proposition 15.26 sets $[V_1, V_1] = 0$, that is $[P_i, P_j] = [P_i, H] = 0$ by Equation (15.23), leaving the Lorentz algebra acting on an abelian ideal of translations: this is Equation (14.99). For σ_{ss} applied to the resulting Poincaré algebra, $[V_1, V_1] = 0$ now reads $[G_i, G_j] = [G_i, P_j] = [P_i, P_j] = 0$ while $[G_i, H] = P_i$ survives, which is the Galilei algebra $(\gamma, \mu, \nu) = (0, 1, 0)$ of Table 14.1. □

15.5.1 The curvature parameter lives in A

Proposition 15.28 (The quadratic family)**.** *For $t \in \mathbb{R}$ let*

$$A^{(t)} := \mathbb{R}[x]/(x^2 - t)\,, \tag{15.26}$$

a two-dimensional commutative unital algebra. Then

$$A^{(t)} \cong \begin{cases} \mathbb{R} \times \mathbb{R}\,, & t > 0 \quad \textit{(split)}\,, \\ \mathbb{C}\,, & t < 0 \quad \textit{(complex)}\,, \\ \mathbb{R}[\varepsilon]/(\varepsilon^2)\,, & t = 0 \quad \textit{(dual)}\,, \end{cases} \tag{15.27}$$

and only the third is non-semisimple: the trace form of $A^{(t)}$ has determinant $4t$ in the basis $\{1, x\}$.

Rests on Definition 15.9 and Corollary 15.11.

Derivation. *Derives Proposition 15.28.*

$t > 0$. Put $\pi_\pm := \frac{1}{2}(1 \pm x/\sqrt{t})$. Using $x^2 = t$,

$$\pi_\pm^2 = \tfrac{1}{4}\left(1 \pm 2x/\sqrt{t} + x^2/t\right) = \tfrac{1}{4}\left(2 \pm 2x/\sqrt{t}\right) = \pi_\pm\,, \qquad \pi_+\pi_- = \tfrac{1}{4}\left(1 - x^2/t\right) = 0\,,$$

and $\pi_+ + \pi_- = 1$: two orthogonal idempotents, so $A^{(t)}$ splits into two copies of $\mathbb{R}$.

$t < 0$. Put $\jmath := x/\sqrt{-t}$; then $\jmath^2 = x^2/(-t) = -1$, so $\mathbb{R}[\jmath] \cong \mathbb{C}$, and $A^{(t)} \cong \mathbb{C}$ by dimension count.

$t = 0$. The relation is $x^2 = 0$: the dual numbers verbatim.

For the trace form, $L_1 = \mathrm{id}$ has trace 2, L_x interchanges $1 \mapsto x \mapsto t x^0$ and has trace 0, and $L_{x^2} = L_t$ has trace $2t$; the Gram matrix of $\tau(ab)$ in $\{1, x\}$ is therefore $\mathrm{diag}(2, 2t)$, of determinant $4t$. Corollary 15.11 then gives semisimplicity of $\mathfrak{g}_{A^{(t)}}$ for $t \neq 0$ and its failure at $t = 0$. □

Theorem 15.29 (The three constant-curvature kinematics, in every dimension)**.** *Let $\mathfrak{g} = \mathfrak{so}(D-1,2)$ with the space–time grading σ_{st} of Equation (15.24), and inside $\mathfrak{g}_{A^{(t)}}$ consider the resonant subalgebra with respect to the basis $\{1, x\}$ of $A^{(t)}$,*

$$\mathfrak{R}^{(t)} := (1 \otimes V_0) \oplus (x \otimes V_1)\,, \tag{15.28}$$

which has the same dimension as $\mathfrak{g}$. *Then* $\mathfrak{R}^{(t)}$ *is a Lie subalgebra, its brackets are those of Equations* (15.22) *and* (15.23) *with* ς *replaced by* $-t$, *and*

$$\boxed{\mathfrak{R}^{(t)} \cong \begin{cases} \mathfrak{so}(D-1,2)\,, & t>0\,,\ \textit{anti-de Sitter,} \\ \mathfrak{so}(D,1)\,, & t<0\,,\ \textit{de Sitter,} \\ \mathfrak{iso}(D-1,1)\,, & t=0\,,\ \textit{Poincaré.} \end{cases}} \tag{15.29}$$

Rests on Propositions 15.23 and 15.28 and Definitions 15.3 and 15.24.

Derivation. *Derives Theorem 15.29.*

Closure and brackets. For $J, \tilde{J} \in V_0$ and $P, \tilde{P} \in V_1$ write $J' := 1 \otimes J$ and $P' := x \otimes P$. Then Equation (15.2) gives

$$\left[J', \tilde{J}'\right] = 1 \otimes \left[J, \tilde{J}\right]\,, \qquad [J', P'] = x \otimes [J, P]\,, \qquad \left[P', \tilde{P}'\right] = t\left(1 \otimes \left[P, \tilde{P}\right]\right)\,, \tag{15.30}$$

the last using $x^2 = t$ together with $[V_1, V_1] \subseteq V_0$. All three stay in $\mathfrak{R}^{(t)}$, so it is a subalgebra; and since $\left[P, \tilde{P}\right]$ is given by Equation (15.23) with $\varsigma = -1$, the brackets of the primed generators are those of Equations (15.22) and (15.23) with ς replaced by $-t$. Explicitly

$$\left[P'_i, P'_j\right] = t\, J'_{ij}\,, \qquad [P'_i, H'] = -t\, G'_i\,, \tag{15.31}$$

with all other brackets independent of t.

The three cases. For $t > 0$ rescale $P'_i \mapsto P'_i/\sqrt{t}$, $H' \mapsto H'/\sqrt{t}$: an invertible change of basis which turns Equation (15.31) into $[P, P] = J$, $[P, H] = -G$, that is Equation (15.23) with $\varsigma = -1$, the anti-de Sitter algebra $\mathfrak{so}(D-1,2)$. For $t < 0$ rescale by $\sqrt{-t}$ instead, obtaining $[P, P] = -J$, $[P, H] = +G$, which is Equation (15.23) with $\varsigma = +1$: by Proposition 15.23 this is the algebra with $\eta_{DD} = +1$, namely $\mathfrak{so}(D,1)$. For $t = 0$ both brackets in Equation (15.31) vanish and the translations become an abelian ideal, giving $\mathfrak{iso}(D-1,1)$ — which is also Corollary 15.27, since $A^{(0)}$ is the algebra of dual numbers and $\mathfrak{R}^{(0)} = \mathfrak{R}_1^{\sigma_{\mathrm{st}}}$. □

Remark 15.30 (What is special about $D = 3$, and what is not). Theorem 15.29 holds in every dimension because it works inside the resonant subalgebra, which has the same dimension as $\mathfrak{g}$. The *ambient* algebra $\mathfrak{g}_{A^{(t)}}$ is twice as large, and only in $D = 3$ does it coincide with a kinematical algebra: there $\dim V_0 = \dim V_1 = 3$, the resonant subalgebra is all of $\mathfrak{g}_{A^{(t)}}$ up to the Hodge dualisation $J_{ij} \leftrightarrow \epsilon_{ijk} J^k$, and one has the accidental isomorphisms

$$\mathfrak{so}(2,1) \oplus \mathfrak{so}(2,1) \cong \mathfrak{so}(2,2)\,, \qquad \mathbb{C} \otimes_{\mathbb{R}} \mathfrak{so}(2,1) \cong \mathfrak{sl}_2(\mathbb{C}) \cong \mathfrak{so}(3,1)\,, \tag{15.32}$$

so that the split, complex and dual branches of Proposition 15.28 are directly the anti-de Sitter, de Sitter and Poincaré algebras of three-dimensional spacetime. In $D = 4$ this fails: $\mathfrak{so}(3,2) \oplus \mathfrak{so}(3,2)$ is a ten-plus-ten dimensional algebra and is not a kinematical algebra at all, and the correct statement in 3+1 is the one Theorem 15.29 makes about $\mathfrak{R}^{(t)}$. Much of the gauge-gravity literature is written in $D = 3$ precisely because of Equation (15.32), and reading its statements as dimension-independent is a mistake this treatise is careful not to make.

15.5.2 The (anti-)de Sitter radius and the cosmological constant

The parameter t has a name in physics. To give it, one needs the relation between the curvature radius of a maximally symmetric spacetime and the cosmological constant, a short computation from the field equations.

Proposition 15.31 (Radius and cosmological constant)**.** *Let (M, g) be a D-dimensional maximally symmetric solution of the vacuum field equations with cosmological constant,*

$$R_{\mu\nu} - \tfrac{1}{2} R\, g_{\mu\nu} + \Lambda\, g_{\mu\nu} = 0\,, \tag{15.33}$$

so that its Riemann tensor has the constant-curvature form

$$R_{\mu\nu\rho\sigma} = K\,(g_{\mu\rho}g_{\nu\sigma} - g_{\mu\sigma}g_{\nu\rho})\,, \qquad K \in \mathbb{R}\,. \tag{15.34}$$

Then, for $D > 2$,

$$\boxed{K = \frac{2\Lambda}{(D-1)(D-2)}}\,, \tag{15.35}$$

and, writing $|K| = 1/\ell^2$ for the curvature radius ℓ,

$$\boxed{\Lambda = \pm\frac{(D-1)(D-2)}{2\ell^2}}\,, \tag{15.36}$$

the upper sign for de Sitter space ($K > 0$, isometry algebra $\mathfrak{so}(D,1)$) and the lower for anti-de Sitter space ($K < 0$, isometry algebra $\mathfrak{so}(D-1,2)$). In $D = 4$ this reads $\Lambda = \pm 3/\ell^2$, and in $D = 3$, $\Lambda = \pm 1/\ell^2$.

Derivation. *Derives Proposition 15.31.*

Trace Equation (15.33) with $g^{\mu\nu}$, using $g^{\mu\nu}g_{\mu\nu} = D$:

$$R - \tfrac{D}{2}R + D\Lambda = 0 \quad\Longrightarrow\quad R = \frac{2D\Lambda}{D-2}\,,$$

valid for $D \neq 2$. Substituting back,

$$R_{\mu\nu} = \left(\tfrac{1}{2}R - \Lambda\right) g_{\mu\nu} = \left(\frac{D\Lambda}{D-2} - \Lambda\right) g_{\mu\nu} = \frac{2\Lambda}{D-2}\, g_{\mu\nu}\,. \tag{15.37}$$

From Equation (15.33) (trace with $g^{\mu\nu}$ and substitute the resulting scalar curvature back).

Contracting Equation (15.34) with $g^{\mu\rho}$,

$$R_{\nu\sigma} = g^{\mu\rho}R_{\mu\nu\rho\sigma} = K\,(D\, g_{\nu\sigma} - g_{\nu\sigma}) = (D-1)K\, g_{\nu\sigma}\,.$$

Comparing with Equation (15.37) gives Equation (15.35), and Equation (15.36) follows on setting $|K| = 1/\ell^2$. The isometry algebras are those of the hyperboloids $\eta_{AB}X^A X^B = \pm\ell^2$ in $\mathbb{R}^{D,1}$ and $\mathbb{R}^{D-1,2}$ respectively, which is how $\mathfrak{so}(D,1)$ and $\mathfrak{so}(D-1,2)$ were introduced in Chapter 14. □

Corollary 15.32 (The expansion parameter is the cosmological constant)**.** *With the identifications of Theorem 15.29 and Proposition 15.31,*

$$\boxed{t = -\frac{2\Lambda}{(D-1)(D-2)} = \begin{cases} +1/\ell^2\,, & \textit{anti-de Sitter, } \Lambda < 0\,, \\ -1/\ell^2\,, & \textit{de Sitter, } \Lambda > 0\,, \\ 0\,, & \textit{Minkowski, } \Lambda = 0\,. \end{cases}} \tag{15.38}$$

The flat limit $\ell \to \infty$ is the degeneration $t \to 0$ of Proposition 15.28, and the sign of the cosmological constant is minus the sign of the discriminant of the quadratic that defines A.

Rests on Theorem 15.29 and Propositions 15.28 and 15.31.

Derivation. *Derives Corollary 15.32.*

Theorem 15.29 makes $\mathfrak{R}^{(t)}$ the anti-de Sitter algebra for $t > 0$ and the de Sitter algebra for $t < 0$; Proposition 15.31 attaches $\Lambda < 0$ to the first and $\Lambda > 0$ to the second, with $|\Lambda| = (D-1)(D-2)/(2\ell^2)$ in both cases. Since the rescaling used in the proof of Theorem 15.29 sends $|t|$ to 1, the parameter t carries the units of $|K| = 1/\ell^2$ and the sign that distinguishes the branches, which is Equation (15.38) once Equation (15.35) is used. □

Remark 15.33 (Which branch nature is on). The cosmological constant inferred from the expansion history is *positive*. With the Planck values $\Omega_\Lambda = 0.6889(56)$ and $H_0 = (67.66 \pm 0.42)\,\mathrm{km/(s\,Mpc)}$ [2348], in $D = 4$,

$$\Lambda = \frac{3\Omega_\Lambda H_0^2}{c^2} \approx 1.1 \times 10^{-52}/\mathrm{m}^2\,, \qquad \ell = \sqrt{3/\Lambda} \approx 1.6 \times 10^{26}\,\mathrm{m}\,, \tag{15.39}$$

a radius of about 17 Gyr in light-travel terms; the independent distance-ladder determination of H_0 [2407] moves the number by a few per cent without touching its sign. By Corollary 15.32 the observed universe therefore sits on the branch $t < 0$, whose kinematical algebra is $\mathfrak{so}(4,1)$ — *de Sitter*, not the anti-de Sitter algebra $\mathfrak{so}(3,2)$ around which most of the gauge-gravity literature is built. That is the verdict already recorded in Table 14.1, now expressed as a property of A: the sign of a measured quantity selects which of the three two-dimensional commutative algebras of Proposition 15.28 describes nature.

One qualification, which the treatise insists on: Equation (15.39) is an inference from the expansion history of Chapter 48 within a ΛCDM parametrisation, not a measurement of a symmetry algebra. A symmetry is measured by testing its transformations, as in Chapter 37. What is being claimed here is that *if* the asymptotic geometry is de Sitter, the algebra is $\mathfrak{so}(4,1)$ and $t < 0$; the antecedent is observational and the consequent is a theorem.

Rests on Corollary 15.32 and Proposition 15.31.

15.6 Central charges as the top level of an expansion

We come to the point of the chapter. In Section 14.4 a central extension was a choice: a class in $H^2(\mathfrak{g}, \mathbb{R})$, imposed on the algebra from outside. Here it is not a choice but a consequence. Adding one level to the truncation adds generators, and the generators added at the top of the tower are the central charge.

Theorem 15.34 (The truncation tower is a tower of central extensions)**.** *Fix a $\mathbb{Z}_2$-grading σ of $\mathfrak{g}$ and let $\mathfrak{R}^\sigma_N$ be as in Definition 15.20. Let $\pi_N : A_{N+1} \to A_N$ be the algebra epimorphism $\varepsilon^{N+1} \mapsto 0$. Then:*

1. *π_N induces a surjective Lie homomorphism $p_N : \mathfrak{R}^\sigma_{N+1} \to \mathfrak{R}^\sigma_N$ with kernel*

$$\mathfrak{k}_N := \varepsilon^{N+1} \otimes V_{[N+1]}\,, \tag{15.40}$$

   and $\mathfrak{k}_N$ is an abelian ideal.

2. *An element $\varepsilon^{N+1} \otimes X$ of $\mathfrak{k}_N$ is central in $\mathfrak{R}^\sigma_{N+1}$ if and only if*

$$\boxed{[X, V_0] = 0}\,, \tag{15.41}$$

   so $\mathfrak{k}_N$ is central exactly when $\big[V_{[N+1]}, V_0\big] = 0$.

3. *When $\mathfrak{k}_N$ is central,*

$$0 \longrightarrow \mathfrak{k}_N \longrightarrow \mathfrak{R}^\sigma_{N+1} \xrightarrow{\;p_N\;} \mathfrak{R}^\sigma_N \longrightarrow 0 \tag{15.42}$$

is a central extension in the sense of Definition 14.72, and the 2-cocycle determined by the level-preserving splitting is

$$\boxed{c\left(\varepsilon^i \otimes X,\ \varepsilon^j \otimes Y\right) = \delta_{i+j,\,N+1}\,\varepsilon^{N+1} \otimes [X,Y]}\,. \tag{15.43}$$

Rests on Definitions 14.72, 15.18 and 15.20 and Propositions 14.73 and 15.6.

Derivation. *Derives Theorem 15.34.*

(1) By Proposition 15.6, $\pi_N \otimes \mathrm{id}$ is a Lie homomorphism $\mathfrak{g}\langle N+1\rangle \to \mathfrak{g}\langle N\rangle$; it maps $\varepsilon^k \otimes V_{[k]}$ identically onto itself for $k \le N$ and kills the top level, so it carries $\mathfrak{R}^\sigma_{N+1}$ onto $\mathfrak{R}^\sigma_N$ with kernel Equation (15.40). That kernel is abelian because

$$\left[\varepsilon^{N+1} \otimes X, \varepsilon^{N+1} \otimes Y\right] = \varepsilon^{2N+2} \otimes [X,Y] = 0\,,$$

since $2N+2 > N+1$ for every $N \ge 0$.

(2) Let $X \in V_{[N+1]}$ and let $\varepsilon^j \otimes Y$ be an arbitrary element of $\mathfrak{R}^\sigma_{N+1}$, so $Y \in V_{[j]}$. Then

$$\left[\varepsilon^{N+1} \otimes X, \varepsilon^j \otimes Y\right] = \varepsilon^{N+1+j} \otimes [X,Y]\ , \tag{15.44}$$

which vanishes automatically for every $j \ge 1$, since then $N+1+j > N+1$. Only $j=0$ can survive; there $Y \in V_0$ and Equation (15.44) equals $\varepsilon^{N+1} \otimes [X,Y]$. By Equation (15.17), $[X,Y] \in V_{[N+1]}$, and $\varepsilon^{N+1} \otimes Z$ vanishes for $Z \in V_{[N+1]}$ only if $Z = 0$. So the bracket vanishes for all Y exactly when $[X,V_0] = 0$.

(3) The sequence Equation (15.42) is exact by (1) and central by (2), so it is a central extension. Define the splitting $s : \mathfrak{R}^\sigma_N \to \mathfrak{R}^\sigma_{N+1}$ by $s(\varepsilon^i \otimes X) = \varepsilon^i \otimes X$ for $i \le N$, a linear section of p_N. For $i, j \le N$,

$$\left[s(\varepsilon^i \otimes X), s(\varepsilon^j \otimes Y)\right] = \varepsilon^{i+j} \otimes [X,Y]$$

in $\mathfrak{R}^\sigma_{N+1}$, whereas the same bracket computed in $\mathfrak{R}^\sigma_N$ and then lifted by s gives $\varepsilon^{i+j} \otimes [X,Y]$ when $i+j \le N$ and 0 when $i+j > N$. The two agree except when $i+j = N+1$, where the difference is $\varepsilon^{N+1} \otimes [X,Y] \in \mathfrak{k}_N$; for $i+j > N+1$ both vanish. Comparing with Equation (14.120), the defect is the cocycle Equation (15.43), which satisfies Equation (14.121) automatically by Proposition 14.73. □

Corollary 15.35 (Where the central charges are)**.** *Let $\mathfrak{z}(V_0) = \{X \in V_0 : [X,V_0] = 0\}$ be the centre of the subalgebra V_0, and let $V_1^{V_0} = \{Y \in V_1 : [Y,V_0] = 0\}$ be the subspace of V_1 on which V_0 acts trivially. Then the top level of $\mathfrak{R}^\sigma_N$ contributes to the centre exactly*

$$\boxed{\varepsilon^N \otimes \begin{cases} \mathfrak{z}(V_0)\,, & N \text{ even,} \\ V_1^{V_0}\,, & N \text{ odd.} \end{cases}} \tag{15.45}$$

In particular, if V_1 carries no trivial V_0-summand — which holds for every grading of Equation (15.24), since V_0 always contains the rotations and V_1 always consists of spatial vectors — central charges occur only at even truncation order*, and they are the level-N copies of the centre of V_0.*

Rests on Theorem 15.34 and Definition 15.24.

Derivation. *Derives Corollary 15.35.*

Apply Theorem 15.34(2) with N in place of $N+1$: an element $\varepsilon^N \otimes X$ with $X \in V_{[N]}$ is central precisely when $[X,V_0] = 0$. For N even, $V_{[N]} = V_0$ and the condition defines $\mathfrak{z}(V_0)$; for N odd, $V_{[N]} = V_1$ and it defines $V_1^{V_0}$. For the last sentence, in each grading of Equation (15.24) the space V_1 is spanned by generators carrying a free spatial index, on which the rotations $J_{ij} \subseteq V_0$ act as on the vector representation of $\mathfrak{so}(D-1)$; that representation has no trivial summand for $D \ge 3$, so $V_1^{V_0} = 0$. □

Corollary 15.36 (Bargmann in 3+1, extended Bargmann in 2+1)**.** *Let* $\mathfrak{g} = \mathfrak{iso}(D-1,1)$ *be the Poincaré algebra and let* σ_{ss} *be the speed–space grading* $V_0 = \langle J_{ij}, H\rangle$, $V_1 = \langle G_i, P_i\rangle$. *Then:*

1. $\mathfrak{R}_1^{\sigma_{\mathrm{ss}}}$ *is the* Galilei *algebra, with no central charge;*
2. *at* $N = 2$ *the element* $M := \varepsilon^2 \otimes H$ *is central in every dimension, and*

$$\boxed{[\varepsilon \otimes G_i, \varepsilon \otimes P_j] = \delta_{ij}\, M}\,. \tag{15.46}$$

   For $D \geq 4$ *the centre of* V_0 *is exactly* $\langle H\rangle$*, so* M *is the* only *central charge; quotienting* $\mathfrak{R}_2^{\sigma_{\mathrm{ss}}}$ *by the ideal* $\varepsilon^2 \otimes \langle J_{ij}\rangle$ *leaves the* Bargmann *algebra, the central extension of Galilei by the mass. In* 3+1 *this is the physically relevant case;*
3. *only for* $D = 3$ *is* $\langle J_{ij}\rangle$ *abelian, so that* $\mathfrak{z}(V_0) = \langle H, J\rangle$ *and* $\mathfrak{R}_2^{\sigma_{\mathrm{ss}}}$ *is the* extended Bargmann *algebra with the two central charges* $M = \varepsilon^2 \otimes H$ *and* $S = \varepsilon^2 \otimes J$*; this is the algebra on which the three-dimensional constructions of [2425] are built.*

In every dimension the cocycle Equation (15.43) *of this extension is* $c(G_i, P_j) = \delta_{ij}$*, which is the class computed from the Poisson brackets of a free particle in Proposition 25.14 and named in Example 14.80.*

Rests on Theorem 15.34, Corollaries 15.27 and 15.35, and Proposition 25.14.

Derivation. *Derives Corollary 15.36.*

The grading. The Poincaré brackets are Equations (15.22) and (15.23) with $\varsigma = 0$, that is $[P_i, P_j] = [P_i, H] = 0$, the rest unchanged (Corollary 15.27). Then $[G_i, G_j] = J_{ij}$, $[G_i, P_j] = \delta_{ij} H$ and $[P_i, P_j] = 0$ lie in V_0; $[J, G] \sim G$, $[J, P] \sim P$, $[H, G_i] = -P_i$ and $[H, P_i] = 0$ lie in V_1; and $[J, J] \sim J$, $[J_{ij}, H] = 0$ lie in V_0. Equation (15.17) holds, and $V_0 = \mathfrak{so}(D-1) \oplus \langle H\rangle$ as a direct sum of ideals.

(1). $\mathfrak{R}_1^{\sigma_{\mathrm{ss}}} = (1 \otimes V_0) \oplus (\varepsilon \otimes V_1)$, and by Equation (15.15) every bracket of two level-1 generators lands at level $2 > 1$ and vanishes, so $[G_i, G_j] = [G_i, P_j] = [P_i, P_j] = 0$ while $[H, G_i] = -P_i$ survives: the Galilei algebra $(\gamma, \mu, \nu) = (0, 1, 0)$. It has no central charge because Corollary 15.35 with $N = 1$ odd gives $V_1^{V_0} = 0$ for $D \geq 3$.

(2). By Equation (15.43) with $i = j = 1$ and $N + 1 = 2$,

$$[\varepsilon \otimes G_i, \varepsilon \otimes P_j] = \varepsilon^2 \otimes [G_i, P_j] = \delta_{ij}\, \varepsilon^2 \otimes H\,,$$

which is Equation (15.46). That M is central follows from Equation (15.41): $[H, V_0] = 0$ because $[H, J_{ij}] = 0$ and $[H, H] = 0$. By Corollary 15.35 the full central content of the top level is $\varepsilon^2 \otimes \mathfrak{z}(V_0)$, and $\mathfrak{z}(V_0) = \langle H\rangle \oplus \mathfrak{z}(\mathfrak{so}(D-1)) = \langle H\rangle$ for $D - 1 \geq 3$, since $\mathfrak{so}(n)$ has trivial centre for $n \geq 3$. The subspace $\mathfrak{i} := \varepsilon^2 \otimes \langle J_{ij}\rangle$ is an ideal: its brackets with levels ≥ 1 vanish for degree reasons, and at level 0 one has $[J_{ij}, J_{kl}] \in \langle J\rangle$ and $[J_{ij}, H] = 0$. Quotienting by $\mathfrak{i}$ kills $[\varepsilon \otimes G_i, \varepsilon \otimes G_j] = \varepsilon^2 \otimes J_{ij}$ and leaves the generators J_{ij}, H, G_i, P_i, M with M central and Equation (15.46) the only new bracket: the Bargmann algebra.

(3). For $D = 3$ the rotation sector is one-dimensional, spanned by a single J, so $[V_0, V_0] = 0$ and $\mathfrak{z}(V_0) = V_0$. Then $S := \varepsilon^2 \otimes J$ is central as well and $[\varepsilon \otimes G_1, \varepsilon \otimes G_2] = \varepsilon^2 \otimes J_{12}$ is a multiple of S: the two-fold central extension called the extended Bargmann algebra. That $D = 3$ is the only such case is the computation of $\mathfrak{z}(V_0)$ in part (2).

The cocycle. Reading Equation (15.46) against Equation (14.120) with central generator $Z = M$ gives $c(G_i, P_j) = \delta_{ij}$ and $c = 0$ on every other pair, which is Equation (25.18) with the mass scaled to unity. □

Remark 15.37 (Why this answers the structural question)*.* Equation (15.43) says that the 2-cocycle of the extension is nothing other than *the part of the bracket that the truncation discards.* Going from $\mathfrak{R}_N^{\sigma}$ to $\mathfrak{R}_{N+1}^{\sigma}$ does not add a central charge by fiat; it stops throwing

one away. The cohomological obstruction of Definition 14.75 and the level structure of the expansion are two descriptions of one object, and the reason a contraction can create a central charge (Proposition 14.84) becomes visible: a contraction is the truncation $\varepsilon^2 = 0$ (Proposition 15.26), and truncation is what makes a bracket land on a central generator instead of a genuine one.

Remark 15.38 (The same statement seen three ways). The mass of a nonrelativistic particle has now been obtained as a central charge three times in this treatise, and the three are one fact in three languages. In Proposition 25.14 it is the constant defect of the Poisson-bracket realisation of the Galilei algebra, computed from the explicit generators of a free particle in 3+1. In Example 14.80 it is the generator of the one-dimensional H^2 of the Galilei algebra. Here it is the level-2 time translation $\varepsilon^2 \otimes H$ of an expansion truncated one order later than the contraction that produces Galilei. The third is the most structural, because Corollary 15.35 also settles how many other central charges there are: in 3+1 exactly none, since $\mathfrak{z}(V_0) = \langle H \rangle$; the second charge S of the extended Bargmann algebra is a feature of 2+1 alone.

Table 15.1: The four gradings of Equation (15.24) applied to $\mathfrak{so}(D-1,2)$ (or to its flat contraction where indicated), the algebra obtained at truncation $N = 1$ by Proposition 15.26, and the central charges carried by the top level at $N = 2$, computed from Corollary 15.35 as $\varepsilon^2 \otimes \mathfrak{z}(V_0)$. Names in the third column are those of Table 14.1. The last row is the exception of Proposition 15.25: it is not a symmetric coset, so the resonant subspace closes only at $N = 1$, where every level-2 bracket vanishes for degree reasons anyway.

Grading	V_0	Algebra at $N = 1$	$\mathfrak{z}(V_0)$	Central charges at $N = 2$
space–time	J_{ij}, G_i	Poincaré $\mathfrak{iso}(D-1,1)$	0, the Lorentz algebra being simple	none; the top level $\varepsilon^2 \otimes V_0$ is the Maxwell tensor Z_{ab}
speed–space	J_{ij}, H	Newton–Hooke from $\mathfrak{so}(D-1,2)$, Galilei from $\mathfrak{iso}(D-1,1)$	$\langle H \rangle$ for $D \geq 4$; $\langle H, J \rangle$ for $D = 3$	the mass $M = \varepsilon^2 \otimes H$; in 2+1 also $S = \varepsilon^2 \otimes J$
speed–time	J_{ij}, P_i	para-Poincaré from $\mathfrak{so}(D-1,2)$, Carroll from $\mathfrak{iso}(D-1,1)$	0: the rotations act nontrivially on P_i for $D \geq 3$	none from the top level
general	J_{ij}	static	—	— (not a symmetric coset)

Remark 15.39 (What the construction does and does not prove). Table 15.1 *produces* four of the eleven algebras of Table 14.1 in one step; the remaining seven arise by iterating the construction (Galilei from Poincaré rather than from $\mathfrak{so}(D-1,2)$, as in Corollary 15.27) or by the sign of t in Corollary 15.32, which is what splits de Sitter from anti-de Sitter and Newton–Hooke from its anti- partner. Section 15.7 does the bookkeeping properly: the four splittings turn out to be three edge directions and one main diagonal of a single cube, and all eleven algebras are accounted for at once. What the construction does *not* do is prove that the list of eleven is complete. That is the separate theorem Theorem 14.90, proved in Appendix A.10 by exhausting the possible structure constants, and nothing here replaces it. The two results answer different questions: one says which algebras exist, the other says how they are generated from one another and what central charges accompany them.

15.7 The Bacry–Lévy-Leblond cube as a lattice of algebras

Theorem 14.90 organises the eleven kinematical algebras by three constants (γ, μ, ν) and observes that setting each of them to zero is a contraction, so that the eight vanishing patterns are the vertices of a cube. The cube is drawn there as a picture. It is in fact an algebraic object, and this section identifies it: the eight vertices are the eight degenerations of one four-dimensional commutative algebra, the three constants are its three structure constants, and the count $8 + 3 = 11$ falls out of the sign data of that algebra — including the two entries that hypothesis (H3) of Theorem 14.90 removes.

15.7.1 The kinematical generators are a Klein grading

The gradings of Equation (15.24) were introduced one at a time. They are not independent.

Lemma 15.40 (The three symmetric cosets are one $\mathbb{Z}_2 \times \mathbb{Z}_2$-grading)**.** *Let $\theta_{\mathrm{st}}, \theta_{\mathrm{ss}}, \theta_{\mathrm{sT}}$ be the involutions of Lemma 15.19 attached to the three symmetric-coset gradings of Proposition 15.25. Then*

$$\theta_{\mathrm{st}}\,\theta_{\mathrm{ss}} = \theta_{\mathrm{sT}}\,, \tag{15.47}$$

so the three generate a Klein four-group $\Gamma \cong \mathbb{Z}_2 \times \mathbb{Z}_2$, and the common eigenspaces of Γ are

$$\boxed{\mathfrak{g}_{00} = \langle J_{ij} \rangle\,, \quad \mathfrak{g}_{01} = \langle G_i \rangle\,, \quad \mathfrak{g}_{10} = \langle H \rangle\,, \quad \mathfrak{g}_{11} = \langle P_i \rangle}\,, \tag{15.48}$$

a Γ-grading of $\mathfrak{so}(D-1,2)$: $[\mathfrak{g}_\alpha, \mathfrak{g}_\beta] \subseteq \mathfrak{g}_{\alpha+\beta}$. The three symmetric-coset gradings are exactly the three surjections $\Gamma \to \mathbb{Z}_2$, and the fourth splitting σ_{gen} of Equation (15.24) is not induced by any group homomorphism — which is why it fails Equation (15.17).

Rests on Lemma 15.19, Definition 15.24, and Propositions 15.23 and 15.25.

Derivation. *Derives Lemma 15.40.*

Evaluate $\theta_{\mathrm{st}}\theta_{\mathrm{ss}}$ on each generator, using $\theta = +1$ on V_0 and -1 on V_1 for the corresponding splitting of Equation (15.24):

$$J_{ij} : (+)(+) = +\,, \quad G_i : (+)(-) = -\,, \quad H : (-)(+) = -\,, \quad P_i : (-)(-) = +\,,$$

which is $+1$ on $\langle J_{ij}, P_i \rangle$ and -1 on $\langle G_i, H \rangle$: that is θ_{sT}, proving Equation (15.47). Each θ squares to the identity, so the group they generate is $\mathbb{Z}_2 \times \mathbb{Z}_2$.

Label a character of Γ by its values on $(\theta_{\mathrm{st}}, \theta_{\mathrm{ss}})$, writing 0 for $+1$ and 1 for -1. Reading the signs above gives Equation (15.48). That this is a grading is Equations (15.22) and (15.23) read modulo two: $[G, P] \propto H$ is $01 + 11 = 10$; $[G, H] \propto P$ is $01 + 10 = 11$; $[H, P] \propto G$ is $10 + 11 = 01$; $[G, G]$ and $[P, P]$ are proportional to J, which is $01 + 01 = 00$ and $11 + 11 = 00$; and $[J, \cdot\,]$ preserves each component because J has degree 00.

A surjection $\Gamma \to \mathbb{Z}_2$ has a two-element kernel, so it makes exactly one of the three nontrivial components even together with $\mathfrak{g}_{00}$; the three kernels $\{00, 01\}$, $\{00, 10\}$, $\{00, 11\}$ give $V_0 = \langle J, G \rangle$, $\langle J, H \rangle$, $\langle J, P \rangle$, which are σ_{st}, σ_{ss}, σ_{sT}. The splitting σ_{gen} has $V_0 = \langle J_{ij} \rangle$ alone; no subgroup of Γ has order one and index two, so it is not the kernel of any homomorphism, and Proposition 15.25 showed directly that it is not a symmetric coset. □

15.7.2 Resonance for a graded group

Definition 15.20 matched the parity of a power of ε to the parity of a grading. Nothing in the argument used $\mathbb{Z}_2$.

Proposition 15.41 (Resonance, general form)**.** *Let Γ be a finite abelian group, let $\mathfrak{g} = \bigoplus_{\alpha\in\Gamma} \mathfrak{g}_\alpha$ be a Γ-graded Lie algebra and $A = \bigoplus_{\alpha\in\Gamma} A_\alpha$ a Γ-graded commutative associative algebra. Then*

$$\boxed{\mathfrak{R} := \bigoplus_{\alpha\in\Gamma} A_\alpha \otimes \mathfrak{g}_\alpha} \tag{15.49}$$

is a Lie subalgebra of $\mathfrak{g}_A$.
Rests on Definition 15.3.

Derivation. *Derives Proposition 15.41.*
By Equation (15.2),

$$[A_\alpha \otimes \mathfrak{g}_\alpha, A_\beta \otimes \mathfrak{g}_\beta] \subseteq (A_\alpha A_\beta) \otimes [\mathfrak{g}_\alpha, \mathfrak{g}_\beta] \subseteq A_{\alpha+\beta} \otimes \mathfrak{g}_{\alpha+\beta} \subseteq \mathfrak{R}\,,$$

the middle inclusion being the two grading hypotheses. Proposition 15.21 is the case $\Gamma = \mathbb{Z}_2$ with $A = A_N$ graded by the parity of the exponent of ε. □

15.7.3 The algebra with three structure constants

Proposition 15.42 (The Γ-graded commutative algebras)**.** *Let* $\Gamma = \mathbb{Z}_2 \times \mathbb{Z}_2$ *and let* A *be a commutative associative unital* Γ*-graded* $\mathbb{R}$*-algebra whose four graded pieces are all one-dimensional, spanned by* 1 *and by* e_G, e_H, e_P *in the degrees of Equation (15.48). Then* A *is determined by three real numbers* $\lambda_G, \lambda_H, \lambda_P$ *through*

$$\boxed{e_G e_P = \lambda_H e_H\,, \qquad e_G e_H = \lambda_P e_P\,, \qquad e_H e_P = \lambda_G e_G}\,, \tag{15.50}$$

and associativity then forces *the squares to be*

$$\boxed{e_G^2 = \lambda_H \lambda_P\,, \qquad e_H^2 = \lambda_G \lambda_P\,, \qquad e_P^2 = \lambda_G \lambda_H}\,. \tag{15.51}$$

Write $A(\lambda)$ *for this algebra. Its trace form Equation (15.7) is* $\operatorname{diag}(4,\, 4e_G^2,\, 4e_H^2,\, 4e_P^2)$ *in the basis* $\{1, e_G, e_H, e_P\}$*, so* $A(\lambda)$ *is semisimple if and only if* $\lambda_G \lambda_H \lambda_P \neq 0$.
Rests on Lemma 15.40, Definition 15.9, and Corollary 15.11.

Derivation. *Derives Proposition 15.42.*
The three products Equation (15.50) are forced to lie in the stated one-dimensional pieces by the grading, since $01 + 11 = 10$, $01 + 10 = 11$ and $10 + 11 = 01$; the mnemonic is that λ_X is the constant of the product that *lands on* e_X. A square has degree 00, so e_X^2 is a multiple of 1.

For Equation (15.51), impose associativity on $e_G\, e_G\, e_P$:

$$(e_G e_G)\, e_P = e_G^2\, e_P\,, \qquad e_G\, (e_G e_P) = e_G\, (\lambda_H e_H) = \lambda_H \lambda_P\, e_P\,,$$

whence $e_G^2 = \lambda_H \lambda_P$; the other two squares follow by the same computation on $e_H e_H e_G$ and $e_P e_P e_G$. With Equation (15.51) in force every remaining triple product is associative: the only case not already used is $(e_G e_H)\, e_P = \lambda_P e_P^2 = \lambda_G \lambda_H \lambda_P$ against $e_G\, (e_H e_P) = \lambda_G e_G^2 = \lambda_G \lambda_H \lambda_P$, which agree.

For the trace form, L_{e_G} sends $1 \mapsto e_G$, $e_G \mapsto e_G^2 \cdot 1$, $e_H \mapsto \lambda_P e_P$ and $e_P \mapsto \lambda_H e_H$: no basis vector is sent to a multiple of itself, so $\tau_A(e_G) = 0$, and likewise for e_H, e_P. Hence $\tau_A(e_X e_Y) = 0$ for $X \neq Y$, while $\tau_A(1) = \operatorname{tr}(\mathrm{id}) = 4$ and $\tau_A(e_X^2) = 4e_X^2$. The determinant is $4^4\, (\lambda_G \lambda_H \lambda_P)^2$ by Equation (15.51), and Corollary 15.11 converts nondegeneracy into semisimplicity. □

15.7.4 The cube

Theorem 15.43 (The three constants of the classification are the three structure constants of A)**.** *Let* $\mathfrak{g}$ *be a kinematical algebra with constants* (γ, μ, ν) *in the sense of Equations (14.128) and (14.129), graded by* Γ *as in Equation (15.48), and let* $\mathfrak{R}(\lambda) \subseteq \mathfrak{g}_{A(\lambda)}$ *be the resonant subalgebra Equation (15.49), of the same dimension as* $\mathfrak{g}$*. Then* $\mathfrak{R}(\lambda)$ *is again a kinematical algebra in the sense of Theorem 14.90, with constants*

$$\boxed{(\gamma, \mu, \nu) \longmapsto (\gamma\lambda_H,\ \mu\lambda_P,\ \nu\lambda_G)}\,, \tag{15.52}$$

and the two brackets that the Jacobi identity fixes, Equation (14.129)*, come out consistent with Equation* (15.52) *without further assumption.*

Rests on Propositions 15.41 and 15.42, Lemma 15.40, and Equations (14.128) *and* (14.129)*.*

Derivation. *Derives Theorem 15.43.*

Write the resonant generators as $J' := 1 \otimes J_{ij}$, $G' := e_G \otimes G_i$, $H' := e_H \otimes H$, $P' := e_P \otimes P_i$; by Proposition 15.41 they span a subalgebra, of dimension $\dim \mathfrak{g}$ because each graded piece of $A(\lambda)$ is one-dimensional. Now apply Equation (15.2) to each bracket of Equation (14.128), using Equation (15.50):

$$\left[G'_i, P'_j\right] = (e_G e_P) \otimes [G_i, P_j] = \lambda_H e_H \otimes \gamma\, \delta_{ij} H = (\gamma\lambda_H)\, \delta_{ij} H' \,, \tag{15.53}$$

$$[G'_i, H'] = (e_G e_H) \otimes [G_i, H] = \lambda_P e_P \otimes \mu P_i = (\mu\lambda_P)\, P'_i \,, \tag{15.54}$$

$$[P'_i, H'] = (e_P e_H) \otimes [P_i, H] = \lambda_G e_G \otimes \nu G_i = (\nu\lambda_G)\, G'_i \,, \tag{15.55}$$

which is Equation (15.52). For the two remaining brackets, use Equation (15.51):

$$\left[G'_i, G'_j\right] = e_G^2 \otimes [G_i, G_j] = \lambda_H\lambda_P\,(-\gamma\mu\,\epsilon_{ijk} J_k) = -\,(\gamma\lambda_H)\,(\mu\lambda_P)\,\epsilon_{ijk} J'_k \,,$$

$$\left[P'_i, P'_j\right] = e_P^2 \otimes [P_i, P_j] = \lambda_G\lambda_H\,(\gamma\nu\,\epsilon_{ijk} J_k) = (\gamma\lambda_H)\,(\nu\lambda_G)\,\epsilon_{ijk} J'_k \,,$$

which are exactly Equation (14.129) written for the image constants. The rotations act on G'_i and P'_i as before, since $1 \cdot e_X = e_X$, so hypothesis (H1) of Theorem 14.90 survives, and the parity and time-reversal automorphisms of hypothesis (H2) commute with the grading because they act on each of J, G, H, P by a sign.

The consistency of the last display is the point worth noting: the squares Equation (15.51) were forced on $A(\lambda)$ by *associativity*, and the relations Equation (14.129) were forced on $\mathfrak{g}$ by the *Jacobi identity*. That the first reproduces the second is the factorisation of Equation (15.3) doing its work. □

Corollary 15.44 (The eight vertices are the eight degenerations of A)**.** *Under Equation* (15.52) *the vanishing pattern of* (γ, μ, ν) *is the vanishing pattern of* $(\lambda_H, \lambda_P, \lambda_G)$*, so the eight vertices of the contraction cube of Theorem 14.90 are the eight vanishing patterns of the structure constants of* $A(\lambda)$*, and each edge of the cube is a degeneration* $\lambda_X \to 0$ *of the algebra. Moreover the four splittings of Equation* (15.24) *sit in the cube as follows:*

- σ_{st} *sets* $\lambda_G = 0$*, killing* ν*;*
- σ_{ss} *sets* $\lambda_H = 0$*, killing* γ*;*
- σ_{sT} *sets* $\lambda_P = 0$*, killing* μ*;*
- σ_{gen} *sets all three to zero at once.*

The first three are the three edge directions; the fourth is the main diagonal, from the vertex where nothing vanishes to the vertex where everything does.

Rests on Theorems 14.90 and 15.43, Proposition 15.26, and Definition 15.24.

Derivation. *Derives Corollary 15.44.*

The first sentence is Equation (15.52). For the identification of the splittings, Proposition 15.26 sets $[V_1, V_1] = 0$ and leaves the other brackets alone. Reading Equation (15.24) against Equations (14.128) and (14.129): for σ_{st}, $V_1 = \langle H, P_i \rangle$ and the brackets inside it are $[P_i, H] = \nu G_i$ and $[P_i, P_j] = \gamma\nu\epsilon_{ijk} J_k$, so exactly $\nu \to 0$, while $[G, P]$ and $[G, H]$ are $[V_0, V_1]$ brackets and are untouched. For σ_{ss}, $V_1 = \langle G_i, P_i \rangle$ contains $[G_i, P_j] = \gamma\delta_{ij} H$, so $\gamma \to 0$. For σ_{sT}, $V_1 = \langle G_i, H \rangle$ contains $[G_i, H] = \mu P_i$, so $\mu \to 0$. For σ_{gen}, $V_1 = \langle G_i, H, P_i \rangle$ contains all three of Equation (14.128), so all three vanish together. □

Remark 15.45 (A correction this forces: the general splitting gives the static algebra, not Carroll). Corollary 15.44 settles a point on which the literature is easy to misread, and on which an earlier draft of Table 15.1 was wrong. The splitting σ_{gen} is often labelled "Carroll". It cannot be: it sets γ, μ and ν to zero simultaneously, and Carroll is $(*, 0, 0)$ with $\gamma \neq 0$. What σ_{gen} produces is the *static* algebra $(0, 0, 0)$, the opposite vertex of the cube. Carroll is reached instead by σ_{sT} applied to the *Poincaré* algebra: Poincaré is $(*, *, 0)$ and σ_{sT} kills μ, leaving $(*, 0, 0)$. In cube terms Carroll is two edges away from de Sitter, so no single graded contraction of $\mathfrak{so}(D-1, 2)$ can reach it; the labelling that suggests otherwise is comparing a one-step move with a two-step one. Reaching Carroll in one step from de Sitter would need a rescaling that is not constant on V_1 — a generalized contraction rather than a graded one — which is outside Definition 15.18.

Rests on Corollary 15.44.

Corollary 15.46 (The sign data, hypothesis (H3), and the count eleven)**.** *Rescaling the basis of $A(\lambda)$ by $e_X \mapsto s_X e_X$ with $s_X \neq 0$ is the rescaling freedom of Theorem 14.90, and its invariants are the three signs*

$$\operatorname{sgn}\left(e_G^2\right), \qquad \operatorname{sgn}\left(e_H^2\right), \qquad \operatorname{sgn}\left(e_P^2\right). \tag{15.56}$$

They obey the following, and the eleven of Theorem 14.90 follows.

1. *If all three λ vanish nowhere then the product of the three signs is $\operatorname{sgn}\left(\lambda_G \lambda_H \lambda_P\right)^2 = +1$, so an even number of them is negative.*
2. *Hypothesis (H3) — each boost generates a noncompact one-parameter subgroup — is exactly the condition*
$$\boxed{e_G^2 \geq 0}, \tag{15.57}$$
the boost direction of A being split or nilpotent but never complex.
3. *Counting the vertices with their surviving free signs gives $2+1+2+2+1+1+1+1 = 11$.*

Rests on Theorems 14.90 and 15.43 and Propositions 15.23 and 15.42.

Derivation. *Derives Corollary 15.46.*

Under $e_X \mapsto s_X e_X$ the products Equation (15.50) give $\lambda_H \mapsto \lambda_H s_G s_P / s_H$ and cyclically, so no λ has an invariant sign; but $e_X^2 \mapsto s_X^2 e_X^2$, so the three signs Equation (15.56) are invariant, and by Equation (15.51) they are the pairwise sign products of the λ — which is the same invariant that Corollary A.59 extracts from the generators. Claim (1) is Equation (15.51) multiplied out.

(2) By Theorem 15.43 the boosts of $\mathfrak{R}(\lambda)$ satisfy $\left[G'_i, G'_j\right] = -\left(\gamma\lambda_H\right)\left(\mu\lambda_P\right)\epsilon_{ijk} J'_k = -\gamma\mu\, e_G^2\, \epsilon_{ijk} J'_k$. Together with $[J, J] \sim J$ and $[J, G'] \sim G'$, the span $\langle J_{ij}, G'_i \rangle$ closes on $\mathfrak{so}(4)$ when the coefficient of $\epsilon_{ijk} J'_k$ is positive and on $\mathfrak{so}(3,1)$ when it is negative; in the first case the boosts generate a circle and in the second a line, and when the coefficient vanishes they span an abelian $\mathbb{R}^3$, again noncompact. So (H3) holds precisely when $-\gamma\mu\, e_G^2 \leq 0$. Normalising the parent by Proposition 15.23, which has $\gamma\mu = 1 > 0$, this is Equation (15.57).

(3) Go through the eight vertices, using Equation (15.51) to see which of the three signs survives. Where all three λ are nonzero, (H3) fixes $\operatorname{sgn}(e_G^2) = +$ and then (1) forces $\operatorname{sgn}(e_H^2) = \operatorname{sgn}(e_P^2)$: one free sign, *two* algebras (de Sitter and anti-de Sitter). Where $\lambda_G = 0$ we have $e_H^2 = e_P^2 = 0$ and $e_G^2 = \lambda_H \lambda_P$, whose sign is fixed by (H3): *one* algebra (Poincaré). Where $\lambda_H = 0$, only $e_H^2 = \lambda_G \lambda_P$ survives and (H3) says nothing about it: *two* (Newton–Hooke and anti-Newton–Hooke). Where $\lambda_P = 0$, only $e_P^2 = \lambda_G \lambda_H$ survives, again unconstrained: *two* (para-Poincaré and inhomogeneous SO(4)). At each of the three vertices with two vanishing λ all three squares vanish, so no sign is left: *one* each (Galilei, Carroll, para-Galilei). At the origin, *one* (static). The total is $2+1+2+2+1+1+1+1 = 11$, and

the three vertices carrying two entries are the three at which a sign survives unconstrained — which is Theorem 14.90. □

Table 15.2: The contraction cube of Theorem 14.90 read off the structure constants of the algebra $A(\lambda)$ of Proposition 15.42. The first column is the vanishing pattern of $(\lambda_G, \lambda_H, \lambda_P)$, $*$ marking a nonzero one; by Equation (15.52) it determines the pattern of (γ, μ, ν) in the second, which is the third column of Table 14.1. The fourth column is the sign left free after hypothesis (H3), by Corollary 15.46, and it is what doubles three of the eight rows into eleven algebras. Names are those of Table 14.1.

$(\lambda_G, \lambda_H, \lambda_P)$	(γ, μ, ν)	Algebra	Free sign after (H3)
$(*, *, *)$	$(*, *, *)$	de Sitter, anti-de Sitter	$\operatorname{sgn} e_H^2 = \operatorname{sgn} e_P^2$; $\operatorname{sgn} e_G^2 = +$ fixed
$(0, *, *)$	$(*, *, 0)$	Poincaré	none: only e_G^2 survives, and (H3) fixes it
$(*, 0, *)$	$(0, *, *)$	Newton–Hooke, anti-Newton–Hooke	$\operatorname{sgn} e_H^2$
$(*, *, 0)$	$(*, 0, *)$	para-Poincaré, inhomogeneous SO(4)	$\operatorname{sgn} e_P^2$
$(0, 0, *)$	$(0, *, 0)$	Galilei	none: every square vanishes
$(0, *, 0)$	$(*, 0, 0)$	Carroll	none
$(*, 0, 0)$	$(0, 0, *)$	para-Galilei	none
$(0, 0, 0)$	$(0, 0, 0)$	static	none

Remark 15.47 (Which axis is the cosmological constant). The three axes of Table 15.2 are not interchangeable, because they carry different physics. The λ_G axis is the one along which Theorem 15.29 runs: setting $\lambda_G = 0$ is the space–time grading, the flat limit, and by Corollary 15.32 the parameter there is $t = -2\Lambda/((D-1)(D-2))$. So λ_G *is* the cosmological constant up to that positive factor, its sign is the sign of $-\Lambda$, and the split/complex/dual trichotomy of Proposition 15.28 along this one axis is the anti-de Sitter/de Sitter/flat trichotomy. The other two axes are the two velocity limits: $\lambda_H \to 0$ is $c \to \infty$ and $\lambda_P \to 0$ is $c \to 0$, and by Remark 15.33 the measured $\Lambda > 0$ puts nature at $\lambda_G < 0$ with the other two coordinates nonzero — the de Sitter vertex, of which the Poincaré algebra of Part IV is the $\lambda_G \to 0$ edge and the Galilei algebra of Part III the further $\lambda_H \to 0$ edge.

Remark 15.48 (What this does and does not replace). Corollary 15.46 reproduces the number eleven, and it explains *why* three of the eight vertices carry two entries: those are the three at which one of the signs Equation (15.56) survives without being fixed by (H3). It does not replace Appendix A.10. That appendix proves that the hypotheses of Theorem 14.90 admit *no other* algebras — that the five unknown constants reduce to three, and that the list is exhaustive — and it does so by exhausting the possible structure constants, which is a statement about all kinematical algebras and not about the ones this construction happens to produce. The present section takes the parametrisation (γ, μ, ν) as given by that theorem and shows that the cube it organises is the degeneration lattice of a single four-dimensional commutative algebra. Existence here, exhaustiveness there.

15.8 Invariant forms and why the extension is needed

An algebra is of no use for building an action unless it carries a nondegenerate invariant form: that is the input to the invariant polynomials of Section 14.7 and to the Chern–Simons

construction of Section 14.8.1. The forms on $\mathfrak{g}_A$ factorise just as the brackets do, and the criterion that results explains why the extension of Corollary 15.36 is not optional.

Definition 15.49 (k-trace)**.** A symmetric k-linear map $\Phi : A^k \to \mathbb{R}$ is a *k-trace* if

$$\Phi(ab, c_2, \dots, c_k) = \Phi(a, bc_2, c_3, \dots, c_k) \tag{15.58}$$

for all arguments.

Lemma 15.50 (Every k-trace comes from a linear functional)**.** *Let A be commutative associative and unital. Then Φ is a k-trace if and only if*

$$\Phi(a_1, \dots, a_k) = \phi\,(a_1 \cdots a_k) \tag{15.59}$$

for a unique linear functional $\phi : A \to \mathbb{R}$, namely $\phi(x) = \Phi(x, 1_A, \dots, 1_A)$.
Rests on Definition 15.49.

Derivation. *Derives Lemma 15.50.*

If Φ has the form Equation (15.59) then Equation (15.58) holds because $(ab)c_2 \cdots c_k = a(bc_2)c_3 \cdots c_k$ by associativity, and Φ is symmetric by commutativity.

Conversely let Φ be a k-trace and put $\phi(x) := \Phi(x, 1_A, \dots, 1_A)$. Using Equation (15.58) with $b = a_2$ and $c_2 = 1_A$,

$$\Phi(a_1, a_2, a_3, \dots, a_k) = \Phi(a_1, a_2 1_A, a_3, \dots, a_k) = \Phi(a_1 a_2, 1_A, a_3, \dots, a_k)\,,$$

which absorbs one argument into the first slot and replaces it by the unit. Symmetry of Φ then allows the next argument to be moved into the second slot, and repeating $k-1$ times gives $\Phi(a_1, \dots, a_k) = \Phi(a_1 \cdots a_k, 1_A, \dots, 1_A) = \phi(a_1 \cdots a_k)$. Uniqueness follows by evaluating Equation (15.59) at $a_2 = \cdots = a_k = 1_A$. □

Theorem 15.51 (Factorisation of invariant forms)**.** *Let ω be an invariant symmetric k-linear form on $\mathfrak{g}$,*

$$\sum_{i=1}^{k} \omega\,(X_1, \dots, [Y, X_i], \dots, X_k) = 0 \qquad \text{for all } Y, X_1, \dots, X_k \in \mathfrak{g}\,, \tag{15.60}$$

and let $\phi : A \to \mathbb{R}$ be linear. Then

$$\boxed{\langle a_1 \otimes X_1, \dots, a_k \otimes X_k \rangle_{\omega,\phi} := \phi\,(a_1 \cdots a_k)\ \omega\,(X_1, \dots, X_k)} \tag{15.61}$$

is an invariant symmetric k-linear form on $\mathfrak{g}_A$.
Rests on Definition 15.3.

Derivation. *Derives Theorem 15.51.*

Symmetry and k-linearity of Equation (15.61) are clear, the first using commutativity of A. For invariance take $Y = a \otimes X$ and compute with Equation (15.2):

$$\sum_{i=1}^{k} \langle a_1 \otimes X_1, \dots, [a \otimes X, a_i \otimes X_i], \dots \rangle_{\omega,\phi} = \sum_{i=1}^{k} \phi\,(a_1 \cdots (a a_i) \cdots a_k)\, \omega\,(X_1, \dots, [X, X_i], \dots)\,.$$

By commutativity and associativity of A the argument of ϕ equals $a\,a_1 \cdots a_k$ for every i, so that scalar factors out of the sum and leaves Equation (15.60), which vanishes. □

The natural question is whether Equation (15.61) exhausts the invariant forms of $\mathfrak{g}_A$. It does not, and identifying exactly which forms it produces is worth doing precisely, because the answer decides what Theorem 15.58 below is a statement about.

Definition 15.52 (Balanced form)**.** $\mathfrak{g}_A$ carries an action of A by $a \cdot (b \otimes X) = (ab) \otimes X$, for which the bracket is A-bilinear: $[a \cdot u, v] = a \cdot [u, v]$. A k-linear form Ω on $\mathfrak{g}_A$ is *balanced* if moving that action between slots costs nothing,

$$\boxed{\Omega\left(a \cdot u_1, u_2, \dots, u_k\right) = \Omega\left(u_1, a \cdot u_2, u_3, \dots, u_k\right)} \tag{15.62}$$

for all $a \in A$ and all $u_i \in \mathfrak{g}_A$.

Rests on Definition 15.3.

Theorem 15.53 (The factorised forms are exactly the balanced ones)**.** *Write* $\mathrm{Inv}^k(\mathfrak{g})$ *for the space of invariant symmetric* k*-linear forms on* $\mathfrak{g}$*, and let* $\{\omega_1, \dots, \omega_r\}$ *be a basis of it. An invariant symmetric* k*-linear form* Ω *on* $\mathfrak{g}_A$ *is balanced if and only if there are linear functionals* $\phi_1, \dots, \phi_r : A \to \mathbb{R}$ *with*

$$\boxed{\Omega\left(a_1 \otimes X_1, \dots, a_k \otimes X_k\right) = \sum_{j=1}^{r} \phi_j\left(a_1 \cdots a_k\right)\, \omega_j\left(X_1, \dots, X_k\right)}, \tag{15.63}$$

that is, if and only if Ω *is a sum of forms Equation* (15.61)*.*

Rests on Definition 15.52, Theorem 15.51, and Lemma 15.50.

Derivation. *Derives Theorem 15.53.*

Only if. For $\Omega = \langle\cdot\rangle_{\omega,\phi}$ both sides of Equation (15.62) equal $\phi\left(a\, a_1 \cdots a_k\right) \omega(X_1, \dots, X_k)$, by commutativity and associativity of A; sums of such forms inherit the property.

If. For $x \in A$ define a k-linear form on $\mathfrak{g}$ by

$$\omega^{(x)}(X_1, \dots, X_k) := \Omega\left(x \otimes X_1,\, 1_A \otimes X_2, \dots, 1_A \otimes X_k\right) . \tag{15.64}$$

It is *symmetric*: balancedness moves x from the first slot to the second, and the symmetry of Ω then exchanges the two slots entirely, so $\omega^{(x)}$ is unchanged under the transposition $X_1 \leftrightarrow X_2$; the same argument applies to any transposition, and transpositions generate the symmetric group. It is *invariant*: since $[1_A \otimes Z, a \otimes X] = a \otimes [Z, X]$, the invariance of Ω under $1_A \otimes Z$, evaluated on the arguments of Equation (15.64), is literally $\sum_i \omega^{(x)}(X_1, \dots, [Z, X_i], \dots, X_k) = 0$. So $x \mapsto \omega^{(x)}$ is a linear map $A \to \mathrm{Inv}^k(\mathfrak{g})$, and expanding in the basis gives functionals $\phi_j \in A^*$ with $\omega^{(x)} = \sum_j \phi_j(x)\, \omega_j$.

It remains to see that Ω is recovered from $\omega^{(\cdot)}$ by multiplying the A-arguments. Fix $X_1, \dots, X_k$ and set $\Phi(a_1, \dots, a_k) := \Omega(a_1 \otimes X_1, \dots, a_k \otimes X_k)$, a k-linear map $A^k \to \mathbb{R}$. Balancedness is exactly Equation (15.58), so Φ is a k-trace, and Lemma 15.50 gives $\Phi(a_1, \dots, a_k) = \Phi(a_1 \cdots a_k, 1_A, \dots, 1_A) = \omega^{(a_1 \cdots a_k)}(X_1, \dots, X_k)$, which is Equation (15.63). □

Theorem 15.54 (Completeness in the bilinear case)**.** *Let* $\mathfrak{g}$ *be* absolutely simple*, that is* $\mathfrak{g} \otimes_{\mathbb{R}} \mathbb{C}$ *simple. Then every invariant symmetric* bilinear *form on* $\mathfrak{g}_A$ *is automatically balanced, and therefore*

$$\Omega\left(a \otimes X,\, b \otimes Y\right) = \phi(ab)\, \kappa_{\mathfrak{g}}(X, Y) \tag{15.65}$$

for a unique linear $\phi : A \to \mathbb{R}$*. The space of invariant symmetric bilinear forms on* $\mathfrak{g}_A$ *has dimension* $\dim A$*.*

Rests on Definition 15.3 and Lemmas 15.50 and A.63.

Full derivation: Appendix A.11 (p. 2804).

Derives Theorem 15.54.

Proposition 15.55 (Completeness fails for $k \geq 3$)**.** *Let* $\mathfrak{g}$ *be simple,* $A = A_1 = \mathbb{R}[\varepsilon]/(\varepsilon^2)$*, and let* $B := \langle\cdot,\cdot\rangle_{\kappa_{\mathfrak{g}},\phi_1}$ *be the invariant bilinear form of Example 15.57 with* $\phi_1(1) = 0$ *and* $\phi_1(\varepsilon) = 1$*. Then the symmetric* 4*-linear form*

$$Q(u_1, u_2, u_3, u_4) := \tfrac{1}{3}\Big(B(u_1, u_2)B(u_3, u_4) + B(u_1, u_3)B(u_2, u_4) + B(u_1, u_4)B(u_2, u_3)\Big) \tag{15.66}$$

is invariant, and is not *of the shape Equation* (15.61)*. Hence the converse of Theorem 15.51 is false for $k = 4$, and by taking products with further copies of B for every even $k \geq 4$.*
Rests on Theorems 15.51 and 15.53 and Example 15.57.

Derivation. *Derives Proposition 15.55.*

Q is the full polarisation of the invariant polynomial $u \mapsto B(u,u)^2$; a product of invariant polynomials is invariant, because the invariants form a ring, so Q is an invariant symmetric 4-linear form. Now evaluate it on $u_1 = 1 \otimes X_1$, $u_2 = 1 \otimes X_2$, $u_3 = \varepsilon \otimes X_3$, $u_4 = \varepsilon \otimes X_4$. From $\phi_1(1) = 0$, $\phi_1(\varepsilon) = 1$ and $\phi_1(\varepsilon^2) = \phi_1(0) = 0$,

$$B(u_1,u_2) = 0\,, \quad B(u_3,u_4) = 0\,, \quad B(u_1,u_3) = \kappa_{\mathfrak{g}}(X_1,X_3)\,, \quad B(u_2,u_4) = \kappa_{\mathfrak{g}}(X_2,X_4)\,,$$

and similarly for the last pairing, so that

$$Q(u_1,u_2,u_3,u_4) = \tfrac{1}{3}\Big(\kappa_{\mathfrak{g}}(X_1,X_3)\kappa_{\mathfrak{g}}(X_2,X_4) + \kappa_{\mathfrak{g}}(X_1,X_4)\kappa_{\mathfrak{g}}(X_2,X_3)\Big)\,, \tag{15.67}$$

which is nonzero — take $X_3 = X_1$ and $X_4 = X_2$ with $\kappa_{\mathfrak{g}}(X_1,X_1) \neq 0 \neq \kappa_{\mathfrak{g}}(X_2,X_2)$, possible because $\kappa_{\mathfrak{g}}$ is nondegenerate. But the product of the four algebra elements is $1 \cdot 1 \cdot \varepsilon \cdot \varepsilon = \varepsilon^2 = 0$, so any form of the shape Equation (15.61) would give $\phi(0)\,\omega = 0$ there. Hence Q is not of that shape; equivalently, by Theorem 15.53, Q is not balanced. □

Remark 15.56 (Why $k = 2$ is special, and what the failure costs). The counterexample is not exotic. Taking $\mathfrak{g} = \mathfrak{so}(2,1)$ makes $\mathfrak{g}_{A_1}$ the three-dimensional Poincaré algebra $\mathfrak{iso}(2,1)$ — one of the low-dimensional coincidences of Remark 15.30, since there the adjoint and vector representations of $\mathfrak{so}(2,1)$ agree — and B its familiar pairing $\langle J_a, P_b\rangle = \eta_{ab}$ with $\langle J_a, J_b\rangle = \langle P_a, P_b\rangle = 0$; then Equation (15.67) is $Q(J_a,J_b,P_c,P_d) = \frac{1}{3}\,(\eta_{ac}\eta_{bd} + \eta_{ad}\eta_{bc})$. The mechanism is visible there: Q is a *product* of two invariant forms, and $\phi_1(a_1a_2)\phi_1(a_3a_4)$ is not a function of $a_1a_2a_3a_4$.

That also says why $k = 2$ escapes. A product of invariants of lower degree needs factors of degree at least one, and $\mathrm{Inv}^1(\mathfrak{g}) = (\mathfrak{g}^*)^{\mathfrak{g}} = 0$ for $\mathfrak{g}$ perfect, so an invariant *bilinear* form cannot be a product and has no room to be unbalanced — Theorem 15.54 then makes that precise.

Nothing in this book needs more: the only invariant forms used anywhere are bilinear, and Theorem 15.58 classifies those completely. What is lost is the stronger reading under which Equation (15.61) would describe every invariant tensor of an expanded algebra in any degree. It does not, and a Chern–Simons construction in five or more dimensions, which needs an invariant form of degree three or higher, must therefore determine its invariant tensors directly rather than by quoting a factorisation.

Example 15.57 ($k = 2$ recovers the Killing form)**.** Taking $k = 2$, $\omega = \kappa_{\mathfrak{g}}$ and $\phi = \tau_A$ in Equation (15.61) reproduces Equation (15.8). The Killing form is thus one member of a family parametrised by the linear functionals on A, and not always the useful member: on $\mathfrak{g}\langle 1\rangle$ the choice $\phi(\varepsilon) \neq 0$ gives a nondegenerate invariant form even though τ_{A_1} is degenerate, since the Gram matrix of $\phi(ab)$ in the basis $\{1,\varepsilon\}$ is

$$\begin{pmatrix} \phi(1) & \phi(\varepsilon) \\ \phi(\varepsilon) & 0 \end{pmatrix}\,,$$

of determinant $-\phi(\varepsilon)^2$. This is the invariant form that makes the Poincaré algebra usable as a gauge algebra.
Rests on Theorems 15.10 and 15.51.

Theorem 15.58 (Nondegeneracy on a resonant subalgebra)**.** *Let $\mathfrak{g}$ be simple with a $\mathbb{Z}_2$-grading σ, let $\omega = \kappa_{\mathfrak{g}}$, and let $\phi : A_N \to \mathbb{R}$ be linear. The restriction of $\langle\cdot,\cdot\rangle_{\kappa,\phi}$ to $\mathfrak{R}_N^{\sigma}$ is nondegenerate if and only if*

$$\boxed{N \text{ is even and } \phi(\varepsilon^N) \neq 0}\,. \tag{15.68}$$

Rests on Theorem 15.51, Definition 15.20, and Lemma 15.19.

Derivation. *Derives Theorem 15.58.*

By Equation (15.61), for $X \in V_{[i]}$ and $Y \in V_{[j]}$,

$$\left\langle \varepsilon^i \otimes X,\ \varepsilon^j \otimes Y \right\rangle_{\kappa,\phi} = \phi\left(\varepsilon^{i+j}\right) \kappa_{\mathfrak{g}}(X, Y)\,. \tag{15.69}$$

From Equation (15.61) (take $\omega = \kappa_{\mathfrak{g}}$ and $a_1, a_2 = \varepsilon^i, \varepsilon^j$).

Two things can make this vanish. By Lemma 15.19, $\kappa_{\mathfrak{g}}(V_0, V_1) = 0$, so Equation (15.69) is zero unless $[i] = [j]$, that is unless $i + j$ is even; and $\phi(\varepsilon^{i+j}) = 0$ whenever $i + j > N$.

Suppose N is odd and take $i = N$, so $[N] = 1$. Any j with $[j] = [N]$ is odd, whence $i + j \geq N + 1 > N$ and Equation (15.69) vanishes for every j. The whole top level $\varepsilon^N \otimes V_1$ lies in the radical of the form, which is therefore degenerate.

Suppose N is even and $\phi(\varepsilon^N) \neq 0$. Order the levels $0, 1, \ldots, N$ and consider the block matrix of the form. Every entry with $i + j > N$ vanishes, and the anti-diagonal entries $i + j = N$ are $\phi(\varepsilon^N)\,\kappa_{\mathfrak{g}}|_{V_{[i]}}$, which are nondegenerate: $[i] = [N - i]$ because N is even, so the two levels pair within the same graded piece, and $\kappa_{\mathfrak{g}}$ restricted to V_0 and to V_1 is nondegenerate by Lemma 15.19. A block matrix whose entries vanish strictly beyond the anti-diagonal and whose anti-diagonal blocks are invertible has determinant equal, up to sign, to the product of those blocks' determinants, hence is invertible.

Finally, if N is even but $\phi(\varepsilon^N) = 0$ the whole anti-diagonal vanishes and the top level is again in the radical: pairing $\varepsilon^N \otimes X$ with $\varepsilon^j \otimes Y$ requires $j = 0$ and gives $\phi(\varepsilon^N)\kappa(X, Y) = 0$. □

Remark 15.59 (The central charge is what makes the form work). Theorem 15.58 is the precise reason the Galilei algebra admits no invariant pairing of the kind an action needs, while the Bargmann algebra does. Galilei is $\mathfrak{R}_1^\sigma$ with $N = 1$ odd: its level-1 sector, the boosts and the space translations, is orthogonal to everything. The extension to $N = 2$ supplies exactly the partner each generator needs, and Equation (15.69) names the pairings: level 1 with level 1, and level 0 with the top level — that is, H with the mass M. The central charge is not decoration; it is the dual partner without which the invariant form has a radical. This also explains why the useful truncations are the even ones.

What follows mathematically is the existence of the corresponding invariant tensors. What does *not* follow is any physical claim: three-dimensional Chern–Simons gravity built on these algebras has no propagating degrees of freedom and no observational status, as Remark 14.94 records, and this treatise carries the construction as mathematics only.

Rests on Theorem 15.58 and Corollary 15.36.

15.9 Relation to the published expansion methods

Two constructions in the literature are special cases of Definition 15.3, and saying which is the honest way to place this material.

Proposition 15.60 (Semigroup expansion)**.** *Let S be a finite abelian semigroup and $\mathbb{R}[S]$ its semigroup algebra, with basis $\{\lambda_\alpha\}_{\alpha \in S}$ and $\lambda_\alpha \lambda_\beta = \lambda_{\alpha\beta}$. Then $\mathfrak{g}_{\mathbb{R}[S]}$ is the S-expanded algebra of [2096]: writing $\lambda_\alpha \lambda_\beta = K_{\alpha\beta}{}^\gamma \lambda_\gamma$,*

$$[\lambda_\alpha \otimes T_a, \lambda_\beta \otimes T_b] = K_{\alpha\beta}{}^\gamma C_{ab}{}^c\ \lambda_\gamma \otimes T_c\,. \tag{15.70}$$

Rests on Definition 15.3 and Theorem 15.4.

Derivation. *Derives Proposition 15.60.*

$\mathbb{R}[S]$ is commutative because S is abelian and associative because S is, and it is unital after adjoining a unit if S has none. Its structure constants in the basis $\{\lambda_\alpha\}$ are $\mu_{\alpha\beta}{}^\gamma = K_{\alpha\beta}{}^\gamma$, so Equation (15.3) is Equation (15.70). □

Example 15.61 (Maurer–Cartan expansion)**.** The expansion of [2044] rescales the group parameters by a formal variable, expands the Maurer–Cartan forms in powers of it and truncates at a chosen order. In the present language the expansion variable is ε, truncation at order N is the passage to $A_N = \mathbb{R}[\varepsilon]/(\varepsilon^{N+1})$, and the resulting algebra is $\mathfrak{g}\langle N\rangle$ of Definition 15.14 or, when the method's grading conditions are imposed, the resonant subalgebra $\mathfrak{R}^{\sigma}_{N}$ of Definition 15.20. The Maurer–Cartan method is thus the case $A = A_N$ of Definition 15.3, and Proposition 15.60 is the case $A = \mathbb{R}[S]$.

Rests on Definitions 15.3, 15.14 and 15.20.

Proposition 15.62 (The complex line is not a semigroup algebra)**.** *There is no finite abelian semigroup S with $\mathbb{R}[S] \cong \mathbb{C}$.*

Derivation. *Derives Proposition 15.62.*

Suppose $\mathbb{R}[S] \cong \mathbb{C}$. Then $|S| = 2$ and the image of S in $\mathbb{C}$ is an $\mathbb{R}$-basis closed under multiplication. In a finite semigroup every element has an idempotent power, and the only idempotents of $\mathbb{C}$ are 0 and 1, so each element of the image is 0 or a root of unity. If 0 lies in the image the two elements span a real line, not $\mathbb{C}$. So the image is $\{z, w\}$ with z, w roots of unity, $\mathbb{R}$-independent, and $z^2 \in \{z, w\}$.

If $z^2 = z$ then $z = 1$, and $w^2 \in \{1, w\}$ forces $w = \pm 1$, dependent on z. If $z^2 = w$ then $w^2 = z^4 \in \{z, w\}$: from $z^4 = z$ we get $z^3 = 1$ with $z \neq 1$, so $w = z^2$ and $zw = z^3 = 1 \notin \{z, w\}$, contradicting closure; and from $z^4 = w = z^2$ we get $z^2 = 1$, hence $w = 1$ and $z = -1$, again dependent. No case survives. □

Remark 15.63 (The inclusion is strict, and the strictness is physical)*.* The tensor product construction is genuinely larger than semigroup expansion. By Proposition 15.28 the algebra $A^{(t)}$ is isomorphic to $\mathbb{C}$ exactly on the branch $t < 0$, which by Corollary 15.32 is the branch with $\Lambda > 0$ — the one the measured cosmological constant selects (Remark 15.33) — and by Proposition 15.62 that two-dimensional algebra is not $\mathbb{R}[S]$ for any S.

The tempting next sentence is that the de Sitter side is therefore outside the reach of semigroup expansion altogether. That sentence is false, and the counterexample is small: for the cyclic group $\mathbb{Z}_3$,

$$\mathbb{R}[\mathbb{Z}_3] \cong \mathbb{R}[x]/(x^3 - 1) \cong \mathbb{R} \times \mathbb{C}\,, \tag{15.71}$$

since $x^3 - 1 = (x-1)(x^2+x+1)$ and the quadratic factor is irreducible over $\mathbb{R}$; more generally $\mathbb{R}[\mathbb{Z}_n]$ acquires a complex factor for every $n \geq 3$. So a large enough semigroup does supply one. What Proposition 15.62 establishes is the narrow statement only: the *two-dimensional* complex algebra is not a semigroup algebra, so the de Sitter branch of the quadratic family Equation (15.26) is not itself a semigroup expansion. Whether the particular semigroup families used in the gauge-gravity literature reach it is a separate question, not settled here, and worth care — Equation (15.71) says the answer is not automatically no.

Rests on Propositions 15.28 and 15.62 and Corollary 15.32.

15.10 The construction in 3+1 dimensions

Every statement above is dimension-independent unless marked otherwise. Because the physical parts of this treatise work in $D = 4$, the instantiation is collected here in one place.

Example 15.64 (The whole chapter at $D = 4$)**.** Take $\mathfrak{g} = \mathfrak{so}(3, 2)$, of dimension 10, with generators J_{ij} $(i, j = 1, 2, 3)$, G_i, H, P_i as in Notation 15.22.

1. *Curvature.* Theorem 15.29 gives the ten-dimensional resonant subalgebra $\mathfrak{R}^{(t)} \subseteq \mathfrak{so}(3, 2)_{A^{(t)}}$, isomorphic to $\mathfrak{so}(3, 2)$ for $t > 0$, to $\mathfrak{so}(4, 1)$ for $t < 0$ and to $\mathfrak{iso}(3, 1)$ for $t = 0$. By Proposition 15.31, $\Lambda = \pm 3/\ell^2$ and, by Corollary 15.32, $t = -\Lambda/3$.

2. *Which branch.* $\Lambda \approx 1.1\times10^{-52}/\mathrm{m}^2 > 0$ from Equation (15.39), so $t \approx -3.7\times10^{-53}/\mathrm{m}^2$ and the algebra is $\mathfrak{so}(4,1)$: de Sitter. The curvature radius $\ell \approx 1.6\times10^{26}\,\mathrm{m}$ is some 19 orders of magnitude larger than the solar system, which is why the flat truncation $t=0$ — the Poincaré algebra of Part IV — is an excellent approximation for every laboratory and astrophysical test in this book.

3. *Central charges.* With the speed–space grading, $V_0 = \langle J_{ij}, H\rangle = \mathfrak{so}(3)\oplus\langle H\rangle$ has centre $\langle H\rangle$, because $\mathfrak{so}(3)$ is simple. By Corollary 15.35 the level-2 truncation therefore carries *exactly one* central charge, $M = \varepsilon^2\otimes H$, and Corollary 15.36 identifies it with the mass: the Bargmann algebra in 3+1, the algebra actually realised by the Poisson brackets of Proposition 25.14. The second charge S of the extended Bargmann algebra requires $\mathfrak{so}(D-1)$ to be abelian and so exists only in 2+1.

4. *What does not carry over from $D=3$.* The isomorphisms Equation (15.32) are false at $D=4$: $\mathfrak{so}(3,2)\oplus\mathfrak{so}(3,2)$ has dimension 20 and is not an isometry algebra of any four-dimensional maximally symmetric spacetime, and $\mathbb{C}\otimes_{\mathbb{R}}\mathfrak{so}(3,2)$ has dimension 20 rather than the 10 of $\mathfrak{so}(4,1)$. Only the resonant subalgebra $\mathfrak{R}^{(t)}$ has the right dimension, which is why Theorem 15.29 is stated in terms of it.

Rests on Theorem 15.29, Proposition 15.31, and Corollaries 15.32, 15.35 and 15.36.

Remark 15.65 (What is proved here and what is borrowed). For the record, since Remark 15.2 withholds the manuscript's authority: every numbered statement of this chapter is proved above or in Appendix A.11, from Definition 15.3 and standard linear algebra. Reproving rather than quoting earned its keep twice. Remark 15.45 corrects the labelling of the general splitting, which gives the static algebra and not Carroll; and Proposition 15.55 *disproves* the manuscript's factorisation converse for $k\geq 3$ by an explicit invariant form on the three-dimensional Poincaré algebra, so what appears here as Theorems 15.53 and 15.54 is a corrected statement and not a weakened one. The only results quoted without proof are Lemma 15.12, textbook commutative algebra, and the real spectral theorem for a commuting family of symmetric operators used in Theorem 15.68, textbook linear algebra; neither is used in a claim about a specific algebra — the three cases of Proposition 15.28, the tower of Section 15.3, the branches of Theorem 15.29 and the central charges of Corollary 15.36 are each verified by direct computation on the algebra concerned.

15.11 Positive functionals and the Gram form

Theorem 15.58 decided when the factorised bilinear form of Theorem 15.51 is *nondegenerate*. As soon as such a form is used as a kinetic pairing — a quadratic form on fields whose sign decides whether a mode carries positive energy — the finer question is when it is *positive*. On the A-side the question concerns a single object, already met as the Gram matrix of Example 15.57. No spacetime dimension enters anywhere in this section; the $D=4$ instantiation is used by the companion manuscript as recorded in Remark 15.71.

Definition 15.66 (Gram form of a functional)**.** For a linear functional $\psi : A\to\mathbb{R}$, the *Gram form* of ψ is the symmetric bilinear form

$$B_\psi(a,b) := \psi(ab) \tag{15.72}$$

on A. By Theorem 15.54, when $\mathfrak{g}$ is absolutely simple every invariant symmetric bilinear form on $\mathfrak{g}_A$ is $B_\psi\otimes\kappa_{\mathfrak{g}}$ for a unique ψ, so the sign structure of the invariant bilinear forms of $\mathfrak{g}_A$ is exactly the sign structure of the Gram forms of A.

Rests on Theorem 15.54.

Lemma 15.67 (Multiplication is B_ψ-self-adjoint)**.** *For every $a \in A$ the multiplication operator $M_a : u \mapsto au$ satisfies $B_\psi(M_a u, v) = B_\psi(u, M_a v)$ for all $u, v \in A$, and the radical $I_\psi := \{u \in A : \psi(uA) = 0\}$ of B_ψ is an ideal of A.*

Rests on Definition 15.66.

Derivation. *Derives Lemma 15.67.*

$B_\psi(au, v) = \psi((au)v) = \psi(u(av)) = B_\psi(u, av)$, by commutativity and associativity of A. If $\psi(uA) = 0$ then for every $b \in A$ also $\psi((bu)A) = \psi(u(bA)) \subseteq \psi(uA) = 0$, so I_ψ absorbs multiplication. □

Theorem 15.68 (Positive Gram forms are positive sums of real characters)**.** *Let A be a finite-dimensional commutative associative unital $\mathbb{R}$-algebra and let $\psi : A \to \mathbb{R}$ be linear. The following are equivalent.*

1. *B_ψ is positive semidefinite.*
2. *There are finitely many* distinct $\mathbb{R}$*-algebra homomorphisms $\chi_1, \dots, \chi_r : A \to \mathbb{R}$ and weights $w_1, \dots, w_r > 0$ with*

$$\boxed{\psi = \sum_{i=1}^{r} w_i \, \chi_i}\,. \tag{15.73}$$

In that case $r = \operatorname{rank} B_\psi$. In particular ψ annihilates $\operatorname{rad}(A)$, and it annihilates every factor of the Artin decomposition (Lemma 15.12) whose residue field is $\mathbb{C}$.

Rests on Definition 15.66 and Lemmas 15.12 and 15.67.

Derivation. *Derives Theorem 15.68.*

(2)⇒(1). Multiplicativity of each χ_i gives

$$B_\psi(a, a) = \sum_{i=1}^{r} w_i \, \chi_i(a)^2 \geq 0\,.$$

(1)⇒(2). If $\psi = 0$ take $r = 0$. Otherwise let I_ψ be the radical of B_ψ, an ideal by Lemma 15.67. For a positive semidefinite form the radical coincides with the null cone — $B_\psi(u, u) = 0$ forces $u \in I_\psi$ by the Cauchy–Schwarz inequality, which holds for every positive semidefinite symmetric form — so the induced form $\bar{B}$ on $\bar{A} := A/I_\psi$ is positive *definite*, and $\bar{A}$ is again commutative, associative and unital ($1_A \notin I_\psi$ because $\psi \neq 0$). Each multiplication operator $\bar{M}_a$ on $\bar{A}$ is $\bar{B}$-symmetric by Lemma 15.67, and the family $\left\{\bar{M}_a : a \in A\right\}$ commutes, so the spectral theorem for commuting symmetric operators on a real inner-product space provides a $\bar{B}$-orthogonal basis $\bar{u}_1, \dots, \bar{u}_r$ of $\bar{A}$ and functions $\chi_i : A \to \mathbb{R}$ with $\bar{M}_a \bar{u}_i = \chi_i(a)\, \bar{u}_i$. Each χ_i is linear in a, multiplicative because $\bar{M}_{ab} = \bar{M}_a \bar{M}_b$, and unital because $\bar{M}_{1_A} = \mathrm{id}$. Writing $\bar{1} = \sum_i c_i \bar{u}_i$ and using $\psi(a) = B_\psi(a, 1_A)$,

$$\psi(a) = \bar{B}\left(\bar{M}_a \bar{1}, \bar{1}\right) = \sum_i c_i^2 \, \chi_i(a)\, \bar{B}\,(\bar{u}_i, \bar{u}_i) =: \sum_i w_i \, \chi_i(a)\,,$$

with every $w_i \geq 0$; discard the terms with $w_i = 0$ and merge repeated characters, which keeps the surviving weights positive and the characters distinct.

Distinct characters are linearly independent: a shortest nontrivial relation $\sum_i \lambda_i \chi_i = 0$ with all $\lambda_i \neq 0$ and $r \geq 2$ admits some a_0 with $\chi_1(a_0) \neq \chi_2(a_0)$; evaluating the relation at $a_0 a$ and subtracting $\chi_1(a_0)$ times the relation at a kills the first term but not the second, producing a shorter nontrivial relation. Hence $B_\psi(a, b) = \sum_i w_i \chi_i(a) \chi_i(b)$ has rank exactly r.

Finally, a character kills every nilpotent element, since $\chi(n)^k = \chi(n^k) = 0$, hence kills $\operatorname{rad}(A)$. And let A_j be a factor of the Artin decomposition, with unit the idempotent π_j and

residue field $\mathbb{C}$. From $\chi(\pi_j)^2 = \chi(\pi_j)$ either $\chi(\pi_j) = 0$, and then χ kills $A_j = \pi_j A_j$ outright, or $\chi(\pi_j) = 1$, and then χ restricts to a unital $\mathbb{R}$-algebra homomorphism $A_j \to \mathbb{R}$, which kills the maximal ideal of A_j (it is nilpotent) and so induces an $\mathbb{R}$-algebra homomorphism $\mathbb{C} \to \mathbb{R}$ — impossible, since $\chi(i)^2 = \chi(-1) = -1$ has no real solution. So every ψ of the form Equation (15.73) vanishes on $\mathrm{rad}(A)$ and on every factor whose residue field is $\mathbb{C}$. □

Corollary 15.69 (No positive pairing survives the radical)**.** *On the truncated algebra $A_N = \mathbb{R}[\varepsilon]/(\varepsilon^{N+1})$ with $N \geq 1$:*

1. *B_ψ is positive semidefinite if and only if $\psi(1) \geq 0$ and $\psi(\varepsilon^k) = 0$ for all $k \geq 1$; its rank is then at most one.*
2. *For N even and $\psi(\varepsilon^N) \neq 0$ — the nondegenerate case of Theorem 15.58 — the signature of B_ψ is $\left(\frac{N}{2}+1, \frac{N}{2}\right)$ or $\left(\frac{N}{2}, \frac{N}{2}+1\right)$: maximally split up to one unit. Nondegeneracy and positivity are therefore incompatible on every truncation with $N \geq 1$: already by (1), a positive semidefinite B_ψ has rank at most one on an algebra of dimension $N+1 \geq 2$.*

Rests on Theorems 15.58 and 15.68 and Definition 15.14.

Derivation. *Derives Corollary 15.69.*

(1) is Theorem 15.68 applied to A_N: an $\mathbb{R}$-character must send ε to a real root of $x^{N+1} = 0$, that is to 0, so the only character is evaluation at $\varepsilon = 0$ and Equation (15.73) reads $\psi = w_1 \,\mathrm{ev}_0$ with $w_1 = \psi(1)$, allowing also $w_1 = 0$ for $\psi = 0$. (2): in the basis $1, \varepsilon, \ldots, \varepsilon^N$ the matrix of B_ψ has entries $\psi(\varepsilon^{i+j})$, zero for $i + j > N$ and equal to $\psi(\varepsilon^N) \neq 0$ along the anti-diagonal $i+j = N$, so $\det B_\psi = \pm\psi(\varepsilon^N)^{N+1} \neq 0$ and the form is nondegenerate. The span of $\{\varepsilon^i : 2i > N\}$ has all pairwise products of total degree exceeding N, hence is totally isotropic for B_ψ, of dimension $N/2$; a nondegenerate symmetric form on a space of dimension $N+1$ with a totally isotropic subspace of dimension $N/2$ has at least $N/2$ positive and $N/2$ negative directions, which fixes the signature up to the placement of the one remaining unit. □

Corollary 15.70 (Signature of the factorised form)**.** *If B_ψ has signature (p_ψ, n_ψ) and $\kappa_{\mathfrak{g}}$ has signature (p_κ, n_κ), the factorised form $B_\psi \otimes \kappa_{\mathfrak{g}}$ on $\mathfrak{g}_A$ has signature*

$$(p_\psi p_\kappa + n_\psi n_\kappa,\ p_\psi n_\kappa + n_\psi p_\kappa)\ . \tag{15.74}$$

A positive definite B_ψ reproduces the sign pattern of $\kappa_{\mathfrak{g}}$ on every summand; each negative direction of B_ψ instead contributes a full reversed copy $-\kappa_{\mathfrak{g}}$, so a split B_ψ carries the reversed pattern on half of its directions.

Rests on Definition 15.66 and Theorem 15.51.

Derivation. *Derives Corollary 15.70.*

Diagonalise each factor and count the signs of the products of diagonal entries. □

Remark 15.71 (What this classifies, and its place in the chapter). Theorem 15.68 is the positivity companion of Theorem 15.58, and the two pull in opposite directions. Nondegeneracy can always be arranged on an even truncation, but it is bought with the anti-diagonal level pairing, and Corollary 15.69 shows that pairing to be maximally split; positivity, by contrast, sees only the totally real semisimple part of A — at most one positive weight per real point of the Artin decomposition — and nothing of the radical. The Gram matrix of Example 15.57, whose determinant is $-\phi(\varepsilon)^2$, is the sharpest instance: split, usable as a nondegenerate pairing, never positive. In the language of moment sequences, B_ψ on A_N is the Hankel matrix of the finite sequence $\psi(\varepsilon^k)$, and Theorem 15.68 is the statement that a positive Hankel form on a truncated polynomial algebra comes from a nonnegative measure

supported on the real spectrum — here the single point $\varepsilon = 0$. As everywhere in this chapter (Remark 15.59), the statement is mathematics: the companion manuscript [2434] applies it at $D = 4$ as a kinetic-positivity criterion, with the weights w_i playing the role of coupling constants, one per summand; no physical claim is made or needed here.

Chapter 16

Calculus of Variations

Functionals, the Euler–Lagrange equations, constraints and Lagrange multipliers, and second-variation conditions — the mathematical engine behind every action principle in this treatise. Where the differential calculus of Chapter 7 finds the points at which a function of finitely many variables is stationary, the calculus of variations finds the *curves* and *fields* at which a functional — a number assigned to a whole function, typically an integral $J[y] = \int f(x, y, y')\,\mathrm{d}x$ — is stationary. The subject was created for isolated geometric problems, systematised by Euler [23] and Lagrange [31], and became the common language of physics when mechanics, optics, field theory and general relativity were each recast as the statement that an action is stationary. The canonical modern exposition is Gelfand–Fomin [1254].

This chapter is consumed three times over: the Euler–Lagrange machinery feeds Lagrangian mechanics in Chapter 21, the several-variable theory feeds classical field theory in Chapter 64 and the gravitational action of Chapter 44, and Noether's theorem [456] is the origin of every conservation law derived in this book.

The logical order of the chapter inverts its historical order. The three problems of Section 16.1 were posed and solved before any general method existed; they are stated first because they fix what the method must deliver, and each is solved here with the machinery of Section 16.2, to which the solutions forward-refer. A reader who wants the theory before the examples may begin at Section 16.2.1 and return.

16.1 The classical problems

16.1.1 The brachistochrone

In the June 1696 issue of the *Acta Eruditorum* Johann Bernoulli challenged the mathematicians of Europe to find, among all curves joining two given points in a vertical plane, the one down which a bead sliding without friction under gravity travels in least time [17]. Solutions by Newton, Leibniz, l'Hôpital and both Bernoullis appeared in the issue of May 1697. The problem is the origin of the subject: it asks for a minimum not over a finite list of numbers but over a whole class of curves.

Set the starting point at the origin, let x be measured horizontally and y vertically *downwards*, and let the bead of mass m start from rest. Energy conservation gives the speed at depth y as $v = \sqrt{2gy}$, with $g = 9.806\,65\,\mathrm{m/s^2}$ the standard acceleration of gravity; the mass cancels, which is why the answer is a purely geometric curve. Since the arc element is $\mathrm{d}s = \sqrt{1 + y'^2}\,\mathrm{d}x$, the transit time to the endpoint (x_1, y_1) is the functional

$$T[y] = \int_0^{x_1} \sqrt{\frac{1 + y'(x)^2}{2g\,y(x)}}\,\mathrm{d}x\,, \tag{16.1}$$

measured in seconds, over curves with $y(0) = 0$, $y(x_1) = y_1 > 0$ and $y > 0$ on $(0, x_1]$.

Proposition 16.1 (The curve of quickest descent is a cycloid)**.** *Every extremal of Equation* (16.1) *is an arc of the cycloid*

$$x = a\,(\theta - \sin\theta)\,, \qquad y = a\,(1 - \cos\theta)\,, \qquad \theta \in [0, \theta_1]\,, \tag{16.2}$$

traced by a point on the rim of a circle of radius a rolling on the underside of the horizontal line $y = 0$. The radius a, of dimension length, and the terminal parameter θ_1 are fixed by the endpoint condition.

Rests on Equations (16.1) *and* (16.25)*.*

Proof. *Derives Proposition 16.1.*

The integrand $f(y, y') = \left[(1+y'^2)/(2gy)\right]^{1/2}$ carries no explicit x, so the Beltrami identity Equation (16.25) applies: $f - y'\,\partial f/\partial y'$ is constant along any extremal. With

$$\frac{\partial f}{\partial y'} = \frac{y'}{\sqrt{2gy}\,\sqrt{1 + y'^2}}\,,$$

the combination collapses to

$$f - y'\,\frac{\partial f}{\partial y'} = \frac{1}{\sqrt{2gy}\,\sqrt{1 + y'^2}} = C > 0\,,$$

that is, $y\,(1 + y'^2) = 2a$ with $2a = 1/(2gC^2)$, a length. Hence $y' = \sqrt{(2a - y)/y}$, and y never exceeds $2a$. Substituting $y = a(1 - \cos\theta) = 2a\sin^2(\theta/2)$ gives $2a - y = 2a\cos^2(\theta/2)$ and therefore $y' = \cos(\theta/2)/\sin(\theta/2)$, so that

$$\mathrm{d}x = \frac{\mathrm{d}y}{y'} = 2a\sin\frac{\theta}{2}\cos\frac{\theta}{2}\,\frac{\sin(\theta/2)}{\cos(\theta/2)}\,\mathrm{d}\theta = a\,(1 - \cos\theta)\,\mathrm{d}\theta\,.$$

Integrating from $\theta = 0$, where $x = y = 0$, gives Equation (16.2). □

Remark 16.2 (Two features the answer alone does not show)*.* The cycloid leaves the origin vertically ($y' \to \infty$ as $\theta \to 0$), so the bead is accelerated hardest at the start; and for y_1 small enough relative to x_1 the optimal arc dips *below* the endpoint and climbs back, which no argument from symmetry would suggest. That the extremal is in fact a minimum, and not merely a stationary curve, follows from the sufficiency theory of Section 16.4.4: the integrand is convex in y', so the Weierstrass excess function is nonnegative.

Remark 16.3 (Bernoulli's optical solution)*.* Johann Bernoulli's own solution did not use any variational calculus. He replaced the falling bead by a light ray crossing a medium of continuously varying refractive index $n \propto 1/v \propto y^{-1/2}$, and applied the law of refraction in the form $\sin\vartheta/v = \text{const}$ that Section 16.7.3 derives from Fermat's principle. Writing $\sin\vartheta = 1/\sqrt{1 + y'^2}$ reproduces $y(1 + y'^2) = \text{const}$ immediately. The step is not a trick: the identity of the two problems is the optical–mechanical analogy, made systematic by Hamilton and Jacobi in Chapter 23, and in this chapter it reappears as the common form of Lemma 16.100, which serves both Maupertuis' principle and Fermat's.

16.1.2 The catenary and the isoperimetric problem

A uniform flexible chain of linear mass density ρ_ℓ (in kg/m) and fixed total length ℓ hangs from two points. Galileo asserted that the resulting curve is a parabola [4]; it is not, as Huygens showed while still a teenager, and the correct curve was found in 1691 by Johann Bernoulli, Leibniz and Huygens independently [16]. In the variational reading the chain minimises its gravitational potential energy

$$U[y] = \rho_\ell\, g \int_{x_0}^{x_1} y(x)\sqrt{1 + y'(x)^2}\,\mathrm{d}x\,, \tag{16.3}$$

in joules, with y measured upwards, subject to the fixed-length constraint

$$L[y] = \int_{x_0}^{x_1} \sqrt{1 + y'(x)^2}\, \mathrm{d}x = \ell\,. \tag{16.4}$$

Proposition 16.4 (The hanging chain is a catenary)**.** *Let ℓ exceed the distance between the suspension points. Every extremal of Equation (16.3) subject to Equation (16.4) is a catenary*

$$y(x) = a \cosh\!\left(\frac{x - x_c}{a}\right) - \lambda\,, \tag{16.5}$$

with $a > 0$ a length, x_c a horizontal offset, and λ the Lagrange multiplier of the length constraint. It is not a parabola.

Rests on Equations (16.3) and (16.25) and Theorem 16.43.

Proof. *Derives Proposition 16.4.*

By Theorem 16.43 the constrained extremals are the free extremals of $U + \rho_\ell g \lambda L$, whose integrand, after division by the constant $\rho_\ell g$, is $F(y, y') = (y + \lambda)\sqrt{1 + y'^2}$. It carries no explicit x, so the Beltrami identity Equation (16.25) gives

$$F - y' \frac{\partial F}{\partial y'} = (y + \lambda)\sqrt{1 + y'^2} - \frac{(y + \lambda)\, y'^2}{\sqrt{1 + y'^2}} = \frac{y + \lambda}{\sqrt{1 + y'^2}} = a\,,$$

a constant with the dimension of length. Writing $u = y + \lambda$ this reads $u = a\sqrt{1 + u'^2}$, that is $u' = \pm\sqrt{u^2/a^2 - 1}$, whose separated form $\mathrm{d}u/\sqrt{u^2 - a^2} = \pm \mathrm{d}x/a$ integrates to $u = a\cosh\big((x - x_c)/a\big)$. Restoring $y = u - \lambda$ gives Equation (16.5). Expanding, $a\cosh(z/a) = a + z^2/(2a) + z^4/(24a^3) + \dots$, so the catenary agrees with a parabola only to second order about its vertex — which is why Galileo's conjecture survived casual inspection. □

Remark 16.5 (The same curve, three problems). The catenary is the extremal of $\int y\sqrt{1 + y'^2}\, \mathrm{d}x$ with the length fixed, and also, with no constraint at all, of the surface-area functional of Proposition 16.9; the two integrands differ only by the additive constant λ that the constraint supplies. This is not a coincidence but the multiplier rule at work: fixing the length of the generating curve and dropping the constant term are the same modification of the integrand.

Queen Dido's problem — to enclose the greatest area with a boundary of given length — is the oldest extremal problem on record and the prototype of the class that carries its name.

Proposition 16.6 (The isoperimetric extremal is a circle)**.** *Among closed, simple, piecewise-C^1 plane curves of fixed length ℓ, every extremal of the enclosed area is a circle of radius $\ell/2\pi$.*

Rests on Theorems 16.35 and 16.43.

Proof. *Derives Proposition 16.6.*

Parametrise the curve by arclength $s \in [0, \ell]$, so that $\dot{x}^2 + \dot{y}^2 = 1$, and use the shoelace form of the area, $A = \frac{1}{2}\oint (x\,\mathrm{d}y - y\,\mathrm{d}x)$. By Theorem 16.43 the constrained extremals are the free extremals of the integrand

$$F(x, y, \dot{x}, \dot{y}) = \tfrac{1}{2}(x\dot{y} - y\dot{x}) + \lambda\sqrt{\dot{x}^2 + \dot{y}^2}\,.$$

The system Equation (16.30) gives, after imposing the arclength gauge $\sqrt{\dot{x}^2 + \dot{y}^2} = 1$,

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(-\tfrac{1}{2}y + \lambda\dot{x}\right) = \tfrac{1}{2}\dot{y}\,, \qquad \frac{\mathrm{d}}{\mathrm{d}s}\left(\tfrac{1}{2}x + \lambda\dot{y}\right) = -\tfrac{1}{2}\dot{x}\,,$$

that is $\lambda\ddot{x} = \dot{y}$ and $\lambda\ddot{y} = -\dot{x}$. These are consistent with the gauge, since $\frac{\mathrm{d}}{\mathrm{d}s}(\dot{x}^2 + \dot{y}^2) = 2\dot{x}\dot{y}/\lambda - 2\dot{y}\dot{x}/\lambda = 0$, and they give the signed curvature

$$\kappa = \dot{x}\ddot{y} - \dot{y}\ddot{x} = -\frac{\dot{x}^2 + \dot{y}^2}{\lambda} = -\frac{1}{\lambda}\,,$$

constant. A plane curve of constant nonzero curvature is a circular arc, and a closed one of length ℓ is the full circle of radius $|\lambda| = \ell/2\pi$. □

Remark 16.7 (What the extremal argument does not prove). Proposition 16.6 identifies the only candidate; it does not by itself show that a maximiser exists. Steiner's nineteenth-century symmetrisation arguments have the same gap, and Weierstrass's criticism of them is what forced the existence question into the open (Section 16.5.1). Existence for this problem follows from the direct method of Section 16.5.2; Euler's founding monograph [23] treats the whole class systematically, and the general multiplier rule for integral constraints is Theorem 16.43.

16.1.3 Minimal surfaces and soap films

A surface spanning a given closed contour and minimising area is a *minimal surface*. Plateau realised such surfaces physically as soap films, whose surface tension makes the area the energy, and catalogued both the surfaces and the junction rules that films obey [186]: three films always meet along an edge at equal angles of 120 degrees, and four such edges meet at a vertex at the tetrahedral angle $\arccos(-1/3) \approx 109.47$ degrees. The physics of the surface tension that enforces this belongs to Chapter 31; the geometry is here.

For a surface given as a graph $z = u(x, y)$ over a region $\Omega \subset \mathbb{R}^2$, the area functional is

$$A[u] = \int_{\Omega} \sqrt{1 + |\nabla u|^2}\, \mathrm{d}x\, \mathrm{d}y\,. \tag{16.6}$$

Proposition 16.8 (The minimal-surface equation). *The Euler–Lagrange equation of Equation (16.6) is*

$$\left(1 + u_y^2\right) u_{xx} - 2u_x u_y\, u_{xy} + \left(1 + u_x^2\right) u_{yy} = 0\,, \tag{16.7}$$

equivalently the vanishing of the mean curvature of the graph.

Rests on Equations (16.6) and (16.32).

Proof. *Derives Proposition 16.8.*

Apply Equation (16.32) with two independent variables and the integrand $f = W \equiv \sqrt{1 + u_x^2 + u_y^2}$, which does not depend on u itself. Since $\partial f/\partial u_x = u_x/W$ and $\partial f/\partial u_y = u_y/W$, the equation reads

$$\frac{\partial}{\partial x}\left(\frac{u_x}{W}\right) + \frac{\partial}{\partial y}\left(\frac{u_y}{W}\right) = 0\,. \tag{16.8}$$

Carrying out the differentiations, with $W_x = (u_x u_{xx} + u_y u_{xy})/W$ and $W_y = (u_x u_{xy} + u_y u_{yy})/W$, and multiplying through by W^3,

$$\begin{aligned} W^2\,(u_{xx} + u_{yy}) - u_x\,(u_x u_{xx} + u_y u_{xy}) - u_y\,(u_x u_{xy} + u_y u_{yy}) \\ = \left(1 + u_y^2\right) u_{xx} - 2u_x u_y u_{xy} + \left(1 + u_x^2\right) u_{yy} = 0\,, \end{aligned}$$

which is Equation (16.7). The left-hand side of Equation (16.8) is $2H$, twice the mean curvature of the graph in the sense of Chapter 13, so the equation says $H = 0$. □

Proposition 16.9 (The catenoid)**.** *Among surfaces of revolution about the x axis generated by a positive profile $y(x)$, the extremals of the area $A = 2\pi \int y\sqrt{1+y'^2}\,\mathrm{d}x$ are the catenoids*

$$y(x) = a\cosh\!\left(\frac{x - x_c}{a}\right), \qquad a > 0\,. \tag{16.9}$$

Rests on Equations (16.6) and (16.25).

Proof. *Derives Proposition 16.9.*

The integrand $f = y\sqrt{1+y'^2}$ carries no explicit x, so Equation (16.25) gives $f - y' f_{y'} = y/\sqrt{1+y'^2} = a$, a constant length. This is the equation solved in Proposition 16.4 with $\lambda = 0$, whose solution is Equation (16.9). □

Remark 16.10 (Two classical surfaces and one warning). The catenoid and the helicoid $(v\cos u,\ v\sin u,\ cu)$ are the two minimal surfaces known to Euler and Meusnier, and they are locally isometric. Neither existence nor uniqueness is automatic: for two coaxial rings whose separation exceeds about 1.3255 times their radius no catenoid joins them at all, and the soap film jumps to the two flat discs — a physical realisation of the failure of a minimising sequence to converge inside the admissible class. The general existence statement for an arbitrary contour is the Plateau problem of Section 16.5.4.

16.2 Functionals and the first variation

16.2.1 Function spaces and functionals

Definition 16.11 (Admissible class; functional)**.** Fix $a < b$ in $\mathbb{R}$, boundary values $y_a, y_b \in \mathbb{R}$, and let

$$\mathcal{A} = \Big\{ y \in C^1[a,b] \;\Big|\; y(a) = y_a,\ y(b) = y_b \Big\} \tag{16.10}$$

be the *admissible class* with fixed endpoints. A *functional* on $\mathcal{A}$ is a map $J : \mathcal{A} \longrightarrow \mathbb{R}$. The functionals of this chapter are integrals

$$J[y] = \int_a^b f\left(x, y(x), y'(x)\right)\mathrm{d}x\,, \tag{16.11}$$

with the *integrand* (or *Lagrangian*) f of class C^2 on $[a,b]\times\mathbb{R}\times\mathbb{R}$.

Rests on Definition 7.66 and Theorem 7.40.

Definition 16.12 (The two norms; weak and strong neighbourhoods)**.** On $C^1[a,b]$ define

$$\|y\|_0 = \max_{a\le x\le b} |y(x)|\,, \tag{16.12}$$

$$\|y\|_1 = \max_{a\le x\le b} |y(x)| + \max_{a\le x\le b} |y'(x)|\,. \tag{16.13}$$

A *strong* ε-neighbourhood of y is $\{\tilde y \in \mathcal{A} \mid \|\tilde y - y\|_0 < \varepsilon\}$; a *weak* one is $\{\tilde y \in \mathcal{A} \mid \|\tilde y - y\|_1 < \varepsilon\}$. Both are metric notions in the sense of Definition 6.24, and since $\|\cdot\|_0 \le \|\cdot\|_1$ the weak ε-neighbourhood is contained in the strong one of the same radius, so every strong neighbourhood contains a weak one: the weak topology on the admissible class is the finer of the two. The containment is strict in the worst possible way — no strong neighbourhood of 0 is contained in any weak one, since $y_n(x) = n^{-1}\sin(n^2 x)$ has $\|y_n\|_0 \to 0$ while $\|y_n\|_1 \to \infty$.

Rests on Definitions 6.24 and 16.11.

Definition 16.13 (Weak and strong extrema)**.** $y \in \mathcal{A}$ is a *weak local minimum* of J if $J[y] \le J[\tilde y]$ for every $\tilde y$ in some weak neighbourhood of y, and a *strong local minimum* if the same holds in some strong neighbourhood. Every strong minimum is a weak minimum; the converse fails.

Rests on Definition 16.12.

Example 16.14 (A weak minimum that is not strong)**.** Take $a = 0$, $b = 1$, $y_a = y_b = 0$ and $f(x, y, y') = y'^2 + y'^3$, and let $y \equiv 0$, so that $J[y] = 0$. If $\|\tilde{y}\|_1 < \frac{1}{2}$ then $|\tilde{y}'| < \frac{1}{2}$ pointwise and the integrand is $\tilde{y}'^2\,(1 + \tilde{y}') \geq \frac{1}{2}\tilde{y}'^2 \geq 0$, so $J[\tilde{y}] \geq 0$: the zero function is a weak minimum. It is not a strong one. Fix $A > 1$ and $0 < \delta < 1$, and let $\tilde{y}$ be the piecewise-linear function of slope $-A$ on a set of total length δ and slope $A\delta/(1-\delta)$ on the rest, so that $\tilde{y}(0) = \tilde{y}(1) = 0$ and $\|\tilde{y}\|_0 \leq A\delta$. Then

$$J[\tilde{y}] = \delta\left(A^2 - A^3\right) + (1-\delta)\left[\left(\frac{A\delta}{1-\delta}\right)^2 + \left(\frac{A\delta}{1-\delta}\right)^3\right],$$

whose first term is negative and of order δA^3 while the second is of order $\delta^2 A^2$; taking $A = \delta^{-1/2}$ makes $J[\tilde{y}] < 0$ while $\|\tilde{y}\|_0 \leq \delta^{1/2} \to 0$. So arbitrarily small *strong* perturbations lower the functional. The distinction is exactly what separates the Legendre and Jacobi conditions of Sections 16.4.1 and 16.4.2 from the Weierstrass condition of Section 16.4.3 [1254].

Rests on Definition 16.13.

Definition 16.15 (Variation; the first variation)**.** A *variation* at $y \in \mathcal{A}$ is a function $\eta \in C^1[a,b]$ with $\eta(a) = \eta(b) = 0$; the set of such η is written $C_0^1[a,b]$, and $y + \varepsilon\eta \in \mathcal{A}$ for every $\varepsilon \in \mathbb{R}$. The *first variation* (the Gâteaux variation) of J at y in the direction η is

$$\delta J[y;\eta] = \lim_{\varepsilon \to 0} \frac{J[y+\varepsilon\eta] - J[y]}{\varepsilon} = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} J[y+\varepsilon\eta]\Big|_{\varepsilon=0}, \tag{16.14}$$

when the limit exists. It is the functional derivative of Equation (7.138) paired with η.

Rests on Definition 16.11 and Equation (7.138).

Proposition 16.16 (Stationarity is necessary)**.** *If $y \in \mathcal{A}$ is a weak local extremum of J and $\delta J[y;\eta]$ exists, then $\delta J[y;\eta] = 0$ for every $\eta \in C_0^1[a,b]$.*

Rests on Definitions 16.13 and 16.15 and Lemma 7.33.

Proof. *Derives Proposition 16.16.*

Fix $\eta \neq 0$ and set $\varphi(\varepsilon) = J[y + \varepsilon\eta]$. Because $\|(y+\varepsilon\eta) - y\|_1 = |\varepsilon|\,\|\eta\|_1$, every ε with $|\varepsilon| < \varepsilon_0/\|\eta\|_1$ places $y + \varepsilon\eta$ inside the weak ε_0-neighbourhood in which y extremises J. Hence φ has a local extremum at $\varepsilon = 0$ and is differentiable there, so $\varphi'(0) = \delta J[y;\eta] = 0$ by Fermat's lemma (Lemma 7.33). □

Proposition 16.17 (The first variation of an integral functional)**.** *For J as in Equation (16.11) with f of class C^2, and for every $\eta \in C^1[a,b]$,*

$$\delta J[y;\eta] = \int_a^b \left[\frac{\partial f}{\partial y}\,\eta(x) + \frac{\partial f}{\partial y'}\,\eta'(x)\right]\mathrm{d}x, \tag{16.15}$$

the partial derivatives being evaluated at $(x, y(x), y'(x))$.

Rests on Equations (16.11) and (16.14).

Proof. *Derives Proposition 16.17.*

The map $(\varepsilon, x) \longmapsto f\,(x, y + \varepsilon\eta, y' + \varepsilon\eta')$ is C^1 on the compact set $[-1,1] \times [a,b]$, so its ε-derivative is continuous there and differentiation under the integral sign is legitimate. Carrying it out and using the chain rule (Proposition 7.31),

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_0 f\,(x, y+\varepsilon\eta, y'+\varepsilon\eta') = \frac{\partial f}{\partial y}\,\eta + \frac{\partial f}{\partial y'}\,\eta',$$

which integrates to Equation (16.15). □

16.2.2 The fundamental lemma

Equation (16.15) turns the stationarity condition into the statement that a certain integral vanishes for *every* variation. Converting that into a pointwise differential equation is the work of the following three lemmas. The first is the fundamental lemma of the calculus of variations; the second is du Bois-Reymond's sharpening [204], which is what makes it unnecessary to assume in advance that the extremal is twice differentiable.

Lemma 16.18 (Fundamental lemma)**.** *Let $g \in C^0[a,b]$ satisfy*

$$\int_a^b g(x)\,\eta(x)\,\mathrm{d}x = 0 \qquad \text{for every } \eta \in C_0^1[a,b]\,. \tag{16.16}$$

Then $g \equiv 0$ on $[a,b]$.
Rests on Definition 16.15 and Theorem 7.40.

Proof. *Derives Lemma 16.18.*

Suppose $g(x_0) \neq 0$ for some x_0; replacing g by $-g$ if necessary, assume $g(x_0) > 0$, and by continuity x_0 may be taken interior. By continuity again (Definition 7.20) there is $\delta > 0$ with $[x_0-\delta, x_0+\delta] \subset (a,b)$ and $g > \tfrac{1}{2}g(x_0)$ on that interval. Put

$$\eta(x) = \begin{cases} \left[(x-x_0)^2 - \delta^2\right]^2, & |x-x_0| \le \delta\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{16.17}$$

This η is of class C^1 — both it and $\eta' = 4(x-x_0)\left[(x-x_0)^2-\delta^2\right]$ vanish at $x = x_0 \pm \delta$ — and lies in $C_0^1[a,b]$. It is strictly positive on the open interval, so

$$\int_a^b g\,\eta\,\mathrm{d}x \ge \frac{g(x_0)}{2}\int_{x_0-\delta}^{x_0+\delta}\eta\,\mathrm{d}x > 0\,,$$

contradicting Equation (16.16). □

Lemma 16.19 (du Bois-Reymond)**.** *Let $h \in C^0[a,b]$ satisfy*

$$\int_a^b h(x)\,\eta'(x)\,\mathrm{d}x = 0 \qquad \text{for every } \eta \in C_0^1[a,b]\,. \tag{16.18}$$

Then h is constant on $[a,b]$.
Rests on Equation (16.18) and Theorem 7.42.

Proof. *Derives Lemma 16.19.*

Let $c = (b-a)^{-1}\int_a^b h\,\mathrm{d}x$ and define

$$\eta(x) = \int_a^x \left[h(t) - c\right]\mathrm{d}t\,.$$

Then $\eta(a) = 0$, and $\eta(b) = \int_a^b h\,\mathrm{d}x - c(b-a) = 0$ by the choice of c, so $\eta \in C_0^1[a,b]$; it is of class C^1 with $\eta' = h - c$ by the fundamental theorem of calculus (Theorem 7.42). Substituting into Equation (16.18),

$$0 = \int_a^b h\,(h-c)\,\mathrm{d}x = \int_a^b (h-c)^2\,\mathrm{d}x + c\int_a^b (h-c)\,\mathrm{d}x = \int_a^b (h-c)^2\,\mathrm{d}x\,,$$

the last integral vanishing by the definition of c. A continuous nonnegative integrand with zero integral vanishes identically, so $h \equiv c$. □

Lemma 16.20 (Mixed form)**.** *Let $M, N \in C^0[a,b]$ satisfy*

$$\int_a^b \left[M(x)\,\eta(x) + N(x)\,\eta'(x) \right] \mathrm{d}x = 0 \qquad \text{for every } \eta \in C_0^1[a,b]\,. \tag{16.19}$$

Then N is of class C^1 and $N' = M$ on $[a,b]$.
Rests on Lemma 16.19 and Corollary 7.44.

Proof. *Derives Lemma 16.20.*
Set $A(x) = \int_a^x M(t)\,\mathrm{d}t$, which is C^1 with $A' = M$ (Theorem 7.42). Integrating by parts (Equation (7.28)) and using $\eta(a) = \eta(b) = 0$,

$$\int_a^b M\,\eta\,\mathrm{d}x = \left[A\,\eta \right]_a^b - \int_a^b A\,\eta'\,\mathrm{d}x = -\int_a^b A\,\eta'\,\mathrm{d}x\,.$$

So Equation (16.19) says $\int_a^b (N - A)\,\eta'\,\mathrm{d}x = 0$ for every admissible η, whence $N - A$ is a constant by Lemma 16.19. Thus $N = A + \text{const}$ is of class C^1 with $N' = A' = M$. □

Remark 16.21 (Broken extremals). Nothing above requires the competing curves to be smooth: the same three lemmas hold verbatim when C_0^1 is enlarged to the piecewise-C^1 functions vanishing at the endpoints, since the bump Equation (16.17) and the primitive built in Lemma 16.19 already lie in the smaller class. An extremal in the wider class may have corners, and the two conditions that must then hold across a corner are the Weierstrass–Erdmann conditions proved as Corollary 16.33 [1254].

16.2.3 The Euler–Lagrange equation

Theorem 16.22 (Euler–Lagrange)**.** *Let f be of class C^2 and let $y \in \mathcal{A}$ be a weak local extremum of Equation (16.11). Then $x \longmapsto \partial f/\partial y'\,(x, y, y')$ is of class C^1 and*

$$\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'} \right) = 0 \qquad \text{on } [a,b]\,. \tag{16.20}$$

A solution of Equation (16.20) is called an extremal *of J.*
Rests on Proposition 16.16, Equation (16.15), and Lemma 16.20.

Proof. *Derives Theorem 16.22.*
By Proposition 16.16 and Proposition 16.17,

$$\int_a^b \left[\frac{\partial f}{\partial y}\,\eta + \frac{\partial f}{\partial y'}\,\eta' \right] \mathrm{d}x = 0 \qquad \text{for every } \eta \in C_0^1[a,b]\,.$$

Both coefficients are continuous in x, because f is C^2 and $y \in C^1$. Lemma 16.20 with $M = \partial f/\partial y$ and $N = \partial f/\partial y'$ then says that N is C^1 with $N' = M$, which is Equation (16.20). □

Two forms of the same statement are worth separating, because they carry different smoothness assumptions.

Corollary 16.23 (du Bois-Reymond form)**.** *Under the hypotheses of Theorem 16.22 there is a constant C with*

$$\frac{\partial f}{\partial y'}\,(x, y, y') = \int_a^x \frac{\partial f}{\partial y}\,(t, y, y')\,\mathrm{d}t + C \qquad \text{for all } x \in [a,b]\,. \tag{16.21}$$

Rests on Equation (16.20) and Theorem 7.43.

Proof. *Derives Corollary 16.23.*
Integrate Equation (16.20) from a to x, using Theorem 7.43 on the C^1 function $\partial f/\partial y'$. □

Remark 16.24 (When the extremal is twice differentiable). Expanding the total derivative in Equation (16.20) gives the second-order form

$$\frac{\partial f}{\partial y} - \frac{\partial^2 f}{\partial x\,\partial y'} - \frac{\partial^2 f}{\partial y\,\partial y'}\,y' - \frac{\partial^2 f}{\partial y'^2}\,y'' = 0\,, \tag{16.22}$$

which presupposes y''. It is legitimate wherever $\partial^2 f/\partial y'^2 \neq 0$: at such a point Equation (16.21) may be solved for y' as a C^1 function of x by the implicit function theorem (Chapter 7), so $y \in C^2$ there. Where $\partial^2 f/\partial y'^2$ vanishes the extremal may genuinely have a corner, and only the first-order form Equation (16.21) survives. This is why the du Bois-Reymond argument matters: the classical derivation, which integrates by parts a second time, assumes the conclusion.

Remark 16.25 (Stationary is not minimal). Theorem 16.22 is a necessary condition only. An extremal may be a minimum, a maximum, or neither: the extremals of $\int_0^b (y'^2 - y^2)\,\mathrm{d}x$ with $y(0) = y(b) = 0$ include $y \equiv 0$, which minimises for $b < \pi$ and does not for $b > \pi$, although the equation $y'' + y = 0$ is the same in both cases. The extra information is second order and is the subject of Section 16.4; $b = \pi$ is the first conjugate point in the sense of Section 16.4.2. In physics this is the reason Hamilton's principle is a principle of *stationary*, not least, action (Section 16.7.1).

Proposition 16.26 (Covariance under change of dependent variable)**.** *Let $y = \Psi(x, u)$ with Ψ of class C^2 and $\partial\Psi/\partial u \neq 0$, and let*

$$F\,(x, u, u') = f\left(x, \Psi(x, u), \frac{\partial \Psi}{\partial x} + \frac{\partial \Psi}{\partial u}\,u'\right).$$

Then the Euler–Lagrange expressions are related by

$$\frac{\partial F}{\partial u} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial F}{\partial u'}\right) = \frac{\partial \Psi}{\partial u}\left[\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\right], \tag{16.23}$$

so u is an extremal of $\int F\,\mathrm{d}x$ if and only if $y = \Psi(x, u)$ is an extremal of $\int f\,\mathrm{d}x$.
Rests on Equation (16.20) and Proposition 7.31.

Proof. *Derives Proposition 16.26.*

Write Ψ_x, Ψ_u, Ψ_{xu}, Ψ_{uu} for the partials of Ψ and note that $y' = \Psi_x + \Psi_u u'$, whence $\partial y'/\partial u' = \Psi_u$ and $\partial y'/\partial u = \Psi_{xu} + \Psi_{uu}u' = \mathrm{d}\Psi_u/\mathrm{d}x$, the total derivative along the curve. The chain rule then gives

$$\frac{\partial F}{\partial u'} = \frac{\partial f}{\partial y'}\,\Psi_u\,, \qquad \frac{\partial F}{\partial u} = \frac{\partial f}{\partial y}\,\Psi_u + \frac{\partial f}{\partial y'}\,\frac{\mathrm{d}\Psi_u}{\mathrm{d}x}\,,$$

and therefore

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial F}{\partial u'}\right) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\Psi_u + \frac{\partial f}{\partial y'}\,\frac{\mathrm{d}\Psi_u}{\mathrm{d}x}\,.$$

Subtracting, the two $\partial f/\partial y' \cdot \mathrm{d}\Psi_u/\mathrm{d}x$ terms cancel and Equation (16.23) remains. Since $\Psi_u \neq 0$, one side vanishes exactly when the other does. □

Remark 16.27 (Why covariance matters here). Equation (16.23) says that the Euler–Lagrange expression transforms as a covector under changes of the dependent variables: the *equation* is coordinate independent even though its individual terms are not. This is what licenses the use of arbitrary generalized coordinates in Chapter 21 — polar, spherical, body-fixed — without rederiving the equations of motion each time, and what makes the same formalism usable on the curved configuration spaces of Chapter 13 and on the spacetimes of Chapter 43. Newton's equations enjoy no such property: the components of $m\ddot{\boldsymbol{x}}$ are not the components of anything in curvilinear coordinates.

16.2.4 First integrals and natural boundary conditions

The Euler–Lagrange equation is of second order, and integrating it once is exactly what makes the classical problems soluble. Two circumstances permit it, and both are instances of Noether's theorem (Section 16.6.1): a coordinate absent from the integrand, and the independent variable absent from the integrand.

Proposition 16.28 (Cyclic coordinate)**.** *If f does not depend on y, that is $\partial f/\partial y \equiv 0$, then along every extremal*

$$\frac{\partial f}{\partial y'}\,(x, y') = \mathit{const}\,. \tag{16.24}$$

Rests on Equation (16.20)*.*

Proof. *Derives Proposition 16.28.*

Immediate from Equation (16.20): the first term vanishes, so the total derivative of $\partial f/\partial y'$ is zero, and a C^1 function with vanishing derivative on an interval is constant (Corollary 7.36). □

Theorem 16.29 (Beltrami identity)**.** *If f does not depend explicitly on x, that is $\partial f/\partial x \equiv 0$, then along every extremal of class C^2*

$$f\,(y, y') - y'\,\frac{\partial f}{\partial y'}\,(y, y') = \mathit{const}\,. \tag{16.25}$$

Rests on Equation (16.20) *and Proposition 7.31.*

Proof. *Derives Theorem 16.29.*

Differentiate the left-hand side along the curve, using the chain rule and $\partial f/\partial x = 0$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left(f - y'\,\frac{\partial f}{\partial y'}\right) &= \frac{\partial f}{\partial y}\,y' + \frac{\partial f}{\partial y'}\,y'' - y''\,\frac{\partial f}{\partial y'} - y'\,\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\\ &= y'\left[\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\right] = 0\,,\end{aligned}$$

the bracket vanishing by Equation (16.20). □

Remark 16.30 (The Beltrami constant is an energy). The quantity $h = y'\,\partial f/\partial y' - f$ conserved by Equation (16.25) is the Legendre transform of f with respect to y'. When x is the time and f the Lagrangian of a mechanical system, h is the energy in joules and Equation (16.25) is its conservation (Chapter 21); when f is convex in y' the same transform produces the Hamiltonian of Chapter 22, and the convexity hypothesis it needs is the strengthened Legendre condition of Section 16.4.1.

So far the endpoints have been held fixed. Releasing one of them costs nothing in the derivation and produces a boundary condition instead of a boundary datum.

Theorem 16.31 (Natural boundary condition)**.** *Let J be extremised over $\mathcal{A}_b = \{y \in C^1[a,b] \mid y(a) = y_a\}$, the value $y(b)$ being free. Then a weak extremum satisfies Equation* (16.20) *on $[a,b]$ together with*

$$\frac{\partial f}{\partial y'}\,(b, y(b), y'(b)) = 0\,. \tag{16.26}$$

Rests on Equation (16.15)*, Theorem 16.22, and Lemma 16.18.*

Proof. *Derives Theorem 16.31.*

Variations η now need only satisfy $\eta(a) = 0$. Those that also vanish at b form the class of Theorem 16.22, so Equation (16.20) holds and $\partial f/\partial y'$ is C^1. For a general admissible η, integrating Equation (16.15) by parts now keeps the boundary term:

$$0 = \delta J[y;\eta] = \int_a^b \left[\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\right]\eta\,\mathrm{d}x + \left[\frac{\partial f}{\partial y'}\,\eta\right]_a^b = \left.\frac{\partial f}{\partial y'}\right|_{x=b}\eta(b)\,, \tag{16.27}$$

the integral vanishing by Equation (16.20) and the term at a by $\eta(a) = 0$. Choosing η with $\eta(b) = 1$ gives Equation (16.26). □

Theorem 16.32 (Transversality)**.** *Let the right-hand endpoint be required only to lie on a C^1 curve $y = \psi(x)$, the abscissa b being free as well. Then at the point $(b, y(b))$ where the extremal meets the curve,*

$$\left[f + (\psi' - y')\,\frac{\partial f}{\partial y'}\right]_{x=b} = 0\,. \tag{16.28}$$

Rests on Theorem 16.31 and Equation (16.20).

Proof. *Derives Theorem 16.32.*

Embed the extremal in a one-parameter family of admissible curves y_ε with terminal abscissa $b_\varepsilon = b + \varepsilon\beta + O(\varepsilon^2)$ and terminal ordinate $y_\varepsilon(b_\varepsilon) = \psi(b_\varepsilon)$, and write η for the vertical variation $\partial y_\varepsilon/\partial\varepsilon|_0$ at fixed x, so $\eta(a) = 0$. Differentiating $J_\varepsilon = \int_a^{b_\varepsilon} f\,\mathrm{d}x$ at $\varepsilon = 0$, and using $\mathrm{d}\left(\int_a^{b_\varepsilon}\right)/\mathrm{d}\varepsilon = \beta f(b)$ for the moving limit,

$$0 = \beta f\big|_b + \int_a^b \left[\frac{\partial f}{\partial y}\eta + \frac{\partial f}{\partial y'}\eta'\right]\mathrm{d}x = \beta f\big|_b + \left.\frac{\partial f}{\partial y'}\right|_b \eta(b)\,,$$

the integral collapsing as in Equation (16.27). The endpoint constraint relates β and $\eta(b)$: differentiating $y_\varepsilon(b_\varepsilon) = \psi(b_\varepsilon)$ gives $\eta(b) + \beta\,y'(b) = \beta\,\psi'(b)$, that is $\eta(b) = \beta\,(\psi' - y')_b$. Substituting and dividing by the arbitrary $\beta \neq 0$ yields Equation (16.28). □

Corollary 16.33 (Weierstrass–Erdmann corner conditions)**.** *Let y be a piecewise-C^1 extremal with a corner at an interior point x_0, and write $y'(x_0^-)$ and $y'(x_0^+)$ for the one-sided slopes. Then both*

$$\left.\frac{\partial f}{\partial y'}\right|_{x_0^-} = \left.\frac{\partial f}{\partial y'}\right|_{x_0^+}, \qquad \left[f - y'\,\frac{\partial f}{\partial y'}\right]_{x_0^-} = \left[f - y'\,\frac{\partial f}{\partial y'}\right]_{x_0^+} \tag{16.29}$$

hold: the momentum $\partial f/\partial y'$ and the energy $y'\,\partial f/\partial y' - f$ are continuous across a corner even though the slope is not.

Rests on Theorem 16.32 and Corollary 16.23.

Proof. *Derives Corollary 16.33.*

The first condition follows from Equation (16.21), which holds on each smooth arc with the *same* constant C — the derivation of Lemma 16.20 never used differentiability of y — so $\partial f/\partial y'$ equals a continuous function of x throughout $[a, b]$.

For the second, regard the corner (x_0, y_0) as a common free endpoint of the two arcs, and vary it. Repeating the computation of Theorem 16.32 for the left arc, whose terminal point moves by (β, γ) with $\gamma = \partial y_\varepsilon(x_0)/\partial\varepsilon|_0 + \beta y'(x_0^-)$, contributes $\beta f|_{x_0^-} + (\partial f/\partial y')_{x_0^-}\left(\gamma - \beta y'(x_0^-)\right)$ to δJ; the right arc, for which the same point is an *initial* point, contributes the same expression at x_0^+ with the opposite sign. Stationarity for all (β, γ) forces both brackets to match: the coefficient of γ gives the first condition again, and the coefficient of β gives $[f - y'\,\partial f/\partial y']$ continuous, the second. □

Remark 16.34 (Where the boundary terms go in physics). The boundary term $[(\partial f/\partial y')\,\eta]_a^b$ discarded in Theorem 16.22 is not always discardable. In field theory it becomes a surface integral over $\partial\Omega$ whose vanishing is a boundary condition on the fields (Chapter 64); in general relativity it does not vanish for the Einstein–Hilbert action, and a compensating surface term must be added by hand (Chapter 44). The elementary case here — a free end of a string forces $\partial f/\partial y' = 0$, which for a taut string is the statement that no transverse force is applied there — is the same phenomenon in one dimension.

16.2.5 Several functions, several variables, higher derivatives

Theorem 16.35 (System of Euler–Lagrange equations). *Let $J[y^1,\dots,y^n] = \int_a^b f\,(x, y^1,\dots,y^n, y'^1,\dots,y'^n)\,\mathrm{d}x$ with f of class C^2, over n-tuples of C^1 functions with all $2n$ endpoint values fixed. At a weak extremum,*

$$\frac{\partial f}{\partial y^i} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'^i}\right) = 0\,, \qquad i = 1,\dots,n\,. \tag{16.30}$$

Rests on Theorem 16.22 and Lemma 16.20.

Proof. *Derives Theorem 16.35.*

Fix i and vary only y^i, holding the other $n-1$ functions fixed: J restricted to that one-parameter family is a functional of the type of Equation (16.11), and Theorem 16.22 gives the i-th equation. Repeating for each i gives the system, and no further information is available from simultaneous variations, since the first variation is linear in $(\eta^1,\dots,\eta^n)$. □

Theorem 16.36 (Euler–Lagrange equations for several independent variables). *Let $\Omega \subset \mathbb{R}^D$ be a bounded domain with piecewise-C^1 boundary, and let*

$$S[\varphi] = \int_\Omega \mathcal{L}\left(x^\mu, \varphi^A, \partial_\mu \varphi^A\right) \mathrm{d}^D x\,, \qquad A = 1,\dots,n\,,\ \mu = 1,\dots,D\,, \tag{16.31}$$

be extremised over fields $\varphi^A \in C^2(\overline{\Omega})$ with prescribed values on $\partial\Omega$. Then at a weak extremum

$$\frac{\partial \mathcal{L}}{\partial \varphi^A} - \partial_\mu\left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi^A)}\right) = 0\,, \qquad A = 1,\dots,n\,, \tag{16.32}$$

with the repeated index μ summed.

Rests on Equation (16.31), Theorem 7.101, and Lemma 16.18.

Proof. *Derives Theorem 16.36.*

Let $\eta^A \in C^1(\overline{\Omega})$ vanish on $\partial\Omega$ and put $\varphi_\varepsilon = \varphi + \varepsilon\eta$. Differentiating under the integral sign as in Proposition 16.17,

$$\delta S[\varphi;\eta] = \int_\Omega \left[\frac{\partial \mathcal{L}}{\partial \varphi^A}\,\eta^A + \frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi^A)}\,\partial_\mu \eta^A\right] \mathrm{d}^D x\,.$$

Write $\pi^\mu_A = \partial\mathcal{L}/\partial(\partial_\mu\varphi^A)$, which is C^1 because $\mathcal{L}$ is C^2 and $\varphi \in C^2$. Then $\pi^\mu_A\,\partial_\mu\eta^A = \partial_\mu\left(\pi^\mu_A \eta^A\right) - \eta^A\,\partial_\mu \pi^\mu_A$, and the divergence theorem (Theorem 7.101) turns the first piece into a surface integral over $\partial\Omega$, where $\eta^A = 0$. Hence

$$\delta S[\varphi;\eta] = \int_\Omega \left[\frac{\partial \mathcal{L}}{\partial \varphi^A} - \partial_\mu \pi^\mu_A\right] \eta^A\,\mathrm{d}^D x\,. \tag{16.33}$$

Stationarity makes this vanish for every such η. The bracket is continuous, and the multidimensional form of Lemma 16.18 — proved by the same argument, with the bump Equation (16.17) replaced by its product over the D coordinates on a small cube inside Ω — forces it to vanish identically, one component A at a time. □

Example 16.37 (The Dirichlet integral gives Laplace's equation)**.** Take $D = 3$, one field u, and $\mathcal{L} = \tfrac{1}{2}|\nabla u|^2$, so that $S[u] = \tfrac{1}{2}\int_\Omega |\nabla u|^2 \,\mathrm{d}^3x$ is the Dirichlet integral. Then $\partial\mathcal{L}/\partial u = 0$ and $\partial\mathcal{L}/\partial(\partial_\mu u) = \partial_\mu u$, so Equation (16.32) reads

$$\nabla^2 u = 0\,, \tag{16.34}$$

Laplace's equation of Chapter 10. Multiplied by the permittivity of free space, $\varepsilon_0 = 8.854\,187\,812\,8 \times 10^{-12}\,\mathrm{F/m}$, the same functional is the electrostatic field energy $U = \frac{\varepsilon_0}{2}\int |\nabla\Phi|^2 \,\mathrm{d}^3x$ in joules, and the statement becomes the one of Chapter 58: the electrostatic potential is the field of least energy compatible with the boundary data. This is the Dirichlet principle, whose existence gap is the subject of Section 16.5.1. Multiple-integral problems of this kind are the systematic subject of Courant–Hilbert [1222].

Rests on Equation (16.32).

Theorem 16.38 (Euler–Poisson equation)**.** *Let $J[y] = \int_a^b f\left(x, y, y', \dots, y^{(m)}\right)\mathrm{d}x$ with f of class C^{m+1}, be extremised over $y \in C^{2m}[a,b]$ with $y, y', \dots, y^{(m-1)}$ prescribed at both endpoints. Then*

$$\sum_{k=0}^{m} (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left(\frac{\partial f}{\partial y^{(k)}}\right) = 0\,. \tag{16.35}$$

Rests on Equation (16.15), Lemma 16.18, and Corollary 7.44.

Proof. *Derives Theorem 16.38.*

Admissible variations η satisfy $\eta^{(j)}(a) = \eta^{(j)}(b) = 0$ for $j = 0, \dots, m-1$. As in Proposition 16.17,

$$\delta J[y;\eta] = \int_a^b \sum_{k=0}^{m} \frac{\partial f}{\partial y^{(k)}}\,\eta^{(k)}\,\mathrm{d}x\,.$$

Integrate the k-th term by parts k times (Equation (7.28)). Each integration produces a boundary term carrying a factor $\eta^{(j)}$ with $j \le k-1 \le m-1$, all of which vanish, and contributes one factor of -1; after k steps the term is $(-1)^k \int \eta\,\mathrm{d}^k\left(\partial f/\partial y^{(k)}\right)/\mathrm{d}x^k$. Summing over k,

$$\delta J[y;\eta] = \int_a^b \eta \sum_{k=0}^{m} (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left(\frac{\partial f}{\partial y^{(k)}}\right)\mathrm{d}x = 0\,,$$

and Lemma 16.18 gives Equation (16.35). □

Proposition 16.39 (Ostrogradsky instability)**.** *Let $m = 2$, write q for the dependent variable and t for the independent one, and suppose the Lagrangian $\mathcal{L}(q, \dot q, \ddot q)$ is* nondegenerate*, meaning $\partial^2\mathcal{L}/\partial\ddot q^2 \neq 0$, so that $\ddot q$ may be solved for in terms of $(q, \dot q, \partial\mathcal{L}/\partial\ddot q)$. Then the conserved energy associated with time translation is linear in one of its momenta and is therefore unbounded below.*

Rests on Equation (16.35) and Theorem 16.29.

Proof. *Derives Proposition 16.39.*

Introduce $Q_1 = q$, $Q_2 = \dot q$ and the Ostrogradsky momenta

$$P_1 = \frac{\partial\mathcal{L}}{\partial\dot q} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial\mathcal{L}}{\partial\ddot q}\right), \qquad P_2 = \frac{\partial\mathcal{L}}{\partial\ddot q}\,.$$

Nondegeneracy lets $P_2 = \partial\mathcal{L}/\partial\ddot q$ be inverted for $\ddot q = \mathcal{F}(Q_1, Q_2, P_2)$. The Beltrami constant of Theorem 16.29, extended to the higher-derivative case by the same computation, is

$$\mathcal{H} = P_1 Q_2 + P_2\,\mathcal{F}(Q_1, Q_2, P_2) - \mathcal{L}(Q_1, Q_2, \mathcal{F})\,. \tag{16.36}$$

Only the first term contains P_1, and it contains it linearly, so for fixed (Q_1, Q_2, P_2) with $Q_2 \neq 0$ the value of $\mathcal{H}$ runs over all of $\mathbb{R}$ as P_1 does. □

Remark 16.40 (Why every fundamental Lagrangian in this book is first order). Proposition 16.39 is Ostrogradsky's theorem [132]. Its physical content is that a nondegenerate theory whose Lagrangian contains second time derivatives has no stable ground state: the system can lower $\mathcal{H}$ without bound by exciting the P_1 direction, and once it is coupled to anything else the vacuum decays. This — not aesthetics — is why the Lagrangians of Chapter 21, Chapter 64 and Chapter 44 are first order in time derivatives. The Einstein–Hilbert action is the instructive exception: it does contain second derivatives of the metric, but *degenerately* — they enter linearly and can be removed by a total divergence, which is exactly the boundary term of Remark 16.34 — so Ostrogradsky's hypothesis fails and general relativity has second-order field equations.

16.3 Constrained variation

16.3.1 Lagrange multipliers

Constraints come in two kinds, and they need different multipliers. A *pointwise* constraint must hold at every x, and buys one multiplier *function*; an *integral* constraint holds once for the whole curve, and buys one multiplier *constant*. The rule for both descends from the finite-dimensional multiplier theorem of Chapter 7, of which the following is the functional analogue [40].

Theorem 16.41 (Multiplier rule for pointwise constraints)**.** *Let $J[y]=\int_a^b f\,(x,y^i,y'^i)\,\mathrm{d}x$ with $i=1,\dots,n$, and let the competing curves be required to satisfy the $m<n$ pointwise constraints*

$$g^\alpha\left(x,y^1,\dots,y^n\right)=0\,,\qquad \alpha=1,\dots,m\,, \tag{16.37}$$

with g^α of class C^2 and the $m\times n$ matrix $\partial g^\alpha/\partial y^i$ of full rank m along the curve. If y is a weak extremum of J within the constrained class, there are functions $\lambda_\alpha\in C^0[a,b]$ such that, with $F=f+\lambda_\alpha g^\alpha$,

$$\frac{\partial F}{\partial y^i}-\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial F}{\partial y'^i}\right)=0\,,\qquad i=1,\dots,n\,. \tag{16.38}$$

Rests on Equation (16.37), Theorem 16.35, and Lemma 16.18.

Proof. *Derives Theorem 16.41.*

Admissible variations $\eta^i\in C^1_0[a,b]$ are no longer arbitrary: differentiating Equation (16.37) along the family $y+\varepsilon\eta$ at $\varepsilon=0$ gives the m pointwise linear conditions

$$\frac{\partial g^\alpha}{\partial y^i}\,\eta^i(x)=0\,,\qquad \alpha=1,\dots,m\,,\quad x\in[a,b]\,. \tag{16.39}$$

Stationarity gives, after the integration by parts of Theorem 16.22,

$$\int_a^b E_i\,\eta^i\,\mathrm{d}x=0\,,\qquad E_i\equiv\frac{\partial f}{\partial y^i}-\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'^i}\right)\,, \tag{16.40}$$

for every η obeying Equation (16.39). By the rank hypothesis the indices may be relabelled so that the leading $m\times m$ block $\partial g^\alpha/\partial y^\beta$, $\beta=1,\dots,m$, is invertible at every x; call $y^1,\dots,y^m$ dependent and $y^{m+1},\dots,y^n$ free. Define $\lambda_\alpha(x)$ as the unique solution of the linear system

$$E_\beta+\lambda_\alpha\,\frac{\partial g^\alpha}{\partial y^\beta}=0\,,\qquad \beta=1,\dots,m\,,$$

which is continuous in x because the block and the E_β are. With this choice the first m of Equation (16.38) hold by construction, since g^α does not contain y'^i and hence

$\partial F/\partial y'^i = \partial f/\partial y'^i$. Substituting back into Equation (16.40) and using Equation (16.39) to eliminate the dependent η^β,

$$0 = \int_a^b \left(E_k + \lambda_\alpha \frac{\partial g^\alpha}{\partial y^k} \right) \eta^k \, \mathrm{d}x \,, \qquad k = m+1, \dots, n \,,$$

now for *arbitrary* $\eta^k \in C_0^1[a,b]$, since the dependent components can always be recovered from Equation (16.39). Lemma 16.18, applied one free index at a time, gives the remaining $n-m$ equations. □

Remark 16.42 (The multiplier is the constraint force). Comparing Equation (16.38) with the unconstrained equations, the effect of the constraint is the extra term $\lambda_\alpha \, \partial g^\alpha / \partial y^i$ on the right-hand side — a force normal to the constraint surface, of magnitude fixed by λ_α. In mechanics this is exactly the reaction of the constraint, and λ_α carries whatever SI unit makes $\lambda_\alpha g^\alpha$ an energy: for a constraint written as a length, λ is a force in newtons. That reading is developed, with the Lagrange equations of the first kind, in Chapter 21. It is also the practical reason to keep the multipliers rather than eliminate the constraint: the numbers one wants — the tension in the rod, the normal force on the track — are the multipliers themselves.

16.3.2 Isoperimetric constraints

Theorem 16.43 (Euler's rule for integral constraints). *Let $J[y] = \int_a^b f\,(x, y, y')\,\mathrm{d}x$ and $K[y] = \int_a^b g\,(x, y, y')\,\mathrm{d}x$ with f, g of class C^2, and let $y \in \mathcal{A}$ be a weak extremum of J within the class $\{\tilde{y} \in \mathcal{A} \mid K[\tilde{y}] = \ell\}$. If y is* not *an extremal of K alone, there is a constant $\lambda \in \mathbb{R}$ such that y is an unconstrained extremal of $\int_a^b (f + \lambda g)\,\mathrm{d}x$, that is*

$$\frac{\partial}{\partial y}(f + \lambda g) - \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{\partial}{\partial y'}(f + \lambda g) \right] = 0 \,. \tag{16.41}$$

Rests on Definition 16.15, Proposition 16.17, and Lemma 16.20.

Proof. *Derives Theorem 16.43.*

Because y is not an extremal of K, Lemma 16.20 supplies some $\eta_2 \in C_0^1[a,b]$ with $\delta K[y;\eta_2] \neq 0$. Fix an arbitrary $\eta_1 \in C_0^1[a,b]$ and consider the two-parameter family $y_\varepsilon = y + \varepsilon_1 \eta_1 + \varepsilon_2 \eta_2$, setting

$$\Phi\,(\varepsilon_1, \varepsilon_2) = J[y_\varepsilon] \,, \qquad \Psi\,(\varepsilon_1, \varepsilon_2) = K[y_\varepsilon] \,.$$

Both are C^1 near the origin, with $\partial\Phi/\partial\varepsilon_i\big|_0 = \delta J[y;\eta_i]$ and similarly for Ψ, by the argument of Proposition 16.17. Since $\Psi(0,0) = \ell$ and $\partial\Psi/\partial\varepsilon_2\big|_0 = \delta K[y;\eta_2] \neq 0$, the implicit function theorem (Chapter 7) provides a C^1 function $\varepsilon_2\,(\varepsilon_1)$ with $\varepsilon_2(0) = 0$ and $\Psi\,(\varepsilon_1, \varepsilon_2(\varepsilon_1)) = \ell$ for small ε_1. Every curve of this one-parameter family is admissible and lies in a weak neighbourhood of y, so $\varepsilon_1 \longmapsto \Phi\,(\varepsilon_1, \varepsilon_2(\varepsilon_1))$ has an extremum at 0:

$$\delta J[y;\eta_1] + \delta J[y;\eta_2]\,\varepsilon_2'(0) = 0 \,, \qquad \varepsilon_2'(0) = -\frac{\delta K[y;\eta_1]}{\delta K[y;\eta_2]} \,,$$

the second identity coming from differentiating the constraint. Setting $\lambda = -\delta J[y;\eta_2]/\delta K[y;\eta_2]$, a constant independent of η_1, the display becomes $\delta J[y;\eta_1] + \lambda\,\delta K[y;\eta_1] = 0$, that is $\delta\,(J + \lambda K)\,[y;\eta_1] = 0$ for every $\eta_1 \in C_0^1[a,b]$. Lemma 16.20 then yields Equation (16.41). □

Remark 16.44 (The exceptional case is not a technicality). If y *is* an extremal of K alone, the conclusion can fail, and the correct statement is the abnormal one: there are constants $(\lambda_0, \lambda) \neq (0,0)$ with y an extremal of $\lambda_0 f + \lambda g$, and λ_0 may be zero. The finite-dimensional

shadow of this is the failure of the multiplier rule at a critical point of the constraint function, and it is the same phenomenon: the constraint set has no tangent direction along which J can be tested.

Example 16.45 (Two solved isoperimetric problems)**.** Proposition 16.6 and Proposition 16.4 are both applications of Theorem 16.43, with λ playing two different physical roles: for Dido's problem $|\lambda|$ is the radius of the answer, and for the hanging chain $\rho_\ell g\lambda$ is the horizontal tension in the chain, in newtons. Both illustrate Remark 16.42: the multiplier is not bookkeeping, it is the physically interesting number.

Rests on Theorem 16.43.

Remark 16.46 (Eigenvalue problems are isoperimetric problems)*.* Take $f = p\,y'^2 + q\,y^2$ and $g = r\,y^2$ with the normalisation $K[y] = \int_a^b r\,y^2\,\mathrm{d}x = 1$. Then Equation (16.41) reads

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(p\,\frac{\mathrm{d}y}{\mathrm{d}x}\right) + q\,y = -\lambda\,r\,y\,, \tag{16.42}$$

which is the Sturm–Liouville eigenvalue problem Equation (9.55) of Chapter 9 with eigenvalue $-\lambda$. The multiplier of the normalisation constraint *is* the eigenvalue. This is the identification the Rayleigh–Ritz method of Section 16.5.3 exploits, and it is why a variational estimate of a ground-state energy is available at all.

16.3.3 Holonomic and nonholonomic constraints

Definition 16.47 (Holonomic and nonholonomic constraints)**.** A constraint on a curve $x \longmapsto (y^1, \dots, y^n)$ is *holonomic* if it can be written as a finite set of relations $g^\alpha\,(x, y^1, \dots, y^n) = 0$ among the coordinates and the independent variable alone, as in Equation (16.37). A constraint given in Pfaffian form,

$$A^\alpha{}_i\,(x,y)\,\mathrm{d}y^i + A^\alpha{}_0\,(x,y)\,\mathrm{d}x = 0\,, \qquad \alpha = 1, \dots, m\,, \tag{16.43}$$

is holonomic if the system Equation (16.43) is integrable — if there are functions g^α and an invertible matrix $\mu^\alpha{}_\beta$ with $\mu^\alpha{}_\beta\left(A^\beta{}_i\mathrm{d}y^i + A^\beta{}_0\mathrm{d}x\right) = \mathrm{d}g^\alpha$ — and *nonholonomic* otherwise. Integrability is decided by the Frobenius condition of Chapter 13: the ideal generated by the one-forms must be closed under exterior differentiation.

Rests on Equation (16.37) and Definition 21.2.

Example 16.48 (The vertical rolling disc)**.** A thin disc of radius R rolls without slipping on a horizontal plane, staying vertical. Its configuration is $(x, y, \varphi, \vartheta)$: contact point, heading angle and rotation angle. Rolling without slipping means the contact point has zero velocity, which in SI reads

$$\mathrm{d}x - R\cos\varphi\,\mathrm{d}\vartheta = 0\,, \qquad \mathrm{d}y - R\sin\varphi\,\mathrm{d}\vartheta = 0\,, \tag{16.44}$$

with R in metres and the angles in radians. Neither one-form is integrable, alone or together: the disc can be rolled around a closed loop in the plane and returned to its starting point with ϑ changed by an arbitrary amount, so no relation $g\,(x, y, \varphi, \vartheta) = 0$ can hold. The constraint removes two velocity directions but no configuration, and the accessible set is all four dimensions. This is the standard nonholonomic example [2027]; the dynamics belongs to Chapter 21 and Chapter 29.

Rests on Definition 16.47.

Remark 16.49 (What the multiplier method does and does not cover)*.* Theorem 16.41 was proved for *holonomic* constraints, and the restriction is essential. Two inequivalent prescriptions exist for a Pfaffian constraint that is not integrable, and it is a matter of fact, not of convention, which describes a rolling body.

The *vakonomic* prescription applies the theorem literally: it restricts the class of comparison curves to those satisfying Equation (16.43) and makes the action stationary within that class. The *d'Alembert–Chetaev* prescription, which is the one mechanics uses, does not restrict the comparison curves at all; it restricts the *virtual displacements* to the annihilator of the constraint forms, $A^{\alpha}{}_{i}\,\delta y^{i} = 0$, and requires the constraint forces to do no virtual work. For a holonomic constraint the two prescriptions coincide, because a curve tangent to an integrable distribution stays on the leaf. For a nonholonomic one they do not: the vakonomic equations carry extra terms in the derivatives of the multipliers along the curve, and the two sets of trajectories differ already for the rolling disc of Example 16.48.

The observed motion of rolling bodies is the d'Alembert–Chetaev one; that is the content of the principle of virtual work as it is set up in Chapter 21, and it is what the experiments on rolling and spinning rigid bodies in Chapter 29 test. The vakonomic equations are not wrong — they are the correct equations of a different problem, namely the optimal-control problem in which the constraint is imposed on the admissible curves by fiat — but they are not the equations of a rolling disc. Degenerate constraint systems, where the constraints follow from the Lagrangian itself rather than being imposed, are a third matter, treated in Chapter 26.

16.4 The second variation

16.4.1 The Legendre condition

Stationarity is a first-order condition and says nothing about whether an extremal minimises, maximises or does neither (Remark 16.25). The second-order information is carried by the second variation, which is a quadratic functional of the variation.

Proposition 16.50 (The second variation)**.** *Let f be of class C^{3} and y an extremal of class C^{2}. For $\eta \in C^{1}_{0}[a,b]$,*

$$\delta^{2} J[y;\eta] \equiv \frac{\mathrm{d}^{2}}{\mathrm{d}\varepsilon^{2}} J[y+\varepsilon\eta]\bigg|_{\varepsilon=0} = \int_{a}^{b} \Big[P(x)\,\eta'(x)^{2} + Q(x)\,\eta(x)^{2} \Big]\,\mathrm{d}x\,, \tag{16.45}$$

where, with all partial derivatives evaluated along the extremal,

$$P = \frac{\partial^{2} f}{\partial y'^{2}}\,, \qquad Q = \frac{\partial^{2} f}{\partial y^{2}} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial^{2} f}{\partial y\,\partial y'} \right). \tag{16.46}$$

Rests on Equation (16.14), Theorem 7.38, and Corollary 7.44.

Proof. *Derives Proposition 16.50.*

Differentiating twice under the integral sign, as in Proposition 16.17,

$$\delta^{2} J[y;\eta] = \int_{a}^{b} \left[\frac{\partial^{2} f}{\partial y^{2}}\,\eta^{2} + 2\,\frac{\partial^{2} f}{\partial y\,\partial y'}\,\eta\,\eta' + \frac{\partial^{2} f}{\partial y'^{2}}\,\eta'^{2} \right] \mathrm{d}x\,.$$

The cross term is a total derivative up to one integration by parts: $2\eta\eta' = (\eta^{2})'$, so by Equation (7.28) and $\eta(a) = \eta(b) = 0$,

$$\int_{a}^{b} \frac{\partial^{2} f}{\partial y\,\partial y'} \left(\eta^{2}\right)' \mathrm{d}x = -\int_{a}^{b} \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial^{2} f}{\partial y\,\partial y'} \right) \eta^{2}\,\mathrm{d}x\,,$$

which absorbed into the first term gives Equation (16.45) with the coefficients Equation (16.46). □

Theorem 16.51 (Legendre's necessary condition)**.** *If the extremal y is a weak local minimum of J, then*

$$P(x) = \frac{\partial^2 f}{\partial y'^2}\,(x, y(x), y'(x)) \geq 0 \qquad \text{for every } x \in [a,b]\,. \tag{16.47}$$

Rests on Equation (16.45) and Proposition 16.16.

Proof. *Derives Theorem 16.51.*

At a weak minimum $\varepsilon \longmapsto J[y+\varepsilon\eta]$ has a minimum at $\varepsilon = 0$, so $\delta^2 J[y;\eta] \geq 0$ for every $\eta \in C_0^1[a,b]$. Suppose instead that $P(x_0) < 0$ at some x_0, which by continuity may be taken interior, and choose $\delta > 0$ and $k > 0$ with $[x_0-\delta, x_0+\delta] \subset (a,b)$ and $P \leq -k$ there. Take the localized variation

$$\eta(x) = \sin^2\!\left(\frac{\pi\,(x - x_0 + \delta)}{2\delta}\right) \quad \text{on } |x - x_0| \leq \delta\,, \qquad \eta = 0 \text{ elsewhere}\,,$$

which lies in $C_0^1[a,b]$ because η and η' both vanish at $x_0 \pm \delta$. Substituting $t = (x - x_0 + \delta)/2\delta$,

$$\int_a^b \eta'^2\,\mathrm{d}x = \frac{\pi^2}{4\delta}\,, \qquad \int_a^b \eta^2\,\mathrm{d}x = \frac{3\delta}{4}\,. \tag{16.48}$$

Hence, writing $M = \max_{[a,b]} |Q|$,

$$\delta^2 J[y;\eta] \leq -k\,\frac{\pi^2}{4\delta} + M\,\frac{3\delta}{4}\,,$$

which is negative for δ small enough — a contradiction. So $P \geq 0$ everywhere. □

Remark 16.52 (The strengthened condition, and what Legendre missed)*.* The *strengthened* Legendre condition is $P > 0$ on $[a,b]$; the problem is then called *regular*. Legendre's memoir on distinguishing maxima from minima [41] introduced the condition and attempted to prove that it is also sufficient, by completing the square in Equation (16.45) with the help of an auxiliary function. The step is correct wherever the auxiliary function exists, and Legendre did not notice that it may fail to exist on the whole interval. The scaling in Equation (16.48) shows why the P term dominates and hence why P alone controls the *local* behaviour: a variation squeezed into a short interval has a large $\int \eta'^2$ and a small $\int \eta^2$. The gap Legendre left is exactly the possibility that a variation spread over the *whole* interval defeats a positive P, and closing it is Jacobi's work (Section 16.4.2). In physics $P > 0$ is the invertibility of the Legendre transform, that is, of the map from velocities to momenta (Chapter 22).

16.4.2 The Jacobi equation and conjugate points

The functional Equation (16.45) is itself a variational problem, and its own Euler–Lagrange equation is the key to the whole second-order theory [101].

Definition 16.53 (Jacobi accessory equation; conjugate point)**.** The *Jacobi accessory equation* of the extremal y is the Euler–Lagrange equation of the quadratic functional Equation (16.45),

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(P(x)\,\frac{\mathrm{d}u}{\mathrm{d}x}\right) - Q(x)\,u = 0\,. \tag{16.49}$$

A point $c \in (a,b]$ is *conjugate* to a along y if the solution u of Equation (16.49) with $u(a) = 0$, $u'(a) = 1$ satisfies $u(c) = 0$.

Rests on Equations (16.20) and (16.45).

Remark 16.54 (Where the Jacobi equation comes from, and what it means). Applying Equation (16.20) to the integrand $Pu'^2 + Qu^2$ gives $2\mathrm{d}(Pu')/\mathrm{d}x - 2Qu = 0$, which is Equation (16.49). Geometrically the equation is the linearisation of Equation (16.20) about the extremal: if $y(x,\alpha)$ is a one-parameter family of extremals through the same initial point, then $u = \partial y/\partial\alpha$ solves Equation (16.49). A conjugate point is therefore a place where neighbouring extremals leaving a come back together to first order — a focus. For geodesics on a sphere the conjugate point of any point is its antipode, which is why a great-circle arc longer than half the circumference is not the shortest path; in optics the same focusing is a caustic (Chapter 23), and in mechanics a kinetic focus.

Theorem 16.55 (Jacobi's necessary condition)**.** *Let the problem be regular, $P > 0$ on $[a,b]$, and let the extremal y be a weak local minimum. Then no point of the open interval (a,b) is conjugate to a.*

Rests on Definition 16.53, Theorem 16.51, and Corollary 16.33.

Proof. *Derives Theorem 16.55.*

Suppose $c \in (a,b)$ is conjugate to a, with u the solution of Equation (16.49) having $u(a) = u(c) = 0$ and $u'(a) = 1$; then u does not vanish identically, and $u'(c) \neq 0$, since a second-order linear solution vanishing together with its derivative at a point vanishes identically. Define the broken variation

$$\eta(x) = u(x) \text{ on } [a,c]\,, \qquad \eta(x) = 0 \text{ on } [c,b]\,,$$

which is continuous and piecewise C^1 with $\eta(a) = \eta(b) = 0$. Multiplying Equation (16.49) by u and integrating by parts over $[a,c]$,

$$\int_a^c \left(Pu'^2 + Qu^2\right)\mathrm{d}x = \left[P\,u\,u'\right]_a^c = 0\,,$$

because u vanishes at both ends. Hence $\delta^2 J[y;\eta] = 0$.

Because y is a weak minimum, $\delta^2 J[y;\cdot] \geq 0$ on $C_0^1[a,b]$, and hence on the larger class of piecewise-C^1 variations vanishing at the endpoints: a corner may be rounded off inside an interval of length ε, which changes $\int (P\eta'^2 + Q\eta^2)\,\mathrm{d}x$ by $O(\varepsilon)$, so the quadratic form is nonnegative on the closure by continuity. So η *minimises* $\delta^2 J$ over the wider class. A minimiser must satisfy the Euler–Lagrange equation Equation (16.49) on each smooth arc, and by the first Weierstrass–Erdmann condition of Corollary 16.33 — here the continuity of $\partial/\partial\eta'\,(P\eta'^2 + Q\eta^2) = 2P\eta'$ — the product $P\eta'$ must be continuous at c. But $P(c)\eta'(c^-) = P(c)u'(c) \neq 0$ while $P(c)\eta'(c^+) = 0$, a contradiction. □

Theorem 16.56 (Sufficiency of the second variation)**.** *Let $P > 0$ on $[a,b]$ and suppose the Jacobi equation Equation (16.49) has a solution w with no zero on $[a,b]$. Then $\delta^2 J[y;\eta] > 0$ for every $\eta \in C_0^1[a,b]$ that does not vanish identically.*

Rests on Equations (16.45) and (16.49).

Proof. *Derives Theorem 16.56.*

Set $W = -P\,w'/w$, which is C^1 on $[a,b]$ because w has no zero there. Differentiating and using Equation (16.49) for w,

$$W' = -\frac{(Pw')'}{w} + \frac{P\,w'^2}{w^2} = -Q + \frac{W^2}{P}\,. \tag{16.50}$$

For any $\eta \in C_0^1[a,b]$ the quantity $W\eta^2$ vanishes at both endpoints, so $\int_a^b (W\eta^2)'\,\mathrm{d}x = 0$ and this may be added to Equation (16.45) free of charge:

$$\delta^2 J[y;\eta] = \int_a^b \left[P\eta'^2 + 2W\eta\eta' + (Q + W')\,\eta^2\right]\mathrm{d}x = \int_a^b P\left(\eta' + \frac{W}{P}\,\eta\right)^2\mathrm{d}x\,,$$

the completion of the square being exact precisely because Equation (16.50) gives $Q + W' = W^2/P$. The integrand is nonnegative, so $\delta^2 J \geq 0$, and it vanishes only if $\eta' = -W\eta/P$ throughout. That linear equation with $\eta(a) = 0$ has only the zero solution, so $\delta^2 J[y;\eta] > 0$ unless $\eta \equiv 0$. □

Remark 16.57 (The two conditions are two faces of one). The hypothesis of Theorem 16.56 — a Jacobi solution without zeros on the closed interval — is the *strengthened Jacobi condition*, and it is equivalent, for a regular problem, to the absence of any point conjugate to a in $(a, b]$: a solution vanishing at $a-\varepsilon$ for small $\varepsilon > 0$ has, by continuous dependence of solutions on their initial data, no zero in $[a, b]$ exactly when the solution vanishing at a has none in $(a, b]$. The equivalence is proved in Appendix A.40, where the continuous dependence on the initial point is derived from the integral form of the initial value problem by a Grönwall estimate and the Sturm separation theorem from the Lagrange identity Equation (9.56); with it, the necessary condition Theorem 16.55 and the sufficient condition Theorem 16.56 test the same hypothesis, and differ only at the single point $x = b$.

Full derivation: Appendix A.40 (p. 2988).

Derives Remark 16.57.

Remark 16.58 (Counting conjugate points). Conjugate points are not merely an obstruction to be avoided; their number is an invariant. Morse's index theorem states that the index of the second variation — the maximal dimension of a subspace on which $\delta^2 J$ is negative definite — equals the number of points conjugate to a in (a, b), counted with multiplicity [746]. The theorem is quoted here and not used; it returns for geodesics in Chapter 43, where it converts a local calculation into a statement about the topology of the space of paths.

16.4.3 Strong variations and the Weierstrass condition

Everything in Sections 16.4.1 and 16.4.2 tests the extremal against curves whose *slopes* are close to its own, and Example 16.14 shows that this is not enough. A strong variation changes the slope by a finite amount over a short interval, and the quantity that measures its effect is Weierstrass's excess function [591].

Definition 16.59 (Weierstrass excess function)**.** For an integrand f of class C^2,

$$E(x, y, p, q) = f(x, y, q) - f(x, y, p) - (q - p)\,\frac{\partial f}{\partial y'}(x, y, p)\,. \tag{16.51}$$

It is the amount by which f, as a function of the slope alone at fixed (x, y), lies above its tangent line at p: the excess of the value at q over the linear extrapolation from p.

Rests on Definition 16.11.

Theorem 16.60 (Weierstrass' necessary condition)**.** *Let y be a strong local minimum of J. Then*

$$E(x, y(x), y'(x), q) \geq 0 \qquad \text{for every } x \in [a, b] \text{ and every } q \in \mathbb{R}\,. \tag{16.52}$$

Rests on Equation (16.51) and Definition 16.13.

Proof. *Derives Theorem 16.60.*

Fix an interior x_0, write $y_0 = y(x_0)$, fix $q \in \mathbb{R}$ and fix $\Delta > 0$ small enough that $[x_0, x_0 + \Delta] \subset (a, b)$. For $0 \leq h < \Delta$ define the comparison curve Y_h: equal to y outside $(x_0, x_0 + \Delta)$; the segment of slope q from (x_0, y_0) on $[x_0, x_0 + h]$; and on $[x_0 + h, x_0 + \Delta]$ the straight segment rejoining $(x_0 + \Delta, y(x_0 + \Delta))$, whose slope is

$$s_h = \frac{y(x_0 + \Delta) - y_0 - qh}{\Delta - h}\,, \qquad s_0 = s_\Delta \equiv \frac{y(x_0 + \Delta) - y_0}{\Delta}\,.$$

Each Y_h is admissible and piecewise C^1, and $\|Y_h - y\|_0 \longrightarrow 0$ as $h \to 0$ at fixed Δ, so for small h it lies in the strong neighbourhood on which y minimises. Hence $G(h) \equiv J[Y_h] - J[y]$ satisfies $G(0) = 0$ and $G(h) \geq 0$, so $G'(0^+) \geq 0$.

Only the interval $(x_0, x_0 + \Delta)$ contributes. Writing $\ell_h(x) = y_0 + qh + s_h\,(x - x_0 - h)$ for the rejoining segment,

$$J[Y_h]\big|_{(x_0, x_0+\Delta)} = \int_{x_0}^{x_0+h} f\,(x, y_0 + q(x - x_0), q)\,\mathrm{d}x + \int_{x_0+h}^{x_0+\Delta} f\,(x, \ell_h(x), s_h)\,\mathrm{d}x\,.$$

Differentiating at $h = 0$, where $\ell_0(x) = y_0 + s_\Delta(x - x_0)$, the two moving limits give $f\,(x_0, y_0, q) - f\,(x_0, y_0, s_\Delta)$, while

$$\left.\frac{\partial s_h}{\partial h}\right|_0 = \frac{s_\Delta - q}{\Delta}\,, \qquad \left.\frac{\partial \ell_h(x)}{\partial h}\right|_0 = (q - s_\Delta)\left(1 - \frac{x - x_0}{\Delta}\right)\,,$$

so that

$$G'(0^+) = f\,(x_0, y_0, q) - f\,(x_0, y_0, s_\Delta) + (q - s_\Delta)\int_{x_0}^{x_0+\Delta}\left[\frac{\partial f}{\partial y}\left(1 - \frac{x - x_0}{\Delta}\right) - \frac{1}{\Delta}\,\frac{\partial f}{\partial y'}\right]\mathrm{d}x\,, \tag{16.53}$$

the partial derivatives being evaluated at $(x, \ell_0(x), s_\Delta)$. Now let $\Delta \to 0$. The first part of the integral is $O(\Delta)$; the second tends to $\partial f/\partial y'\,(x_0, y_0, y'(x_0))$, since $\Delta^{-1}\int_{x_0}^{x_0+\Delta}$ of a continuous function tends to its value at x_0 and $s_\Delta \to y'(x_0)$ by the mean value theorem (Theorem 7.35). Writing $p = y'(x_0)$, the limit of Equation (16.53) is $f(x_0, y_0, q) - f(x_0, y_0, p) - (q - p)\,\partial f/\partial y'(x_0, y_0, p) = E\,(x_0, y_0, p, q)$. Since $G'(0^+) \geq 0$ for every Δ, the limit is nonnegative, which is Equation (16.52) at x_0; continuity extends it to the endpoints. □

Remark 16.61 (Excess, convexity and the Legendre transform). By Taylor's theorem with Lagrange remainder (Theorem 7.38) applied to $q \longmapsto f\,(x, y, q)$,

$$E\,(x, y, p, q) = \tfrac{1}{2}\,(q - p)^2\,\frac{\partial^2 f}{\partial y'^2}\,(x, y, \bar{q}) \tag{16.54}$$

for some $\bar{q}$ between p and q. Hence $E \geq 0$ for all q if f is convex in the slope, and taking $q \to p$ in Equation (16.54) recovers Legendre's condition $\partial^2 f/\partial y'^2 \geq 0$: Weierstrass' condition contains Legendre's, and the two agree exactly when the slope is varied infinitesimally. This is also the hypothesis under which the Legendre transform to the Hamiltonian formalism of Chapter 22 is invertible, and under which the convexity requirement of Tonelli's existence theorem (Section 16.5.2) holds. Three separate questions — strong minimality, the passage to the Hamiltonian, and existence — turn on the same convexity of f in y'.

16.4.4 Fields of extremals and sufficiency

All the conditions so far are necessary. Turning them into a sufficient condition needs one further construction: the extremal must be embedded in a family covering a whole neighbourhood of it.

Definition 16.62 (Field of extremals; slope function)**.** A *field of extremals* on a simply connected region $R \subset \mathbb{R}^2$ is a one-parameter family of extremals of J such that exactly one member passes through each point of R, depending C^1 on the point. Its *slope function* $p : R \longrightarrow \mathbb{R}$ assigns to each $(x, y) \in R$ the slope at that point of the unique member through it. An extremal is *embedded* in the field if its graph lies in R.

Rests on Theorem 16.22 and Definition 6.17.

Theorem 16.63 (Hilbert's invariant integral)**.** *Let p be the slope function of a field of extremals on R. Then the line integral*

$$I[\Gamma] = \int_\Gamma \left\{ \left[f(x,y,p) - p\,\frac{\partial f}{\partial y'}(x,y,p) \right] \mathrm{d}x + \frac{\partial f}{\partial y'}(x,y,p)\,\mathrm{d}y \right\} \tag{16.55}$$

depends only on the endpoints of the curve $\Gamma \subset R$, not on the curve. Along any member of the field itself, $I = \int f(x,y,y')\,\mathrm{d}x = J$.

Rests on Definition 16.62, Proposition 7.104, and Equation (16.20).

Proof. *Derives Theorem 16.63.*

Write the integrand as $\mathcal{P}\,\mathrm{d}x + \mathcal{Q}\,\mathrm{d}y$ with

$$\mathcal{P}(x,y) = f(x,y,p) - p\,\frac{\partial f}{\partial y'}(x,y,p)\,, \qquad \mathcal{Q}(x,y) = \frac{\partial f}{\partial y'}(x,y,p)\,,$$

both C^1 on R. Path independence on a simply connected region is equivalent to $\partial\mathcal{P}/\partial y = \partial\mathcal{Q}/\partial x$ (Proposition 7.104). Abbreviating f_y, $f_{y'}$, $f_{xy'}$, $f_{yy'}$, $f_{y'y'}$ for the partial derivatives evaluated at $(x,y,p(x,y))$, the chain rule gives

$$\begin{aligned} \frac{\partial \mathcal{Q}}{\partial x} &= f_{xy'} + f_{y'y'}\,\frac{\partial p}{\partial x}\,, \\ \frac{\partial \mathcal{P}}{\partial y} &= f_y + f_{y'}\,\frac{\partial p}{\partial y} - \frac{\partial p}{\partial y}\,f_{y'} - p\left(f_{yy'} + f_{y'y'}\,\frac{\partial p}{\partial y}\right) = f_y - p\,f_{yy'} - p\,f_{y'y'}\,\frac{\partial p}{\partial y}\,. \end{aligned}$$

Each member of the field satisfies Equation (16.20), and along it $\mathrm{d}/\mathrm{d}x = \partial_x + p\,\partial_y + (\partial p/\partial x + p\,\partial p/\partial y)\,\partial_{y'}$, so the Euler–Lagrange equation reads

$$f_{xy'} + p\,f_{yy'} + \left(\frac{\partial p}{\partial x} + p\,\frac{\partial p}{\partial y}\right) f_{y'y'} = f_y\,. \tag{16.56}$$

Subtracting the two displays,

$$\frac{\partial \mathcal{P}}{\partial y} - \frac{\partial \mathcal{Q}}{\partial x} = f_y - \left[p\,f_{yy'} + f_{xy'} + f_{y'y'}\left(\frac{\partial p}{\partial x} + p\,\frac{\partial p}{\partial y}\right)\right] = 0$$

by Equation (16.56). Finally, on a member of the field $\mathrm{d}y = p\,\mathrm{d}x$, so the two $f_{y'}$ terms in Equation (16.55) cancel and the integrand reduces to $f(x,y,p)\,\mathrm{d}x = f(x,y,y')\,\mathrm{d}x$. □

Theorem 16.64 (Weierstrass' sufficient condition)**.** *Let the extremal y^* be embedded in a field of extremals on R with slope function p, joining (a,y_a) to (b,y_b), and suppose*

$$E(x,y,p(x,y),q) \geq 0 \qquad \text{for all } (x,y) \in R \text{ and all } q \in \mathbb{R}\,. \tag{16.57}$$

Then $J[Y] \geq J[y^*]$ for every admissible piecewise-C^1 curve Y whose graph lies in R: y^* is a strong minimum relative to R.

Rests on Theorem 16.63, Equation (16.51), and Definition 16.62.

Proof. *Derives Theorem 16.64.*

Let Y be admissible with graph in R. By Theorem 16.63, I takes the same value on the graph of Y as on the graph of y^*, and on the latter it equals $J[y^*]$. Evaluating I along the graph of Y, where $\mathrm{d}y = Y'\,\mathrm{d}x$,

$$J[y^*] = I[Y] = \int_a^b \left[f(x,Y,p) + (Y' - p)\,\frac{\partial f}{\partial y'}(x,Y,p) \right] \mathrm{d}x\,,$$

with $p = p(x,Y(x))$. Subtracting from $J[Y] = \int_a^b f(x,Y,Y')\,\mathrm{d}x$ and comparing with Equation (16.51),

$$J[Y] - J[y^*] = \int_a^b E\left(x, Y(x), p\left(x,Y(x)\right), Y'(x)\right)\mathrm{d}x \geq 0 \tag{16.58}$$

by Equation (16.57). □

Corollary 16.65 (Sufficiency for a weak minimum)**.** *Let y^* be an extremal satisfying the strengthened Legendre condition $P > 0$ on $[a,b]$ and the strengthened Jacobi condition of Remark 16.57. Then y^* is a strict weak local minimum of J.*

Rests on Theorems 16.51 and 16.56.

Proof. *Derives Corollary 16.65.*

Theorem 16.56 gives $\delta^2 J[y^*;\eta] > 0$ for every admissible $\eta \neq 0$, and that by itself is *not* enough: the infimum of $\delta^2 J[y^*;\cdot]$ over the sphere $\|\eta\|_1 = 1$ is zero. A smoothed tent of width δ and unit maximum slope near an interior point has $\delta^2 J = O(\delta)$ while its C^1 norm tends to 1, so no positive lower bound survives the passage to the limit and the positivity must be upgraded to coercivity first.

Step 1: a coercive lower bound. Let w be the zero-free Jacobi solution supplied by the strengthened Jacobi condition, and let $W = -P\,w'/w$ and $\mu(x) = \exp\int_a^x (W/P)\,\mathrm{d}s$, so that $\mu > 0$ is C^1 on $[a,b]$. The completion of the square in the proof of Theorem 16.56 is the exact identity

$$\delta^2 J[y^*;\eta] = \int_a^b P\,v^2\,\mathrm{d}x\,, \qquad v = \eta' + \frac{W}{P}\,\eta = \frac{(\mu\eta)'}{\mu}\,,$$

valid for every $\eta \in C_0^1[a,b]$. Since $\eta(a) = 0$ we have $\mu(x)\eta(x) = \int_a^x \mu\, v\,\mathrm{d}s$, so the Cauchy–Schwarz inequality bounds $\max|\eta|$ by a constant depending only on μ and $b-a$ times $\left(\int_a^b v^2\right)^{1/2}$; hence $\int \eta^2 \le \kappa_2 \int v^2$, and $\eta' = v - W\eta/P$ then gives $\int \eta'^2 \le \kappa_3 \int v^2$ with κ_3 likewise independent of η. With $P \ge P_{\min} > 0$ this is the coercivity estimate

$$\delta^2 J[y^*;\eta] \ge c \int_a^b \eta'^2\,\mathrm{d}x\,, \qquad c = \frac{P_{\min}}{\kappa_3} > 0\,. \tag{16.59}$$

Step 2: the remainder. For $\eta \in C_0^1[a,b]$, Taylor's theorem (Theorem 7.38) applied to f in its last two arguments, the vanishing of the first variation, and the integration by parts of Proposition 16.50 give

$$J[y^*+\eta] - J[y^*] = \tfrac{1}{2}\,\delta^2 J[y^*;\eta] + R[\eta]\,, \qquad |R[\eta]| \le \omega\left(\|\eta\|_1\right) \int_a^b \left(\eta^2 + \eta'^2\right)\mathrm{d}x\,,$$

where $\omega(t) \to 0$ as $t \to 0$: the second derivatives of f are uniformly continuous on a compact neighbourhood of the graph of y^* because f is C^3, and $2\,|\eta\eta'| \le \eta^2 + \eta'^2$. This is the step the norm on the sphere could not supply — the remainder is now measured against $\int(\eta^2+\eta'^2)$, the same quantity Equation (16.59) controls from below.

Step 3. Because $\eta(a) = 0$, Cauchy–Schwarz also gives $\int \eta^2 \le (b-a)^2 \int \eta'^2$. Writing $\kappa = 1 + (b-a)^2$ and combining the two steps,

$$J[y^*+\eta] - J[y^*] \ge \left[\frac{c}{2} - \kappa\,\omega\left(\|\eta\|_1\right)\right] \int_a^b \eta'^2\,\mathrm{d}x\,.$$

Choose $\varepsilon > 0$ with $\kappa\,\omega(\varepsilon) < c/2$. For every $\eta \neq 0$ with $\|\eta\|_1 < \varepsilon$ the bracket is positive and $\int \eta'^2 > 0$, since an $\eta \in C_0^1[a,b]$ with $\eta' \equiv 0$ vanishes identically. So y^* is a strict weak local minimum. □

Remark 16.66 (Where this construction turns into physics). The slope function of a field is a gradient: on the region R the invariant integral Equation (16.55) defines a function $S(x,y)$, unique up to a constant, with $\partial S/\partial y = \partial f/\partial y'\,(x,y,p)$ and $\partial S/\partial x = f - p\,\partial f/\partial y'$. Those two statements together are the Hamilton–Jacobi equation, and S is Hamilton's characteristic function (Chapter 23). The Weierstrass sufficiency theory and Hamilton–Jacobi theory are therefore the same construction read in two directions: the physicist solves a first-order partial differential equation to generate the trajectories, the analyst builds the family of trajectories to certify a minimum. Hilbert announced the invariant integral in the lecture that also posed his twenty-third problem, a call for exactly this kind of systematic development of the calculus of variations [291].

16.5 Direct methods and existence

Everything so far has assumed that a minimiser exists and asked what equations it must satisfy. The *direct* methods reverse the order: they establish that a minimiser exists before writing any differential equation, and then read the Euler–Lagrange equation off it. The nineteenth century learned the hard way that the first question cannot be skipped.

16.5.1 The Dirichlet principle and its failure

Definition 16.67 (The Dirichlet principle)**.** Let $\Omega \subset \mathbb{R}^3$ be a bounded domain and h a continuous function on $\partial\Omega$. The *Dirichlet principle* is the assertion that the *Dirichlet integral*

$$D[u] = \int_\Omega |\nabla u|^2 \, \mathrm{d}^3x \tag{16.60}$$

attains its infimum over $\left\{ u \in C^1(\overline{\Omega}) \mid u|_{\partial\Omega} = h \right\}$, and that the minimiser is the harmonic function with boundary values h.

Rests on Equations (16.32) *and* (16.34)*.*

The second half of the assertion is a theorem: by Example 16.37 the Euler–Lagrange equation of Equation (16.60) is Laplace's equation, so a minimiser, if there is one, is harmonic. Riemann used the principle in this form as the existence engine of his function theory — it is how he produced the conformal maps and the functions on Riemann surfaces that his geometry required — and named it after Dirichlet, from whose lectures he had it. The first half of the assertion is the problem: an infimum need not be attained.

Proposition 16.68 (Weierstrass' counterexample)**.** *Let*

$$\mathcal{C} = \left\{ y \in C^1[-1,1] \mid y(-1) = -1, \; y(1) = 1 \right\},$$

and

$$J[y] = \int_{-1}^{1} x^2 \, y'(x)^2 \, \mathrm{d}x \,. \tag{16.61}$$

Then $\inf_{\mathcal{C}} J = 0$*, and no* $y \in \mathcal{C}$ *attains it.*

Rests on Equation (16.61) *and Definition 16.11.*

Proof. *Derives Proposition 16.68.*

The integrand is nonnegative, so $J \geq 0$. For $\varepsilon > 0$ put

$$y_\varepsilon(x) = \frac{\arctan\left(x/\varepsilon\right)}{\arctan\left(1/\varepsilon\right)},$$

which lies in $\mathcal{C}$, being smooth and odd with $y_\varepsilon(\pm 1) = \pm 1$. Its derivative is $y_\varepsilon' = \varepsilon[\arctan(1/\varepsilon)]^{-1} \left(\varepsilon^2 + x^2\right)^{-1}$, so, using $x^2 \leq \varepsilon^2 + x^2$,

$$J\left[y_\varepsilon\right] = \frac{\varepsilon^2}{\arctan^2(1/\varepsilon)} \int_{-1}^{1} \frac{x^2 \, \mathrm{d}x}{\left(\varepsilon^2 + x^2\right)^2} \leq \frac{\varepsilon^2}{\arctan^2(1/\varepsilon)} \int_{-1}^{1} \frac{\mathrm{d}x}{\varepsilon^2 + x^2} = \frac{2\varepsilon}{\arctan\left(1/\varepsilon\right)},$$

which tends to 0 as $\varepsilon \to 0^+$ because $\arctan(1/\varepsilon) \to \pi/2$. Hence the infimum is 0.

Suppose now $J[y] = 0$ for some $y \in \mathcal{C}$. The integrand $x^2 y'^2$ is continuous and nonnegative with vanishing integral, so it vanishes identically; hence $y'(x) = 0$ for every $x \neq 0$, and y is constant on $[-1, 0)$ and on $(0, 1]$. Continuity at 0 forces the two constants to agree, so y is constant on $[-1, 1]$, contradicting $y(-1) = -1 \neq 1 = y(1)$. □

Remark 16.69 (What the example destroyed, and what replaced it). Weierstrass presented this argument to the Berlin Academy in 1870 [270], and its effect was immediate: every existence proof resting on "the infimum is attained" was suspect, Riemann's function theory included. The gap is real and not an artefact of the example: the infimum of a functional bounded below is always approached by a minimising sequence, but nothing yet said that the sequence converges, nor that the functional is continuous along it.

Hilbert restored the Dirichlet principle thirty years later [314], not by patching Riemann's argument but by supplying the two missing ingredients under explicit hypotheses on the domain and the boundary data: a minimising sequence with enough compactness to extract a convergent subsequence, and enough semicontinuity of D to conclude that the limit is a minimiser. This is the direct method, and its abstract form is Theorem 16.70. Hilbert's twentieth and twenty-third problems both grew out of it [291], and the modern treatment is Courant–Hilbert [1222]. The stake for physics is not decorative: the existence of solutions to the boundary-value problems of Chapter 10 and Chapter 58 is exactly the statement Riemann assumed and Weierstrass questioned.

16.5.2 Tonelli's existence theory

The direct method has three ingredients, and each answers one of the two objections above.

Theorem 16.70 (The direct method)**.** *Let X be a reflexive Banach space and $J : X \longrightarrow \mathbb{R} \cup \{+\infty\}$ a functional that is not identically $+\infty$ and satisfies*

1. coercivity*: $J[u] \longrightarrow +\infty$ whenever $\|u\| \longrightarrow \infty$;*
2. sequential weak lower semicontinuity*: whenever $u_n \rightharpoonup u$ weakly in X, $J[u] \le \liminf_n J[u_n]$.*

Then J attains its infimum on X.
Rests on Definitions 6.9 and 12.2.

Proof. *Derives Theorem 16.70.*
Let $m = \inf_X J < +\infty$ and pick a minimising sequence (u_n), so $J[u_n] \longrightarrow m$. The sequence $(J[u_n])$ is bounded above, so by coercivity $(\|u_n\|)$ is bounded: otherwise a subsequence would have $\|u_{n_k}\| \to \infty$ and hence $J[u_{n_k}] \to +\infty$. In a reflexive Banach space a bounded sequence has a weakly convergent subsequence — this is the compactness input, the Eberlein–Smulian theorem, whose Hilbert-space case is the one used in this book (Chapter 12) — so $u_{n_k} \rightharpoonup u$ for some $u \in X$. Weak lower semicontinuity then gives

$$J[u] \le \liminf_k J\,[u_{n_k}] = m\,,$$

and $J[u] \ge m$ by the definition of the infimum, so $J[u] = m > -\infty$. □

Remark 16.71 (The three ingredients pull against each other). Coercivity and semicontinuity are demands on the same topology in opposite directions. Making the topology coarser makes compactness easier and semicontinuity harder; making it finer does the reverse. The weak topology of a reflexive space is the compromise that works, and it is why Sobolev spaces — not spaces of continuously differentiable functions — are the right home for the direct method: $C^1[a,b]$ with the norm Equation (16.13) is not reflexive and its bounded sets are not weakly compact. That is the technical form of the objection Weierstrass raised.

The remaining question is which integrands make Equation (16.11) weakly lower semicontinuous. Tonelli identified the answer: convexity in the slope [432].

Lemma 16.72 (Convexity implies weak lower semicontinuity)**.** *Let $1 < r < \infty$, let $f(x,q)$ be continuous, bounded below, of class C^1 in q and convex in q for each x, and suppose $|\partial f/\partial q\,(x,q)| \le C\left(1+|q|^{r-1}\right)$. If $u_n' \rightharpoonup u'$ weakly in $L^r(a,b)$, then*

$$\int_a^b f\,(x,u')\,\mathrm{d}x \le \liminf_n \int_a^b f\,(x,u_n')\,\mathrm{d}x\,.$$

Rests on Theorem 16.70 and Equation (16.11).

Proof. *Derives Lemma 16.72.*

Convexity of a C^1 function means its graph lies above every tangent:

$$f\,(x,q_2) \ge f\,(x,q_1) + \frac{\partial f}{\partial q}\,(x,q_1)\,(q_2-q_1)\ . \tag{16.62}$$

Apply this pointwise with $q_1 = u'(x)$ and $q_2 = u_n'(x)$ and integrate:

$$\int_a^b f\,(x,u_n')\,\mathrm{d}x \ge \int_a^b f\,(x,u')\,\mathrm{d}x + \int_a^b \frac{\partial f}{\partial q}\,(x,u')\,(u_n'-u')\,\mathrm{d}x\,.$$

The growth hypothesis puts $\partial f/\partial q\,(\cdot,u')$ in $L^{r'}(a,b)$ with $r' = r/(r-1)$, which is the dual of L^r, so the last integral tends to 0 by the definition of weak convergence. Taking $\liminf$ of both sides gives the claim. □

Theorem 16.73 (Tonelli)**.** *Let $1 < r < \infty$ and let $f\,(x,y,q)$ be continuous, of class C^1 in q, convex in q for each (x,y), and coercive in the sense $f\,(x,y,q) \ge \alpha\,|q|^r - \beta$ with $\alpha > 0$; assume further the growth bound $|\partial f/\partial q\,(x,y,q)| \le C_R\left(1+|q|^{r-1}\right)$ for $|y| \le R$, which is the hypothesis of Lemma 16.72 made locally uniform in y. Then the functional Equation (16.11) attains its minimum over the class of functions in the Sobolev space $W^{1,r}(a,b)$ with the prescribed endpoint values.*

Rests on Theorem 16.70 and Lemma 16.72.

The derivation is carried out in Appendix A.41. The coercivity estimate and the abstract argument are those of the direct-method theorem and the convexity lemma proved above; the passage from an integrand $f(x,q)$ to one that also depends on y rests on the compact embedding of $W^{1,r}(a,b)$ into the continuous functions, which the appendix proves from the fundamental theorem of calculus, Hölder's inequality and the Arzelà–Ascoli theorem. The growth bound is what lets that embedding replace the Scorza-Dragoni argument of the classical treatments; the appendix states precisely what dropping it would cost, and Tonelli's own hypotheses [432] are the ones without it.

Full derivation: Appendix A.41 (p. 2994).

Derives Theorem 16.73.

Remark 16.74 (What the physical Lagrangians do and do not satisfy). For a mechanical system with kinetic energy $T = \tfrac12 m_{ij}(q)\dot q^i\dot q^j$ the Lagrangian $\mathcal{L} = T - V$ is quadratic in the velocities with a positive definite coefficient matrix, hence convex in them, which is half of what Theorem 16.73 asks. The other half is coercivity, and it is a condition on the potential in the direction opposite to the one a physicist expects: $\mathcal{L} \ge \alpha\,|\dot q|^2 - \beta$ uniformly in q demands that V be bounded *above*, not below. For a confining potential the action is in general not bounded below at all. The harmonic oscillator, $m,k>0$ and $q(0)=q(T)=0$, is the smallest example: on $q(t) = A\sin(\pi t/T)$,

$$S = \int_0^T \left(\tfrac12 m\dot q^2 - \tfrac12 k q^2\right)\mathrm{d}t = \frac{A^2}{4}\left(\frac{m\pi^2}{T} - kT\right), \tag{16.63}$$

which for $T > \pi\sqrt{m/k} = \pi/\omega$, with $\omega = \sqrt{k/m}$, is negative and tends to $-\infty$ as $A \to \infty$: the infimum is $-\infty$ and no minimiser exists in any class. That threshold is exactly the

kinetic focus of Remark 16.98, and the same phenomenon as in the model functional of Remark 16.25. So Theorem 16.73 certifies existence for a mechanical Lagrangian whose potential is bounded above — a free or a repelled particle — and not for a confining one; past the first conjugate point the action of Section 16.7.1 is genuinely unbounded below, which is why that section states Hamilton's principle as one of stationary and not of least action. Convexity is not a technicality that can be dropped either. For

$$J[y] = \int_0^1 \left[\left(y'^2 - 1 \right)^2 + y^2 \right] \mathrm{d}x \,, \qquad y(0) = y(1) = 0 \,, \tag{16.64}$$

the integrand is not convex in y', and $\inf J = 0$: a sawtooth of slope ± 1 and amplitude $1/2n$ makes both terms as small as one likes. No function attains 0, since that would need $y \equiv 0$ and $|y'| \equiv 1$ at once. Minimising sequences oscillate faster and faster — a fine structure with a physical counterpart in the twinned microstructure of martensitic phase transitions.

16.5.3 The Rayleigh–Ritz method

The direct method has a computational face. Instead of minimising over the whole admissible class, minimise over a finite-dimensional family of trial functions: the result is an upper bound on the true minimum, and a linear algebra problem in place of a differential equation.

Definition 16.75 (Rayleigh quotient)**.** For the Sturm–Liouville data $p > 0$, q and $r > 0$ continuous on $[a, b]$, the *Rayleigh quotient* of a nonzero $y \in C_0^1[a, b]$ is

$$\mathcal{R}[y] = \frac{\displaystyle\int_a^b \left(p\, y'^2 + q\, y^2 \right) \mathrm{d}x}{\displaystyle\int_a^b r\, y^2 \, \mathrm{d}x} \,. \tag{16.65}$$

Rests on Equation (16.42) and Definition 9.54.

Theorem 16.76 (Variational characterisation of the lowest eigenvalue)**.** *Let $\lambda_1 \le \lambda_2 \le \cdots$ be the eigenvalues of the Sturm–Liouville problem*

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(p \frac{\mathrm{d}y}{\mathrm{d}x} \right) + q\, y = \lambda\, r\, y \,, \qquad y(a) = y(b) = 0 \,, \tag{16.66}$$

with eigenfunctions y_n orthonormal in the weighted inner product $\langle u, v \rangle_r = \int_a^b r\, u\, v \, \mathrm{d}x$ and complete in the corresponding Hilbert space (Chapter 9, Chapter 12). Then

$$\lambda_1 = \min \left\{ \mathcal{R}[y] \mid y \in C_0^1[a, b],\ y \neq 0 \right\} \,, \tag{16.67}$$

the minimum being attained exactly at the multiples of y_1.

Rests on Equations (16.65) and (16.66).

Proof. *Derives Theorem 16.76.*

Write $B[u, v] = \int_a^b (p\, u'v' + q\, uv) \, \mathrm{d}x$ for the energy form, so that $\mathcal{R}[y] = B[y, y] / \langle y, y \rangle_r$, and fix $K > 0$ with $q + Kr \ge 0$ on $[a, b]$, which is possible because q is continuous and $r > 0$ on a compact interval. Put $B_K[u, v] = B[u, v] + K \langle u, v \rangle_r$. Since $p > 0$ and $q + Kr \ge 0$,

$$B_K[u, u] = \int_a^b \left[p\, u'^2 + (q + Kr)\, u^2 \right] \mathrm{d}x \ge 0 \qquad \text{for every } u \in C^1[a, b] \,. \tag{16.68}$$

Expand $y = \sum_{n \ge 1} c_n y_n$ with $c_n = \langle y, y_n \rangle_r$; by completeness the denominator of Equation (16.65) is $\sum_n c_n^2$.

The one integration by parts needed is against the eigenfunctions, not against y: for $y \in C^1_0[a,b]$ and $y_n \in C^2[a,b]$, Equation (7.28) with $y(a) = y(b) = 0$ moves the derivative onto y_n,

$$B[y, y_n] = \int_a^b y \left[-\frac{\mathrm{d}}{\mathrm{d}x}\left(p\,\frac{\mathrm{d}y_n}{\mathrm{d}x} \right) + q\,y_n \right] \mathrm{d}x = \lambda_n \langle y, y_n \rangle_r = \lambda_n c_n \,, \tag{16.69}$$

by Equation (16.66). Only y_n is differentiated twice, so nothing beyond $y \in C^1_0[a,b]$ is used — which matters, because the trial functions of Proposition 16.79 are only C^1. Hence $B_K[y, y_n] = (\lambda_n + K)\, c_n$, and for the partial sums $S_N = \sum_{n \le N} c_n y_n$ bilinearity and orthonormality give $B_K[y, S_N] = B_K[S_N, S_N] = \sum_{n \le N} (\lambda_n + K)\, c_n^2$. Therefore

$$0 \le B_K[y - S_N, y - S_N] = B_K[y, y] - \sum_{n \le N} (\lambda_n + K)\, c_n^2$$

by Equation (16.68). Taking $u = y_1$ there, together with Equation (16.69) for $y = y_1$, shows $B_K[y_1, y_1] = \lambda_1 + K \ge 0$, so every term of the sum is nonnegative and $N \to \infty$ is legitimate: $B_K[y, y] \ge \sum_n (\lambda_n + K)\, c_n^2 \ge (\lambda_1 + K) \sum_n c_n^2$. Subtracting $K \langle y, y \rangle_r = K \sum_n c_n^2$ from both sides leaves $B[y, y] \ge \lambda_1 \langle y, y \rangle_r$, that is

$$\mathcal{R}[y] \ge \frac{\sum_n \lambda_n c_n^2}{\sum_n c_n^2} \ge \lambda_1 \,. \tag{16.70}$$

If $\mathcal{R}[y] = \lambda_1$ then both inequalities in Equation (16.70) are equalities, and the second forces $c_n = 0$ whenever $\lambda_n > \lambda_1$; the lowest eigenvalue of a regular Sturm–Liouville problem is simple, so y agrees with a multiple of y_1 in the weighted inner product and, both being continuous, pointwise. Taking $y = y_1$ in Equation (16.69) gives $\mathcal{R}[y_1] = \lambda_1$, so the value is attained. □

Remark 16.77 (The inequality is elementary, the identity is not). Only the *inequality* Equation (16.70) was proved, and deliberately so: it is all the variational characterisation needs, and it costs nothing beyond $y \in C^1_0[a,b]$. The sharper statement

$$\int_a^b \left(p\,y'^2 + q\,y^2 \right) \mathrm{d}x = \sum_{n \ge 1} \lambda_n c_n^2 \tag{16.71}$$

is equivalent to $B_K[y - S_N, y - S_N] \to 0$, that is to completeness of the eigenfunctions in the *energy* norm $\|u\|_E = B_K[u,u]^{1/2}$ — a strictly stronger fact than the completeness in the weighted L^2 space assumed in Theorem 16.76, and one proved in Appendix A.42 by realising the inverse of the Sturm–Liouville operator as a compact integral operator with the Green function as its kernel and applying the Hilbert–Schmidt theorem Theorem 12.44 to it. That appendix also proves the weighted L^2 completeness *assumed* here, so the hypothesis of Theorem 16.76 is discharged as well; the inequality Equation (16.70) proved above, which is what the variational characterisation uses, rests on Bessel's inequality alone and is independent of both, so there is no circle.

Full derivation: Appendix A.42 (p. 3001).

Derives Equation (16.71).

The distinction is not pedantry: it is exactly the identity that the error estimate in Corollary 16.78 needs, and the inequality alone bounds that error in the wrong direction.

Corollary 16.78 (The variational bound)**.** *For every admissible trial function $\chi \neq 0$, $\lambda_1 \le \mathcal{R}[\chi]$. A trial function accurate to first order in the deviation from y_1 gives an eigenvalue accurate to second order.*

Rests on Equations (16.67) and (16.71).

Proof. *Derives Corollary 16.78.*

The inequality is Equation (16.67). For the error statement, write $\chi = y_1 + \epsilon$ with $\langle \epsilon, y_1 \rangle_r = 0$; then $c_1 = 1$ and $\sum_{n \geq 2} c_n^2 = \|\epsilon\|_r^2$, so Equation (16.71) gives

$$\mathcal{R}[\chi] - \lambda_1 = \frac{\sum_{n \geq 2} (\lambda_n - \lambda_1)\, c_n^2}{1 + \|\epsilon\|_r^2} = \frac{B[\epsilon, \epsilon] - \lambda_1 \, \|\epsilon\|_r^2}{1 + \|\epsilon\|_r^2} \,,$$

whose numerator is a quadratic form in ϵ: the error is second order in the deviation. It must be measured in the energy norm of Remark 16.77 and not in $\|\cdot\|_r$, since $\lambda_n - \lambda_1$ is unbounded and a deviation small in the weighted L^2 norm may have a large derivative. □

Proposition 16.79 (The Ritz linear system)**.** *Let $\chi_1, \dots, \chi_N$ be linearly independent admissible trial functions and minimise $\mathcal{R}$ over $y = \sum_{k=1}^{N} c_k \chi_k$. Then the stationary values of $\mathcal{R}$ are the roots λ of the generalized eigenvalue problem*

$$\sum_{k=1}^{N} \left(K_{jk} - \lambda \, M_{jk} \right) c_k = 0 \,, \qquad j = 1, \dots, N \,, \tag{16.72}$$

with the symmetric matrices

$$K_{jk} = \int_a^b \left(p \, \chi_j' \chi_k' + q \, \chi_j \chi_k \right) \mathrm{d}x \,, \qquad M_{jk} = \int_a^b r \, \chi_j \chi_k \, \mathrm{d}x \,, \tag{16.73}$$

and the lowest root satisfies $\lambda_1 \leq \lambda^{(N)}$, decreasing as the trial family is enlarged.
Rests on Equation (16.65) and Corollary 16.78.

Proof. *Derives Proposition 16.79.*

Substituting the expansion into Equation (16.65) gives $\mathcal{R} = (c_j K_{jk} c_k) \,/\, (c_j M_{jk} c_k)$, a ratio of quadratic forms in the N numbers c_k. Stationarity with respect to c_j requires

$$\frac{2 \sum_k K_{jk} c_k}{c_l M_{lm} c_m} - \frac{(c_l K_{lm} c_m)\, 2 \sum_k M_{jk} c_k}{(c_l M_{lm} c_m)^2} = 0 \,,$$

which, after multiplying by $\tfrac{1}{2} c_l M_{lm} c_m$ and recognising $\mathcal{R} = \lambda$, is Equation (16.72). The bound $\lambda_1 \leq \lambda^{(N)}$ is Corollary 16.78 applied to the minimising member of the trial family, and enlarging the family can only lower the minimum because the smaller family is a subset of the larger. □

Remark 16.80 (The method, its origins and its descendants). Rayleigh introduced the quotient to estimate the frequencies of vibrating systems [197]: for a string of length L under tension F (in newtons) and of linear density ρ_ℓ (in kg/m) the data are $p = F$, $q = 0$, $r = \rho_\ell$, and $\mathcal{R}$ has the SI dimension $1/\mathrm{s}^2$, so $\lambda_1 = \omega_1^2$ is the square of the fundamental angular frequency. Ritz turned the estimate into a systematic and convergent scheme by enlarging the trial family [359]; his 1908 spectroscopic paper [348] is a different work and should not be cited for this method. Equation (16.72) is the ancestor of the finite element method, of the linear variational method for molecular orbitals, and of the variational principle of quantum mechanics: replacing $\mathcal{R}$ by $\langle \psi | \mathcal{H} | \psi \rangle \,/\, \langle \psi | \psi \rangle$ gives an upper bound on the ground-state energy, which is the content of Chapter 82. The normal-mode analyses of Chapter 28 are the same computation with K the stiffness matrix and M the mass matrix. The minimax characterisation of the higher eigenvalues, which extends Equation (16.67) to λ_n, belongs with the spectral theory of Chapter 12.

16.5.4 The Plateau problem

The classical existence question of the subject is the one the soap films of Section 16.1.3 pose.

Theorem 16.81 (Douglas; Radó)**.** *Let $\Gamma \subset \mathbb{R}^3$ be a rectifiable Jordan curve. Then there exists a continuous map X from the closed unit disc $\overline{B}$ into $\mathbb{R}^3$, harmonic and conformal on the open disc, restricting to a monotone parametrisation of Γ on the boundary circle, and minimising the area among all such disc-type maps. Its image is a minimal surface spanning Γ.*

Rests on Equation (16.7) *and Theorem 16.70.*

Douglas minimises a boundary-value functional over monotone boundary parametrisations, defeating the non-compactness of the conformal group of the disc by a three-point normalisation, and Radó argues instead through polygonal approximation and the Dirichlet integral. Either proof is monograph length and neither is reproduced here. Appendix A.43 gives the architecture of Douglas' argument with every step this treatise can carry out carried out — the domination of the area by the Dirichlet integral with its equality case, the conformal invariance of that integral, the identity between Douglas' boundary functional and the Dirichlet integral of the harmonic extension, the lower semicontinuity, the three-point normalisation, the Courant–Lebesgue estimate and the compactness it buys, and the computation showing that a harmonic conformal map has vanishing mean curvature — and states as displayed theorems the results it does not prove. *The theorem itself therefore remains quoted*: the appendix reduces it to two named imports, the solvability of the Dirichlet problem for the disc and Douglas' inner-variation proof of conformality, and says so explicitly.

Full derivation: Appendix A.43 (p. 3009).

Derives Theorem 16.81.

Remark 16.82 (What is proved and what is not)*.* Theorem 16.81 was proved independently by Douglas [678] and Radó [667], and Douglas' work was recognised with one of the first two Fields Medals. It is worth being exact about the claim, because the statement is often quoted more strongly than it was proved.

1. The surface produced is of *disc type*. A contour may bound minimal surfaces of higher genus or with several boundary components, and those require separate arguments.

2. The map is a minimiser of area *among disc-type maps*. It need not be an embedding: the image may have self-intersections, and it may carry isolated branch points where the differential drops rank, although interior branch points were later excluded for area-minimising discs.

3. Conformality is not an extra hypothesis but part of the conclusion, and it is what converts the geometric area functional into the Dirichlet integral to which the direct method applies. This is the technical heart of both proofs.

The result is the landmark application of the direct method in a genuinely two-dimensional problem, and the differential-geometric apparatus it needs — first and second fundamental forms, mean curvature — is that of Chapter 13.

16.6 Symmetry and conservation: Noether's theorems

Emmy Noether's 1918 paper [456] contains two theorems, and they answer opposite questions. The first says that a symmetry group of finitely many parameters produces one conserved quantity per parameter. The second says that a symmetry group depending on arbitrary

functions produces no new conserved quantity at all, but instead identities among the field equations themselves. Both are consequences of the first variation, and both are proved here.

16.6.1 Noether's first theorem

Definition 16.83 (One-parameter transformation group; variational symmetry)**.** A one-parameter family of transformations of the variables of Equation (16.11),

$$\bar{x} = x + \varepsilon\,\xi\,(x,y) + O(\varepsilon^2)\,, \qquad \bar{y}^{\,i} = y^i + \varepsilon\,\phi^i\,(x,y) + O(\varepsilon^2)\,, \tag{16.74}$$

with ξ, ϕ^i of class C^1, is a *variational symmetry* of J if for every curve and every subinterval

$$\int_{\bar{x}_1}^{\bar{x}_2} f\left(\bar{x}, \bar{y}, \frac{\mathrm{d}\bar{y}}{\mathrm{d}\bar{x}}\right) \mathrm{d}\bar{x} = \int_{x_1}^{x_2} \left[f\,(x, y, y') + \varepsilon\,\frac{\mathrm{d}\Lambda}{\mathrm{d}x} \right] \mathrm{d}x + O(\varepsilon^2) \tag{16.75}$$

for some function $\Lambda\,(x,y)$. When Λ may be taken to vanish the symmetry is *strict*; the general case is the *divergence symmetry* of Bessel-Hagen [472], without which the Galilean boost $\boldsymbol{x} \longmapsto \boldsymbol{x} + \boldsymbol{v}t$ escapes the theorem: it changes a free Lagrangian by the total time derivative of $m\left(\boldsymbol{v}\cdot\boldsymbol{x} - \tfrac{1}{2}\,|\boldsymbol{v}|^2\,t\right)$, and the conserved quantity it then yields is the centre-of-mass theorem.

Rests on Equation (16.11) and Definition 16.15.

Theorem 16.84 (Noether's first theorem, one independent variable)**.** *Let Equation (16.74) be a variational symmetry of J with divergence term Λ. Then along every solution of the Euler–Lagrange system Equation (16.30) the quantity*

$$I = \frac{\partial f}{\partial y'^{i}}\left(\phi^i - y'^{i}\,\xi\right) + f\,\xi - \Lambda \tag{16.76}$$

is constant: $\mathrm{d}I/\mathrm{d}x = 0$.

Rests on Equations (16.30) and (16.75) and Proposition 16.17.

Proof. *Derives Theorem 16.84.*

Write $\bar{y}^i(\bar{x})$ for the transformed curve as a function of the transformed independent variable, and compare it with the original curve *at the same value of* x. From Equation (16.74), $\bar{y}^i(\bar{x}) = y^i(x) + \varepsilon\phi^i$ and $\bar{y}^i(x+\varepsilon\xi) = \bar{y}^i(x) + \varepsilon\xi\,y'^i(x) + O(\varepsilon^2)$, so the *vertical* (evolutionary) variation is

$$\bar{y}^i(x) - y^i(x) = \varepsilon\,\bar{\phi}^i\,, \qquad \bar{\phi}^i = \phi^i - \xi\,y'^i\,. \tag{16.77}$$

Renaming the dummy variable $\bar{x}$ back to x, the left-hand side of Equation (16.75) is $\int_{x_1+\varepsilon\xi_1}^{x_2+\varepsilon\xi_2} f\,(x, \bar{y}(x), \bar{y}'(x))\,\mathrm{d}x$, and moving the limits back costs one boundary term to first order:

$$\int_{\bar{x}_1}^{\bar{x}_2} f\left(\bar{x}, \bar{y}, \frac{\mathrm{d}\bar{y}}{\mathrm{d}\bar{x}}\right) \mathrm{d}\bar{x} = \int_{x_1}^{x_2} f\,(x, \bar{y}, \bar{y}')\,\mathrm{d}x + \varepsilon\,[\xi f]_{x_1}^{x_2} + O(\varepsilon^2)\,.$$

By Proposition 16.17 applied to the vertical variation $\bar{\phi}$, and integrating the $\bar{\phi}'$ term by parts,

$$\int_{x_1}^{x_2} f\,(x, \bar{y}, \bar{y}')\,\mathrm{d}x - \int_{x_1}^{x_2} f\,(x, y, y')\,\mathrm{d}x = \varepsilon \int_{x_1}^{x_2} \left[E_i\,\bar{\phi}^i + \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'^{i}}\,\bar{\phi}^i\right)\right] \mathrm{d}x + O(\varepsilon^2)\,,$$

with E_i the Euler–Lagrange expression of Equation (16.40). Substituting both displays into Equation (16.75), dividing by ε and letting $\varepsilon \to 0$,

$$\int_{x_1}^{x_2} \left\{ E_i\,\bar{\phi}^i + \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{\partial f}{\partial y'^{i}}\,\bar{\phi}^i + f\,\xi - \Lambda\right]\right\} \mathrm{d}x = 0\,. \tag{16.78}$$

This holds for every subinterval $[x_1, x_2] \subseteq [a,b]$ and every curve, so the integrand, which is continuous, vanishes identically. On a solution of Equation (16.30) the first term is zero, leaving $\mathrm{d}I/\mathrm{d}x = 0$ with I as in Equation (16.76), since $\bar{\phi}^i = \phi^i - y'^i\xi$. □

Remark 16.85 (The identity is stronger than the theorem). Equation (16.78) holds without any equation of motion: it says that for a variational symmetry the combination $E_i\bar{\phi}^i$ is a total derivative, identically in the curve. That off-shell statement is what Section 16.6.2 exploits, and it is also the honest form of the converse: a conserved quantity of the form Equation (16.76) corresponds to a variational symmetry, but a conservation law of the equations of motion that does not arise from a symmetry of the *action* is not covered — Noether's theorem is a statement about Lagrangians, not about differential equations.

Remark 16.86 (What a point transformation cannot reach). Equation (16.74) moves the point (x, y) only: ξ and ϕ^i are functions of (x, y) and never of y', and no divergence term repairs that. The Laplace–Runge–Lenz vector of the Kepler problem is the standard witness. For $f = |\boldsymbol{p}|^2/2m - V$ with $\boldsymbol{p} = m\dot{\boldsymbol{x}}$, the part of Equation (16.76) that is quadratic in $\boldsymbol{p}$ is $-\xi(t, \boldsymbol{x})\,|\boldsymbol{p}|^2/2m$, which is isotropic in $\boldsymbol{p}$. The corresponding part of $A_j = |\boldsymbol{p}|^2 x_j - (\boldsymbol{x}\cdot\boldsymbol{p})\,p_j - mk\,x_j/|\boldsymbol{x}|$ is not: it vanishes when $\boldsymbol{p}$ is parallel to $\boldsymbol{x}$ and equals $|\boldsymbol{p}|^2 x_j$ when the two are perpendicular. No choice of ξ reproduces an anisotropy, so $\boldsymbol{A}$ is the Noether charge of no transformation of the form Equation (16.74), divergence term or not. The generator that does produce it, $\delta x_i = \varepsilon_j\,(2\dot{x}_i x_j - x_i\dot{x}_j - \delta_{ij}\,\boldsymbol{x}\cdot\dot{\boldsymbol{x}})$, depends on the velocities: it is a *generalized* symmetry, outside the hypotheses proved here and belonging to the later extension of Noether's argument to transformations acting on the derivatives as well. The vector itself, and the hidden SO(4) symmetry it generates, are put to work in Chapter 81.

Corollary 16.87 (Energy from time translation)**.** *If f has no explicit dependence on x, then $\xi = 1$, $\phi^i = 0$, $\Lambda = 0$ is a strict variational symmetry and Equation (16.76) gives the conserved quantity*

$$h = y'^i\,\frac{\partial f}{\partial y'^i} - f = const\,, \tag{16.79}$$

which for one dependent variable is the Beltrami identity Equation (16.25).

Rests on Equations (16.25) and (16.76).

Proof. *Derives Corollary 16.87.*

With $\bar{x} = x + \varepsilon$ and $\bar{y} = y$ the integrand is unchanged and the integration measure is unchanged, so Equation (16.75) holds with $\Lambda = 0$ precisely because f carries no explicit x. Then $I = (\partial f/\partial y'^i)\,(-y'^i) + f = -h$. □

Corollary 16.88 (Momentum and angular momentum)**.** *Let y^1, y^2, y^3 be Cartesian coordinates of a particle in ordinary space and $x = t$ the time.*

1. *If f does not depend on y^k for some fixed k, then $\xi = 0$, $\phi^i = \delta^i_k$ is a strict symmetry and $\partial f/\partial y'^k$ is conserved: the momentum conjugate to a translation.*

2. *If f is invariant under rotations about the unit vector $\boldsymbol{n}$, so that $\phi^i = (\boldsymbol{n}\times\boldsymbol{y})^i$ and $\xi = 0$ define a strict symmetry, then $\boldsymbol{n}\cdot(\boldsymbol{y}\times\boldsymbol{p})$ is conserved, with $p_i = \partial f/\partial y'^i$.*

Rests on Equation (16.76) and Proposition 16.28.

Proof. *Derives Corollary 16.88.*

In both cases $\xi = 0$ and $\Lambda = 0$, so Equation (16.76) reduces to $I = (\partial f/\partial y'^i)\,\phi^i = p_i\phi^i$. Case (1) gives $I = p_k$, which is also Proposition 16.28. Case (2) gives $I = \boldsymbol{p}\cdot(\boldsymbol{n}\times\boldsymbol{y}) = \boldsymbol{n}\cdot(\boldsymbol{y}\times\boldsymbol{p})$ by the cyclic invariance of the scalar triple product. □

Remark 16.89 (SI dimensions of a Noether charge). With $x = t$ in seconds and $f = \mathcal{L}$ in joules, $\partial\mathcal{L}/\partial\dot{y}^i$ is a momentum in kg m/s when y^i is a length, and I in Equation (16.76) carries the dimension of action, J s, divided by whatever dimension the group parameter ε has. For a dimensionless ε — an angle, say — I is an action: angular momentum in J s is the familiar case. For a time translation ε is a time and I is an energy in joules; for a spatial

translation ε is a length and I a momentum. The rule is that the product $I\,\varepsilon$ is always an action, which is the statement that the conserved charge generates the symmetry — the observation that becomes the Poisson-bracket structure of Chapter 22 and the commutator structure of Chapter 5.

Theorem 16.90 (Noether's first theorem, several independent variables)**.** *Let $S[\varphi]$ be the field action Equation (16.31) on $\Omega \subset \mathbb{R}^D$ and let*

$$\bar{x}^\mu = x^\mu + \varepsilon\,\xi^\mu(x,\varphi) + O(\varepsilon^2)\,, \qquad \bar{\varphi}^A = \varphi^A + \varepsilon\,\Phi^A(x,\varphi) + O(\varepsilon^2) \tag{16.80}$$

leave S invariant up to the divergence of a vector $\Lambda^\mu(x,\varphi)$, in the sense that the analogue of Equation (16.75) holds on every subregion. Then on every solution of Equation (16.32) the Noether current

$$j^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi^A)}\left(\Phi^A - \xi^\nu\,\partial_\nu \varphi^A\right) + \mathcal{L}\,\xi^\mu - \Lambda^\mu \tag{16.81}$$

is conserved, $\partial_\mu j^\mu = 0$.

Rests on Equations (16.32) and (16.33) and Theorem 16.84.

Proof. *Derives Theorem 16.90.*

The argument of Theorem 16.84 carries over word for word with the divergence theorem (Theorem 7.101) in place of the one-dimensional integration by parts. The vertical variation is $\bar{\Phi}^A = \Phi^A - \xi^\nu \partial_\nu \varphi^A$, obtained exactly as Equation (16.77). Moving the domain $\bar{\Omega}$ back to Ω contributes, to first order, the flux of $\mathcal{L}\,\xi^\mu$ through $\partial\Omega$, that is $\varepsilon \int_\Omega \partial_\mu\,(\mathcal{L}\xi^\mu)\,\mathrm{d}^D x$, because the Jacobian of Equation (16.80) is $1 + \varepsilon\,\partial_\mu \xi^\mu + O(\varepsilon^2)$. The vertical part contributes, by Equation (16.33) before the boundary term is discarded,

$$\varepsilon \int_\Omega \left\{ \left[\frac{\partial \mathcal{L}}{\partial \varphi^A} - \partial_\mu \pi^\mu_A \right] \bar{\Phi}^A + \partial_\mu \left(\pi^\mu_A \bar{\Phi}^A \right) \right\} \mathrm{d}^D x\,, \qquad \pi^\mu_A = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi^A)}\,.$$

Invariance up to $\partial_\mu \Lambda^\mu$ therefore gives, for every subregion and hence pointwise,

$$\left[\frac{\partial \mathcal{L}}{\partial \varphi^A} - \partial_\mu \pi^\mu_A \right] \bar{\Phi}^A + \partial_\mu \left[\pi^\mu_A \bar{\Phi}^A + \mathcal{L}\,\xi^\mu - \Lambda^\mu \right] = 0\,. \tag{16.82}$$

On shell the bracket in the first term vanishes by Equation (16.32), leaving $\partial_\mu j^\mu = 0$ with j^μ as in Equation (16.81). □

Corollary 16.91 (The canonical energy–momentum tensor)**.** *Let $D = 4$ with coordinates $x^\mu = (ct, x, y, z)$ on the spacetime of Chapter 39, and let $\mathcal{L}$ carry no explicit dependence on x^μ. Then $\xi^\mu = a^\mu$ constant, $\Phi^A = 0$, $\Lambda^\mu = 0$ is a strict symmetry, and the four conserved currents assemble into the* canonical energy–momentum tensor

$$T^\mu{}_\nu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi^A)}\,\partial_\nu \varphi^A - \delta^\mu_\nu\,\mathcal{L}\,, \qquad \partial_\mu T^\mu{}_\nu = 0\,. \tag{16.83}$$

Rests on Equations (16.32) and (16.81).

Proof. *Derives Corollary 16.91.*

Substituting $\xi^\mu = a^\mu$ and $\Phi^A = 0$ into Equation (16.81) gives $j^\mu = -a^\nu \left[\pi^\mu_A\,\partial_\nu \varphi^A - \delta^\mu_\nu \mathcal{L} \right] = -a^\nu T^\mu{}_\nu$, and $\partial_\mu j^\mu = 0$ for every constant a^ν forces $\partial_\mu T^\mu{}_\nu = 0$ componentwise. □

Remark 16.92 (Where the currents become physics). $T^0{}_0$ is the energy density in $\mathrm{J/m^3}$ and $T^0{}_k$ the momentum density; the conservation law Equation (16.83) integrates over a spatial region to the statement that energy and momentum leave it only through its boundary. The

canonical tensor is not in general symmetric, and symmetrising it — adding an identically conserved improvement term — is what produces the tensor that couples to gravity in Chapter 44. An *internal* symmetry, $\xi^\mu = 0$ with Φ^A a linear transformation of the fields, gives instead a current whose conserved charge is electric charge, or baryon number, or any other quantum number of that kind; the electric case is worked in Chapter 64, and the conservation laws of the mechanics chapters (Chapter 21) are the one-dimensional specialisation Corollaries 16.87 and 16.88.

16.6.2 Noether's second theorem

Theorem 16.93 (Noether's second theorem)**.** *Suppose the action Equation (16.31) is invariant under a group of transformations depending on m arbitrary functions $\epsilon^r(x)$ of compact support in Ω, acting on the fields as*

$$\delta\varphi^A = R^A{}_r\,(x,\varphi)\,\epsilon^r + R^{A\mu}{}_r\,(x,\varphi)\,\partial_\mu\epsilon^r\,, \tag{16.84}$$

with $R^A{}_r$ and $R^{A\mu}{}_r$ of class C^1. Then the Euler–Lagrange expressions $E_A = \partial\mathcal{L}/\partial\varphi^A - \partial_\mu\pi^\mu_A$ obey the m identities

$$R^A{}_r\,E_A - \partial_\mu\left(R^{A\mu}{}_r\,E_A\right) = 0\,, \qquad r = 1,\dots,m\,, \tag{16.85}$$

identically in the fields — not merely on solutions.

Rests on Equations (16.32) and (16.33) and Theorem 16.90.

Proof. *Derives Theorem 16.93.*

Because the ϵ^r have compact support, all boundary terms in Equation (16.33) vanish and invariance of the action reads

$$0 = \delta S = \int_\Omega E_A\,\delta\varphi^A\,\mathrm{d}^D x = \int_\Omega \left[E_A R^A{}_r\,\epsilon^r + E_A R^{A\mu}{}_r\,\partial_\mu\epsilon^r\right]\mathrm{d}^D x$$

for every choice of the ϵ^r, and for arbitrary field configurations, since the invariance is a property of the action and not of its extremals. Integrating the second term by parts, again with no boundary contribution,

$$0 = \int_\Omega \left[R^A{}_r E_A - \partial_\mu\left(R^{A\mu}{}_r E_A\right)\right]\epsilon^r\,\mathrm{d}^D x\,.$$

The bracket is continuous and the ϵ^r are arbitrary, so Lemma 16.18, in its D-dimensional form used in Theorem 16.36, gives Equation (16.85), one identity for each r. □

Corollary 16.94 (The field equations are not independent)**.** *Under the hypotheses of Theorem 16.93 the n Euler–Lagrange equations $E_A = 0$ satisfy m differential identities, so at most $n-m$ of them are independent. The Cauchy problem is correspondingly underdetermined: a solution stays a solution after any gauge transformation Equation (16.84) with ϵ^r vanishing in the past.*

Rests on Equation (16.85).

Proof. *Derives Corollary 16.94.*

Equation (16.85) is a linear relation among the E_A and their first derivatives holding identically, so the map $\varphi \longmapsto (E_1,\dots,E_n)$ has image in a set of codimension at least m; equivalently, once $n-m$ of the equations and suitable initial data are imposed, the remaining m follow by integration. That a gauge transform of a solution is a solution is immediate from invariance of the action. □

Remark 16.95 (The two worked instances). *Electrodynamics.* The action of the electromagnetic potential is invariant under $\delta A_\mu = \partial_\mu \epsilon$, one arbitrary function, so in Equation (16.84) $R = 0$ and $R^\nu{}_\mu = \delta^\nu_\mu$. Equation (16.85) becomes $\partial_\mu E^\mu = 0$ identically, where E^μ is the Euler–Lagrange expression of A_μ. Since E^μ contains the source term, the field equations are consistent only if the source is itself conserved: charge conservation is not an extra postulate but the integrability condition of the Maxwell equations, which is the reading developed in Chapter 64 [472].

General relativity. The Einstein–Hilbert action is invariant under diffeomorphisms, which in infinitesimal form are $\delta g_{\mu\nu} = \nabla_\mu \epsilon_\nu + \nabla_\nu \epsilon_\mu$: four arbitrary functions, with $R^A{}_r$ built from the connection and $R^{A\mu}{}_r$ from the metric. The Euler–Lagrange expression is proportional to the Einstein tensor, and Equation (16.85) is the contracted Bianchi identity $\nabla_\mu G^{\mu\nu} = 0$, which follows from the second Bianchi identity Equation (13.314) of Proposition 13.157 by two contractions; that proposition states the uncontracted identities and does not perform the contraction. So the identity that makes the Einstein equations consistent with local energy–momentum conservation is not a happy accident of Riemannian geometry but a consequence of general covariance, exactly as Chapter 44 presents it.

In both cases the degeneracy of the Lagrangian — the Hessian $\partial^2 \mathcal{L} / \partial \dot\varphi^A \partial \dot\varphi^B$ is singular, because some field velocities never appear — is the origin of the constraint analysis of Chapter 26.

16.7 The variational principles of physics

16.7.1 Hamilton's principle

Definition 16.96 (Action; Hamilton's principle)**.** For a mechanical system with generalized coordinates $q^1, \dots, q^n$ and Lagrangian $\mathcal{L}(q, \dot q, t)$ in joules, the *action* between two instants is

$$S[q] = \int_{t_1}^{t_2} \mathcal{L}\left(q(t), \dot q(t), t\right) \mathrm{d}t \,, \tag{16.86}$$

of SI dimension J s. *Hamilton's principle* is the statement that the motion between prescribed configurations $q(t_1)$ and $q(t_2)$ makes Equation (16.86) stationary [96].

Rests on Equation (16.11) *and Definition 21.31.*

Theorem 16.97 (Equivalence with Newton's second law)**.** *Let a single particle of mass m move in ordinary space under a force derived from a potential $V(\boldsymbol{x}, t)$ in joules, and take $\mathcal{L} = \frac{1}{2} m |\dot{\boldsymbol{x}}|^2 - V$. Then $\boldsymbol{x}(t)$ makes Equation* (16.86) *stationary among curves with the prescribed endpoints if and only if it satisfies*

$$m\, \ddot{\boldsymbol{x}}(t) = -\nabla V\left(\boldsymbol{x}(t), t\right) . \tag{16.87}$$

Rests on Equations (16.30) *and* (16.86)*.*

Proof. *Derives Theorem 16.97.*

The three Euler–Lagrange equations Equation (16.30) with $x = t$ and $y^i = x^i$ read

$$\frac{\partial \mathcal{L}}{\partial x^i} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot x^i}\right) = -\frac{\partial V}{\partial x^i} - \frac{\mathrm{d}}{\mathrm{d}t}\left(m \dot x^i\right) = 0 \,,$$

which is Equation (16.87) componentwise. Conversely a solution of Equation (16.87) satisfies the Euler–Lagrange system, hence makes the first variation vanish by Proposition 16.17 and the integration by parts of Theorem 16.22. □

Remark 16.98 (Stationary, not least). The name "principle of least action" is a historical accident. $\mathcal{L} = T - V$ is convex in the velocities, so Equation (16.47) holds strictly and the

second variation is positive definite over short enough intervals; but by Theorem 16.55 the motion ceases to minimise past the first conjugate point — the *kinetic focus* — and for a harmonic oscillator of angular frequency ω that happens after a time π/ω, half a period. The action is then a saddle point, never a maximum. This is why Definition 16.96 says stationary. The standard mechanics treatments [1585, 2027] develop the principle in the form used throughout Chapter 21, and the path-integral formulation of Chapter 106 is what eventually explains why the stationary path is the one that matters: neighbouring paths interfere destructively except near a stationary point of S.

16.7.2 Maupertuis' principle

Hamilton's principle fixes the two *times* and varies the path. The older principle fixes the *energy* instead, lets the transit time float, and varies only the geometric path.

Definition 16.99 (Abbreviated action)**.** For a system with momenta $p_i = \partial \mathcal{L}/\partial \dot{q}^i$, the *abbreviated action* of a path traversed between two configurations is

$$W[\gamma] = \int_\gamma p_i \, \mathrm{d}q^i = \int_{t_1}^{t_2} p_i \, \dot{q}^i \, \mathrm{d}t \,, \tag{16.88}$$

also of dimension J s.

Rests on Equation (16.86) and Definition 16.96.

Lemma 16.100 (The ray equation)**.** *Let n be a positive C^1 function on a region of ordinary space and consider the parametrisation-invariant functional*

$$\mathcal{W}[\gamma] = \int_\gamma n\,(\boldsymbol{x}) \, \mathrm{d}s = \int_{\sigma_1}^{\sigma_2} n\,(\boldsymbol{x}(\sigma)) \, |\boldsymbol{x}'(\sigma)| \, \mathrm{d}\sigma \,. \tag{16.89}$$

Its extremals, parametrised by arclength s so that $\boldsymbol{T} = \mathrm{d}\boldsymbol{x}/\mathrm{d}s$ is a unit vector, satisfy

$$\frac{\mathrm{d}}{\mathrm{d}s}\,(n\,\boldsymbol{T}) = \nabla n \,. \tag{16.90}$$

Rests on Equations (16.30) and (16.89).

Proof. *Derives Lemma 16.100.*

The integrand $f = n\,(\boldsymbol{x})\,|\boldsymbol{x}'|$ is homogeneous of degree one in $\boldsymbol{x}'$, so the functional does not depend on the parametrisation and the three Euler–Lagrange equations Equation (16.30) are not independent; the arclength gauge $|\boldsymbol{x}'| = 1$ fixes the redundancy. With

$$\frac{\partial f}{\partial x'^i} = n\,\frac{x'^i}{|\boldsymbol{x}'|} \,, \qquad \frac{\partial f}{\partial x^i} = \frac{\partial n}{\partial x^i}\,|\boldsymbol{x}'| \,,$$

setting $|\boldsymbol{x}'| = 1$ after the differentiation turns Equation (16.30) into $\mathrm{d}\,(n\,T^i)\,/\mathrm{d}s = \partial n/\partial x^i$, which is Equation (16.90). □

Theorem 16.101 (Maupertuis–Euler–Jacobi principle)**.** *Let a single particle of mass m move in a time-independent potential $V\,(\boldsymbol{x})$ with total energy $E > V\,(\boldsymbol{x})$ throughout the region considered. Then the unparametrised orbits of energy E are exactly the extremals of the geometric functional*

$$W[\gamma] = \int_\gamma \sqrt{2m\,(E - V\,(\boldsymbol{x}))}\,\mathrm{d}s \,, \tag{16.91}$$

that is, the geodesics of the Jacobi metric $\mathrm{d}\sigma^2 = 2m\,(E - V)\,\mathrm{d}s^2$; and the time parametrisation is recovered from $\mathrm{d}t = \sqrt{m/\,[2\,(E - V)]}\,\mathrm{d}s$.

Rests on Equations (16.87) and (16.88) and Lemma 16.100.

Proof. *Derives Theorem 16.101.*

Along a motion of energy E the speed is $v = |\dot{\boldsymbol{x}}| = \sqrt{2\,(E-V)\,/m}$, so the momentum has magnitude $|\boldsymbol{p}| = mv = \sqrt{2m\,(E-V)}$ and points along $\boldsymbol{T}$. Hence $p_i\,\mathrm{d}q^i = |\boldsymbol{p}|\,\mathrm{d}s$ and Equation (16.88) coincides with Equation (16.91); this is Equation (16.89) with $n = \sqrt{2m\,(E-V)}$, in kg m/s.

By Lemma 16.100 the extremals satisfy $\mathrm{d}\,(n\boldsymbol{T})\,/\mathrm{d}s = \nabla n$. Now $n = |\boldsymbol{p}| = mv$, so $n\boldsymbol{T} = \boldsymbol{p}$, and

$$\nabla n = \nabla\sqrt{2m\,(E-V)} = \frac{-m\,\nabla V}{\sqrt{2m\,(E-V)}} = -\frac{\nabla V}{v}\,.$$

Since $\mathrm{d}/\mathrm{d}s = v^{-1}\mathrm{d}/\mathrm{d}t$ along the motion, the ray equation reads $v^{-1}\mathrm{d}\boldsymbol{p}/\mathrm{d}t = -v^{-1}\nabla V$, that is Equation (16.87). Every step is reversible, so the extremals of Equation (16.91) traversed with the stated time law are precisely the Newtonian orbits of energy E. □

Remark 16.102 (The relation to Hamilton's principle, and to whom the principle is due). The two principles are Legendre transforms of one another. For a motion of constant energy E,

$$S = \int_{t_1}^{t_2} \mathcal{L}\,\mathrm{d}t = \int_{t_1}^{t_2} \left(p_i\dot{q}^i - E\right)\mathrm{d}t = W - E\,(t_2 - t_1)\,, \tag{16.92}$$

so making S stationary at fixed transit time and making W stationary at fixed energy are the same condition read with different quantities held fixed. The hypotheses must be stated exactly or the equivalence fails: Hamilton's principle varies the path with t_1, t_2 fixed and the energy free, Maupertuis' principle varies the path with E fixed and the transit time free, and E must exceed V so that $n > 0$ and the region is accessible. At a turning point, where $E = V$, the Jacobi metric degenerates and the ray equation fails.

On the attribution: Maupertuis stated a principle of least action in 1744 for the refraction of light and in 1746 for mechanics, but as a metaphysical assertion about the economy of nature, with a quantity whose definition varied from case to case [24]. The mathematically correct statement, with the hypothesis of fixed energy and the integral $\int mv\,\mathrm{d}s$, is Euler's, in the second addendum to his founding monograph of the same period [23]. Attributing the sharp principle to Maupertuis alone is the customary usage and is wrong; the name is kept here because it is the one every reader will look up, and the correction is recorded rather than the reference silently changed. The geometric reading, that orbits are geodesics of a conformally rescaled metric, is Jacobi's, and it is taken up again with Hamilton–Jacobi theory in Chapter 23.

16.7.3 Fermat's principle

Definition 16.103 (Optical path length; Fermat's principle)**.** In a medium of refractive index $n\,(\boldsymbol{x})$ — dimensionless, equal to c/v_{phase} with $c = 299\,792\,458$ m/s the speed of light in vacuum — the *optical path length* of a curve is $\mathcal{W}[\gamma] = \int_\gamma n\,\mathrm{d}s$, in metres, and the transit time is $\mathcal{W}/c$ in seconds. *Fermat's principle* is the statement that a light ray follows a path making $\mathcal{W}$ stationary [260].

Rests on Equation (16.89) and Lemma 16.100.

Theorem 16.104 (Snell's law of refraction)**.** *Let two homogeneous media of indices n_1 and n_2 meet at a plane interface. Then a ray crossing the interface obeys*

$$n_1 \sin\vartheta_1 = n_2 \sin\vartheta_2\,, \tag{16.93}$$

with ϑ_1, ϑ_2 the angles between the ray and the normal to the interface on the two sides, the incident ray, the refracted ray and the normal being coplanar.

Rests on Definition 16.103 and Lemma 7.33.

Proof. *Derives Theorem 16.104.*

Within each homogeneous medium $\nabla n = 0$, so Equation (16.90) gives $\mathrm{d}\boldsymbol{T}/\mathrm{d}s = \boldsymbol{0}$: the ray is straight. The path is therefore determined by the single point at which it crosses the interface, and coplanarity follows because moving that point off the plane containing the endpoints and the normal lengthens both straight segments.

Place the interface at $z = 0$, the source at perpendicular distance a on one side and the target at perpendicular distance b on the other, their feet separated by d, and let the crossing point be at distance u from the first foot. The optical length is

$$\mathcal{W}(u) = n_1\sqrt{a^2 + u^2} + n_2\sqrt{b^2 + (d-u)^2}\,,$$

a smooth function of the single real variable u, whose stationary points satisfy, by Lemma 7.33,

$$\mathcal{W}'(u) = \frac{n_1 u}{\sqrt{a^2 + u^2}} - \frac{n_2\,(d-u)}{\sqrt{b^2 + (d-u)^2}} = 0\,.$$

The two fractions are exactly $\sin\vartheta_1$ and $\sin\vartheta_2$, so the condition is Equation (16.93). Since $\mathcal{W}'' > 0$ everywhere, the stationary point is the unique minimum. □

Proposition 16.105 (The optical invariant in a stratified medium)**.** *Let $n = n(z)$ depend on height alone. Then along a ray confined to a vertical plane,*

$$n(z)\,\sin\vartheta(z) = \mathit{const}\,, \tag{16.94}$$

with ϑ the angle to the vertical.

Rests on Theorem 16.29 and Equation (16.89).

Proof. *Derives Proposition 16.105.*

Parametrise the ray in the vertical plane as $x \longmapsto z(x)$, so that $\mathcal{W} = \int f\,\mathrm{d}x$ with $f\,(z, z') = n(z)\sqrt{1 + z'^2}$. The integrand carries no explicit dependence on x, so the Beltrami identity Equation (16.25) applies:

$$f - z'\,\frac{\partial f}{\partial z'} = n(z)\sqrt{1+z'^2} - \frac{n(z)\,z'^2}{\sqrt{1+z'^2}} = \frac{n(z)}{\sqrt{1+z'^2}} = \text{const}\,.$$

The unit tangent of the ray is $(1, z')\,/\sqrt{1+z'^2}$, so the angle ϑ it makes with the vertical satisfies $\sin\vartheta = \left(1 + z'^2\right)^{-1/2}$, and the conserved quantity is $n(z)\sin\vartheta$, which is Equation (16.94). □

Remark 16.106 (What the principle explains and where it comes from)*.* Equation (16.94) is the whole content of atmospheric refraction: warm air near a hot road has a smaller index, so $\sin\vartheta$ must grow and a ray descending towards the road bends back upwards — the inferior mirage. The same identity governs gradient-index fibres and the apparent flattening of the setting Sun. Fermat's principle is not, however, an independent postulate. It is the short-wavelength limit of wave optics: substituting $\boldsymbol{E} = \boldsymbol{E}_0 \exp\left[\mathrm{i}k_0 L(\boldsymbol{x})\right]$ into the wave equation and dropping terms of order k_0^{-1} gives the eikonal equation $|\nabla L|^2 = n^2$, whose characteristics are the rays of Equation (16.90); the derivation is that of Born and Wolf [1967] and belongs to Chapter 61. Comparing Lemma 16.100 as used in Theorem 16.101 with its use here shows the optical–mechanical analogy in its sharpest form: a mechanical orbit of energy E is a light ray in a medium of index $n = \sqrt{2m\,(E - V)}$, and the eikonal L is Hamilton's characteristic function. That correspondence, pursued in Chapter 23, is the historical route to the Schrödinger equation.

16.7.4 Field actions

Definition 16.107 (Field action)**.** For fields φ^A on the spacetime of Chapter 39 the action is

$$S[\varphi] = \int_{t_1}^{t_2} \mathrm{d}t \int_V \mathcal{L}\left(\varphi^A, \partial_\mu \varphi^A, x\right) \mathrm{d}^3x \,, \tag{16.95}$$

with the Lagrangian density $\mathcal{L}$ in $\mathrm{J/m^3}$, so that S carries the dimension J s of Equation (16.86). In covariant notation $\mathrm{d}^4x = c\,\mathrm{d}t\,\mathrm{d}^3x$, so $S = c^{-1}\int \mathcal{L}\,\mathrm{d}^4x$. Its Euler–Lagrange equations are Equation (16.32).

Rests on Equations (16.31) *and* (16.86).

Example 16.108 (The electromagnetic action)**.** With the electric field $\boldsymbol{E}$ in V/m, the magnetic field $\boldsymbol{B}$ in teslas, the charge density ρ in $\mathrm{C/m^3}$ and the current density $\boldsymbol{J}$ in $\mathrm{A/m^2}$, the Lagrangian density

$$\mathcal{L}_{\mathrm{EM}} = \frac{\varepsilon_0}{2}\,|\boldsymbol{E}|^2 - \frac{1}{2\mu_0}\,|\boldsymbol{B}|^2 + \boldsymbol{J}\cdot\boldsymbol{A} - \rho\,\Phi \,, \tag{16.96}$$

with $\varepsilon_0 = 8.854\,187\,812\,8 \times 10^{-12}\,\mathrm{F/m}$ and $\mu_0 = 1.256\,637\,062\,12 \times 10^{-6}\,\mathrm{N/A^2}$, and with $\boldsymbol{E} = -\nabla\Phi - \partial\boldsymbol{A}/\partial t$ and $\boldsymbol{B} = \nabla\times\boldsymbol{A}$, has the dimension $\mathrm{J/m^3}$. Its Euler–Lagrange equations in the potentials $(\Phi, \boldsymbol{A})$ are the two inhomogeneous Maxwell equations, the other two holding identically because the fields are built from potentials. The derivation, the covariant form

$$\mathcal{L}_{\mathrm{EM}} = -\frac{1}{4\mu_0} F_{\mu\nu}F^{\mu\nu} + J^\mu A_\mu \,, \tag{16.97}$$

and the gauge analysis are the business of Chapter 64. The sign of the source term is fixed by the signature: in the $(-,+,+,+)$ convention of Appendix E, with $A^\mu = (\Phi/c, \boldsymbol{A})$ and $J^\mu = (c\rho, \boldsymbol{J})$, one has $A_\mu = (-\Phi/c, \boldsymbol{A})$ and $J^\mu A_\mu = \boldsymbol{J}\cdot\boldsymbol{A} - \rho\,\Phi$, which is exactly the last two terms of Equation (16.96); the opposite sign belongs to the mostly-minus convention, which this book does not use (see Equation (16.99) for the same choice). The $F_{\mu\nu}F^{\mu\nu}$ term is the same in both. What belongs here is only that the Maxwell equations are the Euler–Lagrange equations of a multiple integral, so that everything proved in this chapter — the field equations, the Noether currents, the second theorem's identities — applies to them.

Rests on Equations (16.32) *and* (16.95).

Example 16.109 (The Einstein–Hilbert action)**.** The gravitational action is

$$S_{\mathrm{EH}} = \frac{c^3}{16\pi G}\int_\Omega R\,\sqrt{-g}\,\mathrm{d}^4x \,, \tag{16.98}$$

with $x^0 = ct$, R the Ricci scalar of Chapter 13 in $1/\mathrm{m^2}$, g the determinant of the metric, and $G = 6.674\,30 \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$. The prefactor c^3/G has the SI dimension kg/s, so S_{EH} carries J s as an action must; the frequently seen c^4 goes with a different convention for d^4x. Varying Equation (16.98) with respect to the metric gives the Einstein field equations, and varying the matter action gives the energy–momentum tensor that sources them [431]; both computations, and the boundary term of Remark 16.34 that must be added because the integrand contains second derivatives of the metric, are carried out in Chapter 44. The diffeomorphism invariance of Equation (16.98) is the instance of Theorem 16.93 discussed in Remark 16.95.

Rests on Equation (16.32) *and Theorem 16.93.*

Example 16.110 (Geodesics as extremals of proper time)**.** A free test particle of mass m in a spacetime with metric $g_{\mu\nu}$ extremises

$$S = -mc^2\int \mathrm{d}\tau = -mc\int \sqrt{-g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu} \,, \tag{16.99}$$

in J s, whose Euler–Lagrange equations are the geodesic equations of Chapter 43. The integrand is homogeneous of degree one in $\mathrm{d}x^\mu$, so, exactly as in Lemma 16.100, the functional is parametrisation-invariant and the equations are degenerate until a gauge — here proper time, the analogue of arclength — is fixed. This is the same structure as the abbreviated action of Theorem 16.101: the free relativistic particle is the Maupertuis problem of a metric.

Rests on Equations (16.89) *and* (16.91).

Remark 16.111 (What the variational formulation buys). Three things, and they are the reason this chapter exists.

First, *covariance for free*: by Proposition 16.26 the Euler–Lagrange equations of a scalar action take the same form in every coordinate system, so a theory specified by an action is automatically consistent with whatever symmetry group leaves the action invariant.

Second, *conservation laws on demand*: by Theorem 16.90 every continuous symmetry of the action delivers a conserved current, computed from the Lagrangian by differentiation alone. Energy, momentum, angular momentum and electric charge are all obtained this way, and no separate postulate is needed for any of them — which is why Section 16.6.1 is the result the rest of this book leans on hardest.

Third, *a criterion for admissible theories*: the field equations of a Lagrangian theory are not arbitrary. They are second order when the Lagrangian is first order in derivatives, by Equation (16.32); they are unstable if it is genuinely higher order, by Proposition 16.39; and they are redundant, with constraints, precisely when a gauge group acts, by Corollary 16.94. Those three statements between them shape every fundamental theory in this treatise.

Chapter 17

Fourier Analysis and Integral Transforms

Fourier series and transforms, Laplace transforms, convolution and Green's-function methods, with the convergence theorems stated and proved at the depth cap of Part II. The organising idea is a change of basis: a function of time or space is re-expressed as a superposition of harmonics, in which differentiation becomes multiplication and the linear differential equations of physics become algebraic. Fourier introduced the expansion to solve the heat equation [72], and the analysis, synthesis and convergence of such expansions — settled only in 1966 for square-integrable functions [1327] — created much of modern analysis on the way.

The chapter feeds the separation-of-variables and Green's-function methods of Chapter 10, the orthonormal-basis language of Chapter 12, the normal-mode analysis of Chapter 28, and, through the uncertainty relation between a function and its transform, the wave mechanics of Chapter 77 and Chapter 71. Its computational face — sampling and the fast Fourier transform — is what turns the strain record of a gravitational-wave interferometer or the interferogram of a spectrometer into a spectrum, so the experimental chapters consume it too.

The Sturm–Liouville chapter already exhibits the Fourier series and the Fourier integral as the eigenfunction expansion of the one-dimensional Helmholtz operator (Section 9.7.4, Section 9.7.5) and states, without derivation, the orthonormality, the passage to the integral, the derivative rule, the convolution property and the Parseval relation. All five are proved below, in the same normalisation.

17.1 Fourier series

17.1.1 Trigonometric expansion of periodic functions

The subject began in a quarrel about a string. In 1747 d'Alembert solved the equation of a stretched string of length ℓ fixed at both ends and found its general motion to be a superposition of two waves travelling in opposite directions at the speed $c = \sqrt{\tau/\mu}$ — with the tension τ in N and the linear density μ in kg/m, so that c is in m/s [25]. Six years later Daniel Bernoulli asserted that the same motion is a superposition of the normal modes,

$$y(x,t) = \sum_{n=1}^{\infty} A_n \sin\frac{n\pi x}{\ell}\,\cos\left(\frac{n\pi c\,t}{\ell} + \varphi_n\right), \tag{17.1}$$

and hence — setting $t = 0$ — that *any* admissible initial shape of the string is a sine series [28]. Euler and d'Alembert objected that the right-hand side of Equation (17.1) is built from analytic functions and could not possibly represent the arbitrary, kinked initial profiles

the physical problem allows. The objection was reasonable and wrong, and it was settled in Bernoulli's favour by Fourier, who used precisely such expansions to solve the heat equation and computed the coefficients by the formulas below [72]. What the objectors were right about is that the class of functions represented, and the sense in which the series represents them, both need proof; supplying it occupied a century and is the content of Section 17.1.2.

Definition 17.1 (Fourier coefficients and Fourier series)**.** Let f be complex valued, periodic with period $T > 0$, and absolutely integrable over one period. Write

$$\omega_1 = \frac{2\pi}{T}\,, \qquad \nu_1 = \frac{1}{T} \tag{17.2}$$

for the fundamental angular frequency and the fundamental frequency; if t is a time in s then ω_1 is in rad/s and ν_1 in Hz. The *Fourier coefficients* of f are

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\,\mathrm{e}^{-\mathrm{i}n\omega_1 t}\,\mathrm{d}t\,, \qquad n \in \mathbb{Z}\,, \tag{17.3}$$

and the *Fourier series* of f is the formal sum

$$\sum_{n=-\infty}^{\infty} c_n\,\mathrm{e}^{\mathrm{i}n\omega_1 t}\,, \tag{17.4}$$

whose symmetric partial sums are written $S_N f(t) = \sum_{n=-N}^{N} c_n \mathrm{e}^{\mathrm{i}n\omega_1 t}$.

Equation (17.3) is the *analysis* step and is defined for every absolutely integrable f, whatever the series does; Equation (17.4) is the *synthesis* step, and whether it returns f is the convergence problem. The two are kept separate throughout: it is the failure to separate them that made the eighteenth-century controversy irresolvable.

Lemma 17.2 (Orthogonality of the harmonics)**.** *For $m, n \in \mathbb{Z}$,*

$$\frac{1}{T}\int_{-T/2}^{T/2} \mathrm{e}^{\mathrm{i}m\omega_1 t}\,\overline{\mathrm{e}^{\mathrm{i}n\omega_1 t}}\,\mathrm{d}t = \delta_{mn}\,. \tag{17.5}$$

Rests on Definition 17.1.

Proof. *Derives Lemma 17.2.*

The integrand is $\mathrm{e}^{\mathrm{i}(m-n)\omega_1 t}$. For $m = n$ it is 1 and the average is 1. For $m \neq n$ put $p = m - n \neq 0$; the antiderivative is $\mathrm{e}^{\mathrm{i}p\omega_1 t}/(\mathrm{i}p\omega_1)$, so the integral equals

$$\frac{\mathrm{e}^{\mathrm{i}p\pi} - \mathrm{e}^{-\mathrm{i}p\pi}}{\mathrm{i}p\omega_1} = \frac{2\mathrm{i}\sin(p\pi)}{\mathrm{i}p\omega_1} = 0\,,$$

using $\omega_1 T/2 = \pi$ and $\sin(p\pi) = 0$ for integer p. □

This one computation is the whole mechanism. It is the special case, for the operator $-\mathrm{d}^2/\mathrm{d}t^2$ on a periodic interval, of the orthogonality of eigenfunctions of a Sturm–Liouville operator belonging to distinct eigenvalues (Section 9.4), and the model for the abstract orthonormal bases of Chapter 12: it is what lets a coefficient be extracted from a sum by a single integral. Equation (17.5) is the relation left as a pending derivation at Equation (9.109).

Theorem 17.3 (Euler–Fourier coefficient formulas)**.** *If a series $\sum_n \gamma_n \mathrm{e}^{\mathrm{i}n\omega_1 t}$ converges uniformly on $[-T/2, T/2]$ to a function f, then f is continuous and periodic, and $\gamma_n = c_n$ for every n, with c_n given by Equation (17.3). In particular a function has at most one uniformly convergent trigonometric expansion.*

Rests on Definition 17.1 and Lemma 17.2.

Proof. *Derives Theorem 17.3.*

A uniform limit of continuous functions is continuous, and each term is T-periodic, so f is. Fix m, multiply the series by $\mathrm{e}^{-\mathrm{i}m\omega_1 t}$ — a bounded factor, so convergence stays uniform — and integrate over one period. Uniform convergence permits the exchange of sum and integral (the tail is bounded by $T\sup_t|\text{tail}|$), giving

$$\int_{-T/2}^{T/2} f(t)\,\mathrm{e}^{-\mathrm{i}m\omega_1 t}\,\mathrm{d}t = \sum_n \gamma_n \int_{-T/2}^{T/2} \mathrm{e}^{\mathrm{i}(n-m)\omega_1 t}\,\mathrm{d}t = T\gamma_m$$

by Lemma 17.2. Divide by T. □

Real form. Splitting the n and $-n$ terms of Equation (17.4) with Euler's formula (Proposition 8.4) gives the real trigonometric form

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos(n\omega_1 t) + b_n \sin(n\omega_1 t)\right] , \tag{17.6}$$

with

$$a_n = \frac{2}{T}\int_{-T/2}^{T/2} f(t)\cos(n\omega_1 t)\,\mathrm{d}t\,, \qquad b_n = \frac{2}{T}\int_{-T/2}^{T/2} f(t)\sin(n\omega_1 t)\,\mathrm{d}t\,, \tag{17.7}$$

and the dictionary between the two sets of coefficients is

$$a_n = c_n + c_{-n}\,, \qquad b_n = \mathrm{i}\,(c_n - c_{-n})\,, \qquad c_{\pm n} = \tfrac{1}{2}\,(a_n \mp \mathrm{i}b_n)\,, \quad n \geq 1\,, \tag{17.8}$$

with $a_0 = 2c_0$ twice the mean value of f. For real f one has $c_{-n} = \overline{c_n}$, so the real form carries the same information in half as many symbols; the complex form is used below because the harmonics are then eigenfunctions of $\mathrm{d}/\mathrm{d}t$, which is the property the whole subject exploits. Equation (17.7) is Equation (9.118) written with $T = 2L$, the factor $2/T$ there appearing as $1/L$.

The orthonormality relations of the real system follow from Equation (17.5) and the product formulas of trigonometry: for $m, n \geq 1$,

$$\frac{2}{T}\int_{-T/2}^{T/2} \cos(m\omega_1 t)\cos(n\omega_1 t)\,\mathrm{d}t = \delta_{mn}\,, \qquad \frac{2}{T}\int_{-T/2}^{T/2} \sin(m\omega_1 t)\sin(n\omega_1 t)\,\mathrm{d}t = \delta_{mn}\,, \tag{17.9}$$

and the mixed integral vanishes for all m, n. Indeed $\cos A\cos B = \frac{1}{2}\left[\cos(A-B) + \cos(A+B)\right]$ turns the first into $\delta_{m,n} + \delta_{m,-n}$ by Lemma 17.2, and the second index coincidence is impossible for positive m, n; the sine relation is the same identity with the opposite sign, and the mixed one is the integral of an odd function over a symmetric interval. This is the derivation left pending at Equations (9.114) to (9.116).

Proposition 17.4 (Least squares and Bessel's inequality)**.** *Let f be square integrable over a period. Among all trigonometric polynomials $P(t) = \sum_{|n|\leq N} \gamma_n \mathrm{e}^{\mathrm{i}n\omega_1 t}$ of degree at most N, the mean-square error $\frac{1}{T}\int_{-T/2}^{T/2} |f - P|^2\,\mathrm{d}t$ is smallest exactly when $\gamma_n = c_n$, that is, when $P = S_N f$. Moreover*

$$\sum_{n=-N}^{N} |c_n|^2 \leq \frac{1}{T}\int_{-T/2}^{T/2} |f(t)|^2\,\mathrm{d}t \qquad \textit{for every } N\,, \tag{17.10}$$

so the coefficients are square summable and $c_n \to 0$ as $|n| \to \infty$.

Rests on Definition 17.1 and Lemma 17.2.

Proof. *Derives Proposition 17.4.*

Write $\langle g, h\rangle = \frac{1}{T}\int_{-T/2}^{T/2} g\bar{h}\,\mathrm{d}t$. By Lemma 17.2 the harmonics are orthonormal for $\langle\cdot,\cdot\rangle$ and $\langle f, \mathrm{e}^{\mathrm{i}n\omega_1 t}\rangle = c_n$. Expand:

$$\langle f-P, f-P\rangle = \langle f,f\rangle - \sum_{|n|\le N}(\gamma_n\overline{c_n} + \overline{\gamma_n}c_n) + \sum_{|n|\le N}|\gamma_n|^2 = \langle f,f\rangle - \sum_{|n|\le N}|c_n|^2 + \sum_{|n|\le N}|\gamma_n - c_n|^2\,,$$

by completing the square term by term. Only the last sum depends on the γ_n, and it is minimal — and zero — iff $\gamma_n = c_n$. Setting $\gamma_n = c_n$ leaves $0 \le \langle f,f\rangle - \sum_{|n|\le N}|c_n|^2$, which is Equation (17.10); the bound is uniform in N, so the series of squares converges and its terms tend to zero. □

The partial sum is thus the orthogonal projection of f onto the span of the first $2N+1$ harmonics — the concrete instance of the projection theorem of Chapter 12. Whether the inequality Equation (17.10) is an equality in the limit is exactly the question of whether the harmonics are a *complete* orthonormal system; Section 17.2.2 proves that they are.

Example 17.5 (Square wave and sawtooth)**.** Let f be the odd square wave of unit amplitude, $f(t) = \operatorname{sgn}(\sin\omega_1 t)$. It is odd, so all a_n vanish, and

$$b_n = \frac{4}{T}\int_0^{T/2}\sin(n\omega_1 t)\,\mathrm{d}t = \frac{2}{n\pi}\left[1-(-1)^n\right] = \begin{cases}\dfrac{4}{n\pi}\,, & n \text{ odd}\,,\\ 0\,, & n \text{ even}\,.\end{cases} \tag{17.11}$$

Only odd harmonics appear, with amplitudes falling as $1/n$. For the 50 Hz square wave delivered by a simple inverter this places power at 50 Hz, 150 Hz, 250 Hz and so on, with relative amplitudes 1, 1/3, 1/5; the third harmonic carries $(1/3)^2 = 11.1\,\%$ of the fundamental's power. The sawtooth $f(t) = 2t/T$ on $(-T/2, T/2)$ is also odd, with $b_n = (-1)^{n+1}\,2/(n\pi)$, again a $1/n$ envelope. In both cases the $1/n$ decay is the signature of a jump discontinuity, as Proposition 17.10 will show, and it is what makes these two examples the arena of Section 17.1.3.

17.1.2 Convergence: from Dirichlet to Carleson

Rescaling $x = \omega_1 t$ turns any period into 2π and changes no convergence statement, so this subsection fixes $T = 2\pi$ and writes $c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\mathrm{e}^{-\mathrm{i}nx}\mathrm{d}x$.

Lemma 17.6 (Dirichlet kernel)**.** *Define* $D_N(u) = \sum_{n=-N}^{N}\mathrm{e}^{\mathrm{i}nu}$. *Then, for* $u \notin 2\pi\mathbb{Z}$,

$$D_N(u) = \frac{\sin\left[\left(N+\frac{1}{2}\right)u\right]}{\sin(u/2)}\,, \tag{17.12}$$

D_N *is even and real,* $D_N(0) = 2N+1$, *and*

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} D_N(u)\,\mathrm{d}u = 1\,, \qquad \frac{1}{2\pi}\int_0^{\pi} D_N(u)\,\mathrm{d}u = \frac{1}{2}\,. \tag{17.13}$$

Moreover the partial sum is the average of f *against this kernel,*

$$S_N f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-u)\,D_N(u)\,\mathrm{d}u\,. \tag{17.14}$$

Rests on Definition 17.1 and Lemma 17.2.

Proof. *Derives Lemma 17.6.*

The sum is geometric with ratio $r = \mathrm{e}^{\mathrm{i}u} \neq 1$: $\sum_{n=-N}^{N} r^n = (r^{-N} - r^{N+1})/(1-r)$. Multiply numerator and denominator by $\mathrm{e}^{-\mathrm{i}u/2}$; the numerator becomes $\mathrm{e}^{-\mathrm{i}(N+1/2)u} - \mathrm{e}^{\mathrm{i}(N+1/2)u} = -2\mathrm{i}\sin[(N+\tfrac{1}{2})u]$ and the denominator $\mathrm{e}^{-\mathrm{i}u/2} - \mathrm{e}^{\mathrm{i}u/2} = -2\mathrm{i}\sin(u/2)$, giving Equation (17.12). Evenness and reality are visible in the closed form; $D_N(0) = 2N+1$ by counting terms in the sum. Integrating the sum term by term and using Lemma 17.2 leaves only $n = 0$, which gives the first identity of Equation (17.13), and the second follows by evenness. For Equation (17.14), insert Equation (17.3) into the partial sum and exchange the finite sum with the integral:

$$S_N f(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(s) \sum_{n=-N}^{N} \mathrm{e}^{\mathrm{i}n(x-s)} \, \mathrm{d}s = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(s) \, D_N(x-s) \, \mathrm{d}s \,,$$

and substitute $u = x - s$, the range returning to $[-\pi, \pi]$ because both f and D_N are 2π-periodic. □

Everything about pointwise convergence is now a statement about the oscillatory kernel Equation (17.12). It has unit average but does *not* have unit absolute average — that failure is the subject of Lemma 17.12 — and it oscillates faster and faster as N grows. The instrument for exploiting fast oscillation is the following lemma.

Lemma 17.7 (Riemann–Lebesgue)**.** *Let g be absolutely integrable on $[a, b]$ and Riemann integrable in the sense of Section 7.7. Then*

$$\lim_{\lambda \to \infty} \int_a^b g(u) \, \sin(\lambda u) \, \mathrm{d}u = 0 \,, \tag{17.15}$$

and likewise with $\cos(\lambda u)$ or $\mathrm{e}^{-\mathrm{i}\lambda u}$.

Rests on Definition 7.39.

Proof. *Derives Lemma 17.7.*

First let $g = \mathbf{1}_{\{u \in [\alpha, \beta]\}}$ be the indicator of a subinterval. Then

$$\left| \int_\alpha^\beta \sin(\lambda u) \, \mathrm{d}u \right| = \left| \frac{\cos(\lambda\alpha) - \cos(\lambda\beta)}{\lambda} \right| \leq \frac{2}{\lambda} \longrightarrow 0 \,,$$

so the claim holds for indicators and, by linearity, for every step function. Now let g be Riemann integrable and $\varepsilon > 0$. By the definition of the Darboux integral (Definition 7.39) there is a partition whose lower step function φ satisfies $\int_a^b |g - \varphi| \, \mathrm{d}u < \varepsilon/2$. Then

$$\left| \int_a^b g \sin(\lambda u) \, \mathrm{d}u \right| \leq \int_a^b |g - \varphi| \, \mathrm{d}u + \left| \int_a^b \varphi \sin(\lambda u) \, \mathrm{d}u \right| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$$

for λ large enough, since the second term tends to zero. As ε was arbitrary, Equation (17.15) follows. On an unbounded interval one first truncates: absolute integrability makes the tail contribute less than $\varepsilon/2$ uniformly in λ. □

Corollary 17.8 (Localisation)**.** *Whether $S_N f(x)$ converges, and to what, depends only on the values of f in an arbitrarily small neighbourhood of x. Two integrable functions that agree on $(x-\delta, x+\delta)$ have Fourier series that converge or diverge together at x, to the same limit when they converge.*

Rests on Lemmas 17.6 and 17.7.

Proof. *Derives Corollary 17.8.*

Let $h = f_1 - f_2$ vanish on $(x-\delta, x+\delta)$. By Equation (17.14),

$$S_N h(x) = \frac{1}{2\pi} \int_{\delta \le |u| \le \pi} \frac{h(x-u)}{\sin(u/2)} \, \sin\left[\left(N + \tfrac{1}{2}\right) u\right] \mathrm{d}u \,,$$

and on the domain of integration $|\sin(u/2)| \ge \sin(\delta/2) > 0$, so $h(x-u)/\sin(u/2)$ is absolutely integrable there. Lemma 17.7 sends the integral to zero. □

That a *global* object — the coefficients Equation (17.3) integrate f over the whole period — should have purely local convergence behaviour is Riemann's localisation principle, and it is the reason the theorem below needs hypotheses only at the point examined.

Theorem 17.9 (Dirichlet)**.** *Let f be 2π-periodic and absolutely integrable, and let x be a point at which the one-sided limits $f(x^{\pm})$ exist and the* Dini condition

$$\int_0^{\delta} \frac{|f(x+u) - f(x^+)| + |f(x-u) - f(x^-)|}{u} \, \mathrm{d}u < \infty \tag{17.16}$$

holds for some $\delta > 0$. Then

$$\lim_{N \to \infty} S_N f(x) = \frac{f(x^+) + f(x^-)}{2} \,. \tag{17.17}$$

In particular this holds at every point of a piecewise continuously differentiable function, where the series converges to f at points of continuity and to the midpoint of the jump elsewhere.

Rests on Lemmas 17.6 and 17.7 and Theorem 7.35.

Proof. *Derives Theorem 17.9.*

Split Equation (17.14) at $u = 0$ and use the evenness of D_N together with the half-period normalisation Equation (17.13):

$$\begin{aligned} S_N f(x) - \frac{f(x^+) + f(x^-)}{2} &= \frac{1}{2\pi} \int_0^{\pi} \left[f(x-u) - f(x^-)\right] D_N(u) \, \mathrm{d}u \\ &\quad + \frac{1}{2\pi} \int_0^{\pi} \left[f(x+u) - f(x^+)\right] D_N(u) \, \mathrm{d}u \,. \end{aligned}$$

Insert Equation (17.12) and define, on $(0, \pi]$,

$$g_{\pm}(u) = \frac{f(x \pm u) - f(x^{\pm})}{\sin(u/2)} \,,$$

so that each term is $\frac{1}{2\pi} \int_0^{\pi} g_{\pm}(u) \sin[(N + \frac{1}{2})u] \, \mathrm{d}u$. It remains to check that $g_{\pm}$ is absolutely integrable, for then Lemma 17.7 sends both terms to zero. Away from the origin this is clear, since $\sin(u/2)$ is bounded below there. Near the origin, $\sin(u/2) \ge u/\pi$ for $u \in (0, \pi]$ (concavity of the sine on $[0, \pi/2]$), so

$$\int_0^{\delta} |g_{\pm}(u)| \, \mathrm{d}u \le \pi \int_0^{\delta} \frac{|f(x \pm u) - f(x^{\pm})|}{u} \, \mathrm{d}u < \infty$$

by the Dini condition Equation (17.16). Finally, if f is piecewise C^1 then $|f(x \pm u) - f(x^{\pm})| \le M u$ near 0 by the mean value theorem (Theorem 7.35) applied on each side, the integrand of Equation (17.16) is bounded, and the condition holds automatically. □

Dirichlet's own hypothesis [85] was that f be bounded with finitely many maxima, minima and discontinuities in a period — a condition of *bounded variation* rather than of differentiability, which the proof above accommodates with the same estimate. His paper is the first rigorous convergence theorem in analysis, and the first place where the modern habit of stating hypotheses precisely enough to be checked appears in the subject.

Proposition 17.10 (Smoothness and coefficient decay)**.** *Let f be 2π-periodic and $k-1$ times continuously differentiable, with $f^{(k)}$ piecewise continuous. Then*

$$c_n\Big(f^{(k)}\Big) = (\mathrm{i}n)^k\, c_n(f)\,, \tag{17.18}$$

and consequently $|c_n| = o\Big(|n|^{-k}\Big)$. Conversely, a jump discontinuity in f itself forces c_n to decay no faster than $1/|n|$, as in Example 17.5.

Rests on Definition 17.1, Lemma 17.7, and Equation (7.28).

Proof. *Derives Proposition 17.10.*

Integrate by parts once (Equation (7.28)):

$$c_n(f') = \frac{1}{2\pi}\int_{-\pi}^{\pi} f'(x)\mathrm{e}^{-\mathrm{i}nx}\mathrm{d}x = \frac{1}{2\pi}\Big[f\mathrm{e}^{-\mathrm{i}nx}\Big]_{-\pi}^{\pi} + \frac{\mathrm{i}n}{2\pi}\int_{-\pi}^{\pi} f(x)\mathrm{e}^{-\mathrm{i}nx}\mathrm{d}x = \mathrm{i}n\, c_n(f)\,,$$

the boundary term vanishing because f and $\mathrm{e}^{-\mathrm{i}nx}$ are both 2π-periodic. Iterating k times gives Equation (17.18). Since $f^{(k)}$ is integrable, its own coefficients tend to zero by Lemma 17.7, whence $|n|^k\,|c_n(f)| \to 0$. For the converse, suppose $|c_n| \le C\,|n|^{-1-\varepsilon}$ for some $\varepsilon > 0$; then $\sum_n |c_n| < \infty$, the series converges uniformly by the Weierstrass M-test, and its sum is continuous. A function with a jump — the square wave of Equation (17.11), whose coefficients are exactly of size $1/n$ — therefore admits no such bound. $\square$

Corollary 17.11 (Uniform convergence for C^1 functions)**.** *If f is 2π-periodic and continuously differentiable, then $\sum_n |c_n| < \infty$ and the Fourier series converges to f absolutely and uniformly.*

Rests on Propositions 17.4 and 17.10 and Theorem 17.9.

Proof. *Derives Corollary 17.11.*

By Cauchy–Schwarz in the sequence space and Equation (17.18),

$$\sum_{n\neq 0}|c_n| = \sum_{n\neq 0}\frac{1}{|n|}\,|n\,c_n| \le \left(\sum_{n\neq 0}\frac{1}{n^2}\right)^{1/2}\left(\sum_{n\neq 0}|c_n(f')|^2\right)^{1/2} \le \frac{\pi}{\sqrt{3}}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}|f'|^2\,\mathrm{d}x\right)^{1/2},$$

finite, using $\sum_{n\ge 1} n^{-2} = \pi^2/6$ and Bessel's inequality Equation (17.10) applied to f'. The Weierstrass M-test then makes the series uniformly convergent, and Theorem 17.9 identifies its sum with f at every point. $\square$

So smoothness is read off the spectrum and conversely: this is the practical content of Fourier analysis for the solution of differential equations, and it is why the transform methods of Section 10.2.3 work at all.

The failures. Dirichlet's theorem needs a hypothesis at the point, and no hypothesis of mere continuity will do. The obstruction is measurable, and it is the growth of the kernel's absolute average.

Lemma 17.12 (Divergence of the Lebesgue constants)**.** *Let $L_N = \frac{1}{2\pi}\int_{-\pi}^{\pi}|D_N(u)|\,\mathrm{d}u$. Then*

$$L_N \ge \frac{4}{\pi^2}\,\ln N \longrightarrow \infty\,. \tag{17.19}$$

Rests on Lemma 17.6.

Proof. *Derives Lemma 17.12.*

By evenness, $L_N = \frac{1}{\pi}\int_0^{\pi} |D_N(u)|\,\mathrm{d}u$, and $|\sin(u/2)| \le u/2$ on $[0,\pi]$ gives $|D_N(u)| \ge 2\left|\sin[(N+\tfrac12)u]\right|/u$. Substituting $v = (N+\tfrac12)u$,

$$L_N \ge \frac{2}{\pi}\int_0^{(N+1/2)\pi} \frac{|\sin v|}{v}\,\mathrm{d}v \ge \frac{2}{\pi}\sum_{j=1}^{N}\frac{1}{j\pi}\int_{(j-1)\pi}^{j\pi}|\sin v|\,\mathrm{d}v = \frac{4}{\pi^2}\sum_{j=1}^{N}\frac{1}{j} \ge \frac{4}{\pi^2}\ln N\,,$$

bounding $1/v \ge 1/(j\pi)$ on the j-th half-period, using $\int_{(j-1)\pi}^{j\pi}|\sin v|\,\mathrm{d}v = 2$, and comparing the harmonic sum with $\int_1^N \mathrm{d}j/j$. □

Theorem 17.13 (du Bois-Reymond)**.** *There exists a continuous 2π-periodic function whose Fourier series diverges at a point [192].*

Rests on Lemmas 17.6 and 17.12.

Derivation. *Derives Theorem 17.13.*

By Equation (17.14), $f \mapsto S_N f(0)$ is a linear functional on the continuous 2π-periodic functions with the supremum norm, and its norm is exactly L_N: it is at most L_N by the triangle inequality, and at least L_N because a continuous f with $\|f\|_\infty \le 1$ can be made to follow $\operatorname{sgn} D_N(-u)$ except on arbitrarily short intervals around the finitely many sign changes. Lemma 17.12 shows this family of functionals is not uniformly bounded. The uniform boundedness principle of functional analysis then supplies an f for which $\sup_N |S_N f(0)| = \infty$. □

The uniform boundedness principle is the one piece of functional analysis the argument above borrows, and neither it nor the Baire category theorem it rests on is developed in this treatise. It can be dispensed with: the rest of this subsection writes down an explicit continuous function whose Fourier series diverges at the origin, so that Theorem 17.13 depends on nothing unproved. Two elementary estimates come first.

Lemma 17.14 (Uniformly bounded sine sums)**.** *For every integer $M \ge 1$ and every real x,*

$$\left|\sum_{k=1}^{M}\frac{\sin(kx)}{k}\right| \le 1 + 2\pi\,. \tag{17.20}$$

The bound is uniform in M and in x, although the terms are not absolutely summable.

Rests on Proposition 8.4.

Proof. *Derives Lemma 17.14.*

The sum is odd and 2π-periodic in x and vanishes at $x = 0$, so it suffices to bound it for $0 < x \le \pi$. Write $T_m = \sum_{k=1}^{m}\sin(kx)$. Multiplying the finite geometric progression $\sum_{k=1}^{m} \mathrm{e}^{\mathrm{i}kx}$ by $\mathrm{e}^{\mathrm{i}x} - 1$ telescopes it to $\mathrm{e}^{\mathrm{i}(m+1)x} - \mathrm{e}^{\mathrm{i}x}$, so for $\mathrm{e}^{\mathrm{i}x} \ne 1$ — the case $\mathrm{e}^{\mathrm{i}x} = 1$ being $x = 0$, already excluded — that progression sums to $\mathrm{e}^{\mathrm{i}x}\left(\mathrm{e}^{\mathrm{i}mx} - 1\right)/\left(\mathrm{e}^{\mathrm{i}x} - 1\right)$. Taking imaginary parts (Proposition 8.4),

$$T_m = \Im\frac{\mathrm{e}^{\mathrm{i}x}\left(\mathrm{e}^{\mathrm{i}mx} - 1\right)}{\mathrm{e}^{\mathrm{i}x} - 1}\,, \qquad \text{so} \qquad |T_m| \le \frac{2}{|\mathrm{e}^{\mathrm{i}x} - 1|} = \frac{1}{\sin(x/2)}\,, \tag{17.21}$$

using $|\mathrm{e}^{\mathrm{i}x} - 1| = 2\sin(x/2)$ on that range. There $\sin(x/2) \ge x/\pi$, by concavity of the sine on $[0,\pi/2]$, so $|T_m| \le \pi/x$ for every m.

Let $K = \lfloor 1/x \rfloor$. For $k \le K$ use $|\sin(kx)| \le kx$, which gives

$$\sum_{k=1}^{\min(K,M)}\frac{|\sin(kx)|}{k} \le Kx \le 1\,.$$

For $k > K$, and assuming $M > K$ since otherwise there is nothing left, summation by parts against $T_k - T_{k-1} = \sin(kx)$ gives

$$\sum_{k=K+1}^{M} \frac{\sin(kx)}{k} = \frac{T_M}{M} - \frac{T_K}{K+1} + \sum_{k=K+1}^{M-1} T_k \left(\frac{1}{k} - \frac{1}{k+1} \right) ,$$

in which the coefficients $1/k - 1/(k+1)$ are positive and telescope to $1/(K+1) - 1/M$. Bounding every $|T_k|$ by π/x, the modulus of the right-hand side is at most

$$\frac{\pi}{x} \left(\frac{1}{M} + \frac{1}{K+1} + \frac{1}{K+1} - \frac{1}{M} \right) = \frac{2\pi}{x\,(K+1)} < 2\pi ,$$

the last step because $K + 1 > 1/x$. Adding the two ranges gives Equation (17.20). □

Lemma 17.15 (A polynomial whose partial sum spikes at the origin)**.** *For integers $1 \le N < n$ put*

$$Q_{N,n}(x) := 2\sin(nx) \sum_{k=1}^{N} \frac{\sin(kx)}{k} = \sum_{k=1}^{N} \frac{\cos\left[(n-k)x\right] - \cos\left[(n+k)x\right]}{k} . \tag{17.22}$$

Then $Q_{N,n}$ is a real 2π-periodic trigonometric polynomial, and every frequency m occurring in it satisfies $n - N \le |m| \le n + N$. Moreover

$$|Q_{N,n}(x)| \le 2\,(1 + 2\pi) , \qquad Q_{N,n}(0) = 0 , \qquad S_n Q_{N,n}(0) = \sum_{k=1}^{N} \frac{1}{k} \ge \ln N , \tag{17.23}$$

the first for every real x.

Rests on Lemma 17.14 and Definition 17.1.

Proof. *Derives Lemma 17.15.*

The second expression in Equation (17.22) is the first, term by term, by $2\sin(nx)\sin(kx) = \cos\left[(n-k)x\right] - \cos\left[(n+k)x\right]$. It writes $Q_{N,n}$ as a combination of $\cos(mx)$ for $m \in \{n-N, \dots, n-1\} \cup \{n+1, \dots, n+N\}$, all positive because $N < n$; since $\cos(mx) = \frac{1}{2}\left(\mathrm{e}^{\mathrm{i}mx} + \mathrm{e}^{-\mathrm{i}mx}\right)$, the frequencies present are $\pm m$ for those m, and each satisfies $n - N \le |m| \le n + N$.

The bound in Equation (17.23) is Equation (17.20) multiplied by $|2\sin(nx)| \le 2$, and $Q_{N,n}(0) = 0$ because every $\sin(kx)$ vanishes there. Finally the symmetric partial sum S_n of Definition 17.1 retains exactly the frequencies with $|m| \le n$: of the terms just listed it keeps those with $m = n - k$ and discards those with $m = n + k$, so

$$S_n Q_{N,n}(x) = \sum_{k=1}^{N} \frac{\cos\left[(n-k)x\right]}{k} ,$$

which at $x = 0$ is the harmonic sum $\sum_{k=1}^{N} 1/k$; comparison with $\int_1^{N+1} \mathrm{d}t/t$ bounds it below by $\ln(N+1) > \ln N$. □

Proposition 17.16 (An explicit continuous function with a divergent Fourier series)**.** *Put $N_j = 2^{j^3}$ and $n_j = 2N_j$ for $j \ge 1$, and*

$$f(x) := \sum_{j=1}^{\infty} \frac{1}{j^2}\, Q_{N_j,n_j}(x) \tag{17.24}$$

with $Q_{N,n}$ as in Equation (17.22). Then f is continuous and 2π-periodic, and

$$S_{n_j} f(0) = \frac{1}{j^2} \sum_{k=1}^{N_j} \frac{1}{k} \ge j \ln 2 \longrightarrow \infty , \tag{17.25}$$

so its Fourier series diverges at the origin. This proves Theorem 17.13 without appeal to the uniform boundedness principle.

Rests on Lemma 17.15 and Definition 17.1.

Proof. *Derives Proposition 17.16.*

Continuity. Each term of Equation (17.24) is a trigonometric polynomial, hence continuous and 2π-periodic, and by Equation (17.23) it is bounded in modulus by $2(1+2\pi)/j^2$. Since $\sum_j j^{-2}$ converges, the Weierstrass M-test makes the convergence uniform, and a uniform limit of continuous periodic functions is continuous and periodic.

The coefficients are the termwise ones. Write f_J for the J-th partial sum of Equation (17.24). For any $m \in \mathbb{Z}$, Equation (17.3) gives

$$|c_m(f) - c_m(f_J)| = \left| \frac{1}{2\pi} \int_{-\pi}^{\pi} [f(x) - f_J(x)]\, \mathrm{e}^{-\mathrm{i}mx}\, \mathrm{d}x \right| \le \sup_x |f - f_J| \longrightarrow 0\,,$$

so $c_m(f) = \sum_j j^{-2} c_m\big(Q_{N_j,n_j}\big)$. No rearrangement is involved, because the next paragraph shows at most one summand is nonzero for each m.

The blocks occupy disjoint frequency bands. By Lemma 17.15 the j-th term occupies the frequencies m with $N_j \le |m| \le 3N_j$, since $n_j - N_j = N_j$ and $n_j + N_j = 3N_j$. For $i > j$,

$$\frac{N_i}{N_j} \ge \frac{N_{j+1}}{N_j} = 2^{(j+1)^3 - j^3} = 2^{3j^2+3j+1} \ge 2^7 > 3\,,$$

so the bands are pairwise disjoint and move upward with j.

The partial sum at the origin. Fix j and apply S_{n_j}, which keeps the frequencies with $|m| \le n_j = 2N_j$. A block $i < j$ lies entirely below the cut, since $3N_i \le 3N_{j-1} < N_j < 2N_j$, and contributes $i^{-2} Q_{N_i,n_i}(0) = 0$ by Equation (17.23). A block $i > j$ lies entirely above it, since $N_i > 3N_j > 2N_j$, and contributes nothing. The block $i = j$ contributes $j^{-2} S_{n_j} Q_{N_j,n_j}(0) = j^{-2} \sum_{k=1}^{N_j} k^{-1}$, again by Equation (17.23), and that sum is at least $\ln N_j = j^3 \ln 2$. This is Equation (17.25). Since $\sup_N |S_N f(0)| = \infty$, the series cannot converge at $x = 0$. □

Remark 17.17 (What the explicit example costs and buys). The soft proof of Theorem 17.13 is shorter and says more: it shows that *most* continuous functions, in the sense of Baire category, have divergent Fourier series somewhere, whereas Proposition 17.16 exhibits one. What the explicit construction buys is independence: the treatise nowhere proves the uniform boundedness principle, and a chapter that leaned on it would be resting a stated theorem on an unstated one. The two arguments share their mechanism — the harmonic sum $\sum_{k \le N} 1/k$ of Equation (17.23) is the same logarithm that Equation (17.19) produces — which is why the size of the Lebesgue constants is the real content of the theorem either way.

Two further results mark the boundary of the subject and are stated without proof, at the depth cap of Part II. Kolmogorov constructed an absolutely integrable function whose Fourier series diverges at *every* point [548] — so integrability alone guarantees nothing pointwise. Carleson proved that for square-integrable functions the series converges at almost every point [1327], which settles the case relevant to physics, since the states of Chapter 12 and the finite-energy signals of Section 17.2.2 are square integrable. The theorem was extended shortly afterwards to L^p for every $p > 1$; the endpoint $p = 1$ is Kolmogorov's counterexample, so the classification is complete.

Remark 17.18 (Two theorems quoted, and not proved). Carleson's theorem [1327] and Kolmogorov's example [548] are the only statements of this chapter that carry no derivation, and the omission is a decision rather than a debt. Carleson's argument, and every later proof of the same theorem, turns on the boundedness of a maximal operator obtained by decomposing the partial-sum operator over dyadic families in frequency and discarding

an exceptional set in time; the estimates that control it are the subject matter of a course in harmonic analysis, none of that machinery appears anywhere in this treatise, and introducing it to serve one statement would unbalance Part II. Kolmogorov's counterexample is a construction of the same explicit kind as Proposition 17.16, but far more delicate: the function must be arranged so that the partial sums are unbounded at every point at once. Both are quoted because they fix the boundary of the subject, and a reader needs to know where it lies. Nothing in this chapter, and nothing in the physical parts that consume it, uses either: the working statements are Theorem 17.9, Corollary 17.11, Theorem 17.22 and the mean-square convergence of Section 17.2.2, each proved in full here.

Remark 17.19 (What physics actually needs). None of the pathology touches the working use of Fourier series. The signals of the experimental chapters are square integrable, so Proposition 17.4 and Section 17.2.2 give convergence in the mean — which is the sense in which energies, spectra and expectation values are computed — and the pieces of them that are piecewise smooth converge pointwise by Theorem 17.9. The counterexamples matter because they say what may *not* be assumed: that a continuous waveform is reconstructed pointwise by its harmonics is false, and any argument that needs it needs a hypothesis.

17.1.3 The Gibbs phenomenon and summability

At a jump, Theorem 17.9 promises convergence to the midpoint at each fixed point, but says nothing uniform. Something specific and permanent goes wrong nearby.

Theorem 17.20 (Overshoot at a jump)**.** *Let $f(x) = \operatorname{sgn}(\sin x)$, the odd square wave of Example 17.5 with $T = 2\pi$, whose jump at the origin has height 2. Then the partial sum $S_{2N+1}f$ attains its first maximum to the right of the origin at $x_N = \pi/(2N+2)$, and*

$$\lim_{N\to\infty} S_{2N+1}f(x_N) = \frac{2}{\pi}\int_0^{\pi} \frac{\sin v}{v}\,\mathrm{d}v = \frac{2}{\pi}\,\mathrm{Si}(\pi) = 1.178979\ldots \tag{17.26}$$

The partial sums therefore overshoot the limiting value 1 by $0.178979\ldots$, that is by $8.949\,\%$ of the jump height, and the overshoot does not diminish with N: it merely migrates toward the discontinuity, since $x_N \to 0$.

Rests on Equation (17.11) and Theorem 7.38.

Derivation. Derives Theorem 17.20.

By Equation (17.11) the partial sum through the $(2N+1)$-th harmonic is

$$S_{2N+1}f(x) = \frac{4}{\pi}\sum_{k=0}^{N}\frac{\sin\left[(2k+1)x\right]}{2k+1}, \qquad \text{so} \qquad \frac{\mathrm{d}}{\mathrm{d}x}S_{2N+1}f(x) = \frac{4}{\pi}\sum_{k=0}^{N}\cos\left[(2k+1)x\right].$$

The finite sum is geometric: $\sum_{k=0}^{N} \mathrm{e}^{\mathrm{i}(2k+1)x} = \mathrm{e}^{\mathrm{i}x}\left(1-\mathrm{e}^{2\mathrm{i}(N+1)x}\right)/\left(1-\mathrm{e}^{2\mathrm{i}x}\right) = \mathrm{e}^{\mathrm{i}(N+1)x}\sin\left[(N+1)x\right]/\sin x$, whose real part is $\sin\left[2(N+1)x\right]/(2\sin x)$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}x}S_{2N+1}f(x) = \frac{2}{\pi}\,\frac{\sin\left[2(N+1)x\right]}{\sin x}. \tag{17.27}$$

For $0 < x < \pi/(2N+2)$ the right-hand side is positive and it changes sign at $x_N = \pi/(2N+2)$, so that point is the first maximum. Integrating Equation (17.27) from 0, where the partial sum vanishes, and substituting $v = 2(N+1)x$,

$$S_{2N+1}f(x_N) = \frac{2}{\pi}\int_0^{x_N}\frac{\sin\left[2(N+1)x\right]}{\sin x}\,\mathrm{d}x = \frac{2}{\pi}\int_0^{\pi}\frac{\sin v}{2(N+1)\sin\left[v/(2(N+1))\right]}\,\mathrm{d}v\,.$$

As $N\to\infty$ the denominator tends to v uniformly on $[0,\pi]$ — indeed $0 \le v - 2(N+1)\sin[v/(2(N+1))] \le v^3/(24(N+1)^2)$ from the alternating Taylor bound for the sine (Theorem 7.38) — so the integrand converges uniformly to $\sin v/v$, and the limit is Equation (17.26). Numerically $\mathrm{Si}(\pi) = 1.851937\ldots$, giving $1.178979\ldots$, and the excess over 1 divided by the jump height 2 is $8.949\,\%$. □

Remark 17.21 (The name is wrong, and the record says so). The overshoot was described, computed and published by Henry Wilbraham in 1848 [124], half a century before the exchange in *Nature* in which Josiah Willard Gibbs — correcting his own earlier letter — gave the value quoted above [283]. The customary name "Gibbs phenomenon" therefore misattributes the discovery. This treatise keeps the customary name, because a reader searching the literature needs it, and records the misattribution here rather than passing it on silently.

The overshoot is not a defect of the coefficients, which are exactly right, but of the *summation method*: the sharp truncation $\sum_{|n|\le N}$ multiplies the spectrum by a rectangular window, and Lemma 17.6 shows this is equivalent to smoothing f with an oscillating, sign-changing kernel. Change the window and the pathology goes away.

Theorem 17.22 (Fejér). *Let $\sigma_N f = \frac{1}{N+1}\sum_{n=0}^{N} S_n f$ be the Cesàro average of the partial sums. Then*

$$\sigma_N f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-u)\,K_N(u)\,\mathrm{d}u\,, \qquad K_N(u) = \frac{1}{N+1}\left(\frac{\sin\left[(N+1)u/2\right]}{\sin(u/2)}\right)^2, \tag{17.28}$$

and if f is continuous and 2π-periodic then $\sigma_N f \to f$ uniformly [310].

Rests on Lemma 17.6 and Theorem 7.25.

Proof. *Derives Theorem 17.22.*

Averaging Equation (17.14) gives Equation (17.28) with $K_N = \frac{1}{N+1}\sum_{n=0}^{N} D_n$. For the closed form multiply by $\sin^2(u/2)$ and use $\sin\left[(n+\frac12)u\right]\sin(u/2) = \frac12\left[\cos(nu) - \cos((n+1)u)\right]$; the sum telescopes to $\frac12\left[1-\cos((N+1)u)\right] = \sin^2\left[(N+1)u/2\right]$, which is the assertion.

The closed form exhibits three properties. (i) $K_N \ge 0$, being a square. (ii) $\frac{1}{2\pi}\int_{-\pi}^{\pi} K_N = 1$, by averaging Equation (17.13). (iii) For $\delta \le |u| \le \pi$, $K_N(u) \le \left[(N+1)\sin^2(\delta/2)\right]^{-1}$, which tends to 0 uniformly on that set. A family with these three properties is an *approximate identity*, and the standard argument now applies: given $\varepsilon > 0$, uniform continuity of f on the compact circle (Theorem 7.25) gives δ with $|f(x-u) - f(x)| < \varepsilon/2$ for $|u| < \delta$, uniformly in x. Then, using (ii) to write $f(x) = \frac{1}{2\pi}\int K_N(u) f(x)\,\mathrm{d}u$,

$$|\sigma_N f(x) - f(x)| \le \frac{1}{2\pi}\int_{|u|<\delta} |f(x-u)-f(x)|\,K_N(u)\mathrm{d}u + \frac{1}{2\pi}\int_{\delta\le|u|\le\pi} 2\,\|f\|_\infty\,K_N(u)\,\mathrm{d}u\,,$$

where (i) was used to drop the absolute value on K_N. The first term is below $\varepsilon/2$ by (ii); the second is below $\varepsilon/2$ for N large by (iii). The bound is uniform in x. □

Corollary 17.23 (Density, uniqueness, and no overshoot). *(a) Every continuous periodic function is a uniform limit of trigonometric polynomials. (b) If all Fourier coefficients of a continuous periodic f vanish, then $f \equiv 0$; hence the coefficients determine the function. (c) $\min f \le \sigma_N f \le \max f$ pointwise, so the Cesàro means exhibit no overshoot whatever at a jump.*

Rests on Theorem 17.22.

Proof. *Derives Corollary 17.23.*

(a) $\sigma_N f$ is a trigonometric polynomial of degree N. (b) If every $c_n = 0$ then every $S_n f = 0$, so $\sigma_N f = 0$, and Theorem 17.22 gives $f = \lim \sigma_N f = 0$. (c) $K_N \ge 0$ with unit average, so Equation (17.28) exhibits $\sigma_N f(x)$ as a weighted average of values of f, which cannot leave the range of f. □

Part (b) is what closes Theorem 17.3 into a genuine one-to-one correspondence, and it is used in Section 17.2.2 to prove completeness. Part (c) is the practical lesson: the overshoot belongs to the rectangular window and not to the data.

Example 17.24 (Ringing in a band-limited reproduction)**.** A 1.0 kHz square wave passed through a system with a sharp cutoff at 20 kHz retains harmonics up to the nineteenth, i.e. $N = 9$ in Theorem 17.20. Its reconstruction still overshoots by 9 % of the step, in a ripple whose first peak sits $x_N/\omega_1 = 1/\left[4(N+1)\nu_1\right] = 25\,\mu\text{s}$ after the edge. Increasing the cutoff moves the ripple closer to the edge and makes it narrower; it never makes it smaller. Tapering the cutoff instead — which is what a window function does, and what Theorem 17.22 does in its extreme form — trades the ripple for a slower edge. The same trade reappears as spectral leakage in Section 17.5.2 and as the choice of apodisation in every Fourier-transform spectrometer.

17.1.4 Series in several variables

Everything above concerns a function of one variable. A field periodic in space, a motion on a torus, a crystal potential: each is periodic with respect to a lattice in $\mathbb{R}^n$, and the expansion it admits is indexed by the reciprocal lattice. Nothing new has to be proved for it, which is exactly the point of stating it.

Definition 17.25 (Multiple Fourier series)**.** Let $\Lambda \subset \mathbb{R}^n$ be a lattice with primitive vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$, let C be a primitive cell of volume $|C|$, and let Λ^* be the reciprocal lattice generated by the $\boldsymbol{b}_j$ with $\boldsymbol{a}_i \cdot \boldsymbol{b}_j = 2\pi\delta_{ij}$. For f square integrable on C and Λ-periodic, the *multiple Fourier coefficients* are

$$c_{\boldsymbol{G}} = \frac{1}{|C|}\int_C f(\boldsymbol{x})\, \mathrm{e}^{-\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{x}}\, \mathrm{d}^n x\,, \qquad \boldsymbol{G} \in \Lambda^*\,, \tag{17.29}$$

and the associated series is $\sum_{\boldsymbol{G}\in\Lambda^*} c_{\boldsymbol{G}}\, \mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{x}}$. If f carries an SI unit, so does every $c_{\boldsymbol{G}}$; the exponent is a pure number because $\boldsymbol{G}$ carries $1/\mathrm{m}$.

Rests on Definition 17.1 and Corollary 17.49.

Proposition 17.26 (The lattice harmonics are an orthonormal basis)**.** *The functions $e_{\boldsymbol{G}}(\boldsymbol{x}) = |C|^{-1/2}\, \mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{x}}$, $\boldsymbol{G}$ running over Λ^*, form an orthonormal basis of $L^2(C)$. Consequently, for Λ-periodic square-integrable f: the coefficients Equation (17.29) determine f up to a null set; the series converges to f in the mean square; and Parseval's identity holds,*

$$\frac{1}{|C|}\int_C |f|^2\, \mathrm{d}^n x = \sum_{\boldsymbol{G}\in\Lambda^*} |c_{\boldsymbol{G}}|^2\,. \tag{17.30}$$

If in addition $\sum_{\boldsymbol{G}} |c_{\boldsymbol{G}}| < \infty$, the series converges absolutely and uniformly, and its sum is continuous and equal to f.

Rests on Definition 17.25 and Theorems 12.30 and 17.22.

Proof. *Derives Proposition 17.26.*

Change variables to the lattice coordinates $\boldsymbol{x} = u^i \boldsymbol{a}_i$ with $u^i \in [0,1)$, in which C is the unit cube and $\boldsymbol{G}\cdot\boldsymbol{x} = 2\pi\,\ell_i u^i$ for $\boldsymbol{G} = \ell_i \boldsymbol{b}_i$ with integer ℓ_i. The substitution is linear, so its Jacobian is the constant $|C|$ and cancels the prefactor of Equation (17.29); no change-of-variables theorem beyond the scaling of a volume by a determinant is used. The integral Equation (17.29) then factorises into n copies of the one-dimensional integral of Lemma 17.2, so that

$$\frac{1}{|C|}\int_C \mathrm{e}^{\mathrm{i}(\boldsymbol{G}-\boldsymbol{G}')\cdot\boldsymbol{x}}\, \mathrm{d}^n x = \prod_{i=1}^{n} \delta_{\ell_i \ell'_i}\,, \tag{17.31}$$

and the family $\{e_{\boldsymbol{G}}\}$ is orthonormal in $L^2(C)$.

For completeness, take the product of n copies of the Fejér kernel of Equation (17.28), one in each coordinate u^i, and let $\sigma_N f$ be the corresponding average of f. The product kernel inherits the three properties established in the proof of Theorem 17.22: it is non-negative, its

average over the cube is 1, and outside a cube of side 2δ it is bounded by a quantity tending to zero, because at least one factor is then evaluated at an argument of modulus at least δ while the others integrate to at most 1. It is therefore an approximate identity, and the argument of that proof, run with the cube in place of the circle and with uniform continuity supplied by Theorem 7.25 on the compact torus, gives $\sigma_N f \to f$ uniformly for continuous Λ-periodic f. Each $\sigma_N f$ is a finite combination of the $e_{\boldsymbol{G}}$, so their span is dense in the continuous periodic functions for the supremum norm and hence, C having finite volume, in $L^2(C)$ (Remark 17.28). So the orthonormal family is total, which by Theorem 12.30 makes it a basis, and that theorem supplies the mean-square convergence and Equation (17.30); the uniqueness statement is Equation (17.30) applied to a difference.

The last claim is the Weierstrass criterion: each term has modulus $|c_{\boldsymbol{G}}|$, so an absolutely convergent sum of the moduli forces uniform convergence of a series of continuous functions. Its sum is therefore continuous, and it agrees with f outside a null set because it also converges in the mean square; if f is itself continuous the two agree at every point. □

Remark 17.27 (Where the several-variable case is used). Three places, and in each the value of Proposition 17.26 is that it costs nothing. A quasiperiodic motion with n incommensurate frequencies is by construction a function on an n-torus evaluated along a straight line, so its time dependence is a multiple Fourier series and its power sits on the countable set of integer combinations of the frequencies — which is what Chapter 32 needs when it distinguishes quasiperiodicity from chaos by the presence of a continuous spectral component. A one-electron potential in a crystal is Λ-periodic, and Equation (17.29) defines its structure factors (Chapter 126). And Corollary 17.49 is the Poisson-summation counterpart of the same statement, relating a lattice sum to a sum over Λ^*.

17.2 The Fourier transform

Remark 17.28 (Analytical inputs quoted, not proved). From here on three standard tools of integration theory are used freely and are *not* developed in this treatise, whose integral is the Riemann–Darboux integral of Section 7.7: the Fubini–Tonelli theorem on exchanging the order of a double integral, the dominated convergence theorem, and the density of smooth rapidly decreasing functions in the absolutely integrable and square-integrable classes. Each is quoted explicitly at the point of use rather than absorbed silently. Everything else in this chapter is proved.

17.2.1 Definition and inversion

Let f be absolutely integrable on the whole line and, for the moment, supported in $[-T/2, T/2]$. Its Fourier series on that interval, with $\omega_1 = 2\pi/T$ and $\omega_n = n\omega_1$, is built from the coefficients Equation (17.3). Introduce

$$\hat f(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)\, \mathrm{e}^{-\mathrm{i}\omega t}\, \mathrm{d}t\,, \tag{17.32}$$

so that $c_n = \sqrt{2\pi}\, \hat f(\omega_n)/T$, and note that $\sqrt{2\pi}/T = \Delta\omega/\sqrt{2\pi}$ with $\Delta\omega = \omega_{n+1} - \omega_n = \omega_1$. The synthesis formula Equation (17.4) then reads

$$f(t) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \hat f(\omega_n)\, \mathrm{e}^{\mathrm{i}\omega_n t}\, \Delta\omega\,, \tag{17.33}$$

which is a Riemann sum for $\frac{1}{\sqrt{2\pi}} \int \hat f(\omega) \mathrm{e}^{\mathrm{i}\omega t} \mathrm{d}\omega$ with mesh $\Delta\omega = 2\pi/T \to 0$ as the interval is enlarged. This is the passage left as a pending derivation at Equation (9.121); as an argument it is only heuristic, because the summand depends on T as well as on the mesh,

and Theorem 17.32 below supplies the proof. The discrete spectrum of a periodic function becomes a continuous spectrum, and the spacing $\Delta\nu = 1/T$ of the harmonics — in Hz for a signal in time — becomes the resolution of the transform, a fact that returns quantitatively in Section 17.5.2.

Definition 17.29 (Fourier transform; the treatise convention)**.** For absolutely integrable f the *Fourier transform* is Equation (17.32), also written $\mathcal{F}f = \hat{f}$, with the *inverse transform*

$$\left(\mathcal{F}^{-1}g\right)(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(\omega)\, \mathrm{e}^{+\mathrm{i}\omega t}\, \mathrm{d}\omega\,. \tag{17.34}$$

The symmetric factor $1/\sqrt{2\pi}$ and the sign $\mathrm{e}^{-\mathrm{i}\omega t}$ in the analysis kernel are fixed for the whole treatise, matching Equation (9.121). For a function of position the pair is written with the wavenumber k in place of ω and $\mathrm{e}^{-\mathrm{i}kx}$ in place of $\mathrm{e}^{-\mathrm{i}\omega t}$; if x is in m then k is in 1/m, and if t is in s then ω is in rad/s and $\nu = \omega/2\pi$ in Hz. The transform carries the SI unit of f multiplied by that of t: the transform of a voltage record in V is in V s.

Other normalisations are in use and none is wrong; what is fatal is mixing them, since the factors of 2π move between the transform, the inverse and the convolution theorem. Table 17.1 is the dictionary.

Table 17.1: The three conventions in common use, and the translation between them. Writing $\hat{f}$ for the symmetric transform of Equation (17.32), F for the asymmetric one and $\tilde{f}$ for the ordinary-frequency one, the dictionary is $\hat{f}(\omega) = F(\omega)/\sqrt{2\pi} = \tilde{f}(\omega/2\pi)/\sqrt{2\pi}$. Only the first and third are unitary, so only they make Plancherel's theorem (Theorem 17.36) free of stray factors.

Convention	Analysis	Synthesis	Used by
Symmetric, angular frequency (this treatise)	$\dfrac{1}{\sqrt{2\pi}}\int f\,\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t$	$\dfrac{1}{\sqrt{2\pi}}\int \hat{f}\,\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}\omega$	mathematical physics
Asymmetric, angular frequency	$\displaystyle\int f\,\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t$	$\dfrac{1}{2\pi}\int F\,\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}\omega$	field theory, solid state
Ordinary frequency	$\displaystyle\int f\,\mathrm{e}^{-2\pi\mathrm{i}\nu t}\mathrm{d}t$	$\displaystyle\int \tilde{f}\,\mathrm{e}^{2\pi\mathrm{i}\nu t}\mathrm{d}\nu$	signal processing

Proposition 17.30 (Elementary properties)**.** *Let f, g be absolutely integrable, $a \in \mathbb{R}$, $\lambda \neq 0$. Then $\hat{f}$ is bounded, with $\left\|\hat{f}\right\|_\infty \le \|f\|_1/\sqrt{2\pi}$, uniformly continuous, and $\hat{f}(\omega) \to 0$ as $|\omega| \to \infty$. Moreover*

$$\widehat{f(\cdot - a)}(\omega) = \mathrm{e}^{-\mathrm{i}\omega a}\,\hat{f}(\omega) \qquad \textit{(translation)}\,, \tag{17.35}$$

$$\widehat{\mathrm{e}^{\mathrm{i}\omega_0 \cdot} f}(\omega) = \hat{f}(\omega - \omega_0) \qquad \textit{(modulation)}\,, \tag{17.36}$$

$$\widehat{f(\lambda\,\cdot)}(\omega) = \frac{1}{|\lambda|}\,\hat{f}\!\left(\frac{\omega}{\lambda}\right) \qquad \textit{(scaling)}\,, \tag{17.37}$$

$$\widehat{f'}(\omega) = \mathrm{i}\omega\,\hat{f}(\omega) \qquad \textit{(differentiation)}\,, \tag{17.38}$$

$$\widehat{tf}(\omega) = \mathrm{i}\,\frac{\mathrm{d}\hat{f}}{\mathrm{d}\omega}(\omega) \qquad \textit{(multiplication by } t)\,. \tag{17.39}$$

the last two under the hypotheses that f', respectively $tf(t)$, is itself absolutely integrable and that f vanishes at infinity.

Rests on Definition 17.29, Lemma 17.7, and Equation (7.28).

Proof. *Derives Proposition 17.30.*

The bound is immediate from $\left|\hat f\right| \le \frac{1}{\sqrt{2\pi}}\int|f|$. Uniform continuity: $\left|\hat f(\omega+h)-\hat f(\omega)\right| \le \frac{1}{\sqrt{2\pi}}\int|f(t)|\left|\mathrm{e}^{-\mathrm{i}ht}-1\right|\mathrm{d}t$, which tends to zero with h by dominated convergence, independently of ω. The decay is Lemma 17.7 extended to the line as noted in its proof. Equation (17.35) and Equation (17.36) are the substitutions $t\mapsto t+a$ and a regrouping of exponentials; Equation (17.37) is the substitution $s=\lambda t$, the modulus arising because the limits reverse when $\lambda<0$. For Equation (17.38), integrate by parts (Equation (7.28)); the boundary term $\left[f\mathrm{e}^{-\mathrm{i}\omega t}\right]_{-\infty}^{\infty}$ vanishes because $f\to 0$. For Equation (17.39), differentiate Equation (17.32) under the integral sign, legitimate by dominated convergence when $tf(t)$ is absolutely integrable, which brings down a factor $-\mathrm{i}t$. □

Equation (17.38) is the reason the transform exists as a tool: it converts $\mathrm{d}/\mathrm{d}t$ into multiplication by $\mathrm{i}\omega$, hence a constant-coefficient linear differential operator into a polynomial, and a linear partial differential equation into an algebraic one in the transformed variable. That is precisely the manoeuvre of Section 10.2.3, and it is the pending derivative rule Equation (9.122), now proved.

Example 17.31 (The three standard pairs)**.** *(i) Rectangle and sinc.* For the rectangular pulse $f=\mathbf{1}_{\{|t|\le\tau/2\}}$ of duration τ,

$$\hat f(\omega)=\frac{1}{\sqrt{2\pi}}\int_{-\tau/2}^{\tau/2}\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}t=\frac{\tau}{\sqrt{2\pi}}\,\operatorname{sinc}\!\left(\frac{\omega\tau}{2}\right),\qquad \operatorname{sinc}x=\frac{\sin x}{x}. \tag{17.40}$$

The first zero sits at $\omega=2\pi/\tau$, i.e. at $\nu=1/\tau$: a pulse of duration $1.0\,\mu\mathrm{s}$ has its first spectral null at $1.0\,\mathrm{MHz}$.

(ii) Exponential and Lorentzian. For the causal decay $f(t)=\mathrm{e}^{-\alpha t}$ for $t>0$ and 0 for $t<0$, with $\alpha>0$ in $1/\mathrm{s}$,

$$\hat f(\omega)=\frac{1}{\sqrt{2\pi}}\,\frac{1}{\alpha+\mathrm{i}\omega},\qquad \left|\hat f(\omega)\right|^2=\frac{1}{2\pi}\,\frac{1}{\alpha^2+\omega^2}, \tag{17.41}$$

a Lorentzian of full width at half maximum $\Delta\omega=2\alpha$, i.e. $\Delta\nu=\alpha/\pi$. This is the natural line shape of a decaying emitter and the origin of the linewidths measured in Chapter 90.

(iii) Gaussian and Gaussian. For $f(t)=\mathrm{e}^{-t^2/(2a^2)}$, Equation (17.38) and Equation (17.39) give $\mathrm{i}\omega\hat f=\widehat{f'}=-a^{-2}\widehat{tf}=-\mathrm{i}a^{-2}\hat f'$, i.e. the first-order equation $\hat f'=-a^2\omega\hat f$ with $\hat f(0)=a$ from the Gauss integral. Hence

$$\widehat{\mathrm{e}^{-t^2/(2a^2)}}(\omega)=a\,\mathrm{e}^{-a^2\omega^2/2}, \tag{17.42}$$

a Gaussian of width $1/a$ in ω for a Gaussian of width a in t. The Gaussian is thus an eigenfunction of the transform up to scaling, and Theorem 17.43 will show it is the unique minimiser of the bandwidth product.

Rests on Definition 17.29 and Proposition 17.30.

Theorem 17.32 (Fourier inversion)**.** *Let f be absolutely integrable and continuous, and suppose $\hat f$ is also absolutely integrable. Then for every t*

$$f(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\hat f(\omega)\,\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega. \tag{17.43}$$

Rests on Definition 17.29 and Equation (17.42).

Proof. *Derives Theorem 17.32.*

Fourier stated the integral theorem in 1822 [72]; the proof below is the Gaussian-regularisation argument of Titchmarsh's monograph [802]. For $\varepsilon > 0$ set

$$I_\varepsilon(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat f(\omega)\, \mathrm{e}^{-\varepsilon\omega^2/2}\, \mathrm{e}^{\mathrm{i}\omega t}\, \mathrm{d}\omega\,.$$

Insert Equation (17.32). The double integral converges absolutely, because $|f(s)|\,\mathrm{e}^{-\varepsilon\omega^2/2}$ is integrable over the (s,ω) plane, so Fubini's theorem permits doing the ω integral first:

$$I_\varepsilon(t) = \int_{-\infty}^{\infty} f(s)\, \left[\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-\varepsilon\omega^2/2}\mathrm{e}^{\mathrm{i}\omega(t-s)}\mathrm{d}\omega\right] \mathrm{d}s = \int_{-\infty}^{\infty} f(s)\, g_\varepsilon(t-s)\, \mathrm{d}s\,,$$

where, by the Gaussian pair Equation (17.42) read with $a^2 = 1/\varepsilon$,

$$g_\varepsilon(u) = \frac{1}{\sqrt{2\pi\varepsilon}}\, \mathrm{e}^{-u^2/(2\varepsilon)}\,, \qquad \int_{-\infty}^{\infty} g_\varepsilon(u)\,\mathrm{d}u = 1\,, \qquad g_\varepsilon \ge 0\,.$$

The family g_ε is an approximate identity exactly as in Theorem 17.22: non-negative, of unit mass, and concentrating on $|u| < \delta$ as $\varepsilon \to 0$, since $\int_{|u|\ge\delta} g_\varepsilon \to 0$. Continuity and boundedness of f therefore give $I_\varepsilon(t) \to f(t)$. On the other hand $\left|\hat f(\omega)\mathrm{e}^{-\varepsilon\omega^2/2}\right| \le \left|\hat f(\omega)\right|$, which is integrable by hypothesis, so dominated convergence gives $I_\varepsilon(t) \to \frac{1}{\sqrt{2\pi}}\int \hat f(\omega)\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}\omega$. Two limits of one sequence, hence Equation (17.43). □

Corollary 17.33 (The transform is injective)**.** *Two continuous absolutely integrable functions with the same transform are equal. Hence a spectrum determines its signal.*
Rests on Theorem 17.32.

Proof. *Derives Corollary 17.33.*

Apply Theorem 17.32 to the difference, whose transform vanishes identically and is trivially integrable. □

17.2.2 Parseval and Plancherel

Theorem 17.34 (Parseval's identity for series)**.** *Let f be T-periodic and square integrable over a period. Then the harmonics are a* complete *orthonormal system: $S_N f \to f$ in mean square, and*

$$\frac{1}{T}\int_{-T/2}^{T/2} |f(t)|^2\,\mathrm{d}t = \sum_{n=-\infty}^{\infty} |c_n|^2 = \frac{|a_0|^2}{4} + \frac{1}{2}\sum_{n=1}^{\infty}\left(|a_n|^2 + |b_n|^2\right)\,. \tag{17.44}$$

More generally $\frac{1}{T}\int f\bar g = \sum_n c_n\overline{d_n}$ for the coefficients c_n of f and d_n of g.
Rests on Lemma 17.2, Proposition 17.4, and Theorem 17.22.

Proof. *Derives Theorem 17.34.*

Suppose first that f is continuous. By Theorem 17.22 the Cesàro means converge uniformly, hence in mean square, and $\sigma_N f$ is a trigonometric polynomial of degree N; by the least-squares property of Proposition 17.4 the partial sum does at least as well,

$$\frac{1}{T}\int |f - S_N f|^2 \le \frac{1}{T}\int |f - \sigma_N f|^2 \le \|f - \sigma_N f\|_\infty^2 \longrightarrow 0\,.$$

Expanding the left-hand side with Lemma 17.2 gives $\frac{1}{T}\int|f|^2 - \sum_{|n|\le N}|c_n|^2 \to 0$, which is Equation (17.44). For general square-integrable f, pick continuous $f_j \to f$ in mean square (density, quoted in Remark 17.28); both sides of Equation (17.44) are continuous in that norm — the right side because $|c_n(f) - c_n(f_j)|$ is controlled by Bessel's inequality Equation (17.10) applied to $f - f_j$ — so the identity passes to the limit. The bilinear form follows by polarisation, replacing f by $f+g$, $f-g$, $f+\mathrm{i}g$, $f-\mathrm{i}g$ in turn. The real form is Equation (17.8) substituted term by term. □

Parseval stated the identity in 1806 for power series, as a formal consequence of multiplying two series and integrating [51]; its reading as the Pythagorean theorem for an orthonormal system — the length of a vector is the root of the sum of the squares of its components — came later, and is the statement completed in the abstract in Chapter 12. It is the relation left pending at Section 9.4.4.

Lemma 17.35 (Multiplication formula)**.** *If f and g are absolutely integrable then*

$$\int_{-\infty}^{\infty} \hat{f}(\omega)\, g(\omega)\, \mathrm{d}\omega = \int_{-\infty}^{\infty} f(t)\, \hat{g}(t)\, \mathrm{d}t\,. \tag{17.45}$$

Rests on Definition 17.29.

Proof. *Derives Lemma 17.35.*

Both sides equal $\frac{1}{\sqrt{2\pi}} \iint f(t) g(\omega) \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t\, \mathrm{d}\omega$, and the double integral converges absolutely since $|f(t)g(\omega)|$ is integrable on the plane; Fubini's theorem exchanges the order. □

Theorem 17.36 (Plancherel)**.** *The Fourier transform, defined by Equation (17.32) on functions that are both absolutely and square integrable, satisfies*

$$\int_{-\infty}^{\infty} |f(t)|^2 \, \mathrm{d}t = \int_{-\infty}^{\infty} \left|\hat{f}(\omega)\right|^2 \, \mathrm{d}\omega\,, \qquad \text{and} \qquad \int f \bar{g}\, \mathrm{d}t = \int \hat{f}\bar{\hat{g}}\, \mathrm{d}\omega\,, \tag{17.46}$$

and extends uniquely to a unitary map of the square-integrable functions onto themselves, whose inverse is the extension of Equation (17.34) [363].

Rests on Definition 17.29 and Theorem 17.32.

Proof. *Derives Theorem 17.36.*

Take first f smooth and rapidly decreasing, so that f and $\hat{f}$ are both absolutely integrable and Theorem 17.32 applies. Then, inserting Equation (17.43) and exchanging the order of integration (Fubini, the double integral converging absolutely),

$$\int |f(t)|^2 \, \mathrm{d}t = \int \overline{f(t)} \left[\frac{1}{\sqrt{2\pi}} \int \hat{f}(\omega) \mathrm{e}^{\mathrm{i}\omega t} \mathrm{d}\omega\right] \mathrm{d}t = \int \hat{f}(\omega) \overline{\left[\frac{1}{\sqrt{2\pi}} \int f(t) \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t\right]}\, \mathrm{d}\omega = \int \left|\hat{f}(\omega)\right|^2 \mathrm{d}\omega\,.$$

So $\mathcal{F}$ is norm preserving on a dense subspace of the square-integrable functions (density quoted in Remark 17.28). A norm-preserving linear map on a dense subspace of a complete space extends uniquely and continuously: if $f_j \to f$ then (f_j) is Cauchy, hence so is $(\hat{f}_j)$ by isometry, hence it converges, and the limit is independent of the approximating sequence because the isometry controls the difference. The extension is again an isometry, and Equation (17.34) extends the same way to a two-sided inverse, since $\mathcal{F}^{-1}\mathcal{F} = \mathrm{id}$ holds on the dense subspace by Theorem 17.32 and passes to the limit. An invertible isometry of a complex inner-product space is unitary, and the polarisation identity turns the norm statement into the inner-product statement. □

The theorem is the relation left pending at Equation (9.123), and it is the single most useful statement in the chapter for physics, because it says that a change to the frequency description conserves the quadratic quantity — energy, probability, power — that the physics is about.

Example 17.37 (Energy spectral density in SI)**.** Let $u(t)$, in V, be the voltage across a resistance R in Ω. The total energy dissipated is

$$E = \int_{-\infty}^{\infty} \frac{u(t)^2}{R}\, \mathrm{d}t = \frac{1}{R} \int_{-\infty}^{\infty} |\hat{u}(\omega)|^2 \, \mathrm{d}\omega = \frac{2}{R} \int_{0}^{\infty} |\hat{u}(\omega)|^2 \, \mathrm{d}\omega\,, \tag{17.47}$$

the last step for real u, where $\hat{u}(-\omega) = \overline{\hat{u}(\omega)}$. Dimensionally $\hat{u}$ is in V s, so $|\hat{u}|^2/R$ is in J s/rad and its integral over rad/s is in J: the quantity $2\,|\hat{u}(\omega)|^2/R$ is the *energy spectral density*, telling how many joules sit in each interval of angular frequency. The one-sided form is why laboratory instruments report positive frequencies only, and the factor 2 is the commonest bookkeeping error in noise budgets.

17.2.3 Tempered distributions

Equation (17.32) needs f to be absolutely integrable, and the functions physics most wants to transform are not: a constant, a plane wave $\mathrm{e}^{\mathrm{i}k_0 x}$, a periodic signal, the Heaviside step. Each has a perfectly definite spectrum in practice — a plane wave is one sharp line — but the integral defining it diverges. Dirac introduced the object that the answer requires, a "function" δ with $\int \delta(x)\varphi(x)\mathrm{d}x = \varphi(0)$, and used it exactly so [646]; Schwartz supplied the theory in which the object is legitimate [971]. The idea is to stop asking what the value of δ is at a point and to define it by what it does to test functions.

Definition 17.38 (Schwartz space and tempered distributions)**.** The *Schwartz space* $\mathcal{S}$ consists of the infinitely differentiable $\varphi : \mathbb{R} \to \mathbb{C}$ such that

$$p_{m,k}(\varphi) = \sup_{t\in\mathbb{R}} |t|^m \left|\varphi^{(k)}(t)\right| < \infty \qquad \text{for all } m, k \geq 0\,, \tag{17.48}$$

that is, functions decreasing faster than any power together with all their derivatives. A *tempered distribution* T is a linear map $\mathcal{S} \to \mathbb{C}$, written $\langle T, \varphi\rangle$, continuous in the sense that $|\langle T,\varphi\rangle| \leq C \sum_{m,k\leq M} p_{m,k}(\varphi)$ for some C and M. Every function of at most polynomial growth defines one by $\langle T_f, \varphi\rangle = \int f\varphi\,\mathrm{d}t$, and $\langle \delta, \varphi\rangle = \varphi(0)$ defines the Dirac distribution.

Proposition 17.39 (The transform preserves $\mathcal{S}$)**.** *If $\varphi \in \mathcal{S}$ then $\hat\varphi \in \mathcal{S}$, and $\mathcal{F}$ is a bijection of $\mathcal{S}$ with inverse Equation (17.34).*

Rests on Definition 17.38, Proposition 17.30, and Theorem 17.32.

Proof. *Derives Proposition 17.39.*

Iterating Equation (17.38) and Equation (17.39), both of which apply because every $t^m\varphi^{(k)}$ is absolutely integrable, gives

$$\omega^k \frac{\mathrm{d}^m\hat\varphi}{\mathrm{d}\omega^m}(\omega) = \frac{(-\mathrm{i})^m}{\mathrm{i}^{\,k}} \left[\widehat{\frac{\mathrm{d}^k}{\mathrm{d}t^k}\,(t^m\varphi)}\right](\omega)\,, \tag{17.49}$$

each application of Equation (17.39) contributing one factor $-\mathrm{i}$ and each application of Equation (17.38) one factor i. The right-hand side is the transform of an absolutely integrable function, hence bounded by Proposition 17.30; so every seminorm Equation (17.48) of $\hat\varphi$ is finite and $\hat\varphi \in \mathcal{S}$. Since φ and $\hat\varphi$ are both absolutely integrable, Theorem 17.32 applies and exhibits Equation (17.34) as a two-sided inverse. □

Definition 17.40 (Transform and derivative of a distribution)**.** For $T \in \mathcal{S}'$ define

$$\left\langle \hat T, \varphi\right\rangle = \langle T, \hat\varphi\rangle\,, \qquad \langle T', \varphi\rangle = -\langle T, \varphi'\rangle\,, \qquad \varphi \in \mathcal{S}\,. \tag{17.50}$$

Both right-hand sides are defined by Proposition 17.39 and by $\varphi' \in \mathcal{S}$, and both are continuous, so $\hat T$ and T' are again tempered distributions.

Rests on Definition 17.38 and Proposition 17.39.

The definition is forced, not chosen: for an ordinary absolutely integrable f, Lemma 17.35 says exactly $\int \hat f\varphi = \int f\hat\varphi$, so Equation (17.50) is the unique extension of the classical transform. The same holds for the derivative, where integration by parts with vanishing boundary terms gives $\int f'\varphi = -\int f\varphi'$.

Proposition 17.41 (The identities physics uses)**.** *In the sense of Definition 17.40,*

$$\hat\delta = \frac{1}{\sqrt{2\pi}}\,, \qquad \hat 1 = \sqrt{2\pi}\,\delta\,, \tag{17.51}$$

$$\widehat{\mathrm{e}^{\mathrm{i}\omega_0 t}} = \sqrt{2\pi}\,\delta(\omega-\omega_0)\,, \qquad \widehat{\cos\omega_0 t} = \sqrt{\frac{\pi}{2}}\,[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)]\,, \tag{17.52}$$

and, as an identity between distributions in $t - t'$,

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\omega(t-t')}\,\mathrm{d}\omega = \delta(t-t')\,. \tag{17.53}$$

Furthermore $\theta' = \delta$ for the Heaviside step θ, and $\widehat{T'} = \mathrm{i}\omega\hat{T}$ for every tempered T.
Rests on Definition 17.40 and Theorems 7.43 and 17.32.

Proof. *Derives Proposition 17.41.*
$\left\langle \hat{\delta}, \varphi \right\rangle = \langle \delta, \hat{\varphi}\rangle = \hat{\varphi}(0) = \frac{1}{\sqrt{2\pi}}\int \varphi(t)\mathrm{d}t$, which is the action of the constant $1/\sqrt{2\pi}$. Next, $\left\langle \hat{1}, \varphi \right\rangle = \int \hat{\varphi}(\omega)\mathrm{d}\omega = \sqrt{2\pi}\,\varphi(0)$ by Equation (17.43) evaluated at $t = 0$, which is the action of $\sqrt{2\pi}\delta$; this is Equation (17.53) rewritten, since the left side of Equation (17.53) is $\frac{1}{\sqrt{2\pi}}\hat{1}$ evaluated on the difference variable. Equation (17.52) follows by applying the modulation rule Equation (17.36), valid verbatim in the dual pairing, to $\hat{1}$, and the cosine case by $\cos = \frac{1}{2}(\mathrm{e}^{\mathrm{i}\omega_0 t} + \mathrm{e}^{-\mathrm{i}\omega_0 t})$. For the step, $\langle \theta', \varphi\rangle = -\int_0^\infty \varphi'(t)\mathrm{d}t = \varphi(0)$ by the fundamental theorem of calculus (Theorem 7.43) and $\varphi(\infty) = 0$. Finally $\left\langle \widehat{T'}, \varphi \right\rangle = \langle T', \hat{\varphi}\rangle = -\langle T, (\hat{\varphi})'\rangle = -\left\langle T, \widehat{-\mathrm{i}t\varphi} \right\rangle = \left\langle \hat{T}, \mathrm{i}\omega\varphi \right\rangle$, using Equation (17.39). □

Equation (17.53) is the completeness relation of the continuous plane-wave basis. It is what makes the Green's-function constructions of Section 10.3 legitimate — the fundamental solution of a constant-coefficient operator $P(\partial_t)$ is $G(t) = \frac{1}{2\pi}\int \mathrm{e}^{\mathrm{i}\omega t}/P(\mathrm{i}\omega)\,\mathrm{d}\omega$ precisely because $P(\partial_t)G = \delta$ reduces to Equation (17.53) — and it is the normalisation of the non-normalisable scattering states of Chapter 86, whose rigorous home is the rigged triple of Section 12.6.3. Distributional derivatives are likewise what allows a solution of a partial differential equation to have a kink or a shock: that is the weak formulation of Section 10.6, and $\theta' = \delta$ is its smallest example.

Proposition 17.42 (Sokhotski–Plemelj)**.** *As tempered distributions,*

$$\lim_{\varepsilon\to 0^+} \frac{1}{x \pm \mathrm{i}\varepsilon} = \mathrm{PV}\,\frac{1}{x} \mp \mathrm{i}\pi\,\delta(x)\,, \tag{17.54}$$

where the principal value is $\left\langle \mathrm{PV}\,\frac{1}{x}, \varphi \right\rangle = \lim_{\eta\to 0^+}\int_{|x|>\eta} \varphi(x)\,\mathrm{d}x/x$.
Rests on Definition 17.38 and Theorem 17.32.

Proof. *Derives Proposition 17.42.*
Separate real and imaginary parts, $\frac{1}{x\pm\mathrm{i}\varepsilon} = \frac{x}{x^2+\varepsilon^2} \mp \mathrm{i}\frac{\varepsilon}{x^2+\varepsilon^2}$. The imaginary part is π times the Poisson kernel $\frac{1}{\pi}\frac{\varepsilon}{x^2+\varepsilon^2}$, which is non-negative, has unit integral — the substitution $x = \varepsilon s$ gives $\frac{1}{\pi}\int \mathrm{d}s/(1+s^2) = 1$ — and concentrates at the origin, hence is an approximate identity and converges to δ as in Theorem 17.32. For the real part, use oddness to write

$$\int \frac{x\,\varphi(x)}{x^2+\varepsilon^2}\,\mathrm{d}x = \int_0^\infty \frac{x\,[\varphi(x) - \varphi(-x)]}{x^2+\varepsilon^2}\,\mathrm{d}x\,,$$

whose integrand is bounded near 0 by $2\,\|\varphi'\|_\infty\, x^2/(x^2+\varepsilon^2) \le 2\,\|\varphi'\|_\infty$, uniformly in ε; dominated convergence lets $\varepsilon \to 0$ inside, giving $\int_0^\infty [\varphi(x) - \varphi(-x)]\,\mathrm{d}x/x$, which is the principal value. □

Equation (17.54) is the identity behind the causal "$+\mathrm{i}\epsilon$" prescriptions of propagator theory and behind the dispersion relations of Section 17.4.2; it is stated here because it is a statement about distributions and nothing else.

17.2.4 The uncertainty principle

A function and its transform cannot both be sharply localised: this is a theorem about integrals, provable in a page, and it acquires physical content only when the transform variable is given a physical meaning. Both halves are given below, in that order, because conflating them is the commonest confusion in the subject.

Theorem 17.43 (Bandwidth theorem)**.** *Let f be square integrable with $\|f\|_2^2 = \int |f|^2 \,\mathrm{d}t = 1$, and suppose $tf(t)$ and f' are square integrable. Define the centroids and dispersions*

$$\langle t \rangle = \int t\,|f(t)|^2 \,\mathrm{d}t\,, \quad \sigma_t^2 = \int (t-\langle t\rangle)^2\,|f(t)|^2\,\mathrm{d}t\,, \quad \sigma_\omega^2 = \int (\omega - \langle\omega\rangle)^2 \left|\hat f(\omega)\right|^2 \mathrm{d}\omega\,, \tag{17.55}$$

with $\langle\omega\rangle$ defined from $\left|\hat f\right|^2$ in the same way. Then

$$\sigma_t\,\sigma_\omega \geq \frac{1}{2}\,, \tag{17.56}$$

with equality if and only if f is a Gaussian, $f(t) = C\mathrm{e}^{\mathrm{i}\langle\omega\rangle t}\mathrm{e}^{-(t-\langle t\rangle)^2/(4\sigma_t^2)}$.

Rests on Proposition 17.30, Theorem 17.36, and Equation (17.42).

Derivation. *Derives Theorem 17.43.*

Translating t and modulating by $\mathrm{e}^{-\mathrm{i}\langle\omega\rangle t}$ changes neither σ_t nor σ_ω, by Equation (17.35) and Equation (17.36), so assume $\langle t\rangle = \langle\omega\rangle = 0$. By Equation (17.38) and Plancherel Equation (17.46),

$$\sigma_\omega^2 = \int \omega^2 \left|\hat f\right|^2 \mathrm{d}\omega = \int \left|\widehat{f'}\right|^2 \mathrm{d}\omega = \int |f'(t)|^2 \,\mathrm{d}t\,.$$

Now integrate $\frac{\mathrm{d}}{\mathrm{d}t}\left(t\,|f|^2\right) = |f|^2 + t\left(\bar f f' + f\bar f'\right) = |f|^2 + 2t\,\Re\left(\bar f f'\right)$ over the line. The left-hand side integrates to zero, because $t\,|f(t)|^2 \to 0$ at infinity when tf and f are square integrable, so

$$1 = \int |f|^2 \,\mathrm{d}t = -2\int t\,\Re\left(\bar f(t) f'(t)\right)\mathrm{d}t \leq 2\int |tf(t)|\;|f'(t)|\;\mathrm{d}t \leq 2\,\sigma_t\,\sigma_\omega\,,$$

the last step by the Cauchy–Schwarz inequality applied to tf and f'. That is Equation (17.56).

Equality requires equality in Cauchy–Schwarz, i.e. $f'(t) = \lambda\, t\, f(t)$ for a constant λ, and equality in $\Re(\bar f f') \geq -\left|\bar f f'\right|$, which forces λ real and negative. Integrating the first-order equation gives $f(t) = C\mathrm{e}^{\lambda t^2/2}$, a Gaussian; matching $\sigma_t^2 = -1/(2\lambda)$ from Equation (17.42) and undoing the translation and modulation yields the stated form. □

Corollary 17.44 (Duration and bandwidth in SI)**.** *With t in* s *and $\nu = \omega/2\pi$ in* Hz*, Equation (17.56) reads*

$$\sigma_t\,\sigma_\nu \geq \frac{1}{4\pi} \approx 0.0796\,. \tag{17.57}$$

A transform-limited optical pulse of $\sigma_t = 10\,\mathrm{fs}$ therefore has a spectral width of at least $\sigma_\nu = 7.96\,\mathrm{THz}$; at a centre wavelength of $800\,\mathrm{nm}$, where $\nu_0 = c/\lambda_0 = 375\,\mathrm{THz}$, that is about $2\,\%$ of the carrier frequency, so such a pulse cannot be called monochromatic.

Rests on Theorem 17.43 and Equation (17.37).

Proof. *Derives Corollary 17.44.*

$\sigma_\omega = 2\pi\sigma_\nu$ by Equation (17.37) with $\lambda = 2\pi$; substitute in Equation (17.56). The numbers are $1/(4\pi \times 1 \times 10^{-14}\,\mathrm{s}) = 7.96 \times 10^{12}\,\mathrm{Hz}$ and $c/\lambda_0 = 2.998 \times 10^8\,\mathrm{m/s}/800 \times 10^{-9}\,\mathrm{m}$. □

Example 17.45 (Two purely classical consequences)**.** *Natural linewidth.* An emitter whose amplitude decays as $\mathrm{e}^{-t/(2\tau)}$ has, by Equation (17.41) with $\alpha = 1/(2\tau)$, a Lorentzian spectral profile of full width $\Delta\nu = 1/(2\pi\tau)$. For an excited state of lifetime $\tau = 16\,\mathrm{ns}$ — the order of an allowed optical transition — this is $\Delta\nu = 9.9\,\mathrm{MHz}$, the irreducible floor under every measurement in Chapter 90.

Diffraction. A slit of width a multiplies an incident plane wave by a rectangle, so by Equation (17.40) the transmitted amplitude as a function of transverse wavenumber k_x is a sinc with first zero at $k_x = 2\pi/a$. Since $k_x = k\sin\theta$ with $k = 2\pi/\lambda$, the first dark fringe is at $\sin\theta = \lambda/a$: single-slit diffraction is the rectangle–sinc pair and nothing more. For sodium light, $\lambda = 589\,\mathrm{nm}$, through $a = 1.0\,\mathrm{mm}$, $\theta \approx 5.9\times10^{-4}\,\mathrm{rad}$. The optics is developed in Chapter 61.

Remark 17.46 (The quantum reading, in SI). Wave mechanics assigns to a particle of momentum p the wavenumber $k = p/\hbar$, with $\hbar = 1.054\,571\,817\times10^{-34}\,\mathrm{J\,s}$ exactly by the definition of the SI (Chapter 71). If $\psi(x)$ is a normalised position amplitude and $\tilde\psi(k)$ its transform, then the momentum amplitude is $\tilde\psi(p/\hbar)/\sqrt{\hbar}$, normalised in p, and $\sigma_p = \hbar\,\sigma_k$. Substituting into Equation (17.56) with the pair (x, k) gives

$$\sigma_x\,\sigma_p \geq \frac{\hbar}{2}\,, \tag{17.58}$$

which is Kennard's inequality [580], the sharp form of the uncertainty relation that Heisenberg had argued physically from the disturbance caused by a measurement [576]. The mathematical content is entirely Theorem 17.43; what physics supplies is the single constant $\hbar$ converting an inverse length into a momentum. For an electron localised to $\sigma_x = 1.0\times10^{-10}\,\mathrm{m}$, an atomic diameter, Equation (17.58) gives $\sigma_p \geq 5.27\times10^{-25}\,\mathrm{kg\,m/s}$, hence a kinetic-energy scale $\sigma_p^2/(2m_\mathrm{e}) \approx 1.5\times10^{-19}\,\mathrm{J} \approx 0.95\,\mathrm{eV}$ — the electronvolt scale of atomic physics, obtained from a theorem about Fourier integrals. The operator form of the inequality, valid for any pair of non-commuting observables, belongs to Chapter 77.

Rests on Theorem 17.43.

17.2.5 Poisson summation

Sampling a function and sampling its transform are not independent operations. The exact relation between them is one formula, and it controls both the lattice sums of solid-state physics and the aliasing of every digitised measurement.

Theorem 17.47 (Poisson summation)**.** *Let f be continuous with $|f(t)| \leq C(1+|t|)^{-1-\eta}$ and $\left|\hat f(\omega)\right| \leq C(1+|\omega|)^{-1-\eta}$ for some $\eta, C > 0$. Then for every $T > 0$ and every t,*

$$\sum_{n=-\infty}^{\infty} f(t+nT) = \frac{\sqrt{2\pi}}{T}\sum_{m=-\infty}^{\infty} \hat f\!\left(\frac{2\pi m}{T}\right)\mathrm{e}^{2\pi\mathrm{i}mt/T}\,, \tag{17.59}$$

and in particular, at $t = 0$,

$$\sum_{n=-\infty}^{\infty} f(nT) = \frac{\sqrt{2\pi}}{T}\sum_{m=-\infty}^{\infty} \hat f\!\left(\frac{2\pi m}{T}\right)\,. \tag{17.60}$$

Rests on Definitions 17.1 and 17.29 and Theorem 17.9.

Proof. *Derives Theorem 17.47.*

The *periodisation* $F(t) = \sum_n f(t+nT)$ converges absolutely and uniformly on compact sets by the decay hypothesis (comparison with $\sum n^{-1-\eta}$), so F is continuous, and it is

T-periodic by construction, the sum being over all of $\mathbb{Z}$. Its Fourier coefficients in the sense of Equation (17.3) are

$$C_m = \frac{1}{T}\int_0^T F(t)\,\mathrm{e}^{-2\pi \mathrm{i} m t/T}\,\mathrm{d}t = \frac{1}{T}\sum_{n}\int_0^T f(t+nT)\mathrm{e}^{-2\pi \mathrm{i} m t/T}\mathrm{d}t = \frac{1}{T}\int_{-\infty}^{\infty} f(s)\,\mathrm{e}^{-2\pi \mathrm{i} m s/T}\mathrm{d}s\,,$$

where uniform convergence justifies the exchange and the substitution $s = t+nT$ reassembles the intervals $[nT,(n+1)T]$ into the line, the exponential being unchanged because $\mathrm{e}^{-2\pi \mathrm{i} m n} = 1$. The last integral is $\sqrt{2\pi}\,\hat f(2\pi m/T)/T$. By hypothesis $\sum_m |C_m| < \infty$, so the Fourier series of F converges absolutely and uniformly, and its sum is F by Theorem 17.9 — or by Corollary 17.23(b) applied to the difference. Writing that statement out is Equation (17.59). □

Example 17.48 (Theta identity)**.** Take $f(t) = \mathrm{e}^{-\pi s t^2}$ with $s > 0$. By Equation (17.42) with $a^2 = 1/(2\pi s)$, $\hat f(\omega) = (2\pi s)^{-1/2}\mathrm{e}^{-\omega^2/(4\pi s)}$. Put $T = 1$ in Equation (17.60):

$$\sum_{n=-\infty}^{\infty} \mathrm{e}^{-\pi s n^2} = \frac{1}{\sqrt{s}}\sum_{m=-\infty}^{\infty} \mathrm{e}^{-\pi m^2/s}\,. \tag{17.61}$$

The identity converts a slowly convergent sum at small s into a rapidly convergent one, which is how such lattice sums are evaluated in practice; it is also the functional equation from which the analytic continuation of the Riemann zeta function is obtained.

Corollary 17.49 (Lattice sums and the reciprocal lattice)**.** *Let $\Lambda \subset \mathbb{R}^3$ be a Bravais lattice with primitive vectors $\boldsymbol{a}_1,\boldsymbol{a}_2,\boldsymbol{a}_3$ and cell volume $V = \boldsymbol{a}_1\cdot(\boldsymbol{a}_2\times\boldsymbol{a}_3)$, and let Λ^* be the reciprocal lattice generated by the $\boldsymbol{b}_i$ with $\boldsymbol{a}_i\cdot\boldsymbol{b}_j = 2\pi\delta_{ij}$. Then, for f decaying as above in each variable,*

$$\sum_{\boldsymbol{R}\in\Lambda} f(\boldsymbol{r}+\boldsymbol{R}) = \frac{(2\pi)^{3/2}}{V}\sum_{\boldsymbol{G}\in\Lambda^*} \hat f(\boldsymbol{G})\,\mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{r}}\,, \tag{17.62}$$

with $\hat f$ the three-dimensional transform, the convention of Definition 17.29 applied in each Cartesian coordinate.

Rests on Theorem 17.47 and Definition 17.29.

Proof. *Derives Corollary 17.49.*

Change variables to the coordinates u_i defined by $\boldsymbol{r} = \sum_i u_i\boldsymbol{a}_i$, in which Λ becomes $\mathbb{Z}^3$ and the Jacobian is V; Theorem 17.47 applies in each of the three variables in turn, and the dual variables assemble into Λ^* by the defining relation $\boldsymbol{a}_i\cdot\boldsymbol{b}_j = 2\pi\delta_{ij}$. □

For copper, a face-centred cubic crystal with conventional lattice parameter $a = 0.3615\,\mathrm{nm}$, the reciprocal-lattice scale is $2\pi/a = 17.4/\mathrm{nm}$, which is the size of the Brillouin zone that organises the phonon spectra of Chapter 125 and the band structure of Chapter 126. Equation (17.62) is also the statement that a periodic array of scatterers diffracts only into the reciprocal-lattice directions, which is the Laue condition.

Finally, Equation (17.59) in the frequency variable is the exact statement of what sampling does to a spectrum, and Section 17.5.1 uses it in that form.

17.3 Convolution and correlation

17.3.1 The convolution theorem

Definition 17.50 (Convolution)**.** For f, g absolutely integrable on $\mathbb{R}$,

$$(f * g)(t) = \int_{-\infty}^{\infty} f(s)\,g(t-s)\,\mathrm{d}s\,. \tag{17.63}$$

Proposition 17.51 (Algebra of convolution)**.** *Convolution is bilinear, commutative and associative, it commutes with translations, and*

$$\int_{-\infty}^{\infty} |(f * g)(t)| \, \mathrm{d}t \le \left(\int |f| \right) \left(\int |g| \right) , \tag{17.64}$$

so $f * g$ is again absolutely integrable and is defined for almost every t.
Rests on Definition 17.50.

Proof. *Derives Proposition 17.51.*

Commutativity is the substitution $s \mapsto t - s$; associativity is a change of variables in the resulting double integral. For Equation (17.64), Tonelli's theorem allows the non-negative double integral to be done in either order:

$$\iint |f(s)| \, |g(t-s)| \, \mathrm{d}s \, \mathrm{d}t = \int |f(s)| \left[\int |g(t-s)| \, \mathrm{d}t \right] \mathrm{d}s = \left(\int |f| \right) \left(\int |g| \right) ,$$

finite, which also shows the inner integral in Equation (17.63) converges for almost every t. □

Proposition 17.52 (Linear translation-invariant systems)**.** *Let $\mathcal{A}$ be a linear map on signals that commutes with time translation, $\mathcal{A}\,[x(\cdot - a)] = (\mathcal{A}x)\,(\cdot - a)$ for every a, and that is continuous in the sense of Definition 17.38. Then $\mathcal{A}$ is a convolution,*

$$(\mathcal{A}x)\,(t) = \int_{-\infty}^{\infty} h(t-s)\, x(s) \, \mathrm{d}s = (h * x)(t) , \qquad h = \mathcal{A}\delta , \tag{17.65}$$

with h the impulse response *of the system.*
Rests on Definitions 17.38 and 17.50 and Proposition 17.41.

Derivation. *Derives Proposition 17.52.*

By Proposition 17.41 the sifting property writes any input as a superposition of shifted impulses, $x(t) = \int x(s)\, \delta(t-s) \, \mathrm{d}s$. Apply $\mathcal{A}$ and exchange it with the integral, which continuity permits, treating the integral as a limit of finite sums: $(\mathcal{A}x)\,(t) = \int x(s)\, \mathcal{A}\,[\delta(\cdot - s)]\,(t) \, \mathrm{d}s$. Translation invariance identifies $\mathcal{A}\,[\delta(\cdot - s)]\,(t) = h(t-s)$ with $h = \mathcal{A}\delta$, which is Equation (17.65). The statement made fully rigorous — that every continuous linear map between distribution spaces has a kernel, here forced by translation invariance to depend only on $t - s$ — is Schwartz's kernel theorem [971]. □

Every linear system whose behaviour does not depend on when it is used is therefore a convolution: a filter, an amplifier below saturation, a spectrometer, a lens, the elastic response of a solid. What such a system does to a spectrum is now one line.

Theorem 17.53 (Convolution theorem)**.** *For absolutely integrable f, g,*

$$\widehat{f * g}(\omega) = \sqrt{2\pi}\, \hat{f}(\omega)\, \hat{g}(\omega) , \tag{17.66}$$

and, when both sides are defined,

$$\widehat{f\, g}(\omega) = \frac{1}{\sqrt{2\pi}} \left(\hat{f} * \hat{g} \right) (\omega) . \tag{17.67}$$

Rests on Definitions 17.29 and 17.50, Proposition 17.51, and Theorem 17.32.

Proof. *Derives Theorem 17.53.*

Insert Equation (17.63) into Equation (17.32). The double integral converges absolutely by Equation (17.64), so Fubini's theorem permits exchanging the order:

$$\widehat{f * g}(\omega) = \frac{1}{\sqrt{2\pi}} \iint f(s)\, g(t-s)\, \mathrm{e}^{-\mathrm{i}\omega t} \, \mathrm{d}s \, \mathrm{d}t = \frac{1}{\sqrt{2\pi}} \int f(s)\, \mathrm{e}^{-\mathrm{i}\omega s} \left[\int g(u)\, \mathrm{e}^{-\mathrm{i}\omega u} \, \mathrm{d}u \right] \mathrm{d}s ,$$

substituting $u = t - s$ in the inner integral and factoring $\mathrm{e}^{-\mathrm{i}\omega t} = \mathrm{e}^{-\mathrm{i}\omega s}\mathrm{e}^{-\mathrm{i}\omega u}$. The bracket is $\sqrt{2\pi}\,\hat{g}(\omega)$ and the remaining integral $\sqrt{2\pi}\,\hat{f}(\omega)$, giving Equation (17.66). Equation (17.67) is the same computation run through the inverse transform, or equivalently Equation (17.66) applied to $\hat{f}$ and $\hat{g}$ and inverted with Theorem 17.32. □

This is the pending convolution property of Section 9.7.5, and it is the reason integral transforms solve differential equations. A linear constant-coefficient equation $P(\partial_t)u = w$ has, by Equation (17.38), $P(\mathrm{i}\omega)\hat{u} = \hat{w}$, hence $\hat{u} = \hat{w}/P(\mathrm{i}\omega) = \sqrt{2\pi}\,\hat{G}\hat{w}$ with $\hat{G} = 1/\left[\sqrt{2\pi}P(\mathrm{i}\omega)\right]$, and therefore $u = G * w$: every Green's-function solution of Section 10.3 is a convolution of the source with the fundamental solution, and the Green's function is the impulse response of Equation (17.65). The same statement read backwards is the working fact of experimental spectroscopy.

Example 17.54 (Instrument response and widths in quadrature)**.** Let $g_a(t) = \left(a\sqrt{2\pi}\right)^{-1}\mathrm{e}^{-t^2/(2a^2)}$ be the normalised Gaussian of standard deviation a, so that $\hat{g}_a(\omega) = (2\pi)^{-1/2}\,\mathrm{e}^{-a^2\omega^2/2}$ by Equation (17.42). Then Equation (17.66) gives

$$\widehat{g_a * g_b}(\omega) = \sqrt{2\pi}\,\hat{g}_a\hat{g}_b = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-(a^2+b^2)\omega^2/2} = \hat{g}_{\sqrt{a^2+b^2}}(\omega)\,, \tag{17.68}$$

so $g_a * g_b = g_{\sqrt{a^2+b^2}}$ by injectivity (Corollary 17.33): Gaussian widths add in quadrature, and so do the full widths at half maximum, which are $2\sqrt{2\ln 2}$ times the standard deviations. A spectral line of true Doppler width 1.20 GHz recorded through an instrument of Gaussian response 0.90 GHz is therefore observed with width $\sqrt{1.20^2 + 0.90^2} = 1.50$ GHz, and the physical width is recovered by subtracting in quadrature. Every quoted linewidth in Chapter 90 is a measured width so corrected.

Remark 17.55 (Deconvolution is ill posed)*.* Inverting Equation (17.66) suggests recovering the true signal from the record $y = h * f$ by $\hat{f} = \hat{y}/\left(\sqrt{2\pi}\,\hat{h}\right)$. The formula is correct and the procedure is unusable as it stands. A physical response $\hat{h}$ decays at high frequency — for the Gaussian of Example 17.54, as $\mathrm{e}^{-a^2\omega^2/2}$ — so the division multiplies whatever noise the record carries at frequency ω by $\mathrm{e}^{+a^2\omega^2/2}$. An arbitrarily small perturbation of the data therefore produces an arbitrarily large perturbation of the answer: the solution does not depend continuously on the data, which is exactly the third of Hadamard's conditions for a well-posed problem (Chapter 10). Practical deconvolution restores well-posedness by giving something up — truncating the division at the frequency where the signal-to-noise ratio reaches unity, or adding a penalty term that suppresses large high-frequency content. The same arithmetic returns in Section 17.6.3, where the tomographic ramp filter amplifies high frequencies for the same reason.

17.3.2 Correlation and the Wiener–Khinchin theorem

Definition 17.56 (Correlation)**.** For square-integrable f, g the *cross-correlation* and *auto-correlation* are

$$R_{fg}(\tau) = \int_{-\infty}^{\infty} \overline{f(t)}\,g(t+\tau)\,\mathrm{d}t\,, \qquad R_{ff}(\tau) = \int_{-\infty}^{\infty} \overline{f(t)}\,f(t+\tau)\,\mathrm{d}t\,. \tag{17.69}$$

Correlation differs from convolution only by a reflection and a conjugation, $R_{fg} = \bar{f}^{-} * g$ with $f^{-}(t) = f(-t)$, and that one difference is what makes it measure similarity at a lag rather than smoothing.

Theorem 17.57 (Wiener–Khinchin, finite-energy form)**.** *For square-integrable f,*

$$\widehat{R_{ff}}(\omega) = \sqrt{2\pi}\,\left|\hat{f}(\omega)\right|^2 , \tag{17.70}$$

so the autocorrelation and the energy spectral density are a Fourier pair. In particular $R_{ff}(0) = \int |f|^2$ recovers Plancherel's theorem, and $|R_{ff}(\tau)| \le R_{ff}(0)$.
Rests on Definition 17.56 and Theorem 17.53.

Proof. *Derives Theorem 17.57.*

With $f^-(t) = f(-t)$ one has $\widehat{f^-}(\omega) = \overline{\hat{f}(\omega)}$, since conjugating and reflecting the argument each conjugate the analysis kernel. Then $R_{ff} = \overline{f^-} * f$ by inspection of Equation (17.69), and Equation (17.66) gives Equation (17.70). Evaluating the inverse transform at $\tau = 0$ gives $\int |f|^2 = \int \left|\hat{f}\right|^2$, and the bound on $|R_{ff}|$ is Cauchy–Schwarz applied to Equation (17.69). □

The physically important version concerns signals that do not decay — noise, thermal fluctuations, a stationary source — which have infinite energy but finite *power*. The right object is then a statistical one, and the apparatus is that of Chapter 11.

Definition 17.58 (Stationary process and power spectral density)**.** A random process $X(t)$ is *stationary in the wide sense* if $\langle X(t)\rangle$ is independent of t — taken to be zero below — and the autocorrelation

$$R(\tau) = \left\langle \overline{X(t)}\, X(t+\tau) \right\rangle \tag{17.71}$$

depends only on the lag τ, the angle brackets denoting the ensemble average. Its *power spectral density* is

$$S(\omega) = \lim_{T\to\infty} \frac{1}{T} \left\langle \left|\hat{X}_T(\omega)\right|^2 \right\rangle , \qquad X_T = X\,\mathbf{1}_{\{|t|\le T/2\}} . \tag{17.72}$$

Rests on Definitions 17.29 and 17.56.

Theorem 17.59 (Wiener–Khinchin)**.** *If R is absolutely integrable then the limit Equation (17.72) exists and*

$$S(\omega) = \frac{1}{\sqrt{2\pi}}\,\hat{R}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R(\tau)\,\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau , \tag{17.73}$$

and conversely $R(\tau) = \int S(\omega)\mathrm{e}^{\mathrm{i}\omega\tau}\mathrm{d}\omega$. The total power is $R(0) = \int S(\omega)\,\mathrm{d}\omega$.
Rests on Definitions 17.29 and 17.58 and Theorem 17.32.

Derivation. *Derives Theorem 17.59.*

Expand the definition, using Equation (17.32) on the truncated process:

$$\left\langle \left|\hat{X}_T(\omega)\right|^2 \right\rangle = \frac{1}{2\pi}\int_{-T/2}^{T/2}\!\!\int_{-T/2}^{T/2} \left\langle \overline{X(t)}X(t')\right\rangle \mathrm{e}^{-\mathrm{i}\omega(t'-t)}\,\mathrm{d}t\,\mathrm{d}t' = \frac{1}{2\pi}\int_{-T/2}^{T/2}\!\!\int_{-T/2}^{T/2} R(t'-t)\,\mathrm{e}^{-\mathrm{i}\omega(t'-t)}\,\mathrm{d}t\,\mathrm{d}t' ,$$

the ensemble average passing inside the integrals by stationarity and integrability. Change variables to $\tau = t' - t$ and t; for fixed τ with $|\tau| \le T$ the admissible t form an interval of length $T - |\tau|$, so

$$\frac{1}{T}\left\langle \left|\hat{X}_T(\omega)\right|^2 \right\rangle = \frac{1}{2\pi}\int_{-T}^{T} \left(1 - \frac{|\tau|}{T}\right) R(\tau)\,\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau .$$

The triangular factor is bounded by 1 and tends to 1 pointwise, and R is absolutely integrable, so dominated convergence gives the limit Equation (17.73) as $T \to \infty$. The converse statement is Theorem 17.32 applied to $\hat{R}$, and setting $\tau = 0$ gives the power. □

Wiener obtained the result as part of a general harmonic analysis of non-decaying functions, in which the correlation replaces the divergent transform [673]; Khinchin gave the probabilistic formulation for stationary processes used above [740].

Example 17.60 (Noise spectral densities in SI)**.** For a voltage noise $V(t)$ in V, $R(\tau)$ is in V^2 and, by Equation (17.73), $S(\omega)$ is in $\mathrm{V}^2\,\mathrm{s/rad}$. Laboratory practice quotes instead the *one-sided* density per unit ordinary frequency, $S_V(\nu) = 4\pi S(2\pi\nu)$ in V^2/Hz, so that $\langle V^2\rangle = \int_0^\infty S_V(\nu)\,\mathrm{d}\nu$ with no further factors. The thermal noise of a resistor is white in this measure, with $S_V = 4k_{\mathrm{B}}TR$ — Nyquist's analysis [615], derived from the fluctuation–dissipation theorem in Chapter 124, and quoted here only to fix the units. For $R = 1.0\,\mathrm{k\Omega}$ at $T = 300\,\mathrm{K}$, $S_V = 1.66\times10^{-17}\,\mathrm{V}^2/\mathrm{Hz}$, i.e. an amplitude density $\sqrt{S_V} = 4.1\,\mathrm{nV}$ per root hertz, and in a measurement bandwidth of 100 kHz the root-mean-square noise is $\sqrt{S_V B} = 1.3\,\mu\mathrm{V}$. The whole calculation is Theorem 17.59 plus arithmetic.

One definite integral is worth isolating, because every estimate of the thermal noise floor of a mechanical resonator in this book reduces to it and each would otherwise redo the same contour.

Proposition 17.61 (Mean-square response of a damped resonator)**.** *Let $m > 0$, $b > 0$, $k > 0$ and let*

$$\chi(\omega) = \frac{1}{k - m\omega^2 + \mathrm{i}b\omega} \tag{17.74}$$

be the response of the damped oscillator, with m in kg*, b in* kg/s*, k in* N/m *and χ in* m/N*. Then*

$$\int_{-\infty}^{\infty} |\chi(\omega)|^2\,\mathrm{d}\omega = \frac{\pi}{bk}\,, \qquad \textit{equivalently} \qquad \int_0^\infty |\chi(2\pi\nu)|^2\,\mathrm{d}\nu = \frac{1}{4bk}\,. \tag{17.75}$$

Consequently, by Proposition 17.52 and Theorem 17.59, a resonator driven by a stationary force whose one-sided spectral density S_F in N^2/Hz is constant over the band where χ is appreciable has mean-square displacement $\langle x^2\rangle = S_F/(4bk)$.

Rests on Theorems 8.24 and 17.59 and Proposition 17.52.

Proof. *Derives Proposition 17.61.*

Take first $b^2 \neq 4mk$. For real ω, $|\chi(\omega)|^2 = \chi(\omega)\,\tilde\chi(\omega)$ with $\tilde\chi(\omega) = (k - m\omega^2 - \mathrm{i}b\omega)^{-1}$, and both extend to rational functions of a complex ω decaying like $|\omega|^{-4}$, so the arc at infinity contributes nothing and Theorem 8.24 applies to a contour closed in the upper half-plane. The poles of χ are the roots of $m\omega^2 - \mathrm{i}b\omega - k = 0$,

$$z_\pm = \frac{\mathrm{i}b \pm \Delta}{2m}\,, \qquad \Delta = \sqrt{4mk - b^2}\,, \tag{17.76}$$

and both lie in the upper half-plane: for $4mk > b^2$ each has imaginary part $b/2m > 0$, and for $4mk < b^2$ both are purely imaginary, $\mathrm{i}\left[b \pm \sqrt{b^2 - 4mk}\right]/2m$, with positive imaginary part because $b > \sqrt{b^2 - 4mk}$. The poles of $\tilde\chi$ are $-z_\pm$, in the lower half-plane, so only $z_\pm$ are enclosed. Writing $\chi(\omega) = -\left[m(\omega - z_+)(\omega - z_-)\right]^{-1}$ and using $k - mz_\pm^2 = -\mathrm{i}bz_\pm$, which is the equation the roots satisfy, gives $\tilde\chi(z_\pm) = -(2\mathrm{i}bz_\pm)^{-1}$ and hence, by Proposition 8.23,

$$\operatorname{Res}_{\omega=z_+}\chi\tilde\chi = \frac{1}{2\mathrm{i}bz_+\Delta}\,, \qquad \operatorname{Res}_{\omega=z_-}\chi\tilde\chi = -\frac{1}{2\mathrm{i}bz_-\Delta}\,, \tag{17.77}$$

since $z_+ - z_- = \Delta/m$. Their sum is $(2\mathrm{i}b\Delta)^{-1}(z_- - z_+)/(z_+z_-)$, and the root relations $z_+z_- = -k/m$ and $z_- - z_+ = -\Delta/m$ reduce it to $(2\mathrm{i}bk)^{-1}$. Multiplying by $2\pi\mathrm{i}$ gives the first equality of Equation (17.75); the second follows by substituting $\omega = 2\pi\nu$ and halving, the integrand being even in ω. Both sides of Equation (17.75) are continuous in b at $b^2 = 4mk$, which disposes of the critically damped case. □

Example 17.62 (Noise floor of a torsion balance)**.** A torsion pendulum of stiffness $k = 2.0 \times 10^{-8}\,\mathrm{N\,m/rad}$ and damping $b = 1.0 \times 10^{-11}\,\mathrm{N\,m\,s/rad}$ — the rotational analogues of the constants above, with angle in place of displacement — driven by the thermal torque whose one-sided density is $S_\tau = 4k_{\mathrm{B}}Tb$ at $T = 300\,\mathrm{K}$, has mean-square angle $\langle\theta^2\rangle = S_\tau/(4bk) = k_{\mathrm{B}}T/k$ by Equation (17.75). The damping cancels: the answer is $2.1 \times 10^{-13}\,\mathrm{rad}^2$, an angular fluctuation of 0.46 µrad, and it is equipartition (Chapter 118) recovered from the spectrum. That the cancellation happens is the content of the proposition; that it must happen is why the fluctuation–dissipation relation takes the form it does in Chapter 124.

Theorem 17.63 (Matched filter)**.** *Let the record be $d(t) = s(t) + n(t)$, with s a known signal and n stationary noise of strictly positive power spectral density S. Among all linear filters q, the signal-to-noise ratio*

$$\rho^2 = \frac{\left|\int \overline{\hat{q}(\omega)}\,\hat{s}(\omega)\,\mathrm{d}\omega\right|^2}{\int S(\omega)\,|\hat{q}(\omega)|^2\,\mathrm{d}\omega} \tag{17.78}$$

of the filtered output at the arrival time is maximised by

$$\hat{q}(\omega) = \frac{\hat{s}(\omega)}{S(\omega)}\,, \qquad \textit{giving} \qquad \rho^2_{\max} = \int_{-\infty}^{\infty} \frac{|\hat{s}(\omega)|^2}{S(\omega)}\,\mathrm{d}\omega\,. \tag{17.79}$$

Rests on Definition 17.58.

Proof. *Derives Theorem 17.63.*

Write $\hat{q} = \hat{u}/\sqrt{S}$ and $\hat{\sigma} = \hat{s}/\sqrt{S}$, legitimate since $S > 0$. The numerator of Equation (17.78) is $\left|\int \overline{\hat{u}}\,\hat{\sigma}\,\mathrm{d}\omega\right|^2$ and the denominator is $\int |\hat{u}|^2\,\mathrm{d}\omega$, so $\rho^2 = |\langle\hat{u},\hat{\sigma}\rangle|^2 \,/\, \|\hat{u}\|^2 \le \|\hat{\sigma}\|^2$ by the Cauchy–Schwarz inequality, with equality if and only if $\hat{u}$ is a multiple of $\hat{\sigma}$, i.e. $\hat{q} \propto \hat{s}/S$. Substituting gives $\rho^2_{\max} = \|\hat{\sigma}\|^2$, which is Equation (17.79). □

The optimal filter is therefore the signal itself, weighted inversely by the noise power at each frequency: one trusts the record where the detector is quiet and discounts it where the detector is loud. This is the statistic by which the gravitational-wave strain records of Chapter 47 are searched, and Equation (17.79) is the integral whose value is quoted as the detection significance. The same theorem, read in the frequency domain alone, is why the noise curve of an interferometer — a plot of $\sqrt{S}$ in $1/\mathrm{Hz}^{1/2}$ against ν in Hz — is the complete specification of what the instrument can find.

Remark 17.64 (Spectroscopy by autocorrelation)*.* A two-beam interferometer records, as a function of the path difference x, the intensity $I(x) \propto R_{EE}(x/c)$ of the autocorrelation of the field. By Theorem 17.57 its Fourier transform is the spectrum. A Fourier-transform spectrometer therefore measures no spectrum directly at all: it measures an interferogram and transforms it, which is why its resolution is set by the maximum path difference, $\Delta\nu \approx c/(2x_{\max})$ by Corollary 17.44, and why its signal-to-noise advantage over a scanning instrument is that all frequencies are measured at once. The coherence properties this measures are the subject of Chapter 122.

17.4 The Laplace transform

17.4.1 Definition, convergence and inversion

The Fourier transform requires decay at both ends of the line. Physical initial-value problems supply neither: a system is switched on at $t = 0$ and may grow. The remedy is to keep only $t > 0$ and to insert a convergence factor, which is what Laplace's generating-function transform does [56].

Definition 17.65 (One-sided Laplace transform)**.** For f defined on $[0,\infty)$ and $s \in \mathbb{C}$,

$$F(s) = \mathcal{L}\{f\}(s) = \int_0^\infty f(t)\,\mathrm{e}^{-st}\,\mathrm{d}t\,, \tag{17.80}$$

whenever the improper integral converges. If t is in s then s is in 1/s and F carries the SI unit of f multiplied by s.

Theorem 17.66 (Half-plane of convergence)**.** *If Equation (17.80) converges at s_0, it converges for every s with $\Re s > \Re s_0$, uniformly on every sector $|\arg(s-s_0)| \le \theta_0 < \pi/2$. Hence there is an* abscissa of convergence $\sigma_c \in [-\infty, +\infty]$ *such that the transform exists for $\Re s > \sigma_c$ and fails for $\Re s < \sigma_c$.*

Rests on Definition 17.65 and Equation (7.28).

Proof. *Derives Theorem 17.66.*

Let $g(t) = \int_0^t f(u)\mathrm{e}^{-s_0 u}\,\mathrm{d}u$, which converges to a limit as $t \to \infty$ by hypothesis and is therefore bounded, say $|g| \le M$. Write $z = s - s_0$, with $\Re z > 0$, and integrate by parts (Equation (7.28)), using $g' = f\mathrm{e}^{-s_0 t}$:

$$\int_0^T f(t)\mathrm{e}^{-st}\,\mathrm{d}t = \int_0^T g'(t)\,\mathrm{e}^{-zt}\,\mathrm{d}t = g(T)\,\mathrm{e}^{-zT} + z\int_0^T g(t)\,\mathrm{e}^{-zt}\,\mathrm{d}t\,.$$

The boundary term is bounded by $M\mathrm{e}^{-T\Re z} \to 0$, and the remaining integral converges absolutely as $T \to \infty$ because $\left|g\mathrm{e}^{-zt}\right| \le M\mathrm{e}^{-t\Re z}$. Both estimates depend on z only through $|z|/\Re z \le 1/\cos\theta_0$ and $\Re z$, whence uniform convergence on the sector. The abscissa is the infimum of $\Re s$ over the set of convergence, which the first part shows to be a right half-plane. □

Proposition 17.67 (Holomorphy)**.** *On the half-plane $\Re s > \sigma_a$ of* absolute *convergence, F is holomorphic (Definition 8.5) with*

$$F'(s) = -\int_0^\infty t\,f(t)\,\mathrm{e}^{-st}\,\mathrm{d}t\,, \tag{17.81}$$

and more generally $F^{(n)}(s) = \mathcal{L}\{(-t)^n f\}(s)$.

Rests on Definitions 8.5 and 17.65 and Theorem 7.38.

Proof. *Derives Proposition 17.67.*

Fix s with $\Re s = \sigma > \sigma_a$ and choose $\eta > 0$ with $\sigma - \eta > \sigma_a$. For $|h| < \eta/2$,

$$\frac{F(s+h)-F(s)}{h} + \int_0^\infty t f(t)\mathrm{e}^{-st}\mathrm{d}t = \int_0^\infty f(t)\,\mathrm{e}^{-st}\,\frac{\mathrm{e}^{-ht}-1+ht}{h}\,\mathrm{d}t\,,$$

and the elementary estimate $\left|\mathrm{e}^{-ht}-1+ht\right| \le \tfrac{1}{2}|h|^2\,t^2\mathrm{e}^{|h|t}$, obtained from the Taylor remainder (Theorem 7.38), bounds the integrand by $\tfrac{1}{2}|h|\,t^2\,|f(t)|\,\mathrm{e}^{-(\sigma-\eta/2)t}$. Since $t^2\mathrm{e}^{\eta t/2}$ is dominated by $\mathrm{e}^{\eta t}$ for large t, the bound is integrable and proportional to $|h|$, so the difference quotient converges and Equation (17.81) holds. The higher derivatives follow by induction. □

Proposition 17.68 (The Laplace transform is a Fourier transform)**.** *Let $f_\sigma(t) = f(t)\,\mathrm{e}^{-\sigma t}\,\theta(t)$, the causal signal damped by $\mathrm{e}^{-\sigma t}$. Then for $\sigma > \sigma_a$*

$$F(\sigma + \mathrm{i}\omega) = \sqrt{2\pi}\;\widehat{f_\sigma}(\omega)\,. \tag{17.82}$$

Rests on Definitions 17.29 and 17.65.

Proof. *Derives Proposition 17.68.*

Substitute $s = \sigma + \mathrm{i}\omega$ in Equation (17.80) and read the result as Equation (17.32) applied to f_σ, the step function restricting the range to $t > 0$. □

So the Laplace transform is the Fourier transform of an exponentially damped causal signal, and its values on any vertical line $\Re s = \sigma$ are a Fourier spectrum. The advantage over the Fourier transform is that the damping can be chosen to make the integral converge, and that the resulting function of s is holomorphic, so that its values on one line determine it everywhere in the half-plane by analytic continuation. Inversion follows at once.

Theorem 17.69 (Bromwich inversion integral)**.** *Let $\gamma > \sigma_a$ and suppose $\omega \mapsto F(\gamma + \mathrm{i}\omega)$ is absolutely integrable. Then at every point of continuity of f*

$$f(t) = \frac{1}{2\pi \mathrm{i}} \int_{\gamma - \mathrm{i}\infty}^{\gamma + \mathrm{i}\infty} F(s)\, \mathrm{e}^{st}\, \mathrm{d}s\,, \qquad t > 0\,, \tag{17.83}$$

the integral being taken up the vertical line $\Re s = \gamma$; and the same integral vanishes for $t < 0$.
Rests on Proposition 17.68 and Theorem 17.32.

Proof. *Derives Theorem 17.69.*

By Equation (17.82) the hypothesis says $\widehat{f_\gamma}$ is absolutely integrable, so Theorem 17.32 applies to f_γ:

$$f(t)\, \mathrm{e}^{-\gamma t}\, \theta(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \widehat{f_\gamma}(\omega)\, \mathrm{e}^{\mathrm{i}\omega t}\, \mathrm{d}\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\gamma + \mathrm{i}\omega)\, \mathrm{e}^{\mathrm{i}\omega t}\, \mathrm{d}\omega\,.$$

Multiply by $\mathrm{e}^{\gamma t}$ and substitute $s = \gamma + \mathrm{i}\omega$, so that $\mathrm{d}s = \mathrm{i}\,\mathrm{d}\omega$ and the line is traversed upward:

$$f(t)\, \theta(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\gamma + \mathrm{i}\omega)\, \mathrm{e}^{(\gamma + \mathrm{i}\omega)t}\, \mathrm{d}\omega = \frac{1}{2\pi \mathrm{i}} \int_{\gamma - \mathrm{i}\infty}^{\gamma + \mathrm{i}\infty} F(s)\, \mathrm{e}^{st}\, \mathrm{d}s\,.$$

The factor $\theta(t)$ on the left is the statement that the integral vanishes for $t < 0$. □

The contour of Equation (17.83) is Bromwich's, introduced precisely to justify the operational methods of the next subsection [442]. Evaluating it is a residue computation, and the machinery is that of Section 8.7; the only new ingredient is the estimate that closes the contour.

Lemma 17.70 (Jordan's lemma on the Bromwich contour)**.** *Let C_R be the left-hand semicircle $s = \gamma + R\mathrm{e}^{\mathrm{i}\phi}$, $\phi \in [\pi/2, 3\pi/2]$, and let $M_R = \max_{C_R} |F|$. Then for $t > 0$*

$$\left| \int_{C_R} F(s)\, \mathrm{e}^{st}\, \mathrm{d}s \right| \le \frac{\pi\, \mathrm{e}^{\gamma t}\, M_R}{t}\,, \tag{17.84}$$

so the semicircular contribution vanishes as $R \to \infty$ whenever $M_R \to 0$. For $t < 0$ the same bound holds on the right-hand semicircle.
Rests on Lemma 8.9.

Proof. *Derives Lemma 17.70.*

On C_R, write $\phi = \pi/2 + \psi$ with $\psi \in [0, \pi]$; then $\Re(s - \gamma) = R\cos\phi = -R\sin\psi$ and $|\mathrm{e}^{st}| = \mathrm{e}^{\gamma t}\mathrm{e}^{-Rt\sin\psi}$. The ML estimate (Lemma 8.9) with $|\mathrm{d}s| = R\,\mathrm{d}\psi$ gives

$$\left| \int_{C_R} F\mathrm{e}^{st}\mathrm{d}s \right| \le \mathrm{e}^{\gamma t} M_R\, R \int_0^{\pi} \mathrm{e}^{-Rt\sin\psi}\, \mathrm{d}\psi = 2\,\mathrm{e}^{\gamma t} M_R\, R \int_0^{\pi/2} \mathrm{e}^{-Rt\sin\psi}\, \mathrm{d}\psi\,,$$

and Jordan's inequality $\sin\psi \ge 2\psi/\pi$ on $[0, \pi/2]$ — concavity of the sine there — bounds the last integral by $\int_0^{\pi/2} \mathrm{e}^{-2Rt\psi/\pi}\mathrm{d}\psi \le \pi/(2Rt)$, which is Equation (17.84). □

Corollary 17.71 (Residue evaluation and causality)**.** *Let F be holomorphic except for finitely many poles $s_1, \dots, s_k$, all with $\Re s_j < \gamma$, and let $F(s) \to 0$ uniformly in direction as $|s| \to \infty$. Then*

$$f(t) = \sum_{j=1}^{k} \operatorname{Res}_{s = s_j} \left[F(s)\, \mathrm{e}^{st} \right] \qquad (t > 0)\,, \qquad f(t) = 0 \quad (t < 0)\,. \tag{17.85}$$

Rests on Theorems 8.12, 8.24 and 17.69 and Lemma 17.70.

Proof. *Derives Corollary 17.71.*

For $t > 0$ close the Bromwich line with the left semicircle of Lemma 17.70; the closed contour encircles every pole once counterclockwise, the semicircular contribution vanishes by Equation (17.84), and the residue theorem (Theorem 8.24) evaluates what is left, the $2\pi\mathrm{i}$ cancelling the prefactor of Equation (17.83). For $t < 0$ close to the right instead: the enclosed region contains no singularity, so Cauchy's theorem (Theorem 8.12) gives zero. □

That the same integral gives zero for $t < 0$ is not a technicality: it is causality, built into the transform by the restriction to $t > 0$, and it is the reason the poles of F must lie to the left of the contour.

Example 17.72 (Damped sinusoid)**.** Let $F(s) = \omega_0/\left[(s+\alpha)^2 + \omega_0^2\right]$, with $\alpha, \omega_0 > 0$ in 1/s and rad/s. The poles are simple, at $s_\pm = -\alpha \pm \mathrm{i}\omega_0$, and by Equation (8.16) with denominator derivative $2(s+\alpha)$,

$$\operatorname{Res}_{s=s_+}\left[F\mathrm{e}^{st}\right] = \frac{\omega_0\,\mathrm{e}^{(-\alpha+\mathrm{i}\omega_0)t}}{2\mathrm{i}\omega_0} = \frac{\mathrm{e}^{-\alpha t}\mathrm{e}^{\mathrm{i}\omega_0 t}}{2\mathrm{i}}\,,$$

and the residue at s_- is its complex conjugate with the sign reversed. Adding, $f(t) = \mathrm{e}^{-\alpha t}\sin(\omega_0 t)$ for $t > 0$ — the impulse response of a damped oscillator, recovered from its transform by one contour integral.

Proposition 17.73 (Initial- and final-value theorems)**.** *Let f be continuous on $[0,\infty)$ with f' absolutely integrable. Then*

$$\lim_{s\to\infty} s\,F(s) = f(0^+)\,, \qquad \lim_{s\to 0^+} s\,F(s) = \lim_{t\to\infty} f(t)\,, \tag{17.86}$$

the second whenever the limit on its right exists, the first for s real and increasing without bound.

Rests on Definition 17.65, Equation (7.28), and Theorem 7.43.

Proof. *Derives Proposition 17.73.*

Integrating Equation (17.80) by parts gives $\mathcal{L}\{f'\}(s) = sF(s) - f(0^+)$, so $sF(s) - f(0^+) = \int_0^\infty f'(t)\mathrm{e}^{-st}\mathrm{d}t$. The integrand is dominated by $|f'|$, which is integrable, and tends to 0 pointwise as $s \to \infty$; dominated convergence gives the first limit. As $s \to 0^+$ the same integrand tends pointwise to $f'(t)$ under the same domination, so the integral tends to $\int_0^\infty f' = \lim_{t\to\infty} f(t) - f(0^+)$ by the fundamental theorem of calculus (Theorem 7.43), which rearranges to the second limit. □

17.4.2 Initial-value problems and the operational calculus

Proposition 17.74 (Derivative rule)**.** *If f is n times differentiable with $f^{(n)}$ of exponential order,*

$$\mathcal{L}\left\{f^{(n)}\right\}(s) = s^n F(s) - s^{n-1}f(0) - s^{n-2}f'(0) - \cdots - f^{(n-1)}(0)\,. \tag{17.87}$$

Rests on Definition 17.65 and Equation (7.28).

Proof. *Derives Proposition 17.74.*

For $n = 1$, integrate by parts (Equation (7.28)): $\int_0^\infty f'\mathrm{e}^{-st}\mathrm{d}t = \left[f\mathrm{e}^{-st}\right]_0^\infty + s\int_0^\infty f\mathrm{e}^{-st}\mathrm{d}t = -f(0) + sF(s)$, the boundary term at infinity vanishing for $\Re s$ large. Induction on n, applying the case $n = 1$ to $f^{(n-1)}$, gives Equation (17.87). □

Equation (17.87) is the whole method: it converts a linear ordinary differential equation with constant coefficients into an algebraic equation, *and* it swallows the initial data, which enter as the extra polynomial terms rather than as conditions imposed afterwards. This subsumes the constant-coefficient theory of Section 9.1 and discharges the section reserved for the transform at Section 9.13.

Example 17.75 (The driven damped oscillator, in SI)**.** Consider

$$m\,\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+b\,\frac{\mathrm{d}x}{\mathrm{d}t}+k\,x=f(t)\,,\qquad x(0)=x_0\,,\quad \frac{\mathrm{d}x}{\mathrm{d}t}(0)=v_0\,, \tag{17.88}$$

with m in kg, b in kg/s, k in N/m and f in N. Transforming with Equation (17.87),

$$\left(ms^2+bs+k\right)X(s)=F(s)+m\,(sx_0+v_0)+b\,x_0\,, \tag{17.89}$$

an algebraic equation. Define the *transfer function*

$$G(s)=\frac{1}{ms^2+bs+k}\,, \tag{17.90}$$

in m/N, whose poles are $s_\pm=-\Gamma\pm\mathrm{i}\omega_d$ with $\Gamma=b/(2m)$ and $\omega_d=\sqrt{\omega_0^2-\Gamma^2}$, $\omega_0=\sqrt{k/m}$. By Corollary 17.71 the inverse transform of G is the impulse response $h(t)=\mathrm{e}^{-\Gamma t}\sin(\omega_d t)/(m\omega_d)$, and by the convolution theorem in its Laplace form the forced part of the solution is $x_f=h*f$ — Duhamel's principle, the same statement as the causal Green's function of Section 10.3. For $m=0.50\,\mathrm{kg}$, $k=200\,\mathrm{N/m}$ and $b=2.0\,\mathrm{kg/s}$ one has $\omega_0=20\,\mathrm{rad/s}$, $\Gamma=2.0/\mathrm{s}$, $\omega_d=19.9\,\mathrm{rad/s}$ and a quality factor $Q=m\omega_0/b=5.0$: the free oscillation decays by a factor e in 0.50 s, after about 1.6 cycles.

Corollary 17.76 (Heaviside's expansion theorem)**.** *Let $F=P/Q$ with P,Q polynomials,* $\deg P<\deg Q$*, and Q having only simple zeros s_j. Then*

$$f(t)=\sum_j \frac{P(s_j)}{Q'(s_j)}\,\mathrm{e}^{s_j t}\,,\qquad t>0\,. \tag{17.91}$$

Rests on Corollary 17.71 and Equation (8.16).

Proof. *Derives Corollary 17.76.*

$F(s)\to 0$ as $|s|\to\infty$ because $\deg P<\deg Q$, so Corollary 17.71 applies, and each residue is $P(s_j)\mathrm{e}^{s_j t}/Q'(s_j)$ by Equation (8.16). □

Equation (17.91) is exactly the rule Heaviside used, and taught, without proof: he treated the differentiation operator $p=\mathrm{d}/\mathrm{d}t$ as an algebraic symbol, wrote the solution of $Q(p)x=P(p)w$ as $x=[P(p)/Q(p)]\,w$, expanded the rational function in partial fractions, and interpreted each term by rules justified only by their success [258]. Bromwich's contour integral is what turned the procedure into mathematics [442], and Corollary 17.76 is the translation.

Proposition 17.77 (Poles and stability)**.** *A linear time-invariant system with rational transfer function G proper and with poles s_j has an impulse response that decays to zero for every initial condition if and only if $\Re s_j<0$ for every j; it is bounded but non-decaying if a simple pole sits on the imaginary axis and every other pole is strictly to the left; and it grows without bound otherwise.*

Rests on Corollary 17.71.

Proof. *Derives Proposition 17.77.*

By Corollary 17.71 the impulse response is a finite sum of terms $t^{m_j}\mathrm{e}^{s_j t}$ with m_j below the order of the pole, whose modulus is $t^{m_j}\mathrm{e}^{t\Re s_j}$. Such a term tends to zero iff $\Re s_j<0$; it is bounded and non-decaying iff $\Re s_j=0$ and $m_j=0$; and it diverges otherwise. Distinct exponentials being linearly independent, no cancellation between terms is possible. □

Theorem 17.78 (Causality implies dispersion relations)**.** *Let χ be a causal response function — $\chi(t) = 0$ for $t < 0$ — absolutely integrable, with transform $\hat{\chi}$. Then $\hat{\chi}$ extends to a function holomorphic in the upper half plane $\Im\omega > 0$, and if $\hat{\chi}(\omega) \to 0$ as $|\omega| \to \infty$ there,*

$$\Re\hat{\chi}(\omega) = \frac{1}{\pi}\,\mathrm{PV}\int_{-\infty}^{\infty} \frac{\Im\hat{\chi}(\omega')}{\omega' - \omega}\,\mathrm{d}\omega'\,, \qquad \Im\hat{\chi}(\omega) = -\frac{1}{\pi}\,\mathrm{PV}\int_{-\infty}^{\infty} \frac{\Re\hat{\chi}(\omega')}{\omega' - \omega}\,\mathrm{d}\omega'\,. \tag{17.92}$$

Rests on Definition 17.29, Propositions 17.42 and 17.67, Theorem 8.12, Lemma 8.9, and Equation (8.16).

Derivation. Derives Theorem 17.78.

For $\Im\omega > 0$ the factor $\mathrm{e}^{-\mathrm{i}\omega t}$ decays as $\mathrm{e}^{t\Im\omega}$ on the half-line $t < 0$ — where χ vanishes — and as $\mathrm{e}^{-t\Im\omega}$ for $t > 0$, so $\hat{\chi}(\omega) = \frac{1}{\sqrt{2\pi}}\int_0^\infty \chi(t)\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t$ converges there and, by the argument of Proposition 17.67, is holomorphic. Now integrate $\hat{\chi}(\omega')/(\omega' - \omega)$ for real ω around the closed contour consisting of the real axis indented by a small semicircle of radius η *above* the point $\omega' = \omega$, closed by a large semicircle in the upper half plane. The integrand is holomorphic inside, so Cauchy's theorem (Theorem 8.12) makes the total vanish. The large semicircle contributes nothing by the decay hypothesis and the ML estimate (Lemma 8.9). The small semicircle, traversed clockwise, contributes $-\mathrm{i}\pi\hat{\chi}(\omega)$ — half a residue, by Equation (8.16) and the parametrisation $\omega' = \omega + \eta\mathrm{e}^{\mathrm{i}\phi}$. What remains of the real axis is the principal value. Hence

$$\mathrm{PV}\int_{-\infty}^{\infty} \frac{\hat{\chi}(\omega')}{\omega' - \omega}\,\mathrm{d}\omega' = \mathrm{i}\pi\,\hat{\chi}(\omega)\,,$$

and separating real and imaginary parts gives Equation (17.92). Equivalently, the statement is Equation (17.54) applied to $1/(\omega' - \omega - \mathrm{i}\varepsilon)$. □

Real and imaginary parts of a causal response are thus not independent: knowing how a medium absorbs at all frequencies determines how it refracts at each, and conversely. The physical instances — the Kramers–Kronig relations for the permittivity and the refractive index — belong to Chapter 62; what has been proved here is that they follow from causality and analyticity alone, with no model of the medium.

17.5 Sampling and the discrete transform

17.5.1 The sampling theorem

Every measurement that reaches a computer has been sampled: the continuous output of a detector is read at instants $t_n = n\Delta t$ and everything between is discarded. The question of what is lost has an exact answer.

Definition 17.79 (Band-limited function)**.** f is *band limited* to Ω if it is square integrable and $\hat{f}(\omega) = 0$ for $|\omega| > \Omega$. In terms of ordinary frequency the band is $|\nu| \le \nu_{\max} = \Omega/2\pi$, in Hz.

Rests on Definition 17.29.

Theorem 17.80 (Sampling theorem)**.** *Let f be band limited to Ω and let $\Delta t = \pi/\Omega$. Then f is determined by its samples at the instants $n\Delta t$, and is reconstructed from them exactly by the* cardinal series

$$f(t) = \sum_{n=-\infty}^{\infty} f(n\Delta t)\,\operatorname{sinc}\!\left(\frac{\pi\,(t - n\Delta t)}{\Delta t}\right), \qquad \operatorname{sinc} x = \frac{\sin x}{x}\,. \tag{17.93}$$

Equivalently, in Hz*: a signal containing no frequency above $\nu_{\max}$ is completely determined by samples taken at any rate $\nu_{\mathrm{s}} = 1/\Delta t \ge 2\nu_{\max}$, the* Nyquist rate*.*

Rests on Definitions 17.1 and 17.79 and Theorems 17.32 and 17.34.

Proof. *Derives Theorem 17.80.*

On the interval $[-\Omega,\Omega]$, expand $\hat f$ in a Fourier series of period 2Ω in the variable ω: $\hat f(\omega)=\sum_n \gamma_n \mathrm{e}^{-\mathrm{i}n\pi\omega/\Omega}$ with $\gamma_n=\frac{1}{2\Omega}\int_{-\Omega}^{\Omega}\hat f(\omega)\mathrm{e}^{\mathrm{i}n\pi\omega/\Omega}\mathrm{d}\omega$, the series converging in mean square by Theorem 17.34. Since $\hat f$ vanishes outside the band, the inversion formula Equation (17.43) reads $f(t)=\frac{1}{\sqrt{2\pi}}\int_{-\Omega}^{\Omega}\hat f(\omega)\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}\omega$; evaluating it at $t=n\pi/\Omega=n\Delta t$ identifies the coefficients,

$$f(n\Delta t)=\frac{2\Omega}{\sqrt{2\pi}}\,\gamma_n\,,\qquad\text{i.e.}\qquad \gamma_n=\frac{\sqrt{2\pi}}{2\Omega}\,f(n\Delta t)\,. \tag{17.94}$$

So the samples *are* the Fourier coefficients of the spectrum, up to a constant, and knowing them is knowing $\hat f$, hence f. To get Equation (17.93) explicitly, substitute the series for $\hat f$ into the inversion integral and exchange sum and integral — legitimate because the series converges in mean square on a set of finite measure, so it converges in mean, and $\mathrm{e}^{\mathrm{i}\omega t}$ is bounded:

$$f(t)=\frac{1}{\sqrt{2\pi}}\sum_n\gamma_n\int_{-\Omega}^{\Omega}\mathrm{e}^{\mathrm{i}\omega(t-n\Delta t)}\mathrm{d}\omega=\frac{1}{\sqrt{2\pi}}\sum_n\gamma_n\,\frac{2\sin\left[\Omega\,(t-n\Delta t)\right]}{t-n\Delta t}\,.$$

Insert Equation (17.94); the constants collapse to $1/\Omega$ and leave

$$\sum_n f(n\,\Delta t)\,\frac{\sin\left[\Omega(t-n\,\Delta t)\right]}{\Omega(t-n\,\Delta t)}\,,$$

which is Equation (17.93) because $\Omega=\pi/\Delta t$. Sampling faster than the Nyquist rate only makes the band a proper subinterval of $[-\Omega,\Omega]$, which changes nothing in the argument. □

Whittaker introduced the cardinal series as the interpolation formula distinguished by containing no frequency above the sampling limit [433]; Nyquist, analysing how many independent symbols a telegraph line of given bandwidth can carry, obtained the rate [614]; and Shannon stated and proved the theorem in the form used ever since, as the bridge between a continuous waveform and a finite list of numbers [947].

Corollary 17.81 (Sampling periodises the spectrum; aliasing)**.** *Let $\omega_{\mathrm{s}}=2\pi/\Delta t$. For f satisfying the decay hypotheses of Theorem 17.47,*

$$\sum_{m=-\infty}^{\infty}\hat f(\omega-m\omega_{\mathrm{s}})=\frac{\Delta t}{\sqrt{2\pi}}\sum_{n=-\infty}^{\infty}f(n\Delta t)\,\mathrm{e}^{-\mathrm{i}n\,\omega\,\Delta t}\,. \tag{17.95}$$

The right-hand side depends on f only through its samples; the left is the spectrum with all its translates by multiples of the sampling frequency superposed. If f is band limited to $\Omega<\omega_{\mathrm{s}}/2$ the translates do not overlap and $\hat f$ is recovered by discarding all but the central copy — which is Theorem 17.80 again. If they overlap, a component at ν is indistinguishable from components at $\nu-m\nu_{\mathrm{s}}$ for every integer m: it is aliased.

Rests on Theorems 17.32 and 17.47 and Definition 17.79.

Proof. *Derives Corollary 17.81.*

The left-hand side is periodic in ω with period ω_{s}, and the periodisation argument of Theorem 17.47 applies verbatim in the frequency variable: its Fourier coefficients are

$$C_n=\frac{1}{\omega_{\mathrm{s}}}\int_{-\infty}^{\infty}\hat f(\omega)\,\mathrm{e}^{-\mathrm{i}n\Delta t\,\omega}\,\mathrm{d}\omega=\frac{\sqrt{2\pi}}{\omega_{\mathrm{s}}}\,f(-n\Delta t)\,,$$

using $2\pi/\omega_{\mathrm{s}}=\Delta t$ and the inversion formula Equation (17.43). Summing the series $\sum_n C_n\mathrm{e}^{\mathrm{i}n\Delta t\,\omega}$ and relabelling $n\mapsto -n$ gives Equation (17.95), since $\sqrt{2\pi}/\omega_{\mathrm{s}}=\Delta t/\sqrt{2\pi}$. □

Example 17.82 (Aliasing in the laboratory)**.** Audio is conventionally sampled at $\nu_{\mathrm{s}} = 44.1\,\mathrm{kHz}$, so the Nyquist frequency is $22.05\,\mathrm{kHz}$. A $25.0\,\mathrm{kHz}$ component that reaches the converter is recorded as $|25.0\,\mathrm{kHz} - 44.1\,\mathrm{kHz}| = 19.1\,\mathrm{kHz}$, squarely inside the audible band and indistinguishable, after sampling, from a real signal there. Nothing downstream can undo this, which is why the analogue low-pass filter that removes such components *before* the converter is part of the apparatus and must be described as such. The gravitational-wave strain channel of Chapter 47 is recorded at $16\,384\,\mathrm{Hz}$, giving a Nyquist frequency of $8192\,\mathrm{Hz}$, comfortably above the few-hundred-hertz band where the signals sit.

Remark 17.83 (What the theorem does not say)*.* Three caveats, each of which has cost somebody a measurement. (i) *No signal of finite duration is band limited.* If $\hat{f}$ vanishes outside $[-\Omega, \Omega]$ then $f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} \hat{f}(\omega) \mathrm{e}^{\mathrm{i}\omega t} \mathrm{d}\omega$ converges for complex t and is differentiable there, hence entire; were it to vanish on an interval, all its derivatives would vanish at an interior point, its Taylor series (Theorem 8.20) would vanish identically, and f would be identically zero. Band limitation and finite support are mutually exclusive, so every real record is aliased a little, and the design question is how much. (ii) The reconstruction Equation (17.93) converges slowly: the sinc decays only as $1/t$, so a truncated cardinal series has errors falling as the reciprocal of the number of samples retained. (iii) An error in the sample *times* — jitter — is not an error in the sample values and is not corrected by any amount of averaging; its effect is a phase noise proportional to frequency, which is why timing specifications tighten as the band widens.

17.5.2 The discrete Fourier transform and the FFT

Definition 17.84 (Discrete Fourier transform)**.** For $x = (x_0, \dots, x_{N-1}) \in \mathbb{C}^N$,

$$X_k = \sum_{n=0}^{N-1} x_n \, W_N^{nk}, \qquad W_N = \mathrm{e}^{-2\pi \mathrm{i}/N}, \qquad k = 0, \dots, N-1 . \tag{17.96}$$

Proposition 17.85 (Inversion and Parseval for the DFT)**.** *The roots of unity satisfy*

$$\frac{1}{N} \sum_{n=0}^{N-1} W_N^{n(k-k')} = \delta_{kk'} \qquad \text{for } 0 \le k, k' \le N-1 , \tag{17.97}$$

hence

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k \, W_N^{-nk}, \qquad \text{and} \qquad \sum_{n=0}^{N-1} |x_n|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X_k|^2 . \tag{17.98}$$

Rests on Definition 17.84.

Proof. *Derives Proposition 17.85.*

If $k = k'$ every term of Equation (17.97) is 1. If not, $W_N^{k-k'} \neq 1$ and the geometric sum is $\left(1 - W_N^{N(k-k')}\right) / \left(1 - W_N^{k-k'}\right) = 0$, because $W_N^N = 1$. This is Lemma 17.2 for a finite cyclic group in place of the circle. Substituting Equation (17.96) into the right side of Equation (17.98) and using Equation (17.97) returns x_n; the Parseval identity follows by the same computation applied to $\sum_n |x_n|^2 = \sum_n \bar{x}_n x_n$. □

Proposition 17.86 (The DFT is exact for a band-limited periodic signal)**.** *Let f be periodic with period $T_{\mathrm{w}} = N\Delta t$ and have Fourier coefficients c_n vanishing for $|n| \ge N/2$. Then the DFT of the samples $x_n = f(n\Delta t)$ satisfies $X_k = N c_k$ for $0 \le k < N/2$ and $X_{N-k} = N c_{-k}$ for $0 < k \le N/2$. No information is lost and no approximation is made.*

Rests on Definitions 17.1 and 17.84 and Proposition 17.85.

Proof. *Derives Proposition 17.86.*

Insert $f(n\Delta t) = \sum_{|m|<N/2} c_m \mathrm{e}^{2\pi i m n/N}$, which is Equation (17.4) at $t = n\Delta t$ with $\omega_1 \Delta t = 2\pi/N$, into Equation (17.96): $X_k = \sum_m c_m \sum_{n=0}^{N-1} W_N^{n(k-m)} = N c_{k \bmod N}$ by Equation (17.97), the index being read modulo N. Negative m therefore appear at $N - |m|$, which is the stated indexing. □

For a signal that is neither exactly periodic over the record nor exactly band limited — that is, for every real record — the DFT differs from the continuous transform by exactly two effects, both already derived: *aliasing*, Corollary 17.81, from sampling; and *leakage*, from the finite record length, since truncating to N samples multiplies the signal by a rectangular window and hence, by Equation (17.67), convolves the spectrum with the window's transform.

Theorem 17.87 (Cooley–Tukey)**.** *Let $N = 2^q$. Then Equation (17.96) can be evaluated in $\frac{N}{2}\log_2 N$ complex multiplications and $N \log_2 N$ complex additions, rather than the N^2 multiplications of the direct sum.*

Rests on Definition 17.84.

Derivation. *Derives Theorem 17.87.*

Split the sum into even and odd indices, $n = 2r$ and $n = 2r + 1$, and use $W_N^{2rk} = W_{N/2}^{rk}$:

$$X_k = \underbrace{\sum_{r=0}^{N/2-1} x_{2r} W_{N/2}^{rk}}_{E_k} + W_N^k \underbrace{\sum_{r=0}^{N/2-1} x_{2r+1} W_{N/2}^{rk}}_{O_k}, \tag{17.99}$$

so E and O are the discrete transforms of the even- and odd-indexed halves, each of length $N/2$. They are periodic in k with period $N/2$, and $W_N^{k+N/2} = -W_N^k$, so the second half of the output costs nothing new:

$$X_k = E_k + W_N^k O_k\,, \qquad X_{k+N/2} = E_k - W_N^k O_k\,, \qquad 0 \le k < N/2\,. \tag{17.100}$$

Combining the two half-length transforms therefore costs $N/2$ complex multiplications and N additions, whence the recursion $M(N) = 2M(N/2) + N/2$ with $M(1) = 0$. Dividing by N, $M(N)/N = M(N/2)/(N/2) + 1/2$, so $M(N)/N$ increases by $1/2$ at each of the $\log_2 N$ levels of the recursion, giving $M(N) = \frac{N}{2}\log_2 N$; the addition count follows identically. □

Example 17.88 (What the saving is worth)**.** For $N = 2^{20} = 1\,048\,576$ samples — one minute of a channel sampled at 16 384 Hz — the direct sum needs $N^2 \approx 1.1 \times 10^{12}$ complex multiplications and the fast algorithm $\frac{N}{2}\log_2 N \approx 1.05 \times 10^7$, a factor of about 1.0×10^5. This is the difference between an hour of computation and a tenth of a second, and it is why every measured spectrum quoted in this treatise, from the angular power spectrum of Chapter 51 to the strain spectra of Chapter 47, exists.

The algorithm was published by Cooley and Tukey in 1965 [1304] and transformed numerical practice within a few years. It was not new: Gauss had the same doubling recursion in an unpublished note of 1805, written while interpolating the orbit of the asteroid Pallas, and several nineteenth- and twentieth-century authors rediscovered pieces of it — a history reconstructed by Heideman, Johnson and Burrus [1740]. The treatise records the attribution as they establish it: the algorithm is Gauss's, its impact is Cooley and Tukey's.

Example 17.89 (Leakage, windowing and resolution)**.** A record of N samples at spacing Δt spans $T_{\mathrm{w}} = N\Delta t$ and resolves frequencies separated by

$$\Delta\nu = \frac{1}{T_{\mathrm{w}}} = \frac{\nu_{\mathrm{s}}}{N}\,. \tag{17.101}$$

With $N = 2^{16} = 65\,536$ samples at $\nu_s = 16\,384\,\mathrm{Hz}$ the record is 4.0 s long and the resolution 0.25 Hz: no processing of that record can separate two lines closer than that, because Corollary 17.44 forbids it. A sinusoid whose frequency does not fall exactly on a bin is smeared over neighbouring bins by convolution with the transform of the rectangular window, Equation (17.40), whose first side lobe sits at $-13.3\,\mathrm{dB}$ relative to the peak — the value $20\log_{10}|\operatorname{sinc} x|$ at the first stationary point $x \approx 4.4934$ of the sinc, where $\tan x = x$. A weak line 20 dB below a strong neighbour is therefore buried by the neighbour's leakage. Tapering the record with a smooth window suppresses the side lobes at the price of broadening the main lobe: it is Theorem 17.22 again — a non-negative kernel in place of an oscillating one — and the trade between resolution and dynamic range is not avoidable, only chosen.

17.6 Other integral transforms

17.6.1 Hankel and Mellin transforms

The Fourier kernel $\mathrm{e}^{-\mathrm{i}\omega t}$ diagonalises translations of the line. Other symmetries call for other kernels: rotations about an axis call for Bessel functions, and dilations of the half-line call for powers. Both transforms below are, at bottom, the Fourier transform in adapted coordinates, and both are treated systematically by Titchmarsh [802] and Morse and Feshbach [1023].

Proposition 17.90 (Hankel transform as the axially symmetric Fourier transform)**.** *Work in $\mathbb{R}^3$ with cylindrical coordinates (ρ, φ, z), and let $f(\boldsymbol{r}) = g(\rho)\,\mathrm{e}^{\mathrm{i}m\varphi}\mathrm{e}^{\mathrm{i}k_z z}$. Then its three-dimensional Fourier transform is supported on the plane k_z fixed and equals*

$$\mathrm{i}^{-m}\,\mathrm{e}^{\mathrm{i}m\varphi_k}\,(\mathcal{H}_m g)(k_\perp)\,, \qquad (\mathcal{H}_m g)(k_\perp) = \int_0^\infty g(\rho)\,J_m(k_\perp\rho)\,\rho\,\mathrm{d}\rho\,, \tag{17.102}$$

up to the fixed normalisation of the transverse transform, where J_m is the Bessel function of Section 9.8 and $(k_\perp, \varphi_k)$ are polar coordinates in the transverse plane of wavevectors. The transform Equation (17.102) is its own inverse:

$$g(\rho) = \int_0^\infty (\mathcal{H}_m g)(k_\perp)\,J_m(k_\perp\rho)\,k_\perp\,\mathrm{d}k_\perp\,. \tag{17.103}$$

Rests on Definition 17.29 and Proposition 17.41.

Proof. *Derives Proposition 17.90.*

The z dependence separates and produces the constraint on k_z by Equation (17.52). In the transverse plane, write $\boldsymbol{k}_\perp \cdot \boldsymbol{\rho} = k_\perp\rho\cos(\varphi - \varphi_k)$ and do the angular integral with the Bessel integral representation

$$\int_0^{2\pi} \mathrm{e}^{-\mathrm{i}k_\perp\rho\cos\alpha}\,\mathrm{e}^{\mathrm{i}m\alpha}\,\mathrm{d}\alpha = 2\pi\,\mathrm{i}^{-m}\,J_m(k_\perp\rho)\,,$$

which is the generating-function identity of Section 9.8.5 evaluated on the unit circle. What survives is the radial integral Equation (17.102). For the inversion, run the same computation on the two-dimensional inversion integral; the angular factor $\mathrm{e}^{\mathrm{i}m\varphi}$ is reproduced and the radial measure $k_\perp \mathrm{d}k_\perp$ appears, giving Equation (17.103). □

Example 17.91 (Axially symmetric potential problem)**.** Let Φ satisfy Laplace's equation in the half space $z > 0$ of $\mathbb{R}^3$, with axially symmetric boundary data $\Phi(\rho, 0) = \Phi_0(\rho)$ in V and $\Phi \to 0$ as $z \to \infty$. Separating in cylindrical coordinates (Section 10.2) gives solutions $\mathrm{e}^{-kz}J_0(k\rho)$ for each $k > 0$ in 1/m, so

$$\Phi(\rho, z) = \int_0^\infty A(k)\,\mathrm{e}^{-kz}\,J_0(k\rho)\,k\,\mathrm{d}k\,, \qquad A(k) = (\mathcal{H}_0\Phi_0)(k)\,, \tag{17.104}$$

the coefficient being read off by Equation (17.103) at $z = 0$. The Hankel transform does for a cylindrical boundary what the Fourier transform does for a planar one; the diffraction integrals of Chapter 61 for a circular aperture are the same transform with $m = 0$.

The Mellin transform. Where Fourier analysis diagonalises translation, Mellin analysis diagonalises *dilation*. The multiplicative group of positive reals, with its invariant measure $\mathrm{d}x/x$, has characters $x \mapsto x^{-s}$, and the transform against them is the following.

Definition 17.92 (Mellin transform)**.** For f defined on $(0,\infty)$ and $s \in \mathbb{C}$,

$$\tilde f(s) = (\mathcal{M} f)(s) = \int_0^\infty f(x)\, x^{s-1}\, \mathrm{d}x\,, \tag{17.105}$$

whenever the integral converges absolutely. For $s = N+1$ with N a non-negative integer, $\tilde f(N+1) = \int_0^\infty f(x) x^N \mathrm{d}x$ is the N-th *moment* of f; the Mellin transform is thus the analytic continuation of the sequence of moments to complex order.

Theorem 17.93 (Fundamental strip and holomorphy)**.** *Suppose f is continuous on $(0,\infty)$ with*

$$|f(x)| \le C\, x^{-a} \quad (0 < x \le 1)\,, \qquad |f(x)| \le C\, x^{-b} \quad (x \ge 1)\,, \qquad a < b\,. \tag{17.106}$$

Then Equation (17.105) converges absolutely on the open strip $a < \Re s < b$ and defines a holomorphic function there, bounded on every closed substrip. The strip is called fundamental*, and $\tilde f$ determines f on it.*

Rests on Definition 17.92 and Proposition 17.67.

Proof. *Derives Theorem 17.93.*

Split the integral at $x = 1$. With $\sigma = \Re s$, $\int_0^1 |f|\, x^{\sigma-1}\mathrm{d}x \le C \int_0^1 x^{\sigma-a-1}\mathrm{d}x = C/(\sigma - a)$, finite exactly when $\sigma > a$; and $\int_1^\infty |f|\, x^{\sigma-1}\mathrm{d}x \le C \int_1^\infty x^{\sigma-b-1}\mathrm{d}x = C/(b-\sigma)$, finite exactly when $\sigma < b$. Both bounds are uniform on a closed substrip, so the integral converges uniformly there and the argument of Proposition 17.67 — differentiation under the integral sign, the extra factor $\ln x$ being absorbed by shrinking the substrip — gives holomorphy, with $\tilde f'(s) = \int_0^\infty f(x) \ln x\, x^{s-1}\mathrm{d}x$. Uniqueness follows from Theorem 17.94 below. □

Theorem 17.94 (Mellin inversion)**.** *Under the hypotheses of Theorem 17.93, if for some c in the fundamental strip $\omega \mapsto \tilde f(c + \mathrm{i}\omega)$ is absolutely integrable, then at every point of continuity*

$$f(x) = \frac{1}{2\pi \mathrm{i}} \int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} \tilde f(s)\, x^{-s}\, \mathrm{d}s\,, \tag{17.107}$$

the contour being the vertical line $\Re s = c$, traversed upward. The value of the integral is independent of c within the strip.

Rests on Theorems 8.12, 17.32 and 17.93.

Proof. *Derives Theorem 17.94.*

Substitute $x = \mathrm{e}^{-u}$, so $\mathrm{d}x/x = -\mathrm{d}u$ and the half-line $0 < x < \infty$ becomes the whole u line reversed. Writing $\phi(u) = f(\mathrm{e}^{-u})$ and $s = c + \mathrm{i}\omega$,

$$\tilde f(c+\mathrm{i}\omega) = \int_{-\infty}^{\infty} \phi(u)\, \mathrm{e}^{-(c+\mathrm{i}\omega)u}\, \mathrm{d}u = \sqrt{2\pi}\, \widehat{\phi_c}(\omega)\,, \qquad \phi_c(u) = \phi(u)\, \mathrm{e}^{-cu}\,, \tag{17.108}$$

exactly the two-sided analogue of Equation (17.82): the Mellin transform on the line $\Re s = c$ is the Fourier transform of $f(\mathrm{e}^{-u})\mathrm{e}^{-cu}$. The hypothesis makes $\widehat{\phi_c}$ absolutely integrable, so Theorem 17.32 gives

$$\phi(u)\, \mathrm{e}^{-cu} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde f(c+\mathrm{i}\omega)\, \mathrm{e}^{\mathrm{i}\omega u}\, \mathrm{d}\omega\,,$$

and multiplying by e^{cu}, substituting $s = c + \mathrm{i}\omega$ with $\mathrm{d}s = \mathrm{i}\mathrm{d}\omega$ and returning to $x = \mathrm{e}^{-u}$ — so that $\mathrm{e}^{su} = x^{-s}$ — yields Equation (17.107). Independence of c follows from Cauchy's theorem (Theorem 8.12) applied to the rectangle between two admissible lines, the horizontal sides contributing nothing because $\tilde f$ is bounded on the closed substrip and integrable along the verticals. □

Example 17.95 (Two transforms worth knowing)**.** *(i)* For $f(x)=\mathrm{e}^{-x}$ the bounds Equation (17.106) hold with $a=0$ and any b, so the fundamental strip is $\Re s>0$ and, by the definition of the gamma function,

$$\mathcal{M}\left\{\mathrm{e}^{-x}\right\}(s)=\Gamma(s)\,,\qquad \Re s>0\,. \tag{17.109}$$

Its poles at $s=0,-1,-2,\dots$ lie to the left of the strip, with residues $(-1)^n/n!$; by Proposition 17.96 these poles *are* the small-x expansion $\mathrm{e}^{-x}=\sum(-x)^n/n!$. *(ii)* For $f(x)=1/(1+x)$ the strip is $0<\Re s<1$ and $\mathcal{M}f(s)=\pi/\sin(\pi s)$, a function whose poles at every integer encode both the expansion of f in powers of x at the origin and its expansion in powers of $1/x$ at infinity.

Proposition 17.96 (Poles give asymptotics)**.** *Let $\tilde f$ continue meromorphically to $c'<\Re s<b$ with finitely many poles $s_1,\dots,s_k$ there and no pole on $\Re s=c'$, and suppose $\tilde f\to 0$ uniformly as $|\Im s|\to\infty$ in the closed strip. Then, as $x\to 0^+$,*

$$f(x)=\sum_{j=1}^{k}\operatorname{Res}_{s=s_j}\left[\tilde f(s)\,x^{-s}\right]+O\!\left(x^{-c'}\right)\,, \tag{17.110}$$

and a simple pole at s_j with residue r_j contributes $r_jx^{-s_j}$, a double pole a term in $x^{-s_j}\ln x$.
Rests on Theorems 8.24 and 17.94, Lemma 8.9, and Definition 8.22.

Proof. *Derives Proposition 17.96.*

Apply the residue theorem (Theorem 8.24) to the rectangle with vertical sides $\Re s=c$ and $\Re s=c'$ and horizontal sides at $\Im s=\pm T$. Traversed counterclockwise this runs *up* the line $\Re s=c$ and *down* the line $\Re s=c'$, so

$$\frac{1}{2\pi\mathrm{i}}\int_{(c)}-\frac{1}{2\pi\mathrm{i}}\int_{(c')}=\sum_{j}\operatorname{Res}_{s=s_j}\left[\tilde f(s)x^{-s}\right]\,,$$

the horizontal sides vanishing as $T\to\infty$ by the decay hypothesis and the ML estimate (Lemma 8.9). The first term is $f(x)$ by Equation (17.107); the second is bounded by $x^{-c'}\int\left|\tilde f(c'+\mathrm{i}\omega)\right|\mathrm{d}\omega/(2\pi)$, since $|x^{-s}|=x^{-\Re s}$. The residue of a simple pole is $r_jx^{-s_j}$ by Definition 8.22, and of a double pole $\mathrm{d}/\mathrm{d}s$ of $\tilde f_2(s)x^{-s}$, which produces the logarithm. □

This is why the Mellin transform is the natural instrument for asymptotics: the analytic structure of $\tilde f$ to the left of the fundamental strip is, term by term, the expansion of f near the origin, and the structure to the right is the expansion at infinity.

17.6.2 Multiplicative convolution and the moments

The property that makes the Mellin transform indispensable is that it turns a convolution over the multiplicative group into a product — exactly as Theorem 17.53 does for the additive group.

Definition 17.97 (Multiplicative convolution)**.** For f,g on $(0,\infty)$,

$$(f\star g)(x)=\int_0^\infty f(y)\,g\!\left(\frac{x}{y}\right)\frac{\mathrm{d}y}{y}\,. \tag{17.111}$$

The measure $\mathrm{d}y/y$ is the invariant (Haar) measure of the group $((0,\infty),\times)$: it satisfies $\mathrm{d}(\lambda y)/(\lambda y)=\mathrm{d}y/y$ for every $\lambda>0$, which is what makes Equation (17.111) the exact analogue of Equation (17.63).

Theorem 17.98 (Mellin convolution theorem)**.** *If the fundamental strips of $\tilde f$ and $\tilde g$ overlap, then on the common strip*

$$\mathcal{M}\{f\star g\}(s)=\tilde f(s)\,\tilde g(s)\,. \tag{17.112}$$

The transform is commutative and associative, and $(f\star g)\star h$ has transform $\tilde f\tilde g\tilde h$.

Rests on Definitions 17.92 and 17.97.

Proof. *Derives Theorem 17.98.*

Substitute Equation (17.111) into Equation (17.105) and exchange the order of integration, legitimate by Fubini's theorem because on the common strip $\iint |f(y)|\,|g(x/y)|\,x^{\sigma-1}\mathrm{d}x\,\mathrm{d}y/y$ is finite by the same computation performed with absolute values:

$$\mathcal{M}\{f\star g\}(s)=\int_0^\infty \frac{f(y)}{y}\left[\int_0^\infty g\!\left(\frac{x}{y}\right)x^{s-1}\mathrm{d}x\right]\mathrm{d}y\,.$$

In the inner integral put $x=yu$, so $\mathrm{d}x=y\,\mathrm{d}u$ and $x^{s-1}=y^{s-1}u^{s-1}$; the bracket becomes $y^s\int_0^\infty g(u)u^{s-1}\mathrm{d}u=y^s\tilde g(s)$. What remains is $\tilde g(s)\int_0^\infty f(y)y^{s-1}\mathrm{d}y=\tilde f(s)\tilde g(s)$. Commutativity and associativity follow, as for Proposition 17.51, from the substitution $y\mapsto x/y$ and from Equation (17.112) with uniqueness of the transform. □

Corollary 17.99 (Convolution on the unit interval)**.** *Suppose f and g vanish for $x>1$. Then so does $f\star g$, and for $0<x\le 1$*

$$(f\star g)(x)=\int_x^1 f(y)\,g\!\left(\frac{x}{y}\right)\frac{\mathrm{d}y}{y}\,, \tag{17.113}$$

while the Mellin transform reduces to the moment integral

$$\tilde f(N+1)=\int_0^1 f(x)\,x^N\,\mathrm{d}x=:M_N[f]\,. \tag{17.114}$$

Consequently the moments of a multiplicative convolution are the products of the moments:

$$M_N[f\star g]=M_N[f]\;M_N[g]\qquad \text{for every } N\,. \tag{17.115}$$

Rests on Definitions 17.92 and 17.97 and Theorem 17.98.

Proof. *Derives Corollary 17.99.*

In Equation (17.111) the factor $f(y)$ vanishes unless $y\le 1$ and $g(x/y)$ vanishes unless $x/y\le 1$, i.e. $y\ge x$; so the range is $x\le y\le 1$, which is empty for $x>1$ and gives Equation (17.113) otherwise. Equation (17.114) is Definition 17.92 at $s=N+1$ with the support restriction, and Equation (17.115) is Equation (17.112) evaluated at $s=N+1$, the point lying in the common strip whenever the moments converge. □

Equation (17.115) is the statement the rest of this subsection exists for. A convolution over a multiplicative variable — a fraction, a ratio, a scale — is an integral operator whose kernel couples every value of the variable to every other, and it is intractable as it stands. Taking moments replaces it by ordinary multiplication, one independent equation for each N: the moments *diagonalise* the convolution. The consequence for evolution equations is immediate.

Theorem 17.100 (Moments diagonalise a convolution evolution)**.** *Let $f(x,\tau)$ be supported in $0<x\le 1$ and satisfy the integro-differential equation*

$$\frac{\partial f(x,\tau)}{\partial\tau}=(P\star f)(x,\tau)=\int_x^1 P(y)\,f\!\left(\frac{x}{y},\tau\right)\frac{\mathrm{d}y}{y}\,, \tag{17.116}$$

with a fixed kernel P whose moments $\tilde{P}(N+1)$ exist. Then the moments $M_N(\tau) = \int_0^1 f(x,\tau)x^N \mathrm{d}x$ satisfy the decoupled ordinary differential equations

$$\frac{\mathrm{d}M_N(\tau)}{\mathrm{d}\tau} = \tilde{P}(N+1)\, M_N(\tau)\,, \qquad \textit{whence} \qquad M_N(\tau) = M_N(0)\, \mathrm{e}^{\tilde{P}(N+1)\,\tau}\,, \tag{17.117}$$

and $f(x,\tau)$ is recovered from the moments by the inversion contour Equation (17.107), with $M_N = \tilde{f}(N+1)$ continued to complex order.

Rests on Corollary 17.99 and Theorem 17.94.

Proof. *Derives Theorem 17.100.*

Take the N-th moment of Equation (17.116). On the left, differentiation under the integral sign (dominated convergence, the moments being finite on a neighbourhood of τ) gives $\mathrm{d}M_N/\mathrm{d}\tau$. On the right, Equation (17.115) gives $\tilde{P}(N+1)M_N(\tau)$. That is Equation (17.117), a linear first-order equation with constant coefficient, whose solution is the exponential. Recovery of f is Theorem 17.94 applied to the analytic continuation $\tilde{f}(s,\tau) = \tilde{f}(s,0)\mathrm{e}^{\tilde{P}(s)\tau}$, which agrees with M_N at $s = N+1$. □

Corollary 17.101 (Coupled channels)**.** *If $f = (f_1, \dots, f_p)$ is a vector of densities on $(0,1]$ obeying $\partial_\tau f_i = \sum_j P_{ij} \star f_j$ with a fixed matrix of kernels, then the vectors of moments $\boldsymbol{M}_N(\tau)$ obey*

$$\frac{\mathrm{d}\boldsymbol{M}_N}{\mathrm{d}\tau} = \mathbf{A}_N\, \boldsymbol{M}_N\,, \qquad (\mathbf{A}_N)_{ij} = \tilde{P}_{ij}(N+1)\,, \qquad \boldsymbol{M}_N(\tau) = \mathrm{e}^{\mathbf{A}_N \tau}\, \boldsymbol{M}_N(0)\,. \tag{17.118}$$

If $\mathbf{A}_N$ is diagonalisable, with eigenvalues $\lambda_N^{(\alpha)}$ and left *eigenvectors $\boldsymbol{u}_N^{(\alpha)}$ — those satisfying $(\boldsymbol{u}_N^{(\alpha)})^{\mathsf{T}}\mathbf{A}_N = \lambda_N^{(\alpha)}(\boldsymbol{u}_N^{(\alpha)})^{\mathsf{T}}$ — then the scalar combinations $\boldsymbol{u}_N^{(\alpha)} \cdot \boldsymbol{M}_N$, different at each N, evolve as pure exponentials $\mathrm{e}^{\lambda_N^{(\alpha)}\tau}$ and do not mix.*

Rests on Theorem 17.100 and Corollary 17.99.

Proof. *Derives Corollary 17.101.*

Take moments of each component and apply Equation (17.115) to each $P_{ij} \star f_j$; the result is Equation (17.118), a linear system with constant matrix, solved by the matrix exponential. Contracting Equation (17.118) with a left eigenvector gives $\mathrm{d}\left(\boldsymbol{u}_N^{(\alpha)} \cdot \boldsymbol{M}_N\right)/\mathrm{d}\tau = \lambda_N^{(\alpha)}\, \boldsymbol{u}_N^{(\alpha)} \cdot \boldsymbol{M}_N$, which decouples the system. □

Remark 17.102 (Why this is the useful form). Three features of Theorem 17.100 deserve emphasis. First, the diagonalisation is exact, not perturbative: no approximation has been made about the kernel P beyond the existence of its moments. Second, the eigenvalue $\tilde{P}(N+1)$ is a *number* for each N, computed once from the kernel, so an evolution over any interval in τ costs one exponential per moment. Third, and least often stated, reconstructing f from its moments is the ill-conditioned step: the moment problem amplifies small errors in M_N into large errors in f, for the same reason as Remark 17.55, and the stable route is not to invert the moments numerically but to continue $\tilde{f}(s,\tau)$ analytically in s and evaluate the contour integral Equation (17.107) directly. The scale variable τ in Equation (17.116) is dimensionless, being the logarithm of a ratio of two quantities of the same SI dimension; the moments M_N carry whatever dimension $f\,\mathrm{d}x$ carries and N affects none of it.

17.6.3 The Radon transform

The last transform in this chapter integrates a function not against a kernel but over a family of submanifolds, and asks whether the function can be recovered from the integrals. Radon settled the question in 1917 for lines in the plane and planes in space [449], as pure mathematics; the answer became an instrument half a century later.

Definition 17.103 (Radon transform)**.** For f on $\mathbb{R}^3$, integrable and decaying, and for a unit vector $\boldsymbol{n}$ and a real p,

$$(\mathcal{R}f)(\boldsymbol{n},p)=\int_{\mathbb{R}^3} f(\boldsymbol{x})\,\delta(p-\boldsymbol{n}\cdot\boldsymbol{x})\,\mathrm{d}^3x\,, \tag{17.119}$$

the integral of f over the plane $\boldsymbol{n}\cdot\boldsymbol{x}=p$. In a two-dimensional slice $z=\text{const}$ of the same space the identical formula with d^2x gives the integral along the line $\boldsymbol{n}\cdot\boldsymbol{x}=p$, which is what a transmission measurement records: for a beam of initial intensity I_0 and attenuation coefficient $\mu(\boldsymbol{x})$ in $1/\mathrm{m}$, each element of path removes $\mathrm{d}I=-\mu\,I\,\mathrm{d}\ell$, which integrates along the ray to $\ln(I_0/I)=(\mathcal{R}\mu)(\boldsymbol{n},p)$ — the Beer–Lambert law, whose left-hand side is dimensionless.

Theorem 17.104 (Projection-slice theorem)**.** *Let $\hat{f}$ denote the d-dimensional Fourier transform of f and $\mathcal{F}_1$ the one-dimensional transform in the variable p. Then*

$$\mathcal{F}_1\left[(\mathcal{R}f)(\boldsymbol{n},\cdot)\right](\kappa)=(2\pi)^{(d-1)/2}\,\hat{f}(\kappa\boldsymbol{n})\,, \tag{17.120}$$

in the normalisation of Definition 17.29: the one-dimensional transform of one projection is the d-dimensional transform of f restricted to the line through the origin in the direction $\boldsymbol{n}$.
Rests on Definitions 17.29 and 17.103.

Proof. *Derives Theorem 17.104.*
Insert Equation (17.119) and do the p integral first, the delta collapsing it:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\!\!\int_{\mathbb{R}^d} f(\boldsymbol{x})\,\delta(p-\boldsymbol{n}\cdot\boldsymbol{x})\,\mathrm{e}^{-\mathrm{i}\kappa p}\,\mathrm{d}^dx\,\mathrm{d}p=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}^d} f(\boldsymbol{x})\,\mathrm{e}^{-\mathrm{i}\kappa\boldsymbol{n}\cdot\boldsymbol{x}}\,\mathrm{d}^dx\,,$$

which is $\hat{f}(\kappa\boldsymbol{n})$ times $(2\pi)^{d/2}/\sqrt{2\pi}=(2\pi)^{(d-1)/2}$, the d-dimensional transform carrying one factor $(2\pi)^{-1/2}$ per coordinate. □

So the projections do determine f: they fill the transform space along every ray through the origin, hence everywhere. Inversion is the statement of that fact in a form one can compute with.

Theorem 17.105 (Radon inversion in three-dimensional space)**.** *For f smooth and rapidly decreasing on $\mathbb{R}^3$,*

$$f(\boldsymbol{x})=-\frac{1}{8\pi^2}\int_{S^2}\frac{\partial^2}{\partial p^2}\,(\mathcal{R}f)(\boldsymbol{n},p)\Big|_{p=\boldsymbol{n}\cdot\boldsymbol{x}}\,\mathrm{d}\Omega(\boldsymbol{n})\,, \tag{17.121}$$

the integral running over the unit sphere of directions.
Rests on Theorems 17.32 and 17.104.

Derivation. *Derives Theorem 17.105.*
Write the three-dimensional inversion integral in polar coordinates in $\boldsymbol{k}$, using $\boldsymbol{k}=\kappa\boldsymbol{n}$ with $\kappa\in\mathbb{R}$ and $\boldsymbol{n}$ over the whole sphere, which double-counts and so carries a factor $\tfrac12$:

$$f(\boldsymbol{x})=\frac{1}{(2\pi)^3}\int_{\mathbb{R}^3}\left[\int f\,\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{y}}\mathrm{d}^3y\right]\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\,\mathrm{d}^3k=\frac{1}{2(2\pi)^3}\int_{S^2}\!\!\int_{-\infty}^{\infty}\kappa^2\,g_{\boldsymbol{n}}(\kappa)\,\mathrm{e}^{\mathrm{i}\kappa\boldsymbol{n}\cdot\boldsymbol{x}}\,\mathrm{d}\kappa\,\mathrm{d}\Omega\,,$$

where $g_{\boldsymbol{n}}(\kappa)=\int(\mathcal{R}f)(\boldsymbol{n},p)\mathrm{e}^{-\mathrm{i}\kappa p}\mathrm{d}p$ by Theorem 17.104 in its unnormalised form. Since $\kappa^2\mathrm{e}^{\mathrm{i}\kappa p}=-\partial_p^2\mathrm{e}^{\mathrm{i}\kappa p}$, the inner integral is $-\partial_p^2\left[2\pi\,(\mathcal{R}f)(\boldsymbol{n},p)\right]$ evaluated at $p=\boldsymbol{n}\cdot\boldsymbol{x}$, by one-dimensional Fourier inversion. Collecting the constants, $-2\pi/\left[2(2\pi)^3\right]=-1/(8\pi^2)$, gives Equation (17.121). □

Theorem 17.106 (Filtered back-projection in a plane slice)**.** *In a two-dimensional slice, with* $\boldsymbol{n}(\theta) = (\cos\theta, \sin\theta)$ *and projections* $P_\theta(p) = (\mathcal{R}f)(\boldsymbol{n}(\theta), p)$*,*

$$f(\boldsymbol{x}) = \frac{1}{4\pi^2} \int_0^\pi \int_{-\infty}^{\infty} \left[\int P_\theta(p)\, \mathrm{e}^{-\mathrm{i}\kappa p} \mathrm{d}p \right] |\kappa|\, \mathrm{e}^{\mathrm{i}\kappa \boldsymbol{n}(\theta)\cdot\boldsymbol{x}}\, \mathrm{d}\kappa\, \mathrm{d}\theta\,. \tag{17.122}$$

Reading the formula from the inside out: transform each projection, multiply by the ramp filter $|\kappa|$*, transform back, and smear the result back along the direction it came from (*back-projection*), summing over directions.*

Rests on Theorems 17.32 and 17.104.

Derivation. *Derives Theorem 17.106.*

Repeat the previous computation in two dimensions. The inverse transform in polar coordinates $\boldsymbol{k} = \kappa\boldsymbol{n}(\theta)$ has area element $|\kappa|\,\mathrm{d}\kappa\,\mathrm{d}\theta$ with $\kappa \in \mathbb{R}$ and $\theta \in [0, \pi)$ — the modulus is the Jacobian and the halved angular range compensates the sign of κ — so

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^2} \int_0^\pi \int_{-\infty}^{\infty} \hat{f}_{\mathrm{u}}\left(\kappa\boldsymbol{n}(\theta)\right) \mathrm{e}^{\mathrm{i}\kappa\boldsymbol{n}\cdot\boldsymbol{x}}\, |\kappa|\, \mathrm{d}\kappa\, \mathrm{d}\theta\,,$$

with $\hat{f}_{\mathrm{u}}$ the unnormalised transform, which Theorem 17.104 identifies with the bracket of Equation (17.122). □

The factor $|\kappa|$ is the whole content of the word "filtered": plain back-projection without it reconstructs f blurred by a $1/r$ kernel, because the polar measure over-weights low frequencies. It is also what makes the reconstruction noise-sensitive, by Remark 17.55: the ramp amplifies exactly the high-frequency components in which detector noise is loudest, so a practical filter is $|\kappa|$ rolled off above a cutoff, and the choice of cutoff is the trade between resolution and noise.

Corollary 17.107 (How many projections)**.** *To reconstruct features of size* d *across an object of diameter* D*, the detector spacing must satisfy* $\Delta p \le d/2$ *by Theorem 17.80 applied to* p*, and the angular spacing must satisfy* $D\,\Delta\theta/2 \le d/2$*, so that adjacent projections do not skip a resolution element at the rim. Hence about* $\pi D/(2d)$ *view angles are needed.*

Rests on Theorem 17.80 and Definition 17.103.

Proof. *Derives Corollary 17.107.*

The first condition is the Nyquist criterion for the variable p, whose band limit is set by the finest structure present. For the second, two projections at angles differing by $\Delta\theta$ differ, at radius $D/2$, by an arc $(D/2)\,\Delta\theta$; requiring that this not exceed the resolution d gives $\Delta\theta \le 2d/D$, and the number of views over the half-turn $[0, \pi)$ is $\pi/\Delta\theta \ge \pi D/(2d)$. □

For an object of diameter $D = 0.30\,\mathrm{m}$ reconstructed to $d = 1.0\,\mathrm{mm}$, Corollary 17.107 asks for about 470 view angles and a detector pitch of $0.5\,\mathrm{mm}$ — numbers of the order that clinical scanners use, arrived at from the sampling theorem alone. Cormack rediscovered the inversion independently, and, unlike Radon, carried it through to reconstruction algorithms and to a laboratory demonstration on a phantom [1250]; computed tomography followed. It remains the most direct vindication in applied mathematics of a transform inversion formula proved for its own sake.

Remark 17.108 (Tomography of a quantum state). The same inversion reconstructs a quantum state from measurements of rotated quadratures. For a single mode with position x in m and momentum p in kg m/s, the Wigner function

$$W(x,p) = \frac{1}{\pi\hbar} \int_{-\infty}^{\infty} \overline{\psi(x+y)}\, \psi(x-y)\, \mathrm{e}^{2\mathrm{i}py/\hbar}\, \mathrm{d}y \tag{17.123}$$

[710] has the property that its integral along any line in the (x,p) plane is the probability density of the corresponding rotated quadrature — that is, the measured quadrature

distributions are precisely the Radon transform of W. Inverting by Equation (17.122) returns W, hence the state. This is how the states prepared in Chapter 122 are characterised, and the $\hbar$ in Equation (17.123) is what makes the phase-space area element dimensionless: $\mathrm{d}x\,\mathrm{d}p/(2\pi\hbar)$ counts states.

Three sections of this chapter have now met the same pattern from three directions. The Fourier transform diagonalises translation, the Mellin transform dilation, the Radon transform the family of hyperplanes; in each case a hard problem — a differential equation, a convolution over a ratio, an inverse problem — becomes multiplication, and the price is an inversion contour and a discipline about what may be assumed of the data. That is the whole method, and the rest of the treatise uses it.

Part III

Classical Mechanics

Chapter 18

Kinematics

Kinematics is the description of motion before any inquiry into its causes: it fixes the arena — space, time, frames of reference — and the quantities — position, velocity, acceleration — with which every later dynamical statement of this treatise is phrased. The chapter closes with the itinerary equation for constant acceleration, the first law of motion in this book to be confronted directly with experiment, in Chapter 20.

18.1 Foundations

18.1.1 Fundamental concepts

Motion

Motion is the subject of this part, and it is worth fixing at the outset what kind of statement a statement about motion is. When a body is said to move, what is asserted is never a property of the body by itself: it is a relation between that body and a second one, chosen — freely, and for convenience alone — to count as fixed. A passenger asleep in an aircraft seat is at rest with respect to the cabin, moving at some $250\,\mathrm{m/s}$ with respect to the ground, and moving at about $3 \times 10^4\,\mathrm{m/s}$ with respect to the Sun; no experiment settles which of the three is the truth, because all three are. The second body, together with the rulers and clocks tied to it, is the *frame of reference* of Definition 18.7, and every kinematic quantity in this chapter carries an unwritten "with respect to K".

Definition 18.1 (Motion and rest)**.** A particle is *in motion* with respect to a given frame of reference when the point of space it occupies is not the same at every instant, and is *at rest* in that frame otherwise. Quantitatively, motion is the statement that the itinerary map Equation (18.1) is not constant.

Rests on Definitions 18.7 and 18.9.

Kinematics answers the question *how* a particle moves and declines the question *why*: it takes the trajectory as given and extracts from it the quantities — velocity, acceleration — in which the dynamics of Chapter 19 is afterwards phrased. The division is not merely pedagogical. A kinematic statement is a statement about a curve, and is settled by measurements of positions and times alone; a dynamical statement asserts in addition what *produces* the curve, and needs a law.

The mathematics that the description requires is already in place and is not rebuilt here. A trajectory is a differentiable curve in $\mathbb{R}^3$ in the sense of Definition 13.8; the instantaneous quantities are its derivatives with respect to the parameter, in the sense of Definition 7.26; and the tangent, principal normal and binormal of Definitions 13.12, 13.14 and 13.15 supply the natural frame in which an acceleration is read, as Section 18.3.4 shows.

Remark 18.2 (What is assumed here, and what is measured). That motion is relational is an assumption of this chapter, not a theorem of it. Newton took the opposite view, holding that there exists an absolute space with respect to which a body is really at rest [14]; Mach objected that no experiment of Newton's own mechanics can identify that space — Postulate 18.16 makes every inertial frame equivalent — and that an unobservable should be struck from the theory [225]. This treatise follows Mach on the evidence, with one qualification the reader should carry forward: *rotation* is not relational in the same way. A rotating frame is detectable from inside it, by the tilt of a plumb line, by the precession of a pendulum [134], or by the surface of water in a spinning bucket, and Phenomenon 29.53 records the measurements. What uniform translation lacks, and rotation possesses, is an acceleration.

Rests on Postulate 18.16 and Phenomenon 29.53.

Space and time

To study nature and the motions that occur in it, we may take the universe and regard it as a three-dimensional Euclidean manifold; that is, we shall consider the universe to be $\mathbb{R}^3$, to which we shall refer simply as *space*.

We may further consider a parameter with respect to which the state of a physical system evolves. We shall refer to this parameter simply as *time*: a real parameter that indicates the state of a system at any instant.

To determine the location of objects of the macrocosm, such as stars and galaxies, and of the microcosm, such as atoms and electrons, we shall employ physical and mathematical resources to be defined in the course of this treatise, always within the context of space and time.

Particle

Definition 18.3 (Particle)**.** A *particle* is a point of space that can be isolated for its study; that is, a particle may be studied in $\mathbb{R}^3$ as if it were alone in the universe.

For example, to study the motion of a ball we may consider that the ball, together with the Earth, is alone in the universe, since the interactions due to a galaxy located millions of light-years away can be entirely negligible.

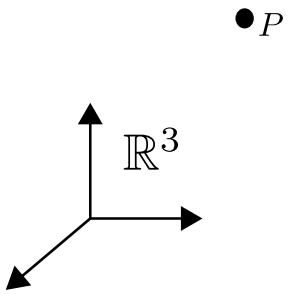


Figure 18.1: A particle: a point of space isolated for its study.

Body

Definition 18.4 (Body)**.** A *body* is a particle with a definite volume in space.

Rests on Definition 18.3.

Remark 18.5. This provisional notion is refined in Chapter 19, where a body is described as a system of many particles occupying a region of space.

Rests on Definition 18.4.

18.1.2 Frames of reference

Free particle

Definition 18.6 (Free particle)**.** A *free particle* is a particle free of any interaction of nature.
Rests on Definition 18.3.

Frame of reference

Definition 18.7 (Frame of reference)**.** A *frame of reference* is the choice of a particle with respect to which the motion of another is observed.
Rests on Definition 18.3.

In order to study distinct frames of reference we shall denote them by K, K', K'', and so on. The choice of a frame of reference allows us to establish the location of the origin of $\mathbb{R}^3$ so as to describe a system mathematically — that is, to introduce distances and angles, which the Euclidean metric permits us to do.

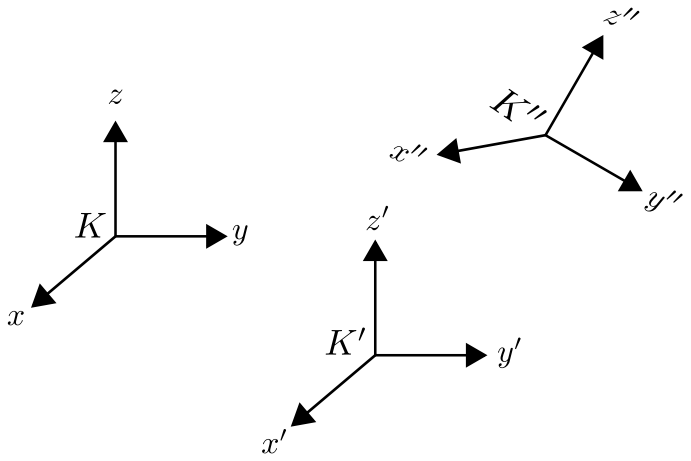


Figure 18.2: A frame of reference: a particle chosen as origin, with respect to which the motion of another particle is observed.

Inertial frame of reference

Definition 18.8 (Inertial frame of reference)**.** An *inertial frame of reference* is a frame of reference located on a free particle.
Rests on Definitions 18.6 and 18.7.

18.1.3 Kinematics of a particle

On the time parameter

We may say that a motion began at the instant t_0 and that, at an arbitrary instant $t \in \mathbb{R}$, the total time of the motion is $\Delta t = t - t_0$.

Just as in space, we may isolate a portion of time in order to study the motion of a particle.

Position

Definition 18.9 (Position)**.** Consider a particle at the point $(x, y, z) \in \mathbb{R}^3$. The *position* of that particle is the vector (as a bilocal object) joining (x, y, z) and the origin, directed towards the particle, which we denote by $\boldsymbol{x}$.
Rests on Definitions 18.3 and 18.7.

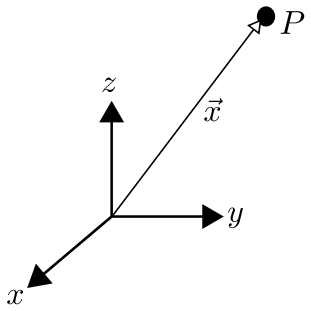


Figure 18.3: The position vector of a particle, joining the origin of the frame of reference to the point occupied by the particle.

We may choose time as a parameter with respect to which the position of a particle varies, that is,

$$\boldsymbol{x} = \boldsymbol{x}(t)\,. \tag{18.1}$$

This equation is called the *itinerary equation* of the particle; it describes a curve in $\mathbb{R}^3$, which we shall call the *trajectory*. In the figure we can see the trajectory of a particle during the time interval $[t_a, t_b]$.

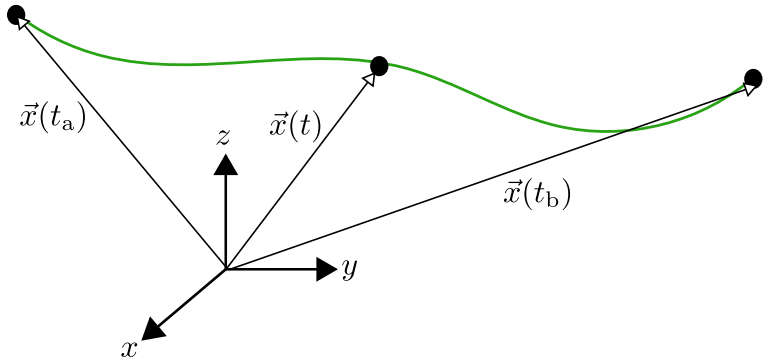


Figure 18.4: The trajectory of a particle: the curve in $\mathbb{R}^3$ described by the itinerary equation $\boldsymbol{x} = \boldsymbol{x}(t)$ over the interval $[t_a, t_b]$.

Velocity

Consider a particle of position $\boldsymbol{x}(t)$ with respect to a given frame of reference. How can we evaluate a change of position between two distinct times t_a and t_b? Define $\Delta\boldsymbol{x} = \boldsymbol{x}(t_b) - \boldsymbol{x}(t_a)$ and $\Delta t = t_b - t_a$. Then the quantity

$$\frac{\Delta\boldsymbol{x}}{\Delta t} = \frac{\boldsymbol{x}(t_b) - \boldsymbol{x}(t_a)}{t_b - t_a} \tag{18.2}$$

is a density, called the *mean velocity*, which indicates how much the position has changed in a given interval of time.

So far we have compared two times. But, given the position of a particle at every time, what quantity can we introduce that indicates the change of position at each instant? To this end, let us recall the definitions of derivative and of vector. When defining a vector on a manifold we concluded that it could not be a bilocal object, and for that purpose we resorted to the definition of the derivative. Something similar happens here: we wish to define a quantity that indicates the change of position at a single point of the time interval — we do not want to compare two points.

We may thus anticipate that the quantity we seek is a vector tangent to a point of the trajectory (conceived as a curve in $\mathbb{R}^3$). The trick that removes the bilocality while retaining

the information corresponding to a single point is to compare two points with the techniques already known, but infinitesimally separated — that is, the derivative.

Definition 18.10 (Velocity)**.** The *velocity* of the particle with respect to the given frame of reference is

$$\boldsymbol{v} = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\,, \tag{18.3}$$

which is indeed a vector in $\mathbb{R}^3$ tangent to the curve $\boldsymbol{x}$, a known result of the theory of curves (Definition 13.9).

Rests on Definitions 7.26, 13.9 and 18.9.

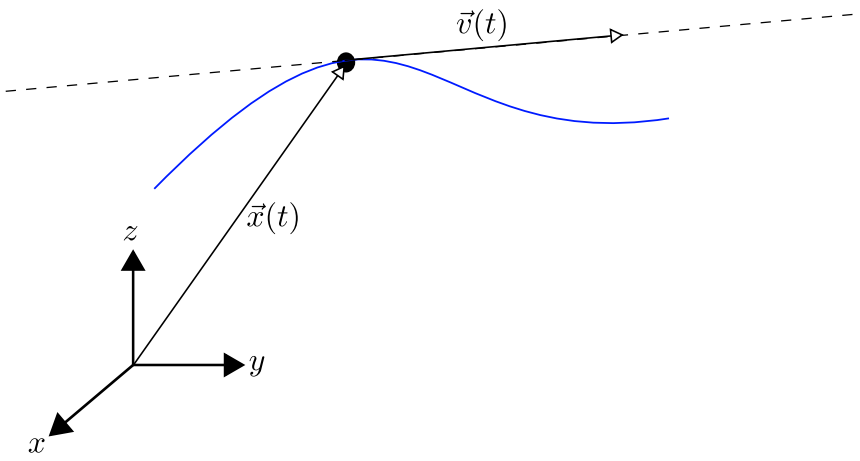


Figure 18.5: The velocity vector: tangent to the trajectory at each of its points.

Acceleration

By the same reasoning as above, we may define a quantity that measures the change of the velocity of a particle in time.

Definition 18.11 (Acceleration)**.** The *acceleration* of the particle with respect to the given frame of reference is

$$\boldsymbol{a} = \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}\,. \tag{18.4}$$

Rests on Definitions 7.26 and 18.10.

18.1.4 The itinerary equation for constant acceleration

Consider a particle with constant acceleration $\boldsymbol{a}_0$. One has

$$\frac{\mathrm{d}^2\boldsymbol{x}}{\mathrm{d}t^2} = \boldsymbol{a}_0\,, \tag{18.5}$$

which is a differential equation from which the itinerary equation is obtained. Considering the initial conditions

$$\boldsymbol{x}(t_0) = \boldsymbol{x}_0\,, \qquad\qquad \boldsymbol{v}(t_0) = \boldsymbol{v}_0\,,$$

and integrating once, we have

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}_0 + \boldsymbol{a}_0(t - t_0)\,,$$

that is, we obtain

$$\boxed{\boldsymbol{v} - \boldsymbol{v}_0 = \boldsymbol{a}_0(t - t_0)}\,. \tag{18.6}$$

Integrating once more, we obtain

$$\boxed{\boldsymbol{x}(t) = \boldsymbol{x}_0 + \boldsymbol{v}_0(t - t_0) + \frac{1}{2}\boldsymbol{a}_0(t - t_0)^2}\,, \tag{18.7}$$

which is known as the *itinerary equation of motion with constant acceleration.*

Phenomenon 18.12 (Uniformly accelerated motion)**.** A body released near the surface of the Earth, with the resistance of the air removed, gains speed in proportion to the elapsed time and falls through a distance proportional to its square; and one and the same constant governs bodies of every mass and composition [4]. The regularity is not a rough one. In a modern ballistic gravimeter the fitted drop reproduces Equation (18.7) at the nanometre level over some 20 cm of fall, and the constant is thereby fixed to about two parts in 10^9 [1906]; the experiment is examined in Chapter 20.

Rests on Equation (18.7) *and Definition 18.11.*

Derivation. *Derives Phenomenon 18.12.*

The kinematical content is the pair of integrations just performed: Equation (18.5) with the stated initial conditions yields Equation (18.6), linear in $t - t_0$, and Equation (18.7), quadratic in it. Nothing else is possible, since prescribing a second derivative fixes a function up to a polynomial of first degree, and the two initial conditions fix that polynomial. What kinematics cannot supply is the observed *universality* of $\boldsymbol{a}_0$: that every body released at the same place receives the same $\boldsymbol{a}_0$ is an experimental input at this stage, and it becomes a statement about mass only in Chapter 19, where the inertial mass of Newton's second law cancels against the gravitational mass in the weight [14]. □

18.2 Galilean relativity

18.2.1 Newton's postulates of absolute space and time

Newton's Scholium to the definitions asserts two absolutes, and this treatise treats them very differently. *Absolute space*, in Newton's words, remains always similar and immovable without relation to anything external, so that a body has a true motion in it as well as the relative motions we measure [14].

Remark 18.13 (Absolute space is stated here, and not adopted). No postulate of absolute space is made in this book. The reason is recorded as Phenomenon 18.17, which is Newton's own fifth corollary: within his mechanics, nothing distinguishes rest in absolute space from uniform motion through it, so the quantity is unmeasurable and the theory is complete without it. What replaces it is Postulate 18.16, which asserts the equivalence of the inertial frames and nothing more. Remark 18.2 states the position and its one qualification — that rotation, unlike uniform translation, *is* locally detectable, which is what makes Newton's bucket a real argument and not a rhetorical one.

Rests on Postulate 18.16 and Phenomenon 18.17.

The postulate on time is a different matter: it is used, constantly and essentially, and it is stated as a postulate of this part.

Postulate 18.14 (Absolute time)**.** Time is an absolute parameter for the whole universe, with respect to any frame of reference.

That is, time elapses equally for every physical system that we can describe in nature [14].

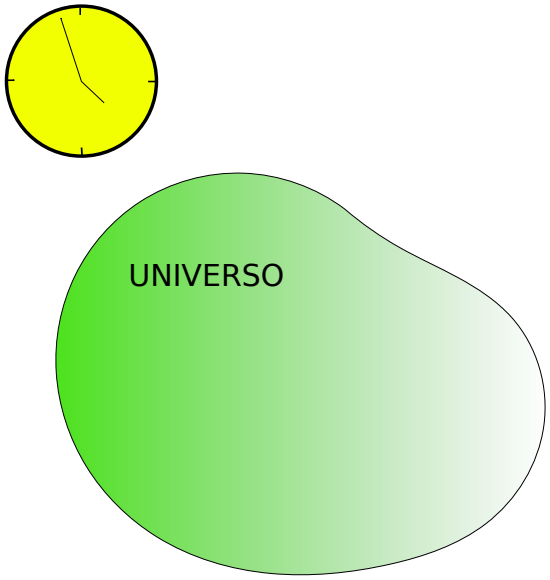


Figure 18.6: Absolute time: a single universal clock, ticking equally for every frame of reference.

Phenomenon 18.15 (A common time, and the speed at which it fails)**.** Clocks of every construction — pendulum, quartz, atomic — that are at rest with respect to one another are observed to keep a common rate, so that a single parameter t serves them all. That, and not more, is the measured content of Postulate 18.14. Clocks in relative motion do *not* keep a common rate: a moving clock runs slow by the factor $\gamma = \left(1 - v^2/c^2\right)^{-1/2}$, as seen in the decay of fast muons [863, 1253], in the transverse Doppler shift of a canal-ray beam [816], and in caesium clocks flown around the world [1476]. Postulate 18.14 is therefore an approximation with a measurable domain of validity, not an identity.

Rests on Postulate 18.14.

Derivation. *Derives Phenomenon 18.15.*

The postulate itself is not derived — it is postulated — but the domain over which it is safe can be quantified. Expanding the factor quoted above for $v \ll c$,

$$\gamma - 1 = \frac{v^2}{2c^2} + O\!\left(\frac{v^4}{c^4}\right), \tag{18.8}$$

so the fractional disagreement between two clocks is second order in their relative speed. At a laboratory speed of 10 m/s it is about 6×10^{-16}; at the 300 m/s of an airliner it is about 5×10^{-13}, which over a two-day circumnavigation accumulates to some tens of nanoseconds — the order of the effect reported in [1476]. The factor γ is derived, rather than quoted, in Chapter 38, and the experiments that exhibit it are collected in Chapter 41. □

18.2.2 Galileo's principle of relativity

We enunciate as a fundamental principle:

Postulate 18.16 (Principle of Galilean relativity)**.** All inertial frames of reference are equivalent.

This is known as *Galileo's principle of relativity*. A consequence is the following: a physical law takes the same form with respect to any inertial frame of reference — or, put differently, the laws of physics do not change according to the observer — a notion known as *Galilean invariance*.

Phenomenon 18.17 (Uniform motion is undetectable from within)**.** No mechanical experiment performed entirely inside a closed cabin discloses whether the cabin is at rest or moving uniformly in a straight line: bodies dropped inside it fall along the same vertical, thrown bodies describe the same arcs, and pendulums keep the same period. Newton records the fact as the fifth corollary to the laws of motion [14]. Whether the same holds of *optical* experiments is a separate question, and an experimental one; the null result of aether-drift interferometry [236] answers it in the affirmative, and it is that extension — not the mechanical statement above — which forces the kinematics of Chapter 38.

Rests on Postulate 18.16 and Definition 18.8.

Derivation. *Derives Phenomenon 18.17.*

Let K' translate with constant velocity $\boldsymbol{V}$ with respect to K, so that the positions the two frames assign to one and the same particle differ by

$$\boldsymbol{x}'(t) = \boldsymbol{x}(t) - \boldsymbol{V}t - \boldsymbol{x}_0 \,. \tag{18.9}$$

Differentiating once gives $\boldsymbol{v}' = \boldsymbol{v} - \boldsymbol{V}$, and differentiating a second time, $\boldsymbol{V}$ being constant,

$$\boldsymbol{a}' = \boldsymbol{a} \,. \tag{18.10}$$

Separations between particles are likewise unchanged, since the terms $\boldsymbol{V}t$ and $\boldsymbol{x}_0$ cancel in a difference $\boldsymbol{x}_1 - \boldsymbol{x}_2$. Every quantity that enters Equation (18.5) — accelerations and relative positions — is therefore identical in the two frames, so an experiment confined to the cabin returns the same numbers whichever frame it is described in. The argument used $\frac{\mathrm{d}\boldsymbol{V}}{\mathrm{d}t} = \boldsymbol{0}$, and it fails at once for a cabin that rotates or accelerates; such motion *is* detectable from inside, as Chapter 29 records. □

18.2.3 Galilean transformations

Postulate 18.16 asserts that all inertial frames are equivalent; it does not say how the coordinates assigned by two of them are related. That relation is fixed by three inputs, each a physical assumption and none of them a mathematical identity: absolute time (Postulate 18.14); the homogeneity of space and of time, that is, that no place and no instant is distinguished; and the isotropy of space, that no direction is distinguished.

Definition 18.18 (Galilean transformation)**.** A *Galilean transformation* is a map of the coordinates $(\boldsymbol{x}, t)$ assigned by an inertial frame K to the coordinates $(\boldsymbol{x}', t')$ assigned by an inertial frame K' of the form

$$\boldsymbol{x}' = R\boldsymbol{x} - \boldsymbol{V}t + \boldsymbol{a} \,, \qquad t' = t + b \,, \tag{18.11}$$

with $R \in \mathrm{SO}(3)$ a rotation, $\boldsymbol{V} \in \mathbb{R}^3$ the velocity of K' with respect to K, $\boldsymbol{a} \in \mathbb{R}^3$ a displacement of the spatial origin and $b \in \mathbb{R}$ a shift of the origin of time. The parameters number $3 + 3 + 3 + 1 = 10$. The pure boost Equation (18.9) is the case $R = \mathbb{1}$, $b = 0$, $\boldsymbol{a} = -\boldsymbol{x}_0$.

Rests on Postulates 18.14 and 18.16 and Definition 18.8.

Remark 18.19 (Why the map is affine, and what is assumed to make it so). The linear-plus-constant form of Equation (18.11) is *assumed* here, not derived, and the assumption should be visible rather than buried. What would derive it is a purely mathematical statement — that a bijection of $\mathbb{R}^3 \times \mathbb{R}$ carrying every uniform rectilinear motion to a uniform rectilinear motion is necessarily affine, the fundamental theorem of affine geometry — and Part II does not at present carry that theorem, so this treatise records the affine form where it is used instead of appealing to a result it has not proved. Physically the assumption is exactly the homogeneity of space and time: were the map not affine, its derivative would differ from

place to place or from instant to instant, which would single out a place or an instant. The same assumption, made in the same place and for the same reason, is what licenses the linear ansatz Equation (38.5) from which the Lorentz transformation is obtained; the two derivations differ in their *second* input, absolute time here and the invariance of c there, and in nothing else.

Rests on Definition 18.18.

Proposition 18.20 (The Galilean group)**.** *The transformations Equation (18.11) form a group of ten parameters under composition, with composition law*

$$(R_2, \boldsymbol{V}_2, \boldsymbol{a}_2, b_2) * (R_1, \boldsymbol{V}_1, \boldsymbol{a}_1, b_1) = (R_2 R_1,\ \boldsymbol{V}_2 + R_2 \boldsymbol{V}_1,\ \boldsymbol{a}_2 + R_2 \boldsymbol{a}_1 - \boldsymbol{V}_2 b_1,\ b_1 + b_2)\ , \tag{18.12}$$

neutral element $(\mathbb{1}, \mathbf{0}, \mathbf{0}, 0)$ *and inverse*

$$(R, \boldsymbol{V}, \boldsymbol{a}, b)^{-1} = \left(R^{-1},\ -R^{-1}\boldsymbol{V},\ -R^{-1}\left(\boldsymbol{a} + \boldsymbol{V} b\right),\ -b\right)\ . \tag{18.13}$$

It is the iterated semidirect product

$$\mathrm{Gal}(3) = \mathbb{R}^4 \rtimes \left(\mathbb{R}^3 \rtimes \mathrm{SO}(3)\right)\ , \tag{18.14}$$

whose normal factor $\mathbb{R}^4$ *consists of the translations of space and time and whose homogeneous factor* $\mathbb{R}^3 \rtimes \mathrm{SO}(3)$ *— boosts and rotations — is isomorphic, as a group, to the Euclidean group* $\mathrm{SE}(3)$ *of Example 4.74.*

Rests on Definitions 4.69 and 18.18 and Example 4.74.

Proof. *Derives Proposition 18.20.*

Apply the transformation labelled 1 and then the one labelled 2. The first gives $\boldsymbol{x}_1 = R_1 \boldsymbol{x} - \boldsymbol{V}_1 t + \boldsymbol{a}_1$ and $t_1 = t + b_1$; substituting these into the second,

$$\begin{aligned} \boldsymbol{x}_2 &= R_2 \boldsymbol{x}_1 - \boldsymbol{V}_2 t_1 + \boldsymbol{a}_2 \\ &= R_2 R_1 \boldsymbol{x} - \left(\boldsymbol{V}_2 + R_2 \boldsymbol{V}_1\right) t + \left(\boldsymbol{a}_2 + R_2 \boldsymbol{a}_1 - \boldsymbol{V}_2 b_1\right)\ , \end{aligned} \tag{18.15}$$

and $t_2 = t + b_1 + b_2$. The result is again of the form Equation (18.11), with $R_2 R_1 \in \mathrm{SO}(3)$ since $\mathrm{SO}(3)$ is a group, and reading off the four slots gives Equation (18.12). Composition of maps is associative, $(\mathbb{1}, \mathbf{0}, \mathbf{0}, 0)$ is plainly neutral, and substituting Equation (18.13) into Equation (18.12) returns the neutral element in each slot, which establishes the group axioms.

For the structure, note first that the translations $T = \{(\mathbb{1}, \mathbf{0}, \boldsymbol{a}, b)\}$ close under Equation (18.12) into $(\boldsymbol{a}_1 + \boldsymbol{a}_2, b_1 + b_2)$, so $T \cong (\mathbb{R}^4, +)$ and is abelian. It is normal: conjugating $(\mathbb{1}, \mathbf{0}, \boldsymbol{a}, b)$ by a general $g = (R, \boldsymbol{V}, \boldsymbol{c}, \beta)$ with Equations (18.12) and (18.13) gives

$$g\,(\mathbb{1}, \mathbf{0}, \boldsymbol{a}, b)\,g^{-1} = (\mathbb{1}, \mathbf{0},\ R\boldsymbol{a} - \boldsymbol{V} b,\ b) \in T\,. \tag{18.16}$$

The complementary subgroup $H = \{(R, \boldsymbol{V}, \mathbf{0}, 0)\}$ closes as $(R_2 R_1, \boldsymbol{V}_2 + R_2 \boldsymbol{V}_1)$, meets T only in the identity, and together with T generates the whole group, since $(R, \boldsymbol{V}, \boldsymbol{a}, b) = (\mathbb{1}, \mathbf{0}, \boldsymbol{a}, b) * (R, \boldsymbol{V}, \mathbf{0}, 0)$. By Theorem 4.72 the group is therefore $T \rtimes H = \mathbb{R}^4 \rtimes H$, which is Equation (18.14) once H is identified. That identification is a comparison of two composition laws: writing the element of H as the pair $(\boldsymbol{V}, R)$, its law reads $(\boldsymbol{V}_2, R_2) * (\boldsymbol{V}_1, R_1) = (\boldsymbol{V}_2 + R_2 \boldsymbol{V}_1, R_2 R_1)$, which is Equation (4.49) verbatim with the boost velocity in the slot the Euclidean translation occupies. Hence $H \cong \mathrm{SE}(3) = \mathbb{R}^3 \rtimes \mathrm{SO}(3)$. □

Proposition 18.21 (Galilean invariants)**.** *Under every transformation Equation (18.11):*

1. *the time interval between two events is unchanged,* $\Delta t' = \Delta t$*;*
2. *simultaneity is absolute, that is,* $\Delta t = 0$ *if and only if* $\Delta t' = 0$*;*

3. the distance between two simultaneous *events is unchanged,* $|\Delta \boldsymbol{x}'| = |\Delta \boldsymbol{x}|$.

Rests on Definition 18.18 and Postulate 18.14.

Proof. *Derives Proposition 18.21.*

For two events labelled A and B, subtract the second equation of Equation (18.11) at A from the same equation at B: the constant b cancels and $\Delta t' = \Delta t$, which is (i), and (ii) is the special case $\Delta t = 0$. Subtracting the first equation likewise gives $\Delta \boldsymbol{x}' = R\,\Delta \boldsymbol{x} - \boldsymbol{V}\Delta t$, so when $\Delta t = 0$ we have $\Delta \boldsymbol{x}' = R\,\Delta \boldsymbol{x}$ and, R being orthogonal, $|R\,\Delta \boldsymbol{x}| = |\Delta \boldsymbol{x}|$, which is (iii). The qualification "simultaneous" in (iii) is not decoration: for $\Delta t \neq 0$ the separation $|\Delta \boldsymbol{x}'|$ depends on $\boldsymbol{V}$, as it must, since the two events then happen at places whose relative position is itself changing. □

Remark 18.22 (Superseded, and in exactly what sense). Equation (18.11) is not the transformation law of nature. The measured law is the Poincaré transformation of Chapter 39, whose homogeneous part is the Lorentz group and under which Proposition 18.21 fails in all three of its parts: time intervals, simultaneity and lengths are all frame dependent. The Galilean group is recovered from the Poincaré group as an Inönü–Wigner contraction in the limit $c \to \infty$ (Example 14.89), which is the precise sense in which the two are related: not as rival descriptions, but as a group and its limit. Since c is finite — $c = 299\,792\,458\,\mathrm{m/s}$ exactly, by the definition of the metre [2333] — the contraction is never exact, and the size of the error is set by v^2/c^2. At the Earth's orbital speed of $2.98 \times 10^4\,\mathrm{m/s}$ that is about 10^{-8}, which is why the whole of Part III may use Equation (18.11) without apology and why the optical experiments of Chapter 37, sensitive at precisely that order, could not.

Rests on Propositions 18.20 and 18.21.

18.2.4 Transformation of velocities and accelerations

Proposition 18.23 (Transformation of velocity and acceleration)**.** *Let K and K' be related by the Galilean transformation Equation (18.11). Then the velocity and the acceleration that the two frames assign to one and the same particle are related by*

$$\boldsymbol{v}' = R\boldsymbol{v} - \boldsymbol{V}\,, \qquad \boldsymbol{a}' = R\boldsymbol{a}\,. \tag{18.17}$$

The acceleration is therefore unaffected by a boost or by a translation of either origin, and is carried by a rotation exactly as a vector is. So too is the relative velocity of two particles, since $\boldsymbol{v}_1' - \boldsymbol{v}_2' = R\,(\boldsymbol{v}_1 - \boldsymbol{v}_2)$ with $\boldsymbol{V}$ cancelling in the difference.

Rests on Definitions 18.10, 18.11 and 18.18.

Proof. *Derives Proposition 18.23.*

By the second equation of Equation (18.11), $\mathrm{d}t'/\mathrm{d}t = 1$, so differentiation with respect to t' and with respect to t are the same operation — this is where Postulate 18.14 enters, and it is the only place it is needed. Differentiating the first equation once, with R, $\boldsymbol{V}$ and $\boldsymbol{a}$ constant,

$$\boldsymbol{v}' = \frac{\mathrm{d}\boldsymbol{x}'}{\mathrm{d}t'} = \frac{\mathrm{d}\boldsymbol{x}'}{\mathrm{d}t} = R\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} - \boldsymbol{V} = R\boldsymbol{v} - \boldsymbol{V}\,,$$

and differentiating a second time removes the constant $\boldsymbol{V}$, leaving $\boldsymbol{a}' = R\boldsymbol{a}$. Both statements are Equation (18.17). The claim about the relative velocity follows by writing Equation (18.17) for each particle and subtracting. □

Corollary 18.24 (Galilean composition of velocities)**.** *If K'' moves with constant velocity $\boldsymbol{V}_2$ with respect to K', and K' with $\boldsymbol{V}_1$ with respect to K, all three sets of axes being parallel, then K'' moves with*

$$\boldsymbol{V} = \boldsymbol{V}_1 + \boldsymbol{V}_2 \tag{18.18}$$

with respect to K: Galilean velocities compose by vector addition, and boosts commute with one another.

Rests on Proposition 18.20.

Proof. *Derives Corollary 18.24.*

Set $R_1 = R_2 = \mathbb{1}$, $\boldsymbol{a}_i = \mathbf{0}$ and $b_i = 0$ in Equation (18.12); the boost slot returns $\boldsymbol{V}_2 + \mathbb{1}\,\boldsymbol{V}_1 = \boldsymbol{V}_1 + \boldsymbol{V}_2$ and every other slot returns its neutral value. The pure boosts are thus a copy of $(\mathbb{R}^3, +)$ inside Gal(3), and addition in $\mathbb{R}^3$ is commutative, so any two boosts commute. □

Remark 18.25 (Where the additivity goes). Two features of Corollary 18.24 fail relativistically, and they fail separately. Additivity is replaced by the composition law Equation (38.29), which returns Equation (18.18) only to leading order and never lets the result exceed c; and commutativity is lost as well, two non-collinear boosts composing to a boost *and* a rotation. Neither failure is visible in the mechanics of this part: the fractional discrepancy in Equation (18.18) is of order $V_1 V_2/c^2$, which for two speeds of 1 km/s is about 10^{-11}.

Rests on Corollary 18.24.

Phenomenon 18.26 (Composition of velocities, and its limit)**.** Two observers in uniform relative motion assign to one and the same body velocities that differ by their relative velocity and by nothing else, and assign to it the *same* acceleration. Everyday mechanics — a ball thrown on a moving ship, a walker on a train — shows no departure from the rule. Light does not obey it. The speed of light in vacuum is found to be the same in frames whose relative velocity is the Earth's orbital velocity of about 29.8 km/s [236], and light in a moving transparent medium is carried along by only a fraction of the medium's speed, not by all of it [133].

Rests on Proposition 18.23 and Corollary 18.24.

Derivation. *Derives Phenomenon 18.26.*

The first claim is Proposition 18.23 with $R = \mathbb{1}$: differentiating Equation (18.9) once and twice gives

$$\boldsymbol{v}' = \boldsymbol{v} - \boldsymbol{V}\,, \qquad \boldsymbol{a}' = \boldsymbol{a}\,, \tag{18.19}$$

the second of which is the invariance of the acceleration already recorded as Equation (18.10).

The second claim is not derived here; it is measured, and the two measurements cited are worth separating. Fizeau sent the two arms of an interferometer through water flowing in opposite directions and found the light dragged not by the full speed u of the water, as Equation (18.19) requires, but by a fraction of it close to Fresnel's $1 - 1/n^2$, which for water at $n = 1.33$ is 0.43 [133]. Michelson and Morley looked for the motion of the Earth through the supposed medium and found no shift at all where Equation (18.19) predicted one of order $v/c \approx 10^{-4}$ in the arrival times [236]. Against those two results, the departure of the exact law of Chapter 38 from Equation (18.19) can be stated once and for all: it is of relative order vV/c^2, which for two speeds of 1 km/s is about 10^{-11} and for light, $v = c$, is of order unity. The failure is thus invisible throughout the mechanics of this part and total in optics; the historical situation this created is the subject of Chapter 37. □

18.3 Circular motion

Circular motion is the simplest motion that is not uniform, and it is the one the rest of mechanics leans on most: a planet held in orbit, a pendulum bob on its arc, an electron in a magnetic field and a point of a spinning body all move, instantaneously, on circles. Everything in this section is kinematics and nothing in it is dynamics — what is computed is the acceleration a circular path *requires*, never the agent that supplies it.

18.3.1 The angular description

Definition 18.27 (Angular position, velocity and acceleration)**.** Let a particle move on a circle of fixed radius r centred on the origin, in the plane $z = 0$, and let $\theta(t)$ be the angle its position makes with a fixed axis of that plane, measured in radians. The *angular velocity* and *angular acceleration* are

$$\omega = \frac{\mathrm{d}\theta}{\mathrm{d}t}\,, \qquad \alpha = \frac{\mathrm{d}\omega}{\mathrm{d}t} = \frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}\,, \tag{18.20}$$

of SI units rad/s and rad/s^2. The arc travelled is $s = r\theta$ and the speed is

$$v = \left|\frac{\mathrm{d}s}{\mathrm{d}t}\right| = r\,|\omega|\ . \tag{18.21}$$

Rests on Definitions 18.9 to 18.11.

Remark 18.28 (The radian carries no dimension). The radian is the coherent SI unit of plane angle and is equal to the number one, being the ratio of two lengths [2333]. It is written in Equation (18.20) because it identifies what is being counted, not because it contributes a dimension: ωr comes out in m/s and $\omega^2 r$ in m/s^2 with no conversion factor. This is also why Equation (18.21) fails if θ is measured in any other angular unit.

Rests on Definition 18.27.

When the motion is periodic it is convenient to name its *period* T, the time of one complete revolution, and its *frequency* $f = 1/T$ in Hz. For constant ω these are

$$T = \frac{2\pi}{|\omega|}\,, \qquad f = \frac{1}{T} = \frac{|\omega|}{2\pi}\,. \tag{18.22}$$

Definition 18.29 (Angular velocity vector)**.** Let $\hat{\boldsymbol{n}}$ be the unit normal to the plane of the motion, oriented by the right-hand rule about the sense of increasing θ. The *angular velocity vector* is $\boldsymbol{\omega} = \omega\,\hat{\boldsymbol{n}}$, and it reproduces the velocity as

$$\boldsymbol{v} = \boldsymbol{\omega} \times \boldsymbol{x}\,. \tag{18.23}$$

Rests on Definitions 18.10 and 18.27.

Remark 18.30 (The angular velocity is an element of $\mathfrak{so}(3)$). Writing Equation (18.23) in components with the basis Equation (14.32) of $\mathfrak{so}(3)$,

$$(\boldsymbol{\omega} \times \boldsymbol{x})_i = \epsilon_{ijk}\omega_j x_k = \omega_k\,(L_k)_{ij}\,x_j\,, \tag{18.24}$$

so the cross product with $\boldsymbol{\omega}$ *is* the antisymmetric matrix $\omega_k L_k$ acting on $\boldsymbol{x}$. The angular velocity is thus not a vector by nature but an element of the Lie algebra of the rotation group, which the three-dimensionality of space allows us to label by three numbers; the finite rotation it generates is $\exp(t\,\omega_k L_k)$, evaluated in closed form by Proposition 14.30. Nothing in this chapter depends on the distinction, but Section 29.1.1 does, and so does every statement about how $\boldsymbol{\omega}$ behaves under a reflection.

Rests on Definition 18.29 and Proposition 14.27.

18.3.2 Uniform circular motion

Phenomenon 18.31 (Uniform circular motion is accelerated motion)**.** A body carried at constant speed v around a circle of radius r is not in uniform motion. It has to be held to the circle by something — a string, a rail, a gravitational attraction — and what is required of that agent is an acceleration of constant magnitude v^2/r directed at every instant towards the centre. Withdraw the agent and the body departs along the tangent, as a stone

released from a sling. The rule v^2/r is Huygens', stated among the theorems on centrifugal force appended to the *Horologium Oscillatorium* [11]; the observation is Newton's point of departure for orbital motion [14].

Rests on Definitions 18.11 and 18.27.

Derivation. *Derives Phenomenon 18.31.*

Take the circle in the plane $z = 0$, with the angular speed $\omega = v/r$ constant, so that the itinerary equation Equation (18.1) reads

$$\boldsymbol{x}(t) = r\,(\cos\omega t,\ \sin\omega t,\ 0)\ . \tag{18.25}$$

Differentiating once, by Equation (18.3),

$$\boldsymbol{v}(t) = r\omega\,(-\sin\omega t,\ \cos\omega t,\ 0)\ ,$$

whose magnitude is the constant $r\omega = v$ and whose scalar product with $\boldsymbol{x}$ vanishes — the velocity is tangent to the circle, as it must be. Differentiating again, by Equation (18.4),

$$\boldsymbol{a}(t) = -r\omega^2\,(\cos\omega t,\ \sin\omega t,\ 0) = -\omega^2\boldsymbol{x}(t)\,, \tag{18.26}$$

so the acceleration points from the body straight at the centre and has the constant magnitude

$$|\boldsymbol{a}| = \omega^2 r = \frac{v^2}{r}\,. \tag{18.27}$$

The result is purely kinematical: it says what acceleration a circular path at constant speed *requires*, and says nothing whatever about what supplies it. Supplying it is the business of Chapter 19, and identifying the supplier for the planets is the business of Chapter 27. □

Two features of Equation (18.26) deserve to be named, because both are constantly misread. First, a motion at constant *speed* is nevertheless an accelerated motion, since it is the velocity *vector* that Definition 18.11 differentiates and that vector is turning. Second, the acceleration Equation (18.26) is directed towards the centre, and there is no outward acceleration anywhere in the calculation: the "centrifugal" term of everyday speech is an artefact of describing the motion in a frame rotating with the body, which is not inertial, and it is derived — with its sign and its companion Coriolis term — in Section 29.5.1.

18.3.3 Non-uniform circular motion

Proposition 18.32 (Tangential and centripetal components)**.** *Let a particle move on the circle of radius r in the plane $z = 0$ with an arbitrary angle $\theta(t)$ of class C^2, and write $\hat{\boldsymbol{r}} = (\cos\theta, \sin\theta, 0)$ and $\hat{\boldsymbol{t}} = (-\sin\theta, \cos\theta, 0)$ for the outward radial and the tangential unit vectors. Then*

$$\boldsymbol{v} = r\omega\,\hat{\boldsymbol{t}}\,, \qquad \boldsymbol{a} = r\alpha\,\hat{\boldsymbol{t}} - r\omega^2\,\hat{\boldsymbol{r}}\,, \tag{18.28}$$

with ω and α as in Equation (18.20). The tangential component $r\alpha = \mathrm{d}v/\mathrm{d}t$ changes the speed; the radial component $-r\omega^2\hat{\boldsymbol{r}}$, of magnitude v^2/r, changes only the direction.

Rests on Definition 18.27 and Phenomenon 18.31.

Proof. *Derives Proposition 18.32.*

The itinerary equation is $\boldsymbol{x}(t) = r\hat{\boldsymbol{r}}(\theta(t))$. Since $\mathrm{d}\hat{\boldsymbol{r}}/\mathrm{d}\theta = \hat{\boldsymbol{t}}$ and $\mathrm{d}\hat{\boldsymbol{t}}/\mathrm{d}\theta = -\hat{\boldsymbol{r}}$, the chain rule (Proposition 7.31) gives $\mathrm{d}\hat{\boldsymbol{r}}/\mathrm{d}t = \omega\hat{\boldsymbol{t}}$ and $\mathrm{d}\hat{\boldsymbol{t}}/\mathrm{d}t = -\omega\hat{\boldsymbol{r}}$. Hence $\boldsymbol{v} = r\omega\hat{\boldsymbol{t}}$, and differentiating once more,

$$\boldsymbol{a} = \frac{\mathrm{d}}{\mathrm{d}t}\left(r\omega\hat{\boldsymbol{t}}\right) = r\alpha\,\hat{\boldsymbol{t}} + r\omega\frac{\mathrm{d}\hat{\boldsymbol{t}}}{\mathrm{d}t} = r\alpha\,\hat{\boldsymbol{t}} - r\omega^2\,\hat{\boldsymbol{r}}\,,$$

which is Equation (18.28). Setting $\alpha = 0$ recovers Equation (18.26), and $|\boldsymbol{a}| = \sqrt{r^2\alpha^2 + r^2\omega^4}$ reduces to Equation (18.27) in that case. □

18.3.4 The intrinsic decomposition of the acceleration

The split of Equation (18.28) into a part along the motion and a part across it is not special to a circle. It holds for every trajectory, with the radius of the circle replaced by the radius of curvature of the curve at the point in question, and it is the reason the curve theory of Section 13.3.1 was set up.

Proposition 18.33 (Intrinsic decomposition of the acceleration)**.** *Let $\boldsymbol{x}(t)$ be a trajectory of class C^2 whose speed $v = |\mathrm{d}\boldsymbol{x}/\mathrm{d}t|$ is nowhere zero, let s be its arc length Equation (13.35), and let $\hat{T}$, $\hat{N}$, $\hat{B}$ be the Frenet frame of Definitions 13.12, 13.14 and 13.15, with curvature $\kappa > 0$. Then*

$$\boldsymbol{v} = v\,\hat{T} \tag{18.29}$$

and

$$\boldsymbol{a} = \frac{\mathrm{d}v}{\mathrm{d}t}\,\hat{T} + \kappa v^2\,\hat{N} = \frac{\mathrm{d}v}{\mathrm{d}t}\,\hat{T} + \frac{v^2}{\rho}\,\hat{N}\,, \tag{18.30}$$

where $\rho = 1/\kappa$ is the radius of curvature. In particular $\boldsymbol{a}\cdot\hat{B} = 0$: the acceleration always lies in the osculating plane, its component across the motion is directed towards the centre of curvature, and it never has a component out of that plane.

Rests on Definitions 13.12, 13.14 and 18.11.

Proof. *Derives Proposition 18.33.*

By Equation (13.36) the arc length satisfies $\mathrm{d}s/\mathrm{d}t = v$, and by Equation (13.38) the unit tangent is $\hat{T} = \mathrm{d}\boldsymbol{x}/\mathrm{d}s$. The chain rule (Proposition 7.31) therefore gives Equation (18.29),

$$\boldsymbol{v} = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}t} = v\,\hat{T}\,.$$

Differentiating that product and using the chain rule once more on $\hat{T}$,

$$\boldsymbol{a} = \frac{\mathrm{d}}{\mathrm{d}t}\left(v\hat{T}\right) = \frac{\mathrm{d}v}{\mathrm{d}t}\,\hat{T} + v\,\frac{\mathrm{d}\hat{T}}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\mathrm{d}v}{\mathrm{d}t}\,\hat{T} + v^2\,\frac{\mathrm{d}\hat{T}}{\mathrm{d}s}\,, \tag{18.31}$$

From Equations (13.36) and (13.38) (differentiating $\boldsymbol{v} = v\hat{T}$ and converting $\mathrm{d}\hat{T}/\mathrm{d}t$ to $\mathrm{d}\hat{T}/\mathrm{d}s$ with $\mathrm{d}s/\mathrm{d}t = v$).

and the first Frenet equation Equation (13.43), $\mathrm{d}\hat{T}/\mathrm{d}s = \kappa\hat{N}$, turns Equation (18.31) into Equation (18.30). The last claim follows because $\hat{T}, \hat{N}, \hat{B}$ are orthonormal by Definition 13.15, so the expansion Equation (18.30) has no $\hat{B}$ component to begin with. □

Remark 18.34 (The two ways an acceleration can be zero)*.* Equation (18.30) makes the classification immediate. The tangential term vanishes when the speed is constant, and the normal term vanishes when the curvature is zero, that is when the path is straight; only when both vanish is the motion unaccelerated, which is uniform motion in a straight line. Uniform circular motion is the second case made non-trivial: $\mathrm{d}v/\mathrm{d}t = 0$ and $\kappa = 1/r$ constant, so Equation (18.30) returns $|\boldsymbol{a}| = v^2/r$, which is Equation (18.27) again, now derived without ever choosing coordinates. Where $\kappa = 0$ the principal normal Equation (13.42) is undefined and the statement degenerates to $\boldsymbol{a} = (\mathrm{d}v/\mathrm{d}t)\,\hat{T}$, which is correct as it stands.

Rests on Proposition 18.33.

18.3.5 Magnitudes

Centripetal accelerations in nature and in the laboratory span some ten orders of magnitude, and it is worth having a few of them in mind.

The first row is not a curiosity: it is measured, and measured easily. A body at rest on the equator is in circular motion, so the free-fall acceleration read by a gravimeter there is smaller than the gravitational attraction by the $3.39 \times 10^{-2}\,\mathrm{m/s^2}$ of Table 18.1 — against an

Table 18.1: Centripetal accelerations $\omega^2 r$ for four circular motions, computed from Equation (18.27). The Earth's spin rate is the sidereal one and r its equatorial radius; the orbital row uses the mean Sun–Earth distance; the car travels at 28 m/s on a curve of 300 m radius; the rotor is a laboratory ultracentrifuge of 5 cm radius. For comparison, the conventional standard acceleration of free fall is $g_0 = 9.806\,65\,\mathrm{m/s^2}$ exactly [2333].

motion	r	ω	$\omega^2 r$
Earth, spin at the equator	$6.378 \times 10^{6}\,\mathrm{m}$	$7.292 \times 10^{-5}\,\mathrm{rad/s}$	$3.39 \times 10^{-2}\,\mathrm{m/s^2}$
Earth, orbit about the Sun	$1.496 \times 10^{11}\,\mathrm{m}$	$1.991 \times 10^{-7}\,\mathrm{rad/s}$	$5.93 \times 10^{-3}\,\mathrm{m/s^2}$
Car on a highway curve	$3.0 \times 10^{2}\,\mathrm{m}$	$9.3 \times 10^{-2}\,\mathrm{rad/s}$	$2.6\,\mathrm{m/s^2}$
Ultracentrifuge rotor	$5.0 \times 10^{-2}\,\mathrm{m}$	$1.05 \times 10^{4}\,\mathrm{rad/s}$	$5.5 \times 10^{6}\,\mathrm{m/s^2}$

instrumental accuracy of about $2 \times 10^{-8}\,\mathrm{m/s^2}$ [1906], which resolves the effect a million times over. The observed equator-to-pole difference is larger still, $5.2 \times 10^{-2}\,\mathrm{m/s^2}$, the remainder coming from the equatorial bulge that the same rotation produces; Phenomenon 29.55 treats both together, and Equation (29.68) is its quantitative form.

The itinerary equation for constant acceleration, Equation (18.7), is the first quantitative law of this treatise, and it is not left as a formal statement: in Chapter 20 it is confronted with experiment — from Galileo's inclined planes to the modern ballistic gravimeters — with every assumption behind it declared and tested.

Chapter 19

Newtonian Dynamics

When we seek what is most fundamental — an idea that governs the whole of mechanics — we meet the concept of force, as the motive of the motion of a particle. In this respect, the laws formulated by Isaac Newton are essential [14]. Even so, as we shall see, there exist further formulations, resting on other axioms, from which Newton's laws can be derived (Chapters 21 and 22).

A particle can translate; and if it has a volume — that is, if it is a body — it can also rotate. We shall introduce the axioms and definitions that allow us to understand the mechanics of a particle and of a body.

19.1 Inertia and mass

Inertia

Consider a particle. On what does its kinematics depend when nature interacts with it? We may suppose a property of the particle that plays a role in the motive of its motion, simply by virtue of its being in nature. We postulate:

Postulate 19.1 (Inertia)**.** If a particle moves in uniform rectilinear motion with respect to some inertial frame of reference, it will continue to do so unless it interacts with nature.

Rests on Definitions 18.3 and 18.8.

This is a property possessed by every particle, known as *inertia*.

Interaction

Definition 19.2 (Interaction)**.** An *interaction* is the motive by which a particle changes its position $\boldsymbol{x}(t) = \mathbf{0}\ \forall t$ to another position $\boldsymbol{x}'$.

Rests on Definition 18.3 and Postulate 19.1.

Mass

One particle can have more inertia than another.

Definition 19.3 (Mass)**.** The *mass* is the scalar quantity that quantifies inertia, in the following sense: the mass of a particle is proportional to how much it costs nature to move it.

Rests on Postulate 19.1 and Definition 18.3.

Remark 19.4. The Fundamentos formulary accordingly sharpens the notion of a particle introduced in Definition 18.3: a particle is a point of space to which a mass m is assigned.

19.2 Newton's laws and the Newtonian formulation: translation

19.2.1 Linear momentum and Newton's first law

Consider the following fact of nature. When a very heavy truck moves at great speed, we feel that it carries great power, implacable in any collision (let us not forget that if the truck moves at great speed it is because our frame of reference is outside the truck; if we were inside it and did not notice what lies outside, for us it would be at rest). Now, if we have a light piece of wood that moves very slowly, we feel that it carries less power than the truck.

On the other hand, we may consider the crossed case. What would happen if the truck moved very slowly while the piece of wood travelled at extravagant speeds close to the speed of light? Clearly we must introduce a quantity that manifests the power of which we speak — one that unifies the mass and the velocity associated with a given motion.

Definition 19.5 (Linear momentum)**.** The *linear momentum* of a particle of mass m moving with velocity $\boldsymbol{v}$ with respect to some frame of reference is

$$\boxed{\boldsymbol{p} = m\boldsymbol{v}}\,. \tag{19.1}$$

Rests on Definitions 18.10 and 19.3.

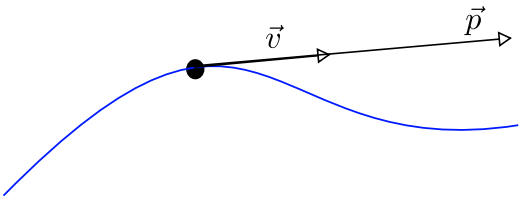


Figure 19.1: Linear momentum: the quantity unifying the mass of a particle with the velocity of its motion.

We enunciate as a fundamental principle:

Postulate 19.6 (Newton's first law)**.** The linear momentum of a free particle is constant, with respect to any inertial frame of reference.

Rests on Definitions 18.8 and 19.5.

Phenomenon 19.7 (Persistence of uniform motion)**.** A body on which nothing acts keeps its velocity: neither its speed nor its direction changes, and no agent whatever is needed to maintain the motion. The fact is never directly visible on Earth, where friction and air resistance are never absent; it is exposed by removing them by degrees — a ball on a horizontal plane runs the farther the smoother the plane, which is Galileo's argument [4] — and it is seen outright in a coasting spacecraft. The contrary Aristotelian expectation, that motion requires a mover and ceases with it, is the oldest falsified theory in mechanics.

Rests on Postulate 19.6 and Equation (19.1)*.*

Derivation. Derives Phenomenon 19.7.

By Postulate 19.6 the linear momentum of a free particle is constant. If in addition the mass of the particle is constant, then from Equation (19.1)

$$\boldsymbol{v} = \frac{\boldsymbol{p}}{m}$$

is constant as well, so the velocity keeps both magnitude and direction, and the trajectory is a straight line traversed uniformly. The observation therefore tests two things at once,

and separating them matters: the constancy of $\boldsymbol{p}$, which is the postulate, and the constancy of m, which fails for a rocket and for a body accumulating or shedding matter. It is only with both that Equation (19.3) collapses to Equation (19.4) and the familiar statement follows. □

19.2.2 Force and Newton's second law

Until now we have used the concept of an interaction of nature with particles. But what, at bottom, is an interaction? We must quantify this concept in order to introduce fundamental equations that govern the entire dynamics of a system. An interaction certainly perturbs the inertia of a particle, and so we introduce as a fundamental postulate:

Postulate 19.8 (Newton's second law)**.** The sum of all the interactions upon a particle is the measure of how much its linear momentum fails to be conserved.

Rests on Definition 19.5 and Postulate 19.6.

We shall call the total *force* upon a particle the sum of all the interactions upon it, and we shall denote it by $\boldsymbol{F}$. Newton's second law is then mathematically expressed by

$$\boxed{\boldsymbol{F} = \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}}\,. \tag{19.2}$$

Newton's second law is fundamental because it provides the equations of motion. That is, given the sum of the forces acting upon a particle, we can compute its trajectory. Moreover, we can finally conclude the following: a particle will continue its rectilinear motion unless a nonvanishing total force acts upon it — this with respect to an inertial frame of reference. The figure illustrates a particle that moves along a straight line and deviates from its trajectory.

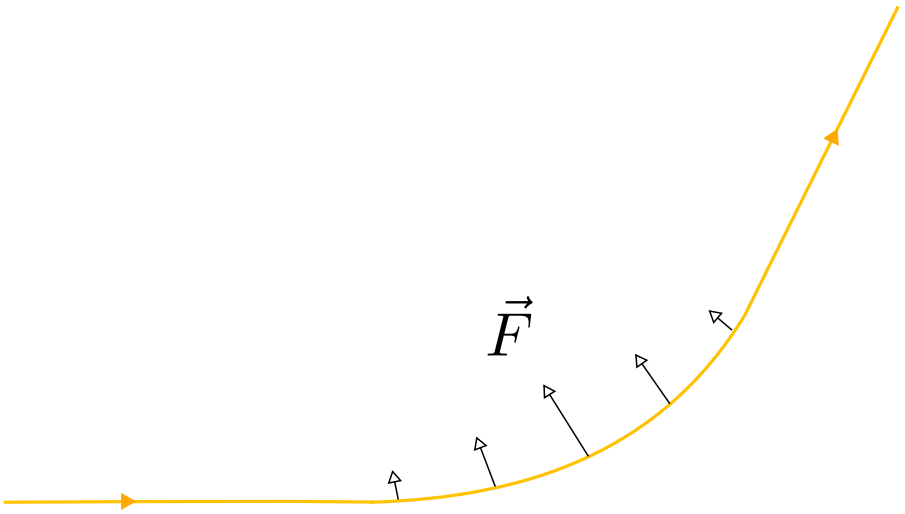


Figure 19.2: A particle moving in a straight line is deflected from its trajectory by a nonvanishing total force.

From Equation (19.1) we see that

$$\boxed{\boldsymbol{F} = \frac{\mathrm{d}m}{\mathrm{d}t}\boldsymbol{v} + m\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}}\,. \tag{19.3}$$

If the mass of the particle is constant, then from Equation (19.3) we see that

$$\boxed{\boldsymbol{F} = m\boldsymbol{a}}\,. \tag{19.4}$$

Remark 19.9. The Fundamentos formulary states the second law in the summary form: the change with respect to time of the linear momentum of a particle is proportional to its mass and to its acceleration, $\boldsymbol{F} = \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = m\boldsymbol{a}$.

Phenomenon 19.10 (Force, mass and acceleration)**.** For a body of fixed mass the acceleration produced is proportional to the applied force and parallel to it; for a fixed applied force it is inversely proportional to the mass. The classical demonstration is Atwood's machine, in which two masses $m_1 > m_2$ hang from the ends of a cord over a light pulley and the pair is observed to move with a uniform acceleration smaller than that of free fall, in the ratio $(m_1 - m_2)/(m_1 + m_2)$. Bringing the two masses close in value makes the acceleration as small as the timing apparatus requires without changing its constancy — the same dilution trick as Galileo's inclined plane [4], and the arrangement by which the second law was first checked quantitatively in the laboratory [14].

Rests on Equation (19.4) *and Postulate 19.8.*

Derivation. *Derives Phenomenon 19.10.*

Idealize the cord as inextensible and massless and the pulley as light and frictionless, so that the tension T is the same throughout and the two masses have accelerations of equal magnitude a and opposite sense. Applying Equation (19.4) to each mass along the vertical, with the descending mass taken positive,

$$m_1 a = m_1 g - T\,, \qquad m_2 a = T - m_2 g\,.$$

Adding the two eliminates the unknown tension, which is the point of the arrangement, and gives

$$a = \frac{m_1 - m_2}{m_1 + m_2}\, g\,, \qquad T = \frac{2 m_1 m_2}{m_1 + m_2}\, g\,. \tag{19.5}$$

The acceleration is constant, so the descent obeys the itinerary equation for constant acceleration of Chapter 18 and can be timed over a measured distance. Two limits check the result: for $m_2 = 0$ it returns $a = g$, free fall; and for $m_1 = m_2$ it returns $a = 0$, equilibrium. What the experiment establishes is that a single number m suffices — the same m that measures the weight mg also measures the resistance to acceleration in Equation (19.4) — which is the assumption whose exactness is tested in Chapter 42. □

Covariance

A law of motion earns the name only if it says the same thing in every inertial frame; were it otherwise, Postulate 18.16 would fail, and an experiment performed inside a closed cabin could disclose the cabin's velocity. That Equation (19.4) passes this test is the content of the present heading. It is worth doing carefully, because the demonstration constrains the admissible *forces* a good deal more than it constrains the equation.

The transformation itself belongs to kinematics and is stated there: Definition 18.18 defines the Galilean transformation

$$\boldsymbol{x}' = R\boldsymbol{x} - \boldsymbol{V}t + \boldsymbol{a}\,, \qquad t' = t + b\,,$$

with $R \in \mathrm{SO}(3)$, $\boldsymbol{V}$ a relative velocity in m/s, $\boldsymbol{a}$ a displacement in m and b a time offset in s, and Proposition 18.20 proves that the ten of them close into the Galilei group with the composition law Equation (18.12). Both are used below without being restated.

Remark 19.11 (Mass is a Galilean scalar). The mass of a particle is the same number in every inertial frame, $m' = m$: mass is a scalar by Definition 19.3, and nothing in Definition 18.18 acts on it. This is an assumption about what mass *is*, not a theorem, and it is exactly the assumption relativity abandons — see Chapter 40, where the inertia of a body depends on its speed.

Rests on Definitions 18.18 and 19.3.

Theorem 19.12 (Galilean covariance of the second law)**.** *Let the mass of a particle be constant and let the total force upon it transform as* $\boldsymbol{F}' = R\boldsymbol{F}$ *under Equation* (18.11)*. Then Equation* (19.4) *holds in* K *if and only if* $\boldsymbol{F}' = m\boldsymbol{a}'$ *holds in* K'*.*
Rests on Definition 18.18, Equation (19.4)*, and Postulate 18.14.*

Proof. *Derives Theorem 19.12.*

By Postulate 18.14 the two frames measure the same time increments, $\mathrm{d}t' = \mathrm{d}t$, so differentiating with respect to t' is differentiating with respect to t. Differentiating the second of Equation (18.11) once, R, $\boldsymbol{V}$ and $\boldsymbol{b}$ being constant,

$$\boldsymbol{v}' = R\boldsymbol{v} - \boldsymbol{V}\,, \tag{19.6}$$

and a second time,

$$\boldsymbol{a}' = R\boldsymbol{a}\,. \tag{19.7}$$

Multiplying Equation (19.7) by the constant mass and subtracting the assumed transformation of the force,

$$\boldsymbol{F}' - m\boldsymbol{a}' = R\,(\boldsymbol{F} - m\boldsymbol{a})\,,$$

and R is invertible, so the left-hand side vanishes exactly when the right-hand side does. □

The hypothesis on the force is not a formality, and it is the half of the statement the source note omits: Equation (19.4) is covariant *provided* the force is the kind of object that transforms as a vector and does not depend on the absolute position, the absolute time or the absolute velocity of the particle. That is a genuine restriction on the interactions nature is allowed to have, and it is satisfied by the ones it does have.

Proposition 19.13 (Which forces are admissible)**.** *Let the force on particle* s *of a system be a sum of pair terms*

$$\boldsymbol{F}_s = \sum_{r \neq s} \boldsymbol{f}\,(\boldsymbol{x}_s - \boldsymbol{x}_r,\, \boldsymbol{v}_s - \boldsymbol{v}_r)\,, \tag{19.8}$$

with $\boldsymbol{f}$ *equivariant under rotations,* $\boldsymbol{f}\,(R\boldsymbol{u}, R\boldsymbol{w}) = R\,\boldsymbol{f}\,(\boldsymbol{u}, \boldsymbol{w})$ *for every* $R \in \mathrm{SO}(3)$*. Then* $\boldsymbol{F}'_s = R\boldsymbol{F}_s$*, so Theorem 19.12 applies to every particle of the system.*
Rests on Definition 18.18 and Theorem 19.12.

Proof. *Derives Proposition 19.13.*

The displacement $\boldsymbol{b}$ and the term $\boldsymbol{V}t$ cancel in a difference of positions, and $\boldsymbol{V}$ cancels in a difference of velocities: by Equations (18.11) and (19.6),

$$\boldsymbol{x}'_s - \boldsymbol{x}'_r = R\,(\boldsymbol{x}_s - \boldsymbol{x}_r)\,, \qquad \boldsymbol{v}'_s - \boldsymbol{v}'_r = R\,(\boldsymbol{v}_s - \boldsymbol{v}_r)\,.$$

Equivariance then returns each pair term as R times its unprimed value, and the finite sum Equation (19.8) with it. □

Corollary 19.14 (Central pair forces are admissible)**.** *A pair force of the form* $\boldsymbol{f}\,(\boldsymbol{u}) = f(|\boldsymbol{u}|)\,\boldsymbol{u}/|\boldsymbol{u}|$ *satisfies the hypothesis of Proposition 19.13, a rotation preserving both* $|\boldsymbol{u}|$ *and the direction of* $\boldsymbol{u}$ *relative to itself. Newtonian gravitation and the Coulomb force are of this kind, and so is a Hooke spring joining two particles; the central case is developed in Chapter 27.*
Rests on Proposition 19.13.

Remark 19.15 (Covariance is not invariance)*.* What the two frames share is the *form* of the law, not the values of the quantities in it. Under a pure boost the momentum of Equation (19.1) changes by $-m\boldsymbol{V}$ and the kinetic energy of Equation (19.18) becomes

$$T' = T - m\,\boldsymbol{v}\cdot\boldsymbol{V} + \tfrac{1}{2}mV^2\,, \tag{19.9}$$

neither of which is unchanged. What is boost-invariant is the acceleration Equation (19.7), together with the separations between particles and their relative velocities — exactly the quantities Equation (19.8) is built from. The distinction returns in Section 19.5.5: a quantity can be conserved in every inertial frame without taking the same value in any two of them.

Remark 19.16 (Variable mass breaks the covariance of the momentum form). Theorem 19.12 was stated for constant mass, and the restriction cannot be lifted. Allowing m to depend on time and transforming under a pure boost,

$$\frac{\mathrm{d}\boldsymbol{p}'}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left[m\left(\boldsymbol{v}-\boldsymbol{V}\right)\right] = \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} - \frac{\mathrm{d}m}{\mathrm{d}t}\,\boldsymbol{V}\,, \tag{19.10}$$

so the two frames disagree by $-\dot{m}\,\boldsymbol{V}$. No transformation law for the force can absorb that term, because it depends on the arbitrary boost velocity $\boldsymbol{V}$ and not on anything physical. Equation (19.2) is therefore Galilean covariant only for a system whose mass is constant. A rocket is not a counterexample but a lesson in bookkeeping: the closed system is the rocket *together with* the exhaust it has already ejected, its total mass is constant, and Equation (19.2) applies to that. The same caveat was recorded in the derivation of Phenomenon 19.7.

Remark 19.17 (Absolute time is doing the work). The proof of Theorem 19.12 leaned on Postulate 18.14 at the very first step: only because t' and t differ by a constant does the second derivative in K' agree with the second derivative in K. If instead the time of K' is allowed to depend on position — as the Lorentz transformations of Chapter 38 require it to — then Equation (19.7) is false, and no redefinition of the force restores Equation (19.4). The mechanics built on that other group is Chapter 40, and the disagreement between the two is of relative order v^2/c^2, which Equation (18.8) bounds.

The observational counterpart of Theorem 19.12 is Phenomenon 18.17: mechanical experiments in a closed cabin do not disclose its uniform velocity. Nothing in the theorem is tested by that observation alone, however, since the covariance was built into the postulates; what is tested is the conjunction of Postulate 18.14 with the assumption that the real forces are of the admissible form Equation (19.8).

19.3 Newton's laws and the Newtonian formulation: rotation

19.3.1 Angular momentum

Let us note the following fact of nature. When we try to push a door, we can clearly observe that it costs much less to push it from a point far from the hinge (for example, the handle) than from a point closer to it. What is happening? We are applying the same force, yet we do not achieve the same motion. In other words, to succeed in opening the door we must apply more or less force depending on where we do it.

This suggests the following. To study the motion of a body we must introduce an analogue of the translational Newton laws, but with a quantity that carries information about the location at which there is a surge of momentum — that is, a quantity that is a function of the position and the momentum of a particle.

Thus we define:

Definition 19.18 (Angular momentum)**.** The *angular momentum* of a particle of position $\boldsymbol{x}$ and momentum $\boldsymbol{p}$ is

$$\boldsymbol{L} = \boldsymbol{x} \times \boldsymbol{p}\,, \tag{19.11}$$

with units of $\mathrm{kg\,m^2/s}$.

Rests on Definitions 18.3 and 19.5.

The principle the source manuscript was reaching for at this point is the rotational counterpart of Postulate 19.6.

Postulate 19.19 (Newton's first law, rotational case)**.** The angular momentum of a free particle, taken about any point at rest in an inertial frame, is constant.

Rests on Definition 19.18 and Postulate 19.6.

Proposition 19.20 (For one particle the rotational first law is not independent)**.** *Postulate 19.19 follows from Postulate 19.6 and Definition 19.18; it is therefore a theorem for a single particle, and only becomes an independent statement for a system of interacting particles (Section 19.5.4).*

Rests on Postulate 19.6 and Definition 19.18.

Proof. *Derives Proposition 19.20.*

Let $\boldsymbol{x}$ be measured from the fixed point. Differentiating Equation (19.11) and using the product rule for the cross product,

$$\frac{\mathrm{d}\boldsymbol{L}}{\mathrm{d}t} = \boldsymbol{v} \times \boldsymbol{p} + \boldsymbol{x} \times \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}\,.$$

The first term vanishes because $\boldsymbol{p} = m\boldsymbol{v}$ is parallel to $\boldsymbol{v}$ by Equation (19.1), and the second vanishes because $\boldsymbol{p}$ is constant for a free particle by Postulate 19.6. Hence $\boldsymbol{L}$ is constant. The argument uses only that the point about which $\boldsymbol{L}$ is taken does not move; about a point in arbitrary motion the statement is false, since then $\mathrm{d}\boldsymbol{x}/\mathrm{d}t$ is no longer $\boldsymbol{v}$. □

For a system the situation is different and the postulate does real work: the internal forces of a system need not cancel in the torque sum merely because they cancel in the force sum, and what is required of them is stated as Postulate 19.61 below.

Phenomenon 19.21 (Equal areas in equal times)**.** A body moving under a force always directed towards one fixed centre sweeps out equal areas about that centre in equal times: it moves quickest where it is nearest and slowest where it is farthest, in exact compensation. This is Kepler's second law, proved from the laws of motion as the first proposition of the *Principia* [14], and it holds for every central force whatever, not only for the inverse square. The Earth's orbit is eccentric enough for the effect to be conspicuous: Newcomb's reduction of the inner-planet observations gives its eccentricity as $e = 0.01675$ at the epoch 1900.0, slowly decreasing [267]. With a semi-major axis of one astronomical unit, $a = 1.495\,978\,707 \times 10^{11}\,\mathrm{m}$ by definition [2333], the perihelion and aphelion distances differ by $2ae \approx 5.0 \times 10^{9}\,\mathrm{m}$, and the Earth runs correspondingly faster in January than in July.

Rests on Postulate 19.19.

Derivation. *Derives Phenomenon 19.21.*

Let the force be central about the origin, $\boldsymbol{F} = F(\boldsymbol{x})\,\boldsymbol{x}/\,|\boldsymbol{x}|$. Then $\boldsymbol{x} \times \boldsymbol{F} = \boldsymbol{0}$, so by Equations (19.14) and (19.15) the angular momentum about that centre is constant. The area swept by the radius vector in a time $\mathrm{d}t$ is that of the triangle spanned by $\boldsymbol{x}$ and $\mathrm{d}\boldsymbol{x}$,

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \frac{1}{2}\,|\boldsymbol{x} \times \boldsymbol{v}| = \frac{|\boldsymbol{L}|}{2m}\,, \tag{19.12}$$

of SI unit m^2/s, which is constant with $|\boldsymbol{L}|$. Since $\boldsymbol{L}$ is moreover fixed in direction, the motion stays in the plane through the centre perpendicular to it. At the two apses the velocity is perpendicular to the radius, so $|\boldsymbol{L}| = mrv$ there and the speeds are in the inverse ratio of the distances,

$$\frac{v_{\mathrm{peri}}}{v_{\mathrm{apo}}} = \frac{r_{\mathrm{apo}}}{r_{\mathrm{peri}}} = \frac{1+e}{1-e}\,, \tag{19.13}$$

which for the eccentricity quoted above is 1.03407. That figure is a prediction computed from e, not a second measurement: the independent test of the areal law, and of the

inverse-square force behind the Earth's particular value of e, belongs to Chapter 27. The derivation assumed only that the force points along $\boldsymbol{x}$; it says nothing at all about how the force depends on distance, which is why the areal law survived every revision of the force law between Kepler and Newton. □

19.3.2 Torque

Analogously to Newton's second law in the translational case, we introduce as a fundamental postulate:

Postulate 19.22 (Newton's second law, rotational case)**.** The sum of all the interactions upon a body at a given point is the measure of how much its angular momentum fails to be conserved.

Rests on Definition 19.18 and Postulate 19.8.

We shall call the total *torque*, or *moment of force*, upon a body the sum of all the forces applied at a given point, and we shall denote it by $\boldsymbol{\tau}$. Newton's second law in the rotational case is then expressed as

$$\boxed{\boldsymbol{\tau} = \frac{\mathrm{d}\boldsymbol{L}}{\mathrm{d}t}}\,, \tag{19.14}$$

with units of $\mathrm{N\,m} = \mathrm{kg\,m^2/s^2}$, dimensionally those of an energy.

Substituting the definition of angular momentum into Equation (19.14), we have

$$\begin{aligned}\boldsymbol{\tau} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\boldsymbol{x}\times\boldsymbol{p}\right)\\ &= \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\times\boldsymbol{p} + \boldsymbol{x}\times\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}\\ &= \boldsymbol{v}\times\boldsymbol{p} + \boldsymbol{x}\times\boldsymbol{F}\,,\end{aligned}$$

where the last step uses Equation (19.2); and since the velocity of a particle is always parallel to its linear momentum, we find the relation

$$\boxed{\boldsymbol{\tau} = \boldsymbol{x}\times\boldsymbol{F}}\,. \tag{19.15}$$

This is a very important relation, for it yields the following conclusion: the torque is a quantity that manifests the total force applied at the place where the particle is. We can use this equation to describe the motion of a body when we apply a force at a specific point of it, as in the example of the door: it costs us more or less to move it depending on where we apply the force.

Phenomenon 19.23 (The lever arm)**.** The same force applied to a door opens it easily at the handle and scarcely at all near the hinge, and a rigid bar pivoted at a point balances two weights when, and only when, the products of weight and distance from the pivot agree on the two sides — Archimedes' law of the lever, the oldest quantitative law in mechanics and the principle of every balance, crowbar and wheelbarrow. What governs rotation is therefore not the force alone but the force together with its point of application, and specifically the product of the force with the perpendicular distance from the axis to its line of action: a force directed straight at the axis has no such distance and never turns anything, which is why pulling a door towards its hinge is futile [14].

Rests on Equations (19.14) and (19.15).

Derivation. *Derives Phenomenon 19.23.*

Equation (19.15) gives $\boldsymbol{\tau} = \boldsymbol{x}\times\boldsymbol{F}$. Let $\boldsymbol{x}$ be the position of the point of application relative to the axis and θ the angle between $\boldsymbol{x}$ and $\boldsymbol{F}$. The magnitude of the torque is then

$$|\boldsymbol{\tau}| = |\boldsymbol{x}|\;|\boldsymbol{F}|\sin\theta = d\;|\boldsymbol{F}|\;, \qquad d := |\boldsymbol{x}|\sin\theta\,, \tag{19.16}$$

where d is the perpendicular distance from the axis to the line of action — the lever arm. It grows in proportion to the distance at which the force is applied, which is the door; and it vanishes when $\boldsymbol{F}$ is parallel or antiparallel to $\boldsymbol{x}$, so that $\sin\theta = 0$, which is the futile pull towards the hinge. For the balanced bar, the angular momentum about the pivot does not change, so by Equation (19.14) the torques must cancel; with the two weights applied on opposite sides and perpendicular to the bar this reads $d_1 F_1 = d_2 F_2$, the law of the lever. Note that the derivation used Equation (19.15), which was itself obtained by discarding $\boldsymbol{v} \times \boldsymbol{p}$; that term vanishes identically, so nothing was approximated. □

Covariance

Theorem 19.12 settled the translational law. The rotational law raises a question the translational one does not, because $\boldsymbol{L}$ and $\boldsymbol{\tau}$ are both defined relative to a chosen point, and a change of frame moves that point.

Remark 19.24 (A transformation law owed by the tensor chapter). The argument below uses the fact that the cross product of two vectors transforms under an orthogonal change of Cartesian axes a_{ij} as

$$(\boldsymbol{u} \times \boldsymbol{w})'_i = (\det a)\, a_{ij}\, (\boldsymbol{u} \times \boldsymbol{w})_j \,, \tag{19.17}$$

that is, as a vector for a rotation and with an extra sign for a reflection — the reason $\boldsymbol{L}$ and $\boldsymbol{\tau}$ are called *axial*, or pseudo-, vectors. Chapter 13 defines the cross product in the index form $(\boldsymbol{u} \times \boldsymbol{w})_i = \varepsilon_{ijk} u_j w_k$ of Equation (13.225) and names it a pseudo-vector, but does not state Equation (19.17); the statement belongs there rather than here. It follows in one line from the Leibniz expansion Equation (5.19) of a determinant, which gives $\varepsilon_{lmn} \det a = \varepsilon_{ijk} a_{il} a_{jm} a_{kn}$, combined with the transformation rule Equation (13.21) for Cartesian tensors. For $R \in \mathrm{SO}(3)$, where $\det R = +1$, Equation (19.17) reads $(R\boldsymbol{u}) \times (R\boldsymbol{w}) = R\,(\boldsymbol{u} \times \boldsymbol{w})$, and that is the form used below.

Theorem 19.25 (Galilean covariance of the rotational second law)**.** *Let a particle of constant mass be subject to a force transforming as $\boldsymbol{F}' = R\boldsymbol{F}$ under Equation (18.11), and let $\boldsymbol{L}'$ and $\boldsymbol{\tau}'$ be formed from the primed position and momentum, that is, about the origin of K'. Then Equation (19.14) holds in K if and only if $\boldsymbol{\tau}' = \mathrm{d}\boldsymbol{L}'/\mathrm{d}t'$ holds in K', even though neither $\boldsymbol{L}$ nor $\boldsymbol{\tau}$ is invariant.*

Rests on Equation (19.14), Definition 18.18, and Theorem 19.12.

Proof. *Derives Theorem 19.25.*

Write $\boldsymbol{c}(t) = \boldsymbol{V}t + \boldsymbol{b}$ for the displacement of the origin of K', so that Equations (18.11) and (19.6) read $\boldsymbol{x}' = R\boldsymbol{x} - \boldsymbol{c}$ and $\boldsymbol{v}' = R\boldsymbol{v} - \boldsymbol{V}$ with $\mathrm{d}\boldsymbol{c}/\mathrm{d}t = \boldsymbol{V}$. Then

$$\begin{aligned} \boldsymbol{\tau}' &= \boldsymbol{x}' \times \boldsymbol{F}' \\ &= (R\boldsymbol{x} - \boldsymbol{c}) \times R\boldsymbol{F} \\ &= R\boldsymbol{\tau} - \boldsymbol{c} \times R\boldsymbol{F} \qquad \text{by Remark 19.24}\,, \end{aligned}$$

using Equation (19.15) in the last line. For the angular momentum, $\boldsymbol{p}' = m\boldsymbol{v}'$ by Equation (19.1) and Definition 18.18, so

$$\begin{aligned} \boldsymbol{L}' &= (R\boldsymbol{x} - \boldsymbol{c}) \times m\,(R\boldsymbol{v} - \boldsymbol{V}) \\ &= R\boldsymbol{L} - m\,(R\boldsymbol{x}) \times \boldsymbol{V} - m\,\boldsymbol{c} \times R\boldsymbol{v} + m\,\boldsymbol{c} \times \boldsymbol{V}\,. \end{aligned}$$

Differentiating with respect to t, which by Postulate 18.14 is the same as differentiating with respect to t', and using $\mathrm{d}\boldsymbol{c}/\mathrm{d}t = \boldsymbol{V}$ together with $\boldsymbol{V} \times \boldsymbol{V} = \boldsymbol{0}$,

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{L}'}{\mathrm{d}t} &= R\boldsymbol{\tau} - m\,(R\boldsymbol{v}) \times \boldsymbol{V} - m\,\boldsymbol{V} \times R\boldsymbol{v} - m\,\boldsymbol{c} \times R\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} \\ &= R\boldsymbol{\tau} - \boldsymbol{c} \times R\boldsymbol{F}\,, \end{aligned}$$

because the two middle terms are equal and opposite by the antisymmetry of the cross product, and $m\,\mathrm{d}\boldsymbol{v}/\mathrm{d}t = \boldsymbol{F}$ by Equation (19.4). The two computations agree, so $\mathrm{d}\boldsymbol{L}'/\mathrm{d}t' = \boldsymbol{\tau}'$; and running the argument with R^{-1}, $-R^{-1}\boldsymbol{V}$ and $-R^{-1}\boldsymbol{b}$ gives the converse. □

Remark 19.26 (The extra terms are not an embarrassment). The displayed expression for $\boldsymbol{L}'$ shows that angular momentum is frame-dependent in a way linear momentum is not: it depends on where the origin is put, on how fast that origin moves, and — through the term $m\,\boldsymbol{c}\times R\boldsymbol{v}$ — on the time. A body at rest at $\boldsymbol{x} = \boldsymbol{0}$ has $\boldsymbol{L} = \boldsymbol{0}$ in K and a nonzero, time-varying $\boldsymbol{L}'$ in any frame whose origin is elsewhere and moving. What survives the change of frame is the *balance* Equation (19.14) between the two sides, which is all a law of motion is ever asked to do.

Remark 19.27 (Parity). The proof of Theorem 19.25 was written for $R \in \mathrm{SO}(3)$, but it does not need $\det R = +1$. For a reflection, Equation (19.17) carries an extra minus sign, so the same two computations return $\boldsymbol{\tau}' = -R\boldsymbol{\tau} - \boldsymbol{c}\times R\boldsymbol{F}$ and $\mathrm{d}\boldsymbol{L}'/\mathrm{d}t = -R\boldsymbol{\tau} - \boldsymbol{c}\times R\boldsymbol{F}$: the sign appears on both sides and cancels between them, and the law is form-invariant under the full orthogonal group $\mathrm{O}(3)$. Newtonian mechanics is therefore exactly parity-symmetric, and the mirror image of a mechanical process is again a mechanical process. That this is a property of mechanics rather than of nature was established only in 1957, when the beta decay of oriented cobalt-60 nuclei was found to distinguish left from right [1133]; the experiment is reported in Chapter 110.

Remark 19.28 (How well is the invariance of space tested?). Theorems 19.12 and 19.25 assert that no experiment can detect the orientation, position, epoch or uniform velocity of a laboratory. Each of the ten parameters of Equation (18.11) therefore carries an experimental bound, and the bounds are severe: the compilation of limits on violations of rotation and boost invariance [2176] collects them across atomic, nuclear and particle physics. They are quoted there for the Lorentz group rather than the Galilei group, which is the correct modern setting; the Galilean statements of this chapter inherit whatever bound applies at speeds where Equation (18.8) is negligible.

19.4 Energy

19.4.1 Kinetic energy

Definition 19.29 (Kinetic energy)**.** The *kinetic energy* of a particle moving with velocity $\boldsymbol{v}$ with respect to a frame of reference K is

$$T = \frac{1}{2}mv^2\,, \tag{19.18}$$

with units of J.

Rests on Definitions 18.10 and 19.3.

19.4.2 Potential energy and conservative forces

Definition 19.30 (Conservative force and potential energy)**.** Consider a force $\boldsymbol{F}_V$ upon a particle. This force is *conservative* if and only if there exists a real function V, which we shall call the *potential energy*, such that

$$\boldsymbol{F}_V = -\boldsymbol{\nabla}V\,. \tag{19.19}$$

Rests on Postulate 19.8 and Equation (19.2).

Let us see that if $\boldsymbol{F}$ is conservative one has

$$\boldsymbol{\nabla}\times\boldsymbol{F} = -\boldsymbol{\nabla}\times\boldsymbol{\nabla}V\,,$$

and therefore

$$\boldsymbol{\nabla} \times \boldsymbol{F} = \boldsymbol{0}\,. \tag{19.20}$$

We may consider that a free particle is subject to a constant potential energy V_0, since the force $\boldsymbol{F}_V$ upon it would still be zero.

On occasion, instead of saying that a particle is subject to a conservative force $\boldsymbol{F}_V$, one says that it is subject to a *potential* V.

19.4.3 Energy

Definition 19.31 (Energy)**.** Consider a particle subject to a potential V. The total *energy* of the particle is

$$E = T + V\,. \tag{19.21}$$

Rests on Definitions 19.29 and 19.30.

19.4.4 The law of conservation of energy

The source manuscript reserves this heading with the instruction to demonstrate that $\mathrm{d}E/\mathrm{d}t = 0$ for every inertial frame. Carried out, the instruction turns out to be right for a closed system and wrong for a single particle in an externally imposed field, and the difference is worth stating precisely.

Theorem 19.32 (Conservation of energy)**.** *Let the total force upon a particle derive from a potential V that carries no explicit dependence on time. Then the energy $E = T + V$ satisfies*

$$\frac{\mathrm{d}E}{\mathrm{d}t} = 0 \tag{19.22}$$

with respect to the inertial frame in which V is time-independent, but not in general with respect to any other. If instead the system is a closed one of N particles of constant masses, whose internal forces derive from a potential U depending only on the mutual distances $|\boldsymbol{x}_s - \boldsymbol{x}_r|$, then Equation (19.22) holds for $E = \sum_s \frac{1}{2} m_s |\boldsymbol{v}_s|^2 + U$ with respect to every *inertial frame.*

Rests on Definition 19.31 and Equations (19.4) and (19.19).

Proof. *Derives Theorem 19.32.*

One particle, one frame. Differentiating Equation (19.18) along the motion and using Equation (19.4), $\mathrm{d}T/\mathrm{d}t = \boldsymbol{F} \cdot \boldsymbol{v}$; and since V carries no explicit time dependence, the chain rule gives $\mathrm{d}V/\mathrm{d}t = \boldsymbol{\nabla} V \cdot \boldsymbol{v} = -\boldsymbol{F} \cdot \boldsymbol{v}$ by Equation (19.19). The two cancel, which is Equation (19.22).

One particle, another frame. Boost to a frame K' moving with constant velocity $\boldsymbol{V}$, so that $\boldsymbol{x} = \boldsymbol{x}' + \boldsymbol{V}t$ and $\boldsymbol{v}' = \boldsymbol{v} - \boldsymbol{V}$ by Equation (19.6). The potential seen there is $V(\boldsymbol{x}' + \boldsymbol{V}t)$, which *does* depend explicitly on the time, and

$$\frac{\mathrm{d}E'}{\mathrm{d}t} = m\,\boldsymbol{v}' \cdot \frac{\mathrm{d}\boldsymbol{v}'}{\mathrm{d}t} + \boldsymbol{\nabla} V \cdot (\boldsymbol{v}' + \boldsymbol{V}) = \boldsymbol{F} \cdot \boldsymbol{v}' - \boldsymbol{F} \cdot (\boldsymbol{v}' + \boldsymbol{V}) = -\boldsymbol{F} \cdot \boldsymbol{V}\,, \tag{19.23}$$

which vanishes only when the force is perpendicular to the boost. The reading is physical rather than paradoxical: in K' the agent maintaining the external field is itself in motion, and it does work at the rate $-\boldsymbol{F} \cdot \boldsymbol{V}$. An external static potential is always an idealization of a second body held fixed, and holding it fixed is what the second frame sees being paid for.

The closed system. Here there is no external agent. Under the general transformation Equation (18.11) the mutual distances are unchanged, since $\boldsymbol{x}'_s - \boldsymbol{x}'_r = R(\boldsymbol{x}_s - \boldsymbol{x}_r)$ and R preserves lengths, so $U' = U$. The kinetic energy transforms by Equation (19.6) as

$$T' = \sum_s \frac{1}{2} m_s |R\boldsymbol{v}_s - \boldsymbol{V}|^2 = T - \boldsymbol{V} \cdot R\boldsymbol{P} + \frac{1}{2} M V^2\,, \tag{19.24}$$

with $\boldsymbol{P} = \sum_s \boldsymbol{p}_s$ the total momentum and M the total mass Equation (19.32). Hence $E' = E - \boldsymbol{V} \cdot R\boldsymbol{P} + \frac{1}{2}MV^2$. Each of the three terms on the right is constant in time — E by the first part of this proof applied to the system, $\boldsymbol{P}$ by Theorem 19.67, and M and $\boldsymbol{V}$ by hypothesis — so E' is constant too. Note what this does *not* say: $E' \neq E$, and the two frames disagree about how much energy the system has. They agree only that the number does not change. □

Phenomenon 19.33 (Mechanical energy is conserved)**.** A pendulum bob released from rest at a given height rises on the far side to that same height, and continues to do so when a nail placed beneath the point of suspension catches the cord in mid-swing and forces the bob onto a shorter arc of a different radius: what the descent confers is fixed by the vertical drop alone and not by the path taken. Galileo reports the interrupted-pendulum experiment for exactly this purpose [4]. The same law governs a ball rolling in a bowl, a mass bouncing on a spring, and a projectile exchanging height for speed; and it fails visibly, as a decaying amplitude, precisely when friction is present — that is, precisely when the total force does not derive from a potential.

Rests on Theorem 19.32 and Equation (19.19)*.*

Derivation. Derives Phenomenon 19.33.

Let the total force on a particle of constant mass derive from a potential that does not depend explicitly on time, $\boldsymbol{F} = -\boldsymbol{\nabla} V$ as in Equation (19.19). Differentiating the kinetic energy Equation (19.18) along the motion and using Equation (19.4),

$$\frac{\mathrm{d}T}{\mathrm{d}t} = m\,\boldsymbol{v} \cdot \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \boldsymbol{F} \cdot \boldsymbol{v}\,,$$

while the potential energy evaluated along the same motion changes, by the chain rule, at the rate

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \boldsymbol{\nabla} V \cdot \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = -\boldsymbol{F} \cdot \boldsymbol{v}\,.$$

Adding the two, the energy Equation (19.21) obeys $\frac{\mathrm{d}E}{\mathrm{d}t} = 0$, which is Equation (19.22). For the pendulum, $V = mgh$ and $T = \frac{1}{2}mv^2$, so a bob released from rest at height h_0 carries $E = mgh_0$ throughout and can be momentarily at rest only where $h = h_0$ — whatever the shape of the arc it has travelled, which is the interrupted-pendulum observation. Two limitations are worth recording. The frame-dependence is real and is settled in Theorem 19.32: for the bob taken alone, in a frame moving horizontally, the energy is *not* constant by Equation (19.23), and constancy is recovered only when the Earth is included and the system is closed. And the argument fails as soon as V depends explicitly on time or the force has a non-conservative part, which is why Equation (19.20) is the criterion to test before invoking the result — the decaying amplitude of a real pendulum is that criterion failing. □

19.4.5 Work

Definition 19.34 (Work)**.** The *work* performed by a particle along a curve C, whose trajectory is due to a total force $\boldsymbol{F}$, is

$$W = \int_C \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{x}\,, \tag{19.25}$$

with units of J.

Rests on Postulate 19.8 and Definition 18.3.

Let us see that if $\boldsymbol{F}$ is conservative and C is a closed curve, one has

$$\begin{aligned} W &= \oint_C \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{x} \\ &= -\oint_C \boldsymbol{\nabla} V \cdot \mathrm{d}\boldsymbol{x} && \text{by Equation (19.19)} \\ &= -\int_S \boldsymbol{\nabla} \times \boldsymbol{\nabla} V \cdot \mathrm{d}\boldsymbol{s} \\ &= 0\,. \end{aligned}$$

Therefore, the work performed by a particle under a conservative total force along a closed curve is null.

The differential of work performed by a particle in a displacement $\mathrm{d}\boldsymbol{x}$, due to a force $\boldsymbol{F}$, is given by

$$\mathrm{d}W = \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{x}\,.$$

Now, if the total force upon the particle comes from a potential V, one has

$$\begin{aligned} \mathrm{d}W &= \boldsymbol{F}_V \cdot \mathrm{d}\boldsymbol{x} \\ &= -\boldsymbol{\nabla} V \cdot \mathrm{d}\boldsymbol{x} \\ &= -\frac{\partial V}{\partial x^i}\,\mathrm{d}x^i \\ &= -\mathrm{d}V\,, \end{aligned}$$

and, integrating, we see that

$$\Delta W = -\Delta V\,. \tag{19.26}$$

Power

Definition 19.35 (Power)**.** The *power* is the rate at which work is performed,

$$P = \frac{\mathrm{d}W}{\mathrm{d}t}\,, \tag{19.27}$$

with units of $\mathrm{W} = \mathrm{J/s}$.

Rests on Definition 19.34.

Work for the magnitude of a torque

The formulary asserts, without derivation, that the differential of work associated with the magnitude of a torque τ through an angle ϕ is

$$\mathrm{d}W = \tau\,\mathrm{d}\phi\,. \tag{19.28}$$

The relation is true under a hypothesis the formulary leaves unsaid, and stating it is the whole of the derivation.

Proposition 19.36 (Work of a torque)**.** *Let a body turn by an infinitesimal rotation $\mathrm{d}\boldsymbol{\phi} = \hat{\boldsymbol{n}}\,\mathrm{d}\phi$ about an axis $\hat{\boldsymbol{n}}$ through the origin, and let a force $\boldsymbol{F}$ be applied at the point $\boldsymbol{x}$. Then the work done is*

$$\mathrm{d}W = \boldsymbol{\tau} \cdot \mathrm{d}\boldsymbol{\phi} = (\hat{\boldsymbol{n}} \cdot \boldsymbol{\tau})\,\mathrm{d}\phi\,, \tag{19.29}$$

in J*, with $\boldsymbol{\tau}$ the torque Equation (19.15) about the origin. The scalar form Equation (19.28) is the special case in which the torque is parallel to the axis of rotation.*

Rests on Definition 19.34 and Equation (19.15).

Proof. *Derives Proposition 19.36.*

Under a rotation by $\mathrm{d}\boldsymbol{\phi}$ about an axis through the origin the point of application moves by $\mathrm{d}\boldsymbol{x} = \mathrm{d}\boldsymbol{\phi} \times \boldsymbol{x}$. This is the generator of $\mathfrak{so}(3)$ acting on a vector: with the basis Equation (14.32) of Proposition 14.27, whose entries are $(L_k)_{ij} = -\epsilon_{kij}$, one has $\left(\mathrm{d}\phi^k L_k \boldsymbol{x}\right)_i = \epsilon_{ikj}\mathrm{d}\phi^k x_j = (\mathrm{d}\boldsymbol{\phi} \times \boldsymbol{x})_i$. Then, by Definition 19.34,

$$\begin{aligned} \mathrm{d}W &= \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{x} \\ &= \boldsymbol{F} \cdot (\mathrm{d}\boldsymbol{\phi} \times \boldsymbol{x}) \\ &= \mathrm{d}\boldsymbol{\phi} \cdot (\boldsymbol{x} \times \boldsymbol{F}) \\ &= \boldsymbol{\tau} \cdot \mathrm{d}\boldsymbol{\phi}\,, \end{aligned}$$

the third line by the cyclic invariance of the scalar triple product, which in the index notation Equation (13.225) is the cyclic symmetry of ε_{ijk}, and the fourth by Equation (19.15). Writing $\mathrm{d}\boldsymbol{\phi} = \hat{\boldsymbol{n}}\,\mathrm{d}\phi$ gives Equation (19.29). The component of $\boldsymbol{\tau}$ perpendicular to $\hat{\boldsymbol{n}}$ does no work: it is taken up by whatever constrains the axis, which is why a door hinge does not absorb energy while the door swings. Consequently Equation (19.28) as the formulary states it holds when, and only when, $\boldsymbol{\tau}$ is parallel to $\hat{\boldsymbol{n}}$ — the case of a body free to turn about a single fixed axis with the force applied in the plane perpendicular to it. □

Corollary 19.37 (Rotational power)**.** *Dividing Equation (19.29) by* $\mathrm{d}t$ *and using Equation (19.27),*

$$P = \boldsymbol{\tau} \cdot \boldsymbol{\omega}\,, \tag{19.30}$$

with $\boldsymbol{\omega} = \mathrm{d}\boldsymbol{\phi}/\mathrm{d}t$ *the angular velocity in* rad/s *and* P *in* W*. This is the relation by which the shaft power of an engine is computed from a dynamometer reading of the torque and a tachometer reading of the rotation rate.*

Rests on Proposition 19.36 and Equation (19.27).

19.4.6 The work–energy theorem

Theorem 19.38 (Work–energy theorem)**.** *The work performed by a particle equals the variation of its kinetic energy,*

$$\boxed{W = \Delta T}\,. \tag{19.31}$$

Rests on Definitions 19.29 and 19.34 and Equation (19.4).

Derivation. *Derives Theorem 19.38.*

Consider a particle under a total force $\boldsymbol{F}$ that performs a work W. From the definition

of work and Equation (19.4),

$$\begin{aligned} W &= \int_C \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{x} \\ &= m \int_{t_a}^{t_b} \left(\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} \cdot \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} \right) \mathrm{d}t \\ &= m \int_{t_a}^{t_b} \left(\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} \cdot \boldsymbol{v} \right) \mathrm{d}t \\ &= \frac{1}{2} m \int_{t_a}^{t_b} \frac{\mathrm{d}(\boldsymbol{v} \cdot \boldsymbol{v})}{\mathrm{d}t} \, \mathrm{d}t \\ &= \frac{1}{2} m \int_{t_a}^{t_b} \mathrm{d} \left| \boldsymbol{v} \right|^2 \\ &= \frac{1}{2} m \left(\left| \boldsymbol{v}(t_b) \right|^2 - \left| \boldsymbol{v}(t_a) \right|^2 \right) . \end{aligned}$$

Therefore, from Equation (19.18), $W = \Delta T$. □

This result is known as the *work–energy theorem*, or the theorem of the *vires vivae* (living forces).

19.5 Systems of particles

In the preceding sections we introduced all the fundamental principles that govern the dynamics of a particle and of a body. We began with particles because Newton's laws take their simplest form there. We did not do the same with bodies of definite volume, since in essence it is not necessary: to speak of particles is not more general, but it is more canonical.

We shall now introduce the concepts, and extend the definitions in the appropriate manner, for systems of many particles and bodies.

19.5.1 Inertia and mass of a system

Body as a system of particles

Definition 19.39 (Body)**.** A *body* is a set of N particles occupying a region V of space. This region of space is called the *volume* of the body.

Rests on Definitions 18.4 and 19.3.

Total mass of the system

Consider a closed system of N particles of masses m_s, with $s = 1, \ldots, N$.

Definition 19.40 (Total mass)**.** The total mass of the system is simply the sum of all the masses,

$$M = \sum_{s}^{N} m_s \, . \tag{19.32}$$

Rests on Definitions 19.3 and 19.39.

Because we have considered the system to be closed, M is constant.

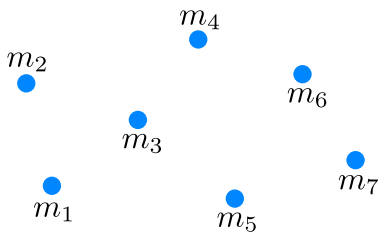


Figure 19.3: A closed system of N particles: the total mass is the sum of the individual masses.

Moment of inertia of the system

Definition 19.3 makes the mass the measure of how much a particle resists being translated. A body resists being *turned* as well, and by an amount that depends on how its mass is distributed about the axis of the turn: this is the everyday fact that a long pole is harder to spin about its middle than a compact one of the same weight.

Definition 19.41 (Moment of inertia about an axis)**.** Let a system of N particles of masses m_s have positions $\boldsymbol{x}_s$ measured from a point O, and let $\hat{\boldsymbol{n}}$ be a unit vector through O. The *moment of inertia* of the system about the axis $(O, \hat{\boldsymbol{n}})$ is

$$I = \sum_{s}^{N} m_s d_s^2 \,, \qquad d_s = |\hat{\boldsymbol{n}} \times \boldsymbol{x}_s| \,, \tag{19.33}$$

with d_s the perpendicular distance from the particle to the axis, of SI unit $\mathrm{kg\,m^2}$. For a body of mass density ρ the sum is replaced by $I = \int_V \rho\, d^2 \,\mathrm{d}V$.

Rests on Definitions 19.3 and 19.39.

The dependence on the axis is quadratic and can be separated out.

Definition 19.42 (Inertia tensor)**.** The *inertia tensor* of the system about the point O is the array

$$I_{ij} = \sum_{s}^{N} m_s \left(|\boldsymbol{x}_s|^2 \, \delta_{ij} - x_{s,i}\, x_{s,j} \right) \,, \tag{19.34}$$

of SI unit $\mathrm{kg\,m^2}$.

Rests on Definitions 13.5 and 19.41.

Proposition 19.43 (The moment of inertia is a quadratic form in the axis)**.** *For every unit vector $\hat{\boldsymbol{n}}$,*

$$I = \hat{n}_i \, I_{ij} \, \hat{n}_j \,, \tag{19.35}$$

and I_{ij} is symmetric, is a Cartesian tensor of rank two in the sense of Definition 13.5, and is positive semidefinite, vanishing on $\hat{\boldsymbol{n}}$ only when every particle of the system lies on the axis.

Rests on Definitions 13.5 and 19.42.

Proof. *Derives Proposition 19.43.*

Since $\hat{\boldsymbol{n}}$ is a unit vector, $d_s^2 = |\hat{\boldsymbol{n}} \times \boldsymbol{x}_s|^2 = |\boldsymbol{x}_s|^2 - (\hat{\boldsymbol{n}} \cdot \boldsymbol{x}_s)^2$, which is Pythagoras' theorem applied to the decomposition of $\boldsymbol{x}_s$ into its components along and across the axis. Substituting into Equation (19.33) and writing $|\boldsymbol{x}_s|^2 = \hat{n}_i \delta_{ij} \hat{n}_j \, |\boldsymbol{x}_s|^2$ and $(\hat{\boldsymbol{n}} \cdot \boldsymbol{x}_s)^2 = \hat{n}_i x_{s,i} x_{s,j} \hat{n}_j$ gives Equation (19.35) with I_{ij} as in Equation (19.34). Symmetry is manifest. Each summand of Equation (19.34) is built from δ_{ij}, which is invariant, and from the outer product $x_{s,i} x_{s,j}$ of a vector with itself, which obeys Equation (13.22) because each factor obeys the vector rule; the finite sum of tensors is a tensor, so I_{ij} satisfies Definition 13.5. Finally Equation (19.33) exhibits I as a sum of terms $m_s d_s^2 \geq 0$, so $I \geq 0$, with equality precisely when every $d_s = 0$. □

Theorem 19.44 (Parallel axes; Steiner)**.** *Let I_{CM} be the moment of inertia of the system about an axis of direction $\hat{\boldsymbol{n}}$ through its centre of mass (Definition 19.47, below), and let I be the moment of inertia about the parallel axis of the same direction through a point at perpendicular distance D from it. Then*

$$\boxed{I = I_{CM} + MD^2}, \tag{19.36}$$

with M the total mass Equation (19.32).

Rests on Definition 19.41 and Equations (19.32) and (19.38).

Proof. *Derives Theorem 19.44.*

Take the origin at the centre of mass and let the second axis pass through the point $\boldsymbol{D}$, chosen perpendicular to $\hat{\boldsymbol{n}}$ without loss of generality, so that $|\boldsymbol{D}| = D$. Write $\boldsymbol{x}_s^{\perp}$ for the component of $\boldsymbol{x}_s$ perpendicular to $\hat{\boldsymbol{n}}$, so that $d_s = \left|\boldsymbol{x}_s^{\perp}\right|$ for the first axis and $\left|\boldsymbol{x}_s^{\perp} - \boldsymbol{D}\right|$ for the second. Then

$$\begin{aligned} I &= \sum_s^N m_s \left|\boldsymbol{x}_s^{\perp} - \boldsymbol{D}\right|^2 \\ &= \sum_s^N m_s \left|\boldsymbol{x}_s^{\perp}\right|^2 - 2\boldsymbol{D}\cdot\sum_s^N m_s \boldsymbol{x}_s^{\perp} + D^2 \sum_s^N m_s \\ &= I_{\mathrm{CM}} + MD^2 \,, \end{aligned}$$

because the middle sum is M times the perpendicular component of the position of the centre of mass Equation (19.38), which vanishes when the origin is the centre of mass itself. The theorem therefore singles the centre of mass out: among all parallel axes of a given direction, the one through the centre of mass has the smallest moment of inertia, and Equation (19.36) says by how much any other exceeds it. □

Example 19.45 (The uniform solid sphere)**.** For a homogeneous sphere of mass M and radius R the symmetry forces $I_{ij} = I\delta_{ij}$, so every axis through the centre gives the same moment of inertia and it may be extracted from the trace of Equation (19.34), which is $I_{ii} = 2\int_V \rho\,|\boldsymbol{x}|^2\,\mathrm{d}V = 3I$. With $\rho = 3M/\left(4\pi R^3\right)$ constant,

$$\int_V \rho\,|\boldsymbol{x}|^2\,\mathrm{d}V = \rho\int_0^R r^2\,4\pi r^2\,\mathrm{d}r = \frac{4\pi\rho R^5}{5} = \frac{3}{5}MR^2\,,$$

whence

$$I = \frac{2}{5}MR^2\,. \tag{19.37}$$

A measured moment of inertia smaller than $\frac{2}{5}MR^2$ is therefore direct evidence that a body's mass is concentrated towards its centre, which is how the internal structure of a planet is read off its precession.

Rests on Definition 19.42 and Equation (19.33).

Remark 19.46 (Principal axes, and what is owed for them). Because I_{ij} is real and symmetric, there is an orthonormal basis of $\mathbb{R}^3$ in which it is diagonal, $I_{ij} = \mathrm{diag}\,(I_1, I_2, I_3)$ with $I_k \geq 0$; the basis directions are the *principal axes* and the I_k the *principal moments*. Theorem 5.75 proves the spectral theorem for a self-adjoint operator on a complex inner-product space, which gives the reality of the I_k at once; the statement actually used here is its real form, that a real symmetric matrix is diagonalized by a *real orthogonal* change of basis, and Part II does not state it separately. It is a standard corollary — the eigenvalues being real, the eigenvectors may be taken real — but it is mathematics and belongs in Chapter 5, not here. Everything that follows from the principal axes, from Euler's equations to the free precession of a symmetric top, is the subject of Chapter 29 [32, 99].

Kinematics of the centre of mass

Consider the preceding system, in which the positions of the particles are $\boldsymbol{x}_s$. How can we assign to the system a quantity that allows us to treat it as a single particle? We can define a point of space that takes an average of the positions, according to the amount of mass assigned to each particle.

Definition 19.47 (Position of the centre of mass)**.** The *position of the centre of mass* of a system of particles is

$$\boldsymbol{x}_{\mathrm{CM}} = \frac{\sum_s^N m_s \boldsymbol{x}_s}{M} \,. \tag{19.38}$$

Rests on Definitions 19.39 and 19.40.

The centre of mass may itself be regarded as a point of space to which the mass M is assigned. By definition, the *velocity of the centre of mass*, with respect to a frame of reference in which its position is $\boldsymbol{x}_{\mathrm{CM}}$, is given by

$$\boldsymbol{v}_{\mathrm{CM}} = \frac{\mathrm{d}\boldsymbol{x}_{\mathrm{CM}}}{\mathrm{d}t} \,, \tag{19.39}$$

and, in the same fashion, the *acceleration of the centre of mass* is given by

$$\boldsymbol{a}_{\mathrm{CM}} = \frac{\mathrm{d}\boldsymbol{v}_{\mathrm{CM}}}{\mathrm{d}t} \,. \tag{19.40}$$

19.5.2 Internal dynamics of the system

Newton's third law

Let us note the following fact of nature. When we push a wall with a given force, directed towards it, it is as if the wall pushed us back with the same force, but in the opposite direction.

Considering again the aforementioned system of particles: what about the interaction among them? Well, since the particles can interact, one of them feels a total force due to the others, and it exerts a force upon the others. What is the relation between these forces? To give an answer, let us take two particles of the system, which we shall call particle 1 and particle 2.

We shall denote by $\boldsymbol{F}_{12}$ the force felt by particle 1 due to particle 2, and by $\boldsymbol{F}_{21}$ the force felt by particle 2 due to particle 1.

On the basis of the philosophical conclusions about *action and reaction*, we may intuit that the magnitudes of these forces must be equal; but we must in addition specify an orientation for these vectors. In any case, owing to the interaction between these two particles they approach or recede from each other; and we may also intuit that this trajectory of approach or recession is in principle a straight line, so that these forces would be parallel.

We then enunciate as a fundamental principle:

Postulate 19.48 (Newton's third law)**.** The force that one particle exerts upon another is equal to the force that the latter exerts upon the former, but with opposite sign,

$$\boldsymbol{F}_{12} = -\boldsymbol{F}_{21} \,. \tag{19.41}$$

Rests on Postulate 19.8 and Definition 19.2.

This is known by the name of *Newton's third law*, or the *principle of action and reaction.*

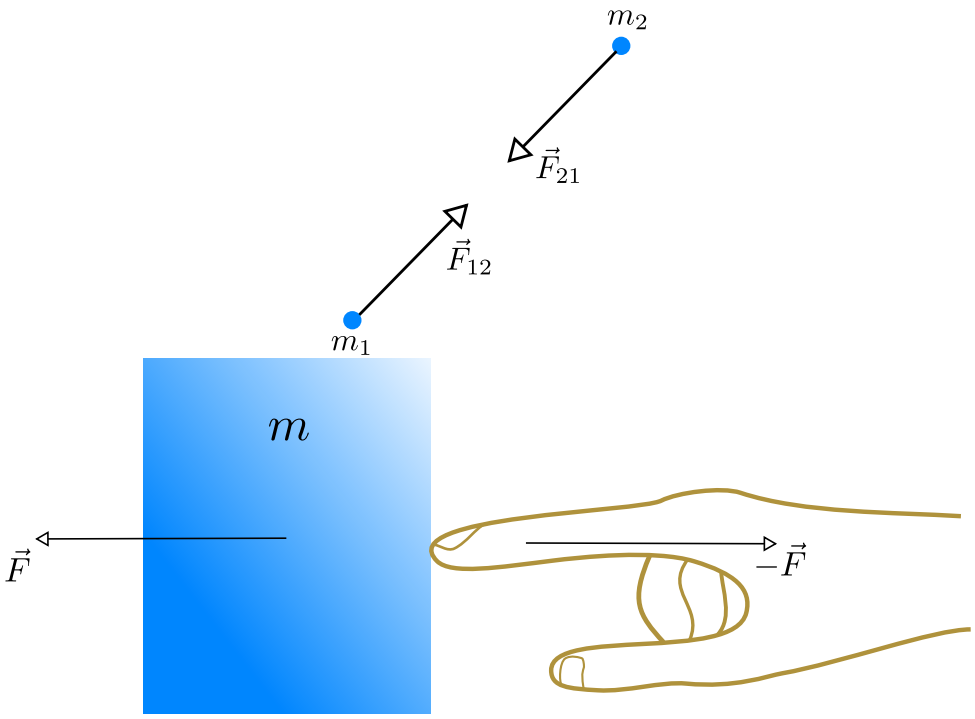


Figure 19.4: Action and reaction: the force one particle exerts upon another is equal and opposite to the force it receives from it.

Net internal force

There are two types of forces to which our system of particles is subject: the internal ones, $\boldsymbol{F}_{\text{INT}}$, which are the interactions among the particles; and the external ones, $\boldsymbol{F}_{\text{EXT}}$, which are those alien to the system — they affect it as a whole, not each particle separately. An example of this can be a gas of molecules inside a balloon: the particles interact with one another repulsively, forming the gas, while an external force could be the action of gravity upon the balloon. We shall assume:

Postulate 19.49 (Net internal force)**.** The internal interactions of the system do not contribute to its motion as a whole,

$$\boldsymbol{F}_{\text{INT}} = \boldsymbol{0}\,, \tag{19.42}$$

Rests on Postulate 19.48.

since for the most part they cancel by Newton's third law. In our example applied to nature, this amounts to the following: the molecular interactions are of no great importance for the motion of the balloon as such.

Phenomenon 19.50 (Recoil and the conservation of momentum)**.** When two bodies interact and nothing external acts on the pair, the vector sum of their linear momenta is the same after the interaction as before. A gun recoils as the shot leaves it; a swimmer moves forward by driving water backward; a rocket accelerates in vacuum with nothing to push against. In a collision on an air track the sum $m_1\boldsymbol{v}_1 + m_2\boldsymbol{v}_2$ is found unchanged whether the bodies rebound or stick together, and whether or not the kinetic energy is conserved — which it usually is not. Newton records the result as the third corollary to the laws of motion [14].

Rests on Postulate 19.48 and Equation (19.2)*.*

Derivation. *Derives Phenomenon 19.50.*

Take two particles subject only to their mutual interaction. Applying Equation (19.2) to each,

$$\frac{\mathrm{d}\boldsymbol{p}_1}{\mathrm{d}t} = \boldsymbol{F}_{12}\,, \qquad \frac{\mathrm{d}\boldsymbol{p}_2}{\mathrm{d}t} = \boldsymbol{F}_{21}\,,$$

and adding the two, Newton's third law Equation (19.41) makes the right-hand side vanish identically:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\boldsymbol{p}_1 + \boldsymbol{p}_2\right) = \boldsymbol{F}_{12} + \boldsymbol{F}_{21} = \boldsymbol{0}\,. \tag{19.43}$$

The total momentum is therefore constant throughout the interaction, however violent and however brief, and — this is the strength of the result — no property of the interaction law enters, so the same conclusion covers collisions whose forces are entirely unknown. For N particles the internal contributions cancel in the same way, pair by pair, which is the content of Postulate 19.49. Energy is not similarly protected: Equation (19.43) holds for an inelastic collision, in which kinetic energy is demonstrably lost to heat and deformation. □

19.5.3 Translational Newton laws for systems of particles

Linear momentum of the centre of mass

Consider a (closed) system of particles, in which the s-th particle has position $\boldsymbol{x}_s$ and linear momentum $\boldsymbol{p}_s$.

Definition 19.51 (Linear momentum of the centre of mass)**.** The *linear momentum of the centre of mass* is

$$\boldsymbol{p}_{\mathrm{CM}} = M\boldsymbol{v}_{\mathrm{CM}}\,. \tag{19.44}$$

Rests on Definitions 19.5 and 19.47.

Completeness

Let us compute:

$$\begin{aligned}
\boldsymbol{v}_{\mathrm{CM}} &= \frac{\mathrm{d}\boldsymbol{x}_{\mathrm{CM}}}{\mathrm{d}t} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\sum_s^N m_s\boldsymbol{x}_s}{M}\right) && \text{by Equation (19.38)} \\
&= \frac{1}{M}\sum_s^N \frac{\mathrm{d}}{\mathrm{d}t}(m_s\boldsymbol{x}_s) \\
&= \frac{1}{M}\sum_s^N \left(\frac{\mathrm{d}m_s}{\mathrm{d}t}\boldsymbol{x}_s + m_s\frac{\mathrm{d}\boldsymbol{x}_s}{\mathrm{d}t}\right)\,,
\end{aligned}$$

so that, by Equation (19.44),

$$\begin{aligned}
\boldsymbol{p}_{\mathrm{CM}} &= \sum_s^N \frac{\mathrm{d}m_s}{\mathrm{d}t}\boldsymbol{x}_s + \sum_s^N m_s\frac{\mathrm{d}\boldsymbol{x}_s}{\mathrm{d}t} \\
&= \sum_s^N \frac{\mathrm{d}m_s}{\mathrm{d}t}\boldsymbol{x}_s + \sum_s^N \boldsymbol{p}_s\,.
\end{aligned}$$

Restricting to the case in which the mass of each particle is constant, we obtain the following *completeness of the total linear momentum* of the system:

$$\boldsymbol{p}_{\mathrm{CM}} = \sum_{s=1}^N \boldsymbol{p}_s\,, \tag{19.45}$$

that is, the sum of the linear momenta of all the particles of the system is equal to the linear momentum of the centre of mass.

Translational Newton's second law for the centre of mass

Let us consider the centre of mass as a particle. Denote the total force upon our system of particles by $\boldsymbol{F}$. Recalling the discussion of the net internal force, one has

$$\boldsymbol{F} = \boldsymbol{F}_{\text{EXT}} + \boldsymbol{F}_{\text{INT}}\,,$$

and, as we have assumed, $\boldsymbol{F}_{\text{INT}} = \boldsymbol{0}$. Thus, applying Newton's second law to our centre of mass, we arrive at the expression

$$\begin{aligned}\boldsymbol{F}_{\text{EXT}} &= \frac{\mathrm{d}\boldsymbol{p}_{\text{CM}}}{\mathrm{d}t} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}(M\boldsymbol{v}_{\text{CM}}) \qquad\qquad \text{by Equation (19.44)}\,,\end{aligned}$$

and, since M is constant, one has

$$\boxed{\boldsymbol{F}_{\text{EXT}} = M\boldsymbol{a}_{\text{CM}}}\,. \tag{19.46}$$

The source flags at this point, without carrying it out, the demonstration that the interaction of the centre of mass with the remaining particles cancels, so that Newton's second law may legitimately be applied to it. That demonstration is the following, and its by-product is that Postulate 19.49 need not have been postulated at all.

Theorem 19.52 (The centre of mass moves as a particle)**.** *Let a closed system of N particles of constant masses have internal forces obeying Equation (19.41). Then the net internal force vanishes,*

$$\boldsymbol{F}_{INT} = \sum_{s}^{N}\sum_{r\neq s}\boldsymbol{F}_{sr} = \boldsymbol{0}\,, \tag{19.47}$$

and consequently Equation (19.46) holds: the centre of mass accelerates as a single particle of mass M acted on by the total external force alone.

Rests on Postulate 19.48 and Equations (19.44) and (19.45).

Proof. *Derives Theorem 19.52.*

Group the double sum in Equation (19.47) into unordered pairs. The pair $\{s, r\}$ contributes $\boldsymbol{F}_{sr} + \boldsymbol{F}_{rs}$, which is $\boldsymbol{0}$ by Equation (19.41); every pair occurs exactly once, so the whole sum vanishes. Now differentiate the completeness relation Equation (19.45) and apply Equation (19.2) to each particle:

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{p}_{\text{CM}}}{\mathrm{d}t} &= \sum_{s}^{N}\frac{\mathrm{d}\boldsymbol{p}_s}{\mathrm{d}t} \\ &= \sum_{s}^{N}\boldsymbol{F}_s \\ &= \boldsymbol{F}_{\text{EXT}} + \boldsymbol{F}_{\text{INT}} = \boldsymbol{F}_{\text{EXT}}\,,\end{aligned}$$

and since M is constant, $\mathrm{d}\boldsymbol{p}_{\text{CM}}/\mathrm{d}t = M\boldsymbol{a}_{\text{CM}}$ by Equations (19.40) and (19.44), which is Equation (19.46). □

Remark 19.53 (A postulate that was a theorem). Equation (19.47) is exactly the statement Postulate 19.49 assumes, so that assumption may be discharged: for a system whose particles interact in pairs obeying the third law, the vanishing of the net internal force is a consequence, not an independent principle. The hedge in the discussion following Postulate 19.49 — "for the most part they cancel" — is therefore too weak, and can be strengthened to an exact cancellation. What it does require is that the forces be *pair* forces: a genuine three-body interaction, one not expressible as a sum over pairs, is not covered by Equation (19.41) and would have to be examined on its own. No such force is known among the fundamental interactions.

Phenomenon 19.54 (The centre of mass ignores internal forces)**.** A body that flies apart in mid-air does not thereby alter the path of its centre of mass: the fragments of a bursting firework, and the two halves of a rod that breaks in flight, continue to straddle the same smooth arc their common centre was already describing, and the arc shows no kink at the instant of the burst. A wrench tossed spinning across a room follows, in its centre of mass, exactly the path of a wrench that does not spin, however complicated the tumbling superposed on it. The observation is what licenses the everyday practice of treating an extended body as a point.

Rests on Equation (19.46) *and Postulate 19.49.*

Derivation. *Derives Phenomenon 19.54.*

By Equation (19.46) the acceleration of the centre of mass answers to $\boldsymbol{F}_{\mathrm{EXT}}$ alone; the internal forces have already dropped out through Postulate 19.49, cancelling pair by pair by Equation (19.41). An explosion, a rotation, or any other rearrangement of the parts is internal by definition, so it alters neither $\boldsymbol{F}_{\mathrm{EXT}}$ nor the total mass M, and therefore leaves $\boldsymbol{a}_{\mathrm{CM}}$ untouched at the instant it occurs and afterwards. With the external force of a body near the ground, $\boldsymbol{F}_{\mathrm{EXT}} = M\boldsymbol{g}$, one has $\boldsymbol{a}_{\mathrm{CM}} = \boldsymbol{g}$ constant, so the centre of mass describes the parabola of Chapter 20 whatever the fragments do individually. Two caveats belong with the result: the cancellation of the internal forces requires the third law in the form Equation (19.41), and M must be constant, which excludes bodies that expel matter altogether rather than merely redistributing it. □

19.5.4 Rotational Newton laws for systems of particles

Angular momentum of the centre of mass

Definition 19.55 (Angular momentum of the centre of mass)**.** The *angular momentum of the centre of mass* is

$$\boldsymbol{L}_{\mathrm{CM}} = \boldsymbol{x}_{\mathrm{CM}} \times \boldsymbol{p}_{\mathrm{CM}} \,. \tag{19.48}$$

Rests on Definitions 19.18 and 19.51.

Completeness

The translational case gave the clean identity Equation (19.45): the momentum of the centre of mass is the sum of the momenta of the particles, with nothing left over. The rotational case does not, and the difference is the single most consequential fact about the mechanics of extended bodies.

Notation 19.56 (Positions and velocities relative to the centre of mass)*.* Throughout this subsection, a primed symbol denotes a quantity referred to the centre of mass:

$$\boldsymbol{x}_s = \boldsymbol{x}_{\mathrm{CM}} + \boldsymbol{x}'_s \,, \qquad \boldsymbol{v}_s = \boldsymbol{v}_{\mathrm{CM}} + \boldsymbol{v}'_s \,, \qquad \boldsymbol{p}'_s = m_s \boldsymbol{v}'_s \,. \tag{19.49}$$

The primes here have nothing to do with the change of inertial frame of Equation (18.11); the centre of mass need not move uniformly.

Lemma 19.57 (The centre of mass is the origin of its own frame)**.** *With the masses constant,*

$$\sum_{s}^{N} m_s \boldsymbol{x}'_s = \boldsymbol{0} \,, \qquad \sum_{s}^{N} \boldsymbol{p}'_s = \boldsymbol{0} \,. \tag{19.50}$$

Rests on Equations (19.38) *and* (19.49)*.*

Proof. *Derives Lemma 19.57.*

By Equations (19.38) and (19.49),

$$\sum_s^N m_s \boldsymbol{x}'_s = \sum_s^N m_s \boldsymbol{x}_s - \left(\sum_s^N m_s\right) \boldsymbol{x}_{\mathrm{CM}} = M\boldsymbol{x}_{\mathrm{CM}} - M\boldsymbol{x}_{\mathrm{CM}} = \boldsymbol{0}\,,$$

using Equation (19.32). Differentiating with respect to time, the masses being constant, gives the second identity. □

Theorem 19.58 (Decomposition of the angular momentum)**.** *The total angular momentum of the system about a fixed origin is*

$$\boxed{\sum_s^N \boldsymbol{x}_s \times \boldsymbol{p}_s = \boldsymbol{L}_{CM} + \boldsymbol{S}}\,, \qquad \boldsymbol{S} := \sum_s^N \boldsymbol{x}'_s \times \boldsymbol{p}'_s\,, \tag{19.51}$$

with $\boldsymbol{L}_{CM}$ the angular momentum Equation (19.48) of the centre of mass and $\boldsymbol{S}$ the internal, *or* spin, *angular momentum, both of SI unit* $\mathrm{kg\,m^2/s}$.

Rests on Definition 19.55, Lemma 19.57, and Equation (19.44).

Proof. *Derives Theorem 19.58.*

Substituting Equation (19.49) into the sum and expanding the four cross products,

$$\begin{aligned}\sum_s^N \boldsymbol{x}_s \times \boldsymbol{p}_s = \boldsymbol{x}_{\mathrm{CM}} \times \left(\sum_s^N m_s\right) \boldsymbol{v}_{\mathrm{CM}} + \boldsymbol{x}_{\mathrm{CM}} \times \sum_s^N \boldsymbol{p}'_s \\ + \left(\sum_s^N m_s \boldsymbol{x}'_s\right) \times \boldsymbol{v}_{\mathrm{CM}} + \sum_s^N \boldsymbol{x}'_s \times \boldsymbol{p}'_s\,.\end{aligned}$$

The two middle sums vanish by Equation (19.50), and the first term is $\boldsymbol{x}_{\mathrm{CM}} \times M\boldsymbol{v}_{\mathrm{CM}} = \boldsymbol{x}_{\mathrm{CM}} \times \boldsymbol{p}_{\mathrm{CM}} = \boldsymbol{L}_{\mathrm{CM}}$ by Equations (19.44) and (19.48). □

Corollary 19.59 (Decomposition of the kinetic energy)**.** *With the same notation,*

$$T = \frac{1}{2} M \left|\boldsymbol{v}_{CM}\right|^2 + \sum_s^N \frac{1}{2} m_s \left|\boldsymbol{v}'_s\right|^2\,. \tag{19.52}$$

Rests on Lemma 19.57 and Equation (19.18).

Proof. *Derives Corollary 19.59.*

Expanding $|\boldsymbol{v}_{\mathrm{CM}} + \boldsymbol{v}'_s|^2$ in $T = \sum_s \frac{1}{2} m_s |\boldsymbol{v}_s|^2$ gives three groups of terms; the cross term is $\boldsymbol{v}_{\mathrm{CM}} \cdot \sum_s \boldsymbol{p}'_s$, which vanishes by Equation (19.50). □

Remark 19.60 (Why there is no rotational completeness). Equation (19.51) is not the analogue of Equation (19.45): the angular momentum of the centre of mass is *not* the total angular momentum, and the defect $\boldsymbol{S}$ does not vanish for any system that is turning. A flywheel spinning in place has $\boldsymbol{L}_{\mathrm{CM}} = \boldsymbol{0}$ and $\boldsymbol{S} \neq \boldsymbol{0}$; the Earth in its orbit carries both, and they are independent — the one associated with the year, the other with the day. The reason the linear case is simpler is that translation has no internal counterpart: there is no way to redistribute the parts of a system so as to give it internal momentum, because Equation (19.50) forbids it, while $\boldsymbol{x}'_s \times \boldsymbol{p}'_s$ can perfectly well sum to something nonzero. The two pieces obey separate equations of motion, which is the subject of the next heading, and it is that separation which lets an extended body be treated as a point for its translation while keeping a full rotational dynamics of its own (Chapter 29).

Rotational Newton's second law for the centre of mass

For the translational law it sufficed that the internal forces cancel in pairs, which is Postulate 19.49. For the rotational law that is not enough: two equal and opposite forces applied at different points cancel as forces but form a couple, and a couple turns things. What is needed in addition is that the pair force act along the line joining the two particles.

Postulate 19.61 (Newton's third law, strong form)**.** The force between two particles is directed along the line joining them,

$$\boldsymbol{F}_{12} \parallel (\boldsymbol{x}_1 - \boldsymbol{x}_2) \ , \tag{19.53}$$

in addition to satisfying Equation (19.41).

Rests on Postulate 19.48 and Corollary 19.14.

Theorem 19.62 (Rotational second law for a system)**.** *Let the internal forces of a system satisfy Postulate 19.61 and let the masses be constant. Then, about any point at rest in an inertial frame,*

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\sum_{s}^{N} \boldsymbol{x}_s \times \boldsymbol{p}_s = \boldsymbol{\tau}_{EXT}}, \qquad \boldsymbol{\tau}_{EXT} := \sum_{s}^{N} \boldsymbol{x}_s \times \boldsymbol{F}_s^{EXT}, \tag{19.54}$$

where $\boldsymbol{F}_s^{EXT}$ is the external force on the s-th particle. Only the external torques appear.

Rests on Postulate 19.61 and Equations (19.2) and (19.14).

Proof. *Derives Theorem 19.62.*

Split the force on particle s into external and internal parts, $\boldsymbol{F}_s = \boldsymbol{F}_s^{\mathrm{EXT}} + \sum_{r\neq s} \boldsymbol{F}_{sr}$, and differentiate the sum, the term $\boldsymbol{v}_s \times \boldsymbol{p}_s$ vanishing for each s as in Proposition 19.20:

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{s}^{N} \boldsymbol{x}_s \times \boldsymbol{p}_s = \sum_{s}^{N} \boldsymbol{x}_s \times \boldsymbol{F}_s^{\mathrm{EXT}} + \sum_{s}^{N}\sum_{r\neq s} \boldsymbol{x}_s \times \boldsymbol{F}_{sr} \ .$$

In the double sum group the terms in pairs. Using Equation (19.41) in the form $\boldsymbol{F}_{rs} = -\boldsymbol{F}_{sr}$, the pair (s, r) contributes

$$\boldsymbol{x}_s \times \boldsymbol{F}_{sr} + \boldsymbol{x}_r \times \boldsymbol{F}_{rs} = (\boldsymbol{x}_s - \boldsymbol{x}_r) \times \boldsymbol{F}_{sr} \ ,$$

which vanishes precisely because Equation (19.53) makes the two factors parallel. Every pair vanishes, so the double sum does, leaving Equation (19.54). □

Theorem 19.63 (The spin obeys the same law about the centre of mass)**.** *Under the same hypotheses, and with $\boldsymbol{S}$ the internal angular momentum of Equation (19.51),*

$$\boxed{\frac{\mathrm{d}\boldsymbol{S}}{\mathrm{d}t} = \boldsymbol{\tau}'_{EXT}}, \qquad \boldsymbol{\tau}'_{EXT} := \sum_{s}^{N} \boldsymbol{x}'_s \times \boldsymbol{F}_s^{EXT}, \tag{19.55}$$

and this holds even when the centre of mass is accelerating, *so that its rest frame is not inertial.*

Rests on Theorem 19.58, Lemma 19.57, and Postulate 19.61.

Proof. *Derives Theorem 19.63.*

Differentiating $\boldsymbol{S}$ as defined in Equation (19.51), the term $\boldsymbol{v}'_s \times m_s \boldsymbol{v}'_s$ vanishes for each s, and by Equations (19.40) and (19.49)

$$m_s \frac{\mathrm{d}\boldsymbol{v}'_s}{\mathrm{d}t} = m_s \left(\boldsymbol{a}_s - \boldsymbol{a}_{\mathrm{CM}}\right) = \boldsymbol{F}_s - m_s \boldsymbol{a}_{\mathrm{CM}} \ ,$$

using Equation (19.4) for the s-th particle. Hence

$$\frac{\mathrm{d}\boldsymbol{S}}{\mathrm{d}t} = \sum_{s}^{N} \boldsymbol{x}'_s \times \boldsymbol{F}_s - \left(\sum_{s}^{N} m_s \boldsymbol{x}'_s \right) \times \boldsymbol{a}_{\mathrm{CM}} \,,$$

and the second term vanishes by Equation (19.50) — which is exactly why the accelerating centre of mass costs nothing here, and why no other point of the body enjoys the same privilege. The internal contributions to the first term cancel pair by pair as in Theorem 19.62, since $\boldsymbol{x}'_s - \boldsymbol{x}'_r = \boldsymbol{x}_s - \boldsymbol{x}_r$ by Equation (19.49), leaving Equation (19.55). □

Corollary 19.64 (The orbital part obeys the point-particle law)**.** *Differentiating Equation* (19.48) *and using Equation* (19.46)*,*

$$\frac{\mathrm{d}\boldsymbol{L}_{CM}}{\mathrm{d}t} = \boldsymbol{x}_{CM} \times \boldsymbol{F}_{EXT} \,, \tag{19.56}$$

so the centre of mass turns about a fixed origin exactly as a single particle of mass M at $\boldsymbol{x}_{CM}$ would. Together with Equation (19.55) *this splits the rotational dynamics of any system into an orbital equation and a spin equation that can be solved separately.*

Rests on Equations (19.46) *and* (19.48)*.*

Phenomenon 19.65 (Spin is conserved when no external torque acts)**.** A skater turning on the spot spins faster the closer the arms are drawn to the axis, and slower as they are extended again, without any push from outside; a diver in the air completes a somersault by tucking; a falling cat that begins with no rotation at all can turn itself over and land on its feet, ending as it began with zero total internal angular momentum. In each case the body changes its own shape, hence its moment of inertia Equation (19.33), while nothing external exerts a torque about the centre of mass.

Rests on Theorem 19.63.

Derivation. *Derives Phenomenon 19.65.*

Gravity acting on a body near the ground contributes $\sum_s \boldsymbol{x}'_s \times m_s \boldsymbol{g} = (\sum_s m_s \boldsymbol{x}'_s) \times \boldsymbol{g} = \boldsymbol{0}$ by Equation (19.50), so a uniform gravitational field exerts no torque about the centre of mass. With the skate pivot and the air resistance also contributing nothing appreciable, $\boldsymbol{\tau}'_{\mathrm{EXT}} = \boldsymbol{0}$ and Equation (19.55) gives $\boldsymbol{S}$ constant. For a body turning about a fixed principal axis (Remark 19.46) the component of $\boldsymbol{S}$ along the axis is $I\omega$ with I the moment of inertia Equation (19.33) about it, so

$$I_1 \omega_1 = I_2 \omega_2 \,, \tag{19.57}$$

and drawing the arms in, which lowers I, raises ω in the same proportion. Note what *is* not conserved: by Equation (19.52) the rotational kinetic energy is $\frac{1}{2} I \omega^2 = \frac{1}{2} S^2 / I$, which *rises* as I falls — the skater's muscles do the work, and the energy budget is what distinguishes this from a free body. The cat is subtler, since its axis is not fixed and I_{ij} changes shape as well as scale; $\boldsymbol{S} = \boldsymbol{0}$ throughout, and the reorientation is possible because the body is not rigid. Both cases are treated properly in Chapter 29 [32]. □

Remark 19.66 (The strong form is a real assumption, and it fails)*.* Postulate 19.61 holds for gravitation, for the Coulomb force and for elastic contact — for every force of the central kind of Corollary 19.14 — and for those Equation (19.54) is exact. It fails for the magnetic force between two moving charges, which is not directed along the line joining them and for which even the weak form Equation (19.41) can fail instantaneously. The angular momentum of the two charges alone is then not conserved; what is conserved is the total including the angular momentum stored in the electromagnetic field, as Chapter 60 shows. This is the characteristic pattern of every later repair to the Newtonian scheme: the conservation law is not abandoned but enlarged, by finding the carrier of what appeared to have been lost.

19.5.5 Conservation laws

Everything a closed system does leaves certain numbers unchanged. There are two ways to see which numbers those are, and this book uses both because they have different reach.

The first is direct: apply the laws already postulated and see what their time derivative turns out to be. This route asks nothing of the forces beyond Equation (19.41), so it survives friction, plastic deformation and every other dissipative process, and it is what justifies the everyday use of momentum conservation in collision problems where the interaction is entirely unknown.

The second is by symmetry. Each of the ten parameters of the Galilei group Equation (18.11) that leaves a system's description unchanged yields one conserved quantity, by Noether's first theorem Theorem 16.84 [456]. That route explains *why* these particular quantities and no others, but it requires the dynamics to come from an action, which excludes the dissipative cases the first route covers.

The three laws, directly

Theorem 19.67 (Conservation of linear momentum)**.** *For a closed system of constant masses subject to no external force, the total momentum*

$$\boldsymbol{P} = \sum_{s}^{N} \boldsymbol{p}_s = M \boldsymbol{v}_{CM}\,, \tag{19.58}$$

of SI unit kg m/s*, is constant in time.*

Rests on Theorem 19.52 and Equation (19.45)*.*

Proof. *Derives Theorem 19.67.*

By Equation (19.45) the sum is $\boldsymbol{p}_{\mathrm{CM}}$, and by Theorem 19.52 $\mathrm{d}\boldsymbol{P}/\mathrm{d}t = \boldsymbol{F}_{\mathrm{EXT}}$, which vanishes by hypothesis. Nothing was assumed about the internal forces except that they obey Equation (19.41); in particular they may be dissipative, which is why the total momentum is conserved in a perfectly inelastic collision, as recorded in Phenomenon 19.50, while the kinetic energy is not. □

Theorem 19.68 (Conservation of angular momentum)**.** *For a closed system of constant masses whose internal forces satisfy Postulate 19.61, and on which the external torque about a fixed point of an inertial frame vanishes, the total angular momentum about that point is constant. If the external torque about the centre of mass vanishes, the spin $\boldsymbol{S}$ is constant separately.*

Rests on Theorems 19.62 and 19.63.

Proof. *Derives Theorem 19.68.*

Immediate from Equation (19.54) and Equation (19.55) with the respective right-hand sides set to zero. The two statements are independent, because Equation (19.51) splits the total into two pieces obeying separate equations: a body may conserve its spin while its orbital angular momentum changes, which is the case of a spinning projectile in flight. □

Theorem 19.69 (Uniform motion of the centre of mass)**.** *For a closed system of constant masses subject to no external force, the quantity*

$$\boldsymbol{G} = \boldsymbol{P}t - M\boldsymbol{x}_{CM}\,, \tag{19.59}$$

of SI unit kg m*, is constant in time. Equivalently, the centre of mass moves in a straight line with constant velocity.*

Rests on Theorem 19.67 and Equation (19.44)*.*

Proof. *Derives Theorem 19.69.*
Differentiating Equation (19.59),

$$\frac{\mathrm{d}\boldsymbol{G}}{\mathrm{d}t} = \boldsymbol{P} + t\,\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} - M\boldsymbol{v}_{\mathrm{CM}} = \boldsymbol{P} - \boldsymbol{P} = \boldsymbol{0}\,,$$

using $\mathrm{d}\boldsymbol{P}/\mathrm{d}t = \boldsymbol{0}$ from Theorem 19.67 and $\boldsymbol{P} = M\boldsymbol{v}_{\mathrm{CM}}$ from Equation (19.44). Solving Equation (19.59) for $\boldsymbol{x}_{\mathrm{CM}}$ gives $\boldsymbol{x}_{\mathrm{CM}}(t) = \boldsymbol{x}_{\mathrm{CM}}(0) + \boldsymbol{v}_{\mathrm{CM}}t$, the second form of the statement. □

Remark 19.70 (The boost law is usually left unnamed). Theorem 19.69 looks like a restatement of Theorem 19.67 and is often passed over as one, but it is a separate conserved quantity: $\boldsymbol{G}$ depends explicitly on the time, and knowing $\boldsymbol{P}$ does not fix the constant $\boldsymbol{x}_{\mathrm{CM}}(0)$ that $\boldsymbol{G}$ encodes. The three components of $\boldsymbol{G}$ are what the three boost parameters of Equation (18.11) contribute, and without them the count in Table 19.1 would come to seven rather than ten. It is also the one of the four whose Noether derivation is not routine, for the reason given in Remark 19.72.

Energy is the remaining kind, and it was settled in Theorem 19.32: for a closed system whose internal forces derive from a potential in the mutual distances, E is constant with respect to every inertial frame, though it takes different values in different ones by Equation (19.24). Unlike the three above, it requires the internal forces to be conservative — exactly the condition Equation (19.20) tests.

Table 19.1: The ten conserved quantities of a closed Newtonian system, one for each parameter of the Galilei group Equation (18.11). The count is the dimension of the group, and it is no coincidence: each entry is the Noether charge Equation (16.76) of the corresponding one-parameter subgroup.

Symmetry	Parameters	Conserved quantity	SI unit
Time translation	1	Energy E	J
Space translation	3	Momentum $\boldsymbol{P}$	kg m/s
Rotation	3	Angular momentum $\boldsymbol{L}$	kg m^2/s
Boost	3	$\boldsymbol{G} = \boldsymbol{P}t - M\boldsymbol{x}_{\mathrm{CM}}$	kg m

The same laws from symmetry

Chapter 16 proves both of Noether's theorems, and the first of them (Theorem 16.84) is stated exactly in the form needed here: one independent variable, several dependent ones. Its two standard corollaries are already the first two entries of Table 19.1 — Corollary 16.87 for energy from invariance under time translation, and Corollary 16.88 for momentum and angular momentum from invariance under space translation and rotation. Nothing of that needs repeating; what belongs here is the identification of the mechanical system to which it applies, and the one case the corollaries do not cover.

Proposition 19.71 (The Newtonian closed system is a Lagrangian system)**.** *Let a closed system of N particles of constant masses have internal forces derived from a potential U depending only on the mutual distances. Then the equations of motion Equation (19.4) for all N particles are the Euler–Lagrange equations of the action Equation (16.86) built from*

$$\mathcal{L} = \sum_{s}^{N} \frac{1}{2} m_s \left|\dot{\boldsymbol{x}}_s\right|^2 - U\left(\{|\boldsymbol{x}_s - \boldsymbol{x}_r|\}_{s<r}\right)\,, \tag{19.60}$$

in J*, and* $\mathcal{L}$ *is invariant under all ten transformations Equation* (18.11) *up to a total time derivative.*

Rests on Theorem 16.97, Equation (19.4)*, and Definition 19.30.*

Proof. *Derives Proposition 19.71.*

Theorem 16.97 establishes the equivalence for one particle in a potential; the N-particle statement is the same computation carried out for each of the $3N$ coordinates in turn, the Euler–Lagrange system Equation (16.30) being one equation per dependent variable. For the invariance: the kinetic term is a sum of squared speeds, unchanged by a rotation and by a time or space translation, and U depends only on the distances $|\boldsymbol{x}_s - \boldsymbol{x}_r|$, which Proposition 19.13 showed to be invariant under the whole group. The one transformation that does not leave $\mathcal{L}$ strictly unchanged is the boost, and it changes it by a total derivative, as Remark 19.72 computes. □

Remark 19.72 (The boost charge needs the divergence term)*.* Under the one-parameter family of boosts $\boldsymbol{x}_s \longmapsto \boldsymbol{x}_s + \varepsilon \boldsymbol{u}\, t$, with $\boldsymbol{u}$ a fixed velocity, the Lagrangian Equation (19.60) is not invariant: to first order in ε it acquires

$$\mathcal{L} \longmapsto \mathcal{L} + \varepsilon \sum_{s}^{N} m_s \dot{\boldsymbol{x}}_s \cdot \boldsymbol{u} = \mathcal{L} + \varepsilon \frac{\mathrm{d}}{\mathrm{d}t} \left(M \boldsymbol{x}_{\mathrm{CM}} \cdot \boldsymbol{u} \right) , \tag{19.61}$$

by Equation (19.38). This is precisely the case Theorem 16.84 admits with a nonzero divergence term Λ, and taking $\xi = 0$, $\phi_s = \boldsymbol{u}t$ and $\Lambda = M\boldsymbol{x}_{\mathrm{CM}} \cdot \boldsymbol{u}$ in the Noether charge Equation (16.76) gives

$$I = \sum_{s}^{N} \boldsymbol{p}_s \cdot \boldsymbol{u}\, t - M\boldsymbol{x}_{\mathrm{CM}} \cdot \boldsymbol{u} = \boldsymbol{u} \cdot \boldsymbol{G} ,$$

with $\boldsymbol{G}$ the quantity Equation (19.59) of Theorem 19.69, and $\boldsymbol{u}$ arbitrary. Had Λ been dropped, the charge would have come out as $\boldsymbol{P} \cdot \boldsymbol{u}\, t$, which is not conserved. The term that cannot be dropped is the same object that appears in Chapter 14 as the central charge of the Galilei algebra — the mass, in Example 14.80 — and in Chapter 25 as the constant defect in the Poisson brackets of a free particle (Proposition 25.14). Three descriptions of one fact: the free particle realizes not the Galilei algebra but its central extension by the mass.

Remark 19.73 (Which route is stronger, and where)*.* The two derivations are not redundant, and neither contains the other.

- The direct route is stronger in *scope*. It needs only Equation (19.41) and yields Theorem 19.67 for internal forces of any kind, dissipative ones included. A perfectly inelastic collision has no Lagrangian description of the colliding bodies alone, yet conserves momentum exactly.

- The symmetry route is stronger in *explanation*. It says why there are ten conserved quantities and not eleven, ties each to a geometrical property of space and time, and generalizes: replacing the Galilei group by the Poincaré group repeats the argument unchanged and delivers the relativistic conservation laws of Chapter 40, while the field-theoretic form Theorem 16.90 delivers those of Chapter 64.

Both are used in what follows. The systematic development of the symmetry route for mechanical systems, with generalized coordinates and constraints, is Chapter 21; the conserved quantities reappear there as first integrals and cyclic coordinates (Proposition 16.28), and again in Chapter 22 as generators of the very transformations they are conserved under.

The Newtonian formulation presented in this chapter provides the equations of motion whenever all the forces are known explicitly. When the motion is restricted by constraints, however, the constraint forces are not known in advance, and the formulation must be recast in terms of generalized coordinates; this is the subject of Chapter 21. The special cases of equilibrium and of central forces — the latter the gateway to celestial mechanics — are developed in Chapter 27.

Chapter 20

Experiment: Free Fall and Projectile Motion

Every experimental chapter of this treatise follows the same arc: a phenomenon is stated as it is observed, the observation is confronted with a derivation from explicitly declared assumptions, and the residual disagreement — there is always one — marks the domain of validity of those assumptions. This chapter is the exemplar. Its subject, the free fall of heavy bodies and the flight of projectiles, is the oldest quantitative experiment in physics, and it remains one of the most precise: the same law of fall that Galileo extracted from a bronze ball and a water clock is today verified, in ballistic gravimeters, to nine significant figures. The derivation given in Section 20.2 is deliberately complete — every assumption numbered, every integration displayed — because it is the pattern to which all later derivations in this book, whether inline, in Appendix A, or pending, are referred.

20.1 Historical context: the Aristotelian expectation and its collapse

For nearly two thousand years the educated expectation about falling bodies was Aristotle's: that a body falls with a speed proportional to its weight and inversely proportional to the resistance of the medium, so that a stone of ten pounds should reach the ground ten times faster than a stone of one pound. That doctrine, together with the companion difficulty of explaining why a projectile keeps moving after it leaves the hand, is rehearsed at length — and demolished — in the book that founded modern mechanics, Galileo's *Discorsi e dimostrazioni matematiche intorno a due nuove scienze*, composed under house arrest at Arcetri and printed by the Elzevirs at Leiden in 1638 [4]. There the Aristotelian position is put in the mouth of Simplicio, and Salviati answers it first with an argument requiring no apparatus at all: if a heavy body fell faster than a light one, then tying the two together should produce a compound body that falls both faster (being heavier than either) and slower (the light body retarding the heavy one) — a contradiction, from which Galileo concludes that all bodies must fall alike once the disturbance of the medium is removed [4].

The argument disposes of the Aristotelian *ratio*, but it does not by itself yield the law of fall. That required measurement, and here lies the historical significance of the experiments described in Section 20.3: free fall was the first phenomenon in history to be captured by a mathematical law that was extracted from, and tested against, quantitative data with an explicit appeal to reproducibility. The Third Day of the *Discorsi* defines uniformly accelerated motion, proves as a theorem that the distances traversed from rest are as the squares of the elapsed times, and derives as a corollary the *odd-number rule* — that the spaces covered in equal successive intervals of time stand in the ratios $1 : 3 : 5 : 7 : \dots$ — before

reporting the inclined-plane measurements that confirm it; the Fourth Day then compounds uniform horizontal motion with naturally accelerated vertical motion and demonstrates that the resulting trajectory is a semi-parabola [4]. The conceptual order — definition, theorem, experiment — is precisely the one this treatise adopts.

What Galileo established kinematically, Newton grounded dynamically. In the *Principia* of 1687 the parabolic trajectory appears no longer as a brute geometrical fact but as the immediate consequence of the second law of motion applied to a uniform downward force, and Newton, in the scholium to the laws, credits Galileo with exactly this pair of discoveries: the squared-time law of descent and the parabola of projectiles [14]. The present chapter retraces that inference in modern notation, and then holds it against four centuries of data.

20.2 The observed phenomenon and its derivation

Phenomenon 20.1. Near the surface of the Earth, and wherever the resistance of the air may be neglected, projectiles follow arcs that are parabolic: the height depends quadratically on the horizontal distance, the axis of the arc is vertical, and the arc opens downward. Moreover a body released from rest falls through distances that grow as the square of the elapsed time, so that the spaces traversed in equal successive intervals stand in the ratios $1 : 3 : 5 : 7 : \dots$ (Galileo's odd-number rule). Both regularities are independent of the weight and composition of the body [4].

Rests on Equation (18.7) *and Phenomenon 18.12.*

This phenomenon admits a short, complete derivation from Newtonian dynamics, which we now give in full; it is the model derivation of the treatise.

Theorem 20.2 (Parabolic trajectory)**.** *Consider a point projectile launched from the origin with speed $v_0 > 0$ at an angle $\theta \in (0, 90^\circ)$ above the horizontal, in a uniform gravitational field and in the absence of air resistance, the ground being treated as flat and non-rotating. Then the motion remains in the vertical plane containing the initial velocity, and the trajectory is the arc of a parabola with vertical axis, opening downward,*

$$z = x \tan\theta - \frac{g\,x^2}{2v_0^2 \cos^2\theta}\,. \tag{20.1}$$

Rests on Postulate 19.8, Equation (18.7)*, and Phenomenon 20.1.*

Derivation. *Derives Theorem 20.2.*

We first state the assumptions explicitly.

1. *Point mass.* The projectile is treated as a point particle of mass m; its extension, internal structure and rotation are ignored.

2. *Uniform field.* The gravitational field is $\boldsymbol{g} = -g\,\hat{\boldsymbol{z}}$, constant in magnitude and direction over the region of flight; numerically g is close to the conventional standard value $g_0 = 9.806\,65\,\mathrm{m/s^2}$ [2333].

3. *No air resistance.* The only force acting during flight is gravity.

4. *Flat, non-rotating Earth.* The laboratory frame is treated as inertial: the curvature of the Earth and the centrifugal and Coriolis accelerations of its rotation are neglected.

Under assumptions 1–4, Newton's second law [14, 2027] reads

$$m\,\ddot{\boldsymbol{r}} = -mg\,\hat{\boldsymbol{z}}\,. \tag{20.2}$$

The mass cancels — an innocuous-looking step whose depth is the subject of Section 20.6 — leaving the kinematic statement $\ddot{\boldsymbol{r}} = -g\,\hat{\boldsymbol{z}}$. Since the right-hand side is constant, two elementary integrations suffice. Integrating once from the initial condition $\dot{\boldsymbol{r}}(0) = \boldsymbol{v}_0$,

$$\dot{\boldsymbol{r}}(t) = \boldsymbol{v}_0 - g t\,\hat{\boldsymbol{z}}\,, \tag{20.3}$$

and integrating again from $\boldsymbol{r}(0) = \boldsymbol{0}$,

$$\boldsymbol{r}(t) = \boldsymbol{v}_0\, t - \tfrac{1}{2} g t^2\,\hat{\boldsymbol{z}}\,. \tag{20.4}$$

Choose the x-axis along the horizontal component of $\boldsymbol{v}_0$, so that $\boldsymbol{v}_0 = (v_0\cos\theta,\, 0,\, v_0\sin\theta)$. The y component of Equation (20.4) vanishes identically: the motion is confined to the vertical plane $y = 0$, as claimed. In that plane the parametric solution is

$$x(t) = v_0\cos\theta\; t\,, \qquad z(t) = v_0\sin\theta\; t - \tfrac{1}{2} g t^2\,. \tag{20.5}$$

Because $\theta \neq 90^\circ$ we have $\cos\theta \neq 0$, and the first equation can be solved for the time, $t = x/(v_0\cos\theta)$. Substituting into the second eliminates t and yields Equation (20.1):

$$z = x\tan\theta - \frac{g\,x^2}{2v_0^2\cos^2\theta}\,.$$

This is a polynomial of degree exactly two in x with constant coefficients; its leading coefficient $-g/(2v_0^2\cos^2\theta)$ is strictly negative. A quadratic graph $z = ax^2 + bx$ with $a < 0$ is a parabola whose axis is parallel to the z-axis and which opens downward — completing the square,

$$z - z_{\mathrm{apex}} = -\frac{g}{2v_0^2\cos^2\theta}\,\bigl(x - x_{\mathrm{apex}}\bigr)^2\,, \tag{20.6}$$

with the apex coordinates given in the corollary below. This proves the first claim of Phenomenon 20.1.

For the second claim, set $\theta = 90^\circ$ and reverse the sign convention so that $s(t)$ denotes the distance fallen from rest; then Equation (20.5) with $v_0 = 0$ gives

$$s(t) = \tfrac{1}{2} g t^2\,. \tag{20.7}$$

Partition time into equal intervals of length τ. The distance covered during the n-th interval is

$$\Delta_n = s(n\tau) - s\bigl((n-1)\tau\bigr) = \tfrac{1}{2} g\tau^2\bigl(n^2 - (n-1)^2\bigr) = \tfrac{1}{2} g\tau^2\,(2n-1)\,, \tag{20.8}$$

so that $\Delta_1 : \Delta_2 : \Delta_3 : \cdots = 1 : 3 : 5 : \ldots$ — Galileo's odd-number rule [4]. Finally, nothing in the derivation refers to m after the cancellation in Equation (20.2), which reproduces the observed independence of the motion from the weight and composition of the body. □

Corollary 20.3 (Range, apex, time of flight)**.** *With launch and landing at the same height, the time of flight, the horizontal range and the apex height are*

$$T = \frac{2v_0\sin\theta}{g}\,, \qquad R = \frac{v_0^2\sin 2\theta}{g}\,, \qquad H = \frac{v_0^2\sin^2\theta}{2g}\,, \tag{20.9}$$

and for fixed v_0 the range is greatest at $\theta = 45^\circ$, where $R_{\max} = v_0^2/g$.

Rests on Theorem 20.2 and Equation (20.5).

Proof. *Derives Corollary 20.3.*

From Equation (20.5), $z(t) = t\,(v_0\sin\theta - \tfrac{1}{2} g t)$ vanishes at $t = 0$ and at $t = T = 2v_0\sin\theta/g$. The range is $R = x(T) = v_0\cos\theta\cdot T = 2v_0^2\sin\theta\cos\theta/g = v_0^2\sin 2\theta/g$, using the double-angle identity. Since $\sin 2\theta \le 1$ with equality precisely when $2\theta = 90^\circ$, the range is maximal at $\theta = 45^\circ$. The apex occurs where $\dot z = v_0\sin\theta - g t = 0$, i.e. at $t = T/2$, by which symmetry of the parabola is recovered; substitution gives $H = v_0^2\sin^2\theta/(2g)$. □

Example 20.4. A ball thrown at $v_0 = 10\,\mathrm{m/s}$ and $\theta = 45^\circ$ has, by Equation (20.9) with $g = g_0$, a time of flight $T \approx 1.44\,\mathrm{s}$, a range $R = v_0^2/g_0 \approx 10.2\,\mathrm{m}$ and an apex height $H \approx 2.55\,\mathrm{m}$. These are the magnitudes of everyday experience, and the fact that everyday experience does *not* flagrantly contradict them is already a statement about air drag: for dense projectiles at such speeds the neglect of resistance in assumption 3 is a percent-level idealisation (Section 20.5).

Rests on Corollary 20.3.

Remark 20.5. The derivation used nothing but the second law and two integrations, yet its logical anatomy repays attention. Each of the four assumptions is independently falsifiable, and each *does* fail at some level of precision: assumption 1 for spinning or extended bodies, assumption 2 over heights where the field gradient matters (Section 20.4), assumption 3 for fast or light projectiles, and assumption 4 for long-range or long-duration motion (Section 20.5). The parabola of Equation (20.1) is therefore not "the" trajectory of a real projectile but the zeroth term of a systematic approximation — which is exactly what an idealised law of nature is.

Rests on Theorem 20.2 and Equation (20.1).

Two experiments anchor the phenomenon, separated by almost four centuries and by eight orders of magnitude in timing precision: Galileo's inclined planes, which established the form of the law, and modern ballistic gravimetry, which turns the law around and uses free fall as a measuring instrument.

20.3 Galileo's Inclined Planes (1604–1638)

Tests Phenomenon 20.1 and Equation (20.8).

Assuming Equation (18.7).

20.3.1 Apparatus

A hard, smooth, very round ball of bronze, and a straight wooden beam or moulding some twelve *braccia* long (roughly 7 m), half a braccio wide and three finger-breadths thick, along whose upper face a channel a little more than one finger in breadth was cut, planed straight, and lined with polished parchment to make it as smooth as possible. One end of the beam could be raised one or two braccia above the horizontal, tilting the groove into an inclined plane of small, adjustable slope. Time was measured with a water clock: a large elevated vessel discharged through a thin pipe into a beaker during the motion, and the collected water was weighed after each run on a sensitive balance, the mass of the effluent serving as the measure of the elapsed time [4].

20.3.2 Procedure

The ball was released from rest at the top of the groove and timed over the full length of the channel; then over one quarter of the length, one half, two thirds, and other fractions; the tilt of the plane was varied; and, in Galileo's own report, each configuration was repeated "a full hundred times" to establish the reproducibility of the timings [4].

20.3.3 Observations and data

The distances traversed from rest were found proportional to the squares of the measured times, at every inclination of the plane: the full length was covered in exactly twice the time of the first quarter, and the spaces covered in equal successive time intervals followed the odd-number progression,

interval n	space in interval	cumulative space
1	1	1
2	3	4
3	5	9
4	7	16

(in units of the distance covered in the first interval), in agreement with Equation (20.8). Galileo states that repeated timings of the same descent never differed by more than one tenth of a pulse beat — of order 0.1 s on a total descent time of several seconds — and that the result did not depend on the weight of the ball [4].

20.3.4 Interpretation

The measurements identify the motion as uniformly accelerated: equal increments of speed in equal times, the definition adopted in the Third Day of the *Discorsi* [4]. The inclined plane is itself the decisive methodological invention: rolling down a slope of inclination α dilutes the effective acceleration by the factor $\sin\alpha$, slowing the motion until a water clock of 0.1 s resolution can resolve it, while leaving the *form* of the law — distance proportional to time squared — untouched. Extrapolating the diluted experiments to vertical inclination, Galileo obtained the law of free fall, Equation (20.7), without ever timing a free vertical drop. The observed independence from the ball's weight is the seed of the universality discussed in Section 20.6. (A modern reader will note what the water clock could not resolve: rolling, as opposed to sliding, reduces the acceleration by a further constant factor — 5/7 for a homogeneous sphere — which drops out of all of Galileo's *ratios* of distances and times, and so left his conclusions intact.)

20.3.5 Primary references

[4]; Newton's dynamical reading of these results is in the scholium to the laws of motion [14].

20.4 Absolute Ballistic Gravimetry (1995–present)

Tests Equations (18.7) *and* (20.7).
Assuming Phenomenon 20.1.

20.4.1 Apparatus

The FG5 class of absolute gravimeters realises Galileo's experiment with interferometric metrology [1906]. The falling body is a corner-cube retroreflector enclosed in a co-falling "drag-free" carriage inside an evacuated drop chamber, the carriage tracking the cube so as to shield it from residual gas drag and to carry the electrostatic environment with it. The falling cube forms one arm of a Michelson-type interferometer; the reference cube is suspended from an actively servo-controlled long-period isolation stage (a "superspring") to decouple it from ground vibration. The length standard is an iodine-stabilised helium–neon laser at 633 nm; the time standard is a rubidium atomic clock [1906].

20.4.2 Procedure

The carriage releases the cube, which falls freely through about 20 cm in about 0.2 s. Interference fringes — one per 316.5 nm of fall, half the laser wavelength — are counted and time-stamped against the atomic clock, and the fringe record is scaled to roughly 200 position–time pairs per drop. Each drop is fitted by least squares to the free-fall trajectory

of Equation (20.5) augmented by the vertical field gradient, and standard corrections are applied for solid-Earth tides, ocean loading, polar motion and local barometric pressure. A measurement campaign comprises hundreds of drops per hour, combined over one or more days; the statistical treatment of the drop-to-drop scatter follows the standard error analysis of repeated measurements [1906, 1945].

20.4.3 Observations and data

quantity	value
conventional standard g_0, exact [2333]	$9.806\,65\,\mathrm{m/s^2}$
FG5 instrumental accuracy [1906]	$\approx 2\times10^{-8}\,\mathrm{m/s^2}$ ($2\,\mu$Gal)
drop-to-drop scatter [1906]	$\approx 5\times10^{-8}\,\mathrm{m/s^2}$
geographic variation of g (equator–pole, altitude)	$\approx \pm 0.03\,\mathrm{m/s^2}$
free-air height gradient [1906]	$\approx -3.1\times10^{-6}\,\mathrm{m/s^2}$ per metre

Here $1\,\mathrm{Gal} = 1\times10^{-2}\,\mathrm{m/s^2}$ is the conventional unit of the gravimetric literature; the mean of a measurement set converges on the drop-to-drop scatter as $1/\sqrt{N}$ in the number of drops [1945]. The fitted trajectories reproduce $s(t) = s_0 + v_0 t + \tfrac{1}{2}gt^2$ at the nanometre level over the whole drop, and the instrument accuracy of $2\times10^{-8}\,\mathrm{m/s^2}$ against $g \approx 9.8\,\mathrm{m/s^2}$ means that the local acceleration of free fall is determined to about two parts in 10^9 — nine significant figures [1906].

20.4.4 Interpretation

Three centuries of refinement have inverted the logic of the experiment. Galileo needed the inclined plane to make the law of fall measurable; the FG5 assumes the law — Newtonian free fall in a locally uniform field with a known gradient — and measures its single parameter g so precisely that the instrument serves geodesy and metrology: monitoring ground-water and magma movements, realising the unit of force, and calibrating the watt-balance experiments that underpin the 2019 redefinition of the SI [2333, 2392]. Equally important is what the fits do *not* show: dropped test bodies of different construction yield the same g within the stated uncertainty, in line with the far sharper torsion-balance (Eötvös-type) tests of the universality of free fall discussed in Chapter 42. The residual systematic budget — laser stability, clock calibration, vertical alignment, residual drag — is itself a textbook exercise in the error analysis of Chapter 2 [1945].

20.4.5 Primary references

[1906, 2392]; conventions for g_0 in [2333].

20.5 Deviations and the domain of validity

The parabola of Equation (20.1) is exact only within assumptions 1–4 of the derivation. Each assumption fails observably, and the pattern of failure is itself lawful. We record here the two deviations that dominate at laboratory and ballistic scales, air resistance and the Earth's rotation, stating in each case what is observed; the corresponding derivations are pending and will follow the standards set by the exemplar above.

Phenomenon 20.6 (Air drag)**.** Real trajectories in air fall short of Equation (20.9) and are asymmetric: the descending branch is steeper than the ascending one, and the range at fixed muzzle speed is maximised at launch angles *below* 45°. A body falling for long enough ceases to accelerate altogether and reaches a constant terminal speed — raindrops arrive at

metres per second, not at the hundreds of metres per second that Equation (20.7) would assign to their fall height [14, 384].

Rests on Theorem 20.2 and Corollary 20.3.

The observed deviations follow from adding to Equation (20.2) a resistive force directed against the velocity,

$$m\ddot{\boldsymbol{r}} = -mg\,\hat{\boldsymbol{z}} - f(|\boldsymbol{v}|)\,\boldsymbol{v}\,, \qquad f > 0\,, \tag{20.10}$$

in which f carries the SI unit kg/s. Two regimes cover the practical cases [1585]. At low Reynolds number the resistance is linear in the velocity, $f = b$ constant; at the Reynolds numbers of a macroscopic projectile in air it is quadratic, $f = c\,|\boldsymbol{v}|$ with $c = \tfrac{1}{2}\rho\,C_{\mathrm{D}}A$ in kg/m, built from the density ρ of the fluid, the frontal area A and a dimensionless drag coefficient C_{D} which is itself a function of the Reynolds number and which drops abruptly at the drag crisis of a sphere [384].

Proposition 20.7 (Terminal speed and the approach to it)**.** *Let a body fall from rest along the vertical under Equation* (20.10)*, and let $v > 0$ be its downward speed. Then v increases monotonically towards the* terminal speed v_{t} *fixed by $f(v_{\mathrm{t}})\,v_{\mathrm{t}} = mg$,*

$$v_{\mathrm{t}} = \frac{mg}{b} \quad \textit{(linear)}\,, \qquad v_{\mathrm{t}} = \sqrt{\frac{mg}{c}} \quad \textit{(quadratic)}\,, \tag{20.11}$$

and the approach to it is

$$v(t) = v_{\mathrm{t}}\left(1 - \mathrm{e}^{-gt/v_{\mathrm{t}}}\right), \qquad v(t) = v_{\mathrm{t}}\tanh\!\left(\frac{gt}{v_{\mathrm{t}}}\right), \tag{20.12}$$

in the linear and quadratic cases respectively. In both the characteristic time is v_{t}/g.

Rests on Equation (20.10) *and Theorem 9.8.*

Proof. *Derives Proposition 20.7.*

Along the vertical, Equation (20.10) reads $m\,\mathrm{d}v/\mathrm{d}t = mg - f(v)\,v$. Dividing by m and using $f(v_{\mathrm{t}})v_{\mathrm{t}} = mg$ to eliminate the constants,

$$\frac{\mathrm{d}v}{\mathrm{d}t} = g\left(1 - \frac{v}{v_{\mathrm{t}}}\right), \qquad \frac{\mathrm{d}v}{\mathrm{d}t} = g\left(1 - \frac{v^2}{v_{\mathrm{t}}^2}\right), \tag{20.13}$$

for $f = b$ and for $f = c\,|v|$ respectively, with Equation (20.11) for v_{t} in each case. Each right-hand side is of class C^1 in v and hence Lipschitz on any bounded interval, so by Theorem 9.8 the initial-value problem with $v(0) = 0$ has exactly one solution; it is therefore enough to check that Equation (20.12) solves it. In the linear case $\mathrm{d}v/\mathrm{d}t = g\,\mathrm{e}^{-gt/v_{\mathrm{t}}} = g\,(1 - v/v_{\mathrm{t}})$; in the quadratic case the identity $\tanh' = 1 - \tanh^2$ gives $\mathrm{d}v/\mathrm{d}t = g\left(1 - \tanh^2(gt/v_{\mathrm{t}})\right) = g\,(1 - v^2/v_{\mathrm{t}}^2)$; and both vanish at $t = 0$. Monotonicity and the limit follow from Equation (20.13), whose right-hand side is strictly positive for $0 \le v < v_{\mathrm{t}}$ and vanishes only at $v = v_{\mathrm{t}}$, which is therefore approached but never attained in finite time. □

Example 20.8 (A raindrop)**.** For a spherical water drop of radius 1.0 mm — mass $m = \tfrac{4}{3}\pi r^3\rho_{\mathrm{w}} = 4.19 \times 10^{-6}$ kg, frontal area $A = \pi r^2 = 3.14 \times 10^{-6}\,\mathrm{m}^2$ — falling in air of density $\rho = 1.2\,\mathrm{kg/m}^3$ with $C_{\mathrm{D}} = 0.5$, Equation (20.11) gives

$$v_{\mathrm{t}} = \sqrt{\frac{2mg}{\rho\,C_{\mathrm{D}}A}} \approx 6.6\,\mathrm{m/s}\,, \tag{20.14}$$

reached to within one per cent after about 1.8 s of fall. Without resistance, Equation (20.7) would give a drop falling from 1 km an arrival speed of $\sqrt{2gh} \approx 140\,\mathrm{m/s}$ — a factor of twenty in speed and some four hundred in kinetic energy, which is the difference between rain and gunfire.

Rests on Proposition 20.7.

Proposition 20.9 (Linear resistance: exact motion, and the lost range)**.** *Let $f = b$ be constant and write $\tau = m/b = v_{\mathrm{t}}/g$. The solution of Equation (20.10) with $\boldsymbol{r}(0) = \mathbf{0}$ and $\dot{\boldsymbol{r}}(0) = \boldsymbol{v}_0$ is*

$$\dot{\boldsymbol{r}}(t) = (\boldsymbol{v}_0 + v_{\mathrm{t}}\hat{\boldsymbol{z}})\,\mathrm{e}^{-t/\tau} - v_{\mathrm{t}}\hat{\boldsymbol{z}}\,, \tag{20.15}$$

$$\boldsymbol{r}(t) = \tau\,(\boldsymbol{v}_0 + v_{\mathrm{t}}\hat{\boldsymbol{z}})\left(1 - \mathrm{e}^{-t/\tau}\right) - v_{\mathrm{t}}\,t\,\hat{\boldsymbol{z}}\,. \tag{20.16}$$

For launch and landing at the same height, and to first order in the small parameter v_0/v_{t}, the range is

$$R = R_0\left(1 - \frac{4}{3}\,\frac{v_0 \sin\theta}{v_{\mathrm{t}}}\right) + O\!\left(\frac{v_0^2}{v_{\mathrm{t}}^2}\right), \qquad R_0 = \frac{v_0^2 \sin 2\theta}{g}\,, \tag{20.17}$$

and $\mathrm{d}R/\mathrm{d}\theta < 0$ at $\theta = 45^\circ$, so that at fixed v_0 the range is greatest at a launch angle strictly below 45°.

Rests on Equation (20.10) and Corollaries 9.9 and 20.3.

Proof. *Derives Proposition 20.9.*

The solution. Differentiating Equation (20.16) returns Equation (20.15), and differentiating once more,

$$\ddot{\boldsymbol{r}} = -\frac{1}{\tau}\,(\boldsymbol{v}_0 + v_{\mathrm{t}}\hat{\boldsymbol{z}})\,\mathrm{e}^{-t/\tau} = -\frac{1}{\tau}\,(\dot{\boldsymbol{r}} + v_{\mathrm{t}}\hat{\boldsymbol{z}}) = -\frac{b}{m}\dot{\boldsymbol{r}} - g\hat{\boldsymbol{z}}\,,$$

since $v_{\mathrm{t}}/\tau = g$; the initial conditions are met by inspection. The system is linear with constant coefficients, so by Corollary 9.9 this is the only solution.

The range. Choose the x-axis along the horizontal component of $\boldsymbol{v}_0$ and write $v_{0x} = v_0\cos\theta$, $v_{0z} = v_0\sin\theta$. Expanding $1 - \mathrm{e}^{-u} = u - u^2/2 + u^3/6 + \dots$ in Equation (20.16) and keeping terms through first order in $1/\tau$, with $v_{\mathrm{t}}/\tau = g$,

$$x(t) = v_{0x}\left(t - \frac{t^2}{2\tau}\right), \qquad z(t) = v_{0z}t - \frac{1}{2}gt^2 - \frac{v_{0z}t^2}{2\tau} + \frac{gt^3}{6\tau}\,. \tag{20.18}$$

The time of flight solves $z(T) = 0$; dividing by T and writing $T = T_0 + T_1$ about the drag-free value $T_0 = 2v_{0z}/g$ of Equation (20.9), the residual at T_0 is

$$-\frac{v_{0z}T_0}{2\tau} + \frac{gT_0^2}{6\tau} = -\frac{v_{0z}^2}{g\tau} + \frac{2v_{0z}^2}{3g\tau} = -\frac{v_{0z}^2}{3g\tau}\,,$$

which against the slope $-g/2$ gives $T_1 = -2v_{0z}^2/\left(3g^2\tau\right)$, that is $T = T_0\,(1 - v_{0z}/3v_{\mathrm{t}})$. Substituting into the first of Equation (20.18),

$$R = v_{0x}T - \frac{v_{0x}T^2}{2\tau} = R_0\left(1 - \frac{v_{0z}}{3v_{\mathrm{t}}}\right) - R_0\frac{v_{0z}}{v_{\mathrm{t}}} = R_0\left(1 - \frac{4}{3}\frac{v_{0z}}{v_{\mathrm{t}}}\right),$$

using $R_0 = v_{0x}T_0 = 2v_{0x}v_{0z}/g$ in both terms, which is Equation (20.17).

The optimal angle. Write $R(\theta) = (v_0^2/g)\sin 2\theta\,(1 - \varepsilon\sin\theta)$ with $\varepsilon = 4v_0/3v_{\mathrm{t}}$. Then

$$\frac{\mathrm{d}R}{\mathrm{d}\theta} = \frac{v_0^2}{g}\left[2\cos 2\theta\,(1 - \varepsilon\sin\theta) - \varepsilon\sin 2\theta\cos\theta\right],$$

and at $\theta = 45^\circ$ the first bracket vanishes with $\cos 2\theta$, leaving $\mathrm{d}R/\mathrm{d}\theta = -\varepsilon v_0^2/\left(g\sqrt{2}\right) < 0$. The stationary point has therefore moved below 45°. □

Proposition 20.10 (The descending branch is the steeper, whatever the resistive law)**.** *Let the motion obey Equation* (20.10) *with any continuous $f > 0$, launched from the origin with $v_{0x} > 0$ and $v_{0z} > 0$ and followed until it returns to $z = 0$. Then v_x is positive and strictly decreasing throughout the flight, and the slope of the trajectory at landing exceeds the launch slope in magnitude,*

$$\left| \left. \frac{\mathrm{d}z}{\mathrm{d}x} \right|_{landing} \right| > \tan\theta\,. \tag{20.19}$$

With $f \equiv 0$ the two are equal, which is the symmetry of the parabola Equation (20.6).
Rests on Equations (20.6) *and* (20.10).

Proof. *Derives Proposition 20.10.*

In components, Equation (20.10) reads $m\dot{v}_x = -f v_x$ and $m\dot{v}_z = -mg - f v_z$. The first is linear and homogeneous in v_x with continuous coefficient, so $v_x(t) = v_{0x} \exp\left(-\int_0^t f/m\,\mathrm{d}t'\right) > 0$, and $\dot{v}_x = -(f/m)\,v_x < 0$: horizontal speed is positive and strictly decreasing. Hence x is strictly increasing and may be used as a parameter along the trajectory. For the slope $p = \mathrm{d}z/\mathrm{d}x = v_z/v_x$,

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{\dot{v}_z v_x - v_z \dot{v}_x}{v_x^2} = \frac{\left(-g - \frac{f}{m} v_z\right) v_x + \frac{f}{m} v_z v_x}{v_x^2} = -\frac{g}{v_x}\,, \qquad \frac{\mathrm{d}p}{\mathrm{d}x} = -\frac{g}{v_x^2}\,, \tag{20.20}$$

the resistance cancelling identically. So p decreases strictly and monotonically from $p = \tan\theta$ at launch, through $p = 0$ at the apex x_a, to $p = -q_\mathrm{f} < 0$ at landing; each value of the slope is attained exactly once, and p may itself be used as parameter, with $\mathrm{d}x = -(v_x^2/g)\,\mathrm{d}p$.

Let $v_\uparrow(u)$ and $v_\downarrow(u)$ denote the horizontal speed at the instants where the slope is $+u$ on the way up and $-u$ on the way down. The apex height is reached and then lost, so the two integrals

$$H = \int_0^{x_\mathrm{a}} p\,\mathrm{d}x = \frac{1}{g}\int_0^{\tan\theta} u\,v_\uparrow(u)^2\mathrm{d}u\,, \qquad H = \int_{x_\mathrm{a}}^{x_\mathrm{f}} (-p)\,\mathrm{d}x = \frac{1}{g}\int_0^{q_\mathrm{f}} u\,v_\downarrow(u)^2\mathrm{d}u \tag{20.21}$$

are equal. Every instant of the descent is later than every instant of the ascent and v_x is strictly decreasing, so $v_\downarrow(u) < v_\uparrow(u)$ wherever both are defined. Were $q_\mathrm{f} \le \tan\theta$, the second integral would be strictly smaller than the first, contradicting their equality; hence $q_\mathrm{f} > \tan\theta$, which is Equation (20.19). With $f \equiv 0$, v_x is constant, both integrands are $u\,v_{0x}^2$, and the two limits coincide. □

Derivation. *Derives Phenomenon 20.6.*

Each clause of Phenomenon 20.6 is now accounted for. The shortfall against Equation (20.9) is Equation (20.17), first order in v_0/v_t and negative. The asymmetry of the two branches is Proposition 20.10, which needs no assumption about the resistive law beyond its sign, and which is why the effect is seen for bodies of every kind. The optimum below 45° is the last claim of Proposition 20.9. The terminal speed is Proposition 20.7, and the numbers for a raindrop are Example 20.8: 6.6 m/s, not the 140 m/s that Equation (20.7) would assign to a fall of 1 km. □

Remark 20.11 (What has no closed form, and what to do about it). Proposition 20.9 is exact only for the linear law. The two-dimensional quadratic problem — Equation (20.10) with $f = c\,|\boldsymbol{v}|$ — couples the two components through $|\boldsymbol{v}| = \sqrt{v_x^2 + v_z^2}$ and has no solution in closed form; the vertical motion of Proposition 20.7 is integrable only because there the motion is one-dimensional and $|\boldsymbol{v}| = v$. What survives without integration is exactly what the three propositions above deliver, and the rest is obtained numerically. Whether a given experiment may ignore resistance is then an error-budget question, settled by comparing the estimated correction with the measurement uncertainty [1945]: Galileo's dense bronze

ball, at speeds of a few m/s on a plane whose water clock resolved 0.1 s, could; a musket ball could not.

Rests on Propositions 20.9 and 20.10.

Phenomenon 20.12 (Coriolis deflection)**.** A body dropped from rest in a tall shaft does not land at the plumb point below its release: it lands displaced towards the *east*, in either hemisphere, by an amount that grows with the fall height — about 1.5 cm for a 100 m drop at latitude 45° [92]. Long-range projectiles are likewise deflected sideways relative to the parabolic plane of Equation (20.1).

Rests on Equation (20.1) and Phenomenon 29.53.

Assumption 4 is the one at fault, and the argument that replaces it can be made without ever leaving an inertial frame — which is worth doing once, because the fictitious-force bookkeeping of Section 29.5 is easy to perform and easy to mistrust.

Proposition 20.13 (Eastward deflection of a body dropped from rest)**.** *Let a body be released from rest with respect to the ground, at a height $h \ll R$ above the surface of the Earth at geographic latitude λ, with R the Earth's radius and $\omega = 7.292 \times 10^{-5}\,\mathrm{rad/s}$ its sidereal rotation rate. To leading order in ω and in h/R it strikes the ground east of the point vertically below its release by*

$$d = \frac{1}{3}\,\omega g\, t_{\mathrm{f}}^{3} \cos\lambda = \frac{\omega \cos\lambda}{3}\sqrt{\frac{8h^{3}}{g}}\,, \qquad t_{\mathrm{f}} = \sqrt{\frac{2h}{g}}\,, \tag{20.22}$$

in either hemisphere, since $\cos\lambda > 0$ at every latitude but the poles, where the deflection vanishes.

Rests on Equations (19.14) and (20.7) and Postulate 19.19.

Proof. *Derives Proposition 20.13.*

Work in a frame that does not rotate with the Earth, with its origin at the Earth's centre, and take cylindrical coordinates $(\varrho, \varphi, \zeta)$ about the polar axis with φ increasing eastward. Treat the Earth's attraction as central, $\boldsymbol{F} = -F(r)\hat{\boldsymbol{r}}$. Then the torque about the centre vanishes identically, $\boldsymbol{x} \times \boldsymbol{F} = \boldsymbol{0}$, so by Equation (19.14) the angular momentum is conserved; in particular its polar component

$$L_{\zeta} = m\varrho^{2}\dot{\varphi} = \text{constant} \tag{20.23}$$

is constant throughout the fall.

At release the body shares the motion of the ground, so $\dot{\varphi}(0) = \omega$, and it stands at radius $r_0 = R + h$, that is at $\varrho_0 = r_0 \cos\lambda$ from the axis. To leading order in ω the fall is radial and obeys Equation (20.7), so $r(t) = r_0 - \tfrac{1}{2}gt^{2}$ at unchanged latitude and $\varrho(t) = r(t)\cos\lambda$. Then Equation (20.23) gives

$$\dot{\varphi}(t) = \omega\left(\frac{r_0}{r(t)}\right)^{2}, \qquad \text{so} \qquad \dot{\varphi} - \omega \simeq \omega\,\frac{2\,(r_0 - r)}{R} = \frac{\omega g t^{2}}{R}\,, \tag{20.24}$$

the body running ahead of the rotation because, as it approaches the axis, conservation of L_{ζ} speeds its azimuthal rate up.

The ground point that lay directly below at release keeps the azimuthal rate ω exactly. The body's lead in azimuth after the flight time t_{f} of Equation (20.7) is therefore

$$\Delta\varphi = \int_{0}^{t_{\mathrm{f}}} (\dot{\varphi} - \omega)\,\mathrm{d}t = \frac{\omega g\, t_{\mathrm{f}}^{3}}{3R}\,, \tag{20.25}$$

and the arc this subtends at the ground, at distance $R\cos\lambda$ from the axis, is $d = R\cos\lambda\,\Delta\varphi$, which is Equation (20.22). Substituting $t_{\mathrm{f}} = \sqrt{2h/g}$ gives the second form.

Three things were dropped and each is small, the numbers being quoted for a hundred-metre drop. Terms of relative order $\omega t_\mathrm{f} \approx 3 \times 10^{-4}$ come from the departure of the fall from radial; terms of relative order $h/R \approx 2 \times 10^{-5}$ come from replacing r_0 by R; and the plumb vertical is tilted from the radial direction by the centrifugal term of Equation (29.68), by about 2×10^{-3} radians, but that tilt lies in the meridian plane and so displaces the plumb point north or south, not east. □

Remark 20.14 (Why the elementary argument gives three halves of this). The argument that suggests itself first — at release the body's eastward speed exceeds the ground's by $\omega h \cos\lambda$, so in a time t_f it should gain $\omega h \cos\lambda\, t_\mathrm{f} = \frac{1}{2}\omega g t_\mathrm{f}^3 \cos\lambda$ — returns three halves of Equation (20.22). It fails on two counts, which is why the discrepancy is not obviously either way: the excess is not held constant, since Equation (20.24) has the azimuthal rate rising during the fall; and what must be compared with the ground is the *angle* swept, not a linear speed measured at a different distance from the axis. The rotating-frame route reaches Equation (20.22) in fewer lines, through the Coriolis acceleration $-2\boldsymbol{\omega} \times \boldsymbol{v}$ of Section 29.5, whose leading contribution $-2\boldsymbol{\omega} \times (-gt\hat{\boldsymbol{z}})$ has magnitude $2\omega g t \cos\lambda$ and points due east, and integrates twice from rest to the same $\frac{1}{3}\omega g t^3 \cos\lambda$ [100, 1585]. The two routes agree, as they must.

Rests on Proposition 20.13.

Derivation. *Derives Phenomenon 20.12.*

Proposition 20.13 is the statement, and the numbers follow from it. For a drop of $h = 100\,\mathrm{m}$ at $\lambda = 45°$, Equation (20.22) gives $t_\mathrm{f} = 4.52\,\mathrm{s}$ and $d = 1.55\,\mathrm{cm}$, which is the figure quoted in Phenomenon 20.12. Reich's measurement is the test: dropping balls down a mine shaft at Freiberg, at latitude about 50.9° and through some 158 m, he reported a mean eastward deviation of about 28 mm [92], against the 27.5 mm that Equation (20.22) gives for those parameters. The lateral deflection of a long-range projectile has the same origin, the Coriolis term acting on a horizontal velocity instead of a vertical one, and the whole effect is the free-fall counterpart of the Foucault pendulum [134] — a direct mechanical signature of assumption 4 failing, that is, of the laboratory frame not being inertial. □

Beyond these two, the remaining assumptions set their own limits. The uniform-field assumption 2 fails at the level of the free-air gradient measured in Section 20.4 — already relevant inside a 20 cm gravimeter drop at nine significant figures — and fails qualitatively for ranges comparable with the Earth's radius, where the trajectory closes into an ellipse about the Earth's centre: Newton's celebrated diagram of a cannonball fired from a mountaintop makes the parabola visibly a local approximation to an orbit [14]. Assumption 1 excludes the spin of real projectiles, whose coupling to the air (the Magnus effect) bends every driven golf ball and curled football; these lie beyond the scope of this chapter.

20.6 From Galileo's ratios to the equivalence principle

One step of the derivation above deserves to be read twice. In Equation (20.2) the mass m multiplies both sides — once as the *inertial* mass, the coefficient of resistance to acceleration in Newton's second law, and once as the *gravitational* mass, the coupling of the body to the field. Their cancellation is what made the trajectory independent of the body, and that independence is not a theorem but an experimental fact: Galileo's inclined planes could not distinguish the bronze ball from a heavier one [4], Newton checked the equality with pendulums of many materials [14], and its modern descendants — Eötvös-type torsion balances and satellite tests reviewed in Chapter 42 — constrain any composition-dependence of free fall at the level of parts in 10^{13} and beyond.

Einstein's insight of 1907, developed into the general theory of relativity [436], was to take this oldest of experimental facts as a founding principle: if all bodies fall alike, then

free fall is not a forced motion at all but the natural, unforced state, and gravitation is a property of spacetime geometry rather than of the falling body. The uniform field $\boldsymbol{g}$ of assumption 2 becomes, in that reading, the first-order description of a curved spacetime in a small laboratory. The bronze ball on the parchment-lined groove thus stands at the head of a straight line of inference that runs through this treatise to Part V; the reader will meet Phenomenon 20.1 again there, restated as the geodesic motion of a test body.

Chapter 21

Lagrangian Mechanics

The Newtonian formulation of Chapter 19 describes a system of particles through the forces acting on each of them, referred to the Cartesian coordinates of an inertial frame. When the motion is restricted by constraints this description carries two burdens: the constraint forces are not known in advance, and the Cartesian coordinates are no longer independent. The present chapter removes both. Constraints and generalized coordinates eliminate the redundant variables; the principle of virtual work and d'Alembert's principle eliminate the constraint forces; and the resulting Euler–Lagrange equations recast the dynamics in terms of a single scalar function of the generalized coordinates, the Lagrangian. The chapter closes with Hamilton's principle, which condenses the equations of motion into the stationarity of one functional, the action [1585, 2027].

Notation 21.1. Throughout this chapter and Chapter 22, the particles of an N-particle system are labelled $s = 1, \dots, N$, carried as a superscript; $i, j = 1, 2, 3$ label Cartesian components; $k = 1, \dots, n$ labels holonomic constraints; and $a, b = 1, \dots, f$ label generalized coordinates. Repeated Cartesian and generalized-coordinate indices are summed over; sums over particles and over constraints are written explicitly, except where the suppression of the particle label is expressly noted. The symbol ∇^s denotes the gradient with respect to the coordinates $\boldsymbol{x}^s$ of the s-th particle. Active forces are denoted $\boldsymbol{F}$, constraint forces $\boldsymbol{\Phi}$, and the total external force on a particle is their sum.

21.1 Generalized coordinates

21.1.1 Constraints

A *constraint* is a condition that relates the coordinates of a particle. Constraints encode, at the level of the description, everything that the environment forbids the system from doing: a bead cannot leave its wire, a ball cannot penetrate the floor. We classify them as follows.

Definition 21.2 (Holonomic constraint)**.** A *holonomic constraint* is a condition imposed on the coordinates of a reference frame, with respect to which the motion of a particle is described, expressible in the form

$$\phi(x^i, t) = C\,, \tag{21.1}$$

where ϕ is the function that describes the constraint and C is a constant.

Example 21.3 (Planar motion)**.** A particle whose motion with respect to some reference frame is two-dimensional is subject to the constraint $z = 0$. In this case $\phi(x^i, t) = z$ and $C = 0$.

Example 21.4 (Circular motion)**.** A particle describing a circular path of radius R is subject to the constraints given by the equations

$$z = 0\,, \qquad x^2 + y^2 = R^2\,. \tag{21.2}$$

Definition 21.5 (Non-holonomic constraint)**.** A *non-holonomic constraint* is a condition imposed on the coordinates of a reference frame, with respect to which the motion of a particle is described, that can be written as

$$\phi(x^i, t) \geq C\,. \tag{21.3}$$

Example 21.6 (Ball on a court)**.** A ball on a tennis court is confined to move only in the region

$$z \geq 0 \tag{21.4}$$

with respect to the ground.

Remark 21.7. More generally, a constraint is called *anholonomic* whenever it cannot be written in the form Equation (21.1); the inequality constraints of Definition 21.5 are one instance.

Remark 21.8 (Scleronomous and rheonomous constraints)*.* A constraint that does not depend explicitly on time is called *scleronomous*; one that does depend explicitly on time is called *rheonomous* [2027].

21.1.2 Degrees of freedom

Definition 21.9 (Degrees of freedom)**.** The number of *degrees of freedom* of a system of particles is the minimum number of independent, time-dependent variables required to describe the state of the system completely.

If we have N particles, $3N$ variables describe the system (three coordinates for each particle, since the motion takes place in $\mathbb{R}^3$). If the system is subject to n holonomic constraints, each removes one variable, and the number of degrees of freedom is

$$f = 3N - n\,. \tag{21.5}$$

Example 21.10. A particle describing circular motion is subject to the two constraints of Equation (21.2); the number of degrees of freedom of the system is therefore $f = 3 - 2 = 1$.

If instead we have a system of N bodies, these can also rotate. Three variables are needed to describe the rotation of each body and three for its translation, so that the total number of degrees of freedom is

$$f = 6N - n\,. \tag{21.6}$$

Example 21.11. A pair of scissors has seven degrees of freedom: three of translation, three of rotation, and one variable describing its opening.

21.1.3 Generalized coordinates

Definition 21.12 (Generalized coordinates)**.** Consider a system of N particles with f degrees of freedom. The f independent, time-dependent variables q^a, $a = 1, \ldots, f$, sufficient to describe the motion of the N particles, are called *generalized coordinates* of the system. There exists, accordingly, a general coordinate transformation of the form

$$\boldsymbol{x}^s = \boldsymbol{x}^s(q^a, t)\,, \qquad s = 1, \ldots, N\,. \tag{21.7}$$

Example 21.13. For circular motion there is a single degree of freedom, to which one may assign as generalized coordinate the angle formed by the line joining the particle to the centre with respect to a fixed line. Alternatively, one may assign as generalized coordinate the arc length traversed by the particle with respect to some fixed point. The choice of generalized coordinates is not unique.

21.1.4 Constraint forces and active forces

Definition 21.14 (Constraint force)**.** Consider a system of N particles with f degrees of freedom. A force acting on the s-th particle is called a *constraint force* if it is due to the action of a constraint; equivalently, it is the force $\boldsymbol{\Phi}$ that replaces the action of a constraint while producing the same effect. We denote by $\boldsymbol{\Phi}_k^s$ the constraint force on the s-th particle due to the k-th constraint.

Definition 21.15 (Active force)**.** A force is called *active* if it is not a constraint force; the active forces are the external forces on a particle, the constraint forces excepted. Denoting the individual active forces on the s-th particle by $\boldsymbol{F}_r^s$, $r = 1, \dots, R$, the total force on that particle is

$$\boldsymbol{F}'^s = \sum_{r=1}^{R} \boldsymbol{F}_r^s + \sum_{k=1}^{n} \boldsymbol{\Phi}_k^s \,. \tag{21.8}$$

By Newton's second law (Chapter 19), for a system of N particles subject to n holonomic constraints and to R active forces,

$$\sum_{r=1}^{R} \boldsymbol{F}_r^s + \sum_{k=1}^{n} \boldsymbol{\Phi}_k^s = m^s \ddot{\boldsymbol{x}}^s \,, \qquad s = 1, \dots, N \,. \tag{21.9}$$

We shall write $\boldsymbol{F}^s = \sum_r \boldsymbol{F}_r^s$ for the total active force and $\boldsymbol{\Phi}^s = \sum_k \boldsymbol{\Phi}_k^s$ for the total constraint force on the s-th particle.

21.2 Virtual work

21.2.1 Virtual displacements

Consider a system of N particles, the position of the s-th particle being $\boldsymbol{x}^s(t)$. The *physical displacement* performed by this particle when the system evolves during a time $\mathrm{d}t$ is the variation $\mathrm{d}\boldsymbol{x}^s$ of the position,

$$\boldsymbol{x}^s(t + \mathrm{d}t) = \boldsymbol{x}^s(t) + \mathrm{d}\boldsymbol{x}^s(t) \,. \tag{21.10}$$

That the system is evolving implies, however, that the constraints may be changing in time. We can construct the idea of another kind of displacement, one performed at fixed time, so that the constraints do not change. As stated this is impossible, since the system does not evolve at a fixed time; nevertheless, in order to adopt principles that are independent of time, we define the *virtual displacement* performed by the s-th particle as a displacement at fixed time, required at least to be compatible with the constraints (which, time being fixed, are not evolving). We denote the virtual displacement by $\delta\boldsymbol{x}^s$.

Definition 21.16 (Compatible and invertible virtual displacements)**.** A virtual displacement is *compatible with the constraints* if it respects them; if it fails to respect some constraint it is called *incompatible*. A compatible virtual displacement is *invertible* if the reversed displacement is also compatible with the constraints; there may exist compatible virtual displacements that are *not invertible*.

21.2.2 The constraint-force hypothesis

We state as a principle that

> *the constraint force on a particle is always perpendicular to the surface of the constraint.*

Thus the constraint force on the s-th particle due to the k-th holonomic constraint, with constraint function $\phi^k(x^i, t)$, is proportional to $\nabla^s \phi^k$; that is,

$$\boldsymbol{\Phi}^s_k = \lambda^k \nabla^s \phi^k \qquad \text{(no sum over } k\text{)}\,. \tag{21.11}$$

Summing all the constraint forces acting on each particle,

$$\boldsymbol{\Phi}^s = \sum_{k=1}^{n} \lambda^k \nabla^s \phi^k\,. \tag{21.12}$$

The quantities λ^k are properties of each surface (that is, of the k-th constraint) and may depend on the coordinates and on time,

$$\lambda^k = \lambda^k(\boldsymbol{x}^s, t)\,. \tag{21.13}$$

Virtual displacements compatible with the constraints

By the constraint-force hypothesis, for virtual displacements compatible with the holonomic constraints—displacements in which the particle moves on the surface of the constraint—the virtual displacement is tangent to that surface, and hence perpendicular to each constraint force. Thus

$$\boldsymbol{\Phi}^s_k \cdot \delta\boldsymbol{x}^s = \lambda^k \nabla^s \phi^k \cdot \delta\boldsymbol{x}^s = 0 \qquad \text{(no sum over } k\text{)}\,, \tag{21.14}$$

and, summing over the constraints,

$$\boldsymbol{\Phi}^s \cdot \delta\boldsymbol{x}^s = 0\,. \tag{21.15}$$

21.2.3 Virtual work and its principle

Definition 21.17 (Virtual work)**.** The *virtual work* performed on the s-th particle by a force $\boldsymbol{F}$ under the virtual displacement $\delta\boldsymbol{x}^s$ is

$$\delta W^s = \boldsymbol{F} \cdot \delta\boldsymbol{x}^s\,. \tag{21.16}$$

Theorem 21.18 (Principle of virtual work)**.** *Consider a system of N particles, subject on each particle to the total active force $\boldsymbol{F}^s$ and the total constraint force $\boldsymbol{\Phi}^s$, and suppose the particles are in equilibrium with respect to an inertial frame, so that the total force on each of them vanishes. Then, for virtual displacements compatible with the constraints, the virtual work of the active forces vanishes:*

$$\sum_{s=1}^{N} \boldsymbol{F}^s \cdot \delta\boldsymbol{x}^s = 0\,. \tag{21.17}$$

Derivation. Equilibrium means $\boldsymbol{F}^s + \boldsymbol{\Phi}^s = \boldsymbol{0}$ for each s. Taking the scalar product with $\delta\boldsymbol{x}^s$ and summing over the particles,

$$\sum_{s=1}^{N} \boldsymbol{F}^s \cdot \delta\boldsymbol{x}^s + \sum_{s=1}^{N} \boldsymbol{\Phi}^s \cdot \delta\boldsymbol{x}^s = 0\,.$$

The second term is the sum of the virtual works performed on the particles by the constraint forces. In general this constraint-force virtual work is different from zero; but if we consider only virtual displacements compatible with the constraints, it vanishes by Equation (21.15), and Equation (21.17) follows. □

21.3 d'Alembert's principle

The principle of virtual work is a statement about statics. d'Alembert's device extends it to dynamics: the accelerations are absorbed into fictitious forces, so that every problem of motion becomes a problem of equilibrium.

Definition 21.19 (Inertial force)**.** The *inertial force* on the s-th particle is defined as

$$\boldsymbol{\varphi}^s = -\dot{\boldsymbol{p}}^s = -m^s \ddot{\boldsymbol{x}}^s \,, \tag{21.18}$$

the force introduced to counteract the fictitious effects that appear when the motion is observed from a non-inertial reference frame.

Theorem 21.20 (d'Alembert's principle)**.** *For a system of N particles, each subject to the total active force $\boldsymbol{F}^s$ and the total constraint force $\boldsymbol{\Phi}^s$, and for virtual displacements compatible with the constraints,*

$$\sum_{s=1}^{N} (\boldsymbol{F}^s - \dot{\boldsymbol{p}}^s) \cdot \delta \boldsymbol{x}^s = 0 \,. \tag{21.19}$$

Derivation. By Newton's second law,

$$\boldsymbol{F}^s + \boldsymbol{\Phi}^s = \dot{\boldsymbol{p}}^s \,.$$

We may pass to a frame in which each particle is in equilibrium by treating the inertial force $\boldsymbol{\varphi}^s = -\dot{\boldsymbol{p}}^s$ as an active force:

$$\boldsymbol{F}^s + \boldsymbol{\Phi}^s + \boldsymbol{\varphi}^s = \boldsymbol{0} \,.$$

Proceeding as in the principle of virtual work, we take the scalar product with $\delta \boldsymbol{x}^s$ and sum over the particles:

$$\sum_{s=1}^{N} \boldsymbol{F}^s \cdot \delta \boldsymbol{x}^s + \sum_{s=1}^{N} \boldsymbol{\Phi}^s \cdot \delta \boldsymbol{x}^s + \sum_{s=1}^{N} \boldsymbol{\varphi}^s \cdot \delta \boldsymbol{x}^s = 0 \,.$$

Again, for virtual displacements compatible with the constraints the second sum vanishes by Equation (21.15), leaving

$$\sum_{s=1}^{N} \boldsymbol{F}^s \cdot \delta \boldsymbol{x}^s + \sum_{s=1}^{N} \boldsymbol{\varphi}^s \cdot \delta \boldsymbol{x}^s = 0 \,,$$

which, on substituting Equation (21.18), is Equation (21.19). □

d'Alembert's principle states that the virtual works performed by the impressed forces and by the inertial forces form a system in equilibrium. Denoting the total virtual work performed by the active forces and by the inertial forces as δW^A and δW^N respectively, it reads

$$\delta W^A + \delta W^N = 0 \,. \tag{21.20}$$

21.4 The Euler–Lagrange equations

21.4.1 Lagrange equations of the first kind

Theorem 21.21 (Lagrange equations of the first kind)**.** *Consider a system of N particles, subject on each particle to the total active force $\boldsymbol{F}^s$ and to the action of n holonomic constraints*

$$\phi^k(x^i, t) = C^k \,, \qquad k = 1, \dots, n \,. \tag{21.21}$$

Then the motion satisfies

$$\boldsymbol{F}^s + \sum_{k=1}^{n} \lambda^k \nabla^s \phi^k = \dot{\boldsymbol{p}}^s \,, \qquad s = 1, \dots, N \,, \tag{21.22}$$

which, together with the conditions Equation (21.21)*, are known as the* Lagrange equations of the first kind.

Derivation. The total virtual work δW^V performed by the constraint forces is, by the constraint-force hypothesis Equation (21.12),

$$\delta W^V = \sum_{s=1}^{N} \boldsymbol{\Phi}^s \cdot \delta \boldsymbol{x}^s = \sum_{s=1}^{N} \sum_{k=1}^{n} \lambda^k \nabla^s \phi^k \cdot \delta \boldsymbol{x}^s \,,$$

and for virtual displacements compatible with the constraints,

$$\sum_{s=1}^{N} \sum_{k=1}^{n} \lambda^k \nabla^s \phi^k \cdot \delta \boldsymbol{x}^s = 0 \,. \tag{21.23}$$

In this case d'Alembert's principle takes the form

$$\delta W^A + \delta W^N + \delta W^V = 0 \,,$$

so that, combining Equation (21.19) and Equation (21.23),

$$\sum_{s=1}^{N} \left(\boldsymbol{F}^s - \dot{\boldsymbol{p}}^s + \sum_{k=1}^{n} \lambda^k \nabla^s \phi^k \right) \cdot \delta \boldsymbol{x}^s = 0 \,.$$

Since the coordinates are linearly independent, the coefficient of each displacement must vanish, which yields Equation (21.22). □

21.4.2 Lagrange multipliers as constraint forces

The multipliers λ^k were introduced by the constraint-force hypothesis Equation (21.12) without being assumed known, and Equation (21.22) does not determine the motion until they are found. That they exist, and that a constrained stationarity problem is equivalent to an unconstrained one for the augmented integrand $\mathcal{L} + \sum_k \lambda^k \phi^k$ whenever the constraint gradients have full rank, is the multiplier rule of Theorem 16.41, proved in Section 16.3.1 and not reproved here [40]. What belongs to mechanics, and is the subject of this subsection, is what the multipliers physically *are*: each one is a constraint force, in newtons, and it is usually the number the experiment measures.

The multipliers are determined by the motion

Proposition 21.22 (Determinacy of the first-kind system)**.** *Let the n constraint functions ϕ^k of Equation* (21.21) *be of class C^2 and let the n vectors of $\mathbb{R}^{3N}$ with components $(m^s)^{-1/2} \nabla^s \phi^k$, $s = 1, \dots, N$, be linearly independent at each point of the motion. Then the multipliers are fixed algebraically by the instantaneous state $(\boldsymbol{x}^s, \dot{\boldsymbol{x}}^s, t)$: the Gram matrix*

$$M^{kl} = \sum_{s=1}^{N} \frac{1}{m^s} \nabla^s \phi^k \cdot \nabla^s \phi^l \tag{21.24}$$

is symmetric positive definite, and

$$\sum_{l=1}^{n} M^{kl} \lambda^l = -\,\Xi^k - \sum_{s=1}^{N} \frac{1}{m^s} \nabla^s \phi^k \cdot \boldsymbol{F}^s \,, \tag{21.25}$$

where Ξ^k collects the terms of $\mathrm{d}^2\phi^k/\mathrm{d}t^2$ that do not contain the accelerations. The first-kind system Equation (21.22)–Equation (21.21) is therefore determinate: $3N+n$ equations for the $3N$ coordinates and the n multipliers.

Rests on Equations (21.12), (21.21) and (21.22).

Derivation. *Derives Proposition 21.22.*

Differentiating $\phi^k(\boldsymbol{x}^s, t) = C^k$ twice along the motion and separating the acceleration terms,

$$\frac{\mathrm{d}^2\phi^k}{\mathrm{d}t^2} = \sum_{s=1}^{N} \nabla^s\phi^k \cdot \ddot{\boldsymbol{x}}^s + \Xi^k = 0\,, \tag{21.26}$$

with

$$\Xi^k = \sum_{s,s'} \dot{\boldsymbol{x}}^s \cdot \left(\nabla^s\nabla^{s'}\phi^k\right)\dot{\boldsymbol{x}}^{s'} + 2\sum_{s} \frac{\partial}{\partial t}\left(\nabla^s\phi^k\right)\cdot\dot{\boldsymbol{x}}^s + \frac{\partial^2\phi^k}{\partial t^2}\,,$$

a function of the state alone (the mixed second derivatives are symmetric by Proposition 7.73). Substituting the accelerations from Equation (21.22),

$$\ddot{\boldsymbol{x}}^s = \frac{1}{m^s}\left(\boldsymbol{F}^s + \sum_{l=1}^{n}\lambda^l\nabla^s\phi^l\right)\,,$$

into Equation (21.26) gives Equation (21.25) with M^{kl} as in Equation (21.24). That matrix is symmetric by inspection, and for any $u \in \mathbb{R}^n$

$$\sum_{k,l=1}^{n} u_k M^{kl} u_l = \sum_{s=1}^{N}\frac{1}{m^s}\left|\sum_{k=1}^{n} u_k\nabla^s\phi^k\right|^2 \geq 0\,,$$

with equality only if $\sum_k u_k\,(m^s)^{-1/2}\,\nabla^s\phi^k = \mathbf{0}$ for every s, that is only if $u = 0$ under the independence hypothesis. So M is invertible and Equation (21.25) has a unique solution λ^l. The count of equations and unknowns follows. □

Remark 21.23 (Failure of the rank condition is physical). The hypothesis is not a technicality. Two constraint surfaces whose gradients become parallel at a point are, at that point, *statically indeterminate*: the same total constraint force can be apportioned between them in infinitely many ways, and no equation of motion can decide the split. The elementary instance is a table with four legs on a rigid floor, where the four normal forces are not determined by rigid-body mechanics and the real answer comes from the elasticity of Chapter 30. Positive definiteness of Equation (21.24) is exactly the statement that this has not happened.

Remark 21.24 (SI dimensions of a multiplier). Since $\boldsymbol{\Phi}^s_k = \lambda^k\nabla^s\phi^k$ is a force, the multiplier carries whatever unit makes the product a newton: $[\lambda^k] = \mathrm{N\,m}/[\phi^k]$, equivalently $[\lambda^k\phi^k] = \mathrm{J}$, which is the mechanical face of the statement in Remark 16.42 that the augmented integrand must remain an energy. For the commonest case, a constraint written as a length — a rod of fixed length, a bead on a wire, a body resting on a surface — ϕ^k is in metres and λ^k is a force in newtons outright. Writing the same constraint as $\phi = r^2 = R^2$ instead of $\phi = r = R$ rescales the multiplier to N/m and halves it; the constraint force $\lambda\nabla\phi$ is of course unchanged. The multiplier is not an invariant of the constraint, only of the constraint *as written*.

A worked case: when does a body leave the surface?

The reason to keep the multipliers rather than eliminate the constraint is that the eliminated description cannot answer questions about the constraint force — and one such question is when the constraint stops acting at all.

Proposition 21.25 (Departure from a smooth sphere)**.** *A point mass m slides without friction on the outside of a fixed sphere of radius R in a uniform gravitational field of magnitude g, starting from rest at polar angle θ_0 measured from the upward vertical. While it remains in contact, the outward normal force is*

$$\lambda(\theta) = mg\,(3\cos\theta - 2\cos\theta_0)\ , \tag{21.27}$$

and the body leaves the surface at the angle θ_* given by

$$\cos\theta_* = \tfrac{2}{3}\cos\theta_0 \qquad \xrightarrow[\theta_0\to 0]{} \qquad \cos\theta_* = \tfrac{2}{3}\,, \quad \theta_* = 48.19^\circ\,, \tag{21.28}$$

independently of m, of R and of g.

Rests on Equations (21.12) and (21.22) and Proposition 21.22.

Derivation. Derives Proposition 21.25.

Take plane polar coordinates (r,θ) in the vertical plane of the motion, θ measured from the upward vertical, and impose the single constraint $\phi = r = R$, so that $\nabla\phi = \hat{\boldsymbol{r}}$ is the outward radial unit vector and λ is the outward normal force in newtons (Remark 21.24). The active force is $\boldsymbol{F} = -mg\hat{\boldsymbol{z}}$, whose radial and tangential components are $-mg\cos\theta$ and $+mg\sin\theta$. With the acceleration in polar coordinates, Equation (21.22) reads

$$m\left(\ddot r - r\dot\theta^2\right) = -mg\cos\theta + \lambda\,, \tag{21.29}$$

$$m\left(r\ddot\theta + 2\dot r\dot\theta\right) = mg\sin\theta\,. \tag{21.30}$$

On the constraint $r = R$, $\dot r = \ddot r = 0$, so Equation (21.29) gives $\lambda = m\left(g\cos\theta - R\dot\theta^2\right)$, which is Equation (21.25) in this one-constraint case. The constraint force is radial and the velocity tangential, so the constraint does no work and the energy of Equation (21.44) is conserved; measuring heights from the centre,

$$\tfrac{1}{2}mR^2\dot\theta^2 = mgR\,(\cos\theta_0 - \cos\theta) \qquad\Longrightarrow\qquad R\dot\theta^2 = 2g\,(\cos\theta_0 - \cos\theta)\ .$$

Substituting gives Equation (21.27). The sphere can push but not pull, so the description holds only while $\lambda \geq 0$; the first angle at which Equation (21.27) vanishes is Equation (21.28), and beyond it the body is a projectile with the free trajectory of Chapter 20. Everything cancels except the ratio $\frac{2}{3}$, so the prediction is a pure number. □

Example 21.26 (The same statement in SI)**.** Take the stipulated values $m = 0.100\,\mathrm{kg}$ and $R = 0.500\,\mathrm{m}$, with the standard gravitational acceleration $g_0 = 9.806\,65\,\mathrm{m/s^2}$ (Appendix B; a local g differs from g_0 by a few parts in 10^3, and since it enters Equation (21.27) linearly it shifts the force by that much and the angle not at all). Release from $\theta_0 \to 0$ then gives a normal force $\lambda(0) = mg_0 = 0.981\,\mathrm{N}$ at the top, departure at $\theta_* = 48.19^\circ$ after a fall of $R/3 = 0.167\,\mathrm{m}$, and a departure speed $v_* = \sqrt{2g_0R/3} = 1.81\,\mathrm{m/s}$, each quoted to the three significant figures of the inputs. Only the angle is falsifiable without reference to those inputs: it depends on nothing that was put in.

Rests on Equation (21.28).

Remark 21.27 (The constraint was an inequality all along). Proposition 21.25 is the promised instance of the distinctions drawn in Definition 21.5 and Definition 21.16. Resting on a surface is the inequality constraint $r \geq R$, not the equality $r = R$; the equality is merely the part of it that is active. At a point of the boundary $r = R$ an outward virtual displacement is *compatible* with $r \geq R$ while the inward one is not, so the compatible displacements there are not invertible, and the argument of Equation (21.14) — which required a displacement tangent to the surface — covers only the tangential ones. What survives is the inequality $\boldsymbol{\Phi}^s \cdot \delta\boldsymbol{x}^s \geq 0$: a one-sided constraint can do no negative virtual work. Its mechanical content

is the sign condition $\lambda \geq 0$, and Equation (21.28) is the instant at which that condition ceases to hold. A model that keeps the equality past θ_* predicts a surface that pulls the body back, which is not a small error but the wrong phenomenon.

Remark 21.28 (Two routes, two purposes). The alternative treatment eliminates the constraint at the outset, taking θ as the single generalized coordinate; the second-kind equations of Section 21.4.3 then reduce to Equation (21.30) with $r = R$, namely $R\ddot{\theta} = g\sin\theta$, and the normal force never appears. That is the right route when only the trajectory is wanted, and the wrong one whenever the quantity of interest is carried by the constraint: the tension in a suspension wire, the load on a bearing, the normal force that decides whether a rolling body slips, the reaction that a structure must be built to withstand. Those are the multipliers themselves, and the price of eliminating them is that they cannot be recovered afterwards from the reduced description.

21.4.3 Lagrange equations of the second kind

The Lagrange equations of the first kind are a manifestation of Newton's second law in the presence of n holonomic constraints, written in Cartesian coordinates. As we saw, each holonomic constraint on a system of N particles removes one degree of freedom, so that fewer variables are needed to describe the system. What we must now do is find a set of f equations that describe the system without reference to the constraints, that is, in terms of the generalized coordinates.

Definition 21.29 (Generalized force)**.** In analogy with the definition of virtual work, we define

$$Q_a = \sum_{s=1}^{N} \boldsymbol{F}^s \cdot \frac{\partial \boldsymbol{x}^s}{\partial q^a} \tag{21.31}$$

as the *total generalized force corresponding to the generalized coordinate* q^a.

Theorem 21.30 (Lagrange equations of the second kind)**.** *For a system of N particles with f degrees of freedom, subject to holonomic constraints and to active forces of total generalized force Q_a, the kinetic energy T satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T}{\partial \dot{q}^a}\right) - \frac{\partial T}{\partial q^a} = Q_a\,, \qquad a = 1, \dots, f\,, \tag{21.32}$$

which are known as the Lagrange equations of the second kind*.*

Derivation. We start from d'Alembert's principle,

$$\sum_{s=1}^{N} (\boldsymbol{F}^s - \dot{\boldsymbol{p}}^s) \cdot \delta \boldsymbol{x}^s = 0\,.$$

Suppressing the particle label (all expressions below carry an implicit sum over the suppressed index s, the mass accompanying its particle), this reads

$$\left(F^i - \dot{p}^i\right) \delta x^i = 0 \qquad \Longleftrightarrow \qquad F^i\, \delta x^i = m\ddot{x}^i\, \delta x^i\,.$$

Since the virtual displacement is performed at fixed time, the variation of Equation (21.7) contains no time term,

$$\delta x^i = \frac{\partial x^i}{\partial q^a}\, \delta q^a\,,$$

and therefore

$$F^i\, \frac{\partial x^i}{\partial q^a}\, \delta q^a = m\ddot{x}^i\, \frac{\partial x^i}{\partial q^a}\, \delta q^a\,. \tag{21.33}$$

We must analyse each side of Equation (21.33) separately.

The left-hand side is, restoring the particle sum, $Q_a\,\delta q^a$ by Equation (21.31).

For the right-hand side we compute, using the product rule,

$$\begin{aligned}
m\ddot{x}^i\,\frac{\partial x^i}{\partial q^a}\,\delta q^a &= m\frac{\mathrm{d}\dot{x}^i}{\mathrm{d}t}\,\frac{\partial x^i}{\partial q^a}\,\delta q^a \\
&= m\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{x}^i\,\frac{\partial x^i}{\partial q^a}\right) - \dot{x}^i\,\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial x^i}{\partial q^a}\right)\right]\delta q^a \\
&= m\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{x}^i\,\frac{\partial x^i}{\partial q^a}\right) - \dot{x}^i\left(\frac{\partial}{\partial q^b}\left(\frac{\partial x^i}{\partial q^a}\right)\frac{\mathrm{d}q^b}{\mathrm{d}t} + \frac{\partial}{\partial t}\left(\frac{\partial x^i}{\partial q^a}\right)\right)\right]\delta q^a \\
&= m\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{x}^i\,\frac{\partial x^i}{\partial q^a}\right) - \dot{x}^i\left(\frac{\partial^2 x^i}{\partial q^b\partial q^a}\dot{q}^b + \frac{\partial^2 x^i}{\partial t\,\partial q^a}\right)\right]\delta q^a \\
&= m\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{x}^i\,\frac{\partial x^i}{\partial q^a}\right) - \dot{x}^i\left(\frac{\partial^2 x^i}{\partial q^a\partial q^b}\dot{q}^b + \frac{\partial^2 x^i}{\partial q^a\partial t}\right)\right]\delta q^a \\
&= m\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{x}^i\,\frac{\partial x^i}{\partial q^a}\right) - \dot{x}^i\,\frac{\partial}{\partial q^a}\left(\frac{\partial x^i}{\partial q^b}\dot{q}^b + \frac{\partial x^i}{\partial t}\right)\right]\delta q^a \\
&= m\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{x}^i\,\frac{\partial x^i}{\partial q^a}\right) - \dot{x}^i\,\frac{\partial \dot{x}^i}{\partial q^a}\right]\delta q^a \\
&= m\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{x}^i\,\frac{\partial x^i}{\partial q^a}\right) - \frac{\partial}{\partial q^a}\left(\frac{1}{2}\dot{x}^i\dot{x}^i\right)\right]\delta q^a\,,
\end{aligned}$$

where the third line applies the chain rule to $\frac{\partial x^i}{\partial q^a}\left(q^b,t\right)$, the fifth uses the symmetry of mixed partial derivatives, and the sixth and seventh recognize the total time derivative of x^i,

$$\dot{x}^i = \frac{\partial x^i}{\partial q^b}\dot{q}^b + \frac{\partial x^i}{\partial t}\,. \tag{21.34}$$

Differentiating Equation (21.34) with respect to $\dot{q}^a$ gives the "cancellation of the dots",

$$\frac{\partial \dot{x}^i}{\partial \dot{q}^a} = \frac{\partial x^i}{\partial q^a}\,, \tag{21.35}$$

so that

$$\begin{aligned}
m\ddot{x}^i\,\frac{\partial x^i}{\partial q^a}\,\delta q^a &= m\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{x}^i\,\frac{\partial \dot{x}^i}{\partial \dot{q}^a}\right) - \frac{\partial}{\partial q^a}\left(\frac{1}{2}\dot{x}^i\dot{x}^i\right)\right]\delta q^a \\
&= \left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{q}^a}\left(\frac{1}{2}m\dot{x}^i\dot{x}^i\right)\right) - \frac{\partial}{\partial q^a}\left(\frac{1}{2}m\dot{x}^i\dot{x}^i\right)\right]\delta q^a\,.
\end{aligned}$$

Since the kinetic energy of the N particles is

$$T = \frac{1}{2}\sum_{s=1}^{N} m^s\,\dot{\boldsymbol{x}}^s\cdot\dot{\boldsymbol{x}}^s$$

(the sum over the suppressed index s being understood throughout), we obtain

$$m\ddot{x}^i\,\frac{\partial x^i}{\partial q^a}\,\delta q^a = \left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T}{\partial \dot{q}^a}\right) - \frac{\partial T}{\partial q^a}\right]\delta q^a\,. \tag{21.36}$$

Substituting Equation (21.31) and Equation (21.36) into Equation (21.33),

$$Q_a\,\delta q^a = \left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T}{\partial \dot{q}^a}\right) - \frac{\partial T}{\partial q^a}\right]\delta q^a\,,$$

and since the δq^a are arbitrary, Equation (21.32) follows. □

21.4.4 The Lagrangian

The total generalized force Q_a may arise from conservative active forces and also from non-conservative ones; in the most general case the potential of the conservative forces may depend on the velocities.

Definition 21.31 (Lagrangian)**.** The *Lagrangian* of a system with kinetic energy T and potential energy V is

$$\mathcal{L} = T - V\,. \tag{21.37}$$

Theorem 21.32 (Lagrange equations)**.** *For a system that feels R active forces, of which D are conservative and derive from the potential V, the equations of motion take the form*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^a}\right) - \frac{\partial \mathcal{L}}{\partial q^a} = \sum_{r=1}^{R-D} Q_a^{(r)}\,, \tag{21.38}$$

where the right-hand side sums the generalized forces of the $R-D$ non-conservative active forces. In particular, when all the active forces are conservative,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^a}\right) - \frac{\partial \mathcal{L}}{\partial q^a} = 0\,. \tag{21.39}$$

Rests on Equations (21.31)*,* (21.32) *and* (21.37)*.*

Derivation. *Derives Theorem 21.32.*

Split the active force on each particle into the resultant $\boldsymbol{F}_{\mathrm{c}}^s$ of the D conservative forces and the remainder $\boldsymbol{F}_{\mathrm{nc}}^s$. By Definition 7.103 the former derives from a potential $V\left(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N, t\right)$ in joules,

$$\boldsymbol{F}_{\mathrm{c}}^s = -\nabla^s V\,,$$

so that the corresponding part of the generalized force Equation (21.31) is, by the chain rule in several variables (Proposition 7.72) applied to $V\left(\boldsymbol{x}^s(q,t), t\right)$,

$$\sum_{s=1}^{N} \boldsymbol{F}_{\mathrm{c}}^s \cdot \frac{\partial \boldsymbol{x}^s}{\partial q^a} = -\sum_{s=1}^{N} \nabla^s V \cdot \frac{\partial \boldsymbol{x}^s}{\partial q^a} = -\frac{\partial V}{\partial q^a}\,. \tag{21.40}$$

Hence $Q_a = -\partial V/\partial q^a + \sum_{r=1}^{R-D} Q_a^{(r)}$, where the sum runs over the non-conservative forces, and Equation (21.32) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T}{\partial \dot{q}^a}\right) - \frac{\partial T}{\partial q^a} + \frac{\partial V}{\partial q^a} = \sum_{r=1}^{R-D} Q_a^{(r)}\,. \tag{21.41}$$

Because V is a function of the coordinates and time only, it contributes nothing to the velocity derivative, $\partial V/\partial \dot{q}^a = 0$, so with $\mathcal{L} = T - V$ of Equation (21.37)

$$\frac{\partial \mathcal{L}}{\partial \dot{q}^a} = \frac{\partial T}{\partial \dot{q}^a}\,, \qquad \frac{\partial \mathcal{L}}{\partial q^a} = \frac{\partial T}{\partial q^a} - \frac{\partial V}{\partial q^a}\,,$$

and Equation (21.41) is exactly Equation (21.38). When every active force is conservative the right-hand side is empty and Equation (21.39) follows. □

Remark 21.33 (Velocity-dependent potentials). The hypothesis $\partial V/\partial \dot{q}^a = 0$ can be weakened, and must be, because the Lorentz force on a charge is velocity-dependent and yet has a Lagrangian. Suppose the generalized force derives from a *generalized potential* $U\left(q, \dot{q}, t\right)$ in joules in the sense

$$Q_a = -\frac{\partial U}{\partial q^a} + \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial U}{\partial \dot{q}^a}\right)\,. \tag{21.42}$$

Substituting Equation (21.42) into Equation (21.32) and collecting terms gives Equation (21.39) with $\mathcal{L} = T - U$, by the same two lines as above; the extra total time derivative in Equation (21.42) is precisely what is needed to cancel the $\partial U/\partial \dot{q}^a$ that now survives in $\partial \mathcal{L}/\partial \dot{q}^a$. The electromagnetic case $U = q_{\mathrm{e}}\,(\Phi - \dot{\boldsymbol{x}} \cdot \boldsymbol{A})$, with q_{e} the charge in coulombs, Φ in volts and $\boldsymbol{A}$ in T m, belongs to Chapter 64; it is also the reason Theorem 21.46 below is stated for a potential $V\,(q, \dot{q}, t)$. Velocity-dependent forces that are *not* of the form Equation (21.42) — sliding friction is the standard example — have no Lagrangian and must be carried on the right-hand side of Equation (21.38) as they stand.

Definition 21.34 (Generalized momentum)**.** The *generalized momentum* corresponding to the generalized coordinate q^a is

$$p_a = \frac{\partial \mathcal{L}}{\partial \dot{q}^a}\,. \tag{21.43}$$

21.4.5 Cyclic coordinates

Definition 21.35 (Cyclic coordinate)**.** A generalized coordinate q^a is said to be *cyclic* if the Lagrangian does not depend on it.

Theorem 21.36 (Conservation of the conjugate momentum)**.** *If q^a is cyclic and every active force is conservative, then the generalized momentum p_a is a constant of the motion.*
Rests on Definition 21.35 and Equations (21.39) *and* (21.43)*.*

Derivation. *Derives Theorem 21.36.*
By Definition 21.35 the index a is one for which $\partial \mathcal{L}/\partial q^a = 0$, so the Lagrange equation Equation (21.39) for that index reduces to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^a}\right) = \frac{\mathrm{d}p_a}{\mathrm{d}t} = 0 \qquad (\text{no sum over } a)\,,$$

and a C^1 function whose derivative vanishes on an interval is constant there (Corollary 7.36). The same statement in the general variational setting is Proposition 16.28. □

Remark 21.37 (A cyclic coordinate is a symmetry)*.* The name is opaque and the content is not. That $\mathcal{L}$ does not depend on q^a says that the translation $q^a \mapsto q^a + \varepsilon$, with the other coordinates held fixed, leaves the Lagrangian — and therefore the action — invariant. Theorem 21.36 is thus the simplest instance of Noether's first theorem (Theorem 16.84), and the conserved quantity it produces is the Noether charge of that one-parameter group. Two cases carry names: invariance under a rigid translation of the whole system conserves the total linear momentum, and invariance under a rigid rotation conserves the total angular momentum (Corollary 16.88). The reading in terms of symmetry is the one that survives into field theory and quantum mechanics, where no coordinate is "absent" in any literal sense.

21.4.6 The energy of a Lagrangian system

The total energy of a system with kinetic energy T and potential energy V is

$$E = T + V\,. \tag{21.44}$$

The function that will turn out to represent it in the Lagrangian formalism is the *Jacobi integral*,

$$h\,(q, \dot{q}, t) := \dot{q}^a \frac{\partial \mathcal{L}}{\partial \dot{q}^a} - \mathcal{L} = \dot{q}^a p_a - \mathcal{L}\,, \tag{21.45}$$

also in joules — the Legendre transform of $\mathcal{L}$ with respect to the velocities, which is what will become the Hamiltonian in Chapter 22, and the mechanical instance of the Beltrami constant of Remark 16.30. That h equals E is a proposition with hypotheses, not a definition.

Proposition 21.38 (Energy in generalized coordinates)**.** *Suppose the coordinate transformation Equation* (21.7) *is scleronomous,* $\partial \boldsymbol{x}^s/\partial t = \mathbf{0}$ *(Remark 21.8), and the potential is velocity-independent,* $V = V(q,t)$*. Then*

$$E = \dot{q}^a \frac{\partial \mathcal{L}}{\partial \dot{q}^a} - \mathcal{L} = \dot{q}^a p_a - \mathcal{L} = h\,. \tag{21.46}$$

Rests on Equations (21.7)*,* (21.44) *and* (21.45)*.*

Derivation. Derives Proposition 21.38.

With $\partial \boldsymbol{x}^s/\partial t = \mathbf{0}$ the velocity relation Equation (21.34) loses its inhomogeneous term, $\dot{\boldsymbol{x}}^s = (\partial \boldsymbol{x}^s/\partial q^a)\,\dot{q}^a$, so the kinetic energy is the quadratic form

$$T = \tfrac{1}{2}\sum_{s=1}^{N} m^s\, \dot{\boldsymbol{x}}^s \cdot \dot{\boldsymbol{x}}^s = \tfrac{1}{2}\, g_{ab}(q)\, \dot{q}^a \dot{q}^b\,, \tag{21.47}$$

with g_{ab} the kinetic metric Equation (21.74), a function of the coordinates alone. Differentiating Equation (21.47) and using $g_{ab} = g_{ba}$,

$$\dot{q}^a \frac{\partial T}{\partial \dot{q}^a} = \dot{q}^a g_{ab} \dot{q}^b = 2T\,, \tag{21.48}$$

which is Euler's relation for a function homogeneous of degree two, obtained here by direct differentiation of the explicit form rather than from a general theorem (Remark 21.40). Since V carries no velocity, $\partial \mathcal{L}/\partial \dot{q}^a = \partial T/\partial \dot{q}^a$, and therefore

$$h = \dot{q}^a \frac{\partial T}{\partial \dot{q}^a} - (T - V) = 2T - T + V = T + V = E\,,$$

which is Equation (21.46). □

Remark 21.39 (The hypothesis actually needed). It is not enough that T carry no explicit t. A single particle in one dimension described from a frame translating at constant speed v, $x = q + vt$, has $T = \tfrac{1}{2}m\,(\dot{q} + v)^2$ — no explicit time dependence whatever — and yet $\dot{q}\,\partial T/\partial \dot{q} - T = \tfrac{1}{2}m\,(\dot{q}^2 - v^2) \neq T$. What fails is homogeneity, not time-independence, and it is the transformation Equation (21.7) rather than T that must be free of t. In the general rheonomous case Equation (21.34) gives $T = T_2 + T_1 + T_0$, homogeneous of degrees two, one and zero in the $\dot{q}^a$, and the same computation yields

$$h = T_2 - T_0 + V\,, \tag{21.49}$$

which is a perfectly good constant of the motion when $\partial \mathcal{L}/\partial t = 0$ but is *not* the energy $T + V$. The rotating frame of Chapter 29 is the case that matters in practice: there h is the Jacobi constant of the restricted three-body problem, T_0 is the centrifugal term, and confusing it with the energy is a standard way to get a wrong answer.

Remark 21.40 (A result Part II does not carry). The general statement behind Equation (21.48) is Euler's theorem on homogeneous functions: if f is differentiable and $f(\mu u) = \mu^k f(u)$ for all $\mu > 0$, then $u^i\, \partial f/\partial u^i = k f(u)$. It is a result of multivariable calculus and belongs in Chapter 7, which does not at present state it. Nothing above rests on it: the kinetic energy is an explicit quadratic form, so Equation (21.48) is one line of differentiation, and Equation (21.49) is obtained the same way term by term. The general theorem is nevertheless wanted elsewhere in this treatise — the virial theorem of Chapter 27 and the thermodynamic Euler relation of Part VI both use it in degrees other than two — and is recorded here as owed.

Proposition 21.41 (Time derivative of the Jacobi integral)**.** *Along any motion obeying Equation* (21.38)*,*

$$\dot{h} = \dot{q}^a Q_a - \frac{\partial \mathcal{L}}{\partial t}\,, \tag{21.50}$$

where $Q_a = \sum_r Q_a^{(r)}$ is the generalized force of the non-conservative active forces, as in Equation (21.38)*. Under the hypotheses of Proposition 21.38, $h = E$ and Equation* (21.50) *is the power balance for the total energy; if in addition the active forces are all conservative,*

$$\dot{E} = -\frac{\partial \mathcal{L}}{\partial t}\,. \tag{21.51}$$

Rests on Equations (21.38)*,* (21.45) *and* (21.46)*.*

Derivation. *Derives Proposition 21.41.*

Differentiate Equation (21.45) along the motion. The total time derivative of $\mathcal{L}(q, \dot{q}, t)$ is, by the chain rule (Proposition 7.72),

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}t} = \frac{\partial \mathcal{L}}{\partial q^a}\dot{q}^a + \frac{\partial \mathcal{L}}{\partial \dot{q}^a}\ddot{q}^a + \frac{\partial \mathcal{L}}{\partial t} = \frac{\partial \mathcal{L}}{\partial q^a}\dot{q}^a + p_a \ddot{q}^a + \frac{\partial \mathcal{L}}{\partial t}\,,$$

while $\mathrm{d}\left(\dot{q}^a p_a\right)/\mathrm{d}t = \ddot{q}^a p_a + \dot{q}^a \dot{p}_a$. The terms in $\ddot{q}^a$ cancel between the two, leaving

$$\dot{h} = \dot{q}^a \left(\dot{p}_a - \frac{\partial \mathcal{L}}{\partial q^a}\right) - \frac{\partial \mathcal{L}}{\partial t}\,,$$

and the bracket is the left-hand side of Equation (21.38), hence equal to Q_a. This is Equation (21.50). If every active force is conservative then $Q_a = 0$, and if the hypotheses of Proposition 21.38 hold as well then $h = E$, giving Equation (21.51). □

Remark 21.42 (Reading the two terms). Equation (21.50) is a power balance in watts, and its two terms are physically distinct. The first, $\dot{q}^a Q_a$, is the rate at which the non-conservative active forces do work on the system — the friction that drains a pendulum, the drive that sustains it. The second, $-\partial\mathcal{L}/\partial t$, is the rate at which work is done *by whatever imposes the explicit time dependence*: the hand that shortens the string of a pendulum, the motor that spins a constraint. Neither is a failure of energy conservation; both record energy crossing the boundary of the system as it was drawn. Note also that the constraint forces appear in neither term, which is the whole economy of the Lagrangian method — they were eliminated at Equation (21.15) precisely because they do no virtual work.

Theorem 21.43 (Conservation of energy)**.** *If the coordinate transformation Equation* (21.7) *is scleronomous, the potential is velocity-independent, every active force is conservative, and the Lagrangian carries no explicit time dependence, $\partial\mathcal{L}/\partial t = 0$, then the total energy $E = T + V$ is a constant of the motion. This is not the Newtonian statement Theorem 19.32 repeated in new notation. That one assumes a potential and reads the constancy of $T + V$ off $\boldsymbol{F} = m\boldsymbol{a}$ directly; this one holds in the presence of constraints, where T is a quadratic form in the generalized velocities with coordinate-dependent coefficients and the conserved quantity is identified from $\partial\mathcal{L}/\partial t$ rather than from a force balance. The logical route is not independent of Newton — the Lagrange equations themselves come from $\boldsymbol{F} = m\boldsymbol{a}$ by d'Alembert's principle, at Equation* (21.36) *— but the hypotheses and the argument are different ones, and the two statements agree, as they must, for an unconstrained particle in a time-independent potential.*

Rests on Equations (21.46) *and* (21.51) *and Theorem 19.32.*

Derivation. *Derives Theorem 21.43.*

The first three hypotheses put Proposition 21.41 in the form Equation (21.51), $\dot{E} = -\partial\mathcal{L}/\partial t$; the fourth makes the right-hand side vanish, and a C^1 function with vanishing derivative is constant (Corollary 7.36). □

Remark 21.44 (Which symmetry this is). $\partial\mathcal{L}/\partial t = 0$ says that the action is invariant under a shift of the origin of time, so Theorem 21.43 is Noether's theorem for time translation (Corollary 16.87), exactly as Theorem 21.36 is Noether's theorem for a translation in a coordinate. The two extra hypotheses — scleronomous constraints and a velocity-independent potential — are not needed for a conserved *quantity*; they are needed only for that quantity to be the energy, per Remark 21.39. It is worth being explicit about what is and is not assumed: nothing here derives conservation of energy from a deeper principle, since time-translation invariance is itself an empirical property of the interactions, tested wherever energy accounting is tested. What the theorem supplies is the exact link between the two.

21.5 Hamilton's principle

The differential principles of this chapter—virtual work and d'Alembert—characterize the motion instant by instant. Hamilton's principle characterizes it globally: among all conceivable histories of the system between two fixed instants, the physical one is singled out by a variational property of a single integral.

21.5.1 The action

Definition 21.45 (Action)**.** The *action* of a system with Lagrangian $\mathcal{L}$ along a history $q^a(t)$ between the instants t_a and t_b is the functional

$$S[q] = \int_{t_a}^{t_b} \mathcal{L}(q, \dot{q}, t)\, \mathrm{d}t\,. \tag{21.52}$$

Since $\mathcal{L}$ is an energy in joules and the integration is over a time, the action carries the SI dimension J s, the dimension of angular momentum and of the Planck constant — a coincidence that stops being one in Chapter 106. The functional is the mechanical instance of Definition 16.96.

Rests on Equation (21.37) *and Definition 16.96.*

21.5.2 Hamilton's principle

Theorem 21.46 (Hamilton's principle)**.** *For a physical system in which the active forces derive from a potential $V = V(q, \dot{q}, t)$ and the constraint forces are produced by holonomic constraints, the line integral*

$$S[q] = \int_{t_a}^{t_b} \mathcal{L}(q, \dot{q}, t)\, \mathrm{d}t$$

has a stationary value on the trajectory of motion of the physical system; that is,

$$\delta S = \delta \int_{t_a}^{t_b} \mathcal{L}(q, \dot{q}, t)\, \mathrm{d}t = 0\,. \tag{21.53}$$

This principle is a necessary and sufficient condition for the Lagrange equations with active forces derived from $V(q, \dot{q}, t)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^a}\right) - \frac{\partial \mathcal{L}}{\partial q^a} = 0\,. \tag{21.54}$$

Rests on Equations (21.38) *and* (21.52) *and Definition 21.12.*

Derivation. *Derives Theorem 21.46.*

The mathematics is Theorem 16.35, the Euler–Lagrange system for a functional of several dependent variables, read with the independent variable x taken to be the time t and the dependent variables y^i taken to be the f generalized coordinates q^a; the first variation is computed in Proposition 16.17, the integration by parts and the passage from $\int E_a \eta^a \mathrm{d}t = 0$

to $E_a = 0$ in Theorem 16.22 and Lemma 16.18. Applied to Equation (21.52) that theorem states exactly that S is stationary among C^1 histories with $q^a(t_a)$ and $q^a(t_b)$ prescribed if and only if Equation (21.54) holds, which is the assertion.

What is physics, and is the reason the hypotheses are stated as they are, is that the theorem may be applied at all. It requires the variations δq^a to be *independent* and otherwise arbitrary, and that is precisely what the two hypotheses secure. Because the constraints are holonomic, Equation (21.5) leaves f independent generalized coordinates and every nearby history obeying the constraints is reached by an unrestricted variation of them — so the comparison class is a full neighbourhood in the q^a, with no condition surviving from the constraints. Because the active forces derive from V, the whole content of the dynamics is inside $\mathcal{L} = T - V$, with no leftover generalized force on the right-hand side of Equation (21.38) to be accounted for. Drop either hypothesis and the equivalence fails: a non-conservative force leaves $\sum_r Q_a^{(r)} \neq 0$, which is not the Euler–Lagrange equation of any functional, and a non-integrable constraint restricts the admissible variations in a way that the unconstrained theorem does not model — the subject of Section 21.6. □

Remark 21.47 (Stationary, not least). The traditional name "principle of least action" overstates the result, and Equation (21.53) says only what is true: the first variation vanishes. For a mechanical Lagrangian the action is indeed a minimum over sufficiently short intervals, but past the first conjugate point — the kinetic focus — it is a saddle, never a maximum; for a harmonic oscillator of angular frequency ω that happens after a time π/ω, half a period. The second-variation machinery that decides this is Section 16.4, and the statement for mechanics specifically is Remark 16.98. The physical reason the *stationary* path is the one that matters is not classical at all: in the path-integral formulation of Chapter 106 every history contributes a phase $\exp(\mathrm{i}S/\hbar)$, and neighbouring histories interfere destructively except where S is stationary.

Remark 21.48. If the potential depends on the coordinates alone, $V = V(q)$, the system is conservative: the hypotheses of Proposition 21.38 and, given scleronomous constraints, of Theorem 21.43 are met, so $h = E$ and E is constant.

Remark 21.49 (Two principles, one content). It is worth recording what has and has not been shown. d'Alembert's principle Equation (21.19) and Hamilton's principle Equation (21.53) both yield Equation (21.54) for a holonomically constrained system with forces derived from a potential, so on that class they are equivalent and the choice between them is one of convenience. They are not equivalent in general. d'Alembert's principle is the more robust of the two: it survives non-conservative forces, which merely appear as Q_a on the right-hand side, and it survives non-integrable constraints, which restrict δq^a without restricting the histories. Hamilton's principle needs a Lagrangian to exist, and needs the admissible histories to form a neighbourhood. The historical order is the reverse of the logical one: Hamilton's essays [96] came ninety years after d'Alembert and half a century after [40], and the variational principle is the later and narrower statement, not the foundation from which the rest descends [1585, 2027].

21.6 Nonholonomic constraints

Every result so far has assumed holonomic constraints. A rolling wheel does not satisfy one. The constraint that a wheel does not slip is a condition on velocities that cannot be integrated to a condition on configurations, and the whole chain of the chapter — eliminate the constraints, count the degrees of freedom, vary freely — breaks at the first link. This section states what survives, and settles a question that is often presented as a matter of taste and is not: given a non-integrable constraint, two inequivalent sets of equations can be written down, and only one of them describes a rolling body.

21.6.1 Pfaffian constraints

Definition 21.50 (Pfaffian constraint)**.** A *Pfaffian* (or *linear kinematic*) constraint on a system with generalized coordinates q^a is a condition of the form

$$A^{\alpha}{}_{a}(q,t)\,\dot{q}^a + A^{\alpha}{}_{0}(q,t) = 0\,, \qquad \alpha = 1,\dots,m\,, \tag{21.55}$$

with the $m \times f$ matrix $A^{\alpha}{}_{a}$ of rank m; throughout this section m counts the Pfaffian constraints and never denotes a mass, the masses appearing only inside the kinetic metric Equation (21.74). It is *integrable*, and the system is then holonomic after all, if the one-forms $A^{\alpha}{}_{a}\mathrm{d}q^a + A^{\alpha}{}_{0}\mathrm{d}t$ admit an invertible matrix of integrating factors turning them into exact differentials $\mathrm{d}g^{\alpha}$; the criterion is the Frobenius condition recorded in Definition 16.47. A non-integrable Equation (21.55) is called *nonholonomic*.

Rests on Definitions 16.47 and 21.2.

Remark 21.51 (What a nonholonomic constraint does and does not remove)*.* A holonomic constraint removes one configuration variable and, with it, one velocity: the counting Equation (21.5) is exact and the system genuinely lives on a lower-dimensional space. A nonholonomic constraint removes a velocity and no configuration at all. The vertical rolling disc of Example 16.48 has four configuration variables and two constraints Equation (16.44), leaving two velocity degrees of freedom — and yet every one of the four configurations is reachable, by rolling the disc round a suitable closed loop. There is therefore no reduced coordinate set to pass to, and Definition 21.9 must be read with care: the number of independent velocities and the dimension of the accessible set are different numbers, which for holonomic systems coincide. This is the mechanical content of non-integrability, and it is why the parking of a car is possible at all.

21.6.2 The d'Alembert–Chetaev prescription

The principle of virtual work is stated for virtual displacements *compatible with the constraints*, and for a Pfaffian constraint that phrase must be given a meaning, since Equation (21.55) constrains velocities and δq^a is taken at fixed time. The answer that mechanics uses is the natural one: a virtual displacement is admissible when it lies in the annihilator of the constraint forms.

Definition 21.52 (Chetaev condition)**.** A virtual displacement is *compatible* with the Pfaffian constraints Equation (21.55) if

$$A^{\alpha}{}_{a}(q,t)\,\delta q^a = 0\,, \qquad \alpha = 1,\dots,m\,. \tag{21.56}$$

The term $A^{\alpha}{}_{0}$ is absent because the displacement is taken at fixed time, exactly as in Equation (21.14).

Rests on Equation (21.55) and Definition 21.16.

Theorem 21.53 (Equations of motion of a nonholonomic system)**.** *Let $q^1,\dots,q^f$ be coordinates on the configuration space that remains after any holonomic constraints have been eliminated, so that they are unrestricted as* configurations*. For a system with Lagrangian $\mathcal{L}$ in those coordinates, subject to the Pfaffian constraints Equation (21.55) whose constraint forces do no virtual work on displacements obeying Equation (21.56), and to non-conservative generalized forces Q_a, the motion satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^a}\right) - \frac{\partial \mathcal{L}}{\partial q^a} = Q_a + \sum_{\alpha=1}^{m} \mu_{\alpha} A^{\alpha}{}_{a}\,, \qquad a = 1,\dots,f\,, \tag{21.57}$$

together with the m constraints Equation (21.55). The $f+m$ equations determine the f coordinates and the m multipliers μ_{α}, and $\sum_{\alpha}\mu_{\alpha}A^{\alpha}{}_{a}$ is the generalized constraint force.

Rests on Equations (21.19), (21.38) and (21.56).

Derivation. *Derives Theorem 21.53.*

d'Alembert's principle Equation (21.19), rewritten in generalized coordinates by the computation of Theorem 21.30 and then split as in Theorem 21.32, states that

$$\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^a}\right) - \frac{\partial \mathcal{L}}{\partial q^a} - Q_a\right] \delta q^a = 0 \tag{21.58}$$

for every δq^a obeying Equation (21.56). Unlike the holonomic case, the δq^a are now *not* arbitrary, so the bracket need not vanish; what Equation (21.58) says is that the covector in brackets annihilates the subspace $\{\delta q \mid A^\alpha{}_a \delta q^a = 0\}$ of $\mathbb{R}^f$ — the kernel of the linear map A, of dimension $f - m$ by Theorem 5.38 and the rank hypothesis. The covectors annihilating that kernel form a subspace of the dual (Definition 5.56) of dimension m, and it contains the m independent rows $A^\alpha{}_a$, so it is exactly their span. Hence there are numbers $\mu_\alpha(t)$ with

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^a}\right) - \frac{\partial \mathcal{L}}{\partial q^a} - Q_a = \sum_{\alpha=1}^{m} \mu_\alpha A^\alpha{}_a \,,$$

which is Equation (21.57). They are determined algebraically: differentiating Equation (21.55) once in time gives m linear relations among the $\ddot{q}^a$, and substituting the accelerations from Equation (21.57) yields an $m \times m$ linear system for the μ_α. For a Lagrangian of the mechanical type $\mathcal{L} = T - V$ with $T = \frac{1}{2} g_{ab} \dot{q}^a \dot{q}^b$, that system has matrix $A^\alpha{}_a g^{ab} A^\beta{}_b$, with g^{ab} inverse to the kinetic metric Equation (21.74) and hence positive definite; the same one-line argument as in Proposition 21.22 makes the matrix positive definite under the rank hypothesis, so the μ_α are unique. □

Remark 21.54 (The initial-value problem). The last step is the one to hold on to. Given a configuration q^a and a velocity $\dot{q}^a$ that satisfies Equation (21.55), the multipliers μ_α follow at once, and Equation (21.57) then fixes the accelerations. The motion of a nonholonomic system is therefore determined by its mechanical state alone, exactly as a Newtonian system is. This is not a formality; it is what the next subsection turns on.

21.6.3 Why the vakonomic prescription is not the observed one

There is a second, formally impeccable way to write equations for a constrained system, and for a nonholonomic constraint it gives different answers.

Definition 21.55 (Vakonomic prescription)**.** The *vakonomic* equations of a system with Lagrangian $\mathcal{L}$ and constraints Equation (21.55) are the Euler–Lagrange equations obtained by making $S[q]$ stationary within the class of histories that *satisfy* Equation (21.55), that is by applying the multiplier rule to the augmented integrand $\mathcal{L} + \sum_\alpha \lambda_\alpha \left(A^\alpha{}_a \dot{q}^a + A^\alpha{}_0\right)$ and varying q^a and λ_α freely.

Rests on Equations (21.53) and (21.55).

Proposition 21.56 (The two prescriptions and their difference)**.** *The vakonomic equations of Definition 21.55 read*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^a}\right) - \frac{\partial \mathcal{L}}{\partial q^a} = -\sum_{\alpha=1}^{m} \dot{\lambda}_\alpha A^\alpha{}_a + \sum_{\alpha=1}^{m} \lambda_\alpha \left(C^\alpha{}_{ab}\, \dot{q}^b + c^\alpha{}_a\right) , \tag{21.59}$$

with the obstruction coefficients

$$C^\alpha{}_{ab} = \frac{\partial A^\alpha{}_b}{\partial q^a} - \frac{\partial A^\alpha{}_a}{\partial q^b} \,, \qquad c^\alpha{}_a = \frac{\partial A^\alpha{}_0}{\partial q^a} - \frac{\partial A^\alpha{}_a}{\partial t} \,. \tag{21.60}$$

If the constraint one-forms are exact, $A^\alpha{}_a = \partial g^\alpha / \partial q^a$ and $A^\alpha{}_0 = \partial g^\alpha / \partial t$, then $C^\alpha{}_{ab} = c^\alpha{}_a = 0$ identically and Equation (21.59) coincides with the d'Alembert–Chetaev equations

Equation (21.57) *with* $Q_a = 0$*, under the identification* $\mu_\alpha = -\dot\lambda_\alpha$*. Otherwise the two differ by the obstruction terms, which are exactly the components of the exterior derivative of the constraint forms — the quantity whose vanishing is the Frobenius integrability condition of Definition 21.50.*

Rests on Definition 21.55, Equation (21.57)*, and Theorem 16.41.*

Derivation. Derives Proposition 21.56.

Write $F = \mathcal{L} + \sum_\alpha \lambda_\alpha G^\alpha$ with $G^\alpha = A^\alpha{}_a \dot q^a + A^\alpha{}_0$. Varying λ_α returns Equation (21.55). For the q^a,

$$\frac{\partial F}{\partial \dot q^a} = \frac{\partial \mathcal{L}}{\partial \dot q^a} + \sum_\alpha \lambda_\alpha A^\alpha{}_a\,, \qquad \frac{\partial F}{\partial q^a} = \frac{\partial \mathcal{L}}{\partial q^a} + \sum_\alpha \lambda_\alpha \left(\frac{\partial A^\alpha{}_b}{\partial q^a}\dot q^b + \frac{\partial A^\alpha{}_0}{\partial q^a} \right),$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial F}{\partial \dot q^a}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot q^a}\right) + \sum_\alpha \dot\lambda_\alpha A^\alpha{}_a + \sum_\alpha \lambda_\alpha \left(\frac{\partial A^\alpha{}_a}{\partial q^b}\dot q^b + \frac{\partial A^\alpha{}_a}{\partial t} \right).$$

Subtracting and rearranging gives Equation (21.59) with Equation (21.60). If $A^\alpha{}_a = \partial_a g^\alpha$ and $A^\alpha{}_0 = \partial_t g^\alpha$ then every entry of Equation (21.60) is a difference of mixed second derivatives of g^α and vanishes by Proposition 7.73, leaving only the $\dot\lambda_\alpha$ term, which is Equation (21.57) with $\mu_\alpha = -\dot\lambda_\alpha$ and $Q_a = 0$. □

Example 21.57 (The difference is not small for a rolling disc)**.** For the vertical rolling disc of Example 16.48, with coordinates $(x, y, \varphi, \vartheta)$ and the constraints Equation (16.44) written as Equation (21.55),

$$A^1{}_x = 1\,, \quad A^1{}_\vartheta = -R\cos\varphi\,, \qquad A^2{}_y = 1\,, \quad A^2{}_\vartheta = -R\sin\varphi\,, \tag{21.61}$$

all other components and all $A^\alpha{}_0$ being zero. The only nonvanishing obstruction coefficients Equation (21.60) are

$$C^1{}_{\varphi\vartheta} = -C^1{}_{\vartheta\varphi} = R\sin\varphi\,, \qquad C^2{}_{\varphi\vartheta} = -C^2{}_{\vartheta\varphi} = -R\cos\varphi\,, \tag{21.62}$$

which cannot all vanish, since $\sin\varphi$ and $\cos\varphi$ have no common zero. The discrepancy between Equation (21.59) and Equation (21.57) in the φ equation is therefore $R\dot\vartheta\,(\lambda_1\sin\varphi - \lambda_2\cos\varphi)$, a term of the same order as the ones beside it whenever the disc is actually rolling ($\dot\vartheta \neq 0$) and the constraint is actually doing something ($\lambda \neq 0$). The two prescriptions are not two approximations to one answer; they are two different answers.

Rests on Equations (16.44) *and* (21.60)*.*

Remark 21.58 (Which one is observed, and how one can tell)*.* The equations of a rolling body are the d'Alembert–Chetaev ones Equation (21.57). Three things settle it, in increasing order of force.

The physical realization of the constraint. Rolling without slipping is enforced by static friction at the contact point: a real force, exerted instant by instant, whose magnitude adjusts to whatever the state requires. It does no work, because the material point of the body at the contact is instantaneously at rest, which is precisely the Chetaev condition Equation (21.56). Nothing in the physical situation restricts which histories the body is permitted to consider, which is what Definition 21.55 assumes.

Determinism. By Remark 21.54 the d'Alembert–Chetaev system fixes the motion from $(q, \dot q)$ alone. The vakonomic system does not: in Equation (21.59) the multipliers obey their own differential equations, so their initial values $\lambda_\alpha(t_0)$ are extra data, and one mechanical initial state gives rise to an m-parameter family of vakonomic motions. A disc released twice from the same place with the same velocity does the same thing twice; that observation alone excludes the vakonomic equations as a description of it, and it can be made in a lecture room.

The two answer different questions. The vakonomic equations are not wrong — they are the correct equations of the optimal-control problem in which the constraint is imposed on the admissible curves by decree and one asks which admissible curve extremizes the action. That is a real and useful problem, and the extra initial data $\lambda_\alpha(t_0)$ are what select a particular optimal trajectory in it. It is simply not the problem a rolling disc solves. The mathematical statement of the dichotomy, and the reason Theorem 16.41 does not cover the nonholonomic case, is Remark 16.49.

Two honest caveats. First, the standard mechanics texts [1585, 2027] present Equation (21.57) as the equations of a nonholonomic system, and the derivation above is theirs in substance; but this treatise's bibliography holds no reference reporting a controlled *experiment* that discriminates the two prescriptions on a rolling body, and the claim of observational adequacy above therefore rests on the accumulated agreement of nonholonomic rigid-body dynamics with measurement — the rolling and spinning experiments of Chapter 29 — rather than on a single cited test. Second, systems whose constraints are not imposed from outside but follow from a degenerate Lagrangian are a third case entirely, governed by the Dirac–Bergmann algorithm of Chapter 26, and nothing here applies to them.

21.7 General coordinate transformations

Generalized coordinates were introduced in Section 21.1.3 as *any* independent set of labels sufficient to locate the system in its configuration space: lengths, angles, or arbitrary mixtures of them, chosen for convenience and carrying no privileged status. That freedom has a consequence which is easy to state and hard to overstate: the equations of motion must not care which labels were chosen. This section proves that they do not, identifies the geometric objects that appear when one changes labels, and—in Section 21.8—shows that the same structure recurs verbatim in special and general relativity. The machinery is the differential geometry of Chapter 13, instantiated here on configuration space.

Remark 21.59 (Indices in this section). The conventions of Notation 21.1 continue to apply: $a, b, \dots$ label the f generalized coordinates. They are *not* the frame indices of Section 13.1, which label directions in spacetime; where spacetime enters, in Section 21.8.1, it carries $\mu, \nu, \dots$ as usual.

21.7.1 Point transformations

Definition 21.60 (Point transformation)**.** A *point transformation* is a change of generalized coordinates

$$Q^a = Q^a\big(q^1, \dots, q^f, t\big) \,, \qquad a = 1, \dots, f \,, \tag{21.63}$$

that is smooth and invertible, i.e. whose Jacobian $J^a{}_b = \frac{\partial Q^a}{\partial q^b}$ is everywhere nonsingular, $\det J \neq 0$. When Equation (21.63) carries an explicit t the transformation is *time-dependent*; a rotating or accelerating frame is of this kind.

Differentiating Equation (21.63) along a trajectory gives the induced transformation of the generalized velocities,

$$\dot{Q}^a = \frac{\partial Q^a}{\partial q^b}\,\dot{q}^b + \frac{\partial Q^a}{\partial t} \,, \tag{21.64}$$

which is affine—not linear—in the velocities whenever the transformation depends on time. Two identities follow from Equation (21.64) and are used repeatedly; the first is the "cancellation of the dots" already met at Equation (21.35), in the special case $q \to x$.

Lemma 21.61 (Cancellation of the dots; commutation of the derivatives)**.** *For any point transformation,*

$$\frac{\partial \dot{Q}^a}{\partial \dot{q}^b} = \frac{\partial Q^a}{\partial q^b}\,, \tag{21.65}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial Q^a}{\partial q^b}\right) = \frac{\partial \dot{Q}^a}{\partial q^b}\,. \tag{21.66}$$

Derivation. The right-hand side of Equation (21.64) contains $\dot{q}^b$ only in the first term, and there linearly with coefficient $\frac{\partial Q^a}{\partial q^b}$, which is a function of q and t alone; differentiating with respect to $\dot{q}^b$ therefore gives Equation (21.65). For Equation (21.66), expand the total time derivative of $\frac{\partial Q^a}{\partial q^b}(q,t)$ by the chain rule and compare with the q^b-derivative of Equation (21.64):

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial Q^a}{\partial q^b} = \frac{\partial^2 Q^a}{\partial q^c \partial q^b}\dot{q}^c + \frac{\partial^2 Q^a}{\partial t\,\partial q^b} = \frac{\partial}{\partial q^b}\left(\frac{\partial Q^a}{\partial q^c}\dot{q}^c + \frac{\partial Q^a}{\partial t}\right) = \frac{\partial \dot{Q}^a}{\partial q^b}\,,$$

the middle step using the symmetry of mixed partial derivatives (Proposition 7.73) and the independence of q^c and $\dot{q}^c$ as coordinates on velocity phase space, so that $\dot{q}^c$ is held fixed by $\partial/\partial q^b$. □

21.7.2 Covariance of the Euler–Lagrange equations

Definition 21.62 (Euler–Lagrange operator)**.** For a Lagrangian $\mathcal{L}$ the *Euler–Lagrange operator* is

$$\mathcal{E}_a[\mathcal{L}] \equiv \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^a}\right) - \frac{\partial \mathcal{L}}{\partial q^a}\,, \tag{21.67}$$

so that the equations of motion Equation (21.39) read $\mathcal{E}_a[\mathcal{L}] = 0$. When a second chart Q^a is in play, $\widetilde{\mathcal{E}}_a[\widetilde{\mathcal{L}}]$ denotes the same operator formed with $Q^a, \dot{Q}^a$ in place of $q^a, \dot{q}^a$.

Theorem 21.63 (Covariance of the Euler–Lagrange equations)**.** *Let $\mathcal{L}$ be a Lagrangian and let $\widetilde{\mathcal{L}}(Q,\dot{Q},t) = \mathcal{L}\Big(q(Q,t), \dot{q}(Q,\dot{Q},t), t\Big)$ be its expression in new coordinates related to the old by a point transformation Equation (21.63). Then the Euler–Lagrange operator transforms as a covector,*

$$\mathcal{E}_b[\mathcal{L}] = \frac{\partial Q^a}{\partial q^b}\,\widetilde{\mathcal{E}}_a[\widetilde{\mathcal{L}}]\,. \tag{21.68}$$

Consequently $\mathcal{E}_b[\mathcal{L}] = 0$ if and only if $\widetilde{\mathcal{E}}_a[\widetilde{\mathcal{L}}] = 0$: the Euler–Lagrange equations hold in one system of generalized coordinates precisely when they hold in every other.

Derivation. Since the transformation is invertible, the defining relation may be read as $\mathcal{L}(q,\dot{q},t) = \widetilde{\mathcal{L}}\Big(Q(q,t), \dot{Q}(q,\dot{q},t), t\Big)$, and it is this composition that is differentiated below. As $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ are the same function of the physical state in different labels, the chain rule gives

$$\frac{\partial \mathcal{L}}{\partial \dot{q}^b} = \frac{\partial \widetilde{\mathcal{L}}}{\partial \dot{Q}^a}\frac{\partial \dot{Q}^a}{\partial \dot{q}^b} = \frac{\partial \widetilde{\mathcal{L}}}{\partial \dot{Q}^a}\frac{\partial Q^a}{\partial q^b}\,, \tag{21.69}$$

using Equation (21.65); note that $\frac{\partial Q^a}{\partial \dot{q}^b} = 0$, because Q depends on q and t only. Differentiating Equation (21.69) totally in time,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^b}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \widetilde{\mathcal{L}}}{\partial \dot{Q}^a}\right)\frac{\partial Q^a}{\partial q^b} + \frac{\partial \widetilde{\mathcal{L}}}{\partial \dot{Q}^a}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial Q^a}{\partial q^b}\right)\,. \tag{21.70}$$

The other term of the operator, again by the chain rule, is

$$\frac{\partial \mathcal{L}}{\partial q^b} = \frac{\partial \widetilde{\mathcal{L}}}{\partial Q^a}\frac{\partial Q^a}{\partial q^b} + \frac{\partial \widetilde{\mathcal{L}}}{\partial \dot{Q}^a}\frac{\partial \dot{Q}^a}{\partial q^b}\,. \tag{21.71}$$

Subtracting Equation (21.71) from Equation (21.70), the two terms carrying $\frac{\partial \widetilde{\mathcal{L}}}{\partial \dot{Q}^a}$ cancel by Equation (21.66), and what remains is

$$\mathcal{E}_b[\mathcal{L}] = \left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \widetilde{\mathcal{L}}}{\partial \dot{Q}^a}\right) - \frac{\partial \widetilde{\mathcal{L}}}{\partial Q^a}\right]\frac{\partial Q^a}{\partial q^b} = \widetilde{\mathcal{E}}_a[\widetilde{\mathcal{L}}]\,\frac{\partial Q^a}{\partial q^b}\,,$$

which is Equation (21.68). The final claim follows because the Jacobian is invertible (Definition 21.60): a covector vanishes in one basis exactly when it vanishes in all. □

Remark 21.64 (Why this is automatic). There is a one-line reason behind the computation. The action Equation (21.52) is a *number* attached to a path—a scalar, independent of how the path is labelled—and Hamilton's principle (Theorem 21.46) characterizes the motion as a stationary point of that number. Stationarity is a property of the functional, not of the chart used to write it, so the resulting equations cannot depend on the chart either. Theorem 21.63 is that observation made explicit, and Equation (21.68) records the precise sense in which it holds: not that the components $\mathcal{E}_a$ are unchanged, but that they transform as the components of a covector, which is what makes the statement $\mathcal{E}_a = 0$ coordinate-independent.

Remark 21.65 (Gauge freedom of the Lagrangian). A second, distinct freedom leaves the equations of motion untouched: adding to $\mathcal{L}$ the total time derivative of any function $F(q,t)$,

$$\mathcal{L}' = \mathcal{L} + \frac{\mathrm{d}F(q,t)}{\mathrm{d}t}\,, \tag{21.72}$$

changes the action only by the boundary term $F(q_b,t_b) - F(q_a,t_a)$, which is fixed once the endpoints are fixed and therefore does not affect the variation. The Lagrangian of a system is thus not unique; only its equivalence class modulo point transformations and total derivatives is physical.

That the equations are untouched can also be seen without mentioning the boundary at all, and this is the sharper statement: the Euler–Lagrange operator annihilates a total time derivative *identically*, on every curve, extremal or not. Writing $G = \mathrm{d}F(q,t)/\mathrm{d}t = \left(\partial F/\partial q^b\right)\dot{q}^b + \partial F/\partial t$ for F of class C^2, the velocity derivative of G is $\partial G/\partial \dot{q}^a = \partial F/\partial q^a$, so

$$\mathcal{E}_a[G] = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial F}{\partial q^a}\right) - \frac{\partial G}{\partial q^a} = \left(\frac{\partial^2 F}{\partial q^b \partial q^a}\dot{q}^b + \frac{\partial^2 F}{\partial t\,\partial q^a}\right) - \left(\frac{\partial^2 F}{\partial q^a \partial q^b}\dot{q}^b + \frac{\partial^2 F}{\partial q^a \partial t}\right) = 0 \tag{21.73}$$

by the symmetry of mixed partial derivatives (Proposition 7.73). Since $\mathcal{E}_a$ is linear in its argument, $\mathcal{E}_a[\mathcal{L}'] = \mathcal{E}_a[\mathcal{L}]$, which is the claim. A function with $\mathcal{E}_a \equiv 0$ is called a *null Lagrangian*; in one independent variable Equation (21.73) has a converse, every null Lagrangian being a total derivative, and it is the freedom Equation (21.72) that reappears as the gauge freedom of the electromagnetic potentials and as the boundary terms that must be added to the Einstein–Hilbert action.

21.7.3 The kinetic metric and the geodesic form

Covariance says the equations keep their *form*. The next question is what objects appear in that form, and the answer is the metric and connection of Chapter 13.

For a system of particles with $T = \frac{1}{2}\sum_i m_i\,\dot{\boldsymbol{x}}_i \cdot \dot{\boldsymbol{x}}_i$ and time-independent constraints, substituting $\dot{\boldsymbol{x}}_i = \frac{\partial \boldsymbol{x}_i}{\partial q^a}\dot{q}^a$ (Equation (21.34) with $\frac{\partial \boldsymbol{x}}{\partial t} = 0$) gives a quadratic form in the generalized velocities.

Definition 21.66 (Kinetic metric)**.** The *kinetic metric* of such a system is the symmetric, positive-definite $(0,2)$ tensor field on configuration space

$$g_{ab}(q) = \sum_i m_i \frac{\partial \boldsymbol{x}_i}{\partial q^a} \cdot \frac{\partial \boldsymbol{x}_i}{\partial q^b}\,, \qquad \text{so that} \qquad T = \tfrac{1}{2}\, g_{ab}(q)\, \dot q^a \dot q^b\,. \tag{21.74}$$

It is a metric in the sense of Definition 13.117, of signature $(f,0)$: configuration space is a Riemannian manifold, and $2T\,\mathrm{d}t^2$ is its line element.

Theorem 21.67 (Free motion is geodesic motion)**.** *For $\mathcal{L} = T = \frac{1}{2} g_{ab}(q)\dot q^a \dot q^b$ with no potential, the Euler–Lagrange equations are the autoparallel equation of the Levi-Civita connection of g_{ab},*

$$\ddot q^c + \mathring{\Gamma}^c{}_{ab}\, \dot q^a \dot q^b = 0\,, \qquad \mathring{\Gamma}^c{}_{ab} = \tfrac{1}{2} g^{cd}\left(\partial_a g_{db} + \partial_b g_{ad} - \partial_d g_{ab}\right)\,, \tag{21.75}$$

with $\mathring{\Gamma}$ exactly the Christoffel symbols of Theorem 13.150.

Derivation. From Equation (21.74), $\frac{\partial \mathcal{L}}{\partial \dot q^c} = g_{cb}\dot q^b$ (using $g_{ab} = g_{ba}$) and $\frac{\partial \mathcal{L}}{\partial q^c} = \frac{1}{2}\,\partial_c g_{ab}\,\dot q^a \dot q^b$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(g_{cb}\dot q^b\right) - \tfrac{1}{2}\,\partial_c g_{ab}\,\dot q^a \dot q^b = g_{cb}\ddot q^b + \partial_a g_{cb}\,\dot q^a \dot q^b - \tfrac{1}{2}\,\partial_c g_{ab}\,\dot q^a \dot q^b = 0\,.$$

The middle term is contracted with the symmetric $\dot q^a \dot q^b$, so only its symmetric part in ab contributes; replacing it by $\frac{1}{2}\left(\partial_a g_{cb} + \partial_b g_{ca}\right)$ and multiplying through by g^{dc},

$$\ddot q^d + \tfrac{1}{2} g^{dc}\left(\partial_a g_{cb} + \partial_b g_{ca} - \partial_c g_{ab}\right)\dot q^a \dot q^b = 0\,,$$

which is Equation (21.75). □

Remark 21.68 (Reading the result). A free system does not travel in straight lines through configuration space; it travels along the *straightest available* lines, the autoparallels of the metric its own kinetic energy defines. With a potential the right-hand side acquires $-g^{cd}\partial_d V$, so V acts as a force pushing the motion off the geodesics. Every statement here is the general apparatus of Section 13.10 specialized to a Riemannian configuration space.

21.7.4 Inertial forces are connection coefficients

The Christoffel symbols in Equation (21.75) are built from derivatives of g_{ab}, and g_{ab} depends on the coordinates chosen. They therefore do not vanish merely because the physics is simple—as the flattest possible example shows.

Example 21.69 (A free particle in plane polar coordinates)**.** For one particle of mass m in a plane, Cartesian coordinates give $T = \frac{m}{2}\left(\dot x^2 + \dot y^2\right)$, so $g_{ab} = m\,\delta_{ab}$, all $\partial_c g_{ab} = 0$, all $\mathring{\Gamma} = 0$, and Equation (21.75) reads $\ddot x = \ddot y = 0$. Passing to polar coordinates $x = r\cos\varphi$, $y = r\sin\varphi$—a point transformation Equation (21.63)—the same kinetic energy becomes

$$T = \frac{m}{2}\left(\dot r^2 + r^2\dot\varphi^2\right)\,, \qquad g_{ab} = m\begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}\,, \tag{21.76}$$

on the chart $r > 0$, where the polar labelling is nondegenerate (at $r = 0$ the metric Equation (21.76) degenerates and g^{ab} fails to exist — a defect of the labels, not of the plane). Its only nonvanishing Christoffel symbols are

$$\mathring{\Gamma}^r{}_{\varphi\varphi} = -r\,, \qquad \mathring{\Gamma}^\varphi{}_{r\varphi} = \mathring{\Gamma}^\varphi{}_{\varphi r} = \frac{1}{r}\,. \tag{21.77}$$

The overall factor m has cancelled, as it must: $\mathring{\Gamma}$ is invariant under a constant rescaling $g \to \lambda g$, the λ from differentiating the metric cancelling the λ^{-1} from its inverse. Mass sets

the scale of the metric but not of the connection. The geodesic equation Equation (21.75) then gives

$$\ddot{r} - r\dot{\varphi}^2 = 0\,, \qquad \ddot{\varphi} + \frac{2}{r}\dot{r}\dot{\varphi} = 0\,, \tag{21.78}$$

in which one recognizes the centripetal term $-r\dot{\varphi}^2$ and the Coriolis-like term $2\dot{r}\dot{\varphi}/r$. These are the same straight lines as before, described in different labels.

Remark 21.70 (The lesson: connection $\neq$ curvature). Example 21.69 is worth pausing on, because it isolates a distinction that the rest of this treatise depends upon. The Christoffel symbols Equation (21.77) are not zero, yet the plane is flat: inserting the symbols Equation (21.77) into the Riemann tensor Equation (13.306) gives $R^a{}_{bcd} = 0$ identically, as it must, since curvature is a tensor (Theorem 13.152) and vanishes in the Cartesian chart.

So *in a flat space* a nonvanishing connection means only one of two things, which are worth separating. Curvilinear coordinates in an inertial frame — Example 21.69 — produce $\mathring{\Gamma}$ terms that are pure bookkeeping: no observer feels anything, and an accelerometer carried along reads zero. A non-inertial frame, obtained by a *time-dependent* point transformation, produces $\mathring{\Gamma}$ terms that an accelerometer does register; these are the *inertial* (fictitious) forces of Definition 21.19 proper — centrifugal, Coriolis and their relatives. What the two cases share is that both are removable everywhere at once, by returning to inertial Cartesian coordinates, and they are removable precisely because $R^a{}_{bcd} = 0$. Curvature is what cannot be removed. Keeping these apart is what makes the equivalence principle a sharp statement rather than a slogan (Chapter 42).

21.8 The universality of general covariance

Nothing in Section 21.7.2 used the number of particles, the signature of the metric, or the presence of gravity. The construction is generic, and it is worth stating plainly how far it reaches, because the same formalism is often presented as though it were the private property of general relativity.

21.8.1 Flat spacetime in arbitrary coordinates

Special relativity is usually written in inertial Cartesian coordinates, in which the metric takes the constant form $\eta_{\mu\nu}$ of Equation (13.1). Nothing compels that choice. Minkowski space is a manifold (Chapter 13), its coordinates are labels, and a general coordinate transformation $x^\mu \to x'^\mu(x)$ is as legitimate there as a point transformation is in mechanics.

Definition 21.71 (Relativistic free particle)**.** Let the worldline be timelike and future-directed, parametrized monotonically by any λ, so that $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu < 0$ in the signature $(-,+,+,+)$ used throughout (Appendix E). Proper time is then defined by $c^2\mathrm{d}\tau^2 = -g_{\mu\nu}\,\mathrm{d}x^\mu\mathrm{d}x^\nu$, and the action of a free massive particle is proportional to the proper time elapsed along the worldline,

$$S = -mc^2\int \mathrm{d}\tau = -mc\int \sqrt{-g_{\mu\nu}(x)\,\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda}}\;\mathrm{d}\lambda\,. \tag{21.79}$$

Two independent invariances hold, for two different reasons. The integrand is a *scalar*, which makes S independent of the coordinates x^μ; and it is *homogeneous of degree one* in $\mathrm{d}x^\mu/\mathrm{d}\lambda$, which makes S independent of the parametrization. The timelike hypothesis is what keeps the square root real.

Remark 21.72 (Sign conventions). With the mostly-plus signature this treatise uses, the interval itself is $\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu$, *negative* on a timelike worldline, so $\mathrm{d}s$ is not the proper

distance travelled by a massive particle and $c^2 \mathrm{d}\tau^2 = -\mathrm{d}s^2$. Writing the action in terms of $\mathrm{d}\tau$ rather than $\mathrm{d}s$, as in Equation (21.79), keeps this consistent with the interval classification of Chapter 38 and with Example 21.74 below. A timelike geodesic *maximizes* proper time, so the leading minus sign is what makes S stationary at a minimum.

Proposition 21.73 (Free motion in flat spacetime, any coordinates)**.** *Extremizing Equation* (21.79) *with $g_{\mu\nu}$ the flat metric written in arbitrary coordinates, and choosing $\lambda = \tau$ the proper time, gives*

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\tau^2} + \mathring{\Gamma}^\mu{}_{\nu\rho} \frac{\mathrm{d}x^\nu}{\mathrm{d}\tau} \frac{\mathrm{d}x^\rho}{\mathrm{d}\tau} = 0\,, \tag{21.80}$$

the same geodesic equation as Equation (21.75)*, now on spacetime. In inertial Cartesian coordinates $\mathring{\Gamma} = 0$ and it reduces to $\mathrm{d}^2 x^\mu / \mathrm{d}\tau^2 = 0$, uniform motion.*

Rests on Equations (21.75) *and* (21.79) *and Theorem 13.150.*

Derivation. *Derives Proposition 21.73.*

Write $\dot{x}^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$ and

$$\mathcal{L}_1 = -mc\,u\,, \qquad u := \sqrt{-g_{\mu\nu}\dot{x}^\mu \dot{x}^\nu} > 0\,, \tag{21.81}$$

so that $S = \int \mathcal{L}_1 \mathrm{d}\lambda$ is Equation (21.79); u carries m/s if λ is a time, and $\mathcal{L}_1 \mathrm{d}\lambda$ is an action in J s whatever λ is, since $u\,\mathrm{d}\lambda = c\,\mathrm{d}\tau$. Differentiating Equation (21.81),

$$\frac{\partial \mathcal{L}_1}{\partial \dot{x}^\mu} = \frac{mc}{u} g_{\mu\nu}\dot{x}^\nu\,, \qquad \frac{\partial \mathcal{L}_1}{\partial x^\mu} = \frac{mc}{2u} \partial_\mu g_{\nu\rho}\, \dot{x}^\nu \dot{x}^\rho\,, \tag{21.82}$$

and the Euler–Lagrange equation $\mathcal{E}_\mu[\mathcal{L}_1] = 0$ of Equation (21.67) reads

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{mc}{u} g_{\mu\nu}\dot{x}^\nu\right) - \frac{mc}{2u}\partial_\mu g_{\nu\rho}\dot{x}^\nu \dot{x}^\rho = 0\,. \tag{21.83}$$

Nothing has been assumed about λ so far, and nothing could be: $\mathcal{L}_1$ is homogeneous of degree one in the $\dot{x}^\mu$, so S is unchanged by any monotone reparametrization, and the four equations Equation (21.83) are not independent — contracting them with $\dot{x}^\mu$ gives $0 = 0$ identically. That is the degeneracy of Corollary 16.94, and it is what leaves one function's worth of freedom to be fixed by a gauge choice, exactly as the arclength gauge is imposed *after* the differentiation in Lemma 16.100.

Fix it by $\lambda = \tau$, legitimate because the worldline is timelike and future-directed, so τ is a strictly increasing function of any admissible λ and $u = c$ identically along the curve. With u constant, Equation (21.83) may be divided by $mc/u = m$:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(g_{\mu\nu}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\right) - \tfrac{1}{2}\,\partial_\mu g_{\nu\rho}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\frac{\mathrm{d}x^\rho}{\mathrm{d}\tau} = 0\,.$$

Expanding the derivative, replacing $\partial_\nu g_{\mu\rho}$ by its part symmetric in $\nu\rho$ (the only part that survives contraction with $\mathrm{d}x^\nu/\mathrm{d}\tau\,\mathrm{d}x^\rho/\mathrm{d}\tau$) and multiplying by $g^{\sigma\mu}$ gives Equation (21.80), with $\mathring{\Gamma}$ the Christoffel symbols of Theorem 13.150 — the identical algebra as in the proof of Theorem 21.67, with g_{ab} replaced by $g_{\mu\nu}$.

Finally, the equivalence with the quadratic Lagrangian $\mathcal{L}_2 = \frac{1}{2} m\, g_{\mu\nu}\dot{x}^\mu \dot{x}^\nu$ used in Theorem 21.67. $\mathcal{L}_2$ has no explicit λ, so Theorem 16.29 makes $\dot{x}^\mu \partial \mathcal{L}_2/\partial \dot{x}^\mu - \mathcal{L}_2 = \mathcal{L}_2$ constant along its extremals; hence u is automatically constant for $\mathcal{L}_2$, and normalizing that constant to $u = c$ selects proper time. Under that normalization the two Lagrangians have the same extremals: $\mathcal{E}_\mu[\mathcal{L}_2] = 0$ is the display above multiplied by m. The square-root form is the one with a geometric meaning (Equation (21.79)); the quadratic form is the one that is easier to vary, and $\mathcal{L}_2$ is what will be used whenever a computation is wanted. Note that $\mathcal{L}_2$ is *not* reparametrization invariant — it fixes λ to be affine — which is exactly the trade being made. □

Example 21.74 (Rindler coordinates)**.** The coordinates natural to a uniformly accelerated observer in flat spacetime are

$$ct = X\sinh\!\left(\frac{aT}{c}\right), \qquad x = X\cosh\!\left(\frac{aT}{c}\right), \qquad X > 0\,, \tag{21.84}$$

covering the wedge $x > |ct|$. Differentiating and substituting into $\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + \mathrm{d}x^2$, the cross terms cancel and the hyperbolic identity $\cosh^2 - \sinh^2 = 1$ leaves

$$\mathrm{d}s^2 = -\frac{a^2X^2}{c^2}\,\mathrm{d}T^2 + \mathrm{d}X^2\,, \tag{21.85}$$

a metric with nonconstant components and hence nonvanishing Christoffel symbols. Both statements can be checked in a few lines, and it is worth doing so, because the example carries the whole distinction of Remark 21.70.

The transverse directions y, z are untouched by Equation (21.84) and contribute $\mathrm{d}y^2+\mathrm{d}z^2$ with constant components, so only the (T, X) block of Equation (21.85) can produce anything. There $g_{TT} = -a^2X^2/c^2$ and $g_{XX} = 1$ are diagonal and depend on X alone, so Equation (21.75) gives, for the only two nonvanishing symbols,

$$\mathring{\Gamma}^X{}_{TT} = -\tfrac{1}{2}g^{XX}\partial_X g_{TT} = \frac{a^2X}{c^2}\,, \qquad \mathring{\Gamma}^T{}_{TX} = \mathring{\Gamma}^T{}_{XT} = \tfrac{1}{2}g^{TT}\partial_X g_{TT} = \frac{1}{X}\,, \tag{21.86}$$

carrying $\mathrm{m/s^2}$ and $1/\mathrm{m}$ respectively, which is what makes every term of Equation (21.80) dimensionally uniform. An observer at fixed X has $\mathrm{d}X/\mathrm{d}\tau = 0$ and, from Equation (21.85), $\mathrm{d}T/\mathrm{d}\tau = c^2/(aX)$; the only nonvanishing component of the four-acceleration $A^\mu = \mathrm{d}^2x^\mu/\mathrm{d}\tau^2 + \mathring{\Gamma}^\mu{}_{\nu\rho}\,(\mathrm{d}x^\nu/\mathrm{d}\tau)\,(\mathrm{d}x^\rho/\mathrm{d}\tau)$ is therefore

$$A^X = \mathring{\Gamma}^X{}_{TT}\left(\frac{\mathrm{d}T}{\mathrm{d}\tau}\right)^2 = \frac{a^2X}{c^2}\cdot\frac{c^4}{a^2X^2} = \frac{c^2}{X}\,, \tag{21.87}$$

whose magnitude, g_{XX} being unity, is c^2/X: the observer must accelerate to stay put, and feels a weight, exactly as the connection terms of Equation (21.80) predict. It equals a only on the worldline $X = c^2/a$, which is also the only one on which Equation (21.85) makes T the proper time.

And yet the curvature vanishes. With only the two symbols Equation (21.86), and both independent of T, the (T, X) block has a single independent curvature component (Proposition 13.154), and Equation (13.306) gives

$$R^T{}_{XTX} = \partial_T\mathring{\Gamma}^T{}_{XX} - \partial_X\mathring{\Gamma}^T{}_{TX} + \mathring{\Gamma}^T{}_{T\lambda}\mathring{\Gamma}^\lambda{}_{XX} - \mathring{\Gamma}^T{}_{X\lambda}\mathring{\Gamma}^\lambda{}_{TX} = 0 + \frac{1}{X^2} + 0 - \frac{1}{X^2} = 0\,, \tag{21.88}$$

the two surviving terms being $-\partial_X(1/X)$ and $-\mathring{\Gamma}^T{}_{XT}\mathring{\Gamma}^T{}_{TX}$. So $R^\mu{}_{\nu\rho\sigma} = 0$ identically, as it must, curvature being a tensor (Theorem 13.152) that vanishes in the inertial Cartesian chart. This is flat spacetime in curvilinear clothing, and $X = 0$ is a horizon where the chart degenerates, not a place where anything happens to the spacetime.

Rests on Equations (21.75) *and* (21.85) *and Theorem 13.152.*

21.8.2 What general relativity adds

The three settings just surveyed share their entire formal apparatus: a manifold of states, a metric, its Levi-Civita connection, geodesics as free motion, and equations of motion that hold in any coordinates whatever. Table 21.1 sets them side by side.

Example 21.75 (Configuration space can itself be curved)**.** It would be tidy, but false, to say that curvature is what separates mechanics from gravitation. A particle constrained

Table 21.1: One formalism, three instantiations. General covariance — the freedom to relabel — is present in all three and therefore distinguishes none of them. Nor does curvature as such (Example 21.75): what distinguishes general relativity is that the curved manifold is spacetime itself, and that its geometry is dynamical rather than fixed in advance.

	Classical mechanics	Special relativity	General relativity
Manifold	configuration space, f degrees of freedom	Minkowski spacetime, $3+1$	spacetime, $3+1$
Metric	kinetic metric $g_{ab}(q)$, signature $(f,0)$ (time-independent constraints)	$\eta_{\mu\nu}$ in any coordinates, $(3,1)$	$g_{\mu\nu}(x)$, $(3,1)$
Relabelling	point transformations $Q(q,t)$; the geometric rows assume $Q(q)$	general coordinate transformations	general coordinate transformations
Free motion	geodesics of g_{ab}, for $V=0$	geodesics of η	geodesics of $g_{\mu\nu}$
Connection	may be nonzero (curvilinear coordinates)	may be nonzero (accelerated frames)	generally nonzero
Curvature	$R^{a}{}_{bcd}$ may be nonzero: constrained systems have genuinely curved configuration spaces	$R^{\mu}{}_{\nu\rho\sigma}=0$ always: spacetime is flat	$R^{\mu}{}_{\nu\rho\sigma}$ generically nonzero and *dynamical*: the field equations fix its Ricci part from $T_{\mu\nu}$, while the Weyl part propagates

to a sphere of radius ρ has $T = \frac{m\rho^2}{2}\left(\dot\theta^2 + \sin^2\theta\,\dot\varphi^2\right)$, so its kinetic metric Equation (21.74) is $m\rho^2$ times the round metric on the unit sphere. Its Gaussian curvature is therefore $K = 1/(m\rho^2)$ — carrying units $1/(\mathrm{kg\,m^2})$, as the SI axiom demands of the curvature of a metric whose components are masses times areas, and *not* the $1/\rho^2$ of the geometric sphere in Example 13.33, which is a different metric. The distinction is one of scale only: the $(1,3)$ tensor $R^a{}_{bcd}$ is invariant under the constant rescaling $g \to \lambda g$, so it is the same as the geometric sphere's and is nonzero either way. Free trajectories are the great circles, geodesics of a genuinely curved space. The same holds for the double pendulum and for rigid-body rotation.

What distinguishes general relativity is therefore not curvature as such, but *which* manifold is curved and what determines it: in mechanics the curvature belongs to an abstract space of configurations and is fixed once the constraints are imposed by hand, whereas in general relativity it belongs to spacetime itself and is dynamical — its Ricci part fixed algebraically by the local energy–momentum, its Weyl part propagating as a field in its own right (Chapter 44).

Three conclusions follow, and they are the point of this section.

1. **General covariance by itself carries little physical content.** The freedom to use arbitrary coordinates is available in Newtonian mechanics (Theorem 21.63) and in flat spacetime (Proposition 21.73) just as much as in curved spacetime, so a theory written covariantly is not thereby a theory of gravity. This is Kretschmann's objection to Einstein, raised in 1917 [454]: essentially any theory can be put in generally covariant form. The price, usually left unsaid, is that covariantizing a theory with fixed background structure requires promoting that structure to explicit fields — Newtonian gravity becomes Newton–Cartan theory — so what does real work is not general covariance but *background independence*, the absence of non-dynamical absolute objects. Whether, and in what form, general covariance retains physical content has been disputed for a century and is not settled here [1869]; the position this treatise takes is the majority one just stated.

2. **The value of the connection at a point carries no invariant information; curvature does.** $\mathring{\Gamma}$ is not a tensor (Equation (13.296)), and for a torsion-free connection its components can always be made to vanish *at any chosen point* — indeed along an entire geodesic, which is what an extended freely falling observer enjoys. What cannot be done, once the space is curved, is to make them vanish throughout a neighbourhood, however small. The Riemann tensor is a tensor (Theorem 13.152): nonzero in one chart, nonzero in all, and no relabelling removes it. Inertial forces belong to the first category and tidal forces to the second, which is why the observable *local* content of gravity is tidal (Section 43.2). "Local" is not idle: a flat connection can still carry global physical content through its holonomy, as the conical geometry around a cosmic string shows, so vanishing curvature guarantees removability only region by region.

3. **The equivalence principle is what licenses describing gravity by a connection at all.** Its empirical core is the universality of free fall: all test bodies fall identically, whatever their mass or composition — the content of the Eötvös-type experiments of Chapter 42. That universality is what makes a gravitational field describable as a property of spacetime rather than as a force with body-dependent couplings, and it is not a theorem but a measured fact. (The stronger forms add local Lorentz invariance and local position invariance, and extend the claim to self-gravitating bodies.)

   Once gravity is represented by a metric connection, the vanishing of $\mathring{\Gamma}$ in a freely falling frame follows as *geometry*, not as further physics: Riemann normal coordinates

make $\mathring{\Gamma}$ vanish at any chosen point, and Fermi normal coordinates make it vanish along an entire freely falling worldline, for any metric whatever. What no choice of chart can do is make $\mathring{\Gamma}$ vanish on an open neighbourhood, because the second derivatives of the metric there carry the curvature, which is a tensor: tidal deviation between neighbouring free-fall trajectories enters at second order in the separation and is chart-independent. That is the precise sense in which the equivalence principle is *local*, and the failure is not merely "global". Gravity is thus the one "inertial force" that relabelling removes along a worldline but never on a neighbourhood of it, and that irremovability *is* spacetime curvature. This is the bridge from the present chapter to Part V.

Remark 21.76 (Where the mathematics lives). Every geometric object used above — manifold, chart, metric, connection, geodesic, curvature — is defined once and in full generality, for dimension $D = p + q$ and signature (p, q), in Chapter 13. This chapter has developed no new mathematics; it has pointed out that the mathematics classical mechanics has used since Lagrange is the mathematics general relativity uses, and that what separates the theories is not the language but the physics it is asked to describe: in mechanics the geometry is a bookkeeping device for constraints imposed by hand, while in general relativity the geometry of spacetime is itself the dynamical variable, determined by its own field equations.

With the Lagrangian formulation complete, the natural next step is to trade the generalized velocities for the generalized momenta Equation (21.43) as independent variables. That exchange—the Legendre transformation of the Lagrangian—produces the Hamiltonian formulation of mechanics, developed in Chapter 22.

Chapter 22

Hamiltonian Mechanics

The Lagrangian formulation of Chapter 21 describes a system with f degrees of freedom by the generalized coordinates q^a and their velocities $\dot{q}^a$. The Hamiltonian formulation trades the velocities for the generalized momenta $p_a = \partial\mathcal{L}/\partial\dot{q}^a$, and in doing so moves the dynamics onto *phase space*, the $2f$-dimensional space of the pairs (q^a, p_a). The reward for the exchange is structural: the equations of motion become a symmetric first-order system, the transformations that preserve their form—the canonical transformations—constitute a far larger class than the point transformations of Lagrangian mechanics, and the whole dynamics is expressed through a single algebraic operation, the Poisson bracket [1585, 2027].

Notation 22.1. The index conventions of Notation 21.1 remain in force; additionally $c, d, e = 1, \dots, 2f$ label phase-space components. Poisson brackets are written with braces, $\{u, v\}$, and Lagrange brackets with square brackets, $[u, v]$, a subscript indicating the coordinate pair in which they are computed.

22.1 From the Lagrangian to the Hamiltonian

22.1.1 The Legendre transformation of the Lagrangian

The generalized momentum conjugate to the coordinate q^a was defined in Definition 21.34 as $p_a = \partial\mathcal{L}/\partial\dot{q}^a$. It is the variable that replaces the generalized velocity in the description that follows.

Definition 22.2 (Hamiltonian)**.** The *Hamiltonian* of a system is the Legendre transformation, in the generalized velocities $\dot{q}^a$, of the Lagrangian of the system:

$$\mathcal{H}(q, p, t) = \dot{q}^a p_a - \mathcal{L}(q, \dot{q}, t)\,, \tag{22.1}$$

the generalized velocities being expressed in terms of (q, p, t) through the inversion of Equation (21.43).

Rests on Definitions 21.31 and 21.34.

The definition presupposes that the inversion can be carried out, and that is a genuine hypothesis rather than a formality: it is the exact point at which the two formulations can fail to be equivalent.

Proposition 22.3 (Invertibility of the Legendre map)**.** *Let $\mathcal{L}$ be of class C^2 and let the* velocity Hessian

$$W_{ab}(q, \dot{q}, t) = \frac{\partial^2 \mathcal{L}}{\partial\dot{q}^a \partial\dot{q}^b} \tag{22.2}$$

be non-singular at a point of the tangent bundle. Then Equation (21.43) can be solved near that point for the velocities, $\dot{q}^a = \dot{q}^a(q, p, t)$, by a C^1 function, and Equation (22.1) defines $\mathcal{H}$ there.

Rests on Equation (21.43) and Corollary A.292.

Proof. *Derives Proposition 22.3.*

At fixed (q,t) the map $\dot q \longmapsto p$ of Equation (21.43) is C^1, and its Jacobian matrix is

$$\frac{\partial p_a}{\partial \dot q^b} = \frac{\partial^2 \mathcal{L}}{\partial \dot q^b \partial \dot q^a} = W_{ab}\,,$$

symmetric by Proposition 7.73. Where $\det W \neq 0$ the inverse function theorem Corollary A.292 supplies a C^1 local inverse, and the dependence on the parameters (q,t) is inherited from that of $\mathcal{L}$. □

Remark 22.4 (When the Legendre map is singular). For the Lagrangians of Section 21.7.3, $\mathcal{L} = \frac{1}{2} g_{ab}(q)\dot q^a \dot q^b - V(q)$, the Hessian is the kinetic metric itself, $W_{ab} = g_{ab}$, which is positive definite, so the inversion is global in the velocities and gives $\dot q^a = g^{ab} p_b$. The condition $\det W \neq 0$ is also the strict form of Legendre's necessary condition Theorem 16.51 for the variational problem the Lagrangian poses, which is why the same convexity governs minimality and the passage to the Hamiltonian. When it fails — as it does for every gauge theory, where the Lagrangian does not depend on some of the velocities at all — the momenta are not independent functions of the velocities, the relations among them are constraints, and the formalism of this chapter must be rebuilt: that is the subject of Chapter 26.

Rests on Proposition 22.3 and Definition 21.66.

Remark 22.5 (Units). $\mathcal{H}$ carries the unit of energy, J, whatever the units of the individual q^a. Each product $\dot q^a p_a$ in Equation (22.1) is an energy, and necessarily so: by Equation (21.43) the unit of p_a is that of $\mathcal{L}$ divided by that of $\dot q^a$, so the two units cancel term by term. For a Cartesian coordinate p_a is a linear momentum, kg m/s; for an angle it is an angular momentum, J s. The one combination whose unit never depends on the choice of coordinates is $p_a\,\mathrm{d}q^a$, which is an action, J s — the unit of $\hbar$, and the reason the symplectic structure built from it in Chapter 24 is measured in the same units as the quantum of action (Remark 5.116).

Rests on Equations (21.43) and (22.1).

22.1.2 Hamilton's equations

The equations that follow are Hamilton's, from the general method in dynamics of [96]; he obtained them as the characteristic-function formulation whose partial differential equation is the subject of Chapter 23, and the canonical pair is what remains when that function is not solved for.

Theorem 22.6 (Hamilton's equations)**.** *The equations of motion of a system with Hamiltonian $\mathcal{H}(q,p,t)$ are*

$$\dot q^a = \frac{\partial \mathcal{H}}{\partial p_a}\,, \tag{22.3}$$

$$\dot p_a = -\frac{\partial \mathcal{H}}{\partial q^a}\,, \tag{22.4}$$

together with the relation

$$\frac{\partial \mathcal{H}}{\partial t} = -\frac{\partial \mathcal{L}}{\partial t}\,. \tag{22.5}$$

Here the active forces are taken to be conservative, so that the Lagrangian obeys Equation (21.39).

Rests on Definition 22.2 and Equations (21.39) and (21.43).

Proof. *Derives Theorem 22.6.*

Differentiate the Lagrangian as a function of its own arguments $(q,\dot q,t)$:

$$\mathrm{d}\mathcal{L} = \frac{\partial \mathcal{L}}{\partial q^a}\,\mathrm{d}q^a + \frac{\partial \mathcal{L}}{\partial \dot q^a}\,\mathrm{d}\dot q^a + \frac{\partial \mathcal{L}}{\partial t}\,\mathrm{d}t = \frac{\partial \mathcal{L}}{\partial q^a}\,\mathrm{d}q^a + p_a\,\mathrm{d}\dot q^a + \frac{\partial \mathcal{L}}{\partial t}\,\mathrm{d}t\,, \tag{22.6}$$

the last step by the definition Equation (21.43) of the momentum. Differentiating Equation (22.1) instead, and subtracting,

$$\mathrm{d}\mathcal{H} = \dot{q}^a\,\mathrm{d}p_a + p_a\,\mathrm{d}\dot{q}^a - \mathrm{d}\mathcal{L} = \dot{q}^a\,\mathrm{d}p_a - \frac{\partial\mathcal{L}}{\partial q^a}\,\mathrm{d}q^a - \frac{\partial\mathcal{L}}{\partial t}\,\mathrm{d}t\,. \tag{22.7}$$

From Equations (22.1) and (22.6) (the two terms in $\mathrm{d}\dot{q}^a$ cancel).

The cancellation of $p_a\,\mathrm{d}\dot{q}^a$ is the whole content of the Legendre transformation: no differential of a velocity survives in Equation (22.7), so $\mathcal{H}$ is genuinely a function of (q, p, t) and of nothing else — which is what Proposition 22.3 licenses us to assume.

Comparing Equation (22.7) with the differential of $\mathcal{H}$ read off its own arguments,

$$\mathrm{d}\mathcal{H} = \frac{\partial\mathcal{H}}{\partial q^a}\,\mathrm{d}q^a + \frac{\partial\mathcal{H}}{\partial p_a}\,\mathrm{d}p_a + \frac{\partial\mathcal{H}}{\partial t}\,\mathrm{d}t\,,$$

and using that $\mathrm{d}q^a$, $\mathrm{d}p_a$ and $\mathrm{d}t$ are independent, gives three identities:

$$\frac{\partial\mathcal{H}}{\partial p_a} = \dot{q}^a\,, \qquad \frac{\partial\mathcal{H}}{\partial q^a} = -\frac{\partial\mathcal{L}}{\partial q^a}\,, \qquad \frac{\partial\mathcal{H}}{\partial t} = -\frac{\partial\mathcal{L}}{\partial t}\,. \tag{22.8}$$

The first is Equation (22.3) and the third is Equation (22.5). Neither uses the equations of motion: both are properties of the transformation alone.

The dynamics enters only now. By the Euler–Lagrange equations Equation (21.39) and Equation (21.43),

$$\dot{p}_a = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial\mathcal{L}}{\partial\dot{q}^a}\right) = \frac{\partial\mathcal{L}}{\partial q^a} = -\frac{\partial\mathcal{H}}{\partial q^a}\,,$$

the last step by the second identity of Equation (22.8). This is Equation (22.4). □

Remark 22.7 (Which half is kinematics and which is dynamics). The asymmetry exposed by the proof is worth keeping in view. Equation (22.3) holds for *any* $\mathcal{H}$ obtained by the Legendre transformation, whether or not the motion satisfies the equations of motion; it merely says that the transformation is invertible and inverts the way a Legendre transformation does. Equation (22.4) is Newton's second law in disguise. The elegance of the pair is therefore partly notational — but only partly, since it is exactly this symmetry of appearance that makes the canonical transformations of Section 22.2 available, and those have no Lagrangian counterpart.

Rests on Theorem 22.6.

Remark 22.8 (Non-conservative forces). If some active forces are not derivable from a potential, the Euler–Lagrange equations carry the generalized force Q_a of Equation (21.38) on their right-hand side, and the last line of the proof gives instead

$$\dot{p}_a = -\frac{\partial\mathcal{H}}{\partial q^a} + Q_a\,, \tag{22.9}$$

Equation (22.3) being untouched, since it does not use the equations of motion. Everything in this chapter that rests on the pair Equations (22.3) and (22.4) — conservation of $\mathcal{H}$, the Poisson bracket form of the evolution, Liouville's theorem — fails in exactly the way Equation (22.9) predicts when $Q_a \neq 0$.

Rests on Theorem 22.6 and Equation (21.38).

Where the Lagrange equations are f second-order equations for the $q^a(t)$, Hamilton's equations are $2f$ first-order equations for the phase-space trajectory $(q^a(t), p_a(t))$.

22.1.3 The Hamiltonian and the energy

Proposition 22.9 (Time derivative of the Hamiltonian)**.** *Along the motion,*

$$\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} = \frac{\partial \mathcal{H}}{\partial t}\,. \tag{22.10}$$

Rests on Equations (22.3) and (22.4).

Proof. *Derives Proposition 22.9.*

$\mathcal{H}$ depends on time both through the trajectory and explicitly, so by the chain rule Proposition 7.72,

$$\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} = \frac{\partial \mathcal{H}}{\partial q^a}\dot{q}^a + \frac{\partial \mathcal{H}}{\partial p_a}\dot{p}_a + \frac{\partial \mathcal{H}}{\partial t} = \frac{\partial \mathcal{H}}{\partial q^a}\frac{\partial \mathcal{H}}{\partial p_a} - \frac{\partial \mathcal{H}}{\partial p_a}\frac{\partial \mathcal{H}}{\partial q^a} + \frac{\partial \mathcal{H}}{\partial t}\,,$$

substituting Equations (22.3) and (22.4). The first two terms are the same sum with opposite signs and cancel. □

Theorem 22.10 (Conservation of the Hamiltonian)**.** *If the active forces of a system are conservative and the Hamiltonian of the system does not depend explicitly on time, then the Hamiltonian is constant.*

Rests on Proposition 22.9.

Proof. *Derives Theorem 22.10.*

The hypothesis on the forces is what puts Equation (22.4) in force and hence Equation (22.10); the hypothesis $\partial\mathcal{H}/\partial t = 0$ then makes its right-hand side vanish, so $\mathrm{d}\mathcal{H}/\mathrm{d}t = 0$ along every motion and $\mathcal{H}$ is constant on each trajectory (Corollary 7.36). By Equation (22.5) the condition $\partial\mathcal{H}/\partial t = 0$ is the same as $\partial\mathcal{L}/\partial t = 0$, so the hypothesis may be tested on whichever of the two functions is in hand. □

Proposition 22.11 (Hamiltonian and total energy)**.** *If the kinetic energy T does not depend explicitly on time, then the Hamiltonian is the total energy of the system,*

$$\mathcal{H} = E\,. \tag{22.11}$$

Rests on Equation (22.1) and Proposition 21.38.

Proof. *Derives Proposition 22.11.*

Under the stated hypothesis Proposition 21.38 gives $E = \dot{q}^a p_a - \mathcal{L}$, and the right-hand side is Equation (22.1) verbatim. The two quantities are therefore the same function, expressed once in the velocities and once in the momenta. □

Remark 22.12 (The hypothesis, stated sharply). "T does not depend explicitly on time" is the wording of Proposition 21.38 and is a shorthand. What the identification $\mathcal{H} = T + V$ actually needs is that T be a homogeneous quadratic form in the generalized velocities and V be independent of them — which follows from Equation (21.74) when the constraints are scleronomous (Remark 21.8), and fails when they are not. The step is Euler's theorem on homogeneous functions: if T is homogeneous of degree k in the $\dot{q}^a$, then $\dot{q}^a\,\partial T/\partial\dot{q}^a = kT$, so for $k = 2$,

$$\dot{q}^a p_a = \dot{q}^a \frac{\partial T}{\partial \dot{q}^a} = 2T\,, \qquad \mathcal{H} = 2T - (T - V) = T + V = E\,. \tag{22.12}$$

A rheonomous constraint makes T the sum of terms of degrees 2, 1 and 0 in the velocities, the theorem then gives $\dot{q}^a p_a = 2T_2 + T_1$, and $\mathcal{H} = T_2 - T_0 + V$ is a conserved quantity that is *not* the energy — the rotating-frame case of Chapter 29, where the difference is the centrifugal term. Conservation of $\mathcal{H}$ and conservation of E are thus two different statements that coincide only under the hypothesis above.

Rests on Proposition 22.11 and Equation (21.74).

Remark 22.13 (Euler's theorem on homogeneous functions is owed). Euler's theorem on homogeneous functions, used in Equation (22.12), is a statement of several-variable differential calculus and belongs in Chapter 7 beside the chain rule Proposition 7.72, which is all its one-line proof requires. It is not stated there, and it is quoted here rather than proved, so that the gap is visible rather than papered over. Nothing else in this chapter depends on it: Proposition 22.11 itself rests only on Proposition 21.38, and Theorem 22.10 does not use it at all.

Rests on Propositions 7.72 and 22.11.

22.2 Canonical transformations

22.2.1 Definition and transformation conditions

Definition 22.14 (Phase-space coordinate transformation)**.** A *transformation of coordinates* of phase space is a transformation of the form

$$Q^a = Q^a(q, p, t)\,, \tag{22.13}$$

$$P_a = P_a(q, p, t)\,. \tag{22.14}$$

It is assumed invertible and twice continuously differentiable, so that (q, p) may equally be regarded as functions of (Q, P, t).

Rests on Definition 22.2.

Definition 22.15 (Canonical transformation)**.** A *canonical transformation* is a transformation of phase-space coordinates that leaves Hamilton's equations invariant in form; that is, one for which

$$\delta S[q, p] = \delta S[Q, P]\,. \tag{22.15}$$

Here $S[q,p] = \int_{t_a}^{t_b} (p_a \dot{q}^a - \mathcal{H})\,\mathrm{d}t$ is the action of Definition 21.45 written in phase-space variables, and $S[Q,P] = \int_{t_a}^{t_b} \left(P_a \dot{Q}^a - K\right)\mathrm{d}t$ is the same functional built from the new variables and from a new Hamiltonian K.

Rests on Definitions 21.45 and 22.14 and Theorem 22.6.

Remark 22.16 (Why the class is larger than the point transformations). Equation (22.13) lets Q^a depend on the momenta, which a point transformation Equation (21.63) of Lagrangian mechanics cannot. The gain is not cosmetic: a canonical transformation may exchange coordinates and momenta outright — $Q = p$, $P = -q$ satisfies every condition below — so the distinction between a "position" and a "momentum" has no invariant meaning in phase space. That is what makes it possible to look for a transformation in which *every* new coordinate is cyclic, which is the programme of Chapter 23.

Rests on Definitions 21.60 and 22.15.

Proposition 22.17 (Transformation conditions)**.** *A transformation Equations (22.13) and (22.14) is canonical if and only if*

$$\frac{\partial Q^a}{\partial q^b} = \frac{\partial p_b}{\partial P_a}\,, \qquad \frac{\partial Q^a}{\partial p_b} = -\frac{\partial q^b}{\partial P_a}\,, \tag{22.16}$$

$$\frac{\partial P_a}{\partial q^b} = -\frac{\partial p_b}{\partial Q^a}\,, \qquad \frac{\partial P_a}{\partial p_b} = \frac{\partial q^b}{\partial Q^a}\,, \tag{22.17}$$

where the derivatives on the left-hand sides are taken in the old variables and those on the right-hand sides in the new. That is: on the left, the variables held fixed are the remaining (q, p); on the right, the remaining (Q, P).

Rests on Definition 22.15 and Equations (22.13) and (22.14).

The conditions are $4f^2$ scalar equations relating derivatives taken in two different sets of variables, and the shortest route to them runs through the generating function of Section 22.2.2. The derivation is therefore deferred to Section 22.2.3, once that function is available.

22.2.2 Generating functions

Postulating a canonical transformation leaves an arbitrary function F at our disposal: since the variational statements Equation (22.15) agree, the two actions may differ by the integral of a total time derivative,

$$S[q,p] = S[Q,P] + \int_{t_a}^{t_b} \frac{\mathrm{d}F}{\mathrm{d}t}\,\mathrm{d}t\,. \tag{22.18}$$

The function F is called the *generating function* of the transformation. Once it is known, the transformation from (q,p) to (Q,P) is determined uniquely. To obtain the transformation one must take

$$F = F(q,p,Q,P)\,, \tag{22.19}$$

which involves the $4f$ phase-space variables, of which only $2f$ are independent, the transformation itself relating the two sets; the dependent variables are eliminated by passing to one of the following mixed forms. In each case K denotes the Hamiltonian that governs the transformed variables (Q,P).

Proposition 22.18 (The mixed generating functions)**.** *Let a canonical transformation admit a generating function depending on one old and one new variable of each conjugate pair, the two being independent. Then the transformation and the new Hamiltonian are given by the following relations.*

Transformation $(q,Q)\to(p,P)$**.**

$$F_1 = F_1(q,Q,t)\,,\qquad p_a = \frac{\partial F_1}{\partial q^a}\,,\qquad P_a = -\frac{\partial F_1}{\partial Q^a}\,,\qquad K = \mathcal{H} + \frac{\partial F_1}{\partial t}\,. \tag{22.20}$$

Transformation $(q,P)\to(p,Q)$**.**

$$F_2 = F_2(q,P,t)\,,\qquad p_a = \frac{\partial F_2}{\partial q^a}\,,\qquad Q^a = \frac{\partial F_2}{\partial P_a}\,,\qquad K = \mathcal{H} + \frac{\partial F_2}{\partial t}\,. \tag{22.21}$$

Transformation $(p,Q)\to(q,P)$**.**

$$F_3 = F_3(p,Q,t)\,,\qquad q^a = -\frac{\partial F_3}{\partial p_a}\,,\qquad P_a = -\frac{\partial F_3}{\partial Q^a}\,,\qquad K = \mathcal{H} + \frac{\partial F_3}{\partial t}\,. \tag{22.22}$$

Rests on Equation (22.18) *and Definition 22.15.*

Proof. *Derives Proposition 22.18.*

Write out Equation (22.18) with the two actions of Definition 22.15. Since the equality of the two variational problems must hold for every comparison path between the same endpoints, and not merely for solutions, the integrands agree identically on phase space:

$$p_a\dot q^a - \mathcal{H} = P_a\dot Q^a - K + \frac{\mathrm{d}F}{\mathrm{d}t}\,. \tag{22.23}$$

Take first $F = F_1(q,Q,t)$, which is legitimate when (q,Q) are $2f$ independent variables. Then

$$\frac{\mathrm{d}F_1}{\mathrm{d}t} = \frac{\partial F_1}{\partial q^a}\dot q^a + \frac{\partial F_1}{\partial Q^a}\dot Q^a + \frac{\partial F_1}{\partial t}\,,$$

and substituting into Equation (22.23) and collecting terms,

$$\left(p_a - \frac{\partial F_1}{\partial q^a}\right)\dot{q}^a - \left(P_a + \frac{\partial F_1}{\partial Q^a}\right)\dot{Q}^a - \mathcal{H} + K - \frac{\partial F_1}{\partial t} = 0\,.$$

At a given instant the $2f$ quantities $\dot{q}^a$ and $\dot{Q}^a$ may be prescribed arbitrarily and independently, because (q, Q) are independent coordinates and the relation holds along every path. An affine expression that vanishes for all values of its variables has vanishing coefficients, which gives the three relations of Equation (22.20).

The remaining forms follow by Legendre transformation in one pair of variables, exactly as $\mathcal{H}$ followed from $\mathcal{L}$. Put $F_2 = F_1 + P_a Q^a$; then

$$\mathrm{d}F_2 = \mathrm{d}F_1 + P_a\,\mathrm{d}Q^a + Q^a\,\mathrm{d}P_a = p_a\,\mathrm{d}q^a + Q^a\,\mathrm{d}P_a + \frac{\partial F_1}{\partial t}\,\mathrm{d}t\,,$$

using $\mathrm{d}F_1 = p_a \mathrm{d}q^a - P_a \mathrm{d}Q^a + (\partial F_1/\partial t)\mathrm{d}t$ from Equation (22.20). No $\mathrm{d}Q^a$ survives, so F_2 is a function of (q, P, t), and reading off its partial derivatives gives Equation (22.21) — the last of them because $\partial F_2/\partial t = \partial F_1/\partial t$. Put instead $F_3 = F_1 - p_a q^a$; the same computation gives $\mathrm{d}F_3 = -q^a \mathrm{d}p_a - P_a \mathrm{d}Q^a + (\partial F_1/\partial t)\mathrm{d}t$, which is Equation (22.22). The fourth member of the family, $F_4 = F_1 + P_a Q^a - p_a q^a$, is obtained by performing both Legendre transformations and satisfies $q^a = -\partial F_4/\partial p_a$, $Q^a = \partial F_4/\partial P_a$, $K = \mathcal{H} + \partial F_4/\partial t$. □

Remark 22.19 (Which form exists). Not every canonical transformation admits every mixed form, and some admit none of the four. The identity transformation $Q^a = q^a$, $P_a = p_a$ is generated by $F_2 = q^a P_a$ but has no F_1 at all, since q and Q are not then independent; the exchange $Q^a = p_a$, $P_a = -q^a$ has $F_1 = q^a Q^a$ but no F_2. The requirement in each case is the independence assumed in Proposition 22.18, which by Corollary A.292 is the non-vanishing of the corresponding Jacobian — $\det\left(\partial^2 F_1/\partial q^a \partial Q^b\right) \neq 0$ for the first form, and correspondingly for the others. Nothing in the results of this chapter depends on a generating function existing in a particular form: Lemma 22.20 below is proved from Equation (22.18) itself, which asks only that *some* F exist.

Rests on Proposition 22.18 and Corollary A.292.

22.2.3 Derivation of the transformation conditions

Assembling the $2f$ new coordinates into a single column, as Equation (22.55) does for the old ones, the transformation Equations (22.13) and (22.14) has at each instant the Jacobian matrix

$$M = \frac{\partial\,(Q, P)}{\partial\,(q, p)} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \qquad \begin{aligned} A_{ab} &= \frac{\partial Q^a}{\partial q^b}\,, & B_{ab} &= \frac{\partial Q^a}{\partial p_b}\,, \\ C_{ab} &= \frac{\partial P_a}{\partial q^b}\,, & D_{ab} &= \frac{\partial P_a}{\partial p_b}\,. \end{aligned} \tag{22.24}$$

Everything below is a statement about M and the symplectic matrix J of Equation (22.56), whose defining properties $J^{\mathsf{T}} = -J$, $J^2 = -\mathbb{1}_{2f}$ and $J^{-1} = -J$ are recorded in Equation (5.173).

Lemma 22.20 (The symplectic condition)**.** *The Jacobian Equation (22.24) of a canonical transformation satisfies, at every point and at every instant,*

$$M^{\mathsf{T}} J\, M = J\,, \tag{22.25}$$

that is, $M \in \mathrm{Sp}(2f, \mathbb{R})$. Conversely, a transformation not involving the time explicitly whose Jacobian satisfies Equation (22.25) is canonical.

Rests on Equation (22.18), Definition 5.130, and Proposition 7.73.

Proof. *Derives Lemma 22.20.*

Necessity. Regard (q, p, t) as the independent variables, so that Q^a, P_a and the generating function F of Equation (22.18) are functions of them. Expanding $\dot{Q}^a$ and $\mathrm{d}F/\mathrm{d}t$ by the chain rule Proposition 7.72 in Equation (22.23) and collecting the coefficients of the independently prescribable $\dot{q}^b$ and $\dot{p}_b$, as in the proof of Proposition 22.18, gives

$$\frac{\partial F}{\partial q^b} = p_b - P_a \frac{\partial Q^a}{\partial q^b}\,, \tag{22.26}$$

$$\frac{\partial F}{\partial p_b} = -P_a \frac{\partial Q^a}{\partial p_b}\,, \tag{22.27}$$

together with $K = \mathcal{H} + P_a\,\partial Q^a/\partial t + \partial F/\partial t$, which plays no part here.

Now impose on F the equality of mixed second partial derivatives (Proposition 7.73). Differentiating Equation (22.26) with respect to p_c and Equation (22.27) with respect to q^b, the terms carrying second derivatives of Q^a cancel between the two — they are the same by Proposition 7.73 applied to Q^a — and what is left is

$$\delta_b^c = \frac{\partial Q^a}{\partial q^b}\frac{\partial P_a}{\partial p_c} - \frac{\partial Q^a}{\partial p_c}\frac{\partial P_a}{\partial q^b} = \left(A^{\mathsf{T}}D - C^{\mathsf{T}}B\right)_{bc}\,. \tag{22.28}$$

Differentiating Equation (22.26) with respect to q^c and comparing with the same expression with b and c exchanged gives, after the same cancellation, that $A^{\mathsf{T}}C$ is symmetric; doing the same with Equation (22.27) gives that $B^{\mathsf{T}}D$ is symmetric. Written out in blocks,

$$M^{\mathsf{T}} J\, M = \begin{pmatrix} A^{\mathsf{T}}C - C^{\mathsf{T}}A & A^{\mathsf{T}}D - C^{\mathsf{T}}B \\ B^{\mathsf{T}}C - D^{\mathsf{T}}A & B^{\mathsf{T}}D - D^{\mathsf{T}}B \end{pmatrix}\,, \tag{22.29}$$

whose four blocks are, in order, 0 by the symmetry of $A^{\mathsf{T}}C$, $\mathbb{1}_f$ by Equation (22.28), $-\mathbb{1}_f$ because that block is minus the transpose of the one above it, and 0 by the symmetry of $B^{\mathsf{T}}D$. That is Equation (22.25).

Sufficiency for a transformation free of the time. With $\zeta = \zeta(\eta)$ and $K(\zeta) = \mathcal{H}(\eta(\zeta))$, the chain rule gives $\partial\mathcal{H}/\partial\eta = M^{\mathsf{T}}\,\partial K/\partial\zeta$, so by the matrix form Equation (22.57) of Hamilton's equations, verified in Proposition 22.41 below,

$$\dot{\zeta} = M\dot{\eta} = MJ\,\frac{\partial\mathcal{H}}{\partial\eta} = MJM^{\mathsf{T}}\,\frac{\partial K}{\partial\zeta}\,.$$

Transposing and inverting Equation (22.25) gives $MJM^{\mathsf{T}} = J$ as well, so $\dot{\zeta} = J\,\partial K/\partial\zeta$: Hamilton's equations keep their form, for every $\mathcal{H}$. The corresponding statement for explicitly time-dependent transformations, and the geometric reading of Equation (22.25) as the preservation of a differential form, are Theorem 24.18. □

Proof. *Derives Proposition 22.17.*

By Lemma 22.20 a transformation is canonical only if $M^{\mathsf{T}}JM = J$, which since $J^{-1} = -J$ is the same as

$$M^{-1} = J^{-1}M^{\mathsf{T}}J = -J\,M^{\mathsf{T}}J = \begin{pmatrix} D^{\mathsf{T}} & -B^{\mathsf{T}} \\ -C^{\mathsf{T}} & A^{\mathsf{T}} \end{pmatrix}\,, \tag{22.30}$$

the last equality by multiplying out the blocks of Equation (22.24) with those of Equation (22.56). But M^{-1} is the Jacobian of the inverse transformation,

$$M^{-1} = \frac{\partial\,(q,p)}{\partial\,(Q,P)} = \begin{pmatrix} \partial q^b/\partial Q^a & \partial q^b/\partial P_a \\ \partial p_b/\partial Q^a & \partial p_b/\partial P_a \end{pmatrix}\,,$$

whose entries are exactly the right-hand sides of Equations (22.16) and (22.17). Reading Equation (22.30) block by block therefore gives $\partial p_b/\partial P_a = A_{ab} = \partial Q^a/\partial q^b$, $\partial q^b/\partial P_a =$

$-B_{ab} = -\partial Q^a/\partial p_b$, $\partial p_b/\partial Q^a = -C_{ab} = -\partial P_a/\partial q^b$ and $\partial q^b/\partial Q^a = D_{ab} = \partial P_a/\partial p_b$, which are the four conditions. Conversely, the four conditions assert Equation (22.30); multiplying by M on the left gives $\mathbb{1}_{2f} = -MJM^{\mathsf{T}}J$, and multiplying by $J^{-1} = -J$ on the right returns $MJM^{\mathsf{T}} = J$, hence Equation (22.25), and Lemma 22.20 closes the equivalence. □

Remark 22.21 (The count). The four families of Equations (22.16) and (22.17) are $4f^2$ scalar equations, but Equation (22.25) shows that they are not independent: $M^{\mathsf{T}}JM$ is antisymmetric for every M, so requiring it to equal J imposes only $f(2f-1)$ conditions. The dimension of $\mathrm{Sp}(2f,\mathbb{R})$ is accordingly $4f^2 - f(2f-1) = f(2f+1)$ (Proposition 5.134), which for one degree of freedom is 3: the canonical transformations of a two-dimensional phase space are, pointwise, the area- and orientation-preserving linear maps.

Rests on Propositions 5.134 and 22.17.

22.3 Invariant integrals

The integral invariants are due to Poincaré, who introduced them in volume III (1899) of *Les méthodes nouvelles de la mécanique céleste* [255].

Theorem 22.22 (Poincaré)**.** *Let S be a two-dimensional surface of phase space, parametrized by (u,v) over a domain D, so that*

$$I = \int_S \mathrm{d}q^a\,\mathrm{d}p_a := \int_D \sum_{a=1}^{f} \frac{\partial\,(q^a,p_a)}{\partial\,(u,v)}\,\mathrm{d}u\,\mathrm{d}v\,. \tag{22.31}$$

Then for a canonical transformation from (q,p) to (Q,P), applied at one instant,

$$I = \int_S \mathrm{d}q^a\,\mathrm{d}p_a = \int_S \mathrm{d}Q^a\,\mathrm{d}P_a\,. \tag{22.32}$$

Rests on Lemma 22.20 and Definition 7.93.

Proof. *Derives Theorem 22.22.*

Write η for the column of Equation (22.55) and ζ for the same column built from (Q,P). Each Jacobian determinant in Equation (22.31) is a difference of two products, and summing over a assembles them into the single matrix product

$$\sum_{a=1}^{f} \frac{\partial\,(q^a,p_a)}{\partial\,(u,v)} = \sum_{a=1}^{f} \left(\frac{\partial q^a}{\partial u}\frac{\partial p_a}{\partial v} - \frac{\partial q^a}{\partial v}\frac{\partial p_a}{\partial u} \right) = \left(\frac{\partial \eta}{\partial u}\right)^{\mathsf{T}} J \left(\frac{\partial \eta}{\partial v}\right), \tag{22.33}$$

as multiplying out the blocks of Equation (22.56) shows. By the chain rule Proposition 7.72, the same surface described in the new variables has $\partial\zeta/\partial u = M\,\partial\eta/\partial u$ and likewise for v, with M the Jacobian Equation (22.24). Hence

$$\left(\frac{\partial \zeta}{\partial u}\right)^{\mathsf{T}} J \left(\frac{\partial \zeta}{\partial v}\right) = \left(\frac{\partial \eta}{\partial u}\right)^{\mathsf{T}} \left(M^{\mathsf{T}} J\, M\right) \left(\frac{\partial \eta}{\partial v}\right) = \left(\frac{\partial \eta}{\partial u}\right)^{\mathsf{T}} J \left(\frac{\partial \eta}{\partial v}\right),$$

by Equation (22.25). The two integrands of Equation (22.32) agree at every point of D, so the integrals agree. □

Remark 22.23 (What the invariant does not depend on). Equation (22.31) defines I through a chosen parametrization of S, and the value is unchanged by any other regular parametrization of the same oriented surface: the integrand Equation (22.33) is bilinear and antisymmetric in $\partial\eta/\partial u$ and $\partial\eta/\partial v$, so a change of parameters multiplies it by the Jacobian determinant of that change, which the transformation of the area element $\mathrm{d}u\,\mathrm{d}v$ then cancels. The cancellation is the change-of-variables theorem for a multiple integral, which Chapter 7

does not state. It is quoted twice in this chapter and nowhere else — here, and for the transformation of the volume element in Corollary 22.24 — and in neither place is it needed for the statement itself: the two integrands of Theorem 22.22 are equal point by point over one and the same D, and Equation (22.34) is the assertion $\det M = 1$ about a determinant.

Rests on Theorem 22.22.

Corollary 22.24 (Invariance of the phase-space volume)**.** *A canonical transformation has* $\det M = 1$*, so the phase-space volume element is invariant,*

$$\mathrm{d}Q^1 \cdots \mathrm{d}Q^f \, \mathrm{d}P_1 \cdots \mathrm{d}P_f = \mathrm{d}q^1 \cdots \mathrm{d}q^f \, \mathrm{d}p_1 \cdots \mathrm{d}p_f \,. \tag{22.34}$$

In particular the flow generated by a Hamiltonian preserves phase-space volume, since the map carrying the state at time t_a to the state at time t_b is itself a canonical transformation.

Rests on Lemma 22.20 and Theorem 5.133.

Proof. *Derives Corollary 22.24.*

By Lemma 22.20 the Jacobian M lies in $\mathrm{Sp}(2f, \mathbb{R})$, and every element of the symplectic group has determinant $+1$ (Theorem 5.133); the volume element transforms by $|\det M|$, whence Equation (22.34).

For the flow, let $M(t)$ be the Jacobian of the map carrying the state at time t_a to the state at time t, and write $\mathcal{H}''_{cd} = \partial^2 \mathcal{H} / \partial \eta^c \partial \eta^d$, symmetric by Proposition 7.73. Differentiating Equation (22.57) with respect to the initial data gives $\dot{M} = J\,\mathcal{H}''M$, and therefore

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(M^{\mathsf{T}} J\, M\right) = \left(J\mathcal{H}''M\right)^{\mathsf{T}} J\, M + M^{\mathsf{T}} J \left(J\mathcal{H}''M\right) = M^{\mathsf{T}}\mathcal{H}'' \left(-J\right) J\, M + M^{\mathsf{T}} J^2 \mathcal{H}''M = 0\,,$$

using $J^{\mathsf{T}} = -J$, $J^2 = -\mathbb{1}_{2f}$ and the symmetry of $\mathcal{H}''$. At $t = t_a$ the map is the identity, for which $M^{\mathsf{T}} J M = J$ holds; hence it holds for all t, and the flow is a canonical transformation by Lemma 22.20. □

Remark 22.25 (This is Liouville's theorem)*.* Corollary 22.24 is Liouville's theorem in coordinates. Its coordinate-free form — that the Hamiltonian flow preserves the $2f$-form built from the symplectic structure, of which the invariance of the volume element is one consequence among several — is Theorem 24.21, and the physical use to which it is put, as the theorem that singles out the phase-space measure of statistical mechanics, is Remark 24.22. The chain Equations (22.32) and (22.34) continues: Poincaré's invariant is the first of a family of f integral invariants, of dimensions $2, 4, \ldots, 2f$, the last of which is the volume. They are the exterior powers of one object (Equation (24.13)).

Rests on Corollary 22.24 and Theorem 24.21.

22.3.1 The Lagrange bracket

Definition 22.26 (Lagrange bracket)**.** The *Lagrange bracket* of two phase-space functions u and v with respect to the coordinates (q, p) is

$$[u, v]_{q,p} = \frac{\partial q^a}{\partial u}\frac{\partial p_a}{\partial v} - \frac{\partial q^a}{\partial v}\frac{\partial p_a}{\partial u}\,, \tag{22.35}$$

the sum over a of the Jacobian determinants $\partial(q^a, p_a)/\partial(u, v)$.

Rests on Definition 22.14 and Proposition 7.72.

Comparing Equation (22.35) with Equation (22.33), the integrand of Poincaré's invariant *is* the Lagrange bracket of the two surface parameters, and in matrix form

$$[u, v]_{q,p} = \left(\frac{\partial \eta}{\partial u}\right)^{\mathsf{T}} J \left(\frac{\partial \eta}{\partial v}\right)\,. \tag{22.36}$$

From Equations (22.33) and (22.35) (the same sum of Jacobian determinants, written as a matrix product).

This is the form in which its properties are read off.

Proposition 22.27 (Properties of the Lagrange bracket)**.**

1. Invariance. *The Lagrange bracket is invariant under canonical transformations:*

$$[u,v]_{q,p} = [u,v]_{Q,P} \ . \tag{22.37}$$

2. Antisymmetry.

$$[u,v]_{q,p} = -\,[v,u]_{q,p} \ . \tag{22.38}$$

3. Fundamental Lagrange brackets.

$$\left[q^a, q^b\right]_{q,p} = 0\,, \qquad [p_a, p_b]_{q,p} = 0\,, \qquad [q^a, p_b]_{q,p} = \delta^a_b\,. \tag{22.39}$$

Rests on Equation (22.36) and Lemma 22.20.

Proof. *Derives Proposition 22.27.*

(1) Invariance. By Equation (22.36) written in the new variables and the chain rule $\partial\zeta/\partial u = M\,\partial\eta/\partial u$,

$$[u,v]_{Q,P} = \left(\frac{\partial\zeta}{\partial u}\right)^{\mathsf{T}} J \left(\frac{\partial\zeta}{\partial v}\right) = \left(\frac{\partial\eta}{\partial u}\right)^{\mathsf{T}} \left(M^{\mathsf{T}} J\, M\right) \left(\frac{\partial\eta}{\partial v}\right) = [u,v]_{q,p}\ ,$$

the last step by Equation (22.25). This is Equation (22.37), and it is the same computation as the one that proved Theorem 22.22 — as it must be, the invariant being the integral of the bracket.

(2) Antisymmetry. Exchanging u and v in Equation (22.35) exchanges the two products and so reverses the sign, which is Equation (22.38); equivalently, $J^{\mathsf{T}} = -J$ in Equation (22.36).

(3) Fundamental brackets. Take the parameters to be two of the phase-space coordinates themselves, so that the derivatives in Equation (22.35) are Kronecker deltas or zero: $\partial q^c/\partial q^a = \delta^c_a$, $\partial p_c/\partial q^a = 0$, $\partial q^c/\partial p_a = 0$, $\partial p_c/\partial p_a = \delta_{ca}$. Then

$$\begin{aligned}
\left[q^a, q^b\right]_{q,p} &= \sum_c \left(\delta^c_a \cdot 0 - \delta^c_b \cdot 0\right) = 0\,, \\
[p_a, p_b]_{q,p} &= \sum_c \left(0 \cdot \delta_{cb} - 0 \cdot \delta_{ca}\right) = 0\,, \\
[q^a, p_b]_{q,p} &= \sum_c \left(\delta^c_a \delta_{cb} - 0 \cdot 0\right) = \delta^a_b\,,
\end{aligned}$$

which is Equation (22.39). □

Remark 22.28 (Where the content of the fundamental brackets lies)*.* As computed, Equation (22.39) is a triviality: it says only that the coordinates are coordinates. The statement with content is the one obtained by combining it with Equation (22.37), namely

$$\left[q^a, q^b\right]_{Q,P} = 0\,, \qquad [p_a, p_b]_{Q,P} = 0\,, \qquad [q^a, p_b]_{Q,P} = \delta^a_b\,, \tag{22.40}$$

for a canonical transformation — and these $f(2f-1)$ independent equations are Equation (22.25) again, written in yet another notation. The same remark applies to the Poisson brackets of Equation (22.44) below, and is why Theorem 24.18 can list the three as equivalent tests for a transformation to be canonical.

Rests on Proposition 22.27 and Lemma 22.20.

22.3.2 The Poisson bracket

Definition 22.29 (Poisson bracket)**.** The *Poisson bracket* of two phase-space functions u and v with respect to the coordinates (q, p) is

$$\{u, v\}_{q,p} = \frac{\partial u}{\partial q^a}\frac{\partial v}{\partial p_a} - \frac{\partial u}{\partial p_a}\frac{\partial v}{\partial q^a}\,. \tag{22.41}$$

Rests on Definitions 7.65 and 22.14.

Where the Lagrange bracket differentiates the coordinates with respect to the functions, the Poisson bracket differentiates the functions with respect to the coordinates. In the matrix notation of Equation (22.36) the two read

$$[u, v]_{q,p} = \left(\frac{\partial \eta}{\partial u}\right)^{\mathsf{T}} J \left(\frac{\partial \eta}{\partial v}\right), \qquad \{u, v\}_{q,p} = \left(\frac{\partial u}{\partial \eta}\right)^{\mathsf{T}} J \left(\frac{\partial v}{\partial \eta}\right), \tag{22.42}$$

where $\partial u/\partial \eta$ is the column of the $2f$ partial derivatives of u. The contrast is the whole of Proposition 22.32 below.

Proposition 22.30 (Properties of the Poisson bracket)**.**

1. Antisymmetry.

$$\{u, v\}_{q,p} = -\{v, u\}_{q,p}\,. \tag{22.43}$$

2. Fundamental Poisson brackets.

$$\left\{q^a, q^b\right\}_{q,p} = 0\,, \qquad \{p_a, p_b\}_{q,p} = 0\,, \qquad \{q^a, p_b\}_{q,p} = \delta^a_b\,. \tag{22.44}$$

3. Invariance. *The Poisson bracket is invariant under canonical transformations:*

$$\{u, v\}_{q,p} = \{u, v\}_{Q,P}\,. \tag{22.45}$$

4. Jacobi identity. *For u, v, and w functions on phase space,*

$$\{u, \{v, w\}\} + \{v, \{w, u\}\} + \{w, \{u, v\}\} = 0\,. \tag{22.46}$$

Rests on Equation (22.42), Lemma 22.20, and Proposition 7.73.

Proof. *Derives Proposition 22.30.*

(1) Antisymmetry. Exchanging u and v in Equation (22.41) exchanges the two products, giving Equation (22.43).

(2) Fundamental brackets. As in the Lagrange case, take the two arguments to be coordinates, so the derivatives are Kronecker deltas or zero:

$$\begin{aligned}
\left\{q^a, q^b\right\}_{q,p} &= \sum_c \left(\delta^a_c \cdot 0 - 0 \cdot \delta^b_c\right) = 0\,,\\
\{p_a, p_b\}_{q,p} &= \sum_c \left(0 \cdot \delta_{cb} - \delta_{ca} \cdot 0\right) = 0\,,\\
\{q^a, p_b\}_{q,p} &= \sum_c \left(\delta^a_c \delta_{cb} - 0\right) = \delta^a_b\,,
\end{aligned}$$

which is Equation (22.44).

(3) Invariance. By the chain rule Proposition 7.72, a function's gradients in the two sets of variables are related by $\partial u/\partial \eta = M^{\mathsf{T}}\,\partial u/\partial \zeta$, so $\partial u/\partial \zeta = \left(M^{-1}\right)^{\mathsf{T}}\,\partial u/\partial \eta$. Hence, by Equation (22.42) in the new variables,

$$\{u, v\}_{Q,P} = \left(\frac{\partial u}{\partial \eta}\right)^{\mathsf{T}} M^{-1} J \left(M^{-1}\right)^{\mathsf{T}} \left(\frac{\partial v}{\partial \eta}\right) = \left(\frac{\partial u}{\partial \eta}\right)^{\mathsf{T}} J \left(\frac{\partial v}{\partial \eta}\right) = \{u, v\}_{q,p}\,,$$

because Equation (22.25) is equivalent to $MJM^{\mathsf{T}} = J$ and hence, multiplying by M^{-1} on the left and by $\left(M^{-1}\right)^{\mathsf{T}}$ on the right, to $M^{-1}J\left(M^{-1}\right)^{\mathsf{T}} = J$. This is Equation (22.45). The invariance is what allows the subscript on a Poisson bracket to be dropped, and it will be from here on.

(4) Jacobi identity. For fixed u the map $D_u : w \longmapsto \{u, w\}$ is, by Equation (22.41), a linear first-order differential operator in w: it involves the first derivatives of w and no higher ones. Now inspect the left-hand side of Equation (22.46), which we call Σ. Each of its three terms is a Poisson bracket one of whose arguments is itself a Poisson bracket, so each term is a sum of products of first derivatives of two of the functions with a *second* derivative of the third — exactly one second-derivative factor in every term.

Collect the terms carrying second derivatives of w. They can arise only from the first two terms of Σ, and by antisymmetry those two are

$$\{u, \{v, w\}\} + \{v, \{w, u\}\} = \{u, \{v, w\}\} - \{v, \{u, w\}\} = (D_u D_v - D_v D_u)\, w\,,$$

the commutator of two first-order operators. In such a commutator the second-order parts cancel: writing $D_u w = B^e \partial_e w$ and $D_v w = A^d \partial_d w$ with coefficients built from the first derivatives of u and v alone, the second-order part of $D_u D_v w$ is $B^e A^d\, \partial_e \partial_d w$ and that of $D_v D_u w$ is $A^e B^d\, \partial_e \partial_d w$, and the two contractions are equal because $\partial_e \partial_d w$ is symmetric in e and d (Proposition 7.73). So $(D_u D_v - D_v D_u)\, w$ is again first order in w. The third term $\{w, \{u, v\}\}$ carries only first derivatives of w to begin with. Hence Σ contains no second derivative of w at all. Since Σ is unchanged by cyclic permutation of u, v, w, it contains no second derivative of u and none of v either. But every term of Σ carries exactly one second derivative, of one of the three functions. Therefore every term cancels and $\Sigma = 0$, which is Equation (22.46). □

Remark 22.31 (The bracket is a Lie algebra structure). Properties (1) and (4) say precisely that the real vector space of smooth functions on phase space, equipped with $\{\cdot, \cdot\}$, is a Lie algebra in the sense of Definition 5.123 — an infinite-dimensional one. The bracket obeys in addition the Leibniz rule $\{u, vw\} = \{u, v\}\, w + v\, \{u, w\}$, immediate from Equation (22.41), so each $D_u = \{u, \cdot\}$ is a derivation of the algebra of observables (Definition 5.124). Both facts are structural rather than dynamical: no Hamiltonian has been named. It is because they hold that the correspondence with the commutator of quantum operators of Chapter 25 can be a correspondence of algebras and not merely an analogy of formulas.

Rests on Proposition 22.30 and Definitions 5.123 and 5.124.

Proposition 22.32 (Duality of the two brackets)**.** *Let $u_1, \dots, u_{2f}$, with $u_c = u_c(q^a, p_a)$, be a set of mutually independent functions on phase space. Then*

$$\sum_{c=1}^{2f} [u_c, u_d]_{q,p}\, \{u_c, u_e\}_{q,p} = \delta_{de}\,. \tag{22.47}$$

Writing $\Lambda_{cd} = [u_c, u_d]_{q,p}$ and $\Pi_{ce} = \{u_c, u_e\}_{q,p}$ for the two matrices, both antisymmetric, Equation (22.47) says $\Lambda^{\mathsf{T}}\Pi = \mathbb{1}_{2f}$; each matrix therefore determines the other, and

$$\Pi = \left(\Lambda^{\mathsf{T}}\right)^{-1} = -\Lambda^{-1}\,. \tag{22.48}$$

Rests on Equation (22.42) and Corollary A.292.

Proof. *Derives Proposition 22.32.*

Let T be the $2f \times 2f$ matrix of the derivatives $T^g{}_c = \partial \eta^g / \partial u_c$. The u_c being $2f$ mutually independent functions of the $2f$ variables η, the map $\eta \longmapsto u$ has a local inverse

(Corollary A.292) and T is invertible, with $\left(T^{-1}\right)_c{}^g = \partial u_c/\partial \eta^g$. The two matrices of Equation (22.42) are then

$$\Lambda = T^{\mathsf{T}} J\, T\,, \qquad \Pi = T^{-1} J \left(T^{-1}\right)^{\mathsf{T}}\,. \tag{22.49}$$

Multiplying them,

$$\Lambda\,\Pi = T^{\mathsf{T}} J\, T\, T^{-1} J \left(T^{\mathsf{T}}\right)^{-1} = T^{\mathsf{T}} J^2 \left(T^{\mathsf{T}}\right)^{-1} = -\mathbb{1}_{2f}\,,$$

using $J^2 = -\mathbb{1}_{2f}$ from Equation (5.173). Since $\Lambda^{\mathsf{T}} = -\Lambda$,

$$\sum_c \Lambda_{cd}\Pi_{ce} = \left(\Lambda^{\mathsf{T}}\Pi\right)_{de} = -\left(\Lambda\Pi\right)_{de} = \delta_{de}\,,$$

which is Equation (22.47), and Equation (22.48) follows on multiplying by $\left(\Lambda^{\mathsf{T}}\right)^{-1}$. □

Remark 22.33 (What the duality is good for). Neither bracket determines a transformation on its own, but each is a complete test. Taking the u_c to be the new coordinates (Q, P) themselves, Equation (22.49) reads $\Lambda = \left(M^{-1}\right)^{\mathsf{T}} J\, M^{-1}$ and $\Pi = MJM^{\mathsf{T}}$; the transformation is canonical when either equals J, and Equation (22.47) shows that one follows from the other. This is why the fundamental Lagrange brackets and the fundamental Poisson brackets are interchangeable as tests (Remark 22.28), a fact used without comment in most of the literature.

Rests on Proposition 22.32 and Lemma 22.20.

22.4 Motion in phase space

Theorem 22.34 (Evolution of a phase-space function)**.** *For an arbitrary phase-space function $U = U(q, p, t)$, along the motion,*

$$\frac{\mathrm{d}U}{\mathrm{d}t} = \{U, \mathcal{H}\}_{q,p} + \frac{\partial U}{\partial t}\,. \tag{22.50}$$

Rests on Equations (22.3), (22.4) and (22.41).

Proof. *Derives Theorem 22.34.*

By the chain rule Proposition 7.72 along a trajectory, and then by Hamilton's equations Equations (22.3) and (22.4),

$$\frac{\mathrm{d}U}{\mathrm{d}t} = \frac{\partial U}{\partial q^a}\dot{q}^a + \frac{\partial U}{\partial p_a}\dot{p}_a + \frac{\partial U}{\partial t} = \frac{\partial U}{\partial q^a}\frac{\partial \mathcal{H}}{\partial p_a} - \frac{\partial U}{\partial p_a}\frac{\partial \mathcal{H}}{\partial q^a} + \frac{\partial U}{\partial t}\,,$$

and the first two terms are Equation (22.41) with $v = \mathcal{H}$. □

Corollary 22.35. *If the function U does not depend explicitly on time, then*

$$\frac{\mathrm{d}U}{\mathrm{d}t} = \{U, \mathcal{H}\}_{q,p}\,. \tag{22.51}$$

Rests on Theorem 22.34.

Proof. *Derives Corollary 22.35.*

Set $\partial U/\partial t = 0$ in Equation (22.50). □

Remark 22.36 (The whole dynamics in one operation). Equation (22.50) is the most compact statement of Hamiltonian mechanics available: it computes the rate of change of *any* observable from the bracket alone, with $\mathcal{H}$ singled out only as the observable that generates the flow. Taking $U = \mathcal{H}$ recovers Equation (22.10), since $\{\mathcal{H}, \mathcal{H}\} = 0$ by Equation (22.43). It is also the classical statement of which the Heisenberg equation of motion is the quantum counterpart, $\{\cdot,\cdot\} \longmapsto [\cdot,\cdot]\,/\,(\mathrm{i}\hbar)$, and the correspondence — with the exact sense in which it holds and the precise sense in which it fails — is Chapter 25. The factor $\mathrm{i}\hbar$ is forced by units: by Remark 22.5 a Poisson bracket of two observables carries their units divided by an action, and $\hbar$ is the action that supplies it.

Rests on Theorem 22.34 and Remark 22.5.

In particular, the equations of motion themselves take the bracket form:

$$\{q^a, \mathcal{H}\}_{q,p} = \dot{q}^a = \frac{\partial \mathcal{H}}{\partial p_a}\,, \tag{22.52}$$

$$\{p_a, \mathcal{H}\}_{q,p} = \dot{p}_a = -\frac{\partial \mathcal{H}}{\partial q^a}\,. \tag{22.53}$$

From Equations (22.44) and (22.51) (taking $U = q^a$ and $U = p_a$, whose brackets with $\mathcal{H}$ reduce to a single term by Equation (22.44)).

Proposition 22.37 (Constants of motion)**.** *If the Poisson bracket of a function that does not depend explicitly on time with the Hamiltonian is zero, then the function is a constant of motion.*

Rests on Corollary 22.35.

Proof. *Derives Proposition 22.37.*

By Equation (22.51), $\mathrm{d}U/\mathrm{d}t = \{U, \mathcal{H}\} = 0$ along every motion, so U is constant on each trajectory (Corollary 7.36). The converse holds as well: a U free of explicit time dependence that is constant along every motion has $\{U, \mathcal{H}\} = 0$ everywhere, since the trajectory through a given point may be chosen to pass through it in any direction the equations allow. □

Theorem 22.38 (Poisson's theorem)**.** *The Poisson bracket of two constants of motion is itself a constant of motion.*

Rests on Equation (22.46) and Theorem 22.34.

Proof. *Derives Theorem 22.38.*

Let u and v be constants of motion, so that by Equation (22.50)

$$\{u, \mathcal{H}\} + \frac{\partial u}{\partial t} = 0\,, \qquad \{v, \mathcal{H}\} + \frac{\partial v}{\partial t} = 0\,. \tag{22.54}$$

Differentiating Equation (22.41) with respect to t at fixed (q, p) gives the Leibniz rule $\partial\,\{u, v\}\,/\partial t = \{\partial u/\partial t, v\} + \{u, \partial v/\partial t\}$. Next, the Jacobi identity Equation (22.46) with $w = \mathcal{H}$, rearranged by antisymmetry Equation (22.43), gives

$$\{\{u, v\}\,, \mathcal{H}\} = \{u, \{v, \mathcal{H}\}\} - \{v, \{u, \mathcal{H}\}\} = -\left\{u, \frac{\partial v}{\partial t}\right\} + \left\{v, \frac{\partial u}{\partial t}\right\}\,,$$

the last step by Equation (22.54). Adding the two results and applying Equation (22.50) to the function $\{u, v\}$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\{u, v\} &= \{\{u, v\}\,, \mathcal{H}\} + \frac{\partial}{\partial t}\{u, v\} \\ &= (-\{u, \partial v/\partial t\} + \{v, \partial u/\partial t\}) + (\{\partial u/\partial t, v\} + \{u, \partial v/\partial t\}) = 0\,,\end{aligned}$$

since $\{v, \partial u/\partial t\} = -\{\partial u/\partial t, v\}$. In particular, when neither constant depends explicitly on time the argument reduces to $\{\{u, v\}\,, \mathcal{H}\} = 0$ and Proposition 22.37. □

Remark 22.39 (Poisson's theorem does not manufacture new constants). The theorem is a genuine tool — the three components of angular momentum close on themselves under the bracket, and knowing two of them to be conserved delivers the third — but it is not a machine for producing integrals. Most often $\{u, v\}$ turns out to be a constant already known, a function of u and v, or simply zero. It says that the constants of motion of a system form a Lie subalgebra of the algebra of Remark 22.31, and the interest lies in which algebra that is: for the Kepler problem it is larger than the rotation algebra, which is the statement that the Laplace–Runge–Lenz vector is conserved (Chapter 27).

Rests on Theorem 22.38 and Remark 22.31.

22.5 Symplectic formulation

The pairs (q^a, p_a) may be assembled into a single object, and Hamilton's equations into a single matrix equation. Define the *coordinate vector* η^c, $c = 1, \dots, 2f$, and the corresponding *vector of derivatives* of the Hamiltonian by

$$\eta^c = \begin{pmatrix} q^a \\ p_a \end{pmatrix}, \qquad \frac{\partial \mathcal{H}}{\partial \eta^c} = \begin{pmatrix} \dfrac{\partial \mathcal{H}}{\partial q^a} \\ \dfrac{\partial \mathcal{H}}{\partial p_a} \end{pmatrix}, \tag{22.55}$$

where the first f components run over the coordinates and the last f over the momenta.

Definition 22.40 (Symplectic matrix)**.** The *symplectic matrix* is the $2f \times 2f$ matrix

$$J = \begin{pmatrix} 0 & \mathbb{1}_f \\ -\mathbb{1}_f & 0 \end{pmatrix}, \qquad J^{\mathsf{T}} = -J\,, \quad J^2 = -\mathbb{1}_{2f}\,, \quad J^{-1} = -J\,. \tag{22.56}$$

It is the matrix of Equation (5.173), the standard non-degenerate antisymmetric form of Proposition 5.115, at $2n = 2f$; the transformations preserving it are the symplectic group $\mathrm{Sp}(2f, \mathbb{R})$ of Definition 5.130.

Rests on Equation (5.173) and Definition 5.130.

Proposition 22.41 (Symplectic form of Hamilton's equations)**.** *Hamilton's equations Equations (22.3) and (22.4) are equivalent to*

$$\dot\eta^c = J^c{}_d \frac{\partial \mathcal{H}}{\partial \eta^d}\,, \qquad \textit{in matrix form} \qquad \dot\eta = J \frac{\partial \mathcal{H}}{\partial \eta}\,. \tag{22.57}$$

Rests on Equations (22.3), (22.4) and (22.56).

Proof. *Derives Proposition 22.41.*

Split the column Equation (22.55) into its two blocks of f components and multiply out, using Equation (22.56):

$$J \frac{\partial \mathcal{H}}{\partial \eta} = \begin{pmatrix} 0 & \mathbb{1}_f \\ -\mathbb{1}_f & 0 \end{pmatrix} \begin{pmatrix} \partial \mathcal{H}/\partial q^a \\ \partial \mathcal{H}/\partial p_a \end{pmatrix} = \begin{pmatrix} \partial \mathcal{H}/\partial p_a \\ -\partial \mathcal{H}/\partial q^a \end{pmatrix},$$

while $\dot\eta$ is the column with blocks $\dot q^a$ and $\dot p_a$. Equating the two columns block by block gives Equation (22.3) from the upper block and Equation (22.4) from the lower, and conversely. The whole effect of J is to interchange the two blocks and change the sign of one of them, which is exactly the asymmetry between Equations (22.3) and (22.4). □

Remark 22.42 (What the matrix form is for). Rewriting two equations as one buys nothing by itself. What it buys is that every statement about canonical transformations becomes a statement about the single matrix J: Equation (22.25) is the condition, Equation (22.42)

are the brackets, and the proofs in Sections 22.2, 22.3 and 22.3.2 above are three lines each because of it. The price is that η mixes quantities of different physical dimension — Remark 22.5 — so J is not a tensor on phase space in any metric sense and η is not a vector; the object that is coordinate-free is the two-form whose matrix J is, and that is the subject of Chapter 24.

Rests on Proposition 22.41 and Remark 22.5.

22.5.1 Symplectic form of the transformation condition

The heading reserved at this point in the source outline asks for the symplectic characterization of canonical transformations: the condition on the Jacobian matrix of the transformation that is equivalent to the $4f^2$ direct conditions of Proposition 22.17. It has already been established, because it was needed before the notation of this section was introduced. It is Lemma 22.20: the single matrix identity $M^{\mathsf{T}} J\, M = J$, stating that the Jacobian Equation (22.24) lies in $\mathrm{Sp}(2f, \mathbb{R})$ at every point and at every instant.

Four apparently different tests for a transformation to be canonical have now been shown to be one test, and it is worth listing them together.

1. The action principles agree, Equation (22.15) — the definition.
2. The direct conditions Equations (22.16) and (22.17) hold.
3. The Jacobian satisfies Equation (22.25).
4. The fundamental Lagrange brackets Equation (22.40), equivalently the fundamental Poisson brackets, are preserved.

That (1) implies (3) is the necessity half of Lemma 22.20; that (3) implies (1) for transformations free of the time is its sufficiency half, and in general Theorem 24.18; the equivalence of (2) and (3) is Proposition 22.17; and (4) is (3) read block by block, Remark 22.28. Only (3) survives unchanged into the coordinate-free language of Chapter 24, where it becomes the statement that the transformation pulls the symplectic form back to itself.

22.6 Beyond the canonical formalism

Three bodies of theory grow directly out of this chapter, and each is given a chapter of its own rather than a section here.

22.6.1 Hamilton–Jacobi theory

Pushed to its limit, the search for a convenient canonical transformation asks for one that makes every new coordinate and momentum constant. The generating function that achieves it satisfies a single first-order partial differential equation, and its level surfaces propagate through configuration space exactly as optical wavefronts propagate through a medium. Chapter 23 develops the equation, the action–angle variables it produces for bounded motion, and the optical–mechanical analogy that leads out of classical mechanics altogether.

22.6.2 Symplectic geometry

The matrix J of Equation (22.56) is the component form of a closed nondegenerate two-form on phase space, and the Poisson bracket, the canonical transformations, the invariant integrals of Section 22.3 and the conservation laws are all statements about it. Chapter 24 develops that geometry.

22.6.3 Symmetries

The symmetry content of the canonical formalism—the correspondence between one-parameter families of canonical transformations, their generators, and the constants of motion—follows from the bracket alone, and with it the correspondence rule that carries the whole formalism into quantum mechanics. Chapter 25 develops both, and Chapter 26 treats the systems for which the Legendre transformation of Definition 22.2 fails to be invertible in the first place—which is to say every gauge theory.

Chapter 23

Hamilton–Jacobi Theory and the Optical–Mechanical Analogy

The canonical transformations of Section 22.2 were introduced as the transformations that preserve the form of Hamilton's equations. Nothing so far has been asked of them beyond that. Hamilton–Jacobi theory asks the extreme question: is there a canonical transformation that makes the equations of motion *trivial*—one whose new coordinates and momenta are all constants? The answer is yes, and the generating function that performs it satisfies a single first-order partial differential equation, the Hamilton–Jacobi equation. Solving that one equation solves the mechanical problem completely.

The reward is not merely computational. The function that appears in the equation is the action itself, and its level surfaces propagate through configuration space exactly as the wavefronts of a wave propagate through an optical medium, with the trajectories playing the part of the rays. This is Hamilton's optical–mechanical analogy: he built the characteristic-function method for optics first [84] and then transplanted it whole into dynamics [96], and Jacobi turned the transplant into the method of integration that carries both their names [101]. It is the reason this chapter sits where it does: it is the last statement classical mechanics can make about itself, and the first sentence of quantum mechanics. The route from Equation (23.2) below to the Schrödinger equation of Chapter 77 is the shortest one in physics, and it is travelled in Section 23.5. Standard treatments are [1585, 2027].

Notation 23.1. The index conventions of Notation 22.1 remain in force: $a, b = 1, \dots, f$ label the degrees of freedom, and repeated indices are summed. The f constants of integration that label a complete integral are written α_a, and the f constants conjugate to them β^a. In the optical sections $\boldsymbol{x}$ denotes a point of physical space and s the arclength along a ray.

23.1 The Hamilton–Jacobi equation

23.1.1 The transformation to constant coordinates

Let $F_2 = S(q, P, t)$ be a generating function of the second kind, in the sense of Equation (22.21), and demand of it that the transformed Hamiltonian vanish identically,

$$K(Q, P, t) = 0\,. \tag{23.1}$$

Hamilton's equations for the new variables then read $\dot{Q}^a = 0$ and $\dot{P}_a = 0$: the new coordinates and momenta are $2f$ constants of the motion. Writing $P_a = \alpha_a$ and $Q^a = \beta^a$ for those constants, and substituting $p_a = \partial S/\partial q^a$ from Equation (22.21) into $K = \mathcal{H} + \partial S/\partial t$, the demand Equation (23.1) becomes a single equation for S.

Definition 23.2 (Hamilton's principal function and the Hamilton–Jacobi equation)**.** *Hamilton's principal function* is a function $S(q^a, \alpha_a, t)$ of the coordinates, of f constants α_a, and of the time, satisfying the *Hamilton–Jacobi equation*

$$\mathcal{H}\left(q^a, \frac{\partial S}{\partial q^a}, t\right) + \frac{\partial S}{\partial t} = 0\,. \tag{23.2}$$

It carries the SI dimension of an action, J s, since it generates the momenta by differentiation with respect to the coordinates.

Rests on Equations (22.21) *and* (23.1) *and Proposition 22.18.*

Equation (23.2) is a first-order partial differential equation in the $f+1$ variables (q^a, t), nonlinear in the derivatives whenever the Hamiltonian is not linear in the momenta. Its characteristic curves are precisely the trajectories of the mechanical system, which is the reading Remark 10.21 of Chapter 10 gives it.

Remark 23.3 (What Part II owes this chapter). Proposition 10.15 treats the *quasilinear* first-order equation $a\,\partial_x u + b\,\partial_y u = c$, in which the derivatives enter linearly. Equation (23.2) is not of that form: the Hamiltonian is quadratic in $\partial S/\partial q^a$ for every mechanical system of interest. The characteristic theory that does cover it—Cauchy's method of characteristic strips for a fully nonlinear equation $F(q, S, \partial S/\partial q) = 0$, in which the momenta are carried along as independent unknowns and the characteristic system is Hamilton's own—is not in Chapter 10, and neither is the theorem that the envelope of a family of solutions is again a solution, which is how a complete integral generates the general one. Both are ordinary theory of first-order partial differential equations and belong there. This chapter does not use them: Theorem 23.6 below is proved directly from the non-degeneracy of the complete integral, which needs only the implicit function theorem Theorem A.286, and the envelope construction is never invoked. The gap is recorded rather than quietly filled.

Rests on Equation (23.2)*, Proposition 10.15, and Theorem A.286.*

Remark 23.4 (What the equation costs and what it buys). The trade is $2f$ coupled ordinary differential equations for one partial differential equation in $f+1$ variables. Considered as a method of solution this is usually a bad bargain, and Hamilton–Jacobi theory is not the way most mechanical problems are solved. Its value is structural: it identifies the action as the object that propagates, it is the natural setting for separation of variables and for the action–angle variables of Section 23.3, and it is the classical limit of wave mechanics.

Rests on Equation (23.2) *and Theorem 22.6.*

23.1.2 Jacobi's theorem

Definition 23.5 (Complete integral)**.** A *complete integral* of Equation (23.2) is a solution $S(q^a, \alpha_a, t)$ depending on f constants α_a, none of which is purely additive, such that

$$\det\left(\frac{\partial^2 S}{\partial q^a\,\partial \alpha_b}\right) \neq 0\,. \tag{23.3}$$

Rests on Equation (23.2)*.*

The constants α_a are f, not $f+1$: Equation (23.2) contains S only through its derivatives, so an additive constant is always available and carries no information, which is what the exclusion of a purely additive constant says.

Theorem 23.6 (Jacobi)**.** *Let $S(q^a, \alpha_a, t)$ be a complete integral of the Hamilton–Jacobi equation Equation* (23.2)*. Then the $2f$ relations*

$$\beta^a = \frac{\partial S}{\partial \alpha_a}\,, \tag{23.4}$$

$$p_a = \frac{\partial S}{\partial q^a}\,, \tag{23.5}$$

with α_a and β^a arbitrary constants, define implicitly a $2f$-parameter family of curves $(q^a(t), p_a(t))$, and every such curve is a solution of Hamilton's equations Equations (22.3) and (22.4). Conversely every solution of Hamilton's equations is obtained in this way for a suitable choice of the constants.

Rests on Definition 23.5 and Theorems 22.6 and A.286.

Proof. Derives Theorem 23.6.

The curves exist. Fix α and β and apply the implicit function theorem Theorem A.286 to the map $F_b(t,q) := \partial S/\partial\alpha_b(q,\alpha,t) - \beta^b$, whose derivative with respect to q is the matrix $\partial^2 S/\partial\alpha_b\partial q^a$, nonsingular by Equation (23.3). Near any solution point Equation (23.4) therefore determines $q^a = q^a(\alpha,\beta,t)$ uniquely and differentiably, and Equation (23.5) then defines $p_a(t)$ along that curve.

The first half of Hamilton's equations. Differentiate Equation (23.4) totally with respect to t, holding α and β fixed:

$$0 = \frac{\partial^2 S}{\partial\alpha_b\,\partial t} + \frac{\partial^2 S}{\partial\alpha_b\,\partial q^a}\,\dot q^a\,. \tag{23.6}$$

Now differentiate the Hamilton–Jacobi equation Equation (23.2) itself with respect to α_b, at fixed q and t. The Hamiltonian depends on α_b only through its momentum slots, so the chain rule Proposition 7.72 gives

$$\frac{\partial\mathcal{H}}{\partial p_a}\,\frac{\partial^2 S}{\partial q^a\,\partial\alpha_b} + \frac{\partial^2 S}{\partial t\,\partial\alpha_b} = 0\,. \tag{23.7}$$

Mixed partial derivatives commute (Proposition 7.73), so subtracting Equation (23.7) from Equation (23.6) leaves

$$\frac{\partial^2 S}{\partial\alpha_b\,\partial q^a}\left(\dot q^a - \frac{\partial\mathcal{H}}{\partial p_a}\right) = 0\,, \tag{23.8}$$

and the matrix is nonsingular by Equation (23.3). Hence $\dot q^a = \partial\mathcal{H}/\partial p_a$, which is Equation (22.3).

The second half. Differentiate Equation (23.5) totally with respect to t:

$$\dot p_a = \frac{\partial^2 S}{\partial q^a\,\partial t} + \frac{\partial^2 S}{\partial q^a\,\partial q^b}\,\dot q^b\,. \tag{23.9}$$

Differentiate Equation (23.2) with respect to q^a, now at fixed α and t, remembering that $\mathcal{H}$ depends on q^a both directly and through the momenta:

$$\frac{\partial\mathcal{H}}{\partial q^a} + \frac{\partial\mathcal{H}}{\partial p_b}\,\frac{\partial^2 S}{\partial q^b\,\partial q^a} + \frac{\partial^2 S}{\partial t\,\partial q^a} = 0\,. \tag{23.10}$$

Substituting $\dot q^b = \partial\mathcal{H}/\partial p_b$, just proved, into Equation (23.9) and adding Equation (23.10) term by term, every second derivative of S cancels and

$$\dot p_a + \frac{\partial\mathcal{H}}{\partial q^a} = 0\,, \tag{23.11}$$

which is Equation (22.4).

The converse. Let $(q^a(t), p_a(t))$ solve Hamilton's equations with data (q_0, p_0) at t_0. By Equation (23.3) and Theorem A.286 again, the f equations $p_{0a} = \partial S/\partial q^a(q_0,\alpha,t_0)$ determine α near any point at which they hold; set $\beta^b := \partial S/\partial\alpha_b(q_0,\alpha,t_0)$. The curve the theorem produces from that (α,β) and the given solution obey the same first-order system with the same data at t_0, so they coincide by the uniqueness half of Theorem 9.8. □

The theorem is what makes the trade of Remark 23.4 worth considering: *one* solution of Equation (23.2) carrying f constants, and no other information, yields the complete solution of the mechanical problem by differentiation and elimination alone—no further integration is performed.

Proposition 23.7 (The principal function is the action)**.** *Along a solution of the equations of motion,*

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mathcal{L}\,, \qquad \textit{so that} \qquad S = \int_{t_a}^{t_b} \mathcal{L}\,\mathrm{d}t + \textit{const}\,. \tag{23.12}$$

Hamilton's principal function is therefore the action of Chapter 16, regarded no longer as a functional of paths but as a function of the endpoint.

Rests on Theorem 23.6, Equation (22.1)*, and Definition 16.96.*

Proof. *Derives Proposition 23.7.*

Along a solution the constants α_a are fixed, so the total time derivative of $S\,(q(t), \alpha, t)$ is

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\partial S}{\partial q^a}\,\dot{q}^a + \frac{\partial S}{\partial t} = p_a \dot{q}^a - \mathcal{H}\,, \tag{23.13}$$

From Equations (23.2) *and* (23.5) *(the first term from the momenta of the complete integral, the second from the equation itself).*

The second equality uses Equation (23.5) for the first term and Equation (23.2) for the second. The right-hand side is the inverse Legendre transformation Equation (22.1) of Chapter 22, that is the Lagrangian. Integrating between two times gives the second equality of Equation (23.12), the constant being the value of S at the lower limit. □

Remark 23.8 (The action as a function, not a functional). Proposition 23.7 is the pivot of the whole chapter and it is easy to read past. In Chapter 16 the action Equation (16.86) is a number assigned to an arbitrary path; here it is a *field* on configuration space, obtained by evaluating that number on the one extremal path that actually reaches the point q at time t. The two readings are related exactly as Hilbert's invariant integral Theorem 16.63 is related to the functional it is built from: given a field of extremals in the sense of Definition 16.62, the value of the action along the member through each point defines a function whose gradient is the momentum, and Equation (23.5) is that statement.

Rests on Proposition 23.7, Theorem 16.63, and Definition 16.62.

23.1.3 Time-independent systems and the characteristic function

Proposition 23.9 (Separation of the time)**.** *If the Hamiltonian does not depend explicitly on the time, the Hamilton–Jacobi equation admits solutions of the form*

$$S(q, \alpha, t) = W(q, \alpha) - Et\,, \tag{23.14}$$

where $E = \alpha_1$ is the conserved value of the Hamiltonian and the function W, called Hamilton's characteristic function*, satisfies the* time-independent Hamilton–Jacobi equation

$$\mathcal{H}\left(q^a, \frac{\partial W}{\partial q^a}\right) = E\,. \tag{23.15}$$

Rests on Equation (23.2)*, Theorem 22.10, and Lemma 10.28.*

Proof. *Derives Proposition 23.9.*

Try the additive ansatz $S(q, \alpha, t) = W(q, \alpha) - h(t)$, which is the general form in which the time enters additively. Then $\partial S/\partial q^a = \partial W/\partial q^a$ and $\partial S/\partial t = -h'(t)$, so Equation (23.2) reads

$$\mathcal{H}\left(q^a, \frac{\partial W}{\partial q^a}\right) = h'(t)\,. \tag{23.16}$$

The left-hand side depends on q alone—this is where the absence of an explicit time dependence in $\mathcal{H}$ is used—and the right-hand side on t alone, so by Lemma 10.28 both equal a constant. Call it E; then $h(t) = Et$ up to the additive constant excluded by Definition 23.5, and Equation (23.16) is Equation (23.15). That the constant is the conserved value of the Hamiltonian is immediate from Equation (23.15) and Equation (23.5), and the conservation itself is Theorem 22.10. Since E is one of the constants of the complete integral it may be named α_1. □

Proposition 23.10 (The characteristic function is the abbreviated action)**.** *The characteristic function is the abbreviated, or Maupertuis, action*

$$W = \int p_a \, \mathrm{d}q^a \,, \tag{23.17}$$

and the variational principle it satisfies at fixed energy,

$$\delta \int p_a \, \mathrm{d}q^a = 0 \,, \tag{23.18}$$

is Maupertuis' principle*. For a Hamiltonian of the form* $\mathcal{H} = |\boldsymbol{p}|^2 / 2m + V(\boldsymbol{x})$ *it reads*

$$\delta \int \sqrt{2m \left(E - V(\boldsymbol{x})\right)} \, \mathrm{d}s = 0 \,, \tag{23.19}$$

$\mathrm{d}s$ *being the element of arclength in configuration space.*
Rests on Proposition 23.9, Definition 16.99, and Theorem 16.101.

Proof. *Derives Proposition 23.10.*
By Equation (23.14) and Proposition 23.7,

$$W = S + Et = \int (\mathcal{L} + E) \, \mathrm{d}t = \int (p_a \dot{q}^a - \mathcal{H} + E) \, \mathrm{d}t = \int p_a \dot{q}^a \, \mathrm{d}t \,, \tag{23.20}$$

since $\mathcal{H} = E$ along the motion; the last integral is $\int p_a \, \mathrm{d}q^a$, which is Equation (23.17) and is exactly the abbreviated action Equation (16.88) of Definition 16.99. The variational principle Equation (23.18) for that functional, with the energy fixed and the transit time free, is Theorem 16.101, which also supplies the metric form: for $\mathcal{H} = |\boldsymbol{p}|^2 / 2m + V$ one has $|\boldsymbol{p}| = \sqrt{2m(E - V)}$ and $p_a \mathrm{d}q^a = |\boldsymbol{p}| \, \mathrm{d}s$, so Equation (23.18) becomes Equation (16.91), which is Equation (23.19). □

Remark 23.11 (Whose principle, and where it fails). Two cautions carried over from Remark 16.102 and not to be lost here. The sharp statement of Equation (23.18), with the hypothesis of fixed energy, is Euler's [23], not Maupertuis' [24]; the name is customary and is kept, the correction recorded. And Equation (23.19) requires $E > V$ throughout the region: at a turning point the integrand vanishes, the Jacobi metric degenerates, and the principle says nothing. That degeneracy is not a technicality—it is the same failure that Section 23.5.2 identifies as the place where the whole geometrical description breaks down.
Rests on Proposition 23.10 and Theorem 16.101.

Remark 23.12 (Two functions, two roles). S generates a transformation to variables that are constant; W generates a transformation to variables in which the new Hamiltonian is E, so that the new momenta are constant and the new coordinates advance linearly in time. The second is the useful one for bounded motion, and it is what Section 23.3 builds on.
Rests on Definition 23.2 and Proposition 23.9.

23.2 Complete integrals and separation of variables

Definition 23.13 (Additive separability)**.** The Hamilton–Jacobi equation is *completely separable* in the coordinates q^a if it admits a complete integral of the form

$$W(q, \alpha) = \sum_{a=1}^{f} W_a \left(q^a; \alpha_1, \ldots, \alpha_f\right) \,, \tag{23.21}$$

each summand depending on a single coordinate.
Rests on Definition 23.5 and Equation (23.15).

Separability turns one partial differential equation into f ordinary ones, each solvable by a quadrature: from Equation (23.21) the momentum $p_a = \mathrm{d}W_a/\mathrm{d}q^a$ is a function of its own coordinate and of the constants, so each W_a is an explicit integral. This is why every mechanical problem that can be "solved in closed form" is, in practice, a separable one.

Proposition 23.14 (Cyclic coordinates)**.** *If q^1 is cyclic—that is, absent from the Hamiltonian, in the sense of Definition 21.35—then the characteristic function separates in it as*

$$W = \alpha_1 q^1 + W'\left(q^2, \dots, q^f; \alpha\right), \tag{23.22}$$

and $\alpha_1 = p_1$ is the conserved momentum conjugate to q^1.

Rests on Definitions 21.35 and 23.13 and Theorem 21.36.

Proof. *Derives Proposition 23.14.*

Substitute Equation (23.22) into Equation (23.15). Since $\partial W/\partial q^1 = \alpha_1$ and q^1 occurs nowhere in $\mathcal{H}$, the resulting equation

$$\mathcal{H}\left(q^2, \dots, q^f;\ \alpha_1, \frac{\partial W'}{\partial q^2}, \dots, \frac{\partial W'}{\partial q^f}\right) = E$$

contains neither q^1 nor any derivative of W' with respect to it: it is an equation for W' in $f-1$ variables, and the ansatz is consistent. Conversely Equation (23.5) gives $p_1 = \partial W/\partial q^1 = \alpha_1$, a constant, in agreement with Theorem 21.36. □

23.2.1 Conditions for separability

Separability is a property of the pair (system, coordinate system), never of the system alone. The Kepler problem separates in spherical coordinates, in parabolic coordinates and in a rotated Cartesian system not at all; the harmonic oscillator separates in Cartesian coordinates and not in polar ones unless it is isotropic. The question this subsection answers is which coordinate systems work, for the class of Hamiltonians in which the kinetic energy is diagonal.

Definition 23.15 (Orthogonal Hamiltonian)**.** A Hamiltonian is *orthogonal* in the coordinates q^a if

$$\mathcal{H}(q,p) = \frac{1}{2}\sum_{a=1}^{f} g^{aa}(q)\, p_a^2 + V(q)\,, \tag{23.23}$$

with no cross terms $p_a p_b$, $a \neq b$. For a single particle of mass m moving in ordinary space with orthogonal curvilinear coordinates of scale factors h_a, so that $\mathrm{d}s^2 = \sum_a h_a^2\,(\mathrm{d}q^a)^2$, one has $g^{aa} = 1/\left(m h_a^2\right)$. The SI dimension of g^{aa} depends on what dimension the coordinate q^a carries—$1/\mathrm{kg}$ when q^a is a length, $1/(\mathrm{kg\,m^2})$ when it is an angle—and in every case $g^{aa}p_a^2$ is an energy.

Rests on Definitions 13.117 and 22.2.

Definition 23.16 (Stäckel matrix; Stäckel system)**.** A *Stäckel matrix* for the coordinates q^a is an $f \times f$ matrix $\Phi = (\phi_{ab})$ whose a-th row depends on the single coordinate q^a, with

$$\det \Phi \neq 0\,. \tag{23.24}$$

The orthogonal Hamiltonian Equation (23.23) is a *Stäckel system* in those coordinates if there exist such a Φ and f functions $u_a(q^a)$, each of one coordinate, with

$$g^{aa} = \left(\Phi^{-1}\right)_{1a}, \qquad V = \sum_{a=1}^{f} \left(\Phi^{-1}\right)_{1a} u_a(q^a)\,. \tag{23.25}$$

Rests on Equation (23.23).

Equation (23.25) is a strong pair of requirements and it is worth reading twice. The first says the diagonal of the inverse metric is the *first row of the inverse* of a matrix built out of one-dimensional functions; it constrains the coordinate system alone. The second says the potential is a combination of one-dimensional potentials with exactly those coefficients; it constrains the system. Dimensionally, ϕ_{a1} carries the reciprocal dimension of g^{aa} and u_a the dimension of p_a^2, so that every term of Equation (23.27) below is a squared momentum.

Theorem 23.17 (Stäckel's conditions are sufficient)**.** *If Equation* (23.25) *holds, then Equation* (23.15) *separates in the coordinates* q^a *and*

$$W(q,\alpha) = \sum_{a=1}^{f} \int^{q^a} \sqrt{2\sum_{b=1}^{f} \alpha_b\, \phi_{ab}(s) - 2u_a(s)}\; \mathrm{d}s \tag{23.26}$$

is a complete integral of it, with $\alpha_1 = E$*.*

Rests on Definitions 23.13 and 23.16 and Equation (23.15)*.*

Proof. *Derives Theorem 23.17.*

The summand of Equation (23.26) depends on the single coordinate q^a, so W has the separated form Equation (23.21) and

$$\left(\frac{\mathrm{d}W_a}{\mathrm{d}q^a}\right)^2 = 2\sum_{b=1}^{f} \alpha_b\, \phi_{ab}(q^a) - 2u_a(q^a)\,. \tag{23.27}$$

Insert this and Equation (23.25) into Equation (23.23). The potential terms cancel against the $-2u_a$ and

$$\mathcal{H} = \sum_a \left(\Phi^{-1}\right)_{1a} \sum_b \alpha_b\, \phi_{ab} = \sum_b \alpha_b \sum_a \left(\Phi^{-1}\right)_{1a} \phi_{ab} = \sum_b \alpha_b\, \delta_{1b} = \alpha_1\,, \tag{23.28}$$

From Equations (23.25) *and* (23.27) *(the first row of the inverse matrix, times the matrix itself).*

The middle step is nothing but $\Phi^{-1}\Phi = \mathbb{1}$ read along its first row. So Equation (23.15) holds with $E = \alpha_1$, identically in q and in the remaining constants.

It remains to check the non-degeneracy Equation (23.3). Differentiating Equation (23.27) with respect to α_b gives $\partial^2 W/\partial q^a \partial \alpha_b = \phi_{ab}/W_a'$, that is the matrix Φ with its a-th row divided by W_a'. Hence

$$\det\left(\frac{\partial^2 W}{\partial q^a\, \partial \alpha_b}\right) = \frac{\det \Phi}{\prod_a W_a'}\,, \tag{23.29}$$

which is nonzero wherever no W_a' vanishes, by Equation (23.24). The zeros of W_a' are exactly the turning points of the a-th degree of freedom, and they are precisely where the construction is expected to fail; that failure is the subject of Section 23.5.2. □

Remark 23.18 (The converse, and what this treatise does not prove)*.* The implication of Theorem 23.17 runs the other way as well: if the orthogonal Hamiltonian Equation (23.23) admits a complete integral of the separated form Equation (23.21) in the coordinates q^a, then a Stäckel matrix and functions u_a satisfying Equation (23.25) exist, so that separability of an orthogonal Hamiltonian is *equivalent* to the Stäckel conditions. That converse is Stäckel's own theorem in the form Levi-Civita gave it, and it is stated here on the authority of [2027]. This treatise does not prove it, and nothing in this chapter or later in the book uses it: every application runs through Theorem 23.17, in the direction that is proved.

It is worth saying where the difficulty sits, because the first two steps are easy and mislead. Write $x_a := \frac{1}{2}\,(W_a')^2$, a function of the single coordinate q^a and of the constants. Differentiating the separated equation $\sum_a g^{aa} x_a + V = \alpha_1$ with respect to α_b gives $\sum_a g^{aa}\phi_{ab} = \delta_{1b}$ with $\phi_{ab} := \partial x_a/\partial \alpha_b$, and Equation (23.29) identifies $\det\Phi \neq 0$ with the

completeness Equation (23.3); so $g^{aa} = (\Phi^{-1})_{1a}$ already. What is *not* immediate is that Φ may be taken independent of the constants—that is, that x_a is affine in α, which is exactly what makes ϕ_{ab} a function of q^a alone and $u_a := \sum_b \alpha_b \phi_{ab} - x_a$ a function of q^a alone. A second differentiation returns only $\sum_a g^{aa}\, \partial\phi_{ab}/\partial\alpha_c = 0$, which says no more than that g^{aa} does not depend on the constants, and admits solutions with $\partial\Phi/\partial\alpha \neq 0$. Removing those requires a change of the separation constants $\alpha \mapsto \tilde{\alpha}(\alpha)$, under which $\Phi \mapsto \Phi\, \partial\alpha/\partial\tilde{\alpha}$, and it is the proof that such a change always exists—not the index bookkeeping—that this treatise leaves to its source.

Rests on Theorem 23.17 and Definition 23.16.

Remark 23.19 (The same conditions govern the Helmholtz equation). Theorem 10.34 of Chapter 10 states that the Helmholtz equation separates in exactly eleven orthogonal coordinate systems of ordinary space, and that the criterion is the Stäckel condition on the scale factors; the derivation is Appendix A.15, and Remark 10.35 there points forward to this chapter. The two statements are not the same statement, and the difference is instructive. The Helmholtz equation is linear and of second order and its separation constants are eigenvalues; Equation (23.15) is nonlinear and of first order and its separation constants are values of conserved quantities. What they share is Equation (23.25)'s first half, which involves only the coordinate system: the eleven systems that separate the Helmholtz equation are exactly the eleven that separate the free-particle Hamilton–Jacobi equation, because for $V = 0$ the second half of Equation (23.25) is vacuous. Adding a potential can only restrict the list further, never extend it. The relation between the two is made precise by the semiclassical expansion of Section 23.5.1, in which the first-order equation is the leading order of the second-order one.

Rests on Theorems 10.34 and 23.17.

Proposition 23.20 (Separability implies integrability)**.** *Let Equation* (23.15) *separate in the coordinates q^a, with complete integral Equation* (23.21)*. Then the f functions $\alpha_a(q,p)$ obtained by solving Equation* (23.5) *for the constants are functionally independent constants of the motion in involution,*

$$\{\alpha_a, \alpha_b\} = 0 \qquad \text{for all } a, b\,, \tag{23.30}$$

so a separable system is completely integrable in the sense of Theorem 23.31. For a Stäckel system they are

$$\alpha_b = \sum_{a=1}^{f} \left(\Phi^{-1}\right)_{ba} \left[\tfrac{1}{2}p_a^2 + u_a(q^a)\right]\,, \tag{23.31}$$

each quadratic in the momenta, and $\alpha_1 = \mathcal{H}$.

Rests on Definition 23.13 and Equations (22.44) *and* (22.45).

Proof. *Derives Proposition 23.20.*

$W(q,\alpha)$ is a generating function of the second kind (Proposition 22.18) with old momenta $p_a = \partial W/\partial q^a$ and new momenta $P_a = \alpha_a$; the non-degeneracy Equation (23.3) is exactly what makes the transformation invertible, so the α_a are genuine functions on phase space and are functionally independent. They are constants of the motion by Theorem 23.6. The Poisson bracket is invariant under a canonical transformation, Equation (22.45), so it may be computed in the new variables, where the fundamental relations Equation (22.44) give $\{P_a, P_b\}_{Q,P} = 0$; that is Equation (23.30).

For the explicit form, Equation (23.27) with $p_a = W_a'$ reads $\frac{1}{2}p_a^2 + u_a(q^a) = \sum_b \alpha_b \phi_{ab}(q^a)$; multiplying by $(\Phi^{-1})_{ca}$ and summing over a inverts it into Equation (23.31). Setting $c = 1$ and using Equation (23.25) recovers Equation (23.23), so $\alpha_1 = \mathcal{H}$. □

Remark 23.21 (The tensorial statement, and what Part II owes it). Equation (23.31) exhibits each separation constant as a quadratic form in the momenta,

$$\alpha_b = \tfrac{1}{2} K_b{}^{cd}(q)\, p_c p_d + U_b(q)\,, \qquad K_b{}^{cd} = \left(\Phi^{-1}\right)_{bc} \delta^{cd}\,. \tag{23.32}$$

That such a quantity is conserved along every trajectory of Equation (23.23) with $V = 0$ is, in the language of Chapter 13, the statement that $K_b{}^{cd}$ is a *Killing tensor* of rank two: a symmetric tensor field obeying $\nabla^{(e} K_b{}^{cd)} = 0$, the rank-two analogue of Killing's equation Equation (13.290). Separability is then the geometric condition that the configuration-space metric admit f commuting Killing tensors with simultaneously diagonalisable, pointwise-distinct eigenvectors—the coordinate system being the one whose surfaces those eigenvectors are normal to.

Part II does not carry this. Section 13.9.2 defines Killing *vectors* Definition 13.139 and derives Killing's equation Proposition 13.140, and stops there: there is no rank-two Killing tensor, no statement that $\xi_\mu p^\mu$ or $K_{\mu\nu} p^\mu p^\nu$ is constant along a geodesic, and no integrability condition for either. The definitions above are therefore stated here for use and are *not* treated as established; every result this chapter actually proves—Theorem 23.17, Proposition 23.20—is proved from Equation (23.25) directly, in coordinates, and needs none of it. The tensorial reading is what makes Carter's constant of Section 23.4.3 intelligible, and that is where the absence is felt.

Rests on Propositions 13.140 and 23.20.

23.2.2 Worked separations

Six complete integrals, in the order of increasing structure. Each ends by recovering, from Theorem 23.6 alone, a result this treatise has already obtained by other means—which is the only honest way to test a method of this generality.

Example 23.22 (Free particle)**.** For $\mathcal{H} = |\boldsymbol{p}|^2/2m$ in Cartesian coordinates all three coordinates are cyclic, so Proposition 23.14 applies three times and

$$S = \boldsymbol{\alpha}\cdot\boldsymbol{x} - \frac{|\boldsymbol{\alpha}|^2}{2m}\, t\,, \tag{23.33}$$

with $\boldsymbol{\alpha}$ in kg m/s and $E = |\boldsymbol{\alpha}|^2/2m$. Equation (23.4) gives $\boldsymbol{\beta} = \partial S/\partial\boldsymbol{\alpha} = \boldsymbol{x} - \boldsymbol{\alpha} t/m$, that is

$$\boldsymbol{x}(t) = \boldsymbol{\beta} + \frac{\boldsymbol{\alpha}}{m}\, t\,, \tag{23.34}$$

uniform rectilinear motion with $\boldsymbol{p} = \boldsymbol{\alpha}$ constant by Equation (23.5). The surfaces $S = \text{const}$ are the planes perpendicular to $\boldsymbol{\alpha}$, advancing at speed $E/|\boldsymbol{\alpha}| = |\boldsymbol{\alpha}|/2m$—which is *half* the particle speed. That factor of two is not an error and is taken up in Remark 23.42.

Rests on Proposition 23.14 and Theorem 23.6.

Example 23.23 (Uniform gravity, and the parabola)**.** Take motion in a vertical plane with $\mathcal{H} = (p_x^2 + p_z^2)/2m + mgz$ and $g = 9.806\,65\,\mathrm{m/s^2}$, the standard value fixed in Chapter 20. Here x is cyclic, so by Propositions 23.9 and 23.14

$$W = \alpha_x x + \int^z \sqrt{2m\left(E - \frac{\alpha_x^2}{2m} - mgz'\right)}\,\mathrm{d}z'\,. \tag{23.35}$$

Write $\alpha_z := E - \alpha_x^2/2m$ for the energy of the vertical motion. Equation (23.4) with respect to α_x and to α_z gives

$$x = \frac{\alpha_x}{m}\,(t - t_0) + x_0\,, \qquad z = z_0 + \frac{\sqrt{2m\alpha_z}}{m}\,(t - t_0) - \frac{g}{2}\,(t - t_0)^2\,, \tag{23.36}$$

which is the parabolic trajectory derived from Newton's second law in Chapter 20, with the initial speeds $u_x = \alpha_x/m$ and $u_z = \sqrt{2\alpha_z/m}$ identified as the components of the launch velocity. The wavefronts $W = \text{const}$ are the curves orthogonal to the family of parabolas of fixed energy—the envelope of that family being the safety parabola, a caustic in the sense of Section 23.5.2.

Rests on Proposition 23.14 and Theorem 23.6.

Example 23.24 (Harmonic oscillator)**.** For $\mathcal{H} = p^2/2m + \tfrac{1}{2}m\omega_0^2 x^2$ (Definition 28.3), Equation (23.15) gives

$$W(x,E) = \int^{x} \sqrt{2mE - m^2\omega_0^2 x'^2}\,\mathrm{d}x' \,, \tag{23.37}$$

whose $\beta = \partial W/\partial E - t$ reduces by the substitution $x = \sqrt{2E/m\omega_0^2}\,\sin\varphi$ to $\arcsin\left(x\sqrt{m\omega_0^2/2E}\right) = \omega_0(t+\beta)$, the solution Equation (28.4). The action variable Equation (23.49) is the phase-plane area of Proposition 28.6 divided by 2π,

$$I = \frac{1}{2\pi}\oint p\,\mathrm{d}x = \frac{E}{\omega_0} \,, \tag{23.38}$$

in J s, so the Hamiltonian in action–angle variables is simply $\mathcal{H}(I) = \omega_0 I$ and Equation (23.52) returns $\dot\theta = \omega_0$: the frequency, read off without solving anything. This is the doorway of Remark 28.7.

Rests on Proposition 28.6, Definition 23.28, and Theorem 23.30.

Example 23.25 (Central force in plane polar coordinates)**.** With $\mathcal{H} = \left(p_r^2 + p_\varphi^2/r^2\right)/2m + V(r)$ the coordinate φ is cyclic, so

$$W = \alpha_\varphi\varphi + \int^{r} \sqrt{2m\,(E - V(r')) - \frac{\alpha_\varphi^2}{r'^2}}\;\mathrm{d}r' \,, \tag{23.39}$$

with $\alpha_\varphi = p_\varphi = L$ the angular momentum, in J s. The two relations Equation (23.4) are

$$t + \beta_E = \int^{r} \frac{m\,\mathrm{d}r'}{\sqrt{2m\,(E-V) - L^2/r'^2}} \,, \tag{23.40}$$

$$\varphi - \beta_\varphi = \int^{r} \frac{L\,\mathrm{d}r'}{r'^2\sqrt{2m\,(E-V) - L^2/r'^2}} \,. \tag{23.41}$$

Equation (23.40) is the radial quadrature of Section 27.3, with $\sqrt{2m(E-V) - L^2/r^2}/m$ the radial velocity Equation (27.34) and the bracket the effective potential Equation (27.32); Equation (23.41) is the orbit equation Equation (27.38). Nothing new is obtained, and that is the point: the two quadratures that Chapter 27 reaches by eliminating the time between two conservation laws fall out here as the two components of a single differentiation.

Rests on Proposition 23.14, Theorem 23.6, and Equation (27.38).

Example 23.26 (The Kepler problem, and the Bohr formula)**.** In spherical coordinates with $V = -k/r$ and $k = e^2/4\pi\varepsilon_0$ in J m the Hamiltonian is $\mathcal{H} = \left[p_r^2 + p_\theta^2/r^2 + p_\varphi^2/(r^2\sin^2\theta)\right]/2m - k/r$, and $W = \alpha_\varphi\varphi + W_\theta(\theta) + W_r(r)$ separates it into

$$\left(\frac{\mathrm{d}W_\theta}{\mathrm{d}\theta}\right)^2 + \frac{\alpha_\varphi^2}{\sin^2\theta} = \alpha_\theta^2 \,, \qquad \left(\frac{\mathrm{d}W_r}{\mathrm{d}r}\right)^2 = 2m\left(E + \frac{k}{r}\right) - \frac{\alpha_\theta^2}{r^2} \,, \tag{23.42}$$

$\alpha_\theta = L$ being the total angular momentum and $\alpha_\varphi = L_z$ its component along the polar axis. For bounded motion, $E < 0$, the three circuits give action variables

$$I_\varphi = \alpha_\varphi \,, \qquad I_\theta = \alpha_\theta - |\alpha_\varphi| \,, \qquad I_r = -\alpha_\theta + k\sqrt{\frac{m}{2\,|E|}} \,, \tag{23.43}$$

the last two by the elementary quadratures $\int_{\theta_1}^{\pi-\theta_1}\sqrt{L^2-L_z^2/\sin^2\theta}\,\mathrm{d}\theta=\pi\,(L-|L_z|)$ and $\int_{r_1}^{r_2} r^{-1}\sqrt{-A^2r^2+Br-C}\,\mathrm{d}r=\pi\left(B/2A-\sqrt{C}\right)$, both evaluated by the residue theorem Theorem 8.24 with branch points at the turning points. Adding the three,

$$E=-\frac{mk^2}{2\,(I_r+I_\theta+I_\varphi)^2}\,, \tag{23.44}$$

so the energy depends on the three actions only through their sum: the Kepler problem is completely degenerate, which is why its orbits close. Imposing $I_a=n_a\hbar$ (Proposition 23.38, in the uncorrected form the old quantum theory used) with $n=n_r+n_\theta+n_\varphi$ gives

$$E_n=-\frac{m_{\mathrm{e}}e^4}{2\,(4\pi\varepsilon_0)^2\,\hbar^2n^2}=-\frac{hcR_\infty}{n^2}\,, \tag{23.45}$$

which is Equation (70.4): with $R_\infty=1.097\,373\,156\,8\times10^7/\mathrm{m}$ the ground state lies $13.605\,693\,\mathrm{eV}$ below the ionization limit [2431]. This is the single computation that made Hamilton–Jacobi theory a working tool of atomic physics for a decade, and the reason Chapter 70 cites this chapter.

Rests on Definition 23.13 and Equations (23.49) and (70.4).

Example 23.27 (Charged particle in a uniform magnetic field)**.** Let a particle of charge q and mass m move in a uniform field $\boldsymbol{B}=B\hat{\boldsymbol{z}}$, described in the gauge $\boldsymbol{A}=(0,Bx,0)$, so that $\mathcal{H}=\left[p_x^2+(p_y-qBx)^2+p_z^2\right]/2m$. Both y and z are cyclic and

$$W=\alpha_y y+\alpha_z z+\int^x\sqrt{2mE_\perp-(\alpha_y-qBx')^2}\,\mathrm{d}x'\,,\qquad E_\perp:=E-\frac{\alpha_z^2}{2m}\,. \tag{23.46}$$

The transverse motion is bounded between the two roots of the radicand and its action variable is, substituting $u=\alpha_y-qBx$ so that the circuit becomes a circle of radius $\sqrt{2mE_\perp}$ in the u-plane,

$$I_x=\frac{1}{2\pi}\oint p_x\,\mathrm{d}x=\frac{1}{2\pi\,|q|\,B}\oint\sqrt{2mE_\perp-u^2}\,\mathrm{d}u=\frac{mE_\perp}{|q|\,B}=\frac{E_\perp}{\omega_{\mathrm{c}}}\,, \tag{23.47}$$

with $\omega_{\mathrm{c}}=|q|\,B/m$ the cyclotron angular frequency in rad/s. For an electron in $B=1.00\,\mathrm{T}$ this is $\omega_{\mathrm{c}}=1.7588\times10^{11}\,\mathrm{rad/s}$, from e/m_{e} built out of Appendix B. The magnetic moment of the gyrating charge is

$$\mu=\frac{E_\perp}{B}=\frac{|q|}{m}\,I_x\,, \tag{23.48}$$

in J/T, and since $|q|\,/m$ is fixed, μ is an adiabatic invariant exactly when I_x is—which Theorem 23.34 establishes. That identification is what Remark 23.37 rests on.

Rests on Proposition 23.14, Equation (23.49), and Theorem 23.34.

23.3 Action–angle variables and integrable systems

For motion that is bounded, the useful canonical variables are not the constants α_a themselves but a particular set of functions of them, chosen so that the conjugate coordinates are angles.

Definition 23.28 (Action variables)**.** Let the motion be bounded and let the Hamilton–Jacobi equation separate in the coordinates q^a, so that each pair (q^a,p_a) traverses a closed curve C_a in its own plane of phase space. The *action variables* are

$$I_a=\frac{1}{2\pi}\oint_{C_a}p_a\,\mathrm{d}q^a\qquad(\text{no sum on }a)\,, \tag{23.49}$$

and the conjugate *angle variables* are

$$\theta^a = \frac{\partial W}{\partial I_a}\,. \tag{23.50}$$

Rests on Definitions 16.99 and 23.13.

Remark 23.29 (Convention). The factor $1/2\pi$ in Equation (23.49) makes the angles 2π-periodic and their rates ordinary angular frequencies. The older literature, and the quantization condition Equation (23.61) below in its original form, use $J_a = \oint p_a\,\mathrm{d}q^a = 2\pi I_a$ with angles of period 1. Both appear in the sources of this treatise; the convention here is the first.

Rests on Definition 23.28.

Theorem 23.30 (Motion in action–angle variables)**.** *In action–angle variables the Hamiltonian depends on the actions alone, $\mathcal{H} = \mathcal{H}(I)$, and the equations of motion integrate immediately:*

$$\dot I_a = 0\,, \tag{23.51}$$

$$\dot\theta^a = \frac{\partial \mathcal{H}}{\partial I_a} =: \omega^a(I)\,, \qquad \theta^a(t) = \omega^a t + \theta^a_0\,. \tag{23.52}$$

The ω^a are the angular frequencies of the motion, obtained without solving for the trajectory: one circuit of C_b advances θ^b by 2π and every other angle not at all, so the period of the b-th degree of freedom is $T_b = 2\pi/\omega^b$.

Rests on Definition 23.28, Proposition 22.18, and Theorem 22.6.

Proof. *Derives Theorem 23.30.*

By Equation (23.21) each momentum $p_a = \mathrm{d}W_a/\mathrm{d}q^a$ is a function of q^a and of the constants α alone, and the circuit C_a is fixed once α is. Hence Equation (23.49) defines f functions $I_a(\alpha)$, which are constants of the motion, and which may be inverted for α wherever $\det(\partial I_a/\partial\alpha_b) \neq 0$. Substituting $\alpha = \alpha(I)$ into W gives $W(q, I)$, a generating function of the second kind (Proposition 22.18) with old momenta $p_a = \partial W/\partial q^a$ and new coordinates $\theta^a = \partial W/\partial I_a$, which is Equation (23.50).

W carries no explicit time dependence, so by Equation (22.21) the new Hamiltonian is the old one re-expressed; by Equation (23.15) that value is $E = \alpha_1(\alpha)$, a function of the constants alone, hence of the actions alone. So $\mathcal{H} = \mathcal{H}(I)$. Theorem 22.6 in the new variables then gives $\dot I_a = -\partial\mathcal{H}/\partial\theta^a = 0$, since $\mathcal{H}$ does not contain the angles, and $\dot\theta^a = \partial\mathcal{H}/\partial I_a$, a function of constants and therefore itself constant; integrating gives Equation (23.52).

For the periods, carry q once around C_b at fixed I. Since $\theta^a = \partial W/\partial I_a$ and the order of the two differentiations may be exchanged,

$$\Delta_b\theta^a = \oint_{C_b} \frac{\partial}{\partial q^b}\left(\frac{\partial W}{\partial I_a}\right)\mathrm{d}q^b = \frac{\partial}{\partial I_a}\oint_{C_b} p_b\,\mathrm{d}q^b = \frac{\partial\,(2\pi I_b)}{\partial I_a} = 2\pi\delta^a_b\,. \tag{23.53}$$

From Equations (23.49) and (23.50) (exchanging the two differentiations under the circuit integral).

Comparing with Equation (23.52), one circuit takes a time $T_b = 2\pi/\omega^b$. □

Theorem 23.31 (Liouville–Arnold)**.** *Let a system with f degrees of freedom possess f functionally independent constants of motion $F_1 = \mathcal{H}, F_2, \dots, F_f$ in involution, that is with*

$$\{F_a, F_b\} = 0 \qquad \text{for all } a, b\,. \tag{23.54}$$

Then each compact connected component of a regular level set $\{F_a = c_a\}$ is diffeomorphic to the f-torus T^f, the motion on it is quasi-periodic with f frequencies, and there exist action–angle variables in a neighbourhood of it. Such a system is called completely integrable.

Rests on Equation (22.41) and Theorems 23.30 and 24.21.

Full derivation: Appendix A.47 (p. 3039).
Derives Theorem 23.31.

Remark 23.32 (What that proof imports). The appendix argument is the classical one. The f commuting integrals generate f Hamiltonian vector fields that commute, are everywhere independent on a regular level set and are tangent to it; on a compact level set they are complete, so their joint flow is an action of $\mathbb{R}^f$, which is transitive on each connected component because its orbits are open. The stabiliser is then a discrete subgroup of $\mathbb{R}^f$ with compact quotient, hence a lattice of rank f, and the component is $\mathbb{R}^f$ modulo that lattice—the torus. The actions are the periods of $p_a\,\mathrm{d}q^a$ over a basis of its cycles, and the angles are conjugate to them.

One ingredient is imported rather than proved: that a discrete subgroup of $\mathbb{R}^f$ with compact quotient is a lattice of full rank—equivalently, that a compact connected manifold carrying f commuting, independent, complete vector fields is diffeomorphic to T^f. That is a theorem of the kind Part II owns and Part II does not carry it, so the debt is recorded here rather than discharged inside a physics appendix. Nothing this chapter proves depends on Theorem 23.31: it is used only through the statement that separable systems are integrable, and Proposition 23.20 establishes that on its own.
Rests on Theorem 23.31 and Proposition 23.20.

Remark 23.33 (Integrability is exceptional). The systems for which Theorem 23.31 applies form a set of measure zero among Hamiltonian systems. What happens to the invariant tori when an integrable system is perturbed—most of them survive, deformed, and the rest break up into the thin chaotic layers that carry the system's sensitivity to initial conditions—is the subject of KAM theory, treated in Section 32.4.2 and stated as Theorem 32.51. The observational foothold is the solar system: the Kirkwood gaps of the asteroid belt (Table 32.2) sit at the mean-motion resonances where the tori fail, and the $3:1$ gap was shown to be chaotic rather than merely resonant by direct integration [1732].
Rests on Theorems 23.31 and 32.51.

23.3.1 Adiabatic invariance

Theorem 23.34 (Adiabatic invariance of the action)**.** *Let a bounded one-dimensional system have Hamiltonian $\mathcal{H}(q,p;\lambda)$ depending on a parameter $\lambda(t)$ that varies slowly, in the sense that its fractional change over one period T of the motion is small,*

$$\varepsilon = \frac{T}{\lambda}\left|\frac{\mathrm{d}\lambda}{\mathrm{d}t}\right| \ll 1\,. \tag{23.55}$$

Then the mean rate of change of the action variable I over one period vanishes,

$$\left\langle \frac{\mathrm{d}I}{\mathrm{d}t} \right\rangle = 0 + \mathcal{O}\!\left(\varepsilon^2\right)\,, \tag{23.56}$$

so that I changes by $\mathcal{O}(\varepsilon)$, not $\mathcal{O}(1)$, over the whole time $\sim T/\varepsilon$ in which λ changes appreciably. The energy, by contrast, is not conserved: it follows λ through $E = E(I,\lambda)$ at fixed I.
Rests on Definition 23.28, Equation (22.10), and Theorem 23.30.

Proof. *Derives Theorem 23.34.*
Freeze λ and let $I(E,\lambda) = (2\pi)^{-1}\oint p\,\mathrm{d}q$ be taken over the closed orbit $\mathcal{H}(q,p;\lambda) = E$, with $p = p(q,E,\lambda)$ the solution of that equation. Two partial derivatives are needed. Differentiating $\mathcal{H}(q,p;\lambda) = E$ at fixed q and λ gives $(\partial\mathcal{H}/\partial p)\,\partial p/\partial E = 1$, and $\partial\mathcal{H}/\partial p = \dot q$ by Equation (22.3), so

$$\frac{\partial I}{\partial E} = \frac{1}{2\pi}\oint \frac{\mathrm{d}q}{\dot q} = \frac{T}{2\pi} = \frac{1}{\omega}\,. \tag{23.57}$$

Differentiating the same relation at fixed q and E gives $\partial\mathcal{H}/\partial\lambda + \dot{q}\,\partial p/\partial\lambda = 0$, whence

$$\frac{\partial I}{\partial \lambda} = -\frac{1}{2\pi}\oint \frac{\partial \mathcal{H}/\partial\lambda}{\dot{q}}\,\mathrm{d}q = -\frac{1}{2\pi}\int_0^T \frac{\partial \mathcal{H}}{\partial \lambda}\,\mathrm{d}t = -\frac{T}{2\pi}\left\langle \frac{\partial \mathcal{H}}{\partial \lambda}\right\rangle , \tag{23.58}$$

the angle brackets denoting the average over one period of the frozen motion.

Along the true motion the energy is not conserved: by Equation (22.10), $\dot{E} = \partial\mathcal{H}/\partial t = \dot{\lambda}\,\partial\mathcal{H}/\partial\lambda$. Hence

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\partial I}{\partial E}\dot{E} + \frac{\partial I}{\partial \lambda}\dot{\lambda} = \frac{\dot{\lambda}}{\omega}\frac{\partial \mathcal{H}}{\partial \lambda} - \frac{T\dot{\lambda}}{2\pi}\left\langle \frac{\partial \mathcal{H}}{\partial \lambda}\right\rangle . \tag{23.59}$$

Both terms are of order ε separately. Averaging Equation (23.59) over one period, during which $\dot{\lambda}$ may be held fixed at the cost of an error of relative order ε, replaces $\partial\mathcal{H}/\partial\lambda$ in the first term by its mean, and $T = 2\pi/\omega$ makes the two terms cancel identically:

$$\left\langle \frac{\mathrm{d}I}{\mathrm{d}t}\right\rangle = \frac{\dot{\lambda}}{\omega}\left\langle \frac{\partial \mathcal{H}}{\partial \lambda}\right\rangle - \frac{2\pi}{\omega}\frac{\dot{\lambda}}{2\pi}\left\langle \frac{\partial \mathcal{H}}{\partial \lambda}\right\rangle = 0\,. \tag{23.60}$$

This is Equation (23.56). Summing over the $\sim 1/\varepsilon$ periods of the slow evolution, the surviving change in I is the accumulated remainder, of order $\varepsilon^2 \times \varepsilon^{-1} = \varepsilon$, whereas an uncompensated first-order drift would have accumulated to order one. The energy statement is immediate: inverting $I = I(E,\lambda)$ at fixed I makes E a function of λ, and $\partial E/\partial I = \omega$ by Equation (23.57), so E tracks λ through the frequency. □

Example 23.35 (The slowly stiffened oscillator)**.** For the oscillator of Example 23.24 with a slowly varying frequency $\omega_0(t)$, Equation (23.38) gives $I = E/\omega_0$, so Theorem 23.34 says E/ω_0 is conserved and therefore $E \propto \omega_0$: shortening a pendulum quasi-statically raises its energy in proportion to its frequency, while its amplitude falls as $\omega_0^{-1/2}$. The energy is supplied by whatever does the shortening, and no amount of care in doing it slowly avoids that.

Rests on Theorem 23.34 and Equation (23.38)*.*

Remark 23.36 (How good the invariance really is)*.* Theorem 23.34 is a first-order statement and is far weaker than what is true. If $\lambda(t)$ is analytic and varies over a finite interval, returning to a constant at both ends, the total change in I is not $\mathcal{O}(\varepsilon)$ but exponentially small, of order $\mathrm{e}^{-c/\varepsilon}$—which is why adiabatic invariants are useful in practice rather than merely asymptotically, and why leaving the statement out would misrepresent how good the invariance is.

That sharper statement is not established in this treatise, and it is recorded as owed rather than asserted, for two reasons. No source in this bibliography states it, so it cannot honestly be cited. And the method it needs—a sequence of averaging transformations, each removing one more order in ε at the cost of shrinking a strip of analyticity, with Cauchy estimates controlling the remainder and the number of steps optimised against the shrinkage—rests on analytic machinery that Part II does not develop; the size of the exponent is then fixed by the distance from the real axis to the nearest complex singularity of λ. Nothing in the book uses it: every application in this chapter and in Chapter 66 needs only the first-order result Equation (23.56) proved above.

Rests on Theorem 23.34.

Remark 23.37 (Why this matters observationally)*.* Adiabatic invariance is the reason the robust quantity of a slowly modulated oscillator is its action and not its energy. The magnetic moment Equation (23.48) of a charged particle gyrating in a slowly varying magnetic field is an adiabatic invariant of exactly this kind: $\mu = E_\perp/B$ is fixed, so as a particle spirals into a region of stronger field its transverse energy must rise, and—the total energy being conserved in a static field—its parallel energy must fall until it reverses. That is the magnetic

mirror, and it is what traps particles in the belts Van Allen identified from the first satellite instruments [1176]. The same invariant governs betatron acceleration and the confinement schemes of laboratory plasmas (Chapter 66).

Rests on Theorem 23.34 and Equation (23.48).

23.3.2 The old quantum theory

Proposition 23.38 (Bohr–Sommerfeld and Einstein–Brillouin–Keller quantization)**.** *The rule that selected the observed atomic states in the old quantum theory is the quantization of the action variables,*

$$\oint_{C_a} p_a \,\mathrm{d}q^a = n_a h\,, \qquad I_a = n_a \hbar\,, \qquad n_a \in \mathbb{N}\,, \tag{23.61}$$

one condition for each separable degree of freedom [441]. Its correct semiclassical form carries an extra term,

$$I_a = \left(n_a + \frac{\mu_a}{4}\right)\hbar\,, \qquad n_a = 0, 1, 2, \dots\,, \tag{23.62}$$

where the integer μ_a—the Maslov index *of the a-th circuit—counts the caustics the circuit crosses, and equals 2 for a smooth one-dimensional well. Equation (23.61) is the case $\mu_a = 0$.*

Rests on Definition 23.28, Theorem 9.78, and Remark 9.79.

Derivation. *Derives Proposition 23.38.*

Neither form can be derived within classical mechanics: both are selection rules imposed from outside, and what classical mechanics supplies is the *object* they quantize, the action variable Equation (23.49), defined by the geometry of phase space and by nothing else.

The justification is semiclassical and is already in this treatise. For one degree of freedom, Theorem 9.78 shows that a solution of the stationary Schrödinger equation Equation (9.59) decaying on both sides of a well exists only for those energies with $\oint P\,\mathrm{d}x = \left(n + \frac{1}{2}\right)h$, which is Equation (23.62) with $\mu = 2$, and Remark 9.79 identifies the two quarters as one $\pi/4$ from each of the two turning points, contributed by the connection formulae Theorem 9.72. Keller extended the same accounting to several degrees of freedom by applying it to each independent circuit of an invariant torus and counting caustics rather than turning points [1150]; that count is μ_a, and its geometric content is Proposition 23.56 below. Substituting $\oint p_a \mathrm{d}q^a = 2\pi I_a$ and $h = 2\pi\hbar$ turns the condition into Equation (23.62). □

The empirical successes and the failures of the uncorrected rule—it gives the hydrogen spectrum Equation (23.45) and the fine structure, and fails for helium—are the subject of Chapter 70. Two structural points survive the failure. First, the quantity being quantized is the action variable of Definition 23.28. Second, the missing $\mu_a/4$ is not a patch but a statement about the classical trajectory: it counts where the projection of the invariant torus onto configuration space folds over. That is why Section 23.5.2 is part of this chapter and not an appendix to it.

23.4 The optical–mechanical analogy

23.4.1 The eikonal limit of wave optics

Definition 23.39 (Eikonal)**.** Let a monochromatic scalar wave of vacuum wavenumber $k_0 = \omega/c$, in 1/m, propagate in a medium of refractive index $n(\boldsymbol{x})$, so that its amplitude satisfies the Helmholtz equation Equation (10.25) with local wavenumber $k = k_0 n(\boldsymbol{x})$,

$$\nabla^2 u + k_0^2 n^2(\boldsymbol{x})\, u = 0\,. \tag{23.63}$$

Write it in the form

$$u(\boldsymbol{x}) = A(\boldsymbol{x})\,\mathrm{e}^{\mathrm{i}k_0\mathcal{S}(\boldsymbol{x})}\,. \tag{23.64}$$

The real function $\mathcal{S}$, in metres, is the *eikonal*; its level surfaces $\mathcal{S} = \text{const}$ are the geometrical wavefronts. The name is Bruns' [263]; the function is Hamilton's characteristic function of [84].

Rests on Equation (10.25) and Definition 16.103.

Theorem 23.40 (Eikonal and transport equations)**.** *In the short-wavelength limit $k_0 \to \infty$, with A and n varying appreciably only over distances large compared with $1/k_0$, the Helmholtz equation reduces at leading order to the* eikonal equation

$$(\boldsymbol{\nabla}\mathcal{S})^2 = n^2(\boldsymbol{x})\,, \tag{23.65}$$

and at the next order to the transport equation

$$\boldsymbol{\nabla}\cdot\left(A^2\boldsymbol{\nabla}\mathcal{S}\right) = 0\,, \tag{23.66}$$

which states the conservation of energy flux along the tubes of rays. The neglected term is $\nabla^2 A/\left(k_0^2 A\right)$, so the approximation is good where $\left|\nabla^2 A\right| \ll k_0^2\,|A|$.

Rests on Definition 23.39 and Equation (23.63).

Proof. *Derives Theorem 23.40.*

Differentiate Equation (23.64) twice:

$$\nabla^2 u = \left[\nabla^2 A + 2\mathrm{i}k_0\,\boldsymbol{\nabla}A\cdot\boldsymbol{\nabla}\mathcal{S} + \mathrm{i}k_0 A\,\nabla^2\mathcal{S} - k_0^2 A\,(\boldsymbol{\nabla}\mathcal{S})^2\right]\mathrm{e}^{\mathrm{i}k_0\mathcal{S}}\,. \tag{23.67}$$

Substituting into Equation (23.63) and cancelling the exponential leaves, with A, $\mathcal{S}$ and n all real,

$$\underbrace{\nabla^2 A - k_0^2 A\left[(\boldsymbol{\nabla}\mathcal{S})^2 - n^2\right]}_{\text{real}} + \mathrm{i}k_0\underbrace{\left[2\,\boldsymbol{\nabla}A\cdot\boldsymbol{\nabla}\mathcal{S} + A\,\nabla^2\mathcal{S}\right]}_{\text{imaginary}} = 0\,, \tag{23.68}$$

and each bracket vanishes separately. The real one, divided by $-k_0^2 A$, reads $(\boldsymbol{\nabla}\mathcal{S})^2 - n^2 = \nabla^2 A/\left(k_0^2 A\right)$, whose right-hand side is the stated neglected term; dropping it gives Equation (23.65). The imaginary one, multiplied by A, is $2A\,\boldsymbol{\nabla}A\cdot\boldsymbol{\nabla}\mathcal{S} + A^2\nabla^2\mathcal{S} = \boldsymbol{\nabla}A^2\cdot\boldsymbol{\nabla}\mathcal{S} + A^2\nabla^2\mathcal{S}$, which is Equation (23.66). □

Proposition 23.41 (Rays)**.** *The curves orthogonal to the wavefronts—the* rays*—satisfy*

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(n\,\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s}\right) = \boldsymbol{\nabla}n\,, \tag{23.69}$$

s being arclength, and they are the stationary curves of Fermat's principle

$$\delta\int n(\boldsymbol{x})\,\mathrm{d}s = 0\,. \tag{23.70}$$

Rests on Theorem 23.40, Lemma 16.100, and Definition 16.103.

Proof. *Derives Proposition 23.41.*

By Equation (23.65) the gradient of the eikonal has magnitude n, so the unit normal to the wavefronts is $\boldsymbol{T} = \boldsymbol{\nabla}\mathcal{S}/n$, and along a ray, parametrised by arclength, $\boldsymbol{T} = \mathrm{d}\boldsymbol{x}/\mathrm{d}s$. Hence $n\,\mathrm{d}\boldsymbol{x}/\mathrm{d}s = \boldsymbol{\nabla}\mathcal{S}$ and

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(n\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s}\right) = (\boldsymbol{T}\cdot\boldsymbol{\nabla})\,\boldsymbol{\nabla}\mathcal{S} = \frac{1}{n}\,(\boldsymbol{\nabla}\mathcal{S}\cdot\boldsymbol{\nabla})\,\boldsymbol{\nabla}\mathcal{S} = \frac{1}{2n}\boldsymbol{\nabla}\,(\boldsymbol{\nabla}\mathcal{S})^2 = \frac{1}{2n}\boldsymbol{\nabla}n^2 = \boldsymbol{\nabla}n\,. \tag{23.71}$$

From Equation (23.65) (a gradient field is curl-free, and its squared magnitude is fixed by the eikonal equation).

The middle step holds because a gradient field has vanishing curl, so that $(\boldsymbol{a}\cdot\boldsymbol{\nabla})\,\boldsymbol{a} = \frac{1}{2}\boldsymbol{\nabla}\,|\boldsymbol{a}|^2$ for $\boldsymbol{a} = \boldsymbol{\nabla}\mathcal{S}$, and the last uses Equation (23.65). Equation (23.69) is Equation (16.90), the Euler–Lagrange equation of the optical path length established in Lemma 16.100; Fermat's principle Equation (23.70) is therefore equivalent to it, which is Definition 16.103 [260]. □

23.4.2 The dictionary

Compare Equation (23.65) with the time-independent Hamilton–Jacobi equation Equation (23.15) for a Hamiltonian $\mathcal{H} = |\boldsymbol{p}|^2/2m + V$, which reads

$$(\boldsymbol{\nabla} W)^2 = 2m\,(E - V(\boldsymbol{x}))\ . \tag{23.72}$$

The two equations are the same equation. Every statement about wavefronts and rays therefore has a mechanical counterpart, obtained by the substitution $n \mapsto \sqrt{2m(E-V)}$.

Table 23.1: Hamilton's optical–mechanical analogy. Each row is a single mathematical statement read in two languages; the identification of the refractive index with $\sqrt{2m(E-V)}$ turns either column into the other.

Geometrical optics	**Classical mechanics**
eikonal $\mathcal{S}(\boldsymbol{x})$	characteristic function $W(\boldsymbol{x})$
refractive index $n(\boldsymbol{x})$	$\sqrt{2m\,(E - V(\boldsymbol{x}))}$
eikonal equation $(\boldsymbol{\nabla}\mathcal{S})^2 = n^2$	Hamilton–Jacobi equation $(\boldsymbol{\nabla} W)^2 = 2m(E-V)$
wavefronts $\mathcal{S} = \text{const}$	surfaces of constant action $W = \text{const}$
rays, orthogonal to the wavefronts	trajectories, orthogonal to the surfaces $W = \text{const}$
Fermat's principle $\delta \int n\,\mathrm{d}s = 0$	Maupertuis' principle $\delta \int \sqrt{2m(E-V)}\,\mathrm{d}s = 0$
Snell's law at an interface	refraction of a trajectory at a potential step
wavelength $\lambda = \lambda_0/n$	de Broglie wavelength $\lambda = h/\lvert\boldsymbol{p}\rvert$
geometrical-optics limit $\lambda \to 0$	classical limit $\hbar \to 0$

The Snell row of Table 23.1 is worth checking, because it is the row a reader is likeliest to take on trust. At a plane interface across which V jumps, the component of $\boldsymbol{p}$ tangent to the interface is conserved—no force acts along it—so $|\boldsymbol{p}_1|\sin\vartheta_1 = |\boldsymbol{p}_2|\sin\vartheta_2$ with ϑ measured from the normal, which under $n = |\boldsymbol{p}|$ is Equation (16.93) exactly (Theorem 16.104 and Phenomenon 61.3). A particle entering a region of lower potential speeds up and bends *towards* the normal, as light does entering a denser medium. This is the correspondence Newton used to argue that light is corpuscular, and the one that made the measured slowing of light in water Phenomenon 61.2 decisive against him.

Remark 23.42 (Phase speed and particle speed). The wavefronts $S = \text{const}$ do *not* move at the speed of the particle. From Equation (23.14), a surface of constant S moves so that $\boldsymbol{\nabla} W \cdot \mathrm{d}\boldsymbol{x} = E\,\mathrm{d}t$, giving a normal speed $u = E/|\boldsymbol{\nabla} W| = E/|\boldsymbol{p}|$, whereas the particle moves at $v = |\boldsymbol{p}|/m$; for a free particle $uv = E/m = v^2/2$, so $u = v/2$, as Example 23.22 found. The wavefront speed is a phase speed and the particle speed a group speed, and their product is fixed. In the wave mechanics of Chapter 71 the same two speeds reappear with the same relation, which is one of the checks de Broglie's hypothesis had to pass [521].

Rests on Equation (23.14) and Example 23.22.

Remark 23.43 (The asymmetry that carries the physics). The dictionary of Table 23.1 is exact, and it is incomplete in one direction only. Geometrical optics is known to be an approximation: it is the short-wavelength limit of a wave theory, and where the limit fails—diffraction at an edge, interference, the caustics where neighbouring rays cross—the wave theory is required, and it is right. The last two rows say that classical mechanics stands to some wave theory in exactly the relation that geometrical optics stands to wave optics. Hamilton wrote the analogy in the 1830s and left it there. The wave theory whose

short-wavelength limit is classical mechanics was found a century later, and the parameter that plays the part of the wavelength is $h/|\boldsymbol{p}|$. That is the content of Chapter 71, and the experiments that settled it—electron diffraction, neutron and atom interferometry—are its evidence.

Rests on Equations (23.65) *and* (23.72).

23.4.3 Where the analogy is not an analogy

The dictionary of Table 23.1 pairs two theories that remain distinct. There is a regime in which the pairing collapses and the two columns become one column: relativistic mechanics, where the eikonal of a massless field and the characteristic function of a massive particle satisfy the same equation with the same metric, differing only in the value of one constant.

Proposition 23.44 (Relativistic Hamilton–Jacobi equation)**.** *For a free particle of rest mass m, with the signature and index conventions of Chapter 39 and $\partial S/\partial x^\mu = -p_\mu$, the Hamilton–Jacobi equation is*

$$\eta^{\mu\nu}\,\partial_\mu S\,\partial_\nu S = m^2c^2\,, \qquad \textit{that is} \qquad \frac{1}{c^2}\left(\frac{\partial S}{\partial t}\right)^2 - (\boldsymbol{\nabla} S)^2 = m^2c^2\,, \tag{23.73}$$

both sides in $\mathrm{kg}^2\,\mathrm{m}^2/\mathrm{s}^2$*. In the presence of an electromagnetic four-potential A_μ and for charge q it becomes*

$$\eta^{\mu\nu}\,(\partial_\mu S + qA_\mu)\,(\partial_\nu S + qA_\nu) = m^2c^2\,, \tag{23.74}$$

and in a spacetime of metric $g_{\mu\nu}$ it is Equation (23.73) *with $\eta^{\mu\nu}$ replaced by $g^{\mu\nu}$. Writing $S = -mc^2t + S'$ and expanding in $1/c^2$ recovers Equation* (23.2) *for S' with $\mathcal{H} = |\boldsymbol{p}|^2/2m$.*

Rests on Equations (23.2) *and* (40.5) *and Definition 13.117.*

Proof. *Derives Proposition 23.44.*

The nonrelativistic relations $p_a = \partial S/\partial q^a$ and $\partial S/\partial t = -\mathcal{H} = -E$ are, in four-vector notation with $x^\mu = (ct, \boldsymbol{x})$ and $p^\mu = (E/c, \boldsymbol{p})$, the single statement $\partial S/\partial x^\mu = -p_\mu$: the time component reads $c^{-1}\partial_t S = -E/c = -p_0$ and the space components $\partial_i S = p^i = -p_i$. Contracting with $\eta^{\mu\nu}$ gives $\eta^{\mu\nu}\partial_\mu S\,\partial_\nu S = \eta^{\mu\nu}p_\mu p_\nu = p^2 = m^2c^2$ by Theorem 40.5, which is Equation (23.73); writing out the components with $\eta = \mathrm{diag}(1,-1,-1,-1)$ gives the second form. Equation (23.74) follows by the minimal substitution $p_\mu \mapsto p_\mu - qA_\mu$ in the same mass-shell condition, and the curved-space form by replacing the flat metric with $g_{\mu\nu}$, which is what makes p^2 the invariant square.

For the limit, put $S = -mc^2t + S'$ in the second form of Equation (23.73). Then $\partial_t S = -mc^2 + \partial_t S'$ and

$$\frac{1}{c^2}\,(\partial_t S)^2 = m^2c^2 - 2m\,\partial_t S' + \frac{1}{c^2}\,(\partial_t S')^2\,,$$

so Equation (23.73) becomes $-2m\,\partial_t S' - (\boldsymbol{\nabla} S')^2 + c^{-2}\,(\partial_t S')^2 = 0$. Dividing by $-2m$ and dropping the last term, of relative order $(\partial_t S')/mc^2$, leaves $\partial_t S' + (\boldsymbol{\nabla} S')^2/2m = 0$. □

Remark 23.45 (Why this is not an analogy). Set $m = 0$ in Equation (23.73): the result, $g^{\mu\nu}\partial_\mu S\,\partial_\nu S = 0$, is the eikonal equation for a massless field in that spacetime, and its characteristics are the null geodesics. Set $m \neq 0$ and the characteristics are the timelike geodesics of Chapter 43. One equation, one geometry, one family of characteristic curves, and the only difference between the ray of light and the trajectory of the particle is the value on the right-hand side. In Table 23.1 the refractive index had to be manufactured out of E and V to make the two columns match; here nothing is manufactured, because there is only one column. The optical–mechanical analogy is, in the relativistic setting, an identity that was hiding inside a nonrelativistic approximation.

Rests on Proposition 23.44 and Theorem 23.40.

Remark 23.46 (Carter's constant). The most consequential separation of Equation (23.73) outside flat space is Carter's. In the Kerr geometry—the rotating black hole of Chapter 45—the equation admits a complete integral of the separated form $S = -Et + L_z\varphi + S_r(r) + S_\theta(\theta)$ in Boyer–Lindquist coordinates, and the separation constant that couples the radial and polar equations is a *fourth* constant of the motion,

$$\mathcal{K} = p_\theta^2 + \cos^2\theta \left[a^2 \left(m^2 c^2 - \frac{E^2}{c^2} \right) + \frac{L_z^2}{\sin^2\theta} \right], \tag{23.75}$$

with a the spin parameter of the hole, a length in m [1366]. Three constants come from symmetry—m from the mass shell, E and L_z from the two Killing vectors of the stationary axisymmetric metric—and the fourth comes from nowhere that Noether's theorem (Theorem 16.84) can see. In the language of Remark 23.21 it is a rank-two Killing tensor, and the separability of Equation (23.73) in that geometry is exactly the Stäckel condition of Definition 23.16 transplanted to a Lorentzian metric. That it exists at all is what makes geodesics around a rotating black hole integrable, and therefore computable; the orbits so computed are what the observed motions of Chapter 50 are matched against. The separation itself is carried out in Part V and is not reproduced here, and Remark 23.21 has already recorded that the Killing-tensor machinery it deserves is missing from Chapter 13.

Rests on Proposition 23.44 and Definition 23.16.

23.5 From Hamilton–Jacobi to the Schrödinger equation

Proposition 23.47 (The Hamilton–Jacobi equation as the classical limit)**.** *Write a solution of the Schrödinger equation of Chapter 77 in the polar form*

$$\psi(\boldsymbol{x}, t) = A(\boldsymbol{x}, t)\, \mathrm{e}^{\mathrm{i}S(\boldsymbol{x},t)/\hbar}, \tag{23.76}$$

with A and S real. Then the real and imaginary parts of the Schrödinger equation are, exactly and without approximation,

$$\frac{\partial S}{\partial t} + \frac{(\boldsymbol{\nabla} S)^2}{2m} + V + Q = 0, \qquad Q := -\frac{\hbar^2}{2m}\frac{\nabla^2 A}{A}, \tag{23.77}$$

$$\frac{\partial A^2}{\partial t} + \boldsymbol{\nabla} \cdot \left(A^2 \frac{\boldsymbol{\nabla} S}{m} \right) = 0. \tag{23.78}$$

Equation (23.77) is the Hamilton–Jacobi equation Equation (23.2) with the single extra term Q, of order $\hbar^2$; Equation (23.78) is the continuity equation for the probability density A^2, and is the exact counterpart of the transport equation Equation (23.66).

Rests on Equations (10.26), (23.2) and (23.66).

Proof. *Derives Proposition 23.47.*

With $\psi = A\mathrm{e}^{\mathrm{i}S/\hbar}$,

$$\frac{\partial \psi}{\partial t} = \left(\frac{\partial A}{\partial t} + \frac{\mathrm{i}A}{\hbar}\frac{\partial S}{\partial t} \right) \mathrm{e}^{\mathrm{i}S/\hbar}, \qquad \nabla^2\psi = \left[\nabla^2 A + \frac{2\mathrm{i}}{\hbar}\boldsymbol{\nabla} A \cdot \boldsymbol{\nabla} S + \frac{\mathrm{i}A}{\hbar}\nabla^2 S - \frac{A}{\hbar^2}(\boldsymbol{\nabla} S)^2 \right] \mathrm{e}^{\mathrm{i}S/\hbar}. \tag{23.79}$$

Substituting into $\mathrm{i}\hbar\,\partial_t\psi = -(\hbar^2/2m)\,\nabla^2\psi + V\psi$ (Equation (10.26)) and cancelling the exponential, the real part is

$$-A\frac{\partial S}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2 A + \frac{A\,(\boldsymbol{\nabla} S)^2}{2m} + VA,$$

which on division by A is Equation (23.77), and the imaginary part is

$$\hbar\frac{\partial A}{\partial t} = -\frac{\hbar}{m}\boldsymbol{\nabla} A \cdot \boldsymbol{\nabla} S - \frac{\hbar A}{2m}\nabla^2 S.$$

Multiplying the latter by $2A/\hbar$ turns the left side into $\partial\left(A^2\right)/\partial t$ and the right into $-m^{-1}\left[\boldsymbol{\nabla}\left(A^2\right)\cdot\boldsymbol{\nabla}S + A^2\nabla^2 S\right]$, which is $-\boldsymbol{\nabla}\cdot\left(A^2\boldsymbol{\nabla}S/m\right)$: that is Equation (23.78). No approximation was made at any step; both equations are exact rewritings. □

Remark 23.48 (What this is and what it is not). Equations (23.77) and (23.78) are an exact algebraic rewriting of the Schrödinger equation, not a different theory and not an interpretation of it. Read in one direction they show that classical mechanics is recovered when the term Q is negligible against the others, which is the precise sense in which the Hamilton–Jacobi equation is the $\hbar \to 0$ limit of wave mechanics. Read in the other direction, as a statement that a particle possesses a definite trajectory guided by Q, they become an interpretive commitment that no experiment discussed in this treatise distinguishes from any other; the treatise's position on such commitments is stated in Chapter 85.

Rests on Equations (23.77) *and* (23.78).

23.5.1 The semiclassical expansion

Proposition 23.47 is exact but not systematic: it isolates one term of order $\hbar^2$ and says nothing about what comes after it. The systematic version is the WKB expansion, and this treatise develops it in full—the expansion itself, its validity criterion, its failure at turning points, the connection formulae that carry a solution across one, the Gamow tunnelling factor and the quantisation condition with its Maslov correction—in one dimension, as mathematics, in Section 9.6 of Chapter 9. That material is not repeated here. What is established here is the one thing that section cannot state, because it is mechanics and not analysis: that the leading order of the expansion *is* the Hamilton–Jacobi equation of this chapter, and the next order the transport equation.

Proposition 23.49 (The Hamilton–Jacobi equation is the leading order)**.** *Write a solution of the stationary Schrödinger equation* $-\left(\hbar^2/2m\right)\nabla^2\psi + V\psi = E\psi$ *in three space dimensions as a single exponential,*

$$\psi(\boldsymbol{x}) = \exp\left[\frac{\mathrm{i}\,\sigma(\boldsymbol{x})}{\hbar}\right], \tag{23.80}$$

with σ *complex and of SI dimension* J s*. Then* σ *satisfies exactly*

$$\left(\boldsymbol{\nabla}\sigma\right)^2 - \mathrm{i}\hbar\,\nabla^2\sigma = 2m\left(E - V\right) = \left|\boldsymbol{P}(\boldsymbol{x})\right|^2, \tag{23.81}$$

and the expansion $\sigma = \sigma_0 + \hbar\sigma_1 + \hbar^2\sigma_2 + \cdots$ *of Equation* (9.64) *gives, at its first two orders,*

$$\left(\boldsymbol{\nabla}\sigma_0\right)^2 = 2m\left(E - V\right), \tag{23.82}$$

$$2\,\boldsymbol{\nabla}\sigma_0\cdot\boldsymbol{\nabla}\sigma_1 = \mathrm{i}\,\nabla^2\sigma_0\,. \tag{23.83}$$

Equation (23.82) *is the time-independent Hamilton–Jacobi equation Equation* (23.72)*, so* $\sigma_0 = W$*; and with* $A := \mathrm{e}^{\mathrm{i}\sigma_1}$*, Equation* (23.83) *is the transport equation* $\boldsymbol{\nabla}\cdot\left(A^2\boldsymbol{\nabla}W\right) = 0$*, that is Equation* (23.66)*.*

Rests on Equations (23.66) *and* (23.72) *and Proposition 9.61.*

Proof. *Derives Proposition 23.49.*

From Equation (23.80), $\boldsymbol{\nabla}\psi = (\mathrm{i}/\hbar)\,\boldsymbol{\nabla}\sigma\,\psi$ and $\nabla^2\psi = \left[(\mathrm{i}/\hbar)\,\nabla^2\sigma - \hbar^{-2}\left(\boldsymbol{\nabla}\sigma\right)^2\right]\psi$. Substituting into the stationary equation, cancelling the nowhere-vanishing ψ and multiplying by $-2m$ gives Equation (23.81); this is the three-dimensional form of the exact Riccati equation Equation (9.63), and $\hbar$ again appears in exactly one place, multiplying the highest derivative. Inserting the series, the squared term contributes $\left(\boldsymbol{\nabla}\sigma_0\right)^2 + 2\hbar\,\boldsymbol{\nabla}\sigma_0\cdot\boldsymbol{\nabla}\sigma_1 + \mathcal{O}(\hbar^2)$ and the second term $-\mathrm{i}\hbar\,\nabla^2\sigma_0 + \mathcal{O}(\hbar^2)$, while the right side is of order $\hbar^0$; equating coefficients of $\hbar^0$ and $\hbar^1$ gives Equations (23.82) and (23.83). Comparing Equation (23.82) with Equation (23.72) identifies σ_0 with the characteristic function.

For the last statement, $\psi = \mathrm{e}^{\mathrm{i}\sigma_1}\mathrm{e}^{\mathrm{i}\sigma_0/\hbar} = A\,\mathrm{e}^{\mathrm{i}W/\hbar}$ with $A = \mathrm{e}^{\mathrm{i}\sigma_1}$, so $\boldsymbol{\nabla}\left(A^2\right) = 2\mathrm{i}A^2\boldsymbol{\nabla}\sigma_1$ and

$$\boldsymbol{\nabla}\cdot\left(A^2\boldsymbol{\nabla}W\right) = A^2\left[2\mathrm{i}\,\boldsymbol{\nabla}\sigma_1\cdot\boldsymbol{\nabla}\sigma_0 + \nabla^2\sigma_0\right],$$

which vanishes precisely when $2\boldsymbol{\nabla}\sigma_0\cdot\boldsymbol{\nabla}\sigma_1 = \mathrm{i}\nabla^2\sigma_0$, that is Equation (23.83). □

Remark 23.50 (The dictionary completed). Proposition 23.49 closes Table 23.1. The eikonal equation of wave optics and the Hamilton–Jacobi equation of mechanics were shown in Section 23.4.2 to be the same equation under $n \mapsto \sqrt{2m(E-V)}$; they are now seen to occupy the same *position* in their respective theories—each is the leading term of a short-wavelength expansion of a linear wave equation, and each is accompanied at the next order by the same transport equation. The role of $1/k_0$ in Theorem 23.40 is played by $\hbar$ here, and the reduced de Broglie wavelength Equation (9.72) of Definition 9.64,

$$\bar{\lambda}(\boldsymbol{x}) = \frac{\hbar}{|\boldsymbol{P}(\boldsymbol{x})|}, \tag{23.84}$$

in metres, is the local wavelength whose smallness is the whole hypothesis. The criterion is $\left|\boldsymbol{\nabla}\bar{\lambda}\right| \ll 1$, which is Equation (9.73) of Proposition 9.65 read in three dimensions: the wavelength must change by much less than itself over one wavelength. For an electron of kinetic energy 100 eV, $\bar{\lambda} = 1.95\times10^{-11}$ m, so the criterion demands a potential smooth on the scale of a few tens of picometres—which is why the semiclassical description fails inside an atom and succeeds for an electron crossing a laboratory field.

Rests on Propositions 9.65 and 23.49.

Remark 23.51 (Where the expansion breaks, and what is owed). The criterion fails wherever $|\boldsymbol{P}| \to 0$, that is on the classical turning surface $V = E$, whatever the smoothness of V: the amplitude $|\boldsymbol{P}|^{-1/2}$ of Equation (9.69) diverges there while the true solution stays finite (Remark 9.66). Within a distance of order

$$\ell = \left(\frac{\hbar^2}{2m\,|\boldsymbol{\nabla}V|}\right)^{1/3} \tag{23.85}$$

of the surface, which is Equation (9.74), the expansion is worthless and the exact local solution is an Airy function (Definition 9.68 and Theorem 9.70); the connection formulae Theorem 9.72 carry a solution through. That analysis is one-dimensional and transverse to the surface, and it applies here without change, because near a smooth turning surface the motion normal to it decouples from the motion along it at leading order. The two things it does *not* settle are what happens at a caustic which is not a turning surface, and how the phases accumulated at successive crossings add up around a closed circuit. Those are the subject of Section 23.5.2, and they are mechanics, not one-dimensional analysis. The series is in any case asymptotic and not convergent (Remark 9.67), so no amount of further orders repairs a place where the leading order is wrong.

Rests on Proposition 9.65 and Theorem 9.72.

23.5.2 Caustics

A complete integral does not describe one trajectory but a family of them, and the geometrical description of Section 23.4.2 presumes that exactly one member of the family passes through each point. Where that fails—where neighbouring members cross—the description fails with it, and the failure is not a defect of the calculation but the signature of the wave theory underneath.

Definition 23.52 (Lagrangian family; caustic)**.** Let $\gamma(\boldsymbol{a};\cdot)$ be a family of trajectories of one energy, smoothly parametrised by $\boldsymbol{a} \in \mathbb{R}^{f-1}$ and issuing orthogonally from a common

wavefront (or from a common point, for a point source), and let $q^b(\boldsymbol{a}, s)$ be the configuration reached along the member $\boldsymbol{a}$ after arclength s of the Jacobi metric Equation (16.91), in which the trajectories of energy E are geodesics by Theorem 16.101. The *Jacobian* of the family is

$$J(\boldsymbol{a}, s) = \det\left(\frac{\partial q^b}{\partial\left(a^1, \ldots, a^{f-1}, s\right)}\right), \tag{23.86}$$

and the *caustic* of the family is the set where $J = 0$: the locus at which the map from the family to configuration space ceases to be a local diffeomorphism. Equivalently, the caustic is the envelope of the family.

Rests on Definitions 16.62 and 23.5.

Proposition 23.53 (A caustic point is a conjugate point)**.** *A point $q\left(\boldsymbol{a}, s_c\right)$ lies on the caustic of the family issuing from a common source if and only if s_c is conjugate, in the sense of Definition 16.53, to the source along that trajectory. Consequently, by Theorem 16.55, a trajectory ceases to minimise the action beyond its first caustic, although it remains an extremal and a genuine solution of the equations of motion.*

Rests on Definitions 16.53 and 23.52 and Theorem 16.55.

Proof. *Derives Proposition 23.53.*

The variation fields of the family, $u^b_{(i)} := \partial q^b / \partial a^i$, are by Remark 16.54 solutions of the Jacobi accessory equation Equation (16.49)—the linearisation of the Euler–Lagrange equations about the trajectory—because the family consists of extremals depending differentiably on the parameter. All members issue from the same source, so $u^b_{(i)} = 0$ there.

Because s is the arclength of the Jacobi metric and the trajectories are its geodesics (Theorem 16.101), each $u_{(i)}$ stays orthogonal to the tangent $\boldsymbol{T} = \partial q / \partial s$: differentiating $\boldsymbol{T} \cdot \boldsymbol{T} = 1$ with respect to a^i gives $\boldsymbol{T} \cdot \partial_s u_{(i)} = 0$, and $\partial_s \boldsymbol{T} = \boldsymbol{0}$ along a geodesic, so $\boldsymbol{T} \cdot u_{(i)}$ is constant along the trajectory and vanishes at the source. The Jacobian Equation (23.86) therefore vanishes at s_c exactly when the $f-1$ transverse vectors $u_{(i)}$ are linearly dependent there, that is when some non-trivial constant combination $u = c^i u_{(i)}$ vanishes at s_c. Such a u is a non-trivial solution of Equation (16.49) vanishing at the source and at s_c, which is the definition of a conjugate point. The converse reads the same equivalence backwards. □

Proposition 23.54 (The geometrical amplitude diverges)**.** *Along a tube of trajectories of one energy, the transport equation Equation (23.66) gives*

$$A^2 \, |\boldsymbol{P}| \, \Sigma = const, \tag{23.87}$$

Σ being the cross-sectional area of the tube and $|\boldsymbol{P}| = \sqrt{2m(E-V)}$ the local momentum. Since $\Sigma \propto |J|$, the geometrical amplitude behaves as $A \propto |J|^{-1/2}$ and diverges on the caustic.

Rests on Equation (23.66), Definition 23.52, and Theorem 7.101.

Proof. *Derives Proposition 23.54.*

Take the region bounded by two cross-sections Σ_1, Σ_2 of the tube and by the tube's side wall, which is built of trajectories. Integrate Equation (23.66) over it and apply the divergence theorem Theorem 7.101. The side wall contributes nothing, because $\boldsymbol{\nabla} W$ is tangent to the trajectories and hence to the wall, so $A^2 \boldsymbol{\nabla} W$ has no normal component there. What remains is that the flux $\int A^2 \boldsymbol{\nabla} W \cdot \mathrm{d}\boldsymbol{\Sigma}$ takes the same value on the two end faces. On a face orthogonal to the trajectories $\boldsymbol{\nabla} W \cdot \mathrm{d}\boldsymbol{\Sigma} = |\boldsymbol{\nabla} W| \, \mathrm{d}\Sigma = |\boldsymbol{P}| \, \mathrm{d}\Sigma$ by Equation (23.72), which for a thin tube is Equation (23.87). For the cross-section, the volume swept by the parameter increments $\mathrm{d}^{f-1}a$ over an arclength $\mathrm{d}s$ is $|J| \, \mathrm{d}^{f-1}a \, \mathrm{d}s$ by Equation (23.86), and it is also $\Sigma \, \mathrm{d}s$, the tangent being a unit vector orthogonal to the cross-section as shown in Proposition 23.53; hence $\Sigma = |J| \, \mathrm{d}^{f-1}a$, proportional to $|J|$ at fixed increments. Solving Equation (23.87) for A gives $A \propto (|\boldsymbol{P}| \, |J|)^{-1/2}$, and on a caustic that is not also a turning surface $|\boldsymbol{P}|$ is finite and non-zero while $J \to 0$.

One bookkeeping point: the flux tube is measured in ordinary space, while Definition 23.52 parametrises the family by the Jacobi arclength. The two differ by the reparametrisation $\mathrm{d}\sigma = |\boldsymbol{P}|\,\mathrm{d}s$, whose factor is positive and finite wherever $E > V$, so the two Jacobians differ by a non-vanishing factor and have the same zero set. The caustic is the same locus in either parametrisation, which is what allows Proposition 23.53 and this proposition to speak of one J. □

Proposition 23.55 (The fold, and the quarter-power law)**.** *Let the caustic be a* fold*: transversally to it, the map from the family parameter to position is quadratic,*

$$\zeta(a) = c\,(a - a_c)^2 + \mathcal{O}\left((a - a_c)^3\right), \qquad c \neq 0\,, \tag{23.88}$$

ζ being the signed distance from the caustic. Then two members of the family reach each point with $c\zeta > 0$ and none reaches a point with $c\zeta < 0$, so that the caustic separates a doubly illuminated side from a dark side, and the geometrical amplitude of each of the two obeys

$$A \propto |\zeta|^{-1/4}\,. \tag{23.89}$$

Rests on Proposition 23.54 and Equation (23.86).

Proof. *Derives Proposition 23.55.*

Solving Equation (23.88) for a gives $a - a_c = \pm\sqrt{\zeta/c}$, real for $c\zeta > 0$ and not otherwise: hence two members on one side and none on the other. The transverse Jacobian is $\mathrm{d}\zeta/\mathrm{d}a = 2c\,(a - a_c) = \pm 2\sqrt{c\zeta}$, so $J \propto |\zeta|^{1/2}$, and Proposition 23.54 gives $A \propto |J|^{-1/2} \propto |\zeta|^{-1/4}$. □

Proposition 23.56 (The phase jump at a fold)**.** *A semiclassical wave crossing a fold caustic acquires, relative to the geometrical continuation, an extra phase $-\pi/2$. Around a closed circuit of an invariant torus the accumulated extra phase is therefore $-\mu\pi/2$, with μ the number of folds crossed, which is the Maslov index appearing in Equation (23.62).*

Rests on Theorem 9.72, Remark 9.79, and Proposition 23.55.

Derivation. *Derives Proposition 23.56.*

Transversally to the fold the problem is one-dimensional: by Proposition 23.55 the two branches merge at $\zeta = 0$ and the amplitude carries the factor $J^{-1/2}$, whose argument shifts by $-\pi/2$ when J changes sign. Which branch is selected is not a matter of taste, and is fixed by the connection formulae Theorem 9.72: a solution decaying on the dark side matches, on the illuminated side, to $2\,|\boldsymbol{P}|^{-1/2}\sin\left[\hbar^{-1}\int|\boldsymbol{P}|\,\mathrm{d}\zeta + \pi/4\right]$, whose two exponential components differ in phase from the naive continuation by $\mp\pi/2$; the outgoing one lags the incoming by $\pi/2$. Adding one such contribution per crossing around a closed circuit gives $-\mu\pi/2$, and demanding single-valuedness of the wave function around the circuit—$\hbar^{-1}\oint p\,\mathrm{d}q - \mu\pi/2 \in 2\pi\mathbb{Z}$—is Equation (23.62). Remark 9.79 records the one-dimensional instance, $\mu = 2$ for a well with two simple turning points, and Keller's extension of the count to several degrees of freedom [1150]. □

Proposition 23.57 (The wave resolution at a turning surface)**.** *Where the caustic is a classical turning surface, $V = E$ with $\boldsymbol{\nabla}V \neq \boldsymbol{0}$, the divergence Equation (23.89) is spurious: the exact solution is finite, and within a few multiples of the scale ℓ of Equation (23.85) it is the Airy function $\mathrm{Ai}\,(-\zeta/\ell)$ of Definition 9.68. The illuminated side carries fringes—the interference of the two branches of Proposition 23.55—whose maxima lie at the extrema of* Ai*, the brightest at $\zeta = 1.0188\,\ell$, and the dark side carries an exponential tail. For $\zeta \gg \ell$ the Airy asymptotics Equation (9.79) reproduce Equation (23.89) exactly.*

Rests on Remark 23.51 and Theorems 9.70 and 9.72.

Derivation. *Derives Proposition 23.57.*

Near a smooth turning surface the potential may be replaced by its tangent plane, $V \approx E + |\boldsymbol{\nabla}V|\,\zeta$, and the stationary Schrödinger equation separates into motion along the

surface, which is free at leading order, and motion normal to it, which is Equation (9.75)—the linearised turning-point equation of Chapter 9. Its solutions are the Airy functions (Proposition 9.69) on the scale Equation (9.74), which is Equation (23.85); the decaying combination on the forbidden side is Ai, and Theorem 9.72 is the statement that it matches the oscillatory semiclassical form on the allowed side. The extrema of $\mathrm{Ai}(-z)$ are the zeros of Ai', the first at $z = 1.0188$, and Equation (9.79) gives $\mathrm{Ai}(-z) \sim \pi^{-1/2} z^{-1/4} \sin\left(\frac{2}{3} z^{3/2} + \frac{\pi}{4}\right)$, whose envelope is the $z^{-1/4}$ of Equation (23.89). □

Remark 23.58 (The caustics that are not turning surfaces). Proposition 23.57 covers the caustic that is a turning surface, where the momentum itself vanishes. Most observed caustics are not of that kind: at the focus of a lens, at the cusp in a coffee cup, at the rainbow angle, the momentum is perfectly finite and it is the *geometry* of the family that folds. The fold Equation (23.88) fixes the transverse scale in that case by itself, at

$$\ell_{\mathrm{f}} = \left(\tfrac{1}{2} R\, \bar{\lambda}^2\right)^{1/3} , \tag{23.90}$$

with R the radius of curvature of the caustic and $\bar{\lambda}$ of Equation (23.84), and the same Airy profile with ζ/ℓ_{f} in place of ζ/ℓ; the peak amplitude then exceeds the smooth geometrical value by a factor of order $\left(R/\bar{\lambda}\right)^{1/6}$. This is Airy's result, obtained for the rainbow [103] and set out in full by Born and Wolf [1967].

Rests on Propositions 23.55 and 23.57.

Derivation. *Derives Equation (23.90).*

Two inputs are used and both should be named. The first is the superposition hypothesis of physical optics: near the caustic the field is written as an integral over the members of the family,

$$\psi(\zeta) = \int \mathrm{d}a\; A(a)\, \exp\!\left[\frac{\mathrm{i}\, S(a,\zeta)}{\bar{\lambda}}\right] , \tag{23.91}$$

$S(a,\zeta)$ being the optical path along the member labelled a to the field point and A a slowly varying amplitude. The reduced wavelength Equation (23.84) is constant across the small region that matters, because at a caustic of this kind $|\boldsymbol{P}|$ is finite and smooth; the phase is therefore $S/\bar{\lambda}$ with S an ordinary length. The geometrical field of Proposition 23.54 is what Equation (23.91) reduces to when the phase is stationary at isolated points, a stationary point in a being exactly a member that passes through the field point. The second input is the integral representation Equation (9.77) of the Airy function.

The phase is fixed by Equation (23.88) alone. Write $s := a - a_c$. The stationary points of S are the two members reaching ζ, which by Proposition 23.55 sit at $s = \pm\sqrt{\zeta/c}$; so $\partial S/\partial s$ vanishes exactly there and, being smooth with simple zeros, equals $\kappa\,(s^2 - \zeta/c)$ to leading order, κ being a constant of the family. Integrating,

$$S = S_0 + \kappa \left(\frac{s^3}{3} - \frac{\zeta}{c}\, s\right) . \tag{23.92}$$

The phase is cubic in the family parameter: that, and no property of the wave equation, is what a fold is.

Substituting Equation (23.92) into Equation (23.91), holding A fixed over the small range of s that contributes and rescaling by $u = \left(\kappa/\bar{\lambda}\right)^{1/3} s$,

$$\psi \propto \int \mathrm{d}u\; \exp\!\left[\mathrm{i}\left(\frac{u^3}{3} - X u\right)\right] , \qquad X := \left(\frac{\kappa}{\bar{\lambda}}\right)^{2/3} \frac{\zeta}{c} . \tag{23.93}$$

From Equations (23.88) and (23.92) (the fold makes the phase cubic, and the rescaling removes every constant from it).

Put $u = \mathrm{i}t$. The integrand becomes the entire integrand of Equation (9.77) with $z = -X$, and the ends of the contour, which run out along the imaginary t axis, may be rotated into the sectors of decay identified in Proposition 9.69 by Corollary 8.15. Hence $\psi \propto \mathrm{Ai}\,(-X)$: oscillatory on the illuminated side $X > 0$, exponentially small on the dark side, with transverse scale

$$\ell_{\mathrm{f}} = c \left(\frac{\bar{\lambda}}{\kappa} \right)^{2/3} . \tag{23.94}$$

It remains to express Equation (23.94) through the radius of curvature. Take the caustic to be locally a circle of radius R and the family to be its tangent lines, labelled by the dimensionless angle s of the point of tangency, with ζ the outward radial distance. The tangents through the point at radius $R + \zeta$ touch where $\cos s = R/\,(R + \zeta)$, so $\zeta = \tfrac{1}{2} R s^2$ and $c = R/2$. The optical path from the tangency point to the field point is $\sqrt{r^2 - R^2}$ and the arc run along the caustic is $R \arccos\,(R/r)$, so the path difference between the two tangents is, at $r = R + \zeta$ and to order $\zeta^{3/2}$,

$$\Delta S = 2 \left[\sqrt{r^2 - R^2} - R \arccos \frac{R}{r} \right] = \frac{4\sqrt{2}}{3} R \left(\frac{\zeta}{R} \right)^{3/2} , \tag{23.95}$$

whereas Equation (23.92) evaluated at its two stationary points gives $\Delta S = \tfrac{4}{3} \kappa \,(\zeta/c)^{3/2}$. Comparing, and using $c = R/2$, gives $\kappa = R/2$ as well, so that Equation (23.94) is $\ell_{\mathrm{f}} = \left(R \bar{\lambda}^2 / 2 \right)^{1/3}$, which is Equation (23.90).

The amplitude then follows without further work. The geometrical amplitude is $|\zeta|^{-1/4}$ by Equation (23.89), and the wave cuts that growth off at $\zeta \sim \ell_{\mathrm{f}}$; relative to the geometrical value a distance of order R from the caustic the peak is larger by $(R/\ell_{\mathrm{f}})^{1/4} = 2^{1/12} \left(R/\bar{\lambda} \right)^{1/6}$, which is the factor quoted above. □

Remark 23.59 (What is observed). Caustics are the sharpest observable signature of the failure of the geometrical limit, and they are visible in every one of the three languages this chapter has used. In optics they are the bright cusped curve on the bottom of a cup, the twinkling of starlight through a turbulent atmosphere, and the supernumerary bows inside a rainbow, whose spacing is the fringe spacing of Proposition 23.57 and whose explanation was Airy's [103]; the rainbow itself is a fold caustic of the family of rays refracted through a spherical drop, at a scattering angle of about 138°, and it is treated in Section 61.8.2. In mechanics they are the safety parabola of Example 23.23—the envelope beyond which no projectile of given speed can reach—and the turning surfaces at which Equation (9.69) diverges. In gravity they are the caustics of gravitational lensing: the images of a background source multiply as it crosses one, and the sharp peak of a microlensing light curve is the crossing itself [1781], the first lensed system having been identified as a pair of quasar images [1645]. What all three have in common is that the divergence is never observed: the wave theory replaces it with a finite peak of the width Equation (23.90), and measuring that width measures the wavelength.

Rests on Propositions 23.54 and 23.57.

Chapter 24

Symplectic Geometry of Phase Space

Chapter 22 assembled the coordinates and momenta into one object η^c and wrote Hamilton's equations as the single matrix equation Equation (22.57). That was a notational convenience. This chapter takes it seriously: the matrix J of Equation (22.56) is the component form of a geometric object living on phase space—a closed, nondegenerate two-form—and almost everything said about Hamiltonian systems is a statement about that form.

The payoff is that structure invisible in coordinates becomes visible. The Poisson bracket, the canonical transformations, Liouville's theorem, the Poincaré invariants of Section 22.3 and the conserved quantities attached to symmetries are not five facts but one fact seen five ways. And several quantities that a coordinate treatment has no reason to single out—the action variables, the adiabatic invariants, the geometric phases—are exactly the quantities that survive perturbation and that experiment therefore finds robust. That is the sense in which the symplectic view lets one see more in nature than the Newtonian one: it says which quantities are structural and which are accidents of the coordinates chosen to write them.

This chapter uses the manifold and differential-form machinery of Chapter 13 and the Lie-group material of Chapter 14; per the dimension rule of this treatise it is stated in general dimension, and the physical instantiation is $2f$ with f the number of degrees of freedom.

Notation 24.1. M denotes a smooth manifold of even dimension $2f$, T_xM its tangent space at x, and $\Omega^k(M)$ the space of differential k-forms. The interior product of a vector field X with a form is written ι_X—it is the operation written i_ξ in Equation (13.276)—the exterior derivative d, and the Lie derivative along X is written $\mathcal{L}_X$: note that in this chapter, and only here, the symbol $\mathcal{L}$ is the Lie derivative and not the Lagrangian, which does not appear. Latin indices $c, d = 1, \dots, 2f$ label phase-space components as in Notation 22.1, and $a, b = 1, \dots, f$ the degrees of freedom. Because ω is reserved throughout for the symplectic form, an angular velocity is written $\boldsymbol{\Omega}$ in Section 24.5 and an ordinary frequency ν in Section 24.7.

24.1 Symplectic vector spaces

Definition 24.2 (Symplectic vector space)**.** A *symplectic vector space* is a pair (V, ω) with V a real vector space and $\omega : V \times V \to \mathbb{R}$ a bilinear form that is

1. *antisymmetric*: $\omega(u, v) = -\omega(v, u)$; and
2. *nondegenerate*: $\omega(u, v) = 0$ for all v implies $u = 0$.

Rests on Proposition 5.115.

Proposition 24.3 (Even dimension and the canonical basis)**.** *A symplectic vector space has even dimension $2f$, and admits a basis $\left(e_1,\dots,e_f,f^1,\dots,f^f\right)$ in which*

$$\omega(e_a,e_b)=0\,,\qquad \omega(f^a,f^b)=0\,,\qquad \omega(e_a,f^b)=\delta_a^{\,b}\,, \tag{24.1}$$

so that the matrix of ω in that basis is the symplectic matrix Equation (22.56). All symplectic vector spaces of the same dimension are therefore isomorphic: unlike a metric, a symplectic form has no local invariants and no signature. Moreover $\mathrm{Sp}(2f,\mathbb{R})$ acts transitively on the set of such bases.

Rests on Definition 24.2, Proposition 5.115, and Equation (22.56).

Proof. *Derives Proposition 24.3.*

The first two statements are Proposition 5.115 of Chapter 5 at $2n=2f$: that proposition proves, for a nondegenerate antisymmetric bilinear form on a real vector space, that the dimension is even and that a basis exists satisfying Equation (5.143), which are Equation (24.1) with the naming used here, and in which the matrix of the form is Equation (5.144)—the matrix J of Equation (22.56). Two symplectic vector spaces of dimension $2f$ are then isomorphic by the linear map carrying one such basis to the other, which preserves ω because it preserves the matrix.

Transitivity is the same observation read inside one space. Let (e,f) and (e',f') be two bases satisfying Equation (24.1) and let M be the matrix of the change of basis. Both bases give ω the matrix J, and a change of basis transforms the matrix of a bilinear form by $[\omega]\mapsto M^{\mathsf{T}}\,[\omega]\,M$ (Equation (5.139)); hence $M^{\mathsf{T}}J\,M=J$ and $M\in\mathrm{Sp}(2f,\mathbb{R})$ by Equation (5.174). Conversely every $M\in\mathrm{Sp}(2f,\mathbb{R})$ carries a canonical basis to a canonical basis, so the action is transitive. Nothing beyond Proposition 5.115 is proved here; it is restated because the naming of the basis vectors is what fixes the block order of J used throughout this chapter. □

Definition 24.4 (Symplectic group)**.** The *symplectic group* $\mathrm{Sp}(2f,\mathbb{R})$ is the group of linear maps $M:V\to V$ preserving ω; in the canonical basis it is the set of $2f\times 2f$ real matrices satisfying

$$M^{\mathsf{T}}J\,M=J\,. \tag{24.2}$$

This is Definition 5.130 of Chapter 5, restated in the notation of this chapter; it is not a second definition.

Rests on Definition 5.130 and Proposition 24.3.

Proposition 24.5 (Properties of the symplectic group)**.** $\mathrm{Sp}(2f,\mathbb{R})$ *is a Lie group of dimension $f(2f+1)$ whose Lie algebra $\mathfrak{sp}(2f,\mathbb{R})$ consists of the matrices A with $A^{\mathsf{T}}J+JA=0$, equivalently $A=JS$ with S symmetric; every element has determinant $+1$; and the group is noncompact for every $f\geq 1$.*

Rests on Definition 24.4, Proposition 5.134, and Theorem 5.133.

Proof. *Derives Proposition 24.5.*

The dimension, the Lie-algebra condition and the fact that $\mathfrak{sp}(2f,\mathbb{R})$ is a Lie algebra are Proposition 5.134 with Equation (5.178) at $n=f$; the determinant is Theorem 5.133. That $A^{\mathsf{T}}J+JA=0$ is the same as $A=JS$ with S symmetric follows by substituting $A=JS$ and using $J^{\mathsf{T}}=-J$, $J^2=-\mathbb{1}_{2f}$ from Equation (22.56):

$$(JS)^{\mathsf{T}}\,J+J\,(JS)=S^{\mathsf{T}}J^{\mathsf{T}}J-S=S^{\mathsf{T}}-S\,,$$

which vanishes exactly when S is symmetric; and every A arises this way with $S=-JA$.

Noncompactness is proved by exhibiting an unbounded one-parameter family. For $\lambda > 0$ put

$$M_\lambda = \begin{pmatrix} \lambda \mathbb{1}_f & 0 \\ 0 & \lambda^{-1}\mathbb{1}_f \end{pmatrix}, \qquad J\,M_\lambda = \begin{pmatrix} 0 & \lambda^{-1}\mathbb{1}_f \\ -\lambda\mathbb{1}_f & 0 \end{pmatrix},$$

so that M_λ^{T}, which multiplies the first block row by λ and the second by λ^{-1}, gives $M_\lambda^{\mathsf{T}}(J\,M_\lambda) = J$: each M_λ lies in $\mathrm{Sp}(2f,\mathbb{R})$. Its entries are unbounded as $\lambda \to \infty$, so $\mathrm{Sp}(2f,\mathbb{R})$ is not a bounded subset of the $4f^2$-dimensional space of matrices and hence, by Theorem 6.12, not compact. Physically this family is the squeeze that trades position spread against momentum spread at constant phase area, and it is the reason a symplectic map can distort a region without limit while preserving its volume. □

Remark 24.6 (Connectedness, and what it rests on). $\mathrm{Sp}(2f,\mathbb{R})$ is also connected, and its maximal compact subgroup is $\mathrm{U}(f)$; the standard argument is the polar decomposition $M = OP$ of a real invertible matrix into an orthogonal and a positive-definite symmetric factor, followed by the observation that both factors of a symplectic matrix are themselves symplectic and that the positive factor can be contracted to the identity along P^t, $t \in [0,1]$. The polar decomposition is a statement of linear algebra and belongs in Chapter 5, which does not at present carry it; the claim is therefore recorded here as one this treatise states without proving. Nothing in this chapter uses it: connectedness enters only when one asks whether a symplectic matrix can be reached continuously from the identity, which is a question about generating canonical transformations by flows and is not needed for any result below.

Rests on Proposition 24.5.

Remark 24.7 (Contrast with the orthogonal group). A metric of signature (p,q) gives the group $\mathrm{SO}(p,q)$ of Chapter 39, of dimension $\frac{1}{2}n(n-1)$ with $n = p+q$; a symplectic form in the same dimension $n = 2f$ gives a group of dimension $f(2f+1) = \frac{1}{2}n(n+1)$, which is larger. Phase space has more symmetry than spacetime, not less, and that is why so much can be transformed away in it.

Rests on Proposition 24.5.

24.2 Symplectic manifolds

Definition 24.8 (Symplectic manifold)**.** A *symplectic manifold* is a pair (M,ω) with M a smooth manifold of dimension $2f$ and $\omega \in \Omega^2(M)$ a two-form that is

1. *closed*: $\mathrm{d}\omega = 0$; and
2. *nondegenerate*: ω_x is a symplectic form on T_xM for every $x \in M$.

This is Definition A.67, restated here because the whole chapter is about it.

Rests on Definitions 13.98, 13.105 and 24.2.

Definition 24.9 (The canonical form on a cotangent bundle)**.** Let Q be the configuration manifold of a mechanical system and $M = T^*Q$ its cotangent bundle—phase space. In coordinates (q^a, p_a) adapted to the bundle, the *canonical one-form* (also called the tautological, Liouville, or symplectic potential) and the *canonical two-form* are

$$\theta = p_a\,\mathrm{d}q^a\,, \qquad \omega = -\mathrm{d}\theta = \mathrm{d}q^a \wedge \mathrm{d}p_a\,. \tag{24.3}$$

Rests on Definitions 13.103 and 24.8.

Proposition 24.10 (The canonical form is symplectic and intrinsic)**.** *The form ω of Equation (24.3) is closed and nondegenerate, and θ is defined without reference to coordinates by $\theta_\xi(v) = \xi(\mathrm{d}\pi(v))$ for $\xi \in T^*Q$ and $v \in T_\xi(T^*Q)$, with $\pi : T^*Q \to Q$ the projection. Phase*

space therefore carries a symplectic structure with no further input: no metric, no connection, and no choice of coordinates.

Rests on Definitions 13.52 and 24.9 and Lemma 13.107.

Proof. *Derives Proposition 24.10.*

Closed. $\omega = -\mathrm{d}\theta$ is exact, hence closed by Lemma 13.107; equivalently $\mathrm{d}\omega = -\mathrm{d}^2\theta = 0$ by Equation (13.245). The second equality in Equation (24.3) is the Leibniz rule for d: $\mathrm{d}\,(p_a \mathrm{d}q^a) = \mathrm{d}p_a \wedge \mathrm{d}q^a = -\mathrm{d}q^a \wedge \mathrm{d}p_a$.

Nondegenerate. In the coordinate basis $(\partial/\partial q^a, \partial/\partial p_a)$ of $T_\xi(T^*Q)$ one reads off from $\omega = \mathrm{d}q^a \wedge \mathrm{d}p_a$ that

$$\omega\left(\frac{\partial}{\partial q^a}, \frac{\partial}{\partial q^b}\right) = 0\,,$$
$$\omega\left(\frac{\partial}{\partial p_a}, \frac{\partial}{\partial p_b}\right) = 0\,,$$
$$\omega\left(\frac{\partial}{\partial q^a}, \frac{\partial}{\partial p_b}\right) = \delta_a^b\,,$$

so the matrix of ω_ξ is J, which is invertible because $J^2 = -\mathbb{1}_{2f}$ by Equation (22.56). A bilinear form with invertible matrix is nondegenerate, and the basis above is a canonical basis in the sense of Equation (24.1).

Intrinsic. Let $\xi \in T^*Q$ lie over $q = \pi(\xi)$ and choose an adapted chart, so that $\xi = p_a\,\mathrm{d}q^a\big|_q$ and a tangent vector $v \in T_\xi(T^*Q)$ has components $v = v_q^a\,\partial/\partial q^a + v_{p,a}\,\partial/\partial p_a$. The projection π forgets the momenta, so $\mathrm{d}\pi(v) = v_q^a\,\partial/\partial q^a\big|_q$, and therefore

$$\xi(\mathrm{d}\pi(v)) = p_a v_q^a = (p_a\,\mathrm{d}q^a)\,(v) = \theta_\xi(v)\,.$$

The left-hand side is built from ξ, v and π alone, none of which mentions a chart; the right-hand side is the coordinate expression Equation (24.3). The two agree in every adapted chart, so θ—and with it $\omega = -\mathrm{d}\theta$—is chart-independent. That the coordinate formula reproduces the definition by evaluating the covector on its own base point is what earns θ the name *tautological*. □

Remark 24.11 (The SI dimension of every object in this chapter). Each product $p_a q^a$ has the dimension of action, and every symplectic object is built from such products. So θ and ω carry the unit J s; the Liouville form Equation (24.13) carries $(\mathrm{J\,s})^f$; a symplectic capacity (Definition 24.33) carries J s; and the Poisson bracket of two quantities u and v carries $[u][v]$ divided by J s, which is why $\{q^a, p_b\}$ in Equation (22.44) is a pure number. A Hamiltonian vector field $X_\mathcal{H}$ built from an energy therefore carries 1/s: it is a rate, as its integral curves being trajectories requires. Two consequences are used later. First, $\omega/\hbar$ is dimensionless, which is what allows it to be the curvature of a connection in Section 24.7; $\hbar$ is the only constant of nature with the right unit to do that job, and Remark 5.116 already notes the coincidence. Second, the column η^c of Equation (22.55) mixes quantities of different dimension—the point of Remark 22.5—so no Euclidean length on phase space is dimensionally meaningful without a declared scale, a point that returns in Remark 24.27.

Rests on Proposition 24.10 and Remark 22.5.

Theorem 24.12 (Darboux)**.** *Every point of a symplectic manifold (M, ω) has a neighbourhood with coordinates (q^a, p_a) in which*

$$\omega = \mathrm{d}q^a \wedge \mathrm{d}p_a\,. \tag{24.4}$$

Rests on Definition 24.8 and Theorem A.68.

Full derivation: Appendix A.12 (p. 2807).
Derives Theorem 24.12.

Theorem 24.12 is Theorem A.68, proved in Appendix A.12 as part of Chapter 5 and not reproved here. The argument there is Moser's deformation trick [1317]: connect ω to its constant-coefficient value at the point by a path of closed nondegenerate forms, use nondegeneracy to solve for the vector field generating the isotopy, and integrate it. Its two hypotheses do exactly one job each—nondegeneracy supplies the vector field, closedness supplies the primitive through the local converse of the Poincaré lemma (Lemma A.77)—and the coordinates it produces are the canonical coordinates of Chapter 22. Corollary A.69 is the statement that any two symplectic manifolds of equal dimension are locally symplectomorphic.

Remark 24.13 (What Darboux's theorem forbids). Darboux's theorem is the exact statement that a symplectic manifold has no local geometry. There is no symplectic analogue of curvature, no invariant distinguishing one point from another, and hence no local measurement that could reveal the symplectic structure of phase space. All symplectic invariants are global, which is why the theorems of this chapter that carry physical content—Liouville, Poincaré recurrence, the Liouville–Arnold tori of Theorem 23.31, the non-squeezing theorem—are statements about whole trajectories or whole regions and never about a neighbourhood of a point. The contrast with the Riemannian case of Chapter 43, where curvature is a local observable and tidal forces measure it, is total.
Rests on Theorem 24.12 and Corollary A.69.

24.3 Hamiltonian vector fields and the flow

Definition 24.14 (Hamiltonian vector field)**.** Let $H \in C^\infty(M)$. Since ω is nondegenerate, there is a unique vector field X_H with

$$\iota_{X_H}\omega = \mathrm{d}H\,. \tag{24.5}$$

It is the *Hamiltonian vector field* of H. In Darboux coordinates

$$X_H = \frac{\partial H}{\partial p_a}\frac{\partial}{\partial q^a} - \frac{\partial H}{\partial q^a}\frac{\partial}{\partial p_a}\,, \tag{24.6}$$

so that its integral curves are exactly the solutions of Hamilton's equations Equations (22.3) and (22.4).
Rests on Definition 24.8 and Equations (22.3) and (22.4).

The coordinate form Equation (24.6) is a one-line check and is worth doing once, because every sign in this chapter descends from it. Write $X = A^a\,\partial/\partial q^a + B_a\,\partial/\partial p_a$. Contracting $\omega = \mathrm{d}q^a \wedge \mathrm{d}p_a$ with X and using the graded Leibniz rule for ι_X,

$$\iota_X\omega = (\iota_X \mathrm{d}q^a)\,\mathrm{d}p_a - \mathrm{d}q^a\,(\iota_X \mathrm{d}p_a) = A^a\,\mathrm{d}p_a - B_a\,\mathrm{d}q^a\,,$$

while $\mathrm{d}H = (\partial H/\partial q^a)\,\mathrm{d}q^a + (\partial H/\partial p_a)\,\mathrm{d}p_a$. Matching the coefficients of $\mathrm{d}p_a$ gives $A^a = \partial H/\partial p_a$ and matching those of $\mathrm{d}q^a$ gives $B_a = -\partial H/\partial q^a$, which is Equation (24.6); these are Equations (22.3) and (22.4) read as the components of a vector field, and they are Equation (22.57) with the matrix J replaced by the form whose matrix it is.

Proposition 24.15 (Poisson bracket from the symplectic form)**.** *With the conventions Equation (24.3) and Equation (24.5), the Poisson bracket of Definition 22.29 is*

$$\{u, v\} = \omega\,(X_u, X_v) = X_v[u]\,, \tag{24.7}$$

and the assignment $u \mapsto X_u$ satisfies

$$X_{\{u,v\}} = -\left[X_u, X_v\right] , \tag{24.8}$$

the bracket on the right being the commutator of vector fields.
Rests on Definitions 22.29 and 24.14 and Equation (22.46).

Proof. *Derives Proposition 24.15.*

The bracket. By Equation (24.5), $\omega(X_u, X_v) = (\iota_{X_u}\omega)(X_v) = \mathrm{d}u(X_v) = X_v[u]$, which is the second equality in Equation (24.7) and holds on any symplectic manifold. For the first, work in Darboux coordinates and use Equation (24.6):

$$\begin{aligned} X_v[u] &= \frac{\partial v}{\partial p_a}\frac{\partial u}{\partial q^a} - \frac{\partial v}{\partial q^a}\frac{\partial u}{\partial p_a} \\ &= \frac{\partial u}{\partial q^a}\frac{\partial v}{\partial p_a} - \frac{\partial u}{\partial p_a}\frac{\partial v}{\partial q^a} , \end{aligned}$$

which is Equation (22.41). In particular $\dot{u} = X_{\mathcal{H}}[u] = \{u, \mathcal{H}\}$ along the flow, so the geometric definition reproduces the evolution law of Chapter 22 with no change of sign, and $\{q^a, p_b\} = \delta^a_b$ as in Equation (22.44).

The homomorphism. Two vector fields are equal when they agree as derivations on every smooth function, so let $w \in C^\infty(M)$ be arbitrary. Using $X_v[w] = \{w, v\}$ twice,

$$\begin{aligned} \left[X_u, X_v\right][w] &= X_u[X_v[w]] - X_v[X_u[w]] \\ &= X_u[\{w, v\}] - X_v[\{w, u\}] = \{\{w, v\}, u\} - \{\{w, u\}, v\} . \end{aligned}$$

On the other side, $X_{\{u,v\}}[w] = \{w, \{u, v\}\}$, and the Jacobi identity Equation (22.46) in the form $\{w, \{u, v\}\} + \{u, \{v, w\}\} + \{v, \{w, u\}\} = 0$ gives, after using antisymmetry on the second term,

$$\begin{aligned} X_{\{u,v\}}[w] &= \{u, \{w, v\}\} - \{v, \{w, u\}\} \\ &= -\{\{w, v\}, u\} + \{\{w, u\}, v\} = -\left[X_u, X_v\right][w] . \end{aligned}$$

Since w was arbitrary, Equation (24.8) follows. The minus sign is not removable: it is the statement that $u \mapsto X_u$ is an *anti*homomorphism of Lie algebras with the conventions fixed above, and reversing it would require reversing the sign in either Equation (24.5) or Equation (22.41). □

Theorem 24.16 (The Hamiltonian flow preserves the symplectic form)**.** *Let ϕ_t be the flow of X_H. Then*

$$\mathcal{L}_{X_H}\omega = 0 , \qquad \textit{equivalently} \qquad \phi_t^*\omega = \omega \quad \textit{for all } t . \tag{24.9}$$

Rests on Definitions 24.8 and 24.14 and Proposition 13.128.

Proof. *Derives Theorem 24.16.*

Cartan's magic formula Equation (13.276) applied to ω and the field X_H reads

$$\mathcal{L}_{X_H}\omega = \mathrm{d}\left(\iota_{X_H}\omega\right) + \iota_{X_H}\left(\mathrm{d}\omega\right) .$$

The first term is $\mathrm{d}\left(\mathrm{d}H\right) = 0$ by Equation (24.5) and the nilpotency Equation (13.245); the second is zero because ω is closed. Each hypothesis of Definition 24.8 is used exactly once, which is why the definition demands both and no more.

For the second form, the Lie derivative is by Equation (13.267) the derivative of the dragged form, so that $\frac{\mathrm{d}}{\mathrm{d}t}\phi_t^*\omega = \phi_t^*\left(\mathcal{L}_{X_H}\omega\right) = 0$; hence $\phi_t^*\omega$ is independent of t and equals its value $\phi_0^*\omega = \omega$ at $t = 0$. □

Definition 24.17 (Symplectomorphism)**.** A diffeomorphism $\psi : M \to M$ with $\psi^*\omega = \omega$ is a *symplectomorphism*, or *canonical transformation*.

Rests on Definitions 24.8 and A.70.

Theorem 24.18 (Symplectic form of the transformation condition)**.** *Let $\eta \mapsto \zeta(\eta)$ be a transformation of phase space in the notation of Equation (22.55), with Jacobian matrix*

$$M^c{}_d = \frac{\partial \zeta^c}{\partial \eta^d}\,. \tag{24.10}$$

The transformation is canonical if and only if

$$M^{\mathsf{T}} J\, M = J\,, \tag{24.11}$$

that is, if and only if its Jacobian lies in $\mathrm{Sp}(2f, \mathbb{R})$ at every point. This single matrix identity is equivalent to the $4f^2$ direct conditions of Proposition 22.17, to the invariance of the fundamental Poisson brackets Equation (22.44), and to the invariance of the fundamental Lagrange brackets Equation (22.39).

Rests on Definition 24.17 and Equations (22.42) and (22.57).

Proof. *Derives Theorem 24.18.*

The condition. Let the transformation be time-independent and invertible, write $\mathcal{H}'(\zeta) = \mathcal{H}(\eta(\zeta))$ for the same energy read in the new variables, and let $\mathcal{H}$ be arbitrary. Differentiating $\zeta = \zeta(\eta)$ along a trajectory and using Equation (22.57) for η,

$$\dot\zeta^c = M^c{}_d\, \dot\eta^d = M^c{}_d\, J^d{}_e\, \frac{\partial \mathcal{H}}{\partial \eta^e}\,, \qquad \frac{\partial \mathcal{H}}{\partial \eta^e} = M^g{}_e\, \frac{\partial \mathcal{H}'}{\partial \zeta^g}\,,$$

the second being the chain rule. In matrix form $\dot\zeta = M\, J\, M^{\mathsf{T}}\, \partial\mathcal{H}'/\partial\zeta$; since $\partial\mathcal{H}'/\partial\zeta$ may be prescribed arbitrarily at any one point by choosing $\mathcal{H}$, the new variables obey Hamilton's equations with $\mathcal{H}'$ for every $\mathcal{H}$ if and only if

$$M\, J\, M^{\mathsf{T}} = J\,. \tag{24.12}$$

That Equation (24.12) and Equation (24.11) are equivalent is algebra. J is invertible, so either identity forces $\det M \neq 0$. Assume Equation (24.12); multiplying on the left by M^{-1} and on the right by $\left(M^{\mathsf{T}}\right)^{-1}$ gives $J = M^{-1} J \left(M^{\mathsf{T}}\right)^{-1}$, and inverting both sides with $J^{-1} = -J$ gives $-J = M^{\mathsf{T}}\,(-J)\,M$, which is Equation (24.11). The same steps run backwards.

The three equivalent tests. All three are Equation (24.12) or Equation (24.11) read in a different bookkeeping. By Equation (22.42) the Poisson bracket of the new variables, computed in the old ones, is

$$\left\{\zeta^c, \zeta^d\right\}_{q,p} = \left(\frac{\partial \zeta^c}{\partial \eta}\right)^{\mathsf{T}} J \left(\frac{\partial \zeta^d}{\partial \eta}\right) = \left(M\, J\, M^{\mathsf{T}}\right)^{cd}\,,$$

so the new variables satisfy the fundamental Poisson brackets Equation (22.44) exactly when Equation (24.12) holds. The Lagrange brackets of Equation (22.36) differentiate the other way round, $\partial\eta/\partial\zeta = M^{-1}$, and give $\left[\zeta^c, \zeta^d\right]_{q,p} = \left((M^{-1})^{\mathsf{T}}\, J M^{-1}\right)^{cd}$, which equals J exactly when Equation (24.11) holds for M^{-1}, hence for M. That the two families of brackets test the transposed identities is the duality recorded in Proposition 22.32. Finally, splitting M into the four $f \times f$ blocks of Equation (22.55) and multiplying out Equation (24.11) block by block returns the four families Equations (22.16) and (22.17), which are $4f^2$ scalar equations; this is Lemma 22.20, and it is what discharges the heading reserved without content in Section 22.5.1.

Coordinate-free reading. Nothing above is more than Definition 24.17 in components: $\psi^*\omega = \omega$ says that the matrix of ω is unchanged when transported by the Jacobian, and the matrix of ω in canonical coordinates is J. □

Corollary 24.19 (Canonical transformations preserve phase volume)**.** $\det M = 1$ *for every canonical transformation, so the phase-space volume element is invariant.*

Rests on Theorems 5.133 and 24.18.

Proof. *Derives Corollary 24.19.*

By Theorem 24.18 the Jacobian lies in $\mathrm{Sp}(2f, \mathbb{R})$ at every point, and every element of that group has determinant $+1$ by Theorem 5.133. The volume element transforms as $\mathrm{d}^{2f}\zeta = (\det M)\,\mathrm{d}^{2f}\eta = \mathrm{d}^{2f}\eta$. Note that taking determinants directly in Equation (24.11) yields only $(\det M)^2 = 1$; the sign is settled by the Pfaffian argument of Theorem 5.133 and not by the naive route. □

24.4 Invariants of the flow

Definition 24.20 (Liouville volume)**.** The *Liouville form* of a symplectic manifold is the $2f$-form

$$\Omega = \frac{1}{f!}\,\omega^{\wedge f} = \mathrm{d}q^1 \wedge \mathrm{d}p_1 \wedge \cdots \wedge \mathrm{d}q^f \wedge \mathrm{d}p_f \tag{24.13}$$

$$= (-1)^{f(f-1)/2}\,\mathrm{d}q^1 \wedge \cdots \wedge \mathrm{d}q^f \wedge \mathrm{d}p_1 \wedge \cdots \wedge \mathrm{d}p_f\,, \tag{24.14}$$

the last two expressions holding in Darboux coordinates. It is a volume form, and every symplectic manifold is therefore orientable.

Rests on Definitions 13.102 and 24.8 and Theorem 24.12.

The two coordinate expressions in Equation (24.13) differ by the sign of the permutation that separates the coordinates from the momenta, and the sign matters only for the choice of orientation, never for a volume. Writing $\omega = \sum_a \omega_a$ with $\omega_a = \mathrm{d}q^a \wedge \mathrm{d}p_a$, each ω_a is a two-form and any two of them commute under $\wedge$, while $\omega_a \wedge \omega_a = 0$; expanding the f-th power therefore leaves only the $f!$ orderings of the distinct factors, so $\omega^{\wedge f} = f!\,\omega_1 \wedge \cdots \wedge \omega_f$, which is the first equality. Regrouping into all the $\mathrm{d}q$ followed by all the $\mathrm{d}p$ requires $f(f-1)/2$ transpositions, which is the second.

Theorem 24.21 (Liouville)**.** *The Hamiltonian flow preserves the Liouville volume:*

$$\mathcal{L}_{X_H}\Omega = 0\,, \qquad \textit{equivalently} \qquad \operatorname{div} X_H = 0\,. \tag{24.15}$$

The volume of any region of phase space is unchanged as the region is carried along by the motion, however violently the region is distorted.

Rests on Definition 24.20, Theorem 24.16, and Proposition 7.73.

Proof. *Derives Theorem 24.21.*

Geometric route. The Lie derivative is a derivation of the exterior algebra, so it obeys the Leibniz rule on wedge products; applied to $\Omega = \omega^{\wedge f}/f!$ this gives

$$\mathcal{L}_{X_H}\Omega = \frac{1}{f!}\sum_{k=1}^{f} \omega^{\wedge(k-1)} \wedge (\mathcal{L}_{X_H}\omega) \wedge \omega^{\wedge(f-k)}$$

$$= \frac{1}{(f-1)!}\,(\mathcal{L}_{X_H}\omega) \wedge \omega^{\wedge(f-1)}\,,$$

using that two-forms commute under $\wedge$, and this vanishes because $\mathcal{L}_{X_H}\omega = 0$ by Theorem 24.16. The Liouville form is built from ω and nothing else, so it inherits every invariance ω has.

Coordinate route. For a volume form Ω the Lie derivative of a vector field satisfies $\mathcal{L}_X\Omega = (\operatorname{div} X)\,\Omega$, which in Darboux coordinates is the ordinary divergence. From Equation (24.6),

$$\begin{aligned}\operatorname{div} X_H &= \frac{\partial}{\partial q^a}\left(\frac{\partial H}{\partial p_a}\right) + \frac{\partial}{\partial p_a}\left(-\frac{\partial H}{\partial q^a}\right)\\ &= \frac{\partial^2 H}{\partial q^a \partial p_a} - \frac{\partial^2 H}{\partial p_a \partial q^a} = 0\,,\end{aligned}$$

the two terms cancelling term by term by the equality of mixed partial derivatives, Proposition 7.73. Note where the antisymmetry does its work: an arbitrary vector field on phase space has no reason to be divergence-free, and it is the opposite signs in Equation (24.6)—that is, the antisymmetry of J—that force the cancellation. □

Remark 24.22 (Liouville's theorem is the foundation of statistical mechanics). Theorem 24.21 is what makes the phase-space volume, and not some other measure, the natural counting measure for a mechanical system; the microcanonical ensemble of Chapter 118 is the uniform measure on an energy shell with respect to it, and it is distinguished by being invariant under the very dynamics whose long-time behaviour it is used to describe. The entropy so defined inherits the additive constant that only quantum mechanics fixes, through the cell size h^f—which is dimensionally the only thing it could be, by Remark 24.11. That this constant is measurable—in the Sackur–Tetrode equation for the absolute entropy of a monatomic gas [380, 393]—is one of the places where phase-space geometry becomes an experimental number.

Rests on Theorem 24.21 and Remark 24.11.

24.4.1 Integral invariants

Theorem 24.23 (Poincaré–Cartan integral invariant)**.** *Let γ_1 and γ_2 be two closed curves encircling the same tube of trajectories in extended phase space (q^a, p_a, t). Then*

$$\oint_{\gamma_1} (p_a\,\mathrm{d}q^a - \mathcal{H}\,\mathrm{d}t) = \oint_{\gamma_2} (p_a\,\mathrm{d}q^a - \mathcal{H}\,\mathrm{d}t)\ . \tag{24.16}$$

If both curves lie at a fixed instant, this reduces to the invariance of $\oint p_a\,\mathrm{d}q^a$, and hence, by Stokes' theorem, to Poincaré's theorem Theorem 22.22.

Rests on Equations (22.3) and (24.3) and Theorem A.311.

Proof. *Derives Theorem 24.23.*

Work on extended phase space $\widetilde{M} = M \times \mathbb{R}$ with coordinates (q^a, p_a, t) and set $\lambda = p_a\,\mathrm{d}q^a - \mathcal{H}\,\mathrm{d}t$, so that $\mathrm{d}\lambda = \mathrm{d}p_a \wedge \mathrm{d}q^a - \mathrm{d}\mathcal{H} \wedge \mathrm{d}t$, that is

$$\mathrm{d}\lambda = \mathrm{d}p_a \wedge \mathrm{d}q^a - \frac{\partial \mathcal{H}}{\partial q^a}\,\mathrm{d}q^a \wedge \mathrm{d}t - \frac{\partial \mathcal{H}}{\partial p_a}\,\mathrm{d}p_a \wedge \mathrm{d}t\,,$$

the term in $\partial\mathcal{H}/\partial t$ dropping out because $\mathrm{d}t \wedge \mathrm{d}t = 0$. The trajectories of the system are the integral curves of

$$Z = \frac{\partial}{\partial t} + \frac{\partial \mathcal{H}}{\partial p_a}\frac{\partial}{\partial q^a} - \frac{\partial \mathcal{H}}{\partial q^a}\frac{\partial}{\partial p_a}\,,$$

that is, of $X_{\mathcal{H}}$ carried along at unit rate in t; this is Equation (24.6) together with $\dot{t} = 1$. Contract $\mathrm{d}\lambda$ with Z, using $\iota_Z(\alpha \wedge \beta) = (\iota_Z\alpha)\,\beta - \alpha\,(\iota_Z\beta)$ on one-forms α, β:

$$\begin{aligned}\iota_Z(\mathrm{d}p_a \wedge \mathrm{d}q^a) &= -\frac{\partial \mathcal{H}}{\partial q^a}\,\mathrm{d}q^a - \frac{\partial \mathcal{H}}{\partial p_a}\,\mathrm{d}p_a\,,\\ -\frac{\partial \mathcal{H}}{\partial q^a}\,\iota_Z(\mathrm{d}q^a \wedge \mathrm{d}t) &= -\frac{\partial \mathcal{H}}{\partial q^a}\frac{\partial \mathcal{H}}{\partial p_a}\,\mathrm{d}t + \frac{\partial \mathcal{H}}{\partial q^a}\,\mathrm{d}q^a\,,\\ -\frac{\partial \mathcal{H}}{\partial p_a}\,\iota_Z(\mathrm{d}p_a \wedge \mathrm{d}t) &= +\frac{\partial \mathcal{H}}{\partial p_a}\frac{\partial \mathcal{H}}{\partial q^a}\,\mathrm{d}t + \frac{\partial \mathcal{H}}{\partial p_a}\,\mathrm{d}p_a\,.\end{aligned}$$

Adding the three lines, the $\mathrm{d}q^a$ terms cancel, the $\mathrm{d}p_a$ terms cancel, and the two $\mathrm{d}t$ terms cancel against each other:

$$\iota_Z \, \mathrm{d}\lambda = 0 \,. \tag{24.17}$$

So the trajectory direction lies in the kernel of $\mathrm{d}\lambda$. Since $\mathrm{d}\lambda$ is a two-form on a space of odd dimension $2f+1$ its kernel is at least one-dimensional; Equation (24.17) identifies it as the direction of the motion.

Now let T be the tube swept by the trajectories through a closed curve, a two-dimensional surface in $\widetilde{M}$, and let γ_1 and γ_2 be two closed curves on T each encircling it once, with $\Sigma \subset T$ the annulus they bound, so that $\partial\Sigma = \gamma_1 - \gamma_2$. At every point of Σ the tangent plane is two-dimensional and contains Z; if v is any second tangent vector there, then $\mathrm{d}\lambda(Z, v) = 0$ by Equation (24.17), so the pullback of $\mathrm{d}\lambda$ to Σ vanishes identically. The general Stokes theorem Theorem A.311 then gives

$$\oint_{\gamma_1} \lambda - \oint_{\gamma_2} \lambda = \int_{\partial\Sigma} \lambda = \int_{\Sigma} \mathrm{d}\lambda = 0 \,,$$

which is Equation (24.16).

If both curves lie in a surface of constant t, then $\mathrm{d}t$ pulls back to zero on them and λ reduces to $\theta = p_a \mathrm{d}q^a$; the statement becomes the invariance of $\oint p_a \, \mathrm{d}q^a$ under the flow, and applying Theorem A.311 once more turns that into the invariance of $\int \omega$ over a surface, which is Theorem 22.22. That the first Poincaré invariant is the only independent one is the content of Equation (24.13): the higher invariants are integrals of $\omega^{\wedge k}$ and follow from the first. □

Theorem 24.24 (Poincaré recurrence)**.** *Let the Hamiltonian flow leave invariant a region $D \subset M$ of finite Liouville volume—as it does, for instance, for bounded motion at fixed energy. Then for every neighbourhood $U \subset D$, almost every point of U returns to U at arbitrarily late times.*

Rests on Theorem 24.21 and Definition 24.20.

Proof. *Derives Theorem 24.24.*

Write μ for the Liouville volume of Definition 24.20, so that $\mu(D) < \infty$, and fix a time step $\tau > 0$; let $\phi = \phi_\tau$ be the time-τ map of the flow. By Theorem 24.21 ϕ preserves μ, and it maps D to itself.

Fix an integer $k \geq 1$ and let

$$B_k = \left\{ x \in U \;:\; \phi^{nk}(x) \notin U \text{ for every integer } n \geq 1 \right\}$$

be the set of points of U that never return under the map ϕ^k. The sets $\phi^{-nk}(B_k)$, $n = 0, 1, 2, \ldots$, are pairwise disjoint. Indeed, suppose x lay in both $\phi^{-nk}(B_k)$ and $\phi^{-mk}(B_k)$ with $n < m$. Then $y = \phi^{nk}(x) \in B_k \subset U$ while $\phi^{(m-n)k}(y) = \phi^{mk}(x) \in B_k \subset U$, so y returns to U after $m - n \geq 1$ steps of ϕ^k, contradicting $y \in B_k$.

Each of these sets lies in D and has the same volume $\mu(B_k)$, because ϕ preserves μ. If $\mu(B_k) > 0$ then the volume of the disjoint union of the first N of them is $N\mu(B_k)$, which exceeds $\mu(D)$ for N large enough—impossible, since they all sit inside D. Hence $\mu(B_k) = 0$ for every k.

Finally, a point of U that returns for arbitrarily large multiples of τ is one lying outside every B_k: if $x \notin B_k$ then x returns to U at some time that is a positive multiple of $k\tau$, and letting k run makes those return times unbounded. The exceptional set is $\bigcup_{k \geq 1} B_k$, a countable union of sets of volume zero and therefore of volume zero. So almost every point of U returns at arbitrarily late times [247]. □

Remark 24.25 (What the recurrence proof assumes about volume). The argument just given treats the Liouville volume as a *measure*: it uses additivity over disjoint sets, monotonicity, invariance under a volume-preserving map, and—in the last step only—the fact that a countable union of null sets is null. Those are the axioms of a measure space, and this treatise does not carry measure theory: the probability chapter of Part II is built on densities and expectations and never defines a σ-algebra or a countably additive measure. The statement is therefore made here on that understanding, and the reader who wants the last step in full rigour needs the countable additivity that Part II owes. Nothing else in this chapter uses it; and the physically relevant half of the theorem—that a positive-volume set of initial conditions must return—needs only finite additivity, which is the elementary property of volume used everywhere else in this part.

Rests on Theorem 24.24.

Remark 24.26 (Recurrence and irreversibility). Theorem 24.24 says that a bounded mechanical system returns arbitrarily close to its initial state, which appears to contradict the second law of Chapter 54. It does not: the recurrence times for a macroscopic number of degrees of freedom exceed the age of the universe by an enormous factor, and the theorem is silent about them. The tension it does expose—between a time-reversible microscopic dynamics and an irreversible phenomenology—is Loschmidt's and Zermelo's objection, and it is treated where it belongs, in Chapter 118.

Rests on Theorem 24.24.

24.4.2 Symplectic capacity and non-squeezing

Liouville's theorem constrains the volume of a phase-space region and nothing else. It is natural to ask whether that is the whole story— whether any volume-preserving deformation of a region can be realised by a Hamiltonian flow. It cannot, and the obstruction is not a refinement of volume but an invariant of a different kind. This subsection states it. It is written as mathematics: this treatise records no experiment that measures a symplectic capacity, and the frequently drawn analogy with the uncertainty relation of Chapter 77 is suggestive rather than derivational, a point made precise in Remark 24.35.

Remark 24.27 (Fixing a scale, so that a "ball" means something). A ball in phase space is not a dimensionally meaningful object: by Remark 24.11 a coordinate and its conjugate momentum carry different units, and no sum of their squares is a length. To speak of balls and cylinders one must first declare a scale. Fix a length $\ell > 0$ and an action $S > 0$ and pass to the dimensionless coordinates

$$u^a = \frac{q^a}{\ell}\,, \qquad v_a = \frac{\ell\, p_a}{S}\,, \qquad \text{so that} \qquad \omega = S\,\mathrm{d}u^a \wedge \mathrm{d}v_a\,. \tag{24.18}$$

All statements below are made for the standard form $\omega_0 = \mathrm{d}u^a \wedge \mathrm{d}v_a$ on $\mathbb{R}^{2f}$, and are converted back by the factor S; every capacity then comes out as a pure number times S, that is, in J s. Different choices of (ℓ, S) give different balls, so the theorem below compares a ball and a cylinder built with the *same* choice. Write $\langle \cdot, \cdot \rangle$ for the Euclidean inner product in these coordinates and $|\cdot|$ for its norm; then, with J the matrix Equation (22.56) and $z, w \in \mathbb{R}^{2f}$ written as columns in the block order of Equation (22.55),

$$\omega_0(z, w) = \langle z, J\, w \rangle\,, \qquad J^{\mathsf{T}} J = \mathbb{1}_{2f}\,, \qquad \langle J\, z, z \rangle = 0\,. \tag{24.19}$$

The last two follow from $J^{\mathsf{T}} = -J$ and $J^2 = -\mathbb{1}_{2f}$: J is an orthogonal matrix and Jz is orthogonal to z for every z.

Rests on Remark 24.11 and Equation (22.56).

Definition 24.28 (Ball and cylinder)**.** In $\left(\mathbb{R}^{2f}, \omega_0\right)$ with the coordinates Equation (24.18), the *ball* of radius ρ and the *cylinder* of radius r over the first conjugate pair are

$$B(\rho) = \left\{ z \in \mathbb{R}^{2f} \;:\; \textstyle\sum_a \left[(u^a)^2 + (v_a)^2 \right] < \rho^2 \right\}, \tag{24.20}$$

$$Z(r) = \left\{ z \in \mathbb{R}^{2f} \;:\; \left(u^1\right)^2 + (v_1)^2 < r^2 \right\}. \tag{24.21}$$

A *symplectic embedding* of one subset of $\mathbb{R}^{2f}$ into another is a smooth embedding φ with $\varphi^* \omega_0 = \omega_0$.

Rests on Remark 24.27 and Definition 24.17.

$Z(r)$ has infinite volume for every $r > 0$, so no volume argument can prevent a ball of any radius from being squeezed into it. And volume preservation alone does not prevent it either:

Example 24.29 (A volume-preserving squeeze)**.** Take $f = 3$—one particle in three-dimensional space—and, for $\lambda > 0$, the linear map

$$\begin{aligned} \left(u^1, v_1\right) &\mapsto \left(\lambda u^1, \lambda v_1\right), \\ \left(u^2, v_2\right) &\mapsto \left(\lambda^{-1} u^2, \lambda^{-1} v_2\right), \\ \left(u^3, v_3\right) &\mapsto \left(u^3, v_3\right). \end{aligned}$$

Its determinant is $\lambda^2 \cdot \lambda^{-2} \cdot 1 = 1$, so it preserves the Liouville volume exactly; and it carries $B(\rho)$ into $Z(\lambda\rho)$, which for small λ is as thin as one likes. It is not symplectic: it pulls ω_0 back to $\lambda^2\, \mathrm{d}u^1 \wedge \mathrm{d}v_1 + \lambda^{-2}\, \mathrm{d}u^2 \wedge \mathrm{d}v_2 + \mathrm{d}u^3 \wedge \mathrm{d}v_3$, which is not ω_0. The squeezing that volume preservation permits is thus achieved by trading area between two *different* conjugate pairs—exactly what a symplectic map may not do.

Rests on Definition 24.28 and Theorem 24.21.

Proposition 24.30 (Linear non-squeezing)**.** *Let $A \in \mathrm{Sp}(2f, \mathbb{R})$ and suppose $A\left(B(\rho)\right) \subseteq Z(r)$. Then $r \geq \rho$.*

Rests on Definitions 24.4 and 24.28 and Equation (24.19).

Proof. *Derives Proposition 24.30.*

Let e_1 and f^1 be the first pair of the canonical basis, so that $\omega_0(e_1, f^1) = 1$ by Equation (24.1). For any $w \in \mathbb{R}^{2f}$ the two coordinates in which the cylinder is defined are

$$u^1(w) = \omega_0\left(w, f^1\right), \qquad v_1(w) = -\omega_0(w, e_1),$$

as one checks by evaluating $\omega_0 = \mathrm{d}u^a \wedge \mathrm{d}v_a$ on the basis. Because A preserves ω_0, $\omega_0(Az, f^1) = \omega_0\left(z, A^{-1} f^1\right)$ and likewise for e_1. Put $a = A^{-1} e_1$ and $b = A^{-1} f^1$, so that

$$\omega_0(a, b) = \omega_0\left(e_1, f^1\right) = 1, \tag{24.22}$$

and the composition of A with the projection onto the cylinder plane is the linear map $L : \mathbb{R}^{2f} \to \mathbb{R}^2$,

$$L(z) = \left(\omega_0(z, b),\, -\omega_0(z, a)\right) = \left(\langle z, J\, b\rangle,\, -\langle z, J\, a\rangle\right),$$

by Equation (24.19). The hypothesis $A\left(B(\rho)\right) \subseteq Z(r)$ says that $|L(z)| < r$ for every z with $|z| < \rho$, hence by continuity $|L(z)| \leq r$ on the closed ball, that is $\rho\, \|L\| \leq r$ with $\|L\|$ the operator norm.

It remains to show $\|L\| \geq 1$. The matrix of L has the two rows $(Jb)^{\mathsf{T}}$ and $-(Ja)^{\mathsf{T}}$, so

$$L\, L^{\mathsf{T}} = \begin{pmatrix} |b|^2 & -\langle a, b\rangle \\ -\langle a, b\rangle & |a|^2 \end{pmatrix}, \quad \det\left(L\, L^{\mathsf{T}}\right) = |a|^2\, |b|^2 - \langle a, b\rangle^2,$$

where $|Jb| = |b|$, $|Ja| = |a|$ and $\langle Jb, Ja\rangle = \langle b, a\rangle$ because J is orthogonal by Equation (24.19). The determinant is the square of the Euclidean area $\mathcal{A}$ of the parallelogram spanned by a and b. Now Jb is orthogonal to b and has the same length, so $Jb/|b|$ is a unit vector perpendicular to b, and the component of a along it is at most the distance from a to the line $\mathbb{R}b$:

$$1 = |\omega_0(a,b)| = |\langle a, Jb\rangle| = |b|\,\left|\left\langle a, \tfrac{Jb}{|b|}\right\rangle\right| \le |b| \cdot \operatorname{dist}(a, \mathbb{R}b) = \mathcal{A}\,,$$

using Equation (24.22). Hence $\det\left(LL^{\mathsf{T}}\right) = \mathcal{A}^2 \ge 1$. The matrix LL^{T} is symmetric positive semidefinite with two eigenvalues whose product is at least 1, so its largest eigenvalue is at least 1; and $\|L\|^2$ is that largest eigenvalue. Therefore $\|L\| \ge 1$, and $r \ge \rho\,\|L\| \ge \rho$. □

The proposition is elementary because a linear symplectic map sends the ball to an ellipsoid and the argument is one about areas of shadows. Gromov's theorem is the same statement without linearity, and it is not elementary at all: it was the first application of pseudoholomorphic curves, and no proof by the methods of this chapter is known.

Theorem 24.31 (Gromov's non-squeezing theorem)**.** *Let $f \ge 2$ and $\rho, r > 0$. If there exists a symplectic embedding $\varphi : B(\rho) \hookrightarrow Z(r)$ in the sense of Definition 24.28, then $r \ge \rho$.*
Rests on Definition 24.28 and Proposition 24.30.

Remark 24.32 (Non-squeezing is quoted, and will not be proved here)*.* Theorem 24.31 carries no derivation in this book, and the omission is a decision rather than a debt. It is due to Gromov, in *Pseudo holomorphic curves in symplectic manifolds* (*Inventiones Mathematicae* **82**, 1985); that paper has no key in this treatise's bibliography, so the attribution is made in words, on the same footing as Darboux's own memoir in Appendix A.12. The proof constructs a pseudoholomorphic disc through a prescribed point of the image and bounds its area by an elliptic-analysis argument—compactness of a moduli space of such discs, and a monotonicity estimate for their area—machinery that appears nowhere else in this treatise and that no other result here needs. Introducing it to serve one statement would unbalance the chapter, and this is the same treatment Part II gives Carleson's theorem in Remark 17.18.

No shortcut through the linear case exists. Proposition 24.30, which *is* proved in full by the methods of this chapter, constrains only linear symplectic maps and says nothing about a nonlinear embedding, and the obvious replacement argument—that a symplectomorphism preserves volume, so it cannot squeeze—is simply false, by Example 24.29. What the theorem is used for below is one statement, that the Gromov width of Equation (24.23) satisfies the third capacity axiom of Definition 24.33, and everything this chapter concludes from that is flagged in Remark 24.35 as mathematics rather than as measured physics.

Definition 24.33 (Symplectic capacity)**.** A *symplectic capacity* is an assignment c of a number in $[0,\infty]$ to every pair $(U, \alpha\omega_0)$ with $U \subseteq \mathbb{R}^{2f}$ and $\alpha \ne 0$ real, such that

1. *monotonicity*: if $(U, \alpha\omega_0) \hookrightarrow (V, \beta\omega_0)$ admits a symplectic embedding, then $c\,(U, \alpha\omega_0) \le c\,(V, \beta\omega_0)$;
2. *conformality*: $c\,(U, \alpha\omega_0) = |\alpha|\; c\,(U, \omega_0)$;
3. *nontriviality*: $c\,(B(1), \omega_0) = c\,(Z(1), \omega_0) = \pi$.

The argument ω_0 is suppressed when it is the standard form. By Remark 24.27 a capacity of a region of physical phase space carries the unit J s.
Rests on Definition 24.28 and Remark 24.27.

The three axioms are those of Ekeland and Hofer. Only the third mentions a specific set, and it is the one that does all the work. Monotonicity and conformality alone are cheap:

the f-th root of the Liouville volume satisfies both, since a symplectic embedding preserves volume and rescaling ω_0 by α rescales the volume by $|\alpha|^f$. What that candidate cannot do is meet the third axiom, because $Z(1)$ has infinite volume as soon as $f \geq 2$; in $f = 1$ the area *is* a capacity, and the whole content of the subject is that the case $f \geq 2$ is not vacuous. The following makes that precise—the existence of a capacity and the non-squeezing theorem are the same statement.

Proposition 24.34 (Capacities exist if and only if non-squeezing holds)**.** *Fix $f \geq 1$. There exists a symplectic capacity on $\left(\mathbb{R}^{2f}, \omega_0\right)$ if and only if the non-squeezing statement of Theorem 24.31 holds in that dimension—for $f = 1$ the elementary statement that an area-preserving embedding of a disc into a disc cannot decrease the radius, for $f \geq 2$ Gromov's theorem. In that case the* Gromov width

$$c_G(U) = \sup\left\{\pi\rho^2 \;:\; B(\rho) \text{ embeds symplectically in } U\right\} \tag{24.23}$$

is a capacity, and it is the smallest one.

Rests on Definition 24.33 and Theorem 24.31.

Proof. *Derives Proposition 24.34.*

First note that any capacity takes the value $\pi\rho^2$ on $B(\rho)$ and on $Z(\rho)$. The dilation $\delta_\rho(z) = \rho z$ carries $B(1)$ onto $B(\rho)$ and satisfies $\delta_\rho^*\omega_0 = \rho^2\omega_0$, so it is a symplectomorphism from $\left(B(1), \rho^2\omega_0\right)$ to $(B(\rho), \omega_0)$; monotonicity applied in both directions and then conformality with $\alpha = \rho^2$ give

$$c\left(B(\rho), \omega_0\right) = c\left(B(1), \rho^2\omega_0\right) = \rho^2 c\left(B(1), \omega_0\right) = \pi\rho^2\,,$$

and the identical argument applies to $Z(r)$, giving πr^2.

A capacity implies non-squeezing. Let c be a capacity and let $\varphi : B(\rho) \hookrightarrow Z(r)$ be a symplectic embedding. Then monotonicity gives $\pi\rho^2 = c\left(B(\rho)\right) \leq c\left(Z(r)\right) = \pi r^2$, hence $r \geq \rho$.

Non-squeezing implies a capacity. Assume Theorem 24.31 and consider c_G of Equation (24.23). Monotonicity is immediate: a symplectic embedding $U \hookrightarrow V$ composes with any $B(\rho) \hookrightarrow U$, so every ρ admissible for U is admissible for V. Conformality follows from the dilation computation above, since precomposing an embedding into $(U, \alpha\omega_0)$ with $\delta_{\sqrt{\alpha}}$ shows that ρ is admissible for $\alpha\omega_0$ exactly when $\rho/\sqrt{\alpha}$ is admissible for ω_0, so the supremum of $\pi\rho^2$ is multiplied by $\alpha > 0$; the case $\alpha < 0$ reduces to it by the reflection $v_a \mapsto -v_a$ in *every* pair, which carries ω_0 to $-\omega_0$ and leaves every ball and every cylinder where it was. For nontriviality, the identity embeds $B(1)$ in itself and in $Z(1)$, so $c_G\left(B(1)\right) \geq \pi$ and $c_G\left(Z(1)\right) \geq \pi$; conversely a symplectic embedding $B(\rho) \hookrightarrow Z(1)$ forces $\rho \leq 1$ by Theorem 24.31, so $c_G\left(Z(1)\right) \leq \pi$, and $B(1) \subset Z(1)$ with monotonicity gives $c_G\left(B(1)\right) \leq \pi$ as well.

Minimality. If c is any capacity and $B(\rho)$ embeds in U, then $\pi\rho^2 = c\left(B(\rho)\right) \leq c(U)$; taking the supremum over such ρ gives $c_G(U) \leq c(U)$. □

Remark 24.35 (What non-squeezing does and does not say about nature)*.* A Hamiltonian flow is a symplectomorphism (Theorem 24.16), so the Gromov width of a phase-space region is conserved along the motion just as its volume is. That is a genuine constraint on classical dynamics, and it is strictly stronger than Liouville's: by Example 24.29 volume preservation permits a squeeze that Theorem 24.31 forbids. The shadow of a Hamiltonian-transported region on any one conjugate plane cannot be made arbitrarily small while the region stays large in the other directions.

Three cautions keep the statement honest. First, this treatise records no experiment that measures a symplectic capacity, and the section is therefore mathematics, not evidence-based

physics. Second, the resemblance to the uncertainty relation of Chapter 77 is formal only: Definition 24.33 assigns a capacity to a *region*, and classical mechanics puts no floor under it—a single point, or a Lagrangian sheet of any extent, has capacity zero, and there is no classical reason for the number S of Remark 24.27 to be $\hbar$ rather than anything else. The uncertainty relation is a theorem about noncommuting operators and states, not about embeddings, and it is the quantum postulate that supplies the scale. Third, what the analogy does establish is a dimensional coincidence worth keeping in view: a capacity carries J s by Remark 24.27, which is the unit of $\hbar$, and Remark 24.11 traces that back to the pairing of a coordinate with its momentum. It is the same coincidence that makes Section 24.7 possible.

Rests on Theorems 24.16 and 24.31 and Remark 24.11.

24.5 Poisson manifolds

Not every space carrying a Poisson bracket is a symplectic manifold. The bracket, not the form, is the more general object, and it is the one mechanics actually needs—rigid-body dynamics lives on a space where the form degenerates.

Definition 24.36 (Poisson manifold)**.** A *Poisson manifold* is a manifold P with a bilinear operation $\{\cdot,\cdot\}$ on $C^\infty(P)$ that is antisymmetric, satisfies the Jacobi identity Equation (22.46), and obeys the Leibniz rule

$$\{u, vw\} = \{u, v\}\, w + v\, \{u, w\}\ . \tag{24.24}$$

Equivalently, it carries a bivector field $\pi \in \Gamma(\Lambda^2 TP)$ with $\{u, v\} = \pi(\mathrm{d}u, \mathrm{d}v)$ whose Schouten–Nijenhuis bracket with itself vanishes, $[\pi, \pi] = 0$; that vanishing is the Jacobi identity.

Rests on Definition 22.29 and Equation (22.46).

Definition 24.37 (Casimir function)**.** A *Casimir* is a function C with $\{C, u\} = 0$ for every u. A Casimir is conserved by every Hamiltonian, whatever it is; on a symplectic manifold only the constants are Casimirs, so a nonconstant Casimir is exactly a signature of degeneracy.

Rests on Definition 24.36 and Proposition 24.15.

That the only Casimirs of a symplectic manifold are the constants is immediate from Proposition 24.15: $\{C, u\} = -X_C[u]$ vanishes for every u only if $X_C = 0$, and then $\mathrm{d}C = \iota_{X_C}\omega = 0$ by Equation (24.5), so C is locally constant. On a Poisson manifold the map $\mathrm{d}C \mapsto X_C$ has a kernel, and the Casimirs are what live in it.

Theorem 24.38 (Symplectic foliation, quoted)**.** *A Poisson manifold is partitioned into immersed submanifolds—the* symplectic leaves*—each of which carries a symplectic form inducing the ambient bracket. The leaves are the level sets of the Casimirs where those are regular, and the dynamics of any Hamiltonian stays on the leaf it starts on.*

Rests on Definitions 24.8, 24.36 and 24.37.

Remark 24.39 (The foliation is quoted; its last clause is proved). Only the existence of the leaves is quoted here, and it is worth being exact about which half that is. The bracket of Definition 24.36 assigns to each point the image of the bivector π, a subspace of the tangent space spanned by the Hamiltonian vector fields; the Jacobi identity makes that family of subspaces involutive, since $[X_u, X_v] = -X_{\{u,v\}}$ is again a Hamiltonian vector field—the computation proving Equation (24.8) uses only the Jacobi identity and the derivation property, both of which Definition 24.36 supplies—and integrating an involutive family into submanifolds is the Frobenius theorem. The manifolds chapter Chapter 13 does not carry it, and even if it did it would not suffice: the rank of π varies from point to point—for the Lie–Poisson bracket Equation (24.26) of the rigid body it is zero at the

origin of $\mathfrak{so}(3)^*$ and two everywhere else—so the distribution is singular, and the tool required is the integrability theorem for singular distributions rather than the constant-rank Frobenius theorem. Weinstein's splitting theorem, the local normal form that generalizes Darboux (Theorem 24.12) to the Poisson case, belongs with the same missing half. All of it is differential topology, and this chapter states it rather than proving it for the reason given in Remark 17.18: the machinery has no other use in this book.

The last clause of Theorem 24.38 is the one the rigid body of Proposition 24.41 actually consumes, and it needs nothing beyond the definitions. If C is a Casimir then $\dot{C} = \{C, \mathcal{H}\} = 0$ along the flow of *any* $\mathcal{H}$ whatever, so every Casimir is a constant of the motion and a trajectory stays in the level set of every Casimir it starts in. Where those level sets are regular they are the leaves, so no application below rests on the quoted half.

Definition 24.40 (Lie–Poisson bracket)**.** Let $\mathfrak{g}$ be a Lie algebra and $\mathfrak{g}^*$ its dual. The *Lie–Poisson bracket* on $\mathfrak{g}^*$ is

$$\{u, v\}_{\pm}(\mu) = \pm \left\langle \mu, \left[\frac{\delta u}{\delta \mu}, \frac{\delta v}{\delta \mu} \right] \right\rangle , \qquad \mu \in \mathfrak{g}^* , \tag{24.25}$$

the inner bracket being that of $\mathfrak{g}$. Both signs give a Poisson structure in the sense of Definition 24.36—antisymmetry and the Leibniz rule are unaffected by an overall sign, and the Jacobi identity is quadratic in the bracket and therefore also unaffected—and which one a given mechanical system requires is fixed by the physics, not by convention. The rigid body referred to body axes takes the minus sign, Proposition 24.41.

Rests on Definitions 5.123 and 24.36.

Proposition 24.41 (The free rigid body is a Lie–Poisson system)**.** *Take $\mathfrak{g} = \mathfrak{so}(3)$, identify $\mathfrak{g}^*$ with $\mathbb{R}^3$ so that $\mu = \boldsymbol{L}$ is the angular momentum referred to body axes, and let $\mathcal{H} = \frac{1}{2} L_i \left(I^{-1}\right)_{ij} L_j$ with I the inertia tensor. Then the minus Lie–Poisson bracket of Equation (24.25) is*

$$\{u, v\}_{-} = -\boldsymbol{L} \cdot (\nabla u \times \nabla v) , \qquad \textit{so that} \qquad \{L_i, L_j\}_{-} = -\varepsilon_{ijk} L_k , \tag{24.26}$$

the gradients being taken with respect to $\boldsymbol{L}$, and the equations of motion it generates are Euler's equations Theorem 29.25 for the torque-free body,

$$\frac{\mathrm{d}\boldsymbol{L}}{\mathrm{d}t} = \boldsymbol{L} \times \boldsymbol{\Omega} , \qquad \boldsymbol{\Omega} = I^{-1} \boldsymbol{L} . \tag{24.27}$$

The function $C = |\boldsymbol{L}|^2$ is a Casimir, the leaves are the spheres $|\boldsymbol{L}| =$ constant, and the trajectories are the intersections of those spheres with the energy ellipsoids $\mathcal{H} =$ constant.

Rests on Definitions 24.37 and 24.40 and Theorem 29.25.

Proof. *Derives Proposition 24.41.*

The bracket. Take the basis e_i of $\mathfrak{so}(3)$ with $[e_j, e_k] = \varepsilon_{jkl} e_l$, so that the structure constants are the Levi-Civita symbol (Definition 14.26 and Lemma 14.62), and identify $\mathfrak{g}^* \cong \mathbb{R}^3$ by $\langle \mu, e_i \rangle = L_i$. For a function u of $\boldsymbol{L}$ the functional derivative is the ordinary gradient, $\delta u / \delta \mu = (\partial u / \partial L_j)\, e_j$. Then

$$\begin{aligned} \{u, v\}_{-} &= -\left\langle \mu, \left[\frac{\partial u}{\partial L_j} e_j, \frac{\partial v}{\partial L_k} e_k \right] \right\rangle = -\varepsilon_{jkl} L_l \frac{\partial u}{\partial L_j} \frac{\partial v}{\partial L_k} \\ &= -\boldsymbol{L} \cdot (\nabla u \times \nabla v) , \end{aligned}$$

since $(\nabla u \times \nabla v)_l = \varepsilon_{ljk} \left(\partial u / \partial L_j\right) \left(\partial v / \partial L_k\right)$ and $\varepsilon_{jkl} = \varepsilon_{ljk}$. Putting $u = L_i$, $v = L_j$ gives $\{L_i, L_j\}_{-} = -\boldsymbol{L} \cdot (\hat{e}_i \times \hat{e}_j) = -\varepsilon_{ijk} L_k$, which is the bracket declared in Remark 29.27.

The equations of motion. With $\nabla\mathcal{H} = I^{-1}\boldsymbol{L} = \boldsymbol{\Omega}$—the body-frame angular velocity, since $\boldsymbol{L} = I\boldsymbol{\Omega}$—and $\nabla L_i = \hat{e}_i$,

$$\begin{aligned}\dot{L}_i &= \{L_i, \mathcal{H}\}_- = -\boldsymbol{L}\cdot(\hat{e}_i \times \boldsymbol{\Omega}) = -\varepsilon_{mik} L_m \Omega_k \\ &= \varepsilon_{imk} L_m \Omega_k = (\boldsymbol{L}\times\boldsymbol{\Omega})_i \ ,\end{aligned}$$

which is Equation (24.27), that is $\mathrm{d}\boldsymbol{L}/\mathrm{d}t + \boldsymbol{\Omega}\times\boldsymbol{L} = 0$— Equation (29.21) with $\boldsymbol{\tau} = \boldsymbol{0}$. In principal axes $L_i = I_i\Omega_i$ and the components read $I_1\dot{\Omega}_1 = (I_2 - I_3)\,\Omega_2\Omega_3$ and its cyclic partners, which are Equations (29.18) to (29.20) with zero torque. Had the plus sign of Equation (24.25) been taken, the same computation would give $\mathrm{d}\boldsymbol{L}/\mathrm{d}t = \boldsymbol{\Omega}\times\boldsymbol{L}$, the time reverse; that is why the sign is part of the physics.

The Casimir and the leaves. For $C = |\boldsymbol{L}|^2$ one has $\nabla C = 2\boldsymbol{L}$, so for every v

$$\{C, v\}_- = -\boldsymbol{L}\cdot(2\boldsymbol{L}\times\nabla v) = 0$$

because the triple product has a repeated factor. Hence C is a Casimir in the sense of Definition 24.37, its regular level sets—the spheres of radius $|\boldsymbol{L}| > 0$—are the symplectic leaves, and every trajectory stays on one of them. Intersecting with the conserved $\mathcal{H}$ gives the closed curves of Proposition 29.26. The odd dimension of $\mathbb{R}^3$ is what forces this structure: no symplectic form exists there, and the two-dimensional leaves are where the symplectic geometry of this chapter actually lives, as Remark 29.27 anticipates. □

Remark 24.42 (The intermediate axis). The leaf picture settles the qualitative theory of free rotation in one figure and one line of algebra. Order the principal moments I_1, I_2, I_3 and intersect the sphere $|\boldsymbol{L}| = L$ with the ellipsoids $\mathcal{H} = E$: near the axes of largest and of smallest moment the intersections are small closed curves encircling the axis, and near the intermediate axis they cross. Linearising Equation (24.27) about steady rotation $\boldsymbol{\Omega} = (\Omega, 0, 0)$ gives

$$I_2\dot{\Omega}_2 = (I_3 - I_1)\,\Omega\,\Omega_3\,, \qquad I_3\dot{\Omega}_3 = (I_1 - I_2)\,\Omega\,\Omega_2\,,$$

and eliminating Ω_3,

$$\ddot{\Omega}_2 = \frac{(I_3 - I_1)\,(I_1 - I_2)}{I_2 I_3}\,\Omega^2\,\Omega_2\,,$$

whose solutions grow exponentially exactly when $(I_3 - I_1)\,(I_1 - I_2) > 0$, that is when I_1 lies strictly between the other two. Rotation about the intermediate axis is unstable and rotation about the other two is stable—the tennis-racket effect, and the reason a thrown book tumbles about one axis and not the other three. The rigid-body chapter treats it in full, Section 29.3.

Rests on Propositions 24.41 and 29.26.

24.6 Symmetry, momentum maps, and reduction

Definition 24.43 (Momentum map)**.** Let a Lie group G with Lie algebra $\mathfrak{g}$ act on (M, ω) by symplectomorphisms, and write ξ_M for the vector field generating the action of $\xi \in \mathfrak{g}$. A *momentum map* is a map $\boldsymbol{J} : M \to \mathfrak{g}^*$ such that, for every ξ,

$$\iota_{\xi_M}\omega = \mathrm{d}\,\langle \boldsymbol{J}, \xi\rangle \ , \tag{24.28}$$

that is, such that the component J_ξ is a Hamiltonian generating the action of ξ.

Rests on Definitions 14.2, 24.14 and 24.17.

Example 24.44 (Linear and angular momentum)**.** Take $M = T^*\mathbb{R}^3$ with $\omega = \mathrm{d}q^i \wedge \mathrm{d}p_i$, the phase space of one particle. For the translation group acting by $\boldsymbol{q} \mapsto \boldsymbol{q} + s\boldsymbol{n}$ with $\boldsymbol{n}$ a fixed unit vector, the generator is $\xi_M = n^i\,\partial/\partial q^i$ and

$$\iota_{\xi_M}\omega = n^i\,\mathrm{d}p_i = \mathrm{d}\left(\boldsymbol{n}\cdot\boldsymbol{p}\right), \qquad \text{so} \qquad J_\xi = \boldsymbol{n}\cdot\boldsymbol{p},$$

the component of linear momentum along $\boldsymbol{n}$, in kg m/s. For the rotation group acting by $\boldsymbol{q} \mapsto R\boldsymbol{q}$, $\boldsymbol{p} \mapsto R\boldsymbol{p}$ about the axis $\boldsymbol{n}$, the generator is $\xi_M = (\boldsymbol{n}\times\boldsymbol{q})^i\,\partial/\partial q^i + (\boldsymbol{n}\times\boldsymbol{p})_i\,\partial/\partial p_i$ and

$$\iota_{\xi_M}\omega = (\boldsymbol{n}\times\boldsymbol{q})^i\,\mathrm{d}p_i - (\boldsymbol{n}\times\boldsymbol{p})_i\,\mathrm{d}q^i = \mathrm{d}\left(\boldsymbol{n}\cdot(\boldsymbol{q}\times\boldsymbol{p})\right),$$

as one checks by expanding $\mathrm{d}\left(\boldsymbol{n}\cdot(\boldsymbol{q}\times\boldsymbol{p})\right) = (\boldsymbol{p}\times\boldsymbol{n})\cdot\mathrm{d}\boldsymbol{q} + (\boldsymbol{n}\times\boldsymbol{q})\cdot\mathrm{d}\boldsymbol{p}$. So $J_\xi = \boldsymbol{n}\cdot\boldsymbol{L}$, the component of angular momentum, in J s. The momentum map of the full rotation group is $\boldsymbol{J} = \boldsymbol{q}\times\boldsymbol{p}$. The name is not an analogy: the momentum map of a symmetry *is* the momentum conjugate to it.

Rests on Definition 24.43 and Equation (24.3).

Theorem 24.45 (Noether, symplectic form)**.** *If G acts by symplectomorphisms preserving H, then the momentum map is conserved along the flow:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{J} = 0\,. \tag{24.29}$$

For G the translation group $\boldsymbol{J}$ is the linear momentum, for the rotation group the angular momentum, and for the one-parameter group generated by the flow itself the Hamiltonian. This is Noether's theorem of Chapter 21, stated on phase space, where it is a single line.

Rests on Definition 24.43, Proposition 24.15, and Equation (24.5).

Proof. *Derives Theorem 24.45.*

Fix $\xi \in \mathfrak{g}$ and differentiate the component J_ξ along the motion. By Equation (24.7) the rate of change of any function is $\dot{J}_\xi = X_H[J_\xi] = \mathrm{d}J_\xi(X_H)$, and by the defining property Equation (24.28) of the momentum map,

$$\begin{aligned}\mathrm{d}J_\xi\left(X_H\right) &= \left(\iota_{\xi_M}\omega\right)\left(X_H\right) = \omega\left(\xi_M, X_H\right) = -\omega\left(X_H, \xi_M\right)\\ &= -\left(\iota_{X_H}\omega\right)\left(\xi_M\right) = -\mathrm{d}H\left(\xi_M\right),\end{aligned}$$

using Equation (24.5) in the last step. The final expression is $-\xi_M[H]$, the rate of change of H along the group direction, and that vanishes because H is invariant under the action. So $\dot{J}_\xi = 0$ for every ξ, which is Equation (24.29). The identifications are Example 24.44; for the one-parameter group generated by X_H itself the defining relation Equation (24.28) is Equation (24.5), so $J = H$, and the conservation statement is that a time-independent Hamiltonian is conserved.

Notice how short the argument is compared with the Lagrangian one of Chapter 21: there the conserved quantity has to be constructed from the variation of the action, here it is handed over by nondegeneracy—every symmetry generator is a Hamiltonian vector field, and its Hamiltonian is the conserved quantity [456]. □

Lemma 24.46 (The level set of the momentum map is the symplectic orthogonal of the orbit)**.** *Let $\boldsymbol{J}$ be a momentum map for the action of G on (M,ω) and let $x \in M$. Then*

$$\ker \mathrm{d}\boldsymbol{J}_x = \left(T_x\left(G\cdot x\right)\right)^\omega, \tag{24.30}$$

where $G\cdot x$ is the group orbit through x and, for a subspace $W \subseteq T_xM$,

$$W^\omega := \{w \in T_xM \;:\; \omega\left(w,z\right) = 0 \text{ for all } z \in W\}$$

is its symplectic orthogonal*. In particular, if μ is a regular value of $\boldsymbol{J}$ then $\dim \boldsymbol{J}^{-1}(\mu) = \dim M - \dim G$ when the action is locally free.*

Rests on Definitions 24.8 and 24.43 and Theorem 13.59.

Proof. *Derives Lemma 24.46.*

The tangent space to the orbit at x is spanned by the values $\xi_M(x)$, $\xi \in \mathfrak{g}$, of the generators. For $w \in T_xM$ and any ξ, pairing the differential of $\boldsymbol{J}$ with ξ and using Equation (24.28),

$$\langle \mathrm{d}\boldsymbol{J}_x(w), \xi \rangle = \mathrm{d}J_\xi(w) = (\iota_{\xi_M}\omega)(w) = \omega(\xi_M(x), w) \,.$$

Hence $\mathrm{d}\boldsymbol{J}_x(w) = 0$ if and only if $\omega(\xi_M(x), w) = 0$ for every ξ, which by antisymmetry is exactly membership of $(T_x(G \cdot x))^\omega$. This is Equation (24.30).

For the dimension count, a locally free action has $\dim T_x(G \cdot x) = \dim G$; nondegeneracy of ω makes $w \mapsto \omega(\cdot, w)$ an isomorphism $T_xM \to T_x^*M$, so the symplectic orthogonal of a k-dimensional subspace has dimension $2f - k$, and therefore $\dim \ker \mathrm{d}\boldsymbol{J}_x = \dim M - \dim G$. Since μ is a regular value, $\boldsymbol{J}^{-1}(\mu)$ is a submanifold with tangent space $\ker \mathrm{d}\boldsymbol{J}_x$ by the regular value theorem Theorem 13.59, of that dimension. □

Theorem 24.47 (Marsden–Weinstein reduction)**.** *Let G act freely and properly on (M, ω) with an equivariant momentum map $\boldsymbol{J}$, and let $\mu \in \mathfrak{g}^*$ be a regular value. Then the quotient*

$$M_\mu = \boldsymbol{J}^{-1}(\mu)/G_\mu \tag{24.31}$$

carries a unique symplectic form pulling back to the restriction of ω, and every G-invariant Hamiltonian on M descends to it. Its dimension is

$$\dim M_\mu = \dim M - \dim G - \dim G_\mu \,. \tag{24.32}$$

Rests on Lemma 24.46, Theorem 13.59, and Definition 24.8.

Full derivation: Appendix A.48 (p. 3044).

Derives Theorem 24.47.

The geometric core of the argument is Lemma 24.46, proved above: the tangent space to the level set of the momentum map is the symplectic orthogonal of the group orbit, from which the degenerate directions of the restricted form turn out to be exactly the orbit directions of the isotropy group G_μ—so that quotienting by that group removes the degeneracy, and removes nothing else. What makes the theorem long is everything around that core: that $\boldsymbol{J}^{-1}(\mu)$ is a submanifold, that its quotient by G_μ is again a smooth manifold with the quotient map a submersion, and that the descended form is well defined, closed and nondegenerate. Appendix A.48 carries it out.

One input there is quoted rather than proved, and is named as such in the appendix: the quotient-manifold theorem, that a free and proper action of a Lie group on a manifold has a smooth quotient with the projection a submersion. It is differential topology and belongs in Chapter 13, which does not yet carry it. Its role is confined to producing the smooth structure on M_μ; the symplectic content of Theorem 24.47—the existence, uniqueness and nondegeneracy of the reduced form, and the dimension count Equation (24.32)—rests on Lemma 24.46 and on nothing quoted.

Remark 24.48 (Reduction is what physicists do without saying so)*.* Eliminating a cyclic coordinate by fixing its conjugate momentum and dropping the pair; passing to the centre-of-mass frame; reducing the two-body problem of Section 27.4 to the radial problem; using the effective potential Equation (27.32)—each of these is Theorem 24.47 applied to a particular group, and the effective potential is exactly the reduced Hamiltonian on M_μ at the fixed value $\mu = L$ of the angular momentum. The Lie–Poisson bracket of Definition 24.40 is itself a reduction: it is what the canonical bracket on T^*G becomes when the whole group is divided out, and that is why the rigid body of Proposition 24.41 lives on three variables rather than the six a cotangent bundle would demand.

Rests on Theorem 24.47 and Proposition 24.41.

24.7 Geometric quantization

The programme of this section attempts to construct a quantum theory from a symplectic manifold intrinsically, using no coordinates and no prescription for ordering operators. It is placed here because it is the sharpest test of how much of quantum mechanics is already present in the geometry of phase space, and its answer is instructive in both directions: the linear structure of quantum mechanics is reproduced exactly, and the obstruction of Theorem 25.38 survives unchanged.

This section is mathematics, and is written as such. No experiment in this treatise tests geometric quantization: what is tested is the quantum mechanics of Chapter 77, which the construction reproduces where it works and does not derive. The scope rule is respected by saying so plainly rather than by omitting the material, because the construction explains where two measured things come from—the Bohr–Sommerfeld rule Equation (23.61) and the half-integer shift in it—and those explanations are checkable.

24.7.1 Prequantization

Definition 24.49 (Prequantum datum)**.** A *prequantum datum* for (M,ω) is a triple (L,h,∇) consisting of a complex line bundle $L\to M$, a Hermitian metric h on it, and a connection ∇ compatible with h, whose curvature

$$F^{\nabla}(X,Y)=\nabla_X\nabla_Y-\nabla_Y\nabla_X-\nabla_{[X,Y]} \tag{24.33}$$

satisfies

$$F^{\nabla}=\frac{\mathrm{i}}{\hbar}\,\omega\,. \tag{24.34}$$

The right-hand side is dimensionless by Remark 24.11, as a curvature must be, and $\hbar$ is the only constant of nature with the unit needed to make it so. Over a chart on which $\omega=-\mathrm{d}\theta$ one may take L trivial and

$$\nabla_X s=X[s]-\frac{\mathrm{i}}{\hbar}\,\theta(X)\,s\,, \tag{24.35}$$

whose curvature is $-(\mathrm{i}/\hbar)\,\mathrm{d}\theta=(\mathrm{i}/\hbar)\,\omega$ as required.

Rests on Definitions 14.103, 14.105 and 24.8.

Definition 24.50 (Prequantum operator)**.** For $u\in C^\infty(M)$ the *prequantum operator* acts on smooth sections of L by

$$\widehat{u}\,s=-\mathrm{i}\hbar\,\nabla_{X_u}s+u\,s\,, \tag{24.36}$$

with X_u the Hamiltonian vector field Equation (24.5). Both terms carry the unit of u, by Remark 24.11.

Rests on Definitions 24.14 and 24.49.

Proposition 24.51 (Prequantization is a Lie-algebra homomorphism)**.** *For all $u,v\in C^\infty(M)$,*

$$[\widehat{u},\widehat{v}]=\mathrm{i}\hbar\,\widehat{\{u,v\}}\,,\qquad\text{and}\qquad\widehat{1}=\mathbb{1}\,. \tag{24.37}$$

That is, prequantization satisfies hypotheses (1) and (2) of Theorem 25.38 on all *of $C^\infty(M)$, with no restriction on the degree of the observable.*

Rests on Definition 24.50 and Equations (24.8) *and* (24.34)*.*

Proof. *Derives Proposition 24.51.*

That $\widehat{1}=\mathbb{1}$ is immediate: $X_1=0$ because $\mathrm{d}1=0$, so Equation (24.36) leaves only multiplication by 1.

For the commutator, expand Equation (24.36) and treat the four terms separately. Multiplication operators commute, so $[u,v]=0$. For the cross terms, the Leibniz rule for a connection gives $[\nabla_X, v]\,s = \nabla_X(vs) - v\nabla_X s = X[v]\,s$, so

$$-\mathrm{i}\hbar\,[\nabla_{X_u}, v] + \mathrm{i}\hbar\,[\nabla_{X_v}, u] = -\mathrm{i}\hbar\,X_u[v] + \mathrm{i}\hbar\,X_v[u] = 2\mathrm{i}\hbar\,\{u,v\}\ ,$$

where the last step uses Equation (24.7) twice: $X_v[u]=\{u,v\}$ and $X_u[v]=\{v,u\}=-\{u,v\}$. For the remaining term, Equation (24.33) rearranged reads $[\nabla_{X_u},\nabla_{X_v}] = F^{\nabla}(X_u,X_v) + \nabla_{[X_u,X_v]}$, so with Equation (24.34), Equation (24.7) and Equation (24.8),

$$\begin{aligned}(-\mathrm{i}\hbar)^2\,[\nabla_{X_u},\nabla_{X_v}] &= -\hbar^2\left[\frac{\mathrm{i}}{\hbar}\,\omega\,(X_u,X_v) + \nabla_{-X_{\{u,v\}}}\right]\\ &= -\mathrm{i}\hbar\,\{u,v\} + \hbar^2\nabla_{X_{\{u,v\}}}\,.\end{aligned}$$

Adding the two displays,

$$\begin{aligned}[\widehat{u},\widehat{v}] &= \hbar^2\nabla_{X_{\{u,v\}}} + \mathrm{i}\hbar\,\{u,v\}\\ &= \mathrm{i}\hbar\left(-\mathrm{i}\hbar\,\nabla_{X_{\{u,v\}}} + \{u,v\}\right) = \mathrm{i}\hbar\,\widehat{\{u,v\}}\,,\end{aligned}$$

which is Equation (24.37). The sign in Equation (24.34) is fixed by this calculation and is not free: with $F^{\nabla} = -(\mathrm{i}/\hbar)\,\omega$ the same steps would produce $3\mathrm{i}\hbar\,\{u,v\}$ in place of $\mathrm{i}\hbar\,\{u,v\}$, and the correspondence would fail on the very first pair of canonical variables. □

Example 24.52 (The prequantum line and the Schrödinger representation)**.** Take one degree of freedom, $M=T^*\mathbb{R}$ with $\omega=\mathrm{d}q\wedge\mathrm{d}p$, L trivial, $\theta=p\,\mathrm{d}q$ and ∇ as in Equation (24.35). From Equation (24.6), $X_q=-\partial/\partial p$ and $X_p=\partial/\partial q$, so Equation (24.36) gives

$$\widehat{q}\,s = \mathrm{i}\hbar\,\frac{\partial s}{\partial p} + q\,s\,,\qquad \widehat{p}\,s = -\mathrm{i}\hbar\,\frac{\partial s}{\partial q}\,, \tag{24.38}$$

the second because the connection term $-(\mathrm{i}/\hbar)\,p\,s$ in $\nabla_{\partial/\partial q}s$ contributes $-p\,s$ once multiplied by $-\mathrm{i}\hbar$, cancelling the $+p\,s$ of Equation (24.36). One checks directly that $[\widehat{q},\widehat{p}]=\mathrm{i}\hbar$, as Proposition 24.51 demands. But these operators act on functions of q *and* p: the prequantum Hilbert space is far too large—twice as many variables as a wavefunction has, and, for the harmonic oscillator, a continuous spectrum where the observed one is discrete. Cutting it down is the job of the next step.

Rests on Proposition 24.51 and Equation (24.6).

24.7.2 Polarization

Definition 24.53 (Real polarization and polarized sections)**.** A *real polarization* of (M,ω) is an integrable distribution $P\subset TM$ of rank f on which ω vanishes—a foliation of M by Lagrangian submanifolds. A section s of L is *polarized* if $\nabla_X s=0$ for every X taking values in P.

Rests on Definitions 24.8 and 24.49.

Proposition 24.54 (The vertical polarization gives wave mechanics)**.** *On $M=T^*Q$ with $\theta=p_a\mathrm{d}q^a$, let P be spanned by the $\partial/\partial p_a$. Then the polarized sections are the functions $\psi(q)$ of the configuration variables alone, and on them*

$$\widehat{q}^{\,a}\psi = q^a\psi\,,\qquad \widehat{p}_a\psi = -\mathrm{i}\hbar\,\frac{\partial\psi}{\partial q^a}\,. \tag{24.39}$$

That is, the Schrödinger representation of Chapter 77 is what geometric quantization of a cotangent bundle in the vertical polarization produces.

Rests on Definitions 24.50 and 24.53 and Equation (24.35).

Proof. *Derives Proposition 24.54.*

P is integrable, being spanned by coordinate fields, and $\omega\,(\partial/\partial p_a, \partial/\partial p_b) = 0$, so it is a real polarization; its leaves are the fibres of $T^*Q \to Q$. Since $\theta\,(\partial/\partial p_a) = 0$, the connection Equation (24.35) reduces on P to the ordinary derivative, so a polarized section satisfies $\partial s/\partial p_a = 0$: it is a function $\psi(q)$.

For the operators, $X_{q^a} = -\partial/\partial p_a$ lies in P, so $\nabla_{X_{q^a}}\psi = 0$ and Equation (24.36) leaves $\widehat{q}^{\,a}\psi = q^a\psi$. For the momenta, $X_{p_a} = \partial/\partial q^a$ and Equation (24.35) gives $\nabla_{X_{p_a}}\psi = \partial\psi/\partial q^a - (\mathrm{i}/\hbar)\,p_a\psi$, whence

$$\begin{aligned}\widehat{p}_a\psi &= -\mathrm{i}\hbar\left(\frac{\partial\psi}{\partial q^a} - \frac{\mathrm{i}}{\hbar}p_a\psi\right) + p_a\psi \\ &= -\mathrm{i}\hbar\,\frac{\partial\psi}{\partial q^a} - p_a\psi + p_a\psi = -\mathrm{i}\hbar\,\frac{\partial\psi}{\partial q^a}\,.\end{aligned}$$

The cancellation is exact and is what makes the answer independent of p, as it must be for an operator on $\psi(q)$. □

Remark 24.55 (What polarization costs). Equation (24.36) defines $\widehat{u}$ for every smooth u, but $\widehat{u}$ maps polarized sections to polarized sections only when the flow of X_u preserves the polarization. On T^*Q with the vertical polarization that restricts the directly quantizable observables to those at most quadratic in the momenta—for $\mathcal{H} = p^2/2m$ one already finds $[\partial/\partial p, (p/m)\,\partial/\partial q] = (1/m)\,\partial/\partial q \notin P$, so the free Hamiltonian moves the polarization and needs a separate treatment, in this case an easy one. Beyond the quadratic observables further devices are required, and none of them evades Theorem 25.38: there is no linear map from all the polynomials to operators satisfying Equation (24.37) irreducibly. Geometric quantization localizes the obstruction rather than removing it— prequantization satisfies the bracket condition on everything but is reducible, polarization restores irreducibility but shrinks the domain, and Theorem 25.38 is the theorem that no construction can have both.

Rests on Proposition 24.54 and Theorem 25.38.

24.7.3 Bohr–Sommerfeld from holonomy

Proposition 24.56 (The Bohr–Sommerfeld condition is a triviality condition on the prequantum holonomy)**.** *Let the motion be bounded and integrable, so that by Theorem 23.31 the level sets of the integrals are tori T carrying the basis cycles C_a of Definition 23.28, and take the polarization tangent to those tori. Then a nowhere-vanishing polarized section of L exists over T if and only if*

$$\oint_{C_a} p_a\,\mathrm{d}q^a = n_a h\,, \qquad n_a \in \mathbb{Z}\,, \tag{24.40}$$

for every a, which is the Bohr–Sommerfeld condition Equation (23.61).

Rests on Definitions 23.28 and 24.49 and Theorem 23.31.

Proof. *Derives Proposition 24.56.*

A Liouville–Arnold torus is Lagrangian: the actions are in involution, so ω restricted to T vanishes. By Equation (24.34) the curvature of ∇ restricted to T therefore vanishes as well, so the connection is flat there and its parallel transport around a loop depends only on the homotopy class of the loop.

Compute that transport. Along a curve γ in T with tangent $\dot\gamma$, the condition $\nabla_{\dot\gamma}s = 0$ reads, by Equation (24.35),

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\mathrm{i}}{\hbar}\,\theta(\dot\gamma)\,s\,, \qquad \text{whence} \qquad s(t) = s(0)\,\mathrm{e}^{\,\Phi(t)}\,, \quad \Phi(t) = \frac{\mathrm{i}}{\hbar}\int_0^t \theta(\dot\gamma)\,\mathrm{d}t'\,.$$

Around a closed cycle C_a the holonomy is therefore $\exp\!\big((\mathrm{i}/\hbar)\oint_{C_a}\theta\big)$, and a globally defined nowhere-vanishing polarized section over T exists exactly when every such holonomy is 1—parallel transport must return a section to itself. That is

$$\frac{1}{\hbar}\oint_{C_a} p_a\,\mathrm{d}q^a \in 2\pi\mathbb{Z}\,, \qquad \text{that is} \qquad \oint_{C_a} p_a\,\mathrm{d}q^a = 2\pi\hbar\,n_a = n_a h\,,$$

using $\theta = p_a \mathrm{d}q^a$ from Equation (24.3) and $h = 2\pi\hbar$. This is Equation (24.40), and it is Equation (23.61) [441]. □

Remark 24.57 (The half-integer, and the honest status of the construction). Proposition 24.56 reproduces the old quantum rule with integers, and the old quantum rule with integers is wrong: the harmonic oscillator's ground state is not at zero energy. The missing piece is the *metaplectic*, or half-form, correction—one quantizes sections of $L \otimes \sqrt{\Lambda}$ rather than of L, where $\sqrt{\Lambda}$ is a square root of the bundle of forms along the polarization—and its effect is to replace n_a by $n_a + \mu_a/4$, with μ_a the Maslov index of the cycle, counting the caustics it crosses. For one degree of freedom in a smooth well $\mu = 2$, so $\oint p\,\mathrm{d}q = \left(n + \frac{1}{2}\right)h$; for the oscillator of frequency ν this integral is E/ν, giving $E_n = \left(n + \frac{1}{2}\right)h\nu$, the observed spectrum with its zero-point energy. It is the same shift the semiclassical limit of wave mechanics produces, and the pending note at Proposition 23.38 records where that derivation belongs.

Two further statements are made here without proof, and are flagged as such. The existence of a prequantum datum at all constrains the cohomology class of ω: running the holonomy argument above over a closed two-surface $\Sigma \subset M$ instead of a loop gives the necessary condition that $(2\pi\hbar)^{-1}\int_\Sigma \omega$ be an integer, and that this condition is also sufficient is Weil's integrality theorem. Its proof needs the cohomology of a good cover of the manifold, which this treatise does not carry; and nothing above rests on it, since the Bohr–Sommerfeld statement just proved uses only the flatness of the connection on one torus. The construction as a whole is due to Kostant and to Souriau, independently, around 1970; neither is in this treatise's bibliography, so the attribution is made in words.

Finally, the scope rule. Geometric quantization is a construction, not a physical theory: it takes a classical system as input and manufactures a candidate quantum one, and Theorem 25.38 guarantees that it cannot do so for every observable. Where it succeeds it reproduces the quantum mechanics postulated in Chapter 77, which is what experiment tests; nothing in this section is itself tested by an experiment recorded in this treatise, and the two checkable statements it touches—Equation (23.61) and the half-integer shift—are tested as spectroscopy, not as geometry.

Rests on Propositions 23.38 and 24.56 and Theorem 25.38.

24.8 What the symplectic view buys

The material above is geometry. This section records, as a checklist for the chapters that will use it, the places where the geometry makes a difference to what is measured. Each item is a pointer, not a derivation.

The measure of statistical mechanics. The Liouville form is the unique flow-invariant measure available without extra structure, which is why the microcanonical ensemble is defined with it (Chapter 118 and Remark 24.22).

Adiabatic invariants. The action variables of Definition 23.28 are the quantities that survive slow modulation, and their conservation is what confines charged particles in magnetic bottles (Chapter 66 and Remark 23.37).

Invariant tori and the stability of orbits. KAM theory (Chapter 32) is a theorem about symplectic maps [1247], and its observational trace is the resonance structure of the solar system, whose inner planets are chaotic with a Lyapunov time of a few million years [1824].

Symplectic integrators. Numerical schemes built to preserve ω exactly—rather than to minimise the local truncation error—have bounded energy error over exponentially long times, which is why planetary ephemerides and accelerator tracking are computed with them. This is a case where a geometric property, not an accuracy estimate, decides the method.

Geometric phases. Holonomy of the Hamiltonian flow over a closed path in parameter space produces a phase that depends on the geometry of the path and not on its rate [1734]; it is observed in polarization rotation in a coiled optical fibre [1783], in molecular spectra near conical intersections, and in the Aharonov–Bohm effect (Chapter 77) [1164]. The construction of Section 24.7 is where such a phase comes from: it is a prequantum holonomy, the same object whose triviality gives Equation (24.40).

Non-squeezing. Liouville's theorem is not the only constraint on what a Hamiltonian flow may do to a region (Theorem 24.31 and Remark 24.35), though no experiment in this treatise measures the finer invariant.

Constrained systems. Degenerate Lagrangians produce presymplectic rather than symplectic structures, and recovering a bracket on the physical subspace is the content of the next chapters (Chapters 25 and 26).

Chapter 25

The Poisson Algebra and the Canonical Bridge to Quantum Mechanics

Chapter 22 introduced the Poisson bracket as a convenient notation and listed its properties; Chapter 24 showed that those properties are the shadow of the symplectic form. This chapter takes the bracket as the primary object and asks what kind of algebraic structure it is. The answer—a Lie algebra acting by derivations—turns out to be the whole of what survives the passage to quantum mechanics.

The route is direct. Every observable generates a one-parameter family of canonical transformations through the bracket; a symmetry is an observable whose generated family preserves the Hamiltonian; and a conserved quantity is the same thing seen from the other side. Rewriting mechanics so that the bracket carries the dynamics leaves a formalism in which the equations of motion are first order and every statement is algebraic, and that formalism has an exact quantum counterpart obtained by replacing the bracket with a commutator. Where the replacement is legitimate, and where it provably is not, is the content of Section 25.5, and it is the reason quantization is a physical hypothesis to be tested rather than a construction to be carried out. Standard treatments of the quantum side are [1593, 1912, 2287].

Nothing in this chapter derives quantum mechanics. The quantum theory is postulated in Chapter 77, and the correspondence set out here is used there — in Section 77.5.1 — to motivate the operator assignments, not to justify them. The direction of logical dependence is stated once, in Remark 25.51, and held to throughout: classical mechanics is a limit of the quantum theory, and the classical structures of Chapters 22 and 24 are what that limit leaves behind.

Notation 25.1. The conventions of Notation 22.1 remain in force. A hat distinguishes an operator on a Hilbert space, $\hat{A}$, from the phase-space function A it is meant to correspond to; the commutator is $\left[\hat{A},\hat{B}\right]$ and the anticommutator $\left\{\hat{A},\hat{B}\right\}$, in the macros of the treatise, and the Poisson bracket keeps its braces $\{u,v\}$. Where a bracket must be qualified by the coordinates in which it is computed the subscript of Equation (22.41) is used.

25.1 The Poisson algebra of observables

Definition 25.2 (Poisson algebra)**.** A *Poisson algebra* is a real vector space A carrying two products: an associative commutative product uv, making it a commutative algebra, and a bracket $\{\cdot,\cdot\}$ making it a Lie algebra, the two being compatible through the Leibniz rule Equation (24.24). The observables $C^\infty(M)$ of a mechanical system, with pointwise

multiplication and the Poisson bracket of Definition 22.29, form one.
Rests on Definitions 22.29 and 24.36.

Theorem 25.3 (The bracket is a Lie bracket)**.** *The Poisson bracket is bilinear, antisymmetric Equation* (22.43) *and satisfies the Jacobi identity Equation* (22.46)*. The observables of a mechanical system therefore form an infinite-dimensional real Lie algebra.*
Rests on Equations (22.41) *and* (22.56)*.*

Derivation. *Derives Theorem 25.3.*

Write the bracket in the phase-space notation of Equation (22.55). With $\partial_c := \partial/\partial\eta^c$ and J^{cd} the components of the symplectic matrix Equation (22.56) — constants, so the position of the index carries no meaning here — the definition Equation (22.41) reads

$$\{u, v\} = J^{cd}\,\partial_c u\,\partial_d v\,, \qquad J^{cd} = -J^{dc}\,, \qquad \partial_e J^{cd} = 0\,. \tag{25.1}$$

Unpacking the blocks of Equation (22.56) returns $\partial_q u\,\partial_p v - \partial_p u\,\partial_q v$, so Equation (25.1) is the same bracket and not a new one.

Bilinearity is immediate, since u and v enter through one derivative each. Antisymmetry follows from $J^{cd} = -J^{dc}$ after relabelling the two summed indices, which is Equation (22.43).

For the Jacobi identity, differentiate once more. Because J is constant it passes through ∂_d, and

$$\{u, \{v, w\}\} = J^{cd}J^{ef}\,\partial_c u\,(\partial_d\partial_e v\,\partial_f w + \partial_e v\,\partial_d\partial_f w)\,. \tag{25.2}$$

Every term of the cyclic sum $\{u,\{v,w\}\}+\{v,\{w,u\}\}+\{w,\{u,v\}\}$ therefore carries exactly one second derivative, of one of the three functions, and first derivatives of the other two. Collect the two terms carrying $\partial\partial u$. They come from the second and third brackets and are

$$A = J^{cd}J^{ef}\,\partial_c v\,\partial_e w\,\partial_d\partial_f u\,, \qquad B = J^{cd}J^{ef}\,\partial_c w\,\partial_f v\,\partial_d\partial_e u\,.$$

In B interchange the names $e \leftrightarrow f$, which turns J^{ef} into $J^{fe} = -J^{ef}$, and then interchange the pairs $(c,d) \leftrightarrow (e,f)$, which merely renames dummies:

$$B = -J^{cd}J^{ef}\,\partial_c w\,\partial_e v\,\partial_d\partial_f u = -J^{cd}J^{ef}\,\partial_c v\,\partial_e w\,\partial_d\partial_f u = -A\,,$$

so the two cancel. The cyclic sum is invariant under $u \to v \to w \to u$, so the same cancellation removes the $\partial\partial v$ and $\partial\partial w$ terms, and nothing is left. This is Equation (22.46).

Two features of the computation are worth naming, because they are what the identity really rests on. Antisymmetry of J is used once, and constancy of J is used once — and the latter is the coordinate shadow of the closedness of the symplectic form (Definition 13.105), which is why the Jacobi identity survives the change to arbitrary coordinates of Theorem 24.12 even though J^{cd} does not stay constant there. A bracket built from an antisymmetric $J^{cd}(\eta)$ that is *not* the inverse of a closed form is bilinear and antisymmetric but fails Jacobi; the Leibniz rule alone does not save it. □

Proposition 25.4 (Leibniz rule and derivations)**.** *For each u, the map $v \mapsto \{v, u\}$ is a derivation of the commutative product; equivalently it is a first-order differential operator, the vector field X_u of Definition 24.14. The Jacobi identity says exactly that $u \mapsto X_u$ is a homomorphism of Lie algebras, up to the sign recorded in Equation* (24.8)*.*
Rests on Theorem 25.3 and Definition 24.14.

Derivation. *Derives Proposition 25.4.*

For the Leibniz rule, put the product vw into the first slot of Equation (25.1) and use the product rule on the single derivative that acts on it:

$$\{vw, u\} = J^{cd}\,\partial_c(vw)\,\partial_d u = w\,J^{cd}\partial_c v\,\partial_d u + v\,J^{cd}\partial_c w\,\partial_d u = w\,\{v,u\} + v\,\{w,u\}\,, \tag{25.3}$$

which is Equation (24.24). A linear map obeying Equation (25.3) is by definition a derivation of the commutative product, and since it involves exactly one derivative of its argument it is a first-order differential operator: reading off the coefficient,

$$\{v,u\} = X_u[v]\,, \qquad X_u = X_u^c\,\partial_c\,, \qquad X_u^c = J^{cd}\,\partial_d u\,, \tag{25.4}$$

which is Equation (24.6). Taking $u = \mathcal{H}$ makes $X_{\mathcal{H}}^c = J^{cd}\partial_d\mathcal{H}$ the right-hand side of Equation (22.57): the Hamiltonian vector field is the velocity field of the motion in phase space.

For the second statement, let w be arbitrary and compute the commutator of the two vector fields on it, using Equation (25.4) twice:

$$[X_u, X_v]\,[w] = X_u[\{w,v\}] - X_v[\{w,u\}] = \{\{w,v\},u\} - \{\{w,u\},v\}\ .$$

Now write the Jacobi identity Equation (22.46) with w in the outer slot and move each inner bracket outwards by antisymmetry, $\{u,\{v,w\}\} = \{\{w,v\},u\}$ and $\{v,\{w,u\}\} = -\{\{w,u\},v\}$:

$$\{w,\{u,v\}\} + \{\{w,v\},u\} - \{\{w,u\},v\} = 0\,.$$

The first term is $X_{\{u,v\}}[w]$ and the remaining two are $[X_u, X_v]\,[w]$, so $X_{\{u,v\}}[w] = -\,[X_u, X_v]\,[w]$ for every w, which is Equation (24.8). The sign is a consequence of the conventions, not a defect: it says that $u \mapsto X_u$ is an antihomomorphism in the convention of Equation (22.41), and it is why some texts define the Hamiltonian vector field with the opposite sign. □

Remark 25.5 (Two products, and which one survives). A Poisson algebra carries a commutative product and a Lie bracket. Their quantum counterparts are the symmetrized product $\frac{1}{2}\left\{\hat{A},\hat{B}\right\}$, which is commutative but no longer associative, and the commutator $\frac{1}{i\hbar}\left[\hat{A},\hat{B}\right]$, which remains a Lie bracket exactly. The asymmetry is not an accident of the formalism: it is why the Lie structure transfers to the quantum theory essentially intact while the pointwise product does not, and it is the algebraic root of the obstruction of Section 25.5.

25.2 Generators, symmetries, and constants of motion

Definition 25.6 (Infinitesimal canonical transformation)**.** Let G be a phase-space function and ε an infinitesimal parameter. The transformation

$$\delta q^a = \varepsilon\,\{q^a, G\}\ , \qquad \delta p_a = \varepsilon\,\{p_a, G\}\ , \qquad \text{that is} \qquad \delta u = \varepsilon\,\{u,G\} \ \text{ for any } u\,, \tag{25.5}$$

is canonical to first order in ε; G is called its *generator*.

Rests on Definitions 22.29 and 24.17.

Proposition 25.7 (Generating function of an infinitesimal transformation)**.** *The transformation Equation (25.5) is generated, in the sense of Equation (22.21), by $F_2 = q^a P_a + \varepsilon G(q,P)$: the identity transformation plus ε times the generator.*

Rests on Equations (22.21) and (25.5).

Derivation. *Derives Proposition 25.7.*

Substitute $F_2 = q^a P_a + \varepsilon G(q,P)$ into the two transformation relations of Equation (22.21):

$$p_a = \frac{\partial F_2}{\partial q^a} = P_a + \varepsilon\,\frac{\partial G}{\partial q^a}\,, \qquad Q^a = \frac{\partial F_2}{\partial P_a} = q^a + \varepsilon\,\frac{\partial G}{\partial P_a}\,. \tag{25.6}$$

Setting $\varepsilon = 0$ gives $P_a = p_a$ and $Q^a = q^a$: the first term alone generates the identity, which is the check the construction needs, since a transformation infinitesimally close to the identity must be generated by a function infinitesimally close to that one.

For $\varepsilon \neq 0$ the relations are implicit, G being a function of (q, P) rather than of (q, p); but the difference between P_a and p_a is itself of order ε by the first of Equation (25.6), so replacing P by p inside a derivative of G costs only $O(\varepsilon^2)$. To first order, therefore,

$$\delta p_a = P_a - p_a = -\varepsilon \frac{\partial G}{\partial q^a}, \qquad \delta q^a = Q^a - q^a = +\varepsilon \frac{\partial G}{\partial p_a}.$$

These are exactly Equation (25.5): by Equation (22.41), $\{q^a, G\} = \partial G/\partial p_a$ and $\{p_a, G\} = -\partial G/\partial q^a$. The statement for an arbitrary observable follows by the chain rule,

$$\delta u = \frac{\partial u}{\partial q^a}\delta q^a + \frac{\partial u}{\partial p_a}\delta p_a = \varepsilon \left(\frac{\partial u}{\partial q^a}\frac{\partial G}{\partial p_a} - \frac{\partial u}{\partial p_a}\frac{\partial G}{\partial q^a} \right) = \varepsilon \, \{u, G\} \, ,$$

so a single function G moves every observable at once, and it does so through the bracket alone. □

Proposition 25.8 (The Hamiltonian generates time translation)**.** *Taking $G = \mathcal{H}$ and $\varepsilon = \mathrm{d}t$ in Equation (25.5) reproduces the evolution equation Equation (22.50). The motion of a mechanical system is therefore itself a one-parameter family of canonical transformations, generated by the Hamiltonian; the time evolution and the canonical transformations are not two subjects but one.*

Rests on Equations (22.50) and (25.5).

Derivation. *Derives Proposition 25.8.*

Put $G = \mathcal{H}$ and $\varepsilon = \mathrm{d}t$ in the last member of Equation (25.5). For an observable u carrying no explicit time dependence this gives $\delta u = \mathrm{d}t \, \{u, \mathcal{H}\}$, that is $\mathrm{d}u/\mathrm{d}t = \{u, \mathcal{H}\}$, which is Equation (22.51). If u does depend explicitly on the time, the transformation still contributes only $\mathrm{d}t \, \{u, \mathcal{H}\}$ — it acts on the phase-space arguments — and the remaining $\partial u/\partial t$ of Equation (22.50) comes from the explicit slot, which the transformation does not touch. Applying the result to $u = q^a$ and $u = p_a$ and using Equation (22.44) returns Hamilton's equations Equations (22.3) and (22.4), so the family generated by $\mathcal{H}$ is the motion itself and not merely a transformation resembling it. □

Theorem 25.9 (Noether, bracket form)**.** *Let G carry no explicit time dependence. Then the following are equivalent:*

1. *G is a constant of the motion;*
2. *$\{G, \mathcal{H}\} = 0$;*
3. *the one-parameter family of canonical transformations generated by G leaves $\mathcal{H}$ invariant.*

Symmetry and conservation are thus the same statement read in two directions, and the object that expresses both is the bracket.

Rests on Equations (22.43), (22.50) and (25.5).

Derivation. *Derives Theorem 25.9.*

(1) $\Leftrightarrow$ (2). With $\partial G/\partial t = 0$ the evolution equation Equation (22.50) reads $\mathrm{d}G/\mathrm{d}t = \{G, \mathcal{H}\}$. If the bracket vanishes identically then G is constant along every motion. Conversely, suppose G is constant along every motion. The bracket $\{G, \mathcal{H}\}$ is a function on phase space, and through each point of phase space there passes a motion — take that point as the initial condition of Equations (22.3) and (22.4) — along which $\mathrm{d}G/\mathrm{d}t$ vanishes at that point. So $\{G, \mathcal{H}\}$ vanishes at every point. The quantifier is what does the work: constancy along a *single* trajectory is a weaker statement and does not give (2).

(2) $\Leftrightarrow$ (3). By Proposition 25.7 the family generated by G moves any observable by $\delta u = \varepsilon \{u, G\}$, so it moves the Hamiltonian by

$$\delta \mathcal{H} = \varepsilon \{\mathcal{H}, G\} = -\varepsilon \{G, \mathcal{H}\} , \tag{25.7}$$

the second equality being the antisymmetry Equation (22.43). The right-hand side vanishes for all ε exactly when $\{G, \mathcal{H}\} = 0$, which is (2); and its vanishing is the statement that the family leaves $\mathcal{H}$ invariant, which is (3). □

Remark 25.10 (What this version of Noether does and does not cover). Theorem 25.9 is the Hamiltonian face of the variational theorem Theorem 24.45, and it is sharper in one respect and narrower in two others. Sharper: the correspondence is a bijection, because Equation (25.7) runs in both directions — every conserved G generates a symmetry and every generated symmetry conserves its generator, with no integrability condition to check. Narrower: it sees only symmetries generated by a phase-space function, so a discrete symmetry such as spatial inversion or time reversal falls outside it entirely — those have conserved quantities in quantum mechanics (Chapter 110) but no classical generator, since a one-parameter family cannot reach them; and it treats the time as a parameter, so a symmetry that reparametrizes the time — the boost of Equation (25.11) is the mild case, general covariance the severe one — needs the extended phase space of Chapter 26 before this argument applies to it.

Definition 25.11 (Generators of the Galilei transformations)**.** Let a free particle of mass m move in three-dimensional Euclidean space, with phase-space coordinates (x_i, p_i) and Hamiltonian $\mathcal{H} = |\boldsymbol{p}|^2 / 2m$. The generators of the Galilei transformations are the phase-space functions

$$\text{time translation:} \quad \mathcal{H} = \frac{p_k p_k}{2m} , \tag{25.8}$$

$$\text{space translation:} \quad p_i , \tag{25.9}$$

$$\text{rotation:} \quad L_i = \epsilon_{ijk} x_j p_k , \tag{25.10}$$

$$\text{boost:} \quad K_i = m\, x_i - t\, p_i . \tag{25.11}$$

All indices are Cartesian components in Euclidean space, so their position carries no meaning and repeated indices are summed; ϵ_{ijk} is the Levi-Civita symbol. The boost generator is the only one carrying an explicit dependence on the time.

Rests on Definitions 22.29 and 25.6.

Proposition 25.12 (The algebra of generators closes)**.** *By Poisson's theorem Theorem 22.38 the constants of motion of a system form a Lie subalgebra of the Poisson algebra. For the free particle in three dimensions the generators Equations* (25.8) *to* (25.11) *satisfy*

$$\{L_i, L_j\} = \epsilon_{ijk} L_k , \qquad \{L_i, p_j\} = \epsilon_{ijk} p_k , \qquad \{p_i, p_j\} = 0 , \tag{25.12}$$

$$\{L_i, K_j\} = \epsilon_{ijk} K_k , \qquad \{K_i, \mathcal{H}\} = p_i , \qquad \{K_i, K_j\} = 0 , \tag{25.13}$$

$$\{L_i, \mathcal{H}\} = 0 , \qquad \{p_i, \mathcal{H}\} = 0 , \qquad \{\mathcal{H}, \mathcal{H}\} = 0 , \tag{25.14}$$

together with

$$\{K_i, p_j\} = m\, \delta_{ij} , \tag{25.15}$$

which is the Galilei algebra of Chapter 14 centrally extended by the mass.

Rests on Definition 25.11, Equation (22.44)*, and Theorem 22.38.*

Derivation. *Derives Proposition 25.12.*

Throughout, the bracket is the one of Definition 22.29 in the Cartesian coordinates (x_i, p_i), whose fundamental values are $\{x_i, p_j\} = \delta_{ij}$ and $\{x_i, x_j\} = \{p_i, p_j\} = 0$; the last of

these is the third entry of Equation (25.12), and the time t is a parameter, not a phase-space variable, so it passes through every bracket as a constant.

Rotations on a vector. Because the bracket is a derivation in each slot, the action of L_i on any function is fixed by its action on x_j and p_j. From the definition Equation (25.10),

$$\{L_i, x_j\} = -\frac{\partial L_i}{\partial p_c}\frac{\partial x_j}{\partial x_c} = -\epsilon_{iac}x_a\delta_{jc} = -\epsilon_{iaj}x_a = \epsilon_{ijk}x_k\,,$$
$$\{L_i, p_j\} = \frac{\partial L_i}{\partial x_c}\frac{\partial p_j}{\partial p_c} = \epsilon_{icb}p_b\delta_{jc} = \epsilon_{ijk}p_k\,,$$

which is the second entry of Equation (25.12): every vector built from $\boldsymbol{x}$ and $\boldsymbol{p}$ rotates as a vector.

Rotations among themselves. Applying the derivation property to $L_j = \epsilon_{jab}x_ap_b$ and using the two results just obtained,

$$\begin{aligned}\{L_i, L_j\} &= \epsilon_{jab}\left(\{L_i, x_a\}\,p_b + x_a\,\{L_i, p_b\}\right) = \epsilon_{jab}\epsilon_{iac}x_cp_b + \epsilon_{jab}\epsilon_{ibc}x_ap_c\\ &= \left(\delta_{ji}\delta_{bc} - \delta_{jc}\delta_{bi}\right)x_cp_b + \left(\delta_{jc}\delta_{ai} - \delta_{ji}\delta_{ac}\right)x_ap_c\\ &= \delta_{ij}\,x_bp_b - x_jp_i + x_ip_j - \delta_{ij}\,x_ap_a = x_ip_j - x_jp_i\,,\end{aligned}$$

the two contractions of Levi-Civita symbols being the standard identity $\epsilon_{iab}\epsilon_{icd} = \delta_{ac}\delta_{bd} - \delta_{ad}\delta_{bc}$ used twice after a cyclic relabelling. Since

$$\epsilon_{ijk}L_k = \epsilon_{ijk}\epsilon_{kab}x_ap_b = \left(\delta_{ia}\delta_{jb} - \delta_{ib}\delta_{ja}\right)x_ap_b = x_ip_j - x_jp_i\,,$$

the two agree, which is the first entry of Equation (25.12).

Rotations on the boost. The boost generator Equation (25.11) is a linear combination of x_i and p_i with coefficients that are not phase-space functions, so it rotates as a vector by linearity:

$$\{L_i, K_j\} = m\,\{L_i, x_j\} - t\,\{L_i, p_j\} = m\,\epsilon_{ijk}x_k - t\,\epsilon_{ijk}p_k = \epsilon_{ijk}\left(m\,x_k - t\,p_k\right) = \epsilon_{ijk}K_k\,,$$

the first entry of Equation (25.13).

Boosts among themselves. Expanding both slots and using the fundamental brackets,

$$\begin{aligned}\{K_i, K_j\} &= \{m\,x_i - t\,p_i, m\,x_j - t\,p_j\}\\ &= m^2\,\{x_i, x_j\} - mt\,\{x_i, p_j\} - tm\,\{p_i, x_j\} + t^2\,\{p_i, p_j\}\\ &= 0 - mt\,\delta_{ij} + mt\,\delta_{ij} + 0 = 0\,,\end{aligned}$$

the two surviving terms cancelling by the antisymmetry $\{p_i, x_j\} = -\delta_{ij}$. This is the third entry of Equation (25.13).

Boosts and the Hamiltonian. Since $\{x_i, \mathcal{H}\} = p_i/m$ and $\{p_i, \mathcal{H}\} = 0$,

$$\{K_i, \mathcal{H}\} = m\,\{x_i, \mathcal{H}\} - t\,\{p_i, \mathcal{H}\} = m\,\frac{p_i}{m} - 0 = p_i\,,$$

the second entry of Equation (25.13).

The remaining brackets with the Hamiltonian. The momenta commute among themselves, so $\{p_i, \mathcal{H}\} = 0$; and

$$\{L_i, \mathcal{H}\} = \frac{1}{m}\,p_b\,\{L_i, p_b\} = \frac{1}{m}\,\epsilon_{ibk}p_bp_k = 0$$

because a symmetric product is contracted with an antisymmetric symbol. Together with $\{\mathcal{H}, \mathcal{H}\} = 0$, which holds by antisymmetry, this is Equation (25.14): energy, momentum and angular momentum are constants of the motion.

Boosts and translations. Finally, and this is the bracket the proposition is about,

$$\{K_i, p_j\} = m\,\{x_i, p_j\} - t\,\{p_i, p_j\} = m\,\delta_{ij}\,,$$

which is Equation (25.15).

Every bracket therefore closes on the span of $\{\mathcal{H}, p_i, L_i, K_i\}$ *together with the constant function* m, and by Theorem 22.38 the ten generators are constants of the motion. For K_i this last claim needs the explicit time dependence, and is worth checking separately: by the evolution equation Equation (22.50),

$$\frac{\mathrm{d}K_i}{\mathrm{d}t} = \frac{\partial K_i}{\partial t} + \{K_i, \mathcal{H}\} = -p_i + p_i = 0\,, \tag{25.16}$$

so that $K_i = m\,x_i(0)$ is conserved. The conservation law carried by the boost generator is the statement that the centre of mass moves uniformly. □

Remark 25.13 (The generator that is not a generator). The right-hand side of Equation (25.15) is m times the constant function 1, and the constant functions are exactly the functions whose bracket with everything vanishes (Definition 24.37). So the algebra of the ten generators does not close on itself: it closes on itself *plus* a central element that generates no transformation at all. The precise sense in which this is an extension, and the proof that it cannot be removed by redefining the generators, are the subject of the next subsection.

25.2.1 Mass as a central charge

The general theory of central extensions of Lie algebras—the cocycle Proposition 14.73, the coboundary, and the cohomology group H^2 that classifies the obstructions (Proposition 14.76)—is developed as mathematics in Section 14.4, where the Heisenberg algebra Example 14.79 is the worked example. This subsection applies it to the one case classical mechanics forces on us; the argument below exhibits a cocycle and shows it is not a coboundary, which is exactly the computation that $H^2 \neq 0$ names.

Proposition 25.14 (The Galilei cocycle is not a coboundary)**.** *Let $\mathfrak{g}$ be the Galilei algebra, in which the abstract generators satisfy $[K_i, P_j] = 0$, and let $\xi \mapsto J_\xi$ be the assignment of Definition 25.11 of a phase-space function to each generator. Then the defect*

$$c(\xi, \eta) = \{J_\xi, J_\eta\} - J_{[\xi,\eta]} \tag{25.17}$$

is a constant function for every pair, and no redefinition $J_\xi \mapsto J_\xi + b(\xi)$ by constants $b(\xi)$ removes it. The realization of the Galilei algebra by Poisson brackets is therefore a realization of a genuine central extension of it, with central charge the mass.

Rests on Equation (25.15), Definition 14.72, and Proposition 14.73.

Derivation. *Derives Proposition 25.14.*

That c takes constant values is read off the brackets computed above: every one of Equations (25.12) to (25.14) reproduces the abstract Galilei bracket exactly, so c vanishes on those pairs, and Equation (25.15) gives

$$c(K_i, P_j) = m\,\delta_{ij} - J_{[K_i,P_j]} = m\,\delta_{ij} - J_0 = m\,\delta_{ij}\,, \tag{25.18}$$

a constant.

Now suppose the generators are redefined by constants, $J_\xi \mapsto \tilde{J}_\xi = J_\xi + b(\xi)$ with b linear. Constants have vanishing bracket with everything, so the left-hand term of Equation (25.17) is unchanged while the right-hand term shifts, and

$$\tilde{c}(\xi, \eta) = c(\xi, \eta) - b\,([\xi, \eta])\ . \tag{25.19}$$

Removing the defect Equation (25.18) would therefore require

$$b\left([K_i, P_j]\right) = m\,\delta_{ij}\,.$$

But in the Galilei algebra $[K_i, P_j] = 0$, so the left-hand side is $b(0) = 0$ by linearity, while the right-hand side is $m\,\delta_{ij} \neq 0$ for a massive particle. No choice of b works, and the extension is not removable. □

Remark 25.15 (Why the relativistic case is different). The same construction applied to the Poincaré algebra of Chapter 39 produces no central charge: every cocycle there *is* a coboundary, so the mass cannot enter that way, and it enters instead as the Casimir invariant that labels a representation (Chapter 94). The two facts are not independent. The Galilei algebra is the Inönü–Wigner contraction of the Poincaré algebra as $c \to \infty$ (Section 14.5), and the central charge of Equation (25.15) is what the Casimir becomes in the limit: a contraction can create a nontrivial cohomology class where the parent algebra had none. That the mass is a label in one theory and a central charge in the other is a single statement about the limit, not a coincidence of two formalisms.

Remark 25.16 (The Bargmann superselection rule, and its limits). Quantum-mechanically a central charge means that the symmetry is realized on the Hilbert space only projectively—up to a phase—and Bargmann's theorem states that for the Galilei group the phase cannot be removed [1029]. The consequence usually drawn is a superselection rule: within nonrelativistic quantum mechanics, states of different mass cannot be coherently superposed, since the relative phase such a superposition would acquire under a boost is not physically meaningful.

Two qualifications keep the statement honest. First, it is a theorem about a symmetry group, not a measurement: nothing in it is tested directly. Second, and more sharply, nature manifestly *does* superpose states of different mass—the neutrino flavour eigenstates of Chapter 114 are coherent superpositions of mass eigenstates, and their oscillation is observed. There is no contradiction, because that is a relativistic regime, where the symmetry is Poincaré and Remark 25.15 says there is no central charge to obstruct anything. The Bargmann rule is a statement about the nonrelativistic limit and holds only there. This is a good example of what an algebraic argument can and cannot deliver: it tells you what a given symmetry permits, and is silent about whether that symmetry is the one nature uses in the regime you are measuring.

Remark 25.17 (Central charges elsewhere in this treatise). The mass is not an isolated case, and it is not even the most important one in this chapter. The canonical commutation relations Equation (25.33) are themselves a central extension: the phase-space translations $\{x_i, p_i\}$ form an abelian algebra classically, and $[\hat{x}_i, \hat{p}_j] = \mathrm{i}\hbar\,\delta_{ij}$ extends it centrally with $\hbar$ as the charge, giving the Heisenberg algebra. On that reading Planck's constant and the mass play the same structural role, and Theorem 25.34 is the statement that the extension has essentially one irreducible representation. Further instances, each cross-referenced where it arises: the Schwinger terms in current-algebra commutators, whose non-removable version is the anomaly of Remark 26.49; the magnetic translation algebra of a charged particle in a uniform magnetic field, centrally extended by the magnetic flux, which is what makes the Landau levels of Chapter 128 degenerate; and the Virasoro central charge of two-dimensional conformal symmetry, which is measurable in the critical behaviour of two-dimensional systems (Chapter 123).

25.2.2 The generators as a bridge

Every entry of Equations (25.12) and (25.13) survives quantization with the bracket replaced by $\frac{1}{\mathrm{i}\hbar}\,[\cdot,\cdot]$; the angular-momentum algebra of Chapter 79 is the first line read that way, and its

half-integer representations—which have no classical counterpart—are the reason spin exists. That the classical bracket algebra of the generators of a symmetry group is reproduced exactly by the quantum commutator algebra, while the algebra of general observables is not (Section 25.5), is the single most important asymmetry in the correspondence.

The reason the generators escape the obstruction is visible in Lemma 25.39: every entry of Equations (25.12), (25.13) and (25.15) is a bracket of polynomials of degree at most two in (x_i, p_i), and degree two is precisely where Theorem 25.38 does *not* bite. A symmetry generator is linear or quadratic because it generates a linear transformation of phase space; the obstruction begins at degree three. This is not a general theorem — a symmetry of an interacting system need not have a low-degree generator — but it accounts for every case in this chapter and for the Poincaré generators of Chapter 94 as well.

25.3 The first-order formalism

Definition 25.18 (Phase-space action)**.** The *phase-space*, or *first-order*, action of a mechanical system is the functional of independently varied paths $(q^a(t), p_a(t))$

$$S\,[q,p] = \int_{t_a}^{t_b} \left(p_a \dot{q}^a - \mathcal{H}(q,p,t)\right) \mathrm{d}t\,. \tag{25.20}$$

Rests on Definition 22.2 and Equation (24.3).

Theorem 25.19 (Hamilton's equations from the first-order action)**.** *Stationarity of Equation (25.20) under independent variations of q^a and p_a, with δq^a vanishing at the endpoints and δp_a unrestricted, is equivalent to Hamilton's equations Equations (22.3) and (22.4).*

Rests on Equation (25.20) and Lemma 16.18.

Derivation. *Derives Theorem 25.19.*

Vary Equation (25.20) in both arguments at once:

$$\delta S = \int_{t_a}^{t_b} \left(\dot{q}^a\, \delta p_a + p_a\, \delta \dot{q}^a - \frac{\partial \mathcal{H}}{\partial q^a}\, \delta q^a - \frac{\partial \mathcal{H}}{\partial p_a}\, \delta p_a \right) \mathrm{d}t\,.$$

Only one term carries a derivative of a variation. Integrating it by parts once,

$$\int_{t_a}^{t_b} p_a\, \delta \dot{q}^a\, \mathrm{d}t = \Big[p_a\, \delta q^a \Big]_{t_a}^{t_b} - \int_{t_a}^{t_b} \dot{p}_a\, \delta q^a\, \mathrm{d}t\,,$$

and the boundary term vanishes because δq^a was required to vanish at the endpoints. Collecting,

$$\delta S = \int_{t_a}^{t_b} \left[\left(\dot{q}^a - \frac{\partial \mathcal{H}}{\partial p_a} \right) \delta p_a - \left(\dot{p}_a + \frac{\partial \mathcal{H}}{\partial q^a} \right) \delta q^a \right] \mathrm{d}t\,. \tag{25.21}$$

The variations δq^a and δp_a are independent and otherwise arbitrary, so by the fundamental lemma Lemma 16.18 applied to each in turn, $\delta S = 0$ for all of them if and only if both coefficients vanish identically, which is Equations (22.3) and (22.4). The converse direction is the same computation read backwards.

The asymmetry of the endpoint conditions is not an oversight. The integrand contains $\dot{q}^a$ but no $\dot{p}_a$, so the integration by parts is needed only for the coordinates, and only they generate a boundary term to be killed. Requiring δp_a to vanish at the ends as well would be harmless but redundant; requiring it of δq^a is not optional. □

Remark 25.20 (Why "first order" matters). The Lagrangian action is second order in time derivatives and its variation needs the positions fixed at both ends; the action Equation (25.20) is first order, and the object multiplying the derivative is the canonical one-form $\theta = p_a \mathrm{d}q^a$ of Equation (24.3). Three consequences run through the rest of this treatise:

1. the symplectic form is read off the action as $\omega = -\mathrm{d}\theta$, so a theory presented in first-order form announces its own phase-space geometry;
2. the quantity exponentiated in the phase-space path integral of Chapter 106 is $\mathrm{e}^{\mathrm{i}S[q,p]/\hbar}$ with S this action, so the measure there is the Liouville measure of Definition 24.20;
3. when the Legendre transformation of Definition 22.2 fails to be invertible, the first-order action is still well defined, and it is the starting point of the constraint analysis of Chapter 26.

25.3.1 The Faddeev–Jackiw form

The action Equation (25.20) singles out a splitting of the phase-space coordinates into q's and p's. Nothing in the variational argument required that splitting: what was used was only that the integrand is linear in the velocities. Dropping the splitting gives a formalism in which the brackets are *read off* the action rather than postulated, and in which the degenerate case — the case that defeats the Legendre transformation — is not an obstacle but the object of study. The construction is due to Faddeev and Jackiw; their 1988 letter is not in this bibliography, and the presentation below is therefore self-contained rather than a report of it.

Definition 25.21 (First-order action in general coordinates)**.** Let ξ^c, $c = 1, \dots, N$, be coordinates on a phase space, let $a_c(\xi)$ be the components of a one-form on it — the *canonical potential* — and let $\mathcal{H}(\xi)$ be the Hamiltonian. The *first-order*, or *Faddeev–Jackiw*, action is

$$S[\xi] = \int_{t_a}^{t_b} \Big(a_c(\xi)\, \dot{\xi}^c - \mathcal{H}(\xi) \Big)\, \mathrm{d}t\,, \tag{25.22}$$

and its *symplectic two-form* is the antisymmetric matrix

$$f_{cd} = \partial_c a_d - \partial_d a_c\,, \qquad \partial_c = \frac{\partial}{\partial \xi^c}\,. \tag{25.23}$$

In the notation of Section 13.6.4 this is $f = \mathrm{d}a$ in components.

Rests on Equation (25.20) and Definition 13.103.

Theorem 25.22 (Faddeev–Jackiw equations and brackets)**.** *Stationarity of Equation (25.22) under variations vanishing at the endpoints is equivalent to*

$$f_{cd}(\xi)\, \dot{\xi}^d = \partial_c \mathcal{H}(\xi)\,. \tag{25.24}$$

If f_{cd} is invertible, with inverse $\left(f^{-1}\right)^{cd}$, the motion is $\dot{\xi}^c = \left(f^{-1}\right)^{cd} \partial_d \mathcal{H}$, and it is generated by the bracket

$$\{u, v\} = \Big(f^{-1}\Big)^{cd}\, \partial_c u\, \partial_d v\,, \qquad \textit{so that} \qquad \Big\{\xi^c, \xi^d\Big\} = \Big(f^{-1}\Big)^{cd}\,. \tag{25.25}$$

This bracket satisfies the Jacobi identity for every choice of a_c.

Rests on Definition 25.21, Lemma 16.18, and Theorem 25.3.

Derivation. *Derives Theorem 25.22.*

The Euler–Lagrange equation of Equation (25.22) is

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial}{\partial \dot{\xi}^c}\Big(a_d \dot{\xi}^d - \mathcal{H}\Big) - \frac{\partial}{\partial \xi^c}\Big(a_d \dot{\xi}^d - \mathcal{H}\Big) = \partial_d a_c\, \dot{\xi}^d - \partial_c a_d\, \dot{\xi}^d + \partial_c \mathcal{H} = 0\,,$$

since $\partial\left(a_d \dot{\xi}^d\right)/\partial \dot{\xi}^c = a_c$ and $\mathrm{d}a_c/\mathrm{d}t = \partial_d a_c\, \dot{\xi}^d$. Rearranged, this is Equation (25.24). Note that a_c enters the equation of motion only through the antisymmetric combination

Equation (25.23): adding a gradient, $a_c \mapsto a_c + \partial_c \chi$, changes the integrand of Equation (25.22) by a total time derivative and leaves the dynamics alone. The potential is a gauge-dependent bookkeeping device; the form is the physics.

If f is invertible, contracting Equation (25.24) with the inverse gives $\dot{\xi}^c = \left(f^{-1}\right)^{cd} \partial_d \mathcal{H}$, and for any observable $\mathrm{d}u/\mathrm{d}t = \partial_c u \, \dot{\xi}^c = \left(f^{-1}\right)^{cd} \partial_c u \, \partial_d \mathcal{H}$, which is $\{u, \mathcal{H}\}$ in the bracket Equation (25.25). The inverse of an antisymmetric matrix is antisymmetric, so the bracket is antisymmetric, and it is a derivation in each slot because it carries one derivative per slot.

For the Jacobi identity the proof of Theorem 25.3 does not apply verbatim, since $\left(f^{-1}\right)^{cd}$ is not constant; what replaces constancy is that $f = \mathrm{d}a$ is exact, hence closed (Definition 13.105), $\partial_c f_{de} + \partial_d f_{ec} + \partial_e f_{cd} = 0$ by inspection of Equation (25.23). Differentiating $\left(f^{-1}\right)^{ce} f_{ed} = \delta^c_d$ gives $\partial_b \left(f^{-1}\right)^{cd} = -\left(f^{-1}\right)^{ce} \left(f^{-1}\right)^{fd} \partial_b f_{ef}$, and substituting this into the cyclic sum turns the three second-derivative groups into three copies of the closedness relation contracted with three inverses; they cancel in the same pattern as before. □

Remark 25.23 (Why this is more than a change of variables). Three things are gained. First, the brackets are an *output*: one writes down a first-order action, inverts a matrix, and the bracket algebra follows, with no need to guess which variables are conjugate to which. Second, Jacobi is automatic — it was bought by writing f as $\mathrm{d}a$, so any theory presented in the form Equation (25.22) carries a legitimate Poisson structure by construction. Third, and this is the point of the construction, when f_{cd} is *not* invertible the equations Equation (25.24) do not determine all the velocities: contracting with a null vector z^c of f, $z^c f_{cd} = 0$, gives $z^c \partial_c \mathcal{H} = 0$, a relation among the coordinates alone — a constraint. Each null direction is either such a constraint or a gauge direction along which the motion is undetermined. Solving the constraints, substituting back into Equation (25.22) and repeating until f is invertible is the Faddeev–Jackiw reduction; it is an alternative to the Dirac algorithm of Section 26.2, and where both apply they give the same final brackets, the Dirac bracket Equation (26.25) being the inverse of f restricted to the constraint surface. The Dirac route classifies constraints first and quantizes second; the Faddeev–Jackiw route eliminates them first and never needs the classification. Neither is uniformly shorter, and the reason to know both is that the bookkeeping of one is often trivial where the other's is not.

Example 25.24 (A charged particle in a strong magnetic field)**.** Take a particle of charge e and mass m confined to the $x_1 x_2$ plane of three-dimensional space, in a uniform magnetic field $\boldsymbol{B} = B\boldsymbol{e}_3$ described in the symmetric gauge $\boldsymbol{A} = \frac{1}{2} B\,(-x_2, x_1, 0)$, with a potential $V(x_1, x_2)$. Its Lagrangian is

$$\mathcal{L} = \frac{m}{2}\left(\dot{x}_1^2 + \dot{x}_2^2\right) + \frac{eB}{2}\left(x_1 \dot{x}_2 - x_2 \dot{x}_1\right) - V\,. \tag{25.26}$$

When the field is strong enough that the cyclotron energy $\hbar eB/m$ dwarfs every other scale in the problem, the kinetic term may be dropped — the formal limit $m \to 0$ at fixed B — and what remains is already of the form Equation (25.22) in the two coordinates $\xi = (x_1, x_2)$, with no momenta at all:

$$a_1 = -\frac{eB}{2} x_2\,, \qquad a_2 = +\frac{eB}{2} x_1\,, \qquad \mathcal{H} = V\,. \tag{25.27}$$

Then $f_{12} = \partial_1 a_2 - \partial_2 a_1 = eB$, which is invertible for $B \neq 0$, and Equation (25.25) gives at once

$$\{x_1, x_2\} = -\frac{1}{eB}\,, \tag{25.28}$$

with SI unit $\mathrm{m}^2/(\mathrm{J\,s})$. The two Cartesian coordinates of the same particle are canonically conjugate to each other: the plane has become its own phase space, of half the dimension one started with, which is what the vanishing of m did to the count.

Rests on Theorem 25.22 and Equation (25.25).

Proposition 25.25 (Noncommuting coordinates and the Landau degeneracy)**.** *Applying Postulate 25.27 to Equation* (25.28) *gives*

$$[\hat{x}_1, \hat{x}_2] = -\mathrm{i}\,\frac{\hbar}{eB} = -\mathrm{i}\,\ell_B^2\,, \qquad \ell_B = \sqrt{\frac{\hbar}{eB}}\,, \tag{25.29}$$

ℓ_B *being the* magnetic length*. The number of independent states per unit area is then bounded by*

$$n = \frac{1}{2\pi\ell_B^2} = \frac{eB}{h} = \frac{B}{\Phi_0}\,, \qquad \Phi_0 = \frac{h}{e} = 4.135\,667\,696 \times 10^{-15}\,\mathrm{Wb}\,, \tag{25.30}$$

which is the observed degeneracy of a Landau level.

Rests on Equation (25.28) *and Postulate 25.27.*

Derivation. *Derives Proposition 25.25.*

Equation (25.29) is Equation (25.31) applied to Equation (25.28). Two operators obeying a commutation relation of the canonical form $\left[\hat{A}, \hat{B}\right] = \mathrm{i}\lambda$ obey the uncertainty relation $\Delta A\,\Delta B \geq |\lambda|/2$ of Section 77.6.3, so here $\Delta x_1\,\Delta x_2 \geq \ell_B^2/2$: a state cannot be localized in the plane to better than an area $2\pi\ell_B^2/2\pi$, and the count of orthogonal states per unit area is bounded by $1/2\pi\ell_B^2$. Substituting $\ell_B^2 = \hbar/eB$ and $\hbar = h/2\pi$ gives Equation (25.30), the flux quantum $\Phi_0 = h/e$ being exact in the SI, since h and e are both defining constants (Appendix B). The bound is an inequality because an uncertainty relation is one; that it is *saturated* — that a Landau level holds exactly eB/h states per unit area and not fewer — does not follow from this argument and requires the explicit eigenfunctions, which are obtained in Chapter 128. What the bracket settles here is the scale, and the scale is what the experiment measures.

Numerically, at $B = 10\,\mathrm{T}$ — a field routinely available in a superconducting laboratory magnet — $\ell_B = 8.11\,\mathrm{nm}$ and $n = 2.418 \times 10^{15}/\mathrm{m}^2$, that is $2.418 \times 10^{11}/\mathrm{cm}^2$, which is the carrier density scale at which the integer quantum Hall plateaux appear (Chapter 128). □

Remark 25.26 (What the example does and does not show)*.* The limit $m \to 0$ is a device for isolating the lowest Landau level, not a physical statement: the electron's mass is $(9.109\,383\,701\,5 \pm 0.000\,000\,002\,8) \times 10^{-31}\,\mathrm{kg}$ and does not vanish. What the Faddeev–Jackiw reading makes visible is that the projection onto one Landau level is a *reduction of phase space*, and that the noncommutativity Equation (25.29) of the surviving coordinates is the bracket of the reduced space rather than a new postulate. The same answer follows from the Dirac algorithm by treating the two momenta as second-class constraints and computing Equation (26.25); the arithmetic there is a 2×2 inversion as well, and the agreement is Remark 25.23 in action rather than a coincidence.

25.4 Canonical quantization: the correspondence rule

Postulate 25.27 (Dirac's correspondence rule)**.** To each classical observable u associate a self-adjoint operator $\hat{u}$ on a Hilbert space, in such a way that the Poisson bracket goes over into the commutator:

$$\{u, v\} \longmapsto \frac{1}{\mathrm{i}\hbar}\,[\hat{u}, \hat{v}]\ . \tag{25.31}$$

Rests on Theorem 25.3 and Equation (22.41)*.*

The status of Postulate 25.27 in this treatise is fixed by Chapter 77, where the quantum theory is set out on its own terms: the postulates there are about states, observables and measurement, and the correspondence rule is used in Section 77.5 as the heuristic that tells one *which* operator to write for a given classical quantity. It is a bridge from a theory that is a limit to the theory it is a limit of, and it is therefore expected to be imperfect. Section 25.5 shows that it is.

Remark 25.28 (The constant on the right-hand side is measured). Antisymmetry, bilinearity and the Jacobi identity are shared by the bracket and the commutator, so a map taking one to the other must carry a constant of proportionality, and dimensional analysis fixes what kind of constant. A Poisson bracket $\{u, v\}$ has the SI dimension of $[u]\,[v]$ divided by an action, since q and p pair to an action; a commutator $[\hat{u}, \hat{v}]$ has the dimension of $[u]\,[v]$ outright. The factor relating them is therefore an action, and Equation (25.31) would be dimensionally consistent for *any* action whatever. That the action in question is Planck's constant,

$$\hbar = \frac{h}{2\pi}\,, \qquad h = 6.626\,070\,15\times10^{-34}\,\mathrm{J\,s}\ \text{(exact)}\,, \qquad \hbar = 1.054\,571\,817\times10^{-34}\,\mathrm{J\,s}\,, \tag{25.32}$$

is not a mathematical statement and cannot be derived here: it is fixed by measurement — historically by the black-body spectrum and the photoelectric effect (Chapters 68 and 69), and today by the definition of the kilogram, which makes h exact and $\hbar$ exact as a defined ratio, quoted above to ten significant figures. The factor i is likewise not free: it is what makes $\hat{u}$ self-adjoint whenever u is real, since the commutator of two self-adjoint operators is anti-self-adjoint.

Proposition 25.29 (Canonical commutation relations)**.** *Applied to the fundamental brackets Equation* (22.44)*, Equation* (25.31) *gives*

$$\left[\hat{q}^a, \hat{q}^b\right] = 0\,, \qquad [\hat{p}_a, \hat{p}_b] = 0\,, \qquad [\hat{q}^a, \hat{p}_b] = \mathrm{i}\hbar\,\delta^a_b\,. \tag{25.33}$$

Rests on Postulate 25.27 and Equation (22.44)*.*

Derivation. *Derives Proposition 25.29.*

Read Equation (22.44) through Equation (25.31) term by term. The first two brackets vanish, so the corresponding commutators do. The third is $\{q^a, p_b\} = \delta^a_b$, a multiple of the constant function 1, whose operator image is $\mathbb{1}$ by the second requirement of Theorem 25.38; hence $[\hat{q}^a, \hat{p}_b] = \mathrm{i}\hbar\,\delta^a_b\,\mathbb{1}$, the identity being suppressed in Equation (25.33) as usual. Both sides carry the SI unit J s, as they must. □

Corollary 25.30 (The relation cannot be realized by matrices)**.** *No pair of operators on a finite-dimensional space, and more generally no pair of* bounded *operators on a Hilbert space, satisfies Equation* (25.33)*. Position and momentum are therefore necessarily unbounded, and necessarily defined on domains smaller than the whole space.*

Rests on Equation (25.33) *and Theorem 12.75.*

Derivation. *Derives Corollary 25.30.*

In finite dimension take the trace of Equation (25.33): the trace of a commutator vanishes, while the trace of $\mathrm{i}\hbar\mathbb{1}$ is $\mathrm{i}\hbar$ dim, which is non-zero. The general bounded statement is Theorem 12.75, and the consequences for domains are Corollary 12.76. This is the first place where the quantum theory demands more than linear algebra, and it is why the Hilbert-space apparatus of Chapter 12 is not optional decoration. □

Proposition 25.31 (The Heisenberg equation is the evolution equation)**.** *Applied to the classical evolution equation Equation* (22.50)*, Equation* (25.31) *gives the Heisenberg equation of motion*

$$\frac{\mathrm{d}\hat{A}}{\mathrm{d}t} = \frac{1}{\mathrm{i}\hbar}\left[\hat{A}, \hat{\mathcal{H}}\right] + \frac{\partial \hat{A}}{\partial t}\,, \tag{25.34}$$

and a quantum constant of motion is an observable commuting with the Hamiltonian, in exact parallel with Proposition 22.37.

Rests on Postulate 25.27 and Equation (22.50)*.*

Derivation. *Derives Proposition 25.31.*

The evolution equation Equation (22.50) reads $\mathrm{d}U/\mathrm{d}t = \{U, \mathcal{H}\} + \partial U/\partial t$. Its first term is a Poisson bracket and goes over into $\left[\hat{A}, \hat{\mathcal{H}}\right]/\mathrm{i}\hbar$ by Equation (25.31); its second is a derivative with respect to the explicit time slot, which the correspondence does not touch, since t is a parameter and not an observable. This gives Equation (25.34). Setting $\partial\hat{A}/\partial t = 0$, the operator is constant in time precisely when it commutes with $\hat{\mathcal{H}}$, which is the operator form of Proposition 22.37.

Two remarks keep the parallel honest. First, this is a transfer of a postulate, not a derivation: Equation (25.34) is equivalent to the Schrödinger equation of Section 77.1.6 and stands or falls with it. Second, the transfer is legitimate here, and not merely suggestive, because $\mathcal{H}$ enters only through a single bracket — so Theorem 25.38 has nothing to obstruct as long as $\hat{\mathcal{H}}$ has been fixed by some other means. □

Theorem 25.32 (Ehrenfest)**.** *For $\hat{\mathcal{H}} = \hat{\boldsymbol{p}}^2/2m + V(\hat{\boldsymbol{x}})$ the expectation values obey*

$$m\frac{\mathrm{d}}{\mathrm{d}t}\langle\hat{\boldsymbol{x}}\rangle = \langle\hat{\boldsymbol{p}}\rangle\,, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\langle\hat{\boldsymbol{p}}\rangle = -\langle\boldsymbol{\nabla}V(\hat{\boldsymbol{x}})\rangle\,. \tag{25.35}$$

The expectation values satisfy the classical equations of motion with the force evaluated as the mean of the force, not *as the force at the mean; the two agree only when the potential is at most quadratic, or when the state is narrow on the scale over which $\boldsymbol{\nabla}V$ varies.*

Rests on Equations (25.33) and (25.34).

Derivation. *Derives Theorem 25.32.*

Both statements are Equation (25.34) evaluated in a state, with $\partial\hat{A}/\partial t = 0$ for $\hat{A} = \hat{x}_i$ and $\hat{A} = \hat{p}_i$. Two commutators are needed. For the first, the commutator with a fixed operator is a derivation of the operator product, $\left[\hat{A}, \hat{B}\hat{C}\right] = \hat{B}\left[\hat{A}, \hat{C}\right] + \left[\hat{A}, \hat{B}\right]\hat{C}$, so with $[\hat{x}_i, \hat{p}_j] = \mathrm{i}\hbar\delta_{ij}$ from Equation (25.33),

$$[\hat{x}_i, \hat{p}_j\hat{p}_j] = \hat{p}_j\,[\hat{x}_i, \hat{p}_j] + [\hat{x}_i, \hat{p}_j]\,\hat{p}_j = 2\mathrm{i}\hbar\,\hat{p}_i\,, \tag{25.36}$$

while $[\hat{x}_i, V(\hat{\boldsymbol{x}})] = 0$ because V is built from the $\hat{x}_j$, which commute among themselves. Hence $\mathrm{d}\langle\hat{x}_i\rangle/\mathrm{d}t = \left\langle\left[\hat{x}_i, \hat{\mathcal{H}}\right]\right\rangle/\mathrm{i}\hbar = \langle\hat{p}_i\rangle/m$, the first of Equation (25.35).

For the second, $[\hat{p}_i, \hat{\boldsymbol{p}}^2] = 0$, and the same derivation property applied repeatedly to a monomial $\hat{x}_{j_1}\cdots\hat{x}_{j_n}$ gives

$$[\hat{p}_i, V(\hat{\boldsymbol{x}})] = -\mathrm{i}\hbar\left(\frac{\partial V}{\partial x_i}\right)(\hat{\boldsymbol{x}})\,, \tag{25.37}$$

since each factor contributes $[\hat{p}_i, \hat{x}_j] = -\mathrm{i}\hbar\delta_{ij}$ in turn, which is exactly the product rule for $\partial/\partial x_i$. The identity therefore holds for polynomials, and extends to any V reachable from them — by a convergent power series, or in general through the functional calculus of Definition 12.60 for a self-adjoint $\hat{\boldsymbol{x}}$. Then $\mathrm{d}\langle\hat{p}_i\rangle/\mathrm{d}t = \left\langle\left[\hat{p}_i, \hat{\mathcal{H}}\right]\right\rangle/\mathrm{i}\hbar = -\langle\partial_i V\rangle$, the second of Equation (25.35).

The final sentence is the substantive one. Expand $\partial_i V$ about the mean position $\langle\hat{\boldsymbol{x}}\rangle$ and take the expectation term by term. The linear term drops, because the mean of $\hat{x}_j - \langle\hat{x}_j\rangle$ is zero by construction, and what is left is

$$\left\langle\frac{\partial V}{\partial x_i}\right\rangle = \left.\frac{\partial V}{\partial x_i}\right|_{\langle\hat{\boldsymbol{x}}\rangle} + \frac{1}{2}\left.\frac{\partial^3 V}{\partial x_i\,\partial x_j\,\partial x_k}\right|_{\langle\hat{\boldsymbol{x}}\rangle}\mathrm{cov}\,(\hat{x}_j, \hat{x}_k) + \cdots\,, \tag{25.38}$$

with $\mathrm{cov}(\hat{x}_j, \hat{x}_k) = \frac{1}{2}\langle\{\hat{x}_j, \hat{x}_k\}\rangle - \langle\hat{x}_j\rangle\langle\hat{x}_k\rangle$. If V is at most quadratic every third derivative vanishes identically and Equation (25.35) becomes the classical equation exactly, whatever the state; otherwise the leading correction is a third derivative of the potential times the variance of the position, and it is not small merely because $\hbar$ is small — it is small only while the wave packet stays narrow. □

Remark 25.33 (Ehrenfest's theorem is not the classical limit). Equation (25.38) is the reason Theorem 25.32 does not, by itself, establish that quantum mechanics reproduces classical mechanics. In an anharmonic potential a packet spreads, the variance in Equation (25.38) grows, and the expectation values then follow no classical trajectory at all; the theorem remains true and says nothing useful. What is needed in addition is a mechanism that keeps the state narrow, and for a macroscopic body that mechanism is decoherence through coupling to an environment [2055], treated in Chapter 87. Ehrenfest's original note [573] claims no more than this: its title asks about the *approximate* validity of classical mechanics within quantum mechanics.

Theorem 25.34 (Stone–von Neumann)**.** *For a system with finitely many degrees of freedom, every irreducible representation of the canonical commutation relations in Weyl's exponentiated form*

$$\mathrm{e}^{\mathrm{i}a\hat{q}}\,\mathrm{e}^{\mathrm{i}b\hat{p}} = \mathrm{e}^{-\mathrm{i}\hbar ab}\,\mathrm{e}^{\mathrm{i}b\hat{p}}\,\mathrm{e}^{\mathrm{i}a\hat{q}} \tag{25.39}$$

is unitarily equivalent to the Schrödinger representation on $L^2(\mathbb{R}^f)$. The quantum kinematics is therefore fixed, up to equivalence, by the classical bracket algebra alone.

Rests on Equation (25.33), Theorem 12.66, and Definition 12.64.

Full derivation: Appendix A.49 (p. 3051).

Derives Theorem 25.34.

Remark 25.35 (The proof is functional analysis, and Part II owes its home). The argument written out in the appendix builds, from the Weyl operators of Equation (25.39), a Gaussian-smeared average that turns out to be a rank-one projector onto a vacuum vector, and then shows that the whole representation space is generated from that vector by the Weyl operators, so that any two irreducible representations are carried onto one another by a unitary. Nothing in it is physics: it is a statement about the Weyl algebra, and it belongs beside the Hilbert-space material of Chapter 12, next to Theorem 12.66, which is what carries the passage between the exponentiated form and the unbounded generators. That Chapter 12 does not carry it is recorded here as a debt of Part II, not as a gap of this chapter; the appendix proof is written so that moving it there would change nothing but its address.

Remark 25.36 (The exponentiated form is not a technicality). Corollary 25.30 forces $\hat{q}$ and $\hat{p}$ to be unbounded, and by Corollary 12.76 an unbounded operator carries a domain that is not the whole space. Two representations of Equation (25.33) can then agree on every vector where both sides are defined and still be inequivalent, because the relation says nothing about what happens outside the common domain. The Weyl form Equation (25.39) repairs this: $\mathrm{e}^{\mathrm{i}a\hat{q}}$ and $\mathrm{e}^{\mathrm{i}b\hat{p}}$ are *unitary*, hence bounded and everywhere defined, and Theorem 12.66 guarantees that a strongly continuous one-parameter unitary group determines a unique self-adjoint generator and conversely. Momentum on a half-line (Example 12.82) and momentum on an interval (Example 12.81) are the standard warnings: the first has no self-adjoint extension at all, the second a whole circle of them, and in neither case does the formal commutation relation notice. The exponentiated form is what makes the uniqueness statement true rather than nearly true.

The proof itself is missing from Part II, where it belongs: Section 12.4.3 establishes Stone's theorem but not the uniqueness of the Weyl representation. Nothing in this chapter rests on the proof — only on the statement, which is used in Theorem 25.38 to say that hypothesis (3) leaves no freedom in the operators of degree at most one.

Remark 25.37 (Where the theorem fails, and why that is physics). Theorem 25.34 assumes finitely many degrees of freedom. A field has infinitely many, and the theorem then fails: inequivalent irreducible representations of the canonical commutation relations exist in profusion, and the choice among them is a physical one. This is not a pathology to be

regretted; it is what allows a quantum field theory to have inequivalent vacua, and it is the origin of superselection sectors, of spontaneous symmetry breaking as a statement about representations (Chapter 104), and of Haag's theorem (Chapter 109).

25.5 The obstruction: what the correspondence cannot be

Postulate 25.27 is stated above as a postulate rather than as a definition or a theorem. That is deliberate, and this section says why: as a definition it would be inconsistent, and as a theorem it is false.

Theorem 25.38 (Groenewold–van Hove)**.** *There is no linear map $u \mapsto \hat{u}$ from the polynomials in (q^a, p_a) to self-adjoint operators satisfying all of*

1. *$\{u, v\} \mapsto \frac{1}{\mathrm{i}\hbar}[\hat{u}, \hat{v}]$ for all u, v;*
2. *$\hat{1} = \mathbb{1}$;*
3. *$\hat{q}^a$ and $\hat{p}_a$ act irreducibly on the Hilbert space.*

The obstruction appears already for polynomials of degree three and four in one degree of freedom: the requirements are consistent on the subalgebra of polynomials of degree at most two and inconsistent on any extension of it.

Rests on Equation (25.31) and Theorems 5.155 and 25.34.

The proof is elementary and takes two steps. First, the hypotheses leave no freedom at all in the images of the polynomials of degree at most three: every one of them is forced to be the symmetrized, or Weyl, ordering. Second, one single classical observable of degree four can be written as a Poisson bracket of degree-three polynomials in two different ways, and the two forced answers disagree.

Lemma 25.39 (The low-degree images are forced)**.** *Take one degree of freedom and write $\hat{q}, \hat{p}$ for the images of q, p. Under hypotheses (1)–(3) of Theorem 25.38,*

$$\widehat{q^2} = \hat{q}^2\,, \qquad \widehat{p^2} = \hat{p}^2\,, \qquad \widehat{qp} = \tfrac{1}{2}\{\hat{q}, \hat{p}\}\,, \tag{25.40}$$

$$\widehat{q^3} = \hat{q}^3\,, \qquad \widehat{p^3} = \hat{p}^3\,, \qquad \widehat{q^2p} = \tfrac{1}{2}\left\{\hat{q}^2, \hat{p}\right\}\,, \quad \widehat{qp^2} = \tfrac{1}{2}\left\{\hat{q}, \hat{p}^2\right\}\,. \tag{25.41}$$

Rests on Theorems 5.155 and 25.34.

Derivation. *Derives Lemma 25.39.*

Throughout, hypothesis (1) is applied only to the pairs written down, and $[\hat{q}, \hat{p}] = \mathrm{i}\hbar$ from Proposition 25.29. The tool used repeatedly is:

Schur step. If an operator D commutes with both $\hat{q}$ and $\hat{p}$ then $D = \lambda\mathbb{1}$ for some complex λ. This is Schur's lemma Theorem 5.155 applied to the irreducible representation of hypothesis (3); Theorem 25.34 says that representation is the Schrödinger one, in which the statement is that an operator commuting with multiplication by x and with differentiation is a constant.

Quadratics. From $\{q^2, q\} = 0$ and $\{q^2, p\} = 2q$, hypothesis (1) requires $\left[\widehat{q^2}, \hat{q}\right] = 0$ and $\left[\widehat{q^2}, \hat{p}\right] = 2\mathrm{i}\hbar\,\hat{q}$. The operator $\hat{q}^2$ satisfies both, so their difference passes the Schur step and $\widehat{q^2} = \hat{q}^2 + \lambda$. Symmetrically $\widehat{p^2} = \hat{p}^2 + \nu$. Writing $\hat{S} := \frac{1}{2}\{\hat{q}, \hat{p}\} = \hat{q}\hat{p} - \frac{1}{2}\mathrm{i}\hbar$, the same argument applied to $\{qp, q\} = -q$ and $\{qp, p\} = p$ gives $\widehat{qp} = \hat{S} + \mu$.

The three constants are removed by three more brackets. From $\{q^2, p^2\} = 4qp$ and $[\hat{q}^2, \hat{p}^2] = 2\mathrm{i}\hbar\,\{\hat{q}, \hat{p}\}$, the additive constants dropping out of a commutator,

$$\widehat{qp} = \frac{1}{4\mathrm{i}\hbar}\left[\widehat{q^2}, \widehat{p^2}\right] = \frac{1}{4\mathrm{i}\hbar}\left[\hat{q}^2, \hat{p}^2\right] = \hat{S}\,,$$

so $\mu = 0$. From $\{q^2, qp\} = 2q^2$ and $\left[\hat{q}^2, \hat{S}\right] = 2\mathrm{i}\hbar\,\hat{q}^2$,

$$\widehat{q^2} = \frac{1}{2\mathrm{i}\hbar}\left[\widehat{q^2}, \widehat{qp}\right] = \frac{1}{2\mathrm{i}\hbar}\left[\hat{q}^2, \hat{S}\right] = \hat{q}^2\,,$$

so $\lambda = 0$; and $\{p^2, qp\} = -2p^2$ with $\left[\hat{p}^2, \hat{S}\right] = -2\mathrm{i}\hbar\,\hat{p}^2$ gives $\nu = 0$ the same way. This is Equation (25.40).

Cubics. From $\{q^3, q\} = 0$ and $\{q^3, p\} = 3q^2$, and using $\widehat{q^2} = \hat{q}^2$ just established, the Schur step gives $\widehat{q^3} = \hat{q}^3 + \kappa$. Then $\{q^3, qp\} = 3q^3$ with $\left[\hat{q}^3, \hat{S}\right] = 3\mathrm{i}\hbar\,\hat{q}^3$ forces $3\hat{q}^3 = 3\,(\hat{q}^3 + \kappa)$, so $\kappa = 0$; symmetrically $\widehat{p^3} = \hat{p}^3$, using $\{p^3, qp\} = -3p^3$. The mixed cubics need no Schur step at all, being brackets of things already fixed: $\{q^3, p^2\} = 6q^2p$ and $[\hat{q}^3, \hat{p}^2] = 3\mathrm{i}\hbar\,\{\hat{q}^2, \hat{p}\}$ give

$$\widehat{q^2p} = \frac{1}{6\mathrm{i}\hbar}\left[\hat{q}^3, \hat{p}^2\right] = \tfrac{1}{2}\left\{\hat{q}^2, \hat{p}\right\}\,,$$

and $\{q^2, p^3\} = 6qp^2$ with $[\hat{q}^2, \hat{p}^3] = 2\mathrm{i}\hbar\,(\hat{q}\hat{p}^2 + \hat{p}\hat{q}\hat{p} + \hat{p}^2\hat{q}) = 6\mathrm{i}\hbar\,(\hat{q}\hat{p}^2 - \mathrm{i}\hbar\hat{p})$ gives $\widehat{qp^2} = \hat{q}\hat{p}^2 - \mathrm{i}\hbar\hat{p} = \frac{1}{2}\,\{\hat{q}, \hat{p}^2\}$. This is Equation (25.41). □

Derivation. *Derives Theorem 25.38.*

Consider the classical observable q^2p^2. It is a Poisson bracket of cubics in two ways:

$$\left\{q^3, p^3\right\} = 9\,q^2p^2\,, \qquad \left\{q^2p, qp^2\right\} = 3\,q^2p^2\,, \tag{25.42}$$

both read off Equation (22.41) directly — for the second, $\partial_q\,(q^2p)\,\partial_p\,(qp^2) - \partial_p\,(q^2p)\,\partial_q\,(qp^2) = 4q^2p^2 - q^2p^2$. Hypothesis (1) applied to each of Equation (25.42), with the images supplied by Lemma 25.39, gives two expressions for the same operator $\widehat{q^2p^2}$.

For the first route, $[\hat{q}^3, \hat{p}^3] = 3\mathrm{i}\hbar\,(\hat{q}^2\hat{p}^2 + \hat{p}\hat{q}^2\hat{p} + \hat{p}^2\hat{q}^2)$, and moving every $\hat{p}$ to the right with $\hat{p}\hat{q}^2 = \hat{q}^2\hat{p} - 2\mathrm{i}\hbar\hat{q}$,

$$\hat{p}\hat{q}^2\hat{p} = \hat{q}^2\hat{p}^2 - 2\mathrm{i}\hbar\,\hat{q}\hat{p}\,, \qquad \hat{p}^2\hat{q}^2 = \hat{q}^2\hat{p}^2 - 4\mathrm{i}\hbar\,\hat{q}\hat{p} - 2\hbar^2\,,$$

so that $[\hat{q}^3, \hat{p}^3] = 3\mathrm{i}\hbar\,(3\hat{q}^2\hat{p}^2 - 6\mathrm{i}\hbar\,\hat{q}\hat{p} - 2\hbar^2)$ and

$$\widehat{q^2p^2} = \frac{1}{9\mathrm{i}\hbar}\left[\hat{q}^3, \hat{p}^3\right] = \hat{q}^2\hat{p}^2 - 2\mathrm{i}\hbar\,\hat{q}\hat{p} - \frac{2}{3}\hbar^2\,. \tag{25.43}$$

For the second route write the two cubics in the compact form supplied by Lemma 25.39, $\widehat{q^2p} = \hat{q}^2\hat{p} - \mathrm{i}\hbar\hat{q}$ and $\widehat{qp^2} = \hat{q}\hat{p}^2 - \mathrm{i}\hbar\hat{p}$. Expanding the commutator in four pieces, and using $[\hat{q}^2\hat{p}, \hat{q}\hat{p}^2] = 3\mathrm{i}\hbar\,\hat{q}^2\hat{p}^2 + 2\hbar^2\hat{q}\hat{p}$ for the leading one,

$$\left[\widehat{q^2p}, \widehat{qp^2}\right] = 3\mathrm{i}\hbar\,\hat{q}^2\hat{p}^2 + 6\hbar^2\hat{q}\hat{p} - \mathrm{i}\hbar^3\,,$$

the three $\hbar^2\hat{q}\hat{p}$ contributions coming one from the leading piece and one from each cross term, and

$$\widehat{q^2p^2} = \frac{1}{3\mathrm{i}\hbar}\left[\widehat{q^2p}, \widehat{qp^2}\right] = \hat{q}^2\hat{p}^2 - 2\mathrm{i}\hbar\,\hat{q}\hat{p} - \frac{1}{3}\hbar^2\,. \tag{25.44}$$

Subtracting Equation (25.44) from Equation (25.43),

$$0 = -\frac{2}{3}\hbar^2 + \frac{1}{3}\hbar^2 = -\frac{\hbar^2}{3}\,\mathbb{1}\,, \tag{25.45}$$

which is false, $\hbar$ being a non-zero action Equation (25.32). No map satisfying (1)–(3) exists.

Nothing in the argument used more than hypothesis (1) on the pairs listed, all of degree at most three, so the conclusion is stronger than the statement: the correspondence already fails if it is demanded only of the polynomials of degree at most three. It does *not* fail

on the polynomials of degree at most two, which close under the bracket — the bracket of a monomial of degree m with one of degree n has degree $m+n-2$ — and on which Equation (25.40) is not merely forced but consistent, as the brackets computed in the proof of Lemma 25.39 verify. Degree two is exactly where the correspondence is exact, and one step beyond it is where it breaks. □

Remark 25.40 (Reading the defect). The mismatch Equation (25.45) is of order $\hbar^2$, not $\hbar$. That is not an accident of the example: an ordering difference of order $\hbar$ would show up already in Equation (25.41), where the two routes agree, and it is the second order that the Poisson bracket cannot see, since it is the first term of the expansion Equation (25.52). Groenewold's 1946 paper and van Hove's 1951 memoir — neither of which is in this bibliography, so the attribution here rests on the computation above and not on a source the reader can check — go further, showing that no subalgebra properly containing the quadratics can be quantized consistently, so that restricting the observables does not rescue the rule either. Note also what is *not* shown: that quantum mechanics is inconsistent, or that $\widehat{q^2p^2}$ has no sensible definition. Both routes produce perfectly good self-adjoint operators. What fails is the demand that one assignment serve every classical observable at once while respecting the bracket.

Corollary 25.41 (Quantization is not a functor)**.** *No prescription assigns operators to all classical observables while preserving the bracket. Every actual quantization therefore either restricts the class of observables it treats, or accepts an ordering ambiguity, or deforms the bracket. Consequently the classical theory does not determine its quantum counterpart: quantization is a hypothesis about nature, and the choices it leaves open are settled by experiment, not by mathematics.*

Rests on Theorem 25.38.

25.5.1 Ordering

Corollary 25.41 says that something must be given up. This subsection makes concrete what is given up when one keeps the operators and abandons the bracket: an *ordering ambiguity*, one for every classical observable in which coordinates and momenta are multiplied together.

Definition 25.42 (Ordering rule)**.** An *ordering rule* is a linear map $u \mapsto \hat{u}$ from the polynomials in (q^a, p_a) to operators which sends 1 to $\mathbb{1}$, sends q^a and p_a to $\hat{q}^a$ and $\hat{p}_a$, and sends a real polynomial to a self-adjoint operator. By Theorem 25.38 no such rule can also satisfy Equation (25.31) for all pairs, so a rule is a choice, and different choices are different physical hypotheses until an experiment separates them.

Rests on Corollary 25.41 and Postulate 25.27.

Proposition 25.43 (The size of the ambiguity)**.** *The classical monomial qp in one degree of freedom has the three obvious candidate images*

$$\hat{q}\hat{p}\,, \qquad \hat{p}\hat{q} = \hat{q}\hat{p} - \mathrm{i}\hbar\,, \qquad \tfrac{1}{2}\{\hat{q},\hat{p}\} = \hat{q}\hat{p} - \tfrac{1}{2}\mathrm{i}\hbar\,, \tag{25.46}$$

of which only the third is self-adjoint. More generally, any two orderings of a monomial $q^m p^n$ differ by a polynomial in $\hat{q},\hat{p}$ of degree at most $m+n-2$, whose coefficients are integer multiples of powers of $\mathrm{i}\hbar$. All orderings therefore agree in the limit $\hbar \to 0$, and none of them can be preferred on dimensional grounds.

Rests on Equation (25.33) and Definition 25.42.

Derivation. *Derives Proposition 25.43.*

The three entries of Equation (25.46) follow from $[\hat{q},\hat{p}] = \mathrm{i}\hbar$; and $(\hat{q}\hat{p})^\dagger = \hat{p}\hat{q} \neq \hat{q}\hat{p}$, so neither of the first two is self-adjoint while the anticommutator manifestly is. For the general statement, any ordering of $q^m p^n$ can be turned into any other by transposing adjacent

factors one at a time, and each transposition replaces $\hat{p}\hat{q}$ by $\hat{q}\hat{p}$ at the cost of an additive $-\mathrm{i}\hbar$ multiplying the remaining $m+n-2$ factors. Iterating, the accumulated difference is a polynomial of degree at most $m+n-2$ with coefficients built from integers and powers of $\mathrm{i}\hbar$. Since $\hbar$ carries the SI unit J s and qp carries the same, the correction terms are dimensionally indistinguishable from the leading one, which is why no dimensional argument can select an ordering. □

Definition 25.44 (Weyl and normal ordering)**.** *Weyl ordering* is the rule that symmetrizes: $q^m p^n$ is sent to the average of the $\binom{m+n}{m}$ distinct orderings of m factors $\hat{q}$ and n factors $\hat{p}$. Equivalently, and more usefully, it is fixed by requiring

$$\mathrm{e}^{\mathrm{i}(\alpha q+\beta p)} \longmapsto \mathrm{e}^{\mathrm{i}(\alpha \hat{q}+\beta \hat{p})} \tag{25.47}$$

for all real α of SI unit $1/\mathrm{m}$ and β of unit $\mathrm{s}/(\mathrm{kg\,m})$, and expanding both sides in powers of α and β. *Normal ordering* is the rule that writes the observable in terms of the ladder combinations $\hat{a}, \hat{a}^\dagger$ of Section 78.4.2 and moves every $\hat{a}^\dagger$ to the left of every $\hat{a}$.

Rests on Definition 25.42.

Example 25.45 (Weyl ordering collapses)**.** Weyl ordering of qp^2 is the average of the three arrangements, and it simplifies:

$$\tfrac{1}{3}\left(\hat{q}\hat{p}^2+\hat{p}\hat{q}\hat{p}+\hat{p}^2\hat{q}\right) = \hat{q}\hat{p}^2 - \mathrm{i}\hbar\hat{p} = \tfrac{1}{2}\left\{\hat{q},\hat{p}^2\right\}, \tag{25.48}$$

using $\hat{p}\hat{q} = \hat{q}\hat{p} - \mathrm{i}\hbar$ twice. This is the same operator that Lemma 25.39 obtained by an entirely different route — from the requirement that Equation (25.31) hold on cubics — and the agreement is the content of the statement that the correspondence is exact up to degree three. It stops being true at degree four, where Equations (25.43) and (25.44) disagree and Weyl ordering picks neither: the Weyl image of q^2p^2 is $\hat{q}^2\hat{p}^2 - 2\mathrm{i}\hbar\hat{q}\hat{p} - \tfrac{1}{2}\hbar^2$, which sits halfway between them.

Rests on Definition 25.44 and Lemma 25.39.

Proposition 25.46 (Normal ordering disagrees already at degree two)**.** *For the harmonic oscillator $\mathcal{H} = p^2/2m + \tfrac{1}{2}m\omega^2 q^2$, Weyl ordering gives $\hat{\mathcal{H}}_W = \hbar\omega\left(\hat{a}^\dagger\hat{a} + \tfrac{1}{2}\right)$ and normal ordering gives $\hat{\mathcal{H}}_N = \hbar\omega\,\hat{a}^\dagger\hat{a}$. The two differ by the zero-point energy*

$$\hat{\mathcal{H}}_W - \hat{\mathcal{H}}_N = \frac{\hbar\omega}{2}\,\mathbb{1}\,. \tag{25.49}$$

Normal ordering is therefore not merely a different arrangement of one monomial: it violates Equation (25.31) at degree two, where Lemma 25.39 shows the correspondence still has no freedom.

Rests on Definition 25.44 and Equation (25.40).

Derivation. *Derives Proposition 25.46.*

Write the classical amplitude $\alpha = \sqrt{m\omega/2\hbar}\,(q + \mathrm{i}p/m\omega)$, whose bracket is $\{\alpha, \bar{\alpha}\} = -\mathrm{i}/\hbar$ by direct substitution into Equation (22.41), so that Equation (25.31) gives $\left[\hat{a}, \hat{a}^\dagger\right] = \mathbb{1}$; and $\mathcal{H} = \hbar\omega\,\bar{\alpha}\alpha$ by expanding the square. Weyl ordering of q^2 and p^2 gives $\hat{q}^2$ and $\hat{p}^2$, whose combination is $\tfrac{1}{2}\hbar\omega\left\{\hat{a}^\dagger, \hat{a}\right\} = \hbar\omega\left(\hat{a}^\dagger\hat{a} + \tfrac{1}{2}\right)$; normal ordering of $\bar{\alpha}\alpha$ gives $\hat{a}^\dagger\hat{a}$ by definition. Subtracting gives Equation (25.49).

For the last sentence, express $\hat{q}^2$ in ladder operators: $\hat{q}^2 = (\hbar/2m\omega)\left(\hat{a}^2 + 2\hat{a}^\dagger\hat{a} + \mathbb{1} + \hat{a}^{\dagger 2}\right)$, while normal ordering of the classical q^2 drops the $\mathbb{1}$. So normal ordering sends q^2 to $\hat{q}^2 - \hbar/2m\omega$, contradicting Equation (25.40). The Dirac rule is not toothless below degree four: there it fixes the answer completely, additive constant included. □

Remark 25.47 (Whether the constant is observable). Equation (25.49) is a multiple of the identity, so it cancels from every transition frequency of a single oscillator and no measurement of that oscillator's spectrum alone can decide between the two rules. It is nevertheless decidable, and it has been decided, in two independent ways. Comparing *two* oscillators that share a force constant and differ in reduced mass makes the half-quantum visible as an isotope displacement of the whole vibrational ladder, which is Mulliken's measurement on the band spectrum of boron monoxide [515] and the evidence behind Phenomenon 78.6. And comparing the same field in two geometries makes the sum of the zero-point energies visible as a force, which is the Casimir effect of Section 99.8. Normal ordering survives in field theory not because it is right but because it is convenient (Section 99.2.4); the constant it discards is infinite there, and the honest statement is that only differences of it are ever used.

Proposition 25.48 (Where the ambiguity does and does not bite)**.** *For a Hamiltonian of the standard mechanical form* $\mathcal{H} = |\boldsymbol{p}|^2/2m + V(\boldsymbol{x})$ *with constant* m *there is no ordering ambiguity whatever: no term multiplies a coordinate by a momentum. The ambiguity appears in exactly three circumstances, and only the third is left open by physical requirements:*

1. Minimal coupling to a magnetic field. *The cross term of* $(\hat{\boldsymbol{p}} - e\boldsymbol{A}(\hat{\boldsymbol{x}}))^2$ *is ambiguous, but*
$$[\hat{p}_i, A_i(\hat{\boldsymbol{x}})] = -\mathrm{i}\hbar\, \boldsymbol{\nabla}\cdot\boldsymbol{A}\,, \tag{25.50}$$
so in the Coulomb gauge every ordering agrees, and in any other gauge the orderings differ by a term that a gauge transformation removes. The ambiguity is not physical.

2. Position-dependent mass. *For* $\mathcal{H} = |\boldsymbol{p}|^2/2m(\boldsymbol{x})$ *the orderings* $\hat{p}_i m^{-1}\hat{p}_i$, $\frac{1}{2}\{m^{-1}, \hat{p}_i\hat{p}_i\}$ *and their one-parameter interpolations are all self-adjoint and inequivalent, differing by terms in* $\hbar^2\boldsymbol{\nabla}^2 m^{-1}$*. Requiring that probability be conserved across a step in* m *selects* $\hat{p}_i m^{-1}\hat{p}_i$*; nothing weaker does.*

3. Curved configuration space. *Quantizing on a configuration manifold, the images of the classical kinetic energy differ by a multiple* $\xi\hbar^2 R$ *of the scalar curvature (Definition 13.153). No mathematical requirement fixes* ξ*.*

Rests on Proposition 25.43 and Equation (25.33)*.*

Derivation. *Derives Proposition 25.48.*

The first sentence is immediate: $|\hat{\boldsymbol{p}}|^2$ involves only momenta, which commute among themselves by Equation (25.33), and $V(\hat{\boldsymbol{x}})$ only coordinates, which do likewise.

(1) Equation (25.50) is Equation (25.37) with V replaced by A_i and the index summed. The two candidate cross terms $\hat{\boldsymbol{p}}\cdot\boldsymbol{A}$ and $\boldsymbol{A}\cdot\hat{\boldsymbol{p}}$ therefore differ by $-\mathrm{i}\hbar\,\boldsymbol{\nabla}\cdot\boldsymbol{A}$, which vanishes identically in the Coulomb gauge $\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0$, and in a general gauge is the divergence of a field that a gauge transformation may set to zero without changing $\boldsymbol{B}$.

(2) Each candidate is self-adjoint because it is built as a symmetric product of self-adjoint factors; they differ because $m^{-1}(\hat{\boldsymbol{x}})$ does not commute with $\hat{\boldsymbol{p}}$, and the difference is computed by applying Equation (25.37) twice, each application producing one factor $\mathrm{i}\hbar$ and one derivative of m^{-1}. The selection rule quoted is the statement that $\hat{p}_i m^{-1}\hat{p}_i$ is the only member of the family whose associated probability current is continuous where m jumps; the others lose or create probability at the step, which is not a matter of taste.

(3) is quoted, not derived: the curvature term is what distinguishes the Laplace–Beltrami operator from the other second-order self-adjoint operators with the same symbol, and the coefficient is a free parameter of the quantization. The point of listing it is that it is the case in which nothing in the mathematics decides, and the honest procedure is the one stated next. □

Remark 25.49 (How the ordering is actually fixed). In every case where the choice has physical consequences, it is settled by requiring agreement with something already measured — a spectrum, a scattering cross-section, a transport coefficient — and not by any property of the classical theory. That is Corollary 25.41 delivering exactly what it promised: the classical theory underdetermines the quantum one, so the residue must be measured. It is also the reason the ambiguity is far less troublesome in practice than in principle. The Hamiltonians of atomic, molecular and nuclear physics are overwhelmingly of the form $|\boldsymbol{p}|^2/2m + V$, for which Proposition 25.48 says there is nothing to choose; the ambiguity becomes a live experimental question mainly in effective theories, where a coordinate-dependent mass or a curved configuration space has been produced by eliminating degrees of freedom, and where the eliminated degrees of freedom are what really fix the answer.

25.5.2 Deformation quantization and the Moyal bracket

Definition 25.50 (Star product and Moyal bracket)**.** For phase-space functions of one Cartesian pair (x, p), the *star product* is the formal series

$$u \star v = \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{\mathrm{i}\hbar}{2}\right)^n \sum_{k=0}^{n} (-1)^k \binom{n}{k} \left(\partial_x^{n-k} \partial_p^{k} u\right) \left(\partial_x^{k} \partial_p^{n-k} v\right) , \tag{25.51}$$

which is associative and noncommutative and which reproduces operator multiplication on Weyl symbols: if u and v are the symbols of $\hat{u}$ and $\hat{v}$ in the sense of Definition 25.44, then $u \star v$ is the symbol of $\hat{u}\hat{v}$. The *Moyal bracket* is

$$\{u, v\}_{\mathrm{M}} = \frac{1}{\mathrm{i}\hbar} (u \star v - v \star u) = \{u, v\} + \mathcal{O}(\hbar^2) , \tag{25.52}$$

the $n = 0$ terms of Equation (25.51) cancelling in the difference, the $n = 1$ terms giving $\mathrm{i}\hbar\,\{u, v\}$, and every even n cancelling because the nth term is symmetric under $u \leftrightarrow v$ for n even and antisymmetric for n odd. The correction to the Poisson bracket is therefore of order $\hbar^2$, never $\hbar$.

Rests on Definition 25.44 and Equation (22.41).

Remark 25.51 (What deformation quantization says). Equation (25.52) exhibits the quantum bracket as a deformation of the Poisson bracket in the parameter $\hbar$, with the classical bracket as the leading term. Read that way the relation between the theories is inverted from the usual telling: the classical Poisson bracket is the $\hbar \to 0$ limit of an exact quantum structure, rather than the quantum commutator being a modification of a classical one. This is the honest logical order—quantum mechanics is not derived from classical mechanics, and classical mechanics is a limit of it—and it is the order this treatise adopts.

25.5.3 Phase-space quantum mechanics

Definition 25.50 deformed the bracket. Deforming the *state* in the same spirit gives a formulation of quantum mechanics carried out entirely on phase space, with no operators and no Hilbert space in sight: observables are functions of $(\boldsymbol{x}, \boldsymbol{p})$, the state is a function of $(\boldsymbol{x}, \boldsymbol{p})$, and the only thing that distinguishes the theory from classical statistical mechanics is that products are taken with $\star$ instead of pointwise. It is an exact rewriting, not an approximation, and its value here is that it shows precisely which classical intuition survives.

Definition 25.52 (Wigner function)**.** Let $\hat{\rho}$ be the density operator of a system of one particle in three-dimensional space, with position-representation kernel $\rho(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x} | \, \hat{\rho} \, | \boldsymbol{x}' \rangle$. Its *Wigner function* is

$$W(\boldsymbol{x}, \boldsymbol{p}) = \frac{1}{(2\pi\hbar)^3} \int_{\mathbb{R}^3} \rho\!\left(\boldsymbol{x} + \frac{\boldsymbol{y}}{2}, \, \boldsymbol{x} - \frac{\boldsymbol{y}}{2}\right) \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{y}/\hbar} \, \mathrm{d}^3 y \, . \tag{25.53}$$

It is the Weyl symbol of $\hat{\rho}$ divided by $(2\pi\hbar)^3$, and carries the SI unit $1/(\mathrm{J}^3\,\mathrm{s}^3)$, the reciprocal of an action cubed, so that $W\,\mathrm{d}^3x\,\mathrm{d}^3p$ is dimensionless.

Rests on Definitions 17.29 and 25.44.

Theorem 25.53 (Properties of the Wigner function)**.** *For any density operator, W is real-valued and normalized, $\int W\,\mathrm{d}^3x\,\mathrm{d}^3p = 1$, and*

$$\int W(\boldsymbol{x},\boldsymbol{p})\,\mathrm{d}^3p = \rho(\boldsymbol{x},\boldsymbol{x})\,, \tag{25.54}$$

$$\int W(\boldsymbol{x},\boldsymbol{p})\,\mathrm{d}^3x = \tilde{\rho}(\boldsymbol{p},\boldsymbol{p})\,, \tag{25.55}$$

the two marginals being exactly the measured position and momentum distributions, $\tilde{\rho}$ denoting the kernel in the momentum representation. For any two states,

$$\operatorname{tr}(\hat{\rho}_1\hat{\rho}_2) = (2\pi\hbar)^3 \int W_1\,W_2\,\mathrm{d}^3x\,\mathrm{d}^3p\,, \tag{25.56}$$

and for a pure state $\hat{\rho} = |\psi\rangle\langle\psi|$,

$$|W(\boldsymbol{x},\boldsymbol{p})| \le \frac{1}{(\pi\hbar)^3} = 2.75\times10^{100}/(\mathrm{J}^3\,\mathrm{s}^3)\,. \tag{25.57}$$

Rests on Equation (25.53), Theorem 17.32, and Proposition 12.4.

Derivation. *Derives Theorem 25.53.*

Reality. Conjugating Equation (25.53) and substituting $\boldsymbol{y}\to-\boldsymbol{y}$ turns the kernel into $\rho(\boldsymbol{x}-\boldsymbol{y}/2,\boldsymbol{x}+\boldsymbol{y}/2)^*$, which is $\rho(\boldsymbol{x}+\boldsymbol{y}/2,\boldsymbol{x}-\boldsymbol{y}/2)$ because $\hat{\rho}$ is self-adjoint, and restores the exponential; so $W^* = W$.

Position marginal. Integrating Equation (25.53) over $\boldsymbol{p}$ and using $\int \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{y}/\hbar}\mathrm{d}^3p = (2\pi\hbar)^3\delta^3(\boldsymbol{y})$, which is the Fourier inversion formula of Section 17.2.1 read as an identity between tempered distributions (Definition 17.38), the delta collapses the $\boldsymbol{y}$ integral at $\boldsymbol{y}=0$ and leaves $\rho(\boldsymbol{x},\boldsymbol{x})$. This is Equation (25.54), and integrating it once more over $\boldsymbol{x}$ gives $\operatorname{tr}\hat{\rho} = 1$, the normalization.

Momentum marginal. Integrate over $\boldsymbol{x}$ and change variables to $\boldsymbol{u} = \boldsymbol{x}+\boldsymbol{y}/2$, $\boldsymbol{v} = \boldsymbol{x}-\boldsymbol{y}/2$, whose Jacobian has modulus 1 in each Cartesian pair. Then $\boldsymbol{y} = \boldsymbol{u}-\boldsymbol{v}$ and

$$\int W\,\mathrm{d}^3x = \frac{1}{(2\pi\hbar)^3}\iint \rho(\boldsymbol{u},\boldsymbol{v})\,\mathrm{e}^{-\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{u}/\hbar}\mathrm{e}^{+\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{v}/\hbar}\,\mathrm{d}^3u\,\mathrm{d}^3v = \tilde{\rho}(\boldsymbol{p},\boldsymbol{p})\,,$$

the double integral being exactly the momentum-representation kernel at coincident arguments, in the convention that carries one factor $(2\pi\hbar)^{-1/2}$ per coordinate. For a pure state this is $\left|\tilde{\psi}(\boldsymbol{p})\right|^2$.

Overlap. Take the product of two Wigner functions with integration variables $\boldsymbol{y}$ and $\boldsymbol{y}'$ and integrate over $\boldsymbol{p}$ first; the same delta identity forces $\boldsymbol{y}' = -\boldsymbol{y}$ and removes one factor $(2\pi\hbar)^{-3}$, leaving

$$\int W_1W_2\,\mathrm{d}^3p = \frac{1}{(2\pi\hbar)^3}\int \rho_1\big(\boldsymbol{x}+\tfrac{\boldsymbol{y}}{2},\boldsymbol{x}-\tfrac{\boldsymbol{y}}{2}\big)\,\rho_2\big(\boldsymbol{x}-\tfrac{\boldsymbol{y}}{2},\boldsymbol{x}+\tfrac{\boldsymbol{y}}{2}\big)\,\mathrm{d}^3y\,.$$

Integrating over $\boldsymbol{x}$ and changing to $(\boldsymbol{u},\boldsymbol{v})$ as before turns this into $(2\pi\hbar)^{-3}\iint \rho_1(\boldsymbol{u},\boldsymbol{v})\rho_2(\boldsymbol{v},\boldsymbol{u})\,\mathrm{d}^3u\,\mathrm{d}^3v$, which is $(2\pi\hbar)^{-3}\operatorname{tr}(\hat{\rho}_1\hat{\rho}_2)$. This is Equation (25.56).

Bound. For a pure state the integrand of Equation (25.53) is $\psi(\boldsymbol{x}+\boldsymbol{y}/2)\psi^*(\boldsymbol{x}-\boldsymbol{y}/2)$; bounding the modulus of the integral by the integral of the modulus and applying Cauchy–Schwarz (Proposition 12.4),

$$|W| \le \frac{1}{(2\pi\hbar)^3}\left(\int\left|\psi(\boldsymbol{x}+\tfrac{\boldsymbol{y}}{2})\right|^2\mathrm{d}^3y\right)^{1/2}\left(\int\left|\psi(\boldsymbol{x}-\tfrac{\boldsymbol{y}}{2})\right|^2\mathrm{d}^3y\right)^{1/2} = \frac{8}{(2\pi\hbar)^3}\,,$$

each integral contributing $8 = 2^3$ from the substitution $\boldsymbol{u} = \boldsymbol{x}\pm\boldsymbol{y}/2$ and the normalization of ψ. That is Equation (25.57). □

Corollary 25.54 (Negativity is unavoidable)**.** *Let ψ be any normalized wavefunction that is odd, $\psi(-\boldsymbol{x}) = -\psi(\boldsymbol{x})$. Then its Wigner function attains at the phase-space origin the most negative value the bound Equation (25.57) permits:*

$$W(\mathbf{0},\mathbf{0}) = -\frac{1}{(\pi\hbar)^3}\,. \tag{25.58}$$

No function on phase space that takes negative values is a probability density, so W is a quasi-*probability, and the failure is not a defect of the construction: Equations (25.54) and (25.55) show that its marginals are genuine, measured distributions.*

Rests on Equations (25.53) and (25.57).

Derivation. *Derives Corollary 25.54.*

Put $\boldsymbol{x} = \mathbf{0}$ and $\boldsymbol{p} = \mathbf{0}$ in Equation (25.53) with $\rho = \psi\psi^*$. The exponential is 1 and the integrand is $\psi(\boldsymbol{y}/2)\psi^*(-\boldsymbol{y}/2) = -\,|\psi(\boldsymbol{y}/2)|^2$ by oddness, so

$$W(\mathbf{0},\mathbf{0}) = -\frac{1}{(2\pi\hbar)^3}\int |\psi(\boldsymbol{y}/2)|^2\,\mathrm{d}^3y = -\frac{8}{(2\pi\hbar)^3} = -\frac{1}{(\pi\hbar)^3}\,,$$

the factor 8 coming from $\mathrm{d}^3y = 8\,\mathrm{d}^3u$ under $\boldsymbol{u} = \boldsymbol{y}/2$. Odd normalized wavefunctions exist in abundance — the first excited state of the harmonic oscillator (Section 78.4) is one — so the value Equation (25.58) is attained by real states and not only permitted. □

Theorem 25.55 (Hudson: which pure states escape)**.** *A pure state has a Wigner function that is non-negative everywhere if and only if its wavefunction is Gaussian, that is of the form $\psi(\boldsymbol{x}) = \exp\left(-\boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x} + \boldsymbol{b}\cdot\boldsymbol{x} + c\right)$ with A a complex symmetric matrix of positive-definite real part. Negativity of W is therefore an exact witness of departure from the Gaussian family, for pure states.*

Rests on Equation (25.53) and Corollary 25.54.

Full derivation: Appendix A.50 (p. 3060).

Derives Theorem 25.55.

Remark 25.56 (What the appendix proof imports, and what the theorem does not say)*.* The easy direction of Theorem 25.55 is a Gaussian integral: the Wigner transform of a Gaussian is a Gaussian, hence positive. The converse is the substantial half, and the appendix proof reduces it to one imported theorem — the Hadamard factorization of an entire function of order at most two, applied to the analytic continuation of ψ. Chapter 8 carries Liouville's theorem and the Cauchy estimates but no theory of the order and genus of an entire function, so that input is quoted there with its place named; the debt is Part II's and is recorded as such.

Two limits of the statement matter for how it is used. It is a theorem about *pure* states: there are mixed states whose Wigner function is non-negative everywhere and which are not mixtures of Gaussians, so "non-negative Wigner function" and "classical" are not synonyms in general, and Corollary 25.54 is a witness only within the pure family. And the 1974 paper in which Hudson proved it has no entry in this bibliography: the attribution above is reported without a source the reader of this book can check, and the appendix supplies the proof precisely so that nothing rests on the attribution.

Theorem 25.57 (The Moyal equation, and its classical limit)**.** *The Wigner function of a state evolving under $\hat{\mathcal{H}} = \hat{p}^2/2m + V(\hat{x})$, in one Cartesian pair, obeys*

$$\frac{\partial W}{\partial t} = -\frac{p}{m}\frac{\partial W}{\partial x} + \frac{\partial V}{\partial x}\frac{\partial W}{\partial p} - \frac{\hbar^2}{24}\frac{\partial^3 V}{\partial x^3}\frac{\partial^3 W}{\partial p^3} + \mathcal{O}(\hbar^4)\,, \tag{25.59}$$

whose first two terms are the classical Liouville equation Equation (24.15). If V is a polynomial of degree at most two, every correction vanishes identically and the Wigner function obeys the classical equation exactly*, for every state.*

Rests on Equations (25.51) and (25.53) and Theorem 24.21.

Derivation. *Derives Theorem 25.57.*

The von Neumann equation $\mathrm{i}\hbar\,\partial_t\hat\rho = \left[\hat{\mathcal{H}},\hat\rho\right]$ becomes, on Weyl symbols, $\partial_t W = \{\mathcal{H},W\}_{\mathrm{M}}$ by Definition 25.50. Evaluate the two pieces of $\mathcal{H}$ separately in Equation (25.51).

For the kinetic term $u = p^2/2m$, every x derivative of u vanishes, so only $k = n$ survives in $u \star W$ and only $k = 0$ in $W \star u$:

$$u\star W - W\star u = \sum_n \frac{1}{n!}\left(\frac{\mathrm{i}\hbar}{2}\right)^n \left[(-1)^n - 1\right]\left(\partial_p^n u\right)\left(\partial_x^n W\right) .$$

Even n cancel; odd n beyond 1 vanish because $\partial_p^3 u = 0$; and $n = 1$ leaves $-\mathrm{i}\hbar\,(p/m)\,\partial_x W$, giving the first term of Equation (25.59) after dividing by $\mathrm{i}\hbar$. The kinetic term is exact — it contributes no $\hbar$ correction at all.

For the potential $u = V(x)$, every p derivative vanishes, so the surviving index is the other one and the sign is reversed:

$$\frac{1}{\mathrm{i}\hbar}\left(V\star W - W\star V\right) = \frac{2}{\mathrm{i}\hbar}\sum_{n\ \text{odd}}\frac{1}{n!}\left(\frac{\mathrm{i}\hbar}{2}\right)^n \left(\partial_x^n V\right)\left(\partial_p^n W\right) .$$

The term $n = 1$ is $(\partial_x V)\,(\partial_p W)$, the classical force term. The term $n = 3$ is $\frac{2}{\mathrm{i}\hbar}\cdot\frac{1}{6}\left(\frac{\mathrm{i}\hbar}{2}\right)^3 \partial_x^3 V\,\partial_p^3 W = -\frac{\hbar^2}{24}\partial_x^3 V\,\partial_p^3 W$, and the next is $O(\hbar^4)$. This is Equation (25.59). If V has degree at most two then $\partial_x^3 V = 0$ and every surviving term of the series beyond $n = 1$ vanishes with it. □

Remark 25.58 (The same statement, three times). That quadratic potentials are the exactly classical case has now appeared three times in this chapter, and it is one fact each time. Theorem 25.32 said the expectation values follow the classical equations when V is at most quadratic; Lemma 25.39 said the correspondence Equation (25.31) is exact on polynomials of degree at most two; Theorem 25.57 says the phase-space density obeys the classical Liouville equation for such a potential. Degree two is where the Poisson algebra and the commutator algebra are isomorphic, and every statement of classicality traces back to that. It is also why the harmonic oscillator, the workhorse of every quantum textbook, is constitutionally unable to display any of the phenomena of this section.

Measuring a Wigner function

The formalism above would be a rewriting and nothing more if W were not accessible. It is, and the route is the Radon transform of Section 17.6.3.

Proposition 25.59 (Homodyne tomography inverts a Radon transform)**.** *Let a mode of frequency ω be described in the dimensionless quadratures*

$$X = x\sqrt{\frac{m\omega}{\hbar}}, \qquad P = \frac{p}{\sqrt{m\omega\hbar}}, \qquad \mathcal{H} = \frac{\hbar\omega}{2}\left(X^2 + P^2\right), \tag{25.60}$$

and let pr_θ be the measured distribution of the rotated quadrature $X_\theta = X\cos\theta + P\sin\theta$. Then

$$\mathrm{pr}_\theta(X_\theta) = (\mathcal{R}W)\left(\boldsymbol{n}(\theta), X_\theta\right), \qquad \boldsymbol{n}(\theta) = (\cos\theta, \sin\theta), \tag{25.61}$$

the Radon transform Equation (17.119) of W over the plane spanned by the two quadratures. The Wigner function is recovered from the family $\{\mathrm{pr}_\theta\}$, $0 \le \theta < \pi$, by filtered back-projection Equation (17.122).

Rests on Theorems 17.106 and 25.57 and Definition 17.103.

Derivation. *Derives Proposition 25.59.*

By Equation (25.54), the distribution of the quadrature X is the integral of W over the conjugate variable P — that is, the integral of W along the line $X = \text{const}$ in the (X, P) plane.

It remains to show that the same holds for every rotated quadrature. The Hamiltonian Equation (25.60) is quadratic, so by Theorem 25.57 the Wigner function evolves under it *exactly* by the classical Liouville flow; and the classical flow of $\mathcal{H} = \frac{1}{2}\hbar\omega(X^2 + P^2)$ in these coordinates is the rigid rotation $(X, P) \mapsto (X\cos\omega t + P\sin\omega t,\ -X\sin\omega t + P\cos\omega t)$, as substitution into Equations (22.3) and (22.4) confirms. So the Wigner function of the state at time $t = \theta/\omega$ is W rotated by θ, its X marginal is the integral of W along the line $X\cos\theta + P\sin\theta = \text{const}$, and that line integral is Equation (17.119) in two dimensions — which is Equation (25.61). In an optical realization the phase θ is set not by waiting but by the phase of a local oscillator, which is the same rotation performed at the detector rather than in the state.

Inverting is then Theorem 17.106 verbatim, with the projections P_θ of Equation (17.122) identified with pr_θ, and Corollary 17.107 says how many phases must be sampled to resolve structure of a given fineness. Nothing in the inversion is special to quantum mechanics: it is the arithmetic of a medical tomograph, applied to a quasi-probability instead of an attenuation coefficient. □

Remark 25.60 (What has been measured). Wigner functions with negative regions have been reconstructed by exactly this procedure. A single-photon state of a light mode is the first excited state of the oscillator Equation (25.60) and so is odd, which by Corollary 25.54 predicts $W(0,0) = -1/\pi\hbar$ in one degree of freedom: the deepest negative value the theory permits. Lvovsky and co-workers reconstructed it by balanced homodyne tomography and found the dip at the origin [2017]; the measured depth is smaller in magnitude than the ideal value, because detection efficiency below unity mixes vacuum into the state and a mixture of a Fock state with vacuum has a shallower dip. Deléglise and co-workers reconstructed the Wigner functions of photon number states and of superpositions of coherent states in a microwave cavity, and watched the negative regions disappear as the field decohered [2121] — which is Remark 25.33 seen directly, the environment turning a quasi-probability into a probability. The apparatus, the calibration and the error budget of these measurements belong with Chapter 122; what belongs here is that the object defined by Equation (25.53) is not a formal device.

Remark 25.61 (What phase-space quantum mechanics settles). The formulation answers, in the sharpest available way, the question this chapter began with. Quantum mechanics *can* be written on the very same phase space as classical mechanics, with the very same observables, and the entire difference is that the pointwise product is replaced by Equation (25.51). The classical limit is then not a change of arena but the vanishing of the $\hbar$-corrections in one product. What is lost in the passage is exactly what Remark 25.5 predicted at the start: the commutative product does not survive, and everything non-classical — the negativity of Corollary 25.54, the $\hbar^2$ term of Equation (25.59), the obstruction of Theorem 25.38 — is a consequence of that one replacement.

25.6 What carries over to field theory

The bracket. For a field theory the phase-space coordinates are a field $\varphi(\boldsymbol{x})$ and its conjugate momentum density $\pi(\boldsymbol{x})$ at each point of space, and the bracket acquires a functional derivative and a delta function, $\{\varphi(\boldsymbol{x}), \pi(\boldsymbol{y})\} = \delta^3(\boldsymbol{x} - \boldsymbol{y})$. The construction and its equal-time commutator counterpart are the subject of Chapters 64 and 99.

The first-order action. Equation (25.20) becomes $\int \mathrm{d}t \int \mathrm{d}^3x\,(\pi\dot{\varphi} - \mathcal{H})$, and it is this form, not the Lagrangian one, that the phase-space path integral of Chapter 106 discretizes. The Faddeev–Jackiw reading Equation (25.22) carries over with ξ^c replaced by the field components at each point of space, and the matrix inversion of Equation (25.25) by the inversion of an operator; where that operator is degenerate the theory is a gauge theory, which is the field-theoretic face of Remark 25.23.

The phase-space formulation. Equation (25.53) generalizes to a functional Wigner distribution, and the truncation of Equation (25.59) at its classical term is what a Boltzmann-equation treatment of a quantum field assumes; the $\hbar^2$ term is the leading quantum correction to it.

The obstruction, again. Theorem 25.38 has a field counterpart, and the ordering ambiguity there is one of the sources of the divergences that renormalization organizes (Chapter 100).

The failure of Stone–von Neumann. Inequivalent representations are the normal situation, per Remark 25.37.

Constraints. Gauge theories have degenerate Lagrangians, their naive brackets do not exist, and the whole apparatus above must be rebuilt on the constraint surface. That is the next chapter.

Chapter 26

Constrained Hamiltonian Systems: the Dirac–Bergmann Formalism

Every construction of Chapter 22 rests on one step: that the definition $p_a = \partial\mathcal{L}/\partial\dot{q}^a$ can be solved for the velocities. When it cannot, the Legendre transformation of Definition 22.2 is not invertible, the momenta are not independent functions on phase space, and the Hamiltonian formalism as stated does not exist. Proposition 22.3 named the condition and Remark 22.4 named the failure; this chapter is what happens next.

This is not a rare degeneracy to be noted and set aside. It is the generic situation for every theory with a gauge symmetry, and therefore for electromagnetism, for the whole Standard Model, and for general relativity. The reason is structural and is worth stating before any formalism: a gauge symmetry means that distinct sets of variables describe the same physical state, so the equations of motion cannot determine the evolution of all the variables—some of them must remain arbitrary—and a Legendre transformation that produced a well-posed evolution for all of them would be a contradiction. Degeneracy of the Lagrangian is the mechanical signature of redundancy in the description.

Dirac's constraint formalism, developed with Bergmann in the years around 1950, is the systematic treatment. It sorts the constraints into those that generate gauge transformations and those that merely restrict phase space, replaces the Poisson bracket by a modified bracket on the physical surface, counts the true degrees of freedom, and hands the result to the quantization rule of Section 25.4. It is stated here, in the classical part, because it is classical mechanics; Parts VII and XI use it rather than redevelop it.

Notation 26.1 (Weak and strong equality). Indices $a, b = 1, \dots, n$ label the configuration variables and $c, d = 1, \dots, 2n$ the phase-space variables, as in Notation 22.1. Constraints are written ϕ_m for primary constraints, χ_α for second-class constraints, γ_A for first-class constraints, and φ_j for the full set. The symbol $\approx$ denotes *weak equality*: $F \approx G$ means that F and G agree on the surface defined by the constraints, but not necessarily off it. The distinction is essential and is not pedantry — a weak equality may not be used inside a Poisson bracket, because the bracket differentiates in directions leaving the surface. Ordinary equality $=$ is called *strong* when the contrast is being drawn.

Rests on Notation 22.1 and Definition 22.29.

Remark 26.2 (The primary literature of this chapter is not yet in the bibliography). The formalism below is due to Dirac and to Bergmann and his collaborators, in papers of 1949–1951 and in Dirac's 1964 lecture notes; the modern standard reference is the monograph of Henneaux and Teitelboim. None of these works currently has an entry in this treatise's bibliography, so the attributions made in the prose of this chapter are, exceptionally, *uncited*. They are recorded in the source file as a citation debt. Every statement below is proved here or carries an explicit pending note, so nothing in the chapter rests on an unchecked appeal to those sources; what is missing is the historical credit, not the evidence.

26.1 Singular Lagrangians and primary constraints

Definition 26.3 (Singular Lagrangian)**.** A Lagrangian $\mathcal{L}(q,\dot q,t)$ is *singular*, or *degenerate*, if its Hessian with respect to the velocities

$$W_{ab}=\frac{\partial^2\mathcal{L}}{\partial\dot q^a\,\partial\dot q^b} \tag{26.1}$$

is not invertible. If $\operatorname{rank}W=R<n$, the Lagrangian is singular with $M=n-R$ degrees of degeneracy.

Rests on Equation (22.2) *and Proposition 22.3.*

Everything below assumes that $\operatorname{rank}W$ is *constant* on the region of TQ under study. This is an assumption about the system, not a triviality: the rank can drop on a lower-dimensional set, and where it does the constraint structure changes and the analysis must be redone patch by patch. All the systems worked in Section 26.6 have constant rank on the region of interest.

Remark 26.4 (The constant-rank theorem is owed by Part II). The proof of Proposition 26.5 needs the following statement, which belongs to multivariable analysis and to the manifolds chapter (Chapters 7 and 13) and is at present in neither. *Constant-rank theorem.* Let $f:U\subseteq\mathbb{R}^N\longrightarrow\mathbb{R}^{N'}$ be C^k, $k\geq1$, with $\operatorname{rank}Df=r$ at every point of U. Then about each point of U and its image there are C^k coordinate changes in which f reads $(u^1,\dots,u^N)\longmapsto(u^1,\dots,u^r,0,\dots,0)$. Consequently $f(U)$ is locally an r-dimensional embedded submanifold of $\mathbb{R}^{N'}$ (Definition 13.54), cut out near each of its points by $N'-r$ functions with linearly independent differentials. Theorem 13.59 is the codimension-one, maximal-rank special case of the last sentence, and Theorem A.286 is the analytic input from which the theorem is proved. The statement is used here and nowhere restated.

Proposition 26.5 (Primary constraints)**.** *If $\mathcal{L}$ is singular with $\operatorname{rank}W=R$ constant, the Legendre map $(q,\dot q)\mapsto(q,p)$ has image of dimension $n+R$, so the momenta satisfy $M=n-R$ independent relations*

$$\phi_m(q,p)\approx0\,,\qquad m=1,\dots,M\,, \tag{26.2}$$

identically in the velocities. These are the primary constraints*; they follow from the definition of the momenta alone, before any equation of motion is used.*

Rests on Definition 26.3 and Remark 26.4.

Proof. *Derives Proposition 26.5.*

Write the Legendre map as $\mathcal{F}:(q^a,\dot q^a)\longmapsto(q^a,\,\partial\mathcal{L}/\partial\dot q^a)$, a map between spaces of dimension $2n$. In the block ordering $(q,\dot q)$ its Jacobian is

$$D\mathcal{F}=\begin{pmatrix}\mathbb{1}_n & 0\\ \dfrac{\partial^2\mathcal{L}}{\partial q^b\,\partial\dot q^a} & W_{ab}\end{pmatrix}. \tag{26.3}$$

The first block column is already independent, and adding it to the second changes no rank, so $\operatorname{rank}D\mathcal{F}=n+\operatorname{rank}W=n+R$, constant by hypothesis. By the constant-rank theorem of Remark 26.4 the image is locally an embedded submanifold of the $2n$-dimensional space of the (q,p), of dimension $n+R$, and is cut out near each of its points by $2n-(n+R)=n-R=M$ functions with independent differentials. Those functions are the ϕ_m, and since the image is by construction the set of values the momenta actually take, $\phi_m\,(q,\partial\mathcal{L}/\partial\dot q)$ vanishes for *every* $(q,\dot q)$: the relations Equation (26.2) hold identically in the velocities, which is what distinguishes them from equations of motion. The q are untouched by the map, so no relation involves the coordinates alone. □

Proposition 26.6 (The canonical Hamiltonian is defined on the constraint surface)**.** *The function*

$$\mathcal{H}_c = p_a \dot{q}^a - \mathcal{L} \tag{26.4}$$

depends on the velocities only through the combinations that the momenta determine: it is a well-defined function of (q, p) on the primary constraint surface, and is ambiguous off it by terms proportional to the primary constraints.

Rests on Proposition 26.5 and Equation (26.4).

Proof. *Derives Proposition 26.6.*

Vary the right-hand side of Equation (26.4) with respect to the velocities at fixed q and fixed p:

$$\delta\mathcal{H}_{\mathrm{c}} = p_a\, \delta\dot{q}^a - \frac{\partial \mathcal{L}}{\partial \dot{q}^a}\, \delta\dot{q}^a = \left(p_a - \frac{\partial \mathcal{L}}{\partial \dot{q}^a} \right) \delta\dot{q}^a = 0\,, \tag{26.5}$$

the last step because $p_a = \partial\mathcal{L}/\partial\dot{q}^a$ is exactly the condition defining the point at which the variation is taken. So $\mathcal{H}_{\mathrm{c}}$ is constant along each fibre of the Legendre map, that is, along each set of velocities carried to the same momentum, and therefore descends to a function on the image — the primary constraint surface of Proposition 26.5. Off that surface it is not defined at all; any extension will do. Two extensions differ by a function vanishing on the surface, hence, by the regularity assumption of Remark 26.7, by a combination $c^m(q,p)\phi_m$. That ambiguity is not a defect to be removed but the origin of the undetermined multipliers in Definition 26.8. □

Remark 26.7 (The regularity assumption, stated once)*.* Throughout this chapter the constraint functions are assumed *regular*: at every point of the surface Σ they cut out, their differentials are linearly independent, and every smooth function vanishing on Σ is a combination $c^j(q,p)\varphi_j$ with smooth coefficients. The first half is what makes Σ a submanifold; the second is what lets "$F \approx 0$" be traded for "$F = c^j\varphi_j$", which is used in almost every proof below. A constraint set that fails regularity — $\phi = q^2$ rather than $\phi = q$ is the standard toy — must be replaced by an equivalent regular one before the formalism applies. This is an assumption about the system, and it is not free: it is exactly what fails at the points where $\operatorname{rank} W$ drops.

Definition 26.8 (Total Hamiltonian)**.** The *total Hamiltonian* is

$$\mathcal{H}_{\mathrm{T}} = \mathcal{H}_{\mathrm{c}} + u^m \phi_m\,, \tag{26.6}$$

with u^m undetermined multipliers. The evolution of any phase-space function is

$$\dot{F} \approx \{F, \mathcal{H}_{\mathrm{T}}\} = \{F, \mathcal{H}_{\mathrm{c}}\} + u^m \{F, \phi_m\}\,. \tag{26.7}$$

Rests on Proposition 26.6 and Definition 22.29.

26.2 The Dirac–Bergmann consistency algorithm

The constraints must be preserved by the motion: a system that starts on the constraint surface must stay on it.

Definition 26.9 (Consistency conditions)**.** The *consistency conditions* are

$$\dot{\phi}_m = \{\phi_m, \mathcal{H}_{\mathrm{c}}\} + u^n \{\phi_m, \phi_n\} \approx 0\,, \qquad m = 1, \dots, M\,. \tag{26.8}$$

Rests on Equations (26.2) and (26.7).

Proposition 26.10 (The four outcomes)**.** *Each condition Equation (26.8) does exactly one of four things:*

1. *it is satisfied identically on the constraint surface, and says nothing;*

2. *it reduces to a relation among the q and p alone, independent of the multipliers—a new,* secondary *constraint;*

3. *it determines some of the multipliers u^m; or*

4. *it is inconsistent, in which case the Lagrangian describes no motion at all.*

The procedure is then repeated on any new constraint, generating tertiary constraints and so on, until no new condition arises.

Rests on Definition 26.9 and Remark 26.7.

Proof. *Derives Proposition 26.10.*

Abbreviate $h_m := \{\phi_m, \mathcal{H}_c\}$ and $C_{mn} := \{\phi_m, \phi_n\}$, both evaluated on the primary constraint surface. Equation (26.8) is then the linear system

$$C_{mn}\, u^n \approx -h_m \,, \tag{26.9}$$

in the M unknowns u^n, with an antisymmetric coefficient matrix. Let $r = \operatorname{rank} C$ on the surface and let $\left\{V_m^{(k)}\right\}$, $k = 1, \dots, M - r$, be a basis of the null vectors of C, so that $V_m^{(k)} C_{mn} \approx 0$.

Decompose Equation (26.9) accordingly. Along the row space of C — r independent combinations — the system is solvable for r combinations of the multipliers, and this is outcome 3; the remaining $M - r$ combinations of the u^n are left arbitrary, which is the point taken up in Theorem 26.14. Contracting Equation (26.9) with a null vector annihilates every multiplier and leaves

$$V_m^{(k)}\, h_m = \left\{ V_m^{(k)} \phi_m, \mathcal{H}_c \right\} \approx 0 \,, \tag{26.10}$$

a condition on the phase-space point alone. There are three possibilities for it, and no others. If the left-hand side vanishes on the surface already, the condition is empty: outcome 1. If it is a function that does not vanish there, it is a new restriction on (q, p) — a secondary constraint: outcome 2. If it reduces to a nonzero constant, no phase-space point satisfies it and the theory has no solutions: outcome 4. The exhaustion is complete because Equation (26.9) is a linear system and every linear system either determines an unknown, or produces a compatibility condition, or is empty.

Repeating the argument with the enlarged constraint set gives the recursion. Note that Equation (26.10) identifies the generator of any new constraint as the bracket of a first-class *primary* combination $V_m^{(k)} \phi_m$ with the canonical Hamiltonian, which is how the algorithm is run in practice. □

Definition 26.11 (The final constraint set)**.** The algorithm terminates on a finite-dimensional phase space, since each round either adds an independent constraint—and there are at most $2n$ of them—or ends. Its output is the full set

$$\varphi_j \approx 0 \,, \qquad j = 1, \dots, J \,, \tag{26.11}$$

comprising the primary constraints and all those generated from them, and the surface $\Sigma \subset M$ they define is the *constraint surface*. The distinction primary/secondary refers to how a constraint was found and has no further significance; the classification that matters is the next one.

Rests on Proposition 26.10 and Remark 26.7.

26.3 First- and second-class constraints

Definition 26.12 (First and second class)**.** A phase-space function F is *first class* if

$$\{F, \varphi_j\} \approx 0 \qquad \text{for every } j\,, \tag{26.12}$$

and *second class* otherwise. The classification applies in particular to the constraints themselves. A *complete separation* of the constraint set is a choice of basis $\{\gamma_A, \chi_\alpha\}$ for it in which the γ_A are a maximal set of independent first-class combinations and the χ_α are the rest, so that no nonzero combination of the χ_α alone is first class.

Rests on Definitions 22.29 and 26.11.

Proposition 26.13 (The first-class functions close)**.** *The Poisson bracket of two first-class functions is first class. The first-class constraints therefore form a Lie algebra under the bracket, weakly; its structure functions need not be constants.*

Rests on Definition 26.12, Equation (22.46), and Remark 26.7.

Proof. *Derives Proposition 26.13.*

Let F and G be first class. By Equation (26.12) and the regularity assumption of Remark 26.7 there are smooth functions with

$$\{F, \varphi_j\} = f_j{}^k \varphi_k\,, \qquad \{G, \varphi_j\} = g_j{}^k \varphi_k\,. \tag{26.13}$$

The Jacobi identity Equation (22.46), rearranged, gives

$$\{\{F, G\}, \varphi_j\} = \{F, \{G, \varphi_j\}\} - \{G, \{F, \varphi_j\}\}\,. \tag{26.14}$$

Insert Equation (26.13) and expand each bracket by the Leibniz rule:

$$\{\{F, G\}, \varphi_j\} = \left\{F, g_j{}^k\right\} \varphi_k + g_j{}^k \{F, \varphi_k\} - \left\{G, f_j{}^k\right\} \varphi_k - f_j{}^k \{G, \varphi_k\}\,. \tag{26.15}$$

Every term on the right carries either an explicit φ_k or a bracket that vanishes weakly by hypothesis, so the whole expression is weakly zero. Hence $\{F, G\}$ is first class.

Applied to the first-class constraints themselves this says $\{\gamma_A, \gamma_B\} \approx 0$, so again by regularity

$$\{\gamma_A, \gamma_B\} = f_{AB}{}^C \gamma_C\,, \tag{26.16}$$

which is closure. The coefficients $f_{AB}{}^C$ are functions on phase space in general; whether they are constants is a property of the particular theory, and Sections 26.6.4 and 26.6.5 are the two cases that matter — constants for Yang–Mills, genuine functions for gravity. Where they are constants, Equation (26.16) is a Lie algebra in the sense of Theorem 25.3; where they are not, it is not, and the difference is not cosmetic. □

Theorem 26.14 (First-class constraints generate gauge transformations)**.** *Let $\gamma_k = V_m^{(k)} \phi_m$ be the first-class* primary *constraints, $V^{(k)}$ the null vectors of Proposition 26.10. Then two solutions of the equations of motion with the same initial data at time t differ at $t + \delta t$ by*

$$\delta_\varepsilon F = \varepsilon^k \{F, \gamma_k\}\,, \tag{26.17}$$

with ε^k arbitrary. The two therefore describe the same physical history, and the variables they differ in carry no physical information.

Rests on Proposition 26.10 and Definition 26.8.

Proof. *Derives Theorem 26.14.*

By Proposition 26.10 the consistency conditions fix the multipliers only up to the null space of C_{mn}: the general solution of Equation (26.9) is

$$u^m = U^m + v^k V_m^{(k)}\,, \tag{26.18}$$

with U^m any particular solution and the v^k arbitrary functions of time. Evolve the same initial point $z(t)$ through δt with two admissible choices u and u'. Both results solve the equations of motion, and by Equation (26.7) they differ by

$$\delta F = \delta t \left(u^m - u'^m\right) \{F, \phi_m\} = \delta t \left(v^k - v'^k\right) V_m^{(k)} \{F, \phi_m\} = \varepsilon^k \{F, \gamma_k\} \,, \tag{26.19}$$

writing $\varepsilon^k := \delta t \left(v^k - v'^k\right)$, which is arbitrary, and using that the $V^{(k)}$ are functions of (q, p) so that $V_m^{(k)} \{F, \phi_m\} = \left\{F, V_m^{(k)} \phi_m\right\}$ weakly. That $\gamma_k = V_m^{(k)} \phi_m$ is first class is Equation (26.10) together with $V_m^{(k)} C_{mn} \approx 0$: its bracket with every primary constraint vanishes weakly by the second, and with the canonical Hamiltonian by the first.

Two evolutions of one initial state cannot be two physical histories, on pain of the theory predicting nothing. So the difference Equation (26.19) is unphysical, and any two phase-space points related by Equation (26.17) describe the same state. □

Remark 26.15 (Dirac's conjecture, and why the theorem stops where it does). Theorem 26.14 is proved for the first-class *primary* constraints, because those are the ones whose multipliers the algorithm demonstrably leaves free. Dirac conjectured that *all* first-class constraints, secondary ones included, generate gauge transformations. The conjecture is true for every system in Section 26.6, and it is what justifies Definition 26.16; the general construction assembles the gauge generator as a chain $G = \varepsilon\gamma + \dot{\varepsilon}\gamma' + \cdots$ built from constraints of successive generations, of which Equation (26.61) below is the worked instance. It is nonetheless *false* in general: counterexamples exist in which a secondary first-class constraint generates no symmetry of the action, and they are not pathological beyond failing the regularity of Remark 26.7 at a point. This treatise does not yet carry a citable reference for either the general construction or the counterexample (Remark 26.2), so the theorem above claims only what is proved here.

Definition 26.16 (Extended Hamiltonian)**.** The *extended Hamiltonian* adds all first-class constraints, primary or not, with independent multipliers:

$$\mathcal{H}_{\mathrm{E}} = \mathcal{H}_{\mathrm{c}} + v^A \gamma_A \,. \tag{26.20}$$

It generates the same motion of gauge-invariant quantities as $\mathcal{H}_{\mathrm{T}}$, and a larger motion of gauge-variant ones.

Rests on Definition 26.8 and Theorem 26.14.

Theorem 26.17 (Counting the physical degrees of freedom)**.** *Let a system have $2n$ phase-space dimensions, F independent first-class constraints and S independent second-class constraints. Then*

$$\dim(\textit{physical phase space}) = 2n - 2F - S \,, \tag{26.21}$$

and the number of physical configuration-space degrees of freedom is half of that, $n - F - \frac{1}{2}S$. A first-class constraint removes two dimensions—itself, and the gauge direction it generates—while a second-class constraint removes one.

Rests on Theorem 26.14, Proposition 26.19, and Remark 26.18.

Proof. *Derives Theorem 26.17.*

By regularity (Remark 26.7) the $F + S$ constraints have independent differentials, so the constraint surface Σ of Equation (26.11) is an embedded submanifold (Definition 13.54) of dimension

$$\dim \Sigma = 2n - F - S \,. \tag{26.22}$$

Consider on Σ the Hamiltonian vector fields $X_A := \{\cdot, \gamma_A\}$ of the first-class constraints. Each is tangent to Σ, because $X_A \varphi_j = \{\varphi_j, \gamma_A\} = -\{\gamma_A, \varphi_j\} \approx 0$ by Equation (26.12), so the flow it generates does not leave the surface; and the F of them are pointwise

independent, because the $\mathrm{d}\gamma_A$ are independent and the symplectic form is non-degenerate. They therefore span an F-dimensional distribution on Σ. It is involutive: by Proposition 25.4 the bracket of Hamiltonian vector fields is the Hamiltonian vector field of the Poisson bracket, $[X_A, X_B] = X_{\{\gamma_A,\gamma_B\}}$, and $\{\gamma_A, \gamma_B\} = f_{AB}{}^C\gamma_C$ by Equation (26.16), whose Hamiltonian vector field on Σ is $f_{AB}{}^C X_C$ — again in the span.

By the Frobenius theorem of Remark 26.18 the distribution is integrable, and Σ is foliated by F-dimensional leaves: the *gauge orbits*. By Theorem 26.14 all points of one orbit describe one physical state, so the physical phase space is the space of leaves, of dimension

$$\dim\Sigma - F = 2n - F - S - F = 2n - 2F - S\,, \tag{26.23}$$

which is Equation (26.21). It is even, because S is even by Proposition 26.19, so half of it is an integer; and it is $2\left(n - F - \frac{1}{2}S\right)$, the doubling that a configuration space and its momenta always produce. Finally the reduced space carries a symplectic form: the pullback of ω to Σ has the X_A in its kernel and nothing else, so it descends to a non-degenerate form on the quotient — which is Theorem 24.47 in the special case where the momentum map is the constraint set. □

Remark 26.18 (The Frobenius theorem is owed by Part II). The proof of Theorem 26.17 needs the following, which belongs to the manifolds chapter (Chapter 13) and is not there. *Frobenius theorem.* A smooth distribution $D \subset TN$ of constant rank k on a manifold N is integrable — through each point there passes a k-dimensional immersed submanifold whose tangent space is D, and N is foliated by such leaves — if and only if D is involutive, that is, closed under the Lie bracket of vector fields. The theorem is invoked by name elsewhere in this part as well: the pending note of Theorem 24.38 attributes it to Part II, and Chapter 16 appeals to it to separate holonomic from nonholonomic constraints. It is stated here because it is used here, not because this is its home.

Proposition 26.19 (Second-class constraints come in pairs)**.** *For a complete separation in the sense of Definition 26.12, the number S of independent second-class constraints is even, and the matrix*

$$C_{\alpha\beta} = \{\chi_\alpha, \chi_\beta\} \tag{26.24}$$

is invertible on the constraint surface.

Rests on Definition 26.12 and Proposition 5.115.

Proof. *Derives Proposition 26.19.*

$C_{\alpha\beta}$ is antisymmetric, by antisymmetry of the Poisson bracket (Equation (22.43)). Suppose it were degenerate at a point of Σ: there would be numbers λ^α, not all zero, with $C_{\alpha\beta}\lambda^\beta \approx 0$. Put $\chi := \lambda^\alpha\chi_\alpha$. Then $\{\chi_\beta, \chi\} \approx 0$ for every β by construction, and $\{\gamma_A, \chi\} \approx 0$ for every A because the γ_A are first class. Since $\{\gamma_A, \chi_\alpha\}$ is a basis of the constraint set, χ has weakly vanishing bracket with every constraint: it is first class. But a complete separation admits no nonzero first-class combination of the χ_α alone, so $\lambda = 0$ — a contradiction. Hence C is non-degenerate on Σ, which is the second assertion.

For the first, C is then the matrix of a non-degenerate antisymmetric bilinear form on the S-dimensional real space spanned by the χ_α at that point. By Proposition 5.115 such a form exists only in even dimension, so $S = 2s$. Equivalently, and this is the one-line version, $\det C = \det\left(-C^{\mathsf{T}}\right) = (-1)^S \det C$, so $\det C = 0$ whenever S is odd. □

26.4 The Dirac bracket

Second-class constraints cannot be imposed strongly inside a Poisson bracket—setting $\chi_\alpha = 0$ in Equation (26.24) would make an invertible matrix vanish. The resolution is to change the bracket.

Definition 26.20 (Dirac bracket)**.** Let χ_α be a complete set of second-class constraints and $(C^{-1})^{\alpha\beta}$ the inverse of the matrix Equation (26.24). The *Dirac bracket* of two phase-space functions is

$$\{A,B\}_{\mathrm{D}} = \{A,B\} - \{A,\chi_\alpha\}\left(C^{-1}\right)^{\alpha\beta}\{\chi_\beta,B\}\,. \tag{26.25}$$

Rests on Proposition 26.19 and Definition 22.29.

Theorem 26.21 (Properties of the Dirac bracket)**.** *The Dirac bracket is bilinear and antisymmetric, obeys the Leibniz rule Equation (24.24) and the Jacobi identity Equation (22.46), and in addition*

$$\{A,\chi_\alpha\}_D = 0 \quad \textit{strongly, for every } A\,, \tag{26.26}$$

$$\{A,B\}_D \approx \{A,B\} \quad \textit{whenever } A \textit{ is first class}\,, \tag{26.27}$$

$$\dot F \approx \{F,\mathcal{H}_T\}_D\,. \tag{26.28}$$

Because of Equation (26.26), the second-class constraints may be set to zero strongly *once the Dirac bracket is adopted: they become identities, and the phase space is effectively reduced.*

Rests on Definitions 26.8 and 26.20 and Proposition 26.19.

Proof. *Derives Theorem 26.21.*

Bilinearity is immediate: both terms of Equation (26.25) are bilinear in (A,B). For antisymmetry, note first that the inverse of an invertible antisymmetric matrix is antisymmetric, since $C^{\mathsf T} = -C$ gives $(C^{-1})^{\mathsf T} = \left(C^{\mathsf T}\right)^{-1} = -C^{-1}$. Exchanging A and B in the correction term and using $\{B,\chi_\alpha\} = -\{\chi_\alpha,B\}$, $\{\chi_\beta,A\} = -\{A,\chi_\beta\}$ and then relabelling $\alpha \leftrightarrow \beta$,

$$-\{B,\chi_\alpha\}\left(C^{-1}\right)^{\alpha\beta}\{\chi_\beta,A\} = -\{A,\chi_\alpha\}\left(C^{-1}\right)^{\beta\alpha}\{\chi_\beta,B\} = +\{A,\chi_\alpha\}\left(C^{-1}\right)^{\alpha\beta}\{\chi_\beta,B\}\,, \tag{26.29}$$

so the correction changes sign with the leading term and $\{A,B\}_{\mathrm{D}} = -\{B,A\}_{\mathrm{D}}$. The Leibniz rule in the first slot follows because $A \longmapsto \{A,B\}$ and $A \longmapsto \{A,\chi_\alpha\}$ are both derivations (Proposition 25.4) and the second factor of the correction does not involve A; antisymmetry then gives it in the second slot.

For Equation (26.26), take $B = \chi_\gamma$:

$$\{A,\chi_\gamma\}_{\mathrm{D}} = \{A,\chi_\gamma\} - \{A,\chi_\alpha\}\left(C^{-1}\right)^{\alpha\beta}C_{\beta\gamma} = \{A,\chi_\gamma\} - \{A,\chi_\alpha\}\,\delta^\alpha_\gamma = 0\,, \tag{26.30}$$

which holds wherever C^{-1} is defined — strongly, not merely on Σ. Equation (26.27) is immediate: if A is first class then $\{A,\chi_\alpha\} \approx 0$ and the whole correction vanishes weakly. For Equation (26.28), the consistency algorithm has fixed the multipliers of the second-class primary constraints, and with that fixing $\mathcal{H}_{\mathrm{T}}$ is first class — it is precisely the statement that the motion preserves every constraint — so $\{\chi_\beta,\mathcal{H}_{\mathrm{T}}\} \approx 0$ and Equation (26.27) applies with $B = \mathcal{H}_{\mathrm{T}}$, giving $\{F,\mathcal{H}_{\mathrm{T}}\}_{\mathrm{D}} \approx \{F,\mathcal{H}_{\mathrm{T}}\} \approx \dot F$.

The Jacobi identity is the one property that is not a short computation; it is what earns the object the name bracket, and it is deferred. □

Full derivation: Appendix A.51 (p. 3069).

Derives Theorem 26.21.

Two routes are written out there. The direct one expands $\{A,\{B,C\}_{\mathrm{D}}\}_{\mathrm{D}}$ and its two cyclic partners: besides the Poisson Jacobi identity itself, this produces terms carrying derivatives of the inverse matrix, which are reduced by differentiating $C^{-1}C = \mathbb{1}$ to express $\left\{A,(C^{-1})^{\alpha\beta}\right\}$ through $\{A,C_{\gamma\delta}\}$, after which everything cancels in triples by the Poisson Jacobi identity applied to the constraints themselves. The structural one observes that the Dirac bracket is the Poisson bracket of the symplectic form induced on the second-class surface — a symplectic form precisely because C is invertible, which is Proposition 26.25 below — and reads the identity off the closure of that form.

Example 26.22 (Holonomic constraints)**.** A particle of mass m moving in $\mathbb{R}^3$ under a potential $V(\boldsymbol{x})$, with $\mathcal{H} = \boldsymbol{p}^2/2m + V$, is confined to the surface $g(\boldsymbol{x}) = 0$ of Definition 21.2. The Dirac bracket built from the resulting second-class pair reproduces exactly the reduced dynamics obtained in Chapter 21 by eliminating the constrained coordinate. The elementary treatment of holonomic constraints there and the formalism here agree, which is the check that fixes all signs.

Rests on Definitions 21.2 and 26.20.

Derivation. *Derives Example 26.22.*

Confinement to the surface is one condition on the coordinates, but a single $\chi_1 = g(\boldsymbol{x})$ is not preserved by the motion: the algorithm of Section 26.2 applied to it returns

$$\chi_2 := \{g, \mathcal{H}\} = \frac{1}{m}\, p_i\, \partial_i g\,, \tag{26.31}$$

the statement that the velocity is tangent to the surface. Their bracket is

$$\{\chi_1, \chi_2\} = \left\{g, \tfrac{1}{m} p_j \partial_j g\right\} = \frac{1}{m}\, \partial_i g\, \partial_i g = \frac{|\nabla g|^2}{m} =: w\,, \tag{26.32}$$

nonzero wherever $\nabla g \neq 0$, which is where the surface is regular. The pair is second class, $S = 2$, $F = 0$, and Equation (26.21) gives $2 \times 3 - 2 = 4$: two configuration degrees of freedom, as Definition 21.9 requires for a point on a surface. With

$$C = \begin{pmatrix} 0 & w \\ -w & 0 \end{pmatrix}, \qquad C^{-1} = \begin{pmatrix} 0 & -w^{-1} \\ w^{-1} & 0 \end{pmatrix}, \tag{26.33}$$

and the elementary brackets $\{x_i, \chi_1\} = 0$, $\{x_i, \chi_2\} = \partial_i g/m$, $\{p_i, \chi_1\} = -\partial_i g$ and $\{p_i, \chi_2\} = -p_k \partial_i \partial_k g/m$, Equation (26.25) gives

$$\{x_i, x_j\}_{\mathrm{D}} = 0\,, \tag{26.34}$$

$$\{x_i, p_j\}_{\mathrm{D}} = \delta_{ij} - \frac{\partial_i g\, \partial_j g}{|\nabla g|^2} = \delta_{ij} - n_i n_j\,, \tag{26.35}$$

$$\{p_i, p_j\}_{\mathrm{D}} = \frac{p_k}{|\nabla g|^2}\left[(\partial_i \partial_k g)\, \partial_j g - (\partial_j \partial_k g)\, \partial_i g\right], \tag{26.36}$$

with $\boldsymbol{n} = \nabla g / |\nabla g|$ the unit normal. Read these three. Equation (26.35) is the projector onto the tangent plane: the momentum conjugate to a coordinate is only the tangential part of $\boldsymbol{p}$, which is exactly what the reduced Lagrangian treatment produces when the normal coordinate is eliminated. The normal component of $\boldsymbol{p}$ has vanishing Dirac bracket with everything, so it is not an independent variable at all — it is χ_2, set strongly to zero by Equation (26.26). And Equation (26.36) says that the momenta no longer commute: the second derivatives of g are the second fundamental form of the surface, so the failure is its extrinsic curvature.

For the sphere $g = |\boldsymbol{x}| - R$ one has $\nabla g = \hat{\boldsymbol{r}}$ and $\partial_i \partial_j g = (\delta_{ij} - \hat{r}_i \hat{r}_j)/r$, and Equation (26.36) collapses on the constraint surface, where $\boldsymbol{p} \cdot \hat{\boldsymbol{r}} \approx 0$, to

$$\{p_i, p_j\}_{\mathrm{D}} \approx \frac{p_i \hat{r}_j - p_j \hat{r}_i}{R} = -\frac{\epsilon_{ijk} L_k}{R^2}\,, \qquad \boldsymbol{L} = \boldsymbol{x} \times \boldsymbol{p}\,, \tag{26.37}$$

so the momenta on a sphere close on the angular momentum. That is the bracket algebra of the free particle on S^2, whose Hamiltonian is $\boldsymbol{L}^2/2mR^2$ — precisely the reduced system obtained in Chapter 21 by writing the Lagrangian in the two angles. The two routes agree, and Equation (26.37) is the classical origin of the operator ordering problem for a particle on a curved configuration space, taken up in Section 25.5.1. □

26.5 Gauge fixing and the reduced phase space

Definition 26.23 (Gauge conditions)**.** Given F first-class constraints γ_A, a set of *gauge conditions* is a set of F functions ψ^A such that

$$\det\left\{\gamma_A, \psi^B\right\} \neq 0 \tag{26.38}$$

on the constraint surface. The combined set $\left\{\gamma_A, \psi^A\right\}$ is then second class, and the Dirac bracket built from it defines the dynamics on the *reduced phase space* of dimension $2n-2F-S$.

Rests on Theorem 26.17 and Definition 26.20.

Condition Equation (26.38) says two things at once, and both are needed. It says the gauge is *attainable*: since $\left\{\psi^B, \gamma_A\right\}$ is the change of ψ^B under the gauge transformation generated by γ_A, a non-vanishing determinant means the gauge parameters can be chosen to reach $\psi = 0$ from anywhere nearby. And it says the gauge is *complete*: no gauge freedom survives, because a residual transformation would leave ψ unchanged and make the determinant vanish. The count then follows from Theorem 26.17 with $F \to 0$ and $S \to S + 2F$.

Proposition 26.24 (The Faddeev–Popov determinant)**.** *Write $\Gamma_a = \left(\gamma_A, \psi^B\right)$ for the combined second-class set of Definition 26.23. Then*

$$\det\{\Gamma_a, \Gamma_b\} \approx \left(\det\left\{\gamma_A, \psi^B\right\}\right)^2 , \tag{26.39}$$

so that the factor $\sqrt{\det\{\Gamma_a, \Gamma_b\}}$ by which the Liouville measure of Definition 24.20 is corrected on passing to the gauge-fixed surface is $\left|\det\left\{\gamma_A, \psi^B\right\}\right|$. This is the object that appears as the Faddeev–Popov determinant in the path-integral quantization of Chapter 106 [1347]. Its exponentiation as an integral over anticommuting fields is what introduces the ghosts of gauge-fixed perturbation theory.

Rests on Definition 26.23 and Proposition 26.19.

Proof. *Derives Proposition 26.24.*

Order the combined set as $\left(\gamma_A, \psi^B\right)$ and write $D_A{}^B := \left\{\gamma_A, \psi^B\right\}$. Because the γ_A are first class, $\{\gamma_A, \gamma_B\} \approx 0$, so on Σ

$$\{\Gamma_a, \Gamma_b\} \approx \begin{pmatrix} 0 & D \\ -D^{\mathsf{T}} & E \end{pmatrix} , \qquad E^{AB} = \left\{\psi^A, \psi^B\right\} , \tag{26.40}$$

with all four blocks $F \times F$ and D invertible by Equation (26.38). Multiply on the left by $\begin{pmatrix} \mathbb{1} & -DE^{-1} \\ 0 & \mathbb{1} \end{pmatrix}$, of determinant 1, where E is invertible; the product is $\begin{pmatrix} DE^{-1}D^{\mathsf{T}} & 0 \\ -D^{\mathsf{T}} & E \end{pmatrix}$, whose determinant is $\det\left(DE^{-1}D^{\mathsf{T}}\right)\det E = (\det D)^2$, in which E has cancelled. Both sides of Equation (26.39) are polynomials in the entries of an *arbitrary* $F \times F$ matrix E — the antisymmetry of E is nowhere used — and they have just been shown equal on the dense open set of invertible E, so they are equal for every E. In particular the identity holds in the common case $E = 0$, where the gauge conditions commute among themselves, and this matters: for odd F an antisymmetric E is never invertible, so the perturbation must be taken outside the antisymmetric matrices, which is what the previous sentence does. Taking the positive square root gives the stated factor. □

The second sentence of Proposition 26.24 — that this determinant is the factor by which the Liouville measure is corrected — is a separate statement, and it is the measure-theoretic face of the Dirac bracket.

Proposition 26.25 (The Liouville measure of a second-class surface)**.** *Let χ_α, $\alpha = 1, \dots, S$, be a second-class set on a phase space of dimension $2n$, so that $C_{\alpha\beta}$ of Equation (26.24) is invertible on the surface Σ_χ they define, and write Θ for the Liouville form Equation (24.13) of the ambient space — the letter Ω being reserved in this chapter for the BRST charge Equation (26.97). Then Σ_χ carries the non-degenerate restriction ω_Σ of the symplectic form, $\det C > 0$, and the Liouville form Θ_Σ built from ω_Σ satisfies*

$$\int_{\Sigma_\chi} F\,\Theta_\Sigma = \int F\,\sqrt{\det C}\,\prod_{\alpha=1}^{S}\delta(\chi_\alpha)\,\Theta \tag{26.41}$$

for every integrable F. The reduced measure is thus the ambient one with a delta function of each constraint and the factor $\sqrt{\det C}$, which for a gauge-fixed set Γ_a is the $\sqrt{\det\{\Gamma_a,\Gamma_b\}}$ of Proposition 26.24.

Rests on Propositions 24.3 and 26.19 and Definition 24.20.

Proof. *Derives Proposition 26.25.*

Write M for the phase space, so that $\dim M = 2n$. Everything is pointwise linear algebra in the tangent space T_xM at a point $x \in \Sigma_\chi$, plus one convention about the delta functions.

The splitting. With the conventions Equations (24.5) and (24.7), $\mathrm{d}\chi_\gamma(X_{\chi_\alpha}) = \{\chi_\gamma, \chi_\alpha\} = C_{\gamma\alpha}$, which is invertible; so the S vectors X_{χ_α} are independent and span a subspace W, and the S covectors $\mathrm{d}\chi_\alpha$ are independent, with common kernel $T_x\Sigma_\chi$ of dimension $2n - S$. The map

$$P(v) = v - X_{\chi_\alpha}\left(C^{-1}\right)^{\alpha\beta}\mathrm{d}\chi_\beta(v) \tag{26.42}$$

gives $\mathrm{d}\chi_\gamma(P(v)) = \mathrm{d}\chi_\gamma(v) - C_{\gamma\alpha}\left(C^{-1}\right)^{\alpha\beta}\mathrm{d}\chi_\beta(v) = 0$, so it projects T_xM onto $T_x\Sigma_\chi$ along W, and $T_xM = W \oplus T_x\Sigma_\chi$. The two summands are ω-orthogonal: for $v \in T_x\Sigma_\chi$, $\omega(X_{\chi_\alpha}, v) = \left(\iota_{X_{\chi_\alpha}}\omega\right)(v) = \mathrm{d}\chi_\alpha(v) = 0$. On W the form has the invertible matrix $\omega\left(X_{\chi_\alpha}, X_{\chi_\beta}\right) = C_{\alpha\beta}$; and a $v \in T_x\Sigma_\chi$ annihilated by ω_Σ is by that orthogonality annihilated by ω on all of T_xM, hence zero. Both restrictions are therefore non-degenerate, $(\Sigma_\chi, \omega_\Sigma)$ is symplectic, and $S = 2s$ is even — which is Proposition 26.19 read geometrically.

The volume factorises. Write $\omega = \omega_1 + \omega_2$, where ω_1 is the pullback of $\omega|_W$ along $v \longmapsto v - P(v)$ and ω_2 the pullback of ω_Σ along P; that there is no cross term is the ω-orthogonality just proved. Two-forms commute under $\wedge$, and $\omega_1^{\wedge k} = 0$ for $k > s$ while $\omega_2^{\wedge k} = 0$ for $k > n - s$, each being pulled back from a space of that dimension. Expanding $\omega^{\wedge n}$ by the binomial theorem therefore leaves a single term:

$$\Theta = \frac{\omega^{\wedge n}}{n!} = \frac{\omega_1^{\wedge s}}{s!} \wedge \frac{\omega_2^{\wedge (n-s)}}{(n-s)!} = \Theta_W \wedge P^*\Theta_\Sigma\,, \tag{26.43}$$

with Θ_W the Liouville form of the symplectic space W.

Two evaluations. Take the adapted basis $(X_{\chi_1}, \dots, X_{\chi_S}, v_1, \dots, v_{2n-S})$ with the v_r a basis of $T_x\Sigma_\chi$. By Proposition 24.3 pick a canonical basis $E_1, \dots, E_S$ of $(W, \omega|_W)$, ordered as in Equation (24.13) so that Θ_W takes the value 1 on it, and write $X_{\chi_\alpha} = L_\alpha{}^a E_a$. Then $C = L\,J\,L^{\mathsf{T}}$ with J the matrix Equation (5.144), whose determinant is 1, so $\det C = (\det L)^2$ — positive, as claimed — while $\Theta_W(X_{\chi_1}, \dots, X_{\chi_S}) = \det L$. Since Θ_W annihilates every v_r and $P^*\Theta_\Sigma$ every X_{χ_α}, only one term of the wedge in Equation (26.43) survives on that basis:

$$\Theta\left(X_{\chi_1}, \dots, v_{2n-S}\right) = \pm\sqrt{\det C}\;\Theta_\Sigma\left(v_1, \dots, v_{2n-S}\right)\,. \tag{26.44}$$

The other top form built from the constraints evaluates on the same basis to

$$\left(\mathrm{d}\chi_1 \wedge \cdots \wedge \mathrm{d}\chi_S \wedge P^*\Theta_\Sigma\right)\left(X_{\chi_1}, \dots, v_{2n-S}\right) = \det C\;\Theta_\Sigma\left(v_1, \dots, v_{2n-S}\right)\,, \tag{26.45}$$

because $\det\big(\mathrm{d}\chi_\alpha(X_{\chi_\beta})\big) = \det C$. Two top forms on a $2n$-dimensional space that agree up to a factor on one basis agree up to that factor everywhere, so, orienting Σ_χ so that the sign in Equation (26.44) is positive,

$$\Theta = \frac{1}{\sqrt{\det C}}\,\mathrm{d}\chi_1 \wedge \cdots \wedge \mathrm{d}\chi_S \wedge P^*\Theta_\Sigma\,. \tag{26.46}$$

The convention. A delta function of a constraint means one thing: if a top form is presented as $\mathrm{d}\chi_1 \wedge \cdots \wedge \mathrm{d}\chi_S \wedge \eta$ for a $(2n-S)$-form η, then integrating $\prod_\alpha \delta(\chi_\alpha)$ against it is integrating η pulled back to Σ_χ. Applied to Equation (26.46), and using that P restricts to the identity on $T\Sigma_\chi$, so that $P^*\Theta_\Sigma$ pulls back to Θ_Σ,

$$\int F\,\sqrt{\det C}\,\prod_\alpha \delta(\chi_\alpha)\,\Theta = \int_{\Sigma_\chi} F\,\sqrt{\det C}\,\frac{\Theta_\Sigma}{\sqrt{\det C}} = \int_{\Sigma_\chi} F\,\Theta_\Sigma\,,$$

which is Equation (26.41). □

Remark 26.26 (What the factor is doing). Equation (26.41) is the reason the Faddeev–Popov determinant is not an arbitrary insertion. The delta functions alone would define a measure on Σ_χ that depends on how the constraints are written: replacing χ_α by $A_\alpha{}^\beta\chi_\beta$ for an invertible matrix of functions leaves the surface untouched but multiplies $\prod\delta(\chi_\alpha)$ by $|\det A|^{-1}$. It multiplies $\sqrt{\det C}$ by $|\det A|$ as well, since $C \mapsto A\,C\,A^{\mathsf{T}}$ on the surface, so the product is invariant — and it is invariant because it equals the intrinsic Liouville volume of the reduced symplectic manifold, which knows nothing about the presentation. The same computation, read on the tangent space rather than on the volume, is what makes the Dirac bracket of Definition 26.20 the Poisson bracket of ω_Σ.

Rests on Propositions 26.24 and 26.25.

Remark 26.27 (Gauge conditions need not exist globally). Equation (26.38) is a local condition. For non-abelian gauge theories no gauge condition satisfies it over the whole space of field configurations—the Gribov ambiguity [1623]—so the reduced phase space is not covered by one chart, and a gauge orbit meets the gauge-fixing surface more than once. This is a genuine obstruction, not a technical inconvenience: it is invisible in perturbation theory about $A = 0$, which is why the Faddeev–Popov procedure works there, and it is exactly what fails when the coupling is strong. Its consequences for the non-perturbative treatment of Chapter 102 are noted there.

26.6 Worked systems

The formalism above is worth exactly as much as the cases it settles. The following are the ones this treatise needs; each is stated with its constraint structure and its degree-of-freedom count.

Notation 26.28 (Field systems). From Section 26.6.2 onward the systems are field theories. A plain $H = \int \mathrm{d}^3x\,\mathcal{H}$ is the Hamiltonian and the script $\mathcal{H}$ the corresponding density; the index j of Equation (26.11) carries a continuous label $\boldsymbol{x}$ and the sums over constraints become integrals; and the Poisson bracket is built from functional derivatives,

$$\{F,G\} = \int \mathrm{d}^3x \left(\frac{\delta F}{\delta\varphi(\boldsymbol{x})}\frac{\delta G}{\delta\pi(\boldsymbol{x})} - \frac{\delta F}{\delta\pi(\boldsymbol{x})}\frac{\delta G}{\delta\varphi(\boldsymbol{x})}\right), \tag{26.47}$$

so that $\{\varphi(\boldsymbol{x}),\pi(\boldsymbol{y})\} = \delta^3(\boldsymbol{x}-\boldsymbol{y})$. The evolution parameter is the coordinate time t of a chosen inertial frame, and $x^0 = ct$; a dot means ∂_t. Momenta conjugate to a field are taken with respect to ∂_t of the *covariant* components A_μ, which is what makes the Gauss constraint come out as Equation (26.56) rather than with the opposite sign. The prices and the limits of the passage from finitely many degrees of freedom to a field are collected in Section 26.8.

Rests on Notation 26.1 and Definition 22.29.

26.6.1 The relativistic free particle

Proposition 26.29 (Reparametrization invariance and the mass-shell constraint)**.** *For the action of a free relativistic particle parametrized by an arbitrary parameter τ,*

$$S = -mc \int \sqrt{\eta_{\mu\nu} \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} \frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}}\, \mathrm{d}\tau\,, \tag{26.48}$$

in the signature of Chapter 39, the Legendre transformation is singular of rank three, the canonical Hamiltonian vanishes identically, and the single primary constraint is the mass shell

$$\phi = \eta^{\mu\nu} p_\mu p_\nu - m^2 c^2 \approx 0\,, \tag{26.49}$$

which is first class. The count is $2 \times 4 - 2 \times 1 = 6$, that is three physical configuration degrees of freedom—the particle's position in space—as it must be.

Rests on Proposition 26.5 and Theorems 26.17 and 40.5.

Proof. *Derives Proposition 26.29.*

Write $\dot{x}^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$ and $\mathcal{L} = -mc\left(\eta_{\alpha\beta}\dot{x}^\alpha\dot{x}^\beta\right)^{1/2}$. The momenta are

$$p_\mu = \frac{\partial \mathcal{L}}{\partial \dot{x}^\mu} = -mc\, \frac{\eta_{\mu\nu}\dot{x}^\nu}{\sqrt{\eta_{\alpha\beta}\dot{x}^\alpha\dot{x}^\beta}}\,, \tag{26.50}$$

and contracting Equation (26.50) with itself the normalization cancels the denominator exactly: $\eta^{\mu\nu}p_\mu p_\nu = m^2c^2$ for *every* $\dot{x}$. That is Equation (26.49), and it holds identically in the velocities, which by Proposition 26.5 makes it a primary constraint. The Hessian

$$W_{\mu\nu} = \frac{\partial^2 \mathcal{L}}{\partial \dot{x}^\mu \partial \dot{x}^\nu} = -\frac{mc}{\sqrt{\dot{x}^2}}\left(\eta_{\mu\nu} - \frac{\dot{x}_\mu \dot{x}_\nu}{\dot{x}^2}\right) \tag{26.51}$$

is a multiple of the projector orthogonal to $\dot{x}$, of rank three, confirming $M = 4 - 3 = 1$: exactly one constraint, no more.

The canonical Hamiltonian is

$$\mathcal{H}_\mathrm{c} = p_\mu \dot{x}^\mu - \mathcal{L} = -mc\frac{\dot{x}^2}{\sqrt{\dot{x}^2}} + mc\sqrt{\dot{x}^2} = 0\,, \tag{26.52}$$

identically, so $\mathcal{H}_\mathrm{T} = u\phi$ by Equation (26.6). Consistency is automatic, $\dot{\phi} = \{\phi, u\phi\} = 0$: no secondary constraint arises and u is never determined. By Theorem 26.14 the constraint is first class — there is nothing for its bracket to fail to annihilate — and u is a gauge parameter. The equations of motion it generates,

$$\dot{x}^\mu = \{x^\mu, \mathcal{H}_\mathrm{T}\} = 2u\, p^\mu\,, \qquad \dot{p}_\mu = 0\,, \tag{26.53}$$

determine the worldline but not its parametrization: changing $u(\tau)$ reparametrizes τ and nothing else. Choosing τ to be the proper time of Definition 38.12 fixes $u = 1/2m$ and returns $p^\mu = m\,\mathrm{d}x^\mu/\mathrm{d}\tau$, the four-momentum of Definition 40.4. Finally $2n = 8$, $F = 1$, $S = 0$, so Equation (26.21) gives $8 - 2 = 6$ and three configuration degrees of freedom. □

Remark 26.30 (A sign to be read once). With the signature $(+,-,-,-)$ and the overall minus in Equation (26.48), Equation (26.50) evaluated in the parametrization $\tau = t$ gives $p_i = \gamma_u m\, u_i$ — the ordinary relativistic momentum, with its familiar sign — but $p_0 = -E/c$. The canonical momenta conjugate to the x^μ are thus the components of $-p_\mu$ in the convention of Definition 40.4. Nothing in the constraint analysis depends on this: Equation (26.49) is quadratic and blind to the overall sign, and it is the same statement as Equation (40.5). The sign is recorded because it is a standing trap in the literature, where both conventions are used without comment.

Remark 26.31 (A vanishing Hamiltonian is not a static system). That $\mathcal{H}_{\mathrm{c}} \approx 0$ says only that the evolution in the arbitrary parameter is pure gauge; the physical motion is the relation between the x^{μ}, which is unaffected. Any theory whose invariance group includes reparametrizations of the evolution parameter has this property, and the confusion it causes when the parameter is called "time" is discussed under Section 26.6.5.

26.6.2 The electromagnetic field

Proposition 26.32 (Constraint structure of the free Maxwell field)**.** *For the Maxwell Lagrangian density of Section 64.5.1,*

$$\mathcal{L} = -\frac{1}{4\mu_0} F_{\mu\nu}F^{\mu\nu} - A_{\mu}J^{\mu}\,, \qquad F_{\mu\nu} = \partial_{\mu}A_{\nu} - \partial_{\nu}A_{\mu}\,, \tag{26.54}$$

the momentum conjugate to A_0 vanishes identically, giving the primary constraint

$$\pi^{0}(\boldsymbol{x}) \approx 0\,, \tag{26.55}$$

whose consistency yields the secondary constraint—Gauss's law—

$$\partial_i \pi^{i}(\boldsymbol{x}) - \rho(\boldsymbol{x}) \approx 0\,. \tag{26.56}$$

Both are first class, and they generate the gauge transformations $A_{\mu} \mapsto A_{\mu} + \partial_{\mu}\Lambda$. The count per point of space is $2 \times 4 - 2 \times 2 = 4$, that is two *physical field degrees of freedom. The algorithm closes after two rounds if and only if the source is conserved, $\partial_{\mu}J^{\mu} = 0$.*

Rests on Propositions 26.5 and 26.10 and Theorem 26.17.

Proof. *Derives Proposition 26.32.*

Since $F_{\mu\nu}$ is antisymmetric, $\mathcal{L}$ contains no $\partial_0 A_0$ at all, and

$$\pi^{\mu} := \frac{\partial \mathcal{L}}{\partial \dot{A}_{\mu}} = -\frac{1}{\mu_0 c} F^{0\mu}\,, \qquad \text{so} \qquad \pi^{0} = 0\,, \quad \pi^{i} = \epsilon_0 E^{i}\,, \tag{26.57}$$

using $F^{0i} = -E^{i}/c$ and $\epsilon_0\mu_0 c^2 = 1$. The first is Equation (26.55), one primary constraint per point; the second identifies the field momentum with the electric displacement, of dimension $\mathrm{C/m^2}$. The Hessian in the four $\dot{A}_{\mu}$ has rank three per point, so there is exactly one.

The canonical Hamiltonian follows from Equation (26.4). Using $\dot{A}_i = E_i + \partial_i \phi$ with $\phi = cA_0$ the electrostatic potential, and integrating the term $\epsilon_0 \boldsymbol{E}\cdot\nabla\phi$ by parts,

$$H_{\mathrm{c}} = \int \mathrm{d}^3x \left[\frac{\pi^{i}\pi^{i}}{2\epsilon_0} + \frac{(\nabla \times \boldsymbol{A})^2}{2\mu_0} - c\,A_0\left(\partial_i \pi^{i} - \rho\right) - \boldsymbol{J}\cdot\boldsymbol{A} \right], \tag{26.58}$$

whose first two terms are the familiar field energy density $\tfrac{1}{2}\epsilon_0 E^2 + B^2/2\mu_0$. The potential A_0 appears *linearly and without derivatives*: it is a multiplier, not a dynamical variable. Consistency of Equation (26.55) is then

$$\dot{\pi}^{0}(\boldsymbol{x}) = \left\{ \pi^{0}(\boldsymbol{x}), H_{\mathrm{T}} \right\} = -\frac{\delta H_{\mathrm{c}}}{\delta A_0(\boldsymbol{x})} = c\left(\partial_i \pi^{i} - \rho\right) \approx 0\,, \tag{26.59}$$

which is outcome 2 of Proposition 26.10: a new constraint containing no multiplier. It is Equation (26.56), and written out it is $\epsilon_0 \nabla\cdot\boldsymbol{E} = \rho$ — Gauss's law appears here as a *constraint on initial data*, not as an equation of motion, which is the structural statement this analysis makes about Maxwell's system.

Running the algorithm once more, the remaining Hamilton equation is $\dot{\pi}^{i} = -\delta H_{\mathrm{c}}/\delta A_i = (\nabla\times\boldsymbol{B})_i/\mu_0 - J_i$, the Ampère–Maxwell law, so

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\partial_i \pi^{i} - \rho\right) = \partial_i \frac{(\nabla\times\boldsymbol{B})_i}{\mu_0} - \nabla\cdot\boldsymbol{J} - \dot{\rho} = -\left(\dot{\rho} + \nabla\cdot\boldsymbol{J}\right), \tag{26.60}$$

the divergence of a curl vanishing identically. The algorithm therefore terminates precisely when $\partial_\mu J^\mu = 0$, and otherwise reaches outcome 4 of Proposition 26.10: a Maxwell field coupled to a non-conserved source has no solutions at all.

Both constraints are first class. π^0 has vanishing bracket with everything built from $\boldsymbol{A}$ and $\boldsymbol{\pi}$; and $\{\mathcal{G}(\boldsymbol{x}), \mathcal{G}(\boldsymbol{y})\} = 0$ identically, where $\mathcal{G} = \partial_i \pi^i - \rho$, because $\mathcal{G}$ depends on the momenta alone. So $F = 2$ per point, $S = 0$, $2n = 8$, and Equation (26.21) gives $8 - 4 = 4$: two configuration degrees of freedom per point of space.

It remains to identify the gauge transformation. Take

$$G[\Lambda] = \int \mathrm{d}^3x \left[\frac{\dot{\Lambda}}{c}\, \pi^0 - \Lambda \left(\partial_i \pi^i - \rho \right) \right] , \tag{26.61}$$

a combination of the primary and secondary constraints with an arbitrary function Λ and its time derivative. Then $\{A_0, G\} = \dot{\Lambda}/c = \partial_0 \Lambda$ and, integrating the second term by parts, $\{A_i, G\} = \partial_i \Lambda$, while $\{\pi^\mu, G\} = 0$. That is $A_\mu \mapsto A_\mu + \partial_\mu \Lambda$ with $\boldsymbol{E}$ and $\boldsymbol{B}$ unchanged, as claimed. □

Remark 26.33 (The count is the observation). The two degrees of freedom of Proposition 26.32 are the two transverse polarizations of light. They are not a formal result: they are what is measured whenever the polarization of an electromagnetic wave is analysed (Chapter 67), and the absence of a third, longitudinal, mode is a statement about nature that the constraint analysis explains rather than assumes. The same count performed for a massless spin-two field gives two polarizations again—the plus and cross modes of Chapter 46, observed as such in the detections of Chapter 47 [2132].

26.6.3 The massive vector field

Proposition 26.34 (Proca constraints are second class)**.** *For the massive vector field of Section 64.6.1 [785],*

$$\mathcal{L} = -\frac{1}{4\mu_0} F_{\mu\nu} F^{\mu\nu} + \frac{1}{2\mu_0} \left(\frac{mc}{\hbar} \right)^2 A_\mu A^\mu , \tag{26.62}$$

the primary constraint is again $\pi^0 \approx 0$ but its consistency returns

$$\partial_i \pi^i + \frac{1}{\mu_0 c} \left(\frac{mc}{\hbar} \right)^2 A_0 - \rho \approx 0 , \tag{26.63}$$

and the two have the nonvanishing bracket

$$\left\{ \pi^0(\boldsymbol{x}), \cdot \right\} \longrightarrow -\frac{1}{\mu_0 c} \left(\frac{mc}{\hbar} \right)^2 \delta^3(\boldsymbol{x} - \boldsymbol{y}) , \tag{26.64}$$

proportional to the mass squared. They are second class rather than first: there is no gauge invariance. The count per point is $2 \times 4 - 2 = 6$, that is three physical degrees of freedom—the three polarizations of a massive spin-one particle.

Rests on Proposition 26.32, Definition 26.12, and Theorem 26.17.

Proof. *Derives Proposition 26.34.*

The mass term of Equation (26.62) carries no derivatives, so the momenta Equation (26.57) are unchanged and $\pi^0 = 0$ is still the only primary constraint. The canonical Hamiltonian Equation (26.58) acquires the extra density $-(2\mu_0)^{-1} (mc/\hbar)^2 (A_0^2 - A_i A_i)$, and now A_0 appears *quadratically*. Consistency of $\pi^0 \approx 0$ therefore gives

$$\dot{\pi}^0 = -\frac{\delta H_{\mathrm{c}}}{\delta A_0} = c \left[\partial_i \pi^i - \rho + \frac{1}{\mu_0 c} \left(\frac{mc}{\hbar} \right)^2 A_0 \right] \approx 0 , \tag{26.65}$$

which is Equation (26.63). Writing $\chi_1 = \pi^0$ and χ_2 for the left-hand side of Equation (26.63), the only term of χ_2 that fails to commute with π^0 is the one containing A_0, and

$$\{\chi_1(\boldsymbol{x}), \chi_2(\boldsymbol{y})\} = -\frac{1}{\mu_0 c}\left(\frac{mc}{\hbar}\right)^2 \delta^3(\boldsymbol{x}-\boldsymbol{y})\,, \tag{26.66}$$

which is Equation (26.64) and is invertible for every $m \neq 0$. By Definition 26.12 the pair is second class, so $S = 2$ and $F = 0$; consistency of χ_2 then determines the multiplier instead of producing a third constraint, and the algorithm stops. Equation (26.21) gives $8 - 0 - 2 = 6$ per point, three configuration degrees of freedom.

The content of Equation (26.63) is worth stating separately: it *determines* A_0 algebraically in terms of the momenta,

$$A_0 = -\mu_0 c\left(\frac{\hbar}{mc}\right)^2 \left(\partial_i \pi^i - \rho\right)\,, \tag{26.67}$$

so A_0 is not free, not gauge, and not dynamical — it is an auxiliary field. Combining Equation (26.67) with the equation of motion $\partial_\mu F^{\mu\nu} + (mc/\hbar)^2 A^\nu = \mu_0 J^\nu$ and taking a divergence returns $\partial_\mu A^\mu = 0$ for a conserved source: the Lorenz condition, here a consequence rather than a choice, exactly as Section 64.6.1 says. □

Remark 26.35 (Why the massive and massless counts differ). The passage from Proposition 26.32 to Proposition 26.34 changes the classification of the constraints, not their number, and that is what changes the count from two to three. Note precisely where the change enters: Equation (26.66) is proportional to m^2, so the bracket that makes the pair second class vanishes discontinuously at $m = 0$; the count is a step function of the mass, and no $m \to 0$ limit of the massive theory has two degrees of freedom. The physical content is that a mass term breaks the gauge invariance and thereby liberates the longitudinal mode. Classically this is the shadow of the van Dam–Veltman–Zakharov discontinuity [1419, 1439], where the corresponding step in the spin-two count leaves a massless limit that disagrees with general relativity. Experimentally it is why "the photon is massless" is a bound and never a measurement: the current limit is $m_\gamma < 1 \times 10^{-18}\,\mathrm{eV}/c^2$ [2427], from laboratory tests of Coulomb's law [1460] and, far more sharply, from planetary and galactic magnetic fields [2157]. And it is the mechanism behind the massive W and Z bosons of Chapter 104, with $M_W c^2 = (80.3692 \pm 0.0133)\,\mathrm{GeV}$ and $M_Z c^2 = (91.1880 \pm 0.0020)\,\mathrm{GeV}$ [2427], where the liberated modes are supplied by the scalar field rather than imposed by hand.

26.6.4 Yang–Mills fields

The non-abelian case is the one the Standard Model actually uses, and it differs from Proposition 26.32 in exactly one place: the Gauss constraints no longer commute with one another. Everything else — the vanishing time-component momentum, the secondary constraint, the count of two polarizations per generator — survives unchanged.

Notation 26.36 (Gauge fields and their SI dimensions). Let the gauge group be compact and simple of dimension N, with structure constants f^{abc} totally antisymmetric in the basis of Notation 102.18, whose conventions are used throughout. The potentials A^a_μ carry the dimension V s/m of the electromagnetic potential and the coupling g the dimension C of a charge, so that

$$\left[\frac{g}{\hbar}\right] = \mathrm{C}/(\mathrm{J\,s})\,, \qquad \left[\frac{g}{\hbar}A^a_\mu\right] = 1/\mathrm{m}\,. \tag{26.68}$$

The ratio $g/\hbar$ is the only combination in which the coupling occurs below, and by Equation (26.68) it is what turns a potential into an inverse length; it is the scalar behind the matrix-valued combination Equation (102.18). The field strength is Equation (102.24),

$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - \frac{g}{\hbar} f^{abc} A^b_\mu A^c_\nu$, of dimension T, and the covariant derivative in the adjoint representation acts on a colour vector X^a as

$$(D_i X)^a = \partial_i X^a - \frac{g}{\hbar} f^{abc} A^b_i X^c \,. \tag{26.69}$$

Rests on Notations 26.28 and 102.18 and Equation (102.24).

Proposition 26.37 (Constraint structure of Yang–Mills theory)**.** *For* $\mathcal{L} = -(4\mu_0)^{-1} F^a_{\mu\nu} F^{a\,\mu\nu}$ *coupled to matter of colour charge density* ρ^a*, the momenta are* $\pi^0_a = 0$ *and* $\pi^i_a = \epsilon_0 E^i_a$*, so there are* N *primary constraints*

$$\pi^0_a(\boldsymbol{x}) \approx 0 \,, \qquad a = 1, \dots, N \,, \tag{26.70}$$

whose consistency yields the N *secondary constraints — the non-abelian Gauss law —*

$$\mathcal{G}^a := \left(D_i \pi^i\right)^a - \rho^a \approx 0 \,. \tag{26.71}$$

All $2N$ *are first class, and the Gauss constraints close on the Lie algebra of the gauge group,*

$$\left\{\mathcal{G}^a(\boldsymbol{x}), \mathcal{G}^b(\boldsymbol{y})\right\} = -\frac{g}{\hbar} f^{abc}\, \mathcal{G}^c(\boldsymbol{x})\, \delta^3(\boldsymbol{x} - \boldsymbol{y}) \,, \tag{26.72}$$

with genuine constants f^{abc} *where Equation (26.56) had zero. The count per point of space is* $2 \times 4N - 2 \times 2N = 4N$*: two physical degrees of freedom per generator, that is two transverse polarizations for each of the* N *gauge bosons.*

Rests on Propositions 26.13 and 26.32 and Notation 26.36.

Proof. *Derives Proposition 26.37.*

$F^a_{\mu\nu}$ is antisymmetric in $\mu\nu$ whatever the nonlinear term does, so no $\partial_0 A^a_0$ occurs in $\mathcal{L}$ and Equation (26.70) follows exactly as Equation (26.55) did. The Legendre transformation gives, by the computation of Proposition 26.32 with ∂_i replaced by D_i throughout,

$$H_{\mathrm{c}} = \int \mathrm{d}^3x \left[\frac{\pi^i_a \pi^i_a}{2\epsilon_0} + \frac{B^i_a B^i_a}{2\mu_0} - c\, A^a_0\, \mathcal{G}^a \right] , \tag{26.73}$$

with $B^i_a = -\frac{1}{2}\epsilon^{ijk} F^a_{jk}$ the chromomagnetic field, in which A^a_0 is again a multiplier, so that consistency of Equation (26.70) produces Equation (26.71). The two energy terms are the abelian ones; what the nonlinearity of Equation (102.24) adds is hidden inside B^i_a, which is cubic and quartic in A, so the free Yang–Mills Hamiltonian already describes a self-interacting field.

For the algebra, smear the constraint with a colour-valued test function, $G[\varepsilon] := \int \mathrm{d}^3x\, \varepsilon^a \mathcal{G}^a$, and take the matter contribution to ρ^a to generate the colour rotation of the matter fields, which it does by construction. Then

$$\frac{\delta G[\varepsilon]}{\delta A^b_i} = -\frac{g}{\hbar} f^{abc} \varepsilon^a \pi^i_c \,, \qquad \frac{\delta G[\varepsilon]}{\delta \pi^i_b} = -\left(D_i \varepsilon\right)^b \,, \tag{26.74}$$

the second by Equation (26.69) and the total antisymmetry of f. Substituting both into Equation (26.47),

$$\{G[\varepsilon], G[\eta]\} = \frac{g}{\hbar} \int \mathrm{d}^3x\; f^{abc} \pi^i_c \left[\varepsilon^a \left(D_i \eta\right)^b - \eta^a \left(D_i \varepsilon\right)^b \right] . \tag{26.75}$$

Integrate the first term by parts. The adjoint derivative is antisymmetric under integration by parts because f is totally antisymmetric, and it is a derivation, so

$$\int f^{abc} \pi^i_c \varepsilon^a \left(D_i \eta\right)^b = -\int \eta^b \left[f^{abc} \left(D_i \varepsilon\right)^a \pi^i_c + f^{abc} \varepsilon^a \mathcal{G}^c \right] , \tag{26.76}$$

where $(D_i\pi^i)^c$ has been recognised as $\mathcal{G}^c$ up to the matter term, which the matter bracket supplies. The first bracket on the right cancels the second term of Equation (26.75) after relabelling $a \leftrightarrow b$ and using $f^{abc} = -f^{bac}$, leaving

$$\{G[\varepsilon], G[\eta]\} = -\frac{g}{\hbar}\int \mathrm{d}^3x\; f^{cab}\varepsilon^a\eta^b\,\mathcal{G}^c = -\frac{g}{\hbar}\,G[[\varepsilon,\eta]]\;, \qquad [\varepsilon,\eta]^c = f^{cab}\varepsilon^a\eta^b\,, \tag{26.77}$$

which unsmeared is Equation (26.72). Since the right-hand side is a combination of the constraints, and $\left\{\pi^0_a, \mathcal{G}^b\right\} = 0$ because $\mathcal{G}$ contains no A^a_0, all $2N$ constraints are first class by Definition 26.12, and Proposition 26.13 is realised with constant structure functions. The count follows from Equation (26.21) with $2n = 8N$, $F = 2N$, $S = 0$. □

Remark 26.38 (What the non-abelian term costs). Three consequences follow from Equation (26.72) and none of them is available in the abelian case. First, the gauge generator Equation (26.61) becomes $G[\Lambda] = \int \left[c^{-1}\dot{\Lambda}^a\pi^0_a - \Lambda^a\mathcal{G}^a\right]$ with Λ colour-valued, and the transformation it generates is Equation (102.23) — so the classical Hamiltonian analysis and the Lagrangian gauge principle Theorem 102.19 agree, which is the check on the whole construction. Second, the Faddeev–Popov determinant of Proposition 26.24 is no longer field independent: for a linear gauge condition $\psi^a = \partial_i A^a_i$ the matrix $\left\{\mathcal{G}^a, \psi^b\right\}$ is $\partial_i D_i$ rather than ∇^2, which depends on A and therefore cannot be dropped as a constant — this is exactly the classical input to [1347] and the reason ghosts are unavoidable in Chapter 102 and absent from abelian Chapter 100. Third, $\partial_i D_i$ has zero modes for large enough A, which is Remark 26.27 said in the canonical language.

Remark 26.39 (The count against the evidence). Two polarizations per generator is a testable statement, and the two places it has been tested are the eight gluons of Chapter 102 — whose count enters the running of α_s through the coefficient of the beta function, measured over three decades of energy — and the electroweak sector, where the three broken generators acquire mass and their third polarization along with it (Chapter 104), while the photon keeps two. Neither count is postulated; both follow from Proposition 26.37 and its massive counterpart Proposition 26.34, and both are what is observed.

26.6.5 General relativity

General relativity is a constrained system of the most extreme kind: its canonical Hamiltonian is a sum of constraints and therefore vanishes weakly, exactly as for the relativistic particle of Proposition 26.29 and for the same reason — the invariance group contains reparametrizations of the evolution parameter. The decomposition that exhibits this is due to Arnowitt, Deser and Misner [1221].

Definition 26.40 (The $3+1$ decomposition)**.** Foliate spacetime by spacelike surfaces $t = \text{const}$ and write, with $x^0 = ct$ and in the signature of Chapter 39,

$$\mathrm{d}s^2 = N^2\left(\mathrm{d}x^0\right)^2 - h_{ij}\left(\mathrm{d}x^i + N^i\mathrm{d}x^0\right)\left(\mathrm{d}x^j + N^j\mathrm{d}x^0\right)\,, \tag{26.78}$$

with h_{ij} the positive-definite Riemannian metric of the surface — so that $h_{ij}\mathrm{d}x^i\mathrm{d}x^j$ is squared proper distance, and the metric g_{ij} induced by Equation (26.78) is $-h_{ij}$ in this signature — and with N, N^i dimensionless. N is the *lapse*, measuring proper time per unit x^0 along the normal, and N^i the *shift*, measuring how the spatial coordinates slide from one surface to the next. Neither is a property of the geometry: both encode the choice of foliation and of coordinates on it. The *extrinsic curvature* of the surface is

$$K_{ij} = \frac{1}{2N}\left(\partial_0 h_{ij} - D_iN_j - D_jN_i\right)\,, \tag{26.79}$$

with $N_i = h_{ij}N^j$ and D_i the Levi-Civita connection of h_{ij}; it carries dimension 1/m and is the only place the time derivative of the metric enters.

Rests on Definition 43.2 and Notation 26.28.

Proposition 26.41 (The ADM form of the Einstein–Hilbert Lagrangian)**.** *With Equation* (26.78) *one has* $\sqrt{|g|} = N\sqrt{h}$ *and*

$$\sqrt{|g|}\,R = N\sqrt{h}\left({}^{(3)}R + K_{ij}K^{ij} - K^2\right) + \text{total derivatives}\,, \tag{26.80}$$

where ${}^{(3)}R$ *is the Ricci scalar of* h_{ij} *and* $K = h^{ij}K_{ij}$*. The Lagrangian of Equation* (43.4) *therefore contains* no *time derivative of* N *or of* N^i*.*

Rests on Definition 26.40 and Equation (43.4)*.*

Full derivation: Appendix A.52 (p. 3073).

Derives Proposition 26.41.

The route taken there is the Gauss–Codazzi decomposition of the four-dimensional Riemann tensor: its projections onto the leaf and onto the unit normal give the Gauss equation, which relates the wholly tangential projection to the intrinsic curvature and to a quadratic in K_{ij}, and the Codazzi equation for the mixed projection; contracting twice assembles Equation (26.80) together with two divergences, one of which is the term whose subtraction is the Gibbons–Hawking–York boundary action. The point of care specific to this treatise is the sign bookkeeping between the mostly-minus signature of Chapter 39 and the positive-definite induced metric of Definition 26.40, which is settled there explicitly rather than inherited from a mostly-plus source. Only the two facts displayed above are used in this chapter: that $\sqrt{|g|} = N\sqrt{h}$ and that no time derivative of N or N^i survives.

Theorem 26.42 (The Hamiltonian structure of general relativity)**.** *The momenta conjugate to the lapse and the shift vanish identically,*

$$\pi_N \approx 0\,, \qquad \pi_i \approx 0\,, \tag{26.81}$$

four primary constraints per point of space, while

$$\pi^{ij} = \frac{\partial \mathcal{L}}{\partial\left(\partial_0 h_{ij}\right)} = \frac{\sqrt{h}}{2\kappa}\left(K^{ij} - K h^{ij}\right)\,, \qquad \kappa = \frac{8\pi G}{c^4}\,, \tag{26.82}$$

is unconstrained. The canonical Hamiltonian is a sum of constraints,

$$H_c = \int \mathrm{d}^3x \left(N\,\mathcal{H}_\perp + N^i\,\mathcal{H}_i\right) + \text{boundary terms}\,, \tag{26.83}$$

with

$$\mathcal{H}_\perp = \frac{2\kappa}{\sqrt{h}}\left(\pi_{ij}\pi^{ij} - \tfrac{1}{2}\pi^2\right) - \frac{\sqrt{h}}{2\kappa}\,{}^{(3)}R\,, \tag{26.84}$$

$$\mathcal{H}_i = -2\,D_j\pi^j{}_i\,, \tag{26.85}$$

so that consistency of Equation (26.81) *yields the four secondary constraints* $\mathcal{H}_\perp \approx 0$ *and* $\mathcal{H}_i \approx 0$*. All eight are first class, the canonical Hamiltonian is weakly zero, and the count per point of space is* $2\times 10 - 2\times 8 = 4$*:* two *physical degrees of freedom.*

Rests on Propositions 26.10 and 26.41 and Theorem 26.17.

Proof. *Derives Theorem 26.42.*

By Proposition 26.41 the Lagrangian $L = (2\kappa)^{-1}\int \mathrm{d}^3x\, N\sqrt{h}\left({}^{(3)}R + K_{ij}K^{ij} - K^2\right)$ contains $\partial_0 N$ and $\partial_0 N^i$ nowhere, so their conjugate momenta vanish identically and Equation (26.81) are primary constraints in the sense of Proposition 26.5. The ten components of h_{ij} and of N, N^i make $2n = 2\times 10$ per point.

For π^{ij}, only K carries $\partial_0 h_{ij}$, and by Equation (26.79) $\partial K_{ij}/\partial\left(\partial_0 h_{kl}\right) = \delta^{(kl)}_{(ij)}/2N$, so

$$\pi^{kl} = \frac{N\sqrt{h}}{2\kappa}\cdot\frac{1}{N}\left(K^{kl} - Kh^{kl}\right), \tag{26.86}$$

which is Equation (26.82). Its trace is $\pi = -\sqrt{h}\,K/\kappa$, so the relation inverts to $K^{ij} = 2\kappa h^{-1/2}\left(\pi^{ij} - \frac{1}{2}\pi h^{ij}\right)$ and

$$K_{ij}K^{ij} - K^2 = \frac{4\kappa^2}{h}\left(\pi_{ij}\pi^{ij} - \tfrac{1}{2}\pi^2\right), \tag{26.87}$$

by direct substitution and collection of the two π^2 terms. Now Legendre transform. Using $\partial_0 h_{ij} = 2NK_{ij} + D_iN_j + D_jN_i$ from Equation (26.79),

$$\pi^{ij}\partial_0 h_{ij} - \mathcal{L} = \frac{N\sqrt{h}}{2\kappa}\left(K_{ij}K^{ij} - K^2\right) - \frac{N\sqrt{h}}{2\kappa}\,{}^{(3)}R + 2\pi^{ij}D_iN_j\,, \tag{26.88}$$

the first term because $2N\pi^{ij}K_{ij}$ is twice the corresponding term of $\mathcal{L}$. Substituting Equation (26.87) into the first term gives $N\mathcal{H}_\perp$ with $\mathcal{H}_\perp$ as in Equation (26.84); integrating the last term by parts gives $-2N_iD_j\pi^{ji}$ plus a surface term, that is $N^i\mathcal{H}_i$ with $\mathcal{H}_i$ as in Equation (26.85). That is Equation (26.83).

Since N and N^i appear in Equation (26.83) linearly and without derivatives, they are multipliers, and consistency of Equation (26.81) reads $\dot\pi_N = -\delta H_c/\delta N = -\mathcal{H}_\perp \approx 0$ and $\dot\pi_i = -\mathcal{H}_i \approx 0$: four secondary constraints, outcome 2 of Proposition 26.10. That they are first class is Equation (26.89) below, whose right-hand sides are combinations of the constraints; the primary constraints commute with everything because no $\mathcal{H}$ contains N or N^i. Hence $F = 8$, $S = 0$, $2n = 20$, and Equation (26.21) gives $20 - 16 = 4$: two configuration degrees of freedom per point of space. Finally, substituting the secondary constraints into Equation (26.83) gives $H_c \approx 0$ up to the boundary terms of Remark 26.45. □

Proposition 26.43 (The hypersurface-deformation algebra)**.** *Smearing $\mathcal{H}_\perp$ with a scalar f and $\mathcal{H}_i$ with a vector ξ^i, and writing $H_\perp[f] = \int f\mathcal{H}_\perp$, $H_\parallel[\xi] = \int \xi^i\mathcal{H}_i$, the constraints close as*

$$\left\{H_\parallel[\xi], H_\parallel[\zeta]\right\} = H_\parallel[[\xi,\zeta]]\,, \tag{26.89}$$

$$\left\{H_\parallel[\xi], H_\perp[f]\right\} = H_\perp\left[\xi^i\partial_i f\right], \tag{26.90}$$

$$\left\{H_\perp[f], H_\perp[g]\right\} = H_\parallel\left[h^{ij}\left(f\partial_j g - g\partial_j f\right)\right]. \tag{26.91}$$

The first two say that the momentum constraints generate spatial diffeomorphisms and that $\mathcal{H}_\perp$ is a scalar density under them. The third is the one that matters: its coefficient is h^{ij}, a function on phase space, so the first-class algebra of general relativity has structure functions *and is not a Lie algebra.*

Rests on Theorem 26.42 and Proposition 26.13.

Full derivation: Appendix A.53 (p. 3082).

Derives Proposition 26.43.

Each bracket is a direct computation from the densities Equations (26.84) and (26.85), and the third is long: both the kinetic and the intrinsic-curvature terms contribute, and the curvature contribution has to be integrated by parts twice before the surviving terms assemble into $\mathcal{H}_i$. The geometric reading, written there beside the computation, is that these brackets reproduce the composition law of deformations of a spacelike surface embedded in a Lorentzian spacetime; the metric appears in the third because a normal deformation at one point becomes a tangential one at a neighbouring point through the geometry of the surface itself, and that is why h^{ij} — a function on phase space, not a constant — sits where a structure constant would.

Remark 26.44 (Three consequences, and one that is out of scope). *The constraints are the initial-value equations.* The four secondary constraints of Theorem 26.42 involve no time derivative of π^{ij}: they are conditions on the data (h_{ij}, π^{ij}) laid down on one surface, and they are precisely the $G^0{}_\mu = \kappa T^0{}_\mu$ components of the Einstein equations. Not every initial datum is admissible, and the constraints say which are — this is the content of Sections 44.9 and 44.9.1, and the well-posedness of the resulting evolution problem is [999], with the constraint-solving method of [1490].

The count matches the waves. Two degrees of freedom per point is the same answer Proposition 26.32 gave for light, and it is the two polarizations of Chapter 46, detected as such in Chapter 47. That a theory with ten metric functions propagates two is not obvious from the field equations and is transparent here.

The Hamiltonian vanishes. $H_{\mathrm{c}} \approx 0$ says, by Remark 26.46, that evolution in the coordinate t is a gauge transformation — a relabelling of the foliation — and not a physical motion. What is physical is the relation between the geometry and the matter in it, exactly as what was physical for the relativistic particle was the worldline and not its parametrization.

What is not claimed. The statement that this vanishing creates a "problem of time" is a statement about a *quantum* theory of gravity, in which Equation (26.92) would read $\hat{\mathcal{H}}_\perp |\psi\rangle = 0$ and leave the state vector with no evolution parameter. No such theory has observational support, so it is outside the evidence base of this treatise and is treated only in Chapter 135. Classically there is no problem of time and nothing to solve.

Remark 26.45 (The energy is a boundary term). The surface terms discarded in Equation (26.83) cannot in fact be discarded. On a region with a boundary, or in an asymptotically flat spacetime, the constraints do not have well-defined functional derivatives unless they are supplemented by surface integrals; the supplement to $\mathcal{H}_\perp$ is the ADM mass and the supplement to $\mathcal{H}_i$ the ADM momentum and angular momentum [1221, 1746]. So the total energy of an isolated gravitating system is a boundary integral at spatial infinity and not the volume integral of any density: the numerical value of the Hamiltonian is carried entirely by the surface term, the bulk being weakly zero. This is the constrained-dynamics origin of the ADM integrals referred to in Chapter 44, and it is why the localization of gravitational energy is not a well-posed question.

Remark 26.46 (A vanishing Hamiltonian, again). Remark 26.31 was stated for the relativistic particle and applies verbatim here, with the reparametrizations of the worldline replaced by the deformations of the spacelike surface. The two systems are the same phenomenon at different sizes, which is why the particle is worth doing first.

26.7 Quantizing a constrained system

Postulate 26.47 (Dirac quantization)**.** Quantize a constrained system by

1. eliminating the second-class constraints, replacing the Poisson bracket by the Dirac bracket Equation (26.25) and imposing $\chi_\alpha = 0$ strongly; then applying the correspondence rule Equation (25.31) to the Dirac bracket, $\{A, B\}_{\mathrm{D}} \mapsto \frac{1}{\mathrm{i}\hbar}\left[\hat{A}, \hat{B}\right]$; and
2. imposing the first-class constraints as conditions on the *states* rather than on the operators,
$$\hat{\gamma}_A |\psi\rangle = 0\,, \tag{26.92}$$
the physical Hilbert space being the space of solutions.

Rests on Postulate 25.27 and Theorems 26.14 and 26.21.

The asymmetry between the two clauses is the whole design and is forced by Theorem 26.21. A second-class constraint has vanishing Dirac bracket with everything, so it may

be set to zero as an operator identity without contradicting any commutator; a first-class constraint may not, because it generates a nontrivial transformation and setting it to zero as an operator would set that generator to zero too. Imposing it on states instead selects the gauge-invariant sector, which is exactly what Theorem 26.14 says the physical states are.

Proposition 26.48 (Consistency of the state conditions)**.** *Equation* (26.92) *is consistent only if the quantum constraint algebra closes,*

$$[\hat{\gamma}_A, \hat{\gamma}_B] = \mathrm{i}\hbar\, \hat{f}_{AB}{}^{C} \hat{\gamma}_C \,, \tag{26.93}$$

with the constraint operators standing to the right*. A failure of Equation* (26.93) *is an* anomaly*: a symmetry of the classical theory that no quantization preserves.*

Rests on Postulate 26.47 and Equation (26.16)*.*

Proof. *Derives Proposition 26.48.*

Let $|\psi\rangle$ satisfy Equation (26.92) for every A. Applying the commutator to it, $[\hat{\gamma}_A, \hat{\gamma}_B]\,|\psi\rangle = \hat{\gamma}_A\hat{\gamma}_B\,|\psi\rangle - \hat{\gamma}_B\hat{\gamma}_A\,|\psi\rangle = 0$, so the commutator must annihilate every physical state. The classical bracket is $f_{AB}{}^{C}\gamma_C$ by Equation (26.16), and its operator image annihilates $|\psi\rangle$ if and only if every $\hat{\gamma}_C$ stands to the right of the corresponding $\hat{f}_{AB}{}^{C}$: the opposite ordering leaves $\hat{\gamma}_C\hat{f}_{AB}{}^{C}\,|\psi\rangle = \big[\hat{\gamma}_C, \hat{f}_{AB}{}^{C}\big]\,|\psi\rangle$, which need not vanish. Where the structure functions are constants the question does not arise, which is why Equation (26.72) is easier to quantize than Equation (26.91). If no operator ordering makes Equation (26.93) hold, the conditions Equation (26.92) are mutually inconsistent and only $|\psi\rangle = 0$ satisfies them.

This is the ordering ambiguity of Definition 25.42 seen from the other end. There, the choice among $\hat{q}\hat{p}$, $\hat{p}\hat{q}$ and their symmetrizations was underdetermined by the correspondence rule and had to be fixed by experiment (Remark 25.49); here the requirement that a gauge symmetry survive quantization fixes part of the same choice, and Theorem 25.38 guarantees that it cannot always be fixed consistently. An anomaly is precisely the case in which it cannot. □

Remark 26.49 (Anomalies are an experimental constraint). An anomaly in a *global* symmetry is a prediction, and a correct one: the observed decay of the neutral pion to two photons is the textbook case [1388, 1391], and its rate counts the colours (Proposition 102.16). An anomaly in a *gauge* symmetry is fatal, because Equation (26.92) then has no solutions and the theory has no physical states. That the gauge anomalies of the Standard Model cancel between quarks and leptons, generation by generation, is therefore not an aesthetic observation but a consistency condition that the observed particle content satisfies and that constrains any proposed addition to it (Chapters 104 and 105). The cancellation is not only of the triangle kind: a global obstruction also exists, and it requires the number of left-handed SU(2) doublets to be even [1707] — which the observed generations satisfy, one quark doublet in three colours plus one lepton doublet making four.

26.7.1 Reduced-phase-space quantization

The alternative to Postulate 26.47 is to do all the classical work first: fix the gauge as in Definition 26.23, construct the reduced phase space of dimension $2n - 2F - S$, take its Dirac bracket, and only then apply the correspondence rule. The two routes are often called *constrain-then-quantize* and *quantize-then-constrain*, and the honest statement is that they are not equivalent.

They differ in three ways, each of which is a known difficulty rather than a technicality.

Ordering. The reduced phase space is in general not a cotangent bundle and carries no distinguished coordinates, so there is no preferred operator ordering on it; Example 26.22 is the smallest example, where Equation (26.37) makes the reduced momenta

non-commuting and the ordering ambiguity of Proposition 25.43 unavoidable. Dirac quantization keeps the ambient linear phase space, where Theorem 25.34 applies and the representation is essentially unique, and pays for it in Section 26.7.2.

Inner product. Equation (26.92) typically has no normalizable solutions — the constraint operator has continuous spectrum at zero — so the physical inner product must be constructed rather than inherited, and different constructions give different theories. The reduced route has an inner product from the start but may have no natural measure.

Spectrum. Where both routes can be carried out, they can disagree on the spectrum of an observable. This is Corollary 25.41 appearing in a new place: quantization is not a functor, so a classical isomorphism between two descriptions of one system does not survive to the quantum theory, and the order of the operations matters.

There is no principle in this treatise that decides between them. Where both have been carried out and compared against measurement they agree on what has been measured: for quantum electrodynamics the Coulomb-gauge reduced construction and the covariant construction of Section 26.7.2 give the same scattering amplitudes [1912], and the agreement is checked to the precision of the electron anomalous moment (Chapter 112). That is an experimental fact about one theory, not a theorem about the two procedures, and it should not be read as one.

26.7.2 Weak imposition: the Gupta–Bleuler construction

For quantum electrodynamics in a manifestly covariant gauge the condition Equation (26.92) cannot be imposed as it stands, and the reason is visible already in Proposition 26.32: the primary constraint is $\pi^0 \approx 0$, so a covariant commutator $\left[\hat{A}_\mu(\boldsymbol{x}), \hat{\pi}^\nu(\boldsymbol{y})\right] = \mathrm{i}\hbar\,\delta^\nu_\mu \delta^3(\boldsymbol{x}-\boldsymbol{y})$ contradicts it outright. The covariant route therefore does not impose the constraint at all. Instead it adds to Equation (26.54) a gauge-fixing term

$$\mathcal{L}_{\mathrm{gf}} = -\frac{1}{2\xi\mu_0}\left(\partial_\mu A^\mu\right)^2 , \tag{26.94}$$

which makes $\pi^0 = -\left(\xi\mu_0 c\right)^{-1}\partial_\mu A^\mu$ nonzero. The Lagrangian is then *non*-singular: there are no constraints, all four components of A_μ propagate, and the canonical formalism of Chapter 22 applies unchanged. The price is paid twice over. The time component has a commutator of the opposite sign to the others, so the Fock space carries an indefinite metric and is not a Hilbert space; and the theory has four polarizations where Proposition 26.32 counted two.

The Gupta–Bleuler construction recovers both [953, 958]. The Lorenz condition $\partial_\mu \hat{A}^\mu\,|\psi\rangle = 0$ is too strong — no state in the Fock space satisfies it — so only its positive-frequency part is required,

$$\left(\partial_\mu \hat{A}^\mu\right)^{(+)}|\psi\rangle = 0\,, \tag{26.95}$$

which is the annihilation half and which does have solutions. The subspace $\mathcal{V}_{\mathrm{phys}}$ they span has a positive *semi*-definite norm: the timelike and longitudinal excitations survive only in the combination that has zero norm, because their opposite-sign contributions cancel exactly. Quotienting by that null subspace,

$$\mathcal{H}_{\mathrm{phys}} = \mathcal{V}_{\mathrm{phys}}/\mathcal{V}_0\,, \tag{26.96}$$

gives a genuine Hilbert space in which only the two transverse polarizations remain — Proposition 26.32's count, recovered. Here and in Section 26.7.3 the symbol $\mathcal{H}_{\mathrm{phys}}$ is a

Hilbert space and not a Hamiltonian density, the one place in this chapter where the script letter carries its other standard meaning. Zero-norm states do not contribute to any expectation value, so observables are insensitive to the choice of representative, and the gauge parameter ξ cancels from every physical amplitude, which is the standard check performed in Chapter 100. The construction is developed with the field operators in Chapter 99.

The essential limitation is that Equation (26.95) splits the constraint operator into positive- and negative-frequency parts, and that split is available only for a free field — a linear equation of motion. For a non-abelian theory the constraint algebra Equation (26.72) is nonlinear in the fields, no such split exists, and the Gupta–Bleuler condition has no non-abelian generalization. That failure is what the next subsection exists to repair.

26.7.3 BRST

The systematic replacement of the constraint conditions Equation (26.92) is due to Becchi, Rouet and Stora and, independently, to Tyutin [1564, 1571]. It trades the F separate conditions for a single one, at the price of enlarging the phase space.

Definition 26.50 (The BRST charge)**.** For each first-class constraint γ_A adjoin a Grassmann-odd pair $\left(\eta^A, \mathcal{P}_A\right)$ — the *ghost* and its momentum — canonically conjugate under the graded bracket, $\left\{\eta^A, \mathcal{P}_B\right\} = -\delta^A_B$, and carrying ghost numbers $+1$ and -1. Under quantization the pair becomes an anticommutator, $\left\{\hat{\eta}^A, \hat{\mathcal{P}}_B\right\} = -\mathrm{i}\hbar\,\delta^A_B$; the overall sign is a convention, fixed by the sign chosen for $\mathcal{P}$, and no statement below depends on it. The *BRST charge* is the Grassmann-odd function of ghost number $+1$

$$\Omega = \eta^A \gamma_A - \tfrac{1}{2}\eta^B \eta^A f_{AB}{}^C \mathcal{P}_C + \cdots , \tag{26.97}$$

determined by the requirement that it be *nilpotent*,

$$\{\Omega, \Omega\} = 0 , \qquad \text{quantum mechanically} \quad \hat{\Omega}^2 = 0 , \tag{26.98}$$

which is a nontrivial condition because the bracket of two odd functions is symmetric. The displayed terms suffice when the structure functions of Equation (26.16) are constants; the omitted terms, of higher order in the ghosts, are needed exactly when they are not.

Rests on Equation (26.16) and Postulate 26.47.

That such an Ω exists at all, for every first-class set and not merely for the ones whose algebra happens to close on constants, is the theorem that makes BRST a general method rather than a device.

Theorem 26.51 (Existence and uniqueness of the BRST charge)**.** *Let γ_A be a regular and irreducible set of first-class constraints in the sense of Remark 26.7, with the algebra Equation (26.16). Then the series Equation (26.97) can be continued, order by order in the ghosts, to a Grassmann-odd function Ω of ghost number $+1$ satisfying Equation (26.98) exactly; and any two such functions with the same leading term are related by a canonical transformation of the extended phase space preserving the ghost number. When the structure functions $f_{AB}{}^C$ are constants the series terminates at the two displayed terms.*

Rests on Definition 26.50, Equation (26.16), and Remark 26.7.

Full derivation: Appendix A.54 (p. 3088).

Derives Theorem 26.51.

The construction there is by induction on the ghost number. Nilpotency at each order is an equation of the form $\delta\,\Omega_{(k)} =$ (known, built from lower orders), where δ is the Koszul–Tate differential of the constraint surface; the obstruction to solving it is a class in the homology of δ at positive degree, which vanishes precisely because the constraints are regular

and irreducible — that is what makes the Koszul–Tate complex a resolution. Uniqueness is the same statement one degree down.

The point of Equation (26.98) is that a nilpotent operator has a cohomology, and

$$\mathcal{H}_{\text{phys}} = \frac{\ker \hat{\Omega}}{\operatorname{im} \hat{\Omega}} \quad \text{at ghost number zero} \tag{26.99}$$

reproduces the physical state space of Postulate 26.47: the kernel condition $\hat{\Omega}\,|\psi\rangle = 0$ contains Equation (26.92) at lowest ghost order, and the quotient by the image removes exactly the zero-norm states that Equation (26.96) removed by hand in the abelian case. Gupta–Bleuler is therefore the abelian shadow of Equation (26.99), and the ghosts that appear here as bookkeeping variables are the same objects that Proposition 26.24 produced as the exponentiation of a Jacobian.

Three things make this the practical form in which the content of this chapter enters a calculation. The gauge-fixed action is $S_{\text{gf}} = S + \{\Omega, \Psi\}$ for a gauge-fixing function Ψ of ghost number -1, and nilpotency makes physical amplitudes independent of Ψ — which is why a gauge may be chosen for convenience and why the ξ of Equation (26.94) cancels. The Ward identities of the abelian theory generalize to the Slavnov–Taylor identities, which is what makes the renormalization of Chapter 102 preserve gauge invariance order by order. And unitarity of the physical subspace follows from Equation (26.99) rather than being assumed, so the unphysical polarizations and the ghosts cancel in every cut. The path-integral development, where this is done in practice, is Chapter 106.

26.8 Constraints in field theory

Everything above was written for finitely many degrees of freedom. The passage to a field theory is not automatic, and the places where it is not are recorded here so that Parts VII and XI can refer to them rather than rediscover them.

Constraints are densities. A field-theoretic constraint $\varphi(\boldsymbol{x}) \approx 0$ is one constraint per point of space, so the index j of Equation (26.11) is continuous and the sums become integrals, as set out in Notation 26.28. The counting theorem Theorem 26.17 is applied per point, which is what makes the counts of Propositions 26.32 and 26.34 statements about polarizations.

The bracket matrix is a kernel. $C_{\alpha\beta}$ of Equation (26.24) becomes an integral kernel containing derivatives of delta functions, and "inverting" it means solving a differential equation with boundary conditions. The Dirac bracket of Equation (26.25) is then nonlocal in general, and its locality in particular cases is a result, not an assumption. The Coulomb-gauge brackets of electrodynamics are the standard example: inverting ∇^2 produces the Coulomb kernel, and the instantaneous interaction it describes is an artifact of the gauge, not of the physics.

Regularization. Products of constraint operators at the same point are singular, so Equation (26.93) is a statement about regularized operators, and the anomaly of Remark 26.49 is precisely an obstruction to finding a regulator that preserves the classical algebra.

Boundary terms. On a region with a boundary, or in an asymptotically flat spacetime, the first-class constraints must be supplemented by surface terms for their brackets to be well defined; the surface terms are the total energy, momentum and angular momentum. This is the origin of the ADM mass (Remark 26.45), and it is why in general relativity the energy is a boundary integral rather than a volume integral of a density.

Rank can vary over field space. The constant-rank hypothesis of Definition 26.3 is a statement about a neighbourhood, and in a field theory the neighbourhood is in an infinite-dimensional space. Where it fails — at reducible connections in Yang–Mills, at degenerate metrics in gravity — the constraint structure changes and the counting theorem does not apply. These are measure- zero sets in the classical theory and are ignored throughout this chapter, which is legitimate for the perturbative statements made here and is exactly what fails to be legitimate non-perturbatively (Remark 26.27).

Chapter 27

Central Forces and Statics

Two special regimes of the Newtonian dynamics of Chapter 19 deserve a chapter of their own. The first is the regime in which nothing moves at all: statics, the theory of equilibrium, whose two conditions are the direct specialization of the translational and rotational laws of motion. The second is the regime in which the force on a particle always points along the line to a fixed centre: central motion, which governs planetary orbits and scattering alike, and which reduces — through its conservation laws — to an equivalent problem in a single radial dimension. Standard extended treatments may be found in the classical references [1585, 1945, 2027].

The two regimes are less distant from one another than they look. Both are governed by what a conservation law removes: statics is what remains when the momenta are held at zero, central motion is what remains when the angular momentum has been used to eliminate one coordinate and the energy to eliminate another. In both, the whole of the physics sits in a function of one variable — a potential energy on a configuration space in the first case, an effective potential of the radius in the second.

Remark 27.1 (Two pieces of mathematics used but not owned here). This chapter uses two elementary facts about $\mathbb{R}^3$ that Chapters 7 and 13 do not state, and states them here rather than proving general mathematics inside a physics chapter. The first is the expansion of the double cross product,

$$\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c}) = \boldsymbol{b}\,(\boldsymbol{a} \cdot \boldsymbol{c}) - \boldsymbol{c}\,(\boldsymbol{a} \cdot \boldsymbol{b}) \,, \tag{27.1}$$

used once, in Theorem 27.38. The second is that the area of a circular sector of radius r subtending an angle ϕ is $\frac{1}{2}r^2\phi$, used once, in Proposition 27.17. Both belong to the elementary vector algebra and plane geometry that Part II owes; neither is in dispute, and the reader who wants them proved will find them in any of [1945, 2027].

27.1 Statics

Definition 27.2 (Translational equilibrium)**.** A particle is said to be in *translational equilibrium* if and only if

$$\sum_m \boldsymbol{F}_m = \boldsymbol{0} \,, \tag{27.2}$$

where the $\boldsymbol{F}_m$ are the m forces upon the particle.

Rests on Equation (19.2) *and Definition 19.5.*

Definition 27.3 (Rotational equilibrium)**.** A particle is said to be in *rotational equilibrium* if and only if

$$\sum_n \boldsymbol{\tau}_n = \boldsymbol{0} \,, \tag{27.3}$$

where the $\boldsymbol{\tau}_n$ are the n torques upon the particle.

Rests on Equation (19.14) *and Definition 19.18.*

Under these conditions, Equations (19.2) and (19.14) state that the linear momentum and the angular momentum of the particle are constant: equilibrium is the regime in which Newton's second laws, translational and rotational, reduce to conservation statements.

A torque is defined about a chosen point, so the second condition looks at first sight to depend on that choice. It does not, provided the first condition already holds — and that is what makes the pair of conditions usable at all.

Proposition 27.4 (Torque balance is independent of the origin)**.** *Let a system of forces $\boldsymbol{F}_m$ act at the points $\boldsymbol{x}_m$ of a body, and suppose Equation (27.2) holds. Then the total torque $\boldsymbol{\tau}_O = \sum_m \boldsymbol{x}_m \times \boldsymbol{F}_m$ takes the same value about every point O. Consequently, if the torque balance Equation (27.3) holds about one point, it holds about all of them.*

Rests on Equations (19.15) and (27.2).

Proof. *Derives Proposition 27.4.*

Let O' be displaced from O by $\boldsymbol{a}$, so that the position of the m-th point of application measured from O' is $\boldsymbol{x}_m - \boldsymbol{a}$. By Equation (19.15) the total torque about O' is

$$\boldsymbol{\tau}_{O'} = \sum_m (\boldsymbol{x}_m - \boldsymbol{a}) \times \boldsymbol{F}_m = \boldsymbol{\tau}_O - \boldsymbol{a} \times \sum_m \boldsymbol{F}_m = \boldsymbol{\tau}_O \,, \tag{27.4}$$

the last step by Equation (27.2). The displacement $\boldsymbol{a}$ was arbitrary. □

The proposition is sharper than it looks, and its converse is false in a way worth stating: when the net force does *not* vanish, the torque does depend on the reference point, and one may always slide O' along the line perpendicular to $\sum \boldsymbol{F}_m$ until the torque about it vanishes. That is why a single unbalanced force is described as acting along a definite *line of action* rather than at a definite point.

Phenomenon 27.5 (Equilibrium requires two conditions, not one)**.** A body stays at rest only if the forces upon it balance *and* their turning effects balance; the two requirements are independent, and Nature displays the difference plainly. Two equal and opposite forces applied at different points of a rigid body — a couple, as when both hands turn a steering wheel — leave the centre of mass exactly where it was and set the body spinning. Conversely a single force whose line of action passes through the centre of mass translates the body without turning it at all. Every static structure, from a balanced beam to a ladder leaning against a wall, stands by the simultaneous satisfaction of both conditions, and fails when either is violated [14].

Rests on Equations (27.2) and (27.3).

Derivation. *Derives Phenomenon 27.5.*

Apply forces $\boldsymbol{F}$ and $-\boldsymbol{F}$ at the points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of a rigid body. Equation (27.2) is satisfied, since the two cancel, so by the second law of Chapter 19 the centre of mass keeps whatever velocity it had. The total torque about the origin, however, is

$$\boldsymbol{\tau} = \boldsymbol{x}_1 \times \boldsymbol{F} + \boldsymbol{x}_2 \times (-\boldsymbol{F}) = (\boldsymbol{x}_1 - \boldsymbol{x}_2) \times \boldsymbol{F} \,, \tag{27.5}$$

which is independent of the choice of origin — a couple has the same moment about every point, which is Proposition 27.4 for the special case of a vanishing resultant — and which vanishes only if $\boldsymbol{F}$ is parallel to the line joining the two points of application. Whenever it does not vanish, Equation (27.3) fails, the angular momentum grows, and the body turns although no net force acts upon it. For the converse case, a single force applied at the centre of mass has $\boldsymbol{x} = \boldsymbol{0}$ measured from there, hence zero moment about it, and the body accelerates without rotating. The two conditions are therefore logically independent, and neither implies the other. □

27.1.1 The centre of gravity

Weight acts on every part of a body, so the gravitational contribution to Equation (27.3) is a sum over the whole mass distribution. In a uniform field that sum collapses to a single term.

Definition 27.6 (Centre of gravity)**.** The *centre of gravity* of a body in a gravitational field is the point $\boldsymbol{X}_g$ such that the total gravitational torque about any point equals the torque of the total weight applied at $\boldsymbol{X}_g$ alone.
Rests on Definition 19.47 and Equation (19.15).

Proposition 27.7 (Centre of gravity in a uniform field)**.** *If the gravitational field $\boldsymbol{g}$ is the same at every point of the body, then $\boldsymbol{X}_g = \boldsymbol{X}$, the centre of mass Equation* (19.38)*, and the total gravitational torque about the origin is*

$$\boldsymbol{\tau}_g = \boldsymbol{X} \times M\boldsymbol{g}\,, \tag{27.6}$$

with M the total mass Equation (19.32)*.*
Rests on Equations (19.15)*,* (19.32) *and* (19.38)*.*

Proof. *Derives Proposition 27.7.*
The weight of the m-th element is $m_m\boldsymbol{g}$ with one and the same $\boldsymbol{g}$, so by Equation (19.15)

$$\boldsymbol{\tau}_g = \sum_m \boldsymbol{x}_m \times m_m\boldsymbol{g} = \left(\sum_m m_m\boldsymbol{x}_m\right) \times \boldsymbol{g} = \boldsymbol{X} \times M\boldsymbol{g}\,, \tag{27.7}$$

using Equation (19.38) in the form $\sum_m m_m\boldsymbol{x}_m = M\boldsymbol{X}$. The right-hand side is the torque of a single force $M\boldsymbol{g}$ applied at $\boldsymbol{X}$, which is Definition 27.6. □

The uniformity hypothesis is not idle. Over a body of vertical extent h near the Earth's surface the field varies by a relative amount $2h/R_\oplus$, about 3×10^{-7} per metre; the resulting separation of the centre of gravity from the centre of mass is what raises the tides and what makes a satellite in orbit settle with its long axis pointing at the planet. For a laboratory body it is utterly negligible, and Proposition 27.7 is exact for every purpose of this section.

Phenomenon 27.8 (The centre of gravity hangs below the support)**.** A flat plate hung freely from a pin comes to rest, after its oscillations have died away, in one definite orientation, and the vertical line through the pin always passes through the same material point of the plate whichever pin is used. Suspending the plate from two different points and marking the two verticals locates that point as their intersection; a third suspension confirms it. This is how the centre of gravity of an irregular body is measured, and it is the principle of every balance and every plumb line [14].
Rests on Equations (27.3) *and* (27.6)*.*

Derivation. *Derives Phenomenon 27.8.*
Take torques about the point of support S, so that the unknown reaction there contributes nothing. The only other force is the weight, whose torque is $(\boldsymbol{X} - \boldsymbol{x}_S) \times M\boldsymbol{g}$ by Equation (27.6). Rotational equilibrium Equation (27.3) therefore demands

$$(\boldsymbol{X} - \boldsymbol{x}_S) \times M\boldsymbol{g} = \boldsymbol{0}\,, \tag{27.8}$$

that is, that $\boldsymbol{X} - \boldsymbol{x}_S$ be parallel to $\boldsymbol{g}$: the centre of gravity lies on the vertical through the support. Two positions satisfy this, one with $\boldsymbol{X}$ below S and one with it above; only the first is stable, by Theorem 27.10, since rotating the body about the pin from that position raises the centre of gravity and so raises the potential energy MgZ, Z being its height, whereas rotating it from the second position lowers Z. Repeating the suspension from a second point gives a second vertical through the same material point, and two non-parallel lines meet in one point. □

27.1.2 Equilibrium as a stationary potential energy

Written out force by force, the equilibrium of a mechanism with many joints is a large system of equations in which most of the unknowns are constraint forces nobody wants to know. The principle of virtual work Equation (21.17) removes them all at once, because a constraint force does no virtual work.

Theorem 27.9 (Equilibrium is stationarity of the potential energy)**.** *Consider a system with holonomic, time-independent constraints, described by generalized coordinates q^a, $a = 1, \dots, n$, whose active forces derive from a potential energy $U(q)$. Then a configuration q_0 is an equilibrium configuration if and only if*

$$\left.\frac{\partial U}{\partial q^a}\right|_{q_0} = 0\,, \qquad a = 1, \dots, n\,. \tag{27.9}$$

Rests on Equations (21.17) *and* (21.31) *and Definition 21.2.*

Proof. *Derives Theorem 27.9.*

By Theorem 21.18 the system is in equilibrium if and only if the virtual work of the active forces vanishes for every virtual displacement compatible with the constraints. Because the constraints are holonomic and time-independent, a compatible virtual displacement is exactly one of the form $\delta \boldsymbol{x}^s = (\partial \boldsymbol{x}^s / \partial q^a)\, \delta q^a$ with the δq^a arbitrary and independent, so that by Equation (21.31) the virtual work is

$$\delta W = \sum_s \boldsymbol{F}^s \cdot \delta \boldsymbol{x}^s = Q_a\, \delta q^a\,, \qquad Q_a = \sum_s \boldsymbol{F}^s \cdot \frac{\partial \boldsymbol{x}^s}{\partial q^a}\,. \tag{27.10}$$

For forces derived from a potential, $\boldsymbol{F}^s = -\nabla_s U$ and the chain rule (Proposition 7.72) gives $Q_a = -\partial U / \partial q^a$. Vanishing of $Q_a \delta q^a$ for arbitrary independent δq^a is vanishing of every Q_a, which is Equation (27.9). □

Equilibrium is thus a stationarity condition, and the character of the stationary point decides whether the equilibrium can be observed at all.

Theorem 27.10 (Lagrange–Dirichlet)**.** *If U has a strict local minimum at the equilibrium configuration q_0, then q_0 is stable: for every neighbourhood N of $(q_0, 0)$ in the space of positions and velocities there is a smaller one from which the motion never leaves N.*

Rests on Equations (19.21) *and* (27.9) *and Theorem 21.43.*

Proof. *Derives Theorem 27.10.*

Set $U(q_0) = 0$ without loss of generality. The kinetic energy T of a system with time-independent constraints is a positive definite quadratic form in the generalized velocities (Definition 21.66), so $T \geq 0$ with equality only at rest, and the conserved energy $E = T + U$ of Theorem 21.43 is a positive definite function of $(q - q_0, \dot{q})$ near the equilibrium, because U has a strict minimum there. Given a neighbourhood N, choose $\varepsilon > 0$ smaller than the least value of U on the boundary of a ball inside N; then any motion starting with $E < \varepsilon$ can never reach that boundary, since doing so would require $U \geq \varepsilon$ and hence $E \geq \varepsilon$. The set $\{E < \varepsilon\}$ is the required smaller neighbourhood. □

Example 27.11 (The ladder against a wall)**.** A uniform ladder of mass $M = 12.0\,\mathrm{kg}$ and length $\ell = 4.00\,\mathrm{m}$ leans at $\vartheta = 65.0°$ to the horizontal, its top against a smooth vertical wall and its foot on a floor that can supply friction. Three unknown constraint forces act: the wall's normal N_w (horizontal), the floor's normal N_f (vertical) and the floor's friction f (horizontal). Taking torques about the foot, where two of the three act and contribute nothing,

$$N_w\, \ell \sin\vartheta = Mg\, \frac{\ell}{2} \cos\vartheta \qquad \Longrightarrow \qquad N_w = \frac{Mg}{2 \tan\vartheta}\,, \tag{27.11}$$

From Equations (27.3) and (27.6) (torques about the foot, the weight acting at the midpoint by Proposition 27.7).

while Equation (27.2) resolved horizontally and vertically gives $f = N_w$ and $N_f = Mg$. With $g = 9.806\,65\,\mathrm{m/s^2}$ the weight is $Mg = 117.7\,\mathrm{N}$, so $N_f = 117.7\,\mathrm{N}$ and $N_w = f = 27.4\,\mathrm{N}$. The ratio the floor is required to supply is

$$\frac{f}{N_f} = \frac{1}{2\tan\vartheta} = 0.233\,, \tag{27.12}$$

independent of both the mass and the length — which is why the question "will it slip?" is answered by the angle alone. Whether the floor can supply it is a question about the surfaces and not about mechanics: the usual empirical bound is $f \le \mu_s N_f$ with a coefficient μ_s that for dry wood on concrete is quoted at a few tenths, comfortably above 0.233. This treatise carries no primary source for tabulated friction coefficients, and none of the mechanics above depends on one: μ_s enters only as an inequality on a constraint force, never as a law of motion.

Rests on Equations (27.2) and (27.3) and Proposition 27.7.

Remark 27.12 (Statically indeterminate systems). Equations (27.2) and (27.3) provide six scalar equations in three dimensions, and a rigid body resting on three supports has exactly three unknown normal reactions: the problem closes. A body on four supports has four, and the equations no longer determine them — which is why a four-legged table rocks and a three-legged one never does. Nothing is wrong with the mechanics; the hypothesis of perfect rigidity has simply been pushed past what it can decide, and the missing equations come from the deformation of the body under load. They are supplied by the theory of elasticity, Chapter 30, and the indeterminacy is resolved by the constitutive relation there rather than by any further principle of statics.

27.2 Central forces

Definition 27.13 (Central force)**.** A *central force* is an active force upon a particle that can be written in the form

$$\boldsymbol{F} = f(r)\,\hat{\boldsymbol{r}}\,. \tag{27.13}$$

Rests on Definitions 19.30 and 21.15.

Definition 27.14 (Central motion)**.** A *central motion* is the motion of a particle due to a central force.

Rests on Definition 27.13.

Everything in this section rests on one kinematic decomposition, which is worth isolating before it is used.

Lemma 27.15 (Velocity and acceleration in plane polar coordinates)**.** *Let a particle move in a fixed plane, and write its position as* $\boldsymbol{x} = r\,\hat{\boldsymbol{r}}$ *with* $\hat{\boldsymbol{r}} = (\cos\phi, \sin\phi)$ *and* $\hat{\boldsymbol{\phi}} = (-\sin\phi, \cos\phi)$*. Then*

$$\dot{\hat{\boldsymbol{r}}} = \dot\phi\,\hat{\boldsymbol{\phi}}\,, \qquad \dot{\hat{\boldsymbol{\phi}}} = -\dot\phi\,\hat{\boldsymbol{r}}\,, \tag{27.14}$$

$$\boldsymbol{v} = \dot r\,\hat{\boldsymbol{r}} + r\dot\phi\,\hat{\boldsymbol{\phi}}\,, \tag{27.15}$$

$$\boldsymbol{a} = \underbrace{\left(\ddot r - r\dot\phi^2\right)}_{a_r}\hat{\boldsymbol{r}} + \underbrace{\left(r\ddot\phi + 2\dot r\dot\phi\right)}_{a_\phi}\hat{\boldsymbol{\phi}}\,. \tag{27.16}$$

Rests on Definitions 18.10 and 18.11 and Proposition 7.31.

Proof. *Derives Lemma 27.15.*

Differentiating the two unit vectors with respect to time by the chain rule (Proposition 7.31), $\dot{\hat{\boldsymbol{r}}} = \dot{\phi}\,(-\sin\phi, \cos\phi) = \dot{\phi}\,\hat{\boldsymbol{\phi}}$ and $\dot{\hat{\boldsymbol{\phi}}} = \dot{\phi}\,(-\cos\phi, -\sin\phi) = -\dot{\phi}\,\hat{\boldsymbol{r}}$, which is Equation (27.14). Then $\boldsymbol{v} = \frac{\mathrm{d}}{\mathrm{d}t}\,(r\hat{\boldsymbol{r}}) = \dot{r}\hat{\boldsymbol{r}} + r\dot{\hat{\boldsymbol{r}}}$, giving Equation (27.15), and differentiating once more,

$$\boldsymbol{a} = \ddot{r}\hat{\boldsymbol{r}} + \dot{r}\dot{\phi}\hat{\boldsymbol{\phi}} + \left(\dot{r}\dot{\phi} + r\ddot{\phi}\right)\hat{\boldsymbol{\phi}} + r\dot{\phi}\,\dot{\hat{\boldsymbol{\phi}}}\,,$$

which on substituting $\dot{\hat{\boldsymbol{\phi}}}$ and collecting terms is Equation (27.16). □

The two "extra" terms in Equation (27.16) are not forces. They are what the derivative of a moving orthonormal frame contributes, and the same quantities appear in Example 21.69 as Christoffel symbols of the plane in polar labelling: Equation (21.78), the free-particle equations there, are exactly $a_r = 0$ and $a_\phi = 0$. That the plane is flat while those symbols are non-zero is the point made in Remark 21.70.

27.2.1 Areal velocity

Definition 27.16 (Areal velocity)**.** The *areal velocity* is the rate of change of the area swept out by the line from the centre to the particle, with respect to an inertial frame of reference,

$$v_a := \frac{\mathrm{d}A}{\mathrm{d}t}\,. \tag{27.17}$$

Rests on Definitions 18.8 and 27.14.

Proposition 27.17 (Areal velocity in polar coordinates)**.** *For a plane motion described in polar coordinates,*

$$v_a = \frac{\mathrm{d}A}{\mathrm{d}t} = \frac{1}{2}r^2\dot{\phi}\,, \tag{27.18}$$

and the transverse component of the acceleration is

$$a_\phi = \frac{2}{r}\frac{\mathrm{d}v_a}{\mathrm{d}t}\,. \tag{27.19}$$

Rests on Equations (27.16) and (27.17).

Proof. *Derives Proposition 27.17.*

Between the times t and $t + \mathrm{d}t$ the radius turns through $\mathrm{d}\phi = \dot{\phi}\,\mathrm{d}t$ and lengthens by $\mathrm{d}r = \dot{r}\,\mathrm{d}t$. The region swept out contains the circular sector of radius r and angle $\mathrm{d}\phi$, and is contained in the sector of the same angle and radius $r + |\mathrm{d}r|$; by the sector area recorded in Remark 27.1 these bound it as

$$\frac{1}{2}r^2\,\mathrm{d}\phi \le \mathrm{d}A \le \frac{1}{2}\,(r + |\mathrm{d}r|)^2\,\mathrm{d}\phi\,,$$

so the two differ by $O(\mathrm{d}t^2)$ and $\mathrm{d}A = \frac{1}{2}r^2\mathrm{d}\phi + O(\mathrm{d}t^2)$. Dividing by $\mathrm{d}t$ and letting $\mathrm{d}t \to 0$ gives Equation (27.18). Differentiating that result,

$$\frac{\mathrm{d}v_a}{\mathrm{d}t} = \frac{1}{2}\left(2r\dot{r}\dot{\phi} + r^2\ddot{\phi}\right) = \frac{r}{2}\left(2\dot{r}\dot{\phi} + r\ddot{\phi}\right) = \frac{r}{2}\,a_\phi\,,$$

by Equation (27.16), which rearranged is Equation (27.19). □

27.2.2 Conservation laws of central motion

Proposition 27.18 (Conservation of angular momentum)**.** *In a central motion the angular momentum $\boldsymbol{L}$ is constant, the motion is confined to the fixed plane through the centre perpendicular to $\boldsymbol{L}$, and in polar coordinates on that plane the angular momentum of the plane central motion has a component only along the z-axis, of magnitude*

$$L = mr^2\dot{\phi}\,. \tag{27.20}$$

Rests on Equations (19.14)*,* (19.15) *and* (27.13)*.*

Proof. *Derives Proposition 27.18.*

By Equation (19.14) the rate of change of the angular momentum is the torque, and by Equation (19.15) the torque is $\boldsymbol{x}\times\boldsymbol{F}$. For a central force Equation (27.13) the force is parallel to $\boldsymbol{x} = r\hat{\boldsymbol{r}}$, so

$$\frac{\mathrm{d}\boldsymbol{L}}{\mathrm{d}t} = \boldsymbol{x}\times\boldsymbol{F} = rf(r)\,\hat{\boldsymbol{r}}\times\hat{\boldsymbol{r}} = \boldsymbol{0}\,, \tag{27.21}$$

and $\boldsymbol{L}$ is a fixed vector. Since $\boldsymbol{L} = \boldsymbol{x}\times m\boldsymbol{v}$ is orthogonal to $\boldsymbol{x}$ by construction, the position stays in the fixed plane through the centre perpendicular to $\boldsymbol{L}$; choosing Cartesian axes with $\boldsymbol{L}$ along z makes that the xy-plane, so $\boldsymbol{L} = L\hat{\boldsymbol{z}}$ with no other component. Finally, by Equation (27.15) the radial part of $\boldsymbol{v}$ is parallel to $\boldsymbol{x}$ and contributes nothing to the cross product, leaving $\boldsymbol{L} = r\hat{\boldsymbol{r}}\times mr\dot{\phi}\hat{\boldsymbol{\phi}} = mr^2\dot{\phi}\,\hat{\boldsymbol{z}}$, which is Equation (27.20). The degenerate case $\boldsymbol{L} = \boldsymbol{0}$ is the rectilinear motion through the centre; the plane is then not unique, and any plane containing the line will serve. □

Phenomenon 27.19 (Kepler's second law: equal areas in equal times)**.** The line drawn from a planet to the Sun sweeps out equal areas in equal times: the planet runs fastest at perihelion and slowest at aphelion, in exactly the proportion that keeps the rate of sweeping constant. The regularity was extracted from Tycho Brahe's naked-eye positions of Mars and holds for every body bound to a centre — planets, comets, the components of a visual binary, a charged particle deflected by a nucleus. Newton showed that it holds if and only if the force is directed at the centre, and made it the first proposition of the *Principia* [14].

Rests on Proposition 27.18 and Equation (27.18)*.*

Derivation. *Derives Phenomenon 27.19.*

By Proposition 27.18 the motion is plane and $L = mr^2\dot{\phi}$ is constant. Comparing with the areal velocity Equation (27.18),

$$v_a = \frac{1}{2}r^2\dot{\phi} = \frac{L}{2m}\,, \tag{27.22}$$

a constant — the equal-area law, and the relation Equation (27.27) read from left to right. Mark what the argument did not use: the form of $f(r)$ never entered. Kepler's second law therefore tests the *centrality* of the force and nothing else, and would hold equally for an inverse-cube or a linear attraction.

The converse also holds, and is what makes the law evidence for anything. If v_a is constant then by Equation (27.19) the transverse acceleration a_ϕ vanishes identically, so by Equation (19.4) the force has no transverse component and is central. Equal areas in equal times and centrality of the force are therefore the same statement. □

Proposition 27.20 (Conservation of energy)**.** *A central force is conservative: with*

$$V(r) := -\int_{r_0}^{r} f(s)\,\mathrm{d}s \tag{27.23}$$

for any fixed $r_0 > 0$, one has

$$\boldsymbol{F} = -\boldsymbol{\nabla}V\,, \tag{27.24}$$

and consequently

$$\boldsymbol{\nabla} \times \boldsymbol{F} = \boldsymbol{0}\,. \tag{27.25}$$

The energy $E = T + V$ of a central motion is therefore constant.
Rests on Equations (19.19) and (27.13) and Definition 19.31.

Proof. *Derives Proposition 27.20.*

The gradient of the radius is $\boldsymbol{\nabla} r = \hat{\boldsymbol{r}}$, since $r = \left(x^2 + y^2 + z^2\right)^{1/2}$ gives $\partial r / \partial x^i = x^i / r$. Hence, by the chain rule (Proposition 7.72) and the fundamental theorem of calculus (Theorem 7.42) applied to Equation (27.23),

$$-\boldsymbol{\nabla} V = -V'(r)\,\hat{\boldsymbol{r}} = f(r)\,\hat{\boldsymbol{r}} = \boldsymbol{F}\,,$$

which is Equation (27.24); Equation (27.25) then follows because the curl of a gradient vanishes, Equation (19.20). Note that no topological hypothesis was needed: the general criterion Proposition 7.104 requires a simply connected domain because it must reconstruct the potential from line integrals, whereas here Equation (27.23) writes it down.

For the conservation, differentiate $E = \frac{1}{2} m \left|\boldsymbol{v}\right|^2 + V$ along the motion and use Equation (19.4) together with Equation (27.24):

$$\frac{\mathrm{d}E}{\mathrm{d}t} = m\boldsymbol{v} \cdot \boldsymbol{a} + \boldsymbol{\nabla} V \cdot \boldsymbol{v} = \boldsymbol{v} \cdot (\boldsymbol{F} - \boldsymbol{F}) = 0\,. \tag{27.26}$$

From Equations (19.4) and (27.24) (the chain rule on $V(\boldsymbol{x}(t))$ and the second law on the kinetic term).

Equivalently, the work Equation (19.25) done between two points depends only on the radii of the endpoints. □

Proposition 27.21 (Conservation of the areal speed)**.** *In a central motion the transverse component of the acceleration is zero, and the areal speed is therefore constant.*
Rests on Equations (27.13) and (27.19).

Proof. *Derives Proposition 27.21.*

A central force Equation (27.13) has no $\hat{\boldsymbol{\phi}}$-component, so by Equation (19.4) neither does the acceleration: $a_\phi = 0$. Equation (27.19) then gives $\mathrm{d}v_a/\mathrm{d}t = \frac{1}{2} r\, a_\phi = 0$. This is the same fact as Equation (27.22), obtained without passing through the angular momentum, and the equality of the two routes is Proposition 27.22. □

Proposition 27.22 (Relation between the constants of the motion)**.** *For a central motion one has*

$$r^2 \dot{\phi} = \frac{L}{m} = 2v_a\,. \tag{27.27}$$

Rests on Equations (27.18) and (27.20).

Proof. *Derives Proposition 27.22.*

The first equality is Equation (27.20) divided by m, and the second is twice Equation (27.18). The content of the chain is that the two conserved quantities of the previous two propositions are not independent: the areal speed is the angular momentum per unit mass, halved. Kepler's second law and the conservation of angular momentum are one law counted twice, and the constant L/m is the one that appears in every formula below. □

27.2.3 The Binet equation

Proposition 27.23 (Binet equation)**.** *In a central motion, the trajectory equation can be found from the following equation:*

$$a_r = -\frac{L^2}{m^2 r^2}\left(\frac{\mathrm{d}^2}{\mathrm{d}\phi^2}\left(\frac{1}{r}\right) + \frac{1}{r}\right)\,. \tag{27.28}$$

Rests on Equations (27.16) and (27.20).

Proof. *Derives Proposition 27.23.*

The device is to use ϕ rather than t as the independent variable, which is legitimate whenever $L \neq 0$, since then $\dot{\phi} = L/(mr^2)$ never vanishes and ϕ is a strictly monotone function of t. Write $u := 1/r$, so that

$$\dot{\phi} = \frac{L}{mr^2} = \frac{Lu^2}{m}\,. \tag{27.29}$$

Then, by the chain rule (Proposition 7.31),

$$\dot{r} = \frac{\mathrm{d}r}{\mathrm{d}\phi}\,\dot{\phi} = -\frac{1}{u^2}\frac{\mathrm{d}u}{\mathrm{d}\phi}\cdot\frac{Lu^2}{m} = -\frac{L}{m}\frac{\mathrm{d}u}{\mathrm{d}\phi}\,,$$

a first simplification worth noticing on its own: the radial velocity is $-L/m$ times $\mathrm{d}u/\mathrm{d}\phi$, with a constant factor. Differentiating once more in the same way,

$$\ddot{r} = -\frac{L}{m}\frac{\mathrm{d}^2u}{\mathrm{d}\phi^2}\,\dot{\phi} = -\frac{L^2u^2}{m^2}\frac{\mathrm{d}^2u}{\mathrm{d}\phi^2}\,,$$

while the centrifugal term of Equation (27.16) is $r\dot{\phi}^2 = u^{-1}L^2u^4/m^2 = L^2u^3/m^2$. Subtracting,

$$a_r = \ddot{r} - r\dot{\phi}^2 = -\frac{L^2u^2}{m^2}\left(\frac{\mathrm{d}^2u}{\mathrm{d}\phi^2} + u\right)\,, \tag{27.30}$$

From Equations (27.16) and (27.29) (eliminating t in favour of ϕ through $\dot{\phi} = Lu^2/m$ and writing $u = 1/r$).

which is Equation (27.28) with u restored to $1/r$. Combined with $ma_r = f(r)$ from Equation (19.4), it turns the dynamical problem into a single second-order equation for the shape of the orbit, with the time eliminated altogether. □

27.3 The equivalent one-dimensional problem

27.3.1 Energy of the system

Proposition 27.24 (Radial form of the energy)**.** *The energy of a particle in central motion is*

$$E = \frac{1}{2}m\dot{r}^2 + \frac{L^2}{2mr^2} + V(r)\,. \tag{27.31}$$

Rests on Equations (27.15), (27.20) and (27.24).

Proof. *Derives Proposition 27.24.*

The motion is plane by Proposition 27.18, and the two terms of Equation (27.15) are orthogonal, so

$$|\boldsymbol{v}|^2 = \dot{r}^2 + r^2\dot{\phi}^2$$

and the kinetic energy Equation (19.18) splits into a radial and a transverse part, $T = \frac{1}{2}m\dot{r}^2 + \frac{1}{2}mr^2\dot{\phi}^2$. Eliminating $\dot{\phi}$ with Equation (27.20),

$$\frac{1}{2}mr^2\dot{\phi}^2 = \frac{1}{2}mr^2\left(\frac{L}{mr^2}\right)^2 = \frac{L^2}{2mr^2}\,,$$

which added to the radial part and to the potential Equation (27.24) gives Equation (27.31). The constancy of E is Proposition 27.20, and the constancy of L is what makes the middle term a function of r alone. □

27.3.2 Effective potential

Definition 27.25 (Effective potential)**.** The *effective potential* of a central motion is

$$V_e = \frac{L^2}{2mr^2} + V(r)\,, \tag{27.32}$$

so that the energy Equation (27.31) takes the one-dimensional form

$$E = \frac{1}{2}m\dot{r}^2 + V_e\,. \tag{27.33}$$

Rests on Equation (27.31).

The reduction is exact, not approximate, and it is the whole yield of the two conservation laws: a motion with three degrees of freedom has been replaced by a motion with one, at the cost of a term in the potential that carries the value of the eliminated angular momentum. The added term $L^2/\left(2mr^2\right)$ is repulsive for every $L \neq 0$, and it is a piece of kinetic energy wearing the mask of a potential — which is why calling it a "centrifugal force" is harmless bookkeeping in the radial equation and false in any other.

27.3.3 Turning points

Considering the radial velocity obtained from Equation (27.33),

$$\dot{r} = \pm\sqrt{\frac{2}{m}\left(E - V_e\right)}\,, \tag{27.34}$$

one must have

$$E \geq V_e\,. \tag{27.35}$$

In the case in which $E = V_e$, one has

$$\dot{r} = 0 \implies r = R\,,$$

with R constant. Therefore a circular potential barrier of radius R is generated, called a *turning point*, which confines the motion.

Phenomenon 27.26 (Bound orbits oscillate between two radii)**.** A body bound to a centre does not spiral into it, however strong the attraction, unless its angular momentum is exactly zero: it approaches to a least distance, turns, and recedes to a greatest one, repeating the excursion indefinitely. Comets return; satellites in eccentric orbits alternate between perigee and apogee; a spacecraft aimed at a planet with any transverse velocity at all swings past it instead of falling in. What turns the body back is not an agency outside it but the angular momentum it already carries, and the same barrier explains why a star cannot collapse merely by attracting itself.

Rests on Equations (27.32) *and* (27.33).

Derivation. *Derives Phenomenon 27.26.*

By Equation (27.33) the radial coordinate obeys a one-dimensional energy equation, $E = \frac{1}{2}m\dot{r}^2 + V_e$, with the effective potential Equation (27.32). Since $\frac{1}{2}m\dot{r}^2 \geq 0$, the motion is confined to the set where $E \geq V_e(r)$, which is Equation (27.35), and the boundary points of that set, where $E = V_e(R)$, are the turning radii at which $\dot{r}$ vanishes and changes sign. Now suppose $L \neq 0$. The centrifugal term $L^2/(2mr^2)$ grows like r^{-2} as $r \to 0$ and so dominates every attractive potential less singular than r^{-2} — the inverse-square attraction among them, whose potential goes only as r^{-1}. Hence $V_e \to +\infty$ at the centre, and the condition $E \geq V_e$ excludes a whole neighbourhood of $r = 0$: the body cannot reach the centre at any finite energy. If instead $L = 0$ the centrifugal term is absent, $V_e = V$, and nothing forbids

the collision — which shows that the barrier is angular momentum and not force. For an attraction that vanishes at infinity, a state with $E < 0$ is trapped between an inner and an outer turning radius and the orbit is bounded; a state with $E \geq 0$ has only an inner one and the body escapes. That the bounded orbit should also *close* is a further and much stronger property, discussed in Phenomenon 27.32 and Theorem 27.41. □

Example 27.27 (Escape from the Earth)**.** Take the Newtonian attraction $V(r) = -GMm/r$, which vanishes at infinity, so that the criterion of the previous derivation reads $E \geq 0$ for escape. A body launched radially ($L = 0$) from the surface $r = R_\oplus$ with speed v has $E = \tfrac{1}{2}mv^2 - GMm/R_\oplus$, and escapes when

$$v \geq v_{\text{esc}} = \sqrt{\frac{2GM}{R_\oplus}}\,. \tag{27.36}$$

The product GM for the Earth is $3.986 \times 10^{14}\,\text{m}^3/\text{s}^2$, obtained from $G = 6.674\,30 \times 10^{-11}\,\text{m}^3/(\text{kg}\,\text{s}^2)$ [2392] and $M = 5.972 \times 10^{24}\,\text{kg}$; with $R_\oplus = 6.371 \times 10^6\,\text{m}$ this gives $v_{\text{esc}} = 1.119 \times 10^4\,\text{m/s}$, that is $11.19\,\text{km/s}$. Two remarks are worth more than the number. First, the escape speed does not depend on the mass of the body or on the direction of launch — only on the sign of E, which is why it is a speed and not a velocity. Second, what enters is the product GM and never the two factors separately, and that is not a notational accident: spacecraft ranging fixes GM for the Earth far more sharply than the laboratory knows G, whose relative uncertainty is 2.2×10^{-5} (Phenomenon 33.16). The mass of the Earth is therefore obtained as GM/G and inherits that uncertainty entire; it is the least well known of the three quantities in Equation (27.36), and celestial mechanics never measures it at all.

Rests on Equations (27.31) *and* (27.35).

27.3.4 The radial quadrature

The one-dimensional problem can be integrated once and for all, which is what makes it worth constructing.

Proposition 27.28 (Quadratures of the central problem)**.** *For a central motion with $L \neq 0$, the time along the orbit and the shape of the orbit are given by*

$$t - t_0 = \int_{r_0}^{r} \frac{\mathrm{d}s}{\sqrt{\dfrac{2}{m}\,(E - V_e(s))}}\,, \tag{27.37}$$

$$\phi - \phi_0 = \int_{r_0}^{r} \frac{L\,\mathrm{d}s}{s^2\sqrt{2m\,(E - V(s)) - \dfrac{L^2}{s^2}}}\,, \tag{27.38}$$

on any interval between successive turning points, the sign of the root being that of $\dot r$ there.

Rests on Equations (27.20) *and* (27.34).

Proof. *Derives Proposition 27.28.*

Equation (27.34) is a separable first-order equation for $r(t)$: dividing by the root and integrating gives Equation (27.37), which is legitimate on any interval where $E > V_e$, that is strictly between turning points, since there the integrand is continuous. For the orbit, divide $\mathrm{d}\phi = \dot\phi\,\mathrm{d}t$ by $\mathrm{d}r = \dot r\,\mathrm{d}t$ and use Equation (27.20) and Equation (27.34):

$$\frac{\mathrm{d}\phi}{\mathrm{d}r} = \frac{\dot\phi}{\dot r} = \frac{L/\,(mr^2)}{\sqrt{(2/m)\,(E - V_e)}} = \frac{L}{r^2\sqrt{2m\,(E - V) - L^2/r^2}}\,,$$

where the last step multiplied numerator and denominator by m and used Equation (27.32). Integrating gives Equation (27.38). □

Definition 27.29 (Apsidal angle)**.** For a bounded orbit with turning radii $r_- < r_+$, the *apsidal angle* is the angle swept between successive turning points,

$$\Phi = \int_{r_-}^{r_+} \frac{L\,\mathrm{d}s}{s^2\sqrt{2m\,(E - V(s)) - \dfrac{L^2}{s^2}}}\,. \tag{27.39}$$

The orbit closes after p radial oscillations and q revolutions if and only if $2\Phi = 2\pi q/p$ with p, q integers — that is, if and only if Φ/π is rational.

Rests on Equation (27.38)*.*

The radial motion is always periodic for a bounded orbit, since r returns to r_- after each excursion; what is not automatic is that the *angle* should also return. The orbit closes only when the two periods are commensurable, and Φ generically depends on E and L, so that a family of orbits under one and the same force law will in general contain no closed orbit at all. Two force laws are the exception, and the fact that Nature's dominant one is among them is the subject of Section 27.4.5.

27.4 The two-body problem and Kepler orbits

Nothing so far has asked what holds the centre of force in place. In the Solar System nothing does: the Sun is attracted by each planet exactly as strongly as it attracts it, by Postulate 19.48, and moves in response. The first business of this section is to show that the two-body problem is nevertheless exactly equivalent to a one-body central problem of the kind already solved — not approximately, and not only when one mass is large.

27.4.1 Reduction to one body

Theorem 27.30 (Reduction of the two-body problem)**.** *Let two particles of masses m_1, m_2 interact through a force that depends only on their separation and acts along the line joining them,*

$$\boldsymbol{F}_{12} = f(r)\,\hat{\boldsymbol{r}} = -\boldsymbol{F}_{21}\,, \qquad \boldsymbol{x} := \boldsymbol{x}_1 - \boldsymbol{x}_2\,, \quad r = |\boldsymbol{x}|\,. \tag{27.40}$$

Then the centre of mass moves uniformly, and the relative coordinate $\boldsymbol{x}$ obeys the one-body central equation

$$\mu\,\ddot{\boldsymbol{x}} = f(r)\,\hat{\boldsymbol{r}}\,, \qquad \mu := \frac{m_1 m_2}{m_1 + m_2}\,, \tag{27.41}$$

with μ the reduced mass*. Moreover the total kinetic energy and the total angular momentum split without cross terms,*

$$T = \frac{1}{2}\,(m_1 + m_2)\left|\dot{\boldsymbol{X}}\right|^2 + \frac{1}{2}\mu\,|\dot{\boldsymbol{x}}|^2\,, \tag{27.42}$$

$$\boldsymbol{L} = (m_1 + m_2)\,\boldsymbol{X} \times \dot{\boldsymbol{X}} + \mu\,\boldsymbol{x} \times \dot{\boldsymbol{x}}\,, \tag{27.43}$$

with $\boldsymbol{X}$ the centre of mass Equation (19.38)*.*

Rests on Postulate 19.48 and Equations (19.4) *and* (19.38)*.*

Proof. *Derives Theorem 27.30.*

Adding the two equations of motion and using Equation (27.40) gives $m_1\ddot{\boldsymbol{x}}_1 + m_2\ddot{\boldsymbol{x}}_2 = \boldsymbol{0}$, that is $\ddot{\boldsymbol{X}} = \boldsymbol{0}$ by Equation (19.40): the centre of mass moves uniformly, as Phenomenon 19.54 states in general. Subtracting instead, after dividing each equation by its own mass,

$$\ddot{\boldsymbol{x}} = \ddot{\boldsymbol{x}}_1 - \ddot{\boldsymbol{x}}_2 = \frac{\boldsymbol{F}_{12}}{m_1} - \frac{\boldsymbol{F}_{21}}{m_2} = \left(\frac{1}{m_1} + \frac{1}{m_2}\right) f(r)\,\hat{\boldsymbol{r}} = \frac{f(r)}{\mu}\,\hat{\boldsymbol{r}}\,, \tag{27.44}$$

From Equations (19.4) and (27.40) (dividing each equation of motion by its mass and subtracting, with $\mu^{-1} = m_1^{-1} + m_2^{-1}$).

which is Equation (27.41). For the splittings, invert the change of variables: with $M = m_1 + m_2$,

$$\boldsymbol{x}_1 = \boldsymbol{X} + \frac{m_2}{M}\boldsymbol{x}\,, \qquad \boldsymbol{x}_2 = \boldsymbol{X} - \frac{m_1}{M}\boldsymbol{x}\,.$$

Substituting into $T = \frac{1}{2}m_1\left|\dot{\boldsymbol{x}}_1\right|^2 + \frac{1}{2}m_2\left|\dot{\boldsymbol{x}}_2\right|^2$, the cross terms carry the factor $m_1m_2/M - m_2m_1/M = 0$ and cancel, while the remaining coefficients of $|\dot{\boldsymbol{x}}|^2$ combine to $\frac{1}{2}\left(m_1m_2^2 + m_2m_1^2\right)/M^2 = \frac{1}{2}\mu$, giving Equation (27.42). The same substitution in $\boldsymbol{L} = m_1\boldsymbol{x}_1 \times \dot{\boldsymbol{x}}_1 + m_2\boldsymbol{x}_2 \times \dot{\boldsymbol{x}}_2$ cancels the cross terms for the same reason and gives Equation (27.43). □

Remark 27.31 (What the reduction does and does not buy). Every result of Sections 27.2 and 27.3 now applies to the two-body problem verbatim, with m read as the reduced mass μ, r as the separation and $\boldsymbol{L}$ as the angular momentum about the centre of mass. When one mass overwhelms the other, $\mu \to m_2$ for $m_1 \gg m_2$ and the centre of mass sits inside the heavy body, which recovers the fixed-centre picture; for the Sun and Jupiter μ differs from Jupiter's mass by one part in 1047, and the common centre of mass lies just outside the solar surface. The reduction is exact for *two* bodies and for no more: the three-body problem admits no such change of variables, has no general solution in closed form, and is where the qualitative theory of Chapter 32 begins.

27.4.2 The orbit of an inverse-square attraction

Phenomenon 27.32 (Kepler's first law: bound orbits are closed ellipses)**.** Each planet describes an ellipse with the Sun at one focus — not at the centre — and the orbit closes upon itself, returning the body to the same point with the same velocity after one revolution. Closure is a strong and very special fact. For a general central force the orbit is bounded, by Phenomenon 27.26, but it does not repeat: successive perihelia occur at different longitudes and the path fills the annulus between the turning radii. The observed closure of planetary orbits is therefore evidence for the inverse-square law itself, and the small residual advances of the perihelia are evidence that the law is not exact [14].

Rests on Equations (27.13) and (27.28) and Theorem 27.30.

Derivation. Derives Phenomenon 27.32.

Take the Newtonian attraction between the two bodies,

$$f(r) = -\frac{k}{r^2}\,, \qquad k := Gm_1m_2 > 0\,, \tag{27.45}$$

whose potential Equation (27.23) is $V(r) = -k/r$ once the additive constant is fixed by requiring $V \to 0$ as $r \to \infty$, and which by Theorem 27.30 makes the separation obey the one-body central equation with mass μ. Insert Equation (27.45) into the Binet equation Equation (27.30), written with μ in place of m and $u = 1/r$, using $\mu a_r = f = -ku^2$:

$$-ku^2 = -\frac{L^2u^2}{\mu}\left(\frac{\mathrm{d}^2u}{\mathrm{d}\phi^2} + u\right),$$

and the factor u^2 cancels on both sides — which happens for the inverse square and for no other power. What is left is linear with constant coefficients,

$$\frac{\mathrm{d}^2u}{\mathrm{d}\phi^2} + u = \frac{\mu k}{L^2}\,. \tag{27.46}$$

From Equations (27.30) and (27.45) (substituting $f = -ku^2$ and cancelling the common factor u^2).

By Theorem 9.13 its general solution is any one solution plus the general solution of the homogeneous equation; the constant $u = \mu k/L^2$ is a solution, and the homogeneous equation is Equation (9.106) with $\lambda = 1$, whose real general solution is Equation (9.108). Choosing the origin of ϕ at the maximum of u — the point of closest approach — absorbs the phase, and writing the remaining constant as $e\,\mu k/L^2$ with $e \geq 0$,

$$\frac{1}{r} = \frac{\mu k}{L^2}\left(1 + e\cos\phi\right), \qquad \text{that is} \qquad r = \frac{p}{1 + e\cos\phi}, \quad p := \frac{L^2}{\mu k}. \tag{27.47}$$

This is the polar equation of a conic of eccentricity e with one focus at the centre of force and semi-latus rectum p (Remark 27.33); for $0 \leq e < 1$ it is an ellipse, and the centre of force is at a focus and not at the centre, since r takes its extreme values $p/(1+e)$ and $p/(1-e)$ at $\phi = 0$ and $\phi = \pi$ and these are unequal unless $e = 0$. The solution is single-valued and 2π-periodic in ϕ, so the orbit closes exactly, and the identification of e with the energy is Proposition 27.34. □

Remark 27.33 (The conic classification is owed by Part II). Equation (27.47) is derived here in full; what is *quoted* is its geometric reading. That $r = p/\left(1 + e\cos\phi\right)$ is the locus of points whose distance from a fixed focus stands in the fixed ratio e to the distance from a fixed directrix, and that this locus is an ellipse, a parabola or a hyperbola according as $e < 1$, $e = 1$ or $e > 1$, is elementary analytic geometry which Chapters 5 and 7 do not state. Nothing in the physics below depends on it: the semi-major axis is defined in Equation (27.49) as half the sum of the two turning radii, which is intrinsic to the orbit, and Kepler's third law is derived from Equation (27.47) directly without any appeal to the geometry of the ellipse. The word "ellipse" is a name for the curve, not a step in an argument.

Proposition 27.34 (Eccentricity, energy and semi-major axis)**.** *For the orbit Equation* (27.47) *of the attraction Equation* (27.45)*,*

$$e = \sqrt{1 + \frac{2EL^2}{\mu k^2}}, \tag{27.48}$$

and, when $e < 1$, half the sum of the two turning radii is

$$a := \frac{r_- + r_+}{2} = \frac{p}{1 - e^2} = -\frac{k}{2E}. \tag{27.49}$$

Rests on Equations (27.31)*,* (27.32) *and* (27.47)*.*

Proof. *Derives Proposition 27.34.*

At a turning point $\dot r = 0$, so by Equation (27.31) with $V = -k/r = -ku$ and μ for m,

$$E = \frac{L^2}{2\mu}u^2 - ku. \tag{27.50}$$

The turning values of u are the extremes of Equation (27.47), namely $u_\pm = \left(\mu k/L^2\right)\left(1 \pm e\right)$. Substituting u_+ into Equation (27.50) and using $p = L^2/\left(\mu k\right)$,

$$E = \frac{L^2}{2\mu}\frac{\mu^2 k^2}{L^4}\left(1+e\right)^2 - k\frac{\mu k}{L^2}\left(1+e\right) = \frac{\mu k^2}{2L^2}\left(1+e\right)\left[\left(1+e\right) - 2\right] = \frac{\mu k^2}{2L^2}\left(e^2 - 1\right),$$

and solving for e gives Equation (27.48); the same substitution with u_- gives the same relation, as it must, since both turning points belong to one orbit. For Equation (27.49), $r_\pm = p/\left(1 \mp e\right)$ so

$$a = \frac{p}{2}\left(\frac{1}{1+e} + \frac{1}{1-e}\right) = \frac{p}{1-e^2},$$

and inserting $1 - e^2 = -2EL^2/\left(\mu k^2\right)$ from Equation (27.48) together with $p = L^2/\left(\mu k\right)$ gives $a = -k/(2E)$, positive precisely when $E < 0$. □

Equation (27.48) classifies the orbits completely by the two constants of the motion, and the classification is collected in Table 27.1. It is worth reading as a statement about *what is observable*: the energy fixes the size of a bound orbit through Equation (27.49) and the angular momentum fixes its shape through p, so that measuring the period and the eccentricity of a planet measures E and L and nothing else is left to determine.

Table 27.1: Classification of the orbits of the inverse-square attraction Equation (27.45) by the energy, through the eccentricity Equation (27.48). The least possible energy $-\mu k^2/\left(2L^2\right)$ is the minimum of the effective potential Equation (27.32) at fixed L; below it no motion exists with that angular momentum. Only the first two rows are bounded, and both close.

Energy	Eccentricity	Curve	Motion
$E = -\mu k^2/\left(2L^2\right)$	$e = 0$	circle	r constant
$-\mu k^2/\left(2L^2\right) < E < 0$	$0 < e < 1$	ellipse	bounded, closed
$E = 0$	$e = 1$	parabola	escapes with zero final speed
$E > 0$	$e > 1$	hyperbola	escapes with finite final speed

27.4.3 Kepler's third law

Phenomenon 27.35 (Kepler's third law)**.** The squares of the orbital periods of the planets stand in the same ratio as the cubes of the semi-major axes of their orbits, with one and the same constant of proportionality for every body bound to the same centre, whatever its own mass. The regularity holds across the Solar System over a range of nearly three orders of magnitude in period, and holds again — with a different constant — for the moons of Jupiter, for artificial satellites of the Earth, and for the components of binary stars. It is by this law that the masses of astronomical bodies are determined at all [14].

Rests on Equations (27.22) *and* (27.47) *and Theorem 27.30.*

Derivation. *Derives Phenomenon 27.35.*

For a circular orbit the result is immediate. A body of mass m carried around a circle of radius r requires, by the kinematics of Chapter 18, the centripetal acceleration $\omega^2 r$, and the attraction of a central mass M supplies it, so that $GM/r^2 = \omega^2 r$ and, with $\omega = 2\pi/T$,

$$T^2 = \frac{4\pi^2}{GM}\, r^3 \,. \tag{27.51}$$

The orbiting mass has cancelled — it appears once as the gravitational charge and once as the inertia — which is why a single constant serves every planet.

For the general bounded orbit, integrate the equal-area law. By Equation (27.22) the areal speed is $L/\left(2\mu\right)$, and by Equation (27.20) the period is

$$T = \int_0^T \mathrm{d}t = \int_0^{2\pi} \frac{\mathrm{d}\phi}{\dot\phi} = \frac{\mu}{L}\int_0^{2\pi} r^2\,\mathrm{d}\phi = \frac{\mu p^2}{L}\int_0^{2\pi} \frac{\mathrm{d}\phi}{\left(1 + e\cos\phi\right)^2} \,,$$

using Equation (27.47) in the last step. The integral is elementary by residues. For $0 < e < \lambda$ the substitution $z = \mathrm{e}^{\mathrm{i}\phi}$, under which $\cos\phi = \left(z + z^{-1}\right)/2$ and $\mathrm{d}\phi = \mathrm{d}z/\left(\mathrm{i}z\right)$, turns $I(\lambda) = \int_0^{2\pi} \left(\lambda + e\cos\phi\right)^{-1} \mathrm{d}\phi$ into a contour integral on the unit circle,

$$I(\lambda) = \oint_{|z|=1} \frac{2\,\mathrm{d}z}{\mathrm{i}\left(ez^2 + 2\lambda z + e\right)} \,,$$

whose two poles have product 1, so that exactly one lies inside, namely $z_+ = \left(-\lambda + \sqrt{\lambda^2 - e^2}\right)/e$; by the residue theorem Theorem 8.24 and Proposition 8.23, $I(\lambda) =$

$2\pi/\sqrt{\lambda^2 - e^2}$, which is also the value at $e = 0$, where the integral is immediate. Differentiating under the integral sign in λ and then setting $\lambda = 1$,

$$\int_0^{2\pi} \frac{\mathrm{d}\phi}{(1 + e\cos\phi)^2} = -I'(1) = \frac{2\pi}{(1 - e^2)^{3/2}}\,. \tag{27.52}$$

Hence, with $a = p/(1 - e^2)$ from Equation (27.49) and $p = L^2/(\mu k)$,

$$T = \frac{2\pi\mu p^2}{L\,(1 - e^2)^{3/2}} = \frac{2\pi\mu}{L}\,p^{1/2}a^{3/2} = \frac{2\pi\mu}{L}\cdot\frac{L}{\sqrt{\mu k}}\,a^{3/2} = 2\pi\sqrt{\frac{\mu}{k}}\,a^{3/2}\,.$$

Squaring, and putting $k = Gm_1m_2$ and $\mu = m_1m_2/(m_1 + m_2)$ from Theorem 27.30,

$$T^2 = \frac{4\pi^2}{G\,(m_1 + m_2)}\,a^3\,. \tag{27.53}$$

From Equations (27.41), (27.49) and (27.52) (integrating the equal-area law over one revolution and eliminating p and L).

This is the exact law. It contains Equation (27.51) as the limit $m_2 \ll m_1$, it replaces the radius by the semi-major axis, and it shows precisely how the "constant" fails to be one: the proportionality factor depends on the *total* mass, so two planets of different masses about the same star obey slightly different laws. The check of Equation (33.19) in the Cavendish chapter is this same relation applied to the Moon. □

Table 27.2: Kepler's third law across the Solar System. The semi-major axes and sidereal periods are the standard J2000 mean Keplerian elements, converted to SI; the last column is the ratio that Equation (27.53) predicts to be $4\pi^2/[G\,(M_\odot + m)]$, which for a massless planet is $2.9747\times10^{-19}\ \mathrm{s^2/m^3}$ with $GM_\odot = 1.327\,12\times10^{20}\ \mathrm{m^3/s^2}$. This treatise carries no citable ephemeris source, and the elements below are therefore uncited.

Body	a (m)	T (s)	T^2/a^3 ($\mathrm{s^2/m^3}$)
Mercury	5.7909×10^{10}	7.6005×10^{6}	$2.974\,73 \times 10^{-19}$
Venus	1.0821×10^{11}	1.9414×10^{7}	$2.974\,73 \times 10^{-19}$
Earth	1.4960×10^{11}	3.1558×10^{7}	$2.974\,72 \times 10^{-19}$
Mars	2.2794×10^{11}	5.9355×10^{7}	$2.974\,90 \times 10^{-19}$
Jupiter	7.7841×10^{11}	3.7436×10^{8}	$2.971\,26 \times 10^{-19}$
Saturn	1.4267×10^{12}	9.2929×10^{8}	$2.973\,61 \times 10^{-19}$
Uranus	2.8710×10^{12}	2.6514×10^{9}	$2.970\,67 \times 10^{-19}$
Neptune	4.4983×10^{12}	5.2004×10^{9}	$2.971\,29 \times 10^{-19}$

Table 27.2 is the law's evidence and also the measure of its limits. Over a 684-fold range in period and a 78-fold range in semi-major axis the ratio T^2a^{-3} is constant to 0.14 %, which is the observation Kepler made and Newton explained. The residual structure is not noise. The four terrestrial planets agree with each other and with $4\pi^2/(GM_\odot)$ to a few parts in 10^5; the giant planets lie systematically *below* that value, which is the sign Equation (27.53) demands, since their own masses enter the denominator. For Jupiter, $m/M_\odot = 9.5 \times 10^{-4}$ predicts a deficit of that relative size and one of about 1.2×10^{-3} is observed; for Saturn, 2.9×10^{-4} predicted against 3.8×10^{-4} observed. Uranus and Neptune depart by more than their masses require, and the honest reading is that a two-body law is not expected to describe them to better than this: their mutual perturbation is large enough to have betrayed Neptune's existence before it was seen, and the mean elements quoted are themselves fits over a span of centuries rather than exact constants of any orbit.

Example 27.36 (A satellite in low Earth orbit)**.** Equation (27.53) with $m_2 \ll m_1$ gives the period of a circular orbit of radius r about the Earth as $T = 2\pi\sqrt{r^3/GM}$. Taking $GM = 3.986 \times 10^{14}\,\mathrm{m^3/s^2}$ as in Example 27.27, and an orbit $4.20 \times 10^5\,\mathrm{m}$ above the mean radius $R_\oplus = 6.371 \times 10^6\,\mathrm{m}$, so that $r = 6.791 \times 10^6\,\mathrm{m}$:

$$T = 2\pi\sqrt{\frac{(6.791 \times 10^6\,\mathrm{m})^3}{3.986 \times 10^{14}\,\mathrm{m^3/s^2}}} = 5.57 \times 10^3\,\mathrm{s}\,,$$

that is 92.8 minutes, against the 92.9 minutes in which the International Space Station is observed to circle at that altitude. The orbital speed is $\sqrt{GM/r} = 7.66 \times 10^3\,\mathrm{m/s}$. Note that the period *decreases* as the orbit is lowered, so that a satellite which loses energy to atmospheric drag speeds up — the characteristic and initially counter-intuitive signature of a bound inverse-square orbit, and a direct consequence of Equation (27.49), in which a more negative energy means a smaller a.

Rests on Equation (27.53).

27.4.4 The Laplace–Runge–Lenz vector

The inverse-square attraction conserves more than energy and angular momentum. It conserves a vector lying in the plane of the orbit, and that extra conservation law is exactly what makes the orbit close.

Definition 27.37 (Laplace–Runge–Lenz vector)**.** For the attraction Equation (27.45), with $\boldsymbol{p} = \mu\dot{\boldsymbol{x}}$ the momentum conjugate to the relative coordinate, the *Laplace–Runge–Lenz vector* is

$$\boldsymbol{A} := \boldsymbol{p} \times \boldsymbol{L} - \mu k\,\hat{\boldsymbol{r}}\,, \tag{27.54}$$

whose SI unit is $\mathrm{kg^2\,m^3/s^2}$, that of a momentum times an angular momentum.

Rests on Equations (27.20), (27.41) *and* (27.45).

Theorem 27.38 (Conservation of the Laplace–Runge–Lenz vector)**.** *Along any motion of the inverse-square attraction,* $\mathrm{d}\boldsymbol{A}/\mathrm{d}t = \boldsymbol{0}$. *The vector lies in the plane of the orbit, since* $\boldsymbol{A} \cdot \boldsymbol{L} = 0$.

Rests on Equations (27.1), (27.41) *and* (27.54).

Proof. *Derives Theorem 27.38.*

Differentiate Equation (27.54). Since $\boldsymbol{L}$ is constant by Proposition 27.18, $\mathrm{d}\left(\boldsymbol{p} \times \boldsymbol{L}\right)/\mathrm{d}t = \dot{\boldsymbol{p}} \times \boldsymbol{L} = \boldsymbol{F} \times \boldsymbol{L}$ by Equation (27.41). With $\boldsymbol{F} = -k\hat{\boldsymbol{r}}/r^2$ and $\boldsymbol{L} = \mu\,\boldsymbol{x} \times \boldsymbol{v}$, the expansion Equation (27.1) of the double cross product gives

$$\boldsymbol{F} \times \boldsymbol{L} = -\frac{k\mu}{r^2}\,\hat{\boldsymbol{r}} \times (\boldsymbol{x} \times \boldsymbol{v}) = -\frac{k\mu}{r^2}\left[\boldsymbol{x}\,(\hat{\boldsymbol{r}} \cdot \boldsymbol{v}) - \boldsymbol{v}\,(\hat{\boldsymbol{r}} \cdot \boldsymbol{x})\right] = k\mu\left(\frac{\boldsymbol{v}}{r} - \frac{\dot{r}\,\boldsymbol{x}}{r^2}\right),$$

using $\hat{\boldsymbol{r}} \cdot \boldsymbol{v} = \dot{r}$ and $\hat{\boldsymbol{r}} \cdot \boldsymbol{x} = r$. What multiplies $k\mu$ on the right is exactly the derivative of the unit radial vector,

$$\frac{\mathrm{d}\hat{\boldsymbol{r}}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\boldsymbol{x}}{r}\right) = \frac{\boldsymbol{v}}{r} - \frac{\dot{r}\,\boldsymbol{x}}{r^2}\,,$$

so $\mathrm{d}\left(\boldsymbol{p} \times \boldsymbol{L}\right)/\mathrm{d}t = \mu k\,\mathrm{d}\hat{\boldsymbol{r}}/\mathrm{d}t$ and the two terms of Equation (27.54) cancel in the derivative. Finally $\boldsymbol{A} \cdot \boldsymbol{L} = (\boldsymbol{p} \times \boldsymbol{L}) \cdot \boldsymbol{L} - \mu k\,\hat{\boldsymbol{r}} \cdot \boldsymbol{L} = 0$, the first term because $\boldsymbol{p} \times \boldsymbol{L}$ is orthogonal to $\boldsymbol{L}$ and the second because $\boldsymbol{L}$ is orthogonal to $\boldsymbol{x}$. □

Proposition 27.39 (The Laplace–Runge–Lenz vector is the orbit)**.** *Measuring* ϕ *from the direction of* $\boldsymbol{A}$,

$$r = \frac{L^2/(\mu k)}{1 + (A/\mu k)\cos\phi}\,, \qquad A^2 = \mu^2 k^2 + 2\mu E L^2\,. \tag{27.55}$$

Thus $\boldsymbol{A}$ *points from the centre of force towards the point of closest approach, and its magnitude is* $\mu k e$ *with* e *the eccentricity Equation* (27.48).

Rests on Equations (27.47), (27.48) *and* (27.54).

Proof. *Derives Proposition 27.39.*

Take the scalar product of Equation (27.54) with $\boldsymbol{x}$. On the left, $\boldsymbol{x}\cdot\boldsymbol{A} = Ar\cos\phi$ with ϕ measured from $\boldsymbol{A}$. On the right, the cyclic property of the scalar triple product gives $\boldsymbol{x}\cdot(\boldsymbol{p}\times\boldsymbol{L}) = \boldsymbol{L}\cdot(\boldsymbol{x}\times\boldsymbol{p}) = L^2$, while $\boldsymbol{x}\cdot\mu k\hat{\boldsymbol{r}} = \mu kr$. Hence $Ar\cos\phi = L^2 - \mu kr$, which rearranged is the first member of Equation (27.55). Comparing with Equation (27.47) identifies $A = \mu ke$, and substituting Equation (27.48) gives $A^2 = \mu^2k^2e^2 = \mu^2k^2 + 2\mu EL^2$. Since r is least where $\cos\phi = 1$, the vector points at the perihelion. □

That the perihelion direction is a constant of the motion is another way of saying that the orbit does not precess, and it is the sharpest statement of Kepler's first law: the extra conserved vector is what distinguishes the inverse-square attraction from every other central force, all of which conserve E and $\boldsymbol{L}$ just as well.

Remark 27.40 (It is not the charge of a point symmetry). It is tempting to look for the symmetry behind Theorem 27.38 among the transformations that move the particle's position, as rotations and translations do. There is none, and Remark 16.86 proves it: the part of the Noether charge Equation (16.76) that is quadratic in the momentum is isotropic in $\boldsymbol{p}$ for any point transformation, whereas the corresponding part of $\boldsymbol{A}$ is not — it vanishes when $\boldsymbol{p}$ is parallel to $\boldsymbol{x}$ and is maximal when they are perpendicular. The generator that does produce $\boldsymbol{A}$ depends on the velocities and is a *generalized* symmetry, outside the hypotheses of Noether's first theorem as proved in Theorem 16.84. So Theorem 27.38 is established here, as it must be, by direct differentiation and not by exhibiting a symmetry of the action. The hidden SO(4) structure the vector generates is not idle, however: it is what makes the hydrogen spectrum degenerate in the orbital angular momentum, and it is put to work in Chapter 81. The vector is far older than its three-barrelled name; what attached Lenz's to it is his use of it in the old quantum theory of the perturbed Kepler motion [514], and Pauli obtained the hydrogen spectrum from it two years later, before Schrödinger's wave equation appeared [553].

27.4.5 Closure, and what its failure measures

Theorem 27.41 (Bertrand)**.** *Among the central forces for which every bounded orbit is closed there are exactly two: the inverse-square attraction $f(r) = -k/r^2$ and the linear attraction $f(r) = -\kappa r$ of the isotropic harmonic oscillator. For every other force law the apsidal angle Equation (27.39) depends on the orbit, and orbits with Φ/π irrational fill the annulus between their turning radii without ever repeating.*

Rests on Equations (27.39) and (27.47).

Full derivation: Appendix A.55 (p. 3094).

Derives Theorem 27.41.

The proof there runs in two stages, and it is worth knowing which does which. The first perturbs a circular orbit of radius r_0 and finds that the perturbation oscillates about it with an angular frequency $\beta(r_0) = \left[3 + r_0 f'(r_0)/f(r_0)\right]^{1/2}$ per unit change of the polar angle, so that the apsidal angle Equation (27.39) tends to π/β. Closure of every nearby orbit makes β rational; β depends continuously on r_0 and a continuous rational-valued function is constant, so β is one fixed rational number, and $rf'/f = \beta^2 - 3$ integrates to the power law $f \propto r^{\beta^2-3}$. That much is a first-order calculation, and it narrows the candidates to a one-parameter family of power laws — it does *not* yet pick out two of them. The second stage supplies the difference, because closure of the orbits infinitesimally near the circle is weaker than closure of all bounded orbits: carrying the expansion to third order produces a solvability condition on β which only $\beta = 1$ and $\beta = 2$ satisfy, and those are $f \propto r^{-2}$ and $f \propto r$.

The two exceptional laws are also the two that Nature exhibits at large scale, which is why the closure of orbits was discovered before the reason for it. The importance of the

theorem is that it converts a qualitative observation into a quantitative test: any departure from r^{-2}, however small, makes the perihelion move.

Phenomenon 27.42 (Planetary orbits do not close exactly)**.** The perihelia of the planets advance slowly along their orbits. Mercury's is observed to advance by some 5600 seconds of arc per century, of which 5025 is the general precession of the equinoxes — a motion of the reference frame and not of the orbit — and 531 is the pull of the other planets, computable within Newtonian mechanics. A residue of about 43 seconds of arc per century survives every such computation. Le Verrier isolated that residue in 1859 and no distribution of matter consistent with the rest of Solar-System astronomy was ever found to produce it [151]; the budget is set out in Table 53.1. By Theorem 27.41 the residue is a measurement: it says that the force governing Mercury is not exactly $-k/r^2$.

Rests on Theorem 27.41 and Equation (27.39).

Derivation. *Derives Phenomenon 27.42.*

Within this chapter the derivation can go only as far as the conditional statement, and that is the honest content of the phenomenon. Theorem 27.41 shows that closure of every bounded orbit is equivalent to the force being $-k/r^2$ or $-\kappa r$, and the second is excluded by Phenomenon 27.32, since a linear attraction puts the centre of force at the *centre* of the ellipse and not at a focus. A non-closing bound orbit therefore falsifies the exact inverse square. What it does not do is say what replaces it. The first attempt was to keep a central force and modify the exponent, writing $f \propto r^{-(2+\delta)}$; for a nearly circular orbit a small δ produces an apsidal angle $\Phi \approx \pi\,(1+\delta/2)$ and hence a slow rotation of the line of apsides, and δ can always be chosen to fit one planet. The proposal fails because the δ that fits Mercury does not fit the others, and because it has no independent motivation whatever.

The correct account is not a modified force law at all. In general relativity the orbit is a geodesic of a curved spacetime and its equation acquires the term shown in Equation (53.2), quadratic in $u = 1/r$ and suppressed by c^{-2}; the resulting advance is Equation (53.1), with no adjustable parameter, and it predicts 42.98 seconds of arc per century for Mercury against the observed residue. That derivation belongs to Chapter 53, where the measurement and its error budget are set out, and to Phenomenon 53.1, which states the result. What this chapter contributes is the logic that makes the measurement mean anything: without Theorem 27.41 a non-closing orbit would be an anomaly of no particular significance, and with it, it is a bound on the force law. □

Remark 27.43 (The repulsive case, and where it is used)*.* Nothing in Section 27.4.2 required $k > 0$. For a repulsive inverse-square force, $k < 0$ in Equation (27.45), the effective potential falls monotonically and has no well, so every motion has $E > 0$ and Equation (27.48) gives $e > 1$. The semi-latus rectum $p = L^2/(\mu k)$ of Equation (27.47) is then negative, and the physical solution is the branch on which $1 + e\cos\phi < 0$: the branch of the hyperbola that curves away from the centre rather than around it. This is the orbit of an alpha particle in the Coulomb field of a nucleus, and integrating Equation (27.38) for it produces the relation between impact parameter and deflection angle from which the Rutherford cross section follows; the calculation, and the scattering experiment that made it famous, are in Section 70.3.2. The same hyperbolic solution, taken in the attractive case with $E > 0$, is the gravitational slingshot: a spacecraft leaves a planet's neighbourhood with the same speed relative to the planet with which it arrived, but in a different direction, and therefore with a different — and usefully larger — speed relative to the Sun.

27.5 The virial theorem

Everything so far has followed individual orbits. For a bound system of many bodies — a star cluster, a gas of self-gravitating masses — the orbits are inaccessible, but a statement about

time averages survives, and it is the tool by which such systems are weighed (Chapter 48).

Theorem 27.44 (Virial theorem)**.** *Let N particles with positions $\boldsymbol{x}_a$ and momenta $\boldsymbol{p}_a$ move under forces $\boldsymbol{F}_a$ such that the motion is bounded: all $|\boldsymbol{x}_a|$ and $|\boldsymbol{p}_a|$ remain finite for all time. Then the long-time averages $\langle f \rangle = \lim_{\tau\to\infty} \tau^{-1}\int_0^\tau f \,\mathrm{d}t$ satisfy*

$$2\,\langle T\rangle = -\Big\langle \sum_{a=1}^{N} \boldsymbol{F}_a \cdot \boldsymbol{x}_a \Big\rangle . \tag{27.56}$$

If the forces derive from a potential V that is a homogeneous function of degree n of the positions, the right side is $n\,\langle V\rangle$, so

$$2\,\langle T\rangle = n\,\langle V\rangle \ , \tag{27.57}$$

and for the inverse-square attraction ($n=-1$)

$$2\,\langle T\rangle = -\,\langle V\rangle \ . \tag{27.58}$$

Rests on Equation (27.41) and Remark 21.40.

Proof. *Derives Theorem 27.44.*

Consider $G = \sum_a \boldsymbol{p}_a \cdot \boldsymbol{x}_a$. Along the motion,

$$\frac{\mathrm{d}G}{\mathrm{d}t} = \sum_a \dot{\boldsymbol{p}}_a \cdot \boldsymbol{x}_a + \sum_a \boldsymbol{p}_a \cdot \dot{\boldsymbol{x}}_a = \sum_a \boldsymbol{F}_a \cdot \boldsymbol{x}_a + 2T \,,$$

since $\boldsymbol{p}_a \cdot \dot{\boldsymbol{x}}_a = m_a\,|\dot{\boldsymbol{x}}_a|^2$ sums to twice the kinetic energy. Averaging over $[0,\tau]$, $\langle \mathrm{d}G/\mathrm{d}t\rangle = [G(\tau)-G(0)]/\tau$, and boundedness of positions and momenta bounds G, so the limit $\tau\to\infty$ vanishes (exactly, already at one period, for periodic motion). This is Equation (27.56). When $\boldsymbol{F}_a = -\partial V/\partial \boldsymbol{x}_a$ with V homogeneous of degree n, Euler's homogeneous-function theorem (Remark 21.40) gives $\sum_a \boldsymbol{x}_a \cdot \partial V/\partial \boldsymbol{x}_a = nV$, whence Equation (27.57); $n=-1$ is the gravitational and Coulomb case. □

Corollary 27.45 (Bound Kepler averages)**.** *For a bound orbit of the inverse-square attraction, the conserved energy $E = T + V$ satisfies*

$$\langle T\rangle = -E \,, \qquad \langle V\rangle = 2E \,, \tag{27.59}$$

so a bound orbit has $E<0$, and adding energy to it slows *the time-averaged motion — the negative specific heat of gravitating systems.*

Rests on Theorem 27.44.

Proof. *Derives Corollary 27.45.*

E is constant, so $E = \langle T\rangle + \langle V\rangle$; combine with Equation (27.58). □

27.6 Where the claims of this chapter are tested

It is worth saying plainly which of the observed facts asserted above are tied to an experiment described in this book and which are not. Kepler's third law is tested in Chapter 33, whose lunar check Equation (33.19) is Equation (27.53) applied to the Earth–Moon system with a laboratory value of G; the failure of exact closure, Phenomenon 27.42, is measured in Chapter 53; and the repulsive branch of Section 27.4.2 is what Section 70.3.2 tests. The remaining five observed statements — Phenomena 27.5, 27.8, 27.19, 27.26 and 27.32 — carry no dedicated experiment chapter here. They are not for that reason unsupported: the first is a commonplace of every static structure, the second is the standard method by which a centre of gravity is located, and the last three are the content of the regularities Kepler extracted from Brahe's catalogue of the positions of Mars. But no measurement with its own uncertainty budget is set out in this book for any of the five, and the honest record of that is this paragraph rather than silence.

Chapter 28

Oscillations and Mechanical Waves

The harmonic oscillator is the most reused model in physics. Near any stable equilibrium the Lagrangian of Chapter 21 is, to lowest order, a quadratic form, so everything close to rest oscillates harmonically: this chapter sits immediately after the general formalisms of Chapters 21 and 22 and before everything in the treatise that vibrates — the lattice phonons of Chapter 125, the electromagnetic modes of Chapter 61, and ultimately the field quanta of Part XI. Its second half lets the number of oscillators grow without bound: coupled oscillators pass to a continuum, the normal-mode problem becomes d'Alembert's wave equation on a string, and superposition, standing waves, Fourier decomposition, dispersion and sound follow, each named below with the primary source that established it. The standard modern treatments are [1585, 2027]; the founding monograph of the wave half is Rayleigh's *Theory of Sound* [198].

The mathematics is not developed here. The eigenvalue problem of Section 28.3.2 is the spectral theorem of Chapter 5; the boundary-value problems of Section 28.4.3 are the Sturm–Liouville theory of Chapter 9; the wave equation and its general solution are Chapter 10; the modal expansions are the Fourier theory of Chapter 17. What this chapter supplies is the physics: which quadratic form, which boundary condition, which measured number.

Notation 28.1 (Symbols reused in this chapter). Three letters do double duty and the split is by half of the chapter. In the oscillator half (Sections 28.1 and 28.2) T is a *period* and k a *spring constant*, in N/m. From Section 28.3.3 onward T is the *tension* of a string, in N, k is a *wavenumber*, in 1/m, and a period is written $2\pi/\omega$. In Section 28.3.2 T is kinetic energy and V potential energy, as everywhere in Chapter 21. Likewise μ is a reduced mass in Section 28.1.3 and a linear mass density, in kg/m, from Section 28.3.3 onward. In Section 28.5 T is a thermodynamic temperature; no tension appears there.

Remark 28.2 (Tabulated material values used below). Four numbers enter the worked examples of this chapter as ordinary handbook values for which this treatise cites no primary source: the mean molar mass of dry air, $M = 28.96 \times 10^{-3}\,\mathrm{kg/mol}$; the thermal diffusivity of air near room temperature, $2.1 \times 10^{-5}\,\mathrm{m^2/s}$; the density of spring steel, $7.85 \times 10^{3}\,\mathrm{kg/m^3}$; and the atomic masses of ^{1}H and ^{35}Cl. They are stated to the precision used and no further, and no conclusion in this chapter turns on their last digit. Everything else — g, c, the molar gas constant, the atomic mass constant — is either an exact SI definition or carries the citation given where it appears.

28.1 The harmonic oscillator

28.1.1 Free oscillations

Hooke announced the linear restoring law in 1678 as the anagram *ut tensio, sic vis* — as the extension, so the force — having sealed it in 1676 to secure priority [13]. It is the whole content of the model.

Definition 28.3 (Simple harmonic oscillator)**.** A particle of mass m moving on a line under a force proportional to its displacement and directed against it,

$$F = -kx\,, \qquad k > 0\,, \tag{28.1}$$

with k the *spring constant*, of SI dimension N/m. Its equation of motion is

$$\ddot{x} + \omega_0^2 x = 0\,, \qquad \omega_0 := \sqrt{\frac{k}{m}}\,, \tag{28.2}$$

and ω_0, of SI dimension rad/s, is the *natural angular frequency*. The Lagrangian and Hamiltonian are

$$\mathcal{L} = \tfrac{1}{2}m\dot{x}^2 - \tfrac{1}{2}kx^2\,, \qquad \mathcal{H} = \frac{p^2}{2m} + \tfrac{1}{2}kx^2\,. \tag{28.3}$$

Rests on Definition 16.96 and Theorem 16.97.

Proposition 28.4 (General solution and isochronism)**.** *Every solution of Equation (28.2) is*

$$x(t) = A\cos\left(\omega_0 t + \varphi\right)\,, \tag{28.4}$$

with $A \geq 0$ and φ fixed by the initial data, and no other solution exists. Its period is $T = 2\pi/\omega_0 = 2\pi\sqrt{m/k}$, independent of A.

Rests on Equation (28.2) and Theorem 9.13.

Proof. *Derives Proposition 28.4.*

Equation (28.2) is linear, homogeneous and of second order with constant coefficients, so by Theorem 9.13 its solution space is two-dimensional, and by Corollary 9.9 every solution extends to all of $\mathbb{R}$. The pair $\cos\omega_0 t$ and $\sin\omega_0 t$ solves it and has Wronskian (Definition 9.14) equal to $\omega_0 \neq 0$, so by Proposition 9.16 it is a basis; writing the general combination in amplitude–phase form gives Equation (28.4). Uniqueness for given $x(0), \dot{x}(0)$ is Theorem 9.8. The period is the least $T > 0$ with $\omega_0 T = 2\pi$, and A appears in Equation (28.4) only as a multiplicative constant, which a linear equation cannot convert into a shift of T: hence the isochronism, and hence also its exactness being no better than the linearity (Section 28.7). □

Proposition 28.5 (Energy, and its equal sharing)**.** *Along Equation (28.4) the total energy is constant,*

$$E = \tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{2}kx^2 = \tfrac{1}{2}m\omega_0^2 A^2\,, \tag{28.5}$$

and the kinetic and potential energies, averaged over one period, are equal and each equal to $E/2$.

Rests on Equations (28.3) and (28.4).

Proof. *Derives Proposition 28.5.*

Substituting Equation (28.4) and $k = m\omega_0^2$,

$$E = \tfrac{1}{2}m\omega_0^2 A^2 \sin^2\left(\omega_0 t + \varphi\right) + \tfrac{1}{2}m\omega_0^2 A^2 \cos^2\left(\omega_0 t + \varphi\right) = \tfrac{1}{2}m\omega_0^2 A^2\,,$$

by the Pythagorean identity (Lemma 7.59), which is also the statement that E does not depend on t. The mean of $\cos^2$ over a period is $\tfrac{1}{2}$ and so is the mean of $\sin^2$, whence $\langle T\rangle = \langle V\rangle = E/2$. The energy oscillates between the two forms at $2\omega_0$ — twice the frequency of the motion — because both forms are quadratic in a quantity oscillating at ω_0. □

The Hamiltonian flow of Equation (28.3) carries every orbit around an ellipse in the phase plane, and the area enclosed is the quantity classical mechanics hands to the old quantum theory.

Proposition 28.6 (Phase-space area and the action variable)**.** *The orbit of energy E is the ellipse $p^2/2m + m\omega_0^2 x^2/2 = E$, and*

$$J := \oint p\,\mathrm{d}x = \frac{2\pi E}{\omega_0}\,, \qquad I := \frac{J}{2\pi} = \frac{E}{\omega_0}\,, \tag{28.6}$$

with I of SI dimension J s.

Rests on Equations (28.3) *and* (28.5).

Proof. *Derives Proposition 28.6.*

The orbit is the level set of $\mathcal{H}$, an ellipse with semi-axes $x_{\max} = \sqrt{2E/m\omega_0^2}$ along x and $p_{\max} = \sqrt{2mE}$ along p. The enclosed area is $\pi x_{\max} p_{\max} = \pi \cdot 2E/\omega_0$, and $\oint p\,\mathrm{d}x$ is that area. The dimension follows from [energy] / [frequency]. □

Remark 28.7 (Why this ellipse is the doorway to quantization). Equation (28.6) says the oscillator's energy is its action variable times its frequency. Imposing $I = n\hbar$ therefore yields $E = n\hbar\omega_0$ at a stroke, which is the Bohr–Sommerfeld rule of Section 9.6.7 and, historically, Planck's hypothesis. The action variable is the invariant of a slowly changing oscillator and the natural coordinate of Chapter 23; the exact quantum spectrum, with its zero-point term, is Chapter 78, and the same ellipse becomes the phase space of a single field mode in Chapter 99.

Rests on Equation (28.6).

Example 28.8 (A laboratory spring, in SI)**.** A mass $m = 0.250\,\mathrm{kg}$ hung on a spring stretches it by $4.90\,\mathrm{cm}$ at rest, so $k = mg/\Delta x = 0.250\,\mathrm{kg} \times 9.806\,65\,\mathrm{m/s^2}/4.90 \times 10^{-2}\,\mathrm{m} = 50.0\,\mathrm{N/m}$. Then $\omega_0 = \sqrt{k/m} = 14.1\,\mathrm{rad/s}$, a frequency of $2.25\,\mathrm{Hz}$ and a period of $0.444\,\mathrm{s}$. Released from $A = 2.00\,\mathrm{cm}$ the energy is $\tfrac{1}{2}kA^2 = 1.00 \times 10^{-2}\,\mathrm{J}$ and the greatest speed is $\omega_0 A = 0.283\,\mathrm{m/s}$. Every one of these numbers is fixed by the two constants m and k alone, which is Phenomenon 28.13 in miniature.

Rests on Equations (28.2) *and* (28.5).

28.1.2 The pendulum

The pendulum is where the harmonic oscillator was first met in nature rather than built, and where its isochronism was first put to work. Galileo reports that a pendulum keeps its period as the swing decays, and proposes it as a timekeeper and as a pulse counter [4]. Huygens turned the observation into the pendulum clock, and in the *Horologium Oscillatorium* of 1673 gave both the exact isochronous curve — the cycloid — and the theory of the compound pendulum, including the centre of oscillation and the reciprocity that makes a reversible pendulum an absolute gravimeter [11]. The measurements themselves are Chapter 34, and the clock is the first entry in the history of frequency standards taken up in Chapter 2.

Phenomenon 28.9 (Isochronism of the small-amplitude pendulum)**.** A pendulum swinging through a small arc keeps the same period whether the arc is wide or narrow: two pendulums of equal length released from different heights stay in step for hundreds of swings. The period depends on the length of the suspension and on the local acceleration of free fall, and — to the accuracy of the observation — on nothing else: not on the amplitude, and not on the mass or the material of the bob. Galileo reports the observation and draws from it the pendulum's use as a timekeeper [4]; it is the fact on which the pendulum clock, the first precision frequency standard, was built.

$$T = 2\pi\sqrt{\frac{\ell}{g}}\,. \tag{28.7}$$

Rests on Equation (28.2) and Proposition 28.4.

Derivation. *Derives Phenomenon 28.9.*

A bob of mass m on a light rod of length ℓ, displaced by an angle θ from the downward vertical, has moment of inertia $m\ell^2$ about the pivot, while the weight exerts about the pivot the moment $-mg\ell\sin\theta$. The rotational law of Chapter 19 then gives

$$\ell\,\ddot{\theta} = -g\sin\theta\,, \tag{28.8}$$

in which the mass has cancelled — which accounts for the observed independence of the bob. For $|\theta| \ll 1$ one has $\sin\theta = \theta + O(\theta^3)$ by Theorem 7.38, and Equation (28.8) reduces to the harmonic equation $\ddot{\theta}+\omega_0^2\theta = 0$ with $\omega_0^2 = g/\ell$, whose general solution is Equation (28.4) with period $T = 2\pi/\omega_0$. That is Equation (28.7), and it contains no A: in a linear equation the amplitude enters only as an overall multiplicative constant, and no linear equation can make its period depend on it (Proposition 28.4). Isochronism is therefore exactly as good as the linearization and no better. Retaining the cubic term makes the period grow with the angular amplitude θ_0 as $T(\theta_0) = T\,(1 + \theta_0^2/16 + O(\theta_0^4))$, derived in Theorem 28.72 and already about one part in 10^3 at an amplitude of 7° — the failure taken up in Section 28.7.1. □

Example 28.10 (The seconds pendulum)**.** A pendulum of period $T = 2\,\mathrm{s}$ — one second to each swing, the beat of a longcase clock — requires, by Equation (28.7), a length

$$\ell = \frac{gT^2}{4\pi^2} = \frac{9.806\,65\,\mathrm{m/s^2}\times 4\,\mathrm{s^2}}{4\pi^2} = 0.9936\,\mathrm{m}\,.$$

The number is close enough to one metre that the seconds pendulum was a serious candidate for the definition of the metre in the eighteenth century, and it was rejected for a reason that belongs to this chapter: ℓ depends on g, which varies over the Earth by some 0.5 %, so the standard would have differed from place to place. That same sensitivity is what makes the pendulum an instrument (Proposition 28.11, Chapter 34).

Rests on Equation (28.7).

Proposition 28.11 (The physical pendulum and its centre of oscillation)**.** *A rigid body of mass m suspended from a pivot a distance d from its centre of mass, with moment of inertia I about that pivot, oscillates for small displacements at*

$$\omega_0^2 = \frac{mgd}{I}\,, \qquad \ell_{eq} := \frac{I}{md}\,, \tag{28.9}$$

with ℓ_{eq} the length of the simple pendulum of the same period. If $I_{cm} = mb^2$ is the moment of inertia about the centre of mass, then $\ell_{eq} = d + b^2/d$, and suspension from the point at distance b^2/d on the far side of the centre of mass gives the same period: the pivot and that point are interchangeable.

Rests on Equations (28.2) and (28.8).

Proof. *Derives Proposition 28.11.*

The moment of the weight about the pivot is $-mgd\sin\theta$ and the angular momentum about it is $I\dot{\theta}$, so the rotational law of Chapter 19 gives $I\ddot{\theta} = -mgd\sin\theta$; linearizing as in Phenomenon 28.9 yields the first of Equation (28.9), and comparison with $\omega_0^2 = g/\ell$ gives the second. The parallel-axis theorem of Chapter 29 gives $I = m(b^2 + d^2)$, whence $\ell_{\mathrm{eq}} = (b^2+d^2)/d = d+b^2/d$. Suspending instead at $d' = b^2/d$ gives $\ell'_{\mathrm{eq}} = d'+b^2/d' = b^2/d+d$, the same number. □

The interchange is Huygens' reciprocity [11], and it is the working principle of Kater's reversible pendulum: adjust the suspension until the two periods agree, measure the distance between the knife edges — which is then ℓ_{eq}, needing no knowledge of I or of the mass distribution — and read g off Equation (28.7). Chapter 34 reports the measurement.

Proposition 28.12 (Huygens' cycloidal pendulum is exactly isochronous)**.** *A particle sliding without friction under gravity on the cycloid*

$$x = a\,(\phi + \sin\phi)\,, \qquad y = a\,(1 - \cos\phi)\,, \qquad |\phi| \le \pi\,, \tag{28.10}$$

with y measured upward, so that $\phi = 0$ is the lowest point, oscillates with period $T = 2\pi\sqrt{4a/g}$ for every *amplitude.*

Rests on Equations (28.2) and (28.8).

Proof. *Derives Proposition 28.12.*

Parametrize the curve by arc length s measured from the lowest point, $\phi = 0$. From Equation (28.10), $\mathrm{d}x = a(1 + \cos\phi)\mathrm{d}\phi$ and $\mathrm{d}y = a\sin\phi\,\mathrm{d}\phi$, so

$$\mathrm{d}s^2 = a^2\left[(1+\cos\phi)^2 + \sin^2\phi\right]\mathrm{d}\phi^2 = 2a^2\,(1+\cos\phi)\,\mathrm{d}\phi^2 = 4a^2\cos^2\!\left(\frac{\phi}{2}\right)\mathrm{d}\phi^2\,,$$

whence $\mathrm{d}s = 2a\cos(\phi/2)\mathrm{d}\phi$ and $s = 4a\sin(\phi/2)$, the cosine being non-negative on $|\phi| \le \pi$. The height above the lowest point is $h = y = a(1-\cos\phi) = 2a\sin^2(\phi/2)$, and eliminating ϕ in favour of s gives $h = s^2/8a$. The potential energy is therefore $V = mgh = mgs^2/8a$, exactly quadratic in the arc length, and the kinetic energy is $\frac{1}{2}m\dot{s}^2$ because s is arc length. So the motion in s is Equation (28.2) with $\omega_0^2 = g/4a$, with no approximation anywhere, and Proposition 28.4 gives a period independent of amplitude. □

The cycloid is the same curve that solves the brachistochrone problem of Proposition 16.1 — one curve answering two questions that have no obvious relation, which is why the coincidence was worth Bernoulli's astonishment. Huygens made a pendulum trace it by hanging the bob between two cycloidal cheeks [11]; the device failed in practice because the string's friction against the cheeks cost more accuracy than the residual anharmonicity of a plain small-arc pendulum cost.

28.1.3 Universality of the harmonic approximation

Phenomenon 28.13 (Universality of small oscillations)**.** Systems that have nothing else in common oscillate sinusoidally when disturbed slightly from a stable equilibrium, and each does so at one frequency fixed by the system and not by the size of the disturbance: a mass on a spring, a tuning fork, a diatomic molecule, an atom in a crystal, a ship rolling, a bead in a bowl. The list is not a coincidence of construction but a consequence of the disturbance being small, and it is why one model is reused throughout this treatise — for lattice vibrations in Chapter 125, for the modes of the electromagnetic field in Chapter 61, and for the field quanta of Chapter 64 [1585, 2027].

Rests on Theorem 7.38 and Equation (28.2).

Derivation. *Derives Phenomenon 28.13.*

Let a system of one degree of freedom have potential energy $V(x)$, twice continuously differentiable, with an equilibrium at x_0, so that $V'(x_0) = 0$, and let that equilibrium be stable, so that $V''(x_0) > 0$. Taylor's theorem (Theorem 7.38) gives

$$V(x) = V(x_0) + \tfrac{1}{2}V''(x_0)\,(x - x_0)^2 + O\!\left((x-x_0)^3\right)\,, \tag{28.11}$$

the linear term being absent precisely because x_0 is an equilibrium. Writing $u := x - x_0$ and keeping the quadratic term, the force is $F = -\frac{\mathrm{d}V}{\mathrm{d}x} = -V''(x_0)\,u$ and Newton's second law becomes

$$m\ddot{u} = -V''(x_0)\,u\,, \qquad \omega_0^2 = \frac{V''(x_0)}{m}\,, \tag{28.12}$$

a harmonic equation with a real frequency exactly because $V''(x_0) > 0$. Nothing about the system entered except the two numbers m and $V''(x_0)$, which is the universality asserted; and

the approximation is controlled by the neglected cubic term, so it degrades as the amplitude grows (Section 28.7). The two excluded cases are instructive. Where $V''(x_0) < 0$ the same expansion gives an exponentially growing solution, which is what makes the equilibrium unstable; where $V''(x_0) = 0$ the leading behaviour is quartic or worse and is not harmonic at all, so the period depends on the amplitude however small the amplitude is made. □

Remark 28.14 (When the approximation fails). Equation (28.11) is useful only while the cubic term is small against the quadratic one over the range actually explored. For an amplitude A that condition reads

$$\frac{|V'''(x_0)|\,A}{3\,V''(x_0)} \ll 1\,, \tag{28.13}$$

a purely local statement with a purely local consequence: the fractional shift of the period is of the same order as the square of the left-hand side (Theorem 28.72). It says nothing about whether the motion remains bounded, and it fails outright at an inflection of V, where $V''(x_0) = 0$ and there is no harmonic regime at any amplitude.

Rests on Equation (28.11).

Example 28.15 (A chemical bond as a spring)**.** The stretching vibration of a diatomic molecule is Equation (28.12) with the reduced mass $\mu = m_1 m_2/(m_1 + m_2)$ in place of m, because the two-body problem separates into a centre-of-mass motion and a relative motion (Chapter 27). For $^1\mathrm{H}^{35}\mathrm{Cl}$ the harmonic constant of the vibrational ladder is $\tilde{\nu}_e \approx 2991/\mathrm{cm}$ [961], that is

$$\nu_e = c\,\tilde{\nu}_e = 8.966 \times 10^{13}\,\mathrm{Hz}\,, \qquad \omega_e = 2\pi\nu_e = 5.634 \times 10^{14}\,\mathrm{rad/s}\,.$$

With $m_1 = 1.0078\,u$ and $m_2 = 34.9689\,u$ the reduced mass is $0.9796\,u$, and with the atomic mass constant $u = (1.660\,539\,068\,92 \pm 0.000\,000\,000\,52) \times 10^{-27}\,\mathrm{kg}$ [2431] this is $\mu = 1.6266 \times 10^{-27}\,\mathrm{kg}$. Hence

$$k = \mu\,\omega_e^2 = 1.6266 \times 10^{-27}\,\mathrm{kg} \times 3.174 \times 10^{29}/\mathrm{s}^2 = 516\,\mathrm{N/m}\,. \tag{28.14}$$

A hydrogen–chlorine bond is a spring ten times stiffer than the laboratory spring of Example 28.8, and the entire difference in their behaviour is the factor 10^{27} between the two masses.

Rests on Equation (28.12).

Remark 28.16 (The same molecule measures the anharmonicity). $\tilde{\nu}_e$ in Example 28.15 is not what a spectrometer reads. The observed fundamental band of $^1\mathrm{H}^{35}\mathrm{Cl}$ lies near $2885/\mathrm{cm}$; $\tilde{\nu}_e$ is the harmonic frequency extrapolated from the ladder of levels, and the difference, $\tilde{\nu}_e - 2\tilde{\nu}_e x_e$ with $\tilde{\nu}_e x_e \approx 52.8/\mathrm{cm}$ [961], is the anharmonicity of the bond. It is 3.5 % — small, which is why Example 28.15 is worth doing, and not zero, which is why Section 28.7 exists. A purely harmonic bond could never break; the observed dissociation of molecules is the same cubic term seen at large amplitude (Chapter 88).

Rests on Equations (28.13) *and* (28.14).

28.2 Damped and driven oscillations

28.2.1 Linear damping

Rayleigh's systematic treatment of dissipation in vibrating systems [198] takes the friction to be linear in the velocity, which is the first term of an expansion in $\dot{x}$ exactly as Equation (28.11) is the first term of an expansion in x.

Definition 28.17 (Damped oscillator)**.**

$$\ddot{x} + 2\beta\dot{x} + \omega_0^2 x = 0\,, \tag{28.15}$$

with $\beta \geq 0$ the *damping rate*, of SI dimension 1/s, arising from a resistive force $-2m\beta\dot{x}$.
Rests on Equation (28.2)*.*

Proposition 28.18 (The three regimes)**.** *The solutions of Equation* (28.15) *are, according to the sign of* $\beta^2 - \omega_0^2$*,*

$$\beta < \omega_0: \quad x = A\mathrm{e}^{-\beta t}\cos(\omega_d t + \varphi)\,, \qquad \omega_d = \sqrt{\omega_0^2 - \beta^2}\,, \tag{28.16}$$

$$\beta = \omega_0: \quad x = (C_1 + C_2 t)\,\mathrm{e}^{-\omega_0 t}\,, \tag{28.17}$$

$$\beta > \omega_0: \quad x = C_+\mathrm{e}^{\lambda_+ t} + C_-\mathrm{e}^{\lambda_- t}\,, \qquad \lambda_\pm = -\beta \pm \sqrt{\beta^2 - \omega_0^2}\,. \tag{28.18}$$

Only the first oscillates; all three decay, and none reaches zero in finite time.
Rests on Equation (28.15) *and Theorem 9.13.*

Proof. *Derives Proposition 28.18.*

Substituting $x = \mathrm{e}^{\lambda t}$ into Equation (28.15) gives the characteristic equation $\lambda^2 + 2\beta\lambda + \omega_0^2 = 0$ with roots $\lambda_\pm = -\beta \pm \sqrt{\beta^2 - \omega_0^2}$. For $\beta < \omega_0$ the roots are $-\beta \pm \mathrm{i}\omega_d$ and the real combinations give Equation (28.16). For $\beta > \omega_0$ both roots are real and negative, since $\sqrt{\beta^2 - \omega_0^2} < \beta$, giving Equation (28.18). At $\beta = \omega_0$ the roots coincide and $\mathrm{e}^{-\omega_0 t}$ alone spans only one dimension of the two-dimensional solution space of Theorem 9.13; a second, independent solution is $t\mathrm{e}^{-\omega_0 t}$, as substitution confirms and as the Wronskian $\mathrm{e}^{-2\omega_0 t} \neq 0$ certifies (Proposition 9.16). Each exponential has strictly negative real part whenever $\beta > 0$, so every solution decays; and an exponential vanishes nowhere, while Equation (28.17) and Equation (28.18) can each cross zero at most once. □

Definition 28.19 (Logarithmic decrement and quality factor)**.** For the underdamped case, the *logarithmic decrement* is the natural logarithm of the ratio of successive maxima,

$$\delta_{\log} = \ln\frac{x(t)}{x\,(t + 2\pi/\omega_d)} = \frac{2\pi\beta}{\omega_d}\,, \tag{28.19}$$

and the *quality factor* is

$$Q := \frac{\omega_0}{2\beta}\,, \tag{28.20}$$

a dimensionless number.
Rests on Equation (28.16)*.*

Proposition 28.20 (Q counts stored energy against energy lost)**.** *For an underdamped oscillator the energy obeys* $E(t) = E_0\mathrm{e}^{-2\beta t}$ *up to terms oscillating at* $2\omega_d$*, and for* $Q \gg 1$

$$Q = 2\pi\,\frac{\textit{energy stored}}{\textit{energy dissipated per cycle}} = \frac{\omega_0}{2\beta}\,, \tag{28.21}$$

so that the motion persists for about Q/π *cycles.*
Rests on Equations (28.5) *and* (28.20)*.*

Proof. *Derives Proposition 28.20.*

Differentiating $E = \tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{2}m\omega_0^2 x^2$ along Equation (28.15),

$$\frac{\mathrm{d}E}{\mathrm{d}t} = m\dot{x}\left(\ddot{x} + \omega_0^2 x\right) = -2m\beta\dot{x}^2\,, \tag{28.22}$$

which is the power of the resistive force and is never positive. Over one cycle $\langle\dot{x}^2\rangle = E/m$ by Proposition 28.5 applied to the slowly decaying envelope, so $\langle \mathrm{d}E/\mathrm{d}t\rangle = -2\beta E$ and $E = E_0\mathrm{e}^{-2\beta t}$. The fractional loss per cycle is then $\Delta E/E = 2\beta\cdot 2\pi/\omega_d \approx 4\pi\beta/\omega_0$ for $Q \gg 1$, and $2\pi E/\Delta E = \omega_0/2\beta = Q$. The amplitude envelope falls by e^{-1} in the time $1/\beta = 2Q/\omega_0$, which is Q/π periods; the energy, decaying twice as fast, does so in $Q/2\pi$ periods. □

Proposition 28.21 (Critical damping returns fastest)**.** *Among all $\beta \geq 0$, the slowest-decaying exponential in Proposition 28.18 has the largest decay rate exactly at $\beta = \omega_0$, where that rate is ω_0.*

Rests on Equations (28.17) *and* (28.18)*.*

Proof. *Derives Proposition 28.21.*

For $\beta < \omega_0$ every solution is bounded by const $\times\, \mathrm{e}^{-\beta t}$ and no better, so the rate is $\beta < \omega_0$. For $\beta > \omega_0$ the slower of the two exponentials in Equation (28.18) decays at

$$|\lambda_+| = \beta - \sqrt{\beta^2 - \omega_0^2} = \frac{\omega_0^2}{\beta + \sqrt{\beta^2 - \omega_0^2}} < \omega_0 \,,$$

a strictly decreasing function of β on (ω_0, ∞). At $\beta = \omega_0$ the rate is ω_0 itself, larger than either branch, the algebraic factor t in Equation (28.17) being no obstacle since $t\mathrm{e}^{-\omega_0 t} \to 0$ faster than $\mathrm{e}^{-\beta' t}$ for any $\beta' < \omega_0$. □

This is why a galvanometer, a door closer and a car suspension are all built close to critical damping: overdamping is not the safe side of the design, it is slower.

Remark 28.22 (The observed range of Q)*.* Quality factors are among the widest-ranging numbers in physics, and Equation (28.21) is what makes them measurable in three independent ways: from the ringdown time, from the fractional energy lost per cycle, and — next — from the width of the resonance curve. That the three agree is the content of the definition, and it is checked every time a resonator is characterized. A tapped wine glass rings for a second or two at a few hundred hertz, which by $Q = \pi f \tau$ is $Q \sim 10^3$; a quartz timekeeping crystal is several orders higher, and the caesium hyperfine resonance on which the SI second is founded [1064] higher still. Individual figures of that kind are engineering data for which this treatise cites no primary source, and none of them is used below. The name and the engineering history of the symbol are recorded by Green [1071].

Rests on Equation (28.21)*.*

28.2.2 The driven oscillator and resonance

Phenomenon 28.23 (Resonance)**.** A lightly damped oscillator driven by a periodic force of fixed amplitude responds far more strongly when the driving frequency is near its own than when it is far away, and the sharper the oscillator — the smaller its damping — the larger that response and the narrower the band of frequencies over which it occurs. A wine glass shatters at one note and not at its neighbours; a radio receiver selects one station out of thousands; a struck string sets an identically tuned string sounding across the room while a mistuned one stays silent, an effect Galileo already describes and attributes to the accumulation of correctly timed impulses [4]. On resonance the response also lags the drive by a quarter of a cycle [1585].

Rests on Equations (28.15) *and* (28.20)*.*

Derivation. *Derives Phenomenon 28.23.*

Take the linearly damped, sinusoidally driven oscillator

$$\ddot{x} + 2\beta\dot{x} + \omega_0^2 x = \frac{F_0}{m}\cos\omega t \,, \tag{28.23}$$

and seek the steady state $x = A(\omega)\cos(\omega t - \delta)$. Substituting and collecting the coefficients of $\cos\omega t$ and $\sin\omega t$ gives

$$A(\omega) = \frac{F_0/m}{\sqrt{\left(\omega_0^2 - \omega^2\right)^2 + 4\beta^2\omega^2}} \,, \qquad \tan\delta = \frac{2\beta\omega}{\omega_0^2 - \omega^2} \,. \tag{28.24}$$

At $\omega = \omega_0$ the first bracket vanishes and $A(\omega_0) = F_0/(2m\beta\omega_0)$, whereas the static response, $\omega = 0$, is $A(0) = F_0/(m\omega_0^2)$. Their ratio is

$$\frac{A(\omega_0)}{A(0)} = \frac{\omega_0}{2\beta} = Q\,, \tag{28.25}$$

the quality factor: a resonator with $Q = 10^4$ answers a drive at its own frequency ten thousand times more strongly than it answers the same drive applied slowly. The denominator of $\tan\delta$ vanishes at $\omega = \omega_0$, so $\delta = \pi/2$ there, which is the quarter-cycle lag. Finally A^2 falls to half its peak where $\left(\omega_0^2 - \omega^2\right)^2 = 4\beta^2\omega^2$, that is $|\omega_0 - \omega| \simeq \beta$ for light damping, so the full width of the resonance curve is $\Delta\omega \simeq 2\beta = \omega_0/Q$. Gain and selectivity are thus the same number: a sharp resonator is necessarily a narrow filter, and a broadband one is necessarily insensitive. □

The steady state is only half the solution. Equation (28.23) is inhomogeneous and linear, so its general solution is the steady state plus an arbitrary solution of Equation (28.15); the latter is the *transient*, it decays as $\mathrm{e}^{-\beta t}$, and it is what carries the initial data. A resonator therefore takes about Q/π cycles to reach the response Equation (28.24) — the same Q again, and the reason a high-Q filter is a slow filter.

Proposition 28.24 (Impulse response of the damped oscillator)**.** *The solution of* $m(\ddot{x} + 2\beta\dot{x} + \omega_0^2 x) = F(t)$ *that vanishes before the force acts is*

$$x(t) = \int_{-\infty}^{t} G(t - t')\, F(t')\, \mathrm{d}t'\,, \qquad G(\tau) = \frac{\mathrm{e}^{-\beta\tau}\sin\omega_d\tau}{m\,\omega_d} \quad (\tau > 0)\,, \tag{28.26}$$

and $G(\tau) = 0$ *for* $\tau < 0$. *G has SI dimension* $\mathrm{m/(N\,s)}$.
Rests on Equation (28.23) and Definition 10.42.

Proof. *Derives Proposition 28.24.*
This is the causal Green's function of Definition 10.42 for the operator on the left, and Equation (28.26) is Duhamel's principle (Theorem 10.52) for it; only G itself has to be identified. An impulse $F = P\delta(\tau)$ leaves x continuous and jumps $\dot{x}$ by P/m, by integrating the equation across $\tau = 0$; the solution of the homogeneous equation with $x(0) = 0$ and $\dot{x}(0) = P/m$ is, from Equation (28.16), $x = (P/m\omega_d)\mathrm{e}^{-\beta\tau}\sin\omega_d\tau$. Dividing by P gives G. Vanishing for $\tau < 0$ is the causality requirement, which selects the retarded solution of Definition 10.50. The dimension follows from displacement per unit impulse per unit time. □

Equation (28.26) makes the oscillator a linear translation-invariant system in the sense of Proposition 17.52, so its response to *any* drive is a convolution and, by the convolution theorem Theorem 17.53, a multiplication in frequency. That multiplier is the susceptibility.

Definition 28.25 (Complex susceptibility)**.** Writing the drive as the real part of $F_0\mathrm{e}^{-\mathrm{i}\omega t}$ and the response as the real part of $\chi(\omega)F_0\mathrm{e}^{-\mathrm{i}\omega t}$,

$$\chi(\omega) = \frac{1}{m\left(\omega_0^2 - \omega^2 - 2\mathrm{i}\beta\omega\right)}\,, \tag{28.27}$$

of SI dimension m/N, with $|\chi| = A/F_0$ and $\arg\chi = \delta$ from Equation (28.24), so that the real part of $\chi F_0\mathrm{e}^{-\mathrm{i}\omega t}$ is $A\cos(\omega t - \delta)$.
Rests on Equation (28.24) and Proposition 17.52.

Remark 28.26 (Causality is visible in the susceptibility). The poles of Equation (28.27) sit at $\omega = -\mathrm{i}\beta \pm \omega_d$, both in the lower half of the complex ω plane, so χ is holomorphic in the upper half plane. That is exactly the analyticity that Theorem 17.78 converts into a dispersion

relation between $\operatorname{Re}\chi$ and $\operatorname{Im}\chi$: the in-phase and out-of-phase parts of the response are not independent, and measuring the absorption of a resonator over all frequencies determines its refraction. The same relation, for the same reason, governs the optical response of matter (Chapter 62) and the scattering amplitudes of Chapter 86. A damping $\beta > 0$ is what pushes the poles off the real axis; an undamped oscillator sits exactly at the boundary, which is why its driven response diverges.

Rests on Equation (28.27) and Theorem 17.78.

Proposition 28.27 (Three resonant frequencies, and only one of them is exact)**.** *The cycle-averaged power delivered by the drive of Equation (28.23) is*

$$\bar{P}(\omega) = \frac{F_0^2 \beta\, \omega^2}{m\left[\left(\omega_0^2 - \omega^2\right)^2 + 4\beta^2\omega^2\right]}, \tag{28.28}$$

which attains its maximum $\bar{P}(\omega_0) = F_0^2/(4m\beta)$ at $\omega = \omega_0$ exactly, and whose two half-power frequencies differ by exactly $2\beta = \omega_0/Q$ — both statements holding for every β, not only in the lightly damped limit in which Equation (28.28) becomes a Lorentzian. The amplitude *peaks instead at $\omega_r = \sqrt{\omega_0^2 - 2\beta^2}$, which exists only for $\beta < \omega_0/\sqrt{2}$.*

Rests on Equations (28.20) and (28.24).

Proof. *Derives Proposition 28.27.*

With $F = F_0 \cos\omega t$ and $\dot{x} = -A\omega \sin(\omega t - \delta)$, the average of $\cos\omega t \sin(\omega t - \delta)$ over a period is $-\frac{1}{2}\sin\delta$, so $\bar{P} = \langle F\dot{x}\rangle = \frac{1}{2}F_0 A\omega \sin\delta$. Substituting A and $\sin\delta = 2\beta\omega/\sqrt{(\omega_0^2-\omega^2)^2 + 4\beta^2\omega^2}$ from Equation (28.24) gives Equation (28.28). Writing D^2 for the bracket, $\bar{P}$ is half its peak where $\beta\omega^2/D^2 = 1/8\beta$, that is $\left(\omega_0^2 - \omega^2\right)^2 = 4\beta^2\omega^2$, or $\omega^2 \pm 2\beta\omega - \omega_0^2 = 0$; the two positive roots are $\sqrt{\beta^2 + \omega_0^2} \mp \beta$ and their difference is 2β with no approximation. For the amplitude, differentiating D^2 and setting the result to zero gives $-4\omega(\omega_0^2 - \omega^2) + 8\beta^2\omega = 0$, hence $\omega_r^2 = \omega_0^2 - 2\beta^2$. Finally, maximizing the velocity amplitude ωA means maximizing ω^2/D^2, whose derivative in $u = \omega^2$ has numerator $(\omega_0^2 - u)\,(\omega_0^2 + u)$; this vanishes at $u = \omega_0^2$ exactly. □

Three "resonant frequencies" therefore have to be kept apart, and they coincide only in the limit of vanishing damping: the power and the velocity peak exactly at ω_0, the displacement peaks slightly below it, and the free ringing runs at $\omega_d = \sqrt{\omega_0^2 - \beta^2}$, lower still. The distinction is invisible at $Q = 10^4$ and unavoidable at $Q = 2$. The Laplace-transform route to the same steady state, with the transient included, is worked in Example 17.75.

28.2.3 Parametric resonance

A pendulum can also be excited without applying any force to it, by varying one of its parameters in time — shortening and lengthening the suspension, or modulating the stiffness. Faraday saw the effect first, in the standing ripples raised on a liquid surface by a vibrating support, and recorded the observation that identifies it: the ripples oscillate at *half* the frequency of the support [88]. The governing equation is Mathieu's [173].

Phenomenon 28.28 (Parametric excitation at half the drive frequency)**.** An oscillator whose stiffness is modulated periodically, with no applied force at all, can grow from rest — and it does so most readily when the modulation runs at *twice* its natural frequency, responding at half the modulation frequency. A liquid surface shaken vertically answers with standing ripples that complete one cycle for every two of the shaking [88]; a child on a swing rises and falls twice per swing and not once; and unlike ordinary resonance the growth is exponential rather than linear in time, and sets in only when the modulation exceeds a threshold fixed by the damping.

Rests on Equations (28.15) and (28.20).

Derivation. *Derives Phenomenon 28.28.*

Modulating the stiffness of Equation (28.15) gives Mathieu's equation with damping,

$$\ddot{x} + 2\beta\dot{x} + \omega_0^2\,(1 + h\cos\gamma t)\,x = 0\,, \qquad 0 < h \ll 1\,. \tag{28.29}$$

The damping can be removed exactly. Putting $x = \mathrm{e}^{-\beta t}y$ turns Equation (28.29) into a Mathieu equation with no first derivative:

$$\ddot{y} + \left[\omega_0^2 - \beta^2 + h\omega_0^2\cos\gamma t\right] y = 0\,. \tag{28.30}$$

From Equation (28.29) (substituting $x = \mathrm{e}^{-\beta t}y$ and cancelling the common exponential).

Hence x grows if and only if y grows faster than $\mathrm{e}^{\beta t}$. Take $\beta \ll \omega_0$ and set $\gamma = 2\omega_0 + \varepsilon$ with $|\varepsilon| \ll \omega_0$. Look for a response at half the modulation frequency, $y = a(t)\cos(\gamma t/2) + b(t)\sin(\gamma t/2)$, with a and b varying slowly enough that $\ddot{a}$ and $\ddot{b}$ may be dropped. Substituting, using $\cos\gamma t\cos(\gamma t/2) = \frac{1}{2}\left[\cos(\gamma t/2) + \cos(3\gamma t/2)\right]$ and $\cos\gamma t\sin(\gamma t/2) = \frac{1}{2}\left[\sin(3\gamma t/2) - \sin(\gamma t/2)\right]$, and discarding the non-resonant third harmonics, the coefficients of $\cos(\gamma t/2)$ and $\sin(\gamma t/2)$ give, with $\omega_0^2 - \gamma^2/4 \approx -\omega_0\varepsilon$ and $\gamma \approx 2\omega_0$,

$$\dot{a} = -\tfrac{1}{2}\left(\varepsilon + \tfrac{1}{2}h\omega_0\right) b\,, \qquad \dot{b} = +\tfrac{1}{2}\left(\varepsilon - \tfrac{1}{2}h\omega_0\right) a\,. \tag{28.31}$$

Eliminating b gives $\ddot{a} = s^2 a$ with

$$s = \tfrac{1}{2}\sqrt{\left(\tfrac{1}{2}h\omega_0\right)^2 - \varepsilon^2}\,, \tag{28.32}$$

real — and the amplitude therefore exponentially growing — exactly in the band $|\varepsilon| < h\omega_0/2$ centred on $\gamma = 2\omega_0$. Restoring the factor $\mathrm{e}^{-\beta t}$, the motion grows only if $s > \beta$, which at the centre of the band ($\varepsilon = 0$, $s = h\omega_0/4$) reads

$$h > \frac{4\beta}{\omega_0} = \frac{2}{Q}\,. \tag{28.33}$$

Every feature asserted follows: the response is at $\gamma/2$ by construction and could not have been at γ, since a term $\cos\gamma t$ multiplying x produces sum and difference frequencies and only the half-frequency component feeds back on itself; the growth is exponential, not the linear-in-t growth of an undamped forced oscillator; and there is a threshold, absent from ordinary resonance, because the modulation multiplies the amplitude instead of adding to it. □

Remark 28.29 (Part II owes Floquet theory). Equation (28.32) is a leading-order result about one instability band. The exact statement — that a linear system with coefficients of period $2\pi/\gamma$ has solutions of the form e^{st} times a function of that same period, so that stability is decided by the eigenvalues of one monodromy matrix, and that the (γ, h) plane is therefore divided into tongues of instability emanating from $\gamma = 2\omega_0/n$ — is Floquet's theorem, a general result about ordinary differential equations with periodic coefficients. It belongs in Chapter 9 alongside the Wronskian theory it extends, and Part II does not at present contain it, nor the Mathieu functions themselves alongside the Bessel, Legendre, Hermite and Laguerre families of that chapter. Nothing above depends on the missing statement: the derivation of Phenomenon 28.28 uses only the slowly-varying-amplitude argument written out there, and the tongue structure is asserted nowhere in this treatise.

Rests on Equation (28.29).

Parametric excitation is not a curiosity. It is how a swing is pumped — the rider raises and lowers the centre of mass twice per swing, modulating ℓ in Equation (28.7) at $2\omega_0$; it is the mechanism of the parametric amplifier, which achieves gain by modulating a reactance rather than by dissipating power in an active element; and it is the standard tool for driving a mechanical resonator without exerting a force on it. The nonlinear saturation that arrests the exponential growth of Equation (28.32), and the fate of the motion once the amplitude leaves the small-oscillation regime, belong to Chapter 32.

28.3 Coupled oscillators and normal modes

28.3.1 Two coupled oscillators and beats

Phenomenon 28.30 (Beats and the periodic exchange of energy)**.** Two tones of nearly equal frequency sounded together are heard not as two tones but as one, of the mean frequency, whose loudness rises and falls at the difference of the two — slowly enough to count when the two are within a few hertz. The mechanical counterpart is as striking: two identical pendulums joined by a weak spring, one started and the other at rest, exchange their motion completely and periodically, the second growing as the first dies away and back again, indefinitely [2027]. Piano tuners set unisons by counting beats, which measures a frequency difference with a stopwatch and no frequency standard at all.

Rests on Equations (28.4) *and* (28.36)*.*

Derivation. *Derives Phenomenon 28.30.*

Superpose two harmonic motions of equal amplitude and frequencies ω_1 and ω_2, and use the sum-to-product identity

$$\cos\omega_1 t + \cos\omega_2 t = 2\cos\left(\frac{\omega_1-\omega_2}{2}\,t\right)\cos\left(\frac{\omega_1+\omega_2}{2}\,t\right)\,. \tag{28.34}$$

The second factor is a rapid oscillation at the mean frequency; the first is a slow one multiplying it. The ear cannot follow the sign of the slow factor, only the size of the envelope, and |cos| has half the period of cos, so the loudness is heard to vary at $|\omega_1-\omega_2|$ — twice the frequency of the slow factor itself. For the coupled pendulums, ω_1 and ω_2 are the frequencies $\omega_\pm$ of Equation (28.36); starting one pendulum alone excites the two modes with equal amplitude, and taking their sum and difference gives one bob proportional to the cosine of the slow factor and the other to its sine. Since $\cos^2+\sin^2=1$, the total energy is constant and passes entirely from one pendulum to the other and back, with period $2\pi/\,|\omega_+-\omega_-|$. The weaker the coupling, the closer the two mode frequencies and the slower the exchange — which is why the effect requires a *weak* spring to be visible at all. □

Proposition 28.31 (The two modes of a pair of coupled pendulums)**.** *Two pendulums of equal mass m and equal natural frequency ω_0, whose bobs are joined by a spring of constant κ, obey*

$$m\ddot{x}_1 = -m\omega_0^2 x_1 - \kappa\,(x_1-x_2)\,,\qquad m\ddot{x}_2 = -m\omega_0^2 x_2 - \kappa\,(x_2-x_1)\,, \tag{28.35}$$

and have exactly two normal modes,

$$x_1+x_2 \;\text{ at }\omega_- = \omega_0\,,\qquad x_1-x_2 \;\text{ at }\omega_+ = \sqrt{\omega_0^2+\frac{2\kappa}{m}}\,, \tag{28.36}$$

whose splitting for a weak spring is $\omega_+-\omega_-\approx\kappa/m\omega_0$.

Rests on Equations (28.2) *and* (28.12)*.*

Proof. *Derives Proposition 28.31.*

Adding and subtracting the two equations of Equation (28.35) gives $\ddot{s}=-\omega_0^2 s$ for $s=x_1+x_2$, the coupling cancelling because the spring exerts equal and opposite forces, and $\ddot{d}=-(\omega_0^2+2\kappa/m)d$ for $d=x_1-x_2$. Each is Equation (28.2), so Equation (28.36) follows from Proposition 28.4. For $\kappa\ll m\omega_0^2$, $\omega_+=\omega_0\sqrt{1+2\kappa/m\omega_0^2}\approx\omega_0+\kappa/m\omega_0$ by Theorem 7.38. □

The symmetric mode leaves the spring unstretched and therefore does not know it is there; the antisymmetric mode stretches it twice per swing and is pushed up in frequency. Two identical oscillators brought into contact always split this way, and the splitting

measures the coupling: this is the mechanical original of level repulsion, of the bonding and antibonding pair of a diatomic molecule (Chapter 88), and of the acoustic and optical branches of a diatomic lattice (Chapter 125). With $m = 0.10\,\mathrm{kg}$, $\omega_0 = 6.0\,\mathrm{rad/s}$ and $\kappa = 0.20\,\mathrm{N/m}$ the splitting is $0.33\,\mathrm{rad/s}$ and the energy crosses from one pendulum to the other in $\pi/(\omega_+ - \omega_-) = 9.4\,\mathrm{s}$ — some nine swings, which is what makes the demonstration watchable.

28.3.2 Small oscillations about equilibrium

Lagrange gave the general theory in the *Mécanique analytique* [40]: expand the Lagrangian to second order about a configuration of equilibrium and the problem becomes an algebraic one, solved once and for all.

Definition 28.32 (The two quadratic forms)**.** Let a system with n degrees of freedom $q_1, \dots, q_n$ have kinetic energy $T = \tfrac{1}{2}\sum a_{jk}(q)\dot q_j \dot q_k$ and potential energy $V(q)$, and let $q^{(0)}$ be an equilibrium, so that $\partial V/\partial q_j = 0$ there. Writing $\eta_j := q_j - q_j^{(0)}$ and keeping the leading term of each energy,

$$T = \tfrac{1}{2}\sum_{j,k} M_{jk}\dot\eta_j\dot\eta_k\,, \qquad V - V\big(q^{(0)}\big) = \tfrac{1}{2}\sum_{j,k} K_{jk}\eta_j\eta_k\,, \tag{28.37}$$

with the constant matrices $M_{jk} = a_{jk}(q^{(0)})$ — the *mass matrix*, symmetric and positive definite — and $K_{jk} = \big(\partial^2 V/\partial q_j \partial q_k\big)\,(q^{(0)})$ — the *stiffness matrix*, symmetric by Proposition 7.73.

Rests on Equation (28.11) and Definition 5.107.

M is positive definite because $T > 0$ for every nonzero velocity; K need not be, and whether it is decides the stability of the equilibrium. The Euler–Lagrange equations (Theorem 16.35) applied to $\mathcal{L} = T - V$ give the linear system

$$M\ddot{\boldsymbol\eta} + K\boldsymbol\eta = \mathbf{0}\,, \tag{28.38}$$

and substituting $\boldsymbol\eta = \boldsymbol u\cos(\omega t - \varphi)$ turns it into an algebraic problem:

$$\big(K - \omega^2 M\big)\,\boldsymbol u = \mathbf{0}\,, \qquad \det\big(K - \omega^2 M\big) = 0\,. \tag{28.39}$$

From Equation (28.37) (Euler–Lagrange for the quadratic Lagrangian, then the harmonic ansatz).

The second condition — the *secular equation* — is necessary and sufficient for a nonzero $\boldsymbol u$ to exist, and it is a polynomial of degree n in ω^2 whose leading coefficient $(-1)^n \det M$ is nonzero.

Theorem 28.33 (Normal modes)**.** *Let M be real symmetric positive definite and K real symmetric, both $n \times n$. Then*

1. *the secular equation Equation (28.39) has exactly n real roots $\omega_1^2 \le \dots \le \omega_n^2$, counted with multiplicity;*
2. *there is a basis $\boldsymbol u^{(1)}, \dots, \boldsymbol u^{(n)}$ of $\mathbb{R}^n$ with $\big(K - \omega_r^2 M\big)\,\boldsymbol u^{(r)} = \mathbf{0}$ and*

$$\big(\boldsymbol u^{(r)}\big)^{\mathsf T} M\,\boldsymbol u^{(s)} = \delta_{rs}\,, \qquad \big(\boldsymbol u^{(r)}\big)^{\mathsf T} K\,\boldsymbol u^{(s)} = \omega_r^2\,\delta_{rs}\,, \tag{28.40}$$

   so that the matrix U with these columns satisfies $U^{\mathsf T} M U = \mathbb{1}$ and $U^{\mathsf T} K U = \operatorname{diag}(\omega_1^2, \dots, \omega_n^2)$;
3. *every $\omega_r^2 > 0$ if and only if K is positive definite, that is, if and only if the equilibrium is a strict local minimum of V to quadratic order.*

Rests on Equation (28.39) and Theorems 5.75 and 5.112.

Proof. *Derives Theorem 28.33.*

Because M is symmetric and positive definite, the bilinear form $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_M := \boldsymbol{x}^{\mathsf{T}} M \boldsymbol{y}$ is an inner product on $\mathbb{R}^n$ in the sense of Definition 5.17. Set $A := M^{-1}K$, which exists because a positive-definite matrix is invertible. For all $\boldsymbol{x}, \boldsymbol{y}$,

$$\langle \boldsymbol{x}, A\boldsymbol{y} \rangle_M = \boldsymbol{x}^{\mathsf{T}} M M^{-1} K \boldsymbol{y} = \boldsymbol{x}^{\mathsf{T}} K \boldsymbol{y} = \boldsymbol{y}^{\mathsf{T}} K \boldsymbol{x} = \langle \boldsymbol{y}, A\boldsymbol{x} \rangle_M = \langle A\boldsymbol{x}, \boldsymbol{y} \rangle_M \,,$$

the middle step because a 1×1 matrix equals its own transpose and $K^{\mathsf{T}} = K$, the last because $\langle \cdot, \cdot \rangle_M$ is symmetric: A is self-adjoint with respect to it.

Theorem 5.75 is stated for a *complex* inner-product space, so apply it to the complexification. Extend the form to $\mathbb{C}^n$ sesquilinearly, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_M := \bar{\boldsymbol{x}}^{\mathsf{T}} M \boldsymbol{y}$, which is an inner product there because M is real, symmetric and positive definite; the computation above, with each transpose replaced by a conjugate transpose, again shows A self-adjoint for it, using only that M and K are real and symmetric. Theorem 5.75 then gives n real eigenvalues and an orthonormal eigenbasis of $\mathbb{C}^n$. Those eigenvectors may be taken real: each ω_r^2 is real and A is a real matrix, so $A - \omega_r^2 \mathbb{1}$ is real and its kernel is spanned by real vectors, and applying Gram–Schmidt (Proposition 5.26) in the M-inner product within each such real kernel — eigenspaces for distinct eigenvalues being already orthogonal — yields a real M-orthonormal eigenbasis of $\mathbb{R}^n$. Since $A\boldsymbol{u} = \omega^2 \boldsymbol{u}$ is equivalent to $(K - \omega^2 M)\,\boldsymbol{u} = \boldsymbol{0}$, and the eigenvalues of A are the roots of Equation (28.39) by Theorem 5.67, parts (i) and (ii) follow; the second identity of Equation (28.40) is the first one with $K\boldsymbol{u}^{(s)} = \omega_s^2 M \boldsymbol{u}^{(s)}$ substituted. For (iii), $U^{\mathsf{T}} K U = \operatorname{diag}(\omega_r^2)$ is a congruence, so by Sylvester's law of inertia Theorem 5.112 the number of positive, zero and negative ω_r^2 equals the corresponding count for K itself; K is positive definite exactly when all n are positive. □

Remark 28.34 (Part II owes the simultaneous diagonalization of two quadratic forms). Theorem 28.33 is a physics statement — it is about T and V — but parts (i) and (ii) rest on a purely linear-algebraic fact: *a pair of real symmetric bilinear forms, one of them positive definite, is simultaneously diagonalized by a single congruence*, equivalently the generalized eigenvalue problem $K\boldsymbol{u} = \lambda M \boldsymbol{u}$ has real eigenvalues and an M-orthonormal eigenbasis. That statement belongs in Chapter 5, beside Theorem 5.75, Theorem 5.112 and Theorem 5.76 — which covers commuting self-adjoint operators, a different hypothesis — and Part II does not at present state it. Two smaller pieces are missing with it. The *real* spectral theorem is one: Theorem 5.75 is stated over $\mathbb{C}$, so the real case a mechanics problem actually presents has to be reached by the complexification argument written into the proof above. The variational companion is the other: Theorem 16.76 characterizes the lowest eigenvalue of a weighted Sturm–Liouville problem, but the finite-dimensional Rayleigh quotient $\boldsymbol{u}^{\mathsf{T}} K \boldsymbol{u} / \boldsymbol{u}^{\mathsf{T}} M \boldsymbol{u}$ used in Proposition 28.38 is not stated there. Rather than prove general linear algebra inside a mechanics chapter, the derivation above exhibits the *reduction* to results Part II does have: the M-inner product, self-adjointness of $M^{-1}K$ in it, Theorem 5.75 unchanged, and Proposition 5.26 to come back to $\mathbb{R}^n$. Nothing in this chapter uses more than that.

Rests on Theorem 28.33.

Definition 28.35 (Normal coordinates)**.** The coordinates ξ_r defined by $\boldsymbol{\eta} = U\boldsymbol{\xi}$, that is $\eta_j = \sum_r U_{jr} \xi_r$, with U the matrix of Theorem 28.33.

Rests on Equation (28.40).

Proposition 28.36 (Normal coordinates decouple the motion)**.** *In normal coordinates*

$$T = \tfrac{1}{2} \sum_r \dot{\xi}_r^2 \,, \qquad V - V\big(q^{(0)}\big) = \tfrac{1}{2} \sum_r \omega_r^2 \xi_r^2 \,, \qquad \ddot{\xi}_r + \omega_r^2 \xi_r = 0 \,, \tag{28.41}$$

so the system is exactly n independent harmonic oscillators, and the total energy is the sum of their separate energies. Each ξ_r has SI dimension $\mathrm{kg}^{1/2}\,\mathrm{m}$.

Rests on Equations (28.38) *and* (28.40).

Proof. *Derives Proposition 28.36.*

Substituting $\boldsymbol{\eta} = U\boldsymbol{\xi}$ into Equation (28.37) gives $T = \tfrac{1}{2}\dot{\boldsymbol{\xi}}^{\mathsf{T}} U^{\mathsf{T}} M U \dot{\boldsymbol{\xi}}$ and the analogous expression for V, and Equation (28.40) reduces $U^{\mathsf{T}}MU$ to $\mathbb{1}$ and $U^{\mathsf{T}}KU$ to the diagonal matrix of the ω_r^2. The Euler–Lagrange equations of the resulting Lagrangian are the n separate equations shown, each of the form Equation (28.2), so by Proposition 28.5 each carries a conserved energy and the total is their sum. The dimension follows from T being an energy and $\dot{\xi}_r^2$ carrying the whole of it. □

Remark 28.37 (Normal coordinates are not lengths). The unit $\mathrm{kg}^{1/2}\,\mathrm{m}$ is not a slip. Absorbing M into the coordinates is what makes Equation (28.41) as simple as it is, and it costs the coordinates their status as displacements. If a displacement is wanted — to compare with a measured amplitude — it must be recovered as $\eta_j = \sum_r U_{jr}\xi_r$, in which the columns of U carry the compensating $1/\mathrm{kg}^{1/2}$. The same square root reappears wherever a mode is normalized this way, in particular in the lattice sums of Chapter 125 and in the field-mode expansions of Chapter 99, and it is one of the places where a book that keeps SI units earns its keep.

Rests on Equation (28.41).

Proposition 28.38 (Rayleigh's bound on the lowest frequency)**.** *For every $\boldsymbol{u} \neq \mathbf{0}$,*

$$\omega_1^2 \le \frac{\boldsymbol{u}^{\mathsf{T}} K \boldsymbol{u}}{\boldsymbol{u}^{\mathsf{T}} M \boldsymbol{u}} \le \omega_n^2 \,, \tag{28.42}$$

with equality on the left exactly for multiples of $\boldsymbol{u}^{(1)}$.

Rests on Equation (28.40) *and Definition 16.75.*

Proof. *Derives Proposition 28.38.*

Expand $\boldsymbol{u} = \sum_r c_r \boldsymbol{u}^{(r)}$ in the M-orthonormal basis of Theorem 28.33. Then $\boldsymbol{u}^{\mathsf{T}} M \boldsymbol{u} = \sum_r c_r^2$ and $\boldsymbol{u}^{\mathsf{T}} K \boldsymbol{u} = \sum_r \omega_r^2 c_r^2$ by Equation (28.40), so the quotient is a weighted average of the ω_r^2 with non-negative weights c_r^2, hence lies between the least and the greatest of them, and attains the least only when every c_r with $\omega_r^2 > \omega_1^2$ vanishes. □

This is the finite-dimensional face of Theorem 16.76, and it is what makes the Rayleigh–Ritz method of Section 16.5.3 work: any trial displacement whatever gives an upper bound on the fundamental frequency, and the bound is second-order accurate in the error of the trial vector because ω_1^2 is a stationary value. Corollary 16.78 is the statement in the continuum, where it is used to estimate the frequencies of structures that cannot be solved exactly.

Example 28.39 (The linear triatomic molecule)**.** Three particles on a line, of masses m, M, m at positions x_1, x_2, x_3, with springs of constant k joining each outer particle to the middle one, have

$$M_{\mathrm{mat}} = \operatorname{diag}(m, M, m)\,, \qquad K = k\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Guessing the modes from the symmetry rather than expanding the determinant: $\boldsymbol{u} = (1,1,1)$ gives $K\boldsymbol{u} = \mathbf{0}$, so $\omega^2 = 0$; $\boldsymbol{u} = (1,0,-1)$ gives $K\boldsymbol{u} = k(1,0,-1)$ and $M_{\mathrm{mat}}\boldsymbol{u} = m(1,0,-1)$, so $\omega^2 = k/m$; and the third mode, orthogonal in the M-metric to both, is $\boldsymbol{u} = (1, -2m/M, 1)$, for which

$$\omega_1^2 = 0\,, \qquad \omega_2^2 = \frac{k}{m}\,, \qquad \omega_3^2 = \frac{k}{m}\left(1 + \frac{2m}{M}\right). \tag{28.43}$$

The three modes are the translation of the whole molecule, the symmetric stretch in which the middle particle does not move, and the antisymmetric stretch in which it moves against the other two. The zero mode is not a defect of the model: V depends only on differences of the x_j, so it has a flat direction, and by Noether's theorem (Theorem 16.84, Corollary 16.88) that translation invariance is the conservation of total momentum. Every isolated molecule and every crystal carries such modes — three translations and, in general, three rotations — and they are what Theorem 28.33(iii) excludes when it asks for a *strict* minimum.

Rests on Theorem 28.33 and Corollary 16.88.

28.3.3 The loaded string and the continuum limit

Let N equal beads of mass m sit at spacing a on a light string of tension T stretched between fixed points, so that the total length is $L = (N+1)a$, and let y_n be the small transverse displacement of the n-th bead. This is Theorem 28.33 with a mass matrix already diagonal, and it is solvable in closed form for every N.

Proposition 28.40 (Modes and dispersion of the loaded string)**.** *With $y_0 = y_{N+1} = 0$, the equations of motion are*

$$m\ddot{y}_n = \frac{T}{a}\left(y_{n+1} - 2y_n + y_{n-1}\right), \qquad n = 1, \dots, N, \tag{28.44}$$

and the N normal modes and frequencies are

$$y_n^{(r)} = \sin\left(k_r n a\right), \quad k_r = \frac{r\pi}{L}, \qquad \omega_r = 2\,\omega_c \sin\!\left(\frac{k_r a}{2}\right), \quad \omega_c := \sqrt{\frac{T}{ma}}, \tag{28.45}$$

for $r = 1, \dots, N$.

Rests on Theorem 28.33 and Equation (28.38).

Proof. *Derives Proposition 28.40.*

For small displacements the string segment from bead n to bead $n+1$ makes an angle whose sine is $(y_{n+1} - y_n)/a$ to first order, and pulls bead n transversely with force $T(y_{n+1} - y_n)/a$; adding the pull of the segment on the other side gives Equation (28.44). Substituting $y_n = u \sin(kna)\mathrm{e}^{-\mathrm{i}\omega t}$, using $\sin(k(n+1)a) + \sin(k(n-1)a) = 2\cos(ka)\sin(kna)$, and cancelling the common factor $u \sin(kna)\mathrm{e}^{-\mathrm{i}\omega t}$,

$$-m\omega^2 = \frac{T}{a}\left(2\cos ka - 2\right) = -\frac{4T}{a}\sin^2\!\left(\frac{ka}{2}\right),$$

which is the dispersion relation in Equation (28.45). The form $\sin(kna)$ already satisfies $y_0 = 0$, and $y_{N+1} = 0$ requires $\sin(k(N+1)a) = 0$, that is $kL = r\pi$. Values r and $r + 2(N+1)$ give the same set of bead displacements, and $r = N+1$ gives none at all, so exactly $r = 1, \dots, N$ are distinct and non-trivial — N modes for N degrees of freedom, as Theorem 28.33 requires. □

Two features of Equation (28.45) outlive the model. First, the frequencies are *not* proportional to r: the chain is dispersive, and the sine bends the spectrum below the straight line $\omega = \omega_c k_r a$ that the low modes follow. Second, there is a highest frequency, $\omega_{\max} = 2\omega_c$, reached at $ka = \pi$; a chain of discrete masses cannot carry a vibration faster than that, because there is nothing between the beads to vibrate. Both features are the mechanical prototype of everything in Chapter 125: the cutoff is the Brillouin-zone edge, and the bending of $\omega(k)$ is what makes a crystal's heat capacity depart from the classical value.

Proposition 28.41 (The continuum limit)**.** *Let $N \to \infty$ and $a \to 0$ with $L = (N+1)a$ and the linear mass density $\mu := m/a$ held fixed. Then Equation (28.44) becomes the wave equation*

$$\frac{\partial^2 y}{\partial t^2} = v^2 \frac{\partial^2 y}{\partial x^2}, \qquad v = \sqrt{\frac{T}{\mu}}, \tag{28.46}$$

and Equation (28.45) becomes $\omega_r = v k_r$ with $k_r = r\pi/L$, for every fixed r.

Rests on Equations (28.44) and (28.45).

Proof. *Derives Proposition 28.41.*

Write $y_n(t) = y(na, t)$ for a twice continuously differentiable y. Taylor's theorem (Theorem 7.38) gives $y_{n\pm1} = y \pm a\,\partial_x y + \tfrac{1}{2}a^2\partial_x^2 y + O(a^3)$, so the bracket in Equation (28.44) is $a^2\partial_x^2 y + O(a^4)$, the odd orders cancelling. Dividing by $m = \mu a$,

$$\partial_t^2 y = \frac{T}{\mu a^2}\left[a^2\partial_x^2 y + O(a^4)\right] = \frac{T}{\mu}\,\partial_x^2 y + O(a^2),$$

which is Equation (28.46) in the limit. For the frequencies, fix r; then $k_r a = r\pi a/L \to 0$, so $2\omega_c \sin(k_r a/2) \to \omega_c k_r a = k_r\sqrt{T/\mu}$, using $\omega_c a = a\sqrt{T/(ma)} = \sqrt{Ta/m} = \sqrt{T/\mu}$. The limit is uniform only on the low modes: for r comparable with N the sine cannot be linearized, which is the statement that a continuum has no cutoff and a lattice does. □

28.4 The vibrating string and the wave equation

28.4.1 d'Alembert's equation and its general solution

Mersenne had established by measurement, a century before any wave theory, that the pitch of a stretched string varies inversely with its length, as the square root of the tension and inversely as the square root of its weight per unit length [2]. d'Alembert wrote down the equation those laws obey and solved it in complete generality [25]; it was the first partial differential equation of physics and the first to be solved.

Proposition 28.42 (The wave equation of a stretched string)**.** *A flexible string of linear mass density μ under tension T, displaced transversely by $y(x,t)$ with $|\partial_x y| \ll 1$, obeys Equation (28.46) with $v = \sqrt{T/\mu}$. The Lagrangian density whose Euler–Lagrange equation this is,*

$$\mathcal{L} = \tfrac{1}{2}\mu\,(\partial_t y)^2 - \tfrac{1}{2}T\,(\partial_x y)^2, \tag{28.47}$$

is the continuum limit of Equation (28.37).

Rests on Equation (28.46) and Theorem 16.36.

Proof. *Derives Proposition 28.42.*

Take the element between x and $x + \mathrm{d}x$, of mass $\mu\,\mathrm{d}x$. The tension pulls along the string at each end; its transverse component is $T\sin\theta \approx T\tan\theta = T\,\partial_x y$ to first order in the slope, and the net transverse force is the difference of that quantity at the two ends, $T\,\partial_x^2 y\,\mathrm{d}x$. Newton's second law gives $\mu\,\mathrm{d}x\,\partial_t^2 y = T\,\partial_x^2 y\,\mathrm{d}x$, which is Equation (28.46). The longitudinal components cancel to the same order, and the tension may be treated as constant along the string because the fractional stretch is $O((\partial_x y)^2)$. For the variational route, the kinetic energy of the element is $\tfrac{1}{2}\mu(\partial_t y)^2\mathrm{d}x$ and the work done in stretching it from $\mathrm{d}x$ to $\mathrm{d}x\sqrt{1+(\partial_x y)^2}$ against constant tension is $\tfrac{1}{2}T(\partial_x y)^2\mathrm{d}x$ to the same order; Theorem 16.36 applied to Equation (28.47) returns Equation (28.46). That Equation (28.47) is the limit of Equation (28.37) is Proposition 28.41 applied term by term. □

Everything about Equation (28.46) as a partial differential equation is Part II's. It is hyperbolic (Definition 10.6,Proposition 10.13), its characteristics are the lines $x \mp vt = \text{const}$ (Corollary 10.12), and its general solution is the pair of travelling waves

$$y(x,t) = f(x - vt) + g(x + vt), \tag{28.48}$$

with the initial-value problem solved explicitly by d'Alembert's formula Theorem 10.55. The physics is what those statements mean here: a disturbance travels at speed v in each direction without changing shape, the value of y at (x,t) depends only on the initial data in the interval $[x-vt, x+vt]$ (Corollary 10.56), and v is fixed by two measurable properties of the string and by nothing else. Mersenne's three laws are $v=\sqrt{T/\mu}$ read together with Equation (28.49). The string is worked as an example of separation of variables in Example 10.38.

Example 28.43 (A guitar string, in SI)**.** A steel string of diameter 0.30 mm and density $7.85\times10^{3}\,\mathrm{kg/m^3}$ has cross-section $\pi d^2/4 = 7.07\times10^{-8}\,\mathrm{m^2}$ and hence $\mu = 5.55\times10^{-4}\,\mathrm{kg/m}$. Sounding 440 Hz over a speaking length $L = 0.650\,\mathrm{m}$ requires, by Equation (28.49) with $n=1$, a wave speed $v = 2Lf_1 = 572\,\mathrm{m/s}$, and therefore a tension

$$T = \mu v^2 = 5.55\times10^{-4}\,\mathrm{kg/m}\times 3.27\times10^{5}\,\mathrm{m^2/s^2} = 182\,\mathrm{N}\,,$$

about the weight of 19 kg. Six such strings put more than a kilonewton on the neck of the instrument, which is why guitars are built the way they are. Note what does *not* enter: the amplitude of the plucking, and the material of the string except through μ.

Rests on Equations (28.46) *and* (28.49).

28.4.2 Superposition and the vibrating-string controversy

Proposition 28.44 (Superposition)**.** *If y_1 and y_2 solve Equation* (28.46) *then so does $c_1y_1+c_2y_2$ for any constants c_1, c_2; and if each satisfies the homogeneous boundary conditions $y(0,t)=y(L,t)=0$, so does the combination.*

Rests on Equation (28.46) *and Definition 10.2.*

Proof. *Derives Proposition 28.44.*

$\partial_t^2 - v^2\partial_x^2$ is a linear operator (Definition 10.2), so it takes a linear combination to the same combination of the images, each of which vanishes. The boundary conditions are likewise linear and homogeneous. □

The proposition is two lines; deciding what may be superposed took half a century. d'Alembert's solution Equation (28.48) says the motion is fixed by two "arbitrary functions"; Euler took that literally, admitting curves drawn freehand, while d'Alembert insisted a function had to be given by a single analytic expression. Daniel Bernoulli then claimed that every motion of the string is a superposition of the sinusoidal modes [28] — which, if true, meant that any admissible initial shape has a trigonometric expansion, a proposition both of the others rejected as obviously false. Bernoulli was right and could not prove it; the proof, and with it the modern notion of a function, had to wait for Fourier [72] and then for Dirichlet.

What the controversy actually settled is the class of admissible initial data, and every part of the answer now lives in Part II. Dirichlet's theorem Theorem 17.9 gives convergence for piecewise smooth data — which covers the plucked string, corner and all. du Bois-Reymond's continuous function with a divergent Fourier series Theorem 17.13 shows that continuity alone is not enough, so Euler's freehand curves are genuinely too general for pointwise convergence; Fejér's theorem Theorem 17.22 rescues them in the mean, and the full convergence story is Section 17.1.2. Data rougher still — a struck string, whose initial velocity is a delta function at the hammer — is handled not by strengthening the convergence but by weakening the notion of solution, which is Definition 10.83 and Section 10.6.1. A corner in the initial shape propagates along a characteristic and stays a corner forever, which is Equation (28.48) again and is exactly what makes the weak formulation necessary.

28.4.3 Standing waves

Phenomenon 28.45 (The harmonic series of a stretched string)**.** A string clamped at both ends does not vibrate at an arbitrary frequency. It sustains a discrete set, $f_n = nf_1$ with $n = 1, 2, 3, \dots$, each with its own stationary pattern of nodes and antinodes, the n-th carrying $n - 1$ nodes between the ends; a bow or a hammer excites a superposition of them, and the ear hears the fundamental with the higher members as timbre. The fundamental is found to vary inversely with the length of the string, as the square root of the tension, and inversely as the square root of the mass per unit length — the three rules by which every stringed instrument is strung and tuned, and which were known from measurement long before any wave theory existed to explain them [2, 4].

$$f_n = \frac{n}{2L}\sqrt{\frac{T}{\mu}}\,, \qquad n \in \mathbb{N}\,. \tag{28.49}$$

Rests on Equation (28.46) *and Theorem 9.57.*

Derivation. Derives Phenomenon 28.45.

Grant the wave equation of Section 28.4.1, $\partial_t^2 y = v^2 \partial_x^2 y$ with $v = \sqrt{T/\mu}$, on the interval $0 \le x \le L$, with the clamped ends imposing $y(0,t) = y(L,t) = 0$. Separating variables (Section 10.2), $y = X(x)\Theta(t)$, gives $X''/X = \Theta''/(v^2\Theta)$; the two sides depend on different variables, so both equal one constant (Lemma 10.28), which must be negative if Θ is to stay bounded — write it $-k^2$. Then $X = a\sin kx + b\cos kx$; the condition $X(0) = 0$ forces $b = 0$, and $X(L) = 0$ then forces $\sin kL = 0$, that is

$$k_n = \frac{n\pi}{L}\,, \qquad n = 1, 2, 3, \dots \tag{28.50}$$

The temporal factor oscillates at $\omega_n = vk_n$, so $f_n = \omega_n/2\pi = nv/(2L)$, which is Equation (28.49); the three empirical rules read off it at once, $f_1 \propto 1/L$, $f_1 \propto \sqrt{T}$ and $f_1 \propto 1/\sqrt{\mu}$. That the modes are the *only* solutions, and that a general motion is a superposition of them, is the Sturm–Liouville theory of Chapter 9: the spatial problem is Section 9.7 with $p = r = 1$ and $q = 0$, its eigenfunctions are mutually orthogonal by Theorem 9.57, and they are complete by Theorem 17.34. That the overtones are *exact* integer multiples of the fundamental is special to the non-dispersive string, in which $\omega = vk$ is linear in k: a stiff bar, a drumhead or a bell has the same discreteness of modes but an inharmonic spectrum, and does not sound a definite pitch in the same way. □

A standing wave is two travelling waves, which is how the boundary condition acts. Adding the incident wave to the wave reflected from a clamped end, both of the form Equation (28.48),

$$A\cos(kx - \omega t) - A\cos(kx + \omega t) = 2A\sin kx\,\sin\omega t\,. \tag{28.51}$$

From Equation (28.48) *(sum-to-product on two counter-propagating waves of equal amplitude).*

The right-hand side is a fixed spatial pattern multiplied by a pure oscillation in time: the nodes at $kx = n\pi$ never move, and no energy crosses them.

Proposition 28.46 (Reflection at an end)**.** *A pulse on a string reflects inverted from a clamped end and upright from an end free to slide transversely.*

Rests on Equations (28.46) *and* (28.48).

Proof. Derives Proposition 28.46.

Take the end at $x = 0$ and write the field as $f(x - vt) + g(x + vt)$, with f the incoming pulse. A clamped end imposes $y(0,t) = 0$ for all t, hence $g(s) = -f(-s)$: the reflected pulse is the incident one reversed in space and in sign. A free end can support no transverse force, so $T\partial_x y(0,t) = 0$, hence $g'(s) = f'(-s)$ and $g(s) = f(-s)$ up to a constant fixed by $y \to 0$ far away: reversed in space and not in sign. □

Melde demonstrated the standing patterns directly, driving one end of a string with a tuning fork and adjusting the tension until a whole number of loops appeared between the ends [155]; the pattern is stationary, visible, and countable, and reading n off it turns Equation (28.49) into a measurement of v. The experiment and its modern repetitions are Chapter 35.

28.4.4 Fourier decomposition

Proposition 28.47 (Modal solution of the initial-value problem)**.** *The solution of Equation (28.46) on $[0,L]$ with $y(0,t)=y(L,t)=0$, $y(x,0)=y_0(x)$ and $\partial_t y(x,0)=v_0(x)$ is*

$$y(x,t)=\sum_{n=1}^{\infty}\left(a_n\cos\omega_n t+b_n\sin\omega_n t\right)\sin k_n x\,,\qquad \omega_n=vk_n\,, \tag{28.52}$$

with $k_n=n\pi/L$ and

$$a_n=\frac{2}{L}\int_0^L y_0(x)\sin k_n x\,\mathrm{d}x\,,\qquad b_n=\frac{2}{L\omega_n}\int_0^L v_0(x)\sin k_n x\,\mathrm{d}x\,. \tag{28.53}$$

Rests on Equation (28.50) and Theorem 17.3.

Proof. *Derives Proposition 28.47.*

Each term of Equation (28.52) solves Equation (28.46) and the boundary conditions, by Phenomenon 28.45, and the sum does too by Proposition 28.44 and termwise differentiation (Theorem 7.51) where the series permits it. Setting $t=0$ gives $y_0=\sum a_n\sin k_n x$ and $v_0=\sum b_n\omega_n\sin k_n x$, which are sine series on $[0,L]$; Equation (28.53) is the Euler–Fourier formula Theorem 17.3 for their coefficients, using the orthogonality $\int_0^L \sin k_n x\sin k_m x\,\mathrm{d}x=(L/2)\delta_{nm}$ of Theorem 9.57. Completeness of the system (Theorem 17.34) is what makes the expansion possible for every square-integrable datum, and uniqueness for the wave equation with given data (Definition 10.24, Theorem 10.55) makes Equation (28.52) the solution and not merely a solution. □

Each mode evolves independently and forever: the string is n uncoupled oscillators, exactly as Proposition 28.36 promised, with n now running to infinity. The energy divides among them in the same way.

Proposition 28.48 (Modal energy)**.** *For Equation (28.52) the total energy is*

$$E=\frac{\mu L}{4}\sum_{n=1}^{\infty}\omega_n^2\left(a_n^2+b_n^2\right)\,, \tag{28.54}$$

constant in time, with the n-th term the energy of the n-th mode.

Rests on Equation (28.52) and Theorem 17.34.

Proof. *Derives Proposition 28.48.*

Write $y=\sum_n q_n(t)\sin k_n x$, and T_{ten} for the tension, since here T is the kinetic energy (Notation 28.1). Then, by the orthogonality used in Proposition 28.47,

$$T=\frac{\mu}{2}\int_0^L(\partial_t y)^2\,\mathrm{d}x=\frac{\mu L}{4}\sum_n\dot q_n^2\,,\qquad V=\frac{T_{\text{ten}}}{2}\int_0^L(\partial_x y)^2\,\mathrm{d}x=\frac{\mu L}{4}\sum_n\omega_n^2 q_n^2\,,$$

the second using $T_{\text{ten}}k_n^2=\mu\omega_n^2$. This is Parseval's identity Theorem 17.34 for the sine system, and it is the continuum form of Equation (28.41). With $q_n=a_n\cos\omega_n t+b_n\sin\omega_n t$ each bracket $\dot q_n^2+\omega_n^2 q_n^2$ equals $\omega_n^2(a_n^2+b_n^2)$ by Proposition 28.5, giving Equation (28.54). □

Example 28.49 (The plucked string, and why a guitar is plucked where it is)**.** Draw the string aside by a height h at the point $x = d$ and release it from rest, so that $v_0 = 0$, every $b_n = 0$, and y_0 is the triangle with apex at d. Evaluating the two integrals of Equation (28.53) over the two straight segments gives

$$a_n = \frac{2hL^2}{\pi^2 n^2\, d\,(L-d)} \sin\!\left(\frac{n\pi d}{L}\right). \tag{28.55}$$

Three things follow. The amplitudes fall as $1/n^2$, which is the general rule that a kink in the data — here the corner at d — fixes the decay rate of the coefficients (Proposition 17.10); the energy, by Equation (28.54), therefore falls as $1/n^2$ as well, since $\omega_n \propto n$. And every harmonic with nd/L an integer is missing outright, because the pluck point is a node of that mode and cannot excite it: plucking at $L/7$ removes the seventh harmonic and its multiples, and plucking at the midpoint removes every even one, leaving a tone of odd partials only. Which partials a player wants is a question of taste rather than of physics, but Equation (28.55) says exactly which ones the choice of d disposes of, and that much is testable on any instrument.

Rests on Equations (28.53) *and* (28.54)*.*

The convergence of Equation (28.52) raises questions this chapter does not settle and does not need to: they are Section 17.1.2, where the transform pair and its convergence theory are developed in full, and Section 10.2.2, where the eigenfunction expansion is carried out for the general boundary-value problem.

28.5 Sound

28.5.1 The acoustic wave equation

Phenomenon 28.50 (The speed of sound in air)**.** Sound travels through still air at about 343 m/s at room temperature, at a speed independent of the loudness of the sound and, across the audible range, of its pitch: a chord fired from a distant gun arrives as a chord, and thunder is not smeared into a glissando. The speed rises with temperature, and at fixed temperature it does not depend on the pressure of the air. Newton computed it in the *Principia* from the elasticity of the air and obtained a value substantially below what was measured [14]; the shortfall stood for more than a century, and was resolved only by recognizing that the compressions in a sound wave are adiabatic rather than isothermal — which makes the speed of sound a measurement of a thermodynamic quantity.

Rests on Equation (28.46)*.*

Derivation. *Derives Phenomenon 28.50.*

Linearize the equations of a compressible fluid about rest, writing $\rho = \rho_0 + \rho'$ and $p = p_0 + p'$ with a velocity u that is first order, and dropping all products of primed quantities. Conservation of mass and Newton's second law for a fluid element give, in one dimension,

$$\frac{\partial \rho'}{\partial t} + \rho_0 \frac{\partial u}{\partial x} = 0\,, \qquad \rho_0 \frac{\partial u}{\partial t} = -\frac{\partial p'}{\partial x}\,.$$

Differentiating the first with respect to t and the second with respect to x and subtracting eliminates u:

$$\partial_t^2 \rho' = \left(\frac{\mathrm{d}p}{\mathrm{d}\rho}\right) \partial_x^2 \rho'\,, \tag{28.56}$$

a wave equation with speed $v^2 = \mathrm{d}p/\mathrm{d}\rho$. The derivative must be taken along the thermodynamic path the gas actually follows, and this is where the history turns. Newton took it isothermally: with $p \propto \rho$ at fixed temperature, $v_T = \sqrt{p_0/\rho_0}$, which for $p_0 = 101\,325\,\mathrm{Pa}$ and $\rho_0 = 1.204\,\mathrm{kg/m^3}$ gives about 290 m/s — some 15 % below the measured value. But a

sound wave compresses the air far faster than heat can diffuse across a wavelength, so the correct path is the adiabatic one, $p \propto \rho^{\gamma}$, whence

$$v = \sqrt{\frac{\gamma p_0}{\rho_0}} = \sqrt{\frac{\gamma R T}{M}}\,, \tag{28.57}$$

and with the diatomic value $\gamma = 1.4$ this is 343 m/s, the observed speed. Because p_0/ρ_0 equals RT/M for an ideal gas, the speed depends on temperature and on composition but not on pressure; and because Equation (28.56) has a k-independent coefficient, it carries no dispersion — both exactly as observed. The thermodynamic input is developed in Chapter 54, and the fluid equations linearized here in Chapter 31. □

Table 28.1: The two thermodynamic paths, computed from the second form of Equation (28.57) with the molar gas constant $R = 8.314\,462\,618$ J/(mol K) — exact by the SI definitions [2431] — the mean molar mass of dry air $M = 28.96 \times 10^{-3}$ kg/mol of Remark 28.2, and $\gamma = 1.4$. Newton's isothermal path is the column with γ omitted. The adiabatic column is what is measured; the isothermal one is low by 15 % at every temperature, because the two differ by the constant factor $\sqrt{\gamma}$ and by nothing else.

Temperature	Isothermal (m/s)	Adiabatic (m/s)
0 °C	280.0	331.4
20 °C	290.1	343.3

Remark 28.51 (Why the compressions really are adiabatic). The choice of path is not a matter of taste, and it can be checked with one number. Heat diffuses a distance $\sqrt{\alpha/\omega}$ in one radian of a cycle, with α the thermal diffusivity; for air, $\alpha \approx 2.1 \times 10^{-5}\,\mathrm{m^2/s}$ (Remark 28.2), so at 1 kHz that distance is 5.8×10^{-5} m against a wavelength of 0.343 m — a ratio of 1.7×10^{-4}. Heat has no time to move between a compression and the rarefaction beside it, so each parcel is compressed adiabatically and Laplace's factor γ is right [59]. The criterion inverts at very high frequency or very low pressure, where the two lengths become comparable and sound becomes both dispersive and strongly absorbed. Read the other way, Equation (28.57) makes a speed measurement into a measurement of $\gamma = c_p/c_V$, and hence of the number of excited degrees of freedom of the molecule (Chapter 55) — a fact about heat extracted from a stopwatch and a distance.

Rests on Equation (28.57).

The same linearization gives the rest of the acoustic vocabulary. With $p' = \rho_0 v\, u$ for a wave travelling in the $+x$ direction, the ratio of acoustic pressure to particle velocity is the *specific acoustic impedance*

$$Z_0 = \rho_0 v = 1.204\,\mathrm{kg/m^3} \times 343\,\mathrm{m/s} = 413\,\mathrm{Pa\,s/m}\,, \tag{28.58}$$

and the intensity — the power crossing unit area — is $I = \langle p' u \rangle = p_{\mathrm{rms}}^2/Z_0$.

Example 28.52 (The threshold of hearing)**.** The reference intensity of the acoustic decibel scale is $I_0 = 1 \times 10^{-12}\,\mathrm{W/m^2}$, chosen as the faintest 1 kHz tone a good ear detects. The corresponding pressure amplitude is

$$p_{\mathrm{rms}} = \sqrt{I_0 Z_0} = \sqrt{1 \times 10^{-12}\,\mathrm{W/m^2} \times 413\,\mathrm{Pa\,s/m}} = 2.0 \times 10^{-5}\,\mathrm{Pa}\,,$$

which is exactly the 20 µPa that the pressure-referenced decibel scale uses — the two references agree because Equation (28.58) relates them. Two orders of magnitude are worth pausing on: 20 µPa is 2×10^{-10} of atmospheric pressure, and the associated displacement of the air, $p_{\mathrm{rms}}/(\omega Z_0)$, is about 8×10^{-12} m — a tenth of an atomic diameter. The ear is a detector working within a factor of a few of the thermal noise of the air itself.

Rests on Equation (28.58).

28.5.2 Acoustic phenomena

Chladni founded experimental acoustics by strewing sand on a metal plate and bowing its edge [39]. The sand collects into sharp curves that change abruptly as the bowing point and the damping finger are moved, and the same plate always gives the same family of figures.

Phenomenon 28.53 (Chladni's figures)**.** A flat plate driven at a slowly increasing frequency responds strongly only near a discrete set of frequencies, and at each of them sand sprinkled on it settles into a stationary pattern of curves — a different pattern for each frequency, reproducible for a given plate, shape and support [39]. Between those frequencies no pattern forms.

Rests on Proposition 10.29 and Theorem 9.57.

Derivation. *Derives Phenomenon 28.53.*

The small transverse displacement $w(\boldsymbol{x}, t)$ of a thin elastic plate obeys a linear equation $\rho_s\, \partial_t^2 w + \mathcal{A} w = 0$, where $\mathcal{A}$ is a linear spatial operator, self-adjoint under the plate's boundary conditions, determined by the elastic moduli (Chapter 30); its explicit form is not needed here. Time separation (Proposition 10.29) reduces the problem to the eigenvalue problem $\mathcal{A} w_r = \rho_s \omega_r^2 w_r$ on a bounded domain, which for a self-adjoint operator has a discrete spectrum with mutually orthogonal eigenfunctions (Theorem 9.57,Section 10.2.2, Theorem 12.44) — so the plate has isolated natural frequencies and, by Phenomenon 28.23 applied to each of them, responds appreciably only near one of them. At $\omega = \omega_r$ the response is essentially the single eigenfunction w_r, whose zero set is a fixed family of curves in the plate, unmoving in time because $w = w_r(\boldsymbol{x}) \cos \omega_r t$ factorizes. A grain of sand on the plate is thrown clear wherever the vertical acceleration $\omega_r^2 \, |w_r|$ exceeds g and is not thrown clear on the zero set; each bounce lands it at random nearby, so it performs a random walk that is absorbed on the nodal curves, where it accumulates. The pattern is therefore the nodal set of w_r, and it changes discontinuously with r because the eigenfunctions do. □

The plate's operator is fourth order rather than second, so its eigenfrequencies are not in the ratios $1 : 2 : 3$ of Equation (28.49) and a plate does not sound a pitch the way a string does; Chladni's figures are the two-dimensional face of Equation (28.50), and the elasticity that fixes $\mathcal{A}$ is Chapter 30.

Helmholtz built the complementary instrument: a rigid cavity with a short open neck, which resonates at one frequency and can therefore be used to detect a single partial in a complex tone [159]. It is the purest acoustic realization of Definition 28.3.

Proposition 28.54 (The Helmholtz resonator)**.** *A cavity of volume V with a neck of cross-section S and effective length L' resonates at*

$$\omega_0 = v \sqrt{\frac{S}{V L'}}\,, \tag{28.59}$$

with v the speed of sound.

Rests on Equations (28.2) and (28.57).

Proof. *Derives Proposition 28.54.*

The air in the neck moves as a slug of mass $\rho_0 S L'$; the air in the cavity is too slow-moving to contribute inertia but supplies the stiffness. Pushing the slug inward a distance x reduces the cavity volume by Sx, and by the adiabatic relation used in Equation (28.57) the pressure rises by $\delta p = -\gamma p_0 \, \delta V / V = \gamma p_0 S x / V$. The restoring force on the slug is $-S \, \delta p = -\gamma p_0 S^2 x / V$, so the effective spring constant is $\gamma p_0 S^2 / V$ and

$$\omega_0^2 = \frac{\gamma p_0 S^2 / V}{\rho_0 S L'} = \frac{\gamma p_0}{\rho_0} \cdot \frac{S}{V L'} = v^2 \frac{S}{V L'}\,,$$

using Equation (28.57). The approximation is that the wavelength greatly exceeds every dimension of the resonator, so that the cavity pressure is uniform and the neck flow incompressible; L' exceeds the geometric neck length by an end correction of order the neck radius, for the same reason a pipe does. □

Example 28.55 (Blowing across a bottle)**.** A wine bottle of $V = 0.75\,\mathrm{L}$ with a neck of radius 9 mm and effective length 80 mm has $S = 2.54 \times 10^{-4}\,\mathrm{m}^2$, $V = 7.5 \times 10^{-4}\,\mathrm{m}^3$ and $S/(VL') = 4.24/\mathrm{m}^2$, whose square root is $2.06/\mathrm{m}$, so Equation (28.59) gives $\omega_0 = 343\,\mathrm{m/s} \times 2.06/\mathrm{m} = 707\,\mathrm{rad/s}$, that is $f_0 = 113\,\mathrm{Hz}$. Emptying the bottle raises V and lowers the pitch — the audible test of Equation (28.59), and the reason the note falls as a bottle is filled is that V is the volume of *air*, not of bottle.

Rests on Equation (28.59).

The last of the classical acoustic effects belongs to the source rather than the resonator. Doppler predicted that motion of source or observer shifts the received frequency [107]; Buys Ballot tested it in 1845 by putting trumpeters on an open railway carriage and stationing trained musicians beside the line to name the note heard as the train passed [111].

Phenomenon 28.56 (The acoustic Doppler shift is not symmetric)**.** The pitch of a whistle rises as its source approaches and falls as it recedes, the change occurring abruptly as the source passes. The shift depends on which of the two — source or listener — is moving through the air, and not only on their relative velocity: a source approaching a still listener at speed u and a listener approaching a still source at the same u give measurably different frequencies [107, 111].

Rests on Equations (28.46) *and* (28.57).

Derivation. *Derives Phenomenon 28.56.*

Sound is a disturbance of the air, and Equation (28.56) holds in the frame in which the air is at rest; that frame is therefore singled out, and the two motions are physically different situations. Let the source emit at frequency f and move toward the listener at u_s, the listener move toward the source at u_o, and the speed of sound be v. In one period $1/f$ the source advances u_s/f toward the wavefront it emitted a period earlier, so the wavelength ahead of it is compressed to $\lambda' = (v - u_s)/f$. The listener, moving into the wave, meets crests at the closing speed $v + u_o$, so

$$f' = \frac{v + u_o}{\lambda'} = f\,\frac{v + u_o}{v - u_s}\,. \tag{28.60}$$

The two velocities enter differently: u_o in the numerator, where it changes the rate of arrival, and u_s in the denominator, where it changes the spacing. Expanding for $u \ll v$, both give $f' \approx f(1 + u/v)$, so the asymmetry is second order and appears only at $O(u^2/v^2)$ — which is why it took a moving train to find it, and why no first-order acoustic experiment could have distinguished the cases. □

Remark 28.57 (The contrast with light is the whole of relativity). Equation (28.60) is asymmetric because the air exists. For light there is no medium and no preferred frame, so the shift can depend only on the relative velocity; the relativistic formula of Section 38.5 is symmetric under exchange of source and observer, and its second-order term is the time dilation rather than an artefact of a medium. The failure of every attempt to find the acoustic-style asymmetry for light is Chapter 37. Note also what Equation (28.60) says at $u_s = v$: the denominator vanishes, all the crests pile onto one surface, and the linearization behind Equation (28.56) fails — which is the shock wave of Section 10.6.3 and Chapter 31.

Rests on Equation (28.60).

Pipes, membranes and rooms complete the classical subject and are Rayleigh's [198]. A pipe open at both ends carries the full harmonic series $f_n = nv/2L$, because both ends are pressure nodes; a pipe closed at one end carries only the odd members, $f_n = (2n-1)v/4L$, because a closed end is a pressure antinode. That is Equation (28.50) with one boundary condition changed, and it is why a stopped organ pipe sounds an octave below an open one of the same length and with a different timbre. The measurements are Chapter 35.

28.6 Dispersion

28.6.1 Phase and group velocity

Definition 28.58 (Dispersion relation; phase and group velocity)**.** For a linear medium supporting waves $\mathrm{e}^{\mathrm{i}(kx-\omega t)}$, the *dispersion relation* is the function $\omega(k)$ the medium imposes. The *phase velocity* and *group velocity* are

$$v_p = \frac{\omega}{k}\,, \qquad v_g = \frac{\mathrm{d}\omega}{\mathrm{d}k}\,, \tag{28.61}$$

and the medium is *non-dispersive* when $\omega = vk$ with v constant, in which case the two coincide.

Rests on Equations (28.45) *and* (28.46)*.*

Phenomenon 28.59 (Groups and crests travel at different speeds)**.** In a dispersive medium a wave group does not travel at the speed of the crests inside it. On deep water the effect is watchable without apparatus: individual crests run forward through a group of swell, appearing at its rear, crossing it and vanishing at its front, so that the group advances at half the speed of its own crests [200]. A wave packet launched in such a medium also spreads as it goes, its length growing without bound. In a non-dispersive medium — a stretched string, air for sound — neither happens: crests and groups keep step and a pulse holds its shape, which is why a spoken word survives the length of a lecture hall.

Rests on Equations (28.34) *and* (28.61)*.*

Derivation. *Derives Phenomenon 28.59.*

Superpose two travelling waves of nearly equal wavenumbers $k \pm \Delta k/2$ and the corresponding frequencies $\omega \pm \Delta\omega/2$. The identity Equation (28.34), applied with the phase $kx - \omega t$ in place of ωt, gives

$$\cos(k_1 x - \omega_1 t) + \cos(k_2 x - \omega_2 t) = 2\cos\!\left(\frac{\Delta k\, x - \Delta\omega\, t}{2}\right)\cos(kx - \omega t)\,,$$

whose second factor — the carrier — has crests moving at the phase velocity $v_p = \omega/k$, while the first — the envelope — has its maxima moving at $\Delta\omega/\Delta k$, which in the limit of a narrow band is

$$v_g = \frac{\mathrm{d}\omega}{\mathrm{d}k}\,. \tag{28.62}$$

The two coincide exactly when $\omega = vk$ is linear in k, which is the non-dispersive case. For gravity waves on deep water the dispersion relation is $\omega = \sqrt{gk}$, derived in Chapter 31, so

$$v_p = \frac{\omega}{k} = \sqrt{\frac{g}{k}}\,, \qquad v_g = \frac{\mathrm{d}\omega}{\mathrm{d}k} = \frac{1}{2}\sqrt{\frac{g}{k}} = \frac{v_p}{2}\,,$$

the observed factor of two. Spreading follows from carrying the same expansion one order further, which is Proposition 28.60: components of a packet with different k travel at different v_g, so the packet's spatial extent grows at a rate set by $\mathrm{d}^2\omega/\mathrm{d}k^2$, which vanishes identically in the non-dispersive case. The discrete chain of Proposition 28.40 is the mechanical prototype of a dispersive medium, and the same kinematics governs the matter waves of Chapter 71. □

Proposition 28.60 (A wave packet translates at v_g and spreads at a rate set by the curvature of $\omega(k)$)**.** *Let $y(x,t) = \int A(k) \mathrm{e}^{\mathrm{i}(kx-\omega(k)t)} \mathrm{d}k/2\pi$ with A concentrated near k_0 within a width Δk, and expand $\omega(k) = \omega_0 + v_g\,(k-k_0) + \frac{1}{2}\omega''\,(k-k_0)^2 + \cdots$. Then to first order the envelope of y is rigidly translated at v_g; the second order broadens an initially Gaussian envelope of width σ_0 to*

$$\sigma(t) = \sigma_0 \sqrt{1 + \left(\frac{\omega'' t}{\sigma_0^2}\right)^2}\,, \tag{28.63}$$

which is constant if and only if $\omega'' = 0$.

Rests on Equation (28.61) and Definition 17.29.

Proof. *Derives Proposition 28.60.*

Substituting the expansion and setting $q = k - k_0$,

$$y = \mathrm{e}^{\mathrm{i}(k_0 x - \omega_0 t)} \int \frac{\mathrm{d}q}{2\pi} A(k_0+q)\, \mathrm{e}^{\mathrm{i}q(x-v_g t)} \mathrm{e}^{-\frac{\mathrm{i}}{2}\omega'' q^2 t}\,,$$

a carrier times an envelope. With ω'' dropped the envelope is a function of $x - v_g t$ alone — rigid translation at v_g, which is the first claim, and which is Proposition 17.30 (a linear phase in q is a translation in x). Retaining ω'' multiplies the transform by a Gaussian of imaginary width; for $A \propto \mathrm{e}^{-\sigma_0^2 q^2/2}$ the two Gaussians combine to one of complex variance $\sigma_0^2 + \mathrm{i}\omega'' t$, whose modulus in x has width $\left|\sigma_0^2 + \mathrm{i}\omega'' t\right| / \sigma_0$, which is Equation (28.63). Since $\sigma(t)$ is strictly increasing unless $\omega'' = 0$, spreading is equivalent to curvature of $\omega(k)$. □

The packet cannot be made both short and monochromatic: σ_0 small forces Δk large, and the trade is exactly the bandwidth theorem Theorem 17.43, with its SI statement in Corollary 17.44. That inequality is a property of Fourier transforms and has nothing to do with quantum mechanics; what Chapter 71 adds to it is the relation $p = \hbar k$, which converts a statement about wave packets into a statement about momentum.

Example 28.61 (Group velocity on the loaded chain)**.** Differentiating Equation (28.45),

$$v_p = \frac{2\omega_c}{k} \sin\!\left(\frac{ka}{2}\right), \qquad v_g = \omega_c a \cos\!\left(\frac{ka}{2}\right).$$

For $ka \ll 1$ both reduce to $\omega_c a = \sqrt{T/\mu}$, the continuum speed of Proposition 28.41. At the other end, $ka \to \pi$, the group velocity vanishes while the phase velocity does not: the mode at the cutoff is a standing oscillation in which alternate beads move oppositely, and it transports no energy at all. A dispersive medium therefore has a band of frequencies it carries and, beyond $\omega_{\max} = 2\omega_c$, a band it cannot carry. This is the mechanical original of every band structure in Chapter 125.

Rests on Equations (28.45) and (28.61).

Remark 28.62 (Neither velocity is a speed limit)*.* In a strongly absorbing medium $\mathrm{d}\omega/\mathrm{d}k$ can exceed the speed of light, or become negative, without anything travelling faster than light: Proposition 28.60 assumed a narrow band and a smooth $\omega(k)$, and near an absorption line neither holds, so v_g stops being the speed of anything. What is bounded is the front of a signal, which is governed by $\omega(k)$ as $k \to \infty$ and, by the same analyticity that gives Theorem 17.78 and Remark 28.26, never exceeds c. The point is taken up where it matters, in Chapter 61 and Chapter 62.

Rests on Equation (28.63) and Theorem 17.78.

28.6.2 Energy transport and impedance

Proposition 28.63 (Energy density, flux, and the conservation law)**.** *For the string of Proposition 28.42 the energy density and energy flux are*

$$u = \tfrac{1}{2}\mu\,(\partial_t y)^2 + \tfrac{1}{2}T\,(\partial_x y)^2\,, \qquad P = -T\,\partial_x y\,\partial_t y\,, \tag{28.64}$$

and they satisfy $\partial_t u + \partial_x P = 0$. *For a travelling wave* $y = A\cos(kx - \omega t)$ *the two energies are equal at every point and*

$$\bar{P} = \tfrac{1}{2} Z \omega^2 A^2 \,, \qquad Z := \mu v = \sqrt{T\mu} \,, \tag{28.65}$$

with the characteristic impedance Z *of SI dimension* kg/s.
Rests on Equation (28.46) *and Proposition 10.57.*

Proof. *Derives Proposition 28.63.*
Differentiating u and using Equation (28.46),

$$\partial_t u = \mu\, \partial_t y\, \partial_t^2 y + T\, \partial_x y\, \partial_x \partial_t y = T\left(\partial_t y\, \partial_x^2 y + \partial_x y\, \partial_x \partial_t y\right) = \partial_x \left(T\, \partial_x y\, \partial_t y\right) = -\partial_x P \,,$$

a conservation law in the sense of Definition 10.93; the same statement in integral form over a backward cone is Proposition 10.57. For a wave travelling in the $+x$ direction, $\partial_t y = -v\, \partial_x y$ by Equation (28.48), so $\tfrac{1}{2}\mu(\partial_t y)^2 = \tfrac{1}{2}\mu v^2 (\partial_x y)^2 = \tfrac{1}{2} T (\partial_x y)^2$: the kinetic and potential densities are equal pointwise, not merely on average as in Proposition 28.5. With $y = A\cos(kx - \omega t)$, $P = TA^2 k\omega \sin^2(kx - \omega t)$, whose mean is $\tfrac{1}{2} TkA^2\omega = \tfrac{1}{2}\mu v \omega^2 A^2$ using $Tk = \mu v \omega$. The dimension of $Z = \sqrt{T\mu}$ is $\sqrt{\mathrm{N} \times \mathrm{kg/m}} = \mathrm{kg/s}$. □

Energy flows at v on the string, since $\bar{P}/\bar{u} = v$ from Equation (28.64); in a dispersive medium the same calculation gives v_g, which is the physical reason Equation (28.62) is the velocity that matters. The impedance Z is what decides what happens when the medium changes.

Proposition 28.64 (Reflection and transmission at a junction)**.** *Two strings of impedances* Z_1 *and* Z_2 *joined at* $x = 0$ *under common tension, with a wave incident from the left, give amplitude coefficients*

$$r = \frac{Z_1 - Z_2}{Z_1 + Z_2} \,, \qquad \tau = \frac{2Z_1}{Z_1 + Z_2} \,, \tag{28.66}$$

and energy fractions r^2 *reflected and* $4Z_1 Z_2 / (Z_1 + Z_2)^2$ *transmitted, which sum to one.*
Rests on Equations (28.48) *and* (28.65).

Proof. *Derives Proposition 28.64.*
Write the field as $\mathrm{e}^{\mathrm{i}k_1 x} + r\mathrm{e}^{-\mathrm{i}k_1 x}$ for $x < 0$ and $\tau \mathrm{e}^{\mathrm{i}k_2 x}$ for $x > 0$, common time factor suppressed. Continuity of the string requires $1 + r = \tau$; continuity of the transverse force $T\,\partial_x y$, which must hold because a massless junction cannot sustain a net force, requires $k_1(1 - r) = k_2 \tau$. Since the tension is common and $k = \omega/v$ with $Z = T/v$, one has $k_i = \omega Z_i / T$, so the second condition reads $Z_1(1 - r) = Z_2 \tau$. Solving the pair gives Equation (28.66). The incident power is $\tfrac{1}{2} Z_1 \omega^2$ by Equation (28.65), the reflected $\tfrac{1}{2} Z_1 \omega^2 r^2$ and the transmitted $\tfrac{1}{2} Z_2 \omega^2 \tau^2$; the ratios are as stated, and $(Z_1 - Z_2)^2 + 4Z_1 Z_2 = (Z_1 + Z_2)^2$ is the conservation of energy. □

Equation (28.66) contains Proposition 28.46 as its two limits: a clamped end is $Z_2 \to \infty$, giving $r = -1$ and total inverted reflection, and a free end is $Z_2 \to 0$, giving $r = +1$. Nothing is reflected when $Z_1 = Z_2$ — *impedance matching*, and the point is that the two strings need not be identical, only equally impeding. That is the mechanical prototype of every matching problem in the treatise: the antireflection coating and the Fresnel coefficients of Chapter 61, the horn that couples a loudspeaker to the air across the factor of 10^3 between their impedances, and the middle ear, whose three bones exist to match the impedance of air to that of the cochlear fluid. A reflection coefficient near ± 1 is also what makes a resonator: the string of Phenomenon 28.45 sustains modes precisely because both ends reflect essentially everything.

28.7 Anharmonicity: a first look

28.7.1 The anharmonic oscillator

Every result of this chapter so far has rested on Equation (28.11) truncated at the quadratic term. This section asks what the next term does, and finds that it does three things a linear equation cannot: it makes the frequency depend on the amplitude, it bends the resonance curve over, and it makes the steady-state response multivalued, so that the amplitude jumps discontinuously as the drive is swept.

Phenomenon 28.65 (The period of a pendulum grows with its amplitude)**.** A pendulum released from a large angle takes measurably longer to complete a swing than the same pendulum released from a small one, and the excess grows as the square of the amplitude. Galileo's own observations of approximate isochronism already show the departure at wide arcs [4], and it is the defect Huygens removed by forcing the bob onto a cycloid (Proposition 28.12) [11]. At an amplitude of 7° the excess is about one part in 10^3; at 30° it is about 1.7 %.

$$T(\theta_0) = 2\pi\sqrt{\frac{\ell}{g}}\left(1 + \frac{\theta_0^2}{16} + O\left(\theta_0^4\right)\right). \tag{28.67}$$

Rests on Equations (28.8) *and* (28.81)*.*

Derivation. *Derives Phenomenon 28.65.*

Expand the exact pendulum equation Equation (28.8) one order beyond the linearization, $\sin\theta = \theta - \theta^3/6 + O(\theta^5)$, giving

$$\ddot{\theta} + \omega_0^2\theta - \frac{\omega_0^2}{6}\theta^3 = 0\,, \qquad \omega_0^2 = \frac{g}{\ell}\,, \tag{28.68}$$

which is Equation (28.70) with $\epsilon = -\omega_0^2/6$. Substituting that value into Equation (28.81) gives

$$\omega = \omega_0 + \frac{3\epsilon\theta_0^2}{8\omega_0} = \omega_0\left(1 - \frac{\theta_0^2}{16}\right),$$

and $T = 2\pi/\omega$ is Equation (28.67) to the order retained. The sign is the physical content: $\epsilon < 0$ because the restoring moment $\sin\theta$ falls short of θ, so the pendulum is a *softening* oscillator and its period lengthens. Numerically, $\theta_0 = 7^\circ = 0.1222\,\text{rad}$ gives $\theta_0^2/16 = 9.3\times10^{-4}$ and $\theta_0 = 30^\circ = 0.5236\,\text{rad}$ gives 1.7×10^{-2}, as asserted. □

Remark 28.66 (The exact period is an elliptic integral)*.* Energy conservation integrates Equation (28.8) exactly: with θ_0 the amplitude, the quarter period is $\int_0^{\theta_0} \mathrm{d}\theta/\sqrt{2\omega_0^2(\cos\theta - \cos\theta_0)}$, and the substitution $\sin(\theta/2) = \sin(\theta_0/2)\sin\phi$ brings it to $\omega_0^{-1}\int_0^{\pi/2}\mathrm{d}\phi/\sqrt{1 - m\sin^2\phi}$ with $m = \sin^2(\theta_0/2)$ — the complete elliptic integral of the first kind, whose expansion in m reproduces Equation (28.67) and every later term. This treatise does not develop the elliptic integrals: they are absent from Chapter 9, which carries the Bessel, Legendre, Hermite and Laguerre families, so the closed form is recorded here as a fact about the integral and is used nowhere. Everything asserted in Phenomenon 28.65 is obtained instead from Theorem 28.72, which needs nothing beyond Theorem 7.38.

Rests on Equation (28.67)*.*

Definition 28.67 (Duffing oscillator)**.** The damped, driven oscillator with a cubic restoring term,

$$\ddot{x} + 2\beta\dot{x} + \omega_0^2 x + \epsilon x^3 = \frac{F_0}{m}\cos\omega t\,, \tag{28.69}$$

whose free undamped case

$$\ddot{x} + \omega_0^2 x + \epsilon x^3 = 0 \tag{28.70}$$

is *hardening* for $\epsilon > 0$ and *softening* for $\epsilon < 0$ [452]. ϵ has SI dimension $1/(\mathrm{m}^2\,\mathrm{s}^2)$.

Rests on Equations (28.11) and (28.23).

Proposition 28.68 (The bent resonance curve)**.** *For $|\epsilon|$, β and F_0 all small, and ω near ω_0, the steady-state amplitude A of Equation (28.69) satisfies*

$$A^2\left[\left(\omega-\omega_0-\frac{3\epsilon A^2}{8\omega_0}\right)^2+\beta^2\right]=\frac{F_0^2}{4m^2\omega_0^2}\,, \tag{28.71}$$

a cubic in A^2. Its peak lies on the backbone *$\omega=\omega_0+3\epsilon A^2/8\omega_0$ of Equation (28.81), bent to higher frequency for $\epsilon>0$ and to lower for $\epsilon<0$; and above a critical drive amplitude Equation (28.71) has three real roots over a range of ω.*

Rests on Equations (28.69) and (28.81).

Proof. *Derives Proposition 28.68.*

Put $\omega=\omega_0+\varepsilon_d$ with $|\varepsilon_d|\ll\omega_0$ and seek $x=A\cos(\omega t-\delta)$ with A and δ slowly varying. Substituting into Equation (28.69), using $\cos^3\psi=\frac{3}{4}\cos\psi+\frac{1}{4}\cos 3\psi$ and discarding the third harmonic — which is non-resonant and enters only at higher order — the coefficients of $\cos(\omega t-\delta)$ and $\sin(\omega t-\delta)$ give, after dividing by $2\omega_0$,

$$\left(\varepsilon_d-\frac{3\epsilon A^2}{8\omega_0}\right)A=\frac{F_0}{2m\omega_0}\cos\delta\,,\qquad \beta A=\frac{F_0}{2m\omega_0}\sin\delta\,.$$

Squaring and adding eliminates δ and gives Equation (28.71). At fixed F_0 the amplitude is greatest where the square bracket is least, that is where the first bracket vanishes, which is the backbone; and Equation (28.71) is a cubic in A^2 whose discriminant changes sign as F_0 grows, so three real positive roots appear above some F_0 and over some interval of ω. Setting $\epsilon=0$ recovers Equation (28.24) near resonance, and setting $\beta=0$ and $F_0=0$ recovers Equation (28.81). □

Where Equation (28.71) has three roots, the middle branch is unstable and the observed amplitude follows whichever outer branch it is already on. Sweeping the drive frequency upward therefore carries the amplitude along the bent nose until the branch ends, at which point it falls abruptly to the lower branch; sweeping downward it climbs the lower branch and jumps abruptly *up* at a different frequency. The response is hysteretic, the two jump frequencies differ, and neither is a property of the oscillator alone — all of which Duffing reports [452] and none of which a linear system can do.

All of it follows from one identity, which is worth isolating because it does the whole job: the determinant that decides the stability of a steady state is the *slope* of the response cubic at that state.

Proposition 28.69 (Which branch is stable, and where it ends)**.** *Write $s=A^2$, $\varepsilon_d=\omega-\omega_0$, $\kappa=3\epsilon/8\omega_0$ and $f=F_0/2m\omega_0$, and let*

$$G(s):=s\left[(\varepsilon_d-\kappa s)^2+\beta^2\right]-f^2 \tag{28.72}$$

be the cubic whose positive roots are the steady amplitudes of Equation (28.71). A steady state is asymptotically stable if $G'(s)>0$ and unstable if $G'(s)<0$. Where three roots exist the outer two are therefore stable and the middle one is a saddle; the two jump frequencies are the detunings at which $G(s)=G'(s)=0$, that is the turning points of the response curve; and three roots exist at all only for drives above

$$F_{0,\mathrm{c}}=\frac{16\,m\,\omega_0^{3/2}\beta^{3/2}}{3^{5/4}\,|\epsilon|^{1/2}}\,, \tag{28.73}$$

first reached at the detuning $\varepsilon_d=\sqrt{3}\,\beta\,\operatorname{sgn}\epsilon$.

Rests on Equation (28.71) and Theorem 32.13.

Proof. *Derives Proposition 28.69.*

Stability is a statement about the *approach* to a steady state, so the amplitude and phase must be allowed to move; the harmonic balance of Proposition 28.68 assumed them constant and cannot answer the question. Replace it by the exact change of variables

$$x = u(t)\cos\omega t + v(t)\sin\omega t\,, \qquad \dot{x} = \omega\left(-u\sin\omega t + v\cos\omega t\right)\,, \tag{28.74}$$

the second line being a constraint on the pair (u, v), equivalent to $\dot{u}\cos\omega t + \dot{v}\sin\omega t = 0$. Differentiating the second of Equation (28.74) gives $\ddot{x} = \omega\left(-\dot{u}\sin\omega t + \dot{v}\cos\omega t\right) - \omega^2 x$, so Equation (28.69) reads

$$\omega\left(-\dot{u}\sin\omega t + \dot{v}\cos\omega t\right) = \left(\omega^2 - \omega_0^2\right)x - 2\beta\dot{x} - \epsilon x^3 + \frac{F_0}{m}\cos\omega t =: S\,. \tag{28.75}$$

Solving Equation (28.75) together with the constraint — two linear equations for $\dot{u}$ and $\dot{v}$, of determinant ω — gives $\dot{u} = -S\sin(\omega t)/\omega$ and $\dot{v} = S\cos(\omega t)/\omega$, still exactly.

Now average over one drive period holding u and v fixed, which is legitimate because Equation (28.69) makes them move on the slow time scales β^{-1} and $\omega_0/|\epsilon|\,A^2$, both long compared with $2\pi/\omega$ under the hypotheses of Proposition 28.68. With $A^2 = u^2 + v^2$ the required averages are $\langle x\sin\omega t\rangle = v/2$, $\langle x\cos\omega t\rangle = u/2$, $\langle \dot{x}\sin\omega t\rangle = -\omega u/2$, $\langle \dot{x}\cos\omega t\rangle = \omega v/2$, $\langle\cos^2\omega t\rangle = 1/2$, and, since $x^3 = \frac{3}{4}A^2 x + (\text{third harmonic})$ by the same identity used above, $\langle x^3\sin\omega t\rangle = \frac{3}{8}A^2 v$ and $\langle x^3\cos\omega t\rangle = \frac{3}{8}A^2 u$. Using $\left(\omega^2 - \omega_0^2\right)/2\omega \approx \varepsilon_d$ and $\omega \approx \omega_0$ elsewhere,

$$\dot{u} = -\beta u - \Lambda v\,, \qquad \dot{v} = \Lambda u - \beta v + f\,, \qquad \Lambda := \varepsilon_d - \kappa A^2\,. \tag{28.76}$$

From Equations (28.69) and (28.75) (averaging the exact variation-of-parameters equations over one drive period).

Setting $\dot{u} = \dot{v} = 0$ and eliminating the phase reproduces $A^2\left(\Lambda^2 + \beta^2\right) = f^2$, which is Equation (28.71): the fixed points of Equation (28.76) are exactly the steady states already found, and nothing has been changed except that they may now be perturbed.

Linearize Equation (28.76) about a fixed point. Since $\partial\Lambda/\partial u = -2\kappa u$ and $\partial\Lambda/\partial v = -2\kappa v$, the Jacobian of Definition 32.12 is

$$J = \begin{pmatrix} -\beta + 2\kappa uv & -\Lambda + 2\kappa v^2 \\ \Lambda - 2\kappa u^2 & -\beta - 2\kappa uv \end{pmatrix}\,, \tag{28.77}$$

whose trace is $-2\beta < 0$ whatever the state, and whose determinant is

$$\det J = \beta^2 - 4\kappa^2 u^2 v^2 + \left(\Lambda - 2\kappa v^2\right)\left(\Lambda - 2\kappa u^2\right) = \Lambda^2 + \beta^2 - 2\kappa\Lambda s\,. \tag{28.78}$$

Differentiating Equation (28.72) gives $G'(s) = (\varepsilon_d - \kappa s)^2 + \beta^2 - 2\kappa s\,(\varepsilon_d - \kappa s) = \Lambda^2 + \beta^2 - 2\kappa\Lambda s$, so

$$\det J = G'(s)\,, \tag{28.79}$$

which is the identity announced. The rest is bookkeeping. A real 2×2 matrix with negative trace has both eigenvalues in the left half plane when its determinant is positive and one of each sign when it is negative, and away from $\det J = 0$ the two eigenvalues are distinct or the matrix is a multiple of the identity, so Theorem 32.13 applies: the state is asymptotically stable exactly when $G'(s) > 0$ and is a saddle when $G'(s) < 0$.

G has leading coefficient $\kappa^2 > 0$ and $G(0) = -f^2 < 0$, so it crosses zero once with positive slope, or three times with slopes positive, negative and positive in order of increasing s. That is the branch structure: outer branches stable, middle branch a saddle, and the transitions occur where a root is double, $G = G' = 0$. Those are the turning points of the response curve and hence the two jump frequencies; parametrized by Λ they are $s = \left(\Lambda^2 + \beta^2\right)/2\kappa\Lambda$ with $\left(\Lambda^2 + \beta^2\right)^2 = 2\kappa\Lambda f^2$ and $\varepsilon_d = \left(3\Lambda^2 + \beta^2\right)/2\Lambda$, which has solutions only for $\kappa\Lambda > 0$ —

the folds lie on the side towards which the backbone leans, high for a hardening spring and low for a softening one.

The two folds merge when G'' vanishes with them. From $G = \kappa^2 s^3 - 2\kappa\varepsilon_d s^2 + (\varepsilon_d^2 + \beta^2)\, s - f^2$, $G'' = 6\kappa^2 s - 4\kappa\varepsilon_d$ vanishes at $s = 2\varepsilon_d/3\kappa$, where $\Lambda = \varepsilon_d/3$ and $G' = \beta^2 - \varepsilon_d^2/3$; requiring $G' = 0$ as well gives $\varepsilon_d = \pm\sqrt{3}\beta$, and positivity of s selects the sign of κ, that is of ϵ. At that point $s = 2\beta/\sqrt{3}\,|\kappa|$ and $f^2 = s\,(\Lambda^2 + \beta^2) = 8\beta^3/3\sqrt{3}\,|\kappa|$, and restoring $\kappa = 3\epsilon/8\omega_0$ and $f = F_0/2m\omega_0$ gives $F_0^2 = 256\, m^2\omega_0^3\beta^3/9\sqrt{3}\,|\epsilon|$, which is Equation (28.73). Below that drive G is monotone and the response is single-valued; above it the two folds separate and the hysteresis loop opens. □

Remark 28.70 (The linear-stability theorem is used before it is proved). Theorem 32.13 belongs to Chapter 32, later in this part, and is used here for a two-dimensional flow whose Jacobian Equation (28.77) is constant-coefficient and generically diagonalizable. Nothing in that chapter depends on the present proposition, so the forward reference carries no circularity; it is placed there because the general theory of fixed points is that chapter's subject, and Remark 32.14 records what Part II still owes it.

28.7.2 Secular terms and the Lindstedt–Poincaré method

The obvious way to treat Equation (28.70) is to expand in ϵ. It fails, and the way it fails is instructive.

Proposition 28.71 (Naive perturbation theory produces a secular term). *Writing $x = x_0 + \epsilon x_1 + O(\epsilon^2)$ in Equation (28.70) with $x_0 = A\cos\omega_0 t$ gives*

$$x_1 = -\frac{3A^3}{8\omega_0}\, t\sin\omega_0 t + \frac{A^3}{32\omega_0^2}\,(\cos 3\omega_0 t - \cos\omega_0 t)\ , \tag{28.80}$$

whose first term grows without bound. The expansion is therefore useless once $\epsilon A^2 t/\omega_0$ is of order one, however small ϵ is.

Rests on Equations (28.2) and (28.70).

Proof. *Derives Proposition 28.71.*

At order ϵ, Equation (28.70) gives $\ddot{x}_1 + \omega_0^2 x_1 = -x_0^3$. With $x_0 = A\cos\omega_0 t$ and $\cos^3\psi = \frac{3}{4}\cos\psi + \frac{1}{4}\cos 3\psi$, the right side is $-A^3\left(\frac{3}{4}\cos\omega_0 t + \frac{1}{4}\cos 3\omega_0 t\right)$. The second piece drives the oscillator off resonance and contributes the bounded terms shown. The first drives it exactly on resonance, and the particular solution of $\ddot{y} + \omega_0^2 y = C\cos\omega_0 t$ is $y = (C/2\omega_0)\, t\sin\omega_0 t$, which with $C = -3A^3/4$ gives the first term of Equation (28.80). Since ϵx_1 must stay small against x_0 for the expansion to mean anything, the bound is $\epsilon A^2 t/\omega_0 \ll 1$. □

The secular term is not a physical instability. The exact motion of Equation (28.70) is bounded, because the energy $\frac{1}{2}\dot{x}^2 + \frac{1}{2}\omega_0^2 x^2 + \frac{1}{4}\epsilon x^4$ is conserved and, for $\epsilon > 0$, confines the orbit. What has gone wrong is that the expansion assumed the frequency was ω_0, and it is not: $t\sin\omega_0 t$ is the first term of the expansion of $\sin\left((\omega_0 + \epsilon\omega_1)\, t\right)$ in ϵ. The cure is to let the frequency be part of what is expanded — Lindstedt's device [219], made into a systematic theory by Poincaré [255].

Theorem 28.72 (Lindstedt–Poincaré frequency shift). *The periodic solutions of Equation (28.70) of amplitude A have frequency*

$$\omega = \omega_0 + \frac{3\epsilon A^2}{8\omega_0} + O\left(\epsilon^2\right)\ , \tag{28.81}$$

and to the same order the motion is $x = A\cos\omega t + (\epsilon A^3/32\omega_0^2)\,(\cos 3\omega t - \cos\omega t)$, free of secular terms.

Rests on Equations (28.70) and (28.80).

Proof. *Derives Theorem 28.72.*

Introduce the stretched time $\tau = \omega t$ with ω unknown, and expand both the solution and the frequency,

$$x = x_0(\tau) + \epsilon x_1(\tau) + \cdots , \qquad \omega = \omega_0 + \epsilon\omega_1 + \cdots ,$$

requiring every x_j to be 2π-periodic in τ — which is the whole of the method: periodicity in τ is imposed, and ω is whatever makes it possible. Since $\mathrm{d}/\mathrm{d}t = \omega\,\mathrm{d}/\mathrm{d}\tau$, Equation (28.70) becomes $\omega^2 x'' + \omega_0^2 x + \epsilon x^3 = 0$ with the prime denoting $\mathrm{d}/\mathrm{d}\tau$. At order ϵ^0, $x_0'' + x_0 = 0$, so $x_0 = A\cos\tau$ after choosing the origin of time. At order ϵ^1,

$$\omega_0^2\,(x_1'' + x_1) = -2\omega_0\omega_1 x_0'' - x_0^3 = 2\omega_0\omega_1 A\cos\tau - A^3\left(\tfrac{3}{4}\cos\tau + \tfrac{1}{4}\cos 3\tau\right) .$$

A term in $\cos\tau$ on the right drives the left-hand operator on resonance and produces exactly the secular term of Proposition 28.71, destroying the periodicity demanded. It is absent if and only if $2\omega_0\omega_1 A = \frac{3}{4}A^3$, that is $\omega_1 = 3A^2/8\omega_0$, which is Equation (28.81). With that choice the remaining equation $x_1'' + x_1 = -\left(A^3/4\omega_0^2\right)\cos 3\tau$ has the bounded periodic solution $x_1 = \left(A^3/32\omega_0^2\right)\cos 3\tau$, to which a multiple of $\cos\tau$ may be added; fixing the amplitude of the fundamental at A makes it $\left(A^3/32\omega_0^2\right)(\cos 3\tau - \cos\tau)$. □

Three things in Theorem 28.72 outlast the example. The frequency shift is quadratic in the amplitude, so an anharmonic oscillator is still isochronous to first order and departs only at second — which is why Phenomenon 28.9 works as well as it does. The response acquires a third harmonic: a nonlinear system driven at one frequency answers at others, which no linear system can do and which is the whole of nonlinear optics (Chapter 61). And the method itself is a renormalization: a parameter of the unperturbed problem is adjusted order by order to absorb what would otherwise be a divergence, which is the same manoeuvre, in a far more elaborate setting, as the mass and charge renormalizations of Chapter 100.

Remark 28.73 (The method's own limits, and where they lead). Theorem 28.72 removes the secular terms order by order and says nothing about whether the resulting series converges. For a single oscillator the answer is benign. For several coupled ones it is not: the analogue of the resonance denominator that produced Equation (28.80) becomes a combination $n_1\omega_1 + \cdots + n_k\omega_k$ that can be made arbitrarily small by choosing the integers n_j, so the corrections are divided by arbitrarily small numbers and the series need not converge at all. Poincaré identified the problem [255]; that the series nevertheless converges for most frequency ratios, and that the corresponding motions survive a small perturbation while the resonant ones do not, is the theorem of Kolmogorov, Arnold and Moser [1247]. Resonance overlap, the destruction of the surviving tori and the onset of chaos are Chapter 32, which is where this chapter hands the anharmonic oscillator on.

Rests on Equation (28.81).

Chapter 29

Rigid Bodies and Rotating Frames

A rigid body is the first mechanical system beyond the point particle that Nature actually supplies: six degrees of freedom, three of them rotational, governed by an inertia tensor rather than by a single mass. This chapter develops rigid-body kinematics and dynamics — the inertia tensor, Euler's equations, tops and gyroscopes — and then turns to motion described from rotating frames, where the centrifugal and Coriolis forces appear and carry the terrestrial evidence of the Earth's rotation: Foucault's pendulum, the eastward deflection of falling bodies, the circulation of the winds, and the Chandler wobble. It builds directly on the dynamics of Chapter 19 and the Lagrangian methods of Chapter 21; the rotation group it works with is treated as algebra in Chapter 4 and as geometry in Chapter 14. Standard treatments are [1585, 2027].

Two conventions are fixed once and used everywhere below. Body axes are always the principal axes of Theorem 29.16, labelled so that I_1, I_2, I_3 are the corresponding moments; and Euler angles are always the z–x–z triple of Notation 29.9. Neither choice is forced by the physics, and a formula quoted from another book will differ from the formulae here if its conventions differ.

29.1 Rigid-body kinematics

29.1.1 Angular velocity and Euler's theorem

Definition 29.1 (Rigid body)**.** A *rigid body* is a system of particles, or a continuous mass distribution, subject to the holonomic constraints

$$|\boldsymbol{r}_a - \boldsymbol{r}_b| = \text{const} \qquad \text{for every pair } a, b\,. \tag{29.1}$$

Rests on Definitions 19.39 and 21.2.

Proposition 29.2 (Six degrees of freedom)**.** *A rigid body containing at least three non-collinear particles has exactly six degrees of freedom: three to locate one of its points and three more to orient it. Its configuration space is* $\mathbb{R}^3 \times \mathrm{SO}(3)$.

Rests on Definitions 21.9 and 29.1.

Proof. *Derives Proposition 29.2.*

Fix three non-collinear particles of the body. Locating the first costs three coordinates. The second is then confined by Equation (29.1) to a sphere about the first, costing two more; the third is confined to a circle about the axis through the first two, costing one more — six in all. Every further particle is then determined, because its distances from three non-collinear points fix it up to a reflection, and a reflection cannot be reached continuously from the identity. The six coordinates split as the position of the chosen point, an element of $\mathbb{R}^3$, and the orientation of the body frame relative to the space frame, which by Definition 14.26 is an element of $\mathrm{SO}(3)$. □

Theorem 29.3 (Euler's rotation theorem)**.** *Every displacement of a rigid body that leaves one point fixed is a rotation through some angle about some axis through that point. That axis is the eigenvector of the displacement matrix belonging to the eigenvalue* $+1$*.*

Rests on Propositions 14.30 and 29.2.

Proof. *Derives Theorem 29.3.*

The displacement preserves all mutual distances and one point, so in Cartesian axes centred on that point it acts as a matrix R with $R^{\mathsf{T}}R = \mathbb{1}$, and it is continuously connected to the identity, so $\det R = +1$: that is, $R \in \mathrm{SO}(3)$ (Definition 14.26). By the Rodrigues formula of Proposition 14.30 the exponential map onto $\mathrm{SO}(3)$ is surjective, so $R = \exp(\theta\, \hat{\boldsymbol{n}} \cdot \boldsymbol{J})$ for some unit $\hat{\boldsymbol{n}}$ and some angle θ; this is by construction the rotation by θ about $\hat{\boldsymbol{n}}$, and $R\hat{\boldsymbol{n}} = \hat{\boldsymbol{n}}$. The general rotation formulae in this form are Euler's [34]. □

Definition 29.4 (Angular velocity)**.** Let the orthonormal body axes be $\hat{\boldsymbol{e}}_i(t) = R(t)\,\hat{\boldsymbol{e}}_i(0)$ with $R(t) \in \mathrm{SO}(3)$. Differentiating $RR^{\mathsf{T}} = \mathbb{1}$ shows that $\dot{R}R^{\mathsf{T}}$ is antisymmetric, hence an element of $\mathfrak{so}(3)$; the *angular velocity* $\boldsymbol{\omega}(t)$ is the vector representing it,

$$\left(\dot{R}R^{\mathsf{T}}\right)_{ij} = -\epsilon_{ijk}\,\omega_k\,, \qquad \text{equivalently} \qquad \dot{R}R^{\mathsf{T}}\boldsymbol{v} = \boldsymbol{\omega} \times \boldsymbol{v} \quad \text{for all } \boldsymbol{v}\,. \tag{29.2}$$

Rests on Definition 14.26, Proposition 14.27, and Lemma 14.62.

The identification of an antisymmetric 3×3 matrix with a vector is the isomorphism $\mathfrak{so}(3) \cong \mathbb{R}^3$ of Proposition 14.27, and it is available in three space dimensions and in no other; that is the technical reason why angular velocity is a vector here and a two-index object in general.

Proposition 29.5 (The velocity field of a rigid body)**.** *If* $\boldsymbol{r}$ *is measured from a point* O *of the body, every material point moves with*

$$\boldsymbol{v} = \boldsymbol{v}_O + \boldsymbol{\omega} \times \boldsymbol{r}\,, \tag{29.3}$$

where $\boldsymbol{v}_O$ *is the velocity of* O *[32].*

Rests on Definition 29.4 and Equation (29.1).

Proof. *Derives Proposition 29.5.*

Write the position of a material point as $\boldsymbol{R}(t) = \boldsymbol{R}_O(t) + R(t)\boldsymbol{r}_0$, with $\boldsymbol{r}_0$ its fixed body coordinates — fixed precisely because of Equation (29.1). Differentiating, $\dot{\boldsymbol{R}} = \boldsymbol{v}_O + \dot{R}\boldsymbol{r}_0 = \boldsymbol{v}_O + \dot{R}R^{\mathsf{T}}(R\boldsymbol{r}_0)$, and by Equation (29.2) the second term is $\boldsymbol{\omega} \times \boldsymbol{r}$ with $\boldsymbol{r} = R\boldsymbol{r}_0$. □

Proposition 29.6 (The angular velocity does not depend on the base point)**.** *If Equation (29.3) holds with respect to* O *and also with respect to a second body point* O'*, the two angular velocities coincide. There is therefore one* $\boldsymbol{\omega}$ *for the body, not one per reference point.*

Rests on Proposition 29.5.

Proof. *Derives Proposition 29.6.*

Let $\boldsymbol{d}$ be the vector from O to O'. Applying Equation (29.3) about each point and subtracting, $(\boldsymbol{\omega} - \boldsymbol{\omega}') \times (\boldsymbol{r} - \boldsymbol{d}) = \boldsymbol{0}$ for every material point. A body with three non-collinear points supplies vectors $\boldsymbol{r} - \boldsymbol{d}$ spanning $\mathbb{R}^3$, and only the zero vector has vanishing cross product with a spanning set, so $\boldsymbol{\omega} = \boldsymbol{\omega}'$. □

Theorem 29.7 (Transport theorem)**.** *Let* $\boldsymbol{Q}(t)$ *be any vector, and let a frame rotate with angular velocity* $\boldsymbol{\omega}$ *relative to an inertial frame. Then*

$$\left(\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}t}\right)_{space} = \left(\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}t}\right)_{body} + \boldsymbol{\omega} \times \boldsymbol{Q}\,, \tag{29.4}$$

where the second derivative is taken holding the rotating basis fixed.

Rests on Definition 29.4 and Proposition 29.5.

Proof. *Derives Theorem 29.7.*

Expand $\boldsymbol{Q} = Q_i \hat{\boldsymbol{e}}_i$ in the rotating basis and differentiate in the inertial frame:

$$\left(\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}t}\right)_{\text{space}} = \dot{Q}_i \hat{\boldsymbol{e}}_i + Q_i \frac{\mathrm{d}\hat{\boldsymbol{e}}_i}{\mathrm{d}t}\,.$$

The first sum is by definition the derivative computed by an observer who regards the basis as fixed. For the second, $\hat{\boldsymbol{e}}_i = R\hat{\boldsymbol{e}}_i(0)$ gives $\mathrm{d}\hat{\boldsymbol{e}}_i/\mathrm{d}t = \dot{R}R^{\mathsf{T}}\hat{\boldsymbol{e}}_i = \boldsymbol{\omega} \times \hat{\boldsymbol{e}}_i$ by Equation (29.2), so the second sum is $\boldsymbol{\omega} \times \boldsymbol{Q}$. □

Theorem 29.7 is the single technical fact behind everything in Sections 29.3 and 29.5: applied to the angular momentum it produces Euler's equations, and applied twice to a position vector it produces the centrifugal and Coriolis terms.

Proposition 29.8 (Chasles: the general displacement is a screw)**.** *Every displacement of a free rigid body can be realised as a rotation about some axis together with a translation* along *that same axis [87].*

Rests on Theorem 29.3 and Proposition 29.5.

Proof. *Derives Proposition 29.8.*

By Theorem 29.3 the displacement is $\boldsymbol{r} \mapsto R\boldsymbol{r} + \boldsymbol{a}$ with R a rotation by θ about a unit vector $\hat{\boldsymbol{n}}$ through the origin. Split $\boldsymbol{a} = a_{\parallel}\hat{\boldsymbol{n}} + \boldsymbol{a}_{\perp}$. Shifting the axis to pass through a point $\boldsymbol{c}$ replaces $\boldsymbol{a}$ by $\boldsymbol{a} + (\mathbb{1} - R)\,\boldsymbol{c}$. Now $\mathbb{1} - R$ annihilates $\hat{\boldsymbol{n}}$ and, for θ not a multiple of 2π, is invertible on the plane orthogonal to $\hat{\boldsymbol{n}}$; choose $\boldsymbol{c}$ in that plane with $(\mathbb{1} - R)\,\boldsymbol{c} = -\boldsymbol{a}_{\perp}$. About the shifted axis the translation is $a_{\parallel}\hat{\boldsymbol{n}}$, parallel to the axis. If θ is a multiple of 2π the displacement is a pure translation, the degenerate screw of infinite pitch. □

29.1.2 Euler angles

Notation 29.9 (The z–x–z Euler angles used in this treatise)*.* Let A be the matrix carrying the components of a vector in the space frame to its components in the body frame. This treatise writes

$$A\,(\varphi, \theta, \psi) = R_3(\psi)\,R_1(\theta)\,R_3(\varphi)\,, \tag{29.5}$$

that is: rotate by φ about the space z axis; then by θ about the new x axis, the *line of nodes*; then by ψ about the new z axis, which is the body axis $\hat{\boldsymbol{e}}_3$. The ranges are $\varphi \in [0, 2\pi)$, $\theta \in [0, \pi]$, $\psi \in [0, 2\pi)$, and the elementary matrices are

$$R_3(\alpha) = \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad R_1(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix}.$$

Multiplying out Equation (29.5),

$$A = \begin{pmatrix} \cos\psi\cos\varphi - \cos\theta\sin\varphi\sin\psi & \cos\psi\sin\varphi + \cos\theta\cos\varphi\sin\psi & \sin\psi\sin\theta \\ -\sin\psi\cos\varphi - \cos\theta\sin\varphi\cos\psi & -\sin\psi\sin\varphi + \cos\theta\cos\varphi\cos\psi & \cos\psi\sin\theta \\ \sin\theta\sin\varphi & -\sin\theta\cos\varphi & \cos\theta \end{pmatrix}. \tag{29.6}$$

Rests on Theorem 29.3 and Definition 14.26.

Remark 29.10 (Other conventions, and why the choice must be stated)*.* Twelve orderings of axes are in use, and both the active and the passive reading of the matrices occur. Landau's treatment uses the same z–x–z triple as Equation (29.5) [1585]; aeronautics prefers a z–y–x triple whose angles are yaw, pitch and roll and are *not* the angles here. A formula lifted from another source into this chapter is therefore wrong unless its convention is checked first. Every Euler-angle formula below — Equations (29.7) and (29.35) in particular — is stated in the convention of Notation 29.9 and in no other.

Rests on Notation 29.9.

Proposition 29.11 (Angular velocity in Euler angles)**.** *In the convention of Notation 29.9, the components of* $\boldsymbol{\omega}$ *along the body axes are*

$$\omega_1 = \dot\varphi \sin\theta \sin\psi + \dot\theta \cos\psi \,, \tag{29.7}$$

$$\omega_2 = \dot\varphi \sin\theta \cos\psi - \dot\theta \sin\psi \,, \tag{29.8}$$

$$\omega_3 = \dot\varphi \cos\theta + \dot\psi \,. \tag{29.9}$$

Rests on Notation 29.9 and Definition 29.4.

Proof. *Derives Proposition 29.11.*

The three elementary rotations are performed about the space z axis, about the line of nodes $\hat{\boldsymbol{e}}_N$, and about the body axis $\hat{\boldsymbol{e}}_3$, so the angular velocities add:

$$\boldsymbol{\omega} = \dot\varphi\, \hat{\boldsymbol{z}} + \dot\theta\, \hat{\boldsymbol{e}}_N + \dot\psi\, \hat{\boldsymbol{e}}_3 \,.$$

It remains to resolve each of the three axes along the body frame. The third column of Equation (29.6) gives the body components of $\hat{\boldsymbol{z}}$, namely $(\sin\theta\sin\psi,\ \sin\theta\cos\psi,\ \cos\theta)$. The line of nodes is the x axis of the intermediate frame, so applying only the last rotation $R_3(\psi)$ to $(1,0,0)$ gives $(\cos\psi,\ -\sin\psi,\ 0)$. Finally $\hat{\boldsymbol{e}}_3 = (0,0,1)$ in the body frame. Collecting components yields Equations (29.7) to (29.9). □

Remark 29.12 (The coordinate singularity, and what it is not)*.* Read as a linear map from $\left(\dot\varphi, \dot\theta, \dot\psi\right)$ to $(\omega_1, \omega_2, \omega_3)$, Equations (29.7) to (29.9) have the matrix

$$M = \begin{pmatrix} \sin\theta\sin\psi & \cos\psi & 0 \\ \sin\theta\cos\psi & -\sin\psi & 0 \\ \cos\theta & 0 & 1 \end{pmatrix}, \qquad \det M = -\sin\theta \,.$$

At $\theta = 0$ and $\theta = \pi$ the map is singular: only the sum $\varphi \pm \psi$ is determined, the two rotations having become rotations about the same axis. This is the defect known in instrument engineering as gimbal lock, and it is a property of the *chart*, not of the body: nothing whatever happens to a top as its axis passes through the vertical. No single chart can do better, because SO(3) is a compact three-manifold (Chapter 14) and therefore admits no global coordinate system; a numerical integration that must pass through $\theta = 0$ changes chart, or abandons Euler angles for the group element itself.

Rests on Propositions 14.31 and 29.11.

29.2 The inertia tensor

29.2.1 Definition and principal axes

Definition 29.13 (Inertia tensor, continuum form)**.** For a body of mass density $\rho(\boldsymbol{r})$ occupying a region V, with $\boldsymbol{r}$ measured from a chosen point O, the *inertia tensor* about O is

$$I_{ik} := \int_V \rho(\boldsymbol{r}) \left(r^2 \delta_{ik} - x_i x_k \right) \mathrm{d}^3 r \,, \tag{29.10}$$

which is Equation (19.34) with the sum over particles replaced by an integral over the body. The two are the same quantity for the same reason and are used interchangeably below, but neither expression is a special case of the other as written: a finite set of point masses is not a bounded density, so the discrete form is not obtained by substituting a ρ into Equation (29.10) — it is the same construction with the mass distribution summed rather than integrated. Where a result below holds for only one of the two, the statement says which. Its SI unit is $\mathrm{kg\,m^2}$. It is symmetric, and under a rotation of the axes its components transform as a Cartesian tensor of second rank (Definition 13.5).

Rests on Definitions 7.93, 13.5 and 19.42.

Proposition 29.14 (Angular momentum and kinetic energy)**.** *If the point O is fixed, or is the centre of mass, then*

$$L_i = I_{ik}\,\omega_k\,, \tag{29.11}$$

and the kinetic energy of rotation is the quadratic form

$$T_{rot} = \tfrac{1}{2} I_{ik}\,\omega_i\omega_k = \tfrac{1}{2}\boldsymbol{\omega}\cdot\boldsymbol{L}\,. \tag{29.12}$$

For a general motion the total kinetic energy separates as $T = \tfrac{1}{2}Mv_{cm}^2 + T_{rot}$, with T_{rot} computed about the centre of mass.

Rests on Definitions 19.55 and 29.13 and Proposition 29.5.

Proof. *Derives Proposition 29.14.*

With O fixed, Equation (29.3) gives $\boldsymbol{v} = \boldsymbol{\omega}\times\boldsymbol{r}$, so

$$\boldsymbol{L} = \int_V \rho\,\boldsymbol{r}\times(\boldsymbol{\omega}\times\boldsymbol{r})\,\mathrm{d}^3r = \int_V \rho\left(r^2\boldsymbol{\omega} - (\boldsymbol{r}\cdot\boldsymbol{\omega})\,\boldsymbol{r}\right)\mathrm{d}^3r\,,$$

using the triple-product expansion $\boldsymbol{a}\times(\boldsymbol{b}\times\boldsymbol{c}) = \boldsymbol{b}(\boldsymbol{a}\cdot\boldsymbol{c}) - \boldsymbol{c}(\boldsymbol{a}\cdot\boldsymbol{b})$, which is the contraction $\epsilon_{ijk}\epsilon_{ilm} = \delta_{jl}\delta_{km} - \delta_{jm}\delta_{kl}$ of the Levi-Civita symbol of Lemma 14.62. That contraction is quoted here and not proved: Part II states the invariance of the symbol but carries no labelled form of the identity, so it is used on its own authority. In components the result is exactly Equation (29.10) contracted with ω_k, which is Equation (29.11). For the energy,

$$T = \tfrac{1}{2}\int_V \rho\,|\boldsymbol{\omega}\times\boldsymbol{r}|^2\,\mathrm{d}^3r = \tfrac{1}{2}\int_V \rho\left(\omega^2r^2 - (\boldsymbol{\omega}\cdot\boldsymbol{r})^2\right)\mathrm{d}^3r = \tfrac{1}{2}I_{ik}\omega_i\omega_k\,,$$

by the same identity. The separation of the general motion is the centre-of-mass decomposition of Chapter 19: the cross term carries the factor $\int_V \rho\,\boldsymbol{r}\,\mathrm{d}^3r$, which vanishes when $\boldsymbol{r}$ is measured from the centre of mass. □

Proposition 29.15 (Positivity)**.** *For any unit vector $\hat{\boldsymbol{n}}$,*

$$\hat{\boldsymbol{n}}\cdot I\hat{\boldsymbol{n}} = \int_V \rho\,d^2\,\mathrm{d}^3r \;\geq\; 0\,,$$

where d is the distance of the mass element from the line through O along $\hat{\boldsymbol{n}}$. The form is strictly positive unless the whole mass lies on that line; a body that is not collinear therefore has a positive-definite inertia tensor.

Rests on Definition 29.13.

Proof. *Derives Proposition 29.15.*

Contract Equation (29.10) with $\hat{\boldsymbol{n}}$: $n_in_k\left(r^2\delta_{ik} - x_ix_k\right) = r^2 - (\boldsymbol{r}\cdot\hat{\boldsymbol{n}})^2 = |\boldsymbol{r}|^2\sin^2\alpha = d^2$, with α the angle between $\boldsymbol{r}$ and $\hat{\boldsymbol{n}}$. The integrand is non-negative, and vanishes identically only where ρ is supported on the line $d = 0$. □

Theorem 29.16 (Principal axes)**.** *For every rigid body and every choice of the point O there exists an orthonormal triad $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \hat{\boldsymbol{e}}_3$ fixed in the body in which I_{ik} is diagonal,*

$$I = \operatorname{diag}(I_1, I_2, I_3)\,, \qquad I_i \geq 0\,. \tag{29.13}$$

These are the principal axes *and the I_i the* principal moments of inertia *[29]. In such axes $\boldsymbol{L} = (I_1\omega_1, I_2\omega_2, I_3\omega_3)$ and $2T_{rot} = \sum_i I_i\omega_i^2$.*

Rests on Definition 29.13, Theorem 5.75, and Proposition 29.15.

Proof. *Derives Theorem 29.16.*

I_{ik} is a real symmetric matrix, hence self-adjoint with respect to the Euclidean inner product. The spectral theorem (Theorem 5.75) supplies an orthonormal basis of eigenvectors with real eigenvalues, and Corollary 5.27 makes the decomposition orthogonal, so the change of basis is itself a rotation once the triad is ordered to be right-handed. The eigenvalues are the values of the form $\hat{\boldsymbol{n}} \cdot I\hat{\boldsymbol{n}}$ on the eigenvectors, and these are non-negative by Proposition 29.15. The remaining formulae are Equations (29.11) and (29.12) evaluated in the diagonalizing basis. Historically the existence of three orthogonal principal axes is Segner's [29]; the statement here is the modern one, and the mathematics belongs to Chapter 5 rather than to mechanics. □

The eigenvalue problem of Theorem 29.16 is the whole content of the passage from "a body" to "three numbers". Note that $\boldsymbol{L}$ and $\boldsymbol{\omega}$ are parallel only when $\boldsymbol{\omega}$ lies along a principal axis, or when two or three of the I_i coincide; in general the angular momentum of a spinning body does not point along its spin.

Proposition 29.17 (Triangle inequalities)**.** *The principal moments satisfy*

$$I_1 + I_2 \geq I_3\,, \qquad I_2 + I_3 \geq I_1\,, \qquad I_3 + I_1 \geq I_2\,,$$

with equality in the first exactly when the whole mass lies in the $\hat{\boldsymbol{e}}_1\hat{\boldsymbol{e}}_2$ plane.
Rests on Theorem 29.16.

Proof. *Derives Proposition 29.17.*

In principal axes $I_1 = \int \rho\,(y^2+z^2)$, $I_2 = \int \rho\,(z^2+x^2)$ and $I_3 = \int \rho\,(x^2+y^2)$, all integrals over V with measure d^3r. Hence $I_1 + I_2 - I_3 = 2\int_V \rho\, z^2\,\mathrm{d}^3r \geq 0$, which vanishes precisely when ρ is supported on $z = 0$. The other two follow by relabelling. □

Definition 29.18 (Classification of tops)**.** A body is a *spherical top* if $I_1 = I_2 = I_3$, a *symmetric top* if exactly two of the principal moments are equal, and an *asymmetric top* if all three differ. For a symmetric top the two equal moments are written A and the third C; the body is called *prolate* if $C < A$ (a cigar) and *oblate* if $C > A$ (a coin).
Rests on Theorem 29.16.

Definition 29.19 (Inertia ellipsoid)**.** The *inertia ellipsoid* of a body at O is the level surface

$$I_{ik}x_i x_k = 1\,, \tag{29.14}$$

a genuine ellipsoid whenever I is positive definite, with semi-axes $I_i^{-1/2}$ along the principal axes.
Rests on Definition 29.13, Proposition 29.15, and Theorem 29.16.

The ellipsoid has a symmetry axis exactly when the body is a symmetric top, and is a sphere exactly for a spherical top. A cube is a spherical top about its centre, so its inertia ellipsoid is a sphere and every axis through the centre is principal — an example worth keeping in mind, because it shows that the symmetry of the inertia tensor can exceed the symmetry of the body.

29.2.2 Computing moments of inertia

Theorem 29.20 (Parallel-axis theorem)**.** *Let I^{cm}_{ik} be the inertia tensor about the centre of mass and $\boldsymbol{a}$ the vector from a point O to the centre of mass. Then the inertia tensor about O is*

$$I_{ik} = I^{cm}_{ik} + M\left(a^2\delta_{ik} - a_i a_k\right)\,. \tag{29.15}$$

In particular the moment about an axis $\hat{\boldsymbol{n}}$ through O exceeds the moment about the parallel axis through the centre of mass by Md^2, with d the distance between the two axes.
Rests on Definitions 19.47 and 29.13.

Proof. *Derives Theorem 29.20.*

Write $\boldsymbol{r} = \boldsymbol{a} + \boldsymbol{r}'$, with $\boldsymbol{r}'$ measured from the centre of mass, so that $\int_V \rho\, \boldsymbol{r}'\, \mathrm{d}^3 r = \boldsymbol{0}$ by Definition 19.47. Substituting into Equation (29.10),

$$I_{ik} = \int_V \rho \left[\left(a^2 + 2\boldsymbol{a}\cdot\boldsymbol{r}' + r'^2 \right) \delta_{ik} - (a_i + x'_i)\,(a_k + x'_k) \right] \mathrm{d}^3 r \,.$$

Every term linear in $\boldsymbol{r}'$ integrates to zero. What remains is the $\boldsymbol{r}'$-quadratic part, which is I^{cm}_{ik}, plus the $\boldsymbol{a}$-quadratic part times the total mass, which is $M\left(a^2\delta_{ik} - a_i a_k\right)$. Contracting Equation (29.15) with $\hat{\boldsymbol{n}}$ twice gives $Ma^2 - (\boldsymbol{a}\cdot\hat{\boldsymbol{n}})^2 M = Md^2$ for the added term. □

Corollary 29.21 (The centre of mass minimizes the moment)**.** *Among all axes of a given direction, the one through the centre of mass has the smallest moment of inertia.*

Rests on Theorem 29.20.

Proof. *Derives Corollary 29.21.*

The added term Md^2 of Theorem 29.20 is non-negative and vanishes only for $d = 0$. □

This is the mathematical content of Huygens' theory of the compound pendulum and of the centre of oscillation [11]; the experimental use made of it, in Kater's reversible pendulum, is Phenomenon 34.8 and Equation (34.21).

Proposition 29.22 (Perpendicular-axis theorem)**.** *For a plane lamina lying in the xy plane,*

$$I_{zz} = I_{xx} + I_{yy} \,. \tag{29.16}$$

Rests on Definition 29.13 and Proposition 29.17.

Proof. *Derives Proposition 29.22.*

With $z = 0$ throughout the body, Equation (29.10) gives $I_{xx} = \int \rho y^2$, $I_{yy} = \int \rho x^2$ and $I_{zz} = \int \rho\,(x^2 + y^2)$, all over V with measure $\mathrm{d}^3 r$; adding the first two gives the third. It is the equality case of Proposition 29.17. □

The disc entries of Table 29.1 illustrate Proposition 29.22: a disc is a lamina, its two diameters are equivalent, and $MR^2/4 + MR^2/4 = MR^2/2$ is the moment about the symmetry axis, as required.

Example 29.23 (The polar moment of the Earth, and what it reveals)**.** The Earth's moment of inertia about its rotation axis is $C = 8.03 \times 10^{37}\,\mathrm{kg\,m^2}$, against a mass $M = 5.972 \times 10^{24}\,\mathrm{kg}$ and an equatorial radius $a = 6.378 \times 10^{6}\,\mathrm{m}$, so that

$$\frac{C}{Ma^2} = 0.331 \,.$$

Table 29.1 gives 0.4 for a homogeneous sphere. The measured value is seventeen per cent smaller, and it cannot be made smaller by any redistribution that keeps the density uniform: the deficit is a direct measurement, made from the outside, that the Earth's mass is concentrated towards its centre. The same number enters Sections 29.3.4 and 29.4.4, where it fixes the free wobble and the precession rate; the numerical values here are the conventional astronomical ones collected in Table 29.2.

Rests on Definition 29.13 and Table 29.1.

Proposition 29.24 (MacCullagh's formula)**.** *Let a body of mass m, centred on its centre of mass, be viewed from a distant point mass M at position $\boldsymbol{R}$ with R much larger than the size of the body. To quadrupole order the mutual gravitational potential energy is*

$$U = -\frac{GMm}{R} - \frac{GM}{2R^3}\left(I_1 + I_2 + I_3 - 3I_R\right) \,, \tag{29.17}$$

where $I_R = \hat{\boldsymbol{R}}\cdot I\hat{\boldsymbol{R}}$ is the moment of inertia of the body about the line joining the two centres.

Rests on Definition 29.13 and Theorem 9.113.

Table 29.1: Principal moments of inertia of homogeneous bodies of total mass M, about axes through the centre of mass. Each entry follows from Equation (29.10) by a single elementary integral; the last column gives the ratio $I/(MR^2)$ or $I/(M\ell^2)$, which is the dimensionless number that actually distinguishes one shape from another. The values for the rod and the lamina are limits, exact when the neglected dimension is small compared with the one retained.

Body	Axis	I	ratio
Thin rod, length ℓ	perpendicular, through centre	$M\ell^2/12$	0.0833
Thin rod, length ℓ	perpendicular, through one end	$M\ell^2/3$	0.3333
Thin ring, radius R	symmetry axis	MR^2	1
Disc, radius R	symmetry axis	$MR^2/2$	0.5
Disc, radius R	diameter	$MR^2/4$	0.25
Solid cylinder, radius R, length ℓ	transverse, through centre	$M\left(3R^2+\ell^2\right)/12$	—
Solid sphere, radius R	any diameter	$2MR^2/5$	0.4
Spherical shell, radius R	any diameter	$2MR^2/3$	0.6667
Rectangular plate, sides a, b	perpendicular, through centre	$M\left(a^2+b^2\right)/12$	—
Solid ellipsoid, semi-axes a, b, c	the a axis	$M\left(b^2+c^2\right)/5$	—

Proof. *Derives Proposition 29.24.*

Write $U=-GM\int_V \rho(\boldsymbol{r})\,|\boldsymbol{R}-\boldsymbol{r}|^{-1}\,\mathrm{d}^3r$ and expand the kernel in Legendre polynomials by the generating function of Theorem 9.113, with $\cos\gamma=\hat{\boldsymbol{R}}\cdot\hat{\boldsymbol{r}}$:

$$\frac{1}{|\boldsymbol{R}-\boldsymbol{r}|}=\frac{1}{R}\sum_{n\ge 0}\left(\frac{r}{R}\right)^n P_n(\cos\gamma)=\frac{1}{R}+\frac{\boldsymbol{r}\cdot\hat{\boldsymbol{R}}}{R^2}+\frac{3\left(\boldsymbol{r}\cdot\hat{\boldsymbol{R}}\right)^2-r^2}{2R^3}+O\!\left(\frac{r^3}{R^4}\right).$$

The $n=0$ term integrates to m/R. The $n=1$ term vanishes because $\boldsymbol{r}$ is measured from the centre of mass. For $n=2$, note from Equation (29.10) that $I_1+I_2+I_3=\operatorname{tr} I=2\int_V \rho\, r^2\,\mathrm{d}^3r$ and that $I_R=\int_V \rho\left(r^2-\left(\boldsymbol{r}\cdot\hat{\boldsymbol{R}}\right)^2\right)\mathrm{d}^3r$, whence

$$\int_V \rho\left(3\left(\boldsymbol{r}\cdot\hat{\boldsymbol{R}}\right)^2-r^2\right)\mathrm{d}^3r=3\left(\tfrac{1}{2}\operatorname{tr} I-I_R\right)-\tfrac{1}{2}\operatorname{tr} I=\operatorname{tr} I-3I_R\,.$$

Multiplying by $-GM/(2R^3)$ gives Equation (29.17). □

Proposition 29.24 is the bridge from this section to Section 29.4.4: it says that a distant mass feels the orientation of a body only through the combination $\operatorname{tr} I-3I_R$, and therefore exerts on it a torque proportional to the departure of the inertia tensor from a multiple of the identity. A spherical top feels no such torque at all, which is why the precession of the equinoxes measures the Earth's equatorial bulge and nothing else.

29.3 Euler's equations and free rotation

29.3.1 Euler's equations

Theorem 29.25 (Euler's equations)**.** *Referred to the principal axes of a body, and about a point that is either fixed or the centre of mass, the rotational equation of motion* $\mathrm{d}\boldsymbol{L}/\mathrm{d}t=\boldsymbol{\tau}$

reads

$$I_1\dot{\omega}_1 - (I_2 - I_3)\,\omega_2\omega_3 = \tau_1\,, \tag{29.18}$$
$$I_2\dot{\omega}_2 - (I_3 - I_1)\,\omega_3\omega_1 = \tau_2\,, \tag{29.19}$$
$$I_3\dot{\omega}_3 - (I_1 - I_2)\,\omega_1\omega_2 = \tau_3\,, \tag{29.20}$$

where every component is taken along the body axes [32].
Rests on Theorems 29.7 and 29.16 and Postulate 19.22.

Proof. *Derives Theorem 29.25.*

The rotational law of Chapter 19 holds in an inertial frame: $(\mathrm{d}\boldsymbol{L}/\mathrm{d}t)_{\text{space}} = \boldsymbol{\tau}$ (Equation (19.14)). The inertia tensor is constant only in the body frame, so transport the derivative there with Equation (29.4):

$$\left(\frac{\mathrm{d}\boldsymbol{L}}{\mathrm{d}t}\right)_{\text{body}} + \boldsymbol{\omega}\times\boldsymbol{L} = \boldsymbol{\tau}\,. \tag{29.21}$$

From Equations (19.14) and (29.4) (transporting the inertial rotational law into the body frame, where the inertia tensor is constant).

In principal axes $L_i = I_i\omega_i$ with constant I_i (Theorem 29.16), so the first term has components $I_i\dot{\omega}_i$; and $(\boldsymbol{\omega}\times\boldsymbol{L})_1 = \omega_2 L_3 - \omega_3 L_2 = (I_3 - I_2)\,\omega_2\omega_3$, with the other two components obtained by cyclic permutation. Substituting gives Equations (29.18) to (29.20). □

Proposition 29.26 (The two integrals of torque-free motion)**.** *When $\boldsymbol{\tau} = \mathbf{0}$, both the rotational kinetic energy and the magnitude of the angular momentum are conserved:*

$$2T = I_1\omega_1^2 + I_2\omega_2^2 + I_3\omega_3^2\,, \qquad L^2 = I_1^2\omega_1^2 + I_2^2\omega_2^2 + I_3^2\omega_3^2\,, \tag{29.22}$$

both constant. The trajectory in the space of body-frame angular velocities is therefore confined to the intersection of an ellipsoid with a second ellipsoid.
Rests on Theorem 29.25 and Equation (29.12).

Proof. *Derives Proposition 29.26.*

Multiply Equations (29.18) to (29.20) by $\omega_1, \omega_2, \omega_3$ and add: the right-hand side vanishes and the products $(I_2 - I_3)\,\omega_1\omega_2\omega_3$ cancel in pairs, leaving $\mathrm{d}\left(\sum_i I_i\omega_i^2\right)/\mathrm{d}t = 0$. Multiplying instead by $I_1\omega_1, I_2\omega_2, I_3\omega_3$ and adding cancels the same products and leaves $\mathrm{d}\left(\sum_i I_i^2\omega_i^2\right)/\mathrm{d}t = 0$. The second sum is $|\boldsymbol{L}|^2$ by Theorem 29.16. Both statements also follow directly in the inertial frame, where $\boldsymbol{L}$ is constant and $\mathrm{d}T/\mathrm{d}t = \boldsymbol{\omega}\cdot\boldsymbol{\tau}$. □

Remark 29.27 (Euler's equations as a Lie–Poisson system). Equation (29.21) is not a Hamiltonian system on a cotangent bundle: its variables are the three components of $\boldsymbol{L}$, an odd number, and no symplectic form exists on an odd-dimensional space. It is instead a Lie–Poisson system on the dual of $\mathfrak{so}(3)$, with bracket $\{L_i, L_j\} = -\epsilon_{ijk}L_k$ and Hamiltonian $T = \frac{1}{2}\sum_i L_i^2/I_i$; the second integral L^2 of Proposition 29.26 is then not an accident but the Casimir function of that bracket, constant on every trajectory of every Hamiltonian. The structure is developed in Chapter 24 (see Definition 24.40 and Definition 24.37), and it is the reason the torque-free problem is integrable for every body whatever.
Rests on Proposition 29.26 and Definitions 24.37 and 24.40.

29.3.2 The free symmetric top

Proposition 29.28 (Free rotation of a symmetric body)**.** *Let $I_1 = I_2 = A$ and $I_3 = C$, and let $\boldsymbol{\tau} = \mathbf{0}$. Then ω_3 is constant, and:*

1. *Seen from the body, $\boldsymbol{\omega}$ keeps a constant angle with $\hat{\boldsymbol{e}}_3$ and its equatorial part rotates uniformly at*
$$\Omega_{body} = \frac{C - A}{A}\,\omega_3\,. \tag{29.23}$$
2. *Seen from space, $\boldsymbol{\omega}$, $\boldsymbol{L}$ and $\hat{\boldsymbol{e}}_3$ are coplanar at every instant, and both $\boldsymbol{\omega}$ and $\hat{\boldsymbol{e}}_3$ describe cones about the fixed direction $\boldsymbol{L}$ at the rate*
$$\Omega_{space} = \frac{|\boldsymbol{L}|}{A}\,. \tag{29.24}$$

Rests on Theorem 29.25 and Definition 29.18.

Proof. *Derives Proposition 29.28.*

With $I_1 = I_2$ the third of Equation (29.20) reads $C\dot{\omega}_3 = 0$, so ω_3 is constant. The first two become $A\dot{\omega}_1 = (A - C)\,\omega_2\omega_3$ and $A\dot{\omega}_2 = (C - A)\,\omega_3\omega_1$. Setting $\zeta := \omega_1 + \mathrm{i}\omega_2$ combines them into the single equation $\dot{\zeta} = \mathrm{i}\Omega_{\mathrm{body}}\zeta$ with Ω_{body} as in Equation (29.23), whose solution $\zeta = \zeta_0 \mathrm{e}^{\mathrm{i}\Omega_{\mathrm{body}}t}$ has constant modulus: the equatorial part of $\boldsymbol{\omega}$ has fixed length and turns uniformly, which is statement (i).

For (ii), decompose $\boldsymbol{\omega} = \boldsymbol{\omega}_\perp + \omega_3\hat{\boldsymbol{e}}_3$, so that $\boldsymbol{L} = A\boldsymbol{\omega}_\perp + C\omega_3\hat{\boldsymbol{e}}_3$ by Theorem 29.16. Eliminating $\boldsymbol{\omega}_\perp$,
$$\boldsymbol{\omega} = \frac{\boldsymbol{L}}{A} + \left(1 - \frac{C}{A}\right)\omega_3\hat{\boldsymbol{e}}_3\,, \tag{29.25}$$
which exhibits $\boldsymbol{\omega}$ as a combination of $\boldsymbol{L}$ and $\hat{\boldsymbol{e}}_3$ — hence the coplanarity. Since $\hat{\boldsymbol{e}}_3$ is carried by the body, $\mathrm{d}\hat{\boldsymbol{e}}_3/\mathrm{d}t = \boldsymbol{\omega} \times \hat{\boldsymbol{e}}_3$, and substituting Equation (29.25) kills the second term:
$$\frac{\mathrm{d}\hat{\boldsymbol{e}}_3}{\mathrm{d}t} = \frac{\boldsymbol{L}}{A} \times \hat{\boldsymbol{e}}_3\,,$$
which is precisely a rotation of $\hat{\boldsymbol{e}}_3$ about the fixed vector $\boldsymbol{L}$ at the rate $|\boldsymbol{L}|/A$. The angle between $\hat{\boldsymbol{e}}_3$ and $\boldsymbol{L}$ is constant because L_3 and $|\boldsymbol{L}|$ are, so the motion is a cone. □

Both rates matter and they are different. Equation (29.23) is what an observer riding on the body would measure — for the Earth it is the free wobble of Section 29.3.4 — while Equation (29.24) is what an outside observer sees, and for a thrown object it is the visible tumbling. The two cones roll on one another: the *body cone*, traced by $\boldsymbol{\omega}$ about $\hat{\boldsymbol{e}}_3$, rolls without slipping on the *space cone*, traced by $\boldsymbol{\omega}$ about $\boldsymbol{L}$, their common line being the instantaneous axis, whose points are momentarily at rest.

Theorem 29.29 (Poinsot's construction)**.** *In torque-free motion the inertia ellipsoid of the body, carried with it, rolls without slipping on a plane that is fixed in space — the* invariable plane*, perpendicular to $\boldsymbol{L}$ and at the distance $\sqrt{2T}/|\boldsymbol{L}|$ from the fixed point [99].*

Rests on Definition 29.19 and Proposition 29.26.

Proof. *Derives Theorem 29.29.*

Put $\boldsymbol{\rho} := \boldsymbol{\omega}/\sqrt{2T}$. Then $\boldsymbol{\rho} \cdot I\boldsymbol{\rho} = (\boldsymbol{\omega} \cdot I\boldsymbol{\omega})/(2T) = 1$ by Equation (29.12), so $\boldsymbol{\rho}$ lies on the inertia ellipsoid Equation (29.14) at every instant.

The outward normal to the ellipsoid at $\boldsymbol{\rho}$ is the gradient of the quadratic form, $\frac{1}{2}\nabla(\boldsymbol{x} \cdot I\boldsymbol{x}) = I\boldsymbol{\rho} = \boldsymbol{L}/\sqrt{2T}$, which is parallel to $\boldsymbol{L}$ — a direction fixed in space, because the motion is torque-free. The tangent plane at the contact point is therefore perpendicular to $\boldsymbol{L}$ for all time. Its distance from the centre is
$$\boldsymbol{\rho} \cdot \frac{\boldsymbol{L}}{|\boldsymbol{L}|} = \frac{\boldsymbol{\omega} \cdot \boldsymbol{L}}{\sqrt{2T}\,|\boldsymbol{L}|} = \frac{2T}{\sqrt{2T}\,|\boldsymbol{L}|} = \frac{\sqrt{2T}}{|\boldsymbol{L}|}\,,$$
constant by Proposition 29.26. A plane of fixed orientation and fixed distance from a fixed point is a fixed plane.

Finally, the contact point $\boldsymbol{\rho}$ is parallel to $\boldsymbol{\omega}$, so it lies on the instantaneous axis of rotation; by Equation (29.3) its material velocity vanishes. A surface touching a plane at a point of zero velocity rolls without slipping. □

Remark 29.30 (Polhode and herpolhode). The curve traced by the contact point on the ellipsoid is the *polhode*, and the curve it traces on the invariable plane the *herpolhode*. The polhode is closed — it is the intersection of the two ellipsoids of Equation (29.22), rescaled — and it is what makes the whole torque-free motion visible at a glance: for a body whose I_i are all distinct the polhodes encircle the axes of largest and of smallest moment and are separated by two curves crossing at the intermediate axis, which is the geometric form of Phenomenon 29.31 below. The herpolhode need not close, and does not in general, since the two frequencies of the motion are incommensurable.

Rests on Theorem 29.29 and Proposition 29.26.

29.3.3 Stability of rotation about the principal axes

Phenomenon 29.31 (The intermediate-axis instability)**.** A body with three distinct principal moments of inertia, thrown spinning about the axis of largest or of smallest moment, holds that axis steadily; thrown about the intermediate axis it tumbles, flipping end over end at a regular rate however carefully it was launched. The demonstration wants nothing more than a tennis racket, a book held shut, or a mobile telephone, and it works every time — the flipping is not a symptom of an imperfect throw. The same instability decides the fate of tumbling satellites and asteroids: internal dissipation drives a freely rotating body towards spin about the axis of *largest* moment, since that is the state of least kinetic energy at fixed angular momentum [1585, 2027].

Rests on Theorems 29.16 and 29.25.

Derivation. *Derives Phenomenon 29.31.*

Take Euler's torque-free equations in the principal frame (Equation (29.18)),

$$I_1\dot{\omega}_1 = (I_2 - I_3)\,\omega_2\omega_3\,, \qquad \text{and cyclically}\,,$$

and perturb the steady rotation $\boldsymbol{\omega} = (\Omega, 0, 0)$ by writing $\omega_1 = \Omega + \varepsilon_1$, $\omega_2 = \varepsilon_2$ and $\omega_3 = \varepsilon_3$ with the ε_i small. To first order the first equation gives $\dot{\varepsilon}_1 = 0$, and the other two give

$$I_2\dot{\varepsilon}_2 = (I_3 - I_1)\,\Omega\,\varepsilon_3\,, \qquad I_3\dot{\varepsilon}_3 = (I_1 - I_2)\,\Omega\,\varepsilon_2\,.$$

Differentiating the first and substituting the second,

$$\ddot{\varepsilon}_2 = \frac{(I_3 - I_1)\,(I_1 - I_2)}{I_2 I_3}\,\Omega^2\varepsilon_2\,. \tag{29.26}$$

The coefficient on the right is negative — so that ε_2 merely oscillates, and the axis is stable — when I_1 is larger than both of I_2, I_3 or smaller than both; and it is positive — so that ε_2 grows exponentially, and the axis is unstable — exactly when I_1 lies strictly between them. The unstable axis is therefore the intermediate one, and only that one. A body with two equal moments has no intermediate axis and shows no such instability, which is why a symmetric top misbehaves in other ways but never in this one.

The growth rate follows from the same coefficient: writing $\varepsilon_2 \propto \mathrm{e}^{\lambda t}$ about the intermediate axis $I_2 < I_1 < I_3$,

$$\lambda = \Omega\sqrt{\frac{(I_3 - I_1)\,(I_1 - I_2)}{I_2 I_3}}\,, \tag{29.27}$$

which is of the order of the spin rate itself unless two of the moments nearly coincide. That is why the flip is fast, and why it recurs at a regular interval rather than settling: the linearization governs only the departure, and the exact motion returns periodically to the neighbourhood of the intermediate axis, as Remark 29.30 shows geometrically. □

Remark 29.32 (What the linearization is entitled to conclude). The argument just given solves the *linearized* equations and reads off the stability of the full nonlinear system. That step is legitimate here because the linearization is hyperbolic in the unstable case — a real positive exponent Equation (29.27), no zero eigenvalue — but the general theorem licensing it is a statement about systems of ordinary differential equations, not about rigid bodies. Part II's chapter on ordinary differential equations treats the scalar equation and its Wronskian theory but carries neither the constant-coefficient linear *system* $\dot{\boldsymbol{x}} = \mathcal{A}\boldsymbol{x}$ nor Lyapunov's first method; those are owed. In the stable case the conclusion is weaker still and is stated as such: the linearization shows bounded oscillation, and it is Theorem 29.29 and Remark 29.30, which are exact, that show the exact motion stays near the axis.

Rests on Phenomenon 29.31 and Theorem 29.29.

Proposition 29.33 (Dissipation drives a free body to its largest moment)**.** *Among all rotations of a given body with a given angular momentum* $\boldsymbol{L}$*, the rotational kinetic energy*

$$T = \frac{1}{2}\left(\frac{L_1^2}{I_1} + \frac{L_2^2}{I_2} + \frac{L_3^2}{I_3}\right) \tag{29.28}$$

is least when $\boldsymbol{L}$ *lies along the axis of largest moment. A freely rotating body that dissipates energy internally, while conserving* $\boldsymbol{L}$ *because the dissipation is internal, therefore ends in rotation about that axis.*

Rests on Proposition 29.26 and Theorem 29.16.

Proof. *Derives Proposition 29.33.*

Equation (29.28) is Equation (29.12) written in terms of $L_i = I_i\omega_i$. At fixed $L^2 = \sum_i L_i^2$ the right-hand side is a weighted average of the $1/I_i$ with weights L_i^2/L^2 summing to one, so it lies between $L^2/(2I_{\max})$ and $L^2/(2I_{\min})$ and attains the lower bound exactly when all the weight sits on the largest moment. Internal forces exert no external torque (Postulate 19.48), so $\boldsymbol{L}$ is conserved while T decreases; a decreasing quantity bounded below by $L^2/(2I_{\max})$ converges, and the only stationary configurations compatible with the bound are rotations about the axis of largest moment. □

An elongated body set spinning about its long axis — the axis of *smallest* moment — is therefore stable against small perturbations by Phenomenon 29.31 but not against dissipation by Proposition 29.33: it is at a maximum of T at fixed L, and any internal flexing bleeds energy away until it tumbles and then settles into an end-over-end rotation about a transverse axis. This is the fate of a spin-stabilized satellite whose structure is not perfectly rigid, and it is also why small asteroids observed in non-principal-axis rotation are taken to have been disturbed recently. No primary reference is cited here for the satellite history; the mechanical statement is Proposition 29.33 and rests on nothing else.

29.3.4 Free nutation of the Earth: the Chandler wobble

Phenomenon 29.34 (Free nutation of the Earth)**.** The Earth's instantaneous axis of rotation is not fixed within the Earth. The pole traces an irregular, roughly circular path a few metres across on the surface, and that motion contains a free component whose period is near 433 days — some two fifths longer than the 305 days that a *rigid* Earth of the observed dynamical ellipticity would show [2027]. The discrepancy is not an error of measurement. It is the signature of the planet yielding elastically to the centrifugal bulge that its own wobble creates, and the size of the lengthening is therefore a measurement of the rigidity of the Earth's interior, made from the outside.

Rests on Proposition 29.28 and Example 29.23.

Derivation. *Derives Phenomenon 29.34.*

For a rigid symmetric body with equatorial moment A and polar moment C, Euler's torque-free equations (Equation (29.18)) with ω_3 constant read

$$A\dot{\omega}_1 = (A - C)\,\omega_2\omega_3\,, \qquad A\dot{\omega}_2 = (C - A)\,\omega_3\omega_1\,.$$

Setting $\zeta := \omega_1 + \mathrm{i}\omega_2$ combines the two into $\dot{\zeta} = \mathrm{i}\Omega_{\text{free}}\zeta$, so the equatorial part of $\boldsymbol{\omega}$ rotates uniformly within the body at the rate

$$\Omega_{\text{free}} = \frac{C - A}{A}\,\omega_3\,, \tag{29.29}$$

that is, one turn in $A/(C - A)$ rotation periods of the body — which is Equation (29.23) specialized to the Earth. The observed dynamical ellipticity is $H := (C - A)/C = 3.274 \times 10^{-3}$ (Table 29.2), so

$$\frac{A}{C - A} = \frac{1 - H}{H} = 304.5\,,$$

and the rigid prediction is a free wobble of about 305 sidereal days — Euler's number [32], and one that no observation matches. Chandler's analysis of the latitude records gave a free period near 427 days [250], and modern determinations place it near 433 days; the amplitude corresponds to a displacement of the pole of a few metres at the surface.

The elastic correction follows from the same equations once the body is allowed to yield. A wobble tilts the rotation axis to $\boldsymbol{\omega} = \Omega\,(m_1, m_2, 1)$, with $|m|$ of order 5×10^{-7} for a pole displacement of a few metres, and the centrifugal potential energy per unit mass, $\Phi = -\frac{1}{2}\,|\boldsymbol{\omega} \times \boldsymbol{r}|^2$, acquires the perturbation

$$\delta\Phi = \Omega^2 z\,(m_1 x + m_2 y)\,, \tag{29.30}$$

in J/kg: a second-degree harmonic of the same shape as a tide, raised by the wobble itself. The Earth responds to it elastically, and Love's number k_2 is defined as the ratio of the extra gravitational potential the deformation produces at the surface to the potential that provoked it [376]. By Equation (29.17) the orientation-dependent part of the external potential of a body is carried entirely by its inertia tensor, and the part of it belonging to the products of inertia c_{13}, c_{23} is $3G\,(c_{13}xz + c_{23}yz)/r^5$. Equating that at $r = a_\oplus$ to $k_2\,\delta\Phi$ gives

$$c_{13} + \mathrm{i}c_{23} = \kappa\,(m_1 + \mathrm{i}m_2)\,, \qquad \kappa := \frac{k_2\,a_\oplus^5\,\Omega^2}{3G}\,, \tag{29.31}$$

with κ carrying $\mathrm{kg\,m^2}$, as an increment of inertia must.

Now redo the free-wobble calculation with $I = I_0 + c$. Angular momentum is conserved in space, so by Equation (29.4) its body components obey $\dot{\boldsymbol{L}} + \boldsymbol{\omega} \times \boldsymbol{L} = \boldsymbol{0}$, and to first order in m and in c one has $L_1 = \Omega\,(Am_1 + c_{13})$, $L_2 = \Omega\,(Am_2 + c_{23})$ and $L_3 = C\Omega$. Writing $m := m_1 + \mathrm{i}m_2$ and combining the first two components as before,

$$(A + \kappa)\,\dot{m} = \mathrm{i}\Omega\,(C - A - \kappa)\,m\,, \tag{29.32}$$

so the free wobble is still circular, but at the rate

$$\Omega_{\text{Ch}} = \Omega\,\frac{C - A - \kappa}{A + \kappa} \simeq \Omega\,\frac{C - A}{A}\left(1 - \frac{k_2}{k_{\text{s}}}\right), \qquad k_{\text{s}} := \frac{3G\,(C - A)}{a_\oplus^5\,\Omega^2}\,, \tag{29.33}$$

From Equations (29.31) and (29.32) (the elastic increment enters only through the ratio of Love numbers).

the approximation using $\kappa/A \sim 1 \times 10^{-3}$. The dimensionless k_{s} is the Love number a body would have if it yielded until its whole bulge were centrifugal; from Table 29.2, with $C - A = HC = 2.629 \times 10^{35}\,\mathrm{kg\,m^2}$ and $a_\oplus^5 = 1.0556 \times 10^{34}\,\mathrm{m^5}$, it is $k_{\text{s}} = 0.938$.

Equation (29.33) therefore lengthens the period by the factor $(1 - k_2/k_s)^{-1}$, and the observed lengthening from 304.5 to 433 days inverts to

$$k_2 = k_s \left(1 - \frac{304.5}{433}\right) = 0.28\,, \tag{29.34}$$

which is the measurement the phenomenon promised: one number describing how far the whole planet yields, read off a period. It is of the size of, and a little below, the degree-two Love number near 0.30 that Earth-tide observations give—no source is cited here for that comparison value—and the difference is where the fluid outer core, which does not follow the mantle's wobble, and the ocean pole tide enter. Neither is in this calculation. Neither is the damping: the observed wobble decays in a few decades and must therefore be re-excited continually, by some redistribution of mass among the atmosphere, the oceans and the solid Earth. Which of those dominates is a question this treatise does not settle. □

The mechanism was identified by Newcomb within a year of Chandler's announcement [254]: a wobbling Earth deforms under the centrifugal field of its own wobble, and the deformation carries part of the equatorial bulge along with the instantaneous pole, so that the effective $C - A$ opposing the motion is smaller than the static one and the period is correspondingly longer. Equation (29.33) is that argument made quantitative, and it deliberately hides the whole interior in a single number: k_2 is what one gets by refusing to model the planet and asking only how much its external field yields. Computing k_2 from a density and a rigidity profile, rather than reading it off a period, is elastostatics and belongs to Chapter 30. What the observation delivers without any of that is a bulk measurement of the rigidity of the mantle that no laboratory sample can provide.

29.4 The heavy top

29.4.1 The Lagrange top

Definition 29.35 (The Lagrange top)**.** A *Lagrange top* is a symmetric rigid body, $I_1 = I_2 = A$ and $I_3 = C$ about a *fixed point* on its symmetry axis, with its centre of mass a distance ℓ from that point along the same axis, moving under uniform gravity $\boldsymbol{g}$. In the Euler angles of Notation 29.9, with the space z axis vertical, its Lagrangian is

$$\mathcal{L} = \frac{A}{2}\left(\dot\theta^2 + \dot\varphi^2 \sin^2\theta\right) + \frac{C}{2}\left(\dot\psi + \dot\varphi\cos\theta\right)^2 - Mg\ell\cos\theta\,. \tag{29.35}$$

Rests on Proposition 29.11, Definition 21.31, and Theorem 29.16.

The kinetic term is Equation (29.12) in principal axes with $\omega_1^2 + \omega_2^2 = \dot\theta^2 + \dot\varphi^2\sin^2\theta$ and $\omega_3 = \dot\psi + \dot\varphi\cos\theta$, both read off Equations (29.7) to (29.9); the potential is Mgz_{cm} with $z_{\mathrm{cm}} = \ell\cos\theta$. The moments A and C are taken about the fixed point, so A already contains the $M\ell^2$ of Theorem 29.20.

Proposition 29.36 (Three first integrals)**.** *The Lagrange top conserves*

$$p_\psi = C\left(\dot\psi + \dot\varphi\cos\theta\right) = C\omega_3 =: L_3\,, \tag{29.36}$$

$$p_\varphi = A\dot\varphi\sin^2\theta + L_3\cos\theta =: L_z\,, \tag{29.37}$$

$$E = \frac{A}{2}\left(\dot\theta^2 + \dot\varphi^2\sin^2\theta\right) + \frac{L_3^2}{2C} + Mg\ell\cos\theta\,. \tag{29.38}$$

The first two are the components of the angular momentum along the body symmetry axis and along the vertical.

Rests on Equation (29.35) and Theorems 21.36 and 21.43.

Proof. *Derives Proposition 29.36.*

Neither ψ nor φ appears in Equation (29.35): both are cyclic (Definition 21.35), so their conjugate momenta are conserved by Theorem 21.36. Computing them from Equation (29.35) gives Equations (29.36) and (29.37). The Lagrangian has no explicit time dependence, so the energy function of Theorem 21.43 is conserved, and since the kinetic term is quadratic in the velocities it equals $T + V$, which is Equation (29.38). That p_ψ and p_φ are the stated components of $\boldsymbol{L}$ follows because ψ generates rotation about $\hat{\boldsymbol{e}}_3$ and φ rotation about $\hat{\boldsymbol{z}}$. □

Theorem 29.37 (Reduction to one degree of freedom)**.** *Write $u := \cos\theta$ and set*

$$a := \frac{L_3}{A}\,, \qquad b := \frac{L_z}{A}\,, \qquad \alpha := \frac{2}{A}\left(E - \frac{L_3^2}{2C}\right), \qquad \beta := \frac{2Mg\ell}{A}\,.$$

Then the motion of the axis obeys

$$\dot{u}^2 = \left(1 - u^2\right)(\alpha - \beta u) - (b - au)^2 =: f(u)\,, \tag{29.39}$$

a cubic in u with $f(\pm 1) = -(b \mp a)^2 \le 0$ and $f \to +\infty$ as $u \to +\infty$. Except in degenerate cases f has exactly two roots $u_1 \le u_2$ in $[-1, 1]$, and the axis nutates periodically between the corresponding polar angles $\theta_2 \le \theta \le \theta_1$, while

$$\dot{\varphi} = \frac{b - au}{1 - u^2}\,, \qquad \dot{\psi} = \frac{L_3}{C} - \frac{u\,(b - au)}{1 - u^2}\,. \tag{29.40}$$

The solution is reduced to the quadrature $t = \int \mathrm{d}u/\sqrt{f(u)}$.

Rests on Proposition 29.36 and Theorem 7.23.

Proof. *Derives Theorem 29.37.*

Solve Equation (29.37) for $\dot{\varphi}$, obtaining the first of Equation (29.40), and Equation (29.36) for $\dot{\psi}$, obtaining the second. Substituting the first into Equation (29.38) and using $\dot{u} = -\sin\theta\,\dot{\theta}$, so that $\dot{u}^2 = (1 - u^2)\,\dot{\theta}^2$, gives

$$\frac{\dot{u}^2}{1 - u^2} + \frac{(b - au)^2}{1 - u^2} = \alpha - \beta u\,,$$

which rearranges to Equation (29.39). Evaluating f at $u = \pm 1$ kills the first product and leaves $-(b \mp a)^2$, which is negative unless the axis reaches the vertical. Since f is a cubic with positive leading coefficient $\beta > 0$, it tends to $+\infty$ at large u; and physical motion requires $f \ge 0$, which the initial condition supplies at some $u_0 \in (-1, 1)$. The intermediate value theorem (Theorem 7.23) then places a root in $(u_0, 1)$ and another in $(-1, u_0)$, and a cubic has no more than three roots, the third lying above 1 and being unphysical. Between the two roots $\dot{u}^2 > 0$, and the motion runs from one to the other and back, periodically, because Equation (29.39) is autonomous and $\dot{u}$ changes sign only at the roots. □

Phenomenon 29.38 (Nutation of a released top)**.** A rapidly spinning top, held with its axis at an angle to the vertical and then released from rest, does not begin to precess smoothly. Its axis first dips, then rises again to the height it started from, and meanwhile it swings around the vertical; the tip of the axis traces a succession of cusps hanging below the circle of release. The dip and the swing are fast and small compared with the mean progress around the vertical, and the faster the top spins the smaller the dip and the quicker the flutter, until the eye no longer resolves it and the motion looks like the smooth precession of Phenomenon 29.46 [2027].

Rests on Theorem 29.37 and Proposition 29.36.

Derivation. *Derives Phenomenon 29.38.*

Release from rest at θ_0 means $\dot\theta = \dot\varphi = 0$ there, so Equation (29.37) gives $b = au_0$ with $u_0 = \cos\theta_0$, and $f(u_0) = 0$ in Equation (29.39) forces $\alpha = \beta u_0$. Writing $u = u_0 + x$ with x small, $b - au = -ax$ and $\alpha - \beta u = -\beta x$, so

$$\dot x^2 = -\beta\left(1 - u_0^2\right)x - a^2x^2 + O\left(x^3\right). \tag{29.41}$$

Differentiating once, $\ddot x = -a^2x - \frac{1}{2}\beta\sin^2\theta_0$: the variable x performs simple harmonic motion at the angular frequency

$$\omega_{\text{nut}} = a = \frac{C\omega_3}{A}, \tag{29.42}$$

about the displaced centre $x = -\beta\sin^2\theta_0/(2a^2)$, with amplitude equal to that displacement. Since x starts at zero and must be negative for Equation (29.41) to be positive, u decreases and θ increases: the top falls first. Converting the full swing in u to an angle through $\Delta u = \sin\theta_0\Delta\theta$,

$$\Delta\theta = \frac{\beta\sin\theta_0}{a^2} = \frac{2AMg\ell\sin\theta_0}{C^2\omega_3^2}, \tag{29.43}$$

which falls off as the inverse square of the spin.

The precession follows from the first of Equation (29.40): $\dot\varphi = -ax/(1-u^2)$, which vanishes exactly where $x = 0$ — at the top of each nutation swing. There the axis is momentarily at rest in both angles, which is the cusp. Averaging over one nutation period, $\langle -x\rangle$ is the centre offset $\beta\sin^2\theta_0/(2a^2)$, so

$$\langle\dot\varphi\rangle = \frac{a}{\sin^2\theta_0}\cdot\frac{\beta\sin^2\theta_0}{2a^2} = \frac{\beta}{2a} = \frac{Mg\ell}{C\omega_3}, \tag{29.44}$$

which is exactly the fast-top rate Equation (29.55) derived independently below. The two calculations agreeing is the check that the smooth precession of Phenomenon 29.46 is the nutation-averaged motion and not a different phenomenon. □

Proposition 29.39 (Steady precession)**.** *A Lagrange top can precess at constant θ and constant $\dot\varphi = \Omega$ provided*

$$A\Omega^2\cos\theta - C\omega_3\Omega + Mg\ell = 0, \qquad \textit{that is} \qquad \Omega_\pm = \frac{C\omega_3 \pm \sqrt{C^2\omega_3^2 - 4AMg\ell\cos\theta}}{2A\cos\theta}. \tag{29.45}$$

For $\theta < \pi/2$ this requires $C^2\omega_3^2 \geq 4AMg\ell\cos\theta$: below that spin no steady precession at the given inclination exists at all. The slow root reduces to $Mg\ell/(C\omega_3)$ when the spin is large.

Rests on Equation (29.35) and Theorem 16.22.

Proof. *Derives Proposition 29.39.*

The Euler–Lagrange equation (Theorem 16.22) for θ applied to Equation (29.35) is

$$A\ddot\theta = A\dot\varphi^2\sin\theta\cos\theta - C\omega_3\dot\varphi\sin\theta + Mg\ell\sin\theta,$$

using $\omega_3 = \dot\psi + \dot\varphi\cos\theta$ from Equation (29.9). Setting $\ddot\theta = 0$, $\dot\varphi = \Omega$ and $\sin\theta \neq 0$ gives the quadratic in Equation (29.45), whose roots are as stated. Reality of the roots is the discriminant condition. For $C\omega_3 \gg \sqrt{AMg\ell}$, expanding the square root to first order gives $\Omega_- \to Mg\ell/(C\omega_3)$ and $\Omega_+ \to C\omega_3/(A\cos\theta)$. □

Proposition 29.40 (The sleeping top)**.** *A top spinning about the exact vertical, $\theta = 0$, is stable against small disturbances if and only if*

$$C^2\omega_3^2 > 4AMg\ell. \tag{29.46}$$

Above that spin the top "sleeps", standing apparently motionless; below it, the vertical position is unstable and the top falls away. Friction slowly reduces ω_3, so a sleeping top wakes when Equation (29.46) fails, and does so abruptly.

Rests on Theorem 29.37 and Proposition 29.36.

Proof. *Derives Proposition 29.40.*

At $\theta = 0$ the two conserved momenta coincide, $L_z = L_3$, since Equation (29.37) at $\sin\theta = 0$ gives $L_z = L_3 \cos 0$. Eliminating $\dot\varphi$ from Equation (29.38) with Equation (29.40) puts the energy in the one-dimensional form $E' = \tfrac{1}{2}A\dot\theta^2 + V_{\text{eff}}(\theta)$ (Definition 27.25) with

$$V_{\text{eff}}(\theta) = \frac{(L_z - L_3\cos\theta)^2}{2A\sin^2\theta} + Mg\ell\cos\theta\,. \tag{29.47}$$

With $L_z = L_3$ the first term is $(L_3^2/2A)\,(1-\cos\theta)\,/\,(1+\cos\theta)$, which for small θ is $L_3^2\theta^2/(8A)$; and $Mg\ell\cos\theta = Mg\ell - Mg\ell\theta^2/2 + O(\theta^4)$. Hence

$$V_{\text{eff}}(\theta) = Mg\ell + \left(\frac{L_3^2}{8A} - \frac{Mg\ell}{2}\right)\theta^2 + O\!\left(\theta^4\right)\,,$$

so $\theta = 0$ is a minimum — and the motion in θ is bounded near it — exactly when $L_3^2 > 4AMg\ell$, which with $L_3 = C\omega_3$ is Equation (29.46). When the coefficient is negative the effective potential falls away from $\theta = 0$ and no bounded motion near the vertical exists. □

Example 29.41 (A toy gyroscope, in numbers)**.** Take a uniform disc of mass $M = 0.100\,\text{kg}$ and radius $R = 0.0300\,\text{m}$ mounted on a light spindle and pivoted at $\ell = 0.0500\,\text{m}$ from its centre, spinning at 50.0/s, that is $\omega_3 = 314\,\text{rad/s}$. From Table 29.1 and Theorem 29.20,

$$C = \tfrac{1}{2}MR^2 = 4.50\times10^{-5}\,\text{kg m}^2\,, \qquad A = \tfrac{1}{4}MR^2 + M\ell^2 = 2.725\times10^{-4}\,\text{kg m}^2\,,$$

and $Mg\ell = 4.905\times10^{-2}\,\text{N m}$ with $g = 9.81\,\text{m/s}^2$. Then:

- the sleeping criterion Equation (29.46) is met with $C^2\omega_3^2/(4AMg\ell) = 3.74$, so this top would stand upright;
- the fast-top precession rate Equation (29.55) is $Mg\ell/(C\omega_3) = 3.47\,\text{rad/s}$, one turn in 1.81 s;
- at $\theta = 60°$ the exact slow root of Equation (29.45) is 3.59 rad/s, so the fast-top formula is low by 3.5 %, and the fast root is 100 rad/s;
- the nutation frequency Equation (29.42) is $C\omega_3/A = 51.9\,\text{rad/s}$, a period of 0.121 s, and the nutation amplitude Equation (29.43) on release at 60° is 0.116 rad, about 6.6°.

The separation of scales that makes Equation (29.55) useful is visible in the last two entries: the flutter is fifteen times faster than the precession and a few degrees wide, which is exactly what one sees.

Rests on Equations (29.43), (29.45) and (29.55).

Remark 29.42 (The closed form, and the mathematics it needs). The quadrature $t = \int \mathrm{d}u/\sqrt{f(u)}$ of Theorem 29.37 can be carried further. Because f is a cubic, the affine substitution that carries its three roots to 0, 1 and $1/k^2$ puts the integral in Legendre's normal form, so that $u = \cos\theta$ is a Jacobi elliptic function of the time; the nutation period $2\int_{u_1}^{u_2}\mathrm{d}u/\sqrt{f(u)}$ is then a complete elliptic integral of the first kind built on the same roots, and the two degenerate cases in which a pair of roots coalesce return steady precession Equation (29.45) and the sleeping top Equation (29.46).

That is a statement about the integral, and it is recorded here and used nowhere. Elliptic integrals and the Jacobi elliptic functions are not developed anywhere in this treatise—the same absence stops Remark 28.66 at the exact period of the finite-amplitude pendulum—so this chapter asserts nothing whose establishment would need the inversion. The nutation frequency Equation (29.42), the nutation amplitude Equation (29.43), the mean precession Equation (29.44), the steady-precession roots Equation (29.45) and the sleeping-top condition Equation (29.46) were all obtained from Equation (29.39) directly, without ever solving it.

Rests on Theorem 29.37 and Remark 28.66.

29.4.2 Integrable cases and the Kovalevskaya top

The heavy top with a fixed point has three degrees of freedom and always possesses three integrals of motion: the energy, the vertical component L_z of the angular momentum, and the trivial $|\boldsymbol{\gamma}| = 1$, where $\boldsymbol{\gamma}$ denotes the upward vertical resolved along the body axes. Complete integrability requires one more. Exactly three families of bodies supply it for arbitrary initial conditions.

Proposition 29.43 (The three integrable heavy tops)**.** *The heavy top with a fixed point admits a fourth independent integral, for all initial data, in the following three cases and — as Kovalevskaya's analysis of the singularity structure of the solutions established — in no others:*

1. *the* Euler case *[32]: the fixed point is the centre of mass, $\ell = 0$, so the motion is torque-free and the fourth integral is $|\boldsymbol{L}|^2$ of Proposition 29.26;*
2. *the* Lagrange case *[40]: $I_1 = I_2$ with the centre of mass on the symmetry axis, and the fourth integral is $L_3 = C\omega_3$ of Equation (29.36);*
3. *the* Kovalevskaya case *[243]: $I_1 = I_2 = 2I_3$ with the centre of mass in the plane of the two equal moments, and the fourth integral is quartic in the angular velocities.*

Rests on Propositions 29.26 and 29.36.

Proof. *Derives Proposition 29.43.*

Cases (i) and (ii) are read off the results already proved: Proposition 29.26 supplies $|\boldsymbol{L}|^2$ when the torque vanishes, and Proposition 29.36 supplies L_3 when the body is symmetric about the axis carrying the centre of mass. The existence of the fourth integral in case (iii), and the completeness of the list, are Kovalevskaya's and are not reproved here; the fourth integral itself is quoted in Remark 29.44. □

Remark 29.44 (The Kovalevskaya integral). With $I_1 = I_2 = 2I_3$, the centre of mass at distance ℓ from the fixed point along the body 1 axis, and $\boldsymbol{\gamma}$ the upward vertical resolved along the body axes, the conserved quantity Kovalevskaya found is

$$K = \left| (\omega_1 + \mathrm{i}\omega_2)^2 + c_0\,(\gamma_1 + \mathrm{i}\gamma_2) \right|^2 , \qquad c_0 = \frac{Mg\ell}{I_3} , \tag{29.48}$$

whose two factors have the same SI dimension, $1/\mathrm{s}^2$, as they must. The sign of c_0 follows the orientation chosen for the body axes and must be fixed before the expression is used.

Rests on Proposition 29.43.

Proof. *Derives Equation (29.48).*

Write $I_3 =: I$, so that $I_1 = I_2 = 2I$. Keep $\boldsymbol{\gamma}$ the upward vertical, and orient the body 1 axis from the centre of mass towards the fixed point, so that the centre of mass sits at $-\ell\hat{\boldsymbol{e}}_1$ and the gravitational torque about the fixed point is $\boldsymbol{\tau} = (-\ell\hat{\boldsymbol{e}}_1) \times (-Mg\boldsymbol{\gamma}) = Mg\ell\,\hat{\boldsymbol{e}}_1 \times \boldsymbol{\gamma} = Mg\ell\,(0, -\gamma_3, \gamma_2)$. That is the convention of the remark; reversing $\hat{\boldsymbol{e}}_1$ reverses

the torque and with it the sign of c_0, which is the freedom recorded there. Euler's equations Equations (29.18) to (29.20) then read

$$\dot{\omega}_1 = \tfrac{1}{2}\omega_2\omega_3\,, \qquad \dot{\omega}_2 = -\tfrac{1}{2}\omega_3\omega_1 - \tfrac{1}{2}c_0\gamma_3\,, \qquad \dot{\omega}_3 = c_0\gamma_2\,, \tag{29.49}$$

with $c_0 = Mg\ell/I$. The vertical is fixed in space, so Equation (29.4) applied to $\boldsymbol{\gamma}$ gives its body components as

$$\dot{\boldsymbol{\gamma}} = \boldsymbol{\gamma} \times \boldsymbol{\omega}\,. \tag{29.50}$$

Set $\zeta := \omega_1 + \mathrm{i}\omega_2$ and $\eta := \gamma_1 + \mathrm{i}\gamma_2$. Combining the first two components of Equation (29.49), and then of Equation (29.50),

$$\dot{\zeta} = -\tfrac{\mathrm{i}}{2}\omega_3\zeta - \tfrac{\mathrm{i}}{2}c_0\gamma_3\,, \tag{29.51}$$

$$\dot{\eta} = -\mathrm{i}\omega_3\eta + \mathrm{i}\gamma_3\zeta\,. \tag{29.52}$$

Now differentiate $\xi := \zeta^2 + c_0\eta$:

$$\dot{\xi} = 2\zeta\dot{\zeta} + c_0\dot{\eta} = -\mathrm{i}\omega_3\zeta^2 - \mathrm{i}c_0\gamma_3\zeta - \mathrm{i}c_0\omega_3\eta + \mathrm{i}c_0\gamma_3\zeta = -\mathrm{i}\omega_3\,\xi\,. \tag{29.53}$$

From Equations (29.51) and (29.52) (the two terms carrying γ_3 cancel identically).

The coefficient is purely imaginary, so $\mathrm{d}\,|\xi|^2/\mathrm{d}t = \bar{\xi}\dot{\xi} + \xi\bar{\dot{\xi}} = -\mathrm{i}\omega_3\,|\xi|^2 + \mathrm{i}\omega_3\,|\xi|^2 = 0$, and $K = |\xi|^2$ of Equation (29.48) is conserved.

The ratio of the moments is exactly what makes the cancellation happen. Repeat the calculation with $I_1 = I_2 = A$ and $I_3 = C$ arbitrary, the centre of mass still on the body 1 axis. Euler's equations now give $\dot{\zeta} = \mathrm{i}k\,\omega_3\zeta - \mathrm{i}\,(Mg\ell/A)\,\gamma_3$ with $k := (C-A)/A$, while Equation (29.50) is unchanged. For a trial quantity $\xi = \zeta^n + \lambda\eta$,

$$\dot{\xi} = \mathrm{i}nk\,\omega_3\zeta^n - \mathrm{i}n\frac{Mg\ell}{A}\gamma_3\zeta^{n-1} - \mathrm{i}\lambda\omega_3\eta + \mathrm{i}\lambda\gamma_3\zeta\,. \tag{29.54}$$

The two terms in γ_3 cancel for no choice of λ unless $n = 2$, and then only for $\lambda = 2Mg\ell/A$. What survives, $\mathrm{i}\omega_3\,(2k\zeta^2 - \lambda\eta)$, is proportional to $\xi = \zeta^2 + \lambda\eta$ only if $2k = -1$, that is only if $A = 2C$; and then $\lambda = 2Mg\ell/(2C) = Mg\ell/I_3 = c_0$. So the quantity displayed in Equation (29.48) is conserved precisely when $I_1 = I_2 = 2I_3$, which is what singles the Kovalevskaya case out. That no *other* fourth integral exists for other ratios is Kovalevskaya's own theorem [243] and is not proved here. □

Remark 29.45 (What integrability buys, and where it stops). For each of the three cases the four integrals are in involution and the machinery of Chapter 23 applies: the motion is confined to a torus and can be written in action–angle variables (Definition 23.28), so the long-time behaviour is quasi-periodic and predictable indefinitely. A heavy top that is not of one of these three kinds has no such description. Its trajectories are not confined to tori, they are sensitive to initial conditions in the sense of Chapter 32, and no closed-form solution exists — which is the general situation, the three integrable cases being exceptional.

Rests on Proposition 29.43 and Definition 23.28.

29.4.3 Gyroscopes

Phenomenon 29.46 (A fast spinning top does not fall)**.** Set a top spinning rapidly and tilt its axis away from the vertical: it does not topple. Its axis instead sweeps out a cone about the vertical at a slow and steady rate, moving at right angles to the direction in which gravity is pulling it. A spinning bicycle wheel held by one end of its axle behaves the same way, and pushes back sideways against the hand that tries to tilt it. The rate of the sweep is inversely proportional to the spin — the faster the top turns, the slower it precesses and the more nearly it appears simply to stand — and, to the accuracy of the observation, it does not depend on how far the axis is tilted [2027].

Rests on Theorem 29.25 and Equation (19.14).

Derivation. *Derives Phenomenon 29.46.*

Let the top be symmetric, spinning at ω_3 about its own axis $\hat{\boldsymbol{e}}_3$, and supported at a point a distance ℓ from its centre of mass along that axis. In the fast-top limit the angular momentum is dominated by the spin, $\boldsymbol{L} \simeq I_3\omega_3\hat{\boldsymbol{e}}_3$. About the support the weight exerts the torque

$$\boldsymbol{\tau} = \ell\,\hat{\boldsymbol{e}}_3 \times M\boldsymbol{g}\,, \qquad |\boldsymbol{\tau}| = Mg\ell\sin\theta\,,$$

with θ the tilt from the vertical. This torque is perpendicular to $\hat{\boldsymbol{e}}_3$ and hence to $\boldsymbol{L}$, so by $\frac{\mathrm{d}\boldsymbol{L}}{\mathrm{d}t} = \boldsymbol{\tau}$ it can turn $\boldsymbol{L}$ but cannot lengthen or shorten it — which is the whole reason the top turns instead of falling. Writing the motion of the axis as a rotation about the vertical at rate Ω, the tip of $\boldsymbol{L}$ travels on a circle of radius $I_3\omega_3\sin\theta$ at speed $\Omega I_3\omega_3\sin\theta$; equating this to $|\boldsymbol{\tau}|$, the factor $\sin\theta$ cancels and

$$\Omega = \frac{Mg\ell}{I_3\omega_3}\,, \tag{29.55}$$

independent of the tilt and inversely proportional to the spin, as observed. The one approximation made was to neglect the angular momentum of the precession itself beside that of the spin, which is what "fast" means; releasing it produces the superposed nutation of Phenomenon 29.38, and Equation (29.44) shows that averaging that nutation returns Equation (29.55) exactly. □

Foucault gave the instrument its name and exhibited it, in 1852, as a second mechanical proof of the Earth's rotation, complementary to the pendulum of the year before [136]: a pendulum keeps its plane against a rotating frame, and a spinning body keeps its axis. The next proposition is the quantitative form of that claim, and it is the principle of the gyrocompass.

Proposition 29.47 (The gyrocompass)**.** *A symmetric rotor spinning at ω_3, whose axis is constrained to remain horizontal but is free in azimuth, is stable with its spin axis pointing north, and oscillates about that direction at the angular frequency*

$$\omega_{gc} = \sqrt{\frac{C\omega_3\,\Omega\cos\lambda}{A}}\,, \tag{29.56}$$

with Ω the Earth's sidereal angular velocity and λ the latitude. The instrument therefore finds true north without magnetism and without sight of the sky, and it fails at the poles, where $\cos\lambda = 0$ [2027].

Rests on Theorem 29.7 and Equation (29.11).

Proof. *Derives Proposition 29.47.*

Use local axes at latitude λ with x east, y north and z up, in which the Earth's angular velocity is $\boldsymbol{\Omega} = \Omega\,(0, \cos\lambda, \sin\lambda)$. Let α be the azimuth of the spin axis measured from north towards east, so that $\hat{\boldsymbol{e}} = (\sin\alpha, \cos\alpha, 0)$. A rotation carrying north into east is a rotation by $-\alpha$ about $\hat{\boldsymbol{z}}$, so the slew contributes $-A\dot{\alpha}\hat{\boldsymbol{z}}$ to the angular momentum and

$$\boldsymbol{L} = C\omega_3\hat{\boldsymbol{e}} - A\dot{\alpha}\hat{\boldsymbol{z}}\,.$$

Apply Equation (29.4) in the frame rotating with the Earth: $(\mathrm{d}\boldsymbol{L}/\mathrm{d}t)_{\text{Earth}} + \boldsymbol{\Omega}\times\boldsymbol{L} = \boldsymbol{\tau}$. The rotor is balanced, so gravity exerts no torque about the support; the constraint that keeps the axis horizontal acts about a horizontal axis and so has no vertical component. Taking the vertical component of the equation,

$$-A\ddot{\alpha} + (\boldsymbol{\Omega}\times\boldsymbol{L})_z = 0\,, \qquad (\boldsymbol{\Omega}\times\boldsymbol{L})_z = \Omega_x L_y - \Omega_y L_x = -\Omega\cos\lambda\, C\omega_3\sin\alpha\,,$$

whence $A\ddot{\alpha} = -C\omega_3\Omega\cos\lambda\sin\alpha$. This is the pendulum equation with $\alpha = 0$ — the spin axis pointing north — as the stable equilibrium, and linearizing about it gives Equation (29.56). At $\cos\lambda = 0$ the restoring term vanishes identically and no direction is preferred. □

Example 29.48 (Order of magnitude of the gyrocompass period)**.** For a disc rotor, $C/A = 2$. Taking a spin of 100/s, that is $\omega_3 = 628\,\mathrm{rad/s}$, at latitude 45° with $\Omega = 7.292 \times 10^{-5}/\mathrm{s}$, Equation (29.56) gives $\omega_{\mathrm{gc}} = 0.255\,\mathrm{rad/s}$, a period of about 25 s. The result is a genuine consequence of the idealization stated in Proposition 29.47 — an axis held exactly horizontal and free in azimuth — and practical instruments differ from it, being made north-seeking by a pendulous suspension and deliberately damped and detuned; no engineering source is cited here for those refinements.

Rests on Equation (29.56).

29.4.4 Precession of the equinoxes and astronomical nutation

Table 29.2: Values adopted for the calculations of Sections 29.3.4, 29.4.4 and 29.5.3. These are the conventional astronomical and geodetic values; no primary reference is cited for them in this treatise, and each is quoted to the precision actually used below rather than to the precision at which it is known. The gravitational constant itself is CODATA 2018, $G = (6.674\,30 \pm 0.000\,15) \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$ [2392], and enters only through the products GM, which are known far better than G and M separately.

Quantity	Symbol	Value
Earth sidereal angular velocity	Ω	$7.292\,12 \times 10^{-5}/\mathrm{s}$
Earth equatorial radius	$a_\oplus$	$6.3781 \times 10^{6}\,\mathrm{m}$
Earth mass	$M_\oplus$	$5.9722 \times 10^{24}\,\mathrm{kg}$
Earth polar moment of inertia	C	$8.03 \times 10^{37}\,\mathrm{kg\,m^2}$
Earth dynamical ellipticity $(C-A)/C$	H	3.2738×10^{-3}
Obliquity of the ecliptic	ε	23.4393°
Heliocentric gravitational constant	$GM_\odot$	$1.327\,12 \times 10^{20}\,\mathrm{m^3/s^2}$
Lunar gravitational constant	GM_{M}	$4.9028 \times 10^{12}\,\mathrm{m^3/s^2}$
Earth orbital semi-major axis	$a_\odot$	$1.495\,98 \times 10^{11}\,\mathrm{m}$
Lunar orbital semi-major axis	a_{M}	$3.8440 \times 10^{8}\,\mathrm{m}$
Inclination of the lunar orbit to the ecliptic	i	5.145°

Phenomenon 29.49 (Precession of the equinoxes)**.** The Earth's axis of rotation is not fixed among the stars. It describes a cone about the pole of the ecliptic whose half-angle is the obliquity, about 23.4°, carrying the equinoctial points westward along the ecliptic by roughly fifty seconds of arc a year and completing one circuit in about $2.6 \times 10^4\,\mathrm{yr}$. Hipparchus detected the drift in the second century BC by comparing his own star positions with earlier ones, and it is why the pole star changes over historical time and why the zodiacal signs no longer coincide with the constellations that named them. Newton identified the cause: the gravitational torque exerted by the Moon and the Sun upon the Earth's equatorial bulge, so that the planet is a very slow and very large version of Phenomenon 29.46 [14]. The modern value of the drift is 50.29 seconds of arc a year, corresponding to a period of $2.577 \times 10^4\,\mathrm{yr}$.

Rests on Proposition 29.24 and Phenomenon 29.46.

Derivation. *Derives Phenomenon 29.49.*

Treat the Earth as a symmetric top with moments $A = B$ and C and symmetry axis $\hat{\boldsymbol{e}}_3$, and a perturber of mass M on a circular orbit of radius r in a plane with unit normal $\hat{\boldsymbol{n}}$. By Proposition 29.24 the orientation dependence of the potential energy is

$$U_{\mathrm{or}} = \frac{GM\,(C-A)}{2r^3}\left(3\,(\hat{\boldsymbol{e}}_3 \cdot \hat{\boldsymbol{r}})^2 - 1\right), \tag{29.57}$$

since for a symmetric top $\operatorname{tr} I - 3I_r = (C-A)\,(1-3\cos^2\Theta)$ with $\cos\Theta = \hat{\boldsymbol{e}}_3\cdot\hat{\boldsymbol{r}}$. The torque on the Earth is minus the rotational gradient of Equation (29.57),

$$\boldsymbol{\tau} = -\hat{\boldsymbol{e}}_3 \times \frac{\partial U_{\text{or}}}{\partial \hat{\boldsymbol{e}}_3} = -\frac{3GM\,(C-A)}{r^3}\,(\hat{\boldsymbol{e}}_3\cdot\hat{\boldsymbol{r}})\,(\hat{\boldsymbol{e}}_3\times\hat{\boldsymbol{r}})\ . \tag{29.58}$$

Now average over one orbit. Let $\hat{\boldsymbol{e}}_3 = \cos\varepsilon\,\hat{\boldsymbol{n}} + \sin\varepsilon\,\hat{\boldsymbol{u}}$ with $\hat{\boldsymbol{u}}$ in the orbit plane, and $\hat{\boldsymbol{v}} = \hat{\boldsymbol{n}}\times\hat{\boldsymbol{u}}$, and write $\hat{\boldsymbol{r}} = \cos\phi\,\hat{\boldsymbol{u}} + \sin\phi\,\hat{\boldsymbol{v}}$. Then $\hat{\boldsymbol{e}}_3\cdot\hat{\boldsymbol{r}} = \sin\varepsilon\cos\phi$ and

$$\hat{\boldsymbol{e}}_3\times\hat{\boldsymbol{r}} = \cos\varepsilon\cos\phi\,\hat{\boldsymbol{v}} - \cos\varepsilon\sin\phi\,\hat{\boldsymbol{u}} + \sin\varepsilon\sin\phi\,\hat{\boldsymbol{n}}\,.$$

Using $\langle\cos^2\phi\rangle = \tfrac{1}{2}$ and $\langle\cos\phi\sin\phi\rangle = 0$, only the $\hat{\boldsymbol{v}}$ term survives:

$$\langle\boldsymbol{\tau}\rangle = -\frac{3GM\,(C-A)}{2r^3}\sin\varepsilon\cos\varepsilon\ \hat{\boldsymbol{v}}\,. \tag{29.59}$$

Insert this into $\mathrm{d}\boldsymbol{L}/\mathrm{d}t = \langle\boldsymbol{\tau}\rangle$ with $\boldsymbol{L} = C\omega_3\hat{\boldsymbol{e}}_3$, exactly as in Phenomenon 29.46. A precession of $\hat{\boldsymbol{e}}_3$ about $\hat{\boldsymbol{n}}$ at rate Ω_p means $\mathrm{d}\hat{\boldsymbol{e}}_3/\mathrm{d}t = \Omega_p\hat{\boldsymbol{n}}\times\hat{\boldsymbol{e}}_3 = \Omega_p\sin\varepsilon\,\hat{\boldsymbol{v}}$, so the common factor $\sin\varepsilon$ cancels and

$$\Omega_p = -\frac{3}{2}\,\frac{GM}{r^3}\,\frac{C-A}{C}\,\frac{\cos\varepsilon}{\omega_3}\,, \tag{29.60}$$

the sign showing the drift to be retrograde — westward along the ecliptic — as observed.

Now put in the numbers of Table 29.2. For the Sun, $GM_\odot/a_\odot^3 = 3.964\times10^{-14}/\mathrm{s}^2$, and with $H = 3.2738\times10^{-3}$, $\cos\varepsilon = 0.91748$ and $\omega_3 = \Omega$,

$$|\Omega_p|_\odot = 2.449\times10^{-12}/\mathrm{s} = 15.94 \text{ seconds of arc per year}\,.$$

For the Moon, $GM_{\mathrm{M}}/a_{\mathrm{M}}^3 = 8.632\times10^{-14}/\mathrm{s}^2$, larger than the Sun's by the factor 2.177, so the coplanar estimate is 34.71 seconds of arc a year. The lunar orbit is not in the ecliptic but inclined by $i = 5.145^\circ$, and its node regresses in 18.6 yr, which is short compared with the precession; averaging Equation (29.57) a second time over that cone multiplies the orientation-dependent part by $1 - \frac{3}{2}\sin^2 i = 0.98794$, giving 34.29. The two contributions add to

$$15.94 + 34.29 = 50.23 \text{ seconds of arc per year}\,,$$

against the observed 50.29: agreement to one part in a thousand, from the mass and orbit of the Moon, the mass and orbit of the Sun, and one number describing the shape of the Earth. The residual is of the size of the terms deliberately dropped — the orbital eccentricities, the higher multipoles of the Earth's figure, and the periodic terms discarded by the averaging, which are the subject of the nutation. □

Proposition 29.50 (The principal nutation)**.** *The orbital averaging carried out in Phenomenon 29.49 discards a periodic remainder. Averaging the lunar torque over the month but* not *over the* 18.6 yr *regression of the lunar nodes, the Earth's axis oscillates about its mean precessional path with that period, with amplitudes*

$$\Delta\varepsilon = \frac{\Omega_M\sin i}{\Omega_N}\,, \qquad \Delta\psi = \frac{\Omega_M\sin i}{\Omega_N}\,\frac{2\cos 2\varepsilon}{\sin 2\varepsilon}\,, \tag{29.61}$$

in obliquity and in ecliptic longitude respectively. Here Ω_M is the magnitude of the mean lunar precession rate Equation (29.60), Ω_N is the rate at which the lunar nodes regress, and i is the inclination of the lunar orbit to the ecliptic. The two terms are in quadrature: the obliquity follows the cosine of the node angle and the longitude its sine.

Rests on Phenomenon 29.49 and Equations (29.57) and (29.60).

Proof. *Derives Proposition 29.50.*

Return to Equation (29.60) before the second averaging. What that equation describes is a precession of $\hat{\boldsymbol{e}}_3$ about the normal $\hat{\boldsymbol{n}}_{\mathrm{M}}$ of the *lunar* orbit, at a rate proportional to the cosine of the angle ε_{M} between the two:

$$\frac{\mathrm{d}\hat{\boldsymbol{e}}_3}{\mathrm{d}t} = \Lambda \cos\varepsilon_{\mathrm{M}}\,(\hat{\boldsymbol{n}}_{\mathrm{M}} \times \hat{\boldsymbol{e}}_3)\ , \qquad \Lambda = -\frac{3}{2}\,\frac{GM_{\mathrm{M}}}{a_{\mathrm{M}}^3}\,\frac{C-A}{C}\,\frac{1}{\omega_3}\,. \tag{29.62}$$

The lunar orbit normal is not the ecliptic pole $\hat{\boldsymbol{n}}$: it lies on a cone of half-angle i about it and travels once round that cone in 18.6 yr as the nodes regress. In axes with $\hat{\boldsymbol{n}} = \hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{e}}_3 = (\sin\varepsilon, 0, \cos\varepsilon)$,

$$\hat{\boldsymbol{n}}_{\mathrm{M}} = \cos i\,\hat{\boldsymbol{n}} + \sin i\,\hat{\boldsymbol{w}}(t)\,, \qquad \hat{\boldsymbol{w}} = (\cos\phi, \sin\phi, 0)\ , \qquad \phi = -\Omega_{\mathrm{N}} t\,, \tag{29.63}$$

so that $\cos\varepsilon_{\mathrm{M}} = \hat{\boldsymbol{e}}_3 \cdot \hat{\boldsymbol{n}}_{\mathrm{M}} = \cos i\cos\varepsilon + \sin i\sin\varepsilon\cos\phi$. To first order in i, Equation (29.62) splits into a secular part and a periodic one,

$$\frac{\mathrm{d}\hat{\boldsymbol{e}}_3}{\mathrm{d}t} = \Omega_{\mathrm{M}}\,(\hat{\boldsymbol{n}} \times \hat{\boldsymbol{e}}_3) + \Lambda \sin i\,[\sin\varepsilon\cos\phi\,(\hat{\boldsymbol{n}} \times \hat{\boldsymbol{e}}_3) + \cos\varepsilon\,(\hat{\boldsymbol{w}} \times \hat{\boldsymbol{e}}_3)]\ , \tag{29.64}$$

with $\Omega_{\mathrm{M}} := \Lambda\cos\varepsilon$ the mean rate already computed. Both periodic terms matter, and they are different things: the second is the wobble of the axis that the torque precesses the Earth about, the first the modulation of the rate at which it does so.

The response is obtained by integrating the periodic terms directly, because $\Omega_{\mathrm{M}}/\Omega_{\mathrm{N}} \approx 5\times10^{-4}$: the displacement $\boldsymbol{\delta}$ they build up varies at the node rate, so the action of the secular term upon $\boldsymbol{\delta}$ is smaller than $\dot{\boldsymbol{\delta}}$ by that factor and may be dropped. With $\hat{\boldsymbol{n}} \times \hat{\boldsymbol{e}}_3 = (0, \sin\varepsilon, 0)$ and $\hat{\boldsymbol{w}} \times \hat{\boldsymbol{e}}_3 = (\sin\phi\cos\varepsilon, -\cos\phi\cos\varepsilon, -\sin\phi\sin\varepsilon)$, the bracket of Equation (29.64) is

$$\left(\sin\phi\cos^2\varepsilon,\ -\cos\phi\cos 2\varepsilon,\ -\sin\phi\sin\varepsilon\cos\varepsilon\right)\ ,$$

and integrating in time, with $\int \sin\phi\,\mathrm{d}t = \cos\phi/\Omega_{\mathrm{N}}$ and $\int\cos\phi\,\mathrm{d}t = -\sin\phi/\Omega_{\mathrm{N}}$ from $\phi = -\Omega_{\mathrm{N}}t$,

$$\boldsymbol{\delta} = \frac{\Lambda\sin i}{\Omega_{\mathrm{N}}}\left(\cos\phi\cos^2\varepsilon,\ \sin\phi\cos 2\varepsilon,\ -\cos\phi\sin\varepsilon\cos\varepsilon\right)\ . \tag{29.65}$$

From Equations (29.63) and (29.64) (integrating the periodic forcing at the node frequency).

The obliquity is read from the third component: $\cos(\varepsilon + \Delta\varepsilon) = \hat{\boldsymbol{z}} \cdot (\hat{\boldsymbol{e}}_3 + \boldsymbol{\delta}) = \cos\varepsilon + \delta_3$ gives $\Delta\varepsilon = -\delta_3/\sin\varepsilon = (\Omega_{\mathrm{M}}\sin i/\Omega_{\mathrm{N}})\cos\phi$, on using $\Lambda\cos\varepsilon = \Omega_{\mathrm{M}}$. The longitude is read from the second, since displacing ψ moves $\hat{\boldsymbol{e}}_3$ by $\sin\varepsilon\,\Delta\psi$ in the y direction: $\Delta\psi = \delta_2/\sin\varepsilon = (\Omega_{\mathrm{M}}\sin i/\Omega_{\mathrm{N}})\,(2\cos 2\varepsilon/\sin 2\varepsilon)\sin\phi$. The amplitudes are Equation (29.61), quoted as magnitudes: Λ is negative, since the precession it drives is retrograde, and its sign fixes only the phase of each term relative to the node. Note where each amplitude came from: the obliquity term needs only the wobble of the torque axis, while the longitude term needs the rate modulation as well, and dropping that would leave $\cot\varepsilon$ in place of $2\cot 2\varepsilon$ and an answer a quarter too large. □

Put in the numbers. The lunar rate is the coplanar 34.71 seconds of arc a year found above, the node rate is $\Omega_{\mathrm{N}} = 2\pi/18.6\,\mathrm{yr}$, that is 69 677 seconds of arc a year, and $i = 5.145°$ from Table 29.2, so that $\sin i = 0.089\,68$. Hence $\Omega_{\mathrm{M}}\sin i/\Omega_{\mathrm{N}} = 4.47\times10^{-5}$ radians, and with $\varepsilon = 23.4393°$,

$$\Delta\varepsilon = 9.21 \text{ seconds of arc}\,, \qquad \Delta\psi = 17.26 \text{ seconds of arc}\,,$$

against the conventional observed principal terms 9.20 and 17.20; as with Table 29.2, no primary reference is cited here for those two modern values. This is the nodding that Bradley extracted from some twenty years of stellar observations and reported in 1748 [26]. The Sun contributes nothing at this period—its orbit normal *is* the ecliptic pole, so the wobble that drives the effect is absent—and its own nutation terms are semiannual and an order of magnitude smaller; they are not computed here.

29.5 Motion in rotating frames

29.5.1 The transport theorem and the fictitious forces

Theorem 29.51 (Equation of motion in a rotating frame)**.** *Let a frame rotate with angular velocity $\boldsymbol{\omega}(t)$ about a fixed origin of an inertial frame. A particle of mass m at position $\boldsymbol{r}$, with velocity $\boldsymbol{v}$ and acceleration $\boldsymbol{a}$ as measured in the rotating frame, obeys*

$$m\boldsymbol{a} = \boldsymbol{F} - 2m\,\boldsymbol{\omega}\times\boldsymbol{v} - m\,\boldsymbol{\omega}\times(\boldsymbol{\omega}\times\boldsymbol{r}) - m\,\dot{\boldsymbol{\omega}}\times\boldsymbol{r}\,, \tag{29.66}$$

where $\boldsymbol{F}$ is the true force. The three added terms are the Coriolis *[100],* centrifugal *[11] and* Euler *forces.*

Rests on Theorem 29.7 and Postulate 19.8.

Proof. *Derives Theorem 29.51.*

Apply Equation (29.4) to the position vector: $(\mathrm{d}\boldsymbol{r}/\mathrm{d}t)_{\text{space}} = \boldsymbol{v} + \boldsymbol{\omega}\times\boldsymbol{r}$. Apply it again, to that whole vector:

$$\left(\frac{\mathrm{d}^2\boldsymbol{r}}{\mathrm{d}t^2}\right)_{\text{space}} = \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)_{\text{rot}}(\boldsymbol{v}+\boldsymbol{\omega}\times\boldsymbol{r}) + \boldsymbol{\omega}\times(\boldsymbol{v}+\boldsymbol{\omega}\times\boldsymbol{r})\,.$$

Expanding, the first bracket gives $\boldsymbol{a} + \dot{\boldsymbol{\omega}}\times\boldsymbol{r} + \boldsymbol{\omega}\times\boldsymbol{v}$, so

$$\left(\frac{\mathrm{d}^2\boldsymbol{r}}{\mathrm{d}t^2}\right)_{\text{space}} = \boldsymbol{a} + 2\,\boldsymbol{\omega}\times\boldsymbol{v} + \boldsymbol{\omega}\times(\boldsymbol{\omega}\times\boldsymbol{r}) + \dot{\boldsymbol{\omega}}\times\boldsymbol{r}\,.$$

Newton's second law holds in the inertial frame (Postulate 19.8), so $m\,(\mathrm{d}^2\boldsymbol{r}/\mathrm{d}t^2)_{\text{space}} = \boldsymbol{F}$; solving for $m\boldsymbol{a}$ gives Equation (29.66). □

Two features of Equation (29.66) are worth stating plainly. The centrifugal term depends on position but not on velocity, so it can be absorbed into a potential — and is, whenever one speaks of "effective gravity". The Coriolis term depends on velocity but not on position, does no work, since it is perpendicular to $\boldsymbol{v}$, and vanishes for a body at rest in the rotating frame; it cannot therefore be absorbed into any potential, and it is the term that carries the detectable signature of rotation for moving bodies.

Example 29.52 (Sizes of the three terms at the Earth's surface)**.** With $\Omega = 7.292\times10^{-5}/\mathrm{s}$ and $a_\oplus = 6.378\times10^{6}\,\mathrm{m}$:

- centrifugal, at the equator: $\Omega^2 a_\oplus = 3.39\times10^{-2}\,\mathrm{m/s^2}$, about $0.35\,\%$ of g;
- Coriolis, for a body moving at $10\,\mathrm{m/s}$: at most $2\Omega v = 1.46\times10^{-3}\,\mathrm{m/s^2}$, some twenty times smaller;
- Euler: the Earth's rotation rate changes by a few milliseconds of day length per century, giving $\dot\Omega \sim 5\times10^{-22}/\mathrm{s^2}$ and $\dot\Omega a_\oplus \sim 3\times10^{-15}\,\mathrm{m/s^2}$, which is seven orders of magnitude below the sensitivity of the best absolute gravimeter [1906] and is neglected everywhere below.

Rests on Equation (29.66).

29.5.2 Absolute rotation: Newton's bucket and Mach's critique

Phenomenon 29.53 (Rotation is locally detectable, uniform motion is not)**.** A sealed laboratory cannot tell whether it is at rest or moving uniformly, but it can always tell whether it is rotating, and how fast, without consulting anything outside. Water in a spinning bucket climbs the walls and takes a curved surface; a plumb line hangs off the

true vertical; a gyroscope holds a direction the walls do not; a pendulum turns its plane of swing. Newton took the bucket as evidence for absolute space [14]; Mach replied that the water's surface may be answering not to space but to the distant matter of the universe [225]. What is not in dispute is the observation itself: the equivalence of frames that Chapter 18 establishes for uniform translation does not extend to rotation.

Rests on Equation (29.66) *and Postulate 18.16.*

Derivation. *Derives Phenomenon 29.53.*

Work in the frame rotating with the bucket at angular velocity ω about the vertical, in which the water is at rest. The centrifugal term of Equation (29.66) is a gradient and contributes to the effective potential energy per unit mass the amount $-\frac{1}{2}\omega^2 r^2$, so a fluid element at height z and distance r from the axis carries

$$\Phi(r, z) = gz - \tfrac{1}{2}\omega^2 r^2 \,.$$

A free surface in equilibrium must be a surface of constant Φ — otherwise there would be an unbalanced force tangent to it and the fluid would not be at rest — so the surface obeys

$$z = z_0 + \frac{\omega^2 r^2}{2g} \,, \tag{29.67}$$

a paraboloid of revolution whose curvature reports ω^2 directly, from measurements made entirely inside the bucket. Since ω enters quadratically, the reading is blind to the sense of the rotation but never vanishes for $\omega \neq 0$: there is no rotation rate that mimics rest. Contrast the derivation of Chapter 18, where the constancy of the relative velocity made the accelerations — and with them every mechanical observable inside the cabin — identical in the two frames. The difference is that the map to a rotating frame is not constant in time, and its second time derivative does not vanish. □

Remark 29.54 (What the bucket does and does not settle). Equation (29.67) establishes that the water responds to *something*, and that the something is not its motion relative to the bucket, since the water and the bucket end up turning together while the surface stays curved. Newton concluded that the response is to absolute space [14]. Mach's objection is that the experiment has never been performed with the distant matter of the universe removed, so the observation cannot distinguish rotation relative to space from rotation relative to that matter [225]. The objection is not answered by any experiment in this chapter, and it is not answered by mechanics at all; what general relativity has to say about it — that local inertial frames are determined by the metric, which is in turn sourced by matter, but that the determination is not the complete one Mach wanted — belongs to Chapter 43.

Rests on Phenomenon 29.53.

29.5.3 Terrestrial evidence of the Earth's rotation

Phenomenon 29.55 (Effective gravity varies with latitude)**.** The acceleration of free fall measured at the Earth's surface is not the same everywhere. It rises smoothly from about $9.780\,\mathrm{m/s^2}$ at the equator to about $9.832\,\mathrm{m/s^2}$ at the poles, and away from those two places a plumb line does not point at the Earth's centre. The variation is enormous by the standards of the instrument that measures it: an absolute gravimeter is accurate to some $2 \times 10^{-8}\,\mathrm{m/s^2}$ [1906], so the equator-to-pole difference is resolved a million times over. The conventional value $9.806\,65\,\mathrm{m/s^2}$ adopted for the SI is accordingly a convention about which latitude to name, not a measurement of a universal constant [2333].

Rests on Equation (29.66).

Derivation. *Derives Phenomenon 29.55.*

In the frame rotating with the Earth, a body at rest on the surface experiences, besides the gravitational attraction, a centrifugal acceleration $\omega^2\rho$ directed away from the rotation axis, where $\rho = R\cos\lambda$ is the distance from that axis at geographic latitude λ. Its component along the local outward vertical is $\omega^2 R\cos^2\lambda$, and the measured free-fall acceleration is reduced by exactly that amount,

$$g(\lambda) \simeq g_{\mathrm{grav}} - \omega^2 R\cos^2\lambda\,, \tag{29.68}$$

largest in effect at the equator and zero at the poles. With the sidereal rotation rate $\omega = 7.292\times10^{-5}/\mathrm{s}$ and the equatorial radius $R = 6.378\times10^{6}\,\mathrm{m}$,

$$\omega^2 R = 3.39\times10^{-2}\,\mathrm{m/s^2}\,,$$

about two thirds of the observed equator-to-pole difference of $5.2\times10^{-2}\,\mathrm{m/s^2}$. The remaining third is not a defect of the calculation but a second consequence of the same rotation: the Earth is not a sphere, its equatorial radius exceeding the polar by some 21 km, so equatorial ground stands farther from the mass that attracts it. There is also a component of the centrifugal acceleration tangent to the surface, $\omega^2 R\sin\lambda\cos\lambda$, which is what tilts the plumb line; it is largest near 45°, where it amounts to about $1.7\times10^{-2}\,\mathrm{m/s^2}$ and deflects the vertical by roughly a tenth of a degree. □

Phenomenon 29.56 (Falling bodies are deflected eastward)**.** A body released from rest and allowed to fall freely does not land at the point vertically below the release. It lands slightly to the east, by an amount that grows as the three-halves power of the drop height and is largest at the equator. Reich measured it in 1831 in a mineshaft at Freiberg, at latitude about 50.9°: from a drop of 158.5 m, repeated 106 times, the mean eastward displacement was 28.3 mm [92]. The effect is a direct mechanical demonstration of the Earth's rotation, and unlike the pendulum it is over in a few seconds.

Rests on Equation (29.66).

Derivation. *Derives Phenomenon 29.56.*

Use local axes at latitude λ with x east, y north and z up, so that $\boldsymbol{\Omega} = \Omega\,(0, \cos\lambda, \sin\lambda)$. The centrifugal term is absorbed into the effective gravity by Equation (29.68), and the Euler term is negligible (Example 29.52), so Equation (29.66) leaves

$$\ddot{\boldsymbol{r}} = -g\hat{\boldsymbol{z}} - 2\,\boldsymbol{\Omega}\times\dot{\boldsymbol{r}}\,.$$

Solve perturbatively in Ω. At zeroth order $\dot{\boldsymbol{r}} = (0, 0, -gt)$; substituting this into the Coriolis term,

$$-2\,\boldsymbol{\Omega}\times\dot{\boldsymbol{r}} = -2\Omega\,(\cos\lambda\,(-gt) - 0,\ 0,\ 0) = (2\Omega gt\cos\lambda,\ 0,\ 0)\ ,$$

which points due east. Integrating twice from rest,

$$x(t) = \tfrac{1}{3}\Omega g\cos\lambda\, t^3 = \tfrac{1}{3}\Omega\cos\lambda\sqrt{\frac{8h^3}{g}}\,, \tag{29.69}$$

using $h = \tfrac{1}{2}gt^2$ for the drop. There is no first-order deflection in the north–south direction: the Coriolis term acting on a purely vertical velocity has no northward component, and the southward term generated at second order is smaller by a further factor Ωt.

For Reich's shaft, $h = 158.5\,\mathrm{m}$ gives $t = 5.68\,\mathrm{s}$, and with $\lambda = 50.9°$ and $\Omega = 7.292\times10^{-5}/\mathrm{s}$, Equation (29.69) predicts 27.6 mm, against the measured mean of 28.3 mm [92] — agreement to 3 %, which for a measurement of a 28 mm displacement at the bottom of a 160 m shaft in 1831 is remarkable. The prediction contains no adjustable parameter: h and λ are properties of the shaft and Ω is the rotation rate of the Earth. □

Proposition 29.57 (The Eötvös effect)**.** *A body moving horizontally over the Earth's surface has its apparent weight changed by the vertical component of the Coriolis force. For an eastward velocity component v_E at latitude λ the measured gravitational acceleration is reduced by*

$$\Delta g = 2\Omega v_E \cos\lambda\,, \tag{29.70}$$

and increased by the same amount for westward motion. There is no such effect for north–south motion at this order.

Rests on Equation (29.66) *and Phenomenon 29.55.*

Proof. *Derives Proposition 29.57.*

In the local axes of Phenomenon 29.56, with $\dot{\boldsymbol{r}} = (v_E, v_N, 0)$,

$$-2\,\boldsymbol{\Omega}\times\dot{\boldsymbol{r}} = 2\Omega\,(\sin\lambda\, v_N,\ -\sin\lambda\, v_E,\ \cos\lambda\, v_E)\ .$$

The vertical component is $2\Omega v_E\cos\lambda$, upward for $v_E > 0$; a gravimeter, which weighs against the local vertical, therefore reads low by that amount, which is Equation (29.70). The component v_N contributes nothing vertical. □

The effect is not small on the scale of the instrument. A ship steaming east at $5\,\mathrm{m/s}$ near the equator carries $\Delta g = 7.3\times10^{-4}\,\mathrm{m/s^2}$, some four orders of magnitude above the $2\times10^{-8}\,\mathrm{m/s^2}$ accuracy of an absolute gravimeter [1906]; marine and airborne gravity surveys must correct for it before any geological signal can be read, and the correction is named after Eötvös, who identified it in shipborne survey data. No primary reference for that attribution is cited here.

Phenomenon 29.58 (Weather systems circulate, and the sense is set by the hemisphere)**.** Air does not flow straight from high pressure to low. In the northern hemisphere it circulates anticlockwise about a low-pressure centre and clockwise about a high, seen from above; in the southern hemisphere both senses reverse; near the equator the organized circulation fails altogether and no tropical cyclone forms within a few degrees of it. The wind blows very nearly *along* the isobars rather than across them, and its speed can be read off a weather chart from the isobar spacing alone. Hadley explained the steadiness of the trade winds by the Earth's rotation in 1735 [21]; the dynamical term responsible is Coriolis's [100].

Rests on Equation (29.66).

Derivation. *Derives Phenomenon 29.58.*

For horizontal motion only the vertical component of $\boldsymbol{\Omega}$ matters, since the horizontal component produces a vertical Coriolis force that the pressure field balances (this is the same separation used in Proposition 29.57). Define the *Coriolis parameter*

$$f := 2\Omega\sin\lambda\,, \tag{29.71}$$

which at 45° is $1.03\times10^{-4}/\mathrm{s}$ and vanishes on the equator. In the local horizontal plane the equation of motion of a fluid element of density ρ is, from Equation (29.66) with the pressure-gradient force of Chapter 31,

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = -\frac{1}{\rho}\nabla_h p - f\,\hat{\boldsymbol{z}}\times\boldsymbol{u}\,.$$

When the flow is steady and slowly varying in space the left-hand side is negligible against either term on the right, and the two balance:

$$f\,\hat{\boldsymbol{z}}\times\boldsymbol{u} = -\frac{1}{\rho}\nabla_h p \qquad\Longrightarrow\qquad \boldsymbol{u} = \frac{1}{\rho f}\,\hat{\boldsymbol{z}}\times\nabla_h p\,. \tag{29.72}$$

This is *geostrophic balance*, and it says three things at once. The velocity is perpendicular to the pressure gradient, hence along the isobars. Its sense is fixed by the sign of f, so it

reverses across the equator. And its magnitude is $|\boldsymbol{u}| = |\nabla_h p|\,/(\rho f)$: for a typical gradient of one hectopascal per 100 km, that is $1 \times 10^{-3}\,\mathrm{N/m^3}$, with $\rho = 1.2\,\mathrm{kg/m^3}$ and f at 45°, the wind is 8.1 m/s.

Whether the balance holds is decided by the ratio of the neglected term to the retained one, the *Rossby number* $\mathrm{Ro} = U/(fL)$ for a flow of speed U over a length L. A mid-latitude weather system has $U \sim 10\,\mathrm{m/s}$ and $L \sim 1 \times 10^6\,\mathrm{m}$, so $\mathrm{Ro} \approx 0.10$ and the balance is good. Near the equator $f \to 0$ and Ro diverges, which is why organized cyclonic circulation does not form there. □

Remark 29.59 (The bathtub, and the limits of the argument). The same estimate disposes of the common claim that a draining basin turns one way in the northern hemisphere and the other in the southern. With $U \sim 0.1\,\mathrm{m/s}$ and $L \sim 0.3\,\mathrm{m}$ the Rossby number is $\mathrm{Ro} \approx 3 \times 10^3$: the Coriolis term is three thousand times smaller than the terms neglected in Equation (29.72), and any residual swirl left in the water, or any asymmetry of the basin, dominates it completely. The effect is real and has been made visible in laboratory tanks left still for a day, but it is not what an ordinary basin shows. The lesson is general: a term in Equation (29.66) being non-zero is not the same as its being the term that decides the motion.

Rests on Phenomenon 29.58.

29.5.4 The Foucault pendulum

Theorem 29.60 (Planar central motion in a rotating frame)**.** *Let a particle move in a plane under a central force $\boldsymbol{F} = F(|\boldsymbol{r}|)\,\boldsymbol{r}$, and let the frame rotate uniformly at $\boldsymbol{\Omega} = \Omega\hat{\boldsymbol{z}}$ about the normal to that plane. Writing the position as the complex number $\zeta = x + \mathrm{i}y$, the rotating-frame equation of motion*

$$\ddot{\zeta} = \frac{F}{m}\zeta - 2\mathrm{i}\Omega\dot{\zeta} + \Omega^2\zeta \tag{29.73}$$

is solved exactly *by $\zeta = \mathrm{e}^{-\mathrm{i}\Omega t}\eta$, where η obeys the non-rotating equation $\ddot{\eta} = (F/m)\,\eta$. The motion in the rotating frame is therefore the inertial motion referred to axes turning at $-\Omega$: the Coriolis and centrifugal terms together amount to a rigid rotation of the orbit and to nothing else.*

Rests on Equation (29.66).

Proof. *Derives Theorem 29.60.*

Equation (29.73) is Equation (29.66) written in the plane: for $\boldsymbol{\Omega} = \Omega\hat{\boldsymbol{z}}$ the Coriolis term $-2\boldsymbol{\Omega} \times \boldsymbol{v}$ becomes $-2\mathrm{i}\Omega\dot{\zeta}$ and the centrifugal term $-\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \boldsymbol{r})$ becomes $+\Omega^2\zeta$. Substituting $\zeta = \mathrm{e}^{-\mathrm{i}\Omega t}\eta$,

$$\dot{\zeta} = \mathrm{e}^{-\mathrm{i}\Omega t}\left(\dot{\eta} - \mathrm{i}\Omega\eta\right), \qquad \ddot{\zeta} = \mathrm{e}^{-\mathrm{i}\Omega t}\left(\ddot{\eta} - 2\mathrm{i}\Omega\dot{\eta} - \Omega^2\eta\right).$$

The right-hand side of Equation (29.73) becomes $\mathrm{e}^{-\mathrm{i}\Omega t}\left[(F/m)\,\eta - 2\mathrm{i}\Omega\dot{\eta} - 2\Omega^2\eta + \Omega^2\eta\right]$, and the factor F is unchanged because $|\zeta| = |\eta|$. Cancelling $\mathrm{e}^{-\mathrm{i}\Omega t}$ and subtracting, the terms in $\dot{\eta}$ and the terms in $\Omega^2\eta$ cancel identically on both sides, and $\ddot{\eta} = (F/m)\,\eta$ remains. No approximation in Ω was made. □

The exact cancellation is worth dwelling on: the centrifugal term is not "small enough to neglect" in the Foucault problem, it is *absorbed*, and its whole effect is already contained in the rotation of the axes. Only the first power of Ω survives in the observable rate. The same theorem, applied to an electron in a magnetic field rather than to a mass in a rotating frame, is Larmor's.

Phenomenon 29.61 (The Foucault pendulum)**.** A long pendulum, set swinging carefully in a plane and then left alone, does not keep that plane with respect to the room. The plane turns steadily — clockwise seen from above in the northern hemisphere, anticlockwise in the southern — at a rate proportional to the sine of the latitude: one full turn in about thirty-two hours at the latitude of Paris, in one sidereal day at either pole, and not at all on the equator. Nothing outside the building is consulted and the room may be sealed, so the experiment measures the rotation of the Earth from inside a laboratory. That is what made Foucault's public demonstration of 1851 the version of the argument that convinced people who had not been convinced by astronomy [134]. The measurement, its apparatus and its data are Section 34.5.

Rests on Theorem 29.60 and Equation (29.66)*.*

Derivation. *Derives Phenomenon 29.61.*

Use the local axes of Phenomenon 29.56, in which $\boldsymbol{\Omega} = \Omega\,(0, \cos\lambda, \sin\lambda)$, and let the bob swing with small amplitude about the lowest point of its suspension. Three reductions bring the problem into the form of Theorem 29.60.

First, the centrifugal term of Equation (29.66) is independent of time and is absorbed once and for all into the effective gravity and the local vertical, exactly as in Equation (29.68); the swing is then about that vertical.

Second, split $\boldsymbol{\Omega}$ into its vertical and horizontal parts, $\Omega_z = \Omega\sin\lambda$ and $\Omega_y = \Omega\cos\lambda$. For small oscillations the bob's velocity is horizontal to first order, since the vertical excursion is second order in the amplitude. The Coriolis force built from Ω_y acting on a horizontal velocity is then vertical, and is taken up by the tension in the wire, which is free to adjust; it does not move the bob horizontally. Only Ω_z survives in the horizontal equations, and this is why the experiment reports the component of $\boldsymbol{\Omega}$ along the local vertical rather than the whole of it.

Third, for small amplitude the horizontal restoring force is central and linear, $\boldsymbol{F} = -m\omega_0^2\boldsymbol{r}$ with $\omega_0^2 = g/\ell$ (Equation (28.7)). Theorem 29.60 now applies verbatim with $\Omega \to \Omega_z$: the substitution $\zeta = \mathrm{e}^{-\mathrm{i}\Omega_z t}\eta$ reduces the motion to $\ddot{\eta} = -\omega_0^2\eta$, a plane pendulum whose plane is fixed. Seen from the room, that plane therefore turns at

$$\Omega_{\mathrm{prec}} = -\Omega\sin\lambda\,, \tag{29.74}$$

clockwise from above for $\lambda > 0$, one turn per sidereal day at the pole and none on the equator. Note what does *not* appear: the length, the mass and the amplitude of the pendulum are all absent, so the effect cannot be confused with a property of the particular apparatus. The numerical value at the latitude of Paris, and the data that confirm it, are Equation (34.39) and Section 34.5. □

Remark 29.62 (The same rate read as a holonomy)*.* There is a second reading of Equation (29.74) that uses no forces at all. Over one sidereal day the pendulum is carried once around its circle of latitude, and its swing plane, which no torque about the vertical acts on, is parallel-transported along that circle in the sense of Definition 13.146. Relative to the ground the plane has turned by $2\pi\sin\lambda$, so relative to a direction carried rigidly with the rotating Earth it has failed to return by the angle

$$2\pi - 2\pi\sin\lambda = 2\pi\,(1 - \sin\lambda)\ ,$$

which is numerically the solid angle subtended at the Earth's centre by the spherical cap that the latitude circle bounds, a cap of colatitude $\frac{\pi}{2} - \lambda$ having solid angle $2\pi\,(1 - \sin\lambda)$. The two computations are independent and they agree. That the agreement is a theorem rather than a coincidence — that the holonomy of parallel transport around a closed curve equals the integrated curvature it encloses — is not proved anywhere in this treatise, and nothing above depends on it; the geometric machinery is that of Chapter 13 and Chapter 14.

Read this way the Foucault pendulum is the oldest laboratory measurement of a geometric phase.

Rests on Phenomenon 29.61 and Definition 13.146.

Chapter 30

Continuum Mechanics and Elasticity

Between the finitely many degrees of freedom of Chapter 29 and the fluids of Chapter 31 stands the deformable solid. This chapter replaces the rigid body's six coordinates by continuous fields: the displacement field and its strain tensor, the stress tensor that transmits force through matter, and the linear law — Hooke's — that couples them. From these follow the equations of elastodynamics, the elastic waves that carry earthquakes across the planet, the engineering theories of rods, beams and plates, and the measured elastic constants that anchor every parameter of the theory to experiment. The continuum limit constructed here is also the pattern for every later field theory in the book, from the fluids of the next chapter to the lattice dynamics of Chapter 125 and the Lagrangian fields of Chapter 64. The canonical monograph is Landau and Lifshitz's *Theory of Elasticity* [1778].

The tensor machinery this chapter uses — the transformation law under orthogonal changes of frame, the algebra of Cartesian tensors, the quotient law — is that of Chapter 13, and is not rebuilt here. What is built here is physics: which tensors the solid carries, what the observed relation between them is, and what follows from it.

Notation 30.1 (Strain and displacement share a letter). Following [1778], u_i with one index denotes a component of the displacement *field* and u_{ik} with two denotes a component of the strain *tensor*; they are different objects, related by Equation (30.4). Stress is σ_{ik}, mass density ρ, and repeated Latin indices are summed over the three spatial directions. All indices are Cartesian, so index position carries no meaning (Definition 13.5); the metric is δ_{ik}.

Rests on Definition 13.5.

30.1 The continuum hypothesis

A cubic centimetre of copper holds of order 10^{23} nuclei and as many electrons, and no theory of matter that tracks them individually will ever predict the sag of a bridge. The whole of this chapter rests on replacing that population by a handful of smooth fields. The replacement is not a mathematical convenience but a physical claim with a domain of validity, and the claim can be stated sharply enough to be tested.

Definition 30.2 (Representative volume element; continuum field)**.** Let a be the mean spacing between the constituent particles of a body and L the scale over which the applied loads and the geometry vary. A *representative volume element* at the point $\boldsymbol{x}$ is a region of linear size ℓ centred there with

$$a \ll \ell \ll L\,. \tag{30.1}$$

A *continuum field* is a quantity assigned to $\boldsymbol{x}$ by averaging the corresponding microscopic quantity over the representative volume element: the mass density $\rho(\boldsymbol{x}) = \Delta m/\Delta V$, the velocity $\boldsymbol{v}(\boldsymbol{x}) = \Delta \boldsymbol{p}/\Delta m$, and so on. The *continuum hypothesis* is the assertion that Equation (30.1) can be satisfied, so that the averages are insensitive to ℓ over a wide range and the resulting fields are smooth on the scale L.

Rests on Definition 7.65.

The hypothesis is quantitative because the insensitivity is quantitative. Density fluctuations in a volume holding N particles are of relative size $N^{-1/2}$ (Chapter 55), so the width of the plateau in ℓ is set by how fast N grows.

Example 30.3 (How small a piece of copper is still a continuum)**.** Copper has mass density $8960\,\mathrm{kg/m^3}$ and molar mass $63.55\,\mathrm{g/mol}$, giving a number density $n = 8.49 \times 10^{28}/\mathrm{m^3}$ and a mean spacing $a = n^{-1/3} = 0.227\,\mathrm{nm}$. A cube of side $100\,\mathrm{nm}$ then contains $N = 8.5 \times 10^7$ atoms and its density is defined to a relative precision $N^{-1/2} = 1.1 \times 10^{-4}$. That cube is already 440 atoms across, so it is large in the sense of Equation (30.1); and it is 10^4 times smaller than a millimetre-scale specimen, so it is also small in that sense. Both inequalities in Equation (30.1) hold with four orders of magnitude to spare, which is why elasticity works so well for ordinary solids and why nothing in this chapter needs to know that copper is made of atoms.

Rests on Definition 30.2.

Definition 30.4 (Material and spatial descriptions)**.** Label each material particle by its position $\boldsymbol{X}$ in a chosen reference configuration. The motion is the map $\boldsymbol{x} = \boldsymbol{\chi}(\boldsymbol{X}, t)$, and the *displacement* is

$$\boldsymbol{u}(\boldsymbol{X}, t) = \boldsymbol{\chi}(\boldsymbol{X}, t) - \boldsymbol{X}\,. \tag{30.2}$$

A field written as a function of $(\boldsymbol{X}, t)$ is in the *material* (Lagrangian) description; written as a function of $(\boldsymbol{x}, t)$, in the *spatial* (Eulerian) description. The two are related by the chain rule, so that for any field f

$$\left.\frac{\partial f}{\partial t}\right|_{\boldsymbol{X}} = \left.\frac{\partial f}{\partial t}\right|_{\boldsymbol{x}} + (\boldsymbol{v}\cdot\boldsymbol{\nabla})\, f\,, \tag{30.3}$$

the second form being the *material derivative*.

Rests on Proposition 7.72 and Definition 30.2.

A solid has a preferred reference configuration — the unloaded one, to which it returns — so solid mechanics is naturally Lagrangian. A fluid has none, which is why Chapter 31 is written in the spatial description throughout; Equation (30.3) is the bridge between the two chapters, and the nonlinear term it carries is the origin of the difficulties of that one.

Remark 30.5 (Where the hypothesis fails, and how one can tell)*.* Equation (30.1) is an inequality between measurable lengths, so its failure is measurable too.

- *Rarefied gases.* The relevant microscopic length is the mean free path ℓ_{mfp}, not the particle spacing. Air at room conditions has $\ell_{\mathrm{mfp}} \approx 68\,\mathrm{nm}$, negligible against any laboratory apparatus; but in a high vacuum, or in the upper atmosphere, ℓ_{mfp} grows to metres and the Knudsen number ℓ_{mfp}/L approaches unity. Beyond $\ell_{\mathrm{mfp}}/L \sim 0.1$ the continuum equations of Chapter 31 fail and the kinetic description of Chapter 55 must be used instead.

- *Nanoscale structures.* A wire ten atoms across has no interior: most of its atoms are surface atoms, with a different coordination and a different local stiffness, so no plateau in ℓ exists and the elastic moduli measured on such a specimen depend on its size.

- *Singular regions inside a continuum.* A shock front is a few mean free paths thick and a crack tip has a radius of a few lattice spacings. In both cases Equation (30.1) fails *locally* while holding everywhere else, which is why the continuum theory can keep them as idealised surfaces and points — and also why it predicts infinite stress there (Section 30.7).

What underwrites the hypothesis where it holds is the kinetic theory of Chapter 55: it is the law of large numbers applied to a mechanical system, and the smallness of $N^{-1/2}$ is the whole of the argument.

Rests on Definition 30.2 and Equation (30.1).

30.2 Deformation: the strain tensor

30.2.1 Displacement and linear strain

A displacement field that is the same everywhere translates the body without deforming it, and one that is linear and antisymmetric rotates it. Deformation is therefore carried by the symmetric part of the displacement *gradient*, and this is what the strain tensor isolates. Cauchy introduced both it and the stress tensor of Section 30.3 in the memoir that founded the subject [73].

Definition 30.6 (Linear strain and infinitesimal rotation)**.** For a displacement field $\boldsymbol{u}$ with small gradients, the *strain tensor* and the *rotation tensor* are the symmetric and antisymmetric parts of $\partial_i u_k$:

$$u_{ik} = \tfrac{1}{2}\left(\partial_i u_k + \partial_k u_i\right), \qquad \omega_{ik} = \tfrac{1}{2}\left(\partial_i u_k - \partial_k u_i\right). \tag{30.4}$$

Both are dimensionless. The *dilatation* is the trace $u_{ll} = \boldsymbol{\nabla}\cdot\boldsymbol{u}$, and the traceless remainder

$$d_{ik} = u_{ik} - \tfrac{1}{3}u_{ll}\delta_{ik} \tag{30.5}$$

is the *deviatoric* or shear part.

Rests on Equation (30.2) and Definition 13.91.

Proposition 30.7 (Strain is a Cartesian tensor of rank two)**.** *Under a change of orthonormal frame $x'_i = a_{ik}x_k$ the nine quantities u_{ik} transform as $u'_{ik} = a_{il}a_{km}u_{lm}$, that is, as the components of a rank-two tensor in the sense of Definition 13.5. The same holds for ω_{ik}, d_{ik}, and the trace u_{ll} is a scalar.*

Rests on Definitions 13.5 and 30.6.

Proof. *Derives Proposition 30.7.*

The displacement is a vector, $u'_k = a_{km}u_m$, and for an orthogonal transformation with constant coefficients the differential operator carries a vector index as well, $\partial'_i = a_{il}\partial_l$. The product $\partial_i u_k$ is therefore a rank-two tensor by the product rule of Section 13.2.2, its symmetric and antisymmetric parts are tensors by the addition and index-permutation rules there, and the trace is a rank-zero tensor by the contraction rule. That the trace is a scalar is the statement that $\boldsymbol{\nabla}\cdot\boldsymbol{u}$ does not depend on the frame in which it is computed, which is what makes Proposition 30.8 a statement about the body rather than about the observer. □

This is the reason the subject is written in tensors at all. A material law must relate the deformation to the force without reference to the orientation of the laboratory; a relation between objects with the transformation law of Equation (13.22) has that property automatically, and one between arbitrary arrays of nine numbers does not.

Proposition 30.8 (The trace is the fractional volume change)**.** *To first order in the displacement gradients, a material element of volume* $\mathrm{d}V$ *becomes*

$$\mathrm{d}V' = (1 + u_{ll})\,\mathrm{d}V\,, \tag{30.6}$$

so that u_{ll} *is the fractional change of volume and the deviatoric part* d_{ik} *deforms at constant volume.*

Rests on Definition 30.6 and Equation (30.2).

Proof. *Derives Proposition 30.8.*

The map $\boldsymbol{X} \mapsto \boldsymbol{X} + \boldsymbol{u}(\boldsymbol{X})$ has Jacobian matrix $F_{ik} = \delta_{ik} + \partial_k u_i$, and a volume element is multiplied by $\det F$. For any matrix A, $\det(\mathbb{1} + A) = 1 + \operatorname{tr} A + O(A^2)$, since the terms of the determinant that are not the product of diagonal entries contain at least two off-diagonal factors. Hence $\det F = 1 + \partial_i u_i + O\left((\partial u)^2\right)$, and $\partial_i u_i = u_{ii} = u_{ll}$ because the antisymmetric part of $\partial_i u_k$ has zero trace. The deviatoric part is traceless by construction, so it changes no volumes to this order. □

Proposition 30.9 (Principal strains)**.** *At each point there is an orthonormal frame in which* u_{ik} *is diagonal. Its three diagonal entries* $u^{(1)}, u^{(2)}, u^{(3)}$ *— the* principal strains *— are the fractional extensions along the three corresponding* principal directions*, and* $u_{ll} = u^{(1)} + u^{(2)} + u^{(3)}$.

Rests on Definition 30.6 and Theorem 5.75.

Proof. *Derives Proposition 30.9.*

u_{ik} is a real symmetric array, hence the matrix of a self-adjoint operator on $\mathbb{R}^3$ with the Euclidean inner product; by Theorem 5.75 it has an orthonormal eigenbasis and real eigenvalues. In that basis a material line element along the j-th eigenvector of length $\mathrm{d}l$ becomes, to first order, one of length $\left(1 + u^{(j)}\right)\mathrm{d}l$, since the antisymmetric part contributes a rotation and not a length change. The trace is basis independent by Proposition 30.7. □

Example 30.10 (Simple shear and pure shear are the same strain)**.** Let $\boldsymbol{u} = (\gamma y, 0, 0)$ with $\gamma = 1 \times 10^{-3}$: the displacement gradient has the single entry $\partial_y u_x = \gamma$, so

$$u_{ik} = \begin{pmatrix} 0 & \gamma/2 & 0 \\ \gamma/2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \omega_{ik} = \begin{pmatrix} 0 & \gamma/2 & 0 \\ -\gamma/2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The principal strains are $\pm\gamma/2$ along the two directions at 45° to the axes, and 0 along z; the trace vanishes, so no volume changes. What distinguishes "simple" shear from "pure" shear is only the accompanying rigid rotation ω_{ik} of magnitude $\gamma/2 = 5 \times 10^{-4}$, and by Proposition 30.7 that rotation cannot enter any material law. A steel bar strained in shear by $\gamma = 1 \times 10^{-3}$ is therefore stressed exactly as one stretched by 5×10^{-4} along one diagonal and compressed by the same amount along the other.

Rests on Proposition 30.9 and Definition 30.6.

30.2.2 Finite strain and compatibility

Definition 30.6 was announced as valid "for small gradients". The exact statement is available and shows precisely what the truncation costs.

Definition 30.11 (Green strain)**.** The *Green (Lagrangian) strain tensor* of the motion Equation (30.2) is

$$E_{ik} = \tfrac{1}{2}\left(\partial_i u_k + \partial_k u_i + \partial_i u_l\, \partial_k u_l\right), \tag{30.7}$$

in which the indices are the material coordinates $\boldsymbol{X}$.

Rests on Equation (30.2) and Definition 30.6.

Proposition 30.12 (The Green strain measures length exactly)**.** *For every material line element* $\mathrm{d}\boldsymbol{X}$*,*

$$\mathrm{d}l'^2 - \mathrm{d}l^2 = 2E_{ik}\,\mathrm{d}X_i\,\mathrm{d}X_k\,, \tag{30.8}$$

with no approximation. Consequently E_{ik} *vanishes identically if and only if the motion is a rigid displacement.*

Rests on Definitions 13.118 and 30.11.

Proof. *Derives Proposition 30.12.*

With $F_{ik} = \partial_k \chi_i = \delta_{ik} + \partial_k u_i$ one has $\mathrm{d}x_i = F_{ik}\mathrm{d}X_k$, so

$$\mathrm{d}l'^2 = \mathrm{d}x_i \mathrm{d}x_i = F_{il}F_{ik}\,\mathrm{d}X_l \mathrm{d}X_k = (\delta_{lk} + \partial_l u_k + \partial_k u_l + \partial_l u_i \partial_k u_i)\,\mathrm{d}X_l \mathrm{d}X_k\,,$$

and subtracting $\mathrm{d}l^2 = \delta_{lk}\mathrm{d}X_l\mathrm{d}X_k$ gives Equation (30.8). If $E_{ik} \equiv 0$ then every material length is preserved, so the motion is an isometry of $\mathbb{R}^3$ and hence (Proposition 13.121) a rotation followed by a translation; the converse is immediate because a rigid motion has $F^{\mathsf{T}}F = \mathbb{1}$. □

Example 30.13 (A rigid rotation forges a strain of $-\theta^2/2$)**.** Rotate a body rigidly about z through θ. Then $\boldsymbol{u} = (R - \mathbb{1})\boldsymbol{X}$ with R the rotation matrix, and the linear strain of Equation (30.4) evaluates to

$$u_{xx} = u_{yy} = \cos\theta - 1 = -\tfrac{1}{2}\theta^2 + O(\theta^4)\,, \qquad u_{xy} = 0\,,$$

while $E_{ik} = 0$ exactly by Proposition 30.12. The linear measure reports a spurious contraction of θ^2 in volume where nothing has been deformed at all. For $\theta = 0.1\,\mathrm{rad}$ — under 6° — that is -5×10^{-3}, four times the elastic strain $\sigma_{\mathrm{y}}/E \approx 1.2 \times 10^{-3}$ at which structural steel yields (Phenomenon 30.73). The linear theory is therefore restricted not by the smallness of the strains but by the smallness of the *rotations*: a thin strip may bend far past $\theta = 0.1$ while every fibre in it stays elastic, and such a problem must be treated with Equation (30.7), or with the special geometrically nonlinear theories of Section 30.6.1.

Rests on Proposition 30.12 and Definition 30.6.

The second question about a strain field is the reverse one. Six functions u_{ik} are prescribed; do they come from any displacement at all? The system Equation (30.4) is six equations for three unknowns, so the answer is no in general, and the obstruction has a name.

Proposition 30.14 (Saint-Venant compatibility is necessary)**.** *If* u_{ik} *is the linear strain of a* C^3 *displacement field, then*

$$\partial_l\partial_k u_{ij} + \partial_i\partial_j u_{kl} - \partial_l\partial_j u_{ik} - \partial_i\partial_k u_{jl} = 0 \tag{30.9}$$

identically. In three dimensions Equation (30.9) has six independent components.

Rests on Definition 30.6 and Proposition 7.73.

Proof. *Derives Proposition 30.14.*

Substitute $u_{ij} = \tfrac{1}{2}(\partial_i u_j + \partial_j u_i)$ into the left-hand side. The first term contributes $\tfrac{1}{2}(\partial_l\partial_k\partial_i u_j + \partial_l\partial_k\partial_j u_i)$ and the other three the analogous expressions; every one of the eight resulting third derivatives appears twice with opposite signs once the partial derivatives are freely commuted, which Proposition 7.73 licenses for a C^3 field. The sum therefore vanishes term by term. The left-hand side is antisymmetric under $i \leftrightarrow l$ and under $j \leftrightarrow k$ and symmetric under the exchange of the pairs, which in three dimensions leaves 3×3 components of which the six symmetric ones are the independent conditions — exactly the index symmetries of a curvature tensor (Proposition 13.154). □

Theorem 30.15 (Compatibility is sufficient on a simply connected body)**.** *Let $\Omega \subset \mathbb{R}^3$ be open, connected and simply connected, and let u_{ik} be a symmetric C^2 field on Ω satisfying Equation (30.9). Then there is a displacement field $\boldsymbol{u}$ on Ω whose linear strain is u_{ik}, unique up to the addition of a rigid displacement $\boldsymbol{a} + \boldsymbol{b} \times \boldsymbol{x}$.*

Rests on Propositions 7.104 and 30.14.

Proof. *Derives Theorem 30.15.*

The construction is Cesàro's: recover first the rotation that the strain does not record, then the displacement. It uses no mathematics beyond Proposition 7.104(3), that a curl-free C^1 field on a simply connected open set is a gradient.

Step 1: the rotation field. Set

$$\Phi_{ijk} := \partial_i u_{jk} - \partial_j u_{ik}\,, \tag{30.10}$$

which is C^1 and antisymmetric in ij. Were u_{ik} already the strain of a displacement, Φ_{ijk} would be $\partial_k \omega_{ij}$ with $\omega_{ij} = \frac{1}{2}(\partial_i u_j - \partial_j u_i)$ the infinitesimal rotation: substituting $u_{jk} = \frac{1}{2}(\partial_j u_k + \partial_k u_j)$ into Equation (30.10) gives $\frac{1}{2}(\partial_i\partial_k u_j - \partial_j\partial_k u_i)$, which is that derivative. Take this as the equation to be solved. For each fixed pair ij the field $k \longmapsto \Phi_{ijk}$ is curl-free exactly when

$$\partial_l \Phi_{ijk} - \partial_k \Phi_{ijl} = \partial_l\partial_i u_{jk} - \partial_l\partial_j u_{ik} - \partial_k\partial_i u_{jl} + \partial_k\partial_j u_{il} \tag{30.11}$$

vanishes, and the right-hand side is the compatibility expression Equation (30.9) with its four indices taken in the order (i,k,l,j) and the sign reversed. It vanishes by hypothesis. By Proposition 7.104(3) on the simply connected Ω there are therefore functions ω_{ij} with $\partial_k \omega_{ij} = \Phi_{ijk}$, each fixed up to a constant. Since $\Phi_{ijk} = -\Phi_{jik}$, the constants may be chosen so that $\omega_{ij} = -\omega_{ji}$, and $\omega_{ii} = 0$.

Step 2: the displacement. For each fixed i consider $k \longmapsto u_{ik} - \omega_{ik}$, which is C^1. Its curl is

$$\begin{aligned} \partial_l\left(u_{ik} - \omega_{ik}\right) - \partial_k\left(u_{il} - \omega_{il}\right) &= \partial_l u_{ik} - \partial_k u_{il} - \Phi_{ikl} + \Phi_{ilk} \\ &= \partial_l u_{ik} - \partial_k u_{il} - (\partial_i u_{kl} - \partial_k u_{il}) \\ &\quad + (\partial_i u_{lk} - \partial_l u_{ik}) = 0\,, \end{aligned} \tag{30.12}$$

every term cancelling in pairs, the last two because $u_{kl} = u_{lk}$. Proposition 7.104(3) again supplies functions u_i with $\partial_k u_i = u_{ik} - \omega_{ik}$. Symmetrising and using the symmetry of u_{ik} with the antisymmetry of ω_{ik},

$$\tfrac{1}{2}\left(\partial_k u_i + \partial_i u_k\right) = \tfrac{1}{2}\left[(u_{ik} - \omega_{ik}) + (u_{ki} - \omega_{ki})\right] = u_{ik}\,, \tag{30.13}$$

so $\boldsymbol{u}$ has the prescribed strain.

Step 3: uniqueness. Let $\boldsymbol{v}$ have zero strain, $\partial_i v_k + \partial_k v_i = 0$. The identity of step 1, applied with $u_{ik} = 0$, gives $\partial_k w_{ij} = \Phi_{ijk} = 0$ for $w_{ij} = \frac{1}{2}(\partial_i v_j - \partial_j v_i)$, so w is a constant antisymmetric array on the connected Ω. The symmetric part of $\partial_j v_k$ vanishes by hypothesis, so $\partial_j v_k = w_{jk}$ is constant and $v_k = a_k + w_{jk}x_j$ with a_k constant; writing the antisymmetric w_{jk} as $\varepsilon_{jkm}b_m$ and using the cyclic symmetry of ε puts this in the form $\boldsymbol{v} = \boldsymbol{a} + \boldsymbol{b} \times \boldsymbol{x}$. Conversely every such field has zero strain. Two displacements with the same strain differ by one of them, which is the stated uniqueness. □

Remark 30.16 (Incompatibility is curvature). The index symmetries found in the proof of Proposition 30.14 are not a coincidence. The deformed body carries the metric $g_{ik} = \delta_{ik} + 2E_{ik}$ by Equation (30.8), and to first order in E the left-hand side of Equation (30.9) is -2 times the linearised Riemann tensor of that metric (Theorem 13.152). Compatibility says that the deformed body is still a piece of flat Euclidean space — that it can be laid down in $\mathbb{R}^3$ without tearing or overlapping. An incompatible strain field is one that cannot: it

describes a body with built-in stresses that no external load produces and no unloading removes. Crystal dislocations are the physical realisation, and their density is precisely the incompatibility Equation (30.9); the microscopic account belongs to Chapter 125.

Rests on Equation (30.9) and Theorem 13.152.

30.3 Stress: the Cauchy tensor

30.3.1 Traction and the existence of the stress tensor

Definition 30.17 (Traction)**.** Cut the body along a surface through the point $\boldsymbol{x}$ with unit normal $\boldsymbol{n}$. The material on the side into which $\boldsymbol{n}$ points exerts on the material on the other side a force $\Delta\boldsymbol{F}$ across an element of area ΔA of the cut. The *traction* is

$$\boldsymbol{t}(\boldsymbol{x},\boldsymbol{n}) = \lim_{\Delta A\to 0} \frac{\Delta \boldsymbol{F}}{\Delta A}\,, \tag{30.14}$$

an SI force per unit area, measured in $\mathrm{N/m^2} = \mathrm{Pa}$. It depends on the orientation of the cut as well as on the point.

Rests on Definitions 7.94 and 30.2.

That the traction depends on $\boldsymbol{n}$ at all is what distinguishes a solid from an ideal fluid, in which the traction is $-p\boldsymbol{n}$ for every $\boldsymbol{n}$. The content of Cauchy's theorem is that the dependence is as simple as it could possibly be.

Theorem 30.18 (Cauchy: the stress tensor exists)**.** *There is a field $\sigma_{ik}(\boldsymbol{x})$, a Cartesian tensor of rank two, such that for every unit vector $\boldsymbol{n}$*

$$t_i(\boldsymbol{x},\boldsymbol{n}) = \sigma_{ik}(\boldsymbol{x})\, n_k\,. \tag{30.15}$$

Its column σ_{ik} at fixed k is the traction across the coordinate plane with normal $\hat{\boldsymbol{e}}_k$.

Rests on Definitions 13.5 and 30.17.

Proof. *Derives Theorem 30.18.*

First, $\boldsymbol{t}(\boldsymbol{x},-\boldsymbol{n}) = -\boldsymbol{t}(\boldsymbol{x},\boldsymbol{n})$. Apply Newton's second law to a flat pillbox of face area A and thickness h straddling the cut: the mass and hence the inertia and the body force are $O(hA)$, while the two faces contribute $[\boldsymbol{t}(\boldsymbol{n}) + \boldsymbol{t}(-\boldsymbol{n})]\,A$ and the rim $O\left(hA^{1/2}\right)$. Dividing by A and letting $h \to 0$ leaves the stated relation.

Now take a tetrahedron with three faces in the coordinate planes, meeting at $\boldsymbol{x}$, and a fourth face of area A with outward unit normal $\boldsymbol{n}$ at perpendicular distance h from $\boldsymbol{x}$. The coordinate faces have areas An_k and outward normals $-\hat{\boldsymbol{e}}_k$, and the volume is $hA/3$. Newton's second law for the tetrahedron reads

$$\frac{\rho h A}{3}\,(\boldsymbol{a}-\boldsymbol{b}) = \boldsymbol{t}(\boldsymbol{n})\,A - \sum_k \boldsymbol{t}(\hat{\boldsymbol{e}}_k)\,An_k\,,$$

where $\boldsymbol{b}$ is the body force per unit mass and the mean values of the continuous fields over the faces have been used. Divide by A and let $h \to 0$: the left side vanishes, and

$$t_i(\boldsymbol{n}) = \sum_k t_i(\hat{\boldsymbol{e}}_k)\, n_k =: \sigma_{ik} n_k\,.$$

This is Equation (30.15). The nine numbers σ_{ik} so defined are a tensor by the quotient law of Section 13.2.2: t_i and n_k are vectors and $\boldsymbol{n}$ is arbitrary, so the array contracting one into the other must transform as a rank-two tensor. Note that no constitutive assumption of any kind has been made — the argument is pure balance of momentum, and Equation (30.15) therefore holds for fluids, plastic solids and viscoelastic media as well. □

Proposition 30.19 (The stress tensor is symmetric)**.** *In the absence of distributed body couples, $\sigma_{ik} = \sigma_{ki}$.*

Rests on Theorem 30.18 and Definition 30.17.

Proof. *Derives Proposition 30.19.*

Apply the balance of angular momentum about its own centre to a cube of side ℓ centred at $\boldsymbol{x}$. The torque about the z axis from the tractions on the four vertical faces is, to leading order, $(\sigma_{yx} - \sigma_{xy})\,\ell^3$; the torque from the body force is $O(\ell^4)$, because $\int (\boldsymbol{r} \times \boldsymbol{f}) \mathrm{d}V$ over a region of size ℓ centred on the reference point carries one factor of ℓ from $\boldsymbol{r}$ and three from $\mathrm{d}V$; and the moment of inertia is $\rho \ell^5/6$. Hence the angular acceleration is

$$\dot{\Omega}_z = \frac{6\,(\sigma_{yx} - \sigma_{xy})}{\rho \ell^2} + O\!\left(\ell^{-1}\right) ,$$

which diverges as $\ell \to 0$ unless $\sigma_{yx} = \sigma_{xy}$. The same argument on the other two axes completes the proof. The hypothesis is essential and is not vacuous: a medium carrying an intrinsic torque per unit volume — a magnetised body in a field (Chapter 129), a liquid crystal — adds a term of order ℓ^3 to the numerator and does support an antisymmetric part. □

Definition 30.20 (Pressure and deviatoric stress)**.** The *pressure* and the *deviatoric stress* are

$$p = -\tfrac{1}{3}\sigma_{ll} , \qquad s_{ik} = \sigma_{ik} + p\,\delta_{ik} , \tag{30.16}$$

so that $\sigma_{ik} = -p\,\delta_{ik} + s_{ik}$ with $s_{ll} = 0$. By Proposition 30.19 and Theorem 5.75 there is an orthonormal frame in which σ_{ik} is diagonal; its entries $\sigma_1 \geq \sigma_2 \geq \sigma_3$ are the *principal stresses*, and $p = -\left(\sigma_1 + \sigma_2 + \sigma_3\right)/3$.

Rests on Proposition 30.19 and Theorem 5.75.

The split is not cosmetic. Phenomenon 30.28 will show that p and s_{ik} are controlled by two independent material constants; Proposition 30.74 will show that plastic failure is governed by s_{ik} alone. A steel specimen at the bottom of the ocean, with $p = 110\,\mathrm{MPa}$ and $s_{ik} = 0$, is compressed by about 6.6×10^{-4} in volume and is in no danger whatever; the same material at $\sigma_1 - \sigma_3 = 500\,\mathrm{MPa}$ has already yielded.

30.3.2 Momentum balance

Theorem 30.21 (Cauchy's equation of motion)**.** *For a continuum of density ρ carrying a body force f_i per unit volume,*

$$\rho\,\partial_t^2 u_i = \partial_k \sigma_{ik} + f_i , \tag{30.17}$$

in the material description and for small displacements. Statics is the case $\partial_t^2 u_i = 0$, giving the equilibrium equations $\partial_k \sigma_{ik} + f_i = 0$.

Rests on Theorems 7.101 and 30.18.

Proof. *Derives Theorem 30.21.*

Let V be any fixed material region with boundary S. Newton's second law for the material in V balances the rate of change of its momentum against the surface tractions and the body force:

$$\int_V \rho\,\partial_t^2 u_i \,\mathrm{d}V = \oint_S t_i \,\mathrm{d}S + \int_V f_i \,\mathrm{d}V = \oint_S \sigma_{ik} n_k \,\mathrm{d}S + \int_V f_i \,\mathrm{d}V ,$$

using Equation (30.15). By the divergence theorem (Theorem 7.101) applied to each of the three vector fields $k \mapsto \sigma_{ik}$,

$$\int_V \left(\rho\,\partial_t^2 u_i - \partial_k \sigma_{ik} - f_i\right) \mathrm{d}V = 0 .$$

The region V was arbitrary and the integrand is continuous, so it vanishes pointwise — the argument of the fundamental lemma Lemma 16.18, applied here with arbitrary subregions rather than arbitrary test functions. □

Definition 30.22 (Boundary conditions of elastostatics)**.** The boundary $\partial\Omega$ is partitioned into a part $\partial\Omega_{\mathrm{u}}$ where the displacement is prescribed, $u_i = \bar{u}_i$, and a part $\partial\Omega_{\mathrm{t}}$ where the traction is prescribed, $\sigma_{ik}n_k = \bar{t}_i$. A free surface is the case $\bar{t}_i = 0$. A *kinematically admissible* field is any C^1 field agreeing with $\bar{u}_i$ on $\partial\Omega_{\mathrm{u}}$.

Rests on Theorem 30.21 and Definition 10.24.

Proposition 30.23 (Virtual work; the weak form)**.** *Let σ_{ik} satisfy equilibrium and the traction condition of Definition 30.22. Then for every field w_i vanishing on $\partial\Omega_{\mathrm{u}}$,*

$$\int_\Omega \sigma_{ik}\, w_{ik}\,\mathrm{d}V = \int_\Omega f_i w_i\,\mathrm{d}V + \int_{\partial\Omega_{\mathrm{t}}} \bar{t}_i w_i\,\mathrm{d}S\,, \tag{30.18}$$

with $w_{ik} = \tfrac{1}{2}(\partial_i w_k + \partial_k w_i)$; and conversely, a stress field satisfying Equation (30.18) for all such w_i is in equilibrium.

Rests on Theorem 30.21 and Definition 30.22.

Proof. *Derives Proposition 30.23.*

Because σ_{ik} is symmetric, $\sigma_{ik}\partial_k w_i = \sigma_{ik}w_{ik}$. Integrate the identity $\partial_k(\sigma_{ik}w_i) = (\partial_k\sigma_{ik})\,w_i + \sigma_{ik}\partial_k w_i$ over Ω and apply Theorem 7.101:

$$\oint_{\partial\Omega} \sigma_{ik}n_k w_i\,\mathrm{d}S = \int_\Omega (\partial_k\sigma_{ik})\,w_i\,\mathrm{d}V + \int_\Omega \sigma_{ik}w_{ik}\,\mathrm{d}V\,.$$

On $\partial\Omega_{\mathrm{u}}$ the field w_i vanishes and on $\partial\Omega_{\mathrm{t}}$ the traction is $\bar{t}_i$; substituting $\partial_k\sigma_{ik} = -f_i$ rearranges the display into Equation (30.18). For the converse, run the computation backwards: Equation (30.18) becomes $\int_\Omega (\partial_k\sigma_{ik} + f_i)\,w_i\mathrm{d}V = \int_{\partial\Omega_{\mathrm{t}}} (\bar{t}_i - \sigma_{ik}n_k)\,w_i\mathrm{d}S$ for all admissible w_i, and choosing w_i supported in the interior and then near the traction boundary kills the two brackets in turn by Lemma 16.18. □

Equation (30.18) is the principle of virtual work of Chapter 27 written for a continuum, and it is the form in which elasticity is actually solved: it asks only one derivative of the displacement instead of two, which is what makes the finite-element method and the Sobolev theory of Definition 10.85 and Theorem 10.88 applicable.

30.4 Linear elasticity: Hooke's law and the moduli

30.4.1 Hooke's law

Hooke published the law in 1676 as the anagram *ceiiinosssttuv* and unscrambled it two years later as *ut tensio, sic vis* — as the extension, so the force [13]. What he had was a statement about springs and wires; the tensorial statement below is what it becomes when the body is described by Definition 30.6 and Theorem 30.18.

Definition 30.24 (The elasticity tensor)**.** The *elastic energy density* $F(u_{ik})$ is the reversible work done per unit volume in bringing the body from its unstrained state to the strain u_{ik}, and the stress is

$$\sigma_{ik} = \frac{\partial F}{\partial u_{ik}}\,. \tag{30.19}$$

The *elasticity tensor* is $C_{iklm} = \partial^2 F/\partial u_{ik}\partial u_{lm}$ evaluated at $u_{ik} = 0$, a rank-four Cartesian tensor with the SI dimension of pressure, Pa.

Rests on Definition 30.6 and Theorem 30.18.

Proposition 30.25 (Twenty-one constants)**.** *C_{iklm} obeys the* minor *symmetries $C_{iklm} = C_{kilm} = C_{ikml}$, which follow from the symmetry of u_{ik} and of σ_{ik}, and the* major *symmetry $C_{iklm} = C_{lmik}$, which follows from the existence of the energy F. A general anisotropic solid therefore has* 21 *independent elastic constants.*

Rests on Definition 30.24 and Proposition 7.73.

Proof. *Derives Proposition 30.25.*

Only the part of C_{iklm} symmetric in (ik) and in (lm) can contribute to $C_{iklm}u_{ik}u_{lm}$, since u_{ik} is symmetric; this is the minor symmetry, and it reduces each index pair to the 6 values $\{11, 22, 33, 23, 13, 12\}$. The major symmetry is $\partial^2 F/\partial u_{ik}\partial u_{lm} = \partial^2 F/\partial u_{lm}\partial u_{ik}$, which is Proposition 7.73 for the smooth function F; without an energy function there would be no reason for it. The array is therefore a symmetric 6×6 matrix, with $6 \times 7/2 = 21$ independent entries. □

Lemma 30.26 (Isotropic Cartesian tensors of rank four)**.** *Every rank-four Cartesian tensor invariant under all of* SO(3) *is a linear combination*

$$T_{iklm} = a\,\delta_{ik}\delta_{lm} + b\,\delta_{il}\delta_{km} + c\,\delta_{im}\delta_{kl} \tag{30.20}$$

of the three products of Kronecker deltas, with a, b, c scalars.

Rests on Theorem 5.128 and Definition 13.5.

Full derivation: Appendix A.56 (p. 3100).

Derives Lemma 30.26.

The classification of the isotropic Cartesian tensors of ranks two, three and four is invariant theory for the rotation group — general mathematics rather than elasticity — and it is proved in Part II, for ranks one to four together, as Theorem 5.128; the ranks are settled in one theorem because they share the same machinery, not because the higher needs the lower. The rank-four case is restated here in the notation of this chapter because it is the one step that turns Proposition 30.25's twenty-one constants into two; the appendix section additionally carries the cubic case, which Part II does not treat.

Remark 30.27 (Two results this chapter borrows from Part II). Two pieces of general mathematics are used below and proved nowhere in this chapter, so they are named here rather than passed over. Both were written into Part II in August 2026 after this chapter reported them missing, and the entries stand as a record of what this chapter rests on. *First*, the classification just stated: the isotropic Cartesian tensors of rank two, three and four are spanned by δ_{ik}, by ε_{ikl} and by the three products Equation (30.20) respectively. It is invariant theory for SO(3), and it is Theorem 5.128, with Section 13.2.2 recording what it does for the tensor algebra. *Second*, the Helmholtz decomposition — that a vector field on a suitable domain splits into a gradient and a curl — which is Theorem 7.105. This chapter does *not* use it: the potential representation of Proposition 30.48 is introduced as an ansatz and verified, which proves that the Rayleigh wave exists and is all that proposition claims. The decomposition would be needed only to prove that *every* plane-strain motion is of that form, a completeness statement made nowhere in this chapter.

Rests on Theorems 5.128 and 7.105.

Phenomenon 30.28 (Hooke's law)**.** For sufficiently small deformations the response of a solid is linear and reversible: the elongation of a loaded spring or bar is proportional to the load and is recovered in full when the load is removed — "ut tensio, sic vis" [13]. In continuum terms the stress is a linear function of the strain, and for an isotropic body it depends on the strain through two material constants alone,

$$\sigma_{ik} = \lambda\,u_{ll}\,\delta_{ik} + 2\mu\,u_{ik} \tag{30.21}$$

[138, 1778]. In structural metals the linear regime extends to strains of order 10^{-3} and ends at the yield point of Phenomenon 30.73.

Rests on Definition 30.24 and Lemma 30.26.

Derivation. *Derives Phenomenon 30.28.*

The undeformed state is an equilibrium and carries no stress, so the expansion of the energy density about $u_{ik} = 0$ has neither a constant nor a linear term:

$$F = \tfrac{1}{2}\, C_{iklm}\, u_{ik} u_{lm} + O(u^3)\,, \tag{30.22}$$

with the constant coefficients of Definition 30.24. By Equation (30.19) the stress is then linear in the strain up to the cubic remainder, $\sigma_{ik} = C_{iklm} u_{lm}$. That is Hooke's law, and its observed domain of validity is precisely the range of strains over which the remainder is negligible — which is why the law is a small-strain statement and not a property of any particular material.

If the body is isotropic, C_{iklm} is invariant under every rotation of the frame, so by Lemma 30.26 it has the form Equation (30.20). The minor symmetries of Proposition 30.25 force $b = c$, since exchanging i and k in Equation (30.20) exchanges the coefficients of $\delta_{il}\delta_{km}$ and $\delta_{im}\delta_{kl}$. Writing $a = \lambda$ and $b = c = \mu$ and contracting with u_{lm} gives $\sigma_{ik} = \lambda u_{ll}\delta_{ik} + 2\mu u_{ik}$, which is Equation (30.21). That two constants suffice for an isotropic solid is a consequence of symmetry, not a further approximation [1778]. □

Remark 30.29 (The Lamé constants and the energy). λ and μ are the *Lamé constants* [138], both measured in Pa. Substituting Equation (30.21) into Equation (30.19) and integrating gives the energy density, which in terms of the split Equation (30.5) reads

$$F = \tfrac{1}{2}\lambda\,(u_{ll})^2 + \mu\, u_{ik} u_{ik} = \tfrac{1}{2} K\,(u_{ll})^2 + \mu\, d_{ik} d_{ik}\,, \qquad K = \lambda + \tfrac{2}{3}\mu\,. \tag{30.23}$$

The second form is the useful one: it separates the energy of a pure volume change from that of a pure shape change, and shows that the two are stored independently. μ is the *shear modulus* and K the *bulk modulus*.

Rests on Equations (30.5) and (30.21).

Proposition 30.30 (Stability bounds)**.** *The unstrained state is a strict minimum of the elastic energy if and only if*

$$\mu > 0 \qquad \text{and} \qquad K = \lambda + \tfrac{2}{3}\mu > 0\,. \tag{30.24}$$

Rests on Equation (30.23) and Theorem 5.112.

Proof. *Derives Proposition 30.30.*

The second form of Equation (30.23) is a sum of two squares in variables that are independent — the trace u_{ll} and the five components of the traceless d_{ik} — so the quadratic form is positive definite exactly when both coefficients are positive, which is Theorem 5.112 read on a form already in normal shape. Each condition is realised by an admissible strain: a pure dilatation $u_{ik} = \epsilon\delta_{ik}$ has $F = \frac{9}{2} K\epsilon^2$ and no shear, and a pure shear as in Example 30.10 has $F = \mu\gamma^2/2$ and no dilatation. A body violating either bound would gain energy by deforming and would not be found in the unstrained state at all. □

30.4.2 The elastic moduli and their measurement

The two constants of Equation (30.21) may be traded for any other independent pair, and the laboratory has good reasons to prefer other pairs: λ corresponds to no experiment anyone performs, whereas the modulus that a tensile machine measures does.

Definition 30.31 (The engineering moduli)**.** For an isotropic solid define

- the *bulk modulus* K by $p = -K\, u_{ll}$ under hydrostatic loading;
- the *shear modulus* μ by $s_{ik} = 2\mu\, d_{ik}$;

- *Young's modulus* E [53] by $\sigma = E\,u_{\parallel}$ under uniaxial stress;
- *Poisson's ratio* ν [86] by $\nu = -u_{\perp}/u_{\parallel}$ under uniaxial stress;
- the *P-wave modulus* $M = \lambda + 2\mu$, the stiffness under uniaxial *strain*.

All but ν are pressures, measured in Pa; ν is dimensionless.
Rests on Equation (30.21) *and Definition 30.20.*

Phenomenon 30.32 (Poisson contraction)**.** A bar stretched along its axis contracts in the two transverse directions. Under uniaxial stress the ratio of transverse contraction to longitudinal extension is a material constant, Poisson's ratio ν [86], and the axial strain is itself proportional to the applied stress, the constant of proportionality being Young's modulus E [53]. The measured value of ν for every ordinary solid lies inside the stability window

$$-1 < \nu < \tfrac{1}{2}\,, \tag{30.25}$$

clustering near $\nu \approx 0.3$ for structural metals and approaching the upper, incompressible end for materials whose shear modulus is small against their bulk modulus [1455].
Rests on Phenomenon 30.28 and Proposition 30.30.

Derivation. *Derives Phenomenon 30.32.*

Take the bar along z and impose the uniaxial stress $\sigma_{zz} = \sigma$ with all other stress components zero. By isotropy the strain is diagonal, with $u_{xx} = u_{yy} \equiv u_{\perp}$ and $u_{zz} \equiv u_{\parallel}$. The transverse components of Equation (30.21) read $0 = \lambda(2u_{\perp} + u_{\parallel}) + 2\mu u_{\perp}$, so

$$\nu \equiv -\frac{u_{\perp}}{u_{\parallel}} = \frac{\lambda}{2(\lambda+\mu)}\,, \tag{30.26}$$

which is positive whenever $\lambda > 0$: the transverse strain has the sign opposite to the axial one, as observed. Substituting back into the axial component gives $\sigma = \mu(3\lambda + 2\mu)u_{\parallel}/(\lambda+\mu)$, so that $E = \mu(3\lambda+2\mu)/(\lambda+\mu)$ — the modulus is not an independent constant but a combination of the two in Equation (30.21). For the bounds, use Proposition 30.30: $\mu > 0$ and $K = \lambda + \frac{2}{3}\mu > 0$. Writing $x = \mu/K$, so that x ranges over $(0, \infty)$, Equation (30.26) becomes

$$\nu = \frac{3 - 2x}{2(3+x)}\,, \tag{30.27}$$

a strictly decreasing function of x with limits $\frac{1}{2}$ as $x \to 0$ and -1 as $x \to \infty$. Its range is therefore the open interval of Equation (30.25). A solid far harder to compress than to shear sits at the top of that interval: it is effectively incompressible and contracts transversely by half of what it extends. □

Proposition 30.33 (Conversion among the moduli)**.** *Any two of* $(\lambda, \mu, K, E, \nu, M)$ *determine the rest. In particular*

$$\mu = \frac{E}{2(1+\nu)}\,, \qquad K = \frac{E}{3(1-2\nu)}\,, \qquad \lambda = \frac{2\mu\nu}{1-2\nu}\,, \tag{30.28}$$

$$E = \frac{9K\mu}{3K+\mu}\,, \qquad \nu = \frac{3K-2\mu}{2(3K+\mu)}\,, \qquad M = K + \tfrac{4}{3}\mu\,. \tag{30.29}$$

Rests on Phenomenon 30.32 and Definition 30.31.

Proof. *Derives Proposition 30.33.*

The derivation of Phenomenon 30.32 gave $\nu = \lambda/\left[2(\lambda+\mu)\right]$ and $E = \mu(3\lambda+2\mu)/(\lambda+\mu)$. Solving the first for λ gives $\lambda(1-2\nu) = 2\mu\nu$, the third relation of Equation (30.28); substituting it into the second gives $E = \mu\left[3\cdot 2\mu\nu/(1-2\nu) + 2\mu\right] / \left[2\mu\nu/(1-2\nu) + \mu\right] = 2\mu(1+\nu)$, the first. Then $K = \lambda + \frac{2}{3}\mu = 2\mu\nu/(1-2\nu) + \frac{2}{3}\mu = \frac{2}{3}\mu(1+\nu)/(1-2\nu) =$

$E/[3(1-2\nu)]$, the second. Eliminating ν between the first two of Equation (30.28) gives $3K(1-2\nu) = 2\mu(1+\nu)$, hence $\nu = (3K-2\mu)/[2(3K+\mu)]$ and, back-substituting, $E = 2\mu(1+\nu) = 9K\mu/(3K+\mu)$. Finally $M = \lambda + 2\mu = \left(K - \tfrac{2}{3}\mu\right) + 2\mu = K + \tfrac{4}{3}\mu$. □

One further consequence of the quadratic energy is needed before the measured values, because it is what turns an assumed deformation into a rigorous *bound* on a modulus.

Proposition 30.34 (Minimum of the potential energy)**.** *Among all kinematically admissible displacement fields, the equilibrium field — and only it, up to a rigid displacement compatible with the constraints — minimises*

$$\Pi[\boldsymbol{v}] = \int_\Omega F\left(v_{ik}\right) \mathrm{d}V - \int_\Omega f_i v_i \,\mathrm{d}V - \int_{\partial\Omega_\mathrm{t}} \bar{t}_i v_i \,\mathrm{d}S\,. \tag{30.30}$$

Rests on Propositions 30.23 and 30.30.

Proof. *Derives Proposition 30.34.*

Let $\boldsymbol{u}$ be the equilibrium field and $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{w}$ any admissible competitor, so that $\boldsymbol{w}$ vanishes on $\partial\Omega_\mathrm{u}$. Because F is the quadratic form Equation (30.22), $F(u_{ik}+w_{ik}) = F(u_{ik}) + \sigma_{ik}w_{ik} + F(w_{ik})$ exactly, the cross term being $C_{iklm}u_{lm}w_{ik} = \sigma_{ik}w_{ik}$ by the major symmetry of Proposition 30.25. Hence

$$\Pi[\boldsymbol{v}] - \Pi[\boldsymbol{u}] = \int_\Omega \sigma_{ik}w_{ik}\,\mathrm{d}V - \int_\Omega f_i w_i \,\mathrm{d}V - \int_{\partial\Omega_\mathrm{t}} \bar{t}_i w_i \,\mathrm{d}S + \int_\Omega F(w_{ik})\,\mathrm{d}V\,.$$

The first three terms cancel by Proposition 30.23, and the last is non-negative by Proposition 30.30, vanishing only if $w_{ik} \equiv 0$, i.e. only if $\boldsymbol{w}$ is a rigid displacement (Proposition 30.12). This is the Dirichlet principle of Proposition 10.91 for the elastic operator, and the same convexity underwrites the direct method Theorem 16.70. □

Table 30.1: Measured room-temperature elastic constants of five cubic single crystals, in the Voigt notation of Definition 30.37, from the compilation of Simmons and Wang [1455]. The spread between independent determinations of these quantities is typically 1 %, which is the uncertainty to attach to every entry. $A = 2C_{44}/(C_{11}-C_{12})$ is the anisotropy factor, equal to 1 for an isotropic solid; the last column tests the Cauchy relation $C_{12} = C_{44}$ of Proposition 30.38, which holds only for the ionic crystal.

Crystal	C_{11} (GPa)	C_{12} (GPa)	C_{44} (GPa)	A	C_{12}/C_{44}
Aluminium	106.8	60.4	28.3	1.22	2.13
Copper	168.4	121.4	75.4	3.21	1.61
Iron	231.4	134.7	116.4	2.41	1.16
Sodium chloride	49.1	12.8	12.8	0.71	1.00
Diamond	1076	125	577	1.21	0.22

Example 30.35 (From single crystals to the isotropic aggregate)**.** A polycrystal is a mass of randomly oriented anisotropic grains and is isotropic on the average, so it must be described by two constants built from the three of Table 30.1. For a cubic crystal the bulk modulus is exact and orientation independent, $K = (C_{11}+2C_{12})/3$. The shear modulus is not, and is bracketed by two estimates: assuming a uniform *strain* throughout the aggregate gives the Voigt value $\mu_\mathrm{V} = (C_{11}-C_{12}+3C_{44})/5$, and assuming a uniform *stress* gives the Reuss value $\mu_\mathrm{R} = 5(C_{11}-C_{12})C_{44}/[4C_{44}+3(C_{11}-C_{12})]$. A uniform strain is a kinematically admissible field but not the equilibrium one, so by Proposition 30.34 it stores more energy

Table 30.2: Isotropic (polycrystalline aggregate) moduli computed from Table 30.1 by the averaging of Example 30.35, with ν and E then following from Equation (30.29). The propagated uncertainty is about 1 % on K, μ and E and about 0.01 on ν, inherited from the single-crystal data; the residual disagreement with handbook values for the corresponding polycrystals is of the same size.

Material	K (GPa)	μ (GPa)	E (GPa)	ν
Aluminium	75.9	26.1	70.4	0.345
Copper	137.1	47.3	127.3	0.345
Iron	166.9	81.8	210.9	0.289
Sodium chloride	24.9	14.7	36.8	0.253
Diamond	442	534	1142	0.069

than the true field does at the same overall deformation: the aggregate looks stiffer than it is, and μ_{V} is a rigorous upper bound. μ_{R} is the corresponding uniform-stress estimate; it falls below μ_{V} for every entry of Table 30.1, and the dual variational principle — minimum complementary energy at fixed stress, which this chapter states without proving — is what makes that ordering general. Their arithmetic mean is a convention, not a theorem, and it is what Table 30.2 reports.

For iron the arithmetic is $K = (231.4 + 269.4)/3 = 166.9\,\mathrm{GPa}$, $\mu_{\mathrm{V}} = (96.7 + 349.2)/5 = 89.2\,\mathrm{GPa}$, $\mu_{\mathrm{R}} = 5(96.7)(116.4)/(465.6 + 290.1) = 74.5\,\mathrm{GPa}$, mean $81.8\,\mathrm{GPa}$; then Equation (30.29) gives $E = 211\,\mathrm{GPa}$ and $\nu = 0.289$. Those are the values quoted for structural steel in every engineering handbook, obtained here from three ultrasonic measurements on a single crystal of pure iron and one averaging assumption. The width of the Voigt–Reuss bracket, 18 % of the mean, is a direct measure of the anisotropy $A = 2.41$; for aluminium, with $A = 1.22$, the same bracket is under 1 % wide, and the Hill convention costs nothing there.

Rests on Propositions 30.33 and 30.34.

Remark 30.36 (How the moduli are actually measured). Three routes, of increasing precision, all traceable to the SI metre, second and kilogram (Chapter 2).

- *Static loading.* Hang a known force on a bar of known cross-section and measure the extension. A steel bar at $\sigma = 200\,\mathrm{MPa}$ — a working stress — stretches by $u_{\parallel} = \sigma/E = 9.5 \times 10^{-4}$, about one part in a thousand, so a metre of bar lengthens by 0.95 mm. That is easy to see and hard to measure to better than 1 %, and the method is further limited by grip effects and by creep.

- *Resonance.* Excite a bar or plate and count its natural frequencies; the moduli follow from the eigenvalue relations of Propositions 30.57 and 30.66. Counting is the one laboratory operation done without error, so this is immediately two orders of magnitude better.

- *Ultrasonic pulse-echo.* Time a pulse across a specimen of measured thickness and obtain v_{P} and v_{S} directly; then by Equation (30.35)

$$\mu = \rho\, v_{\mathrm{S}}^2\,, \qquad \lambda = \rho\left(v_{\mathrm{P}}^2 - 2v_{\mathrm{S}}^2\right). \tag{30.31}$$

  This needs one length, two times and a density, is insensitive to the specimen's shape, and works on a crystal small enough that no static test could grip it — which is why the tabulated single-crystal constants of Table 30.1 are almost all ultrasonic.

Rests on Equation (30.35) and Proposition 30.33.

30.4.3 Anisotropic elasticity

Definition 30.37 (Voigt notation)**.** Map each symmetric index pair to a single index by $11 \mapsto 1$, $22 \mapsto 2$, $33 \mapsto 3$, $23 \mapsto 4$, $13 \mapsto 5$, $12 \mapsto 6$, and write $C_{\alpha\beta}$ for the resulting symmetric 6×6 matrix of Proposition 30.25 [364]. The stress components map directly, σ_α; the strain components carry a factor two on the shear entries, $u_4 = 2u_{23}$, $u_5 = 2u_{13}$, $u_6 = 2u_{12}$, so that $\sigma_\alpha = C_{\alpha\beta} u_\beta$ and $F = \tfrac{1}{2} C_{\alpha\beta} u_\alpha u_\beta$ both hold with no stray factors.

Rests on Proposition 30.25 and Definition 30.24.

Crystal symmetry cuts the 21 constants down. Each symmetry operation of the crystal class must leave $C_{\alpha\beta}$ invariant, and the resulting linear conditions leave 13 constants for the monoclinic classes, 9 for the orthorhombic, 6 or 7 for the tetragonal and trigonal, 5 for the hexagonal, 3 for the cubic — C_{11}, C_{12} and C_{44}, the constants tabulated in Table 30.1 — and 2 for the isotropic limit, where in addition $C_{11} = \lambda + 2\mu$, $C_{12} = \lambda$ and $C_{44} = \mu$, so that $A = 2C_{44}/(C_{11} - C_{12}) = 1$ [364]. The measured A is thus a direct readout of how far a crystal is from isotropy, and Table 30.1 shows it ranging from 0.71 for sodium chloride to 3.21 for copper: a copper single crystal is more than three times stiffer in shear on one set of planes than on another, which is why a textured copper sheet is not an isotropic body.

Proposition 30.38 (The Cauchy relations)**.** *Suppose the atoms of a crystal sit at centres of symmetry of the lattice and interact through central pair potentials $\varphi(r)$, and that the crystal is under no external stress. Then C_{iklm} is symmetric under* every *permutation of its four indices, so that six further relations hold and only* 15 *constants remain. For a cubic crystal the single relation is*

$$C_{12} = C_{44}\,. \tag{30.32}$$

Rests on Definition 30.24 and Proposition 30.25.

Proof. *Derives Proposition 30.38.*

Write $\eta_b = u_{ik} n_i n_k$ for the bond of length r_b and direction $\hat{\boldsymbol{n}}$. By Equation (30.8) the strained length is $r_b' = r_b\,(1 + 2\eta_b)^{1/2} = r_b \left(1 + \eta_b - \tfrac{1}{2}\eta_b^2 + \cdots\right)$. Because every atom sits at a centre of symmetry the bonds come in antiparallel pairs, so no internal relaxation of the sublattices is induced by a homogeneous strain and the energy per unit volume is simply $\sum_b \varphi(r_b')$ over the bonds of one cell divided by its volume. Expanding,

$$\varphi(r_b') = \varphi(r_b) + \varphi'(r_b)\, r_b\, \eta_b + \tfrac{1}{2}\left[\varphi''(r_b) - \frac{\varphi'(r_b)}{r_b}\right] r_b^2\, \eta_b^2 + O(\eta^3)\,.$$

The coefficient of the term linear in u_{ik} is the residual stress, which vanishes by the zero-stress hypothesis. The quadratic term gives

$$C_{iklm} = \sum_b w_b\, n_i n_k n_l n_m\,, \qquad w_b = \frac{1}{V}\left[\varphi''(r_b) - \frac{\varphi'(r_b)}{r_b}\right] r_b^2\,,$$

since $\eta_b^2 = u_{ik} u_{lm} n_i n_k n_l n_m$; the weights w_b depend on the bond length alone, and their exact normalisation — how the bond sum is counted — does not matter for what follows. Every summand is a product of four components of one and the same unit vector, hence invariant under any permutation of $iklm$; so is C_{iklm}. In particular $C_{1122} = C_{1212}$, which in the notation of Definition 30.37 reads $C_{12} = C_{66}$, and $C_{66} = C_{44}$ for a cubic crystal: that is Equation (30.32). The other five relations are the remaining independent identifications among the 21 constants. □

Remark 30.39 (The rari-constant controversy, decided by measurement). Proposition 30.38 was the ground of a nineteenth-century dispute. Navier and Poisson built elasticity on central forces between molecules and so predicted 15 constants, and 1 for an isotropic body

— with $\lambda = \mu$ and therefore $\nu = \frac{1}{4}$ exactly [80, 86]. Green and Lamé built it on the energy alone and predicted 21 and 2 [138]. The question is empirical, and Table 30.1 settles it: the last column should be 1.00 for every crystal if the Cauchy relations held, and it is 1.00 only for sodium chloride, running to 2.13 for aluminium and 0.22 for diamond. Voigt's systematic measurements across the crystal classes are what closed the matter in favour of the multi-constant theory [364]. The physical reading is now clear: an ionic crystal is held together by very nearly central forces between closed shells, so it nearly obeys the relations; a metal is held together by a delocalised electron gas and diamond by directional covalent bonds, and neither is a sum of central pair potentials. That is a statement about the electronic structure, and it is settled in Chapter 126; the elastic constants of Table 30.1 are among the sharpest classical constraints on it, and they reappear as the long-wavelength limit of the lattice dynamics of Chapter 125.

Rests on Proposition 30.38 and Definition 30.37.

30.5 Elastodynamics and elastic waves

30.5.1 The Navier–Cauchy equations

Theorem 30.40 (Navier–Cauchy equations)**.** *A homogeneous isotropic linearly elastic body obeys*

$$\rho\,\partial_t^2 \boldsymbol{u} = (\lambda+\mu)\,\boldsymbol{\nabla}\,(\boldsymbol{\nabla}\cdot\boldsymbol{u}) + \mu\,\nabla^2\boldsymbol{u} + \boldsymbol{f}\,. \tag{30.33}$$

Rests on Theorem 30.21 and Equation (30.21).

Proof. *Derives Theorem 30.40.*

Substitute Equation (30.21), with $u_{ik} = \frac{1}{2}(\partial_i u_k + \partial_k u_i)$, into Cauchy's equation Equation (30.17). Since λ and μ are constants,

$$\partial_k \sigma_{ik} = \lambda\,\partial_i(\boldsymbol{\nabla}\cdot\boldsymbol{u}) + \mu\,\partial_k(\partial_i u_k + \partial_k u_i) = (\lambda+\mu)\,\partial_i(\boldsymbol{\nabla}\cdot\boldsymbol{u}) + \mu\,\nabla^2 u_i\,,$$

where $\partial_k\partial_i u_k = \partial_i\partial_k u_k = \partial_i(\boldsymbol{\nabla}\cdot\boldsymbol{u})$ by Proposition 7.73. This is Equation (30.33). Navier reached the same equation from a molecular model in which the constants were not independent [80]; Cauchy reached it from the stress tensor, in which they are [73], and Remark 30.39 says which of the two the measurements support. □

Proposition 30.41 (Energy balance and its flux)**.** *Define the energy density and the energy flux*

$$e = \tfrac{1}{2}\rho\,\dot{u}_i\dot{u}_i + F\,(u_{ik})\;, \qquad q_k = -\sigma_{ik}\dot{u}_i\,. \tag{30.34}$$

Then $\partial_t e + \partial_k q_k = f_i\dot{u}_i$: elastic energy is locally conserved, and it is transported at the rate at which the internal tractions do work.

Rests on Theorem 30.40 and Equation (30.23).

Proof. *Derives Proposition 30.41.*

Differentiate: $\partial_t e = \rho\dot{u}_i\ddot{u}_i + \sigma_{ik}\dot{u}_{ik}$, using Equation (30.19) for the second term. Because σ_{ik} is symmetric, $\sigma_{ik}\dot{u}_{ik} = \sigma_{ik}\partial_k\dot{u}_i$. Substituting $\rho\ddot{u}_i = \partial_k\sigma_{ik} + f_i$ from Equation (30.17),

$$\partial_t e = \dot{u}_i\partial_k\sigma_{ik} + \sigma_{ik}\partial_k\dot{u}_i + f_i\dot{u}_i = \partial_k(\sigma_{ik}\dot{u}_i) + f_i\dot{u}_i\,,$$

which is the stated balance with q_k as defined. □

Theorem 30.42 (Uniqueness in elastodynamics)**.** *The initial-boundary-value problem for Equation (30.33) — with $\boldsymbol{u}$ and $\dot{\boldsymbol{u}}$ given at $t = 0$, and with either the displacement or the traction prescribed at every point of $\partial\Omega$ at every time — has at most one solution, up to a time-independent rigid displacement when no displacement is prescribed anywhere.*

Rests on Propositions 30.30 and 30.41.

Proof. *Derives Theorem 30.42.*

Let $\boldsymbol{w}$ be the difference of two solutions. It solves Equation (30.33) with $\boldsymbol{f} = 0$, with zero initial data, and with $\boldsymbol{w} = 0$ or $\sigma_{ik}[\boldsymbol{w}]n_k = 0$ at each boundary point. By Proposition 30.41 its total energy $\mathcal{E}(t) = \int_\Omega e \, \mathrm{d}V$ satisfies

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} = -\oint_{\partial\Omega} q_k n_k \, \mathrm{d}S = \oint_{\partial\Omega} \sigma_{ik} n_k \, \dot{w}_i \, \mathrm{d}S = 0 \,,$$

by Theorem 7.101, since on each part of the boundary one factor of the integrand vanishes — the traction where the traction is prescribed, the velocity where the displacement is. Hence $\mathcal{E}(t) = \mathcal{E}(0) = 0$ for all t. But $\mathcal{E}$ is an integral of a sum of two non-negative terms, the second positive definite by Proposition 30.30, so $\dot{\boldsymbol{w}} \equiv 0$ and $w_{ik} \equiv 0$; by Proposition 30.12 $\boldsymbol{w}$ is a rigid displacement, constant in time, and it vanishes if the displacement is prescribed anywhere. This is what makes the problem well posed in the sense of Definition 10.24, and the same energy identity gives the continuous dependence on the data. □

30.5.2 P and S waves

Phenomenon 30.43 (Two bulk wave speeds)**.** An isotropic elastic solid transmits exactly two kinds of bulk wave: a longitudinal, compressional wave (the P phase of seismology) and a transverse, shear wave (the S phase), travelling at the distinct speeds

$$v_\mathrm{P} = \sqrt{\frac{\lambda + 2\mu}{\rho}} \,, \qquad v_\mathrm{S} = \sqrt{\frac{\mu}{\rho}} \,, \tag{30.35}$$

with $v_\mathrm{P} > v_\mathrm{S}$ in every material [86]. The P arrival therefore always precedes the S arrival on a seismogram, and their separation measures the distance to the source [293]. A fluid, which cannot support a static shear, has $\mu = 0$ and carries no transverse branch at all.

Rests on Theorem 30.40 and Proposition 30.30.

Derivation. *Derives Phenomenon 30.43.*

Look for plane-wave solutions of the source-free Equation (30.33), $\boldsymbol{u} = \boldsymbol{A} \exp\left[\mathrm{i}\,(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)\right]$, with $\boldsymbol{A}$ a constant polarization vector. Each spatial derivative brings down $\mathrm{i}k_j$ and each time derivative $-\mathrm{i}\omega$, so $\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{u}) \mapsto -(\boldsymbol{k}\cdot\boldsymbol{A})\,\boldsymbol{k}$ and $\nabla^2\boldsymbol{u} \mapsto -k^2\boldsymbol{A}$, giving

$$\rho\,\omega^2 A_i = \Gamma_{il} A_l \,, \qquad \Gamma_{il} = (\lambda + \mu)\, k_i k_l + \mu k^2 \delta_{il} \,. \tag{30.36}$$

A plane wave exists exactly when $\boldsymbol{A}$ is an eigenvector of the *acoustic tensor* Γ_{il}, with $\rho\omega^2$ the eigenvalue. Γ is real and symmetric, so Theorem 5.75 supplies an orthonormal eigenbasis, and it can be read off: $\hat{\boldsymbol{k}}$ is an eigenvector with eigenvalue $(\lambda+\mu)k^2 + \mu k^2 = (\lambda + 2\mu)k^2$, and every vector orthogonal to $\boldsymbol{k}$ is an eigenvector with eigenvalue μk^2. Those three exhaust $\mathbb{R}^3$, so there are no other polarizations. Hence

$$\omega^2 = \frac{\lambda + 2\mu}{\rho}\, k^2 \quad \text{for } \boldsymbol{A} \parallel \boldsymbol{k} \,, \qquad \omega^2 = \frac{\mu}{\rho}\, k^2 \quad \text{for } \boldsymbol{A} \perp \boldsymbol{k} \,,$$

which are the two speeds Equation (30.35). Both branches are non-dispersive: ω is proportional to k, so a pulse travels undistorted, which is what makes an arrival time a meaningful observable at all.

The physical labels follow from the polarizations. The longitudinal branch has $\boldsymbol{\nabla}\times\boldsymbol{u} = 0$ and carries the dilatation $\boldsymbol{\nabla}\cdot\boldsymbol{u}$; the transverse branch has $\boldsymbol{\nabla}\cdot\boldsymbol{u} = 0$ and carries the rotation, deforming without changing any volume. Finally

$$v_\mathrm{P}^2 - v_\mathrm{S}^2 = \frac{\lambda + \mu}{\rho} = \frac{K + \frac{1}{3}\mu}{\rho} > 0$$

by Proposition 30.30: the ordering of the two speeds is not an accident of the materials that happen to exist but follows from the existence of a stable unstrained state. Setting $\mu = 0$ annihilates the transverse branch and leaves $v_{\mathrm{P}} = \sqrt{K/\rho}$, the sound speed of Chapter 31. □

Proposition 30.44 (The speed ratio is fixed by Poisson's ratio)**.**

$$\frac{v_{\mathrm{P}}}{v_{\mathrm{S}}} = \sqrt{\frac{2(1-\nu)}{1-2\nu}}\,, \tag{30.37}$$

a strictly increasing function of ν on Equation (30.25) with range $\left(2/\sqrt{3}, \infty\right)$. No elastic solid whatever has $v_{\mathrm{P}}/v_{\mathrm{S}} \le 2/\sqrt{3} = 1.1547$, and a measurement of the ratio determines ν without any knowledge of the density.

Rests on Equation (30.35) and Proposition 30.33.

Proof. *Derives Proposition 30.44.*

From Equation (30.35), $v_{\mathrm{P}}^2/v_{\mathrm{S}}^2 = (\lambda + 2\mu)/\mu = \lambda/\mu + 2$, and $\lambda/\mu = 2\nu/(1-2\nu)$ by Equation (30.28); adding gives $(2\nu + 2 - 4\nu)/(1-2\nu) = 2(1-\nu)/(1-2\nu)$, which is Equation (30.37). Differentiating the square, $\mathrm{d}\left[2(1-\nu)/(1-2\nu)\right]/\mathrm{d}\nu = 2/(1-2\nu)^2 > 0$, so it increases; its values at the endpoints $\nu = -1$ and $\nu \to \frac{1}{2}$ are $4/3$ and $+\infty$. The density cancels between the two speeds, which is the practical point: an ultrasonic apparatus that measures only times and one length returns ν directly. □

Example 30.45 (Elastic waves in steel, and the distance to a source)**.** Take the aggregate iron of Table 30.2, $E = 211\,\mathrm{GPa}$, $\nu = 0.289$, with $\rho = 7850\,\mathrm{kg/m^3}$. Then $\mu = E/\left[2(1+\nu)\right] = 81.8\,\mathrm{GPa}$ and $M = \lambda + 2\mu = K + \frac{4}{3}\mu = 276\,\mathrm{GPa}$, so

$$v_{\mathrm{P}} = \sqrt{\frac{2.76 \times 10^{11}\,\mathrm{Pa}}{7850\,\mathrm{kg/m^3}}} = 5.93 \times 10^3\,\mathrm{m/s}\,, \qquad v_{\mathrm{S}} = 3.23 \times 10^3\,\mathrm{m/s}\,,$$

in a ratio 1.84, which Equation (30.37) returns from $\nu = 0.289$ alone. These are the handbook speeds for steel, and the agreement is a closed loop: three ultrasonic constants of a single crystal in Table 30.1, averaged in Example 30.35, reproduce the ultrasonic speeds of the polycrystal.

The separation of the two arrivals then gives the distance. A source at range d sends its P and S phases along the same path, so

$$\Delta t = d\left(\frac{1}{v_{\mathrm{S}}} - \frac{1}{v_{\mathrm{P}}}\right)\,, \tag{30.38}$$

and for the steel above $\Delta t/d = 1.41 \times 10^{-4}\,\mathrm{s/m}$: a delay of one millisecond places the source 7.1 m away. The same arithmetic on a planetary scale, with the phases identified on a seismogram written thousands of kilometres from the epicentre, is Oldham's method and the foundation of seismology [293].

Rests on Proposition 30.44 and Equation (30.35).

Remark 30.46 (Why the plane-wave route, and not the potentials). The derivation of Phenomenon 30.43 was carried out on plane waves, where the completeness of the two polarizations is the spectral theorem for the 3×3 matrix Γ_{il} and nothing more. The usual textbook route instead splits an *arbitrary* displacement field as $\boldsymbol{u} = \boldsymbol{\nabla}\phi + \boldsymbol{\nabla} \times \boldsymbol{\psi}$ with $\boldsymbol{\nabla} \cdot \boldsymbol{\psi} = 0$ and shows that ϕ and $\boldsymbol{\psi}$ separately obey wave equations with the speeds Equation (30.35). That route needs the Helmholtz decomposition theorem — existence and uniqueness of the split, with the conditions at infinity that make it unique — which is general vector analysis and is Theorem 7.105, proved in Appendix A.45. Either route is therefore available. The plane-wave argument is kept because it is self-contained and

because it is what the seismological data are read against; the same theorem is what the potentials of Proposition 30.48 would need if that proposition claimed completeness, which it does not — see Remark 30.27.

Rests on Phenomenon 30.43 and Theorems 7.105 and 30.40.

30.5.3 Surface waves and the seismological evidence

Phenomenon 30.47 (The three seismic phases)**.** A seismogram written far from an earthquake shows three groups of arrivals in a fixed order: a first, small compressional motion, a second and larger shear motion, and last of all a long, slowly decaying train of large-amplitude waves whose period varies through the train [293]. The third group is confined to the neighbourhood of the free surface, travels more slowly than either bulk wave of Phenomenon 30.43, and decays with depth over a distance of the order of a wavelength; its two polarizations are the vertically polarized Rayleigh wave, which exists on a homogeneous half-space [232], and the horizontally polarized Love wave, which requires a surface layer of lower shear speed over a faster substrate [376].

Rests on Phenomenon 30.43 and Proposition 30.49.

Derivation. *Derives Phenomenon 30.47.*

The order of the first two groups is Equation (30.38) with $v_{\mathrm{P}} > v_{\mathrm{S}}$, established in Phenomenon 30.43. That the third group comes last is Proposition 30.49 and Proposition 30.50: both surface branches travel below v_{S}, the Rayleigh wave at $0.919\,v_{\mathrm{S}}$ for a Poisson solid and the Love wave between the two shear speeds of the layered structure. That the third group is the *largest* at great range, though the source radiates mostly body waves, is geometric. A body wave from a point source spreads over a hemisphere of area $2\pi r^2$, so its energy flux falls as r^{-2} and its amplitude as r^{-1}; a surface wave is confined to a skin of fixed thickness of order one wavelength and spreads over a cylinder of area $2\pi r \times \text{const}$, so its flux falls as r^{-1} and its amplitude only as $r^{-1/2}$. At $r = 1000\,\mathrm{km}$ from a source of characteristic size $10\,\mathrm{km}$ the surface wave is therefore favoured by a factor $(r/r_0)^{1/2} = 10$ in amplitude over the body waves, which is why it dominates a distant record.

That the train is long and its period varies through it — the observed *dispersion* — is Proposition 30.50: the Love speed depends on $\omega H/\beta$, so different periods arrive at different times and a single impulsive source is drawn out into a train ordered in period. A homogeneous half-space would give a Rayleigh wave with no dispersion at all (Proposition 30.48 contains no length), so the observed dispersion is itself the evidence that the Earth is layered. Finally, the confinement to the surface: both branches are built from solutions decaying as e^{-qz} with q of order k, so the motion is negligible below a depth of about one wavelength — $35\,\mathrm{km}$ for a $10\,\mathrm{s}$ Love wave travelling at $3.5\,\mathrm{km/s}$. □

Proposition 30.48 (Rayleigh's secular equation)**.** *On the traction-free surface of a homogeneous isotropic half-space there is a wave whose displacement is confined to a plane containing the propagation direction and the normal, decays exponentially with depth, and travels at a speed $c = \xi v_{\mathrm{S}}$ where ξ solves*

$$\left(2-\xi^2\right)^2 = 4\sqrt{1-\xi^2}\,\sqrt{1-\gamma\xi^2}\,, \qquad \gamma = \frac{v_{\mathrm{S}}^2}{v_{\mathrm{P}}^2} = \frac{1-2\nu}{2-2\nu}\,, \tag{30.39}$$

with $0 < \xi < 1$ [232]. The equation contains no length, so the wave is non-dispersive.

Rests on Theorem 30.40 and Definition 30.22.

Proof. *Derives Proposition 30.48.*

Let z measure depth from the free surface, so that the solid occupies $z > 0$, let the propagation be along x, and seek a motion with $u_y = 0$ and no dependence on y. Build it from two potentials,

$$u_x = \partial_x\phi + \partial_z\psi\,, \qquad u_z = \partial_z\phi - \partial_x\psi\,. \tag{30.40}$$

This is an ansatz, not a decomposition theorem, and it is all the proposition needs: it asserts that such a wave *exists*, so it is enough to exhibit one. Since $\boldsymbol{\nabla}\cdot\boldsymbol{u}=\nabla^2\phi$, substituting Equation (30.40) into the source-free Equation (30.33) and collecting gives

$$\partial_x\left[\rho\partial_t^2\phi-(\lambda+2\mu)\nabla^2\phi\right]+\partial_z\left[\rho\partial_t^2\psi-\mu\nabla^2\psi\right]=0$$

for the x component and the same two brackets differentiated by ∂_z and $-\partial_x$ for the z component, so it suffices that each bracket vanish: ϕ obeys the wave equation at v_{P} and ψ at v_{S}, the two speeds Equation (30.35).

Take the surface wave

$$\phi=A\,\mathrm{e}^{-pz}\,\mathrm{e}^{\mathrm{i}(kx-\omega t)}\,,\qquad \psi=B\,\mathrm{e}^{-qz}\,\mathrm{e}^{\mathrm{i}(kx-\omega t)}\,, \tag{30.41}$$

with $k>0$. The two wave equations fix $p^2=k^2-\omega^2/v_{\mathrm{P}}^2$ and $q^2=k^2-\omega^2/v_{\mathrm{S}}^2$, so writing $c=\omega/k=\xi v_{\mathrm{S}}$,

$$p=k\sqrt{1-\gamma\xi^2}\,,\qquad q=k\sqrt{1-\xi^2}\,, \tag{30.42}$$

both real and positive — so that the motion decays with depth over a distance of order $1/k$ — precisely when $0<\xi<1$, since $\gamma<1$ by $v_{\mathrm{S}}<v_{\mathrm{P}}$ (Phenomenon 30.43).

The free surface $z=0$ carries no traction (Definition 30.22), which for this motion is the two conditions $\sigma_{zz}=\sigma_{xz}=0$. From Equation (30.21),

$$\sigma_{zz}=\lambda\nabla^2\phi+2\mu\left(\partial_z^2\phi-\partial_x\partial_z\psi\right)\,, \tag{30.43}$$

$$\sigma_{xz}=\mu\left[2\partial_x\partial_z\phi+\left(\partial_z^2-\partial_x^2\right)\psi\right]\,. \tag{30.44}$$

Insert Equation (30.41). In the first, the coefficient of A is $\lambda\left(p^2-k^2\right)+2\mu p^2=(\lambda+2\mu)p^2-\lambda k^2=2\mu k^2-\rho\omega^2=\mu k^2\left(2-\xi^2\right)$, using $\rho\omega^2=\mu k^2\xi^2$; in the second, the coefficient of B is $\mu\left(q^2+k^2\right)=\mu k^2\left(2-\xi^2\right)$, the same combination. The cross terms are $2\mathrm{i}\mu kq$ in the first and $-2\mathrm{i}\mu kp$ in the second. Dividing throughout by μ, the two conditions at $z=0$ are therefore

$$k^2\left(2-\xi^2\right)A+2\mathrm{i}kq\,B=0\,,\qquad -2\mathrm{i}kp\,A+k^2\left(2-\xi^2\right)B=0\,, \tag{30.45}$$

a homogeneous pair with a non-trivial solution only if its determinant $k^4\left(2-\xi^2\right)^2-4k^2pq$ vanishes. Dividing by k^4 and substituting Equation (30.42) gives Equation (30.39).

It remains to know that an admissible root exists. Put $t=\xi^2$ and $f(t)=(2-t)^2-4\sqrt{(1-t)\,(1-\gamma t)}$. Then $f(0)=0$ and $f'(0)=-4+2\,(1+\gamma)=2\,(\gamma-1)<0$, so f is negative just to the right of the origin, while $f(1)=1>0$; by the intermediate value theorem (Theorem 7.23) f has a root in $(0,1)$, and $\xi=\sqrt{t}$ is the Rayleigh ratio. Finally, no length enters Equation (30.39): ξ depends on γ alone, so $c=\xi v_{\mathrm{S}}$ is independent of k and the wave does not disperse. The depth of penetration, by contrast, is $1/p$ and $1/q$, both of order the wavelength. □

Proposition 30.49 (The Rayleigh speed of a Poisson solid)**.** *For $\nu=\frac{1}{4}$ — the case $\lambda=\mu$, $\gamma=1/3$ — Equation (30.39) has exactly one admissible root, and*

$$\xi^2=2-\frac{2}{\sqrt{3}}\,,\qquad \xi=0.919\,40\,,\qquad c_{\mathrm{R}}=0.9194\,v_{\mathrm{S}}\,. \tag{30.46}$$

Rests on Proposition 30.48.

Proof. *Derives Proposition 30.49.*

Put $t=\xi^2$ and $\gamma=1/3$ and square Equation (30.39):

$$(2-t)^4=16\,(1-t)\left(1-\tfrac{1}{3}t\right)\,.$$

Expanding, $(2-t)^4 = t^4 - 8t^3 + 24t^2 - 32t + 16$ and the right side is $16 - \frac{64}{3}t + \frac{16}{3}t^2$, so the difference is $t^4 - 8t^3 + \frac{56}{3}t^2 - \frac{32}{3}t = \frac{t}{3}\,(3t^3 - 24t^2 + 56t - 32)$. The root $t = 0$ is the trivial static solution. The cubic has $t = 4$ as a root by inspection, and dividing out leaves $3t^2 - 12t + 8 = 0$ with roots $t = 2 \pm 2/\sqrt{3}$. Admissibility requires $0 < t < 1$, since both radicals in Equation (30.39) must be real for the depth dependence to decay: this excludes $t = 4$ and $t = 2 + 2/\sqrt{3} = 3.1547$ and leaves $t = 2 - 2/\sqrt{3} = 0.845\,299$. Squaring can create roots, so verify in the unsquared equation: $(2-t)^2 = 1.333\,33$ and $4\sqrt{1-t}\sqrt{1-t/3} = 4(0.393\,320)(0.847\,487) = 1.333\,33$, so it is genuine. Hence $\xi = \sqrt{0.845\,299} = 0.919\,40$. The Rayleigh wave is thus slower than the S wave by 8 %, which is what puts it last on the seismogram of Phenomenon 30.47, and its speed is a fixed multiple of v_{S} with no free length — so on a homogeneous half-space it does not disperse. □

Proposition 30.50 (Love waves need a slow surface layer)**.** *Let a layer of thickness H, density ρ_1 and shear modulus μ_1 rest on a half-space with ρ_2, μ_2, and write $\beta_j = \sqrt{\mu_j/\rho_j}$. Horizontally polarized surface waves of angular frequency ω and speed $c = \omega/k$ exist if and only if $\beta_1 < c < \beta_2$ and*

$$\tan\left(kH\sqrt{\frac{c^2}{\beta_1^2} - 1}\right) = \frac{\mu_2\sqrt{1 - c^2/\beta_2^2}}{\mu_1\sqrt{c^2/\beta_1^2 - 1}}\,. \tag{30.47}$$

In particular no such wave exists unless $\beta_1 < \beta_2$, and the speed depends on ω: the Love wave is dispersive [376].

Rests on Theorem 30.40 and Definition 30.22.

Proof. *Derives Proposition 30.50.*

Let z measure depth from the free surface, and look for a motion purely along y, $\boldsymbol{u} = (0, v, 0)$ with $v = f(z)\exp[\mathrm{i}(kx - \omega t)]$. Such a field has $\boldsymbol{\nabla}\cdot\boldsymbol{u} = 0$, so Equation (30.33) reduces in each homogeneous medium to $\rho\,\partial_t^2 v = \mu\nabla^2 v$, and hence to

$$f'' + \left(\frac{\omega^2}{\beta^2} - k^2\right) f = 0\,.$$

In the layer take $c > \beta_1$, so $s^2 = \omega^2/\beta_1^2 - k^2 > 0$ and $f = A\cos sz + B\sin sz$. The free surface carries the traction $\sigma_{yz} = \mu_1\partial_z v$, which must vanish at $z = 0$: hence $B = 0$ and $f = A\cos sz$. In the half-space take $c < \beta_2$, so $q^2 = k^2 - \omega^2/\beta_2^2 > 0$ and the only solution bounded at depth is $f = C\mathrm{e}^{-q(z-H)}$. Continuity of the displacement and of the traction $\mu\partial_z v$ across $z = H$ gives

$$A\cos sH = C\,, \qquad -\mu_1 As\sin sH = -\mu_2 qC\,,$$

and dividing the second by the first, $\mu_1 s\tan sH = \mu_2 q$, which is Equation (30.47) once s and q are written out. If $\beta_1 \geq \beta_2$ the two requirements $c > \beta_1$ and $c < \beta_2$ are incompatible and no solution exists — a homogeneous half-space, in particular, carries no Love wave, which is why Love had to introduce the layer [376]. Finally k and ω enter Equation (30.47) separately and not only through c, so c depends on frequency; as $\omega \to \infty$ the wave concentrates in the layer and $c \to \beta_1$, while near the cut-off of each higher branch $c \to \beta_2$. □

30.6 Rods, beams and plates

A rod, a beam, a plate and a shell are not new physics: they are Equation (30.33) on a domain with one or two small dimensions, reduced by exploiting that smallness. The reduction replaces a three-dimensional problem by a one- or two-dimensional one and turns the elastic energy into a functional of a single function, after which the machinery of Chapter 16 does the rest.

30.6.1 The elastica and Euler buckling

Proposition 30.51 (Moment and curvature)**.** *Let a slender rod of cross-section S lie along x and bend in the xz plane with small deflection $w(x)$. If plane cross-sections remain plane and normal to the deformed axis, and if the only non-negligible stress is the axial one, then the fibre at height z carries the strain and stress*

$$u_{xx} = -z\,w''(x)\,, \qquad \sigma_{xx} = -Ez\,w''(x)\,, \tag{30.48}$$

and the internal bending moment about the neutral axis is

$$M = EI\,w''(x)\,, \qquad I = \int_S z^2\,\mathrm{d}S\,, \tag{30.49}$$

with I the second moment of area, of SI dimension m^4*.*
Rests on Equation (30.21) *and Definition 30.31.*

Proof. *Derives Proposition 30.51.*

Under the kinematic hypothesis, a cross-section at x rotates through the angle $w'(x)$, so the axial displacement of the fibre at height z is $u_x = -z\,w'(x)$ and its axial strain is $u_{xx} = \partial_x u_x = -z\,w''$. The lateral faces of the rod are free, so $\sigma_{yy} = \sigma_{zz} = 0$ and the state is uniaxial *stress* — whence $\sigma_{xx} = E\,u_{xx}$ by Definition 30.31, with E and not M appearing. The moment of these stresses about the axis $z = 0$ is $-\int_S \sigma_{xx} z\,\mathrm{d}S = Ew'' \int_S z^2 \mathrm{d}S = EIw''$. The axis $z = 0$ is fixed by requiring zero net axial force, $\int_S \sigma_{xx}\mathrm{d}S = 0$, i.e. $\int_S z\,\mathrm{d}S = 0$: the neutral axis passes through the centroid. □

Phenomenon 30.52 (Euler buckling)**.** A slender straight column under a slowly increased axial compression remains straight, shortening elastically, until at a definite critical load it bows sideways; the deflection appears suddenly and the column carries no more load. The critical load does not depend on the strength of the material but on its stiffness, and falls as the inverse square of the length: for a column of length L, Young's modulus E and area moment of inertia I, pinned at both ends,

$$P_{\mathrm{crit}} = \frac{\pi^2 EI}{L^2}\,. \tag{30.50}$$

Doubling the length of a column therefore quarters the load it can bear [23].
Rests on Proposition 30.51 and Theorem 16.22.

Derivation. *Derives Phenomenon 30.52.*

Let the column occupy $0 \le x \le L$ under the axial compression P and let $y(x)$ be a small lateral deflection, with the ends pinned so that $y(0) = y(L) = 0$ and no moment is transmitted there. By Proposition 30.51 the bending moment at x is $EI\,y''(x)$, and the moment of the applied load about the same section is $-P\,y(x)$; equilibrium of the deflected shape therefore requires

$$EI\,y'' + P\,y = 0\,, \qquad y(0) = y(L) = 0\,. \tag{30.51}$$

This is an eigenvalue problem. Writing $k^2 = P/(EI)$, the general solution is $y = A\sin kx + B\cos kx$; the condition at $x = 0$ forces $B = 0$, and the condition at $x = L$ then requires either $A = 0$ — the straight column, available at every load — or $\sin kL = 0$, i.e. $kL = n\pi$ with n a positive integer. A deflected equilibrium therefore exists only at the discrete loads $P_n = n^2\pi^2 EI/L^2$, and the smallest of them, $n = 1$, is Equation (30.50). Below it the straight column is the only equilibrium; at it a one-parameter family of bent shapes appears, which is what is seen as the sudden onset of buckling. Note that the material enters only through E and the cross-section only through I: the yield stress of Phenomenon 30.73 plays no part, so a slender column fails long before its material does. □

Example 30.53 (A steel rod buckles at a twentieth of its crushing load)**.** Take a circular steel rod of diameter 10 mm and length 1 m, pinned at both ends, with $E = 210\,\mathrm{GPa}$. Its area moment is $I = \pi d^4/64 = 4.91 \times 10^{-10}\,\mathrm{m}^4$ and its area $S = 7.85 \times 10^{-5}\,\mathrm{m}^2$, so Equation (30.50) gives

$$P_{\mathrm{crit}} = \frac{\pi^2 \left(2.10 \times 10^{11}\,\mathrm{Pa}\right)\left(4.91 \times 10^{-10}\,\mathrm{m}^4\right)}{\left(1\,\mathrm{m}\right)^2} = 1.02\,\mathrm{kN}\,,$$

an axial stress of only 13.0 MPa. Crushing the same rod would need $\sigma_{\mathrm{y}} S = 19.6\,\mathrm{kN}$ at a yield stress of 250 MPa: buckling arrives at a nineteenth of that. The dimensionless control parameter is the slenderness L/r with $r = \sqrt{I/S} = d/4 = 2.5\,\mathrm{mm}$ the radius of gyration, here $L/r = 400$, and Equation (30.50) in those terms is $\sigma_{\mathrm{crit}} = \pi^2 E/(L/r)^2$. Yield and buckling change places at $L/r = \pi\sqrt{E/\sigma_{\mathrm{y}}} = 91$: a rod stubbier than that is crushed, a slenderer one buckles. This is the single most consequential formula in structural engineering, and it was the first instability ever computed.

Rests on Phenomenon 30.52 and Equation (30.50)*.*

Remark 30.54 (Other end conditions, and what buckling is variationally)*.* Clamping the ends changes only the length of the half-wave that must fit: Equation (30.50) holds with L replaced by an effective length $L_{\mathrm{eff}} = \kappa L$, with $\kappa = 1$ pinned at both ends, $\kappa = 2$ for a cantilever fixed at one end and free at the other, $\kappa = \frac{1}{2}$ clamped at both, and $\kappa = 0.699$ clamped at one and pinned at the other. The load capacity therefore varies by a factor 16 between the extremes at fixed L.

Read variationally, buckling is the loss of positive-definiteness of a second variation. The energy of the compressed column is

$$\Pi[y] = \int_0^L \left[\tfrac{1}{2} EI\,(y'')^2 - \tfrac{1}{2} P\,(y')^2\right] \mathrm{d}x\,,$$

the second term being the work done by P as the bent column shortens. This is exactly the accessory functional of Proposition 16.50, and Equation (30.51) is its Jacobi equation (Definition 16.53); $P = P_{\mathrm{crit}}$ is the load at which $x = L$ first becomes conjugate to $x = 0$, and Theorem 16.55 is the statement that the straight column ceases to be a minimum there. The two subjects — the sufficiency theory of the calculus of variations and the stability of structures — are the same theory.

Rests on Phenomenon 30.52 and Theorem 16.55.

Remark 30.55 (Beyond the critical load: the elastica)*.* Equation (30.51) is the small-deflection limit. Euler solved the exact problem in the appendix on elastic curves of the *Methodus inveniendi* [23]: parametrising the rod by arc length s and letting $\theta(s)$ be the angle of the tangent, the exact statement that the curvature is proportional to the moment is

$$EI\,\frac{\mathrm{d}^2\theta}{\mathrm{d}s^2} + P\sin\theta = 0\,, \tag{30.52}$$

which is the pendulum equation of Chapter 28 with s in place of time. Its solutions are elliptic functions, and the amplitude of the buckled shape grows as $\theta_{\max} \propto (P/P_{\mathrm{crit}} - 1)^{1/2}$ just above threshold — a supercritical pitchfork bifurcation (Chapter 32). The linear theory predicts the threshold exactly and the amplitude not at all, which is the generic situation for a bifurcation and is why Equation (30.50) is safe to design with and useless for predicting how far a buckled column bends.

Rests on Phenomenon 30.52 and Equation (30.7)*.*

30.6.2 Euler–Bernoulli beam theory

Theorem 30.56 (The beam equation)**.** *A slender rod of bending stiffness EI, mass per unit length ρS, carrying a transverse load $q(x)$ per unit length, has transverse deflection $w(x,t)$ obeying*

$$\rho S\,\partial_t^2 w + \partial_x^2\left(EI\,\partial_x^2 w\right) = q\,. \tag{30.53}$$

At a free end the natural *boundary conditions $EIw'' = 0$ and $(EIw'')' = 0$ hold — vanishing moment and vanishing shear force.*

Rests on Proposition 30.51 and Theorem 16.38.

Proof. *Derives Theorem 30.56.*

By Equation (30.48) the elastic energy density is $\tfrac{1}{2}E\,(zw'')^2$; integrating over the cross-section gives $\tfrac{1}{2}EI\,(w'')^2$ per unit length, so the action of the rod is $\iint \mathcal{L}\,\mathrm{d}x\,\mathrm{d}t$ with Lagrangian density

$$\mathcal{L} = \tfrac{1}{2}\rho S\,(\partial_t w)^2 - \tfrac{1}{2}EI\left(\partial_x^2 w\right)^2 + qw\,.$$

The density depends on a second derivative, so the stationarity condition is the Euler–Poisson equation Theorem 16.38,

$$\frac{\partial \mathcal{L}}{\partial w} - \partial_x\frac{\partial \mathcal{L}}{\partial(\partial_x w)} + \partial_x^2\frac{\partial \mathcal{L}}{\partial(\partial_x^2 w)} - \partial_t\frac{\partial \mathcal{L}}{\partial(\partial_t w)} = 0\,,$$

which evaluates to $q - \partial_x^2\,(EIw'') - \rho S\partial_t^2 w = 0$, i.e. Equation (30.53). The boundary terms produced by the integration by parts are $[EIw''\,\delta w']_0^L$ and $-\left[(EIw'')'\,\delta w\right]_0^L$; where the end is free, δw and $\delta w'$ are unconstrained and each bracket must vanish separately, which is Theorem 16.31. Where the end is clamped, δw and $\delta w'$ vanish and the conditions are $w = w' = 0$ instead — two conditions at each end in every case, as a fourth-order equation requires. □

Proposition 30.57 (Bending waves are dispersive)**.** *Free waves on a uniform beam obey*

$$\omega = k^2\sqrt{\frac{EI}{\rho S}}\,, \tag{30.54}$$

so the phase velocity ω/k grows without bound with k and the group velocity is twice the phase velocity. The natural frequencies of a beam of length L are $\omega_n = (\beta_n L)^2\sqrt{EI/(\rho S L^4)}$ with $\beta_n L$ fixed by the end conditions; for a cantilever the roots of $\cos\beta L\cosh\beta L + 1 = 0$ give $\beta_1 L = 1.8751$.

Rests on Theorem 30.56 and Definition 9.2.

Proof. *Derives Proposition 30.57.*

Substitute $w = \exp[\mathrm{i}\,(kx - \omega t)]$ into the source-free Equation (30.53) with EI constant: $-\rho S\omega^2 + EIk^4 = 0$, which is Equation (30.54). Then $v_{\mathrm{ph}} = \omega/k = k\sqrt{EI/\rho S}$ and $v_{\mathrm{g}} = \mathrm{d}\omega/\mathrm{d}k = 2k\sqrt{EI/\rho S} = 2v_{\mathrm{ph}}$. For a finite beam, separating $w = W(x)\mathrm{e}^{-\mathrm{i}\omega t}$ gives $W'''' = \beta^4 W$ with $\beta^4 = \rho S\omega^2/(EI)$, whose general solution is a combination of $\sin\beta x, \cos\beta x, \sinh\beta x, \cosh\beta x$; imposing $W = W' = 0$ at the clamped end and $W'' = W''' = 0$ at the free one gives a 4×4 homogeneous system whose determinant reduces to $\cos\beta L\cosh\beta L + 1 = 0$, and its smallest positive root is $\beta_1 L = 1.8751$. □

Example 30.58 (A steel ruler)**.** Clamp a steel strip 20 mm wide and 2 mm thick with 300 mm projecting. Then $I = bh^3/12 = 1.33\times 10^{-11}\,\mathrm{m}^4$, $EI = 2.80\,\mathrm{N\,m^2}$ and $\rho S = 0.314\,\mathrm{kg/m}$, so

$$f_1 = \frac{(1.8751)^2}{2\pi}\sqrt{\frac{2.80\,\mathrm{N\,m^2}}{(0.314\,\mathrm{kg/m})\,(0.3\,\mathrm{m})^4}} = 18.6\,\mathrm{Hz}\,,$$

which is what a twanged ruler does. Note the scaling that Proposition 30.57 encodes: $f_1 \propto h/L^2$ at fixed material, since $\sqrt{EI/\rho S} \propto h$ for a rectangle. Halving the overhang quadruples the pitch — two octaves — and this is the measurement route to E of Remark 30.36, since f_1 can be counted to 1×10^{-4} while an extension of a millimetre cannot.

Rests on Proposition 30.57 and Theorem 30.56.

Remark 30.59 (Where Euler–Bernoulli fails). Equation (30.54) is unphysical at large k: the phase velocity grows without limit and eventually exceeds v_{P}, which no disturbance in an elastic solid can do. The reason is that the theory neglects the shear deformation of the cross-section and its rotary inertia, and both matter once the wavelength approaches the depth h. Comparing Equation (30.54) with the shear speed for a rectangular section, $\sqrt{EI/\rho S} = h\sqrt{E/\rho}/\sqrt{12}$, the two agree when $kh = \sqrt{12\mu/E} = \sqrt{6/(1+\nu)} \approx 2.2$: the theory degrades once the wavelength falls below about three times the beam depth. Timoshenko's correction restores both neglected effects and gives a dispersion relation with two branches, both with bounded phase velocity [482]. This is the general pattern of a reduced theory: the reduction assumes a separation of scales, and it fails exactly where the separation does — as in Equation (30.1), one level up.

Rests on Proposition 30.57 and Equation (30.35).

30.6.3 Torsion

Phenomenon 30.60 (Coulomb's law of torsion)**.** A thin metal wire clamped at one end and twisted at the other exerts a restoring torque proportional to the angle of twist, and independent of the angle's history provided the twist is small. Coulomb measured the constant of proportionality and found it proportional to the fourth power of the diameter and inversely proportional to the length,

$$\tau = -\kappa\,\Phi\,, \qquad \kappa = \frac{\pi \mu\, d^4}{32\, L}\,, \tag{30.55}$$

with κ measured in $\mathrm{N\,m/rad}$ [36]. A body of moment of inertia I_0 suspended on such a fibre oscillates with period $T = 2\pi\sqrt{I_0/\kappa}$, which is how κ is measured and how the fibre becomes an instrument.

Rests on Equation (30.21) and Definition 30.31.

Derivation. Derives Phenomenon 30.60.

Take a circular cylinder of radius $a = d/2$ along z and twist it uniformly, so that the section at height z rotates through $\theta' z$ with $\theta' = \Phi/L$ the twist per unit length. In cylindrical coordinates the displacement is $u_\phi = \theta' z r$, $u_r = u_z = 0$, whose only non-vanishing strain component is $u_{\phi z} = \tfrac{1}{2}\theta' r$; the deformation is a pure shear, so no volume changes and only μ can enter. By Equation (30.21) the stress is $\sigma_{\phi z} = 2\mu u_{\phi z} = \mu\theta' r$, and the torque transmitted across a section is

$$\tau = \int_S \sigma_{\phi z}\, r\, \mathrm{d}S = \mu\theta' \int_S r^2 \mathrm{d}S = \mu\theta' J\,, \qquad J = \frac{\pi a^4}{2} = \frac{\pi d^4}{32}\,. \tag{30.56}$$

Setting $\theta' = \Phi/L$ gives Equation (30.55). The d^4 is the content of Coulomb's measurement: a factor d^2 because the sheared area grows so, and a second factor d^2 because the lever arm and the strain each grow with r. The restoring torque is linear in Φ because Hooke's law is linear in the strain, so the torsional pendulum is harmonic and its period is amplitude independent, giving $T = 2\pi\sqrt{I_0/\kappa}$ from Chapter 28. □

Example 30.61 (A torsion fibre as a force sensor)**.** A fused-silica fibre of diameter $50\,\mu\mathrm{m}$ and length $1\,\mathrm{m}$, with shear modulus $\mu = 31\,\mathrm{GPa}$, has

$$\kappa = \frac{\pi\,(3.1\times10^{10}\,\mathrm{Pa})\,(5.0\times10^{-5}\,\mathrm{m})^4}{32\,(1\,\mathrm{m})} = 1.90\times10^{-8}\,\mathrm{N\,m/rad}\,.$$

A bar carrying $I_0 = 1.0 \times 10^{-4}\,\mathrm{kg\,m^2}$ then swings with period $T = 2\pi\sqrt{I_0/\kappa} = 456\,\mathrm{s}$, about 7.6 min. Because κ is read from that period and from a moment of inertia — a mass and a length — the instrument measures torque on an absolute scale, and a deflection of 1 mrad, easily resolved optically, corresponds to $1.9 \times 10^{-11}\,\mathrm{N\,m}$. That sensitivity, and nothing more exotic, is what makes the Cavendish measurement of Chapter 33 possible, and the same fibre reappears in every torsion-balance test of the equivalence principle.

Rests on Phenomenon 30.60 and Equation (30.55).

Remark 30.62 (Saint-Venant's theory: non-circular sections warp). Equation (30.56) used the circular symmetry twice: to assert that plane sections stay plane and to evaluate $\int r^2 \mathrm{d}S$. Neither survives for a general cross-section. Saint-Venant's solution adds an axial displacement $u_z = \theta' \psi(x, y)$ — the *warping function* — and equilibrium then requires $\nabla^2 \psi = 0$ inside the section with a Neumann condition on its boundary, an ordinary potential problem of Section 10.4.3 [144]. The torsional rigidity that results is always *less* than μJ: the circle is the only section that does not warp, and it is the stiffest section of given area in torsion. For an ellipse of semi-axes a and b the exact rigidity is $\mu\pi a^3 b^3/(a^2 + b^2)$, which is $4a^2b^2/(a^2 + b^2)^2$ times μJ — 64 % for a 2 : 1 ellipse, and falling to zero for a slit. Using μJ for a non-circular shaft therefore overestimates its stiffness, sometimes grossly, which is why open thin-walled sections are avoided wherever torsion is carried.

Rests on Phenomenon 30.60 and Equation (30.55).

30.6.4 Plates and shells

Definition 30.63 (Flexural rigidity)**.** For a plate of thickness h made of an isotropic material, the *flexural rigidity* is

$$D = \frac{E h^3}{12\,(1 - \nu^2)}\,, \tag{30.57}$$

with SI dimension N m. The factor $(1 - \nu^2)^{-1}$ distinguishes a plate from a beam: a plate bent in one direction is prevented by its own width from contracting transversely, so the relevant stiffness is the plane-strain modulus $E/(1 - \nu^2)$ and not E.

Rests on Propositions 30.33 and 30.51.

Theorem 30.64 (The Kirchhoff plate equation)**.** *Under the hypotheses that the plate is thin, that normals to the mid-surface remain normal, and that the transverse normal stress is negligible, the transverse deflection $w(x, y, t)$ of a plate of uniform thickness obeys*

$$\rho h\, \partial_t^2 w + D\, \nabla^4 w = q\,, \tag{30.58}$$

with $\nabla^4 = \nabla^2 \nabla^2$ the biharmonic operator and q the transverse load per unit area. At a free edge the two boundary conditions are Kirchhoff's: the bending moment vanishes, and the transverse shear combined with the gradient of the twisting moment vanishes [130].

Rests on Definition 30.63 and Theorem 16.38.

Full derivation: Appendix A.57 (p. 3105).

Derives Theorem 30.64.

The reduction has three stages, all carried out there. The Kirchhoff kinematics — $u_x = -z\,\partial_x w$, $u_z = w$, and the plane-stress condition $\sigma_{zz} = 0$ in place of plane strain — turns the isotropic energy density into a quadratic in the curvatures of the mid-surface, and integrating it across the thickness produces the flexural rigidity Equation (30.57) together with a term proportional to the Gaussian curvature of the deflected surface, which is a pure boundary term whenever the edge is clamped. Varying the resulting functional gives Equation (30.58) by the Euler–Poisson equation Theorem 16.38. The third stage is Kirchhoff's: the variation leaves three boundary quantities at a free edge, one more than a fourth-order equation can carry, and the twisting moment is integrated by parts along the

edge so that its gradient joins the transverse shear, leaving two conditions and a concentrated corner force.

Phenomenon 30.65 (Chladni figures)**.** Sand strewn on a flat plate driven at one of its resonances is swept off the moving regions and collects along sharp curves — the nodal lines of the excited mode. The pattern is not continuous in the driving frequency: it is unchanged through a resonance and jumps to a wholly different figure at the next, so that a given plate and clamping reproduce the same discrete set of frequencies and the same set of figures every time [39]. The mode frequencies of geometrically similar plates of the same material scale in proportion to the thickness and inversely as the square of the lateral dimension [70, 130].

Rests on Theorem 30.64 and Proposition 30.66.

Derivation. Derives Phenomenon 30.65.

Free harmonic motion of the plate is $w = W(x,y)\mathrm{e}^{-\mathrm{i}\omega t}$ in Equation (30.58) with $q = 0$, so W satisfies $D\nabla^4 W = \rho h\omega^2 W$ with the plate's own boundary conditions. That is a self-adjoint eigenvalue problem of the kind of Definition 9.54: its spectrum is a discrete increasing sequence $\omega_1 < \omega_2 < \cdots$, so a resonance can only occur at one of countably many frequencies and the mode shape is locked to the frequency. That is the observed discontinuity: sweeping the driving frequency, nothing changes until the next ω_n is reached, and then the whole figure is replaced by the next eigenfunction.

The sand marks the nodal set $\{W = 0\}$ because a grain leaves the surface whenever the plate's peak downward acceleration exceeds g: with local amplitude A the condition is $\omega^2 A > g$, so at $f = 500\,\mathrm{Hz}$ any region moving by more than $A = g/\omega^2 = 1.0\,\mu\mathrm{m}$ throws its grains clear. They land elsewhere, are thrown again, and random-walk until they reach a region where A falls below the threshold — the neighbourhood of a node. The figure is therefore a picture of $\{W = 0\}$ broadened by the threshold, which is why the curves are sharp: near a nodal line W grows linearly, so the width of the deposit is the threshold amplitude divided by the slope of W, and both are small.

The scaling is Proposition 30.66. Nothing in this argument is peculiar to a plate; the same three statements — discrete spectrum, nodal set, threshold — describe every visualisation of a standing wave in Chapter 35. □

Proposition 30.66 (Chladni scaling)**.** *For geometrically similar plates of the same material, thickness h, lateral dimension a and fixed boundary conditions, the natural frequencies are*

$$\omega_n = \sqrt{\Lambda_n}\,\frac{h}{a^2}\sqrt{\frac{E}{12\rho\,(1-\nu^2)}}\,, \tag{30.59}$$

with Λ_n a pure number depending on the shape and the boundary conditions but on nothing else.

Rests on Theorem 30.64 and Definition 30.63.

Proof. Derives Proposition 30.66.

Write $\boldsymbol{x} = a\boldsymbol{\zeta}$ with $\boldsymbol{\zeta}$ ranging over the fixed reference shape. Then $\nabla^4 = a^{-4}\nabla_\zeta^4$, and the eigenvalue problem $D\nabla^4 W = \rho h\omega^2 W$ becomes $\nabla_\zeta^4 W = \Lambda W$ with $\Lambda = \rho h\omega^2 a^4/D$. The reduced problem involves only the reference shape and the boundary conditions, so its eigenvalues Λ_n are pure numbers. Solving for ω,

$$\omega_n^2 = \Lambda_n \frac{D}{\rho h a^4} = \Lambda_n \frac{E h^2}{12\rho\,(1-\nu^2)\,a^4}\,,$$

using Equation (30.57), and taking the square root gives Equation (30.59): proportional to h, inversely proportional to a^2, exactly as observed in Phenomenon 30.65. The same relation read backwards is a measurement of $E/\left[\rho(1-\nu^2)\right]$ from a frequency, a thickness and a diameter. □

Remark 30.67 (Germain, Kirchhoff, and the edge conditions). Chladni's figures were an open challenge for a generation. Germain won the Academy prize with an essay containing the fourth-order equation Equation (30.58) [70]; her edge conditions were not correct, and neither were Poisson's, who imposed three conditions on a free edge where a fourth-order equation admits only two. Kirchhoff resolved the conflict by deriving the conditions from the variation of the energy rather than from a force balance — the natural boundary conditions of Theorem 16.31 — which combine the transverse shear and the twisting moment into a single *effective* shear [130]. The episode is a clean instance of the general point: for a variational problem, the boundary conditions are not free to be chosen, they are part of the answer.

Rests on Theorems 16.31 and 30.64.

Remark 30.68 (Shells carry load in a different way). Give the mid-surface curvature $1/R$ and the reduction changes character. A transverse load on a flat plate can only be resisted by bending, with stiffness proportional to h^3 through Equation (30.57); on a curved shell it is resisted also by stretching of the mid-surface, with stiffness proportional to h. The ratio of the two is of order $(R/h)^2$, which for an eggshell — $R \approx 20\,\mathrm{mm}$, $h \approx 0.35\,\mathrm{mm}$ — is about 3000. That is why an egg is hard to crush end-on and trivial to break on a rim, and why every efficient structure that spans a distance is curved. The price is that a shell resists load by membrane stress and can therefore lose it catastrophically by buckling, at a load far below the Euler estimate of Equation (30.50) and notoriously sensitive to imperfections.

Rests on Definition 30.63 and Phenomenon 30.52.

30.7 Static problems and contact

Remark 30.69 (Saint-Venant's principle, and its exact status). Every reduced theory of Section 30.6 replaced a real end loading — a clamp, a pin, a hydraulic ram — by a resultant force and moment, and the replacement is licensed by what is called Saint-Venant's principle: two loadings of the same small region with the same resultant force and moment produce stress fields differing appreciably only within a distance of the order of the diameter of that region [144]. The physical content is that a self-equilibrated load — zero resultant, zero moment — produces a field that decays exponentially, and the decay length is set by the only length in the problem, the size of the loaded region.

It is called a principle and not a theorem for good reason. It is provable for a cylinder of compact cross-section, where the decay rates are the eigenvalues of an explicit auxiliary problem, and it *fails* for thin-walled and for strongly anisotropic structures, where a self-equilibrated end load can propagate for many diameters. The honest statement is that it is a very good rule for stocky isotropic bodies, that it is what makes the whole engineering theory of structures possible, and that a designer using it on a thin-walled box section is relying on something known to be false there.

Rests on Proposition 30.23 and Theorem 30.21.

Phenomenon 30.70 (Hertzian contact)**.** Two elastic bodies pressed together touch over a finite area, not at a point, and the force required grows faster than linearly with the approach of their centres. For two spheres of radii R_1, R_2 the measured relation between the normal force F and the mutual approach δ is

$$F = \tfrac{4}{3} E^* R^{1/2} \delta^{3/2}\,, \qquad \frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}\,, \qquad \frac{1}{E^*} = \frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2}\,, \tag{30.60}$$

with contact radius $a = \sqrt{R\delta}$ and peak pressure $p_0 = 3F/(2\pi a^2)$ [218]. The contact is therefore *stiffening*: doubling the approach requires nearly three times the force.

Rests on Proposition 30.23 and Definition 30.31.

Derivation. *Derives Phenomenon 30.70.*

The exponent follows from three geometric and dimensional statements and needs no solution of the elastic problem. Let a be the radius of the contact circle. First, the geometry of two nearly touching spheres gives $\delta = a^2/R$ to leading order in a/R, since the gap between the undeformed surfaces at distance a from the axis is $a^2/2R_1 + a^2/2R_2 = a^2/2R$ and the two surfaces must meet there. Second, the only strain scale in the contact region is the indentation divided by its own size, $\delta/a = a/R$, because a is the only length over which the displacement varies. Third, the stress is then E^* times that strain and acts over an area πa^2, so

$$F \sim E^* \frac{a}{R}\, a^2 = E^* \frac{a^3}{R} = E^* \frac{(R\delta)^{3/2}}{R} = E^* R^{1/2} \delta^{3/2} ,$$

which is Equation (30.60) up to the pure number $4/3$. The exponent $3/2$ is exact and is purely geometric: it says that a Hertzian contact stiffens because the contact area grows with the load, not because any material stiffens. The combination E^* is the plane-strain modulus $E/(1-\nu^2)$ of Definition 30.63, appearing because the material just under a broad contact patch is laterally confined, and the two bodies act as compliances in series — hence the sum of reciprocals. The coefficient $4/3$ and the pressure distribution $p(r) = p_0\sqrt{1 - r^2/a^2}$ require the full elastic solution for a half-space [218]. □

Full derivation: Appendix A.58 (p. 3112).

Derives Phenomenon 30.70.

What is left to the appendix is the part the scaling argument cannot reach: the numerical coefficient $4/3$ and the pressure profile $p(r) = p_0\sqrt{1 - r^2/a^2}$. It is obtained from Boussinesq's solution for a half-space loaded by a prescribed normal pressure over a circle — the surface displacement is a potential-theory integral of the pressure, of the kind of Section 10.4.3 — together with the two contact conditions: inside the circle the two deformed surfaces coincide, which for spheres means the surface displacement must vary as $\delta - r^2/2R$; outside it they separate and the pressure vanishes. The elliptic profile is the unique pressure whose Boussinesq displacement is quadratic in r inside the circle, and integrating it over the contact area gives $F = \frac{2}{3} p_0 \pi a^2$, which with $a = \sqrt{R\delta}$ is Equation (30.60).

Example 30.71 (Contact stresses are enormous)**.** Press two steel balls of radius 10 mm together with 1 N. With $E = 210\,\mathrm{GPa}$ and $\nu = 0.29$, $E^* = 115\,\mathrm{GPa}$ and $R = 5\,\mathrm{mm}$, so Equation (30.60) gives a contact radius $a = [3FR/(4E^*)]^{1/3} = 32\,\mu\mathrm{m}$, an approach $\delta = a^2/R = 0.20\,\mu\mathrm{m}$ and a peak pressure

$$p_0 = \frac{3F}{2\pi a^2} = \frac{3\,(1\,\mathrm{N})}{2\pi\,(3.2\times10^{-5}\,\mathrm{m})^2} = 4.7\times10^8\,\mathrm{Pa} = 470\,\mathrm{MPa}\,.$$

One newton — the weight of an apple — generates half a gigapascal, nearly twice the yield stress of mild steel. The mean contact strain $a/R = 6.4\times10^{-3}$ is nevertheless small, so linear elasticity is self-consistent as a description of the *strain*; what is not small is the stress, because E is large. This is why ball bearings are made of hardened steel with $\sigma_\mathrm{y} \approx 2\,\mathrm{GPa}$ rather than of the structural steel of Example 30.53, and why every rolling contact eventually fails by subsurface fatigue rather than by wear.

Rests on Phenomenon 30.70 and Equation (30.60).

Remark 30.72 (Stress concentration). A uniform tension σ in a large plate is tripled at the edge of a small circular hole: the peak stress is 3σ, independent of the hole's size [1778]. For an elliptical hole of semi-axes a across the tension and b along it, with tip radius of curvature $\varrho = b^2/a$, the peak is

$$\sigma_\mathrm{max} = \sigma\left(1 + \frac{2a}{b}\right) = \sigma\left(1 + 2\sqrt{\frac{a}{\varrho}}\right) , \tag{30.61}$$

which diverges as the ellipse degenerates into a crack, $\varrho \to 0$. Two consequences run through the rest of engineering. First, holes and notches must be given generous radii; a sharp re-entrant corner is a mathematical singularity and a practical crack starter. Second, and more seriously, Equation (30.61) says that linear elasticity predicts *infinite* stress at a crack tip for any applied load whatever, so it cannot by itself say when a cracked body breaks. Resolving that is the content of Proposition 30.76, and the resolution is to abandon the stress criterion for an energy one.

Rests on Equation (30.21) and Theorem 30.18.

30.8 The limits of elasticity

Everything above rests on Equation (30.22): that the energy is quadratic and the stress therefore linear. Four distinct things end that regime, and it is worth being explicit that they are different failures with different signatures.

Phenomenon 30.73 (Yield and brittle fracture)**.** Elasticity has a boundary, and it is reached in two different ways. A ductile metal loaded beyond a critical stress ceases to recover its shape and flows: the flow sets in when the maximum shear stress in the body reaches a material constant, and superposing a hydrostatic pressure — however large — does not by itself produce it [161]. A brittle solid does not flow but breaks, and its measured strength is one to two orders of magnitude below the stress at which the interatomic bonds would part; the strength is set by the largest pre-existing flaw and falls as the inverse square root of its length, so that fine fibres of the same glass are far stronger than bulk specimens [475].

Rests on Propositions 30.74 and 30.76.

Derivation. *Derives Phenomenon 30.73.*

The two halves are Proposition 30.74 and Proposition 30.76 respectively. What the pair says together is that "strength" is not one material property but two, governed by different parts of the stress tensor: ductile flow by the deviatoric part s_{ik} of Equation (30.16), which is why it is insensitive to pressure, and brittle fracture by the elastic energy available to be released, which is why it is sensitive to the flaw population and hence to the specimen's history rather than to its composition alone. That is also why the same material can do both: a steel that flows at room temperature cleaves below its transition temperature, where the shear stress needed for flow rises above the stress needed to run a crack. □

Proposition 30.74 (The maximum shear stress, and its indifference to pressure)**.** *The largest shear traction over all planes through a point is*

$$\tau_{\max} = \tfrac{1}{2}\,(\sigma_1 - \sigma_3)\ , \tag{30.62}$$

attained on the two planes bisecting the extreme principal directions, and it is unchanged by the addition of any hydrostatic stress $\sigma_{ik} \mapsto \sigma_{ik} - p_0\delta_{ik}$. A yield criterion of the form $\tau_{\max} = k$ is therefore independent of pressure [161].

Rests on Definition 30.20 and Theorem 30.18.

Proof. *Derives Proposition 30.74.*

Work in the principal frame of Definition 30.20, so $\sigma = \operatorname{diag}(\sigma_1, \sigma_2, \sigma_3)$ with $\sigma_1 \ge \sigma_2 \ge \sigma_3$. For a unit normal $\boldsymbol{n}$ the traction is $t_i = \sigma_i n_i$ (no sum), its normal component is $\bar\sigma = \sum_i \sigma_i m_i$ with $m_i = n_i^2$, and the shear component has magnitude

$$\tau^2 = |\boldsymbol{t}|^2 - \bar\sigma^2 = \sum_i \sigma_i^2 m_i - \Big(\sum_i \sigma_i m_i\Big)^2 , \qquad m_i \ge 0\,,\ \sum_i m_i = 1\,.$$

Since $\sigma_3 \le \sigma_i \le \sigma_1$ we have $(\sigma_i - \sigma_3)(\sigma_1 - \sigma_i) \ge 0$ for each i; averaging with the weights m_i gives $\sum_i \sigma_i^2 m_i \le (\sigma_1 + \sigma_3)\,\bar\sigma - \sigma_1\sigma_3$. Hence

$$\tau^2 \le (\sigma_1 + \sigma_3)\,\bar\sigma - \sigma_1\sigma_3 - \bar\sigma^2 = (\bar\sigma - \sigma_3)(\sigma_1 - \bar\sigma) \le \left[\frac{\sigma_1 - \sigma_3}{2}\right]^2 ,$$

the last step being the inequality between the geometric and arithmetic means of two non-negative numbers whose sum is $\sigma_1 - \sigma_3$. Both inequalities are equalities when $m = \left(\tfrac12, 0, \tfrac12\right)$, i.e. for $\boldsymbol{n}$ at 45° between the first and third principal directions, giving Equation (30.62). Adding $-p_0\delta_{ik}$ shifts every principal stress by the same $-p_0$ and leaves every difference, hence $\tau_{\max}$, unchanged. □

Remark 30.75 (Two criteria, one uncited). Equation (30.62) is not the only pressure-independent function of the deviator. The alternative in general engineering use replaces it by the second invariant $\sqrt{\tfrac32 s_{ik}s_{ik}}$, which is smooth where Equation (30.62) has corners and fits ductile metals somewhat better; the two differ by at most 15 % for any stress state. This treatise cites no primary source for that criterion, because none was located in the accessible literature during the preparation of this chapter, and the claim is recorded here as uncited rather than attached to a reference that does not support it. Tresca's criterion, by contrast, rests on his own extrusion measurements [161].

Rests on Proposition 30.74 and Definition 30.20.

Proposition 30.76 (Griffith's criterion)**.** *A plate under remote tension σ containing a through-crack of length $2c$ releases elastic energy $\pi\sigma^2c^2/E$ per unit thickness and creates surface energy $4\gamma c$, with γ the surface energy per unit area. The total energy has a maximum at*

$$\sigma_{\mathrm f} = \sqrt{\frac{2E\gamma}{\pi c}}\,, \tag{30.63}$$

so a crack shorter than $2E\gamma/(\pi\sigma^2)$ is stable and a longer one runs away. Strength therefore falls as $c^{-1/2}$ and is a property of the flaw population, not of the bonds [475].

Rests on Remark 30.72 and Equation (30.23).

Proof. *Derives Proposition 30.76.*

Let $U(c)$ be the total energy of the plate at fixed remote displacement, measured from the uncracked state. Introducing the crack relaxes the material near it and releases stored elastic energy; the elastic solution for an elliptical hole degenerating to a crack, which Griffith took over for this purpose, gives that release as $\pi\sigma^2c^2/E$ per unit thickness [475]. Creating the two new faces costs γ per unit area, and the two faces of a crack of length $2c$ have total area $4c$ per unit thickness. Hence

$$U(c) = -\frac{\pi\sigma^2c^2}{E} + 4\gamma c\,, \qquad \frac{\mathrm dU}{\mathrm dc} = -\frac{2\pi\sigma^2 c}{E} + 4\gamma\,.$$

The derivative vanishes at $c^* = 2E\gamma/(\pi\sigma^2)$ and $\mathrm d^2U/\mathrm dc^2 = -2\pi\sigma^2/E < 0$, so c^* is a *maximum* of the energy — an unstable equilibrium. A crack shorter than c^* can only grow by climbing the energy barrier and does not; one longer than c^* releases more energy than it costs at every subsequent length and accelerates without limit. Solving $c = c^*$ for σ gives Equation (30.63). The criterion is an energy balance and not a stress criterion, which is exactly what was needed: Remark 30.72 showed that the stress at a crack tip is infinite for any load, so no stress criterion could ever have distinguished a safe crack from a dangerous one. □

Example 30.77 (Why glass is weak, and thin glass is not)**.** Take a silicate glass with $E = 70\,\mathrm{GPa}$ and a surface energy $\gamma = 0.5\,\mathrm{J/m^2}$, and a surface flaw of half-length $c = 2\,\mu\mathrm{m}$ — a scratch invisible to the eye. Then Equation (30.63) gives

$$\sigma_{\mathrm{f}} = \sqrt{\frac{2\,(7.0\times10^{10}\,\mathrm{Pa})\,(0.5\,\mathrm{J/m^2})}{\pi\,(2.0\times10^{-6}\,\mathrm{m})}} = 1.06\times10^{8}\,\mathrm{Pa} = 106\,\mathrm{MPa}\,,$$

which is squarely inside the range that bulk glass actually shows. The stress the bonds themselves could support is of order $\sqrt{E\gamma/a_0}$ with $a_0 = 0.3\,\mathrm{nm}$ an interatomic spacing, about 11 GPa — a hundredfold higher, and that hundredfold is the whole of the discrepancy the phenomenon reports. Reduce the flaw to $c = 20\,\mathrm{nm}$, which is what drawing and fire-polishing a fine fibre does, and Equation (30.63) predicts 1.1 GPa; this is why Griffith found the strength of his glass fibres rising steeply as their diameter fell, and it is why an optical fibre with a pristine surface outperforms structural steel in tension [475]. Nothing about the glass changed between the two numbers — only the largest scratch on it.

Rests on Proposition 30.76 and Equation (30.63).

Remark 30.78 (Anharmonicity: what the quadratic energy also suppresses)*.* The cubic remainder discarded in Equation (30.22) is not only a correction to Hooke's law; it is the entire origin of two effects that a perfectly harmonic solid does not have. A harmonic solid does not expand when heated — the mean position of every atom sits at the minimum of a symmetric well regardless of the amplitude — and it has infinite thermal conductivity, because its normal modes never scatter off one another. Real solids do both, with a coefficient of linear expansion of order $1\times10^{-5}\,1/\mathrm{K}$ for metals, and the microscopic account is the anharmonic lattice dynamics of Chapter 125. The macroscopic bookkeeping is that the elastic moduli measured adiabatically and isothermally differ, by a term proportional to the square of the expansion coefficient; for metals at room temperature that is a fraction of a percent, which is at the level of the uncertainties in Table 30.1.

Rests on Phenomenon 30.28 and Definition 30.24.

Definition 30.79 (Maxwell viscoelasticity)**.** A *Maxwell material* in shear responds to a deviatoric stress s_{ik} with a deviatoric strain rate

$$\dot{d}_{ik} = \frac{\dot{s}_{ik}}{2\mu} + \frac{s_{ik}}{2\eta}\,, \tag{30.64}$$

the two terms being an elastic and a viscous contribution in series, with η the shear viscosity in Pa s. Its *relaxation time* is

$$t_{\mathrm{r}} = \frac{\eta}{\mu}\,. \tag{30.65}$$

Rests on Equation (30.21) and Definition 30.20.

Proposition 30.80 (A Maxwell material is a solid or a fluid depending on how long you watch)**.** *Under a step of stress applied at $t = 0$ and held, Equation (30.64) gives $d_{ik}(t) = (s_{ik}/2\mu)\,(1 + t/t_{\mathrm{r}})$: the response is elastic for $t \ll t_{\mathrm{r}}$ and viscous for $t \gg t_{\mathrm{r}}$. Under a step of strain held fixed, the stress decays as $s_{ik}(t) = s_{ik}(0)\mathrm{e}^{-t/t_{\mathrm{r}}}$.*

Rests on Definition 30.79.

Proof. *Derives Proposition 30.80.*

For a constant s_{ik} the first term of Equation (30.64) vanishes after the step and integration gives $d_{ik}(t) = d_{ik}(0) + s_{ik}t/(2\eta)$ with the instantaneous elastic response $d_{ik}(0) = s_{ik}/(2\mu)$; factoring out $s_{ik}/2\mu$ and using Equation (30.65) gives the stated form. For constant d_{ik} the left side vanishes and Equation (30.64) becomes $\dot{s}_{ik} = -\,(\mu/\eta)\,s_{ik}$, whose solution is the stated exponential. □

The ratio t_r to the time of observation decides which chapter a material belongs in. Window glass at room temperature has t_r of order 10^{20} years and is a solid by any test that will ever be applied to it; the Earth's mantle has $\eta \approx 1 \times 10^{21}\,\mathrm{Pa\,s}$ and $\mu \approx 1 \times 10^{11}\,\mathrm{Pa}$, so $t_r \approx 10^{10}$ seconds — a few hundred years. It therefore transmits the seismic waves of Phenomenon 30.43 as an elastic solid on a timescale of minutes and convects as a fluid on a timescale of millions of years, and both descriptions are correct at their own frequency. That is the sense in which Chapter 31 is not a different subject but the same one at a different value of ωt_r; the Cauchy momentum equation Equation (30.17) is common to both, and only the constitutive law relating σ_{ik} to the motion changes — from Equation (30.21), linear in the strain, to a law linear in the strain *rate*.

Chapter 31

Fluid Dynamics

A fluid is the continuum of Chapter 30 with one constitutive change that carries an entire subject: it cannot support a shear stress at rest. From that single statement follow the Euler and Navier–Stokes equations, ideal and viscous flow, Bernoulli's theorem, the conservation laws of vorticity, and the turbulence phenomenology that a century and a half of effort has organized but not derived. The chapter is built, like the rest of this part, on the measured record: the viscosities, drag laws and pipe-flow data that test every equation stated here, from Poiseuille's capillaries [105] to Reynolds's dye experiments [226] and the tidal-channel spectra that confirmed Kolmogorov's scaling [1227].

It closes the mechanics of continuous media begun in Chapter 30 and opens three doors: sound and shock waves connect to Chapter 28; the loss of stability and the transition to turbulence are the laboratory of Chapter 32; and electrically conducting fluids return as the magnetohydrodynamics of Chapter 66. The molecular origin of viscosity belongs to Chapter 55. The canonical monograph is [1340], and the measurements that discipline the whole chapter are collected in Chapter 36.

The reader should keep one asymmetry in view from the start. Everything in Sections 31.1 to 31.6.2 is derived: given the constitutive relation and the boundary conditions, the results follow. Everything in Section 31.9 is not: it is organized measurement, dimensional analysis and hypotheses that have been tested but not proved, resting moreover on equations whose well-posedness is an open problem (Section 31.4.5). The chapter marks the boundary between the two wherever it is crossed.

31.1 Kinematics of flow

31.1.1 The continuum hypothesis and the two descriptions

Definition 31.1 (Fluid)**.** A *fluid* is a continuous medium (Section 30.1) whose stress tensor in equilibrium is purely isotropic,

$$\sigma_{ij}^{(\mathrm{rest})} = -p\,\delta_{ij}\,, \tag{31.1}$$

with p the *pressure*, measured in Pa. Equivalently: a fluid at rest exerts no tangential traction on any internal surface, so a shear stress of any magnitude, however small, sets it in motion and keeps it moving.

Rests on Theorem 30.18.

The sign is a convention with content: a fluid at rest pushes on every surface along the inward normal, and $p > 0$ for every fluid treated in this chapter. Solids differ precisely in admitting a static deviatoric part, which is what Chapter 30 is about.

Definition 31.2 (Knudsen number)**.** For a flow of characteristic length L in a gas whose molecular mean free path is λ, the *Knudsen number* is the dimensionless ratio

$$\mathrm{Kn} = \frac{\lambda}{L}\,. \tag{31.2}$$

The field description of this chapter is legitimate when $\mathrm{Kn} \ll 1$.

Rests on Definition 31.1.

Remark 31.3 (When the continuum picture is allowed)*.* The mean free path in air at room temperature and atmospheric pressure is of order $7 \times 10^{-8}\,\mathrm{m}$ (Chapter 55), so $\mathrm{Kn} < 10^{-3}$ for any apparatus larger than a tenth of a millimetre and the continuum description is very safe indeed. It fails, and visibly, in three places this chapter does not treat: rarefied gases, where λ is comparable with the apparatus; the interior of a shock, whose thickness is a few mean free paths (Section 31.10.2); and the last nanometres against a wall, which is where the no-slip condition of Section 31.4.2 is ultimately decided (Section 36.2).

Rests on Definition 31.2.

Granted the hypothesis, there are two ways to label what is moving, and both go back to Euler's memoir [30]. The *material* or Lagrangian description follows an identified fluid particle, labelled by its position $\boldsymbol{X}$ at some reference time and carried to $\boldsymbol{x}(\boldsymbol{X}, t)$; the *spatial* or Eulerian description abandons the particles and records instead the fields $\boldsymbol{v}(\boldsymbol{x}, t)$, $\rho(\boldsymbol{x}, t)$, $p(\boldsymbol{x}, t)$ at fixed points of space. The Eulerian fields are what an instrument in the flow measures, and they are what the equations of this chapter are written for; the material description survives in one construction, the derivative that follows a particle.

Definition 31.4 (Material derivative)**.** The *material derivative* of a field $f(\boldsymbol{x}, t)$ is its rate of change along the motion of a fluid particle,

$$\frac{Df}{Dt} := \frac{\partial f}{\partial t} + (\boldsymbol{v}\cdot\boldsymbol{\nabla})\, f\,. \tag{31.3}$$

Rests on Proposition 7.72 and Equation (7.85).

Derivation. *Derives Definition 31.4.*

Let $\boldsymbol{x}(t)$ be the path of a particle, so that $\mathrm{d}\boldsymbol{x}/\mathrm{d}t = \boldsymbol{v}(\boldsymbol{x}(t), t)$. The chain rule in several variables (Proposition 7.72) gives

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}(t), t) = \frac{\partial f}{\partial t} + \frac{\mathrm{d}x_i}{\mathrm{d}t}\frac{\partial f}{\partial x_i} = \frac{\partial f}{\partial t} + v_i\, \partial_i f\,,$$

which is Equation (31.3). The two terms are physically distinct and it is worth naming them: $\partial_t f$ is the change an instrument at a fixed point records, and $(\boldsymbol{v}\cdot\boldsymbol{\nabla})f$ is the change a particle experiences merely by being carried into a region where f is different. A steady flow, $\partial_t f = 0$ everywhere, can therefore accelerate a particle violently — which is what happens in the Venturi throat of Section 31.3.2, and the reason the acceleration in the equations below is not $\partial_t \boldsymbol{v}$. □

Remark 31.5 (Vector identities this chapter uses, and where they belong)*.* Three identities of the nabla calculus are used repeatedly below:

$$(\boldsymbol{v}\cdot\boldsymbol{\nabla})\,\boldsymbol{v} = \boldsymbol{\nabla}\left(\tfrac{1}{2}v^2\right) - \boldsymbol{v}\times(\boldsymbol{\nabla}\times\boldsymbol{v})\,, \tag{31.4}$$

$$\boldsymbol{\nabla}\times(\boldsymbol{a}\times\boldsymbol{b}) = (\boldsymbol{b}\cdot\boldsymbol{\nabla})\,\boldsymbol{a} - (\boldsymbol{a}\cdot\boldsymbol{\nabla})\,\boldsymbol{b} + \boldsymbol{a}\,(\boldsymbol{\nabla}\cdot\boldsymbol{b}) - \boldsymbol{b}\,(\boldsymbol{\nabla}\cdot\boldsymbol{a})\,, \tag{31.5}$$

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{a}) = \boldsymbol{\nabla}\,(\boldsymbol{\nabla}\cdot\boldsymbol{a}) - \nabla^2\boldsymbol{a}\,, \qquad \boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{a}) = 0\,. \tag{31.6}$$

Each is an exercise in the permutation identity $\epsilon_{ijk}\epsilon_{klm} = \delta_{il}\delta_{jm} - \delta_{im}\delta_{jl}$ together with the equality of mixed partial derivatives. They are stated here rather than proved because they

are general mathematics, not physics: Chapter 7 defines the nabla operator (Equations (7.85) and (7.87)), proves the integral theorems and proves $\boldsymbol{\nabla} \times \boldsymbol{\nabla} f = \mathbf{0}$ (Equation (7.128)), but does not yet carry the product and second-order identities. They belong in that chapter and are recorded here as a debt against it, not as a result of fluid mechanics.

Rests on Equations (7.87) and (7.128).

Definition 31.6 (Streamline, pathline, streakline)**.** At a fixed instant t, a *streamline* is a curve everywhere tangent to $\boldsymbol{v}(\cdot, t)$, i.e. a solution of $\mathrm{d}\boldsymbol{x}/\mathrm{d}s = \boldsymbol{v}(\boldsymbol{x}, t)$ with t held fixed. A *pathline* is the trajectory of one fluid particle, the solution of $\mathrm{d}\boldsymbol{x}/\mathrm{d}t = \boldsymbol{v}(\boldsymbol{x}, t)$ with t running. A *streakline* is the locus at time t of all particles that have passed through a fixed point — the visible line of dye or smoke in an experiment.

Rests on Definition 31.4.

Proposition 31.7 (The three coincide in steady flow)**.** *If $\partial_t \boldsymbol{v} = \mathbf{0}$ everywhere, the streamlines, pathlines and streaklines of the flow are the same family of curves. If the flow is unsteady they need not be, and in general are not.*

Rests on Definition 31.6.

Proof. *Derives Proposition 31.7.*

When $\boldsymbol{v}$ does not depend on t, the two ordinary differential equations of Definition 31.6 are the same autonomous system, and by uniqueness of its solutions (Theorem 9.8) the curve through a given point is the same in both cases. Every particle passing through a fixed point therefore follows the identical curve, so the streakline is that curve too. In unsteady flow the streamline pattern changes while a particle is traversing it and the three families separate: this is why a smoke photograph of an unsteady flow is not a picture of its instantaneous streamlines, and why the dye filament of Section 36.3 is read as a streakline. □

Definition 31.8 (Rate of strain and vorticity)**.** The velocity-gradient tensor $\partial_j v_i$ splits into its symmetric and antisymmetric parts,

$$e_{ij} := \tfrac{1}{2}\left(\partial_i v_j + \partial_j v_i\right), \qquad \boldsymbol{\omega} := \boldsymbol{\nabla} \times \boldsymbol{v}, \tag{31.7}$$

e_{ij} being the *rate-of-strain tensor*, in 1/s, and $\boldsymbol{\omega}$ the *vorticity*, also in 1/s.

Rests on Equation (7.87) and Definition 13.91.

Proposition 31.9 (Cauchy–Stokes decomposition)**.** *The velocity of the fluid at a point $\boldsymbol{x} + \delta\boldsymbol{x}$ near a point $\boldsymbol{x}$ is, to first order in $\delta\boldsymbol{x}$,*

$$\delta v_i = e_{ij}\,\delta x_j + \tfrac{1}{2}\left(\boldsymbol{\omega} \times \delta\boldsymbol{x}\right)_i . \tag{31.8}$$

A fluid element therefore moves as the superposition of a translation, a rigid rotation at angular velocity $\boldsymbol{\omega}/2$, and a pure straining motion with no rotation.

Rests on Definition 31.8.

Proof. *Derives Proposition 31.9.*

Taylor's theorem gives $\delta v_i = (\partial_j v_i)\,\delta x_j$ to first order. Split the gradient into symmetric and antisymmetric parts, $\partial_j v_i = e_{ij} + \Omega_{ij}$ with $\Omega_{ij} = \frac{1}{2}\left(\partial_j v_i - \partial_i v_j\right)$. Now

$$-\tfrac{1}{2}\epsilon_{ijk}\omega_k = -\tfrac{1}{2}\epsilon_{ijk}\epsilon_{klm}\partial_l v_m = -\tfrac{1}{2}\left(\partial_i v_j - \partial_j v_i\right) = \Omega_{ij},$$

so that $\Omega_{ij}\delta x_j = -\frac{1}{2}\epsilon_{ijk}\omega_k \delta x_j = \frac{1}{2}\epsilon_{ikj}\omega_k \delta x_j = \frac{1}{2}\left(\boldsymbol{\omega} \times \delta\boldsymbol{x}\right)_i$, which is Equation (31.8). The factor of one half is the whole content of the statement that vorticity is *twice* the local angular velocity, and it is a standing source of confusion: a fluid in rigid rotation at angular velocity $\boldsymbol{\Omega}$ (Chapter 29) has $\boldsymbol{\omega} = 2\boldsymbol{\Omega}$, not $\boldsymbol{\Omega}$. □

Remark 31.10 (Vorticity is local rotation, not curvature of the path). Two examples fix the distinction, and both are realized in the laboratory. A fluid in rigid rotation, $\boldsymbol{v} = \boldsymbol{\Omega} \times \boldsymbol{x}$, has $\boldsymbol{\omega} = 2\boldsymbol{\Omega}$ everywhere although each particle moves on a circle. The *free vortex* $\boldsymbol{v} = (\Gamma/2\pi r)\,\hat{\boldsymbol{\varphi}}$ has $\boldsymbol{\omega} = \boldsymbol{0}$ everywhere except on the axis, although every particle also moves on a circle: a small paddle carried round by that flow does not turn, because the outer side of it moves more slowly than the inner. Circular streamlines are neither necessary nor sufficient for vorticity.

Rests on Proposition 31.9.

31.1.2 The continuity equation

Lemma 31.11 (Transport theorem for a material volume). *Let $V(t)$ be a volume that moves with the fluid, so that it always consists of the same particles, and let $f(\boldsymbol{x}, t)$ be continuously differentiable. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{V(t)} f \,\mathrm{d}V = \int_{V(t)} \left[\frac{\partial f}{\partial t} + \boldsymbol{\nabla}\cdot(f\boldsymbol{v})\right] \mathrm{d}V = \int_{V(t)} \left[\frac{Df}{Dt} + f\,\boldsymbol{\nabla}\cdot\boldsymbol{v}\right] \mathrm{d}V\,. \tag{31.9}$$

Rests on Theorem 7.101 and Definition 31.4.

Proof. *Derives Lemma 31.11.*

In a time $\mathrm{d}t$ the boundary element $\mathrm{d}S$ with outward unit normal $\boldsymbol{n}$ sweeps the volume $(\boldsymbol{v}\cdot\boldsymbol{n})\,\mathrm{d}S\,\mathrm{d}t$, so the integral changes for two reasons and only two: the integrand changes at fixed points, and the domain moves. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{V(t)} f \,\mathrm{d}V = \int_{V} \frac{\partial f}{\partial t}\,\mathrm{d}V + \oint_{\partial V} f\,(\boldsymbol{v}\cdot\boldsymbol{n})\,\mathrm{d}S\,,$$

and the divergence theorem (Theorem 7.101) turns the surface term into $\int_V \boldsymbol{\nabla}\cdot(f\boldsymbol{v})\,\mathrm{d}V$. Expanding $\boldsymbol{\nabla}\cdot(f\boldsymbol{v}) = \boldsymbol{v}\cdot\boldsymbol{\nabla}f + f\,\boldsymbol{\nabla}\cdot\boldsymbol{v}$ and using Equation (31.3) gives the second form. The first term of the first form requires differentiating an integral with respect to a parameter under a fixed domain, which is elementary here because f is C^1 and V is bounded. □

Theorem 31.12 (Conservation of mass). *For any fluid motion,*

$$\frac{\partial\rho}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\boldsymbol{v}) = 0\,, \qquad \textit{equivalently} \qquad \frac{D\rho}{Dt} + \rho\,\boldsymbol{\nabla}\cdot\boldsymbol{v} = 0\,. \tag{31.10}$$

Rests on Lemma 31.11.

Proof. *Derives Theorem 31.12.*

The mass of a material volume is constant, since the volume contains the same particles at all times and mass is neither created nor destroyed. Setting $f = \rho$ in Equation (31.9) and equating the result to zero,

$$\int_{V(t)} \left[\frac{\partial\rho}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\boldsymbol{v})\right] \mathrm{d}V = 0$$

for *every* material volume. A continuous function whose integral vanishes over every subvolume vanishes identically — if it were positive at a point it would be positive on a ball about that point, by continuity, and the integral over that ball would not vanish — which is Equation (31.10). The second form follows from the first by Equation (31.3). First written as a field equation by Euler [30]. □

Corollary 31.13 (Incompressible flow). *The density of every fluid particle is constant, $D\rho/Dt = 0$, if and only if the velocity field is solenoidal,*

$$\boldsymbol{\nabla}\cdot\boldsymbol{v} = 0\,. \tag{31.11}$$

Rests on Theorem 31.12.

Proof. *Derives Corollary 31.13.*

Immediate from the second form of Equation (31.10), since $\rho > 0$. Note what the statement does *not* say: it does not say the density is uniform. Salt water stratified by depth satisfies Equation (31.11) exactly while ρ varies from place to place, and that is precisely the configuration of the Rayleigh–Taylor and Kelvin–Helmholtz problems of Section 31.8.1. □

Remark 31.14 (When incompressibility is legitimate). Equation (31.11) is an approximation to Equation (31.10), and its validity is a statement about three comparisons, not one. Write c for the sound speed of Section 31.10.1, U and L for the speed and size of the flow, and T for the time over which it changes. Then the fractional density change is of order $\mathrm{Ma}^2 = U^2/c^2$ from the Bernoulli pressure variation, of order L^2/c^2T^2 from unsteadiness, and of order gL/c^2 from the hydrostatic pressure variation. All three must be small. The first is the familiar one and is the reason the flow of water and of air below about 100 m/s may be treated as incompressible: with $c \approx 343\,\mathrm{m/s}$ in air [1340], that is $\mathrm{Ma}^2 \approx 0.085$, so the density varies by under a tenth. The third is why the atmosphere as a whole is not incompressible even when the wind is slow: with L the scale height of Example 31.19, gL/c^2 is of order unity.

Rests on Corollary 31.13.

Example 31.15 (The stream function of a plane flow)**.** Let the flow be independent of one Cartesian coordinate z and have no z component — an idealization of flow past a long cylinder or a wing of large span, and the setting of Section 31.7.1. Then Equation (31.11) reads $\partial_x v_x + \partial_y v_y = 0$, which is satisfied identically by

$$v_x = \frac{\partial \psi}{\partial y}\,, \qquad v_y = -\frac{\partial \psi}{\partial x}\,, \tag{31.12}$$

for a *stream function* $\psi(x, y, t)$ in m^2/s. Since $\boldsymbol{v} \cdot \boldsymbol{\nabla}\psi = 0$, the level curves of ψ are the streamlines, and the difference of ψ between two of them is the volume flux per unit span flowing between them. The existence of ψ on a simply connected domain is the two-dimensional face of Proposition 7.104 applied to the field $(-v_y, v_x)$, whose curl is $-\boldsymbol{\nabla} \cdot \boldsymbol{v} = 0$.

Rests on Corollary 31.13 and Proposition 7.104.

31.2 Hydrostatics

Proposition 31.16 (Hydrostatic balance)**.** *In a fluid at rest in a body-force field $\boldsymbol{g}$ per unit mass,*

$$\boldsymbol{\nabla} p = \rho \boldsymbol{g}\,. \tag{31.13}$$

In a uniform gravitational field $\boldsymbol{g} = -g\hat{\boldsymbol{z}}$ this makes p a function of height alone, $\mathrm{d}p/\mathrm{d}z = -\rho g$, and requires ρ to be a function of height alone as well.

Rests on Definition 31.1 and Theorem 30.21.

Proof. *Derives Proposition 31.16.*

By Definition 31.1 the traction on any internal surface of a fluid at rest is $-p\boldsymbol{n}$. The force on a volume V is therefore $-\oint_{\partial V} p\boldsymbol{n}\,\mathrm{d}S = -\int_V \boldsymbol{\nabla} p\,\mathrm{d}V$ by the divergence theorem (Theorem 7.101), and the body force is $\int_V \rho\boldsymbol{g}\,\mathrm{d}V$. Equilibrium requires the sum to vanish for every V, hence Equation (31.13) pointwise. Taking the curl of Equation (31.13) with $\boldsymbol{g} = -g\hat{\boldsymbol{z}}$ constant gives $\boldsymbol{\nabla}\rho \times \boldsymbol{g} = \boldsymbol{0}$, so $\boldsymbol{\nabla}\rho$ is vertical: a fluid at rest in gravity is necessarily stratified horizontally, which is why a tilted density interface cannot be at rest. □

Phenomenon 31.17 (Pascal's principle)**.** A pressure applied to a confined liquid is transmitted undiminished to every part of the liquid and to the walls of its container [7]. In

consequence a small force applied to a piston of area A_1 balances a large force on a piston of area A_2 in the same closed vessel at the same height,

$$\frac{F_2}{F_1} = \frac{A_2}{A_1}\,, \tag{31.14}$$

which is the hydraulic press. The work done is the same at both pistons, the smaller one moving farther in the ratio A_2/A_1.

Derivation. *Derives Phenomenon 31.17.*

Let the applied pressure change the boundary value from p to $p + \Delta p$ with the fluid remaining at rest and its density unchanged. Both states satisfy Equation (31.13) with the same right-hand side, so the difference field satisfies $\boldsymbol{\nabla}(\Delta p) = \mathbf{0}$ and Δp is the same constant throughout a connected body of liquid — which is Pascal's statement. The force on a piston of area A is then $F = \Delta p\, A$ at each piston, giving Equation (31.14). Conservation of volume for an incompressible liquid, $A_1 d_1 = A_2 d_2$ for the displacements d_i, gives $F_1 d_1 = F_2 d_2$: the press multiplies force and divides distance, and creates no energy. The idealization is the constancy of ρ; a real liquid transmits the change at the sound speed of Section 31.10.1, not instantaneously. □

Phenomenon 31.18 (Archimedes' principle)**.** A body wholly or partly immersed in a fluid at rest is acted on by an upward force equal to the weight of the fluid it displaces,

$$\boldsymbol{F}_\mathrm{b} = -\rho_\mathrm{f} V \boldsymbol{g}\,, \tag{31.15}$$

where V is the immersed volume and ρ_f the density of the fluid; the line of action passes through the centre of mass of the displaced fluid, and neither the magnitude nor the point of application depends on what the body is made of. A body free to move therefore floats at the depth at which the displaced weight equals its own [274].

Rests on Proposition 31.16.

Derivation. *Derives Phenomenon 31.18.*

A fluid at rest carries no shear stress, so the traction on any internal surface is the normal pressure alone and the balance of forces on a fluid element reads $\boldsymbol{\nabla} p = \rho_\mathrm{f} \boldsymbol{g}$. The force exerted on the immersed body is the integral of that pressure over its wetted surface S, directed inwards along the outward normal $\boldsymbol{n}$:

$$\boldsymbol{F}_\mathrm{b} = -\oint_S p\,\boldsymbol{n}\,\mathrm{d}S = -\int_V \boldsymbol{\nabla} p\,\mathrm{d}V = -\int_V \rho_\mathrm{f}\boldsymbol{g}\,\mathrm{d}V = -\rho_\mathrm{f} V \boldsymbol{g}\,,$$

the second step by the divergence theorem (Theorem 7.101). The manoeuvre deserves a word: the pressure field is defined only outside the body, and what the divergence theorem is applied to is its hydrostatic continuation into V. That is legitimate because the surface integral depends only on the boundary values, which are unchanged if the body is replaced by the same volume of fluid — and that replacement is in equilibrium, which is Archimedes' own argument. Taking moments the same way places the resultant at the centroid of V, the centre of buoyancy. Since $\boldsymbol{F}_\mathrm{b}$ involves the body only through V, two bodies of equal volume are buoyed equally however they differ in material, and a body of mean density ρ_b floats with the fraction $\rho_\mathrm{b}/\rho_\mathrm{f}$ of its volume submerged. □

Example 31.19 (The barometric distribution)**.** For an isothermal ideal gas of molar mass M at temperature T, $\rho = Mp/RT$, and Equation (31.13) integrates to

$$p(z) = p(0)\exp\left(-\frac{z}{H}\right), \qquad H = \frac{RT}{Mg}\,. \tag{31.16}$$

For air, $M = 2.896 \times 10^{-2}\,\mathrm{kg/mol}$ and $T = 288\,\mathrm{K}$ with $g = 9.806\,65\,\mathrm{m/s^2}$ (Appendix B) give a scale height $H = 8.4\,\mathrm{km}$, which is the right order for the observed decay of atmospheric pressure with altitude. The isothermal assumption is the weak point — the real troposphere has a lapse rate — and the molecular reading of Equation (31.16), as the Boltzmann distribution in the gravitational potential, belongs to Chapter 55.

Rests on Proposition 31.16.

Phenomenon 31.20 (Capillary rise)**.** A liquid that wets the wall of a narrow vertical tube dipped into it climbs above the outer free surface, to a height inversely proportional to the radius of the tube,

$$h = \frac{2\gamma \cos\theta}{\rho g r}\,, \tag{31.17}$$

where γ is the surface tension of the liquid–air interface, in N/m, and θ the contact angle at the wall. For clean water against glass, $\gamma \approx 7.3 \times 10^{-2}\,\mathrm{N/m}$ at room temperature [1340] and $\theta \approx 0$, so a tube of radius 0.1 mm raises the water by about 0.15 m. A liquid that does not wet the wall, such as mercury against glass, is depressed instead, the same formula with $\cos\theta < 0$.

Derivation. Derives Phenomenon 31.20.

A curved interface supports a pressure difference, the Young–Laplace relation [50],

$$\Delta p = \gamma \left(\frac{1}{R_1} + \frac{1}{R_2} \right)\,, \tag{31.18}$$

the pressure being higher on the concave side; R_1 and R_2 are the principal radii of curvature of the surface (Section 13.3.3). In a tube of radius r wetted at contact angle θ the meniscus is a spherical cap of radius $R = r/\cos\theta$, so both principal radii equal R and the pressure immediately below the meniscus is lower than the atmospheric pressure above it by $2\gamma\cos\theta/r$. Hydrostatic balance (Equation (31.13)) along the liquid from that point down to the level of the flat outer surface, where the pressure is atmospheric, requires the meniscus to stand a height h above it with $\rho g h = 2\gamma\cos\theta/r$, which is Equation (31.17). Two checks are worth making. The rise is not a force exerted by the tube on the liquid — it is the pressure deficit under a curved surface — so it does not depend on the thickness or material of the wall except through θ. And the $1/r$ dependence is the measured signature: a set of tubes of different bores gives a straight line through the origin when h is plotted against $1/r$, from whose slope γ is obtained, which is the classical capillary measurement of surface tension. □

31.3 Ideal flow: the Euler equations

31.3.1 The Euler equations

Definition 31.21 (Ideal fluid)**.** An *ideal* fluid is one whose stress tensor is isotropic *in motion as well as at rest*, $\sigma_{ij} = -p\delta_{ij}$ everywhere and at all times. Equivalently, its viscosity is taken to be exactly zero.

Rests on Definition 31.1.

Theorem 31.22 (Euler's equations of motion)**.** *The motion of an ideal fluid in a body-force field $\boldsymbol{g}$ per unit mass satisfies*

$$\rho \frac{D\boldsymbol{v}}{Dt} = \rho \left[\frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v}\cdot\boldsymbol{\nabla})\,\boldsymbol{v} \right] = -\boldsymbol{\nabla} p + \rho \boldsymbol{g}\,, \tag{31.19}$$

which together with Equation (31.10) and an equation of state closes the system [30].

Rests on Definition 31.21, Lemma 31.11, and Theorem 30.21.

Proof. *Derives Theorem 31.22.*

Apply Lemma 31.11 to the momentum of a material volume, $f = \rho v_i$, and use Equation (31.10):

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{V(t)} \rho v_i \,\mathrm{d}V = \int_V \left[\frac{D(\rho v_i)}{Dt} + \rho v_i\, \boldsymbol{\nabla}\cdot\boldsymbol{v}\right]\mathrm{d}V = \int_V \rho\,\frac{Dv_i}{Dt}\,\mathrm{d}V\,,$$

the cross terms cancelling by Equation (31.10) exactly. The rate of change of momentum equals the applied force, which for an ideal fluid is the pressure traction plus the body force, $-\oint_{\partial V} p n_i \,\mathrm{d}S + \int_V \rho g_i \,\mathrm{d}V = \int_V (-\partial_i p + \rho g_i)\,\mathrm{d}V$ by Theorem 7.101. Equating the two for every material volume, and arguing from continuity as in Theorem 31.12, gives Equation (31.19). The result is Cauchy's momentum equation (Section 30.3.2) with the constitutive relation of Definition 31.21 inserted, so nothing new about mechanics has been assumed: the whole content of Equation (31.19) is Definition 31.21. □

Remark 31.23 (Boundary conditions, and what the ideal model omits). Equation (31.19) is first order in space in the velocity, so it admits only one scalar boundary condition on a solid wall: non-penetration, $\boldsymbol{v}\cdot\boldsymbol{n} = 0$ on the wall, with the tangential velocity left free to slip. The Navier–Stokes equations of Section 31.4.2 are second order and demand one more — no slip, $\boldsymbol{v} = \boldsymbol{0}$ — and that extra condition cannot be imposed on Equation (31.19). The whole difficulty of the inviscid limit lives in that mismatch: the small parameter multiplies the highest derivative, so the limit is singular and a thin layer must appear in which the discarded term is restored. That layer is Prandtl's (Section 31.7.2), and until it was identified the ideal theory returned the flatly false answer of Section 31.3.3.

Rests on Theorem 31.22.

31.3.2 Bernoulli's theorem

Theorem 31.24 (Bernoulli). *Let the flow be steady, let the body force be conservative, $\boldsymbol{g} = -\boldsymbol{\nabla}\Phi$, and let the fluid be ideal and barotropic, $\rho = \rho(p)$, so that a pressure function $P(p) = \int^p \mathrm{d}p'/\rho(p')$ exists. Then the* Bernoulli head

$$B := P + \tfrac{1}{2}v^2 + \Phi \tag{31.20}$$

is constant along each streamline. If the flow is in addition irrotational, $\boldsymbol{\omega} = \boldsymbol{0}$, then B takes the same value on every streamline. If the flow is unsteady and irrotational with $\boldsymbol{v} = \boldsymbol{\nabla}\phi$, the corresponding statement is that

$$\frac{\partial\phi}{\partial t} + P + \tfrac{1}{2}v^2 + \Phi = F(t) \tag{31.21}$$

holds throughout the fluid, with F a function of time alone.

Rests on Theorem 31.22 and Equation (31.4).

Proof. *Derives Theorem 31.24.*

Divide Equation (31.19) by ρ and use Equation (31.4) on the advective term together with $\boldsymbol{\nabla}p/\rho = \boldsymbol{\nabla}P$, valid exactly when $\rho = \rho(p)$:

$$\frac{\partial\boldsymbol{v}}{\partial t} + \boldsymbol{\nabla}\left(P + \tfrac{1}{2}v^2 + \Phi\right) = \boldsymbol{v}\times\boldsymbol{\omega}\,. \tag{31.22}$$

For steady flow the first term vanishes and $\boldsymbol{\nabla}B = \boldsymbol{v}\times\boldsymbol{\omega}$. Contract with $\boldsymbol{v}$: the right-hand side is orthogonal to $\boldsymbol{v}$, so $\boldsymbol{v}\cdot\boldsymbol{\nabla}B = 0$, and B is constant along the integral curves of $\boldsymbol{v}$ — the streamlines (Definition 31.6). If moreover $\boldsymbol{\omega} = \boldsymbol{0}$ then $\boldsymbol{\nabla}B = \boldsymbol{0}$ outright and B is one constant for the whole (connected) fluid. Finally, if $\boldsymbol{v} = \boldsymbol{\nabla}\phi$ then $\partial_t\boldsymbol{v} = \boldsymbol{\nabla}\partial_t\phi$ and $\boldsymbol{v}\times\boldsymbol{\omega} = \boldsymbol{0}$, so Equation (31.22) says that the gradient of the left-hand side of Equation (31.21) vanishes,

which is the claim. Note which hypothesis does which work: barotropy is what makes $\boldsymbol{\nabla} p/\rho$ a gradient, steadiness is what removes $\partial_t \boldsymbol{v}$, and irrotationality is what upgrades "along a streamline" to "everywhere". Dropping any one of them leaves a weaker but still usable statement, and dropping viscosity is what restricts every application to regions outside boundary layers and wakes. □

Phenomenon 31.25 (Pressure falls where the flow speeds up)**.** In the steady flow of a liquid of small viscosity the static pressure is lower where the fluid moves faster. Along a streamline the combination

$$p + \tfrac{1}{2}\rho v^2 + \rho g z = \text{const} \tag{31.23}$$

is observed to be constant [22, 30]. Two instruments measure the statement directly: a Pitot tube, whose mouth faces upstream and brings the flow to rest, reads a pressure exceeding the ambient static pressure by $\tfrac{1}{2}\rho v^2$, and is on that basis the standard airspeed and flow meter [20]; and a Venturi contraction shows a pressure drop at its throat fixed by the ratio of throat to inlet area, from which the volume flux follows [42].

Rests on Theorem 31.24.

Derivation. *Derives Phenomenon 31.25.*

For steady flow of a fluid of constant density in a uniform gravitational field the Euler equation reads $\rho(\boldsymbol{v}\cdot\boldsymbol{\nabla})\boldsymbol{v} = -\boldsymbol{\nabla} p - \rho g \boldsymbol{\nabla} z$. Use the identity Equation (31.4) and divide by ρ:

$$\boldsymbol{\nabla}\left(\frac{p}{\rho} + \frac{v^2}{2} + g z\right) = \boldsymbol{v} \times (\boldsymbol{\nabla} \times \boldsymbol{v})\,.$$

Project onto the local flow direction. The right-hand side is orthogonal to $\boldsymbol{v}$ and its projection vanishes identically, so the directional derivative of the bracket along a streamline is zero and the bracket is constant on each streamline — which is Equation (31.23), the constant-density case of Equation (31.20). Two riders follow at once. If the flow is in addition irrotational, $\boldsymbol{\nabla}\times\boldsymbol{v} = 0$ everywhere and the constant is the same on every streamline. And at a stagnation point, where $v = 0$ at the same height, the pressure exceeds the free-stream value by exactly $\tfrac{1}{2}\rho v^2$: that difference is what a Pitot tube reads [20]. The derivation used the inviscid equation, so the relation holds only where viscous stresses are negligible along the streamline followed — outside boundary layers and away from wakes. □

Example 31.26 (The Venturi meter)**.** A pipe of cross-section A_1 contracts smoothly to a throat of section A_2 and expands again. Mass conservation for an incompressible liquid gives $v_1 A_1 = v_2 A_2$, and Equation (31.23) along the axis at constant height gives $p_1 - p_2 = \tfrac{1}{2}\rho\,(v_2^2 - v_1^2)$. Eliminating the velocities,

$$Q = v_1 A_1 = A_2 \sqrt{\frac{2\,(p_1 - p_2)}{\rho\left[1 - (A_2/A_1)^2\right]}}\,, \tag{31.24}$$

so a differential pressure reading and two areas give the volume flux with no calibration constant whatever [42]. In practice a discharge coefficient near unity is applied, and its departure from unity measures exactly what the derivation left out: the boundary layer on the wall of the contraction. The efflux problem, in which the same theorem is applied between a free surface and a jet, is Phenomenon 36.1; there the ideal theory gets the speed exactly right and the discharge wrong, for the reason set out in Section 36.1.

Rests on Phenomenon 31.25.

Remark 31.27 (What the 1738 source contains). The relation Equation (31.23) is universally called Bernoulli's, and the *Hydrodynamica* [22] is the right source for the physical content — efflux from vessels, the exchange between what he calls the ascending force and the motion

of the liquid. It does not contain the formula in the modern local form, which is a *vis viva* argument for a stream tube rather than a statement about a velocity field, and the field form together with the equations it follows from is Euler's [30]. The distinction is recorded because a reader who checks the 1738 source against Equation (31.23) will not find it there, and should know that this is a fact about the history and not an error in the citation.

Rests on Phenomenon 31.25.

31.3.3 Potential flow and the d'Alembert paradox

Proposition 31.28 (Potential flow reduces to Laplace's equation)**.** *Let an incompressible flow be irrotational on a simply connected region. Then there is a velocity potential ϕ with $\boldsymbol{v} = \boldsymbol{\nabla}\phi$, and it is harmonic (Definition 10.66),*

$$\nabla^2 \phi = 0 \,, \tag{31.25}$$

with the boundary condition $\partial\phi/\partial n = \boldsymbol{U}\cdot\boldsymbol{n}$ on a solid surface moving at velocity $\boldsymbol{U}$. The velocity field is therefore determined by the geometry and the boundary motion alone; the pressure follows afterwards from Equation (31.21), and the density never enters the kinematics.

Rests on Corollary 31.13, Proposition 7.104, and Equation (10.24).

Proof. *Derives Proposition 31.28.*

Irrotationality is $\boldsymbol{\nabla}\times\boldsymbol{v} = \boldsymbol{0}$; on a simply connected region Proposition 7.104 then supplies a potential ϕ with $\boldsymbol{v} = \boldsymbol{\nabla}\phi$. Substituting into Equation (31.11) gives $\boldsymbol{\nabla}\cdot\boldsymbol{\nabla}\phi = \nabla^2\phi = 0$, which is Equation (31.25), an elliptic equation with the full apparatus of Section 10.4.3 available: uniqueness up to a constant for the Neumann problem, the maximum principle (Theorem 10.77) applied to each velocity component, and separation of variables in the coordinate systems of Theorem 10.34. Ellipticity has a physical reading worth stating: the potential at each point depends on the boundary data everywhere, instantaneously, so an incompressible ideal fluid transmits the influence of a moving boundary with infinite speed. That is a property of the incompressible idealization, not of nature, and it is repaired by the finite sound speed of Section 31.10.1. □

Example 31.29 (Flow past a sphere)**.** For a sphere of radius a held in a stream of speed U along $\hat{\boldsymbol{z}}$, the harmonic function satisfying $\partial_r\phi = 0$ at $r = a$ and $\boldsymbol{v} \longrightarrow U\hat{\boldsymbol{z}}$ at infinity is

$$\phi = U\cos\theta\left(r + \frac{a^3}{2r^2}\right) , \tag{31.26}$$

whence, using Equation (7.92),

$$v_r = U\cos\theta\left(1 - \frac{a^3}{r^3}\right) , \qquad v_\theta = -U\sin\theta\left(1 + \frac{a^3}{2r^3}\right) . \tag{31.27}$$

On the surface $v_r = 0$ as required and the speed is $q = \tfrac{3}{2}U\sin\theta$: the fluid at the equator of the sphere moves half again as fast as the free stream, and comes to rest at the two poles. Equation (31.23) then gives the surface pressure

$$p(\theta) = p_\infty + \tfrac{1}{2}\rho U^2\left(1 - \tfrac{9}{4}\sin^2\theta\right) . \tag{31.28}$$

Rests on Proposition 31.28.

Phenomenon 31.30 (Steady ideal flow exerts no drag)**.** A body held in the steady potential flow of an ideal incompressible fluid experiences zero force along the stream [27]. The conclusion is in flat contradiction with every observation of a real fluid, however small

its viscosity is made, and it is not repaired by taking the viscosity smaller: the measured drag coefficient of a sphere tends to a constant of order unity as the Reynolds number grows (Phenomenon 31.55), not to zero. This is the sharpest failure of a limit in classical physics, and its resolution — that the limit of vanishing viscosity is not uniform near a wall — is Phenomenon 36.12.

Derivation. *Derives Phenomenon 31.30.*

Take the sphere of Example 31.29. The drag is the component of the pressure resultant along the stream,

$$F_z = -\oint_S p\, n_z \,\mathrm{d}S = -\int_0^{\pi} p(\theta)\cos\theta \, 2\pi a^2 \sin\theta \,\mathrm{d}\theta \,,$$

and by Equation (31.28) the integrand is a combination of $\cos\theta\sin\theta$ and $\cos\theta\sin^3\theta$, both odd about $\theta = \pi/2$ and both integrating to zero over $(0,\pi)$. Hence $F_z = 0$ exactly. The mechanism is visible in Equation (31.28): the pressure distribution is symmetric fore and aft, so the high pressure at the rear stagnation point pushes the sphere forward exactly as hard as the high pressure at the front pushes it back. The general statement, for a body of arbitrary shape in three dimensions, follows from the same fore-and-aft structure of the far field and is left to the appendix. □

Full derivation: Appendix A.59 (p. 3118).
Derives Phenomenon 31.30.

Remark 31.31 (Two escapes, and only two). The theorem has exactly two loopholes, and both are realized. The first is circulation: if the fluid region is not simply connected — around a body of large span, in the plane transverse to it — the potential need not be single valued, the flow may carry a circulation, and a transverse force appears. That is the lift of Section 31.7.1, and it is still *drag* free, so the paradox survives it. The second is that the flow may fail to be steady or to remain attached, which is what happens in nature and is Section 31.7.2. Lamb's treatise [701] is the classical repository of the potential solutions, and its value is undiminished by the fact that its central prediction about drag is wrong: the outer flow it computes is what the streak photographs of Section 36.5 actually show.
Rests on Phenomenon 31.30.

31.4 Viscous flow: the Navier–Stokes equations

31.4.1 Viscosity

Definition 31.32 (Newtonian fluid; dynamic and kinematic viscosity)**.** A fluid is *Newtonian* if the deviatoric part of its stress tensor is a linear, isotropic, instantaneous function of the rate of strain e_{ij} of Equation (31.7). For the simple shearing motion $\boldsymbol{v} = v(y)\hat{\boldsymbol{x}}$ this reduces to Newton's own statement [14]

$$\sigma_{xy} = \mu \,\frac{\mathrm{d}v}{\mathrm{d}y} \,, \tag{31.29}$$

defining the *dynamic viscosity* μ, whose SI unit is the pascal second, $\mathrm{Pa\,s} = \mathrm{kg/(m\,s)}$. The *kinematic viscosity* is $\nu := \mu/\rho$, in $\mathrm{m^2/s}$.
Rests on Definitions 31.1 and 31.8.

Remark 31.33 (Reading the two viscosities). μ is the constant of the constitutive relation and is what a viscometer measures; ν is the diffusivity of momentum, and it is ν that appears in the equations of motion once they are divided by ρ. The distinction is not pedantry: air is about fifty times less viscous than water in μ and about fifteen times *more* viscous in ν, so momentum diffuses faster in air, and the boundary layer of Section 31.7.2 on a given body

at a given speed is thicker in air than in water. Table 31.1 gives the values used throughout this chapter.

Rests on Definition 31.32.

Table 31.1: Density, dynamic and kinematic viscosity of the two fluids used in the worked numbers of this chapter, at 20 °C and, for air, at a pressure of 101 325 Pa. Values rounded to three significant figures from the tabulation in [1340]; all three quantities vary appreciably with temperature, the viscosity of water by a few percent per kelvin, which is why the capillary measurements of Section 36.2 were made in a controlled bath.

Fluid	ρ (kg/m^3)	μ (Pa s)	ν (m^2/s)
Water	998	1.00×10^{-3}	1.00×10^{-6}
Air	1.20	1.81×10^{-5}	1.51×10^{-5}

Phenomenon 31.34 (The viscosity of a gas does not depend on its density)**.** The dynamic viscosity of a gas is, over a wide range of pressures, independent of the density of the gas: pumping half the air out of a vessel does not halve the damping of a body oscillating inside it. The result was predicted from a molecular model and then measured by Maxwell in a torsional oscillation apparatus [167], and it is counter-intuitive enough that it served as the decisive early test of the kinetic theory — a continuum theory of viscosity has nothing to say about it either way. The independence fails at low enough pressure, where the mean free path becomes comparable with the apparatus and Definition 31.2 is no longer small.

Derivation. Derives Phenomenon 31.34.

The derivation is molecular and belongs to Chapter 55, where it is carried out; the one-line reason is that momentum is carried across a shear layer by molecules crossing it, at a rate proportional to the number density n, each carrying the momentum difference accumulated over one mean free path $\lambda \propto 1/n$, so that $\mu \sim n\lambda \times$ (molecular momentum) is independent of n. What belongs *here* is the consequence: the continuum theory of this chapter takes μ as an empirical coefficient and can neither predict this behaviour nor be embarrassed by it. That is the honest division of labour between Chapter 31 and Chapter 55, and it is why the value of μ appears in Table 31.1 as a measurement. □

Remark 31.35 (Non-Newtonian fluids are out of scope). Many real materials — polymer solutions, suspensions, molten plastics, blood in the smallest vessels — have a deviatoric stress that depends nonlinearly on e_{ij}, or on its history, or both. They are not covered by Definition 31.32 and nothing in this chapter applies to them. Air, water, mercury, glycerol and the light oils are Newtonian to high accuracy over the range of shear rates considered here, which is what makes the agreement reported in Section 36.2 a test of the equations rather than of a fitting form.

Rests on Definition 31.32.

31.4.2 The Navier–Stokes equations

Theorem 31.36 (The Newtonian constitutive relation)**.** *The most general stress tensor that is (i) linear in the velocity gradient, (ii) isotropic, (iii) symmetric, and (iv) reduces to $-p\delta_{ij}$ when the fluid is in rigid-body motion, is*

$$\sigma_{ij} = -p\,\delta_{ij} + 2\mu\left(e_{ij} - \tfrac{1}{3}e_{kk}\delta_{ij}\right) + \zeta\, e_{kk}\,\delta_{ij}\,, \tag{31.30}$$

with exactly two independent coefficients: the shear viscosity μ of Definition 31.32 and the bulk viscosity *ζ, both in* Pa s.

Rests on Definition 31.32 and Proposition 31.9.

Proof. *Derives Theorem 31.36.*

By (i), $\sigma_{ij} + p\delta_{ij} = A_{ijkl}\partial_k v_l$ for a constant fourth-rank tensor A. By (ii) that tensor is isotropic, and every isotropic fourth-rank tensor in three dimensions is a combination of the three available products of Kronecker deltas,

$$A_{ijkl} = \lambda\,\delta_{ij}\delta_{kl} + \mu_1\,\delta_{ik}\delta_{jl} + \mu_2\,\delta_{il}\delta_{jk}\,,$$

a representation this chapter takes from the algebra of Cartesian tensors (Section 13.2.2). Symmetry of σ in (i,j) forces $\mu_1 = \mu_2 =: \mu$, and then

$$\sigma_{ij} + p\delta_{ij} = \lambda\,(\boldsymbol{\nabla}\cdot\boldsymbol{v})\,\delta_{ij} + \mu\,(\partial_i v_j + \partial_j v_i) = \lambda\,e_{kk}\,\delta_{ij} + 2\mu\,e_{ij}\,,$$

which involves only the symmetric part e_{ij} — so requirement (iv) is automatically met, since the antisymmetric part of the velocity gradient, which by Proposition 31.9 is the rigid rotation, has dropped out. Writing $e_{ij} = \left(e_{ij} - \frac{1}{3}e_{kk}\delta_{ij}\right) + \frac{1}{3}e_{kk}\delta_{ij}$ and collecting the isotropic pieces gives Equation (31.30) with $\zeta = \lambda + \frac{2}{3}\mu$. The split is the useful one because the first viscous term is trace free and responds only to shape change, while ζ responds only to volume change and therefore cannot be measured in any incompressible flow. □

Theorem 31.37 (The Navier–Stokes equations)**.** *For a Newtonian fluid,*

$$\rho\frac{D\boldsymbol{v}}{Dt} = -\boldsymbol{\nabla}p + \mu\nabla^2\boldsymbol{v} + \left(\zeta + \tfrac{1}{3}\mu\right)\boldsymbol{\nabla}\,(\boldsymbol{\nabla}\cdot\boldsymbol{v}) + \rho\boldsymbol{g}\,, \tag{31.31}$$

and when the flow is incompressible and μ uniform this reduces to

$$\frac{\partial\boldsymbol{v}}{\partial t} + (\boldsymbol{v}\cdot\boldsymbol{\nabla})\,\boldsymbol{v} = -\frac{1}{\rho}\boldsymbol{\nabla}p + \nu\nabla^2\boldsymbol{v} + \boldsymbol{g}\,, \qquad \boldsymbol{\nabla}\cdot\boldsymbol{v} = 0\,. \tag{31.32}$$

The boundary condition at a solid surface is no slip*: the fluid velocity equals the velocity of the surface, both normal and tangential components.*

Rests on Theorems 31.22 and 31.36.

Proof. *Derives Theorem 31.37.*

Cauchy's momentum equation (Section 30.3.2), derived exactly as in Theorem 31.22 but keeping the full traction, is $\rho\,Dv_i/Dt = \partial_j\sigma_{ij} + \rho g_i$. Insert Equation (31.30) with μ and ζ constant:

$$\partial_j\sigma_{ij} = -\partial_i p + \mu\left(\nabla^2 v_i + \partial_i\partial_j v_j\right) - \tfrac{2}{3}\mu\,\partial_i\,(\partial_j v_j) + \zeta\,\partial_i\,(\partial_j v_j)\,,$$

which collects to Equation (31.31). Under Equation (31.11) every term carrying $\boldsymbol{\nabla}\cdot\boldsymbol{v}$ vanishes and division by ρ gives Equation (31.32). The bulk viscosity has disappeared, which is why it plays no part anywhere in this chapter except in the attenuation of sound.

The equations are second order in space, so they carry one more boundary condition than Equation (31.19) (Remark 31.23). That the extra condition is *no slip* rather than some finite slip is not a theorem: it is a physical hypothesis, and the one Navier's own derivation from a molecular model [75] left open. It was settled by measurement, and the measurement is Poiseuille's — Section 36.2 turns the accuracy of Equation (31.35) into a bound on the slip length of a few tens of nanometres. □

Remark 31.38 (Whose equations, and the Stokes hypothesis). The customary name compresses a sequence. Navier reached Equation (31.32) in 1822 from a molecular model with one coefficient [75]; Cauchy and Poisson rederived it on other grounds; Saint-Venant published the equations in essentially their final continuum form in 1843 [110], two years before Stokes, whose memoir [117] gave the constitutive argument reproduced in Theorem 31.36 and argued the no-slip condition. The label "Navier–Stokes" is therefore a convention rather than a

priority claim, and Saint-Venant's omission from it is the clearest instance in this chapter of a name recording custom rather than history.

Stokes further proposed $\zeta = 0$, which is exact for a monatomic dilute gas and is a good approximation for many others, but is not a consequence of anything in Theorem 31.36. Nothing in this chapter depends on it.

Rests on Theorem 31.37.

31.4.3 Exact solutions: Couette and Poiseuille flow

The nonlinear term of Equation (31.32) is what makes the system intractable, so the solutions that are known in closed form are almost exactly the flows in which it vanishes identically. In a unidirectional flow $\boldsymbol{v} = v(y)\hat{\boldsymbol{x}}$ or $v(r)\hat{\boldsymbol{z}}$ the advective term $(\boldsymbol{v}\cdot\boldsymbol{\nabla})\boldsymbol{v}$ is zero because v does not vary along the direction in which it points. The solutions below are therefore *exact* solutions of the full equations, not approximations, and that is what makes their comparison with measurement a clean test.

Proposition 31.39 (Plane Couette flow)**.** *Between two parallel plates a distance h apart, one at rest and one moving in its own plane at speed U, with no applied pressure gradient, the steady solution of Equation (31.32) is*

$$v(y) = U\,\frac{y}{h}\,, \tag{31.33}$$

and the tangential stress transmitted to each plate is uniform and equal to $\mu U/h$.

Rests on Theorem 31.37.

Proof. *Derives Proposition 31.39.*

With $\boldsymbol{v} = v(y)\hat{\boldsymbol{x}}$ the advective term vanishes, the flow is divergence free, and the steady x component of Equation (31.32) without a pressure gradient reduces to $\mu\,\mathrm{d}^2v/\mathrm{d}y^2 = 0$. Hence v is linear in y; no slip at both plates fixes $v(0) = 0$ and $v(h) = U$, giving Equation (31.33). The stress follows from Equation (31.29). Two features are worth recording: the profile is independent of μ, so a Couette measurement determines μ from the *force* and not from the shape of the profile, and the velocity gradient is uniform, so this is the one flow in which the constitutive relation Equation (31.29) is tested at a single, controlled shear rate. Couette's rotating-cylinder viscometer [246] is the practical realization, with the small gap between the cylinders playing the part of h. □

Proposition 31.40 (Circular Couette flow)**.** *Between coaxial cylinders of radii $R_1 < R_2$ rotating at angular velocities Ω_1 and Ω_2, the steady azimuthal solution is*

$$v_\varphi(r) = Ar + \frac{B}{r}\,, \qquad A = \frac{\Omega_2 R_2^2 - \Omega_1 R_1^2}{R_2^2 - R_1^2}\,, \qquad B = \frac{(\Omega_1 - \Omega_2)\,R_1^2 R_2^2}{R_2^2 - R_1^2}\,, \tag{31.34}$$

and the viscous torque per unit length transmitted between the cylinders is $G = 4\pi\mu B$, independent of r.

Rests on Theorem 31.37 and Proposition 31.39.

Proof. *Derives Proposition 31.40.*

Seek $\boldsymbol{v} = v_\varphi(r)\hat{\boldsymbol{\varphi}}$. The flow is divergence free, and the advective term contributes only the centripetal $-v_\varphi^2/r$ in the radial direction, which is balanced by the radial pressure gradient and does not enter the azimuthal equation. That equation is

$$\frac{\mathrm{d}}{\mathrm{d}r}\left[\frac{1}{r}\frac{\mathrm{d}\,(r v_\varphi)}{\mathrm{d}r}\right] = 0\,,$$

whose general solution is $v_\varphi = Ar + B/r$; the two no-slip conditions $v_\varphi(R_i) = \Omega_i R_i$ give the constants in Equation (31.34). The term Ar is rigid rotation and carries no shear; the shear

stress is $\sigma_{r\varphi} = \mu r\,\mathrm{d}\left(v_\varphi/r\right)/\mathrm{d}r = -2\mu B/r^2$, so the torque per unit length on a cylindrical surface of radius r is $2\pi r^2 \times 2\mu B/r^2 = 4\pi\mu B$, the same at every radius — as it must be, since no torque accumulates in a steady state. This is the working equation of the Couette viscometer [246]. The solution loses stability when the inner cylinder rotates fast enough, in the way computed and measured by Taylor [505] and described in Section 31.8.2. □

Phenomenon 31.41 (The fourth-power law of pipe flow)**.** Steady laminar flow of a viscous liquid along a straight circular tube delivers a volume flux proportional to the applied pressure difference and to the *fourth* power of the tube radius,

$$Q = \frac{\pi R^4}{8\mu}\,\frac{\Delta p}{L}\,, \tag{31.35}$$

inversely proportional to the length of the tube and to the viscosity of the liquid. Halving the radius at fixed pressure drop therefore divides the delivery by sixteen. The law was established by measurement — independently by Hagen in brass tubes [104] and by Poiseuille in fine glass capillaries [105] — before it was derived, and it is accurate enough to serve as a viscometric standard.

Rests on Theorem 31.37.

Derivation. *Derives Phenomenon 31.41.*

Seek a steady, axisymmetric, unidirectional solution $\boldsymbol{v} = v(r)\,\hat{\boldsymbol{z}}$ of the incompressible Navier–Stokes equations in a tube of radius R. Such a field is divergence-free automatically, and its inertial term vanishes identically: $(\boldsymbol{v}\cdot\boldsymbol{\nabla})\boldsymbol{v} = v\,\partial_z(v\,\hat{\boldsymbol{z}}) = 0$, so the nonlinearity plays no part and the solution below is exact rather than approximate. What remains, using the cylindrical Laplacian Equation (7.91), is the axial component

$$0 = -\frac{\mathrm{d}p}{\mathrm{d}z} + \frac{\mu}{r}\frac{\mathrm{d}}{\mathrm{d}r}\left(r\,\frac{\mathrm{d}v}{\mathrm{d}r}\right),$$

in which the first term is independent of r and the second independent of z, so both are the same constant: $\mathrm{d}p/\mathrm{d}z = -\Delta p/L$. Integrating once and discarding the $\ln r$ solution as singular on the axis, then integrating again and imposing the no-slip condition $v(R) = 0$ at the wall,

$$v(r) = \frac{\Delta p}{4\mu L}\left(R^2 - r^2\right), \tag{31.36}$$

the parabolic profile. The flux is

$$Q = \int_0^R v(r)\,2\pi r\,\mathrm{d}r = \frac{\pi\Delta p}{2\mu L}\int_0^R \left(R^2 r - r^3\right)\mathrm{d}r = \frac{\pi R^4}{8\mu}\,\frac{\Delta p}{L}\,,$$

which is Equation (31.35). The fourth power is the signature: two of its powers come from the cross-sectional area and two from the mean speed, which itself grows as R^2 because the wall is the only brake on the flow. The agreement of the measured flux with Equation (31.35), with no fitted constant beyond the viscosity, is the quantitative laboratory test of the no-slip condition: a wall at which the liquid retained a finite tangential speed would deliver a larger flux than this, by the factor computed in Equation (36.8). □

Remark 31.42 (The entry length). Equation (31.36) is the flow far from the inlet. A fluid entering the tube with a flat profile carries the wall's influence inwards by viscous diffusion, over a transverse distance that grows as $\sqrt{\nu t}$ with $t = z/\bar{v}$ the time since entry (Section 31.7.2). The parabolic profile is established when that distance reaches the axis, $\sqrt{\nu z/\bar{v}} \sim R$, i.e. at

$$\frac{L_\mathrm{e}}{D} \sim \mathrm{Re}\,, \qquad \mathrm{Re} = \frac{\rho\bar{v}D}{\mu}\,, \tag{31.37}$$

so the entry length grows in proportion to both the diameter and the Reynolds number. The estimate gives the scaling and not the coefficient, which is measured to be of order 10^{-1}; a tube short enough that L is comparable with L_e delivers less than Equation (31.35), which is the first of the two failures reported in Section 36.2.

Rests on Phenomenon 31.41.

31.4.4 Stokes flow

Definition 31.43 (Creeping flow)**.** When $\mathrm{Re} \ll 1$ the inertial terms of Equation (31.32) are negligible against the viscous term and the equations linearize to the *Stokes equations*

$$\mathbf{0} = -\boldsymbol{\nabla} p + \mu \nabla^2 \boldsymbol{v}\,, \qquad \boldsymbol{\nabla} \cdot \boldsymbol{v} = 0\,, \tag{31.38}$$

in which time appears only through the boundary conditions.

Rests on Theorem 31.37 and Equation (31.42).

Proposition 31.44 (Kinematic reversibility)**.** *If $(\boldsymbol{v}, p)$ solves Equation (31.38) with boundary velocity $\boldsymbol{U}$, then $(-\boldsymbol{v}, -p)$ solves it with boundary velocity $-\boldsymbol{U}$. Reversing the motion of the boundaries therefore reverses the motion of every fluid particle exactly, and a body that repeats a sequence of shapes in reverse order returns the fluid to its initial configuration.*

Rests on Definition 31.43.

Proof. *Derives Proposition 31.44.*

Equation (31.38) is linear and homogeneous in $(\boldsymbol{v}, p)$, so multiplying a solution by -1 gives a solution, and the boundary data are multiplied by -1 with it; uniqueness of the Stokes problem then identifies it as *the* solution. Nothing of the sort holds for Equation (31.32), whose advective term is quadratic and changes sign under $\boldsymbol{v} \longmapsto -\boldsymbol{v}$ in a way the other terms do not. The consequence is observable and is the reason a swimming micro-organism cannot propel itself by any reciprocal stroke: a sequence of shapes retraced in reverse produces no net displacement whatever, however fast it is performed, because the speed cancels out of Equation (31.38) entirely. □

Phenomenon 31.45 (Stokes drag on a sphere)**.** A sphere of radius a moving slowly at speed U through a fluid of viscosity μ experiences a resisting force

$$F = 6\pi \mu a U\,, \tag{31.39}$$

proportional to the first power of the speed and to the first power of the radius, and independent of the density of the fluid [135]. It is the law by which a falling sphere measures a viscosity, by which particles settle out of suspension, and by which the charged oil droplets of Millikan's measurement of the elementary charge were sized [407].

Derivation. *Derives Phenomenon 31.45.*

What can be settled in a few lines is the *form* of Equation (31.39), and it is worth separating that from the numerical factor because the form is what carries the physics. Equation (31.38) is linear in $\boldsymbol{v}$ and the boundary data are linear in U, so the solution, and with it the traction integrated over the sphere, is exactly proportional to U: the drag is linear in the speed, not quadratic, and this is a theorem rather than a small-amplitude approximation. The quantities available are then μ, a and U, whose dimensions $\mathsf{M}\mathsf{L}^{-1}\mathsf{T}^{-1}$, L and $\mathsf{L}\mathsf{T}^{-1}$ admit exactly one combination with the dimensions of force (Theorem D.2), namely $\mu a U$; note that ρ cannot appear, because it has been removed from the equations along with the inertial terms. Hence $F = C\mu a U$ with C a pure number. Solving Equation (31.38) in full for the sphere and integrating the traction gives $C = 6\pi$ [135]; that calculation is longer than a page and is carried out in the appendix. □

Full derivation: Appendix A.60 (p. 3121).

Derives Phenomenon 31.45.

Example 31.46 (Terminal velocity, and Millikan's droplets)**.** A sphere of density ρ_s falling under gravity in a fluid of density ρ_f reaches the speed at which Equation (31.39) balances its weight less the buoyancy of Equation (31.15):

$$U_t = \frac{2a^2\,(\rho_s - \rho_f)\,g}{9\mu}\,. \tag{31.40}$$

For an oil droplet of radius 1 µm and density 900 kg/m^3 in air (Table 31.1) this is $U_t \approx 1.1 \times 10^{-4}$ m/s, about a tenth of a millimetre per second, and the Reynolds number of the motion is then $\rho_f U_t a/\mu \approx 7 \times 10^{-6}$ — amply inside the regime of Definition 31.43, which is what justifies using Equation (31.39) to convert an observed fall speed into a droplet radius [407]. The quadratic dependence on a is what makes sedimentation a size separation: doubling the radius quadruples the settling speed.

Rests on Phenomena 31.18 and 31.45.

Remark 31.47 (The Oseen correction, and why the obvious expansion fails)*.* Equation (31.39) is the leading term of an expansion in Re, and the next term is not obtained by simply iterating Equation (31.38) with the neglected inertia treated as a small forcing. The reason is that the neglect is not uniformly valid: the Stokes velocity disturbance decays only as a/r, so the neglected inertial term $\rho\,(\boldsymbol{v}\cdot\boldsymbol{\nabla})\,\boldsymbol{v}$ falls off more slowly than the retained viscous term $\mu\nabla^2\boldsymbol{v}$, and beyond a distance of order a/Re the ordering is reversed. The naive second approximation then cannot be made to satisfy the condition at infinity at all — Whitehead's paradox. Oseen's repair is to keep the advection by the uniform stream in the far field from the start, giving

$$C_D = \frac{24}{\mathrm{Re}}\left(1 + \frac{3}{16}\mathrm{Re} + \cdots\right), \qquad \mathrm{Re} = \frac{2\rho a U}{\mu}\,, \tag{31.41}$$

with C_D as in Phenomenon 31.55. The structure — an inner region where viscosity dominates and an outer region where it does not, matched to each other — is the same one that resolves the opposite limit in Section 31.7.2, and it is the general method pointed to in Chapter 82.

Rests on Phenomenon 31.45.

31.4.5 Existence and smoothness: an open problem

Remark 31.48 (What is actually proved)*.* Everything in this chapter is a statement about solutions of Equation (31.32), and it is not known that those solutions exist. The position, stated exactly, is this. Leray proved in 1934 that from smooth initial data of finite energy the three-dimensional incompressible equations possess a global *weak* solution — one satisfying the equations in an integrated sense against test functions, and obeying an energy inequality [745]. It is not known whether that solution is unique, and it is not known whether it remains smooth: a weak solution is permitted, as far as anything proved goes, to develop a singularity in finite time at which the velocity gradient blows up. Conversely, a strong solution is known to exist and be unique for a finite time from smooth data, and to persist for all time if the data are small enough; what is missing is the passage between the two regimes. The Clay Millennium problem asks for a proof of global regularity or a counterexample [2094], and neither exists.

Two consequences are owed to the reader. The first is that the turbulence statements of Section 31.9 rest on equations whose well-posedness is open, which is a situation without parallel elsewhere in this treatise — and the empirical adequacy of those equations, which is not in doubt, is no substitute. The second is that the question is not idle: a singularity would be a genuine physical event in the model, at which the continuum hypothesis of Definition 31.2 would have to be abandoned, so the answer matters for what the equations are taken to describe and not only for mathematics. The problem is catalogued with the others in Chapter 133.

Rests on Theorem 31.37.

31.5 Dynamical similarity

31.5.1 The Reynolds number

Phenomenon 31.49 (Dynamical similarity)**.** Two incompressible flows past geometrically similar bodies are observed to be similar to each other — same streamline pattern, same dimensionless force coefficients, same transition points — whenever they agree in the single dimensionless combination

$$\mathrm{Re} = \frac{\rho v L}{\mu}\,, \tag{31.42}$$

whatever the separate values of size, speed, density and viscosity. Drag measurements taken on bodies of very different size and in different fluids collapse onto one curve when plotted against Re [226, 384], and it is this observed collapse, not any argument of principle, that licenses wind-tunnel and towing-tank practice.

Rests on Theorem 31.37.

Derivation. *Derives Phenomenon 31.49.*

Let L and U be the size of the body and the free-stream speed, and measure lengths, velocities, times and pressures in the units L, U, L/U and ρU^2, writing the resulting dimensionless variables with primes. The incompressible Navier–Stokes system

$$\partial_t \boldsymbol{v} + (\boldsymbol{v}\cdot\boldsymbol{\nabla})\boldsymbol{v} = -\frac{1}{\rho}\boldsymbol{\nabla} p + \frac{\mu}{\rho}\nabla^2\boldsymbol{v}\,, \qquad \boldsymbol{\nabla}\cdot\boldsymbol{v} = 0$$

becomes, after dividing throughout by U^2/L,

$$\partial_{t'}\boldsymbol{v}' + (\boldsymbol{v}'\cdot\boldsymbol{\nabla}')\boldsymbol{v}' = -\boldsymbol{\nabla}' p' + \frac{1}{\mathrm{Re}}\nabla'^2\boldsymbol{v}'\,, \qquad \boldsymbol{\nabla}'\cdot\boldsymbol{v}' = 0\,, \tag{31.43}$$

From Equations (31.32) and (31.42) (measuring lengths in L, speeds in U, times in L/U and pressures in ρU^2, then dividing throughout by U^2/L).

with Re as in Equation (31.42) the only parameter left. Geometric similarity of the bodies makes the dimensionless boundary conditions identical as well. Two flows agreeing in Re therefore solve the same dimensionless problem, and if that problem has a unique solution their primed fields coincide; every dimensionless quantity built from them — the drag coefficient, the pressure coefficient, the Strouhal number of any shedding — is a function of Re alone. The same count follows from the Π theorem (Theorem D.2): five dimensional quantities ρ, U, L, μ and the force, in three independent dimensions, leave $5-3=2$ dimensionless groups, one of which may be taken as the drag coefficient and the other as Re [413].

The uniqueness clause is load-bearing and is exactly where the argument stops: two flows at the same Re solve the same problem, but if that problem admits more than one solution — as pipe flow above $\mathrm{Re}\approx 2000$ manifestly does — similarity says nothing about which one is realized. □

Remark 31.50 (Reading the Reynolds number). Equation (31.42) is the ratio of the inertial term $\rho U^2/L$ to the viscous term $\mu U/L^2$ in Equation (31.32), and equivalently the ratio of the time L^2/ν for momentum to diffuse across the flow to the time L/U for it to be carried across. The range spanned in nature is enormous, and it is why one subject contains such different-looking regimes: a swimming bacterium works at $\mathrm{Re}\sim 10^{-5}$, where Proposition 31.44 governs everything; a raindrop falls at $\mathrm{Re}\sim 10^{3}$; a thrown cricket ball flies at $\mathrm{Re}\sim 10^{5}$, close enough to the crisis of Phenomenon 31.55 that its surface finish changes its flight; and an airliner's wing works at $\mathrm{Re}\sim 10^{7}$. No single approximation covers that span, which is the organizing fact of the chapter.

Rests on Phenomenon 31.49.

Definition 31.51 (The other similarity groups)**.** The same nondimensionalization applied to the further terms of the equations of motion produces one dimensionless group per physical effect retained. Those used in this chapter are

$$\mathrm{Fr} = \frac{U}{\sqrt{gL}} \qquad \text{inertia to gravity,} \tag{31.44}$$

$$\mathrm{Ma} = \frac{U}{c} \qquad \text{flow speed to sound speed,} \tag{31.45}$$

$$\mathrm{St} = \frac{fL}{U} \qquad \text{shedding frequency to flow rate,} \tag{31.46}$$

$$\mathrm{We} = \frac{\rho U^2 L}{\gamma} \qquad \text{inertia to surface tension,} \tag{31.47}$$

$$\mathrm{Ra} = \frac{g\alpha\,\Delta T\,d^3}{\nu\kappa} \qquad \text{buoyancy to diffusion,} \tag{31.48}$$

in which c is the sound speed of Equation (31.75), f a shedding or oscillation frequency, γ the surface tension of Equation (31.18), and, for the last, α the thermal expansion coefficient, κ the thermal diffusivity, d the depth of the layer and ΔT the temperature difference imposed across it. All are pure numbers, so that each carries the same value in every system of units (Appendix D).

Rests on Phenomenon 31.49 and Theorem D.2.

Remark 31.52 (Similarity is only as good as the control of the data). Matching Re between a model and a full-scale body is frequently impossible — a model at one tenth scale would need ten times the speed in the same fluid, which changes Ma — and the practical art consists in deciding which group must be matched and which may be violated. Worse, and more instructive: the interior equations do not see the disturbance level of the incoming stream, which is a further dimensionless datum of the experiment. Two flows at the same Re in tunnels of different quality are *not* the same flow, which is the observation Reynolds himself made when patient settling of his tank moved the transition threshold upward without limit [226].

Rests on Phenomenon 31.49.

Phenomenon 31.53 (Transition to turbulence in a pipe)**.** Below a Reynolds number of the order of 2000 the flow of water along a straight pipe is laminar: a filament of dye released on the axis travels the length of the tube as an unbroken coloured thread. Above it the thread breaks up at some point down the tube and mixes across the whole section, and the pressure drop needed to drive a given flux ceases to follow the linear law of Equation (31.35), rising instead roughly as the square of the flux [226]. The threshold is not sharp: with a smooth inlet and a quiet supply the laminar thread survives to much higher Re, so the observed transition depends on the amplitude of the disturbances present and not on Re alone [226, 2169].

Rests on Phenomena 31.41 and 31.49.

Derivation. *Derives Phenomenon 31.53.*

Three statements are contained in the phenomenon and they have three quite different standings. Separating them is the substance of this derivation, because two of the three follow and the third does not.

That the criterion, if there is one, is a condition on a pure number. A straight circular pipe fixes the geometry completely, so by Phenomenon 31.49 the dimensionless problem Equation (31.43) carries the single parameter $\mathrm{Re} = \rho\bar{v}D/\mu$ and nothing else. Any criterion separating the two observed states is therefore a condition on Re alone, and not on the diameter, the speed, the density or the viscosity separately. That is a prediction and it is the one Reynolds tested: he varied the diameter and, by warming the water, the viscosity,

and located the change of state at one value of the combination rather than at one value of any factor in it [226]. What the same argument declines to supply is the number. The uniqueness clause in the proof of Phenomenon 31.49 is exactly what fails here — above the threshold the dimensionless problem possesses two attracting states, and similarity says nothing whatever about which of them is realized.

That the threshold is not the onset of a linear instability. The laminar profile Equation (31.36) is parabolic, so its curvature keeps one sign across the whole section and it possesses no inflexion point. For the plane analogue Theorem 31.77 then forbids inviscid instability outright, and for the circular pipe the computed viscous spectrum is damped at every Reynolds number [2169]: the laminar solution does not lose stability at $\mathrm{Re} \approx 2000$, or anywhere else. Whatever the transition is, it is not a bifurcation of the laminar state, and no linearization about that state can produce the observed threshold. This is the exact complement of Phenomenon 31.78, where the same programme delivers a threshold and a length scale with no adjustable parameter.

That the resistance rises as the square of the flux. The momentum balance on the column of fluid between two sections a distance L apart is exact and assumes no model of the motion: the pressure force on the two faces is carried entirely by the shear on the wall, so

$$\tau_{\mathrm{w}} = \frac{D}{4}\,\frac{\Delta p}{L}\,. \tag{31.49}$$

Below the threshold Equation (31.35) gives $\Delta p/L = 32\mu\bar{v}/D^2$ and hence $\tau_{\mathrm{w}} = 8\mu\bar{v}/D$, linear in the speed because the transverse transport of streamwise momentum is molecular. Above it that transport is carried instead by the Reynolds stress $-\rho\left\langle v_i' v_j' \right\rangle$ of Equation (31.70), in which the viscosity does not appear at all. Grant that the wall stress becomes independent of μ and the quantities left are ρ, $\bar{v}$ and D, from which Theorem D.2 admits exactly one form, $\tau_{\mathrm{w}} = \frac{1}{2} f \rho \bar{v}^2$ with f a pure number; then Equation (31.49) and $Q = \pi D^2 \bar{v}/4$ give

$$\frac{\Delta p}{L} = \frac{2 f \rho \bar{v}^2}{D} = \frac{32 f \rho}{\pi^2 D^5}\, Q^2\,. \tag{31.50}$$

The exponent moves from one to two because the mechanism of transverse momentum transport moves from molecular to inertial, and that is all this argument establishes. It is not exact, and the discrepancy is measurable: f is not quite constant but falls slowly as Re grows in a smooth pipe, so the observed exponent over Reynolds's own range is close to 1.72 rather than 2 [226], and approaches 2 only when the wall is rough enough for the roughness height to replace the viscous scale.

What is not derived is the threshold itself, and Remark 31.54 states what stands in its place. □

Remark 31.54 (This one is not a gap in the writing). The missing third of the derivation above is not a derivation that exists and has not been transcribed. No derivation of the pipe threshold from Equation (31.32) is known. Since the laminar state never loses stability, the turbulent state cannot be reached from it by an infinitesimal disturbance, and the transition is subcritical: it requires a disturbance of finite amplitude, so that the observed threshold is a property of the disturbance environment as much as of the equations. What replaces a stability calculation is a statistical criterion over the localized turbulent patches, or *puffs*, that the flow supports below the threshold. Each puff has a finite probability per unit time of decaying and a finite probability of splitting in two, both measured; the turbulence is sustained indefinitely exactly when splitting outruns decay, and that crossing occurs at $\mathrm{Re} = 2040 \pm 10$ [2169]. That is a sharp and reproducible number, but it is a measured lifetime balance and not a consequence of Equation (31.32) that anyone has derived. The honest statement is that the pipe has two available states and that no theory predicts the

boundary between them; Section 36.3 sets out the measurements a future theory must reproduce.

Rests on Phenomenon 31.53.

31.5.2 Drag laws

Phenomenon 31.55 (Drag, and the failure of ideal flow)**.** Every body moved steadily through a real fluid experiences a force opposing its motion, and the power needed to keep it moving is the product of that force with its speed. Steady potential flow of an inviscid fluid past a finite body predicts exactly zero — Phenomenon 31.30 — so the observation falsifies the inviscid model no matter how small the viscosity of the fluid is made, which is the sharpest failure of a limit in classical physics. The measured drag coefficient of a sphere, $C_{\mathrm{D}} = F/(\frac{1}{2}\rho v^2 \pi a^2)$, falls in inverse proportion to Re in the creeping-flow regime [135], flattens to a plateau of order unity over several decades of Re, and then drops abruptly by a factor of several near $\mathrm{Re} \approx 3 \times 10^5$ — the drag crisis, at which a rougher sphere is measurably *less* resistant than a smooth one [384].

Rests on Phenomena 31.30 and 31.45.

Derivation. *Derives Phenomenon 31.55.*

Three separate statements are contained in the phenomenon, and they have very different standing.

That the ideal model gives zero is Phenomenon 31.30, proved above for the sphere. That the real drag is not zero, and does not tend to zero as $\mu \longrightarrow 0$, follows from the boundary-layer scaling of Section 31.7.2: the viscous stress at the wall is of order $\mu U/\delta$ with $\delta \sim L\,\mathrm{Re}^{-1/2}$, so it falls only as $\mathrm{Re}^{-1/2}$ while the area over which it acts is fixed, and the separated wake contributes a pressure defect that does not fall at all. The limit of vanishing viscosity is not uniform, which is why setting $\mu = 0$ in the equation gives the wrong answer for any μ however small.

That the drag is measurable from the wake alone is the quantitative statement, and it follows from momentum conservation. Take a control volume bounded by a large cylinder around the body, with plane faces far upstream and far downstream where the pressure has returned to p_∞. The flux of x momentum into the volume through the upstream face is $\rho U^2 A$; out through the downstream face it is $\rho \int u^2 \,\mathrm{d}A$; mass conservation requires the deficit $\int (U-u)\,\mathrm{d}A$ to leave through the side, carrying x momentum $\rho U \int (U-u)\,\mathrm{d}A$ with it. The net force on the body is the imbalance,

$$F = \rho \int u\,(U-u)\,\mathrm{d}A\,, \tag{31.51}$$

From Equations (31.10) and (31.19) (balancing the flux of streamwise momentum through a large control surface, the mass deficit leaving through its sides carrying momentum ρU per unit volume).

the integral running over the wake, where alone $u \neq U$. Two things follow: the drag is exactly the momentum deficit the body leaves behind, so a body with no wake has no drag — which is Phenomenon 31.30 again, read backwards — and the drag can be measured by traversing a Pitot tube across the wake, without touching the body at all.

What is *not* derived here is the size of the wake, which requires the boundary layer to be followed to its separation point, and with it both the value of the plateau coefficient and the location and depth of the drag crisis. □

Remark 31.56 (What of the drag curve is owed, and where it is paid). Two of the three things left open above are settled later in this chapter and one is not, and the difference is worth marking here rather than leaving the reader to discover it. The criterion for the wake to exist at all — that the boundary layer can separate only where the external pressure rises along the flow — is derived, from the layer equations at the wall, in Proposition 31.73; the

delay of separation by a layer that has itself become turbulent, and with it the drag crisis and the dimpled golf ball, follow from the same criterion and are given there.

What is not derived, here or there, is any of the numbers. Locating the separation point requires integrating the boundary-layer equations along the surface under the external pressure distribution, and for a bluff body that distribution is itself set by the wake whose size is being sought, so the problem is not closed by the boundary-layer approximation at all. The plateau value $C_{\mathrm{D}} \approx 0.4$ to 0.5 and the Reynolds number $\mathrm{Re} \approx 3 \times 10^5$ at which the crisis occurs are therefore measured quantities [384], reported in Remark 31.57 and in Chapter 36 as measurements, and the chapter claims no derivation of either.

Rests on Phenomenon 31.55 and Proposition 31.73.

Remark 31.57 (The measured curve as the chapter's running test). The drag coefficient of a sphere as a function of Re is one curve spanning nine decades, and each approximation in this chapter is answerable to one stretch of it. Below $\mathrm{Re} \approx 1$ it is $C_{\mathrm{D}} = 24/\mathrm{Re}$, which is Equation (31.39) rewritten, with the Oseen correction Equation (31.41) extending the agreement upward. Through the intermediate decades it flattens to a plateau of order 0.4 to 0.5, which is the momentum-flux estimate $F \sim \rho U^2 a^2$ of Equation (31.51) with a wake comparable to the body. Near $\mathrm{Re} \approx 3 \times 10^5$ it falls abruptly by a factor of several, which is Phenomenon 36.12 operating; that is Eiffel's discovery [384], and the reason a golf ball is dimpled. No part of the curve is derived from first principles across its whole range, and the chapter says so rather than presenting the pieces as a theory.

Rests on Phenomenon 31.55.

31.6 Vorticity

31.6.1 Vortex lines and Helmholtz's theorems

Theorem 31.58 (The vorticity transport equation)**.** *For a fluid of uniform density obeying Equation* (31.32) *with a conservative body force,*

$$\frac{D\boldsymbol{\omega}}{Dt} = (\boldsymbol{\omega} \cdot \boldsymbol{\nabla})\,\boldsymbol{v} + \nu \nabla^2 \boldsymbol{\omega}\,. \tag{31.52}$$

If the density is not uniform the right-hand side acquires the baroclinic *term* $(\boldsymbol{\nabla}\rho \times \boldsymbol{\nabla}p)/\rho^2$*, which vanishes if and only if surfaces of constant density and constant pressure coincide.*

Rests on Theorem 31.37 and Equation (31.5).

Proof. *Derives Theorem 31.58.*

Write Equation (31.32) using Equation (31.4) and $\boldsymbol{g} = -\boldsymbol{\nabla}\Phi$:

$$\frac{\partial \boldsymbol{v}}{\partial t} - \boldsymbol{v} \times \boldsymbol{\omega} = -\boldsymbol{\nabla}\left(\frac{p}{\rho} + \frac{v^2}{2} + \Phi\right) + \nu \nabla^2 \boldsymbol{v}\,.$$

Take the curl. Every gradient dies by Equation (7.128), so the pressure and the body force disappear entirely — the first thing to notice about vorticity is that pressure cannot create it. Expanding $\boldsymbol{\nabla} \times (\boldsymbol{v} \times \boldsymbol{\omega})$ by Equation (31.5) and using $\boldsymbol{\nabla} \cdot \boldsymbol{\omega} = 0$ (a curl is solenoidal, Equation (31.6)) together with $\boldsymbol{\nabla} \cdot \boldsymbol{v} = 0$,

$$\boldsymbol{\nabla} \times (\boldsymbol{v} \times \boldsymbol{\omega}) = (\boldsymbol{\omega} \cdot \boldsymbol{\nabla})\,\boldsymbol{v} - (\boldsymbol{v} \cdot \boldsymbol{\nabla})\,\boldsymbol{\omega}\,,$$

which rearranges to Equation (31.52). For non-uniform density the pressure term is $-\boldsymbol{\nabla} \times (\boldsymbol{\nabla}p/\rho) = -\boldsymbol{\nabla}(1/\rho) \times \boldsymbol{\nabla}p = (\boldsymbol{\nabla}\rho \times \boldsymbol{\nabla}p)/\rho^2$, as stated.

The two terms on the right of Equation (31.52) are the whole dynamics of the subject. The first, $(\boldsymbol{\omega} \cdot \boldsymbol{\nabla})\,\boldsymbol{v}$, is *vortex stretching*: a vortex line lying along a direction in which the fluid is being stretched has its vorticity amplified in proportion, because angular momentum

is conserved while the moment of inertia falls. It is absent in any flow independent of one Cartesian direction, which is why such flows cannot produce the cascade of Section 31.9.2 and why turbulence is an intrinsically three-dimensional phenomenon. The second, $\nu\nabla^2\boldsymbol{\omega}$, diffuses vorticity, with ν as the diffusivity — which is the reading of the kinematic viscosity promised in Remark 31.33. □

Theorem 31.59 (Helmholtz's vortex theorems)**.** *Define the* strength *of a vortex tube as the flux of vorticity across any cross-section of it, equivalently by Theorem 7.100 the circulation around any circuit girdling it once. Then:*

1. *the strength is the same at every cross-section, so a vortex tube cannot end within the fluid but must close on itself or terminate on a boundary;*
2. *in an ideal barotropic fluid with conservative body forces, vortex lines are material lines — they move with the fluid; and*
3. *the strength of a vortex tube is then constant in time.*

The first is kinematic and holds in every flow whatever; the second and third require the hypotheses stated [150].

Rests on Theorems 31.58 and 31.64.

Proof. *Derives Theorem 31.59.*

(1) $\boldsymbol{\nabla}\cdot\boldsymbol{\omega} = 0$ identically by Equation (31.6). Apply the divergence theorem (Theorem 7.101) to the piece of tube between two cross-sections: the side wall is made of vortex lines and carries no flux, so the two end fluxes are equal. If a tube ended in the interior its strength would have to fall to zero there, contradicting that equality.

(2) and (3) follow from Kelvin's theorem (Theorem 31.64), which is proved below independently. Take a material circuit lying in the surface of a vortex tube and not girdling it; the circulation around it is zero at the initial instant, hence zero for all time, so the circuit continues to lie in a surface of vortex lines — the tube surface is a material surface, and since this holds for every tube of every thickness the vortex lines themselves are material. Take next a material circuit girdling the tube once: its circulation is the strength, and Kelvin's theorem says the strength does not change. The three statements are Helmholtz's [150], and they are the reason a vortex ring is a persistent object: it cannot end, it cannot be cut, and its strength cannot decay, until viscosity is restored. □

Example 31.60 (Vortex rings)**.** A puff of fluid ejected from a circular orifice rolls up into a ring of vorticity that then propels itself along its own axis — an arrangement satisfying Theorem 31.59 exactly, since the vortex line closes on itself. The ring's self-induced speed follows from the velocity that the ring's own vorticity induces at its core, and grows as the ring becomes thinner at fixed strength; two coaxial rings pass repeatedly through one another, each contracting and accelerating as it passes through the other and then widening and slowing, a motion which is a direct visual demonstration that the vorticity is being carried by the fluid rather than diffusing through it. The persistence of smoke rings over many diameters is the measured content of Theorem 31.59, and their eventual decay is the measured content of the viscous term in Equation (31.52).

Rests on Theorem 31.59.

Phenomenon 31.61 (The vortex street and its tone)**.** A circular cylinder held across a stream sheds vorticity alternately from its two sides, forming two staggered rows of vortices that travel downstream — the Kármán vortex street. The shedding is periodic, and over the wide range $300 \lesssim \mathrm{Re} \lesssim 2\times10^5$ its frequency f satisfies

$$\mathrm{St} = \frac{f d}{U} \approx 0.2\,, \tag{31.53}$$

nearly independent of Re, where d is the diameter and U the stream speed. Strouhal established the proportionality $f \propto U/d$ by measuring the musical tone emitted by wires in a wind — the Aeolian tone — and obtained a coefficient near 0.185 for fine wires [203]; the value quoted in Equation (31.53) is the modern one for a circular cylinder [1340]. Von Kármán showed that of all the staggered double rows only one spacing is not immediately destroyed by its own induced motion, namely that with the ratio of row separation to streamwise spacing

$$\frac{h}{a} = \frac{1}{\pi}\cosh^{-1}\sqrt{2} \approx 0.281\,, \tag{31.54}$$

which is close to what the photographs show [375].

Rests on Theorem 31.58.

Full derivation: Appendix A.61 (p. 3125).

Derives Phenomenon 31.61.

Remark 31.62 (What the spacing calculation settles, and what it does not). The appendix settles Equation (31.54) and nothing further, and both halves of that sentence matter. It is a theorem about point vortices in an unbounded ideal fluid: the row is idealized as two staggered infinite sequences of line vortices, the cylinder that made them does not appear in the calculation, and what is found is *neutral* stability at one ratio rather than stability, the configuration being unstable at higher order and the observed spacing drifting slowly downstream. That it nonetheless reproduces the photographed ratio to within a few percent is the interesting fact [375].

Equation (31.53) is of a different kind altogether. No derivation from Equation (31.32) predicts the value one fifth. What Phenomenon 31.49 does establish is that the shedding frequency of a given shape can only enter through the dimensionless combination fd/U, and that this must be a function of Re alone; the frequency therefore scales as U/d, which is Strouhal's observation and is why a wire's tone rises with the wind [203]. That the function is nearly flat over three decades of Re, and that its value there is close to 0.2, are measured facts about the coupling between the two boundary layers and the near wake [1340]. They are reported in this chapter as measurements and are not derived in it.

Rests on Phenomena 31.49 and 31.61.

31.6.2 Kelvin's circulation theorem

Definition 31.63 (Circulation)**.** The *circulation* around a closed circuit C is

$$\Gamma := \oint_C \boldsymbol{v}\cdot \mathrm{d}\boldsymbol{x}\,, \tag{31.55}$$

in m^2/s. By Stokes' theorem (Theorem 7.100) it equals the flux of vorticity through any surface spanning C, provided such a surface lies wholly in the fluid.

Rests on Definition 31.8 and Theorem 7.100.

Theorem 31.64 (Kelvin's circulation theorem)**.** *In an ideal barotropic fluid subject to conservative body forces, the circulation around any circuit that moves with the fluid is constant in time,*

$$\frac{D\Gamma}{Dt} = 0\,. \tag{31.56}$$

Rests on Theorem 31.22 and Definition 31.63.

Proof. *Derives Theorem 31.64.*

Let $C(t)$ be a material circuit and parametrize it by a label s carried with the fluid, so that differentiating under the integral sign is legitimate:

$$\frac{D}{Dt}\oint_{C(t)} \boldsymbol{v}\cdot \mathrm{d}\boldsymbol{x} = \oint_C \frac{D\boldsymbol{v}}{Dt}\cdot \mathrm{d}\boldsymbol{x} + \oint_C \boldsymbol{v}\cdot \frac{D\,(\mathrm{d}\boldsymbol{x})}{Dt}\,.$$

The second integral is elementary: a material line element is stretched by the velocity difference across it, so $D(\mathrm{d}\boldsymbol{x})/Dt = \mathrm{d}\boldsymbol{v}$, and $\oint \boldsymbol{v}\cdot\mathrm{d}\boldsymbol{v} = \oint \mathrm{d}\left(\tfrac{1}{2}v^2\right) = 0$ because the circuit is closed. In the first integral substitute Equation (31.19) divided by ρ: with $\rho = \rho(p)$ the pressure term is $-\boldsymbol{\nabla}P$ and with conservative forces the body force is $-\boldsymbol{\nabla}\Phi$, so the integrand is an exact differential and its integral around a closed circuit vanishes by Equation (7.127). Hence Equation (31.56) [176]. □

Corollary 31.65 (Lagrange: an irrotational flow stays irrotational)**.** *If an ideal barotropic fluid under conservative forces is irrotational at one instant throughout a region, it is irrotational in the material image of that region at every later instant.*

Rests on Theorems 31.59 and 31.64.

Proof. *Derives Corollary 31.65.*

By Theorem 7.100 the circulation around every reducible circuit in the region is zero initially, hence by Theorem 31.64 zero at all times, hence the vorticity flux through every small surface vanishes and $\boldsymbol{\omega} = \mathbf{0}$. This is what justifies the potential-flow model of Section 31.3.3 for a body set in motion from rest in still fluid — and, read the other way, it is exactly the difficulty of the lift problem, since it forbids the circulation of Section 31.7.1 from ever appearing. The escape is that the theorem is about a *material* region, and the circuit that would have to be used around an aerofoil sweeps through the boundary layer and the wake, where the hypotheses fail. □

Remark 31.66 (Only three things make circulation)*.* Reading the hypotheses of Theorem 31.64 backwards gives the complete list of ways vorticity can be generated in a fluid initially free of it: viscosity, which enters as the diffusive term of Equation (31.52) and acts overwhelmingly at solid boundaries, where the no-slip condition of Theorem 31.37 is a vorticity source; baroclinicity, when $\boldsymbol{\nabla}\rho$ and $\boldsymbol{\nabla}p$ are not parallel, which is what drives the Rayleigh–Taylor instability of Section 31.8.1 and every buoyancy-driven flow; and non-conservative body forces, which in this chapter do not occur but which in the conducting fluids of Chapter 66 are the Lorentz force. There is no fourth mechanism, and that is a strong statement: every vortex in the atmosphere and the ocean was made at a boundary or by buoyancy.

Rests on Theorem 31.64.

31.7 Lift and the boundary layer

31.7.1 The Kutta–Joukowski theorem

Take the flow to be independent of one Cartesian coordinate, the span, and to have no velocity component along it — the idealization of a wing of large aspect ratio, and the setting in which the complex methods of Chapter 8 apply. Write $z = x + \mathrm{i}y$ for the position in the transverse plane and $w(z) = \phi + \mathrm{i}\psi$ for the complex potential built from the velocity potential of Proposition 31.28 and the stream function of Example 31.15; the Cauchy–Riemann equations (Theorem 8.6) are exactly Equation (31.12) together with $\boldsymbol{v} = \boldsymbol{\nabla}\phi$, so w is holomorphic wherever the flow is irrotational, and

$$\frac{\mathrm{d}w}{\mathrm{d}z} = v_x - \mathrm{i}v_y\,. \tag{31.57}$$

Lemma 31.67 (Blasius' force formula)**.** *Let a body of contour C be held in a steady plane potential flow of a fluid of uniform density ρ, with C traversed once counterclockwise. The force per unit span exerted on it by the fluid pressure is*

$$X - \mathrm{i}Y = \frac{\mathrm{i}\rho}{2}\oint_C \left(\frac{\mathrm{d}w}{\mathrm{d}z}\right)^2 \mathrm{d}z\,. \tag{31.58}$$

Rests on Theorems 8.6 and 31.24.

Proof. *Derives Lemma 31.67.*

For a counterclockwise contour the outward normal is $(\mathrm{d}y, -\mathrm{d}x)\,/\,|\mathrm{d}z|$, so the pressure force on an element is $\mathrm{d}X = -p\,\mathrm{d}y$ and $\mathrm{d}Y = p\,\mathrm{d}x$, whence

$$X - \mathrm{i}Y = -\oint_C p\,(\mathrm{d}y + \mathrm{i}\,\mathrm{d}x) = -\mathrm{i}\oint_C p\,(\mathrm{d}x - \mathrm{i}\,\mathrm{d}y) = -\mathrm{i}\oint_C p\,\overline{\mathrm{d}z}\,.$$

By Equation (31.23) at constant height, $p = B - \tfrac{1}{2}\rho q^2$ with $q^2 = (\mathrm{d}w/\mathrm{d}z)\,\overline{(\mathrm{d}w/\mathrm{d}z)}$; the constant B contributes $-\mathrm{i}B\oint \overline{\mathrm{d}z} = 0$ around a closed contour. Hence

$$X - \mathrm{i}Y = \frac{\mathrm{i}\rho}{2}\oint_C \frac{\mathrm{d}w}{\mathrm{d}z}\,\overline{\left(\frac{\mathrm{d}w}{\mathrm{d}z}\right)}\,\overline{\mathrm{d}z} = \frac{\mathrm{i}\rho}{2}\oint_C \frac{\mathrm{d}w}{\mathrm{d}z}\,\overline{\mathrm{d}w}\,.$$

The body surface is a streamline, so ψ is constant along C and $\mathrm{d}w = \mathrm{d}\phi$ is real there: $\overline{\mathrm{d}w} = \mathrm{d}w = (\mathrm{d}w/\mathrm{d}z)\,\mathrm{d}z$. Substituting gives Equation (31.58). The integrand is holomorphic everywhere in the fluid, so by the deformation theorem (Corollary 8.15) the contour may be pushed out to any circuit enclosing the body — which is what makes the formula usable, since the far field is known even when the body's shape is not. □

Theorem 31.68 (Kutta–Joukowski)**.** *A body held in a steady plane potential flow of speed U along the x axis, around which the circulation is Γ measured counterclockwise, experiences no force along the stream and a transverse force per unit span*

$$Y = -\rho U\Gamma\,, \tag{31.59}$$

that is, a force of magnitude $\rho U\,|\Gamma|$ directed towards the side on which the fluid moves faster [307, 362].

Rests on Lemma 31.67 and Theorem 8.24.

Proof. *Derives Theorem 31.68.*

Far from the body the complex velocity is holomorphic in an annulus and has a Laurent expansion (Theorem 8.21); the constant term is the free stream and the $1/z$ coefficient is fixed by the circulation, since $\oint (\mathrm{d}w/\mathrm{d}z)\,\mathrm{d}z = \Gamma + \mathrm{i}Q$ with Q the volume flux out of the contour, zero for a solid body. Hence

$$\frac{\mathrm{d}w}{\mathrm{d}z} = U - \frac{\mathrm{i}\Gamma}{2\pi z} + \frac{a_2}{z^2} + \cdots\,, \qquad \left(\frac{\mathrm{d}w}{\mathrm{d}z}\right)^2 = U^2 - \frac{\mathrm{i}U\Gamma}{\pi z} + O\!\left(\frac{1}{z^2}\right).$$

Only the $1/z$ term survives the contour integral, by the residue theorem (Theorem 8.24): $\oint (\mathrm{d}w/\mathrm{d}z)^2\,\mathrm{d}z = 2\pi\mathrm{i}\times(-\mathrm{i}U\Gamma/\pi) = 2U\Gamma$. Substituting in Equation (31.58),

$$X - \mathrm{i}Y = \frac{\mathrm{i}\rho}{2}\cdot 2U\Gamma = \mathrm{i}\rho U\Gamma \quad\Longrightarrow\quad X = 0\,, \qquad Y = -\rho U\Gamma\,,$$

which is Equation (31.59). The sign is worth checking physically rather than trusting: a counterclockwise circulation added to a rightward stream speeds the flow up underneath the body and slows it on top, so by Phenomenon 31.25 the pressure is lower underneath and the force is downward — which is what $Y = -\rho U\Gamma < 0$ says for $\Gamma > 0$. That $X = 0$ is Phenomenon 31.30 once more: circulation buys lift and not one newton of drag. □

Remark 31.69 (The Kutta condition, and where the circulation comes from). Theorem 31.68 determines the lift given Γ and says nothing about what Γ is; the potential problem of Proposition 31.28 has a one-parameter family of solutions, one for each circulation, all of them irrotational and all satisfying non-penetration. The physical selection is the *Kutta condition*: of that family, the realized flow is the one for which the velocity at a sharp trailing edge is finite, every other member requiring the fluid to turn the corner at infinite speed. That fixes Γ, and for a thin aerofoil the value it fixes is the content of Proposition 31.70

below — Kutta's computation [307], with the general circulation theorem due to Joukowski, first published in Russian in 1906 [362].

The Kutta condition is an extra physical hypothesis, and its justification is viscous: the boundary layer of Section 31.7.2 cannot negotiate the infinite adverse gradient at the edge and leaves the surface there. Its consistency with Corollary 31.65 is secured by the *starting vortex*: when a wing is set in motion the circulation around a large material circuit enclosing both wing and starting point remains zero, and a vortex of circulation $-\Gamma$ is shed and left behind. It is photographable, and the photographs are the direct evidence that the bound circulation is real.

Rests on Theorem 31.68 and Corollary 31.65.

Proposition 31.70 (The lift-curve slope of a thin aerofoil)**.** *Let a thin, uncambered profile of chord c with a sharp trailing edge be held at a small angle of attack α in a plane potential stream of speed U. Of the one-parameter family of irrotational flows past it, the Kutta condition selects*

$$\Gamma = -\pi c U \alpha\,, \qquad C_{\mathrm{L}} := \frac{Y}{\frac{1}{2}\rho U^2 c} = 2\pi\alpha\,, \tag{31.60}$$

the lift per unit span rising by 2π per radian of incidence, with no dependence whatever on the thickness or on the detailed shape of the section. Thickness, camber and finite span all modify the result, and camber alone shifts the zero-lift angle without altering the slope.

Rests on Theorem 31.68 and Proposition 31.28.

Full derivation: Appendix A.62 (p. 3131).

Derives Proposition 31.70.

31.7.2 The Prandtl boundary layer

Theorem 31.71 (Boundary-layer thickness)**.** *In steady flow of speed U past a body of streamwise extent L at $\mathrm{Re} \gg 1$, viscous effects are confined to a layer against the surface whose thickness satisfies*

$$\frac{\delta}{L} \sim \mathrm{Re}^{-1/2}\,, \tag{31.61}$$

outside which the flow is described by Equation (31.19). The wall shear stress is then of order $\mu U/\delta \sim \rho U^2 \mathrm{Re}^{-1/2}$, which tends to zero with the viscosity, but the dissipation and the drag do not, because the region in which they occur sharpens at the compensating rate [315].

Rests on Theorem 31.37 and Phenomenon 31.30.

Proof. *Derives Theorem 31.71.*

Within the layer the streamwise velocity rises from zero at the wall to U over the transverse distance δ, so transverse gradients exceed streamwise ones by the factor $L/\delta \gg 1$ and the dominant viscous term in Equation (31.32) is $\nu\,\partial_y^2 u \sim \nu U/\delta^2$. The streamwise inertia is $u\,\partial_x u \sim U^2/L$. A layer in which viscosity acts at all is one where these balance, $\nu U/\delta^2 \sim U^2/L$, giving Equation (31.61). The same statement read as a diffusion problem is more transparent: momentum diffuses away from the wall a distance $\sqrt{\nu t}$ in time t, and the time available is the transit time L/U, so $\delta \sim \sqrt{\nu L/U}$, which is Equation (31.61) again — and is the estimate used for the entry length in Remark 31.42.

The scaling settles Phenomenon 31.30 as a matter of principle. The layer disappears as $\mu \longrightarrow 0$, which is why the outer potential solution describes the observed streamlines ever better; but the velocity gradient across it grows as $U/\delta \sim \mathrm{Re}^{1/2}$, so the surface stress falls only as $\mathrm{Re}^{-1/2}$ over a fixed area, and the separated wake it permits contributes a pressure drag that does not fall at all. A limit of that kind cannot be taken by setting the small parameter to zero in the equation, which is exactly what d'Alembert's argument does. Prandtl's water channel made the layer visible and, by removing it through a suction slit, showed that the whole flow around the body is controlled by it (Section 36.5). □

Remark 31.72 (The Blasius solution and its numbers). For the one case that solves completely — a flat plate aligned with a uniform stream, so that the external pressure gradient vanishes — the boundary-layer equations admit a similarity solution in the variable $\eta = y\sqrt{U/\nu x}$, reducing to a single third-order ordinary differential equation [341]. Its numerical solution gives the thickness at which the velocity reaches 99 % of the free stream as $\delta_{99} \approx 5.0\, x\, \mathrm{Re}_x^{-1/2}$, a local friction coefficient $0.664\, \mathrm{Re}_x^{-1/2}$, and a friction drag coefficient $1.328\, \mathrm{Re}_L^{-1/2}$ for one side of a plate of length L. Concretely: a plate 1 m long in water at 1 m/s has $\mathrm{Re}_L = 10^6$ and a layer about 5 mm thick at the trailing edge, half a percent of the plate. The derivation of the similarity reduction is set out in Section 36.5, where the same numbers are used to interpret the streak photographs.

Rests on Theorem 31.71.

Proposition 31.73 (Separation requires an adverse pressure gradient). *In a steady two-dimensional boundary layer the wall curvature of the velocity profile is fixed by the external pressure gradient,*

$$\mu \left(\frac{\partial^2 u}{\partial y^2} \right)_{y=0} = \frac{\mathrm{d}p}{\mathrm{d}x} \,. \tag{31.62}$$

A profile with reversed flow at the wall, and hence a separation point where $(\partial u/\partial y)_{y=0} = 0$, can therefore occur only where $\mathrm{d}p/\mathrm{d}x > 0$ — where the external pressure rises along the flow.

Rests on Theorem 31.71.

Proof. *Derives Proposition 31.73.*

Evaluate the streamwise momentum equation of the layer, $\rho\,(u\partial_x u + v\partial_y u) = -\mathrm{d}p/\mathrm{d}x + \mu\partial_y^2 u$, at the wall, where no slip makes $u = v = 0$: the entire left-hand side vanishes and Equation (31.62) remains. If $\mathrm{d}p/\mathrm{d}x < 0$ the profile is concave towards the wall throughout and $\partial_y u$ cannot vanish inside the layer, so the flow cannot reverse; if $\mathrm{d}p/\mathrm{d}x > 0$ the profile has an inflexion and reversal becomes possible. The criterion is necessary and not sufficient — how far into a rising pressure the layer can go before separating depends on its history — but its necessity already explains the observations of Section 36.5: separation occurs on the rear of a bluff body and in a diverging channel, both of them regions of rising pressure, and never on the accelerating forward face.

The drag crisis follows from the same criterion applied to a layer that has itself become turbulent. A turbulent layer carries fast-moving fluid closer to the wall, so it withstands more adverse pressure before reversing, separates further back, leaves a narrower wake and by Equation (31.51) a smaller drag. That is why roughening a sphere can reduce its resistance [384], and it is Prandtl's explanation of Eiffel's measurement. □

31.8 Instability and transition

31.8.1 Shear and buoyancy instabilities

Theorem 31.74 (The interface dispersion relation). *Let two ideal, incompressible, immiscible fluids be in horizontal layers: fluid 1 of density ρ_1, uniform velocity $U_1\hat{\boldsymbol{x}}$ and depth h_1 above a flat interface at $y = 0$; fluid 2 of density ρ_2, velocity $U_2\hat{\boldsymbol{x}}$ and depth h_2 below it; gravity $-g\hat{\boldsymbol{y}}$; interfacial tension γ; rigid boundaries at $y = h_1$ and $y = -h_2$. Then a small disturbance of the interface proportional to $\exp[\mathrm{i}\,(kx - \omega t)]$, $k > 0$, requires*

$$\rho_1 \coth(kh_1)\,(\omega - kU_1)^2 + \rho_2 \coth(kh_2)\,(\omega - kU_2)^2 = k\left[(\rho_2 - \rho_1)\,g + \gamma k^2\right] . \tag{31.63}$$

The interface is unstable at wavenumber k precisely when this quadratic has a root with $\mathrm{Im}\,\omega > 0$.

Rests on Proposition 31.28 and Theorem 31.24.

Proof. *Derives Theorem 31.74.*

Each layer is irrotational in the undisturbed state and remains so by Corollary 31.65, so write $\boldsymbol{v}_j = U_j \hat{\boldsymbol{x}} + \boldsymbol{\nabla}\phi_j$ with ϕ_j harmonic (Proposition 31.28) and small. Separable harmonic solutions periodic in x with wavenumber k and satisfying $\partial_y \phi = 0$ on the rigid boundaries are

$$\phi_1 = A \cosh\left[k\,(y - h_1)\right] E\,, \qquad \phi_2 = B \cosh\left[k\,(y + h_2)\right] E\,, \qquad E = \mathrm{e}^{\mathrm{i}(kx - \omega t)}\,,$$

and let the interface be $y = \eta = HE$. Write $s_j := \mathrm{i}\,(kU_j - \omega)$, so that $(\partial_t + U_j \partial_x)$ acting on anything proportional to E multiplies it by s_j.

Kinematic condition. A particle on the interface stays on it, so to first order $(\partial_t + U_j \partial_x)\,\eta = \partial_y \phi_j$ at $y = 0$ for each fluid. This gives $s_1 H = -Ak \sinh(kh_1)$ and $s_2 H = Bk \sinh(kh_2)$, whence

$$\phi_1(0) = -\frac{s_1 H}{k} \coth(kh_1)\,, \qquad \phi_2(0) = \frac{s_2 H}{k} \coth(kh_2)\,.$$

Dynamic condition. The unsteady Bernoulli relation Equation (31.21) in each layer, linearized about the uniform stream, gives the pressure at the interface as $p_j = -\rho_j\,(\partial_t + U_j \partial_x)\,\phi_j - \rho_j g \eta + \text{const}$, the constants being equal because the undisturbed interface is in equilibrium. The Young–Laplace relation Equation (31.18) for a surface of small slope reads $p_2 - p_1 = -\gamma\,\partial_x^2 \eta = \gamma k^2 \eta$. Substituting,

$$-\rho_2 s_2 \phi_2(0) + \rho_1 s_1 \phi_1(0) + (\rho_1 - \rho_2)\, gH = \gamma k^2 H\,,$$

and inserting the two expressions for $\phi_j(0)$ together with $s_j^2 = -\,(kU_j - \omega)^2$ gives Equation (31.63) after multiplying through by $-k/H$. The disturbance grows exponentially when ω has a positive imaginary part, and since the two roots of a real quadratic are complex conjugates, instability at a given k is equivalent to the discriminant being negative. □

Corollary 31.75 (Rayleigh–Taylor instability)**.** *For two deep layers at rest, $U_1 = U_2 = 0$ and $kh_j \gg 1$, Equation (31.63) gives*

$$\omega^2 = \frac{k\left[(\rho_2 - \rho_1)\, g + \gamma k^2\right]}{\rho_1 + \rho_2}\,. \tag{31.64}$$

A heavy fluid resting on a light one, $\rho_1 > \rho_2$, is therefore unstable at every wavenumber below the cut-off

$$k_{\mathrm{c}} = \sqrt{\frac{(\rho_1 - \rho_2)\, g}{\gamma}}\,, \qquad \lambda_{\mathrm{c}} = 2\pi \sqrt{\frac{\gamma}{(\rho_1 - \rho_2)\, g}}\,, \tag{31.65}$$

with growth rate $\sqrt{\mathcal{A} g k}$ at long wavelengths, where $\mathcal{A} = (\rho_1 - \rho_2)\,/\,(\rho_1 + \rho_2)$ is the Atwood number [221].

Rests on Theorem 31.74.

Proof. *Derives Corollary 31.75.*

Set $U_j = 0$ and $\coth(kh_j) \longrightarrow 1$ in Equation (31.63) and solve for ω^2, giving Equation (31.64). With $\rho_1 > \rho_2$ the bracket is negative for $k < k_{\mathrm{c}}$, so $\omega^2 < 0$, ω is purely imaginary, and one root grows. Neglecting γ gives $|\omega| = \sqrt{\mathcal{A} g k}$. The cut-off is the physically important part: surface tension stabilizes short wavelengths, so a water surface held above air is stable in a tube narrower than λ_{c}, which for water is a few centimetres — the reason a glass of water inverted over a fine mesh does not empty. □

Corollary 31.76 (Kelvin–Helmholtz instability)**.** *For two deep layers in relative motion, the discriminant of Equation (31.63) is negative — and the interface unstable — precisely when*

$$\rho_1 \rho_2\,(U_1 - U_2)^2 > (\rho_1 + \rho_2)\left[(\rho_2 - \rho_1)\,\frac{g}{k} + \gamma k\right]. \tag{31.66}$$

In particular a vortex sheet in a fluid of uniform density with neither gravity nor surface tension is unstable at every *wavelength, with growth rate* $k\,|U_1 - U_2|\,/2$ *[172, 182].*

Rests on Theorem 31.74.

Proof. *Derives Corollary 31.76.*

Write Equation (31.63) with coth $\longrightarrow 1$ as a quadratic in the phase speed $c = \omega/k$: $(\rho_1 + \rho_2)\,c^2 - 2\,(\rho_1 U_1 + \rho_2 U_2)\,c + \rho_1 U_1^2 + \rho_2 U_2^2 - \gamma k - (\rho_2 - \rho_1)\,g/k = 0$. Its discriminant, divided by four, is

$$(\rho_1 U_1 + \rho_2 U_2)^2 - (\rho_1 + \rho_2)\left(\rho_1 U_1^2 + \rho_2 U_2^2\right) + (\rho_1 + \rho_2)\left[\gamma k + (\rho_2 - \rho_1)\,\frac{g}{k}\right],$$

and the first two terms combine to $-\rho_1\rho_2\,(U_1 - U_2)^2$ identically. The discriminant is negative exactly under Equation (31.66). Setting $\rho_1 = \rho_2 = \rho$ and $g = \gamma = 0$ leaves $c = \frac{1}{2}\,(U_1 + U_2) \pm \frac{\mathrm{i}}{2}\,|U_1 - U_2|$, unstable at every k with growth rate $k\,|U_1 - U_2|\,/2$, which grows without bound as $k \longrightarrow \infty$ — the vortex sheet is an idealization whose short-wave behaviour is regularized in a real fluid by the finite thickness of the shear layer and by viscosity.

Two consequences of Equation (31.66) are worth recording because both are observed. Stable stratification does not confer immunity: with the denser fluid below, the right-hand side is positive but has a minimum over k, so a large enough velocity difference destabilizes some band of wavelengths — which is how billow clouds form on a temperature inversion, and how wind raises waves on water. And with g and γ present the instability sets in first at a definite wavelength, so the billows appear with a selected spacing rather than at all scales at once. □

Theorem 31.77 (Rayleigh's inflexion-point criterion)**.** *Let an ideal fluid flow in a channel with velocity profile* $U(y)$ *parallel to the walls, and consider disturbances proportional to* $\exp\left[\mathrm{i}k\,(x - ct)\right]$ *with transverse amplitude* $\varphi(y)$ *vanishing at both walls. If the flow is unstable* — $\operatorname{Im} c \neq 0$ — *then* $U''(y)$ *must vanish somewhere strictly inside the channel: a necessary condition for instability of an inviscid parallel shear flow is that its velocity profile possess an inflexion point [210].*

Rests on Theorems 31.22 and 31.58.

Proof. *Derives Theorem 31.77.*

Linearizing Equation (31.19) about the parallel flow and eliminating the pressure gives Rayleigh's equation,

$$(U - c)\left(\varphi'' - k^2\varphi\right) - U''\varphi = 0\,, \qquad \varphi\,(y_1) = \varphi\,(y_2) = 0\,. \tag{31.67}$$

Suppose $\operatorname{Im} c \neq 0$. Then $U - c$ never vanishes and the equation may be divided by it. Multiply by the complex conjugate φ^* and integrate across the channel, integrating the first term by parts and using the boundary conditions:

$$-\int_{y_1}^{y_2}\left(|\varphi'|^2 + k^2\,|\varphi|^2\right)\mathrm{d}y - \int_{y_1}^{y_2}\frac{U''\,|\varphi|^2}{U - c}\,\mathrm{d}y = 0\,.$$

The first integral is real. Taking the imaginary part of the second, with $c = c_\mathrm{r} + \mathrm{i}c_\mathrm{i}$ and $1/(U - c) = (U - c_\mathrm{r} + \mathrm{i}c_\mathrm{i})\,/\,|U - c|^2$, leaves

$$c_\mathrm{i}\int_{y_1}^{y_2}\frac{U''\,|\varphi|^2}{|U - c|^2}\,\mathrm{d}y = 0\,.$$

Since $c_\mathrm{i} \neq 0$ by hypothesis and the remaining weight $|\varphi|^2\,/\,|U - c|^2$ is non-negative and not identically zero, U'' must change sign, hence vanish, somewhere in the interval.

The criterion is necessary and not sufficient, and it is worth seeing what it does and does not settle. A free shear layer or a jet has an inflexion point and is violently unstable, which is Corollary 31.76 in a smoothed form. The parabolic pipe and channel profiles of Equation (31.36) have none, so they are stable in the inviscid theory — and they are observed to break down anyway (Phenomenon 31.53), which is the first hint that the pipe transition is not a linear instability at all. □

31.8.2 Rayleigh–Bénard convection and Taylor–Couette flow

Phenomenon 31.78 (Convective onset at a computed threshold)**.** A horizontal fluid layer of depth d heated from below transports heat by conduction alone, at rest, until the imposed temperature difference exceeds a threshold; above it a steady cellular circulation appears whose horizontal periodicity is a fixed multiple of d, independent of the lateral size of the vessel [288]. The control parameter is the Rayleigh number Equation (31.48), and for a layer between stress-free perfectly conducting boundaries Rayleigh's linearized calculation [438] gives

$$\mathrm{Ra}_\mathrm{c} = \frac{27\pi^4}{4} \approx 657.5\,, \qquad \lambda_\mathrm{c} = 2\sqrt{2}\,d \approx 2.83\,d\,, \tag{31.68}$$

with no adjustable parameter whatever; for two rigid boundaries, the commoner laboratory case, the same calculation gives $\mathrm{Ra}_\mathrm{c} \approx 1708$ [1199]. Both threshold and scale are observed.

Rests on Definition 31.51 and Theorem 31.37.

Derivation. *Derives Phenomenon 31.78.*

The reduction of the linearized Boussinesq system to the marginal relation $\mathrm{Ra}(a) = (\pi^2 + a^2)^3 / a^2$ between Rayleigh number and dimensionless horizontal wavenumber, and the minimization of that curve which yields Equation (31.68), are carried out in Section 36.4, where they are set beside the measurement. The point to be taken here is the contrast with Phenomenon 31.53: the same programme — linearize about the base state, seek normal modes, find the parameter value at which the growth rate first reaches zero — delivers a threshold and a length scale in this problem and delivers nothing in the pipe. The failure there is therefore a property of that particular flow, not of the method, and Theorem 31.77 says why: the parabolic profile has no inflexion point, so its instability, if it is one, cannot be of the kind linear theory finds. □

Remark 31.79 (Bénard's own cells, and Taylor's cylinders). Two qualifications belong with Phenomenon 31.78. The first is that Bénard's thin layers had a free upper surface, where the temperature dependence of the surface tension drives a competing instability with its own threshold; the cells he photographed are now attributed largely to that mechanism rather than to buoyancy, so the name is attached to the wrong experiment and the buoyancy instability is properly tested between two plates (Section 36.4). The second is that the calculation predicts only the onset — not the hexagonal plan form, not the amplitude, and nothing about the sequence of further instabilities that follows as the heating is raised, which is the experimental material of Chapter 32.

The analogous problem for the circular Couette flow of Equation (31.34) is Taylor's, and it deserves its priority: rotating the inner cylinder fast enough destabilizes the flow into a stack of toroidal vortices at a critical value of a Taylor number built from the gap, the angular velocity and ν, and Taylor's 1923 comparison of the computed threshold with his measurements was the first quantitative agreement between a hydrodynamic stability calculation and an experiment [505]. The underlying criterion — that the flow is unstable when the angular momentum $r v_\varphi$ decreases outward — is Rayleigh's, and is proved as Proposition 31.80 below; it is stated with its viscous corrections in [1199], and it is not in the convection paper [438] cited above, which is a different problem.

Rests on Phenomenon 31.78 and Proposition 31.40.

Proposition 31.80 (Rayleigh's circulation criterion)**.** *Let an ideal fluid rotate steadily about an axis with azimuthal velocity $v_\varphi(r)$, and write $L(r) := r v_\varphi(r)$ for its angular momentum per unit mass. The motion is unstable to axisymmetric disturbances if the* Rayleigh discriminant

$$\Phi(r) := \frac{1}{r^3}\frac{\mathrm{d}L^2}{\mathrm{d}r} \tag{31.69}$$

is negative anywhere, that is wherever the magnitude of the angular momentum decreases outward. For the circular Couette solution Equation (31.34), $L = Ar^2 + B$ and $\Phi = 4Av_\varphi/r$; so a fluid driven by the inner cylinder with the outer one held at rest, for which $A < 0$ while $v_\varphi > 0$ everywhere except at the outer wall, is unstable at every rotation rate however small.
Rests on Proposition 31.40 and Theorem 31.22.

Proof. *Derives Proposition 31.80.*

Take two thin coaxial rings of equal volume, and hence in an incompressible fluid of equal mass m, at radii $r_1 < r_2$, carrying angular momenta per unit mass L_1 and L_2. Interchange them and leave the rest of the fluid undisturbed. In an ideal axisymmetric motion nothing exerts a torque about the axis on a ring — viscous stresses are absent by hypothesis and the pressure force on a ring is directed through the axis — so each ring carries its own L to its new radius. The two rings have equal volume, so the ambient pressure field does no net work on the pair. The kinetic energy of the azimuthal motion of a ring is $\frac{1}{2}mv_\varphi^2 = \frac{1}{2}mL^2/r^2$, so the interchange costs

$$\Delta E = \frac{m}{2}\left(\frac{L_1^2}{r_2^2} + \frac{L_2^2}{r_1^2} - \frac{L_1^2}{r_1^2} - \frac{L_2^2}{r_2^2}\right) = \frac{m}{2}\left(L_2^2 - L_1^2\right)\left(\frac{1}{r_1^2} - \frac{1}{r_2^2}\right).$$

The second bracket is positive because $r_1 < r_2$. So the interchange must be paid for from outside exactly when L^2 increases outward, and it releases kinetic energy — and will therefore proceed of itself, which is instability — exactly when L^2 decreases outward. Applied to neighbouring radii this is the sign of $\mathrm{d}L^2/\mathrm{d}r$, and dividing by the positive r^3 to make a quantity with the dimensions of the square of a frequency gives Equation (31.69).

For Equation (31.34), $L = r\,(Ar + B/r) = Ar^2 + B$, so $\mathrm{d}L^2/\mathrm{d}r = 2L \cdot 2Ar = 4Ar^2 v_\varphi$ and $\Phi = 4Av_\varphi/r$. With the outer cylinder at rest, $A = -\Omega_1 R_1^2/\left(R_2^2 - R_1^2\right) < 0$ while v_φ falls from $\Omega_1 R_1$ to zero across the gap, so $\Phi < 0$ throughout the interior of the gap and the inviscid flow is unstable for any $\Omega_1 > 0$. Rotating the outer cylinder instead, with the inner one at rest, gives $A > 0$ and $\Phi > 0$: that flow is stable to axisymmetric disturbances at every speed. That asymmetry between driving from inside and driving from outside is what the experiment shows, and it is why Couette's viscometer turns the outer cylinder and measures the torque on the inner one [246]. □

Remark 31.81 (Where the threshold comes from, and what it costs). Proposition 31.80 predicts instability at every rotation rate; the experiment shows a definite threshold, the laminar solution surviving until the inner cylinder is turned fast enough and only then breaking into the stack of toroidal vortices Taylor photographed [505]. The difference is viscosity, which the criterion discards and which both damps the disturbance and diffuses the angular momentum gradient that drives it. Restoring it turns the question into an eigenvalue problem: linearize Equation (31.32) about Equation (31.34) for axisymmetric disturbances proportional to $\mathrm{e}^{\mathrm{i}kz+st}$ and find the rotation rate at which the largest s first reaches zero. In the narrow-gap limit $d := R_2 - R_1 \ll R_1$ the resulting sixth-order system has the same form as the linearized Boussinesq system behind Phenomenon 31.78, with a Taylor number built from d, the angular velocity and ν standing in for the Rayleigh number; its critical value and the axial wavelength of the first vortices — about twice the gap width, so that each vortex is square in section — are computed and tabulated in [1199]. Taylor obtained both and measured both, and the agreement was the first quantitative

confirmation of a hydrodynamic stability calculation [505]. Two things should be carried away. The threshold is a property of the viscous eigenvalue problem and not of the inviscid criterion, which is why no rearrangement of Equation (31.69) will produce it. And it is a computed number rather than a closed formula, which is why this chapter quotes it from the monograph instead of deriving it.

Rests on Proposition 31.80 and Phenomenon 31.78.

31.9 Turbulence

This section is different in kind from those before it, and the difference is stated at the outset rather than left to be inferred. What follows is not a theory of turbulence. It is an averaging procedure that reorganizes Equation (31.32) without closing it, a picture of energy transfer that is suggestive rather than deductive, and a similarity hypothesis that is assumed and then tested. All of it rests on equations whose well-posedness is open (Remark 31.48). The measured regularities are real and are reported; the derivations are labelled by exactly what they assume.

31.9.1 The Reynolds decomposition and the closure problem

Theorem 31.82 (The Reynolds-averaged equations)**.** *Split each field into a mean and a fluctuation, $\boldsymbol{v} = \boldsymbol{V} + \boldsymbol{v}'$ and $p = P + p'$, with $\langle \boldsymbol{v}' \rangle = \boldsymbol{0}$ and the averaging operation linear and commuting with differentiation. Then the mean flow of an incompressible Newtonian fluid obeys*

$$\rho\left(\frac{\partial V_i}{\partial t} + V_j \partial_j V_i\right) = -\partial_i P + \partial_j\left(2\mu E_{ij} - \rho\left\langle v_i' v_j' \right\rangle\right) + \rho g_i\,, \qquad \partial_i V_i = 0\,, \tag{31.70}$$

with E_{ij} the rate of strain of the mean field. The extra term $-\rho\left\langle v_i' v_j' \right\rangle$, the Reynolds stress*, is a new unknown [268].*

Rests on Theorem 31.37.

Proof. *Derives Theorem 31.82.*

Average Equation (31.32) written in conservation form, $\rho\partial_t v_i + \rho\partial_j\left(v_i v_j\right) = -\partial_i p + \mu\nabla^2 v_i + \rho g_i$, which is equivalent to it by Equation (31.11). Every term is linear in the fields except the advective one, and for that

$$\langle v_i v_j \rangle = \left\langle \left(V_i + v_i'\right)\left(V_j + v_j'\right)\right\rangle = V_i V_j + \left\langle v_i' v_j' \right\rangle\,,$$

the cross terms vanishing because $\langle v' \rangle = 0$ and V is already an average. Moving the correlation to the right-hand side and using the averaged continuity equation gives Equation (31.70). Nothing has been approximated: Equation (31.70) is exact. □

Remark 31.83 (The closure problem, stated exactly). Equation (31.70) is four equations for four mean unknowns $(\boldsymbol{V}, P)$ *plus* the six independent components of $\left\langle v_i' v_j' \right\rangle$: ten unknowns, four equations. Deriving an evolution equation for the Reynolds stress from Equation (31.32) is straightforward and produces terms in $\left\langle v_i' v_j' v_k' \right\rangle$; an equation for those produces quadruple correlations, and so on without end. The hierarchy never closes, at any order, and the reason is structural: the nonlinearity of Equation (31.32) is quadratic, so averaging always leaves one more moment than it determines. This is not a temporary gap in knowledge that a cleverer manipulation will fill.

Every practical calculation of a turbulent flow therefore closes the hierarchy by hypothesis. The oldest such hypothesis writes $-\rho\left\langle v_i' v_j' \right\rangle = 2\rho\nu_{\mathrm{T}} E_{ij}$ in imitation of Equation (31.30), defining an *eddy viscosity* ν_{T}. The imitation is formal only. A molecular viscosity is a

property of the fluid, measured once and valid in every flow, because the mean free path is small against the gradients (Definition 31.2); an eddy viscosity is a property of the *flow*, differs from point to point, and is defined by eddies whose size is comparable to the gradients they are supposed to smooth out — the very condition under which the analogy fails. Eddy viscosities and their descendants are engineering correlations fitted to data. They are useful, they are not mechanics, and this treatise does not present them as derived.

Rests on Theorem 31.82.

31.9.2 The energy cascade and Kolmogorov's 1941 theory

Richardson's picture organizes what the closure problem leaves unstructured [494]: energy is fed in at the scale of the geometry that stirs the flow, passed by the nonlinear term to successively smaller eddies with negligible loss, and dissipated by viscosity only at the smallest scales. The picture is consistent with Equation (31.32) — the vortex-stretching term of Equation (31.52) is the mechanism, and its absence in flows independent of one direction is why those flows cascade the other way — but it is not derived from it.

Phenomenon 31.84 (The five-thirds spectrum)**.** In turbulence at high Reynolds number the energy spectrum measured over the intermediate range of scales — small against the geometry that stirs the flow, large against the scale at which viscosity acts — is a power law of exponent $-5/3$,

$$E(k) = C\,\varepsilon^{2/3} k^{-5/3}\,, \tag{31.71}$$

in which ε is the mean rate of energy dissipation per unit mass, in W/kg, and C is a constant of order unity that does not depend on the flow producing the turbulence. The form was predicted from similarity hypotheses [857, 861] and confirmed over a wide range of wavenumbers by the measurements of Grant, Stewart and Moilliet in a tidal channel [1227], whose Reynolds number no laboratory could then reach.

Rests on Theorems 31.58 and D.2.

Derivation. *Derives Phenomenon 31.84.*

The derivation is dimensional, and rests on two hypotheses that are themselves not derived from the Navier–Stokes equations but assumed and then tested: that at scales far below the stirring scale the statistics are locally isotropic and independent of how the turbulence was made, and that in the intermediate range they are independent of the viscosity as well, depending only on the rate ε at which energy is passed down the scales [857]. Granted these, $E(k)$ can depend only on ε and k. In the base dimensions of length L and time T, the spectrum is an energy per unit mass per unit wavenumber and so carries $[\,E\,] = \mathsf{L}^3\mathsf{T}^{-2}$, while ε is an energy per unit mass per unit time, $[\,\varepsilon\,] = \mathsf{L}^2\mathsf{T}^{-3}$, and $[\,k\,] = \mathsf{L}^{-1}$. Setting $E = C\varepsilon^a k^b$ and matching both dimensions,

$$\mathsf{L}^3\mathsf{T}^{-2} = \left(\mathsf{L}^2\mathsf{T}^{-3}\right)^a \left(\mathsf{L}^{-1}\right)^b \implies -3a = -2\,, \quad 2a - b = 3\,,$$

so $a = 2/3$ and $b = -5/3$, which is Equation (31.71); the exponents are fixed uniquely by Theorem D.2 and the only freedom left is the dimensionless C. The same argument locates the lower end of the range: the only length that can be built from ε and the kinematic viscosity ν is $\eta = (\nu^3/\varepsilon)^{1/4}$, below which the second hypothesis must fail because viscosity is precisely what acts there. What the argument does *not* give is any justification of the hypotheses, and measurement has since found small but systematic departures from $-5/3$ at high order — the intermittency corrections — so Equation (31.71) is a leading approximation confirmed to within those, not an identity [1227]. □

Proposition 31.85 (The Kolmogorov scales)**.** *The only combinations of ε and ν with the dimensions of length, velocity and time are*

$$\eta = \left(\frac{\nu^3}{\varepsilon}\right)^{1/4}, \qquad u_\eta = (\nu\varepsilon)^{1/4}, \qquad \tau_\eta = \left(\frac{\nu}{\varepsilon}\right)^{1/2}, \tag{31.72}$$

and the Reynolds number formed from them is exactly unity, which is what identifies η as the scale at which the cascade stops. If the energy is supplied at scale L with velocity U and $\varepsilon \sim U^3/L$, then

$$\frac{L}{\eta} \sim \mathrm{Re}^{3/4}, \qquad \frac{\textit{degrees of freedom}}{\textit{unit volume}} \sim \left(\frac{L}{\eta}\right)^3 \sim \mathrm{Re}^{9/4}. \tag{31.73}$$

Rests on Phenomenon 31.84 and Theorem D.2.

Proof. *Derives Proposition 31.85.*

The dimensions are $[\,\nu\,] = \mathsf{L}^2\mathsf{T}^{-1}$ and $[\,\varepsilon\,] = \mathsf{L}^2\mathsf{T}^{-3}$, and eliminating T between them gives Equation (31.72) uniquely (Theorem D.2). Then $u_\eta\eta/\nu = (\nu\varepsilon)^{1/4}\,(\nu^3/\varepsilon)^{1/4}\,/\nu = 1$. With $\varepsilon \sim U^3/L$ — the statement that the energy U^2 is passed down in the eddy turnover time L/U, which is the cascade hypothesis and not a theorem — $L/\eta = L\,(\varepsilon/\nu^3)^{1/4} \sim (UL/\nu)^{3/4}$, and the number of independent degrees of freedom needed to represent the field is the number of boxes of side η in the volume.

The last estimate is the practical face of the closure problem. At $\mathrm{Re} = 10^6$, a modest laboratory value, Equation (31.73) demands of order 10^{13} degrees of freedom per stirring volume for a computation resolving every scale. For concreteness, in the atmospheric surface layer $\varepsilon \approx 1 \times 10^{-2}\,\mathrm{W/kg}$ and, with air from Table 31.1, Equation (31.72) gives $\eta \approx 0.8\,\mathrm{mm}$ — so the smallest structures in a kilometre-deep flow are under a millimetre across, a scale separation of six decades. □

Remark 31.86 (The one exact result, and the split attribution)*.* The dimensional argument above assumes and does not prove. There is exactly one nontrivial statement about turbulence that *is* derived from Equation (31.32): for homogeneous isotropic turbulence in a statistically steady state, the third moment of the longitudinal velocity increment across a separation r in the inertial range obeys

$$\left\langle \left(\delta u_\parallel\right)^3 \right\rangle = -\tfrac{4}{5}\,\varepsilon\, r\,, \tag{31.74}$$

with no adjustable constant [857]. It follows from the Karman–Howarth relation, itself an exact consequence of the equations under the stated symmetry assumptions, in the limit of vanishing viscosity. Its negative sign is the direction of the cascade made quantitative — energy flows from large scales to small — and it is the one place where the phenomenology touches the equations.

Two attributions should be read correctly. The similarity hypotheses and Equation (31.74) are Kolmogorov's [857]; the spectral form Equation (31.71) in wavenumber is Obukhov's [861], published the same year, and both are customarily filed together under the label "K41". And the constant C in Equation (31.71) is not predicted by any of this: it is measured, close to 1.5 in the three-dimensional normalization used here, and a reader comparing that figure with the literature must first check which of several spectral conventions is in force, since the one-dimensional constant is smaller by a factor of order three.

Rests on Phenomenon 31.84.

Full derivation: Appendix A.63 (p. 3135).

Derives Equation (31.74).

Remark 31.87 (What a theory of turbulence would have to do). Set against the rest of this chapter the position is unusual and should not be softened. The equations are believed correct and are confirmed wherever they can be solved; they are not known to possess smooth solutions; the transition to the turbulent state is not predicted (Phenomenon 31.53); the averaged equations do not close (Remark 31.83); and the statistical description that organizes the measurements is an assumption that measurement has confirmed to leading order and falsified in its details, the intermittency corrections being real. Chapter 36 sets out the specific measured regularities that any future theory will be asked to reproduce before it is asked to predict anything new.

Rests on Remarks 31.48 and 31.83.

31.10 Compressible flow

31.10.1 Sound

Phenomenon 31.88 (Sound travels at the adiabatic speed)**.** Small pressure disturbances propagate through a gas without change of form and without dispersion, at a speed that depends on the state of the gas but not on the amplitude or the frequency of the disturbance. The measured speed in air exceeds by about one part in six the value $\sqrt{p/\rho}$ that Newton obtained by treating the compressions as isothermal [14], and agrees instead with the adiabatic compressibility [59]:

$$c = \sqrt{\left(\frac{\partial p}{\partial \rho}\right)_s} = \sqrt{\frac{\gamma p}{\rho}}\,, \tag{31.75}$$

γ being the ratio of the specific heats, close to $7/5$ for a diatomic gas. The ratio of the two predictions is $\sqrt{\gamma}$, about 1.18 — which is the discrepancy observed.

Rests on Theorems 31.12 and 31.22.

Derivation. *Derives Phenomenon 31.88.*

Write $\rho = \rho_0 + \rho'$, $p = p_0 + p'$ with the primed quantities and the velocity small, and keep first-order terms in the continuity and Euler equations of a fluid at rest:

$$\partial_t \rho' + \rho_0 \,\boldsymbol{\nabla}\cdot\boldsymbol{v} = 0\,, \qquad \rho_0\,\partial_t \boldsymbol{v} = -\boldsymbol{\nabla} p'\,.$$

Take ∂_t of the first and the divergence of the second and subtract: $\partial_t^2 \rho' = \nabla^2 p'$. If the disturbance follows some definite thermodynamic path, $p' = (\mathrm{d}p/\mathrm{d}\rho)\rho'$ along it, and the result is the wave equation $\partial_t^2 \rho' = c^2 \nabla^2 \rho'$ with $c^2 = \mathrm{d}p/\mathrm{d}\rho$ — non-dispersive, as observed, because c carries no reference to the wavelength.

Which path? Newton took the compressions to be isothermal, giving $p \propto \rho$ and $c^2 = p/\rho$ [14]. The justification fails on comparison of two times: heat must diffuse between a compression and the adjacent rarefaction, half a wavelength away, within half a period, and for air at audible frequencies the thermal diffusion length in that time is smaller than the wavelength by orders of magnitude. The compressions are therefore adiabatic, not isothermal. For an ideal gas the adiabat is $p \propto \rho^\gamma$, whence $(\partial p/\partial \rho)_s = \gamma p/\rho$ and Equation (31.75) [59]. Numerically, for air at 20 °C treated as an ideal gas with specific gas constant $287\,\mathrm{J/(kg\,K)}$ and $\gamma = 7/5$, Equation (31.75) gives $c = 343\,\mathrm{m/s}$ against Newton's $290\,\mathrm{m/s}$, and the measured value [1340] is the former. The two answers differ by the factor $\sqrt{\gamma} = \sqrt{7/5} \approx 1.18$ for a diatomic gas, and the measured speed in air selects the adiabatic one. The episode is a clean instance of the pattern this treatise follows throughout: a derivation whose logic is sound fails numerically, and the failure localizes the one assumption — here the thermodynamic path — that has to be replaced. □

Remark 31.89 (The Mach number and the regimes it separates). Equation (31.75) supplies the speed against which Equation (31.45) compares the flow, and the comparison partitions the subject. For $\mathrm{Ma} \ll 1$ the density is effectively constant (Remark 31.14) and the rest of this chapter applies. For $\mathrm{Ma} < 1$ but not small, the flow is compressible but disturbances still travel upstream, so the fluid ahead of a body is forewarned of its approach. For $\mathrm{Ma} > 1$ they cannot: no signal outruns the body, the region it influences is confined to a cone behind it, and the transition between influenced and uninfluenced fluid is the discontinuity of Section 31.10.2. The wave machinery underlying all of this is that of Chapter 28.

Rests on Phenomenon 31.88 and Definition 31.51.

31.10.2 Shock waves

Remark 31.90 (Why a compression wave must steepen). The linearization of Phenomenon 31.88 discards precisely the term responsible for the phenomenon of this subsection. In a one-dimensional simple wave the disturbance at each point travels at $u + c$ rather than at c, and both contributions rise with the local compression: the fluid there is moving forward, and its higher density raises its sound speed. The crest of a compression therefore overtakes the trough ahead of it. Following the characteristics (Proposition 10.15), which for this system are the curves $\mathrm{d}x/\mathrm{d}t = u \pm c$, they cross after a finite time, at which point the smooth solution would become multivalued — an impossibility. What actually forms is a surface across which the fields jump, of thickness a few molecular mean free paths, so thin that the continuum description of Definition 31.2 fails inside it and must be replaced by matching conditions across it.

Rests on Phenomenon 31.88 and Proposition 10.15.

Theorem 31.91 (Rankine–Hugoniot jump conditions)**.** *Let a plane discontinuity be at rest and let the fluid cross it with normal velocity u_1 ahead and u_2 behind. Then conservation of mass, momentum and energy require*

$$\rho_1 u_1 = \rho_2 u_2\,, \qquad p_1 + \rho_1 u_1^2 = p_2 + \rho_2 u_2^2\,, \qquad h_1 + \tfrac{1}{2}u_1^2 = h_2 + \tfrac{1}{2}u_2^2\,, \tag{31.76}$$

with h the specific enthalpy [178, 242]. These three relations, with the equation of state, determine the state behind the shock from the state ahead of it and the shock speed, with no reference whatever to the mechanism operating inside the discontinuity.

Rests on Theorems 31.12 and 31.22.

Proof. *Derives Theorem 31.91.*

Work in the frame in which the discontinuity is stationary and the flow is steady and one-dimensional. Each of mass, momentum and energy obeys a conservation law of the form $\partial_t q + \partial_x F = 0$, which in the steady state reduces to $\mathrm{d}F/\mathrm{d}x = 0$: the flux is the same on both sides however violently q varies between them. Integrating each across a pillbox straddling the surface and shrinking its thickness to zero, only the flux difference survives, and the three fluxes are ρu, $p + \rho u^2$ and $\rho u\left(h + \tfrac{1}{2}u^2\right)$ — the last dividing through by the mass flux, which is common to both sides by the first relation. That gives Equation (31.76). The viscous and conducting terms of Equation (31.31) contribute nothing to the integrated balance because they are gradients of quantities that are equal on the two sides, where the flow is uniform; that is why the result is independent of μ, ζ and the thermal conductivity, although those are exactly what determine the shock's internal structure and thickness. □

Remark 31.92 (Entropy selects the physical branch). Equation (31.76) is symmetric under interchange of the two sides and therefore admits, formally, both compression and rarefaction discontinuities. Only one of them occurs. Computing the entropy from the jump conditions shows that $s_2 > s_1$ for a compression and $s_2 < s_1$ for a rarefaction, so the second law excludes the latter: shocks compress, and a rarefaction spreads out smoothly instead of

steepening — consistently with Remark 31.90, in which it is the compressive part of the wave whose characteristics converge. For a strong shock in a polytropic gas the density ratio saturates at $\rho_2/\rho_1 \longrightarrow (\gamma+1)/(\gamma-1)$, which is 6 for $\gamma = 7/5$: however hard a gas is hit, a single normal shock cannot compress it by more than a factor of six, while the temperature behind it grows without bound.

Rests on Theorem 31.91.

Phenomenon 31.93 (The Mach cone)**.** A body moving through a gas faster than the sound speed is followed by a conical shock surface whose half-angle μ satisfies

$$\sin\mu = \frac{1}{\mathrm{Ma}} = \frac{c}{U}\,, \tag{31.77}$$

so that measuring the angle measures the speed. Mach and Salcher photographed the cone around a supersonic projectile by spark-illuminated schlieren imaging, the first direct observation of a shock in flight [235].

Rests on Remark 31.89 and Theorem 31.91.

Derivation. *Derives Phenomenon 31.93.*

Each point of the trajectory emits a disturbance that spreads as a sphere of radius ct about the point of emission, while the body advances Ut. For $U > c$ the spheres are contained within the envelope that touches all of them, a cone whose apex is the body and whose half-angle satisfies $\sin\mu = ct/Ut$, which is Equation (31.77). For $U < c$ no envelope exists and the disturbance eventually fills the whole space, which is Remark 31.89 restated geometrically. The cone is the locus at which the fluid first learns of the body, so it is where the state jumps, and Theorem 31.91 applies to the component of the velocity normal to it. For a body of finite thickness the attached cone becomes a detached bow shock at speeds close to the sound speed, which is what the photographs of [235] also show. □

31.11 Surface waves and solitary waves

Theorem 31.94 (Gravity–capillary waves on a free surface)**.** *Small-amplitude waves on the free surface of a layer of ideal incompressible liquid of depth h, density ρ and surface tension γ, over a flat rigid bottom, satisfy*

$$\omega^2 = \left(gk + \frac{\gamma k^3}{\rho}\right)\tanh(kh)\,. \tag{31.78}$$

Rests on Theorem 31.74.

Proof. *Derives Theorem 31.94.*

This is Equation (31.63) with the upper fluid removed, $\rho_1 = 0$ and $U_1 = U_2 = 0$, the lower fluid identified as the liquid, $\rho_2 = \rho$ and $h_2 = h$: $\rho\coth(kh)\,\omega^2 = k\,(\rho g + \gamma k^2)$, which rearranges to Equation (31.78). Setting $\gamma = 0$ recovers the classical gravity-wave relation of [701]. The hypotheses inherited from Theorem 31.74 are worth restating because each is visible in the failures of the formula: ideal fluid, so no attenuation is predicted although real waves damp; irrotational motion, so no mean drift; and small amplitude, so no steepening or breaking. □

Phenomenon 31.95 (The observed dispersion of water waves)**.** Waves on deep water are dispersive: long waves travel faster than short ones, so a disturbance spreads into a train with the longest waves at the front, and the energy of the train travels at half the speed of its crests. In shallow water the dispersion disappears and all long waves travel at $\sqrt{gh}$, which is why a tsunami keeps its shape across an ocean. Between the two regimes, surface

tension takes over at short wavelengths and the phase speed passes through a minimum: for a clean water surface

$$c_{\min} = \left(\frac{4g\gamma}{\rho}\right)^{1/4} \approx 0.23\,\mathrm{m/s}\,, \qquad \lambda_{\min} = 2\pi\sqrt{\frac{\gamma}{\rho g}} \approx 1.7\,\mathrm{cm}\,, \tag{31.79}$$

below which the ripples are governed by capillarity and above which by gravity. A steadily moving disturbance slower than $c_{\min}$ therefore generates no waves at all, which is why a slowly drawn wire leaves a still surface.

Rests on Theorem 31.94.

Derivation. *Derives Phenomenon 31.95.*

From Equation (31.78) the phase speed is $c = \omega/k = \sqrt{(g/k + \gamma k/\rho)\tanh(kh)}$. For $kh \gg 1$, $\tanh \longrightarrow 1$; with γ negligible, $c = \sqrt{g/k}$, which decreases with k — long waves are faster — and the group velocity is

$$c_{\mathrm{g}} = \frac{\mathrm{d}\omega}{\mathrm{d}k} = \frac{\mathrm{d}}{\mathrm{d}k}\sqrt{gk} = \tfrac{1}{2}\sqrt{\frac{g}{k}} = \tfrac{1}{2}c\,.$$

For $kh \ll 1$, $\tanh(kh) \approx kh$ and $\omega^2 \approx gh\,k^2$, so $c = c_{\mathrm{g}} = \sqrt{gh}$ independent of k: no dispersion, and the shallow-water speed. For the minimum, the deep-water phase speed $c^2 = g/k + \gamma k/\rho$ is a sum of a decreasing and an increasing term, minimized where their derivatives cancel, $-g/k^2 + \gamma/\rho = 0$, i.e. at $k_{\min} = \sqrt{\rho g/\gamma}$, giving $c^2_{\min} = 2\sqrt{g\gamma/\rho}$ and hence Equation (31.79). With $\gamma = 7.3 \times 10^{-2}\,\mathrm{N/m}$ and $\rho = 998\,\mathrm{kg/m^3}$ these evaluate to $0.23\,\mathrm{m/s}$ and $1.7\,\mathrm{cm}$, both directly observable on a still pond. The factor of one half between energy and crest speed is likewise observable: a group of deep-water waves advances with individual crests appearing at its rear, travelling through it and vanishing at its front. The same factor fixes the half-angle of a ship's wake at $\arcsin(1/3) \approx 19.5°$, independent of the speed of the ship, which is the most easily checked prediction in this section. □

Phenomenon 31.96 (The solitary wave)**.** A single elevation of the free surface of a shallow channel can travel a great distance without change of form and without a train of following waves, at a speed exceeding $\sqrt{gh}$ by an amount fixed by its height. Russell observed such a wave on the Union Canal in 1834, followed it on horseback for over a mile, and reproduced it in a wave tank, where he measured

$$c^2 = g\,(h + a) \tag{31.80}$$

for a wave of amplitude a on water of depth h [116]. The observation was disputed for half a century because linear theory (Theorem 31.94) forbids it: every disturbance should disperse. It is reconciled by the Korteweg–de Vries equation, in which the weak nonlinearity that steepens the wave exactly balances the weak dispersion that spreads it [265].

Rests on Theorem 31.94.

Derivation. *Derives Phenomenon 31.96.*

Long waves of small but finite amplitude on shallow water obey, to leading order in both small parameters a/h and $(h/\ell)^2$ with ℓ the length of the disturbance, the Korteweg–de Vries equation [265]

$$\frac{\partial \eta}{\partial t} + c_0\left(1 + \frac{3\eta}{2h}\right)\frac{\partial \eta}{\partial x} + \frac{c_0 h^2}{6}\,\frac{\partial^3 \eta}{\partial x^3} = 0\,, \qquad c_0 = \sqrt{gh}\,, \tag{31.81}$$

whose two non-trivial terms are exactly the two effects named above: the amplitude-dependent speed $c_0\,(1 + 3\eta/2h)$, which steepens the front, and the third derivative, which is the first correction to $\omega = c_0 k$ obtained by expanding $\tanh(kh)$ in Equation (31.78). Seek a

solution travelling without change of form, $\eta = \eta(x - ct)$. Substituting and integrating twice, with η and its derivatives vanishing at infinity, gives a first-order equation whose solution is

$$\eta = a\,\operatorname{sech}^2\!\left(\frac{x - ct}{\ell}\right), \qquad c = c_0\left(1 + \frac{a}{2h}\right), \qquad \ell^2 = \frac{4h^3}{3a}. \tag{31.82}$$

Two of the three relations are testable and both were tested by Russell. The speed agrees with Equation (31.80) to first order, since $\sqrt{g\,(h+a)} = c_0\sqrt{1 + a/h} \approx c_0\,(1 + a/2h)$, and the expansion is consistent only for $a \ll h$, which is Russell's regime. And the width is *inversely* related to the amplitude, $\ell \propto a^{-1/2}$: a taller solitary wave is narrower and faster, which is the counter-intuitive prediction, and it is what the wave-tank measurements show. The derivation of Equation (31.81) itself from the free-surface equations is longer than a page and is left to the appendix; the modern reading of Equation (31.82) as a soliton, with the collision properties that follow, belongs to Chapter 32. □

Full derivation: Appendix A.64 (p. 3142).
Derives Equation (31.81).

Chapter 32

Nonlinear Dynamics and Chaos

Determinism does not imply predictability. The equations of this part are ordinary differential equations whose solutions exist and are unique, and yet—for the generic nonlinear system—two states closer than any measurement can distinguish separate exponentially fast, so that the deterministic flow forgets its initial condition on a finite horizon. This chapter develops the geometry of phase-space flows, fixed points and their bifurcations, limit cycles, and the Lyapunov exponents that quantify sensitivity; it then follows deterministic chaos from Poincaré's three-body analysis to its experimental record, from convecting fluids and driven pendula to Chua-type circuits. The standard modern text is [1888].

The chapter is the counterpoint to the integrable mechanics of Chapter 23: there the invariant tori organize phase space, and here the KAM theorem states exactly how much of that order survives perturbation, in the symplectic language of Chapter 24. On the dissipative side it supplies what Chapter 31 needs for the transition to turbulence. Chaos is classical determinism at its most honest, and its measured record—period-doubling constants, Lyapunov exponents, fractal dimensions—is as quantitative as any spectroscopy in this book.

Remark 32.1 (What is proved here and what is quoted). Nonlinear dynamics is a branch of mathematics before it is a branch of physics, and this chapter is careful about the difference. Everything that is a computation about a particular system—the amplitude of the van der Pol cycle, the contraction rate of the Lorenz flow, the parameter at which the logistic two-cycle loses stability, the locations of the Kirkwood gaps—is derived here in full. Four general theorems are *quoted*: Poincaré–Bendixson (Theorem 32.30), Hartman–Grobman (Theorem 32.15), the multiplicative ergodic theorem (Theorem 32.74) and KAM (Theorem 32.51). Each is mathematics that belongs in Chapter 9 or in a measure-theoretic chapter that Part II does not yet contain, and each carries a flagged remark saying so where it is used. Three further results—the Hopf bifurcation theorem (Theorem 32.25), the conjugacy of the horseshoe to a shift (Theorem 32.42) and Poincaré's non-integrability theorem (Theorem 32.47)—are likewise quoted, and each says so in its own title. Two further quotations are narrower and are marked where they occur: the existence of the fixed point of the doubling operator, from which Feigenbaum's constants come (Remark 32.88), and the Newhouse–Ruelle–Takens classification of what replaces a destroyed torus (Remark 32.94). Everything quoted is quoted with its primary source where this bibliography has one, and where it does not the attribution is made in words and said to be uncited. Nothing in this chapter's own derivations rests on an unstated result.

Remark 32.2 (Units, and the honest reading of a dimensionless model). The dynamical systems of this subject are almost always written down in dimensionless form—the Lorenz equations, the logistic map, the standard map—and the temptation is to treat the resulting numbers as if they were physical. They are not, and this chapter never quotes one as if it were. Every model below is introduced first in SI variables and then reduced by an *explicit*

scaling, stated in full, so that each dimensionless parameter can be read back as a ratio of measured quantities: the Lorenz σ is the Prandtl number ν/κ of the fluid, its r is the Rayleigh number in units of its critical value, the standard map's K is $K'T^2/I$ built from a kick energy, a kick period and a moment of inertia. Measured quantities—Lyapunov times, cell temperatures, dimensions of attractors reconstructed from data—carry SI units and uncertainties throughout.

32.1 Flows in phase space

32.1.1 Dynamical systems, flows and maps

The object of study is a first-order autonomous system of ordinary differential equations. Any higher-order system is brought to this form by taking the derivatives as further coordinates, and any explicitly time-dependent system by taking the time itself as a further coordinate, so nothing is lost by the restriction.

Definition 32.3 (Dynamical system, phase space, flow)**.** A *continuous dynamical system* on a phase space $U \subseteq \mathbb{R}^n$ is an autonomous first-order system

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}), \qquad \boldsymbol{x} \in U \subseteq \mathbb{R}^n, \tag{32.1}$$

with $\boldsymbol{f}$ continuously differentiable. If $\boldsymbol{x}$ carries the SI units of its physical coordinates, $\boldsymbol{f}$ carries those units divided by s. The *flow* is the map $\varphi_t(\boldsymbol{x}_0) = \boldsymbol{x}(t)$ carrying an initial state to the state t later. A *discrete* dynamical system is an iterated map $\boldsymbol{x}_{n+1} = \boldsymbol{P}(\boldsymbol{x}_n)$; the two are connected by Definition 32.38 below.

Rests on Definitions 9.1 and 13.82.

Proposition 32.4 (The flow is a one-parameter group)**.** *Wherever solutions of Equation* (32.1) *exist for all the times involved,*

$$\varphi_0 = \mathrm{id}, \qquad \varphi_t \circ \varphi_s = \varphi_{t+s}, \qquad \varphi_t^{-1} = \varphi_{-t}. \tag{32.2}$$

Rests on Definition 32.3 and Theorem 9.8.

Proof. *Derives Proposition 32.4.*

Fix $\boldsymbol{x}_0$ and set $\boldsymbol{y}(t) = \varphi_t(\varphi_s(\boldsymbol{x}_0))$ and $\boldsymbol{z}(t) = \varphi_{t+s}(\boldsymbol{x}_0)$. Both solve Equation (32.1)—the second because the system is autonomous, so a time-translated solution is again a solution—and both take the value $\varphi_s(\boldsymbol{x}_0)$ at $t = 0$. Since $\boldsymbol{f}$ is continuously differentiable it is locally Lipschitz (Definition 9.4), so by Picard–Lindelöf (Theorem 9.8) the initial value problem has exactly one solution and $\boldsymbol{y} = \boldsymbol{z}$. The remaining two statements are the cases $s = 0$ and $s = -t$. □

Corollary 32.5 (Trajectories do not cross)**.** *Two distinct trajectories of Equation* (32.1) *never meet, and a single trajectory never crosses itself: a trajectory that returns to a point it has already visited is a closed orbit, traversed periodically forever.*

Rests on Proposition 32.4 and Theorem 9.8.

Proof. *Derives Corollary 32.5.*

If two trajectories shared a point $\boldsymbol{p}$ at times t_1 and t_2, then by Proposition 32.4 each is the flow of $\boldsymbol{p}$ time-translated, hence they are the same trajectory. If one trajectory has $\boldsymbol{x}(t_0+T) = \boldsymbol{x}(t_0)$ with $T > 0$, then φ_T fixes $\boldsymbol{x}(t_0)$ and, by the group property, $\varphi_{t+T} = \varphi_t \circ \varphi_T$ agrees with φ_t on that point for every t: the solution is T-periodic. □

This is why phase portraits in the plane are so rigid, and it is the first hint of the dimension counting that runs through the whole chapter: a bounded, non-self-intersecting curve in a plane has very few places to go.

Theorem 32.6 (Evolution of phase volume)**.** *Let $V(t)$ be the volume of the image $\varphi_t(V_0)$ of a region V_0 of phase space carried along by the flow Equation (32.1). Then*

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \int_{V(t)} \boldsymbol{\nabla}\cdot\boldsymbol{f}\,\mathrm{d}^n x\,. \tag{32.3}$$

Rests on Definition 32.3 and Theorem 7.101.

Proof. *Derives Theorem 32.6.*

In a time $\mathrm{d}t$ a surface element $\mathrm{d}\boldsymbol{S}$ of the boundary is swept through the displacement $\boldsymbol{f}\,\mathrm{d}t$, so it contributes $\boldsymbol{f}\cdot\mathrm{d}\boldsymbol{S}\,\mathrm{d}t$ to the enclosed volume. Summing over the boundary,

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \oint_{\partial V} \boldsymbol{f}\cdot\mathrm{d}\boldsymbol{S}\,,$$

and the divergence theorem (Theorem 7.101) converts the surface integral into Equation (32.3). □

Definition 32.7 (Conservative and dissipative flows)**.** A flow is *volume-preserving*, or *conservative*, where $\boldsymbol{\nabla}\cdot\boldsymbol{f} = 0$, and *dissipative* where $\boldsymbol{\nabla}\cdot\boldsymbol{f} < 0$ throughout the region of interest.

Rests on Theorem 32.6.

Hamiltonian flows are conservative: that is exactly Liouville's theorem (Theorem 24.21), $\operatorname{div} X_{\mathcal{H}} = 0$, and it is what makes the Hamiltonian chaos of Section 32.4 qualitatively different from the dissipative chaos of Section 32.5. A Hamiltonian system can have no attractor at all, because an attractor would be a set of smaller volume than its basin.

Definition 32.8 (Attractor and basin)**.** A compact set A invariant under the flow is an *attractor* if it has a neighbourhood B every point of which is carried into A—$\varphi_t(\boldsymbol{x}) \to A$ as $t \to \infty$ for all $\boldsymbol{x} \in B$—and if A contains a dense orbit, so that it cannot be split into two smaller attractors. The largest such B is the *basin of attraction* of A.

Rests on Definitions 6.9 and 32.3.

Remark 32.9 (The dimension count that locates chaos). Three facts together fix where chaos can live. A one-dimensional flow $\dot{x} = f(x)$ is monotone between its fixed points, so every bounded solution converges to one of them. A two-dimensional flow is confined by Poincaré–Bendixson (Theorem 32.30) to fixed points and closed orbits. Only from three dimensions on can a bounded trajectory avoid both without crossing itself, which by Corollary 32.5 it must. For *maps* the count is one lower, because a map's orbit is a sequence of points and has nothing to cross: a one-dimensional map can already be chaotic, which is what makes the logistic map of Section 32.6.1 the cheapest laboratory in the subject.

The programme itself—study the geometry of the family of solutions rather than hunt for a formula for one of them—is Poincaré's, set out in the memoir on curves defined by a differential equation [216] and carried into celestial mechanics in [247, 255].

32.1.2 Fixed points and Lyapunov stability

Definition 32.10 (Fixed point)**.** A *fixed point* (equilibrium, critical point) of Equation (32.1) is a point $\boldsymbol{x}^*$ with $\boldsymbol{f}(\boldsymbol{x}^*) = \boldsymbol{0}$. It is a trajectory in its own right: $\varphi_t(\boldsymbol{x}^*) = \boldsymbol{x}^*$ for all t.

Rests on Definition 32.3.

Definition 32.11 (Lyapunov stability)**.** A fixed point $\boldsymbol{x}^*$ is *stable in the sense of Lyapunov* if for every $\varepsilon > 0$ there is a $\delta > 0$ such that $|\boldsymbol{x}_0 - \boldsymbol{x}^*| < \delta$ implies $|\varphi_t(\boldsymbol{x}_0) - \boldsymbol{x}^*| < \varepsilon$ for all $t \geq 0$; it is *asymptotically stable* if in addition $\varphi_t(\boldsymbol{x}_0) \to \boldsymbol{x}^*$; and *unstable* otherwise [253].

Rests on Definition 32.10.

The two conditions are independent and both are needed. A point can be stable without being asymptotically stable—the centre of a frictionless oscillator, whose neighbours orbit it forever without approaching—and, less obviously, a point can attract every nearby trajectory while failing to be stable, if the approach takes the trajectory far away first.

Definition 32.12 (Linearization)**.** Write $\boldsymbol{x} = \boldsymbol{x}^* + \delta\boldsymbol{x}$. Taylor's theorem (Theorem 7.38) applied to each component of $\boldsymbol{f}$ gives

$$\dot{\delta\boldsymbol{x}} = A\,\delta\boldsymbol{x} + \mathcal{O}(|\delta\boldsymbol{x}|^2)\,, \qquad A_{ij} = \left.\frac{\partial f_i}{\partial x_j}\right|_{\boldsymbol{x}^*}, \tag{32.4}$$

in which A is the *Jacobian matrix* of $\boldsymbol{f}$ at $\boldsymbol{x}^*$, with SI units of 1/s. The fixed point is *hyperbolic* if no eigenvalue of A has zero real part.

Rests on Definition 32.10 and Theorem 7.38.

Theorem 32.13 (Linear stability)**.** *Let A be diagonalizable with eigenvalues $\lambda_1, \dots, \lambda_n$. If $\Re\lambda_i < 0$ for every i, then $\boldsymbol{x}^*$ is asymptotically stable. If $\Re\lambda_i > 0$ for some i, then $\boldsymbol{x}^*$ is unstable.*

Rests on Definition 32.12 and Theorem 5.73.

Proof. *Derives Theorem 32.13.*

Consider first the linear system $\dot{\boldsymbol{y}} = A\boldsymbol{y}$ alone. Let $\{\boldsymbol{v}_i\}$ be a basis of eigenvectors (Definition 5.65), which exists by hypothesis (Theorem 5.73), and expand $\boldsymbol{y}_0 = \sum_i c_i \boldsymbol{v}_i$. Then

$$\boldsymbol{y}(t) = \sum_i c_i\, \mathrm{e}^{\lambda_i t}\, \boldsymbol{v}_i \tag{32.5}$$

satisfies the system—differentiate term by term and use $A\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$—and takes the value $\boldsymbol{y}_0$ at $t = 0$; by Theorem 9.8 it is the only solution. Its norm is bounded by $C \max_i \mathrm{e}^{(\Re\lambda_i)t}\,|\boldsymbol{y}_0|$ with C the condition number of the eigenbasis, so every solution decays to zero when all $\Re\lambda_i < 0$, and the component along an eigenvector with $\Re\lambda_i > 0$ grows without bound.

The passage from the linear system to Equation (32.4) uses that the neglected terms are $\mathcal{O}(|\delta\boldsymbol{x}|^2)$: choosing δ small enough that the quadratic remainder is smaller than half the linear decay rate on the ball of radius δ, the norm is still strictly decreasing there, which gives both conditions of Definition 32.11. The instability statement follows from the same estimate applied on the unstable eigendirection. □

Remark 32.14 (Two pieces of mathematics quoted here)*.* Theorem 32.13 was stated for diagonalizable A because that is the case Part II supports. The general case—A with a non-trivial Jordan block, whose solutions carry polynomial prefactors $t^k \mathrm{e}^{\lambda t}$—needs the Jordan normal form, which Chapter 5 does not contain. Nothing in this chapter uses the degenerate case: every Jacobian evaluated below is diagonalizable at the parameter values quoted.

Theorem 32.15 (Hartman–Grobman, restated from Part II)**.** *Near a hyperbolic fixed point the flow of Equation (32.1) is topologically conjugate to the flow of its linearization Equation (32.4): there is a homeomorphism of a neighbourhood of $\boldsymbol{x}^*$ onto a neighbourhood of the origin carrying trajectories of the one onto trajectories of the other and preserving both the direction and the parametrization of time [1183]; in particular the fixed point is unstable whenever some eigenvalue of the Jacobian has positive real part. This is Theorem 9.32, written in the notation of this chapter.*

Rests on Theorem 9.32 and Definitions 6.7 and 32.12.

Remark 32.16 (Hartman–Grobman is quoted, not proved)*.* Theorem 32.15 is a theorem of ordinary differential equations and belongs with the Picard–Lindelöf theory of Chapter 9, which states it as Theorem 9.32 and records there, in Remark 9.33, why it is quoted rather

than derived. It is restated here because it is what licenses reading a phase portrait off the eigenvalues, and it is used below only in that qualitative role. Its content is exactly that linearization is trustworthy *and no more than that*: the conjugacy is a homeomorphism, not a diffeomorphism, so it preserves the topology of the portrait—which trajectories go where—but not rates or angles, and it says nothing at all at a non-hyperbolic point. Every bifurcation in Section 32.1.3 happens precisely at a non-hyperbolic point, which is why bifurcation theory cannot be done by linearization alone.

Proposition 32.17 (Classification of planar fixed points)**.** *For $n = 2$ write $\tau = \operatorname{tr} A$ and $\Delta = \det A$. The eigenvalues are*

$$\lambda_{\pm} = \tfrac{1}{2}\left(\tau \pm \sqrt{\tau^2 - 4\Delta}\right), \tag{32.6}$$

and the fixed point is classified as in Table 32.1.

Rests on Definition 32.12 and Theorem 5.67.

Proof. *Derives Proposition 32.17.*

The characteristic polynomial of a 2×2 matrix is $\lambda^2 - \tau\lambda + \Delta$, whose roots (Theorem 5.67) are Equation (32.6). If $\Delta < 0$ the discriminant exceeds τ^2, so the roots are real with opposite signs: a saddle, unstable in one direction and stable in the other. If $\Delta > 0$ and $\tau^2 > 4\Delta$ the roots are real with the sign of τ: a node. If $\Delta > 0$ and $\tau^2 < 4\Delta$ they are a complex conjugate pair with real part $\tau/2$: a focus, spiralling in for $\tau < 0$ and out for $\tau > 0$. If $\tau = 0$ and $\Delta > 0$ they are purely imaginary and the linearized orbits are closed: a centre. Theorem 32.13 then assigns the stabilities. □

Table 32.1: Classification of a hyperbolic fixed point of a planar flow by the trace τ and determinant Δ of its Jacobian, from Equation (32.6). The last row is the one case in which the linearization decides nothing—see Example 32.18.

Condition	Type	Stability
$\Delta < 0$	saddle	unstable
$\Delta > 0$, $\tau^2 > 4\Delta$, $\tau < 0$	node	asymptotically stable
$\Delta > 0$, $\tau^2 > 4\Delta$, $\tau > 0$	node	unstable
$\Delta > 0$, $\tau^2 < 4\Delta$, $\tau < 0$	focus	asymptotically stable
$\Delta > 0$, $\tau^2 < 4\Delta$, $\tau > 0$	focus	unstable
$\Delta > 0$, $\tau = 0$	centre (linear)	undecided

Example 32.18 (A linear centre that is really a stable focus)**.** Take

$$\dot{x} = -y - x^3, \qquad \dot{y} = x - y^3.$$

The Jacobian at the origin is $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, with $\tau = 0$ and $\Delta = 1$: a linear centre. But with $\rho = \tfrac{1}{2}(x^2 + y^2)$,

$$\dot{\rho} = x\dot{x} + y\dot{y} = x(-y - x^3) + y(x - y^3) = -\left(x^4 + y^4\right) < 0$$

everywhere except the origin, so every trajectory spirals in and the origin is asymptotically stable. Changing the sign of the cubic terms makes it unstable, again without touching A. A centre is not hyperbolic, so Theorem 32.15 does not apply, and this is what its hypothesis is protecting against.

Rests on Proposition 32.17 and Theorem 32.15.

Lyapunov's second, or *direct*, method dispenses with linearization altogether [253]. It is the method that survives at a non-hyperbolic point, and in mechanics it is usually the energy.

Definition 32.19 (Lyapunov function)**.** A continuously differentiable V on a neighbourhood N of a fixed point $\boldsymbol{x}^*$ is a *Lyapunov function* if $V(\boldsymbol{x}^*) = 0$, $V > 0$ on $N \setminus \{\boldsymbol{x}^*\}$, and

$$\dot{V} = \boldsymbol{\nabla} V \cdot \boldsymbol{f} \leq 0 \qquad \text{on } N\,. \tag{32.7}$$

It is *strict* if the inequality is strict off $\boldsymbol{x}^*$.

Rests on Definitions 7.65 and 32.10.

Theorem 32.20 (Lyapunov's direct method)**.** *If a Lyapunov function exists on N then $\boldsymbol{x}^*$ is stable; if it is strict, $\boldsymbol{x}^*$ is asymptotically stable.*

Rests on Definitions 32.11 and 32.19.

Proof. *Derives Theorem 32.20.*

Given $\varepsilon > 0$, choose the ball $B_\varepsilon \subseteq N$ and let $m = \min\{V(\boldsymbol{x}) : |\boldsymbol{x} - \boldsymbol{x}^*| = \varepsilon\}$, which is positive because V is positive and continuous on a compact set (Theorem 7.24). By continuity there is $\delta > 0$ with $V < m$ on the ball of radius δ. By Equation (32.7) V never increases along a trajectory, so a trajectory starting inside that ball can never reach the sphere of radius ε, on which $V \geq m$: that is stability. If $\dot{V} < 0$ strictly, then V decreases monotonically along the trajectory and is bounded below, so it converges to some $V_\infty \geq 0$; were $V_\infty > 0$, the trajectory would remain in the compact annulus $\{V_\infty \leq V \leq V(\boldsymbol{x}_0)\}$ on which $\dot{V} \leq -c < 0$, forcing V to fall below V_∞ in finite time. Hence $V_\infty = 0$ and $\boldsymbol{x} \to \boldsymbol{x}^*$. □

Example 32.21 (The damped pendulum)**.** A pendulum of mass m and length ℓ with linear viscous damping of coefficient b (SI units $\mathrm{kg\,m^2/s}$) obeys

$$m\ell^2\ddot{\theta} + b\dot{\theta} + mg\ell\sin\theta = 0\,. \tag{32.8}$$

Take V to be the mechanical energy measured from the hanging position,

$$V = \tfrac{1}{2}m\ell^2\dot{\theta}^2 + mg\ell\,(1 - \cos\theta)\ ,$$

which is positive for $0 < |\theta| < \pi$ and vanishes at the fixed point $(\theta, \dot{\theta}) = (0, 0)$. Differentiating along Equation (32.8),

$$\dot{V} = m\ell^2\dot{\theta}\ddot{\theta} + mg\ell\dot{\theta}\sin\theta = -b\dot{\theta}^2 \leq 0\,,$$

so the hanging position is stable by Theorem 32.20. The inequality is not strict—it fails on the whole line $\dot{\theta} = 0$—so asymptotic stability needs the extra remark that the only trajectory lying entirely in $\{\dot{\theta} = 0\}$ is the fixed point itself, since Equation (32.8) then forces $\sin\theta = 0$. With $b = 0$ the energy is exactly conserved, V is a first integral, and the fixed point is a centre: the isochronous small oscillation of Chapter 28, measured in Chapter 34.

Rests on Theorem 32.20 and Equation (28.8).

32.1.3 Bifurcations

Definition 32.22 (Bifurcation)**.** Let $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}; \mu)$ depend on a control parameter μ. A *bifurcation* occurs at $\mu = \mu_c$ if the qualitative structure of the phase portrait—the number, type or stability of its invariant sets—differs on the two sides of μ_c. By Theorem 32.15 this can happen only where some fixed point or periodic orbit fails to be hyperbolic.

Rests on Definition 32.3 and Theorem 32.15.

The one-dimensional bifurcations come in three normal forms, and the list is short for a reason: expanding $f(x; \mu)$ about the non-hyperbolic point and discarding what does not matter leaves only these possibilities.

Proposition 32.23 (Saddle-node bifurcation)**.** *The system $\dot{x} = \mu - x^2$ has two fixed points $x^* = \pm\sqrt{\mu}$ for $\mu > 0$, of which $+\sqrt{\mu}$ is stable and $-\sqrt{\mu}$ unstable; they merge at $\mu = 0$ and there is no fixed point for $\mu < 0$. The distance between them grows as $\sqrt{\mu}$, and the relaxation rate towards the stable one vanishes as $\sqrt{\mu}$.*

Rests on Definition 32.22 and Theorem 32.13.

Proof. *Derives Proposition 32.23.*

$\mu - x^2 = 0$ has the two real roots $\pm\sqrt{\mu}$ for $\mu > 0$, a double root at $\mu = 0$ and none for $\mu < 0$. The linearization is $f'(x) = -2x$, equal to $-2\sqrt{\mu} < 0$ at $+\sqrt{\mu}$ and $+2\sqrt{\mu} > 0$ at $-\sqrt{\mu}$; Theorem 32.13 in one dimension gives the two stabilities and identifies $2\sqrt{\mu}$ as the relaxation rate, which vanishes at the bifurcation. The slowing down at $\mu \to 0^+$ is generic and observable: it is the *critical slowing down* that announces a saddle-node in a real system before the fixed points actually collide. □

Proposition 32.24 (Transcritical and pitchfork bifurcations)**.** *(i) $\dot{x} = \mu x - x^2$ has fixed points $x^* = 0$ and $x^* = \mu$ for all μ; they exchange stability at $\mu = 0$. (ii) $\dot{x} = \mu x - x^3$ has the single fixed point $x^* = 0$, stable, for $\mu < 0$, and for $\mu > 0$ has $x^* = 0$ unstable together with a symmetric pair $x^* = \pm\sqrt{\mu}$, both stable (*supercritical *pitchfork). (iii) $\dot{x} = \mu x + x^3$ has an unstable pair $x^* = \pm\sqrt{-\mu}$ for $\mu < 0$ which annihilates the stable origin at $\mu = 0$, leaving nothing stable beyond (*subcritical *pitchfork).*

Rests on Definition 32.22 and Theorem 32.13.

Proof. *Derives Proposition 32.24.*

(i) $x(\mu - x) = 0$ gives the two branches; $f'(x) = \mu - 2x$ equals μ at the origin and $-\mu$ at $x^* = \mu$, so the signs are opposite on either side of $\mu = 0$ and the stabilities are exchanged there. (ii) $x(\mu - x^2) = 0$ gives the origin always and $\pm\sqrt{\mu}$ for $\mu > 0$; $f'(x) = \mu - 3x^2$ equals μ at the origin and $-2\mu < 0$ on the outer pair. (iii) is (ii) with $x^3 \to -x^3$: $f' = \mu + 3x^2$ equals μ at the origin and $-2\mu > 0$ on the pair, which therefore exists only for $\mu < 0$ and is unstable. □

The distinction in Proposition 32.24 between the supercritical and subcritical cases is the whole practical content of bifurcation theory in the laboratory. A supercritical transition is continuous and reversible: the new state grows out of the old like $\sqrt{\mu - \mu_c}$, and sweeping the parameter back retraces the same curve. A subcritical transition is discontinuous and hysteretic: the old state loses stability with nothing nearby to fall onto, the system jumps to a distant state, and it stays there when the parameter is swept back past μ_c. Euler buckling (Phenomenon 30.52) is the textbook supercritical pitchfork, its symmetry being the left–right symmetry of the strut; the loss of the convecting rolls in Section 32.5.1 is subcritical, and that is why the Lorenz trajectory has nowhere local to go.

Theorem 32.25 (Hopf bifurcation, quoted)**.** *Let $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}; \mu)$ in $\mathbb{R}^2$ have a fixed point whose Jacobian has a complex conjugate pair of eigenvalues $\lambda_\pm(\mu) = \alpha(\mu) \pm \mathrm{i}\omega(\mu)$ with $\alpha(\mu_c) = 0$, $\omega(\mu_c) \neq 0$ and $\alpha'(\mu_c) \neq 0$. Then, generically, a one-parameter family of periodic orbits is born at μ_c, of period tending to $2\pi/\omega(\mu_c)$ and of amplitude growing as $\sqrt{|\mu - \mu_c|}$ [866].*

Rests on Proposition 32.17 and Definition 32.22.

Proposition 32.26 (The Hopf normal form)**.** *In polar coordinates the normal form of a Hopf bifurcation is*

$$\dot{\rho} = \mu\rho - a\rho^3\,, \qquad \dot{\vartheta} = \omega + b\rho^2\,. \tag{32.9}$$

For $a > 0$ (supercritical) a stable limit cycle of radius $\rho^* = \sqrt{\mu/a}$ exists for $\mu > 0$; for $a < 0$ (subcritical) an unstable cycle of radius $\sqrt{\mu/a}$ exists for $\mu < 0$ and there is nothing local beyond $\mu = 0$.

Rests on Theorem 32.25.

Proof. *Derives Proposition 32.26.*

The radial equation is decoupled and is exactly the pitchfork of Proposition 32.24 restricted to $\rho > 0$: its non-zero fixed point is $\rho^* = \sqrt{\mu/a}$ and its linearization there is $\mu - 3a\rho^{*2} = -2\mu$, negative for $\mu > 0$ when $a > 0$. The angular equation then gives a periodic orbit of angular frequency $\omega + b\mu/a$, which tends to ω as $\mu \to 0$. The $\sqrt{\mu}$ growth of the amplitude in Theorem 32.25 is the $\sqrt{\mu/a}$ here. □

Remark 32.27 (What the Hopf theorem adds, and where its proof lives). Proposition 32.26 is an exact statement about Equation (32.9); Theorem 32.25 is the assertion that a general system satisfying the eigenvalue conditions can be brought to that form by a smooth change of coordinates near the fixed point. That reduction is a theorem of normal-form theory and belongs with the ordinary-differential-equation material of Chapter 9; it is quoted here from [866]. It also requires the implicit function theorem, to follow the fixed point as μ varies; that one *is* available, as Theorem A.286, so the normal-form reduction is the only piece owed. Every Hopf bifurcation located in this chapter is located by the eigenvalue condition $\alpha(\mu_c) = 0$ directly, computed from the actual Jacobian, so none of the derivations below depends on the reduction.

Remark 32.28 (Structural stability and the bifurcation diagram). A system is *structurally stable* if every sufficiently small smooth perturbation of $\boldsymbol{f}$ has a topologically conjugate phase portrait. Hyperbolic fixed points and hyperbolic limit cycles are structurally stable, by Theorem 32.15 and its analogue for periodic orbits; centres and the exact normal forms above are not, which is why a bifurcation is a codimension-one event—it is seen by sweeping one parameter, and not otherwise. What is plotted against that parameter is the *bifurcation diagram*: the location and stability of every invariant set as a function of μ. It is the experimental deliverable of the whole subject. Every measurement in Section 32.7 is a bifurcation diagram in disguise—the Rayleigh number swept in a convection cell, the rotation rate in a Couette cell, the drive amplitude in a circuit—and the universal numbers of Section 32.6.2 are properties of its geometry, not of the system that produced it.

32.2 Limit cycles

32.2.1 The Poincaré–Bendixson theorem

Definition 32.29 (Limit cycle)**.** A *limit cycle* is an isolated closed orbit: a periodic trajectory some neighbourhood of which contains no other closed orbit. It is *attracting* (stable) if all nearby trajectories approach it as $t \to +\infty$, and repelling if they do so as $t \to -\infty$.

Rests on Corollary 32.5 and Definition 32.8.

The isolation is the whole point, and it is what separates a limit cycle from the continuous family of closed orbits surrounding a centre. A conservative system has families; only a system that dissipates in one region and pumps in another can have an isolated one. Limit cycles were named and first analysed in the second instalment of Poincaré's memoir [220].

Theorem 32.30 (Poincaré–Bendixson, restated from Part II)**.** *Let $\boldsymbol{f}$ be continuously differentiable on an open subset of the plane, and let a trajectory of Equation (32.1) remain, for all $t \geq t_0$, inside a compact region containing no fixed point. Then that trajectory either is a closed orbit or approaches one [220, 298]. This is Theorem 9.34, written in the notation of this chapter.*

Rests on Theorem 9.34, Corollary 32.5, and Definition 6.9.

Remark 32.31 (Poincaré–Bendixson is quoted, not proved). The proof is planar topology—a transversal segment, and the Jordan curve theorem, from which the monotonicity of

successive crossings along that segment follows because a trajectory arc together with the chord joining its ends bounds a region the flow cannot re-enter—and it is mathematics that belongs with the ordinary differential equations of Chapter 9, which states it as Theorem 9.34 and sets out in Remark 9.35 exactly which planar facts the proof rests on and which of them this treatise does not carry. It is restated here because the entire architecture of this chapter rests on its corollary: the theorem is what makes the plane too small for chaos, and therefore what sends the subject into three dimensions and into maps.

Corollary 32.32 (No chaos in a planar flow)**.** *A bounded trajectory of a continuously differentiable autonomous flow in the plane cannot be chaotic: as $t \to \infty$ it converges to a fixed point, to a closed orbit, or to a cycle of connecting trajectories between fixed points. Deterministic chaos in a continuous autonomous flow therefore requires at least three phase-space dimensions.*

Rests on Theorem 32.30 and Remark 32.9.

Proof. *Derives Corollary 32.32.*

A bounded trajectory has compact closure. If that closure contains no fixed point, Theorem 32.30 gives convergence to a closed orbit. If it contains fixed points, the limit set consists of those fixed points together with trajectories connecting them, since by Corollary 32.5 nothing else can accumulate without self-intersection in a plane. In neither case do nearby trajectories separate exponentially forever while remaining bounded, which is what Phenomenon 32.72 requires. In three dimensions the argument fails at its first step: a curve may wander indefinitely in a bounded region without crossing itself, because the plane no longer separates it. □

Proposition 32.33 (Bendixson's negative criterion)**.** *If $\boldsymbol{\nabla}\cdot\boldsymbol{f}$ has one strict sign throughout a simply connected region D of the plane, then Equation (32.1) has no closed orbit lying entirely in D [298].*

Rests on Theorem 7.99 and Definition 6.17.

Proof. *Derives Proposition 32.33.*

Let C be a closed orbit in D bounding the region S. Green's theorem (Theorem 7.99) gives

$$\int_S \boldsymbol{\nabla}\cdot\boldsymbol{f}\,\mathrm{d}^2x = \oint_C (f_1\,\mathrm{d}x_2 - f_2\,\mathrm{d}x_1) = \oint_C (f_1\dot{x}_2 - f_2\dot{x}_1)\,\mathrm{d}t = 0\,,$$

because $\dot{x}_1 = f_1$ and $\dot{x}_2 = f_2$ on the orbit make the integrand vanish identically. The left-hand side cannot be zero if the integrand has one strict sign, so no such C exists. The criterion is only negative: it excludes cycles, it never produces one—for that one needs a trapping region and Theorem 32.30. □

Remark 32.34 (Index arguments). A second exclusion principle is topological. The *index* of a closed curve C containing no fixed point is the winding number (Definition 6.20) of the vector field $\boldsymbol{f}$ around it; it is an integer, unchanged by continuous deformation of C (Lemma 6.22), equal to $+1$ for a node, focus or centre and -1 for a saddle, and it is the sum of the indices of the fixed points enclosed. Since a closed orbit is everywhere tangent to $\boldsymbol{f}$, its own index is $+1$. Hence every closed orbit encloses at least one fixed point, and the fixed points it encloses have indices summing to $+1$: a limit cycle cannot surround a single saddle and nothing else, and it cannot lie in a region with no fixed point at all.

32.2.2 Relaxation oscillations: van der Pol

The paradigm of a system with a limit cycle came from radio engineering. A triode circuit in which the effective resistance is negative at small amplitude and positive at large amplitude

obeys, for the voltage v across the tuned circuit,

$$\ddot{v} - \alpha\left(1 - \frac{v^2}{v_0^2}\right)\dot{v} + \omega_0^2 v = 0\,, \qquad \omega_0^2 = \frac{1}{LC}\,, \tag{32.10}$$

in which α has SI units 1/s, v_0 is the voltage at which the damping changes sign, and ω_0 is the resonant angular frequency of the inductance L and capacitance C [555]. Scaling to the dimensionless amplitude $u = v/v_0$ and the dimensionless time $s = \omega_0 t$ puts it in the form used everywhere below,

$$\frac{\mathrm{d}^2 u}{\mathrm{d}s^2} - \varepsilon\left(1 - u^2\right)\frac{\mathrm{d}u}{\mathrm{d}s} + u = 0\,, \qquad \varepsilon = \frac{\alpha}{\omega_0}\,, \tag{32.11}$$

with the single dimensionless control parameter ε. Every statement made about Equation (32.11) converts back by $v = v_0 u$ and $t = s/\omega_0$.

Phenomenon 32.35 (Self-sustained oscillation)**.** A triode circuit whose damping changes sign with amplitude settles into a periodic oscillation whose amplitude and period belong to the circuit alone: started from an arbitrarily small disturbance the oscillation grows until it reaches that amplitude, and started from a larger one it decays to the same amplitude [555]. This is qualitatively unlike the linear oscillator of Chapter 28, whose amplitude is set by the initial conditions and is therefore arbitrary, and unlike a driven resonance, since nothing periodic is applied. In the strongly nonlinear regime the waveform is not sinusoidal at all but a relaxation oscillation of slow charging and abrupt discharge, whose period is set by the slow stage and so scales with the nonlinearity rather than with the resonant frequency of the circuit [555].

Rests on Equation (32.10) *and Definition 32.29.*

Derivation. *Derives Phenomenon 32.35.*

Weak nonlinearity. For $\varepsilon \ll 1$ write Equation (32.11) as a perturbed harmonic oscillator and follow the energy $E = \tfrac{1}{2}(\dot{u}^2 + u^2)$, where the overdot now means $\mathrm{d}/\mathrm{d}s$. Along a solution,

$$\dot{E} = \dot{u}\ddot{u} + u\dot{u} = \varepsilon\left(1 - u^2\right)\dot{u}^2\,. \tag{32.12}$$

Over one cycle the solution is nearly $u = a\cos\psi$ with $\dot{u} = -a\sin\psi$ and ψ advancing by 2π, so averaging Equation (32.12) with $\langle\sin^2\psi\rangle = \tfrac{1}{2}$ and $\langle\cos^2\psi\sin^2\psi\rangle = \tfrac{1}{8}$ gives

$$\left\langle \dot{E} \right\rangle = \varepsilon\left(\frac{a^2}{2} - \frac{a^4}{8}\right)\,,$$

and since $E = a^2/2$ this is an equation for the slowly varying amplitude,

$$\frac{\mathrm{d}a}{\mathrm{d}s} = \frac{\varepsilon a}{2}\left(1 - \frac{a^2}{4}\right)\,. \tag{32.13}$$

Equation (32.13) is the supercritical pitchfork of Proposition 32.24 in the radial variable: for any $a > 0$ the amplitude moves monotonically to $a = 2$, from below if it starts below and from above if it starts above. That is exactly the observation, and in physical variables the cycle has amplitude $v = 2v_0$—twice the voltage at which the damping changes sign, independent of how the oscillation was started and independent of α. Its angular frequency is ω_0 to leading order.

Strong nonlinearity. For $\varepsilon \gg 1$ use the Liénard variable $w = \dot{u}/\varepsilon + F(u)$ with $F(u) = u^3/3 - u$, which turns Equation (32.11) into the first-order pair

$$\dot{u} = \varepsilon\left(w - F(u)\right)\,, \qquad \dot{w} = -\frac{u}{\varepsilon}\,. \tag{32.14}$$

Since ε is large, u moves fast unless $w \approx F(u)$; the trajectory therefore jumps horizontally onto a branch of the cubic $w = F(u)$ in a time of order ε^{-1} and then crawls along it. A branch is attracting when $\partial\dot{u}/\partial u = -\varepsilon F'(u)$ is negative, that is when $F'(u) = u^2 - 1 > 0$: the two outer branches $u > 1$ and $u < -1$. On such a branch $\dot{w} = F'(u)\dot{u} = (u^2 - 1)\dot{u}$ equals $-u/\varepsilon$, so

$$\frac{\mathrm{d}s}{\mathrm{d}u} = -\frac{\varepsilon\,(u^2 - 1)}{u}\,.$$

On the branch $u > 1$ the trajectory enters at $u = 2$—where the previous jump deposited it, since $F(2) = F(-1) = 2/3$—and creeps down to the fold at $u = 1$, at which the branch turns and it jumps across to $u = -2$, because $F(-2) = F(1) = -2/3$. The time for one slow stage is therefore

$$s_{\mathrm{slow}} = \varepsilon \int_1^2 \frac{u^2 - 1}{u}\,\mathrm{d}u = \varepsilon \left[\frac{u^2}{2} - \ln u\right]_1^2 = \varepsilon\left(\frac{3}{2} - \ln 2\right).$$

By the symmetry $u \to -u$, $w \to -w$ of Equation (32.14) the cycle consists of two such stages and two jumps, and the jumps together with the turning regions at the folds contribute only at order $\varepsilon^{-1/3}$. Hence

$$s_{\mathrm{period}} = (3 - 2\ln 2)\,\varepsilon + \mathcal{O}(\varepsilon^{-1/3})\,, \qquad T = \frac{(3 - 2\ln 2)\,\alpha}{\omega_0^2}\,, \tag{32.15}$$

using $3 - 2\ln 2 = 1.6137$. The physical period grows in proportion to the nonlinearity α and falls as the *inverse square* of the resonant angular frequency—it is not the resonant period $2\pi/\omega_0$ at all, which is the observation stated. The waveform is likewise not sinusoidal: two slow ramps of duration $\propto \varepsilon$ separated by jumps of duration $\propto \varepsilon^{-1}$. □

Remark 32.36 (Existence and uniqueness of the cycle). The derivation above establishes the cycle asymptotically at each end of the range of ε. That Equation (32.11) has exactly one closed orbit, attracting every trajectory but the fixed point at the origin, and for *every* $\varepsilon > 0$, is Liénard's theorem (A. Liénard, *Revue générale de l'électricité* **23** (1928) 901–912 and 946–954). **That paper is not in this bibliography and the statement is therefore uncited here**; the reader should treat the global uniqueness as quoted, while the amplitude Equation (32.13) and the period Equation (32.15) are derived above and stand on their own. Note that the key entry `Lienard:1898` in this bibliography is Liénard's electrodynamics paper on the potentials of a moving charge, a different work by the same author, and must not be cited for this theorem.

Phenomenon 32.37 (Frequency demultiplication and the first heard chaos)**.** A neon-tube relaxation oscillator driven by a sinusoidal voltage does not follow the drive: it locks to a submultiple of the drive frequency, and as a circuit parameter is varied smoothly the observed frequency descends in a staircase of whole-number ratios—van der Pol and van der Mark followed it from the fundamental down to one fortieth. At the transitions between two adjacent steps "often an irregular noise is heard in the telephone receivers before the frequency jumps to the next lower value" [590]. That noise is not a defect of the apparatus and not thermal: it is the first recorded laboratory encounter with deterministic chaos.

Rests on Phenomenon 32.35.

Derivation. *Derives Phenomenon 32.37.*

The staircase follows from the mechanism of the oscillator. A neon tube in parallel with a capacitance C charged through a resistance R from a supply V_s has capacitor voltage

$$v(t) = V_s + (V_e - V_s)\,\mathrm{e}^{-t/RC}$$

after each discharge, where V_e is the extinction voltage. The tube fires when v reaches the breakdown voltage V_b, so the free period is

$$T_0 = RC \ln\!\left(\frac{V_s - V_e}{V_s - V_b}\right), \tag{32.16}$$

proportional to C, which is the parameter van der Pol and van der Mark varied. Superposing a small sinusoidal voltage of period T_d modulates the instant at which the threshold is reached: the tube can only fire near a maximum of the drive, so the realized period is an integer multiple nT_d, with n the number of drive periods needed for the ramp to reach V_b. As C is increased, T_0 grows continuously by Equation (32.16) while n can only change by whole numbers, so the observed frequency $1/(nT_d)$ falls in steps: the demultiplication. Each step is a locked periodic orbit, stable over a finite interval of C—the entrainment band, or Arnold tongue, of Section 32.6.3.

The irregular noise sits at the boundaries, where two locking ratios n and $n+1$ are both marginal. There the return map of Section 32.3.1 has no stable fixed point of either period, and the sequence of firing phases neither repeats nor converges. The full account of that regime needs the circle map and is given in Section 32.6.3; what is derived here is the staircase, and the location of the noise at its risers. □

Entrainment of this kind is generic to limit-cycle oscillators and is the nonlinear counterpart of the linear resonance of Chapter 28: a linear oscillator responds at the drive frequency and at no other, whereas a limit-cycle oscillator has a frequency of its own and either captures the drive, is captured by it, or does neither.

32.3 Poincaré sections and discrete dynamics

32.3.1 The return map

Definition 32.38 (Poincaré section and return map)**.** Let Σ be a hypersurface of dimension $n-1$ in the phase space of Equation (32.1), transverse to the flow (that is, $\boldsymbol{f}$ is nowhere tangent to Σ). The *first-return*, or *Poincaré*, *map* $\boldsymbol{P} : \Sigma \to \Sigma$ sends a point $\boldsymbol{x} \in \Sigma$ to the point at which the trajectory through $\boldsymbol{x}$ next meets Σ in the same sense [255]. A fixed point of $\boldsymbol{P}$ is a periodic orbit of the flow; a period-k orbit of $\boldsymbol{P}$ is a periodic orbit piercing Σ k times.

Rests on Definition 32.3 and Corollary 32.5.

This is the device that trades a flow for a map one dimension down, and with it the whole theory of iterated maps becomes available for continuous systems. It is also what makes the numerics of the subject tractable and the figures readable: a three-dimensional flow becomes a scatter of points in a plane.

Proposition 32.39 (Stability of a periodic orbit)**.** *Let $\boldsymbol{x}^* \in \Sigma$ be a fixed point of $\boldsymbol{P}$ and let $\mu_1, \dots, \mu_{n-1}$ be the eigenvalues of $D\boldsymbol{P}(\boldsymbol{x}^*)$, the* characteristic multipliers *of the corresponding periodic orbit. If $|\mu_i| < 1$ for all i the orbit is asymptotically stable; if $|\mu_i| > 1$ for some i it is unstable.*

Rests on Definition 32.38 and Theorem 5.73.

Proof. *Derives Proposition 32.39.*

Linearizing $\boldsymbol{P}$ about $\boldsymbol{x}^*$ as in Definition 32.12 gives $\delta\boldsymbol{x}_{k+1} = D\boldsymbol{P}\,\delta\boldsymbol{x}_k$ to first order, so $\delta\boldsymbol{x}_k = (D\boldsymbol{P})^k \delta\boldsymbol{x}_0$. Expanding in an eigenbasis of $D\boldsymbol{P}$ the i-th component is μ_i^k times its initial value, which tends to zero for $|\mu_i| < 1$ and diverges for $|\mu_i| > 1$. The passage from the linearized map to the full one repeats the estimate at the end of the proof of Theorem 32.13, with the quadratic remainder dominated on a small enough neighbourhood. A multiplier crossing the unit circle is a bifurcation of the periodic orbit: at $\mu = +1$ a saddle-node

of cycles, at $\mu = -1$ a *period-doubling*—the orbit that will drive Section 32.6—and at a complex conjugate pair crossing, the birth of a two-torus. □

Proposition 32.40 (A Hamiltonian section preserves area)**.** *For an autonomous Hamiltonian system with two degrees of freedom, a Poincaré section taken inside a fixed energy surface has a two-dimensional return map that preserves area.*

Rests on Theorems 22.22 and 24.21.

Proof. *Derives Proposition 32.40.*

Take coordinates (q^1, p_1, q^2, p_2) and let Σ be the surface $q^2 = \text{const}$ within $\mathcal{H} = E$, coordinatized by (q^1, p_1). The flow is canonical, so by Poincaré's invariant (Theorem 22.22) the integral $\int \mathrm{d}q^a\,\mathrm{d}p_a$ over a two-surface carried by the flow is conserved. Take a tube of trajectories whose ends are two regions S_0 and $S_1 = \boldsymbol{P}(S_0)$ of Σ. On Σ the coordinate q^2 is constant, so $\mathrm{d}q^2 = 0$ there and the invariant reduces to $\int \mathrm{d}q^1\,\mathrm{d}p_1$; on the lateral surface of the tube the same invariant vanishes because the surface is generated by trajectories. Hence $\int_{S_0} \mathrm{d}q^1\,\mathrm{d}p_1 = \int_{S_1} \mathrm{d}q^1\,\mathrm{d}p_1$: the map preserves area. Liouville's theorem (Theorem 24.21) is the four-dimensional statement of which this is the section. □

Proposition 32.40 is the reason the surfaces of section of Section 32.4.4 look the way they do: an area-preserving map cannot have an attractor, so the points either lie on invariant curves or wander over an area, and never spiral into anything.

32.3.2 The Smale horseshoe and symbolic dynamics

Chaos in a map has a certified mechanism, and it is geometrical rather than analytical: stretch, fold, and put back.

Definition 32.41 (The horseshoe map)**.** Let Q be the unit square. The *horseshoe map* H stretches Q uniformly by a factor $\lambda > 2$ in one direction, contracts it by $\eta < 1/2$ in the other, bends the resulting strip into a horseshoe, and lays it back across Q so that $H(Q) \cap Q$ consists of two full horizontal strips [1359]. The *invariant set* is

$$\Lambda = \bigcap_{k=-\infty}^{\infty} H^k(Q)\,, \tag{32.17}$$

the set of points whose entire forward and backward orbit stays in Q.

Rests on Definition 32.3.

Theorem 32.42 (The horseshoe is a full shift, quoted)**.** *Λ is a compact, totally disconnected, perfect set—a Cantor set—and H restricted to Λ is topologically conjugate to the two-sided shift σ on the space Σ_2 of bi-infinite binary sequences, $(\sigma s)_k = s_{k+1}$, the conjugacy assigning to each point the record of which of the two strips its orbit occupies at each time [1359].*

Rests on Definitions 6.9 and 32.41.

Proposition 32.43 (What the shift does)**.** *The shift σ on Σ_2 has exactly 2^k points of period k for every $k \geq 1$, hence a countable infinity of periodic orbits; it has orbits that are dense in Σ_2; and it has sensitive dependence, two sequences agreeing on $|j| \leq N$ being separated after N steps.*

Rests on Theorem 32.42.

Proof. *Derives Proposition 32.43.*

A sequence with $\sigma^k s = s$ is determined by the k entries $s_0, \ldots, s_{k-1}$ and there are 2^k of these, each giving one such sequence. For a dense orbit, list the finite binary words in order of length—0, 1, 00, 01, 10, 11, 000, ...—and concatenate them into one half-infinite string, completing it to the left arbitrarily; some forward shift of the result begins with any prescribed finite word, and agreement on a long initial word is exactly closeness in Σ_2. For

sensitivity, take s and s' agreeing on $|j| \le N$ and differing at $j = N+1$: after $N+1$ shifts they differ in the zeroth place, which is the coarsest distinction the metric makes. Under the conjugacy of Theorem 32.42 every one of these statements is a statement about H on Λ. □

Remark 32.44 (Topological entropy, and where horseshoes come from). The count 2^k in Proposition 32.43 is the horseshoe's signature: the number of distinguishable orbit segments of length k grows like e^{hk} with $h = \ln 2$, and that exponential growth rate of distinguishable histories is the *topological entropy*. A positive topological entropy is one of the standard definitions of chaos, and for the horseshoe it is exact.

The reason the construction matters physically is that horseshoes are not rare. Wherever the stable and unstable manifolds of a saddle intersect transversally—a *transverse homoclinic point*—some iterate of the map contains a horseshoe, and therefore all of Proposition 32.43. That is the Smale–Birkhoff homoclinic theorem, quoted here from [1359]. Its importance for this chapter is that Poincaré had already found such an intersection, and its infinitely tangled consequences, in the restricted three-body problem [247]: the tangle of Section 32.4.1 and the horseshoe are the same object, seen sixty years apart.

Remark 32.45 (A measure-theoretic gap). Theorem 32.42 calls Λ a Cantor set, and Section 32.5.4 will assign it a dimension. The construction of the Cantor set, its uncountability, and the fact that it has Lebesgue measure zero are results of real analysis and measure theory that Chapters 6 and 7 do not carry; they are used descriptively here and nothing below depends on more than the counting argument of Proposition 32.43, which is elementary.

32.4 Hamiltonian chaos

32.4.1 The three-body problem

Definition 32.46 (The circular restricted three-body problem)**.** Two *primaries* of masses M_1 and M_2 move on a circular orbit about their common centre of mass with angular velocity $\Omega = \left[G(M_1+M_2)/a^3\right]^{1/2}$, by Equation (27.51). A third body of negligible mass m moves in their field. In the frame rotating with the primaries, in which the two are at rest, the Hamiltonian of the third body is

$$\mathcal{H} = \frac{|\boldsymbol{p}|^2}{2m} - \Omega\, L_z - G M_1 m\,\frac{1}{r_1} - G M_2 m\,\frac{1}{r_2}\,, \tag{32.18}$$

with $L_z = (\boldsymbol{r}\times\boldsymbol{p})_z$ and r_1, r_2 the distances to the primaries. Since Equation (32.18) has no explicit time dependence in that frame, it is conserved; the constant is the *Jacobi integral*, of SI unit J.

Rests on Equations (22.1) and (27.51).

The system has three degrees of freedom and one known constant of motion. By Liouville–Arnold (Theorem 23.31) three independent constants in involution would make it integrable, and the question Poincaré was set—and answered in the negative—was whether the remaining two exist.

Theorem 32.47 (Poincaré's non-integrability, quoted)**.** *For the restricted three-body problem written as an integrable system perturbed at order ε in the mass ratio, there is no first integral, other than a function of $\mathcal{H}$ itself, that is analytic in the phase-space variables and in ε and single-valued in the angles [247]. The perturbation series that would construct such an integral diverges, and the geometric reason is the homoclinic tangle: the stable and unstable manifolds of an unstable periodic orbit intersect transversally, and by Corollary 32.5 each intersection forces infinitely many more, so the manifolds fold into a mesh "so complicated that I shall not even try to draw it" [255].*

Rests on Theorem 23.31 and Remark 32.44.

Remark 32.48 (Why the tangle defeats a series solution). An analytic first integral is constant on every trajectory and therefore constant on the closure of any trajectory. In the tangle the closure of a single unstable manifold contains a whole horseshoe (Remark 32.44), whose orbits by Proposition 32.43 come arbitrarily close to every point of a Cantor set of positive topological entropy. An analytic function constant on such a set, and on the infinitely many distinct periodic orbits it contains, is forced to be constant on an open set and hence—by analytic continuation—a function of the energy alone. The contrast with the two-body Kepler problem of Chapter 27 is total: there the motion is integrable, its level sets are the closed ellipses of Phenomenon 27.32, and the extra constants (angular momentum and the Laplace–Runge–Lenz vector) exist and are polynomial. Adding one more body destroys them.

32.4.2 The KAM theorem

Take an integrable Hamiltonian in the action–angle variables of Definition 23.28 and perturb it:

$$\mathcal{H}(I,\theta) = \mathcal{H}_0(I) + \varepsilon \mathcal{H}_1(I,\theta)\,, \tag{32.19}$$

with $\mathcal{H}_1$ 2π-periodic in each angle and ε dimensionless. The unperturbed motion lies on the invariant tori of Theorem 23.31, each labelled by its actions and carrying the frequencies $\omega^a(I) = \partial \mathcal{H}_0/\partial I_a$ of Equation (23.52).

Proposition 32.49 (The small-denominator obstruction)**.** *Seek a canonical transformation removing the perturbation to first order, generated by $S(I',\theta) = I'\cdot\theta + \varepsilon S_1(I',\theta)$. With the Fourier expansion $\mathcal{H}_1 = \sum_{\boldsymbol{k}\in\mathbb{Z}^f} h_{\boldsymbol{k}}(I)\,\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\theta}$, the generator is*

$$S_1(I,\theta) = \sum_{\boldsymbol{k}\neq\boldsymbol{0}} \frac{\mathrm{i}\,h_{\boldsymbol{k}}(I)}{\boldsymbol{k}\cdot\boldsymbol{\omega}(I)}\,\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\theta}\,, \tag{32.20}$$

which is undefined whenever $\boldsymbol{k}\cdot\boldsymbol{\omega}(I) = 0$ for some $\boldsymbol{k}\neq\boldsymbol{0}$, and arbitrarily large near such a resonance.

Rests on Equations (22.18) *and* (23.52).

Proof. *Derives Proposition 32.49.*

The generating function of the second kind (Equation (22.21)) gives $I_a = \partial S/\partial\theta^a = I'_a + \varepsilon\,\partial S_1/\partial\theta^a$, so the new Hamiltonian is

$$\mathcal{H}'(I') = \mathcal{H}_0(I') + \varepsilon\left[\boldsymbol{\omega}(I')\cdot\frac{\partial S_1}{\partial\theta} + \mathcal{H}_1(I',\theta)\right] + \mathcal{O}(\varepsilon^2)\,,$$

using $\partial\mathcal{H}_0/\partial I_a = \omega^a$. Requiring the bracket to be independent of the angles—and taking the angle average of $\mathcal{H}_1$, $h_{\boldsymbol{0}}$, into $\mathcal{H}'$—leaves $\boldsymbol{\omega}\cdot\partial S_1/\partial\theta = -\sum_{\boldsymbol{k}\neq\boldsymbol{0}} h_{\boldsymbol{k}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\theta}$. Inserting the Fourier series for S_1 turns the derivative into a factor $\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\omega}$ and solving mode by mode gives Equation (32.20). The denominators are the resonance conditions; since the rationals are dense, every neighbourhood of every torus contains resonant tori, and this is the difficulty that defeated celestial mechanics for two centuries. □

Definition 32.50 (Diophantine frequency vector)**.** A frequency vector $\boldsymbol{\omega}\in\mathbb{R}^f$ is *Diophantine* of type (γ,τ) if

$$|\boldsymbol{k}\cdot\boldsymbol{\omega}| \geq \frac{\gamma}{|\boldsymbol{k}|^{\tau}} \qquad \text{for all } \boldsymbol{k}\in\mathbb{Z}^f\setminus\{\boldsymbol{0}\}\,, \tag{32.21}$$

with $\gamma>0$ and $\tau > f-1$. The condition says the frequencies are badly approximable by rationals, so the denominators in Equation (32.20) shrink only polynomially in $|\boldsymbol{k}|$ while the numerators $h_{\boldsymbol{k}}$ of an analytic $\mathcal{H}_1$ shrink exponentially.

Rests on Proposition 32.49.

Theorem 32.51 (Kolmogorov–Arnold–Moser, quoted)**.** *Let $\mathcal{H}_0$ be analytic and non-degenerate, in the sense that* $\det(\partial^2\mathcal{H}_0/\partial I_a \partial I_b) \neq 0$*, and let $\mathcal{H}_1$ be analytic. Then for every γ there is $\varepsilon_0(\gamma) > 0$ such that for $|\varepsilon| < \varepsilon_0$ every unperturbed torus whose frequency vector is Diophantine of type (γ, τ) survives as an invariant torus of the perturbed system, merely deformed, carrying the same frequencies. The surviving tori fill a set whose relative measure tends to* 1 *as $\varepsilon \to 0$, the complement being $\mathcal{O}(\sqrt{\varepsilon})$ [1045, 1238, 1247].*

Rests on Definition 32.50 and Theorem 23.31.

Remark 32.52 (KAM is quoted, and will not be proved here). Theorem 32.51 carries no derivation in this book, and the omission is a decision rather than a debt—the same treatment Part II gives Carleson's theorem in Remark 17.18. What the proof does is worth stating, because it is what makes the theorem believable and it is also what makes it long. One cannot iterate the first-order step of Proposition 32.49: each pass leaves a remainder of order ε^2, then ε^3, and the small denominators of Equation (32.20) force one to work on a domain shrunk at every pass, so the successive corrections lose analyticity faster than they gain smallness and the series diverges. Kolmogorov's device is to replace the ordinary perturbation series by a Newton iteration [1045]: each pass solves the linearized problem exactly and therefore *squares* the remainder, $\varepsilon \to \varepsilon^2 \to \varepsilon^4$, which converges fast enough to survive both the shrinking domains and the growth of the denominators permitted by the Diophantine condition Equation (32.21). Arnold supplied the analytic details [1247] and Moser removed the analyticity hypothesis in favour of finite differentiability [1238].

Three ingredients of that argument have no home in this treatise: the Diophantine estimates themselves, the Cauchy estimates on shrinking complex strips that control each pass, and the measure-theoretic accounting that turns "the resonant set is small for each $\boldsymbol{k}$" into the $\mathcal{O}(\sqrt{\varepsilon})$ complement quoted in Theorem 32.51—the last of which needs the Lebesgue measure that Part II does not carry. None of them appears anywhere else in this book. What *is* derived here is the obstruction the theorem overcomes (Proposition 32.49), the condition under which it can be overcome (Definition 32.50), and—in Section 32.4.3—the quantitative estimate of where the tori actually break, which is Chirikov's overlap criterion and is not a corollary of KAM at all. Every number this chapter quotes about the solar system comes from numerical integration [1732, 1824], not from KAM, which is perturbative and silent at those perturbation strengths.

Remark 32.53 (What KAM does and does not guarantee). Four cautions, each of which is routinely overstated in summaries. First, KAM is perturbative: ε_0 is small, and the theorem says nothing at the perturbation strengths of real celestial mechanics, where the question is settled numerically (Section 32.4.5). Second, the surviving tori form a nowhere dense set of large measure—their complement, the resonant gaps, is dense—so "most initial conditions are regular" and "regular initial conditions form a fat Cantor set riddled with chaotic layers" are both true at once. Third, the theorem is about tori, not about the motion in between: what happens in the gaps is the subject of Section 32.4.3. Fourth, and most importantly, the confinement KAM provides is dimension-dependent. In a system with two degrees of freedom the invariant tori are two-dimensional inside a three-dimensional energy surface, so they separate it and a chaotic orbit is trapped between neighbouring tori forever. With three or more degrees of freedom the tori no longer separate the energy surface, the chaotic layers connect, and slow transport along the resonance web—Arnold diffusion—becomes possible. Its existence is established for constructed examples; its *rate* in a given physical system is not settled by any theorem and must be measured or integrated. The primary reference for the diffusion example is not in this bibliography, and the statement is therefore made here without one.

32.4.3 Resonance overlap and the standard map

The local model of a perturbed resonance is a rotor given a periodic kick.

Definition 32.54 (The kicked rotor and the standard map)**.** A rigid rotor of moment of inertia I (SI unit $\mathrm{kg\,m^2}$), free except for an impulsive torque applied once every T, has the Hamiltonian

$$\mathcal{H} = \frac{L^2}{2I} + K' T \cos\vartheta \sum_{n\in\mathbb{Z}} \delta(t - nT)\,, \tag{32.22}$$

with L the angular momentum ($\mathrm{J\,s}$) and K' an energy (J). Integrating Hamilton's equations across one kick and along the free flight between kicks gives the exact stroboscopic map

$$L_{n+1} = L_n + K' T \sin\vartheta_n\,, \tag{32.23}$$

$$\vartheta_{n+1} = \vartheta_n + \frac{T}{I} L_{n+1}\,, \tag{32.24}$$

which in the dimensionless action $l_n = T L_n / I$ becomes the *standard map*

$$l_{n+1} = l_n + K \sin\vartheta_n\,, \qquad \vartheta_{n+1} = \vartheta_n + l_{n+1}\,, \qquad K = \frac{K' T^2}{I}\,. \tag{32.25}$$

Rests on Equations (22.3) *and* (22.4).

The map Equation (32.25) is area-preserving—its Jacobian determinant is 1, as Proposition 32.40 requires—and K is the single dimensionless control parameter, built out of the measured kick energy, kick period and moment of inertia.

Proposition 32.55 (Chirikov's resonance-overlap estimate)**.** *Each primary resonance of Equation* (32.25) *sits at $l = 2\pi m$, $m \in \mathbb{Z}$, and has half-width $\Delta l = 2\sqrt{K}$ in the action. Neighbouring resonances therefore touch when $4\sqrt{K} = 2\pi$, that is at*

$$K_{overlap} = \left(\frac{\pi}{2}\right)^2 = 2.47\,, \tag{32.26}$$

above which no invariant curve can span the cylinder and the action diffuses without bound [1636].

Rests on Equations (32.19) *and* (32.25).

Proof. *Derives Proposition 32.55.*

Near $l = 0$ and for small K the map is the stroboscopic image of the pendulum flow $\dot\vartheta = l$, $\dot l = K\sin\vartheta$, whose energy $\tfrac12 l^2 + K\cos\vartheta$ is conserved. Its separatrix passes through $(\vartheta, l) = (0, 0)$ with energy K, and at $\vartheta = \pi$ reaches $\tfrac12 l^2 - K = K$, so the maximum excursion of the separatrix is $l = 2\sqrt{K}$: that is the half-width. The same holds at every $l = 2\pi m$ by the periodicity of the map in l modulo 2π. Two neighbouring separatrices, of half-width $2\sqrt{K}$ each, span the gap 2π between resonance centres when $2 \cdot 2\sqrt{K} = 2\pi$, which is Equation (32.26).

The criterion is a heuristic and is honest about it: it neglects the higher-harmonic resonances that fill the gaps and the deformation of the separatrices by the neighbouring resonance, both of which lower the threshold. Numerically the last invariant curve of Equation (32.25)—the one carrying the golden-mean winding number, the most irrational and therefore the last to break—disappears at $K_c = 0.9716$, so Equation (32.26) overestimates by a factor of about 2.5 [1636]. That is the accuracy the criterion claims for itself, and it is why the criterion is used for an order of magnitude and never for a number. □

Remark 32.56 (Cantori and slow transport). Just above K_c an invariant curve does not simply vanish: it becomes a *cantorus*, an invariant Cantor set with gaps, which no longer blocks transport but still obstructs it. Orbits leak through the gaps at a rate that vanishes as $K \to K_c^+$, so the diffusion of the action is not the free diffusion of a random walk but is throttled by a hierarchy of partial barriers. This is the mechanism behind the very long timescales of Section 32.4.5: a chaotic orbit can behave regularly for many Lyapunov times before it finds a gap.

The physical realizations of Equation (32.25) are the kicked rotor itself, the periodically driven pendulum of Section 32.7.3, and—the case Chirikov was writing for—any resonance of a nearly integrable system, since Equation (32.22) is what Equation (32.19) reduces to near a single resonance.

32.4.4 Numerical discovery: Hénon–Heiles

Definition 32.57 (The Hénon–Heiles Hamiltonian)**.** A star of mass m moving in the meridional plane of an axisymmetric galactic potential, expanded to cubic order about the circular orbit, has

$$\mathcal{H} = \frac{p_x^2 + p_y^2}{2m} + \frac{1}{2} m\omega^2 \left(x^2 + y^2\right) + m\lambda \left(x^2 y - \frac{y^3}{3}\right), \tag{32.27}$$

with ω in 1/s and λ in $1/(\mathrm{m\,s}^2)$ [1281]. Rescaling by $x = (\omega^2/\lambda)\xi$, $y = (\omega^2/\lambda)\eta$ and $t = s/\omega$ reduces it to

$$\tilde{\mathcal{H}} = \frac{1}{2}\left(\tilde{p}_\xi^2 + \tilde{p}_\eta^2\right) + \frac{1}{2}\left(\xi^2 + \eta^2\right) + \xi^2\eta - \frac{\eta^3}{3}, \tag{32.28}$$

measured in units of the energy $E_* = m\omega^6/\lambda^2$.

Rests on Equation (22.1) *and Theorem 7.38.*

Phenomenon 32.58 (Regular and chaotic motion coexist)**.** A slightly perturbed integrable Hamiltonian system does not become wholly chaotic. Surfaces of section computed for the Hénon–Heiles Hamiltonian show, at low energy, closed invariant curves filling the accessible region — the traces of surviving invariant tori — and, as the energy is raised, islands of such curves embedded in a growing sea of points that wander over a two-dimensional area [1281]. Both kinds of orbit occur at the same energy in the same system, and the chaotic fraction grows continuously with the perturbation: there is no energy at which order gives way to disorder all at once [1045, 1238, 1247]. The numerical experiment also settled the astronomical question it was posed for — whether stellar motion in a galactic disc admits a third isolating integral — in the negative for general orbits, while showing that many individual orbits behave as though it existed [1281].

Rests on Equation (32.27) *and Theorem 32.51.*

Derivation. *Derives Phenomenon 32.58.*

Three things have to be established: that the system has a finite range of bounded motion in which the question makes sense, that a closed curve on the surface of section is the trace of an invariant torus, and that a scattered area is not.

The escape energy. The cubic term in Equation (32.28) eventually overwhelms the quadratic one, so the potential $\tilde{V} = \frac{1}{2}(\xi^2 + \eta^2) + \xi^2\eta - \eta^3/3$ is bounded only inside a finite region. Setting $\boldsymbol{\nabla}\tilde{V} = 0$ gives $\xi(1 + 2\eta) = 0$ and $\eta + \xi^2 - \eta^2 = 0$. From $\xi = 0$ the second equation gives $\eta = 0$ (the minimum, $\tilde{V} = 0$) or $\eta = 1$, at which $\tilde{V} = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$. From $\eta = -\frac{1}{2}$ it gives $\xi^2 = \frac{3}{4}$, and there $\tilde{V} = \frac{1}{2}\left(\frac{3}{4} + \frac{1}{4}\right) - \frac{3}{8} + \frac{1}{24} = \frac{1}{6}$ as well. There are therefore three saddle points at the same height, at the corners of an equilateral triangle, and

$$E_{\mathrm{escape}} = \frac{1}{6} E_* = \frac{m\omega^6}{6\lambda^2}. \tag{32.29}$$

Below this the motion is bounded and the question of integrability is well posed; above it, orbits escape through the three channels.

Reading the section. Take the section $\xi = 0$ with $\pi_\xi > 0$, coordinatized by (η, π_η); by Proposition 32.40 the return map preserves area. If the orbit lies on a two-torus—which is what a third isolating integral would guarantee, by Theorem 23.31—then the intersection of that two-torus with the two-dimensional section inside the three-dimensional energy surface

is a one-dimensional curve, and the successive points must fall on it. Conversely, points that fill a two-dimensional patch cannot lie on any smooth one-dimensional invariant set, so no third integral exists on that orbit. This is what makes the surface of section a decisive diagnostic rather than a picture: the dimension of the point set is the answer.

The measured outcome. At $E = \frac{1}{12}E_*$ the computed section is covered by closed curves and the motion looks integrable; at $E = \frac{1}{8}E_*$ islands of closed curves survive inside a connected region of scattered points; at $E = \frac{1}{6}E_*$ the scattered region occupies almost the whole accessible area, with a few small islands remaining [1281]. The relative area of the scattered region rises continuously from zero, which is the statement of the phenomenon, and Theorem 32.51 is the reason it must: the tori that go first are the resonant ones, and the Diophantine ones of Equation (32.21) survive to larger perturbations, so the transition is graded and not sudden. The astronomical conclusion follows directly: no third isolating integral, analytic in the coordinates, exists for general orbits in such a potential, while the surviving tori explain why so many individual stellar orbits behave as though one did. □

32.4.5 Chaos in the solar system

Phenomenon 32.59 (The solar system is chaotic)**.** The asteroid belt is not populated uniformly: it carries deep gaps at the orbital periods commensurate with Jupiter's — the Kirkwood gaps — and the gap at the 3/1 commensurability coincides with a zone in which the eccentricity of a test asteroid does not oscillate but wanders, on a timescale short against the age of the system, until the orbit crosses those of the inner planets and the body is removed [1732]. The planetary orbits themselves are chaotic in the same sense: numerical integration of the averaged equations gives a positive Lyapunov exponent for the inner solar system, with a Lyapunov time of order 5×10^6 yr, so that an error of 15 m in an initial position grows within about 1×10^8 yr to a distance of the order of the Earth–Sun separation, 1.5×10^{11} m [1824]. The observed regularity of the planets over recorded history is therefore no evidence at all of predictability over the age of the system.

Rests on Equation (32.46) and Theorem 32.51.

Derivation. *Derives Phenomenon 32.59.*

Where the gaps are. Kepler's third law Equation (27.51) with the Sun's mass gives, for a heliocentric orbit, $a = (T/1\,\mathrm{yr})^{2/3}\,\mathrm{au}$. An asteroid in the $p{:}q$ mean-motion resonance with Jupiter has $T = (q/p)\,T_J$ with $T_J = 11.862\,\mathrm{yr}$, so its semi-major axis is fixed by arithmetic alone: Table 32.2 lists the four principal cases against the observed gap positions. The agreement is at the level of the approximation—the derivation used the Sun's mass alone and a circular Jovian orbit—and it identifies the gaps as a resonance phenomenon without yet explaining why a resonance should empty rather than fill.

Why the 3/1 zone empties. At the 3/1 resonance the numerical integration of [1732] shows the eccentricity of a test body not librating but jumping irregularly, in sudden bursts that carry it above $e = 0.3$ on a timescale of order 1×10^6 yr. That range is the one that matters, because Mars has $a_M = 1.5237\,\mathrm{au}$ and $e_M = 0.0934$, hence aphelion $Q_M = a_M(1 + e_M) = 1.666\,\mathrm{au}$; an asteroid at $a = 2.50\,\mathrm{au}$ reaches Mars's orbit when its perihelion $a(1 - e)$ falls to Q_M, that is when

$$e > 1 - \frac{Q_M}{a} = 1 - \frac{1.666}{2.50} = 0.33\,. \tag{32.30}$$

A body that reaches such an eccentricity is removed by a close encounter within an interval negligible against the age of the system, and the gap is the record of that removal. The chaotic zone and the observed gap coincide in width, which is the quantitative content of [1732].

The planets. For the planets themselves the exponent is measured, not derived: [1824] integrates the averaged secular equations for 2×10^8 yr and finds neighbouring solutions

separating exponentially with a Lyapunov time $\lambda^{-1} \approx 5 \times 10^{6}\,\mathrm{yr}$. Feeding that into the horizon formula Equation (32.46), an initial uncertainty $\delta_0 = 15\,\mathrm{m}$ reaches $\Delta = 1.5 \times 10^{11}\,\mathrm{m}$ after

$$T = \lambda^{-1} \ln \frac{\Delta}{\delta_0} = 5 \times 10^{6}\,\mathrm{yr} \times \ln\left(1 \times 10^{10}\right) = 5 \times 10^{6}\,\mathrm{yr} \times 23.0 \approx 1.2 \times 10^{8}\,\mathrm{yr}\,,$$

which is the growth quoted in [1824] to the one significant figure to which the Lyapunov time itself is known. Two conclusions, and only two, follow. Positive λ destroys prediction of the *phase* of the orbits beyond about $1 \times 10^{8}\,\mathrm{yr}$; it does not by itself say that any planet is lost, which is a question about the size of the chaotic zone and not about the exponent, and which the same integrations address separately. The regularity of the planets over the few thousand years of recorded observation is 1×10^{-3} Lyapunov times and constrains nothing. □

Table 32.2: Positions of the principal Kirkwood gaps, computed from Kepler's third law Equation (27.51) with $T_J = 11.862\,\mathrm{yr}$ and no free parameter, against the observed gap centres. The residual reflects the neglect of Jupiter's mass and of the eccentricity of its orbit.

Resonance	T (yr)	a computed (au)	a observed (au)
4:1	2.966	2.06	2.06
3:1	3.954	2.50	2.50
5:2	4.745	2.82	2.82
7:3	5.084	2.96	2.96
2:1	5.931	3.28	3.27

The gap positions in Table 32.2 are arithmetic. What a resonance does to a body sitting in one is dynamics, and the local model is the pendulum of Proposition 32.55 again.

Proposition 32.60 (Libration width of a mean-motion resonance)**.** *Let a test body of mass m_t, negligible beside both other masses, orbit a star of mass M with semi-major axis a, eccentricity e and mean motion $n = \sqrt{GM/a^3}$, perturbed by a planet of mass m on a circular coplanar orbit of mean motion n_J, and let it lie near the p:q commensurability $q\,n = p\,n_J$. Write the single resonant harmonic of the perturbing Hamiltonian as*

$$\mathcal{H}_{res} = -\frac{GMm_t}{a}\,\mu\,S\cos\sigma\,, \qquad \mu = \frac{m}{M}\,, \tag{32.31}$$

with S dimensionless and

$$\sigma = p\lambda_J - q\lambda + (q-p)\varpi \tag{32.32}$$

the resonant angle built from the two mean longitudes and the longitude of pericentre ϖ. Then, while ϖ may be treated as fixed, σ obeys a pendulum equation whose libration angular frequency is

$$\omega_{lib} = q\,n\sqrt{3\mu S}\,, \tag{32.33}$$

and whose separatrix has the half-width

$$\frac{\Delta a}{a} = \frac{4}{\sqrt{3}}\sqrt{\mu S} \tag{32.34}$$

in semi-major axis, independent of p and q except through S. Furthermore S vanishes with the eccentricity as $S = s\,e^{|p-q|}$, with s a function of a/a_J alone and of order unity.

Rests on Proposition 32.55 and Equation (32.19).

Proof. *Derives Proposition 32.60.*

The pendulum. Take as the action conjugate to the mean longitude λ the quantity $\Lambda = m_t\sqrt{GMa}$, for which the Kepler Hamiltonian is $\mathcal{H}_0 = -G^2M^2m_t^3/(2\Lambda^2)$ and

$$\frac{\partial \mathcal{H}_0}{\partial \Lambda} = \frac{G^2M^2m_t^3}{\Lambda^3} = \sqrt{\frac{GM}{a^3}} = n\,, \qquad \frac{\mathrm{d}n}{\mathrm{d}\Lambda} = -\frac{3n}{\Lambda}\,,$$

the second by differentiating the first. Hamilton's equation for Λ picks out the λ dependence of Equation (32.31), which enters only through σ and only as $-q\lambda$:

$$\dot\Lambda = -\frac{\partial \mathcal{H}_{\mathrm{res}}}{\partial \lambda} = -\frac{GMm_t}{a}\mu S\,\sin\sigma\,\frac{\partial\sigma}{\partial\lambda} = q\,\varepsilon\sin\sigma\,, \qquad \varepsilon \equiv \frac{GMm_t}{a}\mu S\,.$$

Differentiating Equation (32.32) twice with ϖ fixed and n_J constant, $\dot\sigma = p\,n_J - q\,n$ and

$$\ddot\sigma = -q\,\frac{\mathrm{d}n}{\mathrm{d}\Lambda}\dot\Lambda = \frac{3q^2n\varepsilon}{\Lambda}\sin\sigma = \omega_{\mathrm{lib}}^2\sin\sigma\,, \tag{32.35}$$

a pendulum with its stable equilibrium at $\sigma = \pi$. Its frequency is Equation (32.33), since $\Lambda = m_t n a^2$ and $\varepsilon = m_t n^2 a^2 \mu S$ give $3q^2n\varepsilon/\Lambda = 3q^2n^2\mu S$.

The width. Equation (32.35) conserves $\frac12\dot\sigma^2 + \omega_{\mathrm{lib}}^2\cos\sigma$. The separatrix passes through the unstable point $\sigma = 0$ with $\dot\sigma = 0$, so its constant is ω_{lib}^2, and at $\sigma = \pi$ it reaches $\frac12\dot\sigma^2 = 2\omega_{\mathrm{lib}}^2$: the maximum excursion is $|\dot\sigma| = 2\omega_{\mathrm{lib}}$. This is the same computation as in Proposition 32.55, on a different pair of variables. Converting to the semi-major axis, $\dot\sigma = -q\,(\mathrm{d}n/\mathrm{d}\Lambda)\,\Delta\Lambda = 3qn\,\Delta\Lambda/\Lambda$, so

$$\frac{\Delta\Lambda}{\Lambda} = \frac{2\omega_{\mathrm{lib}}}{3qn} = \frac{2}{\sqrt3}\sqrt{\mu S}\,,$$

and $\Lambda \propto \sqrt{a}$ makes $\Delta a/a = 2\Delta\Lambda/\Lambda$, which is Equation (32.34). The factor q has cancelled.

Why S carries $e^{|p-q|}$. The perturbing Hamiltonian depends on the two orbits only through their relative geometry, so it is unchanged by rotating the whole system: $\lambda \to \lambda + \phi$, $\lambda_J \to \lambda_J + \phi$, $\varpi \to \varpi + \phi$. The combination Equation (32.32) is invariant under exactly that substitution—its three coefficients sum to $p - q + q - p = 0$—which is what selects the admissible harmonics. Now ϖ is undefined at $e = 0$, and the perturbation is a smooth function of the position of the test body, hence of the two elements $e\cos\varpi$ and $e\sin\varpi$, which are smooth where ϖ is not. A term carrying $\mathrm{e}^{\mathrm{i}j\varpi}$ can be assembled from those two only as a polynomial of degree at least $|j|$ in them, so it vanishes at least as $e^{|j|}$. Here $j = q - p$, giving $S = s\,e^{|p-q|}$ with the remaining factor a function of the semi-major-axis ratio alone. □

Remark 32.61 (The 3/1 zone: what the width says, and what is quoted). Put the numbers in for the 3/1 gap, where $p = 3$, $q = 1$ and the resonance is therefore of second order, $S = s\,e^2$. With $\mu = m_J/M_\odot = 9.5458\times10^{-4}$, Equation (32.34) gives

$$\frac{\Delta a}{a} = \frac{4}{\sqrt3}\sqrt{\mu s}\;e = 0.0714\,\sqrt{s}\;e\,,$$

so a body of the typical belt eccentricity $e \approx 0.1$ librates over $\Delta a \approx 0.018\,\mathrm{au}$ about $a = 2.50\,\mathrm{au}$, and the libration period Equation (32.33) is $2\pi/\omega_{\mathrm{lib}} = 74\,\mathrm{yr}/(\sqrt{s}\,e)$, of order $7\times10^2\,\mathrm{yr}$ there. Two features of those formulas do the work. The width grows *with* the eccentricity, so a body whose e is being pumped finds the resonance widening around it rather than escaping from it; and the libration frequency vanishes with e, while the secular precession rate of ϖ—the very quantity the derivation above froze—is of order μn and does *not* vanish with e. Their ratio is

$$\frac{\omega_{\mathrm{lib}}}{|\dot\varpi|} \sim \frac{n\sqrt{3\mu s}\,e}{\mu n} = \sqrt{\frac{3s}{\mu}}\;e \approx 56\,\sqrt{s}\;e\,,$$

which falls below unity for $e \lesssim 0.02$. Below that eccentricity the single-resonance averaging that produced the pendulum is not valid: the resonant and the secular degrees of freedom run at comparable rates, and two comparable resonances that overlap are precisely the configuration Proposition 32.55 identifies as chaotic. The low-eccentricity part of the 3/1 zone is therefore where the pendulum picture must fail, which is the qualitative statement that a body starting on a nearly circular orbit there does not stay on one.

What is *not* derived here is the rest of [1732]: that the eccentricity actually wanders as far as the Mars-crossing value Equation (32.30) rather than to some smaller ceiling, and that it does so in bursts on a timescale of order 1×10^6 yr. Those are numerical results, obtained by integrating an algebraic mapping of the resonance over millions of orbits, and this chapter quotes them. The timescale is worth one comment, because it is easy to misread: it is *not* the Lyapunov time. The only timescale the local model supplies is the libration period, a few hundred years, and a resonance-overlap chaotic layer has a Lyapunov time of that order; the 1×10^6 yr on which the eccentricity jumps is three orders of magnitude longer, and the gap between them is the throttled transport of Remark 32.56. A chaotic orbit can look regular for a thousand Lyapunov times before it finds a way through.

32.5 Strange attractors

32.5.1 The Lorenz system

Definition 32.62 (The Lorenz system)**.** Truncating the Boussinesq equations for a fluid layer of depth d heated from below to one convective mode and two thermal modes gives the three ordinary differential equations

$$\dot{x} = \sigma(y - x)\,, \qquad \dot{y} = x(r - z) - y\,, \qquad \dot{z} = xy - bz\,, \tag{32.36}$$

in which the dot is $\mathrm{d}/\mathrm{d}s$ with $s = \pi^2(1 + a^2)\kappa t/d^2$ a dimensionless time built from the thermal diffusivity κ (m^2/s), x is proportional to the amplitude of the convective roll, y and z to the horizontal and vertical temperature variations, and the three parameters are

$$\sigma = \frac{\nu}{\kappa}\,, \qquad r = \frac{\mathrm{Ra}}{\mathrm{Ra}_c}\,, \qquad b = \frac{4}{1 + a^2}\,, \tag{32.37}$$

with ν the kinematic viscosity, a the dimensionless horizontal wavenumber of the roll, and

$$\mathrm{Ra} = \frac{g\alpha_T\,\Delta T\,d^3}{\nu\kappa} \tag{32.38}$$

the Rayleigh number, built from the acceleration of gravity g, the thermal expansion coefficient α_T (1/K) and the temperature difference ΔT across the layer [1262]. For stress-free boundaries the critical value is $\mathrm{Ra}_c = 27\pi^4/4 = 657.5$, attained at $a^2 = \tfrac{1}{2}$, whence Lorenz's $b = 8/3$; his other values are $\sigma = 10$ and $r = 28$.

Rests on Theorem 32.6 and Definition 32.7.

Full derivation: Appendix A.65 (p. 3146).

Derives Definition 32.62.

Appendix A.65 carries the reduction in full: the Boussinesq equations for a layer of depth d heated from below, the stream-function form that eliminates the pressure, the linear stability analysis that produces the critical Rayleigh number $\mathrm{Ra}_c = 27\pi^4/4$ at $a^2 = \tfrac{1}{2}$ for stress-free boundaries [438], the Galerkin truncation to one velocity mode and two temperature modes, and the scaling that turns what is left into Equation (32.36) with the three parameters Equation (32.37). It is a page of vector calculus and several pages of algebra, which is why it is not reproduced here, and it is what makes each of σ, r and b a ratio of measured fluid properties rather than a dial: σ is the Prandtl number of the working

fluid, r the Rayleigh number in units of its own critical value, and b a pure function of the horizontal wavenumber of the roll. The physical content that follows the truncation—the dissipation rate, the fixed points, the onset of the strange attractor—is derived in this chapter and does not depend on the appendix.

Remark 32.63 (The Lorenz system is a model, not the fluid). Equation (32.36) retains three of the infinitely many modes of a convecting layer, and Lorenz said plainly that the truncation is not quantitatively faithful to convection at $r = 28$, which is far above the range in which a three-mode truncation can be trusted [1262]. What the system is faithful to is the *mechanism*: a dissipative flow, bounded, with an unstable fixed point and no stable alternative nearby. Its evidential standing rests on Section 32.7.1, where a real convection cell reproduces the universal numbers, and not on any claim that Equation (32.36) describes that cell.

Proposition 32.64 (The Lorenz flow contracts volume)**.** *The divergence of Equation* (32.36) *is the negative constant* $-(\sigma + 1 + b)$*, so every phase volume obeys*

$$V(s) = V(0)\,\mathrm{e}^{-(\sigma+1+b)s}\,, \tag{32.39}$$

which at Lorenz's parameters contracts by a factor e *in every* $3/41 = 0.0732$ *of dimensionless time.*

Rests on Theorem 32.6 and Equation (32.36)*.*

Proof. *Derives Proposition 32.64.*

Differentiating each component of Equation (32.36) with respect to its own variable gives $-\sigma$, -1 and $-b$, whose sum is the constant $-(\sigma + 1 + b)$. Equation (32.3) then reads $\dot V = -(\sigma + 1 + b)V$, which integrates to Equation (32.39). At $\sigma = 10$ and $b = 8/3$ the rate is $41/3$ and its reciprocal is $3/41$. □

Proposition 32.65 (Fixed points of the Lorenz system and their stability)**.** *The origin is a fixed point for every* r*; it is asymptotically stable for* $r < 1$ *and a saddle for* $r > 1$*, the transition at* $r = 1$ *being a supercritical pitchfork that creates the pair*

$$C_{\pm} = \left(\pm\sqrt{b(r-1)},\ \pm\sqrt{b(r-1)},\ r-1\right)\,, \tag{32.40}$$

the two steady convecting rolls. For $\sigma > b + 1$ *the pair loses stability at*

$$r_H = \frac{\sigma(\sigma + b + 3)}{\sigma - b - 1}\,, \tag{32.41}$$

which is $470/19 = 24.7368$ *at* $\sigma = 10$*,* $b = 8/3$*.*

Rests on Equation (32.36) *and Theorem 32.13.*

Proof. *Derives Proposition 32.65.*

Setting the right-hand sides of Equation (32.36) to zero gives $y = x$ from the first, then $x(r - 1 - z) = 0$ from the second and $x^2 = bz$ from the third. Either $x = y = z = 0$, or $z = r - 1$ and $x^2 = b(r - 1)$, which is real only for $r \geq 1$: that is Equation (32.40), and the $\sqrt{r-1}$ growth of the amplitude identifies the pitchfork of Proposition 32.24(ii), the symmetry being the $(x, y) \to (-x, -y)$ invariance of Equation (32.36).

At the origin the Jacobian is block diagonal, with the entry $-b$ and the 2×2 block

$$B = \begin{pmatrix} -\sigma & \sigma \\ r & -1 \end{pmatrix}\,.$$

The block has $\tau = -(\sigma + 1) < 0$ and $\Delta = \sigma(1 - r)$, so by Proposition 32.17 it is a stable node for $r < 1$ and a saddle for $r > 1$.

At $C_\pm$, writing $x_0 = \pm\sqrt{b(r-1)}$, the Jacobian is

$$A = \begin{pmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & -x_0 \\ x_0 & x_0 & -b \end{pmatrix},$$

whose trace is $-(\sigma+1+b)$, whose determinant is $-2\sigma x_0^2 = -2\sigma b(r-1)$, and whose sum of principal 2×2 minors is $0 + \sigma b + (b + x_0^2) = b(\sigma + r)$. Hence

$$\lambda^3 + (\sigma+1+b)\lambda^2 + b(\sigma+r)\lambda + 2\sigma b(r-1) = 0\,. \tag{32.42}$$

A Hopf bifurcation needs a purely imaginary pair. Substituting $\lambda = \mathrm{i}\omega$ in Equation (32.42) and separating real and imaginary parts gives

$$2\sigma b(r-1) = (\sigma+1+b)\,\omega^2\,, \qquad \omega^2 = b(\sigma+r)\,,$$

and eliminating ω^2 leaves $(\sigma+1+b)(\sigma+r) = 2\sigma(r-1)$. Expanding, $\sigma^2+b\sigma+3\sigma+r(1+b-\sigma) = 0$, which solves to Equation (32.41); it is positive only for $\sigma > b+1$. With $\sigma = 10$ and $b = 8/3$, $r_H = 10\cdot\frac{47}{3}/\frac{19}{3} = 470/19$. □

Remark 32.66 (Why $r = 28$ has nowhere to go). Proposition 32.65 places $r = 28$ above $r_H = 24.74$, so all three fixed points are unstable there. The Hopf bifurcation at r_H is subcritical: the periodic orbits it involves exist *below* r_H and are unstable, so they are swallowed by $C_\pm$ rather than emitted by them, in the manner of Proposition 32.24(iii). Above r_H there is therefore no local attractor of any kind, while Proposition 32.64 forces all volumes to zero and a trapping argument keeps trajectories bounded. Something of zero volume and no local structure must attract them: that is the strange attractor of Phenomenon 32.68, and the argument just given is the whole reason to expect one.

32.5.2 Strange attractors and the Ruelle–Takens proposal

Definition 32.67 (Strange attractor)**.** An attractor (Definition 32.8) is *strange* if the dynamics on it has sensitive dependence on initial conditions, that is a positive largest Lyapunov exponent (Definition 32.73). It is a *fractal* attractor if in addition its dimension (Definition 32.79) is not an integer. The two properties are logically independent, and the term coined by Ruelle and Takens [1453] covers the sets—the common case—that have both.

Rests on Definition 32.8 and Phenomenon 32.72.

Phenomenon 32.68 (Strange attractors)**.** A dissipative chaotic system settles onto an attracting set of zero phase-space volume which is nevertheless neither a point, nor a closed curve, nor a smooth surface: it is a set of non-integer dimension, measured from experimental and numerical time series by counting the pairs of points lying closer than a given separation [1720]. The Lorenz system is the standard case: its trajectories are drawn onto a bounded set of two sheets on which they never repeat, and the correlation dimension of that set is measured to be a little above two [1720]. Sets of this kind are what Ruelle and Takens named strange attractors, and their geometry is that of the fractals [1453, 1696].

Rests on Definition 32.67 and Proposition 32.64.

Derivation. *Derives Phenomenon 32.68.*

The zero-volume half of the statement is Proposition 32.64: every volume decays like $\mathrm{e}^{-(\sigma+1+b)s}$, so the attractor can contain no three-dimensional region whatever.

It is equally not a point and not a closed curve. By Remark 32.66 no fixed point and no local periodic orbit is stable at $r = 28$; and on the attractor neighbouring trajectories separate exponentially, by Equation (32.44), while the motion remains bounded, so the set is stretched in one direction and folded back on itself without end. Nor can it be a

smooth surface: a two-dimensional invariant surface in a three-dimensional flow would confine trajectories to a plane-like region in which Corollary 32.32 forbids the observed sensitivity.

What remains is fixed quantitatively by the exponents. The Lorenz attractor at $\sigma = 10$, $b = 8/3$, $r = 28$ has computed Lyapunov spectrum $\lambda_1 = 0.906$, $\lambda_2 = 0$, $\lambda_3 = -14.572$ per unit dimensionless time, whose sum is $-13.667 = -(\sigma + 1 + b)$ exactly as Proposition 32.76 requires—a check on the numbers that costs nothing and is worth making. The Kaplan–Yorke estimate $D = 2 + \lambda_1 / |\lambda_3| = 2.06$ then predicts a dimension just above two, and the correlation dimension measured from the trajectory by pair counting is 2.05 ± 0.01 [1720]. The measured value is what settles the statement; the derivation establishes only that it must lie strictly between two and three. □

Remark 32.69 (The Ruelle–Takens proposal and what displaced Landau). Landau's account of the onset of turbulence had each instability add one incommensurate frequency, so that the flow became turbulent by accumulating an unbounded number of them. Ruelle and Takens proposed instead that after a small number of bifurcations the motion settles on an attractor which is strange in the sense of Definition 32.67, with a continuous power spectrum and sensitive dependence [1453]. The two pictures differ in a directly measurable quantity—the shape of the power spectrum—and that is what the Taylor–Couette measurement of Section 32.7.2 was able to decide.

Example 32.70 (The Hénon map)**.** The tractable prototype is the two-dimensional map

$$x_{n+1} = 1 - \hat{a} x_n^2 + y_n \,, \qquad y_{n+1} = \hat{b} x_n \,, \tag{32.43}$$

with $\hat{a} = 1.4$ and $\hat{b} = 0.3$ [1581]. Its Jacobian matrix is $\begin{pmatrix} -2\hat{a}x & 1 \\ \hat{b} & 0 \end{pmatrix}$, whose determinant is the constant $-\hat{b}$: the map contracts areas by the exact factor 0.3 at every step, whatever the point, which makes it the cleanest available caricature of a dissipative flow. Its largest Lyapunov exponent is $\lambda_1 = 0.419$ per iterate, so $\lambda_2 = \ln \hat{b} - \lambda_1 = -1.623$ and the Kaplan–Yorke estimate is $1 + \lambda_1 / |\lambda_2| = 1.26$; the measured correlation dimension is 1.21 ± 0.01 [1720]. The two differ, as they generally do—the correlation dimension is a lower bound on the information dimension that the Kaplan–Yorke formula estimates—and quoting either as "the" dimension of the attractor without saying which is measured would misreport both.

Rests on Definitions 32.38 and 32.67.

Remark 32.71 (Hyperbolicity, and an honest limit). The horseshoe of Section 32.3.2 is *uniformly hyperbolic*: at every point of its invariant set the tangent space splits into a uniformly contracting and a uniformly expanding direction, and that is what makes its symbolic description exact. Real attractors are not uniformly hyperbolic. The Lorenz attractor contains a fixed point, at which the expansion rate differs from that on the rest of the set, and the Hénon attractor has tangencies between stable and unstable manifolds. That the Lorenz equations possess an attractor at all—rather than a very long transient, or a periodic orbit of enormous period—was for four decades a conjecture, settled eventually by a computer-assisted proof whose primary reference is not in this bibliography and which is therefore not cited here. The practical consequence for a reader is worth stating plainly: the numbers quoted in this section are measurements on trajectories, and their status is that of measurements.

32.5.3 Sensitive dependence and Lyapunov exponents

Phenomenon 32.72 (Sensitive dependence on initial conditions)**.** In a chaotic system two initial states differing by an unmeasurably small amount separate at an exponential rate:

on the attractor the separation of nearby trajectories grows on average as

$$|\delta \boldsymbol{x}(t)| \sim |\delta \boldsymbol{x}(0)|\, \mathrm{e}^{\lambda t}\,, \qquad \lambda > 0\,, \tag{32.44}$$

with λ — the largest Lyapunov exponent — a property of the attractor and not of the particular trajectory followed [1378]. The effect was established in a numerical model of thermal convection, in which two solutions started from imperceptibly different data were followed until they bore no resemblance to one another [1262], and a positive λ is now extracted routinely from a single measured scalar time series [1766]. Determinism is untouched by this; what is destroyed is prediction, on the finite horizon of Equation (32.46).

Rests on Equation (32.4) *and Theorem 32.74.*

Derivation. *Derives Phenomenon 32.72.*

Linearize the flow $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ about a reference trajectory: an infinitesimal separation obeys the variational equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\delta \boldsymbol{x} = D\boldsymbol{f}\big(\boldsymbol{x}(t)\big)\,\delta \boldsymbol{x}\,, \tag{32.45}$$

whose solution is $\delta \boldsymbol{x}(t) = M(t)\,\delta \boldsymbol{x}(0)$ for a fundamental matrix M depending on the reference trajectory alone. For a flow with an invariant measure the multiplicative ergodic theorem (Theorem 32.74) guarantees that the limit

$$\lambda\big(\delta \boldsymbol{x}(0)\big) = \lim_{t\to\infty} \frac{1}{t} \ln \frac{|\delta \boldsymbol{x}(t)|}{|\delta \boldsymbol{x}(0)|}$$

exists for almost every initial point and takes one of finitely many values, the largest of which is attained for every $\delta \boldsymbol{x}(0)$ outside a set of lower dimension [1378]. That is Equation (32.44), and it is why λ can be quoted as a number for the system rather than for the run.

The consequence is quantitative. Suppose the initial state is known to within δ_0 and the prediction is useful while the error stays below a tolerance Δ. Solving $\delta_0 \mathrm{e}^{\lambda T} = \Delta$ gives the horizon

$$T = \frac{1}{\lambda} \ln \frac{\Delta}{\delta_0}\,. \tag{32.46}$$

It depends on the measurement precision only through its *logarithm*: improving the initial data by a factor of ten buys the fixed extra time $\lambda^{-1} \ln 10$ and no more, however great the effort. A system with $\lambda > 0$ is therefore not one whose equations are unknown — Equation (32.45) used them — but one in which prediction is bought at an exponentially rising price. □

Definition 32.73 (Lyapunov spectrum)**.** The *Lyapunov exponents* of a trajectory are the n numbers $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ giving the exponential growth rates of the principal axes of an infinitesimal ball carried by Equation (32.45); equivalently $\lambda_i = \lim_{t\to\infty} t^{-1} \ln \mu_i(t)$ with μ_i the singular values of the fundamental matrix $M(t)$. They carry SI units of 1/s for a flow and are dimensionless per iterate for a map. A bounded, aperiodic trajectory with $\lambda_1 > 0$ is called *chaotic*.

Rests on Equation (32.45)*.*

Theorem 32.74 (Oseledets, quoted)**.** *Let φ_t preserve a probability measure supported on a compact invariant set, and let $\boldsymbol{f}$ be continuously differentiable. Then for almost every point with respect to that measure the limits in Definition 32.73 exist, are the same for almost all points of an ergodic component, and the tangent space at each point splits into subspaces on which the growth rate is exactly λ_i [1378].*

Rests on Definitions 32.8 and 32.73.

Remark 32.75 (Oseledets is quoted, not proved). The multiplicative ergodic theorem is a theorem of ergodic theory, resting on measure theory, on Birkhoff's pointwise ergodic theorem and on the subadditive ergodic theorem. Part II contains none of these: Chapter 11 is built on discrete and Riemann-integrable probability and says so. The theorem is quoted here because it is what converts "two runs diverged" into "the system has an exponent", and every numerical exponent quoted in this chapter relies on it for its meaning. The two propositions that follow are independent of it and are proved.

Proposition 32.76 (The exponents sum to the mean divergence)**.** *For a flow with an ergodic invariant measure on its attractor,*

$$\sum_{i=1}^{n} \lambda_i = \langle \boldsymbol{\nabla} \cdot \boldsymbol{f} \rangle \ , \tag{32.47}$$

the average being over the attractor. In particular a Hamiltonian flow has $\sum_i \lambda_i = 0$*, and the Lorenz flow has* $\sum_i \lambda_i = -(\sigma + 1 + b)$*.*
Rests on Theorem 32.6 and Definition 32.73.

Proof. *Derives Proposition 32.76.*

An infinitesimal ball of radius ϵ is carried by Equation (32.45) into an ellipsoid whose principal axes are $\epsilon\mu_i(t)$, so its volume is $V(t) = V(0)\prod_i \mu_i(t)$ and

$$\frac{1}{t}\ln\frac{V(t)}{V(0)} = \sum_i \frac{1}{t}\ln\mu_i(t) \longrightarrow \sum_i \lambda_i \, .$$

On the other hand Equation (32.3) applied to an infinitesimal volume gives $\dot{V}/V = \boldsymbol{\nabla}\cdot\boldsymbol{f}$ evaluated on the trajectory, so $t^{-1}\ln[V(t)/V(0)]$ is the time average of $\boldsymbol{\nabla}\cdot\boldsymbol{f}$ along it, which for an ergodic measure tends to its space average. Equating the two limits gives Equation (32.47). For a Hamiltonian flow the divergence vanishes identically (Theorem 24.21); for Equation (32.36) it is the constant of Proposition 32.64. □

Proposition 32.77 (A flow has a zero exponent)**.** *Let a trajectory of Equation (32.1) remain in a compact invariant set containing no fixed point. Then it has at least one vanishing Lyapunov exponent, associated with displacement along the flow.*
Rests on Equation (32.45) and Definition 32.73.

Proof. *Derives Proposition 32.77.*

Differentiate $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ with respect to t: $\ddot{\boldsymbol{x}} = D\boldsymbol{f}(\boldsymbol{x})\dot{\boldsymbol{x}}$, so $\delta\boldsymbol{x}(t) = \boldsymbol{f}(\boldsymbol{x}(t))$ is itself a solution of the variational equation Equation (32.45). On a compact attractor containing no fixed point, $|\boldsymbol{f}|$ is bounded above and below by positive constants (Theorem 7.24), so $t^{-1}\ln|\delta\boldsymbol{x}(t)| \to 0$: the exponent along the flow is zero. The hypothesis excludes a fixed point because $|\boldsymbol{f}|$ must be bounded away from zero; the Lorenz attractor does contain one, at the origin, and the conclusion survives there only because a trajectory spends a vanishing fraction of its time near it, which is a statement about the invariant measure and not about the flow alone. This is why a three-dimensional dissipative flow has the spectrum $(+, 0, -)$ with the sum negative—the smallest arrangement that admits chaos—and why a map, which has no such direction, can be chaotic in one dimension. □

Remark 32.78 (Extracting exponents from data). An experiment does not deliver $D\boldsymbol{f}$; it delivers a scalar time series. The standard procedure embeds the series in a higher-dimensional space by delay coordinates, follows a pair of nearby reconstructed points, measures the growth of their separation over a short interval, renormalizes the separation before it leaves the linear regime, and averages the logarithmic growth rates [1766]. The output is λ_1 with an uncertainty dominated by the length of the series and the noise level, and it is on this chain that every experimental claim of a positive exponent in Section 32.7

rests. The embedding step itself—that a scalar observable of a d-dimensional attractor generically reconstructs it in $2d+1$ delay coordinates—is a theorem of Takens whose primary reference is not in this bibliography; it is used here as a method and not cited as a result.

32.5.4 Fractal dimension

Definition 32.79 (Box-counting dimension, restated from Part II)**.** Let $N(\epsilon)$ be the least number of boxes of side ϵ needed to cover a bounded set S. Its *box-counting dimension* is

$$D_0 = \lim_{\epsilon \to 0} \frac{\ln N(\epsilon)}{\ln (1/\epsilon)} , \tag{32.48}$$

when the limit exists. For a point, a smooth curve and a smooth surface it returns 0, 1 and 2. This is Definition 7.110 specialized twice over: the limit is written as though it exists, where that definition keeps the upper and lower limits apart and calls them $\overline{\dim}_{\mathrm{B}}$ and $\underline{\dim}_{\mathrm{B}}$; and the covering family is boxes of a fixed side rather than arbitrary sets of diameter at most δ. The two covering conventions give the same value whenever either limit exists, since a box of side ϵ has diameter $\epsilon\sqrt{3}$ and a set of diameter δ fits in a box of side δ, so the two counts differ by a factor that is constant in ϵ and drops out of the logarithmic ratio. Proposition 7.111 records that the Hausdorff dimension of a set never exceeds the lower box dimension, so a box count bounds the finer quantity from above and is not a substitute for it.

Rests on Definitions 6.5 and 7.110.

Example 32.80 (The middle-thirds Cantor set)**.** Remove the open middle third of $[0,1]$, then the middle thirds of the two remaining intervals, and so on. At stage k the set is covered by $N = 2^k$ intervals of length $\epsilon = 3^{-k}$, so Equation (32.48) gives

$$D_0 = \lim_{k \to \infty} \frac{\ln 2^k}{\ln 3^k} = \frac{\ln 2}{\ln 3} = 0.6309 ,$$

a dimension strictly between that of a point and that of a line. The horseshoe's invariant set Equation (32.17) is the product of two such sets, with 2 and 3 replaced by the map's stretching and contraction factors.

Rests on Definition 32.79 and Theorem 32.42.

Definition 32.81 (Correlation dimension)**.** For N points $\boldsymbol{x}_i$ sampled from a trajectory, let

$$C(\epsilon) = \lim_{N \to \infty} \frac{1}{N^2} \sum_{i \neq j} \Theta(\epsilon - |\boldsymbol{x}_i - \boldsymbol{x}_j|) \tag{32.49}$$

be the fraction of pairs closer than ϵ, with Θ the step function. Where $C(\epsilon) \propto \epsilon^{D_2}$ over a range of ϵ, the exponent D_2 is the *correlation dimension* [1720].

Rests on Definition 32.79.

Equation (32.49) is what makes fractal dimension an experimental quantity rather than a mathematical one: it needs only a list of points, it is computed by counting pairs, and it applies without change to a numerical trajectory and to an attractor reconstructed from a measured scalar series by the embedding of Remark 32.78. The measured values are collected in Table 32.3.

Remark 32.82 (Dimension and exponents, and a measure-theoretic gap). The Kaplan–Yorke estimate used in Table 32.3 is

$$D_{KY} = j + \frac{\lambda_1 + \cdots + \lambda_j}{|\lambda_{j+1}|} ,$$

Table 32.3: Correlation dimensions measured by pair counting Equation (32.49) on numerically generated trajectories, compared with the Kaplan–Yorke estimate built from the Lyapunov spectrum. The two need not agree: D_2 is a lower bound on the information dimension that the Kaplan–Yorke formula estimates.

Attractor	D_2 measured [1720]	Kaplan–Yorke estimate
Lorenz, $\sigma = 10$, $b = 8/3$, $r = 28$	2.05 ± 0.01	2.06
Hénon, $\hat{a} = 1.4$, $\hat{b} = 0.3$	1.21 ± 0.01	1.26

with j the largest index for which the partial sum is still non-negative: the dimension at which the volume growth rate changes sign. It is a conjecture, not a theorem, and its primary reference is not in this bibliography; it is quoted here as an estimate and never as a measured value.

More fundamentally, the canonical definition of a fractal dimension is Hausdorff's, built on an outer measure. Section 7.12 constructs it (Definition 7.108) without any Lebesgue theory, and Proposition 7.111 relates it to the box count: the Hausdorff dimension never exceeds the lower box-counting dimension, and the gap can be total — the rationals of $[0, 1]$ have Hausdorff dimension 0 and box dimension 1. This chapter works with Equation (32.48) and Equation (32.49) because they are what a finite data set can actually return; what they return is therefore an *upper* bound on the Hausdorff dimension of the attractor, never a determination of it. The general fractal vocabulary is Mandelbrot's [1696]. Together with the Lyapunov spectrum, the dimension is the second quantitative fingerprint of a chaotic attractor, and the two are the numbers that an experiment actually returns.

32.6 Universal routes to chaos

32.6.1 The logistic map

Definition 32.83 (The logistic map)**.** A population N_n in generation n, growing at fractional rate ρ per generation and limited by a carrying capacity K, obeys the difference equation $N_{n+1} = N_n\,[1 + \rho\,(1 - N_n/K)]$. Setting $x_n = \rho N_n / [(1+\rho)K]$ and $r = 1 + \rho$ turns it into the *logistic map*

$$x_{n+1} = f_r(x_n) = r\,x_n\,(1 - x_n)\ , \tag{32.50}$$

which maps $[0, 1]$ into itself for $0 < r \le 4$. Both x and r are dimensionless, being a population in units of a population and a rate per generation [1587].

Rests on Definition 32.3.

Proposition 32.84 (Fixed points and the first period doubling)**.** *Equation (32.50) has fixed points $x^* = 0$, with multiplier r, and $x^* = 1 - 1/r$, with multiplier $2 - r$. The origin is stable for $r < 1$; the second fixed point exists and is stable for $1 < r < 3$; at $r_1 = 3$ its multiplier passes through -1 and a stable two-cycle is born, which itself loses stability at*

$$r_2 = 1 + \sqrt{6} = 3.449\,490\,. \tag{32.51}$$

Rests on Equation (32.50) and Proposition 32.39.

Proof. *Derives Proposition 32.84.*

Solving $rx(1 - x) = x$ gives $x^* = 0$ and $x^* = 1 - 1/r$. Since $f_r'(x) = r(1 - 2x)$, the multipliers are $f_r'(0) = r$ and $f_r'(1 - 1/r) = r\,(1 - 2 + 2/r) = 2 - r$; Proposition 32.39 makes each fixed point stable exactly while its multiplier lies in $(-1, 1)$, which gives $r < 1$ and $1 < r < 3$. At $r = 1$ the multiplier is $+1$ and the two branches exchange stability:

the transcritical bifurcation of Proposition 32.24(i). At $r = 3$ it is -1, which is a period doubling.

For the two-cycle $\{p, q\}$, add and multiply the two conditions $q = rp(1-p)$ and $p = rq(1-q)$. Subtracting gives $p+q = (r+1)/r$, and substituting back gives $pq = (r+1)/r^2$. The multiplier of the cycle is the product of the derivatives,

$$f_r'(p)f_r'(q) = r^2(1-2p)(1-2q) = r^2\left[1-2(p+q)+4pq\right] = -r^2+2r+4\,,$$

using the two symmetric functions just found. Setting this to $+1$ returns $r = 3$, the birth of the cycle; setting it to -1 gives $r^2 - 2r - 5 = 0$, whose positive root is Equation (32.51). That is the second period doubling, and the process repeats. □

Remark 32.85 (The cascade, its windows, and the word "chaos"). The doublings continue: $r_3 = 3.544\,090$, $r_4 = 3.564\,407$, $r_5 = 3.568\,759$, and the sequence accumulates geometrically at

$$r_\infty = 3.569\,945\,6\,, \tag{32.52}$$

beyond which the map has orbits of every period and no stable cycle of any period over most of the remaining interval [1587]. The chaotic region is not solid: it is shot through with *windows* of stable periodic behaviour, the widest being the period-three window opening at $r = 1+\sqrt{8} = 3.828\,427$, where the third iterate f_r^3 becomes tangent to the diagonal in a saddle-node bifurcation (Proposition 32.23) creating a stable and an unstable three-cycle at once. The algebra of that tangency is not reproduced here and the value is quoted [1587].

The existence of a three-cycle is not a curiosity. Li and Yorke proved that a continuous map of an interval possessing a periodic point of period three has periodic points of every period, and an uncountable set of points whose orbits neither converge to any periodic orbit nor to each other [1553]; the paper is where the word "chaos" entered the mathematical literature. The theorem is quoted, not proved here.

Proposition 32.86 (The logistic map at $r = 4$ is exactly solvable)**.** *The change of variable $x = \sin^2(\pi\vartheta)$ conjugates f_4 to the doubling map $\vartheta \mapsto 2\vartheta \bmod 1$. Hence $x_n = \sin^2(2^n\pi\vartheta_0)$, the orbit is periodic exactly when ϑ_0 is rational, and the Lyapunov exponent is $\lambda = \ln 2$ per iterate for almost every ϑ_0.*

Rests on Equation (32.50) *and Definition 32.73.*

Proof. *Derives Proposition 32.86.*

With $x = \sin^2(\pi\vartheta)$,

$$f_4(x) = 4\sin^2(\pi\vartheta)\left[1-\sin^2(\pi\vartheta)\right] = \left[2\sin(\pi\vartheta)\cos(\pi\vartheta)\right]^2 = \sin^2(2\pi\vartheta)\,,$$

which is the same function of 2ϑ: the conjugacy holds and iterating gives x_n. Writing ϑ_0 in binary, the doubling map is the shift on its digits, so the orbit is eventually periodic precisely for rational ϑ_0—these form a set of zero length—and for every other starting point the orbit never repeats. The derivative of the doubling map is 2 at every point, so by Definition 32.73 the exponent in the ϑ variable is $\ln 2$ exactly; the conjugacy is smooth with nowhere-vanishing derivative except at the two endpoints, which the orbit visits with probability zero, so the exponent is unchanged in x. This is the cheapest complete example of chaos in the book: an explicit solution, an explicit positive exponent, and sensitivity that is nothing more than the loss of one binary digit of ϑ_0 per step—which is also exactly the symbolic dynamics of Proposition 32.43 with the two symbols the two binary digits. □

32.6.2 Feigenbaum universality

Phenomenon 32.87 (Universal period doubling)**.** As a control parameter is swept, a wide class of systems reaches chaos through an infinite cascade of period-doubling bifurcations

that accumulates at a finite parameter value. The parameter intervals between successive doublings shrink geometrically, and the ratio

$$\delta = \lim_{n\to\infty} \frac{\mu_n - \mu_{n-1}}{\mu_{n+1} - \mu_n} = 4.6692\ldots \tag{32.53}$$

is the same number for the logistic map, for a driven pendulum and for a convecting fluid, provided only that the return map of the system has a smooth quadratic maximum [1587, 1621]. This is the observed fact that a dimensionless constant of a difference equation is measurable in a laboratory: in a mercury Rayleigh–Bénard cell the cascade was followed through several doublings and δ recovered to within a few percent [1693].

Rests on Propositions 32.39 and 32.84.

Derivation. *Derives Phenomenon 32.87.*

Three things have to be shown: that the doublings recur without end, that their parameter intervals shrink geometrically so that the cascade accumulates at a finite parameter value, and that the ratio is a property of the cascade rather than of the map. The first two are derived here; the third is derived only as far as identifying what would have to be true, and the statement that it *is* true is quoted—see Remark 32.88.

The doublings recur. Proposition 32.84 carries the first two exactly: the fixed point $1 - 1/r$ loses stability at $r_1 = 3$ when its multiplier passes through -1, and the two-cycle born there loses stability the same way at $r_2 = 1 + \sqrt{6}$. The mechanism is self-reproducing. Let $\{p_1, \ldots, p_{2^n}\}$ be a stable 2^n-cycle at a parameter just below r_n. The map $f_r^{2^n}$ fixes each p_i, and the chain rule makes its derivative there the product $\prod_j f_r'(p_j)$ over the whole cycle—the same number at every p_i, since the factors are the same and multiply in any order—which is the multiplier that Proposition 32.39 grades and which reaches -1 at r_n. Take a small interval I_i about p_i, so small that $f_r^{2^n}$ maps it into itself; on it that iterate is smooth with negative slope at p_i. It also has a critical point there: the orbit of I_i visits, once per period, the interval containing the single critical point $x = 1/2$ of f_r, so by the chain rule $\left(f_r^{2^n}\right)'$ vanishes at the point of I_i whose orbit lands on $1/2$. That critical point is quadratic, because $f_r''(1/2) = -2r \neq 0$ and the chain rule multiplies it by factors that do not vanish. So $f_r^{2^n}$ restricted to I_i, reflected and rescaled to unit size, is again a unimodal map with a quadratic maximum and a multiplier sweeping downwards through -1: exactly the situation at r_1, one level up. It doubles again, and the argument applies at every level.

The intervals shrink geometrically. Write $\mathcal{R}$ for the operation just described—take f_r, iterate twice, restrict, reflect and rescale—and let $r \mapsto \mathcal{R}(r)$ denote the induced map on the parameter, defined by requiring the rescaled map at $\mathcal{R}(r)$ to look like the original at r. By construction it carries the nth doubling to the $(n-1)$th, so $\mathcal{R}(r_n) = r_{n-1}$, and it fixes the accumulation point, $\mathcal{R}(r_\infty) = r_\infty$, because a map at r_∞ has already doubled infinitely often and rescaling removes one doubling from an infinite supply. If $\mathcal{R}$ is differentiable at r_∞ with derivative δ, then linearizing about the fixed point in the manner of Definition 32.12 gives $r_{n-1} - r_\infty \approx \delta\,(r_n - r_\infty)$, hence

$$r_\infty - r_n \approx C\,\delta^{-n}\,, \qquad \frac{r_n - r_{n-1}}{r_{n+1} - r_n} \longrightarrow \delta\,, \tag{32.54}$$

which is Equation (32.53). Two consequences follow at once. The doubling parameters form a geometric sequence, so their sum converges and the cascade accumulates at a *finite* r_∞ after infinitely many doublings—a cascade that did not shrink geometrically would run off the end of the parameter range instead. And $\delta > 1$, since the doublings crowd together rather than spread apart, so r_∞ is a *repelling* fixed point of $\mathcal{R}$ in the parameter.

The arithmetic checks. Using the exact $r_1 = 3$ and $r_2 = 1 + \sqrt{6} = 3.449\,490$ of Proposition 32.84 and the numerically located $r_3 = 3.544\,090$, $r_4 = 3.564\,407$, $r_5 = 3.568\,759$ of Remark 32.85, the successive ratios in Equation (32.54) are

$$4.7514\,, \qquad 4.6562\,, \qquad 4.6684\,,$$

converging on $\delta = 4.6692$ from alternate sides. Summing the remaining geometric tail from r_5,

$$r_\infty \approx r_5 + \frac{r_5 - r_4}{\delta - 1} = 3.568\,759 + \frac{0.004\,352}{3.6692} = 3.569\,945\,1\,,$$

against the directly computed $r_\infty = 3.569\,945\,6$ of Equation (32.52): agreement in the seventh digit, which is the quantitative test that the convergence really is geometric with that ratio and not merely fast.

Why the ratio is not a property of the logistic map. Nothing in the first step used the quadratic form of Equation (32.50) beyond two facts: that the map has one critical point in the interval, and that the critical point is quadratic. Any system whose return map (Definition 32.38) has those two properties enters the same construction at the first level, and after one application of $\mathcal{R}$ it is described by a rescaled map of the same class. So the cascade of *every* such system is governed by the same operation $\mathcal{R}$, and if $\mathcal{R}$ has a single fixed point attracting that whole class, with one repelling direction of multiplier δ, then every member inherits that δ regardless of what it is made of—a difference equation, a driven pendulum, or a convecting layer of mercury. That is the content of the observation, and Remark 32.90 states the operator whose fixed point it is. Its existence is what is quoted [1621]. □

Remark 32.88 (What is quoted in the cascade, and what is not). The boundary is sharp and worth marking, in the manner of Remark 17.18. Derived above: that the doublings recur, that they accumulate geometrically at a finite parameter value, that the ratio is asymptotically the multiplier of a repelling fixed point of the doubling operation, and that the class of systems entering the construction is fixed by two local properties of the return map and by nothing else. Quoted: that the doubling operator $\mathcal{T}$ of Remark 32.90 actually *has* a fixed point g in a suitable space of unimodal maps, that its linearization there has exactly one eigenvalue outside the unit circle, and that the eigenvalue is $4.669\,201\,609\,102\,990\ldots$. Those three statements are a theorem of analysis, first established by a computer-assisted proof whose primary reference is not in this bibliography; the numbers themselves are quoted from [1621]. No measured quantity in this chapter depends on the quoted half: the comparison in Section 32.7.1 is between a measured ratio and a number computed from the logistic map's own cascade, and Equation (32.54) is what connects them.

Remark 32.89 (The two constants and their measured values). Two numbers govern the cascade [1621]. The first is δ of Equation (32.53), the rate at which the parameter intervals shrink,

$$\delta = 4.669\,201\,609\,102\,990\ldots\,,$$

and the second is the rate at which the attractor itself rescales: successive period-doubled orbits reproduce the previous pattern reduced by

$$\alpha = 2.502\,907\,875\,095\,893\ldots$$

in the state variable. Both are mathematical constants of the doubling operator, so they carry no uncertainty of their own; what carries uncertainty is their measurement in a physical system. For the logistic map the finite-order ratios approach δ from above—the first three, computed from the values in Proposition 32.84 and Remark 32.85, are 4.7514, 4.6562 and 4.6684—so an experiment that resolves only the first few doublings is not limited by that convergence but by its own resolution. In the mercury cell of Section 32.7.1 the measured value is $\delta = 4.4 \pm 0.1$ [1693], about 6 % below the exact constant; the deficit is the resolution of the higher doublings, whose parameter intervals shrink by the factor δ at each step and soon fall below the stability of the control.

Remark 32.90 (Why a difference equation and a fluid share a number). The mechanism is a renormalization group, in exactly the sense of Chapter 108. Define the doubling operator $(\mathcal{T}f)(x) = -\alpha\, f(f(-x/\alpha))$, which composes a map with itself and rescales so that the result is again normalized: it takes the description of a system at one level of the cascade to the description at the next. If $\mathcal{T}$ has a fixed point g, then g describes a system that looks the same after every doubling, and every map in the basin of g flows to it under repeated doubling and therefore inherits its numbers. Linearizing $\mathcal{T}$ about g, the fixed point turns out to have exactly one eigenvalue outside the unit circle, and that eigenvalue is δ; α is the scaling built into the operator. What is universal is thus the *unstable direction of a fixed point in a space of maps*, and what a physical system contributes is only which point of that space it starts from—which is why the only hypothesis needed is a smooth quadratic maximum. The existence of g is not elementary; it was established by a computer-assisted proof whose reference is not in this bibliography, and this chapter quotes the constants rather than deriving them.

32.6.3 Quasiperiodicity and intermittency

Period doubling is one of three routes by which a system is observed to become chaotic, and an experiment must distinguish them.

Definition 32.91 (The circle map)**.** The phase of a limit-cycle oscillator sampled at the period of an external drive obeys, to leading order, the *circle map*

$$\vartheta_{n+1} = \vartheta_n + \Omega - \frac{K}{2\pi}\sin\left(2\pi\vartheta_n\right) \pmod 1\,, \tag{32.55}$$

in which Ω is the ratio of the natural frequency of the oscillator to the drive frequency and K the dimensionless coupling. The *winding number* $W = \lim_{n\to\infty}(\vartheta_n - \vartheta_0)/n$ is the observed frequency ratio.

Rests on Phenomenon 32.37 and Definition 32.38.

Remark 32.92 (Mode locking, Arnold tongues and the torus). For $K < 1$ the winding number of Equation (32.55) is a continuous, non-decreasing function of Ω that is constant on an interval around every rational value—a *devil's staircase*. The intervals are the mode-locked states; plotted in the (Ω, K) plane they are the Arnold tongues, widening from zero width at $K = 0$ until at $K = 1$ they fill the axis and the irrational winding numbers form a set of zero length. That is the staircase heard by van der Pol and van der Mark (Phenomenon 32.37), and the critical line $K = 1$ is where the last quasiperiodic motion is destroyed. In the underlying flow, quasiperiodic motion with two incommensurate frequencies is motion on an invariant two-torus; mode locking is the capture of that motion onto a closed orbit on the torus, and the destruction of the torus at $K = 1$ is the second route to chaos, the one proposed by Ruelle and Takens [1453].

The line $K = 1$ deserves its own statement, because everything the quasiperiodic route claims turns on what happens as it is crossed.

Proposition 32.93 (The circle map loses invertibility at $K = 1$)**.** *For the map Equation (32.55),*

$$\frac{\mathrm{d}\vartheta_{n+1}}{\mathrm{d}\vartheta_n} = 1 - K\cos\left(2\pi\vartheta_n\right)\,. \tag{32.56}$$

Hence: for $K < 1$ the map is an orientation-preserving diffeomorphism of the circle; at $K = 1$ its derivative vanishes at the single point $\vartheta = 0$, where the second derivative vanishes as well, so the critical point is an inflection of cubic type; and for $K > 1$ the derivative is negative on an interval, the map is not injective, and it carries two quadratic critical points.

In the last case $|\mathrm{d}\vartheta_{n+1}/\mathrm{d}\vartheta_n|$ *reaches* $1+K>2$ *at* $\vartheta=\frac{1}{2}$*, so the map both stretches and folds.*

Rests on Definitions 32.38 and 32.91.

Proof. *Derives Proposition 32.93.*

Differentiating Equation (32.55) gives Equation (32.56) at once. Its minimum over the circle is $1-K$, attained at $\vartheta=0$, and its maximum is $1+K$, attained at $\vartheta=\frac{1}{2}$. So the derivative is strictly positive everywhere exactly when $K<1$, and a smooth circle map with everywhere positive derivative is a diffeomorphism: it is a local diffeomorphism, it is strictly increasing as a map of the line commuting with the translation by 1, and it therefore descends to a bijection of the circle preserving orientation.

At $K=1$ the derivative vanishes at $\vartheta=0$ and nowhere else. The second derivative of Equation (32.55) is $2\pi K\sin(2\pi\vartheta)$, which also vanishes at $\vartheta=0$, while the third derivative there is $4\pi^2K\neq 0$. The critical point is therefore cubic, not quadratic—which is why the critical line $K=1$ has universal numbers of its own and *not* those of Section 32.6.2, whose whole hypothesis was a quadratic maximum.

For $K>1$ the derivative is negative wherever $\cos(2\pi\vartheta)>1/K$, an interval about $\vartheta=0$, and positive elsewhere. A map that decreases on one interval and increases on another takes some values twice, so it is not injective; and its two turning points $\vartheta_\pm=\pm(2\pi)^{-1}\arccos(1/K)$ are quadratic, since there $\sin(2\pi\vartheta_\pm)=\mp\sqrt{1-K^{-2}}\neq 0$ makes the second derivative nonzero. The bound on the derivative is read off the maximum computed above. □

Remark 32.94 (What that proposition does and does not settle). The consequences run in both directions. Below $K=1$ the return map is a circle diffeomorphism, and a circle diffeomorphism cannot be chaotic: its orbits are ordered on the circle for all time, so nearby points cannot separate and later re-approach in the way Phenomenon 32.72 requires. The invariant circle of the section—the two-torus of the flow—survives, and the motion on it is either mode-locked or quasiperiodic, exactly the dichotomy of Remark 32.92. Above $K=1$ that protection is gone. Proposition 32.93 supplies precisely the two ingredients Section 32.3.2 identified as sufficient: stretching, by a factor up to $1+K$, and folding, at the two quadratic turning points. Two points with different histories can now share an image, so the past is not recoverable from the present, and the invariant set of the section is no longer a smooth circle—there is no smooth invariant two-torus left to carry the motion. That is the torus breakdown, in the one model where it can be seen explicitly.

Two things are quoted rather than proved. That an orbit of a circle diffeomorphism has a well-defined winding number independent of the starting point, and that an irrational winding number forces the map to be semi-conjugate to the rigid rotation by it, are Poincaré's rotation-number theorem and Denjoy's theorem; both are ordinary differential equations and circle dynamics, and belong with the material of Chapter 9, which does not carry them. And the general classification—that after three generic bifurcations from a fixed point an invariant three-torus is typically replaced by an attractor with a continuous spectrum—is the Newhouse–Ruelle–Takens theorem, quoted here from [1453]. Its proof is a genericity argument in a space of vector fields, machinery that appears nowhere else in this treatise, and Remark 17.18 is the precedent for quoting such a thing rather than owing it. Nothing measured in Section 32.7.2 depends on the quoted half: what the experiment tests is the presence of a continuous spectral component after a small number of transitions, and that is decided by the argument given in Phenomenon 32.97.

Proposition 32.95 (Intermittency: the scaling of the laminar phase)**.** *Let a stable and an unstable fixed point of a return map annihilate in a saddle-node bifurcation (Proposition 32.23) at a parameter* μ_c*, and let the map beyond the bifurcation be* $x_{n+1}=x_n+\epsilon+ax_n^2$ *with* $\epsilon\propto\mu-\mu_c>0$ *and* $a>0$*. Then the orbit spends a mean number of iterations*

$$\langle\ell\rangle\simeq\frac{\pi}{\sqrt{a\epsilon}}\propto(\mu-\mu_c)^{-1/2} \tag{32.57}$$

in the narrow channel where the fixed points used to be, producing long stretches of nearly periodic behaviour interrupted by bursts [1657].

Rests on Proposition 32.23 and Definition 32.38.

Proof. *Derives Proposition 32.95.*

In the channel the increment per iterate is small, so the difference equation may be replaced by the differential equation $\mathrm{d}x/\mathrm{d}n = \epsilon + ax^2$. Separating variables and integrating across the channel,

$$\ell = \int_{-\infty}^{+\infty} \frac{\mathrm{d}x}{\epsilon + ax^2} = \frac{1}{\sqrt{a\epsilon}} \left[\arctan\left(x\sqrt{\frac{a}{\epsilon}} \right) \right]_{-\infty}^{+\infty} = \frac{\pi}{\sqrt{a\epsilon}} ,$$

the limits being taken to the edges of the channel, where the quadratic approximation fails but the integrand is already negligible. That is Equation (32.57). The signature is a distinguishing one: a system approaching chaos by intermittency shows periodic behaviour interrupted by bursts whose mean spacing diverges as an inverse square root of the distance from threshold, and no subharmonics at all. □

Remark 32.96 (The three-route checklist). An experiment sweeping a control parameter therefore has three signatures to look for, and they are mutually exclusive in the power spectrum. *Period doubling*: a sequence of new lines at $f/2$, $f/4$, $f/8$, at parameter intervals shrinking by Equation (32.53). *Quasiperiodicity*: a second line incommensurate with the first, then broadband. *Intermittency*: no new lines at all, but bursts interrupting a periodic signal, with the laminar length scaling as Equation (32.57). Each has been observed. The measurements in Section 32.7 are the record of the first two.

32.7 Experimental chaos

32.7.1 Rayleigh–Bénard convection

The convection cell is the workhorse of experimental chaos because its control parameter, the Rayleigh number Equation (32.38), is set by a temperature difference and can be held and swept with great stability. The background—Bénard's cells and Rayleigh's stability calculation—belongs to Section 31.8.2; for a real cell with rigid upper and lower boundaries the critical value is $\mathrm{Ra}_c = 1708$, not the 657.5 of the stress-free idealization used in Definition 32.62.

Libchaber, Laroche and Fauve used liquid mercury, whose Prandtl number $\sigma = \nu/\kappa$ is about 0.025—from $\nu \approx 1.1 \times 10^{-7}\,\mathrm{m^2/s}$ and $\kappa \approx 4.4 \times 10^{-6}\,\mathrm{m^2/s}$—and is therefore two orders of magnitude below that of water. A small Prandtl number makes the convecting rolls oscillate rather than remain steady at modest Ra, which is exactly what is wanted: the oscillation supplies the periodic orbit whose doubling is to be followed. The cell was small enough to contain only a few rolls, so that the number of active degrees of freedom stayed low, and a horizontal magnetic field was applied to damp the transverse motions that would otherwise introduce further frequencies. The temperature was recorded by a bolometer inside the fluid, and the power spectrum of that signal was computed as Ra was raised [1693].

What the spectrum showed was the cascade of Remark 32.96: the fundamental oscillation, then a line at half its frequency, then at a quarter, then at an eighth, each appearing at a value of Ra closer to the last than the previous interval by a factor near δ. From the ratios of the successive intervals the experiment returns

$$\delta_{\mathrm{measured}} = 4.4 \pm 0.1 , \tag{32.58}$$

against the exact $4.669\ldots$ of Remark 32.89 [1693]. The agreement is at the 6 % level, and the shortfall is a resolution effect: each successive interval is smaller than the last by δ, so

the fourth doubling requires holding the temperature difference to a part in $\delta^3 \approx 100$ of the first interval.

That is the whole evidential point of this chapter. A constant computed from a one-line difference equation with no physical content whatever appears, to a few percent, in the temperature record of a centimetre of mercury—because both systems are governed by the same fixed point of the doubling operator of Remark 32.90, and by nothing else they share.

32.7.2 Taylor–Couette flow

Fluid between concentric cylinders, the inner one rotating, becomes unstable to a stack of toroidal vortices above a critical rotation rate; Taylor computed the threshold and measured it in the same paper, obtaining what was then the first quantitative agreement between a hydrodynamic stability calculation and an experiment [505]. In the narrow-gap limit the criterion is a critical Taylor number of 1708, the same number as the rigid-boundary Rayleigh–Bénard threshold, the two problems having the same linear operator.

Gollub and Swinney took that apparatus into the chaotic regime and measured, by laser-Doppler velocimetry, the radial velocity at a point in the fluid as the rotation rate was increased through and well beyond the Taylor threshold [1545]. Velocimetry is what makes the experiment decisive: it delivers a time series long enough and clean enough for a power spectrum, in a flow that a visualization would only show to be complicated.

Phenomenon 32.97 (Turbulence sets in after a few transitions)**.** In a fluid confined between rotating cylinders, the power spectrum of the velocity acquires, as the rotation rate is raised, first one sharp frequency, then a second incommensurate with it, and then — after only a small number of such steps — a broadband continuous component that grows until it dominates [1545]. The measurement excludes Landau's picture of turbulence as the accumulation of an unbounded sequence of incommensurate frequencies, which would show a spectrum of ever more discrete lines and no continuum, and supports instead a route in which a low-dimensional attractor with a continuous spectrum appears after a few bifurcations [1453]. The underlying instability of the rotating flow is Taylor's [505].

Rests on Remark 32.69 and Definition 32.67.

Derivation. *Derives Phenomenon 32.97.*

What has to be derived is why the shape of the spectrum decides between the two pictures at all, since both predict complicated motion.

Landau's flow after n instabilities is quasiperiodic with n incommensurate frequencies: by construction the velocity at a point is

$$v(t) = V\left(\omega_1 t, \dots, \omega_n t\right) ,$$

with V continuous and 2π-periodic in each of its arguments. Such a V has a multiple Fourier series, so

$$v(t) = \sum_{\boldsymbol{k}\in\mathbb{Z}^n} c_{\boldsymbol{k}}\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\omega}t} , \tag{32.59}$$

and all the power sits on the set $\{\boldsymbol{k}\cdot\boldsymbol{\omega} : \boldsymbol{k}\in\mathbb{Z}^n\}$, which is *countable* for every finite n. A countable set carries no continuous spectral component, whatever its density: the spectrum of a quasiperiodic signal consists of lines and nothing else, and this is true however large n becomes. Landau's route therefore predicts a spectrum that grows ever denser in lines and never acquires a continuum, at any stage.

A continuous component in the spectrum requires an autocorrelation that decays. If v has an integrable autocorrelation then its transform is a continuous function tending to zero at large frequency, by the Riemann–Lebesgue lemma (Lemma 17.7); conversely Equation (32.59) has an autocorrelation $\sum_{\boldsymbol{k}} |c_{\boldsymbol{k}}|^2\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\omega}\tau}$ which is itself quasiperiodic and

never decays. So the appearance of a broadband component is precisely the statement that correlations are being lost—which is what a positive Lyapunov exponent does, by Equation (32.44), and what motion on a torus cannot do.

The measurement therefore reads as follows. Sharp lines appearing one at a time are consistent with both pictures. The appearance of a broadband component, after a number of transitions that can be counted on one hand, is consistent with only one of them. Gollub and Swinney observed exactly that: one frequency at the Taylor transition, a second incommensurate frequency, and then, at a rotation rate of order ten times the critical one, a continuous component that could not be resolved into further lines [1545]. The exact multiple depends on the radius ratio and aspect ratio of the cell and is not a universal number; the number of transitions before the continuum appears is the observable that matters, and it is small. □

The derivation above settles the spectral half of the reading: a quasiperiodic flow of *any* number of frequencies has a pure point spectrum, so a continuous component cannot be Landau's accumulation. The other half—how the invariant two-torus is destroyed once the coupling is strong enough—is Proposition 32.93, whose loss of invertibility at $K = 1$ produces the stretching and folding that a torus cannot support, and Remark 32.94 states exactly which part of the general classification is quoted from [1453] and which is derived. Nothing Gollub and Swinney measured depends on the quoted part.

32.7.3 Circuits and driven oscillators

Electronics is where chaos was first heard, and it remains where it is most easily produced. A circuit has no boundary layers, no aspect ratio and no thermal inertia; its state variables are two voltages and a current, its parameters are components with catalogue values, and its time series is available directly on an oscilloscope.

The first instance is van der Pol and van der Mark's neon-tube oscillator of Phenomenon 32.37, in which the irregular noise heard between locking steps was reported in 1927 and understood half a century later [590]. The modern bench standard is Chua's circuit, an inductor, two capacitors, a linear resistor and one nonlinear resistor with a piecewise-linear characteristic. In SI variables its equations are

$$C_1 \frac{\mathrm{d}v_1}{\mathrm{d}t} = \frac{v_2 - v_1}{R} - g(v_1)\,, \tag{32.60}$$

$$C_2 \frac{\mathrm{d}v_2}{\mathrm{d}t} = \frac{v_1 - v_2}{R} + i_L\,, \tag{32.61}$$

$$L \frac{\mathrm{d}i_L}{\mathrm{d}t} = -v_2\,, \tag{32.62}$$

with v_1, v_2 in V, i_L in A, C_1, C_2 in F, L in H, R in Ω, and the nonlinear element's characteristic

$$g(v) = G_b v + \tfrac{1}{2}\,(G_a - G_b)\,(|v + E| - |v - E|)\,, \tag{32.63}$$

whose slopes G_a, G_b are in S and whose breakpoint E is in V. The system is three-dimensional and autonomous, the minimum Remark 32.9 permits, and it is dissipative wherever the total conductance is positive. Equations (32.60) to (32.62) were shown numerically to possess a chaotic attractor by Matsumoto [1742], and the same attractor is produced on a breadboard in an undergraduate afternoon; sweeping R carries the circuit through a period-doubling cascade of the kind measured in Section 32.7.1.

The mechanical counterpart is the periodically driven, damped pendulum,

$$m\ell^2 \ddot{\theta} + b\dot{\theta} + mg\ell \sin\theta = \Gamma \cos(\omega_d t)\,, \tag{32.64}$$

with the drive torque Γ in N m. Writing the drive phase as a third variable makes Equation (32.64) an autonomous three-dimensional system, and taking a Poincaré section once per drive period (Definition 32.38) reduces it to a two-dimensional map; in the limit of weak damping and impulsive driving that map is the standard map Equation (32.25). The same apparatus that measures the isochronism of Chapter 34 at small amplitude becomes, driven hard, one of the cleanest demonstrations of a strange attractor available.

In every one of these systems the quantitative claim—that the motion is chaotic and not merely complicated—rests on the same measurement chain: record a scalar time series, embed it, extract λ_1 by the algorithm of [1766], and estimate D_2 by pair counting Equation (32.49). A positive exponent and a non-integer dimension, measured from data with stated uncertainties, are what distinguish deterministic chaos from noise. That distinction—and not any of the mathematics above—is what makes chaos an experimental subject.

Chapter 33

Experiment: The Cavendish Torsion Balance

Tests Equation (33.1).

Assuming Phenomenon 33.4 and Equation (33.6).

In 1797–1798 Henry Cavendish measured, in an outbuilding at Clapham, the gravitational attraction between lead spheres on a laboratory bench [44]. The declared aim was the mean density of the Earth; read retrospectively, it is the first laboratory determination of the constant G that fixes the strength of gravitation, and the archetype of every precision weak-force measurement since. Its role in this part is evidential: the inverse-square law of Chapter 27 rests observationally on celestial mechanics, and Cavendish supplies the laboratory half—the same law, with the same constant, between metre-scale masses. That universality is what the tests of Part V will probe to far greater precision.

Two threads run through the chapter and are treated with full honesty. The first is the theory of the torsion pendulum, which is what converts an observed angle into a force, and which does so *absolutely*: the instrument's force scale is fixed by a length, a mass distribution and a time, with no force standard anywhere in the chain. The second is the notorious mutual disagreement of the modern determinations. Six laboratory measurements quoting uncertainties of tens of parts per million scatter over some five hundred parts per million, and the recommended value therefore carries an uncertainty enlarged by a factor of nearly four to cover a spread nobody has explained [2286, 2392, 2431]. G is, by a wide margin, the least precisely known of the fundamental constants, and this chapter says so plainly rather than quoting the recommended value as though it were settled.

Remark 33.1 (Units: the original paper is imperial, this chapter is not). Cavendish reports every length in feet and inches and every mass in pounds, and quotes his result as a pure number—the ratio of the mean density of the Earth to that of water. Editorial rule 3 of this treatise admits SI only, so every apparatus figure below has been converted, and the density ratio has been multiplied by $1.0 \times 10^{3}\,\mathrm{kg/m^{3}}$ to become a density. The conversions are exact by definition (one inch is 25.4 mm, one pound is 0.453 592 37 kg), so nothing is lost; what is gained is that the numbers may be compared directly with the modern ones without a second act of translation. The uncertainty conventions are those of Chapter 2 and [2129]: a parenthesised figure is a standard uncertainty on the last digits quoted.

33.1 Historical context and the prediction under test

The experiment tests a law that had been believed for over a century and never once checked in a laboratory. This section states what the law asserts, shows why the astronomical evidence for it is structurally incapable of yielding the constant that appears in it, and records the one earlier attempt to weigh the Earth by gravitation alone.

33.1.1 Universal gravitation without a constant

Newton's law of universal gravitation, as this treatise states it, is

$$\boldsymbol{F}_{12} = -G\,\frac{m_1 m_2}{r^2}\,\hat{\boldsymbol{r}}_{12}\,, \qquad r = |\boldsymbol{r}_2 - \boldsymbol{r}_1|\,, \tag{33.1}$$

with $\hat{\boldsymbol{r}}_{12}$ the unit vector from body 1 to body 2, so that the force on body 1 points towards body 2. The constant G has the SI dimension $\mathrm{m}^3/(\mathrm{kg\,s}^2)$, which is already an unusual thing for a coupling constant: it is not dimensionless, and no combination of the other constants of nature produces it.

That form is not Newton's. The *Principia* argues throughout in proportions—the force is "as the product of the masses and reciprocally as the square of the distance"—and no symbol for a constant of proportionality appears anywhere in it [14]. Nor could one have been assigned a value, because the eighteenth century possessed no measurement from which to extract it. That is not an accident of technique but a structural feature of the astronomical evidence, and it is worth stating exactly.

Proposition 33.2 (Celestial mechanics determines only the products Gm)**.** *Let a system of bodies of masses m_a move under Equation (33.1) alone. Then every trajectory of the system depends on G and on the masses only through the n products Gm_a. In particular, no observation of orbital motion—however precise, and of however many bodies—can separate G from a mass. For a small body in orbit about a much larger one this specialises to Kepler's third law in the form*

$$T^2 = \frac{4\pi^2 a^3}{G\,(M+m)}\,, \tag{33.2}$$

whose right-hand side contains G nowhere except in that product.

Rests on Equation (33.1).

Proof. *Derives Proposition 33.2.*

The equation of motion of body b is, by Equation (33.1) and the second law of Chapter 19,

$$m_b\,\ddot{\boldsymbol{r}}_b = -\sum_{a\neq b} G\,\frac{m_a m_b}{|\boldsymbol{r}_b - \boldsymbol{r}_a|^2}\,\hat{\boldsymbol{r}}_{ba} \qquad\Longrightarrow\qquad \ddot{\boldsymbol{r}}_b = -\sum_{a\neq b} (Gm_a)\,\frac{\hat{\boldsymbol{r}}_{ba}}{|\boldsymbol{r}_b - \boldsymbol{r}_a|^2}\,. \tag{33.3}$$

The inertial mass m_b has cancelled—itself an experimental fact and not a theorem, as Chapter 20 insists—and what survives on the right is the set of products Gm_a and nothing else. The initial-value problem for Equation (33.3) therefore has the same solution for any (G, m_a) with the products fixed, so no trajectory distinguishes them. For the two-body case, reduction to the relative coordinate replaces Gm_a by $G(M+m)$ and the closed elliptical orbit gives Equation (33.2) (Chapter 27).

The consequence for practice is sharp. Astronomy measures the products GM for the Sun, the Earth and the planets far more precisely than any laboratory measures G — and cannot convert any of them into a mass, because the conversion needs exactly the factor the observations do not contain. To weigh the Earth, one must measure a gravitational force between two bodies *whose masses are known by weighing*, which means bodies small enough to put on a balance. That is the whole rationale of the experiment. □

The eighteenth-century way of putting the same point was to ask for the mean density of the Earth relative to water, a pure number then called the "density of the Earth". Newton himself had guessed it in Book III of the *Principia*, reasoning from the density of surface rocks and the compression expected at depth, at five to six times that of water [14]—a guess that turned out to be remarkably good, and that framed the problem for the next century as one of *weighing the Earth* rather than of measuring a constant.

Phenomenon 33.3 (Gravitation acts between laboratory masses)**.** Two ordinary bodies on a bench attract one another. The attraction is far too weak to notice — a lead sphere of some 160 kg pulls a companion of under a kilogram with a force of order $1.5 \times 10^{-7}\,\mathrm{N}$, some tens of millions of times smaller than the weight of the smaller body — but it is measurable, it is proportional to the product of the two masses and to the inverse square of the distance between their centres, and it is directed along the line joining them. That the law Newton inferred from the Moon and the planets [14] also holds between metre-scale masses in a closed room is not a deduction from the celestial evidence; it is a separate observation, and it is the observation this chapter reports.

Rests on Equation (33.1) *and Proposition 33.2.*

Derivation. *Derives Phenomenon 33.3.*

The magnitude asserted follows from the law together with the measured constant. Taking $G = 6.674\,30 \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$ [2392], a source sphere of $M = 157.9\,\mathrm{kg}$ and a suspended ball of $m = 0.730\,\mathrm{kg}$ whose centres stand $r = 0.2248\,\mathrm{m}$ apart attract with

$$F = \frac{GMm}{r^2} = \frac{6.6743 \times 10^{-11} \times 157.9 \times 0.730}{0.050\,531}\,\mathrm{N} = 1.52 \times 10^{-7}\,\mathrm{N}\,, \tag{33.4}$$

against a weight $mg = 7.16\,\mathrm{N}$ for the very same ball: a ratio of 2.1×10^{-8}. No balance weighs a difference that small directly. Each sphere is treated here as a point mass at its centre, which for a homogeneous sphere is exact and not an approximation (Chapter 27); the correction for the actual inhomogeneity of a cast lead sphere is one of the systematics of Section 33.2.3.

The smallness of the ratio is the design constraint of the whole experiment. It is why the measurement is made as a torque against a torsion fibre rather than as a force against gravity (Section 33.2.1); why the apparatus must be sealed against air currents, whose forces on a 50 mm sphere are orders of magnitude larger (Section 33.2.3); and — ultimately — why G remains the least precisely known of the constants (Section 33.7.4). □

33.1.2 The Schiehallion experiment

One serious attempt preceded Cavendish. In 1774 Nevil Maskelyne, the Astronomer Royal, took a zenith sector to the isolated Perthshire mountain Schiehallion and measured the apparent latitude of two stations, one on the north flank and one on the south, by star transits. The mountain's own attraction pulls the plumb-line sideways at each station, in opposite senses, so the astronomical difference in latitude exceeds the geodetic difference measured by survey. The excess he found was 11.6 arcseconds, that is $5.62 \times 10^{-5}\,\mathrm{rad}$ of total deflection shared between the two sides [33].

Converting a deflection into a density requires the mountain's volume and shape, and it was Charles Hutton who supplied them, from an elaborate survey of the massif; to carry out the integration he invented the device of joining points of equal height by a line, which is the contour of every topographic map since. Hutton's reduction gave a mean density for the Earth of about 4.5 times that of water, that is about $4.5 \times 10^3\,\mathrm{kg/m^3}$ [35].

The result is low by some 18 % against the modern figure, and the reason is instructive: a mountain is an attractor whose mass distribution and internal density are not known, and Hutton had to assume a rock density for the whole massif. The error enters the answer in full. A lead sphere, by contrast, can be weighed on a balance and its dimensions measured with callipers, so that its mass and the position of its centre are known to the precision of ordinary metrology. That, and not any improvement in sensitivity, is what makes the closed-laboratory experiment the decisive one: it replaces an attractor that must be modelled by one that can be weighed.

33.1.3 Michell's apparatus

The instrument was not Cavendish's. It was conceived and built by the Reverend John Michell, who died before he could use it; it passed to Francis John Hyde Wollaston and thence to Cavendish, who rebuilt it substantially before beginning work. Cavendish says so in the opening pages of his paper and credits Michell with the design throughout [44]. The same instrument had been arrived at independently, and for a different purpose, by Coulomb, whose memoir on the torsion of metal wires established the linear restoring law and applied it to electrostatics [36, 38].

Why torsion? Because it is the only elastic degree of freedom soft enough. A body suspended from a fine fibre is held against gravity by the fibre's tension—an enormously stiff constraint—while being almost free to rotate about the fibre's axis, and the two stiffnesses can differ by ten orders of magnitude. Numerically, the balance described in Section 33.3 has a torsional stiffness of order $6\times10^{-5}\,\mathrm{N\,m/rad}$, which turns the $1.5\times10^{-7}\,\mathrm{N}$ of Equation (33.4) into a displacement of a few millimetres at the end of a metre-long beam: a signal that a telescope and a scale can read. The same force applied against the suspension's vertical stiffness would move the ball by a distance far below atomic dimensions.

33.2 Torsion pendulum theory

Everything the instrument reports is an angle. This section sets out the theory that converts angles into forces, and—because that theory is what the modern discrepancy turns on—states carefully which of its ingredients are theorems and which are experimental facts about a piece of wire.

33.2.1 The torsion pendulum

Phenomenon 33.4 (A twisted fibre restores in proportion to the twist)**.** A fine wire or fibre, twisted through a small angle about its own axis, resists with a torque proportional to the angle, and the constant of proportionality is a property of the fibre alone — not of the body hung from it, and not of the amplitude [36]. In consequence a body suspended from such a fibre oscillates torsionally with a period independent of how far it was displaced, and that period, together with the body's moment of inertia, suffices to determine the constant. This is what makes the torsion balance an *absolute* instrument: its force scale is fixed by a length, a mass distribution and a time, with no separate force standard anywhere in the chain.

Rests on Equation (9.106).

Derivation. *Derives Phenomenon 33.4.*

Let the restoring torque be $\tau=-\kappa\theta$, with κ in $\mathrm{N\,m/rad}$, and let the suspended body have moment of inertia I about the fibre axis. The rotational form of the second law (Chapter 29) gives

$$I\ddot{\theta}=-\kappa\theta\,, \tag{33.5}$$

which is Equation (9.106) with $\lambda^2=\kappa/I$ and time in place of the spatial variable. Its solutions are therefore those of Equation (9.108), $\theta(t)=A\cos\omega t+B\sin\omega t$ with $\omega=\sqrt{\kappa/I}$, and every one of them has the same, amplitude-independent period

$$T=2\pi\sqrt{\frac{I}{\kappa}}\,,\qquad\text{so that}\qquad \kappa=\frac{4\pi^2 I}{T^2}\,. \tag{33.6}$$

From Equations (9.108) *and* (33.5) *(reading off* $\omega=\sqrt{\kappa/I}$ *from the harmonic solution and inverting* $T=2\pi/\omega$*).*

Here I is computed from the measured masses and lengths of the beam and its balls, and T is timed; κ follows without any force having been applied to the instrument at all. An equilibrium deflection θ produced by the source masses then reports the gravitational torque as $\kappa\theta$, and hence the force.

Two cautions belong with the result, and both return later. The linearity is an experimental fact about the fibre, not a theorem, and it holds only for small twists; and its slow departures — creep, and the frequency-dependent anelasticity that makes κ differ between a static deflection and an oscillation — are precisely the systematic of Remark 33.10. □

Remark 33.5 (Why the fibre must be fine, quantitatively)*.* For a cylindrical wire of radius a and length ℓ made of a material of shear modulus μ, the elastic theory of Chapter 30 gives

$$\kappa = \frac{\pi \mu a^4}{2\ell}\,. \tag{33.7}$$

The fourth power is the design lever: halving the radius divides the stiffness by sixteen and multiplies the deflection by the same factor, at the cost of a suspension that must still carry the weight of the beam in tension. Cavendish's wire was of copper–silver; Boys later replaced it with drawn quartz [262], which is both finer and far more nearly perfectly elastic — the single largest instrumental advance between 1798 and the twentieth century. The symbol μ is used here for the shear modulus to keep G free for the gravitational constant.

Rests on Equation (33.6)*.*

Real fibres dissipate, and the balance is in practice a damped oscillator: with a retarding torque $-b\dot\theta$ the equation of motion becomes

$$I\ddot\theta + b\dot\theta + \kappa\theta = \tau_{\mathrm{ext}}(t)\,, \tag{33.8}$$

which is Equation (17.88) with the rotational variables in place of the linear ones. Its free solutions decay as $\mathrm{e}^{-\Gamma t}$ with $\Gamma = b/(2I)$, and the *quality factor* $Q = I\omega_0/b = \omega_0/(2\Gamma)$ counts, up to a factor 2π, the number of radians of oscillation in one decay time (Example 17.75). For the fibres used in this experiment Q runs from a few hundred to a few thousand: the free swing persists for hours, which is convenient for timing the period and inconvenient for waiting out a deflection — a tension resolved in Section 33.4.

33.2.2 From deflection to attraction

Proposition 33.6 (The working formula of the torsion balance)**.** *Let a beam carry two small balls of mass m at distance d from the suspension axis, and let two source spheres of mass M be placed one beside each ball, on opposite sides of the beam, with centre-to-centre separation r and the line of centres perpendicular to the beam. Let T be the free torsional period and $s = d\theta$ the linear displacement of a ball from its undisturbed position. Then, in the idealisation that the moment of inertia is that of the two balls alone and that each sphere attracts only its neighbour,*

$$G = \frac{4\pi^2 r^2 s}{M\,T^2}\,. \tag{33.9}$$

Neither the mass m of the suspended balls nor the half-length d of the beam appears: the instrument needs no knowledge of what is hanging from it.

Rests on Equations (33.1) *and* (33.6)*.*

Proof. *Derives Proposition 33.6.*

Each source sphere pulls its neighbouring ball with the force Equation (33.1), directed perpendicular to the beam and therefore wholly tangential; the two forces act at opposite ends and in opposite senses, so their torques add:

$$\tau_g = 2\,\frac{GMm}{r^2}\,d\,. \tag{33.10}$$

At equilibrium this is balanced by the fibre, $\tau_g = \kappa\theta$, and κ is eliminated in favour of the measured period by Equation (33.6) with $I = 2md^2$:

$$\theta = \frac{\tau_g}{\kappa} = \frac{2GMmd}{r^2} \cdot \frac{T^2}{4\pi^2\, 2md^2} = \frac{GMT^2}{4\pi^2 r^2 d}\,. \tag{33.11}$$

From Equations (33.6) and (33.10) (substituting $\kappa = 4\pi^2 I/T^2$ with $I = 2md^2$, so that m cancels between torque and inertia).

Both m and one power of d have cancelled. Writing $s = d\theta$ for the displacement actually read on the scale removes the second, and solving for G gives Equation (33.9).

The cancellation of m is not a coincidence and is worth naming: the gravitational torque is proportional to m and so is the inertia that sets the period, so the suspended mass drops out of the ratio. This is the same cancellation of gravitational against inertial mass that makes free fall universal (Chapter 20), appearing here as a property of the instrument rather than of the trajectory. □

Remark 33.7 (What the idealisation costs, and what repairs it). Equation (33.9) is a two-line formula and it is not the experiment. Four corrections are needed before it becomes one, and each was computed by Cavendish himself [44] and systematised by Boys [262]:

1. *The beam's own inertia.* $I = 2md^2 + I_{\text{beam}}$, and G is proportional to I, so a beam of a few hundred grams changes the answer by several percent. This is the largest of the four and it is measured, not estimated: the beam is weighed and its dimensions taken.

2. *Cross-attraction.* Each source sphere also pulls the *far* ball, along a direction nearly parallel to the beam, which contributes little torque but not none, and in the opposite sense.

3. *The rest of the apparatus.* The case, the mounting and the suspension wire all attract the balls; only the difference between the two source-mass positions survives the reversal of Remark 33.8, but the source spheres' attraction on the *beam* does not cancel and must be computed.

4. *Finite size.* A homogeneous sphere attracts an external point exactly as a point mass at its centre (Chapter 27), so sphericity costs nothing; departures from homogeneity in a cast lead sphere do.

Applied to Cavendish's own headline numbers — $r = 0.2248\,\text{m}$, $s = 4.1 \times 10^{-3}\,\text{m}$, $M = 157.9\,\text{kg}$, $T = 9.0 \times 10^2\,\text{s}$ — Equation (33.9) alone returns $G \approx 6.33 \times 10^{-11}\,\text{m}^3/(\text{kg}\,\text{s}^2)$, five percent below the modern value. Correction 1 has the right sign and the right size to account for essentially all of that shortfall. It is the difference between a five-percent estimate and Cavendish's one-percent result.

Rests on Equation (33.9).

Remark 33.8 (Reversal: the trick that makes the zero irrelevant). The beam has an unknown rest position, set by the fibre's own untwisted orientation, and that position drifts. Cavendish therefore never measures a deflection from a zero: he swings the source spheres from one side of the beam to the other, reversing the sign of τ_g, and measures the *difference* of the two equilibrium positions, which is $2s$. Any static offset, and any drift slow compared with a reversal cycle, cancels identically in the difference, while the signal doubles. Every torsion measurement of G since has been differential in the same way; the modern versions modulate the source masses continuously and detect at the modulation frequency, which is the same idea with the Fourier analysis of Chapter 17 placed after the detector.

Rests on Equation (33.9).

Proposition 33.9 (The time-of-swing method)**.** *Let the source masses be placed instead in two configurations, "near" and "far", chosen so that the gravitational torque on the beam varies with the deflection angle, with gradients $\partial\tau_g/\partial\theta = GC_1$ and GC_2 respectively, where $C_1 \neq C_2$ are geometrical factors carrying the masses and distances. Let ω_1, ω_2 be the measured angular frequencies of free torsional oscillation in the two configurations. Then*

$$G = \frac{I\,(\omega_2^2 - \omega_1^2)}{C_1 - C_2}\,, \tag{33.12}$$

in which the fibre constant κ has cancelled entirely.
Rests on Equations (33.5) *and* (33.10)*.*

Proof. *Derives Proposition 33.9.*

Expanding the total torque about the equilibrium angle, the effective stiffness is the fibre's stiffness reduced by the gravitational gradient,

$$I\ddot{\theta} = -\left(\kappa - \partial_\theta \tau_g\right)\theta \qquad \Longrightarrow \qquad I\omega_i^2 = \kappa - GC_i\,, \quad i = 1, 2\,. \tag{33.13}$$

Subtracting the second instance from the first removes κ identically: $I(\omega_2^2 - \omega_1^2) = G(C_1 - C_2)$, which is Equation (33.12). Nothing about the fibre survives the subtraction — provided there is one κ to subtract. □

The attraction of the method is that it replaces a displacement, read against a scale to perhaps a part in 10^3, by a frequency, which can be timed to a part in 10^6 or better; it is the method of Heyl [657] and of Luther and Towler [1695]. Its weakness is exactly the step that looks like its strength. The cancellation of κ in Equation (33.13) assumes one number κ; if the fibre's stiffness depends on the frequency at which it is interrogated, the two configurations do not share it, and the difference survives as a bias. That is Remark 33.10.

33.2.3 Systematics of the method

Cavendish's paper is, among other things, the first systematic error budget in experimental physics, and the discipline of it is why the result stands. He identified and pursued three nuisances [44].

Magnetism. The lead spheres might carry iron, and the balls might respond to it. He tested by rotating the source spheres about their own axes between runs, which reverses any magnetic moment they carry relative to the beam while leaving the gravitational configuration untouched, and found no change.

Air currents. The room was closed, kept dark, and the observer stayed outside it: the source spheres were swung by a pulley and cord operated from outside, and the beam was read through telescopes sighted at the case through small apertures, the scales lit by lamps. Nothing that could warm or stir the air was allowed inside with the instrument.

Thermal convection. This proved to be the real one, and the manner of its discovery is the model. Cavendish found an attraction that varied between runs in a way the geometry could not explain, and supposed—while thinking it improbable—that a temperature difference between the massive lead spheres and the case was driving a slow current of air past the balls. He then did the decisive thing: rather than argue about it, he made the suspected effect large. He placed lamps beneath the source spheres to heat them deliberately, put thermometers into the spheres themselves and read them through the telescope by light reflected from a convex mirror, and watched the beam move. The effect was there, and it was large. He then established the cause of the ordinary, unheated case: overnight the massive spheres cooled more slowly than the thin case, so a residual temperature difference—and a convection current—was present in every run. Diagnosing a systematic by amplifying it deliberately and then measuring its coefficient is the technique the whole subject still uses.

The modern list is longer and the terms are smaller, but the shape is the same: seismic and acoustic background, the local gravitational background of moving people and vehicles, electrostatic patch potentials between the beam and the case, residual magnetic susceptibility, the metrology of the source masses and their internal density inhomogeneity, and the elastic behaviour of the suspension [2286]. The last of these is the one that has resisted.

Remark 33.10 (Fibre anelasticity, and why it biases one method and not the other). A real fibre is not perfectly elastic: it dissipates, and dissipation and dispersion are not independent. The torsional response $\theta(\omega) = \hat{\chi}(\omega)\tau(\omega)$ is causal, so Theorem 17.78 ties the frequency dependence of the real part of the stiffness to the loss, and a fibre with a frequency-independent loss angle $1/Q$ necessarily has a stiffness that creeps logarithmically with frequency. Kuroda's observation [1901] is that this converts the apparently exact cancellation of κ in Equation (33.12) into an upward fractional bias of the time-of-swing result,

$$\frac{\Delta G}{G} \simeq \frac{1}{\pi Q}\,, \tag{33.14}$$

with Q the quality factor of the torsion mode. For a tungsten fibre with $Q \approx 2 \times 10^3$ that is about 160 parts per million — the same order as the spread among the modern determinations of Section 33.7.4, and an order of magnitude larger than the uncertainty several of them quote. The static-deflection method of Proposition 33.6 does not share the bias, because there κ is not required to be the same at two frequencies; and the angular-acceleration method of Section 33.7.2 avoids the fibre altogether by keeping it untwisted. That three methods with different exposures to one effect still disagree is why the anelastic explanation, though the leading candidate, is not a resolution.

Rests on Theorem 17.78 and Equation (33.12).

33.3 Apparatus

The instrument is a horizontal beam suspended at its midpoint by a fine wire, carrying a small lead ball at each end; two large lead spheres on a swivelling frame can be brought up to the balls from either side without the operator entering the room. The beam, the balls and the scales are enclosed in a narrow mahogany case, and the case stands in a closed shed, read from outside through telescopes. Every dimension is collected in Table 33.1, converted from the inches and pounds of the original per Remark 33.1.

Three features of the design carry the experiment and are worth naming separately. The case is narrow, so that the air inside it has little room to circulate. The source spheres never enter the case: they act through its wall, so that the massive, slowly cooling bodies are thermally outside the volume the balls live in — and even so the residual coupling was Cavendish's dominant systematic. And the readout is remote: telescopes with verniers, sighted from outside the closed room on ivory scales fixed to the ends of the beam and illuminated by lamps, so that the observer's own body heat never reaches the instrument.

33.4 Procedure

A run proceeds as follows. The source spheres are set on one side; the beam, disturbed by the change, begins a slow torsional oscillation about its new equilibrium. Rather than wait for it to come to rest — which at Q of order 10^3 and a period of a quarter of an hour would take days — the successive turning points are recorded through the telescope and the equilibrium is inferred from them. The spheres are then swung to the other side and the procedure repeated, and the difference of the two inferred equilibria is $2s$ (Remark 33.8). Separately, with the spheres removed, the free oscillation is timed over many periods to fix T and hence κ.

Table 33.1: The Cavendish balance, in SI. All values are converted from the imperial figures of [44]; the conversions (25.4 mm to the inch, 0.453 592 37 kg to the pound) are exact, so the number of significant figures shown is that of the original. The two suspension wires are those used before and after the change of wire recorded in Section 33.4; the derived torsional stiffnesses follow from Equation (33.6) with the moment of inertia of the two balls alone, and are quoted to one figure because the beam's own contribution is not reproduced here. The last two rows are the ivory scale's division and the resolution with which it could be read, through a telescope and a vernier, from outside the closed room.

part	quantity	value
beam	length	1.83 m
	half-length d	0.914 m
suspended balls	diameter	5.08×10^{-2} m
	mass m (each)	0.730 kg
source spheres	diameter	0.305 m
	mass M (each)	157.9 kg
	centre separation r	0.2248 m
suspension (i)	torsional period T	9.0×10^{2} s
	stiffness κ	6×10^{-5} N m/rad
suspension (ii)	torsional period T	4.5×10^{2} s
	stiffness κ	2.4×10^{-4} N m/rad
scale	division	1.27×10^{-3} m
	reading resolution	2.5×10^{-4} m

Proposition 33.11 (The equilibrium from three turning points)**.** *Let a lightly damped oscillation about an unknown equilibrium θ_{eq} have successive turning points $\theta_1, \theta_2, \theta_3$, let A be the displacement of the first of them from equilibrium, and let $q = \mathrm{e}^{-\Gamma T/2} \in (0,1)$ be the amplitude ratio between successive extrema. Then*

$$\frac{\theta_1 + 2\theta_2 + \theta_3}{4} = \theta_{eq} + \frac{A}{4}(1-q)^2 , \qquad \frac{\theta_1 + \theta_2}{2} = \theta_{eq} + \frac{A}{2}(1-q) . \tag{33.15}$$

The three-point estimator is therefore in error only at second order in the decay per half period, the two-point estimator at first order.

Rests on Equation (33.8)*.*

Proof. *Derives Proposition 33.11.*

The free solution of Equation (33.8) is $\theta(t) = \theta_{\mathrm{eq}} + A\mathrm{e}^{-\Gamma(t-t_1)}\cos\omega_d(t-t_1)$, whose turning points are separated by half a period, so that with $q = \mathrm{e}^{-\Gamma T/2}$

$$\theta_n = \theta_{\mathrm{eq}} + (-1)^{n-1} A q^{\,n-1} , \qquad n = 1, 2, 3 . \tag{33.16}$$

Substituting the three values into the two estimators,

$$\frac{\theta_1 + 2\theta_2 + \theta_3}{4} = \theta_{\mathrm{eq}} + \frac{A(1 - 2q + q^2)}{4} , \qquad \frac{\theta_1 + \theta_2}{2} = \theta_{\mathrm{eq}} + \frac{A(1-q)}{2} ,$$

and $1 - 2q + q^2 = (1-q)^2$, which is Equation (33.15). To size the gain, note that $\Gamma T/2 = \pi\Gamma/\omega_0 = \pi/(2Q)$ to leading order, so $1 - q \approx \pi/(2Q)$ and the ratio of the two errors is

$$\frac{(A/4)(1-q)^2}{(A/2)(1-q)} = \frac{1-q}{2} \approx \frac{\pi}{4Q} .$$

For $Q \approx 10^3$ that is a suppression by three orders of magnitude, which is why three readings taken over one and a half slow periods do the work of waiting days for the beam to come to rest. □

This is what Cavendish does, in words rather than symbols: he takes the mean of the extremes with the middle one counted twice. It is the reason the experiment was feasible at all with an instrument whose oscillation outlasts the observer's patience by two orders of magnitude.

The published series ran from August 1797 to May 1798 and comprises seventeen experiments yielding twenty-nine determinations of the mean density [44]. The first three experiments, six determinations made in August 1797, used the original suspension wire; Cavendish judged it insufficiently stiff — the beam was too easily disturbed — and replaced it with a wire of about four times the torsional stiffness, on which the remaining fourteen experiments and twenty-three determinations were made. The change is visible in the data as a halving of the period and a roughly fourfold reduction of the deflection — exactly the covariation Equation (33.6) and Equation (33.9) demand of a stiffness raised by a factor of four, and the reason two periods and two deflections appear in Table 33.1. Cavendish tabulates the two groups separately and does not combine them into a weighted mean; the grand average quoted in Section 33.5 is an unweighted one over all twenty-nine determinations.

33.5 Observations and data

Table 33.2: What the Cavendish experiment measured, in SI. The upper block is the raw instrumental signal; the middle block is the result as Cavendish published it and as corrected. The published mean of 5.480 times water contains an arithmetic slip in the averaging, found by Francis Baily in 1821 and recorded in Poynting's review [261]; Cavendish's own twenty-nine numbers average to 5.448. The implied G is not a quantity Cavendish computed — see Section 33.6 — and is obtained here by inverting Equation (33.17) with the modern g and R. The lower block is the modern comparison; the modern mean density is $M/\frac{4}{3}\pi R^3$ with $M = GM_\oplus/G$.

	quantity	value
signal	deflection s, wire (i)	$4.1 \times 10^{-3}\,\mathrm{m}$
	deflection s, wire (ii)	$7.6 \times 10^{-4}\,\mathrm{m}$
	gravitational force per ball	$1.5 \times 10^{-7}\,\mathrm{N}$
	gravitational torque on the beam	$2.8 \times 10^{-7}\,\mathrm{N\,m}$
	determinations (experiments)	29 (17)
result	mean density, as published	$(5.480 \pm 0.038) \times 10^3\,\mathrm{kg/m^3}$
	mean density, corrected	$(5.448 \pm 0.033) \times 10^3\,\mathrm{kg/m^3}$
	relative uncertainty	6.1×10^{-3}
	implied G	$(6.745 \pm 0.041) \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$
modern	mean density of the Earth	$5.5134 \times 10^3\,\mathrm{kg/m^3}$
	G (CODATA 2022)	$(6.674\,30 \pm 0.000\,15) \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$
	relative uncertainty	2.2×10^{-5}
	Cavendish's G is high by	1.1 %

Two things in Table 33.2 deserve comment before the physics is drawn out of them. The first is the arithmetic slip. The number 5.480 printed by Cavendish is not what his own twenty-nine determinations average to; the correct mean is 5.448, as Baily found in 1821

[261]. The book quotes both, because a reader who checks the primary source must be able to see why it disagrees with the modern secondary literature, and because the correction moves the answer *away* from the modern value — his error happened to be in the helpful direction, which is worth knowing when reading claims of eighteenth-century accuracy.

The second is the quoted uncertainty. The ± 0.033 is the standard error of the mean of twenty-nine determinations scattering by a few percent, that is a statistical uncertainty only (Theorem 2.7). It contains no allowance for the systematic effects of Section 33.2.3, and the actual departure from the modern value, 1.1 %, is close to twice the quoted error — which is the honest reading: Cavendish's statistical uncertainty was smaller than his systematic one, as it is for very nearly every measurement of G since.

Phenomenon 33.12 (The Earth is much denser than its crust)**.** The mean density of the Earth is about five and a half times that of water, roughly $5.5 \times 10^3\,\mathrm{kg/m^3}$, whereas the rocks accessible at the surface run between about $2.5 \times 10^3\,\mathrm{kg/m^3}$ and $3.0 \times 10^3\,\mathrm{kg/m^3}$ [44]. The interior must therefore consist of something substantially denser than anything the crust exposes. This was the first quantitative evidence that the planet is differentiated, and the beginning of the line of inference that ends in the seismological picture of an iron core; it was also the result the torsion-balance experiment was performed to obtain, the gravitational constant being a later restatement of the same measurement.

Rests on Equation (33.1) *and Proposition 33.6.*

Derivation. *Derives Phenomenon 33.12.*

Treat the Earth as a sphere of radius R and mass M. A body of mass m at the surface has weight $GMm/R^2 = mg$, so that $M = gR^2/G$ and the mean density is

$$\bar{\rho} = \frac{M}{\frac{4}{3}\pi R^3} = \frac{3g}{4\pi G R}\,. \tag{33.17}$$

From Equation (33.1) *(equating the weight mg to the attraction of a sphere of mass M at radius R and dividing by the volume).*

Every quantity on the right is measured without any assumption about the interior: g by dropping bodies (Chapter 20), R by geodesy, and G — the only difficult one — by this experiment. With $g = 9.806\,65\,\mathrm{m/s^2}$ [2333], $R = 6.371 \times 10^6\,\mathrm{m}$ and $G = 6.674\,30 \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$ [2392],

$$\bar{\rho} = 5.506 \times 10^3\,\mathrm{kg/m^3}\,, \qquad M = \frac{gR^2}{G} = 5.964 \times 10^{24}\,\mathrm{kg}\,.$$

Both are low by 0.14 % against the accepted $5.5134 \times 10^3\,\mathrm{kg/m^3}$ and $5.9722 \times 10^{24}\,\mathrm{kg}$, and the discrepancy is not an error: g_0 is a conventional surface value that already includes the centrifugal reduction of the Earth's rotation, and the planet is an oblate spheroid rather than the sphere assumed here. At the 1 % level at which this chapter works, the spherical idealisation is entirely adequate; at the level of the modern numbers it is not, which is why the modern mean density is computed from $GM_\oplus$ and R rather than from g.

Read in the other direction this is exactly the eighteenth-century programme of "weighing the Earth": a measurement carried out entirely between laboratory masses delivers the mass of the planet, because G is the one factor that no astronomical observation can supply (Proposition 33.2). □

33.6 Interpretation

Cavendish never wrote a value for G, and it would have been an odd thing for him to do: the constant had no name, no symbol and no recognised status as a quantity to be measured.

What he reported was a density ratio. Inverting Equation (33.17) with the modern g and R turns his corrected 5.448 into

$$G = \frac{3g}{4\pi\bar{\rho}R} = (6.745 \pm 0.041) \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}\,, \tag{33.18}$$

which stands 1.1 % above the CODATA 2022 value [2431]. The reframing—from a density of the Earth to a constant of nature—was made a century later, in Boys's paper, which is titled "On the Newtonian constant of gravitation" and treats the Earth's density as a by-product rather than the object [262], and in Poynting's synthesis of the whole line of work [261].

The evidential point for this part of the book is the one Chapter 27 needs and cannot supply: that the law inferred from the heavens holds unchanged between bodies on a bench. It is worth being precise about what is and is not established. The experiment fixes the constant, for lead, at a separation of about 0.2 m. It does *not*, on its own, test the inverse square at laboratory distances to any precision — for that one varies r, which Cavendish did not systematically do, and which the modern short-range experiments of Chapter 42 pursue; nor does it test composition dependence, which had to wait for Heyl's gold, platinum and glass [657] and, far more sharply, for the Eötvös-type balances of the same chapter. What the combination of the bench measurement with the celestial one does test, and sharply, is that a single constant serves both.

Phenomenon 33.13 (One constant serves the bench and the sky)**.** The constant measured between lead spheres in a closed room is the same constant that governs the orbits of the Moon and the planets [44]. No scaling factor, no dependence on the material of the attracting bodies, and no departure from the inverse square is needed to pass from a separation of 0.2 m to one of 3.8×10^8 m — nine orders of magnitude in distance, and more than twenty in mass. Universality of this kind is not implied by the celestial evidence alone, which fixes only the products GM and never G and M separately; it is the laboratory measurement that separates the two and thereby makes the statement testable at all.

Rests on Proposition 33.2 and Phenomenon 33.12.

Derivation. *Derives Phenomenon 33.13.*

Take G from the torsion balance and the Earth's mass $M = 5.964 \times 10^{24}$ kg from Equation (33.17), both obtained without any reference to the Moon. Newton's law together with the circular-orbit condition of Chapter 27 then predicts the lunar month with nothing further adjusted: Equation (33.2) for a circular orbit of radius r about a centre of total mass $M + m$ reads

$$T = 2\pi\sqrt{\frac{r^3}{G\,(M+m)}}\,. \tag{33.19}$$

With $r = 3.844 \times 10^8$ m, $M = 5.964 \times 10^{24}$ kg and the Moon's $m = 7.346 \times 10^{22}$ kg this gives $T = 2.359 \times 10^6$ s, that is 27.30 days, against an observed sidereal month of 27.3217 days. The agreement to 0.08 % — the residual being the eccentricity of the orbit and the solar perturbation, not the constant — is the check that the laboratory G and the celestial one are a single number. Note what the check consumes: the laboratory G enters twice, once in converting g and R into M and once in Equation (33.19) itself, and the two occurrences do not cancel, since the first carries G^{-1} and the second $G^{-1/2}$. The test is therefore genuine and not circular.

That same constant is carried unchanged into the field equations of Chapter 44, where it appears in the coupling $8\pi G/c^4$; the far sharper tests of the universality of gravitational free fall are the subject of Chapter 42. □

33.7 Modern determinations and their discrepancies

33.7.1 Classic redeterminations

Boys reworked the experiment in 1895 with two changes that between them define the modern instrument [262]. The second was the suspension: a drawn quartz fibre in place of drawn metal, which is finer, stronger in tension, and far closer to perfectly elastic — the material answer to Remark 33.5 and, in part, to Remark 33.10. The first was to make the whole instrument small: his beam was a few centimetres long against Cavendish's 1.83 m. He obtained $G \approx 6.658 \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$, and, as the title of his paper announces, he reported it as a constant of nature.

Remark 33.14 (Why shrinking the balance costs no sensitivity). The reason a small instrument does as well is a scaling argument, and it is the one that made Boys's design possible. Take two geometrically similar balances, the second smaller by a factor λ, and load each suspension to the same fraction of the breaking stress of its material. The suspended weight scales as λ^3, so the fibre's cross-sectional area does too and its radius as $\lambda^{3/2}$; by Equation (33.7) the torsional stiffness then scales as $(\lambda^{3/2})^4/\lambda = \lambda^5$. The moment of inertia $2md^2$ scales as $\lambda^3\lambda^2 = \lambda^5$ as well, so by Equation (33.6) the period is *unchanged*; and the gravitational torque Equation (33.10), with $M, m \propto \lambda^3$ and $d, r \propto \lambda$, scales as λ^5 likewise. Hence

$$\theta = \frac{\tau_g}{\kappa} \propto \frac{\lambda^5}{\lambda^5} = \text{const}\,, \qquad T \propto \text{const}\,. \tag{33.20}$$

The *angular* deflection — which is what a mirror and a distant scale read, and which does not care how long the beam is — is independent of the size of the apparatus. Shrinking it therefore costs nothing in signal and buys a great deal in disturbance: less air inside the case to convect, a smaller temperature difference across it, and a smaller and more easily controlled room around it. The strength of quartz is what makes the premise of the argument, a fibre loaded to a fixed fraction of breaking stress, achievable at small scale.

Rests on Equations (33.7) *and* (33.10).

Poynting had meanwhile attacked the same quantity with a common balance rather than a torsion fibre, weighing a mass against the attraction of a large lead sphere placed beneath one pan and then the other; his Adams Prize essay is also where the correction to Cavendish's arithmetic is recorded [261].

Heyl's series at the Bureau of Standards in the 1920s introduced the time-of-swing method of Proposition 33.9 in its modern form and gave $G = (6.670 \pm 0.005) \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$, a relative uncertainty of 7.5×10^{-4} [657]. He also did something nobody had: he ran the measurement with suspended masses of gold, of platinum and of optical glass, an explicit search for a dependence of gravitation on the material of the attracted body. Heyl and Chrzanowski's continuation of the work supplied the value on which the first CODATA recommendation rested [2286].

The reference value for a generation was Luther and Towler's time-of-swing determination with tungsten source masses, $G = (6.6726 \pm 0.0005) \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$, a relative uncertainty of 7.5×10^{-5} [1695]. It dominated the recommended value until the 1990s, and it is a time-of-swing result, which is why Remark 33.10 landed as hard as it did when Kuroda published it in 1995 [1901].

33.7.2 Modern torsion-balance measurements

Gundlach and Merkowitz answered the anelasticity problem by removing the fibre from the measurement chain [1991]. Their balance sits on a turntable whose angular acceleration is servoed so as to keep the fibre's twist at zero: the feedback signal is then the angular acceleration required to follow the gravitational torque, and κ never enters, since the fibre

is never twisted. Neither Equation (33.9) nor Equation (33.12) is used — a third method, whose systematics are dominated by the metrology of the source masses and by the turntable rather than by the suspension. They reported $G = (6.674\,215 \pm 0.000\,092) \times 10^{-11}\ \mathrm{m^3/(kg\,s^2)}$, 14 parts per million, then the most precise value ever obtained.

The BIPM group built a balance that could be operated in two modes on the same apparatus and the same masses: a servo mode, in which an electrostatic torque holds the beam and is itself the measurement, and a Cavendish mode, in which the static deflection is read as in Proposition 33.6. The two agreed with each other and gave $G = (6.675\,54 \pm 0.000\,16) \times 10^{-11}\ \mathrm{m^3/(kg\,s^2)}$, 24 parts per million, after the correction published as an erratum [2222]. This value and Gundlach and Merkowitz's are the canonical irreconcilable pair: they differ by 199 parts per million with a combined uncertainty of 28, that is by more than seven standard deviations.

The HUST group then did what the field most needed, which is not another precise number but two of them from one laboratory: time-of-swing and angular-acceleration feedback, run on the same premises with the same source masses and by the same people, each reaching about 12 parts per million [2314]. The two results are $(6.674\,184 \pm 0.000\,078) \times 10^{-11}\ \mathrm{m^3/(kg\,s^2)}$ and $(6.674\,484 \pm 0.000\,078) \times 10^{-11}\ \mathrm{m^3/(kg\,s^2)}$: they differ from each other by 45 parts per million, about 2.7 standard deviations of the difference. Two methods, one laboratory, and still a disagreement — which is a stronger statement about the state of the subject than any single number in it.

33.7.3 Beyond torsion

Rosi and collaborators measured G with no fibre, no beam and no macroscopic proof mass at all: the test bodies are laser-cooled rubidium atoms in free fall, and the observable is the phase of a matter-wave interferometer, shifted by the gravitational field of tungsten source masses moved between two positions [2237]. The result, $G = (6.671\,91 \pm 0.000\,99) \times 10^{-11}\ \mathrm{m^3/(kg\,s^2)}$ at 150 parts per million, is an order of magnitude less precise than the best torsion values — and it is nevertheless one of the most useful numbers in the set, because its systematics have almost nothing in common with theirs. It lies 544 parts per million below the BIPM value, more than three standard deviations of the difference.

Other routes have been pursued: beam balances weighing a test mass against a movable source, free-fall and pendulum gradiometry, and Fabry–Pérot resonators, all surveyed with their uncertainty budgets in [2286]. The lesson the survey draws is the one this section exists to make. The field does not need another measurement at 10 parts per million that shares a systematic with the measurements already there; it needs measurements whose dominant systematic is different, so that agreement means something. That is why a 150-part-per-million atom-interferometric value carries weight disproportionate to its error bar.

33.7.4 The discrepancy, honestly stated

Example 33.15 (The six values are mutually inconsistent, and by how much)**.** Combine the six measurements of Table 33.3 as independent determinations of one quantity. The minimum-variance combination is the inverse-variance weighted mean Equation (2.4), which gives

$$\hat{G} = (6.674\,382 \pm 0.000\,045) \times 10^{-11}\ \mathrm{m^3/(kg\,s^2)}\,, \tag{33.21}$$

a relative uncertainty of 6.8 parts per million. That number should not be believed, and the reason is visible in the same arithmetic. The residuals about it, in units of each measurement's own uncertainty and in the order of the table, are -3.56, -1.82, $+7.24$,

Table 33.3: Six determinations of G, as published, in units of $\mathrm{m^3/(kg\,s^2)}$. "ToS" is the time-of-swing method of Proposition 33.9, "Cav" the static deflection of Proposition 33.6, "AAF" the angular-acceleration-feedback method, "AI" atom interferometry. The final row is not a measurement but the CODATA 2022 least-squares adjustment over sixteen input values, whose uncertainty is enlarged by an expansion factor of 3.9 to cover the observed scatter [2431]. Relative uncertainties are quoted in parts per million.

year	laboratory and method	$G \times 10^{11}$	$u_r/10^{-6}$	source
1982	NBS, ToS	6.6726(5)	75	[1695]
2000	U. Washington, AAF	6.674215(92)	14	[1991]
2013	BIPM, servo + Cav	6.67554(16)	24	[2222]
2014	LENS, AI	6.67191(99)	150	[2237]
2018	HUST, ToS	6.674184(78)	11.6	[2314]
2018	HUST, AAF	6.674484(78)	11.6	[2314]
2022	CODATA adjustment	6.67430(15)	22	[2431]

-2.50, -2.54 and $+1.31$, and their sum of squares is

$$\chi^2 = \sum_i \frac{\left(G_i - \hat{G}\right)^2}{u_i^2} = 83\,, \tag{33.22}$$

on $6-1=5$ degrees of freedom. That the count is $n-1$ and not n is Lemma 11.83 with the projector onto the complement of the weight vector; the expected value of such a χ^2 is 5 with a standard deviation of $\sqrt{10}$ (Proposition 11.81), so the observed 83 corresponds to a p-value (Definition 11.75) of order 10^{-16}. The data are not a set of independent measurements of one quantity with the stated errors.

The Birge ratio $\sqrt{\chi^2/(n-1)} = 4.1$ is the factor by which every uncertainty would have to be multiplied to make the set self-consistent. CODATA, working over sixteen inputs rather than six, arrives at an expansion factor of 3.9 [2431]. The agreement between the two numbers is a coincidence of arithmetic rather than a reproduction of the adjustment — different input sets, and CODATA chooses its factor to bring all residuals below two rather than to normalise χ^2 — but it shows where the recommended uncertainty comes from. It is not the uncertainty of the measurements. It is the width of their disagreement.

Rests on Equation (2.4), Lemma 11.83, and Proposition 11.81.

Phenomenon 33.16 (The least well determined of the constants)**.** The Newtonian constant of gravitation is known far less precisely than any other fundamental constant. The CODATA recommended value is $G = 6.674\,30 \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$ with a standard uncertainty of $1.5 \times 10^{-15}\,\mathrm{m^3/(kg\,s^2)}$, a relative uncertainty of 2.2×10^{-5} [2392, 2431]; the same adjustment fixes the electron mass and the Rydberg constant to parts in 10^{10} and better. The reason is not that the individual experiments are imprecise. Several claim uncertainties of tens of parts per million, and yet their central values scatter over several hundred parts per million — many times their own quoted errors — so that they cannot all be correct. The recommended uncertainty above is therefore not the uncertainty that any experiment reports: it is an inflated one, enlarged by the adjustment precisely because the inputs are mutually inconsistent [2431].

Rests on Example 33.15 and Equation (33.9).

Derivation. *Derives Phenomenon 33.16.*

Every clause of the statement is arithmetic on Table 33.3, and it is worth doing because the conclusion is often reported as an opinion about the field rather than as a measured property of its data.

Take the scatter first. The extreme values in the table are the BIPM determination at $6.675\,54 \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$ and the LENS atom-interferometric one at $6.671\,91 \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$, whose difference is $3.63 \times 10^{-14}\,\mathrm{m^3/(kg\,s^2)}$, or 544 parts per million of either. Discard the atom interferometer, the one entry whose own uncertainty is large, and the four torsion values still span 203 parts per million, from the same BIPM value down to the HUST time-of-swing one. Against that, the two most precise entries each claim 11.6 parts per million. The spread of the set exceeds the precision claimed within it by a factor between seventeen and forty-seven, which is the first clause of the statement: the experiments are not imprecise, they disagree.

That they disagree beyond their stated errors is Example 33.15, which is quoted here and not repeated: the χ^2 of Equation (33.22) is 83 on five degrees of freedom, whose p-value is of order 10^{-16}. No hypothesis under which the six are independent measurements of one quantity with the stated uncertainties survives that number.

The third clause — that the recommended uncertainty is an inflated one and not the uncertainty of any measurement — follows from a property of the weighted mean, without any knowledge of how CODATA performs its adjustment. By Equation (2.4) the inverse-variance combination of independent measurements has variance $\hat{u}^2 = \left(\sum_i u_i^{-2}\right)^{-1}$, so $\hat{u} \le \min_i u_i$: combining consistent data can only sharpen the answer, never blunt it. Here $\min_i u_i$ is 11.6 parts per million and the six inputs give $\hat{u} = 6.8$ parts per million, as Equation (33.21) records. The CODATA recommended value carries 22 parts per million. That is larger than the best single input, so the recommended uncertainty is provably not the uncertainty of a combination of the data as published; it is that combination enlarged, by a factor of 3.2 on the six values here and by the quoted 3.9 on the sixteen CODATA uses [2431]. What the recommended figure measures is the width of the disagreement.

The remaining clause, that no other fundamental constant is known so poorly, is not derived here: it is read off the same adjustment, which fixes the electron mass and the Rydberg constant to parts in 10^{10} and better [2392, 2431]. Nothing in this chapter establishes it independently.

What the derivation does *not* deliver is the cause. Every step above uses only the published values and their published uncertainties; none of it identifies which measurement is displaced, or by what. That question is open, and Remark 33.17 states how open. □

Remark 33.17 (The discrepancy is not resolved, and the leading candidate does not fit). No accepted account of the scatter exists, and none is constructed here. Resolving it would mean establishing which systematic effect displaces which measurements — frequency-dependent anelasticity of the suspension fibre and the metrology of the source masses are the leading candidates — and a reader is entitled to know how bad the situation is. The first thing to say is that the leading candidate does not fit. If fibre anelasticity (Remark 33.10) were the whole story, the time-of-swing values would sit systematically high and the angular-acceleration values low; instead the two HUST determinations, one of each kind, differ by 45 parts per million in the *opposite* sense to the naive prediction of Equation (33.14), and the BIPM value, which is not a time-of-swing measurement at all, is the highest in the set. The second thing is that no proposal of new physics is on the table here, and none should be: a discrepancy among six laboratory measurements of a laboratory quantity is overwhelmingly likely to be an unidentified systematic in one or more of them, and the honest statement is that nobody yet knows which. This remark is not a placeholder for a derivation the treatise owes. Phenomenon 33.16 asserts that the constant is poorly known, that the inputs are inconsistent and that the recommended uncertainty is inflated, and all three are derived above; the cause of the inconsistency is not asserted anywhere in this book, and so is not

owed.

33.8 Primary references

The experiment itself: [44], which credits Michell for the apparatus and contains the systematic investigations of Section 33.2.3. The precursor: [33] for the Schiehallion observations and [35] for their reduction. The torsion law and the independent invention of the instrument: [36, 38].

The reframing as a measurement of a constant, and the correction to Cavendish's arithmetic: [261, 262]. Modern determinations: [657, 1695, 1991, 2222, 2237, 2314], reviewed with their uncertainty budgets in [2286]. The anelasticity systematic: [1901]. Recommended values and the expansion factor: [2392, 2431]; uncertainty conventions: [1945, 2129]; the conventional standard acceleration g_0: [2333].

Chapter 34

Experiment: The Pendulum

Chapter 28 asserts that a pendulum of length ℓ swinging through a small arc has period $T = 2\pi\sqrt{\ell/g}$, the same whether the arc is wide or narrow and whatever the mass and material of the bob; Chapter 29 asserts that in the frame rotating with the Earth a Coriolis term must turn the plane of that swing at the rate $\Omega \sin\lambda$. Both are predictions, and neither chapter shows either of them being measured. This chapter reports the measurements. It also reports what happens when the small arc is *not* small, which is the first place in this treatise where a linearization is tested rather than assumed, and what happens when the restoring torque comes from a twisted wire instead of from gravity, which is the instrument of Chapter 33.

The pendulum deserves a chapter of its own for a reason beyond the two predictions it settles: it is the first precision instrument in physics. Everything the measurement chapter Chapter 2 requires of an instrument — a reading that is a number, a calibration chain that terminates in base units, a stated uncertainty, and a systematic error budget written down and argued over — appears here first and appears complete. A pendulum is a length and a time and nothing else; that is what makes it an *absolute* instrument, in the technical sense that its scale is fixed by the metre and the second without reference to any other apparatus of the same kind. In four guises it delivered the first precision frequency standard (Section 34.1), the transducer that converts a force too weak to weigh into a visible angle (Section 34.2), the first absolute determination of the acceleration of free fall (Section 34.3), and the first laboratory proof, consulting nothing outside the room, that the Earth turns (Section 34.5).

34.1 Galileo and Huygens: the isochronous pendulum (1602–1673)

Tests Phenomenon 28.9 and Equation (28.7).

Assuming Equation (28.8).

The claim under test is the one stated in Chapter 28 as the isochronism of the small-amplitude pendulum: the period does not depend on the amplitude, nor on the bob. It is an approximate claim, and the interest of the seventeenth-century work is that both halves of it were established — Galileo found the approximation, Huygens found its limits and then removed them exactly, by changing the path from a circle to a cycloid.

34.1.1 Apparatus

Galileo's instrument is a bob on a cord of measured length hung from a fixed point, and, for the comparative observations, two such pendulums of equal length side by side carrying bobs of very different density — cork and lead — so that a dependence on the material of

the bob would show as a growing difference of phase [4]. There is no clock: the counting of swings *is* the clock, and where an independent time is needed it comes from a water vessel weighed before and after, the technique Galileo used for the inclined plane.

Huygens' instrument is a pendulum clock: a bob on a flexible suspension, maintained against friction by a weight-driven escapement, with a dial that accumulates the count. Its distinguishing feature is a pair of rigid metal cheeks flanking the point of suspension, bent to arcs of a cycloid, against which the suspension wraps as the bob swings out [11]. The cheeks are the experiment; the rest is a clock.

34.1.2 Procedure

Release two equal pendulums together from unequal arcs and count the swings until they can be seen to separate: an amplitude dependence of the period appears as a slow accumulation of phase, so that a fractional period difference of ϵ becomes a visible half-swing after about $1/2\epsilon$ swings. The same protocol run with cork and lead bobs tests the independence of the bob. To test the square-root law, compare pendulums whose lengths stand in the ratio $4:1$ and count the swings of each in the same interval.

With the clock, the procedure is to vary the driving weight, and with it the amplitude at which the escapement settles, and read the change in the rate; then to fit the cycloidal cheeks and repeat. The signature sought is not a small rate but a rate that stops moving when the amplitude is changed.

34.1.3 Observations and data

Galileo reports that pendulums of equal length keep step whether the arc is wide or narrow, and that a cork bob and a lead bob of the same suspension length stay together for a great many swings even though the cork is stopped by the air much sooner; and that the period varies as the square root of the length [4]. Huygens reports the residual he was looking for: the circular pendulum runs *slow* at large amplitude, so that a clock's rate depends on how hard it is driven, and the cycloidal constraint removes that dependence entirely [11].

Neither man quotes a number in a unit and neither quotes an uncertainty; both quote a comparison. The quantitative content of the two reports is therefore recovered by asking what a count of swings can resolve, and setting that against the exact period of the circular pendulum, which is the quantity Galileo's claim asserts to be constant and Huygens' measurement shows is not.

Remark 34.1 (The exact period is an elliptic integral, and this treatise does not yet define one). Quadrature of Equation (28.8) at fixed energy gives the period of a swing of angular amplitude θ_0 as

$$T(\theta_0) = \frac{4}{\omega_0}\int_0^{\pi/2} \frac{\mathrm{d}\varphi}{\sqrt{1-\sin^2(\theta_0/2)\sin^2\varphi}}, \qquad \omega_0 = \sqrt{\frac{g}{\ell}}, \tag{34.1}$$

after the substitution $\sin(\theta/2) = \sin(\theta_0/2)\sin\varphi$. The integral is the complete elliptic integral of the first kind of modulus $k = \sin(\theta_0/2)$. That function is not among the results of Part II, so nothing below leans on its properties: the entries of Table 34.1 are quoted as numerical values of Equation (34.1), and every *algebraic* statement made about the amplitude dependence uses instead the small-amplitude series $T/T_0 = 1 + \theta_0^2/16 + 11\theta_0^4/3072 + \cdots$ obtained by the perturbation method of Chapter 28, which reproduces the table to the digits shown up to about 60° and fails by 1.6×10^{-4} at 90°.

Galileo's comparison, reduced. He hangs the two bobs on threads "four or five braccia" long [4]; a Florentine braccio is close to $0.583\,\mathrm{m}$, so the suspension is between $2.33\,\mathrm{m}$ and $2.92\,\mathrm{m}$ and, by Equation (28.7), the period is between $3.06\,\mathrm{s}$ and $3.43\,\mathrm{s}$. The lead bob is

Table 34.1: The amplitude dependence of the circular pendulum: what Galileo's comparison bounds, what Huygens measured, and what the cycloidal cheeks remove. $T_0 = 2\pi\sqrt{\ell/g}$ is the infinitesimal-amplitude period of Equation (28.7) and θ_0 the angular amplitude; the ratio is the numerical value of Equation (34.1), exact to the six figures shown. $N_{1/2}$ is the number of swings after which a pendulum of amplitude θ_0 has fallen half a period behind an identical one of vanishing amplitude, $N_{1/2} = \frac{1}{2}T_0/\left(T - T_0\right)$: that is the point at which the two are seen to be in antiphase, and it is the resolution of a bare count of swings. The last column converts the same fractional excess into the daily rate error of a clock whose escapement holds that amplitude. Nothing in the table depends on ℓ, on the mass, or on the material of the bob.

θ_0	T/T_0	$\left(T - T_0\right)/T_0$	$N_{1/2}$ (swings)	clock error (s per day)
1°	1.000 019	1.9×10^{-5}	2.6×10^{4}	1.6
2°	1.000 076	7.6×10^{-5}	6.6×10^{3}	6.6
5°	1.000 476	4.8×10^{-4}	1.1×10^{3}	41
10°	1.001 907	1.9×10^{-3}	262	165
20°	1.007 669	7.7×10^{-3}	65	663
30°	1.017 409	1.7×10^{-2}	29	1.5×10^{3}
45°	1.039 973	4.0×10^{-2}	13	3.5×10^{3}
60°	1.073 182	7.3×10^{-2}	6.8	6.3×10^{3}
90°	1.180 341	1.8×10^{-1}	2.8	1.6×10^{4}

stated to be more than a hundred times the mass of the cork one. What is counted is swings, and two pendulums can be seen to be in step or not to within perhaps a quarter of a swing; so agreement maintained over N swings bounds the fractional period difference by

$$\left|\frac{\Delta T}{T}\right| < \frac{1}{4N}\,, \tag{34.2}$$

which for the fifty to a hundred swings Galileo reports is 2.5×10^{-3}. That is the entire quantitative content of the observation, and Table 34.1 says exactly what it licenses: the mass independence is established at the 2.5×10^{-3} level, and the amplitude independence is established at that level *only for arcs of order* 10° *or smaller*. The bound is not idle. Galileo states isochronism for wide arcs as well as narrow ones, and there it is simply false: between 50° and 5° the exact period differs by 4.9 %, so the two pendulums would stand in antiphase after ten swings — half a minute, on a suspension of four braccia.

Huygens' measurement, reduced. The signature he reports is that a circular clock's rate follows its driving weight, and stops doing so once the cheeks are fitted [11]. In numbers: an escapement that lets the amplitude wander between 4° and 6° — a small drift, and less than the change produced by a week's accumulation of dust on the pallets — moves the fractional period excess from 3.05×10^{-4} to 6.86×10^{-4}, a change of 3.81×10^{-4}, that is, of 33 s a day. With the cheeks fitted the same drift moves the rate by nothing at all, because Equation (34.6) contains no amplitude. That is a null result of the strongest kind: not a small residual, but a term absent by construction.

Phenomenon 34.2 (Exact isochronism on a cycloid)**.** A body sliding without friction under gravity along an arc of a cycloid, cusps upward, reaches the lowest point of the arc in the same time from every starting height. Unlike the circular pendulum, whose period grows with amplitude, the cycloidal pendulum is isochronous exactly and at every amplitude, up to the cusps; and a pendulum can be made to travel a cycloid by letting its suspension wrap

on two cycloidal cheeks [11]. This is the first case in physics of a nonlinear problem solved exactly by choosing the constraint rather than approximating the motion.

Rests on Proposition 28.12 and Equation (28.7).

Derivation. *Derives Phenomenon 34.2.*

Take the cycloid generated by a circle of radius a rolling on the underside of a horizontal line, in the parametrization

$$x = a\,(\theta + \sin\theta)\,, \qquad y = a\,(1 - \cos\theta)\,, \qquad |\theta| \le \pi\,, \tag{34.3}$$

whose lowest point is $\theta = 0$ and whose cusps are at $\theta = \pm\pi$. Differentiating, $\mathrm{d}x = a\,(1 + \cos\theta)\,\mathrm{d}\theta$ and $\mathrm{d}y = a\sin\theta\,\mathrm{d}\theta$, so that

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 = a^2\,(2 + 2\cos\theta)\,\mathrm{d}\theta^2 = 4a^2\cos^2\!\left(\tfrac{\theta}{2}\right)\mathrm{d}\theta^2\,.$$

Measuring the arc s from the lowest point gives $s = 4a\sin(\theta/2)$, and the height above that point is

$$y = a\,(1 - \cos\theta) = 2a\sin^2\!\left(\tfrac{\theta}{2}\right) = \frac{s^2}{8a}\,. \tag{34.4}$$

The height is therefore *exactly* quadratic in the arc length, with no higher corrections — which is the whole content of the choice of curve. The potential energy per unit mass along the path is $gy = gs^2/8a$, so the tangential equation of motion is

$$\ddot{s} = -\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{gs^2}{8a}\right) = -\frac{g}{4a}\,s\,, \tag{34.5}$$

a harmonic equation with no approximation made anywhere, whose period

$$T = 2\pi\sqrt{\frac{4a}{g}} \tag{34.6}$$

is independent of s and hence of the amplitude, up to the full half-arc $s = 4a$. Comparing with $T = 2\pi\sqrt{\ell/g}$, the cycloidal pendulum swings like a circular one of length $\ell = 4a$. That a suspension of that same length, wrapping on cycloidal cheeks generated by circles of radius a, does travel this curve is the geometric fact that the evolute of a cycloid is a congruent cycloid: the radius of curvature of Equation (34.3) is $\rho = \mathrm{d}s/\mathrm{d}\,(\theta/2) = 4a\cos(\theta/2)$, equal to $4a$ at the vertex and falling to zero at the cusps, exactly as the free length of the wrapping suspension does. □

34.1.4 Interpretation

Two results, not one. Galileo's is that a pendulum has a proper frequency, set by its length and by the local gravity and by nothing else — not by the amplitude, not by the mass, not by the material. That is what makes a swinging weight a *standard* rather than a curiosity, and it is the observational content of the harmonic approximation of Chapter 28: the mass cancels from $\ell\ddot{\theta} = -g\sin\theta$ before any linearization is performed, which is why the cork and the lead keep step, while the amplitude survives in $\sin\theta$ and drops out only when the equation is linearized.

Huygens' result is the sharper one, because it measures the error of that linearization instead of assuming it away. The circular pendulum's period grows with amplitude as $T\,(A) = T\,(1 + A^2/16 + O(A^4))$, derived in Chapter 28; at an amplitude of 7°, that is about one part in 10^3, and a clock whose amplitude wanders by that much gains or loses of order 81 s a day (Table 34.1). A minute and a half a day is not a subtlety for a seventeenth-century navigator, and the cycloidal cheeks are the response: not a smaller error but a structurally zero one. The trade is instructive and recurs throughout this treatise — Huygens bought exactness in the amplitude at the cost of a new systematic, since the cheeks must be cut to the right curve and the suspension must be perfectly flexible, and neither condition is better satisfied in practice than the small-angle condition it replaced.

34.1.5 Primary references

[4, 11]. Huygens' *Horologium Oscillatorium* also contains the theory of the compound pendulum used in Section 34.3 and the proposal for a universal length standard discussed in Section 34.4.

34.2 Coulomb: the torsion pendulum (1784)

Tests Phenomenon 30.60 and Equation (30.55).
Assuming Equations (28.2) *and* (30.21).

Replace gravity by a twisted wire and the pendulum becomes something else: an instrument whose restoring constant can be made as small as the wire is fine, and which therefore responds visibly to forces far below anything a balance can weigh. Coulomb established the law of that restoring torque and the scaling of its constant with the geometry of the wire, in the memoir that also gave him the instrument he used the following year on the electrostatic force.

34.2.1 Apparatus

A fine metal wire, hung vertically from a torsion head at the top of a tall case, carrying at its lower end a horizontal needle or bar of known dimensions — so that its moment of inertia I about the wire axis is computable from measured lengths and masses. The case is graduated in angle and closed against draughts; the torsion head can be turned through a read angle, so that a known twist may be imposed on the wire independently of the needle's position. Wires of several materials, and of systematically varied length L and diameter d, are used in turn [36].

34.2.2 Procedure

Two measurements, run on each wire. First, the free torsional oscillation: displace the needle, release it, and time many periods, repeating at several amplitudes. Second, the static twist: turn the torsion head and read the equilibrium deflection of the needle, comparing the angle applied at the top with the angle taken up at the bottom. Then vary L and d and repeat both. Nothing in either measurement requires a force standard, and that is the point of the instrument.

34.2.3 Observations and data

The restoring torque is proportional to the angle of twist and to nothing else: the torsional period is independent of the amplitude, and the static deflection is linear in the applied twist. The constant of proportionality depends on the wire alone — not on the body hung from it — and varies as the fourth power of the diameter and inversely as the length [36].

Coulomb, like Galileo, states his results as proportionalities rather than as numbers with units. Put into the modern form, they are two: the restoring torque is $-\kappa\varphi$ with κ independent of φ, and $\kappa \propto d^4/L$ with the same two exponents for iron and for brass. Table 34.2 gives the second of these in numbers, and the paragraph after it checks the first against a published instrument; the derivation of both from linear elasticity is Phenomenon 34.3 below, and the two equations used here, Equations (34.8) and (34.9), are established there.

The amplitude independence is the sharper of the two findings, and its reach is worth stating in numbers because it is what separates this instrument from the gravity pendulum of Section 34.1. Table 34.1 shows the gravity pendulum's period rising by 18 % between a vanishing arc and 90°. Over the same range — indeed over several complete revolutions of the needle — Coulomb finds no change in the torsional period at all, and Equation (34.8)

Table 34.2: The torsion constant of a round wire, from Equation (34.8), over the range of geometries a laboratory torsion pendulum uses. The shear moduli are the polycrystalline aggregate values of Chapter 30, $\mu = 81.8\,\mathrm{GPa}$ for iron and 47.3 GPa for copper, each carrying about 1 %, which is the uncertainty on every κ below; the diameters and lengths are exact by construction, being the independent variable. T is the period the suspension would give a needle of moment of inertia $I = 1.7 \times 10^{-5}\,\mathrm{kg\,m^2}$ — a bar of mass 20 g and length 0.10 m — by Equation (34.9). Coulomb's measurement is the statement that the first four rows stand in the ratios 1 : 16 : 256 : 4096 and that rows two and five stand in the ratio 2 : 1: a factor 16 in stiffness for a factor 2 in diameter, and a factor 2 for a factor 2 in length.

material	d (mm)	L (m)	κ (N m/rad)	T (s)
iron	0.05	1.0	5.02×10^{-8}	116
iron	0.10	1.0	8.03×10^{-7}	28.9
iron	0.20	1.0	1.29×10^{-5}	7.2
iron	0.40	1.0	2.06×10^{-4}	1.8
iron	0.10	2.0	4.02×10^{-7}	40.9
copper	0.10	1.0	4.64×10^{-7}	38.0

says why: the constitutive relation is linear in the shear strain, and the strain a torsion pendulum reaches is tiny. Three full turns of the 0.10 mm iron wire of Table 34.2 give a surface shear strain of only $\gamma = (d/2)\,\varphi/L = 9.4 \times 10^{-4}$ — a tenth of a per cent. The angle is enormous and the deformation is not, which is why a wire may be twisted through three revolutions and still answer with a linear torque.

Run the relation backwards and it calibrates. The one torsion instrument in this treatise whose constants are published is Cavendish's: two balls of 0.730 kg on an arm of half-length $a = 0.914\,\mathrm{m}$, so that $I = 2ma^2 = 1.22\,\mathrm{kg\,m^2}$, and torsional periods of $9.0 \times 10^2\,\mathrm{s}$ and $4.5 \times 10^2\,\mathrm{s}$ for the two suspensions he used [44]. Equation (34.9) turns those into $\kappa = 5.9 \times 10^{-5}\,\mathrm{N\,m/rad}$ and $2.4 \times 10^{-4}\,\mathrm{N\,m/rad}$ without any force ever being applied. Inverting Equation (34.8) with the copper modulus asks, for the first of these, a wire of diameter

$$d = \left(\frac{32\kappa L}{\pi\mu}\right)^{1/4} = 0.34\,\mathrm{mm} \qquad \text{at } L = 1.0\,\mathrm{m}\,, \tag{34.7}$$

and the fourth root makes the answer nearly indifferent to the assumed length — halving L gives 0.28 mm. Both are ordinary diameters for drawn wire. The dynamical calibration and the elastic formula therefore agree on the same instrument, which is the consistency Coulomb's memoir asserts and Chapter 33 depends on.

Phenomenon 34.3 (The stiffness of a twisted wire is geometric)**.** A wire twisted about its own axis resists with a torque proportional to the angle, with a constant κ that scales as the fourth power of the wire's diameter and inversely as its length, the remaining factor being a property of the material [36]. Halving the diameter therefore makes the suspension sixteen times softer, which is what allows a torsion pendulum to be tuned over orders of magnitude in sensitivity by a change of wire alone, and what makes it the transducer of choice for forces at the limit of detection.

Rests on Phenomenon 30.60 and Equation (30.55)*.*

Derivation. Derives Phenomenon 34.3.

Twist a circular shaft of radius R, length L and shear modulus μ through a total angle ϕ. Plane cross-sections stay plane and rotate rigidly, so a filament at radius r is sheared

by $\gamma = r\phi/L$ and carries the shear stress $\tau = \mu\gamma = \mu r\phi/L$ — the linear elastic constitutive relation of Chapter 30. Integrating its moment over the section,

$$\text{torque} = \int r\,\tau\,\mathrm{d}A = \frac{\mu\phi}{L}\int r^2\,\mathrm{d}A = \frac{\mu J}{L}\,\phi\,, \qquad J = \int r^2\,\mathrm{d}A = \frac{\pi R^4}{2} = \frac{\pi d^4}{32}\,, \tag{34.8}$$

so that $\kappa = \mu J/L \propto d^4/L$, which is the observed scaling. Two consequences follow at once. The torque is strictly linear in ϕ, so the equation of motion of the suspended body is $I\ddot{\varphi} = -\kappa\varphi$ and the period $T = 2\pi\sqrt{I/\kappa}$ is amplitude-independent, as measured; and, running the argument backwards,

$$\kappa = \frac{4\pi^2 I}{T^2} \tag{34.9}$$

delivers the torque scale of the instrument from a computed moment of inertia and a timed period. No force is ever applied to calibrate it. That is what makes the torsion balance absolute, and it is the reason the same instrument can be turned on gravitation (Chapter 33) and on electrostatics [38] without either measurement borrowing a scale from the other. □

Phenomenon 34.4 (A free torsion pendulum rings for a very long time)**.** Set swinging and left alone, a torsion pendulum on a fine fibre decays far more slowly than any gravity pendulum of comparable period: quality factors of 10^3 and beyond are ordinary for a metal fibre, and much higher values are reached with fused silica. The consequence is practical and dominates the design of every experiment built on the instrument — the equilibrium position cannot be obtained by waiting for the oscillation to stop, because at a period of minutes and a quality factor of 10^3 the amplitude takes of order a day to fall by a factor e. Cavendish already read successive turning points instead of waiting [44], and every later torsion experiment has done some version of the same thing [262, 657, 1991].

Rests on Equation (34.9) *and Definition 28.19.*

Derivation. *Derives Phenomenon 34.4.*

With a linear damping torque the equation of motion is $I\ddot{\varphi} + b\dot{\varphi} + \kappa\varphi = 0$, that is $\ddot{\varphi} + 2\beta\dot{\varphi} + \omega_0^2\varphi = 0$ with $\beta = b/2I$ and $\omega_0^2 = \kappa/I$. The underdamped solution is $\varphi = A\mathrm{e}^{-\beta t}\cos(\omega t + \delta)$ with $\omega = \sqrt{\omega_0^2 - \beta^2}$, so the stored energy, being quadratic in the amplitude, decays as $\mathrm{e}^{-2\beta t}$. Defining the quality factor of Chapter 28 as $Q = \omega_0/2\beta$, the amplitude falls by a factor e in the time

$$\tau_A = \frac{1}{\beta} = \frac{2Q}{\omega_0} = \frac{Q}{\pi}\,T\,, \tag{34.10}$$

that is, after Q/π periods. For a torsion period of a few minutes and $Q \sim 10^3$ this is of order 10^5 seconds, about a day, which is the assertion above. The same Q fixes the sharpness of the resonance, $\Delta\omega \approx \omega_0/Q$. □

It fixes one thing more, and that one decides how a torsion balance should be built. The dissipation that damps the ringdown and the fluctuation that shakes the fibre are the same physical process seen twice, so Q also sets the instrument's noise floor — and the floor is worth having in closed form.

Proposition 34.5 (Thermal torque noise of a torsion pendulum)**.** *A suspension of torsion constant κ, moment of inertia I, resonant angular frequency $\omega_0 = \sqrt{\kappa/I}$ and quality factor Q, in equilibrium at temperature T — the one statement in this chapter in which T is a temperature and not a period — is driven by a fluctuating torque whose one-sided power spectral density is white across the band of the instrument and equal to*

$$S_\tau = 4k_B T b = \frac{4k_B T\kappa}{\omega_0 Q}\,, \tag{34.11}$$

b being the damping coefficient of Equation (34.10). The smallest torque distinguishable in a measurement bandwidth Δf is accordingly

$$\tau_{\min} = \sqrt{S_\tau \, \Delta f} = 2\sqrt{k_B T \, \Delta f} \, \frac{(\kappa I)^{1/4}}{\sqrt{Q}} \, . \tag{34.12}$$

At fixed I and T the floor falls only as the fourth root of the torsion constant, but as the square root of the quality factor: softening the fibre is a weak lever and raising Q is a strong one.

Rests on Equations (34.9) and (34.10).

Derivation. Derives Proposition 34.5.

Add a stochastic torque to the equation of motion used for the ringdown, $I\ddot{\varphi} + b\dot{\varphi} + \kappa\varphi = \tau(t)$. The response to a torque of angular frequency $\omega = 2\pi f$ is $\chi(\omega) = \left(\kappa - I\omega^2 + \mathrm{i}b\omega\right)^{-1}$, so

$$\left\langle \varphi^2 \right\rangle = \int_0^\infty S_\tau \, |\chi(2\pi f)|^2 \, \mathrm{d}f = \frac{S_\tau}{4b\kappa} \, ,$$

the integral being elementary for a quadratic denominator. In thermal equilibrium the suspension holds $\frac{1}{2}\kappa\left\langle\varphi^2\right\rangle = \frac{1}{2}k_B T$ by equipartition (Part XII), that is $\left\langle\varphi^2\right\rangle = k_B T/\kappa$; comparing the two expressions gives $S_\tau = 4k_B T b$. This is the rotational form of Nyquist's relation and an instance of the fluctuation–dissipation theorem [1117]: the coefficient b that damps the ringdown is the same coefficient that drives the fluctuation, because dissipation and fluctuation are one process seen twice. Writing $b = 2I\beta = I\omega_0/Q$ from the definition of Q used in Equation (34.10), and $I\omega_0 = \sqrt{\kappa I} = \kappa/\omega_0$, gives the second form of Equation (34.11) and then Equation (34.12).

Put Cavendish's numbers in it. With $\kappa = 5.9 \times 10^{-5}\,\mathrm{N\,m/rad}$, $I = 1.22\,\mathrm{kg\,m^2}$ and a torsional period of $9.0 \times 10^2\,\mathrm{s}$, so that $\omega_0 = 7.0 \times 10^{-3}\,\mathrm{rad/s}$, and taking $Q = 10^3$ at a room temperature $T = 293\,\mathrm{K}$, Equation (34.11) gives $S_\tau = 1.4 \times 10^{-25}\,\mathrm{N^2\,m^2/Hz}$. Averaging for a day is a bandwidth $\Delta f = 1.2 \times 10^{-5}\,\mathrm{Hz}$, whence $\tau_{\min} = 1.3 \times 10^{-15}\,\mathrm{N\,m}$ — about 5×10^{-9} of the gravitational torque of $2.8 \times 10^{-7}\,\mathrm{N\,m}$ that the instrument had to resolve, as Chapter 33 reconstructs it. Thermal noise was irrelevant to Cavendish by nine orders of magnitude; it is a real limit for the modern instruments [1991, 2135], whose signals are many orders of magnitude smaller. Two cautions belong with Equation (34.11) and are not derived here. It assumes *viscous* damping, a torque proportional to $\dot{\varphi}$; a real fibre dissipates internally, which puts more noise below the resonance than Equation (34.11) predicts and less above it, and is the same physics as the frequency dependence of κ taken up in Phenomenon 34.6. And the relation applies to each dissipation channel separately, with its own b: gas damping, clamp losses and readout damping each contribute a fluctuating torque of their own, so a measured Q that lumps them together is the right number to use only when the fibre dominates. □

34.2.4 Interpretation

The torsion pendulum is the pendulum turned into a force transducer, and Equation (34.9) is why it is an absolute one: a mass distribution, a length and a time fix the torque scale, with no comparison against any other force. That is the property Chapter 33 depends on — Cavendish's gravitational torque is read as $\kappa\theta$ with κ obtained from the instrument's own period — and it is why the same design has served, essentially unchanged, for the electrostatic force law [38], for the gravitational constant [44, 262, 657, 1991], and for the tests of the equality of inertial and gravitational mass [486, 1294, 1469, 2135] that Chapter 42 reports.

The limitation belongs here too, because it is a property of the instrument and not of any one experiment using it.

Phenomenon 34.6 (The torsion constant depends on the frequency)**.** Coulomb's law of torsion is very good but not exact. A real fibre is anelastic: its response lags the applied twist, so that the effective κ measured from a slow static deflection differs from the one inferred from the period of a fast oscillation, by a fractional amount that grows as the quality factor falls. Kuroda showed that this biases the time-of-swing determinations of the gravitational constant upward, by an amount of order $1/\pi Q$ [1901] — some hundreds of parts per million for an ordinary metal fibre, which is the size of the discrepancy that Chapter 33 reports among modern determinations and does not resolve.

Rests on Equations (34.9) *and* (34.10).

Derivation. Derives Phenomenon 34.6.

Only one thing has to be assumed about the material, and it is an observation rather than a model: over the narrow band a torsion balance uses, a metal fibre dissipates with a *loss angle* that is very nearly independent of frequency. Everything else follows from causality.

Let $\hat{\chi}(\omega)$ be the torsional compliance, defined by $\hat{\theta}(\omega) = \hat{\chi}(\omega)\hat{\tau}(\omega)$ with the convention $\theta(t) = \Re\left[\hat{\theta}\,\mathrm{e}^{-\mathrm{i}\omega t}\right]$ of Theorem 17.78, and of SI dimension rad/(N m). The work done per cycle by a torque of amplitude $|\hat{\tau}|$ is $\pi\,|\hat{\tau}|^2\,\Im\hat{\chi}(\omega)$, since the cycle average of $\tau\dot{\theta}$ is $\frac{1}{2}\omega\,|\hat{\tau}|^2\,\Im\hat{\chi}$, so dissipation requires $\Im\hat{\chi} > 0$ for $\omega > 0$; and the loss-angle hypothesis is

$$\frac{\Im\hat{\chi}(\omega)}{\Re\hat{\chi}(\omega)} = \frac{1}{Q}\,, \qquad \omega > 0\,, \tag{34.13}$$

with the same Q as the ring-down of Equation (34.10). To leading order in $1/Q$ the numerator may be evaluated with $\Re\hat{\chi}$ replaced by its value χ_0 at the working frequency, and χ being real in the time domain makes $\Im\hat{\chi}$ odd, so $\Im\hat{\chi}(\omega) = (\chi_0/Q)\,\operatorname{sgn}\omega$.

A constant is not integrable at infinity, so use the once-subtracted form of Equation (17.92), symmetrized over the half line by the parity of the real and imaginary parts. For any reference frequency ω_r,

$$\Re\hat{\chi}(\omega) - \Re\hat{\chi}(\omega_r) = \frac{2}{\pi}\,\mathrm{PV}\!\int_0^\infty \omega'\,\Im\hat{\chi}(\omega')\left[\frac{1}{\omega'^2-\omega^2} - \frac{1}{\omega'^2-\omega_r^2}\right]\mathrm{d}\omega'\,. \tag{34.14}$$

With $\Im\hat{\chi} = \chi_0/Q$ on the half line the substitution $u = \omega'^2$ removes the ω' in the numerator and leaves an elementary integral,

$$\mathrm{PV}\!\int_0^\infty \left[\frac{1}{u-\omega^2} - \frac{1}{u-\omega_r^2}\right]\mathrm{d}u = \left[\ln\left|\frac{u-\omega^2}{u-\omega_r^2}\right|\right]_0^\infty = -2\ln\frac{\omega}{\omega_r}\,, \tag{34.15}$$

the upper limit contributing nothing because the ratio tends to one. Hence $\Re\hat{\chi}(\omega) = \chi_0\,[1 - (2/\pi Q)\ln(\omega/\omega_r)]$, and inverting to the stiffness $\kappa = 1/\Re\hat{\chi}$, to the same order in $1/Q$,

$$\kappa(\omega) = \kappa_r\left[1 + \frac{2}{\pi Q}\ln\frac{\omega}{\omega_r}\right], \qquad \kappa_r = 1/\chi_0\,. \tag{34.16}$$

From Equations (17.92) *and* (34.13) *(a frequency-independent loss angle inserted into the subtracted dispersion relation).*

The fibre is therefore stiffer the faster it is twisted, by $2/\pi Q$ per e-fold of frequency — which is the first assertion of Phenomenon 34.6, with the fractional discrepancy between a static deflection held for a time t_s and an oscillation at ω_0 equal to $(2/\pi Q)\ln(\omega_0 t_s)$: for $Q = 10^3$ and a deflection read after a few hundred seconds, some parts in 10^3.

Now the bias. The time-of-swing determination of G is Equation (33.12): with ω_1 and ω_2 the torsional frequencies in the two source-mass configurations, and $C_1 \neq C_2$ the geometrical gradients, the equation of motion of Equation (33.5) gives $I\omega_i^2 = \kappa - GC_i$ and the fibre

constant cancels in the difference. It cancels, however, only if it is the *same* constant at the two frequencies. With Equation (34.16) it is not, and

$$I\left(\omega_2^2-\omega_1^2\right)=G\left(C_1-C_2\right)+\kappa(\omega_2)-\kappa(\omega_1)\ . \tag{34.17}$$

The two frequencies differ only by the gravitational gradient, a fractional shift of order 10^{-4}, so $\ln\left(\omega_2/\omega_1\right)=\frac{1}{2}\ln\left(\omega_2^2/\omega_1^2\right)\approx\left(\omega_2^2-\omega_1^2\right)/2\omega_0^2$ with $\omega_0^2=\kappa/I$, whence

$$\kappa(\omega_2)-\kappa(\omega_1)=\frac{2\kappa}{\pi Q}\,\frac{\omega_2^2-\omega_1^2}{2\omega_0^2}=\frac{1}{\pi Q}\,I\left(\omega_2^2-\omega_1^2\right)\ , \tag{34.18}$$

which to leading order is $G\left(C_1-C_2\right)/\pi Q$. The experimenter divides the left side of Equation (34.17) by C_1-C_2 and calls the result G, so what is obtained is

$$G_{\text{measured}}=G\left(1+\frac{1}{\pi Q}\right)\ , \tag{34.19}$$

From Equations (34.17) and (34.18) (substituting the residual stiffness difference and dividing by the geometrical factor).

an upward bias, independent of every geometrical factor and of the sense of the difference. That is Equation (33.14), and it is the second assertion of Phenomenon 34.6: at $Q=2\times10^3$ it is 160 parts per million, and at the $Q\approx1\times10^3$ of a bare metal fibre, twice that. □

Remark 34.7 (What is assumed, and what the model cannot be asked). Two things are quoted rather than derived. The first is the loss mechanism: Equation (34.13) is an empirical description of metals over a limited band, and why internal friction in a drawn wire should be so nearly frequency-independent is materials physics that this treatise does not carry. The second is the range of validity of Equation (34.16), which cannot hold down to $\omega=0$: the logarithm diverges, and a genuinely static compliance is finite. What the model is entitled to describe is the *difference* between two frequencies within the band, which is all Equation (34.19) uses, and a static deflection only in the weaker sense of a measurement completed in a finite time t_s, which is why the discrepancy above was quoted with $\ln\left(\omega_0 t_s\right)$ rather than with an infinite one. Kuroda's paper [1901] makes the same restriction.

34.2.5 Primary references

[36]. The instrument's application to the electrostatic force is [38]; its application to gravitation is Chapter 33, following [44]; the anelasticity systematic is [1901].

34.3 Kater: the reversible pendulum (1818)

Tests Equations (28.7) and (28.9).

Assuming Phenomenon 28.9.

A simple pendulum measures g in principle and not in practice: no real pendulum is a point mass on a massless string, and the correction from the actual mass distribution is precisely as uncertain as the mass distribution is. Kater's reversible pendulum removes that difficulty completely, by exploiting a theorem of Huygens to make the awkward quantity drop out of the answer. The result is the first absolute determination of the acceleration of free fall, and the ancestor of every absolute gravimeter since.

34.3.1 Apparatus

A rigid bar, of brass or iron, carrying two knife edges facing each other and separated by a fixed measured distance, with the two edges placed asymmetrically about the centre of

mass; between them, a heavy weight that can be slid along the bar and clamped, and a smaller slider for fine trimming. The bar rests on hard flat planes so that either knife edge in turn serves as the pivot, and the whole hangs from a massive support — the compliance of the support is a systematic in its own right. The comparison clock is an astronomically regulated pendulum clock, and the distance between the knife edges is measured against a length standard [62].

34.3.2 Procedure

Swing the pendulum about one knife edge and time it; invert it, swing it about the other, and time it again. Move the adjustable weight and repeat, iterating until the two periods agree. When they do, the distance between the knife edges — a length that can be measured directly to high accuracy — *is* the length of the simple pendulum having that period, and

$$g = \frac{4\pi^2 L}{T^2} \tag{34.20}$$

follows with no other quantity involved. The period is compared with the clock by the method of coincidences: the experimental pendulum is set swinging in front of the clock pendulum with a mark visible only when the two are aligned, and the slow drift of the coincidences measures the difference of periods far more precisely than either could be timed alone.

One trap must be avoided by construction, and it is the reason the two knife edges are deliberately placed at unequal distances from the centre of mass; the derivation below identifies it.

34.3.3 Observations and data

Kater does not report a value of g. He reports a length: the length of the pendulum vibrating seconds in the latitude of London, reduced to sea level, to vacuum, and to a stated temperature [62]. That is the natural output of the method, since Equation (34.20) with $T = 2\,\mathrm{s}$ makes the two equivalent, and it is the form in which the seconds pendulum was proposed as a length standard — the proposal Section 34.4 takes up. His figure is quoted in inches; editorial rule 3 of this treatise admits SI only, and the conversion at the exact 25.4 mm to the inch gives 0.994 138 m. That is the figure customarily quoted from the memoir, and it is quoted here on that authority: no copy of the 1818 *Philosophical Transactions* paper is held in this treatise's evidence store, so its digits have not been read off the original. Nothing below turns on the last two of them. Table 34.3 carries that through to g and compares it with the modern picture; Table 34.4 is the error budget, rebuilt here in the terms of Chapter 2 and [1945, 2129] rather than reproduced from the memoir.

The gap between the fourth and fifth rows of Table 34.3 is the honest part of the entry. Kater prints seven significant figures, which asserts a resolution of 2.6×10^{-7}; nothing in the apparatus supports that, and the reason is Table 34.4. Every correction he must apply is between 6×10^{-6} and 1.4×10^{-4} of the answer, and a correction is never known to better than a few per cent of itself. The quadrature sum of the corrections is 2.4×10^{-4}; taking the residual uncertainty of each as a tenth of the correction gives an overall $u_g/g \approx 2 \times 10^{-5}$, by Theorem 2.8. That is the number the last row of Table 34.3 should be compared against, and against it the agreement is unremarkable — which is the correct verdict, and a stronger statement about the method than a spuriously perfect one would be.

Phenomenon 34.8 (Suspension and oscillation centres are reciprocal)**.** For any rigid body swinging about a horizontal axis there is a second, parallel axis — on the far side of the centre of mass, along the same perpendicular — about which the body swings with exactly the same period; and the distance between the two axes is the length of the simple

Table 34.3: Kater's determination, in SI. The first block is his result and what Equation (34.20) makes of it; $T = 2\,\mathrm{s}$ is exact by the definition of the seconds pendulum, so $g = 4\pi^2 L/T^2 = \pi^2 L$ and the uncertainty on g is entirely the uncertainty on L. The second block is the comparison. g_n is the conventional standard value, exact by adoption [2431], and carries no claim about London; the reference value against which Kater should be read is the second entry, the two-term latitude formula $g(\lambda) \approx g_{\mathrm{eq}} + (g_{\mathrm{pol}} - g_{\mathrm{eq}}) \sin^2 \lambda$ of Chapter 29 evaluated at Greenwich, $\lambda = 51.48°$, with the equatorial and polar values quoted there. That formula is itself good only to about 1×10^{-5}, which is why the agreement in the last row must not be read as better than that.

	quantity	value
result	length of the seconds pendulum, London	$0.994\,138\,\mathrm{m}$
	period T, exact by definition	$2\,\mathrm{s}$
	implied $g = \pi^2 L$	$9.811\,75\,\mathrm{m/s^2}$
	figures quoted, as a relative resolution	2.6×10^{-7}
	accuracy actually achieved	$\sim 1 \times 10^{-5}$
modern	conventional g_n (exact)	$9.806\,65\,\mathrm{m/s^2}$
	latitude formula at 51.48°	$9.8118\,\mathrm{m/s^2}$
	Kater minus the latitude formula	$-9 \times 10^{-5}\,\mathrm{m/s^2}$
	the same, relative	-9×10^{-6}

Table 34.4: The systematic corrections a reversible-pendulum determination of g must carry, with their sizes recomputed here for Kater's instrument rather than taken from the memoir. Each is quoted as a fractional shift in g; since $g = 4\pi^2 L/T^2$, a fractional error ϵ in the period enters g as 2ϵ and one in the length enters as ϵ (Corollary 2.9). The 1.5° amplitude, the 5 K temperature departure, the 10 µm edge radius and the 20 m elevation are representative working values, not Kater's; brass has a linear expansion coefficient near $1.9 \times 10^{-5}/\mathrm{K}$ and a density near $8.5 \times 10^3\,\mathrm{kg/m^3}$, and air at laboratory conditions $1.20\,\mathrm{kg/m^3}$. Every one of these is larger than the resolution Kater's seven printed figures claim.

correction	mechanism	$\lvert\delta g\rvert / g$
finite arc	period excess $\theta_0^2/16$ at $\theta_0 = 1.5°$, doubled into g	8.6×10^{-5}
air buoyancy	restoring torque reduced by $\rho_{\mathrm{air}}/\rho_{\mathrm{brass}}$	1.4×10^{-4}
entrained air	added moment of inertia of the co-moving air [135]	$\sim 1.4 \times 10^{-4}$
thermal expansion	$\alpha\,\Delta\vartheta$ on the knife-edge separation	9.5×10^{-5}
clock rate	comparison clock in error by 1 s a day	2.3×10^{-5}
knife-edge radius	rolling rather than pivoting, r/L	1.0×10^{-5}
free-air reduction	$2h/R_\oplus$ to sea level	6.3×10^{-6}
in quadrature		2.4×10^{-4}

pendulum that keeps that period. The relation is symmetric: each axis is the other's centre of oscillation. Huygens established it in the *Horologium Oscillatorium* [11], and it converts an otherwise intractable measurement of g into the measurement of one length and one time.

Rests on Equations (28.7) *and* (28.9).

Derivation. *Derives Phenomenon 34.8.*

Let the body have mass m, centre of mass C, and radius of gyration k about the horizontal axis through C parallel to the pivot, so that its moment of inertia about that axis is mk^2. Suspended from a point at distance h from C, it has, by the parallel-axis theorem, moment of inertia $I = m\left(k^2 + h^2\right)$ about the pivot, while gravity exerts the moment $-mgh\sin\vartheta$. The rotational law of Chapter 19 gives $I\ddot{\vartheta} = -mgh\sin\vartheta$, whose small-amplitude period is

$$T = 2\pi\sqrt{\frac{I}{mgh}} = 2\pi\sqrt{\frac{L\left(h\right)}{g}}\,, \qquad L\left(h\right) = \frac{k^2 + h^2}{h} = h + \frac{k^2}{h}\,. \tag{34.21}$$

So the body swings exactly like a simple pendulum of length $L\left(h\right)$, the *equivalent length*. Now ask which pairs of suspension distances give equal periods. Setting $L\left(h_1\right) = L\left(h_2\right)$ and clearing denominators,

$$\left(h_1 - h_2\right)\left(1 - \frac{k^2}{h_1 h_2}\right) = 0\,, \tag{34.22}$$

so either $h_1 = h_2$, or $h_1 h_2 = k^2$. In the second case, with the two pivots on opposite sides of C, their separation is

$$h_1 + h_2 = h_1 + \frac{k^2}{h_1} = L\left(h_1\right)\,, \tag{34.23}$$

which is the assertion: the two axes are conjugate, and the distance between them is the equivalent length. Note what has vanished. Neither m nor k appears in Equation (34.23); the mass distribution of the bar, its knife edges, its adjustable weight and its clamps have all dropped out, and Equation (34.20) needs only the measured separation and the measured period.

The first root of Equation (34.22) is the trap. Two knife edges placed at equal distances from the centre of mass also give equal periods, for a trivial reason, and their separation $2h$ is *not* the equivalent length — using it in Equation (34.20) yields a value of g that is simply wrong, with nothing in the data to signal the error. This is why the two distances are made deliberately unequal, so that the observed equality of periods can only be the conjugate root. □

34.3.4 Interpretation

Equation (34.20) contains a length and a time, and nothing else. That is the whole claim, and it is a strong one: the measurement is *absolute*, in the same sense as the torsion calibration Equation (34.9) — it does not compare the local g against g somewhere else, and it needs no previously calibrated instrument of any kind. The reversible pendulum accordingly became the standard of absolute gravimetry for more than a century, and the worldwide network of g values it established is the empirical base for the latitude dependence that Chapter 29 derives from the centrifugal term and the flattening of the Earth.

The residual difficulties are all systematic and all mechanical: knife edges are not lines, supports are not rigid, air is neither absent nor inert, and the bar is not at the temperature at which it was measured. Each is a correction of the same order as the accuracy sought, which is the recurring shape of a precision measurement and the reason Chapter 2 insists that an uncertainty statement is part of a result rather than a comment on it. The method

was displaced only when a genuinely different one became available: dropping a corner cube in vacuum and reading its fall interferometrically, which replaces the mechanical corrections with optical ones and reaches some $2 \times 10^{-8}\,\mathrm{m/s^2}$ [1906]. The free-fall instrument is treated with the kinematics it tests, in Chapter 20.

34.3.5 Primary references

Kater's memoir on the length of the pendulum vibrating seconds in the latitude of London, read to the Royal Society in 1818, is the primary source for this experiment [62]. The theorem the method rests on is Huygens' [11]; the modern successor instrument is [1906].

34.4 The seconds pendulum and the metre

Before it was an instrument for measuring g, the pendulum was a candidate for defining the metre. The proposal is Huygens' — a universal measure taken from the length of the pendulum that beats seconds, so that any observer anywhere could reconstruct the unit from a clock [11] — and it is the earliest attempt at what Chapter 2 calls a definition by reproducible procedure rather than by artefact. It was not adopted, and the reason it was not adopted is a measurement made with pendulums.

Phenomenon 34.9 (The seconds pendulum is nearly, but not exactly, a metre)**.** A pendulum beating seconds — one swing per second, hence a period of 2 s — is about 0.9936 m long at the conventional acceleration $9.806\,65\,\mathrm{m/s^2}$, within about 0.7 % of the metre. But it is not the same length everywhere: because g runs from about $9.780\,\mathrm{m/s^2}$ at the equator to about $9.832\,\mathrm{m/s^2}$ at the poles, as Chapter 29 derives from the centrifugal term and the flattening of the Earth, the seconds pendulum is some 5 mm longer at the pole than at the equator, a spread of about one part in 190. A length unit defined this way would therefore have depended on where it was realized.

Rests on Equation (28.7) *and Axiom 2.2.*

Derivation. *Derives Phenomenon 34.9.*

Inverting $T = 2\pi\sqrt{\ell/g}$ with $T = 2\,\mathrm{s}$ gives

$$\ell = \frac{gT^2}{4\pi^2} = \frac{g}{\pi^2}\,, \tag{34.24}$$

numerically $\ell = 0.993\,62\,\mathrm{m}$ at $g = 9.806\,65\,\mathrm{m/s^2}$, and $\ell = 0.990\,92\,\mathrm{m}$ and $\ell = 0.996\,19\,\mathrm{m}$ at the equatorial and polar values quoted above — a difference of 5.3 mm. The near-coincidence with the metre is a historical accident of the same kind as the near-coincidence of the metre with the ten-millionth of the meridian quadrant, and carries no physical content; the spread does. Latitude dependence of this size was well established by pendulum surveys before the metre was defined, so the defect was known in advance rather than discovered afterwards. □

The metre was defined instead from the Earth's meridian, then from a platinum–iridium artefact, then from a krypton wavelength, and now from the speed of light and the second [2333]. The pendulum's descendant in the modern system is not the metre but the second: a counted oscillation of a system whose frequency is a property of nature rather than of an artefact. What changed between Huygens' proposal and the present definition is not the idea but the oscillator — from a weight on a wire, whose frequency depends on where you stand, to an atomic transition, which does not.

34.5 Foucault: the pendulum at the Panthéon (1851)

Tests Phenomenon 29.61.

Assuming Equations (28.7) *and* (29.66).

That the Earth rotates was, by 1851, not in doubt; but every argument for it was astronomical, and therefore rested on observations of things outside the laboratory. Foucault produced the missing argument: an experiment performed in a closed room, consulting nothing beyond its walls, whose result is the rotation rate of the Earth.

34.5.1 Apparatus

A heavy sphere — of order 28 kg, of lead clad in brass — hung on a fine steel wire some 67 m long from the crown of the dome of the Panthéon in Paris, at a latitude close to 48.85°. The suspension is designed to be as nearly isotropic as possible, since any asymmetry in the support gives the two horizontal directions slightly different frequencies and thereby drives the swing off a straight line. A pointer fixed to the bob writes in a low ring of damp sand laid on the floor beneath, so that successive extremes of the swing are recorded as a physical mark. Foucault had already run the experiment twice on a smaller scale, first privately in a cellar with a wire of a couple of metres and then publicly at the Paris Observatory, before moving it to the Panthéon where the long wire makes the effect unmistakable within a single session [134].

34.5.2 Procedure

The release is the delicate part, and it is done with a cord: the bob is drawn aside, tied off to a fixed point with a light thread, the system is allowed to come completely to rest, and the thread is then burned through. Pulling the bob aside by hand and letting go imparts an uncontrolled sideways impulse, which turns the intended straight swing into a narrow ellipse; and an elliptical swing precesses on its own account, at a rate set by the area of the ellipse and the anharmonicity of the motion, which is indistinguishable in the data from the effect being sought. Burning the thread releases the bob from rest along the line of the thread, with no transverse velocity. Thereafter the experiment is left alone and the direction of the swing plane is read off the sand at intervals.

34.5.3 Observations and data

The plane of the swing turns steadily, and it turns clockwise seen from above — the same sense in every trial, in Paris. The rate is close to 11.3° an hour, so that a full circuit takes about 31.8 h [134]. The sense reverses in the southern hemisphere and the rate falls to zero on the equator, both of which follow from the derivation below and neither of which is a free parameter.

Table 34.5 collects the apparatus, the prediction and the reading, and then does with them the thing that makes this an absolute measurement: it inverts the observed rate for Ω and recovers the length of the sidereal day from a weight, a wire and a floor.

Two features of the table carry the argument. The predicted and observed rates agree, and the recovered sidereal day sits within a fifth of its own uncertainty of the astronomical value — which is a weak test of the astronomy and a strong one of the mechanics, since nothing astronomical entered the measurement. And the numbers are small: 2.7 mm of advance per swing at the working amplitude, on a trace made in sand. That is why the wire had to be 67 m long. Shorten it and the rate is unchanged — the precession does not know ℓ — but the swings come faster, and the advance per swing shrinks with the period, so the effect stays detectable only for as long as the pendulum keeps swinging. The advance per swing is $A\,\Omega \sin\lambda\, T_0$ and therefore grows as $A\sqrt{\ell}$: a 2 m pendulum turns its plane by

Table 34.5: The Panthéon pendulum: apparatus, prediction and result. The apparatus figures are Foucault's [134]; g at the station is the two-term latitude formula of Chapter 29 at $\lambda = 48.85°$, and the swing period follows from Equation (28.7). The sidereal angular velocity used for the prediction is $\Omega = 2\pi/86\,164\,\mathrm{s}$, the sidereal day being the solar day divided by 366.2422/365.2422 — one extra rotation per orbit. Foucault quotes no uncertainty; the value assigned to the measured rate here is what a sand trace read to about 1° gives over a ten-hour run, and it propagates into the recovered day by Corollary 2.9. The last two rows are the point of the experiment: the rotation period of the Earth, obtained inside a closed building.

	quantity	value
apparatus	wire length ℓ	67 m
	bob mass	28 kg
	latitude λ	48.85°
	swing amplitude (working)	3 m
predicted	g at the station	$9.8095\,\mathrm{m/s^2}$
	swing period $T_0 = 2\pi\sqrt{\ell/g}$	16.42 s
	Ω (sidereal)	$7.2921 \times 10^{-5}\,\mathrm{rad/s}$
	precession rate $\Omega \sin\lambda$	$5.491 \times 10^{-5}\,\mathrm{rad/s}$
	the same, in degrees per hour	11.32°
	one full turn	$1.144 \times 10^{5}\,\mathrm{s}$
	the same, in hours	31.79 h
	plane turned per swing period	$9.0 \times 10^{-4}\,\mathrm{rad}$
	advance of the trace per swing period	2.7 mm
observed	precession rate	11.3° per hour
	one full turn	31.8 h
	assigned uncertainty on the rate	1.2 %
recovered	sidereal day, from the rate and λ	$(8.64 \pm 0.10) \times 10^{4}\,\mathrm{s}$
	sidereal day, adopted	$8.6164 \times 10^{4}\,\mathrm{s}$

11.3° in an hour just as the Panthéon one does, but at a 0.2 m amplitude each of its 2.8 s swings advances the trace by 31 µm — ninety times less — and it is stopped by air in an hour or two.

The one systematic that can imitate the effect is not air, and not the support: it is the shape of the orbit.

Proposition 34.10 (An elliptical swing precesses on its own account)**.** *Let a pendulum of length ℓ traverse, in a frame that is* not *rotating, an ellipse of semi-axes a and b about the vertical, with $a, b \ll \ell$. Then the anharmonicity of the suspension alone makes the ellipse precess, in the sense in which it is traversed, at*

$$\Omega_{ell} = \frac{3}{8}\,\frac{ab}{\ell^2}\,\omega_0\,, \qquad \omega_0 = \sqrt{\frac{g}{\ell}}\,, \tag{34.25}$$

while the frequency of the motion around the ellipse is $\omega = \omega_0\left[1 - \left(a^2+b^2\right)/16\ell^2\right]$. Both reduce to known results at the two ends of the family. $b \to 0$ recovers the arc correction $\omega = \omega_0\left(1 - \theta_0^2/16\right)$ of Table 34.1, with $\theta_0 = a/\ell$. And $a = b$ recovers the conical pendulum: the bob's total angular rate about the vertical is then $\omega + \Omega_{ell} = \omega_0\left(1 + a^2/4\ell^2\right)$, which is the expansion of the exact conical relation $\omega_{con}^2 = g/\left(\ell\cos\vartheta\right)$ at $a = \ell\sin\vartheta$.

Rests on Equations (28.8) and (34.38).

Proof. Derives Proposition 34.10.

Let x and y be horizontal Cartesian coordinates of the bob with origin at the lowest point of the sphere, $\rho^2 = x^2 + y^2$, and $\omega_0^2 = g/\ell$. The constraint puts the bob at height $z = \ell - \sqrt{\ell^2 - \rho^2} = \rho^2/2\ell + \rho^4/8\ell^3 + O(\rho^6)$, and differentiating it gives $\dot z = S/\ell + O(\rho^3)$ with $S := x\dot x + y\dot y$. Both pieces are needed: the quartic term of the potential and the $\dot z^2$ the constraint generates are of the same order, and keeping only one of them is the commonest way to get the coefficient wrong. To fourth order in the displacement,

$$\mathcal{L} = \frac{m}{2}\left(\dot x^2 + \dot y^2\right) + \frac{m}{2\ell^2}S^2 - \frac{m\omega_0^2}{2}\rho^2 - \frac{m\omega_0^2}{8\ell^2}\rho^4\,. \tag{34.26}$$

The Euler–Lagrange equation for x has $\partial\mathcal{L}/\partial\dot x = m\dot x + \left(m/\ell^2\right)Sx$ and $\partial\mathcal{L}/\partial x = \left(m/\ell^2\right)S\dot x - m\omega_0^2 x - \left(m\omega_0^2/2\ell^2\right)\rho^2 x$; the term $\left(m/\ell^2\right)S\dot x$ cancels between the two sides, and with $\dot S = \dot x^2 + \dot y^2 + x\ddot x + y\ddot y = v^2 - \omega_0^2\rho^2$ to the order retained, both coordinates obey the *same* equation,

$$\ddot x + \left[\omega_0^2 + \varepsilon(t)\right]x = 0\,, \qquad \varepsilon = \frac{1}{\ell^2}\left(v^2 - \frac{\omega_0^2}{2}\rho^2\right)\,, \tag{34.27}$$

and likewise for y, with $v^2 = \dot x^2 + \dot y^2$.

Insert the unperturbed ellipse $x = a\cos\omega t$, $y = b\sin\omega t$ with $\omega \approx \omega_0$. Then $\rho^2 = \left(a^2+b^2\right)/2 + \left[\left(a^2-b^2\right)/2\right]\cos 2\omega t$ and $v^2 = \omega_0^2\left\{\left(a^2+b^2\right)/2 - \left[\left(a^2-b^2\right)/2\right]\cos 2\omega t\right\}$, so

$$\varepsilon = \varepsilon_0 + \varepsilon_2\cos 2\omega t\,, \qquad \varepsilon_0 = \frac{\omega_0^2\left(a^2+b^2\right)}{4\ell^2}\,, \qquad \varepsilon_2 = -\frac{3\omega_0^2\left(a^2-b^2\right)}{4\ell^2}\,. \tag{34.28}$$

The perturbation is thus a constant plus a modulation at twice the frequency, and the modulation acts differently on a cosine and a sine. Removing the secular term as in Theorem 28.72: with $x_0 = a\cos\omega t$ the resonant part of $-\varepsilon x_0$ is $-a\left(\varepsilon_0 + \varepsilon_2/2\right)\cos\omega t$, because $\cos 2\omega t\cos\omega t = \frac{1}{2}\left(\cos\omega t + \cos 3\omega t\right)$; with $y_0 = b\sin\omega t$ it is $-b\left(\varepsilon_0 - \varepsilon_2/2\right)\sin\omega t$, because $\cos 2\omega t\sin\omega t = \frac{1}{2}\left(\sin 3\omega t - \sin\omega t\right)$. Hence

$$\omega_x^2 = \omega_0^2 + \varepsilon_0 + \frac{\varepsilon_2}{2}\,, \qquad \omega_y^2 = \omega_0^2 + \varepsilon_0 - \frac{\varepsilon_2}{2}\,. \tag{34.29}$$

From Equations (34.27) and (34.28) (removing the secular terms of the two components separately).

The two components keep their amplitudes — the vertical angular momentum $m\,(x\dot{y}-y\dot{x})$ is conserved exactly, because Equation (34.26) is axisymmetric, and at this order it is $mab\omega$, so the perturbation can redistribute phase but not amplitude — and acquire a growing phase difference.

That difference is a rotation of the axes. Put $\psi=\omega_x t$ and $2\delta=(\omega_y-\omega_x)\,t$, so that $x=a\cos\psi$ and $y=b\sin(\psi+2\delta)$ with δ small over any interval short compared with its own growth time. The same ellipse turned bodily through a small angle Θ and traversed with phase $\psi'=\psi+\eta$ has components $x=a\cos\psi'-\Theta b\sin\psi'$ and $y=\Theta a\cos\psi'+b\sin\psi'$; expanding both to first order in Θ and η and matching the coefficients of $\cos\psi$ and $\sin\psi$ in each component gives $a\eta+\Theta b=0$ from the first and $\Theta a+b\eta=2\delta b$ from the second, whence

$$\Theta=\frac{2\delta\,ab}{a^2-b^2}\,,\qquad \eta=-\frac{\Theta b}{a}\,. \tag{34.30}$$

Now put in the numbers. From Equation (34.29), $\omega_y^2-\omega_x^2=-\varepsilon_2=3\omega_0^2\,(a^2-b^2)\,/4\ell^2$, so $\omega_y-\omega_x=3\omega_0\,(a^2-b^2)\,/8\ell^2$ and $2\delta=(\omega_y-\omega_x)\,t$; substituting in Equation (34.30),

$$\Omega_{\text{ell}}=\dot{\Theta}=\frac{3\omega_0\,(a^2-b^2)}{8\ell^2}\,\frac{ab}{a^2-b^2}=\frac{3}{8}\,\frac{ab}{\ell^2}\,\omega_0\,, \tag{34.31}$$

which is Equation (34.25). The difference of squares cancels, so the rate depends on the *area* of the ellipse alone, and it changes sign with b, that is with the sense of traversal.

The frequency around the ellipse is the rate of ψ', namely $\omega=\omega_x+\dot{\eta}=\omega_x-\Omega_{\text{ell}}b/a=\omega_x-3\omega_0 b^2/8\ell^2$. With $\varepsilon_0+\varepsilon_2/2=\omega_0^2\,(5b^2-a^2)\,/8\ell^2$ from Equation (34.28), so that $\omega_x=\omega_0\,[1+(5b^2-a^2)\,/16\ell^2]$,

$$\omega=\omega_0\left[1+\frac{5b^2-a^2}{16\ell^2}-\frac{6b^2}{16\ell^2}\right]=\omega_0\left[1-\frac{a^2+b^2}{16\ell^2}\right]\,, \tag{34.32}$$

the second statement of the proposition. Both limiting checks are now computations rather than assertions. At $b=0$, Equation (34.32) is $\omega_0\,(1-\theta_0^2/16)$ with $\theta_0=a/\ell$, and Ω_{ell} vanishes: a plane swing does not precess. At $a=b$ the modulation ε_2 vanishes identically, the two components share the frequency $\omega_x=\omega_0\,(1+a^2/4\ell^2)$, and that single number is the whole angular rate of the bob about the vertical, which is what the proposition asserts as $\omega+\Omega_{\text{ell}}$ — Equations (34.31) and (34.32) sum to it. The matching Equation (34.30) is singular at $a=b$, as it must be, since a circle has no axes whose orientation could be tracked; the rate Equation (34.31) is not, and the circular case is settled by the direct computation just given. □

Equation (34.25) is what the burnt thread is for. At the Panthéon, $\omega_0=0.3827\,\text{rad/s}$ and $\ell^2=4489\,\text{m}^2$, so an ellipse of semi-major axis 3 m precesses at the Foucault rate itself once its minor semi-axis reaches

$$b=\frac{8\ell^2\Omega\sin\lambda}{3a\,\omega_0}=0.57\,\text{m}\,, \tag{34.33}$$

and contributes 1 % of it at $b=5.7\,\text{mm}$. Since $b\simeq v_\perp/\omega_0$ for a release from the end of the swing, holding the systematic at 1 % means releasing the bob with a transverse velocity below 2.2 mm/s, against a peak speed of $a\omega_0=1.15\,\text{m/s}$: a fractional tolerance of 2×10^{-3} on the release. No hand meets that, and a burnt thread does. Two further properties of Equation (34.25) are what let the residue be diagnosed rather than merely feared. It is proportional to the *area* of the ellipse, so it decays as the swing decays, while the Foucault rate does not; and it changes sign with the sense in which the ellipse is traversed, while the Foucault rate is clockwise in Paris whatever the release. Repeating the run with the swing started along a different azimuth, and fitting the rate against amplitude rather than quoting a single number, separates the two.

Phenomenon 34.11 (The swing plane precesses at the vertical component of the Earth's rotation)**.** A long pendulum, released from rest and left undisturbed, does not keep its plane of swing with respect to the room. The plane turns at the rate $\Omega \sin\lambda$, where Ω is the Earth's sidereal angular velocity and λ the latitude: clockwise seen from above in the northern hemisphere, anticlockwise in the southern, one turn per sidereal day at either pole, and not at all on the equator. The room may be sealed and nothing outside it consulted, so the experiment measures Ω from within the laboratory [134]. The rate depends on the latitude and not at all on the length of the pendulum, on its mass, or on the amplitude of the swing — which is what distinguishes the effect from every mechanical imperfection that could mimic it.

Rests on Phenomenon 29.61 and Equation (28.7).

Derivation. *Derives Phenomenon 34.11.*

Work in the frame rotating with the Earth at angular velocity $\boldsymbol{\Omega}$, and use local axes at latitude λ with x east, y north and z up, so that $\boldsymbol{\Omega} = \Omega\,(0, \cos\lambda, \sin\lambda)$. The centrifugal term is time-independent and merely redefines the local vertical and the effective g, as Chapter 29 shows; absorbing it, the equation of motion of the bob is

$$\ddot{\boldsymbol{r}} = \boldsymbol{g}_{\text{eff}} + \frac{\boldsymbol{F}_{\text{tension}}}{m} - 2\,\boldsymbol{\Omega} \times \dot{\boldsymbol{r}}\,. \tag{34.34}$$

For small horizontal excursions the vertical velocity is second order, so $\dot{\boldsymbol{r}} \simeq (\dot{x}, \dot{y}, 0)$ and

$$-2\,\boldsymbol{\Omega} \times \dot{\boldsymbol{r}} = 2\Omega\,(\sin\lambda\,\dot{y}, -\sin\lambda\,\dot{x}, \cos\lambda\,\dot{x})\ .$$

The vertical component is taken up by the tension and does not move the bob; only the vertical component of $\boldsymbol{\Omega}$ survives in the horizontal motion. Writing $\Omega_z = \Omega\sin\lambda$ and using the restoring acceleration $-\omega_0^2\,(x, y)$ with $\omega_0^2 = g/\ell$ from Chapter 28,

$$\ddot{x} = -\omega_0^2 x + 2\Omega_z \dot{y}\,, \tag{34.35}$$

$$\ddot{y} = -\omega_0^2 y - 2\Omega_z \dot{x}\,. \tag{34.36}$$

Combine them into $\zeta = x + \mathrm{i}y$:

$$\ddot{\zeta} + 2\mathrm{i}\Omega_z \dot{\zeta} + \omega_0^2 \zeta = 0\,. \tag{34.37}$$

Now substitute $\zeta = \mathrm{e}^{-\mathrm{i}\Omega_z t}\eta$, which is precisely the statement that η is the motion referred to axes turning at the rate Ω_z. Differentiating, $\dot{\zeta} = \mathrm{e}^{-\mathrm{i}\Omega_z t}\,(\dot{\eta} - \mathrm{i}\Omega_z \eta)$ and $\ddot{\zeta} = \mathrm{e}^{-\mathrm{i}\Omega_z t}\,(\ddot{\eta} - 2\mathrm{i}\Omega_z\dot{\eta} - \Omega_z^2\eta)$; substituting into Equation (34.37), the terms in $\dot{\eta}$ cancel identically and there remains

$$\ddot{\eta} + \left(\omega_0^2 + \Omega_z^2\right)\eta = 0\,. \tag{34.38}$$

In the rotating axes the bob is an ordinary plane pendulum, of frequency $\sqrt{\omega_0^2 + \Omega_z^2}$, which for $\Omega_z \ll \omega_0$ is ω_0 to a fractional accuracy of $(\Omega_z/\omega_0)^2/2$ — for the Panthéon pendulum, of order 10^{-9}. Its plane is fixed in those axes; therefore in the room the plane turns at $-\Omega_z$, that is, at $\Omega\sin\lambda$ in the clockwise sense seen from above when $\lambda > 0$. Numerically, with the sidereal day $86\,164\,\mathrm{s} = 2\pi/\Omega$ and $\lambda = 48.85°$, the period of one full turn is

$$\frac{2\pi}{\Omega\sin\lambda} = \frac{86\,164\,\mathrm{s}}{\sin 48.85°} \approx 1.14 \times 10^5\,\mathrm{s}\,, \tag{34.39}$$

about 31.8 h, or 11.3° an hour — the observed value. Two features of the result carry the physical weight. First, ℓ, m and the amplitude are absent, so the effect cannot be confused with a property of the particular pendulum. Second, only $\sin\lambda$ enters: the horizontal component of $\boldsymbol{\Omega}$ drops out entirely, having been absorbed by the tension, which is why the experiment reports the component of the Earth's rotation along the local vertical and not the whole of it. □

34.5.4 Interpretation

The measurement is local in the strict sense: the apparatus is a weight, a wire and a floor, and the result is the rotation rate of the planet they stand on. This is the sharpest available form of the statement of Chapter 29 that rotation, unlike uniform motion, is detectable from inside a sealed laboratory; and it is evidence about the Earth obtained without astronomy, which is what made the public demonstration decisive for people whom the astronomical argument had not moved.

There is a second reading of the same data, and it is the one that matters later in this treatise. The swing plane is not held fixed by any force — Equation (34.38) contains no torque about the vertical — yet after one sidereal day, during which the station has been carried once around its latitude circle, the plane has turned through $\Omega \sin\lambda \times 2\pi/\Omega = 2\pi \sin\lambda$ and not through 2π. The deficit is

$$\Delta = 2\pi - 2\pi \sin\lambda = 2\pi\,(1 - \cos\vartheta)\;, \qquad \vartheta = \frac{\pi}{2} - \lambda\,, \tag{34.40}$$

which is exactly the solid angle subtended at the centre of the Earth by the spherical cap the latitude circle bounds — a purely geometric quantity, independent of ℓ, of m, of g and of Ω itself. That the deficit *is* the holonomy of the Levi-Civita connection of the sphere, so that the pendulum is performing parallel transport (Definition 13.146) rather than merely producing the same number, is a theorem this treatise does not yet carry: it needs the identification of the holonomy around a closed curve with the integral of the Gaussian curvature over the region enclosed, which Chapter 13 does not prove. Taken with that theorem, the Foucault pendulum is the oldest laboratory measurement of a geometric phase and belongs with the fibre-bundle language of Chapter 14; taken without it, Equation (34.40) is an arithmetical coincidence that this chapter has checked and not explained.

Foucault himself drew the operational conclusion and built the instrument that follows from it: if a swinging mass keeps its plane against a rotating frame, a spinning mass will keep its axis, which is the gyroscope [136]. The optical descendant is Sagnac's interferometer [410], in which the rotation appears as a fringe shift rather than a slow turning; ring-laser and fibre gyroscopes are its industrial form, and they measure $\Omega \sin\lambda$ in seconds where the pendulum needs hours.

34.5.5 Primary references

[134]. The gyroscope sequel is [136]; the Coriolis term whose consequence this is was published by Coriolis in 1835 [100]; the optical successor is [410].

34.6 The pendulum as the first precision instrument

Table 34.6 makes the common structure visible. In each case a quantity that cannot be weighed, timed or seen directly — an elastic constant, the strength of gravity, the rotation of a planet — is converted into the period or the azimuth of an oscillation, and the oscillation is counted. Counting is the one laboratory operation that can be done without error, which is why every subsequent increase in precision in this treatise has taken the same form: the Josephson and quantum-Hall standards of Chapter 128, the atomic clocks behind Chapter 90, and the fringe counting of Chapter 47 are all frequency measurements wearing different hardware.

The lineage is also literal. The suspension that isolates a gravitational-wave interferometer's mirrors is a pendulum, chosen for exactly the reason Coulomb's wire was chosen — a low resonant frequency decouples the mass from ground motion above it — and its thermal noise is governed by the quality factor of Equation (34.10) through the same fluctuation–dissipation argument [1117] that gives Equation (34.12), which is why such suspensions are

Table 34.6: The four pendulum experiments of this chapter. Every one of them is *absolute*: its calibration chain terminates in the metre and the second, and it borrows no scale from another instrument of the same kind. That is the property that made the pendulum the first device in physics whose reading could be quoted with an uncertainty.

Experiment	Measures	Result
Galileo, Huygens 1602–1673	period	the period is set by the length and by g alone, to 2.5×10^{-3} in the seventeenth-century data; the amplitude dependence is removed exactly on a cycloid (Table 34.1)
Coulomb 1784	torque	restoring torque linear in the twist, with $\kappa \propto d^4/L$; the torque scale follows from a moment of inertia and a period (Table 34.2)
Kater 1818	g	an absolute value of the local free-fall acceleration from one length and one time: $9.811\,75\,\mathrm{m/s^2}$ at London, good to about 2×10^{-5} (Table 34.3)
Foucault 1851	$\Omega \sin \lambda$	the swing plane turns at about 11.3° an hour at Paris, one circuit in about 31.8 h, whence a sidereal day of $(8.64 \pm 0.10) \times 10^4$ s (Table 34.5)

made of fused silica and not of metal: the lever that matters is Q. The torsion pendulum of Section 34.2 remains the instrument of choice for gravitation at laboratory scale [1991, 2135] and hence for the least well determined of the constants, as Chapter 33 reports. And the reversible pendulum's successor, the free-fall gravimeter, still measures g the way Kater did in principle: a length and a time [1906].

Chapter 35

Experiment: Waves and Acoustics

Chapter 28 reduces a large part of classical wave physics to four sharp claims. A string clamped at both ends sustains a discrete harmonic series $f_n = nf_1$ with $f_1 = (1/2L)\sqrt{T/\mu}$. A bounded elastic sheet has its own discrete spectrum, whose modes are surfaces rather than curves and whose overtones are *not* integer multiples of a fundamental. A compressible fluid carries non-dispersive waves at the speed $\sqrt{\mathrm{d}p/\mathrm{d}\rho}$, whose value depends on which thermodynamic path the compressions follow. And a source in motion through the medium is heard at a shifted frequency. This chapter reports the measurements that established all four.

They are old — 1636 to 1866 — and they are, for that reason, unusually clean. Each is a *counting* experiment: a number of loops on a thread, a number of nodal lines on a plate, a spacing of dust ridges, a musical interval. Ratios of small integers and ratios of lengths are what the apparatus delivers, and ratios are exactly what the theory predicts without adjustable parameters. Two of the experiments do more than confirm: the sound-speed measurement of Section 35.4 exposed a 15 % discrepancy that stood for one hundred and twenty-nine years and was resolved only by recognizing the compressions as adiabatic, so that an acoustic measurement became a thermodynamic one; and the Doppler test of Section 35.6 is the case in which the medium is real and at rest in the laboratory, which makes the source–observer asymmetry a physical fact here and its absence in optics (Chapter 37) a genuine surprise.

35.1 Mersenne: the laws of the vibrating string (1636)

Tests Equation (35.1).

Assuming Equation (28.46).

The pitch of a stretched string was the first quantitative law of acoustics, and it was found a century before there was a wave equation to explain it. Its interest here is not only historical: the three proportionalities are precisely the content of the formula $f_1 = (1/2L)\sqrt{T/\mu}$ that Chapter 28 derives, and they were measured against an instrument of remarkable precision — the trained ear, which judges frequency *ratios* and not frequencies.

35.1.1 Apparatus

A monochord: one string of gut or of drawn brass stretched over a graduated soundboard between a fixed bridge and a movable one, tensioned by a weight hanging over a pulley at one end, so that the tension is known as a weight and not as a turn of a peg. A set of strings of several materials and several diameters, otherwise identical, allows the mass per unit length to be varied at fixed length and tension. For the absolute measurement a second apparatus is needed: a long, heavy cord, metres in length and slack enough that its

fundamental period is a substantial fraction of a second, stretched horizontally between fixed supports.

35.1.2 Procedure

Three comparisons, each holding two of the three variables fixed. Displace the movable bridge and compare the pitch of the shortened string with that of the whole; the comparison is made by ear against the consonances, whose ratios are known in advance and are exact by definition — the octave $2:1$, the fifth $3:2$, the fourth $4:3$. Hang different weights on the same length and find the loads at which the pitch rises by a known interval. Substitute strings of different diameter at the same length and tension and do the same.

The absolute measurement is a separate operation and is the more ingenious of the two. The long cord vibrates slowly enough that its swings can be counted directly against a pendulum or a pulse, giving one frequency in absolute terms. The three laws are then used in the reverse direction, as a transfer standard, to carry that number up to the audible range: shorten the length and reduce the diameter by measured factors, and the frequency scales by factors that a ruler determines.

35.1.3 Observations and data

The three ratio laws, stated without any absolute frequency. Halving the length raises the pitch by an octave, and thirds and quarters of the length give the twelfth and the double octave, so that $f \propto 1/L$ exactly over at least two octaves. Quadrupling the hanging weight raises the pitch by an octave, so that $f \propto \sqrt{T}$. Doubling the diameter of a string of the same material at the same length and tension lowers the pitch by an octave, so that $f \propto 1/d$, and since $\mu \propto d^2$ for a cylinder of fixed density this is $f \propto 1/\sqrt{\mu}$ [2, 4].

Phenomenon 35.1 (Mersenne's laws of the stretched string)**.** The fundamental frequency of a uniform string clamped at both ends varies inversely as its sounding length, as the square root of its tension, and inversely as the square root of its mass per unit length; all three hold as exact proportionalities, over the full range of a musical instrument, before any theory of the vibration exists [2, 4]. The higher partials that the same string sounds together with its fundamental stand to it in the ratios of the whole numbers.

Derivation. *Derives Phenomenon 35.1.*

Take from Chapter 28 the clamped-string spectrum

$$f_n = \frac{n}{2L}\sqrt{\frac{T}{\mu}}\,, \qquad n \in \mathbb{N}\,. \tag{35.1}$$

Holding any two of L, T, μ fixed and varying the third reads off the three laws directly, and setting $n = 2, 3, 4$ at fixed L, T, μ gives the whole-number partials. What must be derived *here* is the transfer, because the absolute measurement is made on one string and reported for another. Let a cord of length L and diameter d be compared with a string of the same material and the same tension, of length L' and diameter d'. Since $\mu = \tfrac{1}{4}\pi\rho d^2$ for a cylinder of density ρ, Equation (35.1) gives

$$\frac{f_1'}{f_1} = \frac{L}{L'}\,\frac{d}{d'}\,, \tag{35.2}$$

in which the tension, the density and the numerical factors have all cancelled. Every quantity on the right is a length measured with a ruler. A frequency of a few hertz, counted by eye on a cord metres long and a millimetre thick, therefore fixes the frequency of a string centimetres long and a tenth of a millimetre thick — three orders of magnitude away — with no clock faster than a pulse and no calibration of the material at all. □

35.1.4 Interpretation

Two things are established, and they are of different kinds. The first is the empirical law itself, which any theory of the vibrating string was thereafter obliged to reproduce; d'Alembert's wave equation [25] and Bernoulli's modal superposition [28] were tested against it a century later, and it is the reason the eigenvalue problem of Chapter 9 was posed at all. The second is methodological, and is the reason this experiment opens the chapter: Equation (35.2) shows that the whole of acoustics can be done in ratios. The ear is a poor absolute instrument and an excellent comparator, resolving a mistuned unison of a fraction of a per cent; the theory predicts ratios; and the apparatus is built so that what is read off is a ratio of lengths. The same discipline governs every experiment below.

One caution belongs with the result. The absolute frequency obtained by the transfer is only as good as the assumption that Equation (35.1) extrapolates unchanged across those three orders of magnitude — that is, that the string is non-dispersive and the vibration linear. That assumption is itself a prediction of the theory, and it is tested independently in Section 35.2, where the same law is checked at fixed frequency across a whole family of modes.

35.1.5 Primary references

[2, 4]. Galileo's *Discorsi* states the same three rules, arrived at independently and published two years later; the two accounts are usually cited together and are so cited here.

35.2 Melde: standing waves on a driven string (1860)

Tests Equation (35.3).

Assuming Equation (35.1).

Mersenne's laws are read from the sounding pitch of a freely vibrating string. Melde inverted the arrangement: drive the string at a *fixed, known* frequency and vary the tension instead, and the standing-wave pattern becomes visible, countable, and present only at discrete settings of the load. The experiment thereby shows the eigenvalue structure of the boundary-value problem directly, in a form that needs no ear at all.

35.2.1 Apparatus

A tuning fork of known frequency, maintained electromagnetically so that it runs continuously at constant amplitude. A light silk or cotton thread tied to one prong, led horizontally over a smooth pulley at a measured distance, and terminated in a scale pan carrying weights, so that the tension is a known weight. The fork is mounted on a rotatable block, so that its prong can be set to vibrate either perpendicular to the thread or along it.

35.2.2 Procedure

With the prong vibrating transversely, the thread is driven at the fork's frequency. Weights are added to the pan and the thread watched; at most loads it thrashes without pattern, but at particular loads it settles into a stationary figure of n clean loops separated by points that do not move. The load is recorded for each n. Because the prong's excursion is a small fraction of the loops' amplitude, the driven end is a node to good approximation, and the thread behaves as if clamped at both ends.

The fork is then rotated through a right angle, so that it pulls the thread *along* its own length, periodically modulating the tension rather than displacing the end. Standing waves appear again, at different loads, and a direct comparison shows the thread completing one transverse cycle for every two cycles of the fork.

35.2.3 Observations and data

Discrete loads, and a pure-number law relating them. The pattern of n loops appears at a tension T_n, and the tensions found for successive n fall as the inverse square of the number of loops, so that four loops require one quarter the load of two and nine loops one ninth the load of three [155]. In the second, longitudinal configuration the response is at half the driving frequency.

Phenomenon 35.2 (Standing waves appear only at discrete tensions)**.** A string of fixed length and mass per unit length, driven transversely at a fixed frequency f, shows a stationary pattern of n loops only at the discrete tensions

$$T_n = \frac{4L^2 f^2 \mu}{n^2}\,, \qquad n \in \mathbb{N}\,, \tag{35.3}$$

and shows no stationary pattern between them; the loads at which successive patterns appear stand in the ratios $1 : \frac{1}{4} : \frac{1}{9} : \ldots$ [155].

Rests on Equation (35.1).

Derivation. *Derives Phenomenon 35.2.*

A stationary pattern with n loops on a string of length L whose ends are both nodes requires $L = n\lambda/2$, so $\lambda = 2L/n$. The wave travels at $v = \sqrt{T/\mu}$ and the pattern oscillates at the driving frequency, so $v = \lambda f$. Eliminating λ and v,

$$\sqrt{\frac{T_n}{\mu}} = \frac{2Lf}{n}\,, \tag{35.4}$$

which on squaring is Equation (35.3). Two features make this a stringent test rather than a demonstration. First, the ratios $T_n/T_1 = 1/n^2$ contain none of L, f or μ: the experiment can be run without knowing the fork's frequency, the thread's mass or even the distance to the pulley, and it still tests the theory, because a scale pan measures a ratio of loads directly. Second, the prediction is as much about the loads at which *nothing* happens as about those at which something does. A continuum of admissible frequencies would give a pattern at every load; the discreteness is the observable, and it comes from the boundary conditions, not from the wave equation. □

Phenomenon 35.3 (Half-frequency response to a modulated tension)**.** When the same string is driven not by displacing its end but by periodically varying its tension at frequency f, standing waves are again established, but the string oscillates at $f/2$ — one transverse cycle for every two cycles of the drive [155]. The response is absent for weak modulation and sets in above a threshold that increases with the damping.

Rests on Equations (35.1) *and* (35.3).

Derivation. *Derives Phenomenon 35.3.*

Rotating the fork removes the transverse drive altogether: the prong now moves *along* the thread, stretching and relaxing it, so the end is not displaced sideways at all and what the drive modulates is the tension,

$$T(t) = T_0\,(1 + \epsilon \cos \omega t)\,, \qquad \omega = 2\pi f\,, \quad 0 < \epsilon \ll 1\,, \tag{35.5}$$

with ϵ the fractional swing of the load and f the fork's frequency. The length, the density and the end conditions are untouched, so the transverse displacement may still be expanded in the clamped-string modes, $y(x,t) = \sum_n q_n(t) \sin(n\pi x/L)$. Substituting that expansion into $\mu\,\partial_t^2 y = T(t)\,\partial_x^2 y$ and using the orthogonality of the sines, each mode separates:

$$\ddot{q}_n + \omega_n^2\,(1 + \epsilon \cos \omega t)\,q_n = 0\,, \qquad \omega_n = \frac{n\pi}{L}\sqrt{\frac{T_0}{\mu}} = 2\pi f_n\,, \tag{35.6}$$

From Equations (35.1) *and* (35.5) *(inserting the modal expansion into the string equation and projecting on* $\sin(n\pi x/L)$*).*

with f_n the free frequency of Equation (35.1) at the mean tension T_0. Each mode of the thread is therefore an oscillator whose stiffness — not whose applied force — is modulated, and Equation (35.6) is Mathieu's equation [173]. That is exactly the system solved in Chapter 28, and its two conclusions may be taken over unchanged: the amplitude grows exponentially in a band of drive frequencies centred on $\omega = 2\omega_n$, so that the response is at half the drive; and, once the mode's damping is restored, the growth occurs only above the threshold $\epsilon > 2/Q_n$, with Q_n the quality factor of that mode. Both features asserted in the phenomenon follow — the half-frequency response by construction, since a term $\cos\omega t$ multiplying q_n feeds back on itself only through its half-frequency component, and the threshold because the modulation multiplies the amplitude instead of adding to it.

The reduction also yields a prediction that a scale pan tests directly. Under the transverse drive the n-loop pattern needs $f_n = f$, which is Equation (35.3); under the longitudinal drive it needs $f_n = f/2$, so that $(n/2L)\sqrt{T_n'/\mu} = f/2$ and

$$T_n' = \frac{L^2 f^2 \mu}{n^2} = \frac{T_n}{4}\,. \tag{35.7}$$

From Equations (35.3) *and* (35.6) *(imposing* $f_n = f/2$ *in place of* $f_n = f$ *and dividing).*

Turn the fork through a right angle and the same figure of n loops reappears at one quarter of the load that held it before. The ratio contains neither the fork's frequency nor the thread's mass, so it is another pure number read off a set of weights — and it is a prediction, not a datum: the loads for the longitudinal configuration are not quoted in the source, which reports the halving of the frequency. □

Remark 35.4 (What the reduction does not settle). Equation (35.6) is linear, so it predicts unbounded exponential growth above threshold and nothing about the steady amplitude actually seen. What arrests the growth is the stretching of the thread by its own transverse excursion, which raises T_0 with amplitude and detunes the mode out of the unstable band; that saturation, and the further instabilities beyond it, belong to Chapter 32. The linear analysis also locates only the principal band; the full tongue structure in the (ω, ϵ) plane is Floquet's theorem, which Chapter 28 records as a result Part II does not yet carry. Nothing above uses it.

Rests on Equation (35.6).

35.2.4 Interpretation

The transverse configuration makes the discreteness of the spectrum visible to the eye and countable, and it does so at fixed frequency, which is the complement of Mersenne's fixed-mode, variable-frequency measurement: between them the law $f_n = (n/2L)\sqrt{T/\mu}$ is checked along both of its independent directions. The nodes, in particular, are not an artefact of the ear or of any interpretation; they are points of a taut thread that visibly do not move while their neighbours a centimetre away sweep through centimetres, and their number is an integer that anyone in the room can count.

The longitudinal configuration is a different phenomenon, and it is the reason this experiment retains a place in modern physics. Driving a system by modulating one of its *parameters* rather than by applying an external force produces a response at half the drive frequency, above a threshold — the mechanical prototype of parametric amplification, and the classical ancestor of the parametric oscillators of Chapter 122. Melde was not the first to see it: Faraday had already reported a half-frequency response in the "crispations" of liquid on a vibrating plate [88], thirty years earlier and in a two-dimensional system. Melde's string is the case in which the mechanism is transparent, because the modulated parameter is a single measurable number, the tension.

35.2.5 Primary references

[155]; [88] for the earlier observation of a half-frequency response.

35.3 Chladni: the nodal figures of a vibrating plate (1787)

Tests Equation (35.12).
Assuming Equation (35.9).

A string has nodes at isolated points; a plate has them along curves, and the curves can be made to draw themselves. Chladni's figures are the first direct images of the eigenfunctions of a two-dimensional boundary-value problem, and they are still the standard demonstration that a bounded elastic body has a discrete spectrum whose modes have shape.

35.3.1 Apparatus

A flat plate of brass or of glass, square or circular, of the order of a tenth of a millimetre to a millimetre thick and some tens of centimetres across, held rigidly at its centre by a clamp on a stand so that the whole boundary is free. Fine, dry, sifted sand. A violin bow, well rosined.

35.3.2 Procedure

Sand is sprinkled evenly over the upper surface. One point of the rim is damped with a fingertip, which forces a node there, and the bow is drawn downwards across the rim at another point, which forces an antinode there. The plate sounds, and within a second or two the sand migrates off the regions in motion and settles in sharp ridges along the curves that are at rest. Changing the damped and bowed points selects a different mode, and the figure and the pitch change together. Each figure is recorded by drawing, and its pitch is recorded by comparison with a monochord or an organ pipe.

35.3.3 Observations and data

A catalogue of figures, each with its own pitch. The nodal sets are curves — for a circular plate, a family of diameters and concentric circles; for a square plate, systems of lines and hyperbola-like arcs that are not the modes of a rectangular membrane and are strikingly harder to guess. The number of nodal diameters m and nodal circles n labels the mode, and the frequencies of the circular plate were found to rise roughly as the square of $m + 2n$ [39]. Crucially, the pitches are *not* in whole-number ratios: a plate has a discrete spectrum, like a string, but an inharmonic one, which is why a struck plate has no definite musical pitch.

Phenomenon 35.5 (Two-dimensional normal modes have shape)**.** A bounded elastic plate vibrates freely only at a discrete set of frequencies; each is accompanied by a stationary pattern whose nodal set is a system of curves on the surface, determined by the shape of the plate and by the boundary conditions and not by how the plate is excited. The frequencies are not integer multiples of the lowest, and for a circular plate labelled by m nodal diameters and n nodal circles they increase approximately as $(m + 2n)^2$ [39].

Derivation. *Derives Phenomenon 35.5.*

Chapter 30 establishes the equation governing the transverse deflection w of a thin plate of density ρ, thickness h and flexural rigidity D, together with the two conditions a free edge imposes; free harmonic motion $w = W\mathrm{e}^{-\mathrm{i}\omega t}$ reduces it to

$$\nabla^4 W = k^4 W\,, \qquad k^4 = \frac{\rho h \omega^2}{D}\,. \tag{35.8}$$

Everything asserted follows from that eigenvalue problem.

Discreteness and shape come first, and they are the cheapest. The operator in Equation (35.8) is self-adjoint on the plate's own domain under the plate's own edge conditions, so its spectrum is a discrete increasing sequence and each eigenvalue carries an eigenfunction determined by the domain and the boundary alone. The bow selects which eigenfunction is excited and how strongly; it has no other freedom. That is exactly the observed independence of the figure from where and how hard the plate is bowed.

The inharmonicity is read off the dispersion relation. Substituting a plane wave $W \propto \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}$ in Equation (35.8),

$$\omega = k^2 \sqrt{\frac{D}{\rho h}}\,, \tag{35.9}$$

From Equation (35.8) (taking ∇^4 of a plane wave, which multiplies it by k^4, and solving for ω).

quadratic in the wavenumber where the string of Equation (35.1) is linear. That single difference is the whole of the inharmonicity: a clamped string has $k_n = n\pi/L$ in exact arithmetic progression, and a dispersion linear in k carries an arithmetic progression of wavenumbers into an arithmetic progression of frequencies — the harmonic series. Squaring destroys it. Whatever sequence $k_1 < k_2 < \dots$ the boundary conditions select, the frequencies go as k^2, so equally spaced wavenumbers give ever-widening frequency intervals and no ratio of small integers survives.

The square law in the labels needs the geometry of the disc. Since $\nabla^4 - k^4 = (\nabla^2 + k^2)\,(\nabla^2 - k^2)$, every solution of Equation (35.8) is the sum of a propagating part solving the Helmholtz equation and an evanescent part solving the modified one. Separating both in polar coordinates on a disc of radius a and discarding what is singular at the centre,

$$W = [A\,J_m(kr) + B\,I_m(kr)]\cos m\,(\theta - \theta_0)\;, \qquad m \in \mathbb{N} \cup \{0\}\,, \tag{35.10}$$

with J_m the Bessel function of Definition 9.93 and I_m its modified counterpart; m is the number of nodal diameters, and the two free-edge conditions fix B/A and quantize k. Because $I_m(kr)$ grows monotonically, roughly as $\mathrm{e}^{kr}/\sqrt{2\pi kr}$, the evanescent term is confined to within a few $1/k$ of the rim once $ka \gg 1$, and the nodal circles in the interior are those of $J_m(kr)$, at $r = j_{m,s}/k$ with $j_{m,s}$ the s-th positive zero. Exactly n of them lie inside the plate when $j_{m,n} \lesssim ka < j_{m,n+1}$, so the mode carrying n interior circles has $ka \approx j_{m,n}$, up to a shift of order unity fixed by the edge condition. The large zeros of J_m are asymptotically spaced by π,

$$j_{m,n} = \left(n + \frac{m}{2} - \frac{1}{4}\right)\pi + O\!\left(\frac{1}{n}\right)\,, \tag{35.11}$$

whence $ka \approx (\pi/2)\,(m + 2n)$ and, by Equation (35.9),

$$\omega_{mn} \approx \frac{\pi^2}{4a^2}\sqrt{\frac{D}{\rho h}}\,(m + 2n)^2\;, \tag{35.12}$$

From Equations (35.9) and (35.11) (substituting $k \approx \pi(m + 2n)/2a$ into the dispersion relation).

which is the measured law. The combination $m + 2n$ is not a coincidence of bookkeeping: Equation (35.11) says a nodal circle advances ka by a full π while a nodal diameter advances it by $\pi/2$, because the angular index enters as the order of the Bessel function and each unit of order displaces the zeros by half a period. Note also that the prefactor of Equation (35.12) carries the thickness and the inverse square of the diameter, which is the scaling Chapter 30 obtains for geometrically similar plates without solving for any mode at all. □

Remark 35.6 (Part II owes the asymptotics of the Bessel zeros). Equation (35.11) is the one step above that is neither physics nor a computation special to this plate: it is a statement about J_m, and it belongs beside the Bessel material of Chapter 9, which at present carries

the series, the recurrences, the generating function and the orthogonality relation but says nothing about where the zeros lie. The missing statement is that the substitution $u = \sqrt{r}\,R$ puts the Bessel equation in the form $u'' + \left[k^2 - \left(m^2 - \tfrac{1}{4}\right)/r^2\right] u = 0$, whose potential term is negligible for $kr \gg m$, so that the zeros are eventually spaced by π/k with the offset quoted. It is used here only for the asymptotic form of Equation (35.12), which is itself asserted as an approximation and is compared with an experiment that reports pitches to a fraction of a tone; nothing in this chapter would change if the offset $-1/4$ were carried differently.

Rests on Equation (35.11) *and Definition 9.93.*

Remark 35.7 (What is not computed here)*.* The exact eigenvalues are not. Imposing the two free-edge conditions on Equation (35.10) gives a transcendental determinant in ka whose roots depend on Poisson's ratio, and neither those roots nor the nodal figures of the *square* plate — which are not separable and which are the harder and more famous half of Chladni's catalogue — follow from the argument above. What is established is the structure the experiment actually reports: a discrete spectrum, figures owned by the plate rather than by the bow, an inharmonic sequence, and the approximate square law in $m + 2n$.

Rests on Equation (35.10)*.*

Phenomenon 35.8 (The figure depends on the tracer)**.** Fine, light powder such as lycopodium does not settle where sand settles. Sprinkled on the same plate sounding the same mode, it collects at the *antinodes* — the regions of greatest motion — producing the photographic negative of the sand figure [88].

Rests on Phenomenon 35.5.

Half of this is already accounted for. Chapter 30 derives the sand figure from a threshold: a grain leaves the surface wherever the peak acceleration $\omega^2 A$ of the plate exceeds g, is thrown, lands, and is thrown again until it reaches a neighbourhood of a node where the local amplitude is too small to eject it. That argument is about a grain coupled to the *plate*, and it predicts the sand figure and nothing else. The lycopodium figure is the complementary case, in which the grain is coupled to the *air*; what carries it is the steady second-order flow that an oscillating boundary drives in the fluid above it, and that flow is not obtained from any first-order acoustic quantity.

Full derivation: Appendix A.66 (p. 3154).

Derives Phenomenon 35.8.

The argument there is a second-order one throughout, and that is the whole reason it does not fit here. Every first-order acoustic quantity oscillates with zero mean, so the steady transport of a tracer cannot appear until the viscous equations of Chapter 31 are expanded to second order in the amplitude. What emerges is Rayleigh's result: the oscillating viscous layer against the plate rectifies the flow above it into a steady slip velocity proportional to the gradient along the surface of the squared first-order amplitude, directed from the antinodes towards the nodes inside the layer and back the other way in the bulk above it. A grain that the plate never throws is carried by the outer branch of that circulation, and collects where the sand does not.

Remark 35.9 (The discriminator between the two powders)*.* It is tempting to separate sand from lycopodium by the ratio of the grain's Stokes relaxation time to the acoustic period, and that is wrong: at audio frequencies the ratio exceeds unity for both powders, so neither follows the oscillating air. The discriminator is whether the surface ejects the grain at all — the threshold $\omega^2 A > g$ of Chapter 30 — and, for a grain it does not eject, which of the steady streaming drag and the residual contact friction wins. That comparison involves the grain's adhesion to the plate, a quantity this treatise does not model, so what the appendix derives is the streaming flow and the direction of its outer branch, not a predicted threshold radius separating the two powders.

35.3.4 Interpretation

The figures are the observational content of the statement that a two-dimensional wave problem has eigenfunctions. A string can be described without that vocabulary; a plate cannot, because its modes have to be drawn. What the sand traces is a level set of the eigenfunction, namely its zero set, and the fact that this set is independent of how hard the plate is bowed, and of where, so long as the same mode is selected, is precisely the statement that the mode belongs to the plate and not to the excitation.

The inharmonicity matters as much as the discreteness, and it is the sharpest way to see that "discrete spectrum" and "harmonic series" are different claims. The clamped string's exact integer ratios are a special consequence of its linear dispersion relation $\omega = vk$; the plate's governing equation is fourth order in space, its dispersion relation is quadratic in k, and the overtones drift away from whole-number ratios. The same distinction returns in the lattice of Chapter 125, where the departure from linearity in $\omega(k)$ is the physical content of dispersion.

Chladni's plates also forced a theory into existence, which is the cleanest possible illustration of the relation this treatise asserts between experiment and derivation. The figures were exhibited across Europe and could not be explained; the Paris Academy set them as a prize problem; and the answer, Sophie Germain's equation for the elastic surface [70], later put on a consistent footing by Kirchhoff [130], is the plate theory used in Chapter 30 to this day.

Finally, Phenomenon 35.8 is retained here as a standing warning about visualization. The sand figure is not the mode; it is what one particular tracer does in the presence of the mode, and a lighter tracer, coupling to the air rather than to the plate, reports the complementary set. An experiment that images a field through a tracer measures the tracer as well as the field.

35.3.5 Primary references

[39, 88]; [70, 130] for the plate theory the figures provoked.

35.4 Newton, Laplace and the speed of sound (1687–1822)

Tests Equations (35.13) *and* (35.16).
Assuming Equation (35.14).

This is the experiment in which a measurement of a mechanical speed turned out to be a measurement of a thermodynamic quantity. Newton derived the speed of sound in the *Principia* from the elasticity of the air and obtained a value well below what the field measurements gave; the discrepancy was not an experimental error, and it did not go away for one hundred and twenty-nine years.

35.4.1 Apparatus

Two stations on a surveyed baseline of the order of ten kilometres, each with a gun and an observer; chronometers, or in the earliest work a seconds pendulum; thermometer, barometer and hygrometer at each station. Newton's own attempt used no baseline at all: he timed the returning echoes from the far end of the cloister of Neville's Court at Trinity College against a pendulum, adjusting the pendulum until its swing kept step with the echoes.

35.4.2 Procedure

The observer at each station records the interval between seeing the flash of the distant gun and hearing its report. Light crosses the baseline in a time utterly negligible on this scale,

so the interval is the transit time of the sound and the speed follows from the surveyed distance. The essential refinement is *reciprocal* firing: guns are discharged alternately from the two ends and the two transit times are averaged, which cancels the component of the wind along the line to first order — the dominant systematic error, and one that no amount of repetition from a single end will remove [71]. Temperature, pressure and humidity are logged so that the result can be reduced to a standard state, and the whole procedure is repeated on several nights.

35.4.3 Observations and data

The speed of sound in dry air at 0 °C is close to 331.3 m/s, rising to about 343 m/s at 20 °C. It is independent of the amplitude of the sound at ordinary intensities, independent of its frequency across the audible band, and — at fixed temperature — independent of the pressure of the air. Newton's theoretical value, computed from the same measured pressure and density, is 279.9 m/s: low by more than 15 %, a discrepancy some fifty times any plausible error in a ten-kilometre baseline [14].

Table 35.1: The sound speed in dry air, and the two theoretical values it is to be compared with. No uncertainty is attached to any entry because none of them is a measurement reported by a source verified for this treatise: the first two rows are the standard state, the middle rows are computed in the derivation below from those two numbers alone, and the last two are the reference values for dry air, quoted to the precision shown. They are given so that the reader can repeat the arithmetic, not so that a fit can be claimed.

Quantity	Value	Origin
Pressure p_0 of the standard atmosphere	101 325 Pa	exact by definition
Density ρ_0 of dry air at 0 °C and p_0	1.293 kg/m^3	reference value
Ratio of specific heats γ of a diatomic gas	1.400	theory
Newton's isothermal speed $\sqrt{p_0/\rho_0}$	279.9 m/s	computed
Laplace's adiabatic speed $\sqrt{\gamma p_0/\rho_0}$	331.2 m/s	computed
Speed of sound, dry air at 0 °C	331.3 m/s	reference value
Speed of sound, dry air at 20 °C	343.2 m/s	reference value

Phenomenon 35.10 (The speed of sound is set by the adiabatic compressibility)**.** Sound propagates through a gas at the speed

$$c = \sqrt{\frac{\gamma p_0}{\rho_0}}\,, \tag{35.13}$$

with $\gamma = c_p/c_V$ the ratio of the specific heats and not, as Newton assumed, at $\sqrt{p_0/\rho_0}$ [59]. Because $p_0/\rho_0 = RT/M$ for an ideal gas, the speed rises as the square root of the absolute temperature, does not depend on the pressure at fixed temperature, and carries no dispersion.

Derivation. *Derives Phenomenon 35.10.*

Chapter 28 linearizes the equations of a compressible fluid about rest and obtains a wave equation with speed $c^2 = \mathrm{d}p/\mathrm{d}\rho$, the derivative to be taken along the thermodynamic path the gas actually follows. Newton took it at constant temperature. For dry air at 0 °C, with

$p_0 = 101\,325\,\mathrm{Pa}$ and $\rho_0 = 1.293\,\mathrm{kg/m^3}$, the isothermal path $p \propto \rho$ gives

$$c_T = \sqrt{\frac{p_0}{\rho_0}} = \sqrt{\frac{101\,325}{1.293}}\,\mathrm{m/s} = 279.9\,\mathrm{m/s}\,. \tag{35.14}$$

A sound wave, however, compresses the air far faster than heat can diffuse across a wavelength, so the correct path is adiabatic, $p \propto \rho^{\gamma}$, whence $\mathrm{d}p/\mathrm{d}\rho = \gamma p_0/\rho_0$ and Equation (35.13) follows. With the diatomic value $\gamma = 1.40$,

$$c = \sqrt{\gamma}\,c_T = 1.1832 \times 279.9\,\mathrm{m/s} = 331.2\,\mathrm{m/s}\,, \tag{35.15}$$

which agrees with the measured $331.3\,\mathrm{m/s}$ to better than one part in a thousand. The temperature dependence follows from $p_0/\rho_0 = RT/M$: at $20\,^{\circ}\mathrm{C}$ the speed is $331.3\,\mathrm{m/s} \times \sqrt{293.15/273.15} = 343.2\,\mathrm{m/s}$, as observed. The absence of dispersion follows because the coefficient in the wave equation contains no k, so all audible frequencies travel at one speed. □

Phenomenon 35.11 (The shortfall is the square root of the specific-heat ratio)**.** The ratio of the measured speed to the isothermal prediction is $\sqrt{\gamma}$, so that the acoustic measurement determines the ratio of the specific heats of the gas without any calorimetry [14, 59].

Rests on Equations (35.13) *and* (35.14).

Derivation. *Derives Phenomenon 35.11.*

Dividing Equation (35.13) by Equation (35.14) gives $c/c_T = \sqrt{\gamma}$, so

$$\gamma = \left(\frac{c}{c_T}\right)^2 = \frac{c^2\rho_0}{p_0}\,. \tag{35.16}$$

With the measured $c = 331.3\,\mathrm{m/s}$ and the same p_0 and ρ_0 as above, $c/c_T = 1.1836$ and $\gamma = 1.401$ — the value of c_p/c_V for a diatomic gas, obtained from a stopwatch, a survey and a barometer. The shortfall is therefore not a defect in the measurement but a datum: it is $1 - 1/\sqrt{\gamma} = 15.5\,\%$, and it *is* the thermodynamics. □

35.4.4 Interpretation

Three separate results sit in this one measurement, and they are worth separating.

The first is the speed itself, which is the input to every acoustic calculation in this part and which fixes the scale against which the Doppler shifts of Section 35.6 are read.

The second is the null result on dispersion, which is among the sharpest in physics and is almost never stated as one. A chord fired from a distant gun arrives as a chord; thunder from a kilometre away is a rumble but not a glissando; an orchestra heard from the back of a hall is not smeared in time by pitch. Each is an observation that $\mathrm{d}c/\mathrm{d}f$ vanishes to the precision of the ear over three decades of frequency, and each is a direct check on the linearity of the wave equation of Chapter 28. The linearity does fail: at large amplitude the compressions steepen into shocks, and that regime belongs to Chapter 31.

The third is the thermodynamic one, and it is the reason this experiment is included in a mechanics part at all. In 1687 there was no thermodynamics, no specific heat and no notion of an adiabatic process; Newton's calculation was correct given his assumption, and the assumption was the error. Equation (35.16) converts a mechanical measurement into a determination of γ, hence of the number of thermally active degrees of freedom per molecule, which is the subject of Chapter 55 and Chapter 54. That a discrepancy of 15 % was allowed to stand, unexplained and undisguised, for well over a century is a better model of scientific practice than the outcome; Newton himself attempted to close it by invoking the finite size of the air particles and the presence of vapour, and that adjustment was wrong.

35.4.5 Primary references

[14, 59]. The field determinations quoted in outline above are those of the eighteenth-century academies and, most carefully, of the Bureau des Longitudes commission that fired reciprocal cannon over a surveyed baseline south of Paris in 1822 and published in the *Annales de Chimie et de Physique*; that memoir is not among the entries of this treatise's bibliography, and no measured figure is attributed to it here. The reference values for the speed and for the density and pressure of dry air are standard and are used above only as inputs to a derivation the reader can repeat.

35.5 Kundt's tube: sound speeds from dust figures (1866)

Tests Equations (35.17) *and* (35.19).

Assuming Equation (35.13).

A ten-kilometre baseline and a pair of cannon are a heavy way to measure a speed, and they cannot be pointed at an arbitrary gas. Kundt reduced the whole measurement to a bench-top apparatus and, more remarkably, to a ruler: his method eliminates the clock entirely, by comparing two wavelengths at one frequency instead of timing one wave.

35.5.1 Apparatus

A horizontal glass tube, closed at one end by a movable piston and at the other by a light disc mounted on the end of a metal rod that is clamped rigidly at its midpoint and passes into the tube through a sliding seal. Fine dry cork dust, or lycopodium powder, spread in a thin line along the floor of the tube. A rosined cloth for exciting the rod, and a supply of whichever gas is to fill the tube.

35.5.2 Procedure

The rod is stroked lengthwise with the rosined cloth until it sounds its longitudinal fundamental: clamped at the middle and free at both ends, it has a displacement node at the clamp and antinodes at its ends. The disc at its inner end therefore drives the gas column at the rod's frequency. The piston is then slid along the tube until the gas column is resonant, at which point the dust leaves the smooth line and springs into a regular series of sharp transverse ridges. The spacing of the ridges is measured with a scale laid along the tube, and the free length of the rod is measured with the same scale. The tube is then flushed and refilled with a different gas and the measurement repeated, the rod, and hence the frequency, being unchanged.

35.5.3 Observations and data

Sharp, evenly spaced ridges of dust, whose spacing depends strongly on the gas filling the tube and not at all on how hard the rod is stroked [166]. The measured quantities are two lengths: the ridge spacing d and the half-length L_r of the rod from clamp to end.

Phenomenon 35.12 (Dust ridges mark the half-wavelength spacing of a standing wave in a gas)**.** A powder light enough to be moved by the gas but too heavy to remain suspended collects at the displacement nodes of a standing sound wave, forming ridges spaced by half a wavelength [166]. The wavelength in the gas and that in the driving rod are therefore both measurable with a scale at one and the same frequency, and their ratio gives the ratio of the sound speeds,

$$\frac{c_{\mathrm{gas}}}{c_{\mathrm{rod}}} = \frac{d}{L_r}\,. \tag{35.17}$$

Rests on Equation (35.13).

Derivation. *Derives Phenomenon 35.12.*

The rod is clamped at its midpoint and free at both ends, so its longitudinal fundamental has a displacement node at the clamp and antinodes at the ends: each half of the rod is a quarter wavelength, and the whole rod is a half wavelength, $\lambda_{\mathrm{rod}} = 2L_r$ with L_r the clamp-to-end length. In the gas the ridges form at the displacement nodes, which in a standing wave are spaced by half a wavelength, so $\lambda_{\mathrm{gas}} = 2d$. Both waves are sustained by the same source and therefore share one frequency f, so

$$\frac{c_{\mathrm{gas}}}{c_{\mathrm{rod}}} = \frac{f\lambda_{\mathrm{gas}}}{f\lambda_{\mathrm{rod}}} = \frac{2d}{2L_r} = \frac{d}{L_r}\,, \tag{35.18}$$

which is Equation (35.17): the frequency cancels, and with it every requirement for a clock, a calibrated fork or a known pitch. Two gases compared with the same rod give, by Equation (35.13),

$$\frac{d_1}{d_2} = \frac{c_1}{c_2} = \sqrt{\frac{\gamma_1 M_2}{\gamma_2 M_1}} \tag{35.19}$$

at a common temperature, so that a ratio of two ruler readings determines a ratio of specific-heat ratios once the molar masses are known. □

35.5.4 Interpretation

Kundt's tube is Chladni's method transplanted into a gas: a powder is used to make a nodal set visible, and the figure is read as a geometric object rather than as a pitch. What the transplant buys is decisive. In the open-air determination of Section 35.4 the measured quantities are a distance of ten kilometres and a time of thirty seconds, and the systematic error is the wind. Here the measured quantities are two lengths of a few centimetres on the same bench, read with the same scale, and Equation (35.17) shows that the frequency — the one quantity that is hard to hold and hard to calibrate — has dropped out of the answer altogether. Reducing a measurement to a ratio of two lengths taken with one instrument is a recurring strategy in this treatise, and it is the same move that Equation (35.2) makes for the string.

The consequence that reached furthest was thermodynamic. Because Equation (35.19) yields γ for any gas that can be put in a tube, the method turned the ratio of specific heats into a routine measurement, including for vapours that no calorimeter of the period could handle. Kundt and Warburg applied it a decade later to mercury vapour and found a ratio of about 1.67 — the value $\frac{5}{3}$ expected for a gas whose molecules have translational degrees of freedom and nothing else [194]. That was the first clear evidence that a molecule can be a structureless point as far as heat is concerned, and it belongs to the equipartition argument of Chapter 55.

35.5.5 Primary references

The primary source is Kundt's memoir on a new kind of acoustic dust figure and its use for determining the speed of sound in solids and in gases [166], together with the later application to vapours by Kundt and Warburg [194], which returns the monatomic ratio for mercury vapour. Both are printed in Poggendorff's *Annalen der Physik und Chemie* — volumes 127 and 157 of that series, which the publisher's continuous numbering records as 203 and 233. The physics the method tests is that of [59], and the dust-figure technique is that of [39].

35.6 Buys Ballot: the Doppler effect on a moving train (1845)

Tests Equations (35.20) *and* (35.21).

Assuming Equation (35.13).

Doppler predicted in 1842 that motion of a source relative to the medium shifts the frequency received, and applied the idea — wrongly — to the colours of double stars [107]. Three years later Buys Ballot tested the acoustic half of the prediction with the fastest vehicle then available and the most sensitive frequency comparators then available: a locomotive, and musicians.

35.6.1 Apparatus

An open flat wagon drawn by a locomotive on a straight, level stretch of the Dutch Rhenish Railway between Utrecht and Maarssen. A group of trumpeters on the wagon, sounding and holding a note agreed in advance. A second group of trumpeters stationed beside the track. Observers of trained absolute pitch, some standing at the trackside, some riding on the wagon, each writing down the note heard. No instrument other than the human ear is used, and none was needed: the quantity to be measured is a frequency ratio, and a musician reports a frequency ratio as an interval.

35.6.2 Procedure

The train is run past the trackside observers at the highest speed the locomotive will hold, with the wagon trumpeters sounding continuously. Each trackside observer records the pitch heard while the source approaches and again while it recedes, expressed as an interval from the note actually being played. The arrangement is then reversed: the trumpeters stand still beside the track and the observers ride past them. Runs are repeated at several speeds, and over several days, the speed of the train being known from the timetable of the run over a measured stretch.

35.6.3 Observations and data

The pitch heard is higher than the pitch sounded while the source approaches and lower while it recedes; the change occurs as the source passes, and the effect reverses with the direction of travel. The size of the shift grows with the speed of the train and is of the order of a semitone at the speeds attainable. Reversing the roles of source and observer gives the same effect to the accuracy of the method [111].

Phenomenon 35.13 (Doppler shift of sound)**.** The frequency received from a source moving through still air differs from the frequency emitted, being raised on approach and lowered on recession by

$$f' = f\,\frac{c}{c \mp v_s}\,, \tag{35.20}$$

with the upper sign for approach; and an observer moving through still air toward or away from a stationary source receives

$$f' = f\,\frac{c \pm v_o}{c}\,. \tag{35.21}$$

Both are observed, and the shift is of order a semitone for the speeds of a nineteenth-century locomotive [107, 111].

Derivation. *Derives Phenomenon 35.13.*

Work in the frame in which the air is at rest, where the wave speed is c and depends on nothing else. Let the source emit f crests per second while moving toward the observer at v_s. In the time $1/f$ between crests the first crest travels c/f while the source advances v_s/f,

so successive crests are separated in space by $\lambda' = (c - v_s)/f$. A stationary observer meets crests at $f' = c/\lambda' = fc/(c - v_s)$, which is Equation (35.20); recession changes the sign of v_s. If instead the source is at rest and the observer advances at v_o, the spacing of the crests in the air is unchanged at $\lambda = c/f$, but the observer sweeps through them at closing speed $c + v_o$, so $f' = (c + v_o)/\lambda = f(c + v_o)/c$, which is Equation (35.21).

The size is what makes the experiment feasible. An equal-tempered semitone is the ratio $2^{1/12} = 1.0595$; from Equation (35.20) an approaching source produces it when $c/(c - v_s) = 1.0595$, that is when

$$v_s = c\left(1 - \frac{1}{1.0595}\right) = 0.0562\,c \approx 19.3\,\mathrm{m/s}\,, \tag{35.22}$$

taking for c the summer-afternoon value 343 m/s of Section 35.4: some 69 km/h, at the upper end of what a locomotive of 1845 could sustain, and a shift that a musician identifies at once because a semitone is the smallest interval the Western scale names. Note also that the shift is a ratio: the observer needs to know neither c nor the absolute pitch of the trumpet, only the interval. □

35.6.4 Interpretation

The experiment confirms Phenomenon 35.13, and it does so in the one arrangement where the confirmation is unambiguous: the source motion is real, mechanical and reversible, the medium is demonstrably at rest, and the effect changes sign as the train goes by. It is the earliest quantitative test of a prediction that is now the basis of radar, of medical flow imaging, of stellar radial velocities and of the expansion of the universe.

Its deeper content is the asymmetry between Equation (35.20) and Equation (35.21). The two describe physically different situations — source moving through the air, observer moving through the air — and they are not the same function. Expanding both for small speeds,

$$\frac{f'_s}{f} = 1 + \frac{v}{c} + \frac{v^2}{c^2} + \dots\,, \qquad \frac{f'_o}{f} = 1 + \frac{v}{c}\,, \tag{35.23}$$

so that they agree at first order and differ at second, by v^2/c^2. For $v/c = 0.056$ that is about three parts in a thousand, roughly five cents of pitch, far below what an ear can judge against a moving background — which is why Buys Ballot found the two arrangements equivalent and why the asymmetry, though real, is not among his results. Stated positively: sound has a preferred frame, the rest frame of the air, and only relative motion *with respect to it* matters; a sufficiently precise acoustic Doppler measurement can therefore detect motion through the medium. That the corresponding optical experiment finds no such frame — and no such second-order asymmetry, only the symmetric relativistic formula — is the content of Chapter 37 and the point of departure for Chapter 38. The contrast is sharpest when the two are set side by side, which is why the acoustic case is worked out here in full.

A word on Doppler's own paper, in the spirit of the honesty rule that governs this treatise. Doppler applied his result to the colours of double stars and concluded that the colours are caused by orbital motion [107]. That conclusion is wrong: the colours of stars are set by their surface temperatures, and the velocities involved are far too small to shift a star out of the visible band. Buys Ballot said so, in the very memoir that confirmed the acoustic effect [111]. The prediction was right, the application was wrong, and the experiment separated them — the optical shift is real but had to wait for spectroscopy, where a shifted absorption line can be measured against a laboratory standard, rather than for the eye.

35.6.5 Primary references

[107, 111].

35.7 Summary of the evidence

Table 35.2: Tests of the wave predictions of Chapter 28 reported in this chapter. Each result is a ratio or a pure number, which is why apparatus of the seventeenth to the nineteenth century suffices to test them.

Experiment	Prediction tested	Result
Mersenne 1636	$f_1 \propto L^{-1}T^{1/2}\mu^{-1/2}$	all three proportionalities confirmed by musical interval; absolute frequency transferred from a slow cord by ratios of lengths
Chladni 1787	discrete inharmonic plate modes	nodal curves drawn by sand; frequencies of a circular plate rise as $(m+2n)^2$, which follows from the plate's quadratic dispersion and the spacing of the Bessel zeros; the spectrum is not harmonic
Melde 1860	$T_n \propto n^{-2}$ at fixed drive	loops appear only at discrete loads in the ratios $1 : 1/4 : 1/9$; a modulated tension gives a half-frequency response, predicted to hold the same figure at one quarter of the load
Newton 1687, Laplace 1816	$c = \sqrt{\gamma p_0/\rho_0}$	measured 331.3 m/s in dry air at 0 °C against 279.9 m/s isothermal; implies $\gamma = 1.401$
Kundt 1866	$c_{\text{gas}}/c_{\text{rod}} = d/L_r$	sound speeds and γ from two ruler readings, with no clock and no known frequency
Buys Ballot 1845	$f' = fc/(c - v_s)$	pitch raised on approach and lowered on recession, of order a semitone, reversing as the train passes

Every entry in Table 35.2 is a consequence of one linear wave equation together with a boundary condition, and none of them carries a free parameter. Two of them reach outside mechanics: the sound speed is a measurement of c_p/c_V and therefore of the internal structure of molecules, and the Doppler shift is the acoustic prototype of the kinematic effect whose optical counterpart forces the transformations of Chapter 38. One derivation is still pending — the acoustic streaming behind Phenomenon 35.8 — and it is of a kind that recurs throughout the chapter: it concerns how the medium moves the tracer, never whether the wave picture holds. Every claim about the waves themselves is derived here or in the theory chapters this one answers to.

Chapter 36

Experiment: Fluid Flow and Turbulence

Chapter 31 sets out the equations of fluid motion and, in the same breath, concedes how much of their content is out of reach: the transition to turbulence is not derived from them, the turbulent state past it is organized by similarity hypotheses rather than deduced, and the existence of smooth three-dimensional solutions is an open problem [745, 2094] catalogued with the rest in Chapter 133. A theory in that condition is held to account by measurement more strictly than usual. This chapter reports the measurements, chosen to span the range in which the equations are trusted and the range in which they are not: efflux from a vessel, where the ideal theory is right to within a factor the ideal theory cannot compute; Poiseuille's capillaries, where the viscous theory is right to viscometric accuracy and thereby measures the no-slip condition; Reynolds's dye filament, where the laminar solution loses its authority at a definite value of a single dimensionless number; Bénard's convection cells, where a threshold computed from the linearized equations meets a measured one; and Prandtl's water channel, where the thin layer that reconciles the inviscid limit with the observed drag was made visible and then manipulated.

The chapter is deliberately not a vindication. Two of the regularities reported below — the value of the pipe transition threshold and the statistical structure of the flow past it — are not reproduced by any derivation from the Navier–Stokes equations, and they are set down here as constraints that any future theory of turbulence must meet rather than as confirmations of the present one. That distinction is the point of the closing section, and it is why an experiment chapter is owed to this part of mechanics more than to most others.

36.1 Efflux from a vessel: Torricelli's law (1738)

Tests Equations (36.2), (36.3) *and* (36.6).

Assuming Equation (36.1).

The oldest quantitative statement about a fluid in motion is that a liquid escapes from a hole in a vessel at the speed a body would acquire in falling freely from the free surface to the hole. It is the simplest possible test of the ideal theory of Chapter 31, it can be read off with a ruler, and — importantly for what follows — it is right about the speed and wrong about the discharge, by a factor that the ideal theory has no way to supply.

36.1.1 Apparatus

An open vessel of horizontal cross-section A, large against the area a of a sharp-edged orifice pierced in a thin wall at depth h below the free surface. Two variants are used. In the first the head is held constant by a supply and an overflow weir, so that a steady jet issues; in

the second the vessel is simply allowed to empty and the falling level is recorded against time. For the range measurement the orifice is in a vertical side wall at a height z above a horizontal floor carrying a scale. Bernoulli's *Hydrodynamica* works throughout with vessels of exactly this kind, efflux from them being its central subject [22].

A third mouthpiece is fitted for the discharge measurement alone: a thin-walled tube of the same bore a, projecting a short way *inward* from the wall into the body of the liquid rather than being flush with it. Its length is a few bore diameters — long enough that the tube's own walls stand clear of the wall of the vessel, short enough that the jet does not reattach inside it. Nothing else is changed: the same vessel, the same head, the same bore.

36.1.2 Procedure

The efflux speed is obtained twice, by independent routes, so that neither reading depends on trusting the other.

The first route is kinematic. A horizontal jet leaving the side orifice is a projectile: it falls the height z in a time $\sqrt{2z/g}$ and lands at a horizontal range $x = v\sqrt{2z/g}$, so that

$$v = x\sqrt{\frac{g}{2z}}\,, \tag{36.1}$$

which reduces the speed measurement to two lengths and the local g — the same kinematics measured directly in Chapter 20. The second route is integral: with the supply shut off, the time taken to empty the vessel from an initial head H is recorded and compared with the prediction derived below.

The discharge is measured separately, by collecting and weighing the liquid delivered in a timed interval, and is compared with the product of the orifice area and the measured jet speed. The jet is also examined a short distance downstream of the plane of the orifice. The whole discharge measurement is then repeated with the re-entrant tube substituted for the flush orifice at the same head and the same bore, which changes the geometry of the approaching flow and nothing else.

36.1.3 Observations and data

Three facts are found, and they are not all favourable to the ideal theory.

First, the jet speed varies as the square root of the head. In the range measurement this appears in a particularly clean form: combining Equation (36.1) with $v = \sqrt{2gh}$ gives $x = 2\sqrt{hz}$, so the landing point is fixed by the two lengths alone and does not depend on g at all. A vessel with orifices at several depths therefore delivers jets whose landing points are predicted without any measurement of gravity, and the prediction is met.

Second, the vessel empties in the time predicted from that speed by mass conservation, which tests the square-root law over the whole range of heads at once rather than at isolated values.

Third, the volume delivered per unit time is *less* than the product of the orifice area with the measured jet speed. The jet is seen to contract downstream of the orifice, reaching a minimum section — the *vena contracta* — within about half an orifice diameter of the wall, and for a sharp-edged orifice in a thin plate the ratio of that minimum section to the orifice area is in the neighbourhood of 0.6. The speed law is exact; the discharge law is not.

Fourth — and this is the observation that makes the third one tractable — the contraction is stronger, and markedly so, when the liquid leaves through the inward-projecting tube instead of through the flush hole. Head, bore and jet speed are unchanged; the delivery is not. The contraction is therefore a property of the *approach* to the orifice and not of the orifice itself, and the two mouthpieces bracket the effect: the re-entrant tube is the geometry in which the approaching flow is most nearly radial, and the discharge is least.

Phenomenon 36.1 (Torricelli's law)**.** A liquid escaping through a small orifice at depth h below the free surface of a large vessel open to the atmosphere issues with the speed

$$v = \sqrt{2gh}\,, \tag{36.2}$$

independent of the density of the liquid, of the size of the orifice and of the direction in which the orifice faces. The volume discharged per unit time is smaller than $a\sqrt{2gh}$, by a factor near 0.6 for a sharp-edged orifice, because the emerging stream contracts to a section smaller than the hole [22, 30].

Derivation. *Derives Phenomenon 36.1.*

Take the vessel large enough that the free surface descends slowly and the flow may be treated as steady. Follow a streamline from a point on the free surface, where the pressure is atmospheric and the speed negligible because $a \ll A$, to a point in the jet just outside the orifice, where the pressure is again atmospheric because the stream is surrounded by air and is straight enough there to carry no transverse pressure gradient. Bernoulli's theorem along that streamline gives

$$p_0 + \rho g h = p_0 + \tfrac{1}{2}\rho v^2\,,$$

whence Equation (36.2). The density cancels, which is why the law holds equally for water and for mercury, and the orifice area has not entered, which is why it holds for any small hole. The result is exactly the free-fall speed from the head, because the argument is the energy balance of a particle descending the same height.

The emptying time follows by mass conservation. With the instantaneous head h, the free surface descends at $\mathrm{d}h/\mathrm{d}t = -a\,v/A$, so

$$\frac{\mathrm{d}h}{\mathrm{d}t} = -\frac{a}{A}\sqrt{2gh} \quad\Longrightarrow\quad t = \frac{A}{a}\int_0^H \frac{\mathrm{d}h}{\sqrt{2gh}} = \frac{A}{a}\sqrt{\frac{2H}{g}}\,,$$

which is A/a times the free-fall time from the initial head. Its agreement with observation tests Equation (36.2) at every head simultaneously.

What the derivation does not give is the discharge. Bernoulli's theorem was applied along one streamline and says nothing about the area of the bundle of streamlines that reaches the jet: the fluid approaches the orifice from all directions across the inner face of the wall, so the streamlines arrive with an inward transverse velocity that they carry past the plane of the hole, and the stream keeps narrowing until that transverse motion has been arrested. The measured contraction ratio is therefore a statement about the geometry of the approaching flow, not about the speed, and the two must be tested separately — which is why the third observation above falsifies nothing. □

There is exactly one mouthpiece for which the ideal theory does supply the missing factor, and it is the re-entrant one. The reason is worth stating before the calculation: the contraction is undetermined because the pressure distribution over the wall around the orifice is undetermined, and the re-entrant tube is the geometry contrived so that no wall carries an unknown pressure at all.

Proposition 36.2 (The contraction coefficient of a re-entrant mouthpiece)**.** *For steady efflux under a head h through a thin tube of bore a projecting into the vessel, the ideal theory gives the jet section exactly:*

$$C_c := \frac{a_j}{a} = \frac{1}{2}\,, \qquad Q = \tfrac{1}{2}a\sqrt{2gh}\,, \tag{36.3}$$

with no empirical input. For a sharp-edged orifice flush with a plane wall the same theory gives a larger coefficient, computed in Proposition 36.3 below.

Rests on Equation (36.2).

Proof. *Derives Proposition 36.2.*

Apply the steady momentum theorem to the liquid inside the vessel, resolved along the axis of the tube. Because the tube projects inward, it is wetted on both faces, so whatever pressure acts on its outer surface acts equally on its inner one and it transmits no net axial force to the liquid. Everywhere else on the interior the liquid is nearly at rest — the vessel is large, and only in the immediate neighbourhood of the mouth does it acquire appreciable speed — so the pressure on the walls is hydrostatic. A hydrostatic pressure distribution on a closed surface exerts no net horizontal force, and the surface here fails to be closed in exactly one place: the mouth of the tube, of area a at depth h, where there is no wall. Equivalently the patch of the opposite wall facing the mouth is unbalanced. The net axial force on the liquid is therefore

$$F = \rho g h\, a\,. \tag{36.4}$$

In the steady state that force equals the rate at which axial momentum leaves the vessel. The jet carries mass at $\rho a_j v$ and moves at v, so the momentum flux is $\rho a_j v^2$, and with Equation (36.2) for v,

$$\rho a_j\,(2gh) = \rho g h\, a\,, \tag{36.5}$$

From Equations (36.2) and (36.4) (equating the momentum flux of the jet to the unbalanced hydrostatic force and substituting $v^2 = 2gh$).

whence $a_j = a/2$, which is Equation (36.3). Note what has and has not been used: Bernoulli's theorem supplied only the speed, and the geometry supplied the rest; ρ, g and h have all cancelled, so the prediction is a pure number, testable by weighing a delivery against the flush orifice at the same head.

The argument also shows precisely why the flush orifice resists it. There the wall surrounding the hole is wetted on one side only, and the liquid accelerates along it, so the pressure over it is *not* hydrostatic and Equation (36.4) acquires an unknown correction. Fixing that correction means knowing the shape of the free surface bounding the jet, which is the free-streamline problem below. □

Proposition 36.3 (Kirchhoff's free-streamline value for the flush slit)**.** *Let an ideal incompressible liquid discharge under a head h through a sharp-edged slit of width a in a plane wall, the jet being bounded by free streamlines on which the pressure is that of the atmosphere and the speed therefore $\sqrt{2gh}$ by Equation (36.2). Then the free-boundary problem has a closed solution and the asymptotic width of the jet is*

$$C_c = \frac{a_j}{a} = \frac{\pi}{\pi + 2} = 0.611\,, \tag{36.6}$$

with no empirical input [172, 175]. The corresponding problem for a circular orifice in three dimensions is not soluble in closed form, and the measured coefficient lies near the same figure.

Rests on Equations (36.2) and (36.3).

Full derivation: Appendix A.67 (p. 3161).

Derives Proposition 36.3.

What the appendix supplies, and what makes the problem tractable at all, is a change of unknown. The difficulty is that the boundary of the jet is not given: it is a streamline whose position is part of the answer, so the domain of the problem depends on its solution. The free-streamline construction removes that by working in the *hodograph* plane, where the velocity rather than the position is the coordinate: on a free streamline the *speed* is known and only the direction varies, and on a solid wall the direction is known and only the speed varies, so an unknown boundary in the physical plane becomes a known one in the hodograph plane. A conformal map then carries that known region onto a half plane, the complex potential is written there by inspection, and Equation (36.6) follows on integrating

back. The one input this treatise does not carry is the conformal mapping itself: Chapter 8 builds the Cauchy theory and the residue calculus but no theory of conformal maps, and the appendix names that debt where it uses it.

36.1.4 Interpretation

The experiment establishes the ideal theory in its own domain and marks that domain's edge in the same measurement. Where the flow is essentially inviscid and no wall is being followed closely — along the interior streamlines, and in the jet — the Bernoulli relation is quantitatively exact. Where the geometry of the flow must be computed rather than assumed, the ideal theory is silent: it predicts the speed and not the area, and the missing factor of roughly 0.6 is supplied by measurement or by the free-streamline theory of discontinuous motions begun by Helmholtz [172, 175], not by the Bernoulli argument. This pattern — an exact statement along a streamline combined with an undetermined global geometry — recurs at every level of the subject, and in its severe form it is the closure problem of turbulence discussed in Section 36.6.

Proposition 36.2 sharpens the diagnosis rather than softening it. The re-entrant mouthpiece is not an easier case of the same calculation; it is a different kind of case, in which the unknown geometry has been arranged out of the problem by making every wetted surface carry a pressure that is either hydrostatic or self-cancelling. Once that is done a momentum balance settles the discharge exactly, with no free constant and no reference to the shape of the jet. The lesson generalizes and is used twice more in this chapter: where an integral conservation law can be applied to a control volume whose boundary is entirely known, it delivers an exact answer without solving the flow — the same move that gives the drag from the wake in Equation (36.25).

36.1.5 Primary references

[22, 30]. The law is customarily attributed to Torricelli's memoir on the motion of heavy bodies of 1644 [5], which is not among the sources verified for this treatise; it is therefore cited here to the *Hydrodynamica*, where efflux from vessels is treated at length, and to Euler's memoir, which first gives the modern local statement from which Equation (36.2) follows in one line [30]. A reader checking the customary attribution should know that the change of source is deliberate and not an error.

36.2 Poiseuille: flow in fine capillaries (1840)

Tests Equations (36.7) *and* (36.8).
Assuming Equation (31.32).

Poiseuille was a physician, and his question was the circulation of blood in the smallest vessels; his answer was the most accurate measurement of a viscous flow made in the nineteenth century, and it remains the calibration standard for viscometry. The law he found was established empirically, before it was derived, and its most exacting consequence is not the flux formula itself but what the formula's accuracy implies about the fluid's behaviour at the wall.

36.2.1 Apparatus

Glass capillaries drawn to fine bore, mounted horizontally, with bores ranging from a few tens of micrometres up to a few hundred — a span of more than an order of magnitude in diameter, which is what makes a power law in the diameter distinguishable from any other dependence. The driving pressure is supplied by a reservoir of compressed air held

at a measured value, rather than by a gravitational head, so that the pressure difference can be varied independently of the geometry and held steady for the duration of a run. The capillary and its supply sit in a bath of controlled temperature, because the coefficient sought is strongly temperature-dependent; for water near room temperature the dynamic viscosity is close to $1.0 \times 10^{-3}\,\mathrm{kg/(m\,s)}$ and changes by a few percent per kelvin. Delivered volume is collected and timed [105]. Hagen, working independently in brass tubes, used the same logic with a gravitational head [104].

36.2.2 Procedure

One variable at a time. With the tube, the temperature and the length fixed, the delivered volume per unit time Q is measured against the applied pressure difference Δp. With Δp and the tube fixed, Q is measured against the length L, shortened by cutting the same capillary so that the bore is not changed between readings — this is the step that removes the largest systematic error, since two nominally identical capillaries are not identical to the fourth power. With Δp, L and the temperature fixed, Q is measured across the family of capillaries, whose bores are determined separately by weighing a thread of mercury of known length drawn into the tube. Finally the bath temperature is varied to isolate the coefficient.

36.2.3 Observations and data

The delivered flux is found to be strictly proportional to the applied pressure difference and inversely proportional to the length of the tube, over the whole range in which the flow remains steady. Against the diameter it rises as the *fourth* power: halving the bore at fixed pressure difference divides the delivery by sixteen. The coefficient depends on the liquid and on the temperature and on nothing else — not on the material of the tube wall, not on the shape of the entry, not on the pressure level, only on the difference [104, 105].

The law fails in two recognizable ways, and the failures are as informative as the law. In tubes made short enough, the measured flux falls below the prediction, because a length of tube is consumed in establishing the flow profile before it becomes uniform along the axis. And above a sufficiently large flux the proportionality to Δp is lost, the flux rising much more slowly than the pressure difference thereafter; that is the transition of Section 36.3, met here from the other side.

Phenomenon 36.4 (The measured fourth-power law and the no-slip condition)**.** Steady flow of a viscous liquid along a straight circular tube of radius R and length L delivers a volume flux

$$Q = \frac{\pi R^4}{8\mu}\,\frac{\Delta p}{L} \tag{36.7}$$

with no adjustable constant beyond the viscosity μ of the liquid, the agreement holding over more than an order of magnitude in R [104, 105]. Because the coefficient in Equation (36.7) is fixed by the condition that the liquid be at rest against the wall, the accuracy of the fit is a quantitative measurement of that condition, and a liquid retaining any appreciable tangential speed at the wall is excluded.

Rests on Equation (36.2).

Derivation. *Derives Phenomenon 36.4.*

Chapter 31 derives Equation (36.7) from the Navier–Stokes equations with the no-slip condition $v(R) = 0$. What is needed here is the sensitivity of that result to the boundary condition itself, since that is the quantity the experiment actually bounds. Replace no-slip by the general linear slip condition, in which the tangential speed at the wall is proportional

to the local shear rate with a coefficient λ of the dimensions of a length, the *slip length*:

$$v(R) = \lambda \left(-\frac{\mathrm{d}v}{\mathrm{d}r} \right)_{r=R} .$$

The interior solution is unchanged, because the differential equation is untouched; only the constant of integration moves. Writing $v(r) = (\Delta p/4\mu L)\,(R^2 - r^2) + v_s$ and evaluating the shear rate of that profile at the wall, $(-\mathrm{d}v/\mathrm{d}r)_R = \Delta p\, R/2\mu L$, the slip condition fixes

$$v_s = \frac{\lambda\, \Delta p\, R}{2\mu L} ,$$

a plug velocity added uniformly across the section. Integrating,

$$Q = \frac{\pi R^4}{8\mu}\,\frac{\Delta p}{L} + v_s \pi R^2 = \frac{\pi R^4}{8\mu}\,\frac{\Delta p}{L}\left(1 + \frac{4\lambda}{R}\right) . \tag{36.8}$$

Three consequences follow, and they are the experimental content of this section. The correction is a pure enhancement, so slip can only make a tube deliver *more* than Equation (36.7), never less — the observed shortfall in short tubes cannot be slip and must be an entry effect. The correction scales as $1/R$, so it is the finest capillaries that carry the information: a family of tubes obeying the same fourth-power law across a wide range of R is direct evidence that λ/R is negligible throughout, which no single tube could establish. And the bound is quantitative: agreement with Equation (36.7) to a fractional accuracy ϵ in a tube of radius R requires $\lambda \le \epsilon R/4$, so agreement at the percent level in a capillary of radius 10 µm bounds the slip length at a few tens of nanometres — a molecular distance. The liquid is at rest against the wall, and Equation (36.7) is how we know it. □

36.2.4 Interpretation

Two separate things are established. The first is the Navier–Stokes equations themselves in their simplest nontrivial exact solution: the inertial term vanishes identically for this geometry, so Equation (36.7) is a consequence of the viscous stress model alone, and its confirmation to viscometric accuracy is a direct confirmation of Newton's linear relation between stress and rate of strain for these liquids. The second is the boundary condition, which is logically independent of the equations and had to be settled by measurement: Navier's own formulation admitted slip, and it is the capillary data, through Equation (36.8), that reduced the slip length to a molecular scale for ordinary liquids on ordinary walls.

The molecular origin of the coefficient μ is not touched here; it belongs to Chapter 55, whose sharpest prediction — that the viscosity of a gas is independent of its density — was confirmed by Maxwell in a separate measurement [167].

36.2.5 Primary references

[104, 105].

36.3 Reynolds: the dye filament and the critical number (1883)

Tests Equation (36.9).

Assuming Equation (36.7).

This is the experiment that gave the subject its organizing parameter and its central unsolved problem in the same afternoon. Its apparatus is a glass tube and a thread of coloured water; its result is that the laminar solution of Section 36.2 ceases to describe the flow above a definite value of one dimensionless combination, and that the value of that threshold is not a property of the equations alone.

36.3.1 Apparatus

A horizontal glass tube leading out of a large glass-walled tank of water, entering the tank through a smoothly rounded bell mouth so that the fluid is not disturbed on the way in. Three such tubes were used, their diameters spanning a factor of several, so that the same transition could be sought at different sizes. A fine nozzle on the axis of the bell mouth introduces a filament of dyed water drawn from a separate reservoir, at a rate adjusted to match the local flow so that the filament neither spreads nor is drawn out. A valve at the far end of the tube sets the discharge, which is measured by collection and timing; a thermometer in the tank gives the temperature, and hence the viscosity. For the study of the broken filament the room is darkened and the tube illuminated by an electric spark, whose brief flash resolves the structure that continuous light averages into a haze [226].

The tank water is left standing for hours before a run, because the result depends on how quiet it is — a fact discovered in the course of the experiment and not anticipated by it.

36.3.2 Procedure

The valve is opened progressively and the filament watched down the length of the tube. At each setting the discharge is measured, the temperature read, and the dimensionless group

$$\mathrm{Re} = \frac{\rho\, \bar{v}\, D}{\mu} \tag{36.9}$$

formed from the mean speed $\bar{v}$, the tube diameter D, and the density and viscosity of the water at that temperature. The procedure is then repeated with a different tube and, since the viscosity is a strong function of temperature, at a different bath temperature, so that the same value of Equation (36.9) is reached by three independent routes: by changing the speed, by changing the size, and by changing the fluid property.

A second, independent series measures the pressure difference required to drive a given flux along a fixed length of pipe, over a range straddling the transition observed optically.

36.3.3 Observations and data

At low speed the dye filament runs the whole length of the tube as a sharp unbroken thread, and the water around it stays clear: the fluid moves in layers that do not interchange. As the speed is raised the thread persists until, at a definite discharge, it breaks up abruptly at some point down the tube and the colour spreads across the whole section beyond that point. Raising the speed further moves the point of breakdown upstream towards the inlet. Under the spark the region past the breakdown is not a uniform mixture but a mass of distinct eddies, which the continuous-light view had smeared into a haze — the flow is not disordered at the molecular level but structured at scales comparable to the tube.

The transition occurs at a value of Equation (36.9) of order 2000 for ordinary conditions of entry and supply, and at substantially the same value in all three tubes and at all bath temperatures — which is the essential observation, since the speed, the diameter and the viscosity individually differ by large factors between those runs. But the threshold is not sharp and not a constant of nature: with the tank water left undisturbed for long enough and the bell mouth kept clean, the unbroken filament survives to values several times larger, and no upper limit was established.

In the pressure series the two regimes obey different laws. Below the transition the pressure gradient required is proportional to the flux, as Equation (36.7) demands. Above it the gradient rises much faster, as a power of the flux near 1.7 — close to, but measurably below, the square. The change of law occurs at the same place as the optical breakdown, so the two observations are of one event [226].

Modern work has replaced the vague "of order 2000" with a sharp number, and in doing so changed what the number *means*. Below a critical Reynolds number a patch of turbulence deposited in the pipe decays; above it, patches split faster than they decay and the turbulence is sustained indefinitely. That critical point — a statistical threshold between two competing rates, not the value at which a filament happens to break — is $\mathrm{Re}_c = 2040 \pm 10$ [2169].

Table 36.1: The pipe transition as it is currently known. The first entry is what Reynolds's filament shows under ordinary conditions of supply and is not a constant of nature; the second is a sharply defined critical point, being the Reynolds number at which the decay rate of a turbulent patch equals its splitting rate. All three quantities are dimensionless.

Quantity	Value	Source
Re at which the dye filament breaks, ordinary inlet	$\sim 2 \times 10^3$	[226]
Re above which turbulence is sustained rather than transient	2040 ± 10	[2169]
Exponent of the flux in the turbulent resistance law	≈ 1.7	[226]

Phenomenon 36.5 (One parameter controls the transition)**.** Whether flow along a straight circular pipe remains laminar is determined, for a given standard of inlet quality, by the single dimensionless combination Equation (36.9) and not by the speed, the diameter, the density or the viscosity separately. Flows agreeing in Re break down at the same place and carry the same dimensionless resistance, however far apart their dimensional parameters lie [226].

Rests on Equations (36.7) *and* (36.9).

Derivation. *Derives Phenomenon 36.5.*

Measure lengths in units of the diameter D, speeds in units of the mean speed $\bar{v}$, times in $D/\bar{v}$ and pressures in $\rho\bar{v}^2$. The incompressible Navier–Stokes system then contains no dimensional quantity whatever, its viscous term carrying the single coefficient $1/\mathrm{Re}$ with Re as in Equation (36.9), and the boundary conditions on a geometrically similar pipe are identical in those units. Two experiments agreeing in Re therefore pose the identical mathematical problem, and every dimensionless property of their solutions — including whether the laminar solution is realized and including the dimensionless resistance — must agree. The observed collapse of three tubes and several temperatures onto one threshold is that statement measured, and it is the licence for all model-scale hydraulic and aerodynamic testing.

Note precisely what the argument does and does not deliver. It shows that *if* there is a threshold it must lie at a fixed value of Re, and it therefore explains the collapse completely. It says nothing about where that value is. And it does not survive the observation that the threshold moves when the inlet is made quieter: the disturbance amplitude entering the pipe is a further dimensionless datum of the experiment which the nondimensionalization of the interior equations does not see, so a flow specified only by Re is not fully specified at all. The similarity argument is exactly as strong as the control of the boundary data. □

The similarity argument explains the collapse and stops there. One further thing can be settled exactly, and it is negative: the breakdown is not the laminar solution losing stability to an infinitesimal disturbance in the way a heated layer does.

Proposition 36.6 (Rayleigh's inflexion criterion)**.** *Let $U(y)$ be a steady parallel flow of an inviscid fluid between rigid walls at y_1 and y_2. A two-dimensional normal-mode disturbance $\propto \mathrm{e}^{\mathrm{i}k(x-ct)}$ can have $\operatorname{Im} c \neq 0$ — that is, can grow — only if U'' vanishes somewhere in (y_1, y_2) [210]. The Poiseuille profile has U'' constant and non-zero, so no inviscid mode of it grows.*

Rests on Equation (36.7).

Proof. *Derives Proposition 36.6.*

Write the disturbance through a streamfunction ψ, with $u' = \partial_y \psi$ and $v' = -\partial_x \psi$, so that continuity is automatic. Linearizing the Euler equations about $U(y)$ and taking the curl removes the pressure, leaving the transport of perturbation vorticity by the base flow against the gradient of the base vorticity $-U'$:

$$(\partial_t + U\partial_x)\,\nabla^2\psi - U''\,\partial_x\psi = 0\,. \tag{36.10}$$

Putting $\psi = \varphi(y)\mathrm{e}^{\mathrm{i}k(x-ct)}$ and cancelling $\mathrm{i}k$ gives Rayleigh's equation,

$$(U-c)\left(\varphi'' - k^2\varphi\right) - U''\varphi = 0\,, \qquad \varphi(y_1) = \varphi(y_2) = 0\,, \tag{36.11}$$

From Equation (36.10) (substituting the normal mode and dividing by the common factor $\mathrm{i}k$).

the boundary conditions expressing that no fluid crosses either wall.

Suppose $c_i := \operatorname{Im} c \neq 0$. Then $U - c$ never vanishes on the real interval, so Equation (36.11) may be divided by it. Multiplying by the complex conjugate $\bar\varphi$, integrating across the channel and integrating $\int \bar\varphi\varphi''$ by parts — the boundary term dies on the wall conditions —

$$\int_{y_1}^{y_2} \left(|\varphi'|^2 + k^2\,|\varphi|^2\right) \mathrm{d}y + \int_{y_1}^{y_2} \frac{U''\,|\varphi|^2}{U-c}\,\mathrm{d}y = 0\,. \tag{36.12}$$

The first integral is real. In the second, $(U-c)^{-1} = (U-\bar c)\,/\,|U-c|^2$, whose imaginary part is $c_i/\,|U-c|^2$, so the imaginary part of Equation (36.12) reads

$$c_i \int_{y_1}^{y_2} \frac{U''\,|\varphi|^2}{|U-c|^2}\,\mathrm{d}y = 0\,. \tag{36.13}$$

From Equation (36.12) (taking the imaginary part, the first integral being real).

If $c_i \neq 0$ the remaining integral must vanish. Its weight $|\varphi|^2\,/\,|U-c|^2$ is non-negative and not identically zero, so U'' cannot keep one sign: it must vanish somewhere. For the parabolic profile of Equation (36.7), U'' is a non-zero constant, and the condition fails. □

Remark 36.7 (What that does and does not prove). Three limits must be stated, or the proposition will be read as more than it is. It is a criterion for the *inviscid* problem, and viscosity does not merely damp — for some profiles it destabilizes, so an inviscidly stable flow may still have a growing viscous mode. It is proved here for a plane channel and for two-dimensional disturbances. And the corresponding statement for the pipe, that the Hagen–Poiseuille solution has no growing eigenmode at *any* Reynolds number, is a computational result which this treatise cites no primary source for; it is recorded here as uncited, and nothing below rests on more than the plane case proved above. What survives is the conclusion that matters: the observed breakdown of Section 36.3 is of a different kind from the onset of Section 36.4, where an infinitesimal disturbance really does grow at a computable threshold.

Rests on Proposition 36.6.

If no disturbance grows, how does a disturbance ever get large? The question sounds rhetorical and is not: decay of every eigenmode forbids growth only in the long run, and the linearized operator here is not normal — its eigenfunctions are not orthogonal — so a superposition of decaying modes may become very large before it becomes small. The mechanism has a name and a simple physical form.

Proposition 36.8 (Algebraic amplification by the lift-up mechanism)**.** *A streamwise-invariant disturbance of a shear flow $U(y)$ — a pair of vortices whose axes lie along the flow — decays in its own cross-stream components while forcing the streamwise component linearly in time. Before viscosity stops it, the streamwise velocity perturbation exceeds the cross-stream one that created it by a factor of order* Re*:*

$$\frac{u_{\max}}{v} \sim \frac{U_0 R}{\nu} \sim \mathrm{Re}\,, \tag{36.14}$$

with R the pipe radius and U_0 the centreline speed. The growth is algebraic and transient, and occurs although every eigenvalue of the problem has negative real part.

Rests on Equation (36.9) *and Proposition 36.6.*

Proof. *Derives Proposition 36.8.*

Linearize about $U(y)\hat{\boldsymbol{x}}$ and take a disturbance independent of x. The cross-stream components (v, w) then satisfy a problem in the cross-stream plane alone: the base flow does not appear in it, because U has no component in that plane and no gradient along x to be advected. That sub-problem is purely viscous and simply decays, on the time

$$\tau_\nu \sim \frac{1}{\nu k_\perp^2} \sim \frac{R^2}{\nu}\,, \tag{36.15}$$

for a vortex filling the pipe, so $k_\perp \sim 1/R$. The streamwise component does not decouple. Its equation retains one coupling term, the advection of the base profile across its own gradient:

$$\left(\partial_t - \nu\nabla^2\right) u = -v\,U'(y)\,. \tag{36.16}$$

This is a *forced* equation, and that asymmetry — u driven by v, with no reaction of u back on v — is the non-normality, written out. Over times short compared with τ_ν the forcing is essentially steady and

$$u \approx -U'(y)\,v\,t\,,$$

growing linearly, not exponentially. Viscosity ends it at $t \sim \tau_\nu$, so with $U' \sim U_0/R$ and Equation (36.15),

$$u_{\max} \sim \frac{U_0}{R}\,v\,\frac{R^2}{\nu} = v\,\frac{U_0 R}{\nu}\,,$$

which is Equation (36.14). The physical content is plain: a streamwise vortex carries slow fluid from the wall towards the axis and fast fluid the other way, and because the profile it is stirring is steep, a small transverse motion redistributes a large streamwise velocity. What it leaves behind are alternating fast and slow *streaks*, elongated along the pipe.

The consequence for the experiment is a matter of arithmetic. At $\mathrm{Re} = 2\times10^3$ the factor in Equation (36.14) is of order a thousand, so a disturbance entering the bell mouth at a thousandth of the mean speed — far below anything the dye filament would reveal — can produce streaks comparable to the mean flow itself, at which point the nonlinear terms that the linear analysis discarded are no longer small. That is why the threshold moves when the tank is left to settle: what the experimenter controls by waiting is not Re but the amplitude the amplifier is fed, and Equation (36.14) says the amplifier's gain rises with Re. A transition governed jointly by a Reynolds number and a disturbance level is exactly what Equation (36.14) predicts and exactly what Table 36.1 reports. □

Remark 36.9 (Where the derivation stops, and why nothing further is owed)*.* Proposition 36.8 carries the account exactly as far as a linear argument can go, and it is worth stating plainly what lies beyond it, so that the boundary is not mistaken for an omission. What follows the algebraic amplification is the finite-amplitude stage: the streaks left by the lift-up develop a secondary instability of their own, and the breakdown of the streaks regenerates the

streamwise vortices that made them, closing a cycle that sustains itself without any growing eigenmode anywhere in it. That picture is a description of solutions found numerically and observed experimentally, not a derivation from the Navier–Stokes equations; the scaling of the critical disturbance amplitude with a negative power of Re is measured, and the near-1.75 power by which the turbulent resistance grows with the flux is an empirical correlation.

No derivation of the threshold from the equations of motion exists. That is a statement about the subject and not about this book, and it is why nothing in this chapter asserts one: the phenomenon stated here is Phenomenon 36.5, the collapse onto a single parameter, which is derived in full above, and the honest account of the threshold itself is given in the interpretation of this experiment below, as two available states whose boundary no theory predicts [2169]. An unwritten derivation would be owed only if a claim rested on it, and none does.

36.3.4 Interpretation

Three distinct claims must be separated, because they are of very different strength.

The collapse onto one parameter is explained, completely, by the derivation above; it is a consequence of the equations and needs nothing else. The existence of a transition is an observation the equations are believed to contain but from which it has never been extracted: by Proposition 36.6 the profile supports no growing inviscid mode, and the laminar solution is not found to be unstable to infinitesimal disturbances at any Reynolds number, so the observed breakdown cannot be the linear instability that the analogous calculation supplies for the convecting layer of Section 36.4. And the value of the threshold is not a property of the equations at all in the form usually quoted, since it depends on the disturbances present — Reynolds's own runs establish this by moving the threshold with nothing but patience.

That last point deserves emphasis because the number 2000 is so often quoted as though it were a constant. What the experiment establishes is a *lower* bound on the laminar range under realistic conditions, and an upper bound that has never been reached. The honest statement of the result is that the pipe has two available states, that which of them is realized depends jointly on Re and on the level of disturbance admitted [2169], and that no theory currently predicts the boundary between them.

What *is* understood is why the two must enter together. Proposition 36.8 supplies an amplifier whose gain grows with Re and which needs no instability to run, so the product of gain and input — not either alone — decides whether the disturbance reaches the amplitude at which the neglected nonlinear terms take over. A threshold in that product is a curve in the (Re, amplitude) plane, not a number; quoting $\mathrm{Re} \approx 2000$ is quoting one point of it, at whatever inlet quality the quoter had. The sharply defined 2040 ± 10 of Table 36.1 escapes that objection only because it is defined by a different question — whether turbulence, once present, persists — which does not refer to the disturbance that started it [2169].

Reynolds returned to the problem a decade later with the decomposition of the velocity field into mean and fluctuating parts, which produces the averaged equations and the Reynolds stresses [268]; that route reorganizes the problem without closing it, and is taken up in Section 36.6.

36.3.5 Primary references

[226, 268].

36.4 Bénard: cellular convection (1900)

Tests Equations (36.21) *and* (36.22).
Assuming Equation (36.17).

Where the pipe gives a transition no calculation predicts, the layer heated from below gives one that a calculation predicts exactly. The two sit in this chapter side by side deliberately: they show that the difficulty in the pipe is specific and not a general failure of hydrodynamic stability theory.

36.4.1 Apparatus

A shallow horizontal layer of liquid spread on a levelled metal plate that is heated from below and held at a uniform temperature, the upper surface left free and open to the air. Bénard used molten spermaceti and paraffin, liquids whose viscosity is high enough to keep the motion slow and orderly and which solidify on cooling, so that the pattern can be fixed and examined at leisure. Layer depths are of the order of a millimetre. The layer is observed from above, photographically and by the optical distortion of a reflected grid, which registers the deformation of the free surface accompanying the flow [288].

36.4.2 Procedure

The plate temperature is raised slowly and the layer watched for the onset of motion. Once cells appear, the temperature difference across the layer is increased further and the pattern followed; the layer depth is then changed and the whole sequence repeated, so that the cell size can be plotted against the depth — the measurement that distinguishes a pattern imposed by the apparatus from one selected by the fluid.

36.4.3 Observations and data

Below a threshold temperature difference nothing moves: the liquid transports heat by conduction alone, and the layer is optically featureless. Above the threshold a steady circulation appears, and it is not disordered but cellular, the layer dividing into a tessellation of nearly regular hexagons with liquid rising in the interior of each cell and descending along the shared boundaries. The cells are stable, persist indefinitely while the heating is maintained, and vanish when the temperature difference is lowered back through the threshold.

The horizontal size of a cell is proportional to the depth of the layer and independent of the lateral extent of the apparatus, which is far larger. This is the decisive quantitative observation, because it identifies the depth as the only length the fluid uses: the pattern is selected by the physics of the layer and not by its container [288].

Phenomenon 36.10 (Convective onset at a threshold, with a selected horizontal scale)**.** A horizontal fluid layer heated from below remains at rest, carrying heat by conduction, until the imposed temperature difference exceeds a threshold; above it a steady cellular circulation sets in whose horizontal periodicity is a fixed multiple of the layer depth and independent of the size of the vessel [288]. The threshold is controlled by the single dimensionless combination

$$\mathrm{Ra} = \frac{g\alpha\,\Delta T\,d^3}{\nu\kappa}\,, \tag{36.17}$$

in which α is the thermal expansion coefficient, ν the kinematic viscosity, κ the thermal diffusivity and d the depth, and it occurs at a critical value of order a thousand [438].

Rests on Equation (36.17)*.*

Derivation. *Derives Phenomenon 36.10.*

Work in the Boussinesq approximation, in which the density variation is kept only where it multiplies g and the fluid is otherwise treated as incompressible with constant ν and κ. The base state is motionless conduction, with the temperature falling linearly at the rate $\beta = \Delta T/d$ from the hot lower plate to the cold upper one. Write the perturbation velocity

$\boldsymbol{u}$, its vertical component w, and the temperature perturbation θ. Linearizing about that state and taking the double curl of the momentum equation — which removes the pressure and, using $\nabla \cdot \boldsymbol{u} = 0$, leaves an equation for w alone — gives the pair

$$\partial_t \nabla^2 w = g\alpha \nabla_h^2 \theta + \nu \nabla^4 w \,, \qquad \partial_t \theta = \beta w + \kappa \nabla^2 \theta \,, \tag{36.18}$$

with $\nabla_h^2 = \partial_x^2 + \partial_y^2$ the horizontal Laplacian. The first equation says that buoyancy, acting only where the temperature varies horizontally, drives vertical motion against viscosity; the second says that a rising parcel carries the conduction gradient with it and loses the excess by diffusion. The two effects compete, and Ra of Equation (36.17) is the ratio in which they stand.

At marginal stability the disturbance neither grows nor decays, so the time derivatives vanish. Eliminating θ between the two steady equations — solve the second for θ and substitute —

$$\nu\kappa \, \nabla^6 w = g\alpha\beta \, \nabla_h^2 w \,. \tag{36.19}$$

From Equation (36.18) (setting the time derivatives to zero, applying $\kappa\nabla^2$ to the first equation and using $\kappa\nabla^2\theta = -\beta w$ from the second).

Now separate. Horizontally, take any planform $f(x, y)$ with $\nabla_h^2 f = -a^2 f / d^2$, so that a is the horizontal wavenumber measured in units of $1/d$; a roll, a square and a hexagon all satisfy this, which is already the statement that the linear problem fixes the periodicity and not the plan form. Vertically, take the boundaries stress-free and perfectly conducting, so that $w = \partial_z^2 w = \theta = 0$ at $z = 0$ and $z = d$. The single half-wave $w = W_0 \sin(\pi z/d) \, f(x, y)$ meets every one of those conditions, together with the further conditions $\partial_z^4 w = 0$ that Equation (36.19) inherits, and it is the gravest such mode. For it,

$$\nabla^2 w = -\frac{\pi^2 + a^2}{d^2} \, w \,, \qquad \nabla_h^2 w = -\frac{a^2}{d^2} \, w \,,$$

so that Equation (36.19) becomes an identity between numbers,

$$\nu\kappa \, \frac{\left(\pi^2 + a^2\right)^3}{d^6} = g\alpha\beta \, \frac{a^2}{d^2} \,.$$

Multiplying by $d^6/\nu\kappa$ and using $\beta d = \Delta T$ to recognize $g\alpha\Delta T d^3/\nu\kappa$ as the Rayleigh number of Equation (36.17),

$$\mathrm{Ra}(a) = \frac{\left(\pi^2 + a^2\right)^3}{a^2} \,, \tag{36.20}$$

From Equations (36.17) and (36.19) (substituting the half-wave mode and collecting the dimensionless group).

which is the marginal-stability relation [438]. Note that no property of the fluid survives in it: the whole of the material physics has gone into the single number Ra.

The layer becomes unstable as soon as the imposed Rayleigh number exceeds Equation (36.20) for *some* wavenumber, so the threshold is the minimum of that curve. Writing $u = a^2$,

$$\frac{\mathrm{dRa}}{\mathrm{d}u} = \frac{3\left(\pi^2 + u\right)^2 u - \left(\pi^2 + u\right)^3}{u^2} = \frac{\left(\pi^2 + u\right)^2 \left(2u - \pi^2\right)}{u^2} \,,$$

which vanishes only at $u = \pi^2/2$ and changes sign from negative to positive there, so the minimum is unique. Substituting,

$$a_\mathrm{c} = \frac{\pi}{\sqrt{2}} \approx 2.221 \,, \qquad \mathrm{Ra}_\mathrm{c} = \frac{27\pi^4}{4} \approx 657.5 \,. \tag{36.21}$$

The critical wavenumber fixes the horizontal periodicity of the pattern that appears at onset:

$$\lambda_\mathrm{c} = \frac{2\pi d}{a_\mathrm{c}} = 2\sqrt{2}\, d \approx 2.83\, d \,, \tag{36.22}$$

so that one rising and one descending region together span a little under three depths, and a single cell is roughly one and a half depths across. This is the measured proportionality of cell size to depth, with the constant predicted and no parameter free. Note that the container never entered the calculation: the only length available to the linearized problem is d, which is why the selected scale cannot depend on the lateral extent. The numerical threshold does depend on the boundary conditions assumed — for the physically commoner case of two rigid boundaries the same calculation gives a critical Rayleigh number near 1708, with a correspondingly larger critical wavenumber [1199] — so the value quoted must always be paired with the boundary conditions it belongs to. □

Remark 36.11 (Why the rigid case is not done the same way). The half-wave $\sin(\pi z/d)$ is what makes the calculation above three lines long, and it works only because a stress-free boundary imposes $\partial_z^2 w = 0$. A rigid boundary imposes $\partial_z w = 0$ instead — no slip — which the half-wave violates, and the vertical problem no longer separates into a single trigonometric mode: the sixth-order operator must be solved with six boundary conditions, the solution is a combination of one trigonometric and two hyperbolic vertical structures, and the eigenvalue condition is a transcendental equation whose lowest root gives the $\mathrm{Ra}_c \approx 1708$ quoted above [1199]. The physics is unchanged and the architecture of the argument is unchanged; only the vertical eigenfunction is no longer elementary. That is why the stress-free value is the one derived in full here and the rigid value is cited.

Rests on Equation (36.20).

36.4.4 Interpretation

The onset of convection is the case in which hydrodynamic stability theory works completely: a threshold and a length scale are computed from the linearized equations with no adjustable parameter, and both are observed. Set beside Section 36.3, this is the useful comparison — the failure to predict the pipe threshold is a failure of a particular flow to lose stability in the linear way, not a failure of the method.

Two honest qualifications are owed. The first concerns Bénard's own configuration. His layers were thin, with a free upper surface, and in that geometry the temperature dependence of the surface tension drives a competing instability with its own threshold and its own preferred scale; later analysis — not among the primary sources verified for this treatise — attributes the cells Bénard actually photographed largely to that mechanism rather than to buoyancy [1082, 1158]. The buoyancy instability that Equation (36.21) describes is properly tested in a layer confined between two plates, where the free surface is absent. The name attached to the phenomenon is therefore, strictly, attached to the wrong experiment, and it is recorded here so that a reader comparing the 1900 photographs with the 1916 calculation is not misled into thinking they are the same system.

The second qualification is that the calculation predicts only the onset. It gives the threshold and the initial pattern; it does not give the amplitude of the motion above threshold, the hexagonal rather than striped plan form, or anything about the sequence of further instabilities that follows as the heating is increased. That sequence — steady cells, then oscillations, then a cascade of period doublings into aperiodic motion — was measured in convecting liquid metal [1693] and in the closely analogous rotating-cylinder flow [505, 1545], and it is the experimental material of Chapter 32. Taylor's rotating cylinders deserve a separate mention as the first case in which a hydrodynamic stability calculation and an experiment agreed quantitatively [505].

36.4.5 Primary references

[288, 438].

36.5 Prandtl: the boundary layer made visible (1904)

Tests Equations (36.23) *and* (36.24).

Assuming Equation (36.9).

The sharpest wrong answer in classical physics is d'Alembert's: steady potential flow of an inviscid fluid past a finite body exerts no drag whatever [27], and the limit of vanishing viscosity therefore contradicts every observation of a real fluid however thin. Prandtl's resolution was not a calculation but a photograph, followed by a calculation. What he showed was that the limit is singular: the viscous term never becomes negligible near a wall, however small the viscosity, because the gradients there grow as fast as the coefficient shrinks.

36.5.1 Apparatus

A small hand-cranked water channel, of the order of a metre long, with a glass side wall for observation and a paddle arrangement returning the water so that a steady stream can be maintained by hand for as long as a run requires. The water is seeded with suspended particles and illuminated from the side, so that in a photograph of finite exposure each particle draws a streak along its path and the streamline pattern is recorded directly. Models are placed in the working section: a flat plate aligned with the stream, a circular cylinder, and a channel with diverging walls. One cylinder is pierced with a narrow slit connected to a suction line [315].

36.5.2 Procedure

The flow past each model is photographed at increasing stream speed, and the streak pattern examined for the line at which the streamlines cease to follow the surface. The diverging channel is used because the same effect occurs there without a curved body, in a geometry where the pressure is known to rise along the flow direction. The decisive run is the last: with the cylinder in place and the flow separated, suction is applied through the slit on one side only, removing the fluid nearest the wall along a single line, and the photograph is repeated.

36.5.3 Observations and data

Outside a thin region adjacent to the surface, the streamline pattern over the forward part of a body agrees with the ideal-flow solution; the fluid there behaves as the inviscid theory says it should. The region in which it does not is thin, and thins further as the speed is raised.

At a definite line on the surface — always in the part of the body where the external pressure is rising along the flow — the streamlines leave the wall. Downstream of that line the photograph shows fluid moving backwards near the surface, and beyond the body a broad wake in place of the smooth closing streamlines that potential theory requires. The same departure appears in the diverging channel, where no curvature is involved, showing that the rising pressure and not the geometry of the body is the operative condition.

With suction applied through the slit, the flow on that side remains attached past the point where it had separated, and the wake becomes strongly asymmetric. Removing a thin sheet of fluid — a small fraction of the discharge, taken from immediately against the wall — changes the entire flow around the body. Nothing else was altered.

Phenomenon 36.12 (The thin viscous layer and its separation)**.** At large Reynolds number the effect of viscosity in the flow past a body is confined to a layer against the surface whose thickness decreases as the Reynolds number rises, the flow outside it being described by

the inviscid equations. The layer detaches from the surface where the external pressure rises along the flow, leaving a wake of reversed flow behind the body; the detachment can be suppressed by removing the fluid within the layer, which demonstrates that the whole downstream flow is controlled by a sheet of fluid of negligible volume [315]. Drag is therefore not zero in the limit of vanishing viscosity, and the d'Alembert paradox is a statement about a limit that is not attained uniformly [27].

Rests on Equation (36.9).

Derivation. *Derives Phenomenon 36.12.*

Consider steady flow of speed U past a body of streamwise extent L, and suppose the viscous effects confined to a layer of thickness $\delta \ll L$. Within it the streamwise velocity must rise from zero at the wall to the external value over the distance δ, so transverse gradients exceed streamwise ones by the factor L/δ and the dominant viscous term is $\nu\,\partial_y^2 u$, of order $\nu U/\delta^2$. The streamwise inertia is $u\,\partial_x u$, of order U^2/L. A layer in which viscosity acts at all exists precisely where these balance:

$$\frac{\nu U}{\delta^2} \sim \frac{U^2}{L} \quad\Longrightarrow\quad \frac{\delta}{L} \sim \left(\frac{\nu}{UL}\right)^{1/2} = \mathrm{Re}^{-1/2}\,. \tag{36.23}$$

This settles the paradox as a matter of principle. The layer thins as $\mathrm{Re}^{-1/2}$ and so disappears in the inviscid limit, which is why the potential solution describes the outer flow ever better; but the velocity gradient across it grows as $U/\delta \sim \mathrm{Re}^{1/2}$, so the viscous stress $\mu\,U/\delta$ on the surface falls only as $\mathrm{Re}^{-1/2}$ while the area over which it acts is fixed. The dissipation does not vanish with the viscosity, because the region in which it occurs sharpens at exactly the compensating rate. A limit of that kind cannot be taken by setting the small parameter to zero in the equation, which is what d'Alembert's argument does.

The scaling is confirmed in detail by the one case that solves completely, the flat plate aligned with a uniform stream, for which the boundary-layer equations reduce to a single ordinary differential equation in the similarity variable $y\sqrt{U/\nu x}$ [341]. Its solution gives the thickness at which the velocity has reached 99 % of the free stream as $\delta_{99} \approx 5.0\,x\,\mathrm{Re}_x^{-1/2}$ and the local wall stress coefficient as $0.664\,\mathrm{Re}_x^{-1/2}$, whence a friction drag coefficient $1.328\,\mathrm{Re}_L^{-1/2}$ for one side of a plate of length L. The numbers are worth making concrete: a plate 1 m long in water of kinematic viscosity $1.0\times10^{-6}\,\mathrm{m^2/s}$ at 1 m/s has $\mathrm{Re}_L = 10^6$ and a layer 5 mm thick at the trailing edge — half a percent of the plate. That is the region which, when removed by suction, changes the flow everywhere. □

The scaling settles the paradox; the two further observations — where the layer leaves the wall, and how much drag it leaves behind — are settled by two arguments that need nothing from the layer equations beyond the thinness just established.

Proposition 36.13 (Separation requires a rising pressure)**.** *In steady flow along a solid wall the curvature of the velocity profile* at *the wall is fixed by the streamwise pressure gradient alone,*

$$\mu\left(\partial_y^2 u\right)_{y=0} = \frac{\mathrm{d}p}{\mathrm{d}x}\,, \tag{36.24}$$

and consequently the wall shear can vanish — which is what separation is — only at a station where $\mathrm{d}p/\mathrm{d}x > 0$*. Wherever the external pressure falls along the flow, the boundary layer stays attached.*

Rests on Equation (36.23).

Proof. *Derives Proposition 36.13.*

Evaluate the streamwise component of the steady incompressible Navier–Stokes equation on the wall itself. There $u = v = 0$, so both convective terms $u\,\partial_x u$ and $v\,\partial_y u$ vanish identically and what remains is

$$0 = -\frac{1}{\rho}\partial_x p + \nu\left(\partial_x^2 u + \partial_y^2 u\right)_{y=0}\,.$$

By Equation (36.23) the profile varies across the layer over the distance δ and along it over the distance L, so $\partial_y^2 u$ exceeds $\partial_x^2 u$ by the factor $(L/\delta)^2 \sim \mathrm{Re}$ and the streamwise derivative may be dropped; the same thinness makes the pressure variation across the layer negligible, so p is the external pressure and depends on x alone. That is Equation (36.24), and note that it is exact at the wall — no approximation has been made to the *form* of the equation, only to the relative size of two of its terms.

Now suppose the flow separates at some station, meaning $(\partial_y u)_{y=0} = 0$ there. Expanding the profile about the wall, where u and its first derivative both vanish,

$$u(y) = \tfrac{1}{2}\left(\partial_y^2 u\right)_{y=0} y^2 + O\left(y^3\right) ,$$

so the sign of the curvature at the wall is the sign of u just above it. The fluid a little way out from the wall is still moving forward, towards the external stream $U > 0$, so $\left(\partial_y^2 u\right)_{y=0} > 0$, and by Equation (36.24) $\mathrm{d}p/\mathrm{d}x > 0$ at that station. Contrapositively, no station at which the external pressure is falling can be a separation point. That is exactly what the photographs report: the streamlines leave the wall only over the rear of the body, and they do so in the diverging channel as well, where there is no curvature at all and the rising pressure is the only thing the two geometries share. □

Proposition 36.14 (The drag is the momentum deficit of the wake)**.** *For a body of unit span held in a uniform stream U of an incompressible fluid at constant pressure far from it, the streamwise force exerted on the body equals the momentum flux missing from the wake,*

$$F = \rho \int u\,(U-u)\,\mathrm{d}y = \rho U^2 \vartheta\,, \qquad \vartheta := \int \frac{u}{U}\left(1-\frac{u}{U}\right)\mathrm{d}y\,, \tag{36.25}$$

the integrals taken across a plane downstream of the body and ϑ being the momentum thickness*. A single traverse of the velocity profile behind a body therefore measures its drag, with no access to the stress at its surface.*

Rests on Equation (36.23)*.*

Proof. *Derives Proposition 36.14.*

Take a control volume of unit span bounded below by the surface, above by a plane $y = Y$ lying wholly outside the wake, upstream by a plane where the flow is still uniform at U, and downstream by the traverse plane. Steady mass conservation fixes what crosses the top: the volume entering upstream is UY per unit span and that leaving downstream is $\int_0^Y u\,\mathrm{d}y$, so the flux through the upper plane is $\rho\left(UY - \int_0^Y u\,\mathrm{d}y\right)$, outward. That fluid crosses the top with streamwise velocity U, because $y = Y$ is outside the wake.

Balance streamwise momentum. The pressure is uniform on every face by hypothesis and contributes nothing, so the only force is the reaction $-F$ of the body on the fluid, and

$$-F = \rho\int_0^Y u^2\,\mathrm{d}y + U\rho\left(UY - \int_0^Y u\,\mathrm{d}y\right) - \rho U^2 Y\,,$$

the three terms being the momentum leaving downstream, that leaving through the top, and that entering upstream. The terms in Y cancel identically — which is what makes the answer independent of where the upper plane was drawn — and

$$-F = \rho\int_0^Y u^2\,\mathrm{d}y - \rho U\int_0^Y u\,\mathrm{d}y = -\rho\int_0^Y u\,(U-u)\,\mathrm{d}y\,,$$

which is Equation (36.25), the integrand vanishing above the wake so that the upper limit may be sent to infinity. Two things are worth noting. The result is exact: it uses conservation of mass and momentum and the uniformity of the pressure, and nothing about viscosity, the layer or the shape of the body. And it makes the content of Phenomenon 36.12 quantitative — a drag that does not vanish with the viscosity is a wake that does not close, and Equation (36.25) is the identity between the two. □

The scaling argument of Equation (36.23) and the two exact identities just proved are as far as order-of-magnitude reasoning and control volumes can take the layer. The constants quoted with them — the 5.0, the 0.664 and the 1.328 — come from somewhere else, and it is worth stating as a proposition what that is.

Proposition 36.15 (Prandtl's layer equations and the Blasius similarity solution)**.** *At large Reynolds number the Navier–Stokes system for steady two-dimensional flow past a body reduces, in the layer, to Prandtl's equations*

$$u\,\partial_x u + v\,\partial_y u = -\frac{1}{\rho}\frac{\mathrm{d}p}{\mathrm{d}x} + \nu\,\partial_y^2 u\,, \qquad \partial_x u + \partial_y v = 0\,, \qquad \partial_y p = 0\,, \tag{36.26}$$

with u the streamwise and v the wall-normal velocity component, in which the pressure is not an unknown of the layer but is handed to it by the external potential flow. For a flat plate aligned with a uniform stream U the pressure gradient vanishes and Equation (36.26) admits the similarity solution $u = U f'(\eta)$ with $\eta = y\sqrt{U/\nu x}$, where f solves the Blasius equation

$$2f''' + f f'' = 0\,, \qquad f(0) = f'(0) = 0\,, \qquad f'(\infty) = 1\,, \tag{36.27}$$

whose numerical solution has $f''(0) = 0.3321$ and $f' = 0.99$ at $\eta = 5.0$ [341]. Those two numbers are what produce the thickness $\delta_{99} \approx 5.0\,x\,\mathrm{Re}_x^{-1/2}$, the local friction coefficient $2f''(0)\,\mathrm{Re}_x^{-1/2} = 0.664\,\mathrm{Re}_x^{-1/2}$ and, on integrating the latter along the plate, the coefficient $1.328\,\mathrm{Re}_L^{-1/2}$ quoted in Phenomenon 36.12.

Rests on Equations (36.9) *and* (36.23).

Full derivation: Appendix A.68 (p. 3167).

Derives Proposition 36.15.

The reduction is a matched asymptotic expansion, and the appendix does it as one: an outer expansion in which the viscous term is dropped and which cannot satisfy the no-slip condition, an inner expansion on the stretched variable $y\,\mathrm{Re}^{1/2}$ in which it cannot be dropped, and a matching condition in the overlap which is what transfers the outer pressure to the inner problem. Two things are worth knowing before reading it. The method of matched asymptotic expansions is not built anywhere in this treatise — Chapter 9 carries the WKB approximation, which is a different singular perturbation with the same flavour, and nothing more general — so the appendix states the matching principle it uses explicitly rather than appealing to a theorem of Part II. And Equation (36.27) has no solution in closed form: the two constants above are quadratures of a numerical integration, quoted from [341] and not computed here.

36.5.4 Interpretation

The experiment does three things at once, and they are usually run together to the detriment of all three.

It rescues the inviscid theory, by showing where it is valid: outside the layer, over the forward part of a body, the potential solution is what the streaks record. It falsifies the inviscid theory's global conclusion, by exhibiting the mechanism — separation — through which a vanishingly thin region imposes a finite drag on the whole flow. And it demonstrates control: the suction run is not an illustration but an intervention, and its success is the strongest evidence that the causal chain runs from the layer outwards and not the other way about.

The practical consequences run through the rest of fluid mechanics. The layer itself undergoes the same laminar-to-turbulent transition as the pipe of Section 36.3, and a turbulent layer, carrying more momentum near the wall, separates later than a laminar one; that is why the measured drag coefficient of a sphere *falls* abruptly as the Reynolds number

rises through the transition of its own boundary layer, and why a roughened sphere can be measurably less resistant than a smooth one [384]. The drag crisis is an experimental consequence of Phenomenon 36.12, and its explanation was Prandtl's.

36.5.5 Primary references

[315, 341].

36.6 What the record establishes, and what it does not

Table 36.2 collects the five results. Read as a whole, they divide sharply into those the equations of Chapter 31 explain and those they do not, and the division does not follow the historical order.

Table 36.2: The five experiments of this chapter, with the status of each result relative to the equations of Chapter 31. The last column is the point of the chapter: three of the five are derivations, two are constraints on a theory that does not yet exist.

Experiment	Regime	Status of the result
Efflux, to 1738	ideal flow	speed derived exactly; discharge derived exactly for the re-entrant mouthpiece by a momentum balance, and undetermined for the flush orifice, where it needs the shape of the free streamline
Poiseuille 1840	Re small	derived exactly; its accuracy measures the no-slip condition and bounds the slip length at a molecular scale
Reynolds 1883	$\mathrm{Re} \sim 10^3$	the collapse onto one parameter is derived; so is the negative result that no infinitesimal disturbance of the profile grows, and the algebraic amplification by Re that takes its place; the threshold value is not derived, and depends on the disturbance level
Bénard 1900	Ra near critical	threshold and cell scale derived in full from linear stability with no free parameter; the plan form and everything above onset are not
Prandtl 1904	Re large	the layer thickness scaling is derived and resolves the d'Alembert paradox; separation is shown to require a rising pressure, and the drag is identified with the wake's momentum deficit; the layer equations themselves are not derived

The honest summary of the turbulent state is that there is no theory of it. What exists is a set of measured regularities, an averaging procedure that reorganizes the equations

without closing them, and a similarity argument whose hypotheses are assumed rather than derived.

The averaging is Reynolds's own [268]: split each field into a mean and a fluctuation and average the equations, and the nonlinear term leaves behind the correlations of the fluctuations acting as additional stresses on the mean flow. Those correlations are new unknowns. Writing an equation for them produces triple correlations, and so on without end: the hierarchy of moment equations never closes, and every practical calculation of a turbulent flow closes it by an assumption that is fitted to data rather than derived. This is a structural feature of the nonlinearity, not a temporary gap.

The similarity argument is Kolmogorov's [857], with the spectral form due to Obukhov [861], resting on Richardson's picture of energy passed down a cascade of scales [494]. Granted that the small scales forget how the turbulence was made and, in an intermediate range, do not feel the viscosity either, dimensional analysis fixes the energy spectrum as a power law of exponent $-5/3$ with a single constant — a result derived in Chapter 31 and confirmed across a wide range of scales by the measurements of Grant, Stewart and Moilliet in a tidal channel, at a Reynolds number no laboratory of the time could reach [1227]. The confirmation is real and the hypotheses remain hypotheses; measured departures at high order — the intermittency corrections — show that they are not exactly true.

Underneath all of it sits an unresolved mathematical question. It is not known whether smooth solutions of the three-dimensional Navier–Stokes equations exist for all time from smooth initial data; what is proved is the existence of weak solutions [745], which are not known to be unique or smooth, and the gap is the substance of the Millennium problem statement [2094]. Every statement in this chapter about the turbulent regime therefore rests on equations whose well-posedness is open — a situation with no parallel elsewhere in this treatise, and one that Chapter 133 records as such.

The five results above are set down in that spirit. A future theory of turbulence, if there is one, will not be asked to predict something new before it has reproduced these: that the transition in a pipe is governed by one dimensionless number, that its threshold moves with the disturbance level and has no established upper bound, that the resistance law changes exponent there, that the flow past the transition is composed of eddies at scales comparable to the geometry rather than being structureless, and that its intermediate-scale energy spectrum follows the five-thirds law. The standard reference for the phenomenology as a whole is Batchelor's monograph [1340].

Part IV

Special Relativity

Chapter 37

Experiments: Light, the Aether, and Time

The kinematics of special relativity is not a philosophical preference; it is forced by measurement. This chapter sets out the experimental base: the aether the nineteenth century postulated as the medium of light, the conflict between that hypothesis and the principle of relativity, and the experiments whose outcomes single out the Lorentz transformations of Chapter 38 over Galilean kinematics.

37.1 Foundations

37.1.1 The vacuum and the aether

The aether hypothesis

Every wave known before the nineteenth century was a disturbance of something: sound of air, ripples of water, elastic waves of a solid. Each has a speed fixed with respect to its medium, and each obeys Galilean velocity addition — sound travels downwind faster than upwind, by exactly the wind speed. When Maxwell's equations (Chapter 60) yielded wave solutions travelling at a speed $c = 1/\sqrt{\varepsilon_0 \mu_0}$ fixed by two laboratory constants, the analogy seemed to settle the matter: light too must be a disturbance of a medium, the *luminiferous aether*, and c must be its speed relative to that medium.

The hypothesis was not idle. It made the aether's mechanical properties calculable and alarming: to carry a transverse wave at c it had to be more rigid than steel, yet offer no measurable resistance to the planets. That tension was tolerated because the alternative — a wave in nothing — seemed worse.

The conflict with the principle of relativity

The aether hypothesis contradicts Postulate 38.1. If light travels at c relative to the aether, then in a laboratory moving at velocity $\boldsymbol{v}$ through it the speed of light must be $c - \boldsymbol{v}$ in the direction of motion and $c + \boldsymbol{v}$ against it. Measuring that anisotropy would determine $\boldsymbol{v}$ by an experiment performed entirely inside the laboratory, and the aether's rest frame would be a physically preferred inertial frame — absolute space, restored by the back door.

So the aether is testable, and the test is quantitative. The Earth's orbital speed is $v_\oplus \approx 29.8\,\mathrm{km/s}$, so $\beta_\oplus \approx 10^{-4}$; even if the Solar System happened to be at rest in the aether at one moment, six months later the Earth's velocity would have reversed and the anisotropy would be $2\beta_\oplus$. A first-order effect is out of reach of an ordinary optical measurement, but a round-trip interference measurement is sensitive at order $\beta_\oplus^2 \approx 10^{-8}$ — and interferometry can resolve that. This is the experiment Michelson and Morley performed.

37.2 The finite speed of light

That light travels at all — rather than arriving instantaneously — is itself an experimental result, and the first quantitative one. Timing the eclipses of Jupiter's moon Io, Rømer found in 1676 that they run late when the Earth is receding from Jupiter and early when it approaches, by an amount consistent with a finite propagation speed [12]. He gave the light-crossing time of the Earth's orbit rather than a speed; with the modern orbital diameter it corresponds to a c within about 25 % of the value now fixed by definition. The relevant point for this part is that c is finite, so that the question of what it is finite *with respect to* can be asked at all.

37.3 Michelson and Morley: the search for the aether wind (1887)

Tests Equation (37.1).

37.3.1 Apparatus

A Michelson interferometer (Figure 37.1) mounted on a sandstone slab about 1.5 m square, floated on a bath of mercury so that the whole instrument could be rotated continuously without flexure or interruption. Multiple reflections folded each arm to an optical path of about 11 m, which the authors quote as 2×10^7 wavelengths of yellow light. A sodium lamp supplied the source; the fringes were viewed through a telescope carried around with the slab.

37.3.2 Procedure

Monochromatic light is divided at a half-silvered plate into two beams travelling along perpendicular arms, retroreflected, and recombined to interfere. If the apparatus moves at speed v through an aether, the round-trip times along the two arms differ, and rotating the instrument by 90° exchanges the arms and reverses the difference. The observable is therefore not a fringe position — which nothing calibrates — but the *shift* of the fringes as the instrument turns, repeated at different times of day and at different seasons so that the Earth's rotation and orbital motion sweep the laboratory velocity through the sky.

37.3.3 Observations and data

For arms of equal length L, expansion of the two round-trip times to second order in $\beta = v/c$ gives a time difference $\Delta t \simeq L\beta^2/c$, hence an expected fringe shift on rotation of

$$\Delta N \simeq \frac{2L}{\lambda}\beta^2 \,, \tag{37.1}$$

which for $L = 11\,\mathrm{m}$, $\lambda = 590\,\mathrm{nm}$ and $\beta = \beta_\oplus$ is about 0.4 of a fringe — some forty times the resolution of the instrument.

The observed displacement was, in the authors' words, "certainly less than the twentieth part of this, and probably less than the fortieth part":

$$\Delta N_{\mathrm{obs}} < 0.02 \text{ fringe (certainly)}\,, \qquad < 0.01 \text{ fringe (probably)}\,. \tag{37.2}$$

Translated into a velocity through Equation (37.1), they concluded that "the relative velocity of the earth and the ether is probably less than one sixth the earth's orbital velocity, and certainly less than one-fourth." The paper states the bound only as that fraction and prints no figure in kilometres per second; with $v_\oplus = 29.8\,\mathrm{km/s}$ the two limits correspond to

about 5 km/s and 7.5 km/s, but those numbers are derived here and are not Michelson and Morley's.

37.3.4 Interpretation

The predicted effect was absent. Contraction hypotheses of the FitzGerald–Lorentz type could be arranged to cancel it, at the cost of a new postulate whose only function was to hide the aether; Kennedy and Thorndike later closed that escape with an unequal-arm version of the experiment, whose null result constrains the velocity dependence that length contraction alone cannot explain away [700]. The economical reading is the one Einstein took: there is no aether wind because there is no aether frame, and the speed of light is the same in every inertial frame — Postulate 38.2.

37.3.5 Primary references

[236]; the unequal-arm follow-up [700].

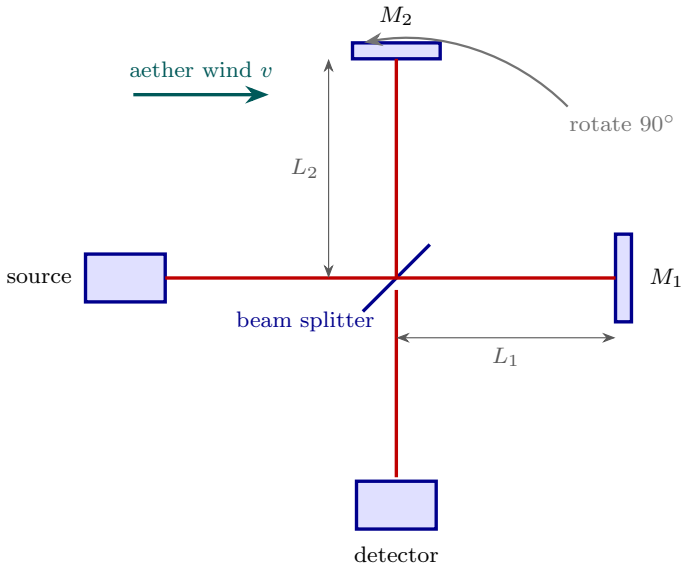


Figure 37.1: The Michelson interferometer of Section 37.3. Light is divided at the beam splitter, travels the two perpendicular arms of lengths L_1 and L_2, is retroreflected at M_1 and M_2, and recombines at the detector. An aether wind of speed v would make the two round-trip times differ at order $(v/c)^2$; rotating the whole instrument through 90° exchanges the arms and doubles the predicted fringe shift.

37.4 Fizeau: light dragged by moving water (1851)

Tests Phenomenon 38.18.

37.4.1 Apparatus

Two parallel tubes, each 1.487 m long with an interior bore of 5.3 mm, through which water was driven at 7.069 m/s. A beam of light was split so that one part traversed both tubes with the current and the other against it, the two being recombined to interfere.

37.4.2 Procedure

The water flow is started and stopped and the fringe displacement recorded. Sending each beam through both tubes — once with the flow and once against — doubles the effect and cancels systematic differences between the two optical paths.

37.4.3 Observations and data

Fizeau reports fringe displacements only. From nineteen sufficiently concurring observations the average was

$$\Delta N_{\text{simple}} = 0.23\,, \qquad \Delta N_{\text{double}} = 0.46\,, \tag{37.3}$$

against the 0.40 predicted by Fresnel's partial-drag hypothesis. No uncertainty is quoted anywhere in the note, and no numerical drag coefficient is printed; the values often attributed to Fizeau are back-derivations by later authors.

37.4.4 Interpretation

Complete dragging of the aether by the water would give a displacement proportional to the full water speed; no dragging at all would give none. The measurement sits between the two, near Fresnel's coefficient $1 - 1/n^2$, which is 0.434 for water at $n = 1.333$. As an aether result this is incoherent — the medium must be dragged partially, by an amount depending on the colour of the light. As a relativistic result it is immediate: the composition law Equation (38.29) applied to $u' = c/n$ gives exactly $c/n + v\,(1 - 1/n^2)$ to first order, with no free parameter and no medium to drag. The derivation is Example 38.20.

37.4.5 Primary references

[133]; the relativistic reading is Einstein's [321].

37.5 Cavity resonators: the modern Michelson–Morley (1979–2015)

Tests Equation (37.1).
Assuming Postulate 38.2.

37.5.1 Apparatus

Two orthogonal electromagnetic resonators — laser-stabilized optical cavities [1635, 2147], or cryogenic sapphire microwave oscillators [2258] — mounted on a turntable and rotated continuously, with their beat frequency recorded.

37.5.2 Procedure

A cavity's resonant frequency depends on its optical length; an anisotropy in the speed of light would make the beat frequency of two perpendicular cavities vary as the pair rotates, at twice the rotation frequency and with sidereal modulation as the Earth turns. The measurement is the amplitude of that modulation, integrated over months.

37.5.3 Observations and data

The searched-for fractional anisotropy is bounded at

$$\frac{\Delta c}{c} \lesssim 10^{-18}\,, \tag{37.4}$$

with the rotating optical cavities of [2147] reaching the 10^{-17} level and the rotating cryogenic microwave experiment of [2258] the 10^{-18} level in the corresponding Lorentz-violating coefficients.

37.5.4 Interpretation

Michelson and Morley bounded the anisotropy near 10^{-8} in β^2; a century of refinement has improved that by ten orders of magnitude and found nothing. Postulate 38.2 is, on present evidence, one of the most stringently tested statements in physics. The modern experiments are usually reported not as bounds on an aether velocity — nobody expects one — but as bounds on the coefficients of a general Lorentz-violating extension of the Standard Model, which is the form in which a null result remains informative (Chapter 133).

37.5.5 Primary references

[1635, 2147, 2258].

37.6 The muon lifetime

The experiments above concern the propagation of light. That the *time* of a moving system is affected as well is shown most directly by unstable particles, whose decay is a clock carried along with them. Muons created by cosmic-ray primaries at an altitude of some 15 km have a proper lifetime near 2.2 µs, so even at the speed of light they should travel only about 660 m before decaying, and essentially none should reach the ground. Very many do.

This is the subject of Chapter 41, where the measurements of Rossi and Hall [863], Frisch and Smith [1253] and the CERN storage ring [1592] are reported in full, once the time-dilation formula Equation (38.27) they test has been derived.

37.7 What the experiments establish

Taken together the results above fix the kinematics uniquely. The Michelson–Morley null result and its modern descendants say the two-way speed of light is isotropic to 10^{-18}, so no preferred frame is detectable optically. Fizeau's measurement says velocities do not add in the Galilean way, and fixes the first-order correction. Kennedy and Thorndike say the effect cannot be removed by contraction alone. The muon experiments say that moving clocks run slow by the factor γ, not by some other function of speed.

What survives is a kinematics in which c is invariant and the transformation between inertial frames is not Galilean. Chapter 38 shows that these two requirements have essentially one solution.

Chapter 38

Lorentz Transformations

A Lorentz transformation relates the space-time coordinates that two inertial frames assign to one and the same event. This chapter derives them from the two postulates of Chapter 37, reads their consequences off the geometry of Minkowski space, and characterizes the most general such transformation as a boost composed with a spatial rotation. The founding paper is Einstein's [321]; the geometric reading followed from Minkowski [357], and the group structure was identified by Poincaré [329]. Standard modern treatments are [1585, 1976].

38.1 The postulates

Everything in this part follows from two statements, both of them experimental claims rather than mathematical ones.

Postulate 38.1 (Principle of relativity)**.** The laws of physics take the same form in every inertial frame. No experiment performed within a freely moving laboratory can determine its velocity.

Postulate 38.2 (Invariance of the speed of light)**.** Light propagates in vacuum with the same speed c in every inertial frame, independently of the motion of its source.

Postulate 38.1 is old — it is Galileo's, and Newtonian mechanics obeys it. Postulate 38.2 is the new one, and it is what the aether experiments of Chapter 37 force upon us: the Michelson–Morley null result (Section 37.3) says that the speed of light does not depend on the motion of the apparatus, and the Fizeau measurement (Section 37.4) says that it does not add to the speed of its medium in the Galilean way. The two postulates are incompatible with Galilean kinematics, in which speeds simply add, and the whole of this chapter is the working out of what must replace it.

Since 1983 the second postulate has been built into the system of units: c is a defined constant, exactly $299\,792\,458\,\mathrm{m/s}$, and the metre is the distance light travels in $1/c$ seconds (Chapter 2). Measuring c is therefore no longer possible; what the experiments now test is the *isotropy* and *source-independence* asserted above, and those are still measured, to the precision reported in Section 37.5.

Notation 38.3 (Standard configuration). Throughout, K and $\bar{K}$ are inertial frames whose origins coincide at $t = \bar{t} = 0$, whose axes are parallel, and whose relative velocity is v along the common x-axis. We write

$$\beta := \frac{v}{c}\,, \qquad \gamma := \frac{1}{\sqrt{1-\beta^2}}\,, \tag{38.1}$$

so that $\gamma \geq 1$, with $\gamma = 1$ only for $\beta = 0$, and $\gamma \to \infty$ as $\beta \to 1$. Coordinates are collected as $x^\mu = (x^0, x^i) = (ct, \boldsymbol{x})$, with Greek indices running over $0, 1, 2, 3$ and Latin over $1, 2, 3$.

38.2 Lorentz boosts

38.2.1 The one-dimensional boost

Theorem 38.4 (Lorentz boost)**.** *In the standard configuration of Notation 38.3, the two postulates force the coordinate transformation*

$$\bar{x} = \gamma\,(x - vt)\ , \tag{38.2}$$

$$\bar{t} = \gamma\left(t - \frac{v}{c^2}x\right)\ , \tag{38.3}$$

$$\bar{y} = y\,, \qquad \bar{z} = z\,. \tag{38.4}$$

Rests on Postulates 38.1 and 38.2 and Notation 38.3.

Proof. *Derives Theorem 38.4.*

Step 1: the transformation is linear. Space and time are homogeneous — no event and no place is privileged — so a uniform velocity in K must appear as a uniform velocity in $\bar{K}$: straight worldlines go to straight worldlines. A map carrying straight lines to straight lines and fixing the coincident origins is linear. Write, for the x and t pair,

$$\bar{x} = Ax + Bt\,, \qquad \bar{t} = Dx + Et\,, \tag{38.5}$$

with coefficients depending on v alone.

Step 2: the origin of $\bar{K}$. The point $\bar{x} = 0$ is by construction the origin of $\bar{K}$, which moves in K along $x = vt$. Substituting into Equation (38.5) gives $Avt + Bt = 0$ for all t, whence $B = -Av$ and

$$\bar{x} = A\,(x - vt)\ . \tag{38.6}$$

From Equation (38.5) (imposing that $\bar{x} = 0$ be the worldline $x = vt$).

Step 3: the invariance of c. Emit a light pulse from the common origin at $t = \bar{t} = 0$ in the $+x$ direction. In K it obeys $x = ct$, in $\bar{K}$ it must obey $\bar{x} = c\bar{t}$ by Postulate 38.2. From Equations (38.5) and (38.6),

$$\bar{x} = A\,(c - v)\,t\,, \qquad \bar{t} = (Dc + E)\,t\,,$$

so $\bar{x} = c\bar{t}$ requires

$$A\,(c - v) = c\,(Dc + E)\ . \tag{38.7}$$

A pulse sent in the $-x$ direction obeys $x = -ct$ and $\bar{x} = -c\bar{t}$, and the same substitution gives

$$A\,(c + v) = c\,(E - Dc)\ . \tag{38.8}$$

Adding Equations (38.7) and (38.8) yields $2Ac = 2cE$, hence $E = A$; subtracting them yields $2Av = -2Dc^2$, hence $D = -Av/c^2$. So far

$$\bar{x} = A\,(x - vt)\ , \qquad \bar{t} = A\left(t - \frac{v}{c^2}x\right)\ . \tag{38.9}$$

From Equations (38.7) and (38.8) (by addition and subtraction).

Step 4: the principle of relativity fixes A. By Postulate 38.1 neither frame is privileged, so expressing (x, t) in terms of $(\bar{x}, \bar{t})$ must reproduce Equation (38.9) with v replaced by $-v$ — from $\bar{K}$, the frame K recedes at $-v$. Moreover A can depend only on $|v|$: space is isotropic, so reflecting the x-axis, which sends $v \to -v$, cannot change a scale factor. Hence $A(-v) = A(v)$, and

$$x = A\,(\bar{x} + v\bar{t})\ , \qquad t = A\left(\bar{t} + \frac{v}{c^2}\bar{x}\right)\ .$$

Substituting Equation (38.9) into the first of these,

$$\begin{aligned} x &= A\left[A\left(x-vt\right)+vA\left(t-\frac{v}{c^2}x\right)\right] \\ &= A^2\left[x-vt+vt-\frac{v^2}{c^2}x\right] = A^2\left(1-\beta^2\right)x\,, \end{aligned}$$

so $A^2(1-\beta^2)=1$. The root is fixed by requiring the transformation to reduce to the identity at $v=0$ rather than to a reflection, so $A=+\left(1-\beta^2\right)^{-1/2}=\gamma$. This is Equations (38.2) and (38.3).

Step 5: the transverse coordinates. Consider a rod of proper length ℓ lying along the y-axis, at rest in $\bar{K}$, and suppose $\bar{y}=\kappa(v)y$ with $\kappa\neq 1$. By isotropy κ depends only on $|v|$, and by the principle of relativity the inverse relation is $y=\kappa(v)\bar{y}$; composing gives $\kappa^2=1$, so $\kappa=1$ (again excluding the reflection). Lengths perpendicular to the motion are unchanged, which is Equation (38.4). □

Setting $c\to\infty$ in Equations (38.2) and (38.3) returns $\bar{x}=x-vt$ and $\bar{t}=t$: the Galilean transformation is the limit in which light is infinitely fast. Every departure from Newtonian kinematics in this part is controlled by the single dimensionless number β, and at $\beta=10^{-6}$ — a satellite in low orbit — $\gamma-1$ is about 5×10^{-13}, which is why the Galilean form survived two centuries of testing.

In matrix form, with $x^\mu=(ct,x,y,z)$,

$$\bar{x}^\mu=\Lambda^\mu{}_\nu x^\nu\,,\qquad [\Lambda]=\begin{pmatrix}\gamma & -\gamma\beta & 0 & 0\\ -\gamma\beta & \gamma & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}. \tag{38.10}$$

From Equations (38.2) to (38.4) (collecting the three coordinate relations into one matrix).

38.2.2 Minkowski space

In a Euclidean space, rotations — linear homogeneous transformations — leave the norm invariant, and with it the Euclidean metric. A boost is also linear and homogeneous and resembles a rotation formally, though not physically.[1] It is natural to ask which metric a boost preserves — equivalently, since the metric defines the space, in which space a boost is a rigid motion.

Proposition 38.5 (The invariant interval)**.** *Under the boost Equations (38.2) to (38.4),*

$$c^2\bar{t}^2-\bar{x}^2-\bar{y}^2-\bar{z}^2=c^2t^2-x^2-y^2-z^2\,. \tag{38.11}$$

Rests on Theorem 38.4 and Equation (38.1).

Proof. *Derives Proposition 38.5.*

The transverse terms are unchanged. For the other two,

$$\begin{aligned} c^2\bar{t}^2-\bar{x}^2 &= \gamma^2\left(ct-\beta x\right)^2-\gamma^2\left(x-\beta ct\right)^2 \\ &= \gamma^2\left[c^2t^2-2\beta ctx+\beta^2x^2-x^2+2\beta ctx-\beta^2c^2t^2\right] \\ &= \gamma^2\left(1-\beta^2\right)\left(c^2t^2-x^2\right)=c^2t^2-x^2\,, \end{aligned}$$

the last step by Equation (38.1). □

[1] As Section 39.2.4 shows, a Euclidean rotation and a Lorentz boost are even topologically distinct: the rotation group is compact and the Lorentz group is not.

Definition 38.6 (Minkowski metric and interval)**.** *Minkowski space* is $\mathbb{R}^4$ equipped with the symmetric bilinear form of signature $(1,3)$

$$[\eta_{\mu\nu}] = \operatorname{diag}(1,-1,-1,-1)\ , \tag{38.12}$$

and the *invariant interval* between two events separated by Δx^μ is

$$\Delta s^2 := \eta_{\mu\nu}\Delta x^\mu \Delta x^\nu = c^2\Delta t^2 - |\Delta \boldsymbol{x}|^2\ . \tag{38.13}$$

Rests on Proposition 38.5.

Proposition 38.5 says precisely that a boost preserves Equation (38.13). Written with Λ, the condition that a linear transformation preserve η is

$$\eta_{\mu\nu}\Lambda^\mu{}_\lambda \Lambda^\nu{}_\rho = \eta_{\lambda\rho}\,, \qquad \text{that is} \qquad \Lambda^{\mathsf{T}}\eta\,\Lambda = \eta\,, \tag{38.14}$$

the exact analogue of the orthogonality condition $O^{\mathsf{T}}O = \mathbb{1}$ that defines a Euclidean rotation, with $\mathbb{1}$ replaced by η. Equation (38.14) is taken as the *definition* of a Lorentz transformation in Section 39.2.1, and the whole of Chapter 39 is the study of the group it defines.

Remark 38.7 (Signature conventions). This treatise uses $(+,-,-,-)$, so that $\Delta s^2 > 0$ for timelike separations and proper time is real. The opposite convention $(-,+,+,+)$ is equally common — Appendix A.1 uses it, following the gravitational-wave literature — and flips the sign of every interval. Nothing physical depends on the choice; every observable is a ratio or a sign comparison in which the convention cancels.

The light cone

The sign of Δs^2 is invariant, so all observers agree on it. It partitions the events of Minkowski space, relative to any chosen event O, into the regions of Figure 38.1.

Definition 38.8 (Causal classification)**.** A separation Δx^μ between two events is *timelike* if $\Delta s^2 > 0$, *null* or *lightlike* if $\Delta s^2 = 0$, and *spacelike* if $\Delta s^2 < 0$.

Rests on Definition 38.6.

Proposition 38.9 (Causal order is absolute)**.** *If two events are timelike or null separated, every inertial observer agrees which came first. If they are spacelike separated, their time order depends on the frame, and there are frames in which they are simultaneous.*

Rests on Equation (38.3) *and Definition 38.8.*

Proof. *Derives Proposition 38.9.*

Take the events at $\Delta x^\mu = (c\Delta t, \Delta x, 0, 0)$ with $\Delta t > 0$. By Equation (38.3), $c\Delta\bar{t} = \gamma\,(c\Delta t - \beta\Delta x)$, which changes sign when β passes through $c\Delta t/\Delta x$. This value is an admissible velocity, $|\beta| < 1$, exactly when $|\Delta x| > c\,|\Delta t|$, that is when $\Delta s^2 < 0$. For timelike and null separations no admissible boost reverses the order; for spacelike separations the boost with $\beta = c\Delta t/\Delta x$ makes $\Delta\bar{t} = 0$ and any larger β reverses the order. □

This is the reason the invariant speed also bounds causal influence. If a signal could travel faster than c, it would connect spacelike-separated events; by Proposition 38.9 some observer would see it arrive before it was sent, and the notion of cause would not survive a change of frame. Nothing in the postulates forbids superluminal *coordinate* speeds that carry no signal — the intersection point of a pair of closing scissor blades is a standard example — and nothing in them permits superluminal signalling.

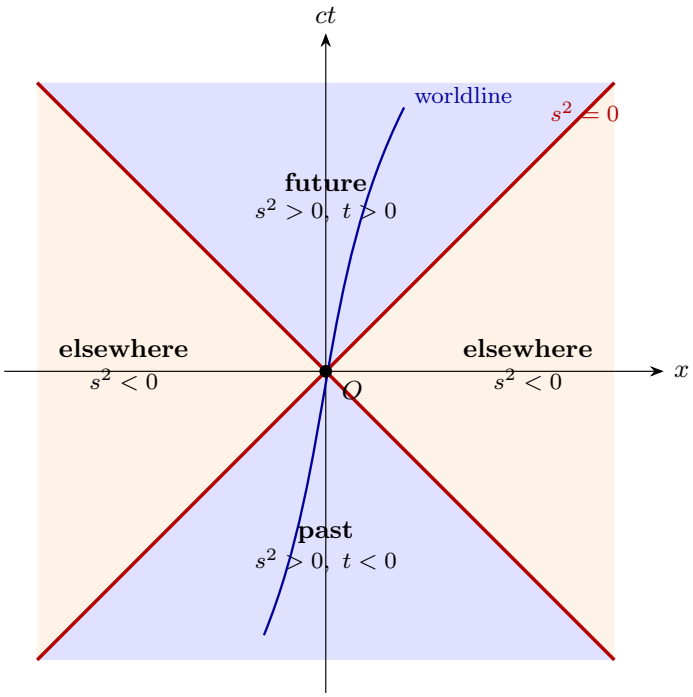


Figure 38.1: The causal structure of Minkowski space, drawn with one spatial dimension suppressed. Light through the event O travels on the 45° lines $x = \pm ct$, where $\Delta s^2 = 0$. Events with $\Delta s^2 > 0$ are timelike separated from O and lie inside the cone, divided into a future and a past that every observer agrees upon; events with $\Delta s^2 < 0$ are spacelike separated and lie outside it. A massive particle's worldline stays everywhere inside the cone, since its speed never reaches c.

38.2.3 The general boost

For a relative velocity $\boldsymbol{v} = c\boldsymbol{\beta}$ in an arbitrary direction, resolve the position into components along and across the motion,

$$\boldsymbol{x}_{\parallel} = \frac{(\boldsymbol{x}\cdot\boldsymbol{\beta})}{|\boldsymbol{\beta}|^2}\boldsymbol{\beta}\,, \qquad \boldsymbol{x}_{\perp} = \boldsymbol{x} - \boldsymbol{x}_{\parallel}\,. \tag{38.15}$$

Theorem 38.4 applies to the longitudinal pair and leaves the transverse part alone, so

$$c\bar{t} = \gamma\,(ct - \boldsymbol{\beta}\cdot\boldsymbol{x})\ , \tag{38.16}$$

$$\bar{\boldsymbol{x}}_{\parallel} = \gamma\left(\boldsymbol{x}_{\parallel} - \boldsymbol{\beta}\,ct\right)\ , \tag{38.17}$$

$$\bar{\boldsymbol{x}}_{\perp} = \boldsymbol{x}_{\perp}\,. \tag{38.18}$$

Recombining $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_{\parallel} + \bar{\boldsymbol{x}}_{\perp}$ and using Equation (38.15),

$$\bar{\boldsymbol{x}} = \boldsymbol{x} + \frac{\gamma-1}{|\boldsymbol{\beta}|^2}\,(\boldsymbol{\beta}\cdot\boldsymbol{x})\,\boldsymbol{\beta} - \gamma\boldsymbol{\beta}\,ct\,. \tag{38.19}$$

From Equations (38.15), (38.17) and (38.18) (recombining the longitudinal and transverse parts).

Reading off the components of $\bar{x}^{\mu} = \Lambda^{\mu}{}_{\nu}(\beta)x^{\nu}$ gives

$$\Lambda^{0}{}_{0}(\beta) = \gamma\,, \tag{38.20}$$

$$\Lambda^{0}{}_{i}(\beta) = -\gamma\beta_i\,, \tag{38.21}$$

$$\Lambda^{i}{}_{0}(\beta) = -\gamma\beta^i\,, \tag{38.22}$$

$$\Lambda^{i}{}_{j}(\beta) = \delta^{i}{}_{j} + \frac{\gamma-1}{|\boldsymbol{\beta}|^2}\beta^i\beta_j\,, \tag{38.23}$$

or, in matrix form,

$$[\Lambda(\beta)] = \begin{pmatrix} \gamma & -\gamma\beta_1 & -\gamma\beta_2 & -\gamma\beta_3 \\ -\gamma\beta^1 & 1+\frac{\gamma-1}{|\boldsymbol{\beta}|^2}\beta^1\beta_1 & \frac{\gamma-1}{|\boldsymbol{\beta}|^2}\beta^1\beta_2 & \frac{\gamma-1}{|\boldsymbol{\beta}|^2}\beta^1\beta_3 \\ -\gamma\beta^2 & \frac{\gamma-1}{|\boldsymbol{\beta}|^2}\beta^2\beta_1 & 1+\frac{\gamma-1}{|\boldsymbol{\beta}|^2}\beta^2\beta_2 & \frac{\gamma-1}{|\boldsymbol{\beta}|^2}\beta^2\beta_3 \\ -\gamma\beta^3 & \frac{\gamma-1}{|\boldsymbol{\beta}|^2}\beta^3\beta_1 & \frac{\gamma-1}{|\boldsymbol{\beta}|^2}\beta^3\beta_2 & 1+\frac{\gamma-1}{|\boldsymbol{\beta}|^2}\beta^3\beta_3 \end{pmatrix}. \tag{38.24}$$

From Equations (38.20) to (38.23) (assembling the components into a matrix).

Taking $\boldsymbol{\beta} = (\beta, 0, 0)$ recovers Equation (38.10), since then $\gamma - 1$ multiplies only the 11 entry and $1 + (\gamma - 1) = \gamma$.

38.3 Spacetime diagrams

The consequences of Theorem 38.4 are algebraically trivial and conceptually treacherous; they become obvious when drawn. A *spacetime diagram* plots ct upward against x rightward, so that light travels on the 45° lines and a particle's history is a curve — its *worldline* — everywhere steeper than 45°.

The boosted frame is drawn in the same picture rather than on a separate sheet. Its time axis is the set $\bar{x} = 0$, which by Equation (38.2) is the line $x = vt$; its space axis is the set $\bar{t} = 0$, which by Equation (38.3) is the line $ct = \beta x$. Both therefore make the same angle $\alpha = \arctan\beta$ with the corresponding unprimed axis, but on opposite sides, as in Figure 38.2: the axes close like a pair of scissors on the light line, and reach it together only as $\beta \to 1$.

One warning must be issued before the diagram is used to read off numbers. The page is Euclidean and the geometry it depicts is not, so distances measured with a ruler on the paper are meaningless. What is meaningful is the invariant Equation (38.13), whose level sets are the hyperbolae of Figure 38.3. Those hyperbolae are the orbits of the boost — a boost slides events along them exactly as a Euclidean rotation slides points along circles — and they, not the ruler, calibrate the axes.

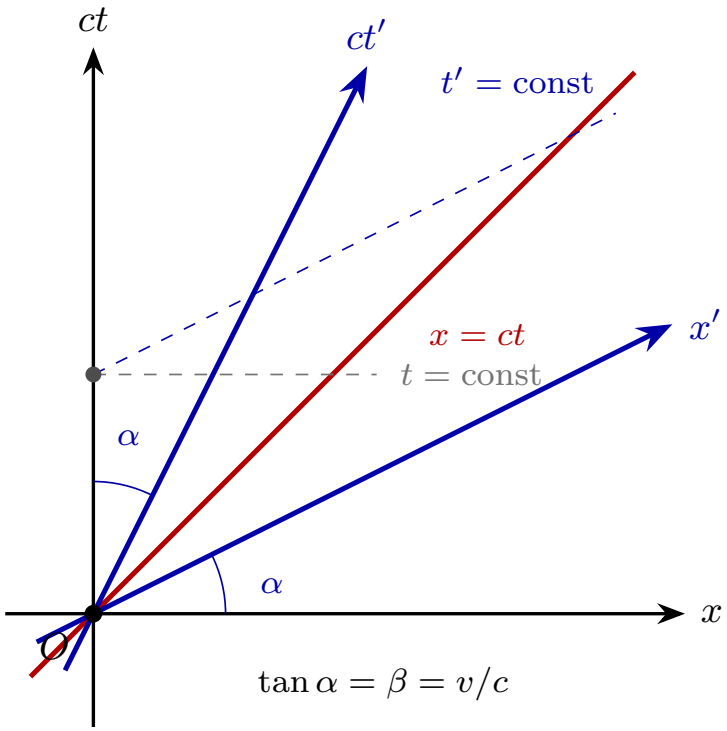


Figure 38.2: The axes of a boosted frame drawn inside the diagram of the unboosted one, for $\beta = 0.5$. The $c\bar{t}$-axis is the worldline of the moving origin and the $\bar{x}$-axis is the locus of events the moving observer calls simultaneous; both make the angle $\alpha = \arctan\beta$ with their unprimed counterparts, on opposite sides of the light line $x = ct$. Because the two axes tilt *towards* the light line rather than rotating rigidly, the surfaces of constant time in the two frames are genuinely different families.

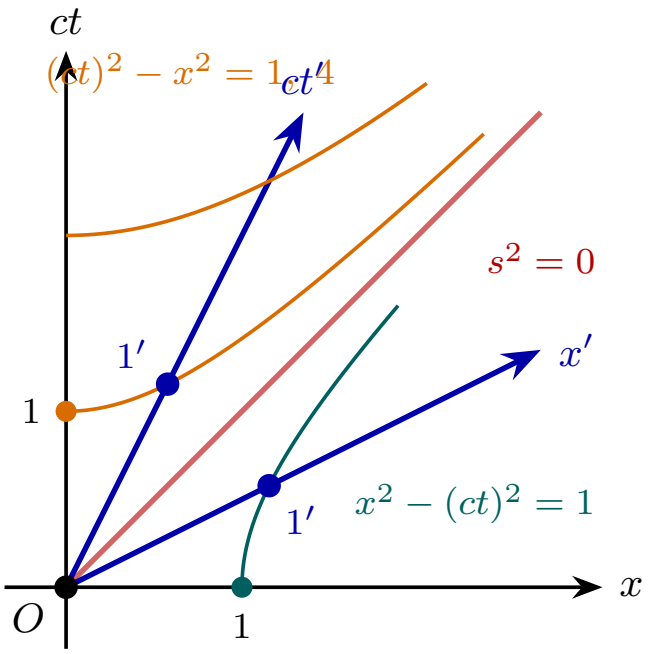


Figure 38.3: Calibration. The hyperbolae $(ct)^2 - x^2 = \pm 1$ are invariant under boosts, so a boost slides events along them, and the unit tick on any frame's time axis is where that axis cuts the timelike hyperbola. The primed ticks are farther from the origin *on the page* than the unprimed ones, and represent the same invariant interval: this is why tick spacings in the two frames may never be compared by eye.

38.4 Consequences of the postulates

38.4.1 Relativity of simultaneity

Phenomenon 38.10 (Relativity of simultaneity)**.** Two spatially separated events that are simultaneous in one inertial frame are not simultaneous in another moving relative to it. For events separated by Δx and simultaneous in K,

$$\Delta \bar{t} = -\gamma \frac{v}{c^2} \Delta x \,. \tag{38.25}$$

Rests on Theorem 38.4 and Equation (38.3).

Derivation. Derives Phenomenon 38.10.

Apply Equation (38.3) to the separation of the two events. With $\Delta t = 0$,

$$\Delta \bar{t} = \gamma \left(\Delta t - \frac{v}{c^2} \Delta x \right) = -\gamma \frac{v}{c^2} \Delta x \,,$$

which vanishes for all Δx only if $v = 0$. □

Figure 38.4 shows why this is not a paradox. "The set of events happening now" is a slice through spacetime, and the two frames slice at different angles. Neither slice is correct; simultaneity is simply not a property of a pair of events, but of a pair of events *and* a frame. Note that Equation (38.25) grows with the separation Δx: disagreement about simultaneity is unmeasurable in a laboratory and enormous over astronomical distances.

Remark 38.11 (Simultaneity is a convention, within limits). That distant simultaneity must be *defined* rather than discovered was Einstein's opening move [321]: he stipulated that light

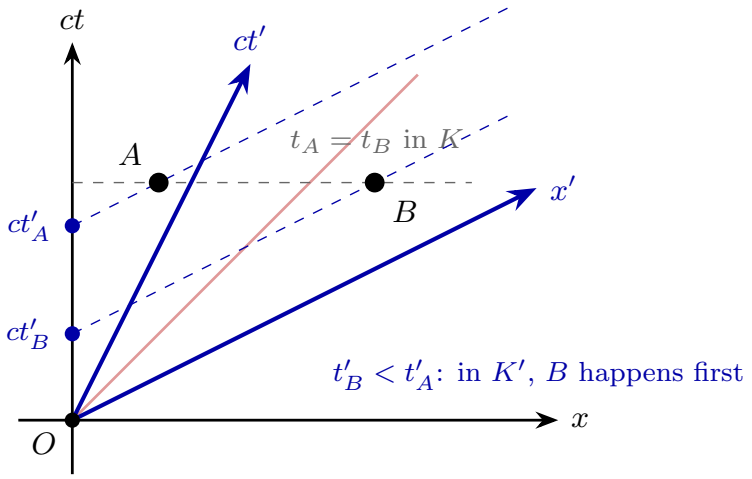


Figure 38.4: Relativity of simultaneity. The events A and B lie on one horizontal line, so K assigns them the same t. The surfaces of constant $\bar{t}$ are tilted, so A and B lie on different ones and $\bar{K}$ assigns them different times — here B happens first. Nothing has moved between the two readings; only the family of slices called "the same instant" has been re-chosen.

takes equal times on the outward and return legs of a round trip, which is what synchronizes separated clocks. The one-way speed of light is not independently measurable — measuring it presupposes synchronized clocks at the two ends — so this is a coordinative definition in the sense of Section 1.3.2, not an experimental result. What *is* measured, and what Chapter 37 reports, is the two-way speed and the isotropy of the round trip.

38.4.2 Time dilation

Definition 38.12 (Proper time)**.** The *proper time* τ along a worldline is the time read by a clock carried along it, defined through the invariant interval by

$$c^2 \mathrm{d}\tau^2 := \mathrm{d}s^2 = c^2 \mathrm{d}t^2 - |\mathrm{d}\boldsymbol{x}|^2 \,, \qquad \text{so} \qquad \mathrm{d}\tau = \frac{\mathrm{d}t}{\gamma} \,. \tag{38.26}$$

Being built from $\mathrm{d}s^2$, proper time is invariant: every frame agrees on what a given clock reads.

Rests on Definition 38.6 and Proposition 38.5.

Phenomenon 38.13 (Time dilation)**.** A clock moving uniformly with speed v relative to a frame K runs slow as judged in K: an interval $\Delta\tau$ of its proper time corresponds to

$$\Delta t = \gamma \Delta\tau > \Delta\tau \tag{38.27}$$

of K-coordinate time [816, 863].

Rests on Theorem 38.4, Equation (38.3), and Definition 38.12.

Derivation. *Derives Phenomenon 38.13.*

The clock is at rest in $\bar{K}$, so the two ticks occur at the same $\bar{x}$ and are separated by $\Delta\bar{t} = \Delta\tau$. Inverting Equation (38.3) — equivalently, applying the boost with $-v$ —

$$\Delta t = \gamma \left(\Delta\bar{t} + \frac{v}{c^2} \Delta\bar{x} \right) = \gamma \Delta\bar{t} = \gamma \Delta\tau \,,$$

since $\Delta\bar{x} = 0$. The same result follows from Equation (38.26) directly. □

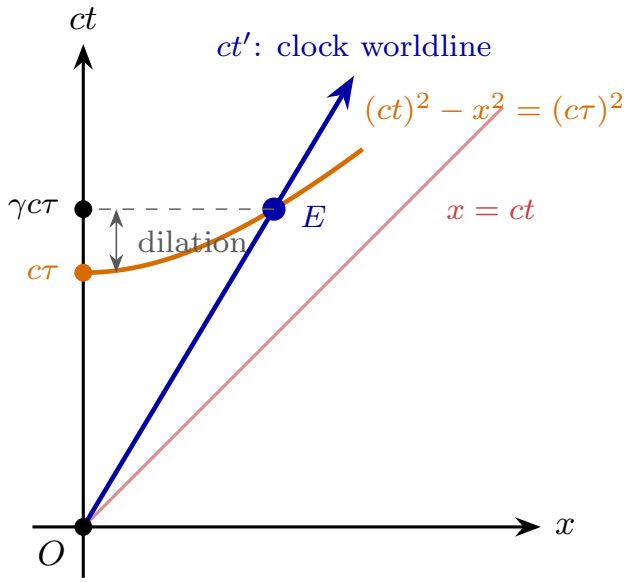


Figure 38.5: Time dilation. The moving clock sits at $\bar{x} = 0$, so its worldline is the $c\bar{t}$-axis; at the event E it has ticked off a proper time τ. The calibration hyperbola through E marks every event lying a proper time τ from O, in any frame, and it is what makes the comparison meaningful. The height of E on the ct-axis is $\gamma c\tau$, larger than $c\tau$.

Remark 38.14 (The effect is reciprocal, and that is consistent). $\bar{K}$ judges K's clocks to run slow by the same factor. The two statements do not contradict each other because each concerns a *different* pair of events: comparing a single moving clock with two stationary ones requires those two to be synchronized, and Phenomenon 38.10 says the frames disagree about that synchronization by exactly the amount needed. An asymmetry appears only when one clock changes frames, which is not a boost; that is the twin paradox, resolved in Section 40.6.

Rests on Phenomena 38.10 and 38.13.

38.4.3 Length contraction

Definition 38.15 (Proper length)**.** The *proper length* L_0 of a body is its length in the frame in which it is at rest. Its length in any other frame is defined as the distance between its two ends *taken at the same time in that frame*; the qualification is essential, because by Phenomenon 38.10 the frames disagree about what it means.

Rests on Phenomenon 38.10.

Phenomenon 38.16 (Length contraction)**.** A body of proper length L_0 moving with speed v along its own length has, in the frame through which it moves, the length

$$L = \frac{L_0}{\gamma} < L_0 \,. \tag{38.28}$$

Dimensions perpendicular to the motion are unaffected.

Rests on Definition 38.15 and Equations (38.2) and (38.4).

Derivation. *Derives Phenomenon 38.16.*

Let the body be at rest in $\bar{K}$ with its ends at $\bar{x}_1$ and $\bar{x}_2$, so $L_0 = \bar{x}_2 - \bar{x}_1$. In K the ends are located simultaneously, at some common t. Applying Equation (38.2) to each end and subtracting,

$$L_0 = \bar{x}_2 - \bar{x}_1 = \gamma\,[(x_2 - x_1) - v\,(t - t)] = \gamma L \,,$$

which is Equation (38.28). The transverse statement is Equation (38.4). □

Figure 38.6 makes the asymmetry with time dilation plain. The body sweeps out a strip in spacetime; the two frames measure its width along different directions, and the strip is what it is. The factor appears in the numerator for lengths and the denominator for times because one measurement is taken along a line of constant t and the other along a worldline.

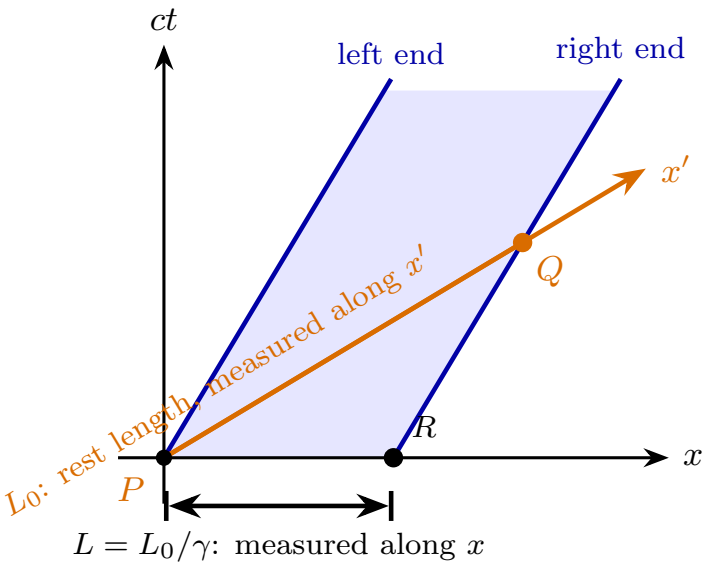


Figure 38.6: Length contraction. A rod at rest in $\bar{K}$ sweeps out the shaded strip between the worldlines of its ends. Its rest length L_0 is the interval PQ measured along the $\bar{x}$-axis, between events $\bar{K}$ calls simultaneous. K instead measures PR along its own line of constant t and obtains L_0/γ. The rod did not shrink; the two observers measured the width of the same strip in different directions.

Remark 38.17 (Contraction is not what a camera records). Phenomenon 38.16 concerns simultaneous positions, not appearances. Light reaching an eye or a plate at one instant left different parts of the body at different times, and when that is accounted for a rapidly moving object looks *rotated* rather than flattened — the Terrell–Penrose effect. Nothing in Equation (38.28) changes; the difference lies in which set of events the measurement collects.

38.5 Composition of velocities

Phenomenon 38.18 (Relativistic composition of velocities)**.** Let a particle have velocity components $u^i = \mathrm{d}x^i/\mathrm{d}t$ in K, and let $\bar{K}$ move with speed v along x. Then in $\bar{K}$ [133]

$$\bar{u}^x = \frac{u^x - v}{1 - u^x v/c^2}\,, \tag{38.29}$$

$$\bar{u}^y = \frac{u^y}{\gamma\,(1 - u^x v/c^2)}\,, \qquad \bar{u}^z = \frac{u^z}{\gamma\,(1 - u^x v/c^2)}\,. \tag{38.30}$$

Rests on Theorem 38.4 and Equations (38.2) to (38.4).

Derivation. *Derives Phenomenon 38.18.*

Differentiate Equations (38.2) to (38.4):

$$\mathrm{d}\bar{x} = \gamma\,(\mathrm{d}x - v\mathrm{d}t)\,, \qquad \mathrm{d}\bar{t} = \gamma\left(\mathrm{d}t - \frac{v}{c^2}\mathrm{d}x\right), \qquad \mathrm{d}\bar{y} = \mathrm{d}y\,.$$

Dividing the first by the second, and cancelling γ and $\mathrm{d}t$,

$$\bar{u}^x = \frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}} = \frac{\mathrm{d}x - v\mathrm{d}t}{\mathrm{d}t - (v/c^2)\,\mathrm{d}x} = \frac{u^x - v}{1 - u^x v/c^2}\,,$$

which is Equation (38.29). Dividing the third by the second keeps one factor of γ and gives Equation (38.30); note that a transverse velocity changes even though a transverse *length* does not, because the time interval is transformed. □

Corollary 38.19 (The speed of light is a fixed point)**.** *If $u^x = c$ then $\bar{u}^x = c$ for every $|v| < c$; and if $|u^x| < c$ and $|v| < c$ then $|\bar{u}^x| < c$.*

Rests on Equation (38.29).

Proof. *Derives Corollary 38.19.*

For the first, substitute $u^x = c$ into Equation (38.29):

$$\bar{u}^x = \frac{c - v}{1 - v/c} = \frac{c\,(1 - v/c)}{1 - v/c} = c\,.$$

For the second, put $u^x = c\beta_1$ and $v = c\beta_2$ with $|\beta_i| < 1$. Then

$$1 - \frac{\bar{u}^x}{c} = 1 - \frac{\beta_1 - \beta_2}{1 - \beta_1\beta_2} = \frac{(1 - \beta_1)\,(1 + \beta_2)}{1 - \beta_1\beta_2} > 0\,,$$

and symmetrically $1 + \bar{u}^x/c > 0$, so $|\bar{u}^x| < c$. □

Corollary 38.19 is the resolution of the puzzle that motivated the whole construction: no matter how fast a source or an observer moves, light still passes them at c, and no composition of subluminal velocities ever reaches c. In the Galilean limit $c \to \infty$ the denominators become 1 and Equation (38.29) returns $\bar{u}^x = u^x - v$.

Example 38.20 (Fizeau's experiment as a first-order effect)**.** Light travelling through water of refractive index n moving at speed v has, in the water's rest frame, the speed $u' = c/n$. In the laboratory, Equation (38.29) gives

$$u = \frac{c/n + v}{1 + v/(nc)} \simeq \frac{c}{n} + v\left(1 - \frac{1}{n^2}\right) + O\!\left(\frac{v^2}{c}\right),$$

expanding to first order in v/c. The coefficient $1 - 1/n^2$ is precisely Fresnel's drag coefficient, which Fizeau measured in 1851 (Section 37.4) and which had stood for half a century as an unexplained empirical rule. It is a first-order consequence of Equation (38.29), requiring no aether and no partial dragging of one.

Rests on Equation (38.29).

38.5.1 Rapidity

The composition law Equation (38.29) is not addition, which makes repeated boosts awkward. One change of variable repairs this.

Definition 38.21 (Rapidity)**.** The *rapidity* φ of a boost is defined by

$$\beta = \tanh\varphi\,, \qquad \text{whence} \qquad \gamma = \cosh\varphi\,, \qquad \gamma\beta = \sinh\varphi\,. \tag{38.31}$$

Rests on Equation (38.1).

The identity $\cosh^2\varphi - \sinh^2\varphi = 1$ is exactly $\gamma^2 - \gamma^2\beta^2 = 1$, so the parametrization is consistent, and φ runs over all of $\mathbb{R}$ as β runs over $(-1, 1)$. Substituting into Equation (38.10), the boost becomes

$$[\Lambda] = \begin{pmatrix} \cosh\varphi & -\sinh\varphi \\ -\sinh\varphi & \cosh\varphi \end{pmatrix} \tag{38.32}$$

in the (ct, x) block — a *hyperbolic rotation*. It is the exact analogue of the circular rotation

$$[R(\theta)] = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

with the trigonometric functions replaced by hyperbolic ones. This is the precise sense in which a boost is a rotation in Minkowski space, and it is why the orbits in Figure 38.3 are hyperbolae rather than circles.

Proposition 38.22 (Rapidities add)**.** *Two collinear boosts of rapidities φ_1 and φ_2 compose into a single boost of rapidity $\varphi_1 + \varphi_2$.*

Rests on Definition 38.21 and Equation (38.32).

Proof. *Derives Proposition 38.22.*

Multiplying two matrices of the form Equation (38.32) and using the hyperbolic addition formulas,

$$\begin{pmatrix} \cosh\varphi_1 & -\sinh\varphi_1 \\ -\sinh\varphi_1 & \cosh\varphi_1 \end{pmatrix} \begin{pmatrix} \cosh\varphi_2 & -\sinh\varphi_2 \\ -\sinh\varphi_2 & \cosh\varphi_2 \end{pmatrix} = \begin{pmatrix} \cosh(\varphi_1+\varphi_2) & -\sinh(\varphi_1+\varphi_2) \\ -\sinh(\varphi_1+\varphi_2) & \cosh(\varphi_1+\varphi_2) \end{pmatrix}.$$

Applying tanh to the sum reproduces Equation (38.29), since

$$\tanh(\varphi_1+\varphi_2) = \frac{\tanh\varphi_1 + \tanh\varphi_2}{1 + \tanh\varphi_1\tanh\varphi_2}.$$

□

Rapidity is therefore the natural parameter: it is additive, unbounded, and identifies the collinear boosts as a one-parameter group isomorphic to $(\mathbb{R}, +)$. That φ ranges over an unbounded set while β is confined to $(-1, 1)$ is the analytic face of the non-compactness proved in Section 39.2.1.

38.6 General Lorentz transformations

38.6.1 Boosts composed with rotations

In the most general case two coordinate systems have a constant relative velocity *and* axes that are not parallel. The coordinates with respect to K and $\bar{K}$ are then related by a transformation Λ containing not only the boosts $\bar{x}^\mu = \Lambda^\mu{}_\nu(\beta)x^\nu$ but also three-dimensional rotations.

For a spatial rotation the coordinate transformation is linear and homogeneous, $\bar{x}^i = R^i{}_j(\theta)x^j$ with $R(\theta) \in \mathrm{SO}(3)$. In the general case we must therefore consider

$$\bar{x}^\mu = \Lambda^\mu{}_\nu(\beta, \theta)x^\nu, \tag{38.33}$$

where $\Lambda(0, \theta)$ is a spatial rotation and $\Lambda(\beta, 0)$ is the boost Equation (38.24).

From Equation (38.33),

$$\bar{x}^0 = \Lambda^0{}_0(0,\theta)x^0 + \Lambda^0{}_i(0,\theta)x^i, \tag{38.34}$$

$$\bar{x}^i = \Lambda^i{}_0(0,\theta)x^0 + \Lambda^i{}_j(0,\theta)x^j, \tag{38.35}$$

and in a spatial rotation the time coordinate is left invariant, $\bar{x}^0 = x^0$. Hence $\Lambda^0{}_0(0,\theta) = 1$ and $\Lambda^0{}_i(0,\theta) = 0$; and for the spatial part to be a rotation, $\Lambda^i{}_0(0,\theta) = 0$ and $\Lambda^i{}_j(0,\theta) = R^i{}_j(\theta)$. In summary the general case is characterized by

$$\begin{aligned} &\Lambda^0{}_0(\beta,0) = \gamma && \Lambda^0{}_0(0,\theta) = 1 \\ &\Lambda^0{}_i(\beta,0) = -\gamma\beta_i && \Lambda^0{}_i(0,\theta) = 0 \\ &\Lambda^i{}_0(\beta,0) = -\gamma\beta^i && \Lambda^i{}_0(0,\theta) = 0 \\ &\Lambda^i{}_j(\beta,0) = \delta^i{}_j + \tfrac{\gamma-1}{|\boldsymbol{\beta}|^2}\beta^i\beta_j && \Lambda^i{}_j(0,\theta) = R^i{}_j(\theta)\,. \end{aligned}$$

Theorem 38.23 (Polar decomposition of a Lorentz transformation)**.** *Every proper orthochronous Lorentz transformation (Definition 39.5) factors uniquely as*

$$\Lambda = \Lambda\,(\boldsymbol{\beta},0)\,R\,(\theta)\ , \tag{38.36}$$

a pure boost followed by a spatial rotation.

Rests on Definition 39.5, Equations (38.14), (38.20) and (38.22), and Proposition 39.6.

Proof. *Derives Theorem 38.23.*

Existence. The first column $\Lambda^\mu{}_0$ is the image of the time direction. Taking $\mu = \nu = 0$ in Equation (38.14) gives $(\Lambda^0{}_0)^2 - \sum_i (\Lambda^i{}_0)^2 = 1$, so that column is a unit timelike vector, and it is future-directed because $\Lambda^0{}_0 \geq 1$ for an orthochronous transformation.

Let $B := \Lambda(\boldsymbol{\beta},0)$ be the pure boost with

$$\gamma := \Lambda^0{}_0\,, \qquad \beta^i := -\frac{\Lambda^i{}_0}{\Lambda^0{}_0}\,, \tag{38.37}$$

which is admissible since $|\boldsymbol{\beta}| < 1$ follows from the unit condition, and which by Equations (38.20) and (38.22) has exactly the same first column as Λ.

Put $R := B^{-1}\Lambda$, a product of Lorentz transformations and hence Lorentz. Its first column is $R^\mu{}_0 = (B^{-1})^\mu{}_\nu B^\nu{}_0 = \delta^\mu{}_0$, so $R^0{}_0 = 1$ and $R^i{}_0 = 0$. Applying the row identity of Proposition 39.6, $(R^0{}_0)^2 = 1 + \sum_i (R^0{}_i)^2$, gives $\sum_i (R^0{}_i)^2 = 0$ and hence $R^0{}_i = 0$. So $R = \operatorname{diag}\left(1, \tilde{R}\right)$ with $\tilde{R}\tilde{R}^{\mathsf{T}} = \mathbb{1}$, that is $\tilde{R} \in \mathrm{O}(3)$; and since $\det B = 1$ — directly from Equation (38.10), $\gamma^2 - \gamma^2\beta^2 = 1$ — while $\det\Lambda = 1$, also $\det\tilde{R} = 1$ and $\tilde{R} \in \mathrm{SO}(3)$. Then $\Lambda = BR$ is the required factorization.

Uniqueness. Suppose $\Lambda = B'R'$ with B' a pure boost and R' a rotation. A rotation fixes the time direction, $R'^\nu{}_0 = \delta^\nu{}_0$, so $\Lambda^\mu{}_0 = B'^\mu{}_\nu R'^\nu{}_0 = B'^\mu{}_0$: the boost B' has the same first column as Λ, and a pure boost is determined by that column through Equation (38.37). Hence $B' = B$ and therefore $R' = B^{-1}\Lambda = R$. □

The counting is $6 = 3 + 3$: three parameters for the boost velocity and three for the rotation, matching the dimension of the Lorentz group computed in Section 39.2.7.

Remark 38.24 (Boosts alone do not form a group)*.* The decomposition Equation (38.36) is not a formality. The product of two boosts in *different* directions is not a boost: it is a boost composed with a nontrivial rotation, the *Thomas rotation*. Boosts are therefore not closed under composition and do not form a subgroup, whereas rotations do. The physical consequence — the Thomas precession of a spin carried around a closed path — is measured in atomic fine structure, where it supplies the factor of $\frac{1}{2}$ in the spin–orbit coupling (Chapter 81).

Rests on Theorem 38.23.

38.6.2 Directions normal to the relative velocity

Equation (38.4) states that lengths perpendicular to the boost are unchanged, but it does not follow that *directions* are unchanged: an angle mixes a transverse with a longitudinal component, and only one of them transforms.

Proposition 38.25 (Aberration of light)**.** *Let a light ray travel in K at an angle ϑ to the x-axis. In $\bar{K}$ its angle $\bar{\vartheta}$ satisfies*

$$\cos\bar{\vartheta} = \frac{\cos\vartheta - \beta}{1 - \beta\cos\vartheta}\,. \tag{38.38}$$

Rests on Equation (38.29) and Corollary 38.19.

Proof. *Derives Proposition 38.25.*

The ray has $u^x = c\cos\vartheta$ and $u^y = c\sin\vartheta$. Substituting into Equation (38.29) and dividing by c,

$$\cos\bar{\vartheta} = \frac{\bar{u}^x}{c} = \frac{c\cos\vartheta - v}{c\,(1 - \beta\cos\vartheta)} = \frac{\cos\vartheta - \beta}{1 - \beta\cos\vartheta}\,,$$

where the speed in $\bar{K}$ is again c by Corollary 38.19. □

Aberration is why the apparent positions of the fixed stars trace small ellipses over a year as the Earth's velocity swings through $\pm 29.8\,\mathrm{km/s}$: the effect was discovered by Bradley in 1728, long before its explanation, and its first-order part $\Delta\vartheta \simeq \beta\sin\vartheta$ is Galilean. The relativistic content of Equation (38.38) lies in the denominator, at second order. At $\beta \to 1$ the formula concentrates almost all directions into a narrow forward cone of half-angle $\sim 1/\gamma$ — the relativistic beaming that makes synchrotron radiation a forward searchlight (Chapter 65).

Chapter 39

Minkowski Space and Its Symmetries

The Lorentz transformations of Chapter 38 are the symmetries of Minkowski space, and this chapter assembles them into groups: the Lorentz group of linear transformations preserving the Minkowski metric, its distinguished subsets, and the Poincaré group obtained by adjoining space-time translations, together with the Lie algebras of both. Standard accounts are found in [1746, 2132]. The source also reserves a chapter for the conserved quantities associated with these symmetries; that heading is retained below as pending.

39.1 Conserved quantities

The reason to assemble the Lorentz transformations into a group is that symmetries and conservation laws are the same thing seen twice. Noether's theorem (Chapter 16) attaches to every one-parameter group of symmetries of an action a quantity conserved along the solutions, and the ten parameters of the Poincaré group of Section 39.3 deliver the ten conservation laws that organize all of mechanics:

1. the four **translations** of spacetime give conservation of the four-momentum p^μ of Definition 40.4 — energy from the time translation, the three momenta from the spatial ones;
2. the three **rotations** give conservation of angular momentum;
3. the three **boosts** give conservation of $\boldsymbol{K} = t\boldsymbol{p} - E\boldsymbol{x}/c^2$, which is the statement that the centre of energy moves uniformly.

The last is unfamiliar only because it is rarely named; it is the relativistic form of the elementary fact that the centre of mass of an isolated system does not accelerate. All ten are components of two tensors, p^μ and the antisymmetric $J^{\mu\nu}$, whose conservation is the content of Equations (39.18) to (39.20). Which quantities are conserved is therefore fixed by the symmetry group of spacetime, and that is the subject of this chapter.

39.2 The Lorentz group

39.2.1 Foundations

Definition 39.1 (Lorentz group)**.** The set of all Lorentz transformations (those that leave the Minkowski metric invariant) is given by

$$L = \left\{ \Lambda \in \mathrm{GL}(4,\mathbb{R}) \mid \Lambda\eta\Lambda^{\mathsf{T}} = \eta \right\} . \tag{39.1}$$

Rests on Definition 38.6.

Since η is a symmetric bilinear form of signature $(1,3)$, the set Equation (39.1) is by definition the orthogonal group of that form, $L = \mathrm{O}(3,1)$.

Proposition 39.2 (L is a Lie group)**.** *L is a group under matrix multiplication and an embedded submanifold of $\mathrm{GL}(4,\mathbb{R})$ of dimension six.*

Rests on Definition 39.1 and Equation (39.2).

Proof. *Derives Proposition 39.2.*

Group. If $\Lambda_1, \Lambda_2 \in L$ then $(\Lambda_1\Lambda_2)^{\mathsf{T}}\eta\Lambda_1\Lambda_2 = \Lambda_2^{\mathsf{T}}\left(\Lambda_1^{\mathsf{T}}\eta\Lambda_1\right)\Lambda_2 = \Lambda_2^{\mathsf{T}}\eta\Lambda_2 = \eta$, so L is closed; $\mathbb{1} \in L$; and taking determinants in Equation (39.2) gives $(\det\Lambda)^2 = 1$, so every $\Lambda \in L$ is invertible, and inverting the defining relation shows $\Lambda^{-1} \in L$.

Manifold. Consider $F : \mathrm{GL}(4,\mathbb{R}) \longrightarrow \mathrm{Sym}(4)$ defined by $F(\Lambda) = \Lambda^{\mathsf{T}}\eta\Lambda - \eta$, whose target is the ten-dimensional space of symmetric matrices. Then $L = F^{-1}(0)$, and the derivative of F at $\Lambda \in L$ sends X to $X^{\mathsf{T}}\eta\Lambda + \Lambda^{\mathsf{T}}\eta X$, which is surjective onto the symmetric matrices — given symmetric S, take $X = \frac{1}{2}\eta^{-1}\left(\Lambda^{\mathsf{T}}\right)^{-1}S$. Since 0 is a regular value, the preimage theorem makes L an embedded submanifold of dimension $16 - 10 = 6$, matching Proposition 39.9. Multiplication and inversion are rational in the entries, hence smooth, so L is a Lie group (Chapter 14). □

Non-compactness of the Lorentz group

The condition satisfied by the elements of the Lorentz group is

$$\Lambda\eta\Lambda^{\mathsf{T}} = \eta \iff \eta_{\lambda\rho}\Lambda^{\lambda}{}_{\mu}\Lambda^{\rho}{}_{\nu} = \eta_{\mu\nu}\,, \tag{39.2}$$

whence, taking the $\mu = \nu = 0$ component and using $\eta_{00} = 1$ and $\eta_{ij} = -\delta_{ij}$,

$$\begin{aligned}
\eta_{\lambda\rho}\Lambda^{\lambda}{}_{0}\Lambda^{\rho}{}_{0} &= \eta_{00}\,,\\
\eta_{00}\Lambda^{0}{}_{0}\Lambda^{0}{}_{0} + \eta_{ij}\Lambda^{i}{}_{0}\Lambda^{j}{}_{0} &= 1\,,\\
\left(\Lambda^{0}{}_{0}\right)^2 - \delta_{ij}\Lambda^{i}{}_{0}\Lambda^{j}{}_{0} &= 1\,,\\
\left(\Lambda^{0}{}_{0}\right)^2 &= 1 + \Lambda^{i}{}_{0}\Lambda^{i}{}_{0}\,,
\end{aligned}$$

from which we see that $\left(\Lambda^{0}{}_{0}\right)^2 \geq 1$ and therefore

$$\left|\Lambda^{0}{}_{0}(\beta,\theta)\right| \geq 1\,. \tag{39.3}$$

From Equation (39.2) (the $\mu = \nu = 0$ component of the defining relation).

The result obtained in Equation (39.3) tells us that, as a manifold, the Lorentz group is not compact.

Derivation. A compact subset of $\mathbb{R}^{16}$ is bounded. The boosts Equation (38.32) form a curve in L parametrized by the rapidity $\varphi \in \mathbb{R}$, along which $\Lambda^{0}{}_{0} = \cosh\varphi \to \infty$ as $\varphi \to \infty$. The entries of Λ are therefore unbounded on L, so L is not compact — and neither is $L^{\uparrow}_{+}$, which contains that whole curve.

The contrast with the rotation group is the point. $\mathrm{SO}(3)$ is compact: its entries are sines and cosines, bounded by 1. Replacing the angle by a rapidity replaces the bounded circular functions by unbounded hyperbolic ones, which is the same statement as the sign difference in the last bracket of Equation (39.15). The consequence matters in Chapter 94: a non-compact group has no finite-dimensional unitary representations except the trivial one, which is why the unitary representations describing particles are infinite dimensional, and why the finite-dimensional representations that classify fields — scalars, spinors, vectors — are not unitary. □

39.2.2 The proper Lorentz group

Definition 39.3 (Proper Lorentz group)**.** The *proper Lorentz group* is

$$L_+ = \{\Lambda \in L \mid \det[\Lambda] = 1\} \,. \tag{39.4}$$

Rests on Definition 39.1.

Evidently we have $L_+ = \mathrm{SO}(3,1)$.

39.2.3 The improper Lorentz group

Taking the determinant of Equation (39.2) gives $(\det\Lambda)^2 \det\eta = \det\eta$, so $\det\Lambda = \pm 1$ and the transformations with $\det\Lambda = -1$ form the complement of L_+ in L.

Definition 39.4 (Improper Lorentz transformations)**.** $L_- := \{\Lambda \in L \mid \det[\Lambda] = -1\}$.
Rests on Definition 39.1 and Equation (39.2).

L_- is not a subgroup — it does not contain the identity, and the product of two of its elements has determinant $+1$ — but it is a coset of L_+, and every element of it is a proper transformation composed with one fixed reflection. The canonical choice is *parity*, the spatial reflection

$$[P] = \mathrm{diag}\,(1,-1,-1,-1) \,, \qquad \det P = -1 \,, \qquad P \in L_-^{\uparrow} \,, \tag{39.5}$$

which reverses the three spatial axes and leaves time alone. That P satisfies Equation (39.2) is immediate, since $[P] = [\eta]$ and $\eta\eta\eta = \eta$.

Parity is a symmetry of gravitation, electromagnetism and the strong interaction, and is *not* a symmetry of the weak interaction: that discovery (Chapter 110) is the reason $L_+^{\uparrow}$ rather than L is the group of physical relevance in particle physics.

39.2.4 The orthochronous Lorentz group

By Equation (39.3) every Lorentz transformation has $|\Lambda^0{}_0| \geq 1$, so the sign of $\Lambda^0{}_0$ splits L in two.

Definition 39.5 (Orthochronous Lorentz group)**.**

$$L^{\uparrow} = \left\{\Lambda \in L \mid \Lambda^0{}_0 \geq 1\right\} \,, \qquad L_+^{\uparrow} = L^{\uparrow} \cap L_+ \,. \tag{39.6}$$

$L_+^{\uparrow}$ is the *proper orthochronous* (or *restricted*) Lorentz group.
Rests on Definition 39.1 and Equation (39.3).

Proposition 39.6 ($L^{\uparrow}$ and $L_+^{\uparrow}$ are subgroups)**.** *$L^{\uparrow}$ is closed under products and inverses, hence so is $L_+^{\uparrow}$. Moreover the two defining conditions, $\det\Lambda = 1$ and $\Lambda^0{}_0 \geq 1$, are open and closed on L, so $L_+^{\uparrow}$ contains every connected subset of L through the identity.*
Rests on Definition 39.5 and Equations (39.2) and (39.3).

Proof. *Derives Proposition 39.6.*
The $\mu = \nu = 0$ component of Equation (39.2) is the column identity $(\Lambda^0{}_0)^2 = 1 + \sum_i \Lambda^i{}_0 \Lambda^i{}_0$ (the computation leading to Equation (39.3)); applied to $\Lambda^{-1} = \eta\Lambda^{\mathsf{T}}\eta \in L$ it becomes the row identity $(\Lambda^0{}_0)^2 = 1 + \sum_i \Lambda^0{}_i \Lambda^0{}_i$. For $\Lambda, \Lambda' \in L^{\uparrow}$,

$$(\Lambda\Lambda')^0{}_0 = \Lambda^0{}_0\, \Lambda'^0{}_0 + \sum_i \Lambda^0{}_i\, \Lambda'^i{}_0 \,,$$

and the Cauchy–Schwarz inequality together with the two identities bounds the cross term strictly:

$$|\textstyle\sum_i \Lambda^0{}_i\, \Lambda'^i{}_0| \le \sqrt{(\Lambda^0{}_0)^2 - 1}\,\sqrt{(\Lambda'^0{}_0)^2 - 1} < \Lambda^0{}_0\, \Lambda'^0{}_0\,,$$

so $(\Lambda\Lambda')^0{}_0 > 0$, hence ≥ 1 by Equation (39.3). For the inverse, $(\Lambda^{-1})^0{}_0 = (\eta\Lambda^{\mathsf{T}}\eta)^0{}_0 = \Lambda^0{}_0 \ge 1$. Clopenness: det and $\Lambda \mapsto \Lambda^0{}_0$ are continuous, $\det = \pm 1$ on L and $|\Lambda^0{}_0| \ge 1$, so each condition singles out a union of connected components; a connected set containing the identity (which satisfies both) therefore stays in $L^{\uparrow}_{+}$. □

That $L^{\uparrow}_{+}$ is itself connected — and is therefore precisely the identity component of L — is proved in Chapter 94, as the continuous image of the connected covering group $\mathrm{SL}(2,\mathbb{C})$.

39.2.5 The antichronous Lorentz group

Definition 39.7 (Antichronous transformations)**.**

$$L^{\downarrow} := \left\{ \Lambda \in L \mid \Lambda^0{}_0 \le -1 \right\}, \tag{39.7}$$

the complement of $L^{\uparrow}$ in L by Equation (39.3).

Rests on Definition 39.5 and Equation (39.3).

An antichronous transformation reverses the direction of time: it maps the future light cone of Figure 38.1 onto the past one. Like L_- it is a coset rather than a subgroup, and its canonical representative is *time reversal*

$$[T] = \mathrm{diag}\,(-1,1,1,1)\,, \qquad \det T = -1\,, \qquad T \in L^{\downarrow}_{-}\,. \tag{39.8}$$

39.2.6 The four components

The two independent signs — $\det\Lambda = \pm 1$ from Definition 39.4 and $\operatorname{sgn}\Lambda^0{}_0 = \pm 1$ from Equation (39.3) — are continuous functions taking discrete values, so neither can change along a path inside L. They therefore partition the group into four pieces that no continuous motion connects.

Proposition 39.8 (Component structure)**.** *L has exactly four connected components,*

$$L^{\uparrow}_{+}\,, \qquad L^{\uparrow}_{-} = P\,L^{\uparrow}_{+}\,, \qquad L^{\downarrow}_{-} = T\,L^{\uparrow}_{+}\,, \qquad L^{\downarrow}_{+} = PT\,L^{\uparrow}_{+}\,, \tag{39.9}$$

and the quotient of L by the identity component is the discrete group

$$\frac{L}{L^{\uparrow}_{+}} \cong \{\mathbb{1}, P, T, PT\} \cong \mathbb{Z}_2 \times \mathbb{Z}_2\,, \tag{39.10}$$

the Klein four-group.

Rests on Definitions 39.4 and 39.5, Equations (39.5) and (39.8), and Proposition 94.10.

Proof. *Derives Proposition 39.8.*

That the four sets are disjoint and exhaust L is the case analysis on the two signs. That each is a coset of $L^{\uparrow}_{+}$ follows because P, T and $PT = -\mathbb{1}$ realize every combination of the two signs, and multiplying by a fixed element is a homeomorphism of L onto itself, so each coset is connected if $L^{\uparrow}_{+}$ is. Connectedness of $L^{\uparrow}_{+}$ is proved in Chapter 94 as the continuous image of $\mathrm{SL}(2,\mathbb{C})$, which is connected. Finally $P^2 = T^2 = \mathbb{1}$ and $PT = TP$, which is the multiplication table of $\mathbb{Z}_2 \times \mathbb{Z}_2$. □

Only $L_+^\uparrow$ can be reached from the identity by a continuous sequence of physical motions, which is why it alone is generated by the Lie algebra of Section 39.2.7. The other three components are reached only by adjoining a discrete operation, and whether Nature respects those operations is an experimental question with a surprising answer: P is violated by the weak interaction (Chapter 110), the combination CP is violated in neutral-meson decays (Chapter 111), and only CPT — with C the charge conjugation of Chapter 105 — is protected by a theorem.

39.2.7 The algebra of the proper Lorentz group

In the context of Minkowski space we must find the generators of the rotations $\Lambda(0,\theta)$ and of the boosts $\Lambda(\beta,0)$, and then the algebra they satisfy.

Proposition 39.9 (The generators)**.** *Write an infinitesimal Lorentz transformation as $\Lambda^\mu{}_\nu = \delta^\mu{}_\nu + \omega^\mu{}_\nu$ with $|\omega| \ll 1$. Then Equation (39.2) holds to first order if and only if $\omega_{\mu\nu} := \eta_{\mu\lambda}\omega^\lambda{}_\nu$ is antisymmetric,*

$$\omega_{\mu\nu} = -\omega_{\nu\mu}\,, \tag{39.11}$$

so the algebra has $\binom{4}{2} = 6$ parameters: three rotations and three boosts.
Rests on Definition 39.1 and Equation (39.2).

Proof. *Derives Proposition 39.9.*
Substituting into $\Lambda^{\mathsf T}\eta\Lambda = \eta$ and keeping first-order terms,

$$\eta_{\mu\nu} = \left(\delta^\lambda{}_\mu + \omega^\lambda{}_\mu\right)\eta_{\lambda\rho}\left(\delta^\rho{}_\nu + \omega^\rho{}_\nu\right) = \eta_{\mu\nu} + \omega_{\nu\mu} + \omega_{\mu\nu} + O\!\left(\omega^2\right)\,,$$

so $\omega_{\mu\nu}+\omega_{\nu\mu}=0$. An antisymmetric 4×4 matrix has six independent entries; the three with both indices spatial generate rotations, and the three with one index equal to 0 generate boosts, as Equation (38.32) shows for a single boost. □

The six parameters are collected into an antisymmetric $\omega^{\mu\nu}$ and the six generators into an antisymmetric $J_{\mu\nu} = -J_{\nu\mu}$, so that a finite transformation is

$$\Lambda = \exp\left(\tfrac{1}{2}\omega^{\mu\nu}J_{\mu\nu}\right)\,. \tag{39.12}$$

Realizing the generators as the differential operators that implement these transformations on functions of x,

$$J_{\mu\nu} = x_\mu\partial_\nu - x_\nu\partial_\mu\,, \tag{39.13}$$

one obtains by direct computation the algebra $\mathfrak{so}(3,1)$:

$$[J_{\mu\nu}, J_{\lambda\rho}] = \eta_{\nu\lambda}J_{\mu\rho} + \eta_{\mu\rho}J_{\nu\lambda} - \eta_{\mu\lambda}J_{\nu\rho} - \eta_{\nu\rho}J_{\mu\lambda}\,, \tag{39.14}$$

in agreement with the general $\mathfrak{so}(p,q)$ relations derived in Section 14.3.2 (Equation (14.90)).

Derivation. *Derives Equation (39.14).*
Insert Equation (39.13) into the commutator and expand, using $\partial_\nu x_\lambda = \eta_{\nu\lambda}$ and the fact that the second-derivative terms cancel in pairs:

$$\begin{aligned}[J_{\mu\nu}, J_{\lambda\rho}] &= [x_\mu\partial_\nu - x_\nu\partial_\mu, x_\lambda\partial_\rho - x_\rho\partial_\lambda] \\ &= x_\mu\,[\partial_\nu, x_\lambda]\,\partial_\rho - x_\mu\,[\partial_\nu, x_\rho]\,\partial_\lambda - x_\nu\,[\partial_\mu, x_\lambda]\,\partial_\rho + x_\nu\,[\partial_\mu, x_\rho]\,\partial_\lambda \\ &\quad + x_\lambda\,[x_\mu\partial_\nu - x_\nu\partial_\mu, \partial_\rho] - x_\rho\,[x_\mu\partial_\nu - x_\nu\partial_\mu, \partial_\lambda]\,.\end{aligned}$$

Each commutator of a derivative with a coordinate is a metric component, $[\partial_\nu, x_\lambda] = \eta_{\nu\lambda}$, and collecting the eight resulting terms in pairs gives

$$[J_{\mu\nu}, J_{\lambda\rho}] = \eta_{\nu\lambda}\,(x_\mu\partial_\rho - x_\rho\partial_\mu) + \eta_{\mu\rho}\,(x_\nu\partial_\lambda - x_\lambda\partial_\nu) - \eta_{\mu\lambda}\,(x_\nu\partial_\rho - x_\rho\partial_\nu) - \eta_{\nu\rho}\,(x_\mu\partial_\lambda - x_\lambda\partial_\mu)\,,$$

each bracket being again a generator Equation (39.13). This is Equation (39.14). The computation nowhere used the signature, only the symmetry of η, which is why the same relations hold for $\mathfrak{so}(p,q)$ generally. □

Splitting the six generators into rotations $J_i = \frac{1}{2}\epsilon_{ijk}J^{jk}$ and boosts $K_i = J^{0i}$ makes the structure familiar: the J_i close among themselves — rotations form a subgroup — while

$$[J_i, J_j] = \epsilon_{ijk}J_k\,, \qquad [J_i, K_j] = \epsilon_{ijk}K_k\,, \qquad [K_i, K_j] = -\epsilon_{ijk}J_k\,. \tag{39.15}$$

From Equation (39.14) (the rotation–boost split just made).

The last bracket is the algebraic statement of Remark 38.24: two boosts do not commute, and their commutator is a rotation — the Thomas rotation. The minus sign distinguishes $\mathfrak{so}(3,1)$ from $\mathfrak{so}(4)$ and is the source of the non-compactness of Equation (39.3).

39.3 The Poincaré group

39.3.1 Foundations

The Lorentz group holds one event fixed: every Λ maps the origin to itself. But no event in Minkowski space is distinguished, so the full symmetry group must include the translations that move the origin.

Definition 39.10 (Poincaré group)**.** The *Poincaré group*, or inhomogeneous Lorentz group $\mathrm{ISO}(3,1)$, is the set of transformations

$$x^\mu \longmapsto \bar{x}^\mu = \Lambda^\mu{}_\nu x^\nu + a^\mu\,, \qquad \Lambda \in L\,, \quad a^\mu \in \mathbb{R}^4\,, \tag{39.16}$$

with the composition law

$$(\Lambda_2, a_2)\,(\Lambda_1, a_1) = (\Lambda_2\Lambda_1,\ \Lambda_2 a_1 + a_2)\ . \tag{39.17}$$

Rests on Definition 39.1.

Equation (39.17) is read off by applying the two transformations in succession, and it shows that the group is not a direct product: the translation part of the composite depends on the Lorentz part of the second factor. It is a *semidirect* product $\mathbb{R}^4 \rtimes L$, with the translations forming a normal subgroup and the Lorentz group acting on them.

Proposition 39.11 (Invariance of the interval)**.** *The Poincaré transformations are exactly the maps of Minkowski space preserving the interval Equation (38.13) between every pair of events.*

Rests on Definition 39.10 and Equations (38.13) and (39.2).

Proof. *Derives Proposition 39.11.*

The separation of two events transforms as $\Delta\bar{x}^\mu = \Lambda^\mu{}_\nu \Delta x^\nu$ — the constant a^μ cancels in the difference — so Equation (39.2) gives invariance of $\eta_{\mu\nu}\Delta x^\mu \Delta x^\nu$. Conversely, a map preserving all intervals preserves in particular those from a fixed event, and the polarization identity then forces it to be affine with linear part satisfying Equation (39.2). □

The group has ten parameters: six of Λ from Proposition 39.9 and four of a^μ. These are the ten symmetries whose conserved charges were listed in Section 39.1, and the classification of the irreducible unitary representations of this group is the classification of the possible kinds of elementary particle — Wigner's theorem, carried out in Chapter 94.

39.3.2 The algebra of the Poincaré group

The commutation relations of the Poincaré algebra are

$$[J_{\mu\nu}, J_{\lambda\rho}] = \eta_{\nu\lambda} J_{\mu\rho} + \eta_{\mu\rho} J_{\nu\lambda} - \eta_{\mu\lambda} J_{\nu\rho} - \eta_{\nu\rho} J_{\mu\lambda} \,, \tag{39.18}$$

$$[J_{\mu\nu}, P_{\lambda}] = \eta_{\nu\lambda} P_{\mu} - \eta_{\mu\lambda} P_{\nu} \,, \tag{39.19}$$

$$[P_{\mu}, P_{\nu}] = 0 \,. \tag{39.20}$$

Derivation. *Derives Equation* (39.19).

The Lorentz generators are Equation (39.13) and the translation generators are

$$P_{\mu} = \partial_{\mu} \,, \tag{39.21}$$

since $\exp(a^{\mu}\partial_{\mu})$ shifts the argument of a function by a^{μ}. Equation (39.18) was derived in Section 39.2.7. For Equation (39.19),

$$\begin{aligned} [J_{\mu\nu}, P_{\lambda}] &= [x_{\mu}\partial_{\nu} - x_{\nu}\partial_{\mu}, \partial_{\lambda}] \\ &= -(\partial_{\lambda} x_{\mu})\,\partial_{\nu} + (\partial_{\lambda} x_{\nu})\,\partial_{\mu} = -\eta_{\lambda\mu}\partial_{\nu} + \eta_{\lambda\nu}\partial_{\mu} = \eta_{\nu\lambda} P_{\mu} - \eta_{\mu\lambda} P_{\nu} \,, \end{aligned}$$

the derivative acting only on the explicit coordinate because the second-derivative terms cancel. Equation (39.20) is the commutativity of partial derivatives, $[\partial_{\mu}, \partial_{\nu}] = 0$: translations in different directions commute, so the translation subgroup is abelian.

The three relations say exactly what Definition 39.10 already said in words: Equation (39.20) that the translations are abelian, Equation (39.19) that they form a normal subgroup transforming as a four-vector under Lorentz transformations, and Equation (39.18) that the Lorentz generators close among themselves. The semidirect structure is visible in the fact that Equation (39.19) does not vanish, while it would for a direct product. □

Remark 39.12 (The Casimir invariants). Two combinations commute with every generator of the algebra:

$$P^2 = \eta^{\mu\nu} P_{\mu} P_{\nu} \,, \qquad W^2 = \eta^{\mu\nu} W_{\mu} W_{\nu} \,, \qquad W^{\mu} := -\tfrac{1}{2} \epsilon^{\mu\nu\lambda\rho} J_{\nu\lambda} P_{\rho} \,, \tag{39.22}$$

the second built from the Pauli–Lubański vector. On an irreducible representation each takes a fixed value, $P^2 = m^2 c^2$ and $W^2 = -m^2 c^2 s\,(s+1)\,\hbar^2$ for $m > 0$. Those two numbers are the mass and the spin: a particle *is* an irreducible representation of the Poincaré group, labelled by the two invariants of its algebra. This is the point at which the kinematics of this part becomes the particle classification of Chapter 94.

Rests on Equations (39.18) *to* (39.20).

Chapter 40

Relativistic Dynamics

Chapter 38 settled how coordinates transform. Dynamics must now be rebuilt on that kinematics, because Newton's second law in the form $\boldsymbol{F} = m\ddot{\boldsymbol{x}}$ is not Lorentz covariant: it singles out a frame's time as the parameter of the motion. The repair is to express everything in objects that transform under Λ in a definite way — four-vectors — and to differentiate with respect to the invariant proper time of Definition 38.12 rather than a coordinate time.

40.1 Four-vectors

Definition 40.1 (Four-vector)**.** A *four-vector* is a quadruple A^μ transforming under a change of inertial frame as the coordinates do,

$$\bar{A}^\mu = \Lambda^\mu{}_\nu A^\nu \,. \tag{40.1}$$

Its *square* $A^2 := \eta_{\mu\nu}A^\mu A^\nu$ is then invariant, by Equation (38.14), and so is the scalar product $A \cdot B = \eta_{\mu\nu}A^\mu B^\nu$ of any two four-vectors.
Rests on Definition 38.6 and Equation (38.14).

Invariants are the currency of the subject: a quantity built as the square of a four-vector has the same value in every frame, so it may be computed in whichever frame makes it easiest and quoted in any other.

Definition 40.2 (Four-velocity)**.** For a particle with worldline $x^\mu(\tau)$ parametrized by its proper time, the *four-velocity* is

$$u^\mu := \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} = \gamma_u \left(c, \boldsymbol{u}\right) , \tag{40.2}$$

where $\boldsymbol{u} = \mathrm{d}\boldsymbol{x}/\mathrm{d}t$ is the ordinary velocity and $\gamma_u = \left(1 - |\boldsymbol{u}|^2/c^2\right)^{-1/2}$.
Rests on Definitions 38.12 and 40.1.

The factor γ_u appears because $\mathrm{d}t = \gamma_u \mathrm{d}\tau$ by Equation (38.26). Since τ is invariant and x^μ is a four-vector, u^μ is one too — which $\mathrm{d}\boldsymbol{x}/\mathrm{d}t$ is not, and that is precisely why the ordinary velocity is the wrong object.

Proposition 40.3 (Normalization of the four-velocity)**.** *$u^2 = c^2$ for every massive particle, in every frame.*
Rests on Equation (40.2) *and Definition 38.6.*

Proof. *Derives Proposition 40.3.*
Directly from Equation (40.2) and $\eta = \mathrm{diag}(1,-1,-1,-1)$,

$$u^2 = \gamma_u^2\left(c^2 - |\boldsymbol{u}|^2\right) = \frac{c^2 - |\boldsymbol{u}|^2}{1 - |\boldsymbol{u}|^2/c^2} = c^2 \,.$$

Alternatively, divide $\mathrm{d}s^2 = c^2\mathrm{d}\tau^2$ by $\mathrm{d}\tau^2$. □

A four-velocity is therefore never zero and always timelike: a particle cannot stand still in spacetime, only redistribute its motion between the time and space directions. At rest it moves through time at rate c; as it speeds up, part of that is traded for spatial motion, and the trade is constrained by Proposition 40.3.

40.2 Four-momentum and mass–energy

Definition 40.4 (Four-momentum)**.** For a particle of *rest mass* m — an invariant, the same in every frame — the *four-momentum* is

$$p^\mu := mu^\mu = (\gamma_u mc,\ \gamma_u m\boldsymbol{u}) =: \left(\frac{E}{c},\ \boldsymbol{p}\right), \tag{40.3}$$

which defines the relativistic energy E and three-momentum $\boldsymbol{p}$:

$$E = \gamma_u mc^2\,, \qquad \boldsymbol{p} = \gamma_u m\boldsymbol{u}\,. \tag{40.4}$$

Rests on Definition 40.2.

Theorem 40.5 (Energy–momentum relation)**.**

$$E^2 = |\boldsymbol{p}|^2 c^2 + m^2c^4\,, \tag{40.5}$$

and equivalently $p^2 = m^2c^2$.

Rests on Definition 40.4 and Proposition 40.3.

Proof. *Derives Theorem 40.5.*

By Definition 40.4 and Proposition 40.3, $p^2 = m^2u^2 = m^2c^2$. Writing the same square in components, $p^2 = E^2/c^2 - |\boldsymbol{p}|^2$, and equating the two gives Equation (40.5). □

Phenomenon 40.6 (Mass–energy equivalence)**.** A body at rest possesses an energy

$$E_0 = mc^2\,, \tag{40.6}$$

and any change in its internal energy changes its rest mass by $\Delta m = \Delta E/c^2$.

Rests on Equation (40.4).

Derivation. *Derives Phenomenon 40.6.*

Set $\boldsymbol{u} = 0$ in Equation (40.4): $\gamma_u = 1$ and $E = mc^2$, which is Equation (40.6). That this is a real energy and not an additive constant follows from expanding Equation (40.4) for $|\boldsymbol{u}| \ll c$,

$$E = mc^2\left(1 - \frac{u^2}{c^2}\right)^{-1/2} = mc^2 + \frac{1}{2}mu^2 + \frac{3}{8}\frac{mu^4}{c^2} + O\left(\frac{u^6}{c^4}\right), \tag{40.7}$$

whose second term is the Newtonian kinetic energy. The relativistic kinetic energy is therefore $T = E - mc^2 = (\gamma_u - 1)\,mc^2$, and E_0 is what remains when all motion is removed.

For the second statement, consider a body at rest emitting two equal light pulses in opposite directions with total energy L, so that it stays at rest. Evaluating the energy balance in a frame moving at $-v$ and expanding to second order in β, the kinetic energies before and after differ by $(L/c^2)\,v^2/2$, which is the kinetic energy of a body whose mass has fallen by L/c^2. This is Einstein's argument [320]; he states the conclusion as a mass change L/V^2 and quotes it numerically as $L/(9\times10^{20})$ with the energy in erg and the mass in gram — that is, $c^2 = 9\times10^{16}\,\mathrm{J/kg}$ to one significant figure. □

Remark 40.7 (There is no "relativistic mass" in this treatise). Some older treatments absorb γ_u into the mass, writing $m_{\text{rel}} = \gamma_u m$ so that $\boldsymbol{p} = m_{\text{rel}}\boldsymbol{u}$ looks Newtonian. The device is consistent but costs more than it saves: the same m_{rel} does not make $\boldsymbol{F} = m_{\text{rel}}\boldsymbol{a}$ true (Section 40.3), and it obscures that m is a Lorentz invariant while $\gamma_u m$ is one component of a four-vector. Throughout this book *mass* means rest mass, and it never depends on speed.
Rests on Equation (40.9).

Example 40.8 (Nuclear binding as a mass deficit)**.** Phenomenon 40.6 is not an exotic effect: it is how nuclear masses are read. The mass of a bound nucleus is less than the sum of its constituents' masses by B/c^2, with B the binding energy; for ^{4}He the deficit is about $28.3\,\text{MeV}/c^2$, close to 0.8 % of the total. The systematics of that deficit across the nuclide chart is what makes both fission and fusion energetically possible (Chapter 107).
Rests on Phenomenon 40.6.

40.3 The relativistic force law

Definition 40.9 (Four-force)**.**

$$f^\mu := \frac{\mathrm{d}p^\mu}{\mathrm{d}\tau}\,, \tag{40.8}$$

whose spatial part is related to the ordinary force $\boldsymbol{F} = \mathrm{d}\boldsymbol{p}/\mathrm{d}t$ by $f^i = \gamma_u F^i$, and whose time component is $f^0 = \gamma_u\,(\boldsymbol{F}\cdot\boldsymbol{u})\,/c$, the rate at which the force does work.
Rests on Definitions 38.12 and 40.4.

Newton's second law is recovered in the covariant form $f^\mu = m\,\mathrm{d}u^\mu/\mathrm{d}\tau$, which reduces to $\boldsymbol{F} = m\boldsymbol{a}$ when $|\boldsymbol{u}| \ll c$. It does *not* reduce to $\boldsymbol{F} = \gamma_u m\boldsymbol{a}$ in general, because γ_u depends on time through $\boldsymbol{u}$:

$$\boldsymbol{F} = \frac{\mathrm{d}}{\mathrm{d}t}\,(\gamma_u m\boldsymbol{u}) = \gamma_u m\boldsymbol{a} + \gamma_u^3\frac{m}{c^2}\,(\boldsymbol{u}\cdot\boldsymbol{a})\,\boldsymbol{u}\,. \tag{40.9}$$

From Equation (40.4) (differentiating $\boldsymbol{p} = \gamma_u m\boldsymbol{u}$ with respect to coordinate time, with γ_u carried along).
Force and acceleration are therefore not parallel unless $\boldsymbol{a}$ is parallel or perpendicular to $\boldsymbol{u}$, in which cases Equation (40.9) collapses to $F_\parallel = \gamma_u^3 m a_\parallel$ and $F_\perp = \gamma_u m a_\perp$ respectively. The two different coefficients were once called the longitudinal and transverse masses; they are a symptom of forcing a Newtonian form onto a non-Newtonian law, and Remark 40.7 applies.

Corollary 40.10 (Nothing massive reaches c)**.** *A body of nonzero rest mass cannot be accelerated to the speed of light by any finite energy.*
Rests on Equation (40.4).

Proof. *Derives Corollary 40.10.*
By Equation (40.4), $E = \gamma_u mc^2 \to \infty$ as $|\boldsymbol{u}| \to c$, since $\gamma_u \to \infty$. The kinetic energy required is $(\gamma_u - 1)\,mc^2$, which diverges. □

This is measured directly, and simply: Bertozzi accelerated electrons through known potentials up to 15 MeV and timed their flight over a measured baseline, finding the speed saturating just below c while the kinetic energy continued to rise in proportion to the applied potential [1272]. Galilean kinematics predicts $v^2 \propto T$ without limit; the data follow $(\gamma - 1)\,mc^2 = T$ instead.

40.4 Massless particles

Setting $m = 0$ in Equation (40.5) gives

$$E = |\boldsymbol{p}|\,c\,, \tag{40.10}$$

a consistent possibility, describing a particle whose four-momentum is null, $p^2 = 0$. Such a particle has no rest frame — there is no boost carrying a null vector to a purely timelike one — and by Equation (40.4) it must travel at exactly c in every frame, since $\boldsymbol{u} = \boldsymbol{p}c^2/E$. The photon is the instance (Chapter 69); the four-momentum $p^\mu = \hbar k^\mu$ ties Equation (40.10) to the wave description, and the Doppler and aberration formulas of Section 38.5 are then statements about how k^μ transforms.

40.5 Collisions

Postulate 40.11 (Conservation of four-momentum)**.** In any isolated process the total four-momentum is conserved:

$$\sum_{\text{in}} p^\mu = \sum_{\text{out}} p^\mu \,. \tag{40.11}$$

Rests on Definition 40.4.

Because Equation (40.11) is an equation between four-vectors, it holds in every frame if it holds in one — which is exactly what a conservation law must do, and what separate conservation of energy and of momentum could not guarantee on its own. Its time component is conservation of energy and its spatial components conservation of momentum; the two are no longer independent statements but components of a single one.

Definition 40.12 (Invariant mass)**.** For a system of particles the *invariant mass* M is defined by

$$M^2c^2 := \left(\sum_a p_a^\mu\right)^2 = \frac{1}{c^2}\left(\sum_a E_a\right)^2 - \left(\sum_a \boldsymbol{p}_a\right)^2 . \tag{40.12}$$

Rests on Definitions 40.1 and 40.4.

The invariant mass of a system is not the sum of its constituents' masses: kinetic energy and binding energy contribute, which is Example 40.8 again. It is, however, invariant and conserved, and it equals the total energy divided by c^2 in the frame where the total momentum vanishes — the centre-of-momentum frame.

Example 40.13 (Threshold energy for pair production)**.** A proton of energy E strikes a proton at rest; what is the least E at which $p + p \to p + p + p + \bar{p}$ can proceed? At threshold the four final particles are produced at rest in the centre-of-momentum frame, so the final invariant mass is $4m_p$. Evaluating Equation (40.12) for the initial state in the laboratory, where $p_1 = (E/c, \boldsymbol{p})$ and $p_2 = (m_p c, \boldsymbol{0})$,

$$M^2c^2 = (p_1 + p_2)^2 = p_1^2 + p_2^2 + 2p_1 \cdot p_2 = 2m_p^2c^2 + 2Em_p \,,$$

using $p_i^2 = m_p^2c^2$. Setting $M = 4m_p$ and solving,

$$E = 7m_pc^2 \,, \qquad T = E - m_pc^2 = 6m_pc^2 \approx 5.6\,\text{GeV} \,. \tag{40.13}$$

From Equations (40.5) and (40.12) (setting $M = 4m_p$ at threshold and solving the laboratory invariant for E).

Six rest masses of kinetic energy are needed to create two, because momentum conservation forbids the products from being at rest in the laboratory. This is why colliders beat fixed targets: with two beams of equal and opposite momentum the centre-of-momentum frame *is* the laboratory, and the whole of the energy is available (Chapter 113).

Rests on Postulate 40.11 and Definition 40.12.

Example 40.14 (Compton scattering)**.** A photon of wavelength λ scatters off an electron at rest and emerges at angle ϑ with wavelength λ'. Conservation Equation (40.11) reads $k + p = k' + p'$; isolating the unobserved electron and squaring,

$$p'^2 = (k - k' + p)^2 = -2k \cdot k' + 2p \cdot (k - k') + p^2 ,$$

and since $p'^2 = p^2 = m_e^2 c^2$ and $k^2 = k'^2 = 0$, the terms rearrange to $k \cdot k' = p \cdot (k - k')$. Evaluating in the laboratory with $p = (m_e c, \mathbf{0})$ and $k^\mu = (\hbar\omega/c)\,(1, \hat{\boldsymbol{n}})$ gives the Compton relation

$$\lambda' - \lambda = \frac{h}{m_e c}\,(1 - \cos\vartheta) , \tag{40.14}$$

whose measurement is reported in Chapter 69. The whole derivation used only Postulate 40.11 and the invariance of the squares — no dynamics of the interaction at all.

Rests on Postulate 40.11, Theorem 40.5, and Equation (40.10)*.*

40.6 The twin paradox

Phenomenon 40.15 (Differential ageing)**.** Of two clocks that separate from a common event and reunite at another, the one that travelled inertially throughout records the greater proper time [1476, 1592].

Rests on Definitions 38.6 and 38.12.

Derivation. Derives Phenomenon 40.15.

Proper time along a worldline is $\tau = \int \mathrm{d}t/\gamma_u$ by Equation (38.26). For the inertial twin, who may be taken at rest, $\gamma_u = 1$ throughout and $\tau_{\text{stay}} = T$, the coordinate time between departure and reunion. For the travelling twin $\gamma_u > 1$ wherever $\boldsymbol{u} \neq 0$, so

$$\tau_{\text{trip}} = \int_0^T \frac{\mathrm{d}t}{\gamma_u} < T = \tau_{\text{stay}} ,$$

strictly, since the traveller must move to return. For a trip out and back at constant speed v, $\tau_{\text{trip}} = T/\gamma$. □

The result is often called a paradox on the grounds that Remark 38.14 makes time dilation reciprocal, so each twin should find the other younger. The objection fails because the situation is not symmetric: the stay-at-home twin occupies one inertial frame, while the traveller occupies two different ones joined by a turnaround. Only the traveller feels an acceleration, and only the traveller's surfaces of simultaneity swing round at the turn, skipping over a long stretch of the stay-at-home's worldline.

No appeal to general relativity is needed. The proper times are path lengths in Minkowski space, and Phenomenon 40.15 is the statement that the straight path between two timelike-separated events is the *longest* in proper time — the reversal, forced by the minus signs in η, of the Euclidean fact that a straight line is the shortest path. The effect has been measured directly on flown atomic clocks [1476, 1477] and, far more precisely, on circulating muons [1592]; both are reported in Chapter 41.

40.7 Covariant electrodynamics in one line

Relativistic dynamics was not invented to repair mechanics; it was extracted from electrodynamics, which had been Lorentz covariant all along without anyone noticing. The Lorentz force law $\boldsymbol{F} = q\,(\boldsymbol{E} + \boldsymbol{u} \times \boldsymbol{B})$ becomes, in the covariant notation of this chapter,

$$\frac{\mathrm{d}p^\mu}{\mathrm{d}\tau} = qF^\mu{}_\nu u^\nu , \tag{40.15}$$

with $F^{\mu\nu}$ the electromagnetic field tensor assembled from $\boldsymbol{E}$ and $\boldsymbol{B}$. That single equation contains both the electric and magnetic force, and the transformation properties of $F^{\mu\nu}$ explain why the split between them is frame-dependent: what one observer calls a magnetic force on a moving charge, another calls an electric force on a charge at rest. The construction belongs with Maxwell's equations and is carried out in Chapters 60 and 64.

Chapter 41

Experiment: Time Dilation and Relativistic Kinematics

Chapter 38 derived time dilation (Phenomenon 38.13) and Chapter 40 the energy–momentum relation. This chapter reports the measurements. They span five orders of magnitude in $\gamma - 1$, from canal-ray ions at $\beta \approx 4 \times 10^{-3}$ to storage-ring muons at $\gamma = 29.33$, and they agree with the same one-parameter formula throughout.

41.1 Ives and Stilwell: the moving atomic clock (1938)

Tests Phenomenon 38.13.

Assuming Definition 38.12.

The cleanest laboratory test uses light itself as the clock. A moving atom radiating at a known rest frequency is a clock whose ticking is read optically; the first-order Doppler shift is a classical effect, but the *second*-order part is pure time dilation and survives when the first-order part is cancelled.

41.1.1 Apparatus

A canal-ray tube producing fast hydrogen ions (H_2^+ and H_3^+) at accelerating potentials from about 6788 V to 18 350 V — the tube sparked above roughly 20 kV. A spectrograph viewed the beam at 7° to its axis, recording simultaneously the direct beam and its image in a mirror behind the tube, so that light emitted forward and backward reached the plate side by side. The line studied was the blue-green Hβ Balmer line at $\lambda_0 = 486.1\,\text{nm}$.

41.1.2 Procedure

Viewing the same emitters both ways gives two lines displaced by the first-order Doppler effect, one to the red and one to the blue. Their *mean* position is what matters: the first-order terms cancel in the average, leaving the second-order shift

$$\Delta\lambda = \frac{\lambda_b + \lambda_f}{2} - \lambda_0 = (\gamma - 1)\,\lambda_0 \simeq \frac{1}{2}\beta^2 \lambda_0\,, \tag{41.1}$$

From Equation (38.27) (written for a moving light source).

The experimental art is that the first-order displacement is some 2 mm on the plate while the second-order shift to be extracted is about 0.005 mm.

41.1.3 Observations and data

Ion speeds ran from $\beta = 2.184\,736 \times 10^{-3}$ to $\beta = 4.433\,244 \times 10^{-3}$, that is v from about $6.55 \times 10^5\,\text{m/s}$ to $1.33 \times 10^6\,\text{m/s}$; the beam velocities inferred from the first-order shifts

agreed with $eV = \tfrac{1}{2}Mv^2$ to within 1 %. The measured second-order shifts for the eight tabulated cases, against the values predicted by Equation (41.1), were

Table 41.1: Measured and predicted second-order Doppler shift in Section 41.1, in picometres (the paper tabulates ångström; 1 Å=100 pm). The ratio scatters between 0.911 and 1.035 about the relativistic value.

$\Delta\lambda$ measured	$\Delta\lambda$ predicted	ratio
1.10	1.16	0.948
1.85	2.03	0.911
2.25	2.38	0.945
2.70	2.75	0.982
2.05	1.98	1.035
3.45	3.52	0.980
2.15	2.33	0.923
4.70	4.78	0.983

41.1.4 Interpretation

The mean displacement is nonzero, of the predicted sign, and of the predicted size — a classical clock would show none at all, since the first-order effects cancel exactly. Ives and Stilwell quote no single headline ratio with an uncertainty, presenting instead graphs of $\Delta\lambda$ against $\tfrac{1}{2}\beta^2\lambda_0$ and concluding that the results are a satisfactory confirmation. A later assessment by Mandelberg and Witten puts the experimental uncertainty of this generation of measurements nearer 10–15 %, and their own repeat found the exponent n in $(1-\beta^2)^n$ to be 0.498 ± 0.025 where special relativity requires exactly $\tfrac{1}{2}$.

41.1.5 Primary references

[816]. The 10–15 % uncertainty estimate and the exponent 0.498 ± 0.025 are from Mandelberg and Witten, *J. Opt. Soc. Am.* **52**, 529 (1962), and are quoted here at second hand.

A muon is a clock with a single tick: it decays, and the distribution of decay times is exponential with a mean proper lifetime $\tau_0 \approx 2.2\,\mu\text{s}$. Muons are created by cosmic-ray primaries in the upper atmosphere and arrive at the ground in numbers that Galilean kinematics cannot account for.

41.2 Rossi and Hall: muon survival with altitude (1941)

Tests Phenomena 38.13 and 115.55.
Assuming Definition 38.12.

41.2.1 Apparatus

A telescope of self-quenching Geiger–Müller tubes of 4 cm internal diameter, in coincidence and anticoincidence, with a permanent absorber equivalent to $186\,\text{g/cm}^2$ of lead above and between the counters and 11.5 cm lead side walls. The assembly, except the variable absorber, sat in a thermostatic box and was carried by truck between stations.

41.2.2 Procedure

Coincidences with anticoincidence in the lower counters select particles that stop in the absorber; inserting $115\,\text{g/cm}^2$ of lead and differencing isolates muons in a chosen residual-range interval, and hence a chosen momentum band. Comparing rates at two altitudes measures the fraction surviving the flight between them. A $200\,\text{g/cm}^2$ iron compensator at

the higher station equalizes the total material traversed, so that the difference in counts reflects decay rather than absorption.

41.2.3 Observations and data

Stations: Echo Lake, Colorado, at 3240 m altitude and 709 g/cm^2 atmospheric depth, and Denver at 1616 m and 856 g/cm^2 — a difference of 1624 m, equivalent to 147 g/cm^2 of air, at practically the same geomagnetic latitude. Survival ratios between the two stations:

$$w_{12} = 0.698 \pm 0.031 \quad \text{(residual range 196–311 g/cm}^2\text{ of Pb)}\,, \tag{41.2}$$

$$W_{12} = 0.883 \pm 0.007 \quad \text{(residual range above 311 g/cm}^2\text{ of Pb)}\,. \tag{41.3}$$

The softer group, of effective momentum 5.0×10^8 eV/c, disintegrates about three times faster than the penetrating group, of effective momentum 1.5×10^9 eV/c. Converting the survival fractions to a mean range before decay, $L = (4.5 \pm 0.6) \times 10^5$ cm, and using a muon mass of 8×10^7 eV/c^2, the authors obtain the proper lifetime

$$\tau_0 = (2.4 \pm 0.3) \times 10^{-6}\,\text{s}\,. \tag{41.4}$$

From Equations (41.2) and (41.3) (the two survival fractions converted to a mean range before decay).

41.2.4 Interpretation

The two momentum groups traverse the same distance, so under Galilean kinematics they would decay in the same proportion; they do not. The faster group survives better, by the amount Equation (38.27) requires for its larger γ. The abstract states the conclusion plainly: the rates are "in agreement with the theoretical predictions based on the relativity change in rate of a moving clock".

41.2.5 Primary references

[863]. "Mesotron" in the original is the modern muon.

41.3 Frisch and Smith: Mount Washington to sea level (1963)

Tests Phenomena 38.13 and 115.55.
Assuming Definition 38.12.

41.3.1 Apparatus

A doped-polystyrene scintillator cylinder 28 cm high and 28 cm in diameter (about 30 g/cm^2), viewed by a 13 cm photomultiplier, with the pulse train displayed on a Tektronix 581 oscilloscope at 8.5 µs sweep and photographed. Iron absorber above the scintillator: 600 g/cm^2 on the mountain and 359 g/cm^2 at sea level (the paper reports the apparatus in inches and feet; SI equivalents are given here per the unit axiom of Chapter 2), the difference compensating the 218 g/cm^2 of air between the stations.

41.3.2 Procedure

The absorber and geometry select muons in a narrow speed band, between 0.9950c and 0.9954c. A muon that stops in the scintillator gives a first pulse; its decay electron gives a second. Counting stopped muons per hour at the two altitudes measures the survival fraction over the descent, and the distribution of intervals between the pulse pairs measures the lifetime in the same apparatus.

41.3.3 Observations and data

Six one-hour runs on the summit of Mount Washington at 1909 m gave 568, 554, 582, 527, 588 and 559 counts; five at Cambridge, Massachusetts, 3 m above sea level, gave 412, 403, 436, 395 and 393. The averages are

$$N_{\text{summit}} = 563 \pm 10 \text{ h}^{-1}, \qquad N_{\text{sea}} = 408 \pm 9 \text{ h}^{-1}. \tag{41.5}$$

The vertical descent is 1907 m, covered in 6.4 µs of earth-frame time at $0.9952c$. With backgrounds of 13 ± 2 and 11 ± 2 subtracted, the survival ratio is

$$\frac{397 \pm 9}{550 \pm 10} = 0.722 \pm 0.021, \tag{41.6}$$

and equating this to $\exp[-1907/(\gamma c \tau_0)]$ with $\tau_0 = (2.211 \pm 0.003) \times 10^{-6}$ s gives the observed dilation factor

$$\gamma_{\text{obs}} = 8.8 \pm 0.8, \tag{41.7}$$

against the predicted effective value 8.4 ± 2.0 for this geometry.

41.3.4 Interpretation

Without dilation the expected sea-level rate would be about 27 per hour rather than 408: the muons survive some fifteen times better than Galilean kinematics permits. The measured factor Equation (41.7) agrees with the prediction within the uncertainties, which the authors are careful to describe as statistical only, the quoted ± 2 on the prediction being "essentially only guessed at". The experiment is included here because it is the most direct possible reading of Phenomenon 38.13: two numbers of counts, one distance, one lifetime.

41.3.5 Primary references

[1253].

41.4 The CERN muon storage ring (1977)

Tests Equations (38.26) *and* (38.27).

41.4.1 Apparatus

The CERN Muon Storage Ring, holding polarized positive and negative muons in a circular orbit at Lorentz factor $\gamma = 29.33$, with detectors counting the decay electrons as a function of time.

41.4.2 Procedure

Muons circulate and decay; the decay-electron rate falls exponentially with a time constant that is the *dilated* lifetime. Measuring it for μ^+ and μ^- separately tests the dilation factor and, by comparison with the at-rest lifetime, also tests CPT.

41.4.3 Observations and data

$$\tau_+ = (64.419 \pm 0.058)\,\mu\text{s}, \qquad \tau_- = (64.368 \pm 0.029)\,\mu\text{s}. \tag{41.8}$$

The dilation factor agrees with the special-relativistic prediction with a fractional error of 2×10^{-3} at 95 % confidence. Dividing, $\tau_-/\gamma = 2.1948\,\mu\text{s}$, consistent with the free-muon lifetime.

41.4.4 Interpretation

This is the most precise direct test of Equation (38.27), and it is performed at a γ thirteen times larger than Frisch and Smith's. Note also what it does *not* disturb: the muons are in circular orbit and therefore permanently accelerated, at some 10^{18} times the acceleration of gravity, and the dilation depends only on their speed. That is the experimental content of the *clock hypothesis* — proper time along a worldline depends on the instantaneous velocity and not on the acceleration — which Equation (38.26) assumes and which the twin-paradox analysis of Section 40.6 needs.

41.4.5 Primary references

[1592]. The derived rest lifetime $\tau_0 = (2.1948 \pm 0.0010)\,\mu s$ is quoted from the abstract record rather than from the article body, which was not consulted; the storage-ring beam momentum and radius are deliberately not quoted here for the same reason. The 1979 final report of the same collaboration states the special-relativistic time transformation valid to $(8 \pm 7) \times 10^{-4}$.

41.5 Hafele and Keating: caesium clocks flown around the world (1972)

Tests Phenomenon 38.13.
Assuming Phenomenon 42.5.

41.5.1 Apparatus

Four caesium-beam atomic clocks, flown as ordinary airline passengers eastward and then westward around the world, and compared before and after with the reference ensemble of the United States Naval Observatory.

41.5.2 Procedure

Both flights circumnavigate the globe; the eastward one moves with the Earth's rotation and the westward against it, so the two have different speeds relative to the non-rotating frame in which the analysis is done. The prediction [1477] was published before the observation [1476].

41.5.3 Observations and data

Predicted and observed time differences relative to the ground clocks, in nanoseconds, are reported for each direction, with the eastward clocks losing time and the westward clocks gaining it — opposite signs, which is the striking part.

41.5.4 Interpretation

This experiment is included with a caveat that matters for the architecture of this treatise: it is *not* a clean test of special relativity. The total effect is the sum of the kinematic dilation of this part, which always slows the flying clock, and the gravitational blueshift of Chapter 42, which speeds it up at altitude. The two are comparable in size at airline altitudes and have opposite signs, which is why the eastward and westward results differ in sign. The experiment therefore tests the *sum* of two predictions from two different theories; it is quoted here because it is the most familiar demonstration that dilation applies to ordinary macroscopic clocks, and its gravitational half is properly treated in Part V.

41.5.5 Primary references

[1477], the advance prediction; [1476], the observation.

41.6 Bertozzi: the speed limit (1964)

Tests Equations (40.4) *and* (40.5).

41.6.1 Apparatus

A Van de Graaff accelerator and a linear accelerator delivering electrons of kinetic energy from about 0.5 MeV to 15 MeV, with a time-of-flight baseline of several metres and a calorimeter to measure the beam energy independently of the applied potential.

41.6.2 Procedure

Electron bunches are timed over the baseline, giving the speed directly; the kinetic energy is known from the accelerating potential and confirmed calorimetrically. Plotting v^2 against T separates the two predictions.

41.6.3 Observations and data

As T rises through the megavolt range, v approaches c and saturates: the measured v^2 flattens while T continues to climb in proportion to the applied potential. The calorimetric energies confirm that the energy really is being delivered to the beam and not lost.

41.6.4 Interpretation

Newtonian mechanics predicts $T = \tfrac{1}{2}mv^2$, hence v^2 linear in T without bound and $v > c$ above 0.25 MeV. The data instead follow $T = (\gamma - 1)\,mc^2$ from Equation (40.7), with $v \to c$ asymptotically. This is Corollary 40.10 measured directly, and it is the most economical demonstration that the relativistic energy–momentum relation Equation (40.5) and not the Newtonian one governs fast particles.

41.6.5 Primary references

[1272].

41.7 Summary of the evidence

No result in Table 41.2 is compatible with Galilean kinematics, and all of them are described by the single factor γ with no adjustable parameter. Together with the isotropy tests of Chapter 37, they close the experimental case for the transformations of Chapter 38.

Table 41.2: Direct tests of relativistic kinematics reported in this chapter. The regime column gives the Lorentz factor or velocity parameter reached.

Experiment	Regime	Result
Ives–Stilwell 1938	$\beta \sim 4 \times 10^{-3}$	second-order Doppler shift present, of predicted size to a few percent
Rossi–Hall 1941	$p \sim 1\,\mathrm{GeV}/c$	momentum-dependent decay rate; $\tau_0 = (2.4 \pm 0.3)\,\mu\mathrm{s}$
Frisch–Smith 1963	$\gamma \approx 8$	$\gamma_{\mathrm{obs}} = 8.8 \pm 0.8$ against 8.4 ± 2.0 predicted
Bailey *et al.* 1977	$\gamma = 29.33$	dilation factor confirmed to 2×10^{-3} at 95 % confidence
Bertozzi 1964	$T \leq 15\,\mathrm{MeV}$	speed saturates at c; $T = (\gamma - 1)\,mc^2$
Hafele–Keating 1972	$\beta \sim 10^{-6}$	flown clocks disagree with ground clocks, kinematic plus gravitational

Part V

General Relativity and Cosmology

Chapter 42

The Equivalence Principle and Classical Tests

General relativity is admitted to this treatise on the same terms as every other theory in it: by its evidence, and evidence first. This chapter collects that evidence in its historical and logical order — the strict proportionality of inertial and gravitational mass, which is what makes gravity universal and first suggests that it is geometry rather than force; the anomalous perihelion advance of Mercury, measured half a century before the theory that explained it; the deflection of starlight and the gravitational redshift; the fourth test found by Shapiro in the radar age; and the engineering test performed daily by every satellite-navigation receiver. The geometric formulation these facts support is constructed in Chapter 43, and the field equations they constrain in Chapter 44.

The organizing principle is the equivalence principle in its three strengths — weak, Einstein, strong — each tested by a different class of experiments and each licensing a different part of the geometric description [2240]. The chapter states the principles precisely, reports the experiments with their numbers and uncertainties, and keeps prediction and postdiction carefully apart. The measurements themselves — apparatus, procedure, observations, error budgets — are set out in Chapter 53, which tests the phenomena stated here and the field equations of Chapter 44.

42.1 The universality of free fall

42.1.1 From Galileo to Newton

That all bodies fall alike is not obvious — everyday experience, where air resistance discriminates between a stone and a feather, argues against it — and it was established as an idealized law before it could be tested cleanly. Galileo's route, set out in the *Discorsi*, was to dilute free fall on an inclined plane until its acceleration could be timed, and to argue away the residual disagreement between bodies as the work of the medium; his logical argument is still the sharpest short statement of the law. Suppose the heavy body fell faster. Tying a light body to it should then retard it, as the light partner drags; but the two tied together are a heavier body still, and should fall faster than either. The supposition contradicts itself, and composition-dependent fall with it [4]. The kinematics of uniformly accelerated fall that this law makes universal is developed in Chapter 18, and its direct measurement — Galileo's plane among the rest — in Chapter 20.

Newton made the statement quantitative, and he could, because his own dynamics is the first in which it is a statement at all: the *inertial* mass in the second law and the *gravitational* mass coupling a body to attraction are logically independent quantities (Chapter 19), and nothing in the theory forces them to be proportional. His instrument was the pendulum

of Chapter 34, with the two masses now kept carefully distinct: a bob of inertial mass m_{I} and gravitational mass m_{G} on a cord of length ℓ obeys $\ell\,\ddot{\vartheta} = -\,(m_{\mathrm{G}}/m_{\mathrm{I}})\,g\sin\vartheta$, so that for small swings

$$T = 2\pi\sqrt{\frac{m_{\mathrm{I}}}{m_{\mathrm{G}}}\cdot\frac{\ell}{g}}\,. \tag{42.1}$$

Two pendulums of equal length therefore beat together if and only if the ratio $m_{\mathrm{G}}/m_{\mathrm{I}}$ is common to their bobs, a fractional difference in the ratio appearing as half of itself in the period. Newton hung identical wooden boxes filled in turn with gold, silver, lead, glass, sand, salt, water, wood and wheat, swung them in pairs beside one another, and observed no difference in their beats; he concluded that the ratio is common to all these materials to about one part in a thousand [14]. In the modern figure of merit — the Eötvös parameter η of Equation (42.2) below — Galileo's argument bounds nothing numerically, and Newton's pendulums give $|\eta| \lesssim 10^{-3}$.

42.1.2 The Eötvös torsion balance

The next six orders of magnitude came not from refining the pendulum but from abandoning it for a null instrument. A pendulum comparison must difference two separately measured periods, each carrying the full value of g with all its local disturbances; a torsion balance measures the *difference* directly, as a torque that vanishes identically when universality holds. The instrument descends from the Cavendish balance of Chapter 33: a horizontal beam hung from a fine wire, with test bodies of different composition at its ends.

The signal is geometrically subtle, and worth stating because it explains the design. On the rotating Earth a suspended body is acted on by the gravitational attraction, proportional to m_{G}, and by the centrifugal acceleration, proportional to m_{I}; the horizontal component of the latter, about $1.7\times10^{-2}\,\mathrm{m/s^2}$ at 45° of latitude, is what a composition dependence would fail to cancel. Two bodies with different ratios then pull the beam ends toward the equator with different horizontal forces, and the beam turns. The full geometry — the torque, its maximum near 45° of latitude, its reversal under a 180° turn of the beam in azimuth — is derived in Section 53.5, as Equation (53.8). The reversal is the crucial operation: Eötvös performed it by hand, differencing the two orientations so that instrumental torques cancel while a genuine composition dependence changes sign.

The campaign of Eötvös, Pekár and Fekete ran from 1906 to 1909 and was published, posthumously for Eötvös, only in 1922. It compared platinum against copper, water, asbestos and several other materials, and bounded $|\eta|$ at the level of a few parts in 10^9 [486] — six orders of magnitude past Newton, with an instrument whose essential parts are a beam, a wire and a mirror.

42.1.3 Modern limits

Every later improvement has been an improvement in modulation — in making the sought signal vary in a known way while the instrument stays still. The Princeton experiment of Roll, Krotkov and Dicke took the Sun as attractor: a violation would then produce a torque that rotates with the Earth, modulated at the diurnal period with no hand ever touching the balance, and the bound reached $|\eta| \lesssim 10^{-11}$ for aluminium against gold [1294]. The Moscow experiment of Braginsky and Panov refined the same scheme to 10^{-12} [1469]. The Eöt-Wash balance in Seattle rotates continuously on a turntable, so the signal appears at the turntable frequency, far above the drift of the fibre, and bounds $|\eta|$ for beryllium against titanium at 10^{-13} [2135].

The current limit belongs to orbit. The MICROSCOPE satellite compared the free fall of concentric titanium and platinum-alloy test masses around the Earth, drag-free and spin-modulated, where the driving acceleration is the full orbital g rather than a small

horizontal residue of it; the first results bounded $|\eta|$ at 10^{-14} [2289] and the final analysis at parts in 10^{15} [2409]. The instruments, procedures and the full table of bounds are reported in Section 53.5 (Table 53.3).

Phenomenon 42.1 (Universality of free fall)**.** Two bodies of different composition, released at the same place in the same external gravitational field, fall with accelerations that agree to within the precision of every experiment yet performed. Writing the fractional differential acceleration of a pair (A, B) as the Eötvös parameter

$$\eta_{AB} := \frac{2\,(a_A - a_B)}{a_A + a_B}\,, \tag{42.2}$$

every measurement is consistent with $\eta_{AB} = 0$: at the 10^{-9} level in the terrestrial torsion balance of Eötvös, Pekár and Fekete [486]; at 10^{-11} with the Sun as attractor [1294]; at 10^{-12} [1469] and 10^{-13} [2135] in later torsion balances; and at the level of parts in 10^{15} for titanium and platinum test masses compared in orbital free fall [2289, 2409].

Rests on Postulate 19.8.

Derivation. *Derives Phenomenon 42.1.*

Newtonian dynamics carries two logically independent masses: the inertial mass m_{I} multiplying the acceleration in the second law, and the gravitational mass m_{G} that couples the body to the field $\boldsymbol{g}$. The equation of motion $m_{\mathrm{I}}\boldsymbol{a} = m_{\mathrm{G}}\boldsymbol{g}$ gives

$$\boldsymbol{a} = \frac{m_{\mathrm{G}}}{m_{\mathrm{I}}}\,\boldsymbol{g}\,,$$

and both bodies sample the same $\boldsymbol{g}$, so on substituting into Equation (42.2) the field cancels and

$$\eta_{AB} = 2\,\frac{(m_{\mathrm{G}}/m_{\mathrm{I}})_A - (m_{\mathrm{G}}/m_{\mathrm{I}})_B}{(m_{\mathrm{G}}/m_{\mathrm{I}})_A + (m_{\mathrm{G}}/m_{\mathrm{I}})_B}\,.$$

Hence η_{AB} vanishes for every pair of bodies if and only if the ratio $m_{\mathrm{G}}/m_{\mathrm{I}}$ is one and the same number for all matter — in which case a choice of units sets it to unity and the two masses may be identified. Nothing in Newtonian mechanics requires this identification; it is precisely the experimental content of the null results above. Its consequence is that the trajectory of a freely falling body depends on nothing but its initial position and velocity, which is what permits gravitation to be described as a property of spacetime rather than as a force with a body-dependent coupling. □

42.2 The equivalence principle in three strengths

42.2.1 Weak, Einstein and strong forms

Einstein's 1907 review took the null result of the pendulum and torsion experiments and elevated it to a principle: for an observer in free fall the gravitational field is locally absent, so a laboratory falling freely in a uniform field is indistinguishable, by any experiment performed within it, from a laboratory floating in empty space — and, equally, a laboratory at rest in the field is indistinguishable from one uniformly accelerated in the absence of gravity [334]. The principle comes in three strengths, distinguished by *which* experiments are admitted to the comparison, and the distinction is not pedantry: each strength is tested by a different class of measurements and licenses a different part of the geometric description [2240].

Definition 42.2 (The equivalence principle in three strengths)**.** (i) *Weak equivalence principle* (universality of free fall): the trajectory of a freely falling test body — one small enough that tidal gradients and its own gravity are negligible — depends only on its initial

position and velocity, not on its composition or internal structure. (ii) *Einstein equivalence principle*: the weak principle holds, and in addition the outcome of any local *non-gravitational* experiment performed in a freely falling laboratory is independent of the laboratory's velocity (local Lorentz invariance) and of where and when in the universe it is performed (local position invariance). (iii) *Strong equivalence principle*: the Einstein principle extended to bodies with non-negligible gravitational self-energy and to local *gravitational* experiments — self-gravitating bodies fall like test bodies, and the locally measured gravitational coupling is the same everywhere and everywhen [2240].

Each strength buys a definite piece of the theory. The weak principle makes the family of free-fall trajectories a property of spacetime alone, which is what allows them to be declared the geodesics of a connection — the content of Postulate 43.1 in Chapter 43. The Einstein principle forces all non-gravitational physics to couple to one and the same metric, which is what makes the theory a *metric* theory and puts the special relativity of Chapter 38 in every freely falling laboratory. The strong principle is the most restrictive: among the metric theories surviving the first two, it is satisfied by general relativity and violated by nearly every proposed alternative, scalar–tensor theories included, which is why its test — lunar laser ranging, below — discriminates where the others cannot [2240].

Remark 42.3 (Schiff's conjecture). The three strengths may not be as independent as their statements suggest. Schiff observed that the gravitational redshift, nominally a test of the Einstein principle, already follows from the weak principle together with energy conservation, and conjectured that any complete, self-consistent theory of gravity satisfying the weak principle necessarily satisfies the Einstein principle in full [1190]. The conjecture remains unproved in generality, but every concrete framework in which it can be checked supports it; on this argument the composition tests of Phenomenon 42.1, at parts in 10^{15}, are also the sharpest available probes of the Einstein form, since a violation of local position invariance would feed back into composition-dependent binding energies and show as a nonzero η [2240].

In the language of this chapter, then: the torsion balances and MICROSCOPE test the weak form; the redshift and clock comparisons of Section 42.3 test the Einstein form directly, and specifically its position invariance; and lunar laser ranging tests the strong form, through the Nordtvedt effect stated next [1377]. The quantitative bookkeeping in which all these tests are compared — the parametrized post-Newtonian expansion and its parameters γ and β — is taken up in Section 42.8.

42.2.2 Lunar laser ranging

Laboratory masses have no usable self-gravity: the gravitational binding energy of a metre-scale test body is a fraction of order 10^{-27} of its rest energy, hopelessly below any balance. The strong form of the principle therefore needs astronomical test bodies, and the Earth–Moon pair falling toward the Sun is the cleanest available: the Earth stores about 4.6×10^{-10} of its rest energy as gravitational binding, the Moon about 1.9×10^{-11}, so the two differ in the relevant quantity by half a part in 10^{9} while being tracked with millimetre-level precision [2067, 2198].

The tracking is lunar laser ranging: timing pulses against the retroreflector arrays left on the Moon by the Apollo 11, 14 and 15 crews and the two Lunokhod rovers, a programme running continuously since 1969, with the APOLLO instrument now returning normal points at the millimetre level [2198]. The apparatus and procedure are reported in Section 53.6. Two relativistic quantities are extracted from the decades-long range series. The first is the Nordtvedt signal of Phenomenon 42.4 below, whose null result is the strong principle's test. The second is a bound on any secular variation of the gravitational constant: $\dot G/G$ comes out consistent with zero to within about 1×10^{-12}/yr [2067]. That number answers a

question older than the data: Dirac's large-numbers hypothesis, which read the coincidence of the great dimensionless ratios of physics as a law and required G to fall with cosmic age at about that rate [812], is excluded by it — a connection developed with the constants of nature in Chapter 2.

Phenomenon 42.4 (The lunar orbit is not polarized toward the Sun)**.** The Earth and the Moon store different fractions of their mass-energy as gravitational binding energy, so if the gravitational mass of a body depended on that fraction the two would fall toward the Sun at different rates and the lunar orbit would be displaced along the Earth–Sun line, with a signature periodic at the synodic month [1377]. Lunar laser ranging to the retroreflectors left on the surface, with residuals at the centimetre level, sees no such displacement: the Nordtvedt parameter is consistent with zero at the level of parts in 10^4 [2067]. Self-gravitating bodies therefore fall like test bodies, which is the observational content of the strong form of the equivalence principle [2240].

Rests on Phenomenon 42.1 and Theorem 29.51.

Derivation. *Derives Phenomenon 42.4.*

Let the gravitational-to-inertial mass ratio of a body depend on its fractional gravitational self-energy,

$$\frac{m_{\mathrm{G}}}{m_{\mathrm{I}}} = 1 + \eta_{\mathrm{N}} \frac{U}{mc^2}\,, \tag{42.3}$$

where $U < 0$ is the body's Newtonian self-energy and the *Nordtvedt parameter* η_{N} vanishes exactly when self-gravitating bodies obey Phenomenon 42.1 [1377]. For a uniform sphere $|U| = 3GM^2/(5R)$ (Equation (49.23)), which gives $|U|/(Mc^2) = 4.2 \times 10^{-10}$ for the Earth and 1.9×10^{-11} for the Moon, both computed here; interior models concentrate the Earth's mass and raise its value to 4.6×10^{-10} [2067]. The two ratios Equation (42.3) therefore differ by about $4.4 \times 10^{-10}\,\eta_{\mathrm{N}}$, and in the Sun's field at the Earth's orbit, $GM_\odot/r_{\mathrm{ES}}^2 = 5.93 \times 10^{-3}\,\mathrm{m/s^2}$, the Moon falls sunward relative to the Earth with the differential acceleration

$$\delta a = 4.4 \times 10^{-10} \times 5.93 \times 10^{-3}\,\mathrm{m/s^2} \times \eta_{\mathrm{N}} = \eta_{\mathrm{N}}\, 2.6 \times 10^{-12}\,\mathrm{m/s^2}\,,$$

directed along the Earth–Sun line (toward the Sun for $\eta_{\mathrm{N}} > 0$, since the Earth, with the larger binding fraction, then couples the more weakly).

What matters is how the lunar orbit responds. Work in the frame rotating about the Earth with the Moon's mean motion $n = 2.662 \times 10^{-6}/\mathrm{s}$, in which the unperturbed Moon rests on a circular orbit of radius r_0 with $n^2 = \mu/r_0^3$, μ the gravitational parameter of the pair. Linearizing the rotating-frame equation of motion Equation (29.66) about this orbit — gravity contributing $+2n^2\xi$ radially and $-n^2\zeta$ along track, the centrifugal term $+n^2\xi$ and $+n^2\zeta$, the Coriolis term coupling the two — gives, for a radial displacement ξ and an along-track displacement ζ,

$$\ddot{\xi} - 2n\dot{\zeta} - 3n^2\xi = \delta a \cos Dt\,, \qquad \ddot{\zeta} + 2n\dot{\xi} = -\delta a \sin Dt\,. \tag{42.4}$$

The forcing rotates at the synodic frequency $D = n - n_\odot = 2.463 \times 10^{-6}/\mathrm{s}$ (synodic month 29.53 days), because the frame rotates at n while the Earth–Sun direction advances at the annual rate $n_\odot = 1.991 \times 10^{-7}/\mathrm{s}$. The ordinary solar tidal terms are omitted: they are independent of η_{N} and already carried by the Newtonian lunar ephemeris.

Substituting the forced solution $\xi = X\cos Dt$, $\zeta = Z\sin Dt$ into Equation (42.4), the along-track equation gives $Z = (\delta a - 2nDX)/D^2$, and eliminating Z from the radial equation leaves $X(n^2 - D^2) = \delta a\,(1 + 2n/D)$, that is

$$X = \frac{\delta a\,(1 + 2n/D)}{(n - D)\,(n + D)}\,. \tag{42.5}$$

The denominator is the physics: $n - D = n_\odot$, so the forcing lies close to the orbit's own frequency — they differ only by the annual rate — and the response is amplified by more than an order of magnitude over the naive estimate $\delta a/n^2$. Numerically the response coefficient $(1 + 2n/D)\,/\,(n^2 - D^2)$ is $3.1 \times 10^{12}\,\mathrm{s}^2$, so

$$X \simeq \eta_{\mathrm{N}} \times 8.2\,\mathrm{m}\,,$$

computed here from the linear response alone. The full lunar theory does better: the classical solar perturbations of the orbit feed the displacement back at the same frequency and amplify it by roughly a further 1.6, to about 13 m per unit η_{N} — a figure quoted from the detailed computations rather than derived here [1377, 2067]. The signature is a range oscillation $X \cos Dt$, largest along the Earth–Sun line and periodic at the synodic month, which is what separates it from the Newtonian terms in the fit.

Ranging residuals at the level of millimetres to centimetres against a 13 m coefficient then bound the parameter at parts in 10^4: the joint fit of the ranging data gives $\eta_{\mathrm{N}} = (4.4 \pm 4.5) \times 10^{-4}$ [2067]. In the parametrized post-Newtonian bookkeeping of Section 42.8 the parameter is, for a metric theory with no preferred-frame or preferred-location effects, the combination

$$\eta_{\mathrm{N}} = 4\beta - \gamma - 3\,, \tag{42.6}$$

a result imported from the post-Newtonian analysis of massive bodies [1377, 2240]. General relativity has $\beta = \gamma = 1$ and predicts $\eta_{\mathrm{N}} = 0$ identically — the null result above — while any theory in which the effective coupling depends on the local gravitational potential predicts $\eta_{\mathrm{N}} \neq 0$ at some level. □

42.3 Gravitational redshift

42.3.1 The prediction

The redshift was the first new phenomenon Einstein extracted from the equivalence principle, in the 1907 review, and it is the one prediction of this chapter that needs no field equation whatever: it follows from the principle together with kinematics, as the derivation below shows, and Einstein reworked it in 1911 with the optics made explicit — a periodic signal climbing a potential difference $\Delta\Phi$ arrives with its frequency shifted by $-\Delta\Phi/c^2$ [334, 371]. Over a laboratory height H in a uniform field the fraction is gH/c^2; from the surface of the Sun to the Earth it is $GM_\odot/\,(c^2 R_\odot) = 2.12 \times 10^{-6}$, computed here from the solar mass parameter and radius.

Because no field equation enters, a redshift measurement tests the equivalence principle itself — in the classification of Definition 42.2, the local position invariance of the Einstein form — and not the dynamics of gravity. A theory that failed it would not need different field equations but a different foundation; conversely, *any* metric theory passes it, which is why the classical redshift experiments, for all their beauty, do not discriminate between general relativity and its metric rivals [1190, 2240].

The obvious place to look first was the Fraunhofer spectrum of the Sun, and the search there stayed inconclusive for half a century: the solar lines are shifted by pressure and by the Doppler effect of convective granulation by amounts comparable to the sought 2.12×10^{-6}, and the early claims scattered on both sides of the prediction [2240]. The clean confirmation came indoors, from an effect discovered in 1958 that turned a nucleus into a frequency standard of quality 10^{13}.

42.3.2 Pound–Rebka and Pound–Snider

Recoil-free gamma emission — the Mössbauer effect — pins the emitting nucleus to its crystal lattice, so that the recoil momentum is taken up by the crystal as a whole and the

emitted line keeps its natural width, a fractional width of about 3×10^{-13} for the 14.4 keV transition of ^{57}Fe [1154]. That is what made a shift of order 10^{-15} measurable at all. Pound and Rebka sent these gamma rays up and down the 22.5 m tower of the Jefferson Physical Laboratory at Harvard, detected the shift as an asymmetry in resonant absorption, and confirmed the prediction to about ten per cent in 1960 [1189]; the 1965 repetition with Snider reached one per cent [1321].

The decisive systematic is thermal: the second-order Doppler shift of nuclei vibrating in their lattices amounts to -2.4×10^{-15}/K of temperature difference between source and absorber — comparable to the entire gravitational effect per kelvin — and controlling it is the limiting art of the experiment. The apparatus, the reversal procedure that cancels almost everything else, and the derivation of that temperature coefficient (Equation (53.4)) are set out in Section 53.3.

Phenomenon 42.5 (Gravitational redshift)**.** A signal emitted by a source at rest low in a static gravitational field and received by an observer at rest higher up arrives with its frequency lowered. Over a height difference H small enough that the field strength g may be treated as uniform,

$$\frac{\Delta\nu}{\nu} = -\frac{gH}{c^2}\,. \tag{42.7}$$

For the 22.5 m tower of the Harvard experiment this is a fractional shift of 2.5×10^{-15}, resolvable only because the recoil-free emission and absorption of gamma rays by nuclei bound in a lattice [1154] supplies a line of comparable relative width. The measurement confirmed Equation (42.7) to about ten per cent [1189] and, in the 1965 repetition, to about one per cent [1321]. A hydrogen maser carried on a suborbital rocket verified it at the 7×10^{-5} level [1660], and optical clocks now resolve it across a height difference of less than one metre [2154].

Rests on Definition 42.2.

Derivation. *Derives Phenomenon 42.5.*

By the Einstein equivalence principle a laboratory held static in a uniform field g is locally indistinguishable from a laboratory in flat spacetime accelerating with proper acceleration g, and the second is the easier to analyse [334, 371]. Work in the momentarily comoving inertial frame in which the emitter is at rest at the instant of emission. Light leaves the floor and reaches the ceiling a height H above it after a time $t = H/c$, up to corrections of higher order in gH/c^2. During that time the receiver, which is accelerating, has acquired the velocity

$$v = gt = \frac{gH}{c}$$

directed away from the source. The receiver is therefore receding, and to first order in v/c the Doppler formula of Chapter 38 gives

$$\frac{\Delta\nu}{\nu} = -\frac{v}{c} = -\frac{gH}{c^2}\,,$$

which is Equation (42.7). Two features of the argument deserve notice. It uses only the equivalence principle, kinematics and the conservation of energy — no field equation enters — so a redshift measurement tests the equivalence principle rather than the dynamics of the gravitational field; and the sign is fixed by the geometry, an upward-travelling signal always being received redshifted. □

42.3.3 Clocks on rockets and in orbit

Everything since Pound and Snider has been a better clock at a bigger height difference. Gravity Probe A flew a hydrogen maser on a suborbital trajectory to 1×10^7 m altitude in

1976, comparing it continuously against a ground maser through a Doppler-cancelling radio link, and verified the combined gravitational and kinematic frequency shift to 7×10^{-5} [1660] — still the sharpest dedicated test of Equation (42.7) on a terrestrial potential difference. At the opposite extreme of scale, optical clocks of fractional instability below 10^{-17} resolve the shift across laboratory height differences: first 0.33 m between two aluminium-ion clocks [2154], and by 2022 across a single millimetre-scale sample of strontium atoms, mapping Equation (42.7) within one atomic cloud [2401].

Two further measurements belong to this family though they are reported elsewhere. The caesium clocks Hafele and Keating flew around the world in 1972 measured the redshift summed with the kinematic time dilation of Phenomenon 38.13, eastward and westward [1476, 1477]; the flights are treated in full in Section 41.5. And the same shift has now been seen in an astronomical spectrum under a potential four orders of magnitude deeper than the Sun's: the star S2, at its 2018 closest approach to the Galactic-centre black hole, showed the predicted combined redshift [2295], a measurement belonging to Chapter 50. The engineering counterpart of all of these — the satellite-navigation clock budget — closes this chapter in Section 42.7.

42.4 The perihelion of Mercury

42.4.1 Le Verrier's anomaly

The oldest evidence in this chapter predates the theory by fifty-six years, and it was found by accounting. Le Verrier, having predicted Neptune from the misbehaviour of Uranus, applied the same perturbation theory — the secular expansion of Chapter 27 — to Mercury, using the accumulated record of transits across the solar disc. The observed advance of Mercury's perihelion, referred to the moving equinox, is about 5600 seconds of arc per century; the precession of the equinoxes accounts for some 5025 of them, and the Newtonian perturbations of Venus, the Earth, Mars, Jupiter and Saturn for another 531. In 1859 Le Verrier announced that the books do not balance: a residual advance remained, near 38 seconds of arc per century in his reduction [151], moved to the modern 43 by Newcomb's re-reduction with improved planetary masses [267]. The full budget, with the numbers above, stands as Table 53.1 in Section 53.1.

Every Newtonian repair was tried and failed. An intramercurial planet — Vulcan, hunted for decades — would have been seen in transit; a sufficient solar oblateness contradicts the observed figure of the Sun; a modified force-law exponent tuned to Mercury disturbs the Moon [2240]. In the terms of Chapter 1 this is the textbook case of an anomaly held open: a small, stubborn, well-measured discrepancy, resistant to parametric patching, waiting half a century for a theory with no free parameter to claim it.

Phenomenon 42.6 (Anomalous perihelion advance of Mercury)**.** After every Newtonian perturbation by the other planets has been computed and subtracted, the perihelion of Mercury is observed to advance in the direction of the orbital motion at a residual rate of about 43 seconds of arc per century — a discrepancy isolated by Le Verrier in 1859 and left unexplained for half a century [151]. General relativity accounts for it with no adjustable parameter: a bound orbit of semi-major axis a and eccentricity e about a mass M has its perihelion advanced by

$$\Delta\varphi = \frac{6\pi GM}{c^2 a\,(1-e^2)} \tag{42.8}$$

per revolution, to lowest order in $GM/(c^2a)$ [428]. Inserting Mercury's orbital elements and the Sun's mass parameter, and multiplying by the 415 revolutions Mercury completes in a century, reproduces the observed residual.

Rests on Postulate 43.1 and Equation (53.2)*.*

Derivation. *Derives Phenomenon 42.6.*

This is carried out in Section 53.1, in the derivation following Phenomenon 53.1. In outline: the timelike geodesics of the Schwarzschild geometry (Chapter 45) give, for the inverse radius $u = 1/r$ in the orbital plane, the orbit equation Equation (53.2), which is the Newtonian equation of Chapter 27 plus the single correction term $3GMu^2/c^2$, smaller than the Newtonian terms by a factor of order $v^2/c^2 \sim 10^{-7}$ for Mercury. Treating it as a perturbation of the Keplerian ellipse, only its $\cos\varphi$ harmonic is resonant with the orbit and accumulates; the ellipse's apsis line turns by $\Delta\varphi = 6\pi G^2M^2/(c^2h^2)$ per revolution, and eliminating the angular momentum h through $h^2 = GMa\,(1-e^2)$ gives Equation (42.8). Numerically, Mercury's elements yield 5.02×10^{-7} radian per revolution, or 42.98 seconds of arc per century — the residual of Table 53.1. □

42.4.2 Einstein's 1915 computation

Einstein computed the advance in the 18 November 1915 communication to the Prussian Academy, from the vacuum field equations in the approximation that was all he had — the exact Schwarzschild solution was still two months away — and obtained the observed 43 seconds of arc per century with nothing to adjust [428]. The same paper carries the doubled light deflection of Section 42.5, so the one communication contains both the postdiction that validated the field equations and the prediction that would test them four years later.

It must be said plainly, as Section 53.1 says it, that this is a postdiction: the number was known, and Einstein knew what he had to match. It is equally important that this costs the result almost nothing. Equation (42.8) contains no parameter that could have been tuned — the field equations were fixed by general covariance and the Newtonian limit, and the orbit followed — whereas every rival explanation on offer was parametric, and each, tuned to Mercury, then contradicted some other observation. A parameter-free postdiction carries the evidential weight of a prediction; the epistemology is Chapter 1's. The same formula has since been confirmed as a genuine prediction far outside the solar system, in the Schwarzschild precession of the star S2 about the Galactic-centre mass [2347], reported with its instrument in Chapter 50.

42.5 The deflection of light

42.5.1 Half and full values

If the equivalence principle holds, light must fall: a horizontal beam crossing a laboratory that accelerates upward arrives lower on the far wall, so the same must happen in the laboratory held static in a gravitational field. The principle alone even fixes the rate, because whatever falls, falls universally. Treat the ray as a trajectory traversed at speed c in the Newtonian field of a mass M, passing at impact parameter b, and integrate the transverse acceleration $GMb/\left(b^2+x^2\right)^{3/2}$ along the undeflected straight line $x = ct$:

$$\Delta\vartheta_{\mathrm{EP}} = \frac{\Delta v_\perp}{c} = \frac{1}{c^2}\int_{-\infty}^{\infty} \frac{GM\,b\,\mathrm{d}x}{\left(b^2+x^2\right)^{3/2}} = \frac{2GM}{c^2 b}\,. \tag{42.9}$$

For a ray grazing the Sun, $b = R_\odot$, this is 4.24×10^{-6} in radian measure, or 0.87 seconds of arc, computed here. Einstein published exactly this value in 1911, from the equivalence principle dressed as optics — the potential-dependent clock rate of Equation (42.7) makes the coordinate speed of light vary with height, and Huygens' principle turns the gradient into a bending of the wavefront — which is the same physics as the falling corpuscle and gives the same number [371].

The completed theory of November 1915 doubled it, to $4GM/\left(c^2 b\right) = 1.75$ seconds of arc at the limb [428]. The second half has a precise pedigree: the equivalence principle

constrains only the time–time part of the metric, but in the full theory the mass curves space as well, and the two contributions to the bending of a null ray are equal. In the parametrized post-Newtonian bookkeeping the deflection is $(1+\gamma)/2$ times the full value, with γ the space-curvature parameter: $\gamma = 0$ reproduces Equation (42.9), and general relativity's $\gamma = 1$ doubles it [2240]. The factor of two is therefore exactly what separates a metric theory from Newtonian gravity dressed in geometry, and it is what the eclipse expeditions were designed to resolve.

42.5.2 The 1919 eclipse expeditions

The measurement asked for stars photographed close to the eclipsed Sun, compared with the same field photographed months away from it. The Greenwich expeditions of 29 May 1919 — Sobral, in northern Brazil, and Príncipe, in the Gulf of Guinea — were placed under an eclipse of unusual length with the Sun standing in the Hyades, and their announced result favoured the full deflection: 1.98 ± 0.12 seconds of arc at the limb from the Sobral 10 cm lens and 1.61 ± 0.30 from Príncipe, bracketing the predicted 1.75 and both far above 0.87 [465]. The apparatus, the plate reduction, and the celebrated difficulty — the Sobral astrographic plates, set aside for a coelostat mirror that had changed figure in the heat — are reported in Section 53.2 (Table 53.2), together with the modern re-measurement of those same plates, which returned values consistent with the full deflection and showed that the rejection had not biased the conclusion [1640].

The announcement of November 1919 made general relativity a public fact, and the treatise's assessment of its evidential weight is the experiment chapter's: the two accepted determinations were designed to separate hypotheses differing by a factor of two, and they did; the precision claimed was modest, and what settles the matter today is not photography at all.

Phenomenon 42.7 (Deflection of light by the Sun)**.** Starlight passing close to the Sun is deflected toward it, so that the apparent positions of stars seen near the eclipsed limb are displaced radially outward. For a ray of impact parameter b passing a mass M with $b \gg GM/c^2$, the total deflection is

$$\Delta\vartheta = \frac{4GM}{c^2 b}\,, \tag{42.10}$$

which for a ray grazing the solar limb amounts to 8.49×10^{-6} in radian measure, or 1.75 seconds of arc. This is exactly twice the value that the equivalence principle alone yields [371], the second half coming from the curvature of space [428]. The eclipse expeditions of 1919 measured a deflection favouring the full value [465]; radio interferometry on very long baselines now confirms it at the 10^{-4} level [2240]. The same bending, exerted by galaxies and clusters, produces the multiple images of gravitational lenses, of which the twin quasar 0957+561 was the first identified [1645].

Rests on Postulate 43.1 and Equation (53.2)*.*

Derivation. *Derives Phenomenon 42.7.*

The full computation is carried out in Section 53.2, in the derivation following Phenomenon 53.2. In outline: for a null geodesic of the Schwarzschild geometry the orbit equation Equation (53.2) loses its Newtonian constant term — that term came from the rest mass — leaving $\mathrm{d}^2u/\mathrm{d}\varphi^2 + u = 3GMu^2/c^2$; perturbing the straight line $u_0 = \sin\varphi/b$ and following the asymptotes where $u = 0$ shows the ray's two asymptotic directions closing by $4GM/\left(c^2 b\right)$, which is Equation (42.10), and numerically 1.751 seconds of arc at the solar limb.

The origin of the factor of two over Equation (42.9) is cleanest in the isotropic form of the weak-field metric used for the Shapiro derivation in Section 53.4: the potential enters

twice, once in the time–time coefficient and once in the spatial one, and the two retard the coordinate speed of light equally. The equivalence principle fixes only the first — it is the clock-rate effect of Equation (42.7) — and by itself reproduces the half value Equation (42.9); the spatial curvature, absent from any account of gravity as a mere universal acceleration, contributes the equal second half. The parametrized statement is that the deflection carries the factor $(1+\gamma)/2$, and the 1919 result is the measurement that first required $\gamma \neq 0$ [2240]. □

42.5.3 Radio interferometry and gravitational lensing

The modern measurement is radio astrometry. Very-long-baseline interferometry locates compact quasars to fractions of a milliarcsecond, and the Sun's deflection field is not a limb-grazing curiosity but a whole-sky distortion: even a source at right angles to the Sun is displaced by about four milliarcseconds, so global astrometry must carry Equation (42.10) everywhere it points [2240]. Dedicated campaigns following quasars as the Sun passes them measure the space-curvature parameter directly: $\gamma - 1$ consistent with zero at the level of a few parts in 10^4 [1903, 2145], four orders of magnitude past the eclipse plates.

The same bending, scaled up to galaxy and cluster masses, is gravitational lensing. The twin quasar 0957+561 — two images, identical spectra, one source — was identified in 1979 as the first gravitational lens [1645], and lensing has since become an astrophysical instrument in its own right: it weighs the dark matter of Chapter 134, and it magnifies and multiplies the strong-field sources of Chapter 50. In this chapter it stands as the demonstration that Equation (42.10) operates identically from $b \sim R_\odot$ to megaparsec impact parameters.

42.6 The Shapiro time delay

42.6.1 Radar ranging to the inner planets

The fourth classical test was proposed, not by Einstein, but by radar. Shapiro observed in 1964 that a pulse whose path passes close to the Sun takes measurably longer to return than the flat-space geometry predicts — an excess of some 200 µs for an Earth–Venus round trip at superior conjunction, growing logarithmically as the ray approaches the limb [1296]. The logarithmic conjunction profile is the effect's signature: it is not degenerate with an ephemeris error, which varies on the orbital timescale rather than with the impact parameter. The first measurement followed within four years, radar echoes from Mercury and Venus timed at Haystack and Arecibo agreeing with the prediction at the level of tens of per cent [1382]; the transmitters, the coded pulse trains and the conjunction-by-conjunction procedure are reported in Section 53.4.

42.6.2 Viking and Cassini

Replacing the passive planetary echo with a spacecraft transponder sharpened the test by orders of magnitude. Ranging to the Viking landers on Mars through several superior conjunctions confirmed the delay to about one part in 10^3 [1644]. The Cassini spacecraft, tracked on its cruise to Saturn in 2002, ran its radio link simultaneously at multiple widely separated frequencies, using the known dispersion of the solar-corona plasma to solve the corona away — the systematic that had limited every predecessor — and delivered $\gamma - 1 = (2.1 \pm 2.3) \times 10^{-5}$ [2046], the tightest solar-system constraint on any metric theory of gravity. The same delay, imposed by a companion star on the timing of a pulsar, is now a standard tool for weighing neutron stars [2155]; that use belongs to Chapter 49.

Phenomenon 42.8 (Shapiro time delay)**.** A radar pulse sent from Earth to an inner planet and reflected back takes measurably longer when its path passes close to the Sun than when the same geometric distance is traversed far from it. For a round trip between points at radii r_1 and r_2 from a mass M past a closest approach b, with $b \ll r_1, r_2$, the excess is

$$\Delta t = \frac{4GM}{c^3} \ln\!\left(\frac{4r_1 r_2}{b^2}\right) , \tag{42.11}$$

of order 200 µs for an Earth–Venus circuit at superior conjunction. Proposed as a fourth test in 1964 [1296] and first measured against Mercury and Venus [1382], the delay was confirmed to one part in 10^3 by ranging to the Viking landers [1644]. The Cassini radio link measured the post-Newtonian parameter γ, which enters Equation (42.11) through an overall factor $(1+\gamma)/2$, as $\gamma - 1 = (2.1 \pm 2.3) \times 10^{-5}$ [2046].

Rests on Postulate 43.1 and Proposition 45.14.

Derivation. *Derives Phenomenon 42.8.*

Carried out in Section 53.4, in the derivation following Phenomenon 53.4. In outline: write the weak-field Schwarzschild metric in isotropic coordinates, whose spatial part is conformally flat, so that the unperturbed ray is a coordinate straight line and the whole effect sits in the coordinate speed of light, $c\,\mathrm{d}t \simeq \mathrm{d}l\,(1 + 2GM/c^2 r)$ — one factor of $GM/c^2 r$ from time dilation and one from spatial curvature, the same equal split as in the deflection. Integrating the excess along the straight path with $r = \sqrt{b^2 + l^2}$ and doubling for the return trip gives Equation (42.11); for an Earth–Venus circuit grazing the limb the logarithm is 11.8 and $\Delta t = 2.3 \times 10^{-4}$ s. The characteristic weakness of the dependence on b — halving the impact parameter adds only 2.7×10^{-5} s — is why the measurement needs a well-sampled conjunction rather than a single well-placed shot. □

42.7 Relativity in the Global Positioning System

Satellite navigation is a clock experiment run as infrastructure. A receiver fixes its position by differencing the arrival times of signals from satellites whose clocks are synchronized to a common timescale; light covers 30 cm per nanosecond, so metre-level positioning demands nanosecond-level timekeeping, held continuously, by caesium and rubidium standards in orbit [2043]. At that precision both relativistic clock effects of this chapter are not corrections but the largest terms in the error budget: the orbiting clocks sit higher in the Earth's potential than the geoid — the redshift of Phenomenon 42.5 makes them run fast — and they move at orbital speed — the time dilation of Phenomenon 38.13 makes them run slow. The budget below is computed here, and the design of the system confirms it daily.

Phenomenon 42.9 (The relativistic clock budget of satellite navigation)**.** The atomic clocks carried by the satellites of the Global Positioning System, in nearly circular orbits of radius $r = 2.656 \times 10^7$ m, run fast with respect to clocks on the geoid by about 38.5 µs per day. The offset is not a small correction: it is designed into the hardware, the flight oscillators being deliberately set low in frequency before launch, with the residual terms modelled in every receiver. Left unaccounted for it would introduce a ranging error $c\,\Delta t$ growing by about 11 km per day, so that positions would become useless within minutes [2043].

Rests on Phenomena 38.13 and 42.5.

Derivation. *Derives Phenomenon 42.9.*

To first order in Φ/c^2 and v^2/c^2, a clock at Newtonian potential Φ moving with speed v in a static weak field accumulates proper time at the rate

$$\frac{\mathrm{d}\tau}{\mathrm{d}t} = 1 + \frac{\Phi}{c^2} - \frac{v^2}{2c^2} , \tag{42.12}$$

the second term being the gravitational redshift of Phenomenon 42.5 written with $\Phi = gH$, and the third the time dilation of Chapter 38. The two have opposite signs: height makes a clock run fast, motion makes it run slow.

For a circular orbit of radius r about a mass M the Newtonian relations $\Phi = -GM/r$ and $v^2 = GM/r$ hold simultaneously, so the two effects combine into one term,

$$\left.\frac{\mathrm{d}\tau}{\mathrm{d}t}\right|_{\text{orbit}} = 1 - \frac{GM}{rc^2} - \frac{GM}{2rc^2} = 1 - \frac{3GM}{2rc^2}\,,$$

while a clock at rest on the ground at radius R has $\mathrm{d}\tau/\mathrm{d}t = 1 - GM/(Rc^2)$, the Earth's rotation and its equatorial bulge being absorbed into the definition of the geoid. The fractional rate at which the orbiting clock gains is therefore

$$\delta = \frac{GM}{c^2}\left(\frac{1}{R} - \frac{3}{2r}\right). \tag{42.13}$$

With $GM = 3.986 \times 10^{14}\,\mathrm{m^3/s^2}$ for the Earth, $R = 6.371 \times 10^{6}\,\mathrm{m}$ and $r = 2.656 \times 10^{7}\,\mathrm{m}$, the bracket is $1.005 \times 10^{-7}\,\mathrm{m^{-1}}$ and $GM/c^2 = 4.435 \times 10^{-3}\,\mathrm{m}$, so $\delta = 4.46 \times 10^{-10}$. Over one day of 86 400 s the accumulated gain is $3.85 \times 10^{-5}\,\mathrm{s}$. Separating the two contributions, $GM/(Rc^2) - GM/(rc^2)$ gives 45.7 µs per day of gravitational blueshift and $GM/(2rc^2)$ gives 7.2 µs per day of velocity dilation, whose difference is the quoted 38.5 µs. Multiplying by c turns the daily offset into 11.5 km of range. □

The system absorbs the constant rate δ in the factory: the fundamental frequency of 10.23 MHz is set before launch to 10.229 999 995 43 MHz, a fractional offset of -4.4647×10^{-10} [2043], agreeing with the δ computed here to a fraction of a per cent — the residue reflecting the rounded orbital radius used above. Two smaller terms remain for the receiver. An orbit of eccentricity e trades height against speed around each revolution, producing a periodic residual of amplitude $2e\sqrt{GMa}/c^2$ [2043] — about 23 ns at $e = 0.01$, computed here — which every receiver applies from the broadcast elements. And a receiver fixed to the rotating Earth is not in an inertial frame: consistent synchronization requires the Sagnac correction of Section 67.6.4,

$$\Delta t_{\text{Sagnac}} = \frac{2\Omega_{\oplus}}{c^2} A_z\,, \tag{42.14}$$

with A_z the equatorial-plane projection of the area swept by the position vector of the signal, counted positive eastward [2043]. For a receiver on the equator watching a satellite near the horizon this reaches about 133 ns, computed here — 40 m of pseudorange, far above the system's accuracy goals.

What the system tests, in the classification of Definition 42.2, is the Einstein principle's position dependence of clock rates summed with special-relativistic time dilation, continuously and publicly; the field equations proper it does not touch, the next order in $GM/\left(rc^2\right)$ entering at 10^{-19} and accumulating femtoseconds per day, far below the flight clocks' stability. Nor is it a precision test — its residuals answer to engineering tolerances, orders of magnitude short of Gravity Probe A [1660] — but it is the test that never stops running: every position fix is a check that Equation (42.12), with both signs, is physics [2043].

42.8 Beyond the classical tests

The tests of this chapter are compared, across theories, in one common language: the parametrized post-Newtonian (PPN) expansion, which writes the metric of any metric theory, to first order beyond Newton, in terms of a small set of dimensionless parameters

and lets each experiment bound a combination of them [2240]. Two parameters carry nearly everything measured here: γ, the space curvature per unit mass, and β, the nonlinearity of superposition. General relativity has $\gamma = \beta = 1$. The deflection and the Shapiro delay each measure $(1+\gamma)/2$, and Cassini's $\gamma - 1 = (2.1 \pm 2.3) \times 10^{-5}$ is the sharpest entry [2046]; the perihelion advance measures $(2 + 2\gamma - \beta)/3$; and the Nordtvedt combination Equation (42.6) closes the system, lunar laser ranging bounding it at parts in 10^4 [2067]. Taken together the solar-system tests pin γ and β to unity at the level of parts in 10^4 or better [2240].

One sector of the weak field is not touched by any test above: the gravitational effect of *rotation*. A spinning mass drags inertial frames around with it, precessing an orbiting gyroscope about the spin axis [455, 458], on top of the larger geodetic precession that any orbiting gyroscope suffers from the curvature of the static field [1191]. Gravity Probe B measured both against a guide star — the geodetic rate to about 0.3 %, the frame-dragging rate to about twenty per cent [2173] — and lunar laser ranging sees the geodetic precession of the lunar orbit independently [2067]. The rates, the apparatus and the honest accounting of Gravity Probe B's error budget are given in Section 53.6.

Everything above is weak-field: the dimensionless potential $GM/(c^2 r)$ never exceeds a few parts in 10^6 in any solar-system test. The evidence beyond that regime opens with the binary pulsar PSR B1913+16, discovered in 1974 [1550]: a clock in an eight-hour relativistic orbit, whose periastron advances four degrees per year where Mercury manages 43 seconds of arc per century, and whose orbit shrinks by energy lost to gravitational radiation at the rate the quadrupole formula of Chapter 46 predicts — confirmed within a few per cent by 1982 [1703] and to a fraction of a per cent over four decades of timing [2275]. The radiation itself was detected directly in 2015, two black holes of thirty-odd solar masses observed in their final orbits by laser interferometry [2264]; that story is Chapter 47's. The strongest fields of all — orbits, spectra and images at a few gravitational radii — are the subject of Chapter 50.

The order of this part follows the logic these tests establish. The universality of free fall licenses geometry (Chapter 43); the classical tests constrain the field equations that the geometry must obey (Chapter 44); and the chapters after them follow the equations into the strong field, confident at the level of parts in 10^5 where the confidence was earned.

Chapter 43

Geometric Formulation of Gravity

The equivalence principle (Chapter 42) motivates, but does not by itself prove, a specific mathematical statement: that gravity is the geometry of the spacetime manifold itself. This chapter states that geometric formulation precisely, reusing throughout the machinery already built, in full generality, in Chapters 13 and 14. Physical spacetime instantiates that general framework at $p+q=3+1$, torsion set to zero: no experiment to date requires $T^a \neq 0$ (Section 13.13.2), and this treatise's scope rule (Chapter 1) accordingly restricts the physical theory to the Levi-Civita connection.

43.1 Geodesics as the trajectories of free fall

Postulate 43.1 (Geodesic motion)**.** A test particle subject to no force other than gravity follows a timelike geodesic of (M,g): an autoparallel curve of the Levi-Civita connection (Definition 13.146), parametrized by its own proper time τ,

$$\frac{\mathrm{d}^2 x^\lambda}{\mathrm{d}\tau^2} + \mathring{\Gamma}^\lambda{}_{\mu\nu}\,\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\,\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau} = 0\,, \tag{43.1}$$

the specialization of Equation (13.297) to the torsion-free connection of Theorem 13.150.

This is the direct generalization of Newton's first law (Chapter 19): in the absence of gravity, $g_{\mu\nu}=\eta_{\mu\nu}$, $\mathring{\Gamma}=0$, and Equation (43.1) reduces to $\mathrm{d}^2x^\lambda/\mathrm{d}\tau^2=0$, uniform motion. Equation (43.1) also follows from extremizing the proper-time functional $\int \mathrm{d}\tau = \int \sqrt{-g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu}/c\,\mathrm{d}\lambda$ (Remark 13.151), the relativistic analogue of Hamilton's principle (Chapter 21).

Derivation. *Derives Equation* (43.1).

Take the action of a free massive particle proportional to its proper time,

$$S[x] = -mc^2\int \mathrm{d}\tau = -mc\int \mathrm{d}\lambda\,\sqrt{-g_{\mu\nu}(x)\,\dot{x}^\mu\dot{x}^\nu}\,, \tag{43.2}$$

with $\dot{x}^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$ and λ an arbitrary parameter. The Euler–Lagrange equations of Chapter 16 applied to $L=\sqrt{-g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu}$ simplify upon choosing $\lambda=\tau$ after the variation, so that $L=c$ is constant along the curve:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(g_{\lambda\nu}\dot{x}^\nu\right) - \frac{1}{2}\,\partial_\lambda g_{\mu\nu}\,\dot{x}^\mu\dot{x}^\nu = 0\,,$$

i.e. $g_{\lambda\nu}\ddot{x}^\nu + (\partial_\mu g_{\lambda\nu} - \tfrac{1}{2}\partial_\lambda g_{\mu\nu})\dot{x}^\mu\dot{x}^\nu = 0$. Symmetrizing $\partial_\mu g_{\lambda\nu} \to \tfrac{1}{2}(\partial_\mu g_{\lambda\nu} + \partial_\nu g_{\lambda\mu})$ (legitimate under the symmetric contraction with $\dot{x}^\mu\dot{x}^\nu$) and raising with $g^{\sigma\lambda}$ produces exactly the Levi-Civita coefficients of Equation (13.301) with $T=0$, which is Equation (43.1).

For the Newtonian limit take a static weak field, $g_{00} = -(1 + 2\Phi/c^2)$ with $|\Phi|/c^2 \ll 1$, and slow motion, $|\mathrm{d}x^i/\mathrm{d}t| \ll c$, so that with $x^0 = ct$ the four-velocity is dominated by $\mathrm{d}x^0/\mathrm{d}\tau \approx c$. The only Christoffel symbol surviving at leading order in Equation (43.1) is $\mathring{\Gamma}^i{}_{00} = -\frac{1}{2}\,\delta^{ij}\partial_j g_{00} = \partial_i\Phi/c^2$, and the spatial components collapse to

$$\frac{\mathrm{d}[}{\mathrm{d}2}]x^i t = -c^2\,\mathring{\Gamma}^i{}_{00} = -\partial_i\Phi\,,$$

which is Newton's law of gravitation $\ddot{\boldsymbol{x}} = -\nabla\Phi$ (Chapter 19): the geodesic postulate contains Newtonian free fall as its slow-motion, weak-field limit. □

43.2 Curvature and tidal forces

Two neighbouring free-falling geodesics, separated by a small connecting vector ξ^μ, obey the geodesic deviation equation

$$\frac{\mathrm{D}^2\xi^\lambda}{\mathrm{d}\tau^2} = -R^\lambda{}_{\mu\nu\rho}\,\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\,\xi^\nu\,\frac{\mathrm{d}x^\rho}{\mathrm{d}\tau}\,, \tag{43.3}$$

with $\mathrm{D}/\mathrm{d}\tau$ the covariant derivative along the geodesic and $R^\lambda{}_{\mu\nu\rho}$ the Riemann tensor of Theorem 13.152 (torsion-free). Curvature, not the connection, is therefore the geometric object with direct physical content: it is observed as the relative (tidal) acceleration of nearby free-falling bodies, exactly as Newtonian tidal forces are governed by the second derivatives of the potential, $-\partial_i\partial_j\Phi$.

Derivation. *Derives Equation (43.3).*

This is the Jacobi equation Equation (13.309) (Proposition 13.155), instantiated on spacetime with the Levi-Civita connection: in the index order of Equation (13.309), $\mathrm{D}^2\xi^\lambda/\mathrm{d}\tau^2 = R^\lambda{}_{\rho\mu\nu}u^\rho u^\mu\xi^\nu$, and the antisymmetry of the Riemann tensor in its last index pair (Proposition 13.154) rewrites this as $-R^\lambda{}_{\mu\nu\rho}\,u^\mu\xi^\nu u^\rho$, the form quoted in Equation (43.3).

For the Newtonian identification take the static weak field of the previous derivation, $u^\mu \approx (c, \mathbf{0})$, and compute from Equation (13.306): at linear order in Φ/c^2 and for time-independent fields,

$$R^i{}_{0j0} = \partial_j\mathring{\Gamma}^i{}_{00} - \partial_0\mathring{\Gamma}^i{}_{j0} = \frac{\partial_i\partial_j\Phi}{c^2}\,,$$

so Equation (43.3) reduces to

$$\frac{\mathrm{d}[}{\mathrm{d}2}]\xi^i t = -c^2 R^i{}_{0j0}\,\xi^j = -\partial_i\partial_j\Phi\;\xi^j\,,$$

the Newtonian tidal equation: the Riemann component $R^i{}_{0j0}$ *is* the tidal tensor $\partial_i\partial_j\Phi/c^2$, which is the sense of the correspondence asserted above. □

43.3 The Einstein–Hilbert action

The field equations governing $g_{\mu\nu}$ itself (Chapter 44) follow from an action principle, stated here in both the variables of this treatise's double geometric language (Section 13.1): the metric $g_{\mu\nu}$, and the vielbein e^a together with the curvature two-form $R^a{}_b$ of Section 13.12.

Definition 43.2 (Einstein–Hilbert action)**.** In $D = p + q$ dimensions,

$$S_{\mathrm{EH}}[g] = \frac{1}{2\kappa}\int_M \mathrm{d}^D x\,\sqrt{|g|}\;R\,, \tag{43.4}$$

with κ a gravitational coupling constant, fixed in $D = 4$ by the Newtonian limit to $\kappa = 8\pi G/c^4$ (Chapter 44). Equivalently, in vielbein-form variables,

$$S_{\mathrm{EH}}[e,\omega] = \frac{1}{2\kappa\,(D-2)!}\int_M \epsilon_{a_1\ldots a_D}\, R^{a_1 a_2}\wedge e^{a_3}\wedge\cdots\wedge e^{a_D}\,, \tag{43.5}$$

with $\omega^a{}_b = \omega^a{}_b(e)$ the torsion-free connection determined by e (Proposition 13.158).

Theorem 43.3 (Equivalence of the two variable sets)**.** *$S_{\mathrm{EH}}[g] = S_{\mathrm{EH}}[e,\omega]$ for $g_{\mu\nu} = \eta_{ab}e^a{}_\mu e^b{}_\nu$ (Equation (13.264)).*

Proof. This is Equation (A.150), proven in Appendix A.8. □

Remark 43.4 (Why the vielbein form matters). The equivalence of Theorem 43.3 shows that Equation (43.5) is not a mere notational variant: it is the *same* physical action, written in the language in which gravity appears as an $\mathrm{SO}(p,q)$ gauge theory (Remark 13.159) — the language required the moment matter includes spinor fields (Chapter 93), which possess no tensorial transformation law under diffeomorphisms and must instead be coupled through the frame and the spin connection [2259]. Varying Equation (43.5) independently with respect to e^a (with ω eliminated via Proposition 13.158, first-order/Palatini formalism) reproduces the same field equations obtained from varying Equation (43.4) with respect to $g_{\mu\nu}$; that variation, and the resulting field equations, are carried out in full in Section 44.1.3: the metric route is Theorem 44.9, and the independent first-order (vielbein–connection) route is Proposition 44.13.

Chapter 44

The Einstein Field Equations

The field equations relate the curvature of spacetime, constructed in Chapter 43 on the mathematical foundation of Chapter 13, to the energy and momentum of matter. This chapter presents them along Einstein's route of November 1915 [427], derives them in full from the Hilbert action [431], establishes energy–momentum conservation as a consequence of the Bianchi identities rather than a separate postulate, fixes the coupling constant by Newtonian correspondence, and states, with Lovelock, that in four dimensions the equations are essentially unique [1451].

The remaining sections supply what every later chapter of this part draws on: the cosmological constant [445], the energy conditions that power the singularity theorems, the initial-value formulation that makes the theory predictive [999], and an overview of the exact solutions — Schwarzschild, Kerr, Friedmann–Lemaître–Robertson–Walker — whose detailed study is Chapter 45 and Chapter 48.

Conventions, fixed once for the whole chapter: the signature is $(-,+,+,+)$ with coordinates $x^\mu = (x^0, x^i) = (ct, \boldsymbol{x})$, as in Chapter 46 and Appendix A.1; the connection is the torsion-free Levi-Civita connection of Theorem 13.150, written $\Gamma^\lambda{}_{\mu\nu}$ without the ring accent of Chapter 43 since no other connection appears; and the curvature conventions are Equation (13.306) and Equation (13.307), for which the round sphere has $R > 0$. The mostly-minus convention of Chapter 39 differs by $g_{\mu\nu} \to -g_{\mu\nu}$; wherever this chapter leans on a result stated in that convention, the statement used is one whose signature bookkeeping is tracked explicitly (the ϵ discipline of Appendix A.52), and the substitution made is said in place.

44.1 The route to the field equations

44.1.1 Requirements on a relativistic theory of gravity

Newtonian gravity fails relativity twice: the Poisson equation propagates influence instantaneously, and the force law singles out a preferred family of frames. What replaces it cannot be guessed freely; by 1915 the constraints on any candidate were sharp, and it is worth stating them as the specification the field equations must meet, because the rest of this chapter checks them off one by one [435].

1. *The equivalence principle.* Free fall is universal to the accuracy of the Eötvös-class experiments (Phenomenon 42.1), so gravity is not a force field on spacetime but a property of spacetime itself: test bodies follow geodesics of a metric $g_{\mu\nu}$ (Postulate 43.1), and the field variable of the theory is the metric.

2. *General covariance.* The equations must be tensor equations, well formed in every coordinate system, since the equivalence principle deprives any coordinate system of

intrinsic meaning (but see the caveat of Remark 44.1).

3. *Newtonian correspondence.* For weak, static fields and slow motion the theory must reduce to the inverse-square dynamics of Chapters 19 and 27, in the field form of the Poisson equation; this is Section 44.4.

4. *Second-order field equations.* The equations should contain no derivatives of $g_{\mu\nu}$ beyond the second, so that the initial-value problem has the shape of every other field theory in this treatise; higher-derivative dynamics generically suffers the Ostrogradsky instability (Proposition 16.39 and Remark 16.40).

5. *A conserved source.* The source must be the energy–momentum content of matter, described by a symmetric tensor $T_{\mu\nu}$ (Section 44.2), and the equations must be compatible with its local conservation — not impose it as an extra axiom.

These five requirements do not merely motivate the Einstein equations; by Lovelock's theorem (Section 44.7) they very nearly *force* them, up to the two constants κ and Λ fixed in Sections 44.4.2 and 44.6.

Remark 44.1 (The Kretschmann objection). General covariance by itself is physically empty: as Kretschmann observed — the argument circulated in 1917, the printed issue is dated 1918 — *any* theory, Newtonian gravity included, can be rewritten in generally covariant form by carrying enough auxiliary structure along [454]. What distinguishes general relativity is not covariance but the absence of prior geometry: the metric is the *only* spacetime structure, and it is dynamical. The modern assessment of the dispute, and of what Einstein's covariance arguments do and do not establish, is reviewed in [1869]. Requirement (ii) above is therefore to be read together with this stronger condition: no non-dynamical background fields.

44.1.2 The November 1915 papers

The final theory arrived in four communications to the Prussian Academy, one week apart. On 4 November Einstein abandoned the restricted covariance of the 1913 Entwurf theory — whose troubled trajectory, including the hole argument against general covariance that he first accepted and then dismantled, is traced in [1869] — and proposed equations covariant under unimodular coordinate transformations [429]. On 11 November an addendum sharpened the hypothesis on the trace of the source [430]. On 18 November he applied the vacuum equations — which the later correction would not touch — to the perihelion of Mercury and obtained the observed 43 seconds of arc per century with no adjustable parameter [428], the computation whose experimental content is Chapter 53. On 25 November the trace term fell into place and the equations took their final, generally covariant form [427]. The episode is a working example of the methodology of Chapter 1: two years of false steps constrained by one hard number.

Definition 44.2 (Einstein tensor)**.** The *Einstein tensor* of (M, g) is

$$G_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{2}\, g_{\mu\nu} R\,, \tag{44.1}$$

with $R_{\mu\nu}$ and R the Ricci contractions of Definition 13.153. It is symmetric, carries SI dimension $1/\mathrm{m}^2$, and is divergence-free as an identity (Theorem 44.23).

Rests on Definition 13.153.

Postulate 44.3 (Einstein field equations)**.** The metric of spacetime is determined by the energy–momentum content of matter through

$$G_{\mu\nu} = \frac{8\pi G}{c^4}\, T_{\mu\nu}\,, \tag{44.2}$$

with $T_{\mu\nu}$ the stress–energy tensor of Section 44.2 [427]. The coupling constant $\kappa = 8\pi G/c^4$ is fixed by Newtonian correspondence in Section 44.4.2.

Rests on Definition 44.2.

Equation (44.2) is ten coupled, quasilinear, second-order partial differential equations for the ten components of $g_{\mu\nu}$ — of which only six are independent, by the four contracted Bianchi identities (Remark 44.25). The same physical postulate can be packaged as an action principle, which is the subject of Section 44.1.3; the postulate proper is then the choice of action.

Proposition 44.4 (Trace-reversed form)**.** *Equation* (44.2) *is equivalent to*

$$R_{\mu\nu} = \kappa \left(T_{\mu\nu} - \tfrac{1}{2}\, g_{\mu\nu} T \right) , \qquad T = g^{\alpha\beta} T_{\alpha\beta} \, . \tag{44.3}$$

In particular the vacuum equations are $R_{\mu\nu} = 0$: vacuum solutions are Ricci-flat, not flat.

Rests on Postulate 44.3.

Proof. *Derives Proposition 44.4.*

Tracing Equation (44.2) with $g^{\mu\nu}$ in $D = 4$ gives $R - 2R = \kappa T$, so $R = -\kappa T$. Substituting back, $R_{\mu\nu} = \kappa T_{\mu\nu} + \frac{1}{2} g_{\mu\nu} R = \kappa (T_{\mu\nu} - \frac{1}{2} g_{\mu\nu} T)$, and the two forms determine each other. With $T_{\mu\nu} = 0$ both give $R_{\mu\nu} = 0$, while the Riemann tensor need not vanish — the Schwarzschild exterior (Section 44.10) is the standing example. This trace-reversed form is the one Einstein printed on 25 November [427], and it is the form the Newtonian limit reads most easily (Section 44.4.2). □

44.1.3 The Hilbert action

Five days before the final field equations, Hilbert presented an action from which they follow by variation [431]: the Einstein–Hilbert action of Definition 43.2, here instantiated at $D = 4$,

$$S_{\mathrm{EH}}[g] = \frac{1}{2\kappa} \int_{\Omega} \mathrm{d}^4 x \, \sqrt{|g|} \, R \, , \tag{44.4}$$

together with a matter action $S_{\mathrm{m}}[\psi; g]$ for the matter fields ψ. This subsection carries out the variation in full — it is the derivation deferred to this chapter by Remark 43.4 — including the boundary term that the usual compact-support argument hides, and the first-order variation in vielbein-form variables.

Remark 44.5 (SI bookkeeping for the action). With $x^0 = ct$ the measure $\mathrm{d}^4 x$ carries m^4, R carries $1/\mathrm{m}^2$ and $\kappa = 8\pi G/c^4$ carries $\mathrm{s}^2/(\mathrm{kg\,m})$, so Equation (44.4) carries J m: it is c times a conventional action in J s, because the time coordinate is measured in metres. Extremals are unaffected by the constant factor. The same convention is applied to S_{m}, so that the tensor defined next comes out in $\mathrm{J/m}^3$, the SI unit of energy density.

Definition 44.6 (Stress–energy tensor)**.** Let $S_{\mathrm{m}}[\psi; g]$ be the action of the matter fields, covariantly coupled to the metric and carrying no other spacetime structure. The *stress–energy tensor* of the matter is its metric variational derivative,

$$T_{\mu\nu} := -\frac{2}{\sqrt{|g|}} \frac{\delta S_{\mathrm{m}}}{\delta g^{\mu\nu}} \, , \qquad \text{equivalently} \qquad \delta S_{\mathrm{m}} = -\frac{1}{2} \int \mathrm{d}^4 x \, \sqrt{|g|} \, T_{\mu\nu} \, \delta g^{\mu\nu} \, , \tag{44.5}$$

the functional derivative being the field-theoretic first variation of Theorem 16.36. $T_{\mu\nu}$ is symmetric by construction, and $T_{\mu\nu}$ carries $\mathrm{J/m}^3$: T_{00} is an energy density, cT_{0i} an energy flux, T_{ij} a stress.

Rests on Theorem 16.36.

Two variational lemmas do all the work. Both are pure algebra plus one classical determinant identity.

Lemma 44.7 (Variation of inverse and determinant)**.** *Under a variation $\delta g_{\mu\nu}$ of the metric,*

$$\delta g^{\mu\nu} = -\, g^{\mu\alpha} g^{\nu\beta}\, \delta g_{\alpha\beta}\,, \tag{44.6}$$

$$\delta\sqrt{|g|} = \tfrac{1}{2}\sqrt{|g|}\; g^{\mu\nu}\, \delta g_{\mu\nu} = -\tfrac{1}{2}\sqrt{|g|}\; g_{\mu\nu}\, \delta g^{\mu\nu}\,. \tag{44.7}$$

Rests on Proposition 5.1.

Proof. *Derives Lemma 44.7.*

Varying $g^{\mu\alpha} g_{\alpha\nu} = \delta^\mu_\nu$ gives $\delta g^{\mu\alpha}\, g_{\alpha\nu} = -g^{\mu\alpha}\, \delta g_{\alpha\nu}$; contracting with $g^{\nu\beta}$ yields Equation (44.6). For the determinant, Jacobi's formula Equation (5.21) applied to the matrix $g_{\mu\nu}$ along any one-parameter family gives $\delta(\det g) = (\det g)\, g^{\mu\nu}\delta g_{\mu\nu}$, and since $|g| = -\det g$ here, $\delta\sqrt{|g|} = \tfrac{1}{2}\,|g|^{-1/2}\;\delta\,|g| = \tfrac{1}{2}\sqrt{|g|}\, g^{\mu\nu}\delta g_{\mu\nu}$. The second equality in Equation (44.7) follows from Equation (44.6). □

Lemma 44.8 (Palatini identity)**.** *The variation of the Levi-Civita connection is the tensor*

$$\delta\Gamma^\lambda{}_{\mu\nu} = \tfrac{1}{2} g^{\lambda\rho}\left(\nabla_\mu \delta g_{\rho\nu} + \nabla_\nu \delta g_{\rho\mu} - \nabla_\rho \delta g_{\mu\nu}\right)\,, \tag{44.8}$$

and the variation of the Ricci tensor is its covariant curl,

$$\delta R_{\mu\nu} = \nabla_\lambda\, \delta\Gamma^\lambda{}_{\nu\mu} - \nabla_\nu\, \delta\Gamma^\lambda{}_{\lambda\mu}\,. \tag{44.9}$$

Consequently the trace is a pure divergence,

$$g^{\mu\nu}\, \delta R_{\mu\nu} = \nabla_\mu v^\mu\,, \qquad v^\mu = g^{\alpha\beta}\, \delta\Gamma^\mu{}_{\alpha\beta} - g^{\mu\beta}\, \delta\Gamma^\lambda{}_{\lambda\beta}\,, \tag{44.10}$$

with, explicitly in terms of the metric variation,

$$v^\mu = g^{\mu\rho} g^{\alpha\beta}\left(\nabla_\alpha \delta g_{\rho\beta} - \nabla_\rho \delta g_{\alpha\beta}\right)\,. \tag{44.11}$$

Rests on Theorems 13.150 and 13.152 and Definition 13.149.

Proof. *Derives Lemma 44.8.*

Step 1: Equation (44.8). Vary the Christoffel formula of Theorem 13.150, $\Gamma^\lambda{}_{\mu\nu} = \tfrac{1}{2} g^{\lambda\rho}(\partial_\mu g_{\rho\nu} + \partial_\nu g_{\rho\mu} - \partial_\rho g_{\mu\nu})$. The $\delta g^{\lambda\rho}$ term contributes, by Equation (44.6) and the same formula read backwards, $-g^{\lambda\alpha}\,\Gamma^\sigma{}_{\mu\nu}\,\delta g_{\alpha\sigma}$. In the $\delta(\partial g)$ terms replace each $\partial\,\delta g$ by $\nabla\,\delta g$ plus its two connection terms; summing the three combinations, four of the six connection terms cancel in pairs by the symmetry of Γ, and the surviving two give $+g^{\lambda\rho}\,\Gamma^\sigma{}_{\mu\nu}\,\delta g_{\rho\sigma}$, which cancels the $\delta g^{\lambda\rho}$ contribution exactly. What remains is Equation (44.8); being built covariantly from the tensor $\delta g_{\mu\nu}$, it is a tensor — as the difference of two connections must be.

Step 2: Equation (44.9). Vary Equation (13.306):

$$\delta R^\lambda{}_{\rho\mu\nu} = \partial_\mu \delta\Gamma^\lambda{}_{\nu\rho} - \partial_\nu \delta\Gamma^\lambda{}_{\mu\rho} + \delta\Gamma^\lambda{}_{\mu\sigma}\Gamma^\sigma{}_{\nu\rho} + \Gamma^\lambda{}_{\mu\sigma}\delta\Gamma^\sigma{}_{\nu\rho} - \delta\Gamma^\lambda{}_{\nu\sigma}\Gamma^\sigma{}_{\mu\rho} - \Gamma^\lambda{}_{\nu\sigma}\delta\Gamma^\sigma{}_{\mu\rho}\,.$$

On the other hand, treating $\delta\Gamma$ as the $(1,2)$ tensor it is,

$$\nabla_\mu \delta\Gamma^\lambda{}_{\nu\rho} - \nabla_\nu \delta\Gamma^\lambda{}_{\mu\rho} = \partial_\mu \delta\Gamma^\lambda{}_{\nu\rho} - \partial_\nu \delta\Gamma^\lambda{}_{\mu\rho} + \Gamma^\lambda{}_{\mu\sigma}\delta\Gamma^\sigma{}_{\nu\rho} - \Gamma^\lambda{}_{\nu\sigma}\delta\Gamma^\sigma{}_{\mu\rho} + \Gamma^\sigma{}_{\nu\rho}\delta\Gamma^\lambda{}_{\mu\sigma} - \Gamma^\sigma{}_{\mu\rho}\delta\Gamma^\lambda{}_{\nu\sigma}\,,$$

the terms in $\Gamma^\sigma{}_{\mu\nu}\delta\Gamma^\lambda{}_{\sigma\rho}$ cancelling between the two covariant derivatives because $\Gamma^\sigma{}_{\mu\nu}$ is symmetric. The right-hand sides agree term by term, so $\delta R^\lambda{}_{\rho\mu\nu} = \nabla_\mu \delta\Gamma^\lambda{}_{\nu\rho} - \nabla_\nu \delta\Gamma^\lambda{}_{\mu\rho}$; contracting λ with μ gives Equation (44.9).

Step 3: Equation (44.10). Contract Equation (44.9) with $g^{\mu\nu}$ and pull the metric through ∇ by metricity (Definition 13.149); the first expression for v^μ results. Substituting Equation (44.8): in the first slot, $g^{\alpha\beta}\delta\Gamma^\mu{}_{\alpha\beta} = g^{\mu\rho}g^{\alpha\beta}(\nabla_\alpha \delta g_{\rho\beta} - \tfrac{1}{2}\nabla_\rho \delta g_{\alpha\beta})$; in the second, $\delta\Gamma^\lambda{}_{\lambda\beta} = \tfrac{1}{2} g^{\lambda\rho}\nabla_\beta \delta g_{\rho\lambda}$ because the first and third terms of Equation (44.8) cancel under the $\lambda\rho$ contraction. Combining the two, the half-terms add up and give Equation (44.11). □

Theorem 44.9 (Variation of the Einstein–Hilbert action)**.** *For any variation $\delta g_{\mu\nu}$,*

$$\delta S_{\mathrm{EH}} = \frac{1}{2\kappa}\int_\Omega \mathrm{d}^4x\,\sqrt{|g|}\,G_{\mu\nu}\,\delta g^{\mu\nu} + \frac{1}{2\kappa}\int_\Omega \mathrm{d}^4x\,\sqrt{|g|}\,\nabla_\mu v^\mu\,, \tag{44.12}$$

with v^μ as in Equation (44.10). For variations supported in a compact subset of the interior of Ω the second integral vanishes, and stationarity of $S_{\mathrm{EH}} + S_{\mathrm{m}}$ for all such variations is equivalent to the field equations Equation (44.2),

$$G_{\mu\nu} = \kappa\, T_{\mu\nu}\,. \tag{44.13}$$

In vacuum (S_{m} absent) the equations are $G_{\mu\nu} = 0$, equivalently $R_{\mu\nu} = 0$ [427, 431].
Rests on Lemmas 44.7 and 44.8 and Definition 44.6.

Proof. *Derives Theorem 44.9.*

Write $\sqrt{|g|}\,R = \sqrt{|g|}\,g^{\mu\nu}R_{\mu\nu}$ and vary the three factors:

$$\delta\Big(\sqrt{|g|}\,R\Big) = \delta\sqrt{|g|}\;R + \sqrt{|g|}\;\delta g^{\mu\nu}\,R_{\mu\nu} + \sqrt{|g|}\;g^{\mu\nu}\,\delta R_{\mu\nu}\,.$$

By Equation (44.7) the first term is $-\frac{1}{2}\sqrt{|g|}\,g_{\mu\nu}R\,\delta g^{\mu\nu}$, which combines with the second into $\sqrt{|g|}\,G_{\mu\nu}\,\delta g^{\mu\nu}$; by Equation (44.10) the third is $\sqrt{|g|}\,\nabla_\mu v^\mu$. That is Equation (44.12).

For the divergence term, note the identity $\Gamma^\lambda{}_{\lambda\mu} = \frac{1}{2}g^{\lambda\rho}\partial_\mu g_{\lambda\rho} = \partial_\mu \ln\sqrt{|g|}$ — Jacobi's formula Equation (5.21) again — whence $\sqrt{|g|}\,\nabla_\mu v^\mu = \partial_\mu(\sqrt{|g|}\,v^\mu)$: a coordinate divergence. If $\delta g_{\mu\nu}$, and with it v^μ, vanishes outside a compact subset of the interior, the integral of $\partial_\mu(\sqrt{|g|}\,v^\mu)$ vanishes by the fundamental theorem of calculus applied coordinate by coordinate (Theorem 7.101 is its three-dimensional integral form). Adding δS_{m} from Equation (44.5),

$$\delta\left(S_{\mathrm{EH}} + S_{\mathrm{m}}\right) = \int_\Omega \mathrm{d}^4x\,\sqrt{|g|}\left[\frac{1}{2\kappa}G_{\mu\nu} - \frac{1}{2}T_{\mu\nu}\right]\delta g^{\mu\nu}\,,$$

and since $\delta g^{\mu\nu}$ is an arbitrary symmetric compactly supported tensor, the fundamental lemma of the calculus of variations in its several-variable form (Lemma 16.18, as used in Theorem 16.36) gives Equation (44.13). The vacuum case and the equivalence $G_{\mu\nu} = 0 \Leftrightarrow R_{\mu\nu} = 0$ are Proposition 44.4. □

The compact-support restriction is not a technicality to be waved away: on a region with boundary the discarded divergence is a genuine obstruction, and its repair is a boundary action in its own right.

Lemma 44.10 (Divergence theorem on (M, g))**.** *Let Ω be a compact coordinate region whose boundary $\partial\Omega$ is piecewise smooth and nowhere null, with outward unit conormal n_μ ($n^\mu n_\mu = \epsilon = \pm 1$), induced metric $h_{\alpha\beta}$ and area element $\sqrt{|h|}\,\mathrm{d}^3y$. Then for any C^1 vector field v^μ,*

$$\int_\Omega \mathrm{d}^4x\,\sqrt{|g|}\,\nabla_\mu v^\mu = \oint_{\partial\Omega} \mathrm{d}^3y\,\sqrt{|h|}\;n_\mu v^\mu\,. \tag{44.14}$$

Rests on Theorem 7.101 and Proposition 5.1.

Proof. *Derives Lemma 44.10.*

As in Theorem 44.9, $\sqrt{|g|}\,\nabla_\mu v^\mu = \partial_\mu(\sqrt{|g|}\,v^\mu)$, and the flat divergence theorem (Theorem 7.101, in its four-dimensional form, which is proved by the identical slicing argument) turns the left side into the flux of $\sqrt{|g|}\,v^\mu$ through $\partial\Omega$. It remains to identify the flux element. On a boundary face given by $x^1 = \text{const}$ (outward direction of increasing x^1, the other faces being identical), the flux is $\int \mathrm{d}^3y\,\sqrt{|g|}\,v^1$. By the adjugate formula for the inverse matrix (Proposition 5.47), g^{11} equals the $(1,1)$ cofactor of $g_{\mu\nu}$ divided by $\det g$; the cofactor is the determinant of the 3×3 block of g on the coordinates (x^0, x^2, x^3) of the face, which is exactly $\det h$. Hence $|g|\;|g^{11}| = |h|$, so $\sqrt{|g|} = \sqrt{|h|}/\sqrt{|g^{11}|}$ — finite precisely because the face is not null, $g^{11} \neq 0$. The outward unit conormal is $n_\mu = \partial_\mu x^1/\sqrt{|g^{11}|}$, with $n^\mu n_\mu = g^{11}/\,|g^{11}| = \epsilon$, and $n_\mu v^\mu = v^1/\sqrt{|g^{11}|}$. Therefore $\sqrt{|g|}\,v^1 = \sqrt{|h|}\;n_\mu v^\mu$, which is Equation (44.14). □

Proposition 44.11 (The boundary term, and the Gibbons–Hawking–York action)**.** *Let $\partial\Omega$ be nowhere null, and consider variations with $\delta g_{\mu\nu} = 0$ on $\partial\Omega$ but with unconstrained normal derivatives there — the boundary data of a second-order variational problem. Then the divergence term of Equation (44.12) does* not *vanish:*

$$n_\mu v^\mu\big|_{\partial\Omega} = -\,h^{\alpha\beta}\,n^\rho\,\partial_\rho \delta g_{\alpha\beta} = -\,2\,\delta K\,, \tag{44.15}$$

with $K = h^{\alpha\beta}\nabla_\alpha n_\beta$ the trace of the extrinsic curvature of $\partial\Omega$. Consequently the action with a well-defined variational principle on Ω is not S_{EH} but

$$S_{\mathrm{EH}} + S_{\mathrm{GHY}}\,, \qquad S_{\mathrm{GHY}} = \frac{1}{\kappa}\oint_{\partial\Omega} \mathrm{d}^3y\,\sqrt{|h|}\;K\,, \tag{44.16}$$

whose total variation under these boundary conditions is $(2\kappa)^{-1}\int\sqrt{|g|}\,G_{\mu\nu}\,\delta g^{\mu\nu}$ [1490, 1596].
Rests on Theorem 44.9 and Lemma 44.10.

Proof. *Derives Proposition 44.11.*

On $\partial\Omega$ every term of v^μ in Equation (44.11) may be written with partial derivatives, since $\nabla\delta g = \partial\,\delta g - \Gamma\,\delta g$ and $\delta g = 0$ there. Decompose $g^{\alpha\beta} = h^{\alpha\beta} + \epsilon^{-1} n^\alpha n^\beta$ along the boundary. In $n_\mu v^\mu = g^{\alpha\beta} n^\rho(\partial_\alpha \delta g_{\rho\beta} - \partial_\rho \delta g_{\alpha\beta})$ the two $n^\alpha n^\beta$ contributions are equal and cancel between the two terms; in the $h^{\alpha\beta}$ part, the first term is a tangential derivative of δg, which vanishes because δg vanishes identically along $\partial\Omega$. That is the first equality of Equation (44.15).

For the second, vary $K = h^{\alpha\beta}\nabla_\alpha n_\beta$ at fixed boundary metric. Since $\delta g_{\mu\nu} = 0$ on $\partial\Omega$, the unit conormal $n_\mu \propto \partial_\mu(\text{boundary function})$ and its normalization are unvaried there, and so is $h^{\alpha\beta}$; moreover δn_β vanishes identically *along* $\partial\Omega$, so its h-projected (tangential) derivatives vanish too, and the only surviving variation is through the connection, $\delta(\nabla_\alpha n_\beta) = -\delta\Gamma^\sigma{}_{\alpha\beta}\,n_\sigma$. With Equation (44.8), again dropping tangential derivatives of δg,

$$\delta K = -\tfrac{1}{2}\,h^{\alpha\beta}n^\rho\,(\partial_\alpha\delta g_{\rho\beta} + \partial_\beta\delta g_{\rho\alpha} - \partial_\rho\delta g_{\alpha\beta}) = +\tfrac{1}{2}\,h^{\alpha\beta}n^\rho\,\partial_\rho\delta g_{\alpha\beta} = -\tfrac{1}{2}\,n_\mu v^\mu\,.$$

By Lemma 44.10 the offending term of Equation (44.12) is $(2\kappa)^{-1}\oint\sqrt{|h|}\,n_\mu v^\mu = -\kappa^{-1}\oint\sqrt{|h|}\,\delta K = -\,\delta S_{\mathrm{GHY}}$, using once more that $\sqrt{|h|}$ is unvaried on the boundary. Adding S_{GHY} therefore cancels it identically. □

Remark 44.12 (Reading the boundary term). The obstruction arises because R contains second derivatives of the metric, so Equation (44.4) is a second-order Lagrangian whose boundary variation involves $\partial_n\,\delta g$; the Gibbons–Hawking–York term subtracts exactly the offending total derivative. The same divergence appears, from the opposite direction, in the $3+1$ split: one of the two total derivatives discarded in the ADM decomposition Equation (26.80) of Proposition 26.41 is this boundary term, as identified in Appendix A.52. Three honest caveats. The construction above assumes a nowhere-null boundary; null segments and the corners where segments meet require refinements that this treatise does not need. On non-compact regions the subtraction of a reference term is needed to make Equation (44.16) finite — that subtraction is where the ADM energy of Remark 26.45 enters. And the conventions here use the outward unit *conormal* throughout; the literature's explicit factors of ϵ [1596] appear when the future-directed normal vector is used on spacelike segments instead, and the two bookkeepings agree.

Proposition 44.13 (First-order variation in vielbein-form variables)**.** *Treat, in the vielbein form Equation (43.5) of the action at $D = 4$,*

$$S[e,\omega] = \frac{1}{4\kappa}\int_M \epsilon_{abcd}\,R^{ab}\wedge e^c\wedge e^d\,, \tag{44.17}$$

the vielbein e^a and the connection ω^{ab} as independent *fields (first-order or Palatini formalism), with compactly supported variations. Stationarity with respect to ω is equivalent to*

the vanishing of the torsion, $T^a = 0$ — so that $\omega = \omega(e)$ by Proposition 13.158 — and stationarity with respect to e then reproduces the vacuum Einstein equations $G_{\mu\nu} = 0$ of Theorem 44.9.

Rests on Theorems 13.156 and 43.3 and Proposition 13.158.

Proof. *Derives Proposition 44.13.*

The ω equation. From the second structure equation Equation (13.312), $\delta R^{ab} = \mathrm{d}\,\delta\omega^{ab} + \delta\omega^a{}_c \wedge \omega^{cb} + \omega^a{}_c \wedge \delta\omega^{cb} = \mathrm{D}\,\delta\omega^{ab}$. The frame Levi-Civita symbol is D-constant: $\mathrm{D}\epsilon_{abcd}$ is totally antisymmetric in $abcd$, hence proportional to ϵ_{abcd}, with coefficient the trace $\omega^e{}_e = \eta_{ef}\,\omega^{ef} = 0$ by the antisymmetry Equation (13.310). The graded Leibniz rule for D, applied to the one-form $\delta\omega^{ab}$, therefore gives

$$\epsilon_{abcd}\,\mathrm{D}\delta\omega^{ab} \wedge e^c \wedge e^d = \mathrm{d}\left(\epsilon_{abcd}\,\delta\omega^{ab} \wedge e^c \wedge e^d\right) + \epsilon_{abcd}\,\delta\omega^{ab} \wedge \mathrm{D}\left(e^c \wedge e^d\right) ,$$

where d replaces D on the full contraction because it is a Lorentz scalar. The exact term integrates to zero, so $\delta_\omega S = (4\kappa)^{-1}\int \epsilon_{abcd}\,\delta\omega^{ab} \wedge \mathrm{D}(e^c \wedge e^d)$. With $\mathrm{D}(e^c \wedge e^d) = T^c \wedge e^d - e^c \wedge T^d$ (first structure equation, Equation (13.311)) and the relabelling symmetry of ϵ_{abcd}, the integrand equals $2\,\epsilon_{abcd}\,\delta\omega^{ab} \wedge T^c \wedge e^d$, so the field equation is $\epsilon_{abcd}\,T^c \wedge e^d = 0$ for every pair ab.

To see that this forces $T^a = 0$, expand $T^c = \frac{1}{2}T^c{}_{ef}\,e^e \wedge e^f$ and wedge the equation with e^g; using $e^e \wedge e^f \wedge e^d \wedge e^g = \epsilon^{efdg}\,e^0 \wedge e^1 \wedge e^2 \wedge e^3$ with the numerical symbol, the content is $\epsilon_{abcd}\,\epsilon^{defg}\,T^c{}_{ef} = 0$. The numerical contraction identity $\epsilon^{defg}\epsilon_{dabc} = \delta^{efg}_{abc}$ (the 3×3 determinant of Kronecker deltas; Lemma A.50 is its twice-contracted consequence) reduces this, after the six contractions, to

$$T^g{}_{ab} = \delta^g{}_b\,T^c{}_{ac} - \delta^g{}_a\,T^c{}_{bc} .$$

Tracing g with b gives $T^c{}_{ac} = 4T^c{}_{ac} - T^c{}_{ac}$, so the trace vanishes, and with it the whole tensor: $T^a = 0$.

The e equation. On the surface $T^a = 0$ the connection is the unique torsion-free $\omega(e)$ of Proposition 13.158, and substituting it back gives the second-order functional $S[e, \omega(e)]$, which by Theorem 43.3 equals $S_{\mathrm{EH}}[g]$ with $g_{\mu\nu} = \eta_{ab}e^a{}_\mu e^b{}_\nu$. Now vary e in the first-order action at $\omega = \omega(e)$: the chain rule splits the variation into the explicit δ_e at fixed ω and the induced δ_ω evaluated on $\delta\omega(e)$ — and the latter vanishes, because $\omega(e)$ is precisely the point where $\delta_\omega S = 0$. So the first-order e equation coincides with the total variation of $S_{\mathrm{EH}}[g(e)]$, which by Theorem 44.9 and $\delta g_{\mu\nu} = \eta_{ab}(e^a{}_\mu\,\delta e^b{}_\nu + e^a{}_\nu\,\delta e^b{}_\mu)$ is $\kappa^{-1}\int \sqrt{|g|}\,G^{\mu\nu}\,\eta_{ab}e^b{}_\nu\,\delta e^a{}_\mu\,\mathrm{d}^4x$. Since the vielbein is invertible, stationarity for all $\delta e^a{}_\mu$ is equivalent to $G^{\mu\nu} = 0$. □

Remark 44.14 (The priority question). Hilbert's Göttingen communication is dated 20 November 1915, five days before Einstein's [427, 431], and for eighty years the field equations were sometimes said to have been found first by Hilbert. The printer's proofs of Hilbert's paper, recovered in 1997, show that the explicit field equations were not present in the version submitted on 20 November and entered only in revision, after Einstein's paper had appeared; the published version also carries a December revision date [1937]. What is securely Hilbert's is the action route itself — the recognition that Equation (44.4) is the scalar whose variation does the work, which this section has followed.

44.2 The stress–energy tensor

44.2.1 Matter models

Definition 44.6 assigns every covariantly coupled matter action a symmetric stress–energy tensor. Three models carry all the matter physics this part needs [1509, 1552]; their special-relativistic particle content — four-momentum and its conservation — is Chapter 40.

Definition 44.15 (Dust)**.** A pressureless fluid — *dust* — with proper mass density ρ (in $\mathrm{kg/m^3}$) and four-velocity field u^μ ($u^\mu u_\mu = -c^2$, Definition 40.2) has

$$T^{\mu\nu} = \rho\, u^\mu u^\nu \,. \tag{44.18}$$

In the local rest frame $T^{00} = \rho c^2$ — the rest energy density of Phenomenon 40.6 — and all other components vanish.

Rests on Definition 40.2 and Phenomenon 40.6.

Definition 44.16 (Perfect fluid)**.** A fluid with proper energy density ρc^2 (internal energy included), proper pressure p (in Pa), and four-velocity u^μ, but no viscosity and no heat flux, has

$$T^{\mu\nu} = \left(\rho + \frac{p}{c^2}\right) u^\mu u^\nu + p\, g^{\mu\nu} \,. \tag{44.19}$$

In the local rest frame $T^{\mu\nu} = \mathrm{diag}(\rho c^2, p, p, p)$: isotropic pressure, no energy flux. Dust is the case $p = 0$; radiation the case $p = \rho c^2/3$, for which the trace $T = -\rho c^2 + 3p$ vanishes.

Rests on Definition 44.15.

Proposition 44.17 (Stress–energy of the electromagnetic field)**.** *The Maxwell action in curved spacetime,*

$$S_{\mathrm{EM}} = -\frac{1}{4\mu_0}\int \mathrm{d}^4x\, \sqrt{|g|}\; F_{\alpha\beta}F^{\alpha\beta} \,, \tag{44.20}$$

with $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ the field tensor of Chapter 60, yields

$$T_{\mu\nu} = \frac{1}{\mu_0}\left(F_\mu{}^\alpha F_{\nu\alpha} - \tfrac{1}{4}\, g_{\mu\nu}\, F_{\alpha\beta}F^{\alpha\beta}\right) . \tag{44.21}$$

It is traceless, $g^{\mu\nu}T_{\mu\nu} = 0$, and its time–time component in any local orthonormal frame is the familiar energy density $\epsilon_0 E^2/2 + B^2/2\mu_0$.

Rests on Definition 44.6 and Lemma 44.7.

Proof. *Derives Proposition 44.17.*

$F_{\mu\nu}$, defined from the potential with lower indices, does not involve the metric; all metric dependence of Equation (44.20) sits in the two raised indices and the volume factor. Varying,

$$\delta S_{\mathrm{EM}} = -\frac{1}{4\mu_0}\int \mathrm{d}^4x \left[\delta\sqrt{|g|}\; F_{\alpha\beta}F^{\alpha\beta} + \sqrt{|g|}\; 2F_{\mu\alpha}F_{\nu\beta}\, g^{\alpha\beta}\, \delta g^{\mu\nu}\right] ,$$

and inserting Equation (44.7) for the first term,

$$\delta S_{\mathrm{EM}} = -\frac{1}{2}\int \mathrm{d}^4x\, \sqrt{|g|}\; \frac{1}{\mu_0}\left(F_\mu{}^\alpha F_{\nu\alpha} - \tfrac{1}{4} g_{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}\right)\delta g^{\mu\nu} \,,$$

which by Equation (44.5) is Equation (44.21). The trace is $\mu_0^{-1}(F^{\nu\alpha}F_{\nu\alpha} - \tfrac{1}{4}\cdot 4\, F_{\alpha\beta}F^{\alpha\beta}) = 0$, a special feature of $D = 4$. For the energy density, work in a local orthonormal frame: $F_0{}^\alpha F_{0\alpha} = g^{ij}F_{0i}F_{0j} = E^2/c^2$ and $F_{\alpha\beta}F^{\alpha\beta} = 2(B^2 - E^2/c^2)$, so

$$T_{00} = \frac{1}{\mu_0}\left[\frac{E^2}{c^2} + \frac{1}{2}\left(B^2 - \frac{E^2}{c^2}\right)\right] = \frac{\epsilon_0 E^2}{2} + \frac{B^2}{2\mu_0} \,,$$

using $\epsilon_0\mu_0 = 1/c^2$; both intermediate quantities are squares, so the identification is independent of the sign convention relating F_{0i} to E_i. □

Remark 44.18 (Hilbert versus canonical tensor). Translation invariance already gives every flat-spacetime field theory a conserved energy–momentum tensor by Noether's first theorem — the canonical tensor Equation (16.83) of Corollary 16.91 — but that tensor is in general not symmetric, and only becomes so after adding an identically conserved improvement term.

The metric definition Equation (44.5) is symmetric from the start, and for the Maxwell field it reproduces exactly the improved (symmetrized) canonical tensor; the systematic statement of that equivalence, the Belinfante construction, belongs to Chapter 64 (Sections 64.4.2 and 64.4.3). It is the metric tensor, not the canonical one, that gravity couples to: the right-hand side of Equation (44.2) is defined by the same variation that produces the left-hand side.

44.2.2 Conservation and Noether's theorem

Theorem 44.19 (Covariant conservation of stress–energy)**.** *Let $S_{\mathrm{m}}[\psi; g]$ be diffeomorphism invariant, and let the matter fields satisfy their own Euler–Lagrange equations. Then*

$$\nabla_\mu T^{\mu\nu} = 0\,. \tag{44.22}$$

Rests on Theorem 16.93, Definition 44.6, and Proposition 13.127.

Proof. *Derives Theorem 44.19.*

Let ξ^μ be an arbitrary compactly supported vector field and drag all fields along its flow. Diffeomorphism invariance says the action does not change: $0 = \delta_\xi S_{\mathrm{m}} = \int (\delta S_{\mathrm{m}}/\delta\psi)\,\delta_\xi\psi + \int (\delta S_{\mathrm{m}}/\delta g^{\mu\nu})\,\delta_\xi g^{\mu\nu}$. On shell the first integral vanishes. For the second, the Lie derivative of the metric is $\delta_\xi g_{\mu\nu} = \nabla_\mu\xi_\nu + \nabla_\nu\xi_\mu$ (Equation (13.275) with $\partial \to \nabla$ by metricity, as in the proof of Killing's equation Equation (13.290)), so by Equations (44.5) and (44.6)

$$0 = +\frac{1}{2}\int \mathrm{d}^4x\,\sqrt{|g|}\;T^{\mu\nu}\,(\nabla_\mu\xi_\nu + \nabla_\nu\xi_\mu) = \int \mathrm{d}^4x\,\sqrt{|g|}\;T^{\mu\nu}\nabla_\mu\xi_\nu\,,$$

using the symmetry of $T^{\mu\nu}$. Integrate by parts with Lemma 44.10 (no boundary term, compact support): $0 = -\int \sqrt{|g|}\,(\nabla_\mu T^{\mu\nu})\,\xi_\nu$, and since ξ_ν is arbitrary, Equation (44.22) follows. This is Noether's second theorem (Theorem 16.93) in action: the symmetry group depends on four arbitrary functions ξ^μ, and what it yields is not four conserved charges but four differential identities [456]. □

Proposition 44.20 (Perfect-fluid equations of motion)**.** *For the perfect fluid Equation (44.19), Equation (44.22) splits along u^μ and orthogonally to it into*

$$\nabla_\mu\,(\rho u^\mu) = -\frac{p}{c^2}\,\nabla_\mu u^\mu\,, \tag{44.23}$$

$$\left(\rho + \frac{p}{c^2}\right) u^\mu \nabla_\mu u^\lambda = -\left(g^{\lambda\nu} + \frac{u^\lambda u^\nu}{c^2}\right)\partial_\nu p\,, \tag{44.24}$$

the relativistic continuity and Euler equations. In the limit $|\boldsymbol{v}| \ll c$, $p \ll \rho c^2$ they reduce to the continuity and Euler equations of Chapter 31.

Rests on Theorem 44.19 and Definition 44.16.

Proof. *Derives Proposition 44.20.*

Insert Equation (44.19) into $\nabla_\mu T^{\mu\lambda} = 0$:

$$\nabla_\mu\Big[\Big(\rho + \tfrac{p}{c^2}\Big)\,u^\mu\Big]\,u^\lambda + \Big(\rho + \tfrac{p}{c^2}\Big)\,u^\mu\nabla_\mu u^\lambda + g^{\lambda\nu}\partial_\nu p = 0\,.$$

Contract with u_λ: since $u_\lambda u^\lambda = -c^2$ is constant, $u_\lambda\nabla_\mu u^\lambda = 0$, so the middle term dies, the first gives $-c^2\nabla_\mu[(\rho + p/c^2)u^\mu]$, and the last gives $u^\mu\partial_\mu p$. Expanding $\nabla_\mu[(p/c^2)u^\mu] = (p/c^2)\nabla_\mu u^\mu + (u^\mu/c^2)\partial_\mu p$, the two pressure-gradient terms cancel and what remains, divided by $-c^2$, is exactly Equation (44.23). Projecting instead with $h^\lambda{}_\sigma = \delta^\lambda{}_\sigma + u^\lambda u_\sigma/c^2$, which annihilates u^λ, kills the first term and produces Equation (44.24). In the nonrelativistic limit $u^\mu \to (c, \boldsymbol{v})$, $\nabla \to \partial$, $\rho + p/c^2 \to \rho$ and the projector becomes spatial; Equation (44.23) becomes $\partial_t\rho + \nabla\cdot(\rho\boldsymbol{v}) = 0$ and Equation (44.24) becomes $\rho(\partial_t + \boldsymbol{v}\cdot\nabla)\boldsymbol{v} = -\nabla p$, the equations of Chapter 31. □

Corollary 44.21 (Dust flows on geodesics)**.** *For dust ($p = 0$), Equation (44.24) reduces to $u^\mu \nabla_\mu u^\lambda = 0$: the integral curves of u^μ are timelike geodesics. The geodesic postulate Postulate 43.1 is therefore not independent of the field equations for this matter model: conservation of the source, itself enforced by the field equations (Corollary 44.24), already moves the matter on geodesics.*

Rests on Proposition 44.20 and Postulate 43.1.

Proof. *Derives Corollary 44.21.*

Set $p = 0$ in Equation (44.24); the right side vanishes and the left is ρ times the geodesic equation Equation (43.1) in its velocity-field form. □

Remark 44.22 (What covariant conservation does and does not conserve)*.* In flat spacetime $\partial_\mu T^{\mu\nu} = 0$ integrates to four conserved charges. The covariant law Equation (44.22) does not: the connection terms act as sources, describing the exchange of energy and momentum between matter and the gravitational field, and no coordinate-independent gravitational energy *density* exists to book the exchange against — by the equivalence principle, the connection can be made to vanish at any one event, taking any candidate density with it. What survives is either of two honest substitutes. With a symmetry: a Killing vector ξ makes $T^{\mu\nu}\xi_\nu$ a genuinely conserved current, by Equation (44.22) and Killing's equation Equation (13.290). With asymptotic flatness: the total energy, momentum and angular momentum of an isolated system are boundary integrals at spatial infinity — the ADM integrals, whose constrained-dynamics origin is Remark 26.45 [1221]. Between the two sits the Landau–Lifshitz pseudotensor [1552], coordinate-dependent by construction, useful because its integrals reproduce the ADM quantities in asymptotically Cartesian coordinates; the energy carried by gravitational waves, computed via the closely related Isaacson average in Appendix A.1, is the one place this part meets the issue quantitatively.

44.3 The Bianchi identities

The differential identities of the curvature tensor are established, in both index and Cartan form, in Chapter 13: the first and second Bianchi identities are Equation (13.308) and, as two lines of form algebra, Proposition 13.157. The customary attribution to Bianchi [302] is loose — the identities were known to Voss in 1880 and within Ricci's school — and this treatise keeps the customary name while recording the fact. What Part II deliberately does not do is contract them (Remark 16.95 notes the omission); the contraction is where the physics lives, and it is performed here.

Theorem 44.23 (Contracted Bianchi identity)**.** *For the Levi-Civita connection,*

$$\nabla^\mu G_{\mu\nu} = 0\,, \qquad \textit{equivalently} \qquad \nabla^\mu R_{\mu\nu} = \tfrac{1}{2}\,\partial_\nu R\,. \tag{44.25}$$

Rests on Proposition 13.154 and Definitions 13.149 and 44.2.

Proof. *Derives Theorem 44.23.*

Write the second Bianchi identity of Equation (13.308) with its cycle explicit,

$$\nabla_\sigma R^\lambda{}_{\rho\mu\nu} + \nabla_\mu R^\lambda{}_{\rho\nu\sigma} + \nabla_\nu R^\lambda{}_{\rho\sigma\mu} = 0\,.$$

Contract λ with μ (legitimate under ∇ by metricity): the first term gives $\nabla_\sigma R_{\rho\nu}$; the third, by antisymmetry in the last index pair, $-\nabla_\nu R_{\rho\sigma}$; the second stays $\nabla_\lambda R^\lambda{}_{\rho\nu\sigma}$:

$$\nabla_\sigma R_{\rho\nu} - \nabla_\nu R_{\rho\sigma} + \nabla_\lambda R^\lambda{}_{\rho\nu\sigma} = 0\,. \tag{44.26}$$

Now contract with $g^{\rho\sigma}$. The first term gives $\nabla^\rho R_{\rho\nu}$ and the second $-\partial_\nu R$. In the third, lower the index, $g^{\rho\sigma} R^\lambda{}_{\rho\nu\sigma} = g^{\rho\sigma} g^{\lambda\alpha} R_{\alpha\rho\nu\sigma}$, and use the pair-exchange symmetry of Equation (13.308) followed by antisymmetry in each pair:

$$g^{\rho\sigma} R_{\alpha\rho\nu\sigma} = g^{\rho\sigma} R_{\nu\sigma\alpha\rho} = g^{\rho\sigma}\,(-1)\,(-1)\,R_{\sigma\nu\rho\alpha} = R^\rho{}_{\nu\rho\alpha} = R_{\nu\alpha}\,,$$

the last step being the Ricci contraction Equation (13.307) of the first and third slots. So the third term is $\nabla_\lambda(g^{\lambda\alpha}R_{\nu\alpha}) = \nabla^\rho R_{\rho\nu}$. Altogether $2\nabla^\rho R_{\rho\nu} = \partial_\nu R$, and

$$\nabla^\mu G_{\mu\nu} = \nabla^\mu R_{\mu\nu} - \tfrac{1}{2}\, g_{\mu\nu}\nabla^\mu R = \tfrac{1}{2}\,\partial_\nu R - \tfrac{1}{2}\,\partial_\nu R = 0\,.$$

□

Corollary 44.24 (Conservation is an integrability condition)**.** *Any metric satisfying Equation (44.2) has $\nabla_\mu T^{\mu\nu} = 0$ automatically. Local energy–momentum conservation is therefore not an extra postulate of general relativity: it is the integrability condition of the field equations, exactly as charge conservation is the integrability condition of the Maxwell equations (Remark 16.95). Conversely, a candidate left-hand side that is* not *identically divergence-free could not consistently couple to conserved matter — the flaw that doomed Einstein's 1913–1915 intermediate theories (Section 44.1.2).*

Rests on Theorem 44.23 and Postulate 44.3.

Proof. *Derives Corollary 44.24.*

Apply ∇^μ to both sides of Equation (44.2) and use Equation (44.25). □

Remark 44.25 (Ten equations, four identities). Equation (44.25) is four differential identities among the ten field equations, so at most six are independent — the situation anticipated in general form by Corollary 16.94: the identities are the Noether identities of the four-parameter diffeomorphism freedom. The count is exactly right, not a defect: four of the ten $g_{\mu\nu}$ are gauge (coordinate choice), and the four equations "lost" to the identities resurface as constraints on initial data rather than evolution equations. That story is Section 44.9.

44.4 The Newtonian correspondence

44.4.1 Geodesics and the potential

Proposition 44.26 (Newtonian limit of geodesic motion)**.** *Let the metric be static and weakly curved,*

$$g_{00} = -\left(1 + \frac{2\Phi}{c^2}\right), \qquad g_{0i} = 0\,, \qquad g_{ij} = \delta_{ij} + O\!\left(\Phi/c^2\right), \tag{44.27}$$

with $|\Phi| \ll c^2$ time-independent, and let a test body move slowly, $|\mathrm{d}x^i/\mathrm{d}t| \ll c$. Then the geodesic equation Equation (43.1) reduces to Newton's second law in the potential Φ,

$$\frac{\mathrm{d}^2 x^i}{\mathrm{d}t^2} = -\,\partial_i\Phi\,, \tag{44.28}$$

identifying Φ (in $\mathrm{m}^2/\mathrm{s}^2$) with the Newtonian gravitational potential of Chapters 19 and 27.

Rests on Postulate 43.1.

Proof. *Derives Proposition 44.26.*

For slow motion $\mathrm{d}x^i/\mathrm{d}\tau \ll \mathrm{d}x^0/\mathrm{d}\tau$ and $\mathrm{d}x^0/\mathrm{d}\tau = c\,\mathrm{d}t/\mathrm{d}\tau \approx c$, so among the $\Gamma^i{}_{\mu\nu}\,\dot{x}^\mu\dot{x}^\nu$ terms of Equation (43.1) only $\Gamma^i{}_{00}(\dot{x}^0)^2$ survives at leading order. For a static metric,

$$\Gamma^i{}_{00} = \tfrac{1}{2}g^{i\rho}\left(2\partial_0 g_{\rho 0} - \partial_\rho g_{00}\right) = -\tfrac{1}{2}\,g^{ij}\partial_j g_{00} = -\tfrac{1}{2}\,\delta^{ij}\partial_j\left(-\frac{2\Phi}{c^2}\right) + O(\Phi^2/c^4) = \frac{\partial_i\Phi}{c^2}\,,$$

where the spatial perturbation in g^{ij} multiplies a quantity already $O(\Phi)$ and drops. Hence $\mathrm{d}^2x^i/\mathrm{d}\tau^2 = -c^2\,\Gamma^i{}_{00} = -\partial_i\Phi$ and, since $\mathrm{d}\tau \approx \mathrm{d}t$ at this order, Equation (44.28). Note that only g_{00} enters: the identification of the potential is insensitive to the spatial part of Equation (44.27). The same g_{00} is what the gravitational-redshift measurements read directly (Phenomenon 42.5). □

44.4.2 Fixing the coupling constant

Theorem 44.27 (Newtonian correspondence of the field equations)**.** *For the static weak field Equation (44.27) sourced by static dust of density ρ, the time–time component of the field equations Equation (44.2) reduces to the Poisson equation*

$$\nabla^2\Phi = \frac{\kappa c^4}{2}\,\rho\,, \tag{44.29}$$

and correspondence with Newtonian gravity, $\nabla^2\Phi = 4\pi G\rho$, fixes

$$\kappa = \frac{8\pi G}{c^4} \approx 2.0766\times 10^{-43}\,\mathrm{s}^2/(\mathrm{kg\,m})\,, \tag{44.30}$$

with G the constant measured in Chapter 33 [427].

Rests on Propositions 44.4 and 44.26 and Definition 44.15.

Proof. *Derives Theorem 44.27.*

Work in the trace-reversed form Equation (44.3). For static dust at rest, Equation (44.18) gives $T_{00} = \rho c^2$, the other components vanishing, and $T = g^{\mu\nu}T_{\mu\nu} = -\rho c^2$ to leading order. Hence

$$R_{00} = \kappa\left(T_{00} - \tfrac{1}{2}g_{00}T\right) = \kappa\left(\rho c^2 - \tfrac{1}{2}(-1)(-\rho c^2)\right) = \tfrac{1}{2}\,\kappa\rho c^2\,.$$

On the geometric side, the Ricci contraction of Equation (13.306) is $R_{00} = \partial_\lambda\Gamma^\lambda{}_{00} - \partial_0\Gamma^\lambda{}_{\lambda 0} + O(\Gamma^2)$; the Γ^2 terms are $O(\Phi^2/c^4)$ and every time derivative vanishes, so with $\Gamma^i{}_{00} = \partial_i\Phi/c^2$ from Proposition 44.26,

$$R_{00} = \partial_i\Gamma^i{}_{00} = \frac{\nabla^2\Phi}{c^2}\,.$$

Equating the two expressions gives Equation (44.29). The Newtonian side is the field form of the inverse-square law: for a point mass, $\Phi = -Gm/|\boldsymbol{x}|$ (Chapter 27), and superposing over a continuous density, $\Phi(\boldsymbol{x}) = -G\int \rho(\boldsymbol{x}')\,|\boldsymbol{x}-\boldsymbol{x}'|^{-1}\,\mathrm{d}^3x'$, the fundamental-solution property of $-1/4\pi\,|\boldsymbol{x}|$ (Proposition 10.67) gives $\nabla^2\Phi = 4\pi G\rho$. Matching coefficients, $\kappa c^4/2 = 4\pi G$, which is Equation (44.30); the numerical value uses the CODATA G of Appendix B, whose laboratory determination is Chapter 33. □

Remark 44.28 (What the Newtonian limit does and does not fix). The correspondence reads only g_{00} and only the trace-reversed time–time equation; it fixes the single constant κ and nothing else. It does not test the spatial part of the weak-field metric — that is done by light deflection and the Shapiro delay, which measure exactly the spatial curvature the Newtonian limit is blind to, and which decided against Einstein's own 1911 half-value (Section 42.5, Chapter 53); in the parametrized post-Newtonian language the limit fixes no PPN parameter at all, while the classical tests fix γ and β [2240]. Nor does the limit exclude a cosmological term: Λ adds to Equation (44.29) a contribution $-\Lambda c^2$ — a constant outward acceleration $\Lambda c^2 r/3$ — utterly negligible at solar-system scales for the measured value of Remark 44.32, which is why Section 44.6 must be settled by cosmology rather than by laboratory gravity.

44.5 The linearized theory

Far from strong sources spacetime is nearly flat, and the field equations linearize. The full derivation — linearized curvature, gauge fixing, wave solutions, energy flux and the quadrupole formula — is carried out once, in Appendix A.1; this section states the equations, since both Chapter 46 and the cosmological perturbation theory of Chapter 48 start from them.

Proposition 44.29 (Linearized field equations)**.** *Write $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ with $|h_{\mu\nu}| \ll 1$, raise and lower indices with η, and set $h = \eta^{\mu\nu} h_{\mu\nu}$. To first order in h:*

1. *The Einstein tensor is*

$$G_{\mu\nu} = \tfrac{1}{2}\left(\partial_\rho\partial_\mu h^\rho{}_\nu + \partial_\rho\partial_\nu h^\rho{}_\mu - \Box h_{\mu\nu} - \partial_\mu\partial_\nu h - \eta_{\mu\nu}\partial_\rho\partial_\sigma h^{\rho\sigma} + \eta_{\mu\nu}\Box h\right), \tag{44.31}$$

   with $\Box = \eta^{\alpha\beta}\partial_\alpha\partial_\beta$.

2. *Under an infinitesimal coordinate change $x^\mu \to x^\mu + \xi^\mu$ (with $\partial\xi$ of order h),*

$$h_{\mu\nu} \;\longrightarrow\; h_{\mu\nu} - \partial_\mu\xi_\nu - \partial_\nu\xi_\mu\,, \tag{44.32}$$

   and Equation (44.31) is invariant, as is the linearized Riemann tensor.

3. *The Lorenz gauge $\partial^\mu \bar{h}_{\mu\nu} = 0$, with $\bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h$ the trace-reversed perturbation, is always attainable, and in it the field equations Equation (44.2) collapse to the inhomogeneous wave equation*

$$\Box\,\bar{h}_{\mu\nu} = -\,2\kappa\,T_{\mu\nu} = -\,\frac{16\pi G}{c^4}\,T_{\mu\nu}\,, \tag{44.33}$$

   which is Equation (46.2) [436, 1509, 1746].

Rests on Postulate 44.3 and Theorem 44.27.

Full derivation: Appendix A.1 (p. 2759).

Derives Proposition 44.29.

The proof is the opening of Appendix A.1: Equation (44.31) is Equation (A.9), assembled from the linearized Christoffel symbols and Riemann tensor Equations (A.4) and (A.6); the gauge transformation Equation (44.32) is Equation (A.16); the attainability of the Lorenz gauge and the reduction to Equation (44.33) are Equations (A.20) and (A.21).

Three comments place the result. First, Equation (44.33) is the whole causal structure of the theory in miniature: disturbances of the metric propagate at speed c, sourced by stress–energy — the qualitative break with the instantaneous Poisson equation that requirement (iii) of Section 44.1.1 preserved only as a limit. Second, the residual gauge freedom inside the Lorenz class reduces the ten $h_{\mu\nu}$ to two physical polarizations; the transverse-traceless construction is Section 46.2, and the same count falls out of the exact Hamiltonian analysis, with no linearization, in Theorem 26.42. Third, the history: Einstein integrated Equation (44.33) approximately in 1916 [436] and corrected the radiation formula in 1918 [453]; the theory of the waves is Chapter 46 and their observation — indirect through the binary pulsar, direct with the LIGO interferometers — is Chapter 47.

44.6 The cosmological constant

Proposition 44.30 (The cosmological term)**.** *For any constant Λ (SI dimension $1/\mathrm{m}^2$), the equations*

$$G_{\mu\nu} + \Lambda\, g_{\mu\nu} = \kappa\, T_{\mu\nu} \tag{44.34}$$

retain every structural property of Equation (44.2): the left side is symmetric, built from g and its first two derivatives, and identically divergence-free, since $\nabla_\lambda g_{\mu\nu} = 0$ by metricity. They follow from the action Equation (44.4) with R replaced by $R - 2\Lambda$. By Theorem 44.33 below, this one-parameter enlargement is the only *one available in four dimensions.*

Rests on Theorems 44.9 and 44.23 and Definition 13.149.

Proof. *Derives Proposition 44.30.*

Divergence-freedom is $\nabla^\mu(\Lambda g_{\mu\nu}) = \Lambda\nabla^\mu g_{\mu\nu} = 0$ together with Equation (44.25). For the action route, the extra term is $-(\Lambda/\kappa)\int \mathrm{d}^4x\sqrt{|g|}$, whose variation by Equation (44.7) is $+(\Lambda/2\kappa)\int \sqrt{|g|}\, g_{\mu\nu}\delta g^{\mu\nu}$; added to Equation (44.12) it shifts $G_{\mu\nu} \to G_{\mu\nu} + \Lambda g_{\mu\nu}$ in Equation (44.13). □

Einstein introduced the term in 1917, to permit a static, spatially closed universe of uniform dust — the balance requires $\Lambda = 4\pi G\rho/c^2$, with the curvature radius $1/\sqrt{\Lambda}$ [445]. The motive did not survive: the balance is unstable, a slight expansion diluting the dust and unbalancing the repulsion further, as Eddington showed in 1930 [649], and the observed expansion of the universe removed the desire for a static solution altogether; the dynamical (Friedmann) analysis in which both statements live belongs to Section 48.2.2. What survived is the term itself.

Proposition 44.31 (de Sitter geometry)**.** *With $T_{\mu\nu} = 0$ and $\Lambda > 0$, the maximally symmetric solution of Equation (44.34) has*

$$R_{\mu\nu} = \Lambda\, g_{\mu\nu}\,, \qquad R = 4\Lambda\,, \qquad R_{\lambda\rho\mu\nu} = \frac{\Lambda}{3}\left(g_{\lambda\mu}g_{\rho\nu} - g_{\lambda\nu}g_{\rho\mu}\right)\,, \tag{44.35}$$

the constant-curvature spacetime of curvature radius $\ell = \sqrt{3/\Lambda}$: de Sitter space, realized as a quadric in five flat dimensions in Section 13.13 [450]. Λ alone, with no matter whatever, curves spacetime.

Rests on Theorem 13.160 and Proposition 44.30.

Proof. *Derives Proposition 44.31.*

By Theorem 13.160 a maximally symmetric metric has $R_{\lambda\rho\mu\nu} = \frac{R}{12}(g_{\lambda\mu}g_{\rho\nu} - g_{\lambda\nu}g_{\rho\mu})$ with R constant, whence $R_{\mu\nu} = \frac{R}{4}g_{\mu\nu}$. Insert into Equation (44.34) with $T = 0$: $\frac{R}{4}g_{\mu\nu} - \frac{R}{2}g_{\mu\nu} + \Lambda g_{\mu\nu} = 0$, so $R = 4\Lambda$, and back-substitution gives Equation (44.35). The curvature constant of Equation (13.318) is $R/12 = \Lambda/3$, giving $\ell = (3/\Lambda)^{1/2}$; existence is the explicit quadric construction of Section 13.13. The case $\Lambda < 0$ gives anti-de Sitter space the same way. □

Remark 44.32 (The measured cosmological constant). Λ is today a measured cosmological parameter, not an option. The accelerating expansion discovered in the type-Ia supernova surveys [1962, 1981] (Phenomenon 48.17) and the CMB-anchored fits of [2348] concordantly give, through $\Lambda = 3\Omega_\Lambda H_0^2/c^2$ with $\Omega_\Lambda \approx 0.685$ and $H_0 \approx 67.4\,\mathrm{km/(s\,Mpc)}$,

$$\Lambda \approx 1.1 \times 10^{-52}/\mathrm{m}^2\,, \tag{44.36}$$

equivalently a vacuum energy density $\rho_\Lambda c^2 = \Lambda c^4/8\pi G \approx 5.3 \times 10^{-10}\,\mathrm{J/m^3}$ — about half a nanojoule per cubic metre. Moved to the right-hand side of Equation (44.34), the term is a perfect fluid with $p = -\rho c^2$, which is what makes it gravitationally repulsive (Remark 44.40). Why it is so small is not known: naive quantum-field-theoretic estimates of vacuum energy exceed the measured value by up to some 120 orders of magnitude, depending on the cutoff assumed — the cosmological-constant problem, stated sharply in [1828]. The observational treatment of Λ, and of the broader dark-energy question, is Chapter 48 (Section 48.6.2).

44.7 Lovelock's uniqueness theorem

The requirements of Section 44.1.1 were presented as motivation; remarkably, they are a characterization. The following theorem is imported: its proof is a classification of tensor concomitants that belongs to the mathematical literature, and this treatise uses only its statement.

Theorem 44.33 (Lovelock's uniqueness theorem; imported)**.** *In four dimensions, let $A^{\mu\nu}$ be a tensor concomitant of the metric and its first and second partial derivatives, $A^{\mu\nu} = A^{\mu\nu}(g_{\alpha\beta}, \partial_\gamma g_{\alpha\beta}, \partial_\gamma\partial_\delta g_{\alpha\beta})$, satisfying $\nabla_\mu A^{\mu\nu} = 0$ identically. Then there exist constants α and λ with*

$$A^{\mu\nu} = \alpha\, G^{\mu\nu} + \lambda\, g^{\mu\nu}\,. \tag{44.37}$$

Neither symmetry of $A^{\mu\nu}$ nor linearity in the second derivatives needs to be assumed: in $D = 4$ both follow [1451, 1479].

Remark 44.34 (Status and scope of the import). The proof is in Lovelock's two papers: the 1971 paper [1451] classifies the divergence-free symmetric concomitants in arbitrary dimension, and the 1972 paper [1479] shows that in four dimensions the symmetry and linearity hypotheses are redundant. The statement is naturally dimension-general in the sense of the Part II conventions: in $D > 4$ the classification admits further terms, one for each even dimension below D — the Lovelock tensors, obtained by varying dimensionally continued Euler densities, of which the first beyond Einstein is the quadratic Gauss–Bonnet combination $R^2 - 4R_{\mu\nu}R^{\mu\nu} + R_{\lambda\rho\mu\nu}R^{\lambda\rho\mu\nu}$. In $D = 4$ that combination is the Euler density itself: a topological term whose variation vanishes identically, contributing nothing to the field equations — which is precisely why the four-dimensional answer Equation (44.37) is so short. Physical spacetime instantiates $D = 4$, and only Equation (44.37) is used in this treatise.

Corollary 44.35 (The field equations are essentially unique)**.** *Any four-dimensional theory meeting the requirements of Section 44.1.1 — metric field variable, tensorial second-order field equations, identically conserved left-hand side coupled to the stress–energy tensor — has field equations of the form $\alpha G_{\mu\nu} + \lambda g_{\mu\nu} = T_{\mu\nu}$; absorbing the normalization into $\kappa = 1/\alpha$ and writing $\Lambda = \lambda/\alpha$, these are Equation (44.34). The Newtonian limit then fixes κ (Theorem 44.27), and observation fixes Λ (Remark 44.32). Nothing else was available.*
Rests on Theorems 44.27 and 44.33.

Proof. *Derives Corollary 44.35.*
Immediate from Theorem 44.33 applied to the left-hand side, plus the two calibrations cited. □

44.8 Energy conditions

The field equations place no restriction whatever on $T_{\mu\nu}$: any metric solves them for *some* source. Physical conclusions therefore need model-independent positivity assumptions about matter, and four standard ones do nearly all the work in this part — above all as hypotheses of the singularity theorems of Section 45.4.2 [1319, 1504].

Definition 44.36 (Energy conditions)**.** Let w^μ range over future-directed timelike vectors with $w^\mu w_\mu = -c^2$, and k^μ over future-directed null vectors. The stress–energy tensor satisfies the

- *null energy condition* (NEC) if $T_{\mu\nu}k^\mu k^\nu \geq 0$ for all k;
- *weak energy condition* (WEC) if $T_{\mu\nu}w^\mu w^\nu \geq 0$ for all w — every observer measures nonnegative energy density;
- *strong energy condition* (SEC) if $(T_{\mu\nu} - \tfrac{1}{2}T g_{\mu\nu})\, w^\mu w^\nu \geq 0$ for all w;
- *dominant energy condition* (DEC) if WEC holds and, for every w, the flux vector $-T^\mu{}_\nu w^\nu$ is causal (non-spacelike) and future-directed — energy does not flow faster than light.

WEC implies NEC by continuity as w tilts to the light cone; SEC also implies NEC the same way, but SEC does *not* imply WEC.

Rests on Definition 44.6.

Proposition 44.37 (Energy conditions for the perfect fluid)**.** *For the perfect fluid Equation (44.19):*

Condition	*Equivalent inequalities*
NEC	$\rho c^2 + p \geq 0$
WEC	$\rho \geq 0$ *and* $\rho c^2 + p \geq 0$
SEC	$\rho c^2 + 3p \geq 0$ *and* $\rho c^2 + p \geq 0$
DEC	$\rho \geq 0$ *and* $\lvert p\rvert \leq \rho c^2$

Rests on Definitions 44.16 and 44.36.

Proof. *Derives Proposition 44.37.*

Write the observer as $w^\mu = \gamma(u^\mu + v^\mu)$ with $u_\mu v^\mu = 0$, v^μ spacelike of norm $v^2 < c^2$; normalization forces $\gamma^2(c^2 - v^2) = c^2$, the Lorentz factor of the relative motion, and $u_\mu w^\mu = -\gamma c^2$. Then from Equation (44.19),

$$T_{\mu\nu} w^\mu w^\nu = \left(\rho + \tfrac{p}{c^2}\right)\gamma^2 c^4 - p\,c^2 = c^2\left[\gamma^2 c^2 \rho + \left(\gamma^2 - 1\right) p\right] .$$

Nonnegativity for all $\gamma \geq 1$ is equivalent to the two endpoint conditions $\gamma = 1$ ($\rho \geq 0$) and $\gamma \to \infty$ ($\rho c^2 + p \geq 0$), since the expression is affine in γ^2: that is WEC. For a null k, $T_{\mu\nu} k^\mu k^\nu = (\rho + p/c^2)(u_\mu k^\mu)^2$ with $u_\mu k^\mu \neq 0$, giving NEC. For SEC, the trace is $T = -\rho c^2 + 3p$, and

$$\left(T_{\mu\nu} - \tfrac{1}{2} T g_{\mu\nu}\right) w^\mu w^\nu = c^2\left[\gamma^2\left(\rho c^2 + p\right) + \tfrac{1}{2}\left(p - \rho c^2\right)\right] ,$$

again affine in γ^2: nonnegative for all γ iff the slope $\rho c^2 + p \geq 0$ and the value at $\gamma = 1$, $\frac{1}{2}(\rho c^2 + 3p)$, is nonnegative. For DEC, compute the flux:

$$f^\mu = -T^\mu{}_\nu w^\nu = \gamma\left(\rho c^2\, u^\mu - p\, v^\mu\right) , \qquad f^\mu f_\mu = \gamma^2\left(-\rho^2 c^6 + p^2 v^2\right) .$$

Causality of f for every $v^2 < c^2$ is equivalent to $p^2 c^2 \leq \rho^2 c^6$, i.e. $\lvert p\rvert \leq \rho c^2$, and future-direction follows from $f^\mu u_\mu = -\gamma \rho c^4 \leq 0$ when $\rho \geq 0$. □

Proposition 44.38 (The electromagnetic field satisfies NEC, WEC, SEC and DEC)**.** *The stress–energy tensor Equation (44.21) satisfies all four conditions of Definition 44.36.*

Rests on Proposition 44.17 and Definition 44.36.

Proof. *Derives Proposition 44.38.*

WEC: for any observer, evaluate in that observer's local orthonormal frame; by Proposition 44.17, $T_{\mu\nu} w^\mu w^\nu / c^2 = T_{\hat{0}\hat{0}} = \epsilon_0 E^2/2 + B^2/2\mu_0 \geq 0$, with E, B the fields that observer measures. NEC follows by continuity. SEC: the tensor is traceless, so the SEC quadratic form equals the WEC form. DEC: in the observer's frame the flux components are $f^{\hat{0}} = c\,T^{\hat{0}\hat{0}} = cu$ and $f^{\hat{\imath}} = c\,T^{\hat{\imath}\hat{0}} = S^{\hat{\imath}}$, where u is the energy density above and $\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{B}/\mu_0$ is the Poynting flux (Chapter 60); causality of f is $c^2 u^2 \geq \lvert\boldsymbol{S}\rvert^2$. With $a = \epsilon_0 E^2$, $b = B^2/\mu_0$,

$$u^2 - \frac{\lvert\boldsymbol{S}\rvert^2}{c^2} = \frac{(a+b)^2}{4} - \frac{\epsilon_0}{\mu_0}\lvert\boldsymbol{E} \times \boldsymbol{B}\rvert^2 \geq \frac{(a+b)^2}{4} - ab = \frac{(a-b)^2}{4} \;\geq\; 0 ,$$

using $\lvert\boldsymbol{E} \times \boldsymbol{B}\rvert \leq \lvert\boldsymbol{E}\rvert\,\lvert\boldsymbol{B}\rvert$ and $\epsilon_0/\mu_0 \cdot E^2 B^2 = ab$. Future-direction is $f^{\hat{0}} = cu \geq 0$. □

Proposition 44.39 (SEC is the attraction condition)**.** *Assume the field equations Equation* (44.34)*. Then for every timelike w,*

$$R_{\mu\nu}w^\mu w^\nu = \kappa\left(T_{\mu\nu} - \tfrac{1}{2}Tg_{\mu\nu}\right)w^\mu w^\nu - \Lambda c^2\,, \tag{44.38}$$

so with $\Lambda = 0$ the SEC is exactly the timelike convergence condition $R_{\mu\nu}w^\mu w^\nu \geq 0$: through the geodesic deviation equation Equation (43.3)*, matter satisfying the SEC focuses congruences of free-falling worldlines — gravity attracts. This is the geometric hypothesis under which the singularity theorems of Penrose and of Hawking–Penrose operate [1319, 1425, 1504] (Section 45.4.2); the null convergence condition used in Penrose's theorem follows the same way from the NEC, with the Λ term dropping out on null vectors.*

Rests on Proposition 44.4 and Definition 44.36.

Proof. *Derives Proposition 44.39.*

Trace-reverse Equation (44.34) as in Proposition 44.4: tracing gives $-R + 4\Lambda = \kappa T$, so $R_{\mu\nu} = \kappa(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T) + \Lambda g_{\mu\nu}$, and contracting twice with w (using $g_{\mu\nu}w^\mu w^\nu = -c^2$) gives Equation (44.38). For a null vector the $g_{\mu\nu}$ terms contract to zero, leaving $R_{\mu\nu}k^\mu k^\nu = \kappa T_{\mu\nu}k^\mu k^\nu$. □

Remark 44.40 (The measured universe violates the SEC). A positive cosmological constant, read as vacuum matter with $p = -\rho_\Lambda c^2$ (Remark 44.32), has $\rho c^2 + 3p = -2\rho_\Lambda c^2 < 0$: it violates the SEC while satisfying NEC, WEC and DEC (marginally, at the $|p| = \rho c^2$ boundary). The observed accelerating expansion [1962, 1981] is precisely the statement that the SEC fails for the universe's present energy budget (Phenomenon 48.17, Section 48.6): cosmic repulsion is measured, not hypothesized. The SEC therefore has the status of a classical-matter hypothesis, valid for the fluids and fields of Section 44.2.1, appropriate to the collapse and black-hole settings of Chapter 45 — not a law of nature. The NEC, by contrast, holds for every matter model in this book, and the modern singularity and area theorems lean on it wherever possible.

44.9 The initial-value problem

General covariance hides the causal structure of the field equations: ten equations for ten $g_{\mu\nu}$, but four identities (Remark 44.25) and four coordinate freedoms. The honest count — what data on a spacelike surface determine the future, and through what evolution equations — is the subject of this section. It is also what makes the theory *predictive*, numerical relativity possible, and the waveform science of Chapter 47 more than curve fitting.

44.9.1 Constraints and evolution

Proposition 44.41 (Four of the ten equations are constraints)**.** *In any coordinates in which x^0 is a time function, the four components $G^{0\nu}$ of the Einstein tensor contain no second x^0-derivative of the metric. The four field equations $G^{0\nu} = \kappa T^{0\nu}$ are therefore not evolution equations but* constraints *on the data induced on each surface $x^0 = \text{const}$.*

Rests on Theorem 44.23.

Proof. *Derives Proposition 44.41.*

The contracted Bianchi identity Equation (44.25), written out,

$$\partial_0 G^{0\nu} = -\,\partial_i G^{i\nu} - \Gamma^\mu{}_{\mu\lambda}G^{\lambda\nu} - \Gamma^\nu{}_{\mu\lambda}G^{\mu\lambda}\,,$$

holds identically in the metric. Every term on the right contains at most second x^0-derivatives of $g_{\mu\nu}$ (G is second order and ∂_i, Γ add no time derivative). If $G^{0\nu}$ contained a second time derivative, say a term $A\,\partial_0^2 g_{\alpha\beta}$ with A built from lower-order quantities, the left side would contain the third derivative $A\,\partial_0^3 g_{\alpha\beta}$, which nothing on the right can supply; since the identity holds for *every* metric, the coefficient A vanishes identically. □

Theorem 44.42 (The constraint equations)**.** *Let Σ be a spacelike hypersurface with future-directed unit normal n^μ ($n^\mu n_\mu = -1$), induced Riemannian metric h_{ij}, and extrinsic curvature $K_{\mu\nu} = h^\alpha{}_\mu h^\beta{}_\nu \nabla_\alpha n_\beta$, where $h^\alpha{}_\mu = \delta^\alpha{}_\mu + n^\alpha n_\mu$ projects onto Σ. Write ${}^{(3)}R$ for the Ricci scalar of h_{ij}, D_i for its Levi-Civita connection, and $K = h^{ij}K_{ij}$. Then the normal–normal and normal–tangential projections of the field equations Equation (44.2) read*

$$ {}^{(3)}R + K^2 - K_{ij}K^{ij} = 2\kappa\, T_{\mu\nu} n^\mu n^\nu \,, \tag{44.39} $$

$$ D_j \left(K^j{}_i - \delta^j{}_i\, K \right) = \kappa\, T_{\mu\nu}\, n^\mu h^\nu{}_i \,, \tag{44.40} $$

the Hamiltonian *and* momentum *constraints: equations on the data (h_{ij}, K_{ij}) alone, with $T_{\mu\nu}n^\mu n^\nu$ the energy density and $-T_{\mu\nu}n^\mu h^\nu{}_i$ the momentum density measured by the observers at rest in Σ.*

Rests on Lemmas A.628 and A.630 and Postulate 44.3.

Proof. *Derives Theorem 44.42.*

The geometric identities are the Gauss and Codazzi equations, proved for arbitrary signature in Appendix A.52 with the normalization $\epsilon = n^\mu n_\mu$ carried as a symbol; in this chapter's signature a spacelike surface has $\epsilon = -1$ and the induced metric $\gamma_{\mu\nu} = g_{\mu\nu} - \epsilon n_\mu n_\nu$ used there is our positive-definite $h_{\mu\nu}$.

Hamiltonian constraint. Contract the Gauss equation Equation (A.985) over its first and third indices, then trace with $h^{\nu\mu}$. The intrinsic side gives ${}^{(3)}R$. The quadratic terms give $-\epsilon(K_{\mu\nu}K^{\mu\nu} - K^2)$. The ambient side is the double h-trace of the projected Riemann tensor; expanding $h^{\mu\nu} = g^{\mu\nu} + n^\mu n^\nu$ in both traces and using the Riemann symmetries, the $nnnn$ term vanishes by antisymmetry, and

$$ h^{\alpha\lambda} h^{\tau\sigma} R_{\alpha\tau\lambda\sigma} = R + 2R_{\mu\nu}n^\mu n^\nu \,. $$

Hence, at $\epsilon = -1$, ${}^{(3)}R = R + 2R_{\mu\nu}n^\mu n^\nu + K_{\mu\nu}K^{\mu\nu} - K^2$. But $2G_{\mu\nu}n^\mu n^\nu = 2R_{\mu\nu}n^\mu n^\nu + R$ since $g_{\mu\nu}n^\mu n^\nu = -1$, so

$$ {}^{(3)}R + K^2 - K_{ij}K^{ij} = 2\, G_{\mu\nu} n^\mu n^\nu \,, $$

and the field equations replace G_{nn} by κT_{nn}, giving Equation (44.39).

Momentum constraint. The contracted Codazzi equation Equation (A.990) is $D_\nu(K^\nu{}_\mu - K\, h^\nu{}_\mu) = h^\alpha{}_\mu R_{\alpha\kappa} n^\kappa$, with no factor of ϵ. Since $h^\alpha{}_\mu g_{\alpha\kappa} n^\kappa = h^\alpha{}_\mu n_\alpha = 0$, the Ricci tensor may be replaced by the Einstein tensor, and the field equations by κT; restricting the free index to Σ gives Equation (44.40).

That these are the four equations of Proposition 44.41 follows because $G_{\mu\nu}n^\nu$ and $G^{0\nu}$ span the same four combinations: $n_\mu \propto \partial_\mu x^0$ on the surface. Neither Equation (44.39) nor Equation (44.40) contains a time derivative of K_{ij}, i.e. a second time derivative of h_{ij}: they restrict the initial data. □

Remark 44.43 (The constrained-Hamiltonian reading, and the count of degrees of freedom)*.* The same split is the backbone of the canonical formulation worked out, as constrained dynamics, in Section 26.6.5: there the lapse and shift appear as multipliers, the vacuum forms of Equations (44.39) and (44.40) are the secondary constraints $\mathcal{H}_\perp \approx 0$, $\mathcal{H}_i \approx 0$ of Theorem 26.42 (stated in the mostly-minus convention; the dictionary is the ϵ bookkeeping of Appendix A.52), all eight constraints are first class with the hypersurface-deformation algebra Proposition 26.43, and the count $2 \times 10 - 2 \times 8 = 4$ leaves *two* configuration degrees of freedom per point of space — the two gravitational-wave polarizations of Section 46.2, with no linearization anywhere in the argument. Given constraint-satisfying data, the remaining six field equations evolve (h_{ij}, K_{ij}) once a lapse and shift — pure gauge — are chosen. Solving the constraints themselves is a nontrivial elliptic problem; the standard method, York's conformal technique, is [1490]. And the constraints, once solved, stay solved: Theorem 44.45 below includes the propagation statement, whose engine is again the contracted Bianchi identity.

44.9.2 Well-posedness

That the remaining equations really do evolve the data — existence, uniqueness, continuous dependence, the Hadamard well-posedness of Definition 10.24 — is not visible until the coordinate freedom is fixed, and the fixing that works is the harmonic (de Donder) gauge.

Proposition 44.44 (Harmonic-gauge reduction)**.** *Define* $\Gamma^\alpha = g^{\mu\nu}\Gamma^\alpha{}_{\mu\nu}$ *and* $\Gamma_\nu = g_{\nu\alpha}\Gamma^\alpha$*. Then, identically,*

$$R_{\mu\nu} = -\tfrac{1}{2}\, g^{\lambda\rho}\, \partial_\lambda \partial_\rho\, g_{\mu\nu} + \tfrac{1}{2}\left(\partial_\mu \Gamma_\nu + \partial_\nu \Gamma_\mu\right) + Q_{\mu\nu}(g, \partial g)\ , \tag{44.41}$$

where $Q_{\mu\nu}$ *is a polynomial in* $g^{\alpha\beta}$ *and* $\partial_\gamma g_{\alpha\beta}$*, quadratic in the first derivatives, containing no second derivative. In* harmonic coordinates *— coordinates each satisfying the curved wave equation, equivalently* $\Gamma^\alpha = 0$ *— the vacuum equations* $R_{\mu\nu} = 0$ *become the* reduced equations

$$g^{\lambda\rho}\, \partial_\lambda \partial_\rho\, g_{\mu\nu} = 2\, Q_{\mu\nu}(g, \partial g)\ , \tag{44.42}$$

ten quasilinear wave equations with a common principal part, the curved d'Alembertian, acting on each component separately.

Rests on Theorems 13.150 and 13.152.

Proof. *Derives Proposition 44.44.*

Only the terms of $R_{\mu\nu}$ carrying second derivatives need tracking; everything else is by definition collected in $Q_{\mu\nu}$. From the Ricci contraction of Equation (13.306), $R_{\mu\nu} = \partial_\lambda \Gamma^\lambda{}_{\nu\mu} - \partial_\nu \Gamma^\lambda{}_{\lambda\mu} + \Gamma\Gamma$, and the Christoffel formula, the second-derivative content is

$$R_{\mu\nu} \ \simeq\ \tfrac{1}{2} g^{\lambda\sigma}\left(\partial_\lambda \partial_\mu g_{\sigma\nu} + \partial_\lambda \partial_\nu g_{\sigma\mu} - \partial_\lambda \partial_\sigma g_{\mu\nu} - \partial_\mu \partial_\nu g_{\lambda\sigma}\right)\ ,$$

$\simeq$ denoting equality modulo first-derivative terms. On the other hand $\Gamma_\nu \simeq g^{\lambda\sigma}(\partial_\lambda g_{\sigma\nu} - \tfrac{1}{2}\partial_\nu g_{\lambda\sigma})$ up to products $(\partial g)(g)$, so

$$\tfrac{1}{2}\left(\partial_\mu \Gamma_\nu + \partial_\nu \Gamma_\mu\right) \ \simeq\ \tfrac{1}{2} g^{\lambda\sigma}\left(\partial_\lambda \partial_\mu g_{\sigma\nu} + \partial_\lambda \partial_\nu g_{\sigma\mu} - \partial_\mu \partial_\nu g_{\lambda\sigma}\right)\ .$$

Subtracting reproduces exactly the missing $-\tfrac{1}{2} g^{\lambda\sigma}\partial_\lambda \partial_\sigma g_{\mu\nu}$, which is Equation (44.41). A coordinate function x^α satisfies $\Box_g x^\alpha = g^{\mu\nu}\nabla_\mu \nabla_\nu x^\alpha = -g^{\mu\nu}\Gamma^\alpha{}_{\mu\nu} = -\Gamma^\alpha$, so harmonicity of all four coordinates is $\Gamma^\alpha = 0$, and Equation (44.42) follows. □

The reduced equations are hyperbolic wherever the metric is Lorentzian, and for *analytic* data the Cauchy–Kovalevskaya theorem (Theorem 10.22) already yields local solutions. Analyticity, however, is poison for relativity — analytic continuation propagates data instantaneously, defeating the causal structure the equations exist to express — and the honest function spaces are Sobolev, where the machinery is energy estimates for quasilinear hyperbolic systems. That machinery lies beyond this treatise, and the two theorems that close the problem are imported.

Theorem 44.45 (Local existence and uniqueness; Choquet-Bruhat; imported)**.** *Let* (Σ, h_{ij}, K_{ij}) *be smooth initial data satisfying the vacuum constraints (Equations* (44.39) *and* (44.40) *with* $T_{\mu\nu} = 0$*). Then there exists a vacuum spacetime* $(M, g_{\mu\nu})$ *— a* development *of the data — containing an embedded hypersurface with induced metric* h_{ij} *and extrinsic curvature* K_{ij}*, obtained by solving the reduced equations Equation* (44.42)*; the harmonic gauge condition and the constraints, imposed at the initial surface, propagate; and any two developments agree, near* Σ*, up to a diffeomorphism fixing* Σ*. The Cauchy problem of general relativity is well posed in the sense of Definition 10.24 [999].*

Theorem 44.46 (Maximal globally hyperbolic development; Choquet-Bruhat and Geroch; imported)**.** *Among all globally hyperbolic vacuum developments of given constraint-satisfying data there is a maximal one, unique up to isometry, into which every other embeds isometrically [1399].*

Remark 44.47 (Status of the imports, and what hangs on them). Theorem 44.45 is Choquet-Bruhat's 1952 theorem, published under the name Fourès-Bruhat [999]; modern proofs, with the Sobolev thresholds and the constraint-propagation computation written out, are in [1746] (ch. 10) and [1504] (ch. 7). The propagation argument is worth stating even unproved, because it closes this chapter's circle: if the reduced equations hold, the contracted Bianchi identity Equation (44.25) forces the four quantities Γ^α to satisfy a linear homogeneous hyperbolic system; data satisfying the constraints start the Γ^α and their first time derivatives at zero; and the unique solution of a homogeneous system with zero data is zero — so the gauge condition maintains itself and the reduced solution solves the full Einstein equations. Theorem 44.46 then removes the arbitrariness of gauge and extent in one stroke: general relativity assigns to each admissible initial datum a unique maximal spacetime. Determinism holds exactly as far as global hyperbolicity does; whether it can fail beyond — Cauchy horizons, naked singularities — is the cosmic-censorship question of Section 45.4.3 [1405]. On these two theorems rests numerical relativity: well-posedness is what guarantees that discretized evolutions converge to the continuum solution, and with it the template waveforms against which the detections of Chapter 47 are read.

44.10 Exact solutions: an overview

The field equations are ten coupled quasilinear PDEs, and their generic solution is numerical. The exactly solved cases are few, each purchased with heavy symmetry, and almost every observational confrontation in this part runs through one of them. They are collected here once, with their working homes; each named solution is derived in the chapter that consumes it.

Table 44.1: The exact solutions this treatise uses. Each is derived and exploited in the chapter of the last column; the citations for the original papers are given in the text below.

Solution	Symmetry	Source	Treated in
Minkowski	maximal, flat	vacuum	Chapter 39
de Sitter / anti-de Sitter	maximal	Λ only	Section 44.6
Schwarzschild exterior	static, spherical	vacuum	Chapter 45
Interior fluid spheres	static, spherical	perfect fluid	Chapter 49
Kerr	stationary, axial	vacuum	Chapter 45
FLRW cosmologies	homogeneous, isotropic	perfect fluid	Chapter 48
Oppenheimer–Snyder	spherical, dynamical	dust	Chapter 45
Linearized waves	none (perturbative)	weak sources	Chapter 46

The Schwarzschild exterior solution — found within weeks of the field equations, from the trenches of the eastern front — describes the vacuum outside any spherical mass [440], with Droste's independent derivation arriving months later in the now-standard coordinates [444]; by Birkhoff's theorem (Section 45.1.2) it is the *unique* spherically symmetric vacuum solution, static or not. It underwrites the classical tests (Chapter 53) and, continued inside its horizon, the black hole (Chapter 45). Its interior companions — Schwarzschild's constant-density sphere and the Tolman–Oppenheimer–Volkoff hydrostatics built on Equation (44.24) — carry relativistic stellar structure [845, 848] and are the business of Chapter 49; the Oppenheimer–Snyder collapse of a dust ball [844] is the exactly solvable model of black-hole formation (Section 45.4.1). The Kerr solution, the stationary rotating vacuum found only in 1963 [1259], is by the no-hair theorems (Section 45.5.3) the endpoint of every gravitational collapse and the metric read directly by the observations of Chapter 50. The Friedmann–Lemaître–Robertson–Walker family — Friedmann's dynamical solutions [490], Lemaître's

independent expansion physics [583], and the Robertson–Walker proof that the metric form is forced by homogeneity and isotropy alone [758, 804] — instantiates the maximally symmetric *spaces* of Section 13.13 as the spatial sections of cosmology, and carries all of Chapter 48. Plane gravitational waves close the list (Chapter 46). The systematic monographs are [1509, 1746].

What the short list teaches is the shape of the subject: a handful of exact geometries, each pinned to data — planetary orbits and light bending for Schwarzschild, X-ray binaries, stellar dynamics and horizon-scale imaging for Kerr, the expansion history for FLRW, interferometer strain for the waves — with the equations themselves, checked in Section 44.4.2 against the Newtonian world and secured as a well-posed dynamical system in Section 44.9, mediating between them. The chapters that follow take up each geometry in turn.

Chapter 45

Schwarzschild Geometry and Black Holes

The first exact solution of the field equations of Chapter 44 was found within weeks of their publication [440], and it carries a remarkable fraction of the theory's observable content: the orbits behind the classical tests of Chapter 42, the innermost stable orbit and photon sphere that set the scales of accretion, and — once its coordinate singularity was understood — the event horizon and the black hole. This chapter derives the solution, studies its geodesics, extends it maximally, and follows gravitational collapse to the singularity theorems that make the outcome generic [1319].

The second half opens the black-hole family: the Kerr rotating solution [1259], its charged relatives, and the uniqueness (no-hair) theorems that reduce every stationary black hole to two observable parameters [1351, 1443]. The shadow of the horizon is derived as a prediction here; the observations — stellar orbits about the galactic centre, the Event Horizon Telescope images, ringdown spectroscopy — are deferred with their full apparatus to the black-hole experiment chapter, Chapter 50, and to Chapter 47. Standard treatments are [1509, 1746].

45.1 The Schwarzschild solution

45.1.1 Derivation and the 1916 paper

A gravitational field is *static* when spacetime admits a timelike Killing vector (Definition 13.139) whose orthogonal distribution is integrable — there is a time coordinate t that the geometry neither depends on nor mixes with space — and *spherically symmetric* when the rotation group acts by isometries with two-dimensional spacelike orbits. On each orbit sphere the metric must be a multiple of the round metric, and the multiple defines the *areal radius* r: the sphere labelled r has area $4\pi r^2$ by construction, whatever "distance to the centre" turns out to mean. Wherever $\mathrm{d}r \neq 0$ these choices reduce the metric to two unknown functions of r [1746],

$$\mathrm{d}s^2 = -\mathrm{e}^{2\alpha(r)}c^2\mathrm{d}t^2 + \mathrm{e}^{2\beta(r)}\mathrm{d}r^2 + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2\right). \tag{45.1}$$

Staticity is assumed here for transparency and *proved* redundant in Section 45.1.2.

Write $x^0 := ct$, so that primes denote $\mathrm{d}/\mathrm{d}r$ throughout. A mechanical application of Equation (13.299) (torsion zero, per Chapter 43) gives the nonvanishing Christoffel symbols

of Equation (45.1),

$$\begin{aligned} \Gamma^0{}_{0r} &= \alpha' , & \Gamma^r{}_{00} &= \alpha' \mathrm{e}^{2(\alpha-\beta)} , & \Gamma^r{}_{rr} &= \beta' , \\ \Gamma^r{}_{\theta\theta} &= -r\mathrm{e}^{-2\beta} , & \Gamma^r{}_{\varphi\varphi} &= -r\mathrm{e}^{-2\beta}\sin^2\theta , & \Gamma^\theta{}_{r\theta} &= \frac{1}{r} , \\ \Gamma^\theta{}_{\varphi\varphi} &= -\sin\theta\cos\theta , & \Gamma^\varphi{}_{r\varphi} &= \frac{1}{r} , & \Gamma^\varphi{}_{\theta\varphi} &= \cot\theta , \end{aligned} \tag{45.2}$$

and contracting the Riemann tensor Equation (13.306) as in Definition 13.153 yields the Ricci components

$$R_{00} = \mathrm{e}^{2(\alpha-\beta)}\left[\alpha'' + \alpha'^2 - \alpha'\beta' + \frac{2\alpha'}{r}\right] , \tag{45.3}$$

$$R_{rr} = -\alpha'' - \alpha'^2 + \alpha'\beta' + \frac{2\beta'}{r} , \tag{45.4}$$

$$R_{\theta\theta} = 1 - \mathrm{e}^{-2\beta}\left[1 + r\left(\alpha' - \beta'\right)\right] , \tag{45.5}$$

with $R_{\varphi\varphi} = \sin^2\theta\, R_{\theta\theta}$ and all off-diagonal components zero.

Theorem 45.1 (Schwarzschild solution)**.** *The unique static, spherically symmetric, asymptotically flat solution of the vacuum field equations $R_{\mu\nu} = 0$ (Chapter 44) exterior to a body of mass M is*

$$\mathrm{d}s^2 = -\left(1 - \frac{r_\mathrm{s}}{r}\right)c^2\mathrm{d}t^2 + \left(1 - \frac{r_\mathrm{s}}{r}\right)^{-1}\mathrm{d}r^2 + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2\right) , \tag{45.6}$$

with the Schwarzschild radius

$$r_\mathrm{s} := \frac{2GM}{c^2} . \tag{45.7}$$

Rests on Equation (13.306)*, Theorem 13.150, and Postulate 43.1.*

Derivation. *Derives Theorem 45.1.*

Vacuum demands that Equations (45.3) to (45.5) vanish. The combination

$$\mathrm{e}^{2(\beta-\alpha)}R_{00} + R_{rr} = \frac{2}{r}\left(\alpha' + \beta'\right) = 0$$

kills every second-derivative term at a stroke: $\alpha + \beta$ is constant, and asymptotic flatness ($\alpha, \beta \to 0$ as $r \to \infty$) sets the constant to zero, so $\beta = -\alpha$. Inserting this into $R_{\theta\theta} = 0$ gives

$$1 = \mathrm{e}^{2\alpha}\left(1 + 2r\alpha'\right) = \left(r\,\mathrm{e}^{2\alpha}\right)' ,$$

which integrates to $r\mathrm{e}^{2\alpha} = r - C$ with C an integration constant:

$$\mathrm{e}^{2\alpha} = 1 - \frac{C}{r} .$$

Direct substitution confirms that $R_{00} = 0$ then holds identically — as it must, since the contracted Bianchi identity (Equation (13.308)) makes the field equations degenerate.

The constant C is fixed by the Newtonian limit. A slow test particle on a geodesic (Postulate 43.1) has $\mathrm{d}t/\mathrm{d}\tau \simeq 1$ and, by Equation (45.2),

$$\frac{\mathrm{d}^2 r}{\mathrm{d}\tau^2} \simeq -c^2\,\Gamma^r{}_{00} = -c^2\alpha'\mathrm{e}^{4\alpha} \simeq -\frac{c^2 C}{2r^2}$$

at large r. Newtonian gravity requires $\ddot{r} = -GM/r^2$ for the same source, so $C = 2GM/c^2 = r_\mathrm{s}$, which is Equations (45.6) and (45.7). Uniqueness within the ansatz is manifest — every step was forced — and the ansatz itself is generic by the reduction above. □

The scale Equation (45.7) is tiny for ordinary bodies: $r_{\mathrm{s}} = 2.95\,\mathrm{km}$ for the Sun ($M_\odot = 1.989\times10^{30}\,\mathrm{kg}$) and 8.87 mm for the Earth, which is why every solar-system test of Chapter 53 probes the far field $r \gg r_{\mathrm{s}}$, where Equation (45.6) deviates from flatness at the level $r_{\mathrm{s}}/r \sim 10^{-6}$ or less.

Remark 45.2 (The 1916 papers, and whose coordinates these are). Schwarzschild found the solution in the winter of 1915–16, on the Russian front, using a radial coordinate for which his "r" was not the areal radius [440]; the form Equation (45.6), in the coordinates now universally called Schwarzschild coordinates, is due independently to Droste [444]. Neither author regarded the locus $r = r_{\mathrm{s}}$ as physical: for every known star it lies far inside the matter, where the vacuum solution does not apply. The interior problem — matching Equation (45.6) to a matter-filled region and the resulting structure equations — is developed with the compact stars that need it, in Chapter 49.

Remark 45.3 (Gravitational redshift from the static metric). Two static observers at radii $r_1 < r_2$ exchange light. Staticity means successive wave crests follow congruent paths, so the *coordinate* time between crests is the same at both ends, while each observer's proper time is $\mathrm{d}\tau = \sqrt{1 - r_{\mathrm{s}}/r}\,\mathrm{d}t$ by Equation (45.6). The received frequency is therefore shifted by

$$\frac{\nu_2}{\nu_1} = \sqrt{\frac{1 - r_{\mathrm{s}}/r_1}{1 - r_{\mathrm{s}}/r_2}} \simeq 1 - \frac{GM}{c^2}\left(\frac{1}{r_1} - \frac{1}{r_2}\right), \tag{45.8}$$

whose weak-field form is the tower redshift measured by Pound, Rebka and Snider and the clock-rate budget of satellite navigation (Chapters 42 and 53). As $r_1 \to r_{\mathrm{s}}$ the shift diverges: the horizon is an infinite-redshift surface for static sources, a fact that controls the appearance of collapse in Section 45.4.1.

45.1.2 Birkhoff's theorem

Staticity was an input above. It is in fact a consequence of spherical symmetry alone — a theorem proved by Jebsen two years before the customary source [476, 497].

Theorem 45.4 (Jebsen–Birkhoff). *Every spherically symmetric solution of the vacuum field equations is locally isometric to the Schwarzschild solution Equation (45.6). In particular the field outside any spherically symmetric body is static, however violently the body itself pulsates, collapses or explodes, so long as it does so spherically.*

Rests on Theorem 45.1 and Equation (13.306).

Proof. *Derives Theorem 45.4.*

Spherical symmetry alone reduces the metric, wherever the areal radius has spacelike gradient, to the form Equation (45.1) with t-dependent functions $\alpha(t,r)$, $\beta(t,r)$: the orbit spheres fix the angular block, the areal radius supplies r, and the residual freedom in relabelling t removes the single cross term g_{tr} [1509, 1746]. The time dependence adds exactly one new nonvanishing Ricci component,

$$R_{0r} = \frac{2}{r}\,\partial_0\beta\,,$$

with $\partial_0 = \partial/\partial x^0$, and adds to R_{00} and R_{rr} terms each of which carries at least one factor of $\partial_0\beta$ or $\partial_0^2\beta$. Vacuum makes R_{0r} vanish, so $\beta = \beta(r)$; all the new terms in R_{00} and R_{rr} then vanish identically, and the static computation of Theorem 45.1 applies verbatim at each instant. It gives $\partial_r(\alpha + \beta) = 0$, i.e. $\alpha = -\beta + f(t)$; the redefinition $\mathrm{d}\tilde{t} = \mathrm{e}^{f(t)}\mathrm{d}t$ absorbs f, and $R_{\theta\theta} = 0$ integrates as before to $\mathrm{e}^{-2\beta} = 1 - C(t)/r$ — with C in fact constant because β is t-independent. The result is Equation (45.6). □

Corollary 45.5 (No monopole gravitational radiation)**.** *A radially pulsating star cannot radiate gravitationally: its exterior field is exactly the static Equation (45.6) at all times, so no disturbance whatever propagates outward. Gravitational radiation begins at the quadrupole (Chapter 46), and Theorem 45.4 is the exact, nonlinear statement of which the absence of monopole terms in the quadrupole formula is the weak-field shadow.*

Rests on Theorem 45.4.

Theorem 45.4 is also the tool that makes collapse computable: whatever happens inside a collapsing spherical star, the exterior is Equation (45.6) with a constant M. That is the key to Section 45.4.1.

45.2 Geodesics and orbits

45.2.1 The effective potential

The metric Equation (45.6) has two obvious Killing vectors, $\boldsymbol{\xi} = \partial_t$ and $\boldsymbol{\eta} = \partial_\varphi$, and each gives a conserved quantity along geodesics (Proposition 13.141). For a timelike geodesic with four-velocity $u^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$, normalized as always to $g_{\mu\nu}u^\mu u^\nu = -c^2$ (Proposition 40.3, in this part's signature), define the conserved specific energy and specific angular momentum

$$\mathcal{E} := \left(1 - \frac{r_\mathrm{s}}{r}\right) c^2 \frac{\mathrm{d}t}{\mathrm{d}\tau}\,, \qquad L := r^2 \sin^2\theta\, \frac{\mathrm{d}\varphi}{\mathrm{d}\tau}\,. \tag{45.9}$$

Spherical symmetry confines each orbit to a plane: with initial conditions $\theta = \pi/2$, $\mathrm{d}\theta/\mathrm{d}\tau = 0$, the θ geodesic equation gives $\mathrm{d}^2\theta/\mathrm{d}\tau^2 = 0$, so the motion stays equatorial, and we take $\theta = \pi/2$, $L = r^2\mathrm{d}\varphi/\mathrm{d}\tau$ throughout.

Proposition 45.6 (Radial equation and effective potential)**.** *Along an equatorial timelike geodesic,*

$$\frac{1}{2}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 + V_\mathrm{eff}(r) = \mathcal{E}_\mathrm{N} := \frac{\mathcal{E}^2 - c^4}{2c^2}\,, \tag{45.10}$$

with the effective potential

$$V_\mathrm{eff}(r) = -\frac{GM}{r} + \frac{L^2}{2r^2} - \frac{GML^2}{c^2 r^3}\,. \tag{45.11}$$

Rests on Theorem 45.1 and Propositions 13.141 and 40.3.

Derivation. *Derives Proposition 45.6.*

Write out the normalization $g_{\mu\nu}u^\mu u^\nu = -c^2$ in the equatorial plane,

$$-\left(1 - \frac{r_\mathrm{s}}{r}\right) c^2 \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 + \left(1 - \frac{r_\mathrm{s}}{r}\right)^{-1} \left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 + r^2 \left(\frac{\mathrm{d}\varphi}{\mathrm{d}\tau}\right)^2 = -c^2\,,$$

and eliminate the cyclic velocities with Equation (45.9): $\mathrm{d}t/\mathrm{d}\tau = \mathcal{E}/(c^2(1 - r_\mathrm{s}/r))$ and $\mathrm{d}\varphi/\mathrm{d}\tau = L/r^2$. Multiplying through by $\tfrac{1}{2}(1 - r_\mathrm{s}/r)$,

$$\frac{1}{2}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 = \frac{\mathcal{E}^2}{2c^2} - \frac{1}{2}\left(1 - \frac{r_\mathrm{s}}{r}\right)\left(c^2 + \frac{L^2}{r^2}\right).$$

Expanding the product and using $r_\mathrm{s}c^2 = 2GM$ gives Equations (45.10) and (45.11) exactly — no weak-field approximation has been made. □

Remark 45.7 (One new term, and everything it changes). The first two terms of Equation (45.11) are precisely the Newtonian effective potential of the Kepler problem (Equation (27.32) with $U = -GM/r$), with $\mathcal{E}_{\mathrm{N}}$ reducing at $r \gg r_{\mathrm{s}}$, where $\mathcal{E} \simeq c^2 + E_{\mathrm{Newt}}$, to the Newtonian specific energy. The whole of general relativity's modification of bound motion is the single attractive term $-GML^2/(c^2 r^3)$, and it changes the phase portrait qualitatively: in Newtonian mechanics any $L \neq 0$ erects an infinite centrifugal barrier (Section 27.3.2), whereas here $V_{\mathrm{eff}} \to -\infty$ as $r \to 0$ for every L. The barrier survives only for $L > L_{\mathrm{crit}} = 2\sqrt{3}\,GM/c$: below that angular momentum a particle is captured from *any* energy — there are orbits that spiral into the centre, which Newtonian gravity forbids.

Proposition 45.8 (Orbit equation)**.** *In terms of the inverse radius $u := 1/r$, an equatorial timelike geodesic with $L \neq 0$ obeys*

$$\frac{\mathrm{d}^2 u}{\mathrm{d}\varphi^2} + u = \frac{GM}{L^2} + \frac{3GM}{c^2}\,u^2 . \tag{45.12}$$

Dropping the u^2 term gives the Binet equation of the Kepler problem (Equation (27.28) with the inverse-square force), whose solutions are the closed conics of Section 27.4.

Rests on Proposition 45.6 and Equation (27.28).

Derivation. *Derives Proposition 45.8.*

Since $\mathrm{d}\varphi/\mathrm{d}\tau = Lu^2$,

$$\frac{\mathrm{d}r}{\mathrm{d}\tau} = \frac{\mathrm{d}r}{\mathrm{d}\varphi}\,\frac{\mathrm{d}\varphi}{\mathrm{d}\tau} = -\frac{1}{u^2}\,\frac{\mathrm{d}u}{\mathrm{d}\varphi}\cdot Lu^2 = -L\,\frac{\mathrm{d}u}{\mathrm{d}\varphi},$$

so the first integral of Proposition 45.6 reads

$$L^2\left(\frac{\mathrm{d}u}{\mathrm{d}\varphi}\right)^2 = \frac{\mathcal{E}^2}{c^2} - (1 - r_{\mathrm{s}}u)\left(c^2 + L^2u^2\right).$$

Differentiating with respect to φ and dividing by $2L^2\,\mathrm{d}u/\mathrm{d}\varphi$ (legitimate except at turning points, which are isolated),

$$\frac{\mathrm{d}^2 u}{\mathrm{d}\varphi^2} = \frac{r_{\mathrm{s}}c^2}{2L^2} - u + \frac{3r_{\mathrm{s}}}{2}\,u^2 ,$$

which is Equation (45.12) on using $r_{\mathrm{s}}c^2 = 2GM$. □

Phenomenon 45.9 (Relativistic advance of the periastron)**.** A bound orbit in the field of a spherical mass M does not close: the periastron advances in the direction of the motion by

$$\Delta\varphi = \frac{6\pi GM}{c^2 a\,(1 - e^2)} \tag{45.13}$$

per revolution, to lowest order in $GM/(c^2 a)$, for an orbit of semi-major axis a and eccentricity e. Two measurements span nine orders of magnitude in central mass and confirm the same formula. For Mercury the advance is the 43 seconds of arc per century that Le Verrier isolated as a residual half a century before there was a theory to explain it [151, 428]; for the star S2 orbiting the compact mass at the Galactic centre it is about twelve minutes of arc per sixteen-year orbit, detected by near-infrared interferometry [2347].

Rests on Equation (45.12) and Proposition 45.8.

Derivation. *Derives Phenomenon 45.9.*

The orbit equation Equation (45.12) is exact; the advance is its lowest-order consequence. The relativistic term is smaller than the linear terms by $3L^2u^2/c^2 \sim v^2/c^2$, so one solves perturbatively about the Keplerian ellipse: substituting the conic $u_0 = (GM/L^2)(1 + e\cos\varphi)$ into the u^2 term produces, among bounded forced terms, one resonant term proportional to

$\cos\varphi$, and its secular growth rotates the apsidal line by Equation (45.13) per revolution. That computation is carried out in full — including the numerical budget that turns it into Mercury's 43 seconds of arc per century — in the derivation of Phenomenon 53.1 in Chapter 53, which takes Equation (45.12) from this chapter as its input; it is not repeated here. For S2 the same formula, with the mass and orbital elements measured in Chapter 50, gives the twelve-minute advance per orbit reported by the GRAVITY interferometer [2347]. □

45.2.2 Circular orbits, the ISCO and the photon sphere

Circular orbits sit at extrema of Equation (45.11), and the systematic treatment of what follows — including the rotating generalization — is [1464].

Proposition 45.10 (Circular orbits and their energy)**.** *Circular timelike geodesics of radius r exist iff $r > 3GM/c^2$, with*

$$L^2 = \frac{GMr^2}{r - 3GM/c^2}\,, \tag{45.14}$$

and carry specific energy

$$\frac{\mathcal{E}}{c^2} = \frac{1 - r_{\mathrm{s}}/r}{\sqrt{1 - 3GM/(c^2 r)}}\,. \tag{45.15}$$

They are stable iff $r \geq r_{\mathrm{ISCO}} = 6GM/c^2$; at the innermost stable circular orbit, $\mathcal{E} = \sqrt{8/9}\,c^2$ and $L = 2\sqrt{3}\,GM/c$.

Rests on Proposition 45.6.

Derivation. *Derives Proposition 45.10.*

Write $m := GM/c^2$. Setting $V'_{\mathrm{eff}}(r) = 0$ in Equation (45.11) and multiplying by r^4,

$$GMr^2 - L^2 r + 3GM\frac{L^2}{c^2} = 0 \qquad \Longrightarrow \qquad L^2\,(r - 3m) = GMr^2\,,$$

which is Equation (45.14); a positive solution requires $r > 3m$, and $L^2 \to \infty$ as $r \to 3m$. For the energy, evaluate $\mathcal{E}^2 = c^4 + 2c^2 V_{\mathrm{eff}}(r)$ on Equation (45.14): putting the terms of Equation (45.11) over the denominator $2r(r - 3m)$,

$$V_{\mathrm{eff}} = -\frac{GM\,(r - 4m)}{2r\,(r - 3m)} \qquad \Longrightarrow \qquad \frac{\mathcal{E}^2}{c^4} = \frac{r^2 - 4mr + 4m^2}{r\,(r - 3m)} = \frac{(r - 2m)^2}{r\,(r - 3m)}\,,$$

whose square root is Equation (45.15). For stability, differentiate Equation (45.11) twice and insert Equation (45.14):

$$V''_{\mathrm{eff}}(r) = \frac{GM}{r^3}\cdot\frac{r - 6m}{r - 3m}\,, \tag{45.16}$$

positive for $r > 6m$, negative for $3m < r < 6m$: the inner branch of circular orbits is unstable, and the marginal case is $r_{\mathrm{ISCO}} = 6m = 6GM/c^2$. There Equations (45.14) and (45.15) give $L^2 = GM \cdot 36m^2/(3m) = 12G^2M^2/c^2$ and $\mathcal{E}/c^2 = (1 - \tfrac{1}{3})/\sqrt{1 - \tfrac{1}{2}} = \tfrac{2}{3}\sqrt{2} = \sqrt{8/9}$. □

Remark 45.11 (The photon sphere as a limit). As $L \to \infty$ the unstable branch radius tends to $3GM/c^2$: the faster the particle, the closer its unstable circular orbit sits to the limiting value, which is the circular orbit of light itself. The same radius reappears in Section 45.7 as the top of the null effective potential, Equation (45.59), where its instability makes it the rim of the shadow rather than a place where light accumulates. Nothing in the Kepler problem corresponds to either radius: both are creatures of the r^{-3} term of Equation (45.11).

Phenomenon 45.12 (Efficiency of accretion onto compact objects)**.** Matter accreting onto a compact object radiates a fraction of order ten per cent of its rest energy, an order of magnitude above the 0.7% that hydrogen burning can release. This is what allows a quasar to outshine its entire host galaxy from a region light-days across [1265], and it is the argument that identified quasars with accreting massive black holes in the nuclei of ordinary galaxies [1403]. The scale is set by the innermost stable circular orbit: circular geodesics of the Schwarzschild geometry exist inward to $r = 3GM/c^2$ but are stable only outside

$$r_{\mathrm{ISCO}} = \frac{6GM}{c^2} = 3r_{\mathrm{s}}\,, \tag{45.17}$$

where a unit mass has specific energy $\sqrt{8/9}\,c^2$. Matter spiralling slowly inward through a thin disc therefore radiates at most

$$\varepsilon = 1 - \sqrt{\tfrac{8}{9}} = 0.057 \tag{45.18}$$

of its rest energy before it crosses the horizon [1464, 1509]. For a rapidly rotating hole the innermost stable orbit moves inward and the figure rises toward about 0.4, which is why the observed efficiencies are informative about spin.

Rests on Proposition 45.10.

Derivation. *Derives Phenomenon 45.12.*

In a geometrically thin, optically thick disc (Chapter 50 for the observational side) the gas drifts inward through a sequence of nearly circular orbits, each lasting many orbital periods, so its specific energy at radius r is Equation (45.15); the difference from c^2 has been carried away by radiation. The drift ends at r_{ISCO}: inside, no stable circular orbit exists (Equation (45.16)), and the remaining infall to the horizon is dynamical, too fast to radiate efficiently, so the energy still aboard is advected in. The radiated fraction is therefore the binding energy at the innermost stable orbit,

$$\varepsilon = 1 - \frac{\mathcal{E}\left(r_{\mathrm{ISCO}}\right)}{c^2} = 1 - \sqrt{\frac{8}{9}} = 0.0572\,,$$

which is Equation (45.18). For a rotating hole the co-rotating ISCO moves inward — to GM/c^2 in the extremal limit — and the same computation on the Kerr geometry gives binding energies up to $1 - 1/\sqrt{3} = 0.42$; those orbits are catalogued in [1464], and the spin dependence is taken up in Section 45.5.1. □

45.2.3 Null geodesics: deflection and delay

Light follows null geodesics, for which proper time is unavailable; let λ be an affine parameter with dimensions of length and define the conserved quantities

$$E := \left(1 - \frac{r_{\mathrm{s}}}{r}\right) c\,\frac{\mathrm{d}t}{\mathrm{d}\lambda}\,, \qquad L := r^2\,\frac{\mathrm{d}\varphi}{\mathrm{d}\lambda} \tag{45.19}$$

for an equatorial ray (E dimensionless, L a length; both are the constants of Proposition 13.141 in convenient normalizations). The null condition $g_{\mu\nu}\,\mathrm{d}x^\mu \mathrm{d}x^\nu = 0$, treated exactly as in the derivation of Proposition 45.6, gives the first integral

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\lambda}\right)^2 = E^2 - \frac{L^2}{r^2}\left(1 - \frac{r_{\mathrm{s}}}{r}\right)\,. \tag{45.20}$$

Passing to $u = 1/r$ as in Proposition 45.8,

$$\left(\frac{\mathrm{d}u}{\mathrm{d}\varphi}\right)^2 + u^2 = \frac{E^2}{L^2} + r_{\mathrm{s}}u^3 \qquad \Longrightarrow \qquad \frac{\mathrm{d}^2u}{\mathrm{d}\varphi^2} + u = \frac{3GM}{c^2}\,u^2\,, \tag{45.21}$$

the orbit equation Equation (45.12) with the mass term absent — it came from the rest mass. As $r \to \infty$ the first form of Equation (45.21) degenerates to $(\mathrm{d}u/\mathrm{d}\varphi)^2 + u^2 = E^2/L^2$, whose straight-line solution passes the centre at perpendicular distance

$$b = \frac{L}{E}\,, \tag{45.22}$$

the *impact parameter*: the ratio of the two conserved quantities is the one geometric datum a distant observer can assign to a ray.

Remark 45.13 (Light deflection at the formula level). Solving Equation (45.21) perturbatively about the straight line $u_0 = \sin\varphi/b$ shifts the two asymptotes of the ray and bends it, for $b \gg r_\mathrm{s}$, through

$$\Delta\vartheta = \frac{4GM}{c^2 b}\,,$$

which is Equation (53.3). The computation — and the observational history from the 1919 eclipse through modern radio interferometry, where the formula is confirmed to a part in 10^4 — is carried out in full in Chapter 53 (derivation of Phenomenon 53.2), taking Equation (45.21) from this chapter as its input [465]. Half the value comes from the equivalence principle alone; the other half is the spatial curvature visible in Equation (45.23) below.

Proposition 45.14 (Isotropic coordinates)**.** *The substitution* $r = \rho\,(1 + r_\mathrm{s}/(4\rho))^2$ *brings Equation* (45.6) *to the isotropic form*

$$\mathrm{d}s^2 = -\left(\frac{1 - r_\mathrm{s}/(4\rho)}{1 + r_\mathrm{s}/(4\rho)}\right)^2 c^2\mathrm{d}t^2 + \left(1 + \frac{r_\mathrm{s}}{4\rho}\right)^4 \left(\mathrm{d}\rho^2 + \rho^2\mathrm{d}\Omega^2\right)\,, \tag{45.23}$$

whose spatial part is conformally flat. To first order in r_s/ρ *it reads* $\mathrm{d}s^2 \simeq -\,(1 - r_\mathrm{s}/\rho)\,c^2\mathrm{d}t^2 + (1 + r_\mathrm{s}/\rho)\,(\mathrm{d}\rho^2 + \rho^2\mathrm{d}\Omega^2)$.

Rests on Theorem 45.1.

Derivation. *Derives Proposition 45.14.*

Write $x := r_\mathrm{s}/(4\rho)$. Differentiating the substitution, $\mathrm{d}r = (1 + x)\,(1 - x)\,\mathrm{d}\rho$, and a two-line computation gives

$$1 - \frac{r_\mathrm{s}}{r} = \frac{(1 + x)^2 - r_\mathrm{s}/\rho}{(1 + x)^2} = \left(\frac{1 - x}{1 + x}\right)^2\,.$$

Hence $(1 - r_\mathrm{s}/r)^{-1}\,\mathrm{d}r^2 = (1 + x)^4\,\mathrm{d}\rho^2$ and $r^2\mathrm{d}\Omega^2 = (1 + x)^4\,\rho^2\mathrm{d}\Omega^2$: every spatial term acquires the same conformal factor, which is Equation (45.23). Expanding $((1 - x)/(1 + x))^2 \simeq 1 - 4x$ and $(1 + x)^4 \simeq 1 + 4x$ gives the first-order form. □

Remark 45.15 (The Shapiro delay at the formula level). In the isotropic form the coordinate speed of light is isotropic and equals $c\,(1 - r_\mathrm{s}/\rho)$ to first order — one factor from time dilation, one from spatial curvature. Integrating the deficit along a straight path between radii r_1 and r_2 passing the mass at distance b gives the round-trip excess delay

$$\Delta t = \frac{4GM}{c^3}\ln\!\left(\frac{4r_1 r_2}{b^2}\right)\,,$$

which is Equation (53.5), proposed by Shapiro as the fourth classical test [1296]. The integration, and the radar and spacecraft measurements that verify it — now at the 2×10^{-5} level via the Cassini link — are in Chapter 53 (derivation of Phenomenon 53.4), which starts from Equation (45.23).

45.3 The horizon

45.3.1 Coordinate and curvature singularities

At $r = r_{\mathrm{s}}$ the metric component g_{rr} of Equation (45.6) diverges and g_{tt} vanishes. Whether anything physical diverges there is a question the coordinates cannot answer; curvature invariants can, because they are scalars.

Proposition 45.16 (Curvature at the horizon and at the centre)**.** *In the orthonormal frame of a static observer the nonvanishing components of the Riemann tensor of Equation* (45.6) *are*

$$R_{\hat{0}\hat{r}\hat{0}\hat{r}} = -\frac{r_{\mathrm{s}}}{r^3}\,, \quad R_{\hat{0}\hat{\theta}\hat{0}\hat{\theta}} = R_{\hat{0}\hat{\varphi}\hat{0}\hat{\varphi}} = \frac{r_{\mathrm{s}}}{2r^3}\,, \quad R_{\hat{r}\hat{\theta}\hat{r}\hat{\theta}} = R_{\hat{r}\hat{\varphi}\hat{r}\hat{\varphi}} = -\frac{r_{\mathrm{s}}}{2r^3}\,, \quad R_{\hat{\theta}\hat{\varphi}\hat{\theta}\hat{\varphi}} = \frac{r_{\mathrm{s}}}{r^3}\,, \tag{45.24}$$

(up to index symmetries), so the Kretschmann scalar is

$$R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} = 4\left[1+\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+1\right]\frac{r_{\mathrm{s}}^2}{r^6} = \frac{12\,r_{\mathrm{s}}^2}{r^6} = \frac{48\,G^2M^2}{c^4r^6}\,. \tag{45.25}$$

It is finite at $r = r_{\mathrm{s}}$ and diverges only at $r = 0$: the horizon is a coordinate singularity, the centre a curvature singularity.

Rests on Theorem 45.1 and Equation (13.306).

Derivation. *Derives Proposition 45.16.*

Mechanical evaluation of Equation (13.306) on Equation (45.2) with $\mathrm{e}^{2\alpha} = \mathrm{e}^{-2\beta} = 1 - r_{\mathrm{s}}/r$, followed by projection onto the orthonormal frame $\hat{e}_0 = (1-r_{\mathrm{s}}/r)^{-1/2}\,c^{-1}\partial_t$, $\hat{e}_r = (1-r_{\mathrm{s}}/r)^{1/2}\partial_r$, $\hat{e}_\theta = r^{-1}\partial_\theta$, $\hat{e}_\varphi = (r\sin\theta)^{-1}\partial_\varphi$, gives Equation (45.24). In the contraction each of the six independent pair components enters with multiplicity four and, in an orthonormal frame, with a positive square, which is the bracket in Equation (45.25). □

Remark 45.17 (Tides at the horizon can be gentle). By geodesic deviation (Equation (43.3)) the frame components Equation (45.24) are tidal accelerations per unit separation: radial stretching $2GM/r^3$, transverse compression GM/r^3, in exact correspondence with the Newtonian tidal tensor $\partial_i\partial_j\Phi$ — and finite at the horizon, where they equal $2GM/r_{\mathrm{s}}^3 = c^6/(4G^2M^2)$ radially. The scale *falls* with mass: for $M = M_\odot$ it is $1.0\times10^{10}/\mathrm{s}^2$ — an infalling body is shredded long before reaching the horizon — while for the $4.3\times10^6\,M_\odot$ object at the Galactic centre it is $5.6\times10^{-4}/\mathrm{s}^2$, milder than the tide in a fast elevator. Locally the horizon of a large black hole is unremarkable; what distinguishes it is global, as the next construction shows.

The diagnosis of $r = r_{\mathrm{s}}$ as an artefact is completed by exhibiting coordinates regular there. Define the *tortoise coordinate* and the *advanced time*

$$r_* := r + r_{\mathrm{s}}\ln\left|\frac{r}{r_{\mathrm{s}}} - 1\right|\,, \qquad v := ct + r_*\,, \tag{45.26}$$

so that $\mathrm{d}r_* = (1-r_{\mathrm{s}}/r)^{-1}\,\mathrm{d}r$ and ingoing radial light rays are $v = \text{const}$. Substituting $c\,\mathrm{d}t = \mathrm{d}v - \mathrm{d}r_*$ into Equation (45.6),

$$\mathrm{d}s^2 = -\left(1-\frac{r_{\mathrm{s}}}{r}\right)\mathrm{d}v^2 + 2\,\mathrm{d}v\,\mathrm{d}r + r^2\mathrm{d}\Omega^2\,, \tag{45.27}$$

the *ingoing Eddington–Finkelstein form*: every component is regular for all $r > 0$, and the determinant $g = -r^4\sin^2\theta$ never vanishes there. The coordinate v was introduced by Eddington [510], its meaning recognized by Lemaître [723] and sharpened by Finkelstein [1145].

Proposition 45.18 (The one-way membrane)**.** *In the geometry Equation (45.27), every future-directed causal curve at $r \le r_{\mathrm{s}}$ has $\mathrm{d}r \le 0$, with equality only for radial null curves on $r = r_{\mathrm{s}}$ itself. The surface $r = r_{\mathrm{s}}$ is a null hypersurface that can be crossed inward but not outward: the* event horizon*.*

Rests on Theorem 45.1.

Derivation. *Derives Proposition 45.18.*

Radial null directions of Equation (45.27) satisfy $\mathrm{d}v\,[(1 - r_{\mathrm{s}}/r)\,\mathrm{d}v - 2\,\mathrm{d}r] = 0$: either $\mathrm{d}v = 0$, the ingoing family, along which r strictly decreases toward the future; or $\mathrm{d}r/\mathrm{d}v = \frac{1}{2}(1 - r_{\mathrm{s}}/r)$, the would-be outgoing family — positive outside the horizon, zero on it, *negative* inside. A causal curve lies inside or on the local light cone spanned by the two, so inside the horizon both bounding directions, and hence every causal direction, have $\mathrm{d}r < 0$. On $r = r_{\mathrm{s}}$ the normal $n_\mu = \partial_\mu (r - r_{\mathrm{s}})$ has $n^\mu n_\mu = g^{rr} = 1 - r_{\mathrm{s}}/r = 0$: the surface is null, generated by the marginally outgoing rays that run along it forever. □

Nothing locally distinguishes the horizon — the curvature there is finite and can be arbitrarily small (Remark 45.17) — and no signal sent from inside ever reaches a distant observer. The horizon is a *globally* defined surface: the boundary of the region from which future null infinity cannot be reached [1504]. Inside it the roles of r and t interchange — $g_{rr} < 0$, $g_{tt} > 0$ in Equation (45.6) — so decreasing r is there a direction in *time*: reaching $r = 0$ is as unavoidable as reaching tomorrow.

45.3.2 The Kruskal–Szekeres extension

Eddington–Finkelstein coordinates extend the exterior across the future horizon. The *maximal* analytic extension, found independently by Kruskal and Szekeres [1185, 1194], is reached by one further exponential substitution.

Proposition 45.19 (Kruskal–Szekeres coordinates)**.** *On the exterior $r > r_{\mathrm{s}}$ define null coordinates $u := ct - r_*$, $v := ct + r_*$ and*

$$U := -\mathrm{e}^{-u/(2r_{\mathrm{s}})}\,, \qquad V := \mathrm{e}^{v/(2r_{\mathrm{s}})}\,, \qquad T := \frac{U+V}{2}\,, \qquad X := \frac{V-U}{2}\,. \tag{45.28}$$

Then

$$\mathrm{d}s^2 = \frac{4r_{\mathrm{s}}^3}{r}\,\mathrm{e}^{-r/r_{\mathrm{s}}}\left(-\mathrm{d}T^2 + \mathrm{d}X^2\right) + r^2\mathrm{d}\Omega^2\,, \tag{45.29}$$

with $r(T,X) > 0$ determined implicitly by

$$X^2 - T^2 = \left(\frac{r}{r_{\mathrm{s}}} - 1\right)\mathrm{e}^{r/r_{\mathrm{s}}}\,, \qquad \frac{T}{X} = \tanh\frac{ct}{2r_{\mathrm{s}}} \quad \textit{(region } X > |T|\,)\,. \tag{45.30}$$

The metric Equation (45.29) is regular on the whole domain $T^2 - X^2 < 1$, which is the maximal analytic extension of the Schwarzschild geometry.

Rests on Theorem 45.1.

Derivation. *Derives Proposition 45.19.*

In null coordinates the (t,r) block of Equation (45.6) is $-(1 - r_{\mathrm{s}}/r)\,\mathrm{d}u\,\mathrm{d}v$. From Equation (45.28),

$$\mathrm{d}U\,\mathrm{d}V = \frac{1}{4r_{\mathrm{s}}^2}\,\mathrm{e}^{(v-u)/(2r_{\mathrm{s}})}\,\mathrm{d}u\,\mathrm{d}v = \frac{1}{4r_{\mathrm{s}}^2}\,\mathrm{e}^{r_*/r_{\mathrm{s}}}\,\mathrm{d}u\,\mathrm{d}v\,,$$

and by Equation (45.26), $\mathrm{e}^{r_*/r_{\mathrm{s}}} = (r/r_{\mathrm{s}} - 1)\,\mathrm{e}^{r/r_{\mathrm{s}}}$ outside the horizon. Hence

$$-\left(1 - \frac{r_{\mathrm{s}}}{r}\right)\mathrm{d}u\,\mathrm{d}v = -\left(1 - \frac{r_{\mathrm{s}}}{r}\right)\frac{4r_{\mathrm{s}}^2\,\mathrm{e}^{-r/r_{\mathrm{s}}}}{r/r_{\mathrm{s}} - 1}\,\mathrm{d}U\,\mathrm{d}V = -\frac{4r_{\mathrm{s}}^3}{r}\,\mathrm{e}^{-r/r_{\mathrm{s}}}\,\mathrm{d}U\,\mathrm{d}V\,,$$

since $(1-r_{\mathrm{s}}/r)\,/\,(r/r_{\mathrm{s}}-1) = r_{\mathrm{s}}/r$. Rewriting $\mathrm{d}U\,\mathrm{d}V$ in (T,X) gives Equation (45.29); the identities $UV = T^2 - X^2$ and $UV = -\mathrm{e}^{r_*/r_{\mathrm{s}}}$ give the first of Equation (45.30), and $-V/U = \mathrm{e}^{ct/r_{\mathrm{s}}}$ the second. The right-hand side of Equation (45.30) is a monotonic function of r mapping $(0,\infty)$ onto $(-1,\infty)$, so r is well defined and positive exactly where $X^2 - T^2 > -1$, and the conformal factor in Equation (45.29) is analytic and nonzero there. The extension is maximal because every geodesic either runs to infinite affine parameter or ends on the curvature singularity $T^2 - X^2 = 1$, i.e. $r = 0$ (Proposition 45.16). □

Remark 45.20 (Anatomy of the extended geometry). The two lines $T = \pm X$ divide the domain into four wedges. Region I ($X > |T|$) is the original exterior, with static observers on hyperbolae of constant $X^2 - T^2$ and constant-t lines through the origin. Region II ($T > |X|$) is the black-hole interior: every future-directed causal curve entering it ends on the singularity $T = \sqrt{1+X^2}$, the interior statement of Proposition 45.18. Region III ($T < -|X|$) is the time-reverse — a *white hole*, out of which everything must pass — and region IV ($X < -|T|$) is a second asymptotically flat exterior, causally disconnected from region I: the two are joined only through the spacelike "throat", and no causal curve crosses from one to the other. The conformal compactification of this diagram is the prototype Penrose diagram [1509]. Regions III and IV are artefacts of exact vacuum eternity: in a black hole formed by collapse the matter's interior occupies the part of the diagram where they would lie, and only regions I and II survive (Section 45.4.1). The white hole and the second universe are not predictions about Nature; the black hole is.

45.4 Gravitational collapse

45.4.1 Oppenheimer–Snyder collapse

The simplest collapse that can be followed exactly is a uniform ball of pressureless dust [844]. Outside, spherical symmetry forces the geometry to be Schwarzschild (Theorem 45.4); the surface of the ball falls along a radial timelike geodesic, so the first computation is radial free fall.

Proposition 45.21 (Radial plunge: the cycloid)**.** *A test particle released from rest at $r = R_0$ falls along the cycloid*

$$r = \frac{R_0}{2}\,(1+\cos\eta)\,, \qquad \tau = \sqrt{\frac{R_0^3}{8GM}}\,(\eta+\sin\eta)\,, \tag{45.31}$$

reaching $r = 0$ at the finite proper time

$$\tau_{\mathrm{collapse}} = \pi\sqrt{\frac{R_0^3}{8GM}}\,, \tag{45.32}$$

identical to the Newtonian free-fall time from the same radius. The horizon is crossed strictly earlier, also at finite proper time.

Rests on Proposition 45.6.

Derivation. *Derives Proposition 45.21.*

Set $L = 0$ in Proposition 45.6; release from rest at R_0 fixes $\mathcal{E} = c^2\sqrt{1 - r_{\mathrm{s}}/R_0}$, so

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 = \frac{\mathcal{E}^2}{c^2} - c^2\left(1-\frac{r_{\mathrm{s}}}{r}\right) = c^2 r_{\mathrm{s}}\left(\frac{1}{r}-\frac{1}{R_0}\right) = 2GM\left(\frac{1}{r}-\frac{1}{R_0}\right).$$

This is *exactly* the Newtonian energy equation — the relativistic corrections have been absorbed into the meaning of r and τ — and the substitution Equation (45.31) solves it:

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 = \frac{(R_0^2/4)\sin^2\eta}{(R_0^3/8GM)\,(1+\cos\eta)^2} = \frac{2GM}{R_0}\cdot\frac{1-\cos\eta}{1+\cos\eta} = 2GM\left(\frac{1}{r}-\frac{1}{R_0}\right),$$

using $\sin^2\eta = (1-\cos\eta)(1+\cos\eta)$ and $R_0/r = 2/(1+\cos\eta)$. The collapse time is $\eta = \pi$ in Equation (45.31); since $r = r_\mathrm{s}$ is reached at some $\eta < \pi$, the horizon crossing is earlier and finite. For a ball of one solar mass released from the Sun's present radius, Equation (45.32) gives $\tau_\mathrm{collapse} = 1.8 \times 10^3\,\mathrm{s}$ — half an hour of proper time from star to singularity. □

Proposition 45.22 (The view from outside: exponential freezing)**.** *To a distant static observer the same infall never ends: the areal radius of the falling surface asymptotes to r_s, and light received from it at late observer time t is redshifted as*

$$1 + z \;\propto\; \exp\!\left(\frac{ct}{2r_\mathrm{s}}\right) = \exp\!\left(\frac{c^3 t}{4GM}\right), \tag{45.33}$$

an e-folding time $4GM/c^3 = 2.0 \times 10^{-5}\,\mathrm{s}$ per solar mass. The luminosity dies on the same timescale: within milliseconds a stellar collapse is, observationally, black.

Rests on Proposition 45.21 and Remark 45.3.

Derivation. *Derives Proposition 45.22.*

Near the horizon write $r = r_\mathrm{s}(1+\epsilon)$, $\epsilon \ll 1$. Dividing the radial equation of Proposition 45.21 by $\mathrm{d}t/\mathrm{d}\tau = \mathcal{E}/(c^2(1-r_\mathrm{s}/r))$ gives $\mathrm{d}r/\mathrm{d}t \simeq -c\,(1-r_\mathrm{s}/r) \simeq -c\,\epsilon$ as $\epsilon \to 0$ (the remaining factors tend to constants), so $r_\mathrm{s}\,\mathrm{d}\epsilon/\mathrm{d}t \simeq -c\,\epsilon$ and $\epsilon \propto \mathrm{e}^{-ct/r_\mathrm{s}}$: the surface approaches the horizon exponentially in coordinate time and never crosses it at any finite t.

For the redshift, compound two factors at emission radius r. Relative to the static observer there, the faller has Lorentz factor $\gamma = -g_{\mu\nu}u^\mu_\mathrm{fall}u^\nu_\mathrm{stat}/c^2 = \mathcal{E}/(c^2\sqrt{1-r_\mathrm{s}/r}) \propto \epsilon^{-1/2}$, so light it emits radially outward is Doppler-shifted by $\sqrt{(1-v/c)/(1+v/c)} \simeq 1/(2\gamma) \propto \sqrt{\epsilon}$; climbing out then costs the gravitational factor $\sqrt{1-r_\mathrm{s}/r} \simeq \sqrt{\epsilon}$ (Equation (45.8)). The received frequency is suppressed by the product, of order ϵ. Finally, the outgoing photon spends a coordinate time $\simeq (r_\mathrm{s}/c)\ln(1/\epsilon)$ escaping the near zone (integrating $\mathrm{d}t = \mathrm{d}r_*/c$), which at $\epsilon \propto \mathrm{e}^{-ct_\mathrm{em}/r_\mathrm{s}}$ doubles the elapsed time: $t \simeq 2t_\mathrm{em}$. Hence $1+z \propto \epsilon^{-1} \propto \mathrm{e}^{ct_\mathrm{em}/r_\mathrm{s}} = \mathrm{e}^{ct/(2r_\mathrm{s})}$, which is Equation (45.33); with $M = M_\odot$, $2r_\mathrm{s}/c = 1.97 \times 10^{-5}\,\mathrm{s}$. The luminosity, proportional to the received photon rate times the energy per photon, falls at least as fast. □

The interior completes the picture, and its computation is the matching of two exact solutions along the falling surface.

Theorem 45.23 (Oppenheimer–Snyder collapse)**.** *Let a uniform, pressureless, momentarily static ball of dust of mass M and initial areal radius $R_0 > r_\mathrm{s}$ collapse under its own gravity. Then: (i) the interior geometry is a portion $0 \le \chi \le \chi_0$ of a closed matter-dominated Friedmann–Lemaître–Robertson–Walker spacetime (Chapter 48),*

$$\mathrm{d}s^2 = -c^2\mathrm{d}\tau^2 + a^2(\tau)\left[\mathrm{d}\chi^2 + \sin^2\chi\,\mathrm{d}\Omega^2\right], \tag{45.34}$$

whose scale factor traverses the collapsing half of the dust cycloid; (ii) the exterior is the Schwarzschild geometry Equation (45.6), and the two match smoothly across the surface — induced metric and extrinsic curvature both continuous — iff

$$R_0 = a_\mathrm{m}\sin\chi_0\,, \qquad r_\mathrm{s} = a_\mathrm{m}\sin^3\chi_0\,, \tag{45.35}$$

with a_m the initial (maximal) scale factor; equivalently $M = \frac{4}{3}\pi\rho_0 R_0^3$ with ρ_0 the initial density — the Schwarzschild mass is exactly the Newtonian mass of the ball; (iii) every dust particle reaches the central singularity at the same finite proper time $\tau = \pi a_\mathrm{m}/(2c)$, the surface crossing the horizon strictly earlier, while a distant observer sees the exponential freezing of Proposition 45.22.

Rests on Theorem 45.4 and Proposition 45.21.

Full derivation: Appendix A.2 (p. 2771).
Derives Theorem 45.23.

Phenomenon 45.24 (Compact objects exist above every cold-support limit)**.** Dark objects are observed with masses far above the maximum that any cold equation of state consistent with causality can hold up. X-ray binaries yield dynamical masses between about $5\,M_\odot$ and $20\,M_\odot$ for the accreting component [2103], and the gravitational-wave merger catalogue reaches tens of solar masses [2264, 2324], against a neutron-star maximum bounded near $3\,M_\odot$ (Chapter 49). No equilibrium configuration is available to such objects, and the idealized collapse of a pressureless homogeneous ball ends at a horizon after finite proper time on the infalling matter, while a distant observer sees the surface asymptote to r_s with exponentially growing redshift and never quite arrive [844]. The observational record of these objects is Chapter 50.
Rests on Theorems 45.23 and 45.27 and Proposition 45.22.

Derivation. *Derives Phenomenon 45.24.*
The absence of equilibria above the maximum mass is the structure theory of Chapter 49. Given that, the collapse statements are Proposition 45.21 (finite proper time to horizon and singularity), Proposition 45.22 (exponential fading and freezing outside), and Theorem 45.23 (the complete interior–exterior solution, proved in Appendix A.2). The idealization — exact sphericity, zero pressure — is dispensable: once the surface is inside r_s, trapped surfaces exist (Remark 45.26), and Theorem 45.27 makes the singular outcome generic, independent of symmetry. What no theorem supplies is the horizon itself in the general case; that is the censorship conjecture of Section 45.4.3. □

45.4.2 The singularity theorems

Is the singularity an artefact of exact spherical symmetry — all the matter aimed at one point? Penrose's 1965 theorem [1319], for which the 2020 Nobel prize in physics was awarded, says no: what forces it is not symmetry but a trapped surface, and trapped surfaces are stable under perturbations.

Definition 45.25 (Trapped surface)**.** Let S be a closed spacelike 2-surface. Through each point of S run two future-directed families of null geodesics orthogonal to S — outgoing and ingoing — and let $\theta_\pm$ be their expansions: the fractional rate of change of the area element dragged along each family. S is *trapped* when both $\theta_+ < 0$ and $\theta_- < 0$: even the outgoing light front loses area. In flat spacetime every round sphere has $\theta_+ > 0$; trapping is the statement that gravity has bent the outgoing light cone back onto itself.

Remark 45.26 (Collapse produces trapped surfaces). In the Eddington–Finkelstein geometry Equation (45.27), consider the sphere $(v, r) = \text{const}$ with $r < r_\mathrm{s}$. The two orthogonal null families are those of Proposition 45.18: the ingoing rays have $\mathrm{d}r < 0$, and the "outgoing" ones have $\mathrm{d}r/\mathrm{d}v = \frac{1}{2}(1 - r_\mathrm{s}/r) < 0$ as well. Since the area is $4\pi r^2$, both expansions are negative: every sphere inside the horizon is trapped. In the Oppenheimer–Snyder spacetime the exterior is exactly this geometry once the surface has passed r_s (Theorem 45.23), so trapped surfaces form in finite proper time — and they persist under any sufficiently small perturbation of the initial data, because $\theta_\pm < 0$ are open conditions.

Theorem 45.27 (Penrose; imported)**.** *Let (M, g) satisfy the null convergence condition $R_{\mu\nu}k^\mu k^\nu \geq 0$ for every null vector k — which the field equations translate into the null energy condition on matter (Section 44.8) — and contain a noncompact Cauchy hypersurface (a spacelike hypersurface met exactly once by every inextendible causal curve). If M contains a trapped surface, then M is future null geodesically incomplete: some outgoing light ray ends at finite affine parameter [1319].*
Rests on Definition 45.25.

Theorem 45.28 (Hawking–Penrose; imported)**.** *Let (M, g) satisfy the timelike and null convergence conditions (strong energy condition via the field equations, Section 44.8), the generic condition (every causal geodesic feels some curvature), and contain no closed timelike curves. If in addition it contains a trapped surface, or a compact achronal hypersurface, or a point whose past light cone starts reconverging, then M contains at least one incomplete timelike or null geodesic [1425].*

Rests on Definition 45.25.

Remark 45.29 (Status of the proofs, and what "singularity" means). Both proofs are global variational arguments — Raychaudhuri focusing plus a topological obstruction — and belong to mathematics; the standard complete accounts are [1319, 1425] and the monograph [1504], and this treatise imports them. What the theorems deliver is deliberately weak: *geodesic incompleteness*, a freely falling history that cannot be continued, not a place of infinite curvature — though in every explicitly known collapse solution the incompleteness is accompanied by divergent curvature, as in Proposition 45.16. Their hypotheses matter: quantum fields violate the pointwise energy conditions (Chapter 109), which is one reason the theorems say nothing about what replaces the singularity, a question quarantined to Chapter 135. The cosmological instance of Theorem 45.28 — the universe's own past incompleteness — belongs to Chapter 48.

45.4.3 Cosmic censorship

The singularity theorems guarantee incompleteness; they do not guarantee a horizon around it. The *weak cosmic censorship conjecture* [1405] asserts that, for generic asymptotically flat initial data with physically reasonable matter, the singularities of gravitational collapse are hidden behind event horizons — no "naked" singularity is visible from infinity, and the exterior remains predictable. Despite half a century of effort it remains a conjecture: no proof exists, and the fine-tuned counterexamples known (critical collapse, certain non-generic dust configurations) are exactly the non-generic cases the conjecture excludes. This treatise therefore treats censorship as a working hypothesis, not a result: it underwrites the astrophysical use of the Kerr geometry below, it is what the area theorem (Theorem 45.47) assumes, and its only observational support to date is indirect — every measured compact-object spin respects the Kerr bound (Phenomenon 45.35). The honest accounting of its status sits with the other open problems in Chapter 133.

45.5 Rotating and charged black holes

45.5.1 The Kerr solution

Astrophysical bodies rotate, and their collapse remnants must carry angular momentum. The stationary axisymmetric vacuum solution that describes them was found by Kerr [1259]; no elementary route to it exists, and its derivation is imported — Kerr's own, via algebraically special metrics, and the systematic construction in [1717].

Theorem 45.30 (Kerr solution; imported)**.** *For each mass M and angular momentum J, define*

$$a := \frac{J}{Mc}\,, \qquad \Sigma := r^2 + a^2\cos^2\theta\,, \qquad \Delta := r^2 - r_{\mathrm{s}} r + a^2\,. \tag{45.36}$$

The metric, in the coordinates of Boyer and Lindquist [1342],

$$\begin{aligned} \mathrm{d}s^2 = &-\left(1 - \frac{r_{\mathrm{s}} r}{\Sigma}\right) c^2 \mathrm{d}t^2 - \frac{2 r_{\mathrm{s}} r\, a \sin^2\theta}{\Sigma}\, c\,\mathrm{d}t\,\mathrm{d}\varphi + \frac{\Sigma}{\Delta}\,\mathrm{d}r^2 + \Sigma\,\mathrm{d}\theta^2 \\ &+ \left(r^2 + a^2 + \frac{r_{\mathrm{s}} r\, a^2 \sin^2\theta}{\Sigma}\right) \sin^2\theta\,\mathrm{d}\varphi^2\,, \end{aligned} \tag{45.37}$$

is a stationary, axisymmetric, asymptotically flat solution of the vacuum field equations [1259, 1342]. At $a = 0$ it reduces to Equation (45.6); at $M = 0$ it is flat spacetime in spheroidal coordinates; at large r its $g_{t\varphi}$ identifies J as the source's angular momentum.
Rests on Theorem 45.1.

That Equation (45.37) solves $R_{\mu\nu} = 0$ is a finite mechanical verification; the discovery was finding it. Its properties, by contrast, are derivable in a few lines each, and they are the physics.

Proposition 45.31 (Horizons and the extremal bound)**.** *The surfaces $\Delta = 0$, at*

$$r_{\pm} = \frac{GM}{c^2} \pm \sqrt{\left(\frac{GM}{c^2}\right)^2 - a^2}\,, \tag{45.38}$$

are null hypersurfaces; r_+ is the event horizon. They exist iff $|a| \le GM/c^2$, i.e. iff the dimensionless spin $\chi = cJ/(GM^2) = ac^2/(GM)$ satisfies $|\chi| \le 1$; at $|\chi| = 1$ (the extremal *hole) the two coincide, and for $|\chi| > 1$ the metric has no horizon at all — its ring singularity at $\Sigma = 0$ would be naked.*
Rests on Theorem 45.30.

Derivation. *Derives Proposition 45.31.*
$\Delta = r^2 - r_{\mathrm{s}} r + a^2$ is a quadratic in r with roots Equation (45.38), real iff $(r_{\mathrm{s}}/2)^2 \ge a^2$, which is $|a| \le GM/c^2$. A surface of constant r has normal $n_\mu = \partial_\mu r$ with norm $g^{\mu\nu} n_\mu n_\nu = g^{rr} = \Delta/\Sigma$, which vanishes exactly on $\Delta = 0$: those surfaces are null, hence one-way membranes for causal curves exactly as in Proposition 45.18. Curvature invariants diverge only on $\Sigma = 0$ — the equatorial ring $r = 0$, $\theta = \pi/2$ — so $r_\pm$ are coordinate singularities removable by Eddington–Finkelstein-type coordinates. With $a = \chi GM/c^2$, the reality condition is $|\chi| \le 1$. □

Proposition 45.32 (Static limit, ergosphere and frame dragging)**.** *(i) Static observers (fixed r, θ, φ) exist only outside the* static limit

$$r_E(\theta) = \frac{GM}{c^2} + \sqrt{\left(\frac{GM}{c^2}\right)^2 - a^2 \cos^2\theta}\,, \tag{45.39}$$

which lies outside r_+ everywhere except at the poles. In the ergosphere *$r_+ < r < r_E$ every observer is dragged around the hole. (ii) An observer with zero angular momentum orbits at*

$$\Omega(r,\theta) := \frac{\mathrm{d}\varphi}{\mathrm{d}t} = \frac{r_{\mathrm{s}} r\, a\, c}{\left(r^2 + a^2\right)^2 - a^2 \Delta \sin^2\theta}\,, \tag{45.40}$$

falling off at large r as $\Omega \simeq 2GJ/(c^2 r^3)$; on the horizon the dragging rate becomes uniform,

$$\Omega_{\mathrm{H}} = \frac{a\,c}{r_+^2 + a^2} = \frac{a\,c}{r_{\mathrm{s}} r_+}\,, \tag{45.41}$$

the angular velocity of the black hole.
Rests on Theorem 45.30.

Derivation. *Derives Proposition 45.32.*
(i) A static worldline has tangent $\propto \partial_t$, timelike iff $g_{tt} < 0$, i.e. $\Sigma > r_{\mathrm{s}} r$; the boundary $r^2 - r_{\mathrm{s}} r + a^2 \cos^2\theta = 0$ gives Equation (45.39), which exceeds r_+ of Equation (45.38) wherever $\sin\theta \neq 0$. Inside it ∂_t is spacelike: remaining at rest with respect to infinity would require moving faster than light against the dragging. (ii) Zero angular momentum means $u_\varphi = g_{\varphi t}\, u^t + g_{\varphi\varphi} u^\varphi = 0$, so $\Omega = u^\varphi/u^t = -g_{t\varphi}/g_{\varphi\varphi}$. Reading the components off Equation (45.37) and using the identity $\Sigma\,(r^2 + a^2) + r_{\mathrm{s}} r\, a^2 \sin^2\theta = \left(r^2 + a^2\right)^2 - a^2 \Delta \sin^2\theta$

gives Equation (45.40); the far-field fall-off follows from $r_s a\, c = 2GJ/c^2$. On $\Delta = 0$ the denominator is $(r_+^2 + a^2)^2$ and $r_s r_+ = r_+^2 + a^2$ (from $\Delta(r_+) = 0$), whence Equation (45.41), independent of θ: the horizon rotates rigidly. The weak-field limit of Equation (45.40) is the Lense–Thirring precession rate whose measurement by Gravity Probe B and lunar laser ranging is reported in Chapter 53. □

Remark 45.33 (The Penrose process). Inside the ergosphere the Killing energy $\mathcal{E} = -\xi_\mu u^\mu$ of a particle can be *negative*, because $\boldsymbol{\xi} = \partial_t$ is there spacelike. A body that splits in the ergosphere can therefore send one fragment into the hole with negative energy and eject the other to infinity with *more* energy than went in: rotational energy is extracted [1405]. The process cannot be repeated indefinitely — each extraction spins the hole down, and the bookkeeping of what is extractable is exactly the irreducible-mass inequality Equation (45.50), quantified in Section 45.6.1: at most $1 - 1/\sqrt{2} \approx 29\%$ of the mass–energy of an extremal hole. Its astrophysical descendants (magnetically mediated extraction from accreting holes) power the relativistic jets observed in Chapter 50.

Remark 45.34 (Carter's constant and integrability). The Kerr metric has only the two Killing vectors ∂_t and ∂_φ — two constants of motion where the Schwarzschild case, with its full rotation group, had three. Geodesic motion is nevertheless integrable: Carter found a fourth invariant by separating the Hamilton–Jacobi equation (Chapter 23, whose separability machinery, Section 23.2, is exactly what is applied) in Boyer–Lindquist coordinates [1366]. The invariant is quadratic in the momenta and derives from a rank-two Killing tensor (Definition 13.142) that is not a product of Killing vectors — the physically important instance of that definition, and a hidden symmetry with no isometry behind it. All Kerr geodesics used below (accretion orbits, photon rings, the spin-dependent ISCO of [1464]) rest on this integrability.

Phenomenon 45.35 (Astrophysical black holes rotate, and stay inside the Kerr bound)**.** The compact objects in accreting binaries and the remnants of gravitational-wave mergers carry angular momentum. Writing the dimensionless spin as

$$\chi := \frac{cJ}{GM^2}\,, \tag{45.42}$$

measured values are non-zero in many systems and in no case exceed $|\chi| = 1$, the value at which the two horizons of the Kerr solution merge and beyond which the solution has no horizon at all. The broad, gravitationally redshifted iron Kα line emitted by the inner accretion disc is one measurement channel [1911]; the remnant of the first detected binary merger, for which $\chi \approx 0.67$, is another [2264]. That the bound is respected wherever it has been tested is the only observational evidence bearing on cosmic censorship (Section 45.4.3).

Rests on Proposition 45.31.

Derivation. *Derives Phenomenon 45.35.*

The bound itself is Proposition 45.31: horizons exist iff $|\chi| \le 1$, the extremal case being the merger of $r_\pm$ at $r = GM/c^2$, so a measured $|\chi| > 1$ would mean either a naked singularity or a non-Kerr object — either way a violation of the framework whose consistency is being tested. The two channels lean on different parts of this chapter: the iron-line profile is set by the spin-dependent inner edge of the disc (Remark 45.36: the co-rotating ISCO sweeps from $6GM/c^2$ at $\chi = 0$ to GM/c^2 at $\chi = 1$, and the gravitational redshift Equation (45.8) from that edge is what broadens the line [1911]); the merger-remnant spin is inferred from the ringdown frequencies, whose dependence on (M, J) is Equation (45.45) below. Both are assembled, with their apparatus, in Chapter 50 and Chapter 47. □

Remark 45.36 (The spin-dependent ISCO). Repeating the circular-orbit analysis of Section 45.2.2 on Equation (45.37) — the computation, using Carter's integrability, is carried

out in [1464] and its results are imported here — gives an ISCO that depends on spin and on orbital sense: for co-rotating equatorial orbits it descends from $6GM/c^2$ at $\chi = 0$ to GM/c^2 at $\chi = 1$, while counter-rotating orbits retreat to $9GM/c^2$. The bounding energy rises correspondingly: the extremal co-rotating ISCO has $\mathcal{E} = c^2/\sqrt{3}$, so thin-disc accretion can radiate up to $1 - 1/\sqrt{3} = 0.42$ of the rest energy — the origin of the "toward 0.4" in Phenomenon 45.12, and the lever by which disc spectroscopy measures spin.

45.5.2 Charge: Reissner–Nordström and Kerr–Newman

Charge is the one other parameter a stationary black hole can carry (Section 45.5.3). The static charged solution was found within two years of Schwarzschild's [439, 457], and unlike Kerr's it is an exercise: the Schwarzschild derivation reruns with the Coulomb field's stress–energy as source.

Proposition 45.37 (Reissner–Nordström solution)**.** *The unique spherically symmetric, asymptotically flat solution of the Einstein–Maxwell equations with charge Q and mass M is*

$$\mathrm{d}s^2 = -f(r)\,c^2\mathrm{d}t^2 + f(r)^{-1}\mathrm{d}r^2 + r^2\mathrm{d}\Omega^2\,, \qquad f(r) = 1 - \frac{r_\mathrm{s}}{r} + \frac{r_Q^2}{r^2}\,, \tag{45.43}$$

with the radial Coulomb field $E = Q/(4\pi\epsilon_0 r^2)$ and the charge length

$$r_Q^2 := \frac{GQ^2}{4\pi\epsilon_0 c^4}\,. \tag{45.44}$$

Horizons sit at $r_\pm = r_\mathrm{s}/2 \pm \sqrt{r_\mathrm{s}^2/4 - r_Q^2}$, existing iff $r_Q \le r_\mathrm{s}/2$.

Rests on Theorem 45.1.

Derivation. *Derives Proposition 45.37.*

Spherical symmetry admits only a radial electrostatic field. The curved-space source-free Maxwell equations reduce, in the ansatz Equation (45.1) with $\alpha + \beta = 0$, to $\partial_r\,(r^2 E) = 0$ — the flux integral — so $E = Q/(4\pi\epsilon_0 r^2)$ exactly as in flat space. Its stress–energy is that of an electromagnetic field with energy density $u_\mathrm{em} = \epsilon_0 E^2/2$ (Section 60.4.2), radial tension equal to the energy density and isotropic transverse pressure: $T^0{}_0 = T^r{}_r = -u_\mathrm{em}$, $T^\theta{}_\theta = T^\varphi{}_\varphi = +u_\mathrm{em}$. Because $T^0{}_0 = T^r{}_r$, the combination that gave $\alpha + \beta = \text{const}$ in the proof of Theorem 45.1 does so again, and asymptotic flatness sets $\beta = -\alpha$. The tensor is traceless, so $R = 0$ and the field equations read $R_{\mu\nu} = (8\pi G/c^4)T_{\mu\nu}$; the $\theta\theta$ component becomes

$$\left(r\mathrm{e}^{2\alpha}\right)' = 1 - \frac{8\pi G}{c^4}\,r^2 u_\mathrm{em} = 1 - \frac{GQ^2}{4\pi\epsilon_0 c^4}\cdot\frac{1}{r^2}\,,$$

which integrates to $r\mathrm{e}^{2\alpha} = r - r_\mathrm{s} + r_Q^2/r$, i.e. Equation (45.43). The horizon condition $f = 0$ is a quadratic with the stated roots. □

Remark 45.38 (Kerr–Newman, and why charge is academic). The rotating charged solution — charge Q grafted onto Equation (45.37) by $\Delta \to r^2 - r_\mathrm{s} r + a^2 + r_Q^2$, with the accompanying electromagnetic field — was found by Newman and collaborators [1318] and is imported here; it completes the stationary family of Section 45.5.3. Astrophysically, charge is negligible by two crushing margins. For charge to matter geometrically requires $r_Q \sim r_\mathrm{s}/2$, i.e. $Q \sim \sqrt{4\pi\epsilon_0 G}\,M \approx 1.7 \times 10^{20}\,\mathrm{C}$ per solar mass. But a hole charged to just $Q_\mathrm{eq} = 4\pi\epsilon_0 G M m_\mathrm{p}/e \approx 1.5 \times 10^2\,\mathrm{C}$ per solar mass already attracts protons electrostatically as strongly as it does gravitationally: surrounding plasma neutralizes any astrophysical hole to about one part in 10^{18} of the geometrically relevant charge. Charge is therefore a logical parameter of the classification, not an observable of astronomy, and (M, J) is the observable list.

45.5.3 The no-hair theorems

Collapse is messy; its endpoint is not. The stationary states available to it are classified by uniqueness theorems of remarkable strength, and they are pure mathematics: this treatise imports them with hypotheses stated.

Theorem 45.39 (Israel; imported)**.** *A static, asymptotically flat vacuum spacetime whose event horizon is a connected, regular, non-degenerate closed 2-surface is isometric to the Schwarzschild solution Equation* (45.6)*; the horizon is then a round sphere. The electrovac counterpart, with Maxwell field, is Reissner–Nordström [1351, 1374].*

Rests on Theorem 45.1 and Proposition 45.37.

Theorem 45.40 (Carter–Robinson; imported)**.** *A stationary, axisymmetric, asymptotically flat vacuum spacetime, regular on and outside a connected non-degenerate event horizon, is a member of the two-parameter Kerr family Equation* (45.37)*, labelled by* (M, J) *with* $|\chi| < 1$*: Carter reduced the problem to two-parameter families [1443], Robinson closed it [1560]. With electromagnetism the endpoint is Kerr–Newman, labelled by* (M, J, Q) *[1697].*

Rests on Theorem 45.30.

Remark 45.41 (Hypotheses, and the honest perimeter)*.* Axisymmetry is not an extra assumption in the end: Hawking's rigidity theorem shows a stationary rotating black hole must be axisymmetric [1504] — though its original proof assumes analyticity of the metric, a technical gap still not closed in full generality. "Non-degenerate" excludes the extremal case $|\chi| = 1$, which needs and has separate treatment. And "vacuum or electrovac" is essential: with other matter — Yang–Mills fields, complex scalars with harmonic time dependence — stationary "hairy" solutions exist mathematically. None is known to be both stable and producible by collapse, and no observation requires one; within the matter content actually observed (Chapter 1's scope rule), the theorems say a stationary black hole is characterized *completely* by (M, J, Q), and astrophysically (Remark 45.38) by (M, J). Wheeler's slogan for this — "a black hole has no hair" — entered the literature through [1454]: every other memory of the progenitor (baryon number, magnetic moments, mountains, history) is radiated away in the ringdown.

Phenomenon 45.42 (Stationary black holes are two-parameter objects)**.** No stationary black hole has yet been observed to carry any property beyond a mass and an angular momentum. The sharpest test available is the ringdown of a merger remnant: for a Kerr hole the frequencies and damping times of the quasi-normal oscillations are fixed functions of M and J alone, so measuring two of them over-determines the remnant and any third measurement is a consistency check. The remnant of the first observed binary black-hole merger rings at the frequency and damping time predicted for a Kerr hole of the mass and spin inferred independently from the inspiral [2264], and the growing merger catalogue has produced no counterexample [2324].

Rests on Theorems 45.39 and 45.40.

Derivation. *Derives Phenomenon 45.42.*

The two-parameter character of the *spectrum* follows from the uniqueness theorems by dimensional analysis, with no wave computation at all. A perturbed remnant settles to Kerr (Theorem 45.40); its perturbations obey linear wave equations on the Kerr background with outgoing radiation at infinity and ingoing at the horizon, and those two boundary conditions make the frequency spectrum discrete and complex — damped resonances, the *quasi-normal modes* [1174, 1438]. The background carries exactly one dimensional scale, GM/c^3 (a time), and one dimensionless parameter, χ; therefore every mode must take the form

$$\omega_{n\ell m} = \frac{c^3}{GM}\, F_{n\ell m}(\chi) \ , \tag{45.45}$$

with $F_{n\ell m}$ *universal* dimensionless functions — computed once and for all, never fit per source. Their values are imported: for Schwarzschild the fundamental quadrupole mode is $\omega = (0.374 - 0.089\,\mathrm{i})\,c^3/(GM)$ [1438, 1717]. The test is then over-determination: one complex frequency (one f, one damping time) inverts Equation (45.45) for (M, χ); any second mode, or the (M, χ) inferred independently from the inspiral, must agree — and for GW150914, ringing at about 250 Hz with the mass and spin the inspiral gave, it does [2264]. A single confirmed extra parameter in the spectrum would falsify the uniqueness theorems' applicability to Nature. □

45.6 Black-hole mechanics

45.6.1 The four laws

Black holes obey mechanical laws with the exact structure of thermodynamics [1493]. The dictionary needs two quantities: the horizon area, and the surface gravity.

Proposition 45.43 (Static acceleration and surface gravity)**.** *A static observer at radius r in Equation (45.6) requires proper acceleration*

$$\mathrm{a}(r) = \frac{GM}{r^2}\left(1 - \frac{r_\mathrm{s}}{r}\right)^{-1/2}, \tag{45.46}$$

which diverges at the horizon. The redshifted acceleration — the force per unit mass an operator at infinity *must supply, holding the observer on an ideal massless tether — stays finite, and its horizon limit defines the* surface gravity*:*

$$\kappa := \lim_{r\to r_\mathrm{s}} \sqrt{1 - \frac{r_\mathrm{s}}{r}}\,\mathrm{a}(r) = \frac{GM}{r_\mathrm{s}^2} = \frac{c^4}{4GM}. \tag{45.47}$$

For Kerr the same construction on the horizon-generating Killing field $\partial_t + \Omega_\mathrm{H}\partial_\varphi$ gives

$$\kappa = \frac{c^2\,(r_+ - r_-)}{2\,(r_+^2 + a^2)}, \tag{45.48}$$

uniform over the horizon (the zeroth law); the Kerr evaluation is imported [1493, 1746].
Rests on Theorems 45.1 and 45.30.

Derivation. *Derives Proposition 45.43.*
The static four-velocity has $u^0 = c\,(1 - r_\mathrm{s}/r)^{-1/2}$ (normalization) and acceleration $a^\mu = u^\nu \nabla_\nu u^\mu$ with single nonvanishing component $a^r = \Gamma^r{}_{00}\,(u^0)^2$. From Equation (45.2) with $\mathrm{e}^{2\alpha} = 1 - r_\mathrm{s}/r$, $\Gamma^r{}_{00} = (r_\mathrm{s}/2r^2)\,(1 - r_\mathrm{s}/r)$, so $a^r = GM/r^2$ and the invariant magnitude is $\mathrm{a} = \sqrt{g_{rr}}\,a^r$, which is Equation (45.46). The tether argument for the redshift factor is the energy bookkeeping of Equation (45.8): work done at infinity is redshifted by $\sqrt{-g_{tt}}/c$ relative to work delivered locally. The limit is elementary. For the Schwarzschild check of Equation (45.48): $r_+ = r_\mathrm{s}$, $r_- = 0$, $a = 0$ gives $c^2/(2r_\mathrm{s}) = c^4/(4GM)$, agreeing with Equation (45.47). Numerically $\kappa = 1.5\times10^{13}\,\mathrm{m/s^2}$ for $M = M_\odot$, scaling as $1/M$ — enormous, yet the temperature it will buy in Section 45.6.2 is tens of nanokelvin. □

Lemma 45.44 (Horizon area)**.** *The area of the Kerr horizon is*

$$A = 4\pi\left(r_+^2 + a^2\right) = 8\pi\,\frac{GM}{c^2}\,r_+ \qquad \left(A = 4\pi r_\mathrm{s}^2 = \frac{16\pi G^2 M^2}{c^4}\ \text{at } a = 0\right). \tag{45.49}$$

Rests on Proposition 45.31.

Derivation. *Derives Lemma 45.44.*

On $r = r_+$, $t = \text{const}$ the induced metric from Equation (45.37) is $\Sigma_+ \mathrm{d}\theta^2 + \left((r_+^2 + a^2)^2 \sin^2\theta/\Sigma_+\right)\mathrm{d}\varphi^2$, writing $\Sigma_+ := \Sigma(r_+, \theta)$ and using the identity from the proof of Proposition 45.32 at $\Delta = 0$. Its area element is $(r_+^2 + a^2)\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi$ — the Σ_+ factors cancel — so $A = 4\pi\,(r_+^2 + a^2)$, and $r_+^2 + a^2 = r_{\mathrm{s}} r_+$ gives the second form. □

Proposition 45.45 (Irreducible mass and the Christodoulou–Smarr mass formula)**.** *Define the* irreducible mass *by* $A =: 16\pi G^2 M_{\mathrm{irr}}^2/c^4$. *Then*

$$M^2 = M_{\mathrm{irr}}^2 + \frac{c^2 J^2}{4G^2 M_{\mathrm{irr}}^2}\,, \tag{45.50}$$

so $M \geq M_{\mathrm{irr}}$, *with equality iff* $J = 0$ *[1417]. In particular, since the area theorem below forbids* M_{irr} *ever to decrease, at most*

$$\frac{M - M_{\mathrm{irr}}}{M} \;\leq\; 1 - \frac{1}{\sqrt{2}} \approx 0.29 \tag{45.51}$$

of a hole's mass–energy is extractable by any process — the bound saturated only from the extremal state, where $M_{\mathrm{irr}} = M/\sqrt{2}$.

Rests on Lemma 45.44 and Proposition 45.31.

Derivation. *Derives Proposition 45.45.*

By Lemma 45.44, $16\pi G^2 M_{\mathrm{irr}}^2/c^4 = 8\pi\,(GM/c^2)\,r_+$, i.e. $r_+ = 2GM_{\mathrm{irr}}^2/\,(Mc^2)$. Substitute this into $r_{\mathrm{s}} r_+ = r_+^2 + a^2$ with $a = J/(Mc)$:

$$\frac{2GM}{c^2}\cdot\frac{2GM_{\mathrm{irr}}^2}{Mc^2} = \frac{4G^2 M_{\mathrm{irr}}^4}{M^2c^4} + \frac{J^2}{M^2c^2}\,,$$

and multiplication by $M^2c^4/\,(4G^2)$ gives $M^2 M_{\mathrm{irr}}^2 = M_{\mathrm{irr}}^4 + c^2J^2/\,(4G^2)$, which is Equation (45.50). At extremality $J = GM^2/c$, and Equation (45.50) becomes $M^2 = M_{\mathrm{irr}}^2 + M^4/\,(4M_{\mathrm{irr}}^2)$, whose solution is $M_{\mathrm{irr}}^2 = M^2/2$; extracting all the angular momentum at fixed M_{irr} leaves a Schwarzschild hole of mass $M/\sqrt{2}$, which is Equation (45.51). □

Proposition 45.46 (The first law)**.** *Along the Kerr family, with* $E = Mc^2$,

$$\mathrm{d}E = \frac{\kappa c^2}{8\pi G}\,\mathrm{d}A + \Omega_{\mathrm{H}}\,\mathrm{d}J\,, \tag{45.52}$$

with κ *and* Ω_{H} *as in Equations (45.41) and (45.48). Moreover the homogeneity of the mass formula yields Smarr's integral form [1513],*

$$E = \frac{\kappa c^2}{4\pi G}\,A + 2\,\Omega_{\mathrm{H}} J\,. \tag{45.53}$$

Rests on Propositions 45.43 and 45.45.

Derivation. *Derives Proposition 45.46.*

Differentiate Equation (45.50) at fixed functional form, using $\mathrm{d}M_{\mathrm{irr}}^2 = c^4\mathrm{d}A/\,(16\pi G^2)$:

$$2M\,\mathrm{d}M = \left(1 - \frac{c^2J^2}{4G^2M_{\mathrm{irr}}^4}\right)\frac{c^4\,\mathrm{d}A}{16\pi G^2} + \frac{c^2 J}{2G^2 M_{\mathrm{irr}}^2}\,\mathrm{d}J\,.$$

Two substitutions, both via $M_{\mathrm{irr}}^2 = Mc^2 r_+/(2G)$ (from Lemma 45.44), turn the coefficients geometric. The full $\mathrm{d}J$ coefficient is

$$\frac{c^2}{2M}\cdot\frac{c^2 J}{2G^2 M_{\mathrm{irr}}^2} = \frac{c^2 J}{2GM^2 r_+} = \frac{c^3 a}{2GMr_+} = \frac{a\,c}{r_{\mathrm{s}} r_+} = \Omega_{\mathrm{H}}$$

by Equation (45.41), using $J = Mca$ and $r_{\mathrm{s}} = 2GM/c^2$. In the $\mathrm{d}A$ coefficient, $c^2J^2/\left(4G^2M_{\mathrm{irr}}^4\right) = a^2/r_+^2$ by the same substitution, and $r_+r_- = a^2$ (product of the roots of Δ) turns the bracket into $1 - r_-/r_+ = (r_+ - r_-)/r_+$; assembling,

$$\left(\frac{\partial E}{\partial A}\right)_J = \frac{c^6\,(r_+ - r_-)}{32\pi G^2 M\, r_+} = \frac{c^2}{8\pi G}\cdot\frac{c^2\,(r_+ - r_-)}{2\,(r_+^2 + a^2)} = \frac{\kappa c^2}{8\pi G}\,,$$

the middle equality by $r_+^2 + a^2 = r_{\mathrm{s}}r_+ = 2GMr_+/c^2$, matching Equation (45.48). This is Equation (45.52). For Equation (45.53): under $(A, J) \to (\lambda^2 A, \lambda^2 J)$ the mass formula gives $E \to \lambda E$; differentiating $E(\lambda^2 A, \lambda^2 J) = \lambda E(A, J)$ at $\lambda = 1$ gives $2A\,\partial E/\partial A + 2J\,\partial E/\partial J = E$, which with the coefficients just computed is Equation (45.53). The Schwarzschild case checks in one line: with Equations (45.47) and (45.49), $\mathrm{d}A = 32\pi G^2 M\,\mathrm{d}M/c^4$ and $(\kappa c^2/8\pi G)\,\mathrm{d}A = c^2\,\mathrm{d}M$ exactly. □

Theorem 45.47 (Area theorem; imported)**.** *In an asymptotically predictable spacetime (cosmic censorship, Section 45.4.3, holds) whose matter satisfies the null energy condition (Section 44.8), the total area of event horizons is non-decreasing toward the future [1447].*

Rests on Proposition 45.18.

Remark 45.48 (The four laws assembled). The parallel with thermodynamics (Chapter 54) is now structural, law by law [1493]:

Law	Black-hole mechanics	Thermodynamics
Zeroth	κ uniform over the horizon of a stationary hole	T uniform at equilibrium
First	$\mathrm{d}E = \dfrac{\kappa c^2}{8\pi G}\,\mathrm{d}A + \Omega_{\mathrm{H}}\,\mathrm{d}J$	$\mathrm{d}U = T\,\mathrm{d}S - p\,\mathrm{d}V$
Second	$\mathrm{d}A \geq 0$ (Theorem 45.47)	$\mathrm{d}S \geq 0$
Third	$\kappa = 0$ (extremality) unattainable in any finite process	$T = 0$ unattainable

The correspondence pairs E with internal energy, κ with temperature and A with entropy, up to constants; Bekenstein argued on information-theoretic grounds that the pairing is physical and the entropy real [1494]. Classically that seems absurd — a classical black hole absorbs everything and emits nothing, so its temperature ought to be exactly zero; the area theorem's own hypotheses (null energy condition) are what quantum fields refuse to satisfy, and quantum theory is exactly what turns the analogy literal (Section 45.6.2). One consequence of Theorem 45.47 is already observable bookkeeping: two merging Schwarzschild holes of mass M each have final mass at least $\sqrt{2}M$ (areas add, $A \propto M^2$), so at most $1 - 1/\sqrt{2} \approx 29\%$ of the total can be radiated — GW150914's measured $3.0\,M_\odot$ out of $65\,M_\odot$, under 5%, sits comfortably inside the bound [2264].

45.6.2 Hawking radiation: prediction without observation

Theorem 45.49 (Hawking temperature; imported)**.** *Quantum fields on the spacetime of a body collapsing to a black hole of surface gravity κ settle, at late times, into a state in which the hole emits thermal radiation to infinity at the temperature*

$$T_{\mathrm{H}} = \frac{\hbar\,\kappa}{2\pi c\,k_{\mathrm{B}}} = \frac{\hbar c^3}{8\pi G M k_{\mathrm{B}}} \qquad (\textit{second form: Schwarzschild})\,, \tag{45.54}$$

and the four laws of Section 45.6.1 become literal thermodynamics with entropy

$$S_{\mathrm{BH}} = \frac{k_{\mathrm{B}}c^3}{4G\hbar}\,A \qquad \left(\frac{S_{\mathrm{BH}}}{k_{\mathrm{B}}} \approx 1.0\times10^{77}\ \textit{for}\ M_\odot\right). \tag{45.55}$$

[1549].
Rests on Propositions 45.43 and 45.46.

Remark 45.50 (Where the prediction comes from, and how firmly). The derivation is imported [1549, 1890]: it is a computation in quantum field theory on the fixed collapse background, comparing the field's in-vacuum with the late-time out-modes. Its structural core, however, is already in this treatise: Proposition 109.74 establishes — as a theorem of axiomatic field theory, via the KMS condition — that the vacuum restricted to a wedge is thermal with respect to the boost flow, at the temperature $T = \hbar a/(2\pi c k_{\mathrm{B}})$ for the observer of proper acceleration a [1590]. Near the horizon a static observer is precisely such an accelerated observer, and Equation (45.54) is the wedge temperature evaluated at the surface gravity and redshifted to infinity (Equation (45.47)); the Euclidean route — regularity of the imaginary-time geometry forcing periodicity $\hbar/(k_{\mathrm{B}}T_{\mathrm{H}})$ — gives the same number [1596]. Consistency of three independent derivations is why the prediction is regarded as robust *given* its premise: quantum field theory remains valid on a classical black-hole background. That premise is untested, and everything that hangs on the answer — the information problem, the microscopic meaning of Equation (45.55) — is quarantined with the quantum-gravity programme in Chapter 135.

Proposition 45.51 (Evaporation: scale and timescale)**.** *A hole radiating thermally at Equation (45.54) from area Equation (45.49) loses mass as* $\mathrm{d}M/\mathrm{d}t \propto -1/M^2$ *and evaporates in*

$$t_{\mathrm{evap}} \simeq \frac{5120\,\pi\,G^2M^3}{\hbar c^4} \approx 2\times 10^{67}\,\mathrm{yr}\left(\frac{M}{M_\odot}\right)^3, \tag{45.56}$$

where the numerical prefactor is the photon-only, geometric-optics Stefan–Boltzmann estimate; emitted-species and greybody corrections change it by factors of order unity [1549].
Rests on Theorem 45.49.

Derivation. *Derives Proposition 45.51.*
Treat the hole as a black body of area $A = 16\pi G^2M^2/c^4$ at temperature Equation (45.54) and apply the Stefan–Boltzmann law (Chapter 54), $P = \sigma A T_{\mathrm{H}}^4$ with $\sigma = \pi^2 k_{\mathrm{B}}^4/(60\hbar^3c^2)$:

$$c^2\frac{\mathrm{d}M}{\mathrm{d}t} = -\sigma A T_{\mathrm{H}}^4 = -\frac{\hbar c^6}{15360\,\pi\,G^2M^2} \qquad\Longrightarrow\qquad t_{\mathrm{evap}} = \frac{5120\,\pi\,G^2M^3}{\hbar c^4},$$

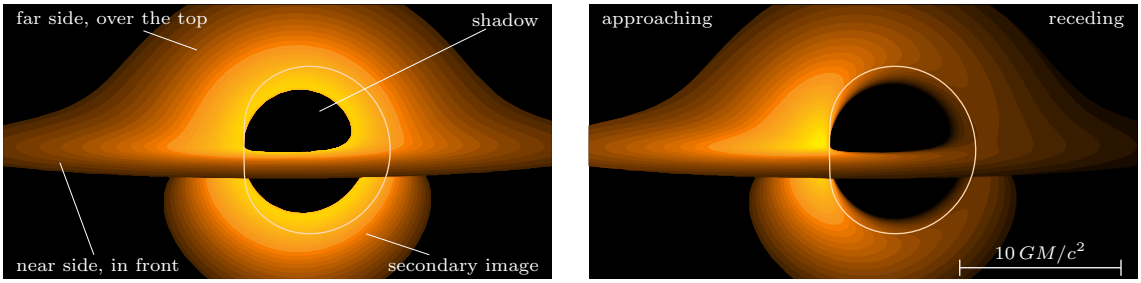
by integrating $M^2\mathrm{d}M$. Inserting $M_\odot$ gives $6.6\times10^{74}\,\mathrm{s}$, i.e. about $2\times10^{67}\,\mathrm{yr}$ — 10^{57} Hubble times. The estimate ignores the frequency-dependent horizon absorptivity and the species count, which is why only its scaling and order of magnitude are asserted. □

The honest evidence status must be stated plainly, in the terms of Chapter 1: *Hawking radiation has never been observed, and for every black hole known to exist it never can be by any foreseeable instrument.* The arithmetic is unforgiving: Equation (45.54) gives $T_{\mathrm{H}} = 6.2\times10^{-8}\,\mathrm{K}$ for one solar mass — colder than the 2.725 K microwave background by nearly eight orders of magnitude, so every astrophysical hole *absorbs* background radiation faster than it emits, and will until cosmic expansion has cooled the background below its temperature. A hole hotter than the background must be lighter than $4.5\times10^{22}\,\mathrm{kg}$ (half a lunar mass, horizon smaller than a tenth of a millimetre); one completing its evaporation today — the potentially detectable terminal γ-ray burst — must have been born primordial at about 10^{11}–$10^{12}\,\mathrm{kg}$ [1549], and searches for such bursts have found nothing attributable. The status is therefore exactly that recorded for the acceleration temperature in Remark 109.75, and for the same structural reason: the KMS mathematics is theorem-solid, the physical consequence unobserved. What *would* count as evidence: the terminal burst of a primordial hole with the predicted spectrum, or — more remotely — thermal spectra from any horizon at the predicted temperature. Analogue systems (sonic horizons in condensates) confirm

the mode-conversion mathematics, not the gravitational claim [1675]. No phenomenon environment appears in this section: the book reserves that environment for observed facts, and this is a prediction — the sharpest untested prediction general relativity and quantum theory jointly make.

45.7 The shadow

Seen from outside, a black hole's mark is what it does to light. Rays with impact parameter below a critical value b_c are captured (Section 45.2.3); an observer looking at a hole backlit or surrounded by emission therefore sees a dark disc of radius b_c — the capture cross-section, not the horizon — bounded by a bright ring where rays that nearly orbited the photon sphere pile up. The escape-cone geometry was computed by Synge [1338], and the first realistic image of a hole with a thin accretion disc — ring, asymmetry and all — by Luminet [1642]. The same construction is carried out in Figure 45.1 for a rapidly spinning hole viewed almost edge-on — the configuration rendered for the film *Interstellar*, whose images leave out the Doppler and gravitational frequency shifts deliberately, so that the disc appears left–right symmetric [2255]; the second panel restores them. The angular size is set by $b_c \approx 2.6\, r_s$: about ten gravitational radii across, only weakly dependent on spin and inclination [1988], which is what makes the measurement a test of the geometry rather than of astrophysical modelling. The imaging observations themselves — and the stellar-orbit mass that calibrates them — are deferred, with the full apparatus, to Chapter 50.



(a) frequency shifts omitted, as filmed

(b) brightness carrying the factor g^4

Figure 45.1: A thin accretion disc around a rapidly spinning black hole, computed by integrating the Kerr null geodesic equations backwards from the observer. The hole has $a = 0.999\,GM/c^2$ and is seen 5° out of the equatorial plane; the disc is opaque, geometrically thin, and reaches from the innermost stable circular orbit at $1.18\,GM/c^2$ out to $20\,GM/c^2$, radiating with the time-averaged flux of a steady thin disc [1531]. Lensing does all the work. Light from the far side is bent over and under the hole, so the far side of the disc appears as an arc *above* the dark region and its underside as a second arc below, while the near side crosses in front as a thin band: a flat disc is wrapped into the shape one sees. The dark region is the capture cross-section, flattened on the left — the side on which co-rotating photons appear — by the dragging of inertial frames, and the thin ring is the analytic boundary of that cross-section, which the disc's own images partly cover and on which the higher-order images pile up too closely to be drawn. Panel (a) omits the Doppler and gravitational frequency shifts, as the film *Interstellar* did [2255]; panel (b) restores them, multiplying the surface brightness by g^4, where g is the redshift factor of a prograde circular orbit. The approaching side is on the left, and outshines the receding one by orders of magnitude. Brightness is logarithmic over four decades and both panels share one normalisation, so they differ by exactly the factor g^4. The first image of this kind was computed by Luminet [1642].

Phenomenon 45.52 (The shadow)**.** A black hole surrounded by an optically thin emitting region appears as a bright ring enclosing a central brightness depression. The depression is bounded by the photon-capture impact parameter, so a Schwarzschild hole of mass M seen from distance D subtends an angular diameter

$$\theta_{\mathrm{sh}} = \frac{2b_{\mathrm{c}}}{D}\,, \qquad b_{\mathrm{c}} = \frac{3\sqrt{3}\,GM}{c^2} = \frac{3\sqrt{3}}{2}\,r_{\mathrm{s}}\,, \tag{45.57}$$

that is $\theta_{\mathrm{sh}} D = 3\sqrt{3}\,r_{\mathrm{s}} \simeq 5.20\,r_{\mathrm{s}}$, some ten gravitational radii and only weakly dependent on spin and on viewing angle [1338, 1988]. Rings of about this size have been imaged at millimetre wavelengths around the nucleus of M87 and around Sgr A* [2328, 2403]; the observations, with their apparatus, are assembled in Chapter 50.

Rests on Equation (45.20)*, Theorem 45.1, and Proposition 13.141.*

Derivation. *Derives Phenomenon 45.52.*

Take the first integral Equation (45.20) of the null geodesic equation, established in Section 45.2.3, which for a ray confined to the equatorial plane reads

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\lambda}\right)^2 = E^2 - \frac{L^2}{r^2}\left(1 - \frac{r_{\mathrm{s}}}{r}\right)\,,$$

with E and L the constants conjugate to the two Killing vectors and $b := L/E$ the impact parameter measured at infinity (Equation (45.22)) [1509]. Radial turning points are the zeros of the right-hand side, so a ray of impact parameter b turns where

$$\frac{1}{b^2} = W(r) := \frac{1}{r^2}\left(1 - \frac{r_{\mathrm{s}}}{r}\right)\,. \tag{45.58}$$

An incoming ray escapes if it meets a turning point outside the horizon and is captured if it does not, so the critical impact parameter separating the two behaviours is fixed by the maximum of W. Differentiating,

$$W'(r) = -\frac{2}{r^3} + \frac{3r_{\mathrm{s}}}{r^4}\,,$$

which vanishes at

$$r_{\mathrm{ph}} = \frac{3}{2}r_{\mathrm{s}} = \frac{3GM}{c^2}\,, \tag{45.59}$$

the photon sphere, and $W'' < 0$ there, so this is the maximum. Its value is

$$W\left(r_{\mathrm{ph}}\right) = \frac{1}{\left(3r_{\mathrm{s}}/2\right)^2}\left(1 - \frac{2}{3}\right) = \frac{4}{9r_{\mathrm{s}}^2}\cdot\frac{1}{3} = \frac{4}{27\,r_{\mathrm{s}}^2}\,,$$

so that $b_{\mathrm{c}}^2 = 27r_{\mathrm{s}}^2/4$ and $b_{\mathrm{c}} = \frac{3}{2}\sqrt{3}\,r_{\mathrm{s}} = 3\sqrt{3}\,GM/c^2$, which is Equation (45.57). The apparent diameter of the capture cross-section is twice this, $2b_{\mathrm{c}} = 3\sqrt{3}\,r_{\mathrm{s}} = 5.196\,r_{\mathrm{s}}$.

Two remarks. The shadow is larger than the horizon by the factor $3\sqrt{3}/2 \approx 2.60$, because the impact parameter is measured at infinity while the horizon radius is a coordinate radius: what is imaged is the capture cross-section, not the horizon. And rays with b slightly above b_{c} orbit many times near r_{ph} before escaping, which piles emission up at the rim and produces the bright ring rather than a mere edge. □

Chapter 46

Gravitational-Wave Theory

The Einstein field equations are nonlinear, but far from strong sources spacetime is very nearly flat, and the equations linearize into a wave equation. The resulting radiative solutions — gravitational waves — are the subject of this short chapter. We state the results and their physical content; the complete derivation, from the linearized field equations through the quadrupole formula, is carried out in Appendix A.1. The experimental record, indirect and direct, occupies Chapter 47.

46.1 Linearized gravity and the wave equation

Consider a metric that deviates only slightly from the Minkowski metric,

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}\,, \qquad |h_{\mu\nu}| \ll 1\,. \tag{46.1}$$

Keeping only terms of first order in $h_{\mu\nu}$, and introducing the trace-reversed perturbation $\bar{h}_{\mu\nu} = h_{\mu\nu} - \tfrac{1}{2}\eta_{\mu\nu}\,h$ with $h = \eta^{\alpha\beta}h_{\alpha\beta}$, one may always choose coordinates satisfying the Lorenz gauge condition $\partial^{\mu}\bar{h}_{\mu\nu} = 0$. In this gauge the linearized Einstein equations collapse to an inhomogeneous wave equation [436, 1509, 1746],

$$\Box\,\bar{h}_{\mu\nu} = -\frac{16\pi G}{c^{4}}\,T_{\mu\nu}\,, \tag{46.2}$$

where $\Box = \eta^{\alpha\beta}\partial_{\alpha}\partial_{\beta}$ is the flat-space d'Alembertian. Every statement in this equation is physical: disturbances of the metric propagate at the speed of light, they are sourced by the stress–energy tensor, and the coupling $G/c^{4} \approx 8.3\times10^{-45}\,\mathrm{s^{2}/(kg\,m)}$ is so small that only astrophysical masses moving at relativistic speeds produce perturbations of detectable size. We do not derive Equation (46.2) here; the derivation is the opening step of Appendix A.1.

46.2 Transverse-traceless gauge and the two polarizations

In vacuum ($T_{\mu\nu} = 0$) the residual gauge freedom within the Lorenz class can be used to set $h_{0\mu} = 0$, $h^{i}{}_{i} = 0$ and $\partial^{i}h_{ij} = 0$: the *transverse-traceless* (TT) gauge. For a plane wave travelling along the z-axis only two independent components survive, the polarization amplitudes h_{+} and $h_{\times}$ [1509, 2132]:

$$\mathrm{d}s^{2} = -c^{2}\mathrm{d}t^{2} + (1+h_{+})\mathrm{d}x^{2} + (1-h_{+})\mathrm{d}y^{2} + 2h_{\times}\,\mathrm{d}x\,\mathrm{d}y + \mathrm{d}z^{2}\,, \tag{46.3}$$

with $h_{+,\times} = h_{+,\times}(t - z/c)$. A gravitational wave therefore carries exactly two physical polarizations, as befits a massless spin-2 field.

The effect on matter is tidal. A circular ring of free test masses in the xy-plane is deformed by the $+$ polarization into an ellipse stretched along x while squeezed along y,

and half a period later the reverse; the $\times$ polarization produces the same pattern rotated by 45°. For two free masses separated by a distance L along a principal axis of the wave, the fractional change in proper separation is

$$\frac{\Delta L}{L} = \frac{h}{2}\,, \tag{46.4}$$

where h is the relevant polarization amplitude. This is precisely the quantity a laser interferometer measures: a differential arm-length oscillation at the wave frequency, of fractional size $h/2$ per arm.

46.3 The quadrupole formula

Phenomenon 46.1. Accelerated masses whose quadrupole moment varies in time radiate gravitational waves. The waves propagate at the speed of light c and carry energy away from the source; at leading order the radiated strain and luminosity are given by the quadrupole formula. This radiation has been observed indirectly, through the orbital decay of binary pulsars [1703], and directly, by the LIGO interferometers in 2015 [2264], exactly as predicted by the linearized theory [436, 453].

Rests on Postulate 44.3.

Full derivation: Appendix A.1 (p. 2759).

Derives Phenomenon 46.1.

We state the results proved there. Let $Q_{ij} = \int \rho\,(x_i x_j - \tfrac{1}{3}\delta_{ij} r^2)\mathrm{d}^3x$ denote the traceless mass quadrupole moment of the source. A distant observer at distance d measures the strain

$$\bar{h}_{ij}(t) = \frac{2G}{c^4 d}\,\ddot{Q}_{ij}\!\left(t - \frac{d}{c}\right)\,, \tag{46.5}$$

and the total power radiated by the source, averaged over several wave periods, is

$$P = \frac{G}{5c^5}\left\langle \dddot{Q}_{ij}\,\dddot{Q}^{ij}\right\rangle\,. \tag{46.6}$$

The prefactor $G/c^5 \approx 2.8\times10^{-53}\,1/\mathrm{W}$ explains at once why gravitational radiation is negligible in the laboratory and why only compact binaries reach detectable luminosities [2132].

For a binary of masses m_1 and m_2 on a circular orbit, the gravitational-wave frequency is twice the orbital frequency, and the entire leading-order evolution is controlled by a single mass combination, the *chirp mass*

$$\mathcal{M} = \frac{(m_1 m_2)^{3/5}}{(m_1+m_2)^{1/5}}\,. \tag{46.7}$$

Energy loss through Equation (46.6) shrinks the orbit and drives the frequency upward according to

$$\frac{\mathrm{d}f_{\mathrm{gw}}}{\mathrm{d}t} = \frac{96}{5}\,\pi^{8/3}\left(\frac{G\mathcal{M}}{c^3}\right)^{5/3} f_{\mathrm{gw}}^{11/3}\,, \tag{46.8}$$

so that a measurement of the frequency and its rate of change during the inspiral yields $\mathcal{M}$ directly, with no astrophysical modelling. This accelerating sweep — the *chirp* — is the unmistakable signature by which compact binaries announce themselves in the data of Chapter 47 [2132, 2264].

46.4 Beyond linear order

The linearized treatment above suffices for the propagation of the waves and for the leading-order emission from slowly moving sources. It does not by itself define the energy carried by the waves — a second-order quantity in $h_{\mu\nu}$ — nor does it describe the late inspiral and merger of black holes, where the full nonlinearity of the field equations takes over and only post-Newtonian expansions and numerical relativity give quantitative waveforms [1509, 1746, 2132].

The first of these gaps is closed in this book: the second-order (Isaacson) analysis — expanding the Ricci tensor to $O(h^2)$, averaging over several wavelengths, and reading the averaged quadratic terms as an effective stress–energy tensor of the radiation — is carried out in Appendix A.1.5, culminating in Equation (A.46); it is that tensor whose flux, integrated over a sphere, yields the luminosity Equation (46.6) quoted above. The second gap is imported knowledge, and honestly so: the post-Newtonian expansion of the two-body problem has been driven to the orders required by detector phase templates [2231], and the merger itself yielded only to numerical solution of the full field equations [2082]. This book states those results where the experiments need them (Chapter 47) and does not rederive them; each is a controlled approximation to, or a numerical solution of, the same field equations derived in Chapter 44, so nothing in the evidential chain rests on machinery outside the theory already stated.

The history of the subject mirrors its subtlety. Einstein's founding 1916 paper on gravitational waves [436] contained a celebrated error of a numerical factor in the radiation formula, which he himself corrected in the 1918 paper that first stated the quadrupole formula in its modern form [453]; decades of confusion about whether the waves were physical or mere coordinate ripples followed, and were settled only in the late 1950s with gauge-invariant treatments of the energy flux [1509, 2132]. That the formula so laboriously secured now agrees with binary-pulsar timing to a fraction of a percent, and with direct interferometry, is the subject of the next chapter: Chapter 47.

Chapter 47

Experiment: Gravitational Waves

The prediction of Chapter 46 — that accelerated masses radiate energy in ripples of spacetime — was confirmed twice, by two utterly different techniques separated by four decades. First indirectly: a binary pulsar, timed for years with radio telescopes, was observed to lose orbital energy at exactly the rate demanded by the quadrupole formula of Appendix A.1. Then directly: kilometre-scale laser interferometers recorded the strain of a passing wave itself. This chapter presents both records in full phenomenological detail; no derivations are performed here.

47.1 The Hulse–Taylor Binary Pulsar (PSR B1913+16) (1974–2016)

Tests Phenomenon 46.1 and Equation (46.6).

A pulsar is a rotating neutron star whose radio beam sweeps past the Earth once per rotation, providing a clock of extraordinary stability. A pulsar bound in a tight orbit around a second neutron star is therefore a moving clock in a strong gravitational field — a ready-made laboratory for relativistic orbital dynamics, including the slow decay of the orbit by gravitational radiation.

47.1.1 Apparatus

The 305 m radio telescope at Arecibo, Puerto Rico, equipped for precision pulsar timing: dispersion-corrected radio receivers and time-tagging of pulse arrival times against maser clocks, later linked to international atomic time. The pulsar PSR B1913+16 was found in a systematic survey for new pulsars conducted by Hulse and Taylor in the summer of 1974 [1550].

47.1.2 Procedure

The pulsar's spin period of about 59 ms was observed to shift periodically by up to 80 µs, revealing orbital Doppler motion: the pulsar orbits an unseen companion, with orbital period 7.75 h and orbital eccentricity $e \approx 0.617$ [1550]. Over the following four decades, pulse arrival times were accumulated and fitted to a fully relativistic timing model, whose post-Keplerian parameters — periastron advance, combined time dilation and gravitational redshift, and the secular decay of the orbital period — are measured without assuming general relativity and only then compared with its predictions [1703, 2275].

47.1.3 Observations and data

The key fitted parameters are collected below [1550, 2275].

Quantity	Value
Pulsar spin period	59 ms
Orbital period P_b	7.75 h
Orbital eccentricity e	0.617
Periastron advance $\langle\dot\omega\rangle$	4.226° per year
Orbital period decay $\dot P_b$	$\approx -2.4\times10^{-12}$ s/s
Observed/predicted decay ratio	0.9983 ± 0.0016

The periastron of the orbit advances by 4.226° per year — in a single day, roughly what Mercury's perihelion accumulates in a century. The orbital period shrinks by about -2.4×10^{-12} s/s, i.e. some 76.5 µs per year. The most vivid presentation of the timing data is the cumulative shift of periastron passage times relative to a constant-period orbit: the data points trace a downward-opening parabola, reaching a cumulative shift of about −40 s after roughly thirty years of observation, with the general-relativistic prediction drawn through them as a solid curve from which no point visibly departs [2275]. After correcting for the small relative acceleration of the solar system and the pulsar in the Galaxy, the ratio of the observed period decay to that predicted by the quadrupole formula is 0.9983 ± 0.0016 [2275].

47.1.4 Interpretation

The orbit decays because the system radiates energy in gravitational waves, at precisely the rate given by the quadrupole formula (Appendix A.1): agreement at the level of 0.16%, with no adjustable parameters. This was the first evidence — indirect but quantitative — for the existence of gravitational radiation, and it earned Hulse and Taylor the 1993 Nobel Prize in Physics. The steadily improving timing residuals of [2275] remain among the most stringent tests of the radiative sector of general relativity.

Every post-Keplerian parameter entering these fits is now derived in this book. The orbital decay $\dot P_b$ — the parameter carrying the gravitational-wave result quoted above — follows from the quadrupole formula proved in Appendix A.1, including the eccentricity enhancement factor $f(e)$ evaluated there at the measured $e = 0.617$. The remaining three rest on the classical-test machinery of Chapter 53 and are short enough to derive here.

Derivation of the remaining post-Keplerian parameters. (i) *Periastron advance.* Equation (53.1) gives the relativistic advance per orbit, $\Delta\varphi = 6\pi GM/[c^2 a(1-e^2)]$; for a binary the relative orbit obeys the same equation with $M = m_1 + m_2$ at leading order. The secular rate is $\dot\omega = \Delta\varphi/P_b$, and eliminating the semi-major axis through Kepler's third law $(2\pi/P_b)^2 a^3 = GM$ gives

$$\langle\dot\omega\rangle = 3\left(\frac{2\pi}{P_b}\right)^{5/3}\left(\frac{G(m_1+m_2)}{c^3}\right)^{2/3}\frac{1}{1-e^2}\,. \tag{47.1}$$

Evaluated at the measured P_b, e and the fitted masses $m_1 = 1.4398\,M_\odot$, $m_2 = 1.3886\,M_\odot$ [2275], this gives 4.23° per year (computed here), against the fitted 4.2266°.

(ii) *Einstein delay.* The pulsar's proper time runs against coordinate time as $\mathrm{d}\tau/\mathrm{d}t = 1 - v_1^2/2c^2 - Gm_2/(c^2 r)$, the two corrections being special-relativistic time dilation and the companion's redshift (Chapter 42). With $v_1 = (m_2/M)\,v$ on the relative orbit and the vis-viva relation $v^2 = GM(2/r - 1/a)$, the $1/r$ terms combine with coefficient $Gm_2(m_1+2m_2)/(Mc^2)$. On the Keplerian ellipse $r = a(1-e\cos E)$ with $nt = E - e\sin E$ one has $\int \mathrm{d}t/r = E/(na)$, whose periodic part is $(e/na)\sin E$ — the constant rates are unobservable, absorbed into the fitted spin period. The arrival times are therefore modulated by $\Delta\tau = -\gamma_{\mathrm E}\sin E$ with

$$\gamma_{\mathrm E} = e\left(\frac{P_b}{2\pi}\right)^{1/3}\left(\frac{G}{c^3}\right)^{2/3}\frac{m_2(m_1+2m_2)}{(m_1+m_2)^{4/3}}\,, \tag{47.2}$$

which at the fitted masses is $\approx 4.3\,\mathrm{ms}$ (computed here), as measured.

(iii) *Shapiro range and shape.* Equation (53.5) applied with the companion as the deflector adds a propagation delay $2(Gm_2/c^3)\ln[\,\cdot\,]$ to pulses crossing the companion's potential; near superior conjunction the logarithm's argument is fixed by the orbital geometry through $\sin i$ alone, so the timing model parametrizes the effect by the *range* $r_{\mathrm{S}} = Gm_2/c^3$ and the *shape* $s = \sin i$. In PSR B1913+16 the inclination makes the effect only marginally visible; it is measured superbly in the double pulsar (Chapter 49). □

Together with $\dot{P}_b$ from Appendix A.1, the timing model is thus derived end to end: three parameters (Equations (47.1) and (47.2), and r_{S}, s) overdetermine the two masses, and the gravitational-wave test is that the fourth, $\dot{P}_b$, then agrees with no freedom left.

47.1.5 Primary references

[1550, 1703, 2275].

47.2 Direct detection: the instrument

The Laser Interferometer Gravitational-Wave Observatory (LIGO) consists of two detectors of identical design, one at Hanford, Washington, and one at Livingston, Louisiana, separated by about 3000 km — a light travel time of about 10 ms that both provides a coincidence veto against local disturbances and encodes the sky direction of a source in the inter-site arrival-time difference [2243].

Each detector is a Michelson interferometer whose two 4 km arms are Fabry–Perot resonant cavities, multiplying the phase accumulated by the laser light in each arm by the cavity finesse. A power-recycling mirror at the input builds up the circulating laser power, and a signal-recycling mirror at the output shapes the instrument's frequency response [2243]. A passing gravitational wave of strain h changes the differential arm length by $\Delta L = hL/2$ (Chapter 46); at the sensitivity reached in 2015 this corresponds to measurable differential displacements of order $10^{-18}\,\mathrm{m}$ — about one thousandth of the diameter of a proton.

The noise budget divides the band into three regimes [2243]. Below roughly 10 Hz rises the seismic wall: ground motion, suppressed by active isolation platforms and by suspending each 40 kg mirror from a four-stage pendulum, still dominates and sets the low-frequency edge of the observation band. In the middle of the band, from tens to a couple of hundred hertz, thermally driven motion of the fused-silica suspension fibres and of the mirror coatings is the limiting noise. At high frequencies the limit is quantum: photon shot noise in the readout, improved by raising the circulating power. The quietest region of the resulting sensitivity curve, around 100 Hz, is precisely where a stellar-mass compact binary sweeps through its final orbits.

47.3 GW150914 (2015)

Tests Phenomenon 46.1 and Equation (46.8).

47.3.1 Apparatus

The two Advanced LIGO interferometers at Hanford and Livingston, described in Section 47.2 [2243], in their first observing run and operating in coincidence [2264].

47.3.2 Procedure

Two independent analyses ran over the coincident data [2264]. A matched-filter search correlated the strain data against a bank of about 250 000 relativistic waveform templates for compact-binary coalescences; a separate, unmodelled burst search looked for coherent excess power in both detectors without waveform assumptions. Both searches recovered the same event, and the significance was established by measuring the background of accidental coincidences with time-shifted data.

47.3.3 Observations and data

On 2015 September 14 at 09:50:45 UTC both detectors recorded the same transient [2264]:

Quantity	Value
Peak strain amplitude	1.0×10^{-21}
Frequency sweep	35–250 Hz in about 0.2 s
Cycles observed	~ 8
Hanford–Livingston delay	$6.9^{+0.5}_{-0.4}$ ms
Combined signal-to-noise ratio	24
False-alarm rate	< 1 per 203 000 yr ($> 5.1\sigma$)
Primary mass m_1	$36^{+5}_{-4}\,M_\odot$
Secondary mass m_2	$29^{+4}_{-4}\,M_\odot$
Final black-hole mass	$62^{+4}_{-4}\,M_\odot$
Radiated energy	$3.0^{+0.5}_{-0.5}\,M_\odot c^2$
Peak luminosity	$3.6^{+0.5}_{-0.4} \times 10^{49}$ W
Luminosity distance	410^{+160}_{-180} Mpc ($z \approx 0.09$)

The signal swept upward in frequency from 35 Hz to 250 Hz over about 0.2 s — roughly eight cycles — increasing in amplitude to a peak strain of 1.0×10^{-21}, then rang down. The Livingston detector recorded it first; Hanford followed $6.9^{+0.5}_{-0.4}$ ms later, comfortably within the 10 ms light travel time between the sites. The waveforms observed in the two detectors, once shifted by this delay and inverted for the relative orientation of the instruments, lie on top of one another [2264].

47.3.4 Interpretation

The chirp-mass relation of Chapter 46 applied to the observed frequency and frequency sweep gives $\mathcal{M} \approx 30\,M_\odot$, and hence a total mass so large, in so small an orbit at merger, that only a pair of black holes fits: at 75 Hz orbital frequency the objects, of 36^{+5}_{-4} and 29^{+4}_{-4} solar masses, were only a few hundred kilometres apart. GW150914 is thus simultaneously the first direct detection of gravitational waves and the first observation of a binary black-hole merger. The full signal — inspiral, merger, and the damped ringdown of the final $62^{+4}_{-4}\,M_\odot$ black hole — matches the inspiral–merger–ringdown waveform computed from general relativity, with about $3.0 \pm 0.5\,M_\odot c^2$ radiated in a fraction of a second at a peak luminosity of $3.6^{+0.5}_{-0.4} \times 10^{49}$ W, briefly exceeding the combined light output of the observable Universe [2264]. The detection was recognized with the 2017 Nobel Prize in Physics.

47.3.5 Primary references

[2264], the discovery; [2243], the detector.

47.4 The multimessenger outlook

Two years after the discovery, on 2017 August 17, the LIGO–Virgo network recorded GW170817, an inspiral lasting about 100 s with component masses of 1–2 $M_\odot$: a binary neutron-star merger [2278]. A short gamma-ray burst arrived 1.7 s after the merger, and the optical counterpart — a kilonova in the galaxy NGC 4993 at about 40 Mpc — was followed across the electromagnetic spectrum for weeks; the near-simultaneity of the gravitational and gamma-ray signals after a journey of more than a hundred million years constrains the speed of gravitational waves to equal c to within a few parts in 10^{15} [2278]. Since then the catalogue of compact binary coalescences has grown from a single event to on the order of a hundred, with black-hole masses ranging from a few to over a hundred solar masses.

Gravitational-wave astronomy has thus passed, within a decade, from heroic first detection to routine observation: a new spectrum, opened on a sector of the Universe — coalescing black holes — that emits no light at all. Its implications for the largest scales, where the same field equations govern the expansion of the Universe itself, are taken up in Chapter 48.

Chapter 48

Evidence-Based Cosmology

Cosmology is the application of the field equations of Chapter 44 to the universe as a whole, and it is admitted here on the strength of four independent data streams: the expansion recorded in galaxy redshifts [627], the light-element abundances fixed in the first minutes [905], the relic microwave background predicted before it was found [906, 1320], and the growth of large-scale structure with its baryon acoustic ruler [2076]. Each is treated with its primary sources; the CMB measurements get their own experiment chapter, Chapter 51.

The chapter's standing claim is that ΛCDM is a *parametrization* of these data, not yet an explanation of them: six numbers fit everything [2348], while the physical nature of two of the ingredients — the dark matter of Chapter 134 and the constant Λ — and the mechanism behind the initial conditions remain unidentified. The chapter states throughout what is measured, what is inferred through the model, and what is not measured at all.

48.1 The expanding universe

48.1.1 Nebular redshifts

The observational record begins not with a theory but with a decade of patient spectroscopy. Between 1912 and 1917, working with the 24-inch refractor of the Lowell Observatory, Slipher measured the Doppler displacements of the spectral lines of the brightest spiral nebulae — objects whose nature, star cloud or island universe, was itself still in dispute [451]. Two facts emerged from his radial velocities and survived every later refinement. First, the velocities were enormous by the standards of stellar astronomy: several hundred kilometres per second, reaching 1100 km/s for the faintest objects measured, an order of magnitude beyond any stellar radial velocity then known. Second, the displacements were overwhelmingly toward the red: of the spirals Slipher had measured by 1917, the receding outnumbered the approaching roughly five to one. A sample drawn from bodies in random motion does not vote five to one for recession.

The observable is dimensionless and spectroscopic. If a line identified in the laboratory at wavelength λ_{em} is received at λ_{obs}, the *redshift* of the source is $z = (\lambda_{\mathrm{obs}} - \lambda_{\mathrm{em}})/\lambda_{\mathrm{em}}$, as restated below in Equation (48.1). At small z the special-relativistic Doppler formula of Chapter 38 reads $z \simeq v/c$ and licenses the word "velocity"; but the derivation of Phenomenon 48.1 will show that a cosmological redshift is not a velocity through space at all — it is a ratio of the sizes of the universe at emission and at reception, and the kinematic reading is legitimate only over distances small enough that the size has not changed. The precise statement is Equation (48.2).

48.1.2 The distance ladder

A redshift is read off a spectrum in an afternoon; a distance must be climbed to. Every extragalactic distance rests on a *ladder* of overlapping calibrations, and the propagation of error up that ladder is the central systematic of observational cosmology, then and now.

The base of the ladder is trigonometric parallax — pure geometry, the apparent semiannual displacement of a nearby star against the distant sky as the Earth traverses its orbit. It defines the parsec ($1\,\mathrm{pc} = 3.0857 \times 10^{16}\,\mathrm{m}$, the distance at which the astronomical unit subtends one second of arc) and it reaches, with modern astrometry, across a usable fraction of the Galaxy — but not remotely to another one.

The rung that first carried distances beyond the Galaxy was found by Leavitt in the Small Magellanic Cloud. Among the periodic variable stars now called Cepheids, she established that the apparent brightness at fixed period was tightly correlated with the period — and since every star in the Cloud sits at essentially one distance, apparent brightness there is absolute brightness up to one common constant [390]. A Cepheid's period, which no distance error can corrupt, therefore reveals its luminosity; comparing luminosity with observed flux gives the distance. Calibrate the zero point on parallax Cepheids in the Galaxy, and the ladder extends as far as individual Cepheids can be resolved. It was by resolving Cepheids in the spiral nebulae that Hubble settled their nature as galaxies and measured the distances that enter Section 48.1.3 [627]. The physics of the pulsation mechanism itself belongs to stellar astrophysics (Chapter 52).

Above the Cepheids sits the brightest standardizable candle known, the type Ia supernova, taken up in Section 48.6: a single event outshines its host galaxy and carries the ladder to redshifts where the expansion history itself becomes measurable. The modern three-rung ladder — parallax, Cepheid, supernova — is the instrument behind the local Hubble-constant measurement of Section 48.7.2 [2407], and each rung multiplies the error budget of the one below it, which is why so much of Section 48.7.2 turns on calibration.

48.1.3 The Hubble–Lemaître law

The linear law joining Slipher's velocities to the new distances has a double history, and the priority is worth recording precisely. In 1927 Lemaître derived, from the field equations of Chapter 44 applied to an expanding homogeneous universe, the prediction that recession velocity grows in proportion to distance, and extracted from Slipher's velocities and Hubble's early distance estimates an expansion rate of about $600\,\mathrm{km/(s\,Mpc)}$ [583]. The paper, in French in the *Annales* of the Brussels scientific society, went unread; when it was translated into English in 1931, the paragraphs deriving the rate were omitted. Hubble established the relation observationally in 1929, from Cepheid-calibrated distances to two dozen galaxies, with a slope near $500\,\mathrm{km/(s\,Mpc)}$ [627], and the law long carried his name alone. Both rates are some seven times the modern value: the error lay not in the velocities but in the first rung of the distance ladder, whose Cepheid zero point was systematically miscalibrated, and unwinding it took three decades [2379].

Phenomenon 48.1 (The Hubble–Lemaître law)**.** The spectral lines of galaxies beyond the Local Group are displaced toward longer wavelengths, and for nearby galaxies the displacement grows in proportion to distance,

$$cz = H_0 d + O\left(z^2\right), \qquad z := \frac{\lambda_{\mathrm{obs}} - \lambda_{\mathrm{em}}}{\lambda_{\mathrm{em}}}\,. \tag{48.1}$$

Slipher's radial velocities of the spiral nebulae established the displacements [451]; Lemaître derived the linear law from the field equations and extracted a rate from those same data [583]; Hubble established the relation observationally from Cepheid distances [627]. The

present value of H_0 is near $70\,\mathrm{km/(s\,Mpc)}$, with a discrepancy between the two ways of measuring it that is the subject of Section 48.7.2.

Rests on Equation (48.3) and Theorem 48.3.

Derivation. *Derives Phenomenon 48.1.*

In the FLRW geometry of Section 48.2 light reaching us radially travels on null curves with $c\,\mathrm{d}t = a(t)\,\mathrm{d}\chi$, where χ is the comoving radial coordinate of the source. Two successive wave crests, emitted at t_e and at $t_\mathrm{e} + \delta t_\mathrm{e}$ and received at t_0 and $t_0 + \delta t_0$, traverse the same comoving interval, so

$$\int_{t_\mathrm{e}}^{t_0} \frac{c\,\mathrm{d}t}{a(t)} = \int_{t_\mathrm{e}+\delta t_\mathrm{e}}^{t_0+\delta t_0} \frac{c\,\mathrm{d}t}{a(t)}\,.$$

Subtracting the common part of the two integrals leaves $\delta t_0/a(t_0) = \delta t_\mathrm{e}/a(t_\mathrm{e})$, and since a wavelength is c times the crest interval,

$$1 + z = \frac{a\,(t_0)}{a\,(t_\mathrm{e})}\,. \tag{48.2}$$

The redshift is therefore a ratio of scale factors, not a velocity: the special-relativistic Doppler formula of Chapter 38 applies only over distances small enough that a may be treated as constant.

For a nearby source expand about the present epoch, $a(t_\mathrm{e}) \simeq a(t_0)\,[1 - H_0\,(t_0 - t_\mathrm{e})]$ with $H_0 := \dot{a}\,(t_0)\,/a\,(t_0)$. Then Equation (48.2) gives $z \simeq H_0\,(t_0 - t_\mathrm{e})$, and to the same order the light-travel time is $t_0 - t_\mathrm{e} \simeq d/c$, whence $cz \simeq H_0 d$. Note what the derivation did not use: no field equation, no matter content, no spatial curvature. Linearity at small z follows from expansion alone, which is why confirming Equation (48.1) establishes that the universe expands without yet constraining what drives the expansion. □

In SI terms the Hubble constant is an inverse time: $H_0 = 67.4\,\mathrm{km/(s\,Mpc)} = 2.18 \times 10^{-18}\,\mathrm{/s}$ (computed here from the definition of the parsec), so that $1/H_0 \approx 14.5 \times 10^9\,\mathrm{yr}$ — the first crude indication, already available in 1929, that the universe has a history of finite depth comparable to the age of the oldest stars.

48.2 The FLRW geometry

48.2.1 Homogeneity, isotropy and the metric

Definition 48.2 (The cosmological principle)**.** On sufficiently large scales, and at each cosmic epoch, the universe is spatially *homogeneous* (no point of space is distinguished) and *isotropic* (no direction is distinguished about any point).

This is a falsifiable idealization, not an axiom of convenience, and the evidence for it is now strong: the microwave background is isotropic to about one part in 10^5 once the kinematic dipole is removed (Section 48.4.2), and galaxy counts approach uniformity when averaged over scales above roughly 100 Mpc [1871, 2348]. Below those scales the universe is conspicuously lumpy — that lumpiness is the subject of Section 48.5 — so the principle governs the *background* on which structure grows.

The mathematics of symmetric spaces was built once, in full generality, in Chapter 13: a metric's continuous symmetries are its Killing vectors (Definition 13.139), a D-dimensional space admits at most $D(D+1)/2$ of them, and a space admitting the full complement has curvature of constant-curvature form (Theorem 13.160). Here that machinery is instantiated on the spatial sections of spacetime.

Theorem 48.3 (The Robertson–Walker metric)**.** *A spacetime obeying Definition 48.2 — foliated by spacelike hypersurfaces each homogeneous and isotropic — admits coordinates (t, r, θ, φ) in which the metric takes the form*

$$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2\right)\right], \qquad k \in \{-1, 0, +1\}\,, \tag{48.3}$$

with r dimensionless, $a(t) > 0$ a function of t alone carrying the dimension of length, and k the sign of the spatial curvature. The uniqueness of this family was established group-theoretically by Robertson and Walker [758, 804].

Rests on Theorem 13.160, Proposition 13.162, and Remark 13.161.

Derivation. *Derives Theorem 48.3.*

Step 1: the slicing and cosmic time. Let the homogeneous slices be labelled by a parameter t and let x^i be coordinates within each slice, carried from slice to slice along the curves of constant x^i. In these coordinates the metric has blocks g_{00}, g_{0i}, g_{ij}. The vector field g_{0i} restricted to a slice is a spatial vector attached to every point; isotropy about every point forbids any nonvanishing spatial vector field defined by the geometry, so $g_{0i} = 0$: the slices are orthogonal to the curves of constant x^i, whose tangents define the *comoving observers*. Homogeneity makes g_{00} a function of t alone, and rescaling t absorbs it: $g_{00} = -c^2$. The label t is then the proper time of every comoving observer, synchronized across each slice — *cosmic time*.

Step 2: each slice has constant curvature. Each slice is a three-dimensional Riemannian space isotropic about every point. By Schur's argument (Remark 13.161), pointwise isotropy in $D = 3 > 2$ already forces the curvature into the maximally symmetric form Equation (13.318) with a single constant K over the slice: the slice is a space of constant curvature, of positive, zero or negative sign.

Step 3: the radial form of a constant-curvature 3-metric. For $K > 0$ the slice is (locally) the 3-sphere of radius $\ell = K^{-1/2}$, realizable by Proposition 13.162 as the quadric $X^2 + Y^2 + Z^2 + W^2 = \ell^2$ in Euclidean 4-space. Introduce on it the radius ρ of the 2-sphere orbits, so that $X^2 + Y^2 + Z^2 = \rho^2$ and $W^2 = \ell^2 - \rho^2$. Then $W\,\mathrm{d}W = -\rho\,\mathrm{d}\rho$, and the induced line element is

$$\mathrm{d}l^2 = \mathrm{d}\rho^2 + \rho^2\mathrm{d}\Omega^2 + \mathrm{d}W^2 = \frac{\mathrm{d}\rho^2}{1-\rho^2/\ell^2} + \rho^2\mathrm{d}\Omega^2\,,$$

with $\mathrm{d}\Omega^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2$. For $K < 0$ the same computation on the quadric of Proposition 13.162 with $\sigma = -1$ (in signature $(3,1)$) flips the sign under ρ^2/ℓ^2; for $K = 0$ the slice is flat and the line element is $\mathrm{d}\rho^2 + \rho^2\mathrm{d}\Omega^2$. Writing $\rho = \ell r$ with r dimensionless unifies the three cases into the bracket of Equation (48.3) with $k = \operatorname{sgn} K$ and an overall factor ℓ^2.

Step 4: the scale factor. The one geometric datum distinguishing slice from slice is the curvature radius ℓ, which homogeneity makes a single number per slice but which may vary with t: call it $a(t)$. Isotropy forbids any direction-dependent rescaling between slices (an anisotropic map of one slice onto the next would define preferred directions), and the sign k cannot change continuously. The metric is therefore Equation (48.3); for $k = 0$ the normalization of a is conventional, since only $\dot a/a$ and $\ddot a/a$ will appear in observables. That this family is the unique one — that no further functional freedom survives the symmetry — is the content of the group-theoretic classification of Robertson and Walker [758, 804], to which we refer for the completeness argument. □

Comoving coordinates deserve one dynamical remark before use: an object released at rest in them stays at rest (its worldline is a geodesic, by the isotropy of Equation (48.3) about it), and an object moving through them slows down.

Proposition 48.4 (Peculiar momentum decays with the expansion)**.** *Let a free particle move through the flat ($k = 0$) FLRW geometry with physical momentum p measured by the comoving observer it passes. Then along its geodesic*

$$p(t)\, a(t) = const\,, \qquad i.e. \qquad p \propto \frac{1}{a}\,. \tag{48.4}$$

Rests on Proposition 13.141 and Equation (48.3).

Derivation. *Derives Proposition 48.4.*

For $k = 0$ the spatial translations ∂_i are Killing vectors of Equation (48.3) (they preserve the flat slice metric, whose overall factor $a^2(t)$ they do not touch), so by Proposition 13.141, for the translation $\boldsymbol{\xi} = \partial_x$ along any fixed comoving direction x, the quantity $\xi_\mu u^\mu = g_{xx} u^x = a^2 \frac{\mathrm{d}x}{\mathrm{d}\tau}$ is constant along the geodesic. The physical momentum component measured locally is $p_x = m\, a \frac{\mathrm{d}x}{\mathrm{d}\tau}$ (the proper distance swept per proper time is $a\,\mathrm{d}x$), so $p_x a = m\, a^2 \frac{\mathrm{d}x}{\mathrm{d}\tau}$ is the conserved quantity, which is Equation (48.4). For a photon the same argument runs with the affine parameter in place of τ, and $p \propto 1/a$ reproduces the redshift Equation (48.2), since $p = h/\lambda$. Peculiar velocities of galaxies decay the same way, which is why the comoving frame is not merely a convention but an attractor. □

48.2.2 The Friedmann equations

Symmetry fixed the metric up to one function $a(t)$ and one sign k; the field equations of Chapter 44 now fix the dynamics of a. The same symmetry first constrains the source: isotropy about every comoving observer forces the stress–energy tensor to have no energy flux and no anisotropic stress in the comoving frame — the only tensors available are $g_{\mu\nu}$ and the observer 4-velocity u^μ — so it must take the perfect-fluid form

$$T_{\mu\nu} = (\varepsilon + p)\, \frac{u_\mu u_\nu}{c^2} + p\, g_{\mu\nu}\,, \tag{48.5}$$

with $\varepsilon(t)$ the energy density ($\mathrm{J/m^3}$) and $p(t)$ the pressure (Pa), both spatially constant by homogeneity.

Theorem 48.5 (The Friedmann equations)**.** *For the metric Equation (48.3) sourced by Equation (48.5), the field equations with cosmological constant Λ reduce to*

$$H^2 := \left(\frac{\dot a}{a}\right)^2 = \frac{8\pi G}{3c^2}\,\varepsilon - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3}\,, \tag{48.6}$$

$$\frac{\ddot a}{a} = -\frac{4\pi G}{3c^2}\,(\varepsilon + 3p) + \frac{\Lambda c^2}{3}\,, \tag{48.7}$$

together with the continuity equation

$$\dot\varepsilon + 3H\,(\varepsilon + p) = 0\,. \tag{48.8}$$

The first two were obtained from the field equations by Friedmann, for positive and for negative spatial curvature [490, 513], and independently by Lemaître [583].

Rests on Equations (48.3) and (48.5) and Postulate 44.3.

Derivation. *Derives Theorem 48.5.*

Write the metric as $g_{00} = -c^2$, $g_{ij} = a^2(t)\gamma_{ij}$, where γ_{ij} is the time-independent bracket of Equation (48.3): a 3-metric of constant curvature k, whose Ricci tensor is $R^{(3)}_{ij} = 2k\gamma_{ij}$ by Equation (13.318) in $D = 3$. The nonvanishing Christoffel symbols of Theorem 13.150 involving time are, by direct computation from $\Gamma^\lambda{}_{\mu\nu} = \frac{1}{2} g^{\lambda\sigma}(\partial_\mu g_{\sigma\nu} + \partial_\nu g_{\sigma\mu} - \partial_\sigma g_{\mu\nu})$,

$$\Gamma^t{}_{ij} = \frac{a\dot a}{c^2}\,\gamma_{ij}\,, \qquad \Gamma^i{}_{tj} = \frac{\dot a}{a}\,\delta^i{}_j\,, \tag{48.9}$$

plus the purely spatial symbols of γ_{ij}. Assembling the Ricci tensor (Definition 13.153) from these,

$$R_{tt} = -3\,\frac{\ddot{a}}{a}\,, \qquad R_{ij} = \left[\frac{a\ddot{a} + 2\dot{a}^2}{c^2} + 2k\right]\gamma_{ij}\,,$$

and the scalar

$$R = g^{tt}R_{tt} + g^{ij}R_{ij} = \frac{6}{c^2}\left[\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{kc^2}{a^2}\right].$$

The Einstein tensor components follow:

$$G_{tt} = 3\left[\frac{\dot{a}^2}{a^2} + \frac{kc^2}{a^2}\right], \qquad G_{ij} = -\frac{1}{c^2}\left[2a\ddot{a} + \dot{a}^2 + kc^2\right]\gamma_{ij}\,.$$

In the comoving frame $u^\mu = (1,0,0,0)$, $u_t = -c^2$, so Equation (48.5) gives $T_{tt} = \varepsilon c^2$ and $T_{ij} = p\,a^2\gamma_{ij}$. The tt component of the field equations $G_{\mu\nu} + \Lambda g_{\mu\nu} = (8\pi G/c^4)\,T_{\mu\nu}$ (Chapter 44) reads

$$3\left[\frac{\dot{a}^2}{a^2} + \frac{kc^2}{a^2}\right] - \Lambda c^2 = \frac{8\pi G}{c^2}\,\varepsilon\,,$$

which is Equation (48.6). The ij components, after dividing out $a^2\gamma_{ij}$ and eliminating $\dot{a}^2/a^2 + kc^2/a^2$ with Equation (48.6), collapse to Equation (48.7).

For Equation (48.8), impose $\nabla_\mu T^{\mu t} = 0$ — a consequence of the field equations via the Bianchi identities (Chapter 44), not a separate assumption. With $T^{tt} = \varepsilon/c^2$ and $T^{ij} = (p/a^2)\gamma^{ij}$,

$$\nabla_\mu T^{\mu t} = \partial_t\frac{\varepsilon}{c^2} + \Gamma^{\mu}{}_{\mu t}\frac{\varepsilon}{c^2} + \Gamma^{t}{}_{ij}\,T^{ij} = \frac{1}{c^2}\left[\dot{\varepsilon} + 3H\,(\varepsilon + p)\right] = 0\,,$$

using $\Gamma^{\mu}{}_{\mu t} = 3\dot{a}/a$ and Equation (48.9). Only two of the three equations are independent: differentiating Equation (48.6) and using Equation (48.8) reproduces Equation (48.7). □

Definition 48.6 (Critical density and the density parameters)**.** The *critical density* at expansion rate H is the energy density that makes the $k = 0$, Λ-free universe balance in Equation (48.6):

$$\varepsilon_{\mathrm{c}} := \frac{3H^2c^2}{8\pi G}\,, \qquad \rho_{\mathrm{c}} := \frac{\varepsilon_{\mathrm{c}}}{c^2} = \frac{3H^2}{8\pi G}\,. \tag{48.10}$$

Each component's *density parameter* is its present share, $\Omega_i := \varepsilon_i(t_0)/\varepsilon_{\mathrm{c}}(t_0)$, with the conventions $\Omega_\Lambda := \Lambda c^2/(3H_0^2)$ and $\Omega_k := -kc^2/(a_0^2H_0^2)$, in terms of which Equation (48.6) evaluated today reads

$$\Omega_{\mathrm{r}} + \Omega_{\mathrm{m}} + \Omega_k + \Omega_\Lambda = 1\,. \tag{48.11}$$

For $H_0 = 67.4\,\mathrm{km/(s\,Mpc)}$ the critical density is $\rho_{\mathrm{c}} = 8.5\times10^{-27}\,\mathrm{kg/m^3}$ (computed here), about five hydrogen atoms per cubic metre — the gulf between that number and any laboratory vacuum is a useful reminder of how empty the universe is on average.

48.2.3 Distances and horizons

In an expanding curved geometry "distance" must be defined operationally, and the two definitions that match what telescopes actually do — comparing fluxes and comparing angles — differ from each other and from the coordinate separation. All are built on the comoving radial coordinate accumulated by a light ray. Define

$$\chi := \int_0^r \frac{\mathrm{d}r'}{\sqrt{1 - kr'^2}}\,, \qquad S_k(\chi) := \begin{cases} \sin\chi & k = +1\,,\\ \chi & k = 0\,,\\ \sinh\chi & k = -1\,, \end{cases} \tag{48.12}$$

so that $r = S_k(\chi)$ and the metric reads $\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + a^2[\mathrm{d}\chi^2 + S_k^2(\chi)\,\mathrm{d}\Omega^2]$. A radial null ray obeys $c\,\mathrm{d}t = a\,\mathrm{d}\chi$, so a source seen today at redshift z sits at

$$\chi(z) = \frac{c}{a_0}\int_0^z \frac{\mathrm{d}z'}{H(z')}\,, \qquad H(z) = H_0\,E(z)\,, \tag{48.13}$$

where the dimensionless expansion history

$$E(z) = \sqrt{\Omega_\mathrm{r}(1+z)^4 + \Omega_\mathrm{m}(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda} \tag{48.14}$$

follows from Equation (48.6) once the scaling of each component with a is known — derived in Section 48.3.1 below.

Proposition 48.7 (Luminosity distance)**.** *Let a source of intrinsic luminosity L (W) at redshift z be observed with energy flux F (W/m^2). Then the* luminosity distance $d_\mathrm{L} := \sqrt{L/(4\pi F)}$ *evaluates to*

$$d_\mathrm{L} = (1+z)\,a_0\,S_k(\chi(z))\,. \tag{48.15}$$

Rests on Equations (48.2) and (48.13) and Theorem 48.3.

Derivation. *Derives Proposition 48.7.*

The photons emitted in all directions at comoving radius χ are, at reception, spread over a sphere whose proper area is $4\pi a_0^2 S_k^2(\chi)$ (set $\mathrm{d}t = \mathrm{d}\chi = 0$ in the metric and integrate the angular part at the observer's epoch). Two further factors of $(1+z)$ deplete the flux: each photon arrives with its energy degraded by $1+z$ (Equation (48.2) applied to $E = h\nu$), and the arrival *rate* of photons is diluted by the same factor (the crest-interval argument in the derivation of Phenomenon 48.1 applies to any clock, photon spacing included). Hence

$$F = \frac{L}{4\pi a_0^2 S_k^2(\chi)\,(1+z)^2}\,,$$

and solving $F = L/(4\pi d_\mathrm{L}^2)$ for d_L gives Equation (48.15). □

Corollary 48.8 (Angular-diameter distance and the duality relation)**.** *An object of proper transverse size ℓ at redshift z subtends the angle $\theta = \ell/d_\mathrm{A}$ with*

$$d_\mathrm{A} = a(t_\mathrm{e})\,S_k(\chi(z)) = \frac{d_\mathrm{L}}{(1+z)^2}\,. \tag{48.16}$$

Rests on Equation (48.15) and Proposition 48.7.

Derivation. *Derives Corollary 48.8.*

The proper size ℓ was laid down at emission: setting $\mathrm{d}t = \mathrm{d}\chi = 0$ *at the emission epoch*, the transverse proper length subtending coordinate angle θ at comoving radius χ is $\ell = a(t_\mathrm{e})S_k(\chi)\,\theta$, whence the first equality; the second follows from Equation (48.15) and $a_0/a(t_\mathrm{e}) = 1+z$. The relation $d_\mathrm{L} = (1+z)^2 d_\mathrm{A}$ holds in any metric theory with photon-number conservation, which makes it a test of the framework itself rather than of a parameter set. □

Two horizons bound what can be seen and what can be reached. The *particle horizon* is the comoving radius light can have crossed since the initial time,

$$\chi_\mathrm{p}(t) = \int_0^t \frac{c\,\mathrm{d}t'}{a(t')}\,, \tag{48.17}$$

finite whenever a grows fast enough near $t = 0$ — in the radiation era $a \propto t^{1/2}$ and the integral converges at the lower end, so at every epoch there *is* a finite sphere of causal

contact, a fact that becomes the horizon problem of Section 48.8. Today the proper particle horizon is about 14 Gpc (computed here from Equation (48.14) with the parameters of Table 48.1). Complementarily, with $\Lambda > 0$ the future integral $\int_t^\infty c\,\mathrm{d}t'/a(t')$ converges (exponential expansion), so an *event horizon* exists: comoving regions beyond it can never be signalled, and galaxies presently leave our future light cone [1871]. The Hubble radius $c/H_0 \approx 4.5$ Gpc (computed) is not a horizon but a curvature scale; it enters the perturbation theory of Section 48.5.1 as the boundary of the Newtonian regime.

48.3 The thermal history

48.3.1 Radiation, matter and Λ eras

Proposition 48.9 (Scaling of the energy components)**.** *Let a component have equation of state $p = w\varepsilon$ with w constant. Then Equation (48.8) integrates to*

$$\varepsilon(a) = \varepsilon_0 \left(\frac{a_0}{a}\right)^{3(1+w)} , \tag{48.18}$$

so nonrelativistic matter ($w = 0$) dilutes as a^{-3}, radiation ($w = 1/3$) as a^{-4}, and the Λ term ($w = -1$, $\varepsilon_\Lambda = \Lambda c^4/8\pi G$ constant) not at all.

Rests on Equation (48.8).

Derivation. Derives Proposition 48.9.

With $p = w\varepsilon$, Equation (48.8) reads $\mathrm{d}\ln\varepsilon = -3(1+w)\,\mathrm{d}\ln a$, which is Equation (48.18). The physics of the two nontrivial cases: matter dilutes with the volume, a^{-3}; radiation dilutes with the volume *and* each quantum loses energy as $1/a$ by Equation (48.4), giving the fourth power. The constancy of the Λ term is what makes it inevitable that, given enough expansion, it dominates — the ordering of the eras below is not an accident of initial conditions. □

Because the four terms of Equation (48.14) scale with four different powers, the history divides into eras with sharp crossovers. Radiation–matter equality sits at $1 + z_{\mathrm{eq}} = \Omega_{\mathrm{m}}/\Omega_{\mathrm{r}} \approx 3400$ (computed here from the densities of Table 48.1, with the photon density from the measured T_0 and the standard three neutrino species); matter–Λ equality at $1 + z = (\Omega_\Lambda/\Omega_{\mathrm{m}})^{1/3}$, i.e. $z \approx 0.30$, and the expansion began accelerating (Equation (48.7) changing sign) at $z \approx 0.6$, both computed from the same table. Solving Equation (48.6) in each pure era gives $a \propto t^{1/2}$ (radiation), $a \propto t^{2/3}$ (matter), and $a \propto \exp\left(\sqrt{\Lambda/3}\,ct\right)$ (Λ) — the last being the reason for the event horizon of Section 48.2.3.

Temperature. The relic radiation cools as $T \propto 1/a$ (Equation (48.24), derived with the CMB in Section 48.4.1); running that backwards makes the early universe arbitrarily hot, and equilibrium thermodynamics (Chapter 120) governs the plasma. Two workhorse relations follow. First, for the coupled relativistic plasma the comoving entropy is conserved: the expansion of Equation (48.8) is adiabatic (no heat crosses the boundary of a comoving volume — there is nowhere for it to come from, by homogeneity), so $s\,a^3 = \text{const}$ with $s \propto g_* T^3$ the entropy density of the relativistic species (Chapter 120), g_* counting their effective degrees of freedom. While g_* is constant this reproduces $T \propto 1/a$; when a species annihilates, its entropy is inherited by the rest, and the photon temperature briefly falls more slowly than $1/a$.

The one clean application is a prediction. Neutrinos decouple from the plasma just before electrons and positrons annihilate, so the $e^\pm$ entropy goes to the photons alone. Before annihilation the electromagnetic sector holds $g_* = 2 + \frac{7}{8} \times 4 = \frac{11}{2}$ (photons plus $e^\pm$);

after, $g_* = 2$. Conservation of $g_*(Ta)^3$ across the transition gives $(T_\gamma a)^3_{\text{after}} = \frac{11}{4}(Ta)^3_{\text{before}}$, while the decoupled neutrinos keep $(T_\nu a)$ fixed, so

$$\frac{T_\nu}{T_\gamma} = \left(\frac{4}{11}\right)^{1/3}, \qquad T_{\nu,0} = 1.95\,\text{K} \quad \text{(computed)}\,. \tag{48.19}$$

The relic neutrino sea itself has never been detected directly; its reality is attested indirectly, through the radiation density it contributes at nucleosynthesis (Section 48.3.2) and in the CMB damping tail [2348] — an honest amber, recorded as such.

Second, in the radiation era Equation (48.6) with $\varepsilon = \frac{g_*}{2}\, a_\text{B} T^4$ ($a_\text{B} = 4\sigma/c$ the radiation constant, Chapter 68) and $a \propto t^{1/2}$ gives the time–temperature relation

$$t = \left(\frac{3c^2}{16\pi G\, g_*\, a_\text{B}}\right)^{1/2} T^{-2}\,, \tag{48.20}$$

so $T = 1 \times 10^{10}\,\text{K}$ is reached at $t \approx 1\,\text{s}$ (with $g_* = 10.75$: photons, $e^\pm$, three neutrino species) and $T = 1 \times 10^9\,\text{K}$ at $t \approx 3\,\text{min}$ (with $g_* = 3.36$ after annihilation); both computed here. The first minutes are nuclear physics territory, taken up next.

Recombination. The plasma remains ionized long after $k_\text{B}T$ falls below the hydrogen binding energy of 13.6 eV, because photons outnumber baryons by $\eta^{-1} \approx 1.6 \times 10^9$ (Section 48.3.2) and the Wien tail keeps ionizing. The ionization balance — the Saha equation [470], an application of the chemical equilibrium of Chapter 120 — drops the free-electron fraction steeply only once $k_\text{B}T \approx 0.26\,\text{eV}$, i.e. $T \approx 3000\,\text{K}$, at $z_* \approx 1090$ (the precise value, $z_* = 1089.9$, is quoted from the fit of Table 48.1). The photons then fly free: that release is the last-scattering surface whose image is the microwave background of Section 48.4.

48.3.2 Big-Bang nucleosynthesis

The proposal that the light elements were assembled in the hot first minutes is the alpha–beta–gamma paper of 1948 [905]; the governing role of the neutron-to-proton ratio and its weak-interaction freeze-out was worked out by Hayashi [960]; and the full reaction network in the modern sense — a stiff system of coupled rate equations integrated through the expansion — was first computed by Wagoner, Fowler and Hoyle [1361], whose descendants remain in use [2357]. The heavier elements are *not* made here: there is no stable nucleus at mass 5 or 8, and the low density starves the triple-α process that stars use to jump the gap (Chapter 52, [1044]); primordial synthesis stalls at ^{7}Li.

Phenomenon 48.10 (A primordial floor of light elements)**.** The least processed material accessible to observation — metal-poor extragalactic gas clouds, and the absorption systems seen against high-redshift quasars — carries a common floor of light elements that stellar processing cannot produce: a helium mass fraction close to 0.25, a deuterium abundance D/H of a few parts in 10^5, and traces of ^{3}He and ^{7}Li. In the hot-big-bang picture these abundances form a one-parameter family governed by the ratio of baryons to photons alone [905, 2357], and the value of that ratio which fits them agrees with the entirely independent determination from the microwave-background power spectrum [2348]. Deuterium is the sharpest of the four: it is destroyed in stellar interiors and made nowhere else, so any observed deuterium is a lower bound on the primordial value.

Rests on Equation (48.20) and Theorem 48.5.

Derivation. *Derives Phenomenon 48.10.*

Step 1: freeze-out of the neutron fraction. Above $k_\text{B}T \approx 1\,\text{MeV}$ the weak reactions $n + \nu_e \leftrightarrow p + e^-$ and $n + e^+ \leftrightarrow p + \bar{\nu}_e$ (Chapter 107) hold neutrons and protons in chemical

equilibrium, so their ratio is the Boltzmann factor of the mass difference $\Delta mc^2 = 1.293\,\mathrm{MeV}$:

$$\frac{n}{p} = \exp\left(-\frac{\Delta m\, c^2}{k_{\mathrm{B}} T}\right) . \tag{48.21}$$

The interconversion rate per nucleon scales as $\Gamma \sim \left[G_{\mathrm{F}}/(\hbar c)^3\right]^2 (k_{\mathrm{B}}T)^5/\hbar$ (the Fermi coupling squared times the phase space of the five-dimensional final-state integral, Chapter 107), while the expansion rate scales as $H \sim \sqrt{g_* G/(\hbar^3 c^5)}\,(k_{\mathrm{B}}T)^2$ by Equation (48.6); their ratio falls as T^3, so equilibrium is abruptly lost. Carrying the numerical factors of the full rates through (we import them from the standard treatment [960, 2357]), $\Gamma = H$ at $k_{\mathrm{B}}T_{\mathrm{f}} \approx 0.8\,\mathrm{MeV}$, where Equation (48.21) gives $n/p \approx 1/5$; residual out-of-equilibrium conversions push the ratio to about $1/6$ over the following seconds.

Step 2: the deuterium bottleneck. Fusion must proceed through deuterium, $p + n \to \mathrm{D} + \gamma$, whose binding energy is only $B_{\mathrm{D}} = 2.224\,\mathrm{MeV}$. Although $k_{\mathrm{B}}T < B_{\mathrm{D}}$ already at freeze-out, each baryon is bathed in $\eta^{-1} \approx 1.6 \times 10^9$ photons, and photodissociation wins as long as the number of photons in the Wien tail above B_{D} exceeds the number of baryons — roughly, while $\eta^{-1}\exp\left(-B_{\mathrm{D}}/k_{\mathrm{B}}T\right) > 1$. Solving with $\eta = 6.1 \times 10^{-10}$ gives $k_{\mathrm{B}}T \approx B_{\mathrm{D}}/21 \approx 0.1\,\mathrm{MeV}$; the full rate balance places the breakout slightly lower, near $k_{\mathrm{B}}T \approx 0.08\,\mathrm{MeV}$, i.e. $T \approx 1 \times 10^9\,\mathrm{K}$, reached at $t \approx 3\,\mathrm{min}$ by Equation (48.20). The delay matters because free neutrons decay with mean life $\tau_n = 878.4\,\mathrm{s}$ (Chapter 107): by breakout the ratio has fallen to $n/p \approx 1/7$.

Step 3: helium. Once deuterium survives, the chain $\mathrm{D}(p,\gamma)^3\mathrm{He}$, $\mathrm{D}(\mathrm{D},n)^3\mathrm{He}$, $^3\mathrm{He}(\mathrm{D},p)^4\mathrm{He}$ runs to completion in minutes, because ^{4}He is by far the most bound light nucleus (Chapter 107). To first approximation *every surviving neutron ends in helium*: with $x := n/p$ at breakout, the helium mass fraction is

$$Y_{\mathrm{p}} = \frac{2x}{1+x} \approx 0.25 \quad \text{for } x = \tfrac{1}{7} . \tag{48.22}$$

The estimate is robust precisely because it rests on a counting argument, not on cross sections; the full network sharpens it to $Y_{\mathrm{p}} = 0.2469 \pm 0.0002$ at the CMB baryon density [2357].

Step 4: the deuterium residue and the baryometer. A trace of deuterium fails to burn, at a level set by the competition between the burning rate (proportional to the baryon density) and the expansion: a *higher* baryon density burns deuterium more completely, so D/H is a steeply decreasing function of η — the baryometer. The network prediction at the CMB-determined density lies between 2.5 and 2.6 parts in 10^5, the spread reflecting nuclear-rate inputs [2357]; the measurement, from damped Lyman-α absorbers along quasar sightlines, is $\mathrm{D/H} = (2.527 \pm 0.030) \times 10^{-5}$ [2308]. The measured helium fraction in metal-poor H II regions, $Y_{\mathrm{p}} = 0.2449 \pm 0.0040$ [2250], agrees with Equation (48.22) and its refinement. The full integration of the coupled rate network through the expansion — dozens of reactions, evaluated against measured cross sections — is imported here by reference [1361, 2357] rather than restated; the analytic skeleton above is what makes its output intelligible. □

The baryon-to-photon ratio used above, $\eta = 6.1 \times 10^{-10}$, is computed here from the CMB baryon density $\Omega_{\mathrm{b}}h^2$ of Table 48.1 and the photon density implied by the measured T_0 — and the point of Phenomenon 48.10 is that the abundances measured in gas clouds return the same number from entirely different physics at an epoch separated by a factor of 10^9 in temperature.

Remark 48.11 (The lithium discrepancy). The account above is honest only if its one failure is stated: the same network that gets D and ^{4}He right predicts a primordial ^{7}Li abundance about a factor of three above what is measured on the Spite plateau of metal-poor halo stars [2357]. Whether the fault lies in stellar depletion of lithium, in the astrophysics of the measurement, or in the network itself is not settled; the discrepancy is recorded here as an open defect of the concordance, not explained away.

48.3.3 The prediction of the relic radiation

The hot early phase required by Section 48.3.2 does not switch itself off: if the universe was once a thermal plasma, its radiation must still be here, cooled by the expansion but preserving its thermal form (the proof of that preservation is given with Phenomenon 48.12 below). In 1948 Gamow set out the framework in which nucleosynthesis fixes the conditions at $T \sim 10^9\,\mathrm{K}$ [914], and Alpher and Herman, correcting and completing his estimate, drew the explicit conclusion: a residual blackbody radiation should fill the universe today, at a temperature they put at about 5 K [906].

The structure of their inference deserves to be exhibited in modern numbers, because it is a genuine prediction — made seventeen years before the discovery, and not being searched for when the discovery came. The deuterium bottleneck fixes the baryon density *at* $T \approx 1\times10^9\,\mathrm{K}$: capture must win against expansion, which by the cross sections requires $\rho_\mathrm{b} \sim 1\times10^{-2}\,\mathrm{kg/m^3}$ there (computed here from the modern network normalization; Alpher and Herman's value was cruder). Astronomy supplies the *present* mean baryon density, $\rho_{\mathrm{b},0} \sim 4\times10^{-28}\,\mathrm{kg/m^3}$. Since $\rho_\mathrm{b} \propto a^{-3}$ while $T \propto a^{-1}$ (Proposition 48.9 and Equation (48.24)), the expansion factor between the two epochs is $(\rho_\mathrm{b}/\rho_{\mathrm{b},0})^{1/3} \approx 4\times10^8$, and the present radiation temperature must be

$$T_0 \approx \frac{1\times10^9\,\mathrm{K}}{4\times10^8} \approx 2.5\,\mathrm{K}\,. \tag{48.23}$$

With the modern inputs the estimate lands on the measured value; with the 1948 inputs it gave 5 K — the right physics and the right order of magnitude, published in *Nature*, and then forgotten for seventeen years. When Penzias and Wilson found the radiation in 1965 (Section 48.4.1) they were unaware of the prediction, and the interpreting paper [1305] cited the hot-big-bang logic afresh. As a case study in how evidence accumulates — prediction, neglect, independent rediscovery — cosmology has no better example (Chapter 1).

48.4 The cosmic microwave background

48.4.1 Discovery

In 1964 Penzias and Wilson, commissioning the 6 m horn antenna at Holmdel for radio astronomy, found an excess antenna temperature of $(3.5 \pm 1.0)\,\mathrm{K}$ at 4080 MHz that no accounting could remove: it was independent of direction, of season, and of every instrumental cause they could test, down to the famous evicted pigeons [1320]. The interpretation — that this is the cooled relic of the hot early universe — was supplied in the companion paper by Dicke, Peebles, Roll and Wilkinson [1305], whose own purpose-built radiometer was months from completion. The spectrum was finally measured absolutely from orbit by COBE/FIRAS, and it is the most perfect Planck spectrum known (Chapter 68); the instruments, data and error budgets belong to the experiment chapter, Chapter 51.

Phenomenon 48.12 (The relic microwave background)**.** The whole sky radiates in the microwave band with a spectrum indistinguishable from a Planck spectrum, at a temperature $T_0 = 2.725\,48\,\mathrm{K}$ known to $0.000\,57\,\mathrm{K}$ [2144], and isotropic to about one part in 10^5 once the dipole of Section 48.4.2 is removed. It was found as an irreducible excess antenna temperature at 4.08 GHz that no instrumental or local source would account for [1320], and identified in the companion paper as the relic of a hot early phase [1305] — a state of affairs predicted seventeen years earlier [906]. Absolute spectrophotometry from orbit found the departures from the Planck form to be below 10^{-4} of the peak intensity [1834, 1922], making this the closest approach to a blackbody ever measured anywhere, laboratory sources included.

Rests on Equation (48.2) *and Theorem 48.3.*

Derivation. *Derives Phenomenon 48.12.*

What must be shown is that a thermal spectrum survives free expansion with its temperature as the only changing parameter — otherwise the observed Planck form could not be a relic, the universe having expanded by a factor of order 10^3 since the photons last scattered. Let the radiation be in equilibrium at temperature T_e when the scale factor is a_e, so that the mean occupation number of a mode of frequency ν is the Planck value of Chapter 68,

$$f(\nu) = \frac{1}{\exp\left(h\nu/\left(k_\mathrm{B}T_\mathrm{e}\right)\right) - 1}\,.$$

Now expand freely to scale factor a. Two things happen, and only two. Every mode redshifts, $\nu \to \nu' = \nu\, a_\mathrm{e}/a$, by Equation (48.2) applied to the mode itself; and the occupation number of each mode is unchanged, since with no emission, absorption or scattering the photons are neither created nor destroyed while the modes themselves are simply carried along by the expansion. Hence

$$f'\left(\nu'\right) = f(\nu) = \frac{1}{\exp\left(h\nu' a/\left(a_\mathrm{e}k_\mathrm{B}T_\mathrm{e}\right)\right) - 1} = \frac{1}{\exp\left(h\nu'/\left(k_\mathrm{B}T\right)\right) - 1}\,,$$

which is again of Planck form, with

$$T = T_\mathrm{e}\frac{a_\mathrm{e}}{a}\,, \qquad \text{equivalently} \qquad T(z) = T_0\left(1+z\right)\,. \tag{48.24}$$

The spectrum is preserved exactly and the single parameter T scales as $1/a$. Two consequences follow. The measured T_0 fixes the radiation temperature at every earlier epoch, which is what makes the thermal history of Section 48.3 calculable; and any measured departure from the Planck form would be a record of energy injected after thermalization ceased, so the spectral limits quoted above are constraints on the entire history since then. □

48.4.2 Anisotropies

The first anisotropy seen was the *dipole*: one side of the sky is hotter than the other by 3.4 mK, detected from a U-2 aircraft in 1977 [1610] and measured by FIRAS at (3.372 ± 0.014) mK [1922]. It is kinematic: a Doppler modulation of Equation (48.24) by our own motion, $\Delta T/T_0 = v/c$, giving $v \approx 370$ km/s (computed) for the barycentre of the Solar System relative to the frame in which the radiation is isotropic — the comoving frame of Section 48.2.1 made tangible. It is removed before the cosmological analysis begins.

Phenomenon 48.13 (The anisotropy spectrum of the relic radiation)**.** Beneath the dipole, the microwave sky carries temperature fluctuations of about one part in 10^5: an rms of roughly 30 µK on ten-degree scales, first detected by the COBE differential radiometers [1858]. At angular scales below a degree the fluctuation power is organized into a series of acoustic peaks, the first near multipole $\ell \approx 220$, measured with high precision by the Planck satellite [2348]. The large-angle plateau was predicted as the gravitational imprint of density perturbations on the last-scattering surface [1357], and the peak structure as the frozen phases of sound waves in the pre-recombination plasma [1431, 1434] — both before any anisotropy had been seen. The apparatus, maps and likelihoods are the subject of Chapter 51.

Rests on Equation (48.24) and Phenomenon 48.12.

Derivation. *Derives Phenomenon 48.13.*

Two mechanisms carry the primordial perturbations onto the sky, and each is derivable in outline here.

The Sachs–Wolfe plateau. Consider a large-scale density perturbation at last scattering — large enough that no causal process has acted on it (Equation (48.17)), so only gravity

speaks. Let Φ be its Newtonian potential ($\Phi < 0$ in an overdensity). Photons climbing out of the well are gravitationally redshifted (Chapter 42): $\Delta T/T|_{\rm grav} = \Phi/c^2$. But the well also retards the local clocks: the proper time elapsed at coordinate time t is shifted by $\delta\tau/\tau = \Phi/c^2$, so the plasma in the well reaches its recombination temperature *later* in cosmic time, by $\delta t/t = -\Phi/c^2$; at that later time the scale factor is larger by $\delta a/a = \frac{2}{3}\,\delta t/t$ (matter era, $a \propto t^{2/3}$), and photons released at fixed local temperature from a larger a suffer less subsequent cooling, arriving hotter by $\Delta T/T|_{\rm delay} = -\frac{2}{3}\Phi/c^2$. The sum is

$$\frac{\Delta T}{T} = \frac{\Phi}{c^2} - \frac{2}{3}\frac{\Phi}{c^2} = \frac{\Phi}{3c^2}\,, \tag{48.25}$$

cold spots over overdensities, at one third of the naive gravitational redshift. The bookkeeping above is heuristic in its treatment of time; the gauge-consistent computation, which confirms the factor 1/3 for adiabatic perturbations in the matter era, is Sachs and Wolfe's [1357] and is imported by reference. With $\Delta T/T \sim 10^{-5}$, Equation (48.25) measures the depth of the primordial potentials, $\Phi/c^2 \sim 3 \times 10^{-5}$ — the normalization every structure-formation calculation in Section 48.5 inherits.

The acoustic peaks. Inside the horizon the pre-recombination plasma is a single tightly coupled fluid, and perturbations in it oscillate as sound waves (the restoring force is photon pressure; the derivation of the sound speed is given with Phenomenon 48.15). All modes of a given wavenumber k start their oscillation together — at horizon entry, with the same phase, because they share the same super-horizon initial condition — so at the instant of decoupling the modes with $k\,r_{\rm s} = n\pi$, where $r_{\rm s}$ is the comoving sound horizon (Equation (48.34)), are caught at the extrema of their oscillation. Those wavenumbers are imprinted as excess variance: projected on the sky through the angular-diameter distance of Corollary 48.8, they appear at multipoles

$$\ell_n \approx n\,\frac{\pi}{\theta_*}\,, \qquad \theta_* = \frac{r_{\rm s}(z_*)}{(1+z_*)\,d_{\rm A}(z_*)}\,, \tag{48.26}$$

up to an order-unity shift from the projection geometry and from the gravitational driving of the oscillations. Computing θ_* here from Equations (48.13) and (48.34) with the parameters of Table 48.1 gives $\theta_* = 1.040 \times 10^{-2}$ rad, against the measured 1.0411×10^{-2} [2348] — agreement at the level of a part in a thousand, which is the concordance of Section 48.7 in miniature. The measured first peak sits at $\ell_1 \approx 220$. On scales smaller still the peaks die away: photons random-walk out of wavelengths shorter than their diffusion length before decoupling completes, erasing the fluctuations — Silk damping, predicted in 1968 [1383] and measured in the high-ℓ tail [2348]. □

What the peak pattern buys, quantitatively, is leverage on nearly every parameter of Section 48.7: the sound horizon is calculable from pre-recombination physics, the peak angle measures the geometry through Equation (48.26), the relative peak heights measure the baryon and matter densities (baryons load the oscillating fluid asymmetrically; matter sets the epoch of equality and the driving). The extraction, with the instruments and likelihoods behind it, is Chapter 51.

48.5 Large-scale structure

48.5.1 Growth of perturbations

The universe of Definition 48.2 is smooth; the universe outside the window is not. The bridge is gravitational instability: the fluctuations of 10^{-5} measured on the last-scattering surface grow under their own gravity into the galaxies and clusters of today. The instability of a self-gravitating medium was analysed by Jeans for a static background [306]; the general-relativistic perturbation theory of an expanding universe is Lifshitz's [892]. On scales

well inside the Hubble radius and for nonrelativistic matter, Newtonian fluid mechanics (Chapter 31) suffices, and that is the regime derived here; the relativistic treatment is imported by reference where super-horizon scales are concerned [892].

Theorem 48.14 (Linear growth of density perturbations)**.** *Let $\delta := \delta\rho/\bar\rho$ be the fractional density perturbation of a pressureless matter component in an FLRW background, on scales well inside the Hubble radius. To linear order each Fourier mode $\delta_{\boldsymbol{k}}$ (comoving wavevector $\boldsymbol{k}$) obeys*

$$\ddot\delta_{\boldsymbol{k}} + 2H\dot\delta_{\boldsymbol{k}} = \left(4\pi G\bar\rho - \frac{c_s^2 k^2}{a^2}\right)\delta_{\boldsymbol{k}}\,, \tag{48.27}$$

with c_s the sound speed of the medium. In the matter era the pressureless growing mode is

$$\delta \propto a\,, \tag{48.28}$$

power-law growth, not the exponential growth of the static case.
Rests on Equations (31.10), (31.19) and (48.6) and Theorem 48.5.

Derivation. *Derives Theorem 48.14.*

Start from the continuity and Euler equations of Chapter 31 (Equations (31.10) and (31.19)) with the gravitational force $-\nabla\Phi$, and the field equation of Newtonian gravity in local form, $\nabla^2\Phi = 4\pi G\rho$ (Chapter 27). The background solution is the expansion itself: $\bar\rho(t) \propto a^{-3}$, velocity field $\bar{\boldsymbol{v}} = H\boldsymbol{r}$, and $\bar\Phi = \frac{2}{3}\pi G\bar\rho r^2$, which satisfies the three equations by virtue of Equations (48.6) and (48.7) restricted to their Newtonian content.

Perturb: $\rho = \bar\rho(1+\delta)$, $\boldsymbol{v} = H\boldsymbol{r} + \boldsymbol{u}$, $\Phi = \bar\Phi + \phi$, and pass to comoving coordinates $\boldsymbol{x} = \boldsymbol{r}/a$, in which $\nabla_r = a^{-1}\nabla_x$ and the time derivative at fixed $\boldsymbol{x}$ replaces that at fixed $\boldsymbol{r}$ via $\partial_t|_r = \partial_t|_x - H\boldsymbol{x}\cdot\nabla_x$. To first order in $(\delta, \boldsymbol{u}, \phi)$ the three equations become

$$\dot\delta + \frac{1}{a}\nabla_x\cdot\boldsymbol{u} = 0\,, \tag{48.29}$$

$$\dot{\boldsymbol{u}} + H\boldsymbol{u} = -\frac{1}{a}\nabla_x\phi - \frac{c_s^2}{a}\nabla_x\delta\,, \tag{48.30}$$

$$\nabla_x^2\phi = 4\pi G\bar\rho\, a^2\delta\,, \tag{48.31}$$

where the pressure term wrote $\nabla p = c_s^2\bar\rho\,\nabla\delta$. Take ∂_t of Equation (48.29), insert Equation (48.30), eliminate $\nabla_x\cdot\boldsymbol{u}$ with Equation (48.29) again and ϕ with Equation (48.31), and Fourier-transform in $\boldsymbol{x}$: the result is Equation (48.27). The Hubble term $2H\dot\delta$ — absent in the static case — acts as friction: expansion carries matter away from the well as fast as gravity gathers it.

Setting $H = 0$, $\bar\rho$ constant recovers Jeans's dichotomy [306]: modes with $k < k_{\mathrm{J}}$, where

$$\frac{c_s^2 k_{\mathrm{J}}^2}{a^2} = 4\pi G\bar\rho\,, \tag{48.32}$$

grow exponentially at rate $\sqrt{4\pi G\bar\rho}$, shorter modes oscillate as sound. In the expanding matter era take $c_s \to 0$, $a \propto t^{2/3}$, so $H = 2/(3t)$ and $4\pi G\bar\rho = \frac{3}{2}H^2 = 2/(3t^2)$ (Einstein–de Sitter values from Equation (48.6)); the ansatz $\delta \propto t^n$ gives $3n^2 + n - 2 = 0$, i.e. $n = 2/3$ or $n = -1$. The growing mode is $\delta \propto t^{2/3} \propto a$, which is Equation (48.28). During radiation domination the rapid expansion (H set by the radiation, not by the matter) overwhelms the self-gravity of the matter and growth stalls logarithmically; once Λ dominates, growth stops [1871]. □

Theorem 48.14 carries an arithmetic consequence that is one of the strongest arguments in this book. From recombination to today the scale factor grows by $1 + z_* \approx 1100$, so a baryonic fluctuation of the measured amplitude $\sim 10^{-5}$ (Phenomenon 48.13) can have

grown by at most that factor — to 10^{-2}, and less once the stalls at both ends are counted. Galaxies, with $\delta \gg 1$, could not exist. The resolution is a matter component that does not couple to the photons: its perturbations, unsupported by radiation pressure, began growing at matter–radiation equality ($z_{\mathrm{eq}} \approx 3400$, Section 48.3.1) and had dug potential wells well before recombination; the baryons, released at z_*, fell into wells already prepared [1699]. This is the structure-formation argument for dark matter cited in Phenomenon 48.16.

The initial spectrum feeding Equation (48.27) was guessed on general grounds before it was measured: Harrison and Zel'dovich argued independently that the only spectrum injecting equal potential fluctuations on every scale — the only one with no built-in preferred length — has spectral index $n_s = 1$ [1424, 1491]. The measured value, $n_s = 0.9649 \pm 0.0042$ [2348], is close to but significantly below that: a small tilt, whose possible origin is taken up in Section 48.8.

48.5.2 Baryon acoustic oscillations

The sound waves that produce the CMB peaks (Phenomenon 48.13) do not vanish at recombination: they freeze into the *matter* distribution too, and their imprint is measurable in galaxy surveys billions of years later — a ruler of known length available at many redshifts.

Phenomenon 48.15 (A standard ruler in the galaxy distribution)**.** The two-point correlation function of galaxies carries a single narrow excess at a comoving separation of about 150 Mpc, riding on an otherwise smooth, nearly power-law clustering. It was first detected in the luminous-red-galaxy sample of the Sloan survey [2076]. The feature sits at the comoving sound horizon frozen in when the photons decoupled from the baryons — the same length that sets the angular spacing of the acoustic peaks in the microwave background [2348] — so it is a ruler of calculable length laid down at a known redshift. Measuring its apparent size along and across the line of sight at several redshifts maps the expansion history geometrically, with no standard candle and no distance ladder involved.

Rests on Equations (48.16) *and* (48.34) *and Corollary 48.8.*

Derivation. *Derives Phenomenon 48.15.*

Step 1: the sound speed of the coupled fluid. Before recombination, Thomson scattering ties the baryons to the photons into a single fluid whose pressure is supplied almost entirely by the photons, $p_\gamma = \varepsilon_\gamma/3$ (Chapter 68), and whose inertia is shared. For adiabatic perturbations the photon and baryon densities fluctuate together (equal number-density contrasts, so $\delta\varepsilon_\gamma/\varepsilon_\gamma = \frac{4}{3}\,\delta\rho_{\mathrm{b}}/\rho_{\mathrm{b}}$), and the sound speed follows from $c_s^2 = c^2\,\delta p/\delta\varepsilon$:

$$c_s^2 = \frac{c^2}{3\,(1+R)}\,, \qquad R := \frac{3\varepsilon_{\mathrm{b}}}{4\varepsilon_\gamma} \propto a\,, \tag{48.33}$$

since only the photons push while both components weigh. Deep in the radiation era $R \ll 1$ and $c_s \to c/\sqrt{3}$; by recombination R is of order unity and the fluid has audibly slowed.

Step 2: the sound horizon. A pressure wave launched at the initial time has travelled, by decoupling, the comoving distance

$$r_{\mathrm{s}} = \int_{t_i}^{t_{\mathrm{d}}} \frac{c_s(t)\,\mathrm{d}t}{a(t)/a_0} = \int_{z_{\mathrm{d}}}^{\infty} \frac{c_s(z)\,\mathrm{d}z}{H(z)}\,, \tag{48.34}$$

the acoustic analogue of Equation (48.17). Evaluating the integral here with Equations (48.14) and (48.33) and the densities of Table 48.1, at the drag epoch $z_{\mathrm{d}} \approx 1060$ (when the baryons, outnumbered 10^9 to one, are finally released by the photons — slightly after the photons are released by the baryons), gives $r_{\mathrm{d}} = 147$ Mpc comoving, in agreement with the fitted value $r_{\mathrm{d}} = (147.09 \pm 0.26)$ Mpc quoted by Planck [2348].

Step 3: survival as a bump. Picture a single pointlike primordial overdensity: dark matter, baryons, photons and neutrinos superposed. The photon–baryon component launches a

spherical sound wave; the dark matter, pressureless, stays and grows by Theorem 48.14. At the drag epoch the wave stalls: the pressure that drove it vanishes with the coupling, and the baryon shell is stranded at radius r_{d}. Thereafter both the central dark-matter peak and the baryon shell grow by the same linear theory — Equation (48.27) is linear, so the configuration is preserved up to overall amplitude, and gravity gradually shares the shell between the components. Since galaxies form where matter is dense, pairs of galaxies separated by r_{d} are slightly more common than the smooth clustering predicts: the correlation function acquires a bump at r_{d}, of relative amplitude set by the baryon fraction $\Omega_{\mathrm{b}}/\Omega_{\mathrm{m}} \approx 1/6$ — a percent-level feature, which is why finding it took a survey of 46 748 luminous red galaxies over $0.72\,h^{-3}\,\mathrm{Gpc}^3$ [2076]. Nonlinear evolution at late times broadens the bump slightly but shifts its centre only at the sub-percent level, which is what qualifies it as a standard ruler [2076, 2348]. □

Used as a ruler, the feature measures $d_{\mathrm{A}}(z)$ transverse to the line of sight and $c/H(z)$ along it (Corollary 48.8); anchored to the same r_{d} that the CMB calibrates, BAO surveys extend the geometric measurement from z_* down through the era of Λ domination, and they are the third leg — with the CMB and the supernovae — of every concordance fit [2348].

48.5.3 Evidence for dark matter

The dark-matter inference is often told as one discovery; it is better told as three independent measurements that agree on a number. Zwicky, applying the virial theorem to the Coma cluster in 1933, found the velocity dispersion of the member galaxies to imply a binding mass vastly exceeding the luminous one [729]. Rubin and Ford, four decades later, traced the rotation of M31 through the emission lines of its H II regions and found no Keplerian decline [1432], a result their survey of field spirals made general [1658]. And the microwave background requires, through the peak structure of Phenomenon 48.13 and the growth argument of Section 48.5.1, a gravitating component that does not scatter light, at five times the baryon density [2348]. Gravitational lensing supplies the most direct image: in the merging cluster 1E 0657−56 the X-ray–emitting gas — most of the baryonic mass — was stripped and slowed by the collision, while the lensing mass sailed through with the galaxies, mapping a dominant collisionless component spatially separated from the baryons [2062, 2093].

Phenomenon 48.16 (Flat rotation curves and the missing mass)**.** The orbital speed of gas and stars in a spiral galaxy does not decline beyond the visible disc: it stays nearly constant out to the largest radius at which any tracer can be found [1432]. The same discrepancy appears on the scale of clusters, where the velocity dispersion of the member galaxies implies a binding mass far above the luminous one [729], and again in the microwave background and in the growth of structure, both of which require a matter component that neither emits nor scatters radiation [2348]. The three arguments are independent, and they agree on the amount.

Rests on Equation (48.27) *and Theorem 48.14.*

Derivation. *Derives Phenomenon 48.16.*

For a test particle on a circular orbit of radius r inside a spherically symmetric mass distribution, Newtonian dynamics gives

$$\frac{v^2}{r} = \frac{GM(r)}{r^2}\,, \qquad \text{that is} \qquad M(r) = \frac{v^2 r}{G}\,,$$

where $M(r)$ is the mass interior to r. Were the luminous matter all the matter, $M(r)$ would be essentially constant outside the optical disc and v would fall off as $r^{-1/2}$, the Keplerian

behaviour familiar from the solar system. The observed constancy of v instead forces

$$M(r) \propto r\,, \qquad \rho(r) = \frac{1}{4\pi r^2}\frac{\mathrm{d}M}{\mathrm{d}r} \propto \frac{1}{r^2}\,, \tag{48.35}$$

an enclosed mass still growing linearly at radii where the surface brightness has long since fallen away exponentially. The inference is purely kinematic: it uses only Newtonian gravity in the weak-field, slow-motion regime together with the assumption of a roughly spherical halo, and it fixes how much unseen mass there is without saying anything whatever about what it is made of.

The cluster version is the same argument with the virial theorem in place of the circular orbit, and the theorem is one line from Newton's equations (Chapter 19). For N bodies with positions $\boldsymbol{r}_i$ and momenta $\boldsymbol{p}_i$, let $\mathcal{G} := \sum_i \boldsymbol{p}_i \cdot \boldsymbol{r}_i$; then $\dot{\mathcal{G}} = 2T + \sum_i \boldsymbol{F}_i \cdot \boldsymbol{r}_i = 2T + U$, the last step by Euler's homogeneous-function identity applied to the gravitational potential energy U, homogeneous of degree -1 in the positions. For a bound system $\mathcal{G}$ stays finite, so the long-time average of $\dot{\mathcal{G}}$ vanishes: $2\langle T\rangle + \langle U\rangle = 0$. A cluster of radius R whose galaxies move with velocity dispersion σ therefore has mass of order $M \sim \sigma^2 R/G$. For Coma, $\sigma \approx 1000\,\mathrm{km/s}$ and R of a few megaparsecs give a mass one to two orders of magnitude above what the starlight accounts for — Zwicky's 1933 arithmetic, essentially unchanged [729]. □

What the particle might be — and the searches that have so far excluded much of the candidate space — is deliberately not discussed here: the evidence above is gravitational and belongs to this chapter; the particle physics is a frontier and belongs to Chapter 134, under the standards of Chapter 1.

48.6 The accelerating universe

48.6.1 Type Ia supernovae

A type Ia supernova is the thermonuclear disruption of a white dwarf driven over its stability limit by accretion (Chapter 52); because the igniting mass is nearly universal, so is the luminosity, and what scatter remains is correlated with the light-curve's decline rate — broader is brighter — tightly enough that a measured light curve standardizes the peak magnitude to about 0.15 magnitudes [1873]. That makes the events standardizable candles visible across half the observable universe, and in the late 1990s two independent teams pushed the Hubble diagram to $z \approx 0.5$–1 expecting to measure the deceleration of the expansion [1962, 1981]. Both found the opposite sign.

Phenomenon 48.17 (The expansion is accelerating)**.** Type Ia supernovae near $z \approx 0.5$, standardized through the correlation between light-curve width and peak luminosity, are observed systematically fainter — and so at systematically larger luminosity distance — than a matter-dominated, decelerating universe predicts, by an amount that grows with redshift [1962, 1981]. Read through the FLRW distance relations of Section 48.2.3, the observation says that $\ddot{a} > 0$ at the present epoch, that is that the deceleration parameter

$$q_0 := -\left.\frac{\ddot{a}a}{\dot{a}^2}\right|_{t_0} \tag{48.36}$$

is negative. The concordance fit, which now folds in the microwave background and the acoustic ruler of Phenomenon 48.15, attributes about 69% of the present energy budget to a smooth component whose equation of state is consistent with $w = -1$ and about 31% to matter [2348]. What has been measured is an expansion history; no mechanism has been measured.

Rests on Equations (48.7) *and* (48.15)*, Proposition 48.7, and Theorem 48.5.*

Derivation. *Derives Phenomenon 48.17.*

Step 1: the kinematic distance–redshift relation. Expand the scale factor about the present,

$$a(t) = a_0 \left[1 + H_0 \, (t - t_0) - \frac{q_0}{2} H_0^2 \, (t - t_0)^2 + \cdots \right] ,$$

which *defines* q_0 via Equation (48.36) and uses no field equation. Inverting Equation (48.2) order by order gives the look-back time in terms of z, and inserting it into the null-ray integral $\chi = \int c \, \mathrm{d}t / a$ and then into Equation (48.15) (curvature enters only at third order) yields

$$d_{\mathrm{L}}(z) = \frac{cz}{H_0} \left[1 + \frac{1 - q_0}{2} \, z + O\!\left(z^2\right) \right] . \tag{48.37}$$

The first term is Equation (48.1); the second is where the dynamics hides. At $z = 0.1$ the two-term formula agrees with the exact integral for the concordance parameters to a quarter of a percent (computed here), so the expansion is trustworthy where the calibrating supernovae live and begins to bend measurably by $z \approx 0.5$.

Step 2: magnitudes. Optical astronomy measures the *distance modulus*

$$\mu := m - M = 5 \log_{10} \frac{d_{\mathrm{L}}}{10 \, \mathrm{pc}} , \tag{48.38}$$

so a fractional distance excess ϵ registers as $\Delta\mu = (5/\ln 10) \, \epsilon \approx 2.17 \, \epsilon$ magnitudes. By Equation (48.37) a change of q_0 shifts the modulus at redshift z by $\Delta\mu \approx 2.17 \, z \, |\Delta q_0| \, /2$: distinguishing the matter-only Einstein–de Sitter value $q_0 = +\frac{1}{2}$ from an accelerating $q_0 \approx -0.5$ requires photometry good to a few tenths of a magnitude at $z \approx 0.5$ — exactly what standardized type Ia photometry provides.

Step 3: what was measured. Both teams found the high-redshift supernovae *fainter* than any decelerating model allows. Against the exact integrals of Equation (48.13) (computed here at $z = 0.5$): the concordance universe places a supernova 0.40 magnitudes fainter than Einstein–de Sitter, 0.19 fainter than an open $\Omega_{\mathrm{m}} = 0.2$ universe, and 0.12 fainter than an empty coasting one — and the data preferred distances near the concordance values, 10–15% beyond the open model [1962], with the 42-supernova fit of the second team giving $\Omega_{\mathrm{m}} \approx 0.28$ under flatness [1981]. Through Equations (48.6) and (48.7), evaluated today,

$$q_0 = \frac{\Omega_{\mathrm{m}}}{2} + \Omega_{\mathrm{r}} - \Omega_\Lambda ,$$

so the measured $\Omega_{\mathrm{m}} \approx 0.3$, $\Omega_\Lambda \approx 0.7$ give $q_0 \approx -0.53$ (computed): the expansion accelerates. The systematics that could have mimicked the signal — grey dust along the line of sight, luminosity evolution between the young universe and the old — were the subject of a decade of scrutiny and are bounded by the later data: the Hubble diagram now extends beyond $z = 1$ into the epoch where the matter term of Equation (48.14) dominates and the model predicts the faintness trend must *reverse* into relative brightening, which is observed [2402]. Dust and evolution do not reverse sign on cue. □

48.6.2 The cosmological constant as dark energy

The minimal reading of Phenomenon 48.17 is already in the field equations. The constant Λ was introduced by Einstein in 1917 to allow a static universe [445] — a motivation the redshifts of Section 48.1 demolished — but the term itself is not optional decoration: it is the one addition the structure of the field equations permits (Chapter 44), and the data now demand it. With $\Lambda > 0$, Equation (48.7) turns positive once matter dilutes sufficiently, and the fit of Table 48.1 gives (computed here from Ω_Λ and H_0)

$$\Lambda = \frac{3 \Omega_\Lambda H_0^2}{c^2} = 1.1 \times 10^{-52} / \mathrm{m}^2 , \qquad \varepsilon_\Lambda = \frac{\Lambda c^4}{8 \pi G} = 5.2 \times 10^{-10} \, \mathrm{J/m}^3 , \tag{48.39}$$

about 3.3 GeV per cubic metre: two protons' worth of rest energy in every cubic metre of vacuum. Interpreted as a fluid, the Λ term has $w = -1$ exactly (Proposition 48.9); the fitted equation of state, from the combination of CMB, BAO and supernovae, is consistent with -1 at the three-percent level [2348, 2402]. Nothing in the data yet distinguishes a true constant from a slowly varying impostor.

Two honest difficulties frame the subject. First, the *smallness problem*: quantum field theory assigns the vacuum an energy density, and every naive estimate of it — cutting the zero-point sum off at any scale from the electroweak to the Planck scale — exceeds Equation (48.39) by some 55 to 120 orders of magnitude [1828]. Why the measured value is not zero yet absurdly small compared to its natural scale is unexplained; Weinberg's review remains the standard statement of how resistant the problem is to every class of mechanism proposed. Second, the *anthropic bound*: Weinberg observed, a decade before the acceleration was found, that a Λ much larger than the matter density at the epoch of structure formation would have prevented galaxies from forming at all, so observers can only find themselves in universes with Λ at most of roughly the observed order [1799]. Whether that observation is an explanation is a question about scientific method rather than about data; this treatise records it as an argument, notes that it predicted the right order of magnitude, and does not admit it as evidence (Chapter 1).

What is measured, in sum: an expansion history $H(z)$, consistent across three independent probes with a constant vacuum energy of the value Equation (48.39). What is not measured: any property of the component beyond its density and its pressure, or any mechanism setting either. "Dark energy" is a name for the gap between those two sentences.

48.7 The ΛCDM parametrization

48.7.1 Six numbers

The base model that fits every data stream of this chapter is defined by Equation (48.3) with $k = 0$, the Friedmann dynamics of Theorem 48.5 sourced by baryons, cold dark matter, radiation and Λ, and a power-law spectrum of adiabatic initial perturbations. Six sampled numbers specify it: two densities, one angle, one optical depth, and the amplitude and tilt of the initial spectrum. Everything else — the Hubble constant included — is derived. Table 48.1 lists the six with the 2018 Planck values, and a selection of the derived quantities this chapter has used; here $h := H_0/(100\,\mathrm{km/(s\,Mpc)})$.

The model's economy is its strongest and its weakest feature at once. Strongest: six numbers, fitted to the microwave background alone, then *predict* the BAO scale in galaxy surveys, the supernova Hubble diagram, the light-element abundances through $\Omega_\mathrm{b}h^2$, and the broad shape of the matter power spectrum — and the predictions land (Phenomena 48.10, 48.15 and 48.17). Weakest: of the six, only $\Omega_\mathrm{b}h^2$ names physics whose microscopic identity is known. The model does not say what the cold component is, why Λ has the value Equation (48.39), or what set A_s and n_s. It is a parametrization whose parameters are measured superbly and explained not at all; they take their place in the ledger of unexplained constants in Appendix C.

48.7.2 The Hubble tension

Phenomenon 48.18 (The Hubble tension)**.** The two best determinations of the present expansion rate disagree. Calibrating type Ia supernovae on the Cepheid distance ladder gives $H_0 = 73.0\,\mathrm{km/(s\,Mpc)}$ with an uncertainty near $1.0\,\mathrm{km/(s\,Mpc)}$ [2407]; fitting the microwave-background power spectra within the same six-parameter model that fits everything else in this chapter gives $H_0 = 67.4\,\mathrm{km/(s\,Mpc)}$ with an uncertainty near $0.5\,\mathrm{km/(s\,Mpc)}$ [2348]. The two differ by about five times their combined uncertainty. The disagreement has

Table 48.1: The six sampled parameters of the base ΛCDM model, and selected derived quantities, from the Planck 2018 fit to the temperature, polarization and lensing power spectra [2348]. All uncertainties are the quoted 68% intervals. Every number in this table is a fit through the model; none is a direct measurement of the quantity named.

Parameter	Symbol	Value
Baryon density	$\Omega_{\mathrm{b}}h^2$	$0.022\,37 \pm 0.000\,15$
Cold-dark-matter density	$\Omega_{\mathrm{c}}h^2$	0.1200 ± 0.0012
Acoustic angular scale	$100\,\theta_{\mathrm{MC}}$	$1.040\,92 \pm 0.000\,31$
Reionization optical depth	τ	0.0544 ± 0.0073
Primordial amplitude	$\ln(10^{10}A_{\mathrm{s}})$	3.044 ± 0.014
Scalar spectral index	n_{s}	0.9649 ± 0.0042
Hubble constant	H_0	$(67.36 \pm 0.54)\,\mathrm{km/(s\,Mpc)}$
Matter fraction	Ω_{m}	0.3153 ± 0.0073
Dark-energy fraction	Ω_Λ	0.6847 ± 0.0073
Fluctuation amplitude	σ_8	0.8111 ± 0.0060
Age of the universe	t_0	$(13.797 \pm 0.023)\,\mathrm{Gyr}$
Last-scattering redshift	z_*	1089.92 ± 0.25
Sound horizon, drag epoch	r_{d}	$(147.09 \pm 0.26)\,\mathrm{Mpc}$

survived a decade of scrutiny of both error budgets and of the rungs of the ladder, and it is at present unresolved (Chapter 133). It is recorded here as a measurement in conflict — one of the two determinations, or the model connecting them, is wrong — and not as evidence for any particular resolution.

Rests on Equations (48.15) *and* (48.26)*, Phenomenon 48.13, and Proposition 48.7.*

Derivation. *Derives Phenomenon 48.18.*

What must be exhibited is where each number comes from, since the conflict lives entirely in the inference chains.

The ladder value. The chain of Section 48.1.2: geometric anchors (parallaxes, the megamaser distance to NGC 4258, detached eclipsing binaries in the LMC) calibrate the Cepheid period–luminosity law; Cepheids in 42 host galaxies calibrate the peak luminosity of type Ia supernovae; the calibrated supernovae, read through Equation (48.38) and Equation (48.37) in the redshift range where the q_0 correction is small, give $H_0 = (73.04 \pm 1.04)\,\mathrm{km/(s\,Mpc)}$ [2407]. The measurement is local ($z \lesssim 0.15$) and nearly model-independent: no assumption about early-universe physics enters.

The CMB value. The chain of Phenomenon 48.13: the peak spacing measures θ_* to three parts in 10^4 (Table 48.1); the sound horizon $r_{\mathrm{s}}(z_*)$ is *computed* from Equation (48.34), whose only inputs are the densities $\Omega_{\mathrm{b}}h^2$ and $\Omega_{\mathrm{m}}h^2$ — both fixed by the peak heights — and standard pre-recombination physics; the ratio then fixes the comoving distance to last scattering through Equation (48.26), and that distance is the integral Equation (48.13) of $c/H(z)$ over the *entire* expansion history. Within ΛCDM — flat, with constant Λ — the integral has no freedom left once the densities are set, and H_0 comes out at $(67.36 \pm 0.54)\,\mathrm{km/(s\,Mpc)}$ [2348]. The value is thus an *inference through the model*, anchored at $z = 1090$ and extrapolated across the whole history; the BAO ruler of Phenomenon 48.15, calibrated by the same r_{d}, confirms the extrapolation at intermediate redshifts.

The comparison. The difference is $5.7\,\mathrm{km/(s\,Mpc)}$ against a combined uncertainty of $1.2\,\mathrm{km/(s\,Mpc)}$: 4.8σ as computed here from the two quoted errors, "about five sigma" as commonly stated [2407]. For the disagreement to be an accident of errors, one budget must be wrong by five times its stated size. The candidate systematics have been hunted for a decade: Cepheid crowding photometry, metallicity terms and anchor distances on the ladder

side — partially checked by the independent tip-of-the-red-giant-branch calibration, which lands between the two values, near 70 km/(s Mpc) with larger uncertainty [2379] — and foregrounds and likelihood choices on the CMB side, where independent instruments and analyses reproduce the low value [2348]. The third logical possibility is the model itself: the CMB chain rests on the computed r_s, so any new physics that shrinks the sound horizon by a few percent — extra relativistic energy density before recombination, for instance — would raise the inferred H_0. No such extension is currently supported by the data it would have to leave undisturbed. The tension stands, and this treatise records it as standing (Chapter 133). □

48.8 Inflation: the evidence status

48.8.1 The proposal

The hot big bang of this chapter is spectacularly successful from one second onward and silent about its own initial conditions. Two of those conditions are quantitatively strange, and the arithmetic deserves to be shown.

The horizon problem. The particle horizon at last scattering, Equation (48.17) evaluated through the radiation and matter eras, is a comoving 280 Mpc (computed here); the comoving distance to the last-scattering surface is 13.9 Gpc (computed, Equation (48.13)). A causally connected patch therefore has an angular radius of about 1.2° on the sky, and the visible last-scattering surface comprises on the order of 10^4 such regions that, within the hot big bang alone, have never exchanged a signal (computed). Yet Phenomenon 48.12 shows them at a common temperature to a part in 10^5. Equilibrium without contact is not impossible as an initial condition — initial conditions are free — but it is unexplained structure.

The flatness problem. Write Equation (48.6) as $|\Omega(t) - 1| = |k|\,c^2/(a^2H^2)$: the deviation from critical density scales as the inverse square of aH, which *grows* throughout radiation and matter domination. Running the measured bound backwards (computed here from Equation (48.14)): today's $|\Omega - 1| < 0.002$ (Section 48.8.2) was $\sim 10^{-6}$ at z_{eq} and below 10^{-17} at the weak freeze-out that opens nucleosynthesis ($t \approx 1$ s, Section 48.3.2). The hot big bang must be *handed* a universe tuned to critical density to seventeen decimal places for one to be here measuring it.

Guth's 1981 observation is that a single mechanism voids both problems [1670]: a phase of accelerated expansion *before* the radiation era.

Proposition 48.19 (Accelerated expansion shrinks the comoving Hubble radius)**.** *The comoving Hubble radius $c/(aH) = c/\dot{a}$ decreases exactly when $\ddot{a} > 0$.*

Rests on Equation (48.7).

Derivation. *Derives Proposition 48.19.*

$\frac{\mathrm{d}}{\mathrm{d}t}\,(c/\dot{a}) = -c\,\ddot{a}/\dot{a}^2$, negative iff $\ddot{a} > 0$. During such a phase $|\Omega - 1| = |k|\,c^2/\dot{a}^2$ is driven *toward* zero rather than away — flatness becomes an attractor — and comoving regions vastly larger than the instantaneous Hubble radius were inside it, in causal contact, before the phase began: the uniformity of the last-scattering surface is inherited from a causally connected past. Sixty-odd e-foldings of such expansion suffice to bring the whole presently observable volume within one pre-inflationary causal patch and to flatten it beyond the measured bounds [1670]; the counting is standard and we do not repeat it. □

Guth's own realization stalled in its exit from the accelerated phase; the workable slow-roll versions came a year later [1677, 1694]. What made the framework more than a cure for two aesthetic complaints is a by-product: the quantum fluctuations of the field driving the phase are stretched to super-horizon scales and become a nearly scale-invariant, slightly tilted spectrum of density perturbations — computed by Mukhanov and Chibisov

in 1981 [1672], eleven years before COBE saw any anisotropy at all (Phenomenon 48.13). A framework that predicted the statistical character of the seeds of all structure, in advance, has earned its place in this chapter.

48.8.2 Measured and unmeasured

The discipline of Chapter 1 requires the balance sheet to be explicit.

Measured, and consistent with the framework. Spatial curvature: the CMB combined with BAO gives $\Omega_k = 0.0007 \pm 0.0019$ — flat to two parts in a thousand [2348], where inflation predicts flatness to experimental precision. The spectral tilt: $n_s = 0.9649 \pm 0.0042$, eight standard deviations below exact scale invariance (computed from the quoted values), with the sign and rough size that generic slow roll predicts [1672, 2348]. Adiabaticity and Gaussianity: the fluctuations are adiabatic and Gaussian to the limits of measurement [2348], as the simplest single-field realizations require.

Not measured. The distinguishing prediction of the framework — the one signal no alternative shares — is a background of primordial gravitational waves, imprinted as B-mode polarization in the CMB. None has been detected: the tensor-to-scalar ratio is bounded at $r < 0.036$ at 95% confidence [2369]. With no tensor signal there is no measurement of the energy scale of the proposed phase, no identification of the field that drove it, and no test of the mechanism as opposed to its outputs. The consistency successes above are real but not exclusive: they check predictions that some alternative histories could also produce, whereas the tensor background is the framework's own signature.

The verdict, in this treatise's terms: inflation is mainstream cosmology and belongs in this chapter as the leading account of the initial conditions — an account whose generic outputs (flatness, tilt, adiabatic Gaussian perturbations) are measured and agree, and whose distinguishing prediction remains untested. It is reported in exactly those terms, and its speculative extensions — which trade in what cannot be observed — are excluded from this book by the scope rule of Chapter 1 (Chapter 133 records what a future B-mode detection, or a definitive null, would each settle).

Chapter 49

Compact Stars and Relativistic Astrophysics

A star that has exhausted its nuclear fuel can no longer hold itself up by thermal pressure, and what remains is a contest between gravity and quantum mechanics: the degeneracy pressure of fermions, first invoked for white dwarfs by Fowler [541], is the only cold support matter possesses, and general relativity sets a ceiling on how much of it gravity will tolerate. This chapter follows that contest through the Tolman–Oppenheimer–Volkoff equations [845, 848], the white-dwarf mass limit of Chandrasekhar [676], the neutron stars predicted by Baade and Zwicky [730] and found as pulsars [1373], up to the maximum masses and radii that pulsar timing, X-ray pulse-profile modelling and the binary merger GW170817 [2278] now measure.

The chapter is the meeting point of three others. The stellar models whose endpoints these objects are belong to Chapter 52; the field equations that govern their interiors are those of Chapter 44; and the horizon that forms when every equilibrium fails is the subject of Chapter 45, with its observational record in Chapter 50. The standard monograph is [1727].

49.1 White dwarfs

49.1.1 The observational puzzle: Sirius B

The companion of Sirius was found by Bessel in 1844 from the wobble it imprints on the bright star's proper motion, and seen telescopically in 1862; its mass, from the astrometric orbit, is close to the Sun's. The surprise came in 1915, when Adams obtained its spectrum and found not the cool red star its faintness suggested but a white-hot one [423]. A hot surface radiating so little light must be small: the Stefan–Boltzmann law forces a radius near the Earth's, and therefore a mean density of order $10^{9}\,\mathrm{kg/m^{3}}$ — matter in a state no classical equation of state allows. The gravitational redshift of Sirius B, the classical test it later furnished, is treated with the other tests in Chapter 42.

Phenomenon 49.1 (Matter at a billion kilograms per cubic metre)**.** The faint companion of Sirius shows the spectrum of a hot white star, yet its luminosity is some ten thousand times below its primary's, which for that surface temperature fixes its radius near the Earth's; its mass, from the astrometric orbit, is close to the Sun's [423]. The implied mean density is of order $10^{9}\,\mathrm{kg/m^{3}}$, a million times that of ordinary solids. Thousands of white dwarfs are now catalogued and they occupy the same narrow band of the mass–radius plane. No classical equation of state supports matter in this state; at such densities the electrons are degenerate and the support is quantum-mechanical, surviving to zero temperature [541].

Rests on Proposition 49.2 and Phenomenon 83.1.

Derivation. *Derives Phenomenon 49.1.*

What must be checked is that the electrons really are degenerate at this density — that the exclusion principle of Chapter 83, and not thermal agitation, sets the pressure. Take $\rho = 10^9\,\mathrm{kg/m^3}$ of fully ionized material with two nucleons per electron, so that

$$n_\mathrm{e} = \frac{\rho}{2m_\mathrm{u}} = \frac{10^9\,\mathrm{kg/m^3}}{2 \times 1.661 \times 10^{-27}\,\mathrm{kg}} \approx 3.0 \times 10^{35}/\mathrm{m^3}\,.$$

The Fermi energy of a free electron gas at that density, from Chapter 120, is

$$E_\mathrm{F} = \frac{\hbar^2}{2m_\mathrm{e}} \left(3\pi^2 n_\mathrm{e}\right)^{2/3} \approx 2.6 \times 10^{-14}\,\mathrm{J} \approx 160\,\mathrm{keV}\,,$$

whereas the interior of a white dwarf, at a temperature of order $10^7\,\mathrm{K}$, has $k_\mathrm{B}T \approx 1.4 \times 10^{-16}\,\mathrm{J} \approx 0.9\,\mathrm{keV}$. The ratio $E_\mathrm{F}/\left(k_\mathrm{B}T\right) \approx 200$ settles the question: essentially every state below the Fermi level is occupied, the gas is degenerate, and the pressure it exerts does not vanish as the star cools. That is why a white dwarf can sit in equilibrium with no energy source whatever, which no thermally supported star can do.

The same arithmetic carries a warning. The Fermi energy already reaches about a third of the electron rest energy $m_\mathrm{e}c^2 = 511\,\mathrm{keV}$, so the non-relativistic formula used above is only an estimate, and a denser — that is, a more massive — white dwarf will be relativistic. That is the regime of Section 49.1.3, and it is where the mass limit comes from. □

49.1.2 Electron degeneracy pressure

Fowler resolved the paradox within a year of the new statistics [537, 539]: the electrons of a white dwarf form a degenerate Fermi gas in the sense of Chapter 120, and the pressure of that gas at zero temperature is what carries the star [541]. The two limits of that pressure — the non-relativistic $P \propto n^{5/3}$ and the ultrarelativistic $P \propto n^{4/3}$ — control everything in this chapter, so we derive them in full.

Proposition 49.2 (Ground-state pressure of the ideal Fermi gas)**.** *A gas of spin-$\frac{1}{2}$ fermions of mass m at number density n and zero temperature fills all momentum states up to the Fermi momentum*

$$p_\mathrm{F} = \hbar \left(3\pi^2 n\right)^{1/3}\,, \tag{49.1}$$

and exerts a pressure that survives at zero temperature. In the non-relativistic regime $p_\mathrm{F} \ll mc$,

$$P = \frac{\left(3\pi^2\right)^{2/3}}{5} \frac{\hbar^2}{m}\, n^{5/3} = \frac{2}{5}\, n E_\mathrm{F}\,, \qquad E_\mathrm{F} = \frac{\hbar^2}{2m} \left(3\pi^2 n\right)^{2/3}\,, \tag{49.2}$$

while in the ultrarelativistic regime $p_\mathrm{F} \gg mc$,

$$P = \frac{\left(3\pi^2\right)^{1/3}}{4} \hbar c\, n^{4/3} = \frac{1}{4}\, n E_\mathrm{F}\,, \qquad E_\mathrm{F} = \hbar c \left(3\pi^2 n\right)^{1/3}\,. \tag{49.3}$$

Rests on Phenomenon 83.1 and Theorem 40.5.

Proof. *Derives Proposition 49.2.*

Free-particle states occupy phase space in cells of volume $(2\pi\hbar)^3$, two spin orientations per cell (Chapters 83 and 120). At zero temperature the exclusion principle fills the states of lowest energy: a sphere in momentum space of radius p_F fixed by counting,

$$n = \frac{2}{(2\pi\hbar)^3} \cdot \frac{4\pi}{3} p_\mathrm{F}^3 = \frac{p_\mathrm{F}^3}{3\pi^2\hbar^3}\,,$$

which is Equation (49.1). The energy density is the sum of single-particle energies $\varepsilon(p)$ over the filled sphere,

$$u = \frac{2}{(2\pi\hbar)^3}\int_0^{p_\mathrm{F}} \varepsilon(p)\, 4\pi p^2\,\mathrm{d}p = \frac{1}{\pi^2\hbar^3}\int_0^{p_\mathrm{F}} \varepsilon(p)\, p^2\,\mathrm{d}p\,.$$

With $\varepsilon = p^2/2m$ this gives $u = (3/5)\,nE_\mathrm{F}$ with $E_\mathrm{F} = p_\mathrm{F}^2/2m$; with $\varepsilon = pc$ it gives $u = (3/4)\,nE_\mathrm{F}$ with $E_\mathrm{F} = p_\mathrm{F}c$.

The pressure follows thermodynamically, with no kinetic argument needed. At zero temperature the total energy of N particles in volume V is $E = Vu(N/V)$, and

$$P = -\left.\frac{\partial E}{\partial V}\right|_N = -u + n\,\frac{\mathrm{d}u}{\mathrm{d}n}\,.$$

For any power law $u \propto n^\gamma$ this is $P = (\gamma - 1)\,u$. Non-relativistically $\gamma = 5/3$, so $P = \frac{2}{3}u = \frac{2}{5}nE_\mathrm{F}$, which is Equation (49.2); ultrarelativistically $\gamma = 4/3$, so $P = \frac{1}{3}u = \frac{1}{4}nE_\mathrm{F}$, which is Equation (49.3). Substituting Equation (49.1) gives the explicit coefficients. Note what happened between the two limits: relativity does not remove the pressure, it *softens* it, from $n^{5/3}$ to $n^{4/3}$ — one power of $n^{1/3}$, lost because the velocity saturates at c and can no longer grow with the Fermi momentum. That lost third of a power is the whole story of the Chandrasekhar limit. □

In a white dwarf the electrons supply the pressure and the nuclei the mass. For a composition with mean molecular weight per electron μ_e — the number of nucleons per electron, equal to 2 for the helium, carbon and oxygen that white dwarfs are made of — the mass density is $\rho = \mu_\mathrm{e} m_\mathrm{u} n_\mathrm{e}$, and Equation (49.2) becomes a polytrope $P = K_{5/3}\,\rho^{5/3}$ with

$$K_{5/3} = \frac{(3\pi^2)^{2/3}}{5}\,\frac{\hbar^2}{m_\mathrm{e}\,(\mu_\mathrm{e} m_\mathrm{u})^{5/3}}\,. \tag{49.4}$$

For $\mu_\mathrm{e} = 2$ this evaluates (computed from the CODATA constants) to a pressure of $3.2 \times 10^{21}\,\mathrm{Pa}$ at the Sirius B density $\rho = 10^9\,\mathrm{kg/m^3}$ — some ten orders of magnitude beyond the pressure at the centre of the Earth, from a gas at zero temperature. A star supported by a polytropic pressure is an exactly solvable structure problem, and the machinery deserves stating once, in the form every later section will use.

Proposition 49.3 (Lane–Emden form of hydrostatic equilibrium; restated from the stellar-structure chapter)**.** *A self-gravitating Newtonian fluid sphere with the polytropic equation of state $P = K\rho^{1+1/n}$ has density profile $\rho(r) = \rho_\mathrm{c}\,\theta^n(\xi)$, where $r = a\xi$,*

$$a^2 = \frac{(n+1)\,K}{4\pi G}\,\rho_\mathrm{c}^{\frac{1}{n}-1}\,, \tag{49.5}$$

and θ solves the Lane–Emden equation

$$\frac{1}{\xi^2}\frac{\mathrm{d}}{\mathrm{d}\xi}\left(\xi^2\frac{\mathrm{d}\theta}{\mathrm{d}\xi}\right) = -\theta^n\,, \qquad \theta(0) = 1\,, \quad \theta'(0) = 0\,. \tag{49.6}$$

The surface is the first zero ξ_1 of θ, and the radius and total mass are

$$R = a\,\xi_1\,, \qquad M = 4\pi a^3 \rho_\mathrm{c}\,\xi_1^2\,\left|\theta'(\xi_1)\right|\,. \tag{49.7}$$

Rests on Equation (31.13), Theorem 52.6, and Proposition 52.1.

Proof. *Derives Proposition 49.3.*

This is Theorem 52.6 in the variables of this chapter; the short derivation is repeated so that the Chandrasekhar argument can lean on the surface-slope mass formula Equation (49.7) without a forward excursion. Hydrostatic equilibrium (Proposition 31.16, specialized to the self-gravitating sphere as in Section 52.1.1) reads $\mathrm{d}P/\mathrm{d}r = -Gm\rho/r^2$ with $\mathrm{d}m/\mathrm{d}r = 4\pi r^2\rho$. Eliminating m by differentiating the first and substituting the second,

$$\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r^2}{\rho}\frac{\mathrm{d}P}{\mathrm{d}r}\right) = -4\pi G\rho\,.$$

Insert $P = K\rho^{1+1/n}$ and $\rho = \rho_\mathrm{c}\theta^n$: the left side becomes $\left[(n+1)\,K\rho_\mathrm{c}^{1/n}/r^2\right]\mathrm{d}\left(r^2\,\mathrm{d}\theta/\mathrm{d}r\right)/\mathrm{d}r$, and dividing by $4\pi G\rho_\mathrm{c}$ and rescaling $r = a\xi$ with Equation (49.5) leaves exactly Equation (49.6). The centre is regular ($\theta'(0) = 0$) because $m \propto r^3$ there. For the mass, use the equation itself backwards:

$$M = 4\pi a^3\rho_\mathrm{c}\int_0^{\xi_1}\theta^n\xi^2\,\mathrm{d}\xi = -4\pi a^3\rho_\mathrm{c}\int_0^{\xi_1}\frac{\mathrm{d}}{\mathrm{d}\xi}\left(\xi^2\frac{\mathrm{d}\theta}{\mathrm{d}\xi}\right)\mathrm{d}\xi = 4\pi a^3\rho_\mathrm{c}\,\xi_1^2\,|\theta'(\xi_1)|\,,$$

since $\theta'(\xi_1) < 0$. No integration of the density profile is ever needed: the surface slope carries the whole mass integral. □

Remark 49.4 (The two Lane–Emden constants this chapter needs). Equation (49.6) has closed-form solutions only for $n = 0, 1, 5$; for the two indices that matter here it is integrated numerically, and we quote the constants from the standard tabulation [831]:

n	ξ_1	$\xi_1^2\,\lvert\theta'(\xi_1)\rvert$
3/2	3.653 75	2.714 06
3	6.896 85	2.018 24

The general theory of polytropes, including these integrations, belongs to Section 52.1.2.

Proposition 49.5 (White-dwarf mass–radius relation). *A non-relativistic white dwarf ($P = K_{5/3}\rho^{5/3}$, polytropic index $n = 3/2$) obeys*

$$R \propto M^{-1/3}\,, \qquad R \approx 8.9\times10^6\,\mathrm{m}\left(\frac{M}{M_\odot}\right)^{-1/3} \quad \text{for } \mu_\mathrm{e} = 2\,. \tag{49.8}$$

Adding mass shrinks *a white dwarf — the opposite of every object held up by thermal or material pressure.*

Rests on Equations (49.2) and (49.7).

Proof. *Derives Proposition 49.5.*

For $n = 3/2$, Equation (49.5) gives $a \propto \rho_\mathrm{c}^{-1/6}$, so Equation (49.7) gives $M \propto a^3\rho_\mathrm{c} \propto \rho_\mathrm{c}^{1/2}$ and $R \propto \rho_\mathrm{c}^{-1/6}$: eliminating the central density, $R \propto M^{-1/3}$. Restoring the constants (Equation (49.4) and Remark 49.4),

$$R = \xi_1\left(\frac{5K_{5/3}}{8\pi G}\right)^{1/2}\rho_\mathrm{c}^{-1/6}\,, \qquad M = 4\pi\,\xi_1^2\,|\theta'(\xi_1)|\left(\frac{5K_{5/3}}{8\pi G}\right)^{3/2}\rho_\mathrm{c}^{1/2}\,,$$

and eliminating ρ_c between them gives, for $\mu_\mathrm{e} = 2$, the figure quoted (computed here from the constants; e.g. $M = 0.6\,M_\odot$ gives $\rho_\mathrm{c} \approx 1.5\times10^9\,\mathrm{kg/m^3}$ and $R \approx 1.05\times10^7\,\mathrm{m}$ — precisely the density regime the Sirius B puzzle demanded). The formula overestimates the radius of the more massive white dwarfs, whose electrons are already partly relativistic and whose true radius therefore falls faster than $M^{-1/3}$; the exact interpolating equation of state joins this curve to the terminating sequence of Section 49.1.3. □

49.1.3 The Chandrasekhar mass limit

The inverse mass–radius relation carries its own death sentence. Compression raises the Fermi momentum, and once $p_{\mathrm{F}} \gtrsim m_{\mathrm{e}}c$ the equation of state softens from $\rho^{5/3}$ to the ultrarelativistic $\rho^{4/3}$ of Equation (49.3) — the polytropic index reaches $n = 3$, and at $n = 3$ the Lane–Emden scaling produces a mass that does not depend on the central density at all. That mass is a *limit*: below it equilibria exist, above it none do [676]. Landau's contemporaneous argument [702] reaches the same scale from the energy of a relativistic degenerate gas, with no structure equation solved, and we give it first.

Phenomenon 49.6 (A maximum mass for a cold degenerate star)**.** No white dwarf has been measured above about $1.4\,M_{\odot}$. The observed mass–radius relation does not continue indefinitely: radius falls as mass rises and the sequence terminates [676]. The same ceiling is what makes thermonuclear supernovae reproducible. A white dwarf pushed across it by accretion from a companion has no equilibrium left to settle into, ignites its carbon degenerately, and releases a nearly fixed amount of energy — which is why type Ia events can serve as the standardizable candles of Chapter 48.

Rests on Proposition 49.2 and Theorems 44.27 and 49.7.

Derivation. *Derives Phenomenon 49.6.*

The existence of the limit, and its scale, follow from comparing two energies, with no structure equation solved [702]. Let the star hold N electrons within a radius R, with μ_{e} nucleons per electron, so that its mass is $M = N\mu_{\mathrm{e}}m_{\mathrm{u}}$ and the electron number density is $n \sim N/R^3$.

Once the electrons are relativistic — the regime Phenomenon 49.1 showed to be near — the Fermi energy is $E_{\mathrm{F}} \sim \hbar c\, n^{1/3} \sim \hbar c\, N^{1/3}/R$, so the kinetic energy of the gas is

$$E_{\mathrm{kin}} \sim N E_{\mathrm{F}} \sim \frac{\hbar c\, N^{4/3}}{R}\,,$$

while the gravitational energy is

$$E_{\mathrm{grav}} \sim -\frac{GM^2}{R} = -\frac{G\,(\mu_{\mathrm{e}}m_{\mathrm{u}})^2\,N^2}{R}\,.$$

Both vary as $1/R$, and that is the decisive fact: contracting the star changes the two terms in exactly the same proportion, so no radius is singled out and equilibrium cannot be reached by adjusting R. The sign of the total energy depends on N alone. It turns negative — unbounded contraction, with no configuration to stop at — once

$$G\,(\mu_{\mathrm{e}}m_{\mathrm{u}})^2\,N^2 > \hbar c\, N^{4/3}\,, \qquad \text{that is} \qquad N > N_{\max} \sim \left(\frac{\hbar c}{G}\right)^{3/2} \frac{1}{(\mu_{\mathrm{e}}m_{\mathrm{u}})^3}\,.$$

The corresponding mass is

$$M_{\max} \sim N_{\max}\mu_{\mathrm{e}}m_{\mathrm{u}} = \frac{1}{\mu_{\mathrm{e}}^2 m_{\mathrm{u}}^2}\left(\frac{\hbar c}{G}\right)^{3/2}\,, \tag{49.9}$$

which for $\mu_{\mathrm{e}} = 2$ evaluates to $9.3 \times 10^{29}\,\mathrm{kg}$, about $0.47\,M_{\odot}$. What the argument establishes without any coefficient at all is the substance of the phenomenon: a limit exists, it does not depend on the radius, and it is assembled out of $\hbar$, c, G and the nucleon mass — a stellar mass written in fundamental constants. The exact coefficient, close to three, is supplied by the $n = 3$ polytrope in Theorem 49.7 below. □

Theorem 49.7 (The Chandrasekhar mass)**.** *For the ultrarelativistic equation of state* $P = K_{4/3}\,\rho^{4/3}$ *with*

$$K_{4/3} = \frac{(3\pi^2)^{1/3}}{4}\,\frac{\hbar c}{(\mu_{\mathrm{e}} m_{\mathrm{u}})^{4/3}}\,, \tag{49.10}$$

the polytropic mass Equation (49.7) *is independent of the central density and equals*

$$M_{\mathrm{Ch}} = 4\pi\,\xi_1^2\,|\theta'(\xi_1)|\left(\frac{K_{4/3}}{\pi G}\right)^{3/2} = \frac{5.82}{\mu_{\mathrm{e}}^2}\,M_\odot = 1.46\,M_\odot \quad \textit{for } \mu_{\mathrm{e}} = 2\,. \tag{49.11}$$

No cold star supported by electron degeneracy can exceed it.
Rests on Equations (49.3) *and* (49.7) *and Propositions 49.2 and 49.3.*

Proof. *Derives Theorem 49.7.*

$P = K\rho^{4/3}$ is the polytrope $n = 3$. Equation (49.5) then gives $a = (K/\pi G)^{1/2}\,\rho_{\mathrm{c}}^{-1/3}$, so in Equation (49.7) the central density cancels identically:

$$M = 4\pi a^3 \rho_{\mathrm{c}}\,\xi_1^2\,|\theta'(\xi_1)| = 4\pi\left(\frac{K}{\pi G}\right)^{3/2}\xi_1^2\,|\theta'(\xi_1)|\,.$$

The $n = 3$ sequence is therefore a sequence of one mass at every central density: compressing the star ($R = a\xi_1 \propto \rho_{\mathrm{c}}^{-1/3} \to 0$) changes nothing in the balance, which is the structural restatement of the $1/R$ degeneracy in Landau's argument. A star lighter than M has a subrelativistic core, a stiffer effective $n < 3$, and finds an equilibrium radius; a star heavier has none. Inserting Equation (49.10) with $\mu_{\mathrm{e}} = 2$ and $\xi_1^2\,|\theta'(\xi_1)| = 2.018\,24$ (Remark 49.4),

$$M_{\mathrm{Ch}} = 2.90 \times 10^{30}\,\mathrm{kg} = 1.46\,M_\odot\,,$$

computed here from the CODATA constants; it exceeds the dimensional estimate Equation (49.9) by the factor $\frac{1}{2}\sqrt{3\pi}\,\xi_1^2\,|\theta'(\xi_1)| \approx 3.1$. Chandrasekhar's 1931 paper [676] quotes $0.91\,M_\odot$ because it took $\mu_{\mathrm{e}} = 2.5$, the then-standard guess at stellar composition; the formula is the same, and with the modern $\mu_{\mathrm{e}} = 2$ of helium, carbon and oxygen it gives the value above. General-relativistic corrections and electron capture shave a few per cent off the Newtonian figure, which is why the limit is customarily quoted as $\approx 1.4\,M_\odot$ [831, 1727]. □

49.1.4 White-dwarf cooling and ages

A white dwarf has no fuel: it shines by leaking the thermal energy of its ions through a thin insulating envelope, and its luminosity is therefore a clock. Mestel's theory [1007] turns the statement into a power law, and the population of the Galactic disc into a datable sample.

Phenomenon 49.8 (The white-dwarf luminosity function has an edge)**.** White dwarfs are observed at every luminosity from about $10^{-1}L_\odot$ down to about $10^{-4.5}L_\odot$ — and then stop. The number of white dwarfs per unit luminosity rises steadily toward the faint end and falls off abruptly below $\log_{10}(L/L_\odot) \approx -4.5$ [1801]. Fainter, older white dwarfs do not exist to be seen: the cutoff is the age of the population. Read through cooling theory it dates the local Galactic disc at about (9 ± 2) Gyr [1801], consistent with every other chronometer of Chapter 48. In the cooling sequence itself, *Gaia* photometry of over ten thousand white dwarfs shows a pile-up of stars at the luminosity where the core crystallises — the latent heat of freezing stalls the cooling, and stars accumulate there like water at its freezing point [2341].
Rests on Propositions 31.16, 49.2 and 52.9.

Derivation. *Derives Phenomenon 49.8.*

Two facts organize the problem. First, the degenerate electrons conduct heat so well that the interior is nearly isothermal at a single T_{c}, and their heat capacity is suppressed by the degeneracy factor T/T_{F} (Chapter 120), so the heat reservoir is the *ions*:

$$U = \frac{3}{2}\,\frac{M}{A m_{\mathrm{u}}}\,k_{\mathrm{B}}T_{\mathrm{c}}\,,$$

with A the mean nucleon number ($A \approx 14$ for a carbon–oxygen mixture). Second, the leak is a thin non-degenerate envelope in radiative equilibrium: photons diffuse with mean free path $1/(\kappa\rho)$, giving the radiative-transfer relation

$$L = -\frac{16\pi a c r^2 T^3}{3\kappa\rho}\,\frac{\mathrm{d}T}{\mathrm{d}r}\,,$$

with a the radiation constant and $\kappa = \kappa_0 \rho T^{-7/2}$ the Kramers opacity of Chapter 52. Dividing hydrostatic equilibrium (Equation (31.13), with $m \approx M$ in the thin envelope) by that relation eliminates r, and the ideal gas law of the envelope, $\rho = \mu m_{\mathrm{u}} P/(k_{\mathrm{B}}T)$ with μ its mean molecular weight, closes the system:

$$\frac{\mathrm{d}P}{\mathrm{d}T} = \frac{16\pi a c G M}{3\kappa L}\,T^3 = \frac{16\pi a c G M k_{\mathrm{B}}}{3\kappa_0 \mu m_{\mathrm{u}} L}\,\frac{T^{15/2}}{P} \quad\Longrightarrow\quad P^2 \propto \frac{M}{L}\,T^{17/2} \quad\Longrightarrow\quad \rho^2 \propto \frac{M}{L}\,T^{13/2}\,.$$

The envelope ends where the electrons turn degenerate, $k_{\mathrm{B}}T \approx E_{\mathrm{F}} \propto \rho^{2/3}$, i.e. $\rho \propto T^{3/2}$; and because the degenerate interior is isothermal, the temperature there is T_{c} itself. Setting $\rho^2 \propto T_{\mathrm{c}}^3$ equal to $(M/L)\,T_{\mathrm{c}}^{13/2}$ gives Mestel's law

$$L \propto M\,T_{\mathrm{c}}^{7/2}\,. \tag{49.12}$$

Now the clock: $-\mathrm{d}U/\mathrm{d}t = L$ reads $T_{\mathrm{c}}^{-7/2}\,\mathrm{d}T_{\mathrm{c}} \propto -\mathrm{d}t$, so $T_{\mathrm{c}} \propto t^{-2/5}$ and

$$L \propto t^{-7/5}\,, \qquad t = \frac{2}{5}\,\frac{U}{L} \propto \left(\frac{L}{M}\right)^{-5/7}\,. \tag{49.13}$$

Cooling decelerates: each decade down in luminosity takes about $10^{5/7} \approx 5$ times longer than the last, which is why the luminosity function *rises* toward the faint end — stars pile up where they cool slowly — and why its termination is sharp. The absolute age at the cutoff requires the calibrated constant in Equation (49.12) together with the crystallisation and phase-separation corrections, and we quote the result rather than the numerics: $(9 \pm 2)\,\mathrm{Gyr}$ for the local disc [1801]. Crystallisation itself was predicted theoretically in the 1960s [1386]; its latent heat, $\sim k_{\mathrm{B}}T$ per ion, interrupts Equation (49.13) for of order a gigayear, and the resulting traffic jam in the cooling sequence is what *Gaia* resolved [2341]. □

49.2 Neutron stars

49.2.1 The Baade–Zwicky hypothesis

Within two years of the discovery of the neutron [692], Baade and Zwicky proposed that a supernova is the transition of an ordinary star into a star made of neutrons — packed at nuclear density, bound by gravity, and only a few kilometres across — and that the energy of the explosion is the gravitational binding energy released in the collapse [730]. The proposal was quantitative and, as Phenomenon 49.14 will show, correct in every part. Collapsing $1.4\,M_{\odot}$ to a radius of 12 km releases $\sim \frac{3}{5}GM^2/R \approx 2.6 \times 10^{46}\,\mathrm{J}$ (Equation (49.23) below) — about ten per cent of the star's rest energy, and four orders of magnitude more than the $10^{42}\,\mathrm{J}$ a supernova radiates as light. The visible fireworks are a rounding error on the energy budget; the budget itself was verified, neutrino by neutrino, in 1987. The prediction of the compact remnant waited thirty-four years for its confirmation: a pulsar sitting inside the Crab supernova remnant of AD 1054 [1384].

49.2.2 The Tolman–Oppenheimer–Volkoff equations

A neutron star is relativistic in a way no white dwarf is: its compactness $GM/\left(Rc^2\right)$ is of order 0.2 rather than 10^{-4}, so Newtonian hydrostatics (Proposition 31.16 and Section 52.1.1) is not a first approximation but a wrong theory. The correct structure equations follow from the field equations of Chapter 44,

$$G_{\mu\nu} \equiv R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^4}\, T_{\mu\nu}\,, \tag{49.14}$$

applied to a static, spherically symmetric perfect fluid [845, 848]. This is the chapter's central derivation, and we give it in full.

The most general static, spherically symmetric line element can be written

$$\mathrm{d}s^2 = -\mathrm{e}^{2\alpha(r)} c^2 \mathrm{d}t^2 + \mathrm{e}^{2\beta(r)} \mathrm{d}r^2 + r^2 \left(\mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\varphi^2\right)\,, \tag{49.15}$$

and the matter is a perfect fluid at rest in these coordinates,

$$T^{\mu\nu} = \left(\rho + \frac{P}{c^2}\right) u^\mu u^\nu + P\, g^{\mu\nu}\,, \qquad u^\mu = \left(\mathrm{e}^{-\alpha}, 0, 0, 0\right) \ \text{in}\ (t, r, \theta, \varphi)\,, \tag{49.16}$$

with $\rho(r)$ the mass density, $P(r)$ the pressure, and the normalization $u^\mu u_\mu = g_{tt}\,\mathrm{e}^{-2\alpha} = -c^2$ (Section 44.2.1). In mixed components $T^\mu{}_\nu = \mathrm{diag}\left(-\rho c^2, P, P, P\right)$.

Theorem 49.9 (Tolman–Oppenheimer–Volkoff equations)**.** *Define the mass function $m(r)$ by $\mathrm{e}^{-2\beta} = 1 - 2Gm/\left(rc^2\right)$. Then the field equations Equation (49.14) for Equations (49.15) and (49.16) reduce to*

$$\frac{\mathrm{d}m}{\mathrm{d}r} = 4\pi r^2 \rho\,, \tag{49.17}$$

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\,\frac{G\left(\rho + \dfrac{P}{c^2}\right)\left(m + \dfrac{4\pi r^3 P}{c^2}\right)}{r^2\left(1 - \dfrac{2Gm}{rc^2}\right)}\,, \tag{49.18}$$

$$\frac{\mathrm{d}\alpha}{\mathrm{d}r} = -\,\frac{1}{\rho c^2 + P}\,\frac{\mathrm{d}P}{\mathrm{d}r}\,. \tag{49.19}$$

At the surface $r = R$, where $P = 0$, the solution matches the exterior Schwarzschild geometry of Chapter 45 with total mass $M = m(R)$.

Rests on Equations (49.14) and (49.16), Postulate 44.3, and Definition 44.16.

Proof. *Derives Theorem 49.9.*

The curvature of the metric Equation (49.15) is computed, Christoffel symbols to Ricci tensor, in Chapter 45, where the vacuum case yields the Schwarzschild solution; the two Einstein-tensor components needed here are [848, 1509]

$$G^t{}_t = -\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left[r\left(1 - \mathrm{e}^{-2\beta}\right)\right]\,,$$
$$G^r{}_r = -\frac{1}{r^2}\left(1 - \mathrm{e}^{-2\beta}\right) + \frac{2}{r}\,\mathrm{e}^{-2\beta}\,\alpha'\,,$$

with the prime denoting $\mathrm{d}/\mathrm{d}r$. The angular components carry no new information: by the contracted Bianchi identity (Section 44.3) they are consequences of these two together with stress–energy conservation, which we impose explicitly below.

The tt equation. With $T^t{}_t = -\rho c^2$, Equation (49.14) reads $\mathrm{d}\left[r\left(1 - \mathrm{e}^{-2\beta}\right)\right]/\mathrm{d}r = \left(8\pi G/c^2\right)\rho r^2$. Substituting the definition $r\left(1 - \mathrm{e}^{-2\beta}\right) = 2Gm/c^2$ turns this into Equation (49.17). Regularity at the centre requires $m(0) = 0$, so $m(r)$ is fixed with no further

constant. Note what m is and is not: $M = m(R)$ is the mass that governs the exterior orbits — the mass a distant satellite weighs — but it is *not* the integral of ρ over proper volume, which carries an extra factor $\mathrm{e}^{\beta} > 1$. The deficit is the gravitational binding energy, already subtracted; general relativity weighs the star bound.

The rr equation. With $T^{r}{}_{r} = P$ and $1 - \mathrm{e}^{-2\beta} = 2Gm/(rc^2)$,

$$\frac{8\pi G}{c^4}P = -\frac{2Gm}{r^3c^2} + \frac{2}{r}\left(1 - \frac{2Gm}{rc^2}\right)\alpha' \quad\Longrightarrow\quad \alpha' = \frac{\dfrac{Gm}{c^2} + \dfrac{4\pi G r^3 P}{c^4}}{r^2\left(1 - \dfrac{2Gm}{rc^2}\right)}\,.$$

Conservation. The fluid obeys $\nabla_\mu T^{\mu}{}_{\nu} = 0$ (Section 44.2.2); take the $\nu = r$ component. For the static diagonal stress tensor,

$$\nabla_\mu T^{\mu}{}_{r} = \partial_r T^{r}{}_{r} + \Gamma^{\mu}{}_{\mu\lambda}T^{\lambda}{}_{r} - \Gamma^{\lambda}{}_{\mu r}T^{\mu}{}_{\lambda} = P' + \Gamma^{t}{}_{tr}\left(P + \rho c^2\right) = P' + \left(\rho c^2 + P\right)\alpha'\,,$$

because $\Gamma^{t}{}_{tr} = \alpha'$ while the spatial Christoffel terms multiply $P - P = 0$. Setting this to zero gives Equation (49.19): pressure gradients are what hold the fluid static against the gravitational potential α. Substituting α' from the rr equation yields Equation (49.18).

Matching. At $r = R$ the pressure vanishes and Equation (49.17) makes m constant; the metric functions become $\mathrm{e}^{-2\beta} = 1 - 2GM/(rc^2)$ and, by integrating α' in vacuum, $\mathrm{e}^{2\alpha} = 1 - 2GM/(rc^2)$ — the Schwarzschild exterior of Chapter 45, as it must be by Birkhoff's theorem there. Given an equation of state $P(\rho)$, the system Equations (49.17) and (49.18) integrates outward from a chosen central density ρ_{c} until P reaches zero: one equilibrium star per central density, and the mass–radius *curve* is the image of that one-parameter family. □

Remark 49.10 (Newtonian limit, and why relativity lowers the ceiling). When $P \ll \rho c^2$ and $2Gm \ll rc^2$, Equation (49.18) reduces to $\mathrm{d}P/\mathrm{d}r = -Gm\rho/r^2$ — hydrostatic equilibrium exactly as in Proposition 31.16 and Section 52.1.1. The three relativistic corrections are the three parenthesized factors, and each has a sign: pressure adds to the gravitating density $(\rho + P/c^2)$, pressure gravitates as a source $(m + 4\pi r^3 P/c^2)$, and the denominator $1 - 2Gm/(rc^2)$ strengthens gravity as compactness grows. All three make gravity *stronger* than Newton at the same P and ρ. Pressure, the agent of support, is also a source of the field it fights — so squeezing harder eventually recruits gravity faster than support, and equilibrium sequences terminate at finite mass even for matter far stiffer than a degenerate electron gas. The Chandrasekhar limit was an accident of special relativity in the equation of state; the neutron-star maximum mass is general relativity's own.

Example 49.11 (The uniform-density star)**.** For constant ρ, $m = \frac{4}{3}\pi r^3\rho = Mr^3/R^3$ and Equation (49.18) separates; partial fractions and the boundary condition $P(R) = 0$ give (verified by substitution)

$$P(r) = \rho c^2\,\frac{\sqrt{1 - \dfrac{r_{\mathrm{s}}r^2}{R^3}} - \sqrt{1 - \dfrac{r_{\mathrm{s}}}{R}}}{3\sqrt{1 - \dfrac{r_{\mathrm{s}}}{R}} - \sqrt{1 - \dfrac{r_{\mathrm{s}}r^2}{R^3}}}\,, \qquad r_{\mathrm{s}} \equiv \frac{2GM}{c^2}\,, \tag{49.20}$$

with central value

$$P_{\mathrm{c}} = \rho c^2\,\frac{1 - \sqrt{1 - r_{\mathrm{s}}/R}}{3\sqrt{1 - r_{\mathrm{s}}/R} - 1}\,. \tag{49.21}$$

For a star of $1.4\,M_\odot$ and 12 km ($r_{\mathrm{s}}/R = 0.35$) this gives $P_{\mathrm{c}} = 0.13\,\rho c^2$: the pressure at the centre of a neutron star is a sizeable fraction of its energy density, which no Newtonian star ever approaches.

Rests on Equation (49.18) and Theorem 49.9.

Proposition 49.12 (Buchdahl bound)**.** *No static spherical star of any composition can be more compact than*

$$\frac{GM}{Rc^2} < \frac{4}{9}, \tag{49.22}$$

equivalently $R > \frac{9}{8} r_{\mathrm{s}}$*: equilibrium fails strictly* outside *the would-be horizon at* r_{s}*.*
Rests on Equation (49.21) *and Example 49.11.*

Proof. *Derives Proposition 49.12.*

For the uniform star the bound is read off Equation (49.21): the central pressure diverges when the denominator vanishes, $3\sqrt{1 - r_{\mathrm{s}}/R} = 1$, i.e. at $r_{\mathrm{s}}/R = 8/9$, and no equilibrium exists at or beyond that compactness — infinite pressure is not available, and would only gravitate harder if it were (Remark 49.10). Buchdahl proved that the uniform star is the extreme case: for *any* equation of state whose density does not increase outward, integrating the field equations from centre to surface bounds the central metric function away from zero only when Equation (49.22) holds [1167]. We use the theorem in Buchdahl's generality but exhibit the proof only in the uniform case, which saturates it. Two consequences are observable: the gravitational redshift from the surface of any static star is bounded by $1 + z = (1 - r_{\mathrm{s}}/R)^{-1/2} < 3$, i.e. $z < 2$; and any object more compact than $\frac{9}{8} r_{\mathrm{s}}$ cannot be a star in equilibrium at all — the operational dividing line between the objects of this chapter and those of Chapter 45. □

Remark 49.13 (The Oppenheimer–Volkoff mass). Oppenheimer and Volkoff integrated Equations (49.17) and (49.18) numerically for the coldest matter then computable — an ideal degenerate neutron gas, i.e. Proposition 49.2 with $m = m_{\mathrm{n}}$ — and found the equilibrium sequence to terminate at about $0.7\,M_\odot$, at a radius near 10 km [845]. The number is far below every measured neutron-star mass (Section 49.4.1), and the discrepancy is itself a measurement: at these densities the neutrons are *not* an ideal gas. Their Fermi energy at nuclear saturation density, ~ 60 MeV by Equation (49.2), is comparable to the depth of the nuclear potential, so the repulsive core of the nuclear force stiffens the true equation of state far beyond the ideal-gas $\rho^{5/3}$ and roughly triples the maximum mass (Section 49.2.4). What survives of the calculation is its structure — a maximum mass exists for *every* equation of state, ideal or not — and its method, which is exactly how the modern curves of Section 49.4 are still computed.

49.2.3 Birth: core collapse and supernova neutrinos

A star above about $8\,M_\odot$ ends with an iron core at the Chandrasekhar mass, supported by exactly the electron degeneracy pressure whose limit Theorem 49.7 establishes. When the core exceeds it, collapse is not gradual but catastrophic: electron capture $p + e^- \to n + \nu_e$ removes the supporting electrons as the density rises, softening the equation of state precisely when stiffness is needed, and the core falls freely until nuclear density stiffens it (Section 49.2.4). The rebound drives a shock into the infalling envelope; the shock stalls, and the current understanding — anticipated by Colgate and White [1328] — is that it is revived by the absorption of a small fraction of the outgoing neutrinos, which carry essentially the whole energy release. The nuclear physics of the burning stages belongs to Chapter 107 and Chapter 52; what belongs here is the gravitational energy budget, and its direct detection.

Phenomenon 49.14 (A neutron star is observed being born)**.** Hours before the optical brightening of SN1987A in the Large Magellanic Cloud, two underground detectors on opposite sides of the Earth recorded a coincident burst of neutrino events: eleven at Kamiokande-II [1793] and eight at IMB [1789], spread over about ten seconds, at energies of tens of MeV. Scaled to the distance of the source, the burst carries of order 10^{46} J — some four orders of magnitude more than the light of the whole supernova — which is the

gravitational binding energy of a neutron star. This is the direct detection of stellar-core collapse, thirty-four years after the proposal that a supernova marks the transition of an ordinary star into a neutron star [730, 1328].

Rests on Theorem 44.27 and Phenomenon 101.5.

Derivation. *Derives Phenomenon 49.14.*

Three numbers are derived: the energy, its carrier, and its time-scale. Collapse carries a stellar core of mass M from a radius of order $10^6\,\mathrm{m}$ down to the radius of a neutron star, and the energy released is the difference of gravitational binding energies, which for a uniform sphere is

$$\Delta E \simeq \frac{3}{5}\frac{GM^2}{R}\,, \tag{49.23}$$

the initial term being negligible because the initial radius is two orders of magnitude larger. With $M = 1.4\,M_\odot = 2.8\times10^{30}\,\mathrm{kg}$ and $R = 12\,\mathrm{km}$,

$$\Delta E \simeq \frac{3}{5}\cdot\frac{6.674\times10^{-11}\,\mathrm{m^3/(kg\,s^2)}\times(2.8\times10^{30}\,\mathrm{kg})^2}{1.2\times10^{4}\,\mathrm{m}} \approx 2.6\times10^{46}\,\mathrm{J}\,,$$

about a tenth of the rest energy Mc^2. Set this beside the two things a supernova is usually described by: the kinetic energy of the ejecta, of order $10^{44}\,\mathrm{J}$, and the radiated light, of order $10^{42}\,\mathrm{J}$. The visible supernova is a by-product at the level of a per cent and a hundredth of a per cent respectively; the event itself is a neutrino event, and that is what the SN1987A detections confirmed.

That the energy emerges as neutrinos, over seconds rather than milliseconds, is a statement about opacity. Photons are hopelessly trapped, and even neutrinos are trapped at nuclear density: with the weak-interaction cross section of Chapter 103, $\sigma \sim 1.7\times10^{-48}\,\mathrm{m^2}\,(E_\nu/m_\mathrm{e}c^2)^2$, a 30 MeV neutrino in matter of $n \approx 1.8\times10^{44}/\mathrm{m^3}$ has a mean free path $\lambda = 1/(n\sigma)$ of order one *metre* — against a stellar radius of $10^4\,\mathrm{m}$. The energy therefore leaves by a random walk, in a diffusion time $t \sim R^2/(\lambda c)$, which with these crude numbers is already of order a second; detailed transport, with the cross sections rising steeply toward the hot centre, lengthens this to the observed ten seconds [1328, 1727]. The neutrinos finally escape from a layer where the temperature corresponds to mean energies of order 10 MeV — as observed.

The observed event counts check the budget quantitatively. Spreading ΔE over six neutrino species at $\langle E_\nu\rangle \approx 12\,\mathrm{MeV}$ gives $\sim 10^{58}$ neutrinos; at the LMC distance of $50\,\mathrm{kpc} = 1.5\times10^{21}\,\mathrm{m}$ the fluence at Earth is $\sim 5\times10^{14}/\mathrm{m^2}$ (computed). Kamiokande-II's fiducial $2.14\times10^6\,\mathrm{kg}$ of water holds about 1.4×10^{32} free protons; folding the inverse-beta-decay cross section $\bar\nu_e p \to n e^+$ at the mean energy, $\sigma \sim 1.5\times10^{-45}\,\mathrm{m^2}$, with the $\bar\nu_e$ sixth of the fluence predicts of order ten to twenty detected events — and eleven were seen [1793], with eight more in the larger, higher-threshold IMB tank [1789]. An order-of-magnitude estimate this crude agreeing this well means the energy, the carrier and the time-scale are all as Equation (49.23) demands. The bounds those nineteen events place on neutrino mass, lifetime and speed belong to Chapter 103. □

49.2.4 The dense-matter equation of state

Everything in Theorem 49.9 is known physics except one input: the pressure–density relation $P(\rho)$ of cold matter beyond nuclear saturation density, $\rho_0 \approx 2.7\times10^{17}\,\mathrm{kg/m^3}$ ($n_0 \approx 1.6\times10^{44}/\mathrm{m^3}$). Below saturation the ground state of matter is known essentially completely [1441]. A neutron star's *crust* is a lattice of increasingly neutron-rich nuclei in a degenerate electron gas; at $\rho \approx 4.3\times10^{14}\,\mathrm{kg/m^3}$ the nuclei can hold no more and free neutrons "drip" out, so from there inward a neutron gas — itself superfluid, the point of contact with Chapter 130 — coexists with the lattice; near $\rho_0/2$ the nuclei merge into

uniform neutron-rich matter of neutrons, protons and electrons in β-equilibrium. The nuclear physics of these regimes is that of Chapter 107.

Above a few times ρ_0 — the density of the inner core — the composition itself is unknown: hyperons, meson condensates and deconfined quark matter are all candidates, each softening the pressure in its own way [1727]. Two constraints are model-independent. Thermodynamic stability requires $\mathrm{d}P/\mathrm{d}\rho \geq 0$, and causality requires the sound speed $c_\mathrm{s}^2 = \mathrm{d}P/\mathrm{d}\rho$ (here ρ the total mass–energy density over c^2) not to exceed c^2 — the stiffest admissible equation of state is $P = \rho c^2$ up to constants, and Section 49.4.1 turns exactly that remark into a bound on every neutron-star mass.

Remark 49.15 (The equation of state is a mass–radius curve). Theorem 49.9 maps each candidate $P(\rho)$ to a one-parameter family $(M(\rho_\mathrm{c}), R(\rho_\mathrm{c}))$ — a curve in the mass–radius plane — and the map runs both ways: measuring masses and radii selects the equation of state, which is how a telescope does nuclear physics. The qualitative dictionary: a *stiff* equation of state (pressure rising fast) holds up larger maximum masses at larger radii; a *soft* one gives compact, light stars. Configurations beyond the maximum of $M(\rho_\mathrm{c})$ are unstable to radial collapse, so the curve's summit is the maximum mass of Section 49.4.1. The observational programme of Section 49.4 — Shapiro-delay masses, X-ray radii, tidal deformability — is the inversion of this map, and its current verdict is a pressure at $2\rho_0$ known to within about a factor of two [2294, 2338].

49.3 Pulsars

49.3.1 Discovery and identification

In 1967 Jocelyn Bell, searching survey charts for interplanetary scintillation, found a radio source that ticked: pulses of 1.337 s period, stable to the accuracy of the observatory clock [1373]. Within a year the identification was settled in favour of a rotating magnetized neutron star: Pacini had noted, just before the discovery, that a magnetized neutron star spun down by magnetic dipole radiation could power the Crab nebula [1355], and Gold showed that rotation — unlike the vibrations and orbits also proposed — naturally supplies the observed stability, the small duty cycle of a sweeping lighthouse beam, and a period that can only lengthen [1372]. The clincher came from the Crab remnant itself: a 33 ms pulsar [1384], spinning too fast for any white dwarf (Equation (49.24)) and slowing at a rate whose energy output matches the nebula's luminosity (Proposition 49.17).

Phenomenon 49.16 (Pulsars: clocks at nuclear density)**.** Certain radio sources emit pulses lasting a few per cent of their cycle, repeating with a regularity that rivals atomic clocks; the first found had a period of 1.337 s [1373]. The period always lengthens and never shortens, apart from occasional abrupt spin-ups [1406]. The source inside the Crab supernova remnant pulses at 33 ms [1384]; the first millisecond pulsar found spins at 1.558 ms [1681], and the fastest known at 1.396 ms — 716 rotations per second [2095]. The stability of the period, and its monotonic increase, identify the clock as the rotation of a massive body rather than any vibration or orbit [1355, 1372]; and rotation at these rates is possible only for matter at nuclear density.

Rests on Postulate 19.8 and Theorem 44.27.

Derivation. *Derives Phenomenon 49.16.*

A self-gravitating body rotating with angular velocity ω holds together at its equator only if its own gravity supplies at least the centripetal acceleration required there,

$$\frac{GM}{R^2} \geq \omega^2 R\,.$$

Writing $M = \frac{4}{3}\pi R^3 \rho$ for the mean density and $P = 2\pi/\omega$ for the period, the factor R^3 cancels from both sides and there remains

$$\rho \geq \frac{3\pi}{GP^2}\,. \tag{49.24}$$

The radius has dropped out entirely: the observed period bounds the density alone, with no model of the star required. For the Crab pulsar, $P = 33\,\mathrm{ms}$ gives $\rho \geq 1.3 \times 10^{14}\,\mathrm{kg/m^3}$, five orders of magnitude above the white-dwarf density of Phenomenon 49.1; for the 1.396 ms record holder it gives $\rho \geq 7.3 \times 10^{16}\,\mathrm{kg/m^3}$, within a factor of four of nuclear saturation density. No white dwarf can rotate this fast, and no other known state of matter reaches such densities. The bound is crude — it treats the body as uniform and ignores the distortion that rotation itself produces — but it is model-independent, and it is what identified the pulsars with the neutron stars predicted three decades earlier [730, 1372]. □

Proposition 49.17 (Magnetic-dipole spin-down)**.** *A neutron star of moment of inertia I, rotating at $\omega = 2\pi/P$ with a magnetic dipole moment $m_\perp$ perpendicular to the spin axis, loses rotational energy at the rate*

$$\dot{E} = -\frac{\mu_0\, m_\perp^2\, \omega^4}{6\pi c^3}\,, \tag{49.25}$$

so that $\dot{\omega} \propto -\omega^3$, the product $P\dot{P}$ is constant, and

$$\tau_\mathrm{c} \equiv \frac{P}{2\dot{P}} \tag{49.26}$$

bounds the age from above (equality for a star born spinning much faster than today). The surface polar field satisfies

$$B_\mathrm{p} \sin\theta_m = \sqrt{\frac{3\mu_0 c^3 I\, P\dot{P}}{8\pi^3 R^6}}\,, \tag{49.27}$$

with θ_m the magnetic inclination.

Proof. *Derives Proposition 49.17.*

Electromagnetic theory gives, for any magnetic moment $\boldsymbol{m}(t)$ varying in time, the radiated power $\mu_0 \left|\ddot{\boldsymbol{m}}\right|^2 / (6\pi c^3)$ — the magnetic counterpart of the electric-dipole formula, quoted here from its home in Section 65.3.2 [1976]. For a moment of fixed magnitude inclined at θ_m to the rotation axis, the component along the axis is static and the perpendicular component $m_\perp = m \sin\theta_m$ rotates rigidly at ω, so $|\ddot{\boldsymbol{m}}| = \omega^2 m_\perp$ at every instant, giving Equation (49.25). Setting $\dot{E} = \mathrm{d}\left(\frac{1}{2} I\omega^2\right)/\mathrm{d}t = I\omega\dot{\omega}$ yields $\dot{\omega} = -K\omega^3$ with $K = \mu_0 m_\perp^2 / (6\pi c^3 I)$; in terms of the period, $P\dot{P} = 4\pi^2 K$, constant. Integrating $\dot{\omega} = -K\omega^3$ from birth spin ω_0,

$$t = \frac{P}{2\dot{P}}\left[1 - \left(\frac{P_0}{P}\right)^2\right] \leq \frac{P}{2\dot{P}}\,,$$

which is Equation (49.26). Finally, a dipole of moment m has polar surface field $B_\mathrm{p} = \mu_0 m / (2\pi R^3)$; solving $P\dot{P} = 4\pi^2 K$ for $m_\perp$ gives Equation (49.27).

The Crab pulsar makes every line quantitative. With $P = 33.4\,\mathrm{ms}$, $\dot{P} = 4.2 \times 10^{-13}$ and the canonical $I = 10^{38}\,\mathrm{kg\,m^2}$ (a uniform $1.4\,M_\odot$ sphere of 12 km gives $\frac{2}{5}MR^2 \approx 1.6 \times 10^{38}\,\mathrm{kg\,m^2}$), Equation (49.25) gives $\left|\dot{E}\right| = 4\pi^2 I\dot{P}/P^3 \approx 4.5 \times 10^{31}\,\mathrm{W}$ (computed) — and the Crab nebula is observed to radiate of order $10^{31}\,\mathrm{W}$ across the whole spectrum. The pulsar's slowing powers the nebula, which settled the identification [1372, 1384]. The characteristic age Equation (49.26) is 1260 years, against the 972 years actually elapsed since the supernova of AD 1054: the difference measures the birth period — about 16 ms for pure dipole braking. And Equation (49.27) gives $B_\mathrm{p} \approx 4 \times 10^8\,\mathrm{T}$ — ten million times the strongest steady laboratory magnet, yet a thousandth of a magnetar (Section 49.3.2). □

49.3.2 The pulsar population

Over three thousand pulsars are now catalogued, and Proposition 49.17 organizes the whole population on one diagram, period against period derivative. Each star's position yields its characteristic age Equation (49.26) and field Equation (49.27); lines of constant age and constant field grid the plane, and populations separate cleanly on it.

The bulk of the population sits near $P \sim 1\,\mathrm{s}$, $B_\mathrm{p} \sim 10^8\,\mathrm{T}$, ages of megayears: ordinary pulsars, drifting toward long periods and radio death. In the corner of short period *and* small $\dot{P}$ sit the millisecond pulsars, discovered with the 1.558 ms PSR B1937+21 [1681]: by Equation (49.27) their fields are only $\sim 10^4\,\mathrm{T}$ to $10^5\,\mathrm{T}$ and their characteristic ages exceed $10^8\,\mathrm{yr}$ — old stars, not young ones. They are understood as *recycled*: spun up by accretion from a binary companion, which is why most millisecond pulsars have companions and why their timing is stable enough to serve as the precision clocks of Section 49.3.3. The record spin, 716 rotations per second [2095], is close to but below the breakup rate of Equation (49.24) — itself a loose confirmation of the density.

At the opposite extreme, the magnetars: slowly rotating ($P \sim 2\,\mathrm{s}$ to $12\,\mathrm{s}$) but with spin-down fields of $10^{10}\,\mathrm{T}$ to $10^{11}\,\mathrm{T}$, whose decay — not rotation — powers X-ray and gamma-ray outbursts, as Duncan and Thompson proposed [1852]. And superposed on the smooth spin-down of many young pulsars are *glitches*, sudden fractional spin-ups of order 10^{-6} first seen in Vela [1406]: the crust, spun down electromagnetically, decouples from an interior neutron superfluid whose quantized vortices (Chapter 130) unpin and transfer angular momentum catastrophically — the one observational window on the superfluidity of the interior.

49.3.3 Binary pulsars: precision gravity

A pulsar in a close binary is a moving clock in a strong field, and timing it turns the orbit into a laboratory. The Hulse–Taylor binary PSR B1913+16 [1550] is treated as an experiment, apparatus to error budget, in Chapter 47 (Section 47.1); its orbital decay was the first quantitative evidence for gravitational radiation, and the full derivation of the decay rate — eccentricity enhancement included — is carried out in Appendix A.1. Here the binaries appear for what they contribute to this chapter: the most precise strong-field tests of general relativity, and the precisely measured neutron-star masses that anchor Section 49.4.1.

Phenomenon 49.18 (Orbital decay of a binary pulsar)**.** The orbit of the binary pulsar PSR B1913+16 is shrinking. Its orbital period decreases secularly, a change first measured within seven years of the system's discovery [1550, 1703]. Four decades of timing now give a decay rate agreeing with the quadrupole prediction of general relativity to about 0.16%, once the relative acceleration of the solar system and the binary within the Galaxy is subtracted [2275]. The double pulsar PSR J0737-3039, in which both neutron stars are seen as pulsars [2047], yields five independent post-Keplerian parameters from a single system, all mutually consistent [2098]; sixteen years of its timing test the radiative sector to 1.3×10^{-4} [2383]. This is the evidence that gravitational radiation carries energy away from a system, and it preceded the direct detections by four decades.

Rests on Equation (A.83)*, Phenomenon 46.1, and Postulate 44.3.*

Derivation. *Derives Phenomenon 49.18.*

The orbital-decay physics is proved in this book once, in Appendix A.1: energy balance against the quadrupole luminosity Equation (46.6), carried through the Keplerian ellipse harmonics, gives the period derivative Equation (A.83), whose eccentricity enhancement factor $f(e) = \left(1 + \frac{73}{24}e^2 + \frac{37}{96}e^4\right)\left(1 - e^2\right)^{-7/2}$ evaluates to 11.86 at the measured $e = 0.6171$; with the measured masses the prediction is $\dot{P}_b = -2.40 \times 10^{-12}$, evaluated numerically there. We do not repeat the derivation; what remains to state is the correction named in

the phenomenon. The *observed* $\dot{P}_b$ is not purely intrinsic: the pulsar and the solar system accelerate differently in the Galactic potential, and a relative acceleration along the line of sight produces a secular drift of the apparent period (a kinematic Doppler derivative, with a further "Shklovskii" term from proper motion) that adds to $\dot{P}_b$ at the per-cent level for PSR B1913+16. It is computed from the measured distance, proper motion and a Galactic rotation model, and subtracted; the 0.16% agreement of Section 47.1 is between the prediction and this *intrinsic* decay, and its error budget is now dominated by the Galactic correction, not by the timing [2275].

The double pulsar sharpens every number. PSR J0737-3039 is a 2.45 h orbit holding a 22.7 ms recycled pulsar and (until it precessed out of view) a 2.77 s young one [2047, 2098]. Seeing both clocks fixes the mass ratio directly, so each measured post-Keplerian parameter — the periastron advance of about 16.9° per year (a Mercury-century of precession every six hours), the Einstein delay, the two Shapiro-delay parameters (the companion's orbit is seen nearly edge-on), and $\dot{P}_b$ — is a separate test with no freedom left. All are mutually consistent; the Shapiro shape was tested to 0.05% already in 2006 [2098], and with sixteen years of data the observed orbital decay agrees with the general-relativistic prediction to 1.3×10^{-4} — the most stringent test of the quadrupole formula in existence [2383]. The same fits deliver the component masses to a few parts in 10^4, $m_A \approx 1.34\,M_\odot$ and $m_B \approx 1.25\,M_\odot$ [2383], the kind of measurement Section 49.4.1 builds on. □

49.4 Maximum masses and radii

49.4.1 Bounds on the maximum mass

Every equation of state gives, through Theorem 49.9, a maximum neutron-star mass (Remark 49.15); measuring any single star's mass therefore falsifies every candidate whose maximum falls below it. Three binary pulsars carry the current record. The Shapiro delay — the extra light travel time through the companion's potential, the same effect measured across the solar system in Chapter 42 — gave $1.97 \pm 0.04\,M_\odot$ for PSR J1614-2230 [2155]; the spectroscopic orbit of its white-dwarf companion gave $2.01 \pm 0.04\,M_\odot$ for PSR J0348+0432 [2211]; and Shapiro-delay timing of PSR J0740+6620 gave $2.14^{+0.10}_{-0.09}\,M_\odot$ [2335], refined by further timing to $2.08 \pm 0.07\,M_\odot$ [2378]. From the other side, causality alone bounds the maximum mass of any star built on known low-density physics near $3\,M_\odot$ [1533]. The window left open — roughly $2.1\,M_\odot$ to $3\,M_\odot$ — is where the heaviest neutron star and the lightest black hole divide the sky between them.

Phenomenon 49.19 (Neutron stars reach two solar masses)**.** Several pulsars in binaries have masses measured dynamically — from the Shapiro delay of the pulses passing their companion, or from the companion's spectroscopic orbit — with no assumption whatever about the equation of state, and the values reach or exceed $2\,M_\odot$: PSR J1614-2230 [2155], PSR J0348+0432 [2211] and PSR J0740+6620 [2335, 2378]. Every candidate equation of state whose maximum-mass configuration falls below these values is thereby excluded, which is the sharpest single constraint on matter above nuclear density. From the other side, causality alone — the requirement that the sound speed not exceed c, matched to known physics at low density — bounds the mass of any neutron star by about $3.2\,M_\odot$ [1533], and no static star of any composition may be more compact than the Buchdahl bound $GM/(Rc^2) < 4/9$ (Proposition 49.12).

Rests on Theorem 49.9 and Proposition 49.12.

Derivation. *Derives Phenomenon 49.19.*

The lower half of the statement rests on machinery already proved. The masses are Keplerian and post-Keplerian timing measurements — for the Shapiro delay, the excess delay as a function of orbital phase yields the companion mass and inclination separately, and the

mass function then fixes the pulsar mass — with no reference to the pulsar's structure; the exclusion of soft equations of state is then immediate from Remark 49.15: the summit of $M(\rho_c)$ computed from Equation (49.18) for the candidate $P(\rho)$ falls below the measured mass, and no stable configuration of that matter can be the observed star.

The upper half — the Rhoades–Ruffini bound — needs one further argument, and it is worth exhibiting because it uses ignorance optimally. Suppose the equation of state is *known* up to some matching density ρ_m (nuclear physics vouches for it there) and completely unknown beyond, except for thermodynamic stability $\mathrm{d}P/\mathrm{d}\rho \geq 0$ and causality $\mathrm{d}P/\mathrm{d}\rho \leq c^2$. Among all admissible continuations, the one supporting the largest mass is the stiffest, $P = P_m + (\rho - \rho_m)\,c^2$: any softer choice lowers the pressure everywhere and with it the maximum mass. The bound is therefore computed by integrating Equations (49.17) and (49.18) with the causal continuation, and its scale is fixed by dimensional analysis alone. Above ρ_m the only quantities in the problem are G, c and ρ_m, and the unique mass they form is

$$M_{\max} \propto \frac{c^3}{\sqrt{G^3 \rho_m}} \quad \Longrightarrow \quad M_{\max} \approx 3.2 \left(\frac{\rho_m}{4.6 \times 10^{17}\,\mathrm{kg/m^3}} \right)^{-1/2} M_\odot \,, \tag{49.28}$$

where the dimensionless coefficient (about 0.09 of $c^3/\sqrt{G^3\rho_m}$, computed) is quoted from the numerical integration of Rhoades and Ruffini at their matching density [1533]. The bound moves only as $\rho_m^{-1/2}$ — trusting nuclear physics to twice the density costs merely a factor $\sqrt{2}$ — which is what makes it robust, and it brackets the measured two-solar-mass stars from above with room to spare. The Buchdahl statement was proved as Proposition 49.12. □

49.4.2 Radius measurements: X-ray pulse-profile modelling

Masses come from orbits; radii are harder, because a radius only shows itself through the spacetime curvature it sets. The NICER X-ray telescope on the International Space Station measures exactly that. Millisecond pulsars carry hot spots at their magnetic poles, and the thermal X-ray pulse observed as the star rotates is shaped by relativistic light bending: the more compact the star, the more of its far side is visible, the shallower the pulse modulation. Fitting the energy-resolved pulse profile of PSR J0030+0451 — with the rotational Doppler shift and aberration of the 4.87 ms spin included — two teams analysing the same data with independent codes obtained $R = 13.02^{+1.24}_{-1.06}\,\mathrm{km}$ with $M = 1.44^{+0.15}_{-0.14}\,M_\odot$ [2338], and $R = 12.71^{+1.14}_{-1.19}\,\mathrm{km}$ with $M = 1.34^{+0.15}_{-0.16}\,M_\odot$ [2339]. For the $2.08\,M_\odot$ pulsar J0740+6620 — where the mass is known in advance from radio timing (Section 49.4.1) — the same technique gives $R = 13.7^{+2.6}_{-1.5}\,\mathrm{km}$ [2386] and $R = 12.39^{+1.30}_{-0.98}\,\mathrm{km}$ [2389]: the heaviest well-measured neutron star is no smaller than the canonical ones, which by Remark 49.15 demands a pressure staying stiff deep into the core. An unanticipated by-product: the pulse-profile fits require the hot spots of J0030+0451 to lie in the *same* hemisphere — the magnetic field is far from a centred dipole, a discovery about pulsar magnetospheres made by a radius measurement [2338, 2339].

49.4.3 Tidal deformability: GW170817

The binary neutron-star inspiral GW170817 (Chapter 47) measured neutron-star structure through a fourth, fully independent channel: the tidal distortion each star raises on the other imprints itself on the gravitational waveform in the final orbits [2278]. The measured quantity is a dimensionless number defined as follows.

Definition 49.20 (Tidal deformability)**.** Place a star of mass M and radius R in a weak external tidal field $\mathcal{E}_{ij}$ (in the Newtonian limit, $\mathcal{E}_{ij} = \partial_i \partial_j \Phi_{\mathrm{ext}}$ evaluated at its centre). Its

induced mass quadrupole moment is, to linear order,

$$Q_{ij} = -\lambda\,\mathcal{E}_{ij}\,, \qquad \lambda = \frac{2}{3}\,k_2\,\frac{R^5}{G}\,, \tag{49.29}$$

which defines the tidal Love number k_2; the *dimensionless tidal deformability* is

$$\Lambda \equiv \frac{\lambda\,c^{10}}{G^4 M^5} = \frac{2}{3}\,k_2\left(\frac{c^2 R}{GM}\right)^5 . \tag{49.30}$$

For realistic neutron stars $k_2 \approx 0.05$ to 0.15 [2127], and the fifth power of the inverse compactness makes Λ a few hundred — and makes it, of all waveform parameters, the one that measures the radius.

Phenomenon 49.21 (The radius of a neutron star is near twelve kilometres)**.** Two methods sharing no systematics agree. Modelling the thermal X-ray pulse profile of PSR J0030+0451, whose shape is set by the relativistic bending of light around the star and therefore by its compactness, gives a radius between about 12 km and 13 km; two teams analysed the same data independently and agreed [2338, 2339]. Measuring instead the tidal deformability imprinted on the late inspiral phasing of the binary neutron-star merger GW170817 [2278] gives $\Lambda(1.4\,M_\odot) = 190^{+390}_{-120}$ and a radius between about 11 km and 13.5 km [2294]. Taken with Phenomenon 49.19, these bracket the pressure of matter above nuclear saturation density — the central unknown of Section 49.2.4 — from two directions at once: the masses exclude equations of state that are too soft, the radii those that are too stiff.

Rests on Equation (49.30)*.*

Derivation. *Derives Phenomenon 49.21.*

Each method needs one mechanism derived: light bending for the pulse profile, and the entry of Λ into the waveform for GW170817.

Light bending. A photon leaving the surface at angle α to the outward radial direction carries, by the null-geodesic machinery of Chapter 45, an impact parameter $b = R\sin\alpha/\sqrt{1 - r_\mathrm{s}/R}$, and arrives at infinity at an angle

$$\psi(\alpha) = \int_R^\infty \frac{b\,\mathrm{d}r}{r^2\sqrt{1 - \dfrac{b^2}{r^2}\left(1 - \dfrac{r_\mathrm{s}}{r}\right)}}$$

from the radius of emission. Grazing emission ($\alpha = \pi/2$) fixes the visible cap: everything within colatitude $\psi_\mathrm{max} = \psi(\pi/2)$ of the sub-observer point is seen at once, and $\psi_\mathrm{max} > \pi/2$ whenever $r_\mathrm{s} > 0$ — a compact star shows more than half its surface. The integral is done numerically in the fits; the compact approximation $\cos\alpha \approx u + (1-u)\cos\psi$ with $u = r_\mathrm{s}/R$, accurate to a per cent for $R \gtrsim 2r_\mathrm{s}$ [2022], exhibits the physics: $\cos\psi_\mathrm{max} = -u/(1-u)$, so for PSR J0030+0451 ($u \approx 0.33$) the visible cap extends to $\psi_\mathrm{max} \approx 119^\circ$ — about 74% of the surface (computed). A polar hot spot therefore never fully disappears, and the depth of the pulse modulation falls monotonically with u: the profile measures compactness, while the observed flux and temperature scale, blue- and redshifted by $(1+z) = (1-u)^{-1/2} \approx 1.2$, tie in the radius itself. Inverting energy-resolved profiles for mass, radius, spot geometry and viewing angle is the numerical likelihood analysis of [2338, 2339], run twice with independent codes as the phenomenon records.

Tidal phasing. In a binary at separation r, each star sits in its companion's tidal field, $\mathcal{E}_{ij} \sim Gm_2/r^3$, and by Equation (49.29) acquires $Q_{ij} = -\lambda\mathcal{E}_{ij}$. The induced quadrupole does two things, both scaling with λ: it deepens the mutual potential by the induced-moment interaction $\delta U = -\frac{1}{2}\lambda\mathcal{E}_{ij}\mathcal{E}^{ij} \sim -\lambda G^2 m_2^2/r^6$, and it adds coherently to the quadrupole moment whose time variation radiates (Equation (46.6)). Relative to the point-mass energy

$\sim Gm_1m_2/r$, the tidal term is smaller by $\lambda G/r^5 \sim \Lambda\,(GM/rc^2)^5$: five powers of compactness, i.e. ten powers of v/c — formally a fifth-post-Newtonian correction, utterly negligible except that Λ is a few *hundred*, and that the phase of a chirping binary (Equation (46.8)) accumulates over thousands of cycles. Carrying the energy and flux corrections through the phase evolution gives the leading tidal contribution to the frequency-domain phase, proportional to the single mass-weighted combination

$$\tilde{\Lambda} = \frac{16}{13}\,\frac{(m_1 + 12m_2)\,m_1^4\Lambda_1 + (m_2 + 12m_1)\,m_2^4\Lambda_2}{(m_1+m_2)^5}\,,$$

with a known numerical coefficient, quoted from the standard computation [2124, 2127]. Matched filtering of GW170817's final orbits therefore measures $\tilde{\Lambda}$; the discovery analysis bounded $\Lambda(1.4\,M_\odot) \le 800$ at 90% confidence (low-spin prior) [2278], and the dedicated reanalysis gives $\Lambda(1.4\,M_\odot) = 190^{+390}_{-120}$, which through Equation (49.30) with the relativistic k_2 of [2127] yields $R = 11.9 \pm 1.4$ km for both stars [2294] — the numbers the phenomenon quotes, in agreement with NICER's, from data that share not a single instrument, wavelength or systematic. The same event's electromagnetic counterpart — the kilonova whose lanthanide spectra realized the r-process site of Chapter 52 — and its detection history belong to Chapter 47. □

49.5 The black-hole endpoint

Beyond the maximum mass there is nothing. The sequence of cold equilibria ends (Section 49.4.1), the Buchdahl bound (Proposition 49.12) forbids trading size for survival, and a core pushed past the limit — by accretion, by a merger, or at birth — collapses with nothing left to stop it. Oppenheimer and Snyder solved the idealized problem in the same year as the equilibrium equations: a uniform pressureless cloud collapsing from rest [844]. Each dust element falls along a radial geodesic of the exterior geometry, reaching the gravitational radius r_s and the centre in *finite proper time* — for a neutron star of $2\,M_\odot$ collapsing from 12 km, the free-fall time $\tau = \frac{\pi}{2}\sqrt{R_0^3/\,(2GM)}$ is about a tenth of a millisecond (computed) — while a distant observer, receiving light ever more redshifted by the factor $(1 - r_\mathrm{s}/r)^{-1/2}$ of Chapter 45, sees the surface freeze and fade at r_s. Nothing marks the crossing locally: the horizon is a global feature of the geometry, not a place where anything happens, and the interior solution matches a collapsing cosmology of the very kind Chapter 48 runs forward.

That is where this chapter hands over. The observational record of the objects on the far side of the Buchdahl line — X-ray binary masses, the shadow of a horizon, ringdowns — is Chapter 50; the merger catalogue of Chapter 47, opened by GW150914 [2264], now measures both sides of the boundary at once, heaviest neutron stars against lightest black holes, and the sparsely populated interval between about $2.2\,M_\odot$ and $5\,M_\odot$ — the "mass gap", whether real or observational — is the current form of a question this chapter has asked three times: how much cold matter can gravity tolerate? Every answer so far — Chandrasekhar's $1.4\,M_\odot$, causality's $3\,M_\odot$, Buchdahl's 4/9 — has been a theorem first and a measurement second, which is this book's definition of understanding a thing.

Chapter 50

Experiment: Black-Hole Observations

Tests Phenomenon 42.5 and Equations (42.8) *and* (50.4).
Assuming Equation (50.2).

The stellar orbits around Sgr A*, the Event Horizon Telescope images, and the X-ray binary mass functions: this chapter assembles the observational case that the Schwarzschild and Kerr geometries of Chapter 45 are realized in Nature. The claim under test is specific—a mass confined within its gravitational radius, with no material surface—and it has been established four independent ways: dynamically, by the mass function of the X-ray binary Cygnus X-1 [1466, 1488]; astrometrically, by the stellar orbits around Sgr A* [2125, 2146]; by direct imaging at event-horizon scale in M87* [2328] and Sgr A* [2403]; and by the gravitational-wave mergers of Chapter 47 [2264].

The four channels are treated as one experiment because they test one prediction from four directions, at masses spanning six orders of magnitude and field strengths spanning the gap between a stellar orbit a thousand gravitational radii out and light circling the photon sphere itself. No single channel proves the case alone; the concordance is the measurement, and Section 50.5 states with some care what it does and does not establish.

50.1 Historical context and the prediction under test

50.1.1 From invisible companions to relativistic collapse

The idea is older than the theory it now tests. In 1784 John Michell, reasoning with Newtonian corpuscles of light, computed that a body of the Sun's density and five hundred times its diameter would have an escape speed exceeding the speed of light, so that "all light emitted from such a body would be made to return towards it"—and he added the observation on which this entire chapter rests: such invisible bodies could nevertheless be detected through the motion of luminous ones revolving about them [37]. Every dynamical mass in Section 50.4 is Michell's method carried out.

General relativity replaced the escape-speed argument with a sharper structure. Within weeks of the field equations, Schwarzschild produced the exact exterior solution for a spherical mass [440], with its characteristic radius $r_{\mathrm{s}} = 2GM/c^2$; for decades the surface at r_{s} was read as a coordinate accident or dismissed as physically unreachable. Two 1939 calculations made it unavoidable. Oppenheimer and Volkoff showed that cold neutron matter has a maximum mass, of order a solar mass in their equation of state, above which no static configuration exists at all [845]; and Oppenheimer and Snyder integrated the collapse of a pressureless ball past that point, finding that the matter crosses r_{s} in finite proper time while a distant observer sees the surface freeze and redshift away [844]. Kerr's rotating generalization [1259] and Penrose's theorem that collapse past a trapped surface ends in a singularity regardless

of symmetry [1319] completed the theoretical picture, which Chapter 45 develops in full: if masses above the cold-matter limit exist and lose their support, general relativity permits them no fate but a black hole.

Observation entered by another door. A 1962 sounding-rocket flight instrumented to look for X-ray fluorescence from the Moon instead recorded a bright source in Scorpius and a diffuse background—the birth of X-ray astronomy [1225]; compact accreting objects, it turned out, announce themselves most loudly at kiloelectronvolt energies. In 1963 Schmidt recognized the emission lines of the "radio star" 3C 273 as hydrogen lines redshifted by $z = 0.158$ [1265], which put the object at cosmological distance and its luminosity orders of magnitude beyond what any assembly of stars in so small a volume could supply. Lynden-Bell drew the inference that quasars are accreting supermassive collapsed objects, that dead quasars must therefore litter the present universe, and that the nuclei of ordinary galaxies—our own included—are where they hide [1403]. The compact nonthermal radio source now called Sgr A* was found at the centre of the Milky Way in 1974 [1517]. The four observational programmes reported below are the systematic interrogation of these two populations, stellar and supermassive.

50.1.2 What general relativity predicts

The falsifiable signatures, each derived in Chapter 45, are these.

First, a purely Newtonian instrument: for a spectroscopic binary the period and radial-velocity amplitude of the visible star alone bound the mass of its unseen companion from below, Equation (50.2). The bound involves no model of the companion, so wherever it exceeds the maximum mass of a cold compact star (Chapter 49), the companion can only be a black hole. This inequality is stated and proved with the Cygnus X-1 data in Section 50.4.1.

Second, a horizon at $r_{\mathrm{s}} = 2GM/c^2$ and hence no material surface: no thermonuclear flashes from accumulated fuel, no quiescent re-radiation of the accretion energy deposited on a surface. Accreted energy not radiated on the way down is lost. The comparison that turns this into a measurement is made in Section 50.5.

Third, the strong-field orbital mechanics of the Schwarzschild geometry: an innermost stable circular orbit truncating the accretion disc and fixing its radiative efficiency, gravitational redshift and periastron advance for test bodies—the same advance formula, Equation (42.8), that Mercury tests at solar field strengths, here applied four orders of magnitude deeper.

Fourth, a photon sphere. Light passing a black hole closer than the critical impact parameter is captured, so a black hole seen against its own emission shows a dark disc—a shadow—of angular diameter $3\sqrt{3}\,r_{\mathrm{s}}/D \simeq 5.2\,r_{\mathrm{s}}/D$ at distance D, nearly independent of spin and viewing angle. Falcke, Melia and Agol computed the image for Sgr A* and showed that at millimetre wavelengths, where the accretion flow turns transparent and the atmosphere still permits interferometry, the shadow of the Galactic centre—about thirty microarcseconds for the mass then favoured—is resolvable from the ground [1988]. Inverted, the shadow size weighs the hole: Equation (50.4).

Fifth, gravitational radiation: an inspiralling binary chirps and the merged remnant rings down at the quasinormal frequencies of a Kerr hole, per Chapter 46. That channel has its own chapter, Chapter 47; here it contributes masses.

50.2 Apparatus

50.2.1 X-ray observatories

The stellar-mass channel is X-ray astronomy, because gas spiralling onto an object a few tens of kilometres across is heated to 10^7–10^8 kelvin and radiates at kiloelectronvolt energies, to which the atmosphere is opaque. The first generation of instruments were gas proportional

counters flown above it: windowed tubes in which an incoming X-ray photoionizes the fill gas and the charge avalanche measures its energy, with sensitivity in the few- to tens-of-keV band, no true imaging, and fields of view defined by mechanical collimators [1225]. Rockets gave minutes of exposure; the scanning and pointed satellites that followed gave years, all-sky monitors to catch transient outbursts, and—decisive for Cygnus X-1—timing at millisecond resolution, which revealed the rapid aperiodic flickering that marks an emitting region light-milliseconds across [2103]. Later observatories added CCD and grating spectrometers resolving the iron Kα fluorescence line near 6.4 keV, whose gravitationally redshifted, relativistically broadened profile probes the innermost disc [1911]. The optical half of the apparatus is conventional stellar spectroscopy of the companion star, from which come the period and velocity amplitude that feed the mass function; the modern addition is very-long-baseline radio astrometry of the compact object itself, precise enough to measure its trigonometric parallax [2387].

50.2.2 Adaptive optics and interferometry on the Galactic Centre

The Galactic centre is invisible at optical wavelengths behind about thirty magnitudes of dust extinction, but opens up in the near-infrared at wavelengths near 2.2 µm. There the diffraction limit of a 10 m telescope is $\lambda/D \approx 2.2\times10^{-7}$ rad, about 45 milliarcseconds — fine enough to separate the stars of the central cluster, if the atmosphere can be defeated. Speckle imaging (freezing the turbulence in millisecond exposures and reconstructing afterwards) and then adaptive optics (measuring the wavefront on a reference star or laser beacon and correcting it with a deformable mirror at kilohertz rates) brought the Keck 10 m and VLT 8.2 m telescopes to that limit, and two decades of such imaging, tied to a radio reference frame and combined with spectroscopic radial velocities, produced the stellar orbits of Section 50.4.2 [2125, 2146][2156].

The precision instrument is GRAVITY, the beam combiner that phase-references all four VLT unit telescopes into a single interferometer with baselines up to 130 m [2279]. The fringe spacing at 2.2 µm is $\lambda/B \approx 1.7\times10^{-8}$ rad, some 3.5 milliarcseconds, and because the quantity measured is a fringe phase against a nearby reference source, positions within that fringe are read to a small fraction of it: astrometry at the level of a few tens of microarcseconds, roughly a hundredfold gain on single-telescope imaging. This is the instrument that turned the S2 orbit from a mass measurement into a test of general relativity [2295, 2347].

50.2.3 The Event Horizon Telescope

Resolving an event horizon takes another three orders of magnitude. The two largest black-hole shadows on the sky—Sgr A* and M87*— subtend about fifty microarcseconds, so the required resolution is $\sim 2\times10^{-10}$ rad, which at an observing wavelength of 1.3 mm demands an aperture the size of the Earth: $\lambda/D \approx 1.2\times10^{-10}$ rad, about 25 microarcseconds, for baselines near 1.1×10^{7} m. The Event Horizon Telescope synthesizes exactly that, by very-long-baseline interferometry: in April 2017, eight radio telescopes at six sites from Hawaii to the South Pole observed simultaneously at 1.3 mm, each recording the raw voltage stream against a hydrogen-maser clock, with the phased ALMA array contributing the sensitivity of a 73 m equivalent dish. Petabytes of recorded signal were later cross-correlated at dedicated correlator facilities, fringe-fitted, and calibrated through the chain described in the instrument papers accompanying [2328]. The wavelength is chosen twice over: it is short enough that the synchrotron emission of the accretion flow becomes optically thin, so the line of sight reaches the horizon scale, and long enough that the atmosphere and the antennas still support coherent interferometry [1988].

Gravitational-wave interferometers, the fourth apparatus, are described in Chapter 47 [2243].

50.3 Procedure

Four measurement protocols, one per channel.

Radial velocities. The optical companion's spectrum is observed over many orbits; the Doppler shift of its lines yields the period P and velocity semi-amplitude K, and with them the mass function $f = PK^3/2\pi G$ of Equation (50.1)—a strict lower bound on the unseen mass [1466, 1488]. Converting the bound into a mass requires the inclination and the companion mass, which enter through ellipsoidal light-curve modelling, the spectral classification of the companion, and the distance; the modern Cygnus X-1 solution fixes the distance by VLBI parallax and fits all elements simultaneously [2178, 2387]. For the transient systems the same spectroscopy is performed in quiescence, when the disc is faint and the companion's lines are clean [2103].

Astrometric orbits. Individual stars near Sgr A* are followed for decades in position (imaging, then interferometry) and in radial velocity (infrared spectroscopy). A simultaneous fit of all measured stars solves for each star's six orbital elements plus the central mass M and the Sun–centre distance R_0, which the data determine jointly because the angular orbit scales as $M^{1/3}/R_0$ while the velocities are absolute [2125, 2146]. The relativistic tests are differential: the Keplerian orbit is fitted and the residual at pericentre—in radial velocity for the redshift, in the orbit orientation for the precession—is compared with the prediction, each effect published only after being measured on both sides of pericentre [2295, 2336, 2347].

Interferometric imaging. The EHT data reduce to complex visibilities on every baseline. Calibration leans on closure quantities—combinations of phases around telescope triangles and of amplitudes around quadrangles from which station-dependent errors cancel exactly. Images were reconstructed by four teams working blind of one another with two independent algorithm families, and only then compared; in parallel, geometric crescent models and libraries of general-relativistic magnetohydrodynamic simulations were fitted directly to the visibilities. The ring diameter is the quantity on which all methods converge, and the procedure was validated end-to-end on synthetic data with known answers [2328, 2403].

Matched filtering. Gravitational-wave strain data are correlated against relativistic waveform banks and the significance established on time-shifted background, per Chapter 47 [2264]. This chapter takes from that channel only the masses and the ringdown.

50.4 Observations and data

50.4.1 Cygnus X-1 and the X-ray binaries

Cygnus X-1 was among the brightest X-ray sources in the sky and the most erratic, flickering on timescales down to milliseconds. When a radio counterpart pinned its position in 1971, the field contained the ninth-magnitude B0 supergiant HDE 226868. In early 1972 Webster and Murdin [1488] and, independently, Bolton [1466] showed the supergiant to be a single-lined spectroscopic binary with a period of 5.6 days, and both drew the same conclusion: for any mass consistent with the supergiant's spectral type, the velocity amplitude demands an unseen companion of several solar masses at least—too heavy for the white dwarfs and neutron stars then known, and dark at every wavelength searched. The phrase used was cautious ("a heavy companion"), but this was the first dynamical identification of a stellar black-hole candidate.

The half-century since has only sharpened the numbers. Table 50.1 collects the modern solution: the distance from VLBI parallax, and with it a black-hole mass of $21.2 \pm 2.2\,M_\odot$ [2387]. The mass function computed from the period and the modern velocity amplitude is small — the orbit is seen nearly face-on — which is exactly why the full solution with

inclination and companion mass is needed here, and why the transient systems below, where the mass function alone settles the question, carry independent weight.

Table 50.1: The Cygnus X-1 system. The orbital period is common to the identification papers [1466, 1488] and the modern solution; the velocity semi-amplitude of HDE 226868 is the modern spectroscopic value [2178]; the mass function is computed here from those two by Equation (50.1); the inclination (published uncertainty below one degree), distance and black-hole mass are the parallax-anchored solution of [2387].

quantity	as published	SI
orbital period P	5.600 days	$4.838 \times 10^5\,\mathrm{s}$
velocity semi-amplitude K	75.6 ± 0.7 km/s	$(7.56 \pm 0.07) \times 10^4\,\mathrm{m/s}$
mass function f (computed)	$0.251\,M_\odot$	$5.0 \times 10^{29}\,\mathrm{kg}$
orbital inclination i	about 27.5°	—
distance d	$2.22^{+0.18}_{-0.17}\,\mathrm{kpc}$	$(6.9 \pm 0.6) \times 10^{19}\,\mathrm{m}$
black-hole mass M_X	$21.2 \pm 2.2\,M_\odot$	$(4.22 \pm 0.44) \times 10^{31}\,\mathrm{kg}$

Cygnus X-1 is persistent, fed by the supergiant's wind. The larger population is transient: systems with low-mass companions that outburst for months and then sit quiescent for years, leaving the companion's spectrum clean. In the best of them the inequality proved below needs no help at all: the discovery measurement for V404 Cygni gave a mass function above $6\,M_\odot$ [1850], refined to $6.08 \pm 0.06\,M_\odot$ in the modern tabulation [2103]—the absolute floor on the compact object's mass, from two spectroscopic observables, already twice the causal neutron-star bound of Chapter 49. Some twenty systems now have dynamical masses, running from about $5\,M_\odot$ to about $20\,M_\odot$, and their X-ray spectra and variability are read through the accretion-disc theory of Shakura and Sunyaev [1512][2103]. The gravitationally redshifted iron Kα line, first resolved in the active galaxy MCG–6–30–15, extends the same disc diagnostics to within a few gravitational radii of the horizon [1911].

Phenomenon 50.1 (Dark companions too massive to be neutron stars)**.** In a number of X-ray binaries the spectroscopic orbit of the optical star requires an unseen companion heavier than any neutron star can be. The first such case was Cygnus X-1, identified in 1972 with the supergiant HDE 226868 [1466, 1488]; its modern dynamical mass is $21.2 \pm 2.2\,M_\odot$ [2387]. In the transient black-hole binaries the argument is sharper, because for several of them the mass function alone — a quantity built from the orbital period and the radial-velocity amplitude, and independent of both the inclination and the companion's mass — already exceeds the maximum neutron-star mass [1850][2103]. Dynamically measured masses in this population run from about $5\,M_\odot$ to about $20\,M_\odot$.

Rests on Theorem 27.30 and Phenomenon 27.35.

Derivation. *Derives Phenomenon 50.1.*

The inequality on which every dynamical black-hole mass rests is worth deriving in full. Let the compact object of mass M_X and the optical star of mass M_opt move on a circular orbit of separation a and period P, inclined at i to the plane of the sky. The optical star circles the centre of mass at radius $a_\mathrm{opt} = aM_\mathrm{X}/\left(M_\mathrm{X} + M_\mathrm{opt}\right)$, and spectroscopy measures the semi-amplitude of its line-of-sight velocity,

$$K = \frac{2\pi a_\mathrm{opt}}{P} \sin i\,.$$

Out of the two observables P and K define the *mass function*

$$f := \frac{PK^3}{2\pi G}\,. \tag{50.1}$$

Substituting for K and then eliminating a with Kepler's third law $4\pi^2 a^3 = G\,(M_{\mathrm{X}} + M_{\mathrm{opt}})\,P^2$,

$$f = \frac{P}{2\pi G}\cdot\frac{8\pi^3 a_{\mathrm{opt}}^3 \sin^3 i}{P^3} = \frac{4\pi^2 a^3}{GP^2}\cdot\frac{M_{\mathrm{X}}^3 \sin^3 i}{(M_{\mathrm{X}}+M_{\mathrm{opt}})^3} = \frac{M_{\mathrm{X}}^3 \sin^3 i}{(M_{\mathrm{X}}+M_{\mathrm{opt}})^2}\,.$$

Since $\sin i \le 1$ and $M_{\mathrm{opt}} > 0$, the right-hand side cannot exceed M_{X}, so

$$M_{\mathrm{X}} \ge f = \frac{PK^3}{2\pi G}\,. \tag{50.2}$$

The mass function is therefore a strict lower bound on the mass of the unseen body, extracted from two numbers a spectrograph delivers directly, with no knowledge of the inclination, the distance or the companion star. Wherever f by itself exceeds the causal neutron-star bound of Chapter 49, the identification of the companion as a black hole follows with nothing further assumed — and this is the only one of the four measurement channels of this chapter that requires no relativistic input at all. For Cygnus X-1, $P = 4.838\times10^5\,\mathrm{s}$ and $K = 7.56\times10^4\,\mathrm{m/s}$ give

$$f = \frac{4.838\times10^5\,\mathrm{s}\times(7.56\times10^4\,\mathrm{m/s})^3}{2\pi\times6.674\times10^{-11}\,\mathrm{m^3/(kg\,s^2)}} \approx 5.0\times10^{29}\,\mathrm{kg} \approx 0.25\,M_\odot\,,$$

which is why that system's mass rests on the full orbital solution of Table 50.1 rather than on the bound; for V404 Cygni the bound itself is $6.08\,M_\odot$ [1850][2103] and the conclusion is immediate. □

50.4.2 Sgr A*: stellar orbits

The decisive Galactic-centre observation was the closing of a single stellar orbit. The star S2, tracked from 1992 and followed through its pericentre passage in 2002, was shown to move on a Keplerian ellipse of period near fifteen years with the radio source Sgr A* at its focus—the first time the enclosed mass could be assigned to a volume as small as a single stellar orbit [2038]. Two decades of astrometry from Keck and the VLT then fixed the mass near $4\times10^6\,M_\odot$: the Keck solution gave $(4.1\text{–}4.5)\times10^6\,M_\odot$ depending on the treatment of the distance [2125], the VLT solution $(4.31\pm0.06_{\mathrm{stat}}\pm0.36_{R_0})\times10^6\,M_\odot$ [2146], and the two programmes' agreement, star by star, is reviewed in [2156].

GRAVITY made the orbit relativistic. At the May 2018 pericentre passage—the star moving at 7650 kilometres per second, 2.6% of c—the measured radial velocity exceeded the Keplerian prediction by about 200 km/s, the combined gravitational redshift and transverse-Doppler shift; expressed as a parameter interpolating between Newtonian gravity (0) and general relativity (1), the result was 0.90 ± 0.17 statistical, with a comparable systematic [2295]. The Keck group, with independent instruments, astrometry and analysis, found 0.88 ± 0.17 [2336]. Continued interferometric monitoring through and beyond pericentre then resolved the Schwarzschild precession of the orbit, 12 minutes of arc per revolution in the plane of the orbit and prograde, at 1.10 ± 0.19 of the prediction of Equation (42.8) [2347]. The same fit delivers the current best mass and distance, collected in Table 50.2.

Phenomenon 50.2 (Four million solar masses inside the orbit of a single star)**.** Individual stars in the central arcsecond of the Galaxy have been followed through complete orbits about a common focus coincident with the compact radio source Sgr A*. The star S2, first traced through a pericentre passage in 2002 [2038], closes its orbit in 16.05 yr [2295], reaching a pericentre distance of about $1.8\times10^{13}\,\mathrm{m}$. Two decades of adaptive-optics astrometry from

Table 50.2: The S2 orbit and its relativistic parameters. Orbital elements, mass and distance are the interferometric solution [2295, 2347]; the pericentre distance and speed are computed here from them (the published values agree). The redshift parameter is 0 for Newtonian gravity and 1 for general relativity; the precession parameter multiplies the Schwarzschild advance of Equation (42.8). The Keck redshift value is the independent measurement of [2336].

quantity	as published	SI
orbital period P	16.05 yr	$5.07 \times 10^{8}\,\mathrm{s}$
eccentricity e	0.885	—
angular semi-major axis	125 mas	$6.06 \times 10^{-7}\,\mathrm{rad}$
pericentre distance r_{p}	≈ 120 au	$1.8 \times 10^{13}\,\mathrm{m}$
pericentre speed v_{p}	$\approx 0.026\,c$	$7.65 \times 10^{6}\,\mathrm{m/s}$
enclosed mass M	$4.26 \times 10^{6}\,M_{\odot}$	$8.47 \times 10^{36}\,\mathrm{kg}$
distance R_0	8.25 kpc	$2.545 \times 10^{20}\,\mathrm{m}$
redshift parameter (VLT; Keck)	0.90 ± 0.17; 0.88 ± 0.17	—
precession parameter	1.10 ± 0.19	—

Keck and the VLT [2125, 2146][2156], later sharpened by interferometry [2279], give an enclosed mass of about $4 \times 10^{6}\,M_{\odot}$ within that pericentre. The orbits are Keplerian about a point mass to the precision of the data: no extended distribution of the required mass is consistent with them.

Rests on Phenomenon 27.35 and Theorem 27.30.

Derivation. *Derives Phenomenon 50.2.*

The mass follows from Kepler's third law applied to the fitted orbit, which is what makes it independent of any model of the source. S2's angular semi-major axis is about 125 milliarcseconds; at the measured distance to the Galactic centre, $R_0 \approx 8.2\,\mathrm{kpc} = 2.53 \times 10^{20}\,\mathrm{m}$, and with one milliarcsecond equal to 4.848×10^{-9} in radian measure, this is

$$a \approx 125 \times 4.848 \times 10^{-9} \times 2.53 \times 10^{20}\,\mathrm{m} \approx 1.53 \times 10^{14}\,\mathrm{m}\,.$$

With $P = 16.05\,\mathrm{yr} = 5.07 \times 10^{8}\,\mathrm{s}$,

$$M = \frac{4\pi^2 a^3}{G P^2} = \frac{4\pi^2\,(1.53 \times 10^{14}\,\mathrm{m})^3}{6.674 \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)} \times (5.07 \times 10^{8}\,\mathrm{s})^2} \approx 8.2 \times 10^{36}\,\mathrm{kg}\,,$$

that is about $4.1 \times 10^{6}\,M_{\odot}$; the mass of S2 itself, of order $10\,M_{\odot}$, is negligible against this and has been dropped. It is worth recording what this does and does not establish. The Schwarzschild radius of such a mass is $2GM/c^2 \approx 1.2 \times 10^{10}\,\mathrm{m}$, so S2's pericentre lies some 1500 gravitational radii out. The stellar orbits therefore demonstrate a very large mass in a very small volume — a mean density that no cluster of ordinary stars or stellar remnants could sustain against evaporation and collision for the age of the Galaxy — but they do not by themselves reach the horizon. That is what the imaging of Section 50.4.3 adds. □

Phenomenon 50.3 (Relativistic redshift and precession of a stellar orbit)**.** At S2's 2018 pericentre passage the measured radial velocity departed from the Keplerian prediction by an amount corresponding to about 200 km/s, in the sense and of the magnitude predicted by the combined gravitational and transverse-Doppler shift [2295], a result confirmed by the independent Keck programme [2336]. Continued monitoring then detected the Schwarzschild precession of the orbit itself, an in-plane advance of the pericentre of about twelve minutes of arc per orbit, consistent with general relativity [2347]. Both effects were predicted before they were measured, and neither has a Newtonian counterpart.

Rests on Phenomena 38.13, 42.5 and 42.6.

Derivation. *Derives Phenomenon 50.3.*

The redshift excess is first order in the small quantities and can be checked directly. A clock at radius r from a mass M moving with speed v is shifted, to first order in $GM/\left(rc^2\right)$ and v^2/c^2, by

$$z = \frac{GM}{rc^2} + \frac{v^2}{2c^2}\,, \tag{50.3}$$

the first term the gravitational redshift of Chapter 42 and the second the transverse Doppler shift of Chapter 38. Both are positive and both peak at pericentre, where the star is deepest in the potential and moving fastest. Taking the mass of Phenomenon 50.2, for which $GM/c^2 \approx 6.1 \times 10^9\,\mathrm{m}$, with S2's pericentre radius $r_\mathrm{p} \approx 1.8 \times 10^{13}\,\mathrm{m}$ and pericentre speed $v_\mathrm{p} \approx 7650\,\mathrm{km/s}$,

$$\frac{GM}{r_\mathrm{p}c^2} \approx 3.4 \times 10^{-4}\,, \qquad \frac{v_\mathrm{p}^2}{2c^2} \approx 3.3 \times 10^{-4}\,,$$

so $z \approx 6.7 \times 10^{-4}$ and $cz \approx 200\,\mathrm{km/s}$, which is the measured excess.

That the two contributions come out nearly equal is not a coincidence. For a Keplerian orbit the vis-viva relation gives $v_\mathrm{p}^2 = (1+e)\,GM/r_\mathrm{p}$ at pericentre, so the transverse-Doppler term is exactly $(1+e)/2$ times the gravitational one — a factor 0.94 for S2's eccentricity $e \approx 0.88$. The two therefore cannot be separated by this measurement; what is tested is their sum, and disentangling them would require an orbit of very different eccentricity.

The precession is the classical formula of Equation (42.8) evaluated on this orbit. With $GM/c^2 \approx 6.1 \times 10^9\,\mathrm{m}$, $a \approx 1.53 \times 10^{14}\,\mathrm{m}$ and $1 - e^2 \approx 0.217$,

$$\Delta\varphi = \frac{6\pi GM}{c^2 a\left(1-e^2\right)} \approx \frac{6\pi \times 6.1 \times 10^9\,\mathrm{m}}{1.53 \times 10^{14}\,\mathrm{m} \times 0.217} \approx 3.5 \times 10^{-3}\,,$$

in radian measure per revolution — about twelve minutes of arc, which is what is measured [2347]. Mercury, for comparison, accumulates 5.0×10^{-7} radian per revolution (Chapter 53): the same formula, verified at a relativistic parameter $GM/\left(c^2 a\left(1-e^2\right)\right)$ some four orders of magnitude larger. □

50.4.3 Event-horizon-scale imaging

In April 2019 the Event Horizon Telescope published the first image of a black hole: the nucleus of M87 resolved into a bright, asymmetric ring of 42 ± 3 microarcseconds diameter around a deep central brightness depression [2328]. The ring persisted across four observing nights and across every independent imaging pipeline; its south side is brighter, as Doppler beaming requires for emission orbiting with the small approaching-side inclination inferred from the large-scale jet; and the inferred mass, $(6.5 \pm 0.7) \times 10^9\,M_\odot$, is the photon-capture cross-section of Equation (50.4) read off the sky. The polarized image that followed mapped an ordered, partly radial magnetic-field pattern in the ring, of the kind required to launch the jet and consistent with a dynamically important ("magnetically arrested") field at the horizon [2372].

One honest calibration must be stated. What the interferometer measures is the diameter of the *emission* ring, which simulations place close to, but systematically outside, the critical curve bounding the shadow — by roughly ten per cent, depending on where the emission peaks. The published mass therefore carries a theory-side calibration from the simulation library, and its error bar does [2328]. The Schwarzschild arithmetic below, which ignores this distinction, overshoots the full analysis by about the same ten per cent, exactly as it should.

In May 2022 the same technique, hardened against the source's minute-scale variability (the dynamical time at the horizon of a $4 \times 10^6\,M_\odot$ hole is minutes, not days), yielded the image of Sgr A*: a ring of 51.8 ± 2.3 microarcseconds [2403]. Here the prediction was

available in advance with no freedom at all: the stellar-orbit mass and distance of Table 50.2 give a Schwarzschild shadow of 53 microarcseconds (computed here; spin and inclination lower it by a few per cent). The measured ring agrees. Table 50.3 collects both sources.

Table 50.3: Event-horizon-scale imaging results, quoted from [2328] (M87*) and [2403] (Sgr A*). The last row is computed here: the Schwarzschild shadow diameter $3\sqrt{3}\,r_{\mathrm{s}}/R_0$ for the mass and distance measured by the stellar orbits of Table 50.2, the number the Sgr A* ring was predicted to match before it was imaged.

quantity	as published	SI
M87* ring diameter	$42 \pm 3\,\mu$as	$(2.04 \pm 0.15) \times 10^{-10}$ rad
M87* mass	$(6.5 \pm 0.7) \times 10^{9}\,M_{\odot}$	$(1.29 \pm 0.14) \times 10^{40}$ kg
M87* distance	16.8 Mpc	5.18×10^{23} m
Sgr A* ring diameter	$51.8 \pm 2.3\,\mu$as	$(2.51 \pm 0.11) \times 10^{-10}$ rad
Sgr A* predicted shadow (computed)	$53\,\mu$as	2.57×10^{-10} rad

Phenomenon 50.4 (A ring of light at event-horizon scale)**.** Very-long-baseline interferometry at 1.3 mm resolves the nucleus of M87 into an asymmetric ring of emission surrounding a central brightness depression. The ring diameter is 42 ± 3 microarcseconds, and the depression is deeper than any optically thin foreground could produce [2328]. The same technique applied to Sgr A* returns a ring of the diameter expected for the mass that the stellar orbits of Phenomenon 50.2 determine independently [2403]: one object weighed two ways, by Newtonian orbital dynamics and by the size of its photon capture cross-section, with concordant results.

Rests on Phenomenon 45.52 and Theorem 45.1.

Derivation. *Derives Phenomenon 50.4.*

The ring size measures a mass. From Chapter 45, the photon capture cross-section of a black hole has apparent diameter $2b_{\mathrm{c}} = 3\sqrt{3}\,r_{\mathrm{s}} \simeq 5.196\,r_{\mathrm{s}}$, and this is only weakly dependent on spin and viewing angle, so a measured angular diameter θ_{sh} at distance D gives

$$M = \frac{c^2}{2G} \cdot \frac{\theta_{\mathrm{sh}} D}{3\sqrt{3}}\,. \tag{50.4}$$

For M87, $\theta_{\mathrm{sh}} = 42$ microarcseconds, which is 2.04×10^{-10} in radian measure, and $D = 16.8\,\mathrm{Mpc} = 5.18 \times 10^{23}$ m, so $\theta_{\mathrm{sh}} D \approx 1.06 \times 10^{14}$ m and $r_{\mathrm{s}} \approx 2.03 \times 10^{13}$ m. Then

$$M = \frac{r_{\mathrm{s}} c^2}{2G} = \frac{2.03 \times 10^{13}\,\mathrm{m} \times 8.988 \times 10^{16}\,\mathrm{m^2/s^2}}{2 \times 6.674 \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}} \approx 1.4 \times 10^{40}\,\mathrm{kg}\,,$$

about $7 \times 10^{9}\,M_{\odot}$, in agreement with the $(6.5 \pm 0.7) \times 10^{9}\,M_{\odot}$ obtained from the full image-domain analysis [2328] — the few-per-cent excess of this crude inversion being the emission-ring calibration discussed above. The arithmetic makes plain what is being measured and what is being assumed: an angular size and a distance, converted into a mass by the single dimensionless number $3\sqrt{3}$, which general relativity supplies with no freedom at all. A different theory of the strong field would predict a different number, and the concordance with the orbital mass of Sgr A* is the test: for $M = 8.47 \times 10^{36}$ kg at $R_0 = 2.545 \times 10^{20}$ m,

$$\theta_{\mathrm{sh}} = \frac{3\sqrt{3} \times 2GM}{c^2 R_0} \approx 2.57 \times 10^{-10}\,\mathrm{rad}\,,$$

53 microarcseconds, against a measured ring of 51.8 ± 2.3 [2403]. □

50.4.4 Black-hole masses from gravitational waves

The fourth channel is treated in full in Chapter 47; the discovery event and its parameter table are Section 47.3. What the channel adds to *this* chapter is mass, and a different kind of certainty. The dynamical and imaging channels infer a compact mass from the motion of matter and light outside it; a gravitational-wave chirp is emitted by the masses themselves, and its phase evolution measures them with no astrophysical intermediary. GW150914's component masses of about $36\,M_\odot$ and $29\,M_\odot$ [2264] sit well above every X-ray binary mass, and the catalogue that followed extends the population to component masses of about $50\,M_\odot$ in the first two observing runs [2324] and beyond in later ones—black holes in a mass range that electromagnetic astronomy had never weighed. The remnant's ringdown, consistent in frequency and damping with the least-damped quasinormal mode of a Kerr black hole of the predicted mass and spin [2264], is the one direct probe of the *dynamical* strong-field geometry in this chapter: the other channels test the stationary field.

Phenomenon 50.5 (Black-hole masses from gravitational waves)**.** The first detected gravitational-wave transient, GW150914, is the inspiral, merger and ringdown of two compact objects of about $36\,M_\odot$ and $29\,M_\odot$, leaving a remnant of about $62\,M_\odot$ — roughly $3\,M_\odot c^2$ radiated away as gravitational waves within a fraction of a second [2264]. Neither body can be anything but a black hole: the signal sweeps up to frequencies at which the two are closer together than objects of those masses could be without merging. The catalogue that followed defines a population of black-hole masses and spins reaching well above the range accessible to the X-ray binaries of Phenomenon 50.1 [2324].

Rests on Phenomenon 27.35 and Theorem 45.1.

Derivation. *Derives Phenomenon 50.5.*

The compactness argument requires only Kepler's third law. Take the total mass $M = 70\,M_\odot = 1.39 \times 10^{32}\,\mathrm{kg}$ read off the inspiral, and the moment in the signal at which the gravitational-wave frequency reaches $f_{\mathrm{GW}} = 150\,\mathrm{Hz}$. A binary radiates principally at twice its orbital frequency (Chapter 46), so the orbital angular velocity there is $\omega = \pi f_{\mathrm{GW}} = 471/\mathrm{s}$, and two bodies orbiting with that angular velocity are separated by

$$a = \left(\frac{GM}{\omega^2}\right)^{1/3} = \left(\frac{6.674 \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)} \times 1.39 \times 10^{32}\,\mathrm{kg}}{(471/\mathrm{s})^2}\right)^{1/3} \approx 350\,\mathrm{km}\,.$$

Against this, the Schwarzschild radius of the total mass is

$$r_{\mathrm{s}} = \frac{2GM}{c^2} \approx 210\,\mathrm{km}\,,$$

so the two bodies were still orbiting freely at a separation of only about $1.7\,r_{\mathrm{s}}$. No neutron star can be involved: its radius, which Chapter 49 places near 12 km, is small enough, but its mass cannot approach $30\,M_\odot$. No main-sequence or degenerate star of any kind is remotely this compact at this mass. Only black holes remain. The Newtonian estimate is moreover conservative — the true relativistic separation at a given orbital frequency is smaller still — so the conclusion is not an artefact of applying Kepler's law outside its proper regime. □

50.5 Interpretation

What the ensemble establishes is best stated as three nested claims, in decreasing order of certainty.

First, and beyond reasonable doubt: *dark compact masses exist*, from about $5\,M_\odot$ in the X-ray binaries through $21\,M_\odot$ in Cygnus X-1 and $\sim 60\,M_\odot$ in the merger remnants to $4.3 \times 10^6\,M_\odot$ at the Galactic centre and $6.5 \times 10^9\,M_\odot$ in M87. The stellar-mass cases rest on

the Newtonian bound of Equation (50.2); the Galactic centre on Kepler's law applied to a closed stellar orbit. Each such mass exceeds every cold-equilibrium limit (Chapter 49), and for the S2 volume no non-collapsed alternative — a cluster of remnants, a ball of fermions — survives its own dynamics for the age of the Galaxy (Phenomenon 50.2).

Second, and directly measured: *the gravitational field of these objects is the field general relativity assigns them*, wherever it has been probed. The S2 orbit verifies the redshift-plus-Doppler prediction and the Schwarzschild precession at the ten-to-twenty per cent level, fifteen hundred gravitational radii out (Phenomenon 50.3). The photon-ring diameters of both imaged sources match Equation (50.4), whose only theoretical content is the strong-field number $3\sqrt{3}$ — and for Sgr A* the mass on the right-hand side was measured first, by an independent Newtonian method, so the image was a prediction cashed (Phenomenon 50.4). The GW150914 ringdown is consistent with the quasinormal spectrum of the Kerr geometry (Phenomenon 50.5). Nothing yet observed requires any parameter beyond the mass and spin of the Kerr family.

Third, and here the honesty matters most: *no material surface is seen*, where every known class of surface would show itself. This is the closest observation comes to the horizon, and it is an absence, argued quantitatively below (Phenomenon 50.6).

Against these stand the things the data do not establish. No observation reaches the horizon itself: an event horizon is a globally defined null surface, and no electromagnetic or gravitational-wave measurement of finite duration can certify one — what is measured is light and matter behaving, down to a few gravitational radii, exactly as if one were there, with no sign of anything else. The interior, the singularity theorems' conclusion, and any Planck-scale modification of the horizon are untested by everything in this chapter. Exotic horizonless alternatives — objects contrived to be as compact as a black hole without its causal structure — are constrained by the ringdown, by the shadow depth and by the surface arguments, but the most contrived of them are not excluded [2328]; the honest assessment of that frontier, and of what would settle it, is continued in Chapter 135. Interpretation throughout leans on the theory of Chapter 45 and Chapter 43.

Phenomenon 50.6 (No material surface)**.** Accreting neutron stars betray their surfaces twice over: the fuel accumulating on them ignites periodically in thermonuclear flashes — the type I X-ray bursts — and in quiescence they re-radiate the accretion energy deposited on them. The black-hole candidates of Phenomenon 50.1 do neither. No burst has ever been recorded from one, and at comparable accretion rates their quiescent luminosities fall orders of magnitude below those of the neutron-star systems [1941][2103]. The energy that arrives is not returned. That is what a horizon does and what a surface of any composition cannot do, and it is the closest the electromagnetic observations come to a direct test of the defining property of a black hole.

Rests on Theorem 19.32 and Proposition 45.18.

Derivation. *Derives Phenomenon 50.6.*

The comparison is an energy budget, and the budget is Newtonian bookkeeping plus one relativistic fact. Let matter fall steadily at rate $\dot{M}$ onto a compact object of mass M and radius R_*. Each unit of mass, arriving from far away, has given up its gravitational binding energy GM/R_* by the time it settles on the surface, whatever route it took; in steady state that energy can only leave as radiation. An object *with* a surface therefore shines with

$$L_{\mathrm{surf}} \simeq \frac{GM\dot{M}}{R_*} = \eta_* \dot{M} c^2 , \qquad \eta_* := \frac{GM}{R_* c^2} , \tag{50.5}$$

independently of how efficiently the infalling flow itself radiates. For a neutron star, with $M \approx 1.4\, M_\odot$ and $R_* \approx 12\,\mathrm{km}$ (Chapter 49), $\eta_* \approx 0.17$: a fixed seventeen per cent of the accreted rest energy must come back out. If instead the object has a horizon, nothing obliges

the flow to radiate at all: the luminosity is $\epsilon \dot{M} c^2$ with ϵ whatever radiative efficiency the flow manages *before* crossing, and the balance is carried through the horizon and gone.

The comparison becomes decisive in quiescence, precisely because the accretion there is inefficient. At low accretion rates the inflow is a hot, tenuous, advection-dominated flow: the ion gas is too dilute to transfer its heat to the radiating electrons within an infall time, so it carries its thermal energy inward, and $\epsilon \lesssim 10^{-2}$ [1941]. The predicted contrast between two otherwise similar transients, one with a horizon and one with a surface, is then

$$\frac{L_{\rm hole}}{L_{\rm surf}} \simeq \frac{\epsilon}{\eta_*} \lesssim \frac{10^{-2}}{0.17} \approx 6 \times 10^{-2} \,,$$

one to two orders of magnitude, before allowing for the gravitational redshift of the surface emission, which only deepens it. This is what is observed: at matched orbital periods — the available proxy for matched quiescent accretion rate — the black-hole transients sit roughly a factor of one hundred below the neutron-star transients in quiescent luminosity [1941][2103].

The burst argument is the same bookkeeping applied to matter rather than energy. On a surface, accreted hydrogen and helium accumulate, compress, and heat until the layer ignites; the burning is thermally unstable in a thin shell, so it runs away, and the neutron-star systems duly flash — thousands of type I bursts stand recorded. A horizon stores no fuel. No type I burst has ever been observed from any system whose mass function exceeds the neutron-star bound [2103].

Two limits of the argument are stated as plainly as the argument. It is differential — it compares two source classes assuming the same kind of quiescent flow in both, so it inherits the advection-dominated model's assumptions, though not its details: any flow radiating a fraction $\epsilon \ll \eta_*$ before arrival gives the same contrast. And it bounds what a surface would emit; it does not observe the horizon. It is evidence of absence obtained where presence was obligatory, which is the strongest statement an electromagnetic observation can make about a horizon. □

50.6 Modern repetitions and precision

Every one of the four channels is a running programme, and each has a stated next decimal place.

At the Galactic centre, GRAVITY continues to track S2 away from pericentre, where the precession signal accumulates, and has added fainter stars on shorter orbits to the fit; each pericentre passage — S2's next falls in 2034 — multiplies the relativistic leverage [2347]. The two-programme redundancy that checked the redshift [2295, 2336] remains the model: the Keck and VLT groups share neither instruments nor analysis chains. The targets beyond the first post-Newtonian order are the spin of Sgr A* through frame dragging of the innermost orbits, and any deviation of the enclosed mass from a point — both within reach of the existing astrometry extended by a decade (Chapter 45 supplies the predictions).

The Event Horizon Telescope has repeated and extended its campaigns: the polarized image of M87* [2372] turned the ring into a magnetometer, and the Sgr A* result [2403] showed that even a source varying on its minutes-long horizon-crossing time can be imaged. Precision here scales directly with the array: every added station fills the aperture, and a shorter observing wavelength tightens the beam as λ/D — the path to resolving the photon ring's substructure, time-resolved imaging of the flow, and a sharper shadow-diameter test than the current few-per-cent agreement of Table 50.3.

In the X-ray binaries the frontier is spin and population. The relativistically broadened iron line [1911] and the disc-continuum method, both calibrated on the innermost-stable-orbit physics of Chapter 45, are applied across the twenty-odd dynamically confirmed systems

[2103], while VLBI parallaxes of the kind that re-weighed Cygnus X-1 [2387] remove the distance systematics that dominated a generation of mass estimates.

The gravitational-wave catalogue grows by design: from one event [2264] to ten binary black holes in the first two observing runs [2324] to, at the time of writing, of order a hundred (Chapter 47) — a population experiment on black-hole masses and spins, and, as the loudest events accumulate, a spectroscopy of the ringdown: more than one quasinormal mode measured in a single remnant would test the Kerr no-hair property, not merely be consistent with it. The projected sensitivities of the next-generation interferometers are cross-referenced in Chapter 47.

50.7 Primary references

The identification of Cygnus X-1: [1466, 1488]; the modern orbital solution and mass: [2178, 2387]; the transient population and its phenomenology: [1850][2103]; the disc theory and the iron-line diagnostic: [1512, 1911]. The birth of the field: [1225] for X-ray astronomy, [1265, 1403, 1517] for the quasar argument and the Galactic-centre source, [37] for the prehistory.

The Galactic-centre orbits: [2038] for the first orbit, [2125, 2146] for the two long-term programmes, reviewed in [2156]; the interferometric instrument: [2279]; the relativistic tests: [2295, 2336, 2347].

Event-horizon-scale imaging: [2328] (M87*), [2372] (polarization), [2403] (Sgr A*); the shadow prediction: [1988].

Gravitational waves: [2264][2324], with the full experimental record in Chapter 47 [2243].

The horizon-versus-surface argument: [1941]. The theoretical prehistory: [440, 844, 845, 1259, 1319].

Chapter 51

Experiment: The Cosmic Microwave Background

Tests Phenomenon 48.12 and Equation (48.24).

Assuming Equations (38.38) *and* (68.5).

Discovery and precision spectroscopy of the relic radiation: the blackbody spectrum, the anisotropies, and the parameters they fix in the cosmology of Chapter 48. The cosmic microwave background is the oldest light there is and the best-measured blackbody there is: found accidentally by Penzias and Wilson [1320] and identified at once as the relic of a hot early universe [1305], it now carries, in its spectrum [1884, 1922], its anisotropies [1858, 2213, 2348] and its polarization [2031], the tightest constraints in all of cosmology.

This chapter is the evidence file for Chapter 48: the hot-big-bang prediction under test, the radiometers and spectrophotometers that tested it, and the numbers they returned. The blackbody physics being tested is that of Chapter 68; the boost kinematics that reads the dipole is that of Chapter 38; and the general-relativistic perturbation theory that turns the anisotropy spectra into cosmological parameters belongs to Chapter 48 and is quoted here, never silently rederived. What this chapter owns is the measuring: apparatus, procedure, the numbers with their uncertainties, and what they do and do not establish.

51.1 Historical context and the prediction under test

51.1.1 The hot-big-bang prediction

The prediction is a by-product of nucleosynthesis. In 1948 Alpher, Bethe and Gamow proposed that the chemical elements were assembled by neutron capture in a hot, dense early universe [905]; the calculation works only if the baryons are outnumbered by thermal photons by a large, fixed factor, because it is the radiation that sets the temperature–time relation through which the reactions freeze out (Chapter 48). Later the same year Alpher and Herman followed that radiation forward: expansion cools a thermal photon gas as $T \propto 1/a$ while preserving its Planck form — the statement proved in Chapter 48 — so the fireball must survive to the present as a cold blackbody filling all of space, at a temperature they estimated as about 5 K [906].

The prediction is threefold, and each part is separately testable. The relic must have a *Planck spectrum*, because it was thermalized when the universe was dense enough to reach equilibrium and free expansion cannot un-thermalize it; any measured distortion is a record of energy released since. It must be *isotropic* to high accuracy, because it fills the universe rather than emanating from anything in it. And it must carry small *anisotropies*, because the structure that now exists had to grow from seed inhomogeneities that were present at decoupling and must have left their imprint on the light released then. Spectrum, isotropy

and anisotropy are the three sections of the observations below, in that order.

Nobody searched for it. The prediction lay dormant for seventeen years, unknown to the radio astronomers who by the late 1950s had the instruments to test it. In 1964 Dicke's group at Princeton rederived the hot early phase independently and set about the measurement deliberately: Roll and Wilkinson were building a purpose-designed radiometer when word arrived that a pair of radio astronomers fifty kilometres away, calibrating a communications antenna, had already found an excess they could not remove. The two papers were published back to back — the measurement [1320] and its interpretation [1305] — and the interpretation paper's senior author was the same Dicke whose switched radiometer, invented in 1946, is the instrument class behind nearly every measurement in this chapter [888].

51.1.2 Precursors

The radiation had in fact been detected before it was predicted. In 1941 McKellar analysed the interstellar absorption lines of the CN radical seen against hot stars, among them ζ Ophiuchi. The first rotationally excited level of CN lies one photon of wavelength 2.64 mm above the ground state, and the ratio of the absorption from the two levels measures how the molecules are distributed between them; McKellar translated the measured ratio into a "rotational temperature" of interstellar space of about 2.3 K [860]. The molecules are thermometers immersed in the millimetre-wave radiation field, and they were reporting the cosmic microwave background — but with no prediction to connect it to, the number stood in the literature as a curiosity, and only after 1965 could it be read for what it is.

Remark 51.1 (A signal without a theory). The CN temperature is a canonical case of the general lesson: a measurement with no theoretical frame attached does not function as a discovery, however accurate it is. McKellar's 2.3 K sat unexplained for twenty-four years — within thirty per cent of the modern value of Phenomenon 51.3 — while the complementary failure ran in parallel: from 1948 to 1965 the Alpher–Herman prediction [906] was a theory without a signal, unknown to the observers who could have tested it. The discovery of 1965 happened at the moment the two finally met [1305, 1320], and neither half would have sufficed alone.

51.2 Apparatus

51.2.1 Radiometers and the Holmdel horn

Every temperature in this chapter is an *antenna temperature*: the temperature of a matched resistive load that would deliver the same power to the receiver, a linear measure of received power in the Rayleigh–Jeans regime of Chapter 68. The obstacle to measuring a few kelvin of sky is that the receiver itself contributes hundreds or thousands of kelvin of noise, and its gain drifts. Dicke's solution of 1946 defines the instrument class: switch the receiver input many times per second between the sky and a stable reference load, and detect synchronously with the switch, so that receiver noise and gain drift, common to both half-cycles, cancel in the difference [888]. The residual fluctuation of an ideal radiometer of system temperature T_{sys}, predetection bandwidth $\Delta\nu$ and integration time τ is

$$\Delta T_{\mathrm{rms}} \simeq \frac{T_{\mathrm{sys}}}{\sqrt{\Delta\nu\,\tau}}\,, \tag{51.1}$$

within a factor of order unity fixed by the switching scheme [888]. The equation is the budget of the whole field: a receiver of $T_{\mathrm{sys}} = 100\,\mathrm{K}$ and $\Delta\nu = 100\,\mathrm{MHz}$ resolves — computed here from Equation (51.1) — about 10 mK in one second, so the millikelvin dipole of Section 51.4.3

needs minutes, and the microkelvin anisotropies of Section 51.4.4 need years of integration and thousands of detectors.

The discovery instrument was not built for cosmology. The Holmdel antenna at Crawford Hill is a 6 m horn-reflector erected for the Echo and Telstar satellite links, chosen for exactly the property a background measurement needs: the horn geometry shields the aperture from the ground, so the response in the back and side lobes is far below that of a paraboloid. Penzias and Wilson equipped it with a ruby travelling-wave maser preamplifier at 4.08 GHz — a wavelength of 7.35 cm, computed here — and, crucially, with a reference termination immersed in liquid helium, giving the switched radiometer an absolute cold standard against which the sky could be measured rather than merely compared [1320]. A year of systematic elimination followed: the atmosphere was measured by tipping the beam and fitting the secant law in zenith angle, the ohmic loss of the antenna throat was bounded, the response away from the beam was checked against the Sun and the Galaxy, the seams were taped, and the pair of pigeons roosting in the throat was evicted and their deposits removed. The excess survived everything, in every direction, in every season [1320].

51.2.2 Satellites: COBE, WMAP, Planck

The atmosphere emits and absorbs strongly across the millimetre band, so the two precision programmes — the spectrum and the anisotropy — both culminated in space.

COBE (launched 1989) carried both instruments of record. FIRAS is a polarizing Michelson interferometer whose two inputs are the sky and an internal reference blackbody: the interferogram measures the *difference* between the two spectra, and the reference is servoed until that difference nearly vanishes, so the instrument spends its dynamic range on the deviation rather than on the signal. An external calibrator — a re-entrant absorbing cone that can swing over the horn and fill the beam entirely — provides the absolute standard; the accuracy of Section 51.4.2 is the accuracy of that cone. The spectral band of the monopole analysis runs from about 60 GHz to 630 GHz with a 7° beam [1834, 1884, 1922]. The companion DMR consists of differential microwave radiometers at 31.5 GHz, 53 GHz and 90 GHz, each comparing two 7° beams 60° apart on the sky — a Dicke radiometer whose reference load is another patch of sky [1858].

WMAP (2001–2010) took the differential principle to the second Lagrange point, where Sun, Earth and Moon sit together in a small solid angle behind the shields. Pseudo-correlation differencing assemblies at five frequencies from 23 GHz to 94 GHz measured temperature differences between beams separated by about 141°, with angular resolution reaching about 0.2° in the highest band; nine years of data give the full-sky maps and power spectra of [2213]. Planck (2009–2013), from the same orbit, spans 30 GHz to 857 GHz with two instrument families: the LFI radiometers at 30, 44 and 70 GHz, and the HFI bolometers from 100 GHz upward, cooled to 0.1 K and reaching about 0.08° resolution — enough to map the temperature anisotropy over the whole acoustic range [2348].

51.2.3 Balloons and ground stations

Between the discovery and the satellites, and alongside them since, the sub-orbital tier has repeatedly been first. The U-2 aircraft radiometers of 1976–77, flying above most of the atmosphere, found the dipole [1610]. BOOMERanG, a bolometric telescope carried around Antarctica for about ten days at an altitude near 38 km by a long-duration balloon, mapped a patch of sky at frequencies from 90 GHz to 400 GHz with sub-degree resolution and resolved the first acoustic peak [1984]. DASI, a thirteen-element interferometer at 26–36 GHz sited at the South Pole — where the precipitable water vapour is among the lowest on Earth — made the first detection of polarization [2031]. The modern ground tier is the arcminute programme: the 10 m South Pole Telescope and the 6 m Atacama Cosmology

Telescope, with kilopixel arrays of transition-edge bolometers, extend the spectra into the damping tail and supply the secondary-anisotropy science of Section 51.4.7 [2219, 2351]; and the BICEP/Keck small-aperture polarimeters — refractors of tens of centimetres, built for the degree scale where a primordial B mode would peak — carry the search of Section 51.4.6 [2227, 2369].

51.3 Procedure

Four protocols cover every result reported below.

Absolute spectrophotometry. The sky is nulled against a calibrated blackbody and the residual spectrum is measured interferometrically; periodically the external calibrator replaces the sky entirely, transferring the laboratory standard to the celestial measurement. The result is a spectrum in absolute units at each frequency, from which one temperature and the distortion parameters of Section 51.4.2 are fitted [1884, 1922].

Differential mapping. Anisotropy instruments never measure the monopole: they measure temperature *differences* between sky directions, following Equation (51.1), and scan so that every pair of pixels is linked by many differences along many paths. The map is then the least-squares solution of the linear model relating the time-ordered differences to the pixel temperatures — an overdetermined system whose redundancy is also the systematics check, since a miscalibrated channel shows up as inconsistency between scan paths. The estimation machinery is that of Chapter 11 [1858, 2213, 2348].

Component separation. The microwave sky is not only the background: the Galaxy radiates synchrotron and free–free emission, both falling steeply with frequency, and thermal dust emission, rising with frequency. The background alone has the achromatic signature of a temperature perturbation of a blackbody — the same ΔT at every frequency when expressed as thermodynamic temperature. Multi-frequency coverage is therefore not a luxury but the separation variable itself; the Galactic plane is masked, and the components are fitted and removed pixel by pixel [2213, 2348].

Harmonic analysis and likelihood. A full-sky anisotropy map is expanded in the spherical harmonics of Chapter 9,

$$\frac{\Delta T}{T_0}(\hat{\boldsymbol{n}}) = \sum_{\ell \geq 1} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\hat{\boldsymbol{n}}) \,, \qquad C_\ell := \left\langle |a_{\ell m}|^2 \right\rangle \,, \tag{51.2}$$

and the cosmological information is carried by the angular power spectrum C_ℓ (and its polarized counterparts C_ℓ^{TE}, C_ℓ^{EE}, C_ℓ^{BB}), because statistical isotropy makes the individual $a_{\ell m}$ zero-mean random variables with a common variance at each ℓ. That same statement fixes the floor of the whole enterprise. The unbiased estimator averages the $2\ell+1$ independent squared moduli available at multipole ℓ, and for Gaussian fluctuations (Chapter 11) the relative scatter of such an average is

$$\frac{\Delta C_\ell}{C_\ell} = \sqrt{\frac{2}{2\ell+1}} \,, \tag{51.3}$$

the *cosmic variance*: we observe one sky, and no instrument, however perfect, can beat the sample size. At $\ell = 2$ the floor is 63 per cent — computed here from Equation (51.3) — which is why the quadrupole is permanently uncertain while the acoustic peaks, with thousands of modes each, are measured to parts in a thousand. Parameters are then inferred by likelihood fits of model spectra to the measured C_ℓ, per Chapter 11 [2213, 2348].

51.4 Observations and data

Throughout this section, a number with an uncertainty is quoted from the cited primary paper; where a number is instead computed here from quoted inputs, the text says so.

51.4.1 Discovery

Penzias and Wilson measured the total antenna temperature at the zenith at 4.08 GHz and subtracted everything they could account for. The budget is the experiment, and it is short:

Table 51.1: The Holmdel antenna-temperature budget at 4.08 GHz, as reported by Penzias and Wilson [1320]. All entries are quoted from the paper. The remainder on the last line, isotropic, unpolarized and stable through the seasons, is the cosmic microwave background.

Contribution	Antenna temperature
Total zenith measurement	6.7 K
Atmospheric absorption	(2.3 ± 0.3) K
Ohmic losses of the antenna	(0.8 ± 0.4) K
Back-lobe ground pickup	< 0.1 K
Unaccounted excess	(3.5 ± 1.0) K

Phenomenon 51.2 (An isotropic microwave excess)**.** A horn-reflector antenna operating at 4.08 GHz, after every identifiable contribution to the system noise had been measured and subtracted, retained an excess antenna temperature of about 3.5 K, uncertain by about 1.0 K. The excess was isotropic within the sensitivity of the instrument, unpolarized and unchanged with the season, so it could be neither a terrestrial source, nor the Galaxy, nor any known population of discrete extragalactic sources [1320]. The companion paper published alongside identified it as relic radiation from a hot, dense early phase of the universe [1305].
Rests on Equation (68.5)*, Phenomenon 68.5, and Definition 48.2.*

Derivation. *Derives Phenomenon 51.2.*

Two steps connect the reported number to a physical statement about the universe. The first is that an *antenna* temperature is a measure of received power — the temperature of a matched resistive load that would deliver the same power to the receiver — so equating it with the thermodynamic temperature of a blackbody filling the beam presupposes the Rayleigh–Jeans limit of the Planck law of Chapter 68, in which the spectral radiance is linear in T. That limit requires $h\nu \ll k_{\mathrm{B}}T$, and at the Holmdel frequency

$$\frac{h\nu}{k_{\mathrm{B}}T} = \frac{6.626 \times 10^{-34}\,\mathrm{J\,s} \times 4.08 \times 10^{9}\,\mathrm{Hz}}{1.381 \times 10^{-23}\,\mathrm{J/K} \times 3\,\mathrm{K}} \approx 0.065\,,$$

so the departure from linearity is a few per cent, well inside the quoted uncertainty. The identification is therefore legitimate at this frequency — and would not have been at, say, 100 GHz, which is why the discovery had to wait for centimetre-wave receivers of very low noise.

The second step is that isotropy together with the absence of polarization already excludes the alternatives. A Galactic origin would concentrate the signal toward the plane; an unresolved population of extragalactic radio sources would carry their synchrotron spectrum, falling steeply with frequency rather than lying flat in T; and any terrestrial, atmospheric or solar contribution would vary with elevation and with the season. All three were tested and excluded, and what remains must fill space uniformly — which is a cosmological statement, not an astronomical one. □

51.4.2 The blackbody spectrum

One measurement at one frequency establishes an excess; only the spectrum can establish a relic. The prediction under test — the thermal form preserved by expansion, Chapter 48 — was settled by FIRAS in stages, each tightening the last:

Table 51.2: The spectrum campaigns. Every value is quoted from the cited paper. The deviation column bounds the departure of the measured spectrum from a Planck curve, as a fraction of the peak brightness.

Measurement	Year	Temperature T_0	Deviation
FIRAS, first nine minutes [1834]	1990	$(2.735 \pm 0.060)\,\mathrm{K}$	—
FIRAS, calibrated [1884]	1994	$(2.726 \pm 0.010)\,\mathrm{K}$	$< 3 \times 10^{-4}$
FIRAS, full data set [1922]	1996	$(2.728 \pm 0.004)\,\mathrm{K}$	$< 5 \times 10^{-5}$ (rms)
Combined, recalibrated [2144]	2009	$(2.725\,48 \pm 0.000\,57)\,\mathrm{K}$	—

The full-data-set analysis also bounds the two canonical distortion parameters: a chemical potential $|\mu| < 9 \times 10^{-5}$ and a Compton parameter $|y| < 1.5 \times 10^{-5}$, both at 95 per cent confidence [1922]. Each is a null with content, because each would be the fossil of energy released into the radiation: a μ distortion records injection early enough for scattering to redistribute photons but too late to create them, a y distortion records injection later still, so the limits constrain the entire energy-release history of the universe since thermalization ceased (Chapter 48). The 2009 value combines the FIRAS data, recalibrated against later precision dipole measurements, with the other absolute measurements, and it is the temperature used throughout this book [2144].

Phenomenon 51.3 (The most perfect blackbody ever measured)**.** The spectrum of the background radiation is a Planck spectrum. An absolute spectrophotometer flown above the atmosphere, nulling the sky against an on-board blackbody calibrator, found no deviation from the Planck form larger than about fifty parts per million of the peak brightness anywhere in the band it covered [1884, 1922]. The temperature is $T_0 = 2.725\,48\,\mathrm{K}$, known to $0.000\,57\,\mathrm{K}$ [2144]. No source constructed in a laboratory has ever been brought this close to thermal equilibrium, and the fit has no free parameter but the one temperature.

Rests on Equation (68.5)*, Phenomenon 68.5, and Definition 48.2.*

Derivation. Derives Phenomenon 51.3.

Three numbers characterize the measured spectrum, and each follows from T_0 and the Planck law of Chapter 68 alone. The wavelength of maximum spectral radiance per unit wavelength is given by the Wien displacement law,

$$\lambda_{\max} = \frac{b}{T_0} = \frac{2.898 \times 10^{-3}\,\mathrm{m\,K}}{2.725\,48\,\mathrm{K}} \approx 1.06\,\mathrm{mm}\,,$$

so the radiation peaks in the microwave band — which is why a radio receiver found it, and why measuring the spectrum required going above the atmospheric water lines. Per unit *frequency* the maximum instead lies at $\nu_{\max} = 2.821\,k_{\mathrm{B}}T_0/h \approx 160\,\mathrm{GHz}$; the two maxima disagree because a spectral density is defined with respect to a chosen variable, and neither is "the" peak of the radiation.

The energy density follows from the Stefan–Boltzmann law,

$$u = aT_0^4 \approx 4.17 \times 10^{-14}\,\mathrm{J/m^3}\,,$$

with $a = 7.566 \times 10^{-16}\,\mathrm{J\,m^{-3}\,K^{-4}}$ the radiation constant, and dividing by the mean photon energy $\approx 2.70\,k_{\mathrm{B}}T_0$ gives a photon number density of about 4.1×10^8 per cubic metre. That is roughly 10^9 photons for every baryon in the universe, and that ratio is the single free parameter of the nucleosynthesis calculation of Chapter 48 — so the spectrum measured here and the light-element abundances measured there are two determinations of one number. □

51.4.3 The dipole

The first anisotropy found is not primordial. Radiometers flown on a U-2 aircraft in 1976–77 detected a smooth cosine pattern across the sky [1610], since measured to four significant figures by Planck [2352]:

Table 51.3: The dipole anisotropy. The inferred speed on the last line is computed here from the quoted Planck amplitude and the monopole temperature, via the boost formula Equation (51.4).

Quantity	Value	Status
Amplitude, U-2 discovery [1610]	$(3.5 \pm 0.6)\,\mathrm{mK}$	quoted
Amplitude, Planck [2352]	$(3.3621 \pm 0.0010)\,\mathrm{mK}$	quoted
Direction, Galactic [2352]	$(l, b) \approx (264.02^\circ, 48.25^\circ)$	quoted
Inferred solar-system speed	$\approx 370\,\mathrm{km/s}$	computed here

Phenomenon 51.4 (The dipole)**.** The background is not exactly isotropic. Superposed on the monopole is a pure dipole of amplitude about 3.36 mK, first detected from a U-2 aircraft [1610] and now measured to four significant figures [2352]. Its interpretation is kinematic: the solar system moves at about 370 km/s with respect to the frame in which the radiation is isotropic. This is the only large anisotropy of the microwave sky that is not primordial, and it must be removed before the intrinsic fluctuations of Section 51.4.4 can be seen at all — they are two orders of magnitude smaller.

Rests on Theorem 38.4, Proposition 38.25, and Phenomenon 68.5.

Derivation. *Derives Phenomenon 51.4.*

Let the radiation be an isotropic blackbody at temperature T_0 in some frame, and let an observer move through it with speed $v = \beta c$. Radiation arriving from the direction making an angle θ with the motion is Doppler shifted by the factor $[\gamma\,(1 - \beta\cos\theta)]^{-1}$ of Chapter 38. Every frequency in the spectrum is multiplied by that same factor, and the Planck spectrum of Chapter 68 depends on frequency only through the combination $h\nu/\,(k_{\mathrm{B}}T)$; the boosted spectrum is therefore again Planckian in every direction, with a direction-dependent temperature

$$T\,(\theta) = \frac{T_0}{\gamma\,(1 - \beta\cos\theta)} = T_0\,(1 + \beta\cos\theta) + O\!\left(\beta^2\right)\,. \tag{51.4}$$

To first order in β this is exactly a dipole of amplitude $\Delta T = \beta T_0$, with no quadrupole or higher multipole at all — so the observed pattern being a clean dipole is itself evidence for the kinematic reading rather than an assumption of it. Inverting,

$$v = c\,\frac{\Delta T}{T_0} = 2.998 \times 10^8\,\mathrm{m/s} \times \frac{3.36 \times 10^{-3}\,\mathrm{K}}{2.725\,\mathrm{K}} \approx 3.7 \times 10^5\,\mathrm{m/s}\,,$$

that is about 370 km/s. The second-order terms of Equation (51.4) generate a kinematic quadrupole smaller than the dipole by a further factor $\beta \approx 10^{-3}$, which is comparable to the primordial signal and is accounted for explicitly in modern analyses. The same boost also aberrates the arrival directions by the aberration formula Equation (38.38) of Chapter 38, displacing the whole anisotropy pattern by an angle of order $\beta \approx 1.2 \times 10^{-3}$ radians — about four minutes of arc, computed here — toward the apex. □

The frame in which the dipole vanishes is physically distinguished: it is the rest frame of the radiation, and to high accuracy the rest frame of the average matter distribution. That this frame exists is a cosmological statement — the universe has a preferred state of motion at each point, defined by its contents — and its significance for the cosmological principle is discussed in Chapter 48. It contradicts nothing in Chapter 38: the laws are Lorentz-invariant; the matter distribution is not.

51.4.4 Primary anisotropies

Underneath the dipole the sky is smooth to a few parts in 10^5, and finding the intrinsic structure took twenty-seven years of steadily improving null results. The COBE DMR maps of 1992 finally resolved it [1858]:

Table 51.4: The DMR discovery numbers [1858]. The relative amplitude on the last line is computed here from the quoted rms and the monopole temperature.

Quantity	Value	Status
Sky rms, 10° smoothing	$(30 \pm 5)\,\mu\mathrm{K}$	quoted
Quadrupole amplitude	$(13 \pm 4)\,\mu\mathrm{K}$	quoted
Relative amplitude $\Delta T/T_0$	$\approx 1.1 \times 10^{-5}$	computed here

Phenomenon 51.5 (Primary anisotropies at one part in a hundred thousand)**.** Once the dipole of Phenomenon 51.4 is subtracted and the Galactic plane masked, the background retains intrinsic temperature fluctuations of relative amplitude $\Delta T/T \sim 10^{-5}$. They were first detected on angular scales of about ten degrees by the differential radiometers of COBE [1858], twenty-seven years after the discovery of the monopole and after a long series of non-detections that had begun to constrain the models; they are now mapped over the whole sky to arcminute resolution [2213, 2348]. On the largest angular scales the fluctuations track the gravitational potential on the last-scattering surface, the relation worked out by Sachs and Wolfe long before there was a measurement to apply it to [1357]. These are the seed inhomogeneities from which all later structure grew, and their smallness is what makes linear perturbation theory sufficient for the epoch.

Rests on Equations (42.7) *and* (51.2) *and Phenomenon 42.5.*

Derivation. *Derives Phenomenon 51.5.*

Two things are derived here: what the reported numbers mean, and why a temperature fluctuation measures a gravitational potential. The statistical frame is Equation (51.2): a map smoothed to an angular scale θ retains the multipoles up to $\ell \sim \pi/\theta$, so the 10° smoothing of Table 51.4 reads the sky at $\ell \lesssim 20$ — the largest scales, causally disconnected at decoupling. Dividing the quoted rms by the monopole,

$$\frac{\Delta T}{T_0} \approx \frac{30\,\mu\mathrm{K}}{2.725\,\mathrm{K}} \approx 1.1 \times 10^{-5}\,,$$

computed here from quoted inputs.

Why this measures gravity: a photon that last scatters at the bottom of a potential well must climb out, and arrives redshifted by the gravitational redshift of Equation (42.7), written for a general weak potential as $\Delta\nu/\nu = -\Delta\Phi/c^2$. Since a uniform fractional shift of every frequency maps a Planck spectrum to a Planck spectrum — the same argument as in the dipole derivation above — the observed effect is a temperature perturbation $\Delta T/T = \Delta\nu/\nu$: an overdense region, with $\Phi < 0$, is seen *cold*. The naive answer $\Delta T/T = |\Phi|/c^2$ overcounts, however, because the same perturbation also alters the plasma density, and hence the temperature, at the point of emission, and alters the rate of the clocks there. The general-relativistic accounting of all three pieces — carried out first by Sachs and Wolfe [1357], and belonging in this book to Chapter 48 — leaves, for adiabatic perturbations on scales larger than the acoustic scale of Section 51.4.5, the net result

$$\frac{\Delta T}{T_0} = \frac{\Phi}{3c^2}\,, \tag{51.5}$$

one third of the naive redshift. Read through Equation (51.5), the DMR amplitude says $\Phi/c^2 \approx 3 \times 10^{-5}$ on the last-scattering surface, computed here: the metric of the universe at

decoupling departed from homogeneity by a few parts in 10^5. That smallness is the licence for every linear calculation in Chapter 48, and the potentials themselves are the seeds whose growth into the present galaxy distribution is followed there. □

51.4.5 The acoustic peaks

Resolution below a degree turned the anisotropy from an amplitude into a spectroscopy. The peak structure was predicted three decades before it was resolved [1431, 1434], with photon diffusion cutting the spectrum off at small scales [1383]; BOOMERanG resolved the first peak in 2000 [1984], and the satellite spectra now trace the harmonics across the whole acoustic range [2213, 2348].

Table 51.5: The acoustic scale. Quoted values are from the cited papers; computed values follow from the quoted ones via Equation (51.7).

Quantity	Value	Status
First-peak position, BOOMERanG [1984]	$\ell_1 = 197 \pm 6$	quoted
Acoustic angular scale $100\theta_*$ [2348]	1.0411 ± 0.0003	quoted
Comoving sound horizon r_* [2348]	$(144.4 \pm 0.3)\,\mathrm{Mpc}$	quoted
Comoving distance to last scattering	$\approx 13.9\,\mathrm{Gpc}$	computed here
Acoustic multipole $\ell_A = \pi/\theta_*$	≈ 302	computed here
Spatial curvature Ω_K [2348]	0.001 ± 0.002	quoted

Phenomenon 51.6 (Acoustic peaks, and a spatially flat universe)**.** The angular power spectrum of the temperature anisotropy is not featureless: it carries a series of harmonic peaks, the first and highest near multipole $\ell \approx 220$, with the spectrum cut off at high ℓ by photon diffusion out of the shortest wavelengths [1984, 2348][1383]. These are the standing acoustic oscillations of the photon-baryon plasma before decoupling, predicted three decades before they were resolved [1431, 1434]. Their positions and relative heights measure geometry and contents at once: the angular scale of the first peak fixes the spatial curvature, which is consistent with zero with $|\Omega_K|$ bounded below about 0.005; the alternation of odd and even peak heights weighs the baryons; and the envelope constrains the total matter density [2348]. The same sound horizon, seen at low redshift in the galaxy distribution rather than projected on the last-scattering surface, is the acoustic ruler of Chapter 48.

Rests on Theorems 48.3 and 48.5 and Phenomenon 65.3.

Derivation. *Derives Phenomenon 51.6.*

What is derivable at this chapter's level is the ruler and its projection; the dynamics of the oscillations themselves — the driven photon-baryon oscillator, the phase of each harmonic, the alternation of the peak heights and the damping envelope — is the business of Chapter 48, resting on [1383, 1431, 1434].

The ruler. Before decoupling, Thomson scattering (Chapter 65) locks photons and baryons into a single fluid whose pressure is the radiation pressure $p = u/3$ of Chapter 68. For radiation alone the sound speed is $c_s = c\sqrt{\partial p/\partial u} = c/\sqrt{3} \approx 1.73 \times 10^5\,\mathrm{km/s}$, computed here; baryons add inertia without adding pressure and slow the wave, so

$$c_s = \frac{c}{\sqrt{3\,(1+R)}}\,, \qquad R := \frac{3\rho_b}{4\rho_\gamma}\,, \qquad r_s = \int_0^{t_*} \frac{c_s}{a\,(t)}\,\mathrm{d}t\,, \tag{51.6}$$

where r_s, the comoving distance sound travels before the photons are released at t_*, is the farthest any pressure wave ever got: the *sound horizon*. Evaluating the integral needs the full expansion and recombination history and is carried out in Chapter 48; the fitted value is $r_* = (144.4 \pm 0.3)\,\mathrm{Mpc}$ [2348].

The projection. A comoving length r_* on the last-scattering surface, seen across the comoving distance D_M in a spatially flat geometry, subtends

$$\theta_* = \frac{r_*}{D_M}\,, \qquad \ell_A = \frac{\pi}{\theta_*}\,, \tag{51.7}$$

the second relation because a spherical harmonic of degree ℓ oscillates with angular period $2\pi/\ell$ along a great circle (Chapter 9), so structure of angular size θ concentrates its power near $\ell \approx \pi/\theta$. The measured acoustic scale is $100\theta_* = 1.0411 \pm 0.0003$ [2348]; combining the two quoted numbers gives, computed here,

$$D_M = \frac{r_*}{\theta_*} \approx 13.9\,\mathrm{Gpc}\,, \qquad \ell_A = \frac{\pi}{\theta_*} \approx 302\,.$$

The first peak sits not at ℓ_A but below it, at $\ell_1 \approx 220 \approx 0.73\,\ell_A$: the oscillator is driven by the decaying gravitational potentials, which advances the phase of every harmonic by a calculable shift (Chapter 48). The BOOMERanG value $\ell_1 = 197 \pm 6$ [1984] and the satellite value near 220 [2213, 2348] bracket the same physics at very different calibrations and sky coverage.

The curvature. Equation (51.7) as written is the flat-space rule that angular size is length over distance. In a positively curved space light rays converge, and the same ruler at the same distance subtends a *larger* angle, pushing the peaks to lower ℓ; negative curvature does the reverse. With r_* fixed by the pre-decoupling plasma physics and the distance fixed by the measured expansion history, the observed θ_* therefore pins the curvature; the fitted result is $\Omega_K = 0.001 \pm 0.002$ [2348]. Space, on the largest measurable scales, is Euclidean to the accuracy of the measurement — one of the sharpest geometric statements in physics, and an input assumption nowhere. □

51.4.6 Polarization

The background is polarized, and its polarization was found where the atmosphere is driest. The DASI interferometer at the South Pole detected linear polarization at the microkelvin level in 2002, at a significance of 4.9σ [2031]; the correlation between temperature and polarization was then traced across the acoustic range by WMAP and Planck [2213, 2348]; B modes generated by gravitational lensing were first detected with the South Pole Telescope [2219]; and the degree-scale search for a primordial B mode produced the episode recorded in Remark 51.8, with the current bound $r < 0.036$ at 95 per cent confidence [2369].

Phenomenon 51.7 (E-mode polarization, and no primordial B mode)**.** The background is linearly polarized at the level of a few microkelvin — about a tenth of the temperature anisotropy — as first detected by an interferometer at the South Pole [2031]. The polarization field decomposes into a curl-free E mode and a divergence-free B mode. The observed E mode is correlated with the temperature field in the sense and with the phase relation that the acoustic picture of Phenomenon 51.6 requires, which is a sharp consistency test of that picture rather than a new parameter [2213, 2348]; and B modes generated by the gravitational lensing of E modes along the line of sight have been detected [2219]. No *primordial* B mode has been detected. A degree-scale signal reported in 2014 [2227] was shown by joint analysis with multi-frequency data to be polarized emission from Galactic dust [2246], and the tensor-to-scalar ratio is now bounded by $r < 0.036$ [2369].

Rests on Phenomena 51.6 and 65.3.

Derivation. *Derives Phenomenon 51.7.*

The polarization mechanism is Thomson scattering, and its geometry is derived here in full; the amplitude of the signal and the phase relation of C_ℓ^{TE} to the temperature spectrum need the perturbation theory of Chapter 48 and are quoted.

By Phenomenon 65.3, light Thomson-scattered through 90° emerges completely linearly polarized, its electric vector normal to the scattering plane: the driven electron re-radiates as a dipole, and a dipole does not radiate along its own axis (Chapter 65). Sit on an electron at last scattering and let the line of sight to the observer run along $\hat{\boldsymbol{x}}$. Incident light arriving from $\hat{\boldsymbol{z}}$ and scattered to the observer emerges polarized along $\hat{\boldsymbol{y}}$; incident light from $\hat{\boldsymbol{y}}$ emerges polarized along $\hat{\boldsymbol{z}}$. If the incident intensity is the same from the two directions, the two scattered polarizations superpose to none: an electron in an isotropic bath scatters unpolarized light. A dipole anisotropy of the incident field cancels too: the light from $+\hat{\boldsymbol{z}}$ and from $-\hat{\boldsymbol{z}}$ is scattered with the *same* polarization axis, so only the sum $I(\hat{\boldsymbol{z}}) + I(-\hat{\boldsymbol{z}})$ enters, and a dipole leaves every such antipodal sum at its isotropic value — likewise for $\pm\hat{\boldsymbol{y}}$. The lowest moment of the incident radiation that produces any net polarization is the difference between the two transverse axes, $I(\hat{\boldsymbol{z}}) - I(\hat{\boldsymbol{y}})$: a *quadrupole*. The polarized sky is therefore a map of the quadrupole anisotropy seen by each electron at the moment it last scattered.

That is also why the signal is small. Before decoupling, scattering is so frequent that the radiation each electron sees is isotropized and the quadrupole is erased; after decoupling there are no scatterers left. Polarization is generated only in the brief interval in between, as free streaming builds a local quadrupole while some scattering remains, and the outcome is a signal about an order of magnitude below the temperature anisotropy — a few microkelvin, exactly the level DASI detected [2031]. The quantitative treatment belongs to Chapter 48.

The observed field is described by the Stokes parameters Q and U on the sphere — a spin-2 quantity, since a headless polarization axis returns to itself under rotation by 180° — and it decomposes into a curl-free E pattern and a divergence-free B pattern, the tensor analogue of splitting a vector field into gradient and curl parts. At linear order, scalar density perturbations generate E only; gravitational waves generate E and B alike. The decomposition machinery and that theorem are Chapter 48's; what matters for this chapter is the experimental logic it dictates, namely that the E mode and the TE correlation had to appear with the measured acoustic phases (they did [2031, 2213, 2348]), and that any primordial B mode is a direct gravitational-wave observable, which is why the degree scale is watched by dedicated instruments and why the bound $r < 0.036$ [2369] is a statement about the early universe rather than about photons. □

Remark 51.8 (How a systematic is caught: the BICEP2 episode). In March 2014 the BICEP2 collaboration reported a degree-scale B mode well above the lensing expectation [2227], read at the time as primordial. The instrument observed at a single frequency, 150 GHz, so the measurement itself could not discriminate the achromatic background from Galactic foregrounds; the discrimination came from joining the BICEP2/Keck maps with the multi-frequency Planck dust maps, and the joint analysis attributed the signal to polarized thermal emission from Galactic dust, leaving no significant primordial component [2246]. The episode is kept in this book deliberately: it is the cleanest modern instance of the component-separation logic of Section 51.3 acting as the error bar — a single-frequency detection at the sky's faintest signal level is not yet a measurement of the sky.

51.4.7 Secondary anisotropies

Between the last-scattering surface and the receiver the radiation crosses the entire later universe, and three imprints of the crossing are measured. Each is a signal in its own right; none is primordial, and all must be modelled in the fits of Section 51.5.

Inverse-Compton scattering in clusters. Photons traversing the hot ionized gas of a galaxy cluster are Compton up-scattered by the thermal electrons, depleting the low-frequency side of the spectrum and enriching the high side — the Sunyaev–Zel'dovich effect [1485]. In the

Rayleigh–Jeans region the brightness change is

$$\left.\frac{\Delta T}{T_0}\right|_{\mathrm{RJ}} = -2y\,, \qquad y := \int \frac{k_{\mathrm{B}} T_{\mathrm{e}}}{m_{\mathrm{e}} c^2}\, n_{\mathrm{e}}\, \sigma_{\mathrm{T}}\, \mathrm{d}l\,, \tag{51.8}$$

quoted from [1485], with σ_{T} the Thomson cross-section of Equation (65.5) and the integral running through the cluster. For a rich cluster $y \sim 10^{-4}$, giving a decrement of about 0.5 mK — computed here from Equation (51.8) — which is why clusters appear as *cold* spots at low frequency; the distortion changes sign at about 217 GHz (the null of the y spectral function, computed here as $3.83\,k_{\mathrm{B}}T_0/h$), a frequency signature no intrinsic fluctuation shares. Because y does not dilute with distance, the effect finds clusters at any redshift, and the arcminute surveys of ACT and SPT now use it as a routine cluster-finding channel [2351].

Gravitational lensing. The anisotropy pattern is viewed through the deflecting mass of the whole line of sight, which remaps the sky by a few arcminutes coherently over degree scales: the acoustic peaks are slightly smoothed, characteristic higher-order correlations appear from which the projected mass map is reconstructed, and E-mode polarization is partly converted to the B modes detected in [2219]. The lensing reconstruction is an integral part of the modern likelihood [2348], and it measures the growth of structure — the dark-sector census of Chapter 134 rests partly on it.

Late-time scattering and evolving potentials. When the first stars reionized the universe, a fraction of the photons scattered again; the fitted optical depth is $\tau = 0.054 \pm 0.007$ [2348], so a fraction $1 - \mathrm{e}^{-2\tau} \approx 10$ per cent of the anisotropy power — computed here — is suppressed, and the re-scattering generates the large-angle E-mode "reionization bump" at $\ell \lesssim 10$ from which τ is actually measured [2348]. Photons also gain or lose energy crossing potentials that evolve while being crossed — the integrated Sachs–Wolfe effect, the late-time counterpart of Equation (51.5) already identified in the original analysis [1357] — which adds power at the lowest multipoles once dark energy begins to flatten the potentials (Chapter 48).

51.5 Interpretation

The measurements above are jointly fitted by a spatially flat six-parameter model — ΛCDM, the standard cosmology of Chapter 48 — and the fit is both spectacularly economical and honestly incomplete.

Table 51.6: The six-parameter fit and its principal derived quantities, from the final Planck analysis (temperature, polarization and lensing combined) [2348]. All values quoted.

Parameter	Symbol	Value
Baryon density	$\Omega_b h^2$	0.02237 ± 0.00015
Cold-dark-matter density	$\Omega_c h^2$	0.1200 ± 0.0012
Acoustic scale	$100\theta_{\mathrm{MC}}$	1.04092 ± 0.00031
Reionization optical depth	τ	0.054 ± 0.007
Scalar amplitude	$\ln(10^{10} A_s)$	3.044 ± 0.014
Scalar spectral index	n_s	0.9649 ± 0.0042
Hubble constant	H_0	(67.4 ± 0.5) km/(s Mpc)
Matter fraction	Ω_m	0.315 ± 0.007
Age of the universe	t_0	(13.80 ± 0.02) Gyr

Two consistency checks can be computed here from the table alone. First, the age: for a flat universe with the quoted H_0 and Ω_m, the expansion history of Chapter 48 integrates in

closed form to

$$t_0 = \frac{2}{3H_0\sqrt{1-\Omega_m}}\,\operatorname{arsinh}\sqrt{\frac{1-\Omega_m}{\Omega_m}} \approx 13.8\,\mathrm{Gyr}\,,$$

in agreement with the quoted value. Second, the photon-to-baryon ratio: the quoted $\Omega_b h^2$ corresponds to a present baryon number density of about $0.25/\mathrm{m}^3$, and the measured T_0 of Phenomenon 51.3 to a photon density of 4.1×10^8 per cubic metre, so $\eta = n_b/n_\gamma \approx 6.1\times10^{-10}$ — the number the nucleosynthesis of Chapter 48 requires independently from the light-element abundances. Two entirely different epochs of the universe, read with two entirely different instruments, return one number.

What the CMB establishes is the hot big bang itself and the skeleton of the model. The Planck spectrum of Phenomenon 51.3 establishes a hot, dense, thermalized early phase — no non-thermal explanation survives the FIRAS distortion limits. The isotropy at 10^{-5} establishes the large-scale homogeneity assumed by the cosmological principle. The acoustic scale establishes spatial flatness at the few-per-mille level. The peak morphology establishes that the seeds were adiabatic and coherent, with a spectrum close to, but measurably different from, scale-invariant: $n_s = 0.9649 \pm 0.0042$ sits eight of its standard deviations below unity — computed here from the quoted values — which any account of the seeds must reproduce. And the fit *requires* the dark sector: $\Omega_c h^2$ is five times $\Omega_b h^2$, and the peaks cannot be fitted by baryons alone. What the radiation establishes is the gravitational existence of that sector, not its identity — the census and the failed identifications are Chapter 134's subject.

What the CMB does not establish must be said with equal care. There is no detection of a primordial B mode: the bound $r < 0.036$ [2369] constrains, but does not confirm, the inflationary origin of the seeds, and the mechanism that produced them remains unobserved. And the model that fits the CMB internally is in tension externally: the same ΛCDM parameters predict $H_0 = (67.4 \pm 0.5)\,\mathrm{km/(s\,Mpc)}$, while the local distance ladder measures $(73.0 \pm 1.0)\,\mathrm{km/(s\,Mpc)}$ [2407] — a discrepancy of about 5σ, computed here from the two quoted uncertainties, unresolved at the time of writing and discussed further in Chapter 48. This book records it as an open disagreement between two mature measurements, not as a defect of either.

51.6 Modern repetitions and precision

Sixty years separate the entries of the table below, and every line was, in its year, the best measurement in the field. The monopole temperature improved from an uncertainty of 1.0 K in 1965 to 0.57 mK in 2009 — a factor of nearly two thousand, computed here; the anisotropy resolution improved from the 7° beams of DMR to the arcminute beams of the ground programme.

Table 51.7: The precision lineage, instrument by instrument. Resolutions are the approximate beam scales of the instruments as described in Section 51.2.

Years	Instrument	Contribution
1965	Holmdel horn [1320]	discovery; T to 1.0 K
1976–77	U-2 radiometers [1610]	the dipole
1989–96	COBE FIRAS [1884, 1922]	Planck spectrum to 5×10^{-5} of peak
1992	COBE DMR [1858]	anisotropy discovery at 7°
2000	BOOMERanG [1984]	first acoustic peak; flatness
2002	DASI [2031]	first polarization detection
2001–10	WMAP [2213]	full-sky spectra; precision ΛCDM
2009–13	Planck [2348]	acoustic range mapped to its cosmic-variance floor
2007–	ACT, SPT [2219, 2351]	damping tail, lensing, SZ clusters
2010–	BICEP/Keck [2246, 2369]	degree-scale B modes; $r < 0.036$

Two closing observations belong to the record. First, the FIRAS spectral limits of 1996 [1922] are still unsurpassed after three decades: no flown instrument has improved on $|\mu| < 9 \times 10^{-5}$ or $|y| < 1.5 \times 10^{-5}$, and the proposed successor spectrometers are designed precisely against those numbers. A measurement can saturate its era's technology and stand for a generation. Second, the frontier has moved wholly to polarization: the temperature sky is measured to its cosmic-variance floor of Equation (51.3) over the acoustic range, so further temperature integration cannot help, and the open quantities — r, the reionization history, the neutrino-mass imprint on lensing — are all polarization and secondary-anisotropy science [2348, 2351, 2369].

51.7 Primary references

The discovery and interpretation pair [1320][1305]; the prediction [906] and the precursor measurement [860]; the spectrum [1834, 1884, 1922, 2144]; the dipole [1610, 2352]; the anisotropies [1858, 1984, 2213, 2348]; the polarization [2031, 2219, 2227, 2246, 2369]; the theoretical predictions under test [1357, 1383, 1431, 1434, 1485].

Chapter 52

Stellar Structure and Nucleosynthesis

Hydrostatic stellar models, pp/CNO energy production with its solar-neutrino and helioseismic evidence, and stellar nucleosynthesis against the measured element abundances. A star is a self-gravitating fusion reactor in quasi-static equilibrium, and this chapter builds the standard model of one: the structure equations assembled by Eddington [538], the nuclear energy sources identified by Bethe and von Weizsäcker [824, 828], and the synthesis of the elements in stellar interiors codified by Burbidge, Burbidge, Fowler and Hoyle [1105].

What earns the model its place in an evidence-based treatise is that the Sun's interior has been measured twice over, by two channels that did not have to agree: the neutrinos of the pp chain and CNO cycle, from the chlorine deficit [1368] to its resolution by flavour transformation [2021], and the acoustic oscillation spectrum of helioseismology [1234]. The chapter feeds its endpoints to Chapter 49 and its nuclear physics comes from Chapter 107; the modern monograph is [2196].

52.1 The equations of stellar structure

52.1.1 Hydrostatic equilibrium and timescales

A star is a ball of gas whose mechanical state, to an accuracy that the end of this subsection will quantify, is *rest*: pressure balances self-gravity at every radius. The equations governing that balance are the fluid statics of Chapter 31 applied to a body whose gravitational field is its own.

Proposition 52.1 (Hydrostatic equilibrium of a self-gravitating sphere)**.** *A spherically symmetric fluid at rest under its own gravity satisfies*

$$\frac{\mathrm{d}m}{\mathrm{d}r} = 4\pi r^2 \rho \,, \tag{52.1}$$

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{G m \rho}{r^2} \,, \tag{52.2}$$

where $m(r)$ is the mass interior to radius r.

Rests on Proposition 31.16.

Proof. *Derives Proposition 52.1.*

Equation (52.1) is the definition of $m(r)$ differentiated. For Equation (52.2), hydrostatic balance $\nabla P = \rho \boldsymbol{g}$ (Equation (31.13)) needs the local gravitational acceleration, and for a spherically symmetric distribution that is $g = -Gm(r)/r^2$ directed inward: the shells outside r exert no net force and the mass inside acts as if concentrated at the centre —

Newton's shell theorem, which is Gauss's flux theorem applied to the Poisson equation $\nabla^2\Phi = 4\pi G\rho$ over a sphere of radius r. □

Two more equations make the system closed once the material properties are known. In a steady state, the luminosity L_r flowing outward through radius r grows by exactly the nuclear power released in each shell,

$$\frac{\mathrm{d}L_r}{\mathrm{d}r} = 4\pi r^2 \rho\, \varepsilon\,, \tag{52.3}$$

with ε the energy generation rate per unit mass (Section 52.2); and the temperature gradient is fixed by how that luminosity is carried — by radiation or by convection — which is the subject of Section 52.1.3. Together with an equation of state, an opacity $\kappa(\rho, T)$ and a rate $\varepsilon(\rho, T)$, Equations (52.1) to (52.3) and the transport equation are the four structure equations assembled by Eddington [538].

Before anything is solved, Equation (52.2) already yields the theorem that governs the energetics of every star.

Theorem 52.2 (Virial theorem for a self-gravitating sphere)**.** *For any configuration obeying Equation* (52.2) *with* $P(R) = 0$ *at the surface,*

$$3\int_0^M \frac{P}{\rho}\,\mathrm{d}m = -E_{\mathrm{grav}}\,, \qquad E_{\mathrm{grav}} = -\int_0^M \frac{Gm}{r}\,\mathrm{d}m\,. \tag{52.4}$$

Rests on Equation (52.2) *and Proposition 52.1.*

Proof. *Derives Theorem 52.2.*

Multiply Equation (52.2) by $4\pi r^3$ and integrate from centre to surface. The left side integrates by parts,

$$\int_0^R 4\pi r^3 \frac{\mathrm{d}P}{\mathrm{d}r}\,\mathrm{d}r = \Big[4\pi r^3 P\Big]_0^R - 3\int_0^R P\, 4\pi r^2\,\mathrm{d}r = -3\int_V P\,\mathrm{d}V\,,$$

the boundary term vanishing at both ends. The right side is

$$-\int_0^R \frac{Gm\rho}{r^2}\, 4\pi r^3\,\mathrm{d}r = -\int_0^M \frac{Gm}{r}\,\mathrm{d}m = E_{\mathrm{grav}}\,,$$

since $\mathrm{d}m = 4\pi r^2\rho\,\mathrm{d}r$; and $\int P\,\mathrm{d}V = \int (P/\rho)\,\mathrm{d}m$. □

Remark 52.3 (Negative heat capacity). For a monatomic ideal gas the internal energy density is $\frac{3}{2}P$, so Equation (52.4) reads $2U_{\mathrm{int}} = -E_{\mathrm{grav}}$, and the total energy is

$$E = U_{\mathrm{int}} + E_{\mathrm{grav}} = \tfrac{1}{2}E_{\mathrm{grav}} = -U_{\mathrm{int}}\,.$$

A star that *loses* energy therefore grows *hotter*: as E decreases, U_{int} increases, half of the released gravitational energy being radiated and half stored as heat. This negative heat capacity is what makes a contracting protostar heat up until fusion ignites, and — combined with the extreme temperature sensitivity of the nuclear rates derived in Section 52.2.1 — it is what makes the burning stable: a transient over-production of energy expands the star, which *cools* it and throttles the reactions back.

Phenomenon 52.4 (The Sun has shone for billions of years)**.** The Sun radiates $L_\odot = 3.828 \times 10^{26}\,\mathrm{W}$, and the radiometric ages of the oldest meteorites and terrestrial rocks show that it has been doing so, at a comparable rate, for about 4.5 Gyr. Neither chemical energy nor gravitational contraction can supply this for anything like that long. The conflict between the geological record and the physics of the day is what forced the conclusion that stars run on a subatomic energy source [466, 538].

Rests on Theorem 52.2.

Derivation. *Derives Phenomenon 52.4.*

Three timescales settle the matter, and none of them requires solving a structure equation.

The *Kelvin–Helmholtz* time is how long a star could shine on the gravitational energy released in contracting to its present radius. For mass M and radius R that energy is of order GM^2/R (Equation (52.4)), so

$$t_{\mathrm{KH}} \sim \frac{GM^2}{RL} = \frac{6.674\times10^{-11}\,\mathrm{m^3/(kg\,s^2)}\times(1.989\times10^{30}\,\mathrm{kg})^2}{6.957\times10^{8}\,\mathrm{m}\times3.828\times10^{26}\,\mathrm{W}} \approx 1\times10^{15}\,\mathrm{s}\,, \tag{52.5}$$

about 3×10^7 yr. This is the number that set nineteenth-century physics against geology and against Darwin: it falls short of the age of the Earth by more than two orders of magnitude.

The *nuclear* time supposes instead that hydrogen fuses to helium, releasing a fraction 0.007 of the rest energy, and that the fuel available is the inner tenth of the mass, where the temperature is high enough for it to burn:

$$t_{\mathrm{nuc}} \sim \frac{0.007\times0.1\,Mc^2}{L} = \frac{0.0007\times1.989\times10^{30}\,\mathrm{kg}\times8.988\times10^{16}\,\mathrm{m^2/s^2}}{3.828\times10^{26}\,\mathrm{W}} \approx 3\times10^{17}\,\mathrm{s}\,, \tag{52.6}$$

about 1×10^{10} yr. This comfortably exceeds the observed 4.5 Gyr and is of the right order to explain why stars of a solar mass are still on the main sequence in the oldest clusters — so the nuclear hypothesis is not merely sufficient, it is quantitatively right.

The *dynamical* time, $t_{\mathrm{dyn}} \sim (G\bar\rho)^{-1/2}$, is under an hour for the Sun. Its separation from the other two by ten and by fourteen orders of magnitude is what licenses treating the star as exactly hydrostatic while its composition slowly changes, which is the assumption every model in this chapter rests on. □

52.1.2 Polytropes and the Lane–Emden equation

The four structure equations couple mechanics to thermodynamics through the run of temperature. There is one family of models in which the mechanics closes on its own: assume a power-law relation between pressure and density. The classical treatment is [831].

Definition 52.5 (Polytrope)**.** A *polytrope of index* n is a self-gravitating sphere obeying Equations (52.1) and (52.2) with the equation of state

$$P = K\rho^{1+1/n}\,, \tag{52.7}$$

K and n constants.

Real stars realize particular indices exactly: a fully convective star is isentropic, and for a monatomic ideal gas constant entropy means $P\propto\rho^{5/3}$, that is $n=3/2$; a non-relativistic degenerate electron gas has the same exponent, and its relativistic limit $P\propto\rho^{4/3}$ is $n=3$ (Chapter 120, taken up for white dwarfs in Chapter 49); and a star in which the ratio of gas to total pressure is uniform is an $n=3$ polytrope, as shown at the end of this subsection.

Theorem 52.6 (Lane–Emden equation)**.** *Write* $\rho=\rho_{\mathrm c}\theta^n$ *with* $\rho_{\mathrm c}$ *the central density, and* $r=\alpha\xi$ *with*

$$\alpha^2 = \frac{(n+1)K}{4\pi G}\,\rho_{\mathrm c}^{\,\frac{1-n}{n}}\,. \tag{52.8}$$

Then Equations (52.1) *and* (52.2) *with Equation* (52.7) *are equivalent to*

$$\frac{1}{\xi^2}\frac{\mathrm d}{\mathrm d\xi}\left(\xi^2\frac{\mathrm d\theta}{\mathrm d\xi}\right) = -\theta^n\,, \qquad \theta(0)=1\,,\quad \theta'(0)=0\,. \tag{52.9}$$

The surface is the first zero ξ_1 *of* θ*, so* $R=\alpha\xi_1$*.*

Rests on Equations (52.1)*,* (52.2) *and* (52.7)*.*

Proof. *Derives Theorem 52.6.*

Divide Equation (52.2) by ρ, multiply by r^2, differentiate, and use Equation (52.1):

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r^2}{\rho}\frac{\mathrm{d}P}{\mathrm{d}r}\right) = -G\frac{\mathrm{d}m}{\mathrm{d}r} = -4\pi G\rho r^2 \,. \tag{52.10}$$

For Equation (52.7),

$$\frac{1}{\rho}\frac{\mathrm{d}P}{\mathrm{d}r} = K\left(1+\frac{1}{n}\right)\rho^{\frac{1}{n}-1}\frac{\mathrm{d}\rho}{\mathrm{d}r} = (n+1)K\,\frac{\mathrm{d}}{\mathrm{d}r}\left(\rho^{1/n}\right) = (n+1)K\rho_{\mathrm{c}}^{1/n}\frac{\mathrm{d}\theta}{\mathrm{d}r}\,,$$

using $\rho^{1/n} = \rho_{\mathrm{c}}^{1/n}\theta$. Substituting into Equation (52.10) and setting $r = \alpha\xi$,

$$\frac{(n+1)K\rho_{\mathrm{c}}^{1/n}}{\alpha^2}\,\frac{1}{\xi^2}\frac{\mathrm{d}}{\mathrm{d}\xi}\left(\xi^2\frac{\mathrm{d}\theta}{\mathrm{d}\xi}\right) = -4\pi G\rho_{\mathrm{c}}\theta^n\,,$$

which is Equation (52.9) exactly when α is chosen as Equation (52.8). The boundary conditions state that the density is ρ_{c} at the centre and that the pressure gradient vanishes there, as Equation (52.2) requires since $m \sim r^3$. □

Proposition 52.7 (Exact solutions)**.** *Equation (52.9) has closed-form solutions for exactly three indices:*

$$n=0:\ \theta = 1-\frac{\xi^2}{6}\,,\qquad n=1:\ \theta=\frac{\sin\xi}{\xi}\,,\qquad n=5:\ \theta=\left(1+\frac{\xi^2}{3}\right)^{-1/2}\,,$$

with $\xi_1 = \sqrt{6}$, π and ∞ respectively; the $n=5$ sphere has infinite radius but finite mass.

Proof. For $n=1$ substitute $u=\xi\theta$: Equation (52.9) becomes $u''=-u$, so $u=\sin\xi$ selects the solution regular at the origin with $\theta(0)=1$. The $n=0$ and $n=5$ cases are verified by direct differentiation in the same way; for $n=5$, $\theta^5\xi^2\to 0$ fast enough that the mass integral below converges even though θ has no zero. □

For every other index the equation is integrated numerically, which is elementary: it is a regular initial-value problem away from $\xi=0$, started with the series $\theta = 1-\xi^2/6+n\xi^4/120-\cdots$. The two indices this chapter and Chapter 49 need are (computed here, by that integration)

$$\begin{aligned} n=\tfrac{3}{2}:&\quad \xi_1 = 3.654\,,\quad \xi_1^2\,|\theta'(\xi_1)| = 2.714\,,\quad \rho_{\mathrm{c}}/\bar\rho = 5.99\,,\\ n=3:&\quad \xi_1 = 6.897\,,\quad \xi_1^2\,|\theta'(\xi_1)| = 2.018\,,\quad \rho_{\mathrm{c}}/\bar\rho = 54.2\,. \end{aligned} \tag{52.11}$$

The mass follows from Equation (52.9) itself, with no further integration:

$$M = \int_0^R 4\pi r^2\rho\,\mathrm{d}r = 4\pi\alpha^3\rho_{\mathrm{c}}\int_0^{\xi_1}\theta^n\xi^2\,\mathrm{d}\xi = 4\pi\alpha^3\rho_{\mathrm{c}}\,\xi_1^2\,|\theta'(\xi_1)|\,, \tag{52.12}$$

since the integrand is $-\mathrm{d}\left(\xi^2\theta'\right)/\mathrm{d}\xi$. The ratio of central to mean density in Equation (52.11) is $\xi_1/\left(3\,|\theta'(\xi_1)|\right)$, a pure number fixed by n alone.

Eliminating ρ_{c} between Equation (52.12) ($M\propto\rho_{\mathrm{c}}^{(3-n)/2n}$ at fixed K) and $R=\alpha\xi_1\propto\rho_{\mathrm{c}}^{(1-n)/2n}$ gives the polytropic mass–radius relation

$$M \propto R^{\frac{3-n}{1-n}}\,. \tag{52.13}$$

Two cases carry physics. For $n=3/2$, $M\propto R^{-3}$: a non-relativistic degenerate star *shrinks* as mass is added, the inverted mass–radius relation of white dwarfs (Chapter 49). For

$n = 3$ the exponent diverges: the mass is independent of the radius, fixed by K alone — the degenerate ultra-relativistic case, where that unique mass is the Chandrasekhar limit (Equation (49.9)).

Eliminating K and ρ_{c} instead in favour of M and R gives the central pressure,

$$P_{\mathrm{c}} = W_n \frac{GM^2}{R^4}, \qquad W_n = \frac{1}{4\pi(n+1)\theta'(\xi_1)^2}, \tag{52.14}$$

with $W_{3/2} = 0.770$ and $W_3 = 11.05$ (computed here from Equation (52.11)).

Remark 52.8 (The Eddington standard model). Eddington's model of a star [538] is the assumption that the ratio $\beta \equiv P_{\mathrm{gas}}/P$ of gas pressure to total pressure is uniform, the remainder $P_{\mathrm{rad}} = (1-\beta)P = aT^4/3$ being radiation pressure (Equation (68.2); $a = 4\sigma/c$). Eliminating T between $P_{\mathrm{rad}} = aT^4/3 = (1-\beta)P$ and $P_{\mathrm{gas}} = \rho k_{\mathrm{B}} T/(\mu m_{\mathrm{u}}) = \beta P$ gives

$$P = \left[\frac{3(1-\beta)}{a}\right]^{1/3} \left[\frac{k_{\mathrm{B}}}{\beta \mu m_{\mathrm{u}}}\right]^{4/3} \rho^{4/3}, \tag{52.15}$$

a polytrope of index exactly 3, with K a known function of β and the mean molecular weight μ. Because the $n = 3$ mass Equation (52.12) depends on K alone, M determines β: substituting gives *Eddington's quartic equation*, $1-\beta \propto \beta^4 \mu^4 M^2$, which for the Sun ($\mu = 0.61$) evaluates to $1-\beta = 4.1 \times 10^{-4}$ (computed here): radiation pressure is negligible in the Sun. At $100\,M_\odot$ the same equation gives $1-\beta \approx 0.12$ — massive stars are increasingly radiation-supported, the structural face of the Eddington limit derived in Section 52.1.4.

The model's payoff is the solar centre with no nuclear physics and no transport theory at all. From Equations (52.11) and (52.14), computed here,

$$P_{\mathrm{c}} = 11.05 \frac{GM_\odot^2}{R_\odot^4} = 1.2 \times 10^{16}\,\mathrm{Pa}, \qquad \rho_{\mathrm{c}} = 54.2\,\bar{\rho}_\odot = 7.6 \times 10^4\,\mathrm{kg/m^3},$$

and, from the ideal-gas law with $\mu = 0.61$,

$$T_{\mathrm{c}} = \frac{P_{\mathrm{c}}\,\mu m_{\mathrm{u}}}{k_{\mathrm{B}}\,\rho_{\mathrm{c}}} \approx 1.2 \times 10^7\,\mathrm{K}.$$

The detailed standard solar model, with its opacity tables, composition gradient and helioseismic calibration, gives $T_{\mathrm{c}} = 1.57 \times 10^7\,\mathrm{K}$, $\rho_{\mathrm{c}} \approx 1.5 \times 10^5\,\mathrm{kg/m^3}$ and $P_{\mathrm{c}} \approx 2.3 \times 10^{16}\,\mathrm{Pa}$ [2072] (quoted): a one-parameter-family polytrope lands within a factor of two of all three, and the temperature within twenty-five per cent. That $T_{\mathrm{c}} \sim 10^7\,\mathrm{K}$, secure already at this level, is what Section 52.2 needs.

52.1.3 Energy transport: radiation and convection

Deep inside a star, photons are absorbed and re-emitted after mean free paths of millimetres. Energy therefore leaks outward as a diffusive trickle, and the machinery is the transport theory of Chapter 55 with the photon gas of Chapter 120 as the carrier.

Proposition 52.9 (Radiative diffusion)**.** *Where the photon mean free path $\ell = 1/(\kappa\rho)$ is small compared with the temperature scale height, the radiative energy flux is*

$$F = -\frac{c}{3\kappa\rho}\frac{\mathrm{d}}{\mathrm{d}r}\left(aT^4\right) = -\frac{4acT^3}{3\kappa\rho}\frac{\mathrm{d}T}{\mathrm{d}r}, \tag{52.16}$$

equivalently, with $L_r = 4\pi r^2 F$,

$$\frac{\mathrm{d}T}{\mathrm{d}r} = -\frac{3\kappa\rho L_r}{16\pi a c\, r^2 T^3}, \tag{52.17}$$

where κ is the Rosseland mean opacity (below).

Rests on Equation (68.2).

Proof. *Derives Proposition 52.9.*

Carriers of speed c and mean free path ℓ random-walking down a gradient of their energy density u transport the net flux $F = -(c\ell/3)\,\mathrm{d}u/\mathrm{d}r$ — the elementary mean-free-path transport argument of Chapter 55, the factor $\frac{1}{3}$ being the angular average. With matter and radiation locally equilibrated the photon energy density is that of a blackbody at the local temperature, $u = aT^4$ (Equation (68.2)), and $\ell = 1/(\kappa\rho)$ defines the opacity κ as an absorption cross-section per unit mass. Substituting gives Equation (52.16).

Both hypotheses are extravagantly satisfied. In the solar interior $\kappa\rho \sim 10^2/\mathrm{m}$, so $\ell \sim 10^{-2}\,\mathrm{m}$ against a scale height of $\sim 10^8\,\mathrm{m}$: the anisotropy the flux represents, $F/(cu)$, is of order 10^{-11} (computed here with $F \approx 2.5\times10^8\,\mathrm{W/m^2}$ and $T \approx 5\times10^6\,\mathrm{K}$ at mid-radius) — the radiation field is isotropic and Planckian to that accuracy, and the diffusion approximation is among the best in physics. Because real opacity depends on frequency, the κ appearing here must be the *Rosseland mean*: the transport-weighted *harmonic* mean, $1/\kappa = \langle \kappa_\nu^{-1} \rangle$ weighted by $\partial B_\nu/\partial T$, since the flux flows preferentially through the transparent frequency windows (the frequency-by-frequency statement is quoted; see [2196]). □

Two opacity regimes recur. At high temperature the electrons are free and Thomson scattering dominates (Chapter 65): $\kappa_{\mathrm{es}} = \sigma_{\mathrm{T}}(1+X)/2m_{\mathrm{u}} = 0.034\,\mathrm{m^2/kg}$ for hydrogen mass fraction $X = 0.7$ (computed here), independent of ρ and T. At lower temperatures bound–free and free–free absorption dominate, with the law Kramers derived from his X-ray absorption theory, $\kappa \propto \rho T^{-7/2}$ [503]. In cool stellar envelopes the opacity is instead governed by the $\mathrm{H^-}$ ion and *rises* steeply with temperature, which is what Section 52.3.3 will need.

Equation (52.16) also prices the journey: a photon random-walks $N \sim (R/\ell)^2$ steps, taking $t \sim R^2/(\ell c)$, which with the mean solar values above is of order $10^4\,\mathrm{yr}$ (computed here), lengthening to order $10^5\,\mathrm{yr}$ when the centrally concentrated run of $\kappa\rho$ is used (quoted [2196]). Energy generated in the core today emerges from the photosphere in the far future; only the neutrinos of Section 52.5 leave in real time.

Radiative diffusion fails, and convection takes over, when the gradient it demands is too steep to be mechanically stable.

Theorem 52.10 (Schwarzschild criterion)**.** *A chemically homogeneous stratification obeying Equation (52.2) is unstable to convection if and only if*

$$\nabla \equiv \frac{\mathrm{d}\ln T}{\mathrm{d}\ln P} > \nabla_{\mathrm{ad}} \equiv \left(\frac{\partial \ln T}{\partial \ln P}\right)_s , \tag{52.18}$$

where $\nabla_{\mathrm{ad}} = 1 - 1/\Gamma_1 = 2/5$ *for a monatomic ideal gas [330]. Radiative equilibrium is therefore stable exactly where*

$$\nabla_{\mathrm{rad}} \equiv \frac{3\kappa L_r P}{16\pi a c G\, m T^4} \le \nabla_{\mathrm{ad}} , \tag{52.19}$$

∇_{rad} *being the gradient Equation (52.17) written per unit of* $\ln P$ *using Equation (52.2).*
Rests on Equations (52.2) and (52.17).

Proof. *Derives Theorem 52.10.*

Displace a fluid parcel upward by δr, slowly enough to stay in pressure balance with its surroundings (the sound-crossing time of a parcel is far shorter than any transport time) yet fast enough to exchange no heat (the radiative diffusion time across it is far longer); both inequalities hold by margins of many orders of magnitude, which is what makes the adiabatic parcel an accurate idealization. Moving adiabatically at equal pressure, its density becomes

$$\rho_{\mathrm{p}} = \rho + \frac{\rho}{\Gamma_1 P}\frac{\mathrm{d}P}{\mathrm{d}r}\,\delta r ,$$

with $\Gamma_1 = (\partial \ln P/\partial \ln \rho)_s$, while the ambient density at the new position is $\rho + (\mathrm{d}\rho/\mathrm{d}r)\delta r$. The parcel keeps rising — instability — precisely when it arrives lighter than its surroundings:

$$\frac{\rho}{\Gamma_1 P}\frac{\mathrm{d}P}{\mathrm{d}r} < \frac{\mathrm{d}\rho}{\mathrm{d}r}\,.$$

For an ideal gas of uniform composition, $\ln\rho = \ln P - \ln T + \text{const}$, and dividing by $\mathrm{d}\ln P/\mathrm{d}r < 0$ reverses the inequality, turning it into $\nabla > 1 - 1/\Gamma_1 = \nabla_{\mathrm{ad}}$, which is Equation (52.18). In the stable case the restoring buoyancy makes the parcel oscillate at the Brunt–Väisälä frequency

$$N^2 = g\left(\frac{1}{\Gamma_1}\frac{\mathrm{d}\ln P}{\mathrm{d}r} - \frac{\mathrm{d}\ln\rho}{\mathrm{d}r}\right), \tag{52.20}$$

positive exactly when Equation (52.18) fails; these are the buoyancy (g-)modes of a stellar interior, the acoustic counterpart being Section 52.6. In the unstable case the same expression is a growth rate, of order the dynamical time — so a superadiabatic stratification overturns essentially instantly on stellar timescales, and a convection zone sits pinned at ∇ only infinitesimally above ∇_{ad} except in its outermost, tenuous layers. □

Equation (52.19) locates the convection zones. ∇_{rad} is large where the opacity is large or where L_r/m is large. Cool envelopes satisfy the first condition — in the Sun, hydrogen's partial ionization both raises κ enormously (H^-) and depresses ∇_{ad} below 2/5 (energy fed into ionization rather than temperature), so the outer envelope convects, from $0.71\,R_\odot$ to the surface as helioseismology locates it (Section 52.6.2). Stars burning on the CNO cycle satisfy the second: Section 52.2.3 concentrates ε so sharply that L_r saturates near the centre, and such stars have *convective cores* instead. The two configurations exchange roles across the mass scale, and the difference propagates into everything from surface abundances to main-sequence lifetimes.

Remark 52.11 (Mixing-length theory). What convection actually carries, once running, is modelled — not derived. Böhm-Vitense's mixing-length theory [1136] posits parcels that rise a distance $\ell_{\mathrm{m}} = \alpha_{\mathrm{MLT}} H_P$ (H_P the pressure scale height) before dissolving, and carries the flux implied by the buoyancy work over that distance. The single free parameter $\alpha_{\mathrm{MLT}} \approx 1.5$–$2$ is calibrated on the Sun, and this honest phenomenological confession sits inside every stellar evolution code. Deep convection is so efficient that the result barely matters there ($\nabla \to \nabla_{\mathrm{ad}}$, an isentrope, whatever α_{MLT}); it matters within the superadiabatic surface layers, and through them the stellar radius. The turbulence being modelled belongs to Chapter 31.

52.1.4 The mass–luminosity relation and the Eddington limit

With hydrostatics and radiative transport in hand, the structure equations yield their most consequential scaling law before any of them is solved.

Phenomenon 52.12 (The mass–luminosity relation)**.** Among main-sequence stars whose masses are measured directly, from the orbits of eclipsing and visual binaries, the luminosity rises as a steep power of the mass,

$$L \propto M^{\alpha}\,, \qquad \alpha \approx 3 \text{ to } 4\,, \tag{52.21}$$

across the bulk of the main sequence [511]. A star of ten solar masses is some thousands of times the more luminous. Set beside the fuel supply of Phenomenon 52.4, which grows only in proportion to the mass, this makes massive stars enormously short-lived, and it is what turns the main-sequence turnoff of a star cluster into a clock.

Rests on Equations (52.2) *and* (52.17) *and Propositions 52.1 and 52.9.*

Derivation. *Derives Phenomenon 52.12.*

The exponent follows from three of the structure equations without any of them being solved, by comparing the scalings of their terms — a *homology* argument. Write M and R for the mass and radius and P, $\rho \sim M/R^3$, T for characteristic interior values.

Hydrostatic equilibrium (Equation (52.2)) scales as

$$\frac{P}{R} \sim \frac{GM}{R^2} \cdot \frac{M}{R^3}\,, \qquad \text{so} \qquad P \sim \frac{GM^2}{R^4}\,.$$

For an ideal gas of mean molecular weight μ, that is $P = \rho k_{\mathrm{B}} T/(\mu m_{\mathrm{u}})$, and eliminating P and ρ leaves

$$T \sim \frac{G\mu m_{\mathrm{u}}}{k_{\mathrm{B}}} \cdot \frac{M}{R}\,. \tag{52.22}$$

Radiative diffusion carries the energy out (Equation (52.17)), which scales as

$$L \sim \frac{acR^2T^3}{\kappa\rho} \cdot \frac{T}{R} = \frac{acRT^4}{\kappa\rho} \sim \frac{acR^4T^4}{\kappa M}\,.$$

Substituting Equation (52.22), every factor of R cancels:

$$L \sim \frac{ac}{\kappa}\left(\frac{G\mu m_{\mathrm{u}}}{k_{\mathrm{B}}}\right)^4 M^3\,. \tag{52.23}$$

For an opacity independent of the local conditions — electron scattering, which dominates in hot interiors — this is $L \propto M^3$, with no free parameter and no dependence on the radius at all. The observed exponent runs somewhat above three because at lower masses the opacity instead follows Kramers' law $\kappa \propto \rho T^{-7/2}$ [503], which steepens the dependence.

The content of Equation (52.23) is worth stating plainly, because it is counter-intuitive: the luminosity of a main-sequence star is fixed by its mass and composition and does not refer to the energy source at all. A star shines as brightly as its own weight compels it to, and the nuclear reaction rate adjusts, by way of the central temperature, to supply exactly that. □

The same radiation that carries the luminosity pushes on the matter carrying it, and there is a luminosity at which the push wins.

Proposition 52.13 (Eddington luminosity)**.** *For fully ionized hydrogen, hydrostatic equilibrium is impossible above*

$$L_{\mathrm{Edd}} = \frac{4\pi GMm_{\mathrm{p}}c}{\sigma_{\mathrm{T}}} = 1.26\times10^{31}\,\mathrm{W}\times\frac{M}{M_\odot} = 3.3\times10^{4}\,L_\odot\times\frac{M}{M_\odot} \tag{52.24}$$

(computed here) [538].

Rests on Equation (52.16).

Proof. *Derives Proposition 52.13.*

A radiation flux F carries momentum flux F/c, and an electron intercepts it with cross-section σ_{T}, feeling an outward force $\sigma_{\mathrm{T}}F/c$; the protons, electrostatically tied to the electrons, supply almost all the weight, $Gm_{\mathrm{p}}M/r^2$ per electron. With $F = L/4\pi r^2$, both forces fall as $1/r^2$, so their ratio is one number for the whole star:

$$\frac{f_{\mathrm{rad}}}{f_{\mathrm{grav}}} = \frac{\sigma_{\mathrm{T}}L}{4\pi GMm_{\mathrm{p}}c} \equiv \frac{L}{L_{\mathrm{Edd}}}\,.$$

At $L > L_{\mathrm{Edd}}$ the effective gravity is outward at every radius and no hydrostatic atmosphere exists: the star sheds mass instead. Evaluating with the constants of Appendix B gives Equation (52.24). □

Setting Equation (52.23) against Equation (52.24): L grows as M^3 but L_{Edd} only as M, and the two meet near $180\,M_\odot$ (computed here from the calibration $L = L_\odot(M/M_\odot)^3$). Stars approaching that mass are radiation-pressure-dominated (Remark 52.8), loosely bound and violently unstable to mass loss, which is why the observed main sequence ends at masses of order $10^2\,M_\odot$. The same limit, applied to accretion instead of fusion, caps the luminosity of matter falling onto compact objects and calibrates the brightest X-ray sources and quasars (Chapter 50).

52.2 Nuclear energy generation

52.2.1 Tunnelling and the Gamow peak

The temperature that hydrostatics dictates (Remark 52.8) is, by the standards of nuclear physics, absurdly cold. Two protons must approach to a few femtometres to fuse, against a Coulomb barrier of 720 keV (Equation (107.99)), while $k_{\mathrm{B}}T_{\mathrm{c}} = 1.35\,\mathrm{keV}$: the barrier stands five hundred times above the mean thermal energy, and the Boltzmann factor for surmounting it classically, $\mathrm{e}^{-530} \sim 10^{-230}$, forbids fusion in any classical gas of any astrophysical size. The resolution is quantum tunnelling through the barrier [600], applied to stellar interiors by Atkinson and Houtermans within a year [622]. The machinery — the Gamow factor, the astrophysical S factor, and the thermal average over the Maxwell–Boltzmann tail — is built in Chapter 107, whose central result, Theorem 107.85, this section uses as is: the reaction rate is dominated by a narrow window of collision energies (the *Gamow peak*) centred on $E_0 = \left(E_G(k_{\mathrm{B}}T)^2/4\right)^{1/3}$ (Equation (107.101)), of width $\Delta E = 4\sqrt{E_0 k_{\mathrm{B}}T/3}$ (Equation (107.102)), where E_G is the Gamow energy of the colliding pair. For the Sun the worked numbers are Example 107.86: protons fuse at $E_0 \approx 6\,\mathrm{keV}$, from far out on the thermal tail.

What this chapter must add is the *temperature sensitivity*, because everything structural — the segregation of burning stages, the pp/CNO divide, the existence of convective cores, the stability of the whole configuration — hangs on it.

Proposition 52.14 (Temperature exponent of a nonresonant rate)**.** *With $S(E)$ slowly varying, the thermally averaged rate per pair behaves as*

$$\langle \sigma v \rangle \propto T^{-2/3} \exp\left(-\frac{3E_0}{k_{\mathrm{B}}T}\right), \tag{52.25}$$

so its local logarithmic slope is

$$\nu \equiv \frac{\partial \ln \langle \sigma v \rangle}{\partial \ln T} = \frac{\tau - 2}{3}, \qquad \tau \equiv \frac{3E_0}{k_{\mathrm{B}}T}. \tag{52.26}$$

Rests on Equations (107.101) *and* (107.102).

Proof. *Derives Proposition 52.14.*

The thermal average is $\langle \sigma v \rangle \propto (k_{\mathrm{B}}T)^{-3/2} \int S(E)\, \mathrm{e}^{-E/k_{\mathrm{B}}T - \sqrt{E_G/E}}\, \mathrm{d}E$, and the saddle-point evaluation performed in the proof of Theorem 107.85 gives for the integral a Gaussian of height $S(E_0)\,\mathrm{e}^{-3E_0/k_{\mathrm{B}}T}$ (the exponent at the peak is $E_0/k_{\mathrm{B}}T + \sqrt{E_G/E_0} = 3E_0/k_{\mathrm{B}}T$, using $\sqrt{E_G/E_0} = 2E_0/k_{\mathrm{B}}T$ from Equation (107.101)) and of width $\Delta E \propto \sqrt{E_0 k_{\mathrm{B}}T}$ (Equation (107.102)). Since $E_0 \propto T^{2/3}$, the width scales as $T^{5/6}$, and the prefactors assemble to $T^{-3/2} \times T^{5/6} = T^{-2/3}$, which is Equation (52.25). Differentiating: $\tau = 3E_0/k_{\mathrm{B}}T \propto T^{-1/3}$, so $\mathrm{d}\tau/\mathrm{d}\ln T = -\tau/3$ and

$$\frac{\partial \ln \langle \sigma v \rangle}{\partial \ln T} = -\frac{2}{3} + \frac{\tau}{3} = \frac{\tau - 2}{3}.$$

□

Table 52.1: Gamow-peak parameters of the hydrogen-burning reactions at the solar central temperature $T_{\mathrm{c}} = 1.57 \times 10^7\,\mathrm{K}$, that is $k_{\mathrm{B}}T = 1.353\,\mathrm{keV}$: Gamow energy E_G, peak energy E_0 and width ΔE (Equations (107.101) and (107.102)), and the temperature exponent ν of Equation (52.26). All entries computed here.

Reaction	E_G (MeV)	E_0 (keV)	ΔE (keV)	τ	ν
$p + p$	0.493	6.1	6.6	13.5	3.8
$^3\mathrm{He} + {}^4\mathrm{He}$	26.9	23.1	12.9	51.2	16.4
$p + {}^7\mathrm{Be}$	13.8	18.5	11.5	41.0	13.0
$p + {}^{14}\mathrm{N}$	45.1	27.4	14.1	60.8	19.6

Table 52.1 turns Equation (52.26) into the architecture of stellar burning. The pp reaction runs as T^4; the CNO bottleneck as T^{20}; and higher-Z fuels, with their larger E_G, switch on only at successively higher temperatures — which is why burning proceeds in *discrete, well-separated stages* (Section 52.2.4) rather than as a continuous smoulder. The steepness also closes the loop opened in Remark 52.3: a star that momentarily over-burns expands and cools, and a rate falling as the twentieth power of a falling temperature is a thermostat of extraordinary stiffness. A star is stable not despite the violence of its energy source but because of how sharply that source responds to the star's own mechanical state.

52.2.2 The pp chains

In the Sun, hydrogen burns predominantly by the proton–proton chains. The first step,

$$p + p \longrightarrow d + e^+ + \nu_e\,,$$

is a *weak* interaction — two protons form no bound state (Chapter 107), so one must convert to a neutron during the fleeting collision — first computed by Bethe and Critchfield from Fermi's theory [810] (Chapter 101). Its cross-section is some twenty orders of magnitude below direct measurement and is known only from theory; that a weak process gates the whole chain is why the Sun burns for 10^{10} years instead of detonating, and the derivation of the total release $Q = 26.73\,\mathrm{MeV}$ per helium nucleus is Phenomenon 107.87 (Equation (107.104)). The deuteron is consumed within seconds, and the chain then completes along three branches.

Table 52.2: The three pp branches. Q-values are computed here from the measured atomic mass excesses (Chapter 107), positron annihilation included, so each branch totals 26.73 MeV per ^{4}He (Equation (107.104)); neutrino energies are quoted from [2072]. Branch shares of chain terminations in the present Sun (quoted, BS05 [2072]): pp I about 85 %, pp II about 15 %, pp III about 0.02 %.

Branch	Reaction	Q (MeV)	E_ν (MeV)
pp I	$p + p \to d + e^+ + \nu_e$	1.442	≤ 0.420, $\langle E_\nu \rangle = 0.267$
	$d + p \to {}^3\mathrm{He} + \gamma$	5.493	—
	$^3\mathrm{He} + {}^3\mathrm{He} \to {}^4\mathrm{He} + 2p$	12.860	—
pp II	$^3\mathrm{He} + {}^4\mathrm{He} \to {}^7\mathrm{Be} + \gamma$	1.587	—
	$^7\mathrm{Be} + e^- \to {}^7\mathrm{Li} + \nu_e$	0.862	0.862 (90%), 0.384
	$^7\mathrm{Li} + p \to 2\,{}^4\mathrm{He}$	17.346	—
pp III	$^7\mathrm{Be} + p \to {}^8\mathrm{B} + \gamma$	0.137	—
	$^8\mathrm{B} \to 2\,{}^4\mathrm{He} + e^+ + \nu_e$	18.07	$\langle E_\nu \rangle \approx 6.7$

The neutrinos are not detail: they are the audit trail. Each branch loses a different fraction of Q to neutrinos, which never thermalize (computed here from Table 52.2): pp I loses $2 \times 0.267 = 0.53\,\mathrm{MeV}$, or 2.0 %; pp II loses 1.08 MeV, 4.0 %; pp III loses 7.0 MeV, fully 26 % of its Q. Weighted by the branch shares, the Sun's neutrino loss is about 2.3 % (computed here): the photon luminosity corresponds to an effective 26.1 MeV of heat per helium nucleus assembled. Two rare variants complete the bookkeeping: the *pep* reaction $p + e^- + p \to d + \nu_e$, a monochromatic 1.44 MeV line, and the *hep* reaction on ^{3}He, the highest-energy solar neutrinos. Branch by branch, this entire table has now been read out of the Sun in real time by neutrino spectroscopy (Section 52.5.3; [2297]), including the tiny pp III branch whose ^{8}B neutrinos carried the solar-neutrino problem for thirty years (Section 52.5.1).

52.2.3 The CNO cycle

Where carbon, nitrogen and oxygen are present, hydrogen can also burn catalytically, in the cycle proposed independently by von Weizsäcker [824] and worked out quantitatively by Bethe [828]:

Table 52.3: The CN cycle. Q-values computed here from the measured atomic mass excesses, positron annihilation included; they sum to 26.73 MeV, as they must, the C, N and O nuclei being returned unchanged. Neutrino mean energies quoted from [2072].

Reaction	Q (MeV)	E_ν (MeV)
$^{12}\mathrm{C} + p \to {}^{13}\mathrm{N} + \gamma$	1.943	—
$^{13}\mathrm{N} \to {}^{13}\mathrm{C} + e^+ + \nu_e$	2.220	$\langle E_\nu \rangle = 0.707$
$^{13}\mathrm{C} + p \to {}^{14}\mathrm{N} + \gamma$	7.551	—
$^{14}\mathrm{N} + p \to {}^{15}\mathrm{O} + \gamma$	7.297	—
$^{15}\mathrm{O} \to {}^{15}\mathrm{N} + e^+ + \nu_e$	2.754	$\langle E_\nu \rangle = 0.997$
$^{15}\mathrm{N} + p \to {}^{12}\mathrm{C} + {}^4\mathrm{He}$	4.965	—

The cycle's properties follow from Table 52.1. Its slowest link at stellar temperatures is $^{14}\mathrm{N}(p,\gamma)^{15}\mathrm{O}$ — the largest Gamow energy among the captures, compounded by a small S factor — so in equilibrium the catalysts pile up as ^{14}N: the CNO cycle is itself a nucleosynthesis process, and the cosmic abundance of nitrogen is largely its fossil [1105]. The same link sets the temperature exponent, $\nu \approx 20$ (Table 52.1), against $\nu \approx 4$ for pp. Two consequences:

First, the crossover. The pp rate, with its shallow slope, wins at low temperature; the CNO rate, at solar metallicity, overtakes it near $T \approx 1.7 \times 10^7\,\mathrm{K}$ — just above the solar centre — corresponding to main-sequence masses above about $1.3\,M_\odot$ (modern values, quoted [2196]). Bethe's own comparison, using the higher solar central temperature then accepted, placed the Sun itself on the CNO side [828]; improved solar models moved the divide just above it. The Sun sits below the divide: Borexino's direct detection of CNO neutrinos measures the cycle's contribution at about one per cent of the solar luminosity [2349], exactly the subordinate role the crossover predicts (Section 52.5.3).

Second, the core convection of Section 52.1.3. A T^{20} rate concentrates the burning into the innermost few per cent of the mass; L_r there saturates, ∇_{rad} of Equation (52.19) exceeds ∇_{ad}, and every CNO-burning star carries a convective core — continuously refuelled and chemically homogenized, which lengthens its life and reshapes its track in Section 52.3.

52.2.4 Helium burning and beyond

Past helium the chart of nuclides has a double gap: no stable mass 5, no stable mass 8. Salpeter identified the two-step route through the unbound ^{8}Be [1010], and Hoyle turned the observed abundance of carbon into the prediction of a nuclear excited state — the argument, stated from the nuclear-physics side in Remark 107.88, that this chapter records as a phenomenon because its input was astronomical.

Phenomenon 52.15 (Carbon exists, and so does the resonance that makes it)**.** Carbon is among the most abundant elements in the universe, yet the route to it is blocked twice: there is no stable nucleus of mass number 5 or 8, so helium burning must pass through the unstable ^{8}Be, whose lifetime is of order 10^{-16} s [1010]. From the bare fact of carbon's abundance, Hoyle argued that the triple-alpha rate must be resonantly enhanced and therefore that ^{12}C must possess a hitherto unknown excited state of spin-parity 0^{+} lying just above the three-alpha threshold, near 7.65 MeV of excitation. The state was then looked for, and found at that energy [1044]. It remains the clearest case in physics of a nuclear property predicted from an astronomical abundance.

Rests on Definition 107.31 and Proposition 52.16.

Derivation. *Derives Phenomenon 52.15.*

The threshold energy is fixed by the masses alone, and it is what makes the prediction quantitative. Three ^{4}He nuclei have a combined atomic mass of $3 \times 4.002602 = 12.007806$ in unified atomic mass units, while ^{12}C has, by the definition of that unit, a mass of exactly 12. The excess is 0.007806 mass units, and with $1\,\mathrm{u} = 931.494\,\mathrm{MeV}/c^{2}$,

$$Q_{3\alpha} = 0.007806 \times 931.494\,\mathrm{MeV} \approx 7.27\,\mathrm{MeV}\,. \tag{52.27}$$

The three-alpha system therefore enters ^{12}C at 7.27 MeV of excitation, so the state Hoyle required at 7.65 MeV sits 0.38 MeV *above* the threshold.

That placement is the whole of the argument. Below the threshold the state could not act as a resonance for three free alpha particles at all. A few hundred keV higher and the Boltzmann factor at helium-burning temperatures would have suppressed the rate, leaving a universe with very little carbon in it. The window is narrow, the abundance is large, and the state is where the abundance requires it to be — which is why the laboratory confirmation was taken as decisive rather than as a coincidence. □

The resonance also fixes how helium burning *responds* to temperature, which is qualitatively unlike the nonresonant rates of Section 52.2.1.

Proposition 52.16 (Temperature sensitivity of the triple-alpha rate)**.** *With the capture fully resonant through the Hoyle state at $E_{\mathrm{r}} = 380\,\mathrm{keV}$ above the three-alpha threshold, the rate per unit volume behaves as*

$$r_{3\alpha} \propto n_{\alpha}^{3}\, T^{-3} \exp\left(-\frac{E_{\mathrm{r}}}{k_{\mathrm{B}}T}\right)\,, \tag{52.28}$$

so its temperature exponent is $\nu = E_{\mathrm{r}}/k_{\mathrm{B}}T - 3 \approx 41$ at $T = 1 \times 10^{8}\,\mathrm{K}$ (computed here).

Rests on Equation (52.27).

Proof. *Derives Proposition 52.16.*

For a resonance narrow on the thermal scale, the thermally averaged rate per pair follows from inserting a sharply peaked cross-section into the Maxwell–Boltzmann average (Chapter 55): the distribution contributes its value at the resonance, $\propto (k_{\mathrm{B}}T)^{-3/2}\mathrm{e}^{-E/k_{\mathrm{B}}T}$, and the energy integral over the resonance profile contributes a temperature-independent strength, so $\langle \sigma v \rangle \propto T^{-3/2}\mathrm{e}^{-E/k_{\mathrm{B}}T}$ with E the resonance energy in the collision frame.

The triple-alpha route applies this twice. First, $\alpha + \alpha \rightleftharpoons {}^{8}\mathrm{Be}$: the ^{8}Be ground state sits $\Delta E_1 = 92\,\mathrm{keV}$ above two alphas, and because its formation and decay proceed through the same channel, the equilibrium population cancels the resonance strength entirely, leaving the Saha-like factor

$$\frac{n_8}{n_\alpha^2} \propto T^{-3/2} \exp\left(-\frac{\Delta E_1}{k_\mathrm{B} T}\right) .$$

Second, ${}^{8}\mathrm{Be} + \alpha \to {}^{12}\mathrm{C}^{*}$ through the Hoyle state, which lies $\Delta E_2 = 288\,\mathrm{keV}$ above the ${}^{8}\mathrm{Be} + \alpha$ threshold: another factor $T^{-3/2}\mathrm{e}^{-\Delta E_2/k_\mathrm{B}T}$. The product is Equation (52.28) with $E_\mathrm{r} = \Delta E_1 + \Delta E_2 = 380\,\mathrm{keV}$ — necessarily the Hoyle state's height above *three* free alphas, Equation (52.27) subtracted from 7.65 MeV. Differentiating logarithmically: $\nu = E_\mathrm{r}/k_\mathrm{B}T - 3$, and at $T = 1 \times 10^{8}\,\mathrm{K}$, where $k_\mathrm{B}T = 8.62\,\mathrm{keV}$, this is $44 - 3 = 41$. □

A rate scaling as T^{41} (and as the *cube* of the density) is effectively a switch: helium ignites abruptly near $1 \times 10^{8}\,\mathrm{K}$ and nowhere cooler. In stars below about $2\,M_\odot$ the helium core reaches that temperature only after becoming electron-degenerate, and ignition under degeneracy is explosive — pressure barely responds to temperature, the thermostat of Remark 52.3 is disconnected, and the burning runs away as the *helium flash* until degeneracy is lifted (Chapters 49 and 120).

Beyond the triple-alpha reaction, successive α captures climb the ladder ${}^{12}\mathrm{C}(\alpha,\gamma){}^{16}\mathrm{O}$ and onward. The carbon-to-oxygen ratio of the universe is set by the competition between the triple-alpha rate and the ${}^{12}\mathrm{C}(\alpha,\gamma)$ rate, whose S factor at stellar energies remains among the least certain numbers in nuclear astrophysics — an honest caveat carried by every model of white dwarf composition. The later stages then follow in the sequence Table 52.4, each ignited by the next Coulomb barrier per Proposition 52.14, each faster than the last: partly because less energy remains per reaction (Phenomenon 107.32), and mainly because from carbon burning onward the core's thermal neutrino emission, not photon diffusion, carries away the power, decoupling the burning from the surface luminosity entirely.

Table 52.4: Advanced burning stages of a massive star ($\sim 20\,M_\odot$): approximate ignition temperatures and durations, quoted as orders of magnitude from [2196]. The collapse of the timescales from millions of years to a day is driven by neutrino losses.

Stage	Principal products	T (K)	Duration
Hydrogen	${}^{4}\mathrm{He}$, ${}^{14}\mathrm{N}$	$\sim 4 \times 10^{7}$	$\sim 10^{7}\,\mathrm{yr}$
Helium	${}^{12}\mathrm{C}$, ${}^{16}\mathrm{O}$	$\sim 2 \times 10^{8}$	$\sim 10^{6}\,\mathrm{yr}$
Carbon	${}^{20}\mathrm{Ne}$, ${}^{23}\mathrm{Na}$, ${}^{24}\mathrm{Mg}$	$\sim 8 \times 10^{8}$	$\sim 10^{3}\,\mathrm{yr}$
Neon	${}^{16}\mathrm{O}$, ${}^{24}\mathrm{Mg}$	$\sim 1.5 \times 10^{9}$	$\sim 1\,\mathrm{yr}$
Oxygen	${}^{28}\mathrm{Si}$, ${}^{32}\mathrm{S}$	$\sim 2 \times 10^{9}$	$\sim 0.3\,\mathrm{yr}$
Silicon	iron-peak nuclei	$\sim 3.5 \times 10^{9}$	$\sim 10^{-2}\,\mathrm{yr}$

Silicon burning is not a fusion of two silicon nuclei — that barrier is out of reach — but a photodisintegration rearrangement: at $3.5 \times 10^{9}\,\mathrm{K}$ the thermal photons strip nucleons and alphas off nuclei and the ensemble relaxes toward nuclear statistical equilibrium, which populates the binding-energy maximum (Proposition 107.41). There the energy release stops, by Phenomenon 107.32: an iron core can only absorb energy, by photodisintegration and electron capture, and its collapse — with everything that follows, neutron stars, supernovae and stellar black holes — is where this chapter hands the star to Chapter 49.

52.3 The Hertzsprung–Russell diagram and stellar evolution

52.3.1 The diagram

Everything observable about the surface of a distant star reduces, to first approximation, to two numbers: a luminosity and a surface temperature (measured as a colour or a spectral type). Plotting stars in that plane is due to Hertzsprung [373] and, independently, Russell [420], and the result is the single observable summary against which every stellar model is judged.

Phenomenon 52.17 (The Hertzsprung–Russell diagram)**.** Stars do not fill the luminosity–temperature plane. About ninety per cent of them lie on one narrow band — the *main sequence* — running diagonally from hot and luminous to cool and faint; a separate population of *giants* sits cool yet luminous; *white dwarfs* sit hot yet faint, thousands of times below the main sequence [373, 420]. The bands are loci, not scatter: a star's position determines its other properties, and clusters of common age populate the diagram in sharply truncated patterns.

Rests on Equations (52.23) *and* (68.2) *and Phenomenon 52.12.*

Derivation. Derives Phenomenon 52.17.

The plane has a hidden third coordinate. A star radiates approximately as a blackbody, so (Equation (68.2))

$$L = 4\pi R^2 \sigma T_{\text{eff}}^4 \,, \tag{52.29}$$

and lines of constant radius are diagonals of slope four in $\log L$ against $\log T_{\text{eff}}$. Every feature of the diagram is a statement about radii.

The main sequence is the locus of hydrogen burners, traced by mass. Homology fixes both of its coordinates as functions of M alone. The luminosity is Equation (52.23), $L \propto M^3$. The radius follows from energy balance: the nuclear power $L \propto M\bar{\rho}T^\nu \propto M^{2+\nu}R^{-3-\nu}$ must equal the transported $L \propto M^3$, so

$$R \propto M^{\frac{\nu-1}{\nu+3}} \,, \tag{52.30}$$

which for pp burning ($\nu \approx 4$, Table 52.1) is $R \propto M^{0.43}$. Then Equation (52.29) gives $T_{\text{eff}}^4 \propto M^3/R^2 \propto M^{2.14}$, and eliminating M,

$$L \propto T_{\text{eff}}^{5.6} \tag{52.31}$$

(computed here): a single steep curve, parametrized by mass, with essentially no width — because homologous stars of one composition have no free parameter left. The main sequence is a mass sequence, not an age sequence.

The other bands are radius excursions. Giants are stars of ordinary mass at 10–$10^2\,R_\odot$ (Section 52.3.3); white dwarfs are stars of ordinary mass at $\sim 10^{-2}\,R_\odot$, the degenerate configurations of Chapter 49. Their separation from the main sequence by orders of magnitude in L at fixed T_{eff} is just Equation (52.29) read along a vertical line. And a cluster — a population of one age and composition — fills the main sequence only up to the mass whose lifetime (Section 52.3.2) equals the cluster's age, which is the truncation observed. □

52.3.2 Main-sequence lifetimes

The fuel scales as M; the burn rate as M^3 to M^4 (Phenomenon 52.12). So the main-sequence lifetime,

$$t_{\text{MS}} \approx \frac{0.007 \times 0.1\, Mc^2}{L} \propto M^{1-\alpha} \,, \tag{52.32}$$

falls as roughly $M^{-2.5}$: $1 \times 10^{10}\,\text{yr}$ for the Sun (Equation (52.6)), but only about $3 \times 10^7\,\text{yr}$ at $10\,M_\odot$ (computed here with $\alpha = 3.5$). The most massive stars live for astronomically

negligible times, which is why O stars are found only in regions of active star formation, and why the main-sequence *turnoff* of a cluster (Phenomenon 52.17) is a clock: reading the turnoff mass and applying Equation (52.32) dates the cluster. The oldest globular clusters date this way to about thirteen thousand million years (quoted; [2196]) — a lower bound on the age of the universe that cosmology must and does clear (Chapter 48).

What ends the main-sequence phase is not the exhaustion of hydrogen everywhere but the mechanics of the helium ash at the centre.

Proposition 52.18 (Schönberg–Chandrasekhar limit)**.** *An isothermal, non-degenerate helium core of mass M_{c} and molecular weight μ_{c}, at temperature T_{c} set by the surrounding envelope, can support that envelope only if its surface pressure does not exceed*

$$P_{\max} = \frac{c_1}{G^3 M_{\mathrm{c}}^2}\left(\frac{k_{\mathrm{B}} T_{\mathrm{c}}}{\mu_{\mathrm{c}} m_{\mathrm{u}}}\right)^4 , \tag{52.33}$$

c_1 a constant of order unity; matching to the envelope demand caps the core mass fraction at

$$q \equiv \frac{M_{\mathrm{c}}}{M} \;\le\; q_{\mathrm{SC}} \approx 0.37\left(\frac{\mu_{\mathrm{env}}}{\mu_{\mathrm{c}}}\right)^2 , \tag{52.34}$$

about ten per cent for a helium core beneath a hydrogen envelope (coefficient quoted from [868, 2196]).

Rests on Theorem 52.2.

Proof. *Derives Proposition 52.18.*

Repeat the integration by parts of Theorem 52.2 over the core alone: the boundary term no longer vanishes, leaving the virial theorem with a surface term,

$$3P_{\mathrm{s}}V_{\mathrm{c}} = 3\int_0^{M_{\mathrm{c}}} \frac{P}{\rho}\,\mathrm{d}m + E_{\mathrm{grav}} = \frac{3M_{\mathrm{c}}k_{\mathrm{B}}T_{\mathrm{c}}}{\mu_{\mathrm{c}}m_{\mathrm{u}}} - c_0\frac{GM_{\mathrm{c}}^2}{R_{\mathrm{c}}} ,$$

the first equality for an isothermal ideal gas, c_0 of order unity encoding the core's density profile. As a function of the core radius this is $P_{\mathrm{s}} = A/R_{\mathrm{c}}^3 - B/R_{\mathrm{c}}^4$ with $A \propto M_{\mathrm{c}}T_{\mathrm{c}}/\mu_{\mathrm{c}}$ and $B \propto GM_{\mathrm{c}}^2$: zero for a large diffuse core, zero again for a compressed one (gravity wins), with a maximum in between at $R_{\mathrm{c}}^{*} = 4B/3A$ of height $P_{\max} = 27A^4/256B^3$, which is Equation (52.33). The remarkable feature is the *denominator*: the more massive the core, the *less* pressure it can exert.

The envelope, meanwhile, demands of the core's surface a pressure of the star's own interior order, $P_{\mathrm{env}} \sim GM^2/R^4$, at a temperature it also dictates, $k_{\mathrm{B}}T_{\mathrm{c}} \sim G\mu_{\mathrm{env}}m_{\mathrm{u}}M/R$ (Equation (52.22)). Substituting both into $P_{\max} \ge P_{\mathrm{env}}$, all factors of R cancel:

$$\frac{(\mu_{\mathrm{env}}/\mu_{\mathrm{c}})^4\,GM^4}{M_{\mathrm{c}}^2\,R^4} \gtrsim \frac{GM^2}{R^4} \quad\Longrightarrow\quad q^2 \lesssim \left(\frac{\mu_{\mathrm{env}}}{\mu_{\mathrm{c}}}\right)^4 ,$$

which is Equation (52.34) up to the coefficient, obtained by integrating the isothermal core structure exactly [868]. With $\mu_{\mathrm{env}} \approx 0.6$ and $\mu_{\mathrm{c}} \approx 1.3$, $q_{\mathrm{SC}} \approx 0.08$ (computed here from Equation (52.34)). □

When shell burning pushes the helium core past q_{SC}, no isothermal equilibrium remains: the core contracts on its own Kelvin–Helmholtz timescale, and the star leaves the main sequence. (Below about $2\,M_\odot$ the core evades the limit by going degenerate — degeneracy pressure needs no temperature — which is why low-mass stars linger as subgiants while more massive stars cross the gap quickly.) The envelope's response to core contraction is expansion — the "mirror" behaviour every detailed calculation shows [2196] — and the star moves right in the diagram, toward the giant region and the track derived next.

52.3.3 Before and after the main sequence

A star arriving at or leaving the main sequence spends its time fully or largely convective, and for such stars there is a wall in the HR diagram.

Proposition 52.19 (The Hayashi track is nearly vertical)**.** *A fully convective star whose photospheric opacity follows the H^- law $\kappa = \kappa_0 \rho^{1/2} T^9$ (quoted fit, valid near $4 \times 10^3\,\mathrm{K}$ [2196]) satisfies*

$$T_{\mathrm{eff}} \propto M^{7/51}\, L^{1/102}\,, \tag{52.35}$$

so its effective temperature is nearly independent of its luminosity: fully convective stars of given mass form an almost vertical line in the HR diagram, and the region cooler than that line is forbidden [1209].

Rests on Theorems 52.6 and 52.10.

Proof. *Derives Proposition 52.19.*

Four ingredients, each one line.

(i) The interior is an $n = 3/2$ polytrope. Efficient convection enforces $\nabla = \nabla_{\mathrm{ad}}$ (Theorem 52.10 and Remark 52.11): the star is isentropic, $P = K\rho^{5/3}$, and the Lane–Emden mass relation Equation (52.12) with the scale Equation (52.8) fixes the adiabat's constant in terms of the bulk parameters: eliminating ρ_{c} between $M \propto K^{3/2}\rho_{\mathrm{c}}^{1/2}$ and $R \propto K^{1/2}\rho_{\mathrm{c}}^{-1/6}$ leaves $K \propto GM^{1/3}R$.

(ii) The adiabat in (P,T). With the ideal-gas law $\rho = \mu m_{\mathrm{u}} P/k_{\mathrm{B}}T$ substituted into $P = K\rho^{5/3}$,

$$P \propto K^{-3/2}\,T^{5/2} \propto M^{-1/2}R^{-3/2}\,T^{5/2}\,.$$

(iii) The photosphere. Hydrostatic equilibrium in the atmosphere, $\mathrm{d}P = (g/\kappa)\,\mathrm{d}\tau$ per unit optical depth, integrates to $P_{\mathrm{ph}}\,\kappa_{\mathrm{ph}} \simeq \frac{2}{3}\,g = \frac{2}{3}\,GM/R^2$ at the photosphere $\tau = \frac{2}{3}$. Writing the opacity law in (P,T) via the ideal-gas law, $\kappa \propto \rho^{1/2}T^9 \propto P^{1/2}T^{17/2}$, this is $P_{\mathrm{ph}}^{3/2}\,T_{\mathrm{eff}}^{17/2} \propto M/R^2$.

(iv) Match. The photosphere lies on the interior adiabat, so substituting (ii) into (iii):

$$M^{-3/4}R^{-9/4}\,T_{\mathrm{eff}}^{15/4}\cdot T_{\mathrm{eff}}^{17/2} \propto \frac{M}{R^2} \quad\Longrightarrow\quad T_{\mathrm{eff}}^{49} \propto M^7 R\,.$$

Eliminating $R \propto L^{1/2}T_{\mathrm{eff}}^{-2}$ (Equation (52.29)) gives $T_{\mathrm{eff}}^{51} \propto M^7 L^{1/2}$, which is Equation (52.35) (algebra verified numerically, computed here).

The exponent $1/102$ is the result: as L changes a thousandfold, T_{eff} moves by $1000^{1/102} \approx 7$ per cent (computed here). The forbidden region follows because a star pushed cooler than this track would have to be superadiabatic throughout its convective envelope, which convection erases on a dynamical timescale (Theorem 52.10). The extreme steepness of the H^- opacity in T is what pins the wall in place. □

Proposition 52.19 organizes both ends of a star's life. A collapsing protostar becomes fully convective and *descends* the Hayashi line at nearly constant T_{eff} (the Hayashi phase), then, as its interior heats and turns radiative, peels off toward the main sequence — reaching it on the Kelvin–Helmholtz timescale Equation (52.5), which is why 3×10^7 yr is the right answer to a question Kelvin asked about the wrong energy source. After the main sequence, the sequence of Proposition 52.18 runs: the helium core contracts, hydrogen burns in a shell, the envelope expands and cools until it *meets* the Hayashi line, and the star, unable to cross it, is deflected upward along it — the red giant branch, at ever-growing radius and luminosity. Helium ignition follows (Proposition 52.16; as a flash below $\sim 2\,M_\odot$), then core helium burning, and then the double-shell asymptotic giant branch, whose thermal pulses alternately ignite the two shells and drive the dredge-up episodes that carry freshly synthesized material to the surface — the fact that Phenomenon 52.21 turns into proof of

nucleosynthesis in situ. Mass loss then strips the envelope (planetary nebula), leaving the carbon–oxygen white dwarf; stars above about $8\,M_\odot$ instead burn through Table 52.4 to core collapse. Both endpoints, and the remnants they leave, are Chapter 49; the detailed evolutionary calculations behind this narrative are [2196].

52.4 Stellar nucleosynthesis

52.4.1 The abundance record

Nucleosynthesis theory has one empirical target: the measured abundances of the elements, compiled from the solar photosphere, primitive meteorites and stellar spectra into a single curve by Suess and Urey [1098], with the modern solar photospheric revision by Asplund and collaborators [2140]. The light-element floor — hydrogen, helium, and the lithium plateau — is set earlier, by big-bang nucleosynthesis, and belongs to Chapter 48; everything above it is made in stars.

Phenomenon 52.20 (The cosmic abundance curve)**.** Element abundances measured in the solar photosphere, in primitive meteorites and in stellar spectra are not arbitrary: plotted against mass number they trace one reproducible curve with definite features [1098, 2140]. Abundance falls by some ten orders of magnitude from hydrogen to the heaviest nuclei, but the fall is interrupted by a pronounced maximum at the iron group near mass number 56; nuclei that are whole multiples of the alpha particle stand systematically above their neighbours; and beyond iron there are two distinct pairs of peaks, offset from one another. This curve is the empirical target of every theory of the origin of the elements, and it is against it that the processes of [1105] were defined.

Rests on Phenomenon 107.32 and Equation (107.66)*.*

Derivation. *Derives Phenomenon 52.20.*

The iron peak, the most conspicuous feature, follows from the binding energies alone. The binding energy per nucleon B/A rises steeply among the lightest nuclei, where surface effects and nucleon pairing dominate; reaches a broad maximum of about 8.8 MeV in the iron-nickel region, near $A = 56$ to 62; and then declines slowly as the Coulomb repulsion of the growing proton number takes over (Chapter 107). Fusing two nuclei releases energy exactly when it raises B/A, so a self-gravitating fusion reactor liberates energy while it burns toward the iron group and would have to absorb energy to proceed past it.

Two consequences follow, and both are visible in the curve. Material processed through successive burning stages piles up where the energy release stops, which is the iron peak. And elements beyond iron cannot be made by fusion in a star at all: they require neutron capture, which faces no Coulomb barrier and so proceeds at any temperature, followed by beta decay [1105]. Whether capture is slow or fast compared with the intervening beta decays selects a different path across the chart of nuclides, and the two paths stall at different closed neutron shells — which is why the peaks beyond iron come in offset pairs rather than singly. □

52.4.2 B^{2}FH and the eight processes

Burbidge, Burbidge, Fowler and Hoyle organized the synthesis of every nuclide into eight processes, each defined by the feature of Phenomenon 52.20 it exists to explain [1105]; Cameron's independent contemporaneous framework reached essentially the same taxonomy [1106]. Table 52.5 lists them with their modern identifications.

The neutron-capture split deserves its derivation, because the abundance curve records it directly. A nucleus exposed to a neutron flux captures at a rate $1/\tau_n$ set by the flux, and

Table 52.5: The processes of [1105], with the sites and products as identified today. The first four are the burning stages of Sections 52.2 and 52.2.4; the last four are the origin of everything beyond the iron peak and of the light elements.

Process	Mechanism and site	Principal products
H burning	pp chains and CNO cycle (Sections 52.2.2 and 52.2.3)	^{4}He, ^{14}N
He burning	triple-alpha and α captures (Section 52.2.4)	^{12}C, ^{16}O
α	advanced burning stages (Table 52.4)	^{20}Ne to ^{40}Ca
e	nuclear statistical equilibrium in silicon burning and supernovae	iron peak
s	slow neutron capture in AGB stars (Section 52.3.3)	heavy nuclides along the valley; peaks at $A \approx 88$, 138, 208
r	rapid neutron capture in explosive, neutron-rich events (Section 52.4.3)	neutron-rich heavies; peaks at $A \approx 80$, 130, 195; actinides
p	photodisintegration of heavy seeds in supernovae	rare proton-rich isotopes
x	light-element synthesis, left open in 1957; now cosmic-ray spallation (Phenomenon 115.16) and big-bang nucleosynthesis (Chapter 48)	Li, Be, B

beta-decays toward stability at a rate $1/\tau_\beta$. The two limits draw different paths across the chart of nuclides:

Slow ($\tau_n \gg \tau_\beta$): every unstable capture product decays before the next capture, so the path hugs the valley of beta stability. At the magic neutron numbers $N = 50$, 82, 126 (Theorem 107.52) the capture cross-sections drop sharply — a closed shell holds the next neutron only weakly — so material accumulates just *at* the magic nuclides of the valley: abundance peaks at $A \approx 88$, 138, 208. The site is the He-burning shells of AGB stars, where (α, n) reactions on ^{13}C and ^{22}Ne release the neutrons; Phenomenon 52.21 is the proof that this runs where the theory puts it.

Rapid ($\tau_n \ll \tau_\beta$): captures outrun decay and drive the material far to the neutron-rich side, until binding is so weak that photodisintegration balances capture. The path then climbs along *waiting points* at the same magic N — but on the neutron-rich side, where a given N is reached at *lower* A. When the flux ceases, the nuclei decay back to the valley at roughly constant A, leaving peaks shifted below the s-process peaks: $A \approx 80$, 130, 195. The twin offset peaks of Phenomenon 52.20 are therefore a direct spectrogram of the magic numbers, read twice at two neutron excesses — and the offset measures how far from stability the r-process ran.

The r-process demands an environment of extreme neutron density on a timescale of seconds. Burbidge, Burbidge, Fowler and Hoyle assigned it to supernovae; the one site now *observed* is the neutron-star merger of Section 52.4.3, with core-collapse supernovae remaining a possible contributor — an honest open fraction.

52.4.3 Direct evidence

The theory above is constrained by the abundance curve it reproduces, but two observations catch nucleosynthesis in the act.

Phenomenon 52.21 (Technetium in stellar atmospheres)**.** The absorption spectra of certain cool giants of spectral class S contain the lines of technetium [1006]. Technetium has no stable isotope, and the longest-lived of them has a half-life of about 4×10^6 years — three orders of magnitude shorter than the lifetime of the star, and four shorter than the age of the Galaxy. The element cannot have been present when the star formed, and there is no mechanism by which it could have been acquired from outside. It is therefore being manufactured in the interior now and carried to the surface: direct observational proof that nucleosynthesis happens inside stars, published five years before the systematic theory [1105].

Rests on Phenomenon 107.2 and Proposition 52.19.

Derivation. *Derives Phenomenon 52.21.*

Radioactive decay is exponential, $N(t) = N_0 2^{-t/t_{1/2}}$, so with $t_{1/2} = 4 \times 10^6$ years any technetium present when a star of age $t = 10^9$ years formed would by now have been reduced by the factor

$$2^{-t/t_{1/2}} = 2^{-250} \approx 10^{-75}\,.$$

This is not merely a small number: a solar mass holds about 10^{57} nucleons, so the surviving population would be less than one atom in the entire star. The observed lines cannot be a relic of the star's formation under any assumption about its initial composition.

Since the atmosphere is far too cool for nuclear reactions to run there, the technetium must be produced in the deep interior and transported outward on a timescale short compared with $t_{1/2}$. That requirement — a dredge-up operating within a few million years — is itself a constraint on the convective structure of evolved giants, and it is met by the thermal pulses of the asymptotic giant branch (Section 52.3.3). The observation therefore does double duty: it establishes in-situ synthesis and it dates the mixing. □

The r-process has its own caught-in-the-act moment, sixty-five years later. The binary neutron-star merger GW170817 (Section 47.4) was followed by a kilonova whose optical and infrared light curve decayed on the timescales of freshly synthesized radioactive heavy nuclei — the thermal glow of an r-process event of about 10^{-2} solar masses of ejecta, observed while it happened [2278]. Technetium in a giant's atmosphere and radioactive heavy elements in merger debris bracket the two neutron-capture processes of Table 52.5: both have now been watched running. On the longest timescales the same processes write galactic history — each stellar generation enriches the gas that forms the next, which is why the ancient halo stars are metal-poor while the young disc is metal-rich, the abundance–age correlation already read as evolutionary in [1105].

52.5 Solar neutrinos

52.5.1 The chlorine experiment and the solar-neutrino problem

Everything in Sections 52.1 and 52.2 rests on physics happening at the centre of a star, under 2×10^{16} Pa of overburden that no photon crosses in less than millennia (Section 52.1.3). The neutrinos of Table 52.2 are the one probe that leaves the core at the speed of light, and Davis built the detector: six hundred tonnes of perchloroethylene deep underground, counting individual argon atoms produced by $\nu_e + {}^{37}\mathrm{Cl} \to e^- + {}^{37}\mathrm{Ar}$ [1368]. The apparatus — installed 1478 m below ground in the Homestake mine — with its procedure and data is registered with the other oscillation experiments in Section 114.2.2; here the Sun's side of the story is told. The chlorine threshold of 0.814 MeV puts the dominant pp neutrinos out of reach: the measured rate is carried mainly by the rare ^{8}B branch of Table 52.2, with a minority from ^{7}Be — precisely the branches most sensitive to the model. Against Bahcall's contemporaneous standard-solar-model prediction [1365], Davis's first result was already low; twenty-five years of running settled the measured rate at 2.56 ± 0.23 SNU, one SNU being 10^{-36} captures per target atom per second, against a predicted 8.1 ± 1.2 (Table 114.4; [1950, 2072]) — a ratio of 0.32 ± 0.05, quoted.

Phenomenon 52.22 (The solar-neutrino deficit, and its resolution)**.** A radiochemical detector deep underground, counting the argon atoms produced by neutrino capture on chlorine, measured a solar electron-neutrino rate about one third of the standard solar model's prediction [1365, 1368]. Every later experiment confirmed a deficit, and it stood for three decades: either the Sun, or the model of it, or the neutrino was misunderstood. The resolution came from a detector sensitive to both the electron-flavour flux and the flavour-blind total. The total agreed with the solar model, while the electron-flavour fraction was depleted by just the observed factor [2021]. The Sun was right and the stellar model was right; the neutrinos change flavour in flight. What had looked for thirty years like a failure of astrophysics became a precision confirmation of it — and, incidentally, the discovery that neutrinos have mass (Chapter 103).

Rests on Equations (107.101) *and* (114.6)*, Proposition 52.14, and Phenomenon 52.25.*

Derivation. *Derives Phenomenon 52.22.*

Two quantitative points carried the thirty-year argument: why the deficit *could* be blamed on astrophysics for so long, and what measurement could not be so blamed.

The ^8B flux is a thermometer of vicious sensitivity. The branch is fed through $^3\mathrm{He} + {}^4\mathrm{He}$ (temperature exponent $\nu = 16.4$, Table 52.1), which builds the ^{7}Be supply, and drained through $p + {}^7\mathrm{Be}$ ($\nu = 13.0$); the competing destruction of ^{7}Be by electron capture is a weak process with almost no temperature dependence. At fixed density and composition the flux therefore responds with the *sum* of the two Gamow exponents, $\Phi({}^8\mathrm{B}) \propto T_\mathrm{c}^{29}$ or so (computed here; detailed solar-model perturbation studies, which let the structure readjust self-consistently, soften the exponent somewhat without changing what follows).

The lever is enormous: with that exponent, a central temperature just four per cent below the model's gives $0.96^{29} \approx 0.31$ (computed here) — the measured ratio. An error of one part in twenty-five in a fifteen-million-kelvin prediction was the entire "astrophysical solution", and no chlorine measurement could exclude it. It was excluded from two independent directions: the gallium detectors reached the pp neutrinos, whose flux is pinned by the solar luminosity itself (Equation (114.6), the bound of Phenomenon 114.3), and still saw a deficit; and helioseismology (Section 52.6.2) measured the sound speed — $c_s^2 \propto T/\mu$ — through the core to a few parts in a thousand, excluding a temperature error of that size by more than an order of magnitude.

The charged-current to neutral-current ratio is flavour-blind astrophysics-proof. On deuterium two channels are open (Section 114.4.4): the charged-current breakup $\nu_e + d \to p + p + e^-$, which only ν_e can drive at these energies, and the neutral-current breakup $\nu + d \to p + n + \nu$, whose amplitude is identical for ν_e, ν_μ and ν_τ because the Z couples to the three lepton doublets with one strength (Chapter 101). The neutral-current rate therefore measures the *total* active flux whatever the Sun does, the charged-current rate measures its electron-flavour part, and their ratio measures the surviving ν_e fraction with every solar uncertainty cancelling. SNO measured the ratio near one third while the neutral-current total matched the solar-model prediction within its uncertainty (Phenomenon 114.5, Table 114.6; [2021]): the missing electron neutrinos were arriving with the wrong flavour, not missing. The oscillation mechanism, the matter effect that shapes the energy dependence, and the parameter determinations belong to Chapters 103 and 114. □

52.5.2 Resolution: flavour transformation

What SNO delivered in 2002 was the comparison the derivation above calls for: in one target, during one exposure, the electron-flavour flux from the charged-current channel and the flavour-blind total from the neutral-current channel (Section 114.4.4). The electron fraction came out near one third; the total ^{8}B flux agreed with the standard solar model within the model's own sixteen per cent uncertainty (Table 114.2) [2021]. Read as stellar physics, that is a *precision confirmation* delivered by the most temperature-sensitive observable the Sun emits: the $T_c^{\sim 20}$ thermometer of Phenomenon 52.22 returned the model's central temperature to within a fraction of a per cent. The Sun was right; the neutrinos oscillate, with all that follows for particle physics developed in Chapter 103 and measured across Chapter 114.

52.5.3 Complete spectroscopy of solar fusion

The modern standard solar model — updated opacities, equation of state, and helioseismic calibration — is that of Bahcall, Serenelli and Basu [2072], whose predicted fluxes are Table 114.2. Borexino measured them: the pp, pep, ^{7}Be and ^{8}B branches separately and in real time [2297], and in 2020 the CNO neutrinos [2349] (Section 114.6.1, Table 114.11).

Phenomenon 52.23 (The Sun's energy budget audited reaction by reaction)**.** Real-time spectroscopy in an ultra-pure liquid scintillator has measured the neutrino fluxes of the individual branches of the pp chain separately [2297], and detected the neutrinos of the CNO cycle, which supplies about one per cent of the Sun's power [2349]. The nuclear energy generation rate reconstructed from those neutrinos agrees with the luminosity the Sun is observed to radiate in photons. The two are not measurements of the same thing at the same time: neutrinos leave the core immediately while photons diffuse out over some 10^5 years, so their agreement also says that the solar output has been steady across that interval.

Rests on Equation (107.104) *and Phenomenon 52.4.*

Derivation. *Derives Phenomenon 52.23.*

The total neutrino flux is fixed by the luminosity with essentially no model dependence, and that is the quantity these measurements check. The net effect of the pp chain, and equally of the CNO cycle, is

$$4\,{}^{1}\mathrm{H} \longrightarrow {}^{4}\mathrm{He} + 2e^{+} + 2\nu_{\mathrm{e}}\,,$$

releasing $Q = 26.73\,\mathrm{MeV}$ per helium nucleus formed (Equation (107.104)) and producing exactly two neutrinos. Each such reaction therefore contributes Q to the luminosity and two neutrinos to the flux, so at the Earth's distance d

$$\Phi_{\nu} = \frac{2L_{\odot}}{Q\,4\pi d^{2}}\,. \tag{52.36}$$

Numerically, with $Q = 26.73\,\mathrm{MeV} = 4.283 \times 10^{-12}\,\mathrm{J}$ and $d = 1.496 \times 10^{11}\,\mathrm{m}$,

$$\Phi_{\nu} = \frac{2 \times 3.828 \times 10^{26}\,\mathrm{W}}{4.283 \times 10^{-12}\,\mathrm{J} \times 4\pi\,(1.496 \times 10^{11}\,\mathrm{m})^{2}} \approx 6.4 \times 10^{14}/(\mathrm{m^{2}\,s})\,,$$

that is some 6×10^{10} per square centimetre per second — the statement made from the nuclear-physics side as Equation (107.105), and deployed as the luminosity constraint on every oscillation analysis in Equation (114.6).

Observe what went into this: the observed luminosity, the energy release of the net reaction, and the Earth–Sun distance. No opacity, no equation of state, no reaction rate, no central temperature. It is a slight overestimate, by a few per cent, because the neutrinos themselves carry away a small part of Q that never becomes light (the loss budget computed in Section 52.2.2). That the measured pp flux matches it is a check on the entire picture of hydrogen burning; the value of the branch-by-branch spectroscopy lies in what it adds beyond this bound — the *distribution* of the fusion rate among the branches, which does depend on the central temperature and the composition, and which is compared with the model term by term in Table 114.11. □

52.6 Helioseismology

52.6.1 The five-minute oscillations

In 1962 Leighton, Noyes and Simon, Doppler-imaging the photosphere, found its surface heaving: patches thousands of kilometres across oscillating vertically with periods clustered around five minutes [1234]. Ulrich proposed that these are not a surface phenomenon but standing *acoustic waves* — sound, trapped in a resonant cavity inside the Sun [1437] — and the proposal carried a sharp signature: trapped modes exist only on discrete ridges in the wavenumber–frequency plane. Deubner's observations resolved exactly those ridges [1543]. The mechanics is the acoustics of Chapters 28 and 31 on a stratified sphere, and the cavity can be exhibited with WKB tools (Chapter 9).

Proposition 52.24 (The p-mode cavity)**.** *Linear adiabatic acoustic waves of angular frequency ω and spherical-harmonic degree ℓ in a stratified star propagate where*

$$k_{r}^{2} = \frac{\omega^{2} - \omega_{\mathrm{ac}}^{2}}{c_{\mathrm{s}}^{2}} - \frac{\ell(\ell+1)}{r^{2}} > 0 \qquad (\omega \gg N)\,, \tag{52.37}$$

with c_{s} the sound speed (Section 31.10.1), N the buoyancy frequency Equation (52.20), and

$$\omega_{\mathrm{ac}} = \frac{c_{\mathrm{s}}}{2H} \tag{52.38}$$

the acoustic cutoff of the near-isothermal atmosphere, H its pressure scale height. A mode is trapped between the inner turning radius r_{t}, where $c_{\mathrm{s}}(r_{\mathrm{t}})/r_{\mathrm{t}} = \omega/\sqrt{\ell(\ell+1)}$, and the surface layer where $\omega < \omega_{\mathrm{ac}}$; the trapped frequencies satisfy

$$\int_{r_{\mathrm{t}}}^{R} k_r \,\mathrm{d}r = (n+\alpha)\pi\,, \tag{52.39}$$

n the radial order and α a slowly varying surface phase.

Rests on Equation (52.20) and Theorems 9.78, 31.12 and 31.22.

Proof. *Derives Proposition 52.24.*

The cutoff. In an isothermal atmosphere, $\rho_0 \propto \mathrm{e}^{-z/H}$ with constant c_{s} and H. Linearize continuity, momentum and the adiabatic condition about this background for vertical motion, neglect buoyancy ($\omega \gg N$), and eliminate all variables in favour of $\chi \equiv \delta p/\sqrt{\rho_0}$: the exponential factors cancel and the three equations collapse to

$$\frac{\mathrm{d}^2\chi}{\mathrm{d}z^2} + \frac{\omega^2 - \omega_{\mathrm{ac}}^2}{c_{\mathrm{s}}^2}\,\chi = 0\,, \qquad \omega_{\mathrm{ac}} = \frac{c_{\mathrm{s}}}{2H}\,,$$

the $\frac{1}{4H^2}$ arising as the square of the $\rho_0^{-1/2}$ gradient. Below ω_{ac} the wave is evanescent: the thinning atmosphere reflects it back down. With photospheric values $T_{\mathrm{eff}} = 5772\,\mathrm{K}$, $\mu \approx 1.25$, $g = 274\,\mathrm{m/s^2}$: $c_{\mathrm{s}} \approx 8.0\,\mathrm{km/s}$, $H = k_{\mathrm{B}}T/\mu m_{\mathrm{u}} g \approx 140\,\mathrm{km}$, so $\omega_{\mathrm{ac}}/2\pi \approx 5\,\mathrm{mHz}$ — a period near three and a half minutes (computed here). Waves of longer period, the five-minute band included, cannot escape: the photosphere is a mirror.

The turning point. In the interior, WKB with $\nabla_{\mathrm{h}}^2 Y_{\ell m} = -\ell(\ell+1)Y_{\ell m}/r^2$ gives the local dispersion relation $\omega^2 = c_{\mathrm{s}}^2(k_r^2 + k_{\mathrm{h}}^2)$ with $k_{\mathrm{h}}^2 = \ell(\ell+1)/r^2$, i.e. Equation (52.37) (the cutoff term matters only near the surface). Descending, c_{s} rises, refracting the ray horizontal: k_r vanishes where $c_{\mathrm{s}}/r = \omega/\sqrt{\ell(\ell+1)}$ and the wave turns back up. High-ℓ modes skim the envelope; low-ℓ modes plunge nearly to the centre — which is what makes the spectrum an interior probe at every depth.

Quantization. A wave trapped between two turning surfaces interferes constructively when its radial phase integral is an integer-plus-phase multiple of π (Theorem 9.78), which is Equation (52.39): at each n, a discrete curve $\omega_n(\ell)$ — precisely the observed ridges of [1543]. The observed power peaks near 3.3 mHz because the modes are driven stochastically by the granulation of the convection zone (Remark 52.11), which shakes hardest at those frequencies (quoted [2024]). □

52.6.2 The seismic Sun

Millions of mode frequencies $\omega_{n\ell}$ have been measured, many to relative accuracies of 10^{-5}. Through Equation (52.39) each is an integral of $1/c_{\mathrm{s}}$ along a different chord of the interior, and a sufficiently rich set of chords can be *inverted*.

Phenomenon 52.25 (The Sun rings, and its interior can be inverted)**.** The solar surface oscillates, with a velocity field whose power is concentrated near a period of five minutes [1234]. The oscillations are not a surface phenomenon but the superposition of millions of trapped acoustic eigenmodes of the whole star [1437], as the discrete ridges observed in the wavenumber–frequency plane confirmed [1543]. Inverting the measured mode frequencies yields the sound speed as a function of radius; it agrees with the standard solar model to a few parts in 10^3 throughout the radiative interior and locates the base of the convection zone [2024]. One discrepancy remains and is recorded here as unresolved: the inversions prefer the older, higher photospheric abundances to the downward-revised ones [2072, 2140].

Rests on Equation (52.39) and Proposition 52.24.

Derivation. *Derives Phenomenon 52.25.*

Away from the surface layers the cutoff term in Equation (52.37) is negligible, and Equation (52.39) becomes

$$\frac{(n+\alpha)\pi}{\omega} = \int_{r_t}^{R} \sqrt{\frac{1}{c_s^2} - \frac{1}{w^2 r^2}}\, \mathrm{d}r \equiv F(w)\,, \qquad w \equiv \frac{\omega}{\sqrt{\ell(\ell+1)}}\,. \tag{52.40}$$

The left side is data; the right side depends on (n, ℓ, ω) only through the single combination w. That millions of measured frequencies, spanning hundreds of ℓ values, collapse onto *one* function $F(w)$ is itself a sharp, parameter-free test of the acoustic-cavity picture, and the solar spectrum passes it (the collapse is known as the Duvall law; quoted [2024]).

The inversion is then classical. Let $a(r) = c_s/r$, monotonic in the region of interest, so that $1/c_s^2 = 1/(a^2 r^2)$ and the turning point is $a(r_t) = w$. Changing variable from r to a,

$$F(w) = \int_{a_s}^{w} \sqrt{\frac{1}{a^2} - \frac{1}{w^2}}\, G(a)\, \mathrm{d}a\,, \qquad G(a) = -\frac{\mathrm{d}\ln r}{\mathrm{d}a}\,,$$

with a_s the surface value. In the variables $x = w^{-2}$, $y = a^{-2}$ this is $\tilde{F}(x) = \int_x^{y_s} \sqrt{y-x}\, g(y)\, \mathrm{d}y$ with $g \geq 0$, and differentiating once,

$$\tilde{F}'(x) = -\frac{1}{2}\int_x^{y_s} \frac{g(y)}{\sqrt{y-x}}\, \mathrm{d}y\,,$$

an integral equation of Abel type. Its classical solution is

$$g(y) = \frac{2}{\pi}\frac{\mathrm{d}}{\mathrm{d}y}\int_y^{y_s} \frac{-\tilde{F}'(x)}{\sqrt{x-y}}\, \mathrm{d}x\,, \tag{52.41}$$

verified by substituting back, exchanging the order of integration, and using $\int_y^x \mathrm{d}u/\sqrt{(x-u)(u-y)} = \pi$. Integrating G recovers $r(a)$ by a quadrature, hence $a(r)$, hence $c_s(r)$: *the sound speed at every radius, from the mode frequencies alone*, the surface phase $\alpha(\omega)$ being fitted along the way — the honest caveat of the method, confined to the layers every mode shares.

What the inversion returns [2024] (quoted): a sound-speed profile agreeing with the standard solar model to a few parts in a thousand at every radius outside the innermost few per cent; the base of the convection zone — visible as the kink where the stratification leaves the adiabat of Theorem 52.10 — at $r = 0.713\,R_\odot$; the envelope helium abundance, read from the Γ_1 depression in the helium ionization zone; and, from the rotational splitting of the modes, the internal rotation — differential through the convection zone, near-rigid beneath, with the shear concentrated in a thin tachocline at the interface, the presumptive seat of the solar dynamo (Chapter 66). Since $c_s^2 \propto T/\mu$, the profile pins the central temperature at the sub-per-cent level, which is what closed the astrophysical escape route in Phenomenon 52.22: the Sun's interior has been measured twice — acoustically and in neutrinos — and the two channels agree with each other and with the model. The one standing tension, between the seismic profile and the revised photospheric abundances [2072, 2140], is recorded in the phenomenon above as unresolved. □

Chapter 53

Experiment: The Classical Tests of General Relativity

Chapter 42 states the classical tests of general relativity as phenomena, and Chapter 44 supplies the field equations from which they follow. Neither shows the measuring. That is the gap this chapter closes: the book has been asserting for five chapters that Mercury's perihelion advances by some 43 seconds of arc per century, that starlight grazing the Sun bends by 1.75 seconds of arc, that a clock at the top of a tower runs fast, and that a radar echo is delayed by passing the Sun — without once exhibiting the apparatus, the raw numbers or the error budget that make those statements evidence rather than doctrine.

Six experiments are reported. They are not six tests of one prediction: they probe logically separable pieces of the theory, and the separation is the point. The redshift measurement and the free-fall null tests constrain the equivalence principle alone, and would be passed by any metric theory of gravity whatever; the light deflection and the radar delay measure the curvature of space, which is where a metric theory first parts company with Newtonian gravity dressed in geometry; the perihelion advance is the first quantity that tests the field equations themselves; and the gyroscope precessions test the behaviour of a locally inertial frame carried through a curved and rotating spacetime. The experiments also span an unusually wide range of evidential quality, from a single photographic campaign of 1919 whose plate selection is still argued over, to a satellite null test at parts in 10^{15}. Both are reported here at their true weight.

53.1 Le Verrier and the perihelion of Mercury (1859)

Tests Phenomenon 53.1.

This is the oldest of the classical tests and the only one whose anomaly was established before the theory that explains it. Its apparatus is not an instrument but an archive: a century and a half of positional astronomy, reduced by the perturbation theory of Chapter 27 until nothing but a residue remained.

53.1.1 Apparatus

Meridian circles and transit instruments of the great European observatories, and above all the record of the transits of Mercury across the solar disc. A transit fixes the instant at which the planet crosses the limb of the Sun to within seconds of time, and therefore constrains the longitude of Mercury's node and perihelion far more sharply than a meridian observation of a small object seen in twilight close to the Sun. Le Verrier's material was the accumulated series of such transits from the end of the seventeenth century to his own day, together with the ordinary meridian observations of the planet [151]. The analytical half

of the apparatus was the theory of secular perturbations: Lagrange's planetary equations integrated for the perturbations of Mercury by Venus, the Earth, Mars, Jupiter and Saturn.

53.1.2 Procedure

An osculating Keplerian ellipse is fitted to the observations; the computed gravitational perturbations of the other planets are applied; and the observed motion of the perihelion is then compared with the computed one. The comparison must be made in a well-defined reference direction, and the conventional one — the equinox — is itself moving, so the largest term in the budget is not gravitational at all but the general precession of the equinoxes. The procedure is thus a subtraction of three large and well-understood quantities from one large observed quantity, in the hope that the remainder is either zero or interesting. Le Verrier found it was not zero.

53.1.3 Observations and data

The perihelion budget, in the modern determination, stands as follows. All entries are rates in seconds of arc per Julian century, referred to the moving equinox.

Table 53.1: The perihelion budget of Mercury, in seconds of arc per Julian century. The first three lines are observation and Newtonian celestial mechanics; the residual on the fourth line is what Le Verrier isolated in 1859, and the fifth is what general relativity supplies with no adjustable parameter.

Contribution	Rate
Observed advance of the perihelion	5600
General precession of the equinoxes	−5025
Perturbations by the other planets	−531
Unexplained residual	43
General relativity, Equation (53.1)	42.98

Le Verrier's own residual was smaller, near 38 seconds of arc per century; the value settled near 43 only after Newcomb's re-reduction of the observations and the improved planetary masses of the later nineteenth century [267]. Nothing in this chapter turns on the difference: the two agree that a residual exists, of the order of tens of seconds of arc per century, and that it is many times the uncertainty of the observations.

Phenomenon 53.1 (Measured anomalous advance of Mercury's perihelion)**.** After the precession of the equinoxes and every Newtonian planetary perturbation have been computed and subtracted, the perihelion of Mercury still advances in the direction of the orbital motion at a rate of about 43 seconds of arc per century. The residual was isolated by Le Verrier in 1859 and no distribution of matter consistent with the rest of solar-system astronomy was ever found to account for it [151]. General relativity predicts, for a bound orbit of semi-major axis a and eccentricity e about a mass M, an advance of

$$\Delta\varphi = \frac{6\pi GM}{c^2 a\,(1-e^2)} \tag{53.1}$$

per revolution, with no adjustable parameter [428]; for Mercury this is 42.98 seconds of arc per century.

Rests on Postulate 43.1 and Proposition 45.8.

Derivation. *Derives Phenomenon 53.1.*

Take from Chapter 45 the orbit equation of a timelike geodesic of the Schwarzschild field in the equatorial plane, written for the inverse radius $u = 1/r$ with $h = r^2\,\mathrm{d}\varphi/\mathrm{d}\tau$ the conserved specific angular momentum:

$$\frac{\mathrm{d}^2u}{\mathrm{d}\varphi^2} + u = \frac{GM}{h^2} + \frac{3GM}{c^2}\,u^2 \,. \tag{53.2}$$

The final term is the whole of the relativistic correction. Its ratio to the term on the left is $3h^2u^2/c^2$, of order v^2/c^2 and so of order 10^{-7} for Mercury; it may therefore be treated as a perturbation.

Dropping it leaves the Newtonian conic of Chapter 27,

$$u_0 = \frac{GM}{h^2}\left(1 + e\cos\varphi\right) ,$$

a closed ellipse whose perihelion does not move. Substituting u_0 into the perturbing term gives

$$\frac{3GM}{c^2}u_0^2 = \frac{3G^3M^3}{c^2h^4}\left(1 + 2e\cos\varphi + e^2\cos^2\varphi\right) .$$

The constant piece and the piece in $\cos 2\varphi$ hidden in $\cos^2\varphi$ drive the oscillator off resonance and produce only bounded corrections to the shape of the orbit. The piece in $\cos\varphi$ drives it *at* its own frequency, and it is this term alone that accumulates. Solving

$$\frac{\mathrm{d}^2u_1}{\mathrm{d}\varphi^2} + u_1 = \frac{6G^3M^3e}{c^2h^4}\cos\varphi \qquad \text{gives} \qquad u_1 = \frac{3G^3M^3e}{c^2h^4}\,\varphi\sin\varphi \,,$$

so that to first order

$$u \simeq \frac{GM}{h^2}\left[1 + e\cos\varphi + e\,\delta\,\varphi\sin\varphi\right] \simeq \frac{GM}{h^2}\left[1 + e\cos((1-\delta)\,\varphi)\right] , \qquad \delta := \frac{3G^2M^2}{c^2h^2} ,$$

the last step using $\cos\left((1-\delta)\,\varphi\right) \simeq \cos\varphi + \delta\varphi\sin\varphi$, valid because $\delta\varphi \ll 1$ over a single revolution. The radius therefore returns to its minimum not after 2π but after $\varphi = 2\pi/\left(1-\delta\right) \simeq 2\pi\left(1+\delta\right)$, and the perihelion advances by

$$\Delta\varphi = 2\pi\delta = \frac{6\pi G^2M^2}{c^2h^2}$$

per revolution. Eliminating h with the Newtonian relation $h^2 = GMa\left(1-e^2\right)$ yields Equation (53.1).

Inserting $GM_\odot = 1.327\,12 \times 10^{20}\,\mathrm{m^3/s^2}$, $a = 5.7909 \times 10^{10}\,\mathrm{m}$ and $e = 0.205\,63$ gives $\Delta\varphi = 5.02 \times 10^{-7}$ in radian measure per revolution. Mercury's period is 87.969 days, so it completes 415.2 revolutions in a Julian century, and the accumulated advance is 2.084×10^{-4} radian, that is 42.98 seconds of arc per century — the fifth line of Table 53.1. □

53.1.4 Interpretation

Two things must be said plainly. The first is that this is a *postdiction*: the number was known for fifty-six years before the theory that produced it, and Einstein computed it in November 1915 knowing what he had to match [428]. The second is that this costs the result almost nothing, for the reason set out in Chapter 1: Equation (53.1) contains no free parameter that could have been tuned. The theory was not built to fit 43 seconds of arc; the field equations were fixed by general covariance and the Newtonian limit, and the orbit followed. A postdiction with no adjustable parameter carries the evidential weight of a prediction, and it is a fair test precisely because the alternatives on offer were all parametric

— an intramercurial planet of adjustable mass, a solar oblateness of adjustable size, a force law $r^{-(2+\epsilon)}$ with adjustable ϵ — and each of them, tuned to fit Mercury, then contradicted the other planets or the observed shape of the Sun.

The same effect has since been measured far outside the solar system. The star S2 orbiting the compact object at the Galactic centre shows a Schwarzschild precession of its orbit, detected by near-infrared interferometry and consistent with Equation (53.1) at a gravitational potential some four orders of magnitude deeper than Mercury's [2347]; that measurement and its instrument belong to Chapter 50.

53.1.5 Primary references

[151] for the anomaly, [428] for the computation that accounts for it, and [2347] for the Galactic-centre analogue. The modern budget of Table 53.1 is the standard one summarized by [2240]; the numbers on its first three lines are quoted at second hand and rounded, and only the last line is computed here.

53.2 Dyson, Eddington and Davidson: the eclipse of 1919

Tests Phenomenon 53.2.

Light deflection is the first of the classical tests in which the prediction preceded the measurement, and the first that distinguishes general relativity from a theory in which gravity acts on light through the equivalence principle alone. The factor separating them is exactly two, which is what makes a photographic measurement of moderate precision decisive.

53.2.1 Apparatus

Two expeditions of the Greenwich Observatory, to Sobral in northern Brazil and to the island of Príncipe in the Gulf of Guinea [465]. Both used photographic refractors fed by coelostat mirrors, since the instruments had to be shipped and set up in the field and could not be pointed at the Sun directly:

- at Sobral, an astrographic objective of 33 cm aperture and 3.4 m focal length, and beside it a smaller lens of 10 cm aperture and 5.8 m focal length;
- at Príncipe, a second astrographic objective of the same 33 cm pattern.

The original report quotes apertures in inches and focal lengths in feet; they are converted here per the unit axiom of Chapter 2. Detection is photographic, on glass plates measured afterwards on a screw micrometer at Greenwich.

53.2.2 Procedure

During totality the star field surrounding the eclipsed Sun is photographed. The same field is photographed again from the same instrument at another epoch — months earlier or later, at night, when the Sun is elsewhere in the sky — to give an undeflected reference. The two plates are then compared star by star. What is sought is a systematic *radial* displacement of each star away from the centre of the Sun, falling off as the reciprocal of the angular distance from it.

The difficulty is that a plate scale which changed between the two epochs would mimic exactly that pattern for a field of limited extent. The reduction therefore solves simultaneously for the scale, the orientation, the plate centre and the deflection constant, and it is the availability of stars at a range of distances from the limb, whose predicted

deflections differ, that separates the deflection from the scale. The eclipse of 29 May 1919 was chosen years in advance for two reasons: its totality was among the longest of the century, and the Sun stood in front of the Hyades, so that an unusually rich field of bright stars surrounded it.

53.2.3 Observations and data

The competing predictions for a ray grazing the solar limb are 1.75 seconds of arc from the full theory and 0.87 from the equivalence principle alone; zero is the third possibility, if light does not fall. The measured deflections, reduced to the value at the limb, were

Table 53.2: Deflection at the solar limb from the two 1919 stations, in seconds of arc, with the uncertainties as quoted by the authors. The Sobral astrographic plates were set aside in the original reduction; see the text.

Instrument	Deflection at the limb
Sobral, 10 cm lens	1.98 ± 0.12
Príncipe, astrographic	1.61 ± 0.30
Sobral, astrographic (set aside)	0.93
General relativity	1.75
Equivalence principle alone	0.87

The Sobral astrographic plates were rejected in the original reduction on an instrumental ground stated at the time: the coelostat mirror serving that telescope had changed figure in the heat of the day, and the star images on the eclipse plates were visibly out of focus, so that the plate scale could not be transferred reliably from the comparison epoch [465].

Phenomenon 53.2 (Measured deflection of starlight at the solar limb)**.** Starlight passing close to the Sun is deflected toward it, so that stars seen near the eclipsed limb are displaced radially outward. The two accepted 1919 determinations bracket the general-relativistic value of 1.75 seconds of arc and both exclude the value 0.87 obtained from the equivalence principle alone [465]. For an impact parameter b with $b \gg GM/c^2$ the theory gives

$$\Delta\vartheta = \frac{4GM}{c^2 b}\,. \tag{53.3}$$

Very-long-baseline radio interferometry on quasars now confirms Equation (53.3) at the level of one part in 10^4 [1903, 2145][2240], and the same bending on galactic and cluster scales produces the multiple images of gravitational lenses, of which the twin quasar 0957+561 was the first identified [1645].

Rests on Postulate 43.1 and Proposition 45.8.

Derivation. *Derives Phenomenon 53.2.*

For a null geodesic the constant term on the right of Equation (53.2) is absent — it came from the rest mass — and the orbit equation of Chapter 45 reduces to

$$\frac{\mathrm{d}^2 u}{\mathrm{d}\varphi^2} + u = \frac{3GM}{c^2}\,u^2\,.$$

With the right-hand side dropped the solution is a straight line passing at perpendicular distance b from the centre,

$$u_0 = \frac{\sin\varphi}{b}\,,$$

which indeed gives $r\sin\varphi = b$. Substituting u_0 into the right-hand side,

$$\frac{\mathrm{d}^2 u_1}{\mathrm{d}\varphi^2} + u_1 = \frac{3GM}{c^2 b^2}\sin^2\varphi = \frac{3GM}{2c^2 b^2}\,(1 - \cos 2\varphi)\,,$$

whose particular solution is

$$u_1 = \frac{3GM}{2c^2b^2}\left(1 + \tfrac{1}{3}\cos 2\varphi\right) .$$

Nothing here is resonant, and the correction stays bounded; what it does is move the asymptotes. The ray comes from infinity and returns to it where $u = 0$. Near $\varphi = 0$ the equation $u = 0$ reads

$$\frac{\varphi}{b} + \frac{3GM}{2c^2b^2}\cdot\frac{4}{3} = 0 \qquad \Longrightarrow \qquad \varphi = -\frac{2GM}{c^2b} ,$$

and near $\varphi = \pi$, writing $\varphi = \pi + \epsilon$ and using $\sin(\pi + \epsilon) \simeq -\epsilon$, it reads $\epsilon = +2GM/(c^2b)$. The total angle swept between the two asymptotes therefore exceeds π by

$$\Delta\vartheta = \frac{4GM}{c^2b} ,$$

which is Equation (53.3). Half of this comes from the Newtonian-like attraction that the equivalence principle already supplies; the other half comes from the term in u^2, that is from the curvature of space, and it is the half that the 1919 plates were looking for. Numerically, with $GM_\odot = 1.327\,12 \times 10^{20}\,\mathrm{m^3/s^2}$ and $b = R_\odot = 6.957 \times 10^8\,\mathrm{m}$, one finds $\Delta\vartheta = 8.49 \times 10^{-6}$ in radian measure, or 1.751 seconds of arc. □

53.2.4 Interpretation

The logic of the result is not that either station measured 1.75 accurately — neither did, and the quoted uncertainties are large — but that the two accepted determinations lie on opposite sides of 1.75 and that both lie many quoted uncertainties above 0.87. The experiment was designed to separate two hypotheses differing by a factor of two, and it separated them.

The rejection of the Sobral astrographic plates has been argued over ever since, and honesty requires that the argument be reported rather than settled here. The case against the rejection is that a value of 0.93 is close to the half-deflection, that the rejection was decided by people who expected the full value, and that the surviving Príncipe result rests on two usable plates. The case for it is that the defocusing of those plates was an instrumental fact recorded at the time and independent of what they showed, and that later re-measurements of the same Sobral astrographic plates with modern machines returned values consistent with the full deflection [1640] — which is the decisive point, because it means the rejection, whatever its statistical propriety, did not bias the conclusion. What settles the matter beyond argument is not the 1919 plates at all but the radio interferometry of the past four decades, which measures the same deflection with a precision four orders of magnitude better and finds Equation (53.3) [1903, 2145][2240].

53.2.5 Primary references

[465] for the expeditions and their reduction; [371] for the half-deflection and [428] for the full value; [1645] for the first gravitational lens; [2240] for the modern radio limits, quoted at second hand.

53.3 Pound, Rebka and Snider: redshift in a tower (1960–1965)

Tests Phenomenon 53.3.

The gravitational redshift is the one classical test that can be done indoors. It is also the one that tests the least: it follows from the equivalence principle and energy conservation alone, as the derivation in Chapter 42 shows, and no field equation enters. That is precisely why it is worth doing — it isolates the assumption on which everything geometric rests.

53.3.1 Apparatus

The tower of the Jefferson Physical Laboratory at Harvard, giving a height difference of 22.5 m between source and absorber. The clock is a nucleus. A source of ^{57}Co decays to ^{57}Fe, which emits a gamma ray of 14.4 keV; the same transition in a foil of iron enriched in ^{57}Fe absorbs it resonantly. Both source and absorber are bound in a crystal lattice, so that a large fraction of the emissions and absorptions are recoil-free in the sense of Mössbauer [1154] and the line retains its natural width: the excited state of ^{57}Fe has a mean life of about 141 ns, corresponding to a fractional linewidth of about 3×10^{-13}. A proportional counter behind the absorber records the transmitted intensity. The gamma path runs inside a mylar tube filled with helium, to keep absorption in air from destroying the count rate over 22.5 m, and the source is mounted on a magnetic transducer — in effect a loudspeaker cone — which moves it sinusoidally so as to impose a known first-order Doppler shift [1189, 1321].

53.3.2 Procedure

The quantity to be detected is a fractional frequency shift of gH/c^2, which for 22.5 m is 2.46×10^{-15}: some two orders of magnitude smaller than the natural width of the line being used to measure it. It is detected as a small asymmetry in the transmitted intensity between the two halves of the transducer cycle, and calibrated by the transducer velocity itself: the velocity needed to cancel the gravitational shift is $v = c\,\Delta\nu/\nu \simeq 0.74\,\mu\text{m/s}$.

The essential trick is a reversal. The whole measurement is performed once with the source at the bottom and the absorber at the top, and once the other way round. The gravitational shift changes sign; almost every systematic error does not. Differencing the two configurations therefore doubles the signal and cancels, to first order, every constant difference between the two nuclear environments — lattice chemistry, hyperfine fields, sample impurities, counter drift. What survives the reversal is a genuine dependence on the direction of propagation in the gravitational field.

One systematic error is not cancelled by the reversal, because it depends on the temperatures of the two samples rather than on their positions: the second-order Doppler shift produced by the thermal vibration of the emitting and absorbing nuclei in their lattices. Its size is derived below. It is comparable to the entire effect for a temperature difference of one kelvin, and the control of source and absorber temperatures is in practice the limiting art of the experiment.

53.3.3 Observations and data

Phenomenon 53.3 (Gravitational shift of gamma rays over a tower)**.** A 14.4 keV gamma ray propagating upward through a height $H = 22.5$ m in the Earth's field is received with its frequency lowered by the fraction $gH/c^2 = 2.46 \times 10^{-15}$, and a downward ray is received with its frequency raised by the same fraction. The 1960 measurement gave a ratio of observed to predicted shift of 1.05 ± 0.10 [1189], and the 1965 repetition with longer integration and improved control of the sample temperatures gave 0.9990 ± 0.0076 [1321].

Rests on Definition 42.2 and Phenomena 38.13 and 42.5.

The derivation of gH/c^2 itself is given in Chapter 42 and is not repeated here. What is derived instead is the systematic that limits the measurement, because it is what decides

whether the number above can be believed.

Derivation. *Derives Phenomenon 53.3.*

A nucleus bound in a lattice at temperature T is not at rest: it vibrates, and although its mean velocity vanishes its mean square velocity does not. A moving emitter is seen, to second order in v/c, with its frequency lowered by the time-dilation factor of Chapter 38,

$$\frac{\Delta\nu}{\nu} = -\frac{\langle v^2 \rangle}{2c^2} \,.$$

The sign is independent of the direction of the motion, which is why this shift does not average away over a vibration cycle and why it does not reverse when the apparatus is turned upside down. In the classical regime, equipartition over the three vibrational degrees of freedom of a nucleus of mass M gives $\frac{1}{2}M\langle v^2 \rangle = \frac{3}{2}k_{\mathrm{B}}T$, hence

$$\frac{\Delta\nu}{\nu} = -\frac{3k_{\mathrm{B}}T}{2Mc^2} \,, \qquad \frac{\mathrm{d}}{\mathrm{d}T}\left(\frac{\Delta\nu}{\nu}\right) = -\frac{3k_{\mathrm{B}}}{2Mc^2} \,. \tag{53.4}$$

For ^{57}Fe, $M = 9.46 \times 10^{-26}\,\mathrm{kg}$, and the coefficient is $-2.4 \times 10^{-15}/\mathrm{K}$.

That number is the point of the derivation. It is essentially equal to the entire gravitational effect of Phenomenon 53.3, so a systematic temperature difference of one kelvin between source and absorber would mimic or cancel the shift completely. Holding the error from this source below the one per cent finally quoted therefore requires the two samples to agree in temperature to about 10 mK, since $0.01 \times 2.46 \times 10^{-15}$ divided by the coefficient above is 1.0×10^{-2} kelvin. It is also why the effect survives the reversal test: heating one end of the tower does not care which way the gamma ray is travelling. The measurement is in this sense a thermometry experiment as much as a gravitational one. □

53.3.4 Interpretation

The redshift exists, has the predicted sign, and has the predicted size to one per cent. Because the derivation of gH/c^2 uses only the equivalence principle, kinematics and energy conservation, this result does not test the field equations at all; it tests the assumption that a local laboratory in free fall is indistinguishable from an inertial one, which is the assumption that licenses describing gravity as geometry in the first place. A theory that failed here would not need repairing but abandoning.

The measurement has been superseded many times over and always in the same direction. A hydrogen maser flown on a suborbital rocket verified the shift at the level of 7×10^{-5} [1660]; optical clocks now resolve it over height differences of less than a metre [2154] and, more recently, across a millimetre-scale atomic sample [2401]; and the same shift has been detected in the spectrum of the star S2 as it passes its closest approach to the Galactic-centre black hole [2295]. The satellite-navigation clock budget of Chapter 42 is the same effect run continuously as an engineering necessity, and the flown caesium clocks of Chapter 41 measure it summed with the kinematic dilation.

53.3.5 Primary references

[1189, 1321]; the recoil-free emission that makes the measurement possible is [1154]. The later and more precise measurements quoted in the interpretation are [1660, 2154, 2295, 2401].

53.4 Shapiro: the radar time delay (1964–2003)

Tests Phenomenon 53.4.

The fourth classical test was not available to Einstein: it requires a transmitter powerful enough to get an echo back from another planet. Shapiro pointed out in 1964 that a radar pulse whose path passes close to the Sun takes measurably longer to return than the same geometric path taken far from it [1296], and the experiment became possible within four years.

53.4.1 Apparatus

For the first measurements, planetary radar: the Haystack antenna, a steerable dish of about 37 m aperture working in the centimetre band, and the 305 m fixed reflector at Arecibo at metre wavelength, transmitting coded pulse trains at Mercury and Venus and receiving the echo some minutes later [1382]. The timing reference is an atomic frequency standard; the range model is a numerical ephemeris of the solar system.

Later measurements replace the passive planet with a spacecraft transponder, which returns a far stronger and better-defined signal: the Viking landers on Mars, ranged to over several superior conjunctions [1644], and finally the Cassini spacecraft on its cruise to Saturn, whose radio link was operated simultaneously at two uplink and three downlink frequencies [2046].

53.4.2 Procedure

Range to the target as a function of time through a superior conjunction, when the line of sight passes close to the Sun, and compare the observed round-trip delay with the delay computed from the ephemeris in flat spacetime. The excess grows logarithmically as the ray approaches the limb, peaking at conjunction, and this characteristic shape is what identifies it: it is not degenerate with an error in the planetary distance, which varies on the orbital timescale rather than with the impact parameter.

Two systematics dominate. The first is the solar corona, whose plasma delays a radio signal by an amount proportional to the inverse square of the frequency, and which is both large and variable precisely where the relativistic effect is largest. The first-generation experiments had to model it; the Cassini experiment instead cancelled it, by running the link at several widely separated frequencies and using the known dispersion of a plasma to solve the plasma delay away — which is why Cassini improved on Viking by two orders of magnitude. The second is the topography and rotation of a planetary target, which the spacecraft transponder experiments remove by construction.

53.4.3 Observations and data

Phenomenon 53.4 (Excess delay of a radar echo passing the Sun)**.** A signal travelling between two points at radii r_1 and r_2 from the Sun, past a closest approach b with $b \ll r_1, r_2$, takes longer than the same path in flat spacetime by

$$\Delta t = \frac{4GM}{c^3} \ln\!\left(\frac{4r_1 r_2}{b^2}\right) \tag{53.5}$$

for the round trip. For an Earth–Venus circuit grazing the solar limb this is of order 200 µs. The effect was proposed as a fourth test in 1964 [1296] and detected in radar echoes from Mercury and Venus within a few years, in agreement with the prediction at the level of tens of per cent [1382]; ranging to the Viking landers confirmed it to about one part in 10^3 [1644]; and the Cassini radio link, which measures the post-Newtonian parameter γ entering Equation (53.5) through an overall factor $(1+\gamma)/2$, gave $\gamma - 1 = (2.1 \pm 2.3) \times 10^{-5}$ [2046].

Rests on Postulate 43.1 and Proposition 45.14.

Derivation. *Derives Phenomenon 53.4.*

Write the Schwarzschild metric of Chapter 45 in isotropic coordinates and keep only first order in $GM/\left(c^2 r\right)$:

$$\mathrm{d}s^2 = -\left(1 - \frac{2GM}{c^2 r}\right) c^2\,\mathrm{d}t^2 + \left(1 + \frac{2GM}{c^2 r}\right)\left(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2\right) .$$

Isotropic coordinates are the right choice here because their spatial part is conformally flat, so that a coordinate straight line is the unperturbed ray and the whole effect appears in one place. Setting $\mathrm{d}s^2 = 0$ for a light ray and writing $\mathrm{d}l$ for the Euclidean coordinate length element,

$$c\,\mathrm{d}t = \mathrm{d}l\,\frac{\left(1 + GM/c^2 r\right)}{\left(1 - GM/c^2 r\right)} \simeq \mathrm{d}l\left(1 + \frac{2GM}{c^2 r}\right) .$$

The coordinate speed of light is less than c in the presence of the mass, and equally so in every direction; the factor two is one part time dilation and one part spatial curvature, exactly as in Equation (53.3).

Integrate along the unperturbed straight line, measuring l from the point of closest approach, so that $r = \sqrt{b^2 + l^2}$. The excess over the flat-space time $\mathrm{d}l/c$ is, one way,

$$\Delta t_1 = \frac{2GM}{c^3}\int_{-l_1}^{l_2} \frac{\mathrm{d}l}{\sqrt{b^2 + l^2}} = \frac{2GM}{c^3}\left[\operatorname{arcsinh}\frac{l_1}{b} + \operatorname{arcsinh}\frac{l_2}{b}\right] .$$

For $b \ll l_i$ one has $\operatorname{arcsinh}\left(l/b\right) \simeq \ln\left(2l/b\right)$ and $l_i \simeq r_i$, so the bracket is $\ln\left(4r_1 r_2/b^2\right)$; doubling for the return trip gives Equation (53.5).

Numerically, $4GM_\odot/c^3 = 1.97\times10^{-5}\,\mathrm{s}$. For a circuit from the Earth at $r_1 = 1.496\times10^{11}\,\mathrm{m}$ to Venus at $r_2 = 1.082\times10^{11}\,\mathrm{m}$ grazing the limb at $b = 6.96\times10^{8}\,\mathrm{m}$, the logarithm is 11.8 and $\Delta t = 2.3\times10^{-4}\,\mathrm{s}$. Note how weakly the answer depends on the geometry: halving the impact parameter adds only $2.7\times10^{-5}\,\mathrm{s}$, which is why the measurement needs a well-sampled conjunction rather than a single well-placed shot. □

53.4.4 Interpretation

The radar delay is the sharpest of the classical tests, and it is sharp for an unglamorous reason: it measures a time, and time is the quantity physics measures best. The Cassini bound $\gamma - 1 = (2.1 \pm 2.3)\times10^{-5}$ [2046] is the tightest constraint in the solar system on any metric theory of gravity, and every proposed alternative to general relativity must survive it before anything else is discussed. It is worth noticing what the delay and the deflection have in common: both are governed by the same coefficient γ, and both would be exactly half their observed size if space were flat and only time were curved. The two experiments are, at post-Newtonian order, measurements of the same number by entirely different means — one photographic and angular, one electronic and temporal — agreeing to the precision of the weaker.

The same delay appears wherever a signal passes a mass, and it has become a workaday astrophysical tool rather than a test: the Shapiro delay in the timing of a pulsar whose companion passes in front of it is now one of the standard ways of weighing a neutron star, as Chapter 49 describes [2155].

53.4.5 Primary references

[1296] for the proposal, [1382] for the first radar detection, [1644] for the Viking ranging and [2046] for the Cassini determination of γ.

53.5 The Eötvös lineage and MICROSCOPE (1922–2022)

Tests Phenomenon 53.5.

The universality of free fall is the foundation the whole geometric description rests on, and it is the only one of these tests that is a *null* experiment: what is measured is the absence of a differential acceleration, and the result is a bound. The bound has been driven down by six orders of magnitude in a century, by instruments that have almost nothing in common except the quantity they report.

53.5.1 Apparatus

Two families of instrument, both comparing the free fall of two bodies of different composition.

The torsion balance, in the lineage that descends from the Cavendish instrument of Chapter 33. A horizontal beam is suspended by a fine wire with test bodies of different materials at its ends; a composition-dependent horizontal force appears as a torque about the fibre and is read by an optical lever or autocollimator. In the Eötvös–Pekár–Fekete instrument, operated from 1906 to 1909 and published only in 1922, a beam some tens of centimetres long hung from a platinum-iridium wire inside a triple thermal enclosure, and platinum was compared against copper, water, asbestos, tallow and a magnesium-aluminium alloy among others [486]. In the modern Eöt-Wash instrument the whole balance sits on a slowly and continuously rotating turntable, and the test bodies are beryllium and titanium [2135].

The orbiting differential accelerometer. The MICROSCOPE satellite carried two differential electrostatic accelerometers in a sun-synchronous orbit at about 710 km. Each held two coaxial hollow cylindrical test masses whose centres were made to coincide; electrodes around each mass sensed its displacement and applied the electrostatic force needed to hold it centred, and that force is the measurement. One accelerometer held masses of two different alloys — a platinum-rhodium alloy against a titanium alloy — and the other held two masses of the same platinum-rhodium alloy, serving as a null reference that should show no signal whatever [2289, 2409].

53.5.2 Procedure

In both families the measured quantity is the Eötvös parameter

$$\eta_{AB} := \frac{2\,(a_A - a_B)}{a_A + a_B}\,, \tag{53.6}$$

and in both the art is to modulate it. A null experiment cannot be done by measuring once, because everything that is not the signal is also a constant: the signal must be made to vary in a known way while the instrument does not.

In the terrestrial torsion balance the modulation is geometric. The attractor is the Earth, or the Sun, or the Galaxy; the source direction rotates with respect to the instrument as the Earth turns, so a violation appears as a torque varying with a diurnal or annual period. Eötvös modulated by hand, turning the beam through 180° in azimuth and differencing the two readings, which reverses the signal but not the instrumental torques. The Eöt-Wash balance modulates continuously, rotating at a period of minutes so that the signal appears as a torque oscillating at the turntable frequency, far above the drift of the fibre and far from every fixed offset.

In orbit the modulation is faster and cleaner. The satellite is spun about an axis perpendicular to the orbital plane, so that the Earth-pointing direction rotates in the instrument frame at the sum of the orbital and spin frequencies. A violation of universality would appear as a differential acceleration between the two test masses at that one known

frequency and in that one known phase; everything else — thermal drift, radiation pressure, residual gas — has to conspire to appear at exactly the same frequency and phase to be mistaken for it. The satellite is drag-free, its thrusters cancelling the atmospheric drag so that the test masses are truly in free fall.

53.5.3 Observations and data

Every experiment in the lineage has returned a null result. The bounds on η, quoted at the order of magnitude their authors claim:

Table 53.3: Bounds on the Eötvös parameter Equation (53.6), by experiment. Every entry is consistent with zero; the column gives the order of magnitude of the limit, not the central value.

Experiment	Bodies and attractor	Bound on $\lvert\eta\rvert$
Eötvös, Pekár and Fekete, published 1922 [486]	platinum against several, Earth	10^{-9}
Roll, Krotkov and Dicke, 1964 [1294]	aluminium against gold, Sun	10^{-11}
Braginsky and Panov, 1972 [1469]	aluminium against platinum, Sun	10^{-12}
Eöt-Wash rotating balance, 2008 [2135]	beryllium against titanium, Earth	10^{-13}
MICROSCOPE, first results, 2017 [2289]	titanium against platinum alloy, Earth	10^{-14}
MICROSCOPE, final results, 2022 [2409]	titanium against platinum alloy, Earth	10^{-15}

Phenomenon 53.5 (Universality of free fall, as measured)**.** No experiment has ever detected a dependence of gravitational acceleration on composition. The Eötvös parameter Equation (53.6) is consistent with zero in every measurement listed in Table 53.3, and the strongest bound now stands at the level of parts in 10^{15}, set by comparing titanium and platinum-alloy test masses in orbital free fall [2409]. The same experiments constrain composition dependence with the Sun and with the Galactic centre as attractor, so the null result is not an accident of one source [1294, 1469].

Rests on Definition 42.2, Phenomenon 42.1, and Equation (53.8).

Derivation. *Derives Phenomenon 53.5.*

The bound is established by the torque analysis derived below: Equation (53.8) converts any composition dependence of free fall into a torque about the fibre of magnitude $\frac{1}{2}\,\eta_{AB}\,m\,L\,a_{\mathrm{h}}$, in which every factor multiplying η_{AB} is fixed by the geometry of the instrument and by Equation (53.7). A recorded torque consistent with zero therefore bounds $\lvert\eta_{AB}\rvert$ directly. The orbital experiments read the same null through the electrostatic suspension force of a differential accelerometer, the driving acceleration being the full orbital g rather than the horizontal centrifugal residue, and each entry of Table 53.3 is such a null readout divided by the known sensitivity factor of its instrument; no further theory enters. □

The derivation that a null η is equivalent to the identity of inertial and gravitational mass is given in Chapter 42. What is derived here is the signal the torsion balance actually looks for, since the geometry is not obvious and it explains why the instrument is sensitive at all.

Derivation. Work in the frame rotating with the Earth, at geographic latitude λ. A body at rest there is acted on by the true gravitational attraction $\boldsymbol{g}$, directed very nearly at the

centre, and by the centrifugal acceleration of magnitude $\omega^2 R\cos\lambda$ directed away from the rotation axis. The centrifugal term is proportional to the *inertial* mass and the attraction to the *gravitational* mass, which is what makes the arrangement a test. Resolving the centrifugal acceleration into vertical and horizontal parts, its horizontal component points toward the equator and has magnitude

$$a_{\mathrm{h}} = \omega^2 R \sin\lambda\cos\lambda\,. \tag{53.7}$$

With $\omega = 7.292\times10^{-5}\,\mathrm{s}^{-1}$ and $R = 6.371\times10^{6}\,\mathrm{m}$ this is $1.69\times10^{-2}\,\mathrm{m/s^2}$ at latitude 45°, that is 1.7×10^{-3} of g.

Write $m_{\mathrm{G}} = m_{\mathrm{I}}\,(1+\varepsilon)$ for each body, and let the local vertical be defined by the plumb line of a body with $\varepsilon = 0$; that vertical lies along $\boldsymbol{g} + \boldsymbol{a}_{\mathrm{cf}}$, so in the plumb frame the horizontal component of $\boldsymbol{g}$ has magnitude a_{h} and points away from the equator, exactly cancelling the centrifugal horizontal component. A body with $\varepsilon \neq 0$ has effective acceleration $(1+\varepsilon)\,\boldsymbol{g} + \boldsymbol{a}_{\mathrm{cf}}$, which differs from the plumb-frame vertical by the horizontal residue $\varepsilon\, a_{\mathrm{h}}$. Two bodies at the ends of a beam of length L suspended at its centre and oriented east–west therefore feel horizontal forces along the north–south direction differing by $m\eta_{AB}a_{\mathrm{h}}$, and since their position vectors from the fibre are $\pm\,(L/2)$ along east, the net torque about the fibre is

$$N = \tfrac{1}{2}\,\eta_{AB}\,m\,L\,a_{\mathrm{h}} = \tfrac{1}{2}\,\eta_{AB}\,mgL\,\frac{\omega^2 R\sin\lambda\cos\lambda}{g}\,. \tag{53.8}$$

The instrument is therefore looking for a torque smaller than the natural scale mgL by the product of η with 1.7×10^{-3} — and η is the quantity being bounded at 10^{-9} or below. That factor 1.7×10^{-3} is the whole sensitivity penalty of doing the experiment on a rotating planet rather than in orbit, and it is the reason a satellite gains several orders of magnitude the moment it works at all: in orbit the driving acceleration is the full local g, about $8\,\mathrm{m/s^2}$ at MICROSCOPE's altitude, rather than $1.7\times10^{-2}\,\mathrm{m/s^2}$.

Two further features of Equation (53.8) matter experimentally. The torque vanishes at the equator and at the poles and is maximal near 45° of latitude, which is a signature no instrumental effect shares; and it reverses when the beam is turned through 180° in azimuth, which is the differencing Eötvös performed. □

53.5.4 Interpretation

The null result is the evidence that licenses geometry. If $m_{\mathrm{G}}/m_{\mathrm{I}}$ is one universal number, then the trajectory of a freely falling body depends on nothing but where it started and how fast it was going — not on what it is made of — and a family of trajectories with that property can be the geodesics of a connection. If η were nonzero at any level, gravity would be a force with a composition-dependent coupling like every other force, and the geometric description of Chapter 43 would be an approximation valid to whatever precision η had last been bounded at. The bound is therefore not a curiosity but the load-bearing measurement of the entire part.

It should be said what these experiments do *not* do. They test the weak form of the equivalence principle, for laboratory bodies whose own gravitational binding energy is negligible. Whether a self-gravitating body falls the same way is a separate question, tested by lunar laser ranging in the next section. And a null result is a bound, never a confirmation: what Table 53.3 records is that a century of increasingly ingenious experiments has failed to find a violation, which is the strongest statement an experiment of this kind can make.

53.5.5 Primary references

[486, 1294, 1469] and [2135, 2289, 2409]. The bounds of Table 53.3 are quoted from those papers at the order of magnitude only; the central values and their separate statistical and systematic uncertainties are given in the sources.

53.6 Lunar laser ranging and Gravity Probe B (1969–2011)

Tests Phenomena 42.4 and 53.6.

The last two experiments test what happens to a *direction* carried through curved spacetime. A gyroscope in orbit, or the orbit of the Moon itself considered as a gyroscope, is parallel-transported along its worldline, and general relativity says that the transported axis does not return to where it started. There are two such effects: the geodetic precession, produced by the curvature due to the central mass, and the very much smaller frame-dragging precession, produced by the rotation of that mass [455, 458]. Lunar laser ranging also provides the one test in this chapter of the equivalence principle for a self-gravitating body.

53.6.1 Apparatus

Lunar laser ranging uses retroreflector arrays left on the Moon — by the Apollo 11, 14 and 15 crews, and on the two Soviet Lunokhod rovers — consisting of fused-silica corner cubes, which return light along the direction it came from regardless of the array's orientation. The Earth end is a telescope firing short laser pulses and timing the return of single photons against an atomic clock; the round trip takes about 2.5 s, and only a few returned photons are recorded per pulse train [2067, 2198].

Gravity Probe B carried four gyroscopes into a polar orbit at about 642 km altitude. Each rotor was a sphere of fused quartz 3.8 cm in diameter, coated with superconducting niobium, electrostatically suspended in vacuum inside a dewar of superfluid helium at about 1.8 K, and spun up to tens of revolutions per second. The rotors were figured to within a few tens of nanometres of a perfect sphere. The spin direction was read out magnetically: a spinning superconductor develops a London moment along its spin axis, and the field of that moment was sensed by a SQUID magnetometer. An on-board telescope held the satellite pointed at the guide star IM Pegasi, providing the fixed direction against which the gyroscope drift was measured, and the satellite was drag-free [2173].

53.6.2 Procedure

For lunar laser ranging the observable is a time series of Earth–Moon distances accumulated over decades and fitted with a full solar-system and lunar-orbit model. Two relativistic signatures are extracted from it. A violation of the equivalence principle for self-gravitating bodies — the Nordtvedt effect [1377] — would make the Earth and the Moon fall toward the Sun at slightly different rates, polarizing the lunar orbit along the Earth–Sun line with a signature periodic at the synodic month. The geodetic precession appears as a slow rotation of the lunar orbit plane, and is separated from Newtonian precessions by its rate and direction.

For Gravity Probe B the observable is the angle between each gyroscope spin axis and the line to IM Pegasi, tracked for the duration of the science mission. The two relativistic drifts are orthogonal by construction, which is why a polar orbit was chosen: the geodetic drift lies in the orbital plane, north–south, and the frame-dragging drift is perpendicular to it, east–west. The guide star has its own proper motion, which was measured independently by very-long-baseline radio interferometry against distant quasars and subtracted.

53.6.3 Observations and data

Phenomenon 53.6 (Precession of a gyroscope in orbit)**.** A gyroscope in a circular orbit of radius r about a mass M with angular momentum J precesses at the geodetic rate

$$\Omega_{\text{geo}} = \frac{3}{2}\,\frac{GM}{c^2 r}\,\sqrt{\frac{GM}{r^3}} \tag{53.9}$$

in the plane of the orbit, and, for a polar orbit, at the frame-dragging rate

$$\Omega_{\mathrm{fd}} = \frac{GJ}{2c^2 r^3} \tag{53.10}$$

perpendicular to it [1191]. Gravity Probe B measured both against the guide star IM Pegasi and found, in milliarcseconds per year, -6601.8 ± 18.3 for the geodetic drift against a predicted -6606.1, and -37.2 ± 7.2 for the frame-dragging drift against a predicted -39.2 [2173]. Lunar laser ranging sees the geodetic precession of the lunar orbit at the level of a few parts in 10^3 and, independently, finds the Nordtvedt parameter consistent with zero at the level of parts in 10^4, so that self-gravitating bodies fall like test bodies [2067][1377].

Rests on Postulate 43.1 and Theorem 45.30.

Derivation. *Derives Phenomenon 53.6.*

Both rates follow from parallel transport of the spin four-vector in the weak-field metric of a slowly rotating mass — the linearization of the Kerr geometry, which is all the Earth requires. Write, to first order in Φ/c^2 and in the rotation,

$$\mathrm{d}s^2 = -\Big(1 + \frac{2\Phi}{c^2}\Big)c^2\mathrm{d}t^2 + \Big(1 - \frac{2\Phi}{c^2}\Big)\delta_{ij}\,\mathrm{d}x^i\mathrm{d}x^j + 2h_i\,c\,\mathrm{d}t\,\mathrm{d}x^i\,, \tag{53.11}$$

with $\Phi = -GM/r$ and the gravitomagnetic potential

$$\boldsymbol{h} = -\frac{2G}{c^3}\,\frac{\boldsymbol{J}\times\boldsymbol{r}}{r^3}\,. \tag{53.12}$$

A torque-free gyroscope on a geodesic parallel-transports its spin, $\mathrm{d}S^\mu/\mathrm{d}\tau + \Gamma^\mu{}_{\nu\rho}u^\nu S^\rho = 0$, with $S_\mu u^\mu = 0$. To the order required, the Christoffel symbols of Equation (53.11) are $\Gamma^i{}_{00} = \partial_i\Phi/c^2$, $\Gamma^i{}_{0j} = \tfrac{1}{2}(\partial_j h_i - \partial_i h_j)$ and $\Gamma^i{}_{jk} = (-\delta_{ij}\partial_k\Phi - \delta_{ik}\partial_j\Phi + \delta_{jk}\partial_i\Phi)/c^2$, while the constraint gives $S^0 = \boldsymbol{S}\cdot\boldsymbol{v}/c$ at leading order. Assembling the spatial components with $u^\mu \approx (c, \boldsymbol{v})$,

$$\frac{\mathrm{d}S^i}{\mathrm{d}t} = -\frac{c}{2}(\partial_j h_i - \partial_i h_j)S^j + \frac{1}{c^2}\Big[(\boldsymbol{v}\cdot\nabla\Phi)S^i + v^i(\boldsymbol{S}\cdot\nabla\Phi) - 2\,\partial_i\Phi\,(\boldsymbol{v}\cdot\boldsymbol{S})\Big]\,.$$

The coordinate components mix a genuine rotation with the stretching of the coordinate grid; the physical spin is the one referred to the local orthonormal rest frame, $\hat{S}^i = (1 - \Phi/c^2)S^i - \frac{1}{2c^2}v^i(\boldsymbol{v}\cdot\boldsymbol{S})$, the first factor from the spatial metric in Equation (53.11), the second the low-velocity boost to the gyroscope frame. Differentiating, using $\mathrm{d}\boldsymbol{v}/\mathrm{d}t = -\nabla\Phi$ on the geodesic, and keeping first order, every non-rotational term cancels and what survives is a pure precession,

$$\frac{\mathrm{d}\hat{\boldsymbol{S}}}{\mathrm{d}t} = \boldsymbol{\Omega}\times\hat{\boldsymbol{S}}\,, \qquad \boldsymbol{\Omega} = \underbrace{\frac{3}{2}\,\frac{GM}{c^2 r^3}\,\boldsymbol{r}\times\boldsymbol{v}}_{\text{geodetic}} + \underbrace{\frac{G}{c^2 r^3}\big[3(\boldsymbol{J}\cdot\hat{\boldsymbol{r}})\hat{\boldsymbol{r}} - \boldsymbol{J}\big]}_{\text{frame dragging}}\,, \tag{53.13}$$

the second term being $-\frac{c}{2}\nabla\times\boldsymbol{h}$ evaluated from Equation (53.12). For a circular orbit, $|\boldsymbol{r}\times\boldsymbol{v}| = r^2\sqrt{GM/r^3}$, so the geodetic magnitude is $\Omega_{\mathrm{geo}} = \frac{3}{2}(GM/c^2 r)\sqrt{GM/r^3}$, which is Equation (53.9), directed along the orbit normal. For a polar orbit the frame-dragging term must be averaged over the orbit: with $\hat{\boldsymbol{r}}$ sweeping a great circle through the poles, $\langle(\boldsymbol{J}\cdot\hat{\boldsymbol{r}})\hat{\boldsymbol{r}}\rangle = \boldsymbol{J}/2$, so $\langle\boldsymbol{\Omega}_{\mathrm{fd}}\rangle = (G/c^2 r^3)(\frac{3}{2} - 1)\boldsymbol{J} = G\boldsymbol{J}/(2c^2 r^3)$, which is Equation (53.10). The same $\boldsymbol{\Omega}$ applied to the Earth–Moon system as a gyroscope (its orbital angular momentum the spin) gives the geodetic precession of the lunar orbit that lunar laser ranging observes. □

Granted Equations (53.9) and (53.10), the predicted rates follow from the orbit and from the Earth's mass and spin. Taking $r = 7.013\times10^6\,\mathrm{m}$ for an altitude of 642 km, and $GM_\oplus = 3.986\times10^{14}\,\mathrm{m^3/s^2}$, the orbital angular velocity is $1.075\times10^{-3}/\mathrm{s}$ and $GM_\oplus/\,(c^2 r) =$

6.32×10^{-10}, so Equation (53.9) gives 1.02×10^{-12}/s, which is 3.22×10^{-5} radian per year, or about 6640 milliarcseconds per year. With the Earth's spin angular momentum $J_\oplus \simeq 5.86 \times 10^{33}\,\mathrm{kg\,m^2/s}$, Equation (53.10) gives 6.3×10^{-15}/s, about 41 milliarcseconds per year. Both agree with the mission's own predictions to within a few per cent; the residual difference comes from the orbital eccentricity, the Earth's oblateness and the solar geodetic term, all carried by the mission model and none by the two-line estimate made here.

53.6.4 Interpretation

The geodetic result is a clean measurement: a fractional uncertainty of about 0.3 %, agreeing with the prediction. The frame-dragging result must be reported more carefully, and this chapter would be dishonest if it were not.

Gravity Probe B was designed to reach an uncertainty below one milliarcsecond per year on both drifts, which would have made the frame-dragging measurement a one-per-cent test. It did not. The delivered uncertainties are 18.3 and 7.2 milliarcseconds per year, some twenty times the design goal, so the frame-dragging number is a detection at about five times its own uncertainty and a confirmation of the predicted rate at the level of roughly twenty per cent — not one per cent.

The cause is instructive and was not anticipated. The niobium coating on the rotors and the surrounding housing carried patches of varying electric potential, and the interaction of those patches produced classical electrostatic torques on the spinning spheres. Because the rotors' polhode motion — the wobble of the spin axis within the body — drifted slowly in period over the mission, these torques were not constant and could not be removed by a simple fit. The analysis had to build an explicit model of the patch-induced misalignment torque and of the polhode phase for each rotor, and subtract it; the final error budget is dominated not by measurement noise but by the uncertainty of that model [2173]. This is worth stating in a treatise that grades its evidence: the geodetic precession is measured, and the frame-dragging precession is detected but only loosely measured, by this experiment.

What lunar laser ranging adds is different in kind. The Earth and the Moon store different fractions of their mass-energy as gravitational binding energy, so they are the two test bodies of an equivalence experiment that the balances of Section 53.5 cannot perform: laboratory masses have no appreciable self-gravity. The null Nordtvedt signal therefore extends Phenomenon 53.5 from the weak to the strong form of the principle, which no other experiment in this chapter touches [2067].

53.6.5 Primary references

[2173] for Gravity Probe B, [2067] for lunar laser ranging, and [1377] for the effect the latter bounds.

53.7 Summary of the evidence

Three remarks close the chapter. The first is that the tests are not redundant. Every row of Table 53.4 could in principle have failed while the others passed, and the third column says how; a theory that reproduced the redshift and the deflection but not the perihelion advance would have been a live possibility in 1919 and is excluded now by radar and by Mercury together.

The second is that the classical tests are all weak-field. The dimensionless potential $GM/\left(c^2 r\right)$ reaches at most a few parts in 10^6, at the solar limb, in every experiment reported here — the Galactic-centre orbit mentioned in Section 53.1 is the exception, and it belongs to Chapter 50 — and the deviations measured are first-order in it. General relativity is here being tested where it is nearly Newtonian, and passing such tests is necessary but

Table 53.4: The classical tests reported in this chapter. The third column states what would have to be wrong for the measurement to come out differently — which is what distinguishes the tests from one another, since all six agree with general relativity.

Experiment	Precision reached	What it constrains
Mercury perihelion, 1859	a residual of 43 out of 5600 seconds of arc per century	the field equations, at first post-Newtonian order
Eclipse of 1919	enough to resolve a factor of two	the curvature of space; excludes the equivalence principle acting alone
Pound–Rebka–Snider	one per cent	the equivalence principle alone; no field equation enters
Shapiro delay to Cassini	$\gamma - 1$ to 2×10^{-5}	the same curvature coefficient as the deflection, by an independent method
Eötvös to MICROSCOPE	η below 10^{-15}	universality of free fall for laboratory bodies; licenses the geometric description
Ranging and Gravity Probe B	0.3 % geodetic, about 20 % frame-dragging	parallel transport of a direction; the gravitomagnetic sector; the strong form of the principle

very far from sufficient. The strong-field evidence is elsewhere: in the orbital decay and timing of binary pulsars, in the direct gravitational-wave detections of Chapter 47, and in the horizon-scale observations of Chapter 50.

The third is about the shape of the evidence rather than its content. The best of these measurements are not the most famous ones. The 1919 eclipse made general relativity a public fact and is the weakest result in the chapter; the Cassini radio link is the strongest and is known to almost nobody. That asymmetry is a standing hazard for a treatise organized around evidence, and it is why Table 53.4 reports precision reached rather than historical importance.

Part VI

Classical Thermodynamics

Chapter 54

Classical Thermodynamics

Thermodynamics is the physics of heat, work and temperature, stated for macroscopic bodies and without any hypothesis about the constitution of matter. Its raw material is a handful of state variables; its whole content is four laws; and its working machinery is the family of thermodynamic potentials generated from the internal energy by Legendre transforms—the same construction that carries the Lagrangian to the Hamiltonian in Chapter 22. The subject was forced into existence by the steam engine: Carnot's memoir on the motive power of fire [76] predates any correct concept of energy, and the theory built on it has survived every later revolution of physics untouched. The canonical modern exposition is [1749].

This chapter develops the phenomenological theory: equilibrium and temperature, the equations of state, the first and second laws with their processes and cycles, entropy, the potentials and the Maxwell relations, phase equilibria, and the third law. The molecular reading of the same quantities begins in Chapter 55 and is completed in Chapter 118; the experiment that grounds the first law has its own chapter, Chapter 56.

Derivation pending #1

Classical Thermodynamics: all derivations of this chapter are pending.

54.1 Systems, equilibrium and temperature

54.1.1 State variables and equilibrium

[Reserved: thermodynamic systems, walls and reservoirs; extensive and intensive variables; the equilibrium postulate that a finite set of state variables $(U, V, N, \dots)$ suffices to describe a macroscopic body at rest, in the axiomatic form of [1749]; quasi-static and reversible processes; exact and inexact differentials, with heat and work as path functions and the state variables as point functions.]

54.1.2 The zeroth law and empirical temperature

[Reserved: thermal contact and thermal equilibrium; the zeroth law—transitivity of thermal equilibrium—as stated by Maxwell in his textbook of the subject [179] and later named the zeroth law by Fowler and Guggenheim [835]; the resulting existence of an empirical temperature as a label on isotherms; thermometric substances and the gas scale.]

Phenomenon 54.1 (Transitivity of thermal equilibrium)**.** Let three bodies A, B, C be brought into thermal contact in pairs. Without exception it is found that if A is in thermal equilibrium with C, and B is in thermal equilibrium with C, then A and B are already in thermal equilibrium when placed in contact: no further change of any of their state variables

occurs. This is the fact that makes a thermometer possible—C being a small standard instrument read in place of the bodies themselves. It was made the ground of thermometry by Maxwell [179] and named the zeroth law by Fowler and Guggenheim [835].

Derivation. Write $A \sim B$ for "A and B are in thermal equilibrium". The relation is reflexive and symmetric by construction, and the observed regularity is precisely transitivity; hence $\sim$ is an equivalence relation on the set of equilibrium states, and its classes partition that set. For a simple fluid, whose equilibrium states are labelled by the pair (P, V), equilibrium with one fixed reference body is a single condition on two variables, so each class is a curve in the (P, V) plane—an isotherm. Choose a thermometric body together with a coordinate that varies monotonically along its own states, and assign to every class the value θ that coordinate takes when the reference body is equilibrated with a member of the class. By construction

$$\theta(P_A, V_A) = \theta(P_B, V_B) \iff A \sim B\,. \tag{54.1}$$

Any second assignment θ' with the same property labels the same classes, so $\theta' = g(\theta)$ with g strictly monotone: empirical temperature is fixed only up to a monotone relabelling. Removing that residual freedom is the work of Section 54.4.3. □

54.1.3 Thermometry and the kelvin

[Reserved: William Thomson's absolute thermometric scale founded on the Carnot cycle [123], anticipating Section 54.4.3; the constant-volume gas thermometer as its classical realisation; the modern kelvin, defined since 2019 by fixing the Boltzmann constant at $1.380\,649 \times 10^{-23}$ J/K [2333], and its relation to the measurement conventions of Chapter 2.]

54.2 Equations of state

54.2.1 The gas laws

[Reserved: Boyle's law $PV = \text{const}$ at fixed temperature, from the pressure–volume data of the 1662 defence of the spring of the air [6]; Gay-Lussac's law of thermal expansion, with the equal expansivity of all gases [47] and Gay-Lussac's own attribution of earlier unpublished work to Charles; Dalton's law of partial pressures [46]; Avogadro's hypothesis that equal volumes at equal temperature and pressure hold equal numbers of molecules [55].]

54.2.2 The ideal-gas equation of state

[Reserved: the combination of the gas laws into the single equation $PV = nRT$ by Clapeyron [95]; the gas constant, exact since 2019 as the product of the Boltzmann and Avogadro constants, $8.314\,462\,618$ J/(mol K) [2333, 2431]; Regnault's precision measurements of the small departures of real gases from the law [122], the experimental seed of Section 54.6.3; the ideal gas as the universal dilute limit.]

Phenomenon 54.2 (The universal dilute-gas equation of state)**.** Every gas, whatever its chemical species, obeys one and the same equation of state in the limit of low density. Three separate observations enter it: at fixed temperature the product PV of a fixed quantity of gas is constant [6]; at fixed pressure the volume grows with temperature by a coefficient of thermal expansion that is the same for all gases [47]; and at fixed pressure and temperature equal volumes of different gases contain equal numbers of molecules [55]. Together they say that

$$PV = nRT\,, \tag{54.2}$$

with one constant R common to every substance. Departures from Equation (54.2) set in at finite density and were mapped systematically by Regnault [122]; they are the subject

of Section 54.6.3. Since the redefinition of the SI in 2019 the gas constant is exact, $R = 8.314\,462\,618\,\mathrm{J/(mol\,K)}$ [2333, 2431].

Derivation. Fix the amount of gas. Boyle's observation says that on each isotherm the product PV depends on the empirical temperature alone, $PV = f(\theta)$, so it is a label on the classes of Equation (54.1). Gay-Lussac's says that at fixed pressure $V = V_0(1 + \alpha\theta_{\mathrm{C}})$, with θ_{C} the Celsius reading and α the same number for every gas; defining

$$T := \theta_{\mathrm{C}} + \alpha^{-1} \tag{54.3}$$

turns it into $V \propto T$ at fixed P. The two together force f to be proportional to T, so $PV = CT$ with C depending only on how much gas is present. Avogadro's observation fixes that dependence: at given P and T the volume is proportional to the number of molecules, hence $C = nR$ with R independent of the species, which is Equation (54.2). The scale Equation (54.3) so constructed is the ideal-gas temperature; that it coincides with the substance-independent scale of thermodynamics is shown in Section 54.4.3, and only then is T more than a convention. □

54.3 The first law

54.3.1 Work, heat and internal energy

[Reserved: the first law $\mathrm{d}U = \delta Q - \delta W$: internal energy as a state function whose change splits into heat and work, neither of which is one by itself; Mayer's inference of the equivalence of heat and work from the two heat capacities of air [108]; Joule's paddle-wheel measurement of the mechanical equivalent [129], treated in full in Chapter 56; Helmholtz's statement of the conservation of energy across all of physics [121].]

Phenomenon 54.3 (Path independence of adiabatic work)**.** Carry a thermally insulated body from one equilibrium state to another by doing work on it. The work required is found to depend only on the two states and not at all on how it is delivered: a paddle wheel churning water, the same wheel churning mercury, cast iron rubbing on cast iron, and an electric current driven by a falling weight all produce the same change of state for the same work [109, 129]. Equivalently, heat and work are interconvertible at one fixed rate—the mechanical equivalent of heat—first inferred from the two heat capacities of air by Mayer [108] and then measured directly by Joule [129]. The measurement is treated in full in Chapter 56.

Derivation. Let $W_{\mathrm{ad}}(A \to B)$ denote the work done *by* an insulated system in passing from state A to state B; the observation is that this number depends on A and B alone. Fix a reference state O and define

$$U(B) := U(O) - W_{\mathrm{ad}}(O \to B)\,, \tag{54.4}$$

which is well posed precisely because of the observed path independence, and which fixes a state function U up to the additive constant $U(O)$. Join now the same two states by a process that is *not* insulated; the work W done by the system differs from W_{ad}, and the deficit

$$Q := U(B) - U(A) + W \tag{54.5}$$

defines the heat absorbed. It is an amount of energy, measured in joules, and owns no independent unit; in differential form Equation (54.5) is the first law $\mathrm{d}U = \delta Q - \delta W$. The converse shows why the observation is not a triviality: if W_{ad} depended on the path, one could carry the system from A to B along the costlier route and back along the cheaper one, returning it to A while extracting net work from nothing—perpetual motion of the first kind, which no mechanism has ever achieved. □

54.3.2 Heat capacities

[Reserved: heat capacities at constant volume and constant pressure; enthalpy; the Mayer relation $C_P - C_V = nR$ for the ideal gas, the relation from which Mayer extracted his equivalent of heat [108]; Regnault's calorimetric data [122]; the ratio $\gamma = C_P/C_V$ and its appearance in the adiabatic law and in Laplace's correction of Newton's speed of sound [59].]

54.3.3 Processes and cycles

[Reserved: isothermal, isobaric, isochoric and adiabatic processes of the ideal gas; the adiabatic law $PV^\gamma = \text{const}$; work as the area of the indicator diagram, the representation Clapeyron introduced [95]; free expansion into vacuum as the standard irreversible contrast, tested by Gay-Lussac [52] and by Joule, and refined into the throttling experiment of Section 54.6.4.]

54.4 The second law

54.4.1 Carnot's theorem

[Reserved: the Carnot cycle of two isotherms and two adiabats; Carnot's theorem that no engine working between two reservoirs outperforms a reversible one, and that all reversible engines between the same reservoirs share one efficiency [76]—an argument that survived the caloric theory it was written in; Clapeyron's analytic and graphical restatement [95].]

54.4.2 The Clausius and Kelvin statements

[Reserved: Clausius's 1850 reconciliation of Carnot's theorem with the first law, and his statement that heat does not pass by itself from a colder to a hotter body [127]; Thomson's statement that no cyclic process converts heat wholly into work [141]; the proof that the two statements are equivalent; perpetual motion of the second kind.]

Phenomenon 54.4 (The two impossibilities)**.** Two negative statements about cyclic processes have never been contradicted by any device built or proposed.

1. *Clausius.* No process has as its sole result the transfer of heat from a colder to a hotter body [127].
2. *Kelvin.* No cyclic process has as its sole result the complete conversion into work of heat drawn from a single reservoir at one temperature [141].

Neither is a consequence of the first law, which permits both: an engine converting the thermal energy of the ocean entirely into work would conserve energy exactly. They are an independent fact about Nature, and the modern form of the impossibility Carnot had already taken for granted [76].

Derivation. The two statements are equivalent, so the observation is a single one. Suppose Kelvin's statement false: some cyclic device draws heat $Q > 0$ from the cold reservoir and delivers work $W = Q$ with no other effect. Dissipate that work as heat in the hot reservoir by friction. The compound operation is cyclic and its only result is the passage of Q from cold to hot, so Clausius's statement fails. Conversely, suppose Clausius's statement false: some device passes heat Q_c from the cold reservoir to the hot one with no other effect. Beside it run an ordinary engine that takes Q_h from the hot reservoir, delivers work $W = Q_\mathrm{h} - Q_\mathrm{c}$ and rejects exactly Q_c to the cold one. In the compound cycle the cold reservoir is left unchanged, and the sole result is the conversion into work of $Q_\mathrm{h} - Q_\mathrm{c}$ drawn from the hot reservoir alone, so Kelvin's statement fails. Each denial implies the other, hence the two statements stand or fall together. □

54.4.3 Thermodynamic temperature and the Carnot efficiency

[Reserved: the universality of the Carnot efficiency as the definition of a temperature independent of every thermometric substance, with $\eta = 1 - T_2/T_1$ [123, 141]; the proof that this scale coincides with the ideal-gas scale of Section 54.2.2.]

Phenomenon 54.5 (The Carnot limit on heat engines)**.** A device working in a cycle between two reservoirs held at temperatures $T_\mathrm{h} > T_\mathrm{c}$ never converts into work more than the fraction

$$\eta_\mathrm{C} = 1 - \frac{T_\mathrm{c}}{T_\mathrm{h}} \tag{54.6}$$

of the heat it draws from the hot reservoir; real engines fall below it, and every engine that is reversible attains it exactly, whatever its working substance and whatever the mechanism [76, 141]. It is the independence from the working substance that lets Equation (54.6) be read backwards, as the definition of a temperature scale owing nothing to any thermometric material [123].

Derivation. Universality. Let E be reversible and X arbitrary, both cycling between the same two reservoirs, and choose the number of cycles of each so that they deliver the same total work W. Run X forward and E backward. The compound device is cyclic and delivers no net work, so by the first law the heat it takes from the cold reservoir equals the heat it gives to the hot one, namely $Q_\mathrm{h}^E - Q_\mathrm{h}^X$. Now $Q_\mathrm{h}^X = W/\eta_X$ and $Q_\mathrm{h}^E = W/\eta_E$, so if $\eta_X > \eta_E$ that quantity is positive and the sole result of the compound cycle is a passage of heat from the colder reservoir to the hotter, contradicting Phenomenon 54.4. Hence $\eta_X \le \eta_E$; taking X reversible as well and exchanging the roles gives equality. All reversible engines between the same pair of reservoirs therefore share one efficiency, a function of the reservoirs alone.

Evaluation. That efficiency may then be computed with any convenient substance. Take one mole of ideal gas, with C_V constant for simplicity, around the cycle $1 \to 2$ isothermal at T_h, $2 \to 3$ adiabatic, $3 \to 4$ isothermal at T_c, $4 \to 1$ adiabatic. The internal energy of an ideal gas depends on T alone (Chapter 56), so on an isotherm the heat absorbed equals the work done and

$$Q_\mathrm{h} = RT_\mathrm{h} \ln \frac{V_2}{V_1}, \qquad Q_\mathrm{c} = RT_\mathrm{c} \ln \frac{V_3}{V_4}, \tag{54.7}$$

Q_c being the heat rejected. On an adiabat $C_V \mathrm{d}T = -P\mathrm{d}V = -RT\,\mathrm{d}V/V$, which integrates to $TV^{R/C_V} = \text{const}$; the two adiabats give $T_\mathrm{h}V_2^{R/C_V} = T_\mathrm{c}V_3^{R/C_V}$ and $T_\mathrm{h}V_1^{R/C_V} = T_\mathrm{c}V_4^{R/C_V}$, and dividing one by the other leaves $V_2/V_1 = V_3/V_4$. The logarithms in Equation (54.7) therefore cancel in the ratio,

$$\frac{Q_\mathrm{c}}{Q_\mathrm{h}} = \frac{T_\mathrm{c}}{T_\mathrm{h}}, \tag{54.8}$$

and $\eta = 1 - Q_\mathrm{c}/Q_\mathrm{h}$ is Equation (54.6). Since the left-hand side of Equation (54.8) was just shown to be independent of the working substance, while the T on its right is the ideal-gas temperature Equation (54.3), the thermodynamic and ideal-gas scales agree wherever both are defined. □

54.4.4 Entropy and the Clausius inequality

[Reserved: the Clausius equality, that $\delta Q/T$ integrates to zero around any reversible cycle [142]; entropy as the resulting state function, named by Clausius in 1865 together with the summary that the energy of the world is constant and its entropy tends to a maximum [162]; the Clausius inequality and the increase of entropy in isolated systems; the entropy of the ideal gas.]

54.4.5 Carathéodory's axiomatics

[Reserved: the axiomatic route of Carathéodory, in which the second law is the existence, near every state, of states inaccessible by adiabatic processes, and entropy arises as an integrating factor of δQ [353]; its equivalence with the engine statements; what the axiomatisation clarifies and what it hides.]

54.5 Thermodynamic potentials

54.5.1 Legendre transforms and the four potentials

[Reserved: internal energy, enthalpy, the Helmholtz free energy introduced as the "free energy" of chemical processes [217], and the Gibbs free energy, all as Legendre transforms of one fundamental relation, in the systematic form Gibbs gave the subject [193]; the natural variables of each potential, and the parallel with the Legendre transform of Chapter 22.]

54.5.2 Maxwell relations

[Reserved: the four Maxwell relations as equalities of mixed second derivatives of the potentials, in the form given in Maxwell's Theory of Heat [179]; their use to reduce any measured response to derivatives of state variables; worked identities for the ideal and the van der Waals gas.]

54.5.3 Extremum principles, stability and the chemical potential

[Reserved: maximum entropy and minimum energy; minimisation of the free energies at fixed intensive variables; concavity and convexity of the potentials, with positive heat capacities and compressibilities as stability conditions; the chemical potential, the Euler relation and the Gibbs–Duhem identity, all from [193]; exposition following [1749].]

54.6 Phase equilibria

54.6.1 Coexistence and the Clausius–Clapeyron equation

[Reserved: equality of temperature, pressure and chemical potential across a phase boundary; latent heat; the Clausius–Clapeyron equation for the slope of the coexistence curve, from Clapeyron's memoir [95] as put on a first-law footing by Clausius [127]; applications to vapour pressure and to the melting line of ice.]

54.6.2 The Gibbs phase rule

[Reserved: the phase rule $f = c - p + 2$ counting the freedoms of a multicomponent, multiphase system [193]; triple points, with the former role of the water triple point in the definition of the kelvin [2333] noted; azeotropes and eutectics named as applications.]

54.6.3 Real gases: van der Waals and the critical point

[Reserved: Andrews's isotherms of carbon dioxide and the discovery of the critical point, with the continuity of the liquid and gaseous states [174]; the van der Waals equation of state from his 1873 dissertation [187]; the Maxwell equal-area construction [191]; the law of corresponding states; where the mean-field picture fails near the critical point, deferred to Chapter 123.]

Phenomenon 54.6 (The critical point)**.** Above a definite temperature T_c, characteristic of the substance, no pressure whatever liquefies a gas. Following the coexistence curve upwards, the difference in density between the liquid and its vapour shrinks and the meniscus between them fades and disappears at a single point (T_c, P_c, V_c); beyond it the two states pass continuously into one another. Andrews established this for carbon dioxide by mapping its isotherms through the point [174]. Plotted in units of their own critical constants, the isotherms of chemically unrelated substances then very nearly coincide—the law of corresponding states [187].

Derivation. Van der Waals's equation of state for one mole, with $v = V/n$,

$$P = \frac{RT}{v-b} - \frac{a}{v^2}\,, \tag{54.9}$$

corrects the ideal gas Equation (54.2) for the volume b the molecules exclude and for the attraction between them, which lowers the pressure by a/v^2. At high temperature P is monotone in v; below some temperature it is not, and a horizontal tie line then cuts the isotherm in three. The two behaviours are separated by the isotherm on which the stationary points merge, so the critical point is fixed by

$$\left(\frac{\partial P}{\partial v}\right)_T = 0\,, \qquad \left(\frac{\partial^2 P}{\partial v^2}\right)_T = 0\,. \tag{54.10}$$

Differentiating Equation (54.9) these read $RT/(v-b)^2 = 2a/v^3$ and $RT/(v-b)^3 = 3a/v^4$; dividing the second by the first gives $1/(v-b) = 3/(2v)$, hence

$$v_c = 3b\,, \qquad T_c = \frac{8a}{27Rb}\,, \qquad P_c = \frac{a}{27b^2}\,, \tag{54.11}$$

the last two by substituting v_c back. A critical point therefore exists for every substance obeying Equation (54.9), which is the observed fact. Measuring P, v and T in units of Equation (54.11) eliminates a and b from Equation (54.9) altogether,

$$\left(P_r + \frac{3}{v_r^2}\right)(3v_r - 1) = 8T_r\,, \tag{54.12}$$

so all such substances have one reduced equation of state: the law of corresponding states, together with the substance-independent prediction $P_c v_c/(RT_c) = 3/8$. Both are mean-field statements, and both are only approximate; how they fail in the immediate neighbourhood of the critical point is the subject of Chapter 123. □

54.6.4 The Joule–Thomson effect

[Reserved: throttling through a porous plug at constant enthalpy; the Joule–Thomson coefficient, its inversion temperature, and its vanishing for the ideal gas, from the Joule–Thomson experiments of 1852–1862 [137], whose apparatus and data are treated in Chapter 56; gas liquefaction as the application.]

54.7 The third law

[Reserved: Nernst's heat theorem, that entropy differences vanish as the temperature goes to zero [328]; the unattainability of absolute zero in finitely many steps; consequences—vanishing heat capacities and expansivities at low temperature—whose observed forms are quantum-statistical and are derived in Chapter 120.]

Phenomenon 54.7 (Nernst's heat theorem)**.** As the temperature is lowered towards absolute zero, the entropy change accompanying any isothermal change of the remaining variables tends to zero:

$$\lim_{T\to 0}\left[S(T,x_2)-S(T,x_1)\right]=0 \tag{54.13}$$

for any two values x_1, x_2 of those variables. Nernst extracted this from the low-temperature behaviour of the heats and free energies of chemical reactions, where the two quantities were found to converge on one another [328]. Two consequences are themselves observed: the heat capacities of condensed matter fall away as $T \to 0$, in flat contradiction of the classical constant of Chapter 55, and absolute zero is never reached, each successive stage of any cooling technique yielding a smaller temperature drop than the last.

Derivation. Both consequences follow from Equation (54.13) and the positivity of heat capacities. For the first, integrate $\mathrm{d}S = C_x \mathrm{d}T/T$ at fixed x from zero:

$$S(T,x)-S(0,x)=\int_0^T C_x(T')\,\frac{\mathrm{d}T'}{T'}\,. \tag{54.14}$$

The left side is finite, so the integral converges at its lower limit; a heat capacity tending to any nonzero constant would make it diverge logarithmically. Hence $C_x(T) \to 0$ as $T \to 0$, and the Dulong–Petit value cannot survive to low temperature.

For the second, suppose absolute zero were reached in one adiabatic step, carried out reversibly and therefore at constant entropy, from a state (T, x_1) with $T > 0$ to a state $(0, x_2)$. That requires $S(T, x_1) = S(0, x_2)$. But Equation (54.13) makes $S(0, x_1) = S(0, x_2)$, while Equation (54.14) with $C_x > 0$ makes $S(\cdot, x_1)$ strictly increasing, so $S(T, x_1) > S(0, x_1) = S(0, x_2)$ for every $T > 0$—a contradiction. No such step exists, and since the argument applies to each step separately, no finite sequence of them reaches zero either. Only the microscopic ground of Equation (54.13) itself lies outside thermodynamics; it is derived from the non-degeneracy of the ground state in Chapter 120. □

54.8 The bridge to the molecular theory

[Reserved: what the phenomenology cannot fix—the values of heat capacities, the entropy constant, the size of fluctuations—and where each is settled: the kinetic theory of Chapter 55, the ensembles of Chapter 118, and the fluctuation phenomena made visible by Brownian motion in Chapter 119; Boltzmann's statistical reading of entropy [196] anticipated here.]

Chapter 55

Kinetic Theory of Gases

Kinetic theory is the bridge from mechanics to heat. Its hypothesis is that a gas is an enormous number of molecules in rapid, disordered motion, and its programme is to compute the state variables of Chapter 54 from the mechanics of Chapter 19: pressure from momentum transfer at the walls, temperature from the mean kinetic energy, and the transport coefficients from the mean free path between collisions. The programme succeeded so completely that it settled a question older than physics—whether matter is molecular—and put a number on it, Avogadro's constant.

This chapter carries the hypothesis from Bernoulli's first kinetic model to the Boltzmann equation and its H-theorem, together with the Maxwell–Boltzmann distribution and the elementary mean-free-path theory of viscosity, heat conduction and diffusion. The systematic solution theory of the Boltzmann equation—Chapman–Enskog and its successors—belongs to Chapter 124, and the ensemble foundation that frees the results from molecular models to Chapter 118.

Derivation pending #2

Kinetic Theory of Gases: all derivations of this chapter are pending.

55.1 The kinetic hypothesis

55.1.1 From Bernoulli to Clausius

[Reserved: Bernoulli's derivation of Boyle's law from particle impacts in section ten of the Hydrodynamica [22]; the long eclipse of the idea under the caloric theory; Waterston's complete kinetic memoir, refused in 1845 and printed only in 1892 with Rayleigh's introduction [256]; Krönig's short revival [145]; Clausius's "kind of motion we call heat", with its translational, rotational and vibrational shares [147].]

55.1.2 Pressure from molecular impacts

[Reserved: the elementary pressure formula $P = \tfrac{1}{3} n m \langle v^2 \rangle$ of [22, 147]; temperature as mean translational kinetic energy, and the kinetic reading of the ideal-gas law of Chapter 54; the virial theorem of Clausius [177] as the systematic version, and its later use for real gases.]

55.1.3 The mean free path

[Reserved: Buys Ballot's objection that gases interdiffuse far too slowly for molecules moving at hundreds of metres per second, and Clausius's answer—the mean free path, $\ell \sim 1/(n\sigma)$—in the 1858 paper [149]; the concept is customarily dated to the 1857 paper [147], which does

not contain it; collision frequency and effective cross-section; orders of magnitude for air at standard conditions, with ℓ of order 10^{-7} m.]

55.2 The Maxwell velocity distribution

55.2.1 Maxwell's derivation

[Reserved: the 1860 derivation of the equilibrium velocity distribution from isotropy and the independence of the components [154]; mean, most probable and root-mean-square speeds; the stronger 1867 rederivation from detailed balance of collisions [169]; the distribution as the stationary point of the collision term, anticipating Section 55.4.2.]

55.2.2 Molecular-beam verification

[Reserved: Stern's direct measurement of thermal molecular speeds in a silver beam [471]; the velocity-selector measurements of Miller and Kusch on potassium and thallium beams, confirming the Maxwell distribution at the per-cent level [1076]; why a beam samples the flux-weighted distribution, and the correction that implies.]

Phenomenon 55.1 (The distribution of molecular speeds)**.** Molecules leaving a small aperture in an oven held at temperature T arrive at a distant detector spread over a range of speeds, and the spread is the one Maxwell computed and no other. Stern measured it directly in a beam of silver atoms, by the deflection of the beam on a rotating drum [471]; Miller and Kusch, passing beams of potassium and thallium through a slotted rotating cylinder that transmits one speed at a time, confirmed the predicted shape to the per-cent level [1076]. Inside the oven the number of molecules with speeds between v and $v + \mathrm{d}v$ is proportional to

$$F(v)\,\mathrm{d}v \propto v^2 \exp\left(-\frac{mv^2}{2k_{\mathrm{B}}T}\right)\mathrm{d}v\,, \tag{55.1}$$

while the escaping beam carries one further power of v, because fast molecules reach the aperture more often than slow ones.

Derivation. Let $f(\boldsymbol{v})\,\mathrm{d}^3v$ be the number density of molecules with velocity within d^3v of $\boldsymbol{v}$. Two properties of a gas at rest fix its form. Isotropy: no direction is distinguished, so f depends on $\boldsymbol{v}$ only through v^2. Independence: in a dilute gas the Cartesian components are uncorrelated, so $f = \varphi(v_x)\varphi(v_y)\varphi(v_z)$ with the same φ for each. Together,

$$\varphi(v_x)\varphi(v_y)\varphi(v_z) = F\left(v_x^2 + v_y^2 + v_z^2\right)\,. \tag{55.2}$$

Take logarithms in Equation (55.2) and differentiate with respect to v_x at fixed v_y, v_z: $\varphi'(v_x)/\varphi(v_x) = 2v_xF'/F$. The right-hand side is symmetric under permuting the three components, so $\varphi'(v_x)/\left(v_x\varphi(v_x)\right)$ is one and the same constant -2λ for each of them, whence $\varphi(v_x) = A\exp(-\lambda v_x^2)$ and $f \propto \exp(-\lambda v^2)$. Normalisability requires $\lambda > 0$.

The pressure fixes λ. A wall normal to $\hat{x}$ is reached in unit time, per unit area, by those molecules within a distance v_x of it that are moving towards it; each rebounds elastically with its normal momentum reversed, delivering $2mv_x$, so

$$P = \int_{v_x>0} 2mv_x\, v_x f\,\mathrm{d}^3v = nm\left\langle v_x^2\right\rangle = \tfrac{1}{3}nm\left\langle v^2\right\rangle\,, \tag{55.3}$$

the last step by isotropy. Comparison with the observed equation of state $P = nk_{\mathrm{B}}T$ of Chapter 54 gives $\langle v^2\rangle = 3k_{\mathrm{B}}T/m$, while the Gaussian above has $\langle v^2\rangle = 3/(2\lambda)$; hence $\lambda = m/(2k_{\mathrm{B}}T)$. Multiplying f by the area $4\pi v^2$ of the shell of constant speed converts it

into the distribution of speeds, which is Equation (55.1). Finally, the molecules crossing the aperture in unit time are those within a distance v_z of it, so their number carries an extra factor v_z; averaging v_z over the forward hemisphere at fixed speed gives $v/2$, and the beam distribution is proportional to $v^3 \exp(-mv^2/2k_{\mathrm{B}}T)$, as observed. □

55.3 Equipartition and specific heats

55.3.1 The equipartition theorem

[Reserved: the equality of mean translational energies between components and between species in Maxwell's 1860 paper [154]; Boltzmann's general equipartition theorem, with mean energy $\frac{1}{2}k_{\mathrm{B}}T$ per quadratic degree of freedom [171]; what counts as a degree of freedom for monatomic and diatomic gases.]

55.3.2 Specific heats and the classical failure

[Reserved: predicted heat-capacity ratios against measurement; the Dulong–Petit law of solid specific heats [65] as an equipartition statement; the failure for diatomic gases at low temperature and for solids—degrees of freedom frozen out, which classical mechanics cannot freeze—flagged by Maxwell himself [154]; the resolution is quantum and is derived in Chapter 68 and Chapter 125.]

Phenomenon 55.2 (Classical heat capacities, and their failure)**.** At ordinary temperatures the molar heat capacity of a simple solid is close to $3R$ and very nearly the same for every element—the law of Dulong and Petit [65]—while the ratio $\gamma = C_P/C_V$ that Laplace inferred for air from the speed of sound is close to $7/5$ [59]. These are exactly the values equipartition predicts, and their agreement was the first quantitative success of the molecular hypothesis. Both fail on cooling: the heat capacities of solids and the rotational and vibrational shares of molecular gases fall away as the temperature drops, and no mechanical system with a fixed number of quadratic degrees of freedom can behave so. Maxwell named the discrepancy the outstanding difficulty of the theory [154]; the low-temperature measurements are collected in Chapter 125 and the resolution is quantum (Chapter 68, Chapter 120).

Derivation. By the equipartition theorem [171], every coordinate or momentum entering the energy quadratically carries a mean energy $\frac{1}{2}k_{\mathrm{B}}T$ in equilibrium. A monatomic gas has three such terms per molecule, the translational momenta, so $U = \frac{3}{2}nRT$ and $C_V = \frac{3}{2}nR$. For any gas in the dilute limit of Chapter 54 the two heat capacities differ by $C_P - C_V = nR$ (Chapter 56), so

$$\gamma = 1 + \frac{2}{f}\,, \tag{55.4}$$

with f the number of quadratic terms: $\gamma = 5/3$ for $f = 3$. A diatomic molecule at ordinary temperature turns as well as translates, adding two rotational terms about the two axes perpendicular to the bond—the third has negligible moment of inertia—so $f = 5$ and $\gamma = 7/5$, the measured value for air. In a solid each atom oscillates about a fixed site in three directions, contributing three kinetic and three potential quadratic terms, so $f = 6$, $U = 3nRT$ and the molar heat capacity is $3R$: the Dulong–Petit law.

The failure has no derivation here, and that is the point. f in Equation (55.4) counts degrees of freedom that exist or do not; classical mechanics offers no mechanism by which a rotation or a vibration present at one temperature can be absent at a lower one, whatever the interaction assumed. The observation is therefore not a defect of the estimate but a refutation of the mechanics used to make it. □

55.4 The Boltzmann equation

55.4.1 The transport equation

[Reserved: the one-particle distribution function on phase space; the streaming term and the binary-collision integral of Boltzmann's 1872 memoir [183]; the molecular-chaos assumption (Stosszahlansatz), stated explicitly as the point where probability enters the mechanics; the collision invariants and the conservation laws they imply.]

55.4.2 The H-theorem

[Reserved: the functional $H = \int f \ln f \, \mathrm{d}^3 v$ and Boltzmann's theorem that collisions can only decrease it [183]; stationarity exactly on the Maxwell distribution; $-H$ as a kinetic entropy, and the second law of Chapter 54 recovered for dilute gases.]

Phenomenon 55.3 (Irreversible approach to equilibrium)**.** An isolated dilute gas prepared with any velocity distribution whatever relaxes, within a few mean times between collisions, to the Maxwell distribution Equation (55.1) at the temperature its energy fixes, and remains there: the equilibrium form is what every measurement on an undisturbed gas finds [1076]. The reverse has never been seen. No gas in equilibrium has ever been observed to sort itself spontaneously into a non-Maxwellian state, or to unmix after mixing—and this although the underlying molecular mechanics is exactly invariant under reversal of every velocity, so that the reverse of any observed history is itself a mechanically possible history. Boltzmann's transport equation reproduces the observed one-sidedness through the H-theorem [183].

Derivation pending #3

The H-theorem: that the binary-collision integral of the Boltzmann transport equation makes the functional $H = \int f \ln f \, \mathrm{d}^3 v$ non-increasing in time, with stationarity exactly on the Maxwell distribution; together with an explicit statement of the molecular-chaos assumption on which the monotonicity rests, and of what the reversibility and recurrence objections do and do not refute.

55.4.3 Reversibility and recurrence

[Reserved: Loschmidt's reversibility objection [195]; Poincaré's recurrence theorem [247] and Zermelo's recurrence objection [273]; Boltzmann's probabilistic reply—the H-theorem holds for the overwhelming majority of microstates, not for all [196]; how the arrow of time enters through the Stosszahlansatz, with the full discussion deferred to Chapter 118.]

55.4.4 The Maxwell–Boltzmann distribution

[Reserved: Boltzmann's 1877 combinatorial definition of the probability of a macrostate and the distribution over energies it yields [196]; the Boltzmann factor and the barometric formula; the 1877 counting as the origin of the statistical entropy developed in Chapter 118.]

55.5 Transport phenomena

55.5.1 Viscosity

[Reserved: the mean-free-path estimate of the shear viscosity of a gas, and Maxwell's startling prediction that it is independent of density [154]; his own Bakerian-lecture measurements confirming it [167]; the refined transport theory of [169]; hand-off of the continuum description to Chapter 31.]

Phenomenon 55.4 (The viscosity of a gas does not depend on its density)**.** Rarefy a gas and its viscosity does not fall. Over a wide range of pressures the shear viscosity of a gas is independent of its density, and it grows with temperature—the opposite of a liquid, which thins on heating. Maxwell obtained the prediction from the mean free path before anyone had looked, called it the most startling consequence of the theory, and then measured it himself in oscillating-disc experiments [154, 167]. The independence holds only while the mean free path stays small against the apparatus; when the gas is rarefied past that point the viscosity does begin to fall with density, which is the free-molecular regime of [355] and Section 55.6.2.

Derivation. A molecule of collision cross-section σ sweeps, in time t, the volume $\sigma \langle v_{\mathrm{rel}} \rangle t$, and meets in it every molecule whose centre lies inside; with the Maxwell distribution Equation (55.1) the mean relative speed of two molecules is $\sqrt{2}\,\langle v \rangle$, so the number of collisions is $\sqrt{2} n\sigma \langle v \rangle t$ and the mean free path is the distance travelled per collision,

$$\ell = \frac{\langle v \rangle t}{\sqrt{2} n\sigma \langle v \rangle t} = \frac{1}{\sqrt{2}\, n\sigma}\,. \tag{55.5}$$

Now let the gas be sheared, with a flow velocity $u(z)$ along x. Molecules crossing the plane z carry with them the mean x momentum of their last collision, one mean free path away, so the momentum transported per unit area and time is of order $\frac{1}{3} n \langle v \rangle \cdot m\,[u(z-\ell) - u(z+\ell)]$, which identifies the viscosity as

$$\eta \approx \tfrac{1}{3} nm \langle v \rangle \ell = \frac{m \langle v \rangle}{3\sqrt{2}\,\sigma}\,. \tag{55.6}$$

The number density has cancelled between the density of carriers and the distance Equation (55.5) each one carries its momentum: halve the density and each molecule travels twice as far before delivering its load. Since $\langle v \rangle = \sqrt{8 k_{\mathrm{B}} T/\pi m}$ from Equation (55.1), the estimate also gives $\eta \propto \sqrt{mT}/\sigma$, increasing with temperature. Only the numerical factor in Equation (55.6) is unreliable—it depends on how crudely the transport step is modelled—and fixing it is what the systematic expansion of Section 55.5.4 does. The cancellation itself does not depend on that factor, and it fails only when ℓ ceases to be set by Equation (55.5), that is when it grows to the size of the vessel. □

55.5.2 Heat conduction

[Reserved: Fourier's law of conduction from his analytical theory of heat [72]; the mean-free-path estimate of the thermal conductivity of a gas and its relation to the viscosity and the specific heat; density independence and the experimental range over which it holds.]

55.5.3 Diffusion

[Reserved: Graham's law of gaseous diffusion [94]; Fick's phenomenological law [143]; the mean-free-path estimate of the diffusion coefficient; Einstein's fluctuation relation connecting diffusion to molecular agitation [318], with Brown's original observations [82] and the decisive experiments of Chapter 119.]

55.5.4 Chapman–Enskog and the road to hydrodynamics

[Reserved: the systematic expansion of the Boltzmann equation about local equilibrium by Chapman [434] and Enskog [447], which turns the estimates of this section into controlled transport coefficients and derives the Navier–Stokes equations of Chapter 31; developed in Chapter 124 and only named here.]

55.6 Effusion and rarefied gases

55.6.1 Graham's law of effusion

[Reserved: effusion through an orifice small against the mean free path; the rate proportional to the inverse square root of the molar mass, established by Graham for diffusion [94] and for effusion and transpiration [120]; the kinetic effusion flux $\tfrac{1}{4}n\langle v\rangle$; isotope separation named as the application.]

Phenomenon 55.5 (Graham's square-root law)**.** A gas escaping through an aperture small compared with the mean free path does so at a rate inversely proportional to the square root of its molar mass, the pressure and temperature being fixed; a mixture driven through such an aperture is thereby partially separated, the lighter component passing preferentially. Graham established the square-root law first for the passage of gases through a porous plug [94] and then for effusion and for capillary transpiration [120]. That the law involves the mass and not the chemical character of the molecule is the point: it is a statement about mechanics, not chemistry.

Derivation. The aperture being small against ℓ, the escaping molecules are simply those that would have struck that patch of wall, and their distribution inside is undisturbed. The number crossing unit area of the plane in unit time is

$$\Phi = \int_{v_z>0} v_z f \,\mathrm{d}^3v = \int_0^\infty v\,F(v)\,\mathrm{d}v \int_0^{\pi/2} \frac{\cos\vartheta\,\sin\vartheta}{2}\,\mathrm{d}\vartheta = \tfrac{1}{4}n\,\langle v\rangle\ , \tag{55.7}$$

where F is the speed distribution normalised to $\int F\,\mathrm{d}v = n$, and the angular factor is the fraction $\tfrac{1}{2}\sin\vartheta\,\mathrm{d}\vartheta$ of solid angle at polar angle ϑ times the obliquity $\cos\vartheta$. Elementary Gaussian integrals over Equation (55.1) give $\langle v\rangle = \sqrt{8k_{\mathrm{B}}T/\pi m}$, so

$$\Phi = n\sqrt{\frac{k_{\mathrm{B}}T}{2\pi m}} = P\sqrt{\frac{1}{2\pi m k_{\mathrm{B}}T}}\,, \tag{55.8}$$

using $P = nk_{\mathrm{B}}T$. At the same pressure and temperature the number density is the same for both gases, so the ratio of the two effusion rates is $\sqrt{m_2/m_1} = \sqrt{M_2/M_1}$, which is Graham's law. The same factor $\tfrac{1}{4}n\,\langle v\rangle$ governs the rate at which a surface is struck, and so the growth of an adsorbed layer in a vacuum system. □

55.6.2 The Knudsen regime

[Reserved: the Knudsen number; Knudsen's laws of molecular flow through tubes at low pressure [355]; thermal transpiration; where the continuum description of Chapter 31 fails and free-molecular flow replaces it, with vacuum technology as the working application.]

55.7 Molecular reality and Avogadro's constant

55.7.1 Loschmidt's count

[Reserved: Avogadro's hypothesis [55] made quantitative: Loschmidt's 1865 estimate of the molecular diameter and the number density of air, from the mean free path and the density of the liquefied gas [163]; the modern Loschmidt constant, about $2.687\times10^{25}/\mathrm{m}^3$ at standard conditions [2431].]

Phenomenon 55.6 (One molecular count, reached by unrelated routes)**.** The number of molecules in a mole of substance can be inferred by methods sharing no apparatus and no theory beyond the molecular hypothesis itself—from the mean free path together with the

density of the liquefied gas [163], from the Brownian displacements of suspended particles [318, 358], from the charge carried in electrolysis, and from the counting of particles emitted by a radioactive source—and the answers agree. That convergence, and not any single determination, is what established that matter is molecular, since no one route is more than an estimate. The constant is now fixed by definition, $N_{\mathrm{A}} = 6.022\,140\,76 \times 10^{23}/\mathrm{mol}$ [2333, 2431], and it is the measurements that are graded against it rather than the other way about.

Derivation. Loschmidt's route needs two measured numbers and one geometrical assumption. The first number is the mean free path ℓ. Read Equation (55.6) in the form $\eta \approx \frac{1}{3}\rho\,\langle v\rangle\,\ell$, with $\rho = nm$ the mass density of the gas, which is weighed directly, and $\langle v\rangle = \sqrt{8RT/\pi M}$ fixed by the molar mass alone; then $\ell = 3\eta/(\rho\,\langle v\rangle)$ follows from measurements that never mention the number of molecules. The second number is the condensation ratio ε, the volume of liquid obtained from unit volume of gas. Take the molecules to be spheres of diameter d, packed in the liquid so as to fill it; then per unit volume of gas

$$\varepsilon = n\,\frac{\pi d^3}{6}\,, \tag{55.9}$$

while Equation (55.5) with $\sigma = \pi d^2$ gives $\ell = 1/(\sqrt{2}\pi d^2 n)$. These are two equations in the two unknowns n and d; substituting the second into Equation (55.9) leaves

$$\varepsilon = \frac{\pi d^3}{6\sqrt{2}\pi d^2 \ell} = \frac{d}{6\sqrt{2}\,\ell}\,, \qquad d = 6\sqrt{2}\,\varepsilon\ell\,, \tag{55.10}$$

and then $n = 1/(\sqrt{2}\pi d^2\ell)$ and $N_{\mathrm{A}} = nV_{\mathrm{m}}$ with V_{m} the molar volume. The packing assumption is crude, so the result is an order of magnitude and not a measurement; what makes it evidence is that the wholly unrelated Brownian and electrolytic routes land on the same order of magnitude. □

55.7.2 From estimates to an exact constant

[Reserved: the convergence of independent determinations of Avogadro's constant—kinetic, Brownian (Perrin's programme [358], with the experiments of Chapter 119), electrolytic and radiometric—as the historical proof of molecular reality; the 2019 redefinition of the SI, which fixes the constant exactly at $6.022\,140\,76 \times 10^{23}/\mathrm{mol}$ *[2333, 2431].]*

Chapter 56

Experiment: The Mechanical Equivalent of Heat

Joule's paddle-wheel measurements established that heat is energy in transfer: a fixed quantity of mechanical work, however delivered, produces the same rise of temperature in the same body [129]. The result killed the caloric theory, grounded the first law of Chapter 54, and fixed the conversion factor that every calorimetric measurement in this treatise silently uses. Its molecular reading—work stirred into the disordered motion of Chapter 55—came a decade later.

When its content is written, this chapter will use the structured `experiment` environment, with apparatus, procedure, observations with uncertainties in SI units, interpretation and primary references; the sections below reserve that structure and name the sources.

Derivation pending #4

Experiment: The Mechanical Equivalent of Heat: all derivations of this chapter are pending.

56.1 Historical context: heat as substance or motion

56.1.1 Rumford's cannon-boring observations

[Reserved: the caloric theory and its genuine successes; Rumford's observation that boring cannon generates heat without limit, arguing that heat cannot be a conserved substance [43]; his rough estimate of an equivalent from horses and boring dust; why the argument, though correct, did not persuade for half a century.]

Phenomenon 56.1 (Friction generates heat without limit)**.** Rubbing generates heat in a body without limit and without altering the body. Rumford, boring cannon at the Munich arsenal with a deliberately blunted tool driven by horses, found the supply of heat inexhaustible so long as the work continued—enough, in one trial, to bring the surrounding water to the boil—and found further that the metal dust produced had the same heat capacity, weight for weight, as the metal it came from [43].

Derivation. The observation refutes the hypothesis it was aimed at. Suppose heat is a conserved substance resident in bodies, as the caloric theory held. Then the heat obtainable from a given body is bounded by the amount it contains, and heat is released only by lowering that content—by a change of state, or by a reduction in the capacity of the substance to hold caloric. Both consequences are contradicted at once: the yield was unbounded and, so far as could be told, proportional to the work expended rather than to any property of

the metal consumed; and the dust, which on the caloric account had given up its heat by a reduced capacity, showed the same capacity as the solid. Hence what is released is not drawn from a fixed store, and the conserved quantity—if there is one—must be one that mechanical work can supply. Identifying it is the work of Section 56.2; the caloric theory survives only as the special case in which no work is done. □

56.1.2 Mayer's inference

[Reserved: Mayer's 1842 claim of the equivalence of heat and work, with the first numerical equivalent, extracted from the difference of the two heat capacities of air [108]; the assumption the inference rests on (no internal work in the free expansion of a gas); the priority dispute with Joule, stated plainly.]

Phenomenon 56.2 (The two heat capacities of a gas differ by the gas constant)**.** Heating a gas at constant pressure takes more heat than heating the same gas through the same interval at constant volume, and for dilute gases the molar difference is the gas constant,

$$C_P - C_V = nR\,. \tag{56.1}$$

The two capacities of air had been measured before any theory of heat existed, and Mayer saw that their difference is the work the expanding gas performs against the atmosphere; equating the one to the other gave him the first numerical value of the mechanical equivalent of heat, without any apparatus of his own [108].

Derivation. Take n moles. At constant volume the gas does no work, so all the heat supplied raises the internal energy: $C_V = \mathrm{d}U/\mathrm{d}T$, using that U depends on the temperature alone—which is the content of Phenomenon 56.4 below, and the assumption Mayer's inference silently rests on. At constant pressure the same rise of temperature raises U by the same amount, since U depends on nothing else, but the gas additionally expands, by $\Delta V = nR\Delta T/P$ from the equation of state $PV = nRT$, and does work

$$P\Delta V = nR\Delta T\,. \tag{56.2}$$

The heat supplied at constant pressure therefore exceeds that supplied at constant volume by exactly $nR\Delta T$, which is Equation (56.1). The inference now runs backwards. The left side of Equation (56.1) is a difference of two *calorimetric* measurements, expressed in whatever unit of heat the laboratory uses; the right side is a *mechanical* work, expressed in mechanical units. The equality between them is therefore a conversion factor between the two units, obtained without ever converting work into heat in an apparatus. Its weakness is equally plain: it is only as good as the assumption that a gas expanding freely does no internal work, which nothing in the argument tests. Joule's paddle wheel tests it. □

56.1.3 Joule's programme, 1843–1849

[Reserved: the magneto-electrical measurements that opened the series [109]; the rarefaction and condensation experiments [112]; the first paddle-wheel announcement [113]; the convergence of Joule's values as the methods improved, closing on the definitive 1850 series [129].]

56.2 The claim under test

[Reserved: the proposition that one fixed conversion factor connects work and heat, independent of the process, the substance and the rate—the quantitative content of the first law of Chapter 54, stated for all of physics by Helmholtz [121]; what a failure of process independence would have looked like; the kinetic reading of the same claim in Chapter 55.]

Phenomenon 56.3 (One conversion factor, whatever the mechanism)**.** In a thermally insulated calorimeter the rise of temperature is proportional to the mechanical work dissipated inside it, and the constant of proportionality is a property of the calorimeter alone. It does not depend on how the work is dissipated, nor on how fast. Joule obtained the same equivalent from a paddle wheel churning water, from the same wheel churning mercury, from cast-iron discs rubbing against each other under mercury [129], and, in the earlier series, from an electric current driven by a falling weight [109] and from the condensation and rarefaction of air [112]; the paddle-wheel method itself he first announced in 1845 [113]. Nothing in the first law compels this agreement: it is what the measurement found.

Derivation. The work delivered in a run of N descents of two weights of mass M through a height h is

$$W = 2NM\left(gh - \tfrac{1}{2}u^2\right) - W_{\mathrm{f}}\,, \tag{56.3}$$

u being the residual speed of the weights at arrest and W_{f} the work absorbed by the pulleys and the suspension, both of which Joule measured separately and subtracted. No heat crosses the insulated boundary, so by the first law the internal energy of the calorimeter rises by exactly W. The calorimeter ends in a state differing from its initial one only in temperature, so that rise is $C\,\Delta T$ with C the total heat capacity of water, vessel and fittings, and

$$\Delta T = \frac{W}{C}\,. \tag{56.4}$$

Everything on the right of Equation (56.4) is a property of the calorimeter and of the work supplied; the stirring mechanism has dropped out. This is the whole content of the observation, and it is the empirical statement that the adiabatic work between two states depends only on the states—the property that makes the internal energy a state function in Chapter 54. Had the equivalent depended on the mechanism, one could raise the calorimeter's temperature by the cheaper route and lower it again by the dearer, returning every part of the apparatus to its initial state while extracting net work; the measurement is thus also a bound on perpetual motion of the first kind. □

56.3 Apparatus

[Reserved: the 1850 paddle-wheel calorimeter [129]: a brass paddle wheel churning water—and, in parallel series, mercury—inside an insulated vessel fitted with stationary vanes that prevent bulk rotation; the drive by two lead weights of about 13 kg *each falling through about* 1.6 m*, the descent repeated twenty times per run; thermometers readable to about* 3 mK*; all dimensions quoted here in SI, converted on port from Joule's imperial originals.]*

56.4 Procedure

[Reserved: determination of the work input from the weight, the fall height and the descent count, with corrections for the residual kinetic energy of the weights at arrest and for pulley friction; measurement of the temperature rise of the calorimeter; blank trials interleaved to fix the radiation correction; Joule's averaging over runs and series [129].]

56.5 Observations

[Reserved: the data tables to be reproduced here in SI—mass of the weights, total fall, work input, water mass and the water equivalent of the vessel, corrected temperature rise, and the resulting equivalent per run—with the uncertainty budget presented per [2129]; Joule's 1850

mean corresponds to a specific heat of water of about $4157\,\mathrm{J/(kg\,K)}$*, within one per cent of the modern* $4187\,\mathrm{J/(kg\,K)}$ *[129].]*

56.6 Interpretation

[Reserved: heat is not a conserved substance; work and heat are two currencies of one conserved quantity, the internal energy of Chapter 54; process independence checked across the water, mercury and cast-iron friction series [129]; the place of the result in the energy-conservation synthesis of Mayer [108] and Helmholtz [121].]

56.7 The Joule–Thomson companion experiments

[Reserved: free expansion into vacuum, from Gay-Lussac [52] to Joule's twin-vessel experiment [112], showing no temperature change at the attainable precision; the porous-plug experiments of Joule and Thomson, 1852–1862, which resolved the small real-gas cooling the cruder method missed [137]; the Joule–Thomson coefficient and its inversion, whose thermodynamics is developed in Chapter 54; gas liquefaction as the technological descendant.]

Phenomenon 56.4 (No cooling in the free expansion of a dilute gas)**.** Let a gas at moderate pressure expand into an evacuated vessel, the whole assembly being thermally isolated from its surroundings. No change of temperature is found, within the precision of the measurement. Gay-Lussac observed this first [52], and Joule repeated it with twin copper vessels joined by a stopcock and immersed in a common water bath [112]. The conclusion drawn—that the internal energy of a dilute gas depends on its temperature alone—is the assumption on which Mayer's inference Equation (56.1) silently rested. The effect the method was too crude to see is the subject of Phenomenon 56.5.

Derivation. The gas expands into vacuum, so it does no work on anything outside it, $W = 0$; the assembly is insulated, so $Q = 0$. By the first law $\Delta U = 0$: the free expansion is a change of volume at constant internal energy. The measurement reports $\Delta T = 0$ across it, that is

$$\left(\frac{\partial T}{\partial V}\right)_U = 0\,. \tag{56.5}$$

Expanding the derivative at constant U by the chain rule,

$$\left(\frac{\partial T}{\partial V}\right)_U = -\frac{1}{C_V}\left(\frac{\partial U}{\partial V}\right)_T\,, \tag{56.6}$$

since $(\partial U/\partial T)_V = C_V$. The heat capacity is finite and positive, so Equation (56.5) forces $(\partial U/\partial V)_T = 0$ and hence $U = U(T)$. Together with the equation of state $PV = nRT$ this completes the description of the ideal gas and yields Equation (56.1). The honest statement of the result is a bound rather than an identity: the experiment shows only that $(\partial U/\partial V)_T$ is smaller than the thermometry could resolve, the water bath's own heat capacity swamping the effect. □

Phenomenon 56.5 (The Joule–Thomson effect)**.** Force a real gas steadily through a porous plug, from a higher to a lower pressure, along a thermally insulated tube. Its temperature changes by a small but definite amount: most gases at room temperature emerge cooler, hydrogen and helium emerge warmer, and for each gas there is an inversion temperature above which cooling gives way to warming. Joule and Thomson established the effect in the series of 1852–1862 [137]. It is the effect the free expansion of Phenomenon 56.4 was too crude to detect, and, unlike that null result, it is a direct measure of the forces between molecules; the industrial liquefaction of gases descends from it.

Derivation. Follow a fixed quantity of gas through the plug. Upstream, the piston pushing it in does work P_1V_1 on it; downstream, the gas does work P_2V_2 in making room for itself. The tube is insulated, so $Q = 0$, and the first law gives $U_2 - U_1 = P_1V_1 - P_2V_2$, that is

$$H_1 = H_2\,, \qquad H := U + PV\,. \tag{56.7}$$

Throttling is a process at constant enthalpy—not a quasi-static one, but the end states are equilibrium states and that is all Equation (56.7) requires. The measured quantity is therefore

$$\mu := \left(\frac{\partial T}{\partial P}\right)_H = -\frac{1}{C_P}\left(\frac{\partial H}{\partial P}\right)_T\,, \tag{56.8}$$

by the same chain rule as before with $(\partial H/\partial T)_P = C_P$. From $\mathrm{d}H = T\,\mathrm{d}S + V\,\mathrm{d}P$ and the Maxwell relation $(\partial S/\partial P)_T = -(\partial V/\partial T)_P$ of Chapter 54,

$$\mu = \frac{1}{C_P}\left[T\left(\frac{\partial V}{\partial T}\right)_P - V\right]\,. \tag{56.9}$$

For an ideal gas $V = nRT/P$ gives $T(\partial V/\partial T)_P = V$ and $\mu = 0$: the effect vanishes identically, so whatever is measured is a departure from ideality and nothing else. The sign of Equation (56.9) changes where $T(\partial V/\partial T)_P = V$, which is the inversion condition; for the van der Waals gas of Chapter 54 the attraction a cools and the excluded volume b warms, and the inversion temperature is where the two balance—which is why hydrogen and helium, with small a, must be precooled before throttling will liquefy them. □

56.8 Modern repetitions and precision

[Reserved: electrical calorimetry replacing mechanical stirring—the continuous-flow measurement of the specific heat of water by Callendar and Barnes [303], and the National Bureau of Standards determinations of Osborne, Stimson and Ginnings [846]; the retirement of the calorie; since the 2019 SI the joule is the sole unit of heat and work, so the "equivalent" is exact by definition [2333], its measurable content surviving as the specific heat of water, with current values in [2431].]

Chapter 57

Experiment: The Kinetic Theory Tested

Chapter 55 makes a gas out of molecules and then makes predictions with them: a definite distribution of molecular speeds at a given temperature, a viscosity that does not fall when the gas is rarefied, three transport coefficients built from one and the same mean free path, a count of molecules reachable from macroscopic measurements alone, and—by what it omits—an equation of state that can be exact only if the molecules exert no forces on one another at a distance. None of that is self-evident. One item, the density independence of the viscosity, Maxwell himself called the most startling consequence of the theory, and he did not believe it until he had measured it. This chapter reports the measurements.

The six experiments below are ordered by what they test rather than by date. Two are molecular-beam experiments that record the distribution of speeds itself, which is the central assertion of the theory and the one that a purely thermodynamic argument can never reach. One is Maxwell's own measurement of the prediction he found hardest to believe. One collects the transport coefficients and checks the relations kinetic theory forces among them. One turns those same measurements into a count of molecules, which is how the molecular hypothesis first acquired a number. The last, the porous plug of Joule and Thomson, measures what the ideal gas leaves out. The Brownian evidence for the same hypothesis—the most direct of all, since it watches the fluctuations rather than the averages—is treated separately in Chapter 119, and the thermodynamic reading of the porous-plug experiment in Chapter 56.

57.1 Stern: the direct measurement of molecular speeds (1920)

Kinetic theory asserts that the molecules of a gas at 300 K are moving at some hundreds of metres per second. Before 1920 that number had never been measured; it had only been inferred, from the pressure a gas exerts and from the speed of sound, by arguments that assume the molecular picture they are meant to support. Stern's molecular-beam apparatus measured a molecular speed directly, by timing a flight over a known distance [471].

57.1.1 Apparatus

A platinum wire, coated with silver and heated electrically in high vacuum, evaporates silver atoms. The wire lies on the axis of two coaxial cylinders fixed to it: an inner cylinder of radius r_1 carrying a narrow slit parallel to the axis, and an outer cylinder of radius r_2 whose inner surface collects the atoms that pass the slit. The whole assembly—wire, slit and collector—can be spun about the wire axis at an angular speed ω of a few hundred

radians per second. The vacuum must be good enough that the flight from slit to collector is collision-free, that is, the mean free path of Phenomenon 57.7 must exceed $r_2 - r_1$, and the deposit is built up over a long run and then developed, as in the molecular-beam experiment of Chapter 80.

57.1.2 Procedure

With the assembly at rest the atoms travel radially and the deposit forms on the collector directly opposite the slit; that position is the zero of the measurement. With the assembly spinning, the collector moves while the atom is in flight, so the deposit is displaced along the circumference by an amount that depends on the time of flight and hence on the speed. Reversing the sense of rotation displaces it the other way; taking half the separation between the two deposits removes any error in locating the zero, which is the only quantity in the measurement that cannot be read off a scale. The oven temperature, the two radii and the angular speed complete the data.

57.1.3 Observations and data

[Reserved: the primary tables of [471], to be reproduced here in SI—the two cylinder radii in metres, the angular speed in radians per second, the wire temperature in kelvin, the measured displacement of the deposit maximum in metres, and the mean molecular speed inferred from them with its uncertainty, presented per [2129]. For orientation, silver at about $1300\,\mathrm{K}$ *has* $\langle v\rangle = \sqrt{8RT/\pi M} \approx 5 \times 10^{2}\,\mathrm{m/s}$*, so with centimetre-scale radii and* ω *of a few hundred radians per second the displacement is of order a millimetre—small, but far larger than the width of the undisplaced trace.]*

57.1.4 Interpretation

The deposit moves, in the direction and by roughly the amount that a flight time $(r_2 - r_1)/v$ requires with v of the predicted size. It is also *smeared*, and the smear is not a defect of the apparatus: since the displacement varies as $1/v$, the deposit is a map of the distribution of speeds, stretched towards the zero for the fast molecules and away from it for the slow ones. This first version measured essentially one moment of that distribution; recording its shape is the business of Section 57.2.

One caution belongs with any beam measurement and is developed in Equation (57.3): what leaves an aperture is not a sample of the gas inside. Fast molecules reach the aperture more often than slow ones, so the beam is weighted by an extra power of the speed, and its mean speed exceeds the mean speed in the oven by the factor $3\pi/8$. Comparing a beam average against the in-oven $\langle v\rangle$ therefore misses by $18\,\%$ before any experimental error is committed.

57.1.5 Primary references

[471]. The method is the one Stern had developed from 1919 onward and applied two years later in Chapter 80; the diffraction of the same beams, which measures a wavelength rather than a speed, belongs to Chapter 71 [651].

Phenomenon 57.1 (Molecules move at the speeds kinetic theory assigns them)**.** The atoms of a vapour in equilibrium at temperature T cross a known distance in a measurable time, and the speed so obtained is of the order $\sqrt{k_{\mathrm{B}}T/m}$ that kinetic theory requires—some hundreds of metres per second for a heavy metal vapour at a thousand kelvin, and larger for lighter molecules in the same ratio $M^{-1/2}$. Stern measured it by the displacement of a silver deposit on a rotating collector [471]. The measurement is direct in the sense that

matters here: it uses a length and a time, and no property of a gas other than the fact that its molecules fly freely between collisions.

Derivation. Let the slit sit at radius r_1 and the collector at radius r_2 on the same rigid body, turning at angular speed ω. An atom that passes the slit travels radially outward in a straight line—this is the assumption the vacuum has to earn—and reaches the collector after a time $t = (r_2 - r_1)/v$. During that time the collecting surface has advanced along its own circumference by $\omega r_2 t$, so the deposit lands a distance

$$s = \frac{\omega r_2 (r_2 - r_1)}{v} \tag{57.1}$$

from the point opposite the slit. Everything in Equation (57.1) but v is a laboratory constant, so the equation is a measurement of v; and since $s \propto 1/v$, a spread of speeds becomes a spread of deposit positions, with the slow molecules thrown furthest. Reversing ω reverses the sign of s, so half the separation of the two deposits gives $|s|$ without knowing where the zero lies. □

57.2 Miller and Kusch: the shape of the distribution (1955)

Between Stern's measurement and this one stand the rotating-drum velocity analysers of the early 1930s—Zartman's and Ko's—which replaced the single displaced deposit by a deposit profile and so recorded the shape of the distribution rather than one moment of it [690, 742]. Miller and Kusch's slotted-rotor selector is the precision version of the same idea, and it is the experiment that turned the Maxwell distribution from a plausible functional form into a measured curve [1076].

57.2.1 Apparatus

An oven holding potassium or thallium at a controlled and measured temperature, effusing through a slit into high vacuum. Between the oven and the detector sits a velocity selector: a cylinder carrying many slots cut helically through its rim, spun about its axis, so that an atom passes only if its time of flight along the cylinder matches the time the entrance slot takes to rotate into the position of the exit slot. The transmitted band of speeds is therefore proportional to the rotation rate, and its width is set by the machining. Detection is by surface ionisation on a hot wire followed by ion collection, which counts atoms as a current rather than accumulating a deposit, and so turns the measurement from a photographic one into an electrical one.

57.2.2 Procedure

Hold the oven at a measured temperature, scan the rotor speed, and record the detector current at each setting. The result is the transmitted intensity as a function of the selected speed. The raw curve is not yet the distribution: it must be corrected for the finite width of the transmission window, for the higher-order pass bands at submultiples of the rotation rate, and for the response of the ioniser. The corrected curve is then compared with the flux-weighted Maxwell form Equation (57.2), with the oven temperature taken from the thermocouple and *not* adjusted to improve the fit—which is what makes the comparison a test rather than a fit.

57.2.3 Observations and data

[Reserved: the primary curves of [1076], to be tabulated here as transmitted intensity against selected speed in m/s *for potassium and for thallium at their stated oven temperatures*

in kelvin, together with the residuals against Equation (57.2) and their uncertainties per [2129]. The headline result quoted in Chapter 55 is agreement with the predicted shape at the per-cent level; the individual residuals are what show whether the disagreement is random or structured, and they are not reproduced here because the primary tables were not consulted.]

57.2.4 Interpretation

Two things are being tested at once, and they are worth separating. The first is the *shape*: an exponential in v^2 with a polynomial prefactor, over a range of speeds wide enough that the prefactor and the exponential are separately visible. The second is the *parameter*: the temperature that best describes the curve must be the temperature the thermocouple reads, since the distribution has no other adjustable constant. A theory that got the shape right with a free temperature would be much weaker evidence.

The composition of the vapour enters the interpretation and is not a detail: a dimer of twice the molecular mass has the same distribution with v rescaled by $1/\sqrt{2}$, so an unrecognised dimer fraction imitates a distorted distribution of the monomer. Any beam experiment of this kind therefore owes an account of what is actually in the beam, which is one reason for measuring two species of very different mass.

57.2.5 Primary references

[690, 742, 1076]. The rotating-drum analysers come first and the velocity selector of Miller and Kusch is the precision version of the same idea.

Phenomenon 57.2 (The distribution of speeds in a molecular beam)**.** The number of molecules per unit time leaving an oven at temperature T with speeds between v and $v + \mathrm{d}v$ is proportional to $v^3 \exp\left(-mv^2/2k_{\mathrm{B}}T\right) \mathrm{d}v$, and the temperature that describes the measured curve is the temperature of the oven. Miller and Kusch established this for potassium and thallium beams, passing them through a slotted rotating selector, and found the predicted shape confirmed at the per-cent level [1076]. The cube of the speed, rather than the square that describes the gas inside the oven, is not an adjustable feature: it follows from the aperture being small against the mean free path.

Derivation. Inside the oven the number density of molecules with speeds in $\mathrm{d}v$ about v is $F(v)\,\mathrm{d}v \propto v^2 \exp\left(-mv^2/2k_{\mathrm{B}}T\right) \mathrm{d}v$, the factor v^2 being the area of the shell of constant speed in velocity space. A molecule crosses the aperture in unit time only if it lies within a distance v_z of it and is moving towards it, so the flux carries one power of the speed more than the density does:

$$\Phi(v)\,\mathrm{d}v \propto v^3 \exp\left(-\frac{mv^2}{2k_{\mathrm{B}}T}\right) \mathrm{d}v\,. \tag{57.2}$$

The extra power is large. Write $a = m/2k_{\mathrm{B}}T$ and use $\int_0^\infty v^4 \exp\left(-av^2\right) \mathrm{d}v = \frac{3}{8}\sqrt{\pi}\, a^{-5/2}$ together with $\int_0^\infty v^3 \exp\left(-av^2\right) \mathrm{d}v = \frac{1}{2}a^{-2}$; then the mean speed of the molecules in the *beam* is

$$\langle v \rangle_{\mathrm{beam}} = \frac{3}{4}\sqrt{\frac{\pi}{a}} = \frac{3\pi}{8}\,\langle v \rangle \approx 1.18\,\langle v \rangle\ , \tag{57.3}$$

since $\langle v \rangle = \sqrt{8k_{\mathrm{B}}T/\pi m} = 2/\sqrt{\pi a}$ is the mean speed in the oven. The two differ by 18 %. The same argument shows why the aperture must be small against the mean free path: if it were not, the escaping molecules would collide on the way out, the gas would flow hydrodynamically through the hole, and the beam would carry a shifted rather than a weighted distribution. □

57.3 Maxwell: the viscosity of air and its density (1866)

In 1860 Maxwell derived from the mean free path a conclusion he plainly disliked: the shear viscosity of a gas should not depend on its density [154]. Halve the number of carriers and each survivor carries its momentum twice as far, and the two effects cancel exactly. No experiment at the time contradicted it, because no experiment had looked. Maxwell looked, in the Bakerian Lecture of 1866 [167].

57.3.1 Apparatus

Parallel glass discs mounted on a common vertical axis and hung from a torsion fibre, so that the assembly oscillates in its own plane between fixed discs a small distance away. The whole is enclosed in a vessel that can be evacuated to a controlled pressure and heated, the gas temperature being read by a thermometer inside and the pressure by a manometer. The angular amplitude is read optically, by a mirror on the suspension and a distant scale, so that nothing touches the moving system.

57.3.2 Procedure

Set the discs swinging and record the period and the decay of the amplitude. The gas in the narrow gaps resists the motion, and the logarithmic decrement of the oscillation—the constant ratio between successive amplitudes—measures that resistance; with the geometry of the gaps known, it yields the shear viscosity. Repeat at a series of pressures at fixed temperature, which is the density test, and at a series of temperatures at fixed pressure, which is the temperature test. The residual decrement of the suspension itself is measured with the vessel evacuated and subtracted.

57.3.3 Observations and data

[Reserved: the primary tables of [167], to be reproduced here in SI—gas pressure in Pa*, temperature in* K*, period in* s*, logarithmic decrement, and the viscosity in* Pa s *with its uncertainty, converted from Maxwell's units on port per the unit axiom of Chapter 2. Two derived statements are what the table has to support: that the viscosity is constant along the pressure series, and that along the temperature series it rises as T rather than as $\sqrt{T}$. For orientation, the viscosity of air near room temperature is about* 18 µPa s.*]*

57.3.4 Interpretation

The pressure series is the striking one. A liquid's viscosity is a property of how crowded it is; a gas's is not, and over the range of pressures the apparatus could span the measured viscosity did not change. That is the mean-free-path cancellation, observed. It is not unlimited: when the gas is rarefied until the mean free path is comparable with the width of the gaps, the molecules cross from one surface to the other without colliding, the internal friction ceases to be a property of the gas alone, and the viscosity does begin to fall with the density—the free-molecular regime of [355]. The failure is as informative as the success, because it locates the length that the elementary theory says is responsible.

The temperature series is the more consequential one, and it went against the theory as Maxwell had first written it. Hard spheres give $\eta \propto \sqrt{T}$; the measurements gave something much closer to $\eta \propto T$. Maxwell drew the correct conclusion—that molecules are not hard spheres but centres of a repulsive force—and rebuilt the transport theory in 1867 around the one force law, the inverse fifth power, that makes the collision integrals tractable [169]. This is the first quantitative information about an intermolecular force obtained from a bulk

measurement, and it is the same programme that the porous plug of Section 57.6 continues at long range.

57.3.5 Primary references

[167]; the prediction under test is from [154] and the theory rebuilt on the result is [169]. The low-pressure breakdown is [355].

Phenomenon 57.3 (Rarefying a gas does not thin it)**.** Over a wide range of densities the shear viscosity of a gas is independent of its density, so that pumping gas out of a vessel does not reduce the drag on a body moving inside it. Maxwell predicted this from the mean free path before any measurement existed [154] and then measured it himself with oscillating discs [167]. The independence holds only while the mean free path remains small against the apparatus; below that the viscosity does fall with density [355].

Derivation. Molecules crossing a plane in a sheared gas carry the mean flow momentum they acquired at their last collision, one mean free path ℓ away, and they cross at a rate of order $\frac{1}{3}n\langle v\rangle$ per unit area. The transported momentum per unit area and time is therefore of order $\frac{1}{3}nm\langle v\rangle\,\ell$ times the velocity gradient, so

$$\eta \approx \tfrac{1}{3}\rho\langle v\rangle\,\ell = \frac{m\langle v\rangle}{3\sqrt{2}\,\sigma}\,, \tag{57.4}$$

using $\rho = nm$ and $\ell = 1/\left(\sqrt{2}n\sigma\right)$ with σ the collision cross-section. The number density has cancelled: it counts the carriers, and it also fixes how far each carrier goes. Nothing in Equation (57.4) depends on how crudely the transport step was modelled except the numerical factor, so the cancellation is robust where the estimate of the coefficient is not. It fails exactly when ℓ stops being given by the collisions—that is, when the vessel, not the gas, sets the distance a molecule travels between momentum exchanges. □

Phenomenon 57.4 (The viscosity of a gas rises with temperature, and too fast for hard spheres)**.** Heating a gas makes it more viscous—the opposite of a liquid—and for air the rise is close to proportional to the absolute temperature. Maxwell's oscillating-disc measurements gave a temperature dependence much nearer $\eta \propto T$ than the $\eta \propto \sqrt{T}$ that follows from treating molecules as hard spheres [167], and it was this discrepancy, not any theoretical preference, that led him to replace the spheres by point centres repelling as the inverse fifth power of the distance [169].

Derivation. Let the molecules repel along the line of centres with a force K/r^{ν}. Two molecules approaching with relative kinetic energy of order $k_{\mathrm{B}}T$ are deflected through an angle of order unity when the potential energy at closest approach is comparable with that energy, $K/b^{\nu-1} \sim k_{\mathrm{B}}T$, where b is the impact parameter. The effective collision radius therefore shrinks as the gas is heated,

$$b \sim \left(\frac{K}{k_{\mathrm{B}}T}\right)^{1/(\nu-1)}\,, \qquad \sigma \sim \pi b^2 \propto T^{-2/(\nu-1)}\,, \tag{57.5}$$

and substituting into Equation (57.4) with $\langle v\rangle \propto \sqrt{T}$ gives a pure power law,

$$\eta \propto T^{s}\,, \qquad s = \frac{1}{2} + \frac{2}{\nu-1}\,. \tag{57.6}$$

Hard spheres are the limit $\nu \to \infty$, for which $s = \frac{1}{2}$: the cross-section is a constant of the molecule and the whole temperature dependence comes from $\langle v\rangle$. The observed $s \approx 1$ requires $2/(\nu-1) = \frac{1}{2}$, that is $\nu = 5$. So the measurement does not merely disagree with hard spheres; it returns a number, the exponent of the force law, and it is a bulk measurement of a molecular interaction. □

57.4 Graham: diffusion, effusion and the transport coefficients (1833–1846)

The kinetic hypothesis nearly foundered on an objection of Buys Ballot: if the molecules of the air move at hundreds of metres per second, a smell released at one end of a room should reach the other end at once, and it does not. Clausius's answer was the mean free path [149]. The quantitative form of the objection and of its answer is the diffusion coefficient, and Graham had measured the relevant rates before either was written down [94, 120].

57.4.1 Apparatus

For diffusion: a graduated glass tube closed at one end by a plug of plaster of Paris—a septum whose pores are small but very numerous—and standing with its open end under water or mercury, so that the change of volume as the enclosed gas exchanges with the outside air is read directly on the graduations. For transpiration: long capillary tubes of known bore through which the gas is driven by a measured pressure difference. For effusion: a fine aperture in a thin plate, small compared with the mean free path, which is the case the kinetic theory treats exactly.

57.4.2 Procedure

Fill the tube with the test gas, note the initial volume, and let the gas interchange with the atmosphere through the septum until the level has moved by a measured amount; time the process. Repeat with a series of gases under identical conditions of temperature, pressure and apparatus, so that everything except the molecular mass is held fixed and the ratios of the rates carry the whole result. The same discipline governs the capillary and aperture versions.

57.4.3 Observations and data

[Reserved: Graham's rate tables from [94] and [120], to be reproduced here as diffusion rates relative to air for hydrogen, oxygen, nitrogen, carbon dioxide and the other gases he used, set against $M^{-1/2}$, with the residuals and their uncertainties. To be tabulated alongside them, for the comparison of Equation (57.9), are the modern transport coefficients of air near $300\,\mathrm{K}$ *and* $100\,\mathrm{kPa}$*: a viscosity of about* $18\,\mu\mathrm{Pa\,s}$*, a thermal conductivity of about* $26\,\mathrm{mW/(m\,K)}$*, and a self-diffusion coefficient of order* $2\times10^{-5}\,\mathrm{m^2/s}$*.]*

57.4.4 Interpretation

Graham's law—rate proportional to $M^{-1/2}$ at fixed temperature and pressure—is a statement about mass and not about chemistry, and that is its whole force: nothing in a chemical account of gases would predict that carbon dioxide and nitrous oxide, which behave quite differently in every reaction, should diffuse at the same rate because they happen to have the same molecular mass.

The coefficients are the sharper test. Kinetic theory does not merely predict that gases diffuse; it predicts a diffusion coefficient built from the same $\langle v\rangle$ and the same ℓ that appear in the viscosity, and therefore predicts relations *among* the three transport coefficients with no free parameter at all. Those relations hold in order of magnitude, which is the content of Phenomenon 57.6, and fail in their numerical coefficients—the elementary argument gives a heat-conduction to viscosity ratio of 1 where the monatomic gases give about $\tfrac{5}{2}$. The discrepancy is not a failure of the molecular picture but of the averaging: fast molecules carry more energy than slow ones but the same momentum, so the three transports do not sample the same molecules and cannot share one mean free path. Repairing that is the

systematic expansion of Chapman and Enskog [434, 447, 1416], developed in Chapter 124, whose continuum limit is the hydrodynamics of Chapter 31.

57.4.5 Primary references

[94, 120]; the phenomenological law of diffusion is [143] and that of heat conduction [72]. The mean free path that reconciles the rates with the speeds is [149].

Phenomenon 57.5 (Gases interdiffuse far more slowly than they move)**.** The molecules of air travel at some hundreds of metres per second, yet a gas released at one point of a still room takes hours to spread across it by diffusion alone. The measured interdiffusion coefficients of common gases at standard conditions are of order $2 \times 10^{-5}\,\mathrm{m^2/s}$ [94, 143], some ten orders of magnitude below the product of a molecular speed and a room. The discrepancy is not evidence against the kinetic hypothesis but a measurement of the mean free path [149].

Derivation. A molecule does not travel in a straight line but executes a random walk of step ℓ at speed $\langle v\rangle$, so the transport is diffusive with

$$D \approx \tfrac{1}{3}\,\langle v\rangle\,\ell\,, \tag{57.7}$$

by the same carrier argument that gave Equation (57.4), with the number density of a labelled species in place of the momentum density. With $\langle v\rangle \approx 470\,\mathrm{m/s}$ and $\ell \approx 1 \times 10^{-7}\,\mathrm{m}$ for air near 300 K and 100 kPa—the values established in Phenomenon 57.7—this gives $D \approx 1.6 \times 10^{-5}\,\mathrm{m^2/s}$, which is the observed order. The time to cross a distance L follows from $L^2 \approx 2Dt$: about a hundredth of a second for a millimetre, but some hours for a metre, since the time grows as the *square* of the distance. The suppression relative to free flight is the ratio ℓ/L, which is about 10^{-7} for a metre—and that, and not any slowness of the molecules, is the answer to Buys Ballot. □

Phenomenon 57.6 (The three transport coefficients of a gas are not independent)**.** Viscosity, thermal conductivity and diffusion in a dilute gas are carried by the same molecules over the same mean free path, so kinetic theory predicts that the dimensionless combinations $\rho D/\eta$ and $\kappa m/\left(\eta c_V^{(1)}\right)$ are pure numbers of order unity, independent of the gas, of its density and of its temperature. Measurement confirms that they are numbers of order unity, and shows that the elementary estimate of those numbers is wrong: for the monatomic gases the second ratio is close to $\frac{5}{2}$ rather than to 1 [167, 403, 434, 447].

Derivation. Each coefficient is the same transport step applied to a different carried quantity. A molecule crossing a plane carries the value its last collision gave it, one mean free path away, so for any transported density χ the flux is of order $-\frac{1}{3}\,\langle v\rangle\,\ell\,\mathrm{d}\chi/\mathrm{d}z$. Taking χ to be the momentum density, the energy density and the number density of a labelled species in turn,

$$\eta \approx \tfrac{1}{3}\rho\,\langle v\rangle\,\ell\,, \qquad \kappa \approx \tfrac{1}{3} n c_V^{(1)}\,\langle v\rangle\,\ell\,, \qquad D \approx \tfrac{1}{3}\,\langle v\rangle\,\ell\,, \tag{57.8}$$

with $c_V^{(1)}$ the heat capacity of a single molecule. The mean free path and the mean speed—the two quantities that are hard to measure and easy to get wrong—cancel between them, leaving

$$\frac{\rho D}{\eta} \approx 1\,, \qquad f := \frac{\kappa m}{\eta c_V^{(1)}} \approx 1\,. \tag{57.9}$$

That these are pure numbers is a prediction with no adjustable content whatever, and it is confirmed. That they are not equal to unity is equally informative. The elementary argument gave one ℓ to all three transports, but the molecules that dominate the energy

flux are faster than those that dominate the momentum flux, and a faster molecule has both a longer free path and a larger weight; the correct averages therefore differ, and by a factor of order unity that only a solution of the transport equation can supply. □

Derivation pending #5

The Chapman–Enskog transport coefficients: that the first-order solution of the Boltzmann transport equation about local equilibrium fixes the numerical coefficients left undetermined by the mean-free-path argument, and in particular gives the ratio of thermal conductivity to viscosity times specific heat as 5/2 for a monatomic gas, against the value 1 of the elementary estimate.

57.5 Loschmidt: the size of an air molecule (1865)

Every measurement so far in this chapter has been of a bulk property. Loschmidt's determination is the step that converts two of them into a molecular length and a molecular count, and it is the first time anyone put a number on the size of a molecule [163]. It is included among the experiments, although it built no apparatus of its own, because its inputs are measurements and its output is a number about the world; the honest description of it is a determination, not a theory.

57.5.1 Apparatus

None of its own. The determination consumes two independent laboratory results: the viscosity of the gas, from an experiment of the kind described in Section 57.3, and the density of the same substance in its condensed state, from weighing a known volume of the liquid. Both were available in 1865, and neither had been made with molecules in mind.

57.5.2 Procedure

From the measured viscosity and mass density obtain the mean free path, $\ell \approx 3\eta/(\rho\,\langle v\rangle)$, with $\langle v\rangle$ fixed by the molar mass and the temperature alone. From the volume of liquid obtained from unit volume of gas obtain the condensation ratio ε. Assume the molecules are spheres that fill the liquid when packed. Then ℓ and ε are two equations in the two unknowns—the molecular diameter d and the number density n—and both follow.

57.5.3 Observations and data

[Reserved: Loschmidt's own inputs and outputs from [163], converted to SI—the viscosity and condensation ratio he used, the diameter he obtained (about $1\,\mathrm{nm}$*), and the number density that follows. The comparison to record beside them: on modern inputs the same equations give a collision diameter near* $3\times10^{-10}\,\mathrm{m}$*, and the Loschmidt constant, the number density of an ideal gas at* $273.15\,\mathrm{K}$ *and* $101\,325\,\mathrm{Pa}$*, is* $2.6867\times10^{25}/\mathrm{m}^3$ *[2431], while the Avogadro constant has been exact since 2019 at* $6.022\,140\,76\times10^{23}/\mathrm{mol}$ *[2333].]*

57.5.4 Interpretation

The packing assumption is crude and the answer is correspondingly crude—Loschmidt's diameter is several times the modern collision diameter, and since n varies as d^{-2} his number density is low by about an order of magnitude. That is not the point, and treating it as a failed precision measurement misreads what happened. Before 1865 the molecular hypothesis had no scale at all: molecules could as well have been 10^{10} times smaller. Afterwards it had

one, obtained from two measurements that mention no molecules, and the scale was small enough to explain why molecules had never been seen and large enough to be consistent with the mean free path.

What made the number believable was not its accuracy but its corroboration by methods sharing neither apparatus nor theory: the Brownian determinations of Chapter 119 [358, 409], the electrolytic route through the elementary charge [407], and the counting of particles from a radioactive source. Agreement among routes with nothing in common but the hypothesis is what an inference of this kind can be worth, and it is why the molecular constitution of matter stopped being an opinion around 1910.

57.5.5 Primary references

[163]; the viscosity input is of the kind measured in [167] and the mean free path is [149]. The modern values are [2431] and [2333].

Phenomenon 57.7 (The mean free path is a measurable length)**.** The average distance a molecule travels between collisions is not a theoretical fiction: it is fixed by measurements that never mention molecules, since inverting the viscosity relation expresses it through the measured viscosity, the weighed mass density and the molar mass alone [149, 167]. For air near 300 K and 100 kPa it comes out of order 1×10^{-7} m: about three hundred times the molecular diameter, and seven orders of magnitude below the size of any vessel—which is why the gas behaves as a continuum in Chapter 31 and ceases to when it does not [355].

Derivation. Solve Equation (57.4) for the mean free path,

$$\ell \approx \frac{3\eta}{\rho \langle v \rangle}, \qquad \langle v \rangle = \sqrt{\frac{8RT}{\pi M}}, \tag{57.10}$$

and note what stands on the right: η from a viscometer, ρ from a balance, and $\langle v \rangle$ from the molar mass and the thermometer. With $\eta \approx 18\,\mu\text{Pa s}$, $\rho \approx 1.2\,\text{kg/m}^3$ and $M \approx 0.029\,\text{kg/mol}$ at 300 K, so that $\langle v \rangle \approx 470\,\text{m/s}$, this gives $\ell \approx 1 \times 10^{-7}\,\text{m}$. The cross-section follows from $\ell = 1/\left(\sqrt{2} n \sigma\right)$ with $n = P/k_\text{B}T \approx 2.4 \times 10^{25}/\text{m}^3$, giving $\sigma \approx 3 \times 10^{-19}\,\text{m}^2$ and a collision diameter $d = \sqrt{\sigma/\pi} \approx 3 \times 10^{-10}\,\text{m}$. Loschmidt's second equation then closes the system: taking the molecules to be spheres of diameter d packed so as to fill the liquid, the condensation ratio is $\varepsilon = n\pi d^3/6$, and eliminating n between it and $\ell = 1/\left(\sqrt{2}\pi d^2 n\right)$ leaves $d = 6\sqrt{2}\,\varepsilon\ell$—a molecular diameter from a condensation ratio and a viscosity. The Avogadro constant is then $N_\text{A} = nV_\text{m}$ with V_m the molar volume. Every step but the packing assumption is exact; that one step is what makes the result an order of magnitude rather than a measurement. □

57.6 Joule and Thomson: the porous plug and the forces between molecules (1852)

The ideal gas of Chapter 54 is the gas whose molecules do not interact except in the instant of collision. Kinetic theory needs to know whether that is true, and if not, by how much and with what sign. The porous-plug experiment answers all three questions with one thermometer, because the effect it measures vanishes identically for a gas of non-interacting molecules [137].

57.6.1 Apparatus

A long tube, thermally lagged and of low conductivity, with a plug of cotton wool or unspun silk packed tightly into the bore so as to present a large flow resistance without any moving

part. Compressed gas is supplied steadily from a pump through a coil immersed in a bath of known temperature, so that the gas arrives at the plug at a controlled temperature; thermometers sit in the stream immediately upstream and downstream of the plug, and pressure gauges read the pressure on each side.

57.6.2 Procedure

Establish a steady flow and hold it until the temperatures cease to drift, which is what makes the analysis a steady-flow one and lets the enthalpy balance of Chapter 56 apply. Read the upstream and downstream temperatures and pressures. The plug is essential and is not merely a convenient resistance: it destroys the kinetic energy of the stream, so that no part of the temperature change can be attributed to the gas being set in bulk motion. Repeat for different gases, different pressure drops and different bath temperatures, the last being what maps the change of sign.

57.6.3 Observations and data

[Reserved: the primary series of [137] and its sequels, to be reproduced here in SI—upstream and downstream pressures in Pa*, bath temperature in* K*, and the measured temperature drop in* K *with its uncertainty, for each gas and each pressure drop, together with the coefficient* $\mu = (\partial T/\partial P)_H$ *in* K/Pa *derived from them. The qualitative pattern the table must show: air at room temperature cools by a fraction of a kelvin per atmosphere of pressure drop, carbon dioxide cools by several times more, and hydrogen warms instead of cooling.]*

57.6.4 Interpretation

The kinetic reading of the result is the one that concerns this chapter. A gas of point molecules that never interact has an internal energy depending on temperature alone and an equation of state $PV = nRT$, and for such a gas the throttling coefficient is identically zero—not small, but exactly zero, as the derivation below shows. Every kelvin measured here is therefore a direct measurement of the interactions the ideal gas omits. Better, the *sign* separates them: cooling means that work has been done against attraction as the molecules were pulled apart, warming that the molecules were already so crowded that their own volume dominates. A gas that does both, at different temperatures, is telling the experimenter that both a long-range attraction and a short-range repulsion exist, and where they balance.

This complements Section 57.3 exactly. The viscosity measurement probes the repulsive core, because it is collisions that transport momentum, and returned an exponent for the repulsive force. The porous plug probes the attractive tail, because it is the work done against attraction on expansion that cools the gas. Between them they outline the intermolecular potential from a thermometer and a torsion fibre—a fact worth pausing over, since neither apparatus can resolve anything smaller than a millimetre.

57.6.5 Primary references

[137]. The cruder free expansion that failed to see the effect is [112]; the thermodynamic treatment of both, with the enthalpy argument and the inversion curve, is in Chapter 56 and Chapter 54.

Phenomenon 57.8 (Molecules exert forces on one another at a distance)**.** Forcing a real gas steadily through a porous plug from a higher to a lower pressure, along a thermally insulated tube, changes its temperature by a small but definite amount, and the sign of the change depends on the gas and on its temperature: most gases at room temperature emerge

cooler, hydrogen and helium emerge warmer, and each gas has an inversion temperature separating the two behaviours. Joule and Thomson established the effect in the series beginning in 1852 [137]. For a gas of molecules exerting no forces at a distance the effect would be exactly zero at every temperature and every pressure, so its magnitude and sign measure the intermolecular potential directly.

Derivation. The measured quantity is $\mu = (\partial T/\partial P)_H$, since throttling conserves enthalpy—the steady-flow first law of Chapter 56—and by the chain rule together with a Maxwell relation,

$$\mu = \frac{1}{C_P}\left[T\left(\frac{\partial V}{\partial T}\right)_P - V\right]. \tag{57.11}$$

For the ideal gas $V = nRT/P$ gives $T\,(\partial V/\partial T)_P = V$ identically, so $\mu \equiv 0$: nothing whatever is measured, at any pressure or temperature. This is what makes the experiment a clean test. Now let the molecules attract weakly at long range and exclude a volume b per mole at short range, so that to first order in the departures from ideality $V_{\mathrm{m}} \approx RT/P + b - a/RT$. Then $T\,(\partial V_{\mathrm{m}}/\partial T)_P = RT/P + a/RT$, and

$$\mu \approx \frac{1}{C_{P,\mathrm{m}}}\left(\frac{2a}{RT} - b\right). \tag{57.12}$$

The two molecular effects enter with opposite signs, exactly as the observation requires: the attraction a cools, the excluded volume b warms, and the attraction loses its advantage as the gas is heated because the molecules pass too quickly to be held. The coefficient vanishes at the inversion temperature

$$T_{\mathrm{inv}} = \frac{2a}{Rb}, \tag{57.13}$$

above which throttling warms the gas. Hydrogen and helium have small a and therefore inversion temperatures far below room temperature, which is why they warm on throttling and must be precooled before a Joule–Thomson stage will liquefy them—an inference from Equation (57.13) that the liquefaction industry has tested more thoroughly than any laboratory. Finally, b is the same excluded volume that appears as the collision cross-section in Phenomenon 57.7: two independent measurements of one molecular size, which is the consistency check that makes the whole scheme more than a fit. □

57.7 Summary of the evidence

No entry in Table 57.1 can be obtained from thermodynamics, and none of them contains an adjustable parameter beyond the one molecular size and one force law that the whole set shares. Taken together with the Brownian measurements of Chapter 119, which watch the fluctuations that these experiments only average over, they are the experimental case for the hypothesis of Chapter 55: that a gas is a very large number of small bodies in disordered motion, obeying the mechanics of Chapter 19. What they do not settle is the statistical foundation—why the equilibrium distribution is the one observed, independently of any model of the molecules—which is the work of Chapter 118, nor the failure of equipartition at low temperature, which no classical mechanics can repair and which Chapter 125 and Chapter 68 resolve.

57.8 Giauque: adiabatic demagnetization and the third law (1933)

This last section is an addendum to the six tests summarised in Table 57.1 and not a seventh of them: it puts no prediction of Chapter 55 at risk. It is placed here because

Table 57.1: Tests of the kinetic theory of gases reported in this chapter. The middle column names the prediction of Chapter 55 that the experiment puts at risk; in every case a purely thermodynamic account of gases predicts nothing at all for the quantity measured, which is what makes these experiments evidence for molecules rather than merely for heat.

Experiment	Prediction tested	Result
Stern 1920	thermal speeds	flight time over a known distance gives speeds of the predicted order, scaling as $\sqrt{T/M}$
Miller–Kusch 1955	shape of the distribution	beam intensity follows $v^3 \exp(-mv^2/2k_{\mathrm{B}}T)$ at the per-cent level, at the oven temperature
Maxwell 1866	viscosity independent of density	confirmed over the accessible pressure range; and $\eta \propto T$, not $\sqrt{T}$, giving the force exponent $\nu = 5$
Graham 1833–1846	transport from one mean free path	rates scale as $M^{-1/2}$; the coefficients obey the predicted relations in order of magnitude, with the numbers fixed only by Chapman–Enskog
Loschmidt 1865	a molecular scale exists	a diameter and a number density from viscosity plus a condensation ratio, correct to an order of magnitude
Joule–Thomson 1852	the ideal gas omits forces	a nonzero throttling coefficient of both signs, where non-interacting molecules give exactly zero

the paragraph above names the boundary at which the classical account of matter fails, and because crossing that boundary was itself an experiment. The third law is stated in Chapter 54 in the form Nernst gave it [328, 378], and the unattainability of absolute zero is deduced there as a theorem of thermodynamics. Both halves are also measurements, and the measurements came out of one programme: the cooling of paramagnetic salts by adiabatic demagnetization, proposed independently by Debye in 1926 and by Giauque in 1927, achieved by Giauque and MacDougall in 1933, and recognised by the Nobel Prize in Chemistry for 1949. What the technique opened was not merely a lower temperature but a region in which the entropy of matter can be watched approaching its limit.

57.8.1 Apparatus

A paramagnetic salt in which the magnetic ions are held far apart by water of crystallisation: Giauque and MacDougall used hydrated gadolinium sulfate, $\mathrm{Gd_2(SO_4)_3 \cdot 8H_2O}$, whose $\mathrm{Gd^{3+}}$ ion has a half-filled shell, and therefore no orbital moment to be quenched, and a ground multiplet of eight equally spaced states. The salt hangs in a vacuum space inside a bath of liquid helium—available at all only since [343]—pumped below its normal boiling point to reach somewhat under 1 K. An electromagnet supplies a flux density of order a tesla over the sample. The thermal switch between sample and bath is helium exchange gas, admitted to couple them and pumped away to isolate the sample; there is no mechanical link, because at these temperatures any solid support conducts far too well. Thermometry is the sample itself: two coaxial coils around the vacuum space read its magnetic susceptibility by mutual inductance, and the Curie law of Chapter 129 converts a susceptibility into a temperature.

That last step is the subtle one, and the apparatus has to answer for it. The Curie law is the high-temperature limit of the paramagnetic response [264, 323, 2011], and it fails exactly where the experiment is going; what the bridge reads is therefore a *magnetic* temperature T^{*} defined by $\chi = C/T^{*}$, which coincides with the thermodynamic temperature only while the ions remain independent. Converting one into the other needs a separate calorimetric measurement: feed the isolated salt a measured quantity of energy, follow the resulting change in T^{*}, and recover the thermodynamic temperature from $T^{-1} = (\partial S/\partial U)$. A magnetic temperature quoted as a thermodynamic one is the characteristic error of this field, and it is an error in the direction of claiming too much.

57.8.2 Procedure

Four steps, run as a cycle. First, with exchange gas in place and the salt anchored to the bath at T_{i}, raise the field to B_{i}: the moments align, the magnetic entropy falls, and the heat of magnetization passes into the helium, which is what makes this step isothermal. Second, pump the exchange gas away, isolating the salt. Third, lower the field to zero, as slowly as the eddy-current heating of the metal parts requires, and no more slowly than the heat leak allows; the salt cools. Fourth, read the susceptibility bridge repeatedly as the salt warms back up, which yields the final temperature, the heat capacity, and—from the slope of the warm-up—the size of the parasitic heat input that has to be subtracted from it. The whole is repeated over a series of initial fields and initial bath temperatures, because the prediction under test is a relation between a ratio of temperatures and a ratio of fields, not a single number.

57.8.3 Observations and data

[Reserved: the primary series of Giauque and MacDougall (1933) and of de Haas, Wiersma and Kramers in the same year, to be reproduced here in SI—initial bath temperature in K*, initial flux density in* T*, the susceptibility before and after, the magnetic temperature* T^{*}

derived from it, and the thermodynamic temperature with its uncertainty stated per [2129]; alongside them the calibration run, as energy input in J *against entropy in* $\mathrm{J/(mol\,K)}$*, from which the zero-field curve* $S(T)$ *is built. Three orientation figures, computed here from stated inputs rather than taken from the papers: the ground multiplet of* $\mathrm{Gd^{3+}}$ *carries a molar entropy* $R\ln 8 = 17.3\,\mathrm{J/(mol\,K)}$ *per mole of ions; the Bohr magneton divided by the Boltzmann constant is* $0.67\,\mathrm{K/T}$ *[2333, 2431], so at* $1\,\mathrm{T}$ *that multiplet is spread over about* $9\,\mathrm{K}$ *and a bath near* $1\,\mathrm{K}$ *is cold enough to order it; and the lattice entropy competing with it is smaller by orders of magnitude, by the Debye law used in Equation* (57.15)*. The first demagnetizations reached a few tenths of a kelvin, and the more dilute salts used later reach a few millikelvin, which is the regime Chapter 110 works in.]*

57.8.4 Interpretation

Two facts make the method work, and neither is incidental. The first is that at $1\,\mathrm{K}$ a solid has almost no entropy left in its lattice: by the Debye law [382] the vibrational entropy falls as T^3, so for a salt of Debye temperature of order $1\times10^2\,\mathrm{K}$ it stands four or more orders of magnitude below the magnetic entropy. Essentially all the disorder that remains is the orientation of the magnetic ions, and that is precisely the disorder a magnetic field can address. The second is that the ions are dilute: the water of crystallisation holds them far enough apart that they behave independently well below the temperature of the bath, which is what makes the magnetic entropy a function of B/T alone and gives the cooling the simple law of Equation (57.16).

The same dilution sets the limit, and it is there that the experiment stops being a refrigeration technique and becomes a statement about the third law. The ions are not perfectly independent: each sits in the dipolar field of its neighbours and in the electric field of the crystal, and the residual internal field that results cannot be switched off [659, 744]. Demagnetizing to zero *applied* field therefore leaves the moments in that residual field rather than in nothing, and the temperature falls only in the corresponding ratio. Stage a second cycle below the first and the same factor applies again, so the temperature after several stages is a geometric sequence with a positive ratio: it has zero as a limit and never as a value. That is the unattainability statement, and here it is not a postulate but the operating characteristic of a machine, measurable stage by stage.

It is worth pausing on the fact that the two halves of the third law are one fact seen from two sides. Were the ions exactly independent, the residual field would vanish, a single stage would reach absolute zero, and the entropy at zero temperature would be $R\ln(2J+1)$ per mole in zero field and zero in any field however small—so Nernst's theorem, which requires the limiting entropy to be independent of the other variables, would fail in the same breath in which unattainability failed. The interactions that put a floor under the refrigerator are the interactions that give the salt a non-degenerate ground state. A thermodynamic law that reads like a prohibition is, microscopically, the statement that matter has a ground state.

The approach of the entropy to a constant was established by a second route in the same programme, and the route matters because at the bottom it uses no thermodynamics at all. The entropy of a substance can be obtained calorimetrically, by integrating C_P/T upward from the lowest attainable temperature and closing the gap below it with the Debye law; and it can be computed, for a gas, from spectroscopic constants alone by way of the partition function of Chapter 118. These are independent determinations of the same quantity, and the calorimetric one is complete only if the entropy left below the integration limit is zero. For most substances the two agree. Where they do not, the discrepancy is not scatter but a number of a particular kind—close to $R\ln w$ for a small w, with w counting the arrangements a molecule can be frozen into—and that is a far stronger confirmation of the third law than agreement alone would be, because it says what the limiting entropy is and

not merely that there is one.

Finally, the technique outlived the question it was built for. Cooling a paramagnetic salt to a few millikelvin puts the hyperfine field of a nucleus embedded in it to work: once the nuclear Zeeman splitting is comparable with $k_{\mathrm{B}}T$, the nuclei align, and an ensemble of oriented nuclei is a laboratory sample with a definite direction in it. That is exactly how the cobalt-60 of Chapter 110 was polarized—the active layer grown on a cerium magnesium nitrate crystal, demagnetized, and left to warm through the useful window [1101, 1133]. A method devised to test a law of thermodynamics is what made a symmetry of the weak interaction measurable, which is the kind of debt between fields that a treatise organised by subject tends to hide.

57.8.5 Primary references

The primary sources of this section are named here in full because they are not yet in the bibliography of this treatise; the works cited by key below are the supporting ones. The proposal is P. Debye, "Einige Bemerkungen zur Magnetisierung bei tiefer Temperatur", *Annalen der Physik* **386** (1926) 1154–1160, and, independently, W. F. Giauque, "A Thermodynamic Treatment of Certain Magnetic Effects. A Proposed Method of Producing Temperatures Considerably Below 1° Absolute", *Journal of the American Chemical Society* **49** (1927) 1864–1870. The achievement is W. F. Giauque and D. P. MacDougall, "Attainment of Temperatures Below 1° Absolute by Demagnetization of $\mathrm{Gd_2(SO_4)_3 \cdot 8H_2O}$", *Physical Review* **43** (1933) 768, confirmed within months by W. J. de Haas, E. C. Wiersma and H. A. Kramers in *Physica* **1** (1933). The residual entropy of ice is measured in W. F. Giauque and J. W. Stout, *Journal of the American Chemical Society* **58** (1936), and counted in L. Pauling, "The Structure and Entropy of Ice and of Other Crystals with Some Randomness of Atomic Arrangement", *Journal of the American Chemical Society* **57** (1935) 2680–2684, whose argument is reproduced in [1188]. The nuclear extension is N. Kurti, F. N. H. Robinson, F. Simon and D. A. Spohr, "Nuclear Cooling", *Nature* **178** (1956) 450–453.

[Reserved: the eight references named in full above are to be entered in the bibliography and the prose replaced by citations. Until then the supporting keys are [328, 378] for the heat theorem, [382] for the lattice entropy, [264, 323, 2011] for the paramagnetism, [659, 744] for the interactions that set the floor, [343] for the helium that preceded the technique, and [1101, 1133] for the use later made of it.]

Phenomenon 57.9 (Demagnetizing an isolated paramagnetic salt cools it)**.** A paramagnetic salt magnetized isothermally in contact with a bath near 1 K, then thermally isolated and demagnetized, ends colder than the bath, and the drop is fixed by the ratio of the fields rather than by any property of the apparatus: the final temperature is the initial one multiplied by the ratio of the residual field to the applied field. Giauque and MacDougall reached a few tenths of a kelvin from hydrated gadolinium sulfate in 1933, and de Haas, Wiersma and Kramers did the same within months. With more dilute salts the method reaches a few millikelvin [1133] and, applied to nuclear rather than electronic moments, the microkelvin range. Before it, the lowest temperatures in existence were those of pumped liquid helium [343].

Derivation. Take N non-interacting ions of angular momentum J and Landé factor g in a flux density B. Each level of the multiplet has energy $E_m = g\mu_{\mathrm{B}}Bm$ with $m = -J, \dots, J$, so the single-ion partition function

$$Z = \sum_{m=-J}^{J} \exp\left(-\frac{g\mu_{\mathrm{B}}Bm}{k_{\mathrm{B}}T}\right) \tag{57.14}$$

depends on B and T only through their ratio: the field scales every energy in the same proportion, and the temperature is measured against energy. The magnetic entropy inherits

the property, since with $x = g\mu_{\mathrm{B}}B/k_{\mathrm{B}}T$ the free energy is $-Nk_{\mathrm{B}}T\ln Z(x)$ and

$$S_{\mathrm{m}} = Nk_{\mathrm{B}}\left(\ln Z - x\,\frac{\mathrm{d}\ln Z}{\mathrm{d}x}\right) = S_{\mathrm{m}}(B/T)\ . \tag{57.15}$$

A reversible demagnetization of a thermally isolated sample holds the total entropy fixed. The lattice carries almost none of it—by the Debye law the vibrational entropy per mole of atoms is $(4\pi^4/5)\,R\,(T/\Theta_{\mathrm{D}})^3$ [382], which near 1 K is smaller than $R\ln(2J+1)$ by orders of magnitude—so holding the total entropy fixed holds S_{m} fixed, and by Equation (57.15) it holds B/T fixed:

$$\frac{T_{\mathrm{f}}}{T_{\mathrm{i}}} = \frac{B_{\mathrm{f}}}{B_{\mathrm{i}}}\,. \tag{57.16}$$

Nothing in Equation (57.16) refers to the particular salt, which is what makes it a prediction rather than a fit; the salt enters only through how small B_{f} can be made and how much entropy $R\ln(2J+1)$ there was to remove. The two assumptions are equally visible: the ions must be independent, which fails once $k_{\mathrm{B}}T$ falls to their interaction energy, and the process must be slow enough to be reversible, which eddy currents in any metal near the sample threaten. □

Phenomenon 57.10 (The entropy of a substance tends to a constant, and the constant counts ground states)**.** The entropy of a substance measured calorimetrically—by integrating C_P/T upward from the lowest attainable temperature—agrees, for most substances, with the entropy computed from spectroscopic data through the partition function; and the agreement requires that the entropy remaining below the integration limit be zero. Where the two determinations disagree, the discrepancy is not scattered but close to $R\ln w$ with w a small number of arrangements per molecule frozen in as the substance is cooled. For ice it is near $R\ln(3/2) = 3.37\,\mathrm{J/(mol\,K)}$, which is the count of proton placements consistent with two protons per oxygen [1188]; for carbon monoxide, whose two ends are nearly indistinguishable, it is of the order of $R\ln 2 = 5.76\,\mathrm{J/(mol\,K)}$. The entropy therefore tends as $T \to 0$ to a constant fixed by the degeneracy of the ground state and not by the substance's chemistry, which is Equation (54.13) in its sharp form.

Derivation. The calorimetric statement is Equation (54.14): the integral $\int_0^T C_x\,\mathrm{d}T'/T'$ converges, so it defines $S(T) - S(0)$ and the calorimetric route measures entropy *differences* only. The absolute value it is compared against comes from Boltzmann's counting [196], $S = k_{\mathrm{B}}\ln W$, which at zero temperature counts ground states. If each molecule can be frozen into any of w arrangements independently of its neighbours then $W = w^N$ and

$$S(0) = Nk_{\mathrm{B}}\ln w = R\ln w \tag{57.17}$$

per mole: a discrete set of admissible values, which is why a discrepancy between the two routes is diagnostic rather than merely disappointing.

Pauling's count for ice is the worked case, and it needs only the two rules the hydrogen bonds impose. Each oxygen has four tetrahedral neighbours and keeps two protons close to it, which it may place on any two of the four bonds: $\binom{4}{2} = 6$ arrangements per molecule, hence 6^N if the molecules were free of one another. Each of the $2N$ bonds must carry exactly one proton; of the four ways the two molecules sharing a bond may independently place a proton on it or not, two satisfy that rule and two do not, so each bond costs a factor $\frac{1}{2}$. Treating the bonds as independent,

$$W \approx 6^N\left(\tfrac{1}{2}\right)^{2N} = \left(\tfrac{3}{2}\right)^N\,, \qquad S(0) = R\ln\tfrac{3}{2}\,. \tag{57.18}$$

The independence of the bonds is an approximation, and the exact counting on the ice lattice exceeds Equation (57.18) by well under a per cent—far less than the calorimetric uncertainty. The comparison therefore tests the physical claim, that the protons are disordered, rather than the combinatorics. □

Phenomenon 57.11 (Absolute zero is unattainable, and the unattainability is an operating characteristic)**.** Every stage of a demagnetization cycle multiplies the temperature by a ratio smaller than one but bounded away from zero, because the applied field cannot be reduced below the field the magnetic ions exert on one another. A finite number of stages therefore leaves a finite temperature, and the successive stages of every cooling technique are observed to yield diminishing returns, exactly as Phenomenon 54.7 requires. The statement is not a postulate of thermodynamics tested at one remove: the bounding ratio is a measurable property of the salt [659, 744], and the machine displays the prohibition in its own performance.

Derivation. Write the field at an ion as the applied flux density combined with an internal field of magnitude b, arising from the dipolar field of the neighbours and from the crystal-field splitting of the multiplet; b is a property of the lattice and does not respond to the magnet. The smallest field a demagnetization can leave is therefore b, and Equation (57.16) gives

$$T_{\mathrm{f}} \geq \frac{b}{B_{\mathrm{i}}}\, T_{\mathrm{i}}\,, \tag{57.19}$$

so that n identical stages reach $T_n \geq (b/B_{\mathrm{i}})^n\, T_{\mathrm{i}} > 0$ for every finite n. A geometric sequence of positive ratio has zero as a limit and never as a value, which is unattainability in the form the apparatus states it.

The two halves of the third law meet here. Suppose b were zero, so that the bound vanished and one stage sufficed. The ions would then be exactly independent at every temperature, the zero-temperature entropy would be $R \ln (2J+1)$ per mole in zero field and zero in any field however small, and Equation (54.13)—the requirement that the limiting entropy not depend on the other variables—would fail in the same breath. Conversely, an interaction strong enough to order the moments discharges $R \ln (2J+1)$ before absolute zero is reached and makes the limiting entropy independent of the field, at the price of a floor under the cooling. Unattainability and Nernst's theorem are one statement about the ground state, read once as a prohibition and once as a limit. □

Derivation pending #6

The residual internal field of a paramagnetic salt: that the dipolar sum over the lattice of magnetic ions, together with the crystal-field splitting of the ground multiplet, fixes the ordering temperature and hence the floor under a single demagnetization stage, so that the cooling ratio is a computable property of the salt rather than a fitted parameter.

Derivation pending #7

The spectroscopic entropy of a gas: that the translational, rotational, vibrational and electronic partition functions, evaluated from band-spectrum constants and molecular masses, give the absolute molar entropy against which the calorimetric integration is compared, together with the symmetry number and the nuclear-spin terms that cancel between the two routes.

Part VII

Classical Field Theory and Electrodynamics

Chapter 58

Electrostatics

Electrostatics is the theory of the fields produced by charges at rest. The chapter follows the order of the source: the Coulomb force is postulated, the electrostatic field and its potential are constructed by continuous superposition, the local field equations — Gauss's law and the vanishing of the curl — are derived, and the boundary-value methods for obtaining the potential are developed. The Coulomb constant is kept throughout as k_E, so that every formula remains valid in any system of units; the worked examples, ported from the companion problem collection, are written in SI units, where $k_E = 1/(4\pi\epsilon_0)$.

58.1 The electrostatic field

58.1.1 Fundamentals

Point charge

Definition 58.1 (Point charge)**.** A *point charge* is a point of space to which one assigns a constant real scalar quantity q called its *charge*.

The Coulomb force

We state as a fundamental principle:

Postulate 58.2 (Coulomb's law)**.** The force that a charge q' located at $\boldsymbol{x}'$ exerts on a charge q located at $\boldsymbol{x}$ is proportional to the product of the charges and inversely proportional to the square of the distance separating them. That is,

$$\boldsymbol{F} = k_E \, \frac{qq'}{|\boldsymbol{x} - \boldsymbol{x}'|^2} \, \frac{\boldsymbol{x} - \boldsymbol{x}'}{|\boldsymbol{x} - \boldsymbol{x}'|} \,, \tag{58.1}$$

where k_E is a positive real constant.

Remark 58.3 (Unit systems). The constant k_E selects the system of units: in Gaussian units $k_E = 1$, while in SI units $k_E = 1/(4\pi\epsilon_0)$, with ϵ_0 the vacuum permittivity [1976, 2333]. The worked examples of this chapter are written in SI units.

Continuous charge distributions

Definition 58.4 (Continuous charge distribution)**.** Consider a set of infinitesimal charges $\mathrm{d}q'$ located at points $\boldsymbol{x}'$ (with respect to a fixed coordinate origin) which together form a continuum occupying a volume V'. We call this continuum a *continuous charge distribution*.

Consider further a charge q at $\boldsymbol{x}$. The *force felt by q due to a continuous charge distribution* is the continuous sum of the forces that the individual charges $\mathrm{d}q'$ exert on it. By Coulomb's law,

$$\boldsymbol{F} = k_E\, q \int_{V'} \mathrm{d}q'\, \frac{\boldsymbol{x} - \boldsymbol{x}'}{|\boldsymbol{x} - \boldsymbol{x}'|^3}\,. \tag{58.2}$$

Charge densities

Definition 58.5 (Charge densities)**.** The *volume charge density* of a distribution is

$$\rho(\boldsymbol{x}') = \frac{\mathrm{d}q'}{\mathrm{d}V'}\,. \tag{58.3}$$

Equivalently one defines the *surface charge density*

$$\sigma(\boldsymbol{x}') = \frac{\mathrm{d}q'}{\mathrm{d}s'} \tag{58.4}$$

on a surface S', and the *line charge density*

$$\lambda(\boldsymbol{x}') = \frac{\mathrm{d}q'}{\mathrm{d}l'} \tag{58.5}$$

on a curve C'.

Substituting Equation (58.3) into Equation (58.2), the force felt by a charge q due to a continuous distribution of density ρ is

$$\boldsymbol{F} = k_E\, q \int_{V'} \rho(\boldsymbol{x}')\, \frac{\boldsymbol{x} - \boldsymbol{x}'}{|\boldsymbol{x} - \boldsymbol{x}'|^3}\, \mathrm{d}V'\,, \tag{58.6}$$

and analogously for the surface and line densities,

$$\boldsymbol{F} = k_E\, q \int_{S'} \sigma(\boldsymbol{x}')\, \frac{\boldsymbol{x} - \boldsymbol{x}'}{|\boldsymbol{x} - \boldsymbol{x}'|^3}\, \mathrm{d}s'\,, \qquad \boldsymbol{F} = k_E\, q \int_{C'} \lambda(\boldsymbol{x}')\, \frac{\boldsymbol{x} - \boldsymbol{x}'}{|\boldsymbol{x} - \boldsymbol{x}'|^3}\, \mathrm{d}l'\,. \tag{58.7}$$

Comparing Equation (58.6) with Equation (58.1), we see that a point charge q located at $\boldsymbol{x}'$ is the particular distribution whose density is

$$\rho(\boldsymbol{x}) = q\, \delta(\boldsymbol{x} - \boldsymbol{x}')\,. \tag{58.8}$$

The electrostatic field

Definition 58.6 (Electrostatic field)**.** The *electrostatic field due to a continuous charge distribution* is the vector field obtained from the force exerted on a test charge q at $\boldsymbol{x}$ in the limit in which q tends to zero — so that the field generated by the test charge itself does not interfere with the field generated by the distribution:

$$\boldsymbol{E}(\boldsymbol{x}) = \lim_{q \to 0} \frac{\boldsymbol{F}}{q}\,. \tag{58.9}$$

From this definition and Equation (58.6), the electrostatic field due to a continuous charge distribution is

$$\boldsymbol{E}(\boldsymbol{x}) = k_E \int_{V'} \rho(\boldsymbol{x}')\, \frac{\boldsymbol{x} - \boldsymbol{x}'}{|\boldsymbol{x} - \boldsymbol{x}'|^3}\, \mathrm{d}V'\,. \tag{58.10}$$

Conversely, consider a charge q at $\boldsymbol{x}$ subject to an external electrostatic field $\boldsymbol{E}$ produced by some distribution of density ρ: the field is given by Equation (58.10) and the force it exerts on q by Equation (58.6). Comparing the two expressions, the *force felt by a charge in an external electrostatic field* is

$$\boldsymbol{F} = q\, \boldsymbol{E}\,. \tag{58.11}$$

The electrostatic potential

Lemma 58.7. *For $\boldsymbol{x} \neq \boldsymbol{x}'$,*

$$\partial_i \left(\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} \right) = -\frac{x_i - x_i'}{|\boldsymbol{x} - \boldsymbol{x}'|^3} \,. \tag{58.12}$$

Derivation pending #8

Gradient identity Equation (58.12): direct differentiation of $|\boldsymbol{x} - \boldsymbol{x}'|^{-1}$ (asserted without proof in the source).

From Equation (58.10) we have, in components,

$$\begin{aligned} E_i(\boldsymbol{x}) &= k_E \int_{V'} \rho(\boldsymbol{x}') \, \frac{x_i - x_i'}{|\boldsymbol{x} - \boldsymbol{x}'|^3} \, \mathrm{d}V' \\ &= -k_E \int_{V'} \rho(\boldsymbol{x}') \, \partial_i \left(\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} \right) \mathrm{d}V' \qquad (58.13) \\ &= -\partial_i \left(k_E \int_{V'} \frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} \, \mathrm{d}V' \right) , \qquad (58.14) \end{aligned}$$

where Equation (58.12) was used in the second step and the derivative, which acts on $\boldsymbol{x}$ only, was taken outside the integral in the third.

Definition 58.8 (Electrostatic potential)**.** The *electrostatic potential* of a continuous charge distribution is

$$\phi(\boldsymbol{x}) = k_E \int_{V'} \frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} \, \mathrm{d}V' \,. \tag{58.15}$$

With this definition, Equation (58.14) reads

$$\boldsymbol{E}(\boldsymbol{x}) = -\boldsymbol{\nabla} \phi(\boldsymbol{x}) \,. \tag{58.16}$$

Gauss's law

Lemma 58.9. *In the sense of distributions,*

$$\partial_i \partial_i \left(\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} \right) = -4\pi \, \delta(\boldsymbol{x} - \boldsymbol{x}') \,. \tag{58.17}$$

Derivation pending #9

Distributional Laplacian of $|\boldsymbol{x} - \boldsymbol{x}'|^{-1}$, Equation (58.17) (asserted without proof in the source).

Theorem 58.10 (Gauss's law)**.** *The divergence of the electrostatic field due to a distribution of charge density ρ is*

$$\boldsymbol{\nabla} \cdot \boldsymbol{E}(\boldsymbol{x}) = 4\pi k_E \, \rho(\boldsymbol{x}) \,. \tag{58.18}$$

Derivation. Applying ∂_i to Equation (58.13),

$$\begin{aligned} \partial_i E_i(\boldsymbol{x}) &= -k_E \int_{V'} \rho(\boldsymbol{x}') \, \partial_i \partial_i \left(\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} \right) \mathrm{d}V' \\ &= 4\pi k_E \int_{V'} \rho(\boldsymbol{x}') \, \delta(\boldsymbol{x} - \boldsymbol{x}') \, \mathrm{d}V' \\ &= 4\pi k_E \, \rho(\boldsymbol{x}) \,, \end{aligned}$$

where Equation (58.17) was used in the second step. □

Integrating Equation (58.18) over an arbitrary volume V enclosing a total charge q_V, and using that by Equation (58.3) the volume integral of the density is the enclosed charge,

$$\int_V \partial_i E_i \, \mathrm{d}V = 4\pi k_E \int_V \rho \, \mathrm{d}V = 4\pi k_E \, q_V \,,$$

and converting the left-hand side into a surface integral by the divergence theorem, we obtain the flux of the electrostatic field through a closed surface:

$$\oint_{\partial V} \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} = 4\pi k_E \, q_V \,, \tag{58.19}$$

a result known as *Gauss's law in integral form.*

Curl of the electrostatic field

Corollary 58.11. *The electrostatic field satisfies*

$$\boldsymbol{\nabla} \times \boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{0} \,. \tag{58.20}$$

Proof. By Equation (58.14) there exists a scalar function such that the electrostatic field can be written as the gradient Equation (58.16); that is, the electrostatic field is a conservative vector field. Since the curl of a gradient vanishes identically, Equation (58.20) follows. □

58.1.2 Geometry of the field

Electrostatic field lines

Definition 58.12 (Field lines)**.** The *electrostatic field lines* are the curves $\boldsymbol{x} = \boldsymbol{x}(\lambda)$ to which the electrostatic field is tangent at every point:

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\lambda}(\lambda) = \alpha \, \boldsymbol{E}\big(\boldsymbol{x}(\lambda)\big) \,.$$

The constant α can be removed by the reparametrization $\lambda' = \alpha\lambda$, since

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\lambda}(\lambda) &= \alpha \, \boldsymbol{E}\big(\boldsymbol{x}(\lambda)\big) \,, \\
\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\lambda'} \frac{\mathrm{d}\lambda'}{\mathrm{d}\lambda} &= \alpha \, \boldsymbol{E}\big(\boldsymbol{x}(\lambda')\big) \,, \\
\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\lambda'} \, \alpha &= \alpha \, \boldsymbol{E}\big(\boldsymbol{x}(\lambda')\big) \,, \\
\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\lambda'} &= \boldsymbol{E}\big(\boldsymbol{x}(\lambda')\big) \,,
\end{aligned}$$

or, renaming the parameter,

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\lambda}(\lambda) = \boldsymbol{E}\big(\boldsymbol{x}(\lambda)\big) \,, \tag{58.21}$$

which is called the *field-line equation.*

Closed line integrals of the electrostatic field

Proposition 58.13. *The line integral of the electrostatic field around any closed curve vanishes:*

$$\oint_{\partial S} \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = 0 \,. \tag{58.22}$$

Derivation. By Stokes's theorem, for an arbitrary compact surface S with closed boundary ∂S,

$$\oint_{\partial S} \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = \int_S (\boldsymbol{\nabla} \times \boldsymbol{E}) \cdot \mathrm{d}\boldsymbol{s}\,,$$

which vanishes by Equation (58.20). □

Proposition 58.14. *There are no closed electrostatic field lines.*

Derivation. Suppose there were a closed field line C with parameter λ. Then, using the field-line equation Equation (58.21),

$$\begin{aligned}\oint_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{x} &= \oint_C \boldsymbol{E}(\boldsymbol{x}(\lambda)) \cdot \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\lambda}\,\mathrm{d}\lambda \\ &= \oint_C \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\lambda} \cdot \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\lambda}\,\mathrm{d}\lambda \\ &= \oint_C \left|\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\lambda}\right|^2 \mathrm{d}\lambda \neq 0\,,\end{aligned}$$

since along an actual curve

$$\left|\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\lambda}\right|^2 > 0\,.$$

This contradicts Proposition 58.13, by which every closed line integral of the electrostatic field vanishes. Hence no closed field line exists. □

Equipotential surfaces

Definition 58.15 (Equipotential surface)**.** Consider a charge distribution generating an electrostatic potential ϕ. An *equipotential surface* is a surface given by the equation

$$\phi(\boldsymbol{x}) = C \tag{58.23}$$

with C constant.

58.2 Methods for obtaining the potential

58.2.1 Boundary conditions

Line integral of the electrostatic field

Consider two points P_i and P_f joined by a curve C described by $\boldsymbol{x}$. From Equation (58.16), integrating along C,

$$\int_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{x} = -\int_C \boldsymbol{\nabla}\phi \cdot \mathrm{d}\boldsymbol{x} = -\int_C \mathrm{d}\phi\,,$$

so that

$$\phi(P_f) - \phi(P_i) = -\int_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{x}\,.$$

Taking P_i to be the point of coordinates $\boldsymbol{x}$ and P_f a point located at $\boldsymbol{x}_0$, we obtain the following expression for the potential:

$$\phi(\boldsymbol{x}) = \phi(\boldsymbol{x}_0) + \int_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l}\,. \tag{58.24}$$

Continuity condition; interface conditions; ideal conductors

[Headings reserved in the source outline.] The worked examples below make use of the standard boundary conditions of electrostatics [1976]: the potential is continuous across any interface; the tangential component of $\boldsymbol{E}$ is continuous across an interface; the jump of the normal component of the field across a charged surface is fixed by the surface density, so that on the surface of an ideal conductor (in SI units)

$$\sigma = \epsilon_0 \, \boldsymbol{E} \cdot \hat{\boldsymbol{n}}\big|_{\partial V} \,, \tag{58.25}$$

with $\hat{\boldsymbol{n}}$ the outward normal of the conductor; the potential is constant throughout an ideal conductor, and vanishes on a grounded one.

Derivation pending #10

Continuity and interface conditions of the electrostatic field and the ideal-conductor boundary conditions, including Equation (58.25): the source reserves these headings without content, while its worked problems use the results.

58.2.2 The method of images

[Heading reserved in the source outline.] The method replaces a boundary-value problem in the presence of conductors by an equivalent charge configuration — the original sources plus fictitious *image* charges placed outside the physical region — chosen so that the boundary conditions of Section 58.2.1 are satisfied. The source develops the method only through the following worked problems, both of which rely on the ideal-dipole potential recorded in Section 58.4.4.

Example 58.16 (Ideal dipole inclined at 45° before a grounded conducting plane)**.** An ideal dipole of dipole moment $\boldsymbol{p}$ is located at a distance d from an infinite conducting plane connected to ground, and forms an angle of 45° with the normal to the plane. Compute the electrostatic potential at every point and the surface charge density induced on the plane.

The potential of an ideal dipole located at $\boldsymbol{x}'$ is given by

$$\phi(\boldsymbol{x}) = \frac{1}{4\pi\epsilon_0} \, \frac{\boldsymbol{x} - \boldsymbol{x}'}{|\boldsymbol{x} - \boldsymbol{x}'|^3} \cdot \boldsymbol{p} \,.$$

Using the method of images, we place an image dipole $\boldsymbol{p}_I$ on the other side of the plane, at a distance d_I; by the symmetry of the problem we may take $|\boldsymbol{p}_I| = |\boldsymbol{p}|$. The potential at every point is then

$$\phi(\boldsymbol{x}) = \frac{1}{4\pi\epsilon_0} \left(\frac{\boldsymbol{x} - \boldsymbol{x}'}{|\boldsymbol{x} - \boldsymbol{x}'|^3} \cdot \boldsymbol{p} + \frac{\boldsymbol{x} - \boldsymbol{x}_I}{|\boldsymbol{x} - \boldsymbol{x}_I|^3} \cdot \boldsymbol{p}_I \right) .$$

Placing the origin at the point of the plane closest to the dipole, with $\hat{\boldsymbol{x}}$ along the normal,

$$\begin{aligned}
\boldsymbol{x} &= x\,\hat{\boldsymbol{x}} + y\,\hat{\boldsymbol{y}} + z\,\hat{\boldsymbol{z}} \,, \\
\boldsymbol{x}' &= d\,\hat{\boldsymbol{x}} \,, \qquad \boldsymbol{x}_I = d_I\,\hat{\boldsymbol{x}} \,, \\
\boldsymbol{p} &= p\,(\cos 45^\circ\,\hat{\boldsymbol{x}} + \sin 45^\circ\,\hat{\boldsymbol{y}}) = \frac{p}{\sqrt{2}}\,(\hat{\boldsymbol{x}} + \hat{\boldsymbol{y}}) \,, \\
\boldsymbol{p}_I &= p\,(\cos 45^\circ\,\hat{\boldsymbol{x}} - \sin 45^\circ\,\hat{\boldsymbol{y}}) = \frac{p}{\sqrt{2}}\,(\hat{\boldsymbol{x}} - \hat{\boldsymbol{y}}) \,.
\end{aligned}$$

We must now impose the boundary conditions of the problem to determine the distance d_I. Since the conducting plane is grounded, the potential must vanish on it; evaluating at the origin,

$$\frac{1}{4\pi\epsilon_0}\frac{p}{\sqrt{2}}\left(\frac{x-d+y}{((x-d)^2+y^2+z^2)^{3/2}}+\frac{x-d_I-y}{((x-d_I)^2+y^2+z^2)^{3/2}}\right)\Bigg|_{\boldsymbol{x}=\boldsymbol{0}}=0\,,$$
$$\frac{-d}{(d^2)^{3/2}}+\frac{-d_I}{(d_I^2)^{3/2}}=0\,,$$

so that

$$d_I=-d\,,\qquad \boldsymbol{x}_I=-d\,\hat{\boldsymbol{x}}\,.$$

Therefore

$$\phi(\boldsymbol{x})=\frac{p}{4\sqrt{2}\,\pi\epsilon_0}\left(\frac{x-d+y}{((x-d)^2+y^2+z^2)^{3/2}}+\frac{x+d-y}{((x+d)^2+y^2+z^2)^{3/2}}\right)\,, \tag{58.26}$$

and with these values the two numerators cancel on the whole plane $x=0$, so the boundary condition is satisfied everywhere on the conductor.

With the potential in hand we can compute the electric field. In particular,

$$\begin{aligned}E_x&=-\frac{\partial\phi}{\partial x}\\&=-\frac{p}{4\sqrt{2}\,\pi\epsilon_0}\Big(\left((x-d)^2+y^2+z^2\right)^{-3/2}-3(x-d)(x-d+y)\left((x-d)^2+y^2+z^2\right)^{-5/2}\\&\quad+\left((x+d)^2+y^2+z^2\right)^{-3/2}-3(x+d)(x+d-y)\left((x+d)^2+y^2+z^2\right)^{-5/2}\Big)\,.\end{aligned}$$

Finally, by the interface condition Equation (58.25) with $\hat{\boldsymbol{n}}=\hat{\boldsymbol{x}}$, the surface charge density induced on the plane is

$$\begin{aligned}\sigma(y,z)&=\epsilon_0\,\boldsymbol{E}\cdot\hat{\boldsymbol{n}}|_{\partial V}=\epsilon_0\,E_x(0,y,z)\\&=-\frac{p}{4\sqrt{2}\,\pi}\Big(\left(d^2+y^2+z^2\right)^{-3/2}+3d\,(y-d)\left(d^2+y^2+z^2\right)^{-5/2}\\&\quad+\left(d^2+y^2+z^2\right)^{-3/2}-3d\,(d-y)\left(d^2+y^2+z^2\right)^{-5/2}\Big)\\&=\frac{p}{2\sqrt{2}\,\pi}\left(3d\,(d-y)\left(d^2+y^2+z^2\right)^{-5/2}-\left(d^2+y^2+z^2\right)^{-3/2}\right)\,.\end{aligned}$$

Example 58.17 (Ideal dipole normal to a grounded conducting plane)**.** An ideal dipole of dipole moment $\boldsymbol{p}$ is oriented along the normal at a distance d from an infinite conducting plane connected to ground. Compute the potential at every point and the force exerted on the dipole.

The image construction is the same as in Example 58.16: an image dipole $\boldsymbol{p}_I$ with $|\boldsymbol{p}_I|=|\boldsymbol{p}|$ is placed at $\boldsymbol{x}_I=d_I\,\hat{\boldsymbol{x}}$, and now

$$\boldsymbol{x}'=d\,\hat{\boldsymbol{x}}\,,\qquad \boldsymbol{p}=p\,\hat{\boldsymbol{x}}\,,\qquad \boldsymbol{p}_I=p\,\hat{\boldsymbol{x}}\,.$$

Imposing that the potential vanish on the grounded plane — evaluating at the origin as before —

$$\frac{-d}{(d^2)^{3/2}}+\frac{-d_I}{(d_I^2)^{3/2}}=0\quad\Longrightarrow\quad d_I=-d\,,\qquad \boldsymbol{x}_I=-d\,\hat{\boldsymbol{x}}\,.$$

Therefore

$$\phi(\boldsymbol{x})=\frac{p}{4\pi\epsilon_0}\left(\frac{x-d}{((x-d)^2+y^2+z^2)^{3/2}}+\frac{x+d}{((x+d)^2+y^2+z^2)^{3/2}}\right)\,, \tag{58.27}$$

which vanishes on the plane $x = 0$, as required.

The force felt by an ideal dipole $\boldsymbol{p}$ in an external field $\boldsymbol{E}$ is given by

$$\boldsymbol{F} = \boldsymbol{\nabla}\,(\boldsymbol{p}\cdot\boldsymbol{E})$$

(cf. Section 58.4.4). In this case the external field is the one generated by the plane — equivalently, by the image dipole,

$$\boldsymbol{E}(\boldsymbol{x}) = \frac{1}{4\pi\epsilon_0}\,\frac{3(\boldsymbol{p}_I\cdot\hat{\boldsymbol{r}})\hat{\boldsymbol{r}} - \boldsymbol{p}_I}{r^3}\,,$$

with r measured from the image dipole. The dipole sits at distance $r = 2d$ from its image, on the axis $\theta = 0$ (spherical coordinates centred on the image, polar axis $\hat{\boldsymbol{x}}$). Thus, with $\boldsymbol{p}_I = \boldsymbol{p}$,

$$\begin{aligned}
\boldsymbol{F} &= \boldsymbol{\nabla}\left(\boldsymbol{p}\cdot\frac{1}{4\pi\epsilon_0}\,\frac{3(\boldsymbol{p}\cdot\hat{\boldsymbol{r}})\hat{\boldsymbol{r}} - \boldsymbol{p}}{r^3}\right)\Bigg|_{r=2d,\;\theta=0} \\
&= \frac{1}{4\pi\epsilon_0}\,\boldsymbol{\nabla}\left(\frac{3(\boldsymbol{p}\cdot\hat{\boldsymbol{r}})^2 - p^2}{r^3}\right)\Bigg|_{r=2d,\;\theta=0} \\
&= \frac{p^2}{4\pi\epsilon_0}\,\boldsymbol{\nabla}\left(3\,\frac{\cos^2\theta}{r^3} - \frac{1}{r^3}\right)\Bigg|_{r=2d,\;\theta=0} \\
&= \frac{p^2}{4\pi\epsilon_0}\left(3\left(-3\,\frac{\cos^2\theta}{r^4}\right) + \frac{3}{r^4}\right)\Bigg|_{r=2d,\;\theta=0}\hat{\boldsymbol{r}} \\
&= -\frac{1}{4\pi\epsilon_0}\,\frac{6p^2}{(2d)^4}\,\hat{\boldsymbol{x}} = -\frac{3p^2}{32\pi\epsilon_0\,d^4}\,\hat{\boldsymbol{x}}\,,
\end{aligned}$$

where the angular part of the gradient drops out because $\mathrm{d}(\cos^2\theta)/\mathrm{d}\theta = -2\cos\theta\sin\theta$ vanishes at $\theta = 0$. The force is attractive: the dipole is pulled toward the plane.

58.2.3 The Poisson equation

Poisson equation

From Equation (58.16), taking the divergence of both sides and using Gauss's law Equation (58.18),

$$\begin{aligned}
\boldsymbol{E} &= -\boldsymbol{\nabla}\phi\,, && \text{apply } \boldsymbol{\nabla}\cdot(\ )\,, \\
4\pi k_E\,\rho &= -\nabla^2\phi\,,
\end{aligned}$$

so that

$$\nabla^2\phi(\boldsymbol{x}) = -4\pi k_E\,\rho(\boldsymbol{x})\,, \tag{58.28}$$

a result known as the *Poisson equation*.

Laplace equation

In a region where the charge density vanishes, Equation (58.28) reduces to

$$\nabla^2\phi(\boldsymbol{x}) = 0\,, \tag{58.29}$$

known as the *Laplace equation*.

58.2.4 Green functions

[Headings reserved in the source outline: the source function, the general solution of the Poisson equation, and the Dirichlet and Neumann boundary conditions are pending port from a completed source; see [1976] for the standard treatment.]

58.2.5 Multipole expansion

[Headings reserved in the source outline: the expansion of the potential in inverse powers of the distance, the multipole moments of order n, and the ideal-dipole limit are pending port from a completed source. The ideal-dipole formulas used by the worked examples are recorded, as pending statements, in Section 58.4.4.]

58.3 Electrostatic potential energy

[Heading reserved in the source outline; no content to port. The energy functional of a linear dielectric medium, used by the worked examples of Chapter 62, is recorded there.]

58.4 Charge distributions in external fields

58.4.1 Potential energy of a charge distribution in an external field

[Heading reserved in the source outline; no content to port.]

58.4.2 Force on a charge distribution in an external field

[Heading reserved in the source outline; no content to port.]

58.4.3 Torque on a charge distribution in an external field

[Heading reserved in the source outline; no content to port.]

58.4.4 The ideal dipole

The worked examples of Section 58.2.2 employ the following standard results for an ideal (point) dipole of moment $\boldsymbol{p}$ [1976, 2313]. The potential of an ideal dipole located at $\boldsymbol{x}'$ is

$$\phi(\boldsymbol{x}) = k_E \, \frac{(\boldsymbol{x} - \boldsymbol{x}') \cdot \boldsymbol{p}}{|\boldsymbol{x} - \boldsymbol{x}'|^3} \, , \tag{58.30}$$

and the corresponding field, at position $r\hat{\boldsymbol{r}}$ relative to the dipole, is

$$\boldsymbol{E}(\boldsymbol{x}) = k_E \, \frac{3\,(\boldsymbol{p} \cdot \hat{\boldsymbol{r}})\,\hat{\boldsymbol{r}} - \boldsymbol{p}}{r^3} \, . \tag{58.31}$$

In an external electrostatic field $\boldsymbol{E}$, the force felt by an ideal dipole is

$$\boldsymbol{F} = \boldsymbol{\nabla}\,(\boldsymbol{p} \cdot \boldsymbol{E}) \, . \tag{58.32}$$

Derivation pending #11

Ideal-dipole potential Equation (58.30), field Equation (58.31) and force Equation (58.32): to be derived from the multipole expansion of Section 58.2.5 and from the energy of a distribution in an external field, both of which the source reserves without content.

58.5 Electrostatics in macroscopic media

[Heading reserved in the source outline.] The macroscopic model of electrostatics in material media — polarization, the displacement field, and the linear dielectric constitutive relation — together with the two dielectric worked problems of the companion collection, is ported to Chapter 62.

Chapter 59

Magnetostatics

Magnetostatics is the theory of the fields produced by steady electric currents: currents themselves, the Biot–Savart law, the vector potential, and magnetic media. It stands to Chapter 58 as the second static limit of the full Maxwell system assembled in Chapter 60, and it is where electromagnetism became a single subject: before 1820 electricity and magnetism were separate sciences, and this chapter's first task is the experiment that joined them. Its second task is a structural asymmetry with electrostatics that has survived every later refinement of the theory: there are electric charges, and there are—as far as every search to date can tell—no magnetic ones. That absence is not an assumption but an experimental verdict, and it is treated as such in Section 59.6.

Derivation pending #12

Magnetostatics: all derivations of this chapter are pending.

59.1 Steady currents

59.1.1 Current density and charge conservation

[Reserved: the current density $\boldsymbol{J}$, the current through a surface, the continuity equation $\partial\rho/\partial t + \boldsymbol{\nabla}\cdot\boldsymbol{J} = 0$, and the magnetostatic regime $\boldsymbol{\nabla}\cdot\boldsymbol{J} = 0$ as its steady limit; surface and line currents; the standard treatment is [1976]. The dynamical role of the continuity equation—it is what forces the displacement current—is taken up in Chapter 60.]

59.1.2 The SI unit of current

[Reserved: the ampere, historically defined through the force per unit length between parallel currents and since 2019 fixed by the elementary charge [2333]; how current is realised and measured, with the metrological framework of Chapter 2.]

59.2 The discovery of electromagnetism

59.2.1 Magnetism before 1820

[Reserved: lodestones, compass needles, and Gilbert's De Magnete*, which established the Earth itself as a magnet and separated magnetic from electric attraction [1]; the phenomenology of poles, the impossibility of isolating one by cutting a magnet, and the inverse-square law between pole strengths, put on an absolute quantitative footing by Gauss's measurement of the terrestrial field [106].]*

Phenomenon 59.1 (The inseparability of magnetic poles)**.** Every magnet has two poles of opposite character, and no operation performed upon it isolates one of them. A bar magnet broken in two yields two complete magnets, each with its own north and south pole, and the subdivision may be repeated to the limit of what can be handled without ever producing a body that acts upon a compass needle as a single pole would. The Earth itself behaves as one such magnet, which is why the needle points at all [1], and the strength of its field was reduced to absolute measure — to mechanical units, and so ultimately to the SI — by Gauss [106].

Derivation pending #13

Inseparability of magnetic poles: derive, from the Amperian representation of a magnetized body by bound currents, that every subdivision of such a body again carries a complete circulating current and is therefore again a dipole, and connect the statement to the vanishing divergence of the magnetic field.

59.2.2 Ørsted's experiment

[Reserved: the deflection of a compass needle by a nearby current-carrying wire, announced by Ørsted in July 1820 [68]—the first observed connection between electricity and magnetism, and the observation from which the whole of electrodynamics descends. The circular geometry of the effect (the needle sets tangentially around the wire, not toward it) already encodes the curl structure of the field equations. Full experimental detail belongs to Chapter 63.]

Phenomenon 59.2 (The deflection of the needle by a current)**.** A magnetic needle placed near a straight conducting wire is deflected the instant a galvanic current is sent through the wire, and it comes to rest *across* the wire rather than along the line joining the two: the needle sets tangentially to the circle that passes through it and has the wire for its axis. The sense of the deflection reverses with the sense of the current, and reverses again when the needle is carried to the opposite side of the wire [68]. This is the first observed action of electricity upon magnetism, and the first force in physics whose direction is not along the line joining its sources.

Derivation. The derivation runs opposite to the history: it assumes the law for which this observation was the first evidence, namely the Biot–Savart law of Section 59.2.3, by which an element $I\,\mathrm{d}\boldsymbol{l}$ of a steady current contributes

$$\mathrm{d}\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0 I}{4\pi}\,\frac{\mathrm{d}\boldsymbol{l}\times\hat{\boldsymbol{r}}}{r^2}\,, \tag{59.1}$$

where $\boldsymbol{r}$ runs from the element to the field point.

Take the wire along the z-axis and the field point at perpendicular distance s in the plane $z=0$. Each element $\mathrm{d}\boldsymbol{l} = \mathrm{d}z\,\hat{\boldsymbol{z}}$ contributes along $\hat{\boldsymbol{z}}\times\hat{\boldsymbol{r}}$, which is $\hat{\boldsymbol{\varphi}}$ for every z; the contributions therefore do not compete in direction and the total field is purely azimuthal. Writing $z = s\tan\alpha$, so that $r = s\sec\alpha$ and $\mathrm{d}z = s\sec^2\alpha\,\mathrm{d}\alpha$, and noting that the sine of the angle between $\mathrm{d}\boldsymbol{l}$ and $\hat{\boldsymbol{r}}$ is $\cos\alpha$,

$$B_\varphi(s) = \frac{\mu_0 I}{4\pi}\int_{-\pi/2}^{\pi/2}\frac{s\sec^2\alpha\,\cos\alpha}{s^2\sec^2\alpha}\,\mathrm{d}\alpha = \frac{\mu_0 I}{4\pi s}\int_{-\pi/2}^{\pi/2}\cos\alpha\,\mathrm{d}\alpha = \frac{\mu_0 I}{2\pi s}\,. \tag{59.2}$$

The field circles the wire. A compass needle is a magnetic dipole $\boldsymbol{m}$ and feels the torque $\boldsymbol{m}\times\boldsymbol{B}$ (Section 59.4.4), which vanishes only when $\boldsymbol{m}$ lies along $\boldsymbol{B}$; the needle therefore comes to rest tangentially to that circle, reverses with I, and reverses again on the far side of the wire, where $\hat{\boldsymbol{\varphi}}$ points the other way. □

59.2.3 The Biot–Savart law

[Reserved: Biot and Savart's quantitative measurement, within months of Ørsted's announcement, of the force exerted by a long straight current on a magnetic pole, falling off as the inverse distance [67]; the modern differential statement $\mathrm{d}\boldsymbol{B} \propto I\,\mathrm{d}\boldsymbol{l} \times \hat{\boldsymbol{r}}/r^2$ and its integral form for closed circuits; worked fields of the straight wire, the circular loop and the solenoid [1976].]

Phenomenon 59.3 (The field of a long straight current)**.** The magnetic action of a long straight steady current, measured by Biot and Savart within weeks of Ørsted's announcement from the oscillation period of a short magnetic needle suspended at a series of distances from a vertical conducting wire, is proportional to the current I and falls off as the inverse *first* power of the perpendicular distance s from the wire, not as the inverse square [67]:

$$B_\varphi = \frac{\mu_0 I}{2\pi s}\,. \tag{59.3}$$

Since the SI revision of 2019 the magnetic constant μ_0 appearing here is a measured quantity rather than a defined one [2333].

Derivation. This is Equation (59.2), already obtained from the Biot–Savart law Equation (59.1) for exactly this configuration. The inverse first power, against the inverse square of each separate element's contribution, is the signature of a line source of unbounded length: the length of wire lying within a given range of the angle α grows in proportion to s, which cancels one power of the distance. □

59.2.4 Ampère's electrodynamics

[Reserved: Ampère's demonstration that two currents attract or repel one another, his force law between current elements, and the theorem that a closed current loop is magnetically equivalent to a sheet of magnetic dipoles, all deduced from his null and balance experiments in the Théorie des phénomènes électro-dynamiques *[77]; the hypothesis of molecular currents as the origin of all magnetism, vindicated a century later in Chapter 129.]*

Phenomenon 59.4 (Force between parallel currents)**.** Two long straight parallel wires carrying steady currents act on one another mechanically, with no magnets present anywhere: they attract when the currents run the same way and repel when they run opposite ways, and the force per unit length is proportional to each current and inversely proportional to their separation d [66, 77],

$$\frac{F}{L} = \frac{\mu_0 I_1 I_2}{2\pi d}\,. \tag{59.4}$$

For $I_1 = I_2 = 1\,\mathrm{A}$ at $d = 1\,\mathrm{m}$ this is $2 \times 10^{-7}\,\mathrm{N/m}$, the number through which the ampere was defined until the SI revision of 2019 replaced that definition by a fixed value of the elementary charge [2333].

Derivation. Wire 1 produces at the location of wire 2 the azimuthal field Equation (59.2) with $s = d$, of magnitude $B_1 = \mu_0 I_1/2\pi d$ and directed perpendicular both to wire 1 and to the line joining the wires. A length L of wire 2, carrying I_2 along the unit vector $\hat{\boldsymbol{l}}_2$, feels the Ampère force $\boldsymbol{F} = I_2 L\,\hat{\boldsymbol{l}}_2 \times \boldsymbol{B}_1$ (Section 59.4.4). Since $\hat{\boldsymbol{l}}_2$ is parallel to the wires and $\boldsymbol{B}_1$ is perpendicular to them, the magnitudes multiply without an angular factor and $F/L = I_2 B_1 = \mu_0 I_1 I_2/2\pi d$, which is Equation (59.4). Taking $\hat{\boldsymbol{l}}_2$ parallel to $\hat{\boldsymbol{l}}_1$ makes $\hat{\boldsymbol{l}}_2 \times \boldsymbol{B}_1$ point from wire 2 towards wire 1 — attraction — and reversing either current reverses it. With $\mu_0 = 4\pi \times 10^{-7}\,\mathrm{N/A^2}$, as it was by definition before 2019, the numerical value quoted follows by substitution. □

59.3 The field equations of magnetostatics

59.3.1 Gauss's law for the magnetic field

[Reserved: $\nabla \cdot \boldsymbol{B} = 0$ and its integral form, the vanishing of the magnetic flux through every closed surface [185]; its reading as the statement that magnetic field lines do not end, and its empirical status as the absence of magnetic charge, examined against experiment in Section 59.6.]

Phenomenon 59.5 (The magnetic flux through a closed surface vanishes)**.** Every measurement of magnetic flux ever performed, over any closed surface and about any source — a current-carrying circuit, a permanent magnet, the Earth, a star — has given zero net flux:

$$\oint_{\partial V} \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = 0 \qquad \text{for every closed surface } \partial V\,, \tag{59.5}$$

equivalently $\nabla \cdot \boldsymbol{B} = 0$ [185]. The statement is not a mathematical necessity but an experimental one: it says that magnetic charge has never been found (Section 59.6.2, [2427]), and it is the one Maxwell equation whose empirical support is a null record rather than a positive measurement.

Derivation. For fields produced by steady currents the vanishing follows from the Biot–Savart law. Since $(\boldsymbol{x}-\boldsymbol{x}')/|\boldsymbol{x}-\boldsymbol{x}'|^3 = -\nabla\,(1/|\boldsymbol{x}-\boldsymbol{x}'|)$, where ∇ acts on the field point, and since $\boldsymbol{J}(\boldsymbol{x}')$ does not depend on the field point,

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,\mathrm{d}V' = \nabla \times \left[\frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,\mathrm{d}V'\right], \tag{59.6}$$

using $\nabla\times(f\boldsymbol{J}) = (\nabla f)\times\boldsymbol{J}$ for constant $\boldsymbol{J}$. A curl is divergence-free, so $\nabla\cdot\boldsymbol{B} = 0$ identically, and the divergence theorem turns that into Equation (59.5). What the derivation establishes is conditional: *if* every magnetic field has currents for its only source, the flux vanishes. The unconditional statement is the experimental one, and the bracketed expression is the vector potential of Section 59.4. □

59.3.2 Ampère's circuital law

[Reserved: $\nabla\times\boldsymbol{B} = \mu_0\boldsymbol{J}$ in the steady regime and the integral statement that the circulation of $\boldsymbol{B}$ around a closed curve measures the enclosed current [77, 185]; its consistency requirement $\nabla\cdot\boldsymbol{J} = 0$, whose failure for open circuits is the door through which the displacement current of Chapter 60 enters; applications to the solenoid and the torus.]

59.3.3 Boundary conditions

[Reserved: continuity of the normal component of $\boldsymbol{B}$ and the jump of its tangential component across a surface current [1976], in parallel with the electrostatic interface conditions of Chapter 58.]

59.4 The vector potential

59.4.1 Existence and gauge freedom

[Reserved: $\nabla\cdot\boldsymbol{B} = 0$ implies $\boldsymbol{B} = \nabla\times\boldsymbol{A}$; the vector potential introduced by Neumann in his general law of induced currents [115]; the gauge ambiguity $\boldsymbol{A}\to\boldsymbol{A}+\nabla\chi$ and the Coulomb gauge $\nabla\cdot\boldsymbol{A} = 0$; the systematic treatment of gauge freedom is deferred to Chapter 60.]

59.4.2 The Biot–Savart law as the general solution

[Reserved: in Coulomb gauge each component of $\boldsymbol{A}$ obeys a Poisson equation, solved by the same Green function as the electrostatic potential of Chapter 58 (general theory in Chapter 10); recovering the Biot–Savart field [67, 1976] by taking the curl.]

59.4.3 Multipole expansion and the magnetic dipole

[Reserved: the expansion of $\boldsymbol{A}$ for a localised current distribution; the vanishing of the monopole term as the first appearance of Section 59.3.1 in an expansion; the magnetic dipole moment $\boldsymbol{m} = \frac{1}{2}\int \boldsymbol{x}' \times \boldsymbol{J}\,\mathrm{d}V'$ and its field [1976]; the dipole–dipole interaction and the field of the Earth as leading-order dipole [106].]

Phenomenon 59.6 (Every magnetostatic source is a dipole to leading order)**.** Seen from far enough away, no steady current distribution and no magnetized body produces a field falling as $1/r^2$, the magnetic counterpart of the field of a point charge. The leading behaviour is always the $1/r^3$ field of a dipole. The Earth is the case measured first and best: it acts on a compass as a magnet [1], its field strength was reduced to absolute mechanical units by Gauss [106], and to leading order in the distance that field is dipolar, with the higher multipoles appearing as corrections.

Derivation. Expand the vector potential of Equation (59.6) for a current distribution confined to a bounded region, at a field point far outside it. The first term of $1/|\boldsymbol{x}-\boldsymbol{x}'|$ is $1/r$ with $r = |\boldsymbol{x}|$, and it contributes

$$\boldsymbol{A}_{\mathrm{mono}}(\boldsymbol{x}) = \frac{\mu_0}{4\pi r}\int \boldsymbol{J}(\boldsymbol{x}')\,\mathrm{d}V' \,. \tag{59.7}$$

For a steady current this integral vanishes. Indeed $\partial'_j\,(x'_i J_j) = J_i + x'_i\,\boldsymbol{\nabla}'\cdot\boldsymbol{J} = J_i$, because $\boldsymbol{\nabla}\cdot\boldsymbol{J} = 0$ in the magnetostatic regime (Section 59.1.1); integrating over a volume that encloses all the current, where $\boldsymbol{J}$ vanishes on the boundary, the left-hand side integrates to zero by the divergence theorem, so

$$\int J_i\,\mathrm{d}V' = 0\,, \qquad i = 1,2,3\,. \tag{59.8}$$

The monopole term Equation (59.7) therefore disappears identically, the expansion begins at the next order, and that order is the dipole $\boldsymbol{m} = \frac{1}{2}\int \boldsymbol{x}'\times\boldsymbol{J}\,\mathrm{d}V'$ whose potential falls as $1/r^2$ and whose field falls as $1/r^3$. This is the same absence, read in an expansion, that Phenomenon 59.5 reads as a flux. □

59.4.4 Force and torque on a dipole

[Reserved: force $\boldsymbol{\nabla}(\boldsymbol{m}\cdot\boldsymbol{B})$ and torque $\boldsymbol{m}\times\boldsymbol{B}$ on a magnetic dipole in an external field; Ampère's equivalence of the current loop and the dipole sheet [77]; the contrast with the electric dipole of Chapter 58, and the relativistic reading of the magnetic force developed in Chapter 40.]

59.5 Magnetostatics in matter

59.5.1 Magnetization and bound currents

[Reserved: the magnetization field $\boldsymbol{M}$ as dipole moment density; the equivalent bound currents $\boldsymbol{J}_b = \boldsymbol{\nabla}\times\boldsymbol{M}$ in the volume and $\boldsymbol{M}\times\hat{\boldsymbol{n}}$ on the surface, following the macroscopic construction of Maxwell's Treatise *[185]; the microscopic averaging that justifies it is carried out in Chapter 62.]*

59.5.2 The auxiliary field $\boldsymbol{H}$

[Reserved: $\boldsymbol{H} = \boldsymbol{B}/\mu_0 - \boldsymbol{M}$ and the circuital law $\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{J}_{free}$ sourced by free currents alone [185]; interface conditions for $\boldsymbol{H}$; linear media, permeability, and the magnetic scalar potential in current-free regions [1976].]

59.5.3 Permanent magnets and boundary-value problems

[Reserved: the uniformly magnetized sphere and bar magnet; hysteresis named as the phenomenon that makes "the" permeability of a ferromagnet ill-defined, with its phenomenology and microscopic origin deferred to Chapter 62 and Chapter 129; magnetic circuits and shielding [1976].]

59.6 Magnetic monopoles: the empirical status

59.6.1 The Dirac quantization argument

[Reserved: Dirac's demonstration that a single magnetic charge g anywhere in the universe forces electric charge to be quantized, in units inversely proportional to g [677]—the one theoretical result that keeps the monopole question open, since charge quantization is observed; the argument requires the quantum mechanics of Part IX and is stated here only for its bearing on Section 59.3.1.]

59.6.2 Searches and the null record

[Reserved: the honest record of monopole searches. Cabrera's superconducting induction detector registered on 14 February 1982 a single candidate event of exactly one Dirac charge [1685]; the event was never repeated, and larger-exposure induction and track-etch searches, accelerator searches, and cosmic-ray searches have all been null—the current limits are collected in the Review of Particle Physics [2427]. Gauss's law for $\boldsymbol{B}$ therefore stands as an empirical law: one unrepeated candidate against decades of null exposure, stated as such and not better.]

Phenomenon 59.7 (The null record of the monopole searches)**.** No magnetic charge has been observed. A superconducting induction detector registered, on 14 February 1982, a single flux jump of the magnitude a monopole of one Dirac charge would produce [1685]; the event was never repeated, in that apparatus or in any of the far larger induction, track-etch, accelerator and cosmic-ray exposures that followed, and every such search has returned a limit rather than a signal [2427]. The record is therefore one unrepeated candidate against decades of null exposure, and this is the whole of the evidence on which Phenomenon 59.5 rests. The question is kept open, rather than closed, by Dirac's demonstration that a single magnetic charge anywhere would force the quantization of electric charge [677] — and electric charge *is* observed to be quantized.

Derivation pending #14

Monopole searches: derive the flux quantum that an induction detector would register for the passage of one Dirac charge through a superconducting loop, and the conversion of a null exposure into a flux limit per unit area, time and solid angle. The Dirac quantization argument itself requires the quantum mechanics of Part IX and is stated, not derived, here.

Chapter 60

The Maxwell Equations

This chapter unifies electrodynamics: electromagnetic induction, the displacement current, the full Maxwell system, its conservation laws, and the potentials through which it is most efficiently formulated (the Lagrangian formulation itself is developed in Chapter 64). The statics of Chapter 58 and Chapter 59 treated $\boldsymbol{E}$ and $\boldsymbol{B}$ as separate fields with separate sources; here the time derivatives couple them. Faraday's discovery that a changing magnetic flux drives a current supplies one coupling from experiment; Maxwell's displacement current supplies the other from a consistency argument—and turns the system into one that propagates waves at a speed fixed by two electrical constants. That the speed came out equal to the measured speed of light is the single most consequential coincidence in classical physics, and its experimental confirmation by Hertz is treated in Chapter 63.

Derivation pending #15

The Maxwell Equations: all derivations of this chapter are pending.

60.1 Electromagnetic induction

60.1.1 Faraday's discovery

[Reserved: the induction of transient currents by switching a neighbouring circuit, by moving a magnet, and by moving the circuit itself, reported in the First Series of Faraday's Experimental Researches in Electricity *[89]; Henry's independent observation of induction and of self-induction [91]; electromotive force and the flux rule; full experimental detail in Chapter 63.]*

Phenomenon 60.1 (Electromagnetic induction)**.** A current flows in a closed conducting circuit whenever — and only while — the magnetic flux threading it changes. A steady current in a neighbouring circuit produces nothing; making or breaking that current produces a transient. So does moving a magnet through the circuit, and so does moving the circuit itself while the magnet stands still [89, 91]. The electromotive force driving the current is the rate at which the flux changes,

$$\mathcal{E} = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t}\,, \qquad \Phi_B = \int_S \boldsymbol{B}\cdot\mathrm{d}\boldsymbol{S}\,, \tag{60.1}$$

with S any surface bounded by the circuit, and it does not matter which of the three ways the change was brought about.

Derivation pending #16

The flux rule: derive it from the Maxwell–Faraday equation together with the Lorentz force acting on the charge carriers of a moving circuit, using the transport theorem for a surface whose boundary moves with the circuit, and show that the two physically distinct mechanisms — an induced electric field in the transformer case, a magnetic force on carried charge in the motional case — combine into the single total rate of change of the flux.

60.1.2 Lenz's law

[Reserved: the sign of the induced current opposes the change of flux that produces it [97]; its reading as a stability statement and its connection with energy conservation, made precise once Poynting's theorem is available in Section 60.4; eddy currents as its macroscopic manifestation.]

Phenomenon 60.2 (Lenz's rule for the sign)**.** The induced current always circulates in the sense whose own magnetic flux *opposes* the change that produced it [97]. A magnet pushed towards a closed loop is resisted; the same magnet withdrawn is held back; a sheet of copper swung between the poles of a magnet is braked as though moving through a fluid, though nothing touches it. The rule fixes the sign in Equation (60.1) and is what makes that minus sign an experimental statement rather than a convention.

Derivation. Suppose the opposite sign, and let a magnet approach a loop of resistance R. The induced current would then produce a flux reinforcing the change, so the force it exerts on the magnet would act *along* the magnet's motion. The magnet would accelerate; the flux would change faster; the electromotive force and the current would grow with it; and the loop would dissipate $\mathcal{E}^2/R$ at an increasing rate while the magnet's kinetic energy also increased. Both the mechanical energy and the Joule heat would rise without anything supplying them, which no experiment has ever exhibited. The only consistent sign is therefore the observed one: the induced current opposes the change, the force on the magnet resists its motion, and the work done against that force is exactly the energy dissipated in the loop. The local form of the same bookkeeping is Poynting's theorem (Section 60.4.2), which turns this argument from a consistency requirement into an identity of the field equations. □

60.1.3 The Maxwell–Faraday equation

[Reserved: the local form $\boldsymbol{\nabla} \times \boldsymbol{E} = -\partial \boldsymbol{B}/\partial t$ abstracted from the flux rule [89, 164]; the two distinct physical cases hiding in one flux rule—motional emf (magnetic force on carried charges) versus transformer emf (induced electric field)—whose unification is relativistic and is completed in Chapter 38.]

60.2 The displacement current and the complete system

60.2.1 The inconsistency of the static circuital law

[Reserved: Ampère's law of Chapter 59 requires $\boldsymbol{\nabla} \cdot \boldsymbol{J} = 0$ and therefore fails for open circuits —the charging capacitor as the canonical case; charge conservation, $\partial \rho/\partial t + \boldsymbol{\nabla} \cdot \boldsymbol{J} = 0$, as the constraint the corrected law must respect.]

60.2.2 Maxwell's dynamical theory

[Reserved: the displacement current $\epsilon_0 \, \partial \boldsymbol{E}/\partial t$, introduced by Maxwell first within the molecular-vortex model of On Physical Lines of Force *[157] and then rederived, free of*

mechanical scaffolding, in A Dynamical Theory of the Electromagnetic Field *[164]; the complete system of field equations, its rearrangement in the* Treatise *[185], and the counting of its independent equations.]*

60.2.3 The Heaviside vector form

[Reserved: the reduction of Maxwell's original twenty equations in twenty variables to the four vector equations used ever since, carried out by Heaviside in 1885 in the Electrician *series reprinted in his* Electrical Papers *[252]; the modern SI statement of the system [1976, 2333], boxed here as the reference form for the rest of the treatise.]*

60.3 Potentials and gauge freedom

60.3.1 Scalar and vector potentials

[Reserved: the homogeneous pair of Maxwell equations is solved identically by $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$ *and* $\boldsymbol{E} = -\boldsymbol{\nabla}\phi - \partial \boldsymbol{A}/\partial t$*, extending the static potentials of Chapter 58 and Chapter 59; the inhomogeneous pair becomes two coupled second-order equations for* $(\phi, \boldsymbol{A})$ *[164].]*

60.3.2 Gauge transformations

[Reserved: the invariance of the fields under $\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\chi$*,* $\phi \to \phi - \partial\chi/\partial t$*; the Lorenz condition, due to L. V. Lorenz [168]—the Danish physicist, not H. A. Lorentz, a misattribution the literature keeps alive—which decouples the potential equations into wave equations; the Coulomb gauge; gauge freedom as the classical seed of the gauge principle of Chapter 64.]*

60.4 Conservation laws

60.4.1 Charge conservation

[Reserved: the continuity equation as an identity of the Maxwell system—the displacement current makes conservation automatic rather than assumed [164]; global charge conservation and its experimental precision.]

Phenomenon 60.3 (Conservation of electric charge)**.** The total electric charge of an isolated system never changes. Charge appears and disappears only in pairs of opposite sign, and no process has ever been seen to create or destroy a net amount of it. The sharpest test is the stability of the electron: as the lightest charged particle it has nothing charged to decay into, so any decay would destroy charge outright, and no such decay has been observed — the resulting bounds on its mean life exceed the age of the universe by many orders of magnitude [2427]. Locally the statement is a continuity equation,

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot \boldsymbol{J} = 0\,. \tag{60.2}$$

Derivation. Take the divergence of the Ampère–Maxwell law $\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J} + \mu_0 \epsilon_0\, \partial \boldsymbol{E}/\partial t$. The left-hand side is the divergence of a curl and vanishes identically, so

$$0 = \mu_0 \boldsymbol{\nabla} \cdot \boldsymbol{J} + \mu_0 \epsilon_0 \frac{\partial}{\partial t} \left(\boldsymbol{\nabla} \cdot \boldsymbol{E} \right) , \tag{60.3}$$

and substituting Gauss's law $\boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho/\epsilon_0$ gives Equation (60.2) after dividing by μ_0. Integrating over a volume and applying the divergence theorem turns it into the global statement that the charge inside a region changes only by the current crossing its boundary.

The step is worth reading backwards, because it is the whole reason for the displacement current. Without that term the same operation yields $\boldsymbol{\nabla}\cdot\boldsymbol{J}=0$, which is false for any circuit that is open — a capacitor being charged — and it is the demand that charge be conserved, not any new measurement, that forced the correction (Section 60.2.1). Conservation is thereby an identity of the completed system rather than an extra assumption imposed on it. □

60.4.2 Poynting's theorem

[Reserved: the local energy balance of the electromagnetic field, with energy density $u=(\epsilon_0 E^2+B^2/\mu_0)/2$ and flux $\boldsymbol{S}=\boldsymbol{E}\times\boldsymbol{B}/\mu_0$, derived by Poynting [229]; where the energy of a circuit actually flows; the ambiguity of $\boldsymbol{S}$ up to a curl and why no experiment inside electrodynamics resolves it.]

60.4.3 The stress tensor and field momentum

[Reserved: the Maxwell stress tensor [185], the momentum density $\epsilon_0\boldsymbol{E}\times\boldsymbol{B}$, and the momentum balance of field plus matter; radiation pressure predicted by Maxwell [185] and measured by Lebedev [299] and by Nichols and Hull [311]; angular momentum of the field, with its mechanical detection in circularly polarized light by Beth [767]; the covariant packaging of these densities awaits Chapter 64.]

Phenomenon 60.4 (Radiation pressure)**.** Light falling on a surface pushes it. The effect was predicted from the field momentum [185] and measured, against enormous thermal backgrounds, with torsion balances in evacuated vessels by Lebedev [299] and by Nichols and Hull [311]. At normal incidence the pressure on a perfectly absorbing surface is

$$P_{\mathrm{rad}}=\frac{I}{c}\,, \tag{60.4}$$

with I the incident irradiance, and twice that on a perfect reflector. Electromagnetic energy therefore carries mechanical momentum, and the field is a dynamical object rather than a bookkeeping device.

Derivation. The electromagnetic momentum density is $\boldsymbol{g}=\epsilon_0\boldsymbol{E}\times\boldsymbol{B}=\boldsymbol{S}/c^2$, with $\boldsymbol{S}=\boldsymbol{E}\times\boldsymbol{B}/\mu_0$ the Poynting vector (Section 60.4.2) and $\boldsymbol{g}$ the momentum entry of the Maxwell stress tensor (Section 60.4.3). For a plane wave in vacuum $|\boldsymbol{S}|=uc$, where u is the energy density, so $|\boldsymbol{g}|=u/c$: the wave carries one unit of momentum for every c units of energy.

Momentum arrives at an absorbing surface at the rate at which the wave delivers it, that is, the momentum density times the speed at which it advances:

$$P_{\mathrm{rad}}=|\boldsymbol{g}|\;c=u=\frac{I}{c}\,, \tag{60.5}$$

since $I=\langle|\boldsymbol{S}|\rangle=\langle u\rangle\,c$ and all quantities are here understood as time averages. A perfect reflector reverses the momentum instead of absorbing it, so the momentum transferred per unit time is doubled and so is the pressure. □

Phenomenon 60.5 (Angular momentum of circularly polarized light)**.** A circularly polarized beam carries angular momentum about its own direction of propagation, and delivers it to matter as a measurable mechanical torque. A birefringent half-wave plate suspended on a fine quartz fibre reverses the handedness of a beam passing through it, and is observed to experience a steady torque of the sense and magnitude that the transfer of angular momentum requires [767].

Derivation pending #17

Angular momentum of the electromagnetic field: derive the angular momentum density from the field momentum density, and show that for a circularly polarized beam of angular frequency ω the ratio of angular momentum to energy is $1/\omega$, so that reversing the handedness transfers twice that amount per unit energy.

60.5 The electromagnetic wave prediction

60.5.1 The wave equation and the speed c

[Reserved: in vacuum the system yields wave equations for $\boldsymbol{E}$ and $\boldsymbol{B}$ with speed $1/\sqrt{\mu_0\epsilon_0}$; Weber and Kohlrausch's measurement of the ratio of electrical units [146], Maxwell's identification of that ratio with the measured speed of light, and his conclusion that light is an electromagnetic disturbance [164]; the wave-equation machinery is that of Chapter 10.]

Phenomenon 60.6 (Light is an electromagnetic wave)**.** A speed can be extracted from measurements that involve no light at all. Comparing the charge on a Leyden jar, determined electrostatically from the force between charged bodies, with the same charge determined electrodynamically from the deflection it produces on discharge through a galvanometer, gives a ratio of units with the dimensions of a velocity — and that velocity came out equal, within the accuracy of the measurement, to the speed of light determined optically [146]. Electromagnetic disturbances were later generated directly, and found to propagate through air, to reflect, to refract in a prism of pitch, to be polarized by a grid of wires, and to travel at a speed consistent with the optical value [239]. Light is an electromagnetic wave [164], and in vacuum both fields obey

$$\nabla^2 \boldsymbol{E} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2}\,, \qquad c = \frac{1}{\sqrt{\mu_0\epsilon_0}}\,. \tag{60.6}$$

Derivation. In a region free of charge and current take the curl of the Maxwell–Faraday equation and exchange the order of the space and time derivatives on the right:

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{E}) = -\frac{\partial}{\partial t}(\boldsymbol{\nabla}\times\boldsymbol{B}) = -\mu_0\epsilon_0\frac{\partial^2 \boldsymbol{E}}{\partial t^2}\,, \tag{60.7}$$

the last step by the Ampère–Maxwell law with $\boldsymbol{J} = 0$. On the left use the identity $\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{E}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{E}) - \nabla^2\boldsymbol{E}$, in which the first term vanishes because $\boldsymbol{\nabla}\cdot\boldsymbol{E} = 0$ where $\rho = 0$. What remains is Equation (60.6); the identical manipulation starting from the Ampère–Maxwell law gives the same equation for $\boldsymbol{B}$.

The propagation speed is therefore fixed by μ_0 and ϵ_0 alone — two constants measured with magnets, currents and charged spheres — and it is the coincidence of that number with the optical speed of light that identifies the two phenomena. Note that the displacement current is indispensable: delete $\mu_0\epsilon_0\,\partial\boldsymbol{E}/\partial t$ and the right-hand side of Equation (60.7) vanishes, leaving no wave. □

60.5.2 Hertz's confirmation

[Reserved: the generation, propagation, reflection and interference of electromagnetic waves in air by Hertz [239], closing the prediction; standing waves and the measured wave speed; full apparatus-level treatment in Chapter 63, and the developed optics of the waves themselves in Chapter 61.]

60.6 Structure of the theory

60.6.1 Relativity of the field decomposition

[Reserved: the split of the electromagnetic field into $\boldsymbol{E}$ and $\boldsymbol{B}$ depends on the observer; the asymmetry of explanation for induction (magnet moving versus circuit moving) that opens Einstein's 1905 paper [321]; forward references to Chapter 38 and Chapter 39, where the field strength is assembled into a single tensor.]

60.6.2 Duality and the absence of magnetic charge

[Reserved: the vacuum system is symmetric under the rotation of $(\boldsymbol{E}, c\boldsymbol{B})$ into one another; sources break the symmetry because no magnetic charge is observed—the empirical record is that of Chapter 59; what a monopole term would do to the equations [677].]

60.6.3 Limits and the domain of validity

[Reserved: the quasi-static limits recovering Chapter 58 and Chapter 59; circuit theory as a limit; linearity and superposition, exact classically and modified only at the quantum level (photon–photon scattering, treated with QED in Chapter 100); classical electrodynamics as the low-photon-number, high-occupancy limit examined in Chapter 122.]

Chapter 61

Electromagnetic Waves and Optics

This chapter develops the wave solutions of the Maxwell system of Chapter 60: plane waves, polarization, energy transport, and the two great limits of the theory—physical optics, where interference and diffraction rule, and geometrical optics, the short-wavelength limit whose mechanics-side twin was met in Chapter 23. Optics is far older than electrodynamics, so the chapter also runs backwards through history: two centuries of measured optical phenomena—refraction, polarization, interference, diffraction, dispersion, and the speed of light itself—were already quantitative science before Maxwell explained what was waving. Each phenomenon is stated with its founding measurement; the wave-optics experiments are treated in full apparatus detail in Chapter 67, and the failure of the aether reading of these waves in Chapter 37.

Derivation pending #18

Electromagnetic Waves and Optics: all derivations of this chapter are pending.

61.1 Plane electromagnetic waves

61.1.1 Monochromatic plane waves

[Reserved: transversality, the mutual orthogonality of $\boldsymbol{E}$, $\boldsymbol{B}$ and the wavevector, the ratio $E/B = c$, and the dispersion relation $\omega = c\,|\boldsymbol{k}|$ in vacuum [164, 1976]; the electromagnetic spectrum from radio to gamma, with Hertz's waves (Chapter 60) as its first artificial band; superposition and wave packets, with the Fourier machinery of Chapter 17.]

61.1.2 Energy and momentum transport

[Reserved: time-averaged Poynting flux, intensity, and radiation pressure for plane waves, specialising the conservation laws of Chapter 60 [229]; irradiance in SI units and the link to photometric practice.]

61.2 The speed of light

61.2.1 Astronomical determinations

[Reserved: Rømer's inference of a finite speed from the eclipse timings of Io [12], and Bradley's stellar aberration, which gave the first terrestrial-independent ratio of c to the orbital speed of the Earth [19]; aberration reappears as a relativity test in Chapter 37.]

Phenomenon 61.1 (Light propagates at a finite speed)**.** The eclipses of Jupiter's innermost satellite do not recur at strictly equal intervals as observed from the Earth. They fall

progressively late through the part of the year in which the Earth is receding from Jupiter, and progressively early through the part in which it is approaching, and the discrepancy accumulated across half a year amounts to many minutes — far more than any irregularity of the satellite's orbit could produce, and correlated with the Earth's position rather than with Jupiter's [12]. Independently, the apparent position of every star describes a small annual ellipse whose angular size is the same for all stars regardless of their distance [19], which a parallax cannot be. Light therefore takes time to travel, and the second observation fixes the ratio of the Earth's orbital speed to the speed of light.

Derivation. Let the satellite emerge from its planet's shadow at strictly periodic instants t_k, separated by the true orbital period T, and let $d(t)$ be the distance from the event to the observer. The emergence is *seen* at $t_k + d(t_k)/c$, so the observed interval between successive emergences is

$$T_{\text{obs}} = T + \frac{d(t_{k+1}) - d(t_k)}{c}\,, \tag{61.1}$$

which exceeds T while the observer recedes and falls short of it while the observer approaches. Summing Equation (61.1) over all the eclipses of one recession, the intermediate distances cancel in pairs and the total accumulated lag is

$$\Delta t_{\text{total}} = \frac{\Delta d}{c}\,, \tag{61.2}$$

Δd being the total increase in the Earth–Jupiter distance, which is close to the diameter of the Earth's orbit because Jupiter moves slowly by comparison. A finite accumulated lag therefore means a finite c, and Equation (61.2) determines it from two measurable lengths and one measurable time. An infinite speed predicts strictly equal intervals, which is what is refuted.

Aberration is the same finiteness read as an angle: a telescope carried transversely at speed v must be tilted forward by $\tan\theta = v/c$ for the light entering the objective to reach the eyepiece, so the apparent displacement is common to all stars and annual in period — and it measures v/c, not c alone. □

61.2.2 Terrestrial measurements

[Reserved: Fizeau's toothed-wheel measurement over the Suresnes–Montmartre baseline [125]; Foucault's rotating mirror, which both measured c in air and showed light travels slower *in water, deciding against the corpuscular theory [128], and his 1862 absolute determination [158]; Michelson's 1879 Annapolis measurement [215] and his Mount Wilson–San Antonio triangulation [584].]*

Phenomenon 61.2 (Light travels more slowly in a denser medium)**.** Light travels more slowly in water than in air. The rotating-mirror method makes the comparison directly, sending the two beams over equal paths and reading the difference as a displacement of the returned image, and the speed in water is lower by very nearly the factor n by which water refracts [128]. The measurement is decisive rather than merely quantitative: the wave account requires the speed to *fall* in the more refracting medium, and the emission account requires it to rise.

Derivation. In a transparent non-magnetic medium the macroscopic Maxwell equations have the same form as in vacuum with ϵ_0 replaced by $\epsilon = \epsilon_r\epsilon_0$ (Chapter 62), so the wave equation derived in Chapter 60 carries a propagation speed

$$v = \frac{1}{\sqrt{\mu_0\epsilon}} = \frac{c}{\sqrt{\epsilon_r}} = \frac{c}{n}\,, \qquad n := \sqrt{\epsilon_r} \geq 1\,. \tag{61.3}$$

That the same n is the refractive index of Snell's law follows from Phenomenon 61.3: matching the wavefronts across the interface gives $\sin\theta_1/\sin\theta_2 = v_1/v_2$, so a medium in which the ray bends towards the normal is one in which the wave is slower.

The emission theory reverses the conclusion from the same data. There the ray bends because the interface pulls the corpuscle normally, leaving the tangential velocity component unchanged; then $v_1 \sin\theta_1 = v_2 \sin\theta_2$, so $v_2/v_1 = \sin\theta_1/\sin\theta_2 = n > 1$ and light must go *faster* in water. The two predictions differ by the factor $n^2 \approx 1.8$ for water and are not reconcilable; the measurement gives the wave value. □

61.2.3 From measured constant to defined constant

[Reserved: the laser frequency-times-wavelength determination of Evenson and collaborators [1470], the 1983 fixing of c at exactly 299 792 458 m/s*, and the consequent redefinition of the metre [2333]; what it means, in the framework of Chapter 2, for a constant to become a definition.]*

61.3 Reflection and refraction

61.3.1 Snell's law and Fermat's principle

[Reserved: the sine law of refraction, found by Snel around 1621 and first printed in Descartes' Dioptrique *[3]; Fermat's derivation of the same law from least time, in the 1662 letters to Cureau de la Chambre [260]; Huygens' wavefront construction [15]; total internal reflection and the evanescent wave [1967].]*

Phenomenon 61.3 (The sine law of refraction)**.** A ray crossing the plane boundary between two transparent media is bent, and the ratio of the sines of the angles of incidence and refraction, both measured from the normal, is the same for every angle of incidence and depends only on the pair of media [3]:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \,. \tag{61.4}$$

The constancy of the ratio is the content of the observation — the angles themselves are not proportional, and the older tables that assumed they were fail at large incidence.

Derivation. Fermat's principle: of all paths from a point A in medium 1 to a point B in medium 2, light takes the one of stationary transit time [260]. Place the interface along the plane $y = 0$, let $A = (0, a)$ and $B = (d, -b)$ with $a, b > 0$, and let the path cross at $(x, 0)$. The optical time is

$$t(x) = \frac{n_1}{c}\sqrt{x^2 + a^2} + \frac{n_2}{c}\sqrt{(d-x)^2 + b^2}\,, \tag{61.5}$$

using $v_i = c/n_i$ from Equation (61.3). Setting $\mathrm{d}t/\mathrm{d}x = 0$,

$$\frac{n_1 x}{\sqrt{x^2 + a^2}} = \frac{n_2 (d-x)}{\sqrt{(d-x)^2 + b^2}}\,, \tag{61.6}$$

and the two quotients are precisely $\sin\theta_1$ and $\sin\theta_2$, the angles being measured from the normal y-axis. This is Equation (61.4). The second derivative of Equation (61.5) is positive, so the stationary point is a minimum, and it is unique. Huygens' construction gives the same law geometrically, by requiring the wavefronts on the two sides to keep step along the interface [15], and this is the version that identifies n with c/v rather than leaving it an empirical constant of the pair of media. □

61.3.2 The Fresnel equations

[Reserved: amplitude reflection and transmission coefficients for the two polarizations, derived by Fresnel from his transverse-wave theory in the memoir read in 1823 [90]; Brewster's angle of complete polarization by reflection, established empirically by Brewster [58]; phase shifts on reflection and their interferometric consequences; the electromagnetic rederivation from the boundary conditions of Chapter 60 [1976].]

Phenomenon 61.4 (Complete polarization at the Brewster angle)**.** For every transparent dielectric there is one angle of incidence, and one only, at which the reflected light is completely linearly polarized, its electric vector lying in the surface. The tangent of that angle is the refractive index of the material [58]:

$$\tan\theta_B = n\,. \tag{61.7}$$

Equivalently, at θ_B the reflected and refracted rays are at right angles, as follows at once from Equation (61.4). The reflectance for light polarized in the plane of incidence falls smoothly to exactly zero there and rises again beyond it, so the vanishing is a genuine zero and not a minimum.

Derivation pending #19

Brewster's angle: derive the Fresnel amplitude coefficients for the two polarizations from the interface conditions on the tangential fields, and show that the coefficient for light polarized in the plane of incidence has a simple zero exactly where the reflected and refracted directions are perpendicular, which by the sine law is at the angle whose tangent is the refractive index.

61.4 Polarization

61.4.1 Discovery and phenomenology

[Reserved: double refraction of Iceland spar, reported by Bartholin [9] and given a wave construction by Huygens [15]; Malus's discovery of polarization by reflection and his cosine-squared intensity law [54]; Brewster's law relating the polarizing angle to the refractive index [58].]

Phenomenon 61.5 (Malus's law)**.** Light that has been reflected from glass at a suitable angle, or transmitted through a crystal of Iceland spar, acquires a property that depends on the orientation of the beam about its own axis: passed through a second such device, it is transmitted in full at one orientation, extinguished at the orientation at right angles to it, and in between the transmitted intensity follows

$$I(\theta) = I_0\cos^2\theta \tag{61.8}$$

with θ the angle between the two devices [54]. Ordinary light shows no such dependence, which is why the property had to be discovered rather than assumed.

Derivation. Take the beam leaving the first polarizer to be linearly polarized with electric field amplitude $\boldsymbol{E}_0 = E_0\hat{\boldsymbol{e}}_1$. An ideal analyser transmits the component of the field along its own axis $\hat{\boldsymbol{e}}_2$ and rejects the orthogonal component, so the transmitted amplitude is

$$E = \boldsymbol{E}_0\cdot\hat{\boldsymbol{e}}_2 = E_0\cos\theta\,, \tag{61.9}$$

θ being the angle between the two axes. Intensity is proportional to the time-averaged square of the field — it is $\epsilon_0 c E^2/2$ for a plane wave in vacuum, by Chapter 60 — so squaring Equation (61.9) gives Equation (61.8). The whole content of the derivation is that the observable is the square of a *projected vector*: that is why the law is $\cos^2\theta$ rather than $|\cos\theta|$, and it is the first quantitative evidence that the optical disturbance is a directed quantity transverse to the ray. □

61.4.2 Mathematical description

[Reserved: linear, circular and elliptical polarization of a transverse plane wave; the Stokes parameters and the description of partially polarized light [139]; Jones vectors for fully polarized beams [1967]; the measurement of the angular momentum carried by circularly polarized light [767].]

61.4.3 Birefringence and wave plates

[Reserved: ordinary and extraordinary rays, the index ellipsoid, quarter- and half-wave plates, and polarimetry [1967]; induced birefringence is deferred to the electro- and magneto-optics of Chapter 62.]

61.5 Interference

61.5.1 Young's double slit

[Reserved: Young's demonstration that light added to light can give darkness, and his measurement of wavelengths from fringe spacings [49]—the decisive evidence for the wave theory; the sinusoidal two-source pattern and its visibility; apparatus detail in Chapter 67, and the single-quantum version of the experiment in Chapter 71.]

Phenomenon 61.6 (Two-beam interference)**.** Light from a single small source, divided into two beams by a pair of narrow apertures a distance d apart and allowed to fall on a screen at distance $D \gg d$, produces not two patches of light but a set of equally spaced bright and dark bands. Blocking either aperture destroys the bands and leaves a smooth illumination, so the dark bands are places where light added to light gives darkness. The band spacing is

$$\Delta y = \frac{\lambda D}{d}\,, \tag{61.10}$$

proportional to D and inversely proportional to d, and this is what first made the wavelength of light a measured quantity [49].

Derivation. Let the two apertures lie at $\pm d/2$ on a line perpendicular to the axis, and let y measure position on the screen from the axis. The path lengths from the two apertures to the point y differ, for $d \ll D$ and $|y| \ll D$, by

$$\delta = \sqrt{D^2 + \left(y + \tfrac{d}{2}\right)^2} - \sqrt{D^2 + \left(y - \tfrac{d}{2}\right)^2} \simeq \frac{y d}{D}\,, \tag{61.11}$$

by expanding each root to first order in $(y \pm d/2)^2/D^2$. Superposing two disturbances of equal amplitude and of the same frequency, whose phases differ by $2\pi\delta/\lambda$, gives an intensity

$$I(y) = 4I_1 \cos^2\!\left(\frac{\pi\delta}{\lambda}\right) = 4I_1 \cos^2\!\left(\frac{\pi d\, y}{\lambda D}\right)\,, \tag{61.12}$$

because the amplitudes, not the intensities, are what add. This is maximal where δ is an integer multiple of λ and zero where it is a half-odd multiple, so consecutive maxima are

separated by $\Delta y = \lambda D/d$, which is Equation (61.10). Blocking one aperture removes the interference term and leaves the constant I_1, as observed; and Equation (61.10) inverted, $\lambda = d\,\Delta y/D$, is the measurement, since d, Δy and D are all lengths that can be read off with a scale. □

61.5.2 Thin films and interferometers

[Reserved: thin-film and wedge interference, Newton's rings (described in the Opticks *[18]); division of amplitude and the Michelson interferometer, built for the aether-drift measurement [236] and treated as a null relativity test in Chapter 37; multiple-beam interference and the Fabry–Pérot étalon [1967].]*

61.5.3 Coherence

[Reserved: temporal and spatial coherence as the condition for visible fringes; the van Cittert–Zernike theorem relating source size to transverse coherence [736, 826]; Michelson's stellar interferometer and the measured diameter of Betelgeuse [479]; quantum coherence of light is taken up in Chapter 122.]

61.6 Diffraction

61.6.1 Fresnel diffraction and the Arago spot

[Reserved: Fresnel's prize memoir on diffraction, combining Huygens' secondary wavelets with interference [78]; Poisson's reductio—a bright spot at the centre of a circular shadow—observed by Arago and reported with the memoir [78], converting the objection into a confirmation; Fresnel zones and the zone plate.]

Phenomenon 61.7 (The bright spot in the centre of a shadow)**.** A small opaque disc of smooth circular edge, illuminated by a point source, casts a shadow with a bright point at its exact centre; for an ideal disc the intensity there equals the intensity that would be found with the disc removed. The straight edge and the disc are also surrounded by systems of fringes whose positions can be measured on a micrometer and compared, one by one, with the values computed from the diffraction integral [78]. Neither the bright centre nor the fringe positions has any account in a theory of rays.

Derivation pending #20

The axial bright point: derive it from the Huygens–Fresnel superposition of secondary wavelets over the plane of the obstacle, showing that every point of the disc's rim is equidistant from the axial field point, so that the wavelets from the rim arrive exactly in phase whatever the radius of the disc, and evaluate the resulting on-axis amplitude for an ideal circular obstacle.

61.6.2 Fraunhofer diffraction and gratings

[Reserved: far-field diffraction as the Fourier transform of the aperture, with the machinery of Chapter 17; Fraunhofer's diffraction gratings and his wavelength standards [69]; the single slit, the circular aperture, and the Rayleigh resolution criterion [209].]

61.6.3 Kirchhoff's scalar theory

[Reserved: the scalar diffraction integral derived from the wave equation with Kirchhoff's boundary values [224], its internal inconsistency and the Rayleigh–Sommerfeld repair [1967]; the validity conditions under which the earlier Fresnel and Fraunhofer treatments emerge as limits.]

61.7 Geometrical optics

61.7.1 The eikonal limit

[Reserved: the short-wavelength expansion of the wave equation, the eikonal equation $(\nabla \mathcal{S})^2 = n^2$*, and the ray equation, named after Bruns's* Das Eikonal *[263]; Hamilton's characteristic-function optics [84] and its mechanical twin, already developed in Chapter 23; where the limit fails—edges and caustics [1967].]*

61.7.2 Images and instruments

[Reserved: mirrors, thin lenses and the paraxial imaging equations; the telescope and microscope as diffraction-limited instruments, tying the resolution limit of Section 61.6.2 to practice [1967]; aberrations named and classified.]

61.8 Dispersion

61.8.1 The prism and the solar spectrum

[Reserved: Newton's experimentum crucis separating white light into unequally refrangible rays [10] and its systematic account in the Opticks *[18]; Fraunhofer's dark lines in the solar spectrum and his measurement of their wavelengths [60], the founding data of the spectroscopy taken up in Chapter 70.]*

Phenomenon 61.8 (Dispersion and the fixed lines of the solar spectrum)**.** A prism spreads white light into a band of colours in an invariable order, and a colour once separated is not decomposed further by a second prism, nor changed by transmission through matter: the refractive index is a function of the colour, and each colour is a simple constituent of the light rather than a modification produced by the glass [10, 18]. Examined at sufficient resolution, the spectrum of the Sun is crossed by a great number of fixed dark lines, always in the same places, whose refractive positions can be measured far more precisely than any colour can be named — which is what turns dispersion from a qualitative effect into an instrument [60].

Derivation pending #21

Dispersion: derive the frequency dependence of the refractive index from the response of bound charges to the field of the wave, obtaining normal dispersion away from resonance and anomalous dispersion with absorption near it, and recover the empirical Sellmeier form as its low-loss limit. The model belongs to the treatment of electrodynamics in matter, and the discreteness of the solar lines themselves is not classical at all and awaits the atomic chapters.

61.8.2 The rainbow

[Reserved: Descartes' ray-tracing of the primary and secondary bows [3]; Newton's account of their colours [18]; supernumerary bows as an interference effect and Airy's caustic integral [103]—the rainbow as the first quantitative test case where ray optics fails and wave optics is required.]

61.8.3 Phase velocity, group velocity, and media

[Reserved: normal and anomalous dispersion; the Sellmeier relation [180]; Rayleigh's distinction of group from phase velocity [197]; light in moving media—Fresnel's partial-drag coefficient [61] and Fizeau's moving-water interferometry confirming it [133], reread relativistically in Chapter 37; the microscopic theory of dispersion belongs to Chapter 62.]

Chapter 62

Electrodynamics in Matter

Electrodynamics in matter is the macroscopic Maxwell theory: the polarization and magnetization fields, the auxiliary fields $\boldsymbol{D}$ and $\boldsymbol{H}$, the response functions that close the system, and dispersion—the frequency dependence of the response, constrained by causality alone through the Kramers–Kronig relations. The chapter gathers what Chapter 58 and Chapter 59 deferred: how the vacuum equations of Chapter 60, exact at the microscopic level, become equations for smoothed fields with matter-dependent sources, and what measured phenomena—conduction, dielectric response, dia-, para- and ferromagnetism, plasma transparency, the skin effect—the response functions encode. The microscopic explanations are quantum and are deferred to the condensed-matter part; here each phenomenon is stated with its founding measurement and its classical model, and the models' honest failures are marked.

Derivation pending #22

Electrodynamics in Matter: all derivations of this chapter are pending.

62.1 From microscopic to macroscopic fields

62.1.1 Bound charges and the polarization field

[Reserved: spatial averaging of the microscopic equations, following the electron-theory programme of Lorentz [356]; the polarization $\boldsymbol{P}$ as dipole-moment density, bound charge $-\boldsymbol{\nabla}\cdot\boldsymbol{P}$ and surface charge $\boldsymbol{P}\cdot\hat{\boldsymbol{n}}$; the displacement field $\boldsymbol{D}=\epsilon_0\boldsymbol{E}+\boldsymbol{P}$ of Maxwell's Treatise *[185].]*

62.1.2 Bound currents and the magnetization field

[Reserved: the magnetization $\boldsymbol{M}$, bound currents $\boldsymbol{\nabla}\times\boldsymbol{M}$ and $\partial\boldsymbol{P}/\partial t$, and the field $\boldsymbol{H}=\boldsymbol{B}/\mu_0-\boldsymbol{M}$ [185]; the macroscopic Maxwell system in matter and what in it is definition versus physics [1976].]

62.1.3 Constitutive relations and boundary conditions

[Reserved: linear isotropic media, $\boldsymbol{D}=\epsilon\boldsymbol{E}$ and $\boldsymbol{B}=\mu\boldsymbol{H}$; permittivity, permeability, susceptibilities and their SI units; interface conditions for $\boldsymbol{D}$ and $\boldsymbol{H}$, extending Chapter 58 and Chapter 59; energy in linear media [1976].]

62.2 Dielectrics

62.2.1 Polarizability and local fields

[Reserved: molecular polarizability, the Lorentz local-field correction [356], and the Clausius–Mossotti relation connecting polarizability to permittivity [131, 206]; its optical form, the Lorentz–Lorenz equation, derived independently by H. A. Lorentz [213] and L. Lorenz [214]—two different men, the same near-collision of names as the gauge condition of Chapter 60.]

62.2.2 Polar molecules and relaxation

[Reserved: orientation polarization of permanent molecular dipoles and its temperature dependence, after Debye [383]; Debye relaxation and the frequency dependence of the dielectric constant of polar liquids; dielectric loss.]

62.2.3 Piezo- and ferroelectricity

[Reserved: piezoelectricity, discovered in quartz by the Curie brothers [212]; ferroelectricity, discovered in Rochelle salt by Valasek [483], with hysteresis of $\boldsymbol{P}$ mirroring the magnetic case of Section 62.4.1; the microscopic theory is deferred to the condensed-matter part (Chapter 125).]

62.3 Conduction

62.3.1 Ohm's law

[Reserved: Ohm's empirical law of the galvanic circuit [81], its local form $\boldsymbol{J} = \sigma \boldsymbol{E}$, conductivity and resistivity in SI units; the Wiedemann–Franz proportionality of thermal and electrical conductivity of metals [140]; Joule heating and its place in the energy balance of Chapter 60.]

Phenomenon 62.1 (Ohm's law)**.** In a metallic conductor held at constant temperature the current is strictly proportional to the electromotive force driving it, over a range of many decades, with a coefficient that depends on the material and its temperature and on nothing else — not on the magnitude of the current, not on how long it has been flowing, not on its history [81]. Locally,

$$\boldsymbol{J} = \sigma \boldsymbol{E}\,, \tag{62.1}$$

with σ the conductivity, measured in $\mathrm{A/(V\,m)}$. The proportionality is a property of a class of materials and not a law of the field, which is why it may be — and in semiconductors, gas discharges and superconductors is — violated without anything in Chapter 60 failing.

Derivation. Model the conductor as containing n free carriers per unit volume, of charge q and mass m, which are accelerated by the field and lose all memory of their drift in collisions occurring at a mean interval τ [289]. Between collisions the equation of motion is $m\,\mathrm{d}\boldsymbol{v}/\mathrm{d}t = q\boldsymbol{E}$, so a carrier emerging from a collision with a randomly directed velocity has, after a time t, an extra velocity $q\boldsymbol{E}t/m$ along the field. Averaging over the collision statistics, the mean extra velocity — the drift velocity — is

$$\langle \boldsymbol{v} \rangle = \frac{q\tau}{m} \boldsymbol{E}\,, \tag{62.2}$$

which is constant in time: the field accelerates, the collisions brake, and the steady state is a terminal drift rather than a runaway. The current density is the charge carried across unit area per unit time, $\boldsymbol{J} = nq\,\langle \boldsymbol{v} \rangle$, so

$$\boldsymbol{J} = \frac{nq^2\tau}{m} \boldsymbol{E}\,, \qquad \sigma = \frac{nq^2\tau}{m}\,, \tag{62.3}$$

which is Equation (62.1), with the linearity traced to the single assumption that τ does not depend on $\boldsymbol{E}$.

That assumption is also the honest limit of the derivation. It holds when the drift velocity is small compared with the mean thermal speed, which in a metal at ordinary current densities it overwhelmingly is; it fails in a gas discharge, where the field sets the collision rate. Nothing here explains what fixes τ, and the classical estimate of its magnitude — and with it Drude's account of the electronic heat capacity — is wrong, for reasons that are quantum (Chapter 126). □

62.3.2 The Drude model and the Hall effect

[Reserved: Drude's kinetic model of metallic conduction [289]: DC and AC conductivity, the relaxation time, and its account of Wiedemann–Franz; the Hall effect [208] as the classical probe of carrier sign and density, and the model's honest failures (heat capacity, sign of some Hall coefficients), resolved only by the quantum theory of Chapter 126; the quantized Hall effect is treated in Chapter 128.]

Phenomenon 62.2 (The Hall effect)**.** A flat conducting strip carrying a steady current, placed in a magnetic field perpendicular to its face, develops a steady potential difference *across* the strip, perpendicular both to the current and to the field [208]. The transverse field is proportional to the product of the current density and the magnetic field,

$$E_y = R_H J_x B_z\,, \qquad R_H = \frac{1}{nq}\,, \tag{62.4}$$

so the measurement returns both the density and the *sign* of the charge carriers. That sign is negative in most metals, as the electron requires — but it is positive in others, beryllium and zinc among them, and no classical model of conduction accounts for a positive carrier in a metal.

Derivation. Let the current run along x with drift velocity v_x, and let $\boldsymbol{B} = B_z\hat{\boldsymbol{z}}$. Each carrier feels the Lorentz force $q(\boldsymbol{E} + \boldsymbol{v}\times\boldsymbol{B})$, whose magnetic part qv_xB_z pushes it along $-y$ or $+y$ according to the sign of qv_x. Charge therefore accumulates on one edge of the strip until the transverse electric field it creates cancels the magnetic push, after which the current runs straight again. That steady state is the observation, and it is defined by the vanishing of the transverse force:

$$qE_y - qv_xB_z = 0 \quad\Longrightarrow\quad E_y = v_xB_z\,. \tag{62.5}$$

Eliminating the drift velocity with $J_x = nqv_x$ from Equation (62.3) gives $E_y = J_xB_z/nq$, which is Equation (62.4). The sign of R_H is the sign of q, because the drift velocity reverses with the carrier sign while the current does not, so the two reversals compound rather than cancel — which is why the experiment is a measurement of the carrier sign and why the metals that return the wrong one remained a genuine anomaly for decades. □

62.3.3 The skin effect

[Reserved: field penetration into a good conductor, the skin depth $\delta = \sqrt{2/(\mu\sigma\omega)}$, and high-frequency resistance, analysed by Rayleigh [233] and Heaviside [252]; consequences for shielding, waveguides and cables.]

62.4 Magnetic materials

62.4.1 Dia-, para- and ferromagnetism

[Reserved: Faraday's discovery that all matter is magnetic— diamagnetic bismuth repelled, paramagnets attracted [119]; Curie's systematic measurements: temperature independence of

diamagnetism, the $1/T$ law of paramagnetism, and the disappearance of ferromagnetism at a critical temperature [264]; ferromagnetic hysteresis, named and mapped by Ewing [231].]

Phenomenon 62.3 (All matter is magnetic)**.** No substance is indifferent to a magnetic field. Suspended in a field strong enough and inhomogeneous enough to make the force visible, every material tested is either drawn towards the region of stronger field or pushed towards the region of weaker field; the second class, of which bismuth is the type case, had not been suspected before it was looked for, and it includes water, glass and flesh [119]. The division is sharp: the repelled substances are repelled whatever the direction of the field, and their response is very nearly independent of temperature, while the attracted ones weaken as the temperature rises [264].

Derivation pending #23

The magnetic response of matter: derive the Bohr–van Leeuwen theorem, that a classical system of charges in thermal equilibrium has identically zero magnetization because the magnetic field enters the classical Hamiltonian only through a shift of the canonical momenta, which is absorbed by a translation of the integration variables in the partition function. The theorem states that the phenomenon above has no classical explanation at all, so every classical model in this section is a parametrization and not a derivation, and the real account is the quantum one of the magnetism chapter.

Phenomenon 62.4 (Ferromagnetic hysteresis)**.** In iron, nickel, cobalt and their alloys the magnetization is not a function of the applied field. Carrying the field around a closed cycle carries the magnetization around a closed curve of nonzero area: the material retains a magnetization when the field is removed (remanence), requires a reversed field of definite strength to bring that magnetization to zero (coercivity), and approaches a saturation beyond which further field produces almost nothing [231]. The state therefore depends on the history, the cycle dissipates energy, and no number deserves the name of "the" permeability of the material.

Derivation pending #24

Ferromagnetic hysteresis: derive the energy dissipated per unit volume per cycle as the area enclosed by the loop traced in the plane of applied field against magnetization, and identify precisely which hypothesis behind a linear constitutive relation the loop violates. The microscopic origin — domains, exchange coupling and magnetocrystalline anisotropy — is quantum and belongs to the magnetism chapter.

62.4.2 Langevin theory and the Curie–Weiss law

[Reserved: Langevin's statistical theory of dia- and paramagnetism [323], the Langevin function and the Curie law; Weiss's molecular field, the Curie–Weiss law and spontaneous magnetization [339]; what the molecular field really is (exchange) awaits Chapter 129.]

Phenomenon 62.5 (The Curie law and the Curie temperature)**.** Measurements of magnetic susceptibility carried across a wide range of temperature separate three behaviours [264]. Diamagnetic susceptibility is negative and nearly independent of temperature. Paramagnetic susceptibility is positive and falls as the reciprocal of the absolute temperature,

$$\chi = \frac{C}{T}\,, \tag{62.6}$$

with C a constant of the substance. And a ferromagnet loses its ferromagnetism entirely above a temperature characteristic of the material, above which it behaves as an ordinary paramagnet. The transition is sharp, and the temperature at which it happens is a reproducible property of the substance.

Derivation. Suppose the substance to contain n permanent magnetic dipoles per unit volume, each of fixed magnitude μ and free to take any orientation [323]. A dipole at angle θ to the applied field has energy $-\mu B\cos\theta$, so in thermal equilibrium at temperature T the orientations are distributed as $\mathrm{e}^{\mu B\cos\theta/k_BT}$ and the mean component along the field is

$$\langle \mu\cos\theta\rangle = \mu\left(\coth x - \frac{1}{x}\right) =: \mu L(x)\,, \qquad x := \frac{\mu B}{k_BT}\,, \tag{62.7}$$

the angular integrals being elementary in the variable $\cos\theta$. The magnetization is $M = n\mu L(x)$, which saturates at $n\mu$ for $x \gg 1$ — all dipoles aligned — and for $x \ll 1$ has the expansion $L(x) = x/3 + O(x^3)$. Ordinary laboratory conditions are deep in the second regime, so

$$M = \frac{n\mu^2 B}{3k_BT}\,, \qquad \chi = \frac{\mu_0 M}{B} = \frac{\mu_0 n\mu^2}{3k_BT}\,, \tag{62.8}$$

which is Equation (62.6) with $C = \mu_0 n\mu^2/3k_B$. The $1/T$ is the competition of two energies and nothing else: alignment is worth μB and disorder is worth k_BT.

The derivation is not, however, a classical one, and saying so is the point of Phenomenon 62.3. It *postulates* dipoles of fixed magnitude that cannot be destroyed by the field or by the thermal motion, and classical mechanics supplies no such object; without that postulate the equilibrium magnetization is exactly zero. The Curie temperature is outside the model altogether — independent dipoles have no cooperative transition — and requires the molecular field of Section 62.4.2, whose own origin is quantum (Chapter 129). □

62.4.3 The Bohr–van Leeuwen theorem

[Reserved: classical statistical mechanics yields exactly zero equilibrium magnetization [484]—the theorem that makes every classical model of this section provisional and magnetism intrinsically quantum; statement, scope, and its role as a signpost to Chapter 129.]

62.4.4 Superconductors: deferred

[Reserved: perfect conductivity and the Meissner–Ochsenfeld expulsion of magnetic flux [724] are named here only as the extreme magnetic response of matter; the London phenomenology, the microscopic theory and the experimental record are treated in Chapter 130 and Chapter 131.]

62.5 Dispersion and causality

62.5.1 The Lorentz oscillator model

[Reserved: bound electrons as damped driven oscillators [356]; normal and anomalous dispersion, absorption lines, and the empirical Sellmeier form recovered ([180], cf. Chapter 61); oscillator strengths and the f-sum rule; the quantum reinterpretation is noted and deferred to Part IX.]

62.5.2 The Kramers–Kronig relations

[Reserved: causality of the response function alone implies dispersion relations connecting the real and imaginary parts of $\epsilon(\omega)$, derived by Kronig [550] and Kramers [581]; sum rules; their derivation uses the complex analysis of Chapter 8, and their reach extends far beyond optics (scattering amplitudes, Chapter 86).]

Phenomenon 62.6 (Absorption and dispersion occur together)**.** In a spectral region where a substance is transparent, its refractive index rises with frequency, smoothly and monotonically, and the measured curve is reproduced across the whole visible range by an empirical formula with one term for each absorption band lying outside it [180]. Wherever the substance *does* absorb, that behaviour reverses: within the absorption band the index falls with increasing frequency. Anomalous dispersion is found where and only where there is absorption, never elsewhere, and the strength of the one tracks the strength of the other. The refractive index and the absorption coefficient are therefore not independent properties of a material, and measuring either over all frequencies determines the other.

Derivation pending #25

Absorption and dispersion: derive the Kramers–Kronig relations, in which the real and imaginary parts of the permittivity are each the Hilbert transform of the other. The only physical input is causality — no response before the stimulus — which makes the response function analytic in the upper half of the complex frequency plane, whence a contour integral gives the pair of dispersion relations and the sum rules that follow from the high-frequency limit. The argument needs the complex analysis of the corresponding methods chapter, and it is worth noting that it uses no model of the medium at all.

62.5.3 Signal velocity and the front

[Reserved: anomalous dispersion permits phase and group velocities exceeding c; Sommerfeld and Brillouin's resolution—the signal front never travels faster than c, with forerunner precursors [412, 421]; consistency with the causal structure of Chapter 38.]

62.6 Free-carrier response and plasmas

62.6.1 Plasma frequency and transparency

[Reserved: the free-electron permittivity $\epsilon(\omega) = \epsilon_0(1 - \omega_p^2/\omega^2)$ from the Drude model [289]; plasma oscillations identified by Tonks and Langmuir [639]; reflection below and transparency above ω_p—the ultraviolet transparency of metals and the ionospheric reflection of radio; the full dynamics of plasmas belongs to Chapter 66.]

Phenomenon 62.7 (A gas of free carriers reflects below a critical frequency and transmits above it)**.** Two observations, made in different sciences, are the same phenomenon. A metal reflects visible light almost completely and is opaque to it, yet the alkali metals become transparent in the ultraviolet, each above a threshold frequency of its own. And radio waves directed upwards are returned from the upper atmosphere below a critical frequency, while above it they are lost into space — a reflection whose height was measured directly by timing the returned pulse and by interference between the direct and reflected rays [517]. In both cases the medium contains free charges of number density n, and the dividing frequency is

$$\omega_p = \sqrt{\frac{nq_e^2}{\epsilon_0 m_e}}\,. \tag{62.9}$$

Below ω_p the medium reflects; above it, it transmits [289, 639].

Derivation. Take carriers that are free — no restoring force — and, for the moment, collisionless, so that $m_e\,\mathrm{d}\boldsymbol{v}/\mathrm{d}t = q_e\boldsymbol{E}$. For a field oscillating as $\mathrm{e}^{-\mathrm{i}\omega t}$ this gives $\boldsymbol{v} = \mathrm{i}q_e\boldsymbol{E}/m_e\omega$ and a current density $\boldsymbol{J} = nq_e\boldsymbol{v} = \mathrm{i}nq_e^2\boldsymbol{E}/m_e\omega$. The current that drives the magnetic field in the Ampère–Maxwell law is the sum of this and the vacuum displacement current, and the sum is

$$\boldsymbol{J} + \epsilon_0\frac{\partial \boldsymbol{E}}{\partial t} = -\mathrm{i}\omega\epsilon_0\left(1 - \frac{\omega_p^2}{\omega^2}\right)\boldsymbol{E} =: -\mathrm{i}\omega\,\epsilon(\omega)\boldsymbol{E}\,, \tag{62.10}$$

with ω_p as in Equation (62.9): the free carriers act as a medium of permittivity $\epsilon(\omega) = \epsilon_0(1 - \omega_p^2/\omega^2)$. Repeating the wave-equation derivation of Chapter 60 with $\epsilon_0 \to \epsilon(\omega)$ gives the dispersion relation

$$\omega^2 = \omega_p^2 + c^2k^2\,. \tag{62.11}$$

For $\omega > \omega_p$ the wavenumber k is real and the wave propagates, with a phase velocity above c and a group velocity below it. For $\omega < \omega_p$, k is imaginary: the field falls off exponentially into the medium instead of propagating. Since $\epsilon(\omega)$ is real in this model, nothing is absorbed, and an incident wave that cannot propagate and cannot be absorbed must be returned — total reflection. The dividing frequency is a measurement of n through Equation (62.9), which is what makes ionospheric sounding a density diagnostic. □

62.7 Magneto- and electro-optics

[Reserved: field-induced anisotropy of media. The Faraday effect—rotation of the plane of polarization by a magnetic field along the ray, the first observed link between light and electromagnetism [118]; the quadratic electro-optic (Kerr) effect [189]; Verdet constants and applications (optical isolators, polarimetry); connection to the birefringence machinery of Chapter 61.]

Phenomenon 62.8 (The Faraday effect)**.** A block of heavy glass placed in a strong magnetic field rotates the plane of polarization of linearly polarized light sent along the field direction. The rotation is proportional to the field strength and to the path length,

$$\theta = V\,B\,d\,, \tag{62.12}$$

with V a constant of the material, and its sense is fixed by the field alone: sending the beam back through the same block *doubles* the rotation instead of undoing it [118]. The effect is therefore not reciprocal, which distinguishes it sharply from the rotation produced by a sugar solution, and it was the first demonstration that light and magnetism act on one another — the observation that made an electromagnetic theory of light thinkable before there was one.

Derivation pending #26

The Faraday effect: derive the difference between the refractive indices of the two circular polarizations for a medium of bound charges in a magnetic field along the ray, by solving the driven oscillator equation with the magnetic force included; then show that a linear polarization, resolved into two circular components of equal amplitude, emerges rotated by half the accumulated phase difference, which gives the Verdet constant. The non-reciprocity should be derived too, and it follows from the field being an axial vector that does not reverse when the propagation direction does.

Chapter 63

Experiments: Electromagnetism

Electromagnetism is the first branch of physics that was built, experiment by experiment, into a field theory, and this chapter collects the measurements that did the building: Coulomb's torsion balance and Cavendish's null test for the inverse-square law [38, 205], Ørsted's deflected needle [68], Ampère's force between currents [66], Faraday's induction [89], the Weber–Kohlrausch measurement that first tied the two systems of units to the speed of light [146], Hertz's electromagnetic waves [239], and Millikan's quantization of charge [407]. Each is the evidence for a specific structural claim of Chapters 58 and 59 and of Chapters 60 and 61, and each is cross-referenced from the theory it tests.

When written out, every experiment below will use the structured `experiment` environment, with apparatus, procedure, observations (numbers and uncertainties in SI units), interpretation and primary references given in full. The chapter closes with the modern precision descendants: laboratory bounds on a deviation from the Coulomb exponent [1460] and the resulting photon-mass limits compiled by the Particle Data Group [2427].

Derivation pending #27

Experiments: Electromagnetism: all derivations of this chapter are pending.

63.1 The predictions under test

[Reserved: the chain of claims this chapter must test, stated before any apparatus: that the electrostatic force is central and inverse-square (Chapter 58); that currents produce and feel magnetic fields with a definite force law (Chapter 59); that a changing magnetic flux drives an electromotive force; that the completed equations propagate transverse waves at a speed fixed by electric and magnetic measurements alone, the prediction of Maxwell's dynamical theory [164]; and that charge comes in integer multiples of a smallest unit.]

63.2 Coulomb: the torsion balance (1785)

63.2.1 Apparatus and procedure

[Reserved: Coulomb's torsion balance [38]: the silver suspension wire as a calibrated torque standard, the charged pith balls, the twist-angle protocol for repulsion, and the separate torsion-pendulum timing method of the second memoir for attraction. The instrument is the same one that weighs the gravitational constant in Chapter 33, and the shared metrology of torsion fibres will be treated once, there, and cited from here.]

63.2.2 Observations

[Reserved: the tabulated twist angles versus separation from [38], converted to SI — forces of order 10^{-4} N at separations of a few cm *— with Coulomb's own three-point verification that halving the distance quadruples the force, and a modern uncertainty analysis of the instrument's systematics (charge leakage, induction on the enclosure).]*

63.2.3 Interpretation

[Reserved: the inference from the data to the force law of Chapter 58, written as an exponent bound $F \propto r^{-2\pm\delta}$ with Coulomb's own δ of order a few percent; why a torsion measurement tests the law only at laboratory separations; and the statement that the modern form of this inference is the photon-mass bound of Section 63.11.2.]

Phenomenon 63.1 (The electrostatic force is central and inverse-square)**.** Two small bodies at rest carrying static electric charge act on one another along the line joining them, with a force proportional to the product of the charges and falling as the inverse square of the separation. The torsion balance establishes the exponent directly, by showing that halving the distance quadruples the force, to within a few percent [38]. The null method establishes it far more sharply: if the exponent is written $2+\delta$, then and only then does the field inside a closed charged conductor vanish for $\delta = 0$, so a sensitive detector inside a charged shell that registers nothing bounds δ. That line of experiment runs from Cavendish, who reached $|\delta| < 0.02$ a century before the result was printed [205], through Plimpton and Lawton at 2×10^{-9} [783], to lock-in detection between concentric icosahedral shells at $|\delta| \lesssim 3 \times 10^{-16}$ [1460]. Read as a photon mass, the same null record gives $m_\gamma < 1 \times 10^{-18}\,\mathrm{eV}/c^2$ [2427].

Derivation pending #28

The Coulomb exponent: derive the interior field of a uniformly charged spherical shell for a force law with exponent $2+\delta$, show that it vanishes identically only when δ is zero, and give the conversion of a measured null voltage between concentric conductors into a bound on δ. Then relate δ to a photon rest mass through the Yukawa form that a massive vector field gives in place of the Coulomb potential.

63.3 Cavendish: the concentric spheres

[Reserved: Cavendish's unpublished null experiment, performed around 1772 and printed only in Maxwell's 1879 edition of the electrical researches [205]: a charged outer sphere, an inner sphere connected and then isolated, and the null reading on the pith-ball electrometer that bounds the exponent deviation to $|\delta| < 0.02$ — a sharper test than [38], a century earlier than its publication. The null method — zero signal inside a closed conductor if and only if the exponent is exactly 2 — is the template for every later test in Section 63.11.1, and its theory is Gauss's law from Chapter 58.]

63.4 Galvanic circuits: Ohm and Kirchhoff

[Reserved: the quantitative laws of steady currents: Ohm's measurements with thermocouple sources and calibrated wires and the linear relation of current to electromotive force [81], and Kirchhoff's circuit laws with their field-theoretic justification [114]. The material evidence — galvanometer deflections versus wire length in SI units — and the reduction of circuit theory to the field description of Chapter 62 are to be written here.]

63.5 Ørsted: the current and the needle (1820)

[Reserved: Ørsted's observation that a galvanic current deflects a nearby compass needle, circulating around the wire rather than pointing along it [68] — the first coupling of electricity to magnetism, and the first force in physics that is neither central nor along the line joining its sources. The immediate quantitative sequel of Biot and Savart, measuring the field of a long straight current through the oscillation period of a suspended magnet [67], gives the $1/r$ law of Chapter 59. Apparatus, procedure and the deflection data in SI units are to be tabulated when written.]

63.6 Ampère: the force between currents (1820)

[Reserved: Ampère's current balance and his four null experiments, establishing within weeks of Ørsted's announcement that parallel currents attract and antiparallel currents repel [66], and the mature force law of the 1826 treatise deduced, as its title claims, uniquely from experiment [77]. The force per unit length between long parallel wires — $2 \times 10^{-7}\,\mathrm{N/m}$ at $1\,\mathrm{A}$ and $1\,\mathrm{m}$ — defined the SI ampere until the 2019 redefinition fixed the elementary charge instead [2333]; both definitions and the continuity between them belong here, with Chapter 59 carrying the theory.]

63.7 Faraday: electromagnetic induction (1831–1832)

63.7.1 Apparatus and procedure

[Reserved: the induction ring — two coils on a soft iron torus — the galvanometer as ballistic charge integrator, and the three protocols of the First Series of the Experimental Researches [89]: switching the primary current, moving a magnet through a coil, and rotating a copper disc in a magnetic field (the homopolar generator).]

63.7.2 Observations

[Reserved: which quantities are tabulated when written — ballistic galvanometer throws proportional to flux change, in SI $\mathrm{V\,s}$; the transient nature of the effect (current flows only while the flux changes); the sign reversal on reversal of motion; and the proportionality of integrated electromotive force to the change of flux rather than to its rate history [89].]

Phenomenon 63.2 (The induced charge measures the change of flux, not its rate)**.** A ballistic galvanometer in a circuit linked to a changing magnetic flux gives a throw proportional to the *total* charge that has passed, and that charge depends only on the total change of flux and on the resistance of the circuit — not on how quickly the change was made. Switching the primary current of an induction ring slowly or abruptly gives the same throw; so does moving a magnet through a coil quickly or slowly [89]. Only the endpoints matter:

$$Q = \frac{|\Delta \Phi_B|}{R}\,. \tag{63.1}$$

Derivation. The circuit has resistance R and carries the current $I = \mathcal{E}/R$ driven by the induced electromotive force. Integrating over the whole of the transient, and using the flux rule $\mathcal{E} = -\mathrm{d}\Phi_B/\mathrm{d}t$ (Chapter 60),

$$Q = \int I\,\mathrm{d}t = \frac{1}{R}\int \mathcal{E}\,\mathrm{d}t = -\frac{1}{R}\int_{t_1}^{t_2} \frac{\mathrm{d}\Phi_B}{\mathrm{d}t}\,\mathrm{d}t = -\frac{\Phi_B(t_2) - \Phi_B(t_1)}{R}\,, \tag{63.2}$$

by the fundamental theorem of calculus. The integrand's time dependence has been integrated away: whatever path Φ_B took between the two endpoints, and however long it took, only the difference survives. This is Equation (63.1).

The result is what makes the ballistic galvanometer the right instrument, because such an instrument responds to the impulse delivered to its coil — the time integral of the current — rather than to the current itself, provided the transient is short compared with its swing period. It is also why the measurement is a *flux* meter: Equation (63.1) inverted gives $\Delta\Phi_B = QR$ in V s, from a charge and a resistance. □

63.7.3 Interpretation

[Reserved: the induction law as the experimental content of the curl-E Maxwell equation of Chapter 60; Faraday's own field-line language as the origin of the field concept; the motional/transformer distinction and its resolution in the flux rule, whose two mechanisms unify only under the relativity of Chapter 38.]

63.8 Weber and Kohlrausch: the ratio of the units (1856)

[Reserved: the measurement that connected electricity to light before Maxwell wrote a field equation: the ratio of the electrodynamic to the electrostatic unit of charge, measured by discharging a Leyden jar of known electrostatic charge through a ballistic galvanometer [146], yielding a velocity within a few percent of the speed of light. Maxwell's identification of that velocity with the propagation speed of electromagnetic waves [164] turned this number into the central prediction that Section 63.9 tests.]

Phenomenon 63.3 (The ratio of the two units of charge is a speed, and that speed is the speed of light)**.** The quantity of electricity on a charged conductor can be measured in two entirely independent ways: electrostatically, from the mechanical force it exerts on another charged body, and electrodynamically, from the deflection produced when it is discharged through a galvanometer whose constant has been fixed magnetically. The two measures are not the same quantity, and their ratio has the dimensions of a velocity. Weber and Kohlrausch measured that ratio by discharging a Leyden jar of known electrostatic charge, and the velocity that came out agreed, to within the accuracy of the experiment, with the speed of light determined optically [146] — a number obtained with no light anywhere in the apparatus, and the first quantitative link between electromagnetism and optics [164].

Derivation pending #29

The ratio of the units: show that the electrostatic and electromagnetic definitions of the unit of charge differ by a factor with the dimensions of a speed, equal in SI terms to the reciprocal square root of the product of the electric and magnetic constants, and hence that this measurement determines exactly the combination that reappears as the propagation speed in the vacuum wave equation. The error budget of the original determination, and its agreement with the contemporary optical values, are to be reconstructed here.

63.8.1 Convection currents: Rowland (1876)

[Reserved: Rowland's demonstration in Helmholtz's laboratory that a moving electrostatic charge produces the same magnetic field as a conduction current [202] — charge in motion is current, with no distinction of mechanism — closing the conceptual gap between Section 63.2 and Section 63.6 and anticipating the relativity of the field decomposition treated in Chapter 38.]

Phenomenon 63.4 (Charge in motion is current)**.** An insulating disc carrying a static electrostatic charge, spun rapidly about its axis, deflects a magnetometer needle exactly as a circular conduction current of the same magnitude would. The deflection reverses with the sense of rotation and with the sign of the charge, and it is proportional to the product of the charge and the angular velocity [202]. There is therefore no distinction of mechanism in the magnetic effect: it is the motion of charge that matters, not whether that motion is conduction along a wire or the carriage of a charge on a moving body.

Derivation. The current density that appears as the source in the field equations is $\boldsymbol{J} = \rho \boldsymbol{v}$, and this is a definition of what a current *is* — charge crossing unit area per unit time — with nothing in it about the mechanism of the motion. A ring of the disc at radius r, carrying surface charge density σ and turning at angular velocity Ω, therefore constitutes a surface current

$$K(r) = \sigma\, v = \sigma \Omega r\,, \tag{63.3}$$

whose total, integrated over the radius of a disc of radius a carrying total charge $Q = \sigma \pi a^2$, is $I = \int_0^a \sigma \Omega r \,\mathrm{d}r = \sigma \Omega a^2/2 = Q\Omega/2\pi$ — the whole charge passing any radial line once per revolution, as it must. Since the field equations see only $\boldsymbol{J}$, the field produced is the field of that current, and it is proportional to $Q\Omega$ and reverses with either factor, as observed.

The result looks trivial and is not: before it, "current" was a property of galvanic circuits, and it was an open question whether the magnetic action belonged to the conduction process itself. What the experiment shows is that the source term of magnetostatics is the convective flux of charge, which is exactly the statement that relativity later makes structural, since a static charge in one frame is a current in another (Chapter 38). □

63.9 Hertz: electromagnetic waves (1888)

63.9.1 Apparatus and procedure

[Reserved: the spark-gap dipole oscillator driven by an induction coil, the resonant loop receiver with its micrometer spark gap, and the standing-wave protocol: reflecting the radiation from a zinc sheet and mapping nodes and antinodes along the laboratory [239].]

63.9.2 Observations

[Reserved: the tabulated node positions giving wavelengths of several metres at an oscillator frequency near $50\,\mathrm{MHz}$*, hence a propagation speed consistent with the optical* $3 \times 10^8\,\mathrm{m/s}$*; the demonstrations of rectilinear propagation, reflection, refraction in a pitch prism, and polarization by a wire grid [239].]*

Phenomenon 63.5 (Electromagnetic waves in free space)**.** A spark-gap oscillator radiates a disturbance that crosses the laboratory, is reflected from a metal sheet, and forms with its own reflection a stationary pattern whose nodes and antinodes can be located by carrying a small resonant loop through it and watching its spark appear and vanish. The node spacing gives wavelengths of several metres, which with the oscillator's frequency yields a propagation speed agreeing with the optical speed of light to within the accuracy of the determination. The same radiation is reflected from conductors, refracted by a prism of pitch, and extinguished by a grid of parallel wires held one way and transmitted by it held the other [239]. It is therefore a transverse wave, and it behaves in every tested respect as light does.

Derivation. Only the standing-wave measurement needs a derivation; the rest is direct observation. A conducting sheet cannot support a tangential electric field at its surface,

so the reflected wave must cancel the incident one there. Superposing an incident wave travelling towards the sheet at $x = 0$ with its reflection,

$$E(x,t) = E_0 \cos(kx + \omega t) - E_0 \cos(kx - \omega t) = -2E_0 \sin(kx) \sin(\omega t)\,, \tag{63.4}$$

which vanishes for all t wherever $\sin(kx) = 0$, that is at $x = 0,\ \pi/k,\ 2\pi/k, \ldots$. The nodes are therefore spaced by

$$\Delta x = \frac{\pi}{k} = \frac{\lambda}{2}\,, \tag{63.5}$$

so the measured spacing gives $\lambda = 2\Delta x$ with a metre rule and nothing else. The propagation speed then follows from the universal relation $v = f\lambda$, with f the frequency of the oscillator, fixed by its capacitance and inductance. The magnetic field has its *antinodes* where the electric field has its nodes, which is why the receiving loop, sensitive to the changing magnetic flux through it, reads the pattern shifted by a quarter wavelength from a detector sensitive to $\boldsymbol{E}$ — a check on the interpretation, not a complication of it. □

63.9.3 Interpretation

[Reserved: the verdict — electromagnetic disturbances propagate at finite speed as transverse waves, as [164] predicted and as Chapter 61 derives; light is such a wave, and the optical evidence for that identification is the subject of Chapter 67. Action-at-a-distance electrodynamics is dead from this table onward.]

63.10 Millikan: the quantization of charge (1911–1913)

63.10.1 Apparatus and procedure

[Reserved: the oil-drop chamber: atomizer, parallel plates at kilovolt potentials, the microscope and stopwatch protocol of timed rises and falls of a single drop, and the correction to Stokes drag at low pressure worked out in the 1911 paper [377]. The prehistory — the electron's e/m from cathode-ray deflection [276] — fixes what the measured e implies for the electron mass.]

63.10.2 Observations

[Reserved: charge changes of single drops always integer multiples of one unit; the 1913 value $e = 1.592 \times 10^{-19}\,\mathrm{C}$ (in modern units, with its uncertainty dominated by the then-current viscosity of air) [407]; the table of drop radii, fall times and inferred charges to be reproduced in SI when written.]

Phenomenon 63.6 (Electric charge is quantized)**.** The charge carried by a single oil drop, measured one drop at a time by timing its fall under gravity and its rise against gravity in a known electric field, is always an integer multiple of one and the same unit. When a drop captures an ion from the surrounding air its charge changes — and it changes by an integer multiple of that same unit, never by a fraction of it. The unit is independent of the size of the drop, of the substance of which it is made, and of the means by which the ions were produced [377, 407]. Charge is therefore not a continuous quantity, and the elementary unit is that carried by the electron of [276]. Millikan's own value was $e = 1.592 \times 10^{-19}\,\mathrm{C}$ [407]; since 2019 the elementary charge has been exact by definition, and its value fixes the SI ampere [2333].

Derivation. Let W be the weight of the drop less the buoyancy of the air. With no field the drop reaches a terminal speed v_f at which the viscous drag balances W; Stokes' law

makes the drag proportional to the speed, so

$$W = k\,v_f\,, \qquad k = 6\pi\eta a\,, \tag{63.6}$$

with η the viscosity of the air and a the radius of the drop. Now apply a vertical field E strong enough to lift the drop, and let it rise at terminal speed v_r; the balance is now $qE - W = k\,v_r$. Adding the two balances eliminates W:

$$q = \frac{k\,(v_f + v_r)}{E}\,. \tag{63.7}$$

The radius hidden in k is not an independent unknown: writing $W = \tfrac{4}{3}\pi a^3 \Delta\rho\, g$ with $\Delta\rho$ the density of the oil less that of the air, Equation (63.6) gives $a = \sqrt{9\eta v_f / 2\Delta\rho g}$. Every quantity on the right of Equation (63.7) is therefore measured: two speeds obtained with a microscope scale and a stopwatch, a field obtained from a voltage and a plate separation, and tabulated material constants.

What the derivation delivers is a number for each drop; what the *experiment* delivers is the observation that those numbers cluster on integer multiples, and that a drop which changes its charge does so by jumping between multiples. No step of the derivation assumes or imposes this, which is what makes the quantization an observation rather than a construction. The one systematic that mattered is visible in Equation (63.6): Stokes' law fails when the drop is not large compared with the mean free path of the air molecules, and the correction — together with the value adopted for η — is the source of the difference between Millikan's e and the modern one. □

63.10.3 Interpretation

[Reserved: charge is quantized; the elementary unit is carried by the electron of [276]; the viscosity systematic and its later correction as a case study in error analysis for Chapter 2. Since 2019 the elementary charge is exact by definition [2333], and the burden of measurement has moved to the quantities defined through it [2392].]

63.11 Modern repetitions and precision

63.11.1 The Coulomb exponent

[Reserved: the lineage of null tests after Cavendish: Plimpton and Lawton's resonant detection inside charged concentric shells, bounding the exponent deviation at 2×10^{-9} [783], and the Williams–Faller–Hill experiment with concentric icosahedral shells and lock-in detection at 4 MHz, reaching $|\delta| \lesssim 3 \times 10^{-16}$ [1460]. The geometry of each apparatus and the conversion of a null voltage to an exponent bound are to be written here.]

63.11.2 Photon-mass bounds

[Reserved: the modern reading of every exponent test — a Yukawa deviation from Coulomb's law is a photon mass, the Proca theory of Chapter 64 — and the compilation of bounds in the Review of Particle Physics, reaching $m_\gamma < 1 \times 10^{-18}\,\mathrm{eV}/c^2$ from solar-wind magnetohydrodynamics [2427], with the laboratory bound of [1460] translated into the same units. The plasma physics behind the solar-wind limit lives in Chapter 66.]

63.11.3 Charge quantization today

[Reserved: where Millikan's result stands now: the exact elementary charge of the revised SI [2333, 2392]; fractional charges of quarks, never observed free, inferred from deep inelastic

scattering (Chapter 116); and the fractionally charged quasiparticles of the fractional quantum Hall effect (Chapter 128), which quantize collective excitations rather than violate charge quantization.]

Chapter 64

Generalized Classical Field Theory

A field is a dynamical system carrying a degree of freedom at every point of space, and this chapter builds the machinery such systems require: the action functional and its Lagrangian density, the Euler–Lagrange field equations, Noether's theorems and the currents and charges they produce [456], the energy–momentum tensor in its canonical and improved forms [850], and the Hamiltonian formulation. The construction is the continuum limit of the machinery of Chapters 21 and 22: the label a that ran over finitely many coordinates becomes the continuous label $\boldsymbol{x}$, sums become integrals, and derivatives become functional derivatives in the sense of Chapter 16. What survives the limit is everything that matters — a variational principle, a symmetry-to-conservation-law correspondence, and a phase space.

The chapter sits here, at the end of Part VII, because electromagnetism is the classical field theory for which the evidence is overwhelming and which therefore earns the general formalism. Maxwell's equations [164] follow from a single Lorentz-scalar Lagrangian density; gauge invariance of that density forces charge conservation; the same variational calculation delivers the energy density, the Poynting flux [229] and the Maxwell stress [185] that are measured as radiation pressure [299, 311]. Two further systems are stated for later use: the classical Klein–Gordon field [544], the simplest relativistic field of all and the template for every scalar in Chapter 104, and the massive vector field of Proca [785], whose comparison with Maxwell is what turns laboratory electrostatics into a bound on the photon mass. Where the Legendre transformation of Chapter 22 fails — and for every gauge theory it does fail — the treatment is the constraint formalism already developed in Chapter 26, which this chapter uses rather than repeats. Standard treatments are [1552, 1912, 1976].

Derivation pending #30

Generalized Classical Field Theory: all derivations of this chapter are pending.

64.1 Fields as dynamical systems

64.1.1 From many degrees of freedom to a continuum

[Reserved: the loaded string of Chapter 28 taken to the limit of infinitely many beads at fixed mass density, in which the discrete coordinates $q_a(t)$ become a field $\phi(t, \boldsymbol{x})$ and the coupled equations become a wave equation; the field as the physical object rather than a bookkeeping device, which is Maxwell's own reading of the electromagnetic field as a dynamical system with energy stored in space [164]; locality as the statement that the Lagrangian is an integral of a density depending on the field and its first derivatives at one point; and why a nonrelativistic mechanical model of that density — an aether — is neither needed nor tenable (Chapter 37).]

64.1.2 The action functional and the Lagrangian density

[Reserved: the action $S[\phi] = \int \mathcal{L}(\phi, \partial_\mu \phi)\, \mathrm{d}^4x$ over a spacetime region, with $\mathcal{L}$ a Lorentz scalar density in the sense of Chapter 39; the three standing restrictions — locality, at most first derivatives, and invariance under the Poincaré group — and what each buys; the dimensions of $\mathcal{L}$ in SI (energy per unit volume) and the normalization conventions fixed here for the whole treatise; the equivalence class of Lagrangians differing by a total divergence [1552].]

64.1.3 Functional derivatives

[Reserved: the functional derivative $\delta S/\delta\phi(x)$ defined by the first variation, its delta-function algebra $\delta\phi(x)/\delta\phi(y) = \delta^4(x-y)$, and the conditions under which a functional is differentiable, taken from Chapter 16; the field-theoretic version of the Euler operator; the notation used throughout Parts VII and XI is fixed here once [1912].]

64.2 The Euler–Lagrange field equations

64.2.1 Variation, boundary terms and natural conditions

[Reserved: the derivation of $\partial_\mu\left(\partial\mathcal{L}/\partial(\partial_\mu\phi)\right) - \partial\mathcal{L}/\partial\phi = 0$ by variation at fixed data on the boundary of the region; the surface term and the two ways to kill it — fixed variations or natural boundary conditions; why a total divergence added to $\mathcal{L}$ changes nothing; the extension to several fields and to higher derivatives, with the Ostrogradsky instability as the reason the treatise stops at first order [1552].]

64.2.2 The real scalar field: the Klein–Gordon equation

[Reserved: the Lagrangian density $\tfrac{1}{2}\partial_\mu\phi\,\partial^\mu\phi - \tfrac{1}{2}\mu^2\phi^2$ and its field equation $(\Box + \mu^2)\phi = 0$ with $\mu = mc/\hbar$, written down in 1926 by Gordon [544] and by Klein [547] (the latter in a five-dimensional setting whose extra dimension has no evidence and is not used anywhere in this treatise); plane-wave solutions and the dispersion relation $\omega^2 = c^2k^2 + \mu^2c^2$; the Yukawa static solution $\mathrm{e}^{-\mu r}/r$ and its interpretation as a force of finite range [765]; the classical field is developed here, its quantization in Chapter 92.]

64.2.3 The complex scalar field and internal symmetry

[Reserved: the complex field with $\mathcal{L} = \partial_\mu\phi^*\,\partial^\mu\phi - \mu^2\,|\phi|^2$; invariance under the global phase rotation $\phi \mapsto \mathrm{e}^{\mathrm{i}\alpha}\phi$, an internal $\mathrm{U}(1)$ symmetry in the sense of Chapter 14; the associated conserved current and the charge it integrates to, obtained from the theorem of Section 64.3.1; why the current is not a probability density and what that cost the interpretation of the equation before quantization [1912].]

64.3 Symmetries and conservation laws

64.3.1 Noether's first theorem

[Reserved: the theorem of [456] — to every continuous group of transformations leaving the action invariant, with r independent parameters, correspond r currents j^μ_A satisfying $\partial_\mu j^\mu_A = 0$ on shell; the charge $Q_A = \int j^0_A\, \mathrm{d}^3x$ is constant provided the flux through the boundary vanishes; invariance up to a total divergence is enough, which is what covers translations and Galilean boosts; the ambiguity $j^\mu \mapsto j^\mu + \partial_\nu k^{[\nu\mu]}$ that leaves the divergence and the charge untouched, and which the improvements of Section 64.4.2 exploit.]

64.3.2 Poincaré invariance and the ten charges

[Reserved: the four translations giving the energy–momentum tensor and the six Lorentz transformations giving the angular-momentum current $M^{\mu\nu\rho}$, whose conservation splits into orbital and spin parts only after the tensor is symmetrized; the ten conserved charges as the classical counterpart of the Poincaré generators of Chapter 39 and, after quantization, of the Wigner classification [849] used in Chapter 94; conformal invariance of the free massless field and the extra currents Bessel-Hagen extracted for electrodynamics [472].]

64.3.3 Internal symmetry and electric charge

[Reserved: the Noether charge of the internal $\mathrm{U}(1)$ *identified with electric charge in coulombs; global versus local invariance; conservation of charge as an experimental claim rather than a convention, with the sharpest laboratory bound coming from searches for charge-violating electron decay — a lifetime exceeding* $6.6 \times 10^{28}\,\mathrm{yr}$ *for* $e \to \nu\gamma$ *from Borexino [2248]; the compilation of related bounds in [2427].]*

64.3.4 Noether's second theorem and local symmetries

[Reserved: the second theorem of [456] — a symmetry depending on arbitrary functions of spacetime does not give new conservation laws but identities among the field equations, so that the equations are not independent and the evolution is underdetermined; the identity $\partial_\mu \partial_\nu F^{\mu\nu} \equiv 0$ *for electrodynamics and the contracted Bianchi identity for gravitation (Chapter 44) as the two instances that matter here; the corollary that a locally invariant theory has a singular Legendre transformation, which is where Chapter 26 takes over.]*

64.4 The energy–momentum tensor

64.4.1 The canonical tensor and its defects

[Reserved: $T^{\mu\nu}_{can} = (\partial\mathcal{L}/\partial(\partial_\mu\phi))\,\partial^\nu\phi - \eta^{\mu\nu}\mathcal{L}$ *as the Noether current of translations [456]; the interpretation of its components as energy density, momentum density, energy flux and stress, in SI units; the two defects — it is in general neither symmetric, so it cannot serve as the source of Chapter 44, nor gauge invariant, which for electrodynamics makes even its energy density depend on the gauge chosen [1552].]*

64.4.2 The Belinfante–Rosenfeld improvement

[Reserved: the improvement term built from the spin current, $T^{\mu\nu} = T^{\mu\nu}_{can} + \partial_\rho k^{[\rho\mu]\nu}$*, constructed independently by Belinfante [850] and Rosenfeld [855]; that it changes neither the conservation law nor the total charges, being of the superpotential form allowed by Section 64.3.1; that the result is symmetric and, for electrodynamics, gauge invariant; and the resulting clean separation of angular momentum into orbital and spin parts, the spin part being what Beth measured for light [767].]*

64.4.3 The Hilbert tensor and the gravitational source

[Reserved: the definition of $T^{\mu\nu}$ *as the response of the matter action to a variation of the metric [431], which is symmetric and gauge invariant by construction; the theorem that it coincides with the Belinfante tensor for every field of this chapter; its conservation as a consequence of diffeomorphism invariance by Noether's second theorem; and the statement that this is the object appearing on the right-hand side of the Einstein equations of Chapter 44, which is why the choice of improvement is not a matter of taste.]*

64.4.4 What is actually measured

[Reserved: the measured consequences of the electromagnetic stress–energy tensor — radiation pressure on a mirror, detected by Lebedev [299] and by Nichols and Hull [311]; the Maxwell stress on conductors and dielectrics [185]; the angular momentum of circularly polarized light [767]; and the momentum of light in a medium, where the rival Minkowski [346] and Abraham [351] forms differ by exactly a division of the total between field and matter, a question belonging to Chapter 62.]

64.5 The electromagnetic field as a Lagrangian system

64.5.1 The free-field Lagrangian

[Reserved: $\mathcal{L} = -\frac{1}{4\mu_0} F_{\mu\nu}F^{\mu\nu} - A_\mu J^\mu$ as the unique Poincaré- and gauge-invariant density quadratic in first derivatives of A_μ; the Euler–Lagrange equations returning the inhomogeneous Maxwell equations of Chapter 60 while the homogeneous pair is the identity $\partial_{[\rho}F_{\mu\nu]} = 0$ satisfied by any $F = \mathrm{d}A$ — so half of Maxwell's system is a consequence of using a potential at all; the second invariant $F_{\mu\nu}\tilde{F}^{\mu\nu}$, a total divergence classically and parity-odd, connecting to Chapter 105 [1552, 1976].]

64.5.2 Minimal coupling and the gauge principle

[Reserved: the replacement $\partial_\mu \to \partial_\mu + \mathrm{i}qA_\mu/\hbar$ that promotes a global phase symmetry to a local one, introduced for the relativistic charged particle by Fock [540] and raised to a principle by Weyl [641]; the interaction term $-A_\mu J^\mu$ and the Lorentz force recovered from it, matching Chapter 40 and the electron theory of [356]; the honest statement that the gauge principle predicts the form *of the coupling but not the value of q, whose quantization is an experimental fact (Chapter 63).]*

64.5.3 Gauge invariance and charge conservation

[Reserved: the demonstration that invariance of the action under $A_\mu \mapsto A_\mu + \partial_\mu\Lambda$ requires $\partial_\mu J^\mu = 0$ identically, not merely on shell — charge conservation is a consistency condition on the coupling, by Noether's second theorem [456, 472]; gauge fixing as the removal of the resulting redundancy, with the Lorenz condition [168] and the Coulomb and temporal gauges; and the count of physical polarizations deferred to Section 64.7.3.]

64.5.4 The electromagnetic energy–momentum tensor

[Reserved: the canonical tensor of the Maxwell field, which is neither symmetric nor gauge invariant, and the Belinfante improvement [850] that yields $T^{\mu\nu} = \mu_0^{-1}\left(F^{\mu\rho}F^{\nu}{}_{\rho} - \frac{1}{4}\eta^{\mu\nu}F^{\rho\sigma}F_{\rho\sigma}\right)$; identification of its components with the energy density $\frac{1}{2}(\epsilon_0 E^2 + B^2/\mu_0)$, the Poynting vector [229] and the Maxwell stress [185]; tracelessness as the classical shadow of conformal invariance [472]; and the use of the same tensor for the radiated power of Chapter 65.]

64.5.5 The non-abelian generalization

[Reserved: Yang and Mills' extension of the gauge principle to a non-abelian group [1054], with $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + g\,[A_\mu, A_\nu]$ and a Lagrangian of the same form; the decisive structural difference — the field strength is not gauge invariant but covariant, and the field carries its own charge, so the free theory is already nonlinear; the classical statement is made here, and the evidence for it is the strong and electroweak sectors of Chapters 102 and 104.]

64.6 Massive vector fields: the Proca theory

64.6.1 The Proca equation

[Reserved: the Lagrangian $-\frac{1}{4\mu_0}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}\mu^2 A_\mu A^\mu$ of [785] and its field equation $\partial_\mu F^{\mu\nu} + \mu^2 A^\nu = 0$; the mass term breaks gauge invariance, so A_μ is no longer redundant and the Lorenz condition $\partial_\mu A^\mu = 0$ becomes a consequence rather than a choice; three propagating polarizations instead of two; the static potential becomes Yukawa's $\mathrm{e}^{-\mu r}/r$ [765] with range $\hbar/mc$, which is the observable signature; and the discontinuity of the degree-of-freedom count in the massless limit, already analysed in Chapter 26.]

64.6.2 The photon mass as a measured quantity

[Reserved: how a laboratory test of Coulomb's law converts into an upper bound on μ, since a Proca field gives a deviation growing with apparatus size — the modern laboratory limit [1460] and the far sharper bounds from planetary and galactic magnetic fields reviewed in [2157], with the value adopted by this treatise taken from [2427], currently of order $1 \times 10^{-18}\,\mathrm{eV}/c^2$; and the methodological point that "the photon is massless" is a bound, never a measurement, so the correct statement is a number with a confidence level.]

64.7 Hamiltonian field theory

64.7.1 Momentum density and the Hamiltonian

[Reserved: the momentum density $\pi = \partial\mathcal{L}/\partial\dot{\phi}$, the Hamiltonian density $\mathcal{H} = \pi\dot{\phi} - \mathcal{L}$ and the field equations in first-order form; the loss of manifest covariance under the required 3+1 split of Chapter 39, and the fact that the resulting energy nonetheless agrees with T^{00} of Section 64.4.2; the scalar and Maxwell fields worked as the two standing examples [1912].]

64.7.2 Poisson brackets for fields

[Reserved: the equal-time bracket $\{\phi(\boldsymbol{x}), \pi(\boldsymbol{y})\} = \delta^3(\boldsymbol{x} - \boldsymbol{y})$ built from functional derivatives, generalizing Chapter 22; the symplectic structure on the space of solutions; the generators of the Poincaré group realized as brackets, closing on the algebra of Chapter 39; and this bracket as the object that the correspondence rule of Chapter 99 turns into a commutator.]

64.7.3 Singular Lagrangians and constraints

[Reserved: the Maxwell Lagrangian is singular — $\partial\mathcal{L}/\partial\dot{A}_0 \equiv 0$ — so its Hamiltonian formulation is the constrained one of Chapter 26, whose results are used here and not rederived: the primary constraint $\pi^0 \approx 0$, the secondary Gauss constraint, both first class, and the count of two physical polarizations per point of space, against three for the Proca field of Section 64.6.1; the observed absence of a longitudinal mode of light (Chapter 67) is what this count predicts.]

64.7.4 Covariant Hamiltonian formulations

[Reserved: the De Donder–Weyl formalism [752, 764], in which each field carries one polymomentum per spacetime direction and the Hamilton equations are first-order and manifestly covariant; the covariant phase space as the space of solutions with its symplectic form; what these formulations buy (covariance, a clean statement of the Noether charges as forms) and what they do not (a quantization rule of comparable standing to Chapter 99).]

64.8 From classical fields to quanta

[Reserved: the closing section, stating precisely what this chapter hands to Part XI and what it cannot decide. The mode expansion of a free field turns it into a continuum of harmonic oscillators (Chapter 28), which is the entire content of free field quantization; the canonical route is Chapter 99 and the path-integral route Chapter 106; the classification of the resulting one-particle states by mass and spin is Wigner's [849], treated in Chapter 94. What classical field theory already fixes — the number of polarizations, the form of the couplings, the conserved charges — and what it cannot — the value of $\hbar$, the spin–statistics connection, and the very existence of quanta — should be stated plainly here [1912].]

Chapter 65

Radiation and Scattering of Electromagnetic Waves

Accelerate a charge and it radiates. That single sentence, made quantitative, accounts for the light of every antenna, every synchrotron, every atom and every star, and it also destroys the classical atom — which is why this chapter is both the last word of classical electrodynamics and one of the doors into Part VIII. The chapter starts from the retarded solutions of the wave equations of Chapter 60, first written down by Lorenz [168], specializes them to a point charge in arbitrary motion — the Liénard–Wiechert potentials [281, 301] — and extracts from them the radiated power [275], the angular distribution, and the multipole hierarchy that organizes every emitter of finite size [241].

Three further themes follow, each with its own evidence. Radiation carries momentum away from its source, so an accelerated charge feels a back-reaction: the Abraham–Lorentz force [313, 356], its covariant completion by Dirac [813], their notorious runaway and pre-accelerating solutions, and the laser–electron collisions in which the effect has finally been measured [2305, 2317]. Radiation incident on a charge is re-radiated, which is scattering: Thomson's cross-section [332], Rayleigh's λ^{-4} law and the measured blue of the sky [181, 286], Mie's exact solution for a sphere [345], and the Compton shift [499] at which the classical account provably fails. And a charge moving through matter radiates in ways that a charge in vacuum cannot: bremsstrahlung [503], synchrotron emission [392, 900] and the Cherenkov cone [735, 792], which are the working tools of the detectors of Chapter 115. Standard treatments are [1552, 1976].

Derivation pending #31

Radiation and Scattering of Electromagnetic Waves: all derivations of this chapter are pending.

65.1 Fields of a charge in arbitrary motion

65.1.1 Retarded potentials

[Reserved: the wave equations for the potentials in Lorenz gauge and their retarded solutions [168], with the source evaluated at $t - |\boldsymbol{x} - \boldsymbol{x}'|/c$; Riemann's independent retarded-potential paper [170]; the retarded Green function of the d'Alembertian, taken from Chapter 10, and its support on the light cone; the advanced solution as an equally valid solution of the same equation, so that the choice of the retarded one is a boundary condition and not a theorem — the thermodynamic reading of that choice, and the Wheeler–Feynman absorber alternative [884], are to be stated honestly here.]

65.1.2 The Liénard–Wiechert potentials

[Reserved: the potentials of a point charge in arbitrary motion, $\varphi = \frac{q}{4\pi\epsilon_0}\left[\frac{1}{R(1-\boldsymbol{n}\cdot\boldsymbol{\beta})}\right]_{ret}$ and the corresponding vector potential, obtained independently by Liénard [281] and Wiechert [301]; the origin of the factor $(1-\boldsymbol{n}\cdot\boldsymbol{\beta})^{-1}$ as a Jacobian of the retarded-time condition rather than a Doppler factor; and the covariant one-line form in terms of the four-velocity, which makes the Lorentz transformation properties of Chapter 40 manifest.]

65.1.3 The velocity and acceleration fields

[Reserved: the field strengths obtained by differentiating the potentials, splitting into a velocity field falling as $1/R^2$ — a boosted Coulomb field, carrying no energy to infinity — and an acceleration field falling as $1/R$ and proportional to $\boldsymbol{a}$; the theorem that only the $1/R$ part contributes to the flux through a sphere of growing radius, which is what defines *radiation; the transverse and mutually orthogonal character of $\boldsymbol{E}$ and $\boldsymbol{B}$ in the wave zone [1976].]*

65.2 Radiated power

65.2.1 The Larmor formula

[Reserved: Larmor's result $P = q^2a^2/6\pi\epsilon_0c^3$ for a slowly moving charge [275], derived from the Poynting flux of the acceleration field; the $\sin^2\theta$ angular distribution about the acceleration; the numerical consequence that an electron on a Bohr-radius orbit radiates its energy away in about $1\times10^{-11}\,\mathrm{s}$, so that the classical atom cannot exist — the sharpest classical failure in this treatise, taken up in Chapter 70.]

Phenomenon 65.1 (An accelerated charge radiates)**.** A charge in uniform motion carries its field along with it and radiates nothing. A charge whose velocity changes emits energy that never returns. For speeds small compared with c the energy leaves at a rate proportional to the square of the acceleration and independent of the velocity,

$$P = \frac{q^2a^2}{6\pi\epsilon_0c^3}\,, \tag{65.1}$$

and it leaves with a $\sin^2\theta$ distribution about the direction of the acceleration, so that nothing at all is emitted along that direction [275]. Both the pattern and the polarization were mapped directly around an oscillating dipole with a resonant loop detector [241], and Equation (65.1) is the quantitative content of every antenna, every synchrotron and every classical radiating system in this chapter.

Derivation pending #32

The Larmor formula: derive the fields of a point charge in arbitrary motion from the retarded potentials, separate the velocity field, which falls as the inverse square of the distance and carries no energy to infinity, from the acceleration field, which falls as the inverse first power and is proportional to the acceleration, and integrate the Poynting flux of the acceleration field over a sphere of large radius.

Phenomenon 65.2 (Atoms are stable and radiate discrete lines)**.** Matter is made of atoms in which light electrons surround a small heavy nucleus, as the large-angle scattering of alpha particles requires [379]; an electron so bound is in permanent acceleration. Yet atoms do not collapse — they persist indefinitely, and a hydrogen atom today is indistinguishable

from one of any other epoch — and when they do radiate they emit sharp lines at fixed frequencies obeying exact numerical relations [230], not a continuum rising in pitch as the electron falls inward. Both facts contradict classical electrodynamics, and the contradiction is not marginal but catastrophic.

Derivation. Take a single electron of charge magnitude e and mass m_e on a circular orbit of radius r about a proton. The Coulomb attraction supplies the centripetal acceleration,

$$a = \frac{e^2}{4\pi\epsilon_0 m_e r^2}\,, \tag{65.2}$$

and the total energy of the orbit — kinetic plus potential, with the virial relation between them — is $E = -e^2/8\pi\epsilon_0 r$. Substituting Equation (65.2) into Equation (65.1) and equating the radiated power to the rate of loss of orbital energy, $\mathrm{d}E/\mathrm{d}t = -P$, gives

$$\frac{e^2}{8\pi\epsilon_0 r^2}\frac{\mathrm{d}r}{\mathrm{d}t} = -\frac{e^6}{96\pi^3\epsilon_0^3 m_e^2 c^3 r^4} \quad\Longrightarrow\quad \frac{\mathrm{d}r}{\mathrm{d}t} = -\frac{4}{3}\frac{r_e^2 c}{r^2}\,, \tag{65.3}$$

where $r_e = e^2/4\pi\epsilon_0 m_e c^2$ is the classical electron radius. Separating and integrating from an initial radius r_0 down to the nucleus,

$$t_{\mathrm{collapse}} = \frac{r_0^3}{4 r_e^2 c}\,. \tag{65.4}$$

With r_0 the Bohr radius $5.29\times10^{-11}\,\mathrm{m}$ and $r_e = 2.82\times10^{-15}\,\mathrm{m}$ [2392], this is about $1.6\times10^{-11}\,\mathrm{s}$.

Two predictions follow, and both are false. The atom would disappear in a fraction of a nanosecond; and as r shrank the orbital frequency would rise continuously, so the emitted spectrum would be a continuous sweep upward in frequency terminating in a burst, rather than the fixed sharp lines that are observed. This derivation therefore establishes not the phenomenon but its irreducibility: within classical electrodynamics there is no stable atom and no line spectrum, and no adjustment of the orbit, the charge distribution or the radiation law repairs it. The resolution is quantum and is taken up in Chapter 70. □

65.2.2 The relativistic generalization

[Reserved: Liénard's relativistic power law $P = \frac{q^2\gamma^6}{6\pi\epsilon_0 c}\left(\dot{\boldsymbol{\beta}}^2 - (\boldsymbol{\beta}\times\dot{\boldsymbol{\beta}})^2\right)$ [281] and its manifestly invariant form $P = -\frac{q^2}{6\pi\epsilon_0 c^3}\,a^\mu a_\mu$; the sharply different penalties for acceleration parallel and perpendicular to the velocity (γ^6 against γ^4), which is why linear accelerators are long and circular machines lose energy to synchrotron radiation (Section 65.7.1); the invariance of the total radiated energy but not of its angular distribution [1552].]

65.2.3 Angular distribution and beaming

[Reserved: the angular distribution in the wave zone for arbitrary velocity, its forward collimation into a cone of half-angle $\sim 1/\gamma$, and the distinction between power radiated per unit retarded time and per unit observer time; the observational consequences — the sweep of a synchrotron beam past a detector, and relativistic beaming in astrophysical jets (Chapter 115) [1976].]

65.3 Multipole radiation

65.3.1 Electric dipole radiation

[Reserved: the oscillating dipole, the first radiating system ever solved, treated by Hertz from Maxwell's theory [241]; the near, intermediate and radiation zones; the total power

$P = \ddot{p}^2/6\pi\epsilon_0 c^3$ and the ω^4 scaling that governs both antenna efficiency and Rayleigh scattering (Section 65.6.2); the polarization and $\sin^2\theta$ pattern that Hertz measured with his resonator loop.]

65.3.2 Magnetic dipole and electric quadrupole

[Reserved: the next terms in the expansion, suppressed relative to the electric dipole by one power of (size/λ); their distinct angular patterns and parity, connecting to the selection rules of Chapter 70 and Chapter 105; the observational contrast with gravitation, where charge conservation and momentum conservation forbid the monopole and dipole terms outright and the leading emission is quadrupole (Chapter 46) [1976].]

65.3.3 The multipole expansion of the radiation field

[Reserved: the systematic expansion in vector spherical harmonics, with electric and magnetic multipoles of each order, their radiated power and angular momentum content; the expansion parameter and its failure for sources comparable with the wavelength; the connection to the spherical-harmonic machinery of Chapter 9 and its reappearance for photon angular momentum in Chapter 122 [1976].]

65.4 Antennas

65.4.1 Radiation resistance

[Reserved: the short dipole treated as a circuit element, with the radiated power written as $\frac{1}{2}I^2 R_{rad}$ and $R_{rad} = \frac{2\pi}{3} Z_0 (\ell/\lambda)^2$ in terms of the impedance of free space $Z_0 = 376.730\,\Omega$; the half-wave dipole with its measured $73\,\Omega$; radiation as a genuine loss term in the energy balance of a circuit, which is how Chapter 62 meets this chapter [241, 1976].]

65.4.2 Arrays, directivity and reciprocity

[Reserved: linear arrays and the array factor, directivity and gain in dimensionless units, beam steering by phasing; the reciprocity theorem making the transmitting and receiving patterns of an antenna identical; aperture antennas and the diffraction limit shared with Chapter 67; the practical instruments this section must cover are radio telescopes and interferometric arrays, whose resolution argument is the same one used for stellar interferometry there [1976].]

65.5 Radiation reaction

65.5.1 The Abraham–Lorentz force

[Reserved: the self-force $\boldsymbol{F}_{rad} = \frac{q^2}{6\pi\epsilon_0 c^3}\dot{\boldsymbol{a}}$ obtained from the work–energy balance over a period, by Abraham [313] and Lorentz [356]; the characteristic time $\tau = q^2/6\pi\epsilon_0 m c^3$, equal to about $6.3 \times 10^{-24}\,\text{s}$ for an electron, and the length $c\tau$ far below any distance at which classical electrodynamics is tested; the divergent self-energy of a point charge and the failure of every classical extended-electron model to remove it.]

65.5.2 The Lorentz–Abraham–Dirac equation

[Reserved: Dirac's covariant derivation from conservation of energy–momentum in a tube around the world line, with mass renormalization absorbing the divergent piece [813]; the resulting third-order equation of motion and the Schott term; the statement that this is a

classical, not quantum, renormalization, and the honest note that the key `Dirac:1938` *in this treatise's bibliography is the unrelated large-numbers paper.]*

65.5.3 Runaways, pre-acceleration and the domain of validity

[Reserved: the exponentially growing free solutions $a \propto \mathrm{e}^{t/\tau}$; the acausal pre-acceleration that appears when they are excluded by a final condition; the argument that both pathologies live at times of order τ, where the classical theory has no claim to validity because pair creation and the Compton wavelength intervene (Chapter 100); the standing of radiation reaction as an effective, not fundamental, statement [1552].]

65.5.4 The Landau–Lifshitz reduction of order

[Reserved: the substitution of the zeroth-order equation of motion into the small radiation-reaction term, producing a second-order equation free of runaways and pre-acceleration and accurate to the same order in τ [1552]; the criterion for the reduction to be legitimate; and this equation as the one actually integrated in accelerator and laser-plasma modelling.]

65.5.5 Radiation reaction observed

[Reserved: radiation damping as a routine engineering fact in electron storage rings, where the damping times and the equilibrium emittance follow from the same self-force; and the direct laboratory observation in the collision of a laser-wakefield electron beam with an intense laser pulse, where the measured electron energy loss departs from the radiation-free prediction [2305, 2317]; what these data do and do not distinguish among the competing equations of Sections 65.5.2 and 65.5.4.]

65.6 Scattering

65.6.1 Thomson scattering

[Reserved: a free charge driven by a plane wave re-radiating as an oscillating dipole, giving the frequency-independent cross-section $\sigma_T = \frac{8\pi}{3} r_e^2 = 6.6525 \times 10^{-29}\,\mathrm{m}^2$ with $r_e = q_e^2/4\pi\epsilon_0 m_e c^2$ [332]; the $(1+\cos^2\theta)$ angular dependence for unpolarized light and the complete polarization of the scattered beam at 90 degrees; the observational uses — the solar K-corona, the electron density diagnostics of Chapter 66, and the Thomson scattering that sets the last-scattering surface of Chapter 51.]

Phenomenon 65.3 (Thomson scattering)**.** A beam of radiation traversing a gas is scattered by the free electrons it contains, and at low photon energy the cross-section per electron is independent of the frequency of the radiation:

$$\sigma_{\mathrm{T}} = \frac{8\pi}{3} r_e^2 = 6.6525 \times 10^{-29}\,\mathrm{m}^2\,, \qquad r_e = \frac{q_e^2}{4\pi\epsilon_0 m_e c^2}\,, \tag{65.5}$$

a fixed area attached to the electron and to nothing else [332]. The scattered radiation from an unpolarized beam follows a $1+\cos^2\theta$ angular distribution and is completely polarized at right angles to the beam. Because the cross-section scales as the inverse square of the mass, scattering by nuclei is negligible by six orders of magnitude, and a Thomson measurement counts electrons.

Derivation. A free charge in the field of a plane wave is driven with acceleration $\boldsymbol{a} = q_e \boldsymbol{E}/m_e$, the magnetic force being smaller by v/c and negligible for a non-relativistic response.

Substituting into the Larmor formula Equation (65.1) and taking the time average,

$$\langle P \rangle = \frac{q_e^2}{6\pi\epsilon_0 c^3} \frac{q_e^2 \langle E^2 \rangle}{m_e^2} \,. \tag{65.6}$$

The incident irradiance of the same wave is $I = \epsilon_0 c \langle E^2 \rangle$, and the cross-section is by definition the ratio of the power removed from the beam to the power per unit area carried by it:

$$\sigma_{\mathrm{T}} = \frac{\langle P \rangle}{I} = \frac{q_e^4}{6\pi\epsilon_0^2 c^4 m_e^2} = \frac{8\pi}{3} r_e^2 \,, \tag{65.7}$$

the last step by substituting r_e from Equation (65.5). The mean-square field $\langle E^2 \rangle$ cancels between numerator and denominator, so the answer does not depend on the intensity; and no frequency ever entered, because a free charge has no internal time scale — which is precisely why the cross-section is frequency-independent, and why the failure of that independence, in Section 65.6.4, is a failure of the classical account rather than a refinement of it. □

65.6.2 Rayleigh scattering and the colour of the sky

[Reserved: scattering by bound charges below resonance, with $\sigma \propto \omega^4 \propto \lambda^{-4}$ from the driven-oscillator polarizability [181]; Rayleigh's later demonstration that the molecules of the air themselves suffice, so the blue of the sky needs no dust [286]; the fluctuation account that explains why a dense but uniform medium does not scatter at all [361]; the measured evidence — the spectrum of skylight, its strong polarization at 90 degrees from the Sun, and the reddening of the setting Sun.]

Phenomenon 65.4 (The colour and polarization of the sky)**.** The light of a clear daytime sky arrives from every direction and not merely from the Sun, and it is blue: measured against the solar spectrum it is enriched at short wavelengths in a ratio that rises steeply towards the violet, the scattering coefficient of the air varying as the inverse fourth power of the wavelength [181]. The same light is strongly polarized, most strongly along the great circle 90 degrees from the Sun. The molecules of the air suffice to produce the whole effect; no suspended dust is required, and the observed brightness agrees with what the known number of molecules gives [286]. The setting Sun is red for the complementary reason: the short wavelengths have been scattered out of the long atmospheric path.

Derivation. A molecule is not a free charge but a bound one, and well below its resonances the field of the wave induces a dipole moment $\boldsymbol{p} = \alpha \boldsymbol{E}$ with a polarizability α that is very nearly constant in frequency — this is the same low-frequency limit of the oscillator response that makes the static permittivity of a gas frequency-independent. Its second derivative is not constant: for $\boldsymbol{E} \propto \mathrm{e}^{-\mathrm{i}\omega t}$, $\ddot{\boldsymbol{p}} = -\omega^2 \alpha \boldsymbol{E}$. Feeding this into the dipole form of the Larmor formula, $P = \langle \ddot{p}^2 \rangle / 6\pi\epsilon_0 c^3$, and dividing by the incident irradiance $I = \epsilon_0 c \langle E^2 \rangle$ as in Equation (65.7),

$$\sigma = \frac{\omega^4 \alpha^2}{6\pi\epsilon_0^2 c^4} \propto \omega^4 \propto \frac{1}{\lambda^4} \,. \tag{65.8}$$

The whole of the colour is in the two time derivatives, each contributing one power of ω to the amplitude and hence two to the cross-section: an oscillating dipole radiates in proportion to its *acceleration*, not its displacement. Numerically, blue light near 450 nm is scattered some four times more strongly than red near 650 nm, which is the observed ratio.

The polarization follows from the $\sin^2\theta$ pattern of Phenomenon 65.1. The induced dipole lies in the plane transverse to the incident ray, so an observer looking at right angles to the beam sees only the dipole component perpendicular to the plane containing the incident and scattered rays — the other component points at the observer and radiates nothing that way. The scattered light is therefore completely polarized at 90 degrees for an ideal dipole scatterer, and strongly but not completely polarized in the real sky, where multiple scattering and molecular anisotropy dilute it. □

65.6.3 Mie scattering

[Reserved: the exact solution for a homogeneous sphere of arbitrary size and refractive index [345]; the crossover from the Rayleigh regime to the size-independent, colour-neutral regime that makes clouds and fog white; forward-peaked scattering, the resonances of the series, and their use in aerosol sizing and lidar; the connection to the boundary-value machinery of Chapter 62.]

65.6.4 Where the classical account fails: Compton scattering

[Reserved: the measured wavelength shift $\Delta\lambda = \frac{h}{m_e c}(1-\cos\theta)$ [499], which classical scattering forbids outright since a driven charge re-radiates at the driving frequency; the Klein–Nishina cross-section that replaces Thomson's at photon energies approaching $m_e c^2$ [630]; the statement of the boundary — Thomson scattering is the low-energy limit of a quantum process — with the experiment itself treated in Chapter 69 and the theory in Chapter 100.]

Phenomenon 65.5 (The Compton shift)**.** X-rays scattered from light elements contain, besides radiation of the incident wavelength, a component of *longer* wavelength. The shift grows with the scattering angle, vanishing in the forward direction and reaching its maximum on backscattering, and it depends on nothing else: not on the incident wavelength, and not on the scattering material [499]. Its measured value is

$$\Delta\lambda = \frac{h}{m_e c}\,(1-\cos\theta)\,, \qquad \frac{h}{m_e c} = 2.426\times10^{-12}\,\mathrm{m}\,, \tag{65.9}$$

the constant being the Compton wavelength of the electron [2392]. Classical scattering forbids the effect outright: a charge driven at frequency ω re-radiates at ω, so the scattered wavelength must equal the incident one at every angle.

Derivation. The derivation is not classical, and that is its content. Treat the radiation as quanta of energy hc/λ and momentum h/λ, and let one of them strike an electron of mass m_e at rest, scattering through θ while the electron recoils with momentum $\boldsymbol{p}_e$ and energy E_e. Momentum conservation, squared, gives

$$p_e^2 = \frac{h^2}{\lambda^2} + \frac{h^2}{\lambda'^2} - \frac{2h^2}{\lambda\lambda'}\cos\theta\,, \tag{65.10}$$

and energy conservation gives $E_e = m_e c^2 + hc\,(1/\lambda - 1/\lambda')$. Substituting both into the relativistic relation $E_e^2 = p_e^2 c^2 + m_e^2 c^4$ of Chapter 40, the terms in $m_e^2 c^4$ cancel, as do the terms in $h^2c^2\,(1/\lambda^2 + 1/\lambda'^2)$, and what remains is

$$2m_e c^3 h\left(\frac{1}{\lambda} - \frac{1}{\lambda'}\right) = \frac{2h^2c^2}{\lambda\lambda'}\,(1-\cos\theta)\,. \tag{65.11}$$

Writing $1/\lambda - 1/\lambda' = (\lambda'-\lambda)/\lambda\lambda'$, the factor $\lambda\lambda'$ cancels from both sides and Equation (65.9) follows immediately, with $\Delta\lambda = \lambda' - \lambda$.

Every step used conservation of energy and momentum and nothing else. What it did *not* use is any property of the electron beyond its mass, which is why the shift is independent of the material; and what it *did* use is that the radiation carries energy and momentum in localized quanta with $E = pc$, which classical electrodynamics does not provide. The classical result is recovered where the quantum of energy is negligible against $m_e c^2$: there $\Delta\lambda/\lambda \to 0$, the shift disappears, and the cross-section returns to Thomson's Equation (65.5). □

65.7 Radiation from relativistic charges and in matter

65.7.1 Synchrotron radiation

[Reserved: emission by a charge on a circular orbit, worked out classically by Schott [392] long before it was seen; the critical frequency $\omega_c \simeq \frac{3}{2}\gamma^3 c/\rho$, the broad power-law spectrum below it, the γ^4 energy loss per turn and the linear polarization in the orbital plane; the first direct observation, in the General Electric 70 MeV synchrotron [900]; the two consequences that matter — synchrotron light sources as instruments, and synchrotron emission as the diagnostic of relativistic electrons in cosmic magnetic fields (Chapters 66 and 115).]

Phenomenon 65.6 (Synchrotron radiation)**.** Electrons circulating at relativistic energy in a magnetic field emit light that was first seen directly, as an intense spot on the orbit visible through the glass wall of the vacuum chamber of a 70 MeV machine [900]. The emission is not confined to the orbital frequency and its low harmonics, as a non-relativistic calculation would have it: it is spread over an enormous range of harmonics up to a critical frequency far above the orbital one, it is collimated into a narrow cone tangent to the orbit so that a fixed detector sees a brief flash once per turn, and it is linearly polarized in the plane of the orbit. The energy lost per turn grows very steeply with the electron energy, which is the practical limit on circular electron accelerators and the reason such machines are also the brightest laboratory light sources.

Derivation pending #33

Synchrotron radiation: derive the emission of a charge on a circular orbit at arbitrary speed from the acceleration field and the relativistic generalization of the Larmor formula, obtaining the forward collimation into a cone whose half-angle is the reciprocal of the Lorentz factor, the critical frequency proportional to the cube of that factor divided by the orbit radius, the power-law spectrum below it, the energy radiated per turn, and the degree and sense of the polarization.

65.7.2 Bremsstrahlung

[Reserved: radiation emitted when a charge is deflected in the Coulomb field of a nucleus; the classical impact-parameter treatment and Kramers' spectrum for the continuous X-ray background of a tube [503], with its sharp short-wavelength cutoff at the electron kinetic energy — a quantum feature no classical calculation gives; the Bethe–Heitler cross-section and the radiation length [732]; screening; thermal bremsstrahlung as the X-ray emission of hot plasmas (Chapter 66).]

65.7.3 Cherenkov radiation

[Reserved: the faint blue light emitted by a charge moving faster than the phase velocity of light in a medium, discovered and characterized by Cherenkov [735] after Vavilov's suggestion that it was not fluorescence [749]; the Frank–Tamm theory [792] with the cone condition $\cos\theta_c = 1/n\beta$ and the emitted energy per unit path and unit frequency $\propto (1 - 1/\beta^2 n^2)$; the threshold as a velocity measurement, which is what makes ring-imaging Cherenkov detectors and the water tanks of Chapter 114 work.]

Phenomenon 65.7 (Cherenkov radiation)**.** A fast charged particle traversing a transparent medium makes it glow with a faint blue light. The light is not fluorescence: it is undiminished by quenching agents that suppress fluorescence, it appears in every pure liquid tested, and — decisively — it is not emitted isotropically but along a cone about the particle's track,

sharply directed forward, and polarized in the plane containing the track and the direction of emission [735]. The half-angle of the cone is fixed by the speed of the particle and the refractive index of the medium,

$$\cos\theta_c = \frac{1}{n\beta}\,, \qquad \beta = \frac{v}{c}\,, \tag{65.12}$$

and there is a sharp threshold: no light at all unless $\beta > 1/n$ [792]. Measuring the cone angle therefore measures the speed of the particle directly.

Derivation. The medium is polarized by the passing charge and relaxes behind it, so that each point of the track acts as a momentary source of a wavelet spreading at the phase speed c/n. Consider the wavelet emitted at the moment the particle passes a point O, and the position of the particle a time t later. The wavelet has expanded to a sphere of radius ct/n about O, while the particle has advanced a distance $vt = \beta ct$ along the track. When $\beta n > 1$ the particle has outrun its own wavelets, and the spheres emitted at all earlier points of the track possess a common tangent cone with apex at the particle: the wavelets add in phase on that cone and cancel elsewhere. Its half-angle θ_c, measured from the track, satisfies

$$\cos\theta_c = \frac{ct/n}{\beta ct} = \frac{1}{n\beta}\,, \tag{65.13}$$

which is Equation (65.12), the ratio being independent of t so that the construction is consistent along the whole track.

The threshold is contained in the same formula: a cosine cannot exceed unity, so a real cone exists only for $n\beta \geq 1$, and at $n\beta = 1$ it degenerates to $\theta_c = 0$, the light being emitted straight ahead. Nothing here is peculiar to electrodynamics — it is the construction that gives the Mach cone of a supersonic body — and nothing in it violates relativity, since what is exceeded is the phase speed of light in the medium and never c. □

65.7.4 Transition radiation

[Reserved: radiation emitted when a charge crosses a boundary between media of different permittivity, predicted by Ginzburg and Frank [890]; the total energy proportional to γ, which is what makes transition-radiation detectors the standard particle-identification tool at high γ; foil stacks and the X-ray yield; used by the detectors of Chapter 115.]

Chapter 66

Plasmas and Magnetohydrodynamics

Almost all the visible matter in the universe is plasma: an ionized gas in which the long-range Coulomb interaction, rather than binary collisions, organizes the motion. The name is Langmuir's, and so is the first quantitative study of the collective oscillation that distinguishes the state [611, 639]. This chapter is where the field theory of Chapters 60 and 64 meets the many-body mechanics of Chapter 55 and the continuum dynamics of Chapter 31: the fields move the charges, the charges are the sources of the fields, and the closure of that loop produces behaviour that neither theory shows alone — screening [501], collective oscillation at the plasma frequency, collisionless damping [891], and a fluid that carries its own magnetic field frozen into it [865].

The chapter is placed at the end of Part VII because it is applied classical electrodynamics, and it is placed before the quantum parts because nothing in it needs $\hbar$. Its evidence base is unusually broad and unusually direct: laboratory measurements of Landau damping [1290] and of Alfvén waves in liquid metal [939]; spacecraft in the solar wind whose existence Parker predicted from a hydrodynamic argument [1156] and Mariner II confirmed [1239]; magnetometers inside a reconnecting current sheet [2268]; a laboratory dynamo [1989]; and fusion devices measured against a criterion written down in 1957 [1120] and finally exceeded in inertial confinement [2423]. Standard monographs are [951, 1242].

Derivation pending #34

Plasmas and Magnetohydrodynamics: all derivations of this chapter are pending.

66.1 The plasma state

66.1.1 Ionization equilibrium

[Reserved: the Saha equation for the degree of thermal ionization [470], derived from the chemical equilibrium of Chapter 118; the sharpness of the transition from neutral gas to plasma at temperatures well below the ionization potential; non-thermal ionization by radiation and by impact; the worked cases this treatise needs — the solar photosphere and chromosphere (Chapter 52), the terrestrial ionosphere, and a laboratory discharge.]

66.1.2 Debye screening and the plasma parameter

[Reserved: the screened potential $\varphi \propto \mathrm{e}^{-r/\lambda_D}/r$ with $\lambda_D = \sqrt{\epsilon_0 k_B T/nq^2}$, transferred from the electrolyte theory of Debye and Hückel [501]; the plasma parameter $\Lambda = n\lambda_D^3$ and the definition of a plasma as a medium with $\Lambda \gg 1$, so that collective effects dominate binary

collisions; numerical values — $\lambda_D \approx 74\,\mu\mathrm{m}$ for a fusion plasma at $1 \times 10^{20}/\mathrm{m}^3$ and $10\,\mathrm{keV}$, and metres in the solar wind; the Coulomb logarithm $\ln\Lambda$ that will appear in every transport coefficient of Section 66.3.4.]

66.1.3 Quasi-neutrality and the plasma frequency

[Reserved: quasi-neutrality as a consequence of screening rather than an assumption, valid on scales above λ_D; the restoring force of a rigid charge displacement and the electron plasma frequency $\omega_p = \sqrt{nq_e^2/\epsilon_0 m_e}$, i.e. $f_p \approx 8.98\sqrt{n}$ hertz with n in inverse cubic metres, identified experimentally by Langmuir [611] and analysed with Tonks [639]; the three defining conditions of the plasma state (many particles per Debye sphere, system size above λ_D, collision time above $1/\omega_p$).]

Phenomenon 66.1 (Collective oscillation at the plasma frequency)**.** An ionized gas does not respond to a disturbance the way a neutral gas does. Probed in a discharge, the electrons are found to oscillate collectively at a frequency fixed by their number density alone — not by the wavelength of the disturbance, not by the gas species, not by the geometry of the vessel, and, at long wavelength, not by the temperature [611, 639]. The frequency is

$$\omega_p = \sqrt{\frac{nq_e^2}{\epsilon_0 m_e}}\,, \tag{66.1}$$

and it is this collective response, rather than any property of the individual collisions, that makes the ionized gas a distinct state of matter.

Derivation. Consider a slab of plasma and displace every electron rigidly through a small distance x along the slab normal, leaving the much heavier ions in place. The displacement uncovers a layer of positive charge on one face and deposits a layer of negative charge on the other, of surface density $\pm nq_e x$, and a pair of such sheets produces a uniform field between them of magnitude

$$E = \frac{n\,|q_e|\,x}{\epsilon_0}\,, \tag{66.2}$$

directed so as to pull the electrons back. The equation of motion of each electron is then

$$m_e \ddot{x} = -\,|q_e|\,E = -\frac{nq_e^2}{\epsilon_0}x\,, \tag{66.3}$$

which is simple harmonic motion at exactly the frequency Equation (66.1). Three features of the observation are already in this one line. The frequency does not depend on the amplitude, because the restoring force is linear in the displacement; it does not depend on the wavelength, because no gradient of anything entered — the correction that does introduce one comes from the thermal pressure and is the Bohm–Gross term of Section 66.2.1; and it depends on the ion mass not at all, the ions serving only as a stationary neutralizing background.

The same ω_p governs the transparency of metals and the reflection of radio waves treated in Chapter 62, and evaluating Equation (66.1) with the CODATA values of the electron charge and mass [2392] gives the working form $f_p = \omega_p/2\pi \approx 8.98\sqrt{n}$ hertz with n in inverse cubic metres. □

66.1.4 Where plasmas are found

[Reserved: the parameter map in density and temperature, all in SI, spanning the interstellar medium, the solar corona and wind, the ionosphere, laboratory discharges, tokamak and inertial-fusion plasmas, and the degenerate interiors of Chapter 49; the direct evidence for

each — the ionospheric reflection of radio waves demonstrated by Appleton and Barnett [517], and the Zeeman splitting that showed sunspots to be strongly magnetized [342], which is the founding observation of astrophysical magnetism.]

Phenomenon 66.2 (Sunspots are strongly magnetized)**.** Spectral lines formed in a sunspot are split into components whose separation is measurable with a grating spectrograph, and whose polarization is the signature that identifies the splitting as magnetic: viewed along the field the outer components are circularly polarized with opposite handedness, viewed across it they are linearly polarized. The splitting is absent in the surrounding photosphere and largest at the darkest part of the spot. The inferred field strengths are of order a tenth of a tesla, thousands of times the field at the Earth's surface [342]. This is the first measurement of a magnetic field outside the Earth, and it is the observation on which the whole of astrophysical magnetism rests.

Derivation. The inference from a measured splitting to a field strength needs only the classical theory of a bound electron in a magnetic field. Let the electron be bound isotropically with natural frequency ω_0 and place the field along z. Writing the transverse displacement as the complex combination $u = x + \mathrm{i}y$, the equation of motion $m_e\ddot{\boldsymbol{r}} = -m_e\omega_0^2\boldsymbol{r} + q_e\dot{\boldsymbol{r}} \times \boldsymbol{B}$ becomes

$$\ddot{u} + \mathrm{i}\frac{q_e B}{m_e}\dot{u} + \omega_0^2 u = 0\,, \tag{66.4}$$

and substituting $u = \mathrm{e}^{-\mathrm{i}\omega t}$ gives $\omega^2 - (q_e B/m_e)\,\omega - \omega_0^2 = 0$. For fields weak enough that $|q_e|\,B/m_e \ll \omega_0$ — which holds by many orders of magnitude for optical lines in any attainable field — the two roots are

$$\omega \simeq \pm\omega_0 + \frac{q_e B}{2m_e}\,, \qquad |\Delta\omega| = \frac{|q_e|\,B}{2m_e}\,. \tag{66.5}$$

The two transverse modes are circular motions of opposite sense, shifted up and down by $|\Delta\omega|$; the third mode, along z, feels no magnetic force and is unshifted. Hence a triplet, with the outer components circularly polarized when the line of sight is along $\boldsymbol{B}$ and linearly polarized when it is across — which is exactly the polarization pattern that identifies the effect and, with it, the direction of the field. Inverting Equation (66.5),

$$B = \frac{4\pi m_e}{|q_e|}\,\Delta\nu = \frac{4\pi m_e c}{|q_e|\,\lambda^2}\,\Delta\lambda\,, \tag{66.6}$$

so a splitting measured in metres of wavelength returns a field in tesla. The classical account is not the whole story — real lines split into patterns richer than a triplet, for reasons that are quantum and belong to the atomic chapters — but the magnitude, and therefore the inferred field, is the classical one. □

66.2 Waves in a plasma

66.2.1 Langmuir waves

[Reserved: electrostatic electron oscillations, their near-dispersionless character at long wavelength, and the Bohm–Gross correction $\omega^2 = \omega_p^2 + 3k^2v_{th}^2$ from thermal pressure [925]; the standing puzzle that the fluid treatment gives no damping at all, resolved kinetically in Section 66.3.2; the observation of these waves as the electron-plasma oscillations first reported in [611, 639].]

66.2.2 Electromagnetic waves and the cutoff

[Reserved: the transverse dispersion relation $\omega^2 = \omega_p^2 + c^2k^2$, the cutoff at ω_p below which the plasma reflects, and the resulting refractive index below unity, whose consequences for phase and group velocity connect to Chapter 62; the ionospheric application, with f_p of order 9 MHz for $n \approx 1 \times 10^{12}/\mathrm{m}^3$, which is why long-distance short-wave radio works and why satellite links must exceed the cutoff [517]; interferometry and cutoff reflectometry as plasma density diagnostics.]

Phenomenon 66.3 (The ionosphere reflects radio waves below a critical frequency)**.** Radio waves directed upwards return to the ground if their frequency is below a critical value and are lost into space if it is above. The returning wave was identified as such by beating it against the ground wave and varying the transmitter frequency: the resulting fringes give the difference in path length, and hence the height of the reflecting layer, which is found to lie of order a hundred kilometres above the ground and to change between day and night [517]. The critical frequency is not a property of the apparatus but of the medium, and it varies with the hour, the season and the solar cycle. There is therefore a permanent ionized layer in the upper atmosphere, and the critical frequency measures its electron density.

Derivation. The reflection itself is the free-carrier cutoff derived in Chapter 62: a wave of angular frequency ω in a medium of free electrons obeys $\omega^2 = \omega_p^2 + c^2k^2$, so k is imaginary and the wave cannot propagate wherever $\omega < \omega_p$. A wave sent upwards into a layer whose density increases with height therefore travels freely until it reaches the level at which ω_p has risen to ω, and is returned from there — so the reflection height depends on the frequency, and the highest frequency returned at vertical incidence is the plasma frequency of the densest part of the layer.

What makes this a measurement is the numerical size of Equation (66.1). Evaluating it with the CODATA electron charge and mass [2392],

$$f_p = \frac{1}{2\pi}\sqrt{\frac{nq_e^2}{\epsilon_0 m_e}} \approx 8.98\sqrt{n}\ \mathrm{Hz}\,, \tag{66.7}$$

with n in inverse cubic metres; so a critical frequency of a few megahertz corresponds to an electron density of order $10^{12}/\mathrm{m}^3$, which is what the ionosphere has. The inverse statement is the diagnostic: sweeping the transmitter frequency and recording the echo delay returns the density as a function of height, and the same principle, applied at much higher frequency, is cutoff reflectometry in a laboratory plasma. □

66.2.3 Magnetized waves and ion-acoustic waves

[Reserved: propagation across and along a background field, the ordinary and extraordinary modes, cyclotron resonances at $\omega_c = qB/m$, Faraday rotation as a measurement of the line integral of $n_e B_\parallel$ used throughout radio astronomy; the ion-acoustic mode with speed $\sqrt{k_B T_e/m_i}$, sustained by electron pressure and ion inertia, and its Landau damping when $T_e \sim T_i$; the classification of the many modes as a warning that "the plasma dispersion relation" is a family, not a formula [1242].]

66.3 Kinetic theory and Landau damping

66.3.1 The Vlasov equation

[Reserved: the collisionless kinetic equation for $f(\boldsymbol{x}, \boldsymbol{v}, t)$ with the self-consistent field [823], its standing as the leading term of the BBGKY hierarchy of Chapter 55, and the sense in which it is a mean-field rather than an exact description; its conservation laws and the

infinity of Casimir invariants; the Jeans-type equilibria that any function of the constants of motion provides.]

66.3.2 Landau damping

[Reserved: Landau's solution of the initial-value problem by Laplace transform and the contour prescription that produces a damping rate for a wave in a collisionless medium [891]; the damping proportional to $\partial f/\partial v$ at the phase velocity, and the resonant-particle reading (particles slightly slower than the wave gain energy, slightly faster ones lose it); the inverse effect — growth on an inverted distribution — as the bump-on-tail instability; why this is not thermodynamically irreversible, the entropy being carried into fine velocity-space structure, with echo experiments as the evidence.]

66.3.3 The observation

[Reserved: Malmberg and Wharton's measurement of the spatial damping of electron plasma waves in a collisionless column, with the measured damping length agreeing with Landau's prediction over more than a decade in wavenumber [1290]; the plasma-wave echo as the direct demonstration that the information is not lost; and the methodological point that a purely mathematical prescription for a contour was confirmed as physics eighteen years after it was written.]

Phenomenon 66.4 (Waves damp in a collisionless plasma)**.** An electron plasma wave launched into a column so tenuous that a particle crosses it many times between collisions is nevertheless damped, and damped strongly. The damping is spatial and exponential, its rate depends on the wavenumber, and it persists as the collision frequency is reduced — so it is not a collisional effect at all. The measured damping length agrees with the collisionless prediction over more than a decade in wavenumber [1290], confirming an analysis that had been produced eighteen years earlier from a contour prescription in a Laplace transform [891]. A dissipative process therefore occurs in a system whose equations of motion are exactly time-reversible and conserve entropy.

Derivation pending #35

Landau damping: solve the linearized Vlasov–Poisson initial-value problem by Laplace transform, analytically continue the dielectric function below the real axis to obtain the correct prescription for the singularity at the phase velocity, and derive the damping rate proportional to the slope of the distribution function there. Then give the resonant-particle reading — particles slightly slower than the wave gain energy and slightly faster ones lose it, and a Maxwellian has more of the former — and explain why the process is not thermodynamically irreversible, the information being carried into fine structure in velocity space and recoverable as a plasma-wave echo.

66.3.4 Collisions and transport

[Reserved: the Fokker–Planck description of small-angle Coulomb scattering with the Rosenbluth potentials [1129]; the $T^{3/2}$ scaling and the Spitzer resistivity [1027], from which a fusion-grade plasma is a better conductor than copper; collisional relaxation times and their ordering $\tau_{ee} < \tau_{ii} < \tau_{ie}$; the full transport coefficients — resistivity, viscosity, thermal conduction, all anisotropic in a magnetic field — as tabulated by Braginskii [1303], connecting to Chapter 124.]

66.4 Single-particle motion

66.4.1 Gyration and the guiding centre

[Reserved: the helical orbit in a uniform field, the gyrofrequency and gyroradius in SI, and the guiding-centre expansion in the small parameter ρ_L/L [951]; the separation of fast gyration from slow drift as an averaging procedure of the kind used in Chapter 32; the limits of the expansion where the field varies on the gyroradius scale.]

66.4.2 Drifts

[Reserved: the $\boldsymbol{E}\times\boldsymbol{B}$ drift, charge- and mass-independent and therefore not a current; the grad-B, curvature and polarization drifts, which are charge dependent and do drive currents; the resulting ring current and the diamagnetism of a plasma; worked numbers for the magnetosphere; the general statement that a drift is a force divided by qB crossed with $\hat{\boldsymbol{b}}$ [951, 1242].]

66.4.3 Adiabatic invariants and magnetic mirrors

[Reserved: the magnetic moment $\mu = mv_\perp^2/2B$ as an adiabatic invariant in the exact sense of Chapter 23, together with the longitudinal and flux invariants; magnetic mirroring, the loss cone and its opening angle; the mirror machine as a confinement scheme and its failure modes; the observational confirmation in the trapped radiation belts discovered by Van Allen [1176], whose particle populations are the loss-cone distribution made visible.]

Phenomenon 66.5 (Magnetic trapping: the radiation belts)**.** Geiger counters carried above the atmosphere found, instead of the expected slow decline of the cosmic-ray rate, counting rates rising by orders of magnitude in two toroidal regions encircling the Earth [1176]. The particles are not passing through: they are trapped, bouncing along the field lines between turning points in the two hemispheres on time scales of a second, and the belts persist indefinitely although nothing material confines them. What confines them is the geometry of the field, and the confinement is incomplete in a definite way — particles whose velocity is too nearly parallel to the field are not turned back and are lost into the atmosphere at the foot of the line, which is why the trapped distributions are empty inside a cone in velocity space.

Derivation. A particle gyrating in a field that varies slowly over one gyroradius and one gyroperiod conserves its magnetic moment

$$\mu = \frac{mv_\perp^2}{2B}\,, \tag{66.8}$$

as an adiabatic invariant. A static magnetic field does no work, so the speed v is constant as well. Writing the pitch angle α by $v_\perp = v\sin\alpha$, constancy of μ and of v gives

$$\frac{\sin^2\alpha}{B} = \text{constant along the orbit}\,. \tag{66.9}$$

As the particle moves into a region of stronger field, $\sin^2\alpha$ must rise in proportion; when it reaches unity the velocity is entirely perpendicular, the parallel motion has stopped, and the particle is reflected. Starting from a field B_0 with pitch angle α_0, the mirror point is therefore where

$$B_m = \frac{B_0}{\sin^2\alpha_0}\,. \tag{66.10}$$

A dipole field grows towards its poles, so every particle whose B_m lies below the field at the top of the atmosphere turns around before reaching it and is trapped, bouncing between the conjugate mirror points — which is the observed configuration.

Equation (66.10) also fixes what escapes. If $B_{\max}$ is the field at the atmospheric foot of the line, the particles lost are those with

$$\sin^2 \alpha_0 < \frac{B_0}{B_{\max}}\,, \tag{66.11}$$

a cone about the field direction whose half-angle depends only on the ratio of the two field strengths and not at all on the particle's charge, mass or energy. That is the loss cone, and its emptiness in the measured distributions is the signature that the trapping mechanism is this one and no other. The same argument, with $B_{\max}$ set by a coil rather than by the Earth, is the magnetic mirror as a laboratory confinement scheme — and Equation (66.11) is equally the statement of why that scheme leaks. □

66.4.4 Fermi acceleration

[Reserved: stochastic energy gain from collisions with moving magnetic irregularities [929], the second-order $(v/c)^2$ mechanism and its first-order diffusive-shock refinement; the power-law spectrum produced, and its match to the observed cosmic-ray spectrum treated in Chapter 115; the same mechanism as the origin of the relativistic electrons whose synchrotron emission is discussed in Chapter 65.]

66.5 Magnetohydrodynamics

66.5.1 The MHD equations and their validity

[Reserved: the one-fluid equations obtained by taking moments of the kinetic equation and eliminating the displacement current — mass continuity, the momentum equation with the $\boldsymbol{J}\times\boldsymbol{B}$ force, an Ohm's law, and the induction equation [951, 1303]; the ordering assumptions (slow compared with ω_p and ω_c, large compared with λ_D and ρ_L, non-relativistic); the closure problem and the standard ideal, resistive and Hall variants; the decomposition of the Lorentz force into magnetic pressure $B^2/2\mu_0$ and tension.]

66.5.2 Flux freezing and the magnetic Reynolds number

[Reserved: Alfvén's theorem that in a perfectly conducting fluid the magnetic flux through any comoving surface is constant, so field lines are material lines [865]; the induction equation as advection plus diffusion, with the magnetic Reynolds number $R_m = UL/\eta_m$ measuring their ratio; the enormous values of R_m in astrophysics and the consequence that field topology is conserved except in thin layers — which is the whole subject of Section 66.6.1.]

66.5.3 Alfvén and magnetosonic waves

[Reserved: the transverse wave on a field line with speed $v_A = B/\sqrt{\mu_0\rho}$, predicted by Alfvén [865] against considerable scepticism; the fast and slow magnetosonic modes and the Friedrichs diagram of their anisotropic phase speeds; the first laboratory detection, in liquid mercury [939]; and the ubiquitous detection in the solar wind and the magnetosphere, where Alfvénic fluctuations dominate the turbulent spectrum.]

Phenomenon 66.6 (Alfvén waves)**.** A conducting fluid threaded by a magnetic field supports a transverse wave that ordinary hydrodynamics does not have: the field lines

behave as stretched strings, and a displacement propagates along them at

$$v_A = \frac{B}{\sqrt{\mu_0 \rho}}\,, \tag{66.12}$$

independent of the wavelength and of the amplitude, with the fluid velocity and the field perturbation transverse to the background field and correlated with one another. The mode was predicted before it was seen [865], first detected in liquid mercury in a strong magnet [939], and is now measured continuously by spacecraft in the solar wind, where fluctuations of exactly this character dominate the turbulent spectrum.

Derivation pending #36

Alfvén waves: linearize the ideal magnetohydrodynamic equations about a uniform field and a uniform density, retaining the tension part of the Lorentz force, and obtain the transverse mode propagating along the field at the speed above, together with the fast and slow magnetosonic modes and their anisotropic phase speeds. The tension reading should be derived as well — the magnetic force decomposed into a pressure and a tension along the field lines, whence the wave speed is the tension divided by the mass per unit length in exactly the form of a wave on a string.

66.5.4 Equilibria and the pinch

[Reserved: magnetostatic equilibrium $\boldsymbol{J} \times \boldsymbol{B} = \nabla p$, force-free fields, and the plasma beta as the ratio of kinetic to magnetic pressure; the Bennett relation for a self-pinching current channel [731]; the Grad–Shafranov equation for axisymmetric equilibria [1149] and the flux-surface geometry of a tokamak; the safety factor and the equilibria that Section 66.7 will test for stability.]

66.6 Magnetic reconnection

66.6.1 The Sweet–Parker layer

[Reserved: the problem stated — ideal MHD conserves topology, yet solar flares release magnetic energy in minutes; Sweet's neutral point [1163] and Parker's quantitative current-sheet model [1125], giving a reconnection rate scaling as $R_m^{-1/2}$; the arithmetic showing that this rate is still orders of magnitude too slow for a flare, which is the reason the subject did not end in 1958.]

66.6.2 Petschek and fast reconnection

[Reserved: Petschek's slow-shock geometry with a much shorter diffusion region and a rate scaling only as $1/\ln R_m$ [1293]; the objection that the configuration is not sustained with uniform resistivity; the modern resolutions — plasmoid instability of thin sheets, and collisionless Hall and kinetic terms in Ohm's law; the honest statement that the rate problem is settled observationally before it is settled theoretically.]

66.6.3 Reconnection observed

[Reserved: Dungey's open magnetosphere, in which dayside reconnection with the interplanetary field drives the whole convection cycle and the aurora [1203]; in situ measurement of the electron diffusion region by the four Magnetospheric Multiscale spacecraft, resolving the non-ideal electric field and the energy conversion directly [2268]; solar flares and coronal

mass ejections as the large-scale case; laboratory reconnection experiments; the quantities to tabulate are the reconnection rate, the layer thickness and the partition of released energy.]

Phenomenon 66.7 (Magnetic reconnection)**.** Magnetic field topology, which ideal magnetohydrodynamics conserves exactly, is observed to change. The evidence is of three kinds. Solar flares release in minutes an amount of magnetic energy that took days to accumulate, and do so in configurations where oppositely directed fields have been pushed together. The whole convection cycle of the magnetosphere, and with it the aurora, switches on when the interplanetary field turns southward and off when it turns northward — a dependence on the *direction* of the external field that makes sense only if the two field systems join [1203]. And four spacecraft flying in formation through the dayside boundary traversed an electron diffusion region directly, measuring there an electric field that no ideal Ohm's law permits, together with the crescent-shaped electron distributions and the conversion of electromagnetic energy into particle energy that accompany it [2268]. The ideal description therefore fails, and it fails in thin layers while remaining excellent everywhere else.

Derivation pending #37

Magnetic reconnection: derive the Sweet–Parker current sheet from mass and flux conservation across a resistive layer, obtaining a reconnection rate scaling as the inverse square root of the magnetic Reynolds number, and show by substituting coronal values that this rate is orders of magnitude too slow for a flare. Then derive Petschek's slow-shock geometry with its much weaker logarithmic dependence, and state honestly which of the modern resolutions — plasmoid instability of thin sheets, or the collisionless Hall and pressure-tensor terms in Ohm's law — the measurements support.

66.7 Instabilities and turbulence

66.7.1 Ideal MHD instabilities

[Reserved: the interchange or flute mode as the magnetic analogue of the Rayleigh–Taylor instability [221]; kink and sausage modes of a current channel and the Kruskal–Schwarzschild analysis [1046]; the Kruskal–Shafranov limit on the safety factor; the energy principle that decides linear stability from the sign of a quadratic form [1135], which is the Hamiltonian stability criterion of Chapter 22 in field form.]

66.7.2 Microinstabilities and turbulence

[Reserved: velocity-space free energy driving the two-stream and Buneman instabilities [1138] and the drift-wave family; anomalous transport as the practical consequence, since observed confinement times fall far below the collisional predictions of Section 66.3.4; MHD turbulence, its anisotropic cascade and the measured spectra in the solar wind, connecting to the general turbulence discussion of Chapter 31.]

66.8 Fusion and confinement

66.8.1 The Lawson criterion

[Reserved: Lawson's power balance between fusion output, bremsstrahlung loss and the energy invested in heating, giving a condition on $n\tau_E$ [1120]; the modern triple-product form $nT\tau_E$ and the ignition condition when alpha heating suffices; numerical targets for

deuterium–tritium, with $n\tau_E \gtrsim 1 \times 10^{20}\,\mathrm{s/m^3}$ at about 10 keV; the input cross-sections come from Chapter 107, and the bremsstrahlung loss from Chapter 65.]

66.8.2 Magnetic confinement

[Reserved: the tokamak and stellarator as the two solutions to the rotational-transform problem posed by Section 66.5.4; confinement scaling laws as empirical relations rather than theory; heating and current drive; the deuterium–tritium campaigns of JET, with 16 MW of fusion power and a measured gain below unity [1977]; what remains between that and a reactor, stated without advocacy.]

66.8.3 Inertial confinement and ignition

[Reserved: compression of a fuel capsule by laser ablation, the ρR criterion as the inertial form of Lawson's condition, and the hydrodynamic instabilities that limit compression; the National Ignition Facility result of target gain above unity — fusion yield exceeding the laser energy delivered to the target [2423] — with the precise statement of what is and is not included in that ratio, since the distinction between target gain and wall-plug gain is where most public confusion lives.]

66.9 Space and astrophysical plasmas

66.9.1 The solar wind

[Reserved: Biermann's inference from the anti-solar direction of comet ion tails that a continuous corpuscular stream must exist [976]; Parker's hydrodynamic solution showing that a hot corona cannot be in hydrostatic equilibrium and must expand supersonically [1156], with the transonic critical point; the confirmation by Mariner II, which measured a continuous flow at 400 km/s [1239]; the Parker spiral geometry of the interplanetary field; the quantities to tabulate — speed, density, temperature and field strength at 1 astronomical unit.]

Phenomenon 66.8 (The solar wind)**.** The Sun's atmosphere does not stay bound to it. A continuous supersonic stream of ionized gas flows outward past the Earth and beyond. Three independent observations establish it. The ion tails of comets point almost directly away from the Sun whatever the comet's own motion, and respond to solar activity within about a day — which radiation pressure alone cannot produce and a corpuscular stream can [976]. A corona at the observed temperature cannot be in hydrostatic equilibrium against the low pressure of the interstellar medium, and must instead expand, passing through a critical point to supersonic speed [1156]. And a spacecraft carrying a plasma analyser on the way to Venus measured the flow directly and continuously — never absent over months — at speeds of order 400 km/s [1239].

Derivation pending #38

The solar wind: derive Parker's transonic solution. Integrate the steady, spherically symmetric momentum equation for an isothermal corona in the Sun's gravitational field, exhibit the critical point at which the flow speed equals the sound speed, show that of the family of solutions only the transonic branch matches both a subsonic base and a vanishing pressure at large distance, and obtain the asymptotic speed. The contrasting hydrostatic solution must be shown to leave a finite pressure at infinity, which is the argument that rules it out.

66.9.2 The magnetosphere

[Reserved: the bow shock, magnetopause and magnetotail as the solution of the solar wind's encounter with a dipole; Dungey convection [1203]; the trapped belts [1176]; substorms and the aurora as the visible end of the chain; geomagnetic storms as a measured, dated phenomenon with technological consequences.]

66.9.3 The dynamo

[Reserved: the question posed by Larmor — how a rotating body can maintain a magnetic field against ohmic decay [460]; Cowling's anti-dynamo theorem forbidding a strictly axisymmetric solution [717], which is why the subject is hard; Parker's cyclonic α-effect and mean-field dynamo [1078]; the solar cycle and its polarity reversal as the observational anchor, resting on Hale's sunspot magnetograms [342]; the geodynamo, its numerical demonstration with a simulated reversal [1900], and the decisive laboratory evidence from the Riga liquid-sodium experiment, where a self-sustaining field eigenmode grew from a purely hydrodynamic flow [1989].]

Chapter 67

Experiment: Wave Optics

This chapter is the evidence for Chapter 61. It collects the measurements that established, quantitatively and against a competing theory of equal standing, that light is a wave: Young's two-slit fringes and the first wavelength measurements [48, 49]; Fresnel's prize memoir on diffraction with the bright spot at the centre of a circular shadow that Poisson produced as a refutation and Arago found [78]; the Fresnel–Arago laws, whose failure of two perpendicularly polarized beams to interfere forced the conclusion that the vibration is transverse [63, 90]; the ruled grating turned from a curiosity into the primary instrument of wavelength metrology by Fraunhofer [69, 74] and Rowland [222]; and the interferometers, from Michelson's [236] through his stellar beam interferometer [469, 479] to the kilometre-scale instruments of Chapter 47.

Two of these experiments are shared with other chapters and are treated here only for their optics: Fizeau's moving-water measurement [133] and the Michelson–Morley null result [236] are read as interferometry here, and read as evidence about the aether and about relativity in Chapter 37. The chapter closes with the modern repetitions, which is where the classical account runs out: the same two-slit pattern builds up one particle at a time, for photons [360, 1774], for electrons [1212, 1826] and for molecules [1966] — a result that belongs to Parts VIII and IX and is stated here only to mark precisely where wave optics stops being the whole story. When written out, every experiment below will use the structured `experiment` environment, with apparatus, procedure, observations (numbers and uncertainties in SI units), interpretation and primary references given in full. The standard modern reference for the theory being tested is [1967].

Derivation pending #39

Experiment: Wave Optics: all derivations of this chapter are pending.

67.1 The predictions under test

67.1.1 Two accounts of light

[Reserved: the state of the question about 1800. On one side the emission theory of [18], which explains rectilinear propagation and reflection immediately, refraction by an increase of speed in the denser medium, and colour by particle species; on the other the wave account of [15] with its secondary wavelets, which explains refraction by a decrease of speed and gives double refraction in calcite. The prior anomaly both had to face is Grimaldi's diffraction — light bending into the geometrical shadow of a small obstacle [8] — and Newton's own rings [18], periodic in a way no particle account made natural.]

67.1.2 What separates them

[Reserved: the discriminating predictions, stated before any apparatus. Only a wave theory predicts (i) fringes whose spacing scales as $\lambda D/d$ and vanishes when either slit is blocked, (ii) a bright point at the exact centre of the shadow of a circular obstacle, (iii) a speed of light lower *in the denser medium, measured by Foucault and treated in Chapter 61, and (iv) a diffraction limit on the resolving power of every instrument. The transversality of the vibration is not predicted by either and is forced by the polarization data of Section 67.4.]*

67.2 Young: the double slit (1801–1804)

67.2.1 Apparatus and procedure

[Reserved: Young's arrangement — sunlight admitted through a pinhole in a window shutter, a narrow card or a pair of slits splitting the beam, and the fringe pattern received on a screen or observed with a lens [49]; the coherence requirement met by the common source, which is why the pinhole cannot simply be widened; the slit separation, slit-to-screen distance and fringe spacing as the three lengths that must be measured, all in metres; the later, cleaner realizations that removed the finite source size — Fresnel's biprism and mirrors [78] and Lloyd's single-mirror geometry [102], which adds the extra half-wave of reflection.]

67.2.2 Observations

[Reserved: the table to be given — slit separation d, screen distance D and measured fringe spacing Δy with uncertainties, and the wavelength inferred from $\lambda = d\,\Delta y/D$; Young's own values, about 700 nm *at the red end and* 420 nm *at the violet end, to be compared with the modern sodium doublet at* 589.3 nm*; the disappearance of the pattern when either slit is covered; and the checks that the spacing is proportional to D and inversely proportional to d [49].]*

Phenomenon 67.1 (The wavelength of light is a measurable length)**.** The two-slit pattern turns light into a measurable length. Three quantities are read with a scale — the separation d of the apertures, the distance D to the screen, and the spacing Δy of the fringes — and they return a wavelength. The values so obtained run from about 700 nm at the red end of the visible spectrum to about 420 nm at the violet end [49], which is why colour is a physical variable and not merely a sensation. Two checks establish that the reading is what it is claimed to be: the fringe spacing is found proportional to D and inversely proportional to d, as it must be if it is a path-difference effect, and the pattern vanishes entirely when either aperture is covered [48, 49].

Derivation. The theory being used is the two-beam superposition of Chapter 61, which gives a fringe spacing $\lambda D/d$ for $d \ll D$. What the experiment needs is that relation inverted,

$$\lambda = \frac{d\,\Delta y}{D}\,, \tag{67.1}$$

in which every quantity on the right is a length measured with a scale, so the wavelength is obtained without any optical constant, any calibration against another wavelength, or any assumption about the speed of light.

The error budget follows from Equation (67.1) by taking logarithms and differentiating, which turns a product into a sum of relative uncertainties combining in quadrature:

$$\frac{u(\lambda)}{\lambda} = \sqrt{\left(\frac{u(d)}{d}\right)^2 + \left(\frac{u(\Delta y)}{\Delta y}\right)^2 + \left(\frac{u(D)}{D}\right)^2}\,. \tag{67.2}$$

The dominant term is the fringe spacing, which is the smallest of the three lengths, and the standard remedy is in the geometry of the apparatus rather than in the instrument: measure across N fringes and divide, which divides the relative uncertainty of Δy by N while leaving the other two terms alone. This is the reason the two-slit determination remains respectable although each individual reading is crude. □

67.2.3 Interpretation

[Reserved: the general law of interference as Young stated it [48] — superposition of two disturbances from a common source with a path difference — and the inference that light adds amplitudes, not intensities, so that light plus light can give darkness; why no emission theory accommodates this; the residual objections of the period (the finite width of the fringes, the "inflection" account) and what answered them; the theory tested lives in Chapter 61.]

67.3 Fresnel and the Poisson–Arago spot (1818–1819)

67.3.1 Apparatus and procedure

[Reserved: the prize memoir's apparatus [78] — a point source, an opaque circular disc a few millimetres across, and a screen at a distance of order a metre, with fringe positions read on a micrometer; the requirement of a smooth circular edge, which is what makes the null prediction sharp; Fresnel's own half-period zone construction, combining Huygens' wavelets with Young's interference, and the integral form of it given in Chapter 61.]

67.3.2 Observations

[Reserved: the bright point at the exact centre of the shadow of the disc, with an on-axis intensity equal to the unobstructed value for an ideal disc; the measured radii of the external and internal fringes of a straight edge and of the disc, compared point by point with Fresnel's integrals; the tabulated quantities are fringe positions in metres against source and screen distances, with the wavelength as the fitted parameter [78].]

67.3.3 Interpretation

[Reserved: the episode's methodological content, which is why it is in this treatise. Poisson derived the central bright spot from Fresnel's own theory as a reductio; Arago performed the observation and found it; the prediction was therefore made by an opponent and confirmed — the cleanest instance in classical physics of a theory surviving a test designed to kill it [78]. The quantitative content is that the diffraction integral, not merely the qualitative idea of waves, is correct.]

67.4 The Fresnel–Arago laws and transversality

67.4.1 Apparatus and procedure

[Reserved: polarizers by reflection at the Brewster angle [58] and by double refraction in calcite; the two-slit or biprism interference arrangement of Section 67.2.1 with an independently controllable polarization in each arm; the analyser and the photometric comparison method; Malus's original apparatus for the $\cos^2$ *law [54].]*

67.4.2 Observations

[Reserved: the four Fresnel–Arago laws as measured statements [63] — two beams polarized at right angles never interfere; two beams polarized in the same plane interfere as unpolarized light does; two beams from perpendicularly polarized components of unpolarized light do not interfere even when brought to a common polarization; and they do if they derive from a single polarized beam. Alongside them, the transmitted intensity $I = I_0 \cos^2\theta$ [54] and the polarizing angle $\tan\theta_B = n$ [58], both with tabulated angles and intensities and their uncertainties.]

Phenomenon 67.2 (Perpendicularly polarized beams do not interfere)**.** Two beams derived from a common source and superposed do not always produce fringes. If they are linearly polarized at right angles to one another, no fringes appear at any path difference whatever: the screen shows only the sum of the two intensities. If they are polarized in the same plane, they interfere exactly as unpolarized light does. Two beams obtained as the perpendicularly polarized components of unpolarized light do not interfere even after being brought back to a common plane of polarization — but two beams obtained the same way from a single *polarized* beam do [63]. These four statements are measured facts, and taken together they force a conclusion about the nature of the disturbance that no amount of interference data alone could give.

Derivation. What a screen records is the time-averaged square of the total disturbance. If the two beams contribute $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ at a point,

$$I \propto \left\langle |\boldsymbol{E}_1 + \boldsymbol{E}_2|^2 \right\rangle = \left\langle E_1^2 \right\rangle + \left\langle E_2^2 \right\rangle + 2\left\langle \boldsymbol{E}_1 \cdot \boldsymbol{E}_2 \right\rangle , \tag{67.3}$$

and all the interference lives in the last term. It is a scalar product, so it vanishes identically — at every path difference, for every amplitude — whenever the two disturbances are orthogonal as vectors. That is the first law, and with it the second: parallel disturbances give the full cross term back.

The conclusion about transversality is now forced. A longitudinal disturbance in a beam has exactly one direction available to it, along the ray, so two beams travelling together could never be orthogonal and the first law could never be observed. The observation therefore requires the disturbance to lie in the plane perpendicular to the propagation direction, where two independent directions exist — and it requires nothing else, which is what makes the inference so strong.

The third and fourth laws add the second ingredient, coherence. The cross term in Equation (67.3) survives averaging only if the phase relation between the two beams is steady. In unpolarized light the two transverse components carry no fixed phase relation, so $\langle \boldsymbol{E}_1 \cdot \boldsymbol{E}_2 \rangle$ averages to zero however the polarizations are afterwards rotated into alignment; in a polarized beam they do carry one, and the fringes return. □

67.4.3 Interpretation

[Reserved: the inference, drawn by Young and by Fresnel, that the vibration is transverse to the propagation direction, since a longitudinal disturbance has no plane to be at right angles to; the amplitude theory of reflection built on it [90], whose quantitative predictions are the Fresnel coefficients of Chapter 61; the cost to the mechanical aether, which must then be a solid, feeding Chapter 37; and the later mechanical demonstration that a circularly polarized beam carries angular momentum [767], which the transverse-wave account of Chapter 64 predicts quantitatively.]

67.5 Diffraction gratings and wavelength metrology

67.5.1 Fraunhofer's gratings

[Reserved: the wire and ruled gratings with which Fraunhofer turned diffraction into measurement [69, 74]: the grating equation $d(\sin\theta_m - \sin\theta_i) = m\lambda$, the first absolute wavelength determinations, and their attachment to the dark lines of the solar spectrum he had catalogued [60], so that a line acquires a number in metres rather than a letter. The apparatus, the angle measurement with a theodolite, and the resulting wavelengths with uncertainties are the content of this section; the spectroscopy they founded is Chapter 70.]

Phenomenon 67.3 (The grating turns a wavelength into an angle)**.** A screen ruled with a large number of equally spaced parallel apertures does not merely spread light: it concentrates it into a set of sharp orders at angles that depend on the wavelength, and the orders grow sharper — not merely brighter — as more rulings are illuminated. The angles obey

$$d\,(\sin\theta_m - \sin\theta_i) = m\lambda\,, \qquad m \in \mathbb{Z}\,, \tag{67.4}$$

with d the ruling interval and θ_i, θ_m the angles of incidence and diffraction from the normal. Since d is a mechanical length and θ_m is read on a divided circle, the wavelength that results is *absolute* [69]. Applied to the fixed dark lines of the solar spectrum [60], this replaces the letters by which those lines were labelled with numbers in metres.

Derivation. Take a plane wave incident at θ_i on a row of apertures spaced by d, and look in the direction θ_m. Two adjacent apertures are reached by wavefronts whose paths differ by $d\sin\theta_i$ on the incoming side, and the paths from them to a distant point in the direction θ_m differ by $d\sin\theta_m$; the net path difference between the contributions of neighbouring apertures is therefore $d(\sin\theta_m - \sin\theta_i)$, the same for every neighbouring pair. All N contributions arrive in phase exactly when that difference is a whole number of wavelengths, which is Equation (67.4).

The sharpening with N is the part that makes the instrument. Away from an order, the N contributions have phases in arithmetic progression with common difference $\delta = 2\pi(\sin\theta_m - \sin\theta_i)d/\lambda$, and their sum is a geometric series whose magnitude is $|\sin(N\delta/2)/\sin(\delta/2)|$. This has principal maxima of height N where Equation (67.4) holds, and its first zero at $N\delta/2 = \pi$: the angular width of an order therefore falls as $1/N$ while its height rises as N. That is why a grating with many thousands of rulings separates lines that a prism merely displaces, and it is the origin of the resolving power measured in Section 67.5.3. □

67.5.2 Rowland's ruling engine and concave grating

[Reserved: the screw-driven ruling engine and the concave grating that made the spectrograph a precision instrument [222]; line densities of order 600 *per millimetre over apertures of order* 0.1 m *(Rowland's specifications were quoted in lines per inch and are converted here, per the SI rule of the front matter); periodic-error "ghosts" as the signature of an imperfect screw, and the Rowland circle geometry that removes the need for collimating optics; the resulting solar wavelength tables, which were the working standard for decades.]*

67.5.3 Resolving power measured

[Reserved: the chromatic resolving power $\lambda/\Delta\lambda = mN$ of a grating and its verification on known doublets, the sodium pair being the standard test object; the Rayleigh criterion as the convention used to state it [209]; the imaging counterpart, Abbe's diffraction limit $d = \lambda/2n\sin\alpha$ for a microscope [184], with the measured demonstration that no improvement in lens quality passes it; the honest statement that this bound constrains linear *imaging only, which is what Section 67.9.3 exploits.]*

Phenomenon 67.4 (There is a diffraction limit, and it is measured)**.** Resolution is bounded by the wave nature of light and not by the quality of the optics. A grating separates two wavelengths only if they are further apart than a definite fraction of the wavelength, the limit rising in proportion to the order and to the number of illuminated rulings and to nothing else [209]; the sodium doublet is the standard test object on which this is checked. In imaging the same bound appears as a smallest resolvable separation, of the order of the wavelength divided by twice the numerical aperture, and it is reached in practice: beyond that point improving a lens changes nothing, which is a measured statement about real microscopes and not a theoretical caution [184].

Derivation pending #40

The diffraction limit: derive the chromatic resolving power of a grating from the angular width of an order — the wavelength divided by the order times the number of rulings — using the N-aperture sum and the convention that two lines are resolved when the maximum of one falls on the first zero of the other. Then derive the imaging bound from the requirement that at least two diffracted orders from a periodic object be collected by the objective, which gives the smallest resolvable period as the wavelength divided by twice the numerical aperture, and state the hypotheses — linearity in the illumination, far-field collection — on which the bound depends.

67.5.4 The same experiment with X-rays

[Reserved: a crystal as a three-dimensional grating with spacing of order $0.1\,\mathrm{nm}$*, and the Laue photographs that demonstrated both the wave nature of X-rays and the periodicity of crystals in a single exposure [385]; the geometry, the observed spot patterns and the wavelengths inferred; the crystallography that follows belongs to Chapter 125, and the same diffraction argument applied to matter waves is Section 67.9.2.]*

67.6 Interferometry

67.6.1 The Michelson interferometer

[Reserved: the beam-splitter geometry, the compensating plate, and the stone float on mercury that isolated it from vibration [236]; fringe counting as a length measurement in units of $\lambda/2$*; the coherence-length limitation set by the source, which is why the arms had to be near-equal before lasers; the sensitivity estimate in metres per fringe, which is the number that makes the null result of Chapter 37 meaningful.]*

67.6.2 The metre in wavelengths of light

[Reserved: Michelson and Benoit's determination of the international prototype metre in wavelengths of the red cadmium line, 1 *metre* $= 1\,553\,163.5$ *wavelengths, obtained by a chain of étalons each twice the previous [266]; the error budget of that chain; the consequence that a length standard could in principle be reproduced anywhere, realized in the successive SI definitions and superseded by the fixed value of* c *[2333]; the modern chain runs through the frequency measurements of Chapter 90.]*

Phenomenon 67.5 (A spectral line is a reproducible standard of length)**.** The wavelength of light emitted by a specified atomic transition is the same everywhere and at every time, so it can serve as a unit of length that no artefact is needed to embody. The international prototype metre was expressed in such units by counting interference fringes through a chain

of étalons, each twice the length of the one before, and the result — 1 metre = 1 553 163.5 wavelengths of the red cadmium line — was reproducible to a few parts in 10^7 [266]. A material standard could therefore in principle be dispensed with, which is what the successive redefinitions of the SI metre eventually did [2333].

Derivation. In a two-beam interferometer the two arms are traversed twice, so displacing one mirror through a distance x changes the optical path difference by $2x$. The intensity at the output returns to the same value each time that difference increases by one wavelength, so the number N of fringes counted while the mirror moves is

$$N = \frac{2x}{\lambda}, \qquad \text{that is} \qquad x = N\,\frac{\lambda}{2}. \tag{67.5}$$

A length is thereby converted into an integer plus a measured fraction, and the measurement is a *count*, which is why its precision is not limited by the resolution of any scale.

The chain of étalons is forced by the one quantity Equation (67.5) does not control. Fringes remain visible only while the path difference stays within the coherence length of the source, which even for a good spectral line falls far short of a metre — and a metre contains several million half-wavelengths, which cannot be counted in one pass. Each étalon is therefore compared with one of half its length by counting the fringes over the difference, and only the shortest is compared directly with the wavelength; the uncertainty of the whole chain is the accumulation of those successive comparisons, which is the error budget the experiment must present. □

67.6.3 Light in a moving medium

[Reserved: Fizeau's interferometer with counter-propagating beams in two tubes of flowing water [133]; the measured fringe shift, of order a fifth of a fringe for water at a few metres per second over path lengths of order a metre, and the drag coefficient $1 - 1/n^2$ extracted from it; the reading of the result as a wave-optics measurement here, with its relativistic reinterpretation as the velocity addition law reserved for Chapter 37.]

67.6.4 Rotation: the Sagnac effect

[Reserved: a closed interferometric loop on a turntable, in which counter-propagating beams acquire a phase difference proportional to the enclosed area and the rotation rate [410]; the measured fringe shift versus rotation rate; the modern descendants — ring-laser and fibre-optic gyroscopes with resolutions far below the Earth's rotation rate — and the statement that the effect is kinematic and common to every wave, which is why it is not evidence about the aether.]

Phenomenon 67.6 (Rotation shifts the fringes: the Sagnac effect)**.** An interferometer whose two beams travel in opposite senses around a closed loop shows a fringe displacement when the whole apparatus is rotated in its own plane. The shift reverses with the sense of rotation, grows in proportion to the rotation rate and to the area enclosed by the loop, and requires no external reference of any kind — the instrument detects its own rotation while sealed [410]. In terms of the enclosed area A and the rotation rate Ω about the normal to the loop,

$$\Delta\varphi = \frac{8\pi A\Omega}{\lambda c}. \tag{67.6}$$

Modern ring-laser and fibre-optic gyroscopes built on this observation resolve rotation rates far below that of the Earth.

Derivation. Take the loop circular of radius R for simplicity, with the beam splitter fixed on it, and work in the non-rotating frame, in which the light travels at c in both senses. While a beam goes round, the splitter itself moves: the co-rotating beam must cover the circumference plus the distance the splitter has advanced, and the counter-rotating beam the circumference minus it. Their transit times therefore satisfy $ct_{\pm} = 2\pi R \pm \Omega R\, t_{\pm}$, so

$$t_{\pm} = \frac{2\pi R}{c \mp \Omega R}, \qquad \Delta t = t_{+} - t_{-} = \frac{4\pi R^{2}\Omega}{c^{2} - \Omega^{2}R^{2}} \simeq \frac{4A\Omega}{c^{2}}, \tag{67.7}$$

with $A = \pi R^{2}$ and $\Omega R \ll c$ in any practicable instrument. Multiplying by the angular frequency $\omega = 2\pi c/\lambda$ converts the time difference into the phase difference Equation (67.6).

Two things are worth reading off the derivation. Nothing in it refers to a medium: the same argument, with c replaced by the appropriate propagation speed, applies to sound and to matter waves, which is why the effect is kinematic and is *not* evidence for or against an aether — a point on which the original interpretation was mistaken. And the result depends on the enclosed area rather than on the shape of the loop or the position of the axis, which is what makes a coiled optical fibre of many turns a practical gyroscope: N turns multiply A, and therefore the signal, by N. □

67.7 Stellar interferometry

67.7.1 Michelson's beam interferometer

[Reserved: the method of measuring an angular diameter from the baseline at which the fringe visibility first vanishes [469], and its execution on the 2.5 m *Mount Wilson telescope with an outrigger beam of about* 6.1 m *(the original specification was in feet and is converted here); the measured angular diameter of* α *Orionis,* 2.3×10^{-7} *radians — about* 47 *milliarcseconds — the first measurement of the size of a star [479]; the visibility-versus-baseline curve as the observable, and the van Cittert–Zernike coherence theory behind it [736, 826], developed in Chapter 61.]*

67.7.2 Intensity interferometry

[Reserved: Hanbury Brown and Twiss's measurement of the angular diameter of Sirius by correlating intensity fluctuations at two separated detectors rather than amplitudes [1083], which is insensitive to atmospheric phase and so allows long baselines; the measured 3.3×10^{-8} *radians, about* 6.9 *milliarcseconds; and the controversy the result provoked, since photon correlations are exactly the observable that Chapter 122 shows to distinguish classical from non-classical light — the experiment is the origin of quantum optics and is cross-referenced there.]*

67.8 Gabor: holography and the recording of phase (1948, 1962)

Every measurement in this chapter so far has been a measurement of intensity, because that is all a photographic plate, a photocell or a retina can report: a square-law detector integrates the square of the field, and the phase of the wave that reached it is nowhere in the record. An image is what survives that loss, and it is very much less than the wave was. Gabor's proposal was to stop discarding the phase [913]. Superpose the wave scattered by an object on a mutually coherent reference wave *before* the detector, and the phase difference between them becomes the position of a fringe — an intensity, which a plate can record.

The developed plate then holds the whole wave, amplitude and phase together, whence the name; and re-illuminating it releases that wave again.

The immediate aim was not three-dimensional pictures. Gabor wanted to correct the spherical aberration of the electron microscope, which no electron lens of the period could be made free of, by recording the aberrated electron wave and undoing the aberration afterwards, at leisure and optically, on the reconstruction [932]. That programme proved hard, and the light-optical demonstration built to test the principle carried a defect — a second, unwanted image superposed on the first — which kept the method a curiosity for fourteen years. Both obstacles fell when a coherent source became available [1186] and Leith and Upatnieks brought the reference wave in at an angle [1235, 1289], and, independently, when Denisyuk recorded the two waves running in opposite senses through a thick emulsion [1224]. The 1971 Nobel Prize in Physics was awarded to Gabor for the invention and development of the holographic method [1473].

67.8.1 Apparatus

Three arrangements, in the order in which they were built.

In-line. A high-pressure mercury arc filtered to a single line and imaged on a pinhole a few micrometres across, so that the illumination is spatially coherent across the plate. The object is a transparency carrying small opaque marks on a clear ground, standing in the diverging beam a short distance in front of a fine-grain photographic plate. In this geometry the part of the beam that passes undiffracted is the reference wave and the light scattered by the marks is the object wave; source, object and plate lie on one axis. That is not a choice but a constraint: the arc is coherent only over a short path difference, so the two interfering paths must be kept nearly equal, and the only way to keep them so is to let them travel together [932]. For reconstruction the developed plate is returned to a similar diverging beam with the object removed, and the emerging wave is examined with a microscope.

Off-axis. A laser beam divided by a beam splitter into two: one illuminates the object, the other is carried to the plate by mirrors and arrives there at an angle to the object wave. Two things follow from having a source coherent over metres rather than micrometres. The reference no longer has to share a path with the object wave, so it can be inclined; and the object may be a solid body at a working distance, illuminated diffusely, instead of a transparency held against the plate [1235, 1289]. The demands move to the emulsion and to the bench. The plate must resolve the fringes formed by two beams inclined at θ, of period $\lambda/\sin\theta$, which for visible light and an offset of a few tens of degrees is of order a micrometre, some thousand lines per millimetre — far beyond a pictorial emulsion. And nothing may move by an appreciable fraction of a wavelength while the exposure lasts, so the whole arrangement stands on an isolated table.

Counter-propagating. The reference beam enters the emulsion from the front and goes on to illuminate an object placed behind the plate; the light scattered back from the object returns through the emulsion against it. The emulsion is thick — many wavelengths — so that the standing wave the two beams form is recorded as a stack of fringe surfaces through its depth rather than as a pattern on its face [1224]. Nothing else is needed: one beam, no beam splitter, and no second path to be matched in length.

67.8.2 Procedure

The exposure is made with the object in place and both waves reaching the plate. The plate is developed so that its amplitude transmittance falls as nearly as possible linearly with the exposure it received: linearity of the photographic response is a hypothesis of the method and not an incidental convenience, and it is secured by working on the straight portion

of the characteristic curve of the emulsion and checked by repeating the record at several exposure levels.

Reconstruction is the same arrangement with the object taken away and the plate returned to where it stood, illuminated by the reference wave alone. What is observed is the wave emerging on the far side.

Four controls separate the effect from its alternatives.

- Expose with the reference blocked. The plate then records the intensity of the scattered light alone, develops to a featureless fog and reconstructs nothing: the record is made by the interference and not by the object light.
- Displace one mirror through a fraction of a wavelength during the exposure. The fringes wash out and the plate again reconstructs nothing, which locates the stored information in the fringe positions.
- Break the developed plate. Each fragment, illuminated alone, reconstructs the whole scene.
- Illuminate with a wave conjugate to the reference — the reference reversed. The emerging wave then converges instead of diverging and forms a real image in space, which can be received on a screen with no lens of any kind.

67.8.3 Observations and data

Gabor's reconstructions were legible: a transparency bearing the names of Huygens, Young and Fresnel was recorded and read back from the plate [932]. Superposed on every one of them, out of focus, lay a second image of the same object which no adjustment of the apparatus removed. The field was of the order of a millimetre across and the reconstruction had to be viewed with a microscope.

With the reference wave inclined, that second image is absent from the direction in which the reconstruction is viewed: the waves leaving the plate travel in distinguishable directions, and an eye placed in one of them receives only the reconstruction [1235]. With a laser and a diffusely illuminated solid scene the reconstruction shows the scene in depth [1289]. Two observations make that statement operational rather than impressionistic. Moving the eye across the plate changes the perspective continuously — near objects shift against far ones, and their hidden sides come into view — and a camera photographing the reconstruction must be refocused between the near and the far parts of the scene, each setting leaving the rest blurred. Neither is true of a photograph of the same scene.

The plate itself shows nothing. Under a microscope the record of a diffusely illuminated object is a fine irregular structure of fringes bearing no resemblance to the object: the scene is not stored anywhere in it as a picture, which is why a fragment can still deliver the whole of it.

The counter-propagating record reconstructs under ordinary white light, by reflection, and returns the object in the colour of the line that recorded it [1224]; tilting the plate away from the recording geometry shifts the reconstructed colour towards the blue. A processed emulsion may also contract, which shifts the colour the same way, so the two must be told apart before any colour is quoted.

Phenomenon 67.7 (An intensity record can give back a wavefront)**.** A wave scattered by an object and superposed at a photographic plate on a mutually coherent reference wave leaves a record which, re-illuminated by that reference wave alone, emits a wave agreeing with the original object wave in amplitude and in phase up to a single real constant. An observer looking through the plate sees the object where it stood, in depth and with parallax, although the object is absent and the plate carries no picture of it [913, 932, 1289]. Three

Table 67.1: The three recording geometries and what each one settles. All three record an interference pattern and reconstruct by diffraction from it, and all three are described by the same four terms of Equation (67.9); they differ only in where the reference wave comes from, and each difference buys one property that the previous arrangement lacked.

Arrangement	Reference wave	Decisive observation
In-line [913, 932]	the undiffracted part of the illuminating beam, on the same axis	a wavefront is recovered from an intensity record, with the conjugate wave superposed on it
Off-axis [1235, 1289]	a separate laser beam brought to the plate at an angle	the reconstructed waves separate in direction; a diffusely illuminated solid scene reconstructs with full parallax
Counter-propagating [1224]	the illuminating beam entering a thick emulsion from the opposite side	a volume record selects its own wavelength and reconstructs in white light

further waves leave the plate at the same time, one of them carrying the complex conjugate of the object wave.

Derivation. Let O and R be the complex amplitudes of the object and reference waves in the plane of the plate, both derived from one source and so mutually coherent. What the plate integrates is

$$I = |R + O|^2 = |R|^2 + |O|^2 + R^* O + R\,O^* \,, \tag{67.8}$$

of which the first two terms are the separate intensities and the last two are the interference. The phase of O has not been lost: it stands in Equation (67.8) as the phase of the cross terms relative to R, that is, as the position of the fringes, and a displacement of fringes is a difference of intensity from place to place, which is exactly what a square-law detector does record.

Develop the plate on the linear part of its characteristic, so that its amplitude transmittance is $t = t_0 - \beta I$ with t_0 and β real constants, and illuminate it with the reference wave alone. The transmitted field is

$$tR = \left(t_0 - \beta\,|R|^2\right) R - \beta\,|O|^2\,R - \beta\,|R|^2\,O - \beta R^2 O^* \,. \tag{67.9}$$

Four waves therefore leave the plate. The first is the reference wave attenuated; the second is that same wave modulated by the object's own intensity, a halo about it; the third and fourth are the interesting ones.

The third term is the assertion. For a uniform reference $|R|^2$ is a real constant over the plate, so this term is O itself multiplied by a number: every ray direction, every curvature and every relative phase of the original object wave is present in it unaltered. An eye or a lens placed in that wave receives exactly what it would have received from the object, and therefore sees the object.

The fourth term is the price. It carries O^*, the object wave with its phase reversed, which converges where O diverged and so forms a real image at the mirror position of the object; superposed on the reconstruction it is a second, unwanted image. It cannot be developed away, because Equation (67.8) is a real quantity and its two cross terms are complex conjugates of one another: no record of R^*O that a square-law detector is able to make omits $R\,O^*$. □

Phenomenon 67.8 (Inclining the reference separates the reconstruction)**.** In the in-line arrangement the four waves of Equation (67.9) travel in nearly the same direction, and the conjugate wave is seen superposed on the reconstruction as an out-of-focus ghost that no adjustment removes; it is present in every reconstruction reported from that geometry [932]. When the reference wave is brought in at an angle to the object wave the four leave the plate in different directions and may be looked at one at a time, so that the reconstruction is seen alone [1235]. The separation appears only above a threshold inclination, and the threshold grows with the angular extent of the object.

Derivation. Put the plate in the plane $z = 0$ and describe the object wave there by its angular spectrum in the plate coordinate x, supported in $|\nu| \le B$. The bound B is fixed by the angular extent of the object as seen from the plate, since a point of it lying at an angle φ from the axis contributes the spatial frequency $\sin\varphi/\lambda$. Let the reference be a plane wave inclined at θ in the same plane,

$$R = R_0\, \mathrm{e}^{2\pi \mathrm{i} \alpha x}\,, \qquad \alpha = \frac{\sin\theta}{\lambda}\,. \tag{67.10}$$

Multiplying a function by $\mathrm{e}^{2\pi\mathrm{i}\alpha x}$ translates its spectrum by α, so the four terms of Equation (67.9) occupy four bands of spatial frequency. The attenuated reference is a single line at $\nu = \alpha$. The halo carries $|O|^2$, an autocorrelation and hence of width $2B$, on the same carrier: it fills $|\nu - \alpha| \le 2B$. The reconstruction carries O with no carrier at all and fills $|\nu| \le B$. The conjugate carries O^* on the doubled carrier R^2 and fills $|\nu - 2\alpha| \le B$.

The reconstruction is free of the others exactly when its band meets neither of theirs, which requires $\alpha - 2B > B$ from the halo and $2\alpha - B > \alpha + 2B$ from the conjugate. Both reduce to the same condition,

$$\frac{\sin\theta}{\lambda} > 3B\,. \tag{67.11}$$

Setting $\alpha = 0$ recovers the in-line case: all four bands are then centred on zero, they overlap whatever the object, and the ghost is unavoidable — which is the observation.

Two consequences are worth reading off Equation (67.11). The threshold is proportional to B and so to the angular extent of the object, so a larger object demands a more oblique reference; the separation is not free. And the fringe period the emulsion must resolve is $1/\alpha < 1/(3B)$, so the record demands of the emulsion a spatial frequency at least three times the highest the object itself presents. A plate good enough to photograph the object is therefore not good enough to hologram it, by a definite factor — which is why the method needed emulsions that ordinary photography never asked for. □

Phenomenon 67.9 (Every part of the plate records the whole scene)**.** A hologram of a diffusely illuminated scene is not an image of it; the developed plate shows only fine irregular fringes. Any fragment of it, illuminated by the reference wave alone, reconstructs the entire scene — at a coarser resolution and over a narrower range of viewing directions, but complete [1289]. The reconstruction is three-dimensional in the operational sense: moving the eye changes the perspective, and a camera must be refocused between the near and the far parts of the scene.

Derivation. Diffuse illumination is what makes the first statement true, and it is the only ingredient needed for it. A diffusing surface scatters into a wide cone, so light from every point of the object reaches every point of the plate; the plate is therefore not a plane of conjugate image points, as it would be behind a lens, but a surface each point of which carries a contribution from the whole object. Restricting attention to a fragment of width a removes no object point from the record.

What the fragment does remove is stated by the same wave theory that gave the diffraction limit of Phenomenon 67.4. On reconstruction the fragment is the aperture

through which the reconstructed wave passes, so it is the exit pupil of the reconstruction, and the angular resolution with which the scene is seen degrades to about λ/a. The range of directions from which the reconstruction can be viewed is the range that fragment subtends at the object, about a/z for a scene at distance z, so the parallax available shrinks in proportion to the fragment as well. Both are aperture losses; neither is a loss of scene, which is the asymmetry the observation reports.

The depth requires no separate argument. By Phenomenon 67.7 the reconstructed wave *is* the object wave, so it carries the curvature that the wave from each object point had when it arrived: a point at distance z_1 reconstructs as a wave diverging from z_1, and a point at z_2 as one diverging from z_2. An eye or a camera must accommodate separately on the two, and displacing the pupil across the wave changes which rays enter it, which is parallax. Depth in the reconstruction is not a rendering of depth; it is the same optical fact as depth in the scene. □

Phenomenon 67.10 (A volume record reconstructs in white light)**.** When the reference wave and the object wave traverse a thick emulsion in opposite senses, the developed plate reconstructs the object under ordinary white light — a filament lamp, or the sun — and returns it in the colour of the monochromatic light that recorded it; the reconstruction is seen by reflection, on the same side as the source [1224]. Tilting the plate away from the recording geometry moves the reconstructed colour towards the shorter wavelengths.

Derivation. Two plane waves of vacuum wavelength λ running in opposite senses along z in a medium of refractive index n superpose to a standing wave whose intensity varies as $\cos^2(knz)$ with $k = 2\pi/\lambda$. Its antinodes are planes separated by

$$\Lambda = \frac{\pi}{kn} = \frac{\lambda}{2n}\,, \tag{67.12}$$

and in an emulsion many wavelengths deep the development records not one fringe pattern on a surface but a stack of such planes through the volume: a set of weakly reflecting layers of period Λ.

Illuminate that stack with white light at normal incidence. Each layer returns a small reflected amplitude, and the reflections from successive layers differ in path by $2n\Lambda$, so they add in phase only for the wavelength λ' with $2n\Lambda = \lambda'$ — which by Equation (67.12) is $\lambda' = \lambda$, the recording wavelength. Every other wavelength present in the illumination is returned by successive layers with phases that spread over a full cycle within the depth of the emulsion, and cancels. The record selects its own recording wavelength out of a broadband source, so neither a filter nor a monochromatic lamp is needed to read it.

The tilt follows from the same condition taken at an angle ϑ inside the emulsion, where the path difference between successive layers is $2n\Lambda\cos\vartheta$ and therefore $\lambda' = \lambda\cos\vartheta < \lambda$: the reconstruction moves towards the blue, as observed. Any processing that contracts the emulsion reduces Λ and shifts the colour the same way, which is why the two have to be separated before a wavelength is reported. □

Derivation pending #41

Diffraction efficiency of a hologram: what fraction of the reconstructing beam is delivered into the reconstruction, derived for a thin absorption record and a thin phase record by expanding the transmittance of a sinusoidal grating, and for a thick phase record by the coupled-wave treatment in which two waves exchange energy under the Bragg condition. What has to be established is that only a volume record can send almost all of the incident light into a single order — which is what makes a white-light reflection hologram bright enough to look at — while the thin records are bounded well below that.

67.8.4 Interpretation

The result is about detectors, not about light. Nothing in the four phenomena above goes beyond the superposition principle of Section 67.2 and the diffraction theory of Section 67.5: a hologram is an interference pattern, and a reconstruction is diffraction from it. What the experiment settles is that the phase discarded by every intensity detector is not lost information about the world but recoverable information about the wave, and that a medium with no memory of the object's shape — a uniform emulsion — can be made to hold it. This is why the experiment belongs beside Young and Fresnel rather than among the modern repetitions of Section 67.9: it uses no quantum optics, the reconstruction is a classical field, and the derivations above apply to any wave for which a coherent reference can be arranged.

Three points of method are worth separating out.

First, the second image was not a defect of workmanship. The derivation of Phenomenon 67.7 shows that any real record of the interference contains both cross terms, so the conjugate wave is a consequence of using a square-law detector and could not have been removed by better plates or better alignment. It was removed instead by changing the geometry so that the two terms leave the plate in different directions — a solution in the space of the experiment rather than in the space of its components.

Second, the fourteen-year delay was a coherence limit and not a conceptual one. Equation (67.11) requires the reference to be inclined, and an inclined reference travels a path different from the object wave's; a source coherent only over a short path difference cannot supply one. Here the coherence theory of [736, 826] appears as a hard constraint on an apparatus: the visibility of the fringes that constitute the record is the mutual coherence of the two beams, and a record of zero visibility is no record at all. A hologram may be read, on this view, as a measurement of the mutual coherence function carried out at every point of the plate at once.

Third, holography does not evade the diffraction limit of Phenomenon 67.4, and it is worth saying why not, since it is often supposed to. The resolution of a reconstruction is set by the aperture of the plate exactly as the resolution of an image is set by the aperture of a lens — that is the fragment result of Phenomenon 67.9. What holography escapes is a different loss, the loss of phase, which is not an aperture effect at all. Gabor's own aim turned on precisely that distinction: an aberration is a phase error, so a recorded wavefront can have its aberration subtracted after the event, which is impossible once the intensity alone has been kept [932]. Carrying that programme out with electron waves proved harder than the light-optical demonstration and came later; the demonstration itself, and the method built on it, are what the prize recognised [1473].

67.8.5 Primary references

The proposal [913] and its full development with the in-line experiments [932]; the off-axis reference wave and its reading as a carrier-frequency communication problem [1235], and the laser reconstruction of diffusely illuminated solid objects [1289]; the counter-propagating volume record [1224], published as a note in the Doklady of the USSR Academy of Sciences and available in English translation in *Soviet Physics Doklady*; the source that made the off-axis geometry practicable [1186]; and Gabor's own retrospective account of the twenty-three years [1473]. The scalar diffraction theory used throughout is that of [1967], and the coherence theory invoked above is [736, 826].

67.9 Modern repetitions and precision

67.9.1 One photon at a time

[Reserved: Taylor's exposure of a diffraction pattern through smoked-glass attenuators over three months, with at most one photon in the apparatus at a time and no change in the pattern [360]; the modern version with a heralded single-photon source, an anticorrelation measurement at a beam splitter demonstrating particle behaviour and an interferometer demonstrating wave behaviour with the same *source [1774]; the quantities are the second-order correlation at zero delay and the fringe visibility, with their classical bounds stated explicitly; the theory is Chapter 122.]*

Phenomenon 67.11 (The pattern survives one quantum at a time)**.** The interference pattern does not require many quanta to be present together. Attenuated until at most one photon is in the apparatus at any moment, and exposed for months, a diffraction pattern comes out unchanged [360]; with a heralded single-photon source the same beam that shows particle-like anticorrelation at a beam splitter shows full-visibility fringes in an interferometer [1774]. The same holds for matter. Electrons sent one at a time through a biprism arrive as individual localized dots, apparently at random, and the accumulated dots build the fringe pattern [1826]; and molecules of some seven hundred atomic mass units show fringes at the de Broglie wavelength as well [1966]. The pattern is therefore a property of each quantum separately and not of the beam, and this is the precise point at which the wave account of this chapter stops being the whole story.

Derivation pending #42

Single-quantum interference: the derivation is quantum and belongs to the later parts. What must be shown here is negative and can be made precise — that no classical field theory reproduces the combination of a single localized detection with a fringe pattern of full visibility, since the second-order correlation at zero delay is bounded below by unity for any classical field, while the measured value for a heralded single photon falls well below it. The wave account of this chapter must therefore be read as the high-occupancy limit of something else.

67.9.2 Matter-wave double slits

[Reserved: Jönsson's electron interference through micro-ruled multiple slits [1212]; Tonomura's single-electron buildup, in which the pattern emerges dot by dot from a biprism [1826]; interference of C_{60} molecules [1966], molecules of 720 *atomic mass units showing the same fringes; the de Broglie wavelength inferred in each case and its comparison with h/p, which is the measurement rather than the illustration; the theory is Chapter 71.]*

67.9.3 Beyond the diffraction limit

[Reserved: why Abbe's bound [184] is not violated by stimulated-emission-depletion microscopy [1883] and its relatives — the imaging is nonlinear in the illumination, so the linear-optics theorem does not apply; near-field microscopy as the other route; the measured resolutions achieved, in nanometres; and the general lesson about the exact hypotheses a "fundamental limit" rests on.]

67.9.4 Interferometers as detectors

[Reserved: the line from Section 67.6.1 to the kilometre-scale, power-recycled, Fabry–Pérot-armed Michelson interferometers of Advanced LIGO [2243], whose measured strain sensitivity

of order 10^{-21} corresponds to a differential arm-length change of about $1 \times 10^{-18}\,\mathrm{m}$ — a fringe fraction of order 10^{-12} — and the first detection made with it [2264]; the noise budget (seismic, thermal, shot) as the modern form of Michelson's error analysis; the physics measured is Chapter 46 and the experiment proper is Chapter 47.]

Part VIII

The Transition to Quantum Physics

Chapter 68

Black-Body Radiation and Planck's Hypothesis

Heat a cavity, pierce it with a small hole, and the light that emerges has a spectrum fixed by the temperature alone — not by the material of the walls, not by their colour, not by the shape of the cavity. Kirchhoff proved this in 1860 [153] and thereby handed nineteenth-century physics a single universal function of frequency and temperature to compute. Thermodynamics and Maxwell's electrodynamics — the two most secure theories then available (Chapters 54 and 60) — between them fixed the integral of that function [211, 227], its behaviour under a change of temperature [259], and, empirically, its short-wavelength tail [272]. What they could not do was fix the function itself. Applied without reservation they returned a spectral density growing without bound at high frequency, and the divergence was no technical subtlety: it is the equipartition theorem of Chapter 55 applied to a system with infinitely many modes [296, 322].

This chapter opens Part VIII because the resolution — Planck's assumption of December 1900 that the energy exchanged between the cavity walls and a mode of frequency ν comes in elements $h\nu$ [295, 300] — is the first quantitative evidence for quantization found anywhere in nature, and because it was forced by measurement rather than by theory: the Reichsanstalt spectra of Lummer and Pringsheim [284] and the far-infrared points of Rubens and Kurlbaum [297] falsified Wien's distribution before any successor to it existed. The chapter then follows the new constant h out of the cavity: into the specific heats of solids [333, 382], into the statistics of indistinguishable quanta [509], and into the present-day SI, where its numerical value is fixed by definition [2333]. The photon itself is not here. Planck quantized the exchange and not the field, and said so; the light quantum belongs to Chapter 69.

Derivation pending #43

Black-Body Radiation and Planck's Hypothesis: all derivations of this chapter are pending.

68.1 Cavity radiation and Kirchhoff's theorem

68.1.1 Emissive and absorptive power

[Reserved: Kirchhoff's law of thermal radiation [153] — at thermal equilibrium the ratio of the emissive power to the absorptive power of any body is a universal function $K(\nu, T)$ of frequency and temperature. The proof is a thermodynamic one: two bodies exchanging radiation inside a common enclosure must not be able to drive a heat flow between equal temperatures, on pain of violating the second law of Chapter 54. Definition of a black body

as the limiting absorber, and Kirchhoff's own realization of it as a small aperture in a large isothermal cavity. Consequence: the spectrum of the aperture is a property of temperature alone, so a laboratory can measure a function of nature rather than of a material.]

Phenomenon 68.1 (Universality of cavity radiation)**.** The radiation emerging from a small aperture in an isothermal enclosure has a spectral distribution fixed by the temperature T and the frequency ν alone. It does not depend on the material of the walls, on their surface finish or colour, or on the size and shape of the cavity. Equivalently, for every body in equilibrium with that radiation the ratio of the spectral emissive power e_ν to the spectral absorptivity a_ν is one and the same universal function of frequency and temperature,

$$\frac{e_\nu(T)}{a_\nu(T)} = K(\nu, T)\,, \tag{68.1}$$

numerically equal to the emissive power of a perfect absorber [153].

Derivation. Place two bodies inside an evacuated enclosure whose walls are held at a common temperature T, and let them exchange energy only by radiation. Interpose a filter transmitting only a narrow band $\mathrm{d}\nu$ about ν; a filter is a passive optical element and can do no work, so equilibrium must hold in that band separately. Let I_ν be the spectral flux each body receives from the enclosure. Body j emits $e_\nu^{(j)}\mathrm{d}\nu$ and absorbs $a_\nu^{(j)} I_\nu \mathrm{d}\nu$; if the two were unequal the body would heat or cool while its surroundings stayed at T, so that heat would pass spontaneously between objects at the same temperature, which the second law of Chapter 54 forbids. Hence $e_\nu^{(j)} = a_\nu^{(j)} I_\nu$ for each body, and

$$\frac{e_\nu^{(1)}}{a_\nu^{(1)}} = \frac{e_\nu^{(2)}}{a_\nu^{(2)}} = I_\nu\,.$$

The two bodies were arbitrary, so the ratio is a function of ν and T only, which is Equation (68.1). Repeating the argument with two *cavities* at the same temperature, joined through the same narrow-band filter, shows that I_ν itself cannot depend on the walls: were the two fluxes unequal at any frequency, energy would flow in that band between enclosures at equal temperature. Finally a body with $a_\nu = 1$ at every frequency — a black body — emits exactly $K(\nu, T)$, so the universal function is realized by the aperture of any isothermal cavity, whatever it is made of. □

68.1.2 The cavity as a thermodynamic system

[Reserved: the electromagnetic field in a cavity treated as a thermodynamic body, with energy density $u(T)$, spectral density $u_\nu(\nu, T)$ and the relation $I = cu/4$ between energy density and the flux through the aperture. Radiation pressure $p = u/3$, obtained from the Maxwell stress tensor of Chapter 60 and used thermodynamically by Boltzmann [227] following Bartoli's cycle; the adiabatic invariance of ν/T under slow expansion of a perfectly reflecting cavity. The Maxwell relations that carry the argument are those of [1749], and the same system reappears as the photon gas of Chapter 120.]

68.2 What classical thermodynamics could settle

68.2.1 The Stefan–Boltzmann law

[Reserved: Stefan's empirical finding that the total radiated flux scales as the fourth power of the absolute temperature [211], extracted from Tyndall's platinum-wire data; Boltzmann's derivation of the same law from thermodynamics plus the radiation pressure $p = u/3$ [227], the first result in physics obtained by putting Maxwell's field through a Carnot cycle. The law

$I = \sigma T^4$ for a black body, and $I = \varepsilon\sigma T^4$ for a grey one; the Stefan–Boltzmann constant, which in the present SI is not measured but computed, $\sigma = 2\pi^5 k_B^4/15h^3c^2 = 5.670\,374\,419 \times 10^{-8}$ in SI units [2333, 2431]. Worked instances: effective temperatures of stars (Chapter 52) and the planetary equilibrium temperature.]

Phenomenon 68.2 (Stefan–Boltzmann law)**.** The total power radiated per unit area by a black surface, integrated over all frequencies, is proportional to the fourth power of its absolute temperature,

$$M = \sigma T^4 \,. \tag{68.2}$$

Stefan extracted the fourth power from Tyndall's measurements on heated platinum wire [211]; a body of emissivity $\varepsilon < 1$ radiates $\varepsilon\sigma T^4$. In the present SI the constant σ is no longer measured but computed from the defining constants: it is $5.670\,374\,419 \times 10^{-8}$ in SI units of watt per square metre per kelvin to the fourth power [2333, 2431].

Derivation. Treat the radiation filling an isothermal cavity as a thermodynamic fluid. By Phenomenon 68.1 its energy density u is a function of temperature alone, so $U = u(T)V$; take from Section 68.1.2 the isotropic radiation pressure $p = u/3$. For any simple system the energy equation of state reads

$$\left(\frac{\partial U}{\partial V}\right)_T = T\left(\frac{\partial p}{\partial T}\right)_V - p \,, \tag{68.3}$$

a Maxwell relation of Chapter 54. Its left side is u and its right side is $Tu'(T)/3 - u/3$, so

$$u = \frac{T}{3}\frac{\mathrm{d}u}{\mathrm{d}T} - \frac{u}{3} \qquad \Longrightarrow \qquad \frac{1}{u}\frac{\mathrm{d}u}{\mathrm{d}T} = \frac{4}{T} \,,$$

whose solution is $u = aT^4$ with a a constant of integration. The flux escaping through a small aperture is $M = cu/4$, the factor $\frac{1}{4}$ coming from averaging $c\cos\vartheta$ over the outward hemisphere, so $M = (ac/4)\,T^4$: this is Equation (68.2) with $\sigma = ac/4$. Thermodynamics fixes the exponent and leaves a undetermined. Only Phenomenon 68.5 computes it, and that it then agrees with the measured σ is the first quantitative success of the quantum hypothesis. □

68.2.2 Wien's displacement law

[Reserved: Wien's thermodynamic argument from the adiabatic compression of a cavity [259], which forces the universal function into the scaling form $u_\nu = \nu^3 f(\nu/T)$ without determining f. Corollaries: the wavelength of the maximum satisfies $\lambda_{\max}T = b$ with $b = 2.897\,771\,955 \times 10^{-3}\,\mathrm{m\,K}$, and the frequency of the maximum a different constant — the two maxima are not at the same place, because $u_\nu \mathrm{d}\nu$ and $u_\lambda \mathrm{d}\lambda$ are different densities, a point worth a remark since it is a standing source of error. Consistency of the scaling form with Section 68.2.1 on integration.]

Phenomenon 68.3 (Wien's displacement law)**.** The spectrum of a black body keeps its shape as the temperature is changed and merely slides along the wavelength axis: the wavelength at which the spectral radiance is greatest is inversely proportional to the absolute temperature,

$$\lambda_{\max}T = b \,, \tag{68.4}$$

where the constant b is $2.897\,771\,955 \times 10^{-3}\,\mathrm{m\,K}$ in the present SI [259, 2431]. This is why a body glows first dull red and then white as it is heated, and why the temperature of a furnace or of a star can be read from the colour of its continuum alone. The maximum of the *same* spectrum plotted against frequency lies at a different place and obeys a different constant, because $u_\nu \mathrm{d}\nu$ and $u_\lambda \mathrm{d}\lambda$ are densities with respect to different variables; the two maxima do not correspond.

Derivation pending #44

Wien's adiabatic argument: a slow compression of a perfectly reflecting cavity leaves the entropy unchanged while shifting every mode frequency, which forces the universal function of Kirchhoff's theorem into the scaling form given by ν cubed times a function of ν/T alone; and the maximisation of the corresponding wavelength density, which turns that scaling form into a fixed product of peak wavelength and temperature.

68.2.3 Wien's distribution law

[Reserved: Wien's 1896 distribution $u_\nu \propto \nu^3 \mathrm{e}^{-a\nu/T}$ [272], obtained from a Maxwell–Boltzmann analogy between emitting molecules and radiation rather than from thermodynamics, and therefore a guess of a different order from Section 68.2.2. Its excellent agreement with the visible and near-infrared measurements available before 1899, and its status until then as the accepted law. That it is the high-frequency limit of Planck's law is shown in Section 68.5.4, which is why its early success was not evidence of its truth.]

68.3 The classical field theory and its failure

68.3.1 Counting the modes of a cavity

[Reserved: the electromagnetic normal modes of a rectangular cavity with conducting walls, from the boundary-value problem of Chapter 61; the density of modes $8\pi\nu^2/c^3$ per unit volume per unit frequency, with the factor of two for the two transverse polarizations counted in Chapter 26. The result is independent of the shape of the cavity, by Weyl's asymptotic law for the eigenvalue counting function; the mode count is the same object that will reappear for phonons in Chapter 125 and for the free-particle density of states in Chapter 118.]

68.3.2 The Rayleigh–Jeans law

[Reserved: equipartition (Chapter 55) assigns $k_B T$ to each mode, giving $u_\nu = 8\pi\nu^2 k_B T/c^3$. Rayleigh's short note of June 1900 states the $\nu^2 T$ form and, doubting it at high frequency, inserts an exponential cut-off by hand [296]; Jeans supplies the correct numerical factor and argues the classical case in full [322]. Honest history: Rayleigh's factor was eight times too large and Jeans held for some years that the cavity was simply not in equilibrium, so "Rayleigh–Jeans law" names a result neither author asserted in the form now taught. The law agrees with the far infrared measurements of Section 68.4.3 to within their uncertainty.]

68.3.3 The ultraviolet catastrophe

[Reserved: the integral of the Rayleigh–Jeans density over frequency diverges, so a cavity at any temperature would hold infinite energy and no thermal equilibrium between matter and the field would exist. The name is Ehrenfest's [369], coined a decade after the fact, and the section should say so: the divergence played no part in Planck's reasoning and is a retrospective diagnosis. What is genuinely diagnosed is that the failure is not a failure of electrodynamics or of statistical mechanics separately, but of equipartition applied to a system with unboundedly many degrees of freedom — exactly the feature that returns as the ultraviolet divergences of Chapter 100.]

68.4 The measurements that decided the matter

68.4.1 Realizing a black body

[Reserved: the electrically heated cavity radiator built at the Physikalisch-Technische Reichsanstalt by Lummer and Kurlbaum [282], the first laboratory object whose emission could be claimed to be Kirchhoff's universal function; its use as the primary radiometric standard and the ancestry of present-day radiation thermometry (Chapter 2). Apparatus detail belongs here: aperture-to-cavity ratio, isothermality of the walls, bolometric and thermopile detection, and the calibration chain.]

68.4.2 Lummer–Pringsheim and the failure of Wien's law

[Reserved: the spectra measured by Lummer and Pringsheim between 300 K *and* 1650 K *[284] out to about* 8 µm*; systematic deviations from Wien's law [272] appearing at long wavelength and high temperature and growing beyond any plausible experimental error. This is the moment the accepted law died, and it died to data, not to an argument. The observations to be tabulated when written: wavelength, temperature, measured spectral radiance and its uncertainty, with the residual against both Wien and Planck.]*

68.4.3 Rubens–Kurlbaum in the far infrared

*[Reserved: the residual-ray (*Reststrahlen*) technique that isolated wavelengths of* 24 µm *and* 51.2 µm *by repeated reflection from fluorite and rock salt, and the finding of Rubens and Kurlbaum that at those wavelengths the radiance is proportional to T rather than to Wien's exponential [297]. Rubens communicated the result to Planck privately on 7 October 1900; the interpolation of Section 68.5.1 was written the same evening and read to the Physikalische Gesellschaft on 19 October. The chapter should be explicit that this is the causal link, since it is the clearest case in physics of a measurement dictating a formula.]*

Phenomenon 68.4 (Long-wavelength proportionality to temperature)**.** Far beyond the maximum of the spectrum the radiance of a black body is proportional to the absolute temperature itself, not to Wien's exponential in $1/T$. Rubens and Kurlbaum isolated 24 µm and 51.2 µm by repeated reflection from fluorite and rock salt and found the radiance at those wavelengths to follow T across the whole accessible range of furnace temperatures — a departure from the accepted distribution [272] far larger than any plausible experimental error [297]. Deviations of the same sign, smaller because the wavelengths were shorter, had already been reported between 300 K and 1650 K out to about 8 µm [284].

Derivation pending #45

The low-frequency limit of the cavity spectral density: that each mode carries the equipartition energy given by Boltzmann's constant times the temperature, so that the spectral density becomes proportional to the square of the frequency times the temperature, and the radiance at fixed long wavelength is therefore linear in T.

68.5 Planck's law

68.5.1 The interpolation of October 1900

[Reserved: Planck's route through the entropy of an oscillator rather than through its energy; the second derivative $\partial^2 S/\partial U^2$, which is proportional to $1/U^2$ in Wien's limit and to $1/U$ in the Rayleigh–Jeans limit, and the simplest expression reducing to both,

$\partial^2 S/\partial U^2 = -\alpha/(U(\beta+U))$ *[294]. Integrating twice returns the law that bears his name. This step is interpolation and nothing more — Planck called it a lucky guess — and the chapter should present it as such, because the physics enters only at the next step.]*

68.5.2 The combinatorial derivation of December 1900

[Reserved: Planck's derivation of 14 December 1900 [295]: distribute P indivisible energy elements ε over N resonators, count the distributions, apply $S = k \log W$ — an equation Planck wrote here for the first time, naming k after Boltzmann — and demand consistency with Wien's displacement law Section 68.2.2, which forces $\varepsilon = h\nu$. The resulting spectral density $u_\nu \propto \nu^3/(\mathrm{e}^{h\nu/k_B T} - 1)$; fits to the Reichsanstalt data yielding the first values of h and k_B and, through k_B, the first good value of Avogadro's number and of the elementary charge.]

Phenomenon 68.5 (Planck's radiation law)**.** The spectral energy density of the radiation filling an isothermal cavity is

$$u_\nu(\nu, T) = \frac{8\pi h \nu^3}{c^3} \frac{1}{\mathrm{e}^{h\nu/k_\mathrm{B}T} - 1}\,, \tag{68.5}$$

a single expression that reproduces the measured spectrum at every frequency and every temperature at which it has been examined, from the far infrared of Phenomenon 68.4 to the visible [284, 297, 300]. It contains one constant of nature not present in classical physics. Planck read the two constants h and k_B off the Reichsanstalt data of 1900, obtaining $6.55 \times 10^{-34}\,\mathrm{J\,s}$ and $1.346 \times 10^{-23}\,\mathrm{J/K}$ respectively [300]; in the present SI both are fixed by definition, at $6.626\,070\,15 \times 10^{-34}\,\mathrm{J\,s}$ and $1.380\,649 \times 10^{-23}\,\mathrm{J/K}$ [2333], so his 1901 values were already right to about one percent.

Derivation. Take the electromagnetic mode density of Section 68.3.1, $8\pi\nu^2/c^3$ modes per unit volume per unit frequency, and Planck's hypothesis [295] that a resonator of frequency ν exchanges energy with the field only in whole elements ε, so that its accessible energies are $E_n = n\varepsilon$ with $n = 0, 1, 2, \ldots$. In equilibrium at temperature T the level n carries statistical weight $\mathrm{e}^{-n\varepsilon/k_\mathrm{B}T}$, and writing $x := \mathrm{e}^{-\varepsilon/k_\mathrm{B}T} < 1$ the mean energy of one resonator is

$$\langle E \rangle = \varepsilon \frac{\sum_{n\geq 0} n x^n}{\sum_{n\geq 0} x^n} = \varepsilon \frac{x/(1-x)^2}{1/(1-x)} = \frac{\varepsilon x}{1-x} = \frac{\varepsilon}{\mathrm{e}^{\varepsilon/k_\mathrm{B}T} - 1}\,,$$

both geometric series converging because $x < 1$. Multiplying by the number of modes gives the spectral density

$$u_\nu = \frac{8\pi\nu^2}{c^3} \frac{\varepsilon}{\mathrm{e}^{\varepsilon/k_\mathrm{B}T} - 1}\,.$$

Nothing so far fixes ε. That it must be proportional to the frequency is forced by Phenomenon 68.3: only $\varepsilon = h\nu$, with h a universal constant, makes the last expression equal to ν^3 times a function of ν/T alone, which is the scaling Wien's thermodynamic argument establishes. Substituting $\varepsilon = h\nu$ yields Equation (68.5).

The two limits are immediate. For $h\nu \gg k_\mathrm{B}T$ the exponential dominates and Equation (68.5) becomes $(8\pi h\nu^3/c^3)\,\mathrm{e}^{-h\nu/k_\mathrm{B}T}$, which is Wien's distribution [272] — so its early success in the visible was no evidence of its truth. For $h\nu \ll k_\mathrm{B}T$, expanding $\mathrm{e}^{h\nu/k_\mathrm{B}T} - 1 \simeq h\nu/k_\mathrm{B}T$ gives $8\pi\nu^2 k_\mathrm{B}T/c^3$, the equipartition result [296, 322] and the observed proportionality to T of Phenomenon 68.4. Both classical results are limits of one formula, and neither is valid outside its own limit. □

68.5.3 The 1901 Annalen paper

[Reserved: the considered exposition [300], with the two universal constants $h = 6.55 \times 10^{-34}\,\mathrm{J\,s}$ and $k = 1.346 \times 10^{-23}\,\mathrm{J/K}$ as Planck determined them, to be compared in a table with the present exact SI values $h = 6.626\,070\,15 \times 10^{-34}\,\mathrm{J\,s}$ and $k_B = 1.380\,649 \times 10^{-23}\,\mathrm{J/K}$ [2333]; the agreement to a percent from 1901 data is itself an observation worth making. Also Planck's system of natural units, proposed in the same period from h, c and G alone.]

68.5.4 Limits, moments and consistency

[Reserved: the two limits of Planck's law — $h\nu \gg k_B T$ recovering Wien [272], $h\nu \ll k_B T$ recovering Rayleigh–Jeans [296, 322] — and its integral, which reproduces the Stefan–Boltzmann law with σ now computed *rather than measured, a genuine prediction and the first quantitative success of the hypothesis. The maximum reproduces Wien's displacement constant. Every classical result of Section 68.2 is thereby subsumed, and the constant b and the constant σ cease to be independent empirical numbers.]*

68.6 What exactly was quantized

68.6.1 Energy elements

[Reserved: the statement $\varepsilon = h\nu$ as a restriction on the exchange *of energy between the field and a material resonator, which is what Planck asserted [295], against the stronger reading that the field itself is granular, which he resisted for a decade and which is Einstein's (Chapter 69). The distinction is not scholastic: it is the difference between a property of matter and a property of light, and only the second is established by the experiments of the next chapter. Modern restatement: the mean energy of a quantum harmonic oscillator, $h\nu/(\mathrm{e}^{h\nu/k_B T} - 1)$, derived properly in Chapters 78 and 118.]*

68.6.2 The reception, 1900–1911

[Reserved: honest history of a result nobody believed. Planck's law was accepted as a formula and rejected as a hypothesis; the turning point was the first Solvay Council of 1911 and, before it, the independent evidence from specific heats [333]. Lorentz and Jeans argued for years that the cavity was simply not in equilibrium [322]; Ehrenfest's analysis [369] isolated what in the hypothesis is indispensable. The section states plainly which claims were settled by evidence and which by exhaustion.]

68.7 Out of the cavity: specific heats

68.7.1 The Einstein solid

[Reserved: Einstein's transfer of the quantization hypothesis from cavity walls to the vibrations of a crystal [333]; $3N$ independent oscillators of a single frequency ν_E, molar heat capacity falling below the Dulong–Petit value $3R$ below $T \approx h\nu_E/k_B$ and vanishing as $T \to 0$. This was the first prediction of the quantum hypothesis outside radiation, and it was confirmed by Nernst's low-temperature calorimetry — the evidence that converted the physics community. Its known defect: the exponential fall is too fast against the measured one.]

Phenomenon 68.6 (The specific heat of a solid vanishes at low temperature)**.** The molar heat capacity of a crystalline solid is close to the Dulong–Petit value $3R$ at ordinary temperatures, falls steadily as the solid is cooled, and tends to zero as $T \to 0$. The fall sets in around a temperature characteristic of the material, so that a hard, light crystal such as

diamond is already far below $3R$ at room temperature while lead is not. Nernst's calorimetry established the behaviour across many substances down to liquid-hydrogen temperatures [378]; in insulating crystals the low-temperature approach to zero follows $C \propto T^3$ [382]. Classical equipartition (Chapter 55) gives $3R$ at every temperature and cannot produce any of this.

Derivation pending #46

The quantum theory of lattice heat capacities: the mean energy of a quantised oscillator applied to the vibrational degrees of freedom of a crystal, which gives the exponential freeze-out of a single Einstein frequency; and the replacement of that single frequency by the elastic-continuum mode spectrum with a cut-off preserving the total mode count, which produces the observed cubic low-temperature law.

68.7.2 Debye's improvement and the phonon

[Reserved: Debye's replacement of the single frequency by the elastic-continuum mode spectrum, with the cut-off that keeps the mode count at $3N$ [382]; the resulting T^3 law at low temperature, which fits the data, and the Debye temperature as the single material parameter. The mode counting is that of Section 68.3.1 with sound in place of light; the full lattice-dynamical treatment is Chapter 125, and the electronic contribution that Debye's theory omits is Chapter 126.]

68.8 The statistics behind the law

68.8.1 Bose's derivation

[Reserved: Bose's 1924 derivation of Planck's law from the counting of indistinguishable quanta in phase-space cells, with no appeal to classical electrodynamics at any point [509] — the first derivation in which the law follows from the statistics of the field rather than from the mechanics of the walls. Einstein translated the paper, had it published, and extended the counting to material particles, from which follow Bose–Einstein statistics and condensation (Chapters 120 and 121). Forward pointer: the indistinguishability at issue is the symmetrization postulate of Chapter 83.]

68.8.2 Detailed balance and the Einstein coefficients

[Reserved: Einstein's 1917 rederivation [446], in which the equilibrium of a two-level atom with the radiation field requires three processes — absorption, spontaneous emission and stimulated *emission — and the ratios of their rates are fixed by demanding Planck's law at equilibrium. The argument also yields the momentum $h\nu/c$ transferred in each elementary process, which is the photon momentum of Chapter 69. Stimulated emission is the principle of the laser (Chapter 122), and the coefficients are computed from first principles in Chapter 82.]*

68.9 Black bodies in nature

68.9.1 The cosmic microwave background

[Reserved: the most exact black-body spectrum known, measured by the FIRAS instrument on COBE with residuals below one part in 10^4 of the peak and a temperature $T = 2.7255\,\mathrm{K}$ [2144]; the absence of any spectral distortion as a constraint on energy release in the early universe, and the present-day cosmological parameters that rest on it [2348]. The measurement itself

is Chapter 51 and the cosmology Chapter 48; what belongs here is only that Kirchhoff's universal function is realized in nature to a precision no laboratory cavity approaches.]

Phenomenon 68.7 (The cosmic microwave background is a black body)**.** Observed from above the atmosphere, the whole sky carries a nearly isotropic radiation field whose spectrum is that of a black body whose temperature T is $(2.725\,48 \pm 0.000\,57)\,\mathrm{K}$, with no departure from Equation (68.5) detected anywhere across the peak [2144]. It is the most accurately black-body spectrum known, and no laboratory cavity approaches it: the universe as a whole realizes Kirchhoff's universal function better than any apparatus built to realize it.

Derivation pending #47

Why the expanding universe leaves a Planck spectrum Planckian: thermal equilibrium of matter and radiation before recombination fixes the form, and the redshift of every mode by a common factor rescales the spectral density into another Planck spectrum with the temperature inversely proportional to the scale factor, so that no spectral distortion is generated by the expansion itself. The limits on distortions then become limits on energy release after equilibrium was lost.

68.9.2 Stellar spectra and radiation thermometry

[Reserved: stellar continua as approximate Planck spectra, the effective temperature as the parameter of the fit, and the departures — absorption lines, limb darkening, non-grey opacity — that carry the astrophysics of Chapter 52; the Sun at about 5772 K. On the metrological side, radiation thermometry as a primary method of realizing the kelvin, and the role of h as a defining constant of the present SI [2333, 2431], so that Planck's law now enters the definition of the units in which it is verified (Chapter 2).]

Chapter 69

The Photon: Photoelectric and Compton Effects

Chapter 68 ended with a constant h that governs how a cavity wall trades energy with the field, and with Planck's own insistence that this says nothing about the field itself. This chapter is about the evidence that it does. Einstein's 1905 proposal [319] was that light of frequency ν behaves in its emission, propagation and absorption as though it consisted of independent quanta of energy $h\nu$ — a statement about light, not about oscillators, and one that the thermodynamics of the previous chapter does not entail. He offered the photoelectric effect [234, 238, 308] as a test, because its observed dependence on frequency and its *in*dependence of intensity are inexplicable if the energy of a wave arrives spread over a wavefront.

The chapter sets out the evidence in the order in which it accumulated and grades it honestly. Millikan's decade of measurement [437] verified Einstein's linear relation to better than a percent while its author continued to reject the hypothesis behind it; and he was not simply being obstinate, for the photoelectric equation alone can be obtained from a quantized atom absorbing from a classical field [1402]. What settled the matter was momentum: Compton's shift in the wavelength of scattered X-rays [499, 501], and then the coincidence experiments [519, 520] that destroyed the last classical alternative [508] by showing that energy and momentum balance in every *individual* scattering event and not merely on average. The quantum thereby acquires the two attributes of a particle, and the field theory that makes this consistent is Part XI.

Derivation pending #48

The Photon: Photoelectric and Compton Effects: all derivations of this chapter are pending.

69.1 The photoelectric effect

69.1.1 Discovery: Hertz and Hallwachs

[Reserved: Hertz's observation, made while producing the first electromagnetic waves [234], that ultraviolet light on the spark gap lowers the potential at which the spark passes — an unwanted effect in an experiment about something else, and reported as such. Hallwachs' isolation of the phenomenon the following year [238]: a freshly polished zinc plate, negatively charged, is discharged by ultraviolet light and a positively charged one is not, so the carriers are negative. Identification of the carriers with Thomson's corpuscles (Chapter 70) by their charge-to-mass ratio.]

69.1.2 Lenard's measurements and the classical impasse

[Reserved: Lenard's systematic study with a monochromated source and a retarding potential [308], which established the three facts that no wave theory accommodates: the maximum kinetic energy of the emitted electrons is independent of the intensity; it increases with the frequency; and below a threshold frequency no electrons are emitted at any intensity. To be added, the fourth: the emission begins without measurable delay, whereas a classical wavefront needs minutes to deliver an electron's worth of energy to an atomic cross-section — the estimate should be made explicitly here, because it is the sharpest form of the contradiction. Only the photocurrent scales with intensity.]

Phenomenon 69.1 (The photoelectric facts)**.** Ultraviolet light falling on a clean metal surface ejects electrons, and the emission has four properties that no theory treating light as a continuous wave accommodates [234, 238, 308]:

1. below a threshold frequency ν_0, characteristic of the emitting metal, no electrons appear at any intensity and after any exposure;

2. above the threshold the maximum kinetic energy of the emitted electrons is independent of the intensity of the light;

3. that maximum increases with the frequency of the light;

4. the emission begins without measurable delay, even at intensities so feeble that a wave spread uniformly over the surface would need minutes to deliver one electron's binding energy to a region of atomic size.

Only the *number* of electrons emitted per second — the photocurrent — is proportional to the intensity.

Derivation pending #49

The classical accumulation time: the energy an electromagnetic wave of a stated irradiance delivers per second to a target of atomic cross-section, compared with the work function of the metal, giving the delay a wave theory requires before the first electron can be emitted; and the demonstration that no redistribution of energy within the classical field shortens it, which is what makes item 4 above a contradiction rather than an estimate.

69.2 The light-quantum hypothesis

69.2.1 Einstein's heuristic viewpoint of 1905

[Reserved: the structure of Einstein's argument [319], which is thermodynamic and not electrodynamic: in the Wien regime of Chapter 68 the entropy of radiation in a volume V varies with V exactly as that of an ideal gas of $E/h\nu$ independent particles, so monochromatic radiation of low density behaves as if *composed of that many localized quanta. Three consequences were offered as tests — Stokes' rule for fluorescence, the photoelectric effect, and photoionization. The chapter should stress that Einstein derived the quantum from the empirical Wien law, not from Planck's counting, and regarded the two arguments as independent.]*

69.2.2 The photoelectric equation and the work function

[Reserved: $E_{\max} = h\nu - \phi$, with ϕ the work function of the emitting surface [319]; the stopping potential $V_0 = E_{\max}/e$ as the measured quantity, the threshold frequency $\nu_0 = \phi/h$, and the slope h/e of the V_0 against ν line, which is universal — the same for every metal, which is the content of the prediction. Tabulated work functions in eV *(caesium about* 2.1 eV*, potassium* 2.3 eV*, copper* 4.7 eV*, platinum* 5.6 eV*), their sensitivity to surface contamination and crystal face, and hence why the experiment is hard. The many-electron theory of ϕ belongs to Chapter 126.]*

69.2.3 Millikan's verification

[Reserved: the ten-year programme [437] that cut fresh alkali-metal surfaces in vacuum with a magnetically operated knife to defeat oxidation, measured the stopping potential across the visible and near ultraviolet, confirmed the strict linearity, and extracted $h = 6.57 \times 10^{-34}\,\mathrm{J\,s}$ to about 0.5 % — then the best value of Planck's constant by any method. Honest history: Millikan undertook the work to disprove what he called a reckless hypothesis and wrote, on confirming the equation, that the light-quantum theory nonetheless seemed untenable; the section records this because it is a clean case of a prediction surviving a hostile test.]

Phenomenon 69.2 (The photoelectric equation)**.** The stopping potential V_0 that just prevents the fastest photoelectrons from reaching the collector is a strictly linear function of the frequency of the incident light,

$$eV_0 = h\nu - \phi\,, \tag{69.1}$$

and the slope of that line is h/e for *every* emitting metal, the material entering only through the intercept ϕ. Millikan established the linearity across the visible and near ultraviolet on alkali surfaces cut fresh in vacuum, and read from the slope a value of Planck's constant of $6.57 \times 10^{-34}\,\mathrm{J\,s}$, to about 0.5 % [437] — then the best determination by any method, and in agreement with the value obtained from cavity radiation [300], which is a measurement of an entirely different kind.

Derivation. Assume with Einstein [319] that light of frequency ν is emitted and absorbed only in indivisible amounts $h\nu$, and that one such amount is taken up by a single electron. An electron at the top of the occupied band must pay the work function ϕ to cross the surface, and any electron starting deeper pays more, so the fastest electron leaves with

$$E_{\max} = h\nu - \phi\,. \tag{69.2}$$

Three of the four observations of Phenomenon 69.1 follow at once. $E_{\max}$ depends on the size of the quantum and not on how many quanta arrive per second, which is item 2; it rises linearly with ν, which is item 3; and it would be negative — so no electron is emitted — for every $\nu < \nu_0 := \phi/h$ whatever the intensity, which is item 1. Raising the intensity multiplies the number of quanta and therefore the photocurrent, and nothing else. Retarding the emitted electrons through a potential difference V stops all those with $E_{\max} < eV$, so the measured stopping potential satisfies $eV_0 = E_{\max}$, which is Equation (69.1). Its slope h/e contains no property of the metal at all, and that universality, rather than the linearity, is the content of the prediction. □

69.2.4 What the photoelectric effect does not prove

[Reserved: the necessary caveat. A quantized atom coupled to a classical *radiation field reproduces the threshold, the linearity in frequency and the promptness of the emission,*

as shown by Lamb and Scully [1402]; the discreteness lies in the atomic energy levels of Chapter 70, which are independently established. The photoelectric effect is therefore strong evidence for quantized absorbers and weak evidence for quantized light. Naming what an experiment does not decide is editorial rule 1 applied to a case where the textbook claim is routinely overstated; the experiments that do decide it are Sections 69.5 and 69.6.1.]

69.3 Momentum and angular momentum of light

69.3.1 Radiation pressure

[Reserved: the classical prediction $p = u/c$ from the Maxwell stress tensor (Chapter 60) and its measurement by Lebedev with a torsion balance in vacuum [299] and, independently and more accurately, by Nichols and Hull [311]. In quantum language each quantum carries $p = h\nu/c = h/\lambda$, which Einstein obtained from detailed balance in 1917 [446]; the two descriptions agree for the pressure and differ in the fluctuations. Applications: comet tails, radiation pressure in stellar interiors (Chapter 52), optical tweezers and laser cooling (Chapter 122).]

Phenomenon 69.3 (Radiation pressure)**.** A beam of light exerts a mechanical force on the surface it falls upon. At normal incidence the pressure is

$$P = \frac{I}{c} \ \text{on a perfect absorber}, \qquad P = \frac{2I}{c} \ \text{on a perfect reflector}, \tag{69.3}$$

with I the incident irradiance. Lebedev detected the effect on light and dark vanes hung from a torsion fibre in an evacuated vessel, the whole difficulty of the measurement being to separate it from the far larger radiometric and convective forces [299]; Nichols and Hull, independently, confirmed the predicted magnitude to a few percent [311].

Derivation. In the quantum description the beam is a flux of N quanta per unit area per unit time, each of energy $h\nu$ and — by the detailed-balance argument Einstein gave in 1917 [446] — of momentum $h\nu/c$ directed along the propagation direction. The irradiance is the energy flux, $I = Nh\nu$, so the momentum arriving per unit area per unit time is $Nh\nu/c = I/c$. Force per unit area is momentum delivered per unit area per unit time, so an absorber feels $P = I/c$; a reflector returns each quantum with its momentum reversed and therefore receives twice as much, $P = 2I/c$. This is Equation (69.3).

The same two numbers follow classically from the Maxwell stress tensor of Chapter 60, and it must be said plainly that Equation (69.3) is therefore *not* evidence for quanta. The two descriptions agree on the mean pressure and differ in its fluctuations; what the quantum reading adds is that the momentum comes in the same indivisible packets as the energy, in the fixed ratio $1/c$, and it is that ratio which Phenomenon 69.5 tests directly. □

69.3.2 Angular momentum of light

[Reserved: circularly polarized light carries angular momentum $\hbar$ per quantum, measured mechanically by Beth [767], who suspended a birefringent half-wave plate on a quartz fibre and detected the torque exerted as the handedness of the transmitted beam was reversed. This is the direct evidence that the photon has spin one, and with the absence of a longitudinal polarization state (Chapter 26) it fixes the field's representation content under the Poincaré group (Chapter 94).]

Phenomenon 69.4 (Light carries angular momentum)**.** A circularly polarized beam of angular frequency ω and power P transports angular momentum about its own axis at the rate P/ω — one unit $\hbar$ for every $\hbar\omega$ of energy it carries, and of the sign set by the

handedness. Beth suspended a birefringent half-wave plate from a quartz fibre inside an evacuated enclosure and measured the steady torque exerted on it when the plate reversed the handedness of the transmitted beam, finding the predicted magnitude and the predicted reversal of sign with the handedness of the incident light [767]. This is a mechanical measurement of a property of light that has no analogue in the mechanics of a scalar wave.

Derivation pending #50

The angular momentum of a circularly polarized wave: the flux of angular momentum computed from the Maxwell stress tensor for a beam of finite transverse extent, its ratio to the energy flux, and the torque this predicts on a half-wave plate that reverses the handedness; and the identification of that ratio with one unit of the reduced Planck constant per quantum, which is what makes the result a statement about the spin of the photon rather than about the field alone.

69.4 Compton scattering

69.4.1 The 1923 experiment

[Reserved: Compton's measurement of molybdenum Kα X-rays (17.5 keV, 71 pm) scattered from graphite and analysed with a Bragg spectrometer [499]; the scattered radiation contains, besides the unshifted line, a component displaced to longer wavelength by an amount depending on the scattering angle alone and not on the wavelength, the material, or the intensity. Classical Thomson scattering (Chapter 65, [1976]) predicts no shift whatever, since a driven electron re-radiates at the driving frequency. Apparatus, angles, and the observed shift with uncertainties to be tabulated.]

Phenomenon 69.5 (The Compton shift)**.** X-rays scattered from a light element contain, besides a component at the incident wavelength, a component displaced to longer wavelength. The displacement depends on the scattering angle θ and on nothing else — not on the incident wavelength, not on the scattering material, not on the intensity —

$$\Delta\lambda = \lambda_{\mathrm{C}}\,(1-\cos\theta)\,, \qquad \lambda_{\mathrm{C}} = \frac{h}{m_{\mathrm{e}}c}\,, \tag{69.4}$$

where the Compton wavelength λ_{C} is 2.426 31 pm [2431]. Compton measured the shift on the molybdenum Kα line scattered from graphite, analysing the scattered radiation with a Bragg crystal spectrometer [499]. Classical scattering from a charge driven by the incident wave predicts no shift at all, at any angle, because a driven oscillator re-radiates at the frequency that drives it [1976].

Derivation. Treat the process as an elastic collision between a quantum of energy $h\nu$ and momentum $h\nu/c$ — the pair used already in Phenomenon 69.3 — and a free electron of mass m at rest, and impose the relativistic conservation laws of Chapter 40. Let the scattered quantum have frequency ν' and travel at an angle θ to the incident direction. Momentum conservation gives the electron momentum $\boldsymbol{p}_{\mathrm{e}} = \boldsymbol{p} - \boldsymbol{p}'$, whose square is

$$p_{\mathrm{e}}^2 = \frac{h^2\nu^2}{c^2} + \frac{h^2\nu'^2}{c^2} - \frac{2h^2\nu\nu'}{c^2}\cos\theta\,,$$

and energy conservation gives $E_{\mathrm{e}} = h\nu - h\nu' + mc^2$. Substituting both into the mass-shell relation $E_{\mathrm{e}}^2 = p_{\mathrm{e}}^2c^2 + m^2c^4$,

$$(h\nu - h\nu')^2 + 2mc^2\,(h\nu - h\nu') + m^2c^4 = h^2\nu^2 + h^2\nu'^2 - 2h^2\nu\nu'\cos\theta + m^2c^4\,.$$

Expanding the square on the left and cancelling $h^2\nu^2 + h^2\nu'^2 + m^2c^4$ from both sides leaves

$$-2h^2\nu\nu' + 2mc^2 h\,(\nu - \nu') = -2h^2\nu\nu'\cos\theta\,,$$

that is $mc^2\,(\nu - \nu') = h\nu\nu'\,(1 - \cos\theta)$. Dividing through by $mc\,\nu\nu'$ and using $\lambda = c/\nu$,

$$\frac{c}{\nu'} - \frac{c}{\nu} = \frac{h}{mc}\,(1 - \cos\theta) \qquad \Longrightarrow \qquad \lambda' - \lambda = \frac{h}{mc}\,(1 - \cos\theta)\ ,$$

which is Equation (69.4). The result carries no adjustable parameter and no property of the target, exactly as observed; Debye published the same kinematics independently and within weeks [501]. The unshifted component comes from electrons bound too tightly to recoil freely, for which m must be replaced by the mass of the whole atom, some five orders of magnitude larger, and the shift becomes unobservable. □

69.4.2 Kinematics and the Compton wavelength

[Reserved: treat the process as an elastic collision between a quantum of energy $h\nu$ and momentum $h\nu/c$ and a free electron at rest, and impose the relativistic conservation laws of Chapter 40. The result is $\Delta\lambda = \lambda_C(1 - \cos\theta)$ with the Compton wavelength $\lambda_C = h/m_e c = 2.426\,31\,\mathrm{pm}$, a pure prediction with no adjustable parameter. Why the unshifted line survives (scattering from tightly bound electrons, for which the recoiling mass is the whole atom); the recoil electrons, seen in a cloud chamber at the predicted energies and angles; and the identification of λ_C as the length scale at which Chapter 93 takes over from single-particle mechanics.]

69.4.3 Debye's independent derivation

[Reserved: Debye published the same kinematics from the same hypothesis within weeks of Compton and without knowledge of the experiment [501], a fact usually omitted; the relation is occasionally called the Compton–Debye formula on the continent. The section records the double attribution because the treatise's citation rule is to name the source that contains the claim, and here there are two.]

69.4.4 The Klein–Nishina cross-section

[Reserved: the kinematics fix the shift but not the rate. The angular distribution follows from the Dirac equation (Chapter 93) as the Klein–Nishina formula [630], one of the earliest quantitative successes of relativistic quantum mechanics; its low-energy limit is the Thomson cross-section $\sigma_T = 6.652\,458\,7 \times 10^{-29}\,\mathrm{m}^2$ [2431] and its high-energy behaviour falls as $\log(E)/E$. Compton scattering as the dominant photon interaction in matter between about $100\,\mathrm{keV}$ and $10\,\mathrm{MeV}$, hence its role in detector physics and in the inverse-Compton emission of Chapter 115.]

69.5 Coincidence: the last classical alternative

69.5.1 The Bohr–Kramers–Slater proposal

[Reserved: the 1924 attempt by Bohr, Kramers and Slater [508] to keep a continuous classical field by surrendering strict conservation: energy and momentum would hold only statistically, the atom being coupled to a "virtual" radiation field that fixes transition probabilities. It is the sharpest counter-theory the light quantum ever faced, and it makes a definite, falsifiable prediction — that a scattered quantum and its recoil electron are not *correlated event by event.]*

69.5.2 Bothe–Geiger coincidences

[Reserved: Bothe and Geiger placed two needle counters, one sensitive to scattered X-rays and one to recoil electrons, on opposite sides of a hydrogen-filled scattering chamber, and recorded their discharges on a moving photographic tape [519]. Coincident events appeared at a rate far above chance, with a resolving time of about 1×10^{-4} s, so energy and momentum balance in each individual act and the statistical proposal of Section 69.5.1 is excluded. This is the founding use of the coincidence method, for which Bothe was awarded the Nobel Prize, and it is the technique behind every particle detector in Part XI.]

Phenomenon 69.6 (Energy and momentum balance in every individual scattering event)**.** In Compton scattering the scattered quantum and the recoil electron appear *together*, in each single act, at the angles and energies that Equation (69.4) and momentum conservation assign to that act — not merely on the average over many acts. Bothe and Geiger placed one needle counter sensitive to scattered X-rays and one to recoil electrons on opposite sides of a hydrogen-filled scattering chamber and recorded coincidences at a rate far above chance with a resolving time of about 1×10^{-4} s [519]. Compton and Simon photographed both the recoil electron track and the track of a second electron ejected by the scattered quantum in the same cloud chamber, and found the two directions correlated as the kinematics require, within the measurement errors [520]. This excludes the one serious counter-theory, in which conservation would hold only statistically [508], and it was withdrawn at once.

Derivation pending #51

The angular correlation implied by the collision kinematics: the recoil-electron direction as a function of the scattering angle of the quantum, obtained from the same conservation laws that give the wavelength shift; the rate of accidental coincidences to be subtracted from the observed rate, given the counting rates and the resolving time; and the contrasting prediction of statistical-only conservation, under which the two directions are independent and the coincidence rate is the accidental rate alone.

69.5.3 Compton–Simon cloud-chamber pairs

[Reserved: the complementary experiment [520], in which a cloud chamber records the recoil electron's track and the track of a second electron ejected by the scattered quantum, so that both directions are measured in the same event; the angular correlation follows the kinematics of Section 69.4.2 within the measurement errors. Together with Section 69.5.2 this closed the question in 1925, and Bohr withdrew the proposal at once — a model of how a theory should be abandoned.]

69.6 The photon as a countable object

69.6.1 Anticorrelation at a beam splitter

[Reserved: the modern decisive experiment. Grangier, Roger and Aspect prepared single quanta by heralding on the second photon of a calcium radiative cascade, sent them onto a beam splitter and measured the joint detection rate at the two outputs [1774]; the observed anticorrelation parameter is well below the value that any classical field, however weak, can produce, while the same source in a Mach–Zehnder interferometer still shows full-visibility interference. One indivisible quantum, taking both paths: the statement of Chapter 85 rests on data of this kind and not on the photoelectric effect.]

Phenomenon 69.7 (A single quantum is not divided by a beam splitter)**.** Light prepared one quantum at a time and sent onto a beam splitter is detected in one output arm or in the other, and not in both. Writing α for the ratio of the joint detection rate at the two outputs to the product of the two singles rates, every classical field — however weak, and whatever its statistics — obeys $\alpha \geq 1$, because the joint rate is an average of a squared intensity and the Cauchy–Schwarz inequality bounds it from below. Grangier, Roger and Aspect prepared single quanta by heralding on the partner photon of a calcium radiative cascade and measured $\alpha = 0.18 \pm 0.06$, far below the classical bound; the same source in a Mach–Zehnder interferometer nevertheless produced interference of full visibility [1774].

Derivation pending #52

The two bounds that make the measurement decisive: the classical inequality on the normalised joint detection rate, from the Cauchy–Schwarz inequality applied to the intensity correlations of any positive classical intensity distribution; and the quantum prediction that a one-quantum state incident on a beam splitter gives a vanishing joint rate, together with the correction for the imperfect heralding and the accidental coincidences of a real source.

69.6.2 Counting photons

[Reserved: photomultipliers, avalanche photodiodes and transition-edge sensors as devices whose output is a count; the photon-number distribution of thermal, coherent and squeezed light, distinguished by their second-order correlation and developed in Chapter 122; single-photon sources and detectors as the working hardware of the Bell tests of Chapter 89. Historical footnote: the word "photon" is Lewis' [551], coined for an entity quite unlike the one it now names.]

69.7 Creation and annihilation of quanta

69.7.1 Pair production and annihilation

[Reserved: above $2m_e c^2 = 1.022\,\mathrm{MeV}$ a photon in the Coulomb field of a nucleus can convert into an electron and a positron; the discovery of the positron in cosmic rays by Anderson [713] and the first photographs of pairs by Blackett and Occhialini [714], who identified the process as the materialization of a quantum. Annihilation into two back-to-back $511\,\mathrm{keV}$ *quanta, the basis of positron emission tomography and the cleanest demonstration that photon number is not conserved. The experiment chapter is Chapter 95, and the theory is the Dirac equation of Chapter 93.]*

69.7.2 The photon in quantum electrodynamics

[Reserved: what a photon actually is in the settled theory — an excitation of the quantized electromagnetic field, first written down by Dirac [572], a massless spin-one particle with two helicity states and no rest frame, created and destroyed by field operators rather than moved about. Present experimental limits on the photon mass and charge [2427], quoted here as the numbers that make the idealization exact for every purpose in this treatise. The full development is Chapters 99 and 100; the non-technical account in [1752] remains the best statement of what the theory says.]

Chapter 70

Atomic Models and Spectra

Atomic spectra were measured for a century before anybody could say what an atom was. Fraunhofer mapped and lettered the dark lines of the solar spectrum in 1817 [60]; Kirchhoff and Bunsen showed in 1860 that each element carries its own fixed set of lines and thereby turned spectroscopy into chemical analysis [152]. By the end of the century the wavelengths were known to six figures and were the most precise numbers in physics — and they were pure numerology. Balmer found that the visible hydrogen lines fit a two-integer formula [230], Rydberg extended it to whole series in many elements with one constant [248], and Ritz observed that every observed frequency is a difference of two terms drawn from a single list [348]. Nothing in classical physics explains why a bound system should radiate at a discrete set of frequencies at all, still less why those frequencies should be differences.

This chapter follows the two lines of evidence — spectroscopic and mechanical — to the point where they meet. The mechanical line runs from Thomson's electron [276] through the large-angle α scattering of Geiger and Marsden [354, 404] to Rutherford's nuclear atom [379], which is dynamically impossible in classical electrodynamics: an orbiting charge radiates and the atom collapses in about $1 \times 10^{-11}\,\mathrm{s}$. Bohr's 1913 trilogy [396–398] joins the two lines by postulating stationary states, and buys with that postulate the Rydberg constant computed from e, m_{e} and h. The old quantum theory that grew from it [441, 463] explained the X-ray regularity of Moseley [408, 418] and the fine structure, and failed on helium, on the anomalous Zeeman effect [391] and on intensities. It is a scaffold, not a theory — but everything it got right is a constraint that Part IX must reproduce, and the direct demonstration that its energy levels are real is Chapter 72.

Derivation pending #53

Atomic Models and Spectra: all derivations of this chapter are pending.

70.1 Spectra before any model

70.1.1 The Fraunhofer lines

[Reserved: Wollaston's 1802 sighting of dark gaps in the solar spectrum, and Fraunhofer's systematic mapping of some 570 of them with a theodolite telescope and a diffraction grating of his own making, lettered A through K in order of decreasing wavelength [60]. The measurement was metrological in intent — Fraunhofer needed fixed wavelengths to characterize optical glass — and the lines became the first absolute wavelength standards. The coincidence of the D line with the yellow of a sodium flame, noted but not explained.]

70.1.2 Spectral analysis

[Reserved: Kirchhoff and Bunsen's demonstration that every element emits a characteristic line spectrum, that the same element absorbs at the same wavelengths, and that a spectrum therefore identifies a substance [152]; the immediate discovery of caesium and rubidium by this means, and the identification of the Fraunhofer D lines with sodium in the solar atmosphere. Absorption versus emission as an instance of Kirchhoff's radiation theorem [153], so that Chapter 68 and this chapter share a founding result. Astrophysical consequence: stellar composition becomes measurable (Chapter 52).]

Phenomenon 70.1 (Characteristic line spectra)**.** A chemical element heated to incandescence in the vapour phase emits light only at a fixed, discrete set of wavelengths: the same set every time, independent of the temperature, of the chemical compound the element was liberated from and of everything else about its preparation, and different for every element. The same element, interposed cold in front of a continuous source, absorbs at exactly those wavelengths and at no others. A spectrum therefore identifies a substance without touching it, which is how caesium and rubidium were discovered [152]. The dark lines Fraunhofer had mapped and lettered in the solar spectrum are of this kind [60]: the D pair coincides with the emission of a sodium flame, so there is sodium in the solar atmosphere.

Derivation pending #54

Two things must be derived and only the second is classical. That a bound system radiates at a discrete set of frequencies at all follows only from the quantisation of its internal energies, and is taken up in the quantum-mechanics part; that emission and absorption occur at the same wavelengths is the detailed-balance content of Kirchhoff's radiation theorem, and follows from the equilibrium argument of the black-body chapter applied frequency by frequency.

70.1.3 Balmer's formula

[Reserved: Balmer's fit to the four visible hydrogen lines, $\lambda = h_B m^2/(m^2-4)$ with $m = 3,4,5,6$ and $h_B = 364.56\,\mathrm{nm}$, matching the measured wavelengths to five significant figures [230], together with its successful prediction of the fifth line before Balmer was told it had already been seen in stellar spectra. The point to make: this is an exact empirical regularity with integers in it, found sixty-five years before it was understood, and it is the reason hydrogen became the test case for every atomic model.]

70.1.4 The Rydberg formula and the combination principle

[Reserved: Rydberg's generalization from wavelengths to wavenumbers, $\tilde{\nu} = R(1/n_1^2 - 1/n_2^2)$ for hydrogen and the term formula with quantum defects for the alkalis, all series in all elements sharing one constant R [248]. Ritz's combination principle [348]: the wavenumber of every observed line is the difference of two terms, and sums and differences of observed lines are themselves observed lines. The principle is stronger than any particular series formula, it is exact, and it says — before any model — that a spectrum is a set of levels *and not a set of frequencies. The other hydrogen series: Lyman in the ultraviolet [327], Paschen in the infrared [347].]*

Phenomenon 70.2 (The Rydberg formula and the Ritz combination principle)**.** The wavenumbers of the hydrogen lines are given, to the full precision of the measurements, by two integers and one constant,

$$\tilde{\nu} = \frac{1}{\lambda} = R_{\mathrm{H}} \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right) , \qquad n_2 > n_1 \geq 1 \, . \tag{70.1}$$

Balmer fitted the four visible lines with $n_1 = 2$ to five significant figures and predicted the fifth before learning it had already been seen [230]; the series with $n_1 = 1$ lies in the ultraviolet [327] and that with $n_1 = 3$ in the infrared [347]. More generally, and in every element examined, the wavenumber of each observed line is the difference of two members of a single list of terms belonging to that element, and sums and differences of observed wavenumbers are themselves observed lines [248, 348]. The combination principle is exact, it is stronger than any particular series formula, and it says — before any model of the atom exists — that a spectrum is a set of *levels* and not a set of frequencies.

Derivation pending #55

The term structure: that there exists a set of stationary energies of the atom whose differences, divided by Planck's constant, are exactly the observed frequencies. The stationary-state postulate assumed in the Bohr section below turns the combination principle into a law rather than deriving it; the derivation proper is the eigenvalue problem for the atomic Hamiltonian, solved for hydrogen in the quantum-mechanics part.

70.2 The constituents of the atom

70.2.1 The electron

[Reserved: Thomson's cathode-ray experiments [276] — deflection by crossed electric and magnetic fields, the charge-to-mass ratio some 1800 *times that of hydrogen, and its independence of the cathode material and the residual gas, which is what licenses the claim that the corpuscle is a universal constituent of matter. The subsequent measurement of e itself by the oil-drop method, and the current values [2431]. This is the first subatomic particle, and its charge quantization is an observed fact that Part XI still does not explain.]*

70.2.2 Thomson's model and its failure

[Reserved: the "plum pudding" atom of electrons in stable rings inside a uniform sphere of positive charge [316], which was a serious dynamical model and not a straw man: it predicted the periodicity of chemical properties from ring closure, and it predicted small-angle multiple scattering of α particles with a calculable and very small tail. That prediction is quantitative and falsifiable, which is exactly why the experiment of the next section settled the question. What it could not do is produce sharp spectral lines.]

70.3 The nuclear atom

70.3.1 Large-angle alpha scattering

[Reserved: Geiger and Marsden's observation that a small fraction of α particles incident on a gold foil are deflected through more than 90°*, some returning towards the source [354] — in Rutherford's later phrase, as incredible as an artillery shell rebounding from a sheet of tissue paper — against the Thomson model's prediction of essentially none, since a diffuse positive charge exerts no large force anywhere. The quantitative campaign of 1913 [404]: counts against scattering angle, foil thickness, foil material and α velocity, verifying every dependence of the formula below. Apparatus and uncertainties belong here; the modern descendant of the technique is deep inelastic scattering (Chapter 116), where the same logic finds structure inside the proton.]*

Phenomenon 70.3 (Large-angle scattering of alpha particles)**.** A beam of α particles traversing a thin metal foil is overwhelmingly deflected through small angles, but a small, definite and measurable fraction is deflected through more than 90°, some of it returning towards the source [354]. The number scattered into unit solid angle about the angle θ follows

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \left(\frac{Z_1 Z_2 e^2}{16\pi\varepsilon_0 E}\right)^2 \frac{1}{\sin^4(\theta/2)}\,, \tag{70.2}$$

with E the kinetic energy of the incident particle and Z_1, Z_2 the charges of projectile and target in units of e. Geiger and Marsden verified every dependence separately — on the scattering angle, on the foil thickness, on the material of the foil and on the velocity of the α particles [404]. A positive charge spread uniformly through the volume of the atom, as in [316], exerts nowhere a force large enough to produce such deflections, so the positive charge and essentially all the mass of the atom occupy a region smaller than $1 \times 10^{-14}\,\mathrm{m}$ [379].

Derivation pending #56

The Rutherford cross-section: the relation between impact parameter and scattering angle for a repulsive inverse-square force, read off the Kepler hyperbola of the classical two-body problem, and the Jacobian that converts it into a differential cross-section; the summation over foil thickness that gives the counting rate actually measured; and the remark, which must be made honestly, that the same expression is also the exact Born approximation for a Coulomb potential, so that a mechanics now known to be wrong gave the right answer here for a reason no one could have known in 1911.

70.3.2 Rutherford's scattering formula

[Reserved: Rutherford's derivation of the differential cross section $\mathrm{d}\sigma/\mathrm{d}\Omega \propto Z^2/(E^2\sin^4(\theta/2))$ for a point charge [379], from the Kepler hyperbola of Chapter 27, and the inference that the positive charge and essentially all the mass occupy a region smaller than $1 \times 10^{-14}\,\mathrm{m}$. The remarkable accident that the classical result coincides with the Born approximation of Chapter 86, so that a wrong mechanics gave the right answer; and the deviation at high energy that measures the nuclear radius (Chapter 107).]

70.3.3 Why the classical nuclear atom cannot exist

[Reserved: an accelerated charge radiates, by the Larmor formula of Chapter 65. An electron in a classical orbit of radius $5.29 \times 10^{-11}\,\mathrm{m}$ therefore spirals into the nucleus in about $1.6 \times 10^{-11}\,\mathrm{s}$, radiating a continuous spectrum of rising frequency as it goes. Both consequences contradict observation outright: atoms are stable for the age of the universe and their spectra are discrete. The estimate should be done explicitly, because it is the cleanest statement in the whole treatise that classical physics is not merely inaccurate but excluded.]

Phenomenon 70.4 (Atoms are stable, identical and discrete)**.** Every atom of a given element in its ground state is indistinguishable from every other, has the same size to within the precision of any measurement, and persists indefinitely: matter does not run down, and no atom has ever been observed to radiate its electrons into its nucleus. An excited atom does not radiate a continuum of rising frequency as it decays either; it radiates the sharp lines of Phenomenon 70.1. Classical electrodynamics requires the opposite on all three counts, because an electron on any bound orbit about the nucleus of Phenomenon 70.3 is accelerated and therefore radiates by the Larmor formula of Chapter 65, losing energy continuously and shrinking its orbit without limit.

Derivation pending #57

Stability of the atomic ground state: the existence of a lowest energy level, which no classical bound system in an inverse-square potential possesses, and the bound relating the localisation of the electron to its kinetic energy which fixes the size and energy of that level. The derivation belongs to the quantum-mechanics part, where the hydrogen atom is solved; what can be established here is only the negative statement of the remark below.

Remark 70.5 (The classical collapse time). The contradiction is quantitative, not merely qualitative, and the estimate is short enough to be given in full. An electron on a circular orbit of radius r about a proton has centripetal acceleration $a = e^2/4\pi\varepsilon_0 m_{\mathrm{e}} r^2$ and total energy $E = -e^2/8\pi\varepsilon_0 r$, and radiates at the Larmor rate $P = e^2a^2/6\pi\varepsilon_0 c^3$. Setting $\mathrm{d}E/\mathrm{d}t = -P$ and writing $r_{\mathrm{e}} := e^2/4\pi\varepsilon_0 m_{\mathrm{e}} c^2$ for the classical electron radius, every constant collapses into

$$\frac{\mathrm{d}r}{\mathrm{d}t} = -\frac{4}{3}\frac{c\,r_{\mathrm{e}}^2}{r^2}\,, \qquad \text{whence} \qquad t = \frac{r_0^3}{4c\,r_{\mathrm{e}}^2} \tag{70.3}$$

for the time to fall from r_0 to the nucleus. Taking for r_0 the observed atomic size $5.291\,77 \times 10^{-11}\,\mathrm{m}$, and for r_{e} its measured value $2.817\,94 \times 10^{-15}\,\mathrm{m}$ [2431], this is about $1.6 \times 10^{-11}\,\mathrm{s}$. Classical physics does not merely misdescribe the atom; it forbids the atom to exist for longer than a hundredth of a nanosecond, while radiating a continuous spectrum that is never seen.

70.4 The Bohr model

70.4.1 The 1913 trilogy

[Reserved: Bohr's postulates [396] — that an atom possesses stationary states in which it does not radiate, and that radiation accompanies a transition between two of them with $h\nu = E_i - E_f$, which is Ritz's combination principle [348] promoted to a law. Selection of the stationary states by quantized angular momentum $L = n\hbar$, equivalently by the action condition of Chapter 23. Parts II and III of the trilogy [397, 398] extend the scheme to multi-electron atoms and to molecules and are much less successful; the section should say which parts survived.]

70.4.2 What the model gets right

[Reserved: energies $E_n = -13.606\,\mathrm{eV}/n^2$, radii $r_n = n^2 a_0$ with the Bohr radius $a_0 = 5.291\,77 \times 10^{-11}\,\mathrm{m}$, and the Rydberg constant emerging as $R_\infty = m_e e^4/8\varepsilon_0^2 h^3 c = 1.097\,373\,156\,8 \times 10^7/\mathrm{m}$ — computed from independently measured constants and agreeing with the spectroscopic value, which is the single result that forced the model on a sceptical community. Also the Z^2 scaling verified on ionized helium, and the correct identification of the Pickering series in stellar spectra as He^+ rather than hydrogen.]

Phenomenon 70.6 (The hydrogen term values)**.** The terms whose differences Equation (70.1) organizes are, for hydrogen, inversely proportional to the square of an integer, so that the energies of the stationary states are

$$E_n = -\frac{hcR_\infty}{n^2}\,, \qquad n = 1, 2, 3, \ldots\,, \tag{70.4}$$

The Rydberg constant is known spectroscopically as $1.097\,373\,156\,8 \times 10^7/\mathrm{m}$, so the ground state lies $13.605\,693\,\mathrm{eV}$ below the ionization limit [2431]. The same expression with R_∞

replaced by $Z^2 R_\infty$ reproduces the observed spectrum of a one-electron ion, which is what identified the Pickering series seen in stellar spectra as belonging to He^+ and not to hydrogen [396].

Derivation. Take Bohr's two postulates [396]: that the atom possesses stationary states in which it does not radiate, and that radiation accompanies a transition between two of them, carrying away $h\nu = E_i - E_f$ — which is Ritz's combination principle [348] promoted from a regularity to a law. Select the stationary states by requiring the orbital angular momentum to be an integer multiple of $\hbar$. For a circular orbit of radius r about a fixed charge Ze, Newton's second law with the Coulomb force and the selection rule read

$$\frac{m_\mathrm{e} v^2}{r} = \frac{Ze^2}{4\pi\varepsilon_0 r^2}\,, \qquad \text{and} \qquad m_\mathrm{e} v r = n\hbar\,.$$

Eliminating v between them gives the radii

$$r_n = \frac{4\pi\varepsilon_0 \hbar^2}{m_\mathrm{e} Z e^2}\, n^2 = \frac{n^2}{Z}\, a_0\,, \qquad a_0 := \frac{4\pi\varepsilon_0\hbar^2}{m_\mathrm{e} e^2}\,,$$

the Bohr radius a_0 being $5.291\,77 \times 10^{-11}\,\mathrm{m}$, which is the observed size of the hydrogen atom and was not put in. The first of the two displayed relations makes the kinetic energy half the magnitude of the potential energy, so the total energy is $E = -Ze^2/8\pi\varepsilon_0 r$, and substituting r_n,

$$E_n = -\frac{m_\mathrm{e} Z^2 e^4}{8\varepsilon_0^2 h^2}\,\frac{1}{n^2}\,.$$

The wavenumber emitted in a transition is then

$$\tilde{\nu} = \frac{E_{n_2} - E_{n_1}}{hc} = \frac{m_\mathrm{e} Z^2 e^4}{8\varepsilon_0^2 h^3 c}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right),$$

which is Equation (70.1) with $R_\infty = m_\mathrm{e} e^4/8\varepsilon_0^2 h^3 c$, together with the Z^2 scaling. The decisive point is not that the formula has the right shape — Balmer's had that in 1885 — but that R_∞ is here *computed* from m_e, e and h, each of them measured in experiments having nothing to do with spectroscopy, and the number agrees with the spectroscopic value. That agreement is what forced a sceptical community to take the postulates seriously. □

70.4.3 Reduced mass and the isotope shift

[Reserved: replacing the electron mass by the reduced mass of the two-body problem (Chapter 27) shifts every line by a part in 2000 *between hydrogen and infinite nuclear mass, and by a computable amount between hydrogen and its isotopes. Urey's discovery of deuterium was exactly this: the predicted shifted Balmer lines were sought in evaporated liquid hydrogen and found [708]. A quantitative prediction of a new isotope from a spectral formula is a strong argument, and it is worth setting out in full. Present-day accuracy of R_∞ [2431] and the spectroscopy that delivers it (Chapter 90).]*

Phenomenon 70.7 (The hydrogen–deuterium isotope shift)**.** Every hydrogen line is accompanied by a faint companion displaced slightly towards shorter wavelength, in the fixed proportion that follows from replacing the electron mass in the Rydberg constant by the reduced mass of the electron and the nucleus. Urey, Brickwedde and Murphy concentrated the heavy component by evaporating liquid hydrogen near its triple point, looked for the predicted companions of the Balmer lines, and found them at the predicted separations — about 0.179 nm for Hα, whose principal component lies at 656.28 nm — thereby discovering deuterium [708].

Derivation. The nucleus is not infinitely heavy, so the two-body reduction of Chapter 27 replaces m_{e} throughout the derivation of Equation (70.4) by the reduced mass $\mu = m_{\mathrm{e}} M / (m_{\mathrm{e}} + M)$ of the electron and a nucleus of mass M. Every term, and therefore every wavenumber, is multiplied by

$$\frac{R_M}{R_\infty} = \frac{\mu}{m_{\mathrm{e}}} = \frac{1}{1 + m_{\mathrm{e}}/M} \, .$$

For two isotopes of nuclear masses M and $M' > M$ the fractional difference in wavenumber is, to first order in the small ratio m_{e}/M,

$$\frac{\Delta \tilde{\nu}}{\tilde{\nu}} = \frac{m_{\mathrm{e}}}{M} - \frac{m_{\mathrm{e}}}{M'} \, .$$

With M the proton mass and M' the deuteron mass, very nearly twice it, this is about $m_{\mathrm{e}}/2M_{\mathrm{p}}$, that is 2.72×10^{-4}. Applied to Hα it predicts a separation of 0.179 nm, which a grating spectrograph of 1932 could resolve comfortably — the whole difficulty of the experiment was to enrich the sample enough for the companion to be visible at all. The displacement is towards *shorter* wavelength because the heavier nucleus makes μ larger and hence every term value larger. Note that the argument makes a quantitative prediction about a substance nobody had yet observed, and that this is a far stronger test of Equation (70.4) than a fit to lines already measured. □

70.4.4 Where the model fails

[Reserved: no line intensities, no transition rates, no selection rules except by fiat; no account of the helium atom, whose ground-state energy the old quantum theory got badly wrong despite a decade of effort; no molecules; no explanation of why the postulates hold. The orbits themselves are wrong — the ground state has zero orbital angular momentum, not $\hbar$ — and the reason the energies come out right anyway is explained only by Chapter 81. Bohr himself catalogued the failures [463]; the model is retained in this treatise as evidence about levels, not as a picture of an atom.]

70.5 The old quantum theory

70.5.1 Sommerfeld's elliptic orbits and the fine structure

[Reserved: Sommerfeld's extension [441] of the quantization condition to each separable degree of freedom of Chapter 23, giving elliptic orbits with a radial and an azimuthal quantum number, degenerate in the nonrelativistic problem; the relativistic correction to the electron mass lifts the degeneracy and yields a fine-structure splitting of order α^2, with $\alpha = e^2/4\pi\varepsilon_0\hbar c \approx 1/137$ introduced here for the first time. The formula agrees with the observed hydrogen fine structure, and it agrees for the wrong reason — the true splitting involves the electron spin of Chapter 79, absent from Sommerfeld's theory. The coincidence is worth a remark: it is the standard warning against counting agreement alone as confirmation.]

70.5.2 The correspondence principle

[Reserved: Bohr's requirement [463] that in the limit of large quantum numbers the frequencies and intensities of quantum transitions approach those of the classical Fourier components of the orbital motion. It is the only tool the old quantum theory had for intensities and selection rules, it fixed $\Delta l = \pm 1$, and it was the heuristic from which Heisenberg's matrix mechanics was constructed. Its modern residue is the classical limit of Chapter 82 and the Ehrenfest theorem of Chapter 77.]

70.5.3 Space quantization

[Reserved: the quantization of the orientation of the orbital plane in a magnetic field, introduced by Sommerfeld [441] to account for the Zeeman pattern of Section 70.7.1, and taken by contemporaries as an absurdity too far. It is the prediction tested directly by the deflection of a silver beam in an inhomogeneous field [481, 491, 492], whose two-component result confirmed the discreteness while contradicting the theory that predicted it — silver has no orbital angular momentum in its ground state, and the doubling is spin. The experiment and this honest history are Chapter 80 [2050].]

70.6 X-ray spectra and the atomic number

70.6.1 X-rays and their wavelengths

[Reserved: Röntgen's discovery of a penetrating radiation from a discharge tube [269]; the demonstration by Friedrich, Knipping and Laue that crystals diffract it, which established both that X-rays are short-wavelength electromagnetic radiation and that crystals are periodic lattices [385]; the Braggs' reduction of the diffraction condition to $n\lambda = 2d\sin\theta$ and the crystal spectrometer built on it [400], which made X-ray wavelengths measurable to a few parts in a thousand. Without this instrument neither Section 70.6.2 nor the Compton measurement of Chapter 69 would have been possible.]

70.6.2 Moseley's law

[Reserved: Moseley's measurement of the characteristic K and L emission of the elements from aluminium to gold [408, 418] and the finding that the square root of the frequency is linear in an integer that advances by one from element to element, $\sqrt{\nu} \propto (Z - \sigma)$ with $\sigma \approx 1$ for the Kα line. The integer is the nuclear charge, not the atomic weight: the periodic table is thereby ordered by Z, the inversions of the weight ordering (argon and potassium, cobalt and nickel) are resolved, and the number of missing elements is fixed at exactly the gaps found — three below gold, all later filled. Screening as the physical content of σ, explained by Chapter 88.]

Phenomenon 70.8 (Moseley's law)**.** The characteristic X-ray emission of the elements from aluminium to gold consists of a few sharp lines whose frequencies advance smoothly and regularly through the periodic table. The square root of the frequency of the Kα line is a linear function of an integer which increases by exactly one from each element to the next,

$$\sqrt{\nu} = A\,(Z - \sigma)\,, \qquad \sigma \approx 1\,, \tag{70.5}$$

and the same holds, with a different slope and screening constant, for the L series [408, 418]. The integer is not the atomic weight but the charge on the nucleus: ordering the elements by Z resolves the inversions of the weight ordering, argon before potassium and cobalt before nickel, and the gaps in the sequence fix the number of elements then undiscovered — three below gold, all later found.

Derivation. Apply Equation (70.4) to the innermost electrons of a heavy atom. The Kα line is the transition from $n = 2$ to $n = 1$, and the electron making it moves in the field of the nucleus screened by the one remaining K electron, so that the effective charge is $(Z - \sigma)\,e$ with σ of order unity. The Z^2 scaling established with Equation (70.4) then gives

$$\nu = cR_\infty\,(Z - \sigma)^2 \left(\frac{1}{1^2} - \frac{1}{2^2}\right) = \frac{3}{4}\,cR_\infty\,(Z - \sigma)^2\,,$$

so that $\sqrt{\nu}$ is linear in Z with slope $A = (3cR_\infty/4)^{1/2}$, which is Equation (70.5). The inner electrons are the ones for which a hydrogen-like treatment is defensible in a many-electron atom, since the nuclear attraction dominates the mutual repulsion there; σ is what measures the residual screening. The straight line holding across sixty elements, with a slope containing no adjustable parameter, is what establishes that Z and not the atomic weight is the quantity that orders the table. □

70.7 Atoms in external fields

70.7.1 The Zeeman effect

[Reserved: Zeeman's observation that spectral lines broaden and then split in a magnetic field, with the components polarized as Lorentz's electron theory required [277]; the normal triplet with splitting $\Delta\nu = eB/4\pi m_e$, which yielded a charge-to-mass ratio agreeing with Thomson's [276] — an independent determination of the electron from optics. The Bohr magneton $\mu_B = 9.274\,010\,065\,7 \times 10^{-24}$ J/T [2431] as the natural unit; astrophysical use in measuring stellar and sunspot magnetic fields.]

Phenomenon 70.9 (Spectral lines split in a magnetic field)**.** A spectral line emitted by a source placed in a magnetic field B is resolved into several components. Their separations are proportional to B, and their polarizations depend on whether the source is viewed along the field or across it [277]. A minority of lines — the "normal" cases — split into exactly three components, the outer two displaced by

$$\Delta\nu = \frac{eB}{4\pi m_e}\,, \tag{70.6}$$

from which Zeeman obtained a charge-to-mass ratio for the radiating particle agreeing with the value Thomson measured on cathode rays [276] — an identification of the electron made optically and independently. Most lines, however, split into more components than three and in patterns Equation (70.6) does not describe. Paschen and Back mapped these anomalous patterns exhaustively, together with their collapse to the normal triplet in strong fields [391], and no theory available before 1925 fitted any of them.

Derivation pending #58

Two derivations of different standing. The normal triplet follows classically, from the Larmor precession that a magnetic field imposes on a bound charge: the three normal modes of an isotropic oscillator in a field, their frequencies, and the circular and linear polarizations of the light each of them emits. The anomalous patterns do not follow at all without the electron spin and the coupling of spin to orbital angular momentum, and their treatment belongs to the angular-momentum chapter of the quantum-mechanics part; what must be recorded here is that the empirical splitting factor was fitted, not derived.

70.7.2 The anomalous Zeeman effect

[Reserved: most lines do not split into three. Paschen and Back's systematic study of the anomalous patterns and of their collapse to the normal triplet in strong fields [391] left a body of exact data that the old quantum theory could not fit at all, and that required the Landé g factor as a pure empirical parameter. This was the sharpest unsolved problem in atomic physics before 1925, and its solution is electron spin with $g \approx 2$ (Chapter 79), whose own precise value is the most stringently tested prediction in physics (Chapter 112).]

70.7.3 The Stark effect

[Reserved: Stark's splitting of the hydrogen Balmer lines in an electric field of order $1 \times 10^5\,\mathrm{V/cm}$*, produced in the dark space just behind a perforated cathode [422]; the splitting is* linear *in the field for hydrogen and quadratic for other atoms, a distinction with no classical explanation. Linearity requires the accidental degeneracy of the Coulomb problem, so the effect is direct evidence about the symmetry of the hydrogen Hamiltonian, and it is the standard worked example of degenerate perturbation theory in Chapter 82. Stark broadening as a plasma diagnostic (Chapter 66).]*

70.8 Levels made visible

[Reserved: a short closing section stating what all the foregoing has and has not established. The spectroscopic evidence is about differences *of energies, by Ritz's principle [348]; every model in this chapter fits differences, and no optical measurement can show that a level exists otherwise than as a term in a difference. The independent demonstration comes from electron impact, where energy is delivered to an atom mechanically and the threshold is measured directly: that is Chapter 72, and it is placed immediately after this chapter for that reason. The theory that replaces the models assembled here begins at Chapter 77, and the hydrogen atom is solved properly in Chapter 81.]*

Chapter 71

Matter Waves

Chapter 69 established that light, which every nineteenth-century experiment had shown to be a wave, delivers energy and momentum in indivisible amounts. De Broglie's proposal [500, 521] was the converse, and it was made on grounds of symmetry rather than of data: if a wave carries momentum $p = h/\lambda$, then a particle of momentum p should be accompanied by a wave of wavelength h/p. The hypothesis was not idle. It supplies at once the one thing Bohr's model of Chapter 70 had to assume — the quantization of angular momentum becomes the condition that an integer number of wavelengths fit the orbit — and it completes the analogy of Chapter 23, in which classical mechanics stands to an unknown wave theory as geometrical optics stands to wave optics. The unknown wavelength is h/p, and the chapter's business is the evidence that it is real.

That evidence is unusually good and it keeps improving, which is why this chapter is long on experiment. Electrons were diffracted by crystals within three years [571, 589], neutral atoms within five [651]; the interference pattern was later built up one electron at a time [1588, 1826], so that no account in terms of a collective effect survives; neutron interferometers made the phase itself a laboratory observable and measured shifts due to gravity [1542] and to a 360° rotation of a spin-half particle [1566]; and molecules of thousands of atoms interfere today [1966, 2337] with no sign of a boundary beyond which the relation fails. The chapter closes on wave packets and group velocity — the machinery that reconciles a wave with a localized particle — and hands the problem of finding the wave's equation of motion to Chapter 77.

Derivation pending #59

Matter Waves: all derivations of this chapter are pending.

71.1 The de Broglie hypothesis

71.1.1 The relation

[Reserved: de Broglie's short note of 1923 [500] and the thesis that expanded it [521]; the pair of relations $E = h\nu$ and $p = h/\lambda$ asserted for every *carrier of energy and momentum, massive or not. The relativistic argument that motivated them: a particle at rest has an internal frequency mc^2/h, and requiring the phase of that oscillation to stay in step with a wave when the particle is seen from a moving frame (Chapter 38) forces a phase velocity c^2/v and a group velocity v. Numerical scale: an electron accelerated through $54\,\mathrm{V}$ has $\lambda = 0.167\,\mathrm{nm}$, comparable to a lattice spacing, which is what makes the effect observable at all.]*

71.1.2 Phase waves and the Bohr condition

[Reserved: the standing-wave condition $n\lambda = 2\pi r$ on a circular orbit is $L = n\hbar$, so Bohr's postulate [396] ceases to be an arbitrary selection rule and becomes the requirement that the accompanying wave be single-valued. The same reading applies to Sommerfeld's action integrals [441], which become the statement that the phase accumulated round a closed path is a multiple of 2π. This is the first argument in physics that explains an old quantum condition rather than positing it, and it is why the hypothesis was taken seriously before any experiment.]

71.1.3 The optical–mechanical analogy completed

[Reserved: the dictionary of Chapter 23 had one incomplete entry — classical mechanics is the short-wavelength limit of some *wave theory, with the part of the wavelength played by an unidentified quantity. De Broglie identifies it as h/p, and the Hamilton–Jacobi equation becomes the eikonal equation of the matter wave, valid where the potential varies slowly on the scale of λ. The criterion for the classical limit follows, and it is quantitative: it fails for an electron in an atom and holds for a marble. Schrödinger constructed the wave equation whose eikonal limit this is, and said so in his first communication [557].]*

71.2 Electron diffraction

71.2.1 Elsasser's prediction

[Reserved: Elsasser's one-page note [523] pointing out that de Broglie wavelengths of slow electrons match crystal lattice spacings, so a crystal should diffract an electron beam exactly as it diffracts X-rays [400], and that the anomalous scattering already reported by Davisson and Kunsman might be this. The prediction was published two years before the confirming experiments and was unknown to one of the two groups that performed them; the section records it because the priority is routinely assigned to the experimenters alone.]

71.2.2 Davisson and Germer

[Reserved: the Bell Telephone Laboratories experiment [571]. A vacuum accident recrystallized a nickel target into a few large crystals; the scattered intensity then showed sharp maxima at particular angles and energies where before it had been smooth. At 54 eV *the beam scattered from the (111) face peaks near* 50°*, giving* $\lambda = 0.165$ nm *against the de Broglie value* 0.167 nm*. Apparatus, the Faraday-box collector, the angle and voltage scans, and the systematic correction for the inner potential of the crystal are all to be given in full, since this is the primary evidence of the chapter.]*

Phenomenon 71.1 (Electrons are diffracted by crystals)**.** A beam of electrons of well-defined kinetic energy scattered from a crystal produces sharp intensity maxima at particular angles, exactly as X-rays of the same wavelength do on the same crystal, and the wavelength read off the diffraction geometry is h/p. Davisson and Germer, scattering electrons of 54 eV from the (111) face of a nickel single crystal, found a pronounced maximum near 50° corresponding to a wavelength of 0.165 nm, against the de Broglie value 0.167 nm [571]. Thomson and Reid passed electrons of 10–60 keV through thin polycrystalline films of celluloid, gold and aluminium and photographed concentric rings whose radii scale as the inverse square root of the accelerating voltage, independently of the orientation of the target [589].

Derivation. Assume de Broglie's relation $\lambda = h/p$ [500, 521] and compute the wavelength of the beam. An electron accelerated from rest through a potential difference V small enough

that $eV \ll m_{\mathrm{e}}c^2$ acquires $p = (2m_{\mathrm{e}}eV)^{1/2}$, so

$$\lambda = \frac{h}{\sqrt{2m_{\mathrm{e}}eV}}\,, \tag{71.1}$$

which at 54 V is 0.167 nm — comparable with a lattice spacing, which is the whole reason the effect is observable at all. The rest is the geometry of a grating. A row of scatterers of spacing d on the crystal surface reinforces at angles satisfying $d\sin\theta = n\lambda$; the nickel row spacing is 0.215 nm, measured independently by X-ray crystallography [400], and this puts the first order at 50°, where the observed maximum lies. Reading $\lambda = d\sin\theta$ back off that geometry gives 0.165 nm. Nothing was adjusted: d comes from X-rays and λ from the accelerating voltage, and the two independently determined numbers agree.

Equation (71.1) also accounts for the transmission geometry. In a polycrystalline film the crystallites take every orientation, so the reinforcement condition is met on cones about the beam and the pattern is a set of rings whose radii are proportional to λ and therefore to $V^{-1/2}$, which is what is photographed [589]. Because a polycrystal has no preferred orientation, that observation also removes any explanation resting on a surface effect peculiar to one crystal face. □

71.2.3 Thomson's transmission experiment

[Reserved: G. P. Thomson and Reid passed electrons of 10–60 keV *through thin films of celluloid, gold and aluminium and photographed concentric Debye–Scherrer rings whose radii scale as the inverse square root of the accelerating voltage [589] — a polycrystalline pattern, independent of target orientation and therefore of any surface effect, complementing the single-crystal geometry of Section 71.2.2. The historical remark is irresistible and belongs in a remark rather than the body: J. J. Thomson showed the electron is a particle [276] and G. P. Thomson that it is a wave, and both were right.]*

71.3 Interference of atoms and molecules

71.3.1 Helium beams: Estermann and Stern

[Reserved: diffraction of neutral helium atoms and of hydrogen molecules from the cleavage face of lithium fluoride [651], with the wavelength varied by changing the source temperature and checked against h/p over the thermal distribution. This is the decisive extension: helium is electrically neutral and structureless, so no explanation in terms of charge, of the Coulomb field, or of anything peculiar to the electron survives. The molecular-beam technique is Stern's, the same apparatus lineage as Chapter 80.]

Phenomenon 71.2 (Neutral atoms and molecules are diffracted)**.** Beams of helium atoms and of hydrogen molecules reflected from the cleavage face of a lithium fluoride crystal show diffraction maxima at the angles the grating condition assigns to the wavelength h/p, and those maxima move as predicted when the wavelength is changed by changing the temperature, and hence the mean momentum, of the source [651]. Helium is electrically neutral, chemically inert and has no permanent electric or magnetic moment, so no account of the phenomenon in terms of charge, of the Coulomb field, or of anything peculiar to the electron survives this measurement: the wave behaviour belongs to matter as such.

Derivation pending #60

Diffraction of a thermal beam: the grating condition for a two-dimensional surface lattice, and the shape of the observed maxima when the incident beam is not monochromatic but carries the Maxwellian velocity distribution of the oven, each velocity class contributing its own de Broglie wavelength. What must be shown is that the position of each maximum tracks the most probable momentum in the predicted way as the source temperature is varied, which is the quantitative content of the experiment.

71.3.2 The electron double slit

[Reserved: Jönsson's fabrication of slits of about 0.5 µm *width in copper foil by electron lithography, and the resulting two-, three-, four- and five-slit interference patterns formed with* 50 keV *electrons and magnified electron- optically [1212] — Young's experiment (Chapter 67) performed with matter. Why it took thirty-four years after Section 71.2.2: the wavelength is four orders of magnitude below optical, so the slit separation and the detection geometry are the whole difficulty.]*

71.3.3 Building the pattern one particle at a time

[Reserved: the experiment that removes the last collective explanation. Merli, Missiroli and Pozzi [1588] and, with a field-emission source, a biprism and a position-sensitive detector, Tonomura and collaborators [1826] recorded interference with at most one electron in the apparatus at a time; individual detections arrive as points at apparently random places and the fringes emerge only in the accumulated histogram. The wave is therefore a wave of probability amplitude, in Born's sense [533], and not a wave of anything that is spread out — the statement that Chapter 77 formalizes and Chapter 85 declines to explain further.]

Phenomenon 71.3 (The interference pattern is built one particle at a time)**.** When the source is weakened until at most one electron is inside the apparatus at any moment, the detector still registers individual, point-like arrivals, at positions that show no pattern one by one. The interference fringes appear only in the accumulated histogram of many thousands of such arrivals, and they are the same fringes, in the same places and with the same spacing, as those formed at high intensity [1588, 1826]. The pattern is therefore not an effect of electrons acting on one another, and the multiple-slit fringes formed with 50 keV electrons [1212] are a property of each electron separately.

Derivation pending #61

The accumulation statistics: that the arrival positions are independent samples drawn from the squared modulus of the single-particle amplitude, so that the histogram of many arrivals converges on the intensity pattern of the wave while no individual arrival carries any fringe at all. What must be derived is the convergence itself and its rate, and what must be stated is that the probabilistic reading of the amplitude is an additional postulate and not a consequence of the de Broglie relation.

71.4 Neutron interferometry

71.4.1 The perfect-crystal interferometer

[Reserved: Rauch, Treimer and Bonse's interferometer cut from a single silicon crystal, in which three parallel lamellae split, redirect and recombine a thermal-neutron beam by Bragg

reflection [1532], with an arm separation of centimetres and a wavelength near 0.18 nm*. It converts the phase of a matter wave into a countable intensity and thereby makes phase an experimental quantity for a massive particle. Requirements: lattice perfection over the whole device, thermal and vibrational stability; this is the instrument behind the next three subsections.]*

71.4.2 Gravitationally induced phase

[Reserved: rotating the interferometer about the incident beam tilts the plane containing the two paths out of the horizontal, so they run at different heights in the Earth's field and accumulate a relative phase proportional to $m^2 g A \lambda / \hbar^2$*; Colella, Overhauser and Werner observed the predicted fringe count [1542]. Two things make it remarkable and both belong here: the inertial and gravitational masses appear as a product* m^2*, so this is a quantum experiment in which the equivalence principle of Chapter 42 is probed; and* $\hbar$ *and* g *appear in the same formula, which no other laboratory measurement of the period achieved.]*

Phenomenon 71.4 (Gravity shifts the phase of a matter wave)**.** Rotating a neutron interferometer about the direction of the incident beam tilts the plane containing its two paths out of the horizontal, so that the two paths run at different heights in the Earth's gravitational field. The counted intensity in each exit beam then oscillates with the tilt angle α, with a phase difference

$$\Delta\varphi = \frac{2\pi m^2 g A \lambda}{h^2} \sin\alpha \tag{71.2}$$

between the paths, where A is the area they enclose and λ the neutron wavelength. Colella, Overhauser and Werner observed the predicted number of fringes as the interferometer was rotated [1542]. Two features are worth naming, and both are unusual. The mass enters *squared*, once inertially through the wavelength and once gravitationally through the potential, so this is a quantum interference experiment that probes the equivalence principle of Chapter 42; and $\hbar$ and g appear together in a single measured quantity, which no other laboratory experiment of its period achieved.

Derivation pending #62

The gravitationally induced phase: the local wavenumber of a neutron in a linear potential, obtained from the free-particle dispersion relation with the gravitational potential energy subtracted from the total energy; the phase accumulated along each arm as the line integral of that wavenumber; and the reduction of the difference to the stated product of the squared mass, the acceleration of free fall, the enclosed area and the wavelength. The systematic corrections that must accompany it — bending of the crystal under its own weight as the device is rotated, and the Sagnac phase from the Earth's rotation — belong with it.

71.4.3 Rotation by a full turn changes the sign

[Reserved: a magnetic field on one arm precesses the neutron spin; the interference contrast returns to its original value only after a rotation of 720°*, the amplitude changing sign at* 360°*. Measured independently by Werner and collaborators [1566] and by Rauch's group [1558]. This is a direct observation of the double cover* $\mathrm{Spin}(3) \to \mathrm{SO}(3)$ *of Chapter 14, that is of the fact that a spin-half state transforms under a representation of the covering group and not of the rotation group itself (Chapter 79); the sign is unobservable for a whole system and observable in an interferometer, which is the distinction the experiment makes concrete.]*

Phenomenon 71.5 (A spin-half state changes sign under a full rotation)**.** A magnetic field applied on one arm of a neutron interferometer precesses the neutron spin through a controlled angle, and the interference contrast is recorded as a function of that angle. The contrast does not return to its initial value after a precession of 360°: the amplitude has changed sign there, and the original state is recovered only after 720°. The double-turn periodicity was measured independently by two groups, on different instruments [1558, 1566]. The sign is unobservable for an isolated system and becomes observable here only because the interferometer compares it against an amplitude that was not rotated.

Derivation pending #63

The rotation operator for a spin-half state: that the generator of rotations has half-integer eigenvalues, so that a rotation by a full turn acts as minus the identity rather than the identity, and that a spin-half state therefore carries a representation of the covering group and not of the rotation group. The interferometric argument that turns this global sign into a measurable relative phase must accompany it, together with the relation between the applied magnetic field, the transit time and the precession angle.

71.4.4 Bound states of a neutron in the Earth's field

[Reserved: ultracold neutrons above a horizontal mirror form quantized vertical bound states in the linear gravitational potential; the transmission through a slit of variable height rises in steps at the predicted heights, with the lowest state around 13 µm *and an energy of a few* peV *[2034]. Airy functions as the eigenfunctions, worked in Chapter 78; the experiment is the cleanest existing observation of quantized motion in a gravitational field and is cited again in Chapter 135 for what it does* not *test.]*

Phenomenon 71.6 (Quantized bound states of a neutron above a mirror)**.** Ultracold neutrons travelling horizontally above a flat, polished mirror are bound vertically between the mirror below and the linear gravitational potential above, and the states of that vertical motion are discrete. The transmitted flux, measured as an absorber is lowered towards the mirror, does not fall smoothly with the gap but in steps: no neutron passes at all below a gap of about 13 µm, which is the extent of the lowest state, and further steps follow at the heights of the states above it [2034]. The energy of the lowest state is about 1.4 peV. This is the cleanest existing observation of quantized motion in a gravitational field.

Derivation pending #64

The eigenvalue problem for a particle in a linear potential above a hard wall: the equation reduces to Airy's equation, the admissible solutions are the decaying Airy function shifted so as to vanish at the mirror, and the eigenvalues are therefore fixed by its zeros. The classical turning-point heights follow from the eigenvalues, and it is those heights that the absorber measures; the step positions predicted this way are to be compared with the observed ones.

71.5 How large can an interfering object be?

71.5.1 Fullerenes

[Reserved: far-field diffraction of C_{60} *from a nanofabricated grating, with a de Broglie wavelength near* 2.5 pm*, some four hundred times smaller than the molecule itself [1966]; the*

molecules are hot, populating millions of internal vibrational states and radiating thermal photons, and the fringes survive — so internal complexity does not by itself destroy coherence. What does destroy it is the emission or scattering of a photon carrying which-path information, which the same apparatus demonstrates by admitting a background gas, and which is the experimental basis of Chapter 87.]

Phenomenon 71.7 (Large molecules interfere)**.** Interference survives in objects far larger than their own de Broglie wavelength and far more complicated than a point. C_{60} molecules diffracted from a nanofabricated grating produce fringes at a wavelength near 2.5 pm, some four hundred times smaller than the molecule itself, even though the molecules are hot enough to populate millions of internal vibrational states and to radiate thermal photons while in flight [1966]. Near-field interferometry has since shown fringes for tailored molecules built from more than two thousand atoms, with masses above 25 000 atomic mass units, at wavelengths of order 50 fm [2337]. No deviation from $\lambda = h/p$ has been found at any mass yet reached, and no boundary beyond which the relation fails has appeared.

Derivation pending #65

Two things, of which only the first is optics. The near-field Talbot–Lau fringe condition and its scaling with mass and wavelength, which is what fixes the grating periods and the source coherence a given molecular mass demands. And the decoherence criterion: how much which-path information the emission or scattering of a single photon of given wavelength carries, and hence the residual gas pressure and internal temperature at which the fringes must disappear — the quantitative statement of what these experiments do and do not bound.

71.5.2 The present limits

[Reserved: Talbot–Lau near-field interferometry has since shown interference for tailored molecules of more than two thousand atoms and masses above 25 000 atomic mass units [2337], at wavelengths of order 50 fm. No deviation from the de Broglie relation has been seen at any mass. The section should state plainly what this does and does not bound: it constrains proposed objective-collapse modifications of quantum mechanics, which is a genuine experimental frontier, and it is reported honestly in Chapter 133 rather than as a settled matter.]

71.6 Wave packets

71.6.1 Superposition and localization

[Reserved: a single plane wave $e^{i(kx-\omega t)}$ describes a particle of definite momentum and no position at all, so a localized particle must be a superposition — a Fourier integral over k with an envelope, in the sense of Chapter 17. The Gaussian packet as the worked case: its width in x, its width in k, and the product of the two. Nothing in this subsection is specific to quantum mechanics; it is the classical mathematics of Chapter 28 applied to the de Broglie wave, which is exactly why it settles the apparent paradox.]

71.6.2 Phase velocity, group velocity and spreading

[Reserved: the free-particle dispersion relation $\omega = \hbar k^2/2m$ gives a phase velocity $\hbar k/2m$ — half the particle's speed, and for the relativistic relation greater than c, which alarms nobody once it is seen that no signal travels at it — and a group velocity $\mathrm{d}\omega/\mathrm{d}k = \hbar k/m = v$, the

particle's own velocity. The dispersion is quadratic and therefore non-vanishing, so every free matter wave packet spreads, at a rate computed here and observed in the expansion of released cold-atom clouds (Chapter 121). Compare the non-dispersive string of Chapter 28.]

71.6.3 Bandwidth and the uncertainty relation

[Reserved: for any wave, the product of the spatial width of a packet and the width of its spectrum is bounded below by a number of order unity — a theorem about Fourier transforms [2313] that predates quantum mechanics. Insert $p = \hbar k$ and it reads $\Delta x\, \Delta p \gtrsim \hbar/2$: the uncertainty relation is what the de Broglie relation makes of a bandwidth theorem. Heisenberg's own argument was operational, through the disturbance caused by a measurement [576], and the two readings are not the same statement; which is which, and which one the modern inequality proves, is settled in Chapter 77.]

71.7 From a wavelength to a wave equation

71.7.1 What the hypothesis does not supply

[Reserved: $\lambda = h/p$ fixes the wavelength of a free particle and says nothing about a particle in a potential, about the equation the wave obeys, about what the wave is a wave of, or about spin. Each gap is named here with the chapter that closes it: the equation of motion [557] in Chapter 77, the probabilistic reading [533] there too, the spin degree of freedom in Chapter 79, and the relativistic treatment — for which the naive substitution into $E^2 = p^2c^2 + m^2c^4$ is the wrong first guess — in Chapters 92 and 93.]

71.7.2 Matter waves as instruments

[Reserved: the hypothesis is now hardware. Electron and neutron diffraction are the standard probes of crystal structure and magnetic order (Chapter 129); atom interferometers built from optical gratings [1847] serve as gravimeters, gradiometers and gyroscopes; and the recoil frequency measured in an atom interferometer gives h/m and hence the fine-structure constant to parts in 10^{10} [2316, 2363], an input to the tests of Chapter 112. A hypothesis that began as a symmetry argument now calibrates the constants of physics, which is the strongest form of confirmation available.]

Chapter 72

Experiment: Franck–Hertz

Everything in Chapter 70 that speaks of atomic energy levels speaks of them indirectly. Spectroscopy measures frequencies, Ritz's principle [348] shows the frequencies are differences of terms, and Bohr's postulates [396] interpret the terms as energies — but no optical measurement can reach a level except as one end of a difference, and a sceptic in 1913 could hold that the terms were a bookkeeping device. Franck and Hertz's experiment removes the intermediary. It delivers energy to a mercury atom mechanically, by bombarding it with electrons of controlled kinetic energy, and asks at what energy the atom begins to accept any. The answer is: at nothing below 4.9 eV, and then in units of 4.9 eV. Discreteness is measured with a voltmeter and an ammeter, in a tube containing no spectroscope at all [417].

The chapter is placed at the end of Part VIII because it is the point at which the quantized atom stops being a model that fits spectra and becomes a mechanical fact. It is also a lesson in reading one's own data: Franck and Hertz took the 4.9 V threshold to be the *ionization* potential of mercury and said so in print, a reading Bohr corrected within a year [425] and which the experimenters themselves abandoned only after further work [443, 467]. That history is set out here in full rather than tidied away, because the treatise's standard is that a reader must be able to reconstruct what was actually claimed and when. When written out, the chapter will use the structured `experiment` environment of the house style, with apparatus, procedure, observations in SI with uncertainties, interpretation and primary references as separate fields, and the digitized current–voltage curve deposited in the evidence store as `docs/library/data/Franck_1914a.csv`.

Derivation pending #66

Experiment: Franck–Hertz: all derivations and data analysis of this chapter are pending.

72.1 Historical context and the prediction under test

72.1.1 What the Bohr postulates predict for a collision

[Reserved: state the prediction sharply before any apparatus is described. If an atom possesses stationary states [396], an electron of kinetic energy below the first excitation energy $E_1 - E_0$ can only scatter elastically, and because the electron is some 4×10^5 times lighter than a mercury atom it then loses a fraction of order 1×10^{-5} of its energy per collision — that is, nothing measurable. Above that threshold an inelastic channel opens and the electron can lose the whole quantum at once. A classical atom with a continuum of internal energies predicts a gradual onset instead. The two predictions differ qualitatively, which is what makes the experiment decisive.]

72.1.2 What was known in 1913

[Reserved: Franck and Hertz's earlier collision work on noble gases, undertaken to measure ionization potentials and mean free paths, not to test Bohr; the state of the art in electron-impact technique; and the fact — important for the honest history of Section 72.5.2 — that the pair were working within an ionization-potential programme and were not looking for excitation. Bohr's papers [396, 397] had appeared months earlier and are not cited in [417].]

72.2 Apparatus

72.2.1 The tube

[Reserved: an evacuated glass vessel containing a drop of mercury, held in an oven at about $450\,\mathrm{K}$ *so that the saturated vapour pressure gives an electron mean free path much shorter than the cathode–grid distance — the condition that guarantees many collisions and hence a periodic structure rather than a single step. A heated platinum-wire cathode; a wire-gauze grid at a positive accelerating potential; a collecting anode a short distance beyond it. Geometry, materials, dimensions and the oven's temperature stability to be given as measured, together with the vapour-pressure curve that converts oven temperature to number density.]*

72.2.2 Electrical arrangement and the systematics it controls

[Reserved: the accelerating voltage between cathode and grid; the small retarding *bias of a few tenths of a volt between grid and anode, which is the essential element of the design — an electron that has just lost its energy in an inelastic collision cannot climb it and is not collected, so the anode current falls. Measurement of the current with a galvanometer. Systematics to be enumerated and bounded: the contact-potential difference between the electrodes, which offsets the whole curve without changing the spacing; space charge; thermionic energy spread of the cathode; and the possibility of double inelastic collisions.]*

72.3 Procedure

[Reserved: bring the oven to temperature and verify stability; set the retarding bias; sweep the accelerating voltage from zero upwards in steps and record the anode current at each step, repeating in both directions to bound hysteresis; repeat the sweep at several oven temperatures to show that the spacing *of the features is independent of vapour pressure while their sharpness is not; repeat with the retarding bias varied to show the features are not an artefact of it. The analysis procedure is to be fixed in advance: locate each current maximum, fit the maxima against their ordinal number, and take the slope as the excitation energy, so that the contact potential appears only in the intercept and drops out of the result.]*

72.4 Observations

72.4.1 The current–voltage curve

[Reserved: the anode current rises with accelerating voltage, drops abruptly near $4.9\,\mathrm{V}$*, rises again, drops near* $9.8\,\mathrm{V}$*, and so on, giving a series of maxima equally spaced by* $4.9\,\mathrm{V}$ *[417]; the original work followed the sequence far enough to see several periods. To be tabulated: the accelerating voltage at each maximum with its uncertainty, the fitted spacing, the intercept and its interpretation as a contact potential, and the oven temperature and derived vapour*

density for each run — all in SI, with the uncertainty budget assembled in the manner of [1945, 2129].]

Phenomenon 72.1 (Periodic structure of the current–voltage curve)**.** In a tube containing mercury vapour, with electrons accelerated from a hot cathode towards a grid and collected beyond it against a small retarding bias, the anode current rises with the accelerating voltage, drops abruptly near 4.9 V, rises again, drops again near 9.8 V, and continues so for as many periods as the tube will sustain: the current maxima are equally spaced, and the spacing is 4.9 V [417]. The *spacing* does not depend on the vapour pressure, on the geometry of the tube or on the size of the retarding bias — these govern only how sharp the features are and where the first one falls — so it is a property of the mercury atom and not of the apparatus.

Derivation. Suppose with [396] that the atom possesses discrete internal energies and that the lowest excitation costs $\varepsilon = E_1 - E_0$. An electron of kinetic energy below ε cannot excite the atom and can only scatter elastically. Because the electron is lighter than a mercury atom by a factor of about 4×10^5, an elastic collision transfers at most the fraction $4m_\mathrm{e}/M$ of its energy, some parts in 10^5, which is nothing measurable; the electron therefore crosses the tube carrying essentially the whole energy eV it gained from the field, climbs the retarding bias and is collected. The current rises with V.

When eV first reaches ε — which happens just in front of the grid, where the accelerating potential has been fully traversed — the inelastic channel opens. An electron that takes it is left with almost no kinetic energy, cannot climb the retarding bias between grid and anode, and is not collected: the current falls. Raising V further, the electron regains energy after its collision and is collected again, so the current recovers; but as soon as it can acquire a further ε before reaching the grid it suffers a *second* inelastic collision and is again stopped. The condition for the k-th drop is therefore

$$eV_k = k\varepsilon + eV_\mathrm{c}\,, \tag{72.1}$$

with V_c the contact-potential difference between the electrodes, so that successive maxima are equally spaced by

$$\Delta V = \frac{\varepsilon}{e}\,. \tag{72.2}$$

Note what Equation (72.1) says about the systematics: the contact potential enters the intercept and cancels from the spacing, which is why the analysis fits the maxima against their ordinal number and takes the slope rather than reading a single threshold. A short mean free path — that is, a high vapour pressure — is what makes the collisions numerous enough for the structure to be periodic at all, but it does not enter Equation (72.2). A classical atom, whose internal energy could be changed by any amount however small, predicts a gradual onset of energy loss and no periodic structure whatever. □

72.4.2 The emission line at 253.7 nanometres

[Reserved: the companion measurement, and the one that closes the argument [416]. When and only when the electron energy exceeds the threshold, the vapour emits ultraviolet light at a single wavelength of 253.7 nm*. The photon energy $hc/\lambda = 4.89$ eV agrees with the electrical threshold, so the energy the electron loses is the energy the atom radiates — the mechanical and the optical measurements of the same level interval, made in the same tube. This is the point at which the experiment tests [396] and not merely the existence of a threshold.]*

Phenomenon 72.2 (The energy absorbed is re-emitted as a single spectral line)**.** When, and only when, the electron energy exceeds the threshold of Phenomenon 72.1, the mercury vapour in the tube emits ultraviolet light — and at one wavelength only, 253.7 nm [416]. The

energy of a single quantum of that light equals, within the precision of the 1914 measurement, the energy the electron is observed to lose. The mechanical and the optical determinations of the same atomic level interval are thus made on the same atoms in the same apparatus, and they agree.

Derivation. By the Planck–Einstein relation the energy of one quantum of light of wavelength λ is $E = hc/\lambda$. With h and c at their present SI values [2333] the product hc is 1239.84 eV nm, so a wavelength of 253.7 nm corresponds to 4.89 eV. The electrical threshold of Phenomenon 72.1 is 4.9 V, so by Equation (72.2) the two numbers are the same quantity measured two ways.

The identification is Bohr's postulate [396] that a transition between stationary states emits a single quantum whose energy is the difference of the two level energies: the electron delivers ε to the atom by collision, and the atom returns ε to the field as one quantum on decaying. The agreement is what makes this a test of that postulate rather than of discreteness alone — a threshold by itself shows only that the atom accepts energy in a definite amount, whereas the coincidence of the threshold with hc/λ shows that the amount accepted is the energy of a level whose existence the spectrum independently records. □

72.4.3 Uncertainties and the dataset

[Reserved: statistical uncertainty from the current measurement and from locating a maximum on a rounded peak; systematic contributions from the voltage calibration, the electron energy spread, and residual gas. Present-day teaching-laboratory realizations reach a few parts in a thousand on the spacing, which is enough to resolve the discrepancy discussed in Section 72.6.2. The digitized curve and the fitted maxima are the chapter's dataset; the SI traceability of the voltage scale is that of Chapter 2 and rests on the defining constants of [2333].]

72.5 Interpretation

72.5.1 Discrete energy transfer

[Reserved: below threshold the electron loses essentially nothing per collision, so it arrives at the grid with the full energy eV; at threshold it can transfer exactly one quantum and is then stopped by the retarding bias; beyond that it re-accelerates and can transfer a second quantum after a further 4.9 V*, which is why the structure is periodic rather than a single edge. The spacing is therefore an atomic energy interval measured in volts, and its independence of vapour pressure, of tube geometry and of the retarding bias is what identifies it as a property of the mercury atom. This is the phenomenon environment the chapter will carry, with its derivation inline.]*

72.5.2 What the authors concluded, and why it was wrong

[Reserved: Franck and Hertz reported 4.9 V *as the ionization potential of mercury [417], on the reasoning that a sudden onset of energy loss meant an electron had been torn off, and they explicitly denied that their result supported Bohr. The reading is wrong: the true first ionization potential of mercury is* 10.44 V*, and* 4.9 eV *is the excitation energy of the lowest triplet state that radiates at* 253.7 nm*. The section states the error plainly and identifies what would have settled it at the time — the resonance line of Section 72.4.2 was in their own data.]*

72.5.3 Bohr's correction and its experimental confirmation

[Reserved: Bohr's reinterpretation [425], which identified the threshold as a resonance potential — an excitation, not an ionization — and noted that the emitted line is exactly what his postulates require of a return to the ground state. Independent confirmation by Davis and Goucher, who separated excitation from ionization in the same vapour by collecting positive ions [443], and by Franck and Einsporn's refined measurement [467], which resolved further excitation potentials and in which Franck accepted the correction. The episode is a case study for Chapter 1: a correct measurement, a wrong inference, and the further measurement that decided between them.]

72.6 Modern repetitions and precision

72.6.1 The neon variant

[Reserved: neon replaces mercury in most teaching laboratories — no oven is needed and the excitation region is visible as a sequence of orange glow layers that migrate towards the cathode as the voltage is raised, so the periodic structure can be seen as well as plotted. The spacing corresponds to excitation of the neon levels near 18.7 eV*, and because several close-lying levels contribute, the measured spacing depends on the current and on the geometry in a way that the mercury curve does not — a systematic studied and quantified by Rapior, Sengstock and Baev [2102].]*

Phenomenon 72.3 (The neon variant and its visible excitation layers)**.** With neon in place of mercury the current–voltage curve shows the same periodic structure, with a spacing near 18.7 V, and the excitation regions are directly visible: the tube glows in a sequence of orange layers, one more appearing for each further period, each layer migrating towards the cathode as the accelerating voltage is raised. The periodicity can therefore be seen as well as plotted. Because several close-lying neon levels are accessible at once, the measured spacing depends on the discharge current and on the electrode geometry in a way that the mercury spacing does not [2102].

Derivation pending #67

Why the glow layers sit where they do: an electron starting from rest at the cathode reaches the excitation energy after falling through a fixed potential difference, so the luminous shells mark equal potential intervals, and their migration towards the cathode as the voltage is raised follows from the field distribution in the tube. The dependence of the effective spacing on current and geometry requires in addition the branching among the accessible neon levels and the collision cross-section of each.

72.6.2 Fine structure of the mercury curve

[Reserved: the canonical figure hides real structure. Mercury has a metastable level at 4.67 eV *as well as the radiating one at* 4.89 eV*, and Franck and Einsporn resolved both [467]; higher-resolution modern curves show additional steps and a spacing that drifts with the number of the maximum [2102]. The honest statement is that "the" Franck–Hertz spacing is an effective quantity averaging over accessible levels, and that agreement to three figures with a single level energy is a coincidence of the usual apparatus, not a property of the method.]*

Phenomenon 72.4 (The spacing is an effective average over accessible levels)**.** The canonical mercury curve hides real structure. Mercury has a metastable level at 4.67 eV as well as the radiating one at 4.89 eV, and electrons excite both; Franck and Einsporn resolved

the two, together with further excitation potentials above them [467]. Higher-resolution modern curves show additional steps, and a spacing that drifts systematically with the ordinal number of the maximum rather than staying constant [2102]. What the experiment measures is therefore an effective interval averaging over the levels the electrons can reach in the particular tube; its agreement to three figures with a single level energy [2420] is a property of the usual apparatus, not of the method.

Derivation pending #68

The effective spacing: the excitation cross-sections of the competing mercury levels as functions of electron energy, and the weighted mean interval they produce in a tube of given vapour density and geometry; and the mechanism of the drift with maximum number, in which electrons surviving several accelerating stages sample a different mixture of levels from those excited near the first drop.

72.6.3 Present-day level energies

[Reserved: the mercury level energies as now known from precision spectroscopy [2420], quoted to the digits the experiment can be compared against, together with the current values of h and e [2333, 2431]; the electrical measurement of an atomic energy interval is today a demonstration rather than a determination, since the spectroscopic route (Chapter 90) is better by many orders of magnitude. A short table comparing the 1914 value, the 1920 refinement and the modern value belongs here.]

72.7 What the experiment settles

[Reserved: closing assessment. The experiment establishes, by a purely mechanical measurement, that an atom accepts energy only in discrete amounts, and it ties those amounts to the emitted spectrum; it does not establish anything about orbits, and the Bohr model's mechanical picture is not tested by it — which is why the result survives the model's abandonment intact and constrains Chapters 77 and 88 directly. Franck and Hertz received the Nobel Prize in Physics for 1925, awarded in 1926, and Franck's Nobel lecture [543] is also the clearest statement of how he came to accept Bohr's reading. Cross-references to the other experiments of this part (Chapters 67 and 80) and to the inelastic-collision formalism of Chapter 86.]

Chapter 73

Experiment: The Blackbody Spectrum Measured

Chapter 68 makes a prediction of an unusual kind. It does not say what some particular substance will do; it says that the light leaving a small hole in any isothermal cavity is governed by one universal function of frequency and temperature, that thermodynamics fixes the integral of that function and its behaviour under a change of temperature but not the function itself, and that the function is Planck's law — which contains a constant of nature absent from classical physics. That chapter derives the law. It does not show it being measured, and the omission matters more here than anywhere else in this treatise, because in this one case the measurement came *first*. Planck did not predict the spectrum and wait for the laboratory; the laboratory handed him a spectrum that no classical theory could produce, and he wrote down a formula to fit it. This chapter is that measurement.

Five experiments are reported, in the order in which they were made. Stefan reads a fourth-power law out of somebody else's thermopile deflections [211]; Lummer and Kurlbaum build the first object whose emission may honestly be called Kirchhoff's universal function [282]; Lummer and Pringsheim take its spectrum far enough into the infrared to kill the accepted distribution law [284]; Rubens and Kurlbaum, by fixing the wavelength and sweeping the temperature instead, produce the single result that Planck was told about on 7 October 1900 and answered the same evening [294, 297]. The fifth is modern: under the SI of 2019 the same cavity, viewed by a calibrated radiometer, is a primary realization of the kelvin [2333, 2334], so that the constant Planck read off the Reichsanstalt data now defines the units in which the measurement is expressed. Together they are the reason Part VIII exists.

Derivation pending #69

Experiment: The Blackbody Spectrum Measured: what is pending is the tabulation of the measured radiance against wavelength and temperature for each experiment, in SI with its uncertainty budget, together with the residuals against the Wien and the Planck distributions. The phenomena stated below each carry their derivation inline.

73.1 Stefan: the fourth power, from Tyndall's data (1879)

The first quantitative law of thermal radiation was obtained without an experiment. Stefan had no apparatus in this matter; he had Tyndall's published numbers for the total emission of an incandescent platinum wire [165], and he asked what power of the absolute temperature would reproduce them [211]. The experiment of record is therefore Tyndall's and the measurement of record is Stefan's re-analysis of it, a distinction this section keeps because

it is what makes the honest assessment of the result possible.

73.1.1 Apparatus

[Reserved: Tyndall's arrangement — a platinum wire heated electrically to incandescence in air, its temperature judged from its colour against a scale of known incandescent hues, and its total radiant emission collected and measured with a thermopile and galvanometer. To be given: the wire dimensions and its surface condition, the geometry that fixes what fraction of the emitted power reaches the pile, the thermopile's linearity, and above all the provenance and uncertainty of the two temperature values, which dominate everything that follows.]

73.1.2 Procedure

[Reserved: raise the wire through a sequence of temperatures and record the galvanometer deflection at each. The quantity used is a ratio *of deflections between a pair of temperatures, which cancels the absolute calibration of the thermopile and leaves only its linearity to be established — an economy worth stating explicitly, since it is the reason a crude instrument could settle an exponent. Stefan's own step is arithmetical: convert both temperatures to the absolute scale and ask what power of T reproduces the measured ratio.]*

73.1.3 Observations and data

[Reserved: the deflection at each temperature and the derived emission ratios, with Stefan's parallel re-analysis of the cooling measurements of Dulong and Petit, which he treated in the same paper and by the same method. To be tabulated when written: temperature pairs in kelvin, measured emission ratio, exponent implied by each pair, and the propagated uncertainty of that exponent.]

Phenomenon 73.1 (Total emission rises as the fourth power of the absolute temperature)**.** The total power radiated per unit area by a hot body, integrated over all frequencies, rises much faster than in proportion to its temperature. Stefan found that Tyndall's platinum wire, taken from 798 K to 1473 K, radiated 11.7 times as much power at the higher temperature as at the lower, and identified this as the fourth power of the ratio of the absolute temperatures [211]. The thermodynamic derivation of the same exponent, which requires the radiation pressure and no data at all, came five years later [227].

Derivation. Take the law in the form $M = \varepsilon\sigma T^4$ of Chapter 68 and suppose the emissivity ε of the wire to be the same at both temperatures. The absolute calibration then cancels from a ratio,

$$\frac{M_2}{M_1} = \left(\frac{T_2}{T_1}\right)^4 , \tag{73.1}$$

and with $T_1 = 798\,\mathrm{K}$ and $T_2 = 1473\,\mathrm{K}$ the temperature ratio is 1.8459, whose fourth power is 11.61. Against Tyndall's 11.7 that is agreement to better than one percent, which is how the law was found.

It is also better agreement than the data deserve, and the derivation should say so. Solving Equation (73.1) for the exponent instead of assuming it gives

$$n = \frac{\ln\left(M_2/M_1\right)}{\ln\left(T_2/T_1\right)} = 4.01 , \tag{73.2}$$

but the sensitivity of that number to the inputs is severe. From Equation (73.2), a fractional error δ in the measured emission ratio moves n by $\delta/\ln(T_2/T_1) \simeq 1.6\,\delta$, so one percent on the

ratio is worth 0.016 on the exponent and is negligible. An error in the hotter temperature is not: it enters as

$$\delta n = \frac{n}{\ln\left(T_2/T_1\right)} \, \frac{\delta T_2}{T_2} \,, \tag{73.3}$$

which for an error of 100 K in T_2 alone gives $\delta n \approx 0.44$. Tyndall judged his temperature from the colour of the wire, so an error of that size is optimistic rather than pessimistic, and the datum by itself is consistent with any exponent from roughly 3.5 to 4.5. Two further defects work in the same direction: platinum is not a black body, and its emissivity rises with temperature, so the assumption that ε cancels from Equation (73.1) is false in the direction that inflates the apparent exponent. The fourth power is right; it was not *established* here, and the sections that follow are the reason it is now believed. □

73.1.4 Interpretation

[Reserved: what the analysis does and does not establish. It establishes that the emission of a hot solid rises far faster than linearly with temperature and is consistent with a fourth power; it does not establish the exponent to better than a few tenths, nor does it concern a black body at all. The section should state plainly that the law became secure only when Boltzmann derived it from the second law together with the radiation pressure [227], and when the Reichsanstalt cavity of Section 73.2 made it testable on a radiator that really was black. Forward link: the Stefan–Boltzmann constant is today not measured but computed from the defining constants [2333, 2431], so this experiment has become a test of the defining constants rather than a determination of anything.]

73.1.5 Primary references

[211]; the thermodynamic derivation of the same law is [227]. Tyndall's measurements [165] are used here as Stefan reports them and the original was not consulted.

73.2 Lummer and Kurlbaum: a black body one can build (1898)

Kirchhoff's theorem tells a laboratory what to build — a small hole in a large isothermal enclosure — and says nothing about how well any actual hole will do [153]. Until that gap was closed there was no object in the world whose spectrum could be compared with the universal function, and every measurement of thermal radiation was a measurement of a material. The electrically heated cavity radiator of Lummer and Kurlbaum is where that changes [282]: it is the first artefact for which the departure from blackness can be estimated rather than hoped for, and it is the direct ancestor of every radiometric standard in use today.

73.2.1 Apparatus

[Reserved: a cylindrical cavity heated electrically along its whole length rather than from one end, so that the wall temperature is uniform; internal diaphragms and a blackened, roughened inner surface, both of which raise the number of reflections a ray must survive before it can leave; a small aperture at one end and a water-cooled front plate to define it; thermocouples distributed along the wall to establish isothermality rather than assume it. Detection is by the surface bolometer Lummer and Kurlbaum had developed for the purpose, with a thermopile as an independent check. To be given: cavity length and bore, aperture diameter, the resulting ratio of aperture area to total interior area, the wall temperature range and its uniformity, and the bolometer's time constant and noise floor.]

73.2.2 Procedure

[Reserved: bring the radiator to a chosen temperature and hold it there; verify isothermality by traversing a thermocouple along the wall and by comparing the emission through the aperture before and after reversing the heating current. Then the two tests that make the claim of blackness an experimental one rather than a design intention: vary the aperture diameter and show that the radiance emerging is unchanged (a black aperture has no memory of its size, a grey one does), and compare two cavities built of different wall materials at the same temperature and show that they agree. Finally, measure the total emission as a function of temperature and compare with the fourth-power law of Section 73.1, this time on a radiator that is actually black.]

73.2.3 Observations and data

[Reserved: total radiance against wall temperature over the working range, with the fitted exponent and the fitted Stefan–Boltzmann constant; the invariance of the radiance under change of aperture and of wall material, quoted as a bound rather than as an equality; the residual temperature non-uniformity along the wall and what it contributes to the uncertainty. All in SI with uncertainties.]

Phenomenon 73.2 (The aperture of an isothermal cavity is black to a computable accuracy)**.** An electrically heated isothermal enclosure viewed through a small aperture emits, to within a small and calculable correction, the universal function of [153], and its emission is insensitive to the material of the walls: a mediocre absorber makes an excellent cavity provided the aperture is small enough. This is what Lummer and Kurlbaum realized in hardware [282], and it is the reason every spectrum reported in the sections that follow is a measurement of a function of nature rather than of a surface.

Derivation. Let the enclosure be isothermal, let its wall have emissivity ε — equal to its absorptivity by Chapter 68 — and let

$$f := \frac{A_{\text{hole}}}{A_{\text{total}}} \tag{73.4}$$

be the fraction of the interior area occupied by the aperture. Compute the *absorptance* of the aperture, which by Kirchhoff's theorem is also its emissivity. Send unit flux in through the hole and assume, as an isothermal cavity with a rough wall makes reasonable, that after each diffuse reflection the flux is spread uniformly over the whole interior. The first encounter with the wall absorbs ε and reflects $1-\varepsilon$; of what is reflected, the fraction f finds the hole and escapes and the fraction $1-f$ meets the wall again. The survival factor per round trip is therefore $(1-\varepsilon)\,(1-f)$, and summing the geometric series of absorbed contributions,

$$\varepsilon_{\text{eff}} = \varepsilon \sum_{k\geq 0} \left[(1-\varepsilon)\,(1-f)\right]^{k} = \frac{\varepsilon}{1-(1-\varepsilon)\,(1-f)}\,,$$

and expanding the denominator, $1-(1-\varepsilon)(1-f) = \varepsilon + f - \varepsilon f$, gives

$$\varepsilon_{\text{eff}} = \frac{\varepsilon}{\varepsilon + (1-\varepsilon)\,f}\,. \tag{73.5}$$

The design rule follows by inspection. The wall material enters only through the product $(1-\varepsilon)\,f$, so it is the *aperture fraction* and not the blackness of the paint that carries the accuracy: a wall of $\varepsilon = 0.5$ — barely better than polished — with an aperture fraction $f = 10^{-3}$ gives $\varepsilon_{\text{eff}} = 0.999$, while the same wall with the lid off would be half a black body. Diaphragms and a roughened surface work by making the effective f seen by an escaping ray smaller still.

Two honesties are owed here. First, Equation (73.5) is the modern integrating-cavity estimate and not Lummer and Kurlbaum's own formula; they argued the case by construction and by the experimental invariances of the procedure above, which is the stronger argument since it does not assume perfect diffusion. Second, the formula assumes isothermality exactly, and a real cavity is graded in temperature along its length; that error is not covered by Equation (73.5) and has to be bounded separately, which is why the thermocouple traverse is part of the apparatus and not of the diagnostics. □

73.2.4 Interpretation

[Reserved: the significance is metrological before it is theoretical. Kirchhoff's universal function ceases to be an idealization and becomes an object with a serial number, so that a spectrum measured in Berlin can be compared with a formula and with a spectrum measured anywhere else. The cavity radiator became the primary radiometric standard of the Physikalisch-Technische Reichsanstalt and the ancestor of the fixed-point blackbodies of Section 73.5; the traceability chain and its uncertainty budget belong to Chapter 2. Worth recording plainly: the Reichsanstalt built this instrument to serve the German lighting industry, and the overthrow of classical physics was a by-product of a standards laboratory doing standards work.]

73.2.5 Primary references

[282]; the theorem the instrument realizes is [153].

73.3 Lummer and Pringsheim: Wien's law breaks (1899)

With a black body available, the universal function could at last be measured rather than guessed. Wien's distribution [272] was the accepted answer and fitted everything then in evidence; what Lummer and Pringsheim did was to push the measurement into the region where it was not in evidence, at long wavelength and high temperature, and there it failed [284, 292]. This is the point at which the accepted law of thermal radiation died, and it died to data rather than to an argument: no successor to Wien's formula existed for another year.

73.3.1 Apparatus

[Reserved: the cavity radiator of Section 73.2 as source; a spectrometer whose prism and windows are transparent well into the infrared, where glass and quartz are opaque, hence fluorite and rock-salt optics; a linear bolometer at the exit slit, with its galvanometer; and the calibration that the whole result rests on — the dispersion curve of the prism, which converts an angular position into a wavelength and an angular slit width into a bandwidth. To be given: prism material and apex angle, slit widths and the resolving power they imply, the bolometer strip dimensions, and the measured transmission of the optical train as a function of wavelength.]

73.3.2 Procedure

[Reserved: hold the radiator at a chosen temperature and scan the bolometer through the spectrum, recording deflection against angle. Reduce each scan to a spectral radiance: correct for the dispersion of the prism, so that equal angular steps become unequal wavelength intervals; for the transmission of prism, windows and mirrors; and for the spectral response of the bolometer. Repeat at each temperature over the working range. The comparison with

theory is then made by fitting the two constants of Wien's distribution to the short-wavelength part of each isotherm and examining the residuals at long wavelength — a procedure that must be fixed in advance, since a fit over the whole range would distribute the discrepancy and hide it.]

73.3.3 Observations and data

[Reserved: isotherms of spectral radiance against wavelength from about 300 K *to* 1650 K*, reaching about* 8 µm *[284], with the residuals against the fitted Wien distribution. The residuals are the observation: they are of one sign, they grow with the product of wavelength and temperature, and they exceed the estimated uncertainty of the measurement by a wide margin at the long-wavelength end. To be tabulated: wavelength, temperature, measured spectral radiance and its uncertainty, and the ratio of measurement to Wien's formula and to Planck's.]*

Phenomenon 73.3 (The measured spectrum exceeds Wien's distribution at long wavelength)**.** Measured with a black body of known temperature and a spectrometer transparent into the infrared, the spectral radiance of cavity radiation is systematically *greater* than Wien's distribution [272] predicts, and the excess is not random: it has one sign, it is negligible in the visible and near infrared where the formula had been tested, and it grows steadily with the product of wavelength and temperature. Lummer and Pringsheim found it out to about 8 µm over temperatures from about 300 K to 1650 K, far beyond any plausible error of their measurement [284].

Derivation. The excess is exactly what Planck's law of Chapter 68 requires, and its dependence on a single variable is the content of the phenomenon. Write the dimensionless variable

$$x := \frac{hc}{\lambda k_{\mathrm{B}} T} = \frac{c_2}{\lambda T}\,, \qquad c_2 := \frac{hc}{k_{\mathrm{B}}} = 1.438\,777 \times 10^{-2}\,\mathrm{m\,K}\,, \tag{73.6}$$

the second radiation constant, whose value is exact in the present SI because h, c and k_{B} are all defining constants [2333, 2431]. In terms of x the Planck spectral radiance carries the factor $1/(\mathrm{e}^{x}-1)$ and Wien's distribution the factor e^{-x}, with the same prefactor, so their ratio is

$$\frac{L_\lambda^{\mathrm{Planck}}}{L_\lambda^{\mathrm{Wien}}} = \frac{\mathrm{e}^{x}}{\mathrm{e}^{x}-1} = \frac{1}{1-\mathrm{e}^{-x}}\,. \tag{73.7}$$

This is greater than unity for every finite x, so the discrepancy can only have one sign, and it is a decreasing function of x — that is, an increasing function of λT — so it is invisible in the visible and unmistakable in the infrared. That it depends on λ and T only through their product is Wien's own displacement scaling of Chapter 68, which both formulas respect; it means the whole family of isotherms collapses onto one curve, and a residual that does not so collapse would indicate an instrumental error rather than a failure of the law.

Table 73.1 evaluates Equation (73.7). The numbers are computed here from Planck's law, not taken from any measurement, and they say what an experimenter had to reach before Wien's formula could be convicted: below $\lambda T \approx 3000\,\mu\mathrm{m\,K}$ the excess is under one percent and no instrument of 1899 could have seen it, which is why measurements confined to the visible and near infrared continued to support Wien's law while these did not. □

Table 73.1: The factor by which Planck's law exceeds Wien's distribution, from Equation (73.7), as a function of the product of wavelength and temperature. Computed from the formula; not a measurement. The short-wavelength end is the region in which Wien's law had been tested and found good.

λT	excess of Planck over Wien
3000 µm K	0.8 %
6000 µm K	10 %
9000 µm K	25 %
12 000 µm K	43 %

73.3.4 Interpretation

[Reserved: the accepted distribution law is falsified, and nothing replaces it. The section should resist the retrospective tidying that makes this look like a step towards Planck: for a year there was simply a measured curve with no formula, and the candidate interpolations then in circulation were fitted to it rather than derived. Also the methodological point, which Chapter 1 takes up: a law tested only where its predictions are insensitive to the thing at issue has not been tested, and the entire agreement of Wien's formula in the visible was of that kind — Table 73.1 shows the discrepancy there was below the noise. Honest caveat: the size of the deviations was disputed at the time, and settling the dispute needed the different and far more decisive measurement of Section 73.4.]

73.3.5 Primary references

[284, 292]; the distribution under test is [272] and the scaling law both formulas obey is [259].

73.4 Rubens and Kurlbaum: residual rays in the far infrared (1900)

This is the decisive measurement, and it is decisive because of an inversion of method. Every spectral measurement before it fixed the temperature and scanned the wavelength, so that the result depended on a dispersion calibration and a chain of transmission corrections. Rubens and Kurlbaum fixed the *wavelength* — at values far beyond anything a prism could reach — and swept the temperature [297]. What they obtained was a curve of radiance against temperature at one wavelength, needing no spectral calibration at all, and on that curve Wien's law and its successors make qualitatively different predictions. Rubens described the result to Planck privately on 7 October 1900; the formula that fits it was presented to the Deutsche Physikalische Gesellschaft twelve days later [294] and derived on 14 December [295].

73.4.1 Apparatus

*[Reserved: the cavity radiator of Section 73.2 as source, with the furnace extended downward in temperature by cooled baths; the residual-ray (*Reststrahlen*) train, in which the beam is reflected in succession from several polished crystal plates — fluorite for 24 µm, rock salt for 51.2 µm — each of which reflects strongly only in the narrow band near its lattice resonance and weakly elsewhere; and a bolometer or thermopile to measure what survives. To be given: the number of successive reflections, the measured reflectance of each plate in and out of band, the residual short-wavelength contamination of the emerging beam, and the detector's sensitivity and noise floor, which set the lowest temperature reachable.]*

73.4.2 Procedure

[Reserved: form the residual-ray beam and verify its purity, then hold the wavelength fixed and step the source through its temperature range, recording the detector deflection at each step.

Because the wavelength never changes, the dispersion curve, the transmission corrections and the detector's spectral response all drop out of the comparison, and only the linearity of the detector and the accuracy of the source temperature remain — this is the whole methodological advantage over Section 73.3. The candidate formulas of the day are then fitted to the measured curve and their residuals compared over the whole temperature range, not over a favourable part of it.]

73.4.3 Observations and data

[Reserved: radiance against source temperature at 24 µm *and at* 51.2 µm*, over a range extending from below room temperature to the upper limit of the furnace [297]. The observation is the* shape *of that curve: it rises without any sign of the ceiling Wien's distribution requires, and over the upper part of the range it is close to a straight line through the origin. To be tabulated: source temperature, measured radiance and its uncertainty at each wavelength, together with the residuals against each of the formulas Rubens and Kurlbaum tested — Wien's and the interpolations of Thiesen, of Rayleigh and of Planck.]*

Phenomenon 73.4 (At fixed long wavelength the radiance grows without bound, nearly in proportion to the temperature)**.** Far beyond the maximum of the spectrum the radiance of a black body does not saturate as the source is made hotter; it goes on rising, and over the upper part of the accessible range it rises very nearly in proportion to the absolute temperature. Rubens and Kurlbaum isolated 24 µm and 51.2 µm by repeated reflection from fluorite and rock salt and found this behaviour at both wavelengths [297]. Wien's distribution [272] requires the opposite: at fixed wavelength it approaches a finite ceiling as $T \to \infty$. The disagreement is not a matter of percentages, and it is what Planck was told about on 7 October 1900.

Derivation. Compare the two formulas of Chapter 68 at fixed λ as functions of T, using the variable $x = c_2/\lambda T$ of Equation (73.6). Wien's distribution carries the factor e^{-x}; as T grows, $x \to 0$ and $\mathrm{e}^{-x} \to 1$, so

$$L_\lambda^{\text{Wien}} \longrightarrow \frac{2hc^2}{\lambda^5} \qquad (T \to \infty)\ , \tag{73.8}$$

a ceiling fixed by the wavelength alone, which no temperature can exceed. Planck's law carries $1/(\mathrm{e}^x - 1)$ instead, and for $x \ll 1$ the expansion $\mathrm{e}^x - 1 \simeq x$ gives

$$L_\lambda^{\text{Planck}} \simeq \frac{2hc^2}{\lambda^5}\,\frac{1}{x} = \frac{2ck_{\mathrm{B}}T}{\lambda^4}\ , \tag{73.9}$$

in which h has cancelled: the radiance is proportional to T, grows without bound, and is the equipartition result of Chapter 55. One formula predicts a plateau and the other a rising straight line, so a single wavelength and a wide temperature range decide between them without any spectral calibration whatever — which is the point of the method.

The wavelengths chosen put the experiment squarely in the regime where Equation (73.9) applies. The crossover, $x = 1$, falls at $T = c_2/\lambda$, which is close to 600 K at 24 µm and to 281 K at 51.2 µm: above those temperatures the linear behaviour sets in, and at the longer wavelength every furnace temperature lies above it. The size of the resulting discrepancy is brutal. At 51.2 µm and a source at 1500 K one has $x = 0.187$, and by Equation (73.7) the true radiance exceeds Wien's by the factor $1/(1 - \mathrm{e}^{-x}) = 5.9$ — a factor of six, not a discrepancy of percentages, and beyond the reach of any excuse about calibration.

Two qualifications keep the statement honest. The proportionality to T is asymptotic, not exact: the next term of the expansion gives $L_\lambda \simeq (2ck_{\mathrm{B}}T/\lambda^4)(1 - x/2 + \dots)$, so the measured curve bends below the straight line at the cold end, by about 9 % at $x = 0.187$,

and the linear region has to be identified rather than assumed. And Equation (73.9) is a limit of Planck's law, not evidence for it: what the measurement establishes directly is that Wien's exponential is wrong at long wavelength. That a single formula reproduces both this limit and the visible spectrum of Section 73.3 is the additional claim, and it is what Rubens and Kurlbaum tested by fitting the competing interpolations to the same curve.

The purity of the residual-ray beam is what makes any of this measurable, and it is worth one line of arithmetic. A source at 1500 K emits overwhelmingly more power near its maximum, at a few micrometres, than at 51.2 µm, so a leak of short-wave radiation would swamp the signal. If a crystal plate reflects a fraction R_{in} in its residual-ray band and R_{out} outside it, then n successive reflections change the ratio of in-band to out-of-band flux by

$$\left(\frac{R_{\mathrm{in}}}{R_{\mathrm{out}}}\right)^{n}, \tag{73.10}$$

a geometric gain in purity for a merely linear loss of signal: a reflectance contrast of ten, which a strong lattice band easily provides, gives four orders of magnitude of suppression after four reflections. This is why a filter that no one can manufacture is replaced by a mirror used four times. □

73.4.4 Interpretation

[Reserved: the causal chain, which is the clearest instance in physics of a measurement dictating a formula and should be set out step by step with its dates. Rubens communicated the far-infrared result to Planck on 7 October 1900; Planck, who knew that the short-wavelength limit had to be Wien's, now knew that the long-wavelength limit had to be linear in T, and constructed the simplest expression for the second derivative of the entropy with respect to the energy that reduces to both — presented on 19 October [294]. That step is interpolation and Planck said so; the physics enters only with the combinatorial derivation of 14 December [295], in which the energy elements $h\nu$ appear. Fitting the new law to the Reichsanstalt data gave the first values of h and of k_B [300], and through the latter the first good values of the Avogadro constant and of the elementary charge — a bonus of the measurement that is often forgotten and that did more than anything else to make the formula credible before the hypothesis behind it was.]

73.4.5 Primary references

[297]; the formula it produced within the fortnight is [294], its derivation [295] and its considered exposition [300]. The distribution it falsified is [272], and the classical limit it exhibits is that of [296, 322].

73.5 Radiometric realization of the kelvin (2019)

The last experiment in this chapter is not a test of Planck's law but a use of it, and the change of status is the point. Since the revision of the SI that took effect in 2019 the kelvin is defined by fixing the numerical value of the Boltzmann constant, and h and c are fixed as well [2310, 2333]. Planck's law therefore contains no measured constant and no property of any substance: an absolute measurement of the spectral radiance leaving a cavity yields the thermodynamic temperature of its walls directly. The instrument of Section 73.2, which was built to make the universal function observable, has become a primary realization of a base unit.

73.5.1 Apparatus

[Reserved: a cavity radiator of the modern kind — either held at a fixed point of a pure metal or continuously variable — with an effective emissivity established by the argument of Phenomenon 73.2 and by direct reflectance measurement; a filter radiometer defining a narrow spectral band and a known solid angle, whose absolute spectral responsivity is the quantity everything rests on; and the chain that establishes that responsivity, running through a trap detector to an absolute cryogenic radiometer [1762], which equates optical power to electrical power dissipated in the same receiver and so ties the measurement to the watt and thence to the SI base units. To be given: the band centre and width, the aperture geometry and its dimensional metrology, the emissivity budget of the cavity, and the responsivity uncertainty.]

73.5.2 Procedure

[Reserved: calibrate the filter radiometer against the cryogenic radiometer; measure the spectral radiance of the cavity in the chosen band, with the geometry known dimensionally; invert Planck's law for T as in Equation (73.12) below. Cross-checks that make the result a realization rather than a reading: repeat in a second band, where the sensitivity factor of Equation (73.13) differs, and compare with an independent primary thermometer of a different physical kind. The comparison against the defined temperature scale in use, and the difference between a thermodynamic temperature and a scale value, belong to Chapter 2.]

73.5.3 Observations and data

[Reserved: measured spectral radiance in each band, the inferred thermodynamic temperature, and the uncertainty budget — which is dominated by the radiometric scale and by the cavity emissivity, and which shrinks relative to the temperature by the factor computed in the derivation below. To be tabulated when written: band, measured radiance, sensitivity factor, inferred temperature and combined standard uncertainty, in SI throughout.]

Phenomenon 73.5 (Thermodynamic temperature is readable from an absolute spectral radiance)**.** The spectral radiance emerging from a black cavity determines the thermodynamic temperature of its walls through Planck's law alone. In the SI that took effect in 2019 the constants entering that law — h, c and k_{B} — are all fixed by definition [2333, 2431], so no material property and no reference temperature intervene: the measurement is a primary realization of the kelvin [2334]. The same relation that was extracted from the Reichsanstalt spectra now supplies the unit in which those spectra would be reported.

Derivation. Write Planck's law of Chapter 68 as a spectral radiance,

$$L_\lambda(\lambda, T) = \frac{2hc^2}{\lambda^5}\,\frac{1}{\mathrm{e}^{x}-1}\,, \qquad x = \frac{c_2}{\lambda T}\,, \tag{73.11}$$

and solve for the temperature. Rearranging Equation (73.11) gives $\mathrm{e}^{x}-1 = 2hc^2/(\lambda^5 L_\lambda)$, hence x by a logarithm and hence

$$T = \frac{c_2}{\lambda}\left[\ln\left(1+\frac{2hc^2}{\lambda^5 L_\lambda}\right)\right]^{-1}. \tag{73.12}$$

Everything on the right of Equation (73.12) is either a defining constant or a measured quantity: the wavelength, which is metrology of the filter, and the absolute spectral radiance, which is metrology of the radiometer. No thermometer appears.

The design of such a measurement turns on how an error in L_λ propagates into T. Differentiating Equation (73.11) at fixed λ, and using $\mathrm{d}x/x = -\mathrm{d}T/T$,

$$\frac{\mathrm{d}L_\lambda}{L_\lambda} = -\frac{\mathrm{e}^x}{\mathrm{e}^x - 1}\,\mathrm{d}x = \frac{x\mathrm{e}^x}{\mathrm{e}^x - 1}\,\frac{\mathrm{d}T}{T}\,,$$

so that

$$\frac{\mathrm{d}T}{T} = \frac{1 - \mathrm{e}^{-x}}{x}\,\frac{\mathrm{d}L_\lambda}{L_\lambda}\,. \tag{73.13}$$

The factor in front is close to unity when $x \ll 1$ and close to $1/x$ when $x \gg 1$, so radiometric thermometry is done at short wavelengths, on the steep Wien side of the spectrum, where a coarse measurement of radiance buys a fine measurement of temperature. At $\lambda = 650\,\mathrm{nm}$ and $T = 1235\,\mathrm{K}$ — representative of a metal fixed point — Equation (73.6) gives $x = 17.9$, so the sensitivity factor is 0.056 and an error of 1 % in the measured radiance appears as 0.056 of a percent in temperature, about 0.7 K. The same one percent at $\lambda = 51.2\,\mu\mathrm{m}$, where x is small, would appear almost undiminished. The far infrared that decided the physics is the worst possible place to realize the unit, and for exactly the reason that made it decisive: there the spectrum has forgotten h.

One consequence should be stated plainly, because it inverts the logic of every earlier section. Since h, c and k_B are now defined rather than measured, Equation (73.12) cannot be used to test Planck's law — it presupposes it. The law is tested instead by comparing this realization against primary thermometers of a different physical kind, and, on a far more exacting scale, by the astrophysical spectrum of Chapter 51. □

73.5.4 Interpretation

[Reserved: what has and has not changed. The 1898 cavity and the present one are the same instrument serving opposite purposes: then a means of exhibiting an unknown universal function, now a means of realizing a defined unit. The section should be explicit that the present arrangement is not evidence for Planck's law and say what is: the mutual consistency of primary thermometers resting on different physics, and the cosmic microwave background spectrum of Chapter 51, which is the most exactly Planckian spectrum known. Also the closing of the historical circle — the constant Planck read off the Reichsanstalt data in 1900 [300] is now a defining constant of the SI [2333], so that his 1901 values may be quoted as a percent-level verification of a definition.]

73.5.5 Primary references

[2333, 2431]. These are the normative and the recommended-values documents; the primary literature of absolute cryogenic radiometry and of the filter-radiometer realizations is not cited here because `references.bib` carries no entry for it, and no key has been invented to stand in its place.

73.6 What the measurements settle

The sequence in Table 73.2 has a shape that is unusual in this treatise and worth naming. Elsewhere a theory predicts and an experiment checks; here the experiments arrive first and the theory is built to fit them, and the constant h enters physics not as a postulate but as the number required to reconcile two limits that were each measured before either was understood. Wien's exponential is the short-wavelength limit and was measured by 1896; the proportionality to T is the long-wavelength limit and was measured in October 1900; the formula joining them was written within days of the second, and the hypothesis that explains the formula came two months later [294, 295].

Table 73.2: The measurements reported in this chapter and what each establishes. The third column states the claim the experiment is entitled to, which in the first case is markedly weaker than the claim it is usually credited with.

Experiment	Quantity fixed	What it establishes
Stefan 1879	neither	total emission rises as a high power of T, consistent with the fourth but not resolving it
Lummer–Kurlbaum 1898	—	an artefact whose emissivity is calculable and near unity; the universal function becomes observable
Lummer–Pringsheim 1899	T fixed, λ scanned	the accepted distribution law is wrong at long wavelength, by more than any error of measurement
Rubens–Kurlbaum 1900	λ fixed, T swept	the radiance grows without bound and nearly linearly in T; Wien's ceiling does not exist
Radiometric kelvin 2019	all three constants	Planck's law realizes the unit of temperature rather than being tested by it

What the chapter does *not* settle should be as clear. None of these measurements shows that light is granular. They show that the energy exchanged between a cavity wall and the field comes in units proportional to the frequency, which is what Planck claimed and no more; the light quantum is a further and independent step, and the experiments that force it are those of Chapter 69. Nor does the chapter settle the exponent Stefan is remembered for — Phenomenon 73.1 gives it as an inference from a ratio whose uncertainty allows several tenths either way, and it is Boltzmann's derivation [227] together with the cavity work of Section 73.2 that makes it secure. The one thing established here beyond argument is negative and decisive: no distribution law derivable from classical electrodynamics and equipartition can fit the curve that Rubens and Kurlbaum measured at 51.2 µm, because equipartition gives the linear rise at long wavelength only at the cost of an infinite total energy [296, 322]. The measurement that forced quantisation is a measurement of the far infrared, made by a standards laboratory, in the service of the lighting industry.

Chapter 74

Experiment: The Photoelectric Effect and Compton Scattering

Chapter 69 states a hypothesis and derives its consequences: light of frequency ν is emitted and absorbed in indivisible amounts $h\nu$ carrying momentum $h\nu/c$, from which follow the photoelectric equation, the Compton shift, and the requirement that energy and momentum balance in each single scattering act. It names the experiments that bear on all three and reports none of them. This chapter is that report.

It is organized around one question, and the question is sharper than the textbook treatment usually admits: *which* of these measurements actually requires the light to be quantized, as opposed to the matter that absorbs it. The answer is unkind to the most famous of them. The photoelectric effect, in every form Hertz, Lenard and Millikan gave it, is reproduced by a quantized atom absorbing from an entirely classical field [1402], so the first three sections below establish the equation of Phenomenon 69.2 and the constant in it without establishing the light quantum at all. What carries the weight is momentum and indivisibility: the wavelength shift of Section 74.4, which no driven classical oscillator can produce; the coincidence experiment of Section 74.5, which excludes conservation-on-the-average and with it the last serious counter-theory; and the anticorrelation of Section 74.6, which is the modern experiment that no classical field of any strength or statistics can reproduce. The sections run in that order, which is also chronological, and the closing Section 74.7 states what each is and is not entitled to claim.

Derivation pending #70

Experiment: The Photoelectric Effect and Compton Scattering: what is pending is the tabulation of the measured numbers for each experiment, in SI with its uncertainty budget — stopping potential against frequency for each emitting metal, scattered wavelength against scattering angle, coincidence and singles counts against resolving time, and the gated detection probabilities of the beamsplitter experiment. The phenomena stated below carry their derivations inline except where a pending derivation is marked against them individually.

74.1 Hertz and Hallwachs: ultraviolet light discharges a metal (1887–1888)

The effect was found by a man who did not want it. Hertz was engaged in the experiment that produced the first electromagnetic waves and confirmed Maxwell's theory; the photoelectric effect entered his apparatus as an interference, was isolated because it would not go away,

and was reported as an anomaly whose explanation he could not offer [234]. Hallwachs turned it into a phenomenon the following year with an apparatus a school can build [238].

74.1.1 Apparatus

Hertz's arrangement is two spark gaps. An induction coil drives the primary gap, whose oscillatory discharge radiates; a wire loop of adjustable size serves as resonator, and the length of the spark that can be drawn across a micrometer gap in that loop measures the induced electromotive force. The relevant addition is optical, and it is the part that matters here: screens interposed between the two gaps, of glass and of quartz, and a quartz prism with which the light of the primary spark can be dispersed and one region of the spectrum at a time allowed to fall on the micrometer gap. Hallwachs' apparatus is a freshly polished zinc plate mounted on an insulating support and connected to a gold-leaf electroscope, illuminated by the light of a carbon arc.

74.1.2 Procedure

Hertz measured the maximum length of the resonator spark with the micrometer gap exposed to the light of the primary spark and again with it shielded. The diagnostic sequence is the interesting part: enclosing the resonator gap in an opaque box shortens the spark; the effect survives when the box is fitted with a quartz window and disappears when the window is glass; and dispersing the light with a quartz prism locates the active agent beyond the violet end of the visible spectrum. Hallwachs charged his zinc plate negatively, watched the electroscope leaf under illumination, then repeated with the plate charged positively, and finally with the plate initially uncharged.

74.1.3 Observations and data

Illumination of the micrometer gap by ultraviolet light lowers the potential difference at which the spark passes, so that a longer gap sparks at the same drive; glass extinguishes the effect and quartz does not, which places the agent in the ultraviolet [234]. On Hallwachs' plate the sign of the charge decides everything: a negatively charged plate loses its charge under illumination, a positively charged plate does not, and an initially uncharged plate charges up *positively* [238]. The carriers liberated are therefore negative.

[Reserved: the tabulation. For Hertz, the maximum resonator spark length in metres against the screening condition and against the position on the dispersed spectrum, with the reproducibility of the spark-length measurement, which is the whole uncertainty of the experiment. For Hallwachs, the rate of fall of the electroscope leaf against illumination and against the sign and magnitude of the initial charge, converted to a charge loss per second in SI. Neither original was consulted for this chapter and no figures are quoted from memory.]

Phenomenon 74.1 (Ultraviolet light liberates negative charge from a clean metal surface)**.** Light beyond the violet end of the visible spectrum, falling on a clean metal surface, releases negatively charged carriers from it. The release is blocked by glass and passed by quartz, which identifies the agent as ultraviolet [234]. Because the carriers are negative, an isolated illuminated conductor loses negative charge if it carries any and otherwise acquires a positive charge, while a positively charged conductor is not discharged at all [238].

Derivation. Only the sign asymmetry needs an argument, and it is a one-line retarding-field argument that will be used quantitatively in Section 74.2 and Section 74.3. Let the liberated carriers have charge $-e$ and leave the surface with kinetic energies up to some maximum $E_{\max}$, and let the isolated conductor sit at potential V relative to its surroundings. A carrier

escaping to infinity must do work eV against the field of the conductor it has just left more negative — or, for $V < 0$, gains that energy. Hence carriers escape whenever

$$E_{\max} > eV\,. \tag{74.1}$$

If $V < 0$ the condition holds for every carrier and the conductor discharges. If $V = 0$ it holds and the conductor charges positive, which raises V until Equation (74.1) fails and the process stops by itself — the plate reaches a steady potential $V = E_{\max}/e$ and no more. If $V > 0$ and already exceeds $E_{\max}/e$ the carriers cannot leave at all, and a positively charged plate keeps its charge. All three of Hallwachs' observations are the three cases of one inequality, and the same inequality read as a measurement is the stopping-potential method: $E_{\max}$ is obtained by finding the V at which emission just ceases.

Nothing here identifies the carriers or requires anything quantum. What Equation (74.1) does establish is that the interesting quantity is $E_{\max}$ and that it is measurable with an electrometer, which is the whole experimental programme of the next two sections. □

74.1.4 Interpretation

In 1887 this is an anomaly and nothing more; there is no theory it contradicts, because there is no theory of it. Its importance is that it makes a quantity measurable. The identification of the carriers came with the corpuscle [276]: their charge-to-mass ratio is that of cathode rays, so the photoelectric current is a current of electrons and the phenomenon is one of the atom's outer structure, treated in Chapter 70 and, for the metallic case, in Chapter 126. It is worth recording that the discovery is incidental in the strict sense: Hertz was verifying Maxwell's equations, the effect degraded his measurements, and he published it because he could not explain it. Chapter 1 takes up what this pattern implies about the difference between a prediction confirmed and an anomaly recorded.

74.1.5 Primary references

[234, 238]. The identification of the carriers with the electron rests on [276].

74.2 Lenard: the energy of the electrons ignores the intensity (1902)

This is the measurement that made the effect a problem. Hertz and Hallwachs establish that light liberates electrons; Lenard establishes *how much energy* they carry, and the answer contradicts the only theory of light then available [308].

74.2.1 Apparatus

An evacuated tube containing the illuminated cathode and a collecting electrode, with a quartz window admitting the light — quartz because glass absorbs precisely the wavelengths that work, as Section 74.1 had shown. The source is a carbon arc, mounted so that it can be moved towards and away from the window. A variable potential difference is applied between cathode and collector so that the emitted electrons may be retarded, and the current is read on a quadrant electrometer, the currents being of the order of picoamperes rather than milliamperes. A magnetic field can be applied across the electron path so that the deflection of the beam gives the charge-to-mass ratio of the carriers independently.

74.2.2 Procedure

Three variables are separated. First the retarding potential is raised until the collected current vanishes; by Equation (74.1) that potential is $E_{\max}/e$ and is the quantity of interest. Then the intensity is changed — by moving the arc towards or away from the window, which changes the irradiance at the cathode by the inverse square of the distance and changes nothing else — and the stopping potential is measured again. Then the spectral composition is changed, by interposing absorbing screens or by changing the source, and the stopping potential measured again. Separately, the time between the first illumination of a shuttered cathode and the first detectable current is looked for.

74.2.3 Observations and data

The photocurrent is proportional to the intensity of the illumination, as any theory would have it. The stopping potential is not: moving the arc changes the current in the ratio the inverse-square law demands and leaves the potential at which the current vanishes unaltered. Changing the spectral composition towards shorter wavelengths raises it. The charge-to-mass ratio of the carriers agrees with that of cathode rays. No delay between illumination and emission is detectable [308].

[Reserved: the tabulation. Stopping potential in volts against source distance — that is, against irradiance in $\mathrm{W/m^2}$ *at the cathode — over the accessible range, showing the constancy; photocurrent in amperes over the same range, showing the proportionality; stopping potential against the spectral band admitted; and the measured charge-to-mass ratio with its uncertainty, against the cathode-ray value. The upper bound on the emission delay, in seconds, is to be given as the resolving time of the electrometer circuit rather than as a measured delay, since no delay was seen.]*

Phenomenon 74.2 (The maximum electron energy is fixed by the colour of the light and not by its intensity)**.** For light of fixed spectral composition, increasing the irradiance at the emitting surface increases the number of photoelectrons per second in proportion and leaves their maximum kinetic energy unchanged; the maximum energy is raised only by shifting the light towards shorter wavelengths [308]. This is item 2 of Phenomenon 69.1 and it is the observation with which no theory treating light as a continuous wave can be reconciled.

Derivation. The two predictions must be set side by side, because the force of the result lies in their disagreement and not in either alone.

Under the hypothesis of Chapter 69 the energy available to one electron is the energy of one quantum, so Equation (69.2) gives $E_{\max} = h\nu - \phi$, in which the irradiance does not appear. Raising the irradiance multiplies the number of quanta crossing the surface per second and therefore the number of electrons ejected per second; it cannot enlarge the parcel any one of them receives. Proportionality of the current to the intensity and constancy of $E_{\max}$ are then the same statement.

Under any classical field the energy delivered to an electron is set by the work done on it by the field, which is quadratic in the field amplitude and therefore proportional to the irradiance. Doubling the irradiance doubles the rate at which energy arrives at every point of the surface, and there is no mechanism in Maxwell's equations by which an electron can decline the extra energy: $E_{\max}$ must rise with the intensity. It does not. The classical picture also predicts that $E_{\max}$ should be insensitive to the frequency, which is the reverse of what is measured. □

Phenomenon 74.3 (Emission begins at once, however feeble the light)**.** The photocurrent appears as soon as the light does, within the resolving time of the measuring circuit, and no accumulation period is observed at any intensity at which the current is detectable at all [308]. This is item 4 of Phenomenon 69.1.

Derivation. The estimate this observation contradicts is worth doing explicitly, because it is the sharpest form of the contradiction and it is arithmetic. Let a classical wave of irradiance I fall on the surface. An electron can absorb only the energy flowing through its own catchment area; take that area to be atomic, $A \approx \pi a^2$ with $a \approx 10^{-10}\,\mathrm{m}$, so $A \approx 3 \times 10^{-20}\,\mathrm{m}^2$. To escape it must accumulate the work function, $\phi \approx 2\,\mathrm{eV} \approx 3.2 \times 10^{-19}\,\mathrm{J}$ for an alkali surface. The waiting time before the first electron can appear is therefore

$$t \approx \frac{\phi}{IA}\,, \tag{74.2}$$

which for a feeble but perfectly ordinary illumination of $I = 10^{-2}\,\mathrm{W/m^2}$ gives $t \approx 10^{3}\,\mathrm{s}$ — of order a quarter of an hour of darkness before the first count. At $I = 1\,\mathrm{W/m^2}$ it is still of order 10 s. Nothing observed resembles either.

Two objections must be closed, or Equation (74.2) is only a straw man. First, one might let the metal collect over a larger area and funnel the energy to one electron; but a classical field exerts its force locally on each electron, and there is no interaction in classical electrodynamics that concentrates the energy of a wavefront onto a single charge faster than it arrives. Second, one might appeal to electrons that were already energetic before the light arrived; the thermal energy available at room temperature is $k_{\mathrm{B}}T \approx 0.025\,\mathrm{eV}$, two orders of magnitude below ϕ, and an appeal to the tail of the distribution predicts a current that depends violently on temperature, which is not observed.

The quantum reading has no waiting time to explain. One quantum is absorbed whole or not at all, so the first electron may leave at the instant the first quantum arrives, and lowering the intensity reduces the *rate* of emission without introducing any delay in the individual act. The full statement of this argument, including the demonstration that no redistribution within the classical field shortens Equation (74.2), is the pending derivation attached to Phenomenon 69.1 in the theory chapter; what is given here is the order-of-magnitude form that the experiment actually confronts. □

74.2.4 Interpretation

Lenard's result is the birth of the problem and, historically, not of its solution: he interpreted the emission as a triggering of electrons already in motion within the atom, the light merely releasing them, and he was still defending that reading long after it had failed. The reading fails on the frequency dependence, which a trigger cannot explain, and it fails on the threshold. But the sceptical structure of Lenard's argument should be preserved rather than smoothed away, because a great deal of what is loosely credited to the photoelectric effect is not established by it: what Phenomenon 74.2 excludes is a *classical field* delivering energy continuously to a target, and that is not the same as establishing that the field is granular. The energy could be granular because the absorber is. That is exactly the loophole Section 69.2.4 names and Section 74.6 closes.

74.2.5 Primary references

[308].

74.3 Millikan: the slope of the line is h/e (1916)

Millikan spent a decade on this measurement in order to disprove the hypothesis it confirms, and the experiment is the better for it: no one has ever had a stronger motive to find the linear law false [437]. What he obtained was the best value of Planck's constant then available by any method, from an experiment in optics and electrometry that touches no cavity and no thermometer.

74.3.1 Apparatus

The difficulty is entirely one of surfaces. An alkali metal exposed to air acquires an oxide layer in minutes, and the layer changes the work function by more than the effect being measured, so the experiment must present a clean surface to the light and keep it clean for the duration of a run. Millikan's answer was to put a machine shop inside the vacuum: a sealed and evacuated glass vessel containing cylinders of sodium, potassium and lithium mounted on a rotatable wheel, together with a cutting knife operated from outside by an electromagnet, so that a fresh surface could be shaved and turned to face the light without ever admitting air. The optical train selects one line at a time from the spectrum of a mercury arc, spanning the visible and the near ultraviolet from about 550 nm to about 250 nm. The emitted electrons are collected in a Faraday cylinder, the current read on a quadrant electrometer, and a known and variable retarding potential is applied between the illuminated surface and the collector.

74.3.2 Procedure

For each metal and each spectral line: shave a fresh surface, admit the line, and record the collected current as the retarding potential is raised, until it vanishes. The potential at which it does is the stopping potential V_0 of Equation (69.1). Repeat for each line; plot V_0 against the frequency ν of the line; and take the slope of the resulting line, which is the measurement. Repeat the whole procedure for a second and a third metal. Separately, determine the contact potential difference between the emitting surface and the collector, which displaces the whole line vertically and must be known before the intercept can be read as a work function — though, as the derivation below shows, not before the slope can be read as h/e.

74.3.3 Observations and data

For every metal the stopping potential is a strictly linear function of frequency over the whole range examined, with no detectable curvature, and the lines obtained for different metals are parallel: the slope is a constant of nature and the material enters only through the intercept. Reading Planck's constant from that slope, Millikan obtained

$$h = 6.57 \times 10^{-34}\,\mathrm{J\,s}\,, \tag{74.3}$$

to about 0.5 % [437].

[Reserved: the tabulation. For each metal, the wavelength and frequency of each mercury line used, the measured stopping potential in volts with its uncertainty, and the residual from the fitted straight line — the residuals being the actual evidence for linearity and the thing a summary slope conceals. Then the fitted slope in V s *with its standard error, the fitted intercept, the separately measured contact potential, and the work function that follows. All in SI.]*

Phenomenon 74.4 (The photoelectric slope determines Planck's constant)**.** The straight line of Equation (69.1) has a slope $\mathrm{d}V_0/\mathrm{d}\nu = h/e$ that is the same for every emitting metal and is independent of the work function, of the contact potential of the apparatus, and of the intensity of the light. Millikan established the linearity and the parallelism across the visible and near ultraviolet on alkali surfaces cut fresh in vacuum, and read from the common slope the value Equation (74.3), to about 0.5 % [437] — at the time the best determination of h by any method, and in agreement with the entirely independent value obtained from the spectrum of cavity radiation [300].

Derivation. The claim to be established is not the linearity, which Phenomenon 69.2 already derives, but the robustness of the slope: why an experiment plagued by surface chemistry and by an unknown contact potential nevertheless yields a clean value of h/e.

Let the illuminated surface have work function ϕ_{E} and the collector ϕ_{C}. The two are in electrical contact through the external circuit, so their Fermi levels coincide and their vacuum levels differ by a fixed contact potential difference $V_{\mathrm{c}} = (\phi_{\mathrm{C}} - \phi_{\mathrm{E}})/e$, whatever the electrometer reads. An electron leaving the surface with kinetic energy E and crossing to the collector against an applied potential difference V therefore arrives only if $E > eV + \phi_{\mathrm{C}} - \phi_{\mathrm{E}}$. The fastest electron just fails when

$$eV_0 = E_{\max} - (\phi_{\mathrm{C}} - \phi_{\mathrm{E}}) = h\nu - \phi_{\mathrm{E}} - \phi_{\mathrm{C}} + \phi_{\mathrm{E}} = h\nu - \phi_{\mathrm{C}}\,. \tag{74.4}$$

Two conclusions follow, and they pull in opposite directions.

The intercept is worthless as a measurement of the illuminated surface. By Equation (74.4) it reports the work function of the *collector*, and any oxide or adsorbed film on that electrode enters the answer; every early attempt to read ϕ off a photoelectric intercept was measuring the wrong electrode.

The slope is untouched. Neither ϕ_{E} nor ϕ_{C} depends on the frequency of the light, so differentiating Equation (74.4),

$$\frac{\mathrm{d}V_0}{\mathrm{d}\nu} = \frac{h}{e}\,, \tag{74.5}$$

in which no property of either metal survives. A contamination that changes the work function displaces the whole line vertically and does not tilt it; that is why the lines for sodium, potassium and lithium come out parallel, and why parallelism is a stronger check on the hypothesis than the value of any one intercept.

What Equation (74.5) measures is the ratio h/e and not h. In the SI in force since 2019 both constants are fixed by definition [2333, 2431], so the slope of this line is now an exact number,

$$\frac{h}{e} = 4.135\,667 \times 10^{-15}\,\mathrm{V\,s}\,, \tag{74.6}$$

and Millikan's experiment has become a test of a definition rather than a determination of a constant. Comparing Equation (74.3) with the defined $h = 6.626\,070\,15 \times 10^{-34}\,\mathrm{J\,s}$, his value is low by about 0.8 %, which is larger than the 0.5 % he quotes. The discrepancy is not in the slope. To convert h/e into h one must multiply by a separately measured elementary charge, and the charge Millikan used was his own oil-drop value [407], later found low because of the value then adopted for the viscosity of air. The lesson is generic and belongs to Chapter 2: an experiment's own precision bounds only the quantity it actually measures, and every constant imported to convert that quantity into another carries its error into the result undiminished. □

Derivation pending #71

The operational definition of the stopping potential. The measured current does not fall to zero at a sharp potential but approaches it along a tail, because the emitted electrons have a distribution of energies rather than a single maximum, because electrons originate at a range of depths below the surface, and because the collector is not equipotential. What is pending is the derivation of the shape of that current–voltage curve near its foot from the energy distribution of the photoelectrons, and the demonstration that the extrapolation Millikan used recovers the true intercept without bias — the point being that the linearity of the plotted line is a statement about a quantity defined by an extrapolation, and that extrapolation must be shown not to have manufactured it.

74.3.4 Interpretation

The equation is confirmed and its author's hypothesis is not, and Millikan said so. Having verified the linear relation to better than a percent, he continued to regard the light-quantum picture behind it as untenable, and he was not merely being stubborn: the photoelectric equation follows equally from a quantized atom absorbing energy from a classical field, as [1402] later showed explicitly and as Section 69.2.4 records. The discreteness that Phenomenon 74.4 demonstrates is a discreteness of *energy exchange*; the atomic energy levels of Chapter 70 supply it without any assumption about the field. The measurement's real and undisputed content is metrological: the constant that Planck extracted from a cavity spectrum [294, 300], by an argument about oscillators in thermal equilibrium, reappears to half a percent in a table-top electrometer experiment on a shaved sodium surface. Two measurements with nothing in common but the constant is the strongest kind of evidence that the constant is real, and it is why Part VIII treats h as a fact about nature before any interpretation of it is settled.

74.3.5 Primary references

[437]. The elementary charge used to convert the measured slope into h is [407]; the independent value of h from cavity radiation is [300]; the semiclassical account of the same equation is [1402].

74.4 Compton: the wavelength of scattered X-rays changes (1923)

Here the evidence changes character. Everything before this section concerns the energy the light delivers; this concerns the momentum it carries away, and the momentum of a wave packet is a property of the light itself that no rearrangement of the absorber can supply [499].

74.4.1 Apparatus

An X-ray tube with a molybdenum target, whose Kα line at about 17.5 keV and 71 pm provides a nearly monochromatic incident beam. A block of graphite serves as scatterer: a light element, so that most of its electrons are loosely enough bound to recoil freely. Slits define the incident beam and select the direction in which the scattered radiation is examined, fixing the scattering angle θ. The scattered radiation enters a Bragg spectrometer — a crystal of known lattice spacing on a divided circle, with an ionization chamber to measure the reflected intensity — so that a wavelength is read as an angle through Bragg's law [400].

74.4.2 Procedure

Set the scattering angle; scan the spectrometer crystal through its range and record the ionization current against crystal angle, converting angle to wavelength by $n\lambda = 2d\sin\vartheta$; identify the peaks in the resulting spectrum. Repeat at a series of scattering angles from small to large, and repeat with scatterers of different atomic number. The comparison to be made is between the wavelength of each peak and the wavelength of the primary beam measured in the same spectrometer, which cancels the calibration of the crystal spacing from the difference.

74.4.3 Observations and data

At every scattering angle other than zero the scattered spectrum contains two peaks, not one: an unmodified peak at the incident wavelength and a second peak displaced towards longer wavelength. The displacement grows with the scattering angle. It does not depend on the incident wavelength, on the material of the scatterer, or on the intensity of the beam [499].

[Reserved: the tabulation. Scattering angle in degrees; measured wavelength of the modified peak in picometres with its uncertainty; measured wavelength of the unmodified peak, as a control that the spectrometer calibration has not drifted; the difference; and the value predicted by Equation (69.4), together with the ratio of measured to predicted shift and its propagated uncertainty. Also the relative height of the two peaks against the atomic number of the scatterer, which is the evidence for the binding interpretation of the unmodified line.]

Phenomenon 74.5 (The shift depends on the scattering angle and on nothing else)**.** The wavelength displacement of the modified peak is a universal function of the scattering angle alone, given by Equation (69.4) with the Compton wavelength $\lambda_{\mathrm{C}} = h/m_{\mathrm{e}}c = 2.426\,31\,\mathrm{pm}$ [2431], and carries no adjustable parameter and no property of the scattering material [499, 501]. Classical scattering of a wave by a charge predicts no displacement at any angle, because a charge driven at one frequency re-radiates at that frequency [1976]; the observation is therefore not a quantitative refinement of the classical result but a contradiction of it.

Derivation. The relation itself is derived in Chapter 69 from the relativistic conservation laws of Chapter 40 applied to a collision between a quantum of energy $h\nu$ and momentum $h\nu/c$ and a free electron at rest. What is derived here is the experimental design that follows from it, because the reason this measurement was not made in 1905 is contained in the formula.

Evaluate Equation (69.4) at representative angles. The shift is $\lambda_{\mathrm{C}}\,(1-\cos\theta)$, so it vanishes in the forward direction, equals λ_{C} at 90°, and saturates at $2\lambda_{\mathrm{C}}$ in backscattering; the values are collected in Table 74.1. The absolute shift is a few picometres and no more, whatever one does.

The quantity an instrument must resolve is not the shift but the *fractional* shift, and that is where the choice of radiation enters:

$$\frac{\Delta\lambda}{\lambda} = \frac{\lambda_{\mathrm{C}}}{\lambda}\,(1-\cos\theta) = \frac{h\nu}{m_{\mathrm{e}}c^2}\,(1-\cos\theta)\ , \tag{74.7}$$

the second form making the physical content plain: the shift is appreciable only when the quantum energy is not negligible against the rest energy of the electron, 511 keV. For the molybdenum Kα line at 71 pm, Equation (74.7) at 90° gives $\Delta\lambda/\lambda \approx 0.034$, a three-percent change in wavelength — large for a crystal spectrometer, which resolves far better than that. For visible light at 500 nm the same formula gives $\Delta\lambda/\lambda \approx 5\times10^{-6}$, below anything a nineteenth-century spectroscopist could have separated from the width of a line. The Compton effect was not overlooked before 1923; it was inaccessible, and it became accessible when X-ray spectroscopy did [400].

The same formula explains the peak that does not move. An electron bound tightly enough that the whole atom recoils requires m_{e} in Equation (69.4) to be replaced by the atomic mass, which for carbon is some four orders of magnitude larger; the predicted shift falls by the same factor and becomes unobservable. Hence a light element scatters mostly into the modified peak and a heavy one mostly into the unmodified peak, which is a further prediction of the same kinematics and is what the scatterer-dependence measurement tests. □

Table 74.1: Wavelength shift predicted by Equation (69.4) at representative scattering angles, with the fractional shift for the molybdenum Kα line at 71 pm from Equation (74.7). Computed from the formula; these are not measurements.

θ	$\Delta\lambda$	$\Delta\lambda/\lambda$ at 71 pm
45°	0.711 pm	0.010
90°	2.426 pm	0.034
135°	4.142 pm	0.058
180°	4.853 pm	0.068

74.4.4 Interpretation

This is the first result in this chapter that a classical field cannot be made to yield. The photoelectric effect concerns what happens *inside* the absorber, and a quantized absorber suffices; the Compton shift concerns what happens to the radiation, and the radiation leaves with less energy and a changed direction in exactly the proportion that a particle of momentum h/λ would carry away. Two features make the evidence hard to evade. The shift is independent of the scattering material, so it cannot be a property of carbon; and it is independent of the incident intensity, so it cannot be a nonlinear response of the medium.

Three honest qualifications belong here. The kinematics were published independently and within weeks by Debye [501], from the same hypothesis and without knowledge of the experiment, and the treatise's citation rule requires that both be named. The kinematics fix the shift but not the rate: the angular distribution requires the relativistic theory of the electron and emerges as the Klein–Nishina cross-section [630], developed from Chapter 93, whose low-energy limit is the Thomson result of Chapter 65. And — the qualification that mattered most at the time — the experiment measures an *average* over very many scattering acts. A theory in which energy and momentum are conserved only statistically predicts the same mean shift, and one such theory was on the table within a year [508]. Closing that gap is the business of Section 74.5.

74.4.5 Primary references

[499]; the independent derivation of the same kinematics is [501]; the crystal-spectrometer technique the measurement rests on is [400].

74.5 Bothe and Geiger: the two products appear together (1925)

Bohr, Kramers and Slater proposed keeping the continuous classical field at the price of strict conservation: energy and momentum would hold only as statistical averages, the atom being coupled to a virtual radiation field that fixes transition probabilities [508]. It is the most serious counter-theory the light quantum ever faced, and its virtue is that it is sharply falsifiable, since it predicts that a scattered quantum and its recoil electron are uncorrelated event by event. This experiment measures that correlation [519].

74.5.1 Apparatus

Two needle counters — point-discharge counters of the kind Geiger had developed — mounted on opposite sides of a chamber filled with hydrogen, which is used as the scatterer because its single loosely bound electron per atom maximizes the modified scattering of Section 74.4 and minimizes absorption. One counter has a thin window and responds to the recoil

electrons themselves; the other is arranged to respond to the scattered X-ray quanta, which it registers through the secondary electrons they liberate in a foil. A collimated X-ray beam crosses the chamber between them. Each counter's discharge drives a string electrometer whose fibre is imaged onto a photographic film driven at constant speed, so that the two records are written side by side against a common time axis.

74.5.2 Procedure

Irradiate the hydrogen and let the film run. Develop it, and measure for every discharge in one trace the time interval to the nearest discharge in the other. Coincidences are then defined by a resolving time τ — set by how finely two deflections can be separated on the moving film, and here of order $1 \times 10^{-4}\,\mathrm{s}$ — and counted. The counting rate of each channel separately is measured on the same records. The number of chance coincidences expected from those singles rates and that resolving time is then computed and subtracted, and the excess is the quantity of interest. The whole experiment consists in showing that the excess is not zero, so the accidental rate must be computed rather than assumed small.

74.5.3 Observations and data

Coincidences between the two counters occur at a rate far above the rate expected from chance alone, with a resolving time of order $1 \times 10^{-4}\,\mathrm{s}$ [519]. Every recorded coincidence is therefore a single scattering act whose two products were both detected, and the scattered quantum and the recoil electron are produced together.

[Reserved: the tabulation. Singles counting rate in each channel in $1/\mathrm{s}$*; total observation time; number of observed coincidences; the computed accidental number from Equation* (74.8) *with its Poisson uncertainty; the excess and its significance. Also the control runs that establish that the excess vanishes when it should — with the X-ray beam off, and with an absorber interposed so that only one channel can respond.]*

Phenomenon 74.6 (Energy and momentum balance in each individual scattering act)**.** The scattered quantum and the recoil electron of a Compton scattering appear simultaneously, in the same act, and not merely at rates that agree on the average. Bothe and Geiger recorded coincidences between a counter responding to scattered X-rays and a counter responding to recoil electrons at a rate far exceeding the chance rate computable from the two singles rates and the resolving time [519]. The complementary measurement, in which a cloud chamber records the direction of the recoil electron and the direction of the scattered quantum in the same event and finds them correlated as the kinematics require, is [520]. This is Phenomenon 69.6, and it excludes the statistical-only conservation of [508].

Derivation. The two competing predictions are predictions about a *number of counts*, so the derivation is the calculation of that number under each hypothesis.

Let the two counters fire at rates N_1 and N_2, and let the apparatus register a coincidence whenever two discharges fall within a resolving time τ of each other. Suppose first that the two channels are statistically independent. In an observation of duration T, counter 1 produces $N_1 T$ discharges; each of them opens a window of total length 2τ, from τ before to τ after, within which a discharge of counter 2 will be recorded as coincident. Provided the windows do not overlap, that is $2\tau N_2 \ll 1$, the expected number of chance coincidences is

$$N_{\mathrm{acc}} = 2\tau N_1 N_2 T\,, \tag{74.8}$$

with Poisson fluctuation $\sqrt{N_{\mathrm{acc}}}$. This is the entire prediction of the statistical proposal [508]: on that theory the recoil electrons and the scattered quanta are emitted in unrelated acts, so there is no source of coincidences other than chance and the observed number must agree with Equation (74.8) within its fluctuation.

Under strict conservation each scattering act produces exactly one quantum and exactly one electron. If such acts occur at a rate R within the region both counters view, and the counters detect their respective products with efficiencies ε_1 and ε_2, the genuine coincidence number is

$$N_{\text{true}} = \varepsilon_1 \varepsilon_2 R T \,, \tag{74.9}$$

and the prediction is $N_{\text{obs}} = N_{\text{true}} + N_{\text{acc}}$, which exceeds Equation (74.8) by an amount that does not fall to zero however the apparatus is arranged.

The design rule follows by dividing the two. The singles rates are themselves proportional to the scattering rate, $N_i = \varepsilon_i R$, so

$$\frac{N_{\text{true}}}{N_{\text{acc}}} = \frac{\varepsilon_1 \varepsilon_2 R}{2\tau \varepsilon_1 \varepsilon_2 R^2} = \frac{1}{2\tau R} \,. \tag{74.10}$$

The genuine rate is linear in the beam intensity and the accidental rate is quadratic, so the discrimination *improves* as the source is weakened and as the resolving time is shortened — the opposite of the usual instinct to increase the signal. This is why the experiment is possible at all with counters of poor efficiency and a resolving time as long as $1 \times 10^{-4}\,\text{s}$, and Equation (74.10) is the reason the coincidence method went on to become the basic technique of particle detection in Part XI. □

Derivation pending #72

The angular correlation measured by the cloud-chamber experiment. What is pending is the direction of the recoil electron as a function of the scattering angle of the quantum, obtained from the same conservation laws that give the wavelength shift, together with the angular spread introduced by the initial motion of the struck electron and by multiple scattering of the recoil track in the chamber gas — the latter being what sets the accuracy with which the predicted correlation can be confirmed, and hence what the measurement is really comparing.

74.5.4 Interpretation

The proposal of Section 69.5.1 was withdrawn at once, and the manner of the withdrawal deserves recording: Bohr accepted the result immediately and without attempting to save the theory, which is the behaviour a falsifiable proposal earns. With it went the last scientifically serious attempt to keep a continuous classical electromagnetic field while accounting for the quantum phenomena, and conservation of energy and momentum was re-established as exact rather than statistical — a conclusion that reaches far beyond optics.

What the experiment establishes should be stated precisely, because it is often overstated. It establishes that the two products of a Compton scattering are correlated in time, event by event, and hence that the process is a collision between two objects and not a redistribution among many. It does not by itself establish that the incident radiation consists of localized quanta before the collision: a field that interacts with matter only in whole quanta, but propagates as a wave, survives this test intact. That last distinction — indivisibility of the propagating field itself — is what Section 74.6 measures.

Methodologically this experiment founded the coincidence technique, for which Bothe was later awarded the Nobel Prize, and Equation (74.10) is the design principle every coincidence measurement in this treatise still uses, from the cosmic-ray telescopes of Section 41.2 to the detectors of Part XI.

74.5.5 Primary references

[519]; the complementary cloud-chamber measurement of the angular correlation is [520]; the theory both of them exclude is [508].

74.6 Grangier, Roger and Aspect: a single quantum is not split (1986)

Everything so far is consistent with a classical field that happens to exchange energy in lumps. This experiment is not, and it is included because it is the one measurement in this chapter that no classical field of any strength or statistics can reproduce [1774]. The quantity it measures is bounded below for every classical field, by an inequality that follows from nothing but the positivity of intensity, and the measured value lies far beneath the bound.

74.6.1 Apparatus

A beam of calcium atoms, excited by two lasers to a state that decays by a two-photon radiative cascade, so that photons are emitted in pairs: one near 551 nm and one near 423 nm, well separated in colour and therefore separable by filters. One member of each pair is sent to a photomultiplier whose pulse opens an electronic gate of fixed duration — it is not measured further, it merely announces that a partner is on its way. The partner is directed onto a beamsplitter, and each of the two output ports carries its own photomultiplier. Coincidence electronics count, within each gate, the detections at the first output port, at the second, and at both. The same source can alternatively feed a Mach–Zehnder interferometer whose fringe visibility is recorded.

74.6.2 Procedure

Run the source at a rate low enough that the probability of two cascades falling in one gate is small, and accumulate three numbers over N gates: N_1 and N_2, the counts at the two output ports, and N_{c}, the number of gates in which both fired. Form the anticorrelation parameter

$$\alpha := \frac{N_{\mathrm{c}} N}{N_1 N_2}\,, \tag{74.11}$$

which is the joint detection probability per gate divided by the product of the single detection probabilities. Then replace the beamsplitter arrangement by the interferometer, without changing the source, and record the fringe visibility. The two configurations put the same light to the two incompatible tests — indivisibility and interference — and the point of the experiment is that both are passed.

74.6.3 Observations and data

The measured anticorrelation parameter is

$$\alpha = 0.18 \pm 0.06\,, \tag{74.12}$$

far below the classical lower bound $\alpha \geq 1$ derived below, and the same source in the interferometer produces interference fringes of high visibility [1774].

[Reserved: the tabulation. Gate duration in seconds; cascade rate in 1/s*; the three counts N_1, N_2 and N_c with their accumulation time; the accidental contribution computed as in Equation (74.8) from the same gate width; the resulting α with its statistical uncertainty; and the fringe visibility with its uncertainty. Also the measured dependence of α on the cascade*

rate, which is the control that identifies the residual α as double-cascade contamination rather than as a failure of the prediction.]

Phenomenon 74.7 (No classical field can give $\alpha < 1$; the measured value is far below it)**.** For light described by any classical field whatever — of any intensity, however feeble, and with any statistics — the anticorrelation parameter Equation (74.11) satisfies $\alpha \geq 1$. Light prepared one quantum at a time violates this: the measured value is Equation (74.12), some thirteen standard deviations below the bound [1774]. The same light, in an interferometer, nevertheless interferes with high visibility. This is Phenomenon 69.7.

Derivation. Take the classical bound first. Let the field arriving at the beamsplitter during a gate have intensity I, treated as a random variable over an ensemble of gates with any distribution whatever, subject only to $I \geq 0$. Let the beamsplitter transmit a fraction T and reflect $R = 1 - T$, and let each detector respond with a probability proportional to the intensity reaching it, with efficiency η. Then the single detection probabilities per gate are

$$P_1 = \eta T \langle I \rangle \,, \qquad P_2 = \eta R \langle I \rangle \,, \tag{74.13}$$

and, because the two detectors respond independently to the *same* field realization, the joint probability is

$$P_{\mathrm{c}} = \eta^2 T R \left\langle I^2 \right\rangle \,. \tag{74.14}$$

Note where the physics sits: the average of the product is the average of I^2, not the product of averages, precisely because a classical wave is present at both outputs at once. Dividing,

$$\alpha = \frac{P_{\mathrm{c}}}{P_1 P_2} = \frac{\langle I^2 \rangle}{\langle I \rangle^2} = 1 + \frac{\mathrm{Var}\,(I)}{\langle I \rangle^2} \geq 1 \,, \tag{74.15}$$

the last step because a variance cannot be negative — equivalently, by the Cauchy–Schwarz inequality applied to I and the constant function. Equality holds only for a field of perfectly steady intensity; any fluctuation makes α larger, never smaller. The efficiencies and the splitting ratio have cancelled, so the bound is not a statement about the apparatus. Neither is it a statement about brightness: making the beam weaker scales I and leaves the ratio in Equation (74.15) untouched. This is the crucial point, and it is what defeats the intuition that a sufficiently dim classical wave would behave like particles.

Now the quantum prediction. A one-photon state incident on port a of the beamsplitter goes over into a superposition of one photon in the transmitted mode and one photon in the reflected mode. That superposition has no component whatever in which both output modes are occupied, since there is only one excitation to distribute, so the joint detection probability vanishes identically and

$$\alpha = 0 \tag{74.16}$$

for an ideal single-photon input — again independently of the splitting ratio and of the detector efficiencies, which appear in P_{c} and in $P_1 P_2$ to the same order and cancel.

The measured Equation (74.12) is neither the classical bound nor the ideal quantum value, and the discrepancy from zero is understood and testable. The heralding is imperfect: if two cascades occur within one gate, the gate contains two photons, which can perfectly well be detected one at each output. The rate of such double-cascade gates grows with the square of the source rate while the single-cascade rate grows linearly, so α should fall towards zero as the source is weakened — which is why the dependence of α on the cascade rate is part of the data and not a diagnostic afterthought.

Finally, the quantity in Equation (74.15) is the normalized second-order correlation function at zero delay, and the inequality $\alpha \geq 1$ is the classical bound on it. The same inequality is violated by resonance fluorescence from a single atom, which was the first observation of the effect [1598], and its systematic theory belongs to Chapter 122. □

74.6.4 Interpretation

Two results are reported together and neither means much without the other. The anticorrelation says that the light is not divided by the beamsplitter: it goes one way or the other, and the classical field picture, in which an amplitude goes both ways and both detectors are illuminated, is excluded by Equation (74.15) at any intensity. The interference says that the very same light, given two paths that are later recombined, produces fringes — which the naive particle picture forbids, since a particle that took one arm cannot know the length of the other. The pair of observations, on one source, is the experimental content of what Chapter 85 discusses, and it should be noted that the discussion rests on data of *this* kind, not on the photoelectric effect.

The historical placement matters as much as the result. Feeble-light interference had been demonstrated as early as 1909, with an exposure of months at an intensity so low that, on a quantum accounting, the apparatus contained one photon at a time [360] — and it established nothing about granularity, because a weak classical wave interferes exactly as well as a strong one. The complementary error is equally old: photon-counting statistics, and even the correlations between the photocurrents of two detectors viewing one source [1083], are reproduced by a classical field of fluctuating intensity, as Equation (74.15) shows for the special case $\alpha \geq 1$. The first experiments to draw the distinction sharply were the tests of the classical field-theoretic predictions for photodetection [1521] and the observation of antibunched fluorescence from a single atom [1598]; what [1774] adds is a heralded single-photon source, so that the two tests — indivisibility and interference — are applied to the same prepared state. The same hardware, and the same calcium cascade, is what makes the Bell tests of Chapter 89 possible.

A closing note on the word. "Photon" was coined by Lewis [551] for an entity quite unlike the one this section measures; what the experiment establishes is the indivisibility of a one-excitation state of the quantized field, and the honest definition of that object is the one given in Chapter 69 and completed in Part XI.

74.6.5 Primary references

[1774]. Precursors that bear directly on the same distinction are [1521, 1598]; the feeble-light interference that does *not* establish granularity is [360]; the intensity correlations that a classical field does reproduce are [1083].

74.7 What the measurements settle

The chapter has a structure that Table 74.2 makes visible and that the standard textbook order conceals. The first three experiments are decisive against *one* classical picture — a continuous field pouring energy into a passive absorber — and are silent about the alternative in which the field stays classical and the absorber is quantized. That alternative is not a philosophical reservation: it reproduces the threshold, the linearity in frequency, the independence of intensity and the promptness of the emission [1402], which is to say every one of the four observations of Phenomenon 69.1. Anyone who says the photoelectric effect proves the existence of photons is asserting something the photoelectric effect does not contain, and Millikan, who measured it best, said so at the time.

What does the work is the second half. Section 74.4 shows that the radiation carries momentum in the fixed ratio $1/c$ to its energy and delivers it in a collision, which is a property of the light and not of the target — the shift is the same for graphite, for hydrogen and for paraffin. Section 74.5 shows that the collision is a single event with two products and not an average over many, which removes the only counter-theory that had survived [508]. And Section 74.6 shows that the propagating field is itself indivisible, by violating a

Table 74.2: The measurements reported in this chapter, and the claim each is entitled to. The third column is the point of the chapter: the experiments most often cited as proving that light is granular do not, and the ones that do are less famous.

Experiment	Quantity measured	What it establishes
Hertz–Hallwachs 1887–1888	sign of the carriers	ultraviolet light liberates electrons from a clean metal; an anomaly, not yet a test of anything
Lenard 1902	stopping potential	the electron energy ignores the intensity and follows the colour; no continuous classical field can deliver energy this way
Millikan 1916	slope $\mathrm{d}V_0/\mathrm{d}\nu$	the slope is h/e for every metal, giving $h = 6.57 \times 10^{-34}\,\mathrm{J\,s}$; quantized *exchange*, not quantized light
Compton 1923	$\Delta\lambda$ against θ	the radiation carries momentum h/λ; classical scattering predicts no shift at all
Bothe–Geiger 1925	coincidence excess	energy and momentum balance in each single act; statistical-only conservation excluded
Grangier *et al.* 1986	α at a beamsplitter	$\alpha = 0.18 \pm 0.06$ against the classical bound $\alpha \geq 1$; the field itself is indivisible

bound Equation (74.15) that follows from the positivity of a classical intensity and from nothing else. Only the third of these speaks about the field between the source and the detector; the first two speak about what it does when it arrives.

Two things this chapter does not settle should be named. It does not settle what a photon *is* — the answer is an excitation of the quantized electromagnetic field, with two helicity states and no rest frame, and it belongs to Part XI. And it does not settle the rate of any of these processes, only their kinematics: the angular distribution of Compton scattering requires Chapter 93 and appears as the Klein–Nishina cross-section [630], and the photoelectric cross-section requires the atomic structure of Chapter 70. What is established here, beyond argument and by measurements a century apart, is that light delivers energy in units $h\nu$, carries momentum $h\nu/c$, conserves both in every individual act, and cannot be divided at a beamsplitter.

Chapter 75

Experiment: Matter Waves Observed

Chapter 71 states a hypothesis that was made on grounds of symmetry and not of data: a carrier of momentum p is accompanied by a wave of wavelength h/p. That chapter draws out what follows — Bohr's quantization condition ceases to be a postulate, the optical–mechanical analogy of geometrical optics and classical mechanics acquires its missing entry, and a localized particle becomes a packet with a group velocity. It names the experiments in passing. It does not show them being done, and the omission is worth repairing at length, because the hypothesis was so nearly baseless when it was made: de Broglie had one relation transposed from light to matter and no measurement whatever to support it.

Six measurements are reported, in the order in which each removes an alternative explanation rather than in strict order of date. Davisson and Germer scatter slow electrons from a single crystal and find the wavelength written into the diffraction geometry [571]; Thomson and Reid send fast electrons through a thin film and photograph rings, so the effect is not peculiar to a surface [589]; Estermann and Stern diffract helium atoms and hydrogen molecules, so it is not peculiar to charge [651]; Jönsson opens two slits and Merli, Missiroli and Pozzi and later Tonomura and collaborators build the pattern one electron at a time, so it is not a collective effect [1212, 1588, 1826]; Colella, Overhauser and Werner tilt a neutron interferometer and read gravity off the fringe count, so the phase of a matter wave is an instrument reading [1542]; and molecular interferometry carries the relation to objects of thousands of atoms and shows what does end it [1966, 2337]. The prediction was published before any of them [500, 521], and the suggestion that a crystal would serve as the grating was published two years before the first two [523].

The arithmetic below uses the CODATA values of h, e, m_e and the neutron mass [2431].

Derivation pending #73

Experiment: Matter Waves Observed: what is pending is the apparatus and procedure of each measurement and the tabulation of its data — scattered intensity against angle and accelerating voltage, ring radii against voltage, diffraction angle against source temperature, fringe visibility against count, neutron intensity against tilt angle, and molecular fringe contrast against internal temperature and residual gas — each in SI with its uncertainty budget. The phenomena stated below each carry their derivation inline.

75.1 Davisson and Germer: electron diffraction by a crystal (1927)

The experiment was an accident twice over. Davisson and Kunsman had been measuring the angular distribution of electrons scattered from a nickel target and finding it smooth; a vacuum failure oxidized the target, and the prolonged heating used to reduce it recrystallized the nickel into a few large crystals. When the measurements resumed, the smooth distribution had been replaced by sharp maxima at particular angles and particular beam energies. What makes the result a measurement rather than a curiosity is that both quantities entering the comparison — the wavelength predicted from the accelerating voltage and the wavelength read off the diffraction angle — are fixed independently, one by the electrical circuit and one by X-ray crystallography, and neither is adjusted.

75.1.1 Apparatus

[Reserved: the sealed vacuum vessel; the electron gun delivering a collimated beam of a few tens of electronvolts with the accelerating potential adjustable and measurable; the nickel target, a single crystal presenting its (111) face, mounted so that it can be rotated about the beam axis and about the axis perpendicular to it; the movable Faraday-box collector, with a retarding potential on its aperture so that only elastically scattered electrons are counted, and the galvanometer that reads its current. To be given: the beam current and its stability, the collector acceptance angle — which sets the angular resolution and therefore, by Equation (75.3), the accuracy of the whole measurement — and the working pressure, since the recrystallization that made the experiment possible was itself a vacuum accident.]

75.1.2 Procedure

[Reserved: two scans, and the logic of taking both. At fixed accelerating voltage the collector is swept in angle and the elastic current recorded, giving the angular distribution; at fixed collector angle the voltage is swept, giving the excitation curve. A maximum must appear in both scans at the place where the grating condition and Equation (75.1) agree, and the joint appearance is the test: a feature of the target alone would not move with voltage, and a feature of the beam alone would not move with the crystal azimuth. The azimuthal scan, in which the crystal is rotated about the beam and the maxima reappear with the three-fold symmetry of the (111) face, is what identifies the scatterer as the lattice.]

75.1.3 Observations and data

[Reserved: the elastic scattered current against collector angle at each accelerating voltage, and against voltage at each collector angle, with the resulting map of maxima in the angle–voltage plane. To be tabulated when written: for each maximum, the accelerating voltage and its uncertainty, the collector angle and its uncertainty, the order assigned, the wavelength deduced from the geometry, the wavelength predicted by Equation (75.1), and their ratio — the whole set, not the single case worked below, since the strength of the result lies in the maxima moving together as the voltage is changed.]

Phenomenon 75.1 (The wavelength of an electron beam can be read off a crystal)**.** A beam of electrons of well-defined kinetic energy scattered from a single crystal produces sharp maxima at particular angles, as X-rays of the same wavelength do on the same crystal, and the wavelength deduced from the diffraction geometry agrees with h/p computed from the accelerating voltage. Davisson and Germer, scattering electrons of 54 eV from the (111) face of a nickel single crystal, found a pronounced maximum near 50°, corresponding to a wavelength of 0.165 nm against the de Broglie value 0.167 nm [571]. The lattice spacing

entering the first number was measured with X-rays [400] and the voltage entering the second with a voltmeter; no quantity was fitted.

Derivation. Two numbers are compared. The first is the wavelength the hypothesis assigns to the beam. An electron accelerated from rest through a potential difference V has kinetic energy $T = eV$ and momentum given by $(pc)^2 = T^2 + 2Tm_ec^2$, so

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2m_eeV}} \left(1 + \frac{eV}{2m_ec^2}\right)^{-1/2} . \tag{75.1}$$

The bracket is the relativistic correction. At $V = 54\,\mathrm{V}$ it departs from unity by 2.6×10^{-5} and may be dropped, leaving $\lambda = 0.167\,\mathrm{nm}$; it may not be dropped in Section 75.2, where the same electrons are accelerated a thousand times harder.

The second number is the wavelength the diffraction geometry assigns to the beam. A row of scatterers of spacing d in the surface reinforces at angles satisfying $d\sin\theta = n\lambda$, and the row spacing of nickel, $d = 0.215\,\mathrm{nm}$, is known from X-ray crystallography [400] — that is, from a measurement in which no electron appears. The observed first-order maximum at $\theta = 50^\circ$ therefore reads

$$\lambda = d\sin\theta = 0.215\,\mathrm{nm} \times 0.766 = 0.165\,\mathrm{nm}\,, \tag{75.2}$$

which agrees with Equation (75.1) to 1.2 %.

It is worth differentiating the comparison, because that says which measurement limits it. From Equation (75.1), $\delta\lambda/\lambda = -\tfrac{1}{2}\,\delta V/V$, so a beam energy known to 1 % fixes the predicted wavelength to 0.5 %. From Equation (75.2),

$$\frac{\delta\lambda}{\lambda} = \cot\theta\,\delta\theta + \frac{\delta d}{d}\,, \tag{75.3}$$

and $\cot 50^\circ = 0.84$, so an error of one degree in locating the maximum is worth 1.5 % of the wavelength. The angular measurement, not the beam energy, sets the accuracy of the test, and the residual 1.2 % is smaller than the spread that a degree of angular uncertainty alone permits. It is also of the size of a known systematic effect that any comparison of this kind must confront: the crystal presents a mean inner potential to the electron, which is accelerated on entering the solid and therefore has a shorter wavelength inside it than outside. The reduction of that effect is the business of the interpretation, not of the arithmetic here; what the arithmetic establishes is that a wavelength computed from a voltage and a wavelength read off a lattice agree to about one percent, which no theory of the electron as a corpuscle predicts at all. □

75.1.4 Interpretation

[Reserved: what the maxima are and what they are not. The azimuthal symmetry identifies the lattice as the scatterer; the joint motion of the maxima in angle and voltage identifies the wavelength as the quantity being diffracted. The section must then treat the two systematic effects honestly — refraction by the mean inner potential of the crystal, which shifts the deduced wavelength, and the distinction between the two-dimensional surface-grating reading and the three-dimensional Bragg reading of the same maxima, which the experiment does not by itself settle. The forward link is that Elsasser's prediction [523] preceded the observation by two years and was unknown to one of the two groups that confirmed it, so the priority customarily assigned to the experimenters alone is incomplete.]

75.1.5 Primary references

[571]. The prediction is [523]; the independently measured lattice spacing is X-ray crystallography [400]; the hypothesis under test is [500, 521].

75.2 Thomson and Reid: transmission rings from thin films (1927)

A reflection experiment on a single crystal leaves one evasion open: the maxima might be produced by something happening in the outermost atomic layer of that particular face of that particular metal. The transmission geometry closes it. Fast electrons sent through a thin polycrystalline film meet crystallites at every orientation, emerge on the far side, and land on a plate as a set of concentric rings; the target has no preferred direction, so there is no surface for the explanation to hide in, and the pattern is the electron counterpart of the Debye–Scherrer powder rings that X-rays make in the same apparatus.

75.2.1 Apparatus

[Reserved: the discharge tube and accelerating column delivering electrons at some tens of kilovolts, with the potential measured independently of the beam; the film target — a self-supporting celluloid film in the first communication, metal films evaporated or beaten thin in the later work — of thickness small enough to transmit and large enough to diffract; the drift space of known length between target and plate; the photographic plate; and the magnetic shielding, since a stray field bends the beam and the whole measurement is of a radius. To be given: film thickness and its uniformity, the drift length and its uncertainty, and the plate calibration used to convert blackening into a ring radius.]

75.2.2 Procedure

[Reserved: record the ring pattern at each of a series of accelerating voltages spanning the available range, then repeat with a target of a different material. Two comparisons are then available and neither needs an absolute calibration: at fixed voltage, the ratios of the ring radii within one pattern must equal the ratios of the inverse lattice spacings of the target, which are known from X-ray work; and at fixed target, every radius must scale with voltage as Equation (75.4) requires. Rotating or replacing the film must change nothing, which is the control that distinguishes a lattice effect from a surface or instrumental one.]

75.2.3 Observations and data

[Reserved: ring radii against accelerating voltage for each target material, with the plate scale and the drift length that convert them to angles. To be tabulated when written: voltage, ring order, measured radius, deduced scattering angle, deduced lattice spacing, and the spacing known from X-ray diffraction on the same material; and separately the test of the scaling law, as the measured radius times the square root of the voltage, which must be constant along a column to the precision of the plate measurement.]

Phenomenon 75.2 (Ring radii scale as the inverse square root of the accelerating voltage)**.** Electrons of some tens of kilovolts transmitted through a thin polycrystalline film produce concentric rings on a plate beyond it. The radii are set by the lattice spacings of the target material and scale with the accelerating voltage as $V^{-1/2}$, independently of the orientation of the film; the pattern is that produced by X-rays of the same wavelength on the same material [589]. Because a polycrystalline film has no preferred orientation and the beam passes through it, no account resting on a surface effect peculiar to one crystal face survives.

Derivation. For each set of lattice planes of spacing d, the crystallites oriented so as to satisfy $2d\sin\theta = n\lambda$ lie on a cone about the beam, and the plate at a distance L downstream records a ring of radius $r = L\tan 2\theta$. The angles are small, because the wavelength is far below the spacing: Equation (75.1) at $V = 40\,\mathrm{kV}$ gives $\lambda = 6.0\,\mathrm{pm}$, so for $d = 0.2\,\mathrm{nm}$ the

first order sits at $\theta = 0.86^\circ$ and the ring at $2\theta = 1.7^\circ$, which on a drift of 0.3 m is a radius of 9 mm. To that accuracy $\tan 2\theta \simeq 2\theta \simeq \lambda/d$, so

$$r \simeq \frac{L\lambda}{d} \propto \frac{1}{\sqrt{V}}\,, \tag{75.4}$$

the proportionality following from Equation (75.1) with the relativistic bracket set to unity.

Equation (75.4) is a parameter-free scaling law, and that is what makes it a test. The drift length L and the plate scale enter every radius alike and cancel from any ratio, so the prediction that $r\sqrt{V}$ is constant can be checked without calibrating the instrument at all. The lattice spacing d is likewise known from X-rays, so the ratios of radii within a single pattern are predicted too, and they are predicted differently for celluloid than for a metal.

One correction has become visible in moving from Section 75.1 to this geometry, and it is worth recording because it is a second, independent piece of physics appearing in the same data. The relativistic bracket of Equation (75.1), negligible at 54 V, shortens the wavelength by 0.5 % at 10 kV and by 2.8 % at 60 kV. An exact $V^{-1/2}$ law therefore cannot hold across the range: the rings at the top of it must come in about two percent tighter than the non-relativistic scaling predicts. At this precision the departure is at the edge of what plate measurements resolve, and the honest statement is that these experiments establish the scaling and are consistent with the correction rather than measuring it. □

75.2.4 Interpretation

[Reserved: the loophole closed — transmission through a target with no preferred orientation removes any surface explanation of Section 75.1 — and the loophole left open, which is that the electron might be diffracted as a corpuscle by a periodic force field rather than as a wave. That one is closed not here but by Section 75.4, where the interference is formed by two apertures in free space with no lattice at all. The historical remark belongs in a remark and not in the body: J. J. Thomson showed that the electron is a particle and G. P. Thomson that it is a wave, and both were right.]

75.2.5 Primary references

[589]. That first communication reports the effect for a celluloid film; the metal films that made the result independent of the chemistry of the target came in the fuller papers of the following year [621].

75.3 Estermann and Stern: helium and hydrogen diffracted (1930)

Everything so far has been done with electrons, and an electron is a charged point that a lattice of ions can grip in several ways. The decisive extension is to a projectile that offers nothing to grip. A ground-state helium atom is electrically neutral, chemically inert, and carries neither an electric nor a magnetic dipole moment; if such a thing is diffracted by a crystal surface, the wave behaviour belongs to matter as such and not to charge. The instrument is Stern's molecular-beam apparatus, the same lineage as Chapter 80, and the grating is the cleavage face of a lithium fluoride crystal.

75.3.1 Apparatus

[Reserved: the beam source — an oven or gas reservoir with a fine aperture, effusing into high vacuum, whose temperature is the control variable of the whole experiment and must therefore be measured, not merely set; the collimating slits defining the beam; the lithium

fluoride crystal, cleaved to present a fresh (100) face, mounted on a rotatable stage; and the detector, a hot-wire or manometric gauge sensitive to a beam flux of neutral atoms, movable in angle. To be given: the source temperatures used, the beam divergence — which sets the angular resolution and hence what fraction of the predicted separation between maxima can be resolved — and the provision for cooling the source, since the temperature scan is the test.]

75.3.2 Procedure

[Reserved: at fixed source temperature, sweep the detector in angle and record the beam flux, locating the specular peak and the diffracted maxima either side of it. Then change the species, from helium to molecular hydrogen, and repeat; then change the source temperature and repeat again. The measurement is of the positions of the maxima and not of their heights, and the quantities compared are ratios — helium against hydrogen at one temperature, and one temperature against another for one species — because the ratios are free of the beam-averaging factor that the absolute positions are not.]

75.3.3 Observations and data

[Reserved: beam flux against detection angle for each species and each source temperature, with the specular peak and the first-order maxima marked. To be tabulated when written: species, source temperature in kelvin, angle of the observed maximum with its uncertainty, the wavelength deduced from the grating condition, and the wavelength predicted by Equation (75.5); and separately the two ratios that the experiment really measures, helium to hydrogen at fixed temperature and cold source to warm source at fixed species.]

Phenomenon 75.3 (Neutral atoms and molecules are diffracted, with the wavelength their temperature assigns)**.** Beams of helium atoms and of hydrogen molecules reflected from the cleavage face of a lithium fluoride crystal show diffraction maxima at the angles the grating condition assigns to the wavelength h/p, and those maxima move as predicted when the wavelength is changed by changing the temperature, and hence the mean momentum, of the source [651]. Helium is electrically neutral, chemically inert and carries no permanent electric or magnetic moment, so no account of the diffraction of Sections 75.1 and 75.2 in terms of charge, of the Coulomb field, or of anything peculiar to the electron survives this measurement.

Derivation. Write $\lambda = h/mv$ and give the particle the mean thermal kinetic energy $\frac{3}{2}k_{\mathrm{B}}T$ of its source. Then

$$\lambda_T = \frac{h}{\sqrt{3mk_{\mathrm{B}}T}}\,, \tag{75.5}$$

which for helium, of mass 4.003 atomic mass units, at $T = 295\,\mathrm{K}$ is 0.073 nm, and for molecular hydrogen, of mass 2.016, at the same temperature 0.104 nm. Both are of the order of the row spacing of a lithium fluoride cleavage face, which is the reason the crystal serves as a grating for them at all — exactly the reason it served for the 54 eV electrons of Section 75.1.

Equation (75.5) makes two predictions that no adjustable quantity can absorb, and both are ratios. The first is the mass scaling at fixed temperature: the mass ratio of helium to molecular hydrogen is 1.99, so their wavelengths must stand in the ratio $\sqrt{1.99} = 1.41$, and the diffraction angles of the two beams from the same face must differ by that factor and by nothing else. The second is the temperature scaling at fixed species: $\lambda_T \propto T^{-1/2}$, so cooling the source from 295 K to about 100 K must lengthen the wavelength by 1.7 and carry every maximum out to the angle the grating condition assigns to the longer wave.

Neither ratio depends on the calibration of anything, and that is fortunate, because the absolute positions do depend on a factor this derivation cannot supply. An effusive source does not emit a monochromatic beam: it emits a flux-weighted Maxwell distribution, so the beam carries a spread of wavelengths whose most probable value differs from Equation (75.5) by a factor of order unity fixed by the weighting and by the aperture geometry. That factor is the same for helium as for hydrogen and the same at 295 K as at 100 K, so it cancels from both ratios above. The spread itself does not cancel; it broadens each maximum, and the broadening is predicted too, since each velocity class contributes at its own angle.

The physical conclusion needs no arithmetic. Helium presents a lattice with nothing to couple to but its momentum, and it is diffracted. Whatever produced the maxima of Section 75.1 is therefore not a property of the charge of the electron. □

75.3.4 Interpretation

[Reserved: the generalization achieved. The de Broglie relation is confirmed for a composite, neutral, chemically inert object, which is a far stronger statement than its confirmation for an elementary charged one, and it is confirmed through a temperature scan rather than at a single point. The section should also be candid about what the surface adds: helium scatters from the outermost layer alone, so this is a two-dimensional grating and the maxima carry information about the surface corrugation as well as about the wavelength — which is why the technique survives today as a probe of surfaces rather than as a test of the relation.]

75.3.5 Primary references

[651]. The molecular-beam method is Stern's, developed for the velocity measurement of [471] and for the experiment of Chapter 80.

75.4 The electron double slit, one electron at a time (1961–1989)

Two explanations still stand after the diffraction experiments. The first is that the pattern is somehow a property of the lattice rather than of the projectile; the second is that it is a collective effect of many electrons in flight together. Both are removed here, and by two distinct measurements taken twenty-eight years apart. Jönsson cut slits in a foil and formed interference with no crystal anywhere in the apparatus [1212]; Merli, Missiroli and Pozzi and, later and more famously, Tonomura and collaborators reduced the source until the apparatus was almost always empty, and watched the fringes assemble from individual point-like arrivals [1588, 1826]. This is Young's experiment of Chapter 67 performed with matter, and then performed one particle at a time.

75.4.1 Apparatus

[Reserved: two instruments, described separately. Jönsson's is an electron-optical bench — a 50 kV *gun, a copper foil carrying slits of about* 0.5 µm *width made by electron-beam lithography and electroplating, in sets of two, three, four and five, and a magnifying electron lens system between the slits and the recording plate, without which the fringes are too fine to record. Tonomura's is a field-emission electron microscope with an electron biprism in place of the slits — a fine charged filament between two earthed plates, which superposes two virtual sources — and a position-sensitive detector reading out single arrivals. To be given for each: the accelerating voltage, the effective separation of the two paths, the drift length, and for the second the emission rate, which is the quantity the whole argument turns on.]*

75.4.2 Procedure

[Reserved: for the multi-slit measurement, record the pattern for two, three, four and five slits at fixed voltage, and verify that the principal maxima do not move while the subsidiary structure between them changes with the slit count, as it does for light. For the single-particle measurement, reduce the beam current until the mean number of electrons in the instrument is far below one, record the individual arrivals with their times, and store successive frames of the accumulating histogram. The essential control is the occupancy estimate of Equation (75.6), which must be made from the measured emission rate and not assumed.]

75.4.3 Observations and data

[Reserved: the recorded interference patterns for each slit count, with fringe spacing against accelerating voltage; and the accumulation sequence, as frames at increasing total counts, with the fringe contrast extracted from each frame. To be tabulated when written: total counts per frame, measured contrast, and the contrast expected from Poisson noise alone at that count, so that the emergence of the pattern can be compared with the statistical prediction rather than merely displayed.]

Phenomenon 75.4 (Each particle interferes with itself)**.** Electrons passing two or more apertures in free space form an interference pattern with no crystal present [1212]. When the source is weakened until the mean number of electrons in the apparatus is far below one, the detector still registers individual, point-like arrivals at positions that show no pattern one by one; the fringes appear only in the accumulated histogram of many thousands of arrivals, and they are the same fringes, in the same places and with the same spacing, as those formed at high intensity [1588, 1826]. The pattern is therefore not an effect of electrons acting on one another.

Derivation. Three estimates, of which the first explains why the experiment took thirty-four years after Section 75.1.

The fringe spacing of a two-aperture pattern of separation a observed at a distance L is $\Delta x = \lambda L/a$. At $V = 50\,\mathrm{kV}$, Equation (75.1) gives $\lambda = 5.4\,\mathrm{pm}$, five orders of magnitude below optical wavelengths and smaller than any atom. Slits ruled a millimetre apart, generous for light, would put the fringes 1.6 nm apart and nothing could record them. Only when the separation is pushed down to a few micrometres — which is what lithography in a foil achieves and what ruling does not — does the spacing reach about 1 µm at a drift of a few tenths of a metre, and even then it must be magnified electron-optically by some four orders of magnitude before a photographic plate can resolve it. The difficulty of the experiment is entirely in these numbers.

The second estimate is the occupancy, and it is what turns the single-particle version into an argument. A 50 keV electron has $\gamma = 1.098$ and hence $v = 0.41c$, so it crosses an instrument of length $\ell \approx 1.5\,\mathrm{m}$ in

$$t_{\mathrm{transit}} = \frac{\ell}{v} \approx 1.2 \times 10^{-8}\,\mathrm{s}\,, \tag{75.6}$$

while a source emitting at a rate R of order 10^3 electrons per second sends them in at mean intervals of $10^{-3}\,\mathrm{s}$. The mean occupancy is $Rt_{\mathrm{transit}} \approx 1.2 \times 10^{-5}$: for each electron in flight, some hundred thousand transit times pass with the apparatus empty. Two electrons are present together with a probability of order the square of that, so an explanation of the fringes as an interaction between electrons fails by ten orders of magnitude and not by one.

The third estimate is what the accumulation itself establishes, and it is less than it is often said to be. If the arrivals are independent draws from a fixed density $p(x)$ — which is

what "one at a time" means operationally — then the count in a detector bin is Poisson, with standard deviation $\sqrt{\bar{N}}$ about its mean $\bar{N}$, so the relative noise on the histogram falls as $\bar{N}^{-1/2}$ and a pattern of fringe contrast C emerges from the noise once $\bar{N} \gtrsim C^{-2}$. For fringes of high contrast a few tens of counts per bin suffice, which is why the published frame sequences show nothing at a hundred arrivals and a clean pattern at a few tens of thousands. The derivation stops there, and the stopping point must be stated plainly: that the density $p(x)$ is the squared modulus of a single-particle amplitude is Born's postulate [533], an addition to the de Broglie relation and not a consequence of it. What these experiments establish is that whatever the fringes are, they are a property of each particle separately; that they are probabilities is an interpretation, formalized in Chapter 77 and left unexplained by Chapter 85. □

75.4.4 Interpretation

[Reserved: the two loopholes closed and the one that opens. The multi-slit series closes the lattice explanation — there is no lattice — and the progression from two to five slits closes any explanation appealing to a property of a particular pair of apertures, since the principal maxima stay put while the subsidiary structure changes exactly as scalar wave optics requires. The occupancy estimate closes the collective explanation. What opens is the measurement problem: a wave passes both apertures and a point is recorded at one place, which is the tension Chapter 77 formalizes. Related neutron measurements with a single slit and a double slit extend the same demonstration to a massive neutral particle [1814].]

75.4.5 Primary references

[1212, 1588, 1826].

75.5 Colella, Overhauser and Werner: gravity shifts a phase (1975)

Everything above measures a wavelength. This experiment measures a *phase*, which is a stronger thing to have: once the phase along each of two separated paths is an instrument reading, anything that acts differently on the two paths becomes measurable. The instrument is the perfect-crystal neutron interferometer of Rauch, Treimer and Bonse [1532], cut from a single silicon ingot so that three parallel lamellae split, redirect and recombine a thermal-neutron beam by Bragg reflection, with the two paths separated by centimetres and the lattice coherent across the whole device. What Colella, Overhauser and Werner then do with it is to rotate it about the incident beam, so that the two paths run at different heights in the Earth's field, and count fringes.

75.5.1 Apparatus

[Reserved: the reactor neutron source and monochromator setting the incident wavelength; the interferometer itself, a monolithic silicon crystal with three lamellae machined from one ingot, its lattice perfect over the whole device; the goniometer that rotates it about the incident beam axis through the angle α; the two neutron counters at the exit beams; and the thermal and vibrational isolation, since a temperature gradient across the crystal or a sub-nanometre relative displacement of the lamellae destroys the interference. To be given: the incident wavelength and its spread, the enclosed area of the two paths, and the mechanical stiffness of the device, which enters the systematic corrections directly.]

75.5.2 Procedure

[Reserved: rotate the interferometer through α in steps from one side of the horizontal to the other, counting neutrons in each exit beam for a fixed time at each step. The counted intensity oscillates in $\sin\alpha$, and the quantity extracted is the number of oscillations between the extreme settings, compared with the prediction of Equation (75.10), which contains no free parameter once the wavelength and the enclosed area are measured. Two systematic corrections must be applied and must be described rather than merely named: the elastic bending of the crystal under its own weight as it is rotated, which changes the path geometry, and the Sagnac phase from the Earth's rotation.]

75.5.3 Observations and data

[Reserved: neutron counts in each exit beam against tilt angle, with counting statistics; the extracted fringe count over the full rotation; and the predicted count from the measured wavelength and area. To be tabulated when written: tilt angle, counts and uncertainty in each beam, and the residual phase after the bending and Sagnac corrections, which is the number that carries the result.]

Phenomenon 75.5 (Gravity shifts the phase of a matter wave)**.** Rotating a neutron interferometer about the direction of the incident beam tilts the plane containing its two paths out of the horizontal, so that the paths run at different heights in the Earth's gravitational field. The counted intensity in each exit beam then oscillates with the tilt angle α, with a phase difference between the paths proportional to $\sin\alpha$, to the area they enclose, to the neutron wavelength, and to the square of the neutron mass. Colella, Overhauser and Werner observed the predicted number of fringes as the interferometer was rotated [1542]. This is a laboratory quantity in which Planck's constant and the acceleration of free fall appear together.

Derivation. Work in the regime where the wave has a local wavenumber and the phase along a path is $\int k\,\mathrm{d}l$. A neutron of total energy E at height z in a uniform field has kinetic energy $E - m_{\mathrm{g}}gz$, so with m_{i} the inertial and m_{g} the gravitational mass,

$$k(z) = \frac{\sqrt{2m_{\mathrm{i}}\,(E - m_{\mathrm{g}}gz)}}{\hbar}\,. \tag{75.7}$$

The expansion parameter is minute. A thermal neutron of wavelength 0.1 nm has $E \approx 0.08\,\mathrm{eV}$, while $m_{\mathrm{g}}gz$ over a height of 1 cm is about $1\times10^{-9}\,\mathrm{eV}$: a ratio of 10^{-8}. To first order, using $E = \hbar^2k_0^2/2m_{\mathrm{i}}$,

$$k(z) \simeq k_0\left(1 - \frac{m_{\mathrm{g}}gz}{2E}\right) = k_0 - \frac{m_{\mathrm{i}}m_{\mathrm{g}}gz}{\hbar^2k_0}\,. \tag{75.8}$$

The neutron is slower at the top, and the phase it accumulates per unit length there is correspondingly less.

The two paths of the interferometer form a parallelogram. One path takes its horizontal leg at the lower height and rises afterwards; the other rises first and takes its horizontal leg at the greater height. The two rising legs are congruent and contribute equally, so they cancel from the difference, and what remains is one leg of length L traversed at height 0 on one path and at height H on the other:

$$\Delta\varphi = [k(H) - k(0)]\,L = -\frac{m_{\mathrm{i}}m_{\mathrm{g}}g}{\hbar^2k_0}HL = -\frac{m_{\mathrm{i}}m_{\mathrm{g}}gA}{\hbar^2k_0}\,, \tag{75.9}$$

with $A = HL$ the area enclosed by the paths. Rotating the instrument by α about the incident beam tips that area out of the vertical, so that the rise becomes $H\sin\alpha$ while the

two rising legs remain congruent with each other; writing $k_0 = 2\pi/\lambda$ and $m_{\mathrm{i}} = m_{\mathrm{g}} = m$ then gives

$$|\Delta\varphi| = \frac{2\pi m^2 g A \lambda}{h^2}\,|\sin\alpha| \ . \tag{75.10}$$

Three things should be read off Equation (75.10).

It is large. With an enclosed area of order $10^{-3}\,\mathrm{m}^2$, a wavelength of 0.14 nm and the neutron mass, the coefficient is about 55 radians, so the intensity passes through some nine fringes as the instrument goes from horizontal to vertical and through twice that over the full rotation. Gravity is the feeblest interaction available in a laboratory and the phase it produces is enormous, because $\hbar^2$ sits in the denominator. The smallness that makes the effect delicate is not in the phase but in the fringe spacing in α, which is what demands the mechanical stability.

It measures a product of two masses. The wavelength carries the inertial mass through $\lambda = h/m_{\mathrm{i}}v$ and the potential carries the gravitational mass, so the observable is $m_{\mathrm{i}}m_{\mathrm{g}}$ and not m. This is a quantum interference experiment sensitive to the ratio that the torsion-balance experiments of Chapter 42 bound classically, and it is the reason the result is quoted as a test of the equivalence principle for a quantum object rather than merely as a demonstration.

It is not a test of general relativity. Equation (75.7) treats gravity as a Newtonian potential energy in a non-relativistic wave equation, and at this accuracy that is all the experiment is sensitive to. What it establishes is that a matter wave responds to a gravitational potential exactly as a wave with that potential in its dispersion relation must, which is a statement about quantum mechanics in a weak static field and about nothing beyond it. □

75.5.4 Interpretation

[Reserved: what the fringe count establishes and the two systematics that limit it. The result confirms Equation (75.10) in form and in magnitude; the residual discrepancies reported by this and later runs, at the level of a percent or so of the phase, are attributed to crystal bending and are themselves a measurement of the elastic response of the device, which is worth stating because it is a case where the systematic became the object of study. The section should also record the companion result obtained on the same class of instrument, that the interference contrast returns only after a rotation of 720° *of the neutron spin and not after* 360° *[1558, 1566] — the double cover of the rotation group of Chapters 14 and 79 made into a fringe count. The 1974 proposal that preceded the measurement is [1530].]*

75.5.5 Primary references

[1542]. The interferometer is [1532]; the spinor-periodicity measurements on the same class of instrument are [1558, 1566].

75.6 Molecule interferometry and the edge of coherence (1999–2019)

The last question the relation invites is where it stops. If matter of every kind has a wavelength h/p, then a chair has one; the reason a chair does not interfere must be found somewhere, and the honest way to look for it is to increase the mass of the interfering object until something fails. Two decades of molecular interferometry have done exactly that, from a molecule of sixty atoms [1966] to tailored molecules of more than two thousand [2337]. Nothing has failed in the relation. What fails, reproducibly and on demand, is the isolation

of the molecule from its surroundings — and that failure is measurable, which makes this section the experimental entrance to Chapter 87.

75.6.1 Apparatus

[Reserved: two generations. The far-field arrangement — an effusive oven sublimating fullerene at high temperature, a velocity selector, a nanofabricated free-standing grating of about 100 nm *period, and a scanning ionization detector. The near-field arrangement — a three-grating Talbot–Lau interferometer, in which the first grating prepares the transverse coherence the source lacks, the second diffracts, and the third serves as a scanning mask, with a source of large molecules that must be volatilized without being destroyed. To be given for each: grating period and open fraction, grating separations, the velocity distribution admitted, and the vacuum and thermal environment, which are the experimental variables of the decoherence measurements rather than mere hygiene.]*

75.6.2 Procedure

[Reserved: record the detected molecular flux against transverse detector position, extract the fringe visibility, and repeat as a function of the two quantities that carry the physics — the internal temperature of the molecules, set by the sublimation and by laser heating, and the residual gas pressure. The velocity selection matters and must be described: the fringe period depends on the wavelength, so an unselected thermal beam washes out its own pattern, and the observed visibility as a function of the width of the velocity window is itself a check on Equation (75.11)*.]*

75.6.3 Observations and data

[Reserved: detected flux against detector position, with the fitted visibility and its uncertainty, for each molecular species. To be tabulated when written: species, mass in atomic mass units, selected speed, deduced de Broglie wavelength, grating period, predicted and observed fringe period, and observed visibility; and separately the two decoherence scans, visibility against internal temperature and against residual gas pressure, with the predictions of Equation (75.12) *alongside.]*

Phenomenon 75.6 (Objects of thousands of atoms interfere at the de Broglie wavelength)**.** Interference survives in objects far larger and far more complicated than the wavelength they carry. Fullerene molecules diffracted from a nanofabricated grating produce fringes at a wavelength of about 2.5 pm, some hundreds of times smaller than the molecule itself, even though the molecules are hot enough to populate very large numbers of internal vibrational states and to radiate thermal photons while in flight [1966]. Near-field interferometry has since shown fringes for tailored molecules built from more than two thousand atoms, with masses beyond 25 000 atomic mass units, at wavelengths of order 50 fm [2337]. No deviation from $\lambda = h/p$ has appeared at any mass reached, and no boundary beyond which the relation fails has been found.

Derivation. The wavelengths first, since they are what make the claim surprising. A fullerene of sixty carbon atoms has mass $m = 720$ atomic mass units, and an oven beam of it selected near 220 m/s has

$$\lambda = \frac{h}{mv} = 2.5\,\mathrm{pm}\,, \tag{75.11}$$

some hundreds of times smaller than the molecule that carries it. For a molecule of 25 000 atomic mass units at a beam speed of order 300 m/s the same relation gives λ of order 53 fm: seven orders of magnitude below the wavelength of visible light, and five below the size of

the object itself. Nothing in $\lambda = h/p$ objects to either, and nothing in the observed fringe periods does either.

Why, then, does a chair not interfere? Not because it is heavy. Suppose a molecule, while in a superposition of two paths separated by d, emits a single photon of wavenumber k in a random direction. The photon state emitted from one path differs from that emitted from the other by the phase $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{d}}$, and the fringe visibility is multiplied by the overlap of the two photon states, averaged over emission direction:

$$|\langle\gamma_1|\gamma_2\rangle| = \left|\frac{1}{4\pi}\int \mathrm{d}\Omega\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{d}}\right| = \left|\frac{\sin kd}{kd}\right| . \tag{75.12}$$

Coherence survives while $kd \ll 1$ — while the emitted wavelength is long compared with the separation of the paths, so that the photon cannot say which path was taken — and is lost when kd approaches unity.

Two consequences follow, and they are the content of the modern experiments. The criterion contains no mass at all: a heavy object is not harder to interfere for being heavy, but for being harder to keep dark and harder to keep from being touched. And because a molecule in flight emits many photons, the visibility is the product of Equation (75.12) over all of them, so it does not decline gently but collapses, the number of emitted photons rising steeply with internal temperature while each factor falls below unity. Both knobs have been turned deliberately. Heating the molecules until their thermal emission reaches short enough wavelengths destroys the fringes [2060]; admitting a background gas, so that a single collision localizes the molecule to within a distance fixed by the momentum transfer, destroys them by the same argument and with the predicted pressure dependence [2052]. In neither case does the wavelength change, and in neither case is $\lambda = h/p$ at fault. □

Phenomenon 75.7 (What ends the interference is information, not mass)**.** The fringe visibility of a molecular interferometer is destroyed controllably by two means, both of which allow the environment to record which path the molecule took: raising the internal temperature of the molecules until the thermal photons they emit have wavelengths comparable with the path separation [2060], and admitting a background gas so that molecules are scattered from [2052]. In both cases the visibility falls with the measured dependence on temperature or pressure, and in both cases the de Broglie wavelength is unchanged. The transition from quantum to classical behaviour observed here is a loss of isolation and not a breakdown of the wave relation.

Derivation pending #74

The quantitative decoherence rates: the thermal photon emission spectrum of a hot molecule and the resulting product of which-path factors as a function of internal temperature, giving the predicted visibility curve; and the collisional case, in which the localization length follows from the momentum transfer distribution of the residual gas at a given pressure and temperature, giving the predicted decay of visibility with pressure. Both are stated qualitatively in the derivation above from the single-photon overlap; what must be supplied is the master equation whose decoherence rate they are, together with the statement of which of its assumptions the data actually test.

75.6.4 Interpretation

[Reserved: what the mass record does and does not bound. It does not establish that there is no limit; a null result at 25 000 *atomic mass units is a lower bound on where any limit lies. It does constrain proposed objective-collapse modifications of quantum mechanics, which predict a mass-dependent breakdown of superposition, and that constraint is a live experimental*

frontier reported as such in Chapter 133 rather than as a settled matter. The section should also record the honest reading of the decoherence measurements: they show a mechanism sufficient to explain why large objects do not interfere, and they do not show that it is the only one — which is the distinction Chapter 87 and Chapter 85 respectively make and decline to resolve.]

75.6.5 Primary references

[1966, 2337]. The decoherence measurements are [2052, 2060].

75.7 What the measurements settle

Table 75.1: The measurements of this chapter, and the alternative explanation each removes. The third column states what the experiment is entitled to claim, which in the last row is markedly weaker than what it is often credited with.

Experiment	Wavelength	What it establishes
Davisson–Germer 1927	0.165 nm	a wavelength computed from a voltage and one read off a lattice agree to about one percent
Thomson–Reid 1927	5–12 pm	ring radii scale as $V^{-1/2}$ in transmission through an unoriented target; not a surface effect
Estermann–Stern 1930	0.07 nm	neutral, inert helium is diffracted, and the maxima track the source temperature; not an effect of charge
Jönsson 1961; Tonomura 1989	5.4 pm	interference with no lattice, and with the apparatus almost always empty; not a collective effect
Colella *et al.* 1975	0.14 nm	the phase is an instrument reading and gravity shifts it by the predicted amount, in $m_{\mathrm{i}}m_{\mathrm{g}}$
Arndt 1999; Fein 2019	0.05–2.5 pm	no deviation up to 25 000 atomic mass units; what ends the fringes is which-path information

The rows of Table 75.1 are not repetitions of one another. Each was performed because the previous one left an explanation standing, and the sequence is the removal of those explanations one at a time until nothing is left but the relation itself. Taken together they confirm $\lambda = h/p$ over more than seven orders of magnitude in mass, from the electron to a molecule of two thousand atoms, and over three in wavelength, with no adjustable parameter anywhere and no observed departure. That is a stronger evidential position than the relation enjoyed when Chapter 71 adopted it, and stronger than most relations in this treatise enjoy.

Three things are not settled here, and the chapter is worth closing on them. The first is the meaning of the wave: the accumulation experiments establish that each particle interferes with itself and say nothing about what the amplitude is, and the identification of its squared modulus with a probability density is a postulate added by Born [533] and examined in Chapter 77. The second is the equation of motion: a wavelength for a free particle is not a wave equation, and everything about a particle in a potential — including

the bound states that Chapter 70 needed — is outside what any measurement in this chapter reaches. The third is the upper limit. No mass at which the relation fails has been found, but no experiment can establish that none exists; what the molecular interferometers deliver is a bound that moves upward with each generation, and the proposals it constrains are named honestly, and without endorsement, in Chapter 133.

A closing observation of a different kind. The relation has stopped being a hypothesis under test and become an instrument: electron and neutron diffraction are the standard probes of crystal and magnetic structure, and atom interferometers built on the same principle [1847] serve as gravimeters and gyroscopes and deliver h/m, and thence the fine-structure constant, to a precision that feeds back into the tests of the constants themselves. A hypothesis proposed from a symmetry, with no data behind it, now calibrates the measurements by which other hypotheses are judged; and the neutron bound states in the Earth's field [2034], in which the same wave is quantized by gravity alone, are the most recent reminder that the instrument is still finding new things to weigh.

Chapter 76

Experiment: X-ray Spectra and the Charge of the Nucleus

Chapter 70 makes one of the strongest claims in the book. The elements are ordered not by their atomic weights, as chemistry had assumed for a century, but by a whole number that counts elementary charges on the nucleus; and the evidence for this is spectroscopic, drawn from radiation far outside the visible range. The chapter states the law, derives its slope from the hydrogenic level formula with no adjustable parameter, and stops there. It never says where the radiation came from, how anyone knew it was radiation at all, or how a wavelength shorter than a visible one by four powers of ten was measured in the first place.

This chapter supplies that chain. It is unusual among the experiment chapters of this treatise in that no single measurement carries the verdict: what is on trial is the whole instrument, and the instrument had to be built from nothing in eighteen years. Röntgen found the radiation and could say almost nothing about its nature. Barkla showed that it is transverse — hence a wave in the electromagnetic field of Chapter 60 and not the longitudinal aether pulse Röntgen had guessed — and that each element, struck hard enough, answers with a radiation of its own. Friedrich, Knipping and von Laue diffracted it from a crystal, which established in one photograph both that X-rays are waves of wavelength comparable to an atomic spacing and that a crystal is a three-dimensional periodic lattice. The Braggs turned that observation into a measuring instrument by reducing the diffraction condition to a reflection law and mounting the crystal on a divided circle. Moseley put the elements one after another into that instrument. Siegbahn then drove its accuracy to the point where the X-ray lines became a term system as sharp as the optical one.

The reader will find the six experiments below in that order, each with its apparatus, its procedure, its observations and its verdict. What they establish jointly is stated at the end (Section 76.7). Five of the six were recognised by a Nobel Prize — Röntgen with the first prize in physics ever awarded, then von Laue, the Braggs, Barkla and Siegbahn in a run of four consecutive years of X-ray work. Moseley, who is missing from that list, was killed at Gallipoli in 1915 at the age of twenty-seven, and the prize is not awarded posthumously.

76.1 Röntgen: a new kind of rays (1895)

The discovery was made on 8 November 1895 with apparatus that a dozen laboratories in Europe already possessed. What Röntgen added was the black cardboard shield and the willingness to look at a fluorescent screen he had no reason to expect to glow.

Table 76.1: The six experiments of this chapter, what each settled, and its recognition. The prize year is the year the prize is *for*; several of these were reserved and awarded the following year.

Experiment	Year	What it settled	Prize
Röntgen	1895	A penetrating radiation exists; it is not cathode rays and is not refracted by matter	1901
Barkla	1905–1911	X-rays are transverse waves; every element has its own K and L radiation	1917
Friedrich, Knipping, von Laue	1912	Crystals diffract X-rays: X-rays are waves, crystals are lattices	1914
W. H. and W. L. Bragg	1913	The reflection law and the spectrometer built on it; absolute wavelengths	1915
Moseley	1913–1914	The ordering integer of the elements is the nuclear charge	—
Siegbahn	1916–1925	X-ray lines form a precise term system with resolved subshells	1924

76.1.1 Apparatus

A Hittorf–Crookes discharge tube — a partially evacuated glass vessel with two electrodes — driven by a Ruhmkorff induction coil, the whole tube wrapped in a close-fitting shroud of black cardboard that passed no visible light. A screen coated with barium platinocyanide served as the detector; photographic plates in light-tight holders served as the recording detector. A darkened room. For the tests of the nature of the radiation: sheets of paper, wood, ebonite, aluminium, lead, platinum and other metals of graded thickness; prisms of water, carbon disulfide, ebonite and aluminium; polished surfaces; a strong magnet; and an electroscope.

76.1.2 Procedure

Run the discharge in the darkened room with the shrouded tube and observe the screen at increasing distances and orientations. Interpose absorbers of known material and thickness between tube and screen and compare the transmitted brightness. Attempt to refract the radiation with the prisms and to reflect it regularly from polished surfaces. Place the magnet across the path and look for a deflection. Bring a charged electroscope into the irradiated region. Expose plates behind absorbers of graded density, including a hand.

76.1.3 Observations and data

The screen fluoresces brightly whenever the discharge runs, at distances of the order of 2 m from the tube and with the cardboard shroud in place; a sheet of paper, a bound book of about a thousand pages, deal boards, or thin sheets of aluminium interposed in the beam weaken but do not extinguish the effect, while a fraction of a millimetre of lead or platinum stops it. Transmission is governed chiefly by the mass of material traversed, and for equal masses the denser elements absorb more strongly. The radiation proceeds in straight lines from the spot on the tube wall struck by the cathode rays and casts sharp shadows; a hand held in the beam casts a shadow in which the bones stand out darker than the flesh.

Photographic plates are blackened. Charged bodies in the irradiated air lose their charge. No regular refraction could be detected with any of the prisms and no regular reflection from any polished surface; a magnet across the path produced no deflection whatever [269].

Phenomenon 76.1 (A penetrating radiation from the discharge tube)**.** The place where cathode rays strike a solid emits a radiation that travels in straight lines, is not deflected by a magnetic field, excites fluorescence, blackens a photographic plate, discharges electrified bodies, and is absorbed in a way governed chiefly by the mass of matter traversed. It shows no measurable regular refraction or reflection at the surfaces of ordinary bodies [269].

Derivation pending #75

X-ray spectra and the charge of the nucleus: the classical account of the emission itself, namely the bremsstrahlung continuum radiated by cathode-ray electrons decelerated in the anticathode, and its short-wavelength cut-off.

The absence of refraction is not a null result but a measurement, and it is worth extracting the number it contains.

Phenomenon 76.2 (The refractive index of matter for X-rays)**.** For X-rays the refractive index of ordinary matter differs from unity by so little that no prism available in 1895 could reveal it [269]; the deviation is of order 10^{-5} or smaller, and — unlike the case of visible light — the index is slightly *less* than one.

Derivation. Model the electrons of the medium as bound oscillators of resonance frequency ω_0, as in the dispersion theory of Chapter 61. For a number density N of such electrons the refractive index satisfies

$$n^2 = 1 + \frac{Ne^2}{\varepsilon_0 m_\mathrm{e}} \frac{1}{\omega_0^2 - \omega^2} \,. \tag{76.1}$$

Every atomic resonance lies in the ultraviolet or below, so for X-rays $\omega \gg \omega_0$ and the binding drops out: the electrons respond as if free,

$$n^2 \simeq 1 - \frac{\omega_\mathrm{p}^2}{\omega^2} \,, \qquad \omega_\mathrm{p}^2 = \frac{Ne^2}{\varepsilon_0 m_\mathrm{e}} \,, \tag{76.2}$$

so that $n < 1$ and $1 - n \simeq \omega_\mathrm{p}^2 / 2\omega^2$. A solid holds of order $10^{30}/\mathrm{m}^3$ electrons, which puts ω_p in the far ultraviolet, while an X-ray angular frequency is two or three orders of magnitude larger again; the quantity $\omega_\mathrm{p}^2/2\omega^2$ is then the small number quoted. A prism deviates a beam by an angle of order $(n-1)$ times its apex angle, and 10^{-5} of a right angle is a few seconds of arc — beyond the reach of a fluorescent screen viewed by eye. The sign is the physically interesting part: it says the phase velocity in matter exceeds c for X-rays, which is why X-ray optics uses grazing-incidence total external reflection and why no ordinary lens focuses them. □

76.1.4 Interpretation

Röntgen's own conclusion was deliberately negative. The radiation is not cathode rays, because a magnet does not deflect it. It is not ultraviolet light, because it is not refracted, not regularly reflected, and could not be polarised by any means he tried. Lacking any positive evidence of transversality he suggested, tentatively, that the rays might be longitudinal waves in the aether. That guess was wrong, and it is the honest reason this chapter must continue: the question "what are X-rays?" was left open by the man who found them, and was closed by Section 76.2 and Section 76.3.

76.1.5 Primary references

[269].

76.2 Barkla: characteristic radiation and the transversality of X-rays (1905–1911)

Barkla worked with an electroscope and an ionisation chamber, without ever resolving a spectral line, and settled two questions that resisted every more direct attack: whether X-rays are transverse, and whether the radiation an element emits belongs to the element.

76.2.1 Apparatus

For the polarisation experiment: an X-ray bulb giving a primary beam; a first scatterer — a block of a light element, paper or carbon, which scatters without appreciable absorption — placed in that beam; a second scatterer placed in the secondary beam that leaves the first at right angles; and an ionisation chamber with an electrometer, mounted so that it can be carried around the axis of the secondary beam while remaining at right angles to it.

For the characteristic-radiation experiments: the same primary bulb, with the beam falling on a radiator consisting of the element under test or one of its compounds; an ionisation chamber viewing the radiator at right angles to the primary beam; and a set of aluminium sheets of accurately known thickness to be interposed one at a time in the secondary beam.

76.2.2 Procedure

In the polarisation experiment, record the ionisation current produced by the twice-scattered beam as the chamber is carried around the axis of the once-scattered beam, and plot it against the azimuth.

In the absorption experiments, record the ionisation current as successive aluminium sheets are added, and read the mass absorption coefficient off the exponential decline. A radiation is called *homogeneous* when the decline is a single exponential over the whole range of thicknesses used. Repeat with radiator elements taken in order through the periodic table, and repeat each radiator with primary beams of different hardness, obtained by varying the tube.

76.2.3 Observations and data

The twice-scattered intensity is not the same in all azimuths: it passes through a maximum and a minimum ninety degrees apart, the maximum lying in the direction of the original primary beam. The modulation is partial rather than complete, as the finite apertures and the multiple scattering in the radiators require [317].

For radiators heavier than about calcium the secondary beam contains, besides the scattered primary, a component that is homogeneous. Its mass absorption coefficient in aluminium is a property of the radiator element and not of the primary beam; it decreases monotonically as one proceeds up the periodic table, so that heavier elements emit a more penetrating characteristic radiation. The component appears only when the primary beam is itself harder than the characteristic radiation it excites — there is a threshold — and for the heavier elements a second, much less penetrating homogeneous radiation appears alongside the first [340, 352]. Barkla named the more penetrating series K and the softer one L, choosing letters in the middle of the alphabet in case further series should be found on either side [367]. They were: M, N and O followed.

Phenomenon 76.3 (X-rays are transverse)**.** Radiation scattered by a light element at right angles to the primary beam is partially polarised: on being scattered a second time through a right angle its intensity depends on the azimuth, with a maximum along the direction of the original primary beam and a minimum at right angles to it [317]. A longitudinal disturbance has no transverse vector to orient and can produce no such dependence, so X-rays are transverse waves.

Derivation. Let the primary beam travel along $\hat{\boldsymbol{z}}$ and be unpolarised. Its electric vector lies in the xy plane, so the electrons of the first radiator are accelerated in that plane. An oscillating charge radiates nothing along its own line of acceleration and, at an angle Θ from it, with an intensity proportional to $\sin^2\Theta$ (Chapter 65). Observe the scattered beam along $\hat{\boldsymbol{x}}$: the x component of the acceleration contributes nothing in that direction, so the surviving radiation is that driven by the y component alone. The beam leaving the first radiator at right angles is therefore linearly polarised, with its electric vector along $\hat{\boldsymbol{y}}$.

That beam drives the electrons of the second radiator along $\hat{\boldsymbol{y}}$. Put the chamber in the plane perpendicular to $\hat{\boldsymbol{x}}$, in the direction $\hat{\boldsymbol{n}} = \hat{\boldsymbol{y}}\sin\varphi + \hat{\boldsymbol{z}}\cos\varphi$, where φ is measured from the original primary direction. Then

$$I(\varphi) \propto \sin^2\Theta = 1 - (\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{y}})^2 = \cos^2\varphi\,, \tag{76.3}$$

which is maximal at $\varphi = 0$, along the original primary beam, and vanishes at $\varphi = \pi/2$. The observed modulation has exactly this period and this phase; it is incomplete because real apertures average Equation (76.3) over a range of scattering angles. For a longitudinal wave the displacement would lie along $\hat{\boldsymbol{z}}$ for every ray of the primary beam, the once-scattered beam would carry no azimuthal marker at all, and $I(\varphi)$ would be flat. □

Phenomenon 76.4 (Characteristic radiation of the elements)**.** An element struck by a sufficiently hard primary beam emits, in addition to scattered radiation, one or more homogeneous radiations whose absorption coefficients are properties of that element alone and are independent of the primary beam that excites them. Their penetrating power increases monotonically along the periodic table. Each such radiation has an excitation threshold: it appears only when the exciting radiation is harder than the radiation excited. For the heavier elements two series are present, the penetrating K series and the softer L series [340, 352, 367].

Derivation pending #76

Characteristic X-radiation: the quantum account, in which the primary beam ejects an electron from an inner shell and the radiation is emitted when an outer electron fills the vacancy, so that the excitation threshold is the binding energy of the shell and the monotone hardening along the table is the growth of that binding energy with nuclear charge.

76.2.4 Interpretation

The polarisation result closed the question Röntgen had left open. X-rays are transverse, hence electromagnetic waves of Chapter 60, and the aether pulse can be dropped. The characteristic-radiation result is the more consequential of the two for what follows. It says that the radiation carries the signature of the emitting atom, and therefore that a spectroscopy of the elements in this range is possible in principle. Barkla could rank the elements by absorption coefficient but could not attach a wavelength to any of them, because he had no dispersing element; his K and L are ordinal labels, not measurements. The instrument that converted his ordering into wavelengths arrives in the next two sections,

and Moseley's law is the result of running it over Barkla's own K and L radiations. It is worth noticing that the monotone ordering Barkla found already contained Moseley's law in disguised form: absorption in aluminium increases steeply with wavelength, so "more penetrating for heavier elements" is "shorter wavelength for larger nuclear charge".

76.2.5 Primary references

[317, 340, 352, 367].

76.3 Friedrich, Knipping and von Laue: diffraction by a crystal (1912)

The experiment was proposed on a guess and worked at once. Von Laue reasoned that if X-rays were waves of about the size of an atom, and if a crystal were a periodic array of atoms, then the crystal was a ready-made three-dimensional diffraction grating that nobody would have to rule. Neither premise was established at the time; the photograph established both together.

76.3.1 Apparatus

A gas X-ray tube; a lead-lined box carrying a series of pinhole apertures which cut from the emitted radiation a narrow pencil a fraction of a millimetre across; a crystal mounted on a goniometer at the exit of the collimator, so that a chosen crystallographic direction could be set parallel to the pencil; and photographic plates in light-tight holders placed behind the crystal, and in the first trials also to the sides and in front of it, to catch radiation scattered in any direction. The first crystal was copper sulfate, chosen because a good specimen was to hand; the decisive photographs were taken with zinc blende, which is cubic and of high symmetry.

76.3.2 Procedure

Set the crystal in the pencil, expose the plate for a period of hours, and develop. Then rotate the crystal so that a symmetry axis lies along the pencil, and repeat. Compare the pattern with that obtained when the crystal is replaced by an amorphous body of similar absorbing power, and with that obtained when the crystal is displaced out of the beam.

76.3.3 Observations and data

The plate behind the crystal shows, in addition to the trace of the undeviated pencil, a set of discrete, sharp spots at well-defined angles. The spots are arranged about the central trace with the point symmetry of the crystal in the direction of the beam: with the pencil along a symmetry axis of zinc blende the pattern reproduces the symmetry of that axis. Rotating the crystal moves the spots as a rigid figure, and displacing the crystal out of the pencil extinguishes them all. An amorphous absorber in the same place gives a diffuse distribution and no spots [385]. Von Laue's companion paper computes the directions of the maxima from a lattice model and reproduces the observed positions [389].

Phenomenon 76.5 (Crystal diffraction of X-rays)**.** A narrow pencil of X-rays traversing a single crystal is scattered into a set of discrete, sharp maxima whose directions are fixed by the orientation of the crystal and whose arrangement carries the symmetry of the crystal. An amorphous body in the same geometry produces no such maxima [385, 389].

Derivation. Discrete maxima at fixed angles are the signature of interference between many scatterers held in a periodic arrangement, and the argument fixes both the wave nature of the radiation and the periodicity of the crystal at once. Let $\hat{\boldsymbol{s}}_0$ and $\hat{\boldsymbol{s}}$ be unit vectors along the incident and the scattered directions, and let the scattering centres sit at the lattice points

$$\boldsymbol{R} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2 + n_3\boldsymbol{a}_3\,, \qquad n_i \in \mathbb{Z}\,, \tag{76.4}$$

with $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ the primitive vectors of the lattice. Measured between a plane wavefront before the crystal and one after it, the optical path through the centre at $\boldsymbol{R}$ is shorter than the path through the centre at the origin by $\boldsymbol{R}\cdot(\hat{\boldsymbol{s}}_0 - \hat{\boldsymbol{s}})$. The wavelets from all the centres arrive in phase precisely when this is an integer multiple of the wavelength for every $\boldsymbol{R}$, and since the n_i range independently over $\mathbb{Z}$ that is equivalent to the three simultaneous conditions

$$\boldsymbol{a}_i \cdot (\hat{\boldsymbol{s}}_0 - \hat{\boldsymbol{s}}) = h_i\lambda\,, \qquad h_i \in \mathbb{Z}\,, \quad i = 1,2,3\,. \tag{76.5}$$

Three equations constrain the two free angles of $\hat{\boldsymbol{s}}$ together with λ, so for a crystal at a general orientation in a *monochromatic* beam there is no solution and no spot appears. The resolution is that the beam of a gas X-ray tube is not monochromatic: it carries the continuous bremsstrahlung spectrum, and each spot on the plate selects out of that continuum the one wavelength for which Equation (76.5) can be satisfied in that direction. Given the observed angles and a lattice spacing of order the atomic diameter, Equation (76.5) returns wavelengths of order 100 pm, four orders of magnitude below visible light. □

76.3.4 Interpretation

One photograph settled two disputed questions in different fields. For physics, X-rays interfere, so they are waves, and their wavelength is of the order of an atomic spacing — which explains at a stroke why Röntgen found no refraction and no reflection with ordinary optics, and why Phenomenon 76.2 holds. For crystallography, the sharp maxima require the scatterers to sit on a lattice with long-range three-dimensional periodicity: the lattice hypothesis, until then an inference from the external shapes and cleavage of crystals, became a direct observation. Everything the treatise says about periodic solids in Chapter 125 and Chapter 126 rests on this measurement and on the industry of structure determination it founded.

76.3.5 Primary references

[385, 389].

76.4 The Braggs: the reflection law and the X-ray spectrometer (1913)

Laue's spots proved that X-rays are waves but were an awkward tool: a Laue photograph mixes all wavelengths and all lattice planes at once. The Braggs' contribution was to reorganise the same geometry so that one wavelength and one set of planes are isolated, and then to hang the whole arrangement on a divided circle. What had been a demonstration became an instrument, and every measurement in the two sections that follow was made with it.

76.4.1 Apparatus

An X-ray tube with a metal anticathode; a pair of lead slits defining a beam of small angular width; a cleaved crystal — rock salt in the first work, later calcite and mica — mounted

on the table of a spectrometer with a divided circle; and, on an arm turning about the same axis, an ionisation chamber filled with a heavy gas so as to absorb the beam efficiently, connected to a quadrant electrometer. The crystal table and the chamber arm are geared or set so that when the crystal face makes a glancing angle with the incident beam, the chamber sits at twice that angle from the beam — the specular position.

76.4.2 Procedure

Set the crystal face at a small glancing angle and the chamber at the corresponding specular angle; record the ionisation current; advance both by a small step and repeat. The record of current against glancing angle is the spectrum of the tube as dispersed by that crystal face. Change the anticathode metal and repeat, to determine which features belong to the target and which to the crystal or the apparatus.

76.4.3 Observations and data

The current against glancing angle is a smooth continuous background on which sharp peaks stand out. The peaks come in families: within a family the glancing angles θ_n satisfy $\sin\theta_n = n\sin\theta_1$ for $n = 1, 2, 3, \dots$, so that a family is a single feature seen in successive orders. The angular positions of the families change when the anticathode metal is changed, and do not change when the crystal is exchanged for another face of the same crystal at the appropriate spacing; the peaks therefore belong to the target metal and are Barkla's characteristic radiations, now resolved and measurable [399, 400].

Phenomenon 76.6 (The Bragg reflection law)**.** A crystal illuminated by X-rays at a glancing angle θ to a set of lattice planes of spacing d returns a strong specularly reflected beam only at those angles satisfying

$$n\lambda = 2d\sin\theta\,, \qquad n \in \mathbb{N}\,, \tag{76.6}$$

where λ is the wavelength; at intermediate angles the reflected intensity is negligible [399, 400].

Derivation. Group the lattice points of Equation (76.4) into a family of parallel planes with perpendicular spacing d. Consider two rays of the incident beam striking adjacent planes and leaving them at the same glancing angle θ on the far side of the normal, that is, specularly. Dropping perpendiculars from the point of incidence on the upper plane to the incident and emergent rays of the lower one shows the extra path travelled by the lower ray to be $2d\sin\theta$; constructive interference between the two requires that this be a whole number of wavelengths, which is Equation (76.6). Reflection at a single plane is specular for any angle because the scatterers within a plane are dense compared with the wavelength; it is the interference *between* planes that selects the discrete set.

This is not a second law but the first one rewritten. In Equation (76.5) the vector $\hat{\boldsymbol{s}}_0 - \hat{\boldsymbol{s}}$ is common to all three conditions; satisfying them simultaneously forces it to lie along the normal of some lattice plane family, and for a specular geometry at glancing angle θ its length is $2\sin\theta$. Taking $\boldsymbol{R}$ along that normal with $|\boldsymbol{R}| = d$ turns Equation (76.5) into Equation (76.6).

What makes Equation (76.6) an instrument rather than a relation is that d can be obtained without any X-ray measurement at all. In rock salt the planes parallel to a cube face carry sodium and chlorine alternately, so their spacing is half the edge a of the conventional cubic cell, and that cell contains four formula units. Hence

$$d = \frac{a}{2} = \frac{1}{2}\left(\frac{4M}{\rho N_{\mathrm{A}}}\right)^{1/3}\,, \tag{76.7}$$

with M the molar mass, ρ the density and N_{A} Avogadro's constant. Using $M = 58.44 \times 10^{-3}\,\mathrm{kg/mol}$, $\rho \approx 2.16 \times 10^{3}\,\mathrm{kg/m^{3}}$ and the present-day exact value $N_{\mathrm{A}} = 6.022\,140\,76 \times 10^{23}/\mathrm{mol}$ of Chapter 2, Equation (76.7) gives $a \approx 564\,\mathrm{pm}$ and $d \approx 282\,\mathrm{pm}$. Every wavelength in this chapter is ultimately referred to a number of this kind, so the X-ray wavelength scale is only as good as the value of N_{A} used to set it — a debt that came due in Section 76.6. □

76.4.4 Interpretation

The spectrometer converts an angle read on a divided circle into a wavelength, absolutely, through Equation (76.6) and Equation (76.7). Three things follow immediately. Barkla's homogeneous radiations become spectral lines with wavelengths. The crystal becomes a monochromator, so that a structure determination can be run the other way round — known wavelength in, unknown spacings out — which is what the younger Bragg did for the alkali halides, diamond and much else, founding X-ray crystallography as a working science. And the incident wavelength of the Compton measurement of Chapter 69 becomes a known quantity, which is what makes the shift there a measurement rather than an impression.

76.4.5 Primary references

[399, 400].

76.5 Moseley: the high-frequency spectra of the elements (1913–1914)

Moseley took the finished instrument and did the obvious thing that nobody had done: he put the elements through it one after another, in order, and looked at the sequence rather than at any one spectrum. The law he found, and its derivation from the hydrogenic level formula, are stated in Chapter 70 and are not repeated here; what follows is the measurement.

76.5.1 Apparatus

A Bragg-type crystal spectrometer with a crystal of potassium ferrocyanide, recording photographically rather than by ionisation chamber so that a whole spectrum is captured in one exposure. The essential addition was to the source: inside the evacuated tube ran a small trolley carrying a dozen target elements in a row, drawn along by a thread led out through a seal, so that one element after another could be brought into the cathode-ray beam without letting the tube up to air. Wavelengths were referred to the known spacing of the crystal.

76.5.2 Procedure

Bring a target into the beam, expose the plate through a chosen range of glancing angles, advance the trolley, and expose the next element on the same plate so that neighbouring elements appear side by side and can be compared directly. Measure the positions of the lines on the developed plate, reduce them to glancing angles, and convert to wavelength with Equation (76.6). Identify in each spectrum the strong pair of lines belonging to Barkla's K series, and for the heavier elements the group belonging to the L series.

76.5.3 Observations and data

Every element examined shows the same small number of strong lines, and the lines move steadily and by regular steps as one passes from element to element along the plate. Plotting

the square root of the frequency of the strongest K line against the ordinal number of the element in the periodic table gives a straight line, of a slope that the hydrogenic formula reproduces without adjustment; the same holds, with a different slope, for the L series. The first paper covered the range from calcium to zinc; the second extended it from aluminium to gold, using the K series for the lighter half of the range and the L series for the heavier [408, 418]. Between aluminium and gold the sequence of ordinal numbers is unbroken except at three places, corresponding to elements not then known and later found.

Phenomenon 76.7 (The ordering integer is the nuclear charge)**.** The integer that orders the characteristic X-ray spectra of the elements, advancing by exactly one from each element to the next, is the same integer that alpha-particle scattering measures as the charge of the nucleus in units of the elementary charge [408, 418, 464]. It is not the atomic weight: the ordering by weight has inversions — argon and potassium, cobalt and nickel, tellurium and iodine — and the ordering by X-ray spectra has none.

Derivation. Two independent routes to the same integer, one spectroscopic and one mechanical, agree.

The spectroscopic route is Moseley's, derived in Chapter 70: the innermost electrons of a heavy atom move in the almost unscreened Coulomb field of the nucleus, the hydrogenic levels of that field scale as the square of the nuclear charge, and the Kα transition therefore has $\sqrt{\nu}$ proportional to the nuclear charge less a screening constant of order unity. What identifies the plotted integer with a charge, rather than merely with a serial position, is that the slope of the straight line is fixed — it must be $(3cR_\infty/4)^{1/2}$, containing only the Rydberg constant already measured in the optical spectrum of hydrogen. A serial number with no physical content could reproduce the straightness but not the slope.

The mechanical route uses no spectroscopy at all. The Rutherford cross-section for the scattering of alpha particles of kinetic energy E by a nucleus of charge Ze (Chapter 107, [379]) is

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \left(\frac{2Ze^2}{16\pi\varepsilon_0 E}\right)^2 \frac{1}{\sin^4(\theta/2)}\,, \tag{76.8}$$

in which Z enters as a square and nothing else about the atom enters at all. Counting the alphas scattered into a known solid angle at a known angle and energy therefore weighs the nuclear charge directly. Chadwick did this for copper, silver and platinum and obtained values agreeing with their ordinal numbers to within an experimental uncertainty of about one per cent [464]. The hypothesis that the two integers are one integer, put forward by van den Broek while Moseley was at work [411], is thus a measured fact and not a convention. □

76.5.4 Interpretation

The consequences are structural rather than numerical. The periodic table acquires an exact ordinal parameter, so the three weight inversions that had embarrassed it since Mendeleev cease to be anomalies and become confirmations: the X-ray spectra put argon before potassium, cobalt before nickel and tellurium before iodine, which is the chemically correct order in every case. The number of elements between any two is fixed, so the count of missing elements ceases to be a matter of opinion; Moseley's range left exactly three gaps below gold, which is exactly what was later filled. The number of rare-earth elements, contested for decades, is likewise settled by counting the gaps in their L spectra. And the nuclear model of the atom acquires its missing parameter: after Moseley an element is specified by an integer Z, the atomic weight being a separate and, as Chapter 107 shows, a compound quantity. The isotope — same Z, different weight — becomes intelligible in the same stroke.

76.5.5 Primary references

[408, 418]. The independent scattering determination of the nuclear charge is [464], resting on [379]; the hypothesis tested is [411].

76.6 Siegbahn: precision X-ray spectroscopy (1916–1925)

Moseley's lines were sufficient for counting but not for anatomy. What looked like one line at his resolution is generally a group, and the groups are what a theory of the many-electron atom actually has to predict. Siegbahn spent a decade rebuilding the spectrometer around that requirement.

76.6.1 Apparatus

A vacuum spectrograph: the whole path from anticathode to plate is evacuated, so that soft radiation is not absorbed in air and the range of the instrument extends far beyond the K lines of the heavy elements. A demountable, water-cooled tube with interchangeable anticathodes, run at high load to obtain usable intensity in a narrow line. Precisely ruled circles carrying cleaved crystals of accurately known spacing — calcite, gypsum, sugar and mica were used for different ranges. Narrow slits and long crystal-to-plate distances for high dispersion. Photographic recording with comparison lines on the same plate. In the later work, ruled concave gratings used at grazing incidence, which measure a wavelength against a ruling interval and so make no assumption whatever about crystal structure.

76.6.2 Procedure

Photograph the spectrum of each anticathode element, measure the line positions against the comparison lines with a travelling microscope, reduce them to glancing angles, and convert to wavelength through Equation (76.6) with the tabulated crystal spacing. Repeat with different crystals to test for systematic error in the spacings. Separately, measure selected lines with the grazing-incidence grating and compare the two wavelength scales.

76.6.3 Observations and data

Each of Barkla's series resolves into a group of components: the K group into a close doublet and further lines, the L group into a set that requires three distinct L levels to explain it, the M group into a set requiring five. Every measured line frequency is a difference of two terms drawn from a table belonging to the element, so that the combination principle familiar from optical spectra holds here as well; the terms are the K, L_1–L_3, M_1–M_5 levels and their successors. The absorption edges measured in the same instrument fall at the term values themselves, just above the emission lines of the corresponding series, which identifies each term as the energy required to remove an electron from that level. Line wavelengths are reproducible to a few parts in ten thousand and better, which is what makes such an analysis possible at all [462, 531].

The comparison of the crystal scale with the grating scale exposed a systematic discrepancy of about two parts in a thousand. It was traced not to either instrument but to Equation (76.7): the crystal spacings had been computed with the value of N_{A} then current, which rested in turn on the value of the elementary charge then current. Siegbahn's X unit, defined by reference to a crystal spacing, is therefore not exactly $10^{-13}\,\mathrm{m}$, and X-ray wavelengths quoted in it must be corrected by that factor before being compared with anything measured optically.

Phenomenon 76.8 (The X-ray term system)**.** The characteristic X-ray lines of an element are differences of a small set of terms belonging to that element, and satisfy the combination principle. The terms occur in groups — one K level, three L levels, five M levels — and each term equals, to within the accuracy of the measurement, the absorption edge of the corresponding series, that is, the energy needed to remove an electron from that level [462, 531].

Derivation pending #77

The X-ray term system: the derivation of the K, L and M level multiplicities from the quantum numbers of the many-electron atom, with the L and M splittings as the combined effect of screening and of spin-orbit coupling, and the identification of an absorption edge with a one-electron binding energy.

76.6.4 Interpretation

Two results of lasting consequence come out of this work. The first is that the inner structure of the atom is directly readable. The multiplicities one, three, five are not fitted to the data; they are the counts of distinct inner levels, and any theory of the atom must reproduce them — as the quantum numbers of Chapter 88 and the spin-orbit coupling of Chapter 79 do. The identification of terms with absorption edges makes each of these levels a measurable binding energy, which is the quantity the excitation threshold of Phenomenon 76.4 had been signalling since Barkla.

The second is methodological, and it is the reason the discrepancy above is reported here rather than buried. A wavelength scale built on Equation (76.7) inherits every uncertainty in N_{A}, so an X-ray wavelength was never a purely spectroscopic quantity: it was tied to a mass, a density and a counting constant. That is no longer a limitation, since N_{A} is exact by definition in the present SI (Chapter 2); but the episode is a clean example of a systematic error that no amount of repetition inside one method can reveal, and that only an independent method — here the ruled grating — can expose.

76.6.5 Primary references

[462, 531].

76.7 What the X-ray spectra established

Taken in order, the six experiments make a single argument, and it is worth setting the conclusion down without hedging.

There exists a penetrating radiation emitted where fast electrons are stopped (Phenomenon 76.1). It is a transverse wave (Phenomenon 76.3), hence a solution of the field equations of Chapter 60, with a frequency so far above every atomic resonance that matter is very nearly transparent to it and its refractive index is barely below one (Phenomenon 76.2). Its wavelength is comparable to the spacing of atoms in a solid, which both crystal diffraction (Phenomenon 76.5) and the reflection law built on it (Phenomenon 76.6) make measurable in absolute units. Each element, struck hard enough, emits a characteristic radiation of its own (Phenomenon 76.4), resolvable into a term system with a definite inner-shell structure (Phenomenon 76.8). And the integer that orders those spectra is the charge on the nucleus (Phenomenon 76.7).

The last statement is the one that reorganised chemistry. Before Moseley the periodic table was an empirical arrangement whose ordering principle — atomic weight — was known

to fail in at least three places and whose length was unknown. After Moseley the table is indexed by an integer with a physical meaning, the number of protons in the nucleus, and the failures of the weight ordering are explained rather than tolerated: weight is a nuclear quantity that depends on neutron number as well, and two samples of one element may differ in it. Chapter 107 takes that up. What matters here is that the identification was not an interpretation placed on the data but a measurement, made twice by unrelated means — a spectroscopic plate and a count of scattered alpha particles — and agreeing.

It should be said plainly that this chapter also records how much can be achieved without a theory. Barkla ranked the elements correctly by a property he could not name, using an electroscope and a stack of aluminium foil. Von Laue's photograph settled a question in crystallography with a piece of zinc blende and a photographic plate. Moseley's decisive instrument was a spectrometer and a trolley pulled by a thread. In each case the quantum theory that explains the result arrived afterwards, and in Barkla's and von Laue's cases the discoverer guessed wrongly about the mechanism while being entirely right about the fact. The facts are what this treatise records; the explanations are checked against them, not the other way round.

Part IX

Quantum Mechanics

Chapter 77

The Postulates of Quantum Mechanics

Quantum mechanics is not derived from classical mechanics; it rests on its own postulates, which state what a physical state is, which quantities can be measured, what the possible outcomes of a measurement are and with what probabilities they occur, how a measurement alters the system, and how composite systems are described. This chapter collects the conventional postulates in the state-vector formulation, together with the canonical quantization rules and the notion of uncertainty on which Heisenberg's principle is built. The mathematical arena — Hilbert spaces and the operators acting on them — is developed in Chapter 12. Standard modern treatments of this material are found in [1593, 2287, 2313]. The source for this chapter is an early outline: its completed content is ported in full below, and the headings it reserves without content are marked as pending rather than filled in with material the source does not contain.

77.1 The conventional postulates

77.1.1 The state postulate

Let $\mathcal{H}$ be a Hilbert space over $\mathbb{C}$ and let $|\psi\rangle \in \mathcal{H}$. The relation

$$\mathrm{e}^{\mathrm{i}\phi}\,|\psi\rangle \sim \mathrm{e}^{\mathrm{i}\gamma}\,|\psi\rangle \iff \phi,\gamma \in \mathbb{R}$$

is an equivalence relation. We say that an equivalence class is a *ray* in $\mathcal{H}$.

Postulate 77.1 (Quantum physical states)**.** A quantum physical state is represented by a ray of normalized vectors, that is, vectors satisfying

$$\langle\psi|\psi\rangle = 1\,. \tag{77.1}$$

If no ambiguity is possible, we shall say that the state of a system is one of its representatives with vanishing global phase; that is, the state will be $|\psi\rangle$.

77.1.2 The observable postulate

Let $O \in \mathcal{L}(\mathcal{H})$ be a linear operator on $\mathcal{H}$.

Postulate 77.2 (Observable quantities)**.** To every measurable quantity there corresponds a Hermitian operator O.

We shall call such an operator an *observable*. It must be noted that not every Hermitian operator is an observable.

77.1.3 The measurement postulate

Consider the spectral decomposition of an observable O,

$$O = \sum_i \lambda_i \, |\lambda_i\rangle\langle\lambda_i| \; . \tag{77.2}$$

Postulate 77.3 (Measurement outcomes)**.** The possible results of a measurement of O are its eigenvalues λ_i.

77.1.4 The probability postulate: the Born rule

Consider a quantum system in the state $|\psi\rangle$, and let O be an observable with eigenvalues λ_i.

Postulate 77.4 (Born rule)**.** The probability of obtaining λ_i in a measurement of O is given by

$$P(\lambda_i, |\psi\rangle) = |\langle\lambda_i|\psi\rangle|^2 \; . \tag{77.3}$$

We must note that the probability is determined on the whole ray of $|\psi\rangle$, since $\left|\mathrm{e}^{\mathrm{i}\phi}\right| = 1$. Moreover, by the theory of probability, it must hold that

$$\sum_i P(\lambda_i, |\psi\rangle) = 1 \, . \tag{77.4}$$

Expanding $|\psi\rangle$ in the eigenbasis $\{|\lambda_i\rangle\}_{i=1}^{\dim \mathcal{H}}$ as $|\psi\rangle = \sum_j a_j \, |\lambda_j\rangle$, we have

$$\begin{aligned} P(\lambda_i, |\psi\rangle) &= |\langle\lambda_i|\psi\rangle|^2 \\ &= \left| \langle\lambda_i| \left(\sum_j a_j \, |\lambda_j\rangle \right) \right|^2 \\ &= \left| \sum_j a_j \, \langle\lambda_i|\lambda_j\rangle \right|^2 \\ &= \left| \sum_j a_j \delta_{ij} \right|^2 \\ &= |a_i|^2 \; , \end{aligned}$$

so the probability of obtaining λ_i in a measurement is the squared modulus of the i-th component of $|\psi\rangle$ in the eigenbasis.

77.1.5 The collapse postulate

Consider a quantum system in the state $|\psi\rangle$, and let O be an observable with eigenvalues λ_i.

Postulate 77.5 (Alteration of quantum systems)**.** If a measurement of O yields the eigenvalue λ_i, then immediately after the measurement the state of the system is $|\lambda_i\rangle$.

77.1.6 The time-evolution postulate

The source records for this postulate only the stationary eigenvalue problem of the Hamiltonian,

$$H \, |\psi\rangle = E \, |\psi\rangle \; . \tag{77.5}$$

[Heading reserved in the source outline; the statement of the time-dependent Schrödinger equation governing the evolution of quantum states is pending port from a completed source.]

77.1.7 The composite-system postulate

Consider a quantum system composed of N individual systems whose states are rays in the Hilbert spaces $\mathcal{H}_1, \ldots, \mathcal{H}_N$.

Postulate 77.6 (Composite quantum systems)**.** The state of the composite system is a ray in the Hilbert space

$$\mathcal{H}_T = \mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_N \,. \tag{77.6}$$

77.2 Density-matrix postulates

[Heading reserved in the source outline; the reformulation of the postulates in terms of density matrices, covering statistical mixtures as well as pure states, is pending port from a completed source.]

77.3 On Planck's constant

[Heading reserved in the source outline; the discussion of Planck's constant, its role as the quantum of action, and its measured value is pending port from a completed source.]

77.4 Pictures of time evolution

77.4.1 The Schrödinger picture

[Heading reserved in the source outline; no content to port.]

77.4.2 The Heisenberg picture

[Heading reserved in the source outline; no content to port.]

77.5 Canonical quantization rules

77.5.1 From the Poisson bracket to the commutator

[Heading reserved in the source outline, together with its subheadings on conjugate coordinates and conserved quantities; the rule mapping classical Poisson brackets to quantum commutators, and with it the canonical commutation relations between conjugate coordinates, is pending port from a completed source.]

Derivation pending #78

From Poisson brackets to commutators: statement of the canonical quantization rule and derivation of the canonical commutation relations between conjugate coordinates, which are used throughout the following chapters.

77.5.2 Linear momentum as a differential operator

In the coordinate representation, the components of the linear momentum act on wave functions as the differential operators

$$p_i = -\mathrm{i}\hbar\,\partial_i \,. \tag{77.7}$$

Derivation pending #79

Justification of the coordinate-representation rule $p_i = -\mathrm{i}\hbar\,\partial_i$ from the canonical commutation relations; the source asserts the rule without derivation.

77.6 The Heisenberg uncertainty principle

77.6.1 Mean value of an observable

Definition 77.7 (Mean value)**.** Consider an observable A. We define the *mean* (or *expected*) *value* of the operator in the state $|\psi\rangle$ as

$$\langle A\rangle_\psi = \langle\psi|\,A\,|\psi\rangle\ . \tag{77.8}$$

77.6.2 Uncertainty

Definition 77.8 (Uncertainty)**.** In the same context as above, we define the *uncertainty* of the observable A in the state $|\psi\rangle$ as

$$\Delta A = \sqrt{\left\langle \Big(A - \langle A\rangle_\psi\,\mathbb{1}\Big)^2 \right\rangle_\psi}\ . \tag{77.9}$$

Let us observe the following. We have

$$\begin{aligned}
(\Delta A)^2 &= \left\langle \Big(A - \langle A\rangle_\psi\,\mathbb{1}\Big)^2 \right\rangle_\psi \\
&= \left\langle \Big(A - \langle A\rangle_\psi\,\mathbb{1}\Big)\Big(A - \langle A\rangle_\psi\,\mathbb{1}\Big) \right\rangle_\psi \\
&= \left\langle A^2 + \langle A\rangle_\psi^2\,\mathbb{1} - 2\,\langle A\rangle_\psi\,A \right\rangle_\psi \\
&= \left\langle \psi\middle| \Big(A^2 + \langle A\rangle_\psi^2\,\mathbb{1} - 2\,\langle A\rangle_\psi\,A\Big)\middle|\psi\right\rangle \\
&= \left\langle \psi\middle| A^2 \middle|\psi\right\rangle + \langle A\rangle_\psi^2\,\langle\psi|\psi\rangle - 2\,\langle A\rangle_\psi\,\langle\psi|A|\psi\rangle \\
&= \left\langle A^2\right\rangle_\psi + \langle A\rangle_\psi^2 - 2\,\langle A\rangle_\psi^2\ ,
\end{aligned}$$

where we used the normalization Equation (77.1). Hence the uncertainty of A is given by

$$\Delta A = \sqrt{\langle A^2\rangle_\psi - \langle A\rangle_\psi^2}\,. \tag{77.10}$$

77.6.3 The generalized uncertainty principle

[Heading reserved in the source outline.]

Derivation pending #80

Statement and derivation of the generalized uncertainty relation bounding the product of the uncertainties of two non-commuting observables by the mean value of their commutator.

Chapter 78

Elementary Quantum Systems

The postulates of Chapter 77 earn their keep on the small family of potentials for which the Schrödinger equation can be solved in closed form: the free particle, the square wells, the delta potential, the barrier, the harmonic oscillator and the periodic lattice. These systems are not toys. Between them they contain the three structural surprises of quantum mechanics—discrete bound states, zero-point energy, and the penetration of classically forbidden regions—and each surprise is attached below to the experiment that observed it, from the systematics of alpha decay [600] to the scanning tunnelling microscope [1682], which turned the most counterintuitive of the three into a routine laboratory instrument.

The operator method for the harmonic oscillator [646] is developed in full, both because the oscillator is the single most reused calculation in physics—phonons in Chapter 125, photons in Chapter 122, field modes in Chapter 99—and because its ladder-operator algebra is the template for the angular-momentum algebra of Chapter 79. Coherent states, the operator method's classical limit, close the chapter where Schrödinger himself left them [556].

Derivation pending #81

Elementary Quantum Systems: all derivations of this chapter are pending.

78.1 The free particle

78.1.1 Plane waves and the continuum

[Reserved: the free Schrödinger equation, its plane-wave solutions $\mathrm{e}^{\mathrm{i}(kx-\omega t)}$ *with* $\hbar\omega = \hbar^2k^2/2m$*, the quadratic dispersion relation, and the fact that the spectrum is purely continuous with no normalizable eigenstates—the rigged-Hilbert-space setting of Chapter 12. The relation* $p = \hbar k$ *that the dispersion encodes carries the matter-wave evidence of Chapter 71.]*

78.1.2 Wave packets, group velocity and spreading

[Reserved: Gaussian wave packets, phase versus group velocity, and the free spreading of a packet first computed by Darwin [570]: at large times the width grows linearly, at a rate fixed by the initial momentum spread. Minimum-uncertainty packets saturate the Heisenberg inequality of Chapter 77.]

78.1.3 Ehrenfest's theorem

[Reserved: Ehrenfest's theorem [573], the exact equations $\mathrm{d}\langle x\rangle/\mathrm{d}t = \langle p\rangle/m$ *and* $\mathrm{d}\langle p\rangle/\mathrm{d}t = -\langle V'(x)\rangle$*, the precise sense in which packet centroids obey Newton's law, and the failure*

of the classical correspondence when $\langle V'(x)\rangle$ differs from $V'(\langle x\rangle)$—the quantum face of the classical-limit bridge built in Chapter 23.]

78.2 Potential wells

78.2.1 The infinite square well

[Reserved: the particle in a box, the n^2 spectrum, node counting and the oscillation theorem inherited from Sturm–Liouville theory (Chapter 9), orthonormality and completeness of the eigenbasis, and the approach to the classical distribution at high quantum numbers [2313].]

Phenomenon 78.1 (Confinement alone quantizes energy)**.** An electron confined to a thin layer by barriers on both sides has a discrete set of energies whose lowest member lies *above* the bottom of the layer, and whose values grow as n^2 for $n = 1, 2, 3, \ldots$ while falling as the inverse square of the layer thickness. Dingle, Wiegmann and Henry saw the levels directly, as a staircase of steps in the optical absorption edge of gallium arsenide layers a few tens of nanometres thick grown between aluminium gallium arsenide barriers; the steps moved with layer thickness in the predicted way and their positions stood in the ratios $1 : 4 : 9 : \ldots$ [1524]. Confinement by itself, with no attractive well at all, produces both a discrete spectrum and a ground-state energy that is not zero.

Derivation. Idealize the barriers as impenetrable, so that the electron moves freely on $0 < x < L$ and ψ vanishes at both walls. Inside,

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} = E\psi\,, \qquad \text{so} \qquad \psi(x) = A\sin kx + B\cos kx\,, \quad E = \frac{\hbar^2 k^2}{2m}\,.$$

The condition $\psi(0) = 0$ removes the cosine, and $\psi(L) = 0$ forces $\sin kL = 0$, that is $kL = n\pi$ with n an integer. The value $n = 0$ gives $\psi \equiv 0$, which is not a state, and negative n repeat the positive ones up to a sign, so

$$E_n = \frac{\pi^2\hbar^2}{2mL^2}\,n^2\,, \qquad n = 1, 2, 3, \ldots \tag{78.1}$$

Three features of Equation (78.1) are the three observed ones. The spectrum is discrete because only a countable set of wavelengths fits between the walls. The ground state has $E_1 > 0$ because $n = 0$ is excluded, and E_1 grows as L^{-2}: squeezing a particle costs energy, which is the uncertainty relation of Chapter 77 in its most elementary form. And the ratios $E_n/E_1 = n^2$ do not involve m or L, which is what makes the observed staircase a signature rather than a fit. A barrier of finite height lets the wavefunction penetrate a little way into the walls, so the effective width exceeds L and every level is pushed slightly down — the correction treated in Section 78.2.2. □

78.2.2 The finite square well

[Reserved: bound states from the transcendental matching conditions, the guaranteed existence of a bound state in one dimension, penetration of the wavefunction into the classically forbidden region, and the direct spectroscopic observation of particle-in-a-box levels in thin semiconductor layers by Dingle, Wiegmann and Henry [1524]—the quantum well, ancestor of the heterostructures of Chapter 127.]

78.2.3 The delta potential

[Reserved: the attractive delta well and its single bound state, the jump condition on the derivative, scattering states and the transmission coefficient, the delta function as the zero-range limit of a deep narrow well [2313]; the double delta well as the minimal model of the covalent bond, taken up for real molecules in Chapter 88.]

78.2.4 The linear potential: the quantum bouncer

[Reserved: the linear potential and its Airy-function solutions (Chapter 9), the spectrum set by the Airy zeros, and the observation of gravitationally bound quantum states of ultracold neutrons bouncing on a mirror by Nesvizhevsky et al. [2034], with level spacings of order $1\,\mathrm{peV}$ *—the Schrödinger equation solved by Nature in the Earth's gravitational field.]*

Phenomenon 78.2 (Gravitationally bound states of the neutron)**.** Ultracold neutrons falling onto a horizontal mirror do not settle onto it continuously. Passing a beam through the gap between the mirror and an absorbing ceiling placed a variable height above it, Nesvizhevsky and collaborators found that the transmitted rate does not grow smoothly from zero with the gap: it stays consistent with zero until the gap exceeds roughly $15\,\mu\mathrm{m}$, and only then grows [2034]. The vertical motion of a neutron in the Earth's field above a hard floor is quantized, with a ground state of energy of order a picoelectronvolt and a classical turning height of order ten micrometres — the Schrödinger equation solved by Nature in the field of Chapter 19.

Derivation. Above the mirror the potential is $V(z) = mgz$ for $z > 0$, and the mirror is a hard wall, $\psi(0) = 0$. The stationary equation

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}z^2} + mgz\,\psi = E\psi$$

rearranges to $\mathrm{d}^2\psi/\mathrm{d}z^2 = (2m^2g/\hbar^2)\,(z - E/mg)\,\psi$, which the substitution

$$u := \frac{z - E/mg}{z_0}\,, \qquad z_0 := \left(\frac{\hbar^2}{2m^2 g}\right)^{1/3} \tag{78.2}$$

turns into the Airy equation $\mathrm{d}^2\psi/\mathrm{d}u^2 = u\psi$, since $(2m^2g/\hbar^2)\,z_0 = z_0^{-2}$ and $\mathrm{d}^2/\mathrm{d}z^2 = z_0^{-2}\mathrm{d}^2/\mathrm{d}u^2$. Of the two solutions of that equation only $\mathrm{Ai}(u)$ decays as $u \to +\infty$, so $\psi(z) \propto \mathrm{Ai}\,(u(z))$, and the wall condition $\psi(0) = 0$ requires $u(0) = -E/(mgz_0)$ to be a zero of Ai. Writing those zeros as $-\lambda_n$, with $\lambda_1 = 2.3381$ and $\lambda_2 = 4.0879$,

$$E_n = mg\,z_0\,\lambda_n\,. \tag{78.3}$$

The spectrum is discrete and unequally spaced, and its whole scale is carried by z_0. For the neutron, $m = 1.675 \times 10^{-27}\,\mathrm{kg}$ and $g = 9.81\,\mathrm{m/s^2}$ give $z_0 = 5.87\,\mu\mathrm{m}$ and $mgz_0 = 0.60\,\mathrm{peV}$, hence $E_1 = 1.4\,\mathrm{peV}$; the classical turning height of the ground state, E_1/mg, is $13.7\,\mu\mathrm{m}$, which is the ceiling height below which the apparatus transmits nothing. □

78.3 Tunnelling

78.3.1 Steps, barriers and the tunnel effect

[Reserved: reflection and transmission at a potential step, scattering by a rectangular barrier, the exact transmission coefficient and its exponential suppression in the opaque limit, transmission resonances at which the barrier turns transparent, and their observed counterpart, the Ramsauer–Townsend transparency of noble-gas atoms to slow electrons [480]. The semiclassical Gamow factor is derived by the WKB method of Chapter 82.]

Phenomenon 78.3 (Ramsauer–Townsend transparency)**.** Argon, krypton and xenon become nearly transparent to slow electrons. As the electron energy is lowered through a few electronvolts the measured elastic cross-section does not rise, as a classical picture of a longer interaction time demands, but passes through a deep minimum at an energy below about $1\,\mathrm{eV}$, where it falls by more than an order of magnitude below its value a few

electronvolts higher [480]. The lighter noble gases helium and neon show no such minimum. A target can therefore be more transparent to a slower projectile, which no picture of a particle following a trajectory through a force field allows.

Derivation pending #82

Resonant transmission by an attractive well: the condition under which the amplitudes reflected from its two edges cancel, making it perfectly transparent, and the three-dimensional form of the same statement as an s-wave phase shift passing through a half turn at the Ramsauer minimum

78.3.2 Alpha decay

[Reserved: the Geiger–Nuttall systematics relating alpha-decay half-life to particle energy [372]; its explanation as tunnelling through the Coulomb barrier by Gamow [600] and independently by Gurney and Condon [603, 625]— the first quantitative success of quantum mechanics in the nucleus, reproducing half-lives spanning more than twenty orders of magnitude with one exponential formula. The nuclear setting is Chapter 107; the same barrier penetration, run in reverse, sets the thermonuclear reaction rates of Chapter 52 [622].]

Phenomenon 78.4 (The Geiger–Nuttall systematics of alpha decay)**.** The alpha-emitting members of one radioactive series obey a single relation between the decay constant λ and the range — hence the energy — of the emitted particle: $\log \lambda$ is a linear function of $\log R$, and equivalently of $E^{-1/2}$ [372]. The relation is extraordinarily steep: across the naturally occurring emitters a factor of about two in alpha energy carries the half-life over more than twenty orders of magnitude, from microseconds to times comparable with the age of the Earth. Worse for classical mechanics, the alpha particle leaves with less energy than the top of the Coulomb barrier it must cross — a height that the scattering of alpha particles by the very same nuclei measures independently. There is no classical path out of the nucleus, and the nucleus decays all the same [600].

Derivation pending #83

The Gamow factor: semiclassical penetration of the Coulomb barrier by a particle whose energy lies below its top, the exponential of the integrated imaginary momentum across the forbidden region, and the linear dependence of the logarithm of the decay constant on the inverse square root of the alpha energy that follows from it

78.3.3 Field emission and the scanning tunnelling microscope

[Reserved: cold emission of electrons from metals in strong fields and the Fowler–Nordheim current–field law [599]; the scanning tunnelling microscope of Binnig and Rohrer [1682], whose exponential sensitivity of the tunnel current to a tip–surface gap of a few tenths of a nanometre resolves single atoms—tunnelling made visible; and the Josephson effect [1231], pair tunnelling between superconductors, developed with Chapter 130 and measured in Chapter 131.]

Phenomenon 78.5 (The tunnel current and its exponential gap law)**.** A current flows between a metal tip and a metal surface separated by a vacuum gap that no classical electron can cross, and it falls exponentially with the width of the gap — by about an order of magnitude for every 0.1 nm. Binnig and Rohrer measured that dependence and then

used it: holding the current constant while the tip is scanned holds the gap constant to a few picometres, which is what makes individual surface atoms visible [1682]. The most counterintuitive prediction of the theory is thereby also the most sensitive length gauge in surface physics.

Derivation. Treat the vacuum gap as a barrier whose top lies a distance ϕ — the work function — above the energy of a tip electron at the Fermi level. Inside the gap the Schrödinger equation reads

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}z^2} = \kappa^2\psi\,, \qquad \kappa = \frac{\sqrt{2m_\mathrm{e}\phi}}{\hbar}\,,$$

so that the solution which stays bounded as the gap widens is $\psi(z) \propto \mathrm{e}^{-\kappa z}$: in a classically forbidden region the amplitude does not oscillate, it decays. Matching this decaying solution to propagating waves on the two metal sides multiplies it by factors that depend on ϕ and on the energy, but not exponentially on the width, so the whole width dependence of the transmitted probability — and with it of the current — is that of $|\psi(d)|^2$:

$$I(d) \propto \mathrm{e}^{-2\kappa d}\,. \tag{78.4}$$

The rate is fixed by the work function alone. For a clean metal $\phi \approx 4\,\mathrm{eV}$, whence $\kappa \approx 1.0 \times 10^{10}/\mathrm{m}$ and $2\kappa \times 0.1\,\mathrm{nm} \approx 2$: the current changes by a factor $\mathrm{e}^{-2} \approx 0.14$ for each $0.1\,\mathrm{nm}$ of retraction, which is the observed decade per angstrom. Inverting the statement gives the instrument, since a current held to $1\,\%$ holds d to about $0.5\,\mathrm{pm}$. □

78.4 The harmonic oscillator

78.4.1 Analytic solution

[Reserved: the Hermite-function solution, the equally spaced spectrum $E_n = \hbar\omega(n + 1/2)$, obtained in wave mechanics in Schrödinger's second communication [558] and in matrix mechanics by Born, Heisenberg and Jordan [534]; the Hermite equation as a Sturm–Liouville problem (Chapter 9).]

78.4.2 Ladder operators

[Reserved: factorization of the Hamiltonian by creation and annihilation operators with $\left[a, a^\dagger\right] = \mathbb{1}$, the algebraic derivation of the spectrum in Dirac's manner [646], number states, and the matrix elements that feed every application from the phonons of Chapter 125 to the quantized field modes of Chapter 99.]

78.4.3 Zero-point energy

[Reserved: the ground-state energy $\hbar\omega/2$ as a theorem of the uncertainty relation; its experimental reality in the isotope band spectra of boron monoxide, where Mulliken found the half-quantum offset demanded by the data [515]; and its descendants, the zero-point motion that keeps helium liquid at absolute zero (Chapter 130) and the vacuum energy of Chapter 99.]

Phenomenon 78.6 (Zero-point energy)**.** A bound oscillator does not come to rest at the bottom of its well. The vibrational levels of a diatomic molecule are

$$E_v = \hbar\omega\left(v + \tfrac{1}{2}\right)\,, \qquad v = 0, 1, 2, \ldots\,, \tag{78.5}$$

and the half-quantum in Equation (78.5) is observable rather than a choice of origin. Two isotopic forms of one molecule share a force constant and differ only in reduced mass, so their ω stand in a known ratio; the whole vibrational ladder of one is therefore displaced relative to the other by an amount that depends on whether the ladder starts at $\hbar\omega/2$ or at zero. Measuring that isotope displacement in the band spectrum of boron monoxide, Mulliken found the offset demanded by the data to be the half-quantum [515].

Derivation. The ground-state energy follows from the uncertainty relation alone. For $\mathcal{H} = p^2/2m + \frac{1}{2}m\omega^2x^2$ the potential is even, so the ground state may be taken with $\langle x\rangle = \langle p\rangle = 0$; then $\langle x^2\rangle = (\Delta x)^2$ and $\langle p^2\rangle = (\Delta p)^2$, and writing $a := (\Delta x)^2$ and using $\Delta x\,\Delta p \geq \hbar/2$,

$$E = \frac{(\Delta p)^2}{2m} + \frac{m\omega^2}{2}\,a \geq \frac{\hbar^2}{8ma} + \frac{m\omega^2}{2}\,a\,.$$

The right-hand side is a function of the single positive variable a which diverges at both ends of its range; its derivative $-\hbar^2/(8ma^2) + m\omega^2/2$ vanishes at $a = \hbar/(2m\omega)$, where the two terms are equal, so

$$E \geq \frac{\hbar\omega}{4} + \frac{\hbar\omega}{4} = \frac{\hbar\omega}{2}\,. \tag{78.6}$$

The bound is attained, so it is the ground-state energy and not merely a limit on it. The Gaussian $\psi_0(x) \propto \mathrm{e}^{-m\omega x^2/2\hbar}$ has exactly $\langle x^2\rangle = \hbar/(2m\omega)$ and saturates the uncertainty relation, and substituting it into the eigenvalue equation confirms this: the coefficient of x^2 cancels for precisely that exponent, and the constant left over is $(\hbar^2/2m)\,(m\omega/\hbar) = \hbar\omega/2$, which is Equation (78.5) at $v = 0$. Nothing in the argument used the excited states or the ladder algebra. A nonzero ground-state energy is forced by the impossibility of a particle being at once at rest and at a definite place, and it is therefore as general as the uncertainty relation itself. □

78.4.4 Coherent states

[Reserved: eigenstates of the annihilation operator, packets of minimum uncertainty that oscillate without spreading, constructed by Schrödinger in 1926 as the classical limit of the oscillator [556] and rediscovered by Glauber as the states of the coherent radiation field [1255]; overcompleteness; the working states of the laser physics of Chapter 122.]

78.5 Periodic potentials: a preview

[Reserved: Bloch's theorem for a periodic potential [623]; the Kronig–Penney model of a periodic array of wells, solved exactly, with its bands of allowed energies separated by gaps [683]; the band gap as the origin of the metal–insulator distinction. The full machinery—crystal momentum, Brillouin zones, effective mass—is developed in Chapter 126 and applied in Chapter 127.]

Chapter 79

Angular Momentum and Spin

Rotational invariance is the most productive symmetry in quantum mechanics, and this chapter extracts everything it gives. The generators of rotations satisfy $[J_i, J_j] = \mathrm{i}\hbar\varepsilon_{ijk}J_k$, and that single commutator—taken alone, with no reference to any particular system—fixes the entire spectrum: eigenvalues $\hbar^2 j(j+1)$ and $\hbar m$, multiplets of $2j+1$ states, and half-integer j alongside integer j. The derivation is the ladder-operator argument already rehearsed for the oscillator in Chapter 78, carried out here in the abstract, so that the orbital angular momentum of an electron, its spin, and the total angular momentum of a nucleus are one subject and not three. The chapter then treats the orbital realisation by the spherical harmonics, the intrinsic realisation—spin—that no function on a sphere can supply, the rule for adding two angular momenta with its Clebsch–Gordan coefficients, and the Wigner–Eckart theorem, which reduces every matrix element of a rotationally covariant operator to one number times a purely geometric factor.

Spin was forced on physics by data, not deduced. The doublet structure of the alkali spectra and the anomalous Zeeman effect [277, 391] defeated every model built from orbital motion alone; Landé's empirical g factor [478] fitted the splittings with a half-integer quantum number; Pauli's "two-valuedness" [529] named the extra degree of freedom without interpreting it; and Uhlenbeck and Goudsmit [532, 562] identified it as an intrinsic angular momentum of the electron, at the price of a factor two that Thomas [561] removed by a relativistic kinematic argument. The cleanest measurement—a silver beam splitting into exactly two [492]—is Chapter 80, the experiment chapter that follows this one. Pauli's spin theory [587] is the formalism used below; Chapter 93 later obtains spin-1/2 from relativistic invariance rather than postulating it, and Chapter 83 shows that the same half-integrality decides the statistics [854]. Standard treatments are [751, 1593, 2287].

Derivation pending #84

Angular Momentum and Spin: all derivations of this chapter are pending.

79.1 The angular-momentum algebra

79.1.1 Rotations and their generators

[Reserved: rotations as a symmetry of the dynamics; the unitary representation $U(R) = \mathrm{e}^{-\mathrm{i}\theta\,\boldsymbol{n}\cdot\boldsymbol{J}/\hbar}$ guaranteed by Wigner's theorem on symmetry operations [1269] (Chapter 77); the commutator $[J_i, J_j] = \mathrm{i}\hbar\varepsilon_{ijk}J_k$ read off from the composition law of $\mathrm{SO}(3)$ [687]; $\boldsymbol{J}$ as the conserved charge of rotational invariance, the quantum counterpart of the classical theorem of Chapter 21; the Casimir $\boldsymbol{J}^2$ and the commuting pair $\{\boldsymbol{J}^2, J_z\}$ as the maximal set of simultaneously measurable rotational observables. Lie-algebraic setting: Chapter 14.]

79.1.2 The spectrum from the algebra alone

[Reserved: the raising and lowering operators $J_\pm = J_x \pm \mathrm{i}J_y$ with $[J_z, J_\pm] = \pm\hbar J_\pm$; positivity of $\boldsymbol{J}^2 - J_z^2$ terminating the ladder at both ends; the resulting eigenvalues $\hbar^2 j(j+1)$ with $j \in \left\{0, \tfrac{1}{2}, 1, \tfrac{3}{2}, \dots\right\}$ and $m = -j, \dots, j$; the matrix elements $J_\pm \,|j,m\rangle = \hbar\sqrt{j(j+1) - m(m\pm1)}\,|j, m\pm1\rangle$. The algebraic derivation is that of matrix mechanics [534], recast in the operator language of [646]; it is the same argument that fixed the oscillator spectrum in Chapter 78, and its conclusion—that half-integer j is allowed—is the opening this chapter needs for spin.]

79.1.3 Representation matrices and Euler angles

[Reserved: the $(2j+1)$-dimensional irreducible representations $D^{(j)}_{m'm}(\alpha,\beta,\gamma)$, Wigner's small d matrices and their construction [687]; unitarity, orthogonality and the composition law; the character $\chi^{(j)}(\theta) = \sin\!\left[(j+\tfrac{1}{2})\theta\right] / \sin(\theta/2)$ used below to decompose products; the Condon–Shortley phase convention [751], which is what makes published tables of coefficients comparable at all, and which this treatise adopts.]

79.1.4 SU(2), SO(3) and the double cover

[Reserved: SU(2) *as the simply connected group with the same algebra as* SO(3) *and the two-to-one homomorphism onto it; integer j as the representations of* SO(3) *proper and half-integer j as the genuinely double-valued ones, so that a 2π rotation returns a fermionic state to minus itself; the topological reason ($\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$), developed in Chapter 6 and Chapter 14. The sign is observable: neutron interferometry shows the interference pattern restored only after 4π [1558, 1566]. The relativistic completion, where the same double cover appears as* Spin(1, 3)*, is Chapter 94.]*

Phenomenon 79.1 (A neutron must be turned twice to come back)**.** Split a neutron beam coherently, precess the spin of one branch through an angle χ by sending it through a magnetic field, and recombine the two branches. The interference pattern at the exit does not repeat with period 2π in χ: it repeats with period 4π, and at $\chi = 2\pi$ — one full turn of the spin — it is inverted, maxima standing where minima stood. Rauch and collaborators in Vienna and Werner and collaborators in Michigan measured the period independently, in neutron interferometers cut from single silicon crystals, and both found 4π [1558, 1566]. A rotation by 2π is the identity on every classical object and is not the identity on a neutron.

Derivation. A rotation by χ about the unit vector $\boldsymbol{n}$ acts on a spin-$\frac{1}{2}$ state as $U = \mathrm{e}^{-\mathrm{i}\chi\,\boldsymbol{n}\cdot\boldsymbol{\sigma}/2}$. From the algebra $\sigma_i\sigma_j = \delta_{ij}\mathbb{1} + \mathrm{i}\varepsilon_{ijk}\sigma_k$ one gets $(\boldsymbol{n}\cdot\boldsymbol{\sigma})^2 = \mathbb{1}$, so the exponential series splits into even and odd parts exactly as for a scalar:

$$U(\chi) = \cos\frac{\chi}{2}\,\mathbb{1} - \mathrm{i}\sin\frac{\chi}{2}\,\boldsymbol{n}\cdot\boldsymbol{\sigma}\,. \tag{79.1}$$

Hence $U(2\pi) = -\mathbb{1}$ and $U(4\pi) = +\mathbb{1}$: the state, not merely the direction of the spin, changes sign after one turn and is restored only after two.

That sign is observable because the interferometer compares the rotated branch with an unrotated one. If each branch carries amplitude A and the geometry contributes a relative phase δ, the recombined intensity for an incident beam of spin state $|s\rangle$ is

$$I \propto \left\| \left(\mathbb{1} + \mathrm{e}^{\mathrm{i}\delta}U\right)|s\rangle \right\|^2 = 2 + 2\,\mathrm{Re}\left[\mathrm{e}^{\mathrm{i}\delta}\,\langle s|\,U\,|s\rangle\right].$$

For an unpolarized beam the spin factor is the average over any orthonormal pair, that is $\frac{1}{2}\operatorname{tr}U$, and since $\operatorname{tr}\mathbb{1} = 2$ while $\operatorname{tr}\sigma_i = 0$, Equation (79.1) gives $\frac{1}{2}\operatorname{tr}U = \cos(\chi/2)$ and

$$I \propto 1 + \cos\frac{\chi}{2}\cos\delta\,. \tag{79.2}$$

The fringe contrast is periodic in χ with period 4π and reverses sign at $\chi = 2\pi$, as observed. The same double-valuedness is why the rotation group of space has a double cover: half-integer j represents SU(2) faithfully and SO(3) only up to a sign. □

79.2 Orbital angular momentum

79.2.1 Spherical harmonics

[Reserved: $\boldsymbol{L} = \boldsymbol{r} \times \boldsymbol{p}$ and its differential form on the sphere; single-valuedness of a wavefunction of position restricting ℓ to integers; the spherical harmonics $Y_{\ell m}(\theta, \varphi)$ as the simultaneous eigenfunctions of $\boldsymbol{L}^2$ and L_z, their orthonormality, completeness and parity $(-1)^\ell$; the associated Legendre functions and the addition theorem; the connection to the Laplace equation and to Sturm–Liouville theory Chapter 9. Conventions and tables follow [751]; standard modern treatments [1593, 2287].]

79.2.2 Central potentials and the radial reduction

[Reserved: separation of the Schrödinger equation in a central potential, the effective radial equation with the centrifugal term $\hbar^2\ell(\ell+1)/2mr^2$, and the $(2\ell+1)$-fold degeneracy that rotational invariance guarantees for every central problem; why that degeneracy is exactly what an external field lifts (Chapter 82); the worked solutions—free particle in spherical waves, spherical well, isotropic oscillator—and the Coulomb case, whose extra degeneracy is the subject of Chapter 81.]

79.2.3 Rotational spectra of molecules

[Reserved: the rigid rotor, energies $E_J = \hbar^2 J(J+1)/2I$ and the equally spaced $\Delta J = \pm 1$ lines in the far infrared and microwave, the direct spectroscopic reading of the $J(J+1)$ law and hence of the algebra of Section 79.1.2 [961]; moments of inertia and bond lengths obtained from line spacings; centrifugal distortion and rotation–vibration coupling; nuclear-spin statistics of homonuclear molecules, where ortho and para intensity alternation is a direct measurement of the spin–statistics connection (Chapter 83); the classical rotor is Chapter 29, the molecular structure Chapter 88.]

Phenomenon 79.2 (The rotational line series of a molecule)**.** The pure rotational spectrum of a heteronuclear diatomic molecule, observed in the far infrared and the microwave, is a series of very nearly equally spaced lines. Two things about it are sharp. The spacing is constant, to a slow correction that grows with the rotational quantum number and is the centrifugal stretching of the bond; and the series extrapolates to zero at zero, so that the first line stands one full spacing above the origin rather than half a spacing [961]. For carbon monoxide the first line lies near 115 GHz and the following ones near two and three times that frequency. The rotational energies therefore go as $J(J+1)$ and not as J^2: the spectrum reads the eigenvalue of $\boldsymbol{J}^2$ off the sky.

Derivation. A rigid rotor has $\mathcal{H} = \boldsymbol{J}^2/2I$ with I its moment of inertia, so by the multiplet argument used for Phenomenon 79.3 its levels are $E_J = \hbar^2 J(J+1)/2I$ with J a non-negative integer, integral rather than half-integral because the rotor's angular momentum is orbital (Section 79.2.1). Electric-dipole radiation carries one unit of angular momentum, so $\Delta J = \pm 1$ (Section 79.2.4), and absorption from J to $J+1$ occurs at

$$\begin{aligned} h\nu_J = E_{J+1} - E_J &= \frac{\hbar^2}{2I}\left[(J+1)(J+2) - J(J+1)\right] \\ &= \frac{\hbar^2}{I}(J+1)\,, \end{aligned} \tag{79.3}$$

that is $\nu_J = \nu_0\,(J+1)$ with $\nu_0 = \hbar/2\pi I$. The lines are equally spaced by ν_0 and the first of them lies at ν_0, one whole spacing from the origin, as observed.

The alternative worth excluding is the naive $E_J \propto J^2$, which a reading of $\boldsymbol{J}^2$ as the square of a vector of length $\hbar J$ would give. It predicts $E_{J+1} - E_J \propto 2J+1$: lines in the ratios $1:3:5:\ldots$, still equally spaced but with the first line at *half* the spacing. The two laws are distinguished by one measurement of the lowest line against the spacing, and the measurement selects $J(J+1)$. Since ν_0 fixes I, the same spectrum then returns the bond length; and centrifugal stretching enters only as a correction of relative order $(\hbar J/I\omega_{\mathrm{vib}})^2$, which is why the leading pattern survives it. □

79.2.4 Parity and selection rules

[Reserved: parity as a discrete symmetry commuting with rotations, the parity of $Y_{\ell m}$, and the resulting electric-dipole selection rules $\Delta\ell = \pm 1$, $\Delta m = 0, \pm 1$ with Laporte's rule as their empirical form [751]; the intensity consequences observed in every atomic spectrum (Chapter 70); higher multipoles and the metastable states that survive because the dipole matrix element vanishes—the $2s$ level of hydrogen and the forbidden nebular lines. The general statement is the Wigner–Eckart theorem of Section 79.5.2; parity as a symmetry in its own right, and its violation by the weak interaction, is Chapter 105.]

79.3 Spin

79.3.1 The evidence for an intrinsic angular momentum

[Reserved: the four independent lines of evidence, each with the measurement that carries it—alkali doublets and the anomalous Zeeman effect [277, 391] with Landé's g factor [478]; Pauli's exclusion principle and its fourth quantum number [529]; the Stern–Gerlach splitting of a ground-state silver beam into exactly two components [481, 491, 492], treated in full in Chapter 80; and the gyromagnetic ratio measured mechanically by magnetization on rotation [424]. The proposal itself [532, 562], and why a literal spinning sphere is excluded by the equatorial speed it would require.]

Phenomenon 79.3 (Two beams, hence a half-integral angular momentum)**.** A collimated beam of ground-state silver atoms crossing a strongly inhomogeneous magnetic field emerges neither as one broadened trace, which is what a classical moment of arbitrary orientation would give, nor as an odd number of traces, which is what any orbital angular momentum would give, but as exactly two, placed symmetrically about the undeflected position [481, 492]. The size of the deflection gives a moment of one Bohr magneton [491]. The full apparatus, its systematics and its sequential versions are Chapter 80; what is used here is the bare count. An even number of components cannot come from orbital motion, so the silver atom carries an angular momentum that is not orbital and whose projection quantum number is half-integral.

Derivation. An atom of magnetic moment $\boldsymbol{\mu}$ in a field $\boldsymbol{B}$ has energy $-\boldsymbol{\mu}\cdot\boldsymbol{B}$ and therefore feels the force $\boldsymbol{F} = \nabla\,(\boldsymbol{\mu}\cdot\boldsymbol{B})$; with the field along z and varying along z, this is $F_z = \mu_z\,\partial B_z/\partial z$. The beam is split into one trace for each distinct value of μ_z, so counting traces counts the spectrum of μ_z, and since $\boldsymbol{\mu}$ is proportional to $\boldsymbol{J}$, the spectrum of J_z.

That spectrum follows from the commutator $[J_i, J_j] = \mathrm{i}\hbar\varepsilon_{ijk}J_k$ and nothing else. Let $|\lambda, m\rangle$ satisfy $\boldsymbol{J}^2\,|\lambda,m\rangle = \hbar^2\lambda\,|\lambda,m\rangle$ and $J_z\,|\lambda,m\rangle = \hbar m\,|\lambda,m\rangle$, and put $J_\pm = J_x \pm \mathrm{i}J_y$. Then $[J_z, J_\pm] = \pm\hbar J_\pm$, so $J_\pm$ carries an eigenvector of J_z to one with m raised or lowered by one, while $[\boldsymbol{J}^2, J_\pm] = 0$ leaves λ alone. From $J_\mp J_\pm = \boldsymbol{J}^2 - J_z^2 \mp \hbar J_z$,

$$\|J_\pm\,|\lambda,m\rangle\|^2 = \hbar^2\,[\lambda - m\,(m\pm 1)] \geq 0\,, \tag{79.4}$$

so m is bounded above and below and both ladders must terminate. Call the largest value j and the smallest j'. Termination means $J_+ |\lambda, j\rangle = 0$ and $J_- |\lambda, j'\rangle = 0$, which by Equation (79.4) give $\lambda = j(j+1)$ and $\lambda = j'(j'-1)$; eliminating λ leaves $j' = -j$, the root $j' = j+1$ being excluded by $j' \leq j$. The two ends are joined by a whole number of unit steps, so $2j$ is a non-negative integer and the multiplet contains

$$2j+1 \text{ states, with } m = -j, -j+1, \ldots, j\,. \tag{79.5}$$

Two traces therefore mean $2j+1 = 2$, that is $j = \frac{1}{2}$. This is the point at which the observation forces the theory's hand: Equation (79.5) permits half-integer j on purely algebraic grounds, no function of position on a sphere realizes it (Section 79.2.1), and the silver beam shows Nature using it anyway. □

79.3.2 Pauli matrices and spinors

[Reserved: the $j = \frac{1}{2}$ representation written out, $\boldsymbol{S} = \frac{\hbar}{2}\boldsymbol{\sigma}$, with the algebra $\sigma_i\sigma_j = \delta_{ij}\mathbb{1} + \mathrm{i}\varepsilon_{ijk}\sigma_k$, tracelessness, and completeness of $\{\mathbb{1}, \sigma_i\}$ on 2×2 matrices; two-component spinors, the Pauli equation with its $g = 2$ magnetic coupling [587]; the Bloch-sphere parametrization of a pure spin-1/2 state and its density matrix (Chapter 77); why the mean of $\boldsymbol{\sigma}$ behaves classically while its components never commute. Reobtained from the Dirac equation in Chapter 93.]

79.3.3 Magnetic moments and the g factor

[Reserved: the Bohr magneton $\mu_\mathrm{B} = e\hbar/2m_\mathrm{e}$ and the nuclear magneton smaller by the mass ratio; $\boldsymbol{\mu} = g\,\mu_\mathrm{B}\boldsymbol{J}/\hbar$ with $g_\ell = 1$ for orbital motion and $g_s \approx 2$ for spin [587]; the anomaly $g_s - 2$, first measured by Kusch and Foley from the ratio of atomic g values [916], now the most precisely tested prediction in physics (Chapter 112, with the radiative theory in Chapter 100); nuclear moments in nuclear magnetons (Chapter 107); the magnetism of matter as the bulk consequence (Chapter 129).]

Phenomenon 79.4 (The anomalous Zeeman effect)**.** In a weak magnetic field a spectral line of an atom splits into several components, equally spaced in frequency [277]. The pattern that a purely orbital moment predicts — a symmetric triplet whose splitting unit is $\mu_\mathrm{B}B/h$ and is the same for every line of every atom — is the exception, not the rule. Most lines split into more components than three, with a unit peculiar to the levels involved and in general not an integer multiple of $\mu_\mathrm{B}B/h$; the sodium D lines, the classic case, split into four and six [391]. Landé found that the whole body of data is reproduced by attaching to each level a single number g, computable from its three quantum numbers, and that fitting the observed patterns requires those quantum numbers to take half-integer values [478].

Derivation pending #85

Landé's g factor from the projection theorem: that the matrix elements of any vector operator inside one multiplet are proportional to those of the total angular momentum, so that the observed splitting unit becomes a closed formula in the three quantum numbers of the level, and that fitting it to the data requires the electron's own gyromagnetic factor to be twice the orbital one

79.3.4 Thomas precession

[Reserved: the spin–orbit coupling computed naively in the electron rest frame is twice the observed one; Thomas' resolution [561]—the rest frame is not inertial, and the composition of

two non-collinear boosts leaves a rotation, so the spin precesses at $\boldsymbol{\omega}_{\mathrm{T}}$ and halves the coupling in the low-velocity limit. The kinematics is that of Chapter 39, where the Wigner rotation is derived; the numerical consequence is the observed fine structure of Chapter 81, and the exact relativistic statement is the Bargmann–Michel–Telegdi equation for the polarization of a particle in an external field [1165], which is what the $g-2$ storage rings of Chapter 112 measure against. This is the worked case where a factor two decided between a right and a wrong theory.]

79.3.5 Spin beyond one half

[Reserved: spin-1 and the photon, whose two observed polarizations rather than three follow from masslessness (Chapter 94) and whose angular momentum was measured mechanically as a torque on a birefringent plate [767]; spin-1 massive bosons (Chapter 104); spin-3/2 and spin-2 states in nuclei and in the hadron spectrum [2427]; the spin-2 graviton polarizations of Chapter 46. What the evidence base does and does not contain about higher spins is Chapter 97.]

Phenomenon 79.5 (Light carries spin angular momentum)**.** A half-wave plate hung on a fine quartz fibre and traversed along its axis by circularly polarized light feels a steady torque about that axis, and the torque reverses when the handedness of the light is reversed. Beth built the apparatus, with the beam passed twice through the suspended plate to double the effect and a resonant detection scheme to pull the deflection out of the Brownian motion of the fibre, and measured a torque per unit incident power equal to $2/\omega$ [767]. Angular momentum is carried by a beam of light independently of any orbital motion of anything, and the amount per photon is one unit of $\hbar$.

Derivation. A half-wave plate converts circularly polarized light of one handedness into the other and transmits it. Suppose each photon of angular frequency ω carries a spin projection $s\hbar$ along its direction of propagation. Each transmitted photon has that projection reversed, so it gives up $2s\hbar$ to the plate, and a beam of power P delivers $P/\hbar\omega$ photons per second. The torque is therefore

$$\tau = 2s\hbar \times \frac{P}{\hbar\omega} = \frac{2sP}{\omega}\,. \tag{79.6}$$

Two things are worth noting about Equation (79.6). First, $\hbar$ cancels: the predicted torque is a classical quantity, which is why a torsion fibre can measure it, and the experiment therefore determines the integer s without needing to know $\hbar$. Second, the measured value $\tau = 2P/\omega$ gives $s = 1$. The electromagnetic field carries spin one, a fact read here off a suspended crystal and recovered from the transformation properties of the field in Chapter 94; that the photon nevertheless shows two polarization states rather than the three of a massive spin-1 particle is a consequence of its masslessness, established there. □

79.4 Addition of angular momenta

79.4.1 The coupling problem and the Clebsch–Gordan series

[Reserved: two commuting angular momenta $\boldsymbol{J}_1$, $\boldsymbol{J}_2$ and their sum; the uncoupled basis $|j_1 m_1\rangle\,|j_2 m_2\rangle$ against the coupled basis $|j_1 j_2; JM\rangle$; the decomposition $j_1 \otimes j_2 = |j_1 - j_2| \oplus \cdots \oplus (j_1 + j_2)$, proved by counting states or by multiplying characters [687]; dimensional check $(2j_1+1)(2j_2+1) = \sum_J (2J+1)$; why the coupled basis is the useful one whenever the interaction is rotationally invariant but the individual momenta are not separately conserved—spin–orbit coupling, LS and jj coupling schemes in atoms (Chapter 88).]

79.4.2 Clebsch–Gordan coefficients

[Reserved: the coefficients $\langle j_1 m_1 j_2 m_2 | J M\rangle$ as the unitary matrix between the two bases; their construction by repeated application of J_- to the stretched state, the selection rules $M = m_1 + m_2$ and the triangle inequality, reality in the Condon–Shortley convention, orthogonality and recursion relations [687, 751]; the closed Racah form of the general coefficient [867]; the tabulated special cases $\frac{1}{2}\otimes\frac{1}{2}$ and $\ell\otimes\frac{1}{2}$ that every atomic calculation reuses.]

79.4.3 Two spins one half: singlet and triplet

[Reserved: the explicit decomposition $\frac{1}{2}\otimes\frac{1}{2} = 0\oplus 1$, the antisymmetric singlet and the symmetric triplet, and the exchange operator written as $\frac{1}{2}(\mathbb{1} + \boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_2)$; the singlet–triplet splitting observed as ortho- and para-helium (Chapter 88) and as the hydrogen hyperfine splitting of Chapter 81; the exchange interaction behind ferromagnetic order (Chapter 129); and the singlet as the state whose correlations violate the Bell inequalities (Chapters 84 and 89) [1271].]

79.4.4 $3j$, $6j$ and $9j$ symbols

[Reserved: Wigner's $3j$ symbol as the symmetrized Clebsch–Gordan coefficient, with the full permutation symmetry that the coefficient hides [687]; the recoupling of three angular momenta and the $6j$ symbol relating the two coupling orders, of four and the $9j$ symbol [867, 873]; sum rules and orthogonality relations; why this apparatus is not bookkeeping but the practical content of atomic and nuclear structure calculations (Chapters 88 and 107), and the same recoupling algebra reappears for isospin.]

79.5 Tensor operators and the Wigner–Eckart theorem

79.5.1 Spherical tensor operators

[Reserved: the definition of an irreducible tensor operator $T^{(k)}_q$ by its commutators with $\boldsymbol{J}$, equivalently by its transformation under $U(R)$ with the matrices $D^{(k)}$; scalars, vectors and rank-two tensors as the cases $k = 0, 1, 2$; the decomposition of a Cartesian tensor into irreducible parts—trace, antisymmetric part, symmetric traceless part—and its physical reading as monopole, vector and quadrupole [751, 867]; products of tensor operators coupled with Clebsch–Gordan coefficients.]

79.5.2 The Wigner–Eckart theorem

[Reserved: the theorem $\langle j' m'|\, T^{(k)}_q\,|jm\rangle = \langle jm\, kq | j' m'\rangle\ \langle j'\|T^{(k)}\|j\rangle\,/\sqrt{2j+1}$, separating a geometric factor fixed entirely by rotational invariance from one reduced matrix element carrying all the dynamics, stated by Eckart [648] and by Wigner [687]; the immediate corollaries—selection rules $|j-k|\le j'\le j+k$ and $m' = m+q$, the proportionality of the matrix elements of any vector operator within a multiplet to those of $\boldsymbol{J}$ (the projection theorem), and hence the derivation of Landé's empirical g factor [478] from first principles.]

79.5.3 What the theorem buys

[Reserved: worked consequences, each attached to a measurement— intensity ratios within a spectral multiplet, fixed by geometry alone and independent of the atomic model [751]; the vanishing of a static electric dipole moment for a state of definite parity, and the searches for the electron electric dipole moment as tests of Chapter 105; nuclear electric quadrupole

moments requiring $I \geq 1$, so that a spin-1/2 nucleus has none (Chapter 107); multipole expansions of radiation and their selection rules (Chapter 65).]

79.6 Spin dynamics and magnetic resonance

79.6.1 Larmor precession

[Reserved: the Hamiltonian $-\boldsymbol{\mu}\cdot\boldsymbol{B}$, the exact solution for a spin in a static field, and the precession of $\langle \boldsymbol{S}\rangle$ at $\omega_{\mathrm{L}} = g\mu_{\mathrm{B}}B/\hbar$; Ehrenfest's theorem (Chapter 78) making the quantum mean obey the classical torque equation exactly for this system, one of the few cases where it does so with no correction whatever; the Zeeman splitting as the energy form of the same statement [277]; the observed precession of a free neutron's spin measured interferometrically [1566]; the muon precession that the $g-2$ experiments of Chapter 112 count.]

Phenomenon 79.6 (Magnetic resonance)**.** A sample of matter, or a beam of atoms or molecules, placed in a static magnetic field B_0 and irradiated with a weak transverse field oscillating at frequency ν, exchanges energy with the radiation only in a narrow band of frequencies, centred on a value strictly proportional to B_0 and independent of the intensity of the radiation. Rabi's molecular beams turned the resonance into the first precise measurements of nuclear magnetic moments [819]; Purcell, Torrey and Pound found the same absorption in a block of paraffin and Bloch, Hansen and Packard found it by nuclear induction in water, both in 1946 [887, 894]. The proportionality constant is a property of the species and of nothing else, which is what makes the resonance an instrument.

Derivation. Write the moment as $\boldsymbol{\mu} = \gamma \boldsymbol{S}$, so that the Hamiltonian in the static field is $\mathcal{H} = -\gamma\, \boldsymbol{S}\cdot\boldsymbol{B}$. For any observable the Ehrenfest relation of Chapter 78 gives $\mathrm{d}\langle A\rangle/\mathrm{d}t = (\mathrm{i}/\hbar)\langle[\mathcal{H}, A]\rangle$, and here the commutator can be evaluated in closed form from the algebra alone:

$$[\mathcal{H}, S_i] = -\gamma B_j\,[S_j, S_i] = -\mathrm{i}\hbar\gamma\,\varepsilon_{jik}B_j S_k\,,$$

so that, with $\varepsilon_{jik}B_j S_k = -(\boldsymbol{B}\times\boldsymbol{S})_i$,

$$\frac{\mathrm{d}\langle \boldsymbol{S}\rangle}{\mathrm{d}t} = \gamma\,\langle \boldsymbol{S}\rangle \times \boldsymbol{B}\,. \tag{79.7}$$

Equation (79.7) is exact — no expansion, no truncation — and it is the classical equation for a top of angular momentum $\langle \boldsymbol{S}\rangle$ carrying the moment $\gamma\langle \boldsymbol{S}\rangle$. Its solution is a precession about $\boldsymbol{B}$ at the angular frequency

$$\omega_{\mathrm{L}} = |\gamma|\,B_0\,, \tag{79.8}$$

with the component along $\boldsymbol{B}$ constant. A transverse driving field is therefore in step with the moment through every cycle when, and only when, it turns at ω_{L}, which is the observed resonance.

The energy form of the same statement is the sharper one. The eigenvalues of $\mathcal{H}$ are $-\gamma\hbar m B_0$ with $m = -j, \ldots, j$ by Equation (79.5), so adjacent levels are separated by $\hbar\,|\gamma|\,B_0$ and a photon is absorbed when $h\nu = \hbar\,|\gamma|\,B_0$, that is when $2\pi\nu = \omega_{\mathrm{L}}$. Because the levels are equally spaced the whole multiplet resonates at one frequency, which is why the absorption line is single and why measuring it measures γ alone; the observed resonance frequency divided by B_0 is the gyromagnetic ratio, with no model of the nucleus intervening. □

79.6.2 Rabi's resonance method

[Reserved: a rotating transverse field added to a static one; the exact two-level solution and the Rabi formula for the transition probability, with its resonance at the Larmor frequency and its generalized frequency off resonance [798]; the molecular-beam magnetic-resonance

method and the first precise nuclear moments [819]; Ramsey's method of separated oscillatory fields [968] as the narrowing that made the caesium frequency standard possible, feeding the SI definition of the second (Chapter 2); the perturbative version of the same result appears as the golden rule in Chapter 82, and the exact solution here is the check on it.]

79.6.3 Nuclear magnetic resonance

[Reserved: resonance in bulk matter rather than a beam, detected by absorption [894] and by induction [887]; relaxation times T_1 and T_2 and the Bloch equations, the open-system description of which is Chapter 87; the chemical shift and its use in molecular structure (Chapter 88); spatial encoding by field gradients and the resulting imaging [1507]—an angular-momentum commutator turned into a clinical instrument; electron spin resonance as the analogue at the electron magneton.]

79.6.4 Spin as the elementary two-state system

[Reserved: spin-$1/2$ as the minimal quantum system and the template for every two-level problem—ammonia inversion, neutral-kaon and neutrino flavour oscillation (Chapters 103 and 111), the qubit; observables that do not commute, so no state is dispersion-free in all three components, and the resulting spin-projection statistics that the Stern–Gerlach chapter measures directly (Chapter 80); the singlet correlations of Chapter 84 as the sharpest statement that these statistics are not classical [1271].]

Chapter 80

Experiment: Stern–Gerlach

Few experiments in the history of physics have delivered so unambiguous a verdict with so modest an apparatus. In the winter of 1921–1922, in a cramped laboratory of the Institute for Theoretical Physics in Frankfurt am Main, Otto Stern and Walther Gerlach sent a beam of neutral silver atoms through a strongly inhomogeneous magnetic field and found that the beam split cleanly into two — not into the continuous fan that every classical intuition demanded. The experiment made *space quantization* visible on a glass plate: the projection of an atomic magnetic moment onto a field axis takes only discrete values. That the discreteness was at first attributed to the wrong angular momentum — orbital rather than spin — takes nothing away from the observation itself, which stands to this day as the prototype of a quantum measurement. This chapter presents the experiment in full phenomenological detail; the quantum-mechanical derivations it calls for are developed in Chapter 79 and are marked pending here.

80.1 Historical context, 1921–1922

By 1921 the old quantum theory was an uneasy patchwork of classical mechanics and quantization rules. Bohr's planetary model of 1913 had been generalized by Sommerfeld, who quantized not only the size and shape of electron orbits but also their *orientation*: in the presence of a magnetic field defining an axis, the component of the orbital angular momentum along that axis was restricted to a discrete set of values, a doctrine Sommerfeld called *Richtungsquantelung* — direction quantization, or space quantization [441]. The rule accounted economically for the normal Zeeman splitting of spectral lines, but it carried an outrageous corollary: an atom entering a magnetic field must instantaneously find its angular momentum pointing only along the quantized directions, however the atom was oriented beforehand. To most physicists trained in the classical tradition this seemed a bookkeeping fiction, useful for spectroscopy but surely not a statement about the actual disposition of atoms in space.

Otto Stern, then at Frankfurt, was among the sceptics — and he possessed the one instrument capable of settling the matter. In the molecular-beam method he had been developing since 1919, atoms fly through high vacuum in nearly collision-free straight lines, so that a force acting on a single atom is recorded directly as a deflection of the beam. In 1921 Stern published the proposal in full: pass a beam of neutral atoms through an inhomogeneous magnetic field; if the atomic magnetic moments are oriented at random, as classical mechanics holds, the beam broadens into a continuous smear, whereas if space quantization is real the beam must split into discrete components [481]. The experiment was conceived as a decisive test between two sharply distinct predictions — and, by Stern's own account, he expected and hoped to refute the quantum prediction. The documented recollections collected by Friedrich and Herschbach preserve the flavour of the wager: Stern

is reported to have vowed to quit physics if Bohr's "nonsense" of spatial quantization turned out to be correct, and even sympathetic colleagues regarded the quantized orientations as a computational timetable rather than physical reality [2050]. It should be remembered that both sides of the wager were arguing within the old quantum theory: the modern resolution — electron spin — was not on anyone's table.

The circumstances were as unfavourable as the idea was sharp. Postwar Germany was sliding into inflation, and the Frankfurt institute had almost no money for equipment. The experiment survived on improvised patronage: Max Born, Stern's chief, raised funds by charging admission to popular lectures on relativity; private donations arrived from abroad; and a grant arranged through the committee of Einstein's Kaiser Wilhelm Institute for Physics — which Einstein then directed, largely as a fund-dispensing body — helped pay for the electromagnet [2050]. Gerlach, an accomplished experimentalist who had joined Frankfurt in 1920, carried the brunt of the laboratory work through repeated failures of vacuum, alignment, and detection, with Stern contributing the crucial design insight that the field gradient, not the field strength, produces the deflecting force.

One further story belongs here, explicitly flagged as *anecdote*. Stern recounted, decades later, that when he and Gerlach first examined the detector plate they saw nothing — until, as they peered at it, the silver trace slowly blackened before their eyes. Stern's explanation was that the sulfurous smoke of their cheap cigars (professorial salaries being too thin for good ones) had converted the invisibly thin silver deposit into visible black silver sulfide, in the manner of a photographic developer. The tale is charming and was told by Stern himself, but its literal role in the discovery is uncertain: the plates were in any case developed chemically as a matter of routine, and Friedrich and Herschbach, who traced the story and even re-enacted the chemistry, present it as a delightful embellishment rather than an established causal link in the chain of discovery [2050]. We repeat it in that spirit.

80.2 Stern–Gerlach (1921–1922)

The experiment ran in Frankfurt from 1921 to the decisive plates of February 1922. Its logic is worth stating before its hardware: a neutral atom carries no net charge, so the Lorentz force cannot deflect it; but if the atom carries a magnetic moment $\boldsymbol{\mu}$, an *inhomogeneous* field exerts a net translational force on it. The deflection of a beam of neutral atoms therefore reads off the distribution of magnetic-moment components along the field gradient — a direct image, atom by atom, of the quantity that space quantization declares to be discrete.

80.2.1 Apparatus

An electrically heated oven evaporates metallic silver at a temperature of about $1300\,\mathrm{K}$; atoms effuse from a small aperture into high vacuum, which must be maintained continuously over runs lasting several hours (the working pressures were of the order of $10^{-3}\,\mathrm{Pa}$; a single cracked gasket could destroy a night's run). Two collimating slits, a few hundredths of a millimetre wide, select from the effusing vapour a thin ribbon-shaped beam. The beam then passes along the edge of a deflecting electromagnet whose pole pieces have the now-famous asymmetric profile — one pole machined to a sharp knife edge, the facing pole carrying a channel-shaped groove — so that the field between them, of magnitude on the scale of $B \approx 0.1\,\mathrm{T}$ near the beam, varies strongly across the gap, with a vertical gradient of order

$$\frac{\partial B_z}{\partial z} \sim 10^{3}\,\mathrm{T/m}\,, \tag{80.1}$$

sustained over a beam path through the field of about $3.5\,\mathrm{cm}$. Detection is cumulative: the silver atoms condense on a cold glass plate placed beyond the magnet, and the invisibly

thin deposit built up over hours is developed photographically to render the trace visible [491, 492].

80.2.2 Procedure

The oven temperature fixes the velocity distribution of the beam: silver atoms at 1300 K leave the source with a broad thermal spread of speeds whose beam-averaged value is of order 660 m/s — a kinetic-theory estimate, its precise value depending on how the effusive flux weights the Maxwell distribution. With the magnet off, a run records the undeflected beam profile, a single narrow line reproducing the collimator geometry. With the magnet energized, the atoms traverse the gradient region and each acquires a transverse deflection determined by the component μ_z of its magnetic moment along the gradient. The classical expectation is unequivocal: magnetic moments emerging from a hot oven point in random directions, and in the field they precess about the field axis at the Larmor frequency, which preserves μ_z; the ensemble therefore presents a continuum of μ_z values from $-|\boldsymbol{\mu}|$ to $+|\boldsymbol{\mu}|$, and the recorded trace should be a single trace, broadened into a smear that is densest at the centre and fades toward maximal deflection. Any discrete multiplicity of traces is irreconcilable with a classical moment distribution.

80.2.3 Observations and data

The developed plates of February 1922 showed the beam split into *two* discrete traces, with no detectable intensity at the undeflected centre where the classical smear would peak. The pattern has the shape of a pair of lips: at the centre of the beam ribbon, where the gradient is strongest near the knife edge, the two traces reach their maximal separation of about 0.2 mm, and they close toward the ends of the ribbon where the field is more nearly uniform [492]. On 8 February 1922 Gerlach mailed a postcard to Niels Bohr carrying a photograph of the split trace and congratulations on the confirmation of his theory — an image that has become one of the icons of quantum physics [2050].

The deflection is governed by the force on a magnetic dipole in an inhomogeneous field, which for the geometry of the experiment reduces to

$$F_z = \mu_z \frac{\partial B_z}{\partial z} \,. \tag{80.2}$$

Derivation pending #86

Force on a magnetic dipole in an inhomogeneous field

Taking Equation (80.2) together with the measured geometry, gradient, and thermal velocity, the observed splitting yields the magnitude of the silver atom's magnetic moment. As an order-of-magnitude check: a moment of one Bohr magneton in the gradient Equation (80.1) feels a force $F_z \approx 9 \times 10^{-21}\,\mathrm{N}$, which acting on a silver atom (mass about $1.8 \times 10^{-25}\,\mathrm{kg}$) moving at 660 m/s over 3.5 cm produces a deflection of order a tenth of a millimetre — consistent with the observed trace separation. Gerlach and Stern's quantitative analysis, published later in 1922, concluded that the moment equals one Bohr magneton within their experimental uncertainty of about ten percent [491]; in modern units,

$$\mu_{\mathrm{B}} = \frac{e\hbar}{2m_e} = 9.274 \times 10^{-24}\,\mathrm{J/T} \tag{80.3}$$

by the current CODATA adjustment [2392].

80.2.4 Interpretation

The immediate conclusion was announced in the title of the discovery paper: "experimental proof of direction quantization in the magnetic field" [492]. The projection of the atomic magnetic moment onto the field axis is not continuously distributed; it takes exactly two values, $\mu_z = \pm\mu_\mathrm{B}$, for silver. Space quantization is a fact of nature, and Stern lost his wager.

Yet the 1922 reading of that fact was wrong in nearly every detail. Within the Bohr–Sommerfeld theory the splitting was ascribed to an orbital angular momentum of one unit, whose projection was held (on Bohr's argument) to take the values $\pm\hbar$ with the value zero forbidden — hence two traces. The modern account is entirely different: the ground state of silver is ${}^2S_{1/2}$; its single valence electron has orbital angular momentum $\ell = 0$ and hence *no* orbital magnetic moment at all. What the plates recorded is the intrinsic spin of that electron, $s = 1/2$, whose projection $S_z = \pm\hbar/2$ couples to the field through a gyromagnetic factor $g_s \approx 2$, so that $\mu_z = \mp g_s \mu_\mathrm{B}/2 \approx \mp\mu_\mathrm{B}$. The numerical coincidence $g_s s \approx 1$ made the spin-1/2 doublet mimic exactly the (doubly erroneous) old-quantum-theory prediction, and so the experiment was celebrated for years as confirming a model that could not have produced its result. Spin was proposed only in 1925 and the Stern–Gerlach pattern correctly reinterpreted in its terms by 1927 [2050, 2287]. The observation outlived its first explanation — a pattern this treatise will meet again.

80.2.5 Primary references

[481, 491, 492].

80.3 The phenomenon: space quantization

Stripped of its 1922 vocabulary, the experimental fact is the following.

Phenomenon 80.1 (Space quantization)**.** The component of an atom's angular momentum — equivalently, of its magnetic moment — measured along any axis singled out by a magnetic field takes only discrete values. For silver atoms in their ground state the measured moment component takes exactly two values, $\mu_z = \pm\mu_\mathrm{B}$ to within experimental accuracy [491, 492], whatever the prior history of the atom and however the field axis is chosen.

Derivation pending #87

Quantum derivation of the two-valued splitting from spin-1/2 (to Chapter 79)

Two features of Phenomenon 80.1 deserve emphasis, because they carry the whole weight of the quantum theory that Chapter 79 builds. First, the discreteness is *two*-fold. Any orbital angular momentum ℓ yields an odd number $2\ell+1$ of projections; an even multiplicity cannot arise from orbital motion at all and forces a half-integer angular momentum, $s = 1/2$, of no classical provenance. Second, the discreteness holds for *any* choice of axis. The magnet's orientation is at the experimenter's disposal, yet along whatever direction one measures, the answer is $\pm\hbar/2$ — never a continuum, never a projection factor $\cos\theta$. No classical vector quantity can behave this way, and the resolution — that the measured component is an operator with a two-point spectrum, not a pre-existing number — is the substance of the angular-momentum algebra developed in Chapter 79 and of the measurement postulates of Chapter 77.

80.4 Sequential Stern–Gerlach experiments

The full quantum-mechanical lesson of the apparatus emerges only when several Stern–Gerlach magnets are chained in series, with beam blocks used to select one output of each stage as the input of the next. In this form the experiment is best known as the opening thought experiment of modern quantum-mechanics pedagogy [1593, 2287] — but it is a *gedanken* experiment made real: each stage is precisely the apparatus of Section 80.2, and versions of the cascade have been realized with atomic and neutron beams. We write SG$\hat{z}$ for a Stern–Gerlach magnet whose gradient axis is $\hat{z}$, and label its two output beams by the measured spin component $S_z = \pm\hbar/2$.

Three arrangements suffice to dismantle classical expectations. In the first, an SG$\hat{z}$ stage selects the $S_z = +\hbar/2$ beam and feeds it to a second SG$\hat{z}$ stage: only the $+\hbar/2$ output fires. Measurement is *repeatable*; the first stage has prepared a state that the second confirms with certainty. In the second arrangement the selected $S_z = +\hbar/2$ beam enters an SG$\hat{x}$ stage: both outputs fire, with equal intensities. A definite value of S_z evidently does not carry with it any definite value of S_x. The third arrangement is the decisive one: the $S_x = +\hbar/2$ output of the second stage is passed to a final SG$\hat{z}$ magnet — and *both* S_z outputs fire, again with equal intensities. The beam entering the SG$\hat{x}$ stage contained, by construction, no $S_z = -\hbar/2$ component whatsoever; the intervening measurement of S_x has destroyed the previously established value of S_z. Table 80.1 summarizes the three cases.

Table 80.1: Sequential Stern–Gerlach cascades. Each stage transmits the beam indicated and blocks the other; probabilities refer to the outputs of the final stage, conditional on an atom entering it.

Cascade	Beam admitted to final stage	Final outcomes	Probabilities
SG$\hat{z} \to$ SG$\hat{z}$	$S_z = +\hbar/2$	$S_z = +\hbar/2$ only	1, 0
SG$\hat{z} \to$ SG$\hat{x}$	$S_z = +\hbar/2$	$S_x = \pm\hbar/2$	1/2, 1/2
SG$\hat{z} \to$ SG$\hat{x} \to$ SG$\hat{z}$	$S_x = +\hbar/2$	$S_z = \pm\hbar/2$	1/2, 1/2

In the language of the postulates of Chapter 77, each magnet performs a projective measurement of one spin component; the transmitted beam is the corresponding eigenstate; and the probabilities in Table 80.1 are Born-rule overlaps between eigenstates of S_z and S_x. The impossibility of jointly sharp values of S_z and S_x is the physical face of their non-commutativity,

$$[S_z, S_x] = \mathrm{i}\hbar S_y \neq 0\,, \tag{80.4}$$

so that no filtering protocol, however elaborate, can produce a beam that passes both an SG$\hat{z}$ and an SG$\hat{x}$ test with certainty [1593, 2287].

Derivation pending #88

Sequential SG probabilities from the measurement postulates

80.5 Legacy

The Stern–Gerlach experiment founded a lineage of beam techniques whose descendants populate the rest of this treatise. Stern's Hamburg laboratory turned the molecular-beam method into precision physics: diffraction of helium and hydrogen beams from crystal surfaces confirmed the de Broglie wavelength for whole atoms, and beam deflection measurements yielded the proton's magnetic moment — anomalously large, a first hint of nucleon substructure. Rabi's molecular-beam magnetic-resonance method, built directly on

Stern–Gerlach state selection, led through nuclear magnetic resonance to the atomic clock and the maser; a Stern–Gerlach magnet sorting state populations sat at the heart of the first ammonia maser. The historical thread from the Frankfurt magnet to these developments is traced in detail by Friedrich and Herschbach [2050].

The experiment has also been repeated and refined in modern settings — longitudinal and matter-wave-interferometric variants with neutrons and cold atoms among them — always with the same two-valued verdict; and its conceptual role has, if anything, grown. A Stern–Gerlach magnet acting on a spin-1/2 atom is the prototype of a *qubit measurement*: a two-level quantum system, an apparatus that entangles the qubit with a macroscopically distinct pointer variable (here, the atom's transverse position), and a readout that yields one bit with Born-rule statistics. Every projective measurement in today's quantum technologies is, conceptually, a Stern–Gerlach experiment in different hardware, and the sequential cascades of Section 80.4 prefigure the state-preparation-and-measurement protocols of Chapter 84. It is fitting that the sharpest early test of quantum discreteness — undertaken by a sceptic hoping to refute it — should have become the standard mental model for the theory's most characteristic act, the measurement [2050].

Chapter 81

The Hydrogen Atom

Hydrogen is the atom the Schrödinger equation solves exactly, and the one against which almost everything else in quantum physics is calibrated. One electron in the Coulomb field of one proton gives the spectrum $E_n = -13.6\,\mathrm{eV}/n^2$, a degeneracy of n^2 that is larger than rotational invariance alone can explain, and a length scale—the Bohr radius $a_0 = 5.29 \times 10^{-11}\,\mathrm{m}$—built from $\hbar$, the electron mass and the elementary charge. This chapter solves the problem twice: once in wave mechanics, following Schrödinger's first communication [557], and once algebraically through the conserved Runge–Lenz vector, following Pauli [553], whose calculation preceded the wave equation. The second route explains the extra degeneracy: the Coulomb problem has a hidden SO(4) symmetry, identified by Fock [755] and Bargmann [766], and the accidental degeneracy is not accidental at all.

The chapter then peels the exact solution apart layer by layer, because each correction to it is a separate piece of physics measured to its own precision: relativistic and spin–orbit fine structure, closed by the Dirac equation of Chapter 93; the Lamb shift [901], which splits levels the Dirac theory holds degenerate and thereby started renormalized quantum electrodynamics (Chapter 100); the hyperfine coupling to the proton spin, whose ground-state splitting radiates the 21 cm line [981, 984] that mapped the Galaxy; and the finite size of the proton, read off from muonic hydrogen [2164, 2210]. Behaviour in external fields—the Zeeman and Stark effects—is worked as the standard application of perturbation theory in Chapter 82 and only summarized here. Because the theory is exact and the experiments are the most precise in atomic physics, hydrogen is where the fundamental constants are fixed [2392, 2431] and where a discrepancy of a few parts in 10^5 counts as a crisis.

Derivation pending #89

The Hydrogen Atom: all derivations of this chapter are pending.

81.1 The spectrum before the theory

81.1.1 Balmer's formula and the Rydberg constant

[Reserved: the four visible hydrogen lines and Balmer's empirical fit $\lambda = hm^2/(m^2-4)$ [230]; Rydberg's generalization to a difference of two terms, valid across the alkali spectra as well [248], and the Rydberg–Ritz combination principle [348], which made term differences rather than lines the primitive objects; the named series and the modern value $R_\infty = 1.097\,373\,156\,8 \times 10^7/\mathrm{m}$ [2392]. What these formulas were: an exact, purely empirical regularity with no mechanism whatever, and the standing challenge that Chapter 70 recounts.]

Phenomenon 81.1 (The Balmer series and the Rydberg formula)**.** The visible emission spectrum of atomic hydrogen is neither a continuum nor an arbitrary set of lines. Balmer

found that the four lines then measured obey

$$\lambda = b\,\frac{m^2}{m^2-4}\,, \qquad m = 3,4,5,6\,, \tag{81.1}$$

with the single constant $b = 3.6456\times10^{-7}\,\mathrm{m}$, and used Equation (81.1) to predict further members of the series that were afterwards found [230]. Rydberg showed that the same regularity governs the whole hydrogen spectrum, and the alkali spectra besides, in the form

$$\frac{1}{\lambda} = R_\infty\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right), \qquad n_2 > n_1\,, \tag{81.2}$$

with one constant serving every series of the atom [248]; Equation (81.1) is the case $n_1 = 2$. The constant is now $R_\infty = 1.097\,373\,156\,8\times10^{7}/\mathrm{m}$, one of the most accurately known quantities in physics [2431]. Classical physics predicts no discrete spectrum at all, still less one indexed by two integers, and least of all a radiating electron that does not spiral into the nucleus.

Derivation pending #90

The bound-state eigenvalues of the Coulomb problem: the radial equation for the potential that falls as the inverse first power of the distance, the termination of its series solution which quantizes the energy as minus a constant over the square of an integer, and the identification of that constant with a combination of the electron mass, the elementary charge and Planck's constant

Phenomenon 81.2 (The Rydberg–Ritz combination principle)**.** The wavenumbers of the lines of an atom are not independent of one another. Each is a difference of two members of a single list of *terms* characteristic of the atom, so that the sum of the wavenumbers of two observed lines is very often itself the wavenumber of an observed line [348]. The rule holds for atoms whose spectra no formula of Balmer's kind describes, it was established empirically two decades before any mechanism for it existed, and it is what made the terms, rather than the lines, the primitive objects of spectroscopy.

Derivation. Assume, with the postulates of Chapter 77, that the atom possesses a set of stationary states of energies E_i, and that radiation is emitted or absorbed only in transitions between them, one photon of energy $h\nu = E_i - E_f$ per transition. Define the *term* of a state as $T_i := -E_i/hc$, a positive number for a bound state. The wavenumber of the line joining states i and f is then

$$\tilde\nu_{if} = \frac{E_i - E_f}{hc} = T_f - T_i\,, \tag{81.3}$$

a difference of two members of one list, as observed; and for any three states,

$$\tilde\nu_{if} + \tilde\nu_{fg} = (T_f - T_i) + (T_g - T_f) = T_g - T_i = \tilde\nu_{ig}\,,$$

so that two lines sharing a level combine into a third. That third line is *observed* only when the transition from i to g is itself allowed by the selection rules of Chapter 79, which is exactly why the empirical rule holds very often rather than always — a discrepancy that was noticed long before it was explained. Equation (81.2) is the special case of Equation (81.3) in which the terms of one atom happen to form the sequence $T_n = R_\infty/n^2$; the content of Phenomenon 81.2 is that terms exist at all, which is a statement about stationary states and holds whatever the sequence. □

81.1.2 What the old quantum theory got right

[Reserved: Bohr's quantized orbits [395] reproducing Balmer's formula and predicting R_∞ from $\hbar$, m and e; Sommerfeld's elliptical orbits and the relativistic fine-structure formula in α^2 [441], numerically right for reasons later shown to be wrong; the Bohr–Sommerfeld quantization as the action condition of Chapter 23; and the failures that ended the programme—no helium, no intensities, no ground for the selection rules. Detailed account in Chapter 70.]

81.2 The exact solution

81.2.1 Two bodies, one Coulomb potential

[Reserved: separation of the centre of mass, the reduced mass $\mu = m_e m_p/(m_e + m_p)$, and the resulting mass-dependent shift of every level; the isotope shift as its measurable consequence, and the discovery of deuterium by exactly that shift in the Balmer lines [708]; the Coulomb potential taken from Chapter 58 and its $1/r$ form as the input that everything below depends on; why the same reduction serves positronium and muonic hydrogen in Section 81.7.2.]

Phenomenon 81.3 (The isotope shift, and the discovery of deuterium)**.** Every line of the hydrogen spectrum has a faint companion displaced to shorter wavelength by a fixed *relative* amount of about 2.7×10^{-4}; for the red Balmer line at 656.3 nm the companion stands about 0.18 nm away. Urey, Brickwedde and Murphy found the companions, showed that they strengthened when the sample was enriched by evaporating liquid hydrogen nearly to exhaustion, and identified them as the spectrum of an isotope of mass number two [708]. The nucleus is therefore not an infinitely heavy centre of force: the spectrum knows the nuclear mass, and the amount by which it knows it is the electron-to-nucleus mass ratio.

Derivation. Separating the centre of mass of two bodies interacting through a potential that depends only on their separation leaves a one-body problem in the relative coordinate, with the reduced mass $\mu = m_e M/(m_e + M)$ in place of the electron mass, M being the nuclear mass. Every energy of the Coulomb problem is proportional to the mass appearing in the kinetic term, so the Rydberg constant of a nucleus of mass M is

$$R_M = R_\infty \frac{\mu}{m_e} = \frac{R_\infty}{1 + m_e/M}\,, \tag{81.4}$$

and wavelengths, being inversely proportional to it, satisfy $\lambda_M = \lambda_\infty (1 + m_e/M)$ to first order in the small ratio. Comparing a nucleus of one proton with one of twice that mass,

$$\frac{\lambda_H - \lambda_D}{\lambda} \simeq \frac{m_e}{m_p} - \frac{m_e}{2m_p} = \frac{m_e}{2m_p} = \frac{1}{2 \times 1836.15} = 2.72 \times 10^{-4}\,,$$

with the mass ratio taken from [2431]. Applied to the red Balmer line this is a displacement of 0.18 nm, the observed one; and it falls on the *short* side because the heavier nucleus gives the larger reduced mass, hence by Equation (81.4) the larger Rydberg constant and the larger wavenumber. The same substitution run to its limit gives positronium, whose reduced mass is $m_e/2$ and whose levels are therefore half of hydrogen's (Section 81.7.2). □

81.2.2 The radial equation and its quantization

[Reserved: separation in spherical coordinates using the angular solution of Chapter 79; the effective radial equation with the centrifugal barrier; the asymptotic factorization $e^{-\rho/2}\rho^{\ell+1}$ and the termination of the series that quantizes the energy; associated Laguerre polynomials, the principal quantum number $n \geq \ell + 1$, and the spectrum $E_n = -\mu e^4/(32\pi^2\varepsilon_0^2\hbar^2 n^2)$ [557]. The radial problem as a Sturm–Liouville system, with node counting, is Chapter 9.]

81.2.3 Eigenfunctions, orbitals and expectation values

[Reserved: the normalized wavefunctions $\psi_{n\ell m} = R_{n\ell}(r)Y_{\ell m}(\theta,\varphi)$; radial probability densities and their nodes; $\langle r\rangle$, $\langle 1/r\rangle$ and $\langle 1/r^3\rangle$, the last of which is what the fine-structure and hyperfine calculations below actually need; the $\ell = 0$ states alone non-vanishing at the origin, which is why they alone carry the Fermi contact and Lamb terms; virial theorem $\langle T\rangle = -\frac{1}{2}\langle V\rangle$; the continuum states above threshold and the photoionization edge (Chapter 65).]

81.2.4 Levels, degeneracy and scales

[Reserved: the level diagram, the ionization energy 13.6 eV, the Bohr radius as $\langle r\rangle$ of the ground state, and the fine-structure constant α entering as the ratio of the electron's orbital velocity to c—which is why every correction below arrives as a further power of α, and why the numerical value of α is fixed by measurements of exactly this kind (Chapter 90). The degeneracy: $2\ell+1$ from rotational invariance, but a total of n^2 orbital states independent of ℓ, which no rotational argument gives. That surplus is the subject of the next section.]

Phenomenon 81.4 (The Coulomb degeneracy exceeds what rotation gives)**.** To the accuracy at which fine structure may be neglected, the bound levels of hydrogen depend only on the principal quantum number and not on the orbital one: all the terms with $\ell = 0, 1, \ldots, n-1$ belonging to one n coincide, giving $\sum_{\ell=0}^{n-1}(2\ell+1) = n^2$ degenerate orbital states. Rotational invariance guarantees only the $2\ell+1$ states of a single ℓ, so the surplus is unexplained by any symmetry so far introduced. Two observations show that the coincidence is exact and not merely close. Hydrogen alone among atoms has a Stark effect *linear* in the applied electric field [422], and a linear response requires states of opposite parity to be exactly degenerate, since a field mixes them in first order only when it costs no energy to do so. And the alkali spectra, whose inner electrons screen the nucleus and spoil the exact inverse-square-law force, show precisely the term splittings by ℓ that hydrogen lacks [248]: it is the $1/r$ potential that is responsible, not the one-electron character of the atom.

Derivation pending #91

The ℓ-independence of the Coulomb spectrum: the conserved Runge–Lenz vector peculiar to the inverse-square-law force, the closure of its commutators with the angular momentum on the algebra of SO(4) *for bound states, and the identification of the levels with irreducible representations of that group whose dimension is the square of the principal quantum number*

81.3 Degeneracy and the hidden symmetry

81.3.1 The Runge–Lenz vector and the algebraic solution

[Reserved: the classical conserved vector of the Kepler problem, pointing along the major axis and expressing the closure of the orbits (Chapter 27), reintroduced into the quantum theory by Lenz [514]; its Hermitian symmetrization $\boldsymbol{A} = \frac{1}{2\mu}(\boldsymbol{p}\times\boldsymbol{L} - \boldsymbol{L}\times\boldsymbol{p}) - e^2\boldsymbol{r}/(4\pi\varepsilon_0 r)$, the commutators of $\boldsymbol{A}$ with $\boldsymbol{L}$ and with itself, and Pauli's derivation of the whole spectrum from that algebra alone [553]—published before the wave equation and reproducing Balmer exactly.]

81.3.2 SO(4) and Fock's construction

[Reserved: the rescaled generators closing on $\mathfrak{so}(4) \cong \mathfrak{su}(2)\oplus\mathfrak{su}(2)$ for bound states (and $\mathfrak{so}(3,1)$ for scattering states); Fock's demonstration that stereographic projection of mo-

mentum space onto the three-sphere turns the bound-state equation into the equation for spherical harmonics on S^3, so that the n^2 degeneracy is the dimension of an $\mathrm{SO}(4)$ irreducible representation [755], with the algebraic completion by Bargmann [766]. Group-theoretic setting: Chapters 5 and 14. The lesson—that an unexplained degeneracy signals a symmetry not yet identified—is used again for the isotropic oscillator and in Chapter 102.]

81.4 Fine and hyperfine structure

81.4.1 Relativistic corrections and spin–orbit coupling

[Reserved: the three terms of order α^2 relative to the Bohr energies—the relativistic kinetic correction, the spin–orbit coupling with its Thomas factor (Chapter 79) [561], and the Darwin contact term for $\ell = 0$ [595]; their sum depending only on n and j, so that $2s_{1/2}$ and $2p_{1/2}$ remain degenerate; the exact Dirac result reproducing Sommerfeld's formula [441, 596] and derived in Chapter 93; the observed $2p_{3/2}$–$2p_{1/2}$ splitting of about 10.9 GHz as the measured scale.]

81.4.2 The Lamb shift

[Reserved: the microwave measurement showing $2s_{1/2}$ lying about 1058 MHz above $2p_{1/2}$, in flat contradiction with the Dirac degeneracy [901]; Bethe's non-relativistic estimate, the first successful renormalized calculation, obtained within weeks [897]; the physical content—vacuum polarization and the electron's self-interaction with its own radiation field—and the full treatment in Chapter 100; the modern value and its role, together with the electron $g-2$ (Chapter 112), as the sharpest confrontation between quantum field theory and measurement.]

Phenomenon 81.5 (The Lamb shift)**.** The states $2s_{1/2}$ and $2p_{1/2}$ of hydrogen are not degenerate, although the Dirac theory makes them exactly so, its energies depending only on n and j. Driving microwave transitions directly between the two in a beam of metastable atoms, and using the metastable atoms themselves as the detector, Lamb and Retherford found $2s_{1/2}$ lying above $2p_{1/2}$ by about 1000 MHz [901]; the modern value is close to 1058 MHz. The shift is only about one part in 10^6 of the binding energy of the level, and it falsified the best relativistic one-particle theory then available. Bethe's estimate of it, made within weeks, was the first calculation in which the infinities of the radiation theory were removed by subtracting a free-particle self-energy [897].

Derivation pending #92

The radiative level shift: the interaction of the bound electron with the fluctuating quantized electromagnetic field, the difference between the bound and free self-energies which is what remains finite after regularization, and the vacuum-polarization contribution of opposite sign that accompanies it

81.4.3 Hyperfine structure

[Reserved: the coupling of the electron spin to the proton magnetic moment, dominated for $\ell = 0$ by the Fermi contact term proportional to $|\psi(0)|^2$ [652]; the total $\boldsymbol{F} = \boldsymbol{I} + \boldsymbol{J}$ and the singlet–triplet splitting of the ground state, smaller than the fine structure by the electron-to-proton mass ratio—about 5.9 μeV, that is 1420.4058 MHz; the proton magnetic moment measured independently by magnetic resonance (Chapter 79); the hyperfine anomaly as a probe of nuclear magnetization distribution (Chapter 107).]

81.4.4 Hydrogen in external fields

[Reserved: a summary, with the derivations deliberately placed in Chapter 82 where the perturbative machinery is built—the linear Stark effect, peculiar to hydrogen because the ℓ degeneracy of Section 81.2.4 survives [422, 560]; the Zeeman effect in the weak field with Landé's g factor and the Paschen–Back regime in the strong field [277, 391]; diamagnetic hydrogen in very strong fields as a laboratory for quantum chaos (Chapter 32).]

81.5 The 21-centimetre line

81.5.1 A prediction made before the instrument existed

[Reserved: van de Hulst's wartime calculation that the hyperfine ground-state transition of neutral hydrogen should be detectable as a radio line near 21 cm*, despite a spontaneous-emission lifetime of order* 10^7 *years, because the Galaxy contains enough atoms [880]; the magnetic-dipole character of the transition and the selection rules of Chapter 79 that make it so slow; why a line this narrow and this weak is nonetheless the most useful in radio astronomy.]*

81.5.2 Detection

[Reserved: the independent detections announced together in 1951—Ewen and Purcell at Harvard [981] and Muller and Oort at Kootwijk [984]—with the receiver technique, the frequency-switching that pulled the line out of the noise, and the measured rest frequency now known to 1420.4058 MHz*. This is the worked case of a quantum-mechanical energy difference of a few* μeV *becoming an astronomical instrument.]*

Phenomenon 81.6 (The 21-centimetre line)**.** Neutral atomic hydrogen in its electronic ground state radiates one spectral line in the radio band, at a rest frequency of 1420.4058 MHz — a vacuum wavelength close to 21 cm. Ewen and Purcell detected it in emission from the Milky Way with a horn antenna mounted out of a laboratory window and a receiver that switched frequency rapidly so as to subtract its own noise [981]; Muller and Oort confirmed it within weeks at Kootwijk and began mapping the Galaxy with it [984]. The transition joins the two hyperfine components of the ground state, split by the interaction of the electron's magnetic moment with the proton's. The splitting is about 5.9 μeV, smaller than the fine structure by roughly the electron-to-proton mass ratio, and the transition is a magnetic dipole one with a spontaneous lifetime of order 10^7 years — so weak that nothing smaller than a galaxy could have been the source in which it was first seen.

Derivation pending #93

The ground-state hyperfine splitting of hydrogen: the Fermi contact interaction between the electron and nuclear magnetic moments, its expectation value in the ground state, which is proportional to the probability density of the electron at the origin, and the resulting separation of the singlet from the triplet

81.5.3 What the line measures

[Reserved: Doppler-shifted 21 cm *emission mapping the rotation curve and spiral structure of the Galaxy [984]; column densities of neutral hydrogen; the flat rotation curves of spiral galaxies as one of the primary lines of evidence catalogued in Chapter 134; absorption against background sources; and the redshifted line as a probe of the neutral intergalactic medium*

(Chapter 48). The complementary Lyman-α absorption forest in quasar spectra [1452] uses the $n = 1 \to 2$ transition of the same atom.]

81.6 Precision hydrogen and the constants

81.6.1 The Rydberg constant and optical frequency metrology

[Reserved: two-photon Doppler-free spectroscopy of the $1s$–$2s$ transition, whose 1 Hz-scale natural width makes it one of the sharpest resonances in physics; the measurement 2.4661×10^{15} Hz to a few parts in 10^{15} using an optical frequency comb referenced to a caesium standard [2179]; the least-squares adjustment in which hydrogen data, together with the proton radius, determine R_∞ [2392, 2431]; the SI traceability chain of Chapter 2 and the experiment chapter Chapter 90.]

81.6.2 The proton radius

[Reserved: the finite-size correction shifting s levels in proportion to the mean square charge radius; muonic hydrogen, in which the muon's larger mass concentrates the wavefunction and amplifies the effect by roughly $(m_\mu/m_\mathrm{e})^3$, giving $r_\mathrm{p} = (0.841\,84 \pm 0.000\,67)$ fm [2164], confirmed and sharpened by the $2s$–$2p$ transitions of [2210]—about 7σ below the then-accepted electronic and scattering value: the "proton radius puzzle". Its resolution by new electronic-hydrogen measurements [2281, 2331] and the revised CODATA value [2431]; the honest reading, that a 7σ discrepancy was settled by remeasurement rather than by new physics.]

Phenomenon 81.7 (The spectrum measures the size of the proton)**.** The s levels of hydrogen lie slightly higher than a point nucleus would place them, by an amount proportional to the mean square charge radius of the proton, and the excess is large enough to be measured. Replacing the electron by a muon concentrates the wavefunction on the nucleus and amplifies the effect by millions: laser spectroscopy of the $2s$–$2p$ transitions of muonic hydrogen gave $r_\mathrm{p} = (0.841\,84 \pm 0.000\,67)$ fm [2164], and a second transition sharpened it [2210]. That value sat about 7σ below the electronic and electron-scattering value then accepted, and the discrepancy was closed by remeasuring the electronic transitions rather than by new physics [2281, 2331, 2431]. A property of the nucleus is thus read out of the energy levels of the electron cloud around it.

Derivation. Only the amplification factor is derived here; the coefficient belongs with the perturbation theory of Chapter 82. Inside a nucleus of finite size the potential departs from the point-charge Coulomb form by some $\delta V(r)$, positive because the enclosed charge is reduced, and supported on a region of the size of the proton. To first order the shift of a level is $\Delta E = \int |\psi|^2 \, \delta V \, \mathrm{d}^3 r$, and since that region is smaller than the atom by five orders of magnitude, $|\psi|^2$ may be replaced by its value at the origin:

$$\Delta E \simeq |\psi(0)|^2 \int \delta V \, \mathrm{d}^3 r \propto |\psi(0)|^2 \left\langle r_\mathrm{p}^2 \right\rangle . \tag{81.5}$$

Two consequences follow. Only $\ell = 0$ states shift at all, because $\psi(0) = 0$ for every $\ell > 0$ — which is why the muonic measurement compares an s level with a p level. And $|\psi(0)|^2 = 1/\pi a^3$ for the ground state, with the Bohr radius $a = 4\pi\varepsilon_0\hbar^2/\mu e^2$ inversely proportional to the reduced mass, so that $\Delta E \propto \mu^3 \left\langle r_\mathrm{p}^2 \right\rangle$. Muonic hydrogen has $\mu \simeq 186\, m_\mathrm{e}$, so its finite-size shift exceeds the electronic one by a factor of about 6×10^6. This is why one laser transition in an exotic atom could outweigh decades of electronic spectroscopy, and equally why a systematic error in that one measurement would have had nothing to check it against — which is the reason the discrepancy was taken seriously rather than dismissed. □

81.6.3 Antihydrogen and the tests it carries

[Reserved: trapped antihydrogen and laser spectroscopy of its $1s$–$2s$ transition, agreeing with hydrogen to a relative precision of about 2×10^{-12} [2299]—at present the sharpest spectroscopic test of CPT invariance (Chapter 105); the antihydrogen hyperfine structure; and the gravitational free-fall measurement, which belongs with the equivalence-principle tests of Chapter 42. Positron and antiproton production and the trapping technique are the experimental content.]

81.7 Hydrogen-like systems

81.7.1 One electron, other cores

[Reserved: hydrogenic ions Z^2-scaled, up to hydrogen-like uranium where $Z\alpha$ approaches unity and the expansion in Section 81.4.1 fails; Rydberg atoms, in which a single highly excited electron sees a screened core and reproduces the hydrogen spectrum with a quantum defect, with polarizabilities scaling as n^7 [1881] (used in the quantum optics of Chapter 122); and the Wannier exciton in a semiconductor, a hydrogen-like bound state of electron and hole with the dielectric constant and effective masses substituted [805], taken up in Chapter 127.]

81.7.2 Exotic atoms

[Reserved: positronium, an electron–positron bound state with reduced mass $m_e/2$ and no nucleus at all, whose formation and decay were first observed by Deutsch [980]—the purely leptonic hydrogen, where the annihilation channels test Chapter 100; muonium, observed through its Larmor precession [1184]; muonic hydrogen [2164, 2210], whose result reshaped the constants; and antiprotonic and pionic atoms, whose level shifts measure the strong interaction at threshold (Chapter 107). Each is the same equation with one substitution, which is what makes hydrogen the reference problem of the treatise.]

Chapter 82

Approximation Methods

The list of quantum systems that can be solved exactly is short: the free particle, the square wells, the oscillator and the periodic chain of Chapter 78, the rigid rotor of Chapter 79, and the Coulomb problem of Chapter 81. Everything else in this treatise—helium, every molecule, every solid, every scattering cross-section, every decay rate—is reached by approximation, and the approximations are therefore not a technical appendix to quantum mechanics but the greater part of how it is compared with experiment. This chapter builds the four that carry that load: stationary perturbation theory, in the form Schrödinger gave it in his third communication [560] on the model of Rayleigh's treatment of vibrating systems; the variational method [359], which bounds rather than expands; the semiclassical WKB approximation [535, 549, 563], which is the $\hbar \to 0$ limit of Chapter 23 made systematic; and time-dependent perturbation theory [559, 572], whose leading term is the transition-rate formula that Fermi later christened the golden rule [954].

Two habits of the chapter deserve stating in advance. First, every method is presented with its *failure* mode named: perturbation series in quantum mechanics are generically asymptotic and divergent [998, 1392], the WKB form breaks down exactly at the classical turning points, and a variational bound says nothing about how close it is. Second, each method is anchored to a measurement it actually explains—the Stark [422] and Zeeman [277] splittings, the helium ground-state energy [628], alpha-decay lifetimes, spectral-line intensities—so that the accuracy claimed can be checked rather than asserted. The chapter closes with the two opposite limits of a time-dependent Hamiltonian: the adiabatic theorem [594], whose geometric remainder is Berry's phase [1734], observed directly in polarized light carried along a helical fibre [1783]; and the sudden approximation at the other extreme. Standard treatments are [1593, 1599, 2287].

Derivation pending #94

Approximation Methods: all derivations of this chapter are pending.

82.1 Stationary perturbation theory

82.1.1 The Rayleigh–Schrödinger series

[Reserved: the split $H = H_0 + \lambda V$ with H_0 solved; the expansion of eigenvalue and eigenvector in powers of λ; the first-order shift $E_n^{(1)} = \langle n^{(0)} | V | n^{(0)} \rangle$ and the second-order shift $\sum_{m \neq n} |\langle m| V |n\rangle|^2 / (E_n - E_m)$, always negative for the ground state; normalization conventions and the intermediate-normalization form; the small parameter is the ratio of a matrix element to a level spacing, not λ itself, which is what makes near-degenerate levels the dangerous case [560]. Rigorous statement in terms of analytic families of operators

[1335], with the functional-analytic setting of Chapter 12.]

82.1.2 Degenerate and near-degenerate levels

[Reserved: why the naive series diverges term by term when $E_n = E_m$, and the resolution—diagonalize V within the degenerate subspace first; the secular determinant, the resulting good quantum numbers, and the identification of the symmetry that the perturbation leaves unbroken; complete and partial lifting of a degeneracy; the crossing rule that levels of the same symmetry repel rather than cross, with avoided crossings as the observable signature (their dynamical version is Section 82.6.3); the worked cases used later are the Stark and Zeeman effects of Section 82.2 [560].]

82.1.3 Does the series converge? Usually not

[Reserved: the honest statement that a perturbation series is generally asymptotic rather than convergent—optimal truncation, and error of the order of the first omitted term; the anharmonic oscillator whose coefficients grow factorially, computed exactly by Bender and Wu [1392]; Dyson's argument that the quantum-electrodynamic series cannot converge, because the theory is unstable for negative coupling [998] (see Chapter 100); Borel and Padé resummation; and the cases where honest convergence does hold, for which the analytic perturbation theory of [1335] gives criteria.]

82.2 Perturbation theory at work

82.2.1 The Stark effect

[Reserved: the observation of electric-field splitting of the Balmer lines by Stark [422], independently by Lo Surdo [406]; the quadratic shift for a non-degenerate level and the atomic polarizability it defines; the linear *Stark effect, peculiar to hydrogen because states of opposite parity are degenerate (Chapter 81), computed by Schrödinger in the paper that introduced the method [560]; parabolic coordinates; and the honest caveat that in a uniform field no bound state strictly exists—the levels are long-lived resonances that eventually field-ionize, as Oppenheimer showed [616].]*

82.2.2 The Zeeman effect

[Reserved: the splitting of spectral lines in a magnetic field [277]; the weak-field regime, where the spin–orbit coupling dominates, $\boldsymbol{J}$ is the good quantum number and the splitting is set by the Landé factor [478] derived from the projection theorem of Chapter 79; the strong-field Paschen–Back regime, where $\boldsymbol{L}$ and $\boldsymbol{S}$ decouple [391]; the crossover as a two-by-two degenerate problem solved exactly; why the "anomalous" effect was anomalous only before spin. Applications: stellar magnetic fields measured by Zeeman splitting (Chapter 52), and magnetometry.]

82.2.3 Fine structure, hyperfine structure and isotope shifts

[Reserved: the standard worked example of degenerate perturbation theory—the α^2 corrections to hydrogen (relativistic kinetic, spin–orbit with the Thomas factor [561], and Darwin) collected in Chapter 81 and derived exactly in Chapter 93; the hyperfine contact term [652]; nuclear volume and mass shifts. Each is a perturbation whose smallness is set by a physical ratio—α^2, the magneton ratio, the nuclear radius over a_0—which is the model for how the parameter should be identified in general.]

82.2.4 The Hellmann–Feynman and virial theorems

[Reserved: $\partial E/\partial\lambda = \langle\psi|\partial H/\partial\lambda|\psi\rangle$ for a normalized eigenstate depending on a parameter [833], and its use in converting a spectrum into forces—the electrostatic theorem for molecular binding (Chapter 88); the quantum virial theorem for a homogeneous potential and its check $\langle T\rangle = -\frac{1}{2}\langle V\rangle$ in hydrogen; the two theorems as exact constraints that any approximate calculation can be tested against, which is what makes them worth stating in a chapter about approximations.]

82.3 The variational method

82.3.1 The Rayleigh–Ritz principle

[Reserved: the theorem that $\langle\psi|H|\psi\rangle \geq E_0$ for any normalized trial state, with equality only for the ground state, and its stationarity form—the Schrödinger equation is the Euler–Lagrange equation of the energy functional (Chapter 16); the systematic method of minimizing over a finite basis, reducing the problem to a generalized eigenvalue determinant [359]; the min-max characterization of the higher eigenvalues (Chapter 12), which turns the same construction into bounds on excited states; the practical warning that a good energy does not imply a good wavefunction, because the error in E is second order in the error in ψ.]

82.3.2 Helium

[Reserved: the two-electron atom, the system on which the old quantum theory failed outright (Chapter 70); the crude one-parameter trial function with an effective nuclear charge screening the second electron; and Hylleraas' explicit inclusion of the interelectronic distance r_{12} in the trial function, which reproduced the measured ionization energy to spectroscopic accuracy [628]—the first quantitative proof that quantum mechanics governs many-electron atoms. Ortho- and para-helium as the singlet–triplet structure of Chapters 79 and 83.]

82.3.3 Self-consistent fields

[Reserved: Hartree's self-consistent field, in which each electron moves in the averaged potential of the others and the equations are iterated to consistency [604]; Fock's antisymmetrized version, whose exchange term is a direct consequence of the statistics of Chapter 83 [653]; the variational character of the whole construction, and Koopmans' theorem identifying an orbital energy with an ionization energy [743]; what is left out—electron correlation—and where it is recovered (Chapter 88); density-functional methods as the working tool of Chapter 126.]

82.4 The semiclassical approximation

82.4.1 The WKB expansion

[Reserved: the ansatz $\psi = \mathrm{e}^{\mathrm{i}S/\hbar}$ with $S = S_0 + \hbar S_1 + \cdots$, the leading order reproducing the Hamilton–Jacobi equation of Chapter 23 and the first order the amplitude $1/\sqrt{p(x)}$ that conserves flux; the validity condition $|\mathrm{d}\lambda/\mathrm{d}x| \ll 2\pi$ on the local de Broglie wavelength $\lambda(x) = h/p(x)$, that is, a wavelength varying slowly on its own scale; the near-simultaneous papers of Wentzel [563], Kramers [549] and Brillouin [535], and the earlier purely mathematical treatment of the same equation by Jeffreys [526], which is why the method is also called JWKB.]

82.4.2 Turning points and connection formulae

[Reserved: the breakdown of the expansion where $p(x) \to 0$, since the assumed short wavelength diverges there; linearization of the potential about a turning point and the Airy-function solution (Chapter 9); the connection formulae joining oscillatory and exponential regions and the $\pi/4$ phase they carry; the classically forbidden region and the exponentially small tail; and caustics as the same phenomenon in the optical language of Chapter 23.]

82.4.3 Semiclassical quantization

[Reserved: two turning points and the quantization condition $\oint p\,\mathrm{d}x = (n + \tfrac{1}{2})h$, the Bohr–Sommerfeld rule [441] corrected by the Maslov half-integer that the connection formulae supply; its exactness for the harmonic oscillator and its accuracy for high n as a form of the correspondence principle; the extension to non-separable systems and the general index counting [1150]; where semiclassical quantization fails altogether—classically chaotic systems, whose spectra are the subject of Chapter 32.]

82.4.4 Barrier penetration

[Reserved: the transmission exponential $\exp\left(-\frac{2}{\hbar}\int |p|\,\mathrm{d}x\right)$ as the standard product of the method; the Gamow factor for the Coulomb barrier [600], which reproduces the empirical Geiger–Nuttall relation between alpha-decay half-life and energy over more than twenty orders of magnitude (Chapter 78, Chapter 107); field ionization of atoms [616] and cold emission from metals; thermonuclear reaction rates as the same integral run inwards (Chapter 52). The exponential sensitivity is what makes an approximate exponent adequate and an approximate prefactor irrelevant.]

82.5 Time-dependent perturbation theory

82.5.1 Variation of constants and the interaction picture

[Reserved: the expansion of the state in eigenstates of H_0 with time-dependent coefficients, introduced for wave mechanics in Schrödinger's fourth communication [559] and cast in the form used ever since by Dirac [572]; the interaction picture and the Dyson series [927], whose field-theoretic use is Chapter 100; first-order amplitude as the Fourier transform of the perturbation at the Bohr frequency; unitarity and the limits of validity—the formulas below fail as soon as the depletion of the initial state matters.]

82.5.2 Fermi's golden rule

[Reserved: the transition rate $\Gamma = (2\pi/\hbar)\,|\langle f|\,V\,|i\rangle|^2\,\rho(E_f)$ into a continuum, with the $\sin^2$ kernel becoming a delta function of energy at long times and the resulting rate constant in time—the origin of exponential decay. The result is Dirac's [572]; the name is Fermi's, whose Chicago lecture notes list it as the second of two "golden rules" [954]. The chapter says which is which rather than repeating the customary attribution of both to Fermi. Second-order form for transitions through virtual intermediate states; applications to beta decay (Chapter 107), spontaneous and stimulated emission (Chapter 122), and scattering rates (Chapter 86).]

82.5.3 Periodic perturbations and the two-level check

[Reserved: a harmonic perturbation, absorption and stimulated emission at $\omega \approx \omega_{fi}$, the resonance denominators and the rotating-wave approximation; oscillator strengths and the Thomas–Reiche–Kuhn sum rule [527], which fixes the total absorption of an atom, summed

over lines, independently of its structure; the exactly solvable two-level system in a rotating field [798] used as the check on the perturbative answer, showing where first order is right and where Rabi flopping takes over (Chapter 79, Chapter 122).]

82.5.4 The Born approximation

[Reserved: the golden rule applied to scattering by a potential; the first Born amplitude as the Fourier transform of the potential in the momentum transfer [533], the Rutherford cross-section [379] recovered from a Coulomb potential, and the form factor as the Fourier transform of a charge distribution—which is how nuclear and nucleon radii are measured (Chapters 107 and 116). Validity at high energy and weak coupling, the second Born term, and the non-perturbative alternative—partial waves and phase shifts—in Chapter 86.]

82.6 Adiabatic and sudden limits

82.6.1 The adiabatic theorem

[Reserved: Born and Fock's theorem [594]: a system in a non-degenerate eigenstate of a slowly varying Hamiltonian remains in the instantaneous eigenstate, with transition probability vanishing in the limit of infinitely slow change; the condition as a ratio of the rate of change to the square of the level spacing; the dynamical phase; the classical counterpart, adiabatic invariance of the action, in Chapter 23; uses—adiabatic passage in spectroscopy, adiabatic cooling, and the slow ramps of Chapter 121.]

82.6.2 Berry's geometric phase

[Reserved: the extra, non-integrable phase left over when the Hamiltonian is carried around a closed loop in parameter space—equal to the flux of a curvature through the loop and independent of how slowly the circuit is traversed, so it is geometric and not dynamical [1734]; the spin-1/2 example giving half the solid angle subtended, and the degeneracy point acting as a monopole of the curvature; the optical precursor in the interference of polarized beams [1093]; the non-adiabatic generalization [1786]. Observed: rotation of the plane of polarization of light in a helically wound optical fibre, matching the predicted solid angle [1783]. The fibre-bundle language is Chapter 14; the same curvature governs the quantized Hall conductance of Chapter 128.]

82.6.3 Landau–Zener transitions

[Reserved: the failure of adiabaticity at an avoided crossing, and the exact two-level transition probability $\exp(-2\pi\gamma)$ *for a linear sweep through it [711], with* γ *the ratio of the squared gap to the sweep rate; the two limits—slow sweep following the adiabatic branch, fast sweep diabatically crossing it; applications to atomic and molecular collisions (Chapter 88), to the Mikheyev–Smirnov–Wolfenstein matter resonance [1631] that carries solar neutrinos through the Sun's density gradient (Chapters 103 and 114), and to control sequences in two-level systems.]*

82.6.4 The sudden approximation

[Reserved: the opposite extreme—a change of Hamiltonian fast compared with every relevant period, so that the state vector has no time to evolve and the transition probabilities are simply the overlaps $|\langle f_{new}|i_{old}\rangle|^2$*; the validity criterion in terms of the switching time against* $\hbar/\Delta E$ *[1599]; the worked case of nuclear beta decay, where the sudden change of nuclear*

charge leaves the atomic electrons excited or ionized—shake-up and shake-off, measurable in the electron spectrum (Chapter 107).]

82.6.5 Born–Oppenheimer separation

[Reserved: the adiabatic theorem applied to nuclei and electrons, with the expansion parameter $(m_e/M)^{1/4}$; electronic energies computed at fixed nuclei become the potential surface on which the nuclei then move [564]; the resulting separation of electronic, vibrational and rotational scales that organizes every molecular spectrum (Chapter 88, Chapter 79); where the separation fails—conical intersections between surfaces, at which the geometric phase of Section 82.6.2 becomes observable in the vibrational spectrum.]

Chapter 83

Identical Particles

Nothing in the postulates of Chapter 77 yet says what happens when two particles of a system cannot be told apart. Classically the question barely arises: two billiard balls may agree in every measurable property and still be labelled, because each carries its own trajectory. Quantum mechanically there is no trajectory to carry a label, the labels are therefore unobservable, and the demand that no prediction depend on them turns out to be among the strongest constraints in physics. It splits matter into exactly two families, forbids two electrons from occupying the same state [529], supplies the effective force that orders magnets [537, 545], fixes the size of atoms and the stability of bulk matter [1345, 1554], and holds white dwarfs up against their own gravity [676]. Its classical shadow is already visible in the Gibbs paradox of Chapter 118 [304].

The chapter sits after the general apparatus of Chapters 77 and 79 and Chapter 82 because it needs all three: a Hilbert space to act on with permutations, spin to build the two families out of, and perturbation theory to turn the exchange term into a number. It is the last piece of one-particle quantum mechanics and the first piece of every many-body theory that follows — the multi-electron atoms and molecules of Chapter 88, the electron gas of Chapter 126, the quantum statistics of Chapter 120, and the condensates of Chapter 121. Helium is the worked example throughout. The deepest statement of the chapter, the spin–statistics connection [854], is stated here and proved only in Chapter 109, since its hypotheses are relativistic and field-theoretic and are not available at this point in the book. Standard treatments are [1593, 2287].

Derivation pending #95

Identical Particles: all derivations of this chapter are pending.

83.1 Indistinguishability and the exchange operator

83.1.1 What indistinguishability asserts

[Reserved: the physical assertion — that no measurement can distinguish particle 1 *from particle* 2*, so that every observable* A *of an* N*-particle system commutes with every permutation of the labels; the contrast with the classical case, where identical particles are still individuated by their trajectories, and the resulting overcounting that Gibbs identified in the entropy of mixing [304] and that the* $1/N!$ *of Chapter 118 patches by hand; the statement that "identical" means agreement in every intrinsic quantum number — mass, charge, spin, lepton and baryon number — and the experimental fact that electrons agree in all of them to the precision of Chapter 112.]*

83.1.2 The permutation group on the Hilbert space

[Reserved: the symmetric group S_N acting on $\mathcal{H}^{\otimes N}$ by permuting tensor factors; that this is a unitary representation, reducible in general, whose irreducible components are labelled by Young diagrams; the group-theoretic language being that of Chapter 5; the observation that the totally symmetric and totally antisymmetric one-dimensional representations are distinguished among all of them by being the only ones in which every transposition acts by a single number.]

83.1.3 The exchange operator and its eigenvalues

[Reserved: the transposition operator P_{12}, its unitarity and its involutive property $P_{12}^2 = \mathbb{1}$, hence eigenvalues ± 1; that $[P_{12}, \mathcal{H}] = 0$ for a Hamiltonian symmetric under exchange, so exchange parity is a constant of the motion and a state cannot change family under any evolution; the resulting superselection rule, first formulated as an independent postulate by Messiah and Greenberg [1291].]

83.1.4 The symmetrization postulate

[Reserved: the postulate that the physical states of N identical particles span only the totally symmetric or the totally antisymmetric subspace of $\mathcal{H}^{\otimes N}$, never a mixed-symmetry one; its statement in Dirac's 1926 paper [537] and in Heisenberg's resonance paper [545]; the careful modern formulation, which separates the postulate from the spin–statistics theorem that later explains it [1291]; that this is a restriction on the state space*, not a restriction on the dynamics, and that it is logically independent of everything in Chapter 77.]*

83.2 Bosons and fermions

83.2.1 Bose–Einstein statistics

[Reserved: the symmetric sector; Bose's counting of photon states [509] and Einstein's extension to a material ideal gas, with the prediction of condensation below a critical temperature [522]; the enhancement factor $\sqrt{n+1}$ for adding a particle to an occupied mode, which is the origin of stimulated emission (Chapter 122); pointers to the thermodynamics in Chapter 120 and to the observation of condensation in Chapter 121.]

83.2.2 Fermi–Dirac statistics and the exclusion principle

[Reserved: the antisymmetric sector; Pauli's exclusion principle, announced as an empirical rule that closes atomic shells and explains the periodic table [529], and its restatement as antisymmetry of the wavefunction; the ideal gas obeying it, worked out independently by Fermi [539] and Dirac [537]; the Fermi sea, Fermi energy and Fermi surface as they will be used in Chapters 126 and 127; the statement that exclusion is a property of the state space and not a force — nothing pushes.]

Phenomenon 83.1 (Exclusion, and the lengths of the periods)**.** The chemical elements repeat their properties in periods of $2, 8, 8, 18, 18, 32$; the ionization energy climbs through each period and falls abruptly at the start of the next; and the elements ending each period are chemically inert. Pauli extracted from the spectroscopic evidence the rule behind it: no two electrons of an atom occupy states agreeing in all four quantum numbers, so that a shell of principal quantum number n holds at most $2n^2$ electrons [529]. Nothing in Chapter 77 forbids the alternative. Were the electrons free to share the lowest orbital, every atom would have the ground configuration $1s^N$, all elements would be chemically alike, and there would be no chemistry and no periodic table to explain.

Derivation. Take the exclusion rule as stated — at most one electron per one-electron state — and count the states. In a central field a one-electron state is labelled by n, by $\ell = 0, 1, \ldots, n-1$, by $m_\ell = -\ell, \ldots, \ell$ and by the two values of m_s, the ranges of ℓ and m_ℓ being those established in Chapter 79. The number of states of a given n is therefore

$$2\sum_{\ell=0}^{n-1}(2\ell+1) = 2\left[2\cdot\frac{(n-1)\,n}{2}+n\right] = 2n^2\,, \tag{83.1}$$

which is $2, 8, 18, 32$ for $n = 1, 2, 3, 4$: the observed period lengths, and the observed inert gases are the elements at which a group of these states is exactly filled. The qualification the data impose is that the ℓ degeneracy of Chapter 81 is broken in a many-electron atom, so the subshells fill out of order — $4s$ before $3d$ — and a period of 18 is first entered in the fourth row; and the periods of 8 are the s and p subshells of one n, which hold $2+6$ between them. What Equation (83.1) does not contain is a force. No term in the Hamiltonian was modified; the counting restricts which states exist, and that restriction alone produces the periodic table. □

83.2.3 Parastatistics, and why nature does not use it

[Reserved: the mixed-symmetry representations of S_N are not excluded by any principle of Chapter 77, and Green's generalized field quantization [1019] realizes them consistently as parabosons and parafermions of order p; the observational situation is that every known particle has $p = 1$; the theoretical statement that in a local relativistic field theory parastatistics of order p is equivalent to ordinary statistics with a hidden $\mathrm{SU}(p)$ degree of freedom, which is how colour entered Chapter 102; the experimental bounds are those of Section 83.5.2.]

83.2.4 Exchange in an effectively two-dimensional system

[Reserved: the argument that only two families exist uses the topology of the configuration space of N points, and in three space dimensions the exchange paths fall into two homotopy classes; Leinaas and Myrheim showed that for particles confined to a plane the relevant group is the braid group and a continuum of phases is allowed [1603], with Wilczek naming the resulting excitations anyons [1706]. This treatise instantiates $3+1$ dimensions only, so the statement belongs here as a property of an effectively two-dimensional system *inside ordinary space — a two-dimensional electron gas in a strong magnetic field — and not as a statement about spacetime. The evidence is the fractional quantum Hall effect of Chapter 128 and the direct anyon-collision measurement of [2354].]*

83.3 Constructing many-particle states

83.3.1 Slater determinants and permanents

[Reserved: the antisymmetrizer applied to a product of N orthonormal orbitals, giving the determinant of the matrix $\psi_i(x_j)$ normalized by $1/\sqrt{N!}$, introduced by Slater in his theory of complex spectra [637]; that it vanishes identically when two orbitals coincide, which is the exclusion principle read off an algebraic identity; the bosonic counterpart, the permanent, and why it is computationally hard where the determinant is easy; matrix elements of one- and two-body operators between determinants — Slater's rules — as the working tool of Chapter 88.]

83.3.2 Occupation numbers and second quantization

[Reserved: the passage from labelled particles to occupation numbers of modes; Fock space and its construction [698]; creation and annihilation operators with the commutator algebra

of Jordan and Klein for bosons [579] and the anti *commutator algebra of Jordan and Wigner for fermions [609], the latter written down precisely to make the Pauli exclusion principle an operator identity; that symmetrization is then automatic rather than imposed; the observation that this formalism is already field theory and is taken up as such in Chapter 99.]*

83.3.3 The self-consistent field

[Reserved: the variational determination of the best single Slater determinant; Hartree's self-consistent field [604] and Fock's antisymmetrized correction [653], giving the Hartree–Fock equations with their direct and exchange terms; the definition of correlation energy as what the method misses; helium and the light atoms as the test cases, and the precision benchmark of Hylleraas' explicitly correlated calculation [628]; forward pointer to Chapter 88 for atoms and molecules and to Chapter 126 for the extended case.]

83.4 The exchange interaction

83.4.1 Exchange energy has no classical analogue

[Reserved: the two-electron matrix element of the Coulomb repulsion between symmetrized and antisymmetrized states splits into a direct term J and an exchange term K, with the sign of K entering with the exchange parity; Heisenberg's resonance treatment [545] and Dirac's derivation of the same splitting from antisymmetry [537]; the essential point that no new interaction has been introduced — the exchange term is the ordinary electrostatic interaction evaluated on a constrained state space, which is why it is large, of order electronvolts, while the magnetic dipole–dipole energy between the same two spins is smaller by some three orders of magnitude.]

83.4.2 Helium: ortho and para

[Reserved: the worked example of the chapter. The two-electron state factorizes into a spatial and a spin part whose exchange parities must multiply to -1: the spin singlet takes the symmetric spatial function (parahelium) and the spin triplet the antisymmetric one (orthohelium). Consequences to be derived: the absence of a triplet ground state; the $2\,^3S$–$2\,^1S$ splitting of about $0.80\,\mathrm{eV}$, entirely exchange in origin; the near-absence of intercombination lines, and hence the nineteenth-century appearance of helium as two distinct gases; the metastability of $2\,^3S$. Perturbative treatment via Chapter 82, variational benchmark via [628], and modern spectroscopic precision via Chapter 90.]

Phenomenon 83.2 (Helium has no triplet ground state)**.** The spectrum of helium falls into two nearly independent systems, singlet and triplet, with transitions between them so weak that the nineteenth century took them for the spectra of two different gases. The lowest term of the entire spectrum is a singlet: no triplet state of the $1s^2$ configuration exists at all, though the corresponding singlet does. Where both do exist — the $1s2s$ configuration — the triplet lies lower, by about $0.80\,\mathrm{eV}$. That separation cannot be magnetic: the magnetic interaction of two electron moments a tenth of a nanometre apart is of order $5\times10^{-5}\,\mathrm{eV}$, four orders of magnitude too small. Heisenberg and Dirac identified it instead as an ordinary electrostatic energy, evaluated on states of definite exchange symmetry [537, 545].

Derivation. Electrons are fermions, so the two-particle state changes sign under the exchange of the pair of labels carrying both position and spin. Neglecting spin-dependent forces, the Hamiltonian does not couple space to spin and a state may be taken as a product $\Psi = \psi(\boldsymbol{r}_1,\boldsymbol{r}_2)\,\chi(1,2)$ of a spatial and a spin factor, each of definite exchange parity. Antisymmetry

of the product then requires the two parities to be opposite:

$$\begin{aligned} \psi \text{ symmetric} &\iff \chi \text{ antisymmetric: the spin singlet}\,, \\ \psi \text{ antisymmetric} &\iff \chi \text{ symmetric: the spin triplet}\,. \end{aligned} \tag{83.2}$$

For the configuration $1s^2$ both electrons occupy the same spatial orbital, so $\psi = \varphi_{1s}(\boldsymbol{r}_1)\,\varphi_{1s}(\boldsymbol{r}_2)$ is necessarily symmetric, and by Equation (83.2) the spin state must be the antisymmetric singlet. A triplet ground state is not merely higher in energy; there is no such state to have an energy.

The sign of the $1s2s$ splitting follows from the same parity. An antisymmetric ψ vanishes wherever $\boldsymbol{r}_1 = \boldsymbol{r}_2$ and so keeps the two electrons apart, while a symmetric one does not — indeed it is enhanced there. Evaluating the Coulomb repulsion $e^2/4\pi\varepsilon_0\,|\boldsymbol{r}_1 - \boldsymbol{r}_2|$, a positive operator, on the two states therefore costs more in the symmetric case, so the singlet lies above the triplet, as observed. No spin operator appeared anywhere in the argument: the spin merely selected which spatial function was admissible, and the energy difference is electrostatic. That is why it is of the order of an electronvolt while the magnetic interaction between the same two moments is smaller by four orders of magnitude, and it is the mechanism carried over to magnetic order in Section 83.4.3. □

83.4.3 From exchange to magnetic order

[Reserved: Dirac's replacement of the exchange term by an effective spin–spin coupling $-2K\,\boldsymbol{S}_1 \cdot \boldsymbol{S}_2$ [537], and Heisenberg's use of it to explain ferromagnetism as a cooperative alignment driven by electrostatics [605]; the sign of K selecting ferromagnetic or antiferromagnetic order; that the Curie temperatures of iron, cobalt and nickel are of the electronvolt scale precisely because the mechanism is Coulomb and not magnetic; Hund's first rule as the atomic case; full development in Chapter 129, and the covalent bond as the same computation with the opposite sign convention in Chapter 88.]

83.5 The spin–statistics connection

83.5.1 Statement of the theorem

[Reserved: the empirical regularity that half-integer-spin particles are fermions and integer-spin particles are bosons, and its elevation to a theorem by Pauli [854] on the hypotheses of relativistic invariance, positivity of the energy and locality (the vanishing of commutators of observables at spacelike separation); the sharpened axiomatic version of Lüders and Zumino [1151]. The theorem is stated *here and proved in Chapter 109, because none of its hypotheses is available in nonrelativistic quantum mechanics — which is itself the finding: the symmetrization postulate of Section 83.1.4 is a genuine extra postulate at this level of the theory and stops being one only in a relativistic field theory. Its companion, the CPT theorem, is in Chapter 105. Composite particles inherit the statistics of the parity of their constituent fermion number, which is why ^{4}He condenses and ^{3}He must pair before it can (Chapter 130).]*

83.5.2 Experimental tests of the symmetrization postulate

[Reserved: the postulate is testable and has been tested. The Ramberg–Snow experiment passed a large current through a copper strip and searched for the X-ray line that a new electron making a Pauli-forbidden transition to a filled K shell would emit, bounding the violation probability at the level of 10^{-26} [1837]; the underground VIP-2 continuation improves it [2405]. Complementary bounds come from the non-observation of symmetric two-photon states and from nuclear spectroscopy. The section will tabulate the quantity bounded, the

bound, and the assumption each experiment tests, in the sense of Chapter 1: the point is that "electrons are identical" is a measured statement, not a definition.]

Phenomenon 83.3 (No electron has been caught violating exclusion)**.** The symmetrization postulate is a testable statement about the world and it has been tested directly. Ramberg and Snow passed a large current through a thin copper strip and searched its X-ray spectrum for the line that a newly arrived electron would emit on making a transition into an already filled K shell — a transition forbidden by exclusion, and displaced in energy from the ordinary $K\alpha$ line by the extra screening that the doubly occupied shell provides. No such line appeared above background, which bounds the probability that a fresh electron enters a filled state at the level of 10^{-26} [1837]. The VIP-2 apparatus, the same experiment moved underground so that cosmic rays no longer set the background, has improved the limit [2405]. That "electrons are identical" is therefore an experimental result carrying a number, and not a convention of the formalism.

Derivation pending #96

The Ramberg–Snow bound: the energy displacement of the anomalous X-ray line produced by the extra screening of a doubly occupied inner shell, the number of fresh electrons that a given charge delivers into the target lattice, and the conversion of a null count and its background into an upper limit on the violation probability

83.6 Two-particle interference

83.6.1 The Hong–Ou–Mandel dip

[Reserved: two indistinguishable photons entering the two input ports of a balanced beam splitter always leave together, because the two amplitudes for coincident detection cancel for bosons; the measurement of the resulting coincidence dip by Hong, Ou and Mandel [1794], whose width measures the photon wave-packet duration on a subpicosecond scale and whose depth measures indistinguishability directly. The derivation is a two-line calculation with the bosonic commutators of Section 83.3.2 and belongs inline. Uses: it is the standard indistinguishability metric in Chapter 122, and the entanglement-swapping step of Chapter 84 depends on it.]

Phenomenon 83.4 (Identical photons leave a beam splitter together)**.** Send one photon into each of the two input ports of a balanced beam splitter. If the two are made indistinguishable — same polarization, same spectrum, arrival times matched — the detectors at the two output ports never fire together: the coincidence rate falls to zero. Delaying one photon by more than its wave-packet duration restores the coincidences to the value expected of two independent particles, so the coincidence rate plotted against delay shows a dip. Hong, Ou and Mandel measured that dip and read the photon wave-packet duration off its width, on a subpicosecond scale that no detector of theirs could have resolved directly [1794]. Each photon separately divides evenly between the ports; only the pair refuses to.

Derivation. Let $a^\dagger$ and $b^\dagger$ create a photon in the two input modes and $c^\dagger$, $d^\dagger$ in the two output modes. A lossless balanced beam splitter acts on the creation operators as

$$a^\dagger \to \frac{c^\dagger + d^\dagger}{\sqrt{2}}, \qquad b^\dagger \to \frac{c^\dagger - d^\dagger}{\sqrt{2}},$$

the relative minus sign being forced by unitarity: without it the map would not preserve

the commutators. The input with one photon in each port is $a^\dagger b^\dagger |0\rangle$, so

$$\begin{aligned} a^\dagger b^\dagger |0\rangle &\to \tfrac{1}{2}\left(c^\dagger + d^\dagger\right)\left(c^\dagger - d^\dagger\right)|0\rangle \\ &= \tfrac{1}{2}\left(c^\dagger c^\dagger - d^\dagger d^\dagger + d^\dagger c^\dagger - c^\dagger d^\dagger\right)|0\rangle \ . \end{aligned} \tag{83.3}$$

For bosons $\left[c^\dagger, d^\dagger\right] = 0$, so the last two terms of Equation (83.3) cancel identically and, since $c^\dagger c^\dagger |0\rangle = \sqrt{2}\,|2_c, 0_d\rangle$,

$$a^\dagger b^\dagger |0\rangle \to \frac{1}{\sqrt{2}}\left(|2_c, 0_d\rangle - |0_c, 2_d\rangle\right) \ . \tag{83.4}$$

Every surviving term of Equation (83.4) has both photons in one port: the amplitude for one photon in each output is exactly zero, which is the observed vanishing of the coincidence rate. The cancellation is between the two histories that end in the same final state — both photons reflected, both transmitted — and it survives only while nothing distinguishes them, so anything that labels the photons, a difference in arrival time included, fills the dip in. Its depth is therefore a direct measurement of indistinguishability.

Repeating the calculation for fermions changes one thing and reverses the conclusion. The anticommutator $\left\{c^\dagger, d^\dagger\right\} = 0$ makes the cross terms add rather than cancel, while $c^\dagger c^\dagger = d^\dagger d^\dagger = 0$ kills the terms that survived above, leaving $-c^\dagger d^\dagger |0\rangle$: the two particles leave in *opposite* ports with certainty. Exchange parity is read directly off the sign of a two-particle coincidence rate. □

83.6.2 Intensity correlations and the fermionic counterpart

[Reserved: the Hanbury Brown–Twiss observation that photons from a thermal source arrive in correlated pairs [1083], its initial reception as a paradox, and its resolution as bosonic bunching — an interference of two-particle amplitudes with no one-particle analogue; the corresponding antibunching of electrons measured in a mesoscopic beam splitter [1975]; the sign of the correlation as a direct readout of exchange parity, and its use as a source-size measurement in stellar intensity interferometry and in heavy-ion collisions.]

Phenomenon 83.5 (Bosons bunch, fermions antibunch)**.** Photons from a thermal source do not arrive independently of one another. Splitting the light of a source between two photodetectors and correlating their output currents, Hanbury Brown and Twiss found an excess of coincidences at zero delay over the rate that statistically independent arrivals would give, decaying away over the coherence time of the light [1083]. The result was disputed on the ground that photons detected at separate places should know nothing of each other, and it is nevertheless there. The same measurement made on electrons leaving a mesoscopic beam splitter gives the opposite sign, a *deficit* of coincidences [1975]. The correlation is a two-particle effect with no one-particle counterpart: neither detector alone shows anything unusual, and the sign of the correlation between them reports the exchange parity of the particles directly.

Derivation pending #97

The two-particle correlation of a chaotic source: the sum over the two indistinguishable histories by which one particle from each of two source points reaches the two detectors, the factor of two it produces in the zero-delay coincidence rate for symmetric states, and its replacement by zero for antisymmetric ones, together with the loss of the effect once the path difference exceeds the coherence length

83.7 Consequences for bulk matter

83.7.1 Stability of matter

[Reserved: that the ground-state energy of N electrons and N nuclei is bounded below by a constant times N — so that bulk matter neither collapses nor releases unbounded energy on assembly — is a theorem, and it is false without the exclusion principle. Dyson and Lenard proved it first [1345]; Lieb and Thirring gave the proof that yields a realistic constant, via their kinetic-energy inequality [1554]. The section states the theorem, records that bosonic "matter" would instead have energy growing like $-N^{7/5}$, and refers the proof to Appendix A. This is the sharpest answer available to the question of why atoms have a size and why matter is extended and nearly incompressible.]

83.7.2 Degeneracy pressure

[Reserved: the pressure of a degenerate Fermi gas at zero temperature, derived from the exclusion principle alone; its nonrelativistic and ultrarelativistic forms and the resulting Chandrasekhar limit of about 1.4 *solar masses for a white dwarf [676]; the neutron-star analogue; the same mechanism, at laboratory densities, as the incompressibility of solids and the bulk modulus of a metal. Astrophysical development in Chapters 49 and 52, thermodynamic development in Chapter 120.]*

Phenomenon 83.6 (White dwarfs, and a limiting mass)**.** There exist stars whose mass is of the order of the Sun's and whose radius is of the order of the Earth's, so that their mean density exceeds that of ordinary matter by some six orders of magnitude. They have exhausted their nuclear fuel, they radiate, and they do not contract; and as they cool they still do not contract, so whatever supports them is not thermal pressure. What supports them is the pressure of a degenerate electron gas, which is a consequence of the exclusion principle alone and does not vanish at zero temperature. There is also a ceiling. No white dwarf is observed above about 1.4 times the mass of the Sun, and Chandrasekhar showed that none can exist, because at the densities a larger mass would require the electrons move relativistically and the equation of state softens until it can no longer hold the star up [676].

Derivation pending #98

The degenerate electron gas: its zero-temperature pressure obtained from the exclusion principle and the filling of momentum space up to the Fermi momentum, the change of the exponent relating pressure to density once that momentum exceeds the electron mass times the speed of light, and the resulting limiting mass, which is of the order of the cube of the Planck mass divided by the square of the nucleon mass

Chapter 84

Entanglement and Bell Tests

The tensor-product structure of composite systems was written down in Chapter 77 without comment. Its consequences are the subject of this chapter, and they are the least classical thing in physics. Almost every state of a composite system fails to factorize; Schrödinger called such states entangled and said of them, correctly, that they are not one but *the* characteristic trait of quantum mechanics [760]. Einstein, Podolsky and Rosen had just argued from the same states that the quantum description must be incomplete [753]. For thirty years the dispute was regarded as philosophical. It stopped being philosophical in 1964, when Bell showed that *any* theory in which the outcomes are fixed by local variables carried by the particles obeys an inequality that quantum mechanics violates [1271]. The question became experimental, and it has been answered.

This chapter develops the theory: what entanglement is, what Bell's theorem does and does not prove, what quantum mechanics may not do with entanglement (signal, clone, share it freely), and how much of it a state contains. The experimental record — Freedman and Clauser, Aspect, and the loophole-free tests of 2015 — lives in Chapter 89, and the reader who wants the evidence before the formalism should read that chapter first. What this chapter insists on is the logic connecting the two: a measured violation excludes a precisely specified class of theories and nothing wider, and the treatise's position on the wider claims is stated in Chapter 85. Standard references are [1788, 1872, 2161]; the modern review is [2148, 2232].

Derivation pending #99

Entanglement and Bell Tests: all derivations of this chapter are pending.

84.1 Entangled states

84.1.1 Tensor products and separability

[Reserved: the composite Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$ and the dimension count that makes product states a measure-zero subset; the definition of a pure entangled state as one that is not a product; the extension to mixed states, where separable means a convex combination of products, due to Werner [1829], together with his example of a mixed state that is entangled yet admits a local hidden-variable model for projective measurements — so that entanglement and Bell violation are not *the same property; the two-qubit Bell basis as the working example throughout.]*

84.1.2 The Schmidt decomposition

[Reserved: every pure bipartite state can be written $\sum_i \sqrt{p_i}\,|a_i\rangle \otimes |b_i\rangle$ with orthonormal local bases, the coefficients being the singular values of the coefficient matrix; the underlying decomposition is Schmidt's [338] and the linear algebra is that of Chapter 5; the Schmidt rank as the first, coarsest measure of entanglement; that the two reduced states share their spectrum, which is why bipartite entanglement entropy is unambiguous.]

84.1.3 Reduced states, purification and steering

[Reserved: the partial trace and the reduced density matrix as the unique object reproducing all local statistics; entanglement as the statement that the parts are mixed while the whole is pure; purification and its essential uniqueness up to a local unitary. Schrödinger's observation that the distant party can steer the local ensemble into any decomposition of the reduced state [759], made precise as the Hughston–Jozsa–Wootters theorem [1868]; that steering changes the ensemble description but not a single measurable local probability, which is the seed of Section 84.4.1.]

84.2 The EPR argument and the replies

84.2.1 Einstein, Podolsky and Rosen

[Reserved: the 1935 argument in its own terms [753]: the sufficient criterion of reality, the locality assumption, the position–momentum correlated two-particle state, and the conclusion that either the description is incomplete or reality depends on what is measured elsewhere; a careful statement of the premises, since almost every later dispute is about which of them a Bell test refutes; Bohr's reply of the same year, which denied the criterion of reality rather than the locality assumption; the reading of the paper Einstein himself preferred, given in his later correspondence, in which the target is separability rather than completeness.]

84.2.2 Verschränkung, the cat, and the amplification argument

[Reserved: Schrödinger's response in Die Naturwissenschaften *[760], which introduced the term* Verschränkung*, framed entanglement as the failure of the whole to be determined by its parts, and used the cat as an* amplification *argument — that nothing in the linear dynamics stops a microscopic superposition from being coupled to a macroscopic degree of freedom; the companion Cambridge paper with the steering theorem [759]; the point, still worth making, that the cat is an argument about the reach of linearity and not a joke about observers, and that its modern quantitative version is the interference of massive molecules and the collapse bounds of Chapter 85.]*

84.2.3 Bohm's spin version

[Reserved: Bohm's replacement of the continuous EPR variables by two spin-$\frac{1}{2}$ particles in the singlet [978], which is what made the argument experimentally tractable: a finite-dimensional Hilbert space, a rotationally invariant state, and the correlation $-\cos\theta$ between spin components along directions at angle θ, with the spin formalism of Chapter 79 and the analyser of Chapter 80. Every subsequent inequality and every experiment uses this version.]

Phenomenon 84.1 (Singlet correlations)**.** Two spin-$\frac{1}{2}$ particles prepared in the singlet state — and the polarization-entangled photon pairs that stand in for them in the laboratory — display two properties at once. Each particle taken by itself is unpolarized: whatever direction $\hat{\boldsymbol{a}}$ its analyser is set to, the outcomes ± 1 occur with equal frequency, so a one-sided

record carries no trace of the setting used on the other side. Taken in pairs the outcomes are correlated, and the correlation depends on nothing but the angle θ between the two analyser directions,

$$E(\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}) = \langle \Psi^- | \, (\boldsymbol{\sigma}\cdot\hat{\boldsymbol{a}}) \otimes (\boldsymbol{\sigma}\cdot\hat{\boldsymbol{b}}) \, | \Psi^- \rangle = -\hat{\boldsymbol{a}}\cdot\hat{\boldsymbol{b}} = -\cos\theta\,, \tag{84.1}$$

so that parallel analysers give opposite results on every single pair while orthogonal analysers give no correlation at all. This is the Einstein–Podolsky–Rosen situation [753] in the two-outcome form that made it experimentally decidable, and Equation (84.1) is the curve against which every measurement of Chapter 89 is read.

Derivation. Write $|\Psi^-\rangle = (|{\uparrow\downarrow}\rangle - |{\downarrow\uparrow}\rangle)/\sqrt{2}$, the unique two-spin state annihilated by the total spin $\boldsymbol{S} = \frac{\hbar}{2}(\boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2)$. From $(\boldsymbol{\sigma}_1+\boldsymbol{\sigma}_2)^2\,|\Psi^-\rangle = 0$ and $\boldsymbol{\sigma}^2 = 3\,\mathbb{1}$ on each factor, $6 + 2\,\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_2$ annihilates $|\Psi^-\rangle$, so

$$\langle \Psi^- | \, \boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_2 \, | \Psi^- \rangle = -3\,. \tag{84.2}$$

The singlet carries zero total angular momentum and is therefore invariant under simultaneous rotations of both factors. Consequently the tensor $T_{ij} = \langle \Psi^- | \, \sigma_1^i \sigma_2^j \, | \Psi^- \rangle$ is an invariant tensor of SO(3) of rank two, and the only such tensor is δ_{ij} up to a factor: $T_{ij} = \lambda\,\delta_{ij}$. Taking the trace and using Equation (84.2) gives $3\lambda = -3$, hence $\lambda = -1$ and

$$E(\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}) = a_i b_j T_{ij} = -\hat{\boldsymbol{a}}\cdot\hat{\boldsymbol{b}}\,,$$

which is Equation (84.1). For the one-sided statistics, the reduced state is $\rho_A = \mathrm{tr}_B\,|\Psi^-\rangle\langle\Psi^-| = \frac{1}{2}\mathbb{1}$, so $\mathrm{tr}(\rho_A\,\boldsymbol{\sigma}\cdot\hat{\boldsymbol{a}}) = 0$ for every $\hat{\boldsymbol{a}}$: each analyser on its own sees a fair coin, which is what makes the correlation invisible until the two records are compared. □

84.3 Bell's theorem

84.3.1 Local hidden variables and the 1964 inequality

[Reserved: the precise hypotheses — a shared variable λ with some distribution, outcome functions $A(a,\lambda)$ and $B(b,\lambda)$ depending only on the local setting (parameter independence) and not on the distant outcome (outcome independence), and settings statistically independent of λ (measurement independence) — and the derivation of Bell's original three-setting inequality on the perfect-anticorrelation singlet [1271]; the quantum prediction $-\cos\theta$ violating it; the fact, essential to the argument's force, that no feature of the quantum formalism is used in deriving the bound.]

84.3.2 CHSH: the form that is actually measured

[Reserved: the four-correlator combination $S = E(a,b) + E(a,b') + E(a',b) - E(a',b')$ with $|S| \le 2$ for any local model, due to Clauser, Horne, Shimony and Holt [1400], which drops the perfect-correlation idealization and so applies to real detectors; the Clauser–Horne form written for detection probabilities rather than correlations [1522]; Fine's theorem, that satisfying the full set of CHSH inequalities is equivalent to the existence of a joint distribution for all four observables [1687], which identifies exactly what a violation refutes; the quantum maximum $2\sqrt{2}$ at the optimal analyser angles; Gisin's theorem that every entangled pure state violates some CHSH inequality [1845].]

Phenomenon 84.2 (Violation of the CHSH bound)**.** Let each station of a two-particle experiment choose between two analyser settings, recording an outcome ± 1, and let $E(a,b)$ be the measured correlation for the setting pair (a,b). The combination

$$S = E(a,b) + E(a,b') + E(a',b) - E(a',b') \tag{84.3}$$

cannot exceed 2 in modulus in any theory in which each outcome is fixed by variables carried by the particles and is unaffected by the setting chosen at the distant station [1271]. Values of S above 2 are what is measured. They were first reported from atomic-cascade photon pairs in the early 1970s, sharpened with two-channel polarizers and with analysers switched while the photons were in flight in the early 1980s, and have since been reproduced with independent sources, analysers and detector technologies; the record, with its uncertainties and its residual assumptions, is Chapter 89. Nature therefore admits no description of the assumed kind.

Derivation. The bound is the whole of the argument, and it uses no quantum mechanics whatever. Suppose the outcomes are functions $A(a,\lambda)$, $A(a',\lambda)$, $B(b,\lambda)$, $B(b',\lambda)$ taking the values ±1, each depending on the setting on its own side and on a variable λ distributed with some density $p(\lambda) \geq 0$, $\int \mathrm{d}\lambda\, p(\lambda) = 1$, so that $E(a,b) = \int \mathrm{d}\lambda\, p(\lambda)\, A(a,\lambda) B(b,\lambda)$. For a fixed λ collect the four products and factor,

$$S(\lambda) = A\big[B + B'\big] + A'\big[B - B'\big]\,, \tag{84.4}$$

suppressing the arguments. Since B and B' each take only the values ±1, either $B = B'$, in which case $B - B' = 0$ and $|B + B'| = 2$, or $B = -B'$, in which case $B + B' = 0$ and $|B - B'| = 2$. In both cases exactly one bracket vanishes and the other has modulus 2; with $|A| = |A'| = 1$ this gives $|S(\lambda)| = 2$ for every λ. Averaging preserves the bound,

$$|S| = \left| \int \mathrm{d}\lambda\, p(\lambda)\, S(\lambda) \right| \leq \int \mathrm{d}\lambda\, p(\lambda)\; |S(\lambda)| = 2\,.$$

Quantum mechanics is not a theory of that form and is not bound by it. Insert the singlet correlation Equation (84.1) with coplanar analyser directions at angles $a = 0$, $a' = \pi/2$, $b = \pi/4$ and $b' = -\pi/4$: then $E(a,b) = E(a,b') = E(a',b) = -1/\sqrt{2}$ while $E(a',b') = -\cos(3\pi/4) = +1/\sqrt{2}$, and Equation (84.3) gives $S = -4/\sqrt{2} = -2\sqrt{2}$. The predicted excess over the local bound is a factor $\sqrt{2}$, not a marginal effect — which is why the question could be settled at all with the photomultipliers of a few percent quantum efficiency available when it was first asked. □

84.3.3 The quantum bound

[Reserved: Tsirelson's proof that quantum mechanics itself caps $|S|$ at $2\sqrt{2}$ [1650], so that the theory is nonlocal in Bell's sense but not maximally so; the Popescu–Rohrlich box, a hypothetical device that reaches the algebraic maximum 4 without permitting signalling [1886], which shows that no-signalling alone does not single out quantum correlations and turns "why $2\sqrt{2}$?" into a live question with several proposed information-theoretic answers; the experimental status, namely that no measured correlation has ever exceeded the Tsirelson bound.]

Phenomenon 84.3 (The data stop at $2\sqrt{2}$)**.** No experiment has produced a CHSH parameter above $2\sqrt{2}$. This is not forced by relativity: the algebraic maximum of Equation (84.3) is 4, and correlation tables reaching 4 can be written down that leave both marginals uniform and so permit no signalling. Nor is it forced by the existence of a violation, which only requires $|S| > 2$. What is observed is that the correlations of Nature stop precisely where the quantum formalism says they stop, so that the measured $|S|$ is bounded above and below by two independent theoretical statements and lies strictly between them [2253, 2254, 2260].

Derivation. Represent the two alternative settings on one side by Hermitian operators A, A' and those on the other side by B, B'. Outcomes ±1 mean $A^2 = A'^2 = B^2 = B'^2 = \mathbb{1}$, hence

$\|A\| = \|A'\| = \|B\| = \|B'\| = 1$; operators belonging to different tensor factors commute. The CHSH observable is $\hat{S} = A \otimes (B + B') + A' \otimes (B - B')$, and $S = \mathrm{tr}(\rho \hat{S})$. Squaring, and using $(B \pm B')^2 = 2\,\mathbb{1} \pm (BB' + B'B)$ together with $(B + B')(B - B') = [B', B]$ and $(B - B')(B + B') = [B, B']$, the anticommutators cancel between the two diagonal terms and the cross terms combine into a single commutator product,

$$\hat{S}^2 = 4\,\mathbb{1} - [A, A'] \otimes [B, B'] \ . \tag{84.5}$$

Each commutator obeys $\|[A, A']\| \leq 2\,\|A\|\,\|A'\| = 2$, so $\left\|\hat{S}^2\right\| \leq 4 + 2 \cdot 2 = 8$ and therefore

$$|S| \leq \left\|\hat{S}\right\| = \sqrt{\left\|\hat{S}^2\right\|} \leq 2\sqrt{2}\,,$$

for every state and every choice of settings. Note where the bound comes from. If the two alternative settings at a station commuted, Equation (84.5) would read $\hat{S}^2 = 4\,\mathbb{1}$ and give back $|S| \leq 2$, the local bound. The whole violation, and its exact size, is therefore the non-commutativity of the two measurements a single station chooses between — not any influence travelling between the stations. □

84.3.4 GHZ: contradiction without inequalities

[Reserved: the three-particle state for which local realism and quantum mechanics disagree on a single *run rather than on an average, so that the refutation is algebraic rather than statistical [1821], in the accessible presentation of [1832]; the sign contradiction obtained from four perfectly correlated measurement settings; the Mermin inequality as its statistical version and its exponentially growing violation with particle number; the experimental realization belongs to Chapter 89.]*

84.3.5 What a violation excludes — and what it does not

[Reserved: the disciplined statement, which this treatise treats as the whole point of the chapter. A measured violation, closing the locality and detection loopholes, refutes the conjunction of the hypotheses of Section 84.3.1. It does not *establish superluminal causation, since Section 84.4.1 forbids it; it does not select any interpretation of Chapter 85; it does not refute determinism as such, only local determinism. The residual assumptions — measurement independence, or "freedom of choice" — are not removable in principle, and the cosmic-Bell experiments of Chapter 89 only push the correlation back in time. The honest summary is that local hidden-variable theories are excluded as a class, which is a statement about theories and is the strongest one the data support.]*

84.4 What entanglement may not do

84.4.1 No-signalling

[Reserved: the theorem that the local statistics at A are independent of the setting chosen at B, because the reduced state is unchanged by any local operation at B — proved directly from the partial trace and the trace-preserving property of local channels; the argument as given by Ghirardi, Rimini and Weber [1653] and the careful analysis of the several distinct locality conditions by Eberhard [1619]; the consequence that entanglement carries no message and that relativistic causality of Chapter 39 is intact; that this is a theorem about quantum mechanics, not a postulate imposed on it.]

Phenomenon 84.4 (No signal accompanies the correlation)**.** In every experiment on entangled pairs, the singles rate recorded at one station is independent, within statistics, of

which setting the distant station chose. The check is made directly in the data of a Bell test as a control on the apparatus [2253, 2254, 2260], and it has never failed. However strongly the two records are correlated, neither party can send the other a message by choosing a setting; the correlation appears only when the two records are brought together and compared, which requires a classical channel and therefore takes at least the light travel time between the stations.

Derivation. Everything one party can measure is computed from the reduced state $\rho_A = \operatorname{tr}_B \rho$. Let the distant party act on their factor alone: the most general such operation is a trace-preserving map $\mathcal{E}_B$, and choosing a setting and performing a measurement whose outcome is not communicated is one instance of it. The joint state becomes $(\mathrm{id}_A \otimes \mathcal{E}_B)\rho$. Then for any observable X of the first factor,

$$\operatorname{tr}\big[(X \otimes \mathbb{1})\,(\mathrm{id}_A \otimes \mathcal{E}_B)\rho\big] = \operatorname{tr}\big[X\ \operatorname{tr}_B \rho\big] = \operatorname{tr}\big[(X \otimes \mathbb{1})\,\rho\big]\,, \tag{84.6}$$

because $\operatorname{tr}_B\big[(\mathrm{id}_A \otimes \mathcal{E}_B)\rho\big] = \operatorname{tr}_B \rho$ whenever $\mathcal{E}_B$ preserves the trace. Hence ρ_A, and with it every local probability, is untouched by anything done at B. The argument uses only linearity and trace preservation, so it holds for every state, every setting and every separation. No-signalling is thus a theorem of the formalism, not a relativistic postulate bolted onto it — which is why Phenomenon 84.2 creates no conflict with the causal structure of Chapter 39. □

84.4.2 No-cloning

[Reserved: the impossibility of a unitary that copies an unknown state, proved in two lines from linearity, given independently by Wootters and Zurek [1708] and by Dieks [1686], both motivated by closing a proposed superluminal-signalling scheme; the no-broadcasting and no-deleting extensions; the constructive consequences — unconditional security of quantum key distribution [1843] and the necessity of teleportation rather than copying in Section 84.6.1; the relation to the distinguishability of non-orthogonal states.]

Phenomenon 84.5 (Unknown states are not copied)**.** No apparatus has ever been built that accepts one system in an unknown state and returns two systems in that same state. The regularity shows itself wherever a quantum signal is amplified: a laser amplifier, an optical repeater or an eavesdropper's tap always adds noise, and the minimum noise it adds is exactly the amount required to stop the output being a faithful pair of copies. It is the reason a key carried by single quanta cannot be intercepted and forwarded undetected, and the reason that the transfer of an unknown state in Section 84.6.1 must destroy the original.

Derivation. Suppose a unitary U on system plus blank register satisfied $U\big(|\psi\rangle \otimes |0\rangle\big) = |\psi\rangle \otimes |\psi\rangle$ for every state $|\psi\rangle$ of the system. Apply it to two states $|\psi\rangle$ and $|\phi\rangle$ and take the inner product of the two results. A unitary preserves inner products, and the inner product of a tensor product is the product of the inner products, so with $z = \langle\psi|\phi\rangle$,

$$z = \langle\psi|\phi\rangle \cdot \langle\psi|\phi\rangle = z^2\,, \tag{84.7}$$

whence $z \in \{0, 1\}$: the only families a single unitary can copy are sets of mutually orthogonal states, which is to say classical alphabets, since a state drawn from an orthogonal set can be measured without disturbance and then rewritten at will. Nothing but linearity and unitarity was used, so the conclusion is exactly as robust as the superposition principle of Chapter 77; in particular it cannot be evaded by a cleverer device. □

84.4.3 Monogamy

[Reserved: entanglement cannot be freely shared. The Coffman–Kundu–Wootters inequality bounds the sum of the squared concurrences of A with B and with C by the squared concurrence

of A with the pair, the deficit defining the three-tangle [1986]; the extreme case, that a maximally entangled pair is uncorrelated with everything else, which is exactly what makes device-independent key distribution possible; the general N-party statement and the role of monogamy in bounding how many parties can jointly violate a Bell inequality.]

84.5 Quantifying entanglement

84.5.1 Separability criteria

[Reserved: the positive-partial-transpose necessary condition of Peres [1928] and the Horodecki proof that it is also sufficient in dimensions 2×2 and 2×3, via the theory of positive but not completely positive maps [1923]; bound entangled states in higher dimensions, which are entangled yet have positive partial transpose and cannot be distilled; entanglement witnesses as the operational form of the criterion, being the observables a laboratory actually measures.]

84.5.2 Entropy of entanglement

[Reserved: for a pure bipartite state, the von Neumann entropy of the reduced state, $S = -\operatorname{tr} \rho_A \log \rho_A$, as the unique measure satisfying the natural axioms; its operational meaning as the asymptotic rate of Bell-pair distillation and of entanglement dilution [1915]; the maximum $\log d$ and the maximally entangled states attaining it; area laws in extended systems and the pointer to Chapter 109 for the field-theoretic case, where the same quantity is ultraviolet divergent and its regularization is a substantive physical question.]

84.5.3 Mixed-state measures

[Reserved: entanglement of formation and distillable entanglement and the gap between them [1915]; Wootters' closed formula for the entanglement of formation of an arbitrary two-qubit state via the concurrence [1964]; the negativity and logarithmic negativity, computable from the partial transpose [2040]; the axioms an entanglement monotone must satisfy — invariance under local unitaries, non-increase under local operations and classical communication — and the fact that no single number orders mixed states completely; the review [2148] as the reference for the full landscape.]

84.6 Entanglement as a resource

84.6.1 Teleportation and entanglement swapping

[Reserved: the protocol of Bennett and co-workers [1860], in which a shared Bell pair plus two classical bits transfer an unknown qubit exactly, destroying the original — and so respecting both Section 84.4.2 and Section 84.4.1, since without the classical message the output is the maximally mixed state; the first photonic realization [1935]; entanglement swapping, which entangles two systems that never interacted and which relies on the two-photon interference of Chapter 83; the resource accounting that makes "one ebit" a physical unit.]

Phenomenon 84.6 (Teleportation of an unknown state)**.** A qubit in a state unknown to everyone involved can be transferred from one system to a second system that never interacted with it, using a previously shared entangled pair and two bits sent over an ordinary classical channel. What is observed is a reconstructed state whose fidelity with the input exceeds 2/3, the highest that any strategy of measuring the input and transmitting a classical description of the result can achieve; the original is destroyed in the process. The protocol has been realized with photon polarization [1935], with trapped ions and atoms,

and with superconducting circuits, and the transfer succeeds only after the classical message arrives.

Derivation. Let the unknown state be $|\psi\rangle = a\,|0\rangle + b\,|1\rangle$ on system 1, and let systems 2 and 3 share $|\Phi^+\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$, with 1 and 2 held by the sender and 3 by the receiver. Expand the product $|\psi\rangle_1 \otimes |\Phi^+\rangle_{23}$ in the Bell basis of the pair $(1,2)$, using $|00\rangle = (|\Phi^+\rangle + |\Phi^-\rangle)/\sqrt{2}$ and its three companions. Collecting terms,

$$|\psi\rangle_1 \otimes \left|\Phi^+\right\rangle_{23} = \tfrac{1}{2}\Big[\left|\Phi^+\right\rangle_{12} \otimes (a\,|0\rangle + b\,|1\rangle)_3 + \left|\Phi^-\right\rangle_{12} \otimes (a\,|0\rangle - b\,|1\rangle)_3 + \left|\Psi^+\right\rangle_{12} \otimes (a\,|1\rangle + b\,|0\rangle)_3 + \left|\Psi^-\right\rangle_{12} \otimes (a\,|1\rangle - b\,|0\rangle)_3\Big]. \tag{84.8}$$

The four Bell states are orthonormal and enter with equal weight, so a measurement of $(1,2)$ in that basis yields each of the four outcomes with probability $1/4$ — a distribution independent of a and b, which is why the sender learns nothing about the state and why the outcome, taken alone, is useless. Conditioned on the outcome, system 3 is left in $\mathbb{1}\,|\psi\rangle$, $\sigma_z\,|\psi\rangle$, $\sigma_x\,|\psi\rangle$ or $\sigma_x\sigma_z\,|\psi\rangle$ respectively, each of which the receiver undoes by applying the inverse of that operator once the two bits naming the outcome have arrived. The correction is a fixed unitary chosen from a list of four; no knowledge of a and b enters it.

Three consistency checks close the argument. Before the message arrives, the receiver's state is the average of the four branches, $\frac{1}{4}\sum_k \sigma_k\,|\psi\rangle\langle\psi|\,\sigma_k = \frac{1}{2}\mathbb{1}$, independent of $|\psi\rangle$ — so nothing has been signalled, in accordance with Phenomenon 84.4, and the classical channel really is indispensable. The input system emerges maximally entangled with the sender's half of the pair and no longer carries $|\psi\rangle$, so nothing has been copied, in accordance with Phenomenon 84.5. And the two classical bits are not a disguised description of $|\psi\rangle$, which would need infinitely many bits: they are a description of the sender's measurement outcome, whose distribution Equation (84.8) shows to be uniform. □

84.6.2 Device-independent protocols

[Reserved: the observation that a Bell violation certifies properties of a device without any assumption about its internals — the strongest practical consequence of Section 84.3; certified randomness generation from a measured CHSH value [2163], and device-independent key distribution descending from Ekert's protocol [1843]; self-testing, whereby a maximal violation forces the state and measurements to be the singlet and the optimal analysers up to a local isometry; the honest caveat that the certification assumes no communication between the devices, which is a physical assumption about the laboratory and not a mathematical one.]

84.7 The experimental record

[Reserved: a short bridging section, not a duplicate of Chapter 89, which owns the apparatus, the data tables and the uncertainties. It records only the logical sequence: the first violation with atomic cascade photons by Freedman and Clauser [1471]; the switching experiment that first closed the locality loophole in a time-varying way [1679]; the loophole-free experiments of 2015, the first to close locality and detection simultaneously [2254]; and the resulting status of local realism as an excluded class of theories. Each entry names the loophole it addressed, so that the chapter's logical claim in Section 84.3.5 can be checked against what was actually measured.]

Chapter 85

Interpretations (Evidence-Anchored)

The formalism of Chapter 77 predicts, with no known exception, the outcome statistics of every experiment in this book. What it does not do is say what happens in a single run. The unitary evolution is linear and deterministic; the measurement postulate is nonlinear and probabilistic; and the theory as usually stated does not say which of the two applies when, nor where the boundary between system and apparatus lies. That is the measurement problem, and it is the only reason this chapter exists. It is placed after entanglement (Chapter 84) because the sharpest constraints on the question are Bell-type theorems, and before decoherence (Chapter 87) because decoherence is the piece of the answer that is not interpretation at all but ordinary quantum dynamics of an open system.

This chapter is written under a rule that the rest of the treatise applies silently and that is worth stating aloud here. *What experiment constrains is physics; what it does not constrain is interpretation, and this book will label it as such.* A great deal of the subject is genuinely empirical — Gleason's theorem [1111], Kochen–Specker contextuality [1353], Bell's analysis of von Neumann's impossibility proof [1324], the PBR theorem [2199], Leggett–Garg tests [1759], the whole decoherence programme [1440, 2055], and above all dynamical collapse models [1772], which make numerically different predictions and are being ruled out region by region [2366, 2377]. The remainder — whether the wavefunction is a thing, whether the other branches exist, whether probabilities are credences — has, at present, no experiment attached. Both kinds of statement appear below, and the ledger in Section 85.6 says which is which. Standard references are [1731, 1788, 1872]; for decoherence, [2112]; for collapse models, [2212].

Derivation pending #100

Interpretations (Evidence-Anchored): all derivations of this chapter are pending.

85.1 The measurement problem

85.1.1 The Born rule

[Reserved: Born's probabilistic reading of the wavefunction, introduced in a footnote of the first collision paper [534] and developed in the second [533]; the statement $p = |\langle a|\psi\rangle|^2$ and its generalization to positive operator-valued measures; the fact that this is the only postulate of Chapter 77 that mentions probability, and the question — taken up in Section 85.3 — of whether it can be derived from the others rather than assumed; the experimental tests of the exponent, via three-slit interference bounding higher-order terms in the Born rule, listed with

their bounds.]

85.1.2 The von Neumann chain

[Reserved: von Neumann's model of measurement as a unitary system–apparatus coupling that produces an entangled state, and the separate "process 1" that reduces it [706]; the chain that follows, in which including the next apparatus only moves the cut; the demonstration that the cut may be placed anywhere without changing predictions, which is the reason the problem is hard rather than the reason it is solved; von Neumann's impossibility proof for hidden variables and Bell's identification of the assumption that makes it inapplicable [1324], a case study in Chapter 1 of a theorem being widely believed to say more than it says.]

85.1.3 Three problems, stated separately

[Reserved: the useful decomposition into the problem of outcomes (why one result), the problem of the preferred basis (why this observable is definite), and the problem of statistics (why these probabilities), following Maudlin's separation [1905]; the observation that decoherence solves the second, is silent on the first, and gives at best a circular account of the third; that each proposal in Section 85.3 is best classified by which of the three it addresses and at what cost.]

85.2 What experiment has already settled

85.2.1 Gleason's theorem

[Reserved: that in a Hilbert space of dimension at least three the only probability assignment on the lattice of projections that is additive on orthogonal families is $\mathrm{tr}(\rho P)$ *[1111]; the consequence that the Born rule is not an independent choice once non-contextual additivity is granted, and the equally important consequence that dimension two is exempt; the modern Busch–Caves–Fuchs–Schack strengthening to POVMs, which removes the dimension restriction; the status of this as a mathematical theorem constraining any interpretation, not as an interpretation.]*

85.2.2 Contextuality

[Reserved: the Kochen–Specker theorem, that no assignment of definite values to all observables can reproduce the functional relations among commuting sets [1353]; the original 117*-vector construction and the small modern sets; the distinction from Bell's theorem — contextuality needs no spatial separation and no locality assumption, only compatibility — and the resulting statement that value-definiteness, not merely locality, is what fails; the experimental tests of state-independent contextuality inequalities, and the loopholes (finite precision, compatibility) that keep them weaker than Bell tests.]*

85.2.3 Bell and Leggett–Garg constraints

[Reserved: what the measured Bell violations of Chapters 84 and 89 exclude, restated as a constraint on interpretations rather than on theories: any account must be either non-local, or non-value-definite, or must abandon measurement independence. The temporal analogue of Leggett and Garg [1759], whose assumptions are macroscopic realism and non-invasive measurability, and its violation in superconducting circuits and nuclear spins; the honest caveat that the non-invasiveness assumption is far harder to enforce than spatial separation, so a Leggett–Garg violation is weaker evidence than a Bell violation.]

85.2.4 PBR and the status of the quantum state

[Reserved: the Pusey–Barrett–Rudolph theorem [2199], that if a system has a real physical state determining measurement statistics, and if independently prepared systems have independent physical states, then distinct quantum states cannot correspond to overlapping distributions — so the wavefunction is ontic, not merely a state of knowledge, within that framework; the preparation-independence assumption, which is exactly where epistemic accounts resist; the ontological-models framework of Harrigan and Spekkens in which the statement is formulated; the experimental tests, which bound the degree of overlap rather than settle the question.]

85.2.5 Decoherence

[Reserved: environmental monitoring destroys local coherence on timescales far shorter than any dissipative timescale, as first argued by Zeh [1440], computed for a mass coupled to scattering particles by Joos and Zeh [1757], and developed into einselection and quantum Darwinism by Zurek [2055, 2151]; the derivation belongs to Chapter 87 and is referenced, not repeated, here. The observations: the controlled decay of a mesoscopic cavity superposition [1916], interference of C_{60} molecules [1966], and its controlled destruction by gas collisions [2052]. The crucial negative statement, which this chapter will make plainly: decoherence explains why the interference terms are unobservable and which basis is stable, and it does not *explain why one outcome occurs — the state after decoherence is still an improper mixture.]*

85.3 The interpretive programmes

85.3.1 Copenhagen and complementarity

[Reserved: Bohr's complementarity as announced at Como [593] — that wave and particle descriptions are mutually exclusive and jointly exhaustive, and that the apparatus must be described classically; Heisenberg's uncertainty paper and its microscope argument [576], with the essential correction that the modern inequality is a property of states, proved by Kennard [580] and in general form by Robertson [636], and not a statement about measurement disturbance — a misattribution worth correcting explicitly, since the distinction is measurable and has been measured in weak-measurement experiments; the quantitative wave–particle duality relations that replaced qualitative complementarity; what "Copenhagen" does and does not commit one to, given that Bohr and Heisenberg disagreed.]

85.3.2 Pilot-wave theory

[Reserved: de Broglie's pilot wave as presented at Solvay [567] and Bohm's two-part revival [995, 996]; particles with definite positions guided by the wavefunction through the velocity field of the probability current, with the quantum potential of the polar decomposition already written in Chapter 23; equivariance of the $|\psi|^2$ distribution, so that the theory reproduces every prediction of Chapter 77 exactly and is therefore not distinguishable by experiment; the explicit non-locality of the guidance equation for many particles, which is what Bell's theorem says any such theory must have; the genuine costs — a preferred foliation, and the unsettled relativistic and field-theoretic extension.]

85.3.3 Everett's relative states

[Reserved: Everett's proposal that the unitary evolution is all there is and that measurement outcomes are relative to a branch [1107]; the modern form in which decoherence supplies

the branch structure, so that the preferred-basis problem is answered by Section 85.2.5; the remaining and hardest problem, the status of probability in a theory where every outcome occurs, and the proposed answers — Deutsch's decision-theoretic derivation [1972] and Zurek's envariance argument [2086]; the statement that no experiment currently distinguishes this from any other interpretation, and the fair note that its advocates regard that as its virtue.]

85.3.4 Consistent histories

[Reserved: the reformulation of quantum mechanics in terms of sequences of projections with a decoherence functional whose off-diagonal terms vanish, giving families of histories to which probabilities may consistently be assigned; Griffiths' formulation [1738], Omnès' development [1857], and the Gell-Mann–Hartle decoherent-histories version aimed at closed systems and cosmology [1865]; the set-selection problem — many mutually incompatible consistent families exist and the formalism does not choose — which is where its critics locate the measurement problem in new clothes.]

85.3.5 Epistemic and relational accounts

[Reserved: QBism, in which the quantum state is an agent's personal degree of belief and the Born rule a normative coherence condition, as set out by Fuchs, Mermin and Schack [2233]; Rovelli's relational account, in which state assignments are always relative to a physical system [1929]; the Frauchiger–Renner argument that a set of natural-looking assumptions about agents reasoning about agents is inconsistent [2312], and the several ways different programmes escape it; the treatise's assessment that these are positions on the meaning of probability, on which no experiment discussed in this book bears.]

85.4 Dynamical collapse: interpretation made testable

Collapse models are the reason this chapter is not merely a survey of opinions. They modify the Schrödinger equation itself, so they predict different numbers, and the numbers are being measured.

85.4.1 GRW and continuous spontaneous localization

[Reserved: the Ghirardi–Rimini–Weber model [1772], in which each constituent undergoes spontaneous localization at a rate λ with width r_C, so that an N-particle centre-of-mass superposition collapses N times faster, so that the transition from micro to macro becomes a computed crossover rather than a postulated cut; the parameters originally proposed, a rate λ of $1 \times 10^{-16}/\mathrm{s}$ at a localization length r_C of $1 \times 10^{-7}\,\mathrm{m}$; the continuous stochastic reformulation of Pearle [1825] and its mass-proportional identical-particle version [1831], which is the model actually confronted with data; the unavoidable side effect, that the noise heats matter and makes charged particles radiate, which is what the experiments of Section 85.4.3 exploit; the unresolved problems — energy non-conservation and no relativistic version with a white-noise field.]

85.4.2 Gravitationally induced collapse

[Reserved: Diósi's proposal that the noise is gravitational and the collapse rate therefore fixed by Newton's constant, leaving only a short-distance regulator free [1820], and Penrose's independent argument from the ill-definedness of the time-translation Killing field for a superposition of spacetimes, giving a decay time set by the gravitational self-energy difference

[1926]; the resulting model is parameter-free up to the regulator, which makes it the most falsifiable proposal in the chapter; the relation to Chapter 135, where the question of whether gravity must be quantized at all is examined against evidence.]

85.4.3 The experimental bounds

[Reserved: the section that makes the chapter empirical. The searches use the predicted spontaneous X-ray emission from germanium and the predicted heating of mechanical oscillators. A cantilever-based force sensor at millikelvin temperatures excludes the mass-proportional CSL rate over a wide band of localization lengths [2366]; an underground germanium experiment excludes the parameter-free Diósi–Penrose model outright, requiring a regulator larger than about $5 \times 10^{-11}\,\mathrm{m}$, which is bigger than the size of a nucleus and therefore not a natural choice [2377]; the review [2212] for the full parameter plane. The section will tabulate, in SI, the excluded region of the (λ, r_C) plane with its uncertainties, and state plainly what remains open — which is the point: this is the one part of the subject where a treatise on evidence-based physics has data to report.]

85.5 Adjacent experiments that constrain the discussion

85.5.1 Delayed choice and quantum erasure

[Reserved: Wheeler's delayed-choice gedanken experiment [1629] and its realization with a single-photon interferometer whose output beam splitter is inserted after the photon has entered [2108]; the correct reading — that which-path information and interference are complementary properties of the whole experimental arrangement — and the incorrect reading, that the past is retroactively altered, which the no-signalling theorem of Chapter 84 already excludes; the quantum eraser as the same statement about post-selected subensembles.]

85.5.2 Weak measurement

[Reserved: the weak-value formalism of Aharonov, Albert and Vaidman [1803], in which a weakly coupled pointer, post-selected on a final state, registers a complex quantity outside the eigenvalue spectrum; the reconstruction of average photon trajectories in a two-slit interferometer [2174], which agree with the Bohmian guidance field of Section 85.3.2 — and the careful statement that this confirms a calculation both formalisms share, and selects neither; the use of weak measurements to separate measurement disturbance from the state-intrinsic uncertainty relation of Section 85.3.1.]

85.6 The ledger

[Reserved: the closing section, a two-column table rather than an argument. One column lists the statements this chapter regards as physics because an experiment bears on them — the Born rule and its tested exponent, Gleason, Kochen–Specker, Bell and Leggett–Garg violations, PBR under its stated assumptions, decoherence rates measured against computed ones, and the excluded regions of collapse-model parameter space. The other lists the statements no experiment in this treatise distinguishes: whether an unobserved branch exists, whether the wavefunction is a thing or a credence, whether a particle has a trajectory between measurements. The chapter closes by naming what would move an entry from the second column to the first — an observed deviation from linear evolution, a measured Born-rule violation, or a macroscopic interference experiment beyond the collapse bounds of Section 85.4.3 — and by noting that these are experiments now under construction, not thought experiments.]

Chapter 86

Scattering Theory

Bound states are the exception. Most of what is known about the microscopic world was learned by throwing something at something else and counting what came out at each angle, and scattering theory is the dictionary between that count and the interaction responsible for it. The subject begins with Rutherford's analysis of large-angle alpha deflection [379], which turned a measured angular distribution into the nuclear atom; it acquired its quantum form with Born's collision papers [533, 534], in which the probability interpretation of Chapter 77 and the approximation scheme that still bears his name were introduced in the same breath; and it supplies the language — cross section, phase shift, resonance width, S-matrix — in which nuclear physics (Chapter 107), particle physics (Chapter 116) and cold-atom physics (Chapter 121) all report their results.

The chapter is placed after the approximation methods of Chapter 82, whose time-dependent perturbation theory it uses, and after identical particles (Chapter 83), because the interference between direct and exchange amplitudes is measurable and was measured [663]. It is deliberately nonrelativistic and potential-based; the relativistic S-matrix, its analyticity and its diagrammatic evaluation belong to Chapters 99 and 100, and the formal structure stated here is what those chapters build on. The standard monographs are [1486, 1698]; the elementary treatments are [2287, 2313].

Derivation pending #101

Scattering Theory: all derivations of this chapter are pending.

86.1 What a scattering experiment measures

86.1.1 Cross sections

[Reserved: the differential cross section defined as scattered flux per unit solid angle divided by incident flux per unit area, so that it carries the dimensions of an area per unit solid angle and is quoted in m^2 *per steradian; the integrated and total cross sections; the practical unit of nuclear and particle physics, the barn, defined as exactly* $1 \times 10^{-28}\,\mathrm{m}^2$ *and quoted in this treatise with that conversion given once; luminosity and event rate, connecting the theoretical quantity to what a detector counts; the transport and momentum-transfer cross sections, which are the ones that enter the kinetic theory of Chapter 55 and the transport coefficients of Chapter 124.]*

86.1.2 Kinematics, frames and flux

[Reserved: the reduction of the two-body problem to one body of reduced mass in a fixed potential, as in Chapter 27; the laboratory-to-centre-of-mass transformation of angles and of

the cross section, with the Jacobian written out, since almost every experimental table needs it; elastic, inelastic and reactive channels; the energy variables and the threshold condition; a remark that the relativistic kinematics of Chapter 40 replaces this section wherever the beam energy demands it.]

86.1.3 The scattering amplitude

[Reserved: the asymptotic form $\psi \sim \mathrm{e}^{\mathrm{i}kz} + f(\theta,\phi)\,\mathrm{e}^{\mathrm{i}kr}/r$ *and the identification* $\mathrm{d}\sigma/\mathrm{d}\Omega = |f|^2$ *by computing the probability current of Chapter 77; that the amplitude is complex while the cross section is not, so that a measured angular distribution determines* $|f|$ *and leaves the phase to be inferred — the phase problem that recurs throughout the chapter, and the reason interference experiments such as Section 86.6.2 carry information that a rate measurement cannot.]*

86.2 The stationary formulation

86.2.1 Asymptotic conditions and scattering states

[Reserved: what it means for an interacting state to approach a free state in the far past and far future; the Møller wave operators $\Omega_{\pm}$ *and their isometry [881]; asymptotic completeness — that the ranges of the wave operators together with the bound states exhaust the Hilbert space — stated as the theorem it is and not assumed, with the potentials for which it is proved; the requirement of a potential falling faster than* $1/r$*, which the Coulomb case violates and which is why Section 86.6 is treated separately.]*

86.2.2 Green functions and the Lippmann–Schwinger equation

[Reserved: the outgoing free Green function $-\mathrm{e}^{\mathrm{i}k|\boldsymbol{x}-\boldsymbol{x}'|}/4\pi\,|\boldsymbol{x}-\boldsymbol{x}'|$ *and the* $\mathrm{i}\varepsilon$ *prescription that selects it, with the contour integration of Chapter 8; the integral equation of Lippmann and Schwinger [963] in both wavefunction and operator form, and the transition operator* T *defined by it; the equivalence to the differential equation with a boundary condition, and why the integral form is the useful one; the formal theory of Gell-Mann and Goldberger [1018] relating the stationary and time-dependent pictures, and Dirac's golden rule [572] as the first-order rate that Chapter 82 supplies.]*

86.2.3 The Born series and the Born approximation

[Reserved: iteration of the Lippmann–Schwinger equation; the first Born approximation, in which the amplitude is the Fourier transform of the potential in the momentum transfer $\boldsymbol{q}$ *[533, 534]; the immediate consequences — that a measured angular distribution at high energy is a direct image of the potential, which is the logic of form factors and of Chapter 116; the worked cases, Yukawa and screened Coulomb, with the screened Coulomb limit reproducing the Rutherford formula exactly; the validity criterion, the convergence of the series, and its failure at low energy and for strong potentials, where Section 86.3 takes over.]*

86.3 Partial waves and phase shifts

86.3.1 The partial-wave expansion

[Reserved: expansion of the plane wave and the amplitude in Legendre polynomials for a central potential, with the special functions of Chapter 9; the radial equation, the centrifugal barrier, and the phase shift δ_ℓ *defined by the asymptotic form of the radial function; the amplitude* $f(\theta) = k^{-1}\sum_\ell (2\ell+1)\mathrm{e}^{\mathrm{i}\delta_\ell}\sin\delta_\ell P_\ell(\cos\theta)$*; the systematic use of this expansion for*

slow electrons in gases by Faxén and Holtsmark [575]; the sign of the phase shift as the sign of the potential, and the semiclassical estimate $\ell_{\max} \approx kR$ that says how many partial waves a given energy needs.]

86.3.2 Unitarity and the optical theorem

[Reserved: conservation of probability as unitarity of the partial wave amplitudes, $|S_\ell| = 1$ for purely elastic scattering and $|S_\ell| < 1$ when channels are open; the optical theorem $\sigma_{tot} = (4\pi/k)\,\mathrm{Im}\, f(0)$, derived for potential scattering by Feenberg [697] and stated in the general form covering inelastic channels by Bohr, Peierls and Placzek [829]; the physical reading, that the total removal of flux from the beam is the shadow cast by forward interference; the unitarity bound $\sigma_\ell \le 4\pi(2\ell+1)/k^2$ and its use as a consistency check on measured cross sections.]

86.3.3 Low energy: scattering length and effective range

[Reserved: the dominance of the s wave as $k \to 0$, the scattering length a defined by $\delta_0 \to -ka$, and the low-energy cross section $4\pi a^2$; the effective-range expansion $k\cot\delta_0 = -1/a + \tfrac{1}{2}r_0k^2 + \cdots$, whose model-independence Bethe established [924]; the neutron–proton system as the worked case, with the triplet and singlet scattering lengths of about $5.4\,\mathrm{fm}$ and $-23.7\,\mathrm{fm}$, the deuteron as the near-threshold bound state the first implies and the virtual state the second implies, and the resulting hugely spin-dependent cross section — a number measured before it was understood (Chapter 107); the Feshbach-resonance tuning of a in cold gases [1956, 2153], which is what makes Chapters 121 and 130 quantitative.]

Phenomenon 86.1 (Low-energy scattering sees only one length)**.** As the collision energy is lowered, elastic scattering becomes isotropic — the measured angular distribution flattens until no direction is preferred — and the total elastic cross section approaches a constant,

$$\sigma \longrightarrow 4\pi a^2 \qquad (k \to 0)\,, \tag{86.1}$$

fixed by a single length a characteristic of the colliding pair. Two potentials of quite different shape that happen to share the value of a cannot be told apart by any low-energy measurement. The length need not resemble the range of the force: it may be many times larger, and it may be negative. Both extremes are observed. The neutron–proton system has scattering lengths of opposite sign and very different magnitude in its two spin channels, which is why the low-energy neutron cross section on hydrogen depends so violently on the relative spin orientation; and in a cold atomic gas a is swept through zero and through infinity by an applied magnetic field tuned across a Feshbach resonance [1956], so that the same atoms behave as an ideal gas or as a strongly interacting fluid on demand.

Derivation. Let the potential have finite range R. For $r > R$ the radial function of angular momentum ℓ is a combination of free solutions, and its amplitude inside the range is suppressed by the centrifugal barrier: the regular free solution behaves as $(kr)^{\ell+1}$ near the origin, so the potential is sampled with weight $(kR)^{2\ell+1}$ and

$$\delta_\ell(k) = O\big((kR)^{2\ell+1}\big) \qquad (k \to 0)\,. \tag{86.2}$$

Only $\ell = 0$ survives the limit. Defining $a := -\lim_{k\to 0}\delta_0(k)/k$, the partial-wave amplitude of Section 86.3.1 gives

$$f(\theta) = \frac{1}{k}\mathrm{e}^{\mathrm{i}\delta_0}\sin\delta_0 + O(k^2) \longrightarrow -a\,,$$

which is independent of θ: the scattering is isotropic, and $\sigma = \int |f|^2\,\mathrm{d}\Omega = 4\pi a^2$, which is Equation (86.1). The geometric meaning of a follows from the same limit. Outside the range

the s-wave radial function is $u \propto \sin(kr + \delta_0) \to k\,(r - a)$, so a is the intercept at which the exterior wave function, extrapolated inwards as a straight line, crosses zero. Everything the interior does to the collision is compressed into that one intercept, and a large $|a|$ signals a bound or virtual state lying close to threshold rather than a large force. □

86.3.4 Levinson's theorem

[Reserved: the relation $\delta_\ell(0) - \delta_\ell(\infty) = n_\ell \pi$ between the total change of a phase shift and the number of bound states of that angular momentum [938], with the half-integer modification for a zero-energy resonance in the s wave; the statement that a scattering measurement therefore counts bound states it cannot produce; the inverse-scattering problem, in which the potential is reconstructed from the phase shifts and the bound-state data, solved by Gel'fand and Levitan [982], together with the non-uniqueness that phase-equivalent potentials exhibit.]

86.4 The S-matrix

86.4.1 The S operator and its symmetries

[Reserved: $S = \Omega_-^\dagger \Omega_+$ as the map from incoming to outgoing asymptotic states, with the name and the resonating-group construction due to Wheeler [806] and its elevation to the primary object of the theory — observable quantities only — by Heisenberg [872]; unitarity from probability conservation, symmetry from time reversal and the resulting reciprocity or detailed-balance relation between forward and reverse reactions, and the constraints from rotational invariance, parity and the discrete symmetries of Chapter 105; the observation that this is where Chapter 109 begins.]

86.4.2 Analyticity, poles and dispersion relations

[Reserved: continuation of the partial-wave amplitude to complex energy or complex momentum; poles on the positive imaginary k axis as bound states, poles on the negative imaginary axis as virtual states, and poles on the second sheet as resonances; the branch cuts and the physical sheet; dispersion relations as Cauchy's theorem (Chapter 8) applied with causality, in the form given for light by Kronig [550] and Kramers [581] and transplanted to scattering amplitudes; complex angular momentum and Regge trajectories [1174], named here because the linear mass-squared–spin trajectories of the observed hadron spectrum are their evidence.]

86.5 Resonances

86.5.1 The Breit–Wigner form

[Reserved: the resonant partial-wave amplitude $\Gamma/2$ over $(E_R - E) - \mathrm{i}\Gamma/2$, giving a Lorentzian peak in the cross section of height set by the unitarity bound and width Γ, first written for slow-neutron capture by Breit and Wigner [771]; the mean life $\tau = \hbar/\Gamma$ and the phase shift sweeping through $\pi/2$; partial widths and branching ratios; the worked examples — neutron resonances in heavy nuclei (Chapter 107), the Δ resonance in pion–nucleon scattering, and the Z line shape whose measured width counts the light neutrino species (Chapter 104).]

Phenomenon 86.2 (Resonances)**.** Cross sections are not smooth functions of energy. Superposed on a slowly varying background are peaks whose shape is, to good accuracy, a Lorentzian of full width Γ at half maximum centred on an energy E_R. Slow neutrons incident on a heavy nucleus show hundreds of them within the first few electronvolts; pion–nucleon scattering shows one dominant peak near the first excited state of the nucleon; and the cross section for electron–positron annihilation into hadrons shows the Z. Wherever the

corresponding unstable state can also be timed independently, its mean life is found to be $\tau = \hbar/\Gamma$, and the widths tabulated for hundreds of such states [2427] are read as lifetimes on exactly that understanding. A resonance is therefore not a new kind of object but a state that lives long enough to be counted and too briefly to be caught.

Derivation. Consider a state that, once formed, decays exponentially: its amplitude is $a(t) = \mathrm{e}^{-\mathrm{i}E_R t/\hbar}\mathrm{e}^{-\Gamma t/2\hbar}$ for $t > 0$ and zero before, so that $|a(t)|^2 = \mathrm{e}^{-\Gamma t/\hbar}$ and the mean life is $\tau = \hbar/\Gamma$. Its energy content is the Fourier transform of Chapter 17,

$$\tilde{a}(E) = \frac{1}{2\pi\hbar}\int_0^\infty \mathrm{d}t\, \mathrm{e}^{\mathrm{i}Et/\hbar} a(t) = \frac{1}{2\pi}\,\frac{\mathrm{i}}{(E - E_R) + \mathrm{i}\Gamma/2}\,, \tag{86.3}$$

the integral converging because of the damping factor. Hence the probability of finding the state at energy E is

$$|\tilde{a}(E)|^2 = \frac{1}{4\pi^2}\,\frac{1}{(E - E_R)^2 + \Gamma^2/4}\,, \tag{86.4}$$

a Lorentzian whose full width at half maximum is exactly Γ. The cross section for forming the state through a given entrance channel is proportional to this same distribution, so the observed line shape and the observed lifetime are two readings of one complex pole at $E = E_R - \mathrm{i}\Gamma/2$; the relation $\tau\Gamma = \hbar$ between them is not an approximation but the statement that the two are Fourier conjugates, the time–energy form of the uncertainty relation of Chapter 77. In partial-wave language the same pole appears as an elastic amplitude proportional to $(\Gamma/2)\big[(E_R - E) - \mathrm{i}\Gamma/2\big]^{-1}$, whose modulus attains the unitarity bound of Section 86.3.2 at $E = E_R$ and whose phase shift sweeps through $\pi/2$ there — which is how a resonance is distinguished from a mere bump in the background. □

86.5.2 Time delay

[Reserved: the Eisenbud–Wigner delay $2\hbar\,\mathrm{d}\delta_\ell/\mathrm{d}E$ as the time a wave packet lingers in the interaction region, with Wigner's lower bound on it — causality forbids arbitrarily large negative delay, that is arbitrarily early emission [1081]; Smith's lifetime matrix generalizing this to many channels [1192]; the equivalence of a long delay and a narrow resonance; the modern measurement of photoionization delays on the attosecond scale as the direct observation of this quantity.]

86.5.3 Fano profiles and Feshbach resonances

[Reserved: interference between a resonant and a direct amplitude producing the asymmetric line shape parametrized by Fano's q [1204], with autoionizing states of helium as the original observation; Feshbach's unified theory of nuclear reactions, in which a closed channel supports a quasi-bound state coupled to an open one [1142]; the optical model, in which the average effect of many open channels is a complex potential [1036], and the absorption cross section it predicts; the same physics used as a control knob in ultracold gases [1956, 2153].]

86.6 Coulomb scattering

86.6.1 Rutherford, classically and quantum mechanically

[Reserved: the classical derivation of the $\csc^4(\theta/2)$ law from the hyperbolic orbits of Chapter 27 [379], and the Geiger–Marsden measurements that tested every one of its dependences [404]; the exact quantum solution in parabolic coordinates by Gordon [602], giving the same cross section but a logarithmically distorted phase and a modified asymptotic form, so that the long range of the Coulomb potential invalidates Section 86.2.1 while leaving the measured

rate unchanged — a coincidence of the inverse-square law that deserves to be flagged rather than passed over; the Mott cross section, which adds the spin of a relativistic projectile [633] and whose deviation from Rutherford at large angle is a measurement of nuclear size.]

Phenomenon 86.3 (Large-angle Coulomb scattering)**.** A beam of alpha particles crossing a thin metal foil is deflected, and a small but perfectly reproducible fraction is deflected through more than a right angle — some of it back towards the source [354]. The number counted per unit solid angle at a scattering angle θ obeys

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \left(\frac{Z_1 Z_2 e^2}{16\pi\epsilon_0 E}\right)^2 \frac{1}{\sin^4(\theta/2)}\,, \tag{86.5}$$

with E the kinetic energy in the centre-of-mass frame. Four dependences are asserted here and all four were measured separately: the fourth power of the sine of the half angle, the inverse square of the energy, the square of the charge number of the target, and — for a foil thin enough that a particle meets at most one scatterer — strict proportionality to the foil thickness [404]. Together they establish that the positive charge of an atom is not spread through it but concentrated in a region small enough to permit the close approaches that large angles require.

Derivation. Let the projectile of charge $Z_1 e$ approach a fixed charge $Z_2 e$ with reduced mass μ, asymptotic speed v, energy $E = \tfrac{1}{2}\mu v^2$ and impact parameter b, and write $\kappa = Z_1 Z_2 e^2/4\pi\epsilon_0$ so that the potential is κ/r. The orbit is the repulsive hyperbola of Chapter 27, whose asymptotes make the deflection angle θ given by

$$b = \frac{\kappa}{2E}\cot\frac{\theta}{2}\,. \tag{86.6}$$

The map $b \mapsto \theta$ is monotone, so particles entering the annulus between b and $b + \mathrm{d}b$ leave between θ and $\theta + \mathrm{d}\theta$, and equating the two fluxes, $2\pi b\,|\mathrm{d}b| = (\mathrm{d}\sigma/\mathrm{d}\Omega)\,2\pi\sin\theta\,|\mathrm{d}\theta|$, gives the classical cross-section formula

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{b}{\sin\theta}\left|\frac{\mathrm{d}b}{\mathrm{d}\theta}\right|\,. \tag{86.7}$$

Differentiating Equation (86.6) gives $|\mathrm{d}b/\mathrm{d}\theta| = (\kappa/4E)\sin^{-2}(\theta/2)$, while $\sin\theta = 2\sin(\theta/2)\cos(\theta/2)$ turns the first factor into $b/\sin\theta = (\kappa/4E)\sin^{-2}(\theta/2)$ as well. Their product is

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \left(\frac{\kappa}{4E}\right)^2 \frac{1}{\sin^4(\theta/2)}\,,$$

which is Equation (86.5). The thickness dependence follows from counting: with n scatterers per unit volume in a foil of thickness t, the fraction of the beam deflected into $\mathrm{d}\Omega$ is $nt\,(\mathrm{d}\sigma/\mathrm{d}\Omega)\mathrm{d}\Omega$, linear in t so long as $nt\sigma \ll 1$.

Two remarks belong with the result rather than after it. First, nothing in the derivation refers to $\hbar$, and yet the first Born approximation of Section 86.2.3 and the exact quantum solution in parabolic coordinates both return Equation (86.5) unchanged. That agreement is a peculiarity of the inverse-square law and must not be read as a general licence to compute cross sections classically. Second, the total cross section obtained by integrating Equation (86.5) diverges, because $\sin^{-4}(\theta/2)$ is not integrable at $\theta = 0$: an unscreened Coulomb potential deflects every particle, however distant. Real targets are screened by their own electrons, which cuts the divergence off at an angle of order the inverse of the screening length times the momentum, and it is that screening — not the formula — that makes a total cross section finite. □

86.6.2 Identical particles: Mott scattering

[Reserved: when projectile and target are identical, the direct and exchange amplitudes must be added with the exchange parity of Chapter 83 before squaring, giving an interference term with no classical analogue and a cross section symmetric about 90 *degrees in the centre-of-mass frame; Mott's calculation for two electrons [663] — customarily cited to his 1929 paper, which is a different result — and its confirmation in alpha-on-helium scattering [645], where the interference maximum is a factor of two above the classical sum; the spin-averaged version, weighting singlet and triplet, as the cleanest laboratory demonstration that exchange symmetry is a measurable property of nature and not a bookkeeping convention.]*

Phenomenon 86.4 (Interference between identical colliding particles)**.** When projectile and target are identical, the measured centre-of-mass angular distribution is symmetric about 90° — as it must be, since the two particles arriving at the detectors cannot be told apart — and it is not the sum of the two contributions that a trajectory picture would add. For identical spinless particles, such as an alpha particle scattering on helium, the counting rate at exactly 90° is twice the classical sum, with oscillations on either side of it [645]. For unpolarized identical spin-half particles the interference enters with the opposite sign in the triplet, which has three times the weight of the singlet, and the measured rate at 90° is instead *half* the classical sum. The sign of a term in a cross section is thus a directly measurable consequence of the exchange symmetry postulated in Chapter 83.

Derivation. Detecting a particle at angle θ in the centre-of-mass frame can happen in two ways: the projectile was deflected by θ, amplitude $f(\theta)$, or the projectile was deflected by $\pi-\theta$ and the recoiling target reached the detector, amplitude $f(\pi-\theta)$. The two possibilities end in the same final configuration and are indistinguishable in principle, so the amplitudes are added before squaring, with the relative sign fixed by the required symmetry of the spatial state:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = |f(\theta) \pm f(\pi-\theta)|^2 = |f(\theta)|^2 + |f(\pi-\theta)|^2 \pm 2\,\mathrm{Re}\big[f(\theta) f^*(\pi-\theta)\big]\,, \tag{86.8}$$

the upper sign for a symmetric spatial state — identical spinless bosons, or spin-half particles in the spin singlet — and the lower sign for an antisymmetric one. The classical prediction is the first two terms alone; the third has no classical counterpart, and both expressions are manifestly symmetric under $\theta \mapsto \pi-\theta$.

At $\theta = \pi/2$ the two amplitudes coincide, $f(\theta) = f(\pi-\theta)$, and the arithmetic is immediate. The symmetric case gives $|2f|^2 = 4\,|f|^2$ against a classical $2\,|f|^2$: a factor of two. The antisymmetric case gives exactly zero, however strong the interaction. For unpolarized spin-half particles the four spin states are equally likely, so the singlet carries weight $\frac{1}{4}$ and the triplet $\frac{3}{4}$; since a symmetric spin state (triplet) forces an antisymmetric spatial state and conversely, the spin-averaged rate at 90° is

$$\tfrac{1}{4}\cdot 4\,|f|^2 + \tfrac{3}{4}\cdot 0 = |f|^2\,,$$

one half of the classical $2\,|f|^2$. The measured ratio at 90° therefore distinguishes bosons from fermions with no adjustable parameter, and it does so through the value of the cross section and not through any spin-dependent force. □

86.7 Observed effects the formalism must reproduce

86.7.1 The Ramsauer–Townsend effect

[Reserved: the near-vanishing of the cross section for slow electrons in argon, krypton and xenon near 0.7 eV*, reported independently by Ramsauer [480] and by Townsend and Bailey*

[495]; its explanation as an s-wave phase shift passing through π, so that the leading partial wave contributes nothing — a purely quantum result with no classical counterpart, and historically one of the earliest; the corresponding transparency window and its consequence for electron transport in noble gases; the analogous zero-crossing of the scattering length used to make non-interacting Fermi gases.]

86.7.2 Threshold behaviour

[Reserved: Wigner's threshold laws, giving the energy dependence of a cross section just above the opening of a channel as a power fixed by the angular momentum and the range of the interaction [922]; the $1/v$ law for exothermic reactions on which slow-neutron absorption and the operation of every thermal reactor rest (Chapter 107); cusps at thresholds as a unitarity consequence rather than a dynamical one; the modification for long-range potentials.]

Phenomenon 86.5 (The inverse-velocity absorption law)**.** For an exothermic reaction — one whose outgoing channel is open with a large energy release, radiative neutron capture being the standard case — the cross section measured at low incident energy rises as the inverse of the incident speed over several decades, so that the product σv is constant. Absorption of slow neutrons in boron, in cadmium and in most heavy nuclei follows this law from thermal energies downwards, interrupted only by the resonances of Phenomenon 86.2. The law is why a reactor moderates its neutrons before it asks them to be absorbed: slowing a neutron buys cross section in exact proportion to the slowing (Chapter 107).

Derivation. At low incident energy only the $\ell = 0$ wave reaches the target, by Equation (86.2), so the reaction cross section is

$$\sigma_{\mathrm{r}} = \frac{\pi}{k^2}\left(1 - |S_0|^2\right). \tag{86.9}$$

Write the exterior s-wave radial function as $u(r) = \mathrm{e}^{-\mathrm{i}kr} - S_0\mathrm{e}^{\mathrm{i}kr}$ and let $f = R\,u'(R)/u(R)$ be its logarithmic derivative at the surface $r = R$ of the interacting region. Matching to the interior solution makes f complex — its imaginary part is the flux disappearing into the reaction channel — and, crucially, the reaction channel is open by an energy of nuclear size, so f is insensitive to the tiny incident energy and may be treated as a constant $f = f_1 + \mathrm{i}f_2$ with $f_2 < 0$. Solving the matching condition for S_0,

$$S_0 = \mathrm{e}^{-2\mathrm{i}kR}\,\frac{f + \mathrm{i}kR}{f - \mathrm{i}kR}, \tag{86.10}$$

so that

$$1 - |S_0|^2 = \frac{(f_2 - kR)^2 - (f_2 + kR)^2}{f_1^2 + (f_2 - kR)^2} = \frac{-4f_2\,kR}{f_1^2 + (f_2 - kR)^2},$$

which is positive and, as $kR \to 0$, proportional to k. Substituting into Equation (86.9),

$$\sigma_{\mathrm{r}} \longrightarrow \frac{\pi}{k^2}\cdot\frac{-4f_2R}{f_1^2 + f_2^2}\,k = \frac{\text{constant}}{k} \propto \frac{1}{v}, \tag{86.11}$$

the constant being fixed by the interior alone. The physical content is worth separating from the algebra: the two powers of $1/k$ in Equation (86.9) are the geometric size of the incident s wave, which grows without limit as the wavelength does, while the single factor of k is the small amplitude that wave has at the target. An endothermic reaction behaves quite differently, because there the outgoing momentum vanishes at threshold and supplies its own energy dependence. □

86.7.3 High energy: the eikonal and Glauber regime

[Reserved: the eikonal approximation, in which the phase accumulated along a straight-line path replaces the full solution — the same short-wavelength limit as the optical–mechanical analogy of Chapter 23, here applied to the quantum amplitude; the profile function and the impact-parameter representation; Glauber's multiple-scattering theory for composite targets [1068], which is how the deuteron and then nuclei are treated as collections of scatterers; the black-disc limit and the factor-of-two total cross section, whose shadow term is the optical theorem of Section 86.3.2 in its most physical form; the pointer to the relativistic treatment in Chapters 100 and 102.]

Phenomenon 86.6 (An opaque target removes twice its own area)**.** At wavelengths short compared with the target, and for a target that absorbs nearly everything striking it, the measured total cross section is not the geometric area πR^2 but twice it. One half is accounted for by the absorption itself; the other half is elastic scattering confined to a narrow forward cone of angular width of order $1/kR$ and carrying diffraction minima, from whose spacing R can be read off. The pattern is standard in fast-neutron scattering from nuclei and in high-energy hadron collisions. It has no classical counterpart whatever: a classical opaque disc removes πR^2 of flux, and nothing more.

Derivation. Model the target as a black disc: in the partial-wave expansion of Section 86.3.1 set $S_\ell = 0$ for $\ell \le L$, with $L = kR$, and $S_\ell = 1$ for $\ell > L$. The two standard sums of Section 86.3.2 then read

$$\sigma_{\mathrm{el}} = \frac{\pi}{k^2}\sum_{\ell}(2\ell+1)\,|1-S_\ell|^2 = \frac{\pi}{k^2}\sum_{\ell=0}^{L}(2\ell+1) = \frac{\pi}{k^2}(L+1)^2 \approx \pi R^2\,, \tag{86.12}$$

$$\sigma_{\mathrm{abs}} = \frac{\pi}{k^2}\sum_{\ell}(2\ell+1)(1-|S_\ell|^2) = \frac{\pi}{k^2}(L+1)^2 \approx \pi R^2\,, \tag{86.13}$$

using $\sum_{\ell=0}^{L}(2\ell+1) = (L+1)^2$ and $L = kR \gg 1$. Their sum is $\sigma_{\mathrm{tot}} = 2\pi R^2$. The same answer follows from the optical theorem: the forward amplitude is $f(0) = (2\mathrm{i}k)^{-1}\sum_\ell(2\ell{+}1)(S_\ell{-}1) = \mathrm{i}(L{+}1)^2/2k \approx \mathrm{i}kR^2/2$, so $\sigma_{\mathrm{tot}} = (4\pi/k)\,\mathrm{Im}\,f(0) = 2\pi R^2$, in agreement with Equations (86.12) and (86.13).

The physical reading is the one worth keeping. Behind an absorbing target there is a shadow, and a shadow is not an absence of wave but the superposition of the incident wave with a second wave that cancels it there. That cancelling wave must be produced by the target, and it carries as much flux as the absorbed wave removed. Far downstream the shadow wave spreads out of the forward direction — by the diffraction angle $1/kR$ — and is then counted by a detector at finite angle as elastic scattering. The elastic half of the cross section is thus not an independent process but the necessary accompaniment of absorption, which is why σ_{el} cannot be made smaller than σ_{abs} for a fully absorbing target however the interaction is arranged. □

Chapter 87

Open Quantum Systems and Decoherence

Every system treated so far in this part has been closed: a state vector, a Hamiltonian, and unitary evolution for all time. No system in a laboratory is closed. A trapped ion radiates, a superconducting circuit shares its substrate with two-level defects, a molecule in flight scatters background gas and emits thermal photons, and an apparatus is by construction coupled to whatever will record its reading. This chapter is the systematic theory of what remains when the rest of the world is traced away: the density operator [582, 585], the completely positive maps that carry it forward [1080, 1450, 1541], the master equations that generate those maps in continuous time [1579, 1586], and the loss of interference — decoherence — that they describe [1440, 1676].

Its place in the book is fixed by what it explains. The entanglement of Chapter 84 is here turned from a resource into a nuisance: coupling to an environment entangles a system with degrees of freedom nobody monitors, and the interference terms that make quantum mechanics quantum are not destroyed but exported, becoming inaccessible to any local measurement. That process is fast, and it is quantitative — the measured decay of a mesoscopic superposition in a microwave cavity [1916] and of interference fringes of hot molecules [2060] are the two cleanest instances, and this chapter states what will be derived and what was measured in each. Decoherence explains why the classical world of Chapter 22 looks classical, and this treatise records plainly what it does not do: it does not select an outcome, and the residual question belongs to Chapter 85. The standard monographs are [2023, 2112].

Derivation pending #102

Open Quantum Systems and Decoherence: all derivations of this chapter are pending.

87.1 States of an open system

87.1.1 The density operator

[Reserved: the density operator introduced independently by Landau, from the damping problem in wave mechanics [582], and by von Neumann in his probabilistic reconstruction of quantum mechanics [585, 706]; its defining properties $\rho = \rho^\dagger$, $\rho \geq 0$, $\operatorname{tr}\rho = 1$; expectation values as $\langle A\rangle = \operatorname{tr}(\rho A)$; the von Neumann equation $\mathrm{i}\hbar\,\mathrm{d}\rho/\mathrm{d}t = [H,\rho]$ as the state-operator form of the Schrödinger evolution of Chapter 77; pure states as the projectors $\rho^2 = \rho$ and the purity $\operatorname{tr}\rho^2$ as the scalar that measures departure from them.]

87.1.2 Reduced states and the partial trace

[Reserved: the partial trace as the unique map reproducing the statistics of every observable of the subsystem; the Schmidt decomposition of a bipartite pure state [338] and the equality of the two reduced spectra; purification, so that every mixed state is the restriction of a pure state on a larger space, which is the state-space face of Stinespring's dilation theorem [1080]; the bipartite machinery of Chapter 84 reused here with the roles reversed.]

87.1.3 Proper and improper mixtures

[Reserved: the distinction, due in its sharp form to d'Espagnat [1577], between a statistical mixture prepared by a classical randomizer and the reduced state of one half of an entangled pair; the two are described by the same operator and are indistinguishable by any measurement on the subsystem alone, and they are not the same physical situation — the second can be undone by recohering the environment, as the spin echo of Section 87.7.2 demonstrates. Why this is the point at which decoherence stops short of solving the measurement problem (Section 87.4.6).]

87.1.4 Entropy and mixedness

[Reserved: the von Neumann entropy $S = -\operatorname{tr}(\rho \ln \rho)$ [706], its vanishing exactly on pure states, its invariance under unitary evolution — hence the impossibility of a closed system's entropy growing, which is why an environment is mandatory; subadditivity and the strong subadditivity inequality of Lieb and Ruskai [1508]; entanglement entropy of a bipartition as the reduced-state entropy; the relation to the thermodynamic entropy of Chapter 118.]

87.2 Quantum channels

87.2.1 Complete positivity

[Reserved: why positivity of a map is not enough — the transpose is positive and its extension to half of an entangled pair is not, so a physical map must be completely positive; Stinespring's dilation theorem for completely positive maps on operator algebras [1080]; Choi's finite-dimensional criterion, that a map is completely positive if and only if its Choi matrix is positive semidefinite [1541]; trace preservation, and the CPTP maps as exactly the reduced dynamics obtainable from a unitary evolution on system plus environment in a product initial state.]

87.2.2 The Kraus representation

[Reserved: Kraus's operator-sum form $\rho \mapsto \sum_k K_k \rho K_k^\dagger$ with $\sum_k K_k^\dagger K_k = \mathbb{1}$ [1450]; the non-uniqueness of the operator set up to an isometry, and what is therefore physical about it; the reading of the K_k as measurement outcomes on the environment, which connects the formalism to the generalized measurements of Chapter 77; the number of Kraus operators bounded by the square of the dimension.]

87.2.3 The standard channels

[Reserved: the qubit channels that every later section uses — dephasing, which destroys coherences in a fixed basis while leaving populations untouched; amplitude damping, the channel of spontaneous emission (Chapter 122); the depolarizing channel; and the two relaxation times T_1 and T_2 they define, with the inequality $T_2 \le 2T_1$ that separates energy relaxation from pure dephasing [2161].]

87.2.4 Process tomography

[Reserved: reconstruction of an unknown channel from the states it produces on a tomographically complete input set, the prescription of Chuang and Nielsen [1936]; its first full application to a two-qubit gate [2064]; the scaling of the required number of settings, and randomized benchmarking as the practical substitute; a measured process matrix is what makes "decoherence" a number rather than a word.]

87.3 Master equations

87.3.1 Projection-operator derivations

[Reserved: the exact Nakajima–Zwanzig equation obtained by projecting the total von Neumann equation onto the relevant part of the state [1155, 1197], with its memory kernel and inhomogeneous initial term; Redfield's earlier relaxation theory for magnetic resonance [1128]; the Born and Markov approximations, weak coupling and short environment correlation time, stated as inequalities on time scales rather than as slogans; the secular (rotating-wave) approximation, without which the resulting equation can produce negative probabilities.]

87.3.2 The Lindblad–Gorini–Kossakowski–Sudarshan generator

[Reserved: the theorem fixing the most general generator of a completely positive trace-preserving semigroup, proved for bounded generators by Lindblad [1586] and, in N-level systems, by Gorini, Kossakowski and Sudarshan [1579]: a Hamiltonian commutator plus a dissipator built from jump operators L_k with non-negative rates. Davies's derivation of the semigroup as the weak coupling limit [1523], which is what ties the abstract theorem to a physical bath; the detailed-balance condition that makes the Gibbs state stationary; why no Markovian generator can be exact.]

87.3.3 Quantum Brownian motion

[Reserved: the influence functional of Feynman and Vernon [1251], integrating out a linear bath exactly; the Caldeira–Leggett model of a particle bilinearly coupled to a continuum of oscillators, its Ohmic spectral density and its high-temperature master equation [1716], together with the demonstration that dissipation suppresses macroscopic tunnelling [1666]; the position-diffusion term whose coefficient sets the decoherence rate of a spatial superposition; the spin–boson problem and its phase diagram [1795]; the classical limit returning the Langevin equation of Chapter 119 and the transport coefficients of Chapter 124.]

87.3.4 Unravellings and quantum trajectories

[Reserved: the Monte Carlo wave-function method, in which the Lindblad evolution is recovered as the average over stochastic pure state trajectories punctuated by jumps [1851], and the diffusive alternative of quantum-state diffusion [1854]; that the unravelling is not unique but is fixed by what the environment is actually measured with; the observation of individual quantum jumps in a single trapped ion by electron shelving [1780], which turned an ensemble formalism into a record of single events.]

Phenomenon 87.1 (Individual quantum jumps)**.** The fluorescence of a single trapped ion driven hard on a strong transition, and coupled weakly to a long-lived third level, is not a smoothly decaying signal. It is a random telegraph: the ion is either fully bright or completely dark, it switches between the two abruptly on the time scale of the detector, and the intervals it spends in each are distributed exponentially [1780]. Averaged over many

such records the signal reproduces the smooth exponential relaxation of ensemble theory exactly. The measurement therefore establishes something an ensemble experiment cannot: that the exponential decay law of an ensemble carries no information about whether the individual system changes gradually or in a single event, and that here it does the latter.

Derivation. Let the ion have a bright configuration B, cycling on the strong transition and scattering photons at a rate the detector resolves, and a dark configuration D, shelved in the metastable level and scattering none; let $\Gamma_\uparrow$ be the rate of the transition $B \to D$ and $\Gamma_\downarrow$ that of $D \to B$. Markovian dynamics — the environment forgetting between events, which is the approximation stated as an inequality on time scales in Section 87.3.1 — means the switching probability per unit time is constant, so the probability of still being bright after a time t obeys $\mathrm{d}P/\mathrm{d}t = -\Gamma_\uparrow P$ and

$$P_{\text{no jump}}(t) = \mathrm{e}^{-\Gamma_\uparrow t}\,, \tag{87.1}$$

an exponential distribution of bright intervals with mean $1/\Gamma_\uparrow$, and likewise for the dark ones with $\Gamma_\downarrow$. Every individual record is therefore a two-valued telegraph with exponentially distributed dwell times, which is what is seen. The ensemble probability $p(t)$ of being dark obeys the rate equation

$$\dot{p} = \Gamma_\uparrow(1-p) - \Gamma_\downarrow p\,, \tag{87.2}$$

whose solution relaxes smoothly and exponentially to $\Gamma_\uparrow/(\Gamma_\uparrow+\Gamma_\downarrow)$. The smooth ensemble curve and the discontinuous single record are thus not in conflict — the first is the mean of the second — and no measurement of the first could have decided between them. That is the general content of an unravelling: a stochastic pure-state trajectory whose average is the Lindblad evolution of Section 87.3.2. It is also why the unravelling is not unique. Which trajectories are realized is fixed by what the environment is monitored with, and here it is fixed by the fact that the experiment counts photons. □

87.3.5 Beyond Markov

[Reserved: structured environments in which the memory kernel cannot be collapsed — solid-state qubits with $1/f$ noise, photonic band gaps, strongly coupled molecular aggregates; measures of non-Markovianity based on the backflow of distinguishability [2143] and on divisibility of the dynamical map [2165], and the fact that the two do not agree; why non-Markovian dynamics is not merely a correction but changes what the coherence time means.]

87.4 Decoherence

87.4.1 Environment-induced superselection

[Reserved: Zeh's argument that macroscopic systems are never isolated and that this, not any modification of the theory, is what makes their superpositions unobservable [1440]; Zurek's pointer-basis analysis [1676] and environment-induced superselection [1709], in which the interaction with the environment singles out a preferred basis and forbids superpositions across it; the review that consolidated the programme [2055]; the honest statement that einselection derives a superselection rule from ordinary unitary dynamics and postulates nothing new.]

87.4.2 Pointer states

[Reserved: the pointer states as those least perturbed by the system–environment coupling — exact eigenstates of the interaction in the ideal case, and predictability-sieve selected states in general [1676, 2055]; why position is the pointer observable for a particle scattering photons

and air molecules while energy eigenstates are pointer states for a weakly damped oscillator; coherent states as the pointer states of the harmonic oscillator, the classical limit already met in Chapters 23 and 78.]

87.4.3 Decoherence rates

[Reserved: the scattering formula of Joos and Zeh [1757] for the decay of spatial coherence, linear in the scattering rate and quadratic in the separation for small separations; the resulting estimates, which give decoherence times shorter than relaxation times by enormous factors — a dust grain in air loses spatial coherence over micrometre separations in far less than a nanosecond, while a single atom in ultrahigh vacuum stays coherent for seconds; the collisional, thermal-photon and gravitational channels compared; why the numbers, not the mechanism, are what make classicality inevitable.]

87.4.4 Quantum Darwinism

[Reserved: Zurek's account of objectivity as redundancy — the pointer observable is the one whose record is imprinted many times over in disjoint fragments of the environment, so that many observers can learn it independently without disturbing it [2151]; the redundancy plateau in the partial information plot as the signature; the first laboratory signatures of the plateau in photonic cluster states [2304]; the claim's status as a consequence of unitary dynamics plus a factorized environment, not a new postulate.]

87.4.5 Thermalization of an isolated system

[Reserved: the complementary route to classical statistics, in which the "bath" is the rest of the same closed system — Deutsch's random-matrix argument [1842] and Srednicki's eigenstate thermalization hypothesis [1887], under which each individual energy eigenstate already encodes thermal expectation values; its direct test in an isolated ultracold Bose gas, where the growth of entanglement entropy was measured while the global state stayed pure [2272]; the link forward to the statistical mechanics of Chapter 118 and back to the scattering description of Chapter 86.]

Phenomenon 87.2 (An isolated system thermalizes while staying pure)**.** A small, completely isolated quantum system — an ultracold gas of a few atoms in an optical lattice, prepared in a pure state and shielded from every bath — relaxes so that each local observable takes the value a thermal ensemble would assign it, while the global state remains pure and its evolution remains unitary. What grows during the relaxation is the entanglement entropy of a subregion, measured by interfering two identical copies of the system; it rises and saturates at the value the thermodynamic entropy of that subregion would have, while the entropy of the whole stays zero throughout [2272]. Thermalization therefore needs no external environment: a system large enough to contain its own complement is its own bath, and the environment of Section 87.2 may be internal.

Derivation pending #103

Eigenstate thermalization: that the expectation of a local observable in an individual many-body energy eigenstate is a smooth function of the energy density alone, so that any superposition confined to a narrow energy window relaxes to the microcanonical value of every local observable; and that the entanglement entropy of a small subregion then grows to the thermodynamic entropy that subregion would carry at the corresponding temperature

87.4.6 What decoherence does not settle

[Reserved: the precise statement of the limit — decoherence converts an improper mixture in a preferred basis into a state operationally indistinguishable from a classical ensemble, and does not select which member is realized; the circularity objection, that the interference terms are still present in the global state; why this leaves the measurement problem exactly where Chapter 85 takes it up, and why no experiment in this treatise distinguishes the readings. The section is to say this plainly rather than allow the formalism to suggest more than it proves.]

87.5 Decoherence observed

87.5.1 Cavity quantum electrodynamics

[Reserved: the Paris experiment in which a mesoscopic superposition of two coherent states of a microwave field, separated by a few photons in phase space, was created with a circular Rydberg atom and its decoherence followed in real time with a second atom [1916]; the observed scaling of the decoherence rate with the square of the separation, the first direct measurement of the rate rather than of its consequences; later reconstruction of the full Wigner function of such a state and of its decay [2121]; the apparatus belongs with Chapter 122.]

Phenomenon 87.3 (Decoherence outruns dissipation by the square of the separation)**.** A superposition of two coherent states of a microwave cavity field, separated by a few photons in phase space, loses its interference much faster than the field loses its energy, and the ratio of the two rates is the square of the phase-space separation. In the Paris experiment a circular Rydberg atom prepared such a superposition and a second atom, sent after a controlled delay, read out the coherence that survived [1916]; the coherence decayed exponentially in the delay, with a time constant shorter than the cavity's own energy damping time by the measured factor, and the dependence on the separation was measured by varying it. The observation is a direct one — the rate itself, read against a delay — and not classicality inferred after the fact from an outcome that looks classical.

Derivation. Grant the generator of Section 87.3.2 for a cavity losing photons to a zero-temperature environment at rate κ, with the single jump operator $L = \sqrt{\kappa}\, a$:

$$\dot\rho = \kappa\Big(a\rho a^\dagger - \tfrac{1}{2}a^\dagger a\rho - \tfrac{1}{2}\rho\, a^\dagger a\Big)\,. \tag{87.3}$$

Coherent states are eigenstates of the annihilation operator, $a\,|\alpha\rangle = \alpha\,|\alpha\rangle$ and $\langle\beta|\,a^\dagger = \beta^*\,\langle\beta|$, so on the coherence $\rho = |\alpha\rangle\,\langle\beta|$ every term of Equation (87.3) acts by multiplication:

$$\dot\rho = \kappa\Big(\alpha\beta^* - \tfrac{1}{2}\,|\alpha|^2 - \tfrac{1}{2}\,|\beta|^2\Big)\rho\,. \tag{87.4}$$

The imaginary part of the bracket only turns the phase. Its real part is $\mathrm{Re}(\alpha\beta^*) - \frac{1}{2}\,|\alpha|^2 - \frac{1}{2}\,|\beta|^2 = -\frac{1}{2}\,|\alpha-\beta|^2$, so the modulus of the coherence decays at the rate

$$\Gamma_{\mathrm{dec}} = \tfrac{1}{2}\kappa\;|\alpha-\beta|^2\;, \tag{87.5}$$

whereas the mean photon number obeys $\mathrm{d}\langle n\rangle/\mathrm{d}t = -\kappa\langle n\rangle$ and so relaxes at the rate κ. The ratio is

$$\frac{\Gamma_{\mathrm{dec}}}{\kappa} = \tfrac{1}{2}\,|\alpha-\beta|^2\;, \tag{87.6}$$

the square of the phase-space separation. At a separation of ten photon amplitudes the interference is gone fifty times faster than the energy; for a separation of macroscopic size,

where $|\alpha-\beta|^2$ counts the quanta by which the two branches differ, the factor is astronomical. This is the quantitative content of the claim that large objects look classical: nothing in Equation (87.3) changes with size except one number in an exponent, and no new physics is invoked at any size. □

87.5.2 Engineered reservoirs in ion traps

[Reserved: superpositions of motional states of a single trapped ion exposed to a deliberately applied, characterized noise, so that the environment is a control parameter rather than a nuisance [2001, 2003]; the measured decay of coherence against the amplitude of the applied noise and against the size of the superposition, confirming the predicted scaling over an order of magnitude; why an engineered reservoir is the cleanest possible test of the Lindblad form of Section 87.3.2.]

87.5.3 Matter-wave interferometry

[Reserved: interference of large molecules [1966] as the platform (theory in Chapter 71); controlled collisional decoherence, in which fringe contrast decays exponentially with background gas pressure exactly as predicted [2052]; thermal decoherence, in which the molecules are heated until they emit their own blackbody photons and destroy their own interference [2060] — decoherence with no external environment at all, the environment being the internal vibrational degrees of freedom; both are quantitative confirmations of the Joos–Zeh formula [1757].]

Phenomenon 87.4 (Interference destroyed by a record kept elsewhere)**.** A beam of large molecules crossing a grating interferometer produces fringes whose contrast can be read directly. Admitting background gas reduces that contrast exponentially in the gas pressure, with no shift of the fringe positions and no loss of molecules beyond ordinary attenuation [2052]. Heating the molecules until they emit their own thermal photons reduces it in the same way, the environment in that case being the molecule's own internal vibrational degrees of freedom, with nothing external involved at all [2060]. In both cases the contrast returns when the disturbance is removed, and in neither does anything about the molecule's own dynamics change. What is measured is therefore the export of coherence into a record kept somewhere else, not a modification of the Schrödinger equation of Chapter 77.

Derivation. Let the interferometer put the molecule in a superposition of two paths separated by Δx, and let a single environment particle scatter from it. Write the environment state after the scattering as $|\chi_1\rangle$ or $|\chi_2\rangle$ according to which path the molecule took. Linearity of the joint evolution gives

$$\tfrac{1}{\sqrt{2}}\big(|1\rangle+|2\rangle\big)\otimes|\chi_0\rangle \longmapsto \tfrac{1}{\sqrt{2}}\big(|1\rangle\otimes|\chi_1\rangle+|2\rangle\otimes|\chi_2\rangle\big)\,, \tag{87.7}$$

and tracing out the environment multiplies the molecule's off-diagonal element by $\langle\chi_1|\chi_2\rangle$. Nothing has been done to the molecule — its two path amplitudes are untouched, and its momentum is barely changed — yet the fringe contrast is now $|\langle\chi_1|\chi_2\rangle|$. When the scattered particle's wavelength is short compared with Δx the two environment states are essentially orthogonal, that overlap is essentially zero, and a single event suffices. The visibility after a time t is then simply the probability that no such event has occurred,

$$V(t)=\mathrm{e}^{-\Gamma t}\,, \qquad \Gamma = n\sigma v_{\mathrm{rel}}\,, \tag{87.8}$$

with n the number density of the gas, proportional to its pressure at fixed temperature, σ the collision cross section of Chapter 86 and v_{rel} the mean relative speed. Hence the observed exponential in pressure, with a decay constant that is an ordinary collision rate

and contains no new parameter. Two features deserve marking. This is not attenuation: the molecule is deflected by a negligible angle and still arrives, but arrives without its fringe. And the total state in Equation (87.7) stays pure throughout, so the coherence has been relocated rather than destroyed — which is the distinction of Section 87.1.3, and the reason the refocusing of Section 87.7.2 can sometimes retrieve it. □

87.5.4 Solid-state qubits

[Reserved: coherent oscillations in a single-Cooper-pair box [1980] and the coherence times of superconducting and spin qubits since — from nanoseconds to hundreds of microseconds, an improvement of five orders of magnitude achieved by removing environments one at a time; $1/f$ charge and flux noise and two-level defects as the dominant channels; why this is the branch of physics in which decoherence theory is used as an engineering tool.]

87.6 The quantum Zeno effect

87.6.1 The Misra–Sudarshan theorem

[Reserved: the theorem that for a state with a Hamiltonian of bounded second moment the short-time survival probability is quadratic in the elapsed time, $1-(\Delta E\,t/\hbar)^2$, so that N ideal projective measurements in a fixed interval leave a survival probability tending to one as $N\to\infty$ — a continuously observed unstable system never decays [1606]; the conditions under which the quadratic regime is destroyed, and the anti-Zeno regime in which measurement accelerates decay; the relation to the projection postulate of Chapter 77.]

87.6.2 Observations

[Reserved: the trapped-ion demonstration of Itano and collaborators, in which the induced transition between two hyperfine levels was progressively suppressed as the number of interrogating optical pulses grew [1833]; the objection that the result follows from ordinary unitary dynamics of the three-level system and requires no collapse postulate [1841], which this treatise states as part of the result rather than as a footnote; observation of both the Zeno and the anti-Zeno regime in the tunnelling decay of atoms from an accelerating optical lattice [2012], where the crossover between them was measured.]

Phenomenon 87.5 (Frequent interrogation suppresses a transition)**.** A driven transition between two levels of a trapped ion, which transfers the whole population in a fixed time when left alone, is progressively suppressed as the number of interrogating pulses applied during that time is increased: the population surviving in the initial level rises monotonically towards unity with the number of pulses [1833]. The effect is real and reproducible. Its interpretation is not settled by the observation, and this treatise says so as part of the result rather than in a footnote: the same numbers follow from the ordinary unitary dynamics of the three-level system with its driving and probe fields, so the experiment does not by itself demonstrate a collapse of the state, and what it does and does not decide belongs with Chapter 85.

Derivation. Let $|\psi\rangle$ be the initial state and H the generator of its evolution. Expanding $|\psi(t)\rangle = \mathrm{e}^{-\mathrm{i}Ht/\hbar}\,|\psi\rangle$ to second order in t and taking the overlap with $|\psi\rangle$, the terms linear in t cancel between the amplitude and its conjugate and there remains

$$P(t) = |\langle\psi|\psi(t)\rangle|^2 = 1 - \frac{(\Delta E)^2 t^2}{\hbar^2} + O(t^4)\,, \qquad (\Delta E)^2 = \langle H^2\rangle - \langle H\rangle^2\,. \tag{87.9}$$

The initial decay is quadratic, not linear, and that is the whole of the effect: a genuine rate would give $P(t)=1-\lambda t$, for which the argument below does nothing at all. Divide a fixed

interval T into N equal parts and project onto $|\psi\rangle$ after each. If each projection resets the state, the survival probability is the product of N independent factors,

$$P_N(T) = \left[1 - \frac{(\Delta E\, T/N)^2}{\hbar^2}\right]^N \longrightarrow \exp\left[-\frac{(\Delta E\, T)^2}{N\hbar^2}\right] \longrightarrow 1 \qquad (N \to \infty)\,, \tag{87.10}$$

so a continuously watched system never leaves its initial state. The hypotheses carry the result and must be stated with it. The second moment ΔE must be finite, which fails for a Hamiltonian unbounded in the relevant sense and is exactly the case of a decay that is exponential at all times; and the interval T/N must still lie inside the quadratic regime of Equation (87.9), which for a decay into a broad continuum is extremely short. When the second condition fails and the probe instead broadens the level into the continuum, the same analysis gives an *increase* of the decay rate — the anti-Zeno regime, which is measured in the same systems. □

87.7 Fighting decoherence

87.7.1 Decoherence-free subspaces

[Reserved: symmetry of the system–environment coupling leaving a subspace on which the dissipator acts trivially [1957]; the singlet of two qubits under collective dephasing as the minimal example; its realization as a memory in two trapped ions, with a measured coherence time an order of magnitude longer than that of a bare qubit [2015].]

87.7.2 Dynamical decoupling

[Reserved: Hahn's spin echo [959], in which a refocusing pulse recovers coherence lost to static inhomogeneity — the direct demonstration that such loss is not irreversible, and the sharpest illustration of Section 87.1.3; the generalization to pulse sequences that average an arbitrary slow environment to zero [1983]; the filter-function picture relating a sequence to the noise spectrum it suppresses, which turns a qubit into a spectrometer of its own environment.]

Phenomenon 87.6 (Coherence recovered by a refocusing pulse)**.** An ensemble of spins precessing in a slightly inhomogeneous static field loses its transverse magnetization in a time T_2^* fixed by the spread of the field. A single π pulse applied at time τ makes the magnetization reappear at time 2τ, essentially undiminished [959]. The echo can be repeated many times, and the envelope of successive echoes decays on a much longer time T_2. The disappearance of the signal is therefore not by itself irreversible — only the residual envelope is — and the two time scales measure two different things: the inhomogeneity of the field, and the rate at which that inhomogeneity itself fluctuates.

Derivation. Give each spin of the ensemble a detuning δ from the mean Larmor frequency, constant in time but differing from spin to spin with some distribution $p(\delta)$. In the frame rotating at the mean frequency its transverse component acquires the phase $\phi(t) = \delta t$, so the ensemble-averaged transverse magnetization at time τ is

$$\int \mathrm{d}\delta\, p(\delta)\, \mathrm{e}^{\mathrm{i}\delta\tau}\,, \tag{87.11}$$

the Fourier transform of the detuning distribution, which decays on the time T_2^* set by the reciprocal of its width. No individual spin has lost anything at that point; their phases have merely fanned out. A π pulse about an axis in the transverse plane conjugates the

accumulated phase, $\phi \mapsto -\phi$. Evolving for a further time t gives the total phase $-\delta\tau + \delta t$, and

$$\phi_{\text{total}}(2\tau) = 0 \qquad \text{for every } \delta\,, \tag{87.12}$$

simultaneously and whatever $p(\delta)$ may be, so the whole ensemble rephases and Equation (87.11) returns to its initial value. The only property of δ used is that it is the same before and after the pulse. Whatever part of δ fluctuates on the time scale τ accumulates different phases in the two halves of the interval, and for that part the cancellation is incomplete — which is precisely the separation between T_2^* and T_2, and the reason a longer τ recovers less. The echo is the sharpest available demonstration of Section 87.1.3: a state operationally indistinguishable from a classical ensemble can nevertheless be brought back, so looking like a mixture is not the same as being one. □

87.7.3 Quantum error correction

[Reserved: Shor's nine-qubit code [1909] and Steane's seven-qubit code [1931], which encode one logical qubit so that the syndrome of an error can be measured without measuring the state; the discretization of errors, whereby correcting a finite set of Pauli errors corrects a continuum; the threshold theorem as a statement about rates rather than about principles; the experimental demonstration that a logical qubit's error rate falls as the code distance grows [2415], which is the point at which the theory of this chapter became a construction rather than a limit.]

Chapter 88

Atoms and Molecules

The hydrogen atom of Chapter 81 is the last atom that can be solved exactly. Adding a single electron adds an interaction term that couples all coordinates and destroys separability, and with it the closed-form spectrum; everything beyond helium is approximation plus computation. What is remarkable is how much survives. The shell structure that the exclusion principle imposes on the one-electron orbitals [529] reproduces the periodic table — the largest single body of chemical regularity ever assembled — from counting alone, and the self-consistent field of Hartree and Fock [604, 653] turns the counting into numbers that match measured ionization energies. This chapter is where the quantum mechanics of Parts VIII and IX meets the evidence of chemistry.

Its second half joins nuclei to electrons. The Born–Oppenheimer separation [564] exploits the mass ratio to define a potential energy surface on which nuclei move, and that surface is what a chemical bond is: the electronic ground-state energy of H_2^+ as a function of internuclear distance [568], or of H_2 in the exchange treatment of Heitler and London [577]. Rotations, vibrations and electronic transitions on that surface are the molecular spectra [961], and the residual attraction between closed-shell atoms with no bond at all is London's dispersion force [661], whose retarded form [910] is the bridge to the vacuum of Part XI. The chapter is placed after Chapter 83, whose antisymmetry postulate it uses on every page, and before the precision spectroscopy of Chapter 90, which measures what is computed here.

Derivation pending #104

Atoms and Molecules: all derivations of this chapter are pending.

88.1 The many-electron problem

88.1.1 The Coulomb Hamiltonian

[Reserved: the nonrelativistic Hamiltonian of Z electrons and a nucleus, with its one-body Coulomb attraction and its pairwise repulsion $\sum_{i<j} e^2/4\pi\epsilon_0\,|\boldsymbol{r}_i-\boldsymbol{r}_j|$; why the repulsion term forbids separation of variables and hence any closed-form solution; the infinite-nuclear-mass approximation and the reduced-mass correction; the atomic units in which the literature is written, with the statement of their SI conversion required by the units axiom (Chapter 2).]

88.1.2 Antisymmetry and Slater determinants

[Reserved: the exclusion principle in its original spectroscopic formulation [529] and its restatement as antisymmetry of the total state (Chapter 83); the Slater determinant as the antisymmetric product built from one-electron spin-orbitals [637], and the Slater–Condon

rules for matrix elements of one- and two-body operators between determinants; the exchange term as a consequence of antisymmetry rather than of any force.]

88.1.3 Helium: exchange and the ortho–para splitting

[Reserved: helium as the smallest instance of everything — Heisenberg's resonance treatment, which identified the singlet–triplet splitting as the exchange integral and explained why ortho-helium and parahelium behave as two non-combining species [545]; first-order perturbation theory and the variational estimate of the ground-state energy (Chapter 82); Hylleraas's explicitly correlated wave function [628], the first calculation to reach spectroscopic accuracy and still the template for benchmark work; the measured ionization energy 24.587 eV *as the number to be reproduced.]*

88.2 Self-consistent field theory

88.2.1 The Hartree method

[Reserved: Hartree's non-Coulomb central field, in which each electron moves in the spherically averaged field of the nucleus and of all the others, solved by iteration to self-consistency [604]; the resulting effective potential interpolating between $-Ze^2/r$ near the nucleus and $-e^2/r$ far away, which is what lifts the ℓ-degeneracy of hydrogen and orders the subshells; the Hartree product's failure to be antisymmetric, and what that costs.]

Phenomenon 88.1 (The accidental degeneracy of hydrogen does not survive)**.** In hydrogen the energy depends on n alone. In every other atom it depends on n and ℓ separately, and always in the same direction: for fixed n, $E_{ns} < E_{np} < E_{nd} < E_{nf}$. The alkali spectra display this with no interpretation required [248]. Their terms are described to high accuracy by

$$E_{n\ell} = -\frac{hcR}{(n-\mu_\ell)^2}\,, \tag{88.1}$$

with a quantum defect μ_ℓ that is nearly independent of n along a series, decreases rapidly with ℓ, and is essentially zero for the highest ℓ: in sodium the s terms are displaced by more than one whole unit of n, while the d and f terms lie almost exactly on the hydrogen values. The ordering is severe enough to reorder shells — the $4s$ level lies below $3d$ — which is why the fourth period of the table is eighteen elements long and not eight.

Derivation. Model the field seen by one outer electron as a central potential $V(r)$ tending to $-Ze^2/4\pi\epsilon_0 r$ as $r \to 0$, where the nucleus is unscreened, and to $-e^2/4\pi\epsilon_0 r$ as $r \to \infty$, where the other $Z-1$ electrons screen all but one unit of charge. Interpolating between those two limits is the whole content of the self-consistent field. Write

$$\delta V(r) = V(r) + \frac{e^2}{4\pi\epsilon_0 r} \tag{88.2}$$

for the departure from a hydrogenic problem of unit effective charge. Since V is more attractive than the fully screened form everywhere and equals it only asymptotically, $\delta V(r) < 0$ at every finite r, and it is concentrated at small r, falling off with the density of the core.

First-order perturbation theory (Chapter 82) gives $\Delta E_{n\ell} = \int_0^\infty \mathrm{d}r\; |u_{n\ell}(r)|^2\, \delta V(r)$, negative for every $n\ell$. The dependence on ℓ follows from one fact about the radial functions: near the origin the regular solution behaves as $u_{n\ell}(r) \propto r^{\ell+1}$, because the centrifugal term $\hbar^2\ell(\ell+1)/2mr^2$ dominates there. So for fixed n the weight $|u_{n\ell}|^2$ carried into the region where δV lives falls monotonically as ℓ grows, and with it the size of the shift,

$$\Delta E_{ns} < \Delta E_{np} < \Delta E_{nd} < \cdots < 0\,,$$

which is the observed ordering. Rewriting the shifted level in Rydberg form gives Equation (88.1) with $\mu_\ell > 0$ decreasing in ℓ; and μ_ℓ is nearly independent of n because the small-r amplitude of $u_{n\ell}$ scales as $n^{-3/2}$ along a series, so that the shift and the level spacing scale together.

What this does *not* give is the order in which subshells fill across the table, which requires comparing different n as well as different ℓ; that ordering remains the empirical rule of Section 88.4.2, and the chapter says so there rather than pretending otherwise. □

88.2.2 Hartree–Fock

[Reserved: Fock's variational derivation from an antisymmetrized trial state [653], independently obtained by Slater [671], giving the Hartree–Fock equations with their nonlocal exchange operator; the self-consistency loop; Koopmans' theorem identifying the orbital energy with minus the ionization energy at frozen orbitals [743], and its direct test in photoelectron spectroscopy; restricted and unrestricted forms, and the basis-set expansions that make the equations a matrix problem.]

88.2.3 Correlation energy

[Reserved: the correlation energy defined as the difference between the exact nonrelativistic energy and the Hartree–Fock limit [1075]; its magnitude — a fraction of a percent of the total energy, and of the same order as the chemical bond energies that must be predicted, which is why the definition matters; configuration interaction, and the systematic hierarchies that converge to the exact answer at a computational cost growing factorially.]

88.2.4 Density functional theory

[Reserved: the Hohenberg–Kohn theorems, that the ground-state density determines the external potential up to a constant and that a universal functional of the density is minimized by it [1284]; the Kohn–Sham construction, which restores an orbital picture by mapping onto a fictitious non-interacting system with the same density [1314]; the exchange–correlation functional as the one unknown, and the honest statement that its approximations are calibrated against experiment and computation rather than derived; the direct descent of this line from the Thomas–Fermi model of Section 88.3.]

88.3 The Thomas–Fermi model

[Reserved: the statistical treatment of the electron cloud as a degenerate Fermi gas at each point, due to Thomas [588] and Fermi [598], with the local relation between density and Fermi momentum taken from Chapter 120; the resulting nonlinear equation for the screened potential and its universal dimensionless solution; the predicted total energy scaling as $Z^{7/3}$, which the exact atomic energies do follow; the proof by Lieb and Simon that the model is the exact large-Z limit [1604]; and Teller's theorem that it cannot bind molecules [1243] — a model that is asymptotically exact for atoms and predicts no chemistry at all, which is the cleanest available lesson about the reach of a limit theorem.]

Phenomenon 88.2 (Atomic binding energies grow as $Z^{7/3}$)**.** The energy required to strip a neutral atom completely — the sum of all its ionization energies — is not proportional to Z, nor to Z^2 as a naive hydrogenic count of Z electrons each bound by an energy of order Z^2 would suggest, but grows very nearly as $Z^{7/3}$ across the whole periodic table [2420]. Equivalently, the bulk of the electron cloud contracts: the radius containing a fixed fraction of the charge falls as $Z^{-1/3}$. This is why atomic volumes vary by less than one order of

magnitude from hydrogen to uranium while Z varies by ninety, and hence why chemistry takes place on a shell whose radius is almost independent of the element.

Derivation. The Thomas–Fermi energy functional has three terms: the kinetic energy of a locally degenerate Fermi gas (Chapter 120), the attraction to the nucleus, and the electrostatic self-repulsion of the cloud,

$$E[n] = C_{\mathrm{F}} \int n^{5/3} \mathrm{d}^3 r - \frac{Z e^2}{4\pi\epsilon_0} \int \frac{n(\boldsymbol{r})}{r} \mathrm{d}^3 r + \frac{e^2}{8\pi\epsilon_0} \iint \frac{n(\boldsymbol{r})\, n(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} \mathrm{d}^3 r\, \mathrm{d}^3 r' \,, \tag{88.3}$$

C_{F} being a pure number times $\hbar^2/m$. Try the scaling

$$n(\boldsymbol{r}) = Z^2\, \nu(Z^{1/3} \boldsymbol{r}) \,, \tag{88.4}$$

which is the unique power-law choice preserving the electron count: substituting $\boldsymbol{u} = Z^{1/3}\boldsymbol{r}$ gives $\int n\, \mathrm{d}^3 r = Z \int \nu\, \mathrm{d}^3 u$, so ν is normalized once and for all. Each of the three terms then carries the same power of Z. The kinetic term collects $Z^{10/3}$ from $n^{5/3}$ and Z^{-1} from the volume element: $Z^{7/3}$. The nuclear term collects Z explicitly, Z^2 from n, $Z^{1/3}$ from $1/r$ and Z^{-1} from the volume element: $Z^{7/3}$. The repulsion collects Z^4 from the two densities, $Z^{1/3}$ from the kernel and Z^{-2} from the two volume elements: $Z^{7/3}$ again. Hence

$$E[n] = Z^{7/3}\, \mathcal{E}[\nu] \,, \tag{88.5}$$

with $\mathcal{E}$ carrying no Z at all: minimizing over ν once settles every atom, and the ground-state energy is a negative constant times $Z^{7/3}$. The same substitution makes the length scale $Z^{-1/3}$, which is the contraction stated above.

Two limits bound how far this may be pressed. It is not a fit but the exact leading behaviour of the true nonrelativistic Coulomb ground-state energy as $Z \to \infty$, which Lieb and Simon later proved as a theorem. And the same model, exact in that limit, cannot bind two atoms into a molecule at any Z — Teller's theorem — so that a description asymptotically correct for atomic energies predicts no chemistry whatever. The pair of statements is the cleanest available warning that an asymptotic theorem constrains the quantity it is about and nothing else. □

88.4 The periodic table

88.4.1 The building-up principle

[Reserved: Bohr's building-up principle, in which the elements are constructed by adding electrons to successive orbits and the periodicity of chemical behaviour is read off the closing of groups [498]; Pauli's exclusion principle supplying the missing reason why a group closes at all [529], with the counting $2(2\ell+1)$ per subshell; the shell capacities 2, 8, 18, 32 and the observed period lengths; the noble gases and the alkali metals as the extreme cases of the same count.]

88.4.2 The Madelung rule and its exceptions

[Reserved: the empirical ordering of subshell filling by increasing $n+\ell$, and by increasing n within equal $n+\ell$, stated by Madelung [780]; the partial rationalizations in terms of the effective potential of Section 88.2.1 [1649]; the roughly twenty elements that violate it, chromium and copper among them, and the fact that no derivation of the rule from the Schrödinger equation exists — stated here as an open item rather than glossed [2111].]

88.4.3 Periodic trends and their evidence

[Reserved: first ionization energies across a period and down a group, with their sawtooth at the half-filled subshells; electron affinities; atomic and ionic radii; electronegativity as a derived rather than a primitive quantity [1188]; Moseley's X-ray law, which established the atomic number as the ordering principle and closed the gaps in the table [408], with the pointer to Chapter 70 where the measurement sits. Each trend is to be given as measured data with uncertainties, not as a schematic arrow.]

88.5 Angular-momentum coupling

88.5.1 LS coupling and atomic terms

[Reserved: Russell–Saunders coupling, in which the orbital and spin angular momenta separately couple to L and S before spin–orbit combines them into J [530]; the term symbol $^{2S+1}L_J$; the addition of angular momenta of Chapter 79 applied to equivalent electrons, where antisymmetry eliminates most of the naive terms; the fine-structure interval rule and its measured departures; the alkali doublets as the simplest observed case.]

88.5.2 Hund's rules

[Reserved: Hund's empirical ordering of the terms of a configuration — maximum S, then maximum L, then minimum J for a less-than-half-filled shell [524]; their validity range, which is the ground configuration and not the excited ones; and the honest correction to the standard textbook explanation, since the detailed calculations show the first rule to be driven by reduced screening and increased electron–nucleus attraction rather than by reduced electron–electron repulsion [1736].]

88.5.3 jj coupling and the intermediate regime

[Reserved: the opposite limit, in which the spin–orbit coupling — growing roughly as Z^4 for the inner electrons — dominates the residual electrostatic interaction, so that each electron's j is good and the j values couple to J; the intermediate coupling that describes most heavy atoms, with term mixing making L and S approximate labels; the observable consequence, namely intercombination lines that are strictly forbidden in pure LS coupling and are nonetheless measured, including the clock transitions used in Chapter 90.]

88.5.4 Atoms in external fields

[Reserved: the Zeeman effect in a many-electron atom, with the Landé g-factor derived from the LS coupling scheme [478]; the Paschen–Back regime in which a strong field decouples L and S and the anomalous pattern reverts to the normal one [391]; the quadratic Stark effect for non-degenerate levels and the linear one for hydrogenic degeneracies (Chapter 81); the light shift, which is the same physics at optical frequency and is the systematic that dominates the error budgets of Chapter 90.]

88.5.5 Hyperfine structure

[Reserved: the nuclear-spin hypothesis introduced to explain the observed line satellites [516]; Fermi's contact interaction between the nuclear magnetic moment and the s-electron density at the nucleus [652], with the resulting splitting and the interval rule; the hydrogen ground-state splitting of 1420.4 MHz *and its astronomical detection as the 21-centimetre line [981], which is the survey tool of Chapter 48; isotope shifts, both mass and field, as a probe of nuclear size, feeding Chapter 90.]*

88.6 The Born–Oppenheimer separation

88.6.1 The adiabatic approximation

[Reserved: Born and Oppenheimer's expansion in the fourth root of the electron-to-nucleus mass ratio [564], giving the electronic problem at clamped nuclei, the potential energy surface as its eigenvalue, and the nuclear Schrödinger equation on that surface; the ordering of electronic, vibrational and rotational energy scales in the ratio $1 : \sqrt{m/M} : m/M$*, which is what makes a molecular spectrum look the way it does; the adiabatic theorem underlying it (Chapter 82).]*

Phenomenon 88.3 (Three separated spectral scales)**.** The spectrum of a molecule is not one ladder but three, cleanly separated: electronic transitions in the visible and ultraviolet, vibrational transitions in the infrared, rotational transitions in the microwave. Their spacings stand in the ratios

$$E_{\mathrm{el}} : E_{\mathrm{vib}} : E_{\mathrm{rot}} \approx 1 : \sqrt{m/M} : m/M \,, \tag{88.6}$$

with m the electron mass and M a nuclear mass, so that each scale is smaller than the one above it by a factor of order forty for a light molecule and more for a heavy one. This separation is the reason an electronic band carries a vibrational progression, why each vibrational band carries rotational fine structure, and why the three regions of a molecular spectrum are recorded with three different kinds of apparatus and were historically three different subjects.

Derivation. Let a be the size of the molecule. An electron confined to that size has momentum of order $\hbar/a$ by the uncertainty relation of Chapter 77, hence an energy scale

$$E_{\mathrm{el}} \sim \frac{\hbar^2}{m a^2} \,, \tag{88.7}$$

which fixes both the depth and the width of the potential energy surface, since both are electronic quantities. Displacing the nuclei by an amount of order a therefore changes the electronic energy by an amount of order E_{el}, so the curvature of the surface at its minimum is $k \sim E_{\mathrm{el}}/a^2$. Nuclei of mass M vibrating in that curvature have

$$E_{\mathrm{vib}} = \hbar\omega \sim \hbar\sqrt{\frac{k}{M}} \sim \hbar\sqrt{\frac{E_{\mathrm{el}}}{M a^2}} = E_{\mathrm{el}}\sqrt{\frac{\hbar^2}{M a^2 E_{\mathrm{el}}}} = E_{\mathrm{el}}\sqrt{\frac{m}{M}} \,, \tag{88.8}$$

the last step by Equation (88.7). Rotation of the whole molecule about its centre of mass has moment of inertia $I \sim M a^2$ and level spacing

$$E_{\mathrm{rot}} \sim \frac{\hbar^2}{M a^2} = E_{\mathrm{el}}\,\frac{m}{M} \,, \tag{88.9}$$

again by Equation (88.7). Together these are Equation (88.6).

The same small parameter justifies the separation that produced the surface in the first place. The nuclear zero-point amplitude is $x_0 = \sqrt{\hbar/M\omega}$, and dividing by a and using Equations (88.7) and (88.8) gives $x_0/a \sim (m/M)^{1/4}$. That is why the Born–Oppenheimer expansion proceeds in the *fourth* root of the mass ratio and not in the ratio itself: what must be small is the excursion of the nuclei measured against the scale on which the electronic state changes. □

88.6.2 Where the surface fails

[Reserved: the Jahn–Teller theorem, that any non-linear molecule in an orbitally degenerate electronic state distorts to remove the degeneracy [794], with its observed signatures in the

spectra of octahedral complexes; conical intersections, where two surfaces touch and the separation breaks down entirely, controlling radiationless transitions and photochemistry; the geometric phase acquired on a circuit around such an intersection, and its relation to the general Berry phase of Chapter 82.]

88.7 The chemical bond

88.7.1 The one-electron bond

[Reserved: H_2^+*, the only molecule whose electronic problem separates exactly, in prolate spheroidal coordinates; Burrau's numerical solution [568], giving the bonding and antibonding curves, the equilibrium separation* 106 pm *and the dissociation energy* 2.79 eV*; the bonding mechanism read off the exact solution — accumulation of electron density between the nuclei — with the virial analysis that attributes it correctly.]*

88.7.2 Valence bond theory

[Reserved: Heitler and London's treatment of H_2 *[577], in which a symmetrized product of atomic orbitals binds and an antisymmetrized one does not, so that the covalent bond is exchange plus spin pairing and the singlet–triplet splitting is the bond; the resulting binding energy compared with the measured* 4.478 eV*; Pauling's hybridization and resonance [686], and the systematization of directed valence [1188].]*

Phenomenon 88.4 (The covalent bond is spin-selected electrostatics)**.** Two hydrogen atoms in their ground states attract and form a stable molecule when their electron spins are opposed, and do not bind at all when the spins are parallel: the singlet potential curve has a deep minimum near an internuclear separation of 74 pm, while the triplet curve is repulsive at every distance. The difference cannot be magnetic. The bond energy is several electronvolts, whereas the magnetic interaction of two electron moments that far apart is of order 1×10^{-4} eV, four to five orders of magnitude too small. The spin therefore does not act on anything; it selects, through the antisymmetry requirement of Chapter 83, which spatial state is permitted, and the two permitted spatial states have very different electrostatic energies.

Derivation pending #105

The Heitler–London calculation: that the symmetrized and the antisymmetrized products of two hydrogenic ground-state orbitals have expectation energies differing by twice an exchange integral; that this integral has the sign, over the relevant range of internuclear distance, which puts the singlet below the separated atoms and the triplet above them; and the size of the discrepancy that remains between so simple a variational estimate and the measured dissociation energy

88.7.3 Molecular orbitals and LCAO

[Reserved: the alternative in which electrons occupy orbitals delocalized over the whole nuclear framework, developed independently by Hund [607] and Mulliken [613], with the linear-combination-of-atomic-orbitals approximation and the secular equations set out by Lennard-Jones [631]; bonding and antibonding combinations, symmetry labels, correlation diagrams, and bond order; the prediction that O_2 *has a triplet ground state — confirmed by its paramagnetism — which valence bond theory in its simple form gets wrong.]*

Phenomenon 88.5 (Molecular oxygen is paramagnetic)**.** Liquid oxygen poured between the poles of a magnet is held there. Gaseous oxygen has a magnetic susceptibility of the size characteristic of two unpaired electron spins per molecule [264], and its microwave spectrum shows the fine structure of a triplet. Every straightforward pairing of the sixteen electrons of O_2 into a closed-shell structure with a double bond predicts a diamagnetic singlet, and is wrong. This is one of the few questions on which two descriptions usually called equivalent give different answers — the simple valence-bond structure fails and the molecular-orbital picture succeeds — and it is settled by a magnet.

Derivation. Grant the molecular-orbital ordering, which puts the last two electrons of O_2 into the antibonding π_g pair. Those two orbitals are degenerate by the cylindrical symmetry of the molecule: they differ only by a rotation about the internuclear axis, so no electrostatic term can separate them. Call them ϕ_1 and ϕ_2. Two electrons in two degenerate orbitals, with the total state obliged to be antisymmetric, have exactly two possibilities: the symmetric spatial combination $\phi_1(1)\phi_2(2) + \phi_2(1)\phi_1(2)$ paired with the spin singlet, or the antisymmetric combination $\phi_1(1)\phi_2(2) - \phi_2(1)\phi_1(2)$ paired with the spin triplet. Their Coulomb expectations are $J \pm K$ with

$$K = \frac{e^2}{4\pi\epsilon_0} \iint \frac{\phi_1^*(\boldsymbol{r}_1)\phi_2^*(\boldsymbol{r}_2)\,\phi_2(\boldsymbol{r}_1)\phi_1(\boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \mathrm{d}^3 r_1\, \mathrm{d}^3 r_2\,, \tag{88.10}$$

the symmetric state at $J + K$ and the antisymmetric at $J - K$. The antisymmetric spatial function vanishes identically whenever $\boldsymbol{r}_1 = \boldsymbol{r}_2$, so two electrons in it are never found at the same point, their mutual repulsion is the smaller, and $K > 0$. The triplet therefore lies lower, O_2 has $S = 1$, and it is paramagnetic. Nothing magnetic entered the argument: the splitting is electrostatic, and the spin only labels which spatial state is allowed.

One qualification belongs with the result. The argument holds the orbitals fixed, and at that level it is exact. When the orbitals are allowed to relax the ordering survives, but its accounting changes — the dominant gain is then in the electron–nucleus attraction rather than in the electron–electron repulsion, which is the correction recorded in Section 88.5.2. The conclusion is robust; the usual one-line explanation of it is not. □

88.7.4 What the two pictures agree on

[Reserved: the demonstration that valence bond and molecular orbital descriptions of H_2 *are the same wave function in different bases once ionic and covalent terms are treated on an equal footing, and that they differ only in what they get right first — molecular orbital theory at equilibrium, valence bond theory at dissociation; the practical consequence for how molecular calculations are set up, and the standing reminder that a picture is a basis choice, not a physical claim [961, 1188].]*

88.8 Molecular spectra

88.8.1 Rotational spectra

[Reserved: the rigid rotor, its $BJ(J+1)$ *ladder and the equally spaced lines of a pure rotational spectrum in the microwave region, with centrifugal distortion as the first correction [961]; the selection rule* $\Delta J = \pm 1$ *and the requirement of a permanent dipole moment; bond lengths determined to four significant figures from measured rotational constants; the ammonia inversion doublet as the transition that ran the first maser [1070] and opened the road to Chapter 90.]*

88.8.2 Vibrational spectra

[Reserved: the harmonic approximation about the minimum of the potential energy surface, giving the normal modes of Chapter 28 applied to a molecule; the Morse potential and its exactly solvable anharmonic spectrum [632], with the convergence of levels to the dissociation limit and the Birge–Sponer extrapolation; overtones and combination bands; rotational fine structure and the P, Q and R branches of a vibration–rotation band [961]; the greenhouse-active modes of CO_2 *and* H_2O *as the worked case with the largest external stakes.]*

88.8.3 Electronic transitions

[Reserved: band systems and their vibrational progressions; the Franck–Condon principle, argued classically from the slowness of nuclear motion [542] and given its quantum form as the overlap of vibrational wave functions on the two surfaces [536], with the observed intensity distributions as its test; predissociation and photodissociation; molecular electronic spectra as the identification tool of interstellar chemistry (Chapter 52).]

88.8.4 Raman scattering

[Reserved: Smekal's prediction of inelastic light scattering with shifted frequencies [504]; its observation in liquids by Raman and Krishnan [618] and, independently and almost simultaneously, in quartz by Landsberg and Mandelstam [610] — an attribution this treatise states in full; Stokes and anti-Stokes lines and the intensity ratio as a thermometer; the polarizability selection rule, complementary to the dipole rule, so that Raman and infrared spectroscopy see different modes of a centrosymmetric molecule; the scattering formalism itself belongs to Chapter 65.]

Phenomenon 88.6 (Light scattered with a molecular frequency shift)**.** Light scattered by a transparent liquid, gas or crystal contains, besides the unshifted Rayleigh line, weak lines displaced from it by fixed frequency intervals characteristic of the scattering substance and independent of the frequency of the illumination [610, 618]. The displacements coincide with vibrational and rotational intervals of the molecule — which is what distinguishes the effect from fluorescence, whose emitted frequency is fixed rather than its shift. The lines appear in pairs, one on each side of the exciting line, and the pair is not symmetric in intensity: the line displaced to lower frequency is much the stronger, and the ratio of the two intensities depends on the temperature alone, so that the two lines together are a thermometer requiring no calibration.

Derivation. The shift itself is conservation of energy. The scattering event leaves the molecule in a different vibrational or rotational state, so $h\nu_{\text{in}} = h\nu_{\text{out}} + \Delta E$ and the displacement $\nu_{\text{in}} - \nu_{\text{out}} = \Delta E/h$ is a property of the molecule, not of the illumination. A displacement downwards — the Stokes line — requires only that the molecule start in the lower state; a displacement upwards — the anti-Stokes line — requires that it already be in the upper one.

The intensity ratio follows from that asymmetry. By microscopic reversibility the two processes have the same squared matrix element, so their rates differ only through the populations of the initial states and through the frequency dependence of dipole radiation (Chapter 65), which supplies a factor ν_{out}^4. In thermal equilibrium (Chapter 118) the population ratio is the Boltzmann factor, so

$$\frac{I_{\text{anti-Stokes}}}{I_{\text{Stokes}}} = \left(\frac{\nu_{\text{in}} + \Delta E/h}{\nu_{\text{in}} - \Delta E/h}\right)^4 \mathrm{e}^{-\Delta E/k_{\text{B}}T}\,. \tag{88.11}$$

At room temperature and a vibrational interval of a few hundred reciprocal centimetres the exponential is small, which is why the anti-Stokes lines are the weak ones and why they

vanish altogether in a cold sample. Inverting Equation (88.11) gives T; and because the two lines are recorded in one spectrum through one optical path, the detector response and the collection efficiency cancel from the ratio, which is what makes the thermometer self-calibrating. □

88.9 Intermolecular forces

88.9.1 London dispersion

[Reserved: second-order perturbation theory in the dipole–dipole interaction between two neutral, non-polar atoms, giving an attraction falling as R^{-6} with a coefficient fixed by the atomic polarizabilities [650, 661]; that the force exists between closed shells with no bond, no permanent moments and no overlap, and that it is a correlation effect invisible to Hartree–Fock; the Lennard-Jones form combining it with exchange repulsion, and the van der Waals equation of state of Chapter 54 as its macroscopic signature.]

Phenomenon 88.7 (Closed-shell atoms attract)**.** Two neutral atoms with no permanent electric moments, no unpaired electrons and no chemical affinity for one another nevertheless attract at distances well beyond the reach of their electron clouds, with an interaction energy falling as the inverse sixth power of the separation. The evidence is everywhere: the noble gases liquefy, the second virial coefficient of every real gas is negative at low temperature (Chapter 54), and even helium — the least polarizable atom there is — forms a dimer [1990]. The force is absent from any description in which each electron moves in the average field of the others, so the Hartree–Fock method of Section 88.2.2 returns exactly zero for it by construction. It is a correlation effect: an interaction between the fluctuations of two charge distributions, not between their means.

Derivation. Place the atoms at a separation R large compared with their radii, so that the charge distributions do not overlap and the interaction may be expanded in multipoles. The leading term is dipole–dipole,

$$W = \frac{1}{4\pi\epsilon_0 R^3}\Big[\boldsymbol{d}_1\cdot\boldsymbol{d}_2 - 3(\boldsymbol{d}_1\cdot\hat{\boldsymbol{R}})(\boldsymbol{d}_2\cdot\hat{\boldsymbol{R}})\Big]\,, \tag{88.12}$$

with $\boldsymbol{d}_i$ the dipole operators of the two atoms. Each atom in its ground state is spherically symmetric, so $\langle \boldsymbol{d}_i\rangle = 0$ and the first-order energy $\langle W\rangle$ vanishes identically: on the average there is nothing there to interact. Second-order perturbation theory (Chapter 82) does not vanish, and for the ground state it is negative term by term,

$$\Delta E = -\sum_{k\neq 0}\frac{|\langle k|W|0\rangle|^2}{E_k - E_0} = -\frac{C_6}{R^6}\,, \qquad C_6 > 0\,, \tag{88.13}$$

the sixth power following immediately because $W \propto R^{-3}$ and the energy is quadratic in W, and C_6 being a double sum over the dipole transitions of the two atoms — that is, an integral over their polarizabilities. Every step used only that an atom has transitions with nonvanishing dipole matrix elements, which every atom has: the attraction is universal, and it is attractive for every pair.

Two consequences are worth recording. Being second order, the process excites both atoms virtually at once, which is precisely what a one-electron mean field cannot represent, and is why the effect is invisible to Hartree–Fock and is the standard benchmark for anything claiming to include correlation. And the derivation assumed the interaction to be instantaneous. Once R exceeds the wavelength of the transitions doing the work, the light-crossing time is no longer negligible against the period of the fluctuation, the correlation is partly lost, and the attraction weakens to R^{-7} — the retarded regime of Section 88.9.2, and the point at which the electromagnetic vacuum has to enter the account. □

88.9.2 Retardation and the Casimir–Polder force

[Reserved: Casimir and Polder's result that beyond a distance of order the transition wavelength, the finite speed of light weakens the interaction to R^{-7} between two atoms and to z^{-4} between an atom and a conducting wall [910]; the direct measurement of the atom–surface potential in the retarded regime by deflection of a sodium beam through a micrometre-scale cavity [1875]; the pointer forward to the field-theoretic reading of the same result in Chapter 99.]

88.9.3 Weakly bound dimers

[Reserved: the helium dimer as the extreme test case — a molecule bound by dispersion alone, with a binding energy of order $100\,\mathrm{neV}$ *and a mean internuclear separation of about* $5\,\mathrm{nm}$*, larger than the classical turning point, measured by diffraction of a helium beam from a transmission grating [1990]; why so weakly bound an object is the sharpest available check on a computed dispersion coefficient; Efimov states and the three-body physics that follows, noted here and left to the few-body literature.]*

Chapter 89

Experiment: Bell Tests

Bell's theorem [1271] converted a philosophical dispute about the completeness of quantum mechanics [753] into an inequality between measurable correlation rates, and thereby into an experiment. This chapter is the record of that experiment: the cascade-photon tests of the 1970s [1471, 1663], the switched-analyzer measurement that first enforced relativistic separation between the choice of setting at one station and the detection at the other [1679], the two experiments that closed the locality and detection loopholes one at a time [1963, 2019], and the three independent 2015 experiments that closed them simultaneously [2253, 2254, 2260]. The theory that the data test — the derivation of the inequality, the meaning of a violation, and what is and is not entailed — is developed in Chapter 84; this chapter supplies the apparatus, the numbers and the error budgets, and states the residual assumptions without softening them.

When written the chapter will use the structured *experiment* environment, with apparatus, procedure, observations, interpretation and references as separate fields, and the tabulated CHSH parameters with their uncertainties lodged as a dataset in the evidence store. Three features distinguish a Bell test from every other experiment in this treatise and shape the sections below. Its result is a bound rather than a value, so the figure of merit is a p-value under an explicitly stated null hypothesis, not an agreement with a predicted number. Its validity rests on the spacetime layout of the apparatus, so the timing diagram is part of the physics and not part of the engineering. And one assumption — that the setting choices are not themselves correlated with the hidden variables — cannot be removed by any experiment, only pushed back in time, which is what the cosmic Bell tests do [2283, 2318]. The 2022 Nobel Prize recognized this line of work [2408]; the reviews are [2232, 2235].

Derivation pending #106

Experiment: Bell Tests: all derivations of this chapter are pending.

89.1 Historical context and the prediction under test

89.1.1 From EPR to a testable inequality

[Reserved: the Einstein–Podolsky–Rosen argument that quantum mechanics is either incomplete or non-local [753]; Bohm's reformulation with two spin-half particles in a singlet, which replaced continuous variables by a two-outcome measurement and made the problem experimental in principle [978]; Bell's derivation of an inequality obeyed by every local hidden-variable model and violated by quantum mechanics [1271], with his own later commentary on what the assumptions mean [1788]; the reduction of the whole dispute to a counting experiment.]

Phenomenon 89.1 (Perfect anticorrelation at aligned analysers)**.** When the two analysers of a Bell apparatus are set to the same direction, the two recorded outcomes are found to be opposite on every detected pair, up to the imperfections of source and detectors, while each outcome taken alone is a fair coin. This is the observation on which the Einstein–Podolsky–Rosen argument rests [753]: the result at one station can be predicted with certainty from the result at the other, at any separation, without the second particle having been touched. It is also the observation a Bell test must reproduce as a calibration before its inequality means anything, and it is measured as a matter of routine in every experiment of this chapter.

Derivation. Write the singlet $|\Psi^-\rangle = (|{\uparrow\downarrow}\rangle - |{\downarrow\uparrow}\rangle)/\sqrt{2}$ in the eigenbasis of the spin component along the common analyser direction $\hat{\boldsymbol{a}}$. Both surviving terms have the two spins opposite and no term with them parallel appears at all, so a joint measurement of that component returns $(+,-)$ or $(-,+)$ and never $(+,+)$ or $(-,-)$: the outcomes are opposite on every pair, and $E(\hat{\boldsymbol{a}},\hat{\boldsymbol{a}}) = -1$. The two terms carry equal amplitude, so each occurs with probability $\frac{1}{2}$ and the marginal recorded at either station is uniform, carrying no information about which direction was chosen. Finally, the singlet is the unique two-spin state of zero total angular momentum and is therefore invariant under simultaneous rotations of both spins, so the same expansion holds in the eigenbasis of *any* direction: the anticorrelation is not a property of a preferred axis.

Note what this does not establish. Nothing in the argument is inconsistent with local hidden variables — a pair of particles carrying opposite predetermined values reproduces every statement above exactly. That is why the settings in Section 89.1.2 must be unequal, and why the dispute could not be decided until Bell found an inequality that unequal settings violate. □

89.1.2 CHSH: the inequality that is measured

[Reserved: the Clauser–Horne–Shimony–Holt inequality $|S| \le 2$, built from four correlation measurements at two settings per side and requiring no perfect anticorrelation [1400]; the Clauser–Horne form, which uses only singles and coincidence rates and so needs no auxiliary no-enhancement assumption [1522]; Eberhard's inequality, optimized for low detection efficiency and used in the 2015 photonic experiments [1862]; the quantum maximum $S = 2\sqrt{2} \approx 2.828$ and Tsirelson's proof that no quantum state exceeds it [1650], which fixes the scale against which every measured S is read.]

89.1.3 The three loopholes

[Reserved: the detection or fair-sampling loophole, exhibited constructively by Pearle, who built a local model reproducing the quantum correlations when undetected pairs are discarded [1430], with the efficiency thresholds — about 83 % for CHSH with a maximally entangled state, about 67 % for the Eberhard inequality with a partially entangled one [1862]; the locality loophole, closed only when the setting choice at one station is spacelike separated from the detection at the other; the freedom-of-choice loophole, and the honest statement that superdeterminism is not testable and that the experiments therefore constrain, and never eliminate, it [2235].]

89.1.4 The first tests

[Reserved: Kocher and Commins's measurement of the polarization correlation of the calcium cascade photons, made for other reasons and later recognized as the first relevant data [1354]; Freedman and Clauser's dedicated test [1471], which measured $\delta = 0.050 \pm 0.008$ against the local bound $\delta \le 0$ — the first reported violation; the conflicting null result of Holt and Pipkin

and its resolution, together with Clauser's own re-measurement that traced the anomaly to the source [1576] and Fry and Thompson's fast confirmation with an isotopically selected mercury cascade [1578]. The episode is kept because a disagreement resolved by finding a systematic is better evidence than unanimity.]

89.2 Apparatus

89.2.1 Atomic cascade sources

[Reserved: the two-photon radiative cascade in calcium $4p^2\,{}^1S_0 \to 4s4p\,{}^1P_1 \to 4s^2\,{}^1S_0$, whose angular-momentum selection rules correlate the polarizations of the two emitted photons; the resonant two-photon laser excitation that raised the rate by orders of magnitude in Aspect's apparatus [1663]; the intrinsic limitation of the geometry — the correlation is degraded by the finite solid angle, since the photons are not emitted back to back — and why this made the detection loophole unclosable with cascade sources.]

89.2.2 Parametric down-conversion

[Reserved: spontaneous parametric down-conversion in a noncentrosymmetric crystal, in which one pump photon splits into a correlated pair; the type-II geometry with its two intersecting cones, which produces polarization-entangled pairs directly and brightly [1902]; phase matching, spectral filtering, and the collection optics that determine the heralding efficiency — the single parameter on which the detection loophole turns; the source used in Section 89.6.1 and, with far higher collection efficiency, in the 2015 experiments.]

89.2.3 Detectors

[Reserved: photomultipliers with quantum efficiencies of a few percent in the cascade era, which is why every experiment before 2001 assumed fair sampling; avalanche photodiodes; transition-edge sensors operating at millikelvin temperatures, which reach detection efficiencies above $90\,\%$ in the near infrared [2131] and are what made the photonic loophole-free experiments possible; state readout by resonance fluorescence in trapped ions and nitrogen–vacancy centres, which is essentially deterministic and closes the detection loophole by construction [2019, 2254].]

89.2.4 Entangled matter systems

[Reserved: entanglement between two trapped ions in the same trap [2019]; entanglement swapping through a Bell-state measurement on photons emitted by remote emitters, the event-ready scheme proposed by Żukowski and collaborators [1878] and used to entangle nitrogen–vacancy centres $1.3\,\mathrm{km}$ apart [2254]; entangled atoms in separate laboratories [2285]; superconducting circuits entangled through a cryogenic waveguide over $30\,\mathrm{m}$ [2422], where the whole spacelike-separation budget is metres rather than kilometres because the readout is nanoseconds rather than microseconds.]

89.2.5 Setting generators

[Reserved: the progression of setting choices — fixed orientations in the earliest tests; the acousto-optic switches of Aspect, Dalibard and Roger, which redirected the light periodically rather than randomly [1679]; physical random-number generators driven by a beam splitter [1963]; photons from Milky Way stars, whose emission moved the last possible common cause back centuries [2283]; high-redshift quasars, moving it back billions of years [2318]; and the

choices of about one hundred thousand human participants [2307]. Each is a statement about when *the setting was fixed, which is the only thing that matters.]*

89.3 Procedure

[Reserved: the spacetime layout, drawn as a diagram with the two stations' light cones — setting choice, analyzer setting completed, detection, and the requirement that each station's setting event be spacelike separated from the other's detection event; the coincidence window and the pitfalls of defining coincidences by post-selection on arrival times; the run structure, with the number of trials fixed in advance; the pre-registered analysis and the martingale-based hypothesis test that avoids assuming independent and identically distributed trials; how the p-value is computed and why it, and not a number of standard deviations, is the honest statistic [2253, 2260].]

89.4 Observations

[Reserved: the quantities to be tabulated for each experiment, in SI with uncertainties per [2129] — source type, station separation, detection and heralding efficiencies, trial count, the measured CHSH parameter and its uncertainty, and the p-value under local realism. The series to be given: Freedman and Clauser's $\delta = 0.050 \pm 0.008$ *[1471]; Aspect's* $S = 2.697 \pm 0.015$ *with two-channel polarizers [1680] and* $S = 2.404 \pm 0.080$ *with time-varying analyzers [1679]; Weihs's* $S = 2.73 \pm 0.02$ *under strict Einstein locality [1963]; Rowe's* $S = 2.25 \pm 0.03$ *with unit detection efficiency [2019]; the 2015 results,* $S = 2.42 \pm 0.20$ *with* $p = 0.039$ *over* $1.3\,\mathrm{km}$ *[2254],* $p \leq 3.74 \times 10^{-31}$ *[2253] and* $p \leq 2.3 \times 10^{-7}$ *[2260]. Every entry is to be read against the quantum bound* $2\sqrt{2}$ *of Section 89.1.2.]*

Phenomenon 89.2 (The measured CHSH parameter exceeds the local bound)**.** Coincidence counting on entangled pairs, with two alternative analyser settings at each station, yields values of

$$S = E(a,b) + E(a,b') + E(a',b) - E(a',b') \tag{89.1}$$

exceeding 2 in modulus — which no local hidden-variable model can produce — and consistent with the quantum prediction, whose maximum is $2\sqrt{2}$. The result has been obtained with atomic cascade photons [1471, 1680], with parametric down-conversion [1963, 2253, 2260], with trapped ions [2019], with nitrogen–vacancy centres [2254] and with superconducting circuits [2422]; at station separations from micrometres to kilometres; and with settings fixed by hand, by periodic switches, by physical random-number generators, by starlight, by quasar light and by human choices. The excess over 2 is therefore not one number from one apparatus but a stable feature of Nature reproduced on every platform on which it has been sought.

Derivation. Both numbers being compared are short to obtain. For the local bound, suppose the outcomes are functions $A(a,\lambda), A(a',\lambda), B(b,\lambda), B(b',\lambda) \in \{-1,+1\}$ of the setting on their own side and of a shared variable λ with density $p(\lambda)$, so that $E(a,b) = \int \mathrm{d}\lambda\, p(\lambda) A(a,\lambda) B(b,\lambda)$. At fixed λ, group the four terms as $S(\lambda) = A[B + B'] + A'[B - B']$. Since $B, B' \in \{-1,+1\}$, one bracket vanishes and the other has modulus 2, so $|S(\lambda)| = 2$ for every λ, and averaging gives $|S| \leq 2$. No quantum mechanics is used, which is what gives the bound its force; the same argument is set out at greater length in Chapter 84.

For the quantum value, take the singlet, whose correlation between analyser directions at angle θ is $-\cos\theta$, and choose coplanar directions $a = 0$, $a' = \pi/2$, $b = \pi/4$, $b' = -\pi/4$. Then $E(a,b) = E(a,b') = E(a',b) = -1/\sqrt{2}$ while $E(a',b') = -\cos(3\pi/4) = +1/\sqrt{2}$, so Equation (89.1) gives $S = -4/\sqrt{2} = -2\sqrt{2}$. The two statements are the whole experiment:

the apparatus is asked to produce a number, and the number it produces must fall on one side or the other of 2. Because the predicted excess is a factor $\sqrt{2}$ rather than a marginal effect, the question was answerable with the detectors of the 1970s once fair sampling was granted, and the fifty years since have gone into removing that grant rather than into improving the statistics. □

89.5 Interpretation

89.5.1 What the violations exclude

[Reserved: the exact statement — no theory in which each outcome is determined by variables carried by the particles and unaffected by the distant setting can reproduce the observed rates; that this is a constraint on explanations*, not the observation of a signal, since the marginal distributions are uniform and no information is transmitted; the no-signalling property checked directly in the data as a control; and the standing warning that "non-locality" names the failure of an assumption, not a mechanism [2232].]*

Phenomenon 89.3 (The marginals carry no message)**.** In every Bell experiment the singles rate recorded at one station is compatible, within its statistical uncertainty, with being independent of the setting chosen at the other. The comparison is made directly on the recorded data and reported as a control, so this is an experimental result and not only a theorem. Consequently a violation of Equation (89.1) is not the observation of a signal: nothing detectable at one station changes when the distant setting changes, and the correlation becomes visible only once the two records are brought together, which needs a classical channel and therefore takes at least the light travel time between the stations. The word "non-locality" in this subject names the failure of an assumption, not the discovery of an influence.

Derivation. Everything one station can record is computed from its reduced state $\rho_A = \operatorname{tr}_B \rho$. The most general thing the distant station can do to its own factor is a trace-preserving map $\mathcal{E}_B$, of which "set the analyser to b and measure, without telling anyone the result" is one instance. Since $\operatorname{tr}_B\big[(\mathrm{id}_A \otimes \mathcal{E}_B)\rho\big] = \operatorname{tr}_B \rho$ whenever $\mathcal{E}_B$ preserves the trace, the reduced state ρ_A is unchanged, and with it

$$\operatorname{tr}\big[(X \otimes \mathbb{1})\,(\mathrm{id}_A \otimes \mathcal{E}_B)\rho\big] = \operatorname{tr}\big[(X \otimes \mathbb{1})\,\rho\big] \tag{89.2}$$

for every local observable X. Only linearity and trace preservation were used, so the conclusion holds for any state and any separation. The reason the check is nonetheless performed on the data is that the *apparatus* might signal even though the physics does not — through cross-talk between the two arms, a shared clock, or a setting generator that drifts in step with a detector. A measured dependence of the marginals on the distant setting would therefore indicate a fault, and finding none is a statement about the experiment. □

89.5.2 What remains open

[Reserved: the alternatives that survive — explicitly nonlocal hidden-variable theories, retrocausal accounts, and superdeterministic models in which the setting choices are correlated with the state at preparation; the honest observation that the cosmic Bell tests constrain the last of these only by pushing the required conspiracy back in time [2283, 2318], and that no experiment can close it. The interpretive consequences belong to Chapter 85; what this chapter contributes there is the datum that any interpretation must accommodate.]

89.5.3 Why a violation is not fragile

[Reserved: the quantitative role of decoherence (Chapter 87) — entanglement degrades with photon loss, mode mismatch and dephasing, so a measured S below the Tsirelson bound is an apparatus statement rather than a physics statement; the visibility threshold $1/\sqrt{2}$ below which CHSH cannot be violated at all; why every reported violation is therefore also a bound on the imperfection of the source and detectors, and how the two are separated in the error budget.]

Phenomenon 89.4 (A violation requires a threshold quality)**.** The measured S falls below $2\sqrt{2}$ in every real apparatus, and it falls monotonically with the interference visibility of the source. No violation at all is seen once that visibility drops below $1/\sqrt{2}$, about 71 %. Photon loss, mode mismatch, dark counts, dephasing and drift all act through this one parameter, and a reported S is therefore two statements at once: one about Nature and one about the apparatus. Separating them is what the error budget of a Bell test is for, and it is the reason a value of S short of $2\sqrt{2}$ is a statement about the equipment rather than a discrepancy with the theory.

Derivation. Model the imperfection by mixing the intended singlet with unpolarized noise,

$$\rho = V\,\left|\Psi^-\right\rangle\left\langle\Psi^-\right| + (1-V)\,\tfrac{1}{4}\mathbb{1}\,, \tag{89.3}$$

where V is the visibility read off the correlation fringe. The maximally mixed part contributes nothing to any correlator, since $\operatorname{tr}\big[\mathbb{1}\,(\boldsymbol{\sigma}\cdot\hat{\boldsymbol{a}})\otimes(\boldsymbol{\sigma}\cdot\hat{\boldsymbol{b}})\big] = \operatorname{tr}(\boldsymbol{\sigma}\cdot\hat{\boldsymbol{a}})\,\operatorname{tr}(\boldsymbol{\sigma}\cdot\hat{\boldsymbol{b}}) = 0$, so every correlator is simply scaled: $E(\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}) = -V\cos\theta$. Substituting into Equation (89.1) at the optimal angles multiplies the ideal value by V,

$$S_{\max}(V) = 2\sqrt{2}\,V\,, \tag{89.4}$$

so $|S|>2$ demands $V > 1/\sqrt{2}\approx 0.707$. Below that visibility no choice of settings violates the inequality, however many pairs are counted — the deficiency cannot be bought off with statistics — and for $V \le \tfrac{1}{2}$ there is in addition an explicit local hidden-variable model reproducing every projective measurement on Equation (89.3), so the state is then provably useless for the purpose. Above the threshold the number of pairs required grows only as the inverse square of the excess of S over 2, which is why the loophole-free experiments could afford to trade brightness for efficiency. □

89.5.4 The 2022 Nobel citation

[Reserved: the award to Clauser, Aspect and Zeilinger for experiments with entangled photons establishing the violation of Bell inequalities and pioneering quantum information science [2408]; the scientific background document as a compact history of the fifty years covered by this chapter; the observation, worth stating, that the prize was given for closing loopholes rather than for a discovery of a new effect — the discipline of eliminating alternatives is itself the achievement.]

89.6 Modern repetitions and precision

89.6.1 Closing the locality loophole

[Reserved: the Innsbruck experiment, with independent quantum random-number generators, fast electro-optic modulators and a 400 m fibre separation, in which every setting choice was spacelike separated from the distant detection [1963]; the timing budget that makes the claim, given as numbers; the extension to 144 km between the Canary Islands, which also addressed the freedom-of-choice loophole [2166].]

89.6.2 Closing the detection loophole

[Reserved: the trapped-ion experiment in which every prepared pair was detected, so that no fair-sampling assumption was needed [2019] — at the cost of a separation of a few micrometres, which leaves the locality loophole wide open; why the two loopholes pulled the design in opposite directions for fourteen years, and what had to improve in detector efficiency [2131] and in entanglement distribution before both could be closed at once.]

89.6.3 The loophole-free experiments

[Reserved: the three independent 2015 experiments — Delft, with nitrogen–vacancy centres 1.3 km *apart and event-ready entanglement swapping, 245 trials and* $p = 0.039$ *[2254]; Vienna, with a high-efficiency photonic source and the Eberhard inequality [2253]; and Boulder, with transition-edge sensors and a fully pre-registered analysis [2260]; the differences in their residual assumptions, and why three experiments with different systematics are worth more than any one of them repeated.]*

Phenomenon 89.5 (Locality and detection closed at the same time)**.** In 2015 three independent experiments [2253, 2254, 2260] observed violations of Equation (89.1) while enforcing together the two conditions that had previously been met only one at a time: each station's setting was chosen and its measurement completed in a region spacelike separated from the distant station's detection event, and essentially every prepared pair yielded a recorded outcome, so that no assumption of fair sampling was needed. The three used different physical systems — nitrogen–vacancy centres on a kilometre-scale baseline with event-ready entanglement swapping at Delft, and two high-efficiency photonic sources with pre-registered analyses at Vienna and at Boulder — and therefore different systematics, so that no single misunderstood effect can account for all three. Their combined conclusion is a statement about a class of theories, and it is the strongest statement the data support: no local hidden-variable theory reproduces the observed rates. The one assumption that remains, that the setting choices are not themselves correlated with the pair at preparation, cannot be removed by any experiment and can only be pushed back in time (Section 89.6.4).

Derivation pending #107

The detection-efficiency threshold: the least heralding efficiency at which the observed coincidence and singles rates cannot be reproduced by any local model free to decide which pairs go undetected, worked out both for the CHSH inequality with a maximally entangled state and for the Eberhard inequality with a partially entangled one, the latter threshold being the lower and being the reason the photonic experiments deliberately used a non-maximally entangled state

89.6.4 Pushing back the setting choice

[Reserved: cosmic Bell tests using the colour of photons from Milky Way stars [2283] and from quasars at redshifts above 3 [2318] to choose the settings; the BIG Bell Test, in which about one hundred thousand people supplied the setting bits for thirteen experiments on five continents in a single day [2307]; what each achieves precisely — an upper bound on the time available for a common cause — and what none of them can achieve.]

89.6.5 Other platforms and longer baselines

[Reserved: atoms in separate buildings [2285]; superconducting circuits in a cryogenic link [2422], the first loophole-free test with massive solid-state systems; satellite-distributed

entanglement over 1200 km*, which extends the correlations to space without closing the loopholes [2290]; multipartite tests, where the Greenberger–Horne–Zeilinger argument turns the statistical contradiction into an all-or-nothing one [1821] and its three-photon realization confirms the prediction [2002].]*

89.6.6 From test to instrument

[Reserved: the inversion of the logic, in which a measured violation certifies rather than tests — Ekert's entanglement-based key distribution [1843]; device-independent randomness generation, certified by the impossibility of superluminal signalling [2303]; the first device-independent quantum key distribution between two trapped-ion nodes [2404]; and the point that a violation used as an instrument must be loophole-free for exactly the reasons catalogued in Section 89.1.3.]

Chapter 90

Experiment: Precision Spectroscopy and Atomic Clocks

Frequency is the quantity physics measures best. Nothing else is known to eighteen significant figures, and the reason is structural: a frequency is a count of cycles against a clock, so its uncertainty falls with averaging time rather than being set by the resolution of an instrument. This chapter is the experimental record of that advantage. It begins with the microwave measurement by Lamb and Retherford of a splitting the Dirac equation of Chapter 93 forbids [901], whose explanation by Bethe [897] started renormalized quantum electrodynamics (Chapter 100). It continues through the hydrogen $1S$–$2S$ two-photon transition, whose natural linewidth of about 1.3 Hz on a carrier of 2.5×10^{15} Hz makes it the sharpest transition in the simplest atom [2179, 2221], and through the femtosecond frequency comb [1987, 1993, 2039], which replaced building-sized frequency chains with a single laser and made optical clocks possible. It ends with clocks whose systematic uncertainty is below 10^{-18} [2315, 2332, 2401].

When written the chapter will use the structured *experiment* environment, with apparatus, procedure, observations, interpretation and references as separate fields, and the measured frequencies with their uncertainty budgets lodged as a dataset in the evidence store. Its place in the book is as the quantitative test bench for Chapters 81 and 88: the atomic theory of those chapters is where the numbers come from, and the disagreements recorded here — above all the proton-radius puzzle [2164, 2210], which moved the accepted proton charge radius by seven standard deviations — are what a 10^{-15} measurement can do to a theory. The chapter also supplies evidence used elsewhere: the constancy of the fundamental constants (Chapter 2) [2134], the gravitational redshift of Chapter 42 measured across a millimetre [2401], and the CPT test of Chapter 105 in antihydrogen [2299]. The review is [2256].

Derivation pending #108

Experiment: Precision Spectroscopy and Atomic Clocks: all derivations of this chapter are pending.

90.1 Historical context and the prediction under test

90.1.1 The degeneracy the Dirac equation predicts

[Reserved: the Dirac spectrum of hydrogen (Chapter 93), in which the energy depends only on n and j, so that $2S_{1/2}$ and $2P_{1/2}$ are exactly degenerate; the pre-war optical evidence, which was consistent with the degeneracy to within its resolution; why an optical measurement

could not settle it and a radio-frequency measurement between the two levels could.]

90.1.2 The Lamb shift

[Reserved: the Lamb–Retherford experiment, in which a beam of metastable $2S_{1/2}$ hydrogen atoms was driven to $2P$ by a microwave field and the quenching detected as a drop in the beam current, giving a splitting of about 1000 MHz *where theory predicted zero [901]; Bethe's nonrelativistic estimate, made within weeks, which located the effect in the electron's interaction with its own radiation field and got the number essentially right by a mass subtraction [897]; the modern value near* 1057.8 MHz*, and the pointer to Chapters 100 and 112.]*

90.1.3 What a frequency measures

[Reserved: the chain from a measured transition frequency to a constant of nature — the Rydberg constant $R_\infty = (10\,973\,731.568\,157 \pm 0.000\,012)/\mathrm{m}$ as the most precisely known of the CODATA inputs [2392, 2431], the proton charge radius entering through the finite-size shift of the S levels, and the fine-structure constant entering through the QED terms; that a single transition determines a combination and not a constant, so at least two independent measurements are always needed; the definition of the second in the SI [2333] as the fixed point of the whole construction.]

Phenomenon 90.1 (Frequency uncertainty falls as the inverse square root of the averaging time)**.** The fractional frequency instability of an atomic clock — the deviation between successive averages of its output — falls as $\tau^{-1/2}$ with the averaging time τ, over many decades, before flattening on to a floor set by systematic effects [2315, 2332]. Nothing else in physics is measured this way. The resolution of the instrument is not the limit; a longer measurement buys precision at a known and verifiable rate, and the rate itself is checked by plotting the deviation against τ and reading off the slope. This is why frequency is the best-measured quantity there is, and why a precision spectroscopy paper is mostly about its systematic budget rather than about its statistics.

Derivation. Interrogate N uncorrelated atoms with a Ramsey sequence of free-evolution time T inside a cycle of total duration $T_\mathrm{c} \geq T$, and steer the oscillator to the centre of the resulting fringe. Each atom yields one of two outcomes, so the transition probability is estimated with the binomial uncertainty $1/(2\sqrt{N})$ at the half-height point of the fringe, where its slope is steepest. By Phenomenon 90.2 that slope is πT per unit frequency, so one cycle determines the frequency to

$$\delta\nu \sim \frac{1}{2\pi T\sqrt{N}}\,. \tag{90.1}$$

Successive cycles are statistically independent, so after a time τ containing $M = \tau/T_\mathrm{c}$ of them the error falls as $M^{-1/2}$ and the fractional instability is

$$\sigma_y(\tau) \sim \frac{1}{2\pi\nu_0 T\sqrt{N}}\sqrt{\frac{T_\mathrm{c}}{\tau}}\,, \tag{90.2}$$

with ν_0 the transition frequency. Three readings of Equation (90.2) between them account for the shape of the whole field. The $\tau^{-1/2}$ law is nothing but the independence of the cycles, so it persists exactly until some drift correlates them, and the τ at which the measured curve departs from the slope is therefore a diagnostic. The instability is inversely proportional to ν_0, so replacing a microwave transition near 1×10^{10} Hz by an optical one near 1×10^{15} Hz buys five orders of magnitude at equal T and N — which is the entire case for optical clocks,

and the reason the counting problem of Section 90.2.5 had to be solved before they could be used at all. And T is in practice limited by the coherence of the interrogating laser rather than of the atom, which is why Section 90.2.4 is a section about lasers in a chapter about atoms. □

90.2 Apparatus

90.2.1 Resonance methods

[Reserved: Rabi's molecular-beam magnetic resonance, which turned a frequency into a deflection and a null [819]; Ramsey's method of separated oscillatory fields, whose interference fringe is narrower than the transit-time limit of a single interaction region and which is the operating principle of every atomic clock since [968]; the hydrogen maser as an active oscillator on the ground-state hyperfine transition [1182]; the microwave method of [901] as the same idea applied to an excited state.]

Phenomenon 90.2 (Separated fields give a fringe narrower than the interaction)**.** Splitting the interrogation into two short pulses separated by a free interval T, instead of applying one continuous pulse, produces a transition probability that oscillates with the detuning with period $1/T$. The central fringe of that pattern is narrower than the resonance obtained from a single interaction of the same total duration, and narrower than the transit-time width of either pulse taken alone: the width is fixed by the time the atom spends *undisturbed* between the pulses, not by the time it spends interacting. Every atomic clock built since 1950 rests on this, and the SI second [2333] is realized with it.

Derivation. Treat the two levels as a spin one-half in the frame rotating at the drive frequency, where a detuning δ in angular measure appears as a field along the polar axis and the drive as a field in the equatorial plane. A pulse of area $\pi/2$, short enough that δ may be neglected during it, carries the initial state from the pole to the equator, $|0\rangle \mapsto (|0\rangle + |1\rangle)/\sqrt{2}$. During the free interval the two components acquire a relative phase δT,

$$\tfrac{1}{\sqrt{2}}\big(|0\rangle + |1\rangle\big) \longmapsto \tfrac{1}{\sqrt{2}}\big(|0\rangle + \mathrm{e}^{\mathrm{i}\delta T}\,|1\rangle\big)\,, \tag{90.3}$$

and a second identical pulse converts that phase back into a population,

$$P_1(\delta) = \tfrac{1}{2}\big[1 + \cos(\delta T)\big]\,. \tag{90.4}$$

The fringes are periodic in δ with period $2\pi/T$, that is $1/T$ in ordinary frequency; the central fringe has full width at half maximum $1/2T$; and the steepest slope, at the half-height point, is πT per unit frequency, which is the number Phenomenon 90.1 needs. Nothing in Equation (90.4) refers to the pulse duration at all.

Two consequences constrain the apparatus. The pattern is periodic, so the resonance must first be located to better than $1/T$ before the central fringe can be identified, which is why an interrogation is stepped down in T from a broad line. And any phase acquired between the pulses that is not δT — from a stray field, a residual motion, or an imperfect drive — is indistinguishable from a frequency offset. That is the reason the systematic budget of Section 90.3 is, item for item, a list of phases. □

90.2.2 Doppler-free laser spectroscopy

[Reserved: the first-order Doppler width, which at room temperature exceeds the natural width of an optical transition by four orders of magnitude and is the obstacle every technique here addresses; saturated-absorption spectroscopy; Doppler-free two-photon excitation, in which counter-propagating beams cancel the linear shift for every velocity class at once — applied to hydrogen $1S$–$2S$ by Hänsch and collaborators [1546]; the residual second-order (time-dilation) shift, which is not *cancelled and which is why the atoms must also be slowed.]*

90.2.3 Cooling and trapping

[Reserved: the proposal to cool a gas by resonant radiation pressure [1547]; deceleration of an atomic beam by frequency-chirped light [1700]; three-dimensional optical molasses [1750]; the magneto-optical trap; the optical lattice at the magic wavelength, at which the light shifts of the two clock levels are equal so that trapping does not move the clock transition [2053]; cryogenic hydrogen beams and the nozzle temperature that sets the residual second-order Doppler shift in [2179].]

Phenomenon 90.3 (A trap that does not move the clock)**.** Atoms held in an optical lattice have both clock levels displaced by the trapping light, and at a general trap wavelength the two displacements differ, so a trapped atom does not have the transition frequency of a free one. At one particular wavelength, different for each species and found by measurement, the two displacements are equal [2053, 2084]: the transition frequency of the trapped sample then agrees with the free-atom value, and what residual dependence on the trap depth remains is small enough to be extrapolated to zero. This is what permits some 10^4 atoms to be interrogated simultaneously while confined tightly enough that the recoil and the first-order Doppler shift both vanish, without the confinement itself displacing the clock.

Derivation. A non-resonant field of amplitude E_0 and angular frequency ω displaces a level i by the second-order energy

$$U_i(\omega) = -\tfrac{1}{4}\alpha_i(\omega)\,E_0^2\,, \tag{90.5}$$

where α_i is the dynamic polarizability, a sum over the dipole transitions out of level i weighted by $(\omega_{ki}^2 - \omega^2)^{-1}$. The displacement of the clock transition is the difference,

$$\Delta U(\omega) = -\tfrac{1}{4}\big[\alpha_e(\omega) - \alpha_g(\omega)\big]E_0^2\,, \tag{90.6}$$

which vanishes wherever the two polarizability curves cross, and does so at every intensity: the condition $\alpha_e = \alpha_g$ fixes a wavelength and says nothing whatever about the trap depth. Such a crossing exists because the two levels are connected by dipole transitions to different sets of states, so their polarizabilities have poles at different frequencies and their difference must change sign between consecutive poles; a usable magic wavelength is a zero of that difference lying far enough from every pole that the induced scattering loss is tolerable.

The distinction that makes this an instrument rather than a trick is worth stating. A cancellation arranged by tuning an intensity would have to be maintained, and every fluctuation of the intensity would reappear as a frequency shift. Here the cancellation is a property of the frequency of the light alone — a quantity that can itself be stabilized to the precision of the clock — so the residual sensitivity to depth is second order and the leading systematic disappears from Section 90.3 rather than being corrected for. □

90.2.4 Stable lasers and optical cavities

[Reserved: the Pound–Drever–Hall scheme for locking a laser to a reference cavity [1718], the same technique that holds the interferometers of Chapter 47 on resonance; the thermal noise floor of the mirror coatings, and the cryogenic single-crystal silicon cavity that reached a linewidth below 40 mHz *[2195]; why the interrogation laser, not the atom, has been the limiting element of optical clocks.]*

90.2.5 The frequency comb

[Reserved: the spectrum of a mode-locked femtosecond laser as a comb of lines at $f_n = n f_{rep} + f_{ceo}$, so that an optical frequency is measured by counting two radio frequencies; self-referencing by octave-spanning supercontinuum, which measures f_{ceo} directly [1993]; the first direct microwave-to-optical link [1987]; the review that established the method [2039];

comb transfer over a 920 km *fibre link between Garching and Braunschweig, which let the hydrogen transition be compared with a primary caesium standard without moving either [2221].]*

Phenomenon 90.4 (A pulse train is a comb of exactly equidistant optical lines)**.** The spectrum of a mode-locked femtosecond laser consists of hundreds of thousands of sharp lines spanning hundreds of terahertz, and their spacing is uniform across the whole span to the precision with which it can be compared. Each line sits at

$$f_n = n\, f_{\mathrm{rep}} + f_{\mathrm{ceo}}\,, \tag{90.7}$$

with n an integer of order 10^6 and f_{rep} and f_{ceo} both radio frequencies that an ordinary counter measures directly. An optical frequency is therefore determined by counting two radio frequencies and one integer [1987, 1993]. Before this, the gap between a caesium standard and an optical transition had to be bridged by a chain of phase-locked oscillators that filled a building and was operated by a team; the comb replaced it with one laser, and that is what made optical clocks usable rather than merely conceivable.

Derivation. Let $A(t)$ be the complex field of a single pulse — envelope times carrier — and let the train repeat with period $T = 1/f_{\mathrm{rep}}$. The carrier is not locked to the envelope: group and phase velocities in the laser cavity differ, so the carrier slips under the envelope by a fixed phase φ from one pulse to the next, and

$$E(t) = \sum_{m} A(t - mT)\, \mathrm{e}^{\mathrm{i}m\varphi}\,. \tag{90.8}$$

Taking the Fourier transform of Chapter 17,

$$\tilde{E}(\omega) = \tilde{A}(\omega) \sum_{m} \mathrm{e}^{\mathrm{i}m(\varphi - \omega T)}\,, \tag{90.9}$$

and the sum is a periodic comb of delta functions: it vanishes unless $\omega T - \varphi$ is an integer multiple of 2π, that is unless $\omega_n = (2\pi n + \varphi)/T$. In ordinary frequency this is Equation (90.7), with

$$f_{\mathrm{rep}} = \frac{1}{T}\,, \qquad f_{\mathrm{ceo}} = \frac{\varphi}{2\pi}\, f_{\mathrm{rep}}\,. \tag{90.10}$$

The pulse shape enters only through $\tilde{A}(\omega)$, which decides which lines carry appreciable amplitude and never where a line sits. The spacing is therefore exactly f_{rep} everywhere in the spectrum, with no error accumulating across the span, because it descends from a repetition period rather than from a dispersion relation.

Measuring f_{rep} is trivial. Measuring the offset is not, and the trick that makes it possible requires the comb to span an octave: frequency-double the line n and beat it against the line $2n$. Their difference is $2(n f_{\mathrm{rep}} + f_{\mathrm{ceo}}) - (2n f_{\mathrm{rep}} + f_{\mathrm{ceo}}) = f_{\mathrm{ceo}}$, in which the unknown integer has cancelled identically. That self-referencing is why an octave-spanning supercontinuum is not a convenience but the enabling step. □

90.3 Procedure

[Reserved: the measurement cycle — prepare, interrogate with a Ramsey sequence, read out, steer the laser to the line centre, repeat; line-centre determination and the statistical uncertainty falling as the inverse square root of averaging time; the systematic budget, which is the real content of a precision measurement and is to be tabulated in full: second-order Doppler, recoil, DC and AC Stark, Zeeman, blackbody radiation, collisional, servo and gravitational shifts, each with its measured or bounded size [2179, 2256]; extrapolation to zero density and to zero light intensity; the practice of splitting the analysis between groups and of publishing the budget before the value.]

90.4 Observations

[Reserved: the quantities to be tabulated in SI with uncertainties per [2129]. The hydrogen $1S$–$2S$ centroid frequency $(2\,466\,061\,413\,187\,035 \pm 10)$ Hz, a fractional uncertainty of 4.2×10^{-15} [2179], and its confirmation through the fibre link [2221]; the $2S$–$4P$ and $1S$–$3S$ transitions as the second input needed to separate R_∞ from the proton radius [2281, 2311]; the $2S$–$2P$ Lamb shift measured in electronic hydrogen [2331] and in muonic hydrogen [2164, 2210]; the derived root-mean-square charge radius, $(0.841\,84 \pm 0.000\,67)$ fm from muonic hydrogen against the then-accepted (0.8768 ± 0.0069) fm from electronic hydrogen and electron scattering; clock fractional uncertainties — caesium fountains at a few parts in 10^{16} [2321], Al^+ at 9.4×10^{-19} [2332], ytterbium at 1.4×10^{-18} [2315].]

90.5 Interpretation

90.5.1 Tests of quantum electrodynamics

[Reserved: hydrogen as the system in which a bound-state QED calculation and a measurement meet with no adjustable parameters beyond R_∞ and the proton radius; the hierarchy of contributions — self-energy, vacuum polarization, recoil, two-loop terms — and the order at which theory rather than experiment currently limits the comparison; the complementary tests in the electron anomalous moment [2256] and in Chapter 112, and in helium (Section 90.6.3).]

90.5.2 The proton-radius puzzle

[Reserved: the discrepancy as it arose — muonic hydrogen giving a radius smaller by about 4 %, some seven standard deviations from the CODATA value, with the muon's larger overlap with the nucleus amplifying the finite-size shift by roughly $(m_\mu/m_e)^3$ [2164, 2210]; the subsequent electronic measurements that moved toward the small radius [2281, 2331, 2358], the low-momentum-transfer scattering experiment that agreed with them [2344], and the $1S$–$3S$ result that did not [2311]; the CODATA adjustment that followed [2431]; the honest verdict, that the puzzle is largely resolved in favour of the small radius and that the remaining tension is a datum, not an embarrassment to be smoothed.]

90.5.3 Constants, and whether they are constant

[Reserved: the comparison of two optical clocks with different sensitivity to the fine-structure constant as a null test of its variation, and the aluminium–mercury ratio measurement that bounded $\dot\alpha/\alpha$ at the 10^{-17} level per year [2134]; the improved bounds from clock networks and the associated limits on violations of local position invariance [2385]; the relation to the astrophysical claims and to the Oklo bound already assessed in [2183]; why a laboratory null result at this precision is the strongest available statement.]

Phenomenon 90.5 (The fine-structure constant does not measurably change)**.** Two optical clocks whose transition frequencies depend differently on the fine-structure constant have been compared over long periods, and their ratio is found not to drift. The comparison of an aluminium-ion clock with a mercury-ion clock bounds the fractional rate of change of α at the level of parts in 10^{17} per year [2134], and it is consistent with the geological and astrophysical bounds assessed in [2183]. What is measured is a ratio of two frequencies in one laboratory, so no assumption about the constancy of any unit enters anywhere — which is the point, since a dimensional constant has no observer-independent rate of change at all and only a dimensionless one can be said to drift.

Derivation. Any atomic transition frequency may be written

$$\nu_i = c\,R_\infty\,A_i\,F_i(\alpha)\,, \tag{90.11}$$

with A_i a pure number fixed by the level structure and F_i collecting the relativistic and radiative corrections, which are the only place α enters once cR_∞ has been factored out. Define the sensitivity $K_i = \partial \ln F_i / \partial \ln \alpha$, computed once per species and large for heavy atoms, where the relativistic corrections grow steeply with nuclear charge. In the ratio of two such frequencies the factor cR_∞ cancels, and with it every dimensional quantity in the problem, leaving

$$\frac{\mathrm{d}}{\mathrm{d}t} \ln \frac{\nu_1}{\nu_2} = (K_1 - K_2)\,\frac{\dot\alpha}{\alpha}\,. \tag{90.12}$$

A bound on the drift of the measured ratio therefore converts directly into a bound on $\dot\alpha/\alpha$, divided by the difference of the two sensitivities — which is why the experiment pairs species chosen to make $K_1 - K_2$ as large as possible rather than to make either clock individually best.

Two features of Equation (90.12) carry the argument. Because it is a ratio it is a dimensionless statement, immune to the objection that the second might be drifting along with everything else. And because it is tested against zero rather than against a predicted value, its reach improves twice over: through the statistical uncertainty of each comparison, which falls by Phenomenon 90.1, and through the span of time over which the comparisons are separated, which enters the slope directly. □

90.5.4 Relativity measured with clocks

[Reserved: the Gravity Probe A hydrogen-maser rocket flight, confirming the gravitational redshift to about $0.007\,\%$ *[1660]; the demonstration of both the velocity and the height effect with optical clocks at everyday scales — speeds of* $10\,\mathrm{m/s}$ *and height differences of* $0.33\,\mathrm{m}$ *[2154]; the resolution of the redshift across a single millimetre of a strontium sample [2401]; the consequence that relativistic geodesy is now a measurement technique (Chapters 41 and 42), and that a clock comparison at* 10^{-18} *requires the geoid height of each clock to a centimetre.]*

Phenomenon 90.6 (Two clocks one metre apart run at different rates)**.** Two clocks of identical construction, at rest in the laboratory and separated vertically, do not keep the same time: the higher runs faster, by the fractional amount

$$\frac{\Delta\nu}{\nu} = \frac{g\,\Delta h}{c^2}\,, \qquad \frac{g_0}{c^2} = 1.09 \times 10^{-16}/\mathrm{m}\,, \tag{90.13}$$

and the effect has been resolved at height differences of a fraction of a metre and, most recently, across a single millimetre of one trapped atomic sample [2154, 2401]. It is not a correction to be applied at the end of an analysis. At the 10^{-18} level of the best clocks the height of each clock above the geoid must be known to about a centimetre before two of them can be compared at all, so that a clock comparison has become a way of measuring that height (Chapter 42).

Derivation. Compare a clock at height h with one at the floor, in a field treated as uniform over that height, and use the equivalence principle: the laboratory is locally indistinguishable from a frame accelerating upwards with acceleration g. A signal of frequency ν emitted downwards takes a time h/c to arrive, and during that time the receiver has acquired an upward speed $v = gh/c$ relative to the frame in which the emitter was at rest at the moment of emission. The first-order Doppler shift then makes the received frequency higher by

$$\frac{\Delta\nu}{\nu} = \frac{v}{c} = \frac{gh}{c^2}\,, \tag{90.14}$$

which is Equation (90.13). Inserting the conventional standard value $g_0 = 9.806\,65\,\mathrm{m/s^2}$ and the exactly defined $c = 299\,792\,458\,\mathrm{m/s}$ [2333] gives $g_0/c^2 = 1.09 \times 10^{-16}/\mathrm{m}$, the number quoted — a shift that the best clocks resolve in seconds and that was, within living memory, at the edge of what a Mössbauer experiment could reach over the height of a tower.

Two limits on the reading. The derivation is first order in gh/c^2 and uses only the equivalence principle together with special relativity, so what it tests is the universality of free fall and not any field equation; the general-relativistic treatment, and what a measured violation would mean, belong to Chapter 42. And the comparison is one of *rates*, not of readings, so it is unaffected by the arbitrary epoch at which either clock was started — which is exactly why an atomic transition, whose frequency is fixed by physics rather than by a convention, is the right instrument for the measurement. □

90.6 Modern repetitions and precision

90.6.1 Optical clocks

[Reserved: the two families — single trapped ions, with essentially no collisional shift but a small signal [2134, 2332], and neutral atoms in a magic wavelength lattice, with 10^4 atoms interrogated at once [2053, 2084, 2315, 2401]; quantum logic spectroscopy, which reads out a clock ion through a co-trapped logic ion [2332]; the blackbody shift as the dominant systematic for most species and the cryogenic environments that suppress it; the stability–accuracy distinction, which is routinely conflated and must not be here [2256].]

90.6.2 Redefining the second

[Reserved: Essen and Parry's caesium standard [1064] and the 1967 definition it enabled [2333]; the accuracy of present primary fountains [2321], now two orders of magnitude behind the best optical clocks; the recommended frequency values for secondary representations of the second [2319] and the criteria a redefinition must satisfy; the practical obstacle, which is not clock performance but the comparison links between national laboratories.]

90.6.3 Two-electron and exotic atoms

[Reserved: helium as the two-electron benchmark, with the $2\,^3S_1 \to 2\,^1S_0$ transition measured in a quantum degenerate gas [2182] and the $2\,^3S$–$2\,^3P$ interval as a route to the fine-structure constant [2291]; muonic helium, giving the alpha-particle charge radius [2384]; positronium, a purely leptonic atom with no nuclear structure at all, in the $1\,^3S$–$2\,^3S$ interval [1864] and in the $n = 2$ fine structure [2359]; muonium, in the ground-state hyperfine interval [1978] and in the $1S$–$2S$ transition [1999]. Each removes one source of theoretical uncertainty that hydrogen carries.]

90.6.4 Antihydrogen

[Reserved: trapped antihydrogen and the measurement of its $1S$–$2S$ transition, consistent with hydrogen at the level of 2×10^{-12} [2299] — the sharpest existing test of CPT symmetry in a bound system (Chapter 105); the production, trapping and laser-cooling chain that makes a spectroscopy experiment out of a few thousand atoms per run; the gravitational free-fall measurement as the complementary programme.]

90.6.5 Toward a nuclear clock

[Reserved: the anomalously low-lying isomeric state of $^{229}\mathrm{Th}$, at about $8.4\,\mathrm{eV}$, the only nuclear transition within reach of a laser; its direct laser excitation [2428] and the measurement

of its frequency ratio against a strontium optical clock [2429]; why a nuclear transition is expected to be far less sensitive to external fields and far more sensitive to a variation of the constants of Section 90.5.3 than any electronic one; the status, which is a demonstrated transition and not yet a clock.]

Chapter 91

Experiment: Quantum Tunnelling Observed

Chapter 78 makes a prediction that has no classical counterpart at all. A particle whose energy lies below the top of a potential barrier still appears on the far side of it, with a probability that falls exponentially in the width of the barrier and in the square root of its height. Nothing in Newtonian mechanics permits this: a classical particle turns around at the point where its kinetic energy vanishes, and the region beyond is not merely improbable but inaccessible. The quantum statement is not a small correction to that picture; it contradicts it outright.

This chapter reports the measurements. They are unusually varied for a single prediction. The barriers range from the Coulomb repulsion of a heavy nucleus, a few times $10^{-14}\,\mathrm{m}$ across and tens of megaelectronvolts high, to a vacuum gap of a few tenths of a nanometre and a few electronvolts high; the particles are alpha particles, electrons at a metal surface, electrons in a semiconductor band, and quasiparticles in a superconductor; the observables are half-lives, current–field curves, current–voltage curves, and pictures of individual atoms. Between the shortest and the longest alpha-decay half-life lies a factor of about 10^{24}, and one exponential accounts for it. Two of these experiments, moreover, did not merely confirm the effect: they inverted it into instruments — tunnelling spectroscopy of the superconducting gap and the scanning tunnelling microscope — so that the least intuitive prediction of quantum mechanics is today a routine laboratory tool. The pair tunnelling of Cooper pairs across the same kind of junction is the Josephson effect, measured in Chapter 131.

91.1 The prediction under test

Every experiment in this chapter measures the same quantity in a different guise, so it is worth writing it once. Let a particle of mass m and energy E meet a one-dimensional barrier $V(z) > E$ between the turning points z_1 and z_2. In the forbidden region the Schrödinger equation has no oscillating solution; the wavefunction decays with the local rate

$$\kappa(z) = \frac{\sqrt{2m\,[V(z) - E]}}{\hbar}\,, \tag{91.1}$$

and the semiclassical transmission probability of Chapter 82 is

$$T \simeq \mathrm{e}^{-G}\,, \qquad G = 2\int_{z_1}^{z_2} \kappa(z)\,\mathrm{d}z\,. \tag{91.2}$$

The whole of what follows is the exponent G evaluated for different barriers. Three features of Equation (91.2) make it testable, and each of the five experiments below turns on one or more of them.

First, T is exponential in the barrier area $\int \kappa \, \mathrm{d}z$, so a modest change in the barrier produces an enormous change in the rate. That is what makes the Geiger–Nuttall relation as steep as it is, and what makes the tunnel current a length gauge of picometre resolution.

Second, T contains no temperature. A thermally activated process carries the Arrhenius factor $\mathrm{e}^{-\Delta E/k_{\mathrm{B}}T}$ and dies as the sample is cooled; a tunnelling process does not. Cold emission from a metal and the peak current of a tunnel diode both survive cooling, and that is how they are distinguished from their classical competitors.

Third, the rate depends on the states available on the far side. This is invisible in the one-body formula but decisive in a solid: the current through a junction counts the empty states the particle can arrive in, so a tunnel junction reads out a density of states. Giaever's measurement of the superconducting gap and Esaki's negative differential resistance are both consequences.

91.2 Alpha decay and the Geiger–Nuttall relation (1911, 1928)

The oldest evidence for tunnelling was in hand for seventeen years before anyone could read it. Alpha radioactivity presented two facts that classical mechanics could not hold together: the alpha particle leaves the nucleus with an energy far below the top of the electrostatic barrier it must cross, and the time it takes to do so depends on that energy with a violence that no ordinary rate law approaches.

91.2.1 Apparatus

Two independent measurements are combined, both made with the standard radioactive-source technique of the period. The range of the alpha particles from a thin deposit of a separated activity was measured in air at known temperature and pressure, either by the ionization produced in a shallow chamber as an absorbing gap was widened, or by counting scintillations on a zinc-sulfide screen as the screen was withdrawn from the source; the range in air at standard conditions is a monotone function of the emission energy and was used as its proxy. The decay constant of the same activity was obtained from the decay of its ionization current with time, or, for the long-lived parents, from the equilibrium activity of a known mass. The sources were members of the three natural series — uranium, thorium and actinium — separated by the radiochemistry of the day [372].

The second, independent, apparatus is the alpha-scattering geometry of Chapter 107: a collimated alpha beam from a strong source, a thin metal foil, and a scintillation screen viewed through a microscope on a rotating arm [379]. It does not measure a decay rate at all; it measures the height of the barrier, and it is what turns the decay data into a paradox.

91.2.2 Procedure

For each emitter the range R in air and the decay constant λ were tabulated, and $\log \lambda$ plotted against $\log R$. The claim under test was a statistical one about the family: whether the members of a series, which differ in nuclear charge, mass and structure, follow a single relation between how fast they decay and how energetically.

The scattering measurement runs the process backwards. An alpha particle fired at a heavy nucleus is turned around by Coulomb repulsion at the distance of closest approach d, fixed by

$$E = \frac{1}{4\pi\varepsilon_0}\frac{2Ze^2}{d}\,, \tag{91.3}$$

and the angular distribution follows the Coulomb law exactly, with no sign of the nuclear force, so long as d exceeds the nuclear radius. The barrier top is therefore not a theoretical construct: it is measured by the very nuclei that emit the alpha particles.

91.2.3 Observations and data

Geiger and Nuttall found that the members of each natural series fall on a straight line in the $(\log R, \log \lambda)$ plane, and that the three series give three nearly parallel lines, differing only in intercept [372]. The lines are extraordinarily steep: the observed ranges in air span a factor of about three, from a few centimetres to under a decimetre, while the decay constants span more than twenty powers of ten.

Table 91.1 restates the same fact in modern terms for six familiar emitters, so that the steepness can be read directly. The energies and half-lives are standard tabulated values [2382] quoted here for orientation, not measurements of this chapter, and are rounded; Geiger and Nuttall themselves worked with ranges rather than energies.

Table 91.1: Alpha energy and half-life for six natural alpha emitters. A factor of 2.2 in energy carries the half-life across some 24 orders of magnitude — the observation that Equation (91.4) has to explain. Values are standard tabulated ones [2382], rounded, and are not drawn from [372], which reports ranges in air.

Emitter	E_α	half-life
^{232}Th	4.01 MeV	1.40×10^{10} yr
^{238}U	4.20 MeV	4.47×10^{9} yr
^{226}Ra	4.78 MeV	1600 yr
^{222}Rn	5.49 MeV	3.30×10^{5} s
^{214}Po	7.69 MeV	1.64×10^{-4} s
^{212}Po	8.78 MeV	3.0×10^{-7} s

The scattering data supply the second number. For a heavy nucleus the Coulomb barrier evaluated at the nuclear surface reaches several times the energy of the alpha particles that the same nucleus emits, and the Coulomb scattering law is observed to hold down to distances of closest approach well inside the classical turning point of an emitted alpha [379].

Phenomenon 91.1 (Emission from below the barrier)**.** An alpha-emitting nucleus emits alpha particles whose energy lies far below the top of its own Coulomb barrier, as that barrier is measured independently by the scattering of alpha particles from the same nuclei [379]. There is no classical trajectory by which the particle can leave, and the nucleus decays regardless [372, 600].

Derivation. The inequality is established without any theory of the decay, by the scattering measurement alone. Fire alpha particles of energy E_s at nuclei of the species in question. If the observed angular distribution follows the Coulomb law, the two charges never came closer than the distance d of Equation (91.3) evaluated at E_s, and no departure from that law is seen for the fastest natural alpha particles on the heaviest nuclei [379]. The nuclear surface therefore lies inside that distance, $R_\mathrm{n} < d$; and since the Coulomb potential decreases monotonically outward,

$$V\left(R_\mathrm{n}\right) > V\left(d\right) = E_\mathrm{s}\,.$$

The barrier top thus exceeds the energy of the fastest alpha particle available for scattering, which in turn exceeds the emission energy of the great majority of the natural emitters of Table 91.1. The classical turning point of an emitted alpha lies outside the nucleus that

emitted it, with a forbidden region in between. That the nucleus decays all the same is the observation; the rate at which it does so is Phenomenon 91.2, and both are accounted for by Equation (91.4). □

Phenomenon 91.2 (The Geiger–Nuttall relation)**.** Within each of the three natural radioactive series the decay constant and the range of the emitted alpha particle satisfy $\log \lambda = A + B \log R$ with a slope B common to the three series and an intercept A characteristic of each [372]. Equivalently, the logarithm of the decay constant is linear in the reciprocal square root of the alpha energy. Across the natural emitters a factor of about two in energy moves the half-life by more than twenty orders of magnitude.

Derivation. Outside the nuclear radius R_{n} the alpha particle of charge $2e$ sees only the Coulomb field of the daughter nucleus of charge Ze,

$$V(r) = \frac{1}{4\pi\varepsilon_0}\frac{2Ze^2}{r}\,,$$

which exceeds the emission energy E everywhere inside the outer turning point b defined by $V(b) = E$. Substituting in Equations (91.1) and (91.2) and writing $b = 2Ze^2/\left(4\pi\varepsilon_0 E\right)$,

$$G = \frac{2}{\hbar}\int_{R_{\mathrm{n}}}^{b}\sqrt{2m\left(\frac{2Ze^2}{4\pi\varepsilon_0 r} - E\right)}\,\mathrm{d}r = \frac{2\sqrt{2mE}}{\hbar}\int_{R_{\mathrm{n}}}^{b}\sqrt{\frac{b}{r} - 1}\,\mathrm{d}r\,.$$

The remaining integral is elementary. With $x = R_{\mathrm{n}}/b$,

$$G = \frac{2b\sqrt{2mE}}{\hbar}\left[\arccos\sqrt{x} - \sqrt{x\,(1-x)}\right]. \tag{91.4}$$

For a barrier much wider than the nucleus, $x \to 0$, the bracket tends to $\pi/2$ and

$$G \longrightarrow 2\pi\eta\,, \qquad \eta = \frac{1}{4\pi\varepsilon_0}\frac{2Ze^2}{\hbar v}\,, \qquad v = \sqrt{\frac{2E}{m}}\,. \tag{91.5}$$

The decay constant is the number of assaults on the barrier per unit time, f, times the probability that one succeeds, so

$$\ln\lambda = \ln f - G\,,$$

and since $\eta \propto Z/\sqrt{E}$ the logarithm of the decay constant is linear in $ZE^{-1/2}$. Within a single series Z varies only slightly, and the statement collapses to the empirical relation of Phenomenon 91.2, the range in air being a monotone function of E. The prefactor f is a nuclear frequency of order $v/2R_{\mathrm{n}}$ and varies by a factor of a few across the whole family; it cannot produce twenty orders of magnitude, and it does not need to, because G does. □

The size of G is what carries the argument. For the alpha decay of $^{238}\mathrm{U}$ the velocity parameter is $v/c \approx 0.047$ and the daughter charge $Z = 90$, so $\eta \approx 28$ and the point-Coulomb exponent $2\pi\eta$ of Equation (91.5) is of order 170; the finite nuclear radius, through the bracket of Equation (91.4), roughly halves it. An exponent of order 10^2 that scales as $E^{-1/2}$ changes by tens when the energy changes by a factor of two, and e^{-G} therefore changes by tens of orders of magnitude. This is the first quantitative success of quantum mechanics inside the nucleus, and it was obtained twice over within months: by Gamow [600] and, independently, by Gurney and Condon [603, 625].

91.2.4 Interpretation

The two phenomena above are irreconcilable in classical mechanics and are two readings of one formula in quantum mechanics. Phenomenon 91.1 says the particle is where it cannot be; Phenomenon 91.2 says the rate at which it gets there is exponential in the barrier, which is the signature of Equation (91.2) and of nothing else. No rate law built on a classical escape over the barrier can reproduce the observed steepness with a reasonable nuclear temperature, and the alternative of a barrier low enough to be climbed is excluded by the scattering measurement that fixes its height.

Two consequences reach far beyond this chapter. Run the same penetration factor backwards and it governs the fusion of light nuclei at stellar temperatures, where the thermal energies are likewise far below the Coulomb barrier: this is the Gamow-peak argument of Chapter 52, first drawn by Atkinson and Houtermans within a year of Gamow's paper [622]. And the exponential sensitivity is the reason alpha half-lives are useful chronometers over geological time, since a rate this steep in energy is correspondingly insensitive to everything else, including the chemical and thermal environment.

A caution belongs here about attribution. Gurney and Condon's short note in *Nature* [603] and Gamow's paper [600] appeared within weeks of one another, and the full Gurney–Condon treatment followed the next year [625]. The qualitative insight — that radioactive decay is barrier penetration and needs no trigger — is common to both; the explicit evaluation of Equation (91.4) and its confrontation with the Geiger–Nuttall systematics is Gamow's.

91.2.5 Primary references

[372] for the empirical relation; [379] for the independent measurement of the barrier; [600, 603, 625] for the explanation; [622] for the stellar inverse. The energies and half-lives of Table 91.1 are standard tabulated values quoted at second hand and are not taken from any of these papers.

91.3 Fowler and Nordheim: cold field emission from metals (1928)

A metal emits electrons when it is hot: the thermionic current is the tail of the Fermi distribution spilling over the work function, and it vanishes as the metal is cooled. A metal also emits electrons when it is cold, if the field at its surface is strong enough — and that current does not vanish as the metal is cooled. The two emissions obey different laws, and the difference is the whole of the evidence.

91.3.1 Apparatus

A sharply pointed metal cathode — a wire etched to a tip of radius of order 100 nm, tungsten being the usual choice — facing an anode in a sealed and evacuated tube. The geometry is the point of the apparatus: the field at a tip of radius r held at potential V is of order V/r, so a few kilovolts across a centimetre gap becomes several volts per nanometre at the apex, some four orders of magnitude larger than the mean field in the tube. The emitted current is collected at the anode and measured over several decades with an electrometer; the cathode can be heated independently, so that the emission can be followed as a function of temperature at fixed field.

91.3.2 Procedure

The emitted current I is recorded against the applied voltage, and converted to a current density J and a surface field F through the tip geometry. Two tests are then applied. The first is the shape of the curve: the prediction is that $\ln\left(J/F^2\right)$ is linear in $1/F$, a form now universally called a Fowler–Nordheim plot, with a slope fixed by the work function alone. The second is the temperature dependence at fixed field, which separates the two mechanisms decisively: thermionic emission is exponential in $-1/T$ and field emission is, to first approximation, independent of T.

91.3.3 Observations and data

Cold metals emit copiously once the surface field reaches the order of $10^9\,\mathrm{V/m}$, that is a few volts across a nanometre, at temperatures where thermionic emission is unmeasurably small. Plotted as $\ln\left(J/F^2\right)$ against $1/F$ the measured curves are straight over several decades of current, and the slopes are consistent with work functions of a few electronvolts — the same work functions that the photoelectric threshold of Chapter 69 measures on the same metals. Fowler and Nordheim analysed the existing cold-emission data on exactly this basis and obtained the law from the new wave mechanics [599].

Phenomenon 91.3 (Cold field emission)**.** A metal at room temperature or below emits electrons into vacuum when the field at its surface reaches the order of $10^9\,\mathrm{V/m}$, although no electron in it has the energy to surmount the work function. The emitted current density obeys

$$J = \frac{a\,F^2}{\phi}\exp\left(-\frac{4\sqrt{2m_\mathrm{e}}\,\phi^{3/2}}{3\hbar e F}\right), \tag{91.6}$$

so that $\ln\left(J/F^2\right)$ is linear in $1/F$ with a slope fixed by the work function ϕ alone, and the current is essentially independent of temperature [599].

Derivation. Put the metal surface at $z=0$, the electron gas filled to the Fermi level, and the vacuum level a work function ϕ above it. An applied field F tilts the vacuum potential, so that an electron at the Fermi level sees

$$V(z) - E_\mathrm{F} = \phi - eFz \qquad (z>0)\,,$$

a triangular barrier of height ϕ and base ϕ/eF. The barrier does not have to be surmounted, only crossed, and Equation (91.2) gives

$$G = \frac{2}{\hbar}\int_0^{\phi/eF}\sqrt{2m_\mathrm{e}\left(\phi - eFz\right)}\,\mathrm{d}z = \frac{4\sqrt{2m_\mathrm{e}}}{3\hbar eF}\,\phi^{3/2}\,.$$

Multiplying e^{-G} by the rate at which electrons of the degenerate gas arrive at the surface — an integral over the occupied states which is algebraic, not exponential, in F and contributes the factor F^2/ϕ — gives Equation (91.6). Taking logarithms,

$$\ln\frac{J}{F^2} = \ln\frac{a}{\phi} - \frac{4\sqrt{2m_\mathrm{e}}\,\phi^{3/2}}{3\hbar e}\cdot\frac{1}{F}\,,$$

which is the straight line the experiment plots. Note what is absent: the temperature never entered, because the electrons that cross are those at the Fermi level, not those in the thermal tail. Note also the field required. The exponent falls to order 10 only when F reaches some $10^9\,\mathrm{V/m}$ for a work function of a few electronvolts, which is why cold emission is a phenomenon of sharp points and thin gaps and was never seen from a flat electrode. □

91.3.4 Interpretation

This is the first tunnelling experiment in which the barrier is under the experimenter's control. In alpha decay the barrier is whatever the nucleus provides; here the height is chosen by choosing the metal and the width is set by turning a voltage knob, and the predicted functional form in both variables can be checked. The straightness of the Fowler–Nordheim plot tests the exponent's dependence on F; the slope tests its dependence on $\phi^{3/2}$ against an independent photoelectric measurement of ϕ; and the absence of temperature dependence excludes the thermionic alternative outright.

The historical position of [599] is worth stating plainly. It is one of the earliest quantitative applications of the new quantum mechanics to a solid, and it rests on Sommerfeld's Fermi-statistics treatment of the electron gas [620], published the same year. The pre-quantum theories of cold emission, which sought to lower the barrier until classical escape became possible, could not survive a measurement of ϕ by any other route.

Two long descendants follow. The field-emission electron source, being cold, bright and small, became the standard gun of high-resolution electron microscopy; and the recognition that a vacuum gap is simply a barrier whose width can be varied continuously is the seed of Section 91.6.

91.3.5 Primary references

[599]; the underlying electron-gas statistics [620]. The empirical finding that the logarithm of the cold-emission current is linear in the reciprocal field predates the theory and is due to Millikan and his collaborators [612]; it is quoted here at second hand through [599], which analyses that data.

91.4 Esaki: the tunnel diode and negative differential resistance (1958)

Tunnelling in a solid need not cross a vacuum gap. In a semiconductor the forbidden energy gap is itself a barrier, and a p–n junction narrow enough turns it into one a carrier can cross. What Esaki found in doing so was a current that *falls* as the voltage rises — a negative differential resistance, impossible for any junction whose carriers move by drift and diffusion.

91.4.1 Apparatus

Germanium p–n junctions doped degenerately on both sides, so that the Fermi level lies inside the conduction band on the n-side and inside the valence band on the p-side. Degenerate doping is what makes the experiment: the depletion layer of a junction narrows as the doping rises, and at the impurity concentrations used the junction is of order 10 nm wide — thin enough for Equation (91.2) to give an appreciable transmission across the forbidden gap. The junctions were made by alloying, contacted for four-terminal measurement, and mounted in a cryostat so that the characteristic could be taken at several temperatures [1141].

91.4.2 Procedure

The current–voltage characteristic is recorded through forward and reverse bias, at room temperature and on cooling. Three features are looked for. In forward bias, the ordinary junction current is thermal injection over the built-in barrier and rises monotonically and exponentially; any departure from monotonicity is disqualifying for that mechanism. In reverse bias, an ordinary junction passes only a small saturation current until avalanche

breakdown; a tunnelling junction conducts freely from the smallest reverse bias. And on cooling, the thermal current must collapse while a tunnel current must not.

91.4.3 Observations and data

In forward bias the current rises steeply from zero to a peak at a bias of a few tens of millivolts, then *falls* with increasing voltage to a valley at a few hundred millivolts, and only then resumes the ordinary exponential climb of a diode. Between peak and valley the differential resistance $\frac{\mathrm{d}V}{\mathrm{d}I}$ is negative. In reverse bias the junction conducts immediately, with no threshold. The peak current is weakly dependent on temperature while the current beyond the valley follows the strongly temperature-dependent thermal law, so that cooling the sample makes the negative-resistance region more pronounced rather than less [1141].

Phenomenon 91.4 (Negative differential resistance in a degenerate junction)**.** A p–n junction degenerate on both sides and of order 10 nm wide passes, in forward bias, a current that rises to a peak at a bias of a few tens of millivolts and then decreases with increasing voltage before rising again: over a finite range of bias the differential resistance is negative. The effect is weakly temperature-dependent, unlike the thermal injection current that dominates beyond the valley [1141].

Derivation. Elastic tunnelling conserves energy, so a carrier crosses only into an empty state at its own energy; the current is therefore proportional to the overlap, in energy, between occupied states on one side and empty states on the other, weighted by the transmission e^{-G} of the junction.

Degeneracy on both sides is what creates such an overlap. On the n-side the conduction band is occupied up to a Fermi level lying above its edge; on the p-side the valence band is empty down to a Fermi level lying below its edge. At zero bias the two Fermi levels coincide, the forward and reverse flows cancel, and no net current flows. A forward bias V raises the whole n-side ladder by eV, opening a window of width eV in which occupied conduction states on the n-side face empty valence states on the p-side at the same energy: the tunnel current grows with V.

It cannot grow indefinitely. As the n-side rises further, its conduction-band edge is lifted past the p-side valence-band edge, and an electron in the window then faces the forbidden gap of the p-side, where there are no states at its energy at all. The overlap shrinks to zero, and with it the tunnel current. A current that rises from zero, is positive in between, and returns to zero must pass through a maximum and then decrease — and a decreasing current at increasing voltage is negative differential resistance. Beyond the cut-off, only the ordinary thermal injection current of a p–n junction remains, and its exponential climb produces the valley and the recovery.

The reverse-bias behaviour is the same argument with the roles exchanged: reverse bias always increases the overlap between occupied valence states and empty conduction states, so the current rises monotonically from zero bias with no threshold, which is what distinguishes this junction from an ordinary rectifier. □

Derivation pending #109

Magnitude of the interband tunnel current: the transmission of a carrier across the forbidden gap of a uniformly biased semiconductor, the imaginary wavevector inside the gap obtained by continuing the two-band dispersion relation, and the resulting exponential dependence of the peak current on the gap, the effective mass and the junction field

91.4.4 Interpretation

The observation of Phenomenon 91.4 is decisive because negative differential resistance is not a quantitative anomaly but a qualitative impossibility for transport by drift and diffusion, in which carrier flow increases with the driving field by construction. Its appearance requires a mechanism in which raising the bias can *remove* the final states, and interband tunnelling is such a mechanism because it is elastic and the semiconductor gap is empty of states. The temperature behaviour closes the argument: a thermally activated process cannot become more prominent on cooling.

The mechanism itself was not new in 1958 — Zener had proposed interband tunnelling as the cause of dielectric breakdown a quarter of a century earlier [750] — but the breakdown context gave no clean signature, since a runaway current can be produced in several ways. Esaki's contribution was to make a junction narrow enough that tunnelling dominates at low bias, where its consequences are unmistakable, and to read off from the shape of the characteristic that the current is a density-of-states overlap.

The device consequences were immediate and are treated with the band theory of Chapter 126 and the junction physics of Chapter 127: a tunnel current involves no storage of minority carriers and no transit time across a drift region, so the tunnel diode responds far faster than a conventional junction, and a negative resistance is an amplifier and an oscillator. For this chapter its importance is evidential — it is tunnelling observed in a solid, at millivolt energies, with the density of states of the far side written into the shape of the curve.

91.4.5 Primary references

[1141]; the interband tunnelling mechanism [750].

91.5 Giaever: tunnelling spectroscopy of the superconducting gap (1960)

Esaki's junction reads out the states available on the far side of a barrier. Giaever turned that incidental feature into the measurement itself: if the tunnel current counts final states, then a junction with a superconductor on one side is a spectrometer for the superconducting energy gap — a quantity of order a millielectronvolt, far below anything optical spectroscopy could then reach.

91.5.1 Apparatus

A thin-film sandwich junction evaporated on a glass substrate. An aluminium strip is deposited first and its surface oxidized in air or in an oxygen glow discharge to grow an insulating film of aluminium oxide a few nanometres thick; a lead strip is then evaporated across it, the crossing defining the junction area. The oxide is the barrier; its thickness is what sets the junction resistance, and it must be thin enough to pass a measurable current and free enough of pinholes to pass it by tunnelling rather than through metallic shorts. Contacts at the four ends of the two strips allow a four-terminal measurement, which keeps the lead resistances out of the answer. The assembly sits in a liquid-helium cryostat, and a solenoid supplies a magnetic field large enough to drive the lead film normal [1180].

The choice of metals is deliberate. Lead becomes superconducting at about 7.2 K, aluminium at about 1.2 K. At 4.2 K the lead is superconducting and the aluminium is not, so the junction has one superconducting and one normal electrode; and the superconductivity can be switched off in place, without changing the junction, by applying the magnetic field.

91.5.2 Procedure

The current–voltage characteristic of the junction is recorded in two states of the same sample: with the lead film normal, and with it superconducting. The comparison is the experiment. Any feature that appears in the second curve and not in the first belongs to the superconducting state and not to the barrier, the geometry or the contacts — an internal control of exactly the kind an oxide barrier demands, since a pinhole or a thermal effect would show in both. The measurement is then repeated at several temperatures between the base temperature and the transition, so that the temperature dependence of any threshold can be followed.

91.5.3 Observations and data

With both films normal the junction is ohmic: current proportional to voltage over the range of interest, which is itself the signature of tunnelling through a thin barrier rather than conduction through metal bridges. With the lead superconducting the characteristic changes completely at low bias. The current is strongly suppressed — reduced to a small residual set by the temperature — until the bias exceeds a threshold, above which the curve rises steeply and rejoins the normal line, displaced from it by roughly the threshold voltage. The threshold is of order a millivolt, corresponding to an energy of order 1 meV. On warming, the threshold shrinks and vanishes at the transition temperature of the lead film [1180].

Phenomenon 91.5 (A tunnel junction measures the superconducting gap)**.** In a normal-metal–insulator–superconductor tunnel junction the current is suppressed below a threshold bias and the differential conductance above it maps out the excitation spectrum of the superconductor. The threshold measures the energy gap directly; for lead it corresponds to an energy of the order of a millielectronvolt, and it closes continuously as the temperature approaches the transition temperature [1180].

Derivation pending #110

The tunnel current between a normal metal and a superconductor: the elastic-tunnelling current as the convolution of the two densities of states with the difference of the two Fermi functions [1198], the BCS quasiparticle density of states with its square-root edge at the gap, and the consequence that the differential conductance at low temperature is proportional to that density of states evaluated at the bias energy

The scale is set independently by the theory of Chapter 130. The weak-coupling BCS relation between the gap and the transition temperature, $2\Delta(0) = 3.53\,k_{\mathrm{B}}T_{\mathrm{c}}$, gives about 2.2 meV for a transition temperature of 7.2 K; lead is a strong-coupling superconductor and its actual gap lies somewhat above the weak-coupling value. The measured threshold is of that order, which is the point: a prediction of the theory published three years earlier [1102] was confronted with a direct measurement of the quantity it predicts.

91.5.4 Interpretation

Two results are obtained at once, and they support each other. The first is about tunnelling: the current through the junction is not a property of the barrier alone but of the states on either side of it, because switching the far electrode between two states of the same metal — with the barrier untouched — changes the characteristic qualitatively. The second is about superconductivity: the excitation spectrum has a gap, of the predicted size, closing at the transition temperature.

The methodological gain is larger than either. A tunnel junction whose differential conductance is proportional to a density of states is a spectrometer of resolution set by temperature, and at a few kelvin that resolution is a fraction of a millielectronvolt — three orders of magnitude finer than optical spectroscopy of the period. Tunnelling spectroscopy went on to measure the phonon structure imprinted on the superconducting gap, and so to confirm that the pairing interaction is the lattice one.

This experiment also sets the stage for the next step in the same apparatus. Replace the normal electrode with a second superconductor and the single-particle current acquires a larger threshold [1179]; but a second current, carried by pairs and flowing at zero bias, appears alongside it. That is the Josephson effect, predicted in [1231] and observed in [1246], and it is measured in Chapter 131 rather than here, because its content is the macroscopic phase rather than barrier penetration as such.

91.5.5 Primary references

[1180]; the gap prediction [1102]; the pair tunnelling that follows from the same junction [1231, 1246].

91.6 Binnig and Rohrer: the scanning tunnelling microscope (1982)

The last experiment in this chapter is the one in which tunnelling stops being the thing measured and becomes the thing measuring. The transmission through a vacuum gap falls by about an order of magnitude for every 0.1 nm of extra width. Read forwards, that is a prediction to be tested; read backwards, it is a displacement transducer of picometre sensitivity, and holding the current constant while a tip is scanned holds the gap constant to the same precision.

91.6.1 Apparatus

A sharp metal tip mounted on a piezoelectric tripod, which translates it in three orthogonal directions with sub-picometre response to an applied voltage, facing the sample across a vacuum gap of a few tenths of a nanometre. A coarse approach mechanism — in the original instrument a piezoelectric walker, since the gap must be closed from millimetres to angstroms without collision — brings the tip within tunnelling range. A bias of order a volt is applied between tip and sample and the tunnel current, of order a nanoampere, is amplified and compared with a set point; the error signal drives the vertical piezo through a feedback loop. The whole assembly is in ultra-high vacuum and, critically, isolated from vibration: the first instrument floated its stage magnetically above a superconducting bowl, and later designs used stacked springs with eddy-current damping [1682].

The vibration requirement follows from the very sensitivity that makes the instrument work. If the current is to be held to a few percent, the gap must be held to a few picometres, and a building vibrates by far more than that.

91.6.2 Procedure

Two measurements are made with the same hardware. The first tests the prediction: with the feedback loop open, the tunnel current is recorded as the tip is advanced or retracted by a known piezoelectric displacement, and $\ln I$ is plotted against the displacement. The slope gives 2κ of Equation (91.1), and hence an apparent barrier height that can be compared with the work functions of the two electrodes — a control that also detects contamination, since an adsorbed layer lowers the apparent barrier far below the clean-surface value.

The second is imaging. The feedback loop is closed, holding the current and therefore the gap constant, and the tip is raster-scanned across the surface. The voltage applied to the vertical piezo, recorded as a function of the lateral position, is the image: a contour of constant tunnel current, which is a contour of constant gap over a homogeneous surface.

91.6.3 Observations and data

The tunnel current across a vacuum gap is measurable and reproducible, and it varies exponentially with the gap width at a rate corresponding to an apparent barrier height of a few electronvolts — the order of a metal work function, as Equation (91.1) requires. Scanned images resolve monatomic steps, a few tenths of a nanometre in height, as clean terraces, and resolve surface reconstructions on the metal and intermetallic surfaces examined [1682]. Within a year the same instrument resolved the seven-by-seven reconstruction of the silicon (111) surface in real space [1715], ending a long-standing crystallographic controversy.

Phenomenon 91.6 (Vacuum tunnelling made an instrument)**.** A current flows between two electrodes separated by a vacuum gap that no electron of the available energy can cross classically, and its logarithm is linear in the gap width with a slope corresponding to a barrier of the order of the work function. Holding that current constant while a tip is scanned therefore holds the gap constant to a few picometres, which is sufficient to image the topography of a surface at the atomic scale [1682].

Derivation. The gap is a barrier of height equal to the work function ϕ and width equal to the separation d, so Equation (91.1) gives a decay rate $\kappa = \sqrt{2m_{\mathrm{e}}\phi}/\hbar$ that does not depend on d at all, and Equation (91.2) gives $I \propto \mathrm{e}^{-2\kappa d}$: the whole width dependence of the current is that of the squared amplitude of the decaying solution at the far electrode. This is the calculation carried out in Chapter 78, and its numerical content is what the instrument exploits. For a clean metal of work function about $4\,\mathrm{eV}$ the rate is $\kappa \approx 10^{10}/\mathrm{m}$, so that $2\kappa d \approx 2$ for $d = 0.1\,\mathrm{nm}$ and the current changes by roughly a factor e^{-2} per tenth of a nanometre. Inverting the relation gives the instrument: holding I to $1\,\%$ holds d to $\delta d = \delta I/(2\kappa I)$, which is about half a picometre. What this one-dimensional argument does *not* supply is the relation between the current and the surface it is imaging, and that is what the image requires. □

Derivation pending #111

The tunnel current as a probe of the local density of states: the transfer-Hamiltonian expression for the current between two weakly coupled electrodes in terms of the overlap of their surface wavefunctions [1198], and its reduction, for a tip modelled by a single spherically symmetric orbital, to the local density of states of the sample at the position of the tip and at the bias energy [1765]

91.6.4 Interpretation

The measurement of the apparent barrier height is the physics; the images are its proof by demonstration. An exponential current–distance relation with a decay rate matching the work function establishes that the transport is vacuum tunnelling and not field emission, contact conduction, or an adsorbate bridge — each of which would give a different slope or no exponential at all. That the same relation, once established, yields atomic-scale images is a much stronger statement than any single I–d curve, because it requires the exponential to hold locally, reproducibly, and over a whole surface.

It is worth stating what has happened here in the terms of Section 91.1. The exponential sensitivity that made tunnelling seem an exotic curiosity in alpha decay — a rate too small

to matter, too steep to control — is the same property that makes the scanning tunnelling microscope work. Steepness is precision when the exponent is a length you wish to measure. The most counterintuitive prediction in Chapter 78 became, within sixty years of being written down, the finest length gauge in surface physics and the basis of an entire family of scanned-probe instruments.

91.6.5 Primary references

[1682].

91.7 Summary of the evidence

Table 91.2: Tests of barrier penetration reported in this chapter. The barrier column gives its physical origin and rough scale; the last column gives the observable that no classical mechanism reproduces. The five span some seven orders of magnitude in barrier height and five in barrier width, and one exponential describes all of them.

Experiment	Barrier	Decisive observation
Geiger–Nuttall 1911, explained by Gamow and by Gurney and Condon 1928	Coulomb repulsion of the daughter nucleus; tens of MeV high, a few times 10^{-14} m wide	half-life exponential in $ZE^{-1/2}$, spanning some 24 orders of magnitude while the energy changes by a factor of two
Fowler–Nordheim 1928	work function of a metal surface tilted by an applied field; a few eV high, of order 10^{-9} m wide	$\ln(J/F^2)$ linear in $1/F$ with slope fixed by $\phi^{3/2}$, and no temperature dependence
Esaki 1958	forbidden gap of a degenerate germanium junction; of order 1 eV high, of order 10^{-8} m wide	negative differential resistance in forward bias, impossible for drift-diffusion transport
Giaever 1960	aluminium-oxide film a few nanometres thick	current suppressed below a threshold of order 1 meV that closes at the transition temperature of the lead film
Binnig–Rohrer 1982	vacuum gap of a few tenths of a nanometre; a few eV high	current exponential in the gap at the work-function rate; constant-current scanning images monatomic steps

The entries in Table 91.2 have almost nothing in common at the level of the apparatus. They were performed seventy years apart, by nuclear physicists, vacuum physicists, semiconductor engineers, low-temperature physicists and surface scientists, on barriers differing by seven orders of magnitude in height. What they share is Equation (91.2): in every case the measured quantity is exponential in $\int \kappa \, \mathrm{d}z$ across a region the particle classically cannot enter, with the mass, the height and the width entering exactly as Equation (91.1) prescribes and the temperature not entering at all.

That last point deserves the final word, because it is what makes the case unavoidable. Every classical account of these observations has to supply the particle with the missing energy from somewhere, and the only available reservoir is thermal. But the alpha half-life

of a nucleus is insensitive to the temperature of the sample; cold field emission persists at temperatures where thermionic emission is unmeasurable; the tunnel diode's peak current strengthens relative to its surroundings on cooling; Giaever's threshold and Binnig and Rohrer's images are obtained at liquid-helium temperatures and below. The classically forbidden region is entered, and it is not entered by borrowing heat. It is entered because the wavefunction of Chapter 77 does not vanish there, and Equation (91.2) says by how much.

Part X

Relativistic Quantum Mechanics

Chapter 92

The Klein–Gordon Equation

The first relativistic wave equation of quantum mechanics was written down before the nonrelativistic one, and discarded. Schrödinger set up the wave equation of a Coulomb-bound electron in relativistic form, found that it gave the wrong fine structure of hydrogen — the term Sommerfeld's old-quantum formula [441] had already fitted — and published instead the nonrelativistic limit that carries his name [557, 1692], returning to the relativistic form only in his fourth communication [559]. Within months the same equation was published by Klein [547] and, in its minimally coupled form and applied to the Compton effect, by Gordon [544]; it has carried their two names ever since. This chapter opens Part X because the equation is the simplest possible marriage of Part IX with Part IV: it is nothing but the mass-shell relation $E^2 = |\boldsymbol{p}|^2 c^2 + m^2 c^4$ of Chapter 40 with the operator substitutions of Chapter 77, and equivalently the eigenvalue equation of the first Poincaré Casimir $\hat{P}^2 = m^2 c^2$ of Chapter 94 at spin zero.

It is also where that marriage fails. Being second order in time, the equation admits both signs of the energy, and the density built from its conserved current is not positive: neither object can be read as a probability, so the equation is not a one-particle quantum mechanics. Dirac's response was a different equation (Chapter 93); the correct response, given by Pauli and Weisskopf in 1934 [748], was to keep this one and change what it describes — not a wavefunction but a *field*, whose quanta are particles and antiparticles and whose conserved density is charge rather than probability. Read that way the equation is exact and is used everywhere: it is the free part of every scalar field in Part XI, its static source solution is the Yukawa potential of nuclear physics [765], and its observed quanta are the pion [902] and the Higgs boson [2186, 2192]. Standard treatments are [1273, 1912, 2287].

Derivation pending #112

The Klein–Gordon Equation: all derivations of this chapter are pending.

92.1 The equation

92.1.1 Schrödinger's discarded first attempt

[Reserved: the historical order of events, which matters because it explains why a correct equation was abandoned — Schrödinger's relativistic Coulomb calculation of late 1925, its disagreement with the measured fine structure of hydrogen, and the nonrelativistic retreat published as the first communication [557], documented in detail by Kragh [1692]; the relativistic equation with electromagnetic coupling restored in the fourth communication [559]. The disagreement was real and its cause was spin, not relativity, as Chapter 93 shows.]

92.1.2 Klein, Gordon, and the equation itself

[Reserved: the equation $(\Box + m^2c^2/\hbar^2)\,\varphi = 0$ with $\Box = \partial_\mu\partial^\mu$, as published by Klein [547] — whose five-dimensional framing is a unification programme without evidence and is treated as such in Part XIII — and by Gordon [544]; the canonical substitutions $E \mapsto \mathrm{i}\hbar\partial_t$, $\boldsymbol{p} \mapsto -\mathrm{i}\hbar\boldsymbol{\nabla}$ applied to the squared mass-shell relation, and why squaring is the step that costs the probability interpretation; the Compton wavelength $\hbar/mc$ as the length scale the mass term introduces.]

92.1.3 The mass shell and the Poincaré Casimir

[Reserved: the equation restated group-theoretically. Every irreducible unitary representation of the Poincaré group carries $\hat{P}^2 = m^2c^2$ [849], so every *relativistic field of any spin satisfies the Klein–Gordon equation component by component; what distinguishes spin 0 is that this is the* only *condition. The classification is carried out in Chapter 94 and instantiated for spin $\frac{1}{2}$ in Chapter 93 and for higher spin in Chapter 97.]*

Phenomenon 92.1 (The relativistic mass shell)**.** Every free particle observed in Nature, of whatever spin, carries an energy and a momentum related by

$$E^2 = |\boldsymbol{p}|^2 c^2 + m^2c^4\,, \tag{92.1}$$

with a single constant m — its mass — that is the same for every particle of the species and the same in every inertial frame. The relation holds across the whole range of momenta accessible to experiment, and the masses it defines are tabulated for every known species in [2427]; its most direct laboratory face, the saturation of a charged particle's speed at c however much kinetic energy is supplied to it, was measured by Bertozzi [1272].

Derivation. Insert the plane wave $\varphi(x) = A\,\mathrm{e}^{-\mathrm{i}p\cdot x/\hbar}$, with $p\cdot x = \eta_{\mu\nu}p^\mu x^\nu = Et - \boldsymbol{p}\cdot\boldsymbol{x}$, into the Klein–Gordon equation $(\Box + m^2c^2/\hbar^2)\,\varphi = 0$ of Section 92.1.2, where $\Box = \partial_\mu\partial^\mu$. Each derivative brings down a factor, $\partial_\mu\varphi = -\mathrm{i}p_\mu\varphi/\hbar$, so that $\Box\varphi = -p_\mu p^\mu\varphi/\hbar^2$ and

$$\left(\Box + \frac{m^2c^2}{\hbar^2}\right)\varphi = \frac{1}{\hbar^2}\left(m^2c^2 - p_\mu p^\mu\right)\varphi\,.$$

A nonzero solution therefore requires $p_\mu p^\mu = m^2c^2$, which written out is $E^2/c^2 - |\boldsymbol{p}|^2 = m^2c^2$ and hence Equation (92.1). Both sides are Lorentz scalars, so every frame returns the same m. The universality is a group-theoretic fact rather than a property of spin zero: every irreducible unitary representation of the Poincaré group is labelled by the eigenvalue of the Casimir $\hat{P}^2$ [849], so a field of any spin satisfies this same equation component by component, and what singles out spin zero is only that no further condition is imposed. □

92.1.4 Lorentz covariance and the scalar transformation law

[Reserved: $\varphi'(x') = \varphi(x)$ as the defining transformation law of a scalar; invariance of the d'Alembertian under the transformations of Chapter 38; the Lagrangian density $\mathcal{L} = \frac{1}{2}\partial_\mu\varphi\,\partial^\mu\varphi - \frac{1}{2}(mc/\hbar)^2\varphi^2$ whose Euler–Lagrange equation this is, connecting to the field-theoretic variational formalism of Chapter 64; the distinction between a scalar and a pseudoscalar under parity, which is what the observed pion turns out to be (Chapter 105).]

92.2 Negative energies and the indefinite density

The two defects of the equation as a one-particle theory are usually listed separately. They are the same defect.

92.2.1 Both branches of the dispersion relation

[Reserved: plane-wave solutions $\mathrm{e}^{-\mathrm{i}p\cdot x/\hbar}$ *with* $p^0 = \pm\sqrt{|\boldsymbol{p}|^2 + m^2c^2}$*; the negative-energy branch cannot be discarded because the positive-energy solutions alone are not complete — any localized initial datum excites both — and because a Lorentz transformation cannot remove it, the two branches being separate orbits in Chapter 94; the resulting instability of any system coupled to such a spectrum, and Dirac's statement of the difficulty as the motivation for a first-order equation [596].]*

92.2.2 The conserved current and why its density changes sign

[Reserved: the current $j^\mu = \frac{\mathrm{i}\hbar}{2m}\left(\varphi^*\partial^\mu\varphi - \varphi\,\partial^\mu\varphi^*\right)$ *conserved by* $\partial_\mu j^\mu = 0$*; the density* j^0 *containing a time derivative because the equation is second order, so that* φ *and* $\partial_t\varphi$ *are independent initial data and* j^0 *takes either sign; the contrast with the positive-definite* $|\psi|^2$ *of Chapter 77 and with the Dirac current of Chapter 93 [596].]*

Phenomenon 92.2 (What is conserved is charge, not particle number)**.** The quantity conserved in the processes that spin-0 particles undergo is electric charge, which takes both signs, and not the number of particles, which does not survive: charged pions are created singly in weak decays, in pairs of opposite charge in strong collisions, and the neutral pion disappears entirely into two photons. Every charged spin-0 species observed comes with a partner of equal mass and opposite charge, while the neutral pion is its own antiparticle [2427]. No experiment counts a conserved number of spin-0 quanta; every experiment finds the total charge unchanged.

Derivation. That the equation cannot supply a conserved *probability* is immediate. Evaluate the current of Section 92.2.2 on the plane wave $\varphi = A\,\mathrm{e}^{-\mathrm{i}p\cdot x/\hbar}$ of Phenomenon 92.1. With $\partial^0 = c^{-1}\partial_t$ and $\partial_t\varphi = -\mathrm{i}E\varphi/\hbar$,

$$j^0 = \frac{\mathrm{i}\hbar}{2mc}\left(\varphi^*\partial_t\varphi - \varphi\,\partial_t\varphi^*\right) = \frac{\mathrm{i}\hbar}{2mc}\left(-\frac{2\mathrm{i}E}{\hbar}\right)|A|^2 = \frac{E}{mc}\,|A|^2\ .$$

The sign of j^0 is the sign of E, so the two branches of Equation (92.1) carry densities of opposite sign, and no choice of normalization can make j^0 positive on both — while the positive-frequency solutions alone are not a complete set, so the negative branch cannot simply be discarded. A quantity that changes sign cannot be a probability density. It can be a charge density, and multiplying j^μ by the charge q makes it exactly that; the conserved integral is then a charge, which pair creation leaves unchanged because the pair is neutral. This is the reinterpretation proved in Section 92.6, and it is what the observations above record. □

92.2.3 The two problems are one problem

[Reserved: that j^0 *is negative exactly on the negative-frequency solutions, so a theory that could exclude one defect could exclude the other; Feshbach and Villars' explicit exhibition of this correspondence [1143]; the statement, to be proved in Section 92.6, that the resolution is to keep both branches and reinterpret* ej^μ *as an electric current [748], antiparticles being the negative-frequency modes read forwards in time [864, 931].]*

92.3 Minimal coupling and the charged scalar

92.3.1 The covariant derivative and gauge invariance

[Reserved: the substitution $\partial_\mu \mapsto D_\mu = \partial_\mu + \frac{\mathrm{i}q}{\hbar}A_\mu$ *introduced for this equation by Schrödinger [559] and by Gordon [544]; invariance of the coupled equation under* $\varphi \mapsto \mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\varphi$*,* $A_\mu \mapsto$

$A_\mu + \partial_\mu \Lambda$, the potentials being those of Chapter 60; a real field cannot carry charge, so the charged scalar is necessarily complex and its two real components are the particle and the antiparticle; the "seagull" term quadratic in A_μ, absent from the Dirac case, and the fact that gauge invariance forces it.]

92.3.2 The current as charge, not probability

[Reserved: the coupled current $j^\mu = \frac{i\hbar}{2m}\left(\varphi^* D^\mu \varphi - \varphi \left(D^\mu \varphi\right)^*\right)$; its total integral conserved and quantized in units of q; the reinterpretation of a sign-indefinite density as an honest charge density [748]; the corresponding Noether statement in Chapter 64, and the observable consequence that a spin-0 particle and its antiparticle are distinct unless the field is real, as they are for $\pi^\pm$ and are not for π^0 [2427].]

92.3.3 The Compton effect after Gordon

[Reserved: Gordon's original application, the Compton effect computed from the minimally coupled equation [544], set against the kinematic treatment of Chapter 69; what the calculation gets right (the kinematics, the shift $\Delta\lambda = (h/mc)(1-\cos\theta)$) and what it cannot get right (the cross section, which needs spin $\frac{1}{2}$ and the field quantization of Chapter 99); scalar Compton scattering as a genuine process for charged pions.]

Phenomenon 92.3 (The Compton shift)**.** Radiation scattered by a free charged particle of mass m returns with a wavelength longer than that of the incident beam by

$$\Delta\lambda = \frac{h}{mc}\left(1 - \cos\theta\right) , \tag{92.2}$$

θ being the scattering angle. The shift depends on neither the incident wavelength nor the intensity nor the spin of the scatterer, and vanishes in the forward direction. The constant h/mc is the Compton wavelength, equal to $2.426\,31 \times 10^{-12}$ m for the electron [2431]; the measurement is Chapter 69.

Derivation. The shift is pure four-momentum conservation and needs no dynamics — which is why the spinless equation gets it right while getting the cross section wrong. Let the incident photon have four-momentum k, the scattered photon k', and the target be at rest with $p = (mc, \mathbf{0})$. Conservation gives $p' = p + k - k'$, and squaring the invariant $p' \cdot p' = m^2c^2$ while using $k \cdot k = k' \cdot k' = 0$ leaves

$$p \cdot k - p \cdot k' - k \cdot k' = 0 .$$

With $k^0 = E/c$ and $k'^0 = E'/c$ this reads $mE - mE' = \left(EE'/c^2\right)\left(1 - \cos\theta\right)$, so that

$$\frac{1}{E'} - \frac{1}{E} = \frac{1 - \cos\theta}{mc^2} .$$

Substituting $E = hc/\lambda$ and $E' = hc/\lambda'$ turns the left side into $\left(\lambda' - \lambda\right)/hc$ and yields Equation (92.2). The mass enters only through the combination h/mc, so the effect is some 1836 times smaller on a proton than on an electron, and the derivation used nothing about the target beyond its mass — hence the independence from spin. □

92.3.4 Exact solutions in external fields

[Reserved: the Coulomb problem solved exactly, giving the spinless energy levels with their non-integer effective angular momentum and their $Z\alpha$ expansion; the failure against hydrogen and the success against pionic atoms [1035]; the constant magnetic field and the scalar Landau levels; the potential step, whose transmission anomaly is the spin-0 version of the Klein paradox treated in Chapter 93 [629].]

92.4 The Feshbach–Villars form and the nonrelativistic limit

92.4.1 The two-component representation

[Reserved: Feshbach and Villars' rewriting of the second-order equation as a coupled pair of first-order equations for $\varphi \pm \frac{\mathrm{i}\hbar}{mc^2}\partial_t\varphi$ [1143], in which the Hamiltonian is a 2×2 matrix built from Pauli matrices, the norm is the indefinite charge form of Section 92.2.2, and the two components are the particle and antiparticle amplitudes; the Hamiltonian is not Hermitian in the ordinary inner product but is pseudo-Hermitian with respect to that form.]

92.4.2 Recovery of the Schrödinger equation

[Reserved: the substitution $\varphi = \mathrm{e}^{-\mathrm{i}mc^2t/\hbar}\psi$ and the neglect of $\partial_t^2\psi$ against $mc^2\partial_t\psi/\hbar$, which turns the Klein–Gordon equation into the Schrödinger equation of Chapter 77 with the rest energy removed and the antiparticle component suppressed by $|\boldsymbol{p}|^2/m^2c^2$ [1143]; the systematic version of this limit for spin $\frac{1}{2}$ is the Foldy–Wouthuysen transformation of Chapter 93 [955].]

92.4.3 Relativistic corrections

[Reserved: the leading correction $-|\boldsymbol{p}|^4/8m^3c^2$ obtained by expanding the square root, the same term produced by perturbation theory in Chapter 82; the absence of any spin–orbit or Darwin term here, which is precisely why the spinless equation misses the hydrogen fine structure and why its failure is evidence for spin rather than against relativity (Chapter 81).]

92.5 Propagation, causality and the propagator

92.5.1 Green functions of the Klein–Gordon operator

[Reserved: the equation $(\Box + m^2c^2/\hbar^2)\,G(x-y) = -\delta^4(x-y)$ and the pole prescriptions that distinguish its retarded, advanced and Feynman solutions; the contour integral in the complex p^0 plane, using the machinery of Chapter 8; the retarded Green function supported inside the forward light cone, which is the sense in which the classical equation is causal, and its tail inside the cone — absent for $m=0$ — as a massive-wave phenomenon.]

92.5.2 The Feynman prescription

[Reserved: the propagator $\tilde{G}_F(p) = \mathrm{i}\hbar/\,(p^2 - m^2c^2 + \mathrm{i}\epsilon)$; the $\mathrm{i}\epsilon$ read as the instruction to propagate positive frequencies forwards and negative frequencies backwards in time [864, 931], which is the same reinterpretation as Section 92.2.3 in a different language; the propagator as the single object that carries this chapter's content into Chapters 99 and 100, where it becomes an internal line of a Feynman diagram [930].]

92.5.3 Localization and its limits

[Reserved: the Newton–Wigner position operator and the localized states it defines [942]; the amplitude for propagation outside the light cone, which is nonzero but falls as $\mathrm{e}^{-mcr/\hbar}$ over the Compton wavelength; Hegerfeldt's theorem that a state localized in a bounded region at one instant, in any theory with energy bounded below, immediately develops tails outside the light cone [1528]; why this is not a violation of relativistic causality, which constrains commutators of observables rather than amplitudes — the statement made precise in Chapter 109.]

92.5.4 The Yukawa potential

[Reserved: the static equation with a point source, whose solution $\mathrm{e}^{-mcr/\hbar}/4\pi r$ *is the Yukawa potential [765]; the range* $\hbar/mc$ *read off from the exponent,* 1.4 fm *for the observed pion mass, how Yukawa predicted a particle of a few hundred* MeV/c^2 *from the size of the nucleus; the Coulomb limit* $m \to 0$*; the potential's role in the nuclear force models of Chapter 107 and its status there as an effective, not fundamental, description.]*

Phenomenon 92.4 (A massive field mediates a force of finite range)**.** The static potential produced by a point source of a field of mass m is not Coulombic but exponentially cut off,

$$V(r) \propto \frac{\mathrm{e}^{-r/\lambda}}{r}, \qquad \lambda = \frac{\hbar}{mc}, \tag{92.3}$$

so the interaction it carries has a definite reach λ instead of extending indefinitely. The force between nucleons behaves this way, with a range of the order of a femtometre; that is the reduced Compton wavelength of a particle of mass $139.57\,\mathrm{MeV}/c^2$, which is the observed mass of the charged pion [2427]. Electromagnetism, by contrast, has no such cutoff, and the photon has no measured mass.

Derivation. For a time-independent field the Klein–Gordon operator reduces to $-\boldsymbol{\nabla}^2 + \lambda^{-2}$ with $\lambda = \hbar/mc$, so a static point source of strength g gives

$$\left(-\boldsymbol{\nabla}^2 + \frac{1}{\lambda^2}\right)\varphi = g\,\delta^3(\boldsymbol{x})\,.$$

Put $u(r) = \mathrm{e}^{-r/\lambda}/r$. Away from the origin the radial Laplacian is $\boldsymbol{\nabla}^2 u = r^{-1}\mathrm{d}^2(ru)/\mathrm{d}r^2 = r^{-1}\mathrm{d}^2\mathrm{e}^{-r/\lambda}/\mathrm{d}r^2 = u/\lambda^2$, so $\left(-\boldsymbol{\nabla}^2 + \lambda^{-2}\right)u = 0$ for $r > 0$: the mass term is exactly what the exponential decay cancels. Near the origin $u \to 1/r$, whose Laplacian is $-4\pi\delta^3(\boldsymbol{x})$ while the mass term stays integrable, so the source is reproduced by

$$\varphi(r) = \frac{g}{4\pi}\,\frac{\mathrm{e}^{-r/\lambda}}{r}\,,$$

which is Equation (92.3). Setting $m = 0$ removes the exponential and returns the Coulomb form, so the range is finite precisely because the mediating field is massive, and reading λ off a measured range predicts a mass. Note what the argument does *not* do: it gives the range, not the strength, and it treats the nucleon source as a fixed point charge, so the resulting potential is an effective description and not a fundamental one. □

92.6 Pauli–Weisskopf: the equation is a field

92.6.1 Quantization of the scalar field

[Reserved: Pauli and Weisskopf's 1934 construction [748], in which φ *is promoted to an operator expanded in creation and annihilation operators for particles and antiparticles, the negative-frequency coefficients being reinterpreted as creation operators; the Hamiltonian becomes positive definite, the charge operator takes both signs, and the number of particles is no longer conserved; the construction is carried out in full in Chapter 99. Their paper was written to show that Dirac's hole theory was not needed to make sense of negative energies — and it did so for a spin-0 field, where no exclusion principle exists to build a filled sea from.]*

92.6.2 Bose statistics and microcausality

[Reserved: commutators rather than anticommutators are forced for spin 0, so scalar quanta are bosons; the field commutator vanishing at spacelike separation as the precise statement

of relativistic causality [854]; the spin–statistics theorem, proved from Lorentz invariance, positivity of the energy and microcausality, whose axiomatic form belongs to Chapter 109 and whose observational consequences run through Chapters 83 and 120.]

92.6.3 What survives of the one-particle reading

[Reserved: an explicit accounting of which results of this chapter remain true in the field theory and which do not — the equation, the plane waves, the propagator, the Yukawa potential and the nonrelativistic limit all survive unchanged; the position operator, the probability density, the single-particle Hilbert space and the external Coulomb spectrum survive only as approximations valid for energies well below $2mc^2$, the threshold at which pair creation makes particle number ill-defined [748, 1273].]

92.7 The observed spin-0 sector

A relativistic wave equation earns its place by the particles that obey it. Nature supplies few elementary spin-0 particles and many composite ones, and both kinds are listed here with the measurements that fix their quantum numbers.

92.7.1 Pions and the light pseudoscalars

[Reserved: Yukawa's prediction of a meson of nuclear-force range [765]; the charged pion identified in photographic emulsions exposed to cosmic rays by Lattes, Occhialini and Powell [902], with the muon of Chapter 115 as its decay product and as the particle it had been confused with; masses $139.57\,\mathrm{MeV}/c^2$ and $134.98\,\mathrm{MeV}/c^2$ for $\pi^{\pm}$ and π^0, spin 0 and odd parity [2427]; the pions as pseudoscalars, not scalars, and as composite states of quarks rather than elementary fields (Chapter 102).]

Phenomenon 92.5 (The pion triplet)**.** Three particles of spin 0 and odd intrinsic parity, $J^P = 0^-$, are observed: a charged pair $\pi^{\pm}$ of mass $139.57\,\mathrm{MeV}/c^2$ and a neutral π^0 of mass $134.98\,\mathrm{MeV}/c^2$ [2427]. The two charged states are distinct particles, each the antiparticle of the other and of equal mass; the neutral state is its own antiparticle and decays to two photons, which no charged particle can do.

Derivation. The structural half of the statement follows from the current of Phenomenon 92.2 alone. If the field is *real*, $\varphi^* = \varphi$, then $j^\mu \propto \varphi^* \partial^\mu \varphi - \varphi\, \partial^\mu \varphi^*$ vanishes identically. A real scalar field carries no conserved charge at all, so its quantum has nothing by which it could be distinguished from its antiparticle: it is self-conjugate, and necessarily neutral. A charged spin-0 particle therefore requires a complex field, whose overall phase is the symmetry whose Noether charge is the conserved one; the quanta associated with φ and with φ^* are then two distinct states of opposite charge, and of equal mass, because both obey Equation (92.1) with the one mass parameter appearing in the equation. That is exactly the observed pattern: π^+ and π^- share a mass while π^0 has no charged partner. What this chapter does *not* derive is the numerical values of the masses or the odd parity; both are measurements, the parity of the charged pion having been fixed by the capture reaction discussed in Chapter 105, and the values themselves are consequences of the quark substructure of Chapter 102 rather than of any free-field equation. □

92.7.2 Pionic and kaonic atoms

[Reserved: the spin-0 realization of the Coulomb problem of Section 92.3.4 — a negative pion or kaon captured into a hydrogen-like orbit around a nucleus, whose X-ray cascade is described by the Klein–Gordon equation with no spin–orbit structure at all, so that the

equation is tested where it applies rather than where it fails [1413]; the strong-interaction shift and broadening of the lowest level, related to the scattering length by the Deser formula [1035];
with the electronic spectra of Chapter 90.]

92.7.3 The Higgs boson

[Reserved: the only elementary spin-0 particle observed to date, found at the LHC in 2012 by ATLAS [2186] and CMS [2192] with a mass near 125 GeV*, whose free field obeys exactly the equation of this chapter; its measured spin-parity $J^P = 0^+$, distinguishing it from the pseudoscalars above; the experiment is treated in Chapter 113 and the mechanism that requires the field in Chapter 104. Nothing in this chapter predicts the existence of such a particle; the equation says only what a spin-0 particle must satisfy if one exists.]*

Phenomenon 92.6 (An elementary scalar exists)**.** Exactly one elementary particle of spin 0 has been observed. Its mass is close to $125\,\mathrm{GeV}/c^2$; its spin and parity are measured to be $J^P = 0^+$, distinguishing it from the pseudoscalar mesons of Phenomenon 92.5; and its production and decay rates agree with those of the Standard Model Higgs boson within present uncertainties [2427]. Its free field obeys Equation (92.1), and hence the equation of this chapter, with no additional structure whatever — it is the only known particle for which that is the complete free-field description.

Derivation pending #113

Determination of the spin and parity of the boson near 125 GeV from the angular distributions of its four-lepton and diphoton decays, and the electroweak argument that requires an elementary scalar field in the first place. The experiment is the Higgs discovery chapter of Part XII and the mechanism is the electroweak chapter of the same part.

Chapter 93

The Dirac Equation

The Klein–Gordon equation of Chapter 92 fails as a one-particle theory because it is second order in time. Dirac asked in 1928 what a *first*-order relativistic wave equation would have to look like, and found that the coefficients multiplying the derivatives cannot be numbers: they must be matrices satisfying an anticommutation relation, and the smallest such matrices are 4×4 [596, 597]. Everything else follows from that one algebraic demand, and what follows is extraordinary. The wavefunction acquires four components; two of them describe an internal angular momentum of magnitude $\hbar/2$ with a gyromagnetic ratio $g = 2$, so that electron spin — inferred from spectra by Uhlenbeck and Goudsmit [532] and grafted onto the Schrödinger equation by hand by Pauli [587] — ceases to be an assumption and becomes a theorem. The hydrogen spectrum comes out exactly, fine structure included [595, 601], where the spinless equation had failed. And the other two components, carrying negative energy, forced Dirac to predict a particle of the electron's mass and opposite charge [647, 677] — found four years later in a cloud chamber (Chapter 95).

This chapter states the equation and its consequences as relativistic quantum *mechanics*, that is, for a single particle in an external field, which is the domain in which the results above are correct. It also marks the boundary of that domain honestly: hole theory needs an infinite filled sea and does not survive (Chapter 99); the Klein paradox [629] and the Zitterbewegung [669] are one-particle artefacts of a theory that cannot conserve particle number; and the exact Coulomb spectrum is falsified at the ninth significant figure by the Lamb shift [901] and the anomalous moment [916], both of which require the quantized field of Part XI. The group-theoretic reading — spin $\tfrac{1}{2}$ as the $(\tfrac{1}{2},0)\oplus(0,\tfrac{1}{2})$ representation of the Lorentz group — is developed in Chapter 94. Standard treatments are [1140, 1273, 1912, 2287].

Derivation pending #114

The Dirac Equation: all derivations of this chapter are pending.

93.1 Construction of the equation

93.1.1 Dirac's requirement: first order in time

[Reserved: the argument of Dirac's 1928 paper [596] — a probability density free of time derivatives requires an equation first order in ∂_t, and Lorentz covariance then requires it to be first order in ∂_i as well; the ansatz $\mathrm{i}\hbar\partial_t\psi = (c\,\boldsymbol{\alpha}\cdot\hat{\boldsymbol{p}} + \beta mc^2)\,\psi$, and the demand that each component satisfy the Klein–Gordon equation of Chapter 92, which fixes $\alpha_i\alpha_j + \alpha_j\alpha_i = 2\delta_{ij}$, $\alpha_i\beta + \beta\alpha_i = 0$ and $\beta^2 = \mathbb{1}$; why no numbers satisfy these relations and matrices of dimension below four do not either.]

93.1.2 The Clifford algebra of the gamma matrices

[Reserved: the covariant form $(\mathrm{i}\hbar\gamma^\mu\partial_\mu - mc)\,\psi = 0$ *and the defining relation* $\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2\eta^{\mu\nu}\mathbb{1}$*, an instance of the algebra introduced by Clifford [201] and developed as an algebraic structure in Chapter 4; the trace identities, the completeness of the sixteen products* $\{\mathbb{1}, \gamma^\mu, \sigma^{\mu\nu}, \gamma^\mu\gamma_5, \gamma_5\}$*, and* γ_5 *as the chirality element; the dimension count that forces* 4×4 *in* $3+1$ *dimensions.]*

93.1.3 Representations and the choice of basis

[Reserved: Pauli's theorem, that any two irreducible sets of gamma matrices are related by a similarity transformation unique up to a factor, so nothing physical depends on the choice; the Dirac (standard) basis in which β *is diagonal, adapted to the nonrelativistic limit of Section 93.7; the Weyl (chiral) basis in which* γ_5 *is diagonal, adapted to the massless limit of Section 93.8; the Majorana basis in which all gammas are purely imaginary and the equation admits real solutions [796].]*

93.1.4 The conserved positive-definite current

[Reserved: the adjoint $\bar\psi = \psi^\dagger\gamma^0$*, the current* $j^\mu = c\,\bar\psi\gamma^\mu\psi$*, its conservation* $\partial_\mu j^\mu = 0$*, and the density* $j^0/c = \psi^\dagger\psi \geq 0$ *— the property the Klein–Gordon current lacked (Chapter 92) and the one Dirac was hunting [596]; the honest qualification that positivity of the density does* not *rescue the one-particle interpretation, because the energy spectrum is still unbounded below (Section 93.5).]*

93.2 Lorentz covariance and the spinor

93.2.1 The spinor transformation law

[Reserved: covariance requires $\psi'(x') = S(\Lambda)\psi(x)$ *with* $S^{-1}\gamma^\mu S = \Lambda^\mu{}_\nu\gamma^\nu$*; the explicit generators* $\sigma^{\mu\nu} = \frac{\mathrm{i}}{2}[\gamma^\mu, \gamma^\nu]$ *and the finite transformation* $S = \exp\left(-\frac{\mathrm{i}}{4}\omega_{\mu\nu}\sigma^{\mu\nu}\right)$*; that* S *is unitary for rotations and Hermitian for boosts, so the representation is not unitary — the group being non-compact (Chapter 94); the* 4π *periodicity of a spinor under rotation and its experimental confirmation by neutron interferometry [1566], which is why the covering group* $\mathrm{SL}(2,\mathbb{C})$ *and not the Lorentz group itself acts.]*

Phenomenon 93.1 (A spinor changes sign under a full rotation)**.** Rotating a spin-$\frac{1}{2}$ particle through 2π multiplies its state by -1; only a rotation through 4π returns it to itself. The sign is not a convention that could be redefined away: in a neutron interferometer, one arm of which passes through a magnetic field that precesses the spin, the interference intensity varies with the precession angle at a period of 4π and not 2π, so the amplitude demonstrably reverses at a single full turn [1566].

Derivation. In the spinor representation of Section 93.2.1 the generator of rotations about the z-axis is $\frac{1}{2}\Sigma_3$, with $\Sigma_k = \mathrm{diag}\,(\sigma_k, \sigma_k)$ built from the Pauli matrices, so a rotation through an angle θ acts on the spinor as

$$S(\theta) = \exp\left(-\frac{\mathrm{i}}{2}\theta\Sigma_3\right) = \mathbb{1}\cos\frac{\theta}{2} - \mathrm{i}\Sigma_3\sin\frac{\theta}{2}\,, \tag{93.1}$$

the exponential series collapsing because $\Sigma_3^2 = \mathbb{1}$. At $\theta = 2\pi$ this is $-\mathbb{1}$ and at $\theta = 4\pi$ it is $+\mathbb{1}$. The half-angle is the whole content: $\mathrm{SO}(3)$ is not simply connected, its closed loops falling into two classes, and the spinor representation is a representation not of $\mathrm{SO}(3)$ but of its double cover $\mathrm{SU}(2)$, in which the lift of a single full turn is an open path ending at $-\mathbb{1}$. An overall sign is unobservable on a state taken alone, which is why the effect can be seen only by interfering a rotated amplitude against an unrotated one. □

93.2.2 Bilinear covariants

[Reserved: the sixteen bilinears $\bar{\psi}\Gamma\psi$ and their transformation types — scalar, vector, antisymmetric tensor, axial vector and pseudoscalar; their role as the only Lorentz-covariant objects a spinor field can build, hence as the complete list of possible interaction terms, which is what makes the $V-A$ structure of the weak interaction a choice among five *rather than a free invention (Chapters 101 and 110); the Fierz rearrangement identities relating different pairings [790].]*

93.2.3 Plane-wave solutions and spin projectors

[Reserved: the four independent plane waves at each momentum, two with $p^0>0$ and two with $p^0<0$; the spinors $u(p,s)$ and $v(p,s)$, their normalization and orthogonality, the energy projectors $(\pm\gamma^\mu p_\mu + mc)/2mc$ and the spin projector from the Pauli–Lubanski vector of Chapter 94; the completeness relations that make every later trace calculation mechanical.]

93.3 Spin as a consequence, not an assumption

This section is the chapter's principal result: an equation written to be first order in time turns out to describe a particle with intrinsic angular momentum $\hbar/2$ and the right magnetic moment, neither of which was put in.

93.3.1 Orbital angular momentum is not conserved

[Reserved: the computation $\left[\hat{\boldsymbol{L}},\hat{H}\right]\neq 0$ for the Dirac Hamiltonian, and the identification of the operator $\hat{\boldsymbol{S}}=\frac{\hbar}{2}\boldsymbol{\Sigma}$ whose sum with $\hat{\boldsymbol{L}}$ is conserved [596]; the eigenvalues $\pm\hbar/2$, matching the two beams of the Stern–Gerlach experiment (Chapter 80) and the algebra of Chapter 79; the contrast with the spinless equation, where $\hat{\boldsymbol{L}}$ alone is conserved.]

Phenomenon 93.2 (Electron spin)**.** A beam of neutral silver atoms, whose single valence electron is in an S state carrying no orbital angular momentum, splits into exactly two components on traversing an inhomogeneous magnetic field — not into the odd number of components that any integer angular momentum would give, and not into the continuum a classical moment would give [481, 491, 492]. The electron therefore carries an intrinsic angular momentum with just two projections, $\pm\hbar/2$, on any axis; the measurement is Chapter 80. In any central field it is the total $\hat{\boldsymbol{J}}=\hat{\boldsymbol{L}}+\hat{\boldsymbol{S}}$, and not $\hat{\boldsymbol{L}}$, that is found to be conserved.

Derivation. Both halves are theorems about the Hamiltonian $\hat{H}=c\,\boldsymbol{\alpha}\cdot\hat{\boldsymbol{p}}+\beta m_e c^2$ of Section 93.1.1. Using $\left[\hat{L}_i,\hat{p}_j\right]=\mathrm{i}\hbar\epsilon_{ijk}\hat{p}_k$, and noting that β and $\boldsymbol{\alpha}$ are constant matrices that commute with $\hat{\boldsymbol{L}}$,

$$\left[\hat{L}_i,\hat{H}\right]=c\,\alpha_j\left[\hat{L}_i,\hat{p}_j\right]=\mathrm{i}\hbar c\,\epsilon_{ijk}\alpha_j\hat{p}_k=\mathrm{i}\hbar c\,(\boldsymbol{\alpha}\times\hat{\boldsymbol{p}})_i\neq 0\,,$$

so orbital angular momentum is not conserved even for a *free* particle — an absurdity unless something else carries the balance. Set $\hat{\boldsymbol{S}}=\frac{\hbar}{2}\boldsymbol{\Sigma}$ with $\Sigma_k=\mathrm{diag}\,(\sigma_k,\sigma_k)$. Since $[\Sigma_i,\alpha_j]=2\mathrm{i}\epsilon_{ijk}\alpha_k$ and $[\Sigma_i,\beta]=0$,

$$\left[\hat{S}_i,\hat{H}\right]=\frac{\hbar c}{2}\,[\Sigma_i,\alpha_j]\,\hat{p}_j=\mathrm{i}\hbar c\,\epsilon_{ijk}\alpha_k\hat{p}_j=-\mathrm{i}\hbar c\,(\boldsymbol{\alpha}\times\hat{\boldsymbol{p}})_i\,,$$

which cancels the previous line exactly, so $\left[\hat{\boldsymbol{L}}+\hat{\boldsymbol{S}},\hat{H}\right]=0$. The operator $\hat{S}_z$ has eigenvalues $\pm\hbar/2$ and no others, because $\Sigma_3^2=\mathbb{1}$. Nothing about spin was assumed anywhere: it was

forced by the demand that the equation be first order in time, which is what made the coefficients matrices in the first place, and the two Stern–Gerlach beams are the observed count of its eigenvalues. □

93.3.2 The gyromagnetic ratio $g = 2$

[Reserved: minimal coupling $\partial_\mu \mapsto \partial_\mu + \frac{\mathrm{i}e}{\hbar}A_\mu$, squaring the Dirac operator, and the appearance of the term $-\frac{e\hbar}{2mc}\,\sigma^{\mu\nu}F_{\mu\nu}$ that carries a magnetic moment $g = 2$ in units of the Bohr magneton [596] — twice the orbital value, exactly the factor Uhlenbeck and Goudsmit had to postulate [532] and Pauli had to insert [587]; the measured departure $g - 2 \approx 2.3 \times 10^{-3}$ [916] and its computation from field theory [919], deferred to Chapter 112.]

Phenomenon 93.3 (The electron magnetic moment)**.** The magnetic moment associated with the electron's spin is twice as large, per unit of angular momentum, as the moment associated with its orbital motion:

$$\boldsymbol{\mu} = -g\,\frac{\mu_{\mathrm{B}}}{\hbar}\hat{\boldsymbol{S}}\,, \qquad \mu_{\mathrm{B}} = \frac{e\hbar}{2m_e}\,, \qquad g \simeq 2\,, \tag{93.2}$$

where an orbital current of the same angular momentum would give $g = 1$. This factor is what the Zeeman splittings of the alkali doublets demand, and it had to be inserted by hand into the prerelativistic theory. Its measured magnitude is 2.002 319 304 36 [2431] — equal to 2 to eleven significant figures, the residual departure of about one part in 10^3 being the anomaly of Chapter 112.

Derivation. Couple the electron minimally to an external field, writing $\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{p}} - q\boldsymbol{A}$ with $q = -e$, and use $(\boldsymbol{\sigma}\cdot\boldsymbol{a})(\boldsymbol{\sigma}\cdot\boldsymbol{b}) = \boldsymbol{a}\cdot\boldsymbol{b} + \mathrm{i}\boldsymbol{\sigma}\cdot(\boldsymbol{a}\times\boldsymbol{b})$ with $\boldsymbol{a} = \boldsymbol{b} = \hat{\boldsymbol{\pi}}$. The cross term does not vanish, because the components of $\hat{\boldsymbol{\pi}}$ fail to commute:

$$[\hat{\pi}_j, \hat{\pi}_k] = -q\,([\hat{p}_j, A_k] + [A_j, \hat{p}_k]) = \mathrm{i}\hbar q\,(\partial_j A_k - \partial_k A_j) = \mathrm{i}\hbar q\,\epsilon_{jkl}B_l\,,$$

so that $(\hat{\boldsymbol{\pi}}\times\hat{\boldsymbol{\pi}})_i = \frac{1}{2}\epsilon_{ijk}\,[\hat{\pi}_j, \hat{\pi}_k] = \mathrm{i}\hbar q B_i$ and hence

$$(\boldsymbol{\sigma}\cdot\hat{\boldsymbol{\pi}})^2 = \hat{\boldsymbol{\pi}}^2 - q\hbar\,\boldsymbol{\sigma}\cdot\boldsymbol{B}\,. \tag{93.3}$$

Eliminating the lower components of the spinor at leading order in $1/m_e c^2$ (Section 93.3.3) leaves the upper two-component amplitude obeying $\mathrm{i}\hbar\partial_t\chi = \left[(\boldsymbol{\sigma}\cdot\hat{\boldsymbol{\pi}})^2/2m_e + q\phi\right]\chi$, and inserting Equation (93.3) produces the interaction $-(q\hbar/2m_e)\,\boldsymbol{\sigma}\cdot\boldsymbol{B} = -\boldsymbol{\mu}\cdot\boldsymbol{B}$ with

$$\boldsymbol{\mu} = \frac{q\hbar}{2m_e}\boldsymbol{\sigma} = -\frac{e}{m_e}\hat{\boldsymbol{S}} = -2\,\frac{\mu_{\mathrm{B}}}{\hbar}\hat{\boldsymbol{S}}\,,$$

using $q = -e$ and $\hat{\boldsymbol{S}} = \frac{\hbar}{2}\boldsymbol{\sigma}$: this is Equation (93.2) with $g = 2$. The same substitution applied to the orbital term gives $-(q/2m_e)\,\hat{\boldsymbol{L}}\cdot\boldsymbol{B}$, that is $g = 1$ for orbital motion. The factor of two is produced entirely by the algebra of $(\boldsymbol{\sigma}\cdot\boldsymbol{a})^2$; it is not fitted, and its *exactness* at this order is what makes the measured departure from 2 informative rather than merely a correction. □

93.3.3 The Pauli equation as the nonrelativistic limit

[Reserved: elimination of the small components at leading order, yielding the Pauli equation $\mathrm{i}\hbar\partial_t\chi = \left[\frac{1}{2m}\left(\hat{\boldsymbol{p}} - \frac{e}{c}\boldsymbol{A}\right)^2 - \frac{e\hbar}{2mc}\boldsymbol{\sigma}\cdot\boldsymbol{B} + e\phi\right]\chi$ [587], previously an empirical two-component equation and now a derived limit; the identification of the Pauli matrices with $\boldsymbol{\Sigma}$ restricted to the upper components; the systematic continuation of this expansion in Section 93.7.1.]

93.4 The hydrogen atom solved exactly

93.4.1 The Dirac–Coulomb spectrum

[Reserved: separation of the equation in a Coulomb field and the closed-form spectrum obtained independently by Darwin [595] and by Gordon [601], depending on n and on the total angular momentum j only; its expansion in powers of $Z\alpha$, reproducing the Bohr levels of Chapter 70 at leading order and Sommerfeld's fine-structure formula [441] at the next — obtained by Sommerfeld from a relativistic orbit calculation with no spin at all, one of the most instructive coincidences in physics, which the chapter should state and explain rather than repeat; the breakdown at $Z\alpha > 1$ and what it signals.]

Phenomenon 93.4 (Fine structure)**.** The spectral lines of atomic hydrogen are not single. Each level of principal quantum number n is resolved into components separated by intervals of relative order $\alpha^2 \approx 5.3\times10^{-5}$, and the pattern is governed by the total angular momentum j alone: two states sharing n and j but differing in orbital ℓ — $2s_{1/2}$ and $2p_{1/2}$, for instance — are observed to coincide. The splittings were fitted, in advance of any notion of spin, by Sommerfeld's relativistic orbit formula [441], and the Dirac equation reproduces the same expression in closed form,

$$E_{nj} = m_e c^2 \left[1 + \frac{(Z\alpha)^2}{\left(n - \left(j + \tfrac{1}{2} \right) + \sqrt{\left(j + \tfrac{1}{2} \right)^2 - (Z\alpha)^2} \right)^2} \right]^{-1/2} , \tag{93.4}$$

now as a consequence of the spin it does not assume.

Derivation pending #115

Separation of the Dirac equation in a Coulomb potential and solution of the resulting radial system in closed form, together with its expansion in powers of the nuclear charge times the fine-structure constant, which must reproduce the Bohr levels at leading order and the fine-structure intervals at the next.

93.4.2 Fine structure: spin–orbit, Darwin and mass terms

[Reserved: the three order-α^4 terms separated by the Foldy–Wouthuysen expansion — the relativistic kinetic correction, the spin–orbit coupling with the factor $\tfrac{1}{2}$ that Thomas supplied kinematically [561], and the Darwin term [595] that acts only on s states and reflects the smearing of the electron over a Compton wavelength; the resulting degeneracy of states with equal j, e.g. $2s_{1/2}$ and $2p_{1/2}$, which is a prediction *of the exact spectrum and is the one the next subsection reports as false; comparison with the perturbation-theory route of Chapter 82.]*

93.4.3 Where the exact spectrum fails: the Lamb shift

[Reserved: Lamb and Retherford's microwave measurement of a splitting of about 1058 MHz between $2s_{1/2}$ and $2p_{1/2}$ [901], states the Dirac equation makes exactly degenerate; the shift as the first quantitative evidence that the electron interacts with the quantized electromagnetic field, and therefore as the experiment that ended relativistic quantum mechanics as a fundamental theory; its calculation belongs to Chapter 100 and the modern spectroscopy to Chapter 90. This subsection must state the falsification plainly: the chapter's best result is wrong at the level current experiments resolve easily.]

Phenomenon 93.5 (The Dirac–Coulomb degeneracy is broken)**.** The states $2s_{1/2}$ and $2p_{1/2}$ of hydrogen, which Equation (93.4) makes exactly degenerate because they share both n and j, are observed to be split, the s state lying higher [901, 2331]. The interval is driven directly as a microwave resonance between the two levels, so it is measured rather than inferred from a difference of optical frequencies, and it exceeds every uncertainty of that measurement by orders of magnitude. The disagreement with Phenomenon 93.4 is therefore not marginal: the exact spectrum of this chapter is false. The measurements are Chapter 90.

Derivation pending #116

The radiative level shift between the two lowest hydrogen states of equal total angular momentum. Deriving it requires the quantized electromagnetic field, the electron self-energy and the vacuum polarization, and a renormalization prescription, so it lies outside relativistic quantum mechanics altogether and belongs to the quantum-electrodynamics chapter of Part XII.

93.5 Negative energies, holes and the positron

93.5.1 Hole theory

[Reserved: the negative-energy solutions of Section 93.2.3 and the catastrophe they imply — an electron radiating indefinitely into ever more negative states; Dirac's proposal that all such states are filled and that the exclusion principle of Chapter 83 forbids the transition [647]; a hole in the sea behaving as a particle of positive energy and opposite charge; the honest catalogue of what is wrong with the picture — an infinite charge density, no analogue for bosons (where Chapter 92 needed none [748]), and no Lorentz-invariant meaning for "filled".]

93.5.2 The prediction of the anti-electron

[Reserved: Dirac's first identification of the hole with the proton [647] and its refutation — Weyl's symmetry argument and Oppenheimer's observation that the hydrogen atom would then annihilate; the retraction and the correct prediction of an "anti-electron", a particle of the same mass and opposite charge, in 1931 [677]; the prediction of the antiproton in the same sentence, confirmed only in 1955; the timeline against the observation of Chapter 95, which this chapter should give exactly, because the case is the treatise's cleanest example of a mathematically forced prediction confirmed afterwards.]

Phenomenon 93.6 (Antiparticles)**.** There exists a particle of the electron's mass and of the opposite charge, the positron; it was first seen in cosmic radiation (Chapter 95), and the same has since proved true of every charged fermion examined. The masses and the charge magnitudes of a particle and of its antiparticle agree to within the precision of the comparisons tabulated in [2427], and no charged particle without such a partner has ever been found.

Derivation. The pairing is a symmetry of the equation itself, independent of hole theory. Write the minimally coupled equation of Section 93.1.2 for charge q,

$$[\gamma^{\mu}\,(\mathrm{i}\hbar\partial_{\mu} - qA_{\mu}) - mc]\,\psi = 0\,,$$

and take its complex conjugate, A_{μ} being real:

$$[\gamma^{\mu *}\,(-\mathrm{i}\hbar\partial_{\mu} - qA_{\mu}) - mc]\,\psi^{*} = 0\,.$$

In every representation of the Clifford algebra there is an invertible matrix C with $C\gamma^{\mu *}C^{-1} = -\gamma^{\mu}$; in the Dirac basis one may take $C = \mathrm{i}\gamma^2$, since there $\gamma^0, \gamma^1, \gamma^3$ are real and γ^2 is purely imaginary. Multiplying the conjugated equation on the left by C and inserting $C^{-1}C$ before ψ^* turns it into

$$[\gamma^{\mu}\,(\mathrm{i}\hbar\partial_{\mu} + qA_{\mu}) - mc]\,\psi^{c} = 0\,, \qquad \psi^{c} := C\psi^{*}\,, \tag{93.5}$$

which is the original equation with q replaced by $-q$ and m untouched. Solutions therefore come in pairs of equal mass and opposite charge, and since the map is invertible neither member of a pair is the more fundamental. An equation that describes an electron cannot avoid also describing a positron — which is why the prediction was forced rather than proposed, and why its confirmation four years later is the strongest single argument of this chapter. The argument says nothing about *why* the electron's mass takes the value it does, and nothing about the cosmological asymmetry between matter and antimatter, which Equation (93.5) does not address. □

93.5.3 The Stueckelberg–Feynman reading

[Reserved: the reinterpretation of negative-energy solutions as antiparticles propagating forwards in time, due to Stueckelberg [864] and Feynman [931]; the propagator prescription and the diagrammatic rules it generates [930]; why this reading and not hole theory is the one that survives into Chapter 99, and why it applies equally to the spin-0 case where no sea can be built.]

93.5.4 Charge conjugation and the Majorana alternative

[Reserved: the operation $\psi \mapsto \psi^{c} = C\bar{\psi}^{\mathsf{T}}$ mapping solutions of charge e to solutions of charge $-e$, and the CPT statement that its existence is not optional (Chapter 105); Majorana's observation that a neutral spin-$\tfrac{1}{2}$ particle may satisfy $\psi^{c} = \psi$ and so be its own antiparticle [796]; that this is a live experimental question only for neutrinos (Chapters 96 and 103), and that neutrinoless double beta decay has not been observed [2427] — so the alternative is stated as open, not adopted.]

93.6 Phenomena of the one-particle theory

93.6.1 Zitterbewegung

[Reserved: Schrödinger's 1930 analysis of free motion [669], in which the velocity operator $c\boldsymbol{\alpha}$ has eigenvalues $\pm c$ and the position expectation oscillates at angular frequency $2mc^2/\hbar$ with amplitude of order the Compton wavelength; the origin of the effect in interference between positive- and negative-energy components, so that it is absent for a state built from one branch alone; the verdict that it is an artefact of the one-particle reading rather than an observable motion of an electron, and the simulations in trapped-ion and cold-atom systems that reproduce the mathematics without vindicating the physics.]

93.6.2 The Klein paradox

[Reserved: Klein's calculation of reflection at a step of height exceeding $2mc^2$, where the transmitted current does not vanish and the reflected current exceeds the incident one [629]; the resolution as spontaneous pair creation by a field strong enough to supply $2mc^2$ over a Compton wavelength, hence as another statement that particle number is not conserved; the Schwinger critical field $m^2c^3/e\hbar \approx 1.3 \times 10^{18}\,\mathrm{V/m}$ as the scale [988], far above anything achieved in the laboratory, so this remains a prediction and must be labelled as one; the spin-0 version noted in Chapter 92.]

93.6.3 Thomas precession and spin transport

[Reserved: the kinematic precession of a spin carried along an accelerated worldline, derived by Thomas [561] from special relativity alone (Chapter 40) and responsible for the factor $\tfrac{1}{2}$ in the spin–orbit term of Section 93.4.2; the covariant equation of spin transport in an external field of Bargmann, Michel and Telegdi [1165], which is what the storage-ring measurements of Chapter 112 are actually analysed with.]

93.7 The systematic nonrelativistic expansion

93.7.1 The Foldy–Wouthuysen transformation

[Reserved: the unitary transformation that decouples upper and lower components order by order in $1/mc^2$, constructed by Foldy and Wouthuysen [955]; the transformed Hamiltonian and the mean position operator, which coincides with the Newton–Wigner operator of Chapter 92 [942] and in terms of which the Zitterbewegung disappears; the free case, where the transformation is exact, against the external-field case, where it is an asymptotic series.]

93.7.2 The corrections order by order

[Reserved: the expansion producing, at order $1/m^2c^2$, exactly the three fine-structure terms of Section 93.4.2 — kinetic, spin–orbit and Darwin — with their coefficients fixed rather than fitted; the method as the standard route to effective nonrelativistic Hamiltonians for atomic and molecular structure (Chapter 88) and for the relativistic corrections in heavy atoms, where the expansion parameter $Z\alpha$ is no longer small.]

93.8 The massless limit and the boundary of the theory

93.8.1 Weyl spinors and chirality

[Reserved: at $m = 0$ the equation decouples into two independent two-component equations for the chirality eigenstates, written by Weyl [641] and long rejected because each violates parity; their adoption after parity violation was observed (Chapter 110) and the neutrino helicity was measured (Chapter 98); the Dirac field as the direct sum of two Weyl fields coupled by the mass term, which is the statement developed in Chapter 96.]

93.8.2 Where relativistic quantum mechanics ends

[Reserved: a closing accounting of the chapter's results by status — the equation, the gamma algebra, the spinor transformation law, $g = 2$ at tree level and the Coulomb spectrum to order α^4 survive intact into field theory; hole theory, the single-particle Hilbert space, the position operator and the exact Coulomb spectrum do not [748, 931]. The energy scale at which the one-particle reading fails is $2mc^2$, and every phenomenon in Section 93.6 is a symptom of that failure. The continuation is Chapter 99.]

Chapter 94

Particles as Poincaré Representations

What *is* a particle? Every earlier part of this treatise uses the word operationally — a silver atom in a Stern–Gerlach magnet (Chapter 80), a photoelectron, a positron track — but relativistic quantum mechanics is where the word acquires a definition. The answer, found by Wigner in 1939 [849], is that a particle species is an *irreducible unitary representation of the Poincaré group*: relativity supplies the symmetry group of spacetime (Section 39.3), quantum mechanics demands that symmetries act unitarily on rays (Postulate 77.1), and irreducibility is the statement that the system has no smaller relativistically invariant part — that it is *elementary*. The two labels that classify these representations, mass and spin, are exactly the two numbers experiment attaches to every particle ever observed. This chapter carries out Wigner's programme in full; the standard modern accounts are [1912, 2287].

Throughout, the metric signature is $(+,-,-,-)$ with $\eta = \mathrm{diag}(1,-1,-1,-1)$ as in Chapter 39, Greek indices run over $0,1,2,3$ and Latin ones over $1,2,3$, the totally antisymmetric symbol is fixed by $\epsilon_{0123} = +1$, and $\hbar$ and c are kept explicit (Section 77.3).

Remark 94.1 (Level of rigor). The generators used below are unbounded operators, defined on dense domains by Stone's theorem for the one-parameter unitary subgroups they generate; commutators are understood on the common invariant dense domain of smooth vectors of the representation. This chapter manipulates them algebraically, as the physics literature does [1912]; the functional-analytic scaffolding (domains, essential self-adjointness, integrability of the algebra to the group) is classical representation theory and is not restated here. Two standing assumptions are recorded once: every representation in this chapter is *strongly continuous* — $(\Lambda, a) \mapsto U(\Lambda, a)\,|\psi\rangle$ is continuous for each state, which is what Stone's theorem consumes — and Schur's lemma, cited from the finite-dimensional Theorem 5.153, is used in its unitary infinite-dimensional (spectral-projection) form [1912].

94.1 Symmetries act unitarily: Wigner's theorem

Quantum states are unit rays, not vectors: $|\psi\rangle$ and $\mathrm{e}^{\mathrm{i}\alpha}\,|\psi\rangle$ describe the same physical state (Postulate 77.1), and every measurable prediction is built from transition probabilities $|\langle\varphi|\psi\rangle|^2$ (Postulate 77.4). A symmetry operation must therefore be, in the first instance, a map of *rays* that preserves those probabilities.

Definition 94.2 (Symmetry transformation)**.** Let $\mathcal{H}$ be the Hilbert space of a quantum system and write $[\psi]$ for the unit ray of a normalized $|\psi\rangle$. A *symmetry transformation* is a bijection T of the set of unit rays such that

$$|\langle\varphi'|\psi'\rangle|^2 = |\langle\varphi|\psi\rangle|^2 \qquad \text{whenever } [\psi'] = T[\psi],\ [\varphi'] = T[\varphi]\,. \tag{94.1}$$

Theorem 94.3 (Wigner). *Let* $\dim \mathcal{H} \geq 2$. *Every symmetry transformation* T *is implemented by an operator* U *on* $\mathcal{H}$ *—* $T[\psi] = [U\psi]$ *for every unit ray — which is either unitary and linear, or antiunitary and antilinear.* U *is unique up to an overall phase, and the two cases are mutually exclusive.*

Full derivation: Appendix A.9 (p. 2797).

Derives Theorem 94.3.

The theorem is due to Wigner [687]; the complete proof given in Appendix A.9 follows Bargmann [1269].

Corollary 94.4 (Connected symmetry groups act unitarily). *Let* G *be a connected Lie group acting by symmetry transformations* T_g*, and let* $U(g)$ *implement* T_g *as in Theorem 94.3. Then every* $U(g)$ *is unitary.*

Proof. Ray composition gives $U(g)U(h) = \omega(g,h)\,U(gh)$ for a phase $\omega(g,h)$, since both sides implement T_{gh} and Theorem 94.3 fixes the implementation up to a phase. In a connected Lie group every element is a finite product of exponentials, $g = \exp X_1 \cdots \exp X_n$: the image of the exponential map contains a neighbourhood of the identity, and the subgroup generated by any neighbourhood of the identity is open, hence closed — its complement is a union of its own open cosets — hence all of the connected group. Each factor is a square: $\exp X = \left(\exp \frac{X}{2}\right)^2$. Hence $U(g)$ equals, up to a phase, a product of operators $U(h_i)^2$. The square of a unitary operator is unitary, and so is the square of an antiunitary one (two antilinear maps compose to a linear map, and isometry is preserved); phases do not affect (anti)unitarity. A product of unitaries is unitary. □

The phases $\omega(g,h)$ cannot in general be removed by rephasing the $U(g)$: one speaks of a *projective* unitary representation. For the Poincaré group the obstruction was computed by Bargmann [1029]:

Theorem 94.5 (Bargmann). *Every* continuous *projective unitary representation of the proper orthochronous Poincaré group arises, after a choice of phases, from an ordinary strongly continuous unitary representation of its universal covering group* $\mathbb{R}^4 \rtimes \mathrm{SL}(2,\mathbb{C})$.

Derivation pending #117

Bargmann's theorem for the Poincaré group: the proof computes the second cohomology of the Lie algebra $\mathfrak{iso}(3,1)$ *and shows every local phase cocycle is a coboundary, then lifts to the simply connected cover; see [1029]. The cohomological machinery is not yet developed in this treatise.*

Remark 94.6 (The cover is physical). Theorem 94.5 is why the classification below runs over representations of the *cover* — in which a rotation by 2π need not be the identity, only its square is — and this is not a mathematical nicety but an observed fact: a neutron interferometer splits a beam, rotates the spin of one arm through 2π with a magnetic field, and finds the recombined intensity *reversed*, returning to itself only after 4π, exactly as the sign $U(2\pi) = -\mathbb{1}$ of a spin-$\frac{1}{2}$ representation demands [1566]. By contrast, for the Galilei group of Part III the obstruction does *not* vanish: its central charge is the mass itself [1029], which is why nonrelativistic quantum mechanics superselects mass while relativistic quantum mechanics treats it as a spectrum label.

94.2 The unitary generators of the Poincaré group

Proposition 94.7 (Composition law). *The transformations* $x \mapsto \Lambda x + a$ *of Minkowski space, with* Λ *in the proper orthochronous Lorentz group* $L_+^\uparrow$ *(Definition 39.5; Proposition 39.6*

makes it a group) and $a \in \mathbb{R}^4$, form a group with

$$(\Lambda', a')\,(\Lambda, a) = (\Lambda'\Lambda,\ a' + \Lambda' a)\,. \tag{94.2}$$

Proof. $x \mapsto \Lambda x + a \mapsto \Lambda'(\Lambda x + a) + a' = (\Lambda'\Lambda)x + (\Lambda' a + a')$; the identity is $(\mathbb{1}, 0)$ and $(\Lambda, a)^{-1} = (\Lambda^{-1}, -\Lambda^{-1}a)$. □

This is the Poincaré group of Section 39.3, restricted to its identity component. By Corollary 94.4 a relativistic quantum system carries unitary operators $U(\Lambda, a)$, projectively composing by Equation (94.2); by Theorem 94.5 we may and do take a true representation of the cover. Stone's theorem then provides Hermitian generators: for translations and for Lorentz transformations respectively,

$$U(\mathbb{1}, a) = \exp\left(-\frac{\mathrm{i}}{\hbar}\, a^\mu \hat{P}_\mu\right), \qquad U(\Lambda(\omega), 0) = \exp\left(-\frac{\mathrm{i}}{2\hbar}\, \omega_{\mu\nu} \hat{J}^{\mu\nu}\right), \tag{94.3}$$

with $\omega_{\mu\nu} = -\omega_{\nu\mu}$ the parameters of Section 39.2.7 and $\hat{J}^{\mu\nu} = -\hat{J}^{\nu\mu}$. The factor $\hbar$ makes $\hat{P}_\mu$ a momentum and $\hat{J}^{\mu\nu}$ an angular momentum; the physical dictionary is

$$\hat{H} = c\,\hat{P}^0\,, \qquad \hat{J}_i = \tfrac{1}{2}\,\epsilon_{ijk}\hat{J}^{jk}\,, \qquad \hat{K}_i = \hat{J}^{i0}\,, \tag{94.4}$$

energy, angular momentum and boost generator (cf. the nonrelativistic operators of Section 77.5).

Proposition 94.8 (The Poincaré algebra of the generators)**.** *On the common dense domain of Remark 94.1,*

$$\left[\hat{J}^{\mu\nu}, \hat{J}^{\lambda\rho}\right] = \mathrm{i}\hbar\left(\eta^{\mu\lambda}\hat{J}^{\rho\nu} + \eta^{\nu\rho}\hat{J}^{\lambda\mu} + \eta^{\mu\rho}\hat{J}^{\nu\lambda} + \eta^{\nu\lambda}\hat{J}^{\mu\rho}\right), \tag{94.5}$$

$$\left[\hat{J}^{\mu\nu}, \hat{P}^{\lambda}\right] = \mathrm{i}\hbar\left(\eta^{\nu\lambda}\hat{P}^{\mu} - \eta^{\mu\lambda}\hat{P}^{\nu}\right), \tag{94.6}$$

$$\left[\hat{P}^{\mu}, \hat{P}^{\nu}\right] = 0\,. \tag{94.7}$$

Proof. Let $\{X_a\}$ be a basis of the abstract Lie algebra with real structure constants, $[X_a, X_b] = f_{ab}{}^c X_c$, and let $\mathrm{d}U$ denote the induced representation of the algebra, so that $U(\exp tX_a) = \exp\left(-\frac{\mathrm{i}}{\hbar}t\,T_a\right)$ defines the Hermitian generator $T_a = \mathrm{i}\hbar\,\mathrm{d}U(X_a)$. Since $\mathrm{d}U$ preserves brackets,

$$[T_a, T_b] = (\mathrm{i}\hbar)^2\,[\mathrm{d}U(X_a), \mathrm{d}U(X_b)] = (\mathrm{i}\hbar)^2 f_{ab}{}^c\,\mathrm{d}U(X_c) = \mathrm{i}\hbar\, f_{ab}{}^c\, T_c\,.$$

The structure constants of $\mathfrak{iso}(3,1)$ were derived in Section 14.3.2 from the realization $J_{AB} = x_A\partial_B - x_B\partial_A$, $P_C = \partial_C$ (Equation (14.86)): the brackets are Equations (14.90) and (14.97) together with $[P_A, P_B] = 0$, restated for the physical signature in Equations (39.18) to (39.20) — granting the identification of that realization with the Lie algebra of $L_+^\uparrow$ in the parameterization Equation (94.3), a construction recorded as pending at Equation (14.86) and in Section 39.3.2. Multiplying each by $\mathrm{i}\hbar$ and raising indices with η gives Equations (94.5) to (94.7). □

Corollary 94.9 (Rotations, boosts, translations)**.** *Specializing indices in Equations (94.5) and (94.6) via Equation (94.4),*

$$\begin{aligned}
&\left[\hat{J}_i, \hat{J}_j\right] = \mathrm{i}\hbar\,\epsilon_{ijk}\hat{J}_k\,, \quad \left[\hat{J}_i, \hat{K}_j\right] = \mathrm{i}\hbar\,\epsilon_{ijk}\hat{K}_k\,, \quad \left[\hat{K}_i, \hat{K}_j\right] = -\mathrm{i}\hbar\,\epsilon_{ijk}\hat{J}_k\,,\\
&\left[\hat{J}_i, \hat{P}_j\right] = \mathrm{i}\hbar\,\epsilon_{ijk}\hat{P}_k\,, \quad \left[\hat{K}_i, \hat{P}_j\right] = \frac{\mathrm{i}\hbar}{c}\,\delta_{ij}\hat{H}\,, \qquad \left[\hat{K}_i, \hat{H}\right] = \mathrm{i}\hbar c\,\hat{P}_i\,,
\end{aligned} \tag{94.8}$$

and $\hat{H}$ commutes with $\hat{P}_i$ and $\hat{J}_i$. The relative sign between the first and third brackets — rotations close on rotations, boosts close on rotations with the opposite *sign — is the non-compactness of the Lorentz group (Equation (39.3)) seen at the algebra level.*

Proof. Direct substitution. As a sample, $\left[\hat{K}_1, \hat{P}_1\right] = \left[\hat{J}^{10}, \hat{P}^1\right] = \mathrm{i}\hbar\left(\eta^{01}\hat{P}^1 - \eta^{11}\hat{P}^0\right) = \mathrm{i}\hbar\,\hat{P}^0 = \frac{\mathrm{i}\hbar}{c}\hat{H}$, and $\left[\hat{J}_1, \hat{J}_2\right] = \left[\hat{J}^{23}, \hat{J}^{31}\right] = \mathrm{i}\hbar\,\eta^{33}\hat{J}^{12}\cdot(-1)^0 = \mathrm{i}\hbar\,\hat{J}^{12} = \mathrm{i}\hbar\,\hat{J}_3$, all other η factors vanishing because the index pairs share no common value. The remaining entries are the same computation with permuted indices. □

94.3 The universal cover $\mathrm{SL}(2,\mathbb{C})$

Theorem 94.5 directs us to the covering group. It is concretely available: to each four-vector associate the Hermitian 2×2 matrix

$$X = x^\mu\sigma_\mu = \begin{pmatrix} x^0 + x^3 & x^1 - \mathrm{i}x^2 \\ x^1 + \mathrm{i}x^2 & x^0 - x^3 \end{pmatrix}, \qquad \sigma_\mu = (\mathbb{1}, \sigma_1, \sigma_2, \sigma_3)\,, \tag{94.9}$$

with σ_i the Pauli matrices (Chapter 80). The map is a linear bijection onto Hermitian matrices, and

$$\det X = (x^0)^2 - \boldsymbol{x}^2 = x_\mu x^\mu\,. \tag{94.10}$$

Proposition 94.10 ($\mathrm{SL}(2,\mathbb{C})$ covers the proper orthochronous Lorentz group)**.** *For $A\in\mathrm{SL}(2,\mathbb{C})$ define $\Lambda(A)$ by $(\Lambda(A)x)^\mu\,\sigma_\mu = A\,(x^\mu\sigma_\mu)\,A^\dagger$. Then $A\mapsto\Lambda(A)$ is a group homomorphism onto $L_+^\uparrow$ with kernel $\{\mathbb{1},-\mathbb{1}\}$: the map is two-to-one, and $\mathrm{SL}(2,\mathbb{C})$, being simply connected, is the universal cover.*

Proof. $AXA^\dagger$ is Hermitian when X is, so $\Lambda(A)$ is a well-defined real-linear map of $\mathbb{R}^4$; it preserves $x_\mu x^\mu$ by Equation (94.10) because $\det(AXA^\dagger) = |\det A|^2\det X = \det X$, hence $\Lambda(A)$ is a Lorentz transformation (Definition 39.1). Homomorphism: $\Lambda(AB)x \leftrightarrow AB\,X\,B^\dagger A^\dagger \leftrightarrow \Lambda(A)\Lambda(B)x$. *Kernel:* if $AXA^\dagger = X$ for all Hermitian X, then $X=\mathbb{1}$ gives $AA^\dagger=\mathbb{1}$, so A is unitary and commutes with every σ_i; a 2×2 matrix commuting with all Pauli matrices is a multiple $\lambda\mathbb{1}$ (write it in the basis $\{\mathbb{1},\sigma_i\}$ and use $\sigma_i\sigma_j = \delta_{ij}\mathbb{1} + \mathrm{i}\epsilon_{ijk}\sigma_k$), and $\det A = \lambda^2 = 1$ forces $\lambda = \pm1$. *Connectedness and image:* polar decomposition writes every $A\in\mathrm{SL}(2,\mathbb{C})$ uniquely as $A = V\mathrm{e}^M$ with $V\in\mathrm{SU}(2)$ and M traceless Hermitian; $\mathrm{SU}(2)$ is the three-sphere $|a|^2+|b|^2=1$ of rows $(a,b),(-\bar b,\bar a)$ and the traceless Hermitian M form $\mathbb{R}^3$, so $\mathrm{SL}(2,\mathbb{C})$ is connected and simply connected ($S^3\times\mathbb{R}^3$; the polar factors depend continuously on A, the positive square root and logarithm being continuous on positive matrices, and $\pi_1(S^3)=0$). Its continuous image $\Lambda(\mathrm{SL}(2,\mathbb{C}))$ is then a connected subgroup of L containing $\mathbb{1}$, hence lands in $L_+^\uparrow$ (Proposition 39.6: both defining conditions are clopen). That it is *all* of $L_+^\uparrow$ follows from Lemmas 94.11 to 94.13: every rotation and every boost is attained, and every $\Lambda\in L_+^\uparrow$ is a boost times a rotation. Finally, a surjective continuous homomorphism of Lie groups with discrete kernel is a covering map — a standard fact used here without restatement. In particular $L_+^\uparrow$, as the continuous image of the connected $\mathrm{SL}(2,\mathbb{C})$, is itself connected: it is the identity component of L, as announced in Section 39.2.4. □

Lemma 94.11 (Unitary elements give rotations)**.** *For a unit vector $\boldsymbol{n}$ and $V = \exp\left(-\mathrm{i}\frac{\theta}{2}\boldsymbol{n}\cdot\boldsymbol{\sigma}\right)\in\mathrm{SU}(2)$, the map $\Lambda(V)$ fixes x^0 and rotates $\boldsymbol{x}$ by the angle θ about $\boldsymbol{n}$. Every rotation arises this way, and $V(\theta=2\pi) = -\mathbb{1}$.*

Proof. Since $(\boldsymbol{n}\cdot\boldsymbol{\sigma})^2 = \mathbb{1}$ (from $\sigma_i\sigma_j = \delta_{ij}\mathbb{1} + \mathrm{i}\epsilon_{ijk}\sigma_k$ and $|\boldsymbol{n}|=1$), the exponential series (Equation (7.40)) splits into $V = \cos\frac{\theta}{2}\,\mathbb{1} - \mathrm{i}\sin\frac{\theta}{2}\,\boldsymbol{n}\cdot\boldsymbol{\sigma}$. A direct computation with the same product identity gives, for $X = x^0\mathbb{1} + \boldsymbol{x}\cdot\boldsymbol{\sigma}$,

$$VXV^\dagger = x^0\mathbb{1} + \left[\cos\theta\,\boldsymbol{x} + \sin\theta\,(\boldsymbol{n}\times\boldsymbol{x}) + (1-\cos\theta)(\boldsymbol{n}\cdot\boldsymbol{x})\,\boldsymbol{n}\right]\cdot\boldsymbol{\sigma}\,,$$

which is Rodrigues' rotation formula about $\boldsymbol{n}$ by θ. Every rotation has an axis and an angle: for $R\in\mathrm{SO}(3)$, $\det(R-\mathbb{1}) = \det\left((R-\mathbb{1})^{\mathsf{T}}\right) = \det(R^{-1}-\mathbb{1}) = \det(R^{-1})\det(\mathbb{1}-R) =$

$-\det(R-\mathbb{1})$, so $\det(R-\mathbb{1})=0$ and R fixes a unit vector $\boldsymbol{n}$; on the invariant plane orthogonal to $\boldsymbol{n}$, R is an orientation-preserving linear isometry, i.e. a rotation by some θ. Hence all of SO(3) is attained; at $\theta=2\pi$, $V=\cos\pi\,\mathbb{1}=-\mathbb{1}$. □

Lemma 94.12 (Hermitian elements give boosts)**.** *For $H=\exp\left(\frac{\eta}{2}\boldsymbol{m}\cdot\boldsymbol{\sigma}\right)$ with $\boldsymbol{m}$ a unit vector and $\eta\in\mathbb{R}$, the map $\Lambda(H)$ is the pure boost of rapidity η along $\boldsymbol{m}$: for $\boldsymbol{m}=\boldsymbol{e}_3$,*

$$x^0\mapsto x^0\cosh\eta+x^3\sinh\eta\,,\quad x^3\mapsto x^0\sinh\eta+x^3\cosh\eta\,,\quad x^{1,2}\mapsto x^{1,2}\,. \tag{94.11}$$

The dictionary to the velocity parameterization of Chapter 38 is $\beta=\tanh\eta$, $\gamma=\cosh\eta$; since η ranges over all of $\mathbb{R}$ and $\boldsymbol{m}$ over all directions, every boost is attained.

Proof. For $\boldsymbol{m}=\boldsymbol{e}_3$ we have $H=\operatorname{diag}\left(\mathrm{e}^{\eta/2},\mathrm{e}^{-\eta/2}\right)$, and in Equation (94.9) the conjugation $HXH^\dagger=HXH$ multiplies the diagonal entries $x^0\pm x^3$ of X by $\mathrm{e}^{\pm\eta}$ while leaving the off-diagonal entries $x^1\mp\mathrm{i}x^2$ fixed ($\mathrm{e}^{\eta/2}\mathrm{e}^{-\eta/2}=1$): $x^0+x^3\mapsto\mathrm{e}^{\eta}(x^0+x^3)$, $x^0-x^3\mapsto\mathrm{e}^{-\eta}(x^0-x^3)$, which is Equation (94.11). A general axis $\boldsymbol{m}$ is $R\,\boldsymbol{e}_3$ for a rotation R; conjugating H by the corresponding V of Lemma 94.11 rotates the boost axis, since Λ is a homomorphism. □

Lemma 94.13 (Boost–rotation decomposition)**.** *Every $\Lambda\in L_+^\uparrow$ factors as $\Lambda=BR$ with B a pure boost and R a rotation.*

Proof. Let $u=\Lambda e_0$, where $e_0=(1,0,0,0)$. Then $u_\mu u^\mu=1$ and $u^0=\Lambda^0{}_0\geq 1$ (Definition 39.5), so $u=(\cosh\eta,\ \sinh\eta\,\boldsymbol{m})$ for some rapidity $\eta\geq 0$ and unit vector $\boldsymbol{m}$. Let B be the boost of Lemma 94.12 along $\boldsymbol{m}$ with rapidity η; then $Be_0=u$, so $R=B^{-1}\Lambda$ satisfies $Re_0=e_0$. A Lorentz transformation fixing e_0 preserves the spatial slice $x^0=0$ and the Euclidean norm on it (both statements from $\Lambda\eta\Lambda^{\mathsf{T}}=\eta$, Equation (39.2): the first column of R is e_0, hence so is the first row), so $R\in\mathrm{O}(3)$; and $\det R=\det\Lambda=1$ places $R\in\mathrm{SO}(3)$. □

Corollary 94.14 (SU(2) covers SO(3); half-integer spin is projective)**.** *Restricting Proposition 94.10 to the rotation subgroup: $\mathrm{SU}(2)\to\mathrm{SO}(3)$ is a two-to-one covering with kernel $\pm\mathbb{1}$. A representation of $\mathrm{SU}(2)$ in which $-\mathbb{1}\mapsto-\mathbb{1}$ — half-integer spin, as classified in Lemma 94.28 below — is therefore a* projective *representation of the rotation group itself, with precisely the 4π periodicity observed in [1566] (Remark 94.6).*

94.4 The invariants: mass and spin

Which combinations of the generators are the same in every inertial frame? By Schur's lemma (Theorem 5.153) an operator commuting with the whole representation acts as a number on an irreducible one: such *Casimir operators* are the invariant labels we are looking for.

Definition 94.15 (Pauli–Lubanski vector)**.**

$$\hat{W}_\mu=\tfrac{1}{2}\,\epsilon_{\mu\nu\rho\sigma}\,\hat{J}^{\nu\rho}\hat{P}^\sigma\,,\qquad \epsilon_{0123}=+1\,. \tag{94.12}$$

Lemma 94.16 (Invariance identity for ϵ)**.** *For every $\omega\in\mathfrak{so}(3,1)$, i.e. $\omega_{\mu\nu}=-\omega_{\nu\mu}$,*

$$\omega^\alpha{}_\mu\epsilon_{\alpha\nu\rho\sigma}+\omega^\alpha{}_\nu\epsilon_{\mu\alpha\rho\sigma}+\omega^\alpha{}_\rho\epsilon_{\mu\nu\alpha\sigma}+\omega^\alpha{}_\sigma\epsilon_{\mu\nu\rho\alpha}=0\,. \tag{94.13}$$

Proof. The left side is totally antisymmetric in $\mu\nu\rho\sigma$: swapping any two of them exchanges two terms and flips the sign of the remaining two through the antisymmetry of ϵ. A totally antisymmetric four-index array in four dimensions is a multiple of $\epsilon_{\mu\nu\rho\sigma}$; contracting the left side with $\epsilon^{\mu\nu\rho\sigma}$ shows the multiple is proportional to $\omega^\alpha{}_\alpha=\eta^{\alpha\beta}\omega_{\beta\alpha}=0$. □

Lemma 94.17 (Properties of $\hat{W}$)**.** *1.* $\left[\hat{W}_\mu,\hat{P}_\lambda\right]=0$;

2. $\hat{W}_\mu \hat{P}^\mu = \hat{P}^\mu \hat{W}_\mu = 0$;

3. $\hat{W}$ *is a four-vector:* $\left[\hat{J}_{\mu\nu}, \hat{W}_\lambda\right] = \mathrm{i}\hbar\left(\eta_{\nu\lambda}\hat{W}_\mu - \eta_{\mu\lambda}\hat{W}_\nu\right)$.

Proof. (i) By Equations (94.6) and (94.7), $\left[\hat{J}^{\nu\rho}\hat{P}^\sigma, \hat{P}_\lambda\right] = \left[\hat{J}^{\nu\rho}, \hat{P}_\lambda\right]\hat{P}^\sigma = \mathrm{i}\hbar\left(\delta^\rho_\lambda \hat{P}^\nu - \delta^\nu_\lambda \hat{P}^\rho\right)\hat{P}^\sigma$; contracted with $\epsilon_{\mu\nu\rho\sigma}$ each term dies because $\hat{P}^\nu\hat{P}^\sigma$ (respectively $\hat{P}^\rho\hat{P}^\sigma$) is symmetric in indices that ϵ antisymmetrizes.

(ii) $\hat{W}_\mu\hat{P}^\mu = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma}\hat{J}^{\nu\rho}\hat{P}^\sigma\hat{P}^\mu = 0$ for the same symmetry reason, using Equation (94.7); the reversed order follows from (i).

(iii) Insert Equations (94.5) and (94.6) into $\left[\hat{J}_{\mu\nu}, \frac{1}{2}\epsilon_{\lambda\alpha\beta\gamma}\hat{J}^{\alpha\beta}\hat{P}^\gamma\right]$ and use the Leibniz rule for commutators; the resulting four $\epsilon\eta$-contractions combine, by Lemma 94.16 applied with ω the antisymmetrized pair $(\mu\nu)$, into exactly the two terms $\mathrm{i}\hbar(\eta_{\nu\lambda}\hat{W}_\mu - \eta_{\mu\lambda}\hat{W}_\nu)$. □

Theorem 94.18 (The Casimir invariants of the Poincaré group)**.** $\hat{P}^2 = \hat{P}_\mu\hat{P}^\mu$ *and* $\hat{W}^2 = \hat{W}_\mu\hat{W}^\mu$ *commute with every generator of Proposition 94.8. On an irreducible representation each is a real number.*

Proof. $\left[\hat{P}^2, \hat{P}_\lambda\right] = 0$ by Equation (94.7), and $\left[\hat{P}^2, \hat{J}_{\mu\nu}\right] = 0$ because Equation (94.6) makes $\hat{P}$ a four-vector and the square of a four-vector is a scalar: $\left[\hat{J}_{\mu\nu}, \hat{P}_\lambda\hat{P}^\lambda\right] = \mathrm{i}\hbar\left(\eta_{\nu\lambda}\hat{P}_\mu - \eta_{\mu\lambda}\hat{P}_\nu\right)\hat{P}^\lambda + \mathrm{i}\hbar\,\hat{P}_\lambda\left(\delta^\lambda_\nu\hat{P}_\mu - \delta^\lambda_\mu\hat{P}_\nu\right) = \mathrm{i}\hbar\left(\hat{P}_\mu\hat{P}_\nu - \hat{P}_\nu\hat{P}_\mu\right) + \mathrm{i}\hbar\left(\hat{P}_\nu\hat{P}_\mu - \hat{P}_\mu\hat{P}_\nu\right) = 0$ identically. The same computation with Lemma 94.17(iii) kills $\left[\hat{J}_{\mu\nu}, \hat{W}^2\right]$, and $\left[\hat{W}^2, \hat{P}_\lambda\right] = 0$ is Lemma 94.17(i). Schur's lemma (Theorem 5.153, with the qualification of Remark 94.1) then makes each a scalar on an irreducible representation; Hermiticity makes the scalar real. □

Definition 94.19 (Mass)**.** On an irreducible representation write

$$\hat{P}^2 = m^2c^2\,\mathbb{1}\,. \tag{94.14}$$

For the representations selected by observation (Remark 94.22 below) $m^2 \geq 0$; the number $m \geq 0$ is the *mass* of the particle. The value of $\hat{W}^2$ will emerge as the spin in Theorem 94.29 and Proposition 94.32.

94.5 Orbits, little groups, and induced representations

Since the $\hat{P}^\mu$ commute (Equation (94.7)), diagonalize them: a basis $|p,\sigma\rangle$ with $\hat{P}^\mu|p,\sigma\rangle = p^\mu|p,\sigma\rangle$, where σ enumerates whatever degeneracy remains. Translations then act by $U(\mathbb{1},a)|p,\sigma\rangle = \mathrm{e}^{-\mathrm{i}\,a\cdot p/\hbar}|p,\sigma\rangle$, and by Equation (94.6) a Lorentz transformation carries the eigenvalue p to Λp: $\hat{P}^\mu U(\Lambda)|p,\sigma\rangle = (\Lambda p)^\mu U(\Lambda)|p,\sigma\rangle$ (conjugate $\hat{P}$ through $U(\Lambda)$ with Equations (94.3) and (94.6)). An irreducible representation therefore lives on a single *orbit* of momenta.

Lemma 94.20 (Orthochronous transformations preserve the sign of p^0 on causal vectors)**.** *If* $p_\mu p^\mu \geq 0$, $p \neq 0$, *and* $\Lambda \in L^\uparrow_+$, *then* $\operatorname{sgn}\left((\Lambda p)^0\right) = \operatorname{sgn}\left(p^0\right)$.

Proof. $(\Lambda p)^0 = \Lambda^0{}_0 p^0 + \Lambda^0{}_i p^i$. By the Cauchy–Schwarz inequality of Section 6.3 and the row identity $\left(\Lambda^0{}_0\right)^2 = 1 + \sum_i \left(\Lambda^0{}_i\right)^2$ (the $\mu=\nu=0$ component of Equation (39.2) applied to $\Lambda^{-1} = \eta\Lambda^{\mathsf{T}}\eta$),

$$\left|\Lambda^0{}_i p^i\right| \leq \sqrt{\textstyle\sum_i \left(\Lambda^0{}_i\right)^2}\;|\boldsymbol{p}| < \Lambda^0{}_0\,\left|p^0\right|$$

whenever $|\boldsymbol{p}| \leq |p^0|$ with $p \neq 0$ (causality) — strictly, since either the causal inequality is strict or $\sum_i(\Lambda^0{}_i)^2 < (\Lambda^0{}_0)^2$. Hence the first term dominates and carries the sign of p^0 ($\Lambda^0{}_0 \geq 1$, Definition 39.5). □

Proposition 94.21 (Orbit classification)**.** *Under $L_+^\uparrow$ the nonzero momenta fall into the disjoint orbits*

orbit	*invariants*	*standard momentum k*
massive, positive energy	$p^2 = m^2c^2 > 0$, $p^0 > 0$	$(mc, 0, 0, 0)$
massive, negative energy	$p^2 = m^2c^2 > 0$, $p^0 < 0$	$(-mc, 0, 0, 0)$
massless, positive energy	$p^2 = 0$, $p^0 > 0$	$(\kappa, 0, 0, \kappa)$
massless, negative energy	$p^2 = 0$, $p^0 < 0$	$(-\kappa, 0, 0, \kappa)$
tachyonic	$p^2 = -\mu^2c^2 < 0$	$(0, 0, 0, \mu c)$

together with the fixed point $p = 0$; here $\kappa > 0$ is any fixed reference momentum (different κ label the same orbit).

Proof. p^2 is invariant by definition of L, and $\operatorname{sgn}(p^0)$ is invariant on the causal orbits by Lemma 94.20, so the listed classes are unions of orbits. Each class is a *single* orbit because its standard momentum reaches every member: a rotation (Lemma 94.11) aligns $\boldsymbol{p}$ with $\boldsymbol{e}_3$, and the boost Equation (94.11) with the appropriate rapidity then produces any admissible p^0: $\cosh\eta = p^0/(mc)$ in the massive case, $\mathrm{e}^\eta = p^0/\kappa$ in the massless case, $\sinh\eta = p^0/(\mu c)$ in the tachyonic case, the spatial part then matching automatically — which also shows both signs of p^0 lie on *one* tachyonic orbit, sinh being unbounded in both directions; the negative-energy rows follow by applying the same transformations to their own standard momenta. The composition of these transformations is the *standard boost* $L(p)$, fixed once and for all on each orbit:

$$L(p) = R(\hat{\boldsymbol{p}})\, B_3(\eta_p)\,, \qquad L(p)\, k = p\,, \tag{94.15}$$

with $R(\hat{\boldsymbol{p}})$ a fixed choice of rotation carrying $\boldsymbol{e}_3$ to the unit vector $\hat{\boldsymbol{p}}$ and B_3 the boost Equation (94.11); the conventions are $R(\boldsymbol{e}_3) = \mathbb{1}$ — so that $L(k) = \mathbb{1}$ — and a fixed measurable choice of $R(\hat{\boldsymbol{p}})$ elsewhere (no continuous global choice exists); any such choice serves. □

Remark 94.22 (Observation selects two orbit classes)*.* Nature realizes the massive and massless positive-energy orbits, and only those. Negative-energy representations would make every system unstable against decay into ever-lower energies — the resolution of this instability is the reinterpretation of the negative-frequency modes in quantum field theory, anticipated by the positron (Chapter 95) — and no tachyonic particle ($p^2 < 0$) has ever been observed. This is an *empirical* input to the classification, not a mathematical one, and this treatise records it as such (cf. the front-matter scope rules).

Definition 94.23 (Little group)**.** The *little group* (stabilizer) of the standard momentum k is

$$G_k = \left\{ \Lambda \in L_+^\uparrow \mid \Lambda k = k \right\}. \tag{94.16}$$

Proposition 94.24 (Massive little group)**.** *For $k = (mc, 0, 0, 0)$, $G_k = \mathrm{SO}(3)$, with cover* $\mathrm{SU}(2)$*.*

Proof. $\Lambda k = k$ says $\Lambda e_0 = e_0$; the proof of Lemma 94.13 showed exactly that such a Λ is a rotation, and conversely rotations fix e_0. The cover statement is Corollary 94.14. □

Proposition 94.25 (Massless little group)**.** *For $k = (\kappa, 0, 0, \kappa)$ the Lie algebra of G_k is spanned by*

$$\hat{J}^3 = \hat{J}^{12}\,, \qquad \hat{A} = \hat{J}^{01} + \hat{J}^{13}\,, \qquad \hat{B} = \hat{J}^{02} + \hat{J}^{23}\,, \tag{94.17}$$

with brackets

$$\left[\hat{J}^3, \hat{A}\right] = \mathrm{i}\hbar\, \hat{B}\,, \qquad \left[\hat{J}^3, \hat{B}\right] = -\mathrm{i}\hbar\, \hat{A}\,, \qquad \left[\hat{A}, \hat{B}\right] = 0\,. \tag{94.18}$$

This is the algebra $\mathfrak{iso}(2)$ *of the Euclidean group of the plane — a rotation* ($\hat{J}^3$) *and two commuting translations* ($\hat{A}, \hat{B}$) *— cf. the semidirect structure Equation* (14.99) *in two Euclidean dimensions.*

Proof. An algebra element $\frac{1}{2}\omega_{\mu\nu}\hat{J}^{\mu\nu}$ stabilizes k iff $\omega^{\mu}{}_{\nu}k^{\nu} = 0$, i.e. $\omega^{\mu}{}_{0} + \omega^{\mu}{}_{3} = 0$ for every μ. Writing this out with η: for $\mu = 0$ it gives $\omega_{03} = 0$; for $\mu = 1$, $\omega_{01} = \omega_{13}$; for $\mu = 2$, $\omega_{02} = \omega_{23}$; for $\mu = 3$ it repeats $\omega_{03} = 0$. Three free parameters remain, multiplying exactly the three combinations Equation (94.17). The brackets follow from Equation (94.5) by direct substitution; as a sample, in $\left[\hat{J}^{12}, \hat{J}^{01}\right]$ the only surviving term of Equation (94.5) is $\eta^{\mu\rho}\hat{J}^{\nu\lambda}$ with $(\mu\nu\lambda\rho) = (1, 2, 0, 1)$, giving $\mathrm{i}\hbar\,\eta^{11}\hat{J}^{20} = \mathrm{i}\hbar\,\hat{J}^{02}$, and $\left[\hat{J}^{12}, \hat{J}^{13}\right] = \mathrm{i}\hbar\,\eta^{11}\hat{J}^{32} = \mathrm{i}\hbar\,\hat{J}^{23}$, whose sum is $\mathrm{i}\hbar\,\hat{B}$; the other two brackets are the same bookkeeping, with the two disjoint-index commutators $\left[\hat{J}^{01}, \hat{J}^{23}\right] = \left[\hat{J}^{13}, \hat{J}^{02}\right] = 0$ making $\left[\hat{A}, \hat{B}\right]$ collapse to $\left[\hat{J}^{01}, \hat{J}^{02}\right] + \left[\hat{J}^{13}, \hat{J}^{23}\right] = \mathrm{i}\hbar\left(-\hat{J}^{12} + \hat{J}^{12}\right) = 0$. □

Lemma 94.26 (Lorentz-invariant measure on a mass shell)**.** *On the orbit* $p^2 = m^2c^2 \geq 0$, $p^0 > 0$, *the measure*

$$\mathrm{d}\mu(p) = \frac{\mathrm{d}^3 p}{2p^0}\,, \qquad p^0 = \sqrt{\boldsymbol{p}^2 + m^2c^2}\,, \tag{94.19}$$

satisfies $\int f(\Lambda^{-1}p)\,\mathrm{d}\mu(p) = \int f(p)\,\mathrm{d}\mu(p)$ *for every* $\Lambda \in L_+^{\uparrow}$ *and integrable* f.

Proof. Write the integral over the orbit as a four-dimensional one,

$$\int f(p)\,\frac{\mathrm{d}^3 p}{2p^0} = \int \mathrm{d}^4 p\;\delta\!\left(p_\mu p^\mu - m^2c^2\right)\theta\!\left(p^0\right) f(p)\,,$$

which follows from the composition rule for the Dirac delta (Proposition 7.113): at fixed $\boldsymbol{p}$ the argument $\left(p^0\right)^2 - \boldsymbol{p}^2 - m^2c^2$ has the single positive root Equation (94.19) with derivative $2p^0$. In the four-dimensional form each factor is invariant under $p \mapsto \Lambda p$: the delta because p^2 is, the step function by Lemma 94.20, and $\mathrm{d}^4 p$ because $|\det\Lambda| = 1$. □

Theorem 94.27 (Wigner's induced representations)**.** *Fix a causal positive-energy orbit of Proposition 94.21 — the ones Lemma 94.26 equips with an invariant measure, and the only ones used in the sequel — with standard momentum* k, *and a unitary irreducible representation* $D_{\sigma'\sigma}$ *of (the cover of) the little group* G_k *on a space with basis* $\{\sigma\}$. *For half-integer representations everything is read in the covering group:* $L(p)$ *then denotes the fixed lift of the standard boost Equation* (94.15) *assembled from the* $\mathrm{SL}(2,\mathbb{C})$ *elements of Lemmas 94.11 and 94.12, so that* $W(\Lambda, p)$ *below lies in the cover of* G_k *and* $D(W)$ *is unambiguous; for integer spin the formulas descend to* $L_+^{\uparrow}$. *Define one-particle states by* $|k, \sigma\rangle$ *carrying* D, *and*

$$|p, \sigma\rangle = U(L(p))\,|k, \sigma\rangle\,, \tag{94.20}$$

with $L(p)$ *the standard boost Equation* (94.15) *and the normalization* $\langle p', \sigma'|p, \sigma\rangle = 2p^0\,\delta^3(\boldsymbol{p}' - \boldsymbol{p})\,\delta_{\sigma'\sigma}$. *Then*

$$U(\mathbb{1}, a)\;|p, \sigma\rangle = \mathrm{e}^{-\mathrm{i}\,p\cdot a/\hbar}\;|p, \sigma\rangle\,, \tag{94.21}$$

$$U(\Lambda, 0)\;|p, \sigma\rangle = \sum_{\sigma'} D_{\sigma'\sigma}(W(\Lambda, p))\;|\Lambda p, \sigma'\rangle\,, \qquad W(\Lambda, p) = L(\Lambda p)^{-1}\,\Lambda\,L(p)\,, \tag{94.22}$$

defines a unitary representation of the covering group; the element $W(\Lambda, p)$ *— the* Wigner rotation *— indeed lies in* G_k.

Proof. $W(\Lambda,p)k = L(\Lambda p)^{-1}\Lambda L(p)\,k = L(\Lambda p)^{-1}(\Lambda p) = k$, so $W(\Lambda,p)\in G_k$. Composition: applying Equation (94.22) twice,

$$W(\Lambda',\Lambda p)\,W(\Lambda,p) = L(\Lambda'\Lambda p)^{-1}\Lambda' L(\Lambda p)\,L(\Lambda p)^{-1}\Lambda L(p) = W(\Lambda'\Lambda,p)\,,$$

so the D factors compose correctly and $U(\Lambda')U(\Lambda) = U(\Lambda'\Lambda)$; translations compose by inspection of Equation (94.21), and the mixed products follow from Equation (94.2) applied to Λa. Unitarity: for wave packets $|\psi\rangle = \sum_\sigma \int \mathrm{d}\mu(p)\,\psi_\sigma(p)\,|p,\sigma\rangle$ the chosen normalization gives $\|\psi\|^2 = \sum_\sigma \int \mathrm{d}\mu(p)\,|\psi_\sigma(p)|^2$, and by Equation (94.22) a Lorentz transformation replaces $\psi_\sigma(p)$ by $\sum_{\sigma'} D_{\sigma\sigma'}(W)\,\psi_{\sigma'}(\Lambda^{-1}p)$: the sum over σ is preserved because D is unitary, and the integral because the measure is invariant (Lemma 94.26). □

Derivation pending #118

Completeness of the classification: that the representation defined by Equations (94.21) and (94.22) is irreducible precisely when D is, and that every *irreducible unitary representation of the cover with $\hat{P}$-spectrum on the given orbit is equivalent to an induced one, is Mackey's imprimitivity theorem, specialized by Wigner to this group [849]. The systematic imprimitivity machinery is not yet developed in this treatise.*

94.6 Massive particles: spin

By Proposition 94.24 and Theorem 94.27, classifying massive particles means classifying the unitary irreducible representations of SU(2) — equivalently, of its algebra $\left[\hat{J}_i,\hat{J}_j\right] = \mathrm{i}\hbar\,\epsilon_{ijk}\hat{J}_k$ (Equation (94.8)).

Lemma 94.28 (Unitary irreducible representations of $\mathfrak{su}(2)$)**.** *The finite-dimensional unitary irreducible representations of the algebra $\left[\hat{J}_i,\hat{J}_j\right] = \mathrm{i}\hbar\,\epsilon_{ijk}\hat{J}_k$ are labelled by $s\in\left\{0,\frac{1}{2},1,\frac{3}{2},\dots\right\}$. The representation labelled s has dimension $2s+1$, carries $\hat{\boldsymbol{J}}^2 = \hbar^2 s(s+1)\,\mathbb{1}$, and admits a basis $|s,\sigma\rangle$, $\sigma = -s,-s+1,\dots,s$, with $\hat{J}_3\,|s,\sigma\rangle = \hbar\sigma\,|s,\sigma\rangle$.*

Proof. Let $\hat{J}_\pm = \hat{J}_1 \pm \mathrm{i}\hat{J}_2$, so that $\hat{J}_\pm^\dagger = \hat{J}_\mp$ and, from the algebra,

$$\left[\hat{J}_3,\hat{J}_\pm\right] = \pm\hbar\,\hat{J}_\pm\,,\qquad \left[\hat{J}_+,\hat{J}_-\right] = 2\hbar\,\hat{J}_3\,,\qquad \hat{\boldsymbol{J}}^2 = \hat{J}_\mp\hat{J}_\pm + \hat{J}_3^2 \pm \hbar\hat{J}_3\,. \tag{94.23}$$

$\hat{\boldsymbol{J}}^2$ commutes with each $\hat{J}_i$ (same computation as in Theorem 94.18, three-dimensional case), so on an irreducible representation $\hat{\boldsymbol{J}}^2 = \lambda\hbar^2\mathbb{1}$ with $\lambda\geq 0$ (it is a sum of squares of Hermitian operators). Diagonalize the Hermitian $\hat{J}_3$: eigenvalues $\hbar\sigma$. The first bracket in Equation (94.23) shifts σ by ±1 under $\hat{J}_\pm$, while

$$\left\|\hat{J}_\pm\,|\sigma\rangle\right\|^2 = \langle\sigma|\,\hat{J}_\mp\hat{J}_\pm\,|\sigma\rangle = \hbar^2\left[\lambda-\sigma(\sigma\pm1)\right]\;\geq 0\,. \tag{94.24}$$

In a finite-dimensional space the ladder must terminate in both directions: there are extremal eigenvalues $\sigma_\pm$ with $\hat{J}_+\,|\sigma_+\rangle = 0$ and $\hat{J}_-\,|\sigma_-\rangle = 0$, so Equation (94.24) forces $\lambda = \sigma_+(\sigma_++1) = \sigma_-(\sigma_--1)$, whose only solution with $\sigma_+\geq\sigma_-$ is $\sigma_- = -\sigma_+$. Writing $s=\sigma_+$: the ladder descends from σ_+ to σ_- in unit steps, so $2s\in\mathbb{Z}_{\geq0}$ and $\lambda = s(s+1)$. The states $\{|s,\sigma\rangle\}_{\sigma=-s}^{s}$ span an invariant subspace (the three generators are linear combinations of $\hat{J}_3$ and $\hat{J}_\pm$, which act within it), hence by irreducibility the whole space: the dimension is $2s+1$. Conversely each such s is realized — Equation (94.24) defines the matrix elements, and the brackets Equation (94.23) are verified directly — so the list is exhaustive. (Infinite-dimensional unitary irreducible representations of a *compact* group do not occur: every unitary representation of SU(2) decomposes into finite-dimensional ones — the Peter–Weyl theorem [1912]; the treatise's own account of SU(2), Section 14.2.2, is still pending.) □

Theorem 94.29 (Massive particles)**.** *The irreducible positive-energy representations with $m > 0$ are classified by the pair (m, s) with $2s \in \mathbb{Z}_{\geq 0}$: mass and* spin*. The representation (m, s) has $2s + 1$ polarization states for each momentum, and*

$$\hat{W}^2 = -m^2c^2\,\hbar^2\,s(s+1)\,\mathbb{1}\,. \tag{94.25}$$

Proof. The classification is Theorem 94.27, Proposition 94.24, and Lemma 94.28, modulo the completeness statement recorded as pending after Theorem 94.27 (irreducibility and exhaustiveness of the induced family). For Equation (94.25) evaluate $\hat{W}^2$ — a Casimir, so any convenient states will do (Theorem 94.18) — on the standard-momentum states $|k, \sigma\rangle$, $k = (mc, \mathbf{0})$. There $\hat{P}^\sigma$ may be replaced by its eigenvalue, in both factors of $\hat{W}^2$: $\hat{W}$ preserves the momentum eigenspace by Lemma 94.17(i). In Equation (94.12) only $\hat{P}^0 = mc$ survives,

$$\hat{W}_0\,|k,\sigma\rangle = \tfrac{1}{2}\epsilon_{0\nu\rho0}\,\hat{J}^{\nu\rho}mc\,|k,\sigma\rangle = 0\,, \qquad \hat{W}_i\,|k,\sigma\rangle = \tfrac{mc}{2}\,\epsilon_{i\nu\rho0}\,\hat{J}^{\nu\rho}\,|k,\sigma\rangle = -mc\,\hat{J}_i\,|k,\sigma\rangle\ ,$$

the last step because $\epsilon_{ijk0} = -\epsilon_{0ijk}$ (one cyclic shift of four indices is an odd permutation) and $\hat{J}_i = \frac{1}{2}\epsilon_{ijk}\hat{J}^{jk}$ Equation (94.4). Hence on these states $\hat{W}^2 = \hat{W}_0\hat{W}^0 + \hat{W}_i\hat{W}^i = -\sum_i \left(mc\,\hat{J}_i\right)^2 = -m^2c^2\,\hat{\boldsymbol{J}}^2$, and Lemma 94.28 supplies the eigenvalue $\hbar^2 s(s+1)$: the Pauli–Lubanski square measures the *rest-frame angular momentum* — the spin. □

Phenomenon 94.30 (Quantization of intrinsic angular momentum)**.** The intrinsic angular momentum of every particle species is quantized: measured along any axis it takes exactly $2s + 1$ equally spaced values $\hbar\sigma$, $\sigma = -s, \ldots, s$, with $2s$ a nonnegative integer characteristic of the species. For the electron and for silver atoms, $s = \frac{1}{2}$: a magnetic-gradient analyzer splits an unpolarized beam into exactly *two* components (Section 80.2), never a classical continuum; light transfers angular momentum to matter in units set by $\hbar$ per photon [767].

Derivation. Theorem 94.29: a massive particle species is an irreducible representation (m, s), its polarization space is the $(2s + 1)$-dimensional representation of Lemma 94.28, and any axis may be taken as the 3-axis by Proposition 94.24, so a measurement of the spin projection resolves the $2s+1$ eigenvalues $\hbar\sigma$ of $\hat{J}_3$ (Postulate 77.3). The massless counterpart — helicity $\pm\hbar$ for the photon — is Proposition 94.32 below. The experimental record is analyzed in Chapter 80 and [767]. □

94.7 Massless particles: helicity

Proposition 94.31 (Massless representations)**.** *On the massless positive-energy orbit, the representations realized by observed particles have $\hat{A}$ and $\hat{B}$ (Equation (94.17)) acting as zero. The little group then effectively reduces to the rotations $\mathrm{e}^{-\mathrm{i}\theta\hat{J}^3/\hbar}$ about the momentum axis, whose irreducible unitary representations are one-dimensional,*

$$D_\theta = \mathrm{e}^{-\mathrm{i}h\theta}\,, \qquad \hat{J}^3 = \hbar h\,\mathbb{1}\,, \qquad h \in \tfrac{1}{2}\mathbb{Z}\,, \tag{94.26}$$

and the number h — the helicity *— replaces spin: a massless species carries a single polarization state per momentum.*

Proof. Suppose instead $\hat{A}$ (say) had a nonzero eigenvalue a on some state. Since $\left[\hat{A}, \hat{B}\right] = 0$ Equation (94.18), diagonalize both; conjugating by the rotation $\mathrm{e}^{-\mathrm{i}\theta\hat{J}^3/\hbar}$ rotates the eigenvalue pair (a, b) by the angle θ in the plane (this is Equation (94.18) exponentiated), so the spectrum would contain the *whole circle* of pairs — a continuous infinity of internal states at fixed momentum, Wigner's "continuous spin" representations [849]. No observed massless particle has more than two polarization states, so observation forces $a = b = 0$: $\hat{A} = \hat{B} = 0$

on the physical representation (the same empirical selection as Remark 94.22). What remains of the little group is the one-parameter rotation subgroup; its irreducible unitary representations are one-dimensional because the group is abelian: each D_θ commutes with the whole representation, so by Schur's lemma (Theorem 5.153, with the qualification of Remark 94.1) it is a scalar, every line is then invariant, and irreducibility forces dimension one (cf. the SO(2) analysis of Section 14.2.1). Thus $D_\theta = \mathrm{e}^{-\mathrm{i}h\theta}$ with $\hat{J}^3$ acting as the number $\hbar h$. *Quantization of h:* in the covering group the rotation subgroup about $\boldsymbol{e}_3$ is $V(\theta) = \exp\left(-\mathrm{i}\frac{\theta}{2}\sigma_3\right)$ (Lemma 94.11), which closes only at $\theta = 4\pi$; single-valuedness of D on the cover demands $\mathrm{e}^{-4\pi\mathrm{i}h} = 1$, i.e. $2h \in \mathbb{Z}$. (A rotation by 2π multiplies the state by $\mathrm{e}^{-2\pi\mathrm{i}h} = (-1)^{2h}$, exactly as for massive spin, Corollary 94.14.) □

Proposition 94.32 (Helicity is the projection of angular momentum on the momentum)**.** *On a massless representation with helicity h,*

$$\hat{W}_\mu = \hbar h\, \hat{P}_\mu\,, \qquad \textit{equivalently} \qquad \frac{\hat{\boldsymbol{J}}\cdot\hat{\boldsymbol{P}}}{\left|\hat{\boldsymbol{P}}\right|} = \hbar h\, \mathbb{1}\,, \tag{94.27}$$

and $\hat{W}^2 = 0$.

Proof. Evaluate the components of Equation (94.12) on the standard-momentum states, $p = k = (\kappa, 0, 0, \kappa)$, replacing $\hat{P}$ by its eigenvalue ($k^0 = k^3 = \kappa$):

$$\hat{W}_0 = \tfrac{1}{2}\epsilon_{0\nu\rho3}\hat{J}^{\nu\rho}\kappa = \kappa\,\hat{J}^{12} = \kappa\,\hat{J}^3\,, \qquad \hat{W}_3 = \tfrac{1}{2}\epsilon_{3\nu\rho0}\hat{J}^{\nu\rho}\kappa = -\kappa\,\hat{J}^{12} = -\kappa\,\hat{J}^3\,,$$
$$\hat{W}_1 = -\kappa\left(\hat{J}^{23} + \hat{J}^{02}\right) = -\kappa\,\hat{B}\,, \qquad \hat{W}_2 = \kappa\left(\hat{J}^{13} + \hat{J}^{01}\right) = \kappa\,\hat{A}\,,$$

each line by direct expansion of the ϵ sums with $\epsilon_{0123} = +1$. On the physical representation $\hat{A} = \hat{B} = 0$ and $\hat{J}^3 = \hbar h$ (Proposition 94.31), so $\hat{W}_\mu = \hbar h\,(\kappa, 0, 0, -\kappa)_\mu = \hbar h\, k_\mu$ on these states. Both sides of $\hat{W}_\mu = \hbar h \hat{P}_\mu$ are four-vectors built on the representation (Lemma 94.17), so equality at the standard momentum propagates to the whole orbit by conjugation with $U(L(p))$. Contracting with itself, $\hat{W}^2 = \hbar^2 h^2 \hat{P}^2 = 0$; taking the spatial part at momentum $\boldsymbol{p}$ gives the stated projection formula ($\hat{W}^i$ reduces to $p^0\hat{J}_i$-terms along $\hat{\boldsymbol{p}}$). □

Remark 94.33 (The photon). The photon is massless to spectacular experimental accuracy ($m_\gamma c^2 < 1 \times 10^{-18}$ eV [2427]) and carries helicity $h = \pm 1$: circularly polarized light exerts a measurable torque, $\pm\hbar$ of angular momentum per photon, first detected by Beth with a torsion pendulum of birefringent plates [767]. The two signs are exchanged by spatial reflection, so the parity-symmetric electrodynamics of Part VII bundles $h = +1$ and $h = -1$ into one species, "the photon", with two polarization states — and *only* two: the missing longitudinal state of a massive spin-1 particle is not a small correction but a different orbit of the Poincaré group altogether. That a state $h = 0$ is absent is the group-theoretic seed of gauge invariance, taken up in Part XI. The same counting will reappear for the graviton candidates of Part XIII: helicity ± 2, two polarizations — the two polarizations derived for gravitational waves in Appendix A.1.

94.8 The particle dictionary

Definition 94.34 (Particle)**.** A *particle species* is an irreducible unitary representation of the universal cover $\mathbb{R}^4 \rtimes \mathrm{SL}(2,\mathbb{C})$ of the proper orthochronous Poincaré group, realized on the positive-energy massive or massless orbits, labelled by (m, s) with $m > 0$ and $2s \in \mathbb{Z}_{\geq 0}$, or by $(0, h)$ with $2h \in \mathbb{Z}$.

Every column of the experimental particle tables [2427] is an instance:

species	mc^2	spin/helicity	polarization states
electron	0.510 998 95 MeV [2392]	$s = \frac{1}{2}$	2 (Section 80.2)
proton	938.27 MeV [2427]	$s = \frac{1}{2}$	2 (composite, but still a $(m, \frac{1}{2})$ representation)
photon	0 ($< 1 \times 10^{-18}$ eV [2427])	$h = \pm 1$	2 [767]
Higgs boson	125.2 GeV [2427]	$s = 0$	1

The wave equations of the two preceding chapters are the *field-theoretic realizations* of these representations: the Klein–Gordon equation (Chapter 92) is nothing but the Casimir statement $\hat{P}^2 = m^2c^2$ Equation (94.14) written on wavefunctions, and the Dirac equation (Chapter 93) realizes the $(m, \frac{1}{2})$ representation on a field carrying the parity-paired $(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$ representation of the Lorentz group. Wigner's 1939 paper [849] thus stands at the head of everything Part XI builds: quantum field theory is the machinery that makes interactions compatible with this classification, and the Standard Model is a specific, experimentally determined list of which representations Nature uses.

Chapter 95

Experiment: The Positron

On 2 August 1932 a cosmic-ray track crossed Carl Anderson's cloud chamber at Caltech, curved the wrong way for an electron, and passed through a 6 mm lead plate emerging with less momentum than it entered — which fixed its direction of travel, and so its sign. Anderson announced the particle in *Science* within weeks [691] and published the full analysis the following March under the title "The Positive Electron" [713]. It was the first antiparticle ever seen, and it had been predicted: Dirac's equation (Chapter 93) required a particle of the electron's mass and opposite charge [647, 677], and no other theory then in existence asked for one. This chapter records the measurement in full phenomenological detail — apparatus, procedure, numbers with their uncertainties in SI units, and the inferences the numbers do and do not support — using the structured `experiment` environment of this treatise when the content is written.

Two things make the case worth the space. First, the inference chain is short and checkable: a track's curvature in a known field gives $B\rho$ and hence momentum, its ionization density gives a bound on the charge and mass, and the lead plate gives the direction; nothing else is needed to establish that the particle is positive, light, and not a proton. Second, the confirmation was immediate and independent — Blackett and Occhialini, using a counter-triggered chamber, published photographs of electron–positron showers within months [714], identifying pair production as the mechanism [726, 732] — and the sequel has run for ninety years, through annihilation radiation [741], positronium [980], the antiproton [1057] and trapped antihydrogen [2280], to a hospital imaging technique that works by annihilating antimatter [1563].

Derivation pending #119

Experiment: The Positron: all derivations of this chapter are pending.

95.1 Historical context and the prediction under test

95.1.1 Dirac's hole theory as a falsifiable prediction

[Reserved: the negative-energy solutions of the Dirac equation, Dirac's filled sea and his first, wrong identification of a hole with the proton [647]; the retraction and the 1931 prediction of an "anti-electron" of equal mass and opposite charge [677]; what exactly was predicted — mass equal to m_e, charge $+e$, and production in pairs — and therefore what the experiment had to measure to confirm or refute it; the state of the theory as of Chapter 93, and the honest note that Anderson was not looking for Dirac's particle and did not cite him.]

95.1.2 Cosmic rays as the available source

[Reserved: the penetrating radiation established by Hess's balloon ascents [387], the only source of particles above 100 MeV *available in 1932; the Caltech programme of measuring cosmic-ray energies in a magnetic cloud chamber; why the question "are the primaries positive or negative?" motivated the apparatus, and how that question made the positron an unintended discovery; the cosmic-ray physics itself is Chapter 115.]*

95.1.3 What else the track could have been

[Reserved: the exhaustive list of alternatives Anderson had to exclude [713] — a negative electron travelling upwards, excluded by the lead plate; a proton of the observed curvature, excluded by ionization density and by range, since a proton of 0.21 T m *would have stopped within about* 5 mm *while the track ran over* 5 cm*; a light nucleus; a chance coincidence of two unrelated tracks. The subsection should state each alternative with the observable that kills it, because this is the anatomy of an experimental claim.]*

95.2 Apparatus

95.2.1 The cloud chamber

[Reserved: Wilson's expansion chamber, in which a sudden adiabatic expansion supersaturates a vapour and ions left along a charged particle's path seed visible droplets [394]; Anderson's vertical chamber of about 17 cm *diameter and* 3 cm *depth, its gas filling and expansion ratio, and the stereoscopic camera pair [713]; the physical mechanism — ionization energy loss — linking droplet density to charge and velocity, and its treatment in Chapter 86.]*

95.2.2 The magnet

[Reserved: the water-cooled electromagnet producing a field of about 1.5 T *across the chamber, then among the strongest available; its calibration and uniformity, which set the systematic floor on every momentum quoted; the relation $p = qB\rho$ from Chapter 59 and its relativistic validity (Chapter 40), so that a measured sagitta gives a momentum without any assumption about the particle's mass.]*

95.2.3 The lead plate

[Reserved: the 6 mm *lead plate mounted across the middle of the chamber [713] — the single component that turns an ambiguous photograph into a measurement, since a particle loses energy in crossing it and therefore curves more tightly on the far side, which fixes the direction of travel and hence the sign of the charge; the design as the model for every later use of a degrader to break the sign–direction degeneracy.]*

95.3 Procedure

95.3.1 Expansion, triggering and photography

[Reserved: random expansion at fixed intervals in Anderson's setup, with roughly one usable cosmic-ray track in fifty photographs [713]; the counter-controlled trigger introduced by Blackett and Occhialini, in which coincident Geiger counters above and below fire the expansion so that most photographs contain a track [714] — an efficiency gain of well over an order of magnitude and the reason their shower photographs came so quickly; scanning and selection criteria, and the selection bias each introduces.]

95.3.2 Track reconstruction and momentum from curvature

[Reserved: reprojection of the stereoscopic pair, fitting a circle to the droplet trail, and the propagation of the sagitta uncertainty into $B\rho$ and thence into momentum, following the error-analysis conventions of Chapter 2; the dominant systematics — multiple scattering in the chamber gas, turbulence and droplet diffusion, distortion of the optics — and why they matter more for high momenta, where the track is nearly straight.]

95.4 Observations

The quantities below are what the chapter will tabulate, each with its uncertainty, in SI units.

95.4.1 The photograph of 2 August 1932

[Reserved: the single track that carries the claim [713]: $B\rho = 0.21\,\mathrm{T\,m}$ above the plate and $0.075\,\mathrm{T\,m}$ below, corresponding to momenta of $63\,\mathrm{MeV}/c$ and $23\,\mathrm{MeV}/c$; curvature sense opposite to that of an electron in the same field; visible track length over $5\,\mathrm{cm}$ beyond the plate; droplet density indistinguishable from that of a fast electron. The figure itself is to be re-authored as a labelled TikZ diagram of the track geometry, not reproduced as an image.]

Phenomenon 95.1 (The positive electron)**.** Cosmic radiation contains singly charged particles whose tracks curve in a magnetic field in the sense opposite to an electron's, whose droplet density along the path is that of a fast electron rather than that of a proton of the same curvature, and whose range exceeds by an order of magnitude the few millimetres such a proton could traverse. Their charge is therefore $+e$ and their mass is the electron's, not the proton's. The particle is stable in vacuum, and its mass and the magnitude of its charge equal the electron's to within the precision of present comparisons [2427].

Derivation. One relation turns the photograph into a measurement, and it is exact relativistically. A particle of charge q in a uniform field $\boldsymbol{B}$ obeys $\mathrm{d}\boldsymbol{p}/\mathrm{d}t = q\,\boldsymbol{v} \times \boldsymbol{B}$. The force is perpendicular to $\boldsymbol{v}$, so it does no work; the speed is constant, hence so is γ, hence so is the magnitude of $\boldsymbol{p} = \gamma m \boldsymbol{v}$. Motion in the plane normal to $\boldsymbol{B}$ is then uniform circular motion of some radius ρ, for which $|\mathrm{d}\boldsymbol{p}/\mathrm{d}t| = |\boldsymbol{p}|\,v/\rho$, while the right-hand side has magnitude $|q|\,vB$. Equating and cancelling v,

$$|\boldsymbol{p}| = |q|\;B\rho\,. \tag{95.1}$$

Equation (95.1) contains no reference to the mass, so a measured sagitta in a calibrated field yields a momentum whatever the particle turns out to be. Three further inferences follow, and between them they exhaust the alternatives.

First, the *sense* of the curvature gives the sign of q once the direction of travel is known, and the lead plate supplies that: a particle can only lose energy in crossing it, so by Equation (95.1) it must emerge with the smaller $B\rho$, and the side of the plate carrying the tighter curve is the far side. Second, the mass enters the *other* observables: at fixed momentum the velocity is $v/c = |\boldsymbol{p}|\,c/\sqrt{|\boldsymbol{p}|^2 c^2 + m^2 c^4}$, and ionization per unit path rises steeply as v falls. A proton carrying the measured momentum would be slow, would ionize far more densely than the track shows, and would come to rest within a few millimetres of chamber gas; a particle of electronic mass at the same momentum is ultrarelativistic, near minimum ionizing, and crosses the chamber. Third, an ordinary electron travelling upward would reproduce the curvature but not the energy loss across the plate, and is excluded by the same argument that fixed the direction. Charge $+e$ and mass m_e are what remains. □

95.4.2 The tabulated quantities

[Reserved: a booktabs table, set with `captionof` *inside the experiment box, listing for each of the fifteen positive tracks reported in [713]: momentum above and below the plate in* MeV/c*, inferred energy loss in the lead, specific ionization relative to a minimum-ionizing electron, and the resulting upper bound on the mass, quoted by Anderson as under twenty electron masses and tightened by later work to consistency with* m_e *[2427]; the modern reference values of* m_e *and* e *from CODATA [2431].]*

95.4.3 Uncertainties and systematics

[Reserved: the uncertainty budget — field calibration, sagitta measurement, energy loss in lead computed from the stopping-power formula, and the statistical weight of a single event; the reason the claim survived on one photograph, namely that the alternatives of Section 95.1.3 are excluded by qualitative observables that do not depend on the momentum precision; the contrast with modern discovery thresholds (Chapter 113), which this chapter should draw explicitly rather than leave implicit.]

95.5 Interpretation

95.5.1 Identification with Dirac's anti-electron

[Reserved: the inference from charge sign, mass bound and ionization to a positive particle of electronic mass [691, 713]; the identification with the hole of Chapter 93 [677], made by Blackett and Occhialini rather than by Anderson [714]; the methodological point that a confirmed prediction and an unmotivated discovery here coincide, and that the treatise records who claimed what.]

95.5.2 Independent confirmation and pair production

[Reserved: the counter-triggered showers of Blackett and Occhialini, showing positive and negative tracks in equal numbers and emerging from a common origin [714]; the interpretation as materialization of a photon into an electron–positron pair above the threshold $2m_ec^2 = 1.022\,\mathrm{MeV}$ *[726], with the cross section computed by Bethe and Heitler [732]; positrons from artificial radioactivity observed independently by Curie and Joliot [737], which made the particle available in the laboratory without cosmic rays and connects to the* β^+ *decays of Chapter 107.]*

Phenomenon 95.2 (Pair production)**.** A photon of sufficient energy passing through matter disappears, and an electron and a positron appear together at the same point; in showers the positive and negative tracks emerge from a common vertex in equal numbers. The process has a threshold — no pair appears below a photon energy of $2m_ec^2 = 1.022\,\mathrm{MeV}$ [2431] — and it requires the presence of a nucleus: a photon travelling in empty space never converts, however energetic.

Derivation. Both statements are kinematic and need no cross section. A single photon has $p_\gamma \cdot p_\gamma = 0$, whereas any electron–positron pair has $(p_+ + p_-)\cdot(p_+ + p_-) \geq (2m_ec)^2$, the minimum being reached when the two are at relative rest. The total four-momentum is conserved and its square is invariant, so $\gamma \to e^+e^-$ in vacuum would demand $0 \geq (2m_ec)^2$, which is false: a third body must be present to absorb momentum. Take that body to be a nucleus of mass M initially at rest. The invariant squared energy is

$$s = M^2c^4 + 2E_\gamma Mc^2\,,$$

and the lightest final state allowed by charge and lepton-number conservation is the nucleus together with the pair, all at relative rest, so that at threshold $\sqrt{s} = (M + 2m_e)\,c^2$. Expanding and cancelling M^2c^4,

$$2E_\gamma Mc^2 = 4Mm_ec^4 + 4m_e^2c^4 \qquad \Longrightarrow \qquad E_\gamma^{\mathrm{thr}} = 2m_ec^2\left(1 + \frac{m_e}{M}\right). \tag{95.2}$$

For any nucleus m_e/M is below 10^{-4}, so the threshold is $2m_ec^2$ to far better than a cloud chamber can measure; the correction is the recoil the nucleus must take, and it is the same term that vanishes only in the unphysical limit $M \to \infty$. Finally, the equality of the numbers of positive and negative tracks is charge conservation applied to a neutral initial state. □

95.5.3 Annihilation radiation

[Reserved: Klemperer's measurement of the radiation emitted when positrons stop in matter [741] — two photons of 511 keV *emitted back to back, as energy and momentum conservation require for a pair annihilating at rest; the process as the exact inverse of pair production and as the first direct evidence that matter and antimatter destroy each other; the astrophysical counterpart, the* 511 keV *line from the Galactic centre mapped by INTEGRAL [2079]; the cross section belongs to Chapter 100.]*

Phenomenon 95.3 (Annihilation radiation)**.** Positrons brought to rest in matter vanish, and in their place two photons appear, each of energy $m_ec^2 = 511.0\,\mathrm{keV}$ [2431] and travelling in opposite directions. Matter and antimatter are not merely oppositely charged: brought together they are destroyed entirely, their whole rest energy converted into radiation. No single photon is ever observed from such an event.

Derivation. Take the pair at rest, so that the initial four-momentum is $(2m_ec, \mathbf{0})$. If two photons are produced, conservation of momentum makes their three-momenta sum to zero, so $\boldsymbol{k}_2 = -\boldsymbol{k}_1$; both being massless their energies are then equal, and conservation of energy gives $E_1 = E_2 = m_ec^2$, the two emerging back to back. Exactly one photon is impossible in any frame: a single photon has $p\cdot p = 0$ while the initial state has $p\cdot p = (2m_ec)^2 \neq 0$, and no change of frame alters an invariant — the same obstruction, read backwards, that forbids a free photon from converting in Phenomenon 95.2. Three or more photons are kinematically allowed and do occur, at a rate suppressed by a further power of the fine-structure constant; which channel is permitted at all is decided not by kinematics but by the charge-conjugation quantum number of the annihilating state, a selection rule belonging to Chapter 105. □

95.5.4 The muon: the same apparatus, a different particle

[Reserved: the honest sequel, in which Anderson's own chamber produced tracks too penetrating for electrons and too light for protons, announced as a particle of intermediate mass [797] and long mistaken for Yukawa's nuclear-force meson; the muon's actual identity as a heavy lepton (Chapter 115); included here because it shows the same apparatus and the same style of argument yielding one correct identification and one wrong one, and because it disciplines the reading of Section 95.5.1.]

Phenomenon 95.4 (A second charged lepton)**.** The chambers that showed the positron also showed singly charged particles that penetrate far more matter than an electron of the same momentum and ionize far less than a proton of the same momentum: particles of intermediate mass. They are the muons — charge $\pm e$, spin $\frac{1}{2}$, mass $105.658\,\mathrm{MeV}/c^2$, some 207 times the electron's, and mean life about $2.2\,\mu\mathrm{s}$ [2427]. They are not the quantum of the nuclear force, which had been predicted with roughly that mass; they behave in every respect as a heavy copy of the electron, and nothing in this part of the treatise explains why a second copy should exist.

Derivation pending #120

The identification of the penetrating cosmic-ray component as a heavy lepton rather than as the quantum of the nuclear force: its anomalously weak absorption in nuclei against the strong coupling a nuclear-force carrier must have, and the three-body shape of its decay electron spectrum. The muon's own phenomenology is taken up in the cosmic-ray chapter of Part XII.

95.6 Modern repetitions and precision

95.6.1 Positronium

[Reserved: Deutsch's observation of the bound electron–positron system in gases [980]; the two ground states — the singlet decaying to two photons in about 125 ps *and the triplet to three photons in about* 142 ns *[2427] — as a direct test of the annihilation selection rules and of charge-conjugation symmetry (Chapter 105); positronium as a purely leptonic "atom" whose spectrum tests the bound-state theory of Chapter 81 with no nuclear structure at all.]*

95.6.2 The antiproton and antihydrogen

[Reserved: the antiproton produced at the Bevatron and identified by simultaneous momentum and velocity measurement [1057], closing the argument that antimatter is general rather than a peculiarity of the electron; antihydrogen trapped at CERN and its $1S$–$2S$ *transition measured and found consistent with hydrogen [2280], the sharpest laboratory test of CPT available (Chapter 105); the direct observation that antihydrogen falls downwards [2417], a test of the equivalence principle of Chapter 42 on antimatter.]*

95.6.3 Precision comparisons of positron and electron

[Reserved: the Penning-trap comparison of the g factors of the electron and the positron, agreeing to about two parts in 10^{12} *[1798], which is simultaneously the tightest test of CPT for leptons and a measurement of the anomaly discussed in Chapter 112; the charge-to-mass ratio comparisons; how each such null result is quoted as a bound rather than as a confirmation, per the conventions of Chapter 2.]*

Phenomenon 95.5 (Particle and antiparticle properties are equal and opposite)**.** Every comparison so far made between a particle and its antiparticle has found their masses equal, their lifetimes equal, their magnetic moments equal in magnitude and opposite in sign, and their charges equal in magnitude and opposite in sign. For the electron and the positron the agreement of the g factors and of the charge-to-mass ratios is among the sharpest such tests available, and no significant difference has been reported in any of them [2427].

Derivation pending #121

The CPT theorem — that any local, Lorentz-invariant quantum field theory with a Hermitian Hamiltonian and the normal spin–statistics connection is invariant under the combined operation of charge conjugation, parity and time reversal — together with the equalities of mass, lifetime and magnetic moment that follow from it, and the reasons a null result here bounds rather than confirms.

95.6.4 Applied annihilation: positron emission tomography

[Reserved: the imaging technique built directly on Section 95.5.3 — a β^+ emitter administered to a patient, the two $511\,\mathrm{keV}$ photons detected in coincidence, and the line of response reconstructed tomographically, as realized in the first transaxial positron tomographs [1557, 1563]; the back-to-back emission geometry, whose small departure from π from residual momentum sets a floor on spatial resolution; included because it is the plainest available demonstration that a 1932 cosmic-ray track is now routine engineering.]

Chapter 96

The Weyl Equation and Neutrinos

The Dirac equation of Chapter 93 is reducible: for a massless particle its four components split into two independent two-component equations, one for each chirality. That split is Weyl's equation [641], and its history is the cleanest case in the book of a mathematically impeccable theory being rejected on a physical prejudice and then reinstated by experiment. Weyl's equation violates parity manifestly, so for a quarter of a century it was dismissed as unphysical; when parity turned out not to be a symmetry of the weak interaction [1091, 1133] the same equation became, within months, the accepted description of the neutrino [1118, 1121, 1130], and its central prediction — that the neutrino has one definite helicity — was measured directly a year later [1148]. This chapter therefore sits after Chapters 93 and 94, which supply the spinor representations it decomposes, and immediately before the experiment that tests it, Chapter 98; the parity measurements themselves are Chapter 110.

The chapter closes on the one place where the Weyl picture is now known to be incomplete. Neutrino flavour oscillations [1954, 2021] require at least two nonzero neutrino masses, and a massive fermion cannot be described by a single Weyl field: the mass must be either of Dirac type, pairing two Weyl fields, or of Majorana type [796], pairing a Weyl field with itself and violating lepton number. Which of the two Nature chose is still an open experimental question, and the searches that would settle it have so far returned null results [2350, 2414]. The phenomenology of the massive case belongs to Chapter 103 and its evidence to Chapter 114; what is settled here is the massless two-component theory, its chirality structure, and the exact point at which it fails. Standing references are [1912, 2427].

Derivation pending #122

The Weyl Equation and Neutrinos: all derivations of this chapter are pending.

96.1 The two-component equation

96.1.1 Weyl's equation

[Reserved: the massless two-component equation $\mathrm{i}\sigma^{\mu}\partial_{\mu}\psi_{L}=0$ *with* $\sigma^{\mu}=(\mathbb{1},-\boldsymbol{\sigma})$*, written down by Weyl in the first of his papers on the electron and gravitation [641]; its plane-wave solutions and the dispersion* $E=|\boldsymbol{p}|\,c$*; the conserved current* $j^{\mu}=\psi_{L}^{\dagger}\sigma^{\mu}\psi_{L}$ *and its positive-definite density; why the equation is first order yet has only two components, and why no Lorentz-invariant mass term can be built from* ψ_{L} *alone.]*

96.1.2 Two-component spinors and $\mathrm{SL}(2,\mathbb{C})$

[Reserved: the undotted and dotted spinor calculus of van der Waerden [640]; $\mathrm{SL}(2,\mathbb{C})$ *as the double cover of the proper orthochronous Lorentz group and its two inequivalent*

fundamental representations $(\frac{1}{2},0)$ *and* $(0,\frac{1}{2})$*, per the classification of Chapter 94; the invariant antisymmetric* ϵ *tensor raising and lowering spinor indices; how a four-vector is a bispinor,* $p_\mu\sigma^\mu$*, with* $\det(p_\mu\sigma^\mu) = p^2$*.]*

96.1.3 Chirality and helicity

[Reserved: chirality as the eigenvalue of γ^5*, a Lorentz-invariant label attached to the representation; helicity as the projection* $\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}}$ *of spin on momentum, a Poincaré label in Wigner's classification [849] that is invariant only for a massless particle; the two coincide exactly at* $m = 0$ *and differ at order* m/E *otherwise — the single fact that governs everything else in this chapter and every helicity suppression in Chapter 101; the Weyl equation as the statement that helicity is fixed and the same for all observers.]*

96.1.4 Relation to the Dirac equation

[Reserved: the chiral (Weyl) basis of the Dirac matrices, in which the Dirac equation of [596] becomes two coupled two-component equations with the mass as the only coupling; the decomposition $\psi = \psi_L + \psi_R$ *with* $P_{L,R} = \frac{1}{2}(\mathbb{1} \mp \gamma^5)$*; the massless limit as a decoupling, so that a Dirac field is exactly two Weyl fields when* $m = 0$*; parity as the operation exchanging the two, developed in Chapter 105.]*

96.1.5 Discrete symmetries of a Weyl field

[Reserved: a single Weyl field is not invariant under parity or under charge conjugation separately, but is invariant under their product CP *and, by the theorem of Chapter 109, under* CPT*; the historical objection that this made the equation unphysical, voiced by Pauli and repeated for twenty-five years without ever being put to a test; how the objection was disposed of, not answered, by Chapter 110; the residual* CP *question deferred to Chapter 111.]*

96.2 The neutrino hypothesis

96.2.1 The continuous beta spectrum

[Reserved: Chadwick's magnetic-spectrometer measurement showing that beta rays emerge with a continuous distribution of energies rather than the line spectrum a two-body decay demands [414]; the calorimetric confirmation by Ellis and Wooster, whose measured heat per disintegration of radium E matched the mean and not the endpoint energy [574], closing the escape route that the missing energy was radiated later; the resulting choice between abandoning energy conservation and inventing a new particle.]

Phenomenon 96.1 (The beta spectrum is continuous)**.** The electrons emitted in a given nuclear beta decay do not all carry the same energy. They emerge with a continuous distribution of kinetic energies, running from zero up to a sharp endpoint fixed by the mass difference of the two nuclei [414], and the heat released per disintegration, measured calorimetrically, equals the *mean* of that distribution rather than its endpoint [574, 662] — so the balance is not carried off by radiation emitted later. A decay into two bodies cannot produce such a spectrum.

Derivation. Let a nucleus of mass M decay at rest into exactly two bodies, of masses M' and m. Four-momentum conservation gives $p_M = p_{M'} + p_m$, so $(p_M - p_m)^2 = p_{M'}^2$, that is

$$M^2c^2 - 2\,p_M\cdot p_m + m^2c^2 = M'^2c^2\,.$$

In the parent's rest frame $p_M = (Mc, \mathbf{0})$ and $p_M \cdot p_m = ME$, where E is the light particle's total energy, so

$$E = \frac{\left(M^2 + m^2 - M'^2\right) c^2}{2M} \,. \tag{96.1}$$

This is a single number, fixed entirely by three masses: a two-body decay emits a *line*, and repeating identical decays cannot broaden it beyond the natural width of the levels. If instead a third body leaves with the electron, the recoiling system is no longer a single particle of fixed mass; its invariant mass W may take any value between the sum of the two remaining rest masses and $Mc^2 - mc^2$, and Equation (96.1) with M' replaced by W/c^2 then sweeps out a continuum. The observed spectrum is therefore evidence that a third particle is emitted, and the endpoint — reached when that particle takes away nothing but its rest energy — bounds its mass. What the spectrum alone does not fix is the new particle's spin or its charge; those come from the angular-momentum balance of the decay and from the absence of any observed ionization, and are the content of the hypothesis of Section 96.2.2. □

96.2.2 Pauli's 1930 letter

[Reserved: Pauli's "desperate remedy" — the letter of 4 December 1930 to the Tübingen conference proposing a neutral, spin-$\frac{1}{2}$, weakly interacting constituent of the nucleus with a mass not exceeding that of the electron [665], so that beta decay is a three-body process and the spectrum is continuous; the spin-statistics argument for nitrogen-14 that motivated it as much as the energy balance did; Brown's historical reconstruction of the letter and its reception [1617]; Fermi's renaming of the particle after the discovery of the neutron.]

96.2.3 Fermi's theory of beta decay

[Reserved: Fermi's four-fermion contact interaction [738], built by analogy with the electromagnetic current coupling and giving the first quantitative beta spectrum shape, the Kurie plot and the ft systematics of Chapter 107; the coupling constant G_F, its dimension, and its modern value as tabulated in [2427]; the cross-section estimate that made the neutrino look undetectable, and the extension to the general Lorentz-invariant four-fermion couplings that Section 96.4.3 then reduces to one.]

96.2.4 Detection: Reines and Cowan

[Reserved: the inverse beta reaction $\bar{\nu}_e + p \to n + e^+$ as the detection channel; the 1953 Hanford attempt [1026] and the 1956 Savannah River confirmation with a delayed-coincidence signature of positron annihilation followed by cadmium neutron capture [1085, 1097]; the measured cross-section of order $1 \times 10^{-47}\,\mathrm{m}^2$ per fission antineutrino; Davis's null result for $\bar{\nu}_e + {}^{37}\mathrm{Cl}$ [1062] showing that neutrino and antineutrino are distinct, and lepton number as the bookkeeping [1022].]

96.2.5 How many neutrinos

[Reserved: the two-neutrino experiment at the AGS establishing that $\nu_\mu \neq \nu_e$ [1223]; direct observation of ν_τ interactions by DONUT [2016]; the invisible width of the Z measured at LEP and SLC, giving $N_\nu = 2.9840 \pm 0.0082$ light active species [2104] — the strongest statement that the count is three; flavour mixing itself left to Chapter 103.]

Phenomenon 96.2 (There are three light neutrino species)**.** The neutrinos produced together with the electron, with the muon and with the tau are three distinct particles, and there is no fourth of the same kind. The sharpest statement comes from the Z boson: the

part of its decay width that leaves no trace in any detector, divided by the width the theory assigns to a single neutrino species, gives three to better than one per cent [2427]. The count is a count of species light enough for the Z to decay into and coupling to it in the standard way; it excludes neither heavier neutral leptons nor species that do not couple to the Z at all.

Derivation pending #123

Extraction of the invisible width of the Z from the measured total and visible widths at the resonance peak, and the Standard Model expression for the partial width into one neutrino species, whose ratio is the species count. The measurement requires the electroweak theory of Part XII to convert a width into a number.

96.3 Parity and the reinstatement of the equation

96.3.1 Parity as an assumed symmetry

[Reserved: parity invariance as an inference from atomic and nuclear spectroscopy, where it holds, illegitimately extended to the weak interaction where it had never been tested; Lee and Yang's survey showing that no existing experiment constrained parity in weak processes, together with the concrete experiments that would [1091]; the θ–τ puzzle — two particles alike in mass and lifetime decaying to final states of opposite parity — as the anomaly that forced the question.]

96.3.2 The measurements

[Reserved: the beta asymmetry from polarized ^{60}Co [1133]; the pion–muon–electron chain, in which the muon is produced polarized and decays asymmetrically [1109, 1110]; the longitudinal polarization of beta electrons, measured as v/c with the sign fixed by Mott scattering [1108]; the circular polarization of the gamma rays following beta decay [1131]; full apparatus and uncertainties in Chapter 110.]

96.3.3 Maximal violation

[Reserved: why the observed asymmetries are not merely nonzero but as large as kinematics allows, i.e. the weak current couples to one chirality only; the immediate consequence that the two-component massless equation discarded in Section 96.1.5 is the correct one; the distinction between parity violation as an experimental fact and maximality as a further, separately measured fact, quantified by the Michel parameters of muon decay [965] and their modern values [2170].]

96.4 The two-component theory of the neutrino

96.4.1 Landau, Lee–Yang and Salam

[Reserved: the three independent 1957 papers proposing the massless two-component neutrino — Landau's, framed as the conservation of combined inversion [1118]; Lee and Yang's, which derives the observable consequences for beta decay and pion decay [1121]; and Salam's, which ties the two-component form to the vanishing of the neutrino mass and to γ^5 invariance [1130]; the shared prediction that the neutrino has a single definite helicity, and the sign left open by all three.]

96.4.2 The measured helicity

[Reserved: the Goldhaber–Grodzins–Sunyar measurement of the helicity of the neutrino emitted in K-capture by ^{152m}Eu, which found it negative — the neutrino is left-handed [1148]; the recoil-Doppler resonant-scattering trick and the analysing magnet, given in full in Chapter 98; the mirror statement for the antineutrino inferred from the $\bar{\nu}$ helicity in beta decay, and why no equally direct measurement of it exists.]

96.4.3 The V–A current

[Reserved: the reduction of Fermi's five possible couplings to the single combination $\gamma^{\mu}(\mathbb{1}-\gamma^{5})$ by Feynman and Gell-Mann [1144] and, independently, by Sudarshan and Marshak [1162]; the conserved vector current hypothesis and the universality of G_F; helicity suppression of $\pi^- \to e^-\bar{\nu}_e$ relative to $\pi^- \to \mu^-\bar{\nu}_\mu$ as the sharpest V–A test; the embedding of V–A in the $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ theory of Chapter 104.]

Phenomenon 96.3 (Helicity suppression of the electronic pion decay)**.** The charged pion decays overwhelmingly to a muon and only very rarely to an electron, although the electronic channel has by far the larger phase space available to it. The measured ratio is

$$\frac{\Gamma\left(\pi^- \to e^-\bar{\nu}_e\right)}{\Gamma\left(\pi^- \to \mu^-\bar{\nu}_\mu\right)} \approx 1.23 \times 10^{-4} \tag{96.2}$$

[2427]: the kinematically favoured channel is suppressed by four orders of magnitude, which is the opposite of what phase space alone predicts and is the sharpest available test of the chirality of the charged weak current.

Derivation. The pion has spin 0, so in its rest frame the charged lepton and the antineutrino leave back to back with total angular momentum zero and their spin projections on the decay axis must cancel. Because the two travel in opposite directions, cancelling projections means *equal* helicities. The charged weak current couples only to the left-chiral fields; for a massless antineutrino left chirality is helicity $+1$ exactly, so the charged lepton is forced into helicity $+1$ as well. But the current couples to the left-chiral part of the charged lepton too, and for a lepton of mass m_ℓ, energy E and momentum p the amplitude for the left-chiral field to be found with positive helicity, relative to its amplitude for the negative helicity, is

$$\sqrt{\frac{E - pc}{E + pc}} \simeq \frac{m_\ell c^2}{2E}\,,$$

using $E^2 - p^2c^2 = m_\ell^2 c^4$ and $E \gg m_\ell c^2$. The amplitude therefore vanishes linearly in m_ℓ and the rate quadratically, so the electronic channel is suppressed relative to the muonic one by roughly $(m_e/m_\mu)^2 \approx 2.3 \times 10^{-5}$, which the larger electronic phase space then partly offsets. Carrying the phase-space factors through gives the tree-level prediction

$$\frac{\Gamma\left(\pi \to e\nu\right)}{\Gamma\left(\pi \to \mu\nu\right)} = \frac{m_e^2}{m_\mu^2}\left(\frac{m_\pi^2 - m_e^2}{m_\pi^2 - m_\mu^2}\right)^2 \approx 1.28 \times 10^{-4}\,,$$

which differs from Equation (96.2) by a few per cent, the difference being radiative corrections. A parity-conserving interaction, coupling equally to both chiralities, would impose no such constraint and would leave the electronic mode the commoner of the two; the observed ratio therefore measures the chirality structure directly and does not merely permit it. □

96.5 Chirality in the Standard Model

96.5.1 Left-handed doublets, right-handed singlets

[Reserved: the Standard Model as a chiral gauge theory — the left-handed fields form $\mathrm{SU}(2)_L$ *doublets and the right-handed fields are singlets, so that the fermion content is naturally a list of Weyl fields rather than Dirac fields [1912]; the absence of any right-handed neutrino field in the minimal content, which is exactly why the minimal Standard Model predicts massless neutrinos; hypercharge assignments and the resulting electric charges, cross-referenced to Chapter 104.]*

96.5.2 The chiral anomaly

[Reserved: the axial current, conserved classically for massless fermions, is not conserved on quantization — the triangle anomaly of Adler [1388] and of Bell and Jackiw [1391]; the $\pi^0 \to \gamma\gamma$ *rate as its measured consequence [2427]; anomaly cancellation as a nontrivial constraint on the chiral fermion content of each Standard Model generation, and its failure mode if any charge assignment is altered; renormalization context in Chapters 100 and 108.]*

96.5.3 Chiral fermions on a lattice

[Reserved: the Nielsen–Ninomiya theorem, that a local, Hermitian, translation-invariant lattice action with the correct continuum limit necessarily doubles chiral fermions into pairs of opposite chirality [1673]; the homotopy argument behind it; why this is a genuine obstruction to simulating the weak interaction rather than a technical nuisance, in contrast with the vector-like case of Chapter 102; the same doubling reappearing as a theorem about band structures in Section 96.8.]

96.6 Mass, and where the massless picture fails

96.6.1 Dirac and Majorana mass terms

[Reserved: the two Lorentz-invariant bilinears that can give a Weyl field a mass — the Dirac term $m\bar{\psi}_L\psi_R$*, which requires a second, independent Weyl field and conserves lepton number, and the Majorana term* $\frac{1}{2}m\,\overline{\psi_L^c}\,\psi_L$*, which does not [796]; the general mass matrix mixing both; the statement that a massive neutrino cannot be a single Weyl particle, and that helicity therefore ceases to be Lorentz-invariant.]*

96.6.2 Oscillations as the evidence for mass

[Reserved: flavour oscillation, proposed by Pontecorvo [1127] and given its mixing-matrix form by Maki, Nakagawa and Sakata [1237]; the atmospheric ν_μ *deficit of Super-Kamiokande [1954] and the neutral-current measurement by SNO showing the solar* ν_e *flux is transformed and not lost [2021]; the measured splittings* $\Delta m^2_{21} \approx 7.5\times10^{-5}\,\mathrm{eV}^2$ *and* $|\Delta m^2_{31}| \approx 2.5\times10^{-3}\,\mathrm{eV}^2$ *[2427]; full treatment in Chapters 103 and 114.]*

Phenomenon 96.4 (Neutrinos change flavour, and therefore have mass)**.** A source of neutrinos of one flavour, allowed to propagate, delivers the other flavours as well. The atmospheric muon-neutrino flux is depleted by an amount depending on the path length through the Earth, and the solar electron-neutrino flux is not lost but converted, the total over all three flavours agreeing with what the Sun is calculated to produce. Fitting the dependence on distance and energy yields two independent mass-squared differences,

$$\Delta m^2_{21} \approx 7.5\times10^{-5}\,\mathrm{eV}^2/c^4\,, \qquad \left|\Delta m^2_{31}\right| \approx 2.5\times10^{-3}\,\mathrm{eV}^2/c^4\,, \tag{96.3}$$

both nonzero [2427]. At least two of the three neutrinos are therefore massive, and the massless two-component description of this chapter cannot be the whole of the story.

Derivation. Suppose the states of definite flavour are superpositions of states of definite mass; take two for simplicity, $|\nu_\alpha\rangle = \cos\theta\,|\nu_1\rangle + \sin\theta\,|\nu_2\rangle$ and $|\nu_\beta\rangle = -\sin\theta\,|\nu_1\rangle + \cos\theta\,|\nu_2\rangle$. A state of definite mass and momentum p propagates with the phase $\mathrm{e}^{-\mathrm{i}E_i t/\hbar}$, and the mass-shell relation expanded for $m_i c^2 \ll E$ gives $E_i \simeq pc + m_i^2 c^3/2p$. The common term pc contributes an overall phase and drops out; over a flight of length $L \simeq ct$ the relative phase is

$$\Delta\phi = \frac{(E_2 - E_1)\,t}{\hbar} \simeq \frac{\Delta m^2 c^4 L}{2\hbar c E}\,, \qquad \Delta m^2 := m_2^2 - m_1^2\,.$$

Projecting the propagated state back onto $|\nu_\beta\rangle$,

$$P\left(\nu_\alpha \to \nu_\beta\right) = \sin^2 2\theta\,\sin^2\left(\frac{\Delta m^2 c^4 L}{4\hbar c E}\right). \tag{96.4}$$

The probability oscillates in L/E with an amplitude set by the mixing angle and a period set by Δm^2, and it is identically zero when $\Delta m^2 = 0$, whatever the mixing. An observed oscillation is therefore a proof that the mass eigenvalues differ, and hence that at least one of them is nonzero. Two things follow that the chapter must not overstate. Equation (96.4) is insensitive to the common scale of the masses, which is why Equation (96.3) reports differences only; and it is insensitive to the sign of Δm^2 in vacuum, which is why the ordering of the states is fixed only by measurements involving matter. □

96.6.3 Absolute mass bounds

[Reserved: oscillations fix differences of squared masses and not the scale, so the scale needs independent measurements; the KATRIN tritium endpoint bound $m_\nu < 0.8\,\mathrm{eV}/c^2$ at 90 per cent confidence [2399]; the cosmological bound on $\sum m_\nu$ of order $0.12\,\mathrm{eV}/c^2$ from the CMB and large-scale structure [2348], with its model dependence stated honestly; the resulting hierarchy question, still open.]

Phenomenon 96.5 (The neutrino mass scale is very small)**.** Oscillations fix mass *differences*; the scale itself has never been measured, only bounded. The endpoint region of the tritium beta spectrum constrains the effective electron-neutrino mass to below about $1\,\mathrm{eV}/c^2$ [2427], and the imprint of free-streaming neutrinos on the cosmic microwave background and on the growth of large-scale structure constrains the sum of the three masses to of order $0.1\,\mathrm{eV}/c^2$ [2348] — the latter only within an assumed cosmological model, and so not a laboratory measurement. Either bound places the neutrino masses at least five orders of magnitude below the electron's, a gap that nothing in this part explains.

Derivation pending #124

The shape of the beta spectrum near its endpoint as a function of the neutrino mass, and the reason the observable is the incoherent sum of squared mixing amplitudes times squared masses rather than any single mass; and, separately, the free-streaming suppression of the matter power spectrum by massive neutrinos, whose derivation belongs to the cosmology part and whose model dependence must be stated with the bound.

96.6.4 Where the mass could come from

[Reserved: the dimension-five lepton-number-violating operator that generates a Majorana mass in the effective theory below any new scale [1646], and the fact that it is a parametrization of ignorance rather than a mechanism; the seesaw realization with heavy right-handed singlets [1605], named here with the honest statement that no direct evidence for it exists; the alternative that neutrinos are Dirac particles with very small Yukawa couplings, equally unconstrained; open-problem status in Chapter 133.]

96.7 The Majorana question

96.7.1 Majorana's symmetric theory

[Reserved: Majorana's real representation of the Dirac matrices and the self-conjugate field $\psi = \psi^c$ [796]; the degrees-of-freedom count — two, not four — and the equivalence of a Majorana field to one Weyl field with a mass; Racah's observation that the neutrino–antineutrino identity is testable in nuclear processes [799]; why no charged fermion can be Majorana, and why the neutrino is the only Standard Model candidate.]

96.7.2 Neutrinoless double beta decay

[Reserved: Furry's calculation of the lepton-number-violating transition $(A, Z) \to (A, Z + 2) + 2e^-$ [836]; the sharp two-electron sum-energy peak at the Q value distinguishing it from the allowed two-neutrino mode; the effective mass $m_{\beta\beta} = \left|\sum U_{ei}^2 m_i\right|$ and the nuclear matrix element as the dominant systematic; the Schechter–Valle argument that observing the decay implies a Majorana mass whatever the mechanism [1701].]

96.7.3 Null results and present status

[Reserved: the background-free ^{76}Ge result of GERDA and its half-life limit [2350]; the ^{136}Xe limit from KamLAND-Zen, $T^{0\nu}_{1/2} > 2 \times 10^{26}$ yr, entering the inverted-ordering band [2414]; the honest verdict — the Dirac versus Majorana question is open, the searches so far are null, and no result in this book depends on the answer; what a positive result would change, and what a continued null result can and cannot exclude.]

Phenomenon 96.6 (Neutrinoless double beta decay has not been seen)**.** Several nuclei that cannot decay by single beta emission do decay by emitting two electrons and two antineutrinos at once; this second-order weak process is observed, and the summed energy of its two electrons is continuous, as Phenomenon 96.1 requires of a decay with unobserved particles in the final state. The lepton-number-violating variant, in which only the two electrons are emitted and their summed energy is a sharp line at the full decay energy, has not been observed in any isotope; the best half-life limits exceed 10^{26} years [2427]. Whether the neutrino is its own antiparticle is therefore an open experimental question, and this treatise records it as one.

Derivation pending #125

The two-neutrino and neutrinoless double beta decay rates: the phase-space and nuclear-matrix-element factorization, the effective Majorana mass that the neutrinoless rate measures, and the argument that observing the decay implies a Majorana mass term whatever mechanism drives it. The nuclear matrix elements are the dominant systematic and must be presented as such.

96.8 Weyl fermions as an observed phenomenon

[Reserved: no elementary Weyl fermion is known, since every observed neutrino has mass; but the equation itself is realized as the low-energy excitation spectrum around isolated band-touching points in crystals without inversion symmetry — the Weyl semimetals, observed in TaAs by angle-resolved photoemission through their Fermi arcs [2257, 2262]; the arcs as the surface signature required by the doubling theorem [1673]; the chiral anomaly measured as negative longitudinal magnetoresistance; band-structure background in Chapter 126.]

Chapter 97

Higher-Spin Wave Equations

Wigner's classification [849] assigns a mass and a spin to every unitary irreducible representation of the Poincaré group, and Chapter 94 shows that the list is unbounded in spin. A relativistic wave equation is the covariant field-theoretic carrier of one entry in that list, and the covariant carrier is always larger than the entry: a symmetric tensor or tensor–spinor has far more components than the representation has states, so every higher-spin equation is a field equation plus a set of constraints that remove the surplus. This chapter follows that pattern up the spin ladder — Proca's spin-1 equation [785], the Rarita–Schwinger spin-$\frac{3}{2}$ equation [862], the Fierz–Pauli spin-2 equation [834], and the general constructions of Singh and Hagen [1534] and Fronsdal [1620, 1622] — and it follows the pathologies with equal care, because the constraints are what break first when the field is coupled to anything [1211, 1408, 1409]. It therefore sits after Chapters 92 and 93, which are the spin-0 and spin-$\frac{1}{2}$ members of the same family, and after Chapter 96, which decomposed the spin-$\frac{1}{2}$ case.

The honest experimental summary is short and is the reason the chapter ends where it does. Elementary massive spin-1 particles exist and are measured to high precision — the W and Z bosons discovered at the CERN $\mathrm{p\bar{p}}$ collider [1710, 1711, 1713, 1714], with the Proca mass supplied by the mechanism of Chapter 104. A massless spin-2 field is required by the observed gravitational radiation of Part V, treated in Chapter 46 and measured in Chapter 47, and its interacting completion is general relativity itself. Spin $\frac{3}{2}$ is realized only as a composite hadronic resonance, the $\Delta(1232)$ [994, 2427]. No elementary particle of spin greater than 2 has ever been observed, and Section 97.6 collects the theorems — Weinberg's soft theorems [1299], Weinberg–Witten [1661] and Coleman–Mandula [1343] — which say why that absence is expected rather than accidental. Speculative higher-spin programmes are named in Section 97.7.2 only to record that they have no evidential support, per the scope rule stated in the front matter.

Derivation pending #126

Higher-Spin Wave Equations: all derivations of this chapter are pending.

97.1 The classification problem

97.1.1 From representations to field equations

[Reserved: Wigner's little-group construction and the mass–spin labelling of unitary irreducible representations [849]; the Bargmann–Wigner equations, which build the spin-s representation as a totally symmetric product of $2s$ Dirac indices, each satisfying the Dirac equation [908]; why a massless representation has two helicity states rather than $2s+1$; the counting that every later section must reproduce, with the group theory in Chapter 14.]

97.1.2 What a wave equation must deliver

[Reserved: the four requirements — the correct on-shell degree of freedom count, a mass-shell condition $(\Box + m^2)\phi = 0$ on every component, positive-definite energy, and hyperbolic propagation inside the light cone; auxiliary conditions $\partial^\mu \phi_{\mu\nu\ldots} = 0$ and tracelessness as the mechanism that removes lower-spin content; the pattern, common to [785, 834, 862], that the constraints follow from the field equations only in the free theory — the observation Section 97.3.2 then turns into a no-go.]

97.2 Spin 1: the Proca field

97.2.1 The Proca equation

[Reserved: the Lagrangian $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}m^2 A_\mu A^\mu$ and the field equation $\partial_\mu F^{\mu\nu} + m^2 A^\nu = 0$ written down by Proca [785]; the divergence of the field equation giving the Lorenz condition $\partial_\mu A^\mu = 0$ as a consequence rather than a gauge choice, so that a Proca field carries three states; the propagator and its $k^\mu k^\nu / m^2$ term; the $m \to 0$ discontinuity in the state count against Maxwell theory (Chapter 60), the spin-1 ancestor of Section 97.4.2.]

97.2.2 The Duffin–Kemmer form

[Reserved: the first-order matrix formulation with β-matrices satisfying $\beta^\mu\beta^\nu\beta^\lambda + \beta^\lambda\beta^\nu\beta^\mu = \eta^{\mu\nu}\beta^\lambda + \eta^{\lambda\nu}\beta^\mu$, due to Duffin [814] and Kemmer [841]; its five-dimensional and ten-dimensional representations describing spin 0 and spin 1 in a single Dirac-like formalism; what is gained (Dirac-style manipulations, a clean current) and what is not (the algebra has no inverse element, so the analogy is incomplete).]

97.2.3 Stueckelberg's restoration of gauge invariance

[Reserved: the introduction of a compensating scalar $A_\mu \to A_\mu + m^{-1}\partial_\mu \varphi$ that makes the massive vector theory gauge invariant [821]; the consequences — a smooth massless limit, power counting improved to renormalizable, and the longitudinal mode exhibited as the would-be Goldstone boson; the relation to the Higgs mechanism [1275, 1283] taken up in Chapter 104; why this trick works for spin 1 and has no unproblematic analogue at spin 2.]

97.2.4 Massive vectors in Nature

[Reserved: the $W^\pm$ and Z^0 bosons as the observed elementary Proca fields — discovery in $\mathrm{p\bar{p}}$ collisions by UA1 [1710, 1711] and UA2 [1713, 1714], with the missing-energy and dilepton-mass signatures; the present masses $m_W = 80.37\,\mathrm{GeV}/c^2$ and $m_Z = 91.19\,\mathrm{GeV}/c^2$ [2427]; the photon as the massless limit, with the experimental bound on its mass; the ρ meson as a composite spin-1 state, not a fundamental Proca field.]

97.3 Spin 3/2: the Rarita–Schwinger field

97.3.1 The equation and its constraints

[Reserved: the vector–spinor ψ_μ and the field equation $(\mathrm{i}\gamma^\nu\partial_\nu - m)\psi_\mu = 0$ supplemented by $\gamma^\mu\psi_\mu = 0$ and $\partial^\mu\psi_\mu = 0$, formulated by Rarita and Schwinger [862]; the covariant Lagrangian with its $\gamma^{\mu\nu\rho}$ kinetic operator; the counting from sixteen components to four physical states; the massless limit and its gauge invariance $\psi_\mu \to \psi_\mu + \partial_\mu \epsilon$, stated here as a property of the free wave equation only.]

97.3.2 The Velo–Zwanziger causality problem

[Reserved: coupling ψ_μ minimally to an external electromagnetic field spoils the constraint algebra — the characteristic surfaces of the resulting system lie outside the light cone, so wave fronts propagate superluminally [1408, 1409]; the earlier indefinite-anticommutator inconsistency found by Johnson and Sudarshan [1211]; the general lesson that the constraints of a higher-spin equation are fragile under interaction, and the survey of which couplings evade the problem and at what cost.]

97.3.3 The $\Delta(1232)$ resonance

[Reserved: the first pion–nucleon resonance, found as a peak in the $\pi^+\mathrm{p}$ total cross-section by Anderson, Fermi, Long and Nagle [994]; its quantum numbers $J^P = \frac{3}{2}^+$, $I = \frac{3}{2}$, mass near $1232\,\mathrm{MeV}/c^2$ and width near $117\,\mathrm{MeV}$ [2427]; the point of principle that this is a composite three-quark state described by Chapter 102, so it is evidence for spin $\frac{3}{2}$ existing in Nature but not for an elementary Rarita–Schwinger field; higher-spin baryon and meson recurrences organized on Regge trajectories [1174, 1200].]

97.4 Spin 2: the Fierz–Pauli field

97.4.1 The Fierz–Pauli Lagrangian

[Reserved: the symmetric tensor $h_{\mu\nu}$ and the unique quadratic Lagrangian free of ghosts, with the relative coefficient of h^2 and $h_{\mu\nu}h^{\mu\nu}$ in the mass term fixed to -1 [834]; the constraints $\partial^\mu h_{\mu\nu} = 0$ and $h^\mu{}_\mu = 0$ reducing ten components to five; the massless case with its linearized diffeomorphism invariance $h_{\mu\nu} \to h_{\mu\nu} + \partial_{(\mu}\xi_{\nu)}$ and two helicity states, which is the linearized Einstein equation of Chapter 44.]

97.4.2 The vDVZ discontinuity

[Reserved: the massless limit of Fierz–Pauli theory does not reproduce linearized general relativity — the extra helicity-0 mode survives and couples to the trace of the source, changing the light-bending prediction relative to the Newtonian potential by a factor $\frac{3}{4}$, discovered independently by van Dam and Veltman [1419] and by Zakharov [1439]; why this is a discontinuity in the propagator and not in the spectrum alone; the measured light deflection of Chapter 42 as the datum it contradicts.]

97.4.3 The Boulware–Deser ghost and the Vainshtein radius

[Reserved: Vainshtein's observation that the linear analysis breaks down inside a radius r_V around a source, so the discontinuity may be an artefact of perturbation theory [1487]; Boulware and Deser's counter-result that generic nonlinear completions propagate a sixth, ghost-like mode [1468]; the ghost-free tuning of de Rham, Gabadadze and Tolley [2181] and the review of the subject [2194], named with the honest statement that no observation requires a graviton mass.]

97.4.4 The massless graviton and its evidence

[Reserved: the theorem that a self-coupled massless spin-2 field is general relativity — Gupta's iterative construction [1042] and Deser's one-step derivation from the first-order form [1420]; the evidence, which is the quadrupole energy loss of the binary pulsar [1550, 1703, 2275] and the two transverse-traceless polarizations of the detected waves [2264, 2278]; the graviton mass bound of order $1 \times 10^{-23}\,\mathrm{eV}/c^2$ from the arrival-time comparison of GW170817 with

its electromagnetic counterpart [2278]; full treatment in Part V, quantization status in Chapter 135.]

97.5 General spin

97.5.1 Massive fields of arbitrary spin

[Reserved: the Singh–Hagen Lagrangians, which achieve a consistent massive spin-s description only at the price of a tower of auxiliary traceless fields of ranks $s-2, s-3, \dots, 0$ [1534]; the pattern that the number of auxiliary fields grows with spin; the Fierz–Pauli conditions $\partial^{\mu}\phi_{\mu\dots} = 0$, $\phi^{\mu}{}_{\mu\dots} = 0$ as the general statement; why no such Lagrangian has ever been needed to describe an observed elementary particle.]

97.5.2 Massless fields of arbitrary spin

[Reserved: Fronsdal's gauge-invariant equations for massless integer spin, with doubly-traceless fields and traceless gauge parameters [1622], and the half-integer case of Fang and Fronsdal [1620]; the two helicity states $\pm s$ that survive; the massless limit of the massive theory as a decomposition into helicities, in contrast with the discontinuity of Section 97.4.2; the free theory as the last point at which higher spin is uncontroversial.]

97.5.3 Coupling obstructions

[Reserved: the systematic obstructions to interacting higher-spin fields — loss of the constraint algebra [1408, 1409], non-positive commutators [1211], and the failure of gauge invariance to survive minimal coupling to gravity; the contrast with spins $0, \tfrac{1}{2}, 1, \tfrac{3}{2}, 2$, each of which has a consistent interacting realization; the statement of the problem in the form the no-go theorems of Section 97.6 then answer.]

97.6 No-go theorems

97.6.1 Soft theorems and Weinberg's low-energy argument

[Reserved: Weinberg's covariant Feynman rules for any spin [1298]; the soft-emission analysis showing that Lorentz invariance of the S matrix forces a massless spin-1 particle to couple to a conserved charge and a massless spin-2 particle to couple universally with a single constant — the equivalence principle derived rather than assumed [1299]; the extension to infrared photons and gravitons [1322]; the conclusion that a massless particle of spin ≥ 3 cannot couple to any conserved quantity that survives at low energy.]

97.6.2 Weinberg–Witten

[Reserved: the theorem that a theory with a Lorentz-covariant, conserved current forbids massless particles of helicity $|h| > \tfrac{1}{2}$ carrying that charge, and that a Lorentz-covariant, conserved stress tensor forbids massless particles of helicity $|h| > 1$ [1661]; the precise reading of the hypotheses, without which the theorem appears to exclude the graviton and the gluon — neither has a gauge-invariant covariant current or stress tensor of the required kind; what the theorem does and does not exclude, stated carefully.]

97.6.3 Coleman–Mandula

[Reserved: the theorem that under an S-matrix with a mass gap, analyticity and a finite number of particle types below any mass, the largest symmetry Lie algebra is the direct

product of the Poincaré algebra with internal symmetries [1343]; the corollary that conserved charges carrying more than one spacetime index would force the scattering amplitude to be trivial; the standard graded-algebra escape route, named here only to record that it lies outside the scope of this book.]

97.7 Experimental status

97.7.1 The observed spectrum

[Reserved: the complete list of elementary particles and their spins as measured — spin 0 (the Higgs boson, see Chapter 113), spin $\frac{1}{2}$ (quarks and leptons), spin 1 (photon, gluons, W, Z), and spin 2 inferred for the graviton from the polarization content of gravitational waves [2264, 2278]; the tabulated composite states of higher spin [2427], all hadronic, all with Regge-organized recurrences [1174, 1200]; how a spin assignment is actually measured, from angular distributions and decay correlations.]

97.7.2 Programmes without evidence

[Reserved: an honest closing statement — no elementary particle of spin $\frac{3}{2}$ or of spin greater than 2 has been observed, and searches place limits but no signal [2427]; the theoretical programmes that postulate them, including graded extensions of the Poincaré algebra and infinite higher-spin towers, are named here solely to record that they have no observational support and are therefore outside the scope of this treatise; the open questions that remain within scope are collected in Chapters 133 and 135.]

Chapter 98

Experiment: Neutrino Helicity

The two-component theory of Chapter 96 predicts that the neutrino is produced in a single helicity state, and the three papers that proposed it [1118, 1121, 1130] left the sign open. Goldhaber, Grodzins and Sunyar measured that sign in 1958 [1148], in what remains one of the most economical experiments in physics: the neutrino is never detected at all. Its helicity is read off the circular polarization of a gamma ray emitted by the recoiling nucleus, using a kinematic accident of the ^{152m}Eu decay scheme that makes only those photons emitted antiparallel to the neutrino Doppler-shifted enough to scatter resonantly. Angular-momentum conservation in the K-capture then transfers the neutrino's helicity to the photon, and a magnetized-iron analyser measures the photon's circular polarization by transmission. The answer — negative helicity, a left-handed neutrino — fixed the sign in the V–A current within weeks [1144, 1162].

The chapter is an experiment chapter and, when written, will use the structured `experiment` environment with its `ExpApparatus`, `ExpProcedure`, `ExpObservations`, `ExpInterpretation` and `ExpReferences` fields, all quantities in SI with uncertainties. It sits immediately after the theory chapter it tests and is the companion of Chapter 110, which reports the parity-violation measurements that made the two-component theory admissible in the first place; the nuclear-structure background it relies on is Chapter 107.

Derivation pending #127

Experiment: Neutrino Helicity: all derivations of this chapter are pending.

98.1 Historical context and the prediction under test

98.1.1 From parity violation to the two-component neutrino

[Reserved: the sequence Lee–Yang [1091], the ^{60}Co asymmetry [1133] and the muon-decay asymmetry [1109, 1110], which together made parity violation a fact; the immediate revival of Weyl's equation [1118, 1121, 1130]; the prediction to be tested — that the neutrino has definite helicity ± 1, with the sign undetermined by theory — and the corresponding statement for the antineutrino; cross-reference to Chapters 96 and 110.]

98.1.2 Why direct neutrino polarimetry is impossible

[Reserved: the neutrino interaction cross-section, of order $1 \times 10^{-47}\,\mathrm{m}^2$ at MeV energies [1085], forbidding any analyser placed in the neutrino's path; the consequent strategy of inferring the helicity from a particle that is *detectable, by angular-momentum bookkeeping in a decay where the kinematics selects a single emission direction; the earlier neutrino-recoil*

experiments that proved the technique — Rodeback and Allen's measurement of the ^{37}Ar *recoil following K-capture [1009].]*

98.1.3 The chosen decay scheme

[Reserved: why ^{152m}Eu *is the unique practical choice — a* $J = 0$ *parent of half-life about* 9.3 h *capturing a K electron to a* $J = 1$ *excited state of* ^{152}Sm *that decays by a* 961 keV *gamma to a* $J = 0$ *ground state, so the whole chain is* $0 \to 1 \to 0$ *and the photon must carry the full unit of angular momentum; the two-body kinematics fixing the nuclear recoil momentum from a monoenergetic neutrino; the resulting Doppler shift, of a few electronvolts, comparable to the recoil energy loss on emission.]*

Phenomenon 98.1 (Nuclear resonance fluorescence is recoil-forbidden, and here recoil–compensated)**.** A gamma ray emitted by a free nucleus is not in general resonantly absorbed by a second nucleus of the same species: emission and absorption each cost a recoil energy, and the resulting mismatch is far larger than the width of the nuclear level, so the resonance is lost. In the decay chain used here it is restored — but only for photons emitted into the hemisphere towards which the emitting nucleus is already moving, having recoiled against the neutrino, that is, only for photons emitted antiparallel to the neutrino. Resonant scattering therefore selects a direction of emission, and the selection is what is observed: photons that have not been Doppler-compensated do not scatter resonantly at all [1148].

Derivation. Let the excited nucleus have mass M and level energy E_0. If it is at rest when it emits, energy and momentum conservation give the photon $E_\gamma = E_0 - E_\mathrm{R}$ with recoil energy $E_\mathrm{R} = E_0^2/2Mc^2$; to excite a second nucleus, itself initially at rest, a photon must instead supply $E_0 + E_\mathrm{R}$. The emitted photon is thus short of the absorption energy by

$$2E_\mathrm{R} = \frac{E_0^2}{Mc^2}\,, \tag{98.1}$$

and resonance fails whenever this exceeds the width of the level. For a transition of about 1 MeV in a nucleus of mass number 152, Equation (98.1) is a few electronvolts, orders of magnitude larger than a nuclear level width, so free-nucleus resonance fluorescence simply does not occur.

Now let the emitting nucleus be moving, with speed $v \ll c$, because it recoiled against a neutrino of energy E_ν, so that $v = E_\nu/Mc$. A photon emitted into the direction of that motion is Doppler-shifted upward by

$$\Delta E = \frac{v}{c}E_\gamma = \frac{E_\nu E_\gamma}{Mc^2}\,,$$

and comparison with Equation (98.1) shows the shortfall is made good when $E_\nu \gtrsim E_\gamma$. That is the kinematic accident on which the whole experiment rests: in this decay the neutrino and the photon carry nearly the same energy, and M cancels between the two expressions, so the compensation is a comparison of E_ν with E_γ and of nothing else. A photon emitted at any other angle receives a smaller shift, and one emitted backwards is shifted the wrong way, so only those emitted antiparallel to the neutrino can scatter resonantly. Selecting the resonantly scattered photons therefore selects that direction, with no angular measurement made anywhere in the apparatus. The selection is not perfect: the excited state has a finite lifetime, during which the recoiling nucleus may be deflected in the source, and this dilutes the correlation — it cannot reverse it, since the resonance condition still admits only the forward hemisphere. □

98.2 Apparatus

98.2.1 The source

[Reserved: the ^{152m}Eu source, produced by neutron activation of europium and used within a few half-lives; source strength, geometry and self-absorption; the source mounted at the top of the analysing magnet so that the photons of interest traverse the magnetized iron; the practical constraint that the short half-life imposes on counting time, and hence on the achievable statistical uncertainty.]

98.2.2 The analysing magnet

[Reserved: the magnetized iron block as a circular-polarization analyser — Compton scattering of circularly polarized photons on the aligned electron spins of the iron has a transmission that depends on the relative sign of photon helicity and electron polarization; the magnitude of the effect, a few per cent, and why the measurement is therefore a difference of two large counting rates; the solenoid and its field reversal; the same polarimetric method used for beta-decay gamma rays by Schopper [1131].]

98.2.3 Resonant scatterer and detector

[Reserved: the ring of ^{152}Sm (as $\mathrm{Sm_2O_3}$) placed to scatter transmitted photons into the detector; nuclear resonance fluorescence as the selector, possible only for photons whose Doppler shift compensates the recoil loss, i.e. only for photons emitted opposite to the neutrino; the sodium-iodide scintillation counter and its shielding; the remark that this is recoil-compensated resonance and not the recoil-free absorption that Mössbauer reported the same year [1154].]

98.3 Procedure

98.3.1 Selecting the recoil-aligned photons

[Reserved: the argument, to be given step by step, that resonant scattering selects photons emitted antiparallel to the neutrino; the angular-momentum balance $J_\gamma = -J_\nu$ along that axis for a $0 \to 1 \to 0$ chain, so that the photon helicity equals the neutrino helicity; the finite lifetime of the intermediate state and the resulting partial loss of the recoil direction, which dilutes but does not reverse the correlation; quantitative estimate of that dilution.]

98.3.2 Field reversal and counting

[Reserved: alternating the magnet polarity at fixed geometry so that all geometric efficiencies cancel in the ratio; the counting asymmetry $\delta = (N_+ - N_-)/(N_+ + N_-)$ as the measured quantity; run lengths set by the source half-life; background subtraction using a non-resonant scatterer of the same electron density; the calibration of the analysing power with gamma rays of known circular polarization.]

98.4 Observations

98.4.1 The counting asymmetry

[Reserved: the measured asymmetry, of order one per cent, with its statistical uncertainty; conversion to the circular polarization of the resonantly scattered photons, found to be negative and of magnitude about two thirds of the kinematic maximum, quoted with an

absolute uncertainty of roughly ten per cent [1148]; the resulting neutrino helicity, consistent with -1 and excluding $+1$ by a wide margin; all quantities to be tabulated with SI uncertainties in a booktabs table carried by `captionof` *inside the experiment box, since a float cannot live there.]*

98.4.2 Systematic checks

[Reserved: the checks performed — resonance verified by comparing scatterers, the polarization sign calibrated against a source of known handedness, magnet-reversal cycling against drift, and the insensitivity of the result to detector threshold; the dominant systematic, which is the analysing power of the magnetized iron; the honest statement of what the total uncertainty was and why the result was nonetheless decisive — the two hypotheses differ in sign, not in magnitude.]

98.5 Interpretation

98.5.1 The neutrino is left-handed

[Reserved: the conclusion that the neutrino emitted in K-capture has negative helicity [1148], fixing the sign left open by [1118, 1121, 1130]; the mirror statement for the antineutrino, inferred from beta-decay electron polarization [1108] together with CPT; the caveat, made explicit, that a helicity eigenstate is Lorentz-invariant only for a massless particle, so the result is a statement about the chirality of the current at the energies used.]

Phenomenon 98.2 (The neutrino is left-handed)**.** The neutrino emitted in the K-capture of ^{152m}Eu has negative helicity: its spin points opposite to its momentum. The measurement is wholly indirect — no neutrino is detected — and reads the helicity off the circular polarization of the 961 keV gamma ray emitted by the recoiling ^{152}Sm nucleus, which is found to be left circularly polarized [1148]. The magnitude is consistent with the maximal value -1 and the opposite sign is excluded by a wide margin, so the neutrino is produced in a single helicity state, as a two-component theory requires and as no parity-conserving theory permits.

Derivation. The inference is angular-momentum bookkeeping along a single axis, and it assumes no dynamics at all. Take the axis along the neutrino's momentum, call that direction $+z$, and use the decay scheme of Section 98.1.3: a $J = 0$ parent captures a K electron of spin projection $m_e = \pm\frac{1}{2}$ and emits a neutrino of helicity $h_\nu = \pm 1$, leaving a $J = 1$ nucleus of projection m_N recoiling along $-z$. Conservation of J_z gives

$$m_e = \tfrac{1}{2} h_\nu + m_N \,. \tag{98.2}$$

That nucleus then decays to a $J = 0$ ground state, so the photon must carry away the whole of m_N; and Phenomenon 98.1 says that the photons reaching the detector are those emitted along $-z$. A photon travelling along $-z$ with angular-momentum projection m_γ on $+z$ has helicity $h_\gamma = -m_\gamma$, and a real photon has no projection zero along its own direction of motion, so $m_\gamma = m_N = \pm 1$. With $m_e = \pm\frac{1}{2}$ the only solutions of Equation (98.2) are

$$h_\nu = -1, \quad m_e = +\tfrac{1}{2}, \quad m_N = +1 \qquad \text{and} \qquad h_\nu = +1, \quad m_e = -\tfrac{1}{2}, \quad m_N = -1 \,,$$

the remaining combinations forcing $m_N = 0$, which the photon cannot carry, or $|h_\nu| = 3$, which no spin-$\frac{1}{2}$ particle has. In both surviving cases $h_\gamma = -m_N = h_\nu$: the photon inherits the neutrino's helicity exactly. A measured negative photon helicity therefore reads $h_\nu = -1$ directly, using only angular-momentum conservation, the spins of the three nuclear levels, and the transversality of the photon. The dilution noted at the end of Phenomenon 98.1 reduces the observed magnitude below unity but cannot change its sign, which is why a measurement of modest precision settles a question whose two answers differ by a sign. □

98.5.2 Consequences for V–A

[Reserved: how the measured sign selects $\gamma^\mu(\mathbb{1}-\gamma^5)$ from the five Fermi couplings, in combination with the beta-decay correlation data; the V–A theory of Feynman and Gell-Mann [1144] and of Sudarshan and Marshak [1162], published in the same year; the conserved vector current hypothesis; helicity suppression of $\pi \to e\nu$ as the independent V–A prediction, and its route into Chapters 101 and 104.]

98.5.3 What the result does not say

[Reserved: the result constrains the helicity of the emitted state, not the existence of a right-handed neutrino field of very small coupling; it says nothing about neutrino mass, and it is fully compatible with the massive neutrinos established by Chapter 114; nor does it distinguish Dirac from Majorana neutrinos; a short list of the questions it left open, each pointing to where the book takes them up.]

98.6 Modern repetitions and precision

98.6.1 Chirality tests in leptonic decays

[Reserved: the Michel parameters of muon decay [965] as the modern parametrization of the leptonic current, measured to the part-in-10^3 level by TWIST [2170]; the ratio $\Gamma(\pi \to e\nu)/\Gamma(\pi \to \mu\nu)$, whose measurement by PIENU [2249] agrees with the helicity-suppressed V–A prediction at the 10^{-3} level; the honest note that no repetition of the direct helicity measurement itself has improved on [1148].]

Phenomenon 98.3 (Muon decay has the shape a left-handed current predicts)**.** The energy and angular distribution of the electrons emitted by polarized muons is described completely by four dimensionless parameters. Measurement returns the values that a purely left-handed charged current requires — $\rho = \delta = \frac{3}{4}$ and $\xi = 1$ — at the level of one part in 10^3 or better [2427]. The leptonic weak current is therefore chiral to that precision, in a purely leptonic process with no nuclear structure anywhere in it, and at an energy release two orders of magnitude above that of the K-capture measured in this chapter.

Derivation pending #128

The Michel spectrum: the most general Lorentz-invariant four-fermion matrix element for muon decay, its reduction to four measurable parameters in the decay distribution, and the values those parameters take when the current is the purely left-handed one. Deriving it also shows which combinations of couplings the experiment cannot separate, which is what makes the bounds on right-handed admixtures the shape they are.

98.6.2 Parity violation in electron scattering and atoms

[Reserved: the SLAC E122 measurement of a parity-violating asymmetry in deep-inelastic scattering of polarized electrons [1625], which fixed the weak mixing angle and confirmed the chiral structure of the neutral current; atomic parity non-conservation in caesium and the nuclear anapole moment [1946]; the proton weak charge measured by Qweak [2301]; how each of these tests the same chirality assignment at a different energy scale.]

Phenomenon 98.4 (Parity is violated in the neutral current too)**.** Parity violation is not confined to the charged-current processes of beta decay. Longitudinally polarized electrons scattered from nuclei show a cross-section asymmetry between the two beam helicities; heavy

atoms show optical transitions that parity conservation would forbid, driven by the weak electron–nucleus interaction; and the proton's weak charge has been measured directly. All of these are described by one dimensionless parameter, the weak mixing angle, whose value is close to $\sin^2\theta_W = 0.231$ in the conventional renormalization scheme [2427], and the same value accounts for measurements spanning many decades of momentum transfer.

Derivation pending #129

The parity-violating asymmetry in the scattering of longitudinally polarized electrons, as the interference between the photon-exchange and neutral-current amplitudes, and its expression in terms of the weak mixing angle; together with the analogous expression for the parity-non-conserving amplitude in a heavy atom, where the nuclear weak charge and the atomic structure factorize. Both belong to the electroweak chapter of Part XII.

98.6.3 Limits on right-handed currents

[Reserved: the general framework in which a small admixture of a right-handed charged current is allowed and bounded; the constraints from beta-decay correlation coefficients, from muon decay [2170], and from direct collider searches for a heavy right-handed vector boson [2427]; the current bounds, quoted as limits on the mixing angle and on the mass of any additional charged vector boson; the standing conclusion that the weak current is purely left-handed to within present sensitivity.]

Phenomenon 98.5 (No right-handed charged current has been found)**.** Every measurement sensitive to a right-handed admixture in the charged weak current returns a null result: the beta-decay correlation coefficients, the muon-decay parameters of Phenomenon 98.3, and the direct searches for a heavy charged vector boson coupling to right-handed fields. The outcomes are quoted as limits — on a mixing angle between the two currents, and on the mass of any additional charged vector boson — and no measurement has required a nonzero right-handed component [2427]. The charged current is purely left-handed to within present sensitivity, which makes the Weyl structure a measured fact and not merely the simplest option compatible with the data.

Derivation pending #130

The general left–right parametrization of the charged current, the observables in beta and muon decay that are first order in the right-handed admixture as against those that are second order, and the translation of a null asymmetry into a bound on the mixing angle and on the mass of an additional charged vector boson. The distinction between the two orders is what determines how the sensitivity scales, and it must be stated with the bounds.

Part XI

Quantum Field Theory and the Standard Model

Chapter 99

Canonical Quantization of Fields

This chapter quantizes a classical field the way Chapter 25 quantized a mechanical system: a field is a mechanical system with one degree of freedom at every point of space, so the canonical procedure — conjugate momenta, equal-time brackets promoted to operators — applies unchanged, and the free field turns out to be an infinite collection of the harmonic oscillators of Chapter 28. Everything particle-like is then a statement about those oscillators: a particle is a quantum of excitation, the number of particles is an eigenvalue, and processes that change that number are describable at all only because the field, not the particle, is the primitive object. That is the reason this chapter opens the part rather than Chapter 92 or Chapter 93, whose single-particle relativistic wave equations both fail — negative energies, negative probabilities, the Klein paradox — for exactly the want of a field. The classical input is the Lagrangian field theory of Chapter 64 and the Poincaré representation theory of Chapter 94; the output is Fock space, propagators, microcausality and the spin–statistics connection, on which Chapter 100 and everything after it is built.

The free-field results below are exact and among the very few in this part that are. Interactions are treated perturbatively from Chapter 100 onward, and the path-integral alternative — indispensable for the non-abelian theories of Chapter 102 — is developed in Chapter 106; the axiomatic reformulation that makes the causality statements below precise is Chapter 109. Standard modern treatments are [1654, 1907, 1912].

Notation 99.1 (Conventions of this chapter). The conventions are those fixed once for the whole part in Section 100.1.1 and are repeated here only in the two places where a scalar field needs them. The metric is $\eta_{\mu\nu} = \mathrm{diag}(+1,-1,-1,-1)$; coordinates are $x^\mu = (ct, \boldsymbol{x})$, so that $\mathrm{d}^4x = c\,\mathrm{d}t\,\mathrm{d}^3x$ carries m^4 and $\partial_\mu = (c^{-1}\partial_t, \nabla)$ carries $1/\mathrm{m}$; the d'Alembertian is

$$\Box := \partial_\mu \partial^\mu = \frac{1}{c^2}\partial_t^2 - \nabla^2 \,, \qquad [\Box] = 1/\mathrm{m}^2 \,. \tag{99.1}$$

Wave four-vectors $k^\mu = (\omega/c, \boldsymbol{k})$ carry $1/\mathrm{m}$ and four-momenta $p^\mu = \hbar k^\mu$ carry $\mathrm{kg\,m/s}$, with the mass shell $p^2 = m^2c^2$. A mass m enters every field equation below only through the inverse length

$$\kappa := \frac{mc}{\hbar} \,, \qquad [\kappa] = 1/\mathrm{m} \,, \tag{99.2}$$

the reciprocal of the reduced Compton wavelength; the mass shell reads $k^2 = \kappa^2$. The letter κ never means anything else in this chapter, and μ never means a mass: it occurs only as a Lorentz index and in μ_0, the magnetic constant. As in Notation 100.1, ϵ standing alone is the positive infinitesimal of the Feynman contour, ε_0 the electric constant, and $\boldsymbol{\varepsilon}_\lambda$ a polarization vector. An action is $S = c^{-1}\int \mathrm{d}^4x\,\mathcal{L}$ with $\mathcal{L}$ an energy density, $\mathrm{J/m^3}$; the SI dimension of each field is stated where the field is introduced. Every equation in this chapter, without exception, is written with $\hbar$ and c in place.

Remark 99.2 (Reading this chapter against its sources). The literature cited below works in units in which $\hbar = c = 1$, so that κ and m are the same symbol and a propagator is written $1/(k^2 - m^2)$. The dictionary is the one given in Remark 100.4, applied here to the two objects this chapter introduces: a mass becomes an inverse length by Equation (99.2), and the scalar propagator $1/(k^2 - m^2 + \mathrm{i}\epsilon)$ of those texts is $\mathrm{i}\hbar c/(k^2 - \kappa^2 + \mathrm{i}\epsilon)$ in Equation (99.93) below. No derivation in this chapter is performed in those units; the remark exists so that the reader can check the chapter against its sources, which is a different thing.

99.1 Why fields

Quantum field theory is not a generalization undertaken for its own sake. It is forced, and this section is the argument that forces it. The three steps are independent and each is fatal on its own: a relativistic one-particle wave equation has no consistent probability interpretation; a theory with a fixed number of particles contradicts what is seen in every accelerator and every cloud chamber; and the uncertainty principle forbids the localization on which a one-particle position operator depends. What survives all three is a system with infinitely many degrees of freedom whose excitations are counted, not tracked.

99.1.1 Failure of the single-particle relativistic theories

The Schrödinger equation $\mathrm{i}\hbar\,\partial_t\psi = -(\hbar^2/2m)\nabla^2\psi$ is first order in time and second order in space, so it cannot be covariant under a transformation that mixes the two. The obvious repair is to quantize the relativistic mass shell $E^2 = \boldsymbol{p}^2c^2 + m^2c^4$ instead of the Newtonian $E = \boldsymbol{p}^2/2m$, substituting $E \to \mathrm{i}\hbar\partial_t$ and $\boldsymbol{p} \to -\mathrm{i}\hbar\nabla$. The result is the Klein–Gordon equation [544],

$$-\hbar^2\partial_t^2\psi = -\hbar^2c^2\nabla^2\psi + m^2c^4\psi\,, \qquad \text{that is} \qquad \left(\Box + \kappa^2\right)\psi = 0\,, \tag{99.3}$$

studied as a wave equation in Chapter 92. It is manifestly covariant, and it is manifestly useless as a one-particle theory, for two reasons that turn out to be one reason.

Negative energies. The plane wave $\psi = N\exp[-\mathrm{i}(Et - \boldsymbol{p}\cdot\boldsymbol{x})/\hbar]$ solves Equation (99.3) whenever $E = \pm\sqrt{\boldsymbol{p}^2c^2 + m^2c^4}$, and the negative branch cannot be thrown away. Equation (99.3) is second order in time, so the initial data are ψ and $\partial_t\psi$, freely specifiable; the positive-frequency solutions alone are not a complete set and cannot match arbitrary initial data. Nor is the restriction stable: any interaction that is not diagonal in energy couples the two branches, which is what the Klein paradox [629] exhibits explicitly for a step potential of height greater than $2mc^2$.

Non-positive density. Multiplying Equation (99.3) by ψ^*, subtracting the conjugate equation multiplied by ψ, and dividing by $2\mathrm{i}mc^2/\hbar$ gives a continuity equation $\partial_t\rho + \nabla\cdot\boldsymbol{j} = 0$ with

$$\rho = \frac{\mathrm{i}\hbar}{2mc^2}\left(\psi^*\partial_t\psi - \psi\,\partial_t\psi^*\right)\,, \qquad \boldsymbol{j} = \frac{\hbar}{2\mathrm{i}m}\left(\psi^*\nabla\psi - \psi\nabla\psi^*\right)\,. \tag{99.4}$$

Evaluated on the plane wave above, $\partial_t\psi = -(\mathrm{i}E/\hbar)\psi$ and

$$\rho = \frac{\mathrm{i}\hbar}{2mc^2}\left(-\frac{2\mathrm{i}E}{\hbar}\right)|N|^2 = \frac{E}{mc^2}\,|N|^2\,, \tag{99.5}$$

which is negative on the negative branch. Born's reading of $|\psi|^2$ as a probability density [534] therefore cannot be transferred to Equation (99.4): a probability density is non-negative by definition, and this one is not.

Dirac's response was to insist on an equation first order in time, so that the density would be $\psi^\dagger\psi \geq 0$ by construction [596]. That succeeded — Chapter 93 follows the construction — and it did not remove the negative energies, which are a consequence of the mass shell and not of the order of the equation. The Dirac Hamiltonian has spectrum $\pm\sqrt{\boldsymbol{p}^2c^2 + m^2c^4}$

and is therefore unbounded below: an atomic electron could cascade down the negative continuum without limit, radiating an unbounded amount of energy, and no state of the theory would be stable.

The hole theory [647] repairs this by fiat. Declare that in the vacuum every negative-energy level is occupied; the exclusion principle then forbids the cascade, and the absence of an electron from the filled sea behaves as a particle of opposite charge. What that particle's mass is the 1930 paper does not settle: Dirac there identified the hole with the proton [647], and only in 1931 [677] concluded that it must carry the electron's own mass and so be a new particle unknown to experiment. That prediction, and its confirmation by Anderson [713] (Chapter 95), is one of the great successes of theoretical physics. It is nevertheless untenable as a foundation, for three reasons.

1. It is unavailable for bosons. The sea is held down by the exclusion principle, and a spin-0 particle obeys no such principle; there is no way to fill the negative-energy states of Equation (99.3). Yet the observed spin-0 particles — the charged pions above all — do have antiparticles: π^+ and π^- are listed as a particle–antiparticle pair of equal mass by the Particle Data Group [2427].

2. The vacuum acquires infinite charge and energy density, removed by a subtraction whose only justification is that it is needed.

3. It is a one-particle theory only in name. A state of the sea plus one electron is a state of infinitely many particles; the formalism has already left the single-particle Hilbert space, without admitting it.

The consistent version of the repair was found by Pauli and Weisskopf [748]: reinterpret Equation (99.3) not as an equation for a wave function but as an equation for a classical field, and quantize the field. Then ρ of Equation (99.4) is not a probability density at all but a charge density, whose sign is the sign of the charge, and the energy comes out positive for both signs. That construction is Section 99.3 below, and it works for spin 0 and spin 1/2 alike, with no sea.

The decisive objection is not internal but experimental. A one-particle theory has a Hilbert space $L^2(\mathbb{R}^3)$ (tensored with a finite-dimensional spin space), on which the number of particles is the constant 1; no unitary evolution on that space can produce a two-particle state. Nature produces them routinely.

Phenomenon 99.3 (Electron–positron pair creation)**.** A photon traversing matter converts into an electron–positron pair in the Coulomb field of a nucleus, provided its energy exceeds

$$E_\gamma^{\text{thr}} = 2m_e c^2 \left(1 + \frac{m_e}{M}\right) \longrightarrow 1.0220\,\text{MeV} \quad \text{as} \quad M \to \infty\,, \tag{99.6}$$

where M is the nuclear mass. Pair production is the dominant photon-absorption process in lead above about 5 MeV, and the positrons so produced are the ones Anderson identified [713].

Rests on Definition 40.4 and Theorem 40.5.

Derivation. *Derives Phenomenon 99.3.*

Energy and momentum conservation are the statement that the total four-momentum before and after are equal, so the invariant $s = (p_\gamma + P)^2$ is the same before and after. Before: the photon has $p_\gamma^2 = 0$ and the nucleus, at rest, has $P^\mu = (Mc, \mathbf{0})$, so

$$s = p_\gamma^2 + P^2 + 2p_\gamma \cdot P = M^2c^2 + 2\,\frac{E_\gamma}{c}\,Mc = M^2c^2 + 2ME_\gamma\,. \tag{99.7}$$

After: three particles, and the smallest value s can take for a given set of masses is the one in which they are all at relative rest, $s \geq (M + 2m_e)^2 c^2$. Combining,

$$2ME_\gamma \geq 4Mm_e c^2 + 4m_e^2 c^2 \quad \Longleftrightarrow \quad E_\gamma \geq 2m_e c^2 \left(1 + \frac{m_e}{M}\right) . \tag{99.8}$$

With the CODATA 2022 electron rest energy $m_e c^2 = 0.510\,998\,950\,69\,\mathrm{MeV}$ [2431] the infinite-mass limit is $1.021\,997\,90\,\mathrm{MeV}$. The recoil term m_e/M is why a nucleus must be present at all: for a photon alone $s = 0$, and Equation (99.8) can never be satisfied, so pair creation in empty space is kinematically forbidden however energetic the photon. □

A theory in which the number of particles is a fixed integer cannot describe Phenomenon 99.3, and no reinterpretation of the wave function repairs that. The state before the conversion has one quantum, the state after has three; the two live in different sectors, and only a formalism whose state space contains both, with operators connecting them, can hold the process at all.

The last objection is that even the one-particle sector is not what a single-particle theory needs it to be, because the position of a relativistic particle is not measurable to arbitrary precision without creating others.

Suppose a particle of mass m is confined to a region of linear size Δx. The uncertainty principle then requires $\Delta p \gtrsim \hbar/(2\Delta x)$, and for a relativistic particle $\Delta E \approx c\,\Delta p \gtrsim \hbar c/(2\Delta x)$. The energy uncertainty reaches the threshold $2mc^2$ of Equation (99.6) when

$$\Delta x \lesssim \frac{\hbar}{4mc} = \frac{1}{4\kappa} , \tag{99.9}$$

that is, at a fraction of the reduced Compton wavelength $\kappa^{-1} = \hbar/mc$. For the electron the Compton wavelength is $\lambda_C = h/(m_e c) = 2.426\,310\,235\,38 \times 10^{-12}\,\mathrm{m}$ and its reduced form is

$$\kappa_e^{-1} = \frac{\hbar}{m_e c} = \frac{\lambda_C}{2\pi} = 3.861\,592\,67 \times 10^{-13}\,\mathrm{m} , \tag{99.10}$$

both from CODATA 2022 [2431]. Attempting to pin an electron down to a tenth of that supplies enough energy to make several more, and the apparatus that would do the pinning — a potential well of depth $\gtrsim mc^2$ — is precisely a device for creating pairs. There is therefore no operational meaning to "the position of the electron" below κ_e^{-1}, and any formalism that carries an exact position operator is claiming more than can be measured.

Newton and Wigner [942] asked how much of a position operator survives. Demanding only that the localized states at a fixed time transform correctly under rotations, translations and space inversion, and that they be orthogonal for distinct points, they showed that a unique such operator exists on the positive-energy one-particle space — but that its eigenstates are not localized in any Lorentz-invariant sense. A Newton–Wigner state localized at a point in one frame has support spread over a region of order κ^{-1} in another, and its wave function has tails falling as $\mathrm{e}^{-\kappa r}$ rather than vanishing outside a bounded set. The one-particle position operator exists; what does not exist is a frame-independent notion of where the particle is, to better than κ^{-1}.

Each objection is really the same one seen from a different side. The mass shell has two branches; the second branch cannot be removed covariantly; a theory that keeps it needs states of arbitrarily negative energy unless the number of quanta is allowed to change; and the scale at which the change becomes unavoidable is κ^{-1}, which is also the scale at which localization fails. The resolution is to stop treating the wave equation as an equation for a probability amplitude and to treat it as the classical equation of motion of a field, to be quantized. Then the two branches become creation and annihilation of two species of quantum, the energy is bounded below, the conserved quantity of Equation (99.4) is a charge that may take either sign, and states with any number of quanta live in one space. Everything after this section is the execution of that programme.

99.1.2 Field as a mechanical system with continuously many degrees of freedom

A field is not a new kind of object requiring new mechanics. It is an ordinary mechanical system with an infinite number of coordinates labelled by a continuous index, and the cleanest way to see this is to watch the index become continuous.

Take N equal masses m_0 threaded on a light string of tension T at spacing ℓ, with transverse displacements $q_n(t)$ and fixed ends, the discrete system of Chapter 28. For small displacements the string segment between neighbours pulls with the transverse component $T(q_{n+1}-q_n)/\ell$, so

$$m_0 \ddot{q}_n = \frac{T}{\ell}\,(q_{n+1} - 2q_n + q_{n-1})\,, \tag{99.11}$$

which follows from the Lagrangian

$$L = \sum_n \left[\frac{m_0}{2}\dot{q}_n^2 - \frac{T}{2\ell}\,(q_{n+1}-q_n)^2 \right]. \tag{99.12}$$

Now let $\ell \to 0$ at fixed linear density $\varrho = m_0/\ell$ and fixed T. Write $q_n(t) = u(t, x_n)$ with $x_n = n\ell$; then $(q_{n+1}-q_n)/\ell \to \partial_x u$ and $\sum_n \to \ell^{-1}\int \mathrm{d}x$, so

$$L \longrightarrow \int \mathrm{d}x \left[\frac{\varrho}{2}\,(\partial_t u)^2 - \frac{T}{2}\,(\partial_x u)^2 \right], \tag{99.13}$$

whose Euler–Lagrange equation is the wave equation $\partial_t^2 u = v^2 \partial_x^2 u$ with $v = \sqrt{T/\varrho}$. Nothing conceptual happened in the limit. The discrete label n became the continuous label x; the finitely many coordinates q_n became the one-parameter family $u(x)$; and the Lagrangian, previously a sum, became an integral of a *density*. The coordinate of the system is now the whole function $u(\cdot)$ at one instant, and x is not a dynamical variable but a name for which coordinate one is talking about.

Definition 99.4 (Field, Lagrangian density, conjugate momentum)**.** A (real, scalar) classical field is a map $\phi : \mathbb{R}^4 \to \mathbb{R}$, $x^\mu \mapsto \phi(x)$, regarded as the continuously indexed set of coordinates $\{\phi(t, \boldsymbol{x})\}$ labelled by $\boldsymbol{x}$. A *local* Lagrangian is the integral of a density depending on the field and its first derivatives at one point,

$$L[\phi, \dot{\phi}] = \int \mathrm{d}^3x\, \mathcal{L}(\phi(x), \partial_\mu \phi(x))\,, \qquad [\mathcal{L}] = \mathrm{J/m^3}\,, \tag{99.14}$$

and the momentum density conjugate to ϕ is

$$\pi(x) := \frac{\partial \mathcal{L}}{\partial \dot{\phi}(x)}\,, \qquad \dot{\phi} := \partial_t \phi\,. \tag{99.15}$$

The Euler–Lagrange equations follow from stationarity of $S = c^{-1}\int \mathrm{d}^4x\,\mathcal{L}$ under variations vanishing on the boundary of the region of integration. Expanding $\delta\mathcal{L} = (\partial\mathcal{L}/\partial\phi)\,\delta\phi + (\partial\mathcal{L}/\partial(\partial_\mu\phi))\partial_\mu\delta\phi$ and integrating the second term by parts leaves, once the boundary term is discarded,

$$\delta S = \frac{1}{c}\int \mathrm{d}^4x \left[\frac{\partial \mathcal{L}}{\partial \phi} - \partial_\mu \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \right] \delta\phi\,,$$

which vanishes for arbitrary $\delta\phi$ only if the bracket does. For a local density the equations therefore read

$$\partial_\mu \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} - \frac{\partial \mathcal{L}}{\partial \phi} = 0\,. \tag{99.16}$$

The derivative of an integral with respect to one of continuously many coordinates is a functional derivative, defined by

$$\delta F[\phi] = \int \mathrm{d}^3x \, \frac{\delta F}{\delta \phi(\boldsymbol{x})} \, \delta\phi(\boldsymbol{x}) \,, \qquad \frac{\delta \phi(\boldsymbol{x})}{\delta \phi(\boldsymbol{y})} = \delta^3(\boldsymbol{x}-\boldsymbol{y}) \,, \tag{99.17}$$

the continuum replacement of $\partial q_n / \partial q_m = \delta_{nm}$; the Dirac delta carries $1/\mathrm{m}^3$, which is where the volume factors in every formula below come from.

The relativistic field whose equation of motion is Equation (99.3) is obtained from

$$\mathcal{L} = \frac{1}{2}\left(\partial_\mu \phi \, \partial^\mu \phi - \kappa^2 \phi^2\right) = \frac{1}{2}\left[\frac{1}{c^2}\left(\partial_t \phi\right)^2 - \left(\nabla\phi\right)^2 - \kappa^2\phi^2\right] , \tag{99.18}$$

for which Equation (99.16) gives $\left(\Box + \kappa^2\right)\phi = 0$ directly. Since $[\partial_\mu] = 1/\mathrm{m}$ and $[\mathcal{L}] = \mathrm{J}/\mathrm{m}^3$, the field itself carries

$$[\phi] = (\mathrm{J}/\mathrm{m})^{1/2} = \mathrm{kg}^{1/2}\,\mathrm{m}^{1/2}/\mathrm{s} \,, \tag{99.19}$$

and by Equation (99.15)

$$\pi = \frac{1}{c^2}\partial_t \phi \,, \qquad [\pi] = \mathrm{kg}^{1/2}/\mathrm{m}^{1.5} \,, \tag{99.20}$$

so that $[\phi][\pi] = \mathrm{J\,s}/\mathrm{m}^3$ — an action per unit volume, which is what the equal-time commutator of Section 99.2.1 will need it to be. Legendre-transforming,

$$\mathcal{H} = \pi\dot{\phi} - \mathcal{L} = \frac{1}{2}\left[c^2\pi^2 + \left(\nabla\phi\right)^2 + \kappa^2\phi^2\right] , \qquad H = \int \mathrm{d}^3x \, \mathcal{H} \,, \tag{99.21}$$

manifestly non-negative — a first sign that the field reading cures the disease of Section 99.1.1, since the same mass shell that gave energies of both signs gives a Hamiltonian density that is a sum of squares.

Classically the field obeys Poisson brackets that are the continuum form of $\{q_n, p_m\} = \delta_{nm}$,

$$\{\phi(t,\boldsymbol{x}), \pi(t,\boldsymbol{y})\} = \delta^3(\boldsymbol{x}-\boldsymbol{y}) \,, \qquad \{\phi,\phi\} = \{\pi,\pi\} = 0 \,, \tag{99.22}$$

obtained by putting the functional derivative Equation (99.17) in place of $\partial q_n/\partial q_m = \delta_{nm}$ in the definition of the bracket. Chapter 64 is the general classical field theory built on these brackets.

Equation (99.14) contains a physical hypothesis that is easy to overlook because it is built into the notation: the Lagrangian density at x depends on the field at x and on its first derivatives there, and on nothing else. A term such as $\int \mathrm{d}^3y \, K(\boldsymbol{x}-\boldsymbol{y})\phi(\boldsymbol{x})\phi(\boldsymbol{y})$ with K of finite range would be perfectly well defined and would still give a linear equation of motion; it is excluded by assumption.

The assumption is not free of consequences. Locality plus a second-order-in-time equation makes the field equation hyperbolic with characteristic speed c, so classical disturbances propagate on and inside the light cone; after quantization it is what makes the commutator of the field with itself vanish at spacelike separation, the statement verified explicitly in Section 99.7.1. A nonlocal kernel K of range R would leave a non-vanishing commutator out to spacelike separations of order R, and thus a signal outside the light cone. The bound on such effects is therefore an experimental bound, and the axiomatic programme of Chapter 109 takes locality — not the Lagrangian — as the primitive assumption for exactly this reason.

99.1.3 The historical sequence

The formalism of this chapter was assembled between 1925 and 1932, in a sequence in which almost every step was taken for a reason different from the one that now justifies it.

The first field ever quantized was the free radiation field, in the concluding chapter of the three-author paper of Born, Heisenberg and Jordan of 1926. Their object was not particles: it was to show that the new matrix mechanics reproduced Einstein's 1909 formula for the energy fluctuations of black-body radiation, whose two terms had for fifteen years looked like an unexplained superposition of a wave term and a particle term. Treating each cavity mode as a matrix-mechanical oscillator produced both terms at once. In the same year Born's collision papers [534] fixed the probability interpretation of the wave amplitude that Section 99.1.1 showed cannot survive the relativistic generalization — the two ideas were contemporaneous, and it took another eight years to see that they were in tension.

Dirac's radiation theory of 1927 [572] is the paper in which the method became a method. He quantized the electromagnetic field in Coulomb gauge, obtained the ladder operators for each mode, and computed the emission and absorption rates of an atom coupled to it. The matrix element for emission into a mode already containing n quanta came out proportional to $\sqrt{n+1}$ and that for absorption to $\sqrt{n}$; the ratio of the resulting rates reproduces Einstein's A and B coefficients, with the $+1$ — spontaneous emission — coming from the mode's zero-point term and not from any external perturbation. The consequences for quantum optics are Chapter 122.

Jordan and Wigner [609] then showed that the same apparatus describes fermions if the commutators are replaced by anticommutators, deriving the exclusion principle rather than postulating it. Heisenberg and Pauli [626, 656] gave the general canonical formalism for an arbitrary system of interacting fields; in doing so they met the difficulty that the momentum conjugate to the time component of the electromagnetic potential vanishes identically, the first appearance of the constrained Hamiltonian structure taken up systematically in Chapter 26. Fock [698] supplied the state space in which all of this acts, and Pauli and Weisskopf [748] closed the argument of Section 99.1.1 by quantizing the scalar field and showing that spin 0 has antiparticles without any sea.

Two later entries in Table 99.1 belong to the same story. Wigner's classification of the unitary irreducible representations of the Poincaré group [849] supplied the definition of "elementary particle" that this chapter uses — an irreducible representation labelled by the two invariants of Chapter 94 — and Casimir's calculation of the force between conducting plates [909], undertaken to explain the stability of colloidal suspensions, turned the zero-point energy from an embarrassment into a measurable quantity (Section 99.8).

99.2 The free real scalar field

Everything the rest of the chapter does is done first, and in full, for the simplest field there is: one real number at each point of spacetime, obeying Equation (99.3). It has no spin, no charge and no gauge freedom, so nothing distracts from the two operations being performed — imposing canonical commutators, and diagonalizing the Hamiltonian. The result is that the field is a continuous family of harmonic oscillators and that its excitations are counted by integers.

99.2.1 Canonical commutation relations

The classical system is Equation (99.18) with the conjugate momentum Equation (99.20) and the Poisson brackets Equation (99.22). Quantization is the same step taken in Chapter 25 for a system of finitely many degrees of freedom (Postulate 25.27): replace the classical variables

Table 99.1: The construction of the canonical formalism. The entries are the steps used in this chapter; the physical motivation column records what the authors were after, which was rarely what the step is now used for.

Year	Step	Motivation at the time
1926	Radiation field quantized mode by mode (Born, Heisenberg and Jordan)	Reproduce Einstein's black-body fluctuation formula
1927	Ladder operators, emission and absorption [572]	Derive the Einstein A and B coefficients
1928	Anticommuting field [609]	Derive the exclusion principle from an algebra
1928	First-order wave equation [596]	Obtain a positive probability density
1929	General canonical formalism [626, 656]	Quantize interacting fields covariantly
1930	Hole theory, with the hole taken to be the proton [647]	Stabilize the negative-energy continuum
1931	The hole given the electron's own mass [677]	Remove the proton identification; predict a new particle
1932	Fock space [698]	A state space for variable particle number
1934	Charged scalar field [748]	Show spin 0 needs no sea
1939	Irreducible representations of the Poincaré group [849]	Define "particle" invariantly
1948	Boundary-dependent vacuum energy [909]	Explain colloidal stability
1950	Wick's theorem [974]	Organize the perturbation series

by operators on a Hilbert space and the Poisson bracket by $(\mathrm{i}\hbar)^{-1}$ times the commutator. The continuous label rides along unchanged.

Postulate 99.5 (Canonical quantization of a scalar field)**.** The field ϕ and its conjugate momentum density π become operator-valued distributions on a Hilbert space, satisfying at equal times

$$[\phi(t,\boldsymbol{x}),\pi(t,\boldsymbol{y})] = \mathrm{i}\hbar\,\delta^3(\boldsymbol{x}-\boldsymbol{y})\,, \tag{99.23}$$

$$[\phi(t,\boldsymbol{x}),\phi(t,\boldsymbol{y})] = [\pi(t,\boldsymbol{x}),\pi(t,\boldsymbol{y})] = 0\,, \tag{99.24}$$

and the operators evolve in the Heisenberg picture, $\mathrm{i}\hbar\,\partial_t A = [A,H]$ with H of Equation (99.21).

Both sides of Equation (99.23) carry $\mathrm{J\,s/m^3}$, by Equations (99.19) and (99.20) on the left and because δ^3 carries $1/\mathrm{m}^3$ on the right. The $\hbar$ is not decoration and cannot be scaled away: it fixes the magnitude of the field fluctuations, and the classical field theory of Chapter 64 is recovered as $\hbar \to 0$, where the right-hand side vanishes and the field components commute.

Equation (99.24) deserves a comment. It says that the field at two different points of space *at the same time* commutes. That is not an extra assumption but the continuum version of $[q_n, q_m] = 0$: the two are independent coordinates of one system, and specifying both simultaneously is exactly what a configuration is. The nontrivial statement — that $[\phi(x),\phi(y)]$ vanishes for *all* spacelike separations, not merely equal-time ones — is a theorem, proved in Section 99.7.1.

Proposition 99.6 (The Heisenberg equations reproduce the field equation)**.** *With H of Equation (99.21) and the brackets Equations (99.23) and (99.24),*

$$\partial_t\phi = c^2\pi\,, \qquad \partial_t\pi = \nabla^2\phi - \kappa^2\phi\,, \tag{99.25}$$

and eliminating π gives $(\Box + \kappa^2)\,\phi = 0$.

Rests on Equations (99.21), (99.23) and (99.24).

Proof. *Derives Proposition 99.6.*

Only the term quadratic in π fails to commute with ϕ. Using $[A,B^2] = B\,[A,B] + [A,B]\,B$ and Equation (99.23),

$$[\phi(\boldsymbol{x}),H] = \frac{c^2}{2}\int \mathrm{d}^3y\,\Big[\phi(\boldsymbol{x}),\pi(\boldsymbol{y})^2\Big] = \frac{c^2}{2}\int \mathrm{d}^3y\,2\mathrm{i}\hbar\,\delta^3(\boldsymbol{x}-\boldsymbol{y})\pi(\boldsymbol{y}) = \mathrm{i}\hbar c^2\pi(\boldsymbol{x})\,,$$

which is the first of Equation (99.25). For the second, the gradient and mass terms contribute. The mass term gives $\frac{1}{2}\kappa^2\int \mathrm{d}^3y\;[\pi(\boldsymbol{x}),\phi(\boldsymbol{y})^2] = -\mathrm{i}\hbar\kappa^2\phi(\boldsymbol{x})$, and the gradient term gives

$$\frac{1}{2}\int \mathrm{d}^3y\;[\pi(\boldsymbol{x}),\nabla_y\phi\cdot\nabla_y\phi] = \int \mathrm{d}^3y\,\nabla_y\phi(\boldsymbol{y})\cdot\Big(-\mathrm{i}\hbar\nabla_y\delta^3(\boldsymbol{x}-\boldsymbol{y})\Big) = \mathrm{i}\hbar\,\nabla^2\phi(\boldsymbol{x})\,,$$

the last step an integration by parts, the surface term vanishing for fields falling off at infinity. Adding and dividing by $\mathrm{i}\hbar$ gives the second of Equation (99.25); substituting $\pi = c^{-2}\partial_t\phi$ gives $c^{-2}\partial_t^2\phi - \nabla^2\phi + \kappa^2\phi = 0$. □

Proposition 99.6 is the consistency check that licenses everything else: the quantized field obeys, as an operator equation, exactly the classical equation of motion, so the classical solution theory — plane waves, superposition, the mass shell — can be carried over intact and used to diagonalize H.

One technical point must be stated now rather than discovered later. Equation (99.23) contains a Dirac delta, and a delta is a distribution, not a function. Correspondingly $\phi(x)$ is not an operator: the object $\phi(x)\,|0\rangle$ has infinite norm, since Equation (99.91) below gives

$$\langle 0|\,\phi(x)^2\,|0\rangle = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\frac{\hbar c^2}{2\omega_{\boldsymbol{k}}} = \frac{\hbar c}{4\pi^2}\int_0^\infty \frac{k^2\,\mathrm{d}k}{\sqrt{k^2+\kappa^2}}\,, \tag{99.26}$$

which diverges quadratically at large k. What *is* an operator is the smeared field

$$\phi(f) = \int \mathrm{d}^4x\, f(x)\,\phi(x)\,, \tag{99.27}$$

with f a smooth test function of compact support. Physically this is not a technicality but a statement about measurement: a field cannot be measured at a mathematical point, only averaged over a region, which is precisely the conclusion Bohr and Rosenfeld reached from the side of the apparatus (Section 99.7.3). It is because the smeared objects, and not the pointwise ones, are the well-defined quantities that the axiomatic reformulation of Chapter 109 takes fields to be operator-valued distributions from the start [1099, 1297].

99.2.2 Mode expansion, creation and annihilation operators

The field equation is linear, so it is solved by superposing plane waves; the equal-time commutators then determine the algebra of the coefficients, and the coefficients turn out to be ladder operators.

Write the general real solution of $(\Box + \kappa^2)\,\phi = 0$ as a Fourier integral over the mass shell. With $k\cdot x := k_\mu x^\mu = \omega_{\boldsymbol{k}} t - \boldsymbol{k}\cdot\boldsymbol{x}$ and

$$\omega_{\boldsymbol{k}} = c\sqrt{\boldsymbol{k}^2 + \kappa^2} = \frac{1}{\hbar}\sqrt{\hbar^2 c^2 \boldsymbol{k}^2 + m^2 c^4}\,, \tag{99.28}$$

each Fourier component $\widetilde{\phi}(t,\boldsymbol{k})$ obeys $\ddot{\widetilde{\phi}} = -\omega_{\boldsymbol{k}}^2\widetilde{\phi}$: one harmonic oscillator per wave vector, of frequency $\omega_{\boldsymbol{k}}$. Reality of ϕ relates the components at $\boldsymbol{k}$ and $-\boldsymbol{k}$, so the independent data are one complex amplitude per $\boldsymbol{k}$, written $a(\boldsymbol{k})$.

Definition 99.7 (Mode expansion of the free real scalar field)**.**

$$\phi(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\sqrt{\frac{\hbar c^2}{2\omega_{\boldsymbol{k}}}}\left[a(\boldsymbol{k})\,\mathrm{e}^{-\mathrm{i}k\cdot x} + a^\dagger(\boldsymbol{k})\,\mathrm{e}^{\mathrm{i}k\cdot x}\right], \tag{99.29}$$

$$\pi(x) = -\mathrm{i}\int \frac{\mathrm{d}^3k}{(2\pi)^3}\sqrt{\frac{\hbar\omega_{\boldsymbol{k}}}{2c^2}}\left[a(\boldsymbol{k})\,\mathrm{e}^{-\mathrm{i}k\cdot x} - a^\dagger(\boldsymbol{k})\,\mathrm{e}^{\mathrm{i}k\cdot x}\right]. \tag{99.30}$$

Equation (99.30) is not an independent assumption: it is $c^{-2}\partial_t$ of Equation (99.29). The normalization factor is fixed by the requirement that Equation (99.23) hold, as Proposition 99.8 shows, and it is worth reading before the proof. The amplitude of each mode carries $\sqrt{\hbar}$: as $\hbar \to 0$ the quantum field flattens to zero and only the classical solution built from a macroscopic occupation survives. That $\sqrt{\hbar}$ is the entire quantum content of the free theory.

Dimensionally, $[\mathrm{d}^3k] = 1/\mathrm{m}^3$ and $[\sqrt{\hbar c^2/2\omega}] = \mathrm{J}^{1/2}\,\mathrm{m}$, so $[a] = \mathrm{m}^{3/2}$ if $[\phi]$ is to be Equation (99.19) — which is exactly the dimension demanded by Equation (99.31) below, since $\delta^3(\boldsymbol{k}-\boldsymbol{k}')$ carries m^3.

Proposition 99.8 (The ladder algebra)**.** *Equations* (99.23) *and* (99.24) *hold if and only if*

$$\left[a(\boldsymbol{k}), a^\dagger(\boldsymbol{k}')\right] = (2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}')\,, \qquad [a(\boldsymbol{k}), a(\boldsymbol{k}')] = \left[a^\dagger(\boldsymbol{k}), a^\dagger(\boldsymbol{k}')\right] = 0\,. \tag{99.31}$$

Rests on Equations (99.23) *and* (99.24) *and Definition 99.7.*

Proof. *Derives Proposition 99.8.*

Invert Equations (99.29) and (99.30) at a fixed time. Writing $\alpha_{\boldsymbol{k}} = \sqrt{2\omega_{\boldsymbol{k}}/(\hbar c^2)}$ and $\beta_{\boldsymbol{k}} = \sqrt{2c^2/(\hbar\omega_{\boldsymbol{k}})}$, the combinations

$$a(\boldsymbol{k}) = \frac{\mathrm{e}^{\mathrm{i}\omega_{\boldsymbol{k}}t}}{2}\int \mathrm{d}^3x\,\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\left[\alpha_{\boldsymbol{k}}\phi(t,\boldsymbol{x}) + \mathrm{i}\beta_{\boldsymbol{k}}\pi(t,\boldsymbol{x})\right] \tag{99.32}$$

follow from $\int \mathrm{d}^3x\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\mathrm{e}^{\pm\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{x}} = (2\pi)^3\delta^3(\boldsymbol{k}\mp\boldsymbol{k}')$. Note that $\alpha_{\boldsymbol{k}}$ and $\beta_{\boldsymbol{k}}$ are the reciprocals of the two normalization factors appearing in Equations (99.29) and (99.30), so that the spatial Fourier transforms of those two equations are

$$\alpha_{\boldsymbol{k}} \int \mathrm{d}^3x\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\phi(t,\boldsymbol{x}) = a(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}\omega_{\boldsymbol{k}}t} + a^\dagger(-\boldsymbol{k})\mathrm{e}^{\mathrm{i}\omega_{\boldsymbol{k}}t}\,,$$
$$\mathrm{i}\beta_{\boldsymbol{k}} \int \mathrm{d}^3x\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\pi(t,\boldsymbol{x}) = a(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}\omega_{\boldsymbol{k}}t} - a^\dagger(-\boldsymbol{k})\mathrm{e}^{\mathrm{i}\omega_{\boldsymbol{k}}t}\,,$$

whose sum is $2a(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}\omega_{\boldsymbol{k}}t}$: the $a^\dagger(-\boldsymbol{k})$ terms cancel, which is Equation (99.32). Then

$$\begin{aligned}\left[a(\boldsymbol{k}), a^\dagger(\boldsymbol{k}')\right] &= \frac{\mathrm{e}^{\mathrm{i}(\omega-\omega')t}}{4}\int \mathrm{d}^3x\, \mathrm{d}^3y\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}+\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{y}} \\ &\qquad \times\left(-\mathrm{i}\alpha_{\boldsymbol{k}}\beta_{\boldsymbol{k}'}\left[\phi(\boldsymbol{x}),\pi(\boldsymbol{y})\right] + \mathrm{i}\beta_{\boldsymbol{k}}\alpha_{\boldsymbol{k}'}\left[\pi(\boldsymbol{x}),\phi(\boldsymbol{y})\right]\right) \\ &= \frac{\hbar}{4}\int \mathrm{d}^3x\, \mathrm{e}^{\mathrm{i}(\boldsymbol{k}'-\boldsymbol{k})\cdot\boldsymbol{x}}\left(\alpha_{\boldsymbol{k}}\beta_{\boldsymbol{k}'} + \beta_{\boldsymbol{k}}\alpha_{\boldsymbol{k}'}\right)\,,\end{aligned}$$

the $[\phi,\phi]$ and $[\pi,\pi]$ terms having dropped by Equation (99.24). The remaining integral is $(2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}')$, which sets $\boldsymbol{k}'=\boldsymbol{k}$ in the bracket; there $\alpha_{\boldsymbol{k}}\beta_{\boldsymbol{k}} = \sqrt{(2\omega/\hbar c^2)(2c^2/\hbar\omega)} = 2/\hbar$, so the bracket is $4/\hbar$ and the prefactor $\hbar/4$ cancels it exactly. For $[a(\boldsymbol{k}), a(\boldsymbol{k}')]$ the same expansion gives $\mathrm{i}\alpha_{\boldsymbol{k}}\beta_{\boldsymbol{k}'}\left[\phi(\boldsymbol{x}),\pi(\boldsymbol{y})\right] + \mathrm{i}\beta_{\boldsymbol{k}}\alpha_{\boldsymbol{k}'}\left[\pi(\boldsymbol{x}),\phi(\boldsymbol{y})\right]$, whose two terms are $+\mathrm{i}\hbar$ and $-\mathrm{i}\hbar$ times the same delta and therefore cancel; likewise for the daggered pair. The converse is the same algebra run backwards: substituting Equation (99.31) into Equations (99.29) and (99.30) reproduces Equation (99.23), as the computation in Proposition 99.44 does explicitly for a general time difference. □

Equation (99.31) is the ladder algebra of the quantum harmonic oscillator, one copy for each $\boldsymbol{k}$, with the Kronecker delta of the discrete oscillator replaced by a Dirac delta; it is the algebra Dirac used for the radiation field [572]. The modes it acts on are the continuum limit of the coupled oscillators of Chapter 28, which is a classical chapter and carries the normal modes but not this algebra. What the algebra buys is that the oscillator spectrum applies mode by mode, and the next theorem derives that here rather than importing it.

Theorem 99.9 (The free scalar Hamiltonian)**.** *Substituting Equations* (99.29) *and* (99.30) *into Equation* (99.21)*,*

$$H = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\frac{\hbar\omega_{\boldsymbol{k}}}{2}\left[a^\dagger(\boldsymbol{k})a(\boldsymbol{k}) + a(\boldsymbol{k})a^\dagger(\boldsymbol{k})\right]\,, \tag{99.33}$$

and the momentum operator $\boldsymbol{P} = -\int \mathrm{d}^3x\, \pi\nabla\phi$ *is*

$$\boldsymbol{P} = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\hbar\boldsymbol{k}\,\frac{1}{2}\left[a^\dagger(\boldsymbol{k})a(\boldsymbol{k}) + a(\boldsymbol{k})a^\dagger(\boldsymbol{k})\right]\,. \tag{99.34}$$

Rests on Equations (99.21)*,* (99.29) *and* (99.30)*.*

Proof. *Derives Theorem 99.9.*

Take the three terms of Equation (99.21) in turn. Each is a product of two mode integrals, and $\int \mathrm{d}^3x$ produces $(2\pi)^3\delta^3(\boldsymbol{k}\pm\boldsymbol{k}')$, so every term reduces to a single integral over $\boldsymbol{k}$ in which the partner is either $\boldsymbol{k}$ (for the $aa^\dagger$ and $a^\dagger a$ pieces) or $-\boldsymbol{k}$ (for the aa and $a^\dagger a^\dagger$ pieces).

For the aa and $a^\dagger a^\dagger$ pieces, collect coefficients. From $\frac{1}{2}c^2\pi^2$ the coefficient is $-\frac{1}{2}c^2(\hbar\omega/2c^2) = -\hbar\omega/4$, the minus sign coming from $(-\mathrm{i})^2$. From $\frac{1}{2}(\nabla\phi)^2$, the gradient acting on $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}$ gives $\mathrm{i}\boldsymbol{k}$ and on the partner $\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}$ gives $-\mathrm{i}\boldsymbol{k}$, so the coefficient is $+\frac{1}{2}(\hbar c^2/2\omega)\boldsymbol{k}^2$; the mass term adds $+\frac{1}{2}(\hbar c^2/2\omega)\kappa^2$. Their sum is $\frac{1}{2}(\hbar c^2/2\omega)(\boldsymbol{k}^2+\kappa^2) = \hbar\omega/4$

by Equation (99.28), cancelling the contribution from π^2 exactly. The time-dependent factors $\mathrm{e}^{\mp 2\mathrm{i}\omega t}$ that multiply these pieces therefore never appear, which is the statement that H is conserved.

For the $a^\dagger a$ and $aa^\dagger$ pieces the relative sign is reversed: π^2 contributes $+\hbar\omega/4$ to each and the gradient and mass terms contribute $+\hbar\omega/4$ to each, giving $\hbar\omega/2$ for each of $a^\dagger a$ and $aa^\dagger$, which is Equation (99.33). Equation (99.34) is the same computation with one factor of ∇ in place of one factor of ∂_t/c^2. □

Two consequences follow at once from Equation (99.31) and Equation (99.33):

$$\left[H, a^\dagger(\boldsymbol{k})\right] = \hbar\omega_{\boldsymbol{k}}\, a^\dagger(\boldsymbol{k})\,, \qquad [H, a(\boldsymbol{k})] = -\hbar\omega_{\boldsymbol{k}}\, a(\boldsymbol{k})\,, \tag{99.35}$$

so $a^\dagger(\boldsymbol{k})$ raises the energy by $\hbar\omega_{\boldsymbol{k}}$ and $a(\boldsymbol{k})$ lowers it by the same amount; and by Equation (99.34) the same operators raise and lower the momentum by $\hbar\boldsymbol{k}$. The energy and momentum added by one application of $a^\dagger(\boldsymbol{k})$ satisfy

$$(\hbar\omega_{\boldsymbol{k}})^2 - (\hbar\boldsymbol{k})^2 c^2 = \hbar^2 c^2 \left(\boldsymbol{k}^2 + \kappa^2\right) - \hbar^2 c^2 \boldsymbol{k}^2 = \hbar^2 c^2 \kappa^2 = m^2 c^4\,, \tag{99.36}$$

the relativistic mass shell of a particle of mass m. This is the central identification of the chapter, and it was not put in: it came out of diagonalizing a field Hamiltonian.

Proposition 99.10 (The invariant measure and the relativistic normalization)**.** *The combination $\omega_{\boldsymbol{k}}\,\delta^3(\boldsymbol{k}-\boldsymbol{k}')$ is invariant under proper orthochronous Lorentz transformations, and*

$$\int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\frac{c}{2\omega_{\boldsymbol{k}}}\,F(k) = \int \frac{\mathrm{d}^4k}{(2\pi)^3}\,\delta\!\left(k^2-\kappa^2\right)\theta(k^0)\,F(k) \tag{99.37}$$

for any F. Consequently the states

$$|\boldsymbol{k}\rangle := \sqrt{2\omega_{\boldsymbol{k}}}\,a^\dagger(\boldsymbol{k})\,|0\rangle\,, \qquad \langle \boldsymbol{k}|\boldsymbol{k}'\rangle = 2\omega_{\boldsymbol{k}}(2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}')\,, \tag{99.38}$$

have a Lorentz-invariant normalization.

Rests on Equations (99.28) and (99.31).

Proof. *Derives Proposition 99.10.*

For a boost of rapidity η along the third axis, $k'^3 = \gamma(k^3 - \beta\omega_{\boldsymbol{k}}/c)$ with $\gamma = \cosh\eta$, $\beta = \tanh\eta$, while k^1, k^2 are unchanged. On the mass shell $\omega_{\boldsymbol{k}} = c\sqrt{\boldsymbol{k}^2+\kappa^2}$, so $\partial\omega_{\boldsymbol{k}}/\partial k^3 = c^2k^3/\omega_{\boldsymbol{k}}$ and

$$\frac{\partial k'^3}{\partial k^3} = \gamma\left(1 - \frac{\beta}{c}\,\frac{c^2k^3}{\omega_{\boldsymbol{k}}}\right) = \frac{\gamma\left(\omega_{\boldsymbol{k}} - \beta c k^3\right)}{\omega_{\boldsymbol{k}}} = \frac{\omega_{\boldsymbol{k}'}}{\omega_{\boldsymbol{k}}}\,,$$

the last equality being the transformation of $k^0 = \omega/c$ itself. Hence $\mathrm{d}^3k' = (\omega_{\boldsymbol{k}'}/\omega_{\boldsymbol{k}})\mathrm{d}^3k$, so $\mathrm{d}^3k/\omega_{\boldsymbol{k}}$ is invariant, and a delta transforms with the inverse Jacobian of its argument, so $\omega_{\boldsymbol{k}}\delta^3(\boldsymbol{k}-\boldsymbol{k}')$ is invariant too. For Equation (99.37), write $k^2-\kappa^2 = (k^0)^2 - \boldsymbol{k}^2 - \kappa^2$ and use $\delta(g(k^0)) = \sum_i \delta(k^0-k^0_i)/\,|g'(k^0_i)|$ with the roots $k^0 = \pm\omega_{\boldsymbol{k}}/c$; the step function keeps the positive one, at which $|g'| = 2\omega_{\boldsymbol{k}}/c$. □

Equation (99.37) is the reason every momentum-space formula in this part carries $\mathrm{d}^3k/(2\pi)^3 2\omega_{\boldsymbol{k}}$ rather than d^3k: it is the unique (up to normalization) Lorentz-invariant measure on the mass shell, and using anything else makes covariance invisible. The convention Equation (99.38) is the one Chapter 100 uses, where the external-line spinors of Proposition 100.14 carry the matching normalization $\bar{u}u = 2m_e c$ and $u^\dagger u = 2E/c$; Chapter 86 is deliberately nonrelativistic and does not use it.

99.2.3 Fock space

The spectrum of Equation (99.33) is built exactly as for one oscillator: find the state annihilated by every lowering operator, then climb.

Definition 99.11 (Vacuum, Fock space, number operator)**.** The *vacuum* $|0\rangle$ is the normalized state with

$$a(\boldsymbol{k})\,|0\rangle = 0 \quad \text{for every } \boldsymbol{k}\,, \qquad \langle 0|0\rangle = 1\,. \tag{99.39}$$

The n-quantum states are $a^\dagger(\boldsymbol{k}_1)\cdots a^\dagger(\boldsymbol{k}_n)\,|0\rangle$, their closed span is the n-particle sector $\mathcal{F}_n$, and *Fock space* is the completed direct sum

$$\mathcal{F} = \bigoplus_{n=0}^{\infty} \mathcal{F}_n = \bigoplus_{n=0}^{\infty} \mathrm{Sym}^n(\mathcal{H}_1)\ , \tag{99.40}$$

with $\mathcal{H}_1$ the one-quantum space. The number operator is

$$N = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\, a^\dagger(\boldsymbol{k})a(\boldsymbol{k})\,, \tag{99.41}$$

whose eigenvalue on $\mathcal{F}_n$ is n.

This is Fock's construction [698]. Three properties make it the right state space, and each is a direct consequence of Equation (99.31).

The one-quantum states are particles. By Equation (99.35) and its momentum analogue, $a^\dagger(\boldsymbol{k})\,|0\rangle$ is a simultaneous eigenstate of H and $\boldsymbol{P}$ with eigenvalues $\hbar\omega_{\boldsymbol{k}}$ and $\hbar\boldsymbol{k}$ (measured from the vacuum values, which Section 99.2.4 disposes of), related by Equation (99.36). The space $\mathcal{H}_1$ they span carries a unitary representation of the Poincaré group, and by Proposition 99.25 below that representation is irreducible with $\hat{P}^2 = m^2c^2$ and $\hat{W}^2 = 0$ — mass m, spin 0, in the classification Wigner obtained for the group itself [849] and which Chapter 94 carries out in full. That is what entitles one to call the excitation a particle rather than a wave amplitude.

Multi-quantum states are automatically symmetric. Since $\left[a^\dagger(\boldsymbol{k}), a^\dagger(\boldsymbol{k}')\right] = 0$,

$$a^\dagger(\boldsymbol{k}_1)a^\dagger(\boldsymbol{k}_2)\,|0\rangle = a^\dagger(\boldsymbol{k}_2)a^\dagger(\boldsymbol{k}_1)\,|0\rangle\ , \tag{99.42}$$

so the two-quantum state is unchanged under exchange of its labels. The symmetrization postulate of Chapter 83 is therefore not a postulate here but a consequence of the choice of commutators, and with it the Bose–Einstein counting that Chapter 120 builds the statistical mechanics of photons and phonons on. The point deserves emphasis: identical particles are identical in this formalism because they are not particles at all but quanta of one field, and there is nothing to label. What the label $\boldsymbol{k}$ names is a mode, not an individual.

The construction is exhaustive and stable. Every state of the free theory is reached from $|0\rangle$ by creation operators, because Equation (99.33) is a sum of commuting oscillator Hamiltonians each of which has the familiar nondegenerate ladder spectrum; and $[N, H] = 0$ by Equations (99.31) and (99.33), so the number of quanta of a free field is conserved. It is exactly the addition of an interaction term to H — something cubic or quartic in ϕ, hence not commuting with N — that lets the number change, which is what Section 99.1.1 demanded and what Chapter 100 exploits.

Proposition 99.12 (Normalization of the n-quantum states)**.** *With* $|\boldsymbol{k}_1\ldots\boldsymbol{k}_n\rangle = \prod_i \sqrt{2\omega_{\boldsymbol{k}_i}}\, a^\dagger(\boldsymbol{k}_i)\,|0\rangle$,

$$\langle \boldsymbol{k}_1\ldots\boldsymbol{k}_n|\boldsymbol{k}'_1\ldots\boldsymbol{k}'_n\rangle = \sum_{\sigma\in S_n}\prod_{i=1}^{n} 2\omega_{\boldsymbol{k}_i}(2\pi)^3\delta^3\!\left(\boldsymbol{k}_i - \boldsymbol{k}'_{\sigma(i)}\right)\ , \tag{99.43}$$

and the resolution of the identity on $\mathcal{F}$ is

$$\mathbb{1} = |0\rangle\langle 0| + \sum_{n\geq 1} \frac{1}{n!} \int \prod_{i=1}^{n} \frac{\mathrm{d}^3 k_i}{(2\pi)^3 2\omega_{\boldsymbol{k}_i}} \; |\boldsymbol{k}_1 \dots \boldsymbol{k}_n\rangle\langle \boldsymbol{k}_1 \dots \boldsymbol{k}_n| \; . \tag{99.44}$$

Rests on Equations (99.31) *and* (99.38).

Proof. *Derives Proposition 99.12.*

Move every annihilation operator of the bra to the right through the creation operators of the ket using Equation (99.31). Each annihilator must be contracted with exactly one creator, since any uncontracted annihilator meets $|0\rangle$ and gives zero; the sum over which creator it meets is the sum over permutations, and each contraction supplies one factor $(2\pi)^3\delta^3$, while the normalization factors supply 2ω. Equation (99.44) follows because the $1/n!$ compensates the $n!$ permutations in Equation (99.43) when the identity is applied to a symmetric state. □

Remark 99.13 (Fock space is a choice, not a deduction). Definition 99.11 presupposes that a state annihilated by every $a(\boldsymbol{k})$ exists and is unique. For the free field it does. For an interacting field it is exactly what Haag's theorem (Section 99.9.1) denies: the interacting theory's vacuum is not in the free theory's Fock space, and the two representations of Equation (99.23) are unitarily inequivalent. Nothing in this section is wrong, but its scope is the free field, and the reader should not carry the picture of "the" Fock space into the interacting theory without the warning of Section 99.9.1.

99.2.4 Normal ordering and the vacuum energy

Equation (99.33) was left in symmetric form on purpose. Using Equation (99.31) to move the annihilator to the right,

$$H = \int \frac{\mathrm{d}^3 k}{(2\pi)^3}\, \hbar\omega_{\boldsymbol{k}} \left[a^\dagger(\boldsymbol{k}) a(\boldsymbol{k}) + \frac{1}{2}(2\pi)^3\delta^3(\boldsymbol{0}) \right] , \tag{99.45}$$

and $(2\pi)^3\delta^3(\boldsymbol{0}) = \int \mathrm{d}^3 x = V$ is the volume of space. The vacuum energy is therefore $E_0 = V\varrho_0$ with energy density

$$\varrho_0 = \int \frac{\mathrm{d}^3 k}{(2\pi)^3} \frac{\hbar\omega_{\boldsymbol{k}}}{2} = \frac{\hbar c}{4\pi^2} \int_0^\infty \mathrm{d}k \; k^2 \sqrt{k^2 + \kappa^2} \, , \tag{99.46}$$

which diverges as the fourth power of the upper limit: cutting the integral off at wave number $\Lambda \gg \kappa$ gives

$$\varrho_0 \approx \frac{\hbar c}{4\pi^2} \int_0^\Lambda k^3 \, \mathrm{d}k = \frac{\hbar c \Lambda^4}{16\pi^2} \, . \tag{99.47}$$

It is one half-quantum $\frac{1}{2}\hbar\omega_{\boldsymbol{k}}$ per mode — the zero-point energy of a single quantum oscillator, which Phenomenon 78.6 derives from the uncertainty relation and which is measured in the isotope shift of a band spectrum — summed over infinitely many modes.

Definition 99.14 (Normal ordering)**.** The normal-ordered product $:X:$ of a product of free-field operators is the same product with every creation operator moved to the left of every annihilation operator, the moves performed as if all the operators commuted — that is, discarding the commutators generated.

The prescription is Wick's [974], and by construction $\langle 0| :X: |0\rangle = 0$. Replacing H by

$$:H: = \int \frac{\mathrm{d}^3 k}{(2\pi)^3}\, \hbar\omega_{\boldsymbol{k}}\, a^\dagger(\boldsymbol{k}) a(\boldsymbol{k}) \tag{99.48}$$

removes Equation (99.46) and assigns the vacuum zero energy. It is essential to be honest about what this step is and is not.

What it *is*: a choice of the additive constant in the Hamiltonian. The equations of motion involve only commutators with H, and a c-number added to H commutes with everything, so no Heisenberg equation, no scattering amplitude, no energy difference and no transition rate in flat space depends on the choice. Ordering ambiguities of exactly this kind arise whenever a classical product is promoted to an operator product, and the free field is the case in which the ambiguity is a pure constant.

What it is *not*: a proof that zero-point energy is unphysical. There are two situations in which the constant is not free, and both are experimental questions rather than formal ones.

Boundary-dependent differences. If boundaries restrict which modes exist, ϱ_0 depends on the geometry, and the *difference* between two geometries is finite even though each is infinite. It is measured. That is the Casimir effect, computed and compared with experiment in Section 99.8.

Gravitational coupling. General relativity couples to the total energy–momentum tensor, not to energy differences, so the constant that Equation (99.48) discards is, in principle, gravitationally visible as a cosmological constant. Taking Equation (99.47) at face value with the cutoff at an energy $E_c = \hbar c \Lambda$ gives

$$\varrho_0 \approx \frac{E_c^4}{16\pi^2(\hbar c)^3} = \begin{cases} 1.3 \times 10^{47}\,\mathrm{J/m^3}\,, & E_c = 1\,\mathrm{TeV}\,, \\ 2.9 \times 10^{111}\,\mathrm{J/m^3}\,, & E_c = E_\mathrm{P} = 1.956 \times 10^{9}\,\mathrm{J}\,, \end{cases} \tag{99.49}$$

against an observed dark-energy density of order $1 \times 10^{-9}\,\mathrm{J/m^3}$: a discrepancy of some 56 orders of magnitude if one stops at the highest energy at which the Standard Model has been tested, and of some 120 if one runs to the Planck energy. This is the cosmological constant problem [1828]. No symmetry of the Standard Model forbids the term, no cancellation among the known fields achieves it, and the problem is stated as open in Chapter 48 and Chapter 133; nothing in this chapter solves it, and the reader should treat any text that quietly normal-orders the problem away as having changed the subject.

Remark 99.15 (What the estimate does and does not assume). Equation (99.49) is a cutoff estimate, and a sharp cutoff in $|\boldsymbol{k}|$ is not Lorentz invariant — it treats the modes as if there were a preferred frame, and a Lorentz-invariant regulator gives an energy density with the equation of state $p = -\varrho$ appropriate to a cosmological constant rather than the $p = \varrho/3$ that Equation (99.47) naively suggests. What survives the change of regulator is the *magnitude*, which is set by the fourth power of the largest energy at which the field description is assumed valid. It is the magnitude, not the tensor structure, that is 10^{56} times too large.

99.3 The complex scalar field and antiparticles

A real field describes one species of quantum, which is its own antiquantum and carries no charge. Allowing the field to be complex doubles the classical degrees of freedom, and the doubling shows up in the quantum theory as two species of quantum of equal mass and opposite charge. Antimatter is a prediction of field theory, obtained here without any reference to a filled sea.

99.3.1 Global phase symmetry and conserved charge

Let ϕ be a complex field with Lagrangian density

$$\mathcal{L} = \partial_\mu \phi^\dagger \partial^\mu \phi - \kappa^2 \phi^\dagger \phi = \frac{1}{c^2}\partial_t \phi^\dagger \partial_t \phi - \nabla\phi^\dagger \cdot \nabla\phi - \kappa^2 \phi^\dagger \phi\,, \tag{99.50}$$

with $[\phi]$ again given by Equation (99.19); the factor $\frac{1}{2}$ of Equation (99.18) is absent because ϕ and $\phi^\dagger$ are independent. Both ϕ and $\phi^\dagger$ obey $(\Box + \kappa^2)\,\phi = 0$. The conjugate momenta are

$$\pi = \frac{\partial \mathcal{L}}{\partial \dot{\phi}} = \frac{1}{c^2}\dot{\phi}^\dagger\,, \qquad \pi^\dagger = \frac{1}{c^2}\dot{\phi}\,, \tag{99.51}$$

and the canonical brackets are

$$[\phi(t,\boldsymbol{x}), \pi(t,\boldsymbol{y})] = \left[\phi^\dagger(t,\boldsymbol{x}), \pi^\dagger(t,\boldsymbol{y})\right] = \mathrm{i}\hbar\,\delta^3(\boldsymbol{x}-\boldsymbol{y})\,, \tag{99.52}$$

all others vanishing.

Equation (99.50) is invariant under the global phase rotation

$$\phi \longmapsto \mathrm{e}^{-\mathrm{i}\theta}\phi\,, \qquad \phi^\dagger \longmapsto \mathrm{e}^{\mathrm{i}\theta}\phi^\dagger\,, \qquad \theta \in \mathbb{R} \text{ constant}\,, \tag{99.53}$$

a representation of the group $\mathrm{U}(1)$ of unit-modulus complex numbers. That group is isomorphic to the rotation group $\mathrm{SO}(2,\mathbb{R})$ of Section 14.2.1 by $\mathrm{e}^{\mathrm{i}\theta} \mapsto R(\theta)$ — the section is written for $\mathrm{SO}(2,\mathbb{R})$ and does not mention $\mathrm{U}(1)$, but its representation theory transfers along the isomorphism unchanged. By Noether's theorem [456], in the field form derived in Section 99.4.1 below, the symmetry carries a conserved current.

Proposition 99.16 (The U(1) current and charge)**.** *The current*

$$j^\mu = \mathrm{i}\left(\phi^\dagger \partial^\mu \phi - \phi\,\partial^\mu \phi^\dagger\right)\,, \qquad [j^\mu] = \mathrm{J/m^2}\,, \tag{99.54}$$

satisfies $\partial_\mu j^\mu = 0$ on solutions, and the dimensionless quantity

$$Q = \frac{1}{\hbar c}\int \mathrm{d}^3x\, j^0 = \frac{\mathrm{i}}{\hbar}\int \mathrm{d}^3x\left(\phi^\dagger\pi^\dagger - \pi\phi\right) \tag{99.55}$$

is conserved. The electric charge of a field of charge quantum q is qQ.

Rests on Equations (99.50)*,* (99.51) *and* (99.53)*.*

Proof. *Derives Proposition 99.16.*

Conservation is direct: $\partial_\mu j^\mu = \mathrm{i}(\phi^\dagger\Box\phi - \phi\Box\phi^\dagger) = \mathrm{i}(-\kappa^2\phi^\dagger\phi + \kappa^2\phi\phi^\dagger) = 0$, the cross terms $\partial_\mu\phi^\dagger\partial^\mu\phi$ cancelling identically and the field equation supplying the rest. That Equation (99.54) is the Noether current of Equation (99.53) is Theorem 99.22 applied with $\delta\phi = -\mathrm{i}\phi$, $\delta\phi^\dagger = +\mathrm{i}\phi^\dagger$:

$$j^\mu = \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}(-\mathrm{i}\phi) + \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi^\dagger)}(\mathrm{i}\phi^\dagger) = -\mathrm{i}\phi\,\partial^\mu\phi^\dagger + \mathrm{i}\phi^\dagger\partial^\mu\phi\,.$$

For Equation (99.55), note $j^0 = \mathrm{i}(\phi^\dagger c^{-1}\partial_t\phi - (c^{-1}\partial_t\phi^\dagger)\phi) = \mathrm{i}c(\phi^\dagger\pi^\dagger - \pi\phi)$ by Equation (99.51) — the second factor written to the right of its derivative, as the Noether construction delivers it, since ϕ and π do not commute and the two orderings differ by the c-number $[\pi,\phi]$ that Equation (99.60) normal-orders away. Dividing by $\hbar c$ makes the integrand an inverse volume: $[j^0] = \mathrm{J/m^2}$ and $[\hbar c] = \mathrm{J\,m}$. Conservation of Q follows from $\partial_\mu j^\mu = 0$ by integrating over a spatial slice and discarding the surface term at infinity. □

This settles the interpretive question left open in Section 99.1.1. The density Equation (99.4) that could not be a probability density *is* Equation (99.54) up to normalization, and the reason it may take either sign is that it is a charge density. Charge densities of either sign are unremarkable; probability densities of either sign are nonsense. Nothing had to be repaired — the quantity had been misidentified. The same reading is what makes minimal coupling $\partial_\mu \to \partial_\mu + \mathrm{i}qA_\mu/\hbar$ legitimate, since $-j^\mu A_\mu$ is then the interaction energy of a charge distribution with the potential, exactly as in Theorem 100.5; a probability current has no business coupling to the electromagnetic field.

Notice also what has *not* been explained. Nothing above forces the eigenvalues of Q to be integers times one universal quantum, and nothing forces the electron and proton charges to be equal and opposite to the precision at which that is measured. The U(1) of Equation (99.53) acts on one field, and nothing fixes the phase a second field of a different charge would pick up: the symmetry actually forced is a group of real phases, whose representations are labelled by a real number. That the observed charges are integer multiples of one quantum is an experimental fact — its current status, including the fractional quark charges never seen free, is Section 63.11.3 — which the free field theory does not derive and which nothing in this chapter explains.

99.3.2 Two species of quantum: particle and antiparticle

The field is complex, so its Fourier coefficients at $\boldsymbol{k}$ and the conjugates at $-\boldsymbol{k}$ are no longer related. Two independent families of ladder operators appear.

Definition 99.17 (Mode expansion of the complex scalar)**.**

$$\phi(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\sqrt{\frac{\hbar c^2}{2\omega_{\boldsymbol{k}}}}\left[a(\boldsymbol{k})\,\mathrm{e}^{-\mathrm{i}k\cdot x} + b^\dagger(\boldsymbol{k})\,\mathrm{e}^{\mathrm{i}k\cdot x}\right], \tag{99.56}$$

$$\phi^\dagger(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\sqrt{\frac{\hbar c^2}{2\omega_{\boldsymbol{k}}}}\left[b(\boldsymbol{k})\,\mathrm{e}^{-\mathrm{i}k\cdot x} + a^\dagger(\boldsymbol{k})\,\mathrm{e}^{\mathrm{i}k\cdot x}\right], \tag{99.57}$$

with

$$\left[a(\boldsymbol{k}), a^\dagger(\boldsymbol{k}')\right] = \left[b(\boldsymbol{k}), b^\dagger(\boldsymbol{k}')\right] = (2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}'), \tag{99.58}$$

and every other commutator among $a, a^\dagger, b, b^\dagger$ vanishing.

That Equation (99.58) is equivalent to Equation (99.52) is the computation of Proposition 99.8 repeated: the $[\phi,\pi]$ bracket receives the a contribution with one sign and the b contribution with the other, and the two add rather than cancel because $\left[b^\dagger, b\right] = -\left[b, b^\dagger\right]$; the normalization $\sqrt{\hbar c^2/2\omega}$ is unchanged.

Theorem 99.18 (Energy and charge of the complex scalar)**.** *With normal ordering,*

$$:H: = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\hbar\omega_{\boldsymbol{k}}\left[a^\dagger(\boldsymbol{k})a(\boldsymbol{k}) + b^\dagger(\boldsymbol{k})b(\boldsymbol{k})\right], \tag{99.59}$$

$$:Q: = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\left[a^\dagger(\boldsymbol{k})a(\boldsymbol{k}) - b^\dagger(\boldsymbol{k})b(\boldsymbol{k})\right]. \tag{99.60}$$

The two species have identical dispersion Equation (99.28), hence identical mass, and charges $+1$ *and* -1.

Rests on Definition 99.17, Equation (99.55), and Theorem 99.9.

Proof. *Derives Theorem 99.18.*

The Hamiltonian density is $\mathcal{H} = c^2\pi^\dagger\pi + \nabla\phi^\dagger\cdot\nabla\phi + \kappa^2\phi^\dagger\phi$; substituting Equations (99.56) and (99.57) and integrating over $\boldsymbol{x}$ proceeds exactly as in Theorem 99.9, the aa and $a^\dagger a^\dagger$ pieces cancelling between the π term and the gradient-plus-mass terms, and yields $\int_{\boldsymbol{k}}\hbar\omega_{\boldsymbol{k}}(a^\dagger a + b\,b^\dagger)$, which normal-ordering turns into Equation (99.59). For the charge, insert Equation (99.51) into Equation (99.55):

$$Q = \frac{\mathrm{i}}{\hbar c^2}\int \mathrm{d}^3x\left(\phi^\dagger\dot\phi - \dot\phi^\dagger\phi\right),$$

and substitute the expansions. The spatial integral pairs each mode either with itself or with its reflection, and the terms pairing two annihilation or two creation operators — $b\,a$ carrying $\mathrm{e}^{-2\mathrm{i}\omega t}$ and $a^\dagger b^\dagger$ carrying $\mathrm{e}^{2\mathrm{i}\omega t}$ — occur in $\phi^\dagger\dot\phi$ and in $\dot\phi^\dagger\phi$ with the same coefficient

and cancel in the difference. The surviving terms occur with opposite coefficients in the two products and therefore add: $a^\dagger a$ with $-2\mathrm{i}\omega_{\boldsymbol{k}} f_{\boldsymbol{k}}^2$ and $b\,b^\dagger$ with $+2\mathrm{i}\omega_{\boldsymbol{k}} f_{\boldsymbol{k}}^2$, where $f_{\boldsymbol{k}}^2 = \hbar c^2/2\omega_{\boldsymbol{k}}$. Multiplying by the prefactor $\mathrm{i}/\hbar c^2$ turns these into $+1$ and -1, giving $Q = \int_{\boldsymbol{k}} (a^\dagger a - b\,b^\dagger)$ and Equation (99.60) after normal ordering. □

Both terms of Equation (99.59) are non-negative: the b quanta cost positive energy, not negative. The negative-energy solutions of Section 99.1.1 have not been suppressed, filled or forbidden — they have been reinterpreted. The coefficient of the "negative-frequency" exponential $\mathrm{e}^{+\mathrm{i}k\cdot x}$ in Equation (99.56) is not the amplitude to find a particle of negative energy but the operator that *creates* an antiparticle of positive energy. This is the Stueckelberg–Feynman reading [864, 931], and Section 99.6.1 shows that the same exponential is what makes a single propagator serve for both species.

Remark 99.19 (No sea, and what that costs). Compare with Section 99.1.1. Hole theory required an infinitely occupied vacuum, which is unavailable for bosons, and it predicted antiparticles only for fermions. Here the antiparticle appears for a spin-0 field, with a vacuum containing no quanta at all, and it appears for exactly the reason the hole picture could not accommodate: the field, not the particle, is fundamental, and a complex field has two kinds of excitation because it has two independent real components. That is Pauli and Weisskopf's argument [748], and it is the reason the hole theory [647] is now of historical interest only, correct predictions notwithstanding.

Two experimental anchors are worth naming precisely, since the claim being tested is not the existence of antimatter but the exact degeneracy that Theorem 99.18 asserts. The first is the positron itself: predicted, then found in cloud-chamber tracks whose curvature in a magnetic field gave a positive particle of about the electron mass [713], the subject of Chapter 95. The second is the modern programme of antiproton and antihydrogen measurements pursued in Chapter 90, which tests the equality of particle and antiparticle masses and charges directly rather than inferring it. Both belong to the wider CPT programme of Chapter 105: the degeneracy asserted here follows from CPT, so a measured violation of it would be a violation of one of the assumptions — Lorentz invariance, locality, or the positivity of the energy — on which this whole chapter rests.

99.3.3 Charge conjugation

The doubling of species comes with a discrete symmetry that exchanges them.

Definition 99.20 (Charge conjugation on the free complex scalar)**.** C is the unitary operator with

$$C\,a(\boldsymbol{k})\,C^{-1} = \eta_C\,b(\boldsymbol{k})\,, \qquad C\,b(\boldsymbol{k})\,C^{-1} = \eta_C^*\,a(\boldsymbol{k})\,, \qquad C\,|0\rangle = |0\rangle\,, \tag{99.61}$$

with $|\eta_C| = 1$ a convention-dependent phase, conventionally $\eta_C = 1$.

Proposition 99.21 (Action of C on field and current)**.** *With $\eta_C = 1$,*

$$C\,\phi(x)\,C^{-1} = \phi^\dagger(x)\,, \qquad C\,j^\mu(x)\,C^{-1} = -j^\mu(x)\,, \qquad C\,{:}Q{:}\,C^{-1} = -\,{:}Q{:}\,, \tag{99.62}$$

and $[C, H] = 0$ for the free field.

Rests on Equations (99.60) and (99.61) and Definition 99.17.

Proof. *Derives Proposition 99.21.*

Apply Equation (99.61) term by term to Equation (99.56): the coefficient of $\mathrm{e}^{-\mathrm{i}k\cdot x}$ becomes $b(\boldsymbol{k})$ and that of $\mathrm{e}^{\mathrm{i}k\cdot x}$ becomes $a^\dagger(\boldsymbol{k})$, which is Equation (99.57). Then $j^\mu = \mathrm{i}(\phi^\dagger\partial^\mu\phi - \phi\partial^\mu\phi^\dagger)$ maps to $\mathrm{i}(\phi\partial^\mu\phi^\dagger - \phi^\dagger\partial^\mu\phi) = -j^\mu$, and integrating gives the statement for Q; equivalently, Equation (99.60) is manifestly odd under $a \leftrightarrow b$. That Equation (99.59) is even under the same exchange is $[C, H] = 0$. □

For a *real* scalar field $\phi^\dagger = \phi$, so Equation (99.62) forces $C\phi C^{-1} = \pm\phi$: a real field is its own antifield, has $Q \equiv 0$, and carries an intrinsic C parity. The electromagnetic field has $C = -1$, since C reverses all charges and hence all currents and the potentials they source; a state of n photons therefore has $C = (-1)^n$, which is why the C-even neutral pion decays to two photons and not to three.

Beyond this the account must be deferred. C alone is one of three discrete operations, and the free complex scalar is too simple a system to show what is interesting about them: the weak interaction violates C and P separately and maximally (Chapter 101), the combination CP is violated too but only slightly, and the combination CPT is a theorem of any Lorentz-invariant local field theory with a bounded-below Hamiltonian [1115, 1123] — the same three hypotheses that appear in the spin–statistics theorem of Section 99.7.2. That is not an accident, and it is the subject of Chapter 105.

99.4 Noether charges and the Poincaré algebra

Two results used repeatedly above were stated as they were needed: that a continuous symmetry of the Lagrangian density gives a conserved current, and that the one-quantum states of a free field carry an irreducible representation of the Poincaré group labelled by a mass and a spin. This section proves both. They are what make the phrase "a particle of mass m and spin s" a definition rather than a description.

99.4.1 Noether's theorem in field form

Theorem 99.22 (Noether, for fields)**.** *Let $\mathcal{L}(\phi_a, \partial_\mu \phi_a)$ be a local Lagrangian density for a collection of fields ϕ_a, and let the infinitesimal transformation $\phi_a \mapsto \phi_a + \varepsilon F_a[\phi]$ change $\mathcal{L}$ by at most a total divergence, $\delta\mathcal{L} = \varepsilon\, \partial_\mu K^\mu$. Then*

$$j^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_a)} F_a - K^\mu \tag{99.63}$$

satisfies $\partial_\mu j^\mu = 0$ on solutions of the field equations, and $Q = c^{-1}\int \mathrm{d}^3x\, j^0$ is constant in time whenever the fields fall off fast enough at spatial infinity.

Rests on Definition 99.4 and Equation (99.16).

Proof. *Derives Theorem 99.22.*

Compute $\delta\mathcal{L}$ two ways. Directly from the chain rule, with summation over a,

$$\delta\mathcal{L} = \varepsilon\left[\frac{\partial \mathcal{L}}{\partial \phi_a} F_a + \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_a)} \partial_\mu F_a\right] = \varepsilon\left[\left(\frac{\partial \mathcal{L}}{\partial \phi_a} - \partial_\mu \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_a)}\right) F_a + \partial_\mu\left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_a)} F_a\right)\right].$$

On a solution the first bracket vanishes by Equation (99.16). By hypothesis the same $\delta\mathcal{L}$ equals $\varepsilon \partial_\mu K^\mu$, so $\partial_\mu\big(\partial\mathcal{L}/\partial(\partial_\mu\phi_a)\, F_a - K^\mu\big) = 0$. Integrating $\partial_\mu j^\mu = 0$ over a spatial slice, $\partial_t \int \mathrm{d}^3x\, j^0/c = -\int \mathrm{d}^3x\, \nabla\cdot\boldsymbol{j}$, which is a surface term at infinity. □

In the quantum theory the conserved charge does more than sit there: it generates the symmetry it came from.

Proposition 99.23 (Charges generate their symmetries)**.** *For the* U(1) *charge of Equation (99.55),*

$$[Q, \phi(x)] = -\phi(x)\,, \qquad \mathrm{e}^{\mathrm{i}\theta Q}\phi(x)\,\mathrm{e}^{-\mathrm{i}\theta Q} = \mathrm{e}^{-\mathrm{i}\theta}\phi(x)\,, \tag{99.64}$$

which is Equation (99.53).

Rests on Equations (99.53), (99.58) and (99.60).

Proof. *Derives Proposition 99.23.*

From Equation (99.60) and Equation (99.58), $[:Q:, a(\boldsymbol{k})] = -a(\boldsymbol{k})$ and $\left[:Q:, b^\dagger(\boldsymbol{k})\right] = -b^\dagger(\boldsymbol{k})$; both coefficients in Equation (99.56) are therefore lowered by one unit, giving the first equation. The second follows from the identity $\mathrm{e}^{X} Y \mathrm{e}^{-X} = Y + [X,Y] + \frac{1}{2!}[X,[X,Y]] + \cdots$ with $X = \mathrm{i}\theta Q$, which here terminates into an exponential series because $[Q,\cdot]$ acts as multiplication by -1 on ϕ. □

Applying Theorem 99.22 to spacetime translations $\phi_a(x) \mapsto \phi_a(x + \varepsilon e)$, for which $\delta\mathcal{L} = \varepsilon\, e^\nu \partial_\nu \mathcal{L}$ is a total divergence with $K^\mu = e^\mu \mathcal{L}$, gives four conserved currents assembled into the canonical energy–momentum tensor

$$T^{\mu\nu} = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_a)} \partial^\nu \phi_a - \eta^{\mu\nu}\mathcal{L}\,, \qquad \partial_\mu T^{\mu\nu} = 0\,, \qquad \hat{P}^\nu = \frac{1}{c}\int \mathrm{d}^3x\, T^{0\nu}\,, \tag{99.65}$$

of which $\hat{P}^0 = H/c$ and $\hat{P}^i = \boldsymbol{P}$ recover Equations (99.33) and (99.34): for the real scalar, $T^{00} = \mathcal{H}$ of Equation (99.21) and $T^{0i} = c^{-1}\partial_t\phi\,\partial^i\phi = -c\pi\partial_i\phi$. Lorentz transformations $\delta\phi = \frac{1}{2}\omega_{\rho\sigma}(x^\rho\partial^\sigma - x^\sigma\partial^\rho)\phi$ give six more currents,

$$\mathcal{M}^{\mu\rho\sigma} = x^\rho T^{\mu\sigma} - x^\sigma T^{\mu\rho} \quad \text{(spin 0)}\,, \qquad \hat{J}^{\rho\sigma} = \frac{1}{c}\int \mathrm{d}^3x\, \mathcal{M}^{0\rho\sigma}\,, \tag{99.66}$$

which for fields carrying spin acquire an extra term from the rotation of the field components themselves — the term whose separation from the orbital piece is the Belinfante–Rosenfeld problem [850] treated in Chapter 64.

99.4.2 The Poincaré algebra and the labels of a particle

The ten charges $\hat{P}^\mu, \hat{J}^{\rho\sigma}$ of Equations (99.65) and (99.66) are Hermitian operators on Fock space, and they close on the Poincaré algebra Equations (94.5) to (94.7) of Proposition 94.8, with an $\mathrm{i}\hbar$ in every bracket because the generators are the physical momentum and angular momentum and not the dimensionless differential operators of Section 14.3.2. That the same algebra appears here as an algebra of field integrals, and in Chapter 94 as an algebra of abstract generators, is the whole content of the phrase "the field theory is Poincaré invariant".

Proposition 99.24 (Translations act correctly on the field)**.** $\left[\hat{P}^\mu, \phi(x)\right] = -\mathrm{i}\hbar\,\partial^\mu\phi(x)$*, so that* $\mathrm{e}^{\mathrm{i}a\cdot\hat{P}/\hbar}\phi(x)\mathrm{e}^{-\mathrm{i}a\cdot\hat{P}/\hbar} = \phi(x+a)$*, and* $\left[\hat{P}^\mu, \hat{P}^\nu\right] = 0$*.*

Rests on Proposition 99.6 and Equations (99.31) and (99.34).

Proof. *Derives Proposition 99.24.*

For $\mu = 0$, where $\hat{P}^0 = H/c$, this is the Heisenberg equation $\mathrm{i}\hbar\,\partial_t\phi = [\phi, H]$ of Proposition 99.6. For the spatial components use Equation (99.34) and Equation (99.31): $[\boldsymbol{P}, a(\boldsymbol{k})] = -\hbar\boldsymbol{k}\,a(\boldsymbol{k})$ and $\left[\boldsymbol{P}, a^\dagger(\boldsymbol{k})\right] = +\hbar\boldsymbol{k}\,a^\dagger(\boldsymbol{k})$, so in Equation (99.29) each term acquires the factor that $-\mathrm{i}\hbar\nabla$ would produce on its exponential. The two components commute because both are diagonal in $\boldsymbol{k}$. □

Proposition 99.25 (The one-quantum sector is an irreducible representation)**.** *On* $\mathcal{H}_1 = \mathrm{Sym}^1(\mathcal{H}_1)$ *of the free real scalar field,*

$$\hat{P}^2 = \hat{P}_\mu\hat{P}^\mu = m^2c^2\,\mathbb{1}\,, \qquad \hat{W}^2 = \hat{W}_\mu\hat{W}^\mu = 0\,, \tag{99.67}$$

with $\hat{W}_\mu = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma}\hat{J}^{\nu\rho}\hat{P}^\sigma$ *the Pauli–Lubanski vector of Definition 94.15, and the representation is irreducible: the free real scalar field describes one species of particle of mass* m *and spin* 0*.*

Rests on Proposition 99.24, Theorem 94.18, and Definition 94.15.

Proof. *Derives Proposition 99.25.*

By Proposition 99.24 the state $a^\dagger(\boldsymbol{k})\,|0\rangle$ is a simultaneous eigenstate of $\hat{P}^\mu$ with eigenvalue $\hbar k^\mu = (\hbar\omega_{\boldsymbol{k}}/c, \hbar\boldsymbol{k})$, so $\hat{P}^2$ has eigenvalue $\hbar^2k^2 = \hbar^2\kappa^2 = m^2c^2$ by Equations (99.2) and (99.28) — which is Equation (94.14), the definition of mass. For $\hat{W}^2$, evaluate in the rest frame, legitimate because $\hat{W}^2$ is a Casimir (Theorem 94.18) and therefore frame independent: there $\hat{P}^\mu = (mc, \mathbf{0})$ and $\hat{W}_i = -mc\,\hat{J}_i$, $\hat{W}_0 = 0$, so $\hat{W}^2 = -m^2c^2\hat{\boldsymbol{J}}^2$. For a scalar field $\hat{J}^{\rho\sigma}$ is purely orbital by Equation (99.66), so $\hat{\boldsymbol{J}}$ annihilates a zero-momentum one-quantum state and $\hat{W}^2 = 0$: comparing with $\hat{W}^2 = -m^2c^2\hbar^2 s(s+1)$ of Equation (94.25) gives $s = 0$. Irreducibility: the states $a^\dagger(\boldsymbol{k})\,|0\rangle$ for $\boldsymbol{k}$ on the mass shell form a single orbit of the proper orthochronous Lorentz group (Proposition 94.21), and the little group of a massive momentum is SU(2) (Proposition 94.24), acting here on a one-dimensional space, so there is no invariant subspace. □

Remark 99.26 (Why the classification matters, and where it stops). Wigner's theorem [849] makes "particle" a group-theoretic notion: a species is a unitary irreducible representation of the Poincaré group, labelled by the two invariants of Theorem 94.18, and the free fields of this chapter are devices for realizing those representations on a Fock space where quanta can be created and destroyed. The classification is complete for free particles and is silent about interacting ones: an interacting theory's asymptotic states are again irreducible representations (that is what makes scattering theory possible, Chapter 86), but a confined quark, which never appears asymptotically, is not classified by it at all (Chapter 102). The massless case is different in kind and not merely in the value of m: there $\hat{P}^2 = \hat{W}^2 = 0$ forces $\hat{W}^\mu \propto \hat{P}^\mu$, the proportionality constant is the helicity, and a massless particle of nonzero spin has two states rather than $2s+1$ (Proposition 94.32) — which is why the photon of Section 99.5.2 has two polarizations and the massive vector of Section 99.5.3 has three.

99.5 Fermions and gauge fields

The scalar field is the whole method in miniature, but it is not any field that occurs in Nature. The matter fields of the Standard Model carry spin 1/2, and the force fields carry spin 1 and a gauge redundancy. Each brings a genuinely new feature, and in both cases the new feature is forced rather than chosen.

99.5.1 The Dirac field and anticommutation

The classical Dirac field is a four-component object $\psi_a(x)$ with Lagrangian density

$$\mathcal{L} = \bar{\psi}\left(\mathrm{i}\hbar c\,\gamma^\mu\partial_\mu - mc^2\right)\psi\,, \qquad \bar{\psi} = \psi^\dagger\gamma^0\,, \tag{99.68}$$

with $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}\mathbb{1}$, $(\gamma^0)^\dagger = \gamma^0$ and $(\gamma^i)^\dagger = -\gamma^i$, exactly as constructed in Chapter 93 and used in Equation (100.5). The field carries $[\psi] = 1/\mathrm{m}^{3/2}$ (Definition 100.2), so that $\psi^\dagger\psi$ is a number density and both terms of Equation (99.68) are energy densities. The Euler–Lagrange equation is $(\mathrm{i}\hbar\gamma^\mu\partial_\mu - mc)\psi = 0$, and every component separately satisfies $(\Box + \kappa^2)\psi_a = 0$, obtained by applying $(\mathrm{i}\hbar\gamma^\nu\partial_\nu + mc)$ and using the Clifford algebra.

The conjugate momentum is peculiar. Since $\mathcal{L} \supset \mathrm{i}\hbar c\,\bar{\psi}\gamma^0 c^{-1}\partial_t\psi = \mathrm{i}\hbar\,\psi^\dagger\partial_t\psi$,

$$\pi_a = \frac{\partial\mathcal{L}}{\partial\dot{\psi}_a} = \mathrm{i}\hbar\,\psi_a^\dagger\,, \tag{99.69}$$

that is, the momentum is the field's own conjugate rather than something new: Equation (99.68) is first order in time, so ψ and $\psi^\dagger$ are not independent initial data but together

constitute the phase space. This is a constrained system in the sense of Chapter 26, and the canonical bracket that survives the constraint analysis is the one between ψ and $\psi^\dagger$ alone.

The question this subsection answers is whether that bracket should be a commutator or an anticommutator. The classical theory says nothing: the canonical procedure of Chapter 22 produces a Poisson bracket, and both promotions reduce to it as $\hbar \to 0$. Nature decides, through the requirement that the energy be bounded below.

Definition 99.27 (Mode expansion of the Dirac field)**.** Write $p^\mu = \hbar k^\mu$ and let $u_s(p), v_s(p)$, $s = 1, 2$, be the positive- and negative-frequency spinor solutions of Chapter 93, normalized as in Proposition 100.14 by

$$\bar u_s(p) u_{s'}(p) = 2mc\,\delta_{ss'}\,, \qquad u_s^\dagger(p) u_{s'}(p) = \frac{2E_p}{c}\,\delta_{ss'}\,, \qquad \sum_s u_s(p)\bar u_s(p) = \gamma^\mu p_\mu + mc\,, \tag{99.70}$$

with $E_p = \hbar\omega_{\boldsymbol{k}}$. The negative-frequency spinors are *not* normalized by copying these: the Lorentz scalar changes sign while the norm, being a density, does not, so

$$\bar v_s(p) v_{s'}(p) = -2mc\,\delta_{ss'}\,, \qquad v_s^\dagger(p) v_{s'}(p) = \frac{2E_p}{c}\,\delta_{ss'}\,, \qquad \sum_s v_s(p)\bar v_s(p) = \gamma^\mu p_\mu - mc\,. \tag{99.71}$$

The two are consistent: tracing the completeness relations gives $\sum_s \bar u_s u_s = +4mc$ and $\sum_s \bar v_s v_s = -4mc$. Then

$$\psi(x) = \sum_{s=1,2} \int \frac{\mathrm{d}^3k}{(2\pi)^3} \sqrt{\frac{c}{2\hbar\omega_{\boldsymbol{k}}}} \left[b_s(\boldsymbol{k}) u_s(p)\, \mathrm{e}^{-\mathrm{i}k\cdot x} + d_s^\dagger(\boldsymbol{k}) v_s(p)\, \mathrm{e}^{\mathrm{i}k\cdot x} \right]. \tag{99.72}$$

The normalization is fixed by dimensions: $[\sqrt{c/2\hbar\omega}] = \mathrm{s}^{1/2}/\mathrm{kg}^{1/2}\,\mathrm{m}^{1/2}$ and the spinors carry the inverse of that by Equation (99.70), leaving $[\psi] = 1/\mathrm{m}^{3/2}$ once $[\mathrm{d}^3k]$ and $[b] = \mathrm{m}^{3/2}$ are counted.

Theorem 99.28 (Anticommutation is forced by positivity of the energy)**.** *Let the mode operators of Equation* (99.72) *obey either commutation or anticommutation relations of the canonical form. Then the Hamiltonian is*

$$H = \sum_s \int \frac{\mathrm{d}^3k}{(2\pi)^3}\, \hbar\omega_{\boldsymbol{k}} \left[b_s^\dagger(\boldsymbol{k}) b_s(\boldsymbol{k}) - d_s(\boldsymbol{k}) d_s^\dagger(\boldsymbol{k}) \right], \tag{99.73}$$

which is unbounded below if $\left[d_s(\boldsymbol{k}), d_{s'}^\dagger(\boldsymbol{k}') \right] = \delta_{ss'}(2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}')$*, and equal to a non-negative operator plus a constant if*

$$\left\{ b_s(\boldsymbol{k}), b_{s'}^\dagger(\boldsymbol{k}') \right\} = \left\{ d_s(\boldsymbol{k}), d_{s'}^\dagger(\boldsymbol{k}') \right\} = \delta_{ss'}(2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}')\,, \tag{99.74}$$

all other brackets vanishing. The corresponding equal-time relation for the field is

$$\left\{ \psi_a(t,\boldsymbol{x}), \psi_b^\dagger(t,\boldsymbol{y}) \right\} = \delta_{ab}\,\delta^3(\boldsymbol{x}-\boldsymbol{y})\,, \qquad \{\psi_a, \psi_b\} = \left\{ \psi_a^\dagger, \psi_b^\dagger \right\} = 0\,. \tag{99.75}$$

Rests on Equations (99.68)*,* (99.69) *and* (99.72)*.*

Proof. *Derives Theorem 99.28.*

From Equations (99.68) and (99.69) the Hamiltonian density is $\mathcal{H} = \pi\dot\psi - \mathcal{L} = \psi^\dagger\left(-\mathrm{i}\hbar c\,\gamma^0\gamma^i\partial_i + mc^2\gamma^0\right)\psi$, and the operator in brackets acting on a solution $\mathrm{e}^{\mp\mathrm{i}k\cdot x}$ returns $\pm\hbar\omega_{\boldsymbol{k}}$ — this is just the Dirac equation solved for $\mathrm{i}\hbar\partial_t\psi$. Substituting Equation (99.72) and integrating over $\boldsymbol{x}$, the $\mathrm{e}^{\pm 2\mathrm{i}\omega t}$ cross terms vanish by $u_s^\dagger(p)v_{s'}(\tilde p) = 0$ with $\tilde p = (p^0, -\boldsymbol{p})$, and the diagonal terms are weighted by $u^\dagger u = v^\dagger v = 2E_p/c$, which cancels the $c/2\hbar\omega$ of the

normalization and leaves Equation (99.73), the minus sign on the d term coming from the sign of the eigenvalue on the negative-frequency solutions.

Now compare the two options. With commutators, $-d\,d^\dagger = -d^\dagger d - (2\pi)^3\delta^3(\mathbf{0})$, so

$$H = \sum_s \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\hbar\omega_{\boldsymbol{k}}\left[b^\dagger b - d^\dagger d\right] + \text{const}\,,$$

and $d^\dagger$ applied repeatedly lowers the energy without limit: the spectrum is unbounded below, the vacuum is not the ground state, and no state is stable. With anticommutators, $-d\,d^\dagger = +d^\dagger d - (2\pi)^3\delta^3(\mathbf{0})$ by Equation (99.74), so

$$:H: = \sum_s \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\hbar\omega_{\boldsymbol{k}}\left[b_s^\dagger(\boldsymbol{k})b_s(\boldsymbol{k}) + d_s^\dagger(\boldsymbol{k})d_s(\boldsymbol{k})\right]\,, \tag{99.76}$$

a sum of non-negative operators, since $\left\{d, d^\dagger\right\} \geq 0$ makes $d^\dagger d$ positive semi-definite. Equation (99.75) follows by substituting Equation (99.72) and using the completeness relation

$$\sum_s u_s(p)u_s^\dagger(p) + \sum_s v_s(\tilde{p})v_s^\dagger(\tilde{p}) = (\gamma^\mu p_\mu + mc)\,\gamma^0 + (\gamma^\mu \tilde{p}_\mu - mc)\,\gamma^0 = 2p^0\,\mathbb{1} = \frac{2\hbar\omega_{\boldsymbol{k}}}{c}\,\mathbb{1}\,,$$

which cancels the normalization factor and leaves $\int \mathrm{d}^3k\,(2\pi)^{-3}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})}\delta_{ab}$. That Equation (99.75) is the correct canonical relation is confirmed by Equation (99.69): the naive prescription $\{\psi, \pi\} = \mathrm{i}\hbar\delta^3$ with $\pi = \mathrm{i}\hbar\psi^\dagger$ gives exactly $\left\{\psi, \psi^\dagger\right\} = \delta^3$, without any $\hbar$ surviving — the field is dimensionless in the appropriate sense, and the $\hbar$ has been absorbed into the constraint. □

The choice is therefore not a convention. Given Equation (99.68), commutators produce a theory with no ground state, which is not a theory of anything; anticommutators produce a positive Hamiltonian. This is one half of the spin–statistics connection; the other half — that a scalar quantized with anticommutators is equally impossible, for a different reason — is Section 99.7.2.

Phenomenon 99.29 (The Pauli exclusion principle)**.** No two quanta of a Dirac field occupy the same mode: a state containing two electrons of identical momentum and spin does not exist. The consequences are the periodic table (Chapter 88) and the degeneracy pressure that supports white dwarfs and neutron stars (Chapter 49).

Rests on Theorem 99.28 and Equation (99.74)*.*

Derivation. Derives Phenomenon 99.29.

Setting $s = s'$ and $\boldsymbol{k}' = \boldsymbol{k}$ in the second family of Equation (99.74) gives $\left\{b_s^\dagger(\boldsymbol{k}), b_s^\dagger(\boldsymbol{k})\right\} = 2\left(b_s^\dagger(\boldsymbol{k})\right)^2 = 0$, so

$$\left(b_s^\dagger(\boldsymbol{k})\right)^2 = 0 \tag{99.77}$$

identically as an operator. Any state containing two quanta in the same mode would be $b_s^\dagger(\boldsymbol{k})b_s^\dagger(\boldsymbol{k})\,|\cdots\rangle$, which is the zero vector, not a state. Equivalently, the number operator for one mode, $n = b^\dagger b$, satisfies $n^2 = b^\dagger b b^\dagger b = b^\dagger(1 - b^\dagger b)b = n$ by Equations (99.74) and (99.77), so its eigenvalues satisfy $n^2 = n$ and are 0 or 1. The antisymmetry of the multi-quantum wave function follows from the same relation with distinct labels: $b_s^\dagger(\boldsymbol{k})b_{s'}^\dagger(\boldsymbol{k}')\,|0\rangle = -b_{s'}^\dagger(\boldsymbol{k}')b_s^\dagger(\boldsymbol{k})\,|0\rangle$. □

Phenomenon 99.29 is worth pausing on. In Chapter 83 the antisymmetry of the fermionic wave function is a postulate, added to quantum mechanics because it is observed. Here it is a theorem, following from Equation (99.74), which in turn follows from requiring the energy to be bounded below. Nothing was assumed about identical particles at all — there are no particles in Equation (99.68), only a field, and the quanta are identical because they are excitations of one field.

Remark 99.30 (The Dirac sea is not needed and is not used). Equation (99.76) contains two non-negative terms and a vacuum defined by $b_s(\boldsymbol{k})\,|0\rangle = d_s(\boldsymbol{k})\,|0\rangle = 0$: no level is occupied, and the positron is a d quantum, not a hole. The hole picture of Section 99.1.1 survives only as a mnemonic. It is not merely superfluous: the constant discarded in passing from Equation (99.73) to Equation (99.76) is $-\sum_s \int_{\boldsymbol{k}} \hbar\omega_{\boldsymbol{k}}(2\pi)^3\delta^3(\boldsymbol{0})$, negative where the bosonic zero-point energy Equation (99.46) was positive, so a fermion field's vacuum energy has the opposite sign to a boson's. That sign is a genuine structural fact — it is what makes the cancellation of Equation (99.49) conceivable in principle — and it is invisible in the sea picture, where the negative energy is attributed to occupied states.

99.5.2 The electromagnetic field and gauge redundancy

The electromagnetic field is the first case in which the canonical procedure does not simply run. The Lagrangian density is

$$\mathcal{L} = -\frac{1}{4\mu_0}F_{\mu\nu}F^{\mu\nu} = \frac{\varepsilon_0}{2}\boldsymbol{E}^2 - \frac{1}{2\mu_0}\boldsymbol{B}^2\,, \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu\,, \tag{99.78}$$

with $[A^\mu] = \mathrm{V\,s/m}$ and $[F_{\mu\nu}] = \mathrm{T}$ as in Definition 100.2, and $\boldsymbol{E} = -\nabla A^0 c - \partial_t \boldsymbol{A}$, $\boldsymbol{B} = \nabla\times\boldsymbol{A}$ in terms of $A^\mu = (A^0, \boldsymbol{A})$. The conjugate momenta are

$$\pi^i = \frac{\partial\mathcal{L}}{\partial\dot{A}_i} = +\varepsilon_0 E^i\,, \qquad \pi^0 = \frac{\partial\mathcal{L}}{\partial\dot{A}_0} = 0\,. \tag{99.79}$$

The sign is the index placement and nothing else: differentiating Equation (99.78) with respect to $\dot{A}^i$, the contravariant component appearing in $\boldsymbol{E}$, gives $-\varepsilon_0 E^i$, and $A_i = -A^i$ then flips it. Both indices are read the same way in every commutator below, so that $\boldsymbol{E}$ appears there through its covariant components $E_i = -E^i$.

The vanishing of π^0 is not an accident of a bad variable choice and cannot be repaired by one. $F_{\mu\nu}$ is antisymmetric, so $\dot{A}_0$ appears nowhere in Equation (99.78), and $\pi^0 = 0$ holds identically rather than dynamically: it is a *primary constraint*, the prototype of the structures analysed by Dirac's algorithm in Chapter 26 and the difficulty Heisenberg and Pauli met first [626, 656]. Consistency, $\partial_t\pi^0 = 0$, then generates the secondary constraint $\nabla\cdot\boldsymbol{\pi} = 0$, that is Gauss's law $\nabla\cdot\boldsymbol{E} = 0$ in vacuum. The canonical commutator $[A_0, \pi^0] = \mathrm{i}\hbar\delta^3$ is inconsistent with $\pi^0 = 0$, so Postulate 99.5 cannot be imposed on all four components. The underlying reason is gauge invariance: A_μ and $A_\mu + \partial_\mu\Lambda$ describe the same $F_{\mu\nu}$ and the same physics, so A_μ carries more components than there are degrees of freedom, and the extra ones have no dynamics to quantize.

Two treatments are standard, and they differ in which of covariance and positivity is given up first.

Coulomb gauge. Impose $\nabla\cdot\boldsymbol{A} = 0$. Gauss's law then makes A^0 obey $\nabla^2 A^0 = 0$ in vacuum, so $A^0 = 0$ with sensible boundary conditions: no dynamics, and two independent components of $\boldsymbol{A}$ remain. This is what Born, Heisenberg and Jordan did in 1926 and what Dirac [572] systematized. The expansion is

$$\boldsymbol{A}(x) = \sum_{\lambda=1,2}\int \frac{\mathrm{d}^3k}{(2\pi)^3}\sqrt{\frac{\hbar\mu_0 c^2}{2\omega_{\boldsymbol{k}}}}\left[\boldsymbol{\varepsilon}_\lambda(\boldsymbol{k})a_\lambda(\boldsymbol{k})\,\mathrm{e}^{-\mathrm{i}k\cdot x} + \mathrm{h.c.}\right], \qquad \omega_{\boldsymbol{k}} = c\,|\boldsymbol{k}|\,, \tag{99.80}$$

with $\boldsymbol{\varepsilon}_\lambda\cdot\boldsymbol{k} = 0$ enforcing the gauge condition, $\left[a_\lambda(\boldsymbol{k}), a^\dagger_{\lambda'}(\boldsymbol{k}')\right] = \delta_{\lambda\lambda'}(2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}')$, and the factor μ_0 making the dimensions come out: $[\sqrt{\hbar\mu_0 c^2/2\omega}] = \mathrm{V\,s\,m^{1/2}}$, which with $[\mathrm{d}^3k]$ and $[a]$ gives $\mathrm{V\,s/m}$. The price is that the gauge condition is not Lorentz invariant, and that the equal-time commutator is modified: since $\nabla\cdot\boldsymbol{A} = 0$ holds as an operator equation, $[A_i, \pi_j]$

cannot be proportional to $\delta_{ij}\delta^3$, whose divergence does not vanish. The correct relation is

$$[A_i(t,\boldsymbol{x}),E_j(t,\boldsymbol{y})] = -\frac{\mathrm{i}\hbar}{\varepsilon_0}\,\delta^{\mathrm{T}}_{ij}(\boldsymbol{x}-\boldsymbol{y})\,, \qquad \delta^{\mathrm{T}}_{ij}(\boldsymbol{r}) = \delta_{ij}\delta^3(\boldsymbol{r}) + \partial_i\partial_j\frac{1}{4\pi\,|\boldsymbol{r}|}\,, \tag{99.81}$$

the transverse delta, which is the projector onto divergence-free vector fields and satisfies $\partial_i\delta^{\mathrm{T}}_{ij}=0$. Its long-range second term is the Coulomb interaction in disguise: in this gauge the instantaneous Coulomb field is not a dynamical variable but part of the constraint, which is why the gauge is convenient for bound states and awkward for scattering.

Derivation. *Derives Equation (99.81).*

Only the transverse part of $\boldsymbol{A}$ is a dynamical variable, so the canonical bracket can only be the transverse part of $\mathrm{i}\hbar\,\delta_{ij}\delta^3$; the whole content of Equation (99.81) is which distribution that is. Work in momentum space, writing $\delta^{\mathrm{T}}_{ij}(\boldsymbol{r}) = \int \mathrm{d}^3k\,(2\pi)^{-3}\,\widetilde{\delta}^{\mathrm{T}}_{ij}(\boldsymbol{k})\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}$. The gauge condition $\nabla\cdot\boldsymbol{A}=0$ is $\boldsymbol{k}\cdot\widetilde{\boldsymbol{A}}=0$, so $\widetilde{\delta}^{\mathrm{T}}_{ij}$ must annihilate every longitudinal vector and act as the identity on every transverse one. A symmetric tensor built from δ_{ij} and $\boldsymbol{k}$ with those two properties is unique:

$$\widetilde{\delta}^{\mathrm{T}}_{ij}(\boldsymbol{k}) = \delta_{ij} - \frac{k_ik_j}{\boldsymbol{k}^2}\,, \tag{99.82}$$

which indeed obeys $k_i\widetilde{\delta}^{\mathrm{T}}_{ij}=0$ and $\widetilde{\delta}^{\mathrm{T}}_{ik}\widetilde{\delta}^{\mathrm{T}}_{kj}=\widetilde{\delta}^{\mathrm{T}}_{ij}$, and whose trace is 2: the two transverse polarizations of Equation (99.80).

Now transform back term by term. The δ_{ij} gives $\delta_{ij}\delta^3(\boldsymbol{r})$. For the second term use $\int \mathrm{d}^3k\,(2\pi)^{-3}\boldsymbol{k}^{-2}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} = 1/(4\pi\,|\boldsymbol{r}|)$, which is the Coulomb Green function, together with the fact that $\partial_i\partial_j$ acting on $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}$ brings down $(\mathrm{i}k_i)(\mathrm{i}k_j)=-k_ik_j$. Hence $-k_ik_j/\boldsymbol{k}^2$ transforms to $+\partial_i\partial_j(4\pi\,|\boldsymbol{r}|)^{-1}$, and the sign in Equation (99.81) is a plus.

The position-space statement confirms it independently: $\partial_i\delta^{\mathrm{T}}_{ij} = \partial_j\delta^3(\boldsymbol{r}) + \partial_j\nabla^2(4\pi\,|\boldsymbol{r}|)^{-1} = \partial_j\delta^3(\boldsymbol{r}) - \partial_j\delta^3(\boldsymbol{r}) = 0$, whereas a minus sign would leave $2\partial_j\delta^3(\boldsymbol{r})$. Carrying out the derivatives gives the explicit form $\delta^{\mathrm{T}}_{ij}(\boldsymbol{r}) = \frac{2}{3}\delta_{ij}\delta^3(\boldsymbol{r}) + (3\hat r_i\hat r_j-\delta_{ij})/(4\pi r^3)$, whose second piece has the angular structure of a point dipole's field. Finally, the prefactor: the canonical relation is $\left[A_i,\pi^j\right] = \mathrm{i}\hbar\,\delta^{\mathrm{T}}_{ij}$ with $\pi^j=\varepsilon_0E^j$ by Equation (99.79), and $E_j=-E^j$ supplies the minus sign shown. □

Covariant (Gupta–Bleuler) quantization. Add to Equation (99.78) the gauge-fixing term $-(2\mu_0\xi)^{-1}(\partial_\mu A^\mu)^2$, which makes $\dot A_0$ appear and removes the constraint by hand (Equation (100.16)). All four polarizations are then quantized covariantly, at the cost that the timelike one has the wrong sign:

$$\left[a_\lambda(\boldsymbol{k}),a^\dagger_{\lambda'}(\boldsymbol{k}')\right] = -\eta_{\lambda\lambda'}(2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}')\,, \tag{99.83}$$

so that $\langle 0|\,a_0a_0^\dagger\,|0\rangle<0$ and the state space has an indefinite metric — there are states of negative norm, which cannot be probabilities.

Derivation pending #131

The indefinite-metric mode algebra of covariant quantization, displayed just above: that the gauge-fixing term makes all four components of the potential dynamical, that the covariant equal-time bracket $[A_\mu,\pi^\nu]$ then delivers $-\eta_{\lambda\lambda'}$ where the Coulomb-gauge algebra has $\delta_{\lambda\lambda'}$, and hence that the timelike mode has negative norm. It needs a basis of four polarization vectors orthonormal in the Minkowski metric, which this chapter does not set up, so it belongs in the long-proof appendix.

Gupta and Bleuler [953, 958] showed that the theory is nevertheless consistent if the physical states are restricted by the subsidiary condition

$$(\partial_\mu A^\mu)^{(+)}\,|\mathrm{phys}\rangle = 0\,, \tag{99.84}$$

the superscript denoting the positive-frequency (annihilation) part. On the subspace so defined the norm is non-negative, the timelike and longitudinal quanta enter only in combinations of zero norm, and their contributions to every observable cancel, leaving the two transverse polarizations of Equation (99.80).

Derivation pending #132

Consistency of the Gupta–Bleuler subsidiary condition displayed just above: that the subspace it selects carries a non-negative norm, that timelike and longitudinal quanta occur in it only in zero-norm combinations, and that those combinations contribute nothing to any observable, so that the counting reduces to the two transverse polarizations. It runs through the mode decomposition of the gauge-fixed field and belongs in the long-proof appendix, next to the indefinite-metric algebra above.

The gauge parameter ξ survives in the propagator (Equation (100.17)) and must cancel from every observable, a requirement that is one of the strongest internal checks on a QED calculation (Chapter 100).

That the photon has two states and not three is not a choice of gauge: it is Remark 99.26. A massless particle is classified by helicity, and the two helicities ± 1 are the whole content of the representation; the third polarization of a massive spin-1 particle has nowhere to go as $m \to 0$. Gauge redundancy is the field-theoretic expression of that representation-theoretic fact, and this is the sense in which the difficulty of the constraints is not a technical nuisance but a consequence of masslessness.

What has been suppressed here, and belongs to Chapter 122, is the phenomenology of the photon number operator: the coherent states that are eigenstates of a_λ and approximate a classical wave, the photon-counting statistics that distinguish them from thermal and number states, and the experiments that measure those statistics directly.

99.5.3 Massive vector fields and the general spin case

Adding a mass term to Equation (99.78) gives Proca's theory [785],

$$\mathcal{L} = -\frac{1}{4\mu_0} F_{\mu\nu}F^{\mu\nu} + \frac{\kappa^2}{2\mu_0} A_\mu A^\mu , \qquad \kappa = \frac{m_V c}{\hbar} , \tag{99.85}$$

whose field equation is $\partial_\mu F^{\mu\nu} + \kappa^2 A^\nu = 0$. The mass term breaks gauge invariance, and the constraint structure changes character.

Proposition 99.31 (Three polarizations, and no gauge freedom)**.** *For $\kappa \neq 0$ the Proca field satisfies $\partial_\mu A^\mu = 0$ as a consequence of its equation of motion, not as a gauge choice; each component then obeys $(\Box + \kappa^2)A^\nu = 0$, and the physical polarizations number three.*

Rests on Equation (99.85) and Proposition 94.24.

Proof. *Derives Proposition 99.31.*

Take ∂_ν of the field equation. The first term vanishes identically, $\partial_\nu\partial_\mu F^{\mu\nu} = 0$, being a symmetric derivative contracted with an antisymmetric tensor, so $\kappa^2 \partial_\nu A^\nu = 0$, and for $\kappa \neq 0$ this forces $\partial_\nu A^\nu = 0$. Substituting back, $\partial_\mu(\partial^\mu A^\nu - \partial^\nu A^\mu) = \Box A^\nu$, so $(\Box + \kappa^2)A^\nu = 0$. Four components subject to one covariant condition leave three; equivalently, the little group of a massive momentum is SU(2) (Proposition 94.24) and the spin-1 representation is three-dimensional. □

Since there is no gauge freedom, $\pi^0 = 0$ is still a constraint but it now *determines* A^0 in terms of the other variables rather than generating a redundancy, and the canonical quantization of the three remaining polarizations proceeds as for three scalars. The consequence

that matters downstream is the propagator. Inverting the quadratic form of Equation (99.85) in the same way as Proposition 100.8 does for the photon gives

$$\widetilde{D}^{\mu\nu}(k) \propto \frac{-\eta^{\mu\nu} + k^{\mu}k^{\nu}/\kappa^{2}}{k^{2} - \kappa^{2} + \mathrm{i}\epsilon}\,, \tag{99.86}$$

and the numerator's second term does not fall off at large k: the propagator tends to a constant $\propto k^{\mu}k^{\nu}/(\kappa^{2}k^{2})$ rather than to zero. In a theory in which massive vector bosons are exchanged between currents that are not conserved, amplitudes then grow with energy without limit, violating the unitarity bound at a calculable scale. That growth is the argument — not an aesthetic preference for symmetry — that forces the weak interaction to be a spontaneously broken gauge theory rather than a Proca theory with arbitrary couplings, and it is carried out quantitatively in Chapter 101.

For spin greater than 1 the pattern continues and worsens. A free field of spin s can be constructed as a symmetric traceless tensor subject to auxiliary conditions, and Fierz and Pauli [834] showed that the conditions can be imposed consistently for the free field; but coupling such a field to an external electromagnetic field generically destroys the constraint algebra, producing either fewer propagating modes than the spin requires or propagation faster than light. The observed spectrum contains no elementary particle of spin greater than 1 — the hadronic resonances of higher spin are composite — and the obstructions are summarized in Chapter 97.

99.6 Propagators

A propagator is the vacuum expectation value of a time-ordered product of two fields. It is the object that perturbation theory is built from, it is a Green function of the classical field equation, and the particular Green function it is — which of the possible contours — is fixed by the requirement that positive frequencies propagate forward in time and negative frequencies backward. This section derives it, and derives it by contour integration, using the residue calculus of Section 8.7 rather than rebuilding it.

99.6.1 Time ordering and the Feynman propagator

Definition 99.32 (Time ordering and the two-point functions)**.** For bosonic fields,

$$T\{\phi(x)\phi(y)\} = \theta(x^{0}-y^{0})\,\phi(x)\phi(y) + \theta(y^{0}-x^{0})\,\phi(y)\phi(x)\,, \tag{99.87}$$

and the Feynman propagator, Wightman function and Pauli–Jordan function of the free real scalar are

$$D_{F}(x-y) = \langle 0|\,T\{\phi(x)\phi(y)\}\,|0\rangle\,, \tag{99.88}$$

$$D^{+}(x-y) = \langle 0|\,\phi(x)\phi(y)\,|0\rangle\,, \tag{99.89}$$

$$\Delta(x-y) = \frac{1}{\mathrm{i}\hbar}\,\langle 0|\,[\phi(x),\phi(y)]\,|0\rangle\,. \tag{99.90}$$

Lemma 99.33 (The Wightman function)**.**

$$D^{+}(x-y) = \int \frac{\mathrm{d}^{3}k}{(2\pi)^{3}}\,\frac{\hbar c^{2}}{2\omega_{\boldsymbol{k}}}\,\mathrm{e}^{-\mathrm{i}k\cdot(x-y)} = \hbar c\int \frac{\mathrm{d}^{4}k}{(2\pi)^{3}}\,\delta\!\left(k^{2}-\kappa^{2}\right)\theta(k^{0})\,\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}\,, \tag{99.91}$$

of SI dimension $\mathrm{J/m}$*; it solves* $(\Box+\kappa^{2})\,D^{+}=0$ *and is manifestly Lorentz invariant.*
Rests on Equations (99.29)*,* (99.31) *and* (99.37)*.*

Proof. *Derives Lemma 99.33.*

Substitute Equation (99.29) twice. Of the four terms only $a(\boldsymbol{k})a^\dagger(\boldsymbol{k}')$ survives between vacua, because $a\,|0\rangle = 0 = \langle 0|\,a^\dagger$; using $\langle 0|\,a(\boldsymbol{k})a^\dagger(\boldsymbol{k}')\,|0\rangle = (2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}')$ from Equation (99.31) collapses the double integral to the first form. The second form is Equation (99.37). Both exponents lie on the mass shell, so $(\Box+\kappa^2)$ annihilates the integrand: $\Box \mathrm{e}^{-\mathrm{i}k\cdot x} = -k^2\mathrm{e}^{-\mathrm{i}k\cdot x}$ and $k^2=\kappa^2$ on the support of the delta. The dimension is $[\hbar c^2/\omega]1/\mathrm{m}^3 = \mathrm{J\,m^2}\cdot 1/\mathrm{m}^3$. □

Theorem 99.34 (The Feynman propagator is a Green function)**.**

$$\left(\Box+\kappa^2\right) D_F(x-y) = -\mathrm{i}\hbar c\,\delta^4(x-y)\,, \tag{99.92}$$

where $\delta^4(x)=\delta(x^0)\delta^3(\boldsymbol{x})$ carries $1/\mathrm{m}^4$; both sides carry $\mathrm{J/m^3}$.
Rests on Equations (99.23), (99.24) and (99.87).

Proof. *Derives Theorem 99.34.*

Differentiate Equation (99.87) with respect to $t=x^0/c$, using $\partial_t\theta(x^0-y^0)=\delta(t_x-t_y)$:

$$\partial_t D_F = \delta(t_x-t_y)\,\langle 0|\,[\phi(x),\phi(y)]\,|0\rangle + \theta\,\langle 0|\,\partial_t\phi(x)\phi(y)\,|0\rangle + \theta\,\langle 0|\,\phi(y)\partial_t\phi(x)\,|0\rangle\,,$$

and the first term vanishes by Equation (99.24). Differentiating again, the delta now multiplies the equal-time commutator of $\partial_t\phi = c^2\pi$ with ϕ:

$$\partial_t^2 D_F = \delta(t_x-t_y)\,c^2\,\langle 0|\,[\pi(x),\phi(y)]\,|0\rangle + (\theta\text{-terms}) = -\mathrm{i}\hbar c^2\,\delta(t_x-t_y)\delta^3(\boldsymbol{x}-\boldsymbol{y}) + (\theta\text{-terms})\,,$$

using Equation (99.23). The θ-terms are the time-ordered product of $\partial_t^2\phi$ with ϕ, and the spatial and mass parts of $(\Box+\kappa^2)$ pass through the θ functions untouched, so

$$\left(\Box+\kappa^2\right) D_F = -\mathrm{i}\hbar\,\delta(t_x-t_y)\delta^3(\boldsymbol{x}-\boldsymbol{y}) + \langle 0|\,T\left\{\left(\Box+\kappa^2\right)\phi(x)\,\phi(y)\right\}|0\rangle\,,$$

whose second term vanishes by the field equation. Finally $\delta(t_x-t_y)=c\,\delta(x^0-y^0)$, giving Equation (99.92). □

Theorem 99.35 (Momentum-space form and the Feynman contour)**.**

$$D_F(x-y) = \int \frac{\mathrm{d}^4k}{(2\pi)^4}\,\frac{\mathrm{i}\hbar c}{k^2-\kappa^2+\mathrm{i}\epsilon}\,\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}\,, \tag{99.93}$$

with $\epsilon\to 0^+$, and this — and no other choice of contour — reproduces Equation (99.88). In terms of four-momenta $p=\hbar k$ with the measure $\mathrm{d}^4p/(2\pi\hbar)^4$ the same statement reads

$$\widetilde{D}_F(p) = \frac{\mathrm{i}\hbar^3 c}{p^2-m^2c^2+\mathrm{i}\epsilon}\,, \qquad [\widetilde{D}_F] = \mathrm{J\,m^3}\,. \tag{99.94}$$

Rests on Equations (99.88), (99.91) and (99.92).

Proof. *Derives Theorem 99.35.*

That Equation (99.93) solves Equation (99.92) is immediate: $\Box\to -k^2$ turns the integrand into $\mathrm{i}\hbar c(\kappa^2-k^2)/(k^2-\kappa^2) = -\mathrm{i}\hbar c$, whose Fourier transform is $-\mathrm{i}\hbar c\,\delta^4(x-y)$. The content of the theorem is that the $\mathrm{i}\epsilon$ is the right prescription, and that is decided by doing the k^0 integral.

Write $k^2-\kappa^2 = (k^0)^2 - \boldsymbol{k}^2 - \kappa^2 = (k^0)^2 - (\omega_{\boldsymbol{k}}/c)^2$ by Equation (99.28). The $\mathrm{i}\epsilon$ displaces the two poles of the integrand off the real k^0 axis:

$$(k^0)^2 - \frac{\omega_{\boldsymbol{k}}^2}{c^2} + \mathrm{i}\epsilon = \left(k^0 - \frac{\omega_{\boldsymbol{k}}}{c} + \mathrm{i}\epsilon'\right)\left(k^0 + \frac{\omega_{\boldsymbol{k}}}{c} - \mathrm{i}\epsilon'\right)\,, \tag{99.95}$$

with $\epsilon' = c\epsilon/2\omega_{\boldsymbol{k}} > 0$: the positive-frequency pole sits *below* the real axis and the negative-frequency pole *above* it. Since $x^0 = ct$, the exponential is $\mathrm{e}^{-\mathrm{i}k^0(x^0-y^0)}$, which decays exponentially as $\Im k^0 \to -\infty$ when $x^0 > y^0$ and as $\Im k^0 \to +\infty$ when $x^0 < y^0$. The integrand also falls as $|k^0|^{-2}$, so a large semicircle in the half plane where the exponential decays contributes nothing as its radius grows and the contour may be closed there; the residue theorem of Section 8.7 then evaluates the integral.

For $x^0 > y^0$, closing below encircles the pole at $k^0 = \omega_{\boldsymbol{k}}/c$ clockwise, contributing $-2\pi\mathrm{i}$ times the residue:

$$\int \frac{\mathrm{d}k^0}{2\pi} \frac{\mathrm{i}\hbar c\, \mathrm{e}^{-\mathrm{i}k^0(x^0-y^0)}}{(k^0 - \omega_{\boldsymbol{k}}/c)(k^0 + \omega_{\boldsymbol{k}}/c)} = \frac{-2\pi\mathrm{i}}{2\pi} \cdot \frac{\mathrm{i}\hbar c\, \mathrm{e}^{-\mathrm{i}\omega_{\boldsymbol{k}}(t_x-t_y)}}{2\omega_{\boldsymbol{k}}/c} = \frac{\hbar c^2}{2\omega_{\boldsymbol{k}}}\, \mathrm{e}^{-\mathrm{i}\omega_{\boldsymbol{k}}(t_x-t_y)}\,,$$

which, reinserted into the remaining d^3k integral, is exactly $D^+(x-y)$ of Equation (99.91). For $x^0 < y^0$, closing above encircles $k^0 = -\omega_{\boldsymbol{k}}/c$ counterclockwise and gives $(\hbar c^2/2\omega_{\boldsymbol{k}})\mathrm{e}^{+\mathrm{i}\omega_{\boldsymbol{k}}(t_x-t_y)}$, which after $\boldsymbol{k} \to -\boldsymbol{k}$ in the spatial integral is $D^+(y-x)$. Together these are Equation (99.88).

Any other placement of the poles gives a different function. Displacing both poles downward gives zero for $x^0 < y^0$ — the retarded function of Section 99.6.2 — and both upward gives the advanced one. The Feynman choice is the unique one that propagates the positive-frequency part forward and the negative-frequency part backward in time, which is the analytic statement of the Stueckelberg–Feynman reading of Section 99.3.2 [864, 930, 931]. Finally, $k = p/\hbar$ turns $k^2 - \kappa^2$ into $(p^2 - m^2c^2)/\hbar^2$ and $\mathrm{d}^4k/(2\pi)^4$ into $\mathrm{d}^4p/(2\pi\hbar)^4$, giving Equation (99.94); its dimension follows from $[\hbar^3 c] = \mathrm{J}^3\,\mathrm{s}^3\,\mathrm{m/s}$ divided by $[p^2]$. □

Equation (99.94) is the scalar member of the family whose fermionic and photonic partners are Equations (100.17) and (100.23); the $\mathrm{i}\epsilon$ there is the one fixed here, which is why Proposition 100.8 refers the prescription back to this chapter. Note that the pole is at $p^2 = m^2c^2$, the mass shell Equation (100.1): a propagator carries information about a particle precisely through the position of its pole, which is what makes the pole a definition of mass in the interacting theory (Chapter 100) and what the Källén–Lehmann representation [1003, 1048] generalizes.

Proposition 99.36 (The propagator outside the light cone)**.** *At spacelike separation, evaluated in the frame where the two points are simultaneous and a distance r apart,*

$$D_F(0,\boldsymbol{r}) = D^+(0,\boldsymbol{r}) = \frac{\hbar c\,\kappa}{4\pi^2 r}\, K_1(\kappa r) \underset{\kappa r \gg 1}{\longrightarrow} \frac{\hbar c\,\kappa}{4\pi^2 r}\sqrt{\frac{\pi}{2\kappa r}}\, \mathrm{e}^{-\kappa r}\,, \tag{99.96}$$

with K_1 the modified Bessel function of the second kind: a solution of the modified *Bessel equation $x^2y'' + xy' - (x^2+\nu^2)y = 0$, which is the Bessel equation of Section 9.8 continued to imaginary argument, and which that section does not itself treat. It is small but not zero.*

Rests on Equations (99.88) *and* (99.91)*.*

Proof. *Derives Proposition 99.36.*

At equal times Equation (99.91) reduces to $D^+ = \frac{1}{2}\hbar c \int \mathrm{d}^3k\,(2\pi)^{-3}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}(\boldsymbol{k}^2+\kappa^2)^{-1/2}$. Doing the angular integral, $\int \mathrm{d}^3k\,(2\pi)^{-3} f(|\boldsymbol{k}|)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} = (2\pi^2 r)^{-1}\int_0^\infty \mathrm{d}k\, k \sin(kr) f(k)$, so

$$D^+(0,\boldsymbol{r}) = \frac{\hbar c}{4\pi^2 r}\int_0^\infty \mathrm{d}k\, \frac{k\sin(kr)}{\sqrt{k^2+\kappa^2}} = \frac{\hbar c\,\kappa}{4\pi^2 r}\, K_1(\kappa r)\,,$$

the last step a standard integral representation of K_1; the check $\kappa \to 0$ returns $\int_0^\infty \sin(kr)\mathrm{d}k = 1/r$ and $D^+ \to \hbar c/(4\pi^2 r^2)$, the massless Coulomb-like falloff. The asymptotic form is $K_1(z) \sim \sqrt{\pi/2z}\,\mathrm{e}^{-z}$. That $D_F = D^+$ here is because at spacelike separation the two orderings give the same answer, which is Proposition 99.44. □

The exponential in Equation (99.96) has range $\kappa^{-1} = \hbar/mc$: the reduced Compton wavelength of Equation (99.10) appears once more, now as the distance over which a virtual quantum of mass m can carry an influence. The same function, sourced by a static charge, is the Yukawa potential of Chapter 92. That D_F does not vanish outside the light cone is often reported as a violation of causality. It is not, and Section 99.7.1 explains exactly why.

99.6.2 The family of Green functions

Theorem 99.35 obtained one Green function of the Klein–Gordon operator by one choice of contour. The other choices are equally legitimate solutions of Equation (99.92) — they differ by solutions of the homogeneous equation — and each answers a different physical question. All are built from the same integrand, so the family is best displayed as one formula with four contours.

Definition 99.37 (Retarded and advanced Green functions)**.**

$$D_R(x) = \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{\mathrm{i}\hbar c\, \mathrm{e}^{-\mathrm{i}k\cdot x}}{(k^0 + \mathrm{i}\epsilon)^2 - \boldsymbol{k}^2 - \kappa^2}\,, \tag{99.97}$$

with both poles pushed into the lower half plane, and D_A with $\mathrm{i}\epsilon \to -\mathrm{i}\epsilon$, pushing both into the upper half plane.

Closing the contour as in Theorem 99.35 shows at once that $D_R(x) = 0$ for $x^0 < 0$ (the contour closes in the upper half plane, which is now empty of poles) and $D_A(x) = 0$ for $x^0 > 0$. D_R is therefore the classical response function: it gives the field produced *after* a source acts, and it is the object behind the retarded potentials of Chapter 65, whose mathematics — the retarded and advanced Green functions of the wave operator, and the causality condition that selects the retarded one — is Chapter 10. It is *not* the object that appears in perturbation theory, because a scattering amplitude is not a response to a classical source but a matrix element of a time-ordered product, and Equation (99.87) is what the expansion of $\exp(\mathrm{i}S/\hbar)$ generates.

Proposition 99.38 (Relations among the two-point functions)**.** *With Δ of Equation (99.90),*

$$\mathrm{i}\hbar\,\Delta(x) = D^+(x) - D^+(-x)\,, \tag{99.98}$$

$$D_F(x) = \theta(x^0)D^+(x) + \theta(-x^0)D^+(-x)\,, \tag{99.99}$$

$$D_R(x) - D_A(x) = \mathrm{i}\hbar\,\Delta(x)\,, \tag{99.100}$$

and Δ is a c-number, not an operator.

Rests on Equations (99.90) and (99.91) and Definition 99.37.

Proof. *Derives Proposition 99.38.*

Equation (99.98) is the definition of the commutator written out, and Equation (99.99) is Equation (99.87) between vacua. For Equation (99.100), subtract the two contours of Definition 99.37: the difference is a closed contour encircling both poles clockwise, so it is $-2\pi\mathrm{i}$ times the sum of the residues, which is $\int \mathrm{d}^3k(2\pi)^{-3}(\hbar c^2/2\omega_{\boldsymbol{k}})(\mathrm{e}^{-\mathrm{i}k\cdot x} - \mathrm{e}^{\mathrm{i}k\cdot x}) = D^+(x) - D^+(-x)$ by Equation (99.91), and that is $\mathrm{i}\hbar\,\Delta$ by Equation (99.98). The $\hbar$ and no c is what the dimensions require: D_R, D_A and D^+ all carry J/m while $[\Delta] = 1/(\mathrm{m\,s})$, so only $\hbar\Delta$ matches. That Δ is a c-number is Proposition 99.44: the commutator of two free fields is proportional to the identity, so its vacuum expectation value determines it completely. □

The same construction applies to the other fields, with two differences.

Table 99.2: The two-point functions of the free scalar field. All are built from the same mass-shell integrand; the contour — equivalently the boundary condition — decides which. Only the first two are Green functions of $\Box+\kappa^2$; the Wightman and Pauli–Jordan functions solve the homogeneous equation.

Function	Contour in k^0	Question it answers
D_F	poles at $+\omega/c$ and $-\omega/c$ displaced into the lower and upper half plane respectively	What does a time-ordered product between vacua give? The building block of the perturbation series
D_R	both poles below the axis; vanishes for $x^0<0$	What field does a source produce afterwards? Classical radiation
D_A	both poles above the axis; vanishes for $x^0>0$	What source would have produced a given field?
D^+	positive-frequency mass shell only	What is the vacuum correlation of two field measurements, in a definite order?
Δ	both signs of frequency, no θ	Do measurements at x and y interfere? Zero outside the light cone

Proposition 99.39 (The Dirac propagator)**.** *With the fermionic time ordering* $T\{\psi(x)\bar\psi(y)\} = \theta(x^0-y^0)\psi(x)\bar\psi(y) - \theta(y^0-x^0)\bar\psi(y)\psi(x)$*, the propagator* $S_F(x-y)=\langle 0|\,T\{\psi(x)\bar\psi(y)\}\,|0\rangle$ *satisfies*

$$(\mathrm{i}\hbar\gamma^\mu\partial_\mu - mc)\,S_F(x-y) = \mathrm{i}\hbar\,\delta^4(x-y)\,, \tag{99.101}$$

whose momentum-space solution is Equation (100.23)*,*

$$\widetilde{S}_F(p) = \frac{\mathrm{i}\hbar}{\gamma^\mu p_\mu - mc + \mathrm{i}\epsilon} = \frac{\mathrm{i}\hbar\,(\gamma^\mu p_\mu + mc)}{p^2 - m^2c^2 + \mathrm{i}\epsilon}\,. \tag{99.102}$$

Rests on Equations (99.68) *and* (99.75)*.*

Proof. *Derives Proposition 99.39.*

The minus sign in the fermionic T product is not a convention but what makes the derivative of the θ functions produce an *anti*commutator: differentiating with respect to x^0,

$$\mathrm{i}\hbar\gamma^0\partial_0\left(\theta\psi\bar\psi - \theta'\bar\psi\psi\right) \supset \mathrm{i}\hbar\gamma^0\,\delta(x^0-y^0)\left\{\psi(x),\bar\psi(y)\right\} = \mathrm{i}\hbar\gamma^0\gamma^0\delta^4(x-y) = \mathrm{i}\hbar\,\delta^4(x-y)\,,$$

using $\left\{\psi_a,\bar\psi_b\right\} = \left\{\psi_a,\psi_c^\dagger\right\}\gamma^0_{cb} = \gamma^0_{ab}\delta^3(\boldsymbol{x}-\boldsymbol{y})$ from Equation (99.75) and $(\gamma^0)^2 = \mathbb{1}$. The remaining terms are the T product of the Dirac equation with $\bar\psi$ and vanish. In momentum space $\mathrm{i}\hbar\gamma^\mu\partial_\mu \to \gamma^\mu p_\mu$, and the second equality in Equation (99.102) follows from $(\gamma^\mu p_\mu + mc)(\gamma^\nu p_\nu - mc) = p^2 - m^2c^2$, itself a consequence of $\{\gamma^\mu,\gamma^\nu\} = 2\eta^{\mu\nu}$. Had the T product been defined with a plus sign, the derivative would have produced a commutator, which by Equation (99.75) is not a c-number, and no Green function would have resulted. □

The photon propagator is different again, because there is no unique answer: the field A_μ is not gauge invariant, so its two-point function depends on the gauge condition. In

Coulomb gauge it is non-covariant and contains the instantaneous Coulomb term visible in Equation (99.81); in the covariant gauges of Section 99.5.2 it is Equation (100.17), carrying the free parameter ξ. The resolution is that the gauge dependence cancels between diagrams whenever the external currents are conserved, so that every measurable quantity is gauge independent — a statement proved, along with the Ward–Takahashi identities that organize it, in Chapter 100. That such a cancellation must happen is guaranteed in advance by the fact that the physical states of Equation (99.84) are gauge invariant; that it does happen diagram by diagram is a nontrivial calculation.

99.6.3 Wick's theorem

Perturbation theory expands $\exp(\mathrm{i}S_{\mathrm{int}}/\hbar)$ and produces vacuum expectation values of time-ordered products of many fields. Wick's theorem [974] reduces every such object to propagators, and in doing so turns an operator calculation into combinatorics.

Definition 99.40 (Contraction)**.** Split the free field into its annihilation and creation parts, $\phi = \phi^+ + \phi^-$, where ϕ^+ collects the $a(\boldsymbol{k})$ terms of Equation (99.29) and ϕ^- the $a^\dagger(\boldsymbol{k})$ terms. The *contraction* of two fields is the c-number

$$\overline{\phi(x)\phi(y)} := T\{\phi(x)\phi(y)\} - :\phi(x)\phi(y): = D_F(x-y)\,. \tag{99.103}$$

That the contraction is the Feynman propagator is the two-field case of the theorem and is worth doing explicitly. For $x^0 > y^0$,

$$T\{\phi(x)\phi(y)\} = \phi(x)\phi(y) =: \phi(x)\phi(y): + \left[\phi^+(x), \phi^-(y)\right] ,$$

since normal ordering differs from the plain product only by the one commutator generated in moving $\phi^+(x)$ past $\phi^-(y)$; and $\left[\phi^+(x), \phi^-(y)\right] = \langle 0|\, \phi(x)\phi(y)\, |0\rangle = D^+(x-y)$, because the normal-ordered term has zero vacuum expectation value. For $y^0 > x^0$ the same with $x \leftrightarrow y$. Comparing with Equation (99.99) gives Equation (99.103).

Theorem 99.41 (Wick)**.** *For free fields,*

$$T\{\phi_1\phi_2\cdots\phi_n\} = :\phi_1\cdots\phi_n: + \sum_{\text{single contractions}} :\cdots: + \sum_{\text{double contractions}} :\cdots: + \cdots , \tag{99.104}$$

the sum running over all ways of contracting zero, one, two, … pairs, the uncontracted fields left inside the normal ordering. Consequently

$$\langle 0|\, T\{\phi_1\cdots\phi_n\}\, |0\rangle = \begin{cases} \displaystyle\sum_{\text{pairings}} \prod_{(i,j)} D_F(x_i - x_j)\,, & n \text{ even}, \\ 0\,, & n \text{ odd}, \end{cases} \tag{99.105}$$

with $(2r-1)!!$ *terms for* $n = 2r$.

Rests on Definition 99.14 and Equation (99.103).

Proof. *Derives Theorem 99.41.*

By induction on n. The case $n = 2$ is Equation (99.103). Suppose Equation (99.104) holds for $n-1$ fields and let the fields be labelled so that $x_1^0 > x_2^0 > \cdots > x_n^0$, which is no loss since both sides are symmetric under relabelling. Then $T\{\phi_1\cdots\phi_n\} = \phi_1\, T\{\phi_2\cdots\phi_n\}$, and by hypothesis the second factor is a sum of normal-ordered products of subsets of $\phi_2, \ldots, \phi_n$ times contractions. It therefore suffices to show that

$$\phi_1\, :\phi_2\cdots\phi_m: = :\phi_1\phi_2\cdots\phi_m: + \sum_{j=2}^{m} \overline{\phi_1\phi_j}\ :\phi_2\cdots\widehat{\phi_j}\cdots\phi_m: ,$$

where the hat omits a factor. Write $\phi_1 = \phi_1^+ + \phi_1^-$. The creation part ϕ_1^- is already to the left of everything and contributes the first term. The annihilation part must be moved through the creation operators inside the normal product; each such move generates one commutator $\left[\phi_1^+, \phi_j^-\right]$, and since x_1^0 is the latest time, $\left[\phi_1^+, \phi_j^-\right] = D^+(x_1 - x_j) = D_F(x_1 - x_j) = \overline{\phi_1 \phi_j}$. The commutators are c-numbers, so they factor out of the normal ordering, leaving the sum shown. Equation (99.105) follows because $\langle 0| :X: |0\rangle = 0$ unless X is empty, so only the terms in which every field is contracted survive; the number of ways of pairing $2r$ objects is $(2r-1)!!$. □

Once an interaction is switched on, Equation (99.105) *is* the Feynman rules: each contraction is an internal line carrying Equation (99.94), each interaction vertex is one factor from the expansion of $\exp(\mathrm{i}S_{\mathrm{int}}/\hbar)$, and the sum over pairings is the sum over diagrams, with the symmetry factors counting the pairings that give the same diagram. That translation is carried out in Proposition 100.14.

Proposition 99.42 (Wick's theorem for fermions and the closed-loop sign)**.** *For anticommuting fields the same expansion holds with two changes: the contraction is* $\overline{\psi(x)\bar\psi(y)} = S_F(x-y)$ *of Equation* (99.102)*, and each term carries the sign of the permutation needed to bring the contracted pairs adjacent. A closed loop of* n *fermion propagators therefore contributes an extra factor* -1*.*

Rests on Theorem 99.41 and Equation (99.102)*.*

Proof. *Derives Proposition 99.42.*

The induction of Theorem 99.41 goes through verbatim with commutators replaced by anticommutators, provided every interchange of two fermion operators is accompanied by a minus sign; that is the origin of the permutation signature. For the loop, consider n vertices joined in a cycle, so that the contractions are $\overline{\psi(x_1)\bar\psi(x_2)}, \overline{\psi(x_2)\bar\psi(x_3)}, \dots, \overline{\psi(x_n)\bar\psi(x_1)}$. Bringing each $\bar\psi(x_{i+1})$ next to its partner $\psi(x_i)$ can be done with an even number of transpositions for the first $n-1$ contractions, but the last one must move $\bar\psi(x_1)$ past all the remaining $2n-2$ operators *and* past the closing of the cycle, which is one odd permutation — a cyclic permutation of an even number of anticommuting objects. The net factor is -1, independent of n, and the resulting expression is the Dirac trace around the loop. □

The extra minus sign of Proposition 99.42 is not a bookkeeping detail. It is what makes the vacuum polarization of Chapter 100 screen rather than antiscreen the charge, and hence what makes the fine-structure constant increase with energy while the strong coupling of Chapter 102 decreases: the sign of the leading term in the beta function of Chapter 108 is decided by the competition between fermion loops carrying this -1 and gauge-boson loops which do not.

99.7 Causality and spin–statistics

Special relativity forbids influence outside the light cone. Quantum mechanics makes influence a statement about operators, not about trajectories, so the relativistic requirement has to be restated as an algebraic condition on the fields. The condition is microcausality, it is satisfied by the free fields constructed above, and — this is the remarkable part — it is satisfied only if bosons are quantized with commutators and fermions with anticommutators.

99.7.1 Microcausality

Definition 99.43 (Microcausality)**.** A field theory satisfies microcausality if every pair of local observables commutes at spacelike separation:

$$[\mathcal{O}_1(x), \mathcal{O}_2(y)] = 0 \qquad \text{whenever} \qquad (x-y)^2 < 0\,. \tag{99.106}$$

Proposition 99.44 (The free scalar commutator)**.** *For the free real scalar field,*

$$[\phi(x),\phi(y)] = \hbar c \int \frac{\mathrm{d}^4k}{(2\pi)^3}\,\delta\!\left(k^2-\kappa^2\right)\operatorname{sgn}(k^0)\,\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}\;\mathbb{1}\,, \tag{99.107}$$

a c-number times the identity, Lorentz invariant, and equal to zero for every spacelike separation $(x-y)^2<0$.

Rests on Equations (99.29)*,* (99.31) *and* (99.91)*.*

Proof. *Derives Proposition 99.44.*

Substituting Equation (99.29) twice and using Equation (99.31),

$$[\phi(x),\phi(y)] = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\frac{\hbar c^2}{2\omega_{\boldsymbol{k}}}\left(\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}-\mathrm{e}^{\mathrm{i}k\cdot(x-y)}\right) = D^+(x-y)-D^+(y-x)\,,$$

which is a c-number because Equation (99.31) is: the operators have cancelled. Writing each Wightman function in the covariant form of Equation (99.91) and sending $k\to -k$ in the second gives Equation (99.107), since $\theta(k^0)-\theta(-k^0)=\operatorname{sgn}(k^0)$.

Invariance: d^4k and $\delta(k^2-\kappa^2)$ are invariant under the full Lorentz group, and $\operatorname{sgn}(k^0)$ is invariant under the *orthochronous* subgroup, because a transformation that preserves the direction of time cannot move a four-vector from the forward mass hyperboloid to the backward one — the two sheets $k^0>0$ and $k^0<0$ of $k^2=\kappa^2$ are disconnected and each is preserved (Proposition 94.21).

Vanishing: take first the case of equal times, $x^0=y^0$, and write $\boldsymbol{r}=\boldsymbol{x}-\boldsymbol{y}$. Then

$$[\phi(x),\phi(y)] = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\frac{\hbar c^2}{2\omega_{\boldsymbol{k}}}\left(\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}-\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\right) = 0\,,$$

because $\omega_{\boldsymbol{k}}$ depends only on $|\boldsymbol{k}|$, so substituting $\boldsymbol{k}\to-\boldsymbol{k}$ in the second term turns it into the first and the two cancel. Now let $(x-y)$ be any spacelike vector. There exists a proper orthochronous Lorentz transformation taking it to a purely spatial vector — that is the definition of spacelike — and the left side is invariant under such a transformation by the paragraph above. Hence it vanishes there too.

Note that the same argument fails for timelike separation, where no such transformation exists, and indeed it must fail: two timelike separated measurements can influence one another, and a formalism in which they could not would be describing nothing. □

Corollary 99.45 (No signalling)**.** *Let* $\mathcal{O}_1$ *be an observable localized in a region* R_1 *and let a projective measurement of it be performed. Then the expectation value of any observable* $\mathcal{O}_2$ *localized in a region* R_2 *spacelike to* R_1 *is unchanged.*

Rests on Equation (99.106) *and Proposition 99.44.*

Proof. *Derives Corollary 99.45.*

Let $\{P_i\}$ be the spectral projectors of $\mathcal{O}_1$; they are functions of $\mathcal{O}_1$ and hence localized in R_1, so $[P_i,\mathcal{O}_2]=0$ by Equation (99.106). A measurement whose outcome is not recorded takes the state ρ to $\sum_i P_i\rho P_i$, so

$$\operatorname{tr}\left(\sum_i P_i\rho P_i\,\mathcal{O}_2\right) = \operatorname{tr}\left(\rho\sum_i P_i\mathcal{O}_2P_i\right) = \operatorname{tr}\left(\rho\,\mathcal{O}_2\sum_i P_i\right) = \operatorname{tr}\left(\rho\,\mathcal{O}_2\right)\,,$$

using cyclicity, then $[P_i,\mathcal{O}_2]=0$, then $\sum_i P_i=\mathbb{1}$. □

Corollary 99.45 is what microcausality buys, and it is exactly the right amount. It does *not* say that spacelike separated measurements are uncorrelated — they are correlated, as Bell-type experiments show and as the Reeh–Schlieder theorem [1218] makes vivid by showing that the vacuum itself is entangled across any spacelike cut. It says that no local

operation can change what a spacelike observer will see, which is the only thing relativity requires.

This resolves the apparent paradox of Proposition 99.36. The propagator $D_F(x-y) = \langle 0|\, T\phi(x)\phi(y)\,|0\rangle$ is not zero outside the light cone; it falls only as $\mathrm{e}^{-\kappa r}$. But D_F is not an observable, and it is not a probability: it is one term in a perturbative expansion, and a matrix element of an ordered product. The observable statement is the commutator, and there the two orderings enter with opposite signs and cancel exactly outside the cone. The particle and antiparticle contributions — the two exponentials in Equation (99.107) — destructively interfere. That cancellation is another way of saying that antiparticles are not optional: a theory with a particle and no antiparticle of equal mass would have a nonvanishing spacelike commutator and would signal faster than light.

The axiomatic programme of Chapter 109 takes Equation (99.106) as one of the Wightman axioms [1099, 1297] rather than deriving it from a Lagrangian, which is the right order of business: it is the physical requirement, and the Lagrangian is only one way of meeting it.

99.7.2 The spin–statistics theorem

Two facts have now been established separately. A spin-1/2 field quantized with commutators has an energy unbounded below (Theorem 99.28). A scalar field quantized with commutators satisfies microcausality (Proposition 99.44). Pauli [854] showed that these are two instances of one theorem.

Theorem 99.46 (Spin and statistics)**.** *Let a relativistic field theory satisfy*

1. *invariance under the proper orthochronous Poincaré group;*
2. *a Hamiltonian bounded below, with a Poincaré-invariant state of lowest energy;*
3. *microcausality, Equation* (99.106)*.*

Then fields of integer spin must be quantized with commutators and fields of half-integer spin with anticommutators. Dropping either choice forces the abandonment of one of the three hypotheses.

Rests on Equation (99.106)*.*

The theorem in this generality requires the analytic continuation of the Wightman functions to complex Lorentz transformations, and belongs in the axiomatic setting where the hypotheses can be stated without reference to any Lagrangian. The axiomatic proofs are Burgoyne's [1139] and, independently, Lüders and Zumino's [1151]; both rest on the analyticity domain Jost established in his work on the CPT theorem [1115], which is a different theorem. The standard textbook account is [1297], and Chapter 109 returns to it.

Derivation pending #133

The spin–statistics theorem for arbitrary spin, by analytic continuation of the Wightman functions to the Jost points: this is a long argument and belongs in the long-proof appendix. The two cases relevant to the Standard Model — spin 0 and spin 1/2 — are proved in full in the text below.

What can be done here in full is the two cases that matter, and they already display the mechanism: each wrong choice destroys one hypothesis, and a different one in each case.

Proposition 99.47 (Half-integer spin cannot be quantized with commutators)**.** *Commutators for the Dirac field of Equation* (99.68) *contradict hypothesis (ii).*

Rests on Equation (99.68) *and Theorem 99.28.*

Proof. *Derives Proposition 99.47.*

This is Theorem 99.28: with commutators, Equation (99.73) becomes $\sum_s \int_{\boldsymbol{k}} \hbar\omega_{\boldsymbol{k}}(b^\dagger b - d^\dagger d)$ up to a constant, whose spectrum extends to $-\infty$ because $d^\dagger d$ has arbitrarily large eigenvalues. There is then no state of lowest energy, so no vacuum in the sense of hypothesis (ii). □

Proposition 99.48 (Integer spin cannot be quantized with anticommutators)**.** *Anticommutators for the real scalar field of Equation* (99.18) *contradict hypothesis (iii), and indeed destroy the theory entirely.*

Rests on Equations (99.18) *and* (99.96) *and Proposition 99.44.*

Proof. *Derives Proposition 99.48.*

Suppose $\left\{a(\boldsymbol{k}), a^\dagger(\boldsymbol{k}')\right\} = (2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}')$ with all other anticommutators vanishing, and expand ϕ as in Equation (99.29). Repeating the computation of Proposition 99.44 with anticommutators in place of commutators changes one sign, giving

$$\{\phi(x), \phi(y)\} = D^+(x-y) + D^+(y-x)\,, \tag{99.108}$$

in which the two terms now *add* instead of cancelling. At spacelike separation both equal Equation (99.96), which is strictly positive, so Equation (99.108) does not vanish outside the light cone.

Worse, put $y = x$. Then $\{\phi(x), \phi(x)\} = 2\phi(x)^2$ equals the c-number $2D^+(0)$, so $\phi(x)^2$ is a multiple of the identity and

$$:\!\phi(x)^2\!: \;=\; \phi(x)^2 - \langle 0|\,\phi(x)^2\,|0\rangle = 0$$

identically. Every even local polynomial in ϕ is therefore a c-number and every odd one is linear in ϕ: the mass term in Equation (99.21) is a constant, an interaction $\lambda\phi^4$ is a constant, and the theory has no nontrivial local observable at all. The would-be anticommuting scalar is not a theory with the wrong causal structure; it is not a theory. □

Remark 99.49 (The theorem is about which hypothesis breaks). It is worth being exact about the logic, because Theorem 99.46 is often quoted as though it derived statistics from spin alone. It does not. It says that the three hypotheses are jointly incompatible with the wrong pairing. One *can* build a theory of anticommuting scalars — Faddeev–Popov ghosts are exactly that, and they appear in Chapters 102 and 106 — but such fields are not observables, do not appear as asymptotic states, and exist only to cancel unphysical polarizations inside loops. They violate the spin–statistics theorem precisely by not being subject to its hypotheses.

The theorem has consequences at every scale, and they are among the most thoroughly verified facts in physics.

Atomic structure. Exclusion applied to atomic electrons is what produces shells, the periodic table and chemistry (Chapter 88); without it every electron would occupy the $1s$ orbital and all elements would be chemically alike.

Stellar remnants. The pressure of a degenerate electron gas supports white dwarfs and that of a degenerate neutron gas supports neutron stars, with the Chandrasekhar limit following from the competition between degeneracy pressure and gravity (Chapter 49). Degeneracy pressure is Equation (99.77) expressed thermodynamically.

Direct tests. The exclusion principle is not merely assumed but bounded. Ramberg and Snow [1837] passed a large current through a thin copper strip and searched for the X-rays that would be emitted if a fresh conduction electron made a transition into an already full $1s$ shell — a transition whose energy differs from the ordinary K_α line by a calculable amount, so that the search is background-free in a narrow window. None were seen, bounding the probability of a Pauli-violating transition at $\beta^2/2 < 1.7 \times 10^{-26}$. The VIP-2 experiment

[2405] repeats the measurement in the low-background environment of an underground laboratory with a silicon drift detector of much better energy resolution, improving on that limit. A violation at any level would falsify one of the three hypotheses of Theorem 99.46, which is why the searches are worth the effort even though nobody expects a signal.

The same three hypotheses appear again in the CPT theorem [1115, 1123], and that is not a coincidence: both theorems are proved from the analyticity that Lorentz invariance plus positivity of the energy forces on the Wightman functions. The connection is drawn in Chapter 105.

99.7.3 Measurability of the field operators

Equation (99.27) said that the field at a point is not an operator and must be smeared. That is a mathematical statement. Bohr and Rosenfeld [715] asked the corresponding physical question — what can a real apparatus actually measure of a quantized electromagnetic field? — and found that the two answers agree.

Their apparatus is a charged test body of finite extent and finite mass, whose displacement under the field is read off. Such a body cannot measure $\boldsymbol{E}$ at a point: it necessarily averages over its own volume and over the time it is exposed. Nor can it be made arbitrarily small, because the momentum transfer needed to read a smaller displacement grows and the body's own radiation reacts back on the field being measured. What it measures is

$$\overline{E}_i(V,T) = \frac{1}{VT}\int_V \mathrm{d}^3x \int_T \mathrm{d}t\; E_i(t,\boldsymbol{x})\,, \tag{99.109}$$

an average over a spacetime region, which is Equation (99.27) with the test function equal to the indicator of that region.

Bohr and Rosenfeld's conclusion was that the uncertainties unavoidably introduced by such an apparatus are exactly those demanded by the commutators of the averaged field strengths, no more and no less: the formalism claims neither more measurability than an apparatus can deliver, nor less. The structural part of that conclusion is a corollary of what has already been proved here.

Corollary 99.50 (Compatible field averages)**.** *Let R_1 and R_2 be spacetime regions every point of which is spacelike separated from every point of the other. Then the field strengths averaged over them, in the sense of Equation (99.109), are compatible observables: the averages commute, $\big[\overline{E}_i(R_1),\overline{B}_j(R_2)\big]=0$, and the two averages may be measured simultaneously with no mutual disturbance.*

Rests on Equations (99.107) and (99.109) and Proposition 99.44.

Proof. *Derives Corollary 99.50.*

$\boldsymbol{E}$ and $\boldsymbol{B}$ are built from $F_{\mu\nu}$ and hence from derivatives of the free field, whose commutator is the c-number Equation (99.107) (with $\kappa=0$) differentiated. That c-number vanishes at spacelike separation by Proposition 99.44, and so do its derivatives, since a function vanishing on the whole spacelike region vanishes with all its derivatives there. The averages are integrals of the fields over R_1 and R_2, so their commutator is the double integral of a function that is zero throughout the domain of integration. □

The commutator is nonzero when the regions are in light contact, which is when one measurement can influence the other, and this is what makes the operational reading of the formalism defensible in the sense discussed in Chapter 1: an operator formalism whose commutators did not match the disturbances of a possible apparatus would be claiming a structure that no experiment could probe. Bohr and Rosenfeld's analysis is the check that the two agree. Note also what the analysis licenses: measurement by a *classical* test body of a *quantized* field, which is the hybrid description every real experiment uses.

99.8 Vacuum energy made observable: the Casimir effect

Section 99.2.4 argued that the zero-point energy Equation (99.46) is unobservable in flat empty space because it is an additive constant, but that differences between geometries are not. This section computes such a difference, gives its magnitude in SI units, and reports the measurements.

99.8.1 The prediction

Phenomenon 99.51 (The Casimir force)**.** Two parallel, plane, perfectly conducting plates of area A at separation $a \ll \sqrt{A}$ in vacuum attract, with energy per unit area and force per unit area

$$\frac{E(a)}{A} = -\frac{\pi^2 \hbar c}{720\, a^3}\,, \qquad \frac{F(a)}{A} = -\frac{\pi^2 \hbar c}{240\, a^4} = -\frac{1.300 \times 10^{-27}\,\mathrm{J\,m}}{a^4}\,, \tag{99.110}$$

independent of the charge of the electron and of every property of the plates except that they are perfect conductors. At $a = 1\,\mu\mathrm{m}$ the pressure is $1.30 \times 10^{-3}\,\mathrm{Pa}$, about 10^{-8} of atmospheric.

Rests on Equations (99.46) *and* (99.80) *and Theorem 99.9.*

Derivation. *Derives Phenomenon 99.51.*

Between perfectly conducting plates the tangential electric field vanishes on both surfaces, so the allowed wave vectors normal to the plates are quantized, $k_z = n\pi/a$ with $n = 0, 1, 2, \ldots$; the wave vector parallel to the plates, $\boldsymbol{k}_\parallel$, is continuous. For $n \geq 1$ there are two polarizations, for $n = 0$ only one. The zero-point energy per unit area is therefore

$$\frac{E(a)}{A} = \frac{\hbar c}{2} \sum_n{}' 2 \int \frac{\mathrm{d}^2 k_\parallel}{(2\pi)^2} \sqrt{\boldsymbol{k}_\parallel^2 + \left(\frac{n\pi}{a}\right)^2} = \hbar c \sum_{n=0}^{\infty}{}' f\!\left(\frac{n\pi}{a}\right)\,, \tag{99.111}$$

where the prime means the $n = 0$ term carries weight $\frac{1}{2}$ (compensating its single polarization) and

$$f(b) = \int \frac{\mathrm{d}^2 k_\parallel}{(2\pi)^2} \sqrt{\boldsymbol{k}_\parallel^2 + b^2} = \lim_{s \to -1/2} \frac{1}{4\pi} \frac{b^{2-2s}}{s-1} = -\frac{b^3}{6\pi}\,, \tag{99.112}$$

the middle expression being the analytic continuation of the convergent integral $\int \mathrm{d}^2 k_\parallel (2\pi)^{-2} (\boldsymbol{k}_\parallel^2 + b^2)^{-s} = b^{2-2s}/[4\pi(s-1)]$, valid for $\Re s > 1$, to $s = -\frac{1}{2}$.

The individual terms of Equation (99.111) are infinite and so is their sum; what is finite is the difference between the sum and the same quantity with the plates absent, that is with the k_z spectrum continuous. That difference is what Euler–Maclaurin computes. With $F(n) = f(n\pi/a)$,

$$\sum_{n=0}^{\infty}{}' F(n) - \int_0^\infty F(n)\,\mathrm{d}n = -\frac{1}{12} F'(0) + \frac{1}{720} F'''(0) - \cdots\,, \tag{99.113}$$

the omitted terms involving $F^{(5)}$ and higher. The integral on the left is precisely the free-space zero-point energy of the enclosed volume, since $\mathrm{d}n = (a/\pi)\mathrm{d}k_z$ converts it into $\int \mathrm{d}^3 k$ times a: it is proportional to the volume, and it is cancelled by the identical contribution from the unbounded space outside when the plates are moved. It therefore exerts no net force, which is the honest version of "discard the divergence". By Equation (99.112), $F(n) = -(\pi/a)^3 n^3/6\pi$, so $F(0) = 0$, $F'(0) = 0$, $F'''(0) = -\pi^2/a^3$, and all higher derivatives vanish identically because F is a cubic. Hence

$$\frac{E(a)}{A} = \hbar c \cdot \frac{1}{720}\left(-\frac{\pi^2}{a^3}\right) = -\frac{\pi^2 \hbar c}{720\, a^3}\,,$$

and $F/A = -\partial(E/A)/\partial a = -\pi^2\hbar c/(240a^4)$, negative and therefore attractive. Numerically $\pi^2\hbar c/240 = 1.3001 \times 10^{-27}\,\mathrm{J\,m}$, which at $a = 1 \times 10^{-6}\,\mathrm{m}$ gives $1.300 \times 10^{-3}\,\mathrm{J/m^3} = 1.300 \times 10^{-3}\,\mathrm{Pa}$.

The same answer follows from ζ-function regularization in one line — Equation (99.111) becomes $-(\hbar c\pi^2/6a^3)\zeta(-3)$ with $\zeta(-3) = 1/120$ — but the Euler–Maclaurin route is preferable here because it exhibits the subtraction explicitly instead of hiding it in an analytic continuation. □

The result contains no coupling constant: e, ε_0 and the properties of the metal have all disappeared, leaving only $\hbar$, c and the geometry. That is the signature of a pure vacuum effect, and it is also the signature of the idealization: the plates entered only through the boundary condition "perfect conductor at all frequencies", which no real metal is.

Table 99.3: The Casimir pressure Equation (99.110) between ideal plates, in SI units, over the range accessible to experiment. The a^{-4} dependence spans seven decades over the two decades of separation that have been probed, which is what makes the measurement demanding at both ends.

Separation a	$\lvert F/A\rvert$	Comparison
100 nm	13.0 Pa	the weight of 0.13 g on 1 cm^2
200 nm	0.813 Pa	
500 nm	2.08×10^{-2} Pa	
1 µm	1.300×10^{-3} Pa	10^{-8} atmospheres
6 µm	1.00×10^{-6} Pa	

Three corrections separate Equation (99.110) from what an experiment sees, and all three are larger than the experimental precision.

Temperature. Equation (99.111) counted zero-point energy only. At temperature T the thermally excited modes contribute too, and the scale that decides when they matter is the one at which k_BT becomes comparable with $\hbar c/a$, the thermal wavelength

$$\lambda_T = \frac{\hbar c}{k_B T} = 7.63\,\mu\mathrm{m} \quad \text{at } T = 300\,\mathrm{K}\,, \tag{99.114}$$

using $k_B = 1.380\,649 \times 10^{-23}\,\mathrm{J/K}$ exactly by the SI definition of the kelvin: the thermal correction sets in for $a \gtrsim \lambda_T$. Note that this is *not* the same as demanding that k_BT reach the energy $\pi\hbar c/a$ of the lowest confined mode, which would put the crossover at $\pi\lambda_T = 24\,\mu\mathrm{m}$; the modes that matter are the continuum of parallel wave vectors, not the lowest normal mode alone. Room-temperature experiments at separations of a few micrometres are therefore already approaching the crossover region and must model it.

Finite conductivity. A real metal reflects poorly at frequencies above its plasma frequency, so modes with $k \gtrsim \omega_p/c$ are not confined at all. The relevant scale is the plasma wavelength, of order 100 nm for gold and aluminium, and the correction reduces the force by tens of percent at separations near that scale — precisely the range where the force is largest and easiest to measure.

Surface roughness and patch potentials. Deviations from flatness of order δ change the local separation and, because the force goes as a^{-4}, do not average out; and spatial variation of the work function across a real surface produces electrostatic patch potentials whose force must be measured and subtracted. These are the dominant systematics of the modern experiments.

A closely related effect is the retarded van der Waals force between a neutral atom and a conducting wall, computed by Casimir and Polder [910] in the same year and measured directly by Sukenik and collaborators [1875] by deflecting a sodium beam between gold plates.

The existence of that derivation matters for interpretation: the same force between plates can be obtained by summing retarded intermolecular forces between the constituent atoms, a calculation in which no zero-point mode sum ever appears. The two routes agree, which means that the Casimir force confirms the *relativistic quantized electromagnetic field*, and does not by itself establish that the vacuum carries an absolute energy density. Section 99.2.4 is not overturned by Section 99.8; it is illustrated by it.

99.8.2 Measurements

Casimir's prediction stood for a decade before anyone tried to see it, and for four decades before anyone saw it convincingly. The difficulty is Table 99.3: forces of nanonewtons at separations of micrometres, between surfaces that must be parallel to a small fraction of that separation.

Table 99.4: Measurements of the Casimir force. The parallel-plate geometry of Equation (99.110) was abandoned after Sparnaay because keeping two plates parallel to a fraction of a micrometre is harder than correcting for a curved surface; the sphere–plate results are compared with the proximity-force form Equation (99.115).

Work	Method	Range	Outcome
Sparnaay 1958 [1161]	parallel plates on a spring balance	0.5–2 µm	consistent with theory; uncertainty of order the effect
Lamoreaux 1997 [1939]	torsion pendulum, spherical lens against a flat plate	0.6–6 µm	agreement at the 5 % level
Mohideen and Roy 1998 [1958]	atomic force microscope, metallized sphere against a plate	0.1–0.9 µm	agreement at the 1 % level after conductivity and roughness corrections

Sparnaay [1161] used the geometry of the calculation: two flat plates, one on a spring balance, with the separation measured capacitively. He found an attraction of the right sign and order of magnitude, and stated the result honestly as not contradicting Casimir's prediction — the uncertainties, dominated by residual electrostatic charges and by the difficulty of keeping the plates parallel, were comparable to the effect itself. It was a null result in the useful sense: it excluded the absence of the force, and nothing more.

The modern experiments abandoned the parallel-plate geometry. Lamoreaux [1939] suspended a spherical lens near a flat plate from a torsion pendulum and measured the torque, sweeping the separation from 0.6 µm to 6 µm; the electrostatic force at larger separations, where the Casimir force is negligible, calibrated the apparatus and measured the residual contact potential that had to be compensated. Mohideen and Roy [1958] replaced the pendulum with an atomic force microscope, gluing a metallized sphere to the cantilever, which pushed the accessible range down to 0.1 µm where the force is four orders of magnitude larger.

Proposition 99.52 (The sphere–plate force)**.** *For a sphere of radius R at closest separation $a \ll R$ from a plane, the proximity-force approximation gives*

$$F_{sphere}(a) = 2\pi R\,\frac{E(a)}{A} = -\frac{\pi^3 \hbar c R}{360\,a^3}\,. \tag{99.115}$$

Rests on Equation (99.110).

Proof. *Derives Proposition 99.52.*

Divide the sphere into annuli at transverse radius ρ, whose local separation from the plane is $h(\rho) = a + \rho^2/2R + O(\rho^4/R^3)$. Treating each annulus as a piece of parallel plate — legitimate when $a \ll R$, since the surfaces are then locally parallel over the region that contributes — the energy is

$$E = \int_0^\infty 2\pi\rho \,\mathrm{d}\rho \, \frac{E(h(\rho))}{A} = 2\pi R \int_a^\infty \mathrm{d}h \, \frac{E(h)}{A} \,,$$

using $\rho \,\mathrm{d}\rho = R \,\mathrm{d}h$. Differentiating with respect to a gives $F = -\partial E/\partial a = 2\pi R \, E(a)/A$, and substituting Equation (99.110) gives Equation (99.115). □

Two things should be said about what the measurements establish.

They establish that the mode structure of the quantized electromagnetic field is changed by boundaries in the way Equation (99.111) says, to the percent level over an order of magnitude in separation, with a functional form (a^{-3} for the sphere, a^{-4} for the pressure between plates) that no adjustable parameter was used to fit. That is a genuine confirmation of the quantized field, and of Section 99.2.4's claim that differences of zero-point energy are physical.

They do not establish the absolute normalization of the vacuum energy, for the reason given above: the same force follows from summing retarded van der Waals interactions [910], in which no zero-point sum appears. Nor do they bear on the cosmological constant problem, whose difficulty is precisely that gravity couples to the absolute value that the Casimir experiment never sees. A text that offers the Casimir effect as evidence for the vacuum energy of Equation (99.49) has overstated its case by more than fifty orders of magnitude.

99.9 Limits of the canonical approach

Everything above is exact and everything above is free. The transition to interacting fields is made in the following chapters by perturbation theory, and it is worth ending with the reason that transition is less well founded than its success suggests.

99.9.1 Haag's theorem

The interaction picture, on which the whole of perturbation theory is built, supposes that the interacting field at a fixed time is related to a free field by a unitary transformation — so that the two share a Hilbert space, and the interacting dynamics can be written as a unitary evolution of free-field states. Haag [1072] proved that this cannot be.

Theorem 99.53 (Haag)**.** *Let ϕ_1 and ϕ_2 be two scalar fields at a fixed time, each irreducible, each covariant under spatial translations and rotations with a unique invariant vacuum, and suppose there is a unitary V with $\phi_2 = V\phi_1 V^{-1}$ and $\pi_2 = V\pi_1 V^{-1}$ at that time. If ϕ_1 is a free field of mass m, then ϕ_2 is also free, of the same mass. Consequently the interaction picture exists only for a theory that is not interacting.*

Rests on Postulate 99.5 and Equation (99.23).

Derivation pending #134

Haag's theorem: the proof runs through the equality of all equal-time vacuum expectation values under the unitary, the reconstruction of the theory from them, and the identification of the mass from the two-point function. It is a page or two of distribution-theoretic argument and belongs in the long-proof appendix; the statement and its consequences are used here, and the axiomatic setting in which the hypotheses are stated is the axiomatic quantum field theory chapter.

The mechanism is easy to state even without the proof. A unitary V preserves all vacuum expectation values, so the two theories have the same equal-time correlation functions; the Wightman reconstruction theorem [1099, 1297] says that those functions determine the theory; and the two-point function of a free field determines the mass. There is therefore no room for ϕ_2 to be anything but a free field of the same mass. The sharpest form of the obstruction is that the interacting vacuum is not merely a different vector in the free theory's Fock space — it is not in that space at all, and the two representations of the canonical commutation relations Equation (99.23) are unitarily inequivalent. This is the warning already entered in Remark 99.13, and it is the starting point of the algebraic approach of [1856] and Chapter 109.

Why, then, does perturbative quantum electrodynamics work? Three things must be said, and the third is uncomfortable.

First, every actual calculation is done in a regularized theory in which Haag's hypotheses fail. A momentum cutoff breaks Lorentz invariance; a finite volume breaks translation invariance; dimensional regularization changes the number of dimensions. The theorem does not apply to any of them, and the perturbative series is generated inside the regularization and only then continued.

Second, what perturbation theory computes is not a unitary transformation between Hilbert spaces but a formal power series in the coupling. That series is believed to be asymptotic rather than convergent — Dyson's argument that a theory with negative α would be unstable indicates a zero radius of convergence [998] — so its first several terms can be an excellent approximation while the series as a whole defines nothing.

Third, and this is the uncomfortable part: the justification of the procedure is empirical. The anomalous magnetic moment of the electron, computed to five loops and measured to comparable precision, agrees at the level of parts in 10^{12}, and the comparison is the subject of Chapter 112. That agreement is the best evidence in physics that the perturbative expansion is computing something real. It is not a proof that the object it is expanding around exists. The honest position is that the formalism of this chapter is known to be inconsistent with the interaction picture built on it, that the inconsistency has never produced a wrong answer where the answer could be checked, and that constructing an interacting four-dimensional field theory that satisfies the axioms remains open (Chapter 109).

99.9.2 What comes next

The free field is solved. Everything that makes the Standard Model a theory of anything — scattering, decay, binding, running couplings, confinement — lies in what happens when the Lagrangian acquires a term that is not quadratic. Four routes lead out, and each is a following chapter.

Perturbation theory and renormalization. Expand $\exp(\mathrm{i}S_{\mathrm{int}}/\hbar)$, apply Theorem 99.41, and read off diagrams. The loop integrals diverge, and the divergences are absorbed into a redefinition of the mass, the charge and the field normalization; the result is the most accurately tested theory known, and the machinery is Chapter 100. The dependence of the resulting couplings on the energy scale — itself a consequence of the renormalization

procedure rather than an extra assumption — is Chapter 108.

The path integral. The canonical method fixes a time slice and therefore hides Lorentz invariance, and for non-abelian gauge fields its constraint structure becomes intractable. The alternative quantization, in which the amplitude is a sum over field histories weighted by $\exp(\mathrm{i}S/\hbar)$, keeps covariance manifest and handles gauge fixing systematically; it is Chapter 106, and it is the only practical route to the theory of Chapter 102. The $\hbar$ in the phase is the same $\hbar$ that appears in Equation (99.23): the two quantizations are equivalent for the systems where both apply.

Scattering theory. Field operators are not what a detector records. Converting the one into the other — asymptotic states, the S matrix, cross-sections and decay rates in SI units — is Chapter 86, and the reduction formula that expresses an S-matrix element in terms of the time-ordered products of this chapter is stated in Chapter 109.

Axiomatics. Finally, the results of Sections 99.7.1, 99.7.2 and 99.9.1 are statements that do not depend on any Lagrangian, and they are better stated in a framework that does not begin with one. That framework is Chapter 109, and its principal use in this book is to say precisely which of the results above are theorems and which are properties of a particular model.

Chapter 100

Quantum Electrodynamics and Renormalization

Quantum electrodynamics is the first interacting quantum field theory and remains the most sharply tested theory in physics: the electron magnetic moment measured in Chapter 112 and computed here agree over the whole range that the independently measured value of the fine-structure constant permits us to compare, roughly one part in 10^9 of the anomaly. This chapter therefore carries two burdens at once. The first is physical — to take the free fields of Chapter 99, couple the Dirac field of Chapter 93 to the Maxwell field of Chapter 60 by the single requirement of local phase invariance, and compute. The second is conceptual: every loop integral so obtained diverges, and the theory only becomes predictive through renormalization, whose history runs from the crisis of the 1930s through the Lamb shift of 1947 to the covariant formalisms of Tomonaga, Schwinger, Feynman and Dyson, and whose modern reading — a low-energy effective description valid up to a cutoff — is the subject of Chapter 108.

QED is placed here because it is the template. The weak interaction (Chapter 101), the strong interaction (Chapter 102) and their unification with electromagnetism (Chapter 104) are all gauge theories built by the same construction, and every calculational device introduced below — Feynman rules, regularization, counterterms, Ward identities, running couplings — is reused there without change of principle. Standard treatments are [1654, 1907, 1912]; Feynman's non-technical account [1752] is unmatched as an orientation.

100.1 The QED Lagrangian

100.1.1 Conventions, and the SI dimension of everything

Every standard text on this subject puts the reduced Planck constant and the speed of light equal to unity on its first page, and drops them from every formula thereafter. This book does not, in obedience to the axiom of SI units stated in the front matter, and the price is one page of bookkeeping paid here once. The reward is that every formula below can be checked dimensionally, which is the cheapest error-detector available in a subject where a wrong factor is otherwise invisible, and that the fine-structure constant appears as what it is — a pure number built out of four dimensioned constants — rather than as a symbol whose dimensionlessness has been assumed.

Notation 100.1 (Metric, coordinates, momenta). The metric is $\eta_{\mu\nu} = \mathrm{diag}(+1,-1,-1,-1)$. Coordinates are $x^\mu = (ct, \boldsymbol{x})$, so that $\mathrm{d}^4x = c\,\mathrm{d}t\,\mathrm{d}^3x$ has SI dimension m^4, and $\partial_\mu = \partial/\partial x^\mu = (c^{-1}\partial_t, \nabla)$ has dimension $1/\mathrm{m}$. Four-momenta are $p^\mu = (E/c, \boldsymbol{p})$, of dimension $\mathrm{kg\,m/s}$, so

that the mass-shell condition reads

$$p^2 := p_\mu p^\mu = \frac{E^2}{c^2} - \boldsymbol{p}^2 = m^2 c^2 \,. \tag{100.1}$$

The Mandelstam variables $s = (p_1+p_2)^2$, $t = (p_1-p_3)^2$, $u = (p_1-p_4)^2$ are therefore squared *momenta*, and $\sqrt{s}\,c$ is the centre-of-mass energy. Wave four-vectors are $k^\mu = p^\mu/\hbar = (\omega/c, \boldsymbol{k})$. The symbol ϵ, standing alone, is always the positive infinitesimal of the Feynman contour prescription and ε, standing alone, is always the dimensional regulator of Definition 100.30; ε_0 is the electric constant, $\varepsilon^\mu_{(\lambda)}$ a polarization vector and $\varepsilon^{\mu\nu\rho\sigma}$ the Levi-Civita symbol. The loop momentum is written ℓ.

The action of a field theory is $S = \int \mathrm{d}t\, L = \int \mathrm{d}t\, \mathrm{d}^3x\, \mathcal{L}$, that is

$$S = \frac{1}{c}\int \mathrm{d}^4x\, \mathcal{L}\,, \tag{100.2}$$

where $\mathcal{L}$ is an energy density, $\mathrm{J/m^3}$. The factor c^{-1} is not decoration: it is what makes S an action, J s, and it will reappear in the photon propagator. Quantum mechanics enters through the phase $\exp(\mathrm{i}S/\hbar)$ of Chapter 106 and through the canonical commutators of Chapter 99, each of which carries its own $\hbar$.

Definition 100.2 (Fields and their dimensions)**.** The Dirac field ψ has dimension $1/\mathrm{m}^{3/2}$, so that $\psi^\dagger\psi$ is a number density and $\bar\psi\psi = \psi^\dagger\gamma^0\psi$ has dimension $1/\mathrm{m}^3$. The electromagnetic four-potential $A^\mu = (\phi/c, \boldsymbol{A})$ has dimension V s/m, equivalently kg m/(C s), and $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ has dimension T.

Definition 100.3 (The fine-structure constant)**.** The electromagnetic coupling of a particle of charge q is measured by the dimensionless combination

$$\alpha = \frac{e^2}{4\pi\varepsilon_0\hbar c} = \frac{\mu_0 c e^2}{4\pi\hbar} = (7.297\,352\,564\,3 \pm 0.000\,000\,001\,1)\times 10^{-3} = \frac{1}{137.035999177(21)}\,, \tag{100.3}$$

the second form following from $\varepsilon_0\mu_0 c^2 = 1$, and the numerical value being the CODATA 2022 recommendation [2431]. Its dimensionlessness is a fact about Nature and not a convention: e^2/ε_0 has dimension J m, exactly that of $\hbar c$, so the ratio is a pure number, and it is that number — not e, which is a unit-dependent quantity fixed by definition since 2019 — that every amplitude below depends on. The combination that will actually occur in the Feynman rules is

$$\frac{\mu_0 c e^2}{\hbar} = 4\pi\alpha\,. \tag{100.4}$$

Remark 100.4 (Reading this chapter against the natural-units literature)*.* The references cited throughout work in units where $\hbar = c = 1$. In those units mass, energy, momentum and inverse length are all the same quantity, the Lagrangian density has dimension $(\text{mass})^4$, the Dirac field has dimension $(\text{mass})^{3/2}$ and the gauge field $(\text{mass})^1$, and the elementary charge is dimensionless, $e = \sqrt{4\pi\alpha}$ in Heaviside–Lorentz form. To map a formula of this chapter onto such a text, delete every $\hbar$ and c; to map back, restore them by the rule that (i) a mass m always appears as mc when it stands beside a momentum and as mc^2 when it stands beside an energy, (ii) each four-dimensional loop measure carries $(2\pi\hbar)^{-4}$, (iii) each propagator carries one factor of $\hbar$ and each vertex one factor of $\hbar^{-1}$, and (iv) e^2 is replaced by $4\pi\alpha\hbar c$ in Gaussian and by $4\pi\alpha\hbar/(\mu_0 c)$ in SI form. No derivation in this chapter is carried out in natural units; this remark exists so that the reader can check the chapter against its sources, which is a different thing.

100.1.2 Local phase invariance and minimal coupling

The free Dirac field of Chapter 93 is governed by the Lagrangian density

$$\mathcal{L}_0 = \bar{\psi}\left(\mathrm{i}\hbar c\,\gamma^\mu\partial_\mu - mc^2\right)\psi\,, \tag{100.5}$$

whose Euler–Lagrange equation (Chapter 64) is $(\mathrm{i}\hbar\gamma^\mu\partial_\mu - mc)\psi = 0$. Both terms in Equation (100.5) have dimension $\mathrm{J/m^3}$, as an energy density must: $\hbar c\,\partial_\mu$ is an energy and $\bar{\psi}\psi$ an inverse volume.

This Lagrangian is invariant under the global phase rotation $\psi \mapsto \mathrm{e}^{\mathrm{i}\theta}\psi$ with θ a real constant — a representation of the group U(1), whose structure and irreducible representations are those of Section 14.2.1. By Noether's theorem [456] — the classical apparatus is Chapter 64 — the symmetry carries a conserved current, which for a particle of charge q is written

$$j^\mu = qc\,\bar{\psi}\gamma^\mu\psi\,, \qquad \partial_\mu j^\mu = 0\,, \tag{100.6}$$

of dimension $\mathrm{A/m^2}$, with $j^0 = cq\psi^\dagger\psi$ the charge density times c. The conserved charge $Q = \frac{1}{c}\int \mathrm{d}^3x\, j^0 = q\int \mathrm{d}^3x\,\psi^\dagger\psi$ is the electric charge.

Nothing forces the phase θ to be the same at every point of spacetime. Weyl's proposal [641] — the gauge principle, and the historical root of the fibre-bundle reading of Section 14.6 — is to demand that it need not be, and to see what the demand costs.

Theorem 100.5 (The gauge principle)**.** *Let $\theta : \mathbb{R}^4 \to \mathbb{R}$ be an arbitrary smooth function and require Equation (100.5) to be invariant under*

$$\psi(x) \longmapsto \mathrm{e}^{-\mathrm{i}q\Lambda(x)/\hbar}\psi(x)\,, \tag{100.7}$$

where Λ carries dimension V s so that $q\Lambda/\hbar$ is a pure number. Then invariance is possible if and only if a vector field A_μ is introduced, transforming as

$$A_\mu(x) \longmapsto A_\mu(x) + \partial_\mu\Lambda(x)\,, \tag{100.8}$$

and the ordinary derivative is replaced by the covariant derivative

$$D_\mu := \partial_\mu + \frac{\mathrm{i}q}{\hbar}A_\mu\,. \tag{100.9}$$

The resulting Lagrangian density $\bar{\psi}(\mathrm{i}\hbar c\gamma^\mu D_\mu - mc^2)\psi$ differs from Equation (100.5) by exactly $-j^\mu A_\mu$: the interaction is not chosen, it is forced.

Rests on Equations (100.5) and (100.6).

Proof. *Derives Theorem 100.5.*

Under Equation (100.7) the derivative acquires an inhomogeneous term,

$$\partial_\mu\psi \longmapsto \mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\left(\partial_\mu - \frac{\mathrm{i}q}{\hbar}\partial_\mu\Lambda\right)\psi\,,$$

so that $\mathcal{L}_0 \mapsto \mathcal{L}_0 + qc\,\bar{\psi}\gamma^\mu\psi\,\partial_\mu\Lambda = \mathcal{L}_0 + j^\mu\partial_\mu\Lambda$, which does not vanish for arbitrary Λ. The obstruction is that $\partial_\mu\psi$ compares the field at neighbouring points, whose phase conventions are now unrelated. It is removed by supplying a rule of comparison — a connection, in the language of Section 14.6. Postulate a field A_μ with the transformation law Equation (100.8) and form D_μ as in Equation (100.9). Then

$$\begin{aligned} D_\mu\psi &\longmapsto \left(\partial_\mu + \frac{\mathrm{i}q}{\hbar}A_\mu + \frac{\mathrm{i}q}{\hbar}\partial_\mu\Lambda\right)\mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\psi \\ &= \mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\left(\partial_\mu\psi - \frac{\mathrm{i}q}{\hbar}(\partial_\mu\Lambda)\psi + \frac{\mathrm{i}q}{\hbar}A_\mu\psi + \frac{\mathrm{i}q}{\hbar}(\partial_\mu\Lambda)\psi\right) \\ &= \mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}D_\mu\psi\,, \end{aligned} \tag{100.10}$$

so $\bar\psi\gamma^\mu D_\mu\psi$ is invariant, the two phases cancelling. Conversely, any modification $\partial_\mu \to \partial_\mu + X_\mu$ restoring invariance must satisfy $X_\mu \mapsto X_\mu + \mathrm{i}q\,\partial_\mu\Lambda/\hbar$ by the same computation read backwards, which is Equation (100.8) for $A_\mu = \hbar X_\mu/\mathrm{i}q$. Finally

$$\bar\psi\left(\mathrm{i}\hbar c\gamma^\mu D_\mu - mc^2\right)\psi = \mathcal{L}_0 - qc\,\bar\psi\gamma^\mu\psi A_\mu = \mathcal{L}_0 - j^\mu A_\mu\,.$$

□

The field A_μ so forced into existence is not yet dynamical. Its kinetic term is fixed by the only gauge-invariant object one can build from it: the curvature of the connection,

$$F_{\mu\nu} = \frac{\hbar}{\mathrm{i}q}\,[D_\mu, D_\nu] = \partial_\mu A_\nu - \partial_\nu A_\mu\,, \tag{100.11}$$

which is manifestly invariant under Equation (100.8) because partial derivatives commute. This is the abelian case of the general curvature Equation (14.139); the non-abelian generalization, where the commutator leaves a residual term and the field strength is no longer gauge invariant but only covariant, is what makes Chapter 102 a different theory rather than a relabelled one.

Collecting, the QED Lagrangian density is

$$\mathcal{L}_{\mathrm{QED}} = \bar\psi\left(\mathrm{i}\hbar c\,\gamma^\mu D_\mu - m_e c^2\right)\psi - \frac{1}{4\mu_0}F_{\mu\nu}F^{\mu\nu}\,, \qquad D_\mu = \partial_\mu - \frac{\mathrm{i}e}{\hbar}A_\mu\,, \tag{100.12}$$

the electron carrying charge $q = -e$ with $e > 0$. Each term is an energy density: F^2/μ_0 has dimension $\mathrm{T}^2/\mathrm{N}/\mathrm{A}^2 = \mathrm{J}/\mathrm{m}^3$, and the normalization $-1/4\mu_0$ is the one that reproduces Maxwell's equations of Chapter 60 in the classical limit, since $-\frac{1}{4\mu_0}F_{\mu\nu}F^{\mu\nu} = \frac{\varepsilon_0}{2}\left(E^2 - c^2B^2\right)$.

Two consequences deserve to be stated separately, because they are observations and not conventions.

Proposition 100.6 (The coupling is universal)**.** *The interaction term $-j^\mu A_\mu$ contains the charge q and nothing else about the matter field. Any two species carrying the same q therefore couple to the photon identically, whatever their masses, and the electromagnetic vertex of a lepton is fixed once the vertex of any one charged particle is.*

Rests on Theorem 100.5 and Equations (100.7) and (100.9).

Proof. *Derives Proposition 100.6.*

Immediate from Theorem 100.5: the covariant derivative Equation (100.9) depends on the matter field only through the charge appearing in its transformation law Equation (100.7), and no other parameter of the matter sector enters D_μ. That this survives quantization — that radiative corrections do not make the observed charge mass-dependent — is not automatic and is the content of the Ward–Takahashi identity of Section 100.5.2. □

Phenomenon 100.7 (The photon has no mass)**.** A term $\frac{1}{2\mu_0}(m_\gamma c/\hbar)^2 A_\mu A^\mu$ is the only Lorentz-invariant mass term available to the gauge field, and it is not invariant under Equation (100.8): local phase invariance forbids a photon mass outright. The prediction is that the Coulomb potential is exactly $1/r$ and that electromagnetic waves travel at exactly c in vacuum for every frequency. Experimentally the photon rest mass is bounded by $m_\gamma < 1.6\times10^{-50}\,\mathrm{kg}$, that is about $1\times10^{-14}\,\mathrm{eV}/c^2$, from a laboratory test of Coulomb's law [1460], and by some four further orders of magnitude from the large-scale magnetic fields of the solar system [2157, 2427]. No deviation has ever been seen.

Rests on Equations (100.8) and (100.12).

Derivation. *Derives Phenomenon 100.7.*

Add the mass term to Equation (100.12); under Equation (100.8) it changes by $\frac{1}{\mu_0}(m_\gamma c/\hbar)^2\left(A^\mu\partial_\mu\Lambda + \frac{1}{2}\partial_\mu\Lambda\,\partial^\mu\Lambda\right)$, which is not a total derivative and does not vanish. Hence gauge invariance requires $m_\gamma = 0$ exactly.

To convert this into something measurable, keep the mass term and drop gauge invariance: the field equation becomes the Proca equation $\partial_\mu F^{\mu\nu} + (m_\gamma c/\hbar)^2 A^\nu = \mu_0 j^\nu$, whose static point-charge solution is the Yukawa potential

$$\phi(r) = \frac{Q}{4\pi\varepsilon_0 r}\,\mathrm{e}^{-r/\lambda}\,, \qquad \lambda = \frac{\hbar}{m_\gamma c}\,.$$

The verification is the same contour computation as for any Yukawa kernel: the Fourier transform of $(\boldsymbol{k}^2 + \lambda^{-2})^{-1}$ is $\mathrm{e}^{-r/\lambda}/4\pi r$, obtained by closing the $|\boldsymbol{k}|$ contour in the upper half plane and picking up the pole at $|\boldsymbol{k}| = \mathrm{i}/\lambda$ (Theorem 8.24). A Cavendish-type null experiment with concentric conductors bounds any departure from the $1/r^2$ force law, and hence bounds λ^{-1}; the laboratory bound $\lambda > 3\times 10^7\,\mathrm{m}$ [1460] translates through $m_\gamma = \hbar/\lambda c$ into

$$m_\gamma < \frac{1.0546\times 10^{-34}\,\mathrm{J\,s}}{3\times 10^{7}\,\mathrm{m}\times 2.998\times 10^{8}\,\mathrm{m/s}} = 1.2\times 10^{-50}\,\mathrm{kg}\,,$$

of the order quoted. The bound is a bound and not a measurement: what is established is that if the photon has a mass it is smaller than 10^{-50} of a kilogram, which is 10^{-20} of the electron mass. □

100.1.3 Gauge fixing and the photon propagator

Gauge invariance, having produced the theory, immediately obstructs its quantization. Write the free photon action from Equation (100.12) in the form

$$S_{\mathrm{EM}} = \frac{1}{c}\int \mathrm{d}^4x \left(-\frac{1}{4\mu_0}F_{\mu\nu}F^{\mu\nu}\right) = \frac{1}{2c\mu_0}\int \mathrm{d}^4x\; A^\mu\left(\eta_{\mu\nu}\Box - \partial_\mu\partial_\nu\right)A^\nu\,, \tag{100.13}$$

after an integration by parts, with $\Box = \partial_\mu\partial^\mu$. In momentum space, with the Fourier convention

$$f(x) = \int \frac{\mathrm{d}^4p}{(2\pi\hbar)^4}\,\tilde f(p)\,\mathrm{e}^{-\mathrm{i}p\cdot x/\hbar}\,, \qquad \partial_\mu \longmapsto -\frac{\mathrm{i}p_\mu}{\hbar}\,, \tag{100.14}$$

— the four-dimensional case of the transform of Chapter 17, with the $\hbar$ inserted so that the conjugate variable is a momentum rather than a wave number — the kernel becomes

$$\widetilde{M}_{\mu\nu}(p) = -\frac{1}{\mu_0\hbar^2}\left(p^2\eta_{\mu\nu} - p_\mu p_\nu\right)\,. \tag{100.15}$$

This matrix is singular: $\widetilde{M}_{\mu\nu}p^\nu = 0$ identically, so it has a null eigenvector at every p and cannot be inverted. The statement is not a technical nuisance but the algebraic shadow of Equation (100.8): in momentum space a gauge transformation is $A_\mu \mapsto A_\mu - \mathrm{i}p_\mu\Lambda/\hbar$, so the longitudinal direction is pure gauge and the action does not depend on it. A propagator is an inverse of the kernel, and one cannot invert a map along a direction in which nothing happens.

The cure is to fix the gauge: add to Equation (100.13) a term that is not gauge invariant, chosen so that physical answers do not depend on it. The covariant family is

$$\mathcal{L}_{\mathrm{gf}} = -\frac{1}{2\mu_0\xi}\left(\partial^\mu A_\mu\right)^2\,, \tag{100.16}$$

with ξ a dimensionless parameter. Its effect on Equation (100.15) is to replace $p^2\eta_{\mu\nu} - p_\mu p_\nu$ by $p^2\eta_{\mu\nu} - (1-\xi^{-1})p_\mu p_\nu$, which is invertible for $\xi \neq 0$.

Proposition 100.8 (The photon propagator)**.** *With the gauge-fixing term Equation (100.16) the free photon two-point function is*

$$\widetilde{D}_{\mu\nu}(p) = \frac{-\mathrm{i}\mu_0\hbar^3 c}{p^2 + \mathrm{i}\epsilon}\left(\eta_{\mu\nu} - (1-\xi)\,\frac{p_\mu p_\nu}{p^2}\right)\,, \tag{100.17}$$

of SI dimension $\mathrm{V}^2\,\mathrm{s}^2\,\mathrm{m}^2$. *The choice* $\xi = 1$ *is the Feynman gauge,* $\xi = 0$ *the Landau gauge, in which the propagator is purely transverse.*

Rests on Equations (100.2), (100.15) *and* (100.16).

Proof. *Derives Proposition 100.8.*

Write $P^{\mathrm{T}}_{\mu\nu} = \eta_{\mu\nu} - p_\mu p_\nu/p^2$ and $P^{\mathrm{L}}_{\mu\nu} = p_\mu p_\nu/p^2$; these are orthogonal projectors summing to $\eta_{\mu\nu}$. The gauge-fixed kernel is

$$\widetilde{M}_{\mu\nu} = -\frac{1}{\mu_0\hbar^2}\left(p^2 P^{\mathrm{T}}_{\mu\nu} + \frac{p^2}{\xi} P^{\mathrm{L}}_{\mu\nu}\right),$$

whose inverse on each eigenspace is immediate: $\widetilde{M}^{-1} = -\mu_0\hbar^2\left(P^{\mathrm{T}} + \xi P^{\mathrm{L}}\right)/p^2$. A Gaussian functional integral with phase $\exp(\mathrm{i}S/\hbar)$ and quadratic form $S = \frac{1}{2}\int A\,\mathcal{M}A$ has two-point function $\mathrm{i}\hbar\mathcal{M}^{-1}$; here $\mathcal{M} = \widetilde{M}/c$ by Equation (100.2), so the propagator is $\mathrm{i}\hbar c\,\widetilde{M}^{-1}$, which is Equation (100.17) after restoring $P^{\mathrm{T}} + \xi P^{\mathrm{L}} = \eta_{\mu\nu} - (1-\xi)p_\mu p_\nu/p^2$. The $\mathrm{i}\epsilon$ prescription is the one that propagates positive frequencies forward in time, fixed by the contour argument of Chapter 99. Dimensionally, $\mu_0\hbar^3 c/p^2$ carries $\mathrm{kg}^2\,\mathrm{m}^6/(\mathrm{s}^4\,\mathrm{A}^2)$, which is $[A]^2\mathrm{m}^4$ as the Fourier transform of a product of two potentials must be. □

The parameter ξ is arbitrary, so no measurable quantity may depend on it. The reason it does not is that a photon propagator always joins two currents, and the extra term is proportional to $p_\mu p_\nu$:

$$p_\mu \tilde{j}^\mu(p) = 0 \quad \text{for a conserved current,} \tag{100.18}$$

by Equation (100.6) transformed with Equation (100.14). At tree level this is checked directly — for an on-shell electron line, $q_\mu \bar{u}(p')\gamma^\mu u(p) = \bar{u}(p')(\gamma\cdot p' - \gamma\cdot p)u(p) = (m_e c - m_e c)\bar{u}(p')u(p) = 0$ by the Dirac equation — and to all orders it is the Ward–Takahashi identity of Section 100.5.2. The ξ-independence of renormalized amplitudes is used below as a check on the renormalization of Section 100.5. The systematic reading of the whole construction, in which $\partial\mathcal{L}/\partial\dot{A}^0 = 0$ is a primary constraint and the gauge freedom is generated by a first-class constraint, belongs to Chapter 26; the covariant quantization with an indefinite-metric state space and a subsidiary condition is that of Gupta and Bleuler [953, 958]. In an abelian theory the Faddeev–Popov determinant that accompanies this procedure [1347] is field-independent and can be discarded, which is precisely what fails in Chapter 102.

100.1.4 Discrete symmetries and Furry's theorem

QED conserves C, P and T separately. This is not an assumption: it is forced by Equation (100.12) together with the requirement that the theory be local, Lorentz invariant and hermitian, because the only interaction the gauge principle allows is $-j^\mu A_\mu$ with j^μ the vector current, and the vector current has definite behaviour under all three operations. Had the gauge principle permitted an axial current $\bar{\psi}\gamma^\mu\gamma^5\psi$ to couple as well, parity would be violated — which is exactly what happens in Chapter 101. The full account of C, P, T and the CPT theorem is Chapter 105; what is needed here is one corollary with large calculational consequences.

Definition 100.9 (Charge conjugation on the Dirac field)**.** Charge conjugation is the unitary $\mathcal{C}$ implementing

$$\mathcal{C}\psi(x)\mathcal{C}^{-1} = \mathcal{C}_{\mathrm{D}}\,\bar{\psi}^{\mathsf{T}}(x)\,, \qquad \mathcal{C}A_\mu(x)\mathcal{C}^{-1} = -A_\mu(x)\,, \tag{100.19}$$

where the numerical matrix $\mathcal{C}_{\mathrm{D}}$ satisfies

$$\mathcal{C}_{\mathrm{D}}\,(\gamma^\mu)^{\mathsf{T}}\,\mathcal{C}_{\mathrm{D}}^{-1} = -\gamma^\mu\,, \tag{100.20}$$

(in the Dirac representation $\mathcal{C}_{\mathrm{D}} = \mathrm{i}\gamma^2\gamma^0$). The sign on A_μ is forced: the current Equation (100.6) is odd under Equation (100.20), and $j^\mu A_\mu$ must be even.

Theorem 100.10 (Furry)**.** *Any QED amplitude given by a single closed fermion loop with an odd number of external photon lines attached vanishes identically [793].*

Rests on Equation (100.20) and Definition 100.9.

Proof. *Derives Theorem 100.10.*

Let n be the number of photons attached to the loop. The contribution of the loop is, up to a factor common to both orientations,

$$T^{\mu_1\cdots\mu_n} = -\operatorname{tr}\left[\gamma^{\mu_1} S(p_1)\gamma^{\mu_2} S(p_2)\cdots\gamma^{\mu_n} S(p_n)\right] ,$$

with $S(p) = \mathrm{i}\hbar(\gamma\cdot p + m_e c)/(p^2 - m_e^2 c^2)$ the fermion propagator and the overall minus sign the one attached to every closed fermion loop by the anticommuting Wick contraction of Chapter 99. Every diagram comes in two versions, differing in the direction in which the charge circulates. Take the transpose of the trace, which leaves it unchanged, and insert $\mathcal{C}_{\mathrm{D}}^{-1}\mathcal{C}_{\mathrm{D}}$ between every pair of factors. By Equation (100.20), $\mathcal{C}_{\mathrm{D}}(\gamma^\mu)^{\mathsf{T}}\mathcal{C}_{\mathrm{D}}^{-1} = -\gamma^\mu$ and hence $\mathcal{C}_{\mathrm{D}} S(p)^{\mathsf{T}}\mathcal{C}_{\mathrm{D}}^{-1} = S(-p)$, because the transpose reverses the sign of the $\gamma\cdot p$ term and leaves $m_e c$ alone. The transposed trace is therefore the trace for the reversed orientation, with an extra factor $(-1)^n$ from the n vertices. Summing the two orientations gives $(1 + (-1)^n)\, T^{\mu_1\cdots\mu_n}$, which vanishes for odd n. □

Remark 100.11 (What Furry's theorem buys). Two of the primitively divergent amplitudes counted in Theorem 100.17 — the one- and three-photon vacuum amplitudes — vanish for this reason, so QED has fewer divergences to renormalize than power counting alone predicts. The even cases are untouched, and do not vanish: the two-photon vacuum amplitude is the photon self-energy of Section 100.4.1, the largest radiative effect in this chapter. In particular the one-photon tadpole vanishes, which is why the photon does not acquire a vacuum expectation value, and the three-photon amplitude vanishes, which is why the leading photon self-interaction is the four-photon box of Section 100.8.3 and is suppressed by α^4 rather than α^3. The theorem is a statement about C invariance and therefore fails in the weak sector, where C is violated maximally.

100.2 Perturbation theory and its three formalisms

The interaction $-j^\mu A_\mu$ cannot be treated exactly. It is treated in powers of α, and because $\alpha \approx 1/137$ the expansion is spectacularly good — which is the whole reason QED is the most accurately checked theory there is. Between 1943 and 1949 three superficially different ways of organizing that expansion appeared, and Dyson proved them the same theory.

100.2.1 Tomonaga–Schwinger covariant formulation

The perturbation theory of the 1930s was built on the interaction picture in a fixed Lorentz frame: states evolve by $\mathrm{i}\hbar\,\mathrm{d}\,|\Psi(t)\rangle\,/\mathrm{d}t = H_{\mathrm{int}}(t)\,|\Psi(t)\rangle$, with t the time coordinate of that frame. The scheme worked, in the sense that it produced the Klein–Nishina cross-section and the Bethe–Heitler formula, but it hid Lorentz invariance completely: at intermediate stages energy was not conserved, individual terms were frame-dependent, and the divergences appeared in forms that cancelled between contributions of different apparent order. Nobody could tell which of the infinities were physical.

Tomonaga's resolution [895], obtained in wartime isolation and unknown outside Japan until 1947, was to replace the constant-time slice by an arbitrary spacelike surface σ and ask how the state changes when the surface is deformed at one point.

Definition 100.12 (Tomonaga–Schwinger equation)**.** Let σ be a spacelike hypersurface and $|\Psi[\sigma]\rangle$ the state functional on it. The interaction-picture evolution is

$$\mathrm{i}\hbar c\,\frac{\delta\,|\Psi[\sigma]\rangle}{\delta\sigma(x)} = \mathcal{H}_{\mathrm{int}}(x)\;|\Psi[\sigma]\rangle\;, \tag{100.21}$$

where $\delta/\delta\sigma(x)$ is the functional derivative with respect to a local deformation of the surface at the point x, of dimension $1/\mathrm{m}^4$, and $\mathcal{H}_{\mathrm{int}}(x) = j^{\mu}(x)A_{\mu}(x)$ is the interaction Hamiltonian density, of dimension J/m^3. Both sides then carry J/m^3, since $\hbar c\,\delta/\delta\sigma$ is an energy density.

Every symbol in Equation (100.21) is a scalar, and the surface is arbitrary: manifest covariance is restored. Taking σ to be a family of parallel flat slices recovers the ordinary interaction picture, which is why nothing is lost.

Proposition 100.13 (Integrability is microcausality)**.** *Equation* (100.21) *possesses a solution for every spacelike σ if and only if*

$$[\mathcal{H}_{\mathrm{int}}(x),\mathcal{H}_{\mathrm{int}}(y)] = 0 \qquad \text{for } (x-y)^2 < 0\,. \tag{100.22}$$

Rests on Equation (100.21) *and Definition 100.12.*

Proof. *Derives Proposition 100.13.*

Two deformations of σ at spacelike separated points x and y can be performed in either order and must give the same final surface. Applying Equation (100.21) twice and demanding that the mixed second functional derivative be symmetric gives $\mathcal{H}_{\mathrm{int}}(x)\mathcal{H}_{\mathrm{int}}(y) = \mathcal{H}_{\mathrm{int}}(y)\mathcal{H}_{\mathrm{int}}(x)$, which is Equation (100.22). The condition holds in QED because $\mathcal{H}_{\mathrm{int}}$ is built from local field operators whose commutators vanish outside the light cone (Chapter 99), and it fails for any nonlocal interaction — which is the formal statement of why locality is not optional here. □

Schwinger reached the same equation independently and built around it a complete operator formalism [920], in which the field equations, the commutation relations and the renormalization subtractions are all written covariantly and the mass and charge renormalizations appear as canonical transformations. It was in that framework that he computed the electron anomaly [919], the calculation reproduced in Section 100.4.3. Schwinger's method is powerful and almost unusable by hand for more than one loop; Feynman's, which came next, is the one everybody calculates with.

100.2.2 Feynman's space-time approach and the diagrams

Feynman's starting point was the observation that a negative-energy solution of the Dirac equation propagating backwards in time is kinematically indistinguishable from a positive-energy antiparticle propagating forwards [931], an idea anticipated by Stueckelberg [864]. The gain is not interpretational but computational: the electron and the positron share a single propagator, the one obtained by choosing the Feynman contour, and the separate "electron" and "positron" terms of old-fashioned perturbation theory — which had to be added by hand and whose cancellations were the source of the confusion described above — are automatically combined in one manifestly covariant expression. The companion paper [930] recast the whole perturbation series in space-time terms and stated the diagrammatic rules that carry his name; what follows is those rules, written with every $\hbar$ and c in place.

Proposition 100.14 (The Feynman rules of QED in SI form)**.** *Expand* $\exp(\mathrm{i}S/\hbar)$ *in powers of the interaction, apply Wick's theorem [974] to the resulting time-ordered products, and read off the momentum-space amplitude $\mathcal{M}$ for a given process by the following dictionary.*

1. *Internal fermion line of momentum p:*
$$\widetilde{S}_F(p) = \frac{\mathrm{i}\hbar\,(\gamma^\mu p_\mu + m_e c)}{p^2 - m_e^2 c^2 + \mathrm{i}\epsilon}\,, \tag{100.23}$$
of dimension m.

2. *Internal photon line of momentum p: the propagator Equation* (100.17).

3. *Vertex, for a fermion of charge q:*
$$-\frac{\mathrm{i}q}{\hbar}\gamma^\mu \quad \left(= +\frac{\mathrm{i}e}{\hbar}\gamma^\mu \text{ for the electron}\right), \tag{100.24}$$
of dimension $\mathrm{C}/(\mathrm{J\,s})$.

4. *Each undetermined loop momentum:* $\int \mathrm{d}^4\ell\,(2\pi\hbar)^{-4}$.

5. *Each closed fermion loop: a factor* -1 *and a Dirac trace.*

6. *External lines: $u(p,s)$ for an incoming fermion, $\bar{u}(p,s)$ outgoing, v, $\bar{v}$ for antifermions, and $\varepsilon^{\mu}_{(\lambda)}(p)$, $\varepsilon^{*\mu}_{(\lambda)}(p)$ for incoming and outgoing photons. Spinors are normalized relativistically, $\bar{u}u = 2m_e c$ and $u^\dagger u = 2E/c$, so that they carry dimension* $\mathrm{kg}^{1/2}\,\mathrm{m}^{1/2}/\mathrm{s}^{1/2}$.

7. *A relative minus sign between diagrams differing by the exchange of two external fermion lines.*

With these conventions a $2 \to 2$ amplitude $\mathcal{M}$ carries dimension $\hbar^2$.
Rests on Theorem 100.5, Proposition 100.8, and Equation (100.2).

Derivation of the vertex factor. *Derives Proposition 100.14.*
The interaction part of Equation (100.2) is, by Theorem 100.5,
$$S_{\mathrm{int}} = -\frac{1}{c}\int \mathrm{d}^4x\; j^\mu A_\mu = -q\int \mathrm{d}^4x\;\bar{\psi}\gamma^\mu\psi\,A_\mu\,,$$
the factor c cancelling between Equation (100.2) and Equation (100.6). Expanding $\exp(\mathrm{i}S/\hbar)$ to first order produces $(\mathrm{i}/\hbar)\times(-q\gamma^\mu)$ at each vertex, which is Equation (100.24). The propagators are the inverses of the quadratic forms, computed in Proposition 100.8 for the photon; for the fermion the quadratic form is $\gamma\cdot p - m_e c$ by Equation (100.14), and $\mathrm{i}\hbar(\gamma\cdot p - m_e c)^{-1} = \mathrm{i}\hbar(\gamma\cdot p + m_e c)/(p^2 - m_e^2c^2)$ using $(\gamma\cdot p)^2 = p^2$. □

Remark 100.15 (The dictionary that removes every $\hbar$). Two vertices joined by one photon propagator contribute
$$\left(\frac{q_1}{\hbar}\right)\left(\frac{q_2}{\hbar}\right)\mu_0\hbar^3 c = q_1 q_2 \mu_0 c\hbar = \pm\,4\pi\alpha\hbar^2$$
for $q_1 q_2 = \pm e^2$, by Equation (100.4). This is the single fact that makes every SI expression below reduce to its textbook form: wherever a natural-units treatment writes e^2 in Heaviside–Lorentz form, this chapter writes $4\pi\alpha\hbar^2$ for a tree exchange, and wherever it writes e^2 inside a loop the $\hbar$'s cancel against the measure $(2\pi\hbar)^{-4}$ and two propagators, leaving a bare $4\pi\alpha$. That cancellation is worth watching for: it is the reason the anomalous moment of Section 100.4.3 comes out as a pure number times α and could not have come out as anything else.

Remark 100.16 (A diagram is not a picture). Each diagram is a mnemonic for one term of the Wick expansion, and the internal lines carry momenta that do not satisfy Equation (100.1): an internal line is not a particle. The spacetime reading in which a virtual pair "is created and reabsorbed" has no operational content, because the intermediate state is summed over and never measured; what is measured is the amplitude for the whole process. Statements of the form "the electron emits a photon and reabsorbs it" are useful for remembering which integral to write down and are not to be taken as descriptions of events. Feynman himself was explicit about this [1752].

100.2.3 Dyson's S-matrix and the equivalence of the formalisms

Dyson showed that Tomonaga's and Schwinger's formalism and Feynman's rules are two presentations of one object [927], and then that the object has a closed form [928]. Integrating Equation (100.21) from the infinite past to the infinite future gives the scattering operator

$$S = T\exp\left(-\frac{\mathrm{i}}{\hbar c}\int \mathrm{d}^4x\,\mathcal{H}_{\mathrm{int}}(x)\right) = \sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{-\mathrm{i}}{\hbar c}\right)^n\int \mathrm{d}^4x_1\cdots\mathrm{d}^4x_n\,T\left[\mathcal{H}_{\mathrm{int}}(x_1)\cdots\mathcal{H}_{\mathrm{int}}(x_n)\right], \tag{100.25}$$

with T the time-ordering symbol; the $1/n!$ is what remains after the $n!$ orderings of the nested integrals are absorbed into T, and $\hbar c$ appears because $\int \mathrm{d}t\,H_{\mathrm{int}} = c^{-1}\int \mathrm{d}^4x\,\mathcal{H}_{\mathrm{int}}$. Applying Wick's theorem [974] to each time-ordered product turns it into a sum of normal-ordered terms with contractions, one contraction being one propagator, and the resulting combinatorics is exactly Proposition 100.14. That is the equivalence, and it is why this chapter uses the three names interchangeably.

The second and deeper thing Dyson proved is that QED has only finitely many divergent amplitudes.

Theorem 100.17 (Superficial degree of divergence)**.** *Let a connected QED diagram have E_e external fermion lines and E_γ external photon lines. Its integrand behaves for large loop momentum as $\ell^D\,\mathrm{d}^{4L}\ell/\ell^{4L}$ with*

$$D = 4 - \frac{3}{2}E_e - E_\gamma\,, \tag{100.26}$$

independently of the order in α. Only the amplitudes with $D \geq 0$ can diverge overall, and there are exactly seven of them.

Rests on Equations (100.17) *and* (100.23) *and Theorem 100.10.*

Proof. *Derives Theorem 100.17.*

Let L be the number of loops, V the number of vertices, and P_e, P_γ the numbers of internal fermion and photon lines. Each loop supplies $\mathrm{d}^4\ell$, each internal fermion line falls as ℓ^{-1} by Equation (100.23) and each internal photon line as ℓ^{-2} by Equation (100.17), so

$$D = 4L - P_e - 2P_\gamma\,.$$

The number of independent loops is the number of internal lines less the number of momentum conservation constraints, one per vertex, with one overall constraint left over: $L = P_e + P_\gamma - V + 1$. Every vertex has exactly two fermion ends and one photon end, and every internal line uses two ends while every external line uses one, so

$$2V = 2P_e + E_e\,, \qquad V = 2P_\gamma + E_\gamma\,.$$

Eliminating $P_e = V - E_e/2$ and $P_\gamma = (V - E_\gamma)/2$,

$$D = 4\,(P_e + P_\gamma - V + 1) - P_e - 2P_\gamma = 3P_e + 2P_\gamma - 4V + 4 = 3V - \frac{3}{2}E_e + V - E_\gamma - 4V + 4\,,$$

which is Equation (100.26): V cancels, so D does not grow with the order. Enumerating the non-negative cases, with E_e even by fermion-number conservation:

E_e	E_γ	D	amplitude
0	0	4	vacuum bubble (unobservable normalization)
0	1	3	photon tadpole — vanishes by Theorem 100.10
0	2	2	photon self-energy (Section 100.4.1)
0	3	1	vanishes by Theorem 100.10
0	4	0	light-by-light (Section 100.8.3)
2	0	1	electron self-energy (Section 100.4.2)
2	1	0	vertex correction (Section 100.4.3)

Every other amplitude has $D < 0$ and is superficially convergent. □

Definition 100.18 (Renormalizability)**.** A theory is *renormalizable* when the number of primitively divergent amplitudes is finite, so that a finite number of parameters absorbs every divergence at every order. QED is renormalizable by Theorem 100.17: three of the seven cases vanish or are unobservable and, as Section 100.4.1 and Section 100.8.3 will show, gauge invariance reduces the photon self-energy from quadratic to logarithmic divergence and makes the light-by-light amplitude finite outright. What remains is absorbed into the electron mass, the electron charge and two field normalizations [928].

Remark 100.19 (What "superficial" conceals). Theorem 100.17 controls the overall divergence of a diagram, not the divergences of its subdiagrams: a convergent-looking seven-loop graph may contain a divergent vertex inside it. Making the bookkeeping of nested and, worse, *overlapping* subdivergences work to all orders is the content of the BPHZ theorem of Section 100.5.3, and it took from 1949 to 1969 to complete.

100.3 Tree-level QED

100.3.1 Elementary processes

At lowest order in α the rules of Proposition 100.14 give every elastic and radiative process of charged leptons and photons. The list below is the working inventory of accelerator and detector physics; the numbers it produces are what Chapter 115 uses to interpret a shower and what Chapter 113 uses to normalize a luminosity.

Phenomenon 100.20 (The Compton shift)**.** X-rays scattered by loosely bound electrons emerge with a wavelength longer than that of the incident beam, the increase depending on the scattering angle θ but neither on the incident wavelength nor on the scattering material,

$$\lambda' - \lambda = \frac{h}{m_e c}\,(1 - \cos\theta)\ , \tag{100.27}$$

where $h/m_e c = 2.426\,31 \times 10^{-12}\,\mathrm{m}$ is the electron Compton wavelength [2431]. Compton measured it in 1923 [499]; the experiment is described in Chapter 69. The rate at which it occurs is the Klein–Nishina cross-section, whose non-relativistic limit is the Thomson formula of Section 65.6.1.

Rests on Equation (100.1).

Derivation. *Derives Phenomenon 100.20.*

Treat the radiation as quanta of energy $\hbar\omega$ and momentum $\hbar\boldsymbol{k}$ with $|\boldsymbol{k}| = \omega/c$, and the target as a free electron at rest. Conservation of energy and of momentum give

$$\hbar\omega + m_e c^2 = \hbar\omega' + E\,, \qquad \hbar\boldsymbol{k} = \hbar\boldsymbol{k}' + \boldsymbol{p}\,. \tag{100.28}$$

Squaring the second, $p^2 = \hbar^2(k^2 + k'^2 - 2kk'\cos\theta)$, and substituting it together with $E = \hbar(\omega - \omega') + m_e c^2$ into the mass-shell relation $E^2 = p^2c^2 + m_e^2c^4$ gives

$$\hbar^2 c^2 \left(k - k'\right)^2 + 2\hbar m_e c^3 \left(k - k'\right) = \hbar^2 c^2 \left(k^2 + k'^2 - 2kk'\cos\theta\right) .$$

The terms in k^2 and k'^2 cancel between the two sides, leaving $m_e c\,(k - k') = \hbar k k'\,(1 - \cos\theta)$, that is $1/k' - 1/k = (\hbar/m_e c)(1 - \cos\theta)$. Multiplying by 2π and using $\lambda = 2\pi/k$ gives Equation (100.27). The shift is purely kinematic: it follows from treating light as a particle of energy $\hbar\omega$ and momentum $\hbar\omega/c$, and it involves no property of the electron except its mass. The assumption that fails for tightly bound electrons is that the target is free, which is why an unshifted line survives beside the shifted one. □

The kinematics is one thing and the rate another. Two diagrams contribute to $\gamma e^- \to \gamma e^-$ at order α — the photon absorbed before or after the emission — and the spin-summed squared amplitude is

$$\overline{|\mathcal{M}|^2} = 32\pi^2\alpha^2\hbar^4 \left[\frac{p\cdot k'}{p\cdot k} + \frac{p\cdot k}{p\cdot k'} + 2m_e^2c^2\left(\frac{1}{p\cdot k} - \frac{1}{p\cdot k'}\right) + m_e^4c^4\left(\frac{1}{p\cdot k} - \frac{1}{p\cdot k'}\right)^2\right] , \tag{100.29}$$

where p, k are the incoming electron and photon momenta and k' the outgoing photon momentum. In the electron rest frame this reduces to the Klein–Nishina cross-section [630],

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{r_e^2}{2}\left(\frac{\omega'}{\omega}\right)^2\left(\frac{\omega}{\omega'} + \frac{\omega'}{\omega} - \sin^2\theta\right) , \qquad r_e = \frac{e^2}{4\pi\varepsilon_0 m_e c^2} = \frac{\alpha\hbar}{m_e c} , \tag{100.30}$$

with $r_e = 2.817\,940\,320\,5 \times 10^{-15}\,\mathrm{m}$ the classical electron radius [2431] and ω'/ω fixed by Equation (100.27).

Derivation pending #135

Klein–Nishina cross-section: the spin-summed trace computation of the two-diagram Compton amplitude leading to the squared matrix element and, after the rest-frame kinematics, to the angular distribution; the same computation gives the total cross section as a function of the ratio of photon energy to electron rest energy. It is a page of Dirac-algebra manipulation of the kind carried out in full for the vertex correction below, and belongs in Appendix A.

Proposition 100.21 (Thomson is the soft limit)**.** *For $\hbar\omega \ll m_e c^2$, Equation (100.30) reduces to the classical Thomson formula of Section 65.6.1,*

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} \longrightarrow \frac{r_e^2}{2}\left(1 + \cos^2\theta\right) , \qquad \sigma_{\mathrm{T}} = \frac{8\pi}{3}r_e^2 = 6.652\,458\,7 \times 10^{-29}\,\mathrm{m}^2 . \tag{100.31}$$

Rests on Equations (100.27) and (100.30).

Proof. *Derives Proposition 100.21.*

By Equation (100.27), $\omega'/\omega = \left[1 + (\hbar\omega/m_e c^2)(1 - \cos\theta)\right]^{-1} \to 1$ when $\hbar\omega \ll m_e c^2$. Then $\omega/\omega' + \omega'/\omega \to 2$ and the bracket in Equation (100.30) becomes $2 - \sin^2\theta = 1 + \cos^2\theta$. Integrating, $\int (1 + \cos^2\theta)\,\mathrm{d}\Omega = 4\pi + \frac{4}{3}\pi = \frac{16}{3}\pi$, so $\sigma = \frac{1}{2}r_e^2 \cdot \frac{16}{3}\pi = \frac{8}{3}\pi r_e^2$. Numerically $r_e^2 = 7.9408 \times 10^{-30}\,\mathrm{m}^2$ and $\frac{8}{3}\pi = 8.3776$, giving $6.652 \times 10^{-29}\,\mathrm{m}^2$, the accepted Thomson cross-section. The limit is the quantitative statement that classical electrodynamics is recovered when the photon carries away no appreciable recoil, and the departure from it, first seen as the falling cross-section for hard X-rays, is direct evidence that scattering is a quantum process. □

The remaining tree processes are obtained the same way and are listed here with their defining features rather than their derivations, each of which is the same trace computation.

- *Møller scattering* $e^-e^- \to e^-e^-$ [704]: two diagrams, t- and u-channel photon exchange, related by the antisymmetry of Proposition 100.14(7). In the ultrarelativistic limit

$$\overline{|\mathcal{M}|^2} = 32\pi^2\alpha^2\hbar^4\left[\frac{s^2+u^2}{t^2} + \frac{s^2+t^2}{u^2} + \frac{2s^2}{tu}\right].$$

- *Bhabha scattering* $e^+e^- \to e^+e^-$ [769]: t-channel exchange plus s-channel annihilation,

$$\overline{|\mathcal{M}|^2} = 32\pi^2\alpha^2\hbar^4\left[\frac{u^2+s^2}{t^2} + \frac{u^2+t^2}{s^2} + \frac{2u^2}{st}\right].$$

  The t-channel pole makes small-angle Bhabha scattering enormous and calculable, which is why every e^+e^- collider measures its luminosity with it — and, as Section 100.7.2 records, why the same process measures the running of α.

- *Annihilation* $e^+e^- \to \mu^+\mu^-$: a single s-channel diagram, giving in the massless limit $\overline{|\mathcal{M}|^2} = 32\pi^2\alpha^2\hbar^4(t^2+u^2)/s^2$ and the total cross section

$$\sigma\left(e^+e^- \to \mu^+\mu^-\right) = \frac{4\pi\alpha^2\hbar^2}{3s} = \frac{4\pi\alpha^2\,(\hbar c)^2}{3E_{\mathrm{cm}}^2}, \tag{100.32}$$

  the reference cross-section against which the hadronic ratio R of Chapter 102 is defined.

- *Pair annihilation into photons* $e^+e^- \to \gamma\gamma$, the process by which positrons disappear in matter and the decay channel of positronium in Section 100.8.4. Its non-relativistic cross-section is $\sigma \to \pi r_e^2 c/v$ [1043], diverging as $v \to 0$ in the way characteristic of an exothermic reaction with a Coulomb-attracted initial state; the conceptual origin is Dirac's hole theory [647], though the field-theoretic derivation needs no holes.

- *Bremsstrahlung* in the field of a nucleus, whose Born-approximation cross-section is the Bethe–Heitler formula [732]. Integrated over photon energies it gives the radiation length X_0, the depth over which an electron's energy falls by e^{-1}, and together with pair production it generates the electromagnetic cascade that every calorimeter measures. For lead $X_0 = 5.612 \times 10^{-3}\,\mathrm{m}$ [2427].

Derivation pending #136

Tree-level cross-sections: the spin-summed trace computations for Møller, Bhabha, muon-pair and two-photon annihilation, and the Born-approximation Bethe–Heitler bremsstrahlung cross-section together with the definition of the radiation length that follows from it. Each is an application of the trace technology established in the next subsection and belongs in Appendix A.

100.3.2 From amplitudes to observables

An amplitude is not an observable. Three steps separate them: summing over unobserved spins, which is done with traces; integrating over final-state momenta, which is done with invariant phase space; and dividing by the incident flux.

Lemma 100.22 (Spin sums)**.** *For Dirac spinors normalized as in Proposition 100.14,*

$$\sum_s u(p,s)\bar{u}(p,s) = \gamma^\mu p_\mu + m_e c\,, \qquad \sum_s v(p,s)\bar{v}(p,s) = \gamma^\mu p_\mu - m_e c\,, \tag{100.33}$$

and, for the photon in an amplitude contracted with a conserved current,

$$\sum_\lambda \varepsilon^\mu_{(\lambda)} \varepsilon^{*\nu}_{(\lambda)} \longrightarrow -\eta^{\mu\nu}\,. \tag{100.34}$$

Rests on Proposition 100.14.

Proof. *Derives Lemma 100.22.*

Both spinor sums are projectors onto the positive- and negative-frequency solutions of the Dirac equation, and are verified by evaluating both sides in the rest frame, where $\gamma^\mu p_\mu = m_e c\gamma^0$ and the two-dimensional spin sums give $\frac{1}{2}(1 \pm \gamma^0)\cdot 2m_e c$; the general case follows by boosting, since both sides transform as bispinor matrices. For the photon, the physical sum runs over the two transverse polarizations only and is not covariant. Writing the amplitude as $\mathcal{M} = \varepsilon^*_\mu \mathcal{M}^\mu$ with $k_\mu \mathcal{M}^\mu = 0$ — gauge invariance, established to all orders in Section 100.5.2 — the two transverse terms may be completed to the full $-\eta^{\mu\nu}$ because the extra longitudinal and scalar pieces are proportional to k^μ or k^ν and annihilate $\mathcal{M}^\mu$. The replacement is therefore legitimate inside a gauge-invariant amplitude and nowhere else. □

Lemma 100.23 (Trace theorems)**.** *With $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}\mathbb{1}$ and $\gamma^5 = \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3$, the trace of a product of an odd number of gamma matrices vanishes, as does $\operatorname{tr}\gamma^5$, while*

$$\operatorname{tr}\mathbb{1} = 4\,, \qquad \operatorname{tr}(\gamma^\mu\gamma^\nu) = 4\eta^{\mu\nu}\,, \tag{100.35}$$

$$\operatorname{tr}(\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma) = 4\left(\eta^{\mu\nu}\eta^{\rho\sigma} - \eta^{\mu\rho}\eta^{\nu\sigma} + \eta^{\mu\sigma}\eta^{\nu\rho}\right), \tag{100.36}$$

and the contraction identities, valid in four dimensions,

$$\gamma^\nu\gamma_\nu = 4\,, \quad \gamma^\nu\gamma^\alpha\gamma_\nu = -2\gamma^\alpha\,, \quad \gamma^\nu\gamma^\alpha\gamma^\beta\gamma_\nu = 4\eta^{\alpha\beta}\,, \quad \gamma^\nu\gamma^\alpha\gamma^\beta\gamma^\rho\gamma_\nu = -2\gamma^\rho\gamma^\beta\gamma^\alpha\,. \tag{100.37}$$

Rests on Equation (100.5).

Proof. *Derives Lemma 100.23.*

$\operatorname{tr}(\gamma^\mu\gamma^\nu) = \frac{1}{2}\operatorname{tr}\{\gamma^\mu,\gamma^\nu\} = \eta^{\mu\nu}\operatorname{tr}\mathbb{1} = 4\eta^{\mu\nu}$. For an odd number, insert $(\gamma^5)^2 = \mathbb{1}$ and use $\{\gamma^5,\gamma^\mu\} = 0$ to move one γ^5 through n matrices, picking up $(-1)^n$; cyclicity of the trace then gives $\operatorname{tr}X = (-1)^n \operatorname{tr}X$, which vanishes for odd n. The four-index case follows by anticommuting γ^μ to the right through the other three, generating three two-index traces with alternating signs. For the contractions, $\gamma^\nu\gamma^\alpha\gamma_\nu = \gamma^\nu(2\eta^\alpha_{\ \nu} - \gamma_\nu\gamma^\alpha) = 2\gamma^\alpha - 4\gamma^\alpha = -2\gamma^\alpha$, and the higher ones follow by the same move applied repeatedly. That these identities carry a dimension-dependent number — the 4 in $\gamma^\nu\gamma_\nu = 4$ — is not innocent: it is exactly what must be tracked when the dimension is continued in Section 100.4.4. □

Proposition 100.24 (Cross-section master formula)**.** *For $1 + 2 \to n$ final particles,*

$$\mathrm{d}\sigma = \frac{1}{4\sqrt{(p_1\cdot p_2)^2 - m_1^2 m_2^2 c^4}}\,\overline{|\mathcal{M}|^2}\,\mathrm{d}\Phi_n\,, \tag{100.38}$$

$$\mathrm{d}\Phi_n = (2\pi\hbar)^4\,\delta^4\!\left(p_1 + p_2 - \sum_f p_f\right)\prod_f \frac{c\,\mathrm{d}^3p_f}{(2\pi\hbar)^3\,2E_f}\,. \tag{100.39}$$

For $n = 2$ in the centre-of-mass frame this becomes

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{1}{64\pi^2\hbar^2 s}\,\frac{|\boldsymbol{p}_f|}{|\boldsymbol{p}_i|}\,\overline{|\mathcal{M}|^2}\,. \tag{100.40}$$

Rests on Proposition 100.14 and Lemma 100.22.

Proof. *Derives Proposition 100.24.*

The transition probability per unit time and volume is $|\mathcal{M}|^2$ divided by the normalization of the four external states, and integrated over the available final momenta with the Lorentz-invariant measure $c\,\mathrm{d}^3p/(2\pi\hbar)^3 2E$, which is $\mathrm{d}^4p\,\delta(p^2-m^2c^2)\theta(p^0)$ in disguise and therefore invariant. Dividing by the incident flux $4\sqrt{(p_1\cdot p_2)^2-m_1^2m_2^2c^4}$, which is the invariant form of $|\boldsymbol{v}_1-\boldsymbol{v}_2|$ times the two densities, gives Equation (100.38). For two final particles the delta function removes three of the six momentum integrals and fixes $|\boldsymbol{p}_f|$, leaving

$$\mathrm{d}\Phi_2=\frac{1}{\hbar^2}\frac{|\boldsymbol{p}_f|}{16\pi^2\sqrt{s}}\,\mathrm{d}\Omega\,,$$

and in the centre-of-mass frame the flux factor is $4\sqrt{s}\,|\boldsymbol{p}_i|$, whence Equation (100.40). The dimensional check is the useful one: $\mathrm{d}\Phi_2$ carries $\hbar^{-2}$, $|\mathcal{M}|^2$ carries $\hbar^4$ by Proposition 100.14, and the flux carries a squared momentum, so $\mathrm{d}\sigma$ carries $\hbar^2/p^2$, an area. □

Applied to Equation (100.32), Equation (100.40) gives $\sigma=8.68\times10^{-38}\,\mathrm{m}^2$ at $E_{\mathrm{cm}}=10\,\mathrm{GeV}$. Collider physics customarily quotes this as 0.868 nanobarns, the barn being $1\times10^{-28}\,\mathrm{m}^2$ by definition; it is not an SI unit and the SI value is the one recorded here.

The last step of the chain, and the one that fixes the sign and magnitude of the coupling once and for all, is the non-relativistic limit.

Theorem 100.25 (One-photon exchange is the Coulomb potential)**.** *The tree amplitude for the elastic scattering of two distinguishable charges q_1, q_2 by one-photon exchange reduces, for $|\boldsymbol{v}|\ll c$, to the first Born approximation for the potential*

$$V(r)=\frac{q_1q_2}{4\pi\varepsilon_0 r}\,, \tag{100.41}$$

repulsive for like charges. The Coulomb law of Chapter 58, and with it the hydrogen spectrum of Chapter 81, is therefore a prediction of Equation (100.12) and not an input.

Rests on Proposition 100.14 and Equations (100.14) and (100.17).

Proof. *Derives Theorem 100.25.*

The single t-channel diagram gives, by Proposition 100.14 in Feynman gauge,

$$\mathrm{i}\mathcal{M}=\left(-\frac{\mathrm{i}q_1}{\hbar}\right)\left(-\frac{\mathrm{i}q_2}{\hbar}\right)\frac{-\mathrm{i}\mu_0\hbar^3c\,\eta_{\mu\nu}}{t}\,[\bar u_3\gamma^\mu u_1]\,[\bar u_4\gamma^\nu u_2]\,,$$

that is $\mathcal{M}=q_1q_2\mu_0\hbar c\,t^{-1}\,[\bar u_3\gamma^\mu u_1]\,[\bar u_4\gamma_\mu u_2]$. In the non-relativistic limit the spatial components of $\bar u\gamma^\mu u$ are suppressed by $|\boldsymbol{v}|/c$ while the time component is $u^\dagger u=2E/c\to 2mc$, so the current product becomes $4m_1m_2c^2$, and the momentum transfer is spacelike, $t=-\boldsymbol{q}^2$ with $\boldsymbol{q}$ the three-momentum transfer. Hence

$$\mathcal{M}=-\frac{4q_1q_2\mu_0m_1m_2c^3\hbar}{\boldsymbol{q}^2}\,.$$

On the other side, an ordinary potential V treated in first Born approximation and re-expressed with the relativistic normalization of Proposition 100.14 contributes $\mathcal{M}=-(2m_1c)(2m_2c)\widetilde V(\boldsymbol{q})/\hbar c$, the factor $(\hbar c)^{-1}$ arising from $\exp(\mathrm{i}S/\hbar)$ together with $\int\mathrm{d}t=c^{-1}\int\mathrm{d}x^0$. Equating,

$$\widetilde V(\boldsymbol{q})=\frac{q_1q_2\mu_0\hbar^2c^2}{\boldsymbol{q}^2}=\frac{q_1q_2\hbar^2}{\varepsilon_0\boldsymbol{q}^2}\,,$$

using $\varepsilon_0\mu_0c^2=1$. Inverting the transform Equation (100.14) in three dimensions,

$$V(r)=\frac{q_1q_2\hbar^2}{\varepsilon_0}\int\frac{\mathrm{d}^3q}{(2\pi\hbar)^3}\frac{\mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}/\hbar}}{\boldsymbol{q}^2}=\frac{q_1q_2\hbar^2}{\varepsilon_0}\cdot\frac{1}{8\pi^3\hbar^3}\cdot\frac{2\pi^2\hbar}{r}=\frac{q_1q_2}{4\pi\varepsilon_0 r}\,,$$

the angular integral giving $4\pi\hbar\sin(qr/\hbar)/qr$ and the remaining radial integral $\int_0^\infty \sin(x)\,\mathrm{d}x/x = \pi/2$ after the substitution $x = qr/\hbar$. That integral is half the imaginary part of the principal-value integral of $\mathrm{e}^{\mathrm{i}z}/z$ along the real axis, which is evaluated by closing in the upper half plane and indenting the contour at the origin, where the half-residue contributes $\mathrm{i}\pi$; it is the residue theorem Theorem 8.24 in the indented form carried out in Theorem 17.78, and not the unindented archetype worked in Section 8.7 itself. The sign is the physical one: $q_1 q_2 > 0$ gives a positive, repulsive potential. What produced it is the $\eta_{\mu\nu}$ of the photon propagator contracted with two timelike currents; had the exchanged quantum been a scalar the same computation would have given attraction between like charges, which is the standard argument that the photon must be a vector. □

100.4 One loop: where the infinities are

At order α^2 in a cross-section the amplitude acquires its first loop, and every one of the seven amplitudes listed in Theorem 100.17 diverges. This section computes the three that matter, in full. The tools are assembled first, because all three computations use the same four of them and each is a piece of mathematics that exists elsewhere in this book.

Lemma 100.26 (Feynman parametrization)**.** *For any $A_1, \dots, A_n$ not lying on a common ray through the origin of the negative reals,*

$$\frac{1}{A_1 A_2 \cdots A_n} = (n-1)! \int_0^1 \mathrm{d}x_1 \cdots \mathrm{d}x_n \, \frac{\delta(1 - \sum_i x_i)}{\left[\sum_i x_i A_i\right]^n} \,. \tag{100.42}$$

In particular $1/AB = \int_0^1 \mathrm{d}x\, [xA + (1-x)B]^{-2}$.
Rests on Equation (17.109).

Proof. *Derives Lemma 100.26.*
For $n = 2$, substitute directly:

$$\int_0^1 \frac{\mathrm{d}x}{[xA + (1-x)B]^2} = \left[\frac{-1}{(A-B)\,(xA + (1-x)B)}\right]_0^1 = \frac{1}{A-B}\left(\frac{1}{B} - \frac{1}{A}\right) = \frac{1}{AB} \,.$$

For general n, differentiate Equation (100.42) for $n-1$ with respect to one of the A_i, which multiplies the left-hand side by an extra factor and raises the power on the right; equivalently, use the integral representation $1/A^a = \Gamma(a)^{-1} \int_0^\infty \mathrm{d}t\, t^{a-1} \mathrm{e}^{-tA}$, multiply n of them together, insert $1 = \int_0^\infty \mathrm{d}\lambda\, \delta(\lambda - \sum t_i)$ and rescale $t_i = \lambda x_i$; the λ integral is then a gamma function and the x_i integral is Equation (100.42). The purpose of the identity is geometric: it replaces n propagator poles at different places by a single quadratic form, which after a shift of the loop momentum depends on the external momenta only through one scalar Δ. □

Lemma 100.27 (Wick rotation)**.** *Let f be a rational function of ℓ^0 decaying faster than $(\ell^0)^{-1}$ at infinity, whose only singularities are the propagator poles displaced by the $\mathrm{i}\epsilon$ prescription of Equation (100.23). Then the contour of the ℓ^0 integration may be rotated anticlockwise from the real axis to the imaginary axis, $\ell^0 = \mathrm{i}\ell^0_{\mathrm{E}}$, giving*

$$\int \mathrm{d}^4\ell \; f\left(\ell^2\right) = \mathrm{i} \int \mathrm{d}^4\ell_{\mathrm{E}} \; f\left(-\ell^2_{\mathrm{E}}\right) \,, \qquad \ell^2_{\mathrm{E}} = \left(\ell^0_{\mathrm{E}}\right)^2 + \boldsymbol{\ell}^2 \,, \tag{100.43}$$

so that the Minkowski integral becomes a Euclidean one with positive definite denominator.
Rests on Equation (100.23) and Theorem 8.12.

Proof. *Derives Lemma 100.27.*
The denominator $\ell^2 - \Delta + \mathrm{i}\epsilon$ vanishes at $\ell^0 = \pm\left(\sqrt{\boldsymbol{\ell}^2 + \Delta} - \mathrm{i}\epsilon\right)$: one pole just below the positive real axis and one just above the negative real axis. The quadrant between the

positive real axis and the positive imaginary axis, and the quadrant between the negative real axis and the negative imaginary axis, are therefore free of singularities, and the arcs at infinity contribute nothing by the decay hypothesis. Cauchy's theorem (Theorem 8.12) applied to the closed contour made of the real segment, the arc and the imaginary segment gives Equation (100.43), the factor i coming from $\mathrm{d}\ell^0 = \mathrm{i}\,\mathrm{d}\ell^0_{\mathrm{E}}$ and the sign of $\ell^2 = (\ell^0)^2 - \boldsymbol{\ell}^2 = -\ell^2_{\mathrm{E}}$. The i$\epsilon$ prescription is thus not a convenience: it is what makes the rotation legal, and a different prescription would rotate the contour into a pole and give a different — and acausal — answer. □

Lemma 100.28 (The d-dimensional loop integral)**.** *For $\Delta > 0$ and d a complex parameter with $0 < \Re d < 2n$,*

$$\int \frac{\mathrm{d}^d \ell_{\mathrm{E}}}{(2\pi\hbar)^d} \frac{1}{\left(\ell^2_{\mathrm{E}} + \Delta\right)^n} = \frac{1}{\hbar^d \,(4\pi)^{d/2}} \frac{\Gamma\left(n - \frac{d}{2}\right)}{\Gamma(n)} \,\Delta^{\frac{d}{2}-n}\,, \tag{100.44}$$

and, after undoing the Wick rotation of Lemma 100.27,

$$\int \frac{\mathrm{d}^d \ell}{(2\pi\hbar)^d} \frac{1}{\left(\ell^2 - \Delta + \mathrm{i}\epsilon\right)^n} = \frac{(-1)^n\,\mathrm{i}}{\hbar^d \,(4\pi)^{d/2}} \frac{\Gamma\left(n - \frac{d}{2}\right)}{\Gamma(n)} \,\Delta^{\frac{d}{2}-n}\,, \tag{100.45}$$

$$\int \frac{\mathrm{d}^d \ell}{(2\pi\hbar)^d} \frac{\ell^2}{\left(\ell^2 - \Delta + \mathrm{i}\epsilon\right)^n} = \frac{(-1)^{n-1}\,\mathrm{i}}{\hbar^d \,(4\pi)^{d/2}} \frac{d}{2} \frac{\Gamma\left(n - 1 - \frac{d}{2}\right)}{\Gamma(n)} \,\Delta^{\frac{d}{2}-n+1}\,. \tag{100.46}$$

Moreover $\ell^\mu \ell^\nu$ in a numerator may be replaced by $\eta^{\mu\nu}\ell^2/d$.

Rests on Lemma 100.27 and Equations (17.42) and (17.109).

Proof. *Derives Lemma 100.28.*

Write the Euclidean integral in polar coordinates. The surface area of the unit sphere S^{d-1} follows from the Gaussian integral of Chapter 17 — take Equation (17.42) at zero frequency, giving $\int_{-\infty}^{\infty} \mathrm{e}^{-t^2}\mathrm{d}t = \sqrt{\pi}$ — evaluated in two ways:

$$\pi^{d/2} = \left(\int_{-\infty}^{\infty} \mathrm{e}^{-t^2}\mathrm{d}t\right)^d = \int \mathrm{d}^d \ell_{\mathrm{E}}\, \mathrm{e}^{-\ell^2_{\mathrm{E}}} = \Omega_d \int_0^\infty r^{d-1}\mathrm{e}^{-r^2}\,\mathrm{d}r = \frac{\Omega_d}{2}\,\Gamma\left(\frac{d}{2}\right)\,,$$

the last step by $u = r^2$ and the integral representation of the gamma function, Equation (17.109). Hence $\Omega_d = 2\pi^{d/2}/\Gamma(d/2)$, an expression that is *defined* for complex d even though no sphere is — which is the whole content of dimensional continuation. The radial integral is a beta function: with $u = r^2$,

$$\int_0^\infty \frac{r^{d-1}\mathrm{d}r}{(r^2+\Delta)^n} = \frac{1}{2}\int_0^\infty \frac{u^{\frac{d}{2}-1}\,\mathrm{d}u}{(u+\Delta)^n} = \frac{\Delta^{\frac{d}{2}-n}}{2}\,\frac{\Gamma\left(\frac{d}{2}\right)\Gamma\left(n-\frac{d}{2}\right)}{\Gamma(n)}\,,$$

convergent at the origin for $\Re d > 0$ and at infinity for $\Re d < 2n$. Assembling with the measure $(2\pi\hbar)^{-d}$ and cancelling $\Gamma(d/2)$ gives Equation (100.44). Equation (100.45) follows from Lemma 100.27, which supplies the factor i and turns $(\ell^2 - \Delta)^n$ into $(-1)^n(\ell^2_{\mathrm{E}} + \Delta)^n$; Equation (100.46) follows from the same rotation with $\ell^2 = -\ell^2_{\mathrm{E}}$ in the numerator and the radial integral run with $d \to d+2$. The replacement $\ell^\mu\ell^\nu \to \eta^{\mu\nu}\ell^2/d$ holds because the integral of $\ell^\mu\ell^\nu$ against a function of ℓ^2 is an invariant tensor, hence proportional to $\eta^{\mu\nu}$, and the constant is fixed by tracing. □

Lemma 100.29 (The pole of the gamma function)**.** *Near $\varepsilon = 0$,*

$$\Gamma(\varepsilon) = \frac{1}{\varepsilon} - \gamma_{\mathrm{E}} + O(\varepsilon)\,, \qquad \gamma_{\mathrm{E}} = 0.577\,215\,664\,9\,, \tag{100.47}$$

and more generally $\Gamma(-n+\varepsilon)$ has a simple pole of residue $(-1)^n/n!$ at every non-positive integer.

Rests on Equation (17.109) and Example 17.95.

Proof. *Derives Lemma 100.29.*

From $\Gamma(1+\varepsilon) = \varepsilon\Gamma(\varepsilon)$ and the Weierstrass product, $\ln\Gamma(1+\varepsilon) = -\gamma_{\mathrm{E}}\varepsilon + O(\varepsilon^2)$, so $\Gamma(\varepsilon) = \varepsilon^{-1}\Gamma(1+\varepsilon) = \varepsilon^{-1}\left(1 - \gamma_{\mathrm{E}}\varepsilon + \dots\right)$, which is Equation (100.47). The poles and their residues are recorded, from the Mellin point of view, at Example 17.95. This simple pole is where every ultraviolet divergence of this chapter will appear: an integral that diverges logarithmically in four dimensions converges for $\Re d < 4$ and its analytic continuation has a first-order pole at $d = 4$. □

Definition 100.30 (Dimensional continuation of the loop)**.** Set $d = 4 - 2\varepsilon$ and replace every four-dimensional loop measure by

$$\int \frac{\mathrm{d}^4\ell}{(2\pi\hbar)^4} \;\longmapsto\; \left(\frac{\bar{\mu}}{\hbar}\right)^{4-d} \int \frac{\mathrm{d}^d\ell}{(2\pi\hbar)^d}\,, \tag{100.48}$$

where $\bar{\mu}$ is an arbitrary momentum, the renormalization scale; the literature quotes it as the energy $\mu = \bar{\mu}c$, in GeV. The factor $(\bar{\mu}/\hbar)^{4-d}$ is not cosmetic: both sides of Equation (100.48) then carry $1/\mathrm{m}^4$, so that every amplitude keeps its four-dimensional SI dimension while d varies, and the coupling α stays the dimensionless number Equation (100.3). That $\bar{\mu}$ has entered is the germ of the renormalization group of Chapter 108: no observable may depend on it, and requiring that gives an equation.

100.4.1 Vacuum polarization

The photon propagates through a vacuum that is polarizable, because a photon may convert into a virtual electron–positron pair and back. The amplitude for this is the one-loop photon self-energy.

Theorem 100.31 (One-loop vacuum polarization)**.** *The one-loop photon self-energy is transverse,*

$$\Pi^{\mu\nu}(q) = \frac{q^2\eta^{\mu\nu} - q^\mu q^\nu}{\mu_0\hbar^3 c}\,\Pi_2\!\left(q^2\right)\,, \tag{100.49}$$

with the dimensionless function

$$\Pi_2\!\left(q^2\right) = -\frac{2\alpha}{\pi}\int_0^1 \mathrm{d}x\; x(1-x)\left[\frac{1}{\varepsilon} - \gamma_{\mathrm{E}} + \ln 4\pi - \ln\frac{\Delta}{\bar{\mu}^2}\right]\,, \qquad \Delta = m_e^2c^2 - x(1-x)q^2\,. \tag{100.50}$$

Its divergence is logarithmic, not quadratic as Theorem 100.17 allows, and the photon acquires no mass.

Rests on Proposition 100.14 and Lemmas 100.23 and 100.28.

Proof. *Derives Theorem 100.31.*

By Proposition 100.14, with the closed-fermion-loop factor -1 and the vertex Equation (100.24) evaluated at the electron charge $-e$, so that each vertex is $+\mathrm{i}e\gamma^\mu/\hbar$,

$$\mathrm{i}\Pi^{\mu\nu}(q) = -\left(\frac{\mathrm{i}e}{\hbar}\right)^2\left(\frac{\bar{\mu}}{\hbar}\right)^{4-d}\int\frac{\mathrm{d}^d\ell}{(2\pi\hbar)^d}\,\mathrm{tr}\left[\gamma^\mu\frac{\mathrm{i}\hbar\,(\gamma\cdot\ell + m_ec)}{\ell^2 - m_e^2c^2}\,\gamma^\nu\frac{\mathrm{i}\hbar\,(\gamma\cdot\ell' + m_ec)}{\ell'^2 - m_e^2c^2}\right]\,,$$

with $\ell' = \ell + q$. The numerical prefactor is $-(-e^2/\hbar^2)(\mathrm{i}\hbar)^2 = -e^2$, so

$$\mathrm{i}\Pi^{\mu\nu}(q) = -e^2\left(\frac{\bar{\mu}}{\hbar}\right)^{4-d}\int\frac{\mathrm{d}^d\ell}{(2\pi\hbar)^d}\,\frac{N^{\mu\nu}}{\left(\ell^2 - m_e^2c^2\right)\left(\ell'^2 - m_e^2c^2\right)}\,, \tag{100.51}$$

where, by Lemma 100.23,

$$N^{\mu\nu} = \mathrm{tr}\left[\gamma^\mu\left(\gamma\cdot\ell + m_ec\right)\gamma^\nu\left(\gamma\cdot\ell' + m_ec\right)\right] = 4\left[\ell^\mu\ell'^\nu + \ell^\nu\ell'^\mu - \eta^{\mu\nu}\left(\ell\cdot\ell' - m_e^2c^2\right)\right]\,,$$

the odd-trace terms vanishing. Combine the two denominators with Lemma 100.26 and shift $k = \ell + xq$; the denominator becomes $\left(k^2 - \Delta\right)^2$ with Δ as stated. Dropping the terms odd in k and using $k^\mu k^\nu \to \eta^{\mu\nu} k^2/d$ from Lemma 100.28,

$$N^{\mu\nu} \longrightarrow 4\left[\left(\frac{2}{d} - 1\right) k^2 \eta^{\mu\nu} - 2x(1-x) q^\mu q^\nu + \eta^{\mu\nu}\left(m_e^2 c^2 + x(1-x) q^2\right)\right] .$$

Now apply Equation (100.45) and Equation (100.46) with $n = 2$. The first term gives

$$-\left(\frac{2}{d} - 1\right) \frac{\mathrm{i}}{\hbar^d \left(4\pi\right)^{d/2}} \frac{d}{2} \Gamma\left(1 - \frac{d}{2}\right) \Delta^{\frac{d}{2} - 1} = -\frac{\mathrm{i}}{\hbar^d \left(4\pi\right)^{d/2}} \Gamma\left(2 - \frac{d}{2}\right) \Delta^{\frac{d}{2} - 1} ,$$

the leading minus sign being the $(-1)^{n-1}$ of Equation (100.46) at $n = 2$, and using $\Gamma(1 - \frac{d}{2}) = \Gamma(2 - \frac{d}{2})/(1 - \frac{d}{2})$, which converts the leading $\Gamma(1 - d/2)$ — the quadratic divergence, with a pole already at $d = 2$ — into $\Gamma(2 - d/2)$. The third term gives $+\mathrm{i}\hbar^{-d}(4\pi)^{-d/2}\Gamma(2 - \frac{d}{2}) \left(m_e^2 c^2 + x(1-x) q^2\right) \Delta^{\frac{d}{2} - 2}$, and since $m_e^2 c^2 + x(1-x) q^2 = \Delta + 2x(1-x) q^2$ the two $\eta^{\mu\nu}$ contributions cancel except for $2x(1-x) q^2 \eta^{\mu\nu}$. Collecting with the $q^\mu q^\nu$ term,

$$\mathrm{i}\Pi^{\mu\nu} = -e^2 \left(\frac{\bar{\mu}}{\hbar}\right)^{4-d} \frac{8\mathrm{i}}{\hbar^d \left(4\pi\right)^{d/2}} \left(q^2 \eta^{\mu\nu} - q^\mu q^\nu\right) \int_0^1 \mathrm{d}x \; x(1-x) \, \Gamma\left(2 - \frac{d}{2}\right) \Delta^{\frac{d}{2} - 2} .$$

The cancellation is the whole point: it is forced by gauge invariance (equivalently by Equation (100.18)), it removes the quadratic divergence that would otherwise have generated a photon mass term $\propto \Lambda^2 \eta^{\mu\nu}$, and it is what Section 100.4.4 means by saying that dimensional regularization respects the symmetry while a momentum cutoff does not. Writing $d = 4 - 2\varepsilon$ and stripping the tensor structure according to Equation (100.49),

$$\Pi_2 = -\mu_0 \hbar^3 c \, e^2 \frac{8}{\hbar^4 \left(4\pi\right)^2} \int_0^1 \mathrm{d}x \; x(1-x) \left(\frac{4\pi\bar{\mu}^2}{\Delta}\right)^{\varepsilon} \Gamma(\varepsilon) = -\frac{2\alpha}{\pi} \int_0^1 \mathrm{d}x \; x(1-x) \left[\frac{1}{\varepsilon} + \dots\right] ,$$

where $\mu_0 c e^2/\hbar = 4\pi\alpha$ by Equation (100.4) and $8/(4\pi)^2 = 1/2\pi^2$, so that the numerical factor is $4\pi\alpha/2\pi^2 = 2\alpha/\pi$; the expansion of $\Gamma(\varepsilon)$ is Lemma 100.29. This is Equation (100.50). Note that every $\hbar$ has cancelled: Π_2 is a pure number, as it must be, since it corrects a dimensionless coupling. □

Corollary 100.32 (The dressed photon propagator)**.** *Summing the geometric series of one-particle-irreducible insertions, the full photon propagator in Feynman gauge is*

$$\widetilde{D}_{\mu\nu}(q) = \frac{-\mathrm{i}\mu_0 \hbar^3 c \, \eta_{\mu\nu}}{q^2 \left[1 - \Pi_2(q^2)\right]} + (\textit{terms} \propto q_\mu q_\nu) , \tag{100.52}$$

the omitted terms dropping out of every amplitude by Equation (100.18). The pole stays at $q^2 = 0$: the photon remains massless to all orders in α, which is the statement that $\Pi^{\mu\nu}$ carries an explicit factor q^2.

Rests on Equations (100.18) and (100.49).

Proof. *Derives Corollary 100.32.*

Each insertion contributes $\widetilde{D}\,(\mathrm{i}\Pi)\,\widetilde{D}$; using Equation (100.49) and the transverse projector, the nth term is Π_2^n times the free propagator, so the sum is $(1 - \Pi_2)^{-1}$ times it. A mass would require a pole at $q^2 \neq 0$, that is a term in $\Pi^{\mu\nu}$ surviving at $q = 0$; the explicit $q^2\eta^{\mu\nu} - q^\mu q^\nu$ forbids it. □

The divergent constant $\Pi_2(0)$ is absorbed into the definition of the charge in Section 100.5.1. What survives is the q^2-dependence, and it is measurable.

Proposition 100.33 (Uehling's modification of the Coulomb potential)**.** *For* $|q^2| \ll m_e^2 c^2$ *the renormalized polarization is*

$$\Pi_2\big(q^2\big) - \Pi_2(0) = -\frac{\alpha\, q^2}{15\pi\, m_e^2 c^2} + O\big(q^4\big)\,, \tag{100.53}$$

so that the potential of a nucleus of charge Ze *acquires the short-range Uehling term [762]*

$$\delta V(\boldsymbol{r}) = -\frac{4Z\alpha^2\hbar^3}{15\, m_e^2 c}\,\delta^3(\boldsymbol{r})\,, \tag{100.54}$$

which shifts an S *level of hydrogen by*

$$\delta E_n = -\frac{4\alpha^5 m_e c^2}{15\pi n^3} = -h \times 27.1\,\mathrm{MHz} \quad \textit{for } n = 2\,. \tag{100.55}$$

Rests on Equation (100.50) *and Theorem 100.25.*

Proof. *Derives Proposition 100.33.*

Subtracting Equation (100.50) at $q^2 = 0$ removes the pole and the scale:

$$\Pi_2\big(q^2\big) - \Pi_2(0) = -\frac{2\alpha}{\pi}\int_0^1 \mathrm{d}x\; x(1-x)\,\ln\frac{m_e^2c^2}{m_e^2c^2 - x(1-x)q^2}\,.$$

Expanding the logarithm for small $|q^2|$ gives $x(1-x)q^2/m_e^2c^2$, and $\int_0^1 x^2(1-x)^2\mathrm{d}x = 1/30$, whence Equation (100.53). The effective coupling seen at momentum transfer $\boldsymbol{q}$ is therefore $\alpha\,[1 + \alpha\boldsymbol{q}^2/15\pi m_e^2c^2]$ for spacelike $q^2 = -\boldsymbol{q}^2$, and Theorem 100.25 then gives

$$\widetilde{V}(\boldsymbol{q}) = -\frac{Ze^2\hbar^2}{\varepsilon_0\boldsymbol{q}^2}\left(1 + \frac{\alpha\boldsymbol{q}^2}{15\pi m_e^2c^2}\right),$$

whose second term is independent of $\boldsymbol{q}$ and therefore, on transforming back, proportional to $\delta^3(\boldsymbol{r})$:

$$\delta V(\boldsymbol{r}) = -\frac{Ze^2\hbar^2\alpha}{15\pi\varepsilon_0 m_e^2c^2}\,\delta^3(\boldsymbol{r}) = -\frac{4Z\alpha^2\hbar^3}{15m_e^2c}\,\delta^3(\boldsymbol{r})\,,$$

using $e^2/\varepsilon_0 = 4\pi\alpha\hbar c$. Taking the expectation value in a hydrogenic S state with $Z = 1$, where $|\psi_{n00}(0)|^2 = 1/\pi n^3 a_0^3$ and $a_0 = \hbar/\alpha m_e c$ (Chapter 81),

$$\delta E_n = -\frac{4\alpha^2\hbar^3}{15m_e^2c}\cdot\frac{\alpha^3 m_e^3c^3}{\pi n^3\hbar^3} = -\frac{4\alpha^5 m_e c^2}{15\pi n^3}\,.$$

Numerically, with $\alpha^5 = 2.0693\times10^{-11}$ and $m_ec^2 = 8.1871\times10^{-14}\,\mathrm{J}$, the $n = 2$ value is $\delta E_2 = -1.798\times10^{-26}\,\mathrm{J}$, that is $\delta E_2/h = -27.1\,\mathrm{MHz}$. The sign is the physically important part: the vacuum *screens*, so the attraction is stronger at short distance and the $2S$ level, which has support at the origin, is pulled *down* relative to $2P$, which has none. This is one of the three one-loop contributions to the Lamb shift of Section 100.8.1, and the only one whose sign is opposite to the total; in muonic atoms, where the Bohr radius is smaller by $m_\mu/m_e \approx 207$ and the electron loop is no longer short-ranged compared with the orbit, the same term becomes the *dominant* QED correction, which is why muonic hydrogen [2164, 2210] measures the proton radius so sharply. □

Remark 100.34 (Screening, and its inversion)*.* The physical reading of Equation (100.53) is that the vacuum behaves as a dielectric: virtual pairs align against the field of a charge, so a test charge at large distance sees less than the bare charge and more of it as it penetrates. Because the polarizable quanta carry the same charge they screen, the effect always has this sign in QED, and the coupling grows with momentum transfer (Section 100.7). In a non-abelian theory the gauge bosons themselves carry the charge and contribute with the opposite sign; when they dominate, the coupling *falls* with momentum transfer, which is asymptotic freedom [1501, 1511] and the subject of Chapter 102.

100.4.2 Electron self-energy and mass renormalization

Theorem 100.35 (One-loop electron self-energy)**.** *Write the sum of one-particle-irreducible corrections to the electron propagator as* $-\mathrm{i}\Sigma(p)/\hbar$*, so that the full propagator is* $\mathrm{i}\hbar\,(\gamma\cdot p - m_e c - \Sigma(p))^{-1}$*. At one loop, with an infinitesimal photon mass* m_γ *retained as an infrared regulator,*

$$\Sigma_2(p) = \frac{\alpha}{2\pi}\int_0^1 \mathrm{d}x\ \left(2m_e c - x\,\gamma\cdot p\right)\left[\frac{1}{\varepsilon} - \gamma_{\mathrm{E}} + \ln 4\pi - \ln\frac{\Delta}{\bar\mu^2}\right], \tag{100.56}$$

with $\Delta = (1-x)m_e^2c^2 + x\,m_\gamma^2c^2 - x(1-x)p^2$*. The pairing matters and is fixed by the shift used below: the Feynman parameter* x *weights the* photon *denominator, so the electron mass enters* Δ *with weight* $1-x$ *and the photon mass with weight* x*, while the mass standing in the numerator is the electron's, carried by the fermion line. The divergence is logarithmic, although Theorem 100.17 permits a linear one.*

Rests on Proposition 100.14 and Lemmas 100.26 and 100.28.

Proof. *Derives Theorem 100.35.*

The rules of Proposition 100.14 give

$$-\frac{\mathrm{i}}{\hbar}\Sigma_2(p) = \left(\frac{\mathrm{i}e}{\hbar}\right)^2\left(\frac{\bar\mu}{\hbar}\right)^{4-d}\int\frac{\mathrm{d}^d\ell}{(2\pi\hbar)^d}\,\gamma^\mu\,\frac{\mathrm{i}\hbar\,(\gamma\cdot(p-\ell)+m_e c)}{(p-\ell)^2 - m_e^2c^2}\,\gamma^\nu\,\frac{-\mathrm{i}\mu_0\hbar^3 c\,\eta_{\mu\nu}}{\ell^2 - m_\gamma^2c^2},$$

whose numerical prefactor is $(-e^2/\hbar^2)(\mathrm{i}\hbar)(-\mathrm{i}\mu_0\hbar^3c) = -e^2\mu_0 c\hbar^2 = -4\pi\alpha\hbar^3$, so that

$$\Sigma_2(p) = -4\pi\mathrm{i}\alpha\,\hbar^4\left(\frac{\bar\mu}{\hbar}\right)^{4-d}\int\frac{\mathrm{d}^d\ell}{(2\pi\hbar)^d}\,\frac{\gamma^\mu\,(\gamma\cdot(p-\ell)+m_e c)\,\gamma_\mu}{\left[(p-\ell)^2 - m_e^2c^2\right]\left[\ell^2 - m_\gamma^2c^2\right]}, \tag{100.57}$$

with $\hbar^4(2\pi\hbar)^{-d}(\bar\mu/\hbar)^{4-d}$ reducing to $(2\pi)^{-4}$ at $d=4$ — every $\hbar$ again cancels, and Σ_2 comes out with the dimension of a momentum, as the mass term it corrects requires. Combine the denominators with Lemma 100.26, weight x on the photon line — equivalently $1-x$ on the electron line — and shift $k = \ell - (1-x)p$; completing the square gives $(k^2-\Delta)^2$ with Δ as stated, and the numerator's $\gamma\cdot(p-\ell) = \gamma\cdot(xp-k)$ then supplies the x that multiplies $\gamma\cdot p$. In the numerator use Equation (100.37) in d dimensions, $\gamma^\mu\gamma^\alpha\gamma_\mu = -(2-2\varepsilon)\gamma^\alpha$ and $\gamma^\mu\gamma_\mu = d$, drop the terms odd in k, and there remains $-(2-2\varepsilon)\,x\,\gamma\cdot p + (4-2\varepsilon)\,m_e c$, which at $\varepsilon = 0$ is $-2x\,\gamma\cdot p + 4m_e c$. Applying Equation (100.45) with $n=2$,

$$\Sigma_2 = -4\pi\mathrm{i}\alpha\cdot\frac{\mathrm{i}}{(4\pi)^2}\int_0^1\mathrm{d}x\,(4m_e c - 2x\,\gamma\cdot p)\left(\frac{4\pi\bar\mu^2}{\Delta}\right)^{\varepsilon}\Gamma(\varepsilon),$$

and $4\pi/(4\pi)^2 = 1/4\pi$ together with the factor 2 pulled out of the bracket gives the $\alpha/2\pi$ of Equation (100.56). The logarithm rather than a linear divergence is visible in Equation (100.57) itself: the numerator is linear in ℓ and the denominator quartic, so the would-be linearly divergent piece is odd in the shifted momentum and integrates to zero, leaving a logarithm. □

Corollary 100.36 (Mass renormalization)**.** *The electromagnetic self-interaction shifts the pole of the propagator away from the Lagrangian mass by*

$$\delta m := \frac{1}{c}\,\Sigma_2(\gamma\cdot p = m_e c) = \frac{3\alpha}{4\pi}m_e\left[\frac{1}{\varepsilon} + \dots\right] \ \leftrightarrow\ \frac{3\alpha}{4\pi}m_e\ln\frac{\Lambda^2}{m_e^2c^2} \tag{100.58}$$

in a cutoff scheme with Λ *an upper limit on the loop momentum, so that the pole sits at the measured* m_e *only if the bare mass is* $m_0 = m_e - \delta m$*: the bare electron is* lighter *than the physical one, the difference being the energy stored in its own field. The shift is positive and proportional to* m_e *itself.*

Rests on Equation (100.56).

Proof. *Derives Corollary 100.36.*

On shell, $\gamma \cdot p = m_e c$ and $p^2 = m_e^2 c^2$, so $\Delta \to (1-x)^2 m_e^2 c^2$ when $m_\gamma \to 0$, and Equation (100.56) gives $\Sigma_2 = \frac{\alpha}{2\pi} m_e c \int_0^1 (2-x)\mathrm{d}x \left[\varepsilon^{-1} + \dots\right] = \frac{3\alpha}{4\pi} m_e c \left[\varepsilon^{-1} + \dots\right]$, since $\int_0^1 (2-x)\mathrm{d}x = 3/2$. The correspondence $\varepsilon^{-1} \leftrightarrow \ln\left(\Lambda^2/m_e^2 c^2\right)$ between the two regulators, established in Section 100.4.4, gives the second form. □

Proposition 100.37 (Chiral symmetry protects the electron mass)**.** *Because $\delta m \propto m_e$, the fractional correction*

$$\frac{\delta m}{m_e} = \frac{3\alpha}{2\pi} \ln \frac{\Lambda}{m_e c} \tag{100.59}$$

is small even for Λ at the Planck momentum: with $\Lambda c = 1.22 \times 10^{19}\,\mathrm{GeV}$ and $m_e c^2 = 0.511\,\mathrm{MeV}$ the logarithm is $\ln\left(2.39 \times 10^{22}\right) = 51.5$ and $\delta m/m_e = 0.18$.

Rests on Equation (100.12) and Corollary 100.36.

Proof. *Derives Proposition 100.37.*

Set $m_e = 0$ in Equation (100.12). The Lagrangian is then invariant under the chiral rotation $\psi \mapsto \mathrm{e}^{\mathrm{i}\theta\gamma^5}\psi$, because γ^μ anticommutes with γ^5 while the mass term does not, and the interaction $-j^\mu A_\mu$ is built from γ^μ alone. A symmetry of the classical Lagrangian that survives regularization — and dimensional regularization preserves this one, away from the anomaly of Chapter 105 — forbids the generation of a term that violates it. Hence no mass can be generated when $m_e = 0$, so δm must vanish with m_e, and by dimensional analysis the only remaining possibility is $\delta m \propto m_e \times (\text{logarithm})$. Numerically, $3\alpha/2\pi = 3.484 \times 10^{-3}$ and $51.5 \times 3.484 \times 10^{-3} = 0.179$. The contrast with a scalar field is the whole reason the point is worth making: a scalar mass has no symmetry protecting it, its correction goes as Λ^2 rather than $m \ln \Lambda$, and the resulting sensitivity of the Higgs mass to the cutoff is the naturalness problem recorded in Chapter 104. □

Remark 100.38 (The field-strength constant and the spectral representation). Expanding Σ_2 about the mass shell, $\Sigma_2(p) = c\,\delta m + (\gamma \cdot p - m_e c)\left(Z_2^{-1} - 1\right) + \dots$, defines the field-strength renormalization Z_2, which is also logarithmically divergent and, unlike δm, infrared divergent as $m_\gamma \to 0$ — a first sign of Section 100.6. That Z_2 is the residue of the propagator at its pole, and that $0 < Z_2 \le 1$ with the deficit carried by multiparticle states, is the content of the Källén–Lehmann spectral representation [1003, 1048], which writes the exact two-point function as a positive-weight superposition of free ones. The representation is used in Chapter 109 as a structural fact rather than a perturbative one; neither the representation nor the bound it implies is derived here or there yet.

Derivation pending #137

The field-strength renormalization as a residue, and its bound: the Källén–Lehmann spectral representation of the exact fermion two-point function, obtained by inserting a complete set of eigenstates of the four-momentum and using the spectrum condition; the positivity of the spectral density, which follows because each weight is a squared modulus of a vacuum-to-state matrix element; the isolated one-particle delta whose weight is the field-strength constant, identifying it with the residue of the exact propagator at its pole; and the sum rule from the canonical equal-time anticommutator, which normalizes the total spectral weight to one and hence bounds the constant between zero and one, the deficit being the weight carried by the multiparticle continuum. The subsection of the axiomatic chapter named for this representation is at present a reserved heading, so the derivation is owed there or in Appendix A.

100.4.3 The vertex correction and the anomalous magnetic moment

The third divergent amplitude is the one that made renormalization believable, because after the divergence is removed a finite, parameter-free number remains and it was measured within months.

Definition 100.39 (Form factors)**.** Lorentz invariance, current conservation and parity restrict the exact electromagnetic vertex of an on-shell electron to

$$\bar{u}(p')\,\Gamma^{\mu}(p',p)\,u(p) = \bar{u}(p')\left[\gamma^{\mu}F_1\!\left(q^2\right) + \frac{\mathrm{i}\sigma^{\mu\nu}q_{\nu}}{2m_e c}F_2\!\left(q^2\right)\right]u(p)\,, \tag{100.60}$$

with $q = p' - p$, $\sigma^{\mu\nu} = \frac{\mathrm{i}}{2}[\gamma^{\mu},\gamma^{\nu}]$, and F_1, F_2 dimensionless — the factor $2m_e c$ in the denominator being exactly what makes $\sigma^{\mu\nu}q_{\nu}/2m_e c$ a pure number. At tree level $F_1 = 1$ and $F_2 = 0$.

Lemma 100.40 (Gordon identity)**.** *For on-shell spinors,*

$$2m_e c\,\bar{u}(p')\gamma^{\mu}u(p) = \bar{u}(p')\left[(p+p')^{\mu} + \mathrm{i}\sigma^{\mu\nu}q_{\nu}\right]u(p)\,. \tag{100.61}$$

Rests on Definition 100.39 and Equation (100.5).

Proof. *Derives Lemma 100.40.*

Using $\gamma^{\nu}\gamma^{\mu} = \eta^{\nu\mu} - \mathrm{i}\sigma^{\nu\mu}$, which is the split of a product into its symmetric and antisymmetric parts,

$$2m_e c\,\bar{u}(p')\gamma^{\mu}u(p) = \bar{u}(p')\left(\gamma\cdot p'\,\gamma^{\mu} + \gamma^{\mu}\,\gamma\cdot p\right)u(p)$$

by the Dirac equations $\bar{u}(p')\gamma\cdot p' = m_e c\bar{u}(p')$ and $\gamma\cdot p\,u(p) = m_e c\,u(p)$. Expanding, $\gamma\cdot p'\gamma^{\mu} = p'^{\mu} - \mathrm{i}p'_{\nu}\sigma^{\nu\mu}$ and $\gamma^{\mu}\gamma\cdot p = p^{\mu} - \mathrm{i}p_{\nu}\sigma^{\mu\nu}$, and $\sigma^{\nu\mu} = -\sigma^{\mu\nu}$, so the antisymmetric terms combine into $\mathrm{i}\sigma^{\mu\nu}(p'-p)_{\nu}$. □

Theorem 100.41 (Schwinger)**.** *At one loop the anomalous magnetic moment of the electron is*

$$a_e = F_2(0) = \frac{\alpha}{2\pi} = 1.161\,410\times10^{-3}\,, \tag{100.62}$$

finite, independent of every regulator and of the renormalization scheme, and containing no free parameter beyond α itself [919].

Rests on Definition 100.39, Proposition 100.14, and Lemma 100.40.

Proof. *Derives Theorem 100.41.*

By Proposition 100.14 the one-loop vertex is

$$\frac{\mathrm{i}e}{\hbar}\,\delta\Gamma^{\mu} = \left(\frac{\mathrm{i}e}{\hbar}\right)^3\int\frac{\mathrm{d}^4\ell}{(2\pi\hbar)^4}\,\frac{-\mathrm{i}\mu_0\hbar^3 c\,\eta_{\nu\rho}}{(\ell-p)^2}\,\gamma^{\nu}\,\frac{\mathrm{i}\hbar\,(\gamma\cdot\ell' + m_e c)}{\ell'^2 - m_e^2c^2}\,\gamma^{\mu}\,\frac{\mathrm{i}\hbar\,(\gamma\cdot\ell + m_e c)}{\ell^2 - m_e^2c^2}\,\gamma^{\rho}\,,$$

with $\ell' = \ell + q$. The numerical prefactor is $(-\mathrm{i}e^3/\hbar^3)\,(-\mathrm{i}\mu_0\hbar^3 c)\,(\mathrm{i}\hbar)^2 = e^3\mu_0 c\hbar^2$, so, dividing by $\mathrm{i}e/\hbar$ and using $\mu_0 c e^2/\hbar = 4\pi\alpha$,

$$\delta\Gamma^{\mu} = \frac{-4\pi\mathrm{i}\alpha}{(2\pi)^4}\int\mathrm{d}^4\ell\,\frac{N^{\mu}}{\left[(\ell-p)^2\right]\left[\ell'^2 - m_e^2c^2\right]\left[\ell^2 - m_e^2c^2\right]}\,, \tag{100.63}$$

every $\hbar$ having cancelled between $\hbar^4$ in the prefactor and $(2\pi\hbar)^{-4}$ in the measure. *The answer is already constrained to its final form*: the correction is α times a dimensionless integral over a dimensionless integrand, so $F_2(0)$ can only be α times a pure number.

The numerator, by Equation (100.37), is

$$N^{\mu} = \gamma^{\nu}\,(\gamma\cdot\ell' + m_e c)\,\gamma^{\mu}\,(\gamma\cdot\ell + m_e c)\,\gamma_{\nu} = -2\,\gamma\cdot\ell\,\gamma^{\mu}\,\gamma\cdot\ell' + 4m_e c\,(\ell+\ell')^{\mu} - 2m_e^2c^2\gamma^{\mu}\,.$$

Combine the three denominators by Lemma 100.26 with parameters x on $\ell'^2 - m_e^2c^2$, y on $\ell^2 - m_e^2c^2$ and z on $(\ell - p)^2$, and $x + y + z = 1$. Using $p^2 = p'^2 = m_e^2c^2$ and hence $2p \cdot q = -q^2$, the combined denominator is $(k^2 - \Delta)^3$ after the shift $k = \ell + xq - zp$, with

$$\Delta = -xy\, q^2 + (1 - z)^2\, m_e^2c^2 \,. \tag{100.64}$$

Terms odd in k integrate to zero. The F_2 part of the answer comes entirely from the terms in N^μ proportional to $(p + p')^\mu$, so those are what must be collected. At $k = 0$,

$$\ell = zp - xq = (x + z)\, p - x\, p' \,, \qquad \ell' = zp + (y + z)\, q = -y\, p + (y + z)\, p' \,,$$

using $q = p' - p$. Sandwiching between $\bar u(p')$ and $u(p)$ and using the Dirac equations repeatedly, together with $\gamma \cdot p\, \gamma^\mu \gamma \cdot p = 2p^\mu \gamma \cdot p - m_e^2c^2\gamma^\mu$ and $2p \cdot p' = 2m_e^2c^2 - q^2$, the four cross terms of $-2\,\gamma \cdot \ell\, \gamma^\mu \gamma \cdot \ell'$ reduce to

$$\begin{aligned} &-2(x + z)(-y) \left[2m_e c\, p^\mu - m_e^2c^2\gamma^\mu\right] \\ &-2(x + z)(y + z) \left[2m_e c\, (p + p')^\mu - \left(3m_e^2c^2 - q^2\right)\gamma^\mu\right] \\ &\quad -2(-x)(-y)\; m_e^2c^2\gamma^\mu \\ &\quad -2(-x)(y + z) \left[2m_e c\, p'^\mu - m_e^2c^2\gamma^\mu\right] \,, \end{aligned} \tag{100.65}$$

while the second piece of N^μ contributes $4m_e c\,[2z\, p^\mu + (y + z - x)q^\mu]$. Writing $p^\mu = \frac{1}{2}\,[(p + p')^\mu - q^\mu]$ and $p'^\mu = \frac{1}{2}\,[(p + p')^\mu + q^\mu]$ and collecting the coefficient b of $(p + p')^\mu$,

$$\frac{b}{2m_e c} = (1 - y)y + x(1 - x) - 2(1 - y)(1 - x) + 2z = 3(x + y) + 2z - (x + y)^2 - 2 \,,$$

and with $x + y = 1 - z$ this is $3 - 3z + 2z - (1 - z)^2 - 2 = z - z^2 = z(1 - z)$, so $b = 2m_e c\, z(1 - z)$. By the Gordon identity Equation (100.61) read backwards, $(p + p')^\mu = 2m_e c\,\gamma^\mu - \mathrm{i}\sigma^{\mu\nu}q_\nu$ between the spinors, so the coefficient of $\mathrm{i}\sigma^{\mu\nu}q_\nu/2m_e c$ in N^μ is $-2m_e c\, b = -4m_e^2c^2 z(1 - z)$.

It remains to do the momentum integral. From Lemma 100.26 with $n = 3$ the Feynman combination carries a factor 2, and from Equation (100.45) with $n = 3$, $d = 4$,

$$\frac{1}{(2\pi)^4} \int \mathrm{d}^4k\, \frac{1}{(k^2 - \Delta + \mathrm{i}\epsilon)^3} = \frac{-\mathrm{i}}{32\pi^2\Delta} \,,$$

which is finite: the F_2 part needs no regulator, and this is why a_e is a prediction while F_1 is not. Assembling,

$$\begin{aligned} F_2\left(q^2\right) &= -4\pi\mathrm{i}\alpha \int_0^1 \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z\; \delta\,(x + y + z - 1)\; 2\left(-4m_e^2c^2 z(1 - z)\right) \frac{-\mathrm{i}}{32\pi^2\Delta} \\ &= \frac{\alpha\, m_e^2c^2}{\pi} \int_0^1 \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z\; \delta\,(x + y + z - 1)\, \frac{z(1 - z)}{\Delta} \,, \end{aligned} \tag{100.66}$$

the prefactor being that of Equation (100.63), in which every $\hbar$ has already cancelled. At $q^2 = 0$, Equation (100.64) gives $\Delta = (1 - z)^2 m_e^2c^2$, so

$$F_2(0) = \frac{\alpha}{\pi} \int_0^1 \mathrm{d}z \int_0^{1-z} \mathrm{d}y\, \frac{z(1 - z)}{(1 - z)^2} = \frac{\alpha}{\pi} \int_0^1 \mathrm{d}z\, (1 - z) \frac{z}{1 - z} = \frac{\alpha}{\pi} \int_0^1 z\, \mathrm{d}z = \frac{\alpha}{2\pi} \,,$$

the y integral running over the length $1 - z$ of the simplex. Numerically $\alpha/2\pi = 1.161\,410 \times 10^{-3}$. □

Remark 100.42 (What the calculation used and what it did not). Not one regulator appears in the last three lines: F_2 is ultraviolet finite because the $(p+p')^\mu$ terms of the numerator carry no power of the loop momentum. The divergence of the vertex sits entirely in F_1, in the $k^2\gamma^\mu$ term dropped above, and is removed in Section 100.5.1 by the condition $F_1(0)=1$, which says that the charge measured at large distance is the charge that appears in Coulomb's law. That condition is imposed exactly, so F_1 carries no infrared divergence at $q^2=0$: it is 1 there by definition. The infrared divergence of the vertex sits at *nonzero* momentum transfer, where the subtracted one-loop F_1 carries $\ln\left(-q^2/m_\gamma^2c^2\right)$ and blows up as the photon-mass regulator is removed, the more so the larger $-q^2$; it is not a defect of the form factor but of the exclusive quantity it is being asked to give, and it cancels against the emission of soft real photons in Equation (100.80), which is the subject of Section 100.6. F_2 is infrared finite at every q^2, because the soft region is suppressed by the explicit factor $z(1-z)$. The prediction of Equation (100.62) is therefore clean in a way no other one-loop quantity in this chapter is, and it is measured in Chapter 112 to thirteen figures.

100.4.4 Regularization

A divergent integral is not a number and cannot be manipulated. A regulator is a one-parameter deformation of the theory that makes every integral finite, chosen so that the original is recovered in a limit; the divergences then appear as terms that blow up in that limit and are disposed of by Section 100.5.1. Three regulators are used in practice and it is worth saying exactly what each costs.

- *Momentum cutoff.* Restrict $|\ell_{\mathrm{E}}|<\Lambda$. Simple, physically suggestive — Λ looks like the momentum above which the theory is not to be trusted — and *it violates gauge invariance.* A sharp cutoff is not translation invariant in loop momentum, so the shift $k=\ell+xq$ used in Theorem 100.31 is illegal, and the cancellation that made $\Pi^{\mu\nu}$ transverse fails: one finds a residual $\Lambda^2\eta^{\mu\nu}$, that is a photon mass, contradicting Phenomenon 100.7. The term can be removed by hand, and every pre-1972 calculation did so, but subtracting a symmetry violation by hand is exactly the kind of step that cannot be automated to high order.

- *Pauli–Villars.* Subtract from each propagator a copy with a large mass M and the wrong-sign residue [945], so that $1/(\ell^2-m^2c^2)\to 1/(\ell^2-m^2c^2)-1/(\ell^2-M^2c^2)$, which falls as ℓ^{-4} and makes a logarithmically divergent integral converge; $M\to\infty$ at the end. This preserves Lorentz invariance and, in QED, gauge invariance, and it was the standard tool of the 1950s. Its price is that the regulator fields have negative norm, so the regulated theory is not unitary, and that it does not extend cleanly to non-abelian gauge theories.

- *Dimensional regularization.* Continue the number of spacetime dimensions to $d=4-2\varepsilon$ as in Definition 100.30 [1463, 1465]. Nothing in the derivation of Lemma 100.28 required d to be an integer, and the result is an analytic function of d with poles at $d=4,6,\dots$; the four-dimensional divergence is the residue at $d=4$. It preserves Lorentz invariance and gauge invariance exactly, because the continuation touches only the measure and never the symmetry of the Lagrangian, and it makes the algebra shorter. It is what made the renormalization of non-abelian theories tractable [1448] and hence what Chapter 102 and Chapter 104 use throughout.

Proposition 100.43 (The two regulators agree on what is divergent)**.** *For a logarithmically divergent one-loop integral, the substitution*

$$\frac{1}{\varepsilon}-\gamma_{\mathrm{E}}+\ln 4\pi-\ln\frac{\Delta}{\bar{\mu}^2}\quad\leftrightarrow\quad\ln\frac{\Lambda^2}{\Delta} \tag{100.67}$$

maps the dimensionally regularized result onto the cutoff one, up to terms that vanish as $\varepsilon \to 0$ *and* $\Lambda \to \infty$*. Differences of the same quantity at two kinematic points — which is what every observable is — are independent of the choice.*

Rests on Equation (100.44) *and Lemma 100.29.*

Proof. *Derives Proposition 100.43.*

In the cutoff scheme, $\int^{\Lambda} \mathrm{d}^4\ell_{\mathrm{E}}\,(\ell_{\mathrm{E}}^2+\Delta)^{-2} = 2\pi^2 \int_0^{\Lambda} \ell^3 \mathrm{d}\ell\,(\ell^2+\Delta)^{-2} = \pi^2\left[\ln(\Lambda^2/\Delta) - 1 + O(\Delta/\Lambda^2)\right]$; in dimensional regularization the same integral is $\pi^2\left[\varepsilon^{-1} - \gamma_{\mathrm{E}} + \ln 4\pi - \ln(\Delta/\bar{\mu}^2)\right]$ by Equation (100.44) and Lemma 100.29. The two differ by a Δ-independent constant, which is Equation (100.67); and in a difference at two values of Δ the constant cancels together with Λ and $\bar{\mu}$. □

Remark 100.44 (Regularization is bookkeeping, and must leave no trace). A regulator is not a physical hypothesis. Λ is not the energy at which new physics appears, M is not a particle, and ε is not a statement that spacetime has 3.9999 dimensions. The requirement on a calculation is that no observable depend on which regulator was used, and Proposition 100.43 is the one-loop instance of that requirement. The genuinely physical reading of a cutoff — as the scale beyond which the theory is replaced by another — belongs to the effective-field-theory picture of Chapter 108, and is a different statement about a different object. One warning belongs here: dimensional regularization is not innocent for γ^5, which has no d-dimensional definition compatible with all its four-dimensional properties, and the resulting care is unavoidable in the chiral theories of Chapter 101 and Chapter 104. In QED, which uses no γ^5 in its interaction, the problem does not arise.

100.5 Renormalization

Three amplitudes diverge. The claim of renormalization is that the divergences can be absorbed, once and for all, into the definitions of three quantities that were never predicted in the first place — the electron's mass, its charge, and the normalization of the fields — and that after this is done every remaining prediction is finite. Nothing is thrown away: what is thrown away is the pretence that the parameters written in Equation (100.12) were the measured ones.

100.5.1 Counterterms and renormalization constants

Definition 100.45 (Bare and renormalized quantities)**.** Regard the fields and parameters of Equation (100.12) as *bare*, write them ψ_0, A_0^μ, m_0, e_0, and define renormalized ones by

$$\psi_0 = \sqrt{Z_2}\,\psi\,, \qquad A_0^\mu = \sqrt{Z_3}\,A^\mu\,, \qquad m_0 = m_e - \delta m\,, \qquad e_0 = \frac{Z_1}{Z_2\sqrt{Z_3}}\,e\,, \tag{100.68}$$

where Z_1 is defined so that the interaction term takes the same form in renormalized variables and δm is the positive shift of Equation (100.58), so that the sign here is the one forced by Corollary 100.36: the self-interaction adds to the mass, hence the bare mass is the smaller. Writing $Z_i = 1 + \delta_i$, the Lagrangian splits as

$$\begin{aligned}\mathcal{L} = {}& \bar{\psi}\left(\mathrm{i}\hbar c\gamma^\mu\partial_\mu - m_e c^2\right)\psi - \frac{1}{4\mu_0}F_{\mu\nu}F^{\mu\nu} + ec\,\bar{\psi}\gamma^\mu\psi A_\mu \\ & + \delta_2\,\mathrm{i}\hbar c\,\bar{\psi}\gamma^\mu\partial_\mu\psi - (\delta_2 m_e - Z_2\delta m)\,c^2\bar{\psi}\psi - \frac{\delta_3}{4\mu_0}F_{\mu\nu}F^{\mu\nu} + \delta_1 ec\,\bar{\psi}\gamma^\mu\psi A_\mu\,,\end{aligned} \tag{100.69}$$

the mass counterterm being the whole of $Z_2 m_0 - m_e = \delta_2 m_e - Z_2\delta m$ left over when $-Z_2 m_0 c^2\bar{\psi}\psi$ has $-m_e c^2\bar{\psi}\psi$ removed from it; the first line is the original theory with *measured* parameters and the second a set of *counterterms*, each of which supplies an extra Feynman rule and each of whose coefficients is fixed — order by order in α, and divergent — by a renormalization condition.

Definition 100.46 (On-shell renormalization conditions)**.** The four constants are fixed by requiring that the renormalized parameters be the measured ones:

1. the full electron propagator has its pole at $\gamma\cdot p = m_e c$, with m_e the mass in the mass-shell relation Equation (100.1) — this fixes δm;
2. the residue at that pole is $\mathrm{i}\hbar$, i.e. the same as the free propagator — this fixes δ_2;
3. the full photon propagator has a pole at $q^2=0$ with residue equal to the free one — this fixes δ_3, through $\Pi_2(0)=0$ after subtraction;
4. $F_1(0)=1$, i.e. the charge measured in the long-wavelength limit — in Thomson scattering, or in Coulomb's law by Theorem 100.25 — is e — this fixes δ_1.

At one loop the conditions give counterterms that cancel exactly the divergent parts computed in Theorem 100.31, Theorem 100.35 and Theorem 100.41. For the photon, Equation (100.50) evaluated at $q^2=0$ has divergent part

$$\Pi_2(0)\big|_{\mathrm{div}} = -\frac{2\alpha}{\pi}\int_0^1 \mathrm{d}x\; x(1-x)\,\frac{1}{\varepsilon} = -\frac{\alpha}{3\pi}\,\frac{1}{\varepsilon}\,, \tag{100.70}$$

since $\int_0^1 x(1-x)\mathrm{d}x = 1/6$, and δ_3 is whatever cancels it; δ_1 and δ_2 are likewise proportional to α/ε, equal to each other by Corollary 100.49, and $\delta m = 3\alpha m_e/4\pi\varepsilon + \dots$ by Corollary 100.36. The essential point is not their values but their number. There are four, they are the same four at every order, and Theorem 100.17 guarantees no fifth is ever needed: that is renormalizability (Definition 100.18) and it was Dyson's theorem [928].

Remark 100.47 (What a "physical parameter" means afterwards). After renormalization the symbols m_e and e in Equation (100.69) denote numbers taken from experiment, and the bare ones are formally infinite. This is uncomfortable only if one insists that the bare parameters were meaningful. They were not: no measurement isolates the coupling of a Lagrangian written at infinite momentum. What the theory predicts is every *other* quantity in terms of those two, and the count of the inputs — how many numbers physics has to be told rather than derive — is the subject of the free-parameters appendix, Appendix C.

100.5.2 The Ward–Takahashi identity

Of the four constants, two turn out to be equal, and the equality is not an accident of one-loop arithmetic.

Theorem 100.48 (Ward–Takahashi)**.** *Let $S(p) = \mathrm{i}\hbar\,(\gamma\cdot p - m_e c - \Sigma(p))^{-1}$ be the exact electron propagator and $\Gamma^\mu(p',p)$ the exact vertex function of Equation (100.60), both off shell. Then*

$$q_\mu\,\Gamma^\mu(p+q,p) = \left[\gamma\cdot(p+q) - m_e c - \Sigma(p+q)\right] - \left[\gamma\cdot p - m_e c - \Sigma(p)\right]\,, \tag{100.71}$$

to all orders in α [973, 1132]. In the limit $q\to 0$ this is Ward's identity

$$\Gamma^\mu(p,p) = \frac{\partial}{\partial p_\mu}\left[\gamma\cdot p - m_e c - \Sigma(p)\right]\,. \tag{100.72}$$

Rests on Equations (100.6) and (100.60) and Theorem 100.5.

Proof. *Derives Theorem 100.48.*

Consider the Green function $G^\mu(x;y,z) = \left\langle T\, j^\mu(x)\,\psi(y)\,\bar\psi(z)\right\rangle$ with j^μ the current Equation (100.6). Differentiating, the divergence picks up two contributions: one from $\partial_\mu j^\mu = 0$, which vanishes, and one from differentiating the time-ordering symbol, which produces

equal-time commutators. Using the canonical relation of Chapter 99, $[j^0(x), \psi(y)]\,\delta(x^0 - y^0) = -qc\,\delta^4(x-y)\,\psi(y)$, and similarly for $\bar\psi$ with the opposite sign,

$$\partial_\mu G^\mu(x;y,z) = -qc\left[\delta^4(x-y) - \delta^4(x-z)\right]\left\langle T\,\psi(y)\bar\psi(z)\right\rangle .$$

Fourier transforming with Equation (100.14) turns $\partial_\mu \mapsto -\mathrm{i}q_\mu/\hbar$ on the left and the delta functions into unity on the right, so that

$$q_\mu\, S(p+q)\,\Gamma^\mu(p+q,p)\,S(p) \;\propto\; S(p) - S(p+q)\,,$$

the left-hand side being the amputated vertex reattached to two full propagators. Multiplying by S^{-1} on both sides gives Equation (100.71); the normalization is fixed by comparing the tree-level terms, where $\Gamma^\mu = \gamma^\mu$ and $q_\mu\gamma^\mu = \gamma\cdot(p+q) - \gamma\cdot p$ identically. Equation (100.72) follows by dividing by q and letting $q \to 0$. The only inputs are current conservation, which is Noether's theorem applied to the exact symmetry of Theorem 100.5, and the canonical commutator; neither is perturbative, so the identity holds order by order and also non-perturbatively. □

Corollary 100.49 ($Z_1 = Z_2$, and the universality of charge)**.** *The vertex and field-strength renormalizations of Equation* (100.68) *coincide,*

$$Z_1 = Z_2\,, \qquad \textit{hence} \qquad e_0 = \frac{e}{\sqrt{Z_3}}\,. \tag{100.73}$$

The renormalization of the charge therefore involves only Z_3, which belongs to the photon and knows nothing about the matter field. Consequently the observed charges of the electron, the muon, the tau and the proton are shifted by radiative corrections by the same factor*, and their equality — verified to one part in 10^{21} for the electron and the proton by the neutrality of bulk matter [2427] — survives quantization.*

Rests on Equations (100.68) *and* (100.72) *and Definition 100.46.*

Proof. *Derives Corollary 100.49.*

By Definition 100.46, δ_2 is fixed by the derivative of Σ at the mass shell and δ_1 by $\Gamma^\mu(p,p) = Z_1^{-1}\gamma^\mu$ at $F_1(0) = 1$. Take Equation (100.72) on shell: the right-hand side is $\gamma^\mu - \partial\Sigma/\partial p_\mu$, whose divergent part is $\gamma^\mu(1-\delta_2^{\mathrm{div}})$ by the definition of δ_2, while the left-hand side is $\gamma^\mu(1-\delta_1^{\mathrm{div}})$ by the definition of δ_1. Hence $\delta_1 = \delta_2$ to all orders, so $Z_1 = Z_2$, and Equation (100.68) collapses to Equation (100.73). Proposition 100.6 asserted universality at tree level; this is the statement that it is not destroyed at any order, which is what had to be shown, since Z_1 and Z_2 are separately mass-dependent and only their ratio is not. □

Remark 100.50 (Gauge invariance, current conservation, transversality)*.* The same identity has three further uses in this chapter, all already invoked. It is what makes $q_\mu\mathcal{M}^\mu = 0$ for an external photon, hence what licenses the polarization sum Equation (100.34); it is what makes $\Pi^{\mu\nu}$ transverse, hence what protects Phenomenon 100.7 beyond one loop; and it is what makes the gauge parameter ξ of Equation (100.17) drop out of every observable. In non-abelian theories the corresponding statements are the Slavnov–Taylor identities [1458, 1484], which are more complicated because the gauge-fixing procedure itself introduces the ghost fields of [1347]; they are what makes the proof of renormalizability of Chapter 102 and Chapter 104 possible [1448].

100.5.3 All orders: the BPHZ theorem

One loop is a special case in which every divergence is overall. At higher order a diagram contains divergent *subdiagrams*, and the counterterms generated at lower order must cancel them before the overall divergence is even well defined.

Definition 100.51 (Nested, disjoint and overlapping divergences)**.** Two divergent subdiagrams of a given graph are *disjoint* if they share no line, *nested* if one is contained in the other, and *overlapping* if they share some but not all of their lines. The first two cases are handled by recursion on the containment ordering. The third is not, and it is the reason renormalization to all orders took twenty years: at two loops the electron self-energy already contains two vertex subdiagrams that overlap.

Theorem 100.52 (Bogoliubov–Parasiuk–Hepp–Zimmermann)**.** *Define, recursively, the subtracted integrand* $\bar{R}_G = I_G + \sum_{\gamma \subset G} (\prod \text{counterterms of } \gamma)\, I_{G/\gamma}$ *and* $R_G = (1 - T_G)\,\bar{R}_G$*, where* T_G *is the Taylor expansion in the external momenta to order* $D(G)$ *of Equation* (100.26)*. Then* R_G *is absolutely convergent for every graph* G *of QED [1103, 1331, 1411], and* R_G *is generated by the counterterm Lagrangian Equation* (100.69) *with the four constants of Definition 100.45 alone. Equivalently, Zimmermann's forest formula*

$$R_G = \sum_{F \in \mathcal{F}(G)} \prod_{\gamma \in F} (-T_\gamma)\; I_G \tag{100.74}$$

— the sum running over all forests*, that is all sets of divergent subgraphs that are pairwise nested or disjoint — solves the recursion in closed form. QED is therefore finite to all orders in perturbation theory once* m_e*,* e *and the two field normalizations are fixed.*

Rests on Equations (100.26) *and* (100.69) *and Definition 100.45.*

Derivation pending #138

The Bogoliubov–Parasiuk–Hepp–Zimmermann theorem: the recursive subtraction, the combinatorial proof that the forest formula solves the recursion, and Hepp's sector decomposition establishing absolute convergence of the subtracted integrand, together with the demonstration that overlapping divergences require no counterterm beyond those of the local Lagrangian. This is a long combinatorial argument and belongs in Appendix A; the statement is used here only to license the claim that the four constants suffice at every order.

Remark 100.53 (Why the counterterms are local)*.* The operation T_γ is a Taylor expansion in external momenta, so the counterterm it produces is a *polynomial* in momenta of degree $D(\gamma)$, that is a local operator in position space. This is not a technical remark: it is why a finite number of terms in the Lagrangian can absorb an infinite number of divergent integrals. Had the divergences been non-polynomial in the external momenta, no local counterterm could have removed them and the theory would not have been renormalizable in any useful sense.

100.5.4 Scheme dependence and the meaning of a renormalized parameter

The conditions of Definition 100.46 are one choice among many. A *scheme* is any prescription that fixes the finite parts of the counterterms; the divergent parts are common to all of them.

Definition 100.54 (Two schemes)**.** In the *on-shell* scheme the counterterms are fixed by Definition 100.46, so that m_e is the pole mass and e the Thomson-limit charge. In *minimal subtraction* the counterterms remove only the pole $1/\varepsilon$; in *modified minimal subtraction*, written $\overline{\text{MS}}$, they remove the whole recurring combination

$$\frac{1}{\varepsilon} - \gamma_{\text{E}} + \ln 4\pi\,, \tag{100.75}$$

which by Lemma 100.28 accompanies every pole. The renormalized parameters then depend on the arbitrary scale $\bar{\mu}$ of Definition 100.30: one speaks of the running mass $m_e(\bar{\mu})$ and the running coupling $\alpha(\bar{\mu})$.

Proposition 100.55 (Scheme independence of observables)**.** *Two schemes give the same value for any S-matrix element expressed in terms of measured quantities. They differ only in what the symbols mean.*

Rests on Definitions 100.46 and 100.54.

Proof. *Derives Proposition 100.55.*

Both schemes are related by a finite redefinition $\alpha \mapsto \alpha' = \alpha(1 + c_1\alpha + \dots)$, $m \mapsto m' = m(1 + d_1\alpha + \dots)$ with c_i, d_i finite. An observable computed to order α^n in one scheme, re-expanded in the other's parameters, differs only at order α^{n+1} — that is, by terms of the order that has not been computed. Once the parameters are eliminated in favour of two *measurements*, the prediction for a third quantity is scheme independent at every order at which it is complete. This is the criterion that separates convention from physics in this chapter: α, m_e, ξ, $\bar{\mu}$ and the choice of regulator are conventions; the relation between Equation (100.62) and the Lamb shift is not. □

Remark 100.56 (Two numbers must be measured)*.* Nothing in QED predicts α or m_e. They are two of the free parameters of the Standard Model catalogued in Appendix C, and their values are taken from the CODATA and Particle Data Group compilations [2427, 2431]. Everything else in this chapter is a prediction, which is why the electron anomaly, the Lamb shift and light-by-light scattering are tests rather than fits — and why the several independent routes to α of Section 100.8.4 matter so much: agreement among them is what converts a fitted parameter into a checked one.

100.6 Infrared structure

The divergences of Section 100.4 came from the ultraviolet. QED has a second kind, from the opposite end of the spectrum, and its resolution is physical rather than formal: it says something true about what a detector can measure.

100.6.1 Soft photons and the Bloch–Nordsieck mechanism

Phenomenon 100.57 (No charge is deflected without radiating)**.** Every deflection of a charged particle is accompanied by electromagnetic radiation, and the number of quanta emitted with energy below $\hbar\omega$ grows without bound as $\omega \to 0$: the "elastic" peak recorded in any scattering experiment is therefore not a property of Nature but a statement about the resolution of the apparatus, and every measured cross-section is inclusive over radiation softer than that resolution [1976]. In perturbation theory the same fact appears as an infrared divergence of the elastic amplitude taken by itself [788].

Rests on Equations (100.23), (100.34) and (100.39).

Derivation. *Derives Phenomenon 100.57.*

Let $\mathcal{M}_0$ be the amplitude for a process in which an electron of momentum p is scattered into momentum p', and attach one outgoing photon of momentum k and polarization $\varepsilon_{(\lambda)}$. The photon may be emitted from the outgoing leg or the incoming one. In the first case the electron propagator immediately before emission carries momentum $p' + k$, so by Equation (100.23) it contributes

$$\frac{\mathrm{i}\hbar\,(\gamma\cdot(p'+k)+m_e c)}{(p'+k)^2 - m_e^2c^2} = \frac{\mathrm{i}\hbar\,(\gamma\cdot(p'+k)+m_e c)}{2\,p'\cdot k}\,,$$

using $p'^2 = m_e^2c^2$ and $k^2 = 0$. As $k \to 0$ the denominator vanishes linearly, so this term dominates; in the numerator $\gamma\cdot k$ may be dropped and $(\gamma\cdot p' + m_e c)\,\gamma^\mu \varepsilon^*_\mu u \to 2p'\cdot\varepsilon^*\,u$ by the Dirac equation. The emission from the incoming leg gives the same structure with $p' \to p$

and an opposite sign, because there the propagator carries $p - k$. Hence the amplitude *factorizes*,

$$\mathcal{M}_{\text{soft}} = \mathcal{M}_0 \;\times\; \mathcal{J}^{\mu} \varepsilon_{\mu}^{*} , \qquad \mathcal{J}^{\mu} = q \left(\frac{p'^{\mu}}{p' \cdot k} - \frac{p^{\mu}}{p \cdot k} \right) , \tag{100.76}$$

with $\mathcal{J}^{\mu}$ the eikonal current, of dimension $\mathrm{C\,s/(kg\,m)}$; the whole dependence on the hard process has dropped out. Squaring, summing over the two polarizations with Equation (100.34), and integrating over the photon phase space of Equation (100.39), the expected number of photons emitted is

$$\mathrm{d}\,\langle n \rangle = \frac{\alpha}{4\pi^2} \left[\frac{2\, p \cdot p'}{(p \cdot k)\,(p' \cdot k)} - \frac{m_e^2 c^2}{(p \cdot k)^2} - \frac{m_e^2 c^2}{(p' \cdot k)^2} \right] \frac{\mathrm{d}^3 k}{|\boldsymbol{k}|} , \tag{100.77}$$

which is dimensionless as it must be, the bracket carrying the inverse square of a momentum and $\mathrm{d}^3 k / |\boldsymbol{k}|$ its square. Every term in the bracket scales as $|\boldsymbol{k}|^{-2}$, while $\mathrm{d}^3 k = |\boldsymbol{k}|^2 \,\mathrm{d}\,|\boldsymbol{k}|\,\mathrm{d}\Omega$, so the photon-momentum dependence is exactly

$$\mathrm{d}\,\langle n \rangle = \frac{\mathrm{d}\,|\boldsymbol{k}|}{|\boldsymbol{k}|} \times (\text{a function of angles only}) = \frac{\mathrm{d}\omega}{\omega} \times (\dots) , \tag{100.78}$$

and $\int_0^{\omega_{\max}} \mathrm{d}\omega/\omega$ diverges logarithmically at the lower limit. The number of photons emitted below any energy is infinite; the *energy* they carry, $\int \hbar\omega \,\mathrm{d}\,\langle n \rangle \propto \int \mathrm{d}\omega$, is finite, which is why nothing physical is wrong. The angular integration for an ultrarelativistic deflection through momentum transfer q gives

$$\langle n \rangle = \frac{2\alpha}{\pi} \left(\ln \frac{-q^2}{m_e^2 c^2} - 1 \right) \ln \frac{\omega_{\max}}{\omega_{\min}} , \tag{100.79}$$

the double logarithm of Bloch and Nordsieck [788] and Weinberg [1322]. Since $\langle n \rangle \to \infty$ as $\omega_{\min} \to 0$, the probability of emitting *no* photon at all, which for the Poisson distribution the soft emissions obey is $\mathrm{e}^{-\langle n \rangle}$, is zero. The elastic peak of any real experiment is the set of events in which no photon above the detector threshold was emitted, and it is therefore a property of the threshold. □

Theorem 100.58 (Bloch–Nordsieck cancellation)**.** *The infrared divergence of the virtual correction to the elastic cross-section cancels exactly against the divergence of the real emission cross-section integrated up to the detector resolution E_{res}. Explicitly, with a photon mass m_γ as regulator,*

$$\left| F_1\!\left(q^2\right) \right|^2 = 1 - \frac{\alpha}{\pi}\, \mathcal{B}\!\left(q^2\right) \ln \frac{-q^2}{m_\gamma^2 c^2} + \dots , \qquad \frac{\mathrm{d}\sigma_{\text{real}}}{\mathrm{d}\sigma_0} = + \frac{\alpha}{\pi}\, \mathcal{B}\!\left(q^2\right) \ln \frac{E_{\text{res}}^2}{m_\gamma^2 c^4} + \dots \tag{100.80}$$

with the same function $\mathcal{B}$, so that the sum

$$\mathrm{d}\sigma_{\text{incl}} = \mathrm{d}\sigma_0 \left[1 - \frac{\alpha}{\pi}\, \mathcal{B}\!\left(q^2\right) \ln \frac{-q^2 c^2}{E_{\text{res}}^2} + \dots \right] \tag{100.81}$$

is finite and depends on E_{res}, which is a property of the apparatus. Exponentiating the soft series replaces the bracket by $\exp\left[-(\alpha/\pi)\, \mathcal{B} \ln\left(-q^2 c^2 / E_{\text{res}}^2\right) \right]$, the Sudakov form factor, which vanishes as $E_{\text{res}} \to 0$: a strictly elastic cross-section is not merely divergent, it is zero.

Rests on Phenomenon 100.57 and Equation (100.60).

Derivation pending #139

Bloch–Nordsieck cancellation: the explicit infrared part of the one-loop form factor with a photon mass regulator, the integral of the eikonal emission rate up to the detector resolution, the demonstration that the coefficient functions of the two logarithms are equal and opposite, and the exponentiation of the soft series into the Sudakov form factor. The structure and its physical reading are established above; what is owed is the coefficient function, and it belongs in Appendix A.

Remark 100.59 (The observable is inclusive). The lesson survives its perturbative origin. In QED an exclusive final state containing a definite number of photons is not an observable at all, because states differing by arbitrarily soft quanta are not distinguishable by any apparatus; the physical quantities are sums over degenerate final states. The same statement, with gluons in place of photons, is what makes a *jet* rather than a quark the observable of Chapter 102. It is also, in a different guise, the reason a charged particle has no isolated asymptotic state — the tension with the scattering formalism of Chapter 86, which assumes free asymptotic particles, is real and is one of the foundational loose ends recorded in Chapter 109.

100.6.2 Mass singularities

A second infrared problem appears when a mass is small rather than a momentum. Setting $m_e = 0$ in Equation (100.77) makes the angular integral diverge as well, because a photon emitted exactly collinear with a massless charge costs nothing.

Theorem 100.60 (Kinoshita–Lee–Nauenberg)**.** *Mass singularities of the form* $\ln(m_e^2 c^2/s)$ *cancel in any transition probability summed over all initial and final states degenerate in energy in the massless limit [1232, 1288]. In QED with massive charges the singularities are cut off by the mass and appear as large logarithms rather than divergences; in the massless limit only sufficiently inclusive quantities are finite.*

Rests on Theorem 100.58 and Phenomenon 100.57.

Derivation pending #140

Kinoshita–Lee–Nauenberg theorem: the classification of the singular regions of a Feynman integral in the massless limit — soft and collinear — and the proof that the sum over the degenerate subspace of initial and final states is free of them, together with the counting that shows why an average over degenerate initial states is needed and not only a sum over final ones. It belongs in Appendix A.

Remark 100.61 (Where this reappears). Two consequences are used elsewhere in this part. First, the collinear logarithms that survive when only final states are summed are precisely the ones resummed by the evolution equations of Altarelli and Parisi [1591], which govern the scaling violations measured in Chapter 116. Second, the same structure with colour in place of charge is why the observable of a strong-interaction final state is a jet defined by a resolution parameter, and never a quark: see Chapter 102.

100.7 The running coupling

100.7.1 Gell-Mann–Low and the beta function

Phenomenon 100.62 (The fine-structure constant runs)**.** The strength of the electromagnetic interaction is not one number but a function of the momentum transfer at which

it is measured. At vanishing momentum transfer $\alpha^{-1} = 137.035999\ldots$ [2431], whereas the effective coupling extracted from e^+e^- data at the Z pole, $\sqrt{s}\,c = 91.19\,\mathrm{GeV}$, satisfies $\alpha^{-1} \approx 129$ [2427]. The difference is some six percent, thousands of times any experimental uncertainty. It is measured independently in small-angle Bhabha scattering, in the Z lineshape and in the hadronic contribution to the vacuum polarization, and the three agree.

Rests on Equation (100.52)*, Proposition 100.33, and Definition 100.46.*

Derivation. *Derives Phenomenon 100.62.*

The measured coupling at momentum transfer q is the one that appears in the dressed propagator Equation (100.52): the product of two charges and one photon exchange carries $\alpha/\left[1-\Pi_2(q^2)\right]$ where α is the Thomson-limit value fixed by Definition 100.46. Since the condition there was $\Pi_2(0) = 0$ after subtraction, the effective coupling is

$$\alpha_{\mathrm{eff}}\left(q^2\right) = \frac{\alpha}{1-\left[\Pi_2(q^2)-\Pi_2(0)\right]}\,. \tag{100.82}$$

By Proposition 100.33 the subtracted polarization is

$$\Pi_2(q^2)-\Pi_2(0) = -\frac{2\alpha}{\pi}\int_0^1 \mathrm{d}x\; x(1-x)\ln\frac{m_e^2c^2}{m_e^2c^2 - x(1-x)q^2}\,,$$

and for $-q^2 \gg m_e^2c^2$ the logarithm becomes $\ln\left[m_e^2c^2/\left(-x(1-x)q^2\right)\right]$. With $\int_0^1 x(1-x)\mathrm{d}x = 1/6$ and $\int_0^1 x(1-x)\ln\left[x(1-x)\right]\mathrm{d}x = 2\int_0^1 x(1-x)\ln x\,\mathrm{d}x = 2\left(-\frac{1}{4}+\frac{1}{9}\right) = -\frac{5}{18}$,

$$\Pi_2(q^2)-\Pi_2(0) = \frac{\alpha}{3\pi}\left(\ln\frac{-q^2}{m_e^2c^2} - \frac{5}{3}\right)\,, \tag{100.83}$$

so that the effective coupling *grows* logarithmically with $-q^2$. Numerically, running from $q^2 = 0$ to $-q^2 = m_Z^2c^2$ with only the electron in the loop gives $\ln\left(m_Z/m_e\right)^2 = \ln\left(1.78\times10^5\right)^2 = 24.2$ and $\alpha\left(24.2-5/3\right)/3\pi = 0.0174$, which is the electron's share of the shift. The muon and the tau contribute 0.0092 and 0.0048 by the same formula with their own masses, and the quarks contribute about 0.0276, weighted by squared charge and by three colours and evaluated from data rather than from Equation (100.50) for the reason given below; the total $\Delta\alpha \approx 0.059$ is the observed $\alpha^{-1}(m_Z^2c^2) \approx 129$. The physical reading is Remark 100.34: a probe of higher momentum transfer penetrates further into the polarization cloud and sees a larger charge. □

Proposition 100.63 (The QED beta function)**.** *Defining the running coupling by evaluating Equation* (100.82) *at* $-q^2 = \bar{\mu}^2$*,*

$$\bar{\mu}\,\frac{\mathrm{d}\alpha}{\mathrm{d}\bar{\mu}} = \beta(\alpha) = \frac{2\alpha^2}{3\pi}\sum_f N_c^{(f)}\left(\frac{q_f}{e}\right)^2 + O\left(\alpha^3\right)\,, \tag{100.84}$$

the sum running over the charged fermions with $m_fc^2 \lesssim \bar{\mu}c$ *and* $N_c = 3$ *for quarks,* 1 *for leptons. The sign is positive: QED is not asymptotically free.*

Rests on Equations (100.82) *and* (100.83)*.*

Proof. *Derives Proposition 100.63.*

From Equation (100.83) at one loop, $\alpha(\bar{\mu})^{-1} = \alpha^{-1} - (3\pi)^{-1}\ln\left(\bar{\mu}^2/m_e^2c^2\right)$ for a single unit-charge fermion. Differentiating with respect to $\ln\bar{\mu}$, $-\alpha^{-2}\,\mathrm{d}\alpha/\mathrm{d}\ln\bar{\mu} = -2/3\pi$, which is Equation (100.84). Each additional species contributes an identical loop with e^2 replaced by q_f^2 and multiplied by its number of colours, whence the sum; each decouples below its own mass, which is the decoupling of heavy states [1538], implicit in the mass dependence of Δ in Equation (100.50). The sign is fixed by the fermion loop being a screening medium (Remark 100.34), and it is the opposite of the non-abelian case [1501, 1511]. The systematic apparatus — the Callan–Symanzik equation [1415, 1435], anticipated by Gell-Mann and Low [1038] and, earlier still, by Stueckelberg and Petermann [1028] — is Chapter 108. □

100.7.2 Measured running of α

The running of Equation (100.82) is not a bookkeeping device: it is measured, and it is the cleanest available evidence that renormalization describes something in Nature rather than a way of organizing arithmetic.

- At $q^2 = 0$ the coupling is $\alpha^{-1} = 137.035999177(21)$ [2431], obtained from the electron anomaly and from atom-recoil interferometry (Section 100.8.4).
- Small-angle Bhabha scattering at e^+e^- colliders measures α at spacelike q^2 directly: the t-channel pole of Section 100.3.1 makes the rate large and its QED prediction essentially exact, so a departure of the measured angular distribution from the fixed-coupling prediction *is* the running.
- At the Z pole, $\sqrt{s}\,c = (91.1880 \pm 0.0020)\,\mathrm{GeV}$, the effective coupling is $\alpha^{-1} \approx 128.9$, a shift $\Delta\alpha \approx 0.059$ of which roughly half is leptonic and computable from Equation (100.83) and roughly half hadronic and extracted from measured $e^+e^- \to$ hadrons cross-sections through a dispersion relation [2427].

The hadronic piece deserves a word, because it is where a QED calculation stops being purely perturbative. The contribution of quark loops to Π_2 cannot be computed from Equation (100.50) at low q^2, since the loop momenta are of order the strong-interaction scale. It is instead obtained from data: the analyticity of $\Pi_2(q^2)$ in the cut plane, together with its behaviour at infinity, gives a dispersion relation expressing Π_2 at spacelike q^2 as an integral of its discontinuity across the timelike cut, and the discontinuity is proportional to the measured cross-section for $e^+e^- \to$ hadrons. The contour argument is the one proved in Theorem 17.78, where causality and analyticity alone are shown to relate the real and imaginary parts of a response function, transposed from a function of frequency with a pole on the axis to a function of q^2 with a branch cut; the same technique carries the hadronic contribution to the muon anomaly in Chapter 112.

Derivation pending #141

Hadronic vacuum polarization from data: the once-subtracted dispersion relation for the renormalized polarization function in the cut plane, with the subtraction fixed by the renormalization condition at zero momentum transfer and the convergence of the integral argued from the behaviour of the function at infinity; the optical theorem identifying the discontinuity across the timelike cut with the measured ratio of the annihilation cross section into hadrons to that into muon pairs; and the resulting integral over that measured ratio, which is the number quoted above as the hadronic half of the shift in the coupling at the mass of the Z boson. Only the analyticity input is proved in the treatise, in the Fourier and integral-transforms chapter; the dispersion integral itself is owed, and belongs in Appendix A.

100.7.3 The Landau pole, stated honestly

Proposition 100.64 (Where the one-loop coupling diverges)**.** *Integrating Equation* (100.84) *with only the electron in the loop, the coupling reaches infinity at*

$$\bar{\mu}_{\mathrm{L}} c = m_e c^2 \exp\!\left(\frac{3\pi}{2\alpha}\right) \approx 10^{286}\,\mathrm{eV}\,, \tag{100.85}$$

that is about 10^{277} GeV *[1047].*

Rests on Equation (100.84).

Proof. *Derives Proposition 100.64.*

$\alpha(\bar{\mu})^{-1} = \alpha^{-1} - (3\pi)^{-1} \ln(\bar{\mu}^2/m_e^2c^2)$ vanishes when $\ln(\bar{\mu}^2/m_e^2c^2) = 3\pi/\alpha$, that is at $\bar{\mu} = m_e c \exp(3\pi/2\alpha)$. With $\alpha^{-1} = 137.036$ the exponent is $3\pi \times 137.036/2 = 645.8$, so $\exp(645.8) = 10^{280.4}$, and multiplying by $m_e c^2 = 0.511 \times 10^6$ eV gives Equation (100.85). □

Remark 100.65 (What this does and does not mean). Four things must be said, and the fourth is the one usually omitted.

First, the computation extrapolates a one-loop formula over 280 orders of magnitude in energy, to a regime where the expansion parameter it is built from is of order one. Nothing in perturbation theory licenses that, and the position of the pole is not a prediction of QED in any useful sense.

Second, there is independent, non-perturbative evidence pointing the same way: lattice studies of pure U(1) gauge theory with charged matter find that the continuum limit exists only at zero renormalized coupling — *triviality* — so that a cutoff-free QED would be a free theory. If that is right, QED alone has no continuum limit and the theory is necessarily an effective one with a finite cutoff.

Third, this costs nothing at any accessible energy. The scale Equation (100.85) is larger than the Planck energy 1.22×10^{28} eV by more than 250 orders of magnitude, so the question is academic long before it is physical: some other description takes over first, and the modern reading of a renormalizable theory as a low-energy effective description with a cutoff (Chapter 108) is the honest framing throughout.

Fourth, QED is not a theory on its own. Above the electroweak scale it is a part of the $\mathrm{SU}(2) \times \mathrm{U}(1)$ theory of Chapter 104, whose hypercharge coupling runs differently and whose non-abelian factor runs with the opposite sign. The extrapolation is therefore not merely unjustified but performed on the wrong theory. What survives all four objections is the moral: the statement "$\alpha = 1/137$" is a statement about a measurement at low momentum transfer, and a theory that says nothing about its own ultraviolet completion should not be asked what happens at 10^{286} electronvolts. Programmes that claim to answer such a question — string theory, supersymmetric unification, grand unified theories — have no observational support of the kind this book requires, and are excluded from it for that reason and no other; the position is stated in Chapter 133.

100.8 Precision tests

Everything above is machinery. This section is the reason the machinery is believed.

100.8.1 The Lamb shift

Phenomenon 100.66 (The Lamb shift)**.** In atomic hydrogen the $2S_{1/2}$ and $2P_{1/2}$ levels are not degenerate, although the Dirac equation of Chapter 93 makes them exactly so: the S level lies above the P level by (1057.8298 ± 0.0032) MHz [2331], first resolved by the microwave resonance method of Lamb and Retherford [901]. No rearrangement of the one-particle Dirac theory produces the splitting; it is the first observed consequence of the interaction of a bound electron with the quantized radiation field, and its explanation [897] was the first finite answer extracted from a divergent quantum field theory.

Rests on Theorem 100.35 and Corollary 100.36.

Derivation of the dominant term. *Derives Phenomenon 100.66.*

Bethe's calculation [897] is non-relativistic and treats the electron in the Coulomb field of the proton, coupled to the quantized transverse field of Chapter 99. In the Coulomb

gauge the coupling of an electron of charge $-e$ is, to first order in the vector potential,

$$H' = \frac{e}{m_e}\,\hat{\boldsymbol{p}}\cdot\hat{\boldsymbol{A}}(\boldsymbol{x})\,, \qquad \hat{\boldsymbol{A}}(\boldsymbol{x}) = \sum_{\boldsymbol{k},\lambda}\sqrt{\frac{\hbar}{2\varepsilon_0\omega_k V}}\left(\boldsymbol{\varepsilon}_{(\lambda)}\,a_{\boldsymbol{k}\lambda}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} + \text{h.c.}\right), \tag{100.86}$$

with $\omega_k = c\,|\boldsymbol{k}|$ and V the quantization volume; the normalization is the one that makes the field energy $\sum \hbar\omega_k\left(n_{\boldsymbol{k}\lambda}+\frac{1}{2}\right)$. Second-order perturbation theory for the state $|n\rangle$ with no photons present, summed over intermediate atomic states $|m\rangle$ and one-photon states, gives

$$\Delta E_n = \sum_m \sum_{\boldsymbol{k},\lambda} \frac{|\langle m;\boldsymbol{k}\lambda|\,H'\,|n;0\rangle|^2}{E_n - E_m - \hbar\omega_k} = \frac{e^2\hbar}{2\varepsilon_0 m_e^2}\sum_m\sum_{\boldsymbol{k},\lambda}\frac{\left|\boldsymbol{\varepsilon}_{(\lambda)}\cdot\boldsymbol{p}_{mn}\right|^2}{\omega_k V\,(E_{nm}-\hbar\omega_k)}\,,$$

writing $E_{nm} = E_n - E_m$ and $\boldsymbol{p}_{mn} = \langle m|\,\hat{\boldsymbol{p}}\,|n\rangle$, and using the dipole approximation $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \approx 1$, legitimate because the relevant wavelengths are long compared with the atom. Converting the mode sum, $\sum_{\boldsymbol{k}} \to V(2\pi)^{-3}\int \mathrm{d}^3k$, and using $\sum_\lambda \int \mathrm{d}\Omega\,\left|\boldsymbol{\varepsilon}_{(\lambda)}\cdot\boldsymbol{p}\right|^2 = \frac{8}{3}\pi\,|\boldsymbol{p}|^2$ together with $k^2\mathrm{d}k/\omega = \omega\,\mathrm{d}\omega/c^3$,

$$\Delta E_n = \frac{2\alpha\hbar^2}{3\pi m_e^2 c^2}\sum_m |\boldsymbol{p}_{mn}|^2 \int_0^{\omega_{\max}} \frac{\omega\,\mathrm{d}\omega}{E_{nm}-\hbar\omega}\,, \tag{100.87}$$

where $e^2\hbar/2\varepsilon_0 m_e^2\cdot(3\pi^2c^3)^{-1} = 2\alpha\hbar^2/3\pi m_e^2c^2$ by $e^2/\varepsilon_0 = 4\pi\alpha\hbar c$. The integral diverges linearly, which is the divergence of Theorem 100.35 seen in non-relativistic disguise. Split it:

$$\frac{\omega}{E_{nm}-\hbar\omega} = -\frac{1}{\hbar} + \frac{1}{\hbar}\,\frac{E_{nm}}{E_{nm}-\hbar\omega}\,.$$

The first piece contributes $-\left(2\alpha\hbar/3\pi m_e^2c^2\right)\omega_{\max}\sum_m|\boldsymbol{p}_{mn}|^2 = -\left(2\alpha\hbar\omega_{\max}/3\pi m_e c^2\right)\langle\hat{\boldsymbol{p}}^2/m_e\rangle_n$ by completeness, that is a term proportional to the expectation of the kinetic energy: it is a change in the electron's *mass*, present for a free electron as much as a bound one, and it is removed by the mass renormalization of Corollary 100.36. This subtraction is the whole of Bethe's insight, and it is renormalization performed before anyone had a name for it. What remains is

$$\Delta E_n = \frac{2\alpha}{3\pi m_e^2c^2}\sum_m |\boldsymbol{p}_{mn}|^2\;E_{mn}\;\ln\frac{\hbar\omega_{\max}}{|E_{nm}|}\,,$$

with $E_{mn} = -E_{nm}$, the logarithm arising from $\int^{\omega_{\max}}\mathrm{d}\omega/(E_{nm}-\hbar\omega)$. Bethe cut the integral off at $\hbar\omega_{\max} = m_ec^2$, the energy above which the non-relativistic treatment is meaningless; the covariant calculation [937] later confirmed that this is the right scale and supplied the constant beside the logarithm. Defining the Bethe logarithm by $\ln\langle\Delta E\rangle$ such that the weighted average may be taken outside the sum, the remaining sum is a commutator:

$$\sum_m |\boldsymbol{p}_{mn}|^2\,(E_m-E_n) = \langle n|\,\hat{\boldsymbol{p}}\cdot(H-E_n)\,\hat{\boldsymbol{p}}\,|n\rangle = \frac{1}{2}\,\langle n|\,[\hat{\boldsymbol{p}},[H,\hat{\boldsymbol{p}}]]\,|n\rangle = \frac{\hbar^2}{2}\left\langle\nabla^2V\right\rangle_n\,,$$

using $[H,\hat{\boldsymbol{p}}] = \mathrm{i}\hbar\nabla V$ and $[\hat{p}_j,\mathrm{i}\hbar\partial_jV] = \hbar^2\nabla^2V$. For the Coulomb potential $V = -e^2/4\pi\varepsilon_0 r$ one has $\nabla^2V = (e^2/\varepsilon_0)\,\delta^3(\boldsymbol{r})$, so the sum equals $2\pi\alpha\hbar^3c\,|\psi_n(0)|^2$ and vanishes for every state with $\ell\neq 0$ — which is why the effect separates $2S$ from $2P$ at all. With $|\psi_{n00}(0)|^2 = \alpha^3m_e^3c^3/\pi n^3\hbar^3$ (Chapter 81),

$$\Delta E_{nS} = \frac{4\alpha^5 m_e c^2}{3\pi n^3}\,\ln\frac{m_ec^2}{\langle\Delta E\rangle}\,. \tag{100.88}$$

Numerically, for $n=2$ the prefactor is $4\alpha^5m_ec^2/24\pi = 8.99\times10^{-26}\,\mathrm{J}$, that is $h\times 135.6\,\mathrm{MHz}$; Bethe's estimate $\langle\Delta E\rangle \approx 17.8\,\mathrm{Ry} = 242\,\mathrm{eV}$ gives a logarithm $\ln\left(5.11\times10^5/242\right) = 7.66$

and hence 1040 MHz, against the 1000 MHz Lamb and Retherford had measured. That a divergent theory, subtracted by hand, reproduced a measured microwave frequency to four percent within weeks of the measurement is the event from which the whole of Section 100.5 follows. □

The complete one-loop shift has three pieces, of which the derivation above is the largest:

- the *self-energy* of the bound electron, positive and dominant, Equation (100.88) together with its relativistic completion [937];
- the *vacuum polarization*, negative and small, computed exactly in Proposition 100.33: $-h \times 27.1\,\mathrm{MHz}$, the only piece of the three whose sign is opposite to the total;
- the *vertex* contribution, which at low momentum transfer is the anomalous magnetic moment of Theorem 100.41 acting on the spin–orbit interaction, and is positive.

Modern theory carries the expansion to two loops and beyond and includes the finite size of the proton, whose charge radius enters through the same $\delta^3(\boldsymbol{r})$ structure as Equation (100.54); the comparison with experiment is therefore a simultaneous determination of the proton radius, and the tension between the muonic and electronic determinations that dominated the 2010s is described in Chapter 90.

Derivation pending #142

Lamb shift, covariant completion: the relativistic one-loop self-energy of a bound electron, which supplies the constant accompanying Bethe's logarithm and fixes the cutoff-independence of the result; the binding corrections organized as an expansion in $Z\alpha$; and the finite-nuclear-size term through which the proton charge radius enters the comparison. The non-relativistic leading term and the vacuum-polarization term are derived in full above; this box covers the remainder, and it belongs in Appendix A.

100.8.2 The electron anomalous magnetic moment

Phenomenon 100.67 (The electron is not a pure Dirac particle)**.** The magnetic moment of the electron exceeds the value $g = 2$ that the Dirac equation assigns to a structureless spin-$\frac{1}{2}$ particle. The excess

$$a_e := \frac{g-2}{2} \tag{100.89}$$

is measured in a one-electron quantum cyclotron to a fractional precision of order 10^{-13}, giving $a_e = 0.001\,159\,652\,180\,59 \pm 0.000\,000\,000\,000\,13$ [2419], and the QED series evaluated to tenth order [2190, 2330] reproduces it over the whole range that the independently measured value of α permits. The apparatus is Chapter 112.

Rests on Theorem 100.41 and Definition 100.39.

Derivation. *Derives Phenomenon 100.67.*

The leading term is Theorem 100.41, $a_e = \alpha/2\pi = 1.161\,410 \times 10^{-3}$, derived in full in Section 100.4.3. It already accounts for the measured value to one part in 10^3, and the remainder is the rest of the series. Because the vertex correction was shown there to be α times a dimensionless integral, the whole expansion is a power series in α/π with pure numbers as coefficients,

$$a_e = \sum_{n \geq 1} C_n \left(\frac{\alpha}{\pi}\right)^n + a_e^{\mu,\tau} + a_e^{\mathrm{had}} + a_e^{\mathrm{weak}}\,, \tag{100.90}$$

with, from [2330],

$$C_1 = \tfrac{1}{2}\,, \qquad C_2 = -0.328478965\ldots\,, \quad C_3 = 1.181241456\ldots\,,$$
$$C_4 = -1.912245\ldots\,, \quad C_5 = 6.737(159)\,.$$

These are the mass-independent coefficients $A_1^{(2n)}$ of that reference; the mass-dependent terms are the separate $a_e^{\mu,\tau}$ of Equation (100.90). C_5 is known only numerically, its uncertainty is the error of the numerical integration, and the value quoted supersedes the same group's earlier 6.675(192). C_2 and C_3 are known in closed form; C_4 required the numerical evaluation of 891 diagrams and C_5 of 12 672, and an independent recomputation of the fifth-order term [2343] does not agree with the published value within its quoted uncertainty — a fact recorded here because a single unverified computation would otherwise be load-bearing. The remaining terms are small: muon and tau loops contribute 2.75×10^{-12}, hadronic vacuum polarization and hadronic light-by-light about 1.69×10^{-12}, and the electroweak contribution about 0.03×10^{-12} [2330]. Summing at $\alpha^{-1} = 137.035999177$ reproduces 0.001 159 652 18, in agreement with [2419].

The honest statement of the test requires one more step, and it is not the one usually made. The series is a function of α, so the comparison needs α measured somewhere else. The two best independent determinations, both from photon-recoil interferometry, are

$$\alpha^{-1} = 137.035999046(27)\ \text{(caesium)}\ [2316]\,, \qquad \alpha^{-1} = 137.035999206(11)\ \text{(rubidium)}\ [2363]\,,$$

and they disagree by more than five combined standard uncertainties. Propagating the difference through $a_e = \alpha/2\pi + \ldots$ gives a spread in the predicted a_e of about 1.4×10^{-12}, roughly ten times the experimental uncertainty 1.3×10^{-13} of [2419]. What QED is confirmed to is therefore the precision of its input — about one part in 10^9 of a_e — and not the thirteen figures that the precision of a_e alone would suggest. The arithmetic is carried out in Chapter 112, and the conclusion is worth stating plainly here because the looser claim is repeated everywhere. □

Remark 100.68 (The muon, and why it is the more sensitive probe). The same series with m_e replaced by m_μ gives $a_\mu = \alpha/2\pi + \ldots$ at leading order, since Theorem 100.41 contains no mass. Sensitivity to anything other than QED enters through loops of heavier states, and those contribute in proportion to m_ℓ^2, so the muon is more sensitive than the electron by $(m_\mu/m_e)^2 \approx 4.3 \times 10^4$. The measured value $a_\mu = (116\,592\,059 \pm 22) \times 10^{-11}$ [2368, 2416] is consequently compared against a Standard Model prediction [2353] whose uncertainty is dominated not by QED but by the hadronic vacuum polarization, where a dispersive evaluation from e^+e^- data and a lattice evaluation [2375] do not agree. The status is recorded in Chapter 112; the point for this chapter is that the QED part of a_μ is not in question, and that a discrepancy, if one survives, would be a statement about hadrons or about physics beyond the Standard Model, not about Equation (100.12).

100.8.3 Light-by-light scattering

Phenomenon 100.69 (Light scatters off light)**.** Two photons in vacuum interact. Maxwell's equations are linear and forbid the process outright; in QED it proceeds through a virtual charged loop, and it has been seen in ultraperipheral lead–lead collisions at the LHC, where the coherent photon flux carried by a nucleus of charge Ze is enhanced by Z^2 and brings the tiny cross-section within reach [2277, 2322]. The QED vacuum is therefore a nonlinear optical medium, as the effective Lagrangian of Euler and Heisenberg requires [777].

Rests on Theorems 100.17 and 100.31 and Equation (100.12).

Derivation of the effective Lagrangian and the scale. *Derives Phenomenon 100.69.*

The amplitude is the $E_e = 0$, $E_\gamma = 4$ entry of Theorem 100.17, with superficial degree of divergence $D = 0$: naively logarithmically divergent. It is in fact finite, and for the same reason Theorem 100.31 was only logarithmically divergent — gauge invariance forces each of the four photon legs to appear through its field strength rather than its potential, so the amplitude carries four explicit powers of momentum and the would-be divergence is cancelled. There is accordingly no counterterm for photon–photon scattering, and the process is a parameter-free prediction.

At photon energies well below $m_e c^2$ the electron loop may be integrated out, leaving a local interaction among the field strengths. Lorentz and gauge invariance and the C, P, T symmetry of Section 100.1.4 permit only two independent quartic invariants, and the coefficients follow from expanding the exact one-loop effective action of a constant field [777, 988]:

$$\mathcal{L}_{\mathrm{EH}} = \frac{2\alpha^2 \varepsilon_0^2 \hbar^3}{45\, m_e^4 c^5} \left[\left(E^2 - c^2 B^2 \right)^2 + 7 c^2 \left(\boldsymbol{E} \cdot \boldsymbol{B} \right)^2 \right] . \tag{100.91}$$

The dimensions check: $\varepsilon_0 E^2$ is an energy density, so the bracket with its prefactor ε_0^2 carries $\mathrm{J}^2/\mathrm{m}^6$, and $\hbar^3/m_e^4 c^5$ carries m^3/J, leaving J/m^3 as an addition to Equation (100.12) must. Compare Equation (100.91) with the Maxwell term $\frac{\varepsilon_0}{2}\left(E^2 - c^2 B^2\right)$ for a pure electric field. Their ratio is $4\alpha^2 \varepsilon_0 \hbar^3 E^2 / 45 m_e^4 c^5$, and using $\varepsilon_0 \hbar c / e^2 = 1/4\pi\alpha$ it collapses to

$$\frac{\mathcal{L}_{\mathrm{EH}}}{\mathcal{L}_{\mathrm{Maxwell}}} = \frac{\alpha}{45\pi} \left(\frac{E}{E_{\mathrm{cr}}} \right)^2 , \qquad E_{\mathrm{cr}} = \frac{m_e^2 c^3}{e\hbar} = 1.3233 \times 10^{18}\, \mathrm{V/m} , \tag{100.92}$$

with E_{cr} the Schwinger field [988], at which the work done on an electron over a Compton wavelength equals its rest energy and at which the vacuum becomes unstable against pair creation. Two statements must be kept apart here, because the literature runs them together. E_{cr} is the scale that makes the *expansion parameter* E/E_{cr} of order one, and it is the field at which the weak-field expansion Equation (100.91) stops being an expansion at all, since pair creation sets in with a rate carrying $\exp\left(-\pi E_{\mathrm{cr}}/E\right)$. The two terms of the Lagrangian are *not* equal there: the loop factor $\alpha/45\pi = 5.16 \times 10^{-5}$ makes the ratio that small at $E = E_{\mathrm{cr}}$, and it would reach unity only at $\sqrt{45\pi/\alpha} \approx 139$ times E_{cr}, that is about $1.8 \times 10^{20}\, \mathrm{V/m}$ — a field the theory as written cannot be extrapolated to. Laboratory static fields are some ten orders of magnitude below E_{cr}, which is why the vacuum looks linear.

The corresponding photon–photon cross-section for unpolarized light of energy $\hbar\omega \ll m_e c^2$ in the centre-of-mass frame is

$$\sigma_{\gamma\gamma} = \frac{973}{10125\pi}\, \alpha^4 \left(\frac{\hbar}{m_e c} \right)^2 \left(\frac{\hbar\omega}{m_e c^2} \right)^6 , \tag{100.93}$$

the sixth power of the energy being the direct consequence of the four powers of field strength in Equation (100.91) together with the two-body phase space. The α^4 and the sixth power together make the cross-section utterly negligible for visible light: at $\hbar\omega = 1\,\mathrm{eV}$ it is of order $1 \times 10^{-70}\, \mathrm{m}^2$.

What makes the process observable is a heavy-ion collider. A nucleus of charge Ze passing at small impact parameter carries a coherent electromagnetic field whose equivalent photon flux scales as Z^2, so a photon–photon luminosity in Pb–Pb collisions is enhanced by $Z^4 = 82^4 \approx 4.5 \times 10^7$ over the proton case. ATLAS reported evidence at 4.4 standard deviations [2277] and then observation at 8.2 [2322], with a fiducial cross-section of about $7.8 \times 10^{-36}\, \mathrm{m}^2$ — about 78 nanobarns — consistent with the QED prediction. □

Derivation pending #143

Light-by-light scattering: the explicit four-photon amplitude from a closed charged-fermion loop, the demonstration by gauge invariance that the six box diagrams combine into a finite answer despite a superficial degree of divergence of zero, the expansion of the one-loop effective action of a constant field yielding the coefficients in the Euler–Heisenberg Lagrangian, and the phase-space integration giving the low-energy cross-section. It belongs in Appendix A.

Remark 100.70 (Related nonlinear-vacuum effects and their status). Equation (100.91) also predicts that a strong magnetic field makes the vacuum birefringent, with different refractive indices for light polarized parallel and perpendicular to the field. Laboratory searches using long optical cavities in superconducting magnets have so far produced only upper limits, still above the predicted effect, so this particular consequence is *not* yet observed and must not be listed among the confirmations. Two effects of the same family are: the deflection of a photon in the Coulomb field of a nucleus, which is the same box amplitude with two legs attached to the nucleus, and the radiation reaction and pair-creation phenomena of ultra-intense laser–electron collisions, where evidence for quantum radiation reaction has been reported [2305, 2317]. None of this leaves QED: it is the same Lagrangian Equation (100.12) evaluated where its perturbative expansion in the external field breaks down.

100.8.4 Bound-state QED and the fine-structure constant

Scattering is not the only place QED is tested to many figures. A bound state made only of leptons has no hadronic structure at all, so its spectrum is a pure QED prediction, and the agreement with experiment is limited only by the input value of α and by the order to which the series has been evaluated.

- *Positronium*, the e^+e^- bound state discovered by Deutsch [980]. Its reduced mass is $m_e/2$, so the Bohr formula of Chapter 81 gives binding energies half those of hydrogen, $-\alpha^2 m_e c^2/4n^2$, that is $-6.80\,\mathrm{eV}$ in the ground state. It is unstable: the spin-singlet state annihilates into two photons at the rate $\Gamma_{2\gamma} = \alpha^5 m_e c^2/2\hbar$, a lifetime of 125 ps, while the spin-triplet state cannot decay into two photons — the two-photon system cannot carry $J = 1$ [917, 975] — and goes to three, giving a lifetime longer by roughly $1/\alpha$ times a phase-space factor, of order 140 ns. Both rates and the ground-state hyperfine splitting are calculable to high order and provide a QED test free of any nuclear input.

- *Muonium*, μ^+e^-, likewise purely leptonic. Its ground-state hyperfine interval is measured to $(4\,463\,302\,765 \pm 53)\,\mathrm{Hz}$ [1978], twelve significant figures, and its comparison with theory determines the muon-to-electron mass ratio.

- *Hydrogen*, whose $1S$–$2S$ two-photon transition is known to $(2\,466\,061\,413\,187\,035 \pm 10)\,\mathrm{Hz}$ [2179] and, after transfer over a 920 km fibre link to a caesium fountain, to $(2\,466\,061\,413\,187\,018 \pm 11)\,\mathrm{Hz}$ [2221] — fifteen significant figures, and the most accurately measured transition frequency in physics. Here QED is not the limiting uncertainty: the proton is. The measured frequency determines the Rydberg constant only in combination with the proton charge radius, which enters through a term of the same $\delta^3(\boldsymbol{r})$ form as Equation (100.54), and the discrepancy between the radius extracted from muonic hydrogen [2164, 2210] and the then-accepted electronic value was the proton-radius puzzle. It has since been resolved in favour of the smaller, muonic value by new electronic measurements [2281, 2311, 2331, 2358], and the recommended radius is now $(0.840\,75 \pm 0.000\,64)\,\mathrm{fm}$ [2431]; the experimental account is Chapter 90.

Derivation pending #144

Bound-state QED: the annihilation rates of para- and ortho-positronium quoted above, the ground-state hyperfine splitting of positronium and of muonium, and the systematic expansion of a two-body bound-state energy in powers of α and $\ln\alpha$ that organizes them. The selection rule forbidding a two-photon decay of a $J = 1$ state is cited rather than derived here and is proved in the discrete-symmetry chapter. This box belongs in Appendix A.

Proposition 100.71 (Several roads to α, and why that matters)**.** *The fine-structure constant is not predicted by anything in this book. It is measured, and the value is trustworthy because at least three physically unrelated routes agree:*

1. *from the electron anomaly, by inverting the series Equation (100.90) against the measured a_e of [2419] — a route that assumes QED;*
2. *from photon recoil, $\alpha^2 = (2R_\infty/c)\,(m_{\mathrm{at}}/m_e)\,(h/m_{\mathrm{at}})$, measured by atom interferometry in caesium [2316] and rubidium [2363] — a route that assumes only atomic kinematics and the Rydberg constant;*
3. *from the muonium hyperfine interval [1978] and from bound-state spectroscopy generally.*

The recommended value $\alpha^{-1} = 137.035999177(21)$ [2431] carries an uncertainty inflated to cover the disagreement between the two recoil determinations.

Rests on Phenomenon 100.67 and Equation (100.90).

Proof. *Derives Proposition 100.71.*

Route (i) is Equation (100.90) solved for α; route (ii) is kinematic, since a photon of known frequency transfers a known momentum to an atom of known mass and the resulting recoil frequency shift gives h/m_{at}, from which α follows through the displayed relation with R_∞ and the mass ratio measured independently; route (iii) uses the theoretical expression for the hyperfine interval, which is proportional to α^2 times the magnetic moment ratio. The routes share no systematic error, and two of them do not presuppose the QED series at all. That is what converts α from a parameter fitted to one measurement into a measured constant of Nature, and it is what makes the anomaly a *test* rather than a definition — the distinction insisted on in Remark 100.56. □

Remark 100.72 (A consequence of the 2019 SI for this chapter). Since 2019 the elementary charge e and the Planck constant h have exact defined values, and the speed of light has been exact since 1983. Equation (100.3) then reads $\alpha = \mu_0 c e^2/4\pi\hbar$ with everything on the right exact *except* the magnetic constant μ_0, which is therefore no longer $4\pi \times 10^{-7}\,\mathrm{N/A^2}$ by definition but a measured quantity, $\mu_0 = (1.256\,637\,061\,27 \pm 0.000\,000\,000\,20) \times 10^{-6}\,\mathrm{N/A^2}$ [2431], carrying exactly the fractional uncertainty of α. The experimental content of "α is measured" has been moved, by a change of unit conventions, into a constant that used to be exact. Nothing physical changed; but a reader who takes $\mu_0 = 4\pi \times 10^{-7}$ as exact and then quotes α to ten figures is quoting two mutually inconsistent statements, and the conventions of Chapter 2 are what keep the two apart.

Chapter 101

Weak Interactions

The weak interaction was forced on physics by an anomaly nobody wanted: the electrons emitted in nuclear beta decay come out with a continuous range of energies, although the initial and final nuclei have sharp masses. Everything in this chapter follows from the two responses to that single observation — a new, undetected neutral particle, and a new, extremely short-ranged interaction that does not conserve the quantum numbers the other forces conserve. The chapter runs from Chadwick's spectrum of 1914 to the discovery of the W and Z bosons in 1983, and it is deliberately told in that order, because the weak interaction is the one force whose theoretical structure was extracted almost entirely from experiment: parity violation, the $V - A$ current, universality and the charm quark were each read off measurements before any principle demanded them.

It sits after Chapter 100 because it is where the QED template first fails. Fermi's four-fermion theory is non-renormalizable and violates unitarity at energies of a few hundred GeV; the resolution — a massive intermediate vector boson, and hence a spontaneously broken gauge symmetry — is the subject of Chapter 104, which this chapter exists to motivate. Its immediate neighbours are the parity experiment of Chapter 110, the neutrino-helicity measurement of Chapter 98, the quark-mixing and neutrino-mass material of Chapter 103, and the nuclear phenomenology of Chapter 107. Standard treatments are [1739, 1933]; current values throughout are from [2427].

Conventions are those of Section 100.1.1 without exception: the metric is $\eta_{\mu\nu} = \mathrm{diag}(+1, -1, -1, -1)$, four-momenta are $p^\mu = (E/c, \boldsymbol{p})$ of SI dimension $\mathrm{kg\,m/s}$, the mass shell is $p^2 = m^2c^2$ as in Equation (100.1), so that a propagator denominator reads $p^2 - m^2c^2$ and never $p^2 - m^2$, and the Dirac field carries $1/\mathrm{m}^{3/2}$ by Definition 100.2. Every $\hbar$ and every c is written out. The one place where this chapter's arithmetic differs in appearance from its sources is the Fermi constant, which the literature quotes divided by $(\hbar c)^3$; both forms are given in Definition 101.8 and the difference between them is stated there rather than left to be guessed. Readers coming from the standard references, which drop both constants from every formula, should use the translation rule of Remark 100.4, which is the only place in this part where that convention is discussed.

101.1 The beta-decay crisis

101.1.1 The continuous spectrum

Alpha and gamma emission produce line spectra: a definite parent nucleus and a definite daughter nucleus differ by a definite energy, and the emitted quantum carries it. Beta emission was expected to do the same, and for two decades the experimental record was read as though it did — the sharp *internal-conversion* lines that accompany the beta continuum were mistaken for the primary spectrum, and the continuous background was ascribed to

electrons that had lost energy on their way out of the source.

Chadwick settled the question in 1914 with a magnetic spectrometer read by an ionization chamber rather than by a photographic plate [414]. A plate saturates, and a sharp line on a saturated plate looks like all there is; a counter measures intensity, and Chadwick's counter showed a broad continuous distribution on which the conversion lines sat as small peaks. The primary beta spectrum of radium B and C was continuous.

The escape route left open was that the electron leaves with the full energy and loses part of it in the source or in the surrounding material, so that the missing energy reappears as secondary radiation. Ellis and Wooster closed it by calorimetry [574]. They enclosed a radium-E source in a thick-walled calorimeter that would absorb everything except a neutrino, and measured the heat per disintegration: about 0.35 MeV, against an endpoint near 1 MeV. Had the missing energy been carried by any form of radiation, the calorimeter would have found it. Meitner and Orthmann repeated the measurement with a different calorimeter and a different absolute calibration and obtained the same answer [662], which is what turned a surprising result into an accepted one.

The seriousness of the anomaly is measured by the responses it drew. Bohr was for several years willing to entertain the possibility that energy conservation holds only statistically in nuclear processes — that is, to give up the conservation law rather than postulate an unseen particle. Pauli took the opposite view, and this chapter is the record of his being right; but the point worth keeping is that at the time neither option was obviously the cheaper one.

Phenomenon 101.1 (The beta spectrum is continuous)**.** The electrons emitted in nuclear beta decay do not carry a single energy. They are distributed continuously from zero up to a sharp endpoint fixed by the mass difference of the initial and final nuclei, although both of those masses are sharp [414]. The missing energy is not recovered as secondary radiation: a calorimeter enclosing a radium-E source records a mean energy per disintegration of about 0.35 MeV against an endpoint near 1 MeV [574].

Derivation. *Derives Phenomenon 101.1.*

Suppose the decay had only two products, a daughter nucleus of mass M' and an electron of mass m, and let the parent of mass M be at rest. Momentum conservation makes the two momenta equal and opposite, $\boldsymbol{p}' = -\boldsymbol{p}$, and energy conservation then reads

$$Mc^2 = \sqrt{p^2c^2 + M'^2c^4} + \sqrt{p^2c^2 + m^2c^4}\,. \tag{101.1}$$

The right-hand side is a strictly increasing function of p, so it attains the value Mc^2 for exactly one p: a two-body decay of a state of definite mass yields a *line*, and the electron energy $E = \sqrt{p^2c^2 + m^2c^4}$ is fixed. The observed continuum is therefore, by itself, a proof that at least three bodies share the released energy. Since the daughter and the electron are both detected and account between them for less than the release, the third is not detected: it must be neutral, must interact far more weakly than the electron, and must be light enough to leave the endpoint where the two-body kinematics puts it. That is precisely the particle Pauli proposed in 1930 [665] and Cowan and Reines detected twenty-six years later [1085]. The argument also shows what would have followed had the spectrum been a line with a lower endpoint — energy conservation itself would have been at stake, which is why Bohr was prepared to give it up. □

Remark 101.2 (What the calorimeter measures, quantitatively). The calorimetric result is stronger than "some energy is missing". Write $\langle T_e \rangle$ for the mean electron kinetic energy and T_0 for its endpoint. A three-body decay with a massless third particle distributes the release between electron and neutrino according to the phase-space factor derived in Proposition 101.11; for an allowed transition with $T_0 \approx 1\,\mathrm{MeV}$ that factor gives $\langle T_e \rangle \approx 0.40\,T_0$, and the measured ratio for radium E is about 0.33 [574]. A calorimeter that recorded T_0

per disintegration would have proved the missing energy to be radiation; one that records a third of it proves the energy is carried away by something the calorimeter cannot stop. The residual discrepancy between 0.40 and 0.33 is itself instructive and was not understood until much later: the decay of radium E, that is ^{210}Bi, is first-forbidden and its spectrum is correspondingly softer than the allowed shape. Nothing in the conclusion depends on that refinement — the alternative on offer in 1927 was that the calorimeter should have found the whole release, and it found a third.

101.1.2 Pauli's neutrino

On 4 December 1930 Pauli wrote to a conference at Tübingen he could not attend, addressing the "radioactive ladies and gentlemen" and proposing what he called a "desperate remedy" [665]: a neutral particle of spin $\frac{1}{2}$, obeying the exclusion principle, of mass "of the same order of magnitude as the electron mass" and in any case not larger than about one per cent of the proton mass, emitted together with the electron in every beta decay. He called it the neutron; the name *neutrino* is Fermi's, coined after Chadwick's discovery of the real neutron in 1932 had taken the first name.

The proposal is remarkable for how much it buys at once. Three independent difficulties, which had no reason to be connected, are removed by one hypothesis.

1. *Energy.* The release is shared between two light particles, so the electron energy is continuous with an endpoint at the full release, which is Phenomenon 101.1.

2. *Angular momentum.* Beta decay changes the mass number A by zero, so it changes the number of nucleons not at all and the number of fermions in the nucleus not at all; emitting a single spin-$\frac{1}{2}$ electron would change the total angular momentum of the system by a half-integer, which no rotation can do. A second spin-$\frac{1}{2}$ particle restores integrality.

3. *Statistics.* The nitrogen-14 nucleus was known from the band spectrum of N_2 to be a boson of spin 1. On the proton–electron model of the nucleus, ^{14}N contains 14 protons and 7 electrons, that is 21 fermions, and must be a fermion. The contradiction is what Pauli's letter calls the "wrong statistics".

Proposition 101.3 (The endpoint measures the neutrino mass)**.** *Let a nucleus of mass M at rest decay to a daughter of mass M', an electron of mass m_e and a neutral particle of mass m_ν. Write $Q = (M - M' - m_e)c^2$ for the energy released as kinetic energy. Then the maximum kinetic energy of the electron is*

$$T_e^{\max} = \frac{(Q - m_\nu c^2)\,(2M'c^2 + Q + m_\nu c^2)}{2\,(M'c^2 + m_e c^2 + Q)} = \left(Q - m_\nu c^2\right)\left[1 - \frac{2m_e c^2 + Q - m_\nu c^2}{2M'c^2} + \cdots\right], \tag{101.2}$$

and in the limit $M'c^2 \gg Q$ in which the daughter recoil is negligible, $T_e^{\max} = Q - m_\nu c^2$.

Rests on Equation (101.1).

Proof. *Derives Proposition 101.3.*

The electron takes its greatest energy when the daughter and the neutrino move together, so that the pair may be treated as a single body of mass $M' + m_\nu$ recoiling against the electron. That is a two-body decay, and Equation (101.1) with $M' \to M' + m_\nu$ fixes the electron energy uniquely:

$$E_e^{\max} = \frac{M^2c^4 + m_e^2c^4 - (M' + m_\nu)^2\,c^4}{2Mc^2}\,.$$

Subtract $m_e c^2$ and use $Mc^2 - m_e c^2 = M'c^2 + Q$ to factor the numerator as $(Q - m_\nu c^2)\,(2M'c^2 + Q + m_\nu c^2)$; writing the denominator as $2(M'c^2 + m_e c^2 + Q)$ gives the exact form in Equation (101.2), and expanding in $1/M'c^2$ gives the second. The recoil correction is of relative order $(2m_e c^2 + Q)\,/2M'c^2$ — dominated by the electron mass, not by Q, which is the trap in this expansion — and for tritium ($Q = 18.6\,\mathrm{keV}$, $M'c^2 = 2808\,\mathrm{MeV}$) it is 1.85×10^{-4}, that is 3.4 eV of recoil energy taken by the daughter. That is four times the whole mass sensitivity of Equation (101.23), so it is computed and subtracted rather than neglected. The leading statement is the one that matters here: the endpoint sits a distance $m_\nu c^2$ below the release, so a measured endpoint bounds the neutrino mass without any assumption about the interaction that produced it. □

Pauli's own bound came from exactly this argument applied to the spectra then available, and gave $m_\nu \lesssim m_e$; the modern version of the same measurement, on tritium, gives $m_\nu < 0.8\,\mathrm{eV}/c^2$ [2399], an improvement of six orders of magnitude in the same observable.

Pauli's caution was well judged in one respect. He estimated that his particle, being neutral and weakly coupled, would have a penetrating power far greater than a photon's, and doubted it could ever be observed. That doubt is quantified in Section 101.1.3 and it was very nearly right.

101.1.3 Detection

A particle postulated to explain a missing quantity is a bookkeeping device until it is detected in its own right. The detection reaction is beta decay run backwards,

$$\bar{\nu}_e + p \longrightarrow n + e^+ , \tag{101.3}$$

which has two attractive features: free protons are available in bulk as the hydrogen of water, and the final state carries its own signature in time.

Proposition 101.4 (Threshold of the inverse beta reaction)**.** *Reaction Equation* (101.3) *on a proton at rest requires*

$$E_{\bar{\nu}} \geq \frac{(m_n + m_e)^2\, c^4 - m_p^2 c^4}{2m_p c^2} = 1.806\,\mathrm{MeV}\,. \tag{101.4}$$

Rests on Equation (101.3)*.*

Proof. *Derives Proposition 101.4.*

The invariant $s = (p_{\bar{\nu}} + p_p)^2$ is conserved, and the final state exists only if $\sqrt{s}\,c \geq (m_n + m_e)c^2$, the least energy the pair can have. With the proton at rest and the antineutrino massless, $sc^2 = m_p^2 c^4 + 2m_p c^2 E_{\bar{\nu}}$, and the inequality rearranges to Equation (101.4). Inserting $m_n c^2 = 939.565\,\mathrm{MeV}$, $m_p c^2 = 938.272\,\mathrm{MeV}$ and $m_e c^2 = 0.511\,\mathrm{MeV}$ [2427] gives 1.806 MeV. A fission reactor emits antineutrinos with a spectrum extending to about 8 MeV, so a useful fraction of them are above threshold; this is why the experiment was done at a reactor and not with a natural source. □

The signature is a delayed coincidence. The positron annihilates within nanoseconds against an electron of the target, giving two 511 keV photons back to back; the neutron wanders, thermalizes, and is captured microseconds later on a nucleus with a large capture cross-section — cadmium dissolved in the water — which de-excites by emitting several megaelectronvolts of gamma rays. Two flashes separated by a characteristic time, in the right energy ratio, in a detector shielded against everything else, is a signature no ordinary background reproduces. Cowan, Reines and their collaborators built it at the Savannah River reactor and reported the observation in 1956 [1085, 1097], twenty-six years after Pauli's letter.

Phenomenon 101.5 (The neutrino interacts, and almost never)**.** Free antineutrinos from a reactor induce reaction Equation (101.3) at a rate corresponding to a cross-section of order $10^{-47}\,\mathrm{m}^2$ for a few-MeV antineutrino [1085, 1097]. A block of lead containing that many protons per unit volume would have to be of order a light-year thick to stop half the beam.

Rests on Equation (101.3) *and Proposition 101.4.*

Derivation. *Derives Phenomenon 101.5.*

The cross-section may be predicted from the coupling that Section 101.2.1 extracts from beta decay itself, which is what makes the number a test rather than a fit. At energies far below the mass of the mediating boson the interaction is a contact one, the amplitude is a product of two currents at a point, and by the derivation of Proposition 101.54 — which is performed there in full and is common to every charged-current process of this kind — the cross-section for the nucleon reaction is

$$\sigma\left(\bar{\nu}_e p \to n e^+\right) = \frac{G_F^2\,|V_{ud}|^2\,(1+3\lambda^2)}{\pi\,(\hbar c)^4}\,p_e c\,E_e\,, \tag{101.5}$$

with $E_e = E_{\bar{\nu}} - (m_n - m_p)c^2$ the positron energy, $p_e c = \sqrt{E_e^2 - m_e^2 c^4}$ and $\lambda = g_A/g_V = 1.2754$ the ratio of the axial and vector nucleon couplings measured in neutron decay [2427]; the factor $1 + 3\lambda^2$ is the sum of one Fermi and three Gamow–Teller spin channels, counted in Section 101.2.2. The factor $|V_{ud}|^2 = \cos^2\theta_C = 0.949$ is there because this is a *hadronic* charged current and not a leptonic one: the reaction turns a d quark into a u, and by Equation (101.63) that vertex carries $\cos\theta_C$. Omitting it overstates the rate by five per cent. At $E_{\bar{\nu}} = 3\,\mathrm{MeV}$ this gives $E_e = 1.707\,\mathrm{MeV}$, $p_e c = 1.628\,\mathrm{MeV}$ and

$$\sigma = 2.6 \times 10^{-47}\,\mathrm{m}^2\,,$$

which is the observed order of magnitude. For the penetrating power, lead has $11.34 \times 10^3\,\mathrm{kg/m^3}$ and molar mass $0.2072\,\mathrm{kg/mol}$, hence 3.30×10^{28} nuclei and 2.70×10^{30} protons per cubic metre; the mean free path is

$$\ell = \frac{1}{n\sigma} = \frac{1}{(2.70 \times 10^{30}\,1/\mathrm{m}^3)\,(2.6 \times 10^{-47}\,\mathrm{m}^2)} = 1.4 \times 10^{16}\,\mathrm{m}\,,$$

which is 1.5 yr of travel at the speed of light. The idealization is explicit and is worth stating: a proton bound in a lead nucleus is not a free proton and the reaction on it is not Equation (101.3), so the figure counts protons, not targets. As an order of magnitude for the transparency of matter to neutrinos it is nevertheless the right one, and it is the reason Pauli thought his particle undetectable. □

Two later measurements complete the picture of what a neutrino is. The first is that there is more than one kind. Firing a beam of neutrinos produced in pion decay, $\pi^+ \to \mu^+\nu$, at a spark-chamber detector, Danby and collaborators at the Brookhaven AGS found the reaction products to be muons and never electrons [1223]: the neutrino accompanying a muon makes a muon, not an electron. Since nothing in the kinematics forbids the electron channel — it is in fact the more open of the two — the null result establishes a conserved quantum number distinguishing the two families, which is lepton flavour. The third family's neutrino was observed directly at Fermilab forty years later by the DONUT collaboration, which identified the short kinked track of a tau produced in a nuclear emulsion [2016]. The second completion is the counting: Section 101.6.4 shows that the number of light neutrino species that couple to the Z is three, and that this is measured, not assumed.

101.2 Fermi's theory

101.2.1 The four-fermion interaction

Fermi's 1934 paper [738] is the origin of every equation in this chapter. Its construction is an analogy, and the analogy is with the electromagnetic interaction of Section 100.1. In QED the interaction Lagrangian density is a current contracted with a field, $\mathcal{L}_{\text{int}} = -j^{\mu}A_{\mu}$, and an elastic scattering amplitude is a product of two currents joined by the photon propagator: the interaction is *current–current*, with the two currents at different points and a field carrying the influence between them. Fermi kept the current–current form, kept the vector character of the currents, and threw away the field: his two currents sit at the same spacetime point.

The reasons were empirical. Beta decay had a range too short to measure, and no quantum of the required kind was known; the electron and the neutrino are created in the decay, exactly as a photon is created in an atomic transition, which is what suggested a field theory rather than a rearrangement of pre-existing constituents; and the vector form was the one that reproduced the observed spectrum shape. Nature rejected the paper as "too remote from physical reality". It appeared in *Zeitschrift für Physik* and contains, in the space of a few pages, the coupling constant, the spectrum shape, and the lifetime systematics.

Definition 101.6 (Fermi's interaction)**.** Let $\psi_p, \psi_n, \psi_e, \psi_\nu$ be the Dirac fields of the proton, neutron, electron and neutrino, each of SI dimension $1/\mathrm{m}^{3/2}$ by Definition 100.2. Fermi's interaction is the Lorentz-invariant contact term

$$\mathcal{L}_{\mathrm{F}} = -G\left(\bar{\psi}_p \gamma^{\mu} \psi_n\right)\left(\bar{\psi}_e \gamma_{\mu} \psi_\nu\right) + \text{h.c.}\,, \tag{101.6}$$

in which each factor in parentheses is a *current*, the two are evaluated at the same point x, and G is a constant.

Proposition 101.7 (The Fermi coupling has the dimension of an energy times a volume)**.** *For Equation* (101.6) *to be an energy density,* $\mathrm{J/m^3}$*, the constant* G *must carry*

$$[G] = \mathrm{J\,m^3} = 1/\mathrm{J}^2 \times \left[(\hbar c)^3\right]\,, \tag{101.7}$$

that is, $G/(\hbar c)^3$ *is an inverse squared energy.*
Rests on Equation (101.6) *and Definition 100.2.*

Proof. *Derives Proposition 101.7.*
By Definition 100.2 the bilinear $\bar{\psi}\gamma^{\mu}\psi$ has dimension $1/\mathrm{m}^3$, so the product of two currents carries $1/\mathrm{m}^6$. An energy density is $\mathrm{J/m^3}$. Hence $[G] = \mathrm{J/m^3} \times \mathrm{m}^6 = \mathrm{J\,m^3}$. Since $\hbar c$ has dimension J m, dividing by $(\hbar c)^3$ leaves $1/\mathrm{J}^2$. □

This one line is the most consequential dimensional statement in the chapter. Every coupling in Chapter 100 is dimensionless once expressed as α; Fermi's is not, and cannot be made so. Section 101.7.2 shows that this alone condemns the theory to non-renormalizability, and Section 101.7.1 shows that it also fixes, in advance, the energy at which the theory must fail.

Definition 101.8 (The Fermi constant)**.** The Fermi constant is defined by the muon lifetime through Equation (101.53), and its measured value is

$$\frac{G_F}{(\hbar c)^3} = (1.166\,378\,7 \pm 0.000\,000\,6) \times 10^{-5}/\mathrm{GeV}^2\,, \qquad G_F = 1.435\,85 \times 10^{-62}\,\mathrm{J\,m^3}\,. \tag{101.8}$$

The first form is the one universally quoted, and it is not G_F but G_F divided by the cube of $\hbar c$; the second is G_F itself in SI base units. Confusing them costs a factor $(\hbar c)^3 = 3.1600 \times 10^{-77}\,\mathrm{J^3\,m^3}$, so the distinction is stated here once and relied on thereafter.

Derivation of the conversion. *Derives Definition 101.8.*

$\hbar c = 3.161\,527 \times 10^{-26}\,\mathrm{J\,m}$ from the defined values of $\hbar$ and c, so $(\hbar c)^3 = 3.160\,03 \times 10^{-77}\,\mathrm{J^3\,m^3}$. One gigaelectronvolt is $1.602\,176\,634 \times 10^{-10}\,\mathrm{J}$ exactly, since the elementary charge is fixed by definition since 2019, so $1/\mathrm{GeV}^2 = 3.895\,66 \times 10^{19}/\mathrm{J}^2$. Then

$$G_F = 1.166\,378\,7 \times 10^{-5} \times 3.895\,66 \times 10^{19}/\mathrm{J}^2 \times 3.160\,03 \times 10^{-77}\,\mathrm{J^3\,m^3} = 1.435\,85 \times 10^{-62}\,\mathrm{J\,m^3}\,.$$

Two derived numbers used repeatedly below follow at once: the energy at which the dimensionless combination $G_F E^2/(\hbar c)^3$ reaches unity is

$$E_F = \left(\frac{(\hbar c)^3}{G_F}\right)^{1/2} = 292.8\,\mathrm{GeV}\,, \tag{101.9}$$

and the combination $(\sqrt{2}G_F/(\hbar c)^3)^{-1/2} = 246.2\,\mathrm{GeV}$ is the electroweak scale v that Chapter 104 identifies with the vacuum expectation value of the scalar field. □

Remark 101.9 (Why the customary units are not abandoned here). Editorial rule 3 requires SI, and Equation (101.8) gives it. The reciprocal-squared-energy form is nevertheless retained alongside, for a reason that is physics and not deference: the quantity that controls the size of every weak amplitude is the dimensionless ratio $G_F E^2/(\hbar c)^3$, and quoting $G_F/(\hbar c)^3$ in $1/\mathrm{GeV}^2$ makes that ratio readable at a glance — at $E = 1\,\mathrm{GeV}$ it is 1.17×10^{-5}, at $E = 100\,\mathrm{GeV}$ it is 0.117, and at Equation (101.9) it is one. The bare SI number $1.435\,85 \times 10^{-62}\,\mathrm{J\,m^3}$ conveys none of that. Both are correct; each is quoted where it does work.

The rest of this subsection extracts from Equation (101.6) the two results Fermi's paper is remembered for: the shape of the electron spectrum and the systematics of the lifetime. Both follow from one tool, which is used again in Section 101.4.3 in its relativistic form and is therefore stated here once with its derivation.

Proposition 101.10 (The golden rule)**.** *Let a system be prepared at $t = 0$ in a discrete state $|i\rangle$ of energy E_i and let a time-independent perturbation V connect it to a continuum of final states $|f\rangle$ with density $\rho(E_f)$ states per unit energy. Then, for times long compared with $\hbar/|E_f - E_i|$ over the width of the continuum and short compared with the lifetime, the transition rate is constant and equal to*

$$\Gamma = \frac{2\pi}{\hbar} \sum_f |\langle f|\, V\, |i\rangle|^2\, \delta\left(E_f - E_i\right)\,, \tag{101.10}$$

the sum being an integral $\int \mathrm{d}E_f\, \rho(E_f)$ over the continuum.

Rests on Postulate 77.4.

Proof. *Derives Proposition 101.10.*

To first order in V the amplitude for finding the system in $|f\rangle$ at time t is

$$c_f(t) = -\frac{\mathrm{i}}{\hbar} \int_0^t \mathrm{d}t'\, \langle f|\, V\, |i\rangle\, \mathrm{e}^{\mathrm{i}(E_f - E_i)t'/\hbar} = \langle f|\, V\, |i\rangle\, \frac{1 - \mathrm{e}^{\mathrm{i}\omega_{fi}t}}{\hbar\omega_{fi}}\,,$$

with $\hbar\omega_{fi} = E_f - E_i$. Hence

$$|c_f(t)|^2 = |\langle f|\, V\, |i\rangle|^2\, \frac{4\sin^2\left(\omega_{fi}t/2\right)}{\hbar^2\omega_{fi}^2}\,.$$

The kernel $4\sin^2(\omega t/2)/(\hbar\omega)^2$ has total integral $\int \mathrm{d}\omega\, 4\sin^2(\omega t/2)/(\hbar\omega)^2 = 2\pi t/\hbar^2$ and a width in ω of order $1/t$, so for t large compared with the inverse spread of the continuum it acts as $(2\pi t/\hbar^2)\delta(\omega) = (2\pi t/\hbar)\delta(E_f - E_i)$. Summing over f and dividing by t gives Equation (101.10); the probability grows linearly in t, which is what makes the *rate* well defined and the decay law exponential once the depletion of $|i\rangle$ is fed back in. The result is Dirac's [572]; the name is Fermi's [954], and it is the tool with which he wrote [738]. □

Proposition 101.11 (The allowed beta spectrum)**.** *Let a nucleus decay by an* allowed *transition — one in which the electron and neutrino are emitted in relative s waves, so that the nuclear matrix element does not depend on their momenta. Then the electron energy spectrum is*

$$\frac{\mathrm{d}\Gamma}{\mathrm{d}E_e} \propto F(Z,E_e)\, p_e E_e \,(E_0 - E_e)\sqrt{(E_0-E_e)^2 - m_\nu^2 c^4}\,, \tag{101.11}$$

where E_0 is the total energy available to the lepton pair, $p_e = \sqrt{E_e^2/c^2 - m_e^2 c^2}$, and $F(Z,E_e)$ is the Coulomb correction for the motion of the emitted electron in the field of the daughter nucleus. For $m_\nu = 0$ this reduces to

$$\frac{\mathrm{d}\Gamma}{\mathrm{d}E_e} \propto F(Z,E_e)\, p_e E_e \,(E_0 - E_e)^2\ . \tag{101.12}$$

Rests on Proposition 101.10.

Proof. *Derives Proposition 101.11.*

Apply the golden rule of Proposition 101.10 with two light particles in the final state and the recoiling daughter treated as infinitely heavy, so that it absorbs momentum but no energy. With the matrix element independent of the lepton momenta, the entire energy dependence is the density of final states,

$$\mathrm{d}\rho \propto \mathrm{d}^3p_e\,\mathrm{d}^3p_\nu\,\delta\left(E_0 - E_e - E_\nu\right) = 16\pi^2 p_e^2\,\mathrm{d}p_e\,p_\nu^2\,\mathrm{d}p_\nu\,\delta\left(E_0 - E_e - E_\nu\right)\ ,$$

the angular integrations giving 4π each because an s-wave matrix element is isotropic. Change variables with $p^2\,\mathrm{d}p/\mathrm{d}E = pE/c^2$, valid for either particle, and use the delta function to set $E_\nu = E_0 - E_e$:

$$\frac{\mathrm{d}\rho}{\mathrm{d}E_e} \propto \frac{p_e E_e}{c^2}\cdot\frac{p_\nu E_\nu}{c^2} = \frac{p_e E_e}{c^4}\,(E_0-E_e)\sqrt{(E_0-E_e)^2 - m_\nu^2c^4}\,,$$

where the last step used $p_\nu c = \sqrt{E_\nu^2 - m_\nu^2 c^4}$. That is Equation (101.11). The Coulomb factor is a separate effect: the emitted electron is not a plane wave but a continuum Coulomb wave in the field of the daughter's charge Z, and the rate is proportional to the squared modulus of that wave at the nucleus. Its non-relativistic form is

$$F(Z,E_e) = \frac{2\pi\eta}{1-\mathrm{e}^{-2\pi\eta}}\,, \qquad \eta = \pm\frac{Z\alpha c}{v_e}\,, \tag{101.13}$$

with the upper sign for β^- emission and the lower for β^+, Z being the charge number of the daughter. The sign is the entire physical content of the factor, and the two limits fix it. For β^- the electron is attracted, so $\eta > 0$, and $F \to 2\pi\eta > 1$ as v_e falls: the Coulomb field draws the outgoing wave onto the nucleus and *enhances* the low-energy end of the spectrum. For β^+ the positron is repelled, so $\eta < 0$, and $F \to 2\pi\,|\eta|\,\mathrm{e}^{-2\pi|\eta|} \ll 1$, which is the Gamow barrier factor and *suppresses* the same end. Here α is the fine-structure constant of Equation (100.3) and v_e the electron speed; η is dimensionless, as it must be. □

Proposition 101.12 (Sargent's rule)**.** *For an allowed transition with $E_0 \gg m_e c^2$ and the Coulomb factor neglected, the total rate grows as the fifth power of the energy release,*

$$\Gamma \propto E_0^5\,. \tag{101.14}$$

Rests on Equation (101.12).

Proof. *Derives Proposition 101.12.*

Integrate Equation (101.12) with $F = 1$ and $p_e c \simeq E_e$:

$$\Gamma \propto \int_0^{E_0} E_e^2\,(E_0-E_e)^2\,\mathrm{d}E_e = E_0^5\left(\frac{1}{3}-\frac{1}{2}+\frac{1}{5}\right) = \frac{E_0^5}{30}\,.$$

The fifth power is pure phase space — two particles sharing a release, each contributing a $p^2\mathrm{d}p$ — and it is what Sargent had found empirically in 1933 by plotting the decay constants of the naturally occurring beta emitters against their endpoints on logarithmic axes and obtaining two straight lines of slope five [727]. The two lines, rather than one, are the first appearance in this chapter of a distinction that Section 101.2.2 identifies: allowed and first-forbidden transitions have different matrix elements and therefore different intercepts, but the same slope, because the slope is kinematics. □

Remark 101.13 (Fermi's theory predicted a shape, and was believed because of it). It is worth being explicit about the logic, because it recurs. Equation (101.12) contains no free function: once E_0 is read off the endpoint and Z off the daughter, the entire spectrum is fixed, with the overall normalization the only adjustable number. A theory that merely permitted a continuous spectrum would have explained nothing — the continuum was the datum. Fermi's theory predicted *which* continuum, and the prediction is a one-parameter fit to a curve of several hundred points. That is why the vector form survived until parity violation forced its modification twenty-three years later, and why the modification, when it came, had to preserve the spectrum shape — as Section 101.4.1 shows the $V-A$ form does.

101.2.2 Allowed transitions: Fermi and Gamow–Teller

Fermi's choice of a vector current in Equation (101.6) was one of five possibilities, and Nature uses two of them. The enumeration is fixed by Lorentz invariance and is worth doing, because the same list organizes every experiment in Section 101.4.1 that fixed the interaction empirically.

Proposition 101.14 (There are exactly five bilinear covariants)**.** *The sixteen matrices*

$$\Gamma_i \in \left\{ \mathbb{1},\ \gamma^\mu,\ \sigma^{\mu\nu},\ \gamma^\mu\gamma^5,\ \gamma^5 \right\}, \qquad \sigma^{\mu\nu} = \frac{\mathrm{i}}{2}\left[\gamma^\mu, \gamma^\nu\right], \tag{101.15}$$

form a basis of the complex 4×4 *matrices, and the bilinears* $\bar\psi\Gamma_i\chi$ *built from them transform under the Lorentz group as a scalar (S), a vector (V), an antisymmetric tensor (T), an axial vector (A) and a pseudoscalar (P) respectively. Counting components,* $1+4+6+4+1=16$.

Rests on Lemma 100.23.

Proof. *Derives Proposition 101.14.*

Linear independence: with $\{\gamma^\mu,\gamma^\nu\}=2\eta^{\mu\nu}\mathbb{1}$, every product of gamma matrices reduces to one of Equation (101.15) up to a sign, and the trace relations of Lemma 100.23 give $\operatorname{tr}(\Gamma_i\Gamma_j^{-1})=4\delta_{ij}$ for the sixteen listed matrices, from which $\sum_i c_i\Gamma_i=0$ forces every $c_i=0$. Sixteen independent matrices in a sixteen-dimensional space are a basis. The transformation laws follow from the defining property of the spinor representation, $S^{-1}(\Lambda)\gamma^\mu S(\Lambda)=\Lambda^\mu{}_\nu\gamma^\nu$, established in Chapter 93: $\bar\psi\chi$ is invariant, $\bar\psi\gamma^\mu\chi$ carries one vector index, and so on. That the last two are *pseudo*-quantities is the statement that γ^5 anticommutes with every γ^μ and therefore picks up a sign under the improper transformation P, which is proved in Proposition 101.23. □

The most general Lorentz-invariant contact interaction between two pairs of spin-$\frac12$ fields is then a sum of five terms, one for each covariant, with five complex coefficients — and, once parity is not assumed, ten, since each term may be accompanied by its parity-reflected partner. Determining those ten numbers from experiment is the programme that occupies Sections 101.3 and 101.4.

For nuclear beta decay only two of the five survive, and the reason is angular momentum. In an allowed transition the leptons carry away no orbital angular momentum, so the total change in nuclear spin equals the total spin of the emitted pair, which is 0 or 1.

Proposition 101.15 (Allowed selection rules)**.** *For an allowed transition,*

1. *the S and V couplings emit the lepton pair in a spin singlet, giving* $\Delta J = 0$ *with no change of nuclear parity — the* Fermi *selection rule;*
2. *the T and A couplings emit it in a spin triplet, giving* $\Delta J = 0, \pm 1$ *with no change of parity and with* $0 \to 0$ *forbidden — the* Gamow–Teller *rule [775];*
3. *the P coupling contributes nothing at leading order, its nuclear matrix element being of order* v_{nucleon}/c*.*

The parity of the nucleus cannot change in an allowed transition, because the leptons carry away no orbital angular momentum and the nuclear operators that survive, $\mathbb{1}$ *and* $\boldsymbol{\sigma}$*, are both even under reflection.*

Rests on Proposition 101.14.

Proof. *Derives Proposition 101.15.*

Take the non-relativistic limit of the nucleon current, in which the large components of the nucleon spinors dominate. For the scalar and the time component of the vector, $\bar{\psi}_p \psi_n$ and $\bar{\psi}_p \gamma^0 \psi_n$ both reduce to $\phi_p^\dagger \phi_n$ acting on the two-component nucleon spinors: the operator is the identity in spin space, so the nucleon spin is untouched and by conservation the pair carries spin zero, hence $\Delta J = 0$. For the axial spatial components and the space components of the tensor, $\bar{\psi}_p \gamma^i \gamma^5 \psi_n$ and $\bar{\psi}_p \sigma^{ij} \psi_n$ reduce to $\phi_p^\dagger \sigma^i \phi_n$: the operator is a Pauli matrix, which flips the nucleon spin, so the pair carries spin one and $\Delta J \in \{0, \pm 1\}$. A spin-one pair cannot be emitted in a $0 \to 0$ transition, since there is nothing to absorb its unit of angular momentum. The remaining components — the space part of the vector, the time part of the axial vector and the pseudoscalar — carry an explicit factor of the nucleon velocity in the non-relativistic reduction and are suppressed by $v_{\text{nucleon}}/c \sim 10^{-1}$; they are the leading terms of the *forbidden* transitions.

Parity: in the allowed approximation the lepton wave functions are constants evaluated at the nucleus, so the whole dependence on the nuclear coordinates sits in the two operators just obtained, $\mathbb{1}$ and σ^i, and both are even under spatial reflection — the identity trivially, the Pauli matrix because a spin is an axial vector. A parity-even operator has vanishing matrix element between nuclear states of opposite parity, so the nuclear parity cannot change. The created pair does carry intrinsic parity -1, a fermion and an antifermion having opposite intrinsic parities, but that factor is the same in every beta decay and multiplies every amplitude alike; what distinguishes one transition from another is the orbital factor $(-1)^\ell$, which is $+1$ here and is exactly what changes sign in a first-forbidden transition. □

The decisive datum is that both rules are needed. The decay

$$^6\text{He}\left(0^+\right) \longrightarrow {}^6\text{Li}\left(1^+\right) + e^- + \bar{\nu}_e$$

has $\Delta J = 1$ and is fast, which the Fermi rule forbids outright; the decay $^{14}\text{O}(0^+) \to {}^{14}\text{N}^*(0^+)$ is fast and has $\Delta J = 0$ between two 0^+ states, which the Gamow–Teller rule forbids outright. Gamow and Teller proposed the second coupling in 1936 for exactly this reason [775]: the observed decays do not fit one interaction, and the pair of them is the minimum the data admit.

Definition 101.16 (Comparative half-life)**.** The measured half-life $t_{1/2}$ of a beta transition depends on the energy release through the fifth power of Equation (101.14) and on the nuclear structure through the matrix elements. Dividing out the first isolates the second. Define the dimensionless phase-space integral

$$f\left(Z, E_0\right) = \frac{1}{\left(m_e c^2\right)^5} \int_{m_e c^2}^{E_0} F\left(Z, E\right) \, pc \, E \, \left(E_0 - E\right)^2 \, \mathrm{d}E \,, \tag{101.16}$$

and call $ft_{1/2}$, of dimension s, the *comparative half-life*. A small ft means a fast transition for its energy release, that is a large nuclear matrix element.

Proposition 101.17 (The comparative half-life of an allowed transition)**.** *For an allowed transition,*

$$ft_{1/2} = \frac{K}{G_F^2 \left|V_{ud}\right|^2 \left(\left|M_{\mathrm{F}}\right|^2 + g_A^2 \left|M_{\mathrm{GT}}\right|^2\right)}, \qquad K = \frac{2\pi^3 \hbar \ln 2 \, (\hbar c)^6}{(m_e c^2)^5}, \tag{101.17}$$

where M_{F} and M_{GT} are the Fermi and Gamow–Teller nuclear matrix elements, g_A is the axial coupling of the nucleon and V_{ud} the quark-mixing element of Section 101.5. Numerically

$$\frac{K}{(\hbar c)^6} = 8.120\,276 \times 10^{-7}\ \mathrm{s/GeV^4}, \qquad K = 1.230\,58 \times 10^{-120}\ \mathrm{J^2\,m^6\,s}. \tag{101.18}$$

Rests on Proposition 101.11 and Equations (101.10) and (101.16).

Proof. *Derives Proposition 101.17.*

Collect the factors of Proposition 101.11. The rate is the golden-rule expression Equation (101.10) with the squared coupling $G_F^2 \left|V_{ud}\right|^2$ in front, the nuclear matrix elements squared and summed over the two spin channels with their multiplicities — one for the singlet, three for the triplet, which is the origin of the factor 3 in Equation (101.5) — and the lepton phase space integrated, which is $(m_e c^2)^5 f(Z, E_0)$ by Equation (101.16). Half-life and rate are related by $t_{1/2} = \hbar \ln 2/(\hbar\Gamma) = \ln 2/\Gamma$. Assembling the numerical factors of the two-particle phase space, which are $(2\pi)^{-3}$ per particle from the density of states and $2\pi/\hbar$ from Equation (101.10), leaves $K = 2\pi^3 \hbar \ln 2\,(\hbar c)^6/(m_e c^2)^5$. Its value follows from $\hbar = 6.582\,120 \times 10^{-25}\ \mathrm{GeV\,s}$ and $m_e c^2 = 0.510\,998\,95\ \mathrm{MeV}$:

$$\frac{K}{(\hbar c)^6} = \frac{2\pi^3\,(\ln 2)\,(6.582\,120 \times 10^{-25}\ \mathrm{GeV\,s})}{(5.109\,989\,5 \times 10^{-4}\ \mathrm{GeV})^5} = 8.120\,276 \times 10^{-7}\ \mathrm{s/GeV^4},$$

and multiplying by $(\hbar c)^6 = 9.985\,76 \times 10^{-154}\ \mathrm{J^6\,m^6}$ after converting $1/\mathrm{GeV^4}$ to $1/\mathrm{J^4}$ gives the SI form in Equation (101.18). □

Phenomenon 101.18 (Superallowed decays measure one number)**.** The half-lives, energy releases and branching ratios of the $0^+ \to 0^+$ beta transitions between members of an isospin triplet — ^{10}C, ^{14}O, ^{26m}Al, ^{34}Cl and about a dozen others, spanning a factor of 10^4 in f — yield, after small and calculable corrections for isospin breaking and radiative effects, a single common value

$$\mathcal{F}t = 3072\,\mathrm{s} \tag{101.19}$$

[2427]. Nuclei of different mass, different structure and different lifetime give the same number to a part in 10^4.

Rests on Proposition 101.15 and Equation (101.17).

Derivation. *Derives Phenomenon 101.18.*

A $0^+ \to 0^+$ transition is pure Fermi by Proposition 101.15: the Gamow–Teller operator cannot connect two spinless states. The Fermi operator is the isospin raising or lowering operator, whose matrix element between members of an isospin multiplet is fixed by the algebra and not by the nuclear wave function at all: for a triplet, $\left|M_{\mathrm{F}}\right|^2 = 2$. Hence Equation (101.17) loses every trace of nuclear structure and reduces to

$$\mathcal{F}t = \frac{K}{2G_F^2 \left|V_{ud}\right|^2 (1 + \Delta_R^V)}, \tag{101.20}$$

where Δ_R^V is the transition-independent radiative correction and $\mathcal{F}t$ denotes ft after the small nucleus-dependent corrections have been removed. With $G_F/(\hbar c)^3 = 1.166\,378\,7 \times 10^{-5}/\mathrm{GeV}^2$, $|V_{ud}| = 0.97435$ and $\Delta_R^V = 0.0240$ [2427],

$$\mathcal{F}t = \frac{8.120\,276 \times 10^{-7}\,\mathrm{s/GeV}^4}{2\,(1.166\,378\,7 \times 10^{-5})^2\,(0.97435)^2\,(1.0240)\;1/\mathrm{GeV}^4} = 3070\,\mathrm{s}\,,$$

against the measured 3072 s. The agreement to a part in 10^3 is what turns Equation (101.20) around: read backwards it is the most precise determination of $|V_{ud}|$ available, and hence the first-row test of the unitarity of the mixing matrix in Section 101.5. The constancy of $\mathcal{F}t$ across the multiplet is separately the evidence for the conserved vector current of Section 101.4.2: if the strong interaction renormalized the vector coupling, it would do so differently in a carbon nucleus and an aluminium one, and the fourteen numbers would not agree. □

Away from the superallowed set the same formula is a measurement of nuclear structure rather than of the coupling. Allowed transitions have $\log_{10}(ft/\mathrm{s})$ between about 3 and 6; each unit of forbiddenness — one unit of orbital angular momentum carried by the leptons, costing a factor $(p_e R/\hbar)^2$ in rate, with R the nuclear radius — adds between three and four to the logarithm, which is why ^{40}K lives for 1.25×10^9 yr and ^{6}He for 0.8 s. The systematics belong to Chapter 107; what belongs here is that the ladder exists and is explained, and that the neutron itself sits on the allowed rung with $ft \approx 1.05 \times 10^3$ s, a Fermi and a Gamow–Teller amplitude combining as $|M_\mathrm{F}|^2 + g_A^2\,|M_\mathrm{GT}|^2 = 1 + 3\lambda^2$ with $\lambda = 1.2754$.

101.2.3 The Kurie plot and the neutrino mass

Equation (101.11) is a curve, and curves are hard to compare by eye. Kurie, Richardson and Paxton observed that dividing out the known factors linearizes it [779], turning the comparison of theory with data into the inspection of a straight line — and, more importantly, turning the neutrino mass into a departure from straightness near one end.

Definition 101.19 (Kurie function)**.** For a measured spectrum $\mathrm{d}\Gamma/\mathrm{d}E_e$, the Kurie function is

$$K\,(E_e) = \left[\frac{\mathrm{d}\Gamma/\mathrm{d}E_e}{F\,(Z, E_e)\,p_e E_e}\right]^{1/2} . \tag{101.21}$$

Proposition 101.20 (The Kurie plot is a straight line if and only if the neutrino is massless)**.** *With Equation* (101.11)*,*

$$K\,(E_e) \propto \left[(E_0 - E_e)\,\sqrt{(E_0 - E_e)^2 - m_\nu^2 c^4}\right]^{1/2} , \tag{101.22}$$

which is $E_0 - E_e$ exactly — a straight line meeting the axis at $E_e = E_0$ — if and only if $m_\nu = 0$. For $m_\nu \neq 0$ the line bends downwards and meets the axis vertically at $E_e = E_0 - m_\nu c^2$.

Rests on Equations (101.11) *and* (101.21)*.*

Proof. *Derives Proposition 101.20.*

Substitute Equation (101.11) into Equation (101.21); the factors F, p_e and E_e cancel identically, leaving Equation (101.22). Writing $\epsilon = E_0 - E_e$ for the energy left to the neutrino, the bracket is $\epsilon\sqrt{\epsilon^2 - m_\nu^2 c^4}$, which vanishes at $\epsilon = m_\nu c^2$ rather than at $\epsilon = 0$, and whose derivative there is infinite. Hence the endpoint moves down by $m_\nu c^2$ — which is Proposition 101.3 recovered from the spectrum instead of from the kinematics — and the approach to it becomes vertical. Both features are confined to the last $m_\nu c^2$ of the spectrum, and the fraction of decays falling there is, from Equation (101.12),

$$\frac{\int_{E_0-\Delta}^{E_0} (E_0 - E)^2\,\mathrm{d}E}{\int^{E_0} (E_0 - E)^2\,\mathrm{d}E} \approx \left(\frac{\Delta}{E_0}\right)^3 ,$$

so a neutrino mass of $1\,\mathrm{eV}/c^2$ affects only 2×10^{-13} of the decays of a nucleus with $E_0 = 18.6\,\mathrm{keV}$. That cube is the whole difficulty of the measurement and the whole reason for choosing the smallest available E_0. □

Tritium, $^3\mathrm{H} \to {}^3\mathrm{He} + e^- + \bar{\nu}_e$, is the laboratory of choice on four counts: its release $E_0 = 18.6\,\mathrm{keV}$ is among the smallest known, which maximizes the fraction $(\Delta/E_0)^3$; the transition is superallowed, so the rate is high and the matrix element is structureless; the daughter is a two-electron atom whose final-state excitation spectrum can be computed exactly, which matters because molecular final states smear the endpoint by several electronvolts; and the half-life of 12.3 yr is short enough for a usable source strength and long enough for the source to be stable.

Phenomenon 101.21 (The neutrino mass is smaller than any other fermion mass by six orders of magnitude)**.** The endpoint of the tritium beta spectrum bounds the electron antineutrino mass at

$$m_\nu < 0.8\,\mathrm{eV}/c^2 \qquad (90\,\%\ \text{confidence}) \tag{101.23}$$

[2399]. The next lightest fermion, the electron, has $m_e c^2 = 0.511\,\mathrm{MeV}$: the ratio is below 1.6×10^{-6}.

Rests on Proposition 101.20 and Equation (101.22)*.*

Derivation. *Derives Phenomenon 101.21.*

The bound is Proposition 101.20 applied to a measured spectrum. KATRIN scans the integral spectrum with a magnetic-adiabatic collimation spectrometer of 1 eV resolution, fits the counting rate above a variable retarding potential to the predicted shape convolved with the response function and the calculable molecular final-state distribution, and treats m_ν^2 as the fit parameter — m_ν^2, and not m_ν, because Equation (101.22) depends on the square and because a fit constrained to positive values is biased. The result is consistent with zero, and the interval is converted to the bound Equation (101.23) [2399]. What the experiment measures is strictly an incoherent average, $m_\beta^2 = \sum_i |U_{ei}|^2\, m_i^2$, over the mass eigenstates that make up the electron antineutrino, because the individual masses are far too close together to be resolved as separate kinks; the mixing matrix U is that of Chapter 103. □

The bound is an upper limit, and an upper limit is compatible with zero. That the masses are *not* zero is known from a different kind of experiment entirely — neutrino oscillation, which measures differences of squared masses through an interference phase and is insensitive to the absolute scale. Oscillation gives $|\Delta m_{32}^2| \approx 2.5 \times 10^{-3}\,\mathrm{eV}^2/c^4$ and hence at least one mass above $0.05\,\mathrm{eV}/c^2$; the apparatus and the analysis are Chapter 114 and the theory is Chapter 103. The two measurements are complementary and neither replaces the other: the endpoint bounds the scale from above, oscillation bounds it from below, and the gap between $0.05\,\mathrm{eV}/c^2$ and $0.8\,\mathrm{eV}/c^2$ is where the answer lies.

Remark 101.22 (What a nonzero neutrino mass costs this chapter)*.* Nothing in Section 101.1 or Section 101.2 assumed the neutrino massless. Two later results do, and they are flagged where they occur: the exact identification of chirality with helicity in Proposition 101.32, which acquires a correction of relative size $m_\nu^2 c^4/4E^2$, and the two-component description of Section 101.3.3, which is exact only for a massless field. Both corrections are numerically negligible at any energy at which a neutrino has ever been detected — at $E = 1\,\mathrm{MeV}$ and $m_\nu c^2 = 0.8\,\mathrm{eV}$ the correction is 1.6×10^{-13} — and both are conceptually decisive, because a Standard Model with massless neutrinos has no room for the oscillations that are observed.

101.3 Parity violation

101.3.1 The tau–theta puzzle and the Lee–Yang question

Parity is the operation $P : \boldsymbol{x} \mapsto -\boldsymbol{x}$. Its significance is that it is a symmetry of the Coulomb interaction, of the nuclear force as then understood, and of every optical and spectroscopic selection rule that had ever been checked; atomic states carry a definite parity, and transitions between states of the wrong relative parity are absent to the precision of the measurement. The inference universally drawn was that parity is a symmetry of Nature. The inference is a generalization from two forces to all of them, and nobody had tested it in the third.

Proposition 101.23 (Behaviour of the bilinears under parity)**.** *Under the parity operation, implemented on a Dirac field by $\psi(t,\boldsymbol{x}) \mapsto \gamma^0\psi(t,-\boldsymbol{x})$, the covariants of Proposition 101.14 transform as*

$$\bar\psi\chi \longmapsto +\bar\psi\chi\,, \qquad \bar\psi\gamma^5\chi \longmapsto -\bar\psi\gamma^5\chi\,, \tag{101.24}$$

$$\bar\psi\gamma^\mu\chi \longmapsto +\bar\psi\gamma_\mu\chi\,, \qquad \bar\psi\gamma^\mu\gamma^5\chi \longmapsto -\bar\psi\gamma_\mu\gamma^5\chi\,, \tag{101.25}$$

with the arguments evaluated at the reflected point. Hence a current–current interaction is invariant under P if and only if each of its two currents has a definite and equal parity type: a product $V\times V$ or $A\times A$ is even, and a cross term $V\times A$ is odd.

Rests on Proposition 101.14.

Proof. *Derives Proposition 101.23.*

Use $\gamma^0\gamma^\mu\gamma^0 = \gamma_\mu$ — no sum, and the index is lowered, which is the statement that the time component is unchanged and the space components reverse — together with $(\gamma^0)^2 = \mathbb{1}$ and $\gamma^0\gamma^5\gamma^0 = -\gamma^5$, the last because γ^5 anticommutes with γ^0. Then $\bar\psi\chi = \psi^\dagger\gamma^0\chi \mapsto \psi^\dagger\gamma^0\gamma^0\gamma^0\chi = \bar\psi\chi$, and $\bar\psi\gamma^5\chi \mapsto \psi^\dagger\gamma^0\gamma^0\gamma^5\gamma^0\chi = -\bar\psi\gamma^5\chi$, and similarly for the two vectors. For the last statement, contract two currents: the product of two quantities that both pick up $+1$, or both -1, is invariant, while a product of one of each changes sign. A theory built only from $V\times V$ and $A\times A$ conserves parity; one containing $V\times A$ does not. □

Proposition 101.23 says that a parity-violating interaction is not exotic — it is simply one whose two currents are not both of the same type — and that testing parity means looking for the interference between the even and the odd pieces. Such an interference shows up as a *pseudoscalar* observable: a quantity built from measured vectors that changes sign under reflection, such as $\langle \boldsymbol{J}\rangle \cdot \hat{\boldsymbol{p}}$, the projection of a momentum on a spin. Before 1956 no experiment on a weak process had ever measured one.

What forced the question was a puzzle in the cosmic-ray and accelerator data. Two particles were seen, called at the time θ^+ and τ^+ — the second has nothing to do with the tau lepton, which was discovered twenty years later — decaying as

$$\theta^+ \longrightarrow \pi^+\pi^0\,, \qquad \tau^+ \longrightarrow \pi^+\pi^+\pi^-\,.$$

As the measurements improved, the two came to have the same mass and the same lifetime, to a precision at which a coincidence became absurd. But their decay products have opposite parity.

Proposition 101.24 (The parities of the two final states are opposite)**.** *A state of n pions in relative s waves has parity $(-1)^n$. Hence the $\pi\pi$ final state has parity $+1$ and the $\pi\pi\pi$ final state parity -1, and no single particle of definite parity can decay to both if parity is conserved.*

Rests on Phenomenon 105.5 and Proposition 105.15.

Proof. *Derives Proposition 101.24.*

The pion is a pseudoscalar, $J^P = 0^-$, established from the capture of a negative pion at rest on deuterium: the capture proceeds from an atomic s state to nn, the two neutrons must be in a state antisymmetric overall, and the only such state with $J = 1$ is ${}^3P_1$, of orbital parity -1; equating initial and final parities gives $P_\pi = -1$. The parity of a system of n such particles with no relative orbital angular momentum is the product of the intrinsic parities, $(-1)^n$. For $n = 2$ this is $+1$; for $n = 3$, -1. If the decaying particle has a definite parity P_X and parity is conserved, then $P_X = +1$ from the first mode and $P_X = -1$ from the second, a contradiction. The escape that the three-pion state carries orbital angular momentum, and so is not in relative s waves, is closed by the observed distribution of the three pions over their kinematically allowed region: the density is uniform up to the edges, which is the signature of $\ell = 0$, since any nonzero orbital angular momentum would suppress the configurations in which two pions are collinear. □

The way out was proposed by Lee and Yang in 1956 [1091], and the paper is remarkable less for what it proposed than for what it established first. They surveyed the entire experimental literature of weak decays and found that no measurement had ever been sensitive to a pseudoscalar observable — that the parity invariance of the weak interaction was not a tested fact but an untested assumption inherited from the strong and electromagnetic sectors. Having established that, they listed experiments that could decide, of which the most direct was to align nuclear spins at low temperature and look for a front–back asymmetry in the emitted electrons.

Remark 101.25 (The methodological point). This is worth stating flatly because it recurs throughout physics: a symmetry can hold to extraordinary precision in one interaction and fail completely in another, and the only way to know is to measure it where it matters. The strong and electromagnetic interactions conserve parity, in the second case to a relative precision better than 10^{-8}; the weak interaction violates it maximally. Nothing about the first two facts implies anything about the third, and the inference that was drawn for twenty-five years was a habit, not an argument. The same lesson is recorded for CP in Chapter 105 and for CPT, which unlike the others is a theorem and not an assumption.

101.3.2 Observation

Wu and her collaborators performed the experiment within months [1133]. A thin layer of ^{60}Co was polarized by adiabatic demagnetization of a cerium magnesium nitrate crystal at a few millikelvin, the degree of polarization being monitored by the anisotropy of the gamma rays from the daughter, and the beta electrons counted in an anthracene scintillator whose light was piped out of the cryostat. Reversing the polarizing field reverses $\langle \boldsymbol{J} \rangle$ and so must, if parity holds, leave the counting rate unchanged. It did not: the rate changed by tens of per cent, and the electrons came out preferentially into the hemisphere *opposite* to the nuclear spin. Chapter 110 is the experiment chapter this measurement belongs to. It states the result and proves the parity argument in full; its apparatus, procedure and systematic-check sections — among them the warming curve that shows the asymmetry dying with the polarization — carry their headings and their sources but are not yet written, so the sketch above is for the moment the fuller of the two accounts.

Garwin, Lederman and Weinrich obtained the same conclusion within days by a completely different route [1110]. Stopping positive muons from the decay $\pi^+ \to \mu^+ \nu_\mu$ in carbon and observing the decay positrons, they found the positrons emitted asymmetrically about the muon spin direction — and, since the muons had been polarized by the pion decay itself, the same data show parity violation twice over, once in $\pi \to \mu$ and once in $\mu \to e$. They also measured the muon magnetic moment in the same apparatus by precessing the spin, which is the ancestor of the experiment in Chapter 112.

Phenomenon 101.26 (Parity is violated in weak decays)**.** In the beta decay of polarized ^{60}Co nuclei the electrons are emitted preferentially into the hemisphere opposite to the nuclear spin. The angular distribution has the form

$$W(\theta) \propto 1 + A\,\frac{v}{c}\,\frac{\langle \boldsymbol{J}\rangle \cdot \hat{\boldsymbol{p}}}{|\langle \boldsymbol{J}\rangle|}\,, \tag{101.26}$$

with A negative and of order unity [1133], and the same asymmetry appears twice over in the chain $\pi \to \mu \to e$ [1110]. The apparatus is Chapter 110. No comparable asymmetry occurs in any strong or electromagnetic process.

Rests on Proposition 101.23.

Derivation. *Derives Phenomenon 101.26.*

Under the parity operation a polar vector such as the momentum $\boldsymbol{p}$ reverses while an axial vector such as the angular momentum $\boldsymbol{J}$ does not, so the scalar product $\langle \boldsymbol{J}\rangle \cdot \hat{\boldsymbol{p}}$ appearing in Equation (101.26) is a pseudoscalar and changes sign. The mirror-image experiment therefore predicts $W(\pi - \theta)$ wherever the original predicts $W(\theta)$, and the two disagree unless $A = 0$. A measured $A \neq 0$ is thus in itself a proof that the interaction responsible is not invariant under P: no dynamical input whatever is required, which is why one asymmetry measurement settled a question that decades of spectroscopy had not thought to ask [1091, 1133]. What the further observation $|A| \approx 1$ adds is that the violation is as large as the kinematics permits — the interaction does not merely prefer one handedness, it uses one and ignores the other — and that statement is delivered by Proposition 101.27 below, once the $V - A$ structure of Section 101.4 is in place. □

Proposition 101.27 (The sign and size of the cobalt asymmetry)**.** *For the pure Gamow–Teller transition* $^{60}\mathrm{Co}\,(5^+) \to {}^{60}\mathrm{Ni}\,(4^+)$*, an interaction that emits only left-chiral electrons and right-chiral antineutrinos gives* $A = -1$ *in Equation* (101.26)*: the electrons are emitted opposite to the nuclear spin, with an asymmetry equal in magnitude to the electron's speed in units of* c*.*

Rests on Equation (101.26) *and Propositions 101.32 and 101.34.*

Proof. *Derives Proposition 101.27.*

The transition lowers the nuclear angular momentum by one unit along the polarization axis $\hat{\boldsymbol{z}}$, so the lepton pair must carry away $J_z = +1$: both leptons have their spin projections $+\tfrac{1}{2}$ along $\hat{\boldsymbol{z}}$. Now apply the chirality rule. The antineutrino, being massless to the accuracy of Phenomenon 101.21, has positive helicity exactly by Proposition 101.32, so its spin is along its momentum: with $S_z = +\tfrac{1}{2}$ it must travel along $+\hat{\boldsymbol{z}}$. The electron is emitted in the left-chiral combination, whose helicity expectation value is $-v/c$ by Proposition 101.34; with its spin also along $+\hat{\boldsymbol{z}}$, a negative helicity means momentum along $-\hat{\boldsymbol{z}}$. The electron therefore prefers the hemisphere opposite the nuclear spin, and the preference is complete only in the limit $v \to c$, being diluted by the factor v/c at lower speeds. Writing the distribution as Equation (101.26) identifies $A = -1$ for this transition.

Comparing that with Wu's number requires one factor Equation (101.26) does not contain, and it is worth being careful, because the arithmetic does not otherwise close. The distribution is written for a *fully* polarized ensemble; a real source has a degree of nuclear polarization $f = |\langle \boldsymbol{J}\rangle|\,/J < 1$, and only that fraction of the nuclei contributes to the front–back difference, so the measured counting-rate asymmetry between the two senses of the polarizing field is

$$\varepsilon = \frac{W(0) - W(\pi)}{W(0) + W(\pi)} = A\,\frac{v}{c}\,f\,. \tag{101.27}$$

Wu's electrons had $v/c \approx 0.6$ and the source reached $f \approx 0.6$, giving $|\varepsilon| \approx 0.36$, which is the observed asymmetry of about 0.4. The product $|A|\ v/c$ alone would predict 0.6, half again too large: the polarization factor is not optional, and it is why the experiment had

to measure f independently, from the anisotropy of the daughter's gamma cascade. The same factor supplies the control — the asymmetry disappears as the crystal warms and f decays to zero, on the same time constant as the gamma anisotropy, which is what makes the result an asymmetry and not an instrumental drift [1133]. □

Two further consequences of the same data deserve separate statement.

The first is that charge conjugation is violated as well. If C were a symmetry, the mirror process for antimatter would be obtained by conjugating charges; the $\pi^+ \to \mu^+$ and $\pi^- \to \mu^-$ chains would then show asymmetries of the same sign. They show opposite ones [1110]. What survives is the combination CP: the mirror image of a left-handed neutrino is a right-handed neutrino, which does not exist, but the CP image is a right-handed antineutrino, which does. That CP is itself violated, by a small amount and in a different sector, is the subject of Chapter 105 and Chapter 111.

The second is quantitative and constrains the theory sharply. A parity-violating interaction has, in general, both a P-even and a P-odd part, and the size of the asymmetry measures their relative magnitude. An asymmetry of order unity — not of order 10^{-2}, and not of order 10^{-6} — says that the two parts are of equal size. An interaction whose P-even and P-odd parts are equal is one that couples to a single chirality and not at all to the other. That is the empirical origin of the $V - A$ structure, and it was read off the data in exactly this way.

101.3.3 Neutrino helicity

The most direct statement the parity data suggest is that the neutrino comes in one handedness only. Landau, Lee and Yang, and Salam each proposed independently in 1957 that the neutrino is described by a two-component field — the Weyl equation of Chapter 96 — rather than by a four-component Dirac field with one half of it inert [1118, 1121, 1130]. A two-component field has no mass term available to it and no partner of the opposite handedness, so the proposal makes handedness a defining property rather than a dynamical accident, and it predicts a definite sign which experiment can check.

Checking it means measuring the helicity of a particle that is never detected. Goldhaber, Grodzins and Sunyar found a way [1148]: in the orbital electron capture $^{152m}\mathrm{Eu} + e^- \to {}^{152}\mathrm{Sm}^* + \nu_e$ followed by $^{152}\mathrm{Sm}^* \to {}^{152}\mathrm{Sm} + \gamma$, the photons emitted opposite to the neutrino carry the neutrino's helicity, and their circular polarization can be measured by transmission through magnetized iron. Selecting those photons is done by resonant scattering, which only the ones emitted against the neutrino have enough energy to undergo. Chapter 98 is the experiment chapter for this measurement; it names the decay scheme, the magnetized-iron analyser and the resonant scatterer, and its apparatus and systematic-check sections are still owed. The angular-momentum argument is short enough to give here, and is.

Phenomenon 101.28 (Neutrinos are left-handed)**.** The neutrino emitted in orbital electron capture on $^{152}\mathrm{Eu}$ has negative helicity — its spin points opposite to its momentum — with an uncertainty that leaves no room for an appreciable admixture of the other state [1148]; the antineutrino emitted in beta decay has positive helicity. No right-handed neutrino, and no left-handed antineutrino, has ever been observed to take part in a weak interaction.

Rests on Definition 101.29.

Derivation. *Derives Phenomenon 101.28.*

What has to be established is that the photon the apparatus selects carries the neutrino's helicity, since the photon's circular polarization is the only thing measured. Both steps are angular-momentum bookkeeping along a single axis.

Take $\hat{\boldsymbol{z}}$ along the photon's momentum. The capture $^{152m}\mathrm{Eu}\,(0^-) + e^- \to {}^{152}\mathrm{Sm}^*(1^-) + \nu_e$ proceeds from a K-shell s orbital, so nothing brings orbital angular momentum in, and the

two products leave back to back, the neutrino along $-\hat{\boldsymbol{z}}$ and the recoiling $^{152}\mathrm{Sm}^*$ along $+\hat{\boldsymbol{z}}$. A two-body state collinear with the quantization axis has $L_z = 0$, so along that axis

$$S_z^e = S_z^\nu + J_z^{\mathrm{Sm}^*} . \tag{101.28}$$

Take $S_z^e = -\frac{1}{2}$. A negative-helicity neutrino moving along $-\hat{\boldsymbol{z}}$ has its spin along $+\hat{\boldsymbol{z}}$, that is $S_z^\nu = +\frac{1}{2}$, and Equation (101.28) then forces $J_z^{\mathrm{Sm}^*} = -1$. The excited state decays to a spinless ground state, $^{152}\mathrm{Sm}^*(1^-) \to {}^{152}\mathrm{Sm}\,(0^+) + \gamma$, so the photon carries away the whole of $J_z = -1$; and a photon travelling along $+\hat{\boldsymbol{z}}$ with $J_z = -1$ has helicity -1. Reversing the neutrino's helicity reverses every sign in the chain, so the photon's helicity is not merely correlated with the neutrino's but equal to it.

That fixes the photons emitted *opposite* to the neutrino, and the resonance condition selects exactly those and no others. A photon emitted by a nucleus at rest falls short of the absorption resonance by the recoil energies of emitter and absorber; only a nucleus recoiling towards the emission direction — that is, one whose neutrino went the other way — is moving fast enough for the Doppler shift to make the deficit up, so only those photons scatter resonantly on a $^{152}\mathrm{Sm}$ target. Their circular polarization, measured by transmission through magnetized iron, is therefore the neutrino's helicity, and it came out negative [1148]. The antineutrino statement follows from CP: the CP image of a negative-helicity neutrino is a positive-helicity antineutrino, and CP survives these decays even though C and P separately do not. □

To say what such a measurement establishes, and what it does not, the two notions it is easy to conflate must be separated. They are different in kind: one is a property of a state, the other a property of a field; one is conserved but frame-dependent, the other frame-independent but not conserved; and they coincide only in a limit.

Definition 101.29 (Helicity)**.** The helicity operator for a spin-$\frac{1}{2}$ particle of momentum $\boldsymbol{p}$ is

$$h = \frac{\boldsymbol{\Sigma}\cdot\boldsymbol{p}}{|\boldsymbol{p}|} , \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\sigma} & 0 \\ 0 & \boldsymbol{\sigma} \end{pmatrix} , \tag{101.29}$$

with eigenvalues ± 1. It is the projection of the spin on the direction of motion, and it is a genuine observable: both factors are measurable.

Definition 101.30 (Chirality)**.** The chirality operator is $\gamma^5 = \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3$, with $(\gamma^5)^2 = \mathbb{1}$ and $\{\gamma^5, \gamma^\mu\} = 0$. Its eigenvalues are ± 1 and the associated projectors are

$$P_{\mathrm{L}} = \frac{1}{2}\left(\mathbb{1} - \gamma^5\right) , \qquad P_{\mathrm{R}} = \frac{1}{2}\left(\mathbb{1} + \gamma^5\right) , \tag{101.30}$$

so that $\psi = \psi_{\mathrm{L}} + \psi_{\mathrm{R}}$ with $\psi_{\mathrm{L}} = P_{\mathrm{L}}\psi$. Chirality is a property of the field, defined without reference to any momentum.

Proposition 101.31 (Chirality against helicity)**.** *Chirality is Lorentz invariant but not conserved; helicity is conserved in free motion but not Lorentz invariant. Precisely:*

1. *γ^5 commutes with the generators $\sigma^{\mu\nu}$ of the spinor representation, so P_{L} and P_{R} commute with every Lorentz transformation: chirality is frame-independent.*
2. *A Dirac mass term couples the two chiralities, $\bar\psi\psi = \bar\psi_{\mathrm{L}}\psi_{\mathrm{R}} + \bar\psi_{\mathrm{R}}\psi_{\mathrm{L}}$, so chirality is conserved by the free equation of motion only if $m = 0$.*
3. *h commutes with the free Dirac Hamiltonian, so helicity is conserved in free motion.*
4. *For $m \neq 0$ a particle moves slower than light, so an observer may overtake it; in the overtaking frame $\boldsymbol{p}$ reverses and $\boldsymbol{\Sigma}$ does not, and the helicity changes sign. Helicity is frame-dependent.*

Rests on Definitions 101.29 and 101.30.

Proof. *Derives Proposition 101.31.*

(i) $\sigma^{\mu\nu}$ is a product of two distinct gamma matrices, and γ^5 anticommutes with each, hence commutes with the pair; therefore $[\gamma^5, \sigma^{\mu\nu}] = 0$ and γ^5 commutes with $\exp(-\frac{\mathrm{i}}{4}\omega_{\mu\nu}\sigma^{\mu\nu})$. (ii) $\bar\psi\psi = \bar\psi(P_{\mathrm{L}}+P_{\mathrm{R}})\psi$, and $\bar\psi P_{\mathrm{L}} = \overline{P_{\mathrm{R}}\psi}$ because $\gamma^0\gamma^5 = -\gamma^5\gamma^0$; the diagonal terms therefore vanish and only the cross terms survive. (iii) The free Hamiltonian is $H = c\boldsymbol{\alpha}\cdot\boldsymbol{p} + \beta mc^2$ with $\boldsymbol{\alpha} = \gamma^0\boldsymbol{\gamma}$; a direct computation using $[\Sigma^i, \alpha^j] = 2\mathrm{i}\varepsilon^{ijk}\alpha^k$ and $[\boldsymbol{\Sigma}, \beta] = 0$ gives $[\boldsymbol{\Sigma}\cdot\boldsymbol{p}, H] = 0$. (iv) A boost of speed $u > v$ along $\hat{\boldsymbol{p}}$ sends $\boldsymbol{p} \mapsto -|\boldsymbol{p}'|\,\hat{\boldsymbol{p}}$ while the spin projection on $\hat{\boldsymbol{p}}$ is unchanged, since a boost along an axis does not rotate spin about that axis; hence $h \mapsto -h$. For $m = 0$ no such boost exists and the obstruction disappears. □

The two therefore cannot be the same thing. The following proposition says exactly how close they are, and it is the quantitative content of every statement in this chapter of the form "only left-handed particles participate".

Proposition 101.32 (Chirality and helicity in a state of definite momentum)**.** *Let $u_\lambda(p)$ be a positive-energy Dirac spinor of momentum $\boldsymbol{p}$, energy E and helicity $\lambda = \pm 1$. Then*

$$\frac{|P_{\mathrm{L}}u_\lambda|^2}{|u_\lambda|^2} = \frac{1}{2}\,(1 - \lambda\beta)\,, \qquad \beta = \frac{|\boldsymbol{p}|\,c}{E} = \frac{v}{c}\,. \tag{101.31}$$

In particular, for $\beta \to 1$ the left-chiral projection of a positive-helicity state vanishes as

$$\frac{1}{2}\,(1-\beta) \simeq \frac{m^2c^4}{4E^2}\,, \tag{101.32}$$

so that chirality and helicity agree exactly in the massless limit and to relative accuracy $m^2c^4/4E^2$ otherwise.

Rests on Definition 101.29 and Equation (101.30).

Proof. *Derives Proposition 101.32.*

Work in the Dirac representation, in which $\gamma^5 = \begin{pmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{pmatrix}$ in 2×2 blocks. A positive-energy solution of definite helicity is

$$u_\lambda(p) = N \begin{pmatrix} \sqrt{E + mc^2}\,\chi_\lambda \\ \lambda\sqrt{E - mc^2}\,\chi_\lambda \end{pmatrix}\,, \qquad (\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}})\,\chi_\lambda = \lambda\chi_\lambda\,, \tag{101.33}$$

as follows from the Dirac equation with the lower two components given by $c\,\boldsymbol{\sigma}\cdot\boldsymbol{p}/(E + mc^2)$ times the upper ones and $c\,|\boldsymbol{p}| = \sqrt{(E+mc^2)(E-mc^2)}$. Its norm is $|u_\lambda|^2 = N^2\left[(E+mc^2) + (E-mc^2)\right] = 2N^2E$. Applying $P_{\mathrm{L}} = \frac{1}{2}\begin{pmatrix} \mathbb{1} & -\mathbb{1} \\ -\mathbb{1} & \mathbb{1} \end{pmatrix}$,

$$P_{\mathrm{L}}u_\lambda = \frac{N}{2}\left(\sqrt{E+mc^2} - \lambda\sqrt{E-mc^2}\right)\begin{pmatrix} \chi_\lambda \\ -\chi_\lambda \end{pmatrix}\,,$$

whose squared norm is

$$|P_{\mathrm{L}}u_\lambda|^2 = \frac{N^2}{2}\left(\sqrt{E+mc^2} - \lambda\sqrt{E-mc^2}\right)^2 = N^2\left(E - \lambda\sqrt{E^2 - m^2c^4}\right) = N^2E\,(1-\lambda\beta)\,,$$

using $\lambda^2 = 1$ and $\sqrt{E^2 - m^2c^4} = |\boldsymbol{p}|\,c = \beta E$. Dividing gives Equation (101.31). For the limit, expand $\beta = \sqrt{1 - m^2c^4/E^2} \simeq 1 - m^2c^4/2E^2$. □

Corollary 101.33 (What Goldhaber's experiment measures)**.** *For a neutrino of mass below* $0.8\,\mathrm{eV}/c^2$ *(Phenomenon 101.21) emitted with an energy near* $0.9\,\mathrm{MeV}$*, as in the europium capture, the wrong-helicity admixture in a purely left-chiral coupling is, by Equation* (101.32)*, at most*

$$\frac{m_\nu^2 c^4}{4E^2} < \frac{(0.8\,\mathrm{eV})^2}{4\,(0.9\,\mathrm{MeV})^2} = 2\times 10^{-13}\,.$$

A measurement of the helicity is therefore, to that accuracy, a measurement of the chirality of the coupling — which is the quantity the theory contains and the experiment cannot see directly.

Rests on Phenomenon 101.21 and Equation (101.32)*.*

Proposition 101.34 (The longitudinal polarization of a beta electron)**.** *An interaction that couples only to* ψ_{L} *emits electrons with mean helicity*

$$\langle h\rangle = -\frac{v}{c}\,, \tag{101.34}$$

and positrons with mean helicity $+v/c$*.*

Rests on Equation (101.31)*.*

Proof. *Derives Proposition 101.34.*

By Equation (101.31) the left-chiral projection contains helicity -1 with weight $\frac{1}{2}(1+\beta)$ and helicity $+1$ with weight $\frac{1}{2}(1-\beta)$, and these are the relative probabilities with which the interaction produces the two. Hence $\langle h\rangle = \frac{1}{2}(1-\beta) - \frac{1}{2}(1+\beta) = -\beta$. For the antiparticle the roles of the projectors are exchanged, because the negative-energy spinor v_λ carries the opposite relation between chirality and helicity, and the sign reverses. Equation (101.34) is a prediction with no free parameter: it says that a beta electron of $v/c = 0.5$ is 50 per cent longitudinally polarized and one of $v/c = 0.9$ is 90 per cent polarized, and it is measured directly, by passing the emitted electrons through a polarization-sensitive scatterer. Frauenfelder and collaborators did it for ^{60}Co within months of Wu's result and found the electrons longitudinally polarized, with the predicted sign and a magnitude consistent with $-v/c$ [1108]. It is also what supplies the factor v/c in Equation (101.26), which is therefore not an empirical fudge but the electron mass making itself felt. □

Finally, the same arithmetic explains a rate that would otherwise be inexplicable. The charged pion decays to a muon and not to an electron, although the electronic channel has far more phase space, because the pion is spinless and forces the charged lepton into the helicity that the left-chiral coupling suppresses by $\frac{1}{2}(1-\beta)$; the suppression is governed by the lepton mass and therefore hurts the electron far more than the muon. The derivation and the measured ratio 1.23×10^{-4} are Phenomenon 96.3. It is the sharpest confirmation of Proposition 101.32 available, because it turns a factor of $(m_e/m_\mu)^2$ into a directly counted branching ratio.

101.4 The $V-A$ interaction

101.4.1 Chirality and the left-handed current

Section 101.3 leaves the interaction determined up to ten coefficients, and the observations of maximal asymmetry and of a purely left-handed neutrino determine them. Feynman and Gell-Mann [1144] and, independently and in the same year, Sudarshan and Marshak [1162] proposed that the answer is

$$J^\mu = \bar{\psi}\,\gamma^\mu\left(\mathbb{1}-\gamma^5\right)\chi\,, \tag{101.35}$$

one combination out of the five, with the vector and axial pieces of equal magnitude and opposite sign. Every charged-current process in the Standard Model is built from currents of this form.

Proposition 101.35 (The $V-A$ current is the left-chiral current)**.**

$$\bar{\psi}\gamma^{\mu}\left(\mathbb{1}-\gamma^{5}\right)\chi = 2\,\bar{\psi}_{\mathrm{L}}\gamma^{\mu}\chi_{\mathrm{L}}\,, \tag{101.36}$$

so an interaction built from Equation (101.35) *involves the right-chiral components of the fields not at all.*

Rests on Equations (101.30) *and* (101.35)*.*

Proof. *Derives Proposition 101.35.*

From Equation (101.30), $P_{\mathrm{L}}^{2}=P_{\mathrm{L}}$ and $\gamma^{\mu}P_{\mathrm{L}}=P_{\mathrm{R}}\gamma^{\mu}$, the latter because γ^{5} anticommutes with γ^{μ}. Also $\overline{P_{\mathrm{L}}\psi}=\psi^{\dagger}P_{\mathrm{L}}\gamma^{0}=\bar{\psi}P_{\mathrm{R}}$, using $P_{\mathrm{L}}^{\dagger}=P_{\mathrm{L}}$ and $\gamma^{0}\gamma^{5}=-\gamma^{5}\gamma^{0}$. Hence

$$\bar{\psi}_{\mathrm{L}}\gamma^{\mu}\chi_{\mathrm{L}}=\bar{\psi}P_{\mathrm{R}}\gamma^{\mu}P_{\mathrm{L}}\chi=\bar{\psi}\gamma^{\mu}P_{\mathrm{L}}P_{\mathrm{L}}\chi=\bar{\psi}\gamma^{\mu}P_{\mathrm{L}}\chi=\frac{1}{2}\bar{\psi}\gamma^{\mu}\left(\mathbb{1}-\gamma^{5}\right)\chi\,.$$

The same computation shows that the *scalar* bilinear does the opposite: $\bar{\psi}_{\mathrm{L}}\chi_{\mathrm{L}}=\bar{\psi}P_{\mathrm{R}}P_{\mathrm{L}}\chi=0$, so a chirally projected theory admits vector and axial currents and forbids scalar, pseudoscalar and tensor couplings between fields of the same chirality. Equation (101.35) is therefore not one arbitrary choice among five: it is the only current a theory of purely left-chiral fields can have. □

Proposition 101.36 (Maximal parity violation is exactly $V-A$)**.** *Write a general vector-and-axial current as $\bar{\psi}\gamma^{\mu}(g_V-g_A\gamma^{5})\chi$. Then the P-violating interference between the two, measured by any pseudoscalar observable, is proportional to $g_V g_A$ and the P-conserving part to $g_V^{2}+g_A^{2}$. The ratio of the two is extremal, and equal to unity, if and only if $|g_V|=|g_A|$.*

Rests on Proposition 101.23 and Equation (101.35)*.*

Proof. *Derives Proposition 101.36.*

By Proposition 101.23 the VV and AA terms of the squared amplitude are P-even and the VA cross terms P-odd, and the coefficients are g_V^{2}, g_A^{2} and $2g_Vg_A$ respectively. The asymmetry is therefore $2g_Vg_A/(g_V^{2}+g_A^{2})$, which by the arithmetic–geometric mean inequality is at most 1 in modulus with equality if and only if $g_V=\pm g_A$. The measured asymmetries of Section 101.3.2 are of order unity, which forces the equality, and the observed *sign* — electrons opposite to the nuclear spin, neutrinos left-handed — selects $g_A=+g_V$ in the convention of Equation (101.35), that is $V-A$ and not $V+A$. □

Remark 101.37 (Where the coefficients actually came from)*.* The account above is the logical order, not the historical one. The coefficients were fixed by a decade of experiments on the angular correlation between the electron and the recoiling nucleus, the observable most sensitive to the Lorentz structure and the one that needs no polarization at all. For an allowed transition it has the form

$$\frac{\mathrm{d}\Gamma}{\mathrm{d}\Omega_{e\nu}}\propto 1+a\,\frac{v_e}{c}\cos\theta_{e\nu}\,, \tag{101.37}$$

with $\theta_{e\nu}$ the angle between the electron and antineutrino momenta, inferred in practice from the recoil of the daughter since the antineutrino is not seen. The coefficient a takes a different value for each of the four couplings, so one number decides among them:

coupling	transition type	a
scalar (S)	Fermi	-1
vector (V)	Fermi	$+1$
tensor (T)	Gamow–Teller	$+1/3$
axial (A)	Gamow–Teller	$-1/3$

The values follow from Proposition 101.32 and angular momentum alone; the derivation is Proposition 101.38.

The history is cautionary, and is recorded because the answer was wrong for five years. Feynman and Gell-Mann list the measurements their proposal contradicted — among them a recoil experiment on ^{6}He, a pure Gamow–Teller decay, whose reported correlation favoured tensor over axial — and say in print that they believe those experiments to be wrong [1144]. They were, and the ^{6}He value was later revised to the axial $-1/3$; but the papers reporting and revising it are not in this bibliography, so what is said here rests on Feynman and Gell-Mann's account of them and not on the sources. The lesson is not that theory ran ahead of experiment: it is that one measurement of a small correlation, with an uncontrolled systematic, decided a structural question for five years, and that the theorists who disbelieved it did so on the strength of *other* data — the maximal asymmetries.

Proposition 101.38 (The electron–neutrino correlation coefficients)**.** *For a $V-A$ interaction the coefficient a in Equation* (101.37) *is $+1$ for a Fermi transition and $-1/3$ for a Gamow–Teller transition; for the parity-conjugate choices S and T the signs reverse.*

Rests on Equation (101.37) *and Propositions 101.32 and 101.35.*

Proof. *Derives Proposition 101.38.*

Work in the limit $v_e \to c$, where by Proposition 101.32 the electron emitted by a left-chiral coupling has helicity -1 exactly and the antineutrino has helicity $+1$ exactly; the factor v_e/c in Equation (101.37) is what restores the general case, being the electron's helicity by Equation (101.34). Then the spin of the antineutrino points along its momentum and the spin of the electron points against its own.

A Fermi transition emits the pair in the spin singlet, whose two spins are antiparallel. Let the antineutrino spin be $+\hat{\boldsymbol{z}}$ and the electron spin $-\hat{\boldsymbol{z}}$. The antineutrino momentum is then along $+\hat{\boldsymbol{z}}$ and the electron momentum, being opposite its spin, is also along $+\hat{\boldsymbol{z}}$: the two particles go the same way, $\cos\theta_{e\nu} = +1$, and the distribution is $1 + \cos\theta_{e\nu}$, that is $a = +1$.

A Gamow–Teller transition emits the pair in the spin triplet, and the three substates must be averaged because an unpolarized nucleus populates them equally. For $m = +1$ both spins are $+\hat{\boldsymbol{z}}$; the antineutrino goes along $+\hat{\boldsymbol{z}}$ and the electron along $-\hat{\boldsymbol{z}}$, giving $\cos\theta_{e\nu} = -1$; the same for $m = -1$. The state $m = 0$ is $(|{\uparrow\downarrow}\rangle + |{\downarrow\uparrow}\rangle)/\sqrt{2}$, which is the singlet's partner with antiparallel spins and gives $\cos\theta_{e\nu} = +1$. Averaging, $a = (-1 - 1 + 1)/3 = -1/3$.

Replacing V by S, or A by T, replaces the left-chiral projector on the *electron* by the right-chiral one — this is Proposition 101.35 read backwards, a scalar bilinear connecting opposite chiralities — so the electron's momentum reverses relative to its spin and every sign above flips. The four entries of the table follow. □

Remark 101.39 (Why the spectrum shape did not change)*.* Proposition 101.11 was derived from Fermi's vector coupling and survives its replacement by $V - A$ unaltered, which is why the 1934 spectrum shape remained good evidence after 1957. The reason is that the shape came entirely from phase space, the matrix element being momentum-independent for an allowed transition; and $V - A$, like V, has a momentum-independent allowed matrix element. What changes is the *correlations* — the asymmetry with respect to a polarization, the electron's own polarization, the electron–neutrino angle — none of which anybody had measured before 1956. A theory can be right about everything that has been measured and wrong about its structure, and the way to tell is to measure something else.

101.4.2 Conserved vector current and universality

Two facts about Equation (101.35) are surprising once the strong interaction is remembered. The nucleon is not an elementary field but a bound state of strongly interacting constituents,

so the current a W sees is not the quark current but a dressed object; one would expect the dressing to change the coupling by a factor of order one, and to change it differently for the vector and the axial parts, and differently again for a nucleon and a nucleus. In fact the vector coupling is not renormalized at all, and the axial one is renormalized by only 27 per cent.

Phenomenon 101.40 (The vector coupling is not renormalized)**.** The vector coupling constant extracted from superallowed nuclear beta decay, where the decaying object is a bound state of strongly interacting quarks, agrees with the coupling extracted from muon decay, where nothing strong is involved, once the quark-mixing factor of Section 101.5 is applied. The axial coupling, by contrast, is renormalized to $\lambda = g_A/g_V = 1.2754(13)$ [2427].

Rests on Proposition 100.6 and Phenomenon 101.18.

Derivation. *Derives Phenomenon 101.40.*

The hypothesis that explains the first half is that of Feynman and Gell-Mann [1144]: the weak vector current $V^\mu = \bar\psi_p \gamma^\mu \psi_n$ and the isovector part of the electromagnetic current are three components of one isospin triplet,

$$V_a^\mu = \bar\psi_N \gamma^\mu \frac{\tau_a}{2} \psi_N \,, \qquad a = 1,2,3 \,, \tag{101.38}$$

with the weak charged currents the combinations $V_1^\mu \pm \mathrm{i} V_2^\mu$ and the isovector electromagnetic current V_3^μ. Isospin is a symmetry of the strong interaction, so all three currents are conserved, $\partial_\mu V_a^\mu = 0$, and their charges $T_a = \frac{1}{c}\int \mathrm{d}^3x\, V_a^0$ generate the algebra $[T_a, T_b] = \mathrm{i}\varepsilon_{abc} T_c$.

The argument is now exactly that of Proposition 100.6, transposed. A charge that generates a symmetry algebra takes eigenvalues fixed by the representation theory of that algebra — half-integers, in units set by the commutation relations — and a continuous dynamical effect cannot shift a discrete eigenvalue. The strong interaction commutes with T_a, so it cannot alter the matrix element of V_a^μ between states of definite isospin at zero momentum transfer, any more than the strong interaction alters the electric charge of the proton, which by Proposition 100.6 and the Ward–Takahashi identity of Section 100.5.2 is exactly the positron's. The vector coupling of a nucleon is therefore the same as that of a bare quark, and the constancy of $\mathcal{F}t$ across the superallowed multiplet in Phenomenon 101.18 is the measurement that confirms it: a renormalization that depended on nuclear structure would show up as a spread in those fourteen numbers, and the spread is a part in 10^4.

The axial current is a different matter, because it is *not* conserved: nothing in the strong interaction is invariant under a chiral rotation once the quarks have mass, and, more importantly, the axial symmetry is spontaneously broken. There is accordingly no protection theorem and g_A is renormalized. The measured 27 per cent is small only because the quark masses are small, and its theoretical status belongs to the chiral symmetry of Chapter 102. That the vector coupling is protected and the axial one is not is a structural prediction, and the two measured numbers — 1 and 1.2754 — are the test of it. □

The partial conservation that replaces the exact conservation has a sharp consequence, which is one of the very few quantitative relations in this chapter that connects a weak amplitude to a purely strong one.

Definition 101.41 (The pion decay constant and PCAC)**.** Define f_π by the matrix element of the axial current between the vacuum and a one-pion state,

$$\langle 0 |\, A_a^\mu(0) \,| \pi_b(q) \rangle = \mathrm{i} f_\pi \, \delta_{ab} \, q^\mu \,, \tag{101.39}$$

with $f_\pi c = 92.3\,\mathrm{MeV}$ in the normalization used here [2427], so that f_π has the dimension of a momentum. The hypothesis of a *partially* conserved axial current is that $\partial_\mu A_a^\mu$ is not zero but is proportional to the pion field,

$$\partial_\mu A_a^\mu = f_\pi \left(\frac{m_\pi c}{\hbar} \right)^2 \hbar \, \phi_a \,, \tag{101.40}$$

which is consistent with Equation (101.39) and vanishes in the limit $m_\pi \to 0$: the pion is the almost-Goldstone boson of the broken chiral symmetry.

Proposition 101.42 (Goldberger–Treiman)**.** *Let $g_{\pi N}$ be the pion–nucleon coupling constant measured in strong-interaction scattering and g_A the axial coupling of the nucleon measured in beta decay. Then*

$$f_\pi \, g_{\pi N} = g_A \, m_N c \,, \tag{101.41}$$

up to corrections of order m_π^2/m_N^2.

Rests on Equations (101.39) and (101.40).

Proof. *Derives Proposition 101.42.*

Take the matrix element of $\partial_\mu A_a^\mu$ between two nucleon states of momenta p and p', and evaluate it in two ways.

On the left, use the general decomposition of the axial matrix element consistent with Lorentz invariance and parity,

$$\langle N(p')| \, A_a^\mu \, |N(p)\rangle = \bar u(p') \left[g_A\!\left(q^2\right) \gamma^\mu \gamma^5 + \frac{h_A(q^2)}{m_N c} \, q^\mu \gamma^5 \right] \frac{\tau_a}{2} u(p) \,, \qquad q = p' - p \,, \tag{101.42}$$

with g_A and h_A dimensionless. Contracting with q_μ, and using $\bar u(p')\gamma^\mu q_\mu \gamma^5 u(p) = \bar u(p') \, (\gamma \cdot p' - \gamma \cdot p) \, \gamma^5 u(p) = 2 m_N c \, \bar u(p') \gamma^5 u(p)$ — the Dirac equation applied to each spinor, the sign of the second term flipped by $\gamma^5 \gamma \cdot p = -\gamma \cdot p \, \gamma^5$ so that the two add — gives

$$q_\mu \, \langle N(p')| \, A_a^\mu \, |N(p)\rangle = \left[2 m_N c \, g_A\!\left(q^2\right) + \frac{q^2}{m_N c} h_A\!\left(q^2\right) \right] \bar u(p') \gamma^5 \frac{\tau_a}{2} u(p) \,. \tag{101.43}$$

On the right, use Equation (101.40): the divergence is $f_\pi (m_\pi c/\hbar)^2 \hbar \phi_a$, and the nucleon matrix element of the pion field is the pion propagator times the pion–nucleon vertex, which with the coupling normalized as $\mathcal{L}_{\pi N} = -\mathrm{i} g_{\pi N} \bar\psi_N \gamma^5 \tau_a \psi_N \phi_a$ reads

$$\langle N(p')| \, \hbar \phi_a \, |N(p)\rangle = \frac{2 \hbar \, g_{\pi N}}{q^2 - m_\pi^2 c^2} \, \bar u(p') \gamma^5 \frac{\tau_a}{2} u(p) \cdot \hbar \,.$$

Equating the two expressions and cancelling the common spinor factor,

$$2 m_N c \, g_A\!\left(q^2\right) + \frac{q^2}{m_N c} h_A\!\left(q^2\right) = \frac{2 f_\pi g_{\pi N} \, m_\pi^2 c^2}{q^2 - m_\pi^2 c^2} \,. \tag{101.44}$$

Now set $q^2 = 0$, where the second term on the left vanishes provided h_A has no pole at the origin — which it has not, the pion pole sitting at $q^2 = m_\pi^2 c^2$ — and the right-hand side becomes $-2 f_\pi g_{\pi N}$. Hence $m_N c \, g_A(0) = -f_\pi g_{\pi N}$, which is Equation (101.41) up to the sign convention for $g_{\pi N}$. The mechanism is worth naming: the right-hand side of Equation (101.44) is proportional to m_π^2 and would vanish in the chiral limit, but the pion propagator supplies a compensating $1/m_\pi^2$ at $q^2 = 0$, so the product survives. That is why an exactly conserved axial current would still give a nonzero relation, and why the corrections are of relative order m_π^2/m_N^2 rather than of order one. Numerically, with $f_\pi c = 92.3\,\mathrm{MeV}$, $g_{\pi N} = 13.05$, $m_N c^2 = 938.9\,\mathrm{MeV}$ and $g_A = 1.2754$ [2427],

$$f_\pi c \, g_{\pi N} = 1204\,\mathrm{MeV} \,, \qquad g_A m_N c^2 = 1197\,\mathrm{MeV} \,,$$

agreeing to 0.6 per cent, which is of the expected order $m_\pi^2/m_N^2 \approx 0.02$. The relation is [1147]. What makes it worth the page is that it connects three quantities measured in three different ways — a weak decay rate, a strong scattering length and a leptonic pion decay — and is right to one per cent. □

Phenomenon 101.43 (One coupling governs every charged-current process)**.** The strength of the charged-current interaction extracted from the purely leptonic decay $\mu^- \to e^- \bar{\nu}_e \nu_\mu$, from superallowed nuclear beta decay and from τ decay is one and the same number. The muon lifetime is measured to a part per million, $\tau_\mu = 2.196\,981\,1 \times 10^{-6}\,\mathrm{s}$ [2225], and defines

$$\frac{G_F}{(\hbar c)^3} = 1.166\,378\,7 \times 10^{-5}/\mathrm{GeV}^2 \tag{101.45}$$

[2427]; the same constant reproduces the nuclear rates once the rotation of Section 101.5 is applied to the strangeness-changing part. Universality was read off the data, not imposed: nothing in Fermi's construction required the coupling of a muon and the coupling of a nucleon to agree.

Rests on Equations (101.20) *and* (101.53) *and Phenomenon 101.40.*

Derivation. Derives Phenomenon 101.43.

The three determinations are as follows. Muon decay gives G_F directly by Equation (101.53), with no hadron anywhere in the problem and hence no strong correction beyond the electromagnetic ones. Superallowed nuclear decay gives $G_F\,|V_{ud}|$ by Equation (101.20), and the vector coupling entering it is unrenormalized by Phenomenon 101.40; dividing out $|V_{ud}| = 0.97435$ returns the muon-decay value. Tau decay to $\nu_\tau e \bar{\nu}_e$ is muon decay with a heavier lepton, so Equation (101.53) applies with $m_\mu \to m_\tau$; the predicted partial width, using $m_\tau c^2 = 1776.93\,\mathrm{MeV}$ and the same G_F, reproduces the measured branching fraction of 17.8 % together with the measured lifetime $290.3 \times 10^{-15}\,\mathrm{s}$ [2427]. Three processes whose energy scales differ by a factor 10^4 and whose participants differ in every other respect share one constant. Puppi's observation that the couplings of (μ, ν_μ), (e, ν_e) and (n, p) to one another form a triangle of equal sides, and Konopinski and Mahmoud's version of the same statement [1022], are the origin of the phrase *universal Fermi interaction*; the residual deficit in the strangeness-changing sector, which is what stops the triangle from closing exactly, is Section 101.5. □

101.4.3 Muon decay and the Fermi constant

The muon is the cleanest weak decay in existence. Nothing that feels the strong interaction appears in either the initial or the final state; the final state has three particles, two of which are essentially massless; and the lifetime can be measured by simply counting. It is therefore the process that *defines* G_F, and this subsection carries out the calculation in full.

Definition 101.44 (The effective Lagrangian for muon decay)**.** At energies far below the mass of the W, the decay $\mu^- \to e^- \bar{\nu}_e \nu_\mu$ is governed by the contact interaction

$$\mathcal{L}_\mu = -\frac{G_F}{\sqrt{2}} \left[\bar{\psi}_{\nu_\mu} \gamma^\alpha \left(\mathbb{1} - \gamma^5\right) \psi_\mu\right] \left[\bar{\psi}_e \gamma_\alpha \left(\mathbb{1} - \gamma^5\right) \psi_{\nu_e}\right] + \mathrm{h.c.}\,, \tag{101.46}$$

whose coefficient $G_F/\sqrt{2}$ carries $\mathrm{J\,m^3}$ by Proposition 101.7; the factor $\sqrt{2}$ is a convention, fixed once and for all here, and Section 101.6.1 shows what it becomes when the W is restored. The corresponding invariant amplitude is

$$\mathcal{M} = \frac{G_F}{\sqrt{2}\,\hbar c}\,\mathcal{A}\,, \qquad \mathcal{A} = \left[\bar{u}_{\nu_\mu} \gamma^\alpha \left(\mathbb{1} - \gamma^5\right) u_\mu\right] \left[\bar{u}_e \gamma_\alpha \left(\mathbb{1} - \gamma^5\right) v_{\bar{\nu}_e}\right]\,, \tag{101.47}$$

of SI dimension $\hbar^2$, in agreement with the general statement of Proposition 100.14: the four external spinors contribute a squared momentum, and $G_F/\hbar c$ carries m^2.

Proposition 101.45 (Decay-rate master formula)**.** *For the decay of a particle of energy E_1 into n particles,*

$$\mathrm{d}\Gamma = \frac{c^2}{2E_1}\,\overline{|\mathcal{M}|^2}\,\mathrm{d}\Phi_n\,, \tag{101.48}$$

with the invariant phase space $\mathrm{d}\Phi_n$ *of Equation* (100.39)*. In the rest frame the prefactor is* $1/2m_1$*; in any other frame the rate is smaller by the Lorentz factor* m_1c^2/E_1*, which is time dilation.*

Rests on Equations (100.39) *and* (101.10) *and Proposition 100.14.*

Proof. *Derives Proposition 101.45.*

Quantize in a box of volume V with periodic boundary conditions and normalize each one-particle state to one particle in the box,

$$\psi(x) = \frac{1}{\sqrt{V}}\,\frac{u(p,s)}{\sqrt{2E/c}}\,\mathrm{e}^{-\mathrm{i}p\cdot x/\hbar}\,, \tag{101.49}$$

which is correctly normalized because $u^\dagger u = 2E/c$ by Proposition 100.14, so that $\int_V \mathrm{d}^3x\,\psi^\dagger\psi = 1$; note that the prefactor is dimensionally forced, $1/\mathrm{m}^{3/2}$ on the left and $u/\sqrt{2E/c}$ dimensionless. With $\mathcal{H}_{\mathrm{int}} = -\mathcal{L}_\mu$, which holds because the interaction contains no derivatives, the matrix element between the one-muon state and a three-particle final state is

$$\langle f|\,\mathcal{H}_{\mathrm{int}}\,|i\rangle = \frac{G_F}{\sqrt{2}}\,\frac{1}{V^2}\,\frac{\mathcal{A}}{\prod_{k=1}^{4}\sqrt{2E_k/c}}\;\times V\delta_{\boldsymbol{p}_1,\,\boldsymbol{p}_2+\boldsymbol{p}_3+\boldsymbol{p}_4}\,,$$

the four factors $V^{-1/2}$ coming from the four fields and the factor V from $\int_V \mathrm{d}^3x$, which also enforces momentum conservation as a Kronecker delta. Squaring, using $\delta_{\mathrm{K}}^2 = \delta_{\mathrm{K}}$, and inserting into the golden rule Equation (101.10) with the continuum sum $\sum_f \to \prod_{k=2}^{4} V\mathrm{d}^3p_k/(2\pi\hbar)^3$ and the continuum form of the Kronecker delta, $\delta_{\mathrm{K}} = (2\pi\hbar)^3V^{-1}\delta^3(\boldsymbol{p}_1 - \sum \boldsymbol{p}_f)$, gives

$$\Gamma = \frac{2\pi}{\hbar}\,\frac{G_F^2}{2}\,\frac{c}{2E_1}\int|\mathcal{A}|^2\left[\prod_f \frac{c\,\mathrm{d}^3p_f}{(2\pi\hbar)^3\,2E_f}\right](2\pi\hbar)^3\,\delta^3\!\left(\boldsymbol{p}_1 - \sum_f \boldsymbol{p}_f\right)\delta\!\left(E_1 - \sum_f E_f\right)\,,$$

every power of V having cancelled: V^3 from the three measures against V^{-2} from the squared matrix element and V^{-1} from the delta. Now use $(2\pi\hbar)^4\,\delta^4(p) = (2\pi\hbar)^3\,\delta^3(\boldsymbol{p})\cdot 2\pi\hbar c\,\delta(E)$, so that $2\pi\delta(E)\,(2\pi\hbar)^3\,\delta^3 = (2\pi\hbar)^4\,\delta^4/\hbar c$, to recognize $\mathrm{d}\Phi_3$ of Equation (100.39) entire:

$$\Gamma = \frac{G_F^2}{4\hbar^2E_1}\int|\mathcal{A}|^2\,\mathrm{d}\Phi_3\,.$$

Finally substitute $|\mathcal{A}|^2 = 2\hbar^2c^2\,|\mathcal{M}|^2/G_F^2$ from Equation (101.47), which cancels G_F and $\hbar$ together and leaves Equation (101.48). Nothing in the last two steps used the specific interaction: the coupling and the spinor structure enter only through $\mathcal{M}$, while the factors of V and of $2E/c$ belong to the external legs. Hence the formula holds for any $1 \to n$ decay, which is how it is used again in Section 101.6.4. □

Lemma 101.46 (The spin-summed square)**.** *With the neutrinos massless and the muon spin averaged,*

$$\overline{|\mathcal{A}|^2} = 128\,(p_\mu\cdot p_{\bar\nu_e})\,(p_e\cdot p_{\nu_\mu})\,. \tag{101.50}$$

The muon pairs with the electron antineutrino *and the electron with the* muon neutrino*; the pairing is a consequence of the current structure and not a choice.*

Rests on Equation (100.33) *and Lemma 100.23.*

Proof. *Derives Lemma 101.46.*

Write $\mathcal{A} = L^\alpha M_\alpha$ with $L^\alpha = \bar u_{\nu_\mu}\gamma^\alpha(1-\gamma^5)u_\mu$ and $M_\alpha = \bar u_e\gamma_\alpha(1-\gamma^5)v_{\bar\nu_e}$. Summing over final spins and averaging over the two muon spins, $\overline{|\mathcal{A}|^2} = \tfrac{1}{2}L^{\alpha\beta}M_{\alpha\beta}$ with

$$L^{\alpha\beta} = \mathrm{tr}\left[\gamma\cdot p_{\nu_\mu}\,\gamma^\alpha\left(1-\gamma^5\right)(\gamma\cdot p_\mu + m_\mu c)\,\gamma^\beta\left(1-\gamma^5\right)\right]\,,$$

the spin sums being those of Equation (100.33) and the conjugation using $\gamma^0\left[\gamma^\beta(1-\gamma^5)\right]^\dagger\gamma^0 = \gamma^\beta(1-\gamma^5)$. The mass term drops, because it leaves an odd number of gamma matrices in the trace. Using $(1-\gamma^5)^2 = 2(1-\gamma^5)$ and the trace theorems of Lemma 100.23 together with $\operatorname{tr}(\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma\gamma^5) = -4\mathrm{i}\varepsilon^{\mu\nu\rho\sigma}$,

$$L^{\alpha\beta} = 8\left[p^\alpha_{\nu_\mu}p^\beta_\mu + p^\beta_{\nu_\mu}p^\alpha_\mu - \eta^{\alpha\beta}\left(p_{\nu_\mu}\cdot p_\mu\right) + \mathrm{i}\varepsilon^{\rho\alpha\sigma\beta}p_{\nu_\mu\rho}p_{\mu\sigma}\right],$$

and $M_{\alpha\beta}$ is the same expression with $p_{\nu_\mu}\to p_e$ and $p_\mu\to p_{\bar\nu_e}$ and the indices lowered. Abbreviate $P_1 = p_{\nu_\mu}$, $P_2 = p_\mu$, $P_3 = p_e$, $P_4 = p_{\bar\nu_e}$ — numbered rather than lettered, since c is the speed of light throughout this book — and contract.

The symmetric parts contract to

$$\begin{aligned}&64\big[2(P_1\cdot P_3)(P_2\cdot P_4) + 2(P_1\cdot P_4)(P_2\cdot P_3)\\&\quad - 4(P_1\cdot P_2)(P_3\cdot P_4) + 4(P_1\cdot P_2)(P_3\cdot P_4)\big]\\&= 128\left[(P_1\cdot P_3)(P_2\cdot P_4) + (P_1\cdot P_4)(P_2\cdot P_3)\right],\end{aligned}$$

the two trace terms cancelling exactly against the term in which the two metrics meet. The product of the two Levi-Civita terms, using $\varepsilon^{\rho\alpha\sigma\beta}\varepsilon_{\kappa\alpha\lambda\beta} = -2\left(\delta^\rho_\kappa\delta^\sigma_\lambda - \delta^\rho_\lambda\delta^\sigma_\kappa\right)$ and $\mathrm{i}^2 = -1$, contributes

$$128\left[(P_1\cdot P_3)(P_2\cdot P_4) - (P_1\cdot P_4)(P_2\cdot P_3)\right],$$

while the cross terms between a symmetric and an antisymmetric tensor vanish identically. Adding, the $(P_1\cdot P_4)(P_2\cdot P_3)$ pieces cancel and the others double:

$$L^{\alpha\beta}M_{\alpha\beta} = 256\left(p_{\nu_\mu}\cdot p_e\right)\left(p_\mu\cdot p_{\bar\nu_e}\right), \tag{101.51}$$

which halved is Equation (101.50). This is the single most useful identity in weak-interaction physics: any two $V-A$ currents, contracted and summed over spins, collapse to a product of two dot products with the coefficient 256, and the cancellation of the other pairing is what fixes which particle goes with which. □

Lemma 101.47 (The two-neutrino phase space)**.** *For two massless particles sharing a total four-momentum q,*

$$\int \mathrm{d}\Phi_2\; k^\alpha k'^\beta = \frac{1}{96\pi\hbar^2}\left(q^2\eta^{\alpha\beta} + 2q^\alpha q^\beta\right), \tag{101.52}$$

where $\mathrm{d}\Phi_2$ is the two-body element of Equation (100.39) with $k+k'=q$.

Rests on Equation (100.39) and Proposition 100.24.

Proof. *Derives Lemma 101.47.*

The integral is a Lorentz tensor built from q alone, so $I^{\alpha\beta} = Aq^2\eta^{\alpha\beta} + Bq^\alpha q^\beta$. Contract with $\eta_{\alpha\beta}$: since $k^2 = k'^2 = 0$ gives $2k\cdot k' = q^2$, the left side is $\frac12 q^2\Phi_2$ with $\Phi_2 = \int\mathrm{d}\Phi_2$, so $4A + B = \frac12\Phi_2$. Contract with $q_\alpha q_\beta$: $q\cdot k = k'\cdot k = \frac12 q^2$ and likewise $q\cdot k'$, so the left side is $\frac14 q^4\Phi_2$ and $A+B = \frac14\Phi_2$. Hence $A = \Phi_2/12$ and $B = \Phi_2/6$. The total two-body phase space for massless products follows from Proposition 100.24, where $\mathrm{d}\Phi_2 = \hbar^{-2}\left|\boldsymbol{p}_f\right|\mathrm{d}\Omega/16\pi^2\sqrt{s}$ with $|\boldsymbol{p}_f| = \frac12\sqrt{s}$ here, giving $\Phi_2 = 4\pi/32\pi^2\hbar^2 = 1/8\pi\hbar^2$; note that it carries $\hbar^{-2}$, as every element of $\mathrm{d}\Phi$ must. Inserting gives Equation (101.52). □

Theorem 101.48 (The muon decay rate)**.** *Neglecting the electron mass and radiative corrections, the $V-A$ interaction Equation (101.46) gives*

$$\Gamma_\mu = \frac{1}{\tau_\mu} = \frac{1}{192\pi^3\hbar}\left(\frac{G_F}{(\hbar c)^3}\right)^2\left(m_\mu c^2\right)^5 = \frac{G_F^2\left(m_\mu c^2\right)^5}{192\pi^3\hbar\,(\hbar c)^6}, \tag{101.53}$$

and the normalized electron energy spectrum, with $x = 2E_e/m_\mu c^2 \in [0,1]$,

$$\frac{1}{\Gamma_\mu}\frac{\mathrm{d}\Gamma_\mu}{\mathrm{d}x} = 2x^2\,(3-2x)\ . \tag{101.54}$$

Rests on Equations (101.46), (101.48) *and* (101.50).

Proof. *Derives Theorem 101.48.*

Combine Equation (101.48), Equation (101.47) and Equation (101.50):

$$\overline{|\mathcal{M}|^2} = \frac{G_F^2}{2\hbar^2c^2}\cdot 128\,(p_\mu\cdot p_{\bar\nu_e})\,(p_e\cdot p_{\nu_\mu}) = \frac{64G_F^2}{\hbar^2c^2}\,(p_\mu\cdot p_{\bar\nu_e})\,(p_e\cdot p_{\nu_\mu})\ .$$

The two neutrinos are not observed, so integrate over them first. Writing $q = p_\mu - p_e$ for their total four-momentum and using Equation (101.52) with $k = p_{\nu_\mu}$ and $k' = p_{\bar\nu_e}$,

$$\int \mathrm{d}\Phi_2\,(p_\mu\cdot p_{\bar\nu_e})\,(p_e\cdot p_{\nu_\mu}) = p_{e\alpha}p_{\mu\beta}\int \mathrm{d}\Phi_2\,k^\alpha k'^\beta = \frac{q^2\,(p_e\cdot p_\mu) + 2\,(q\cdot p_e)\,(q\cdot p_\mu)}{96\pi\hbar^2}\ .$$

Put the muon at rest, $p_\mu = (m_\mu c, \mathbf{0})$, and neglect the electron mass, $p_e^2 = 0$. Then, writing $y = p_\mu\cdot p_e = m_\mu E_e$,

$$q^2 = m_\mu^2c^2 - 2y\,,\qquad q\cdot p_e = y\,,\qquad q\cdot p_\mu = m_\mu^2c^2 - y\,,$$

so the numerator is $\left(m_\mu^2c^2-2y\right)y + 2y\left(m_\mu^2c^2-y\right) = m_\mu^2E_e\,(3m_\mu c^2 - 4E_e)$. The remaining electron integral is the one-particle element of Equation (100.39) with $\mathrm{d}^3p_e = 4\pi E_e^2\mathrm{d}E_e/c^3$ for a massless electron:

$$\int \frac{c\,\mathrm{d}^3p_e}{(2\pi\hbar)^3 2E_e}\,m_\mu^2E_e\left(3m_\mu c^2 - 4E_e\right) = \frac{4\pi m_\mu^2}{2c^2\,(2\pi\hbar)^3}\int_0^{m_\mu c^2/2} E_e^2\left(3m_\mu c^2 - 4E_e\right)\mathrm{d}E_e\,,$$

the upper limit being the greatest energy an electron can take when the two neutrinos go together against it. The integral is $m_\mu c^2 W^3 - W^4$ with $W = m_\mu c^2/2$, that is $\left(m_\mu c^2\right)^4/16$, and $(2\pi\hbar)^3 = 8\pi^3\hbar^3$, so the electron integral equals $m_\mu^2\left(m_\mu c^2\right)^4/64\pi^2\hbar^3c^2$. Assembling with the prefactor $c^2/2m_\mu c^2$ of Equation (101.48),

$$\Gamma_\mu = \frac{1}{2m_\mu}\cdot\frac{64G_F^2}{\hbar^2c^2}\cdot\frac{1}{96\pi\hbar^2}\cdot\frac{m_\mu^2\left(m_\mu c^2\right)^4}{64\pi^2\hbar^3c^2} = \frac{G_F^2m_\mu\left(m_\mu c^2\right)^4}{192\pi^3\hbar^7c^4}\,,$$

since $64/(2\cdot 96\cdot 64) = 1/192$. Writing $m_\mu = \left(m_\mu c^2\right)/c^2$ and $\hbar^7c^6 = \hbar\,(\hbar c)^6$ gives Equation (101.53). The spectrum Equation (101.54) is the integrand before the last integration: with $E_e = xm_\mu c^2/2$ the integrand is proportional to $E_e^2\,(3m_\mu c^2 - 4E_e) \propto x^2\,(3-2x)$, and $\int_0^1 2x^2(3-2x)\mathrm{d}x = 2\left(1-\frac{1}{2}\right) = 1$ fixes the normalization. □

Remark 101.49 (What the fifth power is). The $\left(m_\mu c^2\right)^5$ in Equation (101.53) is forced and could have been written down without the calculation. Γ has dimension 1/s; the only quantities available are G_F, $m_\mu c^2$, $\hbar$ and c; and $G_F^2/(\hbar c)^6$ carries $1/\mathrm{J}^4$, so exactly four powers of energy are needed to make it dimensionless and a fifth to convert $\hbar^{-1}$ into a rate. What the calculation supplies is the pure number $1/192\pi^3$, and that number is where the $V-A$ structure and the three-body phase space actually live.

The same argument applies to the tau, and it is worth being exact about how much of the answer it supplies. The tau is 16.82 times heavier than the muon [2427], so the fifth power predicts its *partial* width to $e^-\bar\nu_e\nu_\tau$ to be $16.82^5 = 1.35\times10^6$ times the muon's total width. That alone would give the tau a lifetime of $1.63\times10^{-12}\,\mathrm{s}$. The measured lifetime is $2.903\times10^{-13}\,\mathrm{s}$ [2427], shorter by a further factor 5.6, and that factor is not dimensional

analysis: it is $1/\mathcal{B}(\tau^- \to e^-\bar{\nu}_e\nu_\tau) = 1/0.178$, the hadronic and muonic channels open to a tau and closed to a muon. Putting the two together, $2.196\,981\,1\times10^{-6}\,\mathrm{s}\times0.178/1.35\times10^{6} = 2.91\times10^{-13}\,\mathrm{s}$, which is the measurement. The ratio of *total* lifetimes is 7.6×10^{6}, and quoting that as what the fifth power predicts would credit dimensional analysis with a branching fraction it knows nothing about. It is the same fifth power as Sargent's rule Equation (101.14): two light particles sharing a release.

Phenomenon 101.50 (The muon lifetime measures the Fermi constant)**.** The measured positive-muon lifetime is

$$\tau_\mu = (2.196\,981\,1 \pm 0.000\,002\,2)\times10^{-6}\,\mathrm{s} \tag{101.55}$$

[2225], a determination to one part per million from 2×10^{12} recorded decays. Inserted into Equation (101.53), corrected for the electron mass and for the one-loop electromagnetic radiative correction, it yields Equation (101.45); conversely, that value of G_F predicts the lifetime.

Rests on Equations (101.45) *and* (101.53)*.*

Derivation. Derives Phenomenon 101.50.

Evaluate Equation (101.53) with $G_F/(\hbar c)^3 = 1.166\,378\,7\times10^{-5}/\mathrm{GeV}^2$, $m_\mu c^2 = 105.658\,375\,5\,\mathrm{MeV}$ and $\hbar = 6.582\,120\times10^{-25}\,\mathrm{GeV\,s}$:

$$\Gamma_\mu^{\mathrm{tree}} = \frac{(1.166\,378\,7\times10^{-5})^2\,(0.105\,658\,375\,5)^5}{192\pi^3\,(6.582\,120\times10^{-25}\,\mathrm{GeV\,s})}\,1/\mathrm{GeV}^4\,\mathrm{GeV}^5 = 4.5717\times10^{5}/\mathrm{s}\,,$$

that is $\tau = 2.1873\times10^{-6}\,\mathrm{s}$, which is 0.44 per cent below the measurement. Two known corrections account for the difference and for nothing else.

The first is kinematic. Retaining the electron mass multiplies Equation (101.53) by

$$f(z) = 1 - 8z + 8z^3 - z^4 - 12z^2\ln z\,, \qquad z = \frac{m_e^2}{m_\mu^2} = 2.339\times10^{-5}\,, \tag{101.56}$$

whose value here is $f = 0.99981$, a reduction of 1.87×10^{-4}: the electron mass eats into the available phase space, and the effect is second order in m_e/m_μ because the spectrum vanishes quadratically at low x by Equation (101.54).

The second is electromagnetic. The one-loop QED correction to the contact interaction, computed once and for all by the methods of Section 100.4, multiplies the rate by

$$1 + \Delta q\,, \qquad \Delta q = \frac{\alpha\,(m_\mu c^2)}{2\pi}\left(\frac{25}{4} - \pi^2\right) = -4.24\times10^{-3}\,, \tag{101.57}$$

using $\alpha(m_\mu c^2)^{-1} = 135.9$ for the fine-structure constant of Equation (100.3) run to the muon mass by Proposition 100.63. The combination $f(z)\,(1+\Delta q) = 0.99557$ gives

$$\Gamma_\mu = 4.5515\times10^{5}/\mathrm{s}\,, \qquad \tau_\mu = 2.1971\times10^{-6}\,\mathrm{s}\,,$$

against the measured $2.196\,981\,1\times10^{-6}\,\mathrm{s}$: agreement to 4×10^{-5}, which is the accuracy of the two-loop terms not included here and of the fourth digit of $\alpha(m_\mu c^2)$ used above. It is worth being explicit about the logic of the comparison. G_F is *defined* by inverting this calculation, so the agreement of the central values is a tautology; what is not a tautology, and is the actual content, is that the correction factors $f(z)$ and $1+\Delta q$ — computed from the electron mass and from QED, with no adjustable parameter — move the tree-level answer by exactly the 0.44 per cent by which it was wrong. □

Derivation pending #145

Muon decay, the parts not derived here: the phase-space function $f(z)$ of the electron-mass correction, obtained by repeating the phase-space integral of this section with $p_e^2 = m_e^2c^2$ retained; the one-loop electromagnetic correction $[25/4 - \pi^2]$, which is a QED calculation of the kind carried out in the QED chapter but applied to a contact operator; and the spin-dependent decay distribution $x^2[(3-2x) + P_\mu \cos\theta\,(2x-1)]$ obtained by keeping the muon spin projector in place of the average taken in the trace lemma. All three are Appendix A material and all three are quoted with their numerical values in the derivation above.

Definition 101.51 (Michel parameters)**.** The most general local, lepton-number-conserving, four-fermion interaction gives an electron spectrum of the form

$$\frac{\mathrm{d}\Gamma}{\mathrm{d}x} \propto x^2 \left[3\,(1-x) + \frac{2}{3}\rho\,(4x-3) \right], \tag{101.58}$$

together with three further parameters η, ξ and δ governing the electron-mass term and the dependence on the muon polarization [965]. The $V-A$ interaction predicts $\rho = \frac{3}{4}$, $\eta = 0$, $\xi = 1$, $\delta = \frac{3}{4}$.

Phenomenon 101.52 (The muon decay spectrum has the $V-A$ shape)**.** The measured Michel parameters are $\rho = 0.74979(26)$ and $\delta = 0.75047(34)$ [2427], agreeing with the $V-A$ prediction $\frac{3}{4}$ to a few parts in 10^4; measurements with polarized muons bound the admixture of a right-handed charged current at the same level [2170].

Rests on Equations (101.54) *and* (101.58).

Derivation. *Derives Phenomenon 101.52.*

Setting $\rho = \frac{3}{4}$ in Equation (101.58) gives $x^2\left[3 - 3x + \frac{1}{2}(4x-3)\right] = \frac{1}{2}x^2\,(3-2x)$, which is Equation (101.54) up to normalization; so the derived spectrum is the $\rho = \frac{3}{4}$ member of the family and the measurement is a direct test of the Lorentz structure, independent of the rate and therefore independent of G_F. The value $\frac{3}{4}$ is not generic: a scalar or tensor contact interaction gives $\rho = 0$, and a $V+A$ admixture moves ξ away from 1. What the agreement excludes is stated most sharply in the language of left–right symmetric models, where a right-handed W of mass M_2 mixing with the ordinary one shifts $P_\mu \xi\delta/\rho$ below unity; the measured value of that combination is consistent with 1 and bounds the mixing [2170]. □

The last property of the muon that belongs here is a null result, and it is one of the most stringent in physics.

Phenomenon 101.53 (Charged-current lepton flavour is conserved)**.** The decay $\mu^+ \to e^+\gamma$, which conserves electric charge, total lepton number, energy and angular momentum, and which would occur if the muon and the electron differed only in mass, has never been observed. The branching ratio is bounded by

$$\mathcal{B}\left(\mu^+ \to e^+\gamma\right) < 1.5 \times 10^{-13} \qquad (90\,\%\ \text{confidence}) \tag{101.59}$$

from the first dataset of MEG II combined with the full MEG sample; that dataset alone gives $< 7.5 \times 10^{-13}$ [2424].

Rests on Equation (101.46).

Derivation. *Derives Phenomenon 101.53.*

What the bound establishes is that the quantum numbers L_e and L_μ are separately conserved in charged-current processes, and not merely their sum. Were they not, the same W exchange that produces $\mu \to e\nu\bar{\nu}$ would, with the two neutrinos joined into a

loop and a photon attached, produce $\mu \to e\gamma$ at a rate suppressed only by α and by loop factors — of order 10^{-4} of the total, not 10^{-13}. The observed suppression is therefore a statement about the structure of the current and not about the size of a coupling. In the Standard Model with massive neutrinos the amplitude is not exactly zero but is suppressed by $\left(\Delta m_\nu^2 c^4 / M_W^2 c^4\right)^2 \sim 10^{-50}$, an unobservably small number, so any signal at all at the level of Equation (101.59) would be new physics; this is why the search continues. The experiment stops positive muons in a thin target and looks for a positron and a photon, back to back, each of energy $m_\mu c^2/2 = 52.8\,\mathrm{MeV}$, in time coincidence [2424]. □

The crossed version of muon decay is a scattering process, and its cross-section is needed twice below — once in Phenomenon 101.5 and once in Section 101.7.1.

Proposition 101.54 (Inverse muon decay)**.** *For $\nu_\mu e^- \to \mu^- \nu_e$ at centre-of-mass energy E_{cm} large compared with the muon mass, the $V-A$ contact interaction gives an isotropic cross-section*

$$\sigma = \frac{G_F^2 E_{\mathrm{cm}}^2}{\pi\,(\hbar c)^4}\,. \tag{101.60}$$

Rests on Equations (100.40)*,* (101.47) *and* (101.51)*.*

Proof. *Derives Proposition 101.54.*

The amplitude is Equation (101.47) with the external legs crossed, and Equation (101.51) applies unchanged: summing over final spins and averaging over the two electron spins — the neutrino has only one helicity, so nothing is averaged there —

$$\overline{|\mathcal{A}|^2} = \frac{1}{2}\cdot 256\,(p_{\nu_\mu}\cdot p_e)\,(p_\mu \cdot p_{\nu_e}) = 128\cdot\frac{s}{2}\cdot\frac{s}{2} = 32 s^2\,,$$

using $2p_{\nu_\mu}\cdot p_e = s$ for massless incoming particles and the same for the outgoing pair, with $s = E_{\mathrm{cm}}^2/c^2$ a squared momentum in the conventions of Notation 100.1. Then $\overline{|\mathcal{M}|^2} = 16 G_F^2 s^2/\hbar^2 c^2$, and Equation (100.40) with $|\boldsymbol{p}_f| = |\boldsymbol{p}_i|$ gives

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{1}{64\pi^2\hbar^2 s}\cdot\frac{16 G_F^2 s^2}{\hbar^2 c^2} = \frac{G_F^2 s}{4\pi^2\hbar^4 c^2}\,,$$

independent of angle. Multiplying by 4π and writing $s = E_{\mathrm{cm}}^2/c^2$ gives Equation (101.60). The isotropy is not an accident: the contact interaction has no length scale, so it can produce only s-wave scattering, and that fact is what Section 101.7.1 turns against the theory. Applied to a nucleon rather than to an electron, with the vector and axial nucleon couplings inserted, the same calculation gives Equation (101.5): the factor $1+3\lambda^2$ is the one Fermi channel plus the three Gamow–Teller channels of Proposition 101.15, each weighted by its coupling, and the factor $|V_{ud}|^2$ is the quark mixing of Section 101.5, which has no counterpart in the purely leptonic process computed here. □

101.5 Strangeness-changing currents

101.5.1 The strangeness deficit and the Cabibbo rotation

Phenomenon 101.43 is not quite true, and the size of the discrepancy is the subject of this section. When the same G_F that governs muon decay and superallowed nuclear decay is used to predict the rate of a decay that *changes strangeness* — $K^+ \to \mu^+\nu_\mu$ against $\pi^+ \to \mu^+\nu_\mu$, or $\Lambda \to p e^- \bar{\nu}_e$ against neutron decay — the prediction is too large by a factor of about twenty. Universality holds to a per cent for the strangeness-conserving processes and fails by a factor of twenty for the others.

Phenomenon 101.55 (Strangeness-changing decays are suppressed by one angle)**.** Semileptonic decays with $\Delta S = 1$ proceed at roughly one twentieth of the rate predicted by the coupling measured in $\Delta S = 0$ decays; and the deficit of the $\Delta S = 0$ rate below the purely leptonic strength is exactly what is needed to make up the difference. One angle,

$$\theta_C = 12.96^\circ\,, \qquad \sin\theta_C = 0.2243\,, \qquad \cos\theta_C = 0.9745\,, \tag{101.61}$$

accounts for both [1248, 2427].

Rests on Phenomenon 101.43 and Equation (101.20)*.*

Derivation. Derives Phenomenon 101.55.

Cabibbo's hypothesis is that the state entering the hadronic charged current is not the down quark but a rotation of the down-type mass eigenstates,

$$d' = d\cos\theta_C + s\sin\theta_C\,, \tag{101.62}$$

so that the hadronic current is

$$J^\mu_{\text{had}} = \bar\psi_u\gamma^\mu\left(\mathbb{1} - \gamma^5\right)\psi_{d'} = \cos\theta_C\,\bar\psi_u\gamma^\mu\left(\mathbb{1} - \gamma^5\right)\psi_d + \sin\theta_C\,\bar\psi_u\gamma^\mu\left(\mathbb{1} - \gamma^5\right)\psi_s\,, \tag{101.63}$$

and that the *same* G_F multiplies this current and the leptonic one. The two consequences are then not independent. A $\Delta S = 0$ amplitude carries $\cos\theta_C$ and a $\Delta S = 1$ amplitude carries $\sin\theta_C$, so their rates stand in the ratio

$$\frac{\Gamma\left(\Delta S = 1\right)}{\Gamma\left(\Delta S = 0\right)} = \tan^2\theta_C = 0.0530 = \frac{1}{18.9}\,,$$

which is the observed factor of about twenty; and the strangeness-conserving hadronic rate is suppressed relative to the purely leptonic one by $\cos^2\theta_C = 0.9496$, a deficit of about five per cent, which is precisely what Equation (101.20) sees as $|V_{ud}|^2$. The angle is therefore measured twice, in unrelated experiments: from the kaon rates as $\sin\theta_C = |V_{us}| = 0.2243$, and from the superallowed nuclear rates as $\cos\theta_C = |V_{ud}| = 0.97435$. The two determinations satisfy $\sin^2 + \cos^2 = 1$ to about a part in 10^3, and that residual is presently the sharpest unexplained tension in flavour physics; it is recorded, with the current numbers, in Phenomenon 103.7. □

The conceptual step in Equation (101.62) is larger than its algebra suggests, and it is the one that the whole of Chapter 103 elaborates: *the states that couple to the W are not the states of definite mass.* There are two bases for the down-type quarks, one diagonalizing the mass matrix and one diagonalizing the charged-current coupling, and they are related by a rotation. Nothing requires them to coincide, and they do not. Universality is restored not by making the couplings equal but by recognizing that a single coupling is shared out between channels, with the shares summing to one — which is a statement of orthogonality, and therefore the first appearance of the unitarity that Section 101.5.4 turns into a testable constraint.

101.5.2 Selection rules

Equation (101.63) carries two selection rules that were established empirically before it was written and that are among the sharpest tests of the current structure.

Proposition 101.56 (The $\Delta S = \Delta Q$ rule)**.** *A semileptonic decay of hadrons mediated by Equation* (101.63) *satisfies* $\Delta S = \Delta Q$*, where both refer to the hadronic system.*

Rests on Equation (101.63)*.*

Proof. *Derives Proposition 101.56.*

The strangeness-changing part of Equation (101.63) is $\bar{\psi}_u \Gamma \psi_s$, which annihilates an s quark and creates a u. The s quark carries $S = -1$ and charge $-\frac{1}{3}e$; the u carries $S = 0$ and $+\frac{2}{3}e$. Hence $\Delta S = 0 - (-1) = +1$ and $\Delta Q = +\frac{2}{3} - (-\frac{1}{3}) = +1$ in units of e, and the two are equal. The Hermitian conjugate term gives $\Delta S = \Delta Q = -1$. The rule is a prediction and not a convention: it forbids $\Sigma^+ \to n e^+ \nu_e$, which has $\Delta S = +1$ and $\Delta Q = -1$, while allowing $\Sigma^- \to n e^- \bar{\nu}_e$. The forbidden mode is bounded at the 10^{-5} level and the allowed one occurs at 10^{-3} [2427]: two decays of the same multiplet, differing by two orders of magnitude, with the sign of a charge as the only distinction. □

Proposition 101.57 (The $\Delta I = \frac{1}{2}$ rule, and its limits)**.** *The $\Delta S = 1$ current $\bar{\psi}_u \Gamma \psi_s$ carries isospin $\frac{1}{2}$, since u is a member of an isospin doublet and s an isospin singlet. Semileptonic $\Delta S = 1$ decays therefore obey $\Delta I = \frac{1}{2}$ exactly. For* nonleptonic *decays the current–current product contains both $\Delta I = \frac{1}{2}$ and $\Delta I = \frac{3}{2}$, so the rule is not a consequence of the structure; the observed dominance of the first by a factor of about* 20 *in amplitude is a strong-interaction effect and is not explained here.*

Rests on Equation (101.63).

Proof. *Derives Proposition 101.57.*

The isospin assignment is immediate. For the nonleptonic case, the effective operator is a product of two hadronic currents, each of isospin $\frac{1}{2}$ in the $\Delta S = 1$ sector and each of isospin 1 or 0 in the $\Delta S = 0$ sector; combining $\frac{1}{2} \otimes 1$ gives $\frac{1}{2} \oplus \frac{3}{2}$, so both appear with comparable coefficients at the level of the operator. The measurement is the ratio of the two kaon decay amplitudes $K^0 \to \pi^+\pi^-$ and $K^+ \to \pi^+\pi^0$, the second of which is pure $\Delta I = \frac{3}{2}$ because two pions in the $I = 0$ state cannot carry unit charge; the measured ratio of amplitudes is about $1/22$ [2427]. Honesty requires the statement that this factor is a QCD effect — short-distance renormalization of the operators plus long-distance hadronic dynamics — and is not derived in this book; it is the reason the rule is quoted here as an observation with its magnitude, rather than proved. The related but different question of why the operator basis is what it is belongs to Chapter 108. □

Derivation pending #146

The $[\Delta I = 1/2]$ enhancement in nonleptonic strange decays: why the isospin-one-half amplitude dominates the isospin-three-halves one by a factor near twenty, when the current–current operator produces both with comparable coefficients. The explanation combines short-distance operator mixing under the renormalization group with long-distance hadronic matrix elements, so the derivation belongs with the strong-interaction and renormalization-group material rather than here; the measured ratio is quoted with its source above.

101.5.3 The GIM mechanism

Equation (101.62) solves one problem and creates another, and the second is severe. If the state d' couples to the W, it also appears in whatever *neutral* current the theory possesses — and it appears bilinearly.

Proposition 101.58 (A single rotated doublet produces a flavour-changing neutral current)**.** *With the Cabibbo rotation Equation (101.62) and one down-type state d',*

$$\bar{\psi}_{d'} \Gamma \psi_{d'} = \cos^2\theta_C \, \bar{\psi}_d \Gamma \psi_d + \sin^2\theta_C \, \bar{\psi}_s \Gamma \psi_s + \sin\theta_C \cos\theta_C \left(\bar{\psi}_d \Gamma \psi_s + \bar{\psi}_s \Gamma \psi_d \right) , \tag{101.64}$$

whose last term changes strangeness without changing charge, with a coefficient $\sin\theta_C \cos\theta_C = 0.219$ — of the same order as everything else in the theory.

Rests on Equation (101.62).

Proof. *Derives Proposition 101.58.*
Substitute Equation (101.62) and expand. The cross terms are the last bracket. □

Such a term would drive $K^0 \to \mu^+\mu^-$, K^0–$\bar{K}^0$ mixing and $K^+ \to \pi^+\nu\bar{\nu}$ at rates comparable with ordinary weak processes. The measured branching ratio is $\mathcal{B}(K_L \to \mu^+\mu^-) = 6.84 \times 10^{-9}$ [2427], against $\mathcal{B}(K^+ \to \mu^+\nu_\mu) = 0.636$: the neutral mode is suppressed by some eight orders of magnitude relative to the charged one. Equation (101.64) is therefore not merely unobserved but excluded by a very large factor, and something must cancel it.

Theorem 101.59 (Glashow–Iliopoulos–Maiani)**.** *Let the down-type states be rotated by an orthogonal matrix and supplied to* two *up-type partners, so that the two charged-current doublets are*

$$\begin{pmatrix} u \\ d' \end{pmatrix}, \qquad \begin{pmatrix} c \\ s' \end{pmatrix}, \qquad \begin{aligned} d' &= d\cos\theta_C + s\sin\theta_C\,, \\ s' &= -d\sin\theta_C + s\cos\theta_C\,, \end{aligned} \tag{101.65}$$

with c a fourth quark. Then the neutral current is diagonal:

$$\bar{\psi}_{d'}\Gamma\psi_{d'} + \bar{\psi}_{s'}\Gamma\psi_{s'} = \bar{\psi}_d\Gamma\psi_d + \bar{\psi}_s\Gamma\psi_s\,, \tag{101.66}$$

and no flavour-changing neutral current survives at tree level [1423].
Rests on Proposition 101.58 and Equation (101.62).

Proof. *Derives Theorem 101.59.*
The two rotated states are the components of $\psi_{d'} = R\psi_d$ with R the orthogonal matrix of Equation (101.65). Then $\sum_i \bar{\psi}_{i'}\Gamma\psi_{i'} = \bar{\psi}_j \left(R^{\mathsf{T}}R\right)_{jk}\Gamma\psi_k = \bar{\psi}_j\delta_{jk}\Gamma\psi_k$, which is Equation (101.66). Explicitly, the cross terms are $+\sin\theta_C\cos\theta_C$ from d' and $-\sin\theta_C\cos\theta_C$ from s', and they cancel exactly. Nothing about the value of θ_C was used: the cancellation is the orthogonality of the rotation, not a numerical accident, which is why it survives to all orders in θ_C and generalizes without change to the unitary matrix of Section 101.5.4. □

Remark 101.60 (A particle predicted by a null result). Theorem 101.59 is worth pausing on as a piece of scientific reasoning. The input is the *absence* of a process; the output is the existence of a particle. Glashow, Iliopoulos and Maiani proposed the fourth quark in 1970 [1423] on no other evidence, and it was found four years later as the J/ψ resonance.

The cancellation is exact only at zero momentum transfer and only if the two up-type quarks were degenerate. At one loop, in the box diagram that generates K^0–$\bar{K}^0$ mixing, the two contributions differ by the difference of the squared quark masses, and the surviving amplitude is suppressed by $(m_c^2 - m_u^2)\,c^4/\left(M_W c^2\right)^2$ rather than cancelling outright. The measured K_L–K_S mass difference, $\Delta m_K c^2 = 3.484 \times 10^{-12}$ MeV [2427] — a fractional mass splitting of 7×10^{-15}, one of the smallest measured quantities in physics — therefore *bounds the charm mass*, and the bound obtained before the discovery was of order a few GeV/c^2, which is where the charm quark was found. The estimate is due to Gaillard and Lee, whose paper is not in this book's bibliography and is therefore named without a citation rather than cited to a key that does not exist; the mechanism itself is [1423], and the loop calculation belongs to Chapter 103.

101.5.4 Three generations: the CKM matrix

Kobayashi and Maskawa observed in 1973, before the fourth quark had been seen and while the third generation was entirely hypothetical, that the rotation of Equation (101.65) generalizes to N generations as an $N \times N$ unitary matrix, and that the number of *physically meaningful* parameters it contains depends on N in a way that has an observable consequence [1505].

Definition 101.61 (The quark mixing matrix)**.** With three generations the charged-current interaction of the quarks is

$$J^{\mu}_{\text{had}} = \sum_{i,k} \bar{\psi}_{u_i} \gamma^{\mu} \left(\mathbb{1} - \gamma^{5} \right) V_{ik} \, \psi_{d_k} \,, \tag{101.67}$$

where i runs over (u, c, t), k over (d, s, b), and V is the Cabibbo–Kobayashi–Maskawa matrix [1248, 1505]. Its entries are dimensionless pure numbers.

Theorem 101.62 (The mixing matrix is unitary)**.** *V is unitary, $V^{\dagger}V = VV^{\dagger} = \mathbb{1}$.*

Rests on Propositions 5.148 and 5.150.

Proof. *Derives Theorem 101.62.*

V is the matrix relating two orthonormal bases of one three-dimensional complex internal space — the basis of definite mass and the basis that couples to the W — and a change between orthonormal bases is unitary. That the space carries an inner product for which *both* bases are orthonormal is not automatic and is what has to be supplied: it holds because the flavour symmetry acting on the space can be realized by inner-product-preserving operators, which is Proposition 5.150, and because the space then decomposes cleanly, which is Proposition 5.148. Those two propositions are exactly where the licence to write the word "unitary" here comes from, and the argument is set out in Remark 5.152, which states it for this case and records the hypotheses under which each of the two fails. What unitarity buys is set out there as well: two families of relations, the row relations $\sum_k V_{ik} V^{*}_{jk} = \delta_{ij}$ and the column relations $\sum_i V^{*}_{ik} V_{il} = \delta_{kl}$, whose off-diagonal cases are sums of three complex numbers equal to zero and therefore closed triangles in the complex plane. The one that is measured is the subject of Chapter 111. □

Proposition 101.63 (Parameter counting, and why three generations permit CP violation)**.** *An $N \times N$ unitary mixing matrix contains*

$$\frac{N\,(N-1)}{2} \text{ angles} \qquad \text{and} \qquad \frac{(N-1)\,(N-2)}{2} \text{ phases} \tag{101.68}$$

that cannot be removed by redefining the phases of the quark fields. For $N = 2$ there is one angle and no phase, and the interaction is automatically invariant under CP; for $N = 3$ there are three angles and one phase, and CP invariance is violated unless that phase happens to vanish.

Rests on Theorems 101.59 and 101.62 and Equation (101.67).

Proof. *Derives Proposition 101.63.*

A unitary $N \times N$ matrix has N^2 real parameters, of which $N(N-1)/2$ are the angles of the real orthogonal subgroup and the remaining $N^2 - N(N-1)/2 = N(N+1)/2$ are phases. The quark fields may be rephased freely, $\psi_{u_i} \mapsto \mathrm{e}^{\mathrm{i}\alpha_i}\psi_{u_i}$ and $\psi_{d_k} \mapsto \mathrm{e}^{\mathrm{i}\beta_k}\psi_{d_k}$, which changes $V_{ik} \mapsto \mathrm{e}^{\mathrm{i}(\beta_k - \alpha_i)} V_{ik}$ and leaves every other term of the Lagrangian invariant, the mass terms because they are diagonal and the neutral currents because they are diagonal by Theorem 101.59. There are $2N$ such phases, but the overall common phase $\alpha_i = \beta_k$ acts trivially on V, so $2N - 1$ of the phases in V can be removed. The number surviving is $N(N+1)/2 - (2N-1) = (N-1)(N-2)/2$, which with the angles gives Equation (101.68).

For the CP statement, note that the CP conjugate of Equation (101.67) is the same expression with V replaced by V^{*}. If every entry of V can be made real by rephasing, the interaction is CP invariant; if a phase survives, it cannot, and CP is violated. With $N = 2$, $(N-1)(N-2)/2 = 0$: the Cabibbo matrix is real and Equation (101.62) conserves CP however θ_C is chosen. With $N = 3$ one phase survives. Kobayashi and Maskawa's inference [1505] was made in the reverse direction: CP violation had been observed in the neutral kaon system nine years earlier [1274], no two-generation model could accommodate

it without additional structure, and a third generation could. Two of its three quarks were subsequently found. It is one of the few occasions in this chapter on which theory ran ahead of experiment, and it is worth noting that it did so by counting parameters. □

Remark 101.64 (What this chapter does and does not claim about flavour). The measured magnitudes of the mixing elements, the closure of the unitarity triangle, the size of the CP-violating phase and the whole neutrino sector belong to Chapter 103, and the experiments that establish them to Chapter 111 and Chapter 114. What has been established *here* is narrower and prior to all of it: that the charged-current interaction has one universal strength; that the hadronic states it couples are rotations of the mass eigenstates; that the rotation is unitary, with the licence for the word supplied by Remark 5.152; that unitarity is what removes flavour-changing neutral currents; and that the number of surviving parameters, which is pure counting, decides whether CP can be violated at all. Nothing in the list is a fit. Every one of the five statements is a structural consequence of Equation (101.35) plus the observation that the two bases differ.

101.6 Neutral currents and the intermediate bosons

101.6.1 The intermediate vector boson hypothesis

Fermi's contact term was always understood as an approximation. The electromagnetic interaction it was modelled on is not a contact term: two currents are joined by a photon, and the interaction has infinite range because the photon is massless. Yukawa's insight [765], made for the nuclear force, was that the range of an exchange force is set by the mass of what is exchanged.

Proposition 101.65 (Range and mass)**.** *A field ϕ obeying the Klein–Gordon equation with mass M and a static point source produces the potential*

$$V(r) \propto \frac{\mathrm{e}^{-r/R}}{r}\,, \qquad R = \frac{\hbar}{Mc}\,, \tag{101.69}$$

so an interaction of observed range R requires a quantum of mass $M = \hbar/Rc$; equivalently, a momentum transfer q small compared with Mc cannot resolve the exchange and the interaction appears pointlike.

Rests on Equation (101.6).

Proof. *Derives Proposition 101.65.*

The static field equation is $\left(\nabla^2 - (Mc/\hbar)^2\right)\phi = -g\,\delta^3(\boldsymbol{x})$, whose spherically symmetric solution is obtained by writing $\phi = u(r)/r$, giving $u'' = (Mc/\hbar)^2 u$ and hence $u \propto \mathrm{e}^{-Mcr/\hbar}$ for the solution decaying at infinity. In momentum space the same statement is that the propagator is $\left[q^2 + (Mc)^2\right]^{-1}$ in the Euclidean region, which for $q \ll Mc$ is the constant $(Mc)^{-2}$: the propagator loses all memory of the momentum transfer, and the amplitude reduces to a product of two currents at one point, which is Equation (101.6). That is why Fermi's theory works, and it also says exactly how well: the neglected terms are of relative order $q^2/(Mc)^2$. □

The weak interaction has no measurable range, which by Proposition 101.65 means a very heavy quantum; but that argument alone gives only a lower bound. The quantitative statement comes from G_F itself, and requires knowing the coupling of the current to the boson. Glashow supplied the structure in 1961 [1206]: the charged currents $J_1^\mu \pm \mathrm{i}J_2^\mu$ of Equation (101.35) do not close on themselves under commutation but generate a third, *neutral* current, so a gauge theory of the weak interaction must have a group at least as large as SU(2), and once the electromagnetic U(1) is included the smallest possibility is

$\mathrm{SU(2)_L \times U(1)_Y}$. Weinberg [1362] and Salam [1381] completed it with the mechanism that gives the bosons mass, and 't Hooft proved that the result is renormalizable [1448, 1463] — which is what made the whole construction believable, since without that proof it would have been no better off than the theory it replaced. The derivations belong to Chapter 104; what is needed here is the structure of the couplings and their SI dimensions.

Definition 101.66 (The electroweak couplings in SI)**.** Let W_μ^a $(a = 1, 2, 3)$ and B_μ be the gauge fields of $\mathrm{SU(2)_L}$ and $\mathrm{U(1)}_Y$, each of the dimension of the electromagnetic four-potential, V s/m, and let the covariant derivative be

$$D_\mu = \partial_\mu + \frac{\mathrm{i}}{\hbar}\left(g\, W_\mu^a T^a + g'\, B_\mu \frac{Y}{2}\right), \tag{101.70}$$

in exact parallel with Equation (100.9). Then g and g' carry the dimension of *electric charge*, C, and the dimensionless couplings are

$$\alpha_W = \frac{g^2}{4\pi\varepsilon_0\hbar c}, \qquad \alpha' = \frac{g'^2}{4\pi\varepsilon_0\hbar c}, \tag{101.71}$$

built exactly as α is in Equation (100.3). The generators are $T^a = \frac{1}{2}\sigma^a$ acting on left-chiral doublets and zero on right-chiral singlets, and Y is the weak hypercharge.

Proposition 101.67 (The mixing angle and the fermion couplings)**.** *Let the neutral fields be rotated by an angle θ_W,*

$$W_\mu^3 = \cos\theta_W\, Z_\mu + \sin\theta_W\, A_\mu, \qquad B_\mu = -\sin\theta_W\, Z_\mu + \cos\theta_W\, A_\mu. \tag{101.72}$$

Requiring A_μ to couple to the electric charge Q and to nothing else fixes

$$e = g\sin\theta_W = g'\cos\theta_W, \qquad Q = T^3 + \frac{Y}{2}, \tag{101.73}$$

and then Z_μ couples to

$$\frac{e}{\sin\theta_W\cos\theta_W}\left(T^3 - \sin^2\theta_W\, Q\right) = \frac{g}{2\cos\theta_W}\,\gamma^\mu\left(g_V - g_A\gamma^5\right), \tag{101.74}$$

with, for each fermion,

$$g_V = T^3 - 2Q\sin^2\theta_W, \qquad g_A = T^3, \tag{101.75}$$

T^3 being the third component of weak isospin of the fermion's left-chiral *part. Also $\alpha_W = \alpha/\sin^2\theta_W$.*

Rests on Equation (101.70).

Proof. *Derives Proposition 101.67.*

Substituting Equation (101.72) into the neutral part of Equation (101.70), the coefficient of A_μ is $g\sin\theta_W T^3 + g'\cos\theta_W Y/2$. For this to be proportional to a single generator Q — which is what it means for the photon to couple to electric charge — the two coefficients must agree, $g\sin\theta_W = g'\cos\theta_W =: e$, and then the coefficient is $e\,(T^3 + Y/2)$, giving Equation (101.73). The coefficient of Z_μ is $g\cos\theta_W T^3 - g'\sin\theta_W Y/2$, which on eliminating g and g' in favour of e and substituting $Y/2 = Q - T^3$ becomes

$$\frac{e}{\sin\theta_W\cos\theta_W}\left[\cos^2\theta_W T^3 - \sin^2\theta_W\left(Q - T^3\right)\right] = \frac{e}{\sin\theta_W\cos\theta_W}\left(T^3 - \sin^2\theta_W Q\right).$$

Since T^3 acts only on the left-chiral part, write $T^3 \to T^3 P_\mathrm{L} = \frac{1}{2}T^3\,(\mathbb{1} - \gamma^5)$; then $T^3 P_\mathrm{L} - \sin^2\theta_W Q = \frac{1}{2}\left[(T^3 - 2Q\sin^2\theta_W) - T^3\gamma^5\right]$, which with $e/\sin\theta_W = g$ is Equation (101.74) and Equation (101.75). Finally $e = g\sin\theta_W$ squared and divided by $4\pi\varepsilon_0\hbar c$ gives $\alpha = \alpha_W\sin^2\theta_W$. □

Proposition 101.68 (The low-energy limit reproduces Fermi)**.** *Let the charged-current interaction be*

$$\mathcal{L}_{\mathrm{CC}} = -\frac{gc}{2\sqrt{2}}\left[\bar{\psi}_\nu \gamma^\mu \left(\mathbb{1} - \gamma^5\right)\psi_\ell \, W_\mu^+ + h.c.\right], \tag{101.76}$$

and let the W propagator be that of a massive vector field,

$$\widetilde{D}^W_{\mu\nu}(q) = \frac{-\mathrm{i}\mu_0\hbar^3 c}{q^2 - M_W^2c^2 + \mathrm{i}\epsilon}\left(\eta_{\mu\nu} - \frac{q_\mu q_\nu}{M_W^2c^2}\right), \tag{101.77}$$

which is Equation (100.17) with the mass restored. Then for $|q^2| \ll M_W^2c^2$ the exchange of a W between two such currents is indistinguishable from Equation (101.46) with

$$\frac{G_F}{\sqrt{2}} = \frac{g^2\mu_0\hbar^2}{8M_W^2}, \qquad \textit{equivalently} \qquad \frac{G_F}{\sqrt{2}\,(\hbar c)^3} = \frac{\pi\alpha_W}{2\,(M_Wc^2)^2} = \frac{\pi\alpha}{2\sin^2\theta_W\,(M_Wc^2)^2}. \tag{101.78}$$

Rests on Equations (100.17) and (101.46) and Proposition 100.14.

Proof. *Derives Proposition 101.68.*

The vertex factor read off Equation (101.76) by the rule of Proposition 100.14 — expand $\exp(\mathrm{i}S/\hbar)$ with $S = c^{-1}\int \mathrm{d}^4x\,\mathcal{L}$ — is $-\mathrm{i}g\gamma^\mu(\mathbb{1}-\gamma^5)/2\sqrt{2}\hbar$, of the same SI dimension $\mathrm{C}/(\mathrm{J\,s})$ as the QED vertex Equation (100.24), since g is a charge by Definition 101.66. Two vertices joined by Equation (101.77) contribute, in the limit $q^2 \to 0$,

$$\left(\frac{-\mathrm{i}g}{2\sqrt{2}\hbar}\right)^2 \frac{-\mathrm{i}\mu_0\hbar^3 c\,\eta_{\mu\nu}}{q^2 - M_W^2c^2} \;\longrightarrow\; -\frac{\mathrm{i}g^2\mu_0\hbar}{8M_W^2c}\,\eta_{\mu\nu},$$

the $q_\mu q_\nu$ term being dropped because it is contracted with currents of light external fermions, on which it gives a factor $m_\ell c/M_Wc$ squared by the argument of Equation (100.18). The contact interaction Equation (101.46), treated the same way, contributes $-\mathrm{i}G_F/\sqrt{2}\hbar c$ times the same spinor structure. Setting the two equal,

$$\frac{G_F}{\sqrt{2}\,\hbar c} = \frac{g^2\mu_0\hbar}{8M_W^2c},$$

which is the first form of Equation (101.78). For the second, substitute $g^2 = 4\pi\varepsilon_0\hbar c\,\alpha_W$ from Equation (101.71) and $\mu_0 = 1/\varepsilon_0c^2$:

$$\frac{G_F}{\sqrt{2}} = \frac{4\pi\varepsilon_0\hbar c\,\alpha_W\hbar^2}{8M_W^2\varepsilon_0c^2} = \frac{\pi\alpha_W\hbar^3}{2M_W^2c},$$

and dividing by $(\hbar c)^3$ gives $\pi\alpha_W/2\,(M_Wc^2)^2$. The dimensional check is worth doing once: $g^2\mu_0\hbar^2/M_W^2$ carries $\mathrm{C}^2 \times \mathrm{kg\,m/C}^2 \times \mathrm{J}^2\,\mathrm{s}^2/\mathrm{kg}^2 = \mathrm{J\,m}^3$, which is Equation (101.7). □

Remark 101.69 (The problem has moved, not gone). Equation (101.78) contains two unknowns where Fermi's theory had one, and it makes the dimensionful G_F into the ratio of a dimensionless coupling to a mass. That is progress only if the mass is itself explained, and *a massive vector field is not innocent*: the term $\frac{1}{2}M^2W_\mu W^\mu$ is not invariant under the gauge transformation Equation (100.8), and the $q_\mu q_\nu/M^2c^2$ in Equation (101.77) does not fall off at large q, which destroys the power counting of Theorem 100.17 exactly as G_F does. Inserting the boson by hand therefore buys a finite range and loses renormalizability. What buys both is spontaneous breaking of a gauge symmetry, and that is the subject of Chapter 104; 't Hooft's theorems [1448, 1463] are the statement that it works.

101.6.2 Gargamelle and the weak neutral current

The most immediate prediction of Proposition 101.67 is that a fourth current exists, coupling to the Z, which changes neither charge nor flavour. Nothing in Fermi's theory contains such a thing — the observed weak processes all involve a charge change — so its existence or absence was a decisive test.

Proposition 101.70 (The effective neutral-current interaction)**.** *At $|q^2| \ll M_Z^2 c^2$, and given the tree-level relation $M_W = M_Z \cos\theta_W$ of Phenomenon 104.35, the neutral-current interaction of a neutrino with a fermion f is*

$$\mathcal{L}_{\mathrm{NC}} = -\frac{G_F}{\sqrt{2}} \left[\bar{\psi}_\nu \gamma^\alpha \left(\mathbb{1} - \gamma^5\right) \psi_\nu\right] \left[\bar{\psi}_f \gamma_\alpha \left(g_V^f - g_A^f \gamma^5\right) \psi_f\right] , \tag{101.79}$$

with the same *G_F as the charged current and the couplings of Equation (101.75). The neutral-current strength is therefore not a new parameter; the only new number is $\sin^2\theta_W$.*

Rests on Proposition 101.68 and Equations (101.74) and (101.75).

Proof. *Derives Proposition 101.70.*

Repeat the matching of Proposition 101.68 with the Z vertex $-\mathrm{i}g\gamma^\mu(g_V - g_A\gamma^5)/2\hbar\cos\theta_W$ of Equation (101.74) and the Z propagator, which is Equation (101.77) with $M_W \to M_Z$. The effective coefficient is $(g/2\cos\theta_W)^2\,\mu_0\hbar^2/M_Z^2$ times the two current structures. For the neutrino, $T^3 = +\frac{1}{2}$ and $Q = 0$ give $g_V^\nu = g_A^\nu = \frac{1}{2}$ by Equation (101.75), so its factor $\frac{1}{2}(\mathbb{1} - \gamma^5)$ contributes an extra $\frac{1}{2}$. Hence the coefficient is

$$\frac{1}{2} \cdot \frac{g^2 \mu_0 \hbar^2}{4\cos^2\theta_W M_Z^2} = \frac{g^2\mu_0\hbar^2}{8M_W^2} = \frac{G_F}{\sqrt{2}} ,$$

using $M_Z^2\cos^2\theta_W = M_W^2$ and then Equation (101.78). Two independent factors have conspired: the Z is heavier than the W, which weakens the interaction, and it couples more strongly by $1/\cos\theta_W$, which strengthens it by exactly the same amount. That the two cancel is the content of the relation $\rho = 1$ of Equation (104.45), and it is a prediction of the doublet structure of the symmetry breaking rather than a fitted coincidence. □

Proposition 101.71 (Neutrino–electron elastic scattering)**.** *For $E_\nu \gg m_e c^2$, with the squared centre-of-mass energy $E_{\mathrm{cm}}^2 = 2m_e c^2 E_\nu$ for a neutrino of laboratory energy E_ν on an electron at rest,*

$$\sigma\left(\nu_\mu e^- \to \nu_\mu e^-\right) = \frac{G_F^2 E_{\mathrm{cm}}^2}{\pi\left(\hbar c\right)^4}\left(g_L^2 + \frac{g_R^2}{3}\right) , \qquad \sigma\left(\bar{\nu}_\mu e^- \to \bar{\nu}_\mu e^-\right) = \frac{G_F^2 E_{\mathrm{cm}}^2}{\pi\left(\hbar c\right)^4}\left(g_R^2 + \frac{g_L^2}{3}\right) , \tag{101.80}$$

with $g_L = \frac{1}{2}(g_V + g_A) = -\frac{1}{2} + \sin^2\theta_W$ and $g_R = \frac{1}{2}(g_V - g_A) = \sin^2\theta_W$ for the electron. The ratio to the charged-current process Equation (101.60) is $g_L^2 + g_R^2/3$, which for $\sin^2\theta_W = 0.231$ is 0.090.

Rests on Equations (101.60) and (101.79).

Proof. *Derives Proposition 101.71.*

Split the electron current into its chiral parts, $g_V - g_A\gamma^5 = g_L\left(\mathbb{1} - \gamma^5\right) + g_R\left(\mathbb{1} + \gamma^5\right)$, which is the definition of g_L and g_R. The g_L piece is Equation (101.46) with a different coefficient, so Equation (101.60) applies to it directly and gives an isotropic cross-section proportional to g_L^2.

The g_R piece requires one new ingredient, and it is angular momentum rather than algebra. In the centre-of-mass frame let the neutrino move along $+\hat{\boldsymbol{z}}$. It is left-handed, so $S_z^\nu = -\frac{1}{2}$. A left-chiral electron moving along $-\hat{\boldsymbol{z}}$ has its spin along $+\hat{\boldsymbol{z}}$, $S_z^e = +\frac{1}{2}$, so the pair has $J_z = 0$ and can be in a $J = 0$ state: the amplitude may be isotropic, and is. A right-chiral electron moving along $-\hat{\boldsymbol{z}}$ has $S_z^e = -\frac{1}{2}$, so $J_z = -1$ and the state must have $J \geq 1$; the

amplitude is then proportional to the rotation matrix element $d^{1}_{-1,-1}(\theta^{*}) = \tfrac{1}{2}\,(1+\cos\theta^{*})$, which vanishes in the backward direction because a backward scatter would reverse J_z without anything to absorb the two units. Hence

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\cos\theta^{*}} \propto g_L^2 + g_R^2 \left(\frac{1+\cos\theta^{*}}{2}\right)^2 , \tag{101.81}$$

and integrating over $\cos\theta^{*} \in [-1,1]$ gives $2g_L^2$ and $2g_R^2/3$ respectively, which is the factor $g_L^2 + g_R^2/3$ in Equation (101.80) once the normalization is fixed against Equation (101.60). For the antineutrino every helicity reverses, so g_L and g_R exchange roles. Numerically, at $\sin^2\theta_W = 0.231$, $g_L = -0.269$ and $g_R = 0.231$, giving 0.0901 for the neutrino and 0.0775 for the antineutrino, and

$$\sigma\left(\nu_\mu e^- \to \nu_\mu e^-\right) = 1.55{\times}10^{-46}\,\mathrm{m}^2 {\times} \frac{E_\nu}{1\,\mathrm{GeV}}\,, \qquad \sigma\left(\bar\nu_\mu e^- \to \bar\nu_\mu e^-\right) = 1.34{\times}10^{-46}\,\mathrm{m}^2 {\times} \frac{E_{\bar\nu}}{1\,\mathrm{GeV}}\,,$$

which is the size of the signal Gargamelle was looking for. Note that the g_V and g_A enter the two cross-sections in different combinations, so measuring both determines the sign of g_A as well as the value of $\sin^2\theta_W$ — a single measurement would not. □

Phenomenon 101.72 (Neutrinos scatter without becoming charged leptons)**.** A muon neutrino can transfer energy and momentum to matter and leave the apparatus still a neutrino. Both the leptonic channel $\bar\nu_\mu e^- \to \bar\nu_\mu e^-$ [1503] and the hadronic channel $\nu_\mu N \to \nu_\mu X$ [1502] were seen in the Gargamelle bubble chamber as events carrying hadrons or a recoil electron but no outgoing muon, at a rate a substantial fraction of the charged-current rate. Such a current changes neither electric charge nor flavour, and Fermi's theory contains nothing of the kind.

Rests on Propositions 101.70 and 101.71 and Phenomenon 101.5.

Derivation. *Derives Phenomenon 101.72.*

That the process exists at all is Proposition 101.70: a gauge theory containing the observed charged currents cannot avoid a neutral one, because the commutator of the two charged generators is a third generator. That it occurs at the observed *rate* is Proposition 101.71 applied to the leptonic channel Gargamelle actually saw, which was the *antineutrino* one: its predicted cross-section per unit beam energy is $1.34 \times 10^{-46}\,\mathrm{m}^2/\mathrm{GeV}$ at $\sin^2\theta_W = 0.231$, the neutrino channel being the larger of the two at $1.55 \times 10^{-46}\,\mathrm{m}^2/\mathrm{GeV}$. Gargamelle recorded a single such event in the whole exposure, which was the entire leptonic sample and which no background could produce, since an electron recoiling almost exactly along the beam direction with nothing else in the chamber has no other source [1503]. The hadronic channel supplied the statistics: some hundreds of muonless events, with a ratio to the charged-current rate of roughly one to five [1502].

The obstacle was background, and it is worth recording because the signal is defined by an absence. A neutron entering the chamber from the surrounding material produces hadrons and no muon, which is exactly the signature. The argument that settled it was the *spatial distribution*: neutrons are attenuated as they penetrate the liquid, so a neutron-induced sample is concentrated near the walls and falls off with depth, whereas a neutrino-induced sample is uniform because the chamber is transparent to neutrinos by Phenomenon 101.5. The observed distribution was flat. The same reasoning applied to the ratio of muonless to muonful events as a function of position gave a rate that could not be produced by neutrons at any plausible flux. □

Derivation pending #147

The hadronic neutral-current ratio: the prediction for the ratio of neutral-current to charged-current rates on an isoscalar nuclear target, which is what the Gargamelle hadronic sample actually measured. It requires the quark-parton content of the nucleon and the quark neutral-current couplings derived in this section, and therefore belongs with the deep-inelastic material rather than here; the leptonic ratio, which needs nothing beyond this chapter, is derived above.

101.6.3 Discovery of the W and Z

The bosons themselves were produced ten years later. The difficulty was energy: Equation (101.78) predicted masses near $80\,\mathrm{GeV}/c^2$ and $90\,\mathrm{GeV}/c^2$, and no machine reached the required centre-of-mass energy. Rubbia's proposal was to convert the CERN Super Proton Synchrotron into a proton — antiproton collider, which requires no second ring but does require enough antiprotons; van der Meer's stochastic cooling supplied them [1760], by sensing the deviation of a sample of the beam at one point in the ring and applying a correcting kick at another, thereby reducing the phase-space volume of a stored beam — an operation that looks at first sight like a violation of Liouville's theorem and is not, because the feedback system carries away the entropy.

Phenomenon 101.73 (The weak interaction is carried by heavy bosons)**.** Two massive vector particles are produced in $p\bar{p}$ collisions and observed as resonances: the W, as an isolated high-transverse-momentum electron recoiling against missing transverse energy [1710, 1714], and the Z, as a narrow peak in the invariant mass of charged lepton pairs [1711, 1713]. Their masses are

$$M_W c^2 = 80.37\,\mathrm{GeV}\,, \qquad M_Z c^2 = 91.1880\,\mathrm{GeV} \tag{101.82}$$

[2427]. Because these are enormous compared with the energy released in any nuclear decay, the exchange is indistinguishable from a contact interaction at every energy at which beta decay was ever studied.

Rests on Equation (101.78) and Propositions 100.63 and 101.71.

Derivation. *Derives Phenomenon 101.73.*

The masses were predicted before they were measured, from two quantities determined in entirely different experiments. Inverting Equation (101.78),

$$\left(M_W c^2\right)^2 = \frac{\pi\alpha}{\sqrt{2}\sin^2\theta_W}\cdot\frac{(\hbar c)^3}{G_F}\,, \tag{101.83}$$

and inserting $G_F/(\hbar c)^3 = 1.166\,378\,7\times10^{-5}/\mathrm{GeV}^2$ from the muon lifetime, $\alpha^{-1} = 127.95$ at this scale from Proposition 100.63, and $\sin^2\theta_W = 0.2312$ from the neutral-current measurements of Proposition 101.71 and Phenomenon 104.55,

$$\left(M_W c^2\right)^2 = \frac{\pi\,(1/127.95)}{\sqrt{2}\,(0.2312)\,(1.166\,378\,7\times10^{-5})}\,\mathrm{GeV}^2 = 6438\,\mathrm{GeV}^2\,,$$

that is $M_W c^2 = 80.2\,\mathrm{GeV}$, and $M_Z = M_W/\cos\theta_W$ gives $M_Z c^2 = 91.5\,\mathrm{GeV}$. The measured values are Equation (101.82): the tree-level prediction is right to a few parts in a thousand, the residual being the electroweak radiative correction that Section 101.6.4 turns into a measurement in its own right.

Read the other way, the same relation shows why Fermi's contact theory worked so well for so long. Writing $Q^2 = -q^2c^2$ for the squared four-momentum transfer measured as a

squared energy, the exchange contributes the propagator factor $\left[Q^2 + (M_W c^2)^2\right]^{-1}$, and in a nuclear decay Q is a few MeV, so the neglected ratio $Q^2/(M_W c^2)^2$ is of order 10^{-9}. The corresponding range is $\hbar/M_W c = 2.5 \times 10^{-18}\,\mathrm{m}$, a thousandth of the diameter of a proton, which is why no experiment before 1983 had resolved the exchange. Where it must fail is at Q approaching $M_W c^2$, which is the unitarity problem of Section 101.7.1. □

Remark 101.74 (Why the W has no mass peak and the Z does). The difference is kinematic. The Z decays to e^+e^- or $\mu^+\mu^-$, both measured, so the invariant mass of the pair is formed directly and the Z is a peak in it [1711, 1713]. The W decays to a charged lepton and a neutrino, which is not measured, and at a hadron collider the longitudinal balance is unusable because the colliding partons carry unknown longitudinal momenta and most of the debris escapes down the beam pipe. The transverse balance is usable, the initial transverse momentum being zero: the vector sum of the observed transverse momenta, negated, is the neutrino's. The variable that then carries the mass is the transverse mass

$$M_{\mathrm{T}}^2 c^4 = 2 p_{\mathrm{T}}^{\ell} c\, p_{\mathrm{T}}^{\nu} c\,(1 - \cos\Delta\varphi)\ , \tag{101.84}$$

with $\Delta\varphi$ the azimuthal angle between the two. It satisfies $M_{\mathrm{T}} \le M_W$, with equality for a transverse decay, so its distribution has a sharp edge at M_W — a Jacobian peak, from the vanishing of $\mathrm{d}M_{\mathrm{T}}/\mathrm{d}\theta^*$ there — and equivalently the lepton's own transverse-momentum spectrum has an edge at $M_W c/2$. UA1 and UA2 report exactly this: isolated electrons of $p_{\mathrm{T}} \approx 40\,\mathrm{GeV}/c$ against an equal and opposite missing transverse energy [1710, 1714], which is still, forty years later, the operational signature of a neutrino and of anything else that leaves a detector without interacting.

101.6.4 Precision electroweak measurements

Producing the Z in e^+e^- annihilation, where the initial state is known exactly, converts it from a discovery into a metrological instrument. LEP ran at and around the resonance from 1989 to 1995 and recorded some seventeen million Z decays; the combined analysis of the four experiments [2104] fixes M_Z to two parts in 10^5 and every partial width to a fraction of a per cent. One of those measurements counts particles that were never detected.

Proposition 101.75 (Ratios of Z partial widths)**.** *For final-state fermions light compared with the Z,*

$$\frac{\Gamma\left(Z \to f\bar{f}\right)}{\Gamma\left(Z \to f'\bar{f}'\right)} = \frac{N_c^f \left[\left(g_V^f\right)^2 + \left(g_A^f\right)^2\right]}{N_c^{f'} \left[\left(g_V^{f'}\right)^2 + \left(g_A^{f'}\right)^2\right]}\ , \tag{101.85}$$

with $N_c = 1$ for leptons and 3 for quarks. In absolute terms

$$\Gamma\left(Z \to f\bar{f}\right) = \frac{N_c^f}{6\sqrt{2}\pi\hbar}\,\frac{G_F}{(\hbar c)^3}\left(M_Z c^2\right)^3 \left[\left(g_V^f\right)^2 + \left(g_A^f\right)^2\right]\ . \tag{101.86}$$

Rests on Equations (101.48) and (101.74) and Lemma 101.47.

Proof. *Derives Proposition 101.75.*

By Equation (101.48) the width is $(c^2/2M_Z c^2)\int \overline{|\mathcal{M}|^2}\,\mathrm{d}\Phi_2$, and for massless final states $\int \mathrm{d}\Phi_2 = 1/8\pi\hbar^2$ by Lemma 101.47. The squared amplitude, summed over final spins and averaged over the three Z polarizations, involves the trace

$$T^{\mu\nu} = \mathrm{tr}\left[\gamma\cdot p'\,\gamma^\mu\left(g_V - g_A\gamma^5\right)\gamma\cdot p\,\gamma^\nu\left(g_V - g_A\gamma^5\right)\right]\ ,$$

contracted with the polarization sum $\sum_\lambda \varepsilon_\mu^*\varepsilon_\nu = -\eta_{\mu\nu} + q_\mu q_\nu/M_Z^2 c^2$. Moving the first factor $(g_V - g_A\gamma^5)$ rightwards through $\gamma\cdot p$ and then through γ^ν — each passage flipping the sign

of the γ^5 term — collects the two into $(g_V - g_A\gamma^5)^2 = g_V^2 + g_A^2 - 2g_V g_A\gamma^5$. The γ^5 piece produces a Levi-Civita tensor, antisymmetric in $\mu\nu$, which vanishes against the symmetric polarization sum. Hence the entire dependence on the fermion species is the factor $g_V^2 + g_A^2$, together with the colour multiplicity N_c, and Equation (101.85) follows without evaluating anything else. Carrying the remaining common factor through, with the coupling $g/2\cos\theta_W$ of Equation (101.74) re-expressed through Equation (101.78) and $M_W = M_Z\cos\theta_W$, gives Equation (101.86). As a check, for a neutrino $g_V = g_A = \frac{1}{2}$ and Equation (101.86) gives, as an energy width,

$$\hbar\Gamma_{\nu\bar{\nu}} = \frac{(1.166\,378\,7\times10^{-5})\,(91.1880)^3\,(0.5)}{6\sqrt{2}\pi}\,\text{GeV} = 166\,\text{MeV}\,,$$

against the measured 167.2 MeV [2104]: a tree-level formula good to one per cent, the difference being the radiative corrections. Widths are quoted throughout as energies, $\hbar\Gamma$, which is the convention of the source; the rate itself is Γ and carries 1/s. □

Phenomenon 101.76 (There are three light neutrino species)**.** The Z resonance is wider than the sum of its visible decay channels, and the invisible remainder counts the neutrino species light enough for the Z to decay into. The first scan already gave three to within a few per cent [1819]; the final four-experiment combination gives

$$N_\nu = 2.984 \pm 0.008 \tag{101.87}$$

[2104]. A fourth generation with a neutrino lighter than $M_Z c^2/2$ is excluded at more than fifteen standard deviations.

Rests on Equations (101.75) *and* (101.85).

Derivation. *Derives Phenomenon 101.76.*

Measure the total width Γ_Z from the shape of the resonance in $e^+e^- \to$ anything visible, the hadronic width from the hadronic cross-section, and the leptonic width from the lepton-pair cross-sections. The invisible width is the difference, $\Gamma_{\text{inv}} = \Gamma_Z - \Gamma_{\text{had}} - 3\Gamma_{\ell\ell}$. Then

$$N_\nu = \frac{\Gamma_{\text{inv}}}{\Gamma_{\ell\ell}}\left(\frac{\Gamma_{\ell\ell}}{\Gamma_{\nu\bar{\nu}}}\right)_{\text{SM}}\,, \tag{101.88}$$

which is arranged so that the absolute normalization of Equation (101.86) cancels and only the ratio Equation (101.85) is needed. That ratio is pure algebra: for the neutrino $g_V = g_A = \frac{1}{2}$ gives $g_V^2 + g_A^2 = \frac{1}{2}$, while for a charged lepton $T^3 = -\frac{1}{2}$ and $Q = -1$ give, by Equation (101.75), $g_A = -\frac{1}{2}$ and $g_V = -\frac{1}{2} + 2\sin^2\theta_W = -0.0376$ at $\sin^2\theta_W = 0.23122$, so $g_V^2 + g_A^2 = 0.2514$ and

$$\left(\frac{\Gamma_{\nu\bar{\nu}}}{\Gamma_{\ell\ell}}\right)_{\text{SM}} = \frac{0.5}{0.2514} = 1.989\,.$$

The measured ratio is $\Gamma_{\text{inv}}/\Gamma_{\ell\ell} = 5.943\pm0.016$ [2104], whence $N_\nu = 5.943/1.989 = 2.988$, and the full treatment with radiative corrections gives Equation (101.87). The near-cancellation of g_V for the charged lepton — $-\frac{1}{2} + 2(0.231)$ is small because $\sin^2\theta_W$ happens to lie near $\frac{1}{4}$ — is why $\Gamma_{\ell\ell}$ is almost exactly half $\Gamma_{\nu\bar{\nu}}$, and it is a numerical accident with no significance beyond making the arithmetic memorable.

Two caveats belong with the number. It counts only neutrinos with $m_\nu < M_Z/2$ that couple to the Z with standard strength: a heavier or a sterile neutrino is invisible to this measurement, and Chapter 103 records what other evidence says about those. And it is a *counting* measurement whose result is not an integer, 2.984 ± 0.008; the deficit below three is about two standard deviations and is understood as a residual effect in the Bhabha-scattering normalization, which is exactly the kind of statement Chapter 11 is needed to assess. □

The rest of the LEP programme, and the SLC programme beside it, turned the same resonance into a test of the theory's quantum corrections. The weak mixing angle is measured in half a dozen mutually independent ways — from the forward–backward asymmetries of the Z decays, from the left–right asymmetry with a polarized electron beam, from neutrino–nucleon scattering, from parity violation in atomic caesium [1519], and from polarized electron–deuteron scattering [1625], whose apparatus is Chapter 110 — at momentum transfers spanning six orders of magnitude, and one running, scheme-defined angle describes them all [2427]. That consistency, and not any single number, is the measurement. One entry does not fit — a measurement of the W mass sitting several standard deviations from the rest of the electroweak fit — and it is recorded here, because it is unresolved and this is the chapter that owns the quantity. Chapter 133 is not the place to look for it: that chapter collects the observations with no theoretical explanation at all — the flavour hierarchy, the strong CP problem, dark matter, dark energy, the baryon asymmetry — and does not treat the electroweak fit.

101.7 Why Fermi theory had to be replaced

101.7.1 Unitarity violation

Proposition 101.7 said that G_F carries $\mathrm{J\,m^3}$, so that the dimensionless number governing a weak amplitude at energy E is $G_F E^2/(\hbar c)^3$. A dimensionless expansion parameter that grows with energy is a theory with an expiry date, and the date can be computed.

Phenomenon 101.77 (The contact cross-section grows without bound)**.** The measured cross-section for neutrino scattering rises linearly with the beam energy from 1 MeV to several hundred GeV, as Equation (101.60) requires. The linear rise is a statement about a range and not about all energies, and the range is fixed by this chapter's own formulae: Equation (101.60) is the $M_W \to \infty$ limit, and restoring the W propagator of Equation (101.77) damps the growth once the momentum transfer approaches $M_W c$. For a nucleon target $E_{\mathrm{cm}}^2 = 2m_N c^2 E_\nu$, so the linear regime is $E_\nu \ll (M_W c^2)^2/2m_N c^2 = 3.4\,\mathrm{TeV}$; already at the top of the accelerator range, $E_\nu = 350\,\mathrm{GeV}$, the ratio $E_{\mathrm{cm}}^2/(M_W c^2)^2$ has reached 0.10 and the neglect is no longer safe. The departure is itself observed, in the momentum-transfer dependence of deep-inelastic neutrino scattering (Chapter 116).

Rests on Equations (101.60) *and* (101.77) *and Proposition 101.54.*

Derivation. *Derives Phenomenon 101.77.*

This is Proposition 101.54: the cross-section is $G_F^2 E_{\mathrm{cm}}^2/\pi(\hbar c)^4$, and $E_{\mathrm{cm}}^2 = 2m_{\mathrm{target}} c^2 E_{\mathrm{beam}}$ for a fixed target, so σ is proportional to the beam energy. The growth is not a defect of the calculation but a consequence of the dimension of the coupling: with only G_F, $\hbar$, c and E_{cm} available, and σ an area, the answer must be $G_F^2 E_{\mathrm{cm}}^2/(\hbar c)^4$ times a pure number. Restoring the propagator replaces this by $\sigma \propto G_F^2 E_{\mathrm{cm}}^2/\left[1 + E_{\mathrm{cm}}^2/(M_W c^2)^2\right]$, which stops rising once E_{cm} reaches $M_W c^2$ and tends to the constant $G_F^2 (M_W c^2)^2$ instead; the turnover is the direct observation of the propagator, and hence of the W in a scattering experiment rather than as a resonance. □

Theorem 101.78 (The contact theory violates unitarity)**.** *Probability conservation bounds a cross-section proceeding through a single partial wave J by*

$$\sigma_J \le \frac{4\pi\,(2J+1)}{k^2}\,, \qquad k = \frac{|\boldsymbol{p}|}{\hbar} = \frac{E_{\mathrm{cm}}}{2\hbar c} \quad \textit{(massless)}\,, \tag{101.89}$$

that is $\sigma_{J=0} \le 16\pi(\hbar c)^2/E_{\mathrm{cm}}^2$. Equation (101.60) *is pure $J = 0$ and grows as E_{cm}^2, so it breaches the bound at*

$$E_{\mathrm{cm}} = \left(16\pi^2 \frac{(\hbar c)^6}{G_F^2}\right)^{1/4} = 1.04\,\mathrm{TeV}\,, \tag{101.90}$$

and the theory is inconsistent above it.
Rests on Equation (101.60) and Proposition 101.54.

Proof. *Derives Theorem 101.78.*

The bound is the statement that the S-matrix is unitary. Writing the elastic amplitude in partial waves, $f(\theta) = k^{-1}\sum_J (2J+1)a_J P_J(\cos\theta)$, unitarity of $S_J = 1 + 2ia_J$ forces $|a_J| \le 1$, whence $\sigma_J = 4\pi(2J+1)\,|a_J|^2/k^2 \le 4\pi(2J+1)/k^2$; the partial-wave apparatus is that of Chapter 86. That the contact interaction produces only $J=0$ is Proposition 101.54: the cross-section came out isotropic, and an isotropic amplitude is by definition pure s-wave. Setting $G_F^2 E_{\mathrm{cm}}^2/\pi(\hbar c)^4 = 16\pi(\hbar c)^2/E_{\mathrm{cm}}^2$ and solving gives Equation (101.90); with $G_F/(\hbar c)^3 = 1.166\,378\,7\times10^{-5}/\mathrm{GeV}^2$, $E_{\mathrm{cm}}^4 = 16\pi^2/(1.166\,378\,7\times10^{-5})^2\mathrm{GeV}^4$ and $E_{\mathrm{cm}} = 1038\,\mathrm{GeV}$. □

Remark 101.79 (Which number to quote, and why they differ). The scale at which Fermi theory fails is quoted in the literature as anything between 300 GeV and 1 TeV, and the spread is not carelessness: the two ends of the range are different statements.

- Equation (101.90), about 1 TeV, is the energy at which *this particular channel* saturates the $J=0$ bound exactly. It depends on the channel through the spin and colour counting and on the convention for the bound, and moving either shifts it by tens of per cent.
- Equation (101.9), $E_F = \left((\hbar c)^3/G_F\right)^{1/2} = 292.8\,\mathrm{GeV}$, is the energy at which the dimensionless expansion parameter $G_F E^2/(\hbar c)^3$ reaches unity. That is where the perturbative expansion stops converging, which happens before the leading term saturates the bound; it is channel-independent, and it is the number the phrase "a few hundred GeV" refers to.

Both say the same physics: the theory predicts its own replacement, and predicts it at a definite scale. What is remarkable is that the prediction was right — the W was found at $80.4\,\mathrm{GeV}/c^2$, within a factor of a few of E_F, and the Higgs boson at $125\,\mathrm{GeV}/c^2$ (Chapter 113) — and the argument is now the standard template for looking beyond the Standard Model, as Section 101.7.2 records.

Remark 101.80 (The cure, and its cost). Replacing the contact vertex by W exchange cures the growth: by Equation (101.78) the amplitude acquires the factor $(M_W c^2)^2 / \left[Q^2 + (M_W c^2)^2\right]$, so the cross-section stops rising at $Q \sim M_W c^2$ and falls thereafter, and Equation (101.89) is never approached. It does not, however, cure everything: the process $\nu\bar\nu \to W^+W^-$, with longitudinally polarized Ws whose polarization vectors grow as $q^\mu/M_W c$, has an amplitude that again rises with energy, and the cancellation of *that* growth requires the scalar of Chapter 104 and fixes the shape of its couplings. The unitarity argument therefore does not stop at the W; it is what made a Higgs-like state mandatory below about 1 TeV, which is why the LHC was built to that energy and what it found is Chapter 113.

101.7.2 Non-renormalizability and the effective-theory reading

The second failure is not about any particular energy but about the perturbation series itself, and it was diagnosed by Dyson's power counting [928] — the same analysis that established QED to be renormalizable, applied to a different vertex, giving the opposite answer.

Theorem 101.81 (Superficial degree of divergence of the contact theory). *For a connected diagram built from V four-fermion vertices Equation (101.6) with E external fermion lines, the superficial degree of divergence is*

$$D = 4 - \frac{3}{2}E + 2V\,. \tag{101.91}$$

It grows *with the order of perturbation theory, so infinitely many amplitudes diverge and infinitely many counterterms are needed.*

Rests on Equations (100.23), (100.26) *and* (101.6).

Proof. *Derives Theorem 101.81.*

Let L be the number of loops and P the number of internal fermion lines. Each loop contributes $\mathrm{d}^4\ell$ and each internal fermion propagator falls as ℓ^{-1} by Equation (100.23), so $D = 4L - P$. Every vertex has four fermion ends, every internal line uses two and every external line one, giving $4V = 2P + E$; and the number of independent loops is $L = P - V + 1$. Eliminating,

$$D = 4\,(P - V + 1) - P = 3P - 4V + 4 = 3\left(2V - \frac{E}{2}\right) - 4V + 4 = 2V + 4 - \frac{3}{2}E\,,$$

which is Equation (101.91).

The contrast with Equation (100.26) is the whole point. There, $D = 4 - \frac{3}{2}E_e - E_\gamma$ and V cancelled: the degree of divergence is a property of the *amplitude*, not of the order, so only finitely many amplitudes diverge and finitely many parameters absorb them, which is Definition 100.18. Here V does not cancel, and it enters with a positive sign. The four-point function has $E = 4$ and $D = 2V - 2$: convergent at $V = 1$, quadratically divergent at $V = 2$, quartically at $V = 3$, and so on without end. Each new order requires a new counterterm of a new operator structure, and each new counterterm brings a new undetermined constant, so the theory has no predictive power beyond the order at which it is first fitted. □

Corollary 101.82 (The diagnosis is dimensional)**.** *The extra* $2V$ *in Equation* (101.91) *is exactly the* -2 *mass dimension of the coupling, counted once per vertex. A coupling of negative mass dimension is non-renormalizable and a coupling of non-negative mass dimension is not, and no calculation beyond dimensional analysis is required to tell which.*

Rests on Equation (101.91) *and Proposition 101.7.*

Proof. *Derives Corollary 101.82.*

An amplitude of fixed external content has a fixed overall dimension. If a vertex carries a coupling of mass dimension $-\delta$, then V vertices carry $-V\delta$, and the loop integrals must supply $+V\delta$ more powers of momentum to compensate, which is exactly $D \to D + V\delta$. With $\delta = 2$ for G_F this is the $2V$ of Equation (101.91). Conversely the QED coupling $e/\hbar$ has $\delta = 0$ once the $\hbar$'s are counted — which is the content of α being dimensionless, Equation (100.3) — and no such term appears. □

Remark 101.83 (The modern reading: Fermi's theory is not wrong, it is effective)*.* The verdict of Theorem 101.81 was, for thirty years, that Fermi's theory was a stopgap. The modern reading is different, and it is the frame in which every current search for physics beyond the Standard Model is conducted.

Equation (101.6) is the leading term of a systematic expansion. Start from the theory with the W and *integrate out* the heavy field — expand the propagator Equation (101.77) in powers of $q^2/M_W^2c^2$:

$$\frac{1}{q^2 - M_W^2c^2} = -\frac{1}{M_W^2c^2}\left(1 + \frac{q^2}{M_W^2c^2} + \frac{q^4}{M_W^4c^4} + \cdots\right).$$

The leading term is the contact interaction with the coefficient Equation (101.78); the next is a dimension-eight operator with two extra derivatives and two further powers of M_W in the denominator; and so on. The expansion is in E/M_Wc^2, it is systematic, and its accuracy at any energy can be estimated in advance — which is why Fermi's theory reproduces nuclear beta decay with the margin Phenomenon 101.73 computed.

Non-renormalizability is then not a disease but a *measurement*. An operator whose coefficient has mass dimension -2 announces a scale, and that scale is where the description stops: Fermi's coupling announced Equation (101.9), 292.8 GeV, and the W was found at $80.4\,\mathrm{GeV}/c^2$. How such operators mix under a change of scale, and why the leading ones dominate, is Chapter 108.

The same logic now runs in the other direction: the Standard Model is treated as the leading term, and searches are reported as bounds on the coefficients of the higher-dimension operators that could be added to it, equivalently as bounds on a scale Λ. That is how a null result at a collider becomes a quantitative statement, and it is why the failure recorded in this section is now regarded as the most useful thing Fermi's theory ever did.

Remark 101.84 (What this chapter has and has not established). The chapter began with a continuous spectrum and ends with two massive bosons. Everything between was extracted from measurement: the neutrino from kinematics (Phenomenon 101.1), the contact form and the size of the coupling from spectra and lifetimes (Proposition 101.11 and Phenomenon 101.50), the two nuclear couplings from the observed selection rules (Proposition 101.15), the violation of parity from one asymmetry (Phenomenon 101.26), the $V-A$ structure from the size of that asymmetry together with the neutrino helicity (Proposition 101.36 and Phenomenon 101.28), the quark rotation from a deficit in rate (Phenomenon 101.55), a fourth quark from a null result (Theorem 101.59), a third generation from a parameter count (Proposition 101.63), and the neutral current from events with a missing muon (Phenomenon 101.72).

What has *not* been established is why any of it is so. The gauge group was quoted, not derived; the boson masses were matched to G_F, not explained; the mixing angles and the fermion masses are inputs. Those are the business of Chapters 103 and 104 — and, in the case of the mixing parameters, of nobody at present: they are measured numbers with no known origin, recorded as such in Chapter 133 and in the free-parameter count of the appendices.

Chapter 102

Quantum Chromodynamics

Quantum chromodynamics is the gauge theory of colour, and it is the one part of the Standard Model whose defining property — that its elementary fields have never been seen in isolation — makes the evidence for it indirect at every step. This chapter therefore follows the evidence rather than the formalism: the regularities of the hadron spectrum that produced the quark hypothesis, the spin-statistics paradox of the Δ^{++} that forced a threefold colour degree of freedom, the deep inelastic scattering that showed pointlike constituents inside the proton, the three-jet events that made the gluon visible, and the lattice computations that now reproduce the hadron masses from the Lagrangian alone. Only then is the failure to observe a free quark discussed for what it is — a strong empirical regularity supported by a well-tested numerical demonstration, not a theorem.

QCD is the second gauge theory of the part and reuses the entire apparatus of Chapter 100, but two features have no QED analogue and organize everything below. The gluons carry the charge they mediate, so the beta function changes sign: the coupling is weak at short distance — asymptotic freedom, which is why perturbation theory works at all here — and strong at hadronic scales, where it does not, and Chapter 106 together with lattice methods must take over. And the near-masslessness of the lightest quarks gives an approximate chiral symmetry whose spontaneous breaking, not the Higgs mechanism of Chapter 104, supplies about 99 percent of the mass of ordinary matter. Its experimental anchor is Chapter 116; its downstream users are Chapter 107, Chapter 52 and Chapter 49. Standard treatments are [1739, 1933]; data throughout are from [2427].

Units and conventions are those fixed once for the whole part in Section 100.1.1: the metric $\eta_{\mu\nu} = \mathrm{diag}(+1, -1, -1, -1)$ of Notation 100.1, the mass shell $p^2 = m^2c^2$ of Equation (100.1), the action Equation (100.2) carrying an explicit c^{-1}, and $\hbar$ and c written out in every formula. Nothing below is computed in units where they are set to one; Remark 100.4 gives the dictionary for reading this chapter against the literature, which does.

102.1 Hadron spectroscopy and the quark hypothesis

102.1.1 The eightfold way

Between 1947 and 1960 the list of strongly interacting particles grew from a handful to several dozen. The cosmic-ray and accelerator experiments of the period produced the pion, the kaons, the Λ, the Σ and Ξ hyperons and, once the bubble chamber allowed short-lived states to be reconstructed from their decay products, a succession of resonances — the $\Delta(1232)$, the $\Sigma(1385)$, the $\Xi(1530)$, the ρ, the K^*. None of them could be called elementary with a straight face. The response was not to propose constituents but to look for a symmetry, and the symmetry that worked is the one whose group theory is established, as mathematics and with no physics attached, in Section 14.2.2.

Two additive quantum numbers organize the data. The first is *isospin*: the near-equality of the proton and neutron masses, and of the three pion masses, is described by an SU(2) under which (p, n) is a doublet, and its representation theory is that of Chapter 5. The second is *strangeness* S, introduced by Gell-Mann [1017] and independently by Nakano and Nishijima [1024] to account for the fact that the new particles are produced copiously (strong interaction) but decay slowly (weak interaction): S is conserved by the strong interaction and violated by the weak one, so a strange particle can only be produced in pairs and must then wait for a weak decay. Writing the *hypercharge* $Y = B + S$ with B the baryon number, the observed electric charges of every hadron obey

$$\frac{Q}{e} = I_3 + \frac{Y}{2}\,, \tag{102.1}$$

with I_3 the third isospin component. Equation (102.1) is a fact about hundreds of measured charges, not a definition.

The content of the eightfold way, proposed independently by Gell-Mann [1205] and Ne'eman [1216], is that I_3 and Y are the two commuting quantum numbers of a rank-two group, and that the group is SU(3). Proposition 14.50 gives the rank as $\ell = 2$ and the dimension as 8; the two Cartan generators of Definition 14.51 are T_3 and T_8, and the identification with the measured quantum numbers is

$$I_3 = T_3\,, \qquad Y = \frac{2}{\sqrt{3}}\,T_8\,, \tag{102.2}$$

the factor $2/\sqrt{3}$ being fixed once and for all by Equation (14.65). A hadron multiplet is then an irreducible representation, and the weight diagram of that representation — the set of simultaneous eigenvalues of (T_3, T_8), computed as pure mathematics in Equation (14.79) and Equation (14.80) — is a prediction of which charges and strangenesses occur together and how many states there are.

The prediction is sharp and it is met. The eight lightest spin-$\frac{1}{2}$ baryons (p, n, Λ, $\Sigma^{\pm,0}$, $\Xi^{0,-}$) fill the weight diagram of the adjoint representation **8** exactly: six states on a hexagon and two at the centre, which is the count $6 + 2 = 8$ of Remark 14.56. The eight lightest pseudoscalar mesons ($\pi^{\pm,0}$, $K^{\pm}$, K^0, $\bar{K}^0$, η) fill a second copy of the same diagram. The nine lightest spin-$\frac{3}{2}$ baryons known in 1961 fill nine of the ten places of a decuplet, and the tenth place is the subject of Section 102.1.2.

Phenomenon 102.1 (Hadron masses obey an octet mass formula)**.** Within a multiplet the masses are not equal — SU(3) is broken by the strange quark mass — but the breaking is not arbitrary. For the $J^P = \frac{1}{2}^+$ baryon octet the measured masses satisfy the Gell-Mann–Okubo relation [1205, 1240]

$$2\,(M_N + M_\Xi) = 3M_\Lambda + M_\Sigma \tag{102.3}$$

to better than one percent, and for the $J^P = \frac{3}{2}^+$ decuplet the masses are equally spaced in hypercharge to within a few percent.

Rests on Theorem 5.153 and Equations (14.69) *and* (14.70).

Derivation. *Derives Phenomenon 102.1.*

Write the strong Hamiltonian as $H = H_0 + H_1$, with H_0 SU(3)-symmetric and H_1 the symmetry-breaking piece. The breaking comes from the quark mass term of Section 102.3.3, which for $m_u = m_d \neq m_s$ splits into a piece proportional to the identity — absorbed into H_0 — and a piece proportional to λ_8 of Equation (14.65). So H_1 transforms as the eighth component of an octet. Treat it in first-order perturbation theory (Chapter 82), so that the mass shift of a state $|B\rangle$ is $\langle B|H_1|B\rangle$.

By the Wigner–Eckart theorem for SU(3) — the same statement as for SU(2), and proved in the same way from Schur's lemma Theorem 5.153 — the diagonal matrix elements

of an octet operator inside an octet are determined by two reduced matrix elements, because $\mathbf{8} \otimes \mathbf{8}$ contains $\mathbf{8}$ twice (once symmetrically, through d_{abc} of Equation (14.70), and once antisymmetrically, through f_{abc} of Equation (14.69)). Two reduced matrix elements mean two constants, and the only SU(3) scalars that can be built from T_8 and the state's own quantum numbers, to first order, are Y and the isospin Casimir $I(I+1)$ corrected by the trace term $Y^2/4$. Hence

$$M = M_0 + aY + b\left[I(I+1) - \frac{Y^2}{4}\right], \tag{102.4}$$

with M_0, a and b the same for the whole multiplet.

For the octet, (Y, I) is $(1, \frac{1}{2})$ for N, $(0, 0)$ for Λ, $(0, 1)$ for Σ and $(-1, \frac{1}{2})$ for Ξ, so that $I(I+1) - Y^2/4$ equals $\frac{1}{2}$, 0, 2, $\frac{1}{2}$ respectively, and

$$\begin{aligned} M_N &= M_0 + a + \tfrac{1}{2}b\,, & M_\Lambda &= M_0\,, \\ M_\Sigma &= M_0 + 2b\,, & M_\Xi &= M_0 - a + \tfrac{1}{2}b\,. \end{aligned} \tag{102.5}$$

The three unknowns are eliminated from the four masses in one way: $M_N + M_\Xi = 2M_0 + b$ and $3M_\Lambda + M_\Sigma = 4M_0 + 2b$, whence Equation (102.3). Numerically, with the isospin averages $M_N c^2 = 938.92\,\mathrm{MeV}$, $M_\Lambda c^2 = 1115.68\,\mathrm{MeV}$, $M_\Sigma c^2 = 1193.15\,\mathrm{MeV}$ and $M_\Xi c^2 = 1318.29\,\mathrm{MeV}$ [2427],

$$2\,(M_N + M_\Xi)\,c^2 = 4514.4\,\mathrm{MeV}\,, \qquad (3M_\Lambda + M_\Sigma)\,c^2 = 4540.2\,\mathrm{MeV}\,, \tag{102.6}$$

a discrepancy of 25.8 MeV, that is 0.57 % of either side, against a spread of nearly 400 MeV among the masses being related. For the decuplet the states satisfy $I = 1 + Y/2$, so that $I(I+1) - Y^2/4 = 2 + \frac{3}{2}Y$ and Equation (102.4) collapses to a linear function of Y: consecutive isospin multiplets are equally spaced in mass. That is what Section 102.1.2 exploits. □

Remark 102.2 (What the eightfold way did and did not assert). Neither [1205] nor [1216] proposed constituents. The claim was that an approximate SU(3) organizes the spectrum, and the evidence for it was the multiplet structure and Equation (102.3). That the observed multiplets are **1**, **8** and **10** — and never the **3** that generates them — was at the time an unexplained selection rule. It is explained in Section 102.2.1, and its explanation is colour.

102.1.2 The Ω^- prediction and its confirmation

By 1962 the decuplet contained the four Δ states ($Y = 1$, $I = \frac{3}{2}$), the three $\Sigma(1385)$ states ($Y = 0$, $I = 1$) and the two $\Xi(1530)$ states ($Y = -1$, $I = \frac{1}{2}$): nine of ten. The tenth place, $Y = -2$, $I = 0$, was empty. What Equation (102.4) supplies is not merely the existence of a tenth state but every one of its quantum numbers and, through the equal-spacing rule, its mass.

Phenomenon 102.3 (A particle predicted in full and then found)**.** The equal spacing of the decuplet requires a tenth baryon with strangeness $S = -3$, isospin $I = 0$, charge $-e$, spin-parity $\frac{3}{2}^+$ and a mass about 145 MeV above the $\Xi(1530)$, that is close to $1680\,\mathrm{MeV}/c^2$; being below the threshold for any strangeness-conserving strong decay it must decay weakly, and therefore live long enough to leave a macroscopic track. A single bubble-chamber event at Brookhaven displayed exactly such a particle, with mass $(1686 \pm 12)\,\mathrm{MeV}/c^2$ [1270]. The current world average is $(1672.45 \pm 0.29)\,\mathrm{MeV}/c^2$ [2427], within 0.5 % of the prediction.

Rests on Equations (102.1) *and* (102.4).

Derivation. *Derives Phenomenon 102.3.*

Apply Equation (102.4) to the decuplet, where it reduces to $M = \text{const} + \left(a + \frac{3}{2}b\right) Y$ as shown above. The measured spacings between consecutive isospin multiplets of the decuplet are, with modern isospin-averaged masses [2427],

$$M_{\Sigma^*} - M_{\Delta} = 153\,\text{MeV}/c^2\,, \quad M_{\Xi^*} - M_{\Sigma^*} = 149\,\text{MeV}/c^2\,, \quad M_{\Omega} - M_{\Xi^*} = 139\,\text{MeV}/c^2\,, \tag{102.7}$$

equal to within 5 %, which is the accuracy first-order perturbation theory in a breaking of this size deserves. With the masses available in 1962 the first two spacings were about $145\,\text{MeV}/c^2$ each, and extrapolating one further step placed the missing state near $1680\,\text{MeV}/c^2$.

The quantum numbers follow with no freedom at all. The tenth weight of the decuplet sits at $Y = -2$, $I = I_3 = 0$, so Equation (102.1) gives $Q = -e$, and $Y = B + S$ with $B = 1$ gives $S = -3$. The lifetime argument is kinematic: the lightest $S = -3$ combination reachable by a strong decay, which cannot change strangeness, is ΞK, whose threshold is $(1315 + 494)\,\text{MeV}/c^2 = 1809\,\text{MeV}/c^2$ [2427] — above the predicted mass. Every strong channel is therefore closed, the state must decay through the weak interaction of Chapter 101 with $\Delta S = 1$, and its lifetime must be of the order of $10^{-10}\,\text{s}$ rather than the $10^{-23}\,\text{s}$ of its nine partners. That is the difference between a resonance visible only as a bump in a mass distribution and a particle that travels centimetres and is photographed. □

Remark 102.4 (Why this counts for more than a fit). The eightfold way was fitted to the nine known decuplet members; the tenth was not in the data. What was published in advance was a mass, a charge, a strangeness, a spin, a decay mode and an order of magnitude for a lifetime, and all six were confirmed by one event [1270]. A scheme that reproduces n known numbers with n parameters says nothing; a scheme that predicts an unobserved object's entire specification and is then vindicated has been exposed to refutation and has survived. This is the asymmetry between accommodation and prediction argued in Chapter 1, and the Ω^- is the cleanest instance of it in hadron physics.

102.1.3 Quarks

The representations that occur in Nature are **1**, **8** and **10**. None of them is the fundamental **3**, but all of them are built from it: this is the observation of Gell-Mann [1277] and, in a more concretely constituent reading, of Zweig [1301], whose "aces" are the same objects.

Definition 102.5 (Quarks)**.** A quark is a spin-$\frac{1}{2}$ fermion of baryon number $B = \frac{1}{3}$ transforming in the fundamental representation **3** of flavour SU(3). Reading the weights Equation (14.79) through Equation (102.2) assigns to the three states (u, d, s) the quantum numbers

$$\begin{aligned} u:&\quad I_3 = +\tfrac{1}{2}\,, \quad Y = +\tfrac{1}{3}\,, \quad Q = +\tfrac{2}{3}e\,,\\ d:&\quad I_3 = -\tfrac{1}{2}\,, \quad Y = +\tfrac{1}{3}\,, \quad Q = -\tfrac{1}{3}e\,,\\ s:&\quad I_3 = 0\,, \quad\;\; Y = -\tfrac{2}{3}\,, \quad Q = -\tfrac{1}{3}e\,, \end{aligned} \tag{102.8}$$

the charges following from Equation (102.1) and being, for the first time in physics, fractions of the elementary charge. Antiquarks carry the conjugate representation $\bar{\mathbf{3}}$ and the opposite signs.

Proposition 102.6 (Mesons are $q\bar{q}$, baryons are qqq)**.** *The multiplets observed in Section 102.1.1 are exactly those obtained by combining quarks as*

$$\mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{1} \oplus \mathbf{8}\,, \qquad \mathbf{3} \otimes \mathbf{3} \otimes \mathbf{3} = \mathbf{1} \oplus \mathbf{8} \oplus \mathbf{8} \oplus \mathbf{10}\,. \tag{102.9}$$

Rests on Definition 102.5, Theorem 14.57, and Equation (14.81).

Proof. *Derives Proposition 102.6.*

The first decomposition is Theorem 14.57, proved there both by splitting $\mathrm{End}(\mathbb{C}^3)$ into its trace and traceless parts and by counting weights. For the second, tensor Equation (14.81) once more with **3**, or count directly: $\mathbb{C}^3 \otimes \mathbb{C}^3 \otimes \mathbb{C}^3$ has dimension 27 and decomposes under the symmetric group S_3, which commutes with the SU(3) action, into a totally symmetric part of dimension $\binom{5}{3} = 10$, a totally antisymmetric part of dimension $\binom{3}{3} = 1$, and two copies of a mixed-symmetry part of dimension 8 each; $10 + 1 + 8 + 8 = 27$. Each symmetry class is SU(3)-invariant, and the four pieces are the representations named. Baryon number is additive with $B = \frac{1}{3}$ per quark, so $q\bar{q}$ has $B = 0$ and qqq has $B = 1$, matching mesons and baryons respectively, and Equation (102.1) is reproduced because Q and Y are additive by construction. □

The decuplet is the totally symmetric combination, which is why it contains *uuu*, *ddd* and *sss* — the Δ^{++}, the Δ^{-} and the Ω^{-}. That fact is about to cause trouble.

Phenomenon 102.7 (The nucleon magnetic moments come out right)**.** Treating the nucleon as three quarks in a spatially symmetric ground state, with each quark contributing a Dirac moment inversely proportional to its mass, predicts

$$\frac{\mu_p}{\mu_n} = -\frac{3}{2}\,, \tag{102.10}$$

with no free parameter whatever. The measured ratio is $2.792847/(-1.913043) = -1.45990$ [2431], low by 2.7 %.

Rests on Definition 102.5, Proposition 102.6, and Equation (93.2).

Derivation. *Derives Phenomenon 102.7.*

A Dirac particle of charge $Q_q e$ and mass m_q, with Q_q the dimensionless charge in units of the elementary charge, has magnetic moment $\mu_q = Q_q e\hbar/(2m_q)$, of SI dimension J/T; this is the $g = 2$ result of the Dirac equation (Chapter 93), whose small radiative correction is the subject of Section 100.8.2 and is irrelevant at the accuracy in question here. Write $\hat{\mu} := e\hbar/(2m_q)$ for the common scale, with $m_u = m_d =: m_q$ by isospin symmetry, so that $\mu_u = +\frac{2}{3}\hat{\mu}$ and $\mu_d = -\frac{1}{3}\hat{\mu}$.

The proton is *uud* with total spin $\frac{1}{2}$ and no orbital angular momentum. The spin-flavour wavefunction that is symmetric under exchange of all three quarks — required by Section 102.2.1, and there is exactly one such state with $J = \frac{1}{2}$ — gives, on projecting the magnetic moment operator $\sum_i \mu_i \sigma_z^{(i)}$ onto it,

$$\mu_p = \frac{4\mu_u - \mu_d}{3}\,, \qquad \mu_n = \frac{4\mu_d - \mu_u}{3}\,. \tag{102.11}$$

The combinatorial factors 4 and -1 are the statement that in the symmetric $J = \frac{1}{2}$ state the like pair (uu in the proton) is coupled to spin 1 with weight $\frac{2}{3}$ and the odd quark carries the remaining spin with weight $-\frac{1}{3}$. Substituting, $\mu_p = \hat{\mu}$ and $\mu_n = -\frac{2}{3}\hat{\mu}$, whence Equation (102.10). The unknown m_q has cancelled, which is why the ratio is a prediction and the individual moments are not.

Feeding the measured $\mu_p = 2.792847\,\mu_N$ [2431], with the nuclear magneton $\mu_N = 5.050\,783\,739\,3 \times 10^{-27}\,\mathrm{J/T}$ [2431], back into $\mu_p = e\hbar/(2m_q)$ gives

$$m_q = \frac{m_p}{2.792847} = 5.99 \times 10^{-28}\,\mathrm{kg} = 336\,\mathrm{MeV}/c^2\,, \tag{102.12}$$

almost exactly one third of the nucleon mass. A model in which the nucleon is three objects of mass $m_N/3$ therefore reproduces both the moment ratio and the moment scale. That the number $336\,\mathrm{MeV}/c^2$ has nothing to do with the u and d masses in the Lagrangian, which are $2.16\,\mathrm{MeV}/c^2$ and $4.70\,\mathrm{MeV}/c^2$ [2427], is the first appearance in this chapter of the fact developed in Section 102.8.2: the mass of ordinary matter is not the mass of its quarks. □

Remark 102.8 (The status of the model in 1964). Gell-Mann's paper is explicit that the quarks might be a mathematical device: fractional charge had never been seen, and searches for it were already under way. Zweig, writing at CERN in the same year, took the constituents literally and could not get his paper published in a journal [1301]. The successes above — the multiplet structure, Equation (102.3), Equation (102.10) — are all statements about *symmetry*, and a symmetry can be realized without its generating representation being populated by particles. What converted the scheme into a theory of constituents was the deep inelastic scattering of Section 102.5, which saw the objects directly. The searches for free fractional charge, meanwhile, have never found anything: Section 102.7.3.

102.2 Colour

102.2.1 The statistics paradox

The quark model of Section 102.1.3 works, and it contains a contradiction with a theorem.

Phenomenon 102.9 (The Δ^{++} violates the spin-statistics theorem)**.** The $\Delta^{++}(1232)$ is the lightest baryon of charge $+2e$, with $J^P = \frac{3}{2}^+$, mass $(1232.0 \pm 2.0)\,\mathrm{MeV}/c^2$ and width $(117 \pm 3)\,\mathrm{MeV}$ [2427]. In the quark model its content is uuu: three identical fermions. Being the lightest state of its quantum numbers it has no orbital excitation, so its spatial wavefunction is symmetric; being $J = \frac{3}{2}$ with three spin-$\frac{1}{2}$ constituents, its spin wavefunction in the $J_z = +\frac{3}{2}$ state is $\uparrow\uparrow\uparrow$, also symmetric; and its flavour wavefunction uuu is symmetric. The total wavefunction is therefore symmetric under exchange of any two of three identical fermions, in direct contradiction with the statistics asserted by the spin-statistics theorem, whose proof belongs to the axiomatic chapter.

Rests on Definition 102.5 and Proposition 102.6.

The contradiction is not a technicality that a more careful model might avoid. It is repeated by the Δ^- (ddd) and the Ω^- (sss), and the same symmetric combination is what produced the successful magnetic moments Equation (102.11): the model's failure and its success come from the same wavefunction. Either the spin-statistics theorem fails — and it is a theorem, whose proof belongs to the axiomatic chapter — or a quark carries a quantum number that has not yet been written down.

Greenberg proposed the first escape, replacing Fermi statistics for quarks by *parastatistics* of order three, in which up to three identical quarks may occupy the same state [1278]. Han and Nambu proposed the second, and the one that survived: quarks come in three varieties distinguished by a new exactly conserved quantum number, on which acts a further, exact SU(3) that has nothing to do with the flavour SU(3) of Section 102.1.1 [1309]. The two proposals are related — parastatistics of order three is what an antisymmetrized threefold degeneracy looks like if one refuses to name the degeneracy — but only the second can be gauged, and gauging it is QCD.

Definition 102.10 (Colour)**.** Each quark flavour f carries an additional index $i = 1, 2, 3$ ("red, green, blue"), transforming in the fundamental representation $\mathbf{3}$ of a group $\mathrm{SU}(3)_c$, distinct from and commuting with flavour SU(3) and with the Poincare group. Antiquarks transform in $\bar{\mathbf{3}}$. The generators are the $T_a = \frac{1}{2}\lambda_a$ of Definition 14.51, with the structure constants Equation (14.69) and the normalization $\mathrm{tr}(T_a T_b) = \frac{1}{2}\delta_{ab}$ of Equation (14.66). Colour is dimensionless and is carried by no lepton.

With a colour index available, the Δ^{++} wavefunction is

$$\left|\Delta^{++}, J_z = \tfrac{3}{2}\right\rangle = \frac{1}{\sqrt{6}}\,\epsilon_{ijk}\,\left|u^i\uparrow\right\rangle\left|u^j\uparrow\right\rangle\left|u^k\uparrow\right\rangle\,, \tag{102.13}$$

totally antisymmetric because ϵ_{ijk} is, and the theorem is satisfied. Nothing has been bought cheaply: the same index must now explain why the states it permits are not seen.

Postulate 102.11 (The colour-singlet rule)**.** Every observable asymptotic state is a singlet of $\mathrm{SU}(3)_c$.

Theorem 102.12 (Which quark combinations can be colour singlets)**.** *A state built from m quarks and n antiquarks contains a colour singlet only if $m-n$ is divisible by three. In particular $q\bar{q}$ and qqq each contain exactly one singlet, while qq, $qqq\bar{q}$ and any single quark contain none.*

Rests on Definition 102.10, Theorem 5.153, and Proposition 14.55.

Proof. *Derives Theorem 102.12.*

The centre of SU(3) is the group of the three matrices $\omega^r \mathbb{1}$, $r = 0, 1, 2$, with $\omega = \mathrm{e}^{2\pi\mathrm{i}/3}$; these are unitary, unimodular ($\det(\omega\mathbb{1}) = \omega^3 = 1$) and commute with everything, and no other multiple of the identity is unimodular. On the fundamental $\mathbf{3}$ the element $\omega\mathbb{1}$ acts as multiplication by ω, and on $\bar{\mathbf{3}}$ as multiplication by ω^{-1}. On a tensor product of m copies of $\mathbf{3}$ and n of $\bar{\mathbf{3}}$ it therefore acts as ω^{m-n}, and on the trivial representation it must act as 1. Hence $\omega^{m-n} = 1$, i.e. $3 \mid m-n$. This number, $m-n \bmod 3$, is the *triality*, and it is a representation-theoretic invariant: no equivalence can change it, so the obstruction is absolute and not a matter of which coupling scheme is used.

For the multiplicities: the number of singlets in $V \otimes W$ is $\dim \mathrm{Hom}_G(V^*, W)$, which by Schur's lemma Theorem 5.153 is 1 if W is equivalent to V^* and 0 otherwise. Since $\bar{\mathbf{3}}$ is inequivalent to $\mathbf{3}$ (Proposition 14.55), $\mathbf{3} \otimes \mathbf{3}$ has no singlet, while $\mathbf{3} \otimes \bar{\mathbf{3}}$ has exactly one — as Theorem 14.57 exhibits explicitly, the trace part of a 3×3 matrix. For three quarks, the decomposition Equation (102.9) contains $\mathbf{1}$ once, and the invariant realizing it is ϵ_{ijk}: the unique (up to scale) totally antisymmetric invariant tensor of $\mathrm{SL}(3,\mathbb{C})$, hence of SU(3). □

Corollary 102.13 (The selection rule of the eightfold way)**.** *Combining Postulate 102.11 with Theorem 102.12 and Equation (102.9), the lightest observable hadrons are exactly the flavour multiplets $\mathbf{1} \oplus \mathbf{8}$ from $q\bar{q}$ and $\mathbf{1} \oplus \mathbf{8} \oplus \mathbf{8} \oplus \mathbf{10}$ from qqq: mesons and baryons, and nothing of triality one or two. This is the rule left unexplained in Remark 102.2.*

Rests on Postulate 102.11, Theorem 102.12, and Equation (102.9).

Remark 102.14 (The rule is a postulate here and a consequence later)*.* Postulate 102.11 is stated as a postulate because that is its honest status in this chapter: it is inferred from the observed spectrum and it is not derived from the Lagrangian of Section 102.3.3. The linear potential of Section 102.7.2 makes it plausible — separating a non-singlet pair costs an energy growing without bound — and the lattice computations of Section 102.7.1 exhibit it, but a proof from the continuum theory does not exist. Section 102.7 says so at length rather than disguising the gap.

102.2.2 Counting the colours

Colour was introduced to fix a statistics problem, and a quantum number introduced to fix one problem is worth little until it is measured somewhere else. It is measured in three independent places, all of which count *three*. The counts are independent in the strong sense: they involve different initial states, different interactions and different theoretical inputs, and each would fail by a large factor if N_c were 1.

Phenomenon 102.15 (Three colours, counted in an annihilation ratio)**.** The ratio

$$R := \frac{\sigma(e^+e^- \to \text{hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)} \tag{102.14}$$

is, away from resonances, nearly independent of energy over wide intervals and steps upward at each new quark threshold. The measured plateaux lie close to 2 below the charm threshold and close to 10/3 between the charm and bottom thresholds [2427]. Both values require three colour states per quark flavour: with one colour they would be 2/3 and 10/9.

Rests on Definition 102.10, Equation (100.32), and Theorem 100.58.

Derivation. *Derives Phenomenon 102.15.*

At leading order both processes proceed through a single virtual photon, so they differ only in the charge of the produced pair and in how many distinct pairs can be produced. The denominator is Equation (100.32), $\sigma(e^+e^- \to \mu^+\mu^-) = 4\pi\alpha^2(\hbar c)^2/3E_{\mathrm{cm}}^2$, computed in Section 100.3.1. For a pointlike spin-$\frac{1}{2}$ particle of electric charge $Q_f e$ — here and throughout, Q_f is the dimensionless charge of flavour f in units of e — created well above its threshold, the same computation applies with the single change $e^2 \to Q_f e^2$ at the production vertex, so the cross-section is that of Equation (100.32) times Q_f^2; and the rates for distinguishable final states add incoherently, so a quark flavour existing in N_c colours contributes N_c times, the electromagnetic vertex being blind to the colour label. Dividing by the muon channel, for which $Q = -1$ and there is one internal state,

$$R = N_c \sum_q Q_q^2 \,, \tag{102.15}$$

the sum running over the flavours light enough to be produced. With $Q_u = Q_c = +\frac{2}{3}$ and $Q_d = Q_s = Q_b = -\frac{1}{3}$ this gives $R = \frac{2}{3}N_c$ for u, d, s, $R = \frac{10}{9}N_c$ once charm opens and $R = \frac{11}{9}N_c$ once bottom opens — that is 2, 10/3 and 11/3 for $N_c = 3$, which is what is measured.

Confinement does not spoil the count. The quarks produced at the photon vertex cannot emerge as free particles, but whatever they do afterwards they end as hadrons with unit probability, so the hadronic total is settled at the moment of pair creation and the long-distance physics cancels in the ratio. This is the same inclusiveness argument that makes Theorem 100.58 work, applied to colour instead of to soft photons.

The residual excess of the data over the integers above is calculable. Gluon emission from the produced pair multiplies Equation (102.15) by $1 + \alpha_s(E_{\mathrm{cm}})/\pi + \ldots$, the first term being the same $\mathcal{O}(\alpha_s)$ real-plus-virtual sum that gives the three-jet rate of Section 102.6.2; at $E_{\mathrm{cm}} = 5\,\mathrm{GeV}$, where $\alpha_s \approx 0.2$, this is a 6 % excess, and it is seen. So R is simultaneously a colour counter at the 10 % level and, once the count is granted, one of the determinations of α_s collected in Section 102.4.2. □

The second count is sharper by a factor of ten in precision, and it is carried out in full in Section 106.7.2 rather than repeated here.

Proposition 102.16 (The neutral pion width counts colours squared)**.** *The two-photon width of the π^0 is fixed by the axial anomaly (Theorem 106.87) to be Equation* (106.103)*,*

$$\Gamma(\pi^0 \to \gamma\gamma) = \left(\frac{N_c}{3}\right)^2 \frac{\alpha^2 \left(m_\pi c^2\right)^3}{64\pi^3 F_\pi^2} \,, \tag{102.16}$$

proportional to N_c^2. For $N_c = 3$ the prediction is $(7.79 \pm 0.14)\,\mathrm{eV}$ [2427, 2431] against the measured $(7.802 \pm 0.117)\,\mathrm{eV}$ [2361]: agreement to 0.2 %. For $N_c = 1$ it would be nine times smaller.

Rests on Theorems 106.87 and 106.89 and Equation (106.103)*.*

Proof. *Derives Proposition 102.16.*

The whole content is the colour factor of the triangle graph. The axial current whose divergence carries the anomaly is the third isospin component, $j_3^{\mu 5} = \frac{1}{2}(\bar{u}\gamma^\mu\gamma^5 u - \bar{d}\gamma^\mu\gamma^5 d)$, and each quark in the loop is weighted by the square of its electric charge and summed over colours, the electromagnetic vertices being colour-blind. The sum is

$$N_c \left[Q_u^2 - Q_d^2\right] = N_c \left(\frac{4}{9} - \frac{1}{9}\right) = \frac{N_c}{3} \,, \tag{102.17}$$

the relative minus sign coming from the isospin structure of the current. The amplitude is proportional to Equation (102.17) and the width to its square, which is Equation (102.16);

the remaining factors — the anomaly coefficient itself, the reduction of the amplitude to the pion decay constant through the partial conservation of the axial current, and the two-body phase space — are derived in Section 106.7.2. The reason this count is sharper than R is that it is a *low-energy* theorem: the anomaly receives no corrections at higher order in α_s (Theorem 106.89), so the only uncertainty is the measured F_π, entering squared. □

Remark 102.17 (Drell–Yan as a third, weaker count). Drell and Yan proposed the production of a massive lepton pair in a hadron–hadron collision through the annihilation of a quark from one hadron with an antiquark from the other [1421]. Colour enters with the opposite sign to R: the annihilating pair must be in a colour-singlet configuration to make a virtual photon, and if the quark and the antiquark are drawn independently from their parents the probability of matching colours is $1/N_c$. The cross-section therefore carries a factor $1/N_c$ multiplying parton densities that, being measured in the deep inelastic scattering of Section 102.5, already contain a sum over N_c colours. The absolute normalization is thus a count. It is a weaker one, because the leading-order prediction is corrected by an $\mathcal{O}(\alpha_s)$ factor of about 1.5 to 2 — the so-called K factor, which is not small — so the measurement tests $N_c = 3$ at the tens-of-percent level rather than the percent level. It is reported here as a consistency check with an independent sign structure, which is what it is, and not as a precision determination.

102.3 The QCD Lagrangian

102.3.1 Yang–Mills theory

Colour, as introduced in Definition 102.10, is a global label. Theorem 100.5 showed what happens when a global phase symmetry is required to hold point by point: an interaction is not chosen but forced. Yang and Mills asked the same question of a non-abelian group [1054], taking isospin as their example, and the answer differs from the abelian one in exactly one respect — the mediating field is charged under the symmetry it mediates.

Notation 102.18 (Colour fields and their SI dimensions). The quark field q_f^i carries a flavour label f and a colour index $i = 1, 2, 3$, and has the dimension $1/\mathrm{m}^{3/2}$ of any Dirac field (Definition 100.2). The eight gauge potentials A_μ^a, $a = 1, \dots, 8$, carry the dimension $\mathrm{V\,s/m}$ of the electromagnetic potential, and the colour charge g_s carries the dimension C of an electric charge. The combination that appears everywhere is

$$\mathcal{A}_\mu := \frac{g_s}{\hbar} A_\mu^a T_a \,, \qquad [\mathcal{A}_\mu] = 1/\mathrm{m} \,, \tag{102.18}$$

which is Equation (106.63). The dimensionless measure of the coupling is

$$\alpha_s := \frac{g_s^2}{4\pi\varepsilon_0 \hbar c} = \frac{\mu_0 c\, g_s^2}{4\pi\hbar} \,, \tag{102.19}$$

the exact analogue of Equation (100.3) and equal to $\hat{g}^2/4\pi$ in the notation of Equation (106.66). That α_s is a pure number is, as for α, a fact and not a convention: g_s^2/ε_0 carries $\mathrm{J\,m}$, exactly as $\hbar c$ does.

Theorem 102.19 (The non-abelian gauge principle)**.** *Require the free quark Lagrangian $\mathcal{L}_0 = \bar{q}(\mathrm{i}\hbar c\,\gamma^\mu \partial_\mu - mc^2)q$, summed over colours, to be invariant under*

$$q(x) \longmapsto U(x)\, q(x) \,, \qquad U(x) = \exp[\mathrm{i}\theta^a(x) T_a] \in \mathrm{SU}(3)_c \,, \tag{102.20}$$

with θ^a arbitrary real functions. Then invariance holds if and only if eight vector fields A_μ^a are introduced, the derivative is replaced by

$$D_\mu := \partial_\mu + \frac{\mathrm{i} g_s}{\hbar} A_\mu^a T_a = \partial_\mu + \mathrm{i}\mathcal{A}_\mu \,, \tag{102.21}$$

and the potentials transform as

$$\mathcal{A}_\mu \longmapsto U\mathcal{A}_\mu U^\dagger + \mathrm{i}\,(\partial_\mu U)\,U^\dagger\,, \tag{102.22}$$

which infinitesimally reads

$$\delta A^a_\mu = -\frac{\hbar}{g_s}\,\partial_\mu\theta^a - f^{abc}\theta^b A^c_\mu\,. \tag{102.23}$$

Rests on Definition 102.10 and Equations (14.67) and (100.8).

Proof. *Derives Theorem 102.19.*

Under Equation (102.20) the derivative picks up an inhomogeneous term, $\partial_\mu q \mapsto U\partial_\mu q + (\partial_\mu U)q$, so $\mathcal{L}_0$ is not invariant. Postulate a matrix field $\mathcal{A}_\mu$ valued in the Lie algebra and define D_μ by Equation (102.21). Demanding $D_\mu q \mapsto U(D_\mu q)$ gives

$$\left(\partial_\mu + \mathrm{i}\mathcal{A}'_\mu\right) Uq = (\partial_\mu U)\,q + U\partial_\mu q + \mathrm{i}\mathcal{A}'_\mu U q \overset{!}{=} U\partial_\mu q + \mathrm{i}U\mathcal{A}_\mu q\,,$$

for all q, whence $\mathrm{i}\mathcal{A}'_\mu U = \mathrm{i}U\mathcal{A}_\mu - (\partial_\mu U)$ and, multiplying on the right by $U^\dagger$, Equation (102.22). Conversely Equation (102.22) makes $\bar q\gamma^\mu D_\mu q$ invariant by the same computation read backwards, so the condition is necessary as well as sufficient.

For the infinitesimal form put $U = \mathbb{1} + \mathrm{i}\theta^a T_a$ and keep first order. The first term of Equation (102.22) gives $\mathcal{A}_\mu + \mathrm{i}\theta^b\,[T_b, \mathcal{A}_\mu]$, and with $[T_b, T_c] = \mathrm{i}f_{bca}T_a$ of Equation (14.67) together with the total antisymmetry of f this is $\mathcal{A}_\mu - (g_s/\hbar)f^{abc}\theta^b A^c_\mu T_a$. The second term gives $\mathrm{i}(\mathrm{i}\partial_\mu\theta^a T_a) = -\partial_\mu\theta^a T_a$. Dividing by $g_s/\hbar$ yields Equation (102.23). For one abelian generator the second term of Equation (102.23) is absent and the first reproduces Equation (100.8) with $\Lambda = -\hbar\theta/g_s$. □

The whole difference between QED and QCD sits in the second term of Equation (102.23): the gluon potential does not merely shift under a gauge transformation, it *rotates* in colour space. It carries colour, and a field that carries the charge it mediates interacts with itself.

Definition 102.20 (The gluon field strength)**.**

$$F^a_{\mu\nu} := \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - \frac{g_s}{\hbar} f^{abc} A^b_\mu A^c_\nu\,, \tag{102.24}$$

of dimension T, equivalently $\mathcal{F}_{\mu\nu} = (g_s/\hbar)F^a_{\mu\nu}T_a = -\mathrm{i}\,[D_\mu, D_\nu]$, which is Equation (106.64).

Proposition 102.21 (Curvature and its transformation law)**.** $\mathcal{F}_{\mu\nu} = -\mathrm{i}\,[D_\mu, D_\nu]$ *holds with Equation (102.24), and under Equation (102.22)*

$$\mathcal{F}_{\mu\nu} \longmapsto U\mathcal{F}_{\mu\nu}U^\dagger\,, \tag{102.25}$$

so that $F^a_{\mu\nu}$ transforms in the adjoint representation **8** *and $F^a_{\mu\nu}F^{a\,\mu\nu}$ is gauge invariant. Unlike Equation (100.11), $F^a_{\mu\nu}$ is itself* not *invariant.*

Rests on Equations (14.66), (102.22) and (102.24).

Proof. *Derives Proposition 102.21.*

Compute the commutator on a quark field:

$$\begin{aligned}[D_\mu, D_\nu] &= \mathrm{i}\,(\partial_\mu\mathcal{A}_\nu - \partial_\nu\mathcal{A}_\mu) - [\mathcal{A}_\mu, \mathcal{A}_\nu] \\ &= \mathrm{i}\frac{g_s}{\hbar}\left(\partial_\mu A^a_\nu - \partial_\nu A^a_\mu\right)T_a - \left(\frac{g_s}{\hbar}\right)^2 A^b_\mu A^c_\nu\,\mathrm{i}f_{bca}T_a\,,\end{aligned} \tag{102.26}$$

the derivative terms acting on the quark cancelling between the two orderings. Factoring $\mathrm{i}g_s/\hbar$ and using $f_{bca} = f_{abc}$ gives Equation (102.24). Since $D_\mu \mapsto UD_\mu U^\dagger$ by construction — Equation (102.22) is exactly the statement that $D_\mu q$ transforms like q — the commutator of two such objects transforms as $U\,[D_\mu, D_\nu]\,U^\dagger$, which is Equation (102.25). Finally $F^aF^a \propto \mathrm{tr}(\mathcal{F}_{\mu\nu}\mathcal{F}^{\mu\nu})$ by Equation (14.66), and a trace of a product of two adjoint objects is invariant by cyclicity — it is the Killing form of Proposition 14.53, which is what makes this the only quadratic invariant available. □

Remark 102.22 (A sign convention, and how to convert). Many texts define $D_\mu = \partial_\mu - \mathrm{i}\mathcal{A}_\mu$ and obtain $F^a_{\mu\nu}$ with $+g_s f^{abc} A^b A^c/\hbar$. The two conventions are related by $g_s \to -g_s$, under which every physical quantity — which depends on g_s only through $\alpha_s \propto g_s^2$, or through an even number of vertices — is unchanged. The convention adopted here is the one that makes Equation (102.21) the literal non-abelian copy of Equation (100.9) and agrees with Notation 106.50, so that formulae may be carried between the three chapters without a sign audit.

Proposition 102.23 (The gluon self-couplings)**.** *The Yang–Mills Lagrangian density*

$$\mathcal{L}_{\mathrm{YM}} = -\frac{1}{4\mu_0} F^a_{\mu\nu} F^{a\,\mu\nu}\,, \tag{102.27}$$

of dimension $\mathrm{J/m^3}$*, contains — besides the free quadratic term — a cubic and a quartic self-interaction of the gauge field,*

$$\begin{aligned}\mathcal{L}_{\mathrm{YM}} = &-\frac{1}{4\mu_0}\left(\partial_\mu A^a_\nu - \partial_\nu A^a_\mu\right)\left(\partial^\mu A^{a\nu} - \partial^\nu A^{a\mu}\right)\\ &+\frac{g_s}{2\mu_0\hbar} f^{abc}\left(\partial_\mu A^a_\nu - \partial_\nu A^a_\mu\right) A^{b\,\mu} A^{c\,\nu}\\ &-\frac{g_s^2}{4\mu_0\hbar^2} f^{abc} f^{ade} A^b_\mu A^c_\nu A^{d\,\mu} A^{e\,\nu}\,,\end{aligned} \tag{102.28}$$

both proportional to the structure constants Equation (14.69) *and therefore absent in an abelian theory.*

Rests on Equations (14.69)*,* (100.12) *and* (102.24)*.*

Proof. *Derives Proposition 102.23.*

Write $F^a_{\mu\nu} = G^a_{\mu\nu} - (g_s/\hbar) f^{abc} A^b_\mu A^c_\nu$ with $G^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu$ and square, the cross term appearing twice by the symmetry of the contraction. Multiplying by $-1/4\mu_0$ gives Equation (102.28). For an abelian group all f^{abc} vanish and only the first line survives, which is Equation (100.12); the normalization $1/4\mu_0$ is fixed by that limit. Dimensionally F^2/μ_0 carries $\mathrm{T^2\,A^2/N} = \mathrm{J/m^3}$, as in Equation (100.12). □

Remark 102.24 (Yang and Mills's mass problem). A Proca term $(2\mu_0)^{-1}\,(m_g c/\hbar)^2\, A^a_\mu A^{a\,\mu}$, which is the only gluon mass term available and carries $\mathrm{J/m^3}$ as it must, is not invariant under Equation (102.23) — the inhomogeneous piece $\partial_\mu \theta^a$ does not cancel — so gauge invariance forbids a gluon mass, exactly as it forbids a photon mass in Phenomenon 100.7. This was the objection to [1054] at the time: a massless charged vector particle would have been conspicuous, and none was seen. The two known resolutions are different from each other. For the weak interaction the symmetry is spontaneously broken and the gauge bosons acquire mass through the mechanism of Chapter 104. For the strong interaction the gluon really is massless, and the interaction is nevertheless short-ranged because the theory confines (Section 102.7) — a resolution that requires the non-perturbative behaviour of the very self-couplings Equation (102.28) that caused the problem. It took nineteen years.

102.3.2 Gauge fixing, ghosts and the Slavnov–Taylor identities

The quadratic form of Equation (102.28) is singular for the same reason as Equation (100.15): the action does not depend on the gauge direction, so it cannot be inverted there and there is no propagator. In the abelian case the cure of Section 100.1.3 — add $-(\partial\cdot A)^2/2\mu_0\xi$ and proceed — is complete, because the Jacobian accompanying the change of variables is a field-independent constant. In the non-abelian case it is not, and the difference is a new field.

Theorem 102.25 (Gauge fixing needs a determinant)**.** *Fixing the gauge by a condition* $G^a(\mathcal{A}) = 0$ *inside the functional integral of Chapter 106 requires the insertion of the Faddeev–Popov determinant* $\det\left(\delta G^a/\delta\theta^b\right)$ *[1347]. For the covariant condition* $G^a = \partial^\mu A^a_\mu$ *this operator is* $\partial^\mu D_\mu$*, which depends on the gauge field whenever the group is non-abelian; the determinant is therefore not a constant and is represented by a functional integral over a pair of anticommuting scalar fields* c^a*,* $\bar{c}^a$ *in the adjoint representation, the* ghosts*.*

Rests on Theorem 106.52, Equation (102.23)*, and Corollary 106.53.*

Proof. *Derives Theorem 102.25.*

This is Theorem 106.52, proved there; only the group-theoretic input is quoted here. The variation of the gauge condition under Equation (102.23) is $\delta G^a = -(\hbar/g_s)\partial^\mu\left(\delta^{ab}\partial_\mu + (g_s/\hbar)f^{acb}A^c_\mu\right)\theta^b$, whose kernel is the adjoint covariant derivative. The f^{acb} term is absent for an abelian group, which is Corollary 106.53 and the reason the photon needed no ghosts in Section 100.1.3. □

The ghosts are not particles: they are anticommuting scalars, in violation of the spin-statistics theorem, and they may therefore never appear in an asymptotic state. Their function is negative — they cancel the contributions of the unphysical gluon polarizations. A massless vector field has four components and two physical (transverse) polarizations; in a covariant gauge all four propagate, and the timelike and longitudinal ones would spoil the unitarity of the theory. That the ghost loop cancels them exactly is Proposition 106.55, and it is the reason the practical quantization of QCD is the functional one of Chapter 106 rather than the canonical route of Chapter 99.

What survives gauge fixing is a residual global symmetry mixing gauge field and ghosts: the BRST symmetry of Becchi, Rouet and Stora [1518, 1571] and, independently, Tyutin [1564]. Its generator is nilpotent (Theorem 106.57), the physical state space is its cohomology (Definition 106.59), and it is the modern statement of what gauge invariance becomes after the gauge has been fixed. Its practical consequence is a set of identities among renormalization constants, the Slavnov–Taylor identities [1458, 1484], generalizing the Ward–Takahashi identity Theorem 100.48 of the abelian theory. Where the abelian identity gives the single relation $Z_1 = Z_2$ of Corollary 100.49 — charge renormalization is governed by the photon field renormalization alone — the non-abelian identities relate the quark–gluon, three-gluon, four-gluon and ghost–gluon vertex renormalizations to each other, so that the *one* coupling g_s of Equation (102.28) remains one coupling after renormalization. That is not automatic: three vertices renormalized independently would be three couplings, and the theory would lose its predictive content.

Theorem 102.26 (Non-abelian gauge theories are renormalizable)**.** *A Yang–Mills theory with fermions in any representation, quantized with Theorem 102.25, is renormalizable: all ultraviolet divergences are removed to every order by a finite number of counterterms with the same structure as the terms already present [1448, 1463].*

Rests on Theorems 100.17 and 102.25.

Derivation pending #148

Renormalizability of Yang–Mills theory. The proof combines the power-counting theorem, which for a gauge theory in four dimensions gives superficial degree of divergence four minus the number of external lines weighted by their dimension exactly as in the abelian case, with the Slavnov–Taylor identities, which show that the divergent structures allowed by power counting are constrained to be gauge invariant and therefore already present in the Lagrangian. The regularization must respect the gauge symmetry, and dimensional regularization does. This is a long argument of the same shape as the BPHZ theorem of the electromagnetic chapter and belongs beside it in Appendix A.

102.3.3 The Lagrangian and its parameters

Definition 102.27 (The QCD Lagrangian density)**.**

$$\mathcal{L}_{\mathrm{QCD}} = -\frac{1}{4\mu_0} F^a_{\mu\nu} F^{a\,\mu\nu} + \sum_f \bar{q}_f \left(\mathrm{i}\hbar c\, \gamma^\mu D_\mu - m_f c^2 \right) q_f \,, \tag{102.29}$$

with $F^a_{\mu\nu}$ from Equation (102.24), D_μ from Equation (102.21) acting on the colour triplet index, and the flavour sum running over $f = u, d, s, c, b, t$. Both terms carry $\mathrm{J/m^3}$. This is the Lagrangian written down by Fritzsch, Gell-Mann and Leutwyler [1499]. A further gauge-invariant term is allowed and is deferred to Section 102.10.

Proposition 102.28 (The parameters of QCD)**.** *Equation (102.29) contains seven free parameters: the six quark masses and the coupling g_s, equivalently α_s at one reference scale. Nothing else — not the number of gluons, not their couplings to each other, not the relative strength with which any flavour couples to them — is adjustable.*

Rests on Equation (102.29), Proposition 14.50, and Theorem 102.26.

Proof. *Derives Proposition 102.28.*

The gauge sector is fixed by the group: the number of gauge fields is $\dim \mathfrak{su}(3) = 8$ by Proposition 14.50, and the cubic and quartic couplings of Equation (102.28) carry the structure constants Equation (14.69), which are numbers, multiplied by powers of the same g_s that appears in the quark vertex. The quark sector contributes one mass per flavour. That the coupling is the same for every flavour is the non-abelian version of Proposition 100.6: D_μ depends on the matter field only through its colour representation, and every quark is a $\mathbf{3}$. Renormalizability (Theorem 102.26) is what guarantees the count is not enlarged order by order, and the Slavnov–Taylor identities are what keep the three gauge vertices tied to one number. □

Proposition 102.29 (Feynman rules for QCD in SI form)**.** *The rules of Proposition 100.14 carry over with the following replacements and additions, each factor written with its $\hbar$ and c in place.*

1. *Quark propagator: Equation (100.23) times δ_{ij} in colour.*
2. *Gluon propagator: Equation (100.17) times δ^{ab}.*
3. *Quark–gluon vertex:*

$$-\frac{\mathrm{i} g_s}{\hbar} \gamma^\mu \left(T_a\right)_{ij} \,, \tag{102.30}$$

of dimension $\mathrm{C/(J\,s)}$, identical to Equation (100.24) with $q \to g_s T_a$.

Table 102.1: The seven parameters of Equation (102.29). Quark masses are $\overline{\text{MS}}$ values, those of u, d and s quoted at the scale 2 GeV and those of c, b, t at their own mass, in the scheme of Definition 100.54; the scheme dependence is not an uncertainty but part of the definition, because a quark mass is not the pole of a propagator of anything observable. Values from [2427].

Parameter	Value	SI equivalent
m_u	$(2.16 \pm 0.07)\,\text{MeV}/c^2$	$3.85 \times 10^{-30}\,\text{kg}$
m_d	$(4.70 \pm 0.07)\,\text{MeV}/c^2$	$8.38 \times 10^{-30}\,\text{kg}$
m_s	$(93.5 \pm 0.8)\,\text{MeV}/c^2$	$1.667 \times 10^{-28}\,\text{kg}$
m_c	$(1.273 \pm 0.005)\,\text{GeV}/c^2$	$2.269 \times 10^{-27}\,\text{kg}$
m_b	$(4.183 \pm 0.007)\,\text{GeV}/c^2$	$7.457 \times 10^{-27}\,\text{kg}$
m_t	$(172.57 \pm 0.29)\,\text{GeV}/c^2$	$3.0763 \times 10^{-25}\,\text{kg}$
$\alpha_s(m_Z c^2)$	0.1180 ± 0.0009	dimensionless

4. *Three-gluon vertex: from the second line of Equation (102.28), a factor proportional to $(g_s/\hbar\mu_0)f^{abc}$ times a momentum-dependent tensor antisymmetric in the simultaneous exchange of any two legs.*

5. *Four-gluon vertex: from the third line, proportional to $(g_s/\hbar)^2\mu_0^{-1}$ times sums of products $f^{abe}f^{cde}$.*

6. *Ghost propagator and ghost–gluon vertex from Theorem 102.25, with a factor -1 for each closed ghost loop as for a fermion loop.*

Every colour factor is a contraction of the T_a and f^{abc} computed once and for all in Proposition 14.52. The two that occur constantly are

$$(T_aT_a)_{ij} = C_F\,\delta_{ij}\,, \quad C_F = \frac{4}{3}\,, \qquad f^{acd}f^{bcd} = C_A\,\delta^{ab}\,, \quad C_A = 3\,, \tag{102.31}$$

being the quadratic Casimirs Equation (14.77) and Equation (14.78) of the fundamental and adjoint representations, and $\operatorname{tr}(T_aT_b) = T_F\delta_{ab}$ with $T_F = \frac{1}{2}$ from Equation (14.66).

Rests on Proposition 100.14, Equation (102.28), and Theorem 102.25.

Proof. *Derives Proposition 102.29.*

Rules 1–3 are Proposition 100.14 with the substitution $q\gamma^\mu \to g_s\gamma^\mu T_a$ forced by Equation (102.21), together with the observation that the free parts of Equation (102.29) are diagonal in colour; rules 4–6 are read off the interaction terms of Equation (102.28) and of Theorem 102.25 by the same expansion of $\exp(\mathrm{i}S/\hbar)$. For Equation (102.31), T_aT_a commutes with every generator and is therefore a multiple of the identity on the irreducible **3**; Equation (14.77) evaluates the multiple as $4/3$, and Equation (14.78) does the same for the adjoint, where the generators are $-\mathrm{i}f^{abc}$. □

Remark 102.30 (Flavour independence is a prediction, and it is tested). Proposition 102.28 asserts that one number governs the coupling of the gluon to every quark flavour. The claim is testable because α_s is extracted from processes dominated by different flavours: τ decays involve u, d and s; bottomonium spectroscopy involves b; the Z hadronic width involves all five accessible flavours democratically. That these determinations agree after being evolved to a common scale, as Section 102.4.2 records, is a test of flavour independence and not merely a measurement of a number.

102.4 Asymptotic freedom

102.4.1 The sign of the beta function

Remark 100.34 read the running of α as the polarization of a medium: virtual pairs align against an inserted charge, so that a distant probe sees less charge than a close one, and the effective coupling grows with momentum transfer. The reading is literal, and in SI units it can be pushed further than it usually is, because SI keeps the vacuum's electric and magnetic responses as separate named quantities. That separation is what makes the sign of the non-abelian beta function computable by a short argument instead of a two-loop-sized diagram count.

Lemma 102.31 (The vacuum's two responses are reciprocal)**.** *Let the polarization of the vacuum by virtual pairs be described by an effective permittivity ε_{eff} and permeability μ_{eff}. Lorentz invariance of the vacuum requires*

$$\varepsilon_{\text{eff}}\,\mu_{\text{eff}} = \frac{1}{c^2}\,, \qquad \textit{hence} \qquad \frac{\varepsilon_{\text{eff}}}{\varepsilon_0} = \frac{\mu_0}{\mu_{\text{eff}}}\,. \tag{102.32}$$

Consequently a paramagnetic *vacuum, $\mu_{\text{eff}} > \mu_0$, is one with $\varepsilon_{\text{eff}} < \varepsilon_0$: it* anti*screens.*
Rests on Postulate 38.2 and Equation (102.19).

Proof. *Derives Lemma 102.31.*
In a medium of permittivity ε and permeability μ the phase velocity of light is $(\varepsilon\mu)^{-1/2}$. The vacuum, polarized or not, is Lorentz invariant, and by the postulates of Chapter 38 light propagates in it at c in every frame; a medium with $\varepsilon\mu \neq c^{-2}$ would define a rest frame. Hence Equation (102.32). The second form follows from $\varepsilon_0\mu_0 c^2 = 1$, and the statement about signs is immediate. Finally, the coupling measured at a scale is $\alpha_s = g_s^2/\left(4\pi\varepsilon_{\text{eff}}\hbar c\right)$ by Equation (102.19), so $\varepsilon_{\text{eff}} < \varepsilon_0$ means a *larger* effective coupling at long distance and a smaller one at short distance, which is asymptotic freedom. □

The problem is therefore reduced to a magnetostatic one: is the QCD vacuum paramagnetic or diamagnetic? A magnetic response has two sources, and they compete. Orbital motion of a charged particle in a field is diamagnetic — Lenz's law, and Landau diamagnetism. Alignment of an intrinsic magnetic moment with the field is paramagnetic. For a charged particle of spin s and gyromagnetic ratio g the two enter the energy in a fixed ratio, and the following lemma is where the ratio is fixed.

Lemma 102.32 (Landau levels with $g = 2$)**.** *A massless particle of charge q, spin projection s_z along a uniform magnetic field $B\hat{z}$ and gyromagnetic ratio $g = 2$ has energies*

$$E_n^2(p_z) = c^2 p_z^2 + |q|\,\hbar c^2 B\,(2n + 1 + \kappa)\,, \qquad \kappa := -2s_z\,, \quad n = 0, 1, 2, \dots\,, \tag{102.33}$$

each level carrying the degeneracy $|q|\,B/(2\pi\hbar)$ per unit area transverse to the field. The quantity $|q|\,\hbar c^2 B$ has the dimension of a squared energy.
Rests on Equations (93.2) and (102.28).

Proof. *Derives Lemma 102.32.*
For a spinless particle the transverse motion is a harmonic oscillator of frequency $\omega_c = |q|\,c^2 B/E$, whose quantization gives $E^2 = c^2 p_z^2 + (2n+1)\,|q|\,\hbar c^2 B$ — the relativistic Landau problem, obtained by squaring the wave operator so that $\boldsymbol{p}_\perp^2 \to (2n+1)\,|q|\,\hbar B$. A magnetic moment $\boldsymbol{\mu} = g\,(q/2m)\,\boldsymbol{S}$ adds an interaction energy $-\boldsymbol{\mu}\cdot\boldsymbol{B}$, which in the squared equation appears as $-g\,|q|\,\hbar c^2 B s_z$; putting $g = 2$ gives Equation (102.33). That $g = 2$ is the Dirac value for a spin-$\frac{1}{2}$ particle (Chapter 93), and its small radiative correction (Theorem 100.41) is of higher order in the coupling and irrelevant to a one-loop coefficient. That $g = 2$ also for the charged gluons is a property of Equation (102.28): expanding

$-F^a_{\mu\nu}F^{a\,\mu\nu}/4\mu_0$ to second order about a background field produces, besides the covariant Laplacian, exactly the term $-2\mathrm{i}\mathcal{F}_{\mu\nu}$ acting on the vector index of the fluctuation, and a coefficient 2 in front of the field-strength coupling to the spin generator is the statement $g=2$. The degeneracy $|q|\,B/(2\pi\hbar)$ per unit area is the standard one: one state per flux quantum $2\pi\hbar/|q|$. □

Theorem 102.33 (The one-loop coefficient, from a magnetic response)**.** *Let a gauge theory contain species i of spin s_i carrying charge $x_i g_s$ under a chosen Cartan direction, the label i running over all one-particle states (particles and antiparticles counted separately). Then the vacuum behaves as a medium with*

$$\frac{\varepsilon_{\mathrm{eff}}}{\varepsilon_0}=1+\frac{\alpha_s}{24\pi}\,S\,\ln\frac{\Lambda_E^2}{E^2}\,,\qquad S:=\sum_i x_i^2\,(-1)^{2s_i}\sum_{s_z}\Big(1-12s_z^2\Big)\,, \tag{102.34}$$

with Λ_E an ultraviolet cutoff energy and E the energy scale at which the coupling is measured; and consequently

$$\frac{\mathrm{d}\alpha_s}{\mathrm{d}\ln E}=\frac{S}{12\pi}\,\alpha_s^2+O\Big(\alpha_s^3\Big)\,. \tag{102.35}$$

Inside the spin sum the term 1 is the diamagnetic contribution of the orbital motion and the term $-12s_z^2$ the paramagnetic contribution of the magnetic moment.

Rests on Lemmas 102.31 and 102.32 and Equation (102.19).

Proof. *Derives Theorem 102.33.*

Put the vacuum in a uniform field B pointing along the chosen Cartan direction in colour space, so that every field component has a definite charge $x_i g_s$ and the problem is the abelian one of Lemma 102.32. The vacuum energy density is the sum of the zero-point energies of all modes,

$$u=\frac{1}{2}\sum_i(-1)^{2s_i}\frac{|q_i|\,B}{2\pi\hbar}\sum_{s_z}\sum_{n=0}^{\infty}\int_{-\infty}^{\infty}\frac{\mathrm{d}p_z}{2\pi\hbar}\,E_n(p_z)\,, \tag{102.36}$$

the factor $(-1)^{2s}$ being the sign of a fermionic zero-point energy, $-\frac{1}{2}\hbar\omega$ per mode instead of $+\frac{1}{2}\hbar\omega$.

Cut the longitudinal momentum off at $|p_z|\le P$ and write $\varepsilon_n^2=A\,(2n+1+\kappa)$ with $A=|q|\,\hbar c^2 B$. Then

$$\int_{-P}^{P}\frac{\mathrm{d}p_z}{2\pi\hbar}\sqrt{c^2p_z^2+\varepsilon_n^2}=\frac{cP^2}{2\pi\hbar}+\frac{\varepsilon_n^2}{4\pi\hbar c}-\frac{\varepsilon_n^2}{4\pi\hbar c}\ln\frac{\varepsilon_n^2}{4c^2P^2}+O\Big(P^{-2}\Big)\,, \tag{102.37}$$

the first two terms being independent of B and of n in a way that contributes no $B^2\ln B$, and the third being the one that matters. Split $\ln\varepsilon_n^2=\ln A+\ln(2n+1+\kappa)$: the second piece carries no B and produces a term proportional to B^2 with no logarithm, i.e. a renormalization of the field energy and not a running. Keeping the first piece,

$$u_{\ln}=-\frac{1}{2}\sum_i(-1)^{2s_i}\frac{|q_i|\,B}{2\pi\hbar}\cdot\frac{A_i}{4\pi\hbar c}\,\ln\frac{A_i}{4c^2P^2}\sum_{s_z}\sum_{n=0}^{\infty}(2n+1+\kappa)\,. \tag{102.38}$$

The sum over Landau levels diverges and is defined by analytic continuation: writing $2n+1+\kappa=2(n+x)$ with $x=(1+\kappa)/2$, and using the Hurwitz zeta function $\zeta(\sigma,x)=\sum_{n\ge0}(n+x)^{-\sigma}$ continued to $\sigma=-1$, where $\zeta(-1,x)=-\frac{1}{2}\left(x^2-x+\frac{1}{6}\right)$,

$$\sum_{n=0}^{\infty}(2n+1+\kappa)=2\zeta(-1,x)=-\left(x^2-x+\tfrac{1}{6}\right)=\frac{1-3\kappa^2}{12}=\frac{1-12s_z^2}{12}\,, \tag{102.39}$$

using $\kappa = -2s_z$. The two terms of Equation (102.39) are the two magnetic responses: the 1 comes from the $2n+1$ of the orbital motion and the $-12s_z^2$ from the κ of the moment.

Substituting $A_i = |q_i|\,\hbar c^2 B$ into Equation (102.38) and writing $L_i = \ln\left[4c^2P^2/(|q_i|\,\hbar c^2 B)\right]$,

$$u_{\ln} = \frac{cB^2}{16\pi^2\hbar}\sum_i q_i^2(-1)^{2s_i}L_i\sum_{s_z}\frac{1-12s_z^2}{12}\,. \tag{102.40}$$

Comparing with the field energy density $u = B^2/2\mu_{\text{eff}}$ and using Equation (102.32) together with $q_i^2 = x_i^2 g_s^2$ and $g_s^2/(\varepsilon_0\hbar c) = 4\pi\alpha_s$ from Equation (102.19),

$$\frac{\varepsilon_{\text{eff}}}{\varepsilon_0} = \frac{\mu_0}{\mu_{\text{eff}}} = 1 + \frac{\alpha_s}{24\pi}\sum_i x_i^2(-1)^{2s_i}\left[\sum_{s_z}\left(1-12s_z^2\right)\right]L_i\,. \tag{102.41}$$

The species-dependence of L_i is a B-independent additive constant under the logarithm and is therefore a choice of subtraction scheme, not physics: only the coefficient of $\ln B$ is common to all schemes, which is the standard statement that a one-loop beta-function coefficient is scheme-independent (Proposition 100.55). Identifying the energy scale probed by the field, $E^2 \sim g_s\hbar c^2 B$ — the field sets the transverse momentum of the modes — and writing $\Lambda_E = 2cP$, all L_i become $\ln(\Lambda_E^2/E^2)$ up to constants, which is Equation (102.34).

Finally $\alpha_s(E) = g_s^2/\left(4\pi\varepsilon_{\text{eff}}\hbar c\right)$, so $\alpha_s(E)^{-1} = \alpha_s(\Lambda_E)^{-1} + (S/24\pi)\ln(\Lambda_E^2/E^2)$. Differentiating with respect to $\ln E$ gives $-\alpha_s^{-2}\,\mathrm{d}\alpha_s/\mathrm{d}\ln E = -S/12\pi$, which is Equation (102.35). □

Corollary 102.34 (The master formula reproduces QED)**.** *For a single Dirac fermion of unit charge, Equation* (102.35) *gives* $\mathrm{d}\alpha/\mathrm{d}\ln E = 2\alpha^2/3\pi$*, which is Equation* (100.84)*.*

Rests on Equations (100.84)*,* (102.35) *and* (102.39)*.*

Proof. *Derives Corollary 102.34.*

The one-particle states are the fermion and the antifermion, each of charge magnitude e, so $\sum_i x_i^2 = 2$. The statistics factor is $(-1)^{2s} = -1$ and the spin sum over $s_z = \pm\frac{1}{2}$ is $2\left(1 - 12\cdot\frac{1}{4}\right) = -4$. Hence $S = 2\cdot(-1)\cdot(-4) = 8$ and $\mathrm{d}\alpha/\mathrm{d}\ln E = 8\alpha^2/12\pi = 2\alpha^2/3\pi$. The sign is positive — QED screens — and by Equation (102.39) the reason is arithmetical: for $s_z = \pm\frac{1}{2}$ the paramagnetic term $12s_z^2 = 3$ beats the diamagnetic 1, so the electron vacuum is in fact paramagnetic, and it is the Fermi statistics — the minus sign of a fermion loop — that turns the result back into screening. □

Theorem 102.35 (Asymptotic freedom)**.** *For* $\mathrm{SU}(3)_c$ *with* n_f *quark flavours light enough to be excited at the scale considered,*

$$\frac{\mathrm{d}\alpha_s}{\mathrm{d}\ln E} = -\frac{b_0}{2\pi}\,\alpha_s^2 + O\left(\alpha_s^3\right)\,, \qquad b_0 = \frac{11}{3}C_A - \frac{4}{3}T_F n_f = 11 - \frac{2}{3}n_f\,, \tag{102.42}$$

which is negative — the coupling decreases with energy — for every $n_f \le 16$ *[1501, 1511]. Nature supplies* $n_f = 6$*.*

Rests on Theorem 102.33 and Equations (14.80) *and* (102.31)*.*

Proof. *Derives Theorem 102.35.*

Apply Theorem 102.33 twice.

Quarks. A quark flavour is a colour triplet; taking the Cartan direction to be T_3, its three colour states carry charges $x = +\frac{1}{2}, -\frac{1}{2}, 0$, and each is a Dirac field with its own antiparticle. Hence $\sum_i x_i^2 = 2\,\mathrm{tr}\left(T_3^2\right) = 2T_F = 1$ per flavour, the factor 2 counting antiparticles and $T_F = \frac{1}{2}$ being Equation (14.66). With the spin factor $(-1)\times(-4) = +4$ from Corollary 102.34, quarks contribute $S_q = 4n_f$.

Gluons. The gluons form the adjoint $\mathbf{8}$, whose charges under ad_{T_3} are the first components of the roots Equation (14.80), namely ±1, $\pm\frac{1}{2}$, $\pm\frac{1}{2}$ and two zeros, so that $\sum_a x_a^2 = f^{3cd}f^{3cd} =$

$C_A = 3$ by Equation (102.31). Here the charged components already occur in ± pairs — a gluon of charge $+1$ *is* the antiparticle of the one with charge -1 — so no further doubling is applied. A gluon is massless and has two physical polarizations, $s_z = \pm 1$; the unphysical ones are removed by the ghosts of Theorem 102.25, which is where that machinery earns its keep. The statistics factor is $(-1)^2 = +1$ and the spin sum is

$$\sum_{s_z=\pm 1} \left(1 - 12 s_z^2\right) = 2\,(1 - 12) = -22\,, \tag{102.43}$$

so $S_g = 3 \times (-22) = -66$.

Adding, $S = 4n_f - 66 = -6\left(11 - \frac{2}{3}n_f\right) = -6b_0$, and Equation (102.35) becomes Equation (102.42). Restoring the group factors in place of their numerical values, the quark term is $-\frac{4}{3}T_F n_f$ and the gluon term $+\frac{11}{3}C_A$, which is the form used for any gauge group. $b_0 > 0$ requires $n_f < 33/2$, i.e. $n_f \leq 16$. □

Remark 102.36 (Eleven is twelve minus one). The whole of asymptotic freedom is in Equation (102.43). A gluon carries spin 1, so its paramagnetic term is $12 s_z^2 = 12$ against a diamagnetic term of 1: the moment wins by eleven to one, the vacuum is strongly paramagnetic, and by Lemma 102.31 a paramagnetic vacuum antiscreens. The famous 11 of b_0 is literally $12 - 1$. The electron's moment also wins, three to one, but a fermion loop enters with the opposite sign and the result is screening. So the two theories differ not because the gluon is charged in some vague sense, but because *spin one* triples the moment term relative to spin one half while *Bose statistics* declines to flip its sign. This is the contrast Remark 100.34 promised, made quantitative: QED has $S = 8$ per unit-charge fermion and QCD has $S = -66$ from the gluons alone.

Remark 102.37 (What the short calculation costs). Three things in the derivation above are shortcuts, and none of them affects the coefficient. First, the sum Equation (102.39) is divergent and was assigned a value by analytic continuation. A proper-time or dimensional regulator (Definition 100.30) gives the same coefficient of $\ln B$, which is the only quantity used; the finite parts differ and are scheme, as Equation (102.41) already noted. Second, the constant chromomagnetic field is not the true vacuum: the $n = 0$, $s_z = +1$ gluon mode of Equation (102.33) has $E^2 = c^2 p_z^2 - |q|\,\hbar c^2 B$, which is negative for small p_z, so the configuration is unstable and the effective action acquires an imaginary part [1624]. The instability is real physics but it does not touch the real part, which is where the logarithm and hence b_0 live. Third, only the physical gluon polarizations were counted, the unphysical ones being assumed cancelled by ghosts; the gauge-invariant background-field computation does this explicitly and confirms Equation (102.42) [1463, 1501, 1511]. The argument is given in this form because it exhibits the *mechanism* — a competition between two magnetic responses that SI units keep visibly distinct — rather than only the answer.

Definition 102.38 (The scale of QCD)**.** Integrating Equation (102.42) between two energies gives

$$\frac{1}{\alpha_s(E)} = \frac{1}{\alpha_s(E_0)} + \frac{b_0}{2\pi} \ln \frac{E}{E_0}\,, \tag{102.44}$$

so that the pair (α_s, E_0) may be traded for the single energy Λ_{QCD} at which the right-hand side formally vanishes:

$$\alpha_s(E) = \frac{2\pi}{b_0 \ln\left(E/\Lambda_{\mathrm{QCD}}\right)}\,, \qquad \Lambda_{\mathrm{QCD}} = E_0 \exp\!\left[-\frac{2\pi}{b_0 \alpha_s(E_0)}\right]. \tag{102.45}$$

Phenomenon 102.39 (A mass scale from a dimensionless coupling)**.** The Lagrangian Equation (102.29) with the light quark masses set to zero contains no dimensionful parameter whatever: $\hbar$ and c are conversion factors, and α_s is a pure number by Equation (102.19).

Nevertheless the theory it defines has a scale, and the scale is of the observed size: Λ_{QCD} is a few hundred MeV, corresponding to a length $\hbar c/\Lambda_{\mathrm{QCD}}$ of about $1 \times 10^{-15}\,\mathrm{m}$ — the size of a hadron. Almost all of the mass of ordinary matter is set by this number, which appears in no line of the Lagrangian.

Rests on Equations (102.19), (102.29) *and* (102.45).

Derivation. *Derives Phenomenon 102.39.*

This is *dimensional transmutation*, and Equation (102.45) is the whole of it. A statement of the form "$\alpha_s = 0.118$" is meaningless without naming the scale at which it holds, because Equation (102.44) says the number changes with the scale; and a statement of the form "$\alpha_s = 0.118$ at E_0" contains one dimensionless number and one energy. The combination in Equation (102.45) is invariant under changing E_0: differentiating $\ln \Lambda_{\mathrm{QCD}} = \ln E_0 - 2\pi/\left(b_0 \alpha_s(E_0)\right)$ with respect to $\ln E_0$ gives $1 + \left(2\pi/b_0\alpha_s^2\right) \mathrm{d}\alpha_s/\mathrm{d}\ln E_0 = 1 - 1 = 0$ by Equation (102.42). So the theory has exactly one free parameter in its massless limit, and that parameter is an *energy*. The classical scale invariance of the massless Lagrangian is broken by the quantum theory — it is the trace anomaly, the same phenomenon in another guise — and Λ_{QCD} is what the breaking produces. Numerically, taking $E_0 = m_Z c^2 = 91.188\,\mathrm{GeV}$ and $\alpha_s = 0.1180$ [2427] with $n_f = 5$, so that $b_0 = 11 - \frac{10}{3} = \frac{23}{3}$,

$$\Lambda^{(5)}_{\mathrm{QCD}}\Big|_{\text{one loop}} = 91.188\,\mathrm{GeV} \times \mathrm{e}^{-6.945} = 88\,\mathrm{MeV}\,, \tag{102.46}$$

and at four loops in the $\overline{\mathrm{MS}}$ scheme the same data give about 0.21 GeV for five active flavours and about 0.33 GeV for three [2427]. The factor of two between Equation (102.46) and the four-loop value is not a discrepancy: Λ is defined by a truncation of Equation (102.42) and by a scheme, and changes when either changes, whereas α_s at a stated scale does not. That is why the Particle Data Group quotes $\alpha_s(m_Z c^2)$ and not Λ. What is robust, and what Phenomenon 102.39 asserts, is the order of magnitude and the fact that a scale exists at all. □

Remark 102.40 (The scale that no parameter contains). It is worth pausing on what has happened. The proton mass is $938\,\mathrm{MeV}/c^2$; the quark masses in Table 102.1 that build it sum to $9\,\mathrm{MeV}/c^2$. The remaining 99 % is Λ_{QCD}, and Λ_{QCD} is not a parameter of the theory: it is the scale at which a dimensionless coupling, run down from wherever it was measured, becomes of order one. Had $\alpha_s(m_Z c^2)$ been 0.09 instead of 0.118, Equation (102.45) would have put Λ_{QCD} at a few MeV and nucleons would have been a hundred times lighter and a hundred times larger. The exponential in Equation (102.45) is what converts a modest change in a dimensionless number into an enormous change in a mass, and it is the only mechanism in the Standard Model that generates a scale without being given one.

102.4.2 The measured running of α_s

Phenomenon 102.41 (The strong coupling weakens with energy)**.** Determinations of the strong coupling from τ decays, from lattice computation, from deep inelastic scattering, from hadronic event shapes and from Z-pole observables span more than two decades in momentum transfer and fall on one decreasing curve, with

$$\alpha_s(m_Z c^2) = 0.1180 \pm 0.0009 \tag{102.47}$$

at the Z mass [2427]. The strong interaction is therefore weak at short distance and strong at hadronic distances — the opposite of the behaviour established in Chapter 100, and the reason any perturbative calculation of a strong process is possible at all.

Rests on Equations (102.42) *and* (102.44).

Derivation. *Derives Phenomenon 102.41.*

The curve is Equation (102.44) with Equation (102.42), together with the rule that a quark decouples below its own mass — the decoupling theorem [1538] already used in Proposition 100.63 — so that n_f and hence b_0 change at each flavour threshold and α_s is matched continuously across it. Starting from Equation (102.47) and integrating downwards through the b threshold and upwards through the t threshold with the masses of Table 102.1 gives Table 102.2. The prediction is not one number but the shape of a curve over three decades, and the fact that determinations using entirely different physics land on it is the content of the phenomenon. □

Table 102.2: The one-loop running Equation (102.44) evolved from Equation (102.47), with n_f changing at the quark masses of Table 102.1. The last column is the distance $\hbar c/E$ resolved at that energy. The one-loop truncation is accurate to about one percent above 10 GeV and drifts low by roughly a tenth of its value at the τ mass, where the measured α_s is about 0.31 [2427]; this is the expected size of the $O(\alpha_s^3)$ term of Equation (102.42).

E	n_f	$\alpha_s(E)$ one loop	$\hbar c/E$
1.777 GeV ($m_\tau c^2$)	4	0.279	1.1×10^{-16} m
4.183 GeV ($m_b c^2$)	5	0.212	4.7×10^{-17} m
10 GeV	5	0.173	2.0×10^{-17} m
91.188 GeV ($m_Z c^2$)	5	0.1180	2.2×10^{-18} m
1 TeV	6	0.089	2.0×10^{-19} m

The determinations that must agree come from unrelated physics, and listing them is the point.

- *Hadronic τ decays.* The ratio of the τ's hadronic width to its electronic width is the analogue of R at $E = m_\tau c^2 = 1.777\,\text{GeV}$, and is calculable as a series in α_s using the analyticity of the correlator — a contour argument of the kind established in Chapter 8. It is the lowest-energy determination, and therefore the one with the longest lever arm when evolved to $m_Z c^2$.
- *Lattice QCD.* A short-distance quantity computed on the lattice (Section 102.7.1) and matched to its perturbative expansion determines α_s with no experimental input beyond the hadron masses used to set the scale. This class is currently the most precise.
- *Deep inelastic scattering.* The Q^2 dependence of the structure functions of Section 102.5 is governed by α_s through Section 102.5.2; here α_s is measured by a *slope*, not by a rate.
- *Event shapes and jet rates in e^+e^-.* The three-jet fraction of Section 102.6.2 is proportional to α_s at leading order.
- *Electroweak observables at the Z pole.* The hadronic width of the Z carries the factor $1 + \alpha_s/\pi + \dots$ of Phenomenon 102.15, measured at $E = m_Z c^2$ where no evolution is needed at all.

Remark 102.42 (Why the running is the strongest test of QCD). Any theory with one coupling can be made to fit one measurement of that coupling. What cannot be arranged is that five classes of measurement, at energies differing by a factor of five hundred and using different final states, different detectors and different theoretical machinery, agree after being

evolved by Equation (102.44) with a coefficient b_0 that is not adjustable — it is $11 - \frac{2}{3}n_f$, fixed by Theorem 102.35 from the group and the particle content alone. The agreement of the world data with one curve of predicted shape, rather than the value 0.1180 at one point, is the quantitative evidence for QCD. The general renormalization-group apparatus behind the evolution is the subject of Chapter 108; a systematic survey of the determinations is [2142, 2427].

102.5 Deep inelastic scattering and the parton model

102.5.1 Bjorken scaling and partons

Everything so far has been inference from a spectrum. Deep inelastic scattering is different in kind: it is Rutherford's experiment performed on the proton, and what it sees is not a symmetry but an object.

Notation 102.43 (Kinematics of inelastic lepton scattering). An electron of four-momentum k scatters off a proton of four-momentum P and mass M, emerging with k' and leaving an unobserved hadronic system X. Write $q = k - k'$ for the momentum transferred by the virtual photon and

$$Q^2 := -q^2\,, \qquad \nu := \frac{P\cdot q}{M}\,, \qquad x := \frac{Q^2}{2M\nu} = \frac{Q^2}{2P\cdot q}\,, \qquad y := \frac{P\cdot q}{P\cdot k}\,. \tag{102.48}$$

By Notation 100.1 a four-momentum carries kg m/s, so Q^2 carries $\mathrm{kg}^2\,\mathrm{m}^2/\mathrm{s}^2$ and ν, being $P\cdot q/M$, is an *energy*: in the proton rest frame $\nu = E - E'$, the energy lost by the electron. Both x and y are dimensionless, and $0 \le x \le 1$ because $W^2c^2 := (P+q)^2c^2 \ge M^2c^4$ for any hadronic final state. Elastic scattering is $x = 1$; deep inelastic scattering means $Q^2c^2 \gg M^2c^4$ at fixed x.

Definition 102.44 (Structure functions)**.** Lorentz invariance, current conservation and parity restrict the inclusive cross-section, at one-photon exchange, to two independent functions of the two invariants:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E'} = \sigma_{\mathrm{Mott}}\left[W_2\left(x,Q^2\right) + 2W_1\left(x,Q^2\right)\tan^2\frac{\vartheta}{2}\right]\,, \tag{102.49}$$

with the Mott cross-section for a structureless target

$$\sigma_{\mathrm{Mott}} = \frac{\alpha^2\left(\hbar c\right)^2\cos^2\left(\vartheta/2\right)}{4E^2\sin^4\left(\vartheta/2\right)}\,, \tag{102.50}$$

of dimension m^2, ϑ the electron scattering angle and E its incident energy. The dimensionless structure functions are

$$F_1\left(x,Q^2\right) := Mc^2W_1\,, \qquad F_2\left(x,Q^2\right) := \nu W_2\,, \tag{102.51}$$

W_1 and W_2 each carrying $1/\mathrm{J}$.

Phenomenon 102.45 (The proton contains pointlike constituents)**.** In inelastic electron–proton scattering at large momentum transfer the cross-section falls far more slowly with Q^2 than the elastic one does, and the structure functions depend, to a good first approximation, only on the dimensionless combination $x = Q^2/2M\nu$ — with Q the momentum transferred, ν the energy transferred and M the proton mass — and not on Q^2 separately [1395, 1397], exactly as Bjorken had predicted from current algebra before the data existed [1393]. The two structure functions are moreover not independent:

$$F_2(x) = 2xF_1(x)\,, \tag{102.52}$$

which the data obey [2427]. Scaling says the target contains constituents that carry no scale of their own; Equation (102.52) says those constituents carry spin $\frac{1}{2}$. The experiment is Chapter 116.

Rests on Definition 102.44 and Equation (102.50).

Derivation. *Derives Phenomenon 102.45.*

Scaling. Suppose the photon, at large Q^2, resolves a time short compared with the time over which the proton's constituents interact with one another, so that it scatters elastically off one of them and the rest are spectators. This is the impulse approximation, and Feynman's name for the constituents is *partons* [1401]. Let the struck parton carry a fraction ξ of the proton's four-momentum, $p = \xi P$, and let it remain on its mass shell after the collision, its mass being ξM in the same approximation. Then

$$(\xi P + q)^2 = \xi^2 M^2 c^2 \quad \Longrightarrow \quad \xi^2 M^2 c^2 + 2\xi\, P\cdot q - Q^2 = \xi^2 M^2 c^2 \quad \Longrightarrow \quad \xi = \frac{Q^2}{2P\cdot q} = x\,. \tag{102.53}$$

So the measured variable x *is* the momentum fraction of the struck constituent: this is what makes x physical rather than a convenient abbreviation. The cross-section is then the incoherent sum over constituents of pointlike elastic cross-sections, each of which is the Mott formula Equation (102.50) times the squared charge, and the only place a scale could enter — a form factor, i.e. a size — is absent by hypothesis. Hence

$$F_2(x) = \sum_q Q_q^2\, x\,[q(x) + \bar{q}(x)]\ , \tag{102.54}$$

with $q(x)\mathrm{d}x$ the number of quarks of flavour q carrying momentum fraction in $\mathrm{d}x$: a function of x alone. Any structure of the constituents themselves would reintroduce a length and destroy the scaling, so scaling is the statement that the constituents are pointlike down to the resolution $\hbar c/Qc$ reached — about $1\times10^{-16}\,\mathrm{m}$ at the original SLAC energies, a tenth of the proton radius $8.4075\times10^{-16}\,\mathrm{m}$ [2431]. The contrast is with elastic scattering, whose cross-section carries the squared dipole form factor and therefore falls as Q^{-12}: the same beam sees a soft object elastically and hard objects inelastically, which is precisely Rutherford's situation.

The Callan–Gross relation. Decompose the virtual photon into its three polarization states in the frame where it moves along $+z$ and the struck parton along $-z$. A massless spin-$\frac{1}{2}$ parton coupling through the vector current $\bar{q}\gamma^\mu q$ conserves helicity, because γ^μ commutes with γ^5 up to a sign that leaves the chiral projectors intact and, for a massless fermion, chirality is helicity. After absorbing the photon the parton moves along $+z$; its helicity is unchanged, so its *spin projection on the z axis* has flipped, $\Delta J_z = \pm1$. Only a transverse photon, $\lambda = \pm1$, can supply that. The longitudinal photon, $\lambda = 0$, therefore cannot be absorbed at all:

$$\sigma_L = 0 \qquad \text{for spin-}\tfrac{1}{2}\text{ partons.} \tag{102.55}$$

Expressing the two cross-sections in terms of the structure functions, the transverse one is proportional to F_1 and the longitudinal to the combination $F_L := F_2 - 2xF_1$ (up to terms of order M^2c^4/Q^2c^2, negligible in the deep inelastic region), so Equation (102.55) is exactly Equation (102.52) [1398]. A spin-0 parton gives the opposite conclusion by the same argument — it has no spin to flip, so only $\lambda = 0$ is absorbed, $\sigma_T = 0$ and $F_1 = 0$ — and the data exclude it decisively: the measured ratio $R_L = \sigma_L/\sigma_T$ is small, as QCD predicts; the experimental write-up belongs to Section 116.4.3. □

Proposition 102.46 (The partons have fractional charge)**.** *Comparing charged-lepton and neutrino deep inelastic scattering on an isoscalar target measures the mean square charge of*

the constituents directly:

$$\frac{F_2^{eN}(x)}{F_2^{\nu N}(x)} = \frac{1}{2}\left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = \frac{5}{18}\,, \tag{102.56}$$

which the data confirm [2427].

Rests on Definition 102.5 and Equation (102.54).

Proof. *Derives Proposition 102.46.*

The photon couples to a quark with strength $Q_q e$, so Equation (102.54) weights each flavour by Q_q^2. The charged weak current of Chapter 101 couples with a strength independent of electric charge, so the corresponding neutrino structure function is $F_2^{\nu N}(x) = x\sum_q [q(x) + \bar{q}(x)]$ with unit weights. On an isoscalar target the u and d distributions enter symmetrically, and the ratio of the two sums is the mean of $(2/3)^2$ and $(1/3)^2$, which is Equation (102.56). Had the constituents carried integer charges the ratio would have been of order unity instead of 0.28. This is a measurement of the charges themselves, independent of the spectroscopy of Section 102.1.3, and it is why neutrino scattering was decisive. □

Phenomenon 102.47 (Half the proton's momentum is electrically neutral)**.** Integrating the measured F_2 over x and dividing out the charge weights gives the total momentum fraction carried by charged constituents,

$$\int_0^1 \mathrm{d}x\; x \sum_q \left[q(x) + \bar{q}(x)\right] \approx 0.5\,, \tag{102.57}$$

so that about half the proton's momentum is carried by something that does not couple to the photon or to the weak current [2427]. The natural candidate is the gauge field itself.

Rests on Equations (102.53), (102.54) and (102.56).

Derivation. *Derives Phenomenon 102.47.*

The momentum of the proton is shared among its constituents by definition, $\int_0^1 \mathrm{d}x\, x \sum_{\text{all}} f_i(x) = 1$, since x is the momentum fraction (Equation (102.53)). The measured left-hand side of Equation (102.57) is obtained from Equation (102.54) together with Equation (102.56), the two providing enough information to separate the charge weights from the distributions; the details of the extraction belong in Section 116.5.2. The deficit is a fact about the data and not an inference: whatever carries the missing half is electrically neutral and weakly inert. In QCD it is the gluon, which appears here for the first time as an experimental necessity rather than a theoretical convenience — and it appeared this way, in the momentum sum rule, before it was seen in the three-jet events of Section 102.6.2. □

102.5.2 Scaling violation and the DGLAP equations

Scaling is not exact, and QCD says exactly how it fails. A quark is not free even at short distance: it can radiate a gluon, and the probability of doing so is not small because the emission is logarithmically enhanced when the gluon is soft or collinear — the same enhancement as the soft photon of Section 100.6.1. Raising Q^2 means resolving a smaller transverse distance, hence seeing emissions that were previously unresolved, hence finding more partons, each carrying less momentum. Structure functions therefore drift: down at large x, up at small x, logarithmically in Q^2.

Proposition 102.48 (Collinear emission and its logarithm)**.** *The probability that a quark of momentum fraction y radiates, leaving a quark of fraction $x = zy$, is*

$$\mathrm{d}\mathcal{P} = \frac{\alpha_s}{2\pi}\, P_{qq}(z)\,\mathrm{d}z\,\frac{\mathrm{d}k_T^2}{k_T^2}\,, \qquad P_{qq}(z) = C_F\left[\frac{1+z^2}{1-z}\right]_+\,, \tag{102.58}$$

with k_T the transverse momentum of the emitted gluon, $C_F = 4/3$ from Equation (102.31), and the plus prescription defined by $\int_0^1 \mathrm{d}z\,[g(z)]_+ f(z) = \int_0^1 \mathrm{d}z\, g(z)[f(z) - f(1)]$. Integrating $\mathrm{d}k_T^2/k_T^2$ up to the resolution Q^2 produces $\ln Q^2$, which is the entire origin of the scaling violation.

Rests on Equations (100.76) and (102.31) and Theorem 100.60.

Proof. *Derives Proposition 102.48.*

Two features are derived here and the third is quoted. The factor $\mathrm{d}k_T^2/k_T^2$ is the collinear singularity of a massless emitter: the propagator of the quark after emission has virtuality $(p-k)^2 \propto -k_T^2/(1-z)$, and the squared amplitude carries its inverse square against a two-body phase space $\propto \mathrm{d}k_T^2$, leaving $\mathrm{d}k_T^2/k_T^2$. It is logarithmic precisely because the emitter is massless, and it is the mass singularity of Theorem 100.60 in its QCD form. The factor $1/(1-z)$ is the soft singularity: as $z \to 1$ the gluon energy goes to zero and the amplitude reduces to the eikonal current Equation (100.76), whose square gives exactly this pole, with C_F replacing the squared electric charge because the emission vertex is T_a and $T_a T_a = C_F \mathbb{1}$. The numerator $1 + z^2$ is the spin structure of the vertex, the two terms being the helicity-conserving and helicity-flip configurations of the emitting quark; it is computed from the same trace technology as Section 100.3.2 [1474, 1591].

The plus prescription is not an extra assumption but a normalization forced by conservation of quark number. The number of quarks of a given flavour minus antiquarks is a conserved charge and cannot change with Q^2, so $\int_0^1 \mathrm{d}z\, P_{qq}(z) = 0$; the unregulated $1/(1-z)$ is not integrable, and the divergence is cancelled by the virtual correction, which lives entirely at $z = 1$. The plus prescription is exactly that cancellation written as a distribution: it subtracts, at $z = 1$, whatever the real emission supplies. Equivalently $[(1+z^2)/(1-z)]_+ = (1+z^2)\,[1/(1-z)]_+ + \frac{3}{2}\delta(1-z)$, and the $\frac{3}{2}$ is the virtual term. □

Theorem 102.49 (The DGLAP evolution equations)**.** *The parton distributions depend on the scale at which they are defined, and the dependence is*

$$\frac{\partial q_i(x,Q^2)}{\partial \ln Q^2} = \frac{\alpha_s(Q^2)}{2\pi} \int_x^1 \frac{\mathrm{d}y}{y} \left[P_{qq}\left(\frac{x}{y}\right) q_i(y,Q^2) + P_{qg}\left(\frac{x}{y}\right) g(y,Q^2) \right], \tag{102.59}$$

together with the companion equation for the gluon distribution involving P_{gq} and P_{gg} [1474, 1591, 1595]. The right-hand side is a multiplicative convolution on $(0,1]$ in the sense of Definition 17.97.

Rests on Proposition 102.48, Equation (102.58), and Definition 17.97.

Proof. *Derives Theorem 102.49.*

A parton of fraction x at resolution $Q^2 + \mathrm{d}Q^2$ is either one that was already there at Q^2, or one produced by an emission from a parton of larger fraction $y > x$ that the coarser resolution could not separate. Multiplying Equation (102.58) by the density of emitters $q_i(y)$, changing variables from z to $x = zy$ at fixed y so that $\mathrm{d}z = \mathrm{d}x/y$, and integrating $\mathrm{d}k_T^2/k_T^2$ over the shell between the two resolutions gives the first term of Equation (102.59). The second term is the same statement for a gluon converting to a quark–antiquark pair, with $P_{qg}(z) = T_F\,[z^2 + (1-z)^2]$. That the kernel depends on x and y only through x/y is a consequence of the collinear factorization: the emission probability knows only the momentum fraction taken, not the absolute momenta. Comparing with Equation (102.63) below identifies the integral as $(P \star q)(x)$ with the Haar measure $\mathrm{d}y/y$ of Definition 17.97, which is the structure that Corollary 102.50 exploits. □

Corollary 102.50 (Moments diagonalise the evolution)**.** *Define the moments of a distribution as in Equation (17.114), $M_N[f] = \int_0^1 f(x) x^N \mathrm{d}x$, so that $N = 0$ counts partons and $N = 1$ is the momentum fraction. For a non-singlet combination — one in which the gluon*

term of Equation (102.59) cancels, for instance $u-d$ or $q-\bar q$ — the evolution becomes an ordinary differential equation for each moment separately,

$$\frac{\mathrm{d}M_N}{\mathrm{d}\ln Q^2}=\frac{\alpha_s(Q^2)}{2\pi}\,\gamma_N\,M_N\,,\qquad \gamma_N:=\int_0^1 \mathrm{d}z\; z^N P_{qq}(z)\,, \tag{102.60}$$

whose solution, using Equation (102.42), is a pure power of the coupling:

$$\frac{M_N\left(Q^2\right)}{M_N\left(Q_0^2\right)}=\left[\frac{\alpha_s(Q^2)}{\alpha_s(Q_0^2)}\right]^{-2\gamma_N/b_0}\,. \tag{102.61}$$

The anomalous dimensions are

$$\gamma_N=C_F\left[\frac{3}{2}-H_N-H_{N+2}\right]\,,\qquad H_m:=\sum_{k=1}^{m}\frac{1}{k}\,, \tag{102.62}$$

with $\gamma_0=0$ and $\gamma_1=-\frac{4}{3}C_F=-\frac{16}{9}$.

Rests on Theorem 102.49, Corollary 17.99, and Equation (102.42).

Proof. *Derives Corollary 102.50.*

Write the convolution of Theorem 102.49 in the form

$$(P\star q)\,(x)=\int_x^1 P(y)\,q\!\left(\frac{x}{y}\right)\frac{\mathrm{d}y}{y}\,, \tag{102.63}$$

which is Equation (17.113): both factors vanish for argument greater than one, so the range is $x\le y\le 1$ and the convolution stays supported on the unit interval. By Corollary 17.99, and ultimately by the Mellin convolution theorem Theorem 17.98, the moments of a multiplicative convolution are the products of the moments, Equation (17.115). Taking M_N of Equation (102.59) therefore turns the integral operator into multiplication by $\gamma_N=M_N[P_{qq}]$, giving Equation (102.60). This is the whole reason the Mellin transform is the right instrument here: it diagonalises dilations, and Equation (102.59) is an equation about how momentum fractions dilate.

To solve, divide Equation (102.60) by Equation (102.42) rewritten as $\mathrm{d}\alpha_s/\mathrm{d}\ln Q^2=-b_0\alpha_s^2/4\pi$ — the factor $\frac{1}{2}$ relative to Equation (102.42) because $\ln Q^2=2\ln Q$. Then $\mathrm{d}\ln M_N/\mathrm{d}\alpha_s=-2\gamma_N/(b_0\alpha_s)$, which integrates to $M_N\propto\alpha_s^{-2\gamma_N/b_0}$ and hence to Equation (102.61).

For Equation (102.62), use the plus prescription:

$$\gamma_N=C_F\int_0^1\mathrm{d}z\,\frac{\left(z^N-1\right)\left(1+z^2\right)}{1-z}\,,$$

the subtraction at $z=1$ making the integrand finite. Split the numerator: $\left(z^N-1\right)+\left(z^{N+2}-z^2\right)$, and write the second bracket as $\left(z^{N+2}-1\right)-\left(z^2-1\right)$. With $\int_0^1\left(z^m-1\right)/(1-z)\,\mathrm{d}z=-H_m$ — immediate from $\left(1-z^m\right)/(1-z)=1+z+\cdots+z^{m-1}$ integrated term by term — this gives $\gamma_N=C_F\left[-H_N-H_{N+2}+H_2\right]$ and $H_2=\frac{3}{2}$, which is Equation (102.62). Checking: $\gamma_0=C_F[\frac{3}{2}-0-\frac{3}{2}]=0$, as quark-number conservation demands, and $\gamma_1=C_F[\frac{3}{2}-1-\frac{11}{6}]=-\frac{4}{3}C_F$. □

Example 102.51 (How fast a moment moves)**.** Take the momentum moment $N=1$ of a non-singlet distribution and evolve from $Qc=10\,\mathrm{GeV}$ to $Qc=100\,\mathrm{GeV}$ with $n_f=5$, so $b_0=23/3$ and the exponent in Equation (102.61) is $-2\gamma_1/b_0=(32/9)\,(3/23)=0.464$. From Table 102.2, α_s falls from 0.173 to 0.116 over that range, a ratio 0.673, so

$$\frac{M_1\left(Q^2\right)}{M_1\left(Q_0^2\right)}=0.673^{0.464}=0.83\,. \tag{102.64}$$

The momentum carried by the valence quarks falls by about a sixth over one decade in Q: a large, slow, logarithmic effect, exactly what distinguishes a scaling violation from a scaling failure.

Remark 102.52 (The prediction is a shape, not a number). QCD does not predict $q(x, Q_0^2)$: the distributions at one scale are non-perturbative and must be measured. What it predicts is their entire Q^2 dependence, with no free parameter beyond α_s. Equation (102.61) makes the point sharply: the ratio of two moments' evolution,

$$\frac{\ln\left[M_N(Q^2)/M_N(Q_0^2)\right]}{\ln\left[M_M(Q^2)/M_M(Q_0^2)\right]} = \frac{\gamma_N}{\gamma_M}\,, \tag{102.65}$$

contains neither α_s nor Λ_{QCD} nor the initial distributions — only the two rational numbers of Equation (102.62). Plotting one moment against another on logarithmic axes must give a straight line of predicted slope, and it does. Global fits to data spanning four orders of magnitude in Q^2 and five in x are described by Equation (102.59) with one coupling; this is the most demanding quantitative test QCD has passed, and its experimental side belongs to Section 116.6.4.

102.5.3 Factorization and parton distributions

Equation (102.54) and Equation (102.59) together contain a claim that goes far beyond deep inelastic scattering: that a hadronic cross-section splits into a part depending on the hard process, which is calculable, and a part depending on the hadron, which is not but is *universal*.

Theorem 102.53 (Collinear factorization)**.** *For an inclusive process with a large scale Q, the cross-section can be written to leading power in Λ_{QCD}/Q as*

$$\sigma = \sum_{i,j} \int \mathrm{d}x_1 \mathrm{d}x_2 \; f_i\left(x_1, \mu_F^2\right) f_j\left(x_2, \mu_F^2\right) \hat{\sigma}_{ij}\left(x_1, x_2, Q^2, \mu_F^2, \alpha_s\right)\,, \tag{102.66}$$

where $\hat{\sigma}_{ij}$ is a partonic cross-section computable as a series in α_s and the f_i are parton distributions that depend on the hadron but not on the process. The arbitrary scale μ_F separating the two is a factorization scale; its μ_F-dependence cancels between the two factors order by order, and $\partial f_i/\partial \ln \mu_F^2$ is Equation (102.59).

Rests on Equations (102.54) and (102.59).

Derivation pending #149

Collinear factorization to all orders. What is available inline is the leading-order statement, which is the impulse approximation together with the observation that the collinear logarithms are universal because they come from the region where the emitted parton is nearly on shell and therefore long-lived compared with the hard collision. The all-orders proof requires showing that the leading regions of the loop momenta are exhausted by hard, collinear and soft subgraphs, that the soft gluons cancel or are absorbed into eikonal lines, and that what remains is the same operator matrix element in every process. That analysis, and the counterexamples at subleading power, belong in Appendix A.

The physical content is that the distributions may be *measured in one process and used in another*. They are extracted from deep inelastic scattering on fixed targets and at HERA, and then used — with the evolution Equation (102.59) carrying them from $Q \sim 10\,\mathrm{GeV}/c$ to $Q \sim 1\,\mathrm{TeV}/c$ — to predict cross-sections at hadron colliders. Every Standard Model rate

measured at the LHC, the Higgs production of Chapter 113 included, is computed with Equation (102.66) and would be unpredictable without it. That the same $f_i(x,\mu_F^2)$ describe electron–proton scattering at 10 GeV and proton–proton collisions at 13 TeV is a test of universality across three decades in scale and two different initial states, and it is the reason Theorem 102.53 is stated as a theorem rather than as a modelling assumption [1817].

Remark 102.54 (What the factorization scale is not). μ_F is not a physical boundary between "inside" and "outside" the proton. It is a subtraction point, exactly like the renormalization scale of Definition 100.54: emissions harder than μ_F are put into $\hat{\sigma}$ and those softer into f_i, and the split is a bookkeeping choice. A complete calculation is μ_F-independent; a calculation truncated at order α_s^n has a residual dependence of order α_s^{n+1}, and varying μ_F by a factor of two is the standard — and admittedly crude — estimate of the missing higher orders. A parton distribution is therefore not a probability density in any scheme-independent sense; it is a scheme-dependent object, and quoting one without its scheme and scale is meaningless.

102.6 Jets and the gluon

102.6.1 Two-jet events and the quark spin

Deep inelastic scattering sees the constituents by their recoil. Electron–positron annihilation makes a pair of them directly, and what the detector records is the closest thing to a photograph of a quark that confinement permits.

Proposition 102.55 (Angular distribution of a produced fermion pair). *For $e^+e^-\to f\bar{f}$ through one virtual photon, with all masses negligible,*

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}=\frac{\alpha^2\,(\hbar c)^2}{4E_{\mathrm{cm}}^2}\,Q_f^2\,N_c^{(f)}\left(1+\cos^2\vartheta\right), \tag{102.67}$$

ϑ being the angle between the outgoing fermion and the beam. A produced spin-0 pair would give $\sin^2\vartheta$ instead.

Rests on Equations (100.32) and (100.40) and Definition 102.10.

Proof. *Derives Proposition 102.55.*

Section 100.3.1 gives the spin-averaged squared amplitude for this process as $\overline{|\mathcal{M}|^2}=32\pi^2\alpha^2\hbar^4\,(t^2+u^2)\,/s^2$. For massless kinematics in the centre-of-mass frame, $t=-\tfrac{1}{2}s\,(1-\cos\vartheta)$ and $u=-\tfrac{1}{2}s\,(1+\cos\vartheta)$, so

$$\frac{t^2+u^2}{s^2}=\frac{1}{4}\left[(1-\cos\vartheta)^2+(1+\cos\vartheta)^2\right]=\frac{1+\cos^2\vartheta}{2}\,.$$

Inserting this into the two-body phase space of Equation (100.40) and multiplying by Q_f^2 for the charge and by N_c for the colour multiplicity gives Equation (102.67); integrating over angles returns Equation (100.32) times the same factors, which is Equation (102.15). The $\sin^2\vartheta$ for scalars follows because a scalar pair produced by a vector current must be in a p-wave with $J_z=0$ along the pair axis, and the corresponding Wigner function is $\sin\vartheta$. □

Phenomenon 102.56 (Hadrons emerge as jets, and some events have three). High-energy e^+e^- annihilation does not produce hadrons isotropically. Most events consist of two back-to-back collimated sprays whose common axis follows the $1+\cos^2\vartheta$ distribution characteristic of a produced spin-$\tfrac{1}{2}$ pair [1548], and a calculable minority are planar events with three distinct sprays [1634]. The energy sharing and the angular correlations of the third spray are those of radiation emitted by a coloured source into a massless spin-one quantum. This is the only direct evidence for the gluon, and it is direct only in the sense that a jet, not a parton, is what a detector records.

Rests on Equations (102.58) and (102.67).

Derivation of the two-jet structure. *Derives Phenomenon 102.56.*

Two facts make a jet. First, the parton pair is produced back to back with energy $E_{\mathrm{cm}}/2$ each, and the angular distribution of the pair axis is Equation (102.67) — an observable that survives hadronization because the subsequent radiation is soft. Second, the radiation *is* soft and collinear: by Equation (102.58) the emission probability is concentrated at small k_T, so the hadrons that materialize share the parton's direction to within an angle of order Λ_{QCD}/E. The transverse momentum inside a jet is therefore bounded by a hadronic scale while its longitudinal momentum grows with the beam energy, and the spray narrows as $1/E$. Hanson and collaborators established this at SPEAR by measuring the sphericity tensor and showing that its principal axis is distributed as $1+\cos^2\vartheta$ [1548]: the axis of a spray of hadrons remembers the spin of a particle that was never observed. □

Remark 102.57 (What a jet has to be to be calculable). A jet is defined by an algorithm, and not every algorithm defines a calculable quantity. Theorem 100.58 and Theorem 100.60 say that infrared and collinear divergences cancel between real and virtual corrections only for observables that are *inclusive* over soft emission and over collinear splitting. An observable is called infrared and collinear safe if it is unchanged by adding a zero-energy particle and by replacing any particle with two collinear ones carrying its total momentum [1611]. The jet algorithms in use are constructed to have that property, and quantities computed from algorithms that lack it are not merely imprecise but divergent order by order. This is the practical content of Theorem 100.60 in QCD, and it is why the comparison of Phenomenon 102.56 with theory compares *jet* rates and never parton rates.

102.6.2 Three-jet events: the gluon

Proposition 102.58 (The three-jet energy distribution). *Let $x_i = 2E_i/E_{\mathrm{cm}}$ be the scaled energies of the quark, antiquark and gluon, so that $x_1+x_2+x_3=2$. Then to first order in α_s*

$$\frac{1}{\sigma_0}\frac{\mathrm{d}^2\sigma}{\mathrm{d}x_1\,\mathrm{d}x_2} = \frac{\alpha_s}{2\pi}\,C_F\,\frac{x_1^2+x_2^2}{(1-x_1)\,(1-x_2)} = \frac{2\alpha_s}{3\pi}\,\frac{x_1^2+x_2^2}{(1-x_1)\,(1-x_2)}\,, \tag{102.68}$$

with σ_0 the two-jet cross-section Equation (102.67) integrated over angles [1668]. Had the emitted quantum been a scalar rather than a vector, the numerator would have been x_3^2.

Rests on Proposition 102.48 and Equations (100.76) and (102.31).

Derivation of the singular structure. *Derives Proposition 102.58.*

The two limits that dominate the rate are derivable from results already established, and they already distinguish a vector from a scalar.

Soft. As $x_3\to 0$ both x_1 and x_2 tend to 1, and the amplitude factorizes into the two-jet amplitude times the eikonal current Equation (100.76) of Section 100.6.1, with the electric charge replaced by g_sT_a and the colour factor $T_aT_a=C_F$ of Equation (102.31). Squaring gives the double pole $[(1-x_1)(1-x_2)]^{-1}$ of Equation (102.68) with numerator 2, which is $x_1^2+x_2^2$ at $x_1=x_2=1$. A scalar emitter has no such factorization: a soft scalar does not couple to the eikonal current at all, because the coupling to a fast fermion line is suppressed by the emitted energy, and $x_3^2\to 0$ in the same limit. The observed events pile up at the soft edge, and that alone excludes a scalar.

Collinear. As $x_1\to 1$ at fixed x_2 the gluon becomes collinear with the antiquark, and Equation (102.68) reduces to $(\alpha_s/2\pi)\,C_F\,(1+x_2^2)\,/\,[(1-x_1)(1-x_2)]$, whose x_2 dependence is the splitting function Equation (102.58) with $z=x_2$ — as it must be, since Proposition 102.48 and Equation (102.68) are two views of the same emission. That the same function $C_F(1+z^2)/(1-z)$ governs scaling violation in Section 102.5.2 and the three-jet rate here is a consistency requirement of the theory and is tested by the agreement of the two determinations of α_s in Section 102.4.2. □

Derivation pending #150

The exact order-α_s three-jet distribution. What is derived inline is the soft and collinear structure, which is where the rate is concentrated and which already excludes a scalar gluon. The full matrix element requires the spin-summed trace over the two interfering diagrams in which the gluon is emitted from the quark and from the antiquark, together with the three-body phase space in the Dalitz variables; and the corresponding statement for a scalar emitter, needed to make the exclusion quantitative rather than qualitative. Both are trace computations of the kind carried out for the electromagnetic processes and belong beside them in Appendix A.

The experimental history is short and decisive. Planar three-jet events were reported by TASSO at PETRA in 1979 [1634] and confirmed within months by the other three experiments at the same machine. The interpretation rests on more than the existence of a third spray: the distribution of energy between the three jets follows Equation (102.68), and the orientation of the three-jet plane relative to the beam — and of the jets within it — is that of a vector quantum and not a scalar. The measurement that established this is the distribution in the Ellis–Karliner angle, the angle between the two most energetic jets in the rest frame of the two least energetic, which is sensitive to the spin of the radiated quantum and insensitive to almost everything else.

Remark 102.59 (Testing the non-abelian structure itself). The existence of a spin-one mediator is not yet the existence of a *non-abelian* one. What distinguishes QCD from an abelian theory of coloured quarks and a neutral gluon is the triple-gluon vertex of Equation (102.28), and that vertex first contributes to a four-jet final state, where a gluon splits into two gluons. The angular correlations among four jets — notably the angle between the planes defined by the two hardest and the two softest jets — depend on the ratio C_A/C_F, which the group theory of Equation (102.31) fixes at 9/4. The LEP experiments measured C_A and C_F from four-jet rates and correlations and found values consistent with 3 and 4/3 at the ten-percent level [2427], with the abelian alternative ($C_A = 0$) excluded by many standard deviations. This is the closest thing to a direct measurement of the gluon self-coupling, and therefore of the mechanism that Theorem 102.35 says produces asymptotic freedom.

102.7 Confinement

102.7.1 Lattice gauge theory

Everything in Sections 102.4 and 102.5 was perturbative and therefore restricted to short distance. Confinement is a long-distance statement, and the only systematic method for it is Wilson's [1537]: define the theory on a Euclidean lattice, where the functional integral is a finite-dimensional ordinary integral, and evaluate it. The construction is carried out in Section 106.5.1 and is not repeated here; what is repeated is what it buys and what it does not.

Three of its results are used throughout this chapter. The link variables Equation (106.79) put the gauge field in the *group* rather than the algebra, so gauge invariance is exact at finite spacing and there is no gauge fixing and no ghost (Proposition 106.67). The Wilson action Equation (106.81) reduces to Equation (102.27) as the spacing goes to zero (Theorem 106.68), with lattice coupling $\beta = 6/\hat{g}^2 = 6/(4\pi\alpha_s)$ for SU(3). And the Wilson loop Definition 106.70 measures the static potential, an area law being equivalent to a linearly rising $V(R)$ (Proposition 106.71).

Theorem 102.60 (Confinement at strong coupling, and what it proves)**.** *For $\beta \ll 1$ the Wilson loop obeys an area law with string tension $\sigma = (\hbar c/a^2)\ln(2N^2/\beta)$, which is Equation* (106.84)*. This is an exact statement about the lattice theory at strong bare coupling. It is* not *a proof of confinement in QCD.*

Rests on Theorems 100.25 and 106.72 and Equation (106.84)*.*

Proof. *Derives Theorem 102.60.*

The area law itself is Theorem 106.72. The second sentence is the point at issue and is established by a counterexample. Compact U(1) lattice gauge theory — lattice electrodynamics — obeys the same strong-coupling area law, by the same plaquette-tiling argument, since that argument uses only the group integration formula $\int \mathrm{d}U\, U = 0$ and never the non-abelian structure. But continuum electrodynamics manifestly does not confine: the Coulomb potential of Theorem 100.25 is measured every day. The resolution is that compact U(1) in four dimensions has a genuine phase transition at intermediate β separating a confining strong-coupling phase from a Coulomb weak-coupling phase, and the continuum limit is taken in the latter. Hence a strong-coupling area law establishes nothing about the continuum theory unless one also knows that no such transition intervenes between $\beta \ll 1$ and the continuum limit $\beta \to \infty$ demanded by Remark 106.69. For SU(3) the absence of a transition is a numerical finding, not a theorem. □

The first such demonstration is Creutz's, for the gauge group SU(2) [1651]. Computing the string tension by Monte Carlo across the region where the strong-coupling expansion fails, he found no discontinuity, and found in addition that σa^2 falls with β at the rate that asymptotic freedom demands for that group — *asymptotic scaling*: the lattice spacing shrinks with the bare coupling exactly as the running of the coupling requires, so that σ expressed in physical units is β-independent. That is the statement that the lattice theory and the asymptotically free continuum theory are the same theory. It is a numerical demonstration to finite accuracy. For SU(3) the same continuity and scaling have since been established on the lattice, most sharply by the spectroscopy computations of [2122], and that is what the honest claim in Section 102.7.3 rests on.

Remark 102.61 (Lattice QCD as a source of numbers)*.* Since Creutz the method has become quantitative. With improved actions, physical quark masses, several lattice spacings extrapolated to zero and several volumes extrapolated to infinity, lattice QCD computes: the light hadron spectrum to about one percent [2122]; the quark masses of Table 102.1, which are *defined* by the lattice or by sum rules since no quark is ever isolated; α_s, currently the most precise determination entering Equation (102.47); decay constants and form factors needed to extract the quark mixing parameters of Chapter 103; the equation of state of hot matter (Section 102.9); and the hadronic vacuum polarization entering the muon anomaly [2375]. What it does not do is prove anything: it is a controlled numerical evaluation of a functional integral with quantified statistical and systematic uncertainties, which is a different epistemic object from a theorem and is treated as such throughout this chapter. Its relationship to the statistical mechanics of Chapter 118 is exact — the Euclidean weight $\mathrm{e}^{-S_E/\hbar}$ is a Boltzmann factor — and the continuum limit is a critical point in the sense of Chapter 123.

102.7.2 The static potential and hadron masses

Phenomenon 102.62 (The static potential rises linearly)**.** The energy of a static colour source and its anticolour separated by r is, over the range $0.2 \times 10^{-15}\,\mathrm{m} \lesssim r \lesssim 1 \times 10^{-15}\,\mathrm{m}$ accessible both to lattice computation and to quarkonium spectroscopy,

$$V(r) = -\frac{C_F\,\alpha_s \hbar c}{r} + \sigma r\,, \qquad \sigma \approx 0.9\,\mathrm{GeV/fm} = 1.4 \times 10^5\,\mathrm{N}\,, \tag{102.69}$$

the first term being the short-distance Coulomb-like exchange with the colour factor $C_F = 4/3$ of Equation (102.31) and the second the confining term. The string tension σ is a force, and it is about fourteen tonnes weight.

Rests on Equations (14.77) *and* (102.31) *and Proposition 106.71.*

Derivation of the two terms and a numerical test. *Derives Phenomenon 102.62.*

The Coulomb term is one-gluon exchange between a colour source in $\mathbf{3}$ and one in $\bar{\mathbf{3}}$ coupled to a singlet. The colour factor is $\langle T_a \otimes T_a \rangle_{\mathbf{1}} = \frac{1}{2}\left[C_2(\mathbf{1}) - 2C_2(\mathbf{3})\right] = -C_F$, using $C_2(\mathbf{1}) = 0$ and Equation (14.77); the sign is attraction, and the magnitude is $4/3$ times the electromagnetic case with $\alpha \to \alpha_s$. Note that the same computation for a $\mathbf{3} \otimes \mathbf{3}$ pair coupled to $\bar{\mathbf{3}}$ gives $-C_F/2$, attractive but weaker, and for the symmetric $\mathbf{6}$ gives $+C_F/4$, repulsive: colour singlets are the most tightly bound configurations available, which is Postulate 102.11 made plausible.

The linear term is not derived here — it is measured on the lattice and inferred from the spectrum — but it has a check that costs one page and is worth doing, because it connects Equation (102.69) to a completely independent body of data. Model the confining flux as a relativistic string of tension σ with massless ends rotating rigidly at the speed of light. If the string has total length $2R$, the transverse speed at distance r from the centre is $v = cr/R$, and

$$E = 2\int_0^R \frac{\sigma \, \mathrm{d}r}{\sqrt{1 - r^2/R^2}} = \pi\sigma R\,,$$
$$J = \frac{2}{c^2}\int_0^R \frac{\sigma \, v \, r \, \mathrm{d}r}{\sqrt{1 - r^2/R^2}} = \frac{2\sigma}{cR}\int_0^R \frac{r^2 \mathrm{d}r}{\sqrt{1 - r^2/R^2}} = \frac{\pi\sigma R^2}{2c}\,, \tag{102.70}$$

using $\int_0^1 u^2(1-u^2)^{-1/2}\mathrm{d}u = \pi/4$. Eliminating R and writing $E = Mc^2$ and $J = \hbar j$,

$$j = \frac{M^2 c^3}{2\pi\sigma\hbar}\,, \tag{102.71}$$

a linear *Regge trajectory*: spin proportional to squared mass, with a slope fixed by the string tension alone. The name records Regge's poles in complex angular momentum [1174]; that the observed hadronic trajectories are in fact linear is an empirical statement about the spectrum, tested on the ρ data below [2427]. Numerically, $\sigma = 1.442 \times 10^5\,\mathrm{N}$ gives $j = 0.896$ for $Mc^2 = 1\,\mathrm{GeV}$: by Equation (102.71) the slope of j in the squared rest energy $(Mc^2)^2$ is $1/(2\pi\sigma\hbar c) = 3.49 \times 10^{19}/\mathrm{J}^2$, that is $0.896/\mathrm{GeV}^2$ in the customary units. The measured ρ trajectory — $\rho(775)$ with $j = 1$, $a_2(1320)$ with $j = 2$, $\rho_3(1690)$ with $j = 3$ [2427] — has slope $0.882/\mathrm{GeV}^2$ between the first two states and $0.895/\mathrm{GeV}^2$ between the second and third. Agreement to about one percent, from a string tension measured in an entirely different way, and with no adjustable parameter: the linearity of Equation (102.69) and the linearity of the observed meson trajectories are the same fact. □

Remark 102.63 (Quarkonium as the short-distance probe). Charmonium and bottomonium are non-relativistic bound states of heavy quarks and therefore sample Equation (102.69) at small r, where the Coulomb term dominates. Their spectra test the potential directly. The signature fact is that the spacing between the ground state and the first radial excitation is nearly the same in the two systems — $0.589\,\mathrm{GeV}/c^2$ for $\psi(2S) - J/\psi(1S)$ and $0.563\,\mathrm{GeV}/c^2$ for $\Upsilon(2S) - \Upsilon(1S)$ [2427], with $m_{J/\psi}c^2 = 3.0969\,\mathrm{GeV}$ and $m_\Upsilon c^2 = 9.4604\,\mathrm{GeV}$ — although the reduced masses differ by a factor of three. A pure Coulomb potential would give spacings scaling as the reduced mass; a pure linear potential would give them scaling as $m^{-1/3}$; the observed mass-independence is what an intermediate, roughly logarithmic, potential produces, and Equation (102.69) is exactly such a potential over the relevant range.

Phenomenon 102.64 (Almost none of the proton's mass comes from quark masses)**.** The proton mass is $938.272\,\mathrm{MeV}/c^2$, while the current masses of its three valence quarks sum to

less than $10\,\mathrm{MeV}/c^2$ [2427] — about one percent of the total. The remainder is the energy of the gluon field and of the confined motion of the quarks. That this is not an evasion is shown by computation: lattice evaluations of the light hadron spectrum, taking as input only the quark masses and the coupling, reproduce the measured masses at the percent level [2122].

Rests on Phenomenon 102.39 and Equations (102.29) *and* (106.81).

Derivation. *Derives Phenomenon 102.64.*

The arithmetic is immediate from Table 102.1: $2m_u + m_d = 9.02\,\mathrm{MeV}/c^2$ against $938.272\,\mathrm{MeV}/c^2$, so 99 % of the mass is not quark mass. What has to be explained is not the deficit but the *presence* of a scale at all, and that is Phenomenon 102.39: in the limit $m_u, m_d \to 0$ the Lagrangian Equation (102.29) contains no dimensionful parameter, yet the quantum theory generates Λ_{QCD} through Equation (102.45), and every mass in the spectrum is a pure number times $\Lambda_{\mathrm{QCD}}/c^2$. The proton mass is therefore predicted up to that one overall scale, which is fixed by measuring any single hadron mass; everything else is then a prediction with no freedom.

Turning that into numbers is a computation and not a derivation, and the distinction is worth keeping. The lattice programme [2122] proceeds by: choosing the bare quark masses and β; computing hadron correlators by Monte Carlo evaluation of Equation (106.81) with dynamical quarks; setting the scale by matching one measured quantity, so that a is known in metres; tuning the light quark masses by matching two more, typically m_π and m_K; extrapolating to zero spacing, to infinite volume and, where necessary, to the physical light quark masses; and then *predicting* the remaining spectrum. Three inputs, and the masses of the nucleon, the Δ, the Λ, the Σ, the Ξ and the Ω come out right to a few percent. That is the strongest single piece of evidence that Equation (102.29) is the correct theory of the strong interaction, and it is the reason this chapter can call confinement an observed regularity of a known Lagrangian rather than a mystery. □

102.7.3 No free quark has ever been observed

Phenomenon 102.65 (No free quark has ever been observed)**.** No particle carrying a fractional electric charge has been detected in any experiment. Millikan-type searches for residual fractional charge in bulk matter set limits below 10^{-20} fractionally charged particles per nucleon in the best samples [1992], and accelerator and cosmic-ray searches are likewise null [2427]. The positive counterpart of this null result is that the potential energy between a static colour source and its anticolour grows linearly with their separation instead of falling off, so that pulling them apart creates a new pair rather than liberating a colour charge.

Rests on Equations (102.12) *and* (102.69).

Derivation of the string-breaking mechanism. *Derives Phenomenon 102.65.*

Take Equation (102.69) at face value and try to separate a quark from an antiquark. At separation r the stored energy is σr, and it grows without bound. But the vacuum is not empty: as soon as σr exceeds the energy needed to create a light quark–antiquark pair, roughly twice the constituent mass Equation (102.12), it is energetically favourable for the flux tube to break, the new pair capping the two broken ends. The critical separation is

$$r_{\mathrm{break}} \approx \frac{2m_q c^2}{\sigma} = \frac{2 \times 0.336\,\mathrm{GeV}}{0.9\,\mathrm{GeV/fm}} \approx 0.7 \times 10^{-15}\,\mathrm{m}\,, \tag{102.72}$$

which is smaller than a hadron. So the attempt to isolate a quark produces, at a separation of less than one hadron radius, two hadrons instead: what would have been a free colour source is always screened by a newly created one before it can be separated. This is why the null results above are expected, and it is also why they cannot be strengthened into an observation of the mechanism: the mechanism predicts precisely that nothing anomalous is ever seen. □

Derivation pending #151

Confinement: no analytic derivation from the continuum Lagrangian exists, and none should be implied. What is available is Wilson's area law for the loop expectation value, exact in the strong-coupling lattice expansion, together with the numerical demonstration that the lattice theory joins continuously onto the asymptotically free continuum. Closing the gap between those two statements is an open problem, not an exercise.

Remark 102.66 (The honest status of confinement). Three statements must be kept apart, and conflating them is the commonest overclaim in expositions of QCD. *(i)* No free quark has been observed. This is an experimental fact, Phenomenon 102.65. *(ii)* The lattice theory confines at strong coupling. This is a theorem, Theorem 102.60, and by the compact-U(1) counterexample given in its proof it does not by itself say anything about the continuum theory. *(iii)* Continuum SU(3) Yang–Mills theory has a mass gap and confines. This is neither observed directly nor proved. It is supported by numerical evidence of high quality — the asymptotic scaling first demonstrated for SU(2) [1651] and the SU(3) spectrum of [2122] — and it is one of the recognized open problems of mathematical physics, recorded as such in Chapter 133. This chapter asserts (i) and (ii) and reports (iii) as well supported and unproved.

102.8 Chiral symmetry and its breaking

102.8.1 The chiral limit

Table 102.1 contains a striking accident. The masses of the two lightest quarks, $2.16\,\mathrm{MeV}/c^2$ and $4.70\,\mathrm{MeV}/c^2$, are of the order of one percent of Λ_{QCD}, which is the only other scale in the theory. To that accuracy they may be set to zero, and the resulting theory has a symmetry that the real one nearly has.

Proposition 102.67 (The chiral symmetry of massless quarks)**.** *Write $q = q_L + q_R$ with the projectors $P_{L,R} = \tfrac{1}{2}\left(\mathbb{1} \mp \gamma^5\right)$. In Equation (102.29) the gauge coupling is diagonal in chirality and the mass term is not:*

$$\bar{q}\mathrm{i}\hbar c\gamma^\mu D_\mu q = \bar{q}_L \mathrm{i}\hbar c\gamma^\mu D_\mu q_L + \bar{q}_R \mathrm{i}\hbar c\gamma^\mu D_\mu q_R\,, \qquad \bar{q}mc^2 q = c^2\left(\bar{q}_L m q_R + \bar{q}_R m q_L\right)\,. \tag{102.73}$$

Hence for n_f massless flavours the Lagrangian is invariant under independent unitary rotations of the left- and right-handed fields,

$$\mathrm{SU}(n_f)_L \times \mathrm{SU}(n_f)_R \times \mathrm{U}(1)_V \times \mathrm{U}(1)_A\,, \tag{102.74}$$

of which $\mathrm{U}(1)_V$ is baryon number and $\mathrm{U}(1)_A$ is destroyed by the axial anomaly of Theorem 106.87.

Rests on Equations (102.21) and (102.29) and Theorem 106.87.

Proof. *Derives Proposition 102.67.*

γ^5 anticommutes with every γ^μ, so $P_L\gamma^\mu = \gamma^\mu P_R$ and $\bar{q}_L\gamma^\mu q_R = q^\dagger P_L\gamma^0\gamma^\mu P_R q = q^\dagger\gamma^0 P_R\gamma^\mu P_R q = 0$ because $P_R\gamma^\mu P_R = \gamma^\mu P_L P_R = 0$; the cross terms in the kinetic piece therefore vanish, which is the first half of Equation (102.73). For the mass term the same algebra keeps only the cross terms, since $\bar{q}_L q_L = 0$. The covariant derivative Equation (102.21) contains no γ^5 and acts only on colour, so it commutes with the flavour rotations $q_L \to U_L q_L$, $q_R \to U_R q_R$; with the mass matrix set to zero nothing else connects the two chiralities, giving Equation (102.74). Splitting $\mathrm{U}(n_f) = \mathrm{SU}(n_f) \times \mathrm{U}(1)$ on each side and recombining the two U(1)s into vector ($U_L = U_R$) and axial ($U_L = U_R^\dagger$) combinations names the last

two factors. The vector one is the conservation of baryon number. The axial one is not a symmetry of the quantum theory at all: the measure of the functional integral is not invariant under it, which is Theorem 106.87, and the classical current has an anomalous divergence proportional to the gluon topological density. □

Phenomenon 102.68 (The spectrum has no parity doublets)**.** If $\mathrm{SU}(2)_L \times \mathrm{SU}(2)_R$ were realized in the ordinary way — as a symmetry of the states as well as of the Lagrangian — every hadron would have a partner of equal mass and opposite parity. It does not. The lightest vector meson is the $\rho(775)$ with $J^{PC} = 1^{--}$ and its chiral partner the $a_1(1260)$ with 1^{++}, some $455\,\mathrm{MeV}/c^2$ heavier; the nucleon at $939\,\mathrm{MeV}/c^2$ with $J^P = \frac{1}{2}^+$ has as its partner the $N(1535)$ with $\frac{1}{2}^-$, some $596\,\mathrm{MeV}/c^2$ heavier [2427]. Meanwhile the *vector* subgroup $\mathrm{SU}(2)_V$ — ordinary isospin — is realized in the usual way, the proton and neutron differing by $1.3\,\mathrm{MeV}/c^2$, one part in seven hundred [2427].

Rests on Proposition 102.67 and Equation (102.74)*.*

Derivation. Derives Phenomenon 102.68.

If a symmetry generator Q annihilates the vacuum then it commutes with the Hamiltonian and maps energy eigenstates to degenerate energy eigenstates. The axial generators Q_5^a change parity, because the axial current is a pseudovector; so $Q_5^a\,|h\rangle$ would be a state of the same mass and opposite parity to $|h\rangle$, for every hadron. The splittings quoted are not small corrections to such a degeneracy: they are of order Λ_{QCD}, which is the largest scale available, whereas the explicit breaking by the quark masses is of order 7 MeV. Symmetry breaking by the Lagrangian cannot produce an effect a hundred times its own size. Hence the axial generators do *not* annihilate the vacuum, while the vector ones evidently do — the isospin multiplets are degenerate to the accuracy at which electromagnetic and $m_d - m_u$ effects enter. The symmetry is spontaneously broken, in the pattern

$$\mathrm{SU}(2)_L \times \mathrm{SU}(2)_R \longrightarrow \mathrm{SU}(2)_V\,. \tag{102.75}$$

□

102.8.2 Dynamical symmetry breaking

Definition 102.69 (The quark condensate)**.** The order parameter of Equation (102.75) is the vacuum expectation value

$$\langle \bar{q} q \rangle = \langle \bar{q}_L q_R + \bar{q}_R q_L \rangle \neq 0\,, \tag{102.76}$$

of SI dimension $1/\mathrm{m}^3$ by Definition 100.2: it is a number density, and it is negative.

The mechanism was identified by Nambu and Jona-Lasinio [1215] by explicit analogy with superconductivity. In the BCS theory of superconductivity, treated in the condensed-matter part of this book, an arbitrarily weak attraction between electrons destabilizes the Fermi sea and produces a condensate of pairs, and with it a gap: an excitation that was massless acquires a mass generated by the interaction rather than written into the Hamiltonian. Replace electron pairs by quark–antiquark pairs and the gap by a quark mass, and Equation (102.76) is the same statement. A massless quark propagating through a vacuum full of $\bar{q}q$ pairs acquires an effective mass, the *constituent* mass Equation (102.12) of about $336\,\mathrm{MeV}/c^2$ — a hundred times its Lagrangian mass, and manufactured entirely by the interaction.

Theorem 102.70 (Goldstone)**.** *Let j_a^μ be a conserved current of a symmetry that does not annihilate the vacuum, in the sense that some local operator Φ has $\langle [Q_a, \Phi] \rangle \neq 0$ with $Q_a = c^{-1} \int \mathrm{d}^3x\, j_a^0$. Then the theory contains a massless one-particle state $|\pi_a\rangle$ with $\langle 0 | j_a^\mu | \pi_a \rangle \neq 0$ [1207].*

Rests on Phenomenon 102.68 and Equation (102.75)*.*

Proof. *Derives Theorem 102.70.*

Write $v = \langle [Q_a, \Phi(0)] \rangle \neq 0$ and insert a complete set of states between the operators:

$$v = \frac{1}{c} \int \mathrm{d}^3x \sum_n \left[\left\langle 0|j_a^0(x)|n \right\rangle \langle n|\Phi(0)|0\rangle - \langle 0|\Phi(0)|n\rangle \left\langle n|j_a^0(x)|0 \right\rangle \right] .$$

Translation invariance gives $\langle 0|j_a^0(x)|n\rangle = \langle 0|j_a^0(0)|n\rangle \, \mathrm{e}^{-\mathrm{i}p_n \cdot x/\hbar}$, so the integral over d^3x produces $(2\pi\hbar)^3\delta^3(\boldsymbol{p}_n)$: only states of zero three-momentum contribute. For those the remaining x dependence is $\mathrm{e}^{-\mathrm{i}E_n t/\hbar}$, and therefore

$$v = \frac{(2\pi\hbar)^3}{c} \sum_{n,\, \boldsymbol{p}_n = 0} \left[\left\langle 0|j_a^0|n \right\rangle \langle n|\Phi|0\rangle \, \mathrm{e}^{-\mathrm{i}E_n t/\hbar} - \text{c.c. term} \right] .$$

But Q_a is time independent, because $\partial_\mu j_a^\mu = 0$ and the spatial part integrates to a surface term; so v cannot depend on t. Since $v \neq 0$, at least one term must survive, and a surviving term must have $E_n = 0$ at $\boldsymbol{p}_n = 0$. A state with zero energy at zero momentum is a massless particle, and by construction it has a non-vanishing matrix element with the current. Its quantum numbers are those of the current: for the axial current A_a^μ, which is a pseudovector, the state must be a pseudoscalar, since $\langle 0|A_a^\mu|\pi_b\rangle$ can only be proportional to p^μ and p^μ is the only vector available for a spinless state. Applying this to Equation (102.75): three generators are broken — $\mathrm{SU}(2)_L \times \mathrm{SU}(2)_R$ has six and $\mathrm{SU}(2)_V$ has three — so there are three massless pseudoscalars, an isotriplet. □

Remark 102.71 (Where the mass of ordinary matter comes from). Chapter 104 is often described as explaining the origin of mass. It does not explain the origin of *our* mass. The Higgs mechanism supplies the Lagrangian quark masses of Table 102.1, which contribute $9\,\mathrm{MeV}/c^2$ of the proton's $938.272\,\mathrm{MeV}/c^2$. The remaining 99 % is supplied by Equation (102.76) and Phenomenon 102.39: a dimensionless coupling that runs, a scale generated by that running, and a vacuum that condenses. Two different mechanisms, in two different sectors, and the one that matters for the mass of a table is the strong one.

102.8.3 The pion as a pseudo-Goldstone boson

Phenomenon 102.72 (The pion is anomalously light)**.** The lightest hadron is far lighter than the rest of the spectrum: $m_{\pi^\pm}c^2 = 139.570\,\mathrm{MeV}$ and $m_{\pi^0}c^2 = 134.977\,\mathrm{MeV}$, against 775 MeV for the ρ and 938 MeV for the nucleon [2427]. The gap is not a small effect to be absorbed into a fit; it is a factor of five or six, and it is specific to the pseudoscalar triplet. Furthermore it is m_π^2, and not m_π, that is proportional to the light quark masses, so that the pion alone becomes massless when those masses are set to zero.

Rests on Theorem 102.70, Phenomenon 102.68, and Definition 102.69.

Derivation. *Derives Phenomenon 102.72.*

Theorem 102.70 supplies three massless pseudoscalars in the limit $m_u = m_d = 0$, and Phenomenon 102.68 identifies the broken symmetry as the one whose Goldstone bosons they are. The pions are those states, made massive only by the explicit breaking. It remains to find how their mass depends on the breaking, and the answer is the Gell-Mann–Oakes–Renner relation [1370]:

$$\left(m_\pi c^2\right)^2 F_\pi^2 = -\left(m_u + m_d\right) c^2 \, \langle \bar{q}q \rangle \, (\hbar c)^3 \, , \tag{102.77}$$

where F_π is the pion decay constant of Equation (106.103), an energy, and every factor is written so that both sides carry (energy)^4: $\langle \bar{q}q \rangle$ carries $1/\mathrm{m}^3$ by Definition 102.69 and $(\hbar c)^3$ converts it.

The derivation is a Ward identity. The divergence of the isotriplet axial current, exact in QCD, is $\partial_\mu A_a^\mu = (m_u + m_d)\, c\, P_a$ with $P_a = \mathrm{i}\bar{q}\gamma^5\tau_a q$ the pseudoscalar density; it vanishes in

the chiral limit, which is Proposition 102.67. Take its matrix element between the vacuum and a one-pion state. The left-hand side is fixed by the definition of the decay constant, $\langle 0|A^\mu_a|\pi_b(p)\rangle = \mathrm{i}F_\pi p^\mu \delta_{ab}/c$, so that $\langle 0|\partial_\mu A^\mu_a|\pi_b\rangle = F_\pi \left(m_\pi c\right)^2 \delta_{ab}/\hbar c$ on using $p^2 = m_\pi^2 c^2$. The right-hand side is $\left(m_u + m_d\right) c \langle 0|P_a|\pi_b\rangle$, and the remaining matrix element is related to the condensate by the same chiral rotation that produced the Goldstone boson: acting with the broken charge on the pseudoscalar density returns the scalar density, so $\langle 0|P_a|\pi_b\rangle$ is proportional to $\langle \bar{q}q\rangle / F_\pi$. Assembling the factors gives Equation (102.77). The structural point survives any bookkeeping: m_π^2, not m_π, is linear in the quark masses, because the quark mass enters the Lagrangian linearly and the pion mass enters the Lagrangian squared. □

Example 102.73 (Two numbers out of the relation)**.** *The condensate.* Insert $m_{\pi^0} c^2 = 134.977\,\mathrm{MeV}$, $F_\pi = 92.07\,\mathrm{MeV}$ from $f_\pi = (130.2 \pm 1.2)\,\mathrm{MeV}$, and $\left(m_u + m_d\right) c^2 = 6.86\,\mathrm{MeV}$ [2427] into Equation (102.77). Writing $-\langle \bar{q}q\rangle \left(\hbar c\right)^3 = \Sigma^3$ with Σ an energy,

$$\Sigma = \left[\frac{\left(m_\pi c^2\right)^2 F_\pi^2}{\left(m_u + m_d\right) c^2}\right]^{1/3} = 282\,\mathrm{MeV}\,, \qquad \langle \bar{q}q\rangle = -2.9 \times 10^{45}/\mathrm{m}^3\,, \tag{102.78}$$

about three quark–antiquark pairs per cubic femtometre, and of the order of Λ_{QCD} as Phenomenon 102.39 requires — there being nothing else for it to be made of. Lattice and sum-rule determinations of the condensate agree with Equation (102.78) at the ten-percent level [2427], which is the accuracy the leading order of the chiral expansion deserves.

A ratio with no condensate in it. Equation (102.77) extended to three flavours gives $\left(m_K c^2\right)^2 F_K^2 = -\left(m_d + m_s\right) c^2 \langle \bar{q}q\rangle \left(\hbar c\right)^3$ for the K^0, so that in the limit $F_K \to F_\pi$ the condensate cancels:

$$\frac{m_{K^0}^2}{m_{\pi^0}^2} = \frac{m_d + m_s}{m_u + m_d} = \frac{4.70 + 93.5}{2.16 + 4.70} = 14.3\,, \tag{102.79}$$

against the measured $(497.611/134.977)^2 = 13.6$ [2427]: agreement to 5 %. The neutral mesons are used because the charged ones carry an electromagnetic self-energy that Equation (102.77) does not describe. This is a genuine test, because the quark masses entering it are determined elsewhere — on the lattice and from sum rules — and not fitted to the meson masses.

Remark 102.74 (Chiral perturbation theory, and the ninth meson)*.* Equation (102.77) is the first term of a systematic expansion. Because the pions are the light degrees of freedom at low energy, and because their interactions are constrained by the broken symmetry to vanish as the momenta go to zero, one may write an effective Lagrangian for the pion field alone, organized in powers of momenta and quark masses: chiral perturbation theory [1647, 1737]. It is an effective field theory in the sense developed in the renormalization-group chapter, with a finite number of measurable constants at each order, and it predicts pion–pion scattering lengths, form factors and mass ratios to a few percent.

The pseudoscalar octet has eight members and the Goldstone count for $\mathrm{SU}(3)_L \times \mathrm{SU}(3)_R \to \mathrm{SU}(3)_V$ is eight, so the η' is one too many. It is not a Goldstone boson of $\mathrm{U}(1)_A$, because $\mathrm{U}(1)_A$ is not a symmetry: the anomaly of Theorem 106.87 destroys it, with the gluon topological density of Section 106.6.2 on the right-hand side of the divergence equation. Weinberg's observation [1565] is that if $\mathrm{U}(1)_A$ *had* been a spontaneously broken symmetry the ninth pseudoscalar would have had to weigh at most $\sqrt{3} m_\pi$, about $234\,\mathrm{MeV}/c^2$; the η' weighs $957.78\,\mathrm{MeV}/c^2$ [2427]. Its heaviness is therefore evidence for the anomaly, and through it for the topological structure that Section 102.10 is about.

102.9 QCD under extreme conditions

Asymptotic freedom implies that matter compressed or heated far enough must eventually become a gas of quarks and gluons: at short distances the coupling is weak, and raising the

temperature or the density reduces the mean separation. The transition is not a prediction of perturbation theory — it happens at $\alpha_s \sim 1$ — but of the lattice, and it has been looked for in heavy-ion collisions.

Phenomenon 102.75 (Hadronic matter deconfines when heated)**.** Lattice computation of QCD at finite temperature shows a rapid but smooth rise of the energy density, by a factor of order ten over a narrow range of temperature centred on $k_B T_c \approx 155\,\mathrm{MeV}$, that is

$$T_c \approx \frac{155\,\mathrm{MeV}}{k_B} = 1.8 \times 10^{12}\,\mathrm{K}\,, \tag{102.80}$$

some 10^5 times the temperature at the centre of the Sun. It is a crossover, not a phase transition with a discontinuity. Above it the degrees of freedom are those of quarks and gluons rather than of hadrons.

Rests on Theorem 102.35 and Definition 102.10.

Derivation of the size of the jump. *Derives Phenomenon 102.75.*

The energy density of an ideal ultrarelativistic gas is the standard blackbody result, generalized to g internal states and to fermions by Fermi–Dirac counting:

$$\varepsilon = g_{\mathrm{eff}}\,\frac{\pi^2\,(k_B T)^4}{30\,(\hbar c)^3}\,, \qquad g_{\mathrm{eff}} = g_B + \frac{7}{8}\,g_F\,, \tag{102.81}$$

the factor $7/8$ being the ratio of the Fermi to the Bose integral $\int_0^\infty x^3 \mathrm{d}x/(\mathrm{e}^x \pm 1)$. Counting: a gluon has 8 colours and 2 polarizations, so $g_B = 16$; a quark flavour has 3 colours, 2 spins, and particle and antiparticle, so $g_F = 12$ per flavour and $g_F = 36$ for the three light flavours. Hence $g_{\mathrm{eff}} = 16 + \frac{7}{8}\cdot 36 = 47.5$. Below the transition the degrees of freedom are three pions, $g_{\mathrm{eff}} = 3$. The ratio is nearly sixteen, and that factor — not the precise value of T_c — is what the lattice sees as a steep rise in ε/T^4.

Numerically at $k_B T = 155\,\mathrm{MeV}$, Equation (102.81) gives $\varepsilon \approx 1.2\,\mathrm{GeV}$ per cubic femtometre, that is

$$\varepsilon \approx 1.9 \times 10^{35}\,\mathrm{J/m^3}\,, \tag{102.82}$$

against about $0.4\,\mathrm{GeV}$ per cubic femtometre inside a nucleon at rest. The lattice value near T_c is some tens of percent below the ideal-gas limit Equation (102.81), and the deficit does not go away quickly as the temperature rises: the medium is not a weakly coupled gas, which is exactly what the heavy-ion data below also say. □

The experimental programme is the collision of heavy nuclei at relativistic energies, and its conclusion after the first five years at RHIC [2068] rests on two observations that point the same way.

Jet quenching. A parton produced in a hard scattering inside the collision volume must traverse the medium before fragmenting into the jet of Section 102.6. If the medium is coloured it radiates gluons and loses energy. The observable is the suppression of high transverse-momentum hadrons relative to the same quantity in proton–proton collisions scaled by the number of binary collisions, and it is large — a factor of about five in central gold–gold collisions — while being absent in deuteron–gold collisions, which excludes an initial-state explanation.

Elliptic flow. A non-central collision has an almond-shaped overlap region. If the matter produced behaves as a fluid, the pressure gradients are larger along the short axis and the final momentum distribution is anisotropic in a way that remembers the initial geometry. The measured anisotropy is large and is described by relativistic hydrodynamics with a shear viscosity to entropy density ratio close to the smallest values known for any fluid. The conclusion of [2068] was accordingly not “a weakly coupled quark–gluon gas has been

made" but "a strongly coupled, nearly ideal liquid has been made", which was not what the perturbative picture had led anyone to expect and is the reason the result mattered.

Two further regimes are treated elsewhere. Cold dense quark matter is the possible interior of a neutron star and belongs to Chapter 49. And the whole universe passed through Equation (102.80) downwards at an age of a few microseconds, converting a quark–gluon plasma into hadrons; that is the cosmological quark–hadron transition, and it is the last epoch before the nucleosynthesis of Chapter 52.

102.10 The strong CP problem

Definition 102.27 deferred one term, and it is the only known defect of the theory.

Proposition 102.76 (QCD permits one more term)**.** *Gauge invariance, Lorentz invariance, locality and renormalizability permit, besides Equation* (102.29)*, the term*

$$\mathcal{L}_\theta = \theta\hbar\, q(x) \propto \theta\, \frac{g_s^2}{\hbar} \sum_a \boldsymbol{E}^a \cdot \boldsymbol{B}^a \,, \tag{102.83}$$

with $q(x)$ the gluon topological charge density of Definition 105.43 and θ a dimensionless number, and nothing else. The term violates P and CP (Proposition 105.44), and only the combination $\bar\theta = \theta + \arg\det M_q$ of Equation (105.111) *is physical, because a chiral rotation of the quark fields shifts θ through the anomaly of Theorem 106.87.*

Rests on Equation (102.29)*, Definition 105.43, and Theorem 100.17.*

Proof. *Derives Proposition 102.76.*

Power counting (Theorem 100.17) allows operators of mass dimension at most four. The gauge-invariant, Lorentz-invariant operators of dimension four that can be built from $F^a_{\mu\nu}$ are $F^a_{\mu\nu}F^{a\,\mu\nu}$, already in Equation (102.29), and $\epsilon^{\mu\nu\rho\sigma}F^a_{\mu\nu}F^a_{\rho\sigma}$, which is Equation (102.83); from the quark fields they are the kinetic and mass terms. There is no other. The identification of Equation (102.83) with the topological density, its periodicity in θ, and its origin in the vacuum structure are Corollary 106.84 and Section 106.6.2. The statement that only $\bar\theta$ is observable is Equation (105.111): it ties the gluon sector to the phases of the quark mass matrix, which come from a different sector of the Standard Model altogether. □

Phenomenon 102.77 (The neutron has no measurable electric dipole moment)**.** Every symmetry of QCD permits a term in the Lagrangian proportional to $F\tilde F$; it violates P and CP, and it would give the neutron a permanent electric dipole moment along its spin. None is seen. The best measurement is consistent with zero and bounds $|d_n| < 1.8\times10^{-28}\,e\,\mathrm{m}$ at 90 percent confidence [2346], which forces the coefficient of the permitted term below about 10^{-10}. Nothing in the theory explains why a term it allows should be absent to that accuracy.

Rests on Propositions 102.76 and 105.45 and Equation (105.112)*.*

Derivation. *Derives Phenomenon 102.77.*

The step that converts a measured bound on a dipole moment into a bound on a Lagrangian parameter is the chiral estimate Equation (105.112), derived in Proposition 105.45: $d_n \approx \bar\theta \times 1.6\times10^{-37} - -6.4\times10^{-37}\,\mathrm{C\,m}$, the parametric form following from the requirement that the answer vanish with the quark masses — because a massless quark makes θ unobservable by the chiral rotation of Proposition 102.76 — and the coefficient from a chiral loop. Combining with the ultracold-neutron measurement [2346] gives $\left|\bar\theta\right| \lesssim 10^{-10}$, which is Equation (105.113). The uncertainty of the estimate is a factor of a few, which is why the bound is quoted as an order of magnitude. □

Remark 102.78 (A fine-tuning puzzle, honestly labelled). Nothing is inconsistent. $\bar{\theta} = 0$ is a perfectly good value of a parameter, and QCD with $\bar{\theta} = 0$ describes every measurement in this chapter. What is peculiar is that $\bar{\theta}$ is an angle, periodic with period 2π, assembled from the gluon sector and the phases of the Yukawa couplings of Chapter 104, which are otherwise of order one and which supply the large CP violation of Chapter 105; and the two contributions cancel to ten decimal places. That is a fine-tuning puzzle, not a contradiction, and this book records it as such.

The best-known proposal promotes $\bar{\theta}$ to a dynamical field whose potential is minimized at zero [1609], which predicts a light pseudoscalar particle [1630]. No such particle has been detected in any search to date. In the terms of this book that makes it a proposal and not physics, and the strong CP problem is carried forward unresolved to Chapter 133. It is worth noting what would settle it: detecting the particle, or measuring a nonzero neutron electric dipole moment, which would show $\bar{\theta} \neq 0$ and remove the puzzle by removing its premise.

Chapter 103

Flavour Physics and Neutrinos

Flavour is the part of the Standard Model that nothing explains. The gauge structures of Chapter 100, Chapter 101 and Chapter 102 follow from one principle applied to three groups; the existence of three copies of every fermion, their wildly disparate masses, and the two mixing matrices that relate the states of definite mass to the states that couple to the W are pure measurement. This chapter assembles that measurement. It is also where the treatise's editorial rule bites hardest: flavour is the richest source of proposals without evidence, and what is recorded below is what has been seen — quark mixing, CP violation in three meson systems, and neutrino oscillation — together with an explicit statement of what has not.

The chapter follows Chapter 102 because the hadronic states whose decays carry the information are QCD bound states, and it feeds Chapter 111 and Chapter 114, the two chapters that carry the decisive experimental programmes; the first is still a skeleton of reserved headings and the second is written in part, so where a measurement is owed to one of them this chapter says so. Neutrino oscillation is the only laboratory result that the Standard Model as originally written does not accommodate, since it requires neutrino mass; that makes this chapter the natural bridge to Chapter 133, and its cosmological consequences — the matter–antimatter asymmetry — connect to Chapter 48. Standard treatments are [1739, 1933, 2106]; all quoted values are from [2427] unless another source is named, and every constant is the CODATA 2022 recommendation [2431].

Two pieces of scaffolding are inherited rather than rebuilt. The kinematic and dimensional conventions — metric signature, the mass-shell relation $p^2 = m^2c^2$, the SI dimension carried by a field and by a coupling — are those fixed in Section 100.1.1, and are used here without restatement. The *origin* of the quark mixing matrix, as the residual misalignment between the basis in which the Yukawa couplings are diagonal and the basis in which the charged current is diagonal, is derived in Theorem 104.38 from the Higgs couplings of Chapter 104; Equation (104.48) is the biunitary diagonalization that produces it. This chapter starts from that result and asks what the matrix actually is.

Remark 103.1 (Reading this chapter against the natural-units literature). Every reference cited below writes its formulae with $\hbar = c = 1$, and this book does not: editorial rule 3 requires SI throughout, so $\hbar$ and c appear in every equation and in every intermediate step here. The dictionary is mechanical, and is given once so that a reader can move between the two. A quantity quoted in the literature as an energy to the power n becomes, in SI, that energy divided by the appropriate power of $\hbar c$: a length is $\hbar c/E$, a time is $\hbar/E$, a squared mass is E^2/c^4. The numerical bridge is

$$\hbar c = 1.973\,269\,8 \times 10^{-7}\,\mathrm{eV\,m} = 3.161\,526\,8 \times 10^{-26}\,\mathrm{J\,m}\,, \qquad \hbar = 6.582\,119\,6 \times 10^{-16}\,\mathrm{eV\,s}\,. \tag{103.1}$$

Two conventions deserve naming because they recur. The Fermi constant is universally quoted as $G_F/(\hbar c)^3 = 1.166\,378\,7 \times 10^{-5}/\mathrm{GeV}^2$, which is not G_F; the constant itself is

$G_F = 1.435\,85 \times 10^{-62}\,\mathrm{J\,m^3}$, as Definition 101.8 sets out. And a neutrino squared-mass splitting is quoted as Δm^2 in eV^2 when the quantity meant is $\Delta m^2 c^4$, an energy squared. Both usages are retained below *with the powers of $\hbar$ and c written out*, because the derivations are carried out in SI and not translated into it afterwards. No derivation in this chapter takes place inside this remark.

103.1 The discovery of new quantum numbers

103.1.1 V particles

In October 1946 Rochester and Butler exposed a cloud chamber in a magnetic field to the cosmic radiation at sea level and, in some five thousand photographs, found two in which a track forked without any incoming charged particle at the vertex [903]. The first showed a neutral object decaying into two oppositely charged particles — the forked "V" that named the class — and the second a charged object changing direction with the emission of something neutral. Neither could be a scattering: momentum conservation at the vertex failed unless an unseen neutral particle was created, and the reconstructed mass of the decaying object, about $500\,\mathrm{MeV}/c^2$ in the neutral case, matched nothing then known. The cosmic-ray context, and the instrumental history that made such an exposure possible, belong to Chapter 115.

What made the new particles a problem rather than merely a discovery was a mismatch of time scales, and it is worth putting numbers on it because the whole of flavour physics grows out of that mismatch.

Phenomenon 103.2 (Copious production, slow decay)**.** The new particles are produced in nucleon–nucleon and pion–nucleon collisions with cross sections of hadronic size, of order $1 \times 10^{-30}\,\mathrm{m^2}$, which is the strength of the strong interaction; yet they decay with lifetimes of order $1 \times 10^{-10}\,\mathrm{s}$, which is thirteen to fourteen orders of magnitude longer than a strong decay of the same energy release. The Λ^0 is the cleanest case: its measured width is $2.515 \times 10^{-15}\,\mathrm{GeV}$, corresponding to

$$\tau_\Lambda = \frac{\hbar}{\Gamma_\Lambda} = \frac{6.582\,119\,6 \times 10^{-16}\,\mathrm{eV\,s}}{2.515 \times 10^{-6}\,\mathrm{eV}} = 2.62 \times 10^{-10}\,\mathrm{s}\,, \tag{103.2}$$

against $\tau_\Delta = \hbar/\Gamma_\Delta = 5.6 \times 10^{-24}\,\mathrm{s}$ for the $\Delta(1232)$, a genuine strong decay of comparable energy release [2427].

Rests on Equation (101.53).

Derivation. *Derives Phenomenon 103.2.*

The natural time scale for a decay driven by the interaction that also binds the hadron is the time light takes to cross it, $R/c = 1 \times 10^{-15}\,\mathrm{m}/c = 3.3 \times 10^{-24}\,\mathrm{s}$, and τ_Δ is within a factor of two of it: the Δ decays as fast as it can. The ratio $\tau_\Lambda/\tau_\Delta = 4.7 \times 10^{13}$ therefore cannot be absorbed into phase space, which varies as a power of the energy release and can account for two or three orders of magnitude at most, nor into a centrifugal barrier, which for the observed spins is absent. The only available conclusion is that the interaction responsible for the decay is not the interaction responsible for the production. Comparing τ_Λ with the muon lifetime $2.197 \times 10^{-6}\,\mathrm{s}$ and allowing for the fifth-power dependence of Equation (101.53) on the energy release identifies the decay interaction as the weak one.

The inference is then sharp. *Something* is conserved by the interaction that makes these particles and violated by the interaction that destroys them. That is the whole content of the next subsection, and it is the template for every conserved flavour quantum number in this chapter: exact for two of the three forces, broken by the third. □

A second regularity sharpened the puzzle: the new particles are never produced singly. A π^- striking a proton yields $\Lambda^0 + K^0$ but never $\Lambda^0 + \pi^0$, although the latter is energetically far easier. Pais named this *associated production*, and it is exactly what a conserved additive quantum number, carried with opposite sign by the two partners, would produce.

103.1.2 Strangeness

Definition 103.3 (Strangeness)**.** Strangeness S is an additive quantum number, integer valued, assigned to each hadron so that the electric charge satisfies

$$Q = e\left(I_3 + \frac{B+S}{2}\right), \tag{103.3}$$

with I_3 the third component of isospin and B the baryon number. It is conserved by the strong and electromagnetic interactions and violated by the weak interaction, which changes it by at most one unit in a single transition.

The assignment was made independently by Gell-Mann [1017] and by Nakano and Nishijima [1024], from the observed production and decay patterns alone, several years before any dynamical picture existed. Equation (103.3) is the same relation that reappears in Equation (102.1) as a statement about quark content and, in a wider form, in Equation (104.13) as a statement about the hypercharge of an electroweak multiplet. Its three appearances are one relation seen at three depths.

Proposition 103.4 (Associated production follows from conservation)**.** *If S is conserved by the strong interaction and the initial state of a collision of ordinary matter has $S = 0$, then strange particles can only be produced in pairs of opposite strangeness, and the lightest strange hadron of each sign is stable against strong decay.*

Rests on Definition 103.3 and Phenomenon 103.2.

Proof. *Derives Proposition 103.4.*

Additivity and conservation give $\sum_{\text{final}} S = 0$, so a final state containing one particle with $S = -1$ must contain another with $S = +1$; a final state $\Lambda^0\pi^0$ with $S = -1$ is forbidden however much phase space favours it, which is what is seen. For the second statement, let h be the lightest hadron with $S = -1$. A strong decay must conserve S, so its products together carry $S = -1$; but every hadron with $S = -1$ is at least as heavy as h, and any accompanying non-strange hadron has positive mass, so no such final state is open. The decay must therefore proceed through an interaction that violates S, and by Phenomenon 103.2 that interaction is weak. Applying the argument to Λ^0 ($S = -1$) and to K^+ ($S = +1$) accounts for the long lifetimes of both. □

Example 103.5 (Reading the charge formula)**.** The K^+ is an isodoublet member with $I_3 = +\frac{1}{2}$, $B = 0$; Equation (103.3) then requires $S = +1$. Its antiparticle K^- has $S = -1$, so K^+ and K^- are *not* members of one isodoublet, and the neutral kaons K^0 ($S = +1$) and $\bar{K}^0$ ($S = -1$) are distinct particles rather than one self-conjugate state — the fact that makes the whole of Section 103.5 possible. The Λ^0 is an isosinglet, $I_3 = 0$, $B = 1$, $Q = 0$, hence $S = -1$. The Ω^-, with $I_3 = 0$, $B = 1$, $Q = -e$, requires $S = -3$: three units of a quantum number that the weak interaction can shed only one at a time, which is why it decays in a cascade and why its lifetime is long even by strange standards. That prediction, and its confirmation, are Phenomenon 102.3.

In quark language, settled two decades later, S counts the strange quarks with a sign: $S = -(n_s - n_{\bar{s}})$, the minus sign an accident of the convention fixed before quarks were

known. The same construction repeats for every heavy flavour: charm C, bottom $\tilde{B}$ and top T are defined so that Equation (103.3) generalizes to

$$Q = e\left(I_3 + \frac{B + S + C + \tilde{B} + T}{2}\right), \tag{103.4}$$

and each is conserved by the strong and electromagnetic interactions and violated by the weak. Strangeness is not a special quantum number; it is the first member of a family, and the family is the subject of this chapter.

Remark 103.6 (What was and was not explained). Definition 103.3 is a classification, not a mechanism. It says that some quantity is conserved by two forces and violated by the third; it does not say why, nor how much the violation is, nor why the violating amplitude should be the same size as the one that governs nuclear beta decay. Those questions were answered — the first two by Section 103.2.1, the third only partly, and the deepest of them not at all — and the honest summary is that a selection rule was found in 1953 and the number controlling it was measured in 1963.

103.2 Quark mixing

103.2.1 The Cabibbo angle

By 1963 the weak interaction had a universal strength: the coupling extracted from muon decay and from nuclear beta decay agreed to within a few per cent, which is Phenomenon 101.43 — stated there over muon decay, superallowed nuclear beta decay and τ decay, the third of which came much later. Strangeness-changing decays did not fit. The semileptonic decay $\Lambda^0 \to p\, e^- \bar{\nu}_e$ proceeds at some five per cent of the rate that the same universal coupling predicts, and $K^+ \to \pi^0 e^+ \nu_e$ likewise. There were two ways to read this: either universality is approximate and the weak interaction has a separate, weaker, strangeness-changing piece, or universality is exact and the strength is *shared* between the two channels.

Cabibbo took the second reading [1248], and it is the conceptual step on which everything below rests: *the state that couples to the W is not a state of definite mass.*

Phenomenon 103.7 (Strangeness-changing decays are suppressed, by one angle)**.** Semileptonic decays that change strangeness proceed at roughly one twentieth of the rate of their strangeness-conserving counterparts, and a single rotation angle accounts for the whole of the deficit. The measured magnitudes of the first-row mixing elements are

$$|V_{ud}| = 0.973\,67 \pm 0.000\,32\,, \qquad |V_{us}| = 0.2243 \pm 0.0008\,, \qquad |V_{ub}| = 0.003\,82 \pm 0.000\,20\,, \tag{103.5}$$

obtained from superallowed nuclear beta decay, from kaon decays and from semileptonic B decays respectively [2427]. They satisfy $|V_{ud}|^2 + |V_{us}|^2 + |V_{ub}|^2 = 1$ to about a part in 10^3, with a residual deficit of a few standard deviations that is at present the sharpest unexplained departure from the Standard Model anywhere in flavour physics.

Rests on Phenomenon 101.43 and Definition 103.3.

Derivation. *Derives Phenomenon 103.7.*

Cabibbo's hypothesis is that the down-type state entering the charged current is a rotation of the mass eigenstates, $d' = d\cos\theta_C + s\sin\theta_C$, carrying the same overall coupling G_F as the leptonic current [1248]. A strangeness-conserving amplitude then acquires a factor $\cos\theta_C$ and a strangeness-changing one a factor $\sin\theta_C$, so their rates stand in the ratio $\tan^2\theta_C$ and the two effects are not independent: the suppression of the second is precisely the deficit of the first against the leptonic strength, which is what restores universality. Reading $\sin\theta_C = |V_{us}|$ from Equation (103.5) gives $\theta_C = 12.96°$ and hence $\cos\theta_C = 0.9745$,

which is $|V_{ud}|$ to within one part in a thousand — one angle, measured twice over in unrelated processes. The global fit quoted later as $\theta_{12} = 13.00^\circ$ in Equation (103.22) is the same angle: the 0.04° between the two is the difference between the direct kaon determination of $|V_{us}|$ used here and the fitted $\lambda = 0.225\,01$ used there, and nothing else. The rate ratio it predicts, $\tan^2\theta_C = 0.0530$, is the observed factor of about twenty. That the squares sum to unity is the two-generation case of the unitarity discussed in Section 103.4.3; that they sum to unity only to a part in 10^3 is a measurement, not an approximation, and it is recorded here rather than smoothed away. □

The arithmetic behind the last sentence is worth doing explicitly, because the number it produces is the one genuinely open experimental anomaly in quark flavour physics today.

Proposition 103.8 (First-row unitarity deficit)**.** *With the values of Equation* (103.5)*,*

$$\Delta_{\mathrm{CKM}} := |V_{ud}|^2 + |V_{us}|^2 + |V_{ub}|^2 = 0.998\,36 \pm 0.000\,72\,, \tag{103.6}$$

which is below unity by 2.3 *standard deviations.*
Rests on Equation (103.5)*.*

Proof. *Derives Proposition 103.8.*
$|V_{ud}|^2 = 0.948\,033$, $|V_{us}|^2 = 0.050\,310$ and $|V_{ub}|^2 = 1.46 \times 10^{-5}$, summing to $0.998\,358$. The uncertainty propagates as $\delta\Delta = 2\left[(|V_{ud}|\,\delta|V_{ud}|)^2 + (|V_{us}|\,\delta|V_{us}|)^2\right]^{1/2}$, the V_{ub} term being negligible, which gives $2\left[(3.12\times10^{-4})^2 + (1.79\times10^{-4})^2\right]^{1/2} = 7.2\times10^{-4}$. Hence $1 - \Delta_{\mathrm{CKM}} = 1.64\times10^{-3} \pm 7.2\times10^{-4}$, a deficit of 2.3σ. □

Remark 103.9 (How to read a two-sigma deficit honestly)*.* The significance of Equation (103.6) depends on which determination of $|V_{us}|$ is used — the value from $K \to \pi e\nu$ and the value from the ratio of $K \to \mu\nu$ to $\pi \to \mu\nu$ differ by about 2σ between themselves — and on the radiative corrections applied to superallowed nuclear beta decay, which were revised in the late 2010s and are what moved this quantity from agreement to tension. Quoted significances between 2σ and 3σ therefore all appear in the literature and all are defensible. What is not defensible is either extreme: calling it a discovery, or suppressing it. Unitarity of the first row is a prediction with no free parameters, the measurement is at the level of one part in 10^3, and it does not currently come out right. This treatise records that, and records equally that the most likely resolution is a systematic effect in one of the two inputs.

103.2.2 GIM and the prediction of charm

Cabibbo's rotation solves the charged-current problem and creates a neutral-current one, which Proposition 101.58 states: the same rotated field d' appears bilinearly in any neutral current, and its cross terms change strangeness with a coefficient $\sin\theta_C\cos\theta_C = 0.219$, of ordinary size. Nature says otherwise.

Phenomenon 103.10 (There are no flavour-changing neutral currents)**.** The neutral weak current does not change quark flavour. A flavour-changing neutral coupling of ordinary strength would make $K_L \to \mu^+\mu^-$ comparable in rate with $K^+ \to \mu^+\nu_\mu$; the measured branching fractions are 6.84×10^{-9} and 0.636 respectively [2427], a suppression by eight orders of magnitude, and the mass differences of the neutral kaon and B systems are correspondingly tiny. The suppression is not a numerical accident of one channel: no tree-level neutral flavour transition has ever been observed in any system, in any generation.
Rests on Theorems 101.59 and 104.38.

Derivation. *Derives Phenomenon 103.10.*
The tree-level statement is Theorem 101.59, proved there for two generations and extended to three in Theorem 104.38(ii): the neutral-current coefficient depends only on

T^3 and Q, hence is the same for every generation, hence is proportional to the identity in generation space, and $U^\dagger \mathbb{1} U = \mathbb{1}$ for any unitary U. Nothing survives the rotation to the mass basis. This is the Glashow–Iliopoulos–Maiani mechanism [1423], and it is a consequence of unitarity together with the requirement that every fermion of a given charge and chirality sit in a complete weak multiplet. The two hypotheses are what a search for flavour-changing neutral currents actually tests.

At tree level the cancellation is exact and the observed rate would be zero, which is also wrong. The measured 6.84×10^{-9} is small but not absent, and it is produced at one loop, where the cancellation is only partial. That calculation is Theorem 103.11 below, and it is where the mechanism stops being an algebraic identity and starts predicting a number. □

The one-loop statement is the important one historically, because it converts a null result into a mass.

Theorem 103.11 (GIM suppression at one loop)**.** *Let $A(m_i)$ be the amplitude for a flavour-changing neutral process, computed with a single up-type quark of mass m_i circulating in the loop, and let $\lambda_i = V_{id}V_{is}^*$ be the associated product of mixing factors. Then the physical amplitude is*

$$A = \sum_{i=u,c,t} \lambda_i A(m_i)\,, \qquad \sum_i \lambda_i = 0\,, \tag{103.7}$$

so any part of $A(m_i)$ that is independent of m_i cancels identically. Expanding A in the small parameter $x_i = m_i^2 c^4 / \left(M_W c^2\right)^2$ leaves

$$A \simeq A'(0) \sum_i \lambda_i \frac{m_i^2 c^4}{\left(M_W c^2\right)^2}\,, \tag{103.8}$$

so the surviving amplitude is suppressed by the ratio of a squared mass difference *of the internal quarks to the squared W mass. A flavour-changing neutral amplitude therefore measures the mass of the heaviest quark circulating in the loop whose mixing factor λ_i is not itself negligible.*

Rests on Theorem 101.62.

Proof. *Derives Theorem 103.11.*

The sum rule $\sum_i \lambda_i = \sum_i V_{id}V_{is}^* = \delta_{ds} = 0$ is the (d,s) column relation of Theorem 101.62. Write $A(m_i) = A(0) + A'(0)x_i + O(x_i^2)$, where the expansion is legitimate for $m_i c^2 \ll M_W c^2$ and where $A(0)$ is by construction independent of i. Then

$$A = A(0)\sum_i \lambda_i + A'(0)\sum_i \lambda_i x_i + \ldots = A'(0)\sum_i \lambda_i x_i + \ldots\,,$$

which is Equation (103.8). Two remarks fix its meaning. First, the cancellation is not a fine tuning: it is the statement that the loop, evaluated with degenerate internal masses, is a flavour-diagonal operator, and the mixing matrix cannot rotate a multiple of the identity into anything else. Second, the surviving term is a *difference*, since $\sum_i \lambda_i = 0$ allows m_i^2 to be shifted by any constant; writing the sum over the first two generations alone gives $\lambda_c \left(m_c^2 - m_u^2\right) c^4$, and the mass of the light quark drops out of everything. □

Example 103.12 (The charm mass from the kaon mass difference)**.** The K_L–K_S mass difference is generated by the box diagram in which two W bosons and two internal up-type quarks convert K^0 into $\bar{K}^0$. Keeping the charm term of Equation (103.8), which dominates for a system built from d and s quarks, the short-distance contribution is

$$\Delta m_K c^2 = \frac{G_F^2}{6\pi^2 \left(\hbar c\right)^6} f_K^2 c^4 B_K \left(m_K c^2\right) \left(m_c c^2\right)^2 \left|V_{cd}V_{cs}^*\right|^2\,, \tag{103.9}$$

where $f_K = 155.7\,\mathrm{MeV}/c^2$ is the kaon decay constant and $B_K = 0.717$ the lattice matrix element of the four-quark operator [2427], computed by the methods of Section 102.7.1. The dimensions balance as follows: $G_F/(\hbar c)^3$ carries $1/\mathrm{J}^2$, so its square carries $1/\mathrm{J}^4$, and the five remaining factors — $f_K^2 c^4$, $m_K c^2$ and $(m_c c^2)^2$ — supply J^5, leaving an energy, which is what $\Delta m_K c^2$ is. Inserting $G_F/(\hbar c)^3 = 1.166\,378\,7 \times 10^{-5}/\mathrm{GeV}^2$, $m_K c^2 = 497.611\,\mathrm{MeV}$, $|V_{cd}V_{cs}| = \sin\theta_C \cos\theta_C = 0.2186$ and $m_c c^2 = 1.273\,\mathrm{GeV}$ gives

$$\Delta m_K c^2\big|_{\text{charm box}} = 1.54 \times 10^{-6}\,\mathrm{eV}\,, \tag{103.10}$$

against the measured $\Delta m_K c^2 = 3.484 \times 10^{-6}\,\mathrm{eV}$, a fractional mass splitting of 7.0×10^{-15} and one of the smallest quantities ever measured [2427]. The box accounts for 44% of it; the short-distance QCD correction raises this to about 60% and the remainder is long-distance, that is, two-pion and other hadronic intermediate states that no short-distance expansion reaches.

Run backwards, before the charm quark was known, the same formula bounds its mass: saturating the observed Δm_K with the box requires $m_c c^2 \approx 1.9\,\mathrm{GeV}$, and requiring merely that the box not *exceed* the measurement gives an upper bound of the same order. Gaillard and Lee obtained a few GeV/c^2 by this argument in 1974 [1527]. The particle was found that November at $3.097\,\mathrm{GeV}/c^2$, which for a bound state of two charm quarks is exactly where a $1.5\,\mathrm{GeV}/c^2$ quark puts it.

Derivation pending #152

The short-distance two-W box amplitude for the kaon mass difference. What is carried out inline is the GIM structure of the internal sum, the dimensional check of every factor, and the numerical evaluation; the loop integral itself — the box function in its limit of internal mass small against the W mass, together with the reduction of the four-quark operator to the decay constant and the bag parameter — is not derived here and belongs in Appendix A.

Remark 103.13 (A particle predicted by a null result, quantified). Remark 101.60 records the structure of the inference: the input is the absence of a process and the output is the existence of a particle. Example 103.12 is what makes it a *measurement* rather than a hunch. The absence of $K_L \to \mu^+\mu^-$ at ordinary strength requires the mechanism; the smallness of Δm_K, which is not zero, then fixes the scale of the particle the mechanism needs. Glashow, Iliopoulos and Maiani proposed the fourth quark in 1970 on the first fact [1423]; the second gave its mass four years later, months before it was found. The same logic, applied to B^0–$\bar{B}^0$ mixing in 1987, gave the top mass a decade before the top was observed: see Section 103.5.3.

103.2.3 The November revolution

On 11 November 1974 two groups announced the same particle. At Brookhaven, Ting's group measured e^+e^- pairs from p + Be collisions and found a peak at $3.1\,\mathrm{GeV}/c^2$ whose width was consistent with the apparatus resolution and nothing more [1515]. At SPEAR, Richter's group scanned the e^+e^- annihilation cross section and found it rise by two orders of magnitude over a few MeV of beam energy [1516]. The particle is now written J/ψ, with

$$m_{J/\psi}c^2 = (3.096\,900 \pm 0.000\,006)\,\mathrm{GeV}\,, \qquad \Gamma_{J/\psi} = (92.6 \pm 1.7)\,\mathrm{keV} \tag{103.11}$$

[2427]. It is the width that mattered.

Phenomenon 103.14 (A hadronic resonance three orders too narrow)**.** The J/ψ lies far above the threshold for decay into ordinary hadrons and carries the quantum numbers of

the photon, so nothing forbids it from decaying strongly. Yet its width is 92.6 keV, against 149 MeV for the $\rho(770)$ and 4.25 MeV for the $\phi(1020)$ [2427] — narrower than the ρ by a factor 1.6×10^3 and than the ϕ by a factor 46.

Rests on Theorem 105.9 and Phenomenon 102.41.

Derivation. *Derives Phenomenon 103.14.*

The three resonances differ in one respect: how their valence quarks can reach the final state. The ρ is $u\bar{u}$ and $d\bar{d}$, and its decay $\rho \to \pi\pi$ can proceed with both valence quarks carried into the final-state hadrons, one $q\bar{q}$ pair being created from the vacuum. The ϕ is $s\bar{s}$; its dominant decay $\phi \to K\bar{K}$ likewise carries both valence quarks through, but only just, the phase space being 32 MeV, and its decay to three pions — which requires the $s\bar{s}$ pair to annihilate — is suppressed despite the far larger phase space. The J/ψ is $c\bar{c}$ below the threshold for a pair of charmed mesons, so *every* decay to light hadrons requires the c and $\bar{c}$ to annihilate.

That is the rule of Okubo [1263], Zweig [1301] and Iizuka [1333]: a decay whose quark-line diagram falls apart into two disconnected pieces is suppressed. Stated as an empirical rule it explains the ϕ; stated dynamically it explains the J/ψ and the size of the suppression. The disconnected diagram requires at least three gluons, because one gluon carries colour and cannot make a colour singlet, and two gluons cannot make a state of odd charge conjugation (Theorem 105.9 is the same argument for photons). The rate is therefore of order α_s^6 evaluated at a scale set by the charm mass, and by Phenomenon 102.41 that coupling is already small: $\alpha_s\,(m_c c^2) \approx 0.35$, so $\alpha_s^6 \approx 1.8 \times 10^{-3}$ against a light-quark process running at $\alpha_s \approx 1$. Asymptotic freedom, discovered the year before, is what turns an empirical rule into a number of the right order. □

The interpretation was settled within months by three further facts. The system has radial and orbital excitations with the spacings of a non-relativistic two-body bound state, treatable by the methods of Chapter 82 and testing the static potential of Equation (102.69) directly, as Remark 102.63 sets out. Above 3.73 GeV the annihilation cross section rises to a new plateau, consistent with the opening of a channel with one unit of a new quantum number in each of two final-state hadrons — open charm. And the ratio R of Phenomenon 102.15 steps up by 4/3, which is $3 \times (2/3)^2$: three colours of a quark of charge $+\frac{2}{3}e$. The fourth flavour predicted by Example 103.12 was there, with the charge the GIM mechanism required.

103.3 Three generations

103.3.1 Kobayashi and Maskawa

In 1973, with three quarks known and a fourth conjectural, Kobayashi and Maskawa asked what it would take for the weak interaction to violate CP [1505]. The answer is a counting argument and nothing else, which is what makes it remarkable.

The counting is carried out twice in this book, in Proposition 101.63 and in Theorem 105.25: an $N \times N$ unitary matrix carries $N(N-1)/2$ angles and $(N-1)(N-2)/2$ phases that survive rephasing of the quark fields. It is not repeated here. What belongs here is the consequence and the inference.

Proposition 103.15 (Why the counting forbids CP violation at $N = 2$)**.** *For $N = 2$ the mixing matrix can be brought to real orthogonal form by rephasing, and the charged-current interaction is then invariant under CP; for $N = 3$ one phase survives and CP invariance holds only if that phase is 0 or π.*

Rests on Equations (101.67), (101.68) and (105.16).

Proof. *Derives Proposition 103.15.*

Under CP the charged-current term of Equation (101.67) maps to itself with $V \mapsto V^*$ (Equation (105.16) gives the transformation of the bilinear). If V can be made real by a rephasing that leaves the rest of the Lagrangian invariant, then $V = V^*$ in that basis and the interaction is CP symmetric. Equation (101.68) with $N = 2$ gives $(N-1)(N-2)/2 = 0$ surviving phases, so the Cabibbo matrix is real and CP is conserved however θ_C is chosen; with $N = 3$ one phase survives and cannot be removed, so $V \neq V^*$ in any basis unless the phase is a multiple of π. □

Remark 103.16 (The size of the inference). The measurement that forced this was a single branching ratio of 2×10^{-3} [1274]. From it, and from the requirement that the theory be renormalizable and its gauge structure unchanged, Kobayashi and Maskawa concluded that two further quarks exist. Both were found: the bottom in 1977 [1597] and the top in 1995 [1891, 1892]. It is worth being precise about what was and was not predicted. The *number* three was predicted, as a minimum. The masses were not, and are still not predicted by anything. What the counting delivers is that the observation of CP violation is, on the stated hypotheses, a measurement of the number of generations — and the completely independent measurement of that number from the Z width, Phenomenon 103.22, agrees with it.

Two alternatives existed in 1973 and both are now excluded. A right-handed charged current with a complex coupling could violate CP with two generations, and is excluded by the helicity measurements of Section 101.3.3 and by the direct violation of Section 103.5.2; a superweak interaction acting only in the mixing could violate CP with two generations, and is excluded by $\varepsilon'/\varepsilon \neq 0$. The Kobayashi–Maskawa mechanism is not merely the simplest surviving option, it is the only one of the three that survived.

103.3.2 The tau lepton

In 1975 Perl's group at SPEAR reported events of the form $e^+e^- \to e^\pm\mu^\mp +$ missing energy, with no other charged particle and no photon [1556]. Two features made them a discovery rather than a background. The final state violates electron and muon number separately unless at least two unobserved neutral particles are present; and the events appear only above a threshold near 3.6 GeV in the total energy, which is the signature of pair production of a new particle of mass about $1.8\,\mathrm{GeV}/c^2$.

Phenomenon 103.17 (A third charged lepton)**.** There exists a charged lepton of mass $m_\tau c^2 = (1776.93 \pm 0.09)\,\mathrm{MeV}$ and lifetime $\tau_\tau = 2.903 \times 10^{-13}\,\mathrm{s}$ [2427], with its own neutrino, whose charged-current interaction has the same strength and the same $V - A$ structure as the electron's and the muon's.

Rests on Equation (101.53).

Derivation. *Derives Phenomenon 103.17.*

The universality claim is testable to a fraction of a per cent by one number, and the test is the cleanest in the whole of flavour physics because it involves no hadrons at all. Equation (101.53) gives the rate of a purely leptonic charged-current decay,

$$\Gamma\left(\ell \to \ell' \nu \bar{\nu}\right) = \frac{G_F^2 \left(m_\ell c^2\right)^5}{192\pi^3 \hbar \left(\hbar c\right)^6} f\left(\frac{m_{\ell'}^2}{m_\ell^2}\right), \qquad f(x) = 1 - 8x + 8x^3 - x^4 - 12x^2 \ln x\,, \tag{103.12}$$

whose fifth power of the mass is the dominant factor. Since the τ has hadronic decay channels as well, the prediction is for the *branching fraction*

$$\mathcal{B}\left(\tau \to e \nu \bar{\nu}\right) = \frac{\tau_\tau}{\tau_\mu} \left(\frac{m_\tau}{m_\mu}\right)^5 \frac{f\left(m_e^2/m_\tau^2\right)}{f\left(m_e^2/m_\mu^2\right)}\,. \tag{103.13}$$

Inserting $\tau_\tau = 2.9034 \times 10^{-13}\,\mathrm{s}$, $\tau_\mu = 2.196\,981\,1 \times 10^{-6}\,\mathrm{s}$, $m_\tau c^2 = 1776.93\,\mathrm{MeV}$ and $m_\mu c^2 = 105.658\,375\,5\,\mathrm{MeV}$, so that $(m_\tau/m_\mu)^5 = 1.3453 \times 10^6$ and the phase-space ratio is 1.00019, gives $\mathcal{B} = 0.1778$. The measured value is $\mathcal{B}(\tau \to e\nu\bar{\nu}) = 0.1782(4)$ [2427]. Agreement to two parts in a thousand, over a factor 1.3×10^6 in rate, is what "the same coupling" means quantitatively; expressed as a coupling ratio, g_τ/g_μ differs from unity by less than 2×10^{-3}. □

Derivation pending #153

The phase-space function $f(x) = 1 - 8x + 8x^3 - x^4 - 12x^2 \ln x$ for a purely leptonic charged-current decay with a massive *daughter lepton. The muon-decay rate inherited from the weak-interactions chapter is derived there only for a massless daughter, where $f(0) = 1$; the massive case is the one used above, and the three-body phase-space integral that produces it belongs in Appendix A.*

The τ neutrino was assumed for a quarter of a century and observed directly only in 2000, by the DONUT experiment, which recorded four ν_τ charged-current interactions producing a short track with a kink — the τ decaying within a millimetre of its production point [2016]. Its appearance in an oscillated beam, which is a different measurement, came in 2015 [2247] and belongs to Chapter 114.

Remark 103.18 (The tau as an instrument). The τ is the only lepton heavy enough to decay into hadrons, and that makes it a laboratory for two things this book measures elsewhere. Its hadronic branching fraction, compared with the leptonic one, gives α_s at a scale of 1.78 GeV — the lowest scale at which the coupling is extracted — and the running of that value up to $M_Z c^2$ is one of the entries in Table 102.2. Its decays into $\pi\pi$ measure the same hadronic vacuum polarization that limits the Standard Model prediction of the muon anomaly, which belongs to Section 112.6.3. Neither use was anticipated when the particle was found.

103.3.3 The bottom quark

Lederman's group at Fermilab measured the mass spectrum of muon pairs produced in 400 GeV proton–nucleus collisions and found a structure near $9.5\,\mathrm{GeV}/c^2$ standing above a smooth continuum [1597]. It resolved into the $\Upsilon(1S)$ and $\Upsilon(2S)$, with $m_\Upsilon c^2 = (9.460\,40 \pm 0.000\,10)\,\mathrm{GeV}$ and $\Gamma_\Upsilon = (54.0 \pm 1.3)\,\mathrm{keV}$ [2427]: the same signature as the J/ψ — a resonance far too narrow for a strong decay — one flavour further up.

Three consequences organize the rest of this chapter.

First, bottomonium and charmonium together test the static potential of Equation (102.69) in a way neither can alone. The reduced masses differ by a factor of about three, yet the spacing between the ground state and the first radial excitation is nearly the same in the two systems, $0.589\,\mathrm{GeV}/c^2$ against $0.563\,\mathrm{GeV}/c^2$; Remark 102.63 draws out why only an intermediate, roughly logarithmic, potential does that.

Second, the bottom quark is the one heavy flavour that forms mesons with a *long* lifetime on the scale of a modern detector. A B meson lives $1.52 \times 10^{-12}\,\mathrm{s}$, so at the energies of an asymmetric-energy collider it travels a few hundred micrometres before decaying — far enough to separate its decay vertex from the production point with a silicon detector, and hence to tag which of the two mesons produced in an event decayed first. That single instrumental fact is what makes the time-dependent measurements of Section 103.5.3 possible; without it the CP asymmetry integrates to a much smaller number.

Third, the mixing of the third generation to the first two is small — the b quark *is* the third generation, and what is small is its coupling out of it. The direct determinations are $|V_{cb}| = 0.0408 \pm 0.0014$ and $|V_{ub}| = 0.003\,82 \pm 0.000\,20$ (Table 103.1), and that smallness is

why B mesons live as long as they do. The lifetime and the mixing element are the same measurement seen twice.

103.3.4 The top quark

The top quark was expected from 1977, when the b was found with weak isospin $-\frac{1}{2}$ and therefore in need of a partner, and was not observed until 1995. In between it was measured indirectly, which is the part of the story that belongs to a treatise of evidence.

Phenomenon 103.19 (The top mass was known before the top was seen)**.** Electroweak radiative corrections depend on the top mass quadratically, through the correction $\Delta\rho$ to the ρ parameter of Equation (104.52), and the LEP precision data therefore determined m_t before any top quark was produced. The value inferred from the fit in 1994 was $(178 \pm 20)\,\mathrm{GeV}/c^2$; the direct measurements published the following year gave $(176 \pm 13)\,\mathrm{GeV}/c^2$ [1892] and $(199 \pm 30)\,\mathrm{GeV}/c^2$ [1891], and the current world average is $m_t c^2 = (172.57 \pm 0.29)\,\mathrm{GeV}$ [2427].
Rests on Equation (104.52), Proposition 104.43, and Theorem 104.40.

Derivation. *Derives Phenomenon 103.19.*

The mechanism is Proposition 104.43: a doublet whose two members have different masses breaks the custodial symmetry of Theorem 104.40, and the resulting shift in the ratio of W and Z self-energies grows as the squared mass splitting, $\Delta\rho \propto (m_t^2 - m_b^2)\,c^4$. Because the dependence is quadratic rather than logarithmic, a measurement of the Z lineshape and asymmetries at the per-mille level of Section 104.6.3 pins m_t to about ten per cent. That is a real prediction and it was confirmed: the indirect and direct values agree within their uncertainties, which is a test of the loop structure of the theory and not merely of its tree-level content. The same fit subsequently bounded the Higgs mass, whose contribution is only logarithmic and therefore far weaker — and that bound too was confirmed, in Phenomenon 104.59. □

Two properties of the top are peculiar to it and both follow from its mass.

Proposition 103.20 (The top decays before it hadronizes)**.** *The top width is* $\Gamma_t = 1.42\,\mathrm{GeV}$ *[2427], so its lifetime is*

$$\tau_t = \frac{\hbar}{\Gamma_t} = \frac{6.582\,119\,6 \times 10^{-16}\,\mathrm{eV\,s}}{1.42 \times 10^{9}\,\mathrm{eV}} = 4.6 \times 10^{-25}\,\mathrm{s}\,, \tag{103.14}$$

which is shorter than the time $\hbar/(\Lambda_{\mathrm{QCD}}) \approx 3.3 \times 10^{-24}\,\mathrm{s}$ *that QCD needs to dress a colour charge into a hadron. The top therefore never forms a bound state.*
Rests on Equations (102.45) and (103.1) and Phenomenon 102.65.

Proof. *Derives Proposition 103.20.*

$\Lambda_{\mathrm{QCD}} \approx 0.2\,\mathrm{GeV}$ by Equation (102.45), and the hadronization time is the inverse of that energy in the sense of Equation (103.1), $\hbar/0.2 \times 10^{9}\,\mathrm{eV} = 3.3 \times 10^{-24}\,\mathrm{s}$: seven times Equation (103.14). The width itself scales as m_t^3, since $t \to bW$ is a two-body decay into a real W with the top mass setting both the coupling's phase space and the longitudinal polarization enhancement, so the inequality gets stronger with mass and is a statement about the top specifically. The consequences are experimental and considerable: the top's spin information survives into its decay products, unrandomized by hadronization, and the object reconstructed in a detector is a bare colour triplet in flight — never, by Phenomenon 102.65, an asymptotic state. □

Remark 103.21 (A Yukawa coupling of one). By Theorem 104.38 the coupling of a fermion to the Higgs field is $\hat{y}_f = \sqrt{2}\,m_f c^2/E_v$ with $E_v = 246.22\,\mathrm{GeV}$, so

$$\hat{y}_t = \frac{\sqrt{2} \times 172.57\,\mathrm{GeV}}{246.22\,\mathrm{GeV}} = 0.991\,,$$

a dimensionless number equal to one within its uncertainty, while the electron's is 2.9×10^{-6}. Table 104.2 lists the full range. No principle in the Standard Model relates them, and that the largest of them sits at unity is the kind of fact that looks like a clue and has never been cashed as one. It is listed as open in Section 133.1.

103.3.5 How many generations

Phenomenon 103.22 (There are exactly three light neutrino species)**.** The part of the width of the Z resonance that no observed final state accounts for counts the neutrino species light enough for the Z to decay into them. Combining the scans of the four LEP experiments,

$$N_\nu = 2.984 \pm 0.008 \tag{103.15}$$

[2104]: two standard deviations below three, and more than a hundred standard deviations below four. The count was already unambiguous in the first scan taken [1819]. Any further neutrino must therefore be heavier than $m_Z c^2/2$ or must not couple to the Z at all.

Rests on Equation (101.75).

Derivation. *Derives Phenomenon 103.22.*

The invisible width is the total width minus the sum of the widths into observed final states, and the species count is that quantity divided by the width into one neutrino pair. The latter is not measured; it is computed, and the computation is short.

By Equation (101.75) the Z couples to a fermion f through vector and axial couplings

$$g_V^f = T_f^3 - 2Q_f \sin^2\theta_W\,, \qquad g_A^f = T_f^3\,, \tag{103.16}$$

and the partial width of a vector boson of mass M_Z into a pair of fermions light compared with it is

$$\Gamma\left(Z \to f\bar{f}\right) = N_c^f \,\frac{G_F\left(M_Z c^2\right)^3}{6\pi\sqrt{2}\,\hbar\,(\hbar c)^3}\left[\left(g_V^f\right)^2 + \left(g_A^f\right)^2\right], \tag{103.17}$$

N_c^f being 3 for quarks and 1 for leptons. The dimensional check is worth doing once: $G_F/(\hbar c)^3$ carries $1/\mathrm{J}^2$, multiplying it by an energy cubed leaves an energy, and the remaining $\hbar^{-1}$ converts that energy into a rate — so Equation (103.17) is 1/s, and $\hbar\Gamma$ is the width in energy units usually quoted.

For a neutrino, $T^3 = +\frac{1}{2}$ and $Q = 0$, so $g_V = g_A = \frac{1}{2}$ and the bracket is $\frac{1}{2}$:

$$\hbar\Gamma\left(Z \to \nu\bar{\nu}\right) = \frac{G_F\left(M_Z c^2\right)^3}{12\pi\sqrt{2}\,(\hbar c)^3} = \frac{1.166\,378\,7 \times 10^{-5}/\mathrm{GeV}^2 \times (91.1880\,\mathrm{GeV})^3}{12\pi\sqrt{2}} = 165.9\,\mathrm{MeV}\,, \tag{103.18}$$

the full one-loop value being 167.2 MeV. For a charged lepton, $T^3 = -\frac{1}{2}$ and $Q = -1$, so with $\sin^2\theta_W = 0.2312$ — the $\overline{\mathrm{MS}}$ value at the scale $M_Z c^2$ [2427], the fourth digit being meaningless without that scheme and that scale, which is why Equation (104.62) quotes the same quantity to three figures — the bracket is $(-0.0376)^2 + \left(-\frac{1}{2}\right)^2 = 0.2514$ and $\hbar\Gamma_{\ell\ell} = 83.4\,\mathrm{MeV}$, against the measured $(83.985 \pm 0.086)\,\mathrm{MeV}$ [2104] — the 0.7% shortfall being exactly the radiative correction the tree-level formula omits. The count is then

$$N_\nu = \frac{\Gamma_{\mathrm{inv}}}{\Gamma_{\ell\ell}}\left(\frac{\Gamma_{\ell\ell}}{\Gamma_{\nu\nu}}\right)_{\mathrm{SM}}, \qquad \left(\frac{\Gamma_{\nu\nu}}{\Gamma_{\ell\ell}}\right)_{\mathrm{SM}} = 1.9912\,, \tag{103.19}$$

the tree-level version of which is $0.5/0.2514 = 1.9888$, differing from the full value by one part in a thousand. With the measured $\Gamma_{\mathrm{inv}} = (499.0 \pm 1.5)\,\mathrm{MeV}$ and $\hbar\Gamma_{\ell\ell} = (83.985 \pm 0.086)\,\mathrm{MeV}$ [2104], the ratio is $499.0/83.985 = 5.942$, and dividing by 1.9912 gives $N_\nu = 2.984$, which is Equation (103.15).

The residual 2σ deviation from three deserves a word, because it is not a hint of a fourth neutrino. It is comparable with the radiative corrections just displayed, and the largest of those enters not the widths but the *luminosity*: the t-channel photon exchange in small-angle Bhabha scattering, which is what LEP counted to normalize every cross section. A re-analysis of that effect published in 2019 — not in this book's bibliography, and therefore named without a citation rather than cited to a key that does not exist — raised the central value to 2.996 ± 0.007. The measurement counts three, and the apparent tension was an artefact of the normalization, not of the neutrino sector. □

Derivation pending #154

The tree-level partial width of the Z into a fermion pair. The formula is stated above with its dimensional check and then used to count the neutrino species; its derivation from the neutral-current Lagrangian — the squared matrix element summed over final and averaged over initial polarizations, times the two-body phase space — is not carried out here and belongs in Appendix A.

Remark 103.23 (What the count does and does not exclude). Equation (103.15) counts neutrinos that are *active* — coupling to the Z with the standard strength — and *light*, meaning $m_\nu c^2 < M_Z c^2/2$. A fourth neutrino evading both conditions is not excluded by it, and the searches for one, with their null results, belong to Section 114.6.4, which reserves them. What the count does exclude is a fourth *generation* of the observed kind, since a fourth charged lepton would come with a light active partner. An independent bound of the same strength comes from primordial nucleosynthesis, where the expansion rate at freeze-out — and hence the neutron-to-proton ratio frozen into the light elements — depends on the number of relativistic species. The light-element floor itself is Phenomenon 48.10; that phenomenon states the abundances as a one-parameter family in the baryon-to-photon ratio and does not yet extract a species count from them, so the nucleosynthesis determination of the number of neutrinos is owed to Chapter 48 rather than held there. It agrees with Equation (103.15), which is what makes the number three a fact about Nature rather than about the Z.

And it is a fact without an explanation. Nothing in the Standard Model fixes the number of generations; three is measured, twice, by completely different means, and predicted by nothing. It is listed in Section 133.1.

103.4 The CKM matrix

103.4.1 Structure and parametrization

Definition 103.24 (The CKM matrix)**.** $V_{\mathrm{CKM}} = U_u^\dagger U_d$ is the unitary 3×3 matrix of Theorem 104.38 relating the down-type states that couple to the W to the down-type states of definite mass,

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} . \tag{103.20}$$

Its entries are dimensionless. Four of its nine complex entries are physically independent: three angles and one phase, by Equation (101.68).

The convention that the matrix acts on the down-type states rather than the up-type ones is arbitrary and universal; it makes V_{CKM} the object appearing in Equation (101.67). Two parametrizations are in use, and each is right for a different purpose.

Definition 103.25 (Standard parametrization)**.** With $c_{ij} = \cos\theta_{ij}$, $s_{ij} = \sin\theta_{ij}$ and one phase δ,

$$V = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix} . \tag{103.21}$$

The three angles may be taken in the first quadrant and δ in $[0, 2\pi)$ without loss of generality.

Derivation of the form. *Derives Definition 103.25.*

Write $V = R_{23}(\theta_{23})\,\Phi(\delta)\,R_{13}(\theta_{13})\,\Phi(-\delta)\,R_{12}(\theta_{12})$, where R_{ij} is a real rotation in the ij plane and $\Phi(\delta) = \mathrm{diag}(1, 1, e^{i\delta})$. Any unitary 3×3 matrix can be brought to this form because such a matrix has 9 real parameters — the CKM matrix is unitary with no determinant condition imposed, so the count is that of U(3) and not of SU(3), which has 8 — of which 3 are the angles of the SO(3) subgroup and 6 are phases, and 5 of the phases are removed by rephasing the quark fields, per Proposition 101.63. Multiplying out the five factors gives Equation (103.21). The placement of δ is a convention: any rephasing moves it around the matrix without changing a single observable, since every observable is a rephasing invariant of the type of Equation (105.77). That the phase can be attached to V_{ub}, the smallest element, is what makes the standard parametrization convenient and what makes CP violation small even though δ is of order unity. □

The measured angles are

$$\theta_{12} = 13.00^\circ\,, \qquad \theta_{23} = 2.40^\circ\,, \qquad \theta_{13} = 0.21^\circ\,, \qquad \delta = 65.7^\circ\,, \tag{103.22}$$

[2427]: a strong hierarchy in the angles and a phase of ordinary size. These four numbers are not independent of the four quoted below — they are the same measurement in the other parametrization, and Remark 103.27 carries out the conversion, because a book that quoted both from separate compilations would sooner or later quote two values of one quantity. The hierarchy in the angles is the content of that second parametrization.

Definition 103.26 (Wolfenstein parametrization)**.** Set $\lambda = s_{12}$, $A\lambda^2 = s_{23}$ and $A\lambda^3(\rho - i\eta) = s_{13}e^{-i\delta}$. Then to third order in λ,

$$V = \begin{pmatrix} 1 - \tfrac{1}{2}\lambda^2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1 - \tfrac{1}{2}\lambda^2 & A\lambda^2 \\ A\lambda^3(1 - \rho - i\eta) & -A\lambda^2 & 1 \end{pmatrix} + O\left(\lambda^4\right) , \tag{103.23}$$

with the measured values [2427]

$$\lambda = 0.225\,01 \pm 0.000\,68\,, \quad A = 0.826 \pm 0.017\,, \quad \bar\rho = 0.1591 \pm 0.0094\,, \quad \bar\eta = 0.3523 \pm 0.0076\,, \tag{103.24}$$

where $\bar\rho + i\bar\eta = (\rho + i\eta)\left(1 - \tfrac{1}{2}\lambda^2\right) + O(\lambda^4)$ are the convention-independent combinations that survive to higher order [1733].

Derivation. *Derives Definition 103.26.*

Expand Equation (103.21) with $s_{12} = \lambda$, $s_{23} = A\lambda^2$, $s_{13}e^{-i\delta} = A\lambda^3(\rho - i\eta)$ and $c_{ij} = 1 - \tfrac{1}{2}s_{ij}^2 + \ldots$, keeping terms to $O(\lambda^3)$ in each entry. The $(1,1)$ entry is $c_{12}c_{13} = 1 - \tfrac{1}{2}\lambda^2 + O(\lambda^4)$, the $(1,2)$ entry is $s_{12}c_{13} = \lambda + O(\lambda^5)$, and the $(2,1)$ entry is $-s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} = -\lambda + O(\lambda^6)$, so the upper-left 2×2 block is a rotation through the Cabibbo angle up to $O(\lambda^4)$. The $(3,1)$ entry is $s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} = A\lambda^3 - A\lambda^3(\rho + i\eta) = A\lambda^3(1 - \rho - i\eta)$ to this order, which is the entry carrying the interesting phase. Note that the expansion is not a mathematical approximation imposed for convenience: λ^2 is 0.0506 and λ^4 is 2.6×10^{-3}, so the hierarchy is a property of the measured matrix. □

Remark 103.27 (The standard angles are the apex, converted). Equation (103.22) and Equation (103.24) are one measurement in two coordinate systems, and this book takes the apex as primary because every closure test below is drawn in the $(\bar\rho,\bar\eta)$ plane. Inverting Definition 103.26 with $\rho+\mathrm{i}\eta = (\bar\rho+\mathrm{i}\bar\eta)\left(1-\tfrac{1}{2}\lambda^2\right)^{-1}$ gives $s_{12}=\lambda$, so $\theta_{12}=13.003°$; $s_{23}=A\lambda^2=4.182\times10^{-2}$, so $\theta_{23}=2.397°$; and $s_{13}=A\lambda^3\sqrt{\rho^2+\eta^2}=3.732\times10^{-3}$, so $\theta_{13}=0.2138°$. Note that this uses the *unbarred* pair; the third-order expansion of Example 103.28 uses the barred pair and gets 3.64×10^{-3} for the same quantity, which is the 2.5% discussed there. The phase is fixed by the same relation: $s_{13}\mathrm{e}^{-\mathrm{i}\delta}=A\lambda^3(\rho-\mathrm{i}\eta)$ has the argument of $\bar\rho-\mathrm{i}\bar\eta$, the positive rescaling dropping out, so

$$\delta=\arctan\frac{\bar\eta}{\bar\rho}=\arctan\frac{0.3523}{0.1591}=65.70°\,. \tag{103.25}$$

That is the same arithmetic that produces the angle γ in Phenomenon 103.31, which is why the two agree to within a twentieth of a degree: reconstructing the exact unitary matrix from $(\theta_{12},\theta_{23},\theta_{13},\delta)$ and evaluating $\gamma=\arg(-V_{ud}V_{ub}^*/V_{cd}V_{cb}^*)$ as in Equation (103.31) gives 65.66°, a difference of 0.04° which is the $O(\lambda^4)$ residue. A reader consulting an older compilation will meet $\delta\approx68°$; that number belongs with an earlier apex, and quoting it beside Equation (103.24) would put two mutually inconsistent phases in one chapter.

Example 103.28 (What the λ expansion costs)**.** Four combinations follow from Equation (103.24) once the expansion is truncated at third order:

$$|V_{cb}|=A\lambda^2=0.826\times(0.22501)^2=4.182\times10^{-2}\,, \tag{103.26}$$

$$|V_{ub}|=A\lambda^3\sqrt{\bar\rho^2+\bar\eta^2}=3.64\times10^{-3}\,, \tag{103.27}$$

$$|V_{td}|=A\lambda^3\sqrt{(1-\bar\rho)^2+\bar\eta^2}=8.58\times10^{-3}\,, \tag{103.28}$$

$$J=A^2\lambda^6\bar\eta=3.12\times10^{-5}\,. \tag{103.29}$$

The corresponding numbers from the global fit [2427] — that is, from the exact parametrization of Equation (103.21), not from the expansion — are 4.182×10^{-2}, 3.69×10^{-3}, 8.57×10^{-3} and 3.08×10^{-5}, the last being Equation (105.78).

It is essential to say what this does and does not test. $(\lambda,A,\bar\rho,\bar\eta)$ is a reparametrization of the same four degrees of freedom as $(\theta_{12},\theta_{23},\theta_{13},\delta)$, so nothing here is a test of the CKM matrix, and Equation (103.26) is not a test at all: Definition 103.26 *defines* $A\lambda^2=s_{23}$ and $|V_{cb}|=s_{23}c_{13}$ exactly, so the agreement of the two printed digit for digit is an identity, differing only by $c_{13}=1-7\times10^{-6}$. What the remaining three lines measure is the error of the truncation, and that error is predicted in advance. Writing $\bar\rho,\bar\eta$ in place of ρ,η changes a quantity by a relative $\lambda^2/2=2.5\%$, which is exactly the size of the $O(\lambda^5)$ terms the third-order parametrization drops. In Equations (103.27) and (103.29) the apex coordinates enter directly and the full 2.5% is available; in Equation (103.28) they enter as $1-\bar\rho$, where the same shift is only 0.5%; and in Equation (103.26) they do not enter at all. The line that should be exact is exact, the line that should be accurate to 0.5% is accurate to 0.1%, and the two that should be wrong by up to 2.5% are wrong by 1.4% and 1.3%. The expansion is therefore usable wherever a per-cent statement is wanted, and must be abandoned wherever it is not — which is why every closure test below is stated in terms of the exact matrix.

The genuine test, against numbers that were *not* inputs to the fit, is against the direct determinations of Table 103.1: $|V_{cb}|=0.0408\pm0.0014$ against 4.182×10^{-2} is 0.7σ, $|V_{ub}|=0.003\,82\pm0.000\,20$ against 3.64×10^{-3} is 0.9σ, and $|V_{td}|=0.0086\pm0.0002$ against 8.58×10^{-3} is 0.1σ. Those three comparisons carry the physics; the four lines above carry only the arithmetic of the expansion.

103.4.2 Measured magnitudes

Every element is measured, most of them more than once and by methods sharing no systematics. Table 103.1 collects them with the process each comes from.

Table 103.1: Magnitudes of the CKM elements and the processes that measure them [2427]. The values are the direct determinations, not the outputs of the global unitarity fit, so that the closure tests of Section 103.4.3 are tests. The uncertainty on $|V_{cb}|$ and $|V_{ub}|$ is enlarged over the fit value to cover the inclusive–exclusive tension discussed in the text.

Element	Magnitude	Principal determination
$\lvert V_{ud}\rvert$	$0.973\,67 \pm 0.000\,32$	superallowed $0^+ \to 0^+$ nuclear beta decay; free neutron decay
$\lvert V_{us}\rvert$	0.2243 ± 0.0008	$K \to \pi\ell\nu$ form factor; ratio of $K \to \mu\nu$ to $\pi \to \mu\nu$; hadronic τ decay
$\lvert V_{ub}\rvert$	$0.003\,82 \pm 0.000\,20$	inclusive and exclusive $B \to X_u\ell\nu$
$\lvert V_{cd}\rvert$	0.221 ± 0.004	charm production in neutrino scattering; $D \to \pi\ell\nu$
$\lvert V_{cs}\rvert$	0.975 ± 0.006	$D \to K\ell\nu$; $D_s \to \ell\nu$; W decay to charm
$\lvert V_{cb}\rvert$	0.0408 ± 0.0014	inclusive and exclusive $B \to X_c\ell\nu$
$\lvert V_{td}\rvert$	0.0086 ± 0.0002	B^0–$\bar{B}^0$ oscillation frequency (loop)
$\lvert V_{ts}\rvert$	0.0415 ± 0.0009	B_s–$\bar{B}_s$ oscillation frequency (loop)
$\lvert V_{tb}\rvert$	1.014 ± 0.029	single top production cross section (tree)

Three features of the table matter more than the numbers.

The third row is measured at one loop and at tree level, and the two agree. $|V_{td}|$ and $|V_{ts}|$ come from oscillation frequencies, which are box diagrams of the kind computed in Example 103.12, and therefore assume the loop content of the Standard Model. $|V_{tb}|$ comes from single top production, which is a tree-level charged-current process and assumes nothing of the sort. Consistency between them is a statement that no unknown particle is circulating in the box.

The determination of $|V_{cb}|$ *and* $|V_{ub}|$ *is not settled.* Two methods exist for each. The *inclusive* method sums over all charmed (or charmless) final states and uses an operator product expansion in $1/m_b$; the *exclusive* method measures one channel, $B \to D^*\ell\nu$ or $B \to \pi\ell\nu$, and requires a lattice form factor at the kinematic point of interest. For $|V_{cb}|$ the two differ by about 2 to 3 standard deviations, inclusive high; for $|V_{ub}|$ by a similar amount, again inclusive high. The discrepancy has persisted for two decades, has narrowed as lattice calculations improved without disappearing, and is almost certainly a problem of hadronic theory rather than of the electroweak sector — but it is not resolved, and the enlarged uncertainties in Table 103.1 are an acknowledgement of that rather than a measurement.

Nothing in the table is predicted. Four numbers, and the overall scale of the quark masses, are inputs of the Standard Model. The tests of the next subsection are tests of the *relations* among the nine entries, not of their values.

103.4.3 Unitarity triangle

Unitarity of a 3×3 matrix gives six orthogonality relations between distinct columns or rows. Each is a statement that three complex numbers sum to zero, hence that they close a triangle in the complex plane. Four of the six triangles are extremely flat — one side is smaller than the others by two powers of λ or more — and *two* have all three sides of the

same order. Building the matrix from Equation (103.24) and taking the ratio of the shortest side to the longest gives, for the three column relations (d,s), (d,b), (s,b), the values 0.002, 0.387, 0.020, and for the three row relations (u,c), (u,t), (c,t), the values 0.001, 0.403, 0.046. The (d,b) column triangle, whose sides are all of order $A\lambda^3$, is the one universally called *the* unitarity triangle and is the one defined below; the (u,t) row triangle is if anything marginally less degenerate and has the same area by Proposition 103.30. It is not drawn separately because it carries almost the same information: its angles, computed from the same matrix, are 91.60°, 64.63° and 23.77° against 91.60°, 65.66° and 22.74° for the column triangle, so the largest angle is common to both and the other two differ at $O(\lambda^2)$.

Definition 103.29 (The unitarity triangle)**.** The relation between the first and third columns of Equation (103.20),

$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0\,, \tag{103.30}$$

has three terms of order $A\lambda^3$. Dividing through by $V_{cd}V_{cb}^*$ normalizes the middle side to unit length and puts the triangle's base on the real axis from $(0,0)$ to $(1,0)$; the apex is then at $(\bar\rho,\bar\eta)$. Its interior angles are

$$\alpha = \arg\left(-\frac{V_{td}V_{tb}^*}{V_{ud}V_{ub}^*}\right)\,,\quad \beta = \arg\left(-\frac{V_{cd}V_{cb}^*}{V_{td}V_{tb}^*}\right)\,,\quad \gamma = \arg\left(-\frac{V_{ud}V_{ub}^*}{V_{cd}V_{cb}^*}\right)\,, \tag{103.31}$$

and $\alpha+\beta+\gamma=\pi$ identically.

Proposition 103.30 (Area and the Jarlskog invariant)**.** *The unnormalized triangle of Equation (103.30) has area $J/2$, with J the Jarlskog invariant of Definition 105.26; all six unitarity triangles have the same area. CP is violated if and only if that area is non-zero.*

Rests on Equations (103.30) and (105.77) and Proposition 105.27.

Proof. *Derives Proposition 103.30.*

For three complex numbers $z_1+z_2+z_3=0$, the triangle they close has area $\frac{1}{2}\,|\mathrm{Im}\,(z_1 z_2^*)|$, since the cross product of the two vectors z_1 and $-z_3 = z_1+z_2$ is $\mathrm{Im}\,(z_1\overline{z_1+z_2}) = \mathrm{Im}\,(z_1 z_2^*)$. With $z_1 = V_{ud}V_{ub}^*$ and $z_2 = V_{cd}V_{cb}^*$ this is $\mathrm{Im}\,(V_{ud}V_{ub}^*V_{cd}^*V_{cb})$, which is a Jarlskog quartet and therefore equals $\pm J$ by Equation (105.77). Since Equation (105.77) makes every quartet equal up to sign, and each of the six relations produces one, all six triangles have area $J/2$. A degenerate triangle — three collinear complex numbers — has $J=0$ and by Proposition 105.27 conserves CP. That the area is a rephasing invariant while the individual angles are not is the reason J, and not δ, is the physical measure of CP violation. □

The triangle is now measured many times over, and the measurements are independent in a strong sense: they use different decays, different detectors, and different parts of the theory.

Phenomenon 103.31 (The unitarity triangle closes)**.** The three angles of Equation (103.31) are measured separately, each from a CP asymmetry in a different class of B decay, with results [2427]

$$\alpha = 85^\circ\,,\qquad \beta = 22.2^\circ\,,\qquad \gamma = 66^\circ\,, \tag{103.32}$$

and the two sides, from $|V_{ub}/V_{cb}|$ and from the ratio of B_s to B_d oscillation frequencies, are measured separately again. All of them meet, within uncertainties, at one apex.

Rests on Definition 103.29 and Equations (103.24) and (103.31).

Derivation. *Derives Phenomenon 103.31.*

Two checks, one of the angles and one relating angles to sides.

First the angles. Equation (103.32) sums to $173^\circ \pm 6^\circ$, consistent with 180° at 1.2 standard deviations; and the constraint $\alpha+\beta+\gamma=\pi$ is an identity of Definition 103.29,

not an input to any of the three measurements. That the three numbers, extracted from $B \to \pi\pi$ and $\rho\rho$, from $B \to J/\psi K_S$, and from $B \to DK$ respectively, sum correctly is a non-trivial test.

Now the apex. Definition 103.29 places it at $(\bar\rho, \bar\eta)$, so

$$\tan\beta = \frac{\bar\eta}{1-\bar\rho}\,, \qquad \tan\gamma = \frac{\bar\eta}{\bar\rho}\,.$$

Inserting Equation (103.24) gives $\beta = 22.7°$, hence $\sin 2\beta = 0.713$, and $\gamma = 65.7°$. The directly measured values are $\sin 2\beta = 0.70 \pm 0.02$, from the time-dependent asymmetry in $B^0 \to J/\psi K_S$ [2005, 2010], and $\gamma = 66° \pm 3°$, from interference between $b \to c$ and $b \to u$ amplitudes in $B \to DK$. Both agree. The first of them is the sharper statement, because $\sin 2\beta$ is measured to three per cent in a tree-dominated decay whose theoretical uncertainty is below one per cent — and it agrees with a value obtained from Δm_d, Δm_s, ε_K and $|V_{ub}/V_{cb}|$, that is from three loop processes and one tree process, none of them a CP asymmetry.

What is being tested is the following. There are four parameters. The observables listed number well over a dozen. Any one of them could have landed outside the region the others define, and that would have falsified the hypothesis that a single phase in a unitary matrix accounts for all CP violation in the quark sector. None has. The global fits [2074] quantify the agreement, and the residual freedom in the apex is now a few per cent in each coordinate. □

Remark 103.32 (Where the closure is weakest). Three qualifications keep Phenomenon 103.31 honest. The angle α carries a 5° uncertainty dominated by the hadronic "penguin" amplitudes that must be subtracted from $B \to \pi\pi$, which is theory, not statistics. The side $|V_{ub}/V_{cb}|$ inherits the inclusive–exclusive tension of Section 103.4.2, and the two choices move the apex by about one standard deviation. And the first-row relation Δ_{CKM} of Equation (103.6), which is the most precisely measured unitarity statement of all, is the one that does not come out right. The correct summary is that the triangle closes at the few-per-cent level and the first row does not close at the 10^{-3} level; both are measurements and both are recorded.

103.5 Neutral meson mixing and CP violation

103.5.1 Neutral kaon mixing

A neutral meson carrying a flavour quantum number has an antiparticle distinct from itself, and the weak interaction, which does not conserve that quantum number, connects the two. The system is then a two-state quantum system with decay, and its behaviour is the most sensitive interferometer physics possesses.

Definition 103.33 (Neutral-meson mixing)**.** Let P^0 and $\bar P^0$ be a neutral meson and its antiparticle, distinguished by a flavour quantum number conserved by the strong interaction. In the two-dimensional subspace they span, the effective time evolution is

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\begin{pmatrix} a \\ \bar a \end{pmatrix} = \left(\mathsf{M}c^2 - \frac{\mathrm{i}\hbar}{2}\Gamma\right)\begin{pmatrix} a \\ \bar a \end{pmatrix}\,, \tag{103.33}$$

with M and Γ Hermitian 2×2 matrices, the first carrying kg and the second 1/s. The eigenvalues of the non-Hermitian matrix in brackets define two states of definite mass and width, split by Δm and $\Delta\Gamma$, and the two dimensionless ratios

$$x := \frac{\Delta m\, c^2}{\hbar\bar\Gamma}\,, \qquad y := \frac{\Delta\Gamma}{2\bar\Gamma}\,, \qquad \bar\Gamma := \tfrac{1}{2}\left(\Gamma_1 + \Gamma_2\right)\,, \tag{103.34}$$

decide whether mixing is observable: $x \ll 1$ and $y \ll 1$ means the meson decays before it can oscillate.

The derivation of Equation (103.33) from the Weisskopf–Wigner treatment of an unstable state, the constraints that CPT invariance and CP invariance impose on M and Γ, and the explicit diagonalization are Proposition 105.22 and Proposition 105.23; they are not repeated here. What belongs here is the physics the four measured systems display.

Table 103.2: The four neutral-meson systems, with the mixing parameters of Equation (103.34) [2427]. Only the strange system has $y \approx 1$, because only there is a decay channel open to one CP eigenstate and closed to the other; only the charm system has $x \ll 1$, because there the GIM suppression of Theorem 103.11 runs over down-type quarks, whose masses are all small.

System	Quarks	x	y	Heaviest quark in the box
K^0–$\bar{K}^0$	$d\bar{s}$	0.946	0.997	charm (top suppressed by mixing)
D^0–$\bar{D}^0$	$c\bar{u}$	0.0041	0.0065	bottom
B^0–$\bar{B}^0$	$d\bar{b}$	0.770	0.001	top
B_s–$\bar{B}_s$	$s\bar{b}$	27.0	0.062	top

Gell-Mann and Pais saw the essential point in 1955, before any of it was measured [1067]. If CP were conserved, the states of definite mass and lifetime would be the CP eigenstates

$$|K_1\rangle = \frac{1}{\sqrt{2}}\left(\left|K^0\right\rangle - \left|\bar{K}^0\right\rangle\right), \qquad |K_2\rangle = \frac{1}{\sqrt{2}}\left(\left|K^0\right\rangle + \left|\bar{K}^0\right\rangle\right), \tag{103.35}$$

with $CP = +1$ and -1 respectively in the phase convention of Equation (105.45). The two-pion final state has $CP = +1$ and the three-pion state $CP = -1$ (Equation (105.47)), and the three-pion channel has almost no phase space — $m_K c^2 - 3m_\pi c^2 = 83\,\mathrm{MeV}$ against 219 MeV for two pions. So K_1 should decay fast and K_2 slowly. The prediction was of a *new particle*, of the same mass and opposite CP, and it was found: the measured lifetimes are

$$\tau_S = 8.954 \times 10^{-11}\,\mathrm{s}, \qquad \tau_L = 5.116 \times 10^{-8}\,\mathrm{s}, \qquad \frac{\tau_L}{\tau_S} = 571, \tag{103.36}$$

which is Equation (105.48).

Proposition 103.34 (The mass difference is measured as a phase)**.** *A beam that is pure K^0 at $t = 0$ contains, at proper time t, a $\bar{K}^0$ component of probability*

$$P\left(K^0 \to \bar{K}^0\right)(t) = \frac{1}{4}\left[\mathrm{e}^{-\Gamma_S t} + \mathrm{e}^{-\Gamma_L t} - 2\mathrm{e}^{-(\Gamma_S+\Gamma_L)t/2}\cos\frac{\Delta m_K c^2 t}{\hbar}\right], \tag{103.37}$$

so Δm_K is read off an interference pattern in time, not from a mass measurement. This is why $\Delta m_K c^2 = (3.484 \pm 0.006) \times 10^{-6}\,\mathrm{eV}$ is known to three significant figures although it is a fraction 7.0×10^{-15} of the kaon mass itself.

Rests on Definition 103.33, Equation (103.35)*, and Proposition 101.56.*

Proof. *Derives Proposition 103.34.*

Write $|K^0\rangle = (|K_L\rangle + |K_S\rangle)/\sqrt{2}$ and $\left|\bar{K}^0\right\rangle = (|K_L\rangle - |K_S\rangle)/\sqrt{2}$, which inverts Equation (103.35) up to the CP-violating admixture of order 2×10^{-3} that is neglected here and restored in Section 103.5.2. Each mass eigenstate evolves with its own phase and its own damping, $|K_i(t)\rangle = \mathrm{e}^{-\mathrm{i}m_i c^2 t/\hbar}\mathrm{e}^{-\Gamma_i t/2}|K_i\rangle$, so

$$\left\langle \bar{K}^0 \middle| K^0(t)\right\rangle = \frac{1}{2}\left(\mathrm{e}^{-\mathrm{i}m_L c^2 t/\hbar - \Gamma_L t/2} - \mathrm{e}^{-\mathrm{i}m_S c^2 t/\hbar - \Gamma_S t/2}\right),$$

whose squared modulus is Equation (103.37). The strangeness of the beam at time t is measured directly, by the sign of the charged lepton in a semileptonic decay: the rule $\Delta S = \Delta Q$ (Proposition 101.56) says a K^0 gives ℓ^+ and a $\bar{K}^0$ gives ℓ^-. Counting the two charges as a function of decay position along a beam line therefore traces Equation (103.37), and the cosine's period measures Δm_K. With $x = \Delta m_K c^2/\hbar\bar{\Gamma} = 0.946$ from Table 103.2, roughly one oscillation occurs per lifetime, which is the optimum: much less and there is no interference, much more and it averages away. □

Remark 103.35 (Regeneration). A beam of K_L, purified by letting the K_S component die away, recovers a K_S component on passing through matter. The mechanism is that the strong interaction sees K^0 and $\bar{K}^0$, not K_L and K_S: the $\bar{K}^0$ can convert a nucleon's quark content in ways the K^0 cannot, so the forward scattering amplitudes f and $\bar{f}$ differ. Writing $|K_L\rangle \propto |K^0\rangle + \left|\bar{K}^0\right\rangle$, the transmitted beam is $f\,|K^0\rangle + \bar{f}\left|\bar{K}^0\right\rangle \propto \left(f+\bar{f}\right)|K_L\rangle + \left(f-\bar{f}\right)|K_S\rangle$, so the regenerated amplitude is proportional to $f-\bar{f}$ and vanishes only if the strong interaction is blind to strangeness, which it is not. The effect, Equation (105.64), is the single largest background to the measurement of Section 103.5.2 and had to be eliminated by evacuating the decay volume.

103.5.2 Discovery of CP violation

Phenomenon 103.36 (The long-lived kaon decays to two pions)**.** The neutral kaon system contains a short-lived state decaying to two pions and a long-lived state, some six hundred times longer lived, decaying to three. In a beam allowed to travel far enough that the short-lived component has died away, a small fraction of the surviving decays are nevertheless to $\pi^+\pi^-$: the branching fraction is about 2×10^{-3} [1274, 2427]. The experiment belongs to Chapter 111.

Rests on Definition 103.33 and Equations (103.35) *and* (105.47).

Derivation. Derives Phenomenon 103.36.

The pion has intrinsic parity -1, so a state of two pions with relative orbital angular momentum L has parity $(-1)^2(-1)^L = (-1)^L$; the kaon has spin zero, so $L = 0$ and the parity is $+1$. Charge conjugation exchanges π^+ with π^-, which for $L = 0$ is the identity, so the state also has $C = +1$. The two-pion state is therefore an eigenstate of CP with eigenvalue $+1$. The three-pion state available to the same system has, by the same counting together with the small energy release that suppresses non-zero orbital angular momenta, $CP = -1$. If CP were a symmetry of the decay interaction the two final states would be reached from two orthogonal kaon states, and whichever state decays to three pions could never decay to two. That one and the same long-lived state does both is thus a direct proof that CP is not conserved [1274]. What the argument does not settle is where the violation sits — in the mixing that defines the long-lived state, or in the decay amplitude itself. Separating them required the measurement of ϵ'/ϵ a generation later, and only that measurement excluded a superweak interaction acting solely on the mixing. □

The separation is parametrized by two complex numbers. The CP-violating admixture in the mixing is ε, defined by Equation (105.54); the CP violation in the decay amplitude itself is ε', defined through the two isospin amplitudes of the two-pion final state by Equation (105.72). The measured values are

$$|\varepsilon| = (2.228 \pm 0.011) \times 10^{-3}\,, \qquad \mathrm{Re}\left(\frac{\varepsilon'}{\varepsilon}\right) = (1.66 \pm 0.23) \times 10^{-3}\,, \tag{103.38}$$

the first from the ratio of K_L to K_S two-pion rates and the second from the double ratio of charged to neutral two-pion modes in the two beams — a quantity constructed precisely so that detector acceptance and beam normalization cancel (Proposition 105.24). The

result was established independently by NA48 [1973] and KTeV [1965] after two decades of measurements that disagreed with each other.

Remark 103.37 (What $\varepsilon' \neq 0$ excluded). Wolfenstein's superweak hypothesis [1300] proposed a new $\Delta S = 2$ interaction acting only in the mass matrix, which would produce $\varepsilon \neq 0$ and $\varepsilon' = 0$ exactly. It fits the 1964 measurement perfectly, requires no third generation, and survived for thirty-five years. Equation (103.38) killed it: ε'/ε is small but non-zero at more than seven standard deviations, so CP violation occurs in the decay amplitude and not only in the mixing. This is the point at which the Kobayashi–Maskawa mechanism ceased to have a live competitor. The Standard Model prediction for ε'/ε is a difference of two large and partially cancelling hadronic contributions and remains uncertain at the level of tens of per cent; what the measurement establishes is the *sign and existence* of direct CP violation, not a precision test.

103.5.3 B mixing and the B factories

In 1987 the ARGUS collaboration at DESY, studying $\Upsilon(4S)$ decays to $B\bar{B}$ pairs, counted events in which both B mesons decayed semileptonically with leptons of the *same* sign [1787]. Same-sign dileptons require one of the two mesons to have changed into its own antiparticle. The measured ratio of same-sign to opposite-sign pairs was $r = 0.21$, and the consequence was immediate and unwelcome.

Proposition 103.38 (Time-integrated mixing probability)**.** *For a system with $y \approx 0$, the probability that a meson born as P^0 decays as $\bar{P}^0$, integrated over all decay times, is*

$$\chi = \frac{x^2}{2\,(1+x^2)}\,, \qquad \textit{so} \qquad r = \frac{\chi}{1-\chi} = \frac{x^2}{2+x^2}\,. \tag{103.39}$$

ARGUS's $r = 0.21$ therefore gives $x_d = 0.73$.

Rests on Equations (103.34) *and* (103.37).

Proof. *Derives Proposition 103.38.*

From Equation (103.37) with $\Gamma_L = \Gamma_S = \Gamma$, the mixed and unmixed probabilities are $\tfrac{1}{2}\mathrm{e}^{-\Gamma t}\left[1 \mp \cos\left(\Delta m\, c^2 t/\hbar\right)\right]$. Using $\int_0^\infty \mathrm{e}^{-\Gamma t}\mathrm{d}t = \Gamma^{-1}$ and $\int_0^\infty \mathrm{e}^{-\Gamma t}\cos\omega t\,\mathrm{d}t = \Gamma/\left(\Gamma^2+\omega^2\right)$ with $\omega = \Delta m\, c^2/\hbar = x\Gamma$, and normalizing by the total Γ^{-1},

$$\chi = \frac{1}{2}\left(1 - \frac{1}{1+x^2}\right) = \frac{x^2}{2\,(1+x^2)}\,.$$

Solving $r = \chi/(1-\chi) = 0.21$ for x gives $x^2 = 2r/(1-r) = 0.532$, hence $x = 0.73$. □

Example 103.39 (The top mass from B mixing)**.** The box diagram generating Δm_d is the one of Example 103.12 with the external quarks d and b, and by Theorem 103.11 it is dominated by the heaviest internal quark, which for a b-flavoured meson is the top. The amplitude is no longer in the regime $m_i \ll M_W$, so the full loop function is needed:

$$\Delta m_d c^2 = \frac{G_F^2\left(M_W c^2\right)^2}{6\pi^2\left(\hbar c\right)^6}\,\eta_B\left(m_B c^2\right)\left(f_B\sqrt{B_B}\,c^2\right)^2 |V_{tb}V_{td}^*|^2\, S_0(x_t)\,, \tag{103.40}$$

with $x_t = \left(m_t c^2\right)^2/\left(M_W c^2\right)^2$, $\eta_B = 0.55$ the short-distance QCD factor, and

$$S_0(x) = x\left[\frac{1}{4} + \frac{9}{4\,(1-x)} - \frac{3}{2\,(1-x)^2}\right] - \frac{3x^3\ln x}{2\,(1-x)^3}\,,$$

which reduces to $S_0(x) \to x$ for $x \ll 1$, recovering Equation (103.8). With $m_t c^2 = 172.57\,\text{GeV}$ and $M_W c^2 = 80.369\,\text{GeV}$ one has $x_t = 4.61$ and $S_0 = 2.52$; inserting $f_B\sqrt{B_B}\,c^2 = 210\,\text{MeV}$, $m_B c^2 = 5.2797\,\text{GeV}$ and $|V_{td}| = 8.57 \times 10^{-3}$ gives

$$\Delta m_d c^2 = 3.51 \times 10^{-4}\,\text{eV}\,, \tag{103.41}$$

against the measured $3.336 \times 10^{-4}\,\text{eV}$ — five per cent, which is the accuracy of the lattice input.

In the dimensionless variable of Equation (103.34), with $\tau_B = 1.519 \times 10^{-12}\,\text{s}$, Equation (103.41) corresponds to $x_d = \Delta m_d c^2 \tau_B/\hbar = 0.81$, against the measured 0.770 of Table 103.2 — the same five per cent, now carried into the quantity ARGUS actually reported.

Now run it backwards to 1987. With $m_t c^2 = 30\,\text{GeV}$, the value then expected, $S_0 = 0.129$ and Equation (103.41) falls by a factor 19.5, giving $x_d = 0.042$ against the 0.73 that ARGUS measured. The hadronic inputs were poorly known in 1987, but not by a factor of eighteen in a squared amplitude, and the conclusion drawn at once was that the top quark is far heavier than anyone had supposed.

Read with today's inputs, ARGUS's number becomes a determination of the top mass, and it is worth doing the inversion rather than asserting its result. Since $\Delta m_d \propto S_0(x_t)$ with everything else held fixed, ARGUS's $x_d = 0.73$ requires $S_0 = 2.5226 \times 0.73/x_d^{\text{ref}}$, where x_d^{ref} is whatever the rest of the formula produces today. Anchoring on the formula itself, $x_d^{\text{ref}} = 0.81$, gives $S_0 = 2.27$, hence $x_t = 4.03$ and $m_t c^2 = 161\,\text{GeV}$; anchoring instead on the *measured* $x_d^{\text{ref}} = 0.770$, which removes the five-per-cent lattice offset, gives $S_0 = 2.39$, hence $x_t = 4.30$ and $m_t c^2 = 167\,\text{GeV}$. The two bracket the answer: ARGUS's mixing rate says $m_t c^2$ is between about 160 GeV and 170 GeV, against the 172.57 GeV eventually measured, eight years before the top was produced — the second time in this chapter that a loop measured a quark that could not yet be made. The gap between the two anchors is the honest uncertainty of the inference, and it is far smaller than the uncertainty the hadronic inputs carried in 1987; the 1987 statement was correctly made as an inequality, and only today's lattice numbers turn it into a measurement.

Derivation pending #155

The loop function $S_0(x)$ and the B-meson box amplitude for the mass difference. Both are stated above and then used twice to weigh the top quark, so the derivation is load-bearing: the box integral for an internal mass comparable with the W mass, whose value is S_0, is not carried out here and belongs in Appendix A.

Proposition 103.40 (The mixing ratio measures a ratio of CKM elements)**.** *Taking the ratio of Equation (103.40) for the B_s and B_d systems, the loop function, the QCD factor and the W mass cancel exactly, leaving*

$$\frac{\Delta m_s}{\Delta m_d} = \frac{m_{B_s}}{m_{B_d}}\,\xi^2\left|\frac{V_{ts}}{V_{td}}\right|^2, \qquad \xi := \frac{f_{B_s}\sqrt{B_s}}{f_{B_d}\sqrt{B_d}}\,. \tag{103.42}$$

The lattice quantity $\xi = 1.208$ is a ratio *of decay constants, in which most of the systematic uncertainty of the individual calculations cancels; this is what makes Equation (103.42) the best determination of $|V_{td}/V_{ts}|$.*

Rests on Equation (103.40).

Proof. *Derives Proposition 103.40.*

Divide Equation (103.40) by its B_s counterpart; every factor independent of the light spectator quark cancels. Inverting with the measured oscillation frequencies $\Delta m_d c^2/\hbar =$

$5.069 \times 10^{11}/\mathrm{s}$ and $\Delta m_s c^2/\hbar = 1.7765 \times 10^{13}/\mathrm{s}$ [2202, 2427], that is $\Delta m_d c^2 = 3.336 \times 10^{-4}\,\mathrm{eV}$ and $\Delta m_s c^2 = 1.169 \times 10^{-2}\,\mathrm{eV}$,

$$\left|\frac{V_{td}}{V_{ts}}\right| = \left[\frac{\Delta m_d}{\Delta m_s}\,\frac{m_{B_s}}{m_{B_d}}\right]^{1/2} \xi = \left[\frac{1}{35.05} \times 1.0165\right]^{1/2} \times 1.208 = 0.206\,,$$

against $8.6 \times 10^{-3}/4.15 \times 10^{-2} = 0.207$ from the individual entries of Table 103.1. Note the units. The literature quotes these as angular frequencies $\Delta mc^2/\hbar$ in 1/ps — 0.5069/ps and 17.765/ps — while this book quotes the energy splitting Δmc^2 in eV. The conversion is $\hbar = 6.582\,119\,6 \times 10^{-16}\,\mathrm{eV\,s}$, so 1/ps corresponds to $6.582 \times 10^{-4}\,\mathrm{eV}$, and the ratio the derivation uses is of course the same either way. □

The asymmetric-energy B factories were built for one measurement. An $\Upsilon(4S)$ produced at rest decays to a $B\bar{B}$ pair almost at rest, and the two mesons travel some 30 µm before decaying — too little to separate. Colliding beams of unequal energy gives the pair a Lorentz boost along the beam axis, stretching the separation to a few hundred micrometres, which a silicon vertex detector resolves. The time between the two decays is then measurable, and with it the time-dependent asymmetry of Equation (105.81).

Phenomenon 103.41 (Large CP violation in the B system)**.** In $B^0 \to J/\psi\, K_S$ the time-dependent decay-rate asymmetry between mesons tagged at production as B^0 and as $\bar{B}^0$ is an oscillation of amplitude

$$S_{J/\psi K_S} = \sin 2\beta = 0.70 \pm 0.02\,, \tag{103.43}$$

observed independently by BaBar [2010] and Belle [2005] in 2001. The effect is of order unity, against 2×10^{-3} in the kaon system, and it agrees with the value predicted from measurements containing no CP asymmetry at all.

Rests on Proposition 105.28 and Equations (103.31) *and* (103.40).

Derivation. Derives Phenomenon 103.41.

Proposition 105.28 gives the asymmetry as $S_f \sin(\Delta m_d c^2 t/\hbar) - C_f \cos(\Delta m_d c^2 t/\hbar)$ with $S_f = 2\,\mathrm{Im}\,\lambda/\left(1 + |\lambda|^2\right)$ and $\lambda = (q/p)\left(\bar{\mathcal{A}}_f/\mathcal{A}_f\right)$. For $f = J/\psi K_S$ three things conspire to make this clean. The decay $b \to c\bar{c}s$ is dominated by a single tree amplitude, so $\left|\bar{\mathcal{A}}/\mathcal{A}\right| = 1$ to within a per cent and $C_f \approx 0$; the mixing phase q/p is, from the box of Equation (103.40), twice the phase of $V_{tb}^* V_{td}$; and the K^0–$\bar{K}^0$ mixing in the final state contributes twice the phase of $V_{cs}^* V_{cd}$. Collecting,

$$\lambda = \mathrm{e}^{-2\mathrm{i}\beta}\,, \qquad \beta = \arg\left(-\frac{V_{cd}V_{cb}^*}{V_{td}V_{tb}^*}\right)\,,$$

which is exactly the angle of Equation (103.31); hence $S = \sin 2\beta$ with no hadronic quantity anywhere in the relation. That is the whole reason the measurement is decisive: the hadronic matrix elements cancel between the two interfering paths because both end in the same final state.

Why is the effect a thousand times larger than in the kaon? Because CP violation requires interference between two amplitudes of comparable size and different weak phase. In the kaon, the CP-violating amplitude is a small admixture in a state dominated by CP-conserving physics, so the asymmetry carries the small factor ε. In $B^0 \to J/\psi K_S$ the two interfering paths — decay without mixing, and decay after mixing — have *equal* magnitude by construction, since $|q/p| = 1$ and $|\lambda| = 1$, so the asymmetry is the full sin of the phase difference. Nothing is small except the phase, and the phase is not small. □

Mixing in the B_s system, with $x_s = 27$, oscillates twenty-seven times per lifetime and required a detector with picosecond decay-time resolution; it was resolved at the Tevatron and measured precisely at LHCb, which also observed CP violation in B_s decays [2202]. The value of Δm_s enters Equation (103.42) and is one of the constraints on the apex in Phenomenon 103.31.

103.5.4 Charm and the present picture

The charm system was the last to yield. By Theorem 103.11, a D^0 box runs over d, s and b quarks, all light compared with M_W, so the GIM suppression is severe and x and y are of order 10^{-3} (Table 103.2); the CP-violating part is suppressed further by the small mixing factor $|V_{cb}V_{ub}| \sim 1.5 \times 10^{-4}$.

Phenomenon 103.42 (CP violation in charm decays)**.** The difference of CP asymmetries between $D^0 \to K^+K^-$ and $D^0 \to \pi^+\pi^-$ is

$$\Delta A_{CP} = A_{CP}\left(K^+K^-\right) - A_{CP}\left(\pi^+\pi^-\right) = (-15.4 \pm 2.9) \times 10^{-4}\,, \tag{103.44}$$

non-zero at 5.3 standard deviations [2323]. With it, CP violation is established in all three quark systems in which it can occur.

Rests on Theorem 103.11 and Definition 103.25.

Derivation. *Derives Phenomenon 103.42.*

Taking the difference of two asymmetries rather than either alone is what makes the measurement possible: production asymmetries between D^0 and $\bar{D}^0$ in pp collisions, and detection asymmetries between positive and negative tracks, are common to the two final states and cancel in the difference, while the CP-odd parts do not, since U-spin relates the two channels with opposite sign. The residual is then a genuine asymmetry between a decay and its conjugate. Its magnitude, 1.5×10^{-3}, is of the order the Kobayashi–Maskawa phase predicts once the small mixing factor and the interference with the dominant Cabibbo-favoured amplitude are accounted for — but the hadronic matrix elements of charm decays are not calculable to better than a factor of a few, so the correct statement is that the observation is *consistent with* the Standard Model and does not test it sharply. □

Remark 103.43 (One phase, everything measured)*.* It is worth stating plainly what Section 103.5 has established. Every CP-violating observable measured in the quark sector — ε and ε' in the kaon, $\sin 2\beta$, γ and α in the B mesons, the B_s asymmetries, and Equation (103.44) in charm — is described by the single phase of Definition 103.25, whose rephasing-invariant measure is $J = 3.08 \times 10^{-5}$. Four parameters, well over a dozen independent measurements, one consistent solution. The places where the description is least sharply tested are those where hadronic matrix elements dominate the uncertainty: ε'/ε, the charm asymmetry just quoted, and the angle α. The places where it is sharply tested — $\sin 2\beta$, γ, the ratio $\Delta m_s/\Delta m_d$ — are those where the hadronic physics cancels.

103.5.5 What CP violation cannot yet explain

Phenomenon 103.44 (The universe contains matter and almost no antimatter)**.** No astrophysical concentration of antimatter is observed, and the baryon density inferred from primordial nucleosynthesis agrees with the one inferred from the acoustic peaks of the microwave background on a baryon-to-photon ratio

$$\eta = \frac{n_B - n_{\bar{B}}}{n_\gamma} = (6.12 \pm 0.04) \times 10^{-10} \tag{103.45}$$

[2348]. Generating such an asymmetry from a symmetric initial state requires baryon-number violation, violation of C and of CP, and a departure from thermal equilibrium [1358]. The

CP violation established in Phenomenon 103.36 is genuine but quantitatively far too small — by eleven powers of ten — to produce η.

Rests on Phenomenon 103.36, Theorem 105.41, and Proposition 105.42.

Derivation. *Derives Phenomenon 103.44.*

The three Sakharov conditions are derived as Theorem 105.41 and the value of η from $\Omega_b h^2$ as Proposition 105.42; neither is repeated. What is needed here is the size of the effect the CKM phase can produce, and the argument is dimensional.

By Proposition 105.27, all CP violation in the quark sector is proportional to the invariant $\det\left[M_u M_u^\dagger c^4, M_d M_d^\dagger c^4\right]$, which carries J multiplied by the six squared-mass differences of Equation (105.79) and has SI dimension J^{12}. Any asymmetry generated at a temperature T is a pure number, so it can depend on that invariant only through the dimensionless ratio Equation (105.106), and evaluating it with the running quark masses at $k_B T \approx 100\,\mathrm{GeV}$, where the baryon-number-violating processes are active, gives Equation (105.107),

$$d_{CP} \approx 2 \times 10^{-21}\,, \tag{103.46}$$

to be compared with $\eta = 6.1 \times 10^{-10}$: the ratio $\eta / d_{CP} = 3 \times 10^{11}$, a shortfall of eleven orders of magnitude, and detailed calculations [1882] do not improve it. The reason is structural. CP violation in this theory requires all six quark masses to be distinct, and at 100 GeV five of the six are negligible against the temperature, so the effect is suppressed by twelve powers of small mass ratios. The phase is of order unity; the masses kill it.

The third Sakharov condition fails as well, and independently. A departure from equilibrium requires the electroweak transition to be first order, and for the measured Higgs mass of Equation (104.66) it is not: it is a smooth crossover, as Remark 104.65 records. Two of the three conditions are therefore not met by the Standard Model with its measured parameters. The origin of Equation (103.45) is unknown, and this treatise says so; it is carried to Section 133.5. □

Remark 103.45 (Why this is stated as a failure and not as a puzzle solved). Mechanisms that would generate the observed asymmetry exist in outline — heavy Majorana neutrinos decaying out of equilibrium, a first-order transition driven by additional scalars — and none has any experimental support. They are named in Section 133.5 and in Section 103.7.4 as hypotheses without evidence, which is what they are. What is *measured* is Equation (103.45), and what is *calculated* is Equation (103.46); the eleven orders of magnitude between them are one of the few places where the Standard Model is known, from laboratory numbers plus one cosmological observation, to be incomplete.

103.6 Lepton flavour and neutrino mixing

103.6.1 The PMNS matrix

Pontecorvo asked in 1957 whether the neutral-kaon phenomenon of Section 103.5.1 has a leptonic analogue [1127, 1159]. His original proposal was $\nu \leftrightarrow \bar{\nu}$ oscillation, which is not what happens; but the idea that a neutral particle produced in a weak interaction need not be a state of definite mass, and can therefore convert into another, is exactly what happens. The correct form was written by Maki, Nakagawa and Sakata in 1962 [1237], the year the second neutrino flavour was established [1223]: flavour states are superpositions of mass states, and the superposition matrix plays for leptons the role V_{CKM} plays for quarks.

Definition 103.46 (The PMNS matrix)**.** In the basis in which the charged-lepton mass matrix is diagonal, the neutrino flavour states are

$$|\nu_\alpha\rangle = \sum_{i=1}^{3} U_{\alpha i}^{*}\, |\nu_i\rangle\,, \qquad \alpha \in \{e, \mu, \tau\}\,, \tag{103.47}$$

with U unitary and dimensionless. Written in the parametrization of Equation (103.21) with angles $\theta_{12}, \theta_{23}, \theta_{13}$ and Dirac phase δ_{CP}, it carries in addition a diagonal factor

$$U = U^{(\text{Dirac form})} \times \operatorname{diag}\left(1, \mathrm{e}^{\mathrm{i}\alpha_{21}/2}, \mathrm{e}^{\mathrm{i}\alpha_{31}/2}\right) , \tag{103.48}$$

present if and only if the neutrinos are Majorana particles.

Proposition 103.47 (Majorana phases do not affect oscillation)**.** *The two phases α_{21}, α_{31} of Equation (103.48) cancel from every oscillation probability, and from every process in which lepton number is conserved. No oscillation experiment, however precise, can determine whether the neutrino is a Dirac or a Majorana particle.*

Rests on Definition 103.46 and Equation (103.48).

Proof. *Derives Proposition 103.47.*

Every oscillation amplitude has the structure $\sum_i U_{\beta i} \mathrm{e}^{-\mathrm{i}\varphi_i} U^*_{\alpha i}$, in which each index i appears once on U and once on U^*. Under Equation (103.48), $U_{\alpha i} \mapsto U_{\alpha i} \mathrm{e}^{\mathrm{i}\alpha_{i1}/2}$ for every α, so the product $U_{\beta i} U^*_{\alpha i}$ acquires $\mathrm{e}^{\mathrm{i}\alpha_{i1}/2} \mathrm{e}^{-\mathrm{i}\alpha_{i1}/2} = 1$. The same cancellation occurs in any amplitude with one neutrino line entering and one leaving, that is, in any lepton-number-conserving process. It fails only where two neutrino lines are contracted with each other, which requires a Majorana propagator and violates lepton number by two units — the process of Section 103.7.2, and the reason that is the only practical test.

The counting is worth recording. For N Dirac neutrinos the phase count is $(N-1)(N-2)/2$, as for quarks, because the neutrino fields may be rephased. For N Majorana neutrinos a rephasing $\nu_i \to \mathrm{e}^{\mathrm{i}\phi}\nu_i$ is forbidden — it would multiply the Majorana mass term $\overline{\nu^c}\nu$ by $\mathrm{e}^{2\mathrm{i}\phi}$ — so only the charged-lepton fields may be rephased, $N-1$ of those phases being usable, and the surviving count is $N(N-1)/2$: three phases for $N = 3$, one Dirac and two Majorana. □

Phenomenon 103.48 (Lepton mixing is large)**.** The magnitudes of the PMNS entries are, from the global fit to all oscillation data [2356, 2427],

$$|U| \approx \begin{pmatrix} 0.82 & 0.55 & 0.15 \\ 0.46 & 0.50 & 0.73 \\ 0.33 & 0.67 & 0.67 \end{pmatrix} , \tag{103.49}$$

against the near-diagonal $|V_{\mathrm{CKM}}|$ of Table 103.1. Two of the three lepton mixing angles are of order tens of degrees; all three quark mixing angles are small.

Rests on Definition 103.46 and Equations (103.21) and (103.22).

Derivation. *Derives Phenomenon 103.48.*

Equation (103.49) follows by inserting the measured $\sin^2\theta_{12} = 0.307$, $\sin^2\theta_{23} = 0.546$ and $\sin^2\theta_{13} = 0.0220$ of Table 103.3 into Equation (103.21), with δ_{CP} set to zero, which shifts the entries of the lower two rows by at most a few units in the second decimal. The corresponding angles are

$$\theta_{12} = 33.6^\circ , \qquad \theta_{23} = 47.6^\circ , \qquad \theta_{13} = 8.5^\circ ,$$

to be compared with Equation (103.22), where the largest is 13.0° and the smallest 0.21°. The comparison is not between two numbers but between two *patterns*: the quark matrix is a small perturbation of the identity, with a hierarchy $1 : \lambda : \lambda^3$; the lepton matrix has no hierarchy at all except that one of its nine entries is small. □

Remark 103.49 (Two matrices, one theory, no explanation). The Standard Model requires both matrices to be unitary — the argument is Remark 5.152 in both cases — and fixes nothing else about either. Their entries are Yukawa couplings in disguise, by Theorem 104.38,

and the Yukawa couplings are free parameters. That one matrix comes out nearly diagonal and the other nearly democratic is therefore a fact with no theoretical content in the Standard Model whatsoever. Attempts to relate them exist and none has evidence; the fact is recorded as open in Section 133.1. It should not be dressed up as a hint.

103.6.2 Oscillation in vacuum

Phenomenon 103.50 (Neutrinos change flavour)**.** A neutrino created with one flavour is detected, some distance later, with another. Three independent classes of observation establish it. The flux of atmospheric ν_μ depends on zenith angle, that is on the distance travelled, in a way no production model imitates [1954]. The solar ν_e flux measured by charged-current reactions is about a third of the total active flux measured by the neutral-current reaction on deuterium, and the total agrees with the solar model [2021] — which resolves the deficit Davis reported thirty years earlier [1368] without appeal to solar physics at all. Reactor $\bar{\nu}_e$ disappear and reappear as a function of L/E [2049]. Flavour is therefore not conserved in propagation, and the neutrino masses are not all zero. Lepton mixing is moreover large, two of its three angles being of order tens of degrees [2427], in flat contrast with the small angles of Section 103.4. The experiments are Chapter 114.

Rests on Definition 103.46 and Phenomenon 103.48.

Derivation. *Derives Phenomenon 103.50.*

Take two flavour states related to two states of definite mass $m_1 \neq m_2$ by a single angle,

$$|\nu_\alpha\rangle = \cos\theta\,|\nu_1\rangle + \sin\theta\,|\nu_2\rangle\ , \qquad |\nu_\beta\rangle = -\sin\theta\,|\nu_1\rangle + \cos\theta\,|\nu_2\rangle\ . \tag{103.50}$$

The mass states, not the flavour states, are the ones that propagate with a definite phase. For a neutrino of energy E far above its rest energy, $E_i = \sqrt{p^2c^2 + m_i^2c^4} \approx E + m_i^2c^4/2E$, so after travelling a distance L in a time L/c the component $|\nu_i\rangle$ has acquired the phase

$$\varphi_i = \frac{E_i L}{\hbar c} = \frac{EL}{\hbar c} + \frac{m_i^2 c^3 L}{2\hbar E}\,. \tag{103.51}$$

The first term is common to both components and unobservable. The amplitude to find the other flavour is the projection of the propagated state on $|\nu_\beta\rangle$,

$$A_{\alpha\to\beta} = \sin\theta\cos\theta\left(\mathrm{e}^{-\mathrm{i}\varphi_2} - \mathrm{e}^{-\mathrm{i}\varphi_1}\right)\ ,$$

and with $\left|\mathrm{e}^{-\mathrm{i}\varphi_2} - \mathrm{e}^{-\mathrm{i}\varphi_1}\right|^2 = 2\left[1 - \cos(\varphi_1 - \varphi_2)\right] = 4\sin^2\tfrac{1}{2}(\varphi_1 - \varphi_2)$ its squared modulus is

$$P_{\alpha\to\beta} = \sin^2 2\theta\,\sin^2\!\left(\frac{\Delta m^2 c^3 L}{4\hbar E}\right)\ , \qquad \Delta m^2 := m_2^2 - m_1^2\,. \tag{103.52}$$

Three consequences follow, and all three are what is seen. The effect is an interference between two mass components, so it measures a difference of squared masses and never a mass: the oscillation experiments cannot say how heavy a neutrino is, only that two of them differ. It vanishes if the masses are equal or if $\theta = 0$, so its observation proves both that the mass states are non-degenerate and that the flavour and mass bases are misaligned. And it depends on L and E only through their ratio, which is why the atmospheric, reactor and accelerator data — taken at baselines from tens of metres to thousands of kilometres — collapse onto one curve when plotted against L/E. □

Remark 103.51 (Where the derivation is loose, and why it survives). Equation (103.52) was obtained by giving the two mass components the same momentum and letting them travel the same distance in the same time. Each of those is an approximation, and each

has generated a literature. A neutrino is produced localized — in a nucleus, in a pion decay — so its state is a wave packet, and the components of different mass travel at slightly different group velocities and eventually cease to overlap. The correct statement is that Equation (103.52) is the limit of a wave-packet calculation under two conditions, both of which hold by enormous margins in every experiment discussed here.

Localization. The production and detection regions must be small compared with the oscillation length Equation (103.55): otherwise the phase varies over the source and the pattern washes out. What sets σ_x for a reactor antineutrino is not the size of the decaying nucleus, which is of order $1 \times 10^{-15}\,\mathrm{m}$, but the extent over which the decaying system is coherently localized — a fission fragment bound in the fuel lattice, whose position spread is of atomic order, $\sigma_x \sim 1 \times 10^{-11}\,\mathrm{m}$. Against an oscillation length of 99 km either number is negligible, so the localization condition is never in question. The coherence condition is the one that has to be checked, and there the distinction between the two matters by eight orders of magnitude.

Coherence. The packets must still overlap on arrival, which requires

$$L \lesssim L_{\mathrm{coh}} = \frac{4\sqrt{2}\,\sigma_x E^2}{\Delta m^2 c^4}\,, \tag{103.53}$$

the separation of the packets growing as $L\,\Delta m^2 c^4/2E^2$ because that is the difference of their group velocities. For the same reactor case, $\sigma_x = 1 \times 10^{-11}\,\mathrm{m}$, $E = 3\,\mathrm{MeV}$ and $\Delta m^2 c^4 = 7.53 \times 10^{-5}\,\mathrm{eV}^2$ give $L_{\mathrm{coh}} = 6.8 \times 10^3\,\mathrm{km}$, against KamLAND's 180 km. The margin is a factor of 38, and it is the atomic σ_x that supplies it: inserting the nuclear $1 \times 10^{-15}\,\mathrm{m}$ instead would give $L_{\mathrm{coh}} = 0.68\,\mathrm{km}$, so that coherence would be lost some 260 times over before the detector and no oscillatory dependence on L/E could be seen at all. KamLAND sees one, so the relevant localization is the atomic one — which is a measurement of σ_x, not an assumption about it. Note that Equation (103.53) is not the same statement as the averaging of Equation (103.52) over a spread in L or E: that is a classical loss of resolution, this is a quantum loss of coherence, and in every existing experiment the first sets in long before the second.

Derivation pending #156

The wave-packet coherence length for neutrino oscillation, including its factor $4\sqrt{2}$. What is given inline is the group-velocity separation, which fixes the scaling with the packet width, the energy and the squared-mass splitting; the Gaussian overlap integral that fixes the numerical coefficient, and with it the precise meaning of the packet width, is not carried out here and belongs in Appendix A.

The chapter's SI discipline has a payoff here, because the number every experimental paper uses is a disguised statement about $\hbar$ and c and is almost never derived.

Proposition 103.52 (The oscillation phase in practical units)**.** *With the squared-mass splitting expressed as an energy squared in* eV^2*, the baseline in kilometres and the energy in* GeV*,*

$$\frac{\Delta m^2 c^3 L}{4\hbar E} = \frac{(\Delta m^2 c^4)\,L}{4\hbar c\,E} = 1.26693\,\frac{\left(\Delta m^2 c^4/\mathrm{eV}^2\right)\,(L/\mathrm{km})}{(E/\mathrm{GeV})}\,, \tag{103.54}$$

and the oscillation length, the distance over which the phase advances by π, is

$$L_{\mathrm{osc}} = \frac{4\pi\hbar c E}{\Delta m^2 c^4} = 2.4797\,1\,\mathrm{km}\,\frac{(E/\mathrm{GeV})}{\left(\Delta m^2 c^4/\mathrm{eV}^2\right)}\,. \tag{103.55}$$

Rests on Equations (103.1) *and* (103.52)*.*

Proof. *Derives Proposition 103.52.*

Multiply numerator and denominator of the phase in Equation (103.52) by c: the numerator becomes $\Delta m^2 c^4$, an energy squared, and the denominator $4\hbar c E$, also an energy squared multiplied by a length. Hence

$$\varphi = \frac{(\Delta m^2 c^4)\, L}{4\,\hbar c\, E}\,,$$

which is manifestly dimensionless when L is a length. Now insert $\hbar c = 1.973\,269\,8 \times 10^{-7}\ \mathrm{eV\,m}$ from Equation (103.1), and convert the units of the three inputs: $L = (L/\mathrm{km}) \times 1 \times 10^{3}\ \mathrm{m}$ and $E = (E/\mathrm{GeV}) \times 1 \times 10^{9}\ \mathrm{eV}$. Then

$$\varphi = \frac{1 \times 10^{3}\ \mathrm{m}}{4 \times 1.973\,269\,8 \times 10^{-7}\ \mathrm{eV\,m} \times 1 \times 10^{9}\ \mathrm{eV}} \times \frac{\left(\Delta m^2 c^4/\mathrm{eV}^2\right)\,(L/\mathrm{km})}{(E/\mathrm{GeV})}\,,$$

and the prefactor is the pure number

$$\frac{1 \times 10^{3}\ \mathrm{m}}{4 \times 1.973\,269\,8 \times 10^{-7}\ \mathrm{eV\,m} \times 1 \times 10^{9}\ \mathrm{eV}} \times \mathrm{eV}^2 = \frac{1 \times 10^{3}\ \mathrm{m}}{789.308\ \mathrm{eV}^2\,\mathrm{m}} \times \mathrm{eV}^2 = 1.26693\,.$$

This is the constant that appears as "1.27" in every experimental paper on the subject, where it is quoted rather than derived; it is $(4\hbar c)^{-1}$ in disguise, and nothing else. Setting $\varphi = \pi$ and solving for L gives $\pi/1.26693 = 2.4797$, which is Equation (103.55). □

Example 103.53 (Three baselines)**.** Equation (103.55) explains the design of every experiment treated in Chapter 114. Only two of the three splittings are independent, since $\Delta m^2_{31} = \Delta m^2_{32} + \Delta m^2_{21}$ identically; this book tabulates Δm^2_{21} and Δm^2_{32} in Table 103.3 and writes Δm^2_{31} for the sum, which exceeds $|\Delta m^2_{32}|$ by 3% in the normal ordering — immaterial wherever a two-figure statement is being made, which is everywhere below except Equation (103.75). For the atmospheric splitting $|\Delta m^2_{32}|\,c^4 = 2.455 \times 10^{-3}\ \mathrm{eV}^2$ and $E = 1\ \mathrm{GeV}$, $L_{\mathrm{osc}} = 1010\ \mathrm{km}$: hence accelerator baselines of 295 km (T2K) and 810 km (NOvA), and hence the atmospheric signal appearing as a deficit of upward-going and not downward-going muon neutrinos, since the Earth's diameter is 12 742 km and its atmosphere 15 km thick. For the solar splitting $7.53 \times 10^{-5}\ \mathrm{eV}^2$ at a reactor energy of 3 MeV, $L_{\mathrm{osc}} = 99\ \mathrm{km}$: hence KamLAND at 180 km. For the *atmospheric* splitting at the same reactor energy, $L_{\mathrm{osc}} = 3.0\ \mathrm{km}$: hence Daya Bay's far detectors at 1.6 km, which is where the θ_{13} oscillation is near its first maximum while the solar term is still negligible. The separation of the two splittings by a factor 33 is what allows each to be measured almost independently of the other.

Theorem 103.54 (Three-flavour oscillation probability)**.** *For three flavours mixed by Equation (103.47),*

$$\begin{aligned} P_{\alpha\to\beta} = \delta_{\alpha\beta} &- 4\sum_{i>j} \mathrm{Re}\!\left(U^*_{\alpha i} U_{\beta i} U_{\alpha j} U^*_{\beta j}\right) \sin^2\!\left(\frac{\Delta m^2_{ij} c^4 L}{4\hbar c E}\right) \\ &+ 2\sum_{i>j} \mathrm{Im}\!\left(U^*_{\alpha i} U_{\beta i} U_{\alpha j} U^*_{\beta j}\right) \sin\!\left(\frac{\Delta m^2_{ij} c^4 L}{2\hbar c E}\right)\,, \end{aligned} \tag{103.56}$$

with $\Delta m^2_{ij} = m^2_i - m^2_j$. The second sum is odd under $\alpha \leftrightarrow \beta$ and changes sign between neutrinos and antineutrinos; it is the leptonic CP-violating term.

Rests on Equations (103.47) and (103.51).

Proof. *Derives Theorem 103.54.*

The amplitude is $A_{\alpha\to\beta} = \sum_i U_{\beta i} \mathrm{e}^{-\mathrm{i}m_i^2 c^3 L/2\hbar E} U^*_{\alpha i}$, the common phase of Equation (103.51) having been dropped. Then

$$P = |A|^2 = \sum_{i,j} U_{\beta i} U^*_{\alpha i} U^*_{\beta j} U_{\alpha j} \mathrm{e}^{-\mathrm{i}\Delta m_{ij}^2 c^3 L/2\hbar E} .$$

Split the double sum into $i = j$, which gives $\sum_i |U_{\alpha i}|^2 |U_{\beta i}|^2$, and the off-diagonal terms, which pair (i,j) with (j,i). Writing $W_{ij} = U^*_{\alpha i} U_{\beta i} U_{\alpha j} U^*_{\beta j}$, the coefficient standing in front of $\mathrm{e}^{-\mathrm{i}\varphi_{ij}}$ in the display above is $U_{\beta i} U^*_{\alpha i} U^*_{\beta j} U_{\alpha j} = (U^*_{\alpha i} U_{\beta i}) \left(U_{\alpha j} U^*_{\beta j}\right) = W_{ij}$ itself, and the (j,i) term is its complex conjugate because $W_{ji} = W^*_{ij}$ and $\varphi_{ji} = -\varphi_{ij}$. The paired contribution is therefore $2\,\mathrm{Re}\,(W_{ij}\mathrm{e}^{-\mathrm{i}\varphi_{ij}}) = 2\,\mathrm{Re}\,W_{ij}\cos\varphi_{ij} + 2\,\mathrm{Im}\,W_{ij}\sin\varphi_{ij}$ with $\varphi_{ij} = \Delta m_{ij}^2 c^3 L/2\hbar E$ — the sign of the second term being the sign of the CP-violating term in Equation (103.56), so it is worth stating why it is a plus: $\mathrm{e}^{-\mathrm{i}\varphi}$ contributes $-\mathrm{i}\sin\varphi$, and $\mathrm{Re}(-\mathrm{i}W) = +\,\mathrm{Im}\,W$. Using $\cos\varphi = 1 - 2\sin^2(\varphi/2)$ and the unitarity relation $\sum_i U^*_{\alpha i} U_{\beta i} = \delta_{\alpha\beta}$, whose squared modulus supplies $\sum_i |U_{\alpha i}|^2 |U_{\beta i}|^2 + 2\sum_{i>j} \mathrm{Re}\,W_{ij} = \delta_{\alpha\beta}$, gives Equation (103.56). Replacing U by U^*, which is what CP conjugation does, flips the sign of the imaginary part and leaves the real part alone, which is the last statement. □

Corollary 103.55 (The two-flavour limit is a good approximation, and why)**.** *Because* $|\Delta m^2_{31}|/\Delta m^2_{21} = 34$ *and* $|U_{e3}|^2 = 0.022$ *is small, Equation* (103.56) *reduces to a two-flavour formula in each of two regimes. At* L/E *tuned to the atmospheric splitting the solar phase is negligible and* $P(\nu_\mu \to \nu_\mu) \approx 1 - \sin^2 2\theta_{23} \sin^2(\Delta m^2_{31} c^4 L/4\hbar c E)$ *with* $\sin^2 2\theta_{23} = 0.99$*; at* L/E *tuned to the solar splitting the atmospheric phase averages to one half and* $P(\bar\nu_e \to \bar\nu_e) \approx \cos^4\theta_{13}\left[1 - \sin^2 2\theta_{12}\sin^2(\Delta m^2_{21} c^4 L/4\hbar c E)\right] + \sin^4\theta_{13}$*. This factorization, and not any approximation made for convenience, is why the parameters could be measured one at a time over thirty years.*

Rests on Equation (103.56)*.*

103.6.3 Matter effects

A neutrino crossing ordinary matter forward-scatters coherently off the electrons, protons and neutrons in it. Coherent forward scattering does not change the neutrino's momentum and produces no observable interaction; what it produces is a refractive index, exactly as for light in glass, and hence a shift in the propagation phase. The shift would be irrelevant if it were the same for all three flavours. It is not: matter contains electrons and contains no muons or taus, so ν_e alone can forward-scatter through the *charged* current, and its potential differs from the others'.

Proposition 103.56 (The matter potential)**.** *In a medium of electron number density* n_e*, an electron neutrino acquires the extra potential energy*

$$V_e = \sqrt{2}\, G_F\, n_e\,, \tag{103.57}$$

of SI dimension J*, while* ν_μ *and* ν_τ *acquire none from the charged current. An antineutrino acquires* $-V_e$*. Numerically, at the centre of the Sun where* $n_e = 6\times10^{31}/\mathrm{m}^3$*,*

$$V_e = \sqrt{2}\times 1.435\,85\times10^{-62}\,\mathrm{J\,m^3}\times 6\times10^{31}/\mathrm{m^3} = 1.22\times10^{-30}\,\mathrm{J} = 7.6\times10^{-12}\,\mathrm{eV}\,. \tag{103.58}$$

Rests on Equations (101.6) *and* (101.8) *and Definition 100.2.*

Proof. *Derives Proposition 103.56.*

The relevant term of the effective Lagrangian of Equation (101.6), Fierz-rearranged so that the neutrino fields are paired with each other, is

$$\mathcal{L}_{\mathrm{eff}} = -\frac{G_F}{\sqrt{2}}\left[\bar\psi_{\nu_e}\gamma^\mu\left(\mathbb{1} - \gamma^5\right)\psi_{\nu_e}\right]\left[\bar\psi_e\gamma_\mu\left(\mathbb{1} - \gamma^5\right)\psi_e\right]\,,$$

of dimension $\mathrm{J/m^3}$, since each bilinear carries $1/\mathrm{m^3}$ by Definition 100.2 and G_F carries $\mathrm{J\,m^3}$. Average the electron bilinear over the medium. For an unpolarized, isotropic electron gas at rest, the spatial components of the vector current average to zero and the axial current averages to the net spin density, also zero; the surviving component is $\langle \bar{\psi}_e \gamma^0 \psi_e \rangle = n_e$, the electron number density. The neutrino bilinear with γ^0 counts the neutrino number density, so the average Lagrangian is a potential energy per neutrino of $V_e = \sqrt{2} G_F n_e$: one factor of 2 from the $(\mathbb{1} - \gamma^5)$ acting on a left-chiral neutrino, against the $1/\sqrt{2}$ of the Fermi normalization. The neutral current contributes as well, but identically for all three flavours, so it adds a multiple of the identity to the Hamiltonian and cannot affect oscillation — the reason only the charged-current term appears in Equation (103.57). For an antineutrino, C conjugation reverses the sign of the vector current, giving $-V_e$; this sign is the whole of the mass-ordering sensitivity discussed below. The numerical evaluation is Equation (103.58), using G_F in SI from Equation (101.8) — not $G_F/(\hbar c)^3$, the confusion Definition 101.8 exists to prevent. □

Theorem 103.57 (Mixing in matter and the MSW resonance)**.** *For two flavours with vacuum mixing angle θ and splitting Δm^2, propagation through matter of electron density n_e is governed by an effective mixing angle θ_m and an effective splitting Δm_m^2 given by*

$$\tan 2\theta_m = \frac{\Delta m^2 c^4 \sin 2\theta}{\Delta m^2 c^4 \cos 2\theta - 2 E V_e}\,, \tag{103.59}$$

$$\Delta m_m^2 c^4 = \left[\left(\Delta m^2 c^4 \cos 2\theta - 2 E V_e \right)^2 + \left(\Delta m^2 c^4 \sin 2\theta \right)^2 \right]^{1/2} . \tag{103.60}$$

The mixing is maximal, $\sin^2 2\theta_m = 1$, at the resonance

$$2 E V_e = \Delta m^2 c^4 \cos 2\theta\,, \tag{103.61}$$

however small the vacuum angle θ, provided $\cos 2\theta > 0$; the resonance occurs for neutrinos if $\Delta m^2 > 0$ and for antineutrinos if $\Delta m^2 < 0$ [1631, 1779].

Rests on Equation (103.50) and Proposition 103.56.

Proof. *Derives Theorem 103.57.*

The evolution of the flavour amplitudes with time is $\mathrm{i}\hbar\,\partial_t \Psi = \mathcal{H}\Psi$, and for an ultrarelativistic neutrino of fixed momentum the free part of $\mathcal{H}$ is $\sqrt{p^2c^2 + m_i^2 c^4} \approx pc + m_i^2 c^4 / 2E$ in the mass basis. Dropping the term proportional to the identity and rotating to the flavour basis with Equation (103.50),

$$\mathcal{H} = \frac{\Delta m^2 c^4}{4E} \begin{pmatrix} -\cos 2\theta & \sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{pmatrix} + \begin{pmatrix} V_e & 0 \\ 0 & 0 \end{pmatrix} ,$$

the matter term being diagonal in the *flavour* basis by Proposition 103.56. Subtracting $\tfrac{1}{2} V_e \mathbb{1}$, which shifts both levels equally and changes no probability,

$$\mathcal{H} = \frac{1}{2} \begin{pmatrix} -A & B \\ B & A \end{pmatrix} , \qquad A = \frac{\Delta m^2 c^4 \cos 2\theta}{2E} - V_e\,, \qquad B = \frac{\Delta m^2 c^4 \sin 2\theta}{2E}\,.$$

A real symmetric traceless 2×2 matrix of this form is diagonalized by a rotation through θ_m with $\tan 2\theta_m = B/A$, which after multiplying numerator and denominator by $2E$ is Equation (103.59); its eigenvalue difference is $\sqrt{A^2 + B^2}$, which multiplied by $2E$ is Equation (103.60). The angle is $\pi/4$ exactly when $A = 0$, which is Equation (103.61); and since $V_e > 0$ for matter, A can vanish only if $\Delta m^2 \cos 2\theta > 0$. For an antineutrino $V_e \to -V_e$ and the condition becomes $\Delta m^2 \cos 2\theta < 0$. The resonance therefore occurs for neutrinos or for antineutrinos according to the *sign* of the splitting, and that is the only place in this chapter where a sign, rather than a magnitude, is observable. □

Example 103.58 (The solar resonance energy)**.** Insert the solar values $\Delta m^2_{21}c^4 = 7.53 \times 10^{-5}\,\mathrm{eV}^2$, $\sin^2\theta_{12} = 0.307$ so that $\cos 2\theta_{12} = 1 - 2\sin^2\theta_{12} = 0.386$, and Equation (103.58), into Equation (103.61):

$$E_{\mathrm{res}} = \frac{\Delta m^2 c^4 \cos 2\theta}{2V_e} = \frac{7.53 \times 10^{-5}\,\mathrm{eV}^2 \times 0.386}{2 \times 7.6 \times 10^{-12}\,\mathrm{eV}} = 1.9\,\mathrm{MeV}\,. \tag{103.62}$$

That number organizes the whole solar neutrino programme. Neutrinos produced well below 1.9 MeV — the *pp* neutrinos, with endpoint 420 keV, which dominate the flux — never reach resonance and oscillate essentially as in vacuum, their survival probability averaging to

$$P_{ee}\big|_{\text{low } E} = 1 - \tfrac{1}{2}\sin^2 2\theta_{12} = 0.57\,.$$

Neutrinos produced well above it — the ^{8}B neutrinos, extending to 15 MeV, which are the ones Homestake and SNO detected — are created above resonance, where the matter term dominates and $\theta_m \to \pi/2$, so the neutrino leaves the Sun as the heavier mass eigenstate ν_2 and its survival probability is

$$P_{ee}\big|_{\text{high } E} = |U_{e2}|^2 = \sin^2\theta_{12} = 0.31\,.$$

Both numbers are measured, the second by SNO (Phenomenon 114.5) and the first by Borexino, which detected the *pp* neutrinos in real time [2229] and later the CNO neutrinos [2349]. A single curve, rising from 0.31 to 0.57 as the energy falls through a few MeV, passes through every measured point. Both expressions are the two-flavour limit; restoring the third angle multiplies each by $\cos^4\theta_{13} = 0.957$ and adds $\sin^4\theta_{13} = 4.8 \times 10^{-4}$, moving them to 0.55 and 0.29, which is the four-per-cent accuracy at which the two-flavour statement should be read. It is worth being clear about what this establishes that vacuum oscillation could not: the *sign* of Δm^2_{21}, and hence that ν_2 is heavier than ν_1 — equivalently, that θ_{12} lies in the first octant. Vacuum oscillation is even in Δm^2 and blind to both.

Proposition 103.59 (Adiabaticity)**.** *The neutrino follows the instantaneous eigenstate through the resonance, rather than jumping between eigenstates, provided*

$$\gamma := \frac{\Delta m^2 c^4 \sin^2 2\theta}{2E\cos 2\theta}\,\frac{L_\rho}{\hbar c} \gg 1\,, \qquad L_\rho := \left|\frac{\mathrm{d}\ln n_e}{\mathrm{d}r}\right|^{-1}_{\mathrm{res}}\,, \tag{103.63}$$

and for the Sun, with $L_\rho = 6.6 \times 10^7$ m and the values of Example 103.58, $\gamma = 1.5 \times 10^4$.
Rests on Theorem 103.57 and Equation (103.60).

Proof. *Derives Proposition 103.59.*

The adiabatic theorem states that a system in a non-degenerate instantaneous eigenstate of a slowly varying Hamiltonian stays in it, with a transition probability that vanishes as the variation is slowed; the general statement and its proof belong to Section 82.6.1, which reserves them. Its condition is that the rate of change of the Hamiltonian's eigenbasis be small compared with the level splitting divided by $\hbar$. Here the eigenbasis is fixed by θ_m, and the fastest variation is at resonance. Differentiating $\tan 2\theta_m = B/A$ from the proof of Theorem 103.57, in which only A depends on r,

$$\frac{\mathrm{d}\,(2\theta_m)}{\mathrm{d}r}\bigg|_{A=0} = -\frac{1}{B}\frac{\mathrm{d}A}{\mathrm{d}r} = \frac{V_e}{B}\frac{\mathrm{d}\ln n_e}{\mathrm{d}r} = \cot 2\theta\,\frac{\mathrm{d}\ln n_e}{\mathrm{d}r}\,,$$

the last step using $V_e = \Delta m^2 c^4 \cos 2\theta/2E$ at resonance; hence $\mathrm{d}\theta_m/\mathrm{d}r = \cot 2\theta/\,(2L_\rho)$ in magnitude. The two eigenvalues of the traceless Hamiltonian there are $\pm\Delta m^2_m c^4/4E$ with $\Delta m^2_m c^4 = \Delta m^2 c^4 \sin 2\theta$ by Equation (103.60), so the ratio of the phase rate to the rotation rate is

$$\gamma = \frac{\Delta m^2_m c^4}{4E\hbar c}\left(\frac{\mathrm{d}\theta_m}{\mathrm{d}r}\right)^{-1} = \frac{\Delta m^2 c^4 \sin^2 2\theta}{2E\cos 2\theta}\,\frac{L_\rho}{\hbar c}\,.$$

Numerically, with $\Delta m^2 c^4 \sin^2 2\theta = 7.53 \times 10^{-5}\,\mathrm{eV}^2 \times 0.851 = 6.41 \times 10^{-5}\,\mathrm{eV}^2$, $2E\cos 2\theta = 2 \times 1.91 \times 10^6\,\mathrm{eV} \times 0.386 = 1.47 \times 10^6\,\mathrm{eV}$ and $L_\rho/\hbar c = 6.6 \times 10^7\,\mathrm{m}/1.9733 \times 10^{-7}\,\mathrm{eV\,m} = 3.34 \times 10^{14}/\mathrm{eV}$, the product is 1.5×10^4. The solar density profile is exponential over the relevant region, which is what makes L_ρ a constant, and the margin is four orders of magnitude: the conversion is adiabatic and the survival probabilities of Example 103.58 are exact in that limit. □

Remark 103.60 (Two further matter signatures). Two consequences of Theorem 103.57 are measured or being measured, and both belong to Section 114.5.2, which reserves them.

Day–night asymmetry. Solar neutrinos arriving at night cross the Earth, whose electron density partially regenerates the ν_e component, so the night-time rate should exceed the day-time rate by a few per cent. The effect is at the edge of the sensitivity of existing detectors and is observed at about the 2σ level; it is a prediction of the mechanism with no free parameter beyond the already-measured ones.

The mass ordering. By the sign statement of Theorem 103.57, matter enhances $\nu_\mu \to \nu_e$ and suppresses $\bar\nu_\mu \to \bar\nu_e$ if $\Delta m^2_{31} > 0$, and does the opposite if it is negative. A long-baseline beam run alternately in neutrino and antineutrino mode therefore determines the ordering — but the same comparison is also where δ_{CP} shows up, by Theorem 103.54, and the two effects are degenerate at a single baseline. Breaking the degeneracy requires either two baselines or a baseline long enough ($\gtrsim 1000\,\mathrm{km}$) that the matter effect dominates. That is the design brief of the next generation of experiments, and it is why the ordering is not yet known.

103.6.4 The evidence

The experimental case belongs to Chapter 114 and is summarized here in the order in which it was made, because the order is itself instructive: a thirty-year anomaly, then two measurements that admitted no escape.

The solar deficit, 1968–1998. Davis's chlorine detector, 615 t of perchloroethylene 1478 m underground, counted about one third of the ν_e capture rate the standard solar model predicted [1368, 1950]. Because the ^{8}B flux it was sensitive to depends on roughly the twenty-fifth power of the solar core temperature, the natural reading for twenty years was that the solar model was wrong [1365]. The gallium experiments removed that escape by reaching the *pp* neutrinos, whose flux is fixed by the measured solar luminosity almost independently of the model, and found a deficit there too [1974, 2138]. This is Phenomenon 114.3.

The atmospheric zenith-angle dependence, 1998. Super-Kamiokande measured the ν_μ and ν_e rates as functions of arrival direction, and found the muon neutrinos arriving from below — having crossed 13 000 km of Earth — depleted by about half relative to those from above, at the same energy, while the electron neutrinos showed no such dependence [1954]. Because the same detector measures both, and because the deficit is a function of L/E and not of L or E separately, no production model and no absorption mechanism reproduces it. This is Phenomenon 114.4, whose derivation there uses exactly Equation (103.52).

The neutral-current measurement, 2002. SNO measured, in one detector, the ν_e flux by a charged-current reaction on deuterium and the total active flux by the flavour-blind neutral-current breakup of the deuteron. The first came to about a third of the second, and the second agreed with the solar model [2008, 2021]. The subtraction Equation (114.35) contains no solar-model input at all: the missing electron neutrinos arrive as ν_μ and ν_τ, and Phenomenon 114.5 is the statement.

Reactor disappearance and reappearance, 2003. KamLAND observed electron antineutrinos from Japanese power reactors at an average baseline of 180 km with a survival probability of about 0.6, and — decisively — found that probability to be an *oscillatory* function of L/E rather than a monotone suppression [2049]. A flux normalization error produces a

deficit; only interference produces a wave. This is Phenomenon 114.6, and it converted the solar result from a statement about the Sun into a statement about neutrinos.

The third angle, 2012. Daya Bay and RENO compared near and far identical detectors at reactor complexes and measured $\sin^2 2\theta_{13} \approx 0.09$ [2188, 2189], which is Phenomenon 114.8. Many had expected it to vanish; that it does not is the precondition for ever observing δ_{CP}, since by Theorem 103.54 the CP-odd term is proportional to the leptonic Jarlskog invariant $J_\ell = \frac{1}{8}\sin 2\theta_{12}\sin 2\theta_{23}\sin 2\theta_{13}\cos\theta_{13}\sin\delta_{CP}$ and therefore to $\sin\theta_{13}$.

Accelerator beams, 2006–. K2K [2089] and MINOS [2100] confirmed the atmospheric parameters with a controlled beam over 250 km and 735 km; T2K found $\nu_\mu \to \nu_e$ appearance [2168] and now constrains δ_{CP} [2345]; NOvA measures the same parameters over 810 km [2398]; and OPERA observed ν_τ appearance directly [2247], closing the three-flavour picture by seeing the flavour the others infer.

103.6.5 Measured parameters and what is still unknown

Table 103.3: The six oscillation parameters, from the global fit to all data [2356, 2427]. The splittings are quoted, as universally, as $\Delta m^2 c^4$ in eV^2 — an energy squared, per Remark 103.1. The sign of Δm^2_{32} is not measured; the value listed assumes the normal ordering, and the inverted ordering gives $(-2.529 \pm 0.029) \times 10^{-3}\,\mathrm{eV}^2$. The phase δ_{CP} is constrained, not measured.

Parameter	Value	Principal source
$\Delta m^2_{21} c^4$	$(7.53 \pm 0.18) \times 10^{-5}\,\mathrm{eV}^2$	KamLAND reactor spectrum; solar
$\lvert\Delta m^2_{32}\rvert\, c^4$	$(2.455 \pm 0.028) \times 10^{-3}\,\mathrm{eV}^2$	atmospheric; accelerator disappearance
$\sin^2\theta_{12}$	0.307 ± 0.013	solar survival probability; KamLAND
$\sin^2\theta_{23}$	0.546 ± 0.021	atmospheric and accelerator ν_μ disappearance
$\sin^2\theta_{13}$	0.0220 ± 0.0007	short-baseline reactor near/far ratio
δ_{CP}	$\approx 1.2\pi$, poorly bounded	$\nu_\mu \to \nu_e$ versus $\bar\nu_\mu \to \bar\nu_e$

Three questions remain open within the three-flavour picture itself, and it is worth being exact about what each is.

The mass ordering. The sign of Δm^2_{32} is unknown. *Normal ordering* means $m_1 < m_2 < m_3$, with the close pair at the bottom; *inverted* means $m_3 < m_1 < m_2$, with the close pair at the top. Global fits mildly prefer the normal ordering, at a level between two and three standard deviations that depends on which datasets are combined and is not a measurement. The routes to settling it are those of Remark 103.60.

The octant of θ_{23}. Oscillation of muon neutrinos to tau neutrinos depends on $\sin^2 2\theta_{23}$, which is symmetric under $\theta_{23} \to \pi/2 - \theta_{23}$. Since $\sin^2 2\theta_{23} = 0.99$, the angle is close to $\pi/4$ and the two solutions are close together; distinguishing them requires the appearance channel, where $|U_{\mu 3}|^2 = \sin^2\theta_{23}$ enters unsquared.

The phase δ_{CP}. T2K's data disfavour CP conservation at about the 95 % level [2345] and prefer a value near $-\pi/2$; NOvA's are compatible with CP conservation [2398]. Combining them gives a preference and not a measurement, and this treatise records it as such. A non-zero δ_{CP} would be leptonic CP violation — the leptonic counterpart of Section 103.5 — and would not, by itself, establish anything about the baryon asymmetry of Phenomenon 103.44,

a point taken up in Section 103.7.4.

103.7 Neutrino mass

Oscillation measures squared-mass *differences* and nothing else. Three quite different experiments bound three different combinations of the three masses, and none of them measures a mass eigenvalue. Keeping the three separate is the point of this section, because they are routinely and wrongly quoted as though they were the same number.

Definition 103.61 (Three mass observables)**.**

$$m_\beta^2 := \sum_i |U_{ei}|^2 \, m_i^2 \qquad \text{(beta-decay kinematics)}\,, \tag{103.64}$$

$$m_{\beta\beta} := \left| \sum_i U_{ei}^2 \, m_i \right| \qquad \text{(neutrinoless double beta decay)}\,, \tag{103.65}$$

$$\Sigma := \sum_i m_i \qquad \text{(cosmology)}\,. \tag{103.66}$$

The first is an incoherent sum of squares and is always real and positive; the second involves U_{ei}^2 rather than $|U_{ei}|^2$ and therefore depends on the Majorana phases of Equation (103.48), so cancellations can drive it to zero; the third is a plain sum.

103.7.1 Direct kinematic limits

The only model-independent route to the mass scale is the shape of a beta spectrum near its endpoint, because there the neutrino is non-relativistic and its rest energy competes with its kinetic energy. The method is that of Section 101.2.3: a non-zero mass removes the last $m_\nu c^2$ of electron energy and rounds the approach to the endpoint, changing the Kurie plot from a straight line to a curve that meets the axis vertically.

Phenomenon 103.62 (The electron-neutrino mass is below one electronvolt)**.** Measuring the electron spectrum of molecular tritium decay, $^3\mathrm{H} \to {}^3\mathrm{He}^+ + e^- + \bar{\nu}_e$, with endpoint $Q = 18.574\,\mathrm{keV}$, the KATRIN experiment finds

$$m_\beta < 0.8\,\mathrm{eV}/c^2 \qquad \text{at } 90\,\% \text{ confidence} \tag{103.67}$$

[2399], with m_β the combination Equation (103.64). The bound assumes only energy and momentum conservation and the measured molecular final-state distribution.

Rests on Equations (101.11) *and* (103.64) *and Definition 101.19.*

Derivation. *Derives Phenomenon 103.62.*

By Equation (101.11), the electron spectrum of an allowed transition is

$$\frac{\mathrm{d}\Gamma}{\mathrm{d}E_e} \propto F(Z, E_e)\, p_e E_e \,(Q - T_e)\, \sqrt{(Q - T_e)^2 - m_\beta^2 c^4}\,,$$

T_e being the electron kinetic energy. For $m_\beta = 0$ this is $(Q - T_e)^2$ near the endpoint, so the Kurie function of Definition 101.19 is linear and reaches zero at $T_e = Q$; for $m_\beta \neq 0$ the spectrum terminates at $T_e = Q - m_\beta c^2$ and approaches it with infinite slope. Two consequences fix the design.

First, the sensitivity is to m_β^2, not to m_β: the leading distortion of the rate integrated over the last interval Δ below the endpoint is a relative shift of order $m_\beta^2 c^4/\Delta^2$. Halving the mass sensitivity therefore costs a factor of four in statistics, which is why progress in this field is slow and why an experiment reports a limit on m_β^2 that can and does fluctuate negative.

Second, the region carrying the information is minute. Since $\mathrm{d}\Gamma/\mathrm{d}T_e \propto (Q - T_e)^2$, the fraction of all decays occurring within Δ of the endpoint is

$$\frac{\int_{Q-\Delta}^{Q} (Q-T)^2 \,\mathrm{d}T}{\int_{0}^{Q} (Q-T)^2 \,\mathrm{d}T} = \left(\frac{\Delta}{Q}\right)^3 , \tag{103.68}$$

which for $\Delta = 1\,\mathrm{eV}$ and $Q = 18.574\,\mathrm{keV}$ is 1.6×10^{-13}. An experiment sensitive to $0.2\,\mathrm{eV}/c^2$ must therefore combine a source of enormous activity with a spectrometer of electronvolt resolution at $18.6\,\mathrm{keV}$ — a relative resolution of 5×10^{-5} — and reject every background event in the same window. That is what KATRIN's 10 m-diameter magnetic adiabatic collimation spectrometer exists to do.

Tritium is chosen because Q is the smallest of any convenient superallowed emitter, maximizing Equation (103.68), and because its nuclear matrix element is simple. The residual model dependence is molecular, not nuclear: the daughter $^3\mathrm{HeT}^+$ ion is left in a distribution of rotational, vibrational and electronic excitations spread over some 1 eV, and that distribution is calculated, not measured. It is the leading systematic of Equation (103.67). □

Remark 103.63 (Why a weaker bound is worth more). Equation (103.67) is an order of magnitude weaker than the cosmological bound of Section 103.7.3 and is nevertheless the more valuable number, for the reason that runs through this whole treatise: it is model-independent. It assumes no cosmological model, no growth of structure, no assumption about whether the neutrino is its own antiparticle, and no nuclear matrix element. Its interpretation requires only Equation (103.64) and the measured mixing angles. When a model-dependent bound and a model-independent one disagree, it is the model that is in question, and an experiment that can say so is worth building even at lower sensitivity.

103.7.2 Neutrinoless double beta decay

Double beta decay is the slowest process ever measured, and its existence is a consequence of nuclear structure alone: the pairing term of Equation (107.42) splits the even-A mass parabola, so an even–even nuclide can lie below both its odd–odd neighbours while lying above the even–even nuclide two steps away. Single beta decay is then closed and double beta decay is open. That argument, the isotopes it selects and the Q^{11} phase space are Phenomenon 107.72. What belongs here is the lepton-number-violating variant and what its non-observation means.

Phenomenon 103.64 (Neutrinoless double beta decay has not been observed)**.** Two-neutrino double beta decay is observed, with half-lives between 10^{18} and 10^{24} years [2427] — the range quoted in Phenomenon 107.72, which is the isotope-by-isotope statement. The neutrinoless mode, which would change lepton number by two units and is possible only if the neutrino is its own antiparticle, is not observed: the best searches set half-life limits

$$T_{1/2}^{0\nu}\left(^{76}\mathrm{Ge}\right) > 1.8 \times 10^{26}\,\mathrm{yr}\,, \qquad T_{1/2}^{0\nu}\left(^{136}\mathrm{Xe}\right) > 2.3 \times 10^{26}\,\mathrm{yr}\,, \tag{103.69}$$

at 90 % confidence, from GERDA [2350] and KamLAND-Zen [2414] respectively. This is stated here as what it is — a search that has found nothing — and it leaves the Dirac or Majorana character of the neutrino undetermined.

Rests on Phenomenon 107.72, Proposition 101.32, and Equation (103.65).

Derivation. *Derives Phenomenon 103.64.*

Furry computed the transition $(A, Z) \to (A, Z+2) + 2e^-$ in 1939 [836]: the antineutrino emitted at one nucleon vertex is absorbed as a neutrino at the other. Two conditions must hold for the amplitude not to vanish, and each supplies a factor.

The neutrino must be its own antiparticle, since the particle emitted at the first vertex is what must be absorbed at the second. This is the Majorana condition, and it is what makes the process a test of Section 103.7.4 rather than of anything else.

The neutrino must have mass. The charged current emits a right-helicity antineutrino and absorbs a left-helicity neutrino (Proposition 101.32), so the two vertices demand opposite helicities of the same particle. The amplitude therefore requires the "wrong-helicity" component of the propagating state. Equation (101.32) gives the chirality *weight* of that component as $\frac{1}{2}(1-\beta) \simeq m^2c^4/4E^2$, which is a probability, so the amplitude carries its square root, $m_i c^2/2E$. The internal neutrino carries a virtual momentum of order the inverse internucleon spacing, $\hbar c/r \approx 100\,\mathrm{MeV}$, so the suppression factor is of order $m_i c^2/100\,\mathrm{MeV}$, one factor per exchanged mass eigenstate; the factor of two, and the shape of the neutrino propagator between the two nucleons, are conventionally absorbed into the nuclear matrix element defined below rather than carried separately. Summing coherently over the three mass states with the mixing factors U_{ei} from each vertex, the amplitude is proportional to $\sum_i U_{ei}^2 m_i$ — *squared*, not modulus-squared, because both vertices are of the same type — which is $m_{\beta\beta}$ of Equation (103.65). Hence

$$\frac{1}{T_{1/2}^{0\nu}} = G^{0\nu}\,(Q, Z)\,\left|M^{0\nu}\right|^2\,\left(\frac{m_{\beta\beta}}{m_e}\right)^2 , \tag{103.70}$$

with $G^{0\nu}$ a phase-space factor, carrying 1/yr, calculable exactly from the Q value and the Coulomb field, and $M^{0\nu}$ the dimensionless nuclear matrix element.

Two structural remarks. First, Equation (103.70) is a rate proportional to a squared mass, so a half-life limit improved by a factor 100 improves the mass limit only by a factor 10 — the same quadratic wall that, in the derivation of Phenomenon 103.62, makes every halving of the kinematic mass sensitivity cost a factor of four in statistics. Second, the electron sum-energy spectrum of the neutrinoless mode is a line at Q, sitting on the continuum of the two-neutrino mode, and the two-neutrino mode is therefore an irreducible background whose tail is controlled only by energy resolution. This is why the two leading techniques are a high-purity germanium diode, whose resolution at $Q = 2039\,\mathrm{keV}$ is about 3 keV, and a very large mass of xenon-loaded scintillator, which trades resolution for target mass. □

Remark 103.65 (The matrix element is the limiting uncertainty). Converting Equation (103.69) into a bound on $m_{\beta\beta}$ requires $M^{0\nu}$, and the many-body methods available — the interacting shell model, the quasiparticle random-phase approximation, the interacting boson model, energy density functionals — disagree with one another by a factor of two to three for every isotope of interest, as Phenomenon 107.72 records. Since the rate goes as the square, that is a factor of four to nine in half-life and a factor of two to three in the mass limit. The published bounds are therefore quoted as *ranges*:

$$m_{\beta\beta} < 79 \text{ to } 180\,\mathrm{meV}/c^2 \quad \left({}^{76}\mathrm{Ge}\right) , \qquad m_{\beta\beta} < 36 \text{ to } 156\,\mathrm{meV}/c^2 \quad \left({}^{136}\mathrm{Xe}\right) , \tag{103.71}$$

the spread being theory, not statistics [2350, 2414]. A treatise of evidence must quote the range and not its most favourable end.

Proposition 103.66 (What the searches are approaching)**.** *With the mixing parameters of Table 103.3, $m_{\beta\beta}$ of Equation (103.65) is bounded, over the unknown Majorana phases and the unknown lightest mass, by*

$$m_{\beta\beta}\big|_{\mathrm{IO}} \geq 19\,\mathrm{meV}/c^2 , \qquad m_{\beta\beta}\big|_{\mathrm{NO}} \geq 0 , \tag{103.72}$$

so the inverted ordering predicts a floor that the next generation of experiments can reach, and the normal ordering does not.

Rests on Equations (103.48) and (103.65).

Proof. *Derives Proposition 103.66.*

Write $m_{\beta\beta} = \left| c_{12}^2 c_{13}^2 m_1 + s_{12}^2 c_{13}^2 m_2 \mathrm{e}^{\mathrm{i}\alpha_{21}} + s_{13}^2 m_3 \mathrm{e}^{\mathrm{i}\alpha_{31}} \right|$. In the inverted ordering $m_3 \approx 0$ and $m_1 c^2 \approx m_2 c^2 \approx \sqrt{|\Delta m_{32}^2|\, c^4} = 50\,\mathrm{meV}$, so the third term is negligible and the first two, being nearly equal in mass, can cancel only to the extent that their coefficients differ: the minimum over α_{21} is

$$c_{13}^2 \left(c_{12}^2 - s_{12}^2 \right) \sqrt{|\Delta m_{32}^2|\, c^4}/c^2 = 0.978 \times 0.386 \times 50\,\mathrm{meV}/c^2 = 19\,\mathrm{meV}/c^2 \,,$$

using $\cos 2\theta_{12} = 0.386$ from Table 103.3. The maximum, at $\alpha_{21} = 0$, is $49\,\mathrm{meV}/c^2$. In the normal ordering with $m_1 \to 0$ the three terms are 0, $s_{12}^2 c_{13}^2 \sqrt{\Delta m_{21}^2 c^4}/c^2 = 2.6\,\mathrm{meV}/c^2$ and $s_{13}^2 \sqrt{\Delta m_{31}^2 c^4}/c^2 = 1.1\,\mathrm{meV}/c^2$, which are comparable and can cancel exactly for a suitable phase. There is therefore no floor.

The consequence for the programme is sharp and should be stated without optimism. Equation (103.71) is at the level of the inverted-ordering band, and an experiment covering that band completely would, if it saw nothing, exclude the inverted ordering *provided* the neutrino is a Majorana particle. It would not exclude the Majorana hypothesis, because the normal ordering permits $m_{\beta\beta} = 0$ exactly. The asymmetry is inherent: the experiment can discover, and can exclude one ordering, but can never refute the Majorana nature. □

103.7.3 Cosmological bounds

Neutrinos decouple from the primordial plasma while relativistic, and survive as a relic background of number density fixed by the temperature. Once non-relativistic they contribute to the matter density like any other matter; but unlike cold dark matter they retain a large velocity dispersion, and therefore stream freely out of gravitational potential wells smaller than a characteristic scale, suppressing the growth of structure below it.

Proposition 103.67 (Relic density and free streaming)**.** *Three species of light neutrino of total mass Σ contribute*

$$\Omega_\nu h^2 = \frac{\Sigma}{93.14\,\mathrm{eV}/c^2} \,, \tag{103.73}$$

and suppress the matter power spectrum on scales below the free-streaming length by a relative amount $\Delta P/P \approx -8\Omega_\nu/\Omega_m$. For $\Sigma = 0.12\,\mathrm{eV}/c^2$, $\Omega_\nu h^2 = 1.3 \times 10^{-3}$ and the suppression is at the per-cent level — measurable, and the basis of the bound.

Rests on Equation (103.66).

Proof. *Derives Proposition 103.67.*

The relic number density per species follows from decoupling in thermal equilibrium, the general treatment of which belongs to Section 48.3: a fermion species in equilibrium at temperature T_ν has $n_\nu = \frac{3}{4} \times \frac{2\zeta(3)}{\pi^2} \left(k_B T_\nu/\hbar c \right)^3$ per degree of freedom, and $T_\nu = (4/11)^{1/3}\, T_\gamma$ because the electron–positron annihilation that follows decoupling heats the photons and not the neutrinos. With $T_{\gamma,0} = 2.7255\,\mathrm{K}$ this gives $n_\nu = 1.12 \times 10^8/\mathrm{m}^3$ per species. Dividing the mass density $\Sigma\, n_\nu$ by the critical density $\rho_c = 1.878\,34 \times 10^{-26}\,\mathrm{kg/m^3}\, h^2$ gives Equation (103.73). The naive arithmetic, $\rho_c c^2/n_\nu$ per unit h^2, comes to 94.1 eV; the standard constant 93.14 eV includes the one-per-cent correction from the fact that decoupling is not instantaneous, which leaves the neutrinos slightly hotter than $(4/11)^{1/3}\, T_\gamma$.

For the free streaming: a relic neutrino of mass m has, at redshift z, a typical momentum $p \approx 3.15 k_B T_\nu (1+z)/c$ and hence a speed $v/c \approx 3.15 k_B T_\nu (1+z)/mc^2$. Today $T_\nu = (4/11)^{1/3} \times 2.7255\,\mathrm{K} = 1.9454\,\mathrm{K}$, so $k_B T_\nu = 1.676 \times 10^{-4}\,\mathrm{eV}$ and $3.15 k_B T_\nu = 5.28 \times 10^{-4}\,\mathrm{eV}$; for $mc^2 = 0.05\,\mathrm{eV}$ this gives $v/c = 1.06 \times 10^{-2}$, that is $v = 3.2 \times 10^3\,\mathrm{km/s}$. Over a Hubble time it therefore travels a comoving distance of order v/H, and perturbations of smaller wavelength are erased because the neutrinos simply leave them. The suppression coefficient

8 follows from solving the equation for the growth of density perturbations, which belongs to Section 48.5.1, with a component that does not cluster; the essential point for this chapter is that the effect is proportional to Ω_ν, hence to Σ, and is therefore a measurement of the sum of the masses and of no other combination. □

Phenomenon 103.68 (The sum of the neutrino masses is bounded)**.** Combining the temperature and polarization anisotropies of the cosmic microwave background with the baryon acoustic oscillation scale measured in galaxy surveys,

$$\Sigma = \sum_i m_i < 0.12\,\mathrm{eV}/c^2 \qquad \text{at } 95\,\% \text{ confidence} \tag{103.74}$$

[2348], within the six-parameter ΛCDM model whose parameters belong to Section 48.7.
Rests on Proposition 103.67 and Phenomenon 48.15.

Derivation. *Derives Phenomenon 103.68.*

The microwave background alone constrains Σ only weakly, because a neutrino of mass below about $0.6\,\mathrm{eV}/c^2$ is still relativistic at recombination and its effect on the acoustic peaks is degenerate with other parameters; the constraint comes from combining it with a low-redshift measurement of the expansion history and of the clustering amplitude, which is what the baryon acoustic scale of Phenomenon 48.15 supplies. Extracting Equation (103.74) from those data is a global fit, not a closed-form derivation, and it is quoted here for comparison with Equation (103.67) rather than reproduced.

The comparison is the point. Equation (103.74) is a factor 20 stronger than the kinematic bound — and rests on the assumptions that the dark energy is a cosmological constant, that the primordial spectrum is a power law, and that no other light relic contributes. Each has been examined and none is established independently. Relaxing them typically weakens the bound by a factor of two to three, which is the honest uncertainty to attach to it. □

Remark 103.69 (The bound is close to the inverted-ordering minimum)*.* The minimum values of Σ permitted by the measured splittings are, taking the lightest mass to zero,

$$\Sigma_{\mathrm{NO}} \geq 0.059\,\mathrm{eV}/c^2\,, \qquad \Sigma_{\mathrm{IO}} \geq 0.100\,\mathrm{eV}/c^2\,, \tag{103.75}$$

computed from $m_2c^2 = \sqrt{\Delta m^2_{21} c^4}$ and $m_3c^2 = \sqrt{\Delta m^2_{31} c^4}$ in the normal case, and from $m_1c^2 = \sqrt{|\Delta m^2_{31}|\,c^4}$, $m_2c^2 = \sqrt{|\Delta m^2_{32}|\,c^4}$ in the inverted case. The inverted minimum sits just below Equation (103.74), so cosmology mildly disfavours the inverted ordering — and would exclude it if the bound tightened by a factor of two without a detection. That is a striking situation and it deserves a warning: a *cosmological* bound excluding a *laboratory* mass ordering would be an inference of exactly the model-dependent kind Remark 103.63 cautions against. The ordering should be settled by Remark 103.60, in an experiment, and this treatise will record it as settled only then.

103.7.4 Dirac or Majorana: an open question

Every charged fermion has a Dirac mass, because a Majorana mass term would violate electric charge conservation. The neutrino is the only known fermion for which both options are open, and no experiment has distinguished them.

Definition 103.70 (The two mass terms)**.** For a left-chiral field ν_L and, if it exists, a right-chiral partner ν_R, the Lorentz-invariant mass terms are

$$\mathcal{L}_D = -m_D c^2 \left(\bar{\nu}_L \nu_R + \bar{\nu}_R \nu_L\right)\,, \tag{103.76}$$

$$\mathcal{L}_M = -\tfrac{1}{2} m_M c^2 \left(\overline{\nu^c_L}\,\nu_L + \bar{\nu}_L\,\nu^c_L\right)\,, \tag{103.77}$$

with ν^c the charge-conjugate field of Equation (105.15). The first conserves lepton number and requires a new field; the second violates lepton number by two units and requires none.

The full analysis of these two terms, the degrees-of-freedom count that makes a Majorana field equivalent to one massive Weyl field, and Majorana's original symmetric theory [796] belong to Section 96.6.1 and Section 96.7.1, which reserve them. What belongs here is the experimental status, and it can be stated in four sentences.

Remark 103.71 (The evidential situation, stated plainly). *No experiment distinguishes the two.* Oscillation cannot, by Proposition 103.47; kinematic beta-decay measurements cannot, since Equation (103.64) is phase-independent; cosmology cannot, since Equation (103.66) is too. The only practical discriminator is neutrinoless double beta decay, and by Equation (103.69) it has found nothing.

A positive result would settle it whatever the mechanism. If the decay is observed, a Majorana mass term for the neutrino follows, even if the operator that drives the decay is something else entirely: the observed transition can always be closed into a $\bar{\nu} \to \nu$ self-energy insertion, generating Equation (103.77) at some order. This is the Schechter–Valle argument [1701]. It is a statement about the *existence* of the term, not about its size, which the argument leaves many orders of magnitude below the observed masses — a limitation worth stating, since the theorem is often quoted as though it were quantitative.

The most economical hypothesis has no evidence. If a heavy right-chiral singlet N exists with a Majorana mass M far above the electroweak scale, the light eigenvalue of the combined mass matrix is $m_\nu \approx m_D^2/M$ — small because M is large. This is the seesaw [1605]; with m_D of the order of the top mass, $Mc^2 \sim 1 \times 10^{15}\,\mathrm{GeV}$. In the effective theory below M it is one realization of the unique dimension-five operator that can generate a neutrino mass from Standard Model fields [1646], whose coefficient carries $1/\mathrm{J}$ and is therefore suppressed by one power of a new scale. That operator is a *parametrization* of ignorance and the seesaw is one *model* for its coefficient. There is at present no experimental evidence for either, and this treatise names them as hypotheses and nothing more.

Sterile neutrinos: null. A fourth neutrino not coupling to the Z evades Equation (103.15) and would appear as a fourth oscillation frequency. The LSND excess [2007] and the MiniBooNE low-energy excess have been read that way; the MicroBooNE liquid-argon measurement finds no corresponding excess in the electron-neutrino channel [2397], and the reactor and gallium rate anomalies have been partly resolved by recalculating fluxes. No sterile state is established and the anomalies have not all gone away; Section 114.6.4 is the section reserved for the state of each.

Remark 103.72 (Leptogenesis is named here and is not evidence). The heavy singlets of the seesaw, decaying out of equilibrium with a CP-violating asymmetry, would generate a lepton asymmetry that electroweak processes partly convert into the baryon asymmetry of Equation (103.45). The mechanism is attractive precisely because it needs no new ingredient beyond the one already invoked to explain the neutrino masses. It is nevertheless untested and, with the relevant scale above $1 \times 10^9\,\mathrm{GeV}$, untestable by any foreseeable experiment; and a measurement of δ_{CP} would not establish it, because the phases entering the heavy-neutrino decays are not the phases entering Equation (103.47). It is listed in Section 133.5 as an unexplained observation, not as a solved problem.

103.8 Open questions in flavour

The Standard Model has nineteen free parameters, and thirteen of them — nine charged-fermion masses, three CKM angles and one phase — belong to flavour; the neutrino sector adds at least seven more. Every one is measured and none is predicted. This closing section names the specific questions that are open, distinguishing those where a measurement disagrees with a prediction from those where there is no prediction to disagree with.

The mass hierarchy. The charged-fermion masses span from $0.511\,\mathrm{MeV}/c^2$ to $172.57\,\mathrm{GeV}/c^2$, a factor 3.4×10^5 within one set of gauge quantum numbers, and the

neutrino masses lie a further six orders of magnitude below the electron. Expressed as Yukawa couplings by Remark 103.21 they run from 1 down to 10^{-6} for the charged leptons and to 10^{-12} or less for neutrinos. There is no prediction here to disagree with; there is a pattern with no explanation, and it is Section 133.1.

Charged-lepton flavour violation: not observed. In the Standard Model with massive neutrinos, $\mu \to e\gamma$ proceeds through a loop with the GIM suppression of Theorem 103.11, and because the internal masses are neutrino masses the suppression factor $\left(\Delta m^2 c^4\right)^2 / \left(M_W c^2\right)^4$ — with $\Delta m^2 c^4 = 2.5 \times 10^{-3}\,\mathrm{eV}^2$ and $M_W c^2 = 80.4\,\mathrm{GeV}$, equal to 1.5×10^{-49} — makes it utterly unobservable. Any observation would therefore be new physics, which is why the search is worth pursuing. The MEG II experiment finds

$$\mathcal{B}\left(\mu^+ \to e^+\gamma\right) < 7.5 \times 10^{-13} \tag{103.78}$$

from its first dataset, and 1.5×10^{-13} combining with the earlier MEG result, both at 90 % confidence [2424]. Phenomenon 101.53 records the same null result in its charged-current context and quotes the same two values. Nothing is seen.

The B-decay anomalies: mostly resolved, and the resolution is instructive. For a decade the ratios $R_K = \mathcal{B}\left(B \to K\mu^+\mu^-\right) / \mathcal{B}\left(B \to Ke^+e^-\right)$ and its K^* analogue were measured below the Standard Model value of unity, at the level of 2.5σ each. Lepton universality is a sharp prediction: the ratio is one up to phase-space and radiative corrections of a per cent, because the gauge couplings do not distinguish the flavours. In 2022 LHCb identified a mis-estimated electron-identification efficiency, and the corrected measurements are

$$R_K = 0.949 \pm 0.047\,, \qquad R_{K^*} = 1.027 \pm 0.077\,, \tag{103.79}$$

consistent with unity [2412]. This treatise records the episode rather than quietly dropping it, because it is a clean example of an anomaly that was a systematic effect — and because the related $b \to s\mu\mu$ branching fractions and angular observables remain about 3σ below their Standard Model predictions, with hadronic uncertainties that are disputed. Universality holds; the absolute rates are not settled.

The muon anomalous magnetic moment. The muon's anomaly is sensitive to any flavour-dependent physics at high scale, and its Standard Model prediction is limited by the hadronic vacuum polarization, which belongs to Section 112.6.3. The status — including the disagreement between the data-driven and lattice evaluations of that contribution, which is larger than the effect being tested — belongs to Section 112.6.4. It is quoted here only to record that the one place where a flavour-sensitive precision observable may disagree with theory is a place where the theory prediction is itself contested.

Three generations, and no reason. The number is measured twice, by Equation (103.15) and by nucleosynthesis, and predicted by nothing.

First-row unitarity. Proposition 103.8: the most precisely tested relation in quark flavour physics is currently low by two to three standard deviations. It is the one entry in this list that is a genuine discrepancy between a prediction and a measurement rather than an absence of prediction, and it is most likely a systematic effect in one of the inputs.

Every item above is listed as open in Chapter 133 rather than resolved by a model this treatise would have no evidence to include. The models exist — horizontal symmetries relating the generations, additional gauge bosons distinguishing them, composite fermions, extra spatial dimensions in which the Yukawa hierarchy becomes a geometric one — and every one of them predicts something that has not been seen. Under editorial rule 1 they are named here, once, so that a reader knows they were considered, and are not developed.

Chapter 104

Electroweak Unification and the Higgs Boson

The weak and electromagnetic interactions are two faces of one gauge theory. This chapter builds it: the gauge group $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ of Glashow [1206], its spontaneous breaking to the $\mathrm{U}(1)_{\mathrm{em}}$ of Chapter 60 by a scalar doublet, and the resulting massive $W^{\pm}$ and Z^0 carriers that replace the contact interaction of Fermi [738] used in Chapter 101. The chapter therefore sits after Chapter 64, where the gauge-field formalism belongs, and after the renormalization of Chapter 100, and before the strong sector of Chapter 102 with which it completes the Standard Model. Two threads run through it. The first is theoretical: how a gauge boson acquires mass without destroying the gauge invariance that makes the theory renormalizable — the mechanism of Anderson [1245], Englert and Brout [1275], Higgs [1282, 1283] and Guralnik, Hagen and Kibble [1279], whose renormalizability 't Hooft and Veltman established [1448, 1463].

The second thread is evidential, and it is the reason the theory is in this book at all. Weak neutral currents were seen in Gargamelle [1502, 1503]; the W and Z were produced and weighed at the CERN $p\bar{p}$ collider [1710, 1711, 1713, 1714]; four LEP experiments and SLD measured the Z resonance to parts in 10^4 [2104]; and the scalar itself was found in 2012 — the discovery account belongs to Chapter 113 — and has since been mapped coupling by coupling [2394, 2410]. Where the picture is incomplete — the metastability of the vacuum, the W-mass tension, the origin of the Yukawa hierarchy — the chapter says so. The standard reference treatment is [1933].

104.1 Conventions carried into this chapter

Everything below is written in SI units, in the conventions fixed once in Section 100.1.1: the metric is $\eta_{\mu\nu} = \mathrm{diag}(+1,-1,-1,-1)$, a four-momentum is $p^\mu = (E/c, \boldsymbol{p})$ of dimension $\mathrm{kg\,m/s}$, the mass shell is $p^2 = m^2c^2$ (Equation (100.1)), Mandelstam variables are squared *momenta* so that $\sqrt{s}\,c$ is a centre-of-mass energy, and every Lagrangian density is an energy density, $\mathrm{J/m^3}$, with the action $S = c^{-1}\int \mathrm{d}^4x\,\mathcal{L}$ of Equation (100.2). The gauge fields and their couplings are those of Definition 101.66: W^a_μ and B_μ carry $\mathrm{V\,s/m}$ like the electromagnetic potential of Definition 100.2, and g and g' carry the dimension of electric charge, C.

Three abbreviations recur so often that they are given names here.

Notation 104.1 (Dimensionless couplings and the electroweak scale). For any coupling q of the dimension of charge, write

$$\hat{q} := \frac{q}{\sqrt{\varepsilon_0 \hbar c}}\,, \tag{104.1}$$

a pure number, since q^2/ε_0 and $\hbar c$ both carry J m. In particular $\hat{e}^2 = e^2/(\varepsilon_0 \hbar c) = 4\pi\alpha$ by Equation (100.3), so $\hat{e} = \sqrt{4\pi\alpha} = 0.302\,822$ at zero momentum transfer, and $\hat{g}$, $\hat{g}'$ are the numbers that the references cited below write simply as g and g' (see Remark 104.2).

The scalar doublet Φ carries the dimension of a scalar field, $\mathrm{J}^{1/2}/\mathrm{m}^{1/2}$ (Equation (106.33)), so its vacuum expectation value v does too. The combination

$$E_v := v\sqrt{\hbar c} \tag{104.2}$$

is an energy, and it is what the literature means when it writes "$v = 246\,\mathrm{GeV}$". Its value is computed in Proposition 104.26.

Remark 104.2 (Reading this chapter against the natural-units literature). Every source cited below sets $\hbar = c = 1$. The dictionary is the one given once in Remark 100.4, plus three entries special to this chapter: a coupling of charge dimension maps to its hat by Equation (104.1), a vacuum expectation value maps to E_v by Equation (104.2), and a scalar self-coupling λ — which carries $1/(\mathrm{J\,m})$ here, by Equation (106.34) — maps to $\hat{\lambda} = \lambda\hbar c$. With those three substitutions every formula below becomes the one in the references, and conversely. No derivation in this chapter is performed with $\hbar$ or c set to one; this remark exists so that the reader can check the chapter against its sources.

104.2 From Fermi's theory to a gauge theory

104.2.1 The four-fermion interaction and its breakdown

Fermi's theory of beta decay [738] is a product of two currents evaluated at one point, Equation (101.6), with a single constant G_F in front. Chapter 101 develops it in full and measures it: the definitive determination comes from the positive-muon lifetime, which the MuLan experiment fixed to about one part per million [2225], giving

$$\frac{G_F}{(\hbar c)^3} = (1.166\,378\,7 \pm 0.000\,000\,6) \times 10^{-5}/\mathrm{GeV}^2\,, \qquad G_F = 1.435\,85 \times 10^{-62}\,\mathrm{J\,m}^3 \tag{104.3}$$

by Equation (101.8). The Lorentz structure of the currents is $V - A$, forced on the theory by the observation of maximal parity violation and written down by Feynman and Gell-Mann [1144] and by Sudarshan and Marshak [1162]; Section 101.4 derives it from the data and Chapter 110 is the apparatus.

Three facts about that theory are the reason this chapter exists, and all three are established in Chapter 101. They are collected here because everything the electroweak theory does is a response to them.

Proposition 104.3 (Three failures of the contact theory). *Let G_F be as in Equation (104.3) and let E denote a centre-of-mass energy. Then*

1. *the dimensionless number governing every weak amplitude is $G_F E^2/(\hbar c)^3$, which reaches unity at $E_F = 292.8\,\mathrm{GeV}$ (Equation (101.9)), so the perturbation expansion has an expiry date;*
2. *the $J = 0$ cross-section $\sigma = G_F^2 E^2/\pi(\hbar c)^4$ of Equation (101.60) breaches the partial-wave unitarity bound at $E = 1.04\,\mathrm{TeV}$ (Theorem 101.78);*
3. *the superficial degree of divergence $D = 4 - \frac{3}{2}E_{ext} + 2V$ of Equation (101.91) grows with the order V of perturbation theory, so the theory needs infinitely many counterterms and predicts nothing beyond the order at which it is fitted.*

All three are consequences of one fact: G_F carries $\mathrm{J\,m}^3$ (Proposition 101.7), a dimension no rescaling can remove.

Rests on Equation (104.3), Proposition 101.7, and Theorem 101.78.

Derivation. *Derives Proposition 104.3.*

Each item is proved where it is stated in Chapter 101; what is added here is that the three are the same statement counted three ways. A coupling of dimension $\mathrm{J\,m^3}$ can only enter an amplitude in the combination $G_F E^2/(\hbar c)^3$, because that is the unique dimensionless product of G_F, $\hbar$, c and an energy; hence item (i). A cross-section is an area and the only area available is $G_F^2 E^2/(\hbar c)^4$ times a pure number, which grows while the unitarity bound $16\pi(\hbar c)^2/E^2$ falls, so they must cross; hence item (ii). And in the momentum counting of Theorem 100.17 each vertex contributes its coupling's negative mass dimension to D with a positive sign, which is Corollary 101.82; hence item (iii). □

The obvious repair is Yukawa's: give the interaction a carrier [765]. Two currents joined by a heavy vector boson of mass M_W reproduce the contact term at momentum transfers small compared with $M_W c$ (Proposition 101.65), with the matching condition Equation (101.78),

$$\frac{G_F}{\sqrt{2}\,(\hbar c)^3} = \frac{\pi \alpha_W}{2\,(M_W c^2)^2} = \frac{\hat{g}^2}{8\,(M_W c^2)^2}\,, \tag{104.4}$$

the last form using $\hat{g}^2 = 4\pi\alpha_W$ from Equation (101.71) and Equation (104.1). This trades one dimensionful constant for a dimensionless coupling and a mass, which is progress only if the mass is explained. It is not, and the cure is worse than it looks.

Theorem 104.4 (Inserting a vector boson by hand does not help)**.** *Let W_μ be a vector field with the Proca mass term*

$$\mathcal{L}_M = \frac{1}{\mu_0}\left(\frac{M_W c}{\hbar}\right)^2 W^+_\mu W^{-\mu} \tag{104.5}$$

added by hand to a gauge-invariant kinetic term. Then

1. *$\mathcal{L}_M$ is not invariant under $W_\mu \mapsto W_\mu + \partial_\mu \Lambda$;*
2. *its propagator Equation (101.77) tends to a constant rather than to zero at large momentum, so the power counting of Theorem 100.17 fails and the theory is non-renormalizable;*
3. *the longitudinal polarization vector grows linearly with momentum,*

$$\varepsilon^\mu_{(L)}(k) = \frac{k^\mu}{M_W c} + \mathcal{O}\!\left(\frac{M_W c}{|\boldsymbol{k}|}\right)\,, \tag{104.6}$$

so amplitudes with external longitudinal Ws grow with energy and the unitarity problem of Proposition 104.3(ii) reappears at a higher scale instead of being cured.

Rests on Equation (101.77), Theorem 100.17, and Proposition 104.3.

Proof. *Derives Theorem 104.4.*

(i) Under the shift, $W^+_\mu W^{-\mu}$ acquires $\partial_\mu\Lambda^+ W^{-\mu} + W^+_\mu \partial^\mu \Lambda^- + \partial_\mu\Lambda^+\partial^\mu\Lambda^-$, which no choice of Λ makes vanish identically. The mass term breaks the very invariance that Theorem 100.5 used to construct the interaction.

(ii) The massive vector kernel is $-\mu_0^{-1}\hbar^{-2}\left[(k^2 - M_W^2 c^2)\eta_{\mu\nu} - k_\mu k_\nu\right]$, whose inverse is Equation (101.77). At large k the transverse part falls as k^{-2}, exactly as the photon propagator Equation (100.17) does, but the term $k_\mu k_\nu / M_W^2 c^2$ divided by $k^2 - M_W^2 c^2$ tends to η-independent constant behaviour: it does not fall at all. Repeating the count of Theorem 100.17 with internal boson lines contributing k^0 instead of k^{-2} raises D by two for every such line, and D again grows with the order.

(iii) A massive vector of four-momentum $k^\mu = (E/c, 0, 0, k)$ with $E^2/c^2 - k^2 = M_W^2 c^2$ has three polarization vectors satisfying $k_\mu \varepsilon^\mu = 0$ and $\varepsilon \cdot \varepsilon^* = -1$; the longitudinal one is

$$\varepsilon^\mu_{(L)} = \frac{1}{M_W c}\left(k, 0, 0, \frac{E}{c}\right) = \frac{k^\mu}{M_W c} - \frac{M_W c}{E/c + k}\,(1, 0, 0, -1)\,,$$

as one checks by expanding $E/c = \sqrt{k^2 + M_W^2 c^2}$. The second term is bounded; the first grows without limit, and it is the one that appears in Equation (104.6). Since an amplitude is linear in each external polarization, an amplitude with n longitudinal Ws carries n factors of $k/M_W c$ relative to the transverse one. The physical content is that the longitudinal mode is the piece a genuinely massless gauge field does not have; inserting a mass by hand supplies it without supplying a field for it to come from. □

Remark 104.5 (Where the theory is going). Theorem 104.4 is the pivot of the chapter and it is worth naming its resolution in advance. The longitudinal mode is not manufactured out of nothing: it is a scalar degree of freedom that existed before the symmetry broke, and the growth of Equation (104.6) in one description is the ordinary, non-growing scalar amplitude in another. That equivalence, made exact in Proposition 104.47, is what saves unitarity, and the requirement that it hold fixes the couplings of the scalar (Theorem 104.51) and hence everything the LHC measures in Section 104.7.2. Remark 101.69 states the same conclusion from the other side.

104.2.2 Yang–Mills fields and the mass problem

The gauge principle of Theorem 100.5 takes a global phase symmetry and makes it local, at the cost of one vector field. Yang and Mills [1054] asked what happens when the global group is non-abelian, and the answer is the framework in which the rest of this chapter is written. The general construction belongs to Chapter 64 and its geometry to Section 14.6; what is needed here is the explicit SU(2) case in SI units.

Definition 104.6 (Non-abelian gauge fields in SI). Let $T^a = \tfrac{1}{2}\sigma^a$ be the generators of SU(2), with $\left[T^a, T^b\right] = \mathrm{i}\varepsilon^{abc}T^c$ and $\mathrm{tr}(T^aT^b) = \tfrac{1}{2}\delta^{ab}$ (Section 14.3). Let ψ be a field in some representation with generators T^a, and W_μ^a three vector fields of dimension V s/m. The covariant derivative is

$$D_\mu = \partial_\mu + \frac{\mathrm{i}g}{\hbar} W_\mu^a T^a \,, \tag{104.7}$$

with g of dimension C, and the field strength is

$$W_{\mu\nu}^a = \partial_\mu W_\nu^a - \partial_\nu W_\mu^a - \frac{g}{\hbar}\varepsilon^{abc} W_\mu^b W_\nu^c \,, \qquad \left[W_{\mu\nu}^a\right] = \mathrm{T} \,, \tag{104.8}$$

so that the Yang–Mills Lagrangian density is

$$\mathcal{L}_{\mathrm{YM}} = -\frac{1}{4\mu_0} W_{\mu\nu}^a W^{a\mu\nu} \,, \tag{104.9}$$

of dimension J/m^3, exactly as in Equation (100.12). This is Equation (106.65) with the coupling displayed rather than absorbed.

Proposition 104.7 (Covariance and self-coupling). *Under the local transformation $\psi \mapsto U(x)\psi$ with $U(x) = \exp(-\mathrm{i}\theta^a(x)T^a)$, Equation (104.7) is covariant, $D_\mu\psi \mapsto U D_\mu \psi$, if and only if the potential transforms as*

$$\frac{g}{\hbar} W_\mu \longmapsto U \frac{g}{\hbar} W_\mu U^{-1} + \mathrm{i}\,(\partial_\mu U)\, U^{-1} \,, \qquad W_\mu := W_\mu^a T^a \,, \tag{104.10}$$

and then $W_{\mu\nu} := W_{\mu\nu}^a T^a$ transforms homogeneously, $W_{\mu\nu} \mapsto U W_{\mu\nu} U^{-1}$, so that Equation (104.9) is invariant. Expanding Equation (104.8) inside Equation (104.9) produces, besides the free quadratic term, a cubic vertex of strength $g/\hbar\mu_0$ and a quartic vertex of strength $g^2/\hbar^2\mu_0$: the gauge bosons carry the charge they mediate.

Rests on Equations (104.7) to (104.9).

Proof. *Derives Proposition 104.7.*

The first statement is Equation (100.10) with the phase replaced by a matrix and the extra term coming from the failure of U and W_μ to commute; writing $\mathcal{W}_\mu = gW_\mu/\hbar$ of dimension $1/\mathrm{m}$, the requirement $(\partial_\mu + \mathrm{i}\mathcal{W}'_\mu)U\psi = U(\partial_\mu + \mathrm{i}\mathcal{W}_\mu)\psi$ for all ψ gives $\mathrm{i}\mathcal{W}'_\mu U = U\mathrm{i}\mathcal{W}_\mu - (\partial_\mu U)$, which is Equation (104.10). For the field strength, note that $\mathrm{i}\mathcal{W}_{\mu\nu} = [D_\mu, D_\nu]$ is a commutator of covariant objects, hence a covariant object: it is Equation (106.64), and in components it is Equation (104.8) because $\left[T^b, T^c\right] = \mathrm{i}\varepsilon^{bcd}T^d$ contributes $-\varepsilon^{abc}\mathcal{W}^b_\mu\mathcal{W}^c_\nu$. Since $\mathrm{tr}(W_{\mu\nu}W^{\mu\nu})$ is invariant under conjugation, Equation (104.9) is invariant. Finally, squaring Equation (104.8) gives cross terms $(\partial W)(g\varepsilon WW/\hbar)$ and $(g\varepsilon WW/\hbar)^2$, which are the cubic and quartic self-interactions; they have no counterpart in Equation (100.12), where ε^{abc} has nothing to contract. That the gauge bosons interact with one another is a prediction, and it is measured directly in Phenomenon 104.57. □

The obstruction is now stated in its final form.

Theorem 104.8 (Gauge invariance forbids an explicit mass term)**.** *No term of the form $\frac{1}{2\mu_0}(Mc/\hbar)^2 W^a_\mu W^{a\mu}$ is invariant under Equation (104.10), and no gauge-invariant local function of W^a_μ alone gives a mass to the vector field.*

Rests on Equation (104.10).

Proof. *Derives Theorem 104.8.*

For an infinitesimal transformation, $\delta(gW^a_\mu/\hbar) = \partial_\mu\theta^a + \varepsilon^{abc}\theta^b(gW^c_\mu/\hbar)$, obtained by expanding Equation (104.10) to first order. Then

$$\delta\left(W^a_\mu W^{a\mu}\right) = \frac{2\hbar}{g}W^{a\mu}\partial_\mu\theta^a + 2\varepsilon^{abc}\theta^b W^c_\mu W^{a\mu}\,.$$

The second term vanishes because ε^{abc} is antisymmetric in ac while $W^c_\mu W^{a\mu}$ is symmetric; the first does not, and it cannot be cancelled by anything, since θ^a is an arbitrary function while $W^{a\mu}$ is a field. For the second claim, observe that a mass term is by definition quadratic in the field with no derivatives, and Equation (104.10) contains an inhomogeneous piece $\partial_\mu\theta$ that only a derivative can absorb; every gauge-invariant local functional of W alone is therefore built from $W_{\mu\nu}$ and its covariant derivatives, each of which carries at least one ∂_μ and hence at least two powers of momentum. Such a term contributes to the kinetic operator, never to a mass. □

Remark 104.9 (Gauge invariance does not require a massless quantum)*.* Theorem 104.8 forbids writing a mass down. It does not prove that the quantum is massless, and Schwinger made the distinction sharp [1241]: what fixes the mass is the position of the pole in the full propagator, not the form of a term in the Lagrangian. Write the exact photon two-point function of Corollary 100.32 as

$$\widetilde{D}_{\mu\nu}(k) \propto \frac{\eta_{\mu\nu} - k_\mu k_\nu/k^2}{k^2\left[1 - \Pi(k^2)\right]}\,,$$

which is transverse, and hence gauge invariant, whatever Π is. If the vacuum polarization $\Pi(k^2)$ is regular at $k^2 = 0$, the pole sits at $k^2 = 0$ and the quantum is massless. If instead $\Pi(k^2) \to -M^2c^2/k^2$ as $k^2 \to 0$ — a pole in Π — the propagator's pole moves to $k^2 = M^2c^2$ and the quantum is massive, with the transversality, and therefore the gauge invariance, untouched. Schwinger exhibited this in two-dimensional electrodynamics, where the massless-fermion loop produces exactly such a pole. The physical realization in three space dimensions is the superconductor: the photon inside one is massive, with $M = \hbar/\lambda_L c$ for the London penetration depth λ_L, and nothing about electromagnetic gauge invariance is lost (Chapter 130). What supplies the pole in Π is a massless scalar excitation, and the rest of this section is the search for one.

104.2.3 Glashow's $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$

The charged weak currents of Equation (101.35) do not close under commutation: $[J^+, J^-]$ produces a third, electrically neutral current. A gauge theory containing them must therefore have at least three generators, and the smallest group with three is SU(2). Glashow [1206] took the further step of adjoining a U(1) so that electromagnetism could be accommodated without identifying the photon with the neutral member of the triplet, which is impossible because the photon does not violate parity while the SU(2) currents maximally do. Salam and Ward reached the same structure independently [1295].

Definition 104.10 (The electroweak gauge group and its representations)**.** The gauge group is $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ with gauge fields W_μ^a and B_μ and couplings g, g', acting through Equation (101.70),

$$D_\mu = \partial_\mu + \frac{\mathrm{i}}{\hbar}\left(g W_\mu^a T^a + g' B_\mu \frac{Y}{2} \right) . \tag{104.11}$$

The subscript L is not decoration: T^a acts on the left-chiral projection $\psi_L = \tfrac{1}{2}(\mathbb{1} - \gamma^5)\psi$ only, and annihilates ψ_R. One generation of matter is

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L , \quad e_R , \quad \begin{pmatrix} u \\ d \end{pmatrix}_L , \quad u_R , \quad d_R , \tag{104.12}$$

with the quarks additionally carrying colour, on which $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ acts trivially. The hypercharge Y is assigned so that

$$Q = T^3 + \frac{Y}{2} \tag{104.13}$$

reproduces the observed electric charges, Q measured in units of e.

Table 104.1: Electroweak quantum numbers of one generation, in the convention $Q = T^3 + Y/2$ of Equation (104.13). Right-chiral fields are $\mathrm{SU}(2)_L$ singlets, so $T = T^3 = 0$ and $Y = 2Q$. The hypercharge of the scalar doublet Φ is fixed in Section 104.4.1 by the requirement that its lower component be neutral. The Standard Model as tested contains no ν_R; Chapter 103 takes up what the observation of neutrino mass adds.

Field	T	T^3	Y	Q
ν_{eL}	$\frac{1}{2}$	$+\frac{1}{2}$	-1	0
e_L	$\frac{1}{2}$	$-\frac{1}{2}$	-1	-1
e_R	0	0	-2	-1
u_L	$\frac{1}{2}$	$+\frac{1}{2}$	$+\frac{1}{3}$	$+\frac{2}{3}$
d_L	$\frac{1}{2}$	$-\frac{1}{2}$	$+\frac{1}{3}$	$-\frac{1}{3}$
u_R	0	0	$+\frac{4}{3}$	$+\frac{2}{3}$
d_R	0	0	$-\frac{2}{3}$	$-\frac{1}{3}$
Φ^+	$\frac{1}{2}$	$+\frac{1}{2}$	$+1$	$+1$
Φ^0	$\frac{1}{2}$	$-\frac{1}{2}$	$+1$	0

Remark 104.11 (Two normalizations of hypercharge). This chapter and Chapter 101 write $Q = T^3 + Y/2$, which is the convention of the original papers and of Table 104.1. Proposition 106.96 writes $Q = T^3 + Y$, with every hypercharge half as large — so the left-handed quark doublet carries $Y = 1/6$ there and $Y = 1/3$ here. Both are in wide use and neither is wrong; the ratios that carry physics, and every anomaly condition, are identical. The reader comparing the two must halve or double, and nothing else.

That the assignments of Table 104.1 are consistent at all is a nontrivial statement, since Y must be constant on each $\mathrm{SU}(2)_L$ multiplet while Q is not.

Proposition 104.12 (Hypercharge is determined, not chosen)**.** *Given Equation* (104.13)*, the requirement that Y take one value on each irreducible $\mathrm{SU}(2)_L$ multiplet fixes Y on that multiplet to twice the average electric charge of its members, and is consistent only if the charges within a doublet differ by exactly one unit.*

Rests on Equation (104.13)*.*

Proof. *Derives Proposition 104.12.*

For a doublet with members of charge $Q_\uparrow$, $Q_\downarrow$ and $T^3 = \pm\frac{1}{2}$, Equation (104.13) gives $Y = 2Q_\uparrow - 1 = 2Q_\downarrow + 1$; the two are equal if and only if $Q_\uparrow - Q_\downarrow = 1$, and then $Y = Q_\uparrow + Q_\downarrow$, twice the average. The leptonic doublet (ν_e, e) has $0 - (-1) = 1$ and the quark doublet (u, d) has $\frac{2}{3} - (-\frac{1}{3}) = 1$: both pass, and it is the charged current that requires it, since W^+ carries exactly one unit of charge. The right-chiral fields are singlets, so $T^3 = 0$ and $Y = 2Q$ with no constraint at all. Nothing so far explains *why* the charges come in these ratios; that is the content of Section 104.5.3, where the same numbers are forced by consistency at one loop. □

Two consequences follow immediately and were Glashow's main results. First, since B_μ and W^3_μ are both neutral, the photon must be a combination of them and the orthogonal combination is a second neutral boson: the model predicts a *weak neutral current*, of a kind Fermi's theory does not contain and no experiment had then seen. Proposition 101.67 carries out the rotation and reads off the fermion couplings; the experimental discovery is Section 104.6.1. Second, the neutral current so obtained is flavour diagonal at tree level provided the quark doublets are complete, which is the mechanism of Glashow, Iliopoulos and Maiani [1423] proved as Theorem 101.59 and used again in Section 104.4.3.

Remark 104.13 (What Glashow's paper could not do). The 1961 paper inserts the W and Z masses by hand [1206], and says so: its closing paragraph records that the resulting theory is not renormalizable. By Theorem 104.4 that is not a blemish to be tidied up later but the same disease Fermi's theory had, moved to a different place. The group structure was right and the mass terms were fatal, and the six years between 1961 and 1967 were spent finding out how to have the first without the second.

104.3 Spontaneous symmetry breaking

104.3.1 The superconducting analogy

A symmetry of the dynamics need not be a symmetry of the state the system actually occupies. A ferromagnet below its Curie temperature has a magnetization pointing somewhere, although the exchange Hamiltonian singles out no direction; a crystal has axes, although the interatomic potential is rotationally invariant. Nothing about this is quantum mechanical, and Chapter 123 treats it as the general phenomenon of an ordered phase. What Nambu saw [1187] is that the same thing can happen to the *vacuum* of a relativistic field theory, and that when it does, the consequences are as sharp as in the laboratory.

Definition 104.14 (Spontaneous breaking)**.** Let a theory have an action invariant under a group G and let $|0\rangle$ be its ground state. The symmetry is *spontaneously broken* to the subgroup $H = \{h \in G : U(h)\,|0\rangle = |0\rangle\}$ when $H \neq G$. The set of states $\{U(g)\,|0\rangle : g \in G\}$ is then in bijection with the coset space G/H, all of them degenerate in energy, and the choice among them is not made by the dynamics.

The ancestor of the Higgs field is not an elementary particle at all but the Ginzburg–Landau order parameter of a superconductor [957]: a complex field ψ whose free energy

density, near the transition, is $a\,|\psi|^2 + \frac{1}{2}b\,|\psi|^4$ with a changing sign at T_c. The following proposition is the entire mathematical content of that model, and it will be used unchanged for the electroweak potential.

Proposition 104.15 (The Mexican-hat minimum)**.** *Let ϕ be a complex field of dimension $\mathrm{J}^{1/2}/\mathrm{m}^{1/2}$ and let*

$$V(\phi) = -\mu^2\,|\phi|^2 + \lambda\,|\phi|^4\,, \qquad \left[\mu^2\right] = 1/\mathrm{m}^2\,, \qquad [\lambda] = 1/(\mathrm{J\,m})\,, \tag{104.14}$$

with $\mu^2 > 0$ and $\lambda > 0$. Then $\phi = 0$ is a local maximum*, the minima form the circle*

$$|\phi| = \frac{v}{\sqrt{2}}\,, \qquad v = \mu\sqrt{\frac{1}{\lambda}}\,, \tag{104.15}$$

and the depth of the well is $V_{\min} = -\lambda v^4/4$.
Rests on Definition 104.14.

Proof. *Derives Proposition 104.15.*
Write $|\phi|^2 = \rho$. Then $V = -\mu^2\rho + \lambda\rho^2$, $\frac{\mathrm{d}V}{\mathrm{d}\rho} = -\mu^2 + 2\lambda\rho$, which vanishes at $\rho = \mu^2/2\lambda$, and $\mathrm{d}^2V/\mathrm{d}\rho^2 = 2\lambda > 0$ there, so it is the minimum; $\rho = 0$ gives $\mathrm{d}V/\mathrm{d}\rho = -\mu^2 < 0$, so the origin is a maximum. Setting $\rho = v^2/2$ gives $v^2 = \mu^2/\lambda$, which is Equation (104.15), and $V_{\min} = -\mu^2 v^2/2 + \lambda v^4/4 = -\lambda v^4/4$ on using $\mu^2 = \lambda v^2$. The dimensions check: $[\mu^2\,|\phi|^2] = 1/\mathrm{m}^2 \times \mathrm{J/m} = \mathrm{J/m}^3$, an energy density, and likewise for the quartic. Every point of the circle is a minimum and the phase of ϕ at the minimum is undetermined: that undetermined phase is the degeneracy of Definition 104.14, with $G = \mathrm{U}(1)$ and H trivial. □

Remark 104.16 (Nambu's transfer, and what it explained)*.* In the Bardeen–Cooper–Schrieffer ground state (Chapter 130) the electron number symmetry is spontaneously broken by the condensate, and the single-particle spectrum acquires a gap Δ. Nambu [1187] recognized the gap equation as a self-consistent mass generation and asked what the relativistic analogue would be. With Jona-Lasinio he built it [1215]: a theory of massless fermions with a chirally invariant four-fermion interaction develops, above a critical coupling, a non-zero $\langle\bar\psi\psi\rangle$ and hence a fermion mass that no term in the Lagrangian contains. The model also produced a massless pseudoscalar bound state, which Nambu identified with the pion — the first appearance of what the next subsection proves must always be there. Two lessons carry over. Mass can be a property of the vacuum rather than of the Lagrangian; and a broken continuous symmetry leaves a massless excitation behind. The strong interaction realizes both, and Chapter 102 takes up the pion as an approximate Goldstone boson of chiral symmetry.

104.3.2 Goldstone's theorem

Goldstone's model [1207] is Proposition 104.15 promoted to a field theory. Write $\phi = \frac{1}{\sqrt{2}}(v + h + \mathrm{i}\chi)$ with h, χ real fields measuring the departure from one chosen minimum, radially and tangentially. Substituting into $\mathcal{L} = \partial_\mu\phi^*\partial^\mu\phi - V(\phi)$ and expanding gives

$$\mathcal{L} = \frac{1}{2}\partial_\mu h\,\partial^\mu h - \frac{1}{2}\left(2\lambda v^2\right)h^2 + \frac{1}{2}\partial_\mu\chi\,\partial^\mu\chi - \lambda v\,h\left(h^2+\chi^2\right) - \frac{\lambda}{4}\left(h^2+\chi^2\right)^2 + \frac{\lambda v^4}{4}\,, \tag{104.16}$$

in which the field χ has no mass term at all. By the dictionary of Notation 106.23, a mass term $\frac{1}{2}(mc/\hbar)^2$ means

$$m_h = \frac{\hbar}{c}\sqrt{2\lambda v^2} = \frac{\hbar\mu\sqrt{2}}{c}\,, \qquad m_\chi = 0\,. \tag{104.17}$$

The massless mode is the one that moves along the circle of minima, where the potential is flat by construction. That this is no accident of the quartic potential is Goldstone's theorem.

Theorem 104.17 (Goldstone's theorem, tree level)**.** *Let ϕ_i, $i = 1, \dots, n$, be real scalar fields with a potential $V(\phi)$ invariant under a compact connected group G acting as $\delta\phi_i = \theta^a \, (t^a)_{ij} \, \phi_j$ with real antisymmetric generators t^a, $a = 1, \dots, \dim G$. Let ϕ_0 minimize V and let $H \subseteq G$ be its stabilizer. Then the mass matrix*

$$\left(M^2\right)_{ij} = \left. \frac{\partial^2 V}{\partial\phi_i \, \partial\phi_j} \right|_{\phi_0} \tag{104.18}$$

annihilates every vector $t^a\phi_0$, so the spectrum contains exactly $\dim G - \dim H$ massless scalars, one for each broken generator.

Rests on Definition 104.14.

Proof. *Derives Theorem 104.17.*

Invariance of V under an infinitesimal transformation reads, for every ϕ and every a,

$$\frac{\partial V}{\partial\phi_i} \, (t^a)_{ij} \, \phi_j = 0 \, . \tag{104.19}$$

This is an identity in ϕ, so it may be differentiated. Applying $\partial/\partial\phi_k$ gives

$$\frac{\partial^2 V}{\partial\phi_k \, \partial\phi_i} \, (t^a)_{ij} \, \phi_j + \frac{\partial V}{\partial\phi_i} \, (t^a)_{ik} = 0 \, .$$

Now evaluate at $\phi = \phi_0$. The second term vanishes because ϕ_0 is a stationary point of V, and the first becomes

$$\left(M^2\right)_{ki} \, (t^a\phi_0)_i = 0 \, . \tag{104.20}$$

Hence each vector $t^a\phi_0$ lies in the kernel of M^2. If t^a belongs to the Lie algebra of H then $t^a\phi_0 = 0$ and Equation (104.20) is empty; otherwise $t^a\phi_0 \neq 0$ and it is a genuine null eigenvector. The vectors $\{t^a\phi_0\}$ span a space of dimension $\dim G - \dim H$, because the map $t \mapsto t\phi_0$ from the Lie algebra of G to $\mathbb{R}^n$ has kernel exactly the Lie algebra of H by definition of the stabilizer; and M^2, being a real symmetric matrix, has as many zero eigenvalues as the dimension of its kernel. Since ϕ_0 is a *minimum*, M^2 is positive semi-definite and no eigenvalue is negative. Therefore the spectrum consists of $\dim G - \dim H$ massless modes and $n - \dim G + \dim H$ massive ones.

Geometrically, $t^a\phi_0$ is tangent to the orbit $G\phi_0$ at ϕ_0, and V is constant on that orbit by Equation (104.19): a direction along which the potential does not change is a direction in which the field costs no energy to excite at zero momentum, which is what a massless particle is. □

Remark 104.18 (The exact theorem, and its assumptions)*.* Theorem 104.17 is a statement about a classical potential, and Proposition 106.45 shows how loops modify a potential. The result nevertheless survives: Goldstone, Salam and Weinberg [1226] proved it to all orders, and the proof uses only the existence of a conserved current j^μ with $\langle 0 | \, [Q, \mathcal{O}] \, | 0 \rangle \neq 0$ for some local operator $\mathcal{O}$, together with Lorentz invariance, locality and a positive-definite state space. The spectral representation of Proposition 106.47, applied to $\langle 0 | j^\mu(x) \mathcal{O}(0) | 0 \rangle$, then forces a zero-mass contribution. Each assumption is load-bearing, and their precise statement belongs to Chapter 109. The one that fails in this chapter is the last but one: gauge fixing a gauge theory either breaks manifest Lorentz invariance (radiation gauge) or introduces negative-norm states (covariant gauge), and *that* is the loophole the next subsection walks through. It is not a technicality — it is why the photon in a superconductor can be massive without a massless mode appearing in the spectrum.

104.3.3 The Anderson–Higgs mechanism

Anderson [1245] pointed out, in a non-relativistic setting and before any of the relativistic papers, that a gauge field coupled to a system with a broken symmetry does not leave a massless scalar behind: the would-be Goldstone mode and the gauge field combine, and what is left is a massive vector. In a plasma the resulting mass is the plasma frequency; in a superconductor it is the inverse London penetration depth. The relativistic version is worked here in full, for one abelian gauge field, because every feature of the electroweak case is already present.

Theorem 104.19 (The abelian Higgs mechanism)**.** *Let ϕ be a complex scalar of charge q coupled to a* U(1) *gauge field through $D_\mu = \partial_\mu + \mathrm{i}qA_\mu/\hbar$, with*

$$\mathcal{L} = -\frac{1}{4\mu_0}F_{\mu\nu}F^{\mu\nu} + (D_\mu\phi)^*\,(D^\mu\phi) + \mu^2\,|\phi|^2 - \lambda\,|\phi|^4\ . \tag{104.21}$$

Then the spectrum consists of one real scalar of mass $m_h = \hbar\sqrt{2\lambda v^2}/c$ and one massive *vector of mass*

$$m_A = \frac{qv\sqrt{\mu_0}}{c}\,, \tag{104.22}$$

with no massless scalar. The count of degrees of freedom is unchanged: two (complex scalar) plus two (massless vector) before, one plus three after.

Rests on Equations (100.8)*,* (104.6) *and* (104.16)*.*

Proof. *Derives Theorem 104.19.*

Write $\phi(x) = \frac{1}{\sqrt{2}}\,[v + h(x)]\exp[\mathrm{i}\chi(x)/v]$, which is legitimate wherever $\phi \neq 0$ and which parametrizes the same two real degrees of freedom as before, h radial and χ tangential; χ/v is dimensionless because $[\chi] = [v]$. The potential depends only on $|\phi|$, so it is exactly Equation (104.16) without the χ terms: h gets the mass Equation (104.17) and χ appears nowhere in V.

Now perform the gauge transformation Equation (100.8) with $\Lambda = -\hbar\chi/qv$, under which $\phi \mapsto \mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\phi = \frac{1}{\sqrt{2}}(v + h)$: the phase is gone. This choice is the *unitary gauge*, and it is available precisely because χ shifts inhomogeneously under a gauge transformation, exactly as A_μ does. What was a field is now a gauge parameter. The Lagrangian becomes

$$\begin{aligned}\mathcal{L} &= -\frac{1}{4\mu_0}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}\partial_\mu h\,\partial^\mu h + \frac{q^2}{2\hbar^2}\,(v+h)^2\,A_\mu A^\mu - V(h)\\ &= -\frac{1}{4\mu_0}F_{\mu\nu}F^{\mu\nu} + \frac{q^2v^2}{2\hbar^2}A_\mu A^\mu + \frac{1}{2}\partial_\mu h\,\partial^\mu h - V(h) + \frac{q^2v}{\hbar^2}hA_\mu A^\mu + \frac{q^2}{2\hbar^2}h^2A_\mu A^\mu\,,\end{aligned} \tag{104.23}$$

using $|D_\mu\phi|^2 = \frac{1}{2}(\partial h)^2 + \frac{1}{2}(q/\hbar)^2(v+h)^2A^2$ for a real ϕ. The second term of Equation (104.23) is a mass term. Comparing it with the Proca form $\frac{1}{2\mu_0}(m_Ac/\hbar)^2A^2$ gives

$$\frac{m_A^2c^2}{\mu_0\hbar^2} = \frac{q^2v^2}{\hbar^2}\,, \qquad \text{that is} \qquad m_A = \frac{qv\sqrt{\mu_0}}{c}\,,$$

which is Equation (104.22). Dimensionally, $q^2v^2\mu_0/c^2$ carries $\mathrm{C}^2 \times \mathrm{J/m} \times \mathrm{kg\,m/C^2} \times \mathrm{s^2/m^2} = \mathrm{kg}^2$, so m_A is a mass.

For the counting: before breaking, ϕ has two real components and a massless vector has two transverse polarizations, four in all; after, h has one and a massive vector has three, again four. The Goldstone mode has not disappeared, it has become the longitudinal polarization whose absence from the massless theory Equation (104.6) identified as the problem. The last two terms of Equation (104.23) are the interactions of the surviving scalar with the vector, and their strengths are not free: both are fixed by m_A and v. That is the statement measured in Phenomenon 104.62. □

Corollary 104.20 (The Meissner effect is the same computation)**.** *In the static limit, Equation* (104.23) *gives the field equation* $(\nabla^2 - m_A^2 c^2/\hbar^2)\,\boldsymbol{A} = 0$ *for the transverse potential, so a magnetic field entering the ordered medium is screened exponentially over the length*

$$\lambda_L = \frac{\hbar}{m_A c} = \frac{\hbar}{q v \sqrt{\mu_0}}\,, \tag{104.24}$$

the Compton wavelength of the massive photon. This is the Meissner effect, observed in 1933 [724] and measured as a penetration depth of tens of nanometres in elemental superconductors (Chapter 130).

Rests on Equations (104.22) *and* (104.23).

Proof. *Derives Corollary 104.20.*

Varying Equation (104.23) with respect to A_μ gives $\partial_\mu F^{\mu\nu} = -\mu_0 q^2 v^2 \hbar^{-2} A^\nu$, and in a static configuration with $\partial_0 = 0$ and $\nabla \cdot \boldsymbol{A} = 0$ the spatial part reads $\nabla^2 \boldsymbol{A} = \mu_0 q^2 v^2 \hbar^{-2} \boldsymbol{A} = (m_A c/\hbar)^2 \boldsymbol{A}$ by Equation (104.22). A solution decaying into the half-space $z > 0$ is $\boldsymbol{A} \propto \mathrm{e}^{-z/\lambda_L}$ with λ_L as stated, and $\boldsymbol{B} = \operatorname{curl} \boldsymbol{A}$ decays with it. □

Remark 104.21 (Why this evades Goldstone's theorem). Theorem 104.17 is not contradicted; its hypotheses are not met. In unitary gauge the massless scalar is simply absent from the field content, and the theory manifestly has no massless spin-zero particle — but unitary gauge is not manifestly Lorentz covariant in its propagator structure, and the vector propagator does not fall off (Theorem 104.4(ii)), so the assumptions used in Remark 104.18 fail. In a covariant gauge the Goldstone field χ is present, but so are negative-norm states, and the positivity assumption fails instead. Either way one hypothesis goes. Which one is a matter of bookkeeping, and Section 104.5.1 shows that the freedom to choose is exactly what makes the theory both unitary and renormalizable.

104.3.4 The 1964 papers

Three groups published the relativistic mechanism in 1964, within four months of one another, and they proved different things.

Englert and Brout [1275] computed the vacuum polarization of a gauge field coupled to a scalar with a non-zero expectation value and exhibited the pole in $\Pi(k^2)$ that Remark 104.9 identified as the criterion: their result is that the gauge boson mass emerges from a resummation, and they gave it for a general non-abelian group in the same paper.

Higgs's first paper, in *Physics Letters* [1282], is the negative half of the argument: it shows by explicit construction that the Goldstone theorem's proof breaks in a gauge theory quantized in a Lorentz gauge, because the current is not the generator of a symmetry of the physical state space. His second paper, in *Physical Review Letters* [1283], works the abelian model of Theorem 104.19 and states the consequence the others did not emphasize: besides the massive vector there survives *one massive scalar quantum*, with definite couplings. That particle is the subject of Section 104.7, and it is the reason the mechanism is falsifiable rather than merely consistent. Higgs returned to its properties in 1966 [1332], computing its decays into pairs of gauge bosons.

Guralnik, Hagen and Kibble [1279] worked in radiation gauge, where Lorentz invariance is not manifest, and showed explicitly that the massless mode still occurs in the field algebra but decouples from the physical states: their treatment is the one that makes the evasion of Theorem 104.17 transparent rather than mysterious.

Kibble then supplied what the Standard Model actually uses [1352]: the extension to a non-abelian group, where the vector bosons need not all acquire the same mass and some may acquire none.

Theorem 104.22 (Kibble: the gauge-boson mass matrix)**.** *Let a set of scalars Φ transform in a representation of a gauge group with generators T^a and couplings g_a, coupled through $D_\mu = \partial_\mu + \mathrm{i} g_a A^a_\mu T^a/\hbar$, and let Φ acquire the expectation value Φ_0. Then the gauge fields acquire the mass matrix*

$$\left(M^2\right)^{ab} = \frac{2\mu_0}{c^2}\,\Re\!\left[(g_a T^a \Phi_0)^\dagger \left(g_b T^b \Phi_0\right)\right] \quad \textit{(no sum on } a,\ b\textit{)}, \tag{104.25}$$

whose kernel is spanned by the generators annihilating Φ_0. Hence the gauge bosons of the unbroken subgroup H stay exactly massless, one massive vector appears for each broken generator, and that is also the number of Goldstone modes removed from the scalar spectrum by Theorem 104.17.

Rests on Theorems 104.17 and 104.19.

Proof. *Derives Theorem 104.22.*

Setting $\Phi = \Phi_0$ in the scalar kinetic term, $(D_\mu\Phi_0)^\dagger (D^\mu\Phi_0) = \hbar^{-2}\,(g_a T^a\Phi_0)^\dagger \left(g_b T^b \Phi_0\right) A^a_\mu A^{b\mu}$, since $\partial_\mu \Phi_0 = 0$. Only the part symmetric in ab survives the contraction with $A^a_\mu A^{b\mu}$, which is the real part. Comparing with the Proca form $\frac{1}{2\mu_0}(c/\hbar)^2 (M^2)^{ab} A^a_\mu A^{b\mu}$ gives Equation (104.25).

The matrix is a Gram matrix of the vectors $g_a T^a \Phi_0$, so it is positive semi-definite and a vector n^a lies in its kernel if and only if $n^a g_a T^a \Phi_0 = 0$, that is, if and only if the corresponding generator annihilates the vacuum. The rank of M^2 therefore equals the number of independent broken generators, which by Theorem 104.17 is the number of massless scalars the ungauged theory would have had. Each of them is removed from the scalar spectrum by the unitary-gauge argument of Theorem 104.19 applied direction by direction, and reappears as the longitudinal polarization of the corresponding massive vector: the bookkeeping balances generator by generator, not merely in total. □

Remark 104.23 (What the theorem buys, stated in advance). Theorem 104.22 is why the photon is *exactly* massless in the electroweak theory rather than approximately so. If a single generator combination annihilates Φ_0, the corresponding vector has strictly zero mass to all orders — because the unbroken U(1) remains a genuine gauge symmetry and Theorem 104.8 still applies to it. The experimental bound on the photon mass, $m_\gamma < 1.6 \times 10^{-50}\,\mathrm{kg}$ in the laboratory and some four orders of magnitude tighter from solar-system magnetic fields (Phenomenon 100.7), therefore tests a structural feature of the symmetry-breaking pattern and not a fitted parameter.

104.4 The Glashow–Weinberg–Salam model

104.4.1 The electroweak Lagrangian

Weinberg's 1967 paper [1362] applies Theorem 104.22 to Glashow's group with the smallest scalar multiplet that will do the job, and Salam presented the same construction the following year [1381]. The choice of multiplet is not free: it must break $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ down to a U(1) under which the electron is charged and the neutrino is not, and it must do so leaving exactly one unbroken generator.

Definition 104.24 (The scalar sector)**.** Let Φ be a complex $\mathrm{SU}(2)_L$ doublet of hypercharge $Y = 1$,

$$\Phi = \begin{pmatrix} \Phi^+ \\ \Phi^0 \end{pmatrix}, \qquad [\Phi] = \mathrm{J}^{1/2}/\mathrm{m}^{1/2}, \tag{104.26}$$

with the Lagrangian density

$$\mathcal{L}_\Phi = (D_\mu\Phi)^\dagger (D^\mu\Phi) + \mu^2 \Phi^\dagger\Phi - \lambda\left(\Phi^\dagger\Phi\right)^2, \tag{104.27}$$

D_μ being Equation (104.11) and $\mu^2 > 0$, $\lambda > 0$. By Equation (104.13) the upper component has $Q = +1$ and the lower $Q = 0$: the hypercharge assignment $Y = 1$ is exactly what makes the lower component electrically neutral, and hence what makes it able to acquire an expectation value without breaking electromagnetism. The SI dimensions are $[\mu^2] = 1/\mathrm{m}^2$ and $[\lambda] = 1/(\mathrm{J\,m})$, so that $\hat{\lambda} = \lambda\hbar c$ is dimensionless (Equation (106.34)); $\hbar\mu c$ is an energy.

The whole electroweak Lagrangian is now

$$\mathcal{L}_{\mathrm{EW}} = -\frac{1}{4\mu_0} W^a_{\mu\nu} W^{a\mu\nu} - \frac{1}{4\mu_0} B_{\mu\nu} B^{\mu\nu} + \mathrm{i}\hbar c \sum_\psi \bar{\psi}\gamma^\mu D_\mu \psi + \mathcal{L}_\Phi + \mathcal{L}_{\mathrm{Yuk}}\,, \tag{104.28}$$

each term of dimension $\mathrm{J/m^3}$: the two gauge kinetic terms as in Equation (104.9), the fermion kinetic terms as in Equation (100.5) but with *no mass terms*, the scalar sector Equation (104.27), and the Yukawa couplings of Section 104.4.3. The sum over ψ runs over the chiral fields of Table 104.1, with D_μ carrying that field's own T^a and Y. There is not one free mass in it.

Theorem 104.25 (The breaking pattern)**.** *The potential in Equation (104.27) is minimized on the three-sphere $\Phi^\dagger\Phi = v^2/2$ with $v = \mu/\sqrt{\lambda}$. An $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ rotation brings any point of it to*

$$\Phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \tag{104.29}$$

whose stabilizer is generated by $Q = T^3 + Y/2$ alone. Hence

$$\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y \longrightarrow \mathrm{U}(1)_{em}\,, \tag{104.30}$$

three of the four generators are broken, three of the four real scalar fields become the longitudinal polarizations of $W^\pm$ and Z^0, one real scalar survives as a particle, and the photon is exactly massless.

Rests on Equation (104.27) and Theorems 104.17 and 104.22.

Proof. *Derives Theorem 104.25.*

The potential depends on Φ only through $\rho = \Phi^\dagger\Phi$, so Proposition 104.15 applies verbatim with $|\phi|^2$ replaced by ρ: the minimum is at $\rho = \mu^2/2\lambda = v^2/2$. The set of Φ with $\Phi^\dagger\Phi = v^2/2$ is a sphere S^3 in the four real dimensions of the doublet, and SU(2) acts transitively on S^3 (it *is* S^3 as a manifold), so every minimum is gauge-equivalent to Equation (104.29).

For the stabilizer, compute the action of each generator on Φ_0. With $T^a = \frac{1}{2}\sigma^a$ and $Y = 1$,

$$\begin{aligned} T^1\Phi_0 &= \frac{v}{2\sqrt{2}} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, & T^2\Phi_0 &= \frac{-\mathrm{i}v}{2\sqrt{2}} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\ T^3\Phi_0 &= \frac{-v}{2\sqrt{2}} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, & \frac{Y}{2}\Phi_0 &= \frac{v}{2\sqrt{2}} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \end{aligned} \tag{104.31}$$

none of which vanishes: each of the four generators moves the vacuum. But the combination $Q = T^3 + Y/2$ does annihilate it, by the last two lines of Equation (104.31), and it is the only combination that does, since the first two lines are linearly independent of one another and of the third. The unbroken subgroup is therefore one-dimensional and generated by Q, which is Equation (104.30); that Q is the electric charge is Equation (104.13), so what survives is electromagnetism and not some other U(1).

By Theorem 104.17 the ungauged model would have $\dim G - \dim H = 4 - 1 = 3$ massless scalars, leaving $4 - 3 = 1$ massive one. By Theorem 104.22 the gauge-boson mass matrix has rank 3 and a one-dimensional kernel spanned by the generator Q, so exactly one vector — the photon — stays massless, exactly to all orders by Remark 104.23, and three become massive by absorbing the three Goldstone modes. □

Proposition 104.26 (The vacuum expectation value in SI)**.** *Matching the low-energy limit of W exchange to the Fermi constant, Equation (104.4), fixes*

$$v^2 = \frac{(\hbar c)^2}{\sqrt{2}G_F}\,, \qquad E_v = v\sqrt{\hbar c} = \left(\frac{(\hbar c)^3}{\sqrt{2}G_F}\right)^{1/2} = 246.22\,\mathrm{GeV}\,, \tag{104.32}$$

that is $E_v = 3.9449 \times 10^{-8}\,\mathrm{J}$, $v = 2.2186 \times 10^{5}\,\mathrm{J}^{1/2}/\mathrm{m}^{1/2}$, and an associated length

$$\frac{\hbar c}{E_v} = 8.0143 \times 10^{-19}\,\mathrm{m} = 0.801\,43\,\mathrm{am}\,. \tag{104.33}$$

Rests on Equations (104.3) and (104.4).

Derivation. *Derives Proposition 104.26.*

The W mass is computed in Theorem 104.28 below as $M_W c^2 = \frac{1}{2}\hat{g}E_v$; substituting it into the matching condition Equation (104.4),

$$\frac{G_F}{\sqrt{2}\,(\hbar c)^3} = \frac{\hat{g}^2}{8\,(M_W c^2)^2} = \frac{\hat{g}^2}{8\cdot\frac{1}{4}\hat{g}^2E_v^2} = \frac{1}{2E_v^2}\,,$$

so $E_v^2 = (\hbar c)^3/\sqrt{2}G_F$, which is Equation (104.32); the coupling $\hat{g}$ has cancelled, which is why v is known far more precisely than either boson mass. Numerically, with $\sqrt{2} \times 1.166\,378\,7 \times 10^{-5} = 1.649\,508\,6 \times 10^{-5}$ from Equation (104.3), the reciprocal is $60\,624.1\,\mathrm{GeV}^2$ and its square root is $246.220\,\mathrm{GeV}$. Converting, $246.220\,\mathrm{GeV} \times 1.602\,176\,634 \times 10^{-10}\,\mathrm{J/GeV} = 3.944\,87 \times 10^{-8}\,\mathrm{J}$; dividing by $\sqrt{\hbar c} = 1.778\,068\,3 \times 10^{-13}\,\mathrm{J}^{1/2}\,\mathrm{m}^{1/2}$ gives v; and dividing $\hbar c = 3.161\,527 \times 10^{-26}\,\mathrm{J\,m}$ by E_v gives Equation (104.33). □

Remark 104.27 (What 246 GeV is, and what it is not)*.* E_v is not a mass and not the energy of anything. It is the value of a field, converted to an energy by the two constants available; the physical content of Equation (104.33) is that the electroweak condensate has a structure length of 0.8 am, about a thousandth of a proton radius ($8.4075 \times 10^{-16}\,\mathrm{m}$, [2431]). Every mass in the electroweak sector is E_v times a dimensionless coupling, so what the Standard Model explains is mass *ratios*; the overall scale is one input number, and why it is 10^{-17} of any gravitational scale is Section 104.8.2.

Theorem 104.28 (The gauge boson masses)**.** *With the vacuum Equation (104.29), the quadratic terms in the gauge fields obtained from $(D_\mu\Phi_0)^\dagger\,(D^\mu\Phi_0)$ are*

$$\frac{g^2v^2}{4\hbar^2}W_\mu^+W^{-\mu} + \frac{v^2}{8\hbar^2}\left(gW_\mu^3 - g'B_\mu\right)^2\,, \qquad W_\mu^\pm = \frac{W_\mu^1 \mp \mathrm{i}W_\mu^2}{\sqrt{2}}\,, \tag{104.34}$$

and therefore, defining

$$Z_\mu = \frac{gW_\mu^3 - g'B_\mu}{\sqrt{g^2+g'^2}}\,, \qquad A_\mu = \frac{g'W_\mu^3 + gB_\mu}{\sqrt{g^2+g'^2}}\,, \tag{104.35}$$

the masses

$$M_W c^2 = \frac{1}{2}\hat{g}E_v\,, \qquad M_Z c^2 = \frac{1}{2}\sqrt{\hat{g}^2+\hat{g}'^2}\,E_v\,, \qquad M_\gamma = 0\,, \tag{104.36}$$

with $\hat{g}$, $\hat{g}'$ and E_v as in Notation 104.1.

Rests on Equation (104.29), Notation 104.1, and Theorem 104.22.

Proof. *Derives Theorem 104.28.*

Since $\partial_\mu\Phi_0 = 0$, $D_\mu\Phi_0 = \frac{\mathrm{i}}{\hbar}\left(gW_\mu^aT^a + \frac{1}{2}g'B_\mu\right)\Phi_0$, using $Y = 1$. Then

$$(D_\mu\Phi_0)^\dagger\,(D^\mu\Phi_0) = \frac{1}{\hbar^2}\Phi_0^\dagger\left(gW_\mu^aT^a + \tfrac{1}{2}g'B_\mu\right)\left(gW^{b\mu}T^b + \tfrac{1}{2}g'B^\mu\right)\Phi_0\,.$$

Insert $T^a = \frac{1}{2}\sigma^a$ and $\Phi_0 = \frac{v}{\sqrt{2}}(0,1)^{\mathsf{T}}$. The matrix in the middle is

$$\frac{1}{2}\begin{pmatrix} gW^3 + g'B & g\,(W^1 - \mathrm{i}W^2) \\ g\,(W^1 + \mathrm{i}W^2) & -gW^3 + g'B \end{pmatrix} = \frac{1}{2}\begin{pmatrix} gW^3 + g'B & \sqrt{2}gW^+ \\ \sqrt{2}gW^- & -gW^3 + g'B \end{pmatrix},$$

so acting on Φ_0 it returns $\frac{v}{2\sqrt{2}}\left(\sqrt{2}gW_\mu^+, -gW_\mu^3 + g'B_\mu\right)^{\mathsf{T}}$, whose squared norm is

$$\frac{v^2}{8}\left[2g^2 W_\mu^+ W^{-\mu} + \left(gW_\mu^3 - g'B_\mu\right)^2\right],$$

which divided by $\hbar^2$ is Equation (104.34). Note that B_μ and W_μ^3 enter only through the single combination $gW^3 - g'B$: the orthogonal combination A_μ of Equation (104.35) is absent, so it has no mass, which is the algebraic form of Theorem 104.22 here.

Reading off the masses requires the Proca normalizations. A complex vector has mass term $\mu_0^{-1}\,(M_W c/\hbar)^2\, W_\mu^+ W^{-\mu}$ and a real one $\frac{1}{2\mu_0}\,(M_Z c/\hbar)^2\, Z_\mu Z^\mu$, so

$$M_W^2 = \frac{\mu_0 g^2 v^2}{4c^2}, \qquad M_Z^2 = \frac{\mu_0\,(g^2 + g'^2)\,v^2}{4c^2}.$$

Now convert to the hatted couplings. With $g = \hat{g}\sqrt{\varepsilon_0 \hbar c}$ from Equation (104.1) and $v = E_v/\sqrt{\hbar c}$ from Equation (104.2),

$$M_W c^2 = \frac{c\sqrt{\mu_0}\,g v}{2} = \frac{c\sqrt{\mu_0\varepsilon_0}\,\hat{g}\sqrt{\hbar c}\,E_v}{2\sqrt{\hbar c}} = \frac{\hat{g}E_v}{2},$$

because $c\sqrt{\mu_0\varepsilon_0} = 1$ exactly. The same substitution in M_Z gives the second of Equation (104.36). Dimensionally $\mu_0 g^2 v^2/c^2$ carries $\mathrm{kg\,m/C^2 \times C^2 \times J/m \times s^2/m^2 = kg^2}$, as a squared mass must. □

Corollary 104.29 (The physical scalar and its self-coupling)**.** *Writing* $\Phi = \frac{1}{\sqrt{2}}\,(0, v + h)^{\mathsf{T}}$ *in unitary gauge, the surviving real scalar* h *has*

$$m_h c^2 = \hbar c\, v\sqrt{2\lambda} = \sqrt{2\hat{\lambda}}\,E_v, \qquad \hbar\mu c = \frac{m_h c^2}{\sqrt{2}}, \tag{104.37}$$

and the measured $m_h c^2 = (125.20 \pm 0.11)\,\mathrm{GeV}$ *[2427] gives*

$$\hat{\lambda} = \frac{1}{2}\left(\frac{m_h c^2}{E_v}\right)^2 = 0.1293, \qquad \hbar\mu c = 88.53\,\mathrm{GeV}. \tag{104.38}$$

Rests on Equations (104.14) *and* (104.16) *and Notation 106.23.*

Proof. *Derives Corollary 104.29.*

The potential restricted to the radial direction is Equation (104.14) with $|\phi|^2 = (v + h)^2/2$, so Equation (104.16) applies unchanged: the h mass term is $\frac{1}{2}(2\lambda v^2)h^2$, which by Notation 106.23 means $(m_h c/\hbar)^2 = 2\lambda v^2$. Multiplying by $(\hbar c)^2$, $(m_h c^2)^2 = 2(\hbar c)^2\lambda v^2 = 2(\lambda\hbar c)(v^2\hbar c) = 2\hat{\lambda}E_v^2$, which is Equation (104.37); and $\mu^2 = \lambda v^2$ gives $\hbar\mu c = m_h c^2/\sqrt{2}$. Numerically $m_h c^2/E_v = 125.20/246.22 = 0.508\,48$, whose square halved is $0.129\,28$, and $125.20/\sqrt{2} = 88.53$. □

Remark 104.30 (The energy stored in the vacuum). Proposition 104.15 gives the depth of the well as $V_{\min} = -\lambda v^4/4 = -\hat{\lambda}E_v^4/4(\hbar c)^3$, which with Equation (104.38) is $-2.5 \times 10^{45}\,\mathrm{J/m^3}$. Gravity responds to absolute energy density, and the observed dark-energy density is of order $1 \times 10^{-9}\,\mathrm{J/m^3}$ [2348]: the electroweak condensate alone exceeds it by some 54 orders of magnitude. Nothing in this chapter, or anywhere in the Standard Model, cancels that. It is recorded here because it is a consequence of the mechanism just derived and not of anything speculative, and it is taken up as an open problem in Chapter 133.

104.4.2 Mixing and the weak mixing angle

The rotation Equation (104.35) is a rotation by an angle, and Proposition 101.67 has already carried out the consequences for the fermion couplings. Collecting what Theorem 104.28 adds:

Proposition 104.31 (The tree-level electroweak relations)**.** *Define* θ_W *by*

$$\cos\theta_W = \frac{\hat{g}}{\sqrt{\hat{g}^2+\hat{g}'^2}}\,, \qquad \sin\theta_W = \frac{\hat{g}'}{\sqrt{\hat{g}^2+\hat{g}'^2}}\,. \tag{104.39}$$

Then

$$M_W c^2 = \frac{1}{2}\hat{g}E_v\,, \qquad \frac{M_W}{M_Z} = \cos\theta_W\,, \qquad \hat{e} = \hat{g}\sin\theta_W = \hat{g}'\cos\theta_W\,, \tag{104.40}$$

and, eliminating $\hat{g}$ *in favour of* $\hat{e}$ *and* E_v *and using Equation* (104.32)*,*

$$\left(M_W c^2\right)^2 \sin^2\theta_W = \frac{\pi\alpha\,(\hbar c)^3}{\sqrt{2}G_F} = 1389.8\,\mathrm{GeV}^2\,. \tag{104.41}$$

Four measurable quantities — α, G_F, M_W, M_Z — are thus expressed through three parameters $\hat{g}$, $\hat{g}'$, v, so one relation among them is a prediction.

Rests on Equations (104.32) *and* (104.36) *and Proposition 101.67.*

Proof. *Derives Proposition 104.31.*

The first two of Equation (104.40) are Equation (104.36) with Equation (104.39). The third is Equation (101.73), whose derivation is Proposition 101.67: substituting Equation (104.35) into Equation (104.11), the coefficient of A_μ is $\hat{g}\sin\theta_W T^3 + \hat{g}'\cos\theta_W Y/2$, which is proportional to $Q = T^3 + Y/2$ if and only if the two coefficients coincide, and then their common value is $\hat{e}$. For Equation (104.41), square the first relation and use $\hat{g}\sin\theta_W = \hat{e}$: $(M_W c^2)^2\sin^2\theta_W = \frac{1}{4}\hat{e}^2E_v^2 = \pi\alpha E_v^2$, since $\hat{e}^2 = 4\pi\alpha$; and $E_v^2 = (\hbar c)^3/\sqrt{2}G_F$ by Equation (104.32). Numerically, $\pi \times 7.297\,352\,564\,3 \times 10^{-3} \times 60\,624.1\,\mathrm{GeV}^2 = 1389.8\,\mathrm{GeV}^2$. □

Two numbers can now be extracted from the measured masses.

Proposition 104.32 (The couplings, from the masses)**.** *With* $M_W c^2 = (80.369 \pm 0.013)\,\mathrm{GeV}$ *and* $M_Z c^2 = (91.1880 \pm 0.0020)\,\mathrm{GeV}$ *[2427] and* E_v *from Equation* (104.32)*,*

$$\hat{g} = 0.6528\,, \qquad \hat{g}' = 0.3500\,, \qquad \sin^2\theta_W\big|_{\text{on-shell}} = 1 - \frac{M_W^2}{M_Z^2} = 0.2232\,. \tag{104.42}$$

The value of $\hat{e}$ *implied by Equation* (104.40) *is then* 0.3084*, corresponding to* $\alpha = \hat{e}^2/4\pi = 1/132.1$ *— which is neither* $\alpha(0) = 1/137.036$ *[2431] nor* $\alpha(M_Z c^2) = 1/128.94$ *(Phenomenon 100.62), but lies between them.*

Rests on Equations (104.32) *and* (104.40)*.*

Derivation. *Derives Proposition 104.32.*

$\hat{g} = 2M_W c^2/E_v = 2(80.369)/246.220 = 0.652\,82$, and $\sqrt{\hat{g}^2+\hat{g}'^2} = 2M_Z c^2/E_v = 2(91.1880)/246.220 = 0.740\,70$, whence $\hat{g}'^2 = 0.548\,643 - 0.426\,179 = 0.122\,464$ and $\hat{g}' = 0.349\,95$. Then $\sin^2\theta_W = \hat{g}'^2/(\hat{g}^2+\hat{g}'^2) = 0.122\,464/0.548\,643 = 0.223\,21$, which is also $1 - (80.369/91.1880)^2$ as Equation (104.40) requires. Finally $\hat{e} = \hat{g}\sin\theta_W = 0.652\,82 \times 0.472\,45 = 0.308\,43$ and $\hat{e}^2/4\pi = 0.095\,129/12.566\,37 = 0.007\,569\,9 = 1/132.10$. □

Remark 104.33 (Why the tree relations do not close, and by how much). Proposition 104.32 is the chapter's first honest failure and it is instructive. Four measured quantities were fitted with three parameters, and the leftover is a 3.7 per cent discrepancy in $\hat{e}^2$ — far outside every experimental error. Nothing is wrong: the relations are tree-level, and a one-loop

electroweak theory is not a tree-level one. The discrepancy is conventionally packaged as a single correction Δr defined by replacing Equation (104.41) with

$$\left(M_W c^2\right)^2 \left(1 - \frac{M_W^2}{M_Z^2}\right) = \frac{\pi \alpha \, (\hbar c)^3}{\sqrt{2} G_F \, (1 - \Delta r)} \, , \tag{104.43}$$

and its value follows from the measured masses. Solving Equation (104.43) for M_W with $\Delta r = 0$ and $\alpha = \alpha(0)$ gives

$$\left(M_W c^2\right)^2 = \frac{\left(M_Z c^2\right)^2}{2} \left[1 + \sqrt{1 - \frac{4 \times 1389.8 \, \mathrm{GeV}^2}{\left(M_Z c^2\right)^2}}\right] ,$$

that is $M_W c^2 = 80.94 \, \mathrm{GeV}$, against the measured $80.369 \, \mathrm{GeV}$; matching them requires $\Delta r = 0.036$. Two effects account for most of it, and they act in opposite directions. The running of the electromagnetic coupling from zero momentum transfer to $M_Z c^2$ (Phenomenon 100.62), with $\alpha^{-1}(M_Z c^2) = 128.94$, contributes $\Delta\alpha = 1 - \alpha(0)/\alpha(M_Z c^2) \approx 0.059$, and the top quark contributes $-\left(\cos^2\theta_W / \sin^2\theta_W\right) \Delta\rho \approx -3.48 \times 0.0093 = -0.032$ through Equation (104.52) below; the sum, 0.027, lands within about 0.009 of the required 0.036, the remainder coming from the scalar-mass logarithm, the remaining bosonic self-energies and higher orders. That these two one-loop terms capture the bulk of a 3.7 per cent effect — and that the rest is calculable — is the reason the electroweak fit of Section 104.6.4 can measure the top and scalar masses without producing either particle.

Remark 104.34 (There is no such thing as *the* weak mixing angle). Proposition 104.32 defines $\sin^2\theta_W$ by the mass ratio. Other definitions are in use and they do not agree, because beyond tree level the angle is a renormalized quantity and a renormalized quantity needs a scheme (Definition 100.54). Three are standard:

- the *on-shell* angle, $s_W^2 := 1 - M_W^2/M_Z^2 = 0.2232$, defined to all orders by the physical masses;
- the $\overline{\mathrm{MS}}$ angle at the Z scale, $\hat{s}_Z^2 = 0.231\,29 \pm 0.000\,04$, defined by the ratio of running couplings in the scheme of Equation (100.75);
- the *effective leptonic* angle $\sin^2\theta_{\mathrm{eff}}^{\ell} = 0.231\,55 \pm 0.000\,04$, defined so that the measured Z-pole asymmetries take their tree-level form,

the last two as quoted in the electroweak review of [2427]. The differences, about one per cent, are hundreds of times the experimental errors, so a number quoted without its scheme is not a measurement of anything. This is why the honest statement of the unification test is Phenomenon 104.55: not that some number equals 0.23, but that one running, scheme-defined angle fits every process at every momentum transfer.

The first determination of the angle away from the Z was Prescott's measurement of the parity-violating asymmetry in the scattering of longitudinally polarized electrons from deuterium at SLAC [1625], which gave $\sin^2\theta_W = 0.20 \pm 0.03$, sharpened to 0.224 ± 0.020 by the follow-up scan over beam energies [1643]; the measurement belongs to Chapter 110. Its significance was not the number but that the number agreed with the one Gargamelle had extracted from neutrino scattering in an entirely different process.

Phenomenon 104.35 (The boson masses and the mixing angle are not independent)**.** The measured masses of the two weak bosons,

$$m_W c^2 = (80.369 \pm 0.013) \, \mathrm{GeV} \, , \qquad m_Z c^2 = (91.1880 \pm 0.0020) \, \mathrm{GeV} \tag{104.44}$$

[2427] determine a mixing angle through $\cos\theta_W = m_W/m_Z$, and that angle agrees with the one determined independently from the neutral-current couplings of Phenomenon 104.55. Equivalently the parameter

$$\rho := \frac{m_W^2}{m_Z^2 \cos^2\theta_W} \tag{104.45}$$

is measured to be unity to within the small radiative corrections [2104, 2427]. Three quantities measured in unrelated ways — two masses and a coupling ratio — satisfy one relation that nothing obliged them to satisfy. The statement has content only when the same renormalization scheme defines the angle on both sides: taken as a *definition* of θ_W from the masses, $\rho = 1$ says nothing, and what is tested is the agreement of that angle with the one the couplings give.

Rests on Equations (104.36) *and* (104.39).

Derivation. *Derives Phenomenon 104.35.*

The masses are Equation (104.36), so $m_W^2/m_Z^2 = \hat{g}^2/(\hat{g}^2 + \hat{g}'^2) = \cos^2\theta_W$ by Equation (104.39), and Equation (104.45) gives $\rho = 1$ identically. Two features of that one-line computation carry the physics.

First, E_v has cancelled. The relation is therefore a statement about the *representation* the breaking field occupies and not about how strongly the symmetry is broken; it survives any number of doublets and, as Theorem 104.41 shows, fails for almost any other multiplet.

Second, the angle in Equation (104.45) is defined by Equation (104.39) from the couplings, while the ratio on the left is measured from masses. What is tested is that these two determinations agree, and the test is quantitative only when a scheme is fixed, per Remark 104.34. In the on-shell scheme $\rho = 1$ is an identity and the content moves into the agreement of the mass-derived angle with the asymmetry-derived one; in $\overline{\text{MS}}$ the content stays in ρ itself, whose measured deviation from unity is the few-per-mille effect predicted by Equation (104.52). The measured ρ is in either case evidence that the electroweak symmetry is broken by doublets, which is a fact about the vacuum, not a convention. □

104.4.3 Fermion masses and the Yukawa couplings

Equation (104.28) contains no fermion mass term, and that was not an omission.

Proposition 104.36 (Chiral gauge charges forbid a bare Dirac mass)**.** *For any fermion of Table 104.1, the Dirac mass term* $-mc^2\bar{\psi}\psi = -mc^2\left(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L\right)$ *is not invariant under* $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$.

Rests on Definition 104.10 and Table 104.1.

Proof. *Derives Proposition 104.36.*

Insert $\psi = \psi_L + \psi_R$ with $\psi_{L,R} = \frac{1}{2}(\mathbb{1} \mp \gamma^5)\psi$ and use $\bar{\psi}_L\psi_L = \bar{\psi}_R\psi_R = 0$, which follows from $\gamma^5\gamma^0 = -\gamma^0\gamma^5$: the mass term connects the two chiralities and nothing else. But ψ_L is a component of an $\mathrm{SU}(2)_L$ doublet while ψ_R is a singlet, so $\bar{\psi}_L\psi_R$ transforms as a doublet, not a singlet, and cannot be a term in an invariant Lagrangian. Even setting the $\mathrm{SU}(2)_L$ aside, the hypercharges do not match: for the electron $Y_L = -1$ and $Y_R = -2$, so $\bar{\psi}_L\psi_R$ carries $Y = -2 - (-1) = -1 \neq 0$ and is not $\mathrm{U}(1)_Y$ invariant either. The statement is not that the electron is massless; it is that its mass cannot be a parameter of the symmetric Lagrangian, and must come from the vacuum. □

Definition 104.37 (Yukawa couplings)**.** Let Q_L, L_L be the left-chiral doublets and u_R, d_R, e_R the singlets, each carrying a generation index $i = 1, 2, 3$. The most general gauge-invariant, renormalizable coupling to Φ is

$$\mathcal{L}_{\text{Yuk}} = -\left[y^d_{ij}\bar{Q}_{Li}\Phi\, d_{Rj} + y^u_{ij}\bar{Q}_{Li}\tilde{\Phi}\, u_{Rj} + y^e_{ij}\bar{L}_{Li}\Phi\, e_{Rj}\right] + \text{h.c.}\,, \tag{104.46}$$

where

$$\tilde{\Phi} := \mathrm{i}\sigma^2\Phi^* = \begin{pmatrix} \Phi^{0*} \\ -\Phi^- \end{pmatrix} \tag{104.47}$$

is the conjugate doublet, which transforms as a doublet with $Y = -1$. The couplings y_{ij} are arbitrary complex matrices of dimension $\mathrm{J}^{1/2}\,\mathrm{m}^{1/2}$, so that $\hat{y} := y/\sqrt{\hbar c}$ is dimensionless.

Derivation of the dimension and of Equation (104.47). *Derives Equation* (104.47).

A fermion bilinear $\bar{\psi}\psi$ carries $1/\mathrm{m}^3$ by Definition 100.2 and Φ carries $\mathrm{J}^{1/2}/\mathrm{m}^{1/2}$, so for Equation (104.46) to be an energy density, $[y] = \mathrm{J}/\mathrm{m}^3 \times \mathrm{m}^3 \times \mathrm{m}^{1/2}/\mathrm{J}^{1/2} = \mathrm{J}^{1/2}\,\mathrm{m}^{1/2}$, and dividing by $\sqrt{\hbar c} = \mathrm{J}^{1/2}\,\mathrm{m}^{1/2}$ leaves a pure number. As for $\tilde{\Phi}$: under $\Phi \mapsto U\Phi$ with $U \in \mathrm{SU}(2)$, $\Phi^* \mapsto U^*\Phi^*$, and the identity $\sigma^2 U^* \sigma^2 = U$, valid for every SU(2) matrix because the doublet representation is *pseudoreal* — unitarily equivalent to its own complex conjugate — gives $\mathrm{i}\sigma^2\Phi^* \mapsto U\,(\mathrm{i}\sigma^2\Phi^*)$. Complex conjugation flips the sign of Y, so $\tilde{\Phi}$ has $Y = -1$, and $\bar{Q}_L\tilde{\Phi}u_R$ carries $Y = -\frac{1}{3} - 1 + \frac{4}{3} = 0$ as required, while $\bar{Q}_L\Phi d_R$ carries $Y = -\frac{1}{3} + 1 - \frac{2}{3} = 0$. Without $\tilde{\Phi}$ the up-type quarks could not be given mass by this doublet at all. □

Theorem 104.38 (Masses, mixing, and where the CKM matrix comes from)**.** *Substituting Equation* (104.29) *into Equation* (104.46) *produces mass matrices* $m^f_{ij}c^2 = \hat{y}^f_{ij}E_v/\sqrt{2}$. *Each is a general complex matrix and is brought to non-negative diagonal form by a* biunitary *transformation,*

$$U_f^\dagger m^f V_f = \mathrm{diag}\,(m_{f1}, m_{f2}, m_{f3})\ , \tag{104.48}$$

with U_f *acting on left-chiral and* V_f *on right-chiral fields. In the mass basis so defined,*

1. *the charged current acquires the matrix* $V_{\mathrm{CKM}} = U_u^\dagger U_d$, *which is unitary but not in general diagonal;*
2. *the neutral currents — both the photon's and the* Z*'s — remain exactly diagonal in flavour;*
3. *the couplings of* h *to fermions remain diagonal, and proportional to the masses.*

Rests on Equations (104.29) *and* (104.46) *and Theorem 101.59.*

Proof. *Derives Theorem 104.38.*

With $\Phi \to \Phi_0$ the term $y^d_{ij}\bar{Q}_{Li}\Phi d_{Rj}$ becomes $\left(y^d_{ij}v/\sqrt{2}\right)\bar{d}_{Li}d_{Rj}$, which compared with $-mc^2\bar{\psi}_L\psi_R$ identifies $m^d_{ij}c^2 = \hat{y}^d_{ij}\sqrt{\hbar c}\,v/\sqrt{2} = \hat{y}^d_{ij}E_v/\sqrt{2}$; likewise for u and e, the conjugate doublet supplying $\Phi^{0*} \to v/\sqrt{2}$ for the up type.

For Equation (104.48): given any complex m, the matrix $mm^\dagger$ is Hermitian positive semi-definite, hence $mm^\dagger = U\Sigma^2U^\dagger$ with U unitary and $\Sigma \geq 0$ diagonal, by the spectral theorem for normal operators. Define $V := m^\dagger U\Sigma^{-1}$ where Σ is invertible; then $V^\dagger V = \Sigma^{-1}U^\dagger mm^\dagger U\Sigma^{-1} = \Sigma^{-1}\Sigma^2\Sigma^{-1} = \mathbb{1}$, so V is unitary, and $U^\dagger mV = U^\dagger mm^\dagger U\Sigma^{-1} = \Sigma$. (If Σ has zero entries, restrict to its support and extend V arbitrarily unitarily on the complement.) This is the singular-value decomposition, and the two unitaries are independent precisely because the left- and right-chiral fields are independent — which is possible only because the theory is chiral.

(i) The charged current is $\bar{u}_{Li}\gamma^\mu d_{Li}$ in the gauge basis. Substituting $u_L = U_u u'_L$ and $d_L = U_d d'_L$ with primes denoting mass eigenstates gives $\bar{u}'_L U_u^\dagger\gamma^\mu U_d d'_L = \bar{u}'_L\gamma^\mu\left(U_u^\dagger U_d\right)d'_L$, which is Equation (101.67) with $V_{\mathrm{CKM}} = U_u^\dagger U_d$, manifestly unitary as the product of two unitaries — which is Theorem 101.62 and is what Chapter 5 licenses.

(ii) A neutral current is $\bar{f}_{Li}\gamma^\mu$ (coefficient) f_{Li} with a coefficient depending only on T^3 and Q (Equation (101.75)), hence *the same for every generation*; it is therefore proportional to the identity matrix in generation space, and $U_f^\dagger\mathbb{1}U_f = \mathbb{1}$. Nothing survives the rotation. This is the Glashow–Iliopoulos–Maiani mechanism [1423] in its general form, proved as

Theorem 101.59; note that the argument needs the coefficient to depend only on the gauge quantum numbers, which fails as soon as two fields with the *same* T^3 and Q but different masses can mix — the reason a second scalar doublet, or an extra quark that is not part of a complete doublet, generically produces flavour-changing neutral currents and is excluded by the kaon and B-meson data of Chapter 111.

(iii) The Yukawa term with $\Phi = \frac{1}{\sqrt{2}}(0, v+h)^{\mathsf{T}}$ is the mass term multiplied by $(1 + h/v)$, so the coupling matrix of h is m^f/v — the *same* matrix, diagonalized by the same U_f, V_f. Hence h has no flavour-changing couplings and its coupling to each fermion is $m_f c^2/E_v$ per unit field, which is the statement measured in Phenomenon 104.62. □

Table 104.2: Dimensionless Yukawa couplings $\hat{y}_f = \sqrt{2}\, m_f c^2/E_v$ implied by the measured fermion masses of [2427] and $E_v = 246.22\,\text{GeV}$. The quark masses are scheme-dependent quantities — u, d, s quoted in $\overline{\text{MS}}$ at 2 GeV, c and b at their own scale, the top from direct reconstruction — so these numbers are indicative to a few per cent, not metrological; the point is the range they span. No principle in the Standard Model relates them: they are nine of its free parameters.

Fermion	$m_f c^2$ [GeV]	$\hat{y}_f$
t	172.57	0.991
b	4.183	0.0240
τ	1.776 93	0.0102
c	1.273	0.007 31
μ	0.105 658 4	6.07×10^{-4}
s	0.0935	5.37×10^{-4}
d	0.004 70	2.70×10^{-5}
u	0.002 16	1.24×10^{-5}
e	$5.109\,990 \times 10^{-4}$	2.93×10^{-6}

Remark 104.39 (The one coupling of order unity). Table 104.2 spans a factor of 3.4×10^5 from top to electron. Exactly one entry is of order one, and it is the top's: $\hat{y}_t = 0.991$, within one per cent of unity. Whether that is a coincidence is not known. What is known is that it makes the top quark the dominant virtual particle in every electroweak loop, which is why $\Delta\rho$ in Equation (104.52) goes as m_t^2, why the vacuum stability analysis of Section 104.8.1 turns on the top mass at the 1 GeV level, and why the electroweak fit could locate the top before it was produced. The spread itself is an input, not a prediction: Table 104.2 lists numbers the theory takes from experiment, and Chapter 133 records that no evidence-based explanation of the pattern exists. Neutrino masses are not in the table at all, because Equation (104.46) cannot generate them without a right-chiral neutrino; that the masses are nonetheless non-zero (Chapters 103 and 114) is the one laboratory fact this Lagrangian does not accommodate.

104.4.4 Custodial symmetry and the ρ parameter

$\rho = 1$ came out of Phenomenon 104.35 as an accident of the doublet. It is not an accident: it follows from a global symmetry of the scalar sector that the gauge couplings do not respect, which is why the corrections to it are small and calculable.

Theorem 104.40 (Custodial symmetry)**.** *Assemble the doublet and its conjugate into the* 2×2 *matrix*

$$\mathcal{M} := \begin{pmatrix} \tilde{\Phi} & \Phi \end{pmatrix} = \begin{pmatrix} \Phi^{0*} & \Phi^{+} \\ -\Phi^{-} & \Phi^{0} \end{pmatrix}. \tag{104.49}$$

Then $\Phi^\dagger\Phi = \frac{1}{2}\operatorname{tr}\left(\mathcal{M}^\dagger\mathcal{M}\right)$*, so the scalar potential in Equation* (104.27) *is invariant under the* global *group*

$$\mathcal{M} \longmapsto L\,\mathcal{M}\,R^\dagger\,, \qquad (L,R)\in \mathrm{SU}(2)_L\times\mathrm{SU}(2)_R\,, \tag{104.50}$$

larger than the gauged $\mathrm{SU}(2)_L\times\mathrm{U}(1)_Y$*. The vacuum* $\mathcal{M}_0 = \left(v/\sqrt{2}\right)\mathbb{1}$ *is invariant under the diagonal subgroup* $L=R$*, the* custodial $\mathrm{SU}(2)_V$*, under which* (W^1,W^2,W^3) *transforms as a triplet. Hence in the limit* $g'\to 0$ *the three fields have equal masses and* $\rho = 1$ *exactly.*

Rests on Equations (104.27)*,* (104.35) *and* (104.45)*.*

Proof. *Derives Theorem 104.40.*

That $\Phi^\dagger\Phi = \frac{1}{2}\operatorname{tr}(\mathcal{M}^\dagger\mathcal{M})$ is a computation: $\operatorname{tr}(\mathcal{M}^\dagger\mathcal{M}) = 2\left(|\Phi^0|^2+|\Phi^+|^2\right)$, since the two columns of $\mathcal{M}$ have the same norm. A trace of the form $\operatorname{tr}(\mathcal{M}^\dagger\mathcal{M})$ is invariant under Equation (104.50) for any unitary L, R, and the potential depends on $\mathcal{M}$ only through it. The kinetic term is

$$(D_\mu\Phi)^\dagger\,(D^\mu\Phi) = \frac{1}{2}\operatorname{tr}\left[(D_\mu\mathcal{M})^\dagger\,D^\mu\mathcal{M}\right]\,, \qquad D_\mu\mathcal{M} = \partial_\mu\mathcal{M} + \frac{\mathrm{i}g}{\hbar}W_\mu^a T^a\mathcal{M} - \frac{\mathrm{i}g'}{\hbar}B_\mu\mathcal{M}T^3\,,$$

the relative sign in the last term coming from the opposite hypercharge of $\tilde\Phi$. The $\mathrm{SU}(2)_L$ term is invariant under $\mathrm{SU}(2)_R$, since R acts on the right; the $\mathrm{U}(1)_Y$ term is not, because only T^3 of $\mathrm{SU}(2)_R$ is gauged. So $\mathrm{SU}(2)_R$ is a symmetry of the whole Lagrangian when $g'=0$ and of the scalar potential always.

At the vacuum, $\Phi^+=0$ and $\Phi^0 = v/\sqrt{2}$, so $\mathcal{M}_0 = \left(v/\sqrt{2}\right)\mathbb{1}$ and $L\mathcal{M}_0R^\dagger = \mathcal{M}_0$ if and only if $LR^\dagger = \mathbb{1}$, that is $L=R$: the unbroken global group is the diagonal $\mathrm{SU}(2)_V$. Under it the gauge fields W_μ^a, which carry an adjoint index of $\mathrm{SU}(2)_L$, transform as an $\mathrm{SU}(2)_V$ triplet, and a mass matrix invariant under a group that acts irreducibly on a triplet must be proportional to the identity by Schur's lemma (Chapter 5). Hence $M_{W^1} = M_{W^2} = M_{W^3}$ when $g'=0$. Reinstating g' mixes W^3 with B but changes no mass in the W sector, so $M_W = M_{W^3} = M_Z\cos\theta_W$ by Equation (104.35), and Equation (104.45) gives $\rho = 1$. □

Theorem 104.41 (Ross–Veltman: ρ counts the representation)**.** *Let the symmetry be broken by several complex scalar multiplets labelled* i*, the* i*-th having weak isospin* T_i *and acquiring an expectation value* $v_i/\sqrt{2}$ *in the component of third component* T_i^3*, that component being electrically neutral. Then at tree level*

$$\rho = \frac{\sum_i\left[T_i\,(T_i+1) - \left(T_i^3\right)^2\right]v_i^2}{2\sum_i\left(T_i^3\right)^2 v_i^2}\,. \tag{104.51}$$

For doublets ($T=\frac{1}{2}$*,* $T^3=-\frac{1}{2}$) *this is* 1 *for any number of them; for a complex triplet with* $T^3=\pm1$ *it is* $\frac{1}{2}$*; for a triplet breaking through its* $T^3=0$ *component the denominator vanishes. The measurement* $\rho=1$ *is thus evidence about the multiplet structure of the breaking sector [1561, 1659].*

Rests on Theorem 104.28 and Equation (104.45)*.*

Proof. *Derives Theorem 104.41.*

Repeat the computation of Theorem 104.28 for a general multiplet. Writing $T^\pm = T^1\pm\mathrm{i}T^2$ and $W_\mu^\pm = (W_\mu^1 \mp \mathrm{i}W_\mu^2)/\sqrt{2}$, one has $T^1W_\mu^1 + T^2W_\mu^2 = \frac{1}{\sqrt{2}}\left(T^+W_\mu^+ + T^-W_\mu^-\right)$, so the part of $\hbar^{-2}\left|gW_\mu^aT^a\Phi_0\right|^2$ containing $W_\mu^+W^{-\mu}$ is

$$\frac{g^2}{2\hbar^2}\Phi_0^\dagger\left\{T^+,T^-\right\}\Phi_0\,W_\mu^+W^{-\mu} = \frac{g^2}{\hbar^2}\Phi_0^\dagger\left[T\,(T+1) - \left(T^3\right)^2\right]\Phi_0\,W_\mu^+W^{-\mu}\,,$$

using $\{T^+,T^-\}=2\left[(T^1)^2+(T^2)^2\right]=2\left[\boldsymbol{T}^2-(T^3)^2\right]$ and the eigenvalue $T(T+1)$ of the Casimir (Chapter 14). With $\Phi_0^\dagger\Phi_0=v_i^2/2$ in the i-th multiplet, summing over multiplets and matching to the Proca form gives

$$M_W^2=\frac{\mu_0 g^2}{2c^2}\sum_i\left[T_i\left(T_i+1\right)-\left(T_i^3\right)^2\right]v_i^2\,.$$

For the neutral sector, the coefficient of the vacuum in the neutral covariant derivative is $gT^3W_\mu^3+g'\left(Y/2\right)B_\mu$ with $Y/2=Q-T^3$, and $Q\Phi_0=0$ because the component acquiring the expectation value is neutral; hence the coefficient is $\left(gW_\mu^3-g'B_\mu\right)T^3$, and

$$M_Z^2=\frac{\mu_0\left(g^2+g'^2\right)}{c^2}\sum_i\left(T_i^3\right)^2v_i^2\,.$$

Dividing, with $\cos^2\theta_W=g^2/\left(g^2+g'^2\right)$, every coupling cancels and Equation (104.51) remains. Substituting $T=\frac12$, $T^3=-\frac12$ gives numerator $\left(\frac34-\frac14\right)v^2$ and denominator $2\cdot\frac14v^2$, equal; substituting $T=1$, $T^3=-1$ gives $(2-1)\,v^2$ over $2v^2$, that is $\frac12$; and $T=1$, $T^3=0$ gives $2v^2$ over 0. □

Custodial symmetry is broken by two things in the real theory, and each produces a calculable shift.

Proposition 104.42 (The sources of custodial violation)**.** *$\mathrm{SU}(2)_V$ is violated (i) by the hypercharge coupling g', which gauges only T^3 of $\mathrm{SU}(2)_R$, and (ii) by the Yukawa matrices, which are $\mathrm{SU}(2)_R$ symmetric only if $y^u=y^d$ within each doublet. Since $\hat y_t=0.991$ and $\hat y_b=0.0240$ by Table 104.2, the third generation violates it maximally, and it dominates.*

Rests on Theorem 104.40 and Equation (104.46).

Proof. *Derives Proposition 104.42.*

(i) is proved inside Theorem 104.40: the B_μ term carries $\mathcal{M}T^3$ on the right and is invariant only under the U(1) subgroup of $\mathrm{SU}(2)_R$ generated by T^3. (ii) follows by writing Equation (104.46) in the $\mathcal{M}$ notation: $\bar Q_L\mathcal{M}\,\mathrm{diag}(y^u,y^d)\,Q_R$ with $Q_R=(u_R,d_R)^{\mathsf T}$ is invariant under $\mathcal{M}\mapsto\mathcal{M}R^\dagger$, $Q_R\mapsto RQ_R$ if and only if $\mathrm{diag}(y^u,y^d)$ commutes with every $R\in\mathrm{SU}(2)$, that is if and only if $y^u=y^d$. □

Proposition 104.43 (The top contribution to ρ)**.** *At one loop the leading custodial violation from the third generation is*

$$\Delta\rho:=\rho-1=\frac{3G_F\left(m_tc^2\right)^2}{8\sqrt2\pi^2\left(\hbar c\right)^3}+\cdots=9.3\times10^{-3}\,,\tag{104.52}$$

for $m_tc^2=(172.57\pm0.29)\,\mathrm{GeV}$ [2427], the omitted terms being the bottom mass (negligible) and the logarithmic scalar contribution. The dependence is quadratic *in the top mass, and only logarithmic in m_h — the screening theorem of Veltman [1612].*

Rests on Proposition 104.42 and Equation (104.32).

Derivation of the parametric form. *Derives Proposition 104.43.*

$\Delta\rho$ measures the difference between the W and Z self-energies at zero momentum, $\Delta\rho=\Pi_{WW}(0)/M_W^2c^2-\Pi_{ZZ}(0)/M_Z^2c^2$, each computed from a quark loop. By Proposition 104.42 the difference vanishes when $m_t=m_b$, so it is proportional to the isospin splitting. Each self-energy carries two Yukawa or gauge vertices and one loop factor $1/16\pi^2$ (Equation (100.82) shows the same factor in QED), and the only dimensionful quantities available are the quark masses and E_v; since $\Delta\rho$ is dimensionless and the vertices are the Yukawa couplings $\hat y\propto m_fc^2/E_v$, the leading term must be $N_c\left(m_tc^2\right)^2/16\pi^2E_v^2$ times a pure number, with

$N_c = 3$ for colour. Writing $E_v^2 = (\hbar c)^3/\sqrt{2}G_F$ turns this into Equation (104.52) up to the pure number, which the explicit loop integration of [1612] fixes to exactly 1: Equation (104.52) *is* $N_c\,(m_t c^2)^2/16\pi^2 E_v^2$. Numerically, $3 \times 1.166\,378\,7 \times 10^{-5} \times (172.57)^2 = 1.0421$ and $8\sqrt{2}\pi^2 = 111.66$, giving 9.33×10^{-3}. That the answer grows with m_t^2 rather than $\ln m_t$ is what made the top mass predictable from LEP data before the Tevatron produced it [1891, 1892]. □

Derivation pending #157

The one-loop coefficient in the top contribution to the rho parameter: the W and Z vacuum-polarization integrals with a t–b doublet in the loop, regulated dimensionally, whose difference at zero momentum transfer gives the factor $3/8\sqrt{2}\pi^2$ quoted from Veltman's paper above; and the accompanying logarithmic dependence on the scalar mass, which is what the screening theorem asserts. Belongs in Appendix A.

Definition 104.44 (The oblique parameters)**.** Corrections that enter only through the gauge-boson self-energies — which is where any heavy new state that does not couple directly to light fermions must show up — are parametrized by three numbers S, T, U [1835], normalized to vanish for the Standard Model at a reference point. T is custodial violation, $\alpha T = \Delta\rho$; S measures the difference between the self-energies at $q^2 = M_Z^2 c^2$ and at $q^2 = 0$; U is rarely significant and is usually fixed to zero. Their measured values are consistent with zero at the level of a few hundredths [2427], which is a quantitative statement that no heavy particle with electroweak quantum numbers has yet made itself felt indirectly.

104.5 Renormalizability

104.5.1 't Hooft's proofs

Between 1967 and 1971 the Weinberg–Salam model was cited about as often as a paper is cited when nobody believes it matters. What changed was not a measurement: 't Hooft proved that the theory makes finite predictions at every order. He did it in two papers, and the citation keys used in this book do *not* follow the order of publication — [1449] is the earlier, on *massless* Yang–Mills fields, and [1448] is the later, on the spontaneously broken case; both entries in the bibliography record the fact.

The massless paper establishes that a pure Yang–Mills theory, gauge fixed by the Faddeev–Popov procedure of Theorem 106.52 with its ghosts, has divergences that are absorbed into a redefinition of the finitely many parameters already present. The broken paper is the one this chapter needs, and its instrument is a family of gauges interpolating between two descriptions each of which makes one property manifest and the other obscure.

Theorem 104.45 (The R_ξ gauges)**.** *Take the abelian Higgs model of Theorem 104.19 and, instead of the unitary gauge, write $\phi = \frac{1}{\sqrt{2}}(v + h + \mathrm{i}\chi)$ with h and χ both retained. Then*

1. *the quadratic Lagrangian contains the mixing term $(qv/\hbar)\,A^\mu \partial_\mu \chi$, which prevents the propagators from being read off;*

2. *adding the gauge-fixing term*

$$\mathcal{L}_{\mathrm{gf}} = -\frac{1}{2\mu_0 \xi}\left(\partial^\mu A_\mu - \xi\mu_0 \frac{qv}{\hbar}\chi\right)^2 \tag{104.53}$$

cancels the mixing exactly and gives χ the mass $\sqrt{\xi}\,m_A$;

3. the gauge propagator becomes

$$\widetilde{D}_{\mu\nu}(k) = \frac{-\mathrm{i}\mu_0\hbar^3 c}{k^2 - m_A^2c^2 + \mathrm{i}\epsilon}\left[\eta_{\mu\nu} - (1-\xi)\,\frac{k_\mu k_\nu}{k^2 - \xi m_A^2 c^2}\right], \tag{104.54}$$

which falls as k^{-2} for every finite ξ;

4. the Faddeev–Popov ghost acquires the same mass $\sqrt{\xi}\,m_A$ and does not decouple.

The limit $\xi \to \infty$ returns the unitary gauge and Equation (101.77); $\xi = 1$ is the 't Hooft–Feynman gauge, in which the propagator is simply $-\mathrm{i}\mu_0\hbar^3 c\,\eta_{\mu\nu}/(k^2 - m_A^2c^2)$.

Rests on Theorem 104.19 and Equations (101.77) and (104.22).

Proof. *Derives Theorem 104.45.*

(i) Expanding $|D_\mu\phi|^2$ with $D_\mu = \partial_\mu + \mathrm{i}qA_\mu/\hbar$ and $\phi = \frac{1}{\sqrt{2}}(v + h + \mathrm{i}\chi)$,

$$|D_\mu\phi|^2 = \frac{1}{2}(\partial h)^2 + \frac{1}{2}(\partial\chi)^2 + \frac{q^2}{2\hbar^2}\left[(v+h)^2 + \chi^2\right]A_\mu A^\mu + \frac{q}{\hbar}A^\mu\left[(v+h)\,\partial_\mu\chi - \chi\,\partial_\mu h\right],$$

whose last bracket contains, at quadratic order, exactly $(qv/\hbar)A^\mu\partial_\mu\chi$. A term linear in A and linear in χ makes the quadratic form non-diagonal in field space.

(ii) Expand Equation (104.53):

$$\mathcal{L}_{\mathrm{gf}} = -\frac{1}{2\mu_0\xi}(\partial^\mu A_\mu)^2 + \frac{qv}{\hbar}(\partial^\mu A_\mu)\,\chi - \frac{\xi\mu_0 q^2 v^2}{2\hbar^2}\chi^2 .$$

The middle term integrates by parts to $-(qv/\hbar)A^\mu\partial_\mu\chi$ up to a total derivative, cancelling (i). The last term is a mass term for χ; comparing with $-\frac{1}{2}(m_\chi c/\hbar)^2\chi^2$ and using $\mu_0 q^2 v^2/\hbar^2 = (m_A c/\hbar)^2$ from Equation (104.22) gives $m_\chi = \sqrt{\xi}\,m_A$. The would-be Goldstone boson thus has a gauge-dependent mass, which is the sharpest possible statement that it is not a particle.

(iii) With the mixing gone, the gauge kernel in momentum space (Fourier convention Equation (100.14)) is

$$\widetilde{M}_{\mu\nu} = -\frac{1}{\mu_0\hbar^2}\left[\left(k^2 - m_A^2c^2\right)\eta_{\mu\nu} - \left(1 - \xi^{-1}\right)k_\mu k_\nu\right],$$

which on the transverse projector $P^{\mathrm{T}}_{\mu\nu} = \eta_{\mu\nu} - k_\mu k_\nu/k^2$ is $-(k^2 - m_A^2c^2)/\mu_0\hbar^2$ and on the longitudinal projector P^{L} is $-(k^2 - \xi m_A^2c^2)/\mu_0\hbar^2\xi$. Inverting on each eigenspace and multiplying by $\mathrm{i}\hbar c$, exactly as in Proposition 100.8, gives

$$\widetilde{D}_{\mu\nu} = -\mathrm{i}\mu_0\hbar^3 c\left[\frac{P^{\mathrm{T}}_{\mu\nu}}{k^2 - m_A^2c^2} + \frac{\xi P^{\mathrm{L}}_{\mu\nu}}{k^2 - \xi m_A^2c^2}\right],$$

and combining the two fractions over the common denominator gives Equation (104.54), since $-(k^2 - \xi m_A^2c^2) + \xi(k^2 - m_A^2c^2) = k^2(\xi - 1)$. As $\xi \to \infty$ the bracket tends to $\eta_{\mu\nu} - k_\mu k_\nu/m_A^2c^2$, which is Equation (101.77): the unitary gauge is the infinite-ξ member of the family, and χ becomes infinitely heavy and drops out, which is why it was absent there.

(iv) The Faddeev–Popov determinant is that of $\delta G/\delta\Lambda$ for G the bracket in Equation (104.53). Under a gauge transformation of parameter Λ, $\delta A_\mu = \partial_\mu\Lambda$ and, from $\phi \mapsto \mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\phi$ to first order, $\delta\chi = -q\Lambda(v+h)/\hbar$. Hence

$$\frac{\delta G}{\delta\Lambda} = \Box + \xi\left(\frac{m_A c}{\hbar}\right)^2\left(1 + \frac{h}{v}\right),$$

so the ghost has mass $\sqrt{\xi}\,m_A$ and a coupling to h. In the abelian *unbroken* theory this operator is field-independent and the ghosts may be discarded (Corollary 106.53); here they may not. □

Remark 104.46 (What the two limits buy, and why both are needed). Theorem 104.45 is the technical heart of the renormalizability proof. At $\xi \to \infty$ the field content is manifestly physical — one massive vector with three polarizations, one scalar, no ghosts, no Goldstone — so unitarity is evident and power counting fails (Theorem 104.4(ii)). At finite ξ every propagator falls as k^{-2}, so power counting works exactly as in Theorem 100.17 and the theory is renormalizable by inspection, but the spectrum now contains a Goldstone of mass $\sqrt{\xi} m_A$ and a ghost of the same mass, neither of which is a particle. Since ξ is an arbitrary parameter, no observable can depend on it, and Theorem 106.61 proves that none does — so the two descriptions compute the same numbers, and the theory is unitary *and* renormalizable although no single gauge makes both obvious. The masses coinciding is not a coincidence: the Goldstone and the ghost are two members of a BRST quartet (Proposition 106.60) and cancel each other in every physical cut.

Proposition 104.47 (The Goldstone-boson equivalence theorem)**.** *Let $\mathcal{M}^{\mu_1 \cdots \mu_n}$ be an amplitude with n external gauge bosons of mass M and momenta k_i, and let $\mathcal{M}_\chi$ be the corresponding amplitude with each gauge boson replaced by its would-be Goldstone boson. Then at energies $E \gg Mc^2$,*

$$\mathcal{M}\left(W_L^{(1)} \dots W_L^{(n)}\right) = (\mathrm{i})^n \, \mathcal{M}_\chi + \mathcal{O}\left(\frac{Mc^2}{E}\right) . \tag{104.55}$$

Rests on Theorem 104.45 and Equations (104.6) and (106.76).

Derivation. *Derives Proposition 104.47.*

The Slavnov–Taylor identity Equation (106.76), applied to the broken theory in an R_ξ gauge, says that the BRST variation of any correlator vanishes. Applied to a Green function with one gauge leg and the ghost of that leg, it gives the relation between the longitudinal part of the gauge amplitude and the Goldstone amplitude,

$$k_\mu \mathcal{M}^\mu(k) = \mathrm{i} M c \, \mathcal{M}_\chi(k) ,$$

which is the broken-theory counterpart of the transversality $k_\mu \mathcal{M}^\mu = 0$ of Equation (100.18) — the right-hand side is zero only when $M = 0$. Now use Equation (104.6): the longitudinal polarization is $\varepsilon_L^\mu = k^\mu / Mc + v^\mu$ with v^μ of order $Mc/|\boldsymbol{k}|$ and bounded. Contracting,

$$\mathcal{M}(W_L) = \varepsilon_{L\mu} \mathcal{M}^\mu = \frac{k_\mu \mathcal{M}^\mu}{Mc} + v_\mu \mathcal{M}^\mu = \mathrm{i} \mathcal{M}_\chi + \mathcal{O}\left(\frac{Mc^2}{E}\right) ,$$

and iterating over the n legs gives Equation (104.55). (The identity holds with $\mathcal{M}_\chi$ computed in the same R_ξ gauge, and the corrections are uniform in the scattering angle away from the forward region.) □

Remark 104.48 (What the equivalence theorem resolves). An amplitude with longitudinal gauge bosons appeared, in the unitary-gauge computation, to grow like $(E/Mc^2)^n$ because each polarization vector grows. Proposition 104.47 says that the whole amplitude equals a *scalar* amplitude with no growing polarization vectors at all. The individual diagrams therefore must cancel among themselves down to the size of the scalar amplitude, and that cancellation is not optional — it is forced by the Slavnov– Taylor identity, which is forced by the gauge symmetry. It will be used twice: in Theorem 104.51 to bound the scalar mass, and in Phenomenon 104.57 to read the non-abelian group structure off a measured cross-section.

The identities that organize the cancellations to all orders are the non-abelian generalization of the Ward–Takahashi identity of Theorem 100.48, found by Slavnov [1484] and Taylor [1458], and their modern algebraic form is the BRST master equation Equation (106.76) of Section 106.4.2.

Remark 104.49 (A theorem, not a datum, made the model orthodox). No new measurement stood between the general indifference to [1362] in 1968 and its status as the consensus theory in 1972. What changed was 't Hooft's proof [1448, 1463]. This is worth recording plainly in a book organized around evidence: the criterion applied here was internal consistency, and it was applied *before* the decisive experiments — neutral currents in 1973, the bosons in 1983 — which the theory then survived. Consistency is not evidence, and the chapter's evidential claims all rest on Section 104.6; but a theory that cannot be computed cannot be tested, and renormalizability is what made the tests possible.

104.5.2 Dimensional regularization and gauge cancellations

The regulator used throughout is the one 't Hooft and Veltman introduced for this problem [1463] and which Definition 100.30 states: continue the loop integral to $d = 4-2\varepsilon$ dimensions. Its virtue over a momentum cutoff is precisely that it preserves gauge invariance — a cutoff $|\ell| < \Lambda$ is not invariant under a gauge transformation that shifts the loop momentum, and the resulting non-invariant counterterms would include exactly the photon mass term that Theorem 104.8 forbids. The systematic treatment of the scheme dependence belongs to Chapter 108.

Remark 104.50 (One honest blemish: γ^5 in d dimensions). Dimensional regularization has a defect specific to a chiral theory. γ^5 is defined by $\gamma^5 = \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3$, a four-dimensional construction, and there is no continuation to $d \neq 4$ that is simultaneously anticommuting with all γ^μ and consistent with the cyclicity of the trace. The standard resolution splits the Dirac algebra into four-dimensional and $(d-4)$-dimensional parts, at the cost of extra non-invariant counterterms that must be fixed by imposing the Slavnov–Taylor identities by hand. This is not a hidden inconsistency — the identities can be restored order by order — but it is real work, and it is the reason the anomaly of Section 104.5.3 shows up in this scheme as an unavoidable obstruction rather than as an artefact one could regulate away.

The most consequential gauge cancellation in the theory is the one that made the scalar mandatory before it was found.

Theorem 104.51 (Longitudinal W scattering bounds the scalar mass)**.** *Consider $W_L^+W_L^- \to W_L^+W_L^-$ at $\sqrt{s}\,c \gg M_W c^2$. Without the scalar, the $J=0$ partial wave grows as s and violates unitarity. With the scalar of mass m_h included, the growth cancels and the surviving constant amplitude is*

$$a_0 = -\frac{G_F\,(m_h c^2)^2}{4\sqrt{2}\pi\,(\hbar c)^3}\,, \tag{104.56}$$

so that the requirement $|\Re a_0| \le \tfrac{1}{2}$ gives

$$m_h c^2 \le \left(\frac{2\sqrt{2}\pi\,(\hbar c)^3}{G_F}\right)^{1/2} = 873\,\mathrm{GeV}\,, \tag{104.57}$$

improved to 713 GeV *when the coupled channels W_LW_L, Z_LZ_L, hh, hZ_L are diagonalized [1601] (the result was announced in a companion letter the same year). The measured $m_h c^2 = (125.20 \pm 0.11)\,\mathrm{GeV}$ satisfies it with room to spare.*

Rests on Proposition 104.47 and Equations (101.89) and (104.37).

Proof. *Derives Theorem 104.51.*

By Proposition 104.47 the high-energy amplitude equals that for the eaten Goldstone bosons, $\chi^+\chi^- \to \chi^+\chi^-$, computed from the scalar sector alone: the gauge couplings drop out of the leading behaviour entirely. Write the doublet as

$$\Phi = \begin{pmatrix} \chi^+ \\ \frac{1}{\sqrt{2}}\,(v + h + \mathrm{i}\chi^0) \end{pmatrix}\,, \qquad \Phi^\dagger\Phi = \left|\chi^+\right|^2 + \tfrac{1}{2}\left[(v+h)^2 + \left(\chi^0\right)^2\right]\,,$$

and expand $-\lambda(\Phi^\dagger\Phi)^2$. Two vertices matter: the contact term $-\lambda\left|\chi^+\right|^4$ and the coupling $-2\lambda v\, h\left|\chi^+\right|^2$, in which $2\lambda v = \left(m_h c/\hbar\right)^2/v$ by Equation (104.37). Summing the contact term with h exchange in the s and t channels, and using the Mandelstam conventions of Notation 100.1 in which s and t are squared momenta,

$$\mathcal{M} = -\hbar^2\frac{\left(m_h c^2\right)^2}{E_v^2}\left[\frac{s}{s-m_h^2c^2}+\frac{t}{t-m_h^2c^2}\right],$$

the factor $\hbar^2$ being the dimension an amplitude carries in the normalization of Proposition 100.24, and $(m_hc^2)^2/E_v^2 = 2\hat{\lambda}$ by Equation (104.37). The two propagator poles have absorbed the contact term: each channel contributes the vertex $2\lambda v$ twice with a propagator between them, and the leftover constant is exactly the quartic. For $s \gg m_h^2c^2$, and away from the forward region where $|t|$ is also large, both brackets tend to 1 and the amplitude tends to the *constant* $-2\hbar^2(m_hc^2)^2/E_v^2$. That it is constant rather than growing is the cancellation: had the h been absent, only the contact term would survive and the bracket would be $-s/m_h^2c^2$ in disguise — growth without limit.

The $J=0$ projection is $a_0 = \frac{1}{32\pi\hbar^2}\int_{-1}^{1}\mathcal{M}\,\mathrm{d}(\cos\theta)$, the normalization being the one in which $\sigma_J = 16\pi\hbar^2(2J+1)\left|a_J\right|^2/s$ reproduces Equation (101.89). For a constant amplitude the integral is $2\mathcal{M}$, so $a_0 = \mathcal{M}/16\pi\hbar^2$, that is

$$a_0 = -\frac{\left(m_h c^2\right)^2}{8\pi E_v^2},$$

manifestly dimensionless; and $E_v^2 = (\hbar c)^3/\sqrt{2}G_F$ turns it into Equation (104.56). Imposing $|a_0| \le \frac{1}{2}$,

$$\left(m_h c^2\right)^2 \le \frac{2\sqrt{2}\pi\left(\hbar c\right)^3}{G_F} = 2\sqrt{2}\pi \times 85\,735\,\mathrm{GeV}^2 = 7.618\times10^5\,\mathrm{GeV}^2,$$

whose square root is 873 GeV. The coupled-channel analysis replaces the coefficient 1 by 3/2 in the largest eigenvalue, giving $m_hc^2 \le \left(4\sqrt{2}\pi(\hbar c)^3/3G_F\right)^{1/2} = 713\,\mathrm{GeV}$. □

Remark 104.52 (What a failure of the bound would have meant). Theorem 104.51 is a prediction of a very unusual kind: it does not say what exists, it says that *something* must, below about a TeV, or the theory is inconsistent there. Either a scalar with the couplings the mechanism requires, or a strongly interacting sector in which W_LW_L scattering saturates the bound and the perturbative description fails. The two alternatives are distinguishable, and the LHC was built with enough energy to settle it. It found the first (Chapter 113), and the couplings match (Phenomenon 104.62). Had it found the second, this chapter would look very different; had it found neither, the Standard Model would have been falsified as a description of energies above a TeV. That is the sense in which Equation (104.57) was a genuine risk.

104.5.3 Anomaly cancellation

A symmetry of the classical action need not survive quantization. The axial current of a massless Dirac field is conserved classically and not quantum mechanically, by the triangle anomaly of Adler [1388] and of Bell and Jackiw [1391], derived in Theorem 106.87 and measured in the π^0 lifetime (Section 106.7.2). In a *global* symmetry that is a prediction. In a *gauge* symmetry it is fatal, because the Slavnov–Taylor identities on which Theorem 104.45 and Remark 104.46 depend then fail: the unphysical polarizations stop decoupling, the S-matrix stops being unitary and the divergences stop organizing themselves. That is Theorem 106.94.

The electroweak theory is chiral by construction — T^a acts on ψ_L and not on ψ_R — so it is exactly the case in which the anomaly can occur. The condition for it not to is Theorem 106.95: the symmetrized trace $\mathsf{A}^{abc} = \mathrm{tr}(\{T^a, T^b\} T^c)_L - \mathrm{tr}(\{T^a, T^b\} T^c)_R$ must vanish for every triple of gauge generators. Bouchiat, Iliopoulos and Meyer [1467] checked it for this theory and found that it holds — but only for a complete generation, and only with three colours.

Proposition 104.53 (One generation is anomaly free)**.** *With the hypercharges of Table 104.1, in the convention $Q = T^3 + Y/2$, the four independent conditions are satisfied:*

$$[\mathrm{SU}(2)_L]^2\,\mathrm{U}(1)_Y: \quad N_c \cdot \tfrac{1}{3} + 1 \cdot (-1) = \frac{N_c - 3}{3} = 0\,, \tag{104.58}$$

$$[\mathrm{SU}(3)_c]^2\,\mathrm{U}(1)_Y: \quad 2 \cdot \tfrac{1}{3} - \tfrac{4}{3} - \left(-\tfrac{2}{3}\right) = 0\,, \tag{104.59}$$

$$[\mathrm{U}(1)_Y]^3: \quad \left[6\left(\tfrac{1}{3}\right)^3 + 2\,(-1)^3\right] - \left[3\left(\tfrac{4}{3}\right)^3 + 3\left(-\tfrac{2}{3}\right)^3 + (-2)^3\right] = 0\,, \tag{104.60}$$

$$\textit{gravitational}\ \mathrm{U}(1)_Y: \quad \left[6 \cdot \tfrac{1}{3} + 2\,(-1)\right] - \left[3 \cdot \tfrac{4}{3} + 3\left(-\tfrac{2}{3}\right) - 2\right] = 0\,, \tag{104.61}$$

where the multiplicities count colour and SU(2) *components and the first bracket in each of the last two is the left-chiral contribution, the second the right-chiral one.*

Rests on Table 104.1 and Theorem 106.95.

Proof. *Derives Proposition 104.53.*

$[\mathrm{SU}(2)]^3$ vanishes identically because the symmetrized trace of three SU(2) generators is zero in every representation, SU(2) having only real and pseudoreal representations; and $[\mathrm{SU}(3)]^3$ vanishes because the quarks appear in triplet and antitriplet with respect to colour once the two chiralities are combined. What remains is the four displayed conditions.

For Equation (104.58): only $\mathrm{SU}(2)_L$ doublets contribute, each with $\mathrm{tr}(\{T^a, T^b\}) = \delta^{ab}$, so the condition is $\sum_{\text{doublets}} Y = 0$ with colour multiplicity; the quark doublet contributes $N_c \times \frac{1}{3}$ and the lepton doublet -1. This is where the number of colours enters: the condition fixes $N_c = 3$.

For Equation (104.59): only coloured fields contribute, with $\mathrm{tr}(\{T^a, T^b\}) = \delta^{ab}$ for the fundamental, so the condition is $\sum_L Y - \sum_R Y = 0$ over the quark SU(2) components: the left doublet supplies $2 \times \frac{1}{3}$ and the right singlets $\frac{4}{3}$ and $-\frac{2}{3}$.

For Equation (104.60): the trace is $\sum Y^3$ with all multiplicities. Left: 6 quark states (three colours, two isospin) of $Y = \frac{1}{3}$, giving $6/27$, plus 2 lepton states of $Y = -1$, giving -2; total $-\frac{16}{9}$. Right: $3(\frac{4}{3})^3 = \frac{64}{9}$, $3(-\frac{2}{3})^3 = -\frac{8}{9}$, and $(-2)^3 = -8$; total $\frac{56}{9} - 8 = -\frac{16}{9}$. Equal.

For Equation (104.61), the triangle with two insertions of the energy–momentum tensor requires $\sum Y = 0$ separately for each chirality summed with sign: left gives $2 - 2 = 0$, right gives $4 - 2 - 2 = 0$. This is a statement about the electroweak currents in a fixed classical background metric and not a claim about quantum gravity.

In every case the quarks alone fail and the leptons alone fail; only the complete generation cancels. Proposition 106.96 does the same arithmetic in the halved-hypercharge convention of Remark 104.11, with identical conclusions. □

Remark 104.54 (The condition read backwards predicted the top). Proposition 104.53 is usually presented as a consistency check. Read the other way it is a prediction. When the b quark was found in 1977 [1597], the third generation had a charged lepton, a neutrino and a charge $-\frac{1}{3}$ quark, and Equation (104.58) then reads $N_c Y_Q + Y_L \neq 0$ unless the b has an $\mathrm{SU}(2)_L$ partner of charge $+\frac{2}{3}$. The theory therefore required a top quark — not as an aesthetic preference but because without it the theory does not exist — and it was found eighteen years later [1891, 1892] with the mass the electroweak fit had by then indicated. The same argument forbids a fourth generation consisting of leptons alone, and Phenomenon 101.76 independently excludes a fourth light neutrino.

104.6 Experimental confirmation

104.6.1 Weak neutral currents

The neutral current is the model's first genuinely new prediction. Every weak process known in 1970 changed the charge of the fermions involved; Proposition 101.67 says a fourth current exists which changes neither charge nor flavour, with a strength fixed relative to the charged current by the single angle θ_W. A theory of this kind is falsifiable in the strongest sense: it predicted a class of events nobody had seen.

Gargamelle was a heavy-liquid bubble chamber of about $12\,\mathrm{m}^3$, filled with freon, exposed at CERN to a neutrino beam. The signature of a neutral-current event is what is *not* there: hadrons produced with no outgoing muon. Two publications in 1973 carried the discovery. The leptonic one [1503] rests on a single event, $\bar{\nu}_\mu e^- \to \bar{\nu}_\mu e^-$ — one electron recoiling with nothing else in the chamber — for which the expected background was a small fraction of an event; the hadronic one [1502] reports a sample of muonless hadronic events at a rate of about 0.2 of the charged-current rate. The credibility of the second rested on an argument about backgrounds rather than on the events themselves: neutrons produced by neutrino interactions in the surrounding material could enter the chamber and produce hadrons with no muon. The collaboration showed that such events would cluster near the chamber walls and fall off with depth, whereas the observed muonless events were distributed through the fiducial volume exactly as the charged-current events were. Phenomenon 101.72 states the result and Proposition 101.71 computes the cross-sections it is compared with.

Confirmation in a different process came from SLAC experiment E122 [1625]: longitudinally polarized electrons scattered from deuterium show an asymmetry between the two helicities, of order 10^{-4}, which arises only from the interference of the electromagnetic amplitude with a parity-violating neutral-current one. The measurement gave $\sin^2\theta_W = 0.20 \pm 0.03$ [1625], refined to 0.224 ± 0.020 by the 1979 scan over beam energies [1643], and, crucially, excluded the alternative models then in play — several of which reproduced the Gargamelle neutrino data with a parity-conserving neutral current and were killed by this single number. The apparatus belongs to Chapter 110.

Phenomenon 104.55 (One mixing angle describes every neutral-current measurement)**.** The weak neutral current is not the electromagnetic one. It couples to neutrinos, which carry no electric charge [1502, 1503], and it violates parity, as the asymmetry between the scattering of right- and left-handed polarized electrons from deuterium first showed [1625]. One parameter, the weak mixing angle, fixes its strength relative to electromagnetism in every process, and the value extracted from neutrino scattering, from polarized electron scattering, from atomic parity violation and from the asymmetries on the Z resonance is the same,

$$\sin^2\theta_W \approx 0.231 \tag{104.62}$$

in the conventions of [2427], over momentum transfers spanning several orders of magnitude. That one number should serve for all of them is the empirical content of unification.

Rests on Equations (104.35) *and* (104.40) *and Proposition 101.67.*

Derivation. *Derives Phenomenon 104.55.*

Three steps: the rotation that produces the current, the couplings it gives each fermion, and the reason a scheme must be attached to any number quoted.

The rotation. The neutral part of Equation (104.11) is $\hbar^{-1}\left(gT^3W_\mu^3 + g'\frac{Y}{2}B_\mu\right)$. Substituting Equation (104.35) and demanding that the massless combination A_μ couple to Q and to nothing else forces $\hat{e} = \hat{g}\sin\theta_W = \hat{g}'\cos\theta_W$, which is Equation (104.40); the orthogonal combination then couples to $T^3 - \sin^2\theta_W Q$ with strength $\hat{e}/\sin\theta_W\cos\theta_W$. This is Proposition 101.67, proved there in full.

The couplings. Projecting onto chiralities gives Equation (101.75): for each fermion

$$g_V = T^3 - 2Q\sin^2\theta_W\,, \qquad g_A = T^3\,,$$

with T^3 that of the *left-chiral* component. The whole predictive content follows: the neutrino has $Q = 0$, so $g_V = g_A = \frac{1}{2}$ and its neutral-current coupling is independent of θ_W — which is why the ratio of neutral- to charged-current neutrino cross-sections measures the angle cleanly. The electron has $g_A = -\frac{1}{2}$ and $g_V = -\frac{1}{2} + 2\sin^2\theta_W$, which is small and changes sign near $\sin^2\theta_W = \frac{1}{4}$: a parity-violating electron asymmetry is therefore a sensitive, and sign-carrying, probe. The Z-pole asymmetries measure the combination $\mathcal{A}_f = 2g_V g_A/(g_V^2 + g_A^2)$, which for the electron is a steep function of $\sin^2\theta_W$ for exactly the same reason. Four independent kinds of measurement thus depend on one number in four different functional ways, and that is what makes their agreement a test rather than a fit.

The scheme. Beyond tree level, θ_W is not a parameter of the Lagrangian but a renormalized quantity, and different definitions differ by finite amounts of order α/π — the three standard ones are listed in Remark 104.34 and differ by about one per cent, which is hundreds of times the experimental errors. Worse, in the $\overline{\text{MS}}$ scheme the angle *runs*: the same definition gives about 0.2313 at $\mu = M_Z c^2$ and about 0.2387 at $\mu \to 0$, the difference being real physics — vacuum polarization by charged particles, Section 100.7 — and not a change of convention. Table 104.3 lists the determinations across momentum transfer, and the content of Equation (104.62) is that one running curve passes through all of them. A number quoted without its scheme and its scale is not a measurement. □

Table 104.3: Determinations of the weak mixing angle across six orders of magnitude in momentum transfer, as quoted by the experiments. The low-Q entries are expressed in the $\overline{\text{MS}}$ scheme, in which the angle runs, so they are not expected to equal the Z-pole value: the prediction is the *difference*, about 0.007 between $Q = 0$ and $Q = M_Z c$, and it is what the table tests. Uncertainties are as published and are combined statistical and systematic.

Measurement	Momentum transfer	$\sin^2\theta_W$	Reference
Cesium atomic parity violation	≈ 0	≈ 0.24	[1946]
Polarized Møller scattering, SLAC E158	$Q^2 = 0.026\,\text{GeV}^2/c^2$	0.2397 ± 0.0013	[2069]
Polarized *ep* scattering, Qweak	$Q = 0.157\,\text{GeV}/c$	0.2383 ± 0.0011	[2301]
Polarized *ed* scattering, SLAC E122	$Q^2 \approx 1.5\,\text{GeV}^2/c^2$	$\approx 0.20 \pm 0.03$	[1625]
Z-pole asymmetries, LEP and SLD	$Q = M_Z c$	$0.231\,53 \pm 0.000\,16$	[2104]

104.6.2 Discovery of the W and Z bosons

By 1980 the masses were predicted. Putting $\sin^2\theta_W = 0.23$ from Table 104.3 into Equation (104.41),

$$M_W c^2 = \frac{37.28\,\text{GeV}}{\sin\theta_W} \approx 78\,\text{GeV}\,, \qquad M_Z c^2 = \frac{M_W c^2}{\cos\theta_W} \approx 89\,\text{GeV}\,, \tag{104.63}$$

at tree level, and about 80 GeV and 91 GeV once Δr of Remark 104.33 is included — numbers derived from a neutrino cross-section and a muon lifetime, with no collider involved.

Producing them required a beam of a few hundred GeV in the centre of mass, and CERN got one by converting the SPS into a proton–antiproton collider. The enabling invention was stochastic cooling [1760]: a pickup senses the deviation of a sample of the circulating beam from the mean orbit, and a kicker one straight section downstream applies a correcting impulse, reducing the phase-space volume of the antiproton beam by many orders of magnitude over hours. Without it there is no way to accumulate enough antiprotons.

The signature is a lepton of high transverse momentum with nothing balancing it. UA1 [1710] and UA2 [1714] reported in 1983 events with an isolated electron of $p_T \gtrsim 20\,\mathrm{GeV}/c$ and large missing transverse energy — the neutrino — from $W \to e\nu$. The mass is read from the transverse-mass distribution rather than from the momenta directly, because the neutrino's longitudinal momentum is unmeasured; the observable is Equation (101.84) and the reason it works is a kinematic singularity.

Proposition 104.56 (The Jacobian peak)**.** *Let a boson of mass M at rest decay to two massless particles. The transverse momentum of either with respect to any axis is $p_T = \tfrac{1}{2}Mc\sin\theta^*$, and therefore the distribution in p_T has an integrable singularity at the endpoint,*

$$\frac{\mathrm{d}N}{\mathrm{d}p_T} = \frac{\mathrm{d}N}{\mathrm{d}(\cos\theta^*)}\,\frac{p_T}{\tfrac{1}{2}Mc\sqrt{\left(\tfrac{1}{2}Mc\right)^2 - p_T^2}}\,, \tag{104.64}$$

which diverges as $p_T \to \tfrac{1}{2}Mc$ however smooth the angular distribution is. The position of the edge measures M.

Rests on Equation (101.1).

Proof. *Derives Proposition 104.56.*

In the rest frame each decay product has momentum $\tfrac{1}{2}Mc$, so $p_T = \tfrac{1}{2}Mc\sin\theta^*$ with θ^* the polar angle to the chosen axis. Then $\mathrm{d}p_T/\mathrm{d}(\cos\theta^*) = -\tfrac{1}{2}Mc\,\cos\theta^*/\sin\theta^*$, and inverting and substituting $\sin\theta^* = 2p_T/Mc$, $\cos\theta^* = \sqrt{1-(2p_T/Mc)^2}$ gives Equation (104.64). The Jacobian, not the dynamics, produces the peak; this is why the method is robust against the unknown production mechanism, and why smearing it by the boson's transverse motion and by the detector resolution is the dominant systematic. Applying the same construction to the pair, with the neutrino's transverse momentum inferred from the missing energy, gives the transverse mass Equation (101.84), whose endpoint is Mc^2 itself and which is less sensitive to the boson's motion. □

Later in 1983 both experiments found the neutral partner in the cleaner channel $Z \to e^+e^-$ and $\mu^+\mu^-$, where both leptons are measured and the invariant mass is direct [1711, 1713]: a handful of events at an invariant mass near $95\,\mathrm{GeV}/c^2$, on essentially no background. Phenomenon 101.73 records the measurement and Equation (101.83) the comparison. The masses agreed with Equation (104.63) at the accuracy then available, which is the point: two numbers measured in the 1970s in low-energy experiments predicted where two new particles would be found, and they were found there.

104.6.3 LEP, SLD and the Z resonance

LEP turned the Z from a discovery into an instrument. Four experiments — ALEPH, DELPHI, L3 and OPAL — recorded some seventeen million Z decays between 1989 and 1995, and SLD added a smaller sample with a longitudinally polarized electron beam, which gives access to asymmetries the unpolarized machine cannot reach. The combined analysis [2104] is the most precise test the electroweak theory has had.

Three classes of result matter here.

The lineshape. Scanning the beam energy across the resonance determines M_Z and the total width Γ_Z, and the peak cross-section determines the product of partial widths. The

masses quoted throughout this chapter come from this analysis. The energy calibration was itself a small physics programme: the beam energy was measured by resonant depolarization to a precision of about 1 MeV, at which level the circumference of the ring responds to the tides raised by the Moon and the Sun, to the water level in Lake Geneva, and to leakage currents from a nearby railway line. All three had to be modelled to reach the quoted M_Z.

The invisible width. Subtracting the visible partial widths from the total leaves the width into channels the detector cannot see, and dividing by the calculated width per neutrino species counts them: $N_\nu = 2.984 \pm 0.008$, which is Phenomenon 101.76. That a resonance curve counts the generations of matter is among the most economical measurements in physics.

The asymmetries. Forward–backward asymmetries in $e^+e^- \to f\bar{f}$ and, at SLD, the left–right asymmetry with polarized beams, measure $\sin^2\theta^\ell_{\text{eff}}$ to better than one part in a thousand (Table 104.3). Proposition 101.75 derives the partial widths that these are compared with.

LEP's second phase ran above the W-pair threshold and measured something no fixed-target experiment can: that the gauge bosons couple to each other.

Phenomenon 104.57 (The gauge bosons couple to one another)**.** The cross-section for $e^+e^- \to W^+W^-$, measured at LEP2 from threshold up to 209 GeV, rises to a maximum and then falls. Three amplitudes contribute — neutrino exchange in the t channel, and annihilation through a photon and through a Z — and each grows with energy on its own; the data are reproduced only when all three are present and interfere destructively, the neutrino-exchange term alone exceeding the measurement by a large factor at the highest energies [2223]. The two annihilation diagrams exist only if three gauge bosons meet at a vertex.

Rests on Propositions 104.7 and 104.47 and Equation (104.40)*.*

Derivation. *Derives Phenomenon 104.57.*

The claim to be established is that the three amplitudes must cancel their leading high-energy growth, and that the cancellation fixes the triple-gauge coupling to the Yang–Mills value.

Take the final Ws longitudinal, which is where the growth lives by Equation (104.6), and apply Proposition 104.47: at $\sqrt{s}\,c \gg M_W c^2$ the amplitude equals that for producing the two eaten Goldstone bosons, $e^+e^- \to \chi^+\chi^-$. Now enumerate the diagrams for *that* process.

- The s-channel photon and Z survive, with the couplings of a charged scalar read off Equation (104.11): the $\chi^\pm$ sits in the doublet with $T^3 = \pm\frac{1}{2}$ and $Q = \pm 1$, so it couples to the photon with charge $\hat{e}$ and to the Z with $\hat{e}\,(1 - 2\sin^2\theta_W)\,/2\sin\theta_W\cos\theta_W$, by Equation (101.74) evaluated at those quantum numbers.
- The t-channel neutrino exchange survives only through the *Yukawa* coupling of $\chi^\pm$ to $e\nu$, which by Theorem 104.38 is $\hat{y}_e/\sqrt{2}$, of order 2×10^{-6} from Table 104.2, and is therefore negligible.

The high-energy amplitude is thus the production of a pair of charged scalars through two s-channel vector exchanges. Such an amplitude does *not* grow: an s-channel propagator $1/(s - M^2c^2)$ falls as $1/s$ while the vertices supply at most one power of momentum each, so the amplitude tends to a constant in the $J = 1$ partial wave and the cross-section falls as $1/s$, which is the observed shape.

The consequence is the cancellation. Computed instead in unitary gauge, each of the three diagrams carries two longitudinal polarization vectors and so grows as $s/M_W^2c^2$; but their sum must equal the non-growing scalar amplitude just obtained. Hence the leading terms cancel identically. The cancellation uses the couplings in *exactly* the combinations Equation (104.40) supplies — the γW^+W^- vertex equal to $\hat{e}$ and the ZW^+W^- vertex

equal to $\hat{g}\cos\theta_W$, both of which are forced by Proposition 104.7 because the $WW\gamma$ and WWZ couplings come from the same ε^{abc} in Equation (104.8). Change either coupling by a fraction δ and a term growing as $\delta s/M_W^2c^2$ survives; at $\sqrt{s}\,c = 200\,\mathrm{GeV}$ that factor is $s/M_W^2c^2 \approx 6$, so a few-per-cent change in the vertex is visible in the cross-section. That is why the LEP2 measurement of a cross-section is a measurement of the group structure: the agreement in [2223] bounds the anomalous triple-gauge couplings at the per-cent level, and the neutrino-exchange diagram alone — with the two annihilation diagrams removed — overshoots the data at the highest energies by a large factor, exactly as an uncancelled $s/M_W^2c^2$ requires. □

104.6.4 Precision fits and the W mass

Remark 104.33 showed that four measured quantities are described by three parameters, so one relation is predicted. The electroweak fit generalizes this: several dozen measurements — M_Z, Γ_Z, the partial widths, the asymmetries, M_W, m_t, α, α_s, G_F — are described by a handful of parameters, so the system is heavily over-constrained and its consistency is a test at every point [2228].

The historically decisive property is that the one-loop corrections depend on particles too heavy to produce. $\Delta\rho$ grows as m_t^2 by Equation (104.52), so the LEP data constrained the top mass before the Tevatron produced it, and the indirect value agreed with the direct measurement of [1891, 1892]. The scalar enters only logarithmically — Veltman's screening theorem again — so the constraint on m_h was much weaker, but it was real: the fit before 2012 preferred a scalar mass of order 100 GeV and excluded values above a few hundred GeV [2228], and the particle was found at $125\,\mathrm{GeV}/c^2$ inside that range. Two particles were thus located by loop effects before either was seen.

One entry does not fit, and it is stated here as what it is.

Remark 104.58 (The W-mass tension, unresolved). The CDF collaboration's final analysis of its Tevatron data [2395] reports $M_Wc^2 = (80\,433.5 \pm 9.4)\,\mathrm{MeV}$, an uncertainty of 9.4 MeV — the most precise single measurement of the quantity. It disagrees with the world average (80.369 ± 0.013) GeV [2427] by about four standard deviations, and with the indirect prediction of the electroweak fit, $M_Wc^2 = (80\,357 \pm 6)\,\mathrm{MeV}$ [2427], by about seven: the difference of 76.5 MeV is 6.9 times the combined uncertainty of 11.2 MeV obtained by adding the two errors in quadrature. It also disagrees with the LHC measurements, which are individually less precise but mutually consistent and consistent with the fit.

This is recorded as an unresolved experimental tension and not as evidence for anything. Two measurements of the same quantity that disagree by seven standard deviations mean that at least one uncertainty is underestimated; which one is not known. The relevant systematics — parton distributions, the modelling of the W transverse-momentum spectrum, the momentum-scale calibration — are of the same order as the quoted total error, and no independent reanalysis has reproduced the CDF value. Until it is resolved, the honest statement is that the W mass is known to about 13 MeV with one outlier at four sigma, and no shift in any Standard Model parameter is warranted.

104.7 The Higgs boson

104.7.1 Search and discovery

Everything about the scalar is predicted except its mass. Given m_h, Equation (104.38) fixes λ, and Theorem 104.38(iii) and Equation (104.23) fix every coupling, hence every production cross-section and every branching fraction. The search was therefore a scan in one variable with everything else nailed down — which is why a negative result at any mass is a genuine exclusion, and why the positive result is a test rather than a fit.

LEP searched in $e^+e^- \to Zh$ up to a centre-of-mass energy of 209 GeV and set the limit

$$m_h c^2 > 114.4\,\mathrm{GeV} \quad \text{at 95 per cent confidence} \tag{104.65}$$

[2045], which is about as far as that machine could reach. The Tevatron and then the LHC took the range upward. In July 2012 ATLAS [2186] and CMS [2192] independently reported a resonance near $125\,\mathrm{GeV}/c^2$, each with a local significance of about five standard deviations, in the two channels with the cleanest signature: $h \to \gamma\gamma$, where the mass is reconstructed from two photons, and $h \to ZZ^* \to 4\ell$, where it is reconstructed from four charged leptons. Their first joint mass combination [2242] gave $(125.09 \pm 0.24)\,\mathrm{GeV}$, and the present world average is $(125.20 \pm 0.11)\,\mathrm{GeV}$ [2427]. The apparatus, the trigger, the look-elsewhere effect and the statistical machinery belong to Chapter 113; the statistical foundations belong to Chapter 11.

Phenomenon 104.59 (A neutral scalar boson exists)**.** A narrow resonance is produced in proton–proton collisions and decays to two photons and to four charged leptons. It was observed independently by ATLAS [2186] and CMS [2192] in 2012, at a mass

$$m_H c^2 = (125.09 \pm 0.24)\,\mathrm{GeV} \tag{104.66}$$

in their first combination [2242], now $(125.20 \pm 0.11)\,\mathrm{GeV}$ [2427]. That it decays to two photons excludes spin one [917, 975], and the angular distributions of the four-lepton final state select $J^P = 0^+$ over the remaining alternatives. Direct searches at LEP had already excluded a Standard Model scalar below $114.4\,\mathrm{GeV}/c^2$ [2045]. The full discovery account belongs to Chapter 113.

Rests on Theorem 104.38, Table 104.2, and Equation (101.86).

Derivation of the decay pattern. Derives Phenomenon 104.59.

Once m_h is given, the tree-level partial widths follow from the couplings of Theorem 104.38(iii) with no freedom left. For a fermion pair,

$$\hbar\Gamma\left(h \to f\bar{f}\right) = \frac{N_c \hat{y}_f^2}{16\pi}\, m_h c^2 \left(1 - \frac{4m_f^2}{m_h^2}\right)^{3/2}, \tag{104.67}$$

N_c being 3 for quarks and 1 for leptons, and $\hat{y}_f = \sqrt{2} m_f c^2 / E_v$ from Table 104.2. The structure is that of Equation (101.86) — the width of a particle at rest into two massless ones is a coupling squared times the mass over 16π, times a phase-space factor — and the exponent $3/2$ rather than $1/2$ is the signature of a *scalar*: the β^3 arises because a $J = 0$ state decaying to a fermion pair through a scalar coupling produces them in a P wave, contributing β^2 beyond the β of two-body phase space. A pseudoscalar coupling $\mathrm{i}\bar{f}\gamma^5 f$ would give β^1 instead, which is one of the ways the parity of the state is tested.

Two checks with no strong-interaction corrections. For $\tau^+\tau^-$, Table 104.2 gives $\hat{y}_\tau = 0.010\,21$, so $\hbar\Gamma = 1.0416 \times 10^{-4} \times 125.20\,\mathrm{GeV}/16\pi = 0.259\,\mathrm{MeV}$; for $\mu^+\mu^-$, $\hat{y}_\mu = 6.07 \times 10^{-4}$ gives 0.92 keV. Against the total width of 4.1 MeV these are branching fractions of 6.3×10^{-2} and 2.2×10^{-4}, which are the values in Table 104.4.

The two heaviest available channels need a further remark. At $m_h c^2 = 125.20\,\mathrm{GeV}$ both W^+W^- and ZZ are *below* threshold, since $2M_W c^2 = 160.7\,\mathrm{GeV}$; they proceed with one boson off shell, $h \to VV^* \to Vf\bar{f}'$, a three-body decay whose rate is suppressed by an extra factor of order $\hat{g}^2/16\pi^2$ relative to the on-shell formula and which is not obtained by putting a negative number under a square root. That this suppression is only partial — WW^* is still the second largest channel — is because the coupling grows as the squared vector mass, Equation (104.69) below.

The total width predicted by summing all channels is 4.1 MeV, some 3×10^{-5} of the mass; the measured value is 3.7 MeV with an uncertainty of about 1.7 MeV [2427], in agreement. A width this narrow is itself a prediction: it follows from the couplings being small, which follows from most fermions being light. □

Table 104.4: Standard Model branching fractions of a scalar of mass $125\,\mathrm{GeV}/c^2$, to the precision that matters here. The first, third, fourth, fifth and last are tree-level decays governed by Equation (104.67); WW^* and ZZ^* proceed through one off-shell boson; gg, $\gamma\gamma$ and $Z\gamma$ have no tree-level vertex at all and arise from loops of the heaviest charged and coloured particles. Every entry is fixed once the mass is given: there is no free parameter in this table.

Channel	Mechanism	Branching fraction
$b\bar{b}$	tree, Yukawa	0.58
W^+W^{-*}	tree, gauge	0.21
gg	loop (top)	0.082
$\tau^+\tau^-$	tree, Yukawa	0.063
$c\bar{c}$	tree, Yukawa	0.029
ZZ^*	tree, gauge	0.026
$\gamma\gamma$	loop (W, top)	0.0023
$Z\gamma$	loop (W, top)	0.0015
$\mu^+\mu^-$	tree, Yukawa	2.2×10^{-4}

Derivation pending #158

The loop-induced scalar vertices: the amplitudes for $h\to\gamma\gamma$ and $h\to gg$, from the W and top triangles, which have no tree-level counterpart and which supply the discovery channel and the dominant production mechanism (gluon fusion) respectively; together with the resulting production cross-sections at the LHC. These require the full one-loop tensor reduction and belong in Appendix A. Also owed are the three-body off-shell rates $h\to WW^*$ and $h\to ZZ^*$, whose branching fractions the table quotes without derivation. What is derived above is the two-body tree-level fermionic width formula and the tau and muon entries computed from it; every other entry of the branching table quotes the standard values.

104.7.2 Quantum numbers and couplings

Theorem 104.60 (Landau–Yang)**.** *A massive particle of spin 1 cannot decay into two real photons [917, 975]. Hence the observation of $h\to\gamma\gamma$ excludes $J=1$ for the resonance.*

Rests on Theorem 99.46.

Proof. *Derives Theorem 104.60.*

Work in the rest frame of the decaying particle, where the two photons travel along $\pm\hat{\boldsymbol{k}}$ with polarization vectors $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, both transverse: $\boldsymbol{\varepsilon}_i\cdot\hat{\boldsymbol{k}}=0$. The amplitude is a scalar function linear in each $\boldsymbol{\varepsilon}_i$, and for $J=1$ it must be a *vector* $\boldsymbol{A}$ contracted with the decaying particle's polarization. Bose symmetry requires $\boldsymbol{A}$ to be invariant under exchanging the two photons, which means $\boldsymbol{\varepsilon}_1\leftrightarrow\boldsymbol{\varepsilon}_2$ together with $\hat{\boldsymbol{k}}\to-\hat{\boldsymbol{k}}$.

Enumerate the vectors that are linear in each polarization and built from $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\hat{\boldsymbol{k}}$:

$$\begin{aligned}
\boldsymbol{A}_1 &= \boldsymbol{\varepsilon}_1\times\boldsymbol{\varepsilon}_2 \;\longmapsto\; \boldsymbol{\varepsilon}_2\times\boldsymbol{\varepsilon}_1 = -\boldsymbol{A}_1\,,\\
\boldsymbol{A}_2 &= (\boldsymbol{\varepsilon}_1\cdot\boldsymbol{\varepsilon}_2)\,\hat{\boldsymbol{k}} \;\longmapsto\; (\boldsymbol{\varepsilon}_2\cdot\boldsymbol{\varepsilon}_1)\left(-\hat{\boldsymbol{k}}\right) = -\boldsymbol{A}_2\,,\\
\boldsymbol{A}_3 &= \left(\hat{\boldsymbol{k}}\cdot\boldsymbol{\varepsilon}_2\right)\boldsymbol{\varepsilon}_1 - \left(\hat{\boldsymbol{k}}\cdot\boldsymbol{\varepsilon}_1\right)\boldsymbol{\varepsilon}_2 = 0\,,
\end{aligned}\tag{104.68}$$

the last vanishing by transversality. $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are antisymmetric under the exchange and are therefore forbidden for two identical bosons, and no other structure exists. The

one remaining candidate, $\left[\hat{\boldsymbol{k}}\cdot(\boldsymbol{\varepsilon}_1\times\boldsymbol{\varepsilon}_2)\right]\hat{\boldsymbol{k}}$, is not independent: transversality makes $\boldsymbol{\varepsilon}_1\times\boldsymbol{\varepsilon}_2$ parallel to $\hat{\boldsymbol{k}}$, so it *is* $\boldsymbol{A}_1$, and it is antisymmetric for the same reason. Anything else contains either $\hat{\boldsymbol{k}}\cdot\boldsymbol{\varepsilon}_i$, which vanishes, or a second power of some $\boldsymbol{\varepsilon}_i$, which the linearity forbids. Hence the $J=1$ amplitude vanishes identically.

The argument uses only rotational invariance, Bose statistics and transversality; it does not use parity, so it excludes 1^+ and 1^- alike, and it fails for a massive vector in the final state (which has a longitudinal polarization) and for two *virtual* photons. Together with the four-lepton angular distributions, which distinguish 0^+ from 0^- and from 2^+, this fixes $J^{PC}=0^{++}$. □

The decisive test of the mechanism is not that a scalar exists but that its couplings track mass, because that is what makes it the field that supplies mass rather than merely another particle.

Proposition 104.61 (The couplings are the masses divided by v)**.** *In unitary gauge the interactions of h with the massive fields are*

$$\mathcal{L}_h=-\frac{h}{v}\sum_f m_f c^2\,\bar{f}f+\frac{2h}{v}\left[\frac{M_W^2c^2}{\mu_0\hbar^2}W_\mu^+W^{-\mu}+\frac{M_Z^2c^2}{2\mu_0\hbar^2}Z_\mu Z^\mu\right]+\mathcal{O}\left(h^2\right),\tag{104.69}$$

that is: every coupling is the corresponding mass term divided by v, twice over for the vectors. The vertex strengths are therefore

$$g_{hff}=\frac{m_fc^2}{E_v}\quad\textit{(dimensionless)},\qquad g_{hVV}=\frac{2\left(M_Vc^2\right)^2}{E_v}\quad\textit{(an energy)},\tag{104.70}$$

and the variable that puts both families on one axis — the reduced coupling *of the experimental figures — is g_{hff} for a fermion and $\sqrt{g_{hVV}/2E_v}=M_Vc^2/E_v$ for a vector. Plotted against mass, every Standard Model particle then lies on the single straight line of slope $1/E_v$ through the origin.*

Rests on Theorem 104.38 and Equations (104.34) and (104.46).

Proof. *Derives Proposition 104.61.*

For the fermions, Theorem 104.38(iii): substituting $\Phi=\frac{1}{\sqrt{2}}(0,v+h)^{\mathsf{T}}$ into Equation (104.46) multiplies the mass term by $(1+h/v)$, whose linear part is the first term of Equation (104.69). Dividing by $\sqrt{\hbar c}$ to make it dimensionless, and using $E_v=v\sqrt{\hbar c}$, gives $g_{hff}=m_fc^2/E_v$.

For the vectors, Equation (104.34) was computed at $\Phi=\Phi_0$; restoring the fluctuation replaces v by $v+h$ throughout, so the mass terms are multiplied by $(1+h/v)^2=1+2h/v+h^2/v^2$. The linear part is twice the mass term over v, which is the bracket in Equation (104.69); the *factor of two* is the whole difference between the fermion and vector cases and it comes from the mass term being quadratic in the field's coupling to Φ rather than linear. Hence $g_{hVV}=2(M_Vc^2)^2/E_v$, quadratic in the mass and carrying the dimension of an energy; dividing by $2E_v$ and taking the square root returns M_Vc^2/E_v, the same linear law the fermions obey, which is why the two families can be drawn on one plot. □

Phenomenon 104.62 (The scalar couples in proportion to mass)**.** The strength with which the scalar couples to each particle is measured channel by channel, and it grows with the mass of the particle it couples to: linearly for the fermions and quadratically for the weak bosons. The ten-year coupling maps of ATLAS [2394] and CMS [2410] cover W, Z, t, b, τ and, with weaker significance, μ — some three orders of magnitude in mass — and follow that pattern throughout. This, rather than the mere existence of a scalar, is the test of the mechanism: a scalar whose couplings did not track mass would not be the field that supplies mass.

Rests on Equation (104.70) and Propositions 104.26 and 104.61.

Derivation. *Derives Phenomenon 104.62.*

The prediction is Equation (104.70), derived in Proposition 104.61: a one-parameter family with the parameter, E_v, already fixed by the muon lifetime through Proposition 104.26. There is nothing to adjust.

The measurement is conventionally reported as coupling modifiers κ_i, defined so that $\kappa_i = 1$ reproduces Equation (104.70) and rates scale as κ^2. What is actually measured is a rate, that is a product of a production cross-section and a branching fraction, so a global fit over many final states is required to separate the couplings; the experiments report the results of that fit [2394, 2410], and every κ_i comes out consistent with unity, with uncertainties of roughly a tenth for W, Z, t, b and τ and roughly a fifth for μ. Since $m_\mu c^2 = 0.1057\,\text{GeV}$ and $m_t c^2 = 172.57\,\text{GeV}$, the verified range is a factor of 1630 in mass, over which Equation (104.70) varies by the same factor — so the agreement is a statement about a three-decade lever arm and not about one point.

Two features make the test sharper than a fit to five numbers. The $\gamma\gamma$ and gg vertices have no tree-level counterpart (Table 104.4), so their measured rates test the same couplings at one loop, through a different combination: the $\gamma\gamma$ amplitude is dominated by the W loop with the top loop interfering destructively, and the gg amplitude by the top alone, so agreement in both simultaneously constrains the relative sign of κ_W and κ_t — something no tree-level measurement reaches. And any heavy coloured or charged particle would contribute to those same loops whether or not it can be produced directly, so the agreement bounds such particles indirectly. □

104.7.3 The potential and the self-coupling

Everything verified so far probes the theory at the minimum of the potential. The *shape* of the potential away from the minimum is a separate claim and it is largely unmeasured.

Proposition 104.63 (The self-couplings are predicted, not fitted)**.** *Expanding Equation (104.27) in unitary gauge,*

$$V(h) = \frac{1}{2}\left(\frac{m_h c}{\hbar}\right)^2 h^2 + \frac{(m_h c)^2}{2\hbar^2 v} h^3 + \frac{(m_h c)^2}{8\hbar^2 v^2} h^4 , \tag{104.71}$$

so both the trilinear and the quartic self-coupling are determined by m_h and v, which are already measured. Writing the trilinear coupling as κ_λ times its Standard Model value, the mechanism predicts $\kappa_\lambda = 1$ with no freedom.

Rests on Equations (104.16), (104.27) and (104.37).

Proof. *Derives Proposition 104.63.*

From Equation (104.16) with the Goldstone modes gauged away, $V = \lambda v h^3 + \frac{\lambda}{4} h^4 + \lambda v^2 h^2$, and $\lambda v^2 = \frac{1}{2}(m_h c/\hbar)^2$ by Equation (104.37). Substituting $\lambda = (m_h c)^2/2\hbar^2 v^2$ into each term gives Equation (104.71). The two higher couplings are m_h^2/v and m_h^2/v^2 times pure numbers: one parameter, λ, generates the mass and both self-interactions, which is the content of the claim that the potential is the quartic one and not merely some function with a minimum in the right place. □

Measuring κ_λ requires producing two scalars at once, whose cross-section at the LHC is some three orders of magnitude below single production and which additionally interferes with a box diagram that does not involve the self-coupling. The present constraints [2394] allow κ_λ over a range of order -1 to $+7$ — consistent with the predicted 1, and far too loose to be called a test. Until it narrows, the statement that the vacuum sits at the minimum of a quartic potential is an assumption supported by the mass and the couplings, not a measurement of the potential's shape.

Remark 104.64 (Loops reshape the potential). The tree-level $V(h)$ is not the object whose minimum determines the vacuum. That object is the effective potential of Equation (106.59), the Legendre transform of the generating functional evaluated at a constant field, whose one-loop correction Coleman and Weinberg computed [1495] and Proposition 106.45 reproduces: each field circulating in the loop contributes a term proportional to $\pm M^4(h)\ln M^2(h)$, with a plus sign for bosons and a minus for fermions, $M(h)$ being that field's h-dependent mass. In the Standard Model the top quark's contribution is the largest and it enters with the fermionic minus sign. That single fact is the origin of everything in Section 104.8.1.

Remark 104.65 (The electroweak transition is a crossover). At high temperature the effective potential acquires thermal corrections and the minimum returns to $\Phi = 0$: the symmetry is restored, and the early universe passed through the transition at a temperature of order $100\,\mathrm{GeV}/k_B$. Whether the transition is first order — with bubbles, latent heat and departure from equilibrium — or a smooth crossover depends on the scalar mass, and lattice computations place the endpoint of the first-order line at $m_h c^2 \approx 72\,\mathrm{GeV}$ [1925]. The measured 125.20 GeV is far above it, so the Standard Model transition is a crossover. The consequence is not academic: Sakharov's third condition for generating a baryon asymmetry [1358] is a departure from thermal equilibrium (Theorem 105.41), and a crossover supplies none. The Standard Model therefore cannot account for the observed baryon asymmetry, and Section 105.6.2 quantifies the shortfall. This is a failure of the theory derived from measured parameters, and it is stated here because this chapter measured them.

104.8 What the model does not explain

104.8.1 Vacuum stability

λ is a coupling, so it runs (Chapter 108). Its beta function is dominated by the top Yukawa, and the top's contribution is negative.

Proposition 104.66 (The quartic coupling runs down)**.** *At one loop, in the conventions of Definition 104.24 and with $\hat{y}_t$, $\hat{g}$, $\hat{g}'$ as in Notation 104.1,*

$$16\pi^2 \frac{\mathrm{d}\hat{\lambda}}{\mathrm{d}\ln\mu} = 24\hat{\lambda}^2 + 12\hat{\lambda}\hat{y}_t^2 - 6\hat{y}_t^4 - 3\hat{\lambda}\left(3\hat{g}^2 + \hat{g}'^2\right) + \frac{3}{8}\left[2\hat{g}^4 + \left(\hat{g}^2 + \hat{g}'^2\right)^2\right] \tag{104.72}$$

[2193]. Evaluated at $\mu = m_t c^2$, where the running couplings are $\hat{\lambda} = 0.126$, $\hat{y}_t = 0.94$, $\hat{g} = 0.64$, $\hat{g}' = 0.35$ — each a little below its value at the electroweak scale in Equation (104.38), Table 104.2 and Equation (104.42), because all of them run — the five terms are 0.381, 1.336, −4.684, −0.511 *and* 0.232*, so*

$$\left.\frac{\mathrm{d}\hat{\lambda}}{\mathrm{d}\ln\mu}\right|_{m_t} = \frac{-3.246}{16\pi^2} = -0.0206\,, \tag{104.73}$$

negative, and dominated by the single term $-6\hat{y}_t^4$.

Rests on Definition 104.24, Equation (104.38), and Corollary 106.39.

Derivation of the sign. Derives Proposition 104.66.

The coefficients are quoted from [2193]; what is derived here is why the dominant one is negative and why it is large. By Remark 104.64 the one-loop effective potential receives $-N_c M_t^4(h)\ln M_t^2(h)$ from the top, with the minus sign of a fermion loop — the same sign that makes a closed fermion loop carry (-1) in Corollary 106.39. Since $M_t(h) \propto \hat{y}_t h$, this is a term $\propto -\hat{y}_t^4 h^4 \ln h$, that is a negative contribution to the running quartic, with the colour factor $N_c = 3$ and the fourth power of the coupling. Numerically $\hat{y}_t^4 = 0.78$ while $\hat{\lambda}^2 = 0.016$, so the fermion term beats the bosonic ones by a factor of order fifty, and Equation (104.73) follows by arithmetic. The one number that matters is the sign. □

Remark 104.67 (Metastability, and how fragile the statement is). Integrating Equation (104.72) together with the running of $\hat{y}_t$ and $\hat{g}_3$ — which cannot be done by holding the coefficients fixed, because $\hat{y}_t$ itself falls with scale and $\left|\mathrm{d}\hat{\lambda}/\mathrm{d}\ln\mu\right|$ falls with it — gives $\hat{\lambda}(\mu)$ crossing zero at $\mu \approx 10^{10}$–$10^{11}\,\mathrm{GeV}$ [2193, 2215]. Above that scale the potential turns over and a deeper minimum exists at large field values: the electroweak vacuum is not the global minimum but a metastable one. The computed tunnelling lifetime exceeds the age of the universe by hundreds of orders of magnitude, so nothing observable follows.

Three caveats, all of which the cited analyses state and none of which may be dropped:

- the conclusion sits on a knife edge in m_t. Shifting the top mass by about $1\,\mathrm{GeV}$, or the strong coupling by about one standard deviation, moves the answer between absolute stability, metastability and instability. The measured m_t has an uncertainty of $0.29\,\mathrm{GeV}$ [2427], but the relation between the quantity a hadron collider reconstructs and the short-distance mass entering Equation (104.72) carries a further theoretical uncertainty of comparable size;
- the extrapolation assumes no new particle couples to the scalar anywhere between $10^3\,\mathrm{GeV}$ and $10^{10}\,\mathrm{GeV}$ — sixteen decades in which nothing has been looked at. A single new state changes the running;
- the running is a statement about a coupling in a scheme, not about an observable. Whether "the vacuum is metastable" is a physical statement at all depends on the extrapolation being valid, which is the assumption in question.

The honest summary is that the measured parameters put the Standard Model, extrapolated with no new physics, marginally on the metastable side, and that the margin is smaller than the combined theoretical and experimental uncertainty on the top mass. It is an intriguing near-coincidence, and it is not evidence for anything.

Derivation pending #159

Vacuum stability: the coupled integration of the one-loop renormalization-group equations for the quartic coupling, the top Yukawa and the three gauge couplings from the electroweak scale upward, which is what locates the zero of the quartic near $1\times10^{10}\,\mathrm{GeV}$*; and the bounce action governing the tunnelling rate out of the electroweak minimum. The beta function and its sign are derived in the text above; the integration and the bounce belong in Appendix A.*

104.8.2 Open questions, stated honestly

The electroweak theory explains a great deal from very little: two gauge couplings, one scalar mass and one vacuum expectation value generate the W, the Z, the photon's exact masslessness, the Fermi constant, the neutral current and the whole pattern of boson self-interactions. What it does not do is explain its own inputs.

The Yukawa hierarchy. Table 104.2 is a list of nine numbers spanning 3.4×10^5, taken from experiment. Nothing in the Lagrangian relates them, and no evidence-based mechanism producing them exists. The four parameters of the CKM matrix (Proposition 101.63) are in the same position.

The number of generations. Three, by Phenomenon 101.76 for the light neutrinos and by Proposition 104.53 for the requirement that each be complete. Why three is not known; the theory works for any number.

The scale. $E_v = 246.22\,\mathrm{GeV}$ is one input. Expressed against the only other scale built from the constants of Nature, the Planck energy $\sqrt{\hbar c^5/G} \approx 1.22 \times 10^{19}\,\mathrm{GeV}$, it is smaller

by a factor 2×10^{-17}, and Remark 104.30 records a related and much larger discrepancy in the vacuum energy.

Sensitivity of m_h^2 to heavy thresholds. If the theory is embedded in a larger one containing a particle of mass M coupled to Φ, the scalar's squared mass receives a correction of order M^2 times a loop factor, so reproducing the measured 125.20 GeV requires a cancellation of relative size $(m_h/M)^2$. It should be said plainly what this is and is not. It is an argument about the *form* a more fundamental theory would have to take; it is not an observation, because no such particle is known to exist and the correction is not a measurable quantity. It has motivated a great many proposed extensions — supersymmetry, technicolour, grand unification, extra dimensions — and direct searches at the LHC have so far returned null results for every one of them. None of those programmes has observational support, and by editorial rule 1 they are named here only to record that fact.

Neutrino mass. Equation (104.46) gives the neutrino no mass, because there is no ν_R in Table 104.1. Oscillations show the masses are not zero (Chapter 114). This is the one place where a laboratory measurement contradicts the Lagrangian of Equation (104.28) rather than merely failing to be explained by it, and Chapter 103 takes it up.

Dark matter, dark energy, the baryon asymmetry. None has an explanation here. Remark 104.65 showed that the measured scalar mass positively *forbids* the Standard Model from generating the baryon asymmetry, which is a sharper statement than silence. The cosmological evidence belongs to Chapter 134, and the full inventory of what is observed and unexplained is Chapter 133.

Chapter 105

Discrete Symmetries and CPT

Three transformations lie outside the connected Lorentz group of Chapter 39: space inversion P, charge conjugation C and time reversal T. Each was assumed to be an exact symmetry of nature, and each turned out not to be. Parity fell first, proposed as testable by Lee and Yang [1091] and demolished by Wu's cobalt-60 experiment [1133], the subject of Chapter 110; the combination CP, offered by Landau as the replacement [1118], fell in turn with the 2π decay of the long-lived kaon [1274], the subject of Chapter 111; and direct violation of T was observed in the neutral kaon [1948] and in the neutral B system [2197]. This chapter collects those results and the theory that survives them.

What survives is the product. The CPT theorem of Lüders [1049], Pauli [1079] and Jost [1115] says that any local, Lorentz-invariant quantum field theory with a positive-energy spectrum is invariant under CPT — so particle and antiparticle must have equal masses and lifetimes and opposite magnetic moments. It is therefore the sharpest available test of the axioms of Chapter 109, and the chapter reviews the measurements that probe it, from antihydrogen spectroscopy [2280, 2299] to the antiproton magnetic moment [2288] and charge-to-mass ratio [2400]. It closes on two honest failures: Sakharov's conditions for a baryon asymmetry [1358], which the observed CP violation falls short of by many orders of magnitude, and the strong CP problem, bounded but unexplained by the neutron electric dipole moment [2346].

The conventions are those fixed for quantum electrodynamics in Section 100.1.1, and are used without further comment: the metric is $\eta_{\mu\nu} = \mathrm{diag}(+1,-1,-1,-1)$, coordinates are $x^\mu = (ct, \boldsymbol{x})$, four-momenta are $p^\mu = (E/c, \boldsymbol{p})$ with the mass shell $p^2 = m^2c^2$ of Equation (100.1), and $\hbar$ and c appear explicitly in every formula. Exactly one symbol is reassigned against them, and the reassignment is declared here rather than left to the reader to discover: Notation 100.1 reserves ϵ, standing alone, for the positive infinitesimal of the Feynman contour, whereas from Section 105.3.1 onwards ϵ is the CP-violating mixing parameter of the neutral kaon, which is what the symbol means everywhere in this subject. The contour prescription is accordingly written $+\mathrm{i}0$ in this chapter and never $+\mathrm{i}\epsilon$.

Operators on the Hilbert space are written in calligraphic type — $\mathcal{P}, \mathcal{C}, \mathcal{T}$ — with a single exception: their product is written Θ (Definition 105.33), the capital theta being universal for it. The Dirac matrices γ^μ and γ^5 are the ordinary italic ones of Chapter 93; the numerical 4×4 matrices that implement P, C and T on Dirac indices are written in upright roman, and carry the subscript D where the letter would otherwise collide with an operator — the charge-conjugation matrix $\mathrm{C_D}$ of Definition 100.9 and the time-reversal matrix $\mathrm{T_D}$, against the parity matrix S_P, whose letter collides with nothing. The two kinds must not be confused: $\mathcal{C}$ is a unitary operator on a Fock space, $\mathrm{C_D}$ a matrix one can write down.

105.1 The three discrete operations

The Lorentz group L has four connected components (Proposition 39.8), and the three not containing the identity are reached from it by adjoining the two matrices

$$\Lambda_P = \mathrm{diag}(+1,-1,-1,-1)\,, \qquad \Lambda_T = \mathrm{diag}(-1,+1,+1,+1)\,, \tag{105.1}$$

acting on $x^\mu = (ct, \boldsymbol{x})$. Neither can be reached by any continuous motion, so whether Nature is invariant under them is not settled by the Lie algebra of Section 39.2.7: it is an experimental question, and Section 105.2 records the answer.

Charge conjugation is of a different kind. It is not a spacetime transformation at all: it acts on the internal labels of the fields, exchanging each particle with its antiparticle while leaving momentum and spin untouched. It has no classical counterpart in the sense that P and T do, and it exists only because the relativistic wave equations of Chapter 93 force antiparticles into the theory — which is why the empirical discovery of C had to wait for the positron of Chapter 95.

By Wigner's theorem (Theorem 94.3) each of the three, being a symmetry of the ray space, is implemented by an operator that is either unitary or antiunitary. The theorem does not say which. The answer, derived in Section 105.1.3, is that $\mathcal{P}$ and $\mathcal{C}$ may be taken unitary while $\mathcal{T}$ must be antiunitary, and that single structural difference is responsible for most of what distinguishes T from its two companions: T carries no conserved quantum number, forbids no transition outright, and can be tested only by comparing a process with its motion reverse.

105.1.1 Parity

Parity in quantum mechanics Wigner introduced parity as a conserved quantum number in 1927 [592], reading it off the reflection symmetry of atomic spectra. The argument is short enough to give in full, and it fixes the pattern that the field-theoretic version will repeat.

Let a spinless particle move in a central potential, so that the Hamiltonian is

$$H = \frac{\hat{\boldsymbol{p}}^2}{2m} + V(|\boldsymbol{x}|)\,. \tag{105.2}$$

Define the unitary operator $\mathcal{P}$ on wavefunctions by

$$(\mathcal{P}\psi)(\boldsymbol{x}) = \eta_P\, \psi(-\boldsymbol{x})\,, \tag{105.3}$$

with η_P a phase. Then $\mathcal{P}\hat{\boldsymbol{x}}\mathcal{P}^{-1} = -\hat{\boldsymbol{x}}$ and, because $\hat{\boldsymbol{p}} = -\mathrm{i}\hbar\nabla$ and $\nabla \to -\nabla$ under $\boldsymbol{x} \to -\boldsymbol{x}$, also $\mathcal{P}\hat{\boldsymbol{p}}\mathcal{P}^{-1} = -\hat{\boldsymbol{p}}$. Both terms of Equation (105.2) are therefore invariant, $[\mathcal{P}, H] = 0$, and energy eigenstates may be chosen to be eigenstates of $\mathcal{P}$. Since $\mathcal{P}^2$ acts as the identity on $\boldsymbol{x}$, we may choose $\eta_P^2 = 1$ for a scalar wavefunction, so the eigenvalues are ± 1: even and odd states.

Orbital angular momentum $\hat{\boldsymbol{L}} = \hat{\boldsymbol{x}} \times \hat{\boldsymbol{p}}$ is a product of two odd vectors and is therefore *even*: $\mathcal{P}\hat{\boldsymbol{L}}\mathcal{P}^{-1} = +\hat{\boldsymbol{L}}$. A quantity that behaves this way under P — like a vector under rotations but even under reflection — is called an *axial* (or pseudo-) vector, as against the *polar* vectors $\boldsymbol{x}$ and $\boldsymbol{p}$; the corresponding scalars are called pseudoscalar and scalar. The spherical harmonics inherit the eigenvalue

$$Y_{\ell m}(-\hat{\boldsymbol{n}}) = (-1)^\ell Y_{\ell m}(\hat{\boldsymbol{n}})\,, \tag{105.4}$$

so a state of orbital angular momentum ℓ has parity $(-1)^\ell$ times the intrinsic parities of its constituents.

Proposition 105.1 (Laporte's rule)**.** *In a parity-conserving theory, an electric-dipole transition connects only states of opposite parity. In particular no $E1$ line joins two even or two odd atomic configurations.*

Rests on Equation (105.3)*.*

Proof. *Derives Proposition 105.1.*

The electric-dipole operator is $\hat{\boldsymbol{d}} = -e\sum_a \hat{\boldsymbol{x}}_a$, a sum of polar vectors, so $\mathcal{P}\hat{\boldsymbol{d}}\mathcal{P}^{-1} = -\hat{\boldsymbol{d}}$. If $\mathcal{P}\,|i\rangle = \pi_i\,|i\rangle$ and $\mathcal{P}\,|f\rangle = \pi_f\,|f\rangle$ with $\pi_{i,f} = \pm 1$, then inserting $\mathcal{P}^{-1}\mathcal{P} = \mathbb{1}$ twice gives

$$\langle f|\,\hat{\boldsymbol{d}}\,|i\rangle = \langle f|\,\mathcal{P}^{-1}\left(\mathcal{P}\hat{\boldsymbol{d}}\mathcal{P}^{-1}\right)\mathcal{P}\,|i\rangle = -\pi_i\pi_f\,\langle f|\,\hat{\boldsymbol{d}}\,|i\rangle\ ,$$

so the matrix element vanishes unless $\pi_i\pi_f = -1$. □

This is the selection rule that Laporte extracted empirically from iron and titanium spectra in 1925 [528], two years before Wigner explained it: the terms of a complex spectrum fall into two classes, and lines occur only between classes and never within one. It is the earliest quantitative evidence for parity as a physical quantum number, and it is worth noticing what kind of evidence it is — a systematic *absence*, which is what a symmetry always delivers.

Parity on the fields In field theory $\mathcal{P}$ is a unitary operator on the Fock space, defined by its action on the fields. For a spinless field the possibilities are two,

$$\mathcal{P}\phi(ct,\boldsymbol{x})\mathcal{P}^{-1} = \eta_P\,\phi(ct,-\boldsymbol{x})\,,\qquad \eta_P = \pm 1\,, \tag{105.5}$$

and the field is called scalar for $\eta_P = +1$ and pseudoscalar for $\eta_P = -1$. The number η_P is the *intrinsic parity* of the corresponding particle.

For the electromagnetic potential the transformation is not a matter of choice. The observable fields are $\boldsymbol{E} = -\nabla\phi_{\mathrm{el}} - \partial_t\boldsymbol{A}$ and $\boldsymbol{B} = \nabla\times\boldsymbol{A}$; under $\boldsymbol{x}\to-\boldsymbol{x}$ one has $\nabla\to-\nabla$, so the assignment

$$\mathcal{P}A^0(ct,\boldsymbol{x})\mathcal{P}^{-1} = A^0(ct,-\boldsymbol{x})\,,\qquad \mathcal{P}A^i(ct,\boldsymbol{x})\mathcal{P}^{-1} = -A^i(ct,-\boldsymbol{x}) \tag{105.6}$$

gives $\boldsymbol{E}\to-\boldsymbol{E}$ and $\boldsymbol{B}\to+\boldsymbol{B}$: the electric field is polar, the magnetic field axial. Both signs are forced by Equation (105.6) and neither is conventional. In compact notation $\mathcal{P}A^\mu(x)\mathcal{P}^{-1} = \Lambda_P{}^\mu{}_\nu A^\nu(x_P)$ with $x_P = (ct,-\boldsymbol{x})$; the photon has intrinsic parity -1, since the spatial polarization vector, which is what a physical photon carries, reverses.

For the Dirac field the transformation is likewise forced, up to one phase.

Proposition 105.2 (Parity on a Dirac field)**.** *Let ψ satisfy $(\mathrm{i}\hbar\gamma^\mu\partial_\mu - mc)\psi = 0$. Then*

$$\mathcal{P}\psi(ct,\boldsymbol{x})\mathcal{P}^{-1} = \eta_P\,\gamma^0\psi(ct,-\boldsymbol{x}) \tag{105.7}$$

maps solutions to solutions, and γ^0 is the unique matrix (up to a factor) that does so.

Rests on Equations (100.5) *and* (105.1)*.*

Proof. *Derives Proposition 105.2.*

Write $\psi'(x_P) = \mathrm{S}_P\psi(x)$ and demand that ψ' satisfy the Dirac equation in the reflected coordinates. Since $\partial/\partial x^0_P = \partial_0$ and $\partial/\partial x^i_P = -\partial_i$, the requirement is

$$\left(\mathrm{i}\hbar\gamma^0\partial_0 - \mathrm{i}\hbar\gamma^i\partial_i - mc\right)\mathrm{S}_P\psi = 0\,.$$

Multiplying the original equation on the left by S_P and inserting $\mathrm{S}_P^{-1}\mathrm{S}_P$ before ψ gives

$$\left(\mathrm{i}\hbar\,\mathrm{S}_P\gamma^0\mathrm{S}_P^{-1}\partial_0 + \mathrm{i}\hbar\,\mathrm{S}_P\gamma^i\mathrm{S}_P^{-1}\partial_i - mc\right)\mathrm{S}_P\psi = 0\,,$$

so the two agree if and only if $\mathrm{S}_P\gamma^0\mathrm{S}_P^{-1} = \gamma^0$ and $\mathrm{S}_P\gamma^i\mathrm{S}_P^{-1} = -\gamma^i$. The Clifford relation $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}\mathbb{1}$ gives $\gamma^0\gamma^0\gamma^0 = \gamma^0$ and $\gamma^0\gamma^i\gamma^0 = -\gamma^i$, since $(\gamma^0)^2 = \mathbb{1}$, so $\mathrm{S}_P = \gamma^0$ works. Uniqueness up to a factor follows because any other solution S' makes $\mathrm{S}'\gamma^0$ commute with all four γ^μ, and the Dirac algebra is irreducible, so by Schur's lemma (Chapter 5) $\mathrm{S}'\gamma^0$ is a multiple of the identity. □

The phase η_P is a new label, and it is worth saying exactly where it comes from. Wigner's classification in Chapter 94 sorts one-particle states by the two Casimir invariants of the *connected* Poincaré group, mass and spin (Theorem 94.29); $\mathcal{P}$ is not in that group, so the classification says nothing about it. Enlarging the group by Λ_P enlarges the label set by one sign per irreducible representation, and that sign is the intrinsic parity. It is therefore a representation label of the *full* Lorentz group, and it exists as a quantum number only in a sector where P is a symmetry — which is why the weak interaction does not assign one.

Applying Equation (105.7) twice returns $\eta_P^2\psi$, so $\eta_P^2 = \pm 1$ and $\eta_P \in \{\pm 1, \pm \mathrm{i}\}$: a fermion's intrinsic parity need not be real. This is not a defect but a reflection of the fact that fermion number is superselected, so that a state of one fermion and a state of none are never superposed and the phase never interferes with anything. What *is* measurable is a relative parity, and one relative parity is fixed by the theory rather than by convention.

Proposition 105.3 (A fermion and its antifermion have opposite intrinsic parity)**.** *If the field ψ transforms as Equation (105.7) with phase η_P, the charge-conjugate field $\psi^c = \mathrm{C}_\mathrm{D}\bar\psi^{\mathsf{T}}$ of Definition 100.9 transforms with phase $-\eta_P^*$.*

Rests on Equations (100.20) and (105.7) and Definition 100.9.

Proof. *Derives Proposition 105.3.*

From Equation (105.7), $\mathcal{P}\bar\psi(x)\mathcal{P}^{-1} = \eta_P^*\bar\psi(x_P)\gamma^0$, since $\gamma^{0\dagger} = \gamma^0$ and $\bar\psi = \psi^\dagger\gamma^0$. Hence

$$\mathcal{P}\psi^c(x)\mathcal{P}^{-1} = \mathrm{C}_\mathrm{D}\left(\eta_P^*\bar\psi(x_P)\gamma^0\right)^{\mathsf{T}} = \eta_P^*\,\mathrm{C}_\mathrm{D}\gamma^{0\mathsf{T}}\bar\psi^{\mathsf{T}}(x_P)\,.$$

Equation (100.20) with $\mu = 0$ reads $\mathrm{C}_\mathrm{D}\gamma^{0\mathsf{T}}\mathrm{C}_\mathrm{D}^{-1} = -\gamma^0$, so $\mathrm{C}_\mathrm{D}\gamma^{0\mathsf{T}} = -\gamma^0\mathrm{C}_\mathrm{D}$ and the right-hand side is $-\eta_P^*\gamma^0\psi^c(x_P)$. □

Example 105.4 (Positronium)**.** An electron–positron bound state with relative orbital angular momentum L has parity

$$\eta_P(e^+e^-) = \eta_P(e^-)\,\eta_P(e^+)\,(-1)^L = (-1)^{L+1}\,, \tag{105.8}$$

by Proposition 105.3 and Equation (105.4). The ground states are therefore odd, $J^P = 0^-$ for para-positronium and 1^- for ortho-positronium, which is why the 2γ decay of the former proceeds to a state of odd parity and not, as one might carelessly guess, to an even one.

The bilinears, and where parity can hide Everything that can appear in a local Lagrangian built from a Dirac field is a sum of the sixteen bilinears $\bar\psi\Gamma\psi$ with Γ running over $\mathbb{1}$, γ^5, γ^μ, $\gamma^\mu\gamma^5$ and $\sigma^{\mu\nu} = \frac{\mathrm{i}}{2}[\gamma^\mu, \gamma^\nu]$, possibly carrying derivatives. Their behaviour under P follows from Equation (105.7) by one line of Clifford algebra each. Since $\gamma^0\gamma^\mu\gamma^0 = \gamma_\mu$ — meaning $+\gamma^0$ for $\mu = 0$ and $-\gamma^i$ for $\mu = i$ — and $\gamma^0\gamma^5\gamma^0 = -\gamma^5$ because γ^5 anticommutes with every γ^μ, one obtains

$$\mathcal{P}\left[\bar\psi\psi\right](x)\mathcal{P}^{-1} = +\left[\bar\psi\psi\right](x_P)\,, \tag{105.9}$$

$$\mathcal{P}\left[\bar\psi\gamma^5\psi\right](x)\mathcal{P}^{-1} = -\left[\bar\psi\gamma^5\psi\right](x_P)\,, \tag{105.10}$$

$$\mathcal{P}\left[\bar\psi\gamma^\mu\psi\right](x)\mathcal{P}^{-1} = +\left[\bar\psi\gamma_\mu\psi\right](x_P)\,, \tag{105.11}$$

$$\mathcal{P}\left[\bar\psi\gamma^\mu\gamma^5\psi\right](x)\mathcal{P}^{-1} = -\left[\bar\psi\gamma_\mu\gamma^5\psi\right](x_P)\,. \tag{105.12}$$

The four names — scalar S, pseudoscalar $\tilde{S}$, vector V^μ, axial vector A^μ — are exactly these four lines, and Table 105.1 collects them together with the C and T columns derived below.

The interaction of quantum electrodynamics is $-j^\mu A_\mu$ with $j^\mu = qc\,\bar{\psi}\gamma^\mu\psi$ the vector current of Equation (100.6). Comparing Equation (105.11) with Equation (105.6), the two lower an index in the same way, so the contraction $j^\mu A_\mu$ is invariant and P is an exact symmetry of QED — the statement made without proof in Section 100.1.4 and proved here. The same argument applies verbatim to the quark–gluon coupling of Chapter 102, which is also a vector current, with one exception that is the subject of Section 105.7: the topological term permitted in the QCD Lagrangian is P-odd, and only experiment tells us that its coefficient is tiny.

What Equation (105.12) shows is where parity violation must live. If a Lagrangian contains *both* V^μ and A^μ coupled to the same object, then under P the relative sign of the two flips and the theory is not invariant. Parity is violated exactly when vector and axial currents interfere — which is why the discovery of Section 105.2.2 took the form it did.

Multiplicativity and the parity of the pion Because $\mathcal{P}$ is a unitary operator implementing a $\mathbb{Z}_2$ operation, the parity of a composite state is the product of the intrinsic parities of its constituents times $(-1)^L$ for each relative orbital angular momentum. Parity is therefore a *multiplicative* quantum number, unlike charge or baryon number, which are additive; this is the general fact that a symmetry group $\mathbb{Z}_2$ has one-dimensional representations labelled by a sign.

Intrinsic parities are measurable only relative to a convention, and the convention universally adopted is that the proton, the neutron and the Λ carry $\eta_P = +1$. Everything else is then determined by experiment. The classical determination is the pion's, and it is worth giving in full because it uses nothing but the two counting rules just established plus the antisymmetry of identical fermions.

Phenomenon 105.5 (The charged pion is pseudoscalar)**.** The reaction $\pi^- + d \to n + n$, in which a negative pion is captured at rest from an atomic orbit of pionic deuterium, proceeds at a rate comparable to the competing radiative channel $\pi^- + d \to n + n + \gamma$ [985]. Its occurrence at all requires the pion to have intrinsic parity $\eta_P(\pi) = -1$, given $\eta_P(n) = +1$.
Rests on Equation (105.4).

Derivation. *Derives Phenomenon 105.5.*

A negative pion stopped in liquid deuterium is captured into a Bohr orbit of the pionic deuterium atom and cascades downward. Because the pion is some 273 times heavier than the electron, its Bohr radius is smaller by the same factor and the overlap with the deuteron is enormous; the cascade terminates in the $1S$ state, from which nuclear capture occurs. The initial orbital angular momentum is therefore $L_i = 0$, and since the pion is spinless and the deuteron has spin 1, the initial total angular momentum is $J = 1$. The initial parity is

$$\eta_P(\pi)\,\eta_P(d)\,(-1)^{L_i} = \eta_P(\pi)\,,$$

because the deuteron, a ${}^3S_1$ bound state of two nucleons of parity $+1$ with a small ${}^3D_1$ admixture, has $\eta_P(d) = +1$: both $L = 0$ and $L = 2$ are even.

The final state is two neutrons, identical fermions, so the total wavefunction must be antisymmetric under exchange. Exchanging the two neutrons multiplies the spatial part by $(-1)^L$ and the spin part by $(-1)^{S+1}$ — the triplet $S = 1$ is symmetric and the singlet $S = 0$ antisymmetric — so the requirement is

$$(-1)^{L+S+1} = -1\,, \qquad \text{that is} \qquad L + S \text{ even}\,. \tag{105.13}$$

Enumerate the states with $J = 1$: ${}^1S_0$ has $J = 0$; ${}^3S_1$ has $L + S = 1$, odd, excluded; ${}^1P_1$ has $L + S = 1$, excluded; ${}^3D_1$ has $L + S = 3$, excluded; and ${}^3P_1$ has $L + S = 2$, even, and $J = 1$. Exactly one state survives, ${}^3P_1$, of parity $(-1)^L = (-1)^1 = -1$.

Equating initial and final parity gives $\eta_P(\pi) = -1$. The experimental input is that the reaction occurs: Panofsky, Aamodt and Hadley measured the ratio of the two capture channels and found the non-radiative one present at a comparable rate [985], which cannot happen if the only accessible final state is forbidden by parity. □

The same $J^P = 0^-$ assignment for the neutral pion follows from the observed $\pi^0 \to \gamma\gamma$ angular correlation, and the fact that all three pions form an isospin triplet with the same spin and parity ties them together. A pseudoscalar pion is the reason a two-pion state has parity $(+1)(-1)^L$ rather than $(-1)^L$, and it is what makes $K \to 2\pi$ and $K \to 3\pi$ distinguish parities in Section 105.2.1.

105.1.2 Charge conjugation

The operation on the Dirac field The Dirac equation in an external electromagnetic field,

$$\left[\mathrm{i}\hbar\gamma^\mu\left(\partial_\mu + \frac{\mathrm{i}q}{\hbar}A_\mu\right) - mc\right]\psi = 0\,, \tag{105.14}$$

with D_μ the covariant derivative of Equation (100.9), describes a particle of charge q and mass m. Its solutions of negative frequency were interpreted by Dirac as antiparticles and observed as the positron (Chapter 95). Charge conjugation is the statement that the equation with q replaced by $-q$ has exactly as many solutions, in explicit correspondence.

Proposition 105.6 (The charge-conjugate spinor)**.** *Let* $\mathrm{C_D}$ *satisfy Equation* (100.20)*,* $\mathrm{C_D}(\gamma^\mu)^{\mathsf{T}}\mathrm{C_D}^{-1} = -\gamma^\mu$*. If* ψ *solves Equation* (105.14)*, then*

$$\psi^c := \mathrm{C_D}\,\bar\psi^{\mathsf{T}} \tag{105.15}$$

solves the same equation with $q \to -q$*, with the same mass.*

Rests on Equations (100.20) *and* (105.14)*.*

Proof. *Derives Proposition 105.6.*

Expand Equation (105.14): $\mathrm{i}\hbar\gamma^\mu\partial_\mu\psi - q\gamma^\mu A_\mu\psi - mc\psi = 0$. Take the Hermitian conjugate, using $\gamma^{\mu\dagger} = \gamma^0\gamma^\mu\gamma^0$ and A_μ real:

$$-\mathrm{i}\hbar\,\partial_\mu\psi^\dagger\gamma^{\mu\dagger} - q\,\psi^\dagger\gamma^{\mu\dagger}A_\mu - mc\,\psi^\dagger = 0\,.$$

Multiply on the right by γ^0 and use $\gamma^{\mu\dagger}\gamma^0 = \gamma^0\gamma^\mu$ together with $\bar\psi = \psi^\dagger\gamma^0$:

$$-\mathrm{i}\hbar\,\partial_\mu\bar\psi\,\gamma^\mu - q\,\bar\psi\gamma^\mu A_\mu - mc\,\bar\psi = 0\,.$$

This is a row-spinor equation; transpose it to obtain a column,

$$-\mathrm{i}\hbar\,(\gamma^\mu)^{\mathsf{T}}\,\partial_\mu\bar\psi^{\mathsf{T}} - q\,(\gamma^\mu)^{\mathsf{T}}\,A_\mu\bar\psi^{\mathsf{T}} - mc\,\bar\psi^{\mathsf{T}} = 0\,,$$

and multiply on the left by $\mathrm{C_D}$, inserting $\mathrm{C_D}^{-1}\mathrm{C_D}$ in front of each $\bar\psi^{\mathsf{T}}$. With $\mathrm{C_D}(\gamma^\mu)^{\mathsf{T}}\mathrm{C_D}^{-1} = -\gamma^\mu$ and $\psi^c = \mathrm{C_D}\bar\psi^{\mathsf{T}}$ this reads

$$\mathrm{i}\hbar\gamma^\mu\partial_\mu\psi^c + q\gamma^\mu A_\mu\psi^c - mc\,\psi^c = 0\,,$$

which is Equation (105.14) with $q \to -q$. □

In the Dirac representation one may take $\mathrm{C_D} = \mathrm{i}\gamma^2\gamma^0$, as recorded in Definition 100.9; then $\mathrm{C_D^{\mathsf{T}}} = -\mathrm{C_D}$ and $\mathrm{C_D^\dagger} = \mathrm{C_D}^{-1} = -\mathrm{C_D}$. Two features of Proposition 105.6 deserve emphasis. First, the mass is untouched: charge conjugation relates a particle to something of the same mass, which is a *theorem* of the free theory and, once interactions are switched on, becomes the much deeper statement of Section 105.5. Second, ψ^c is built from $\bar\psi^{\mathsf{T}}$, i.e. from ψ^*, so the operation involves complex conjugation of the wavefunction while remaining, on the Fock space, a *unitary* operator: the antilinearity is absorbed into the exchange of creation and annihilation operators. In terms of modes, $\mathcal{C}$ maps the operator that annihilates an electron of momentum $\boldsymbol{p}$ and spin s into the one that annihilates a positron of the same $\boldsymbol{p}$ and the same s. Momentum and spin are spectators; only the charge-like labels move.

C-parity as a quantum number Since $\mathcal{C}$ exchanges a particle with a different particle, a charged state is never an eigenstate: $\mathcal{C}\,|e^-\rangle \propto |e^+\rangle$, which is orthogonal to it. Only a state that is its own antiparticle can carry a C eigenvalue — the photon, the π^0, the η, positronium, a neutral non-strange meson–antimeson combination.

The photon's value is forced. The interaction $-j^\mu A_\mu$ must be C-even (it is the coupling that generates all of atomic physics, and C is an exact symmetry of QED), while the current $j^\mu = qc\,\bar\psi\gamma^\mu\psi$ is C-odd. To see the latter, apply $\mathcal{C}\psi\mathcal{C}^{-1} = \eta_C \mathrm{C_D}\bar\psi^{\mathsf T}$ and $\mathcal{C}\bar\psi\mathcal{C}^{-1} = -\eta_C^*\psi^{\mathsf T}\mathrm{C_D^{-1}}$ to a bilinear. The two fermion operators must be reordered, which costs a minus sign, and the result is

$$\mathcal{C}\left[\bar\psi\Gamma\psi\right]\mathcal{C}^{-1} = \bar\psi\left(\mathrm{C_D^{-1}}\Gamma\mathrm{C_D}\right)^{\mathsf T}\psi\,. \tag{105.16}$$

For $\Gamma = \gamma^\mu$ the defining relation gives $\mathrm{C_D^{-1}}\gamma^\mu\mathrm{C_D} = -(\gamma^\mu)^{\mathsf T}$, whose transpose is $-\gamma^\mu$, so $V^\mu \to -V^\mu$. Hence

$$\mathcal{C}A^\mu(x)\mathcal{C}^{-1} = -A^\mu(x)\,, \qquad \eta_C(\gamma) = -1\,, \tag{105.17}$$

and an n-photon state has $C = (-1)^n$. The same computation applied to the other four Γ's gives the C column of Table 105.1; the only entry that requires a moment's care is $\gamma^5 = \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3$, for which $\mathrm{C_D^{-1}}\gamma^5\mathrm{C_D} = \mathrm{i}(-\gamma^{0\mathsf T})(-\gamma^{1\mathsf T})(-\gamma^{2\mathsf T})(-\gamma^{3\mathsf T}) = \mathrm{i}\,(\gamma^3\gamma^2\gamma^1\gamma^0)^{\mathsf T} = (\gamma^5)^{\mathsf T}$, because reversing four mutually anticommuting factors costs six transpositions and hence no sign. So the pseudoscalar bilinear is C-even.

Example 105.7 (The neutral pion, and a forbidden decay)**.** The decay $\pi^0 \to \gamma\gamma$ is observed with a branching fraction of $(98.823 \pm 0.034)\,\%$ [2427], so $\eta_C(\pi^0) = (-1)^2 = +1$. The decay $\pi^0 \to 3\gamma$ would require $C = (-1)^3 = -1$ and is forbidden. The experimental limit is a branching fraction below 3.1×10^{-8} at 90 % confidence [2427] — a suppression of more than seven orders of magnitude relative to the allowed channel, where phase space alone would cost only about two. The missing five orders are the symmetry.

Example 105.8 (Positronium as a C meter)**.** For a fermion–antifermion pair of relative orbital angular momentum L and total spin S, exchanging the two particles requires exchanging their positions, their spins, and their charge labels; the first costs $(-1)^L$, the second $(-1)^{S+1}$, the third is $\mathcal{C}$ itself, and the total must be -1 because the constituents are fermions. Hence

$$\eta_C(f\bar f) = (-1)^{L+S}\,. \tag{105.18}$$

Para-positronium ($L = 0$, $S = 0$) has $C = +1$ and decays to two photons; ortho-positronium ($L = 0$, $S = 1$) has $C = -1$ and must decay to three. The observed rates differ by the extra factor of α that the third photon costs, the singlet living 125 ps and the triplet 142 ns, a ratio of about 1100 that no dynamical accident would supply.

Theorem 105.9 (Furry)**.** *Any amplitude in quantum electrodynamics consisting of a single closed fermion loop with an odd number of external photons vanishes identically [793].*

Rests on Theorem 100.10 and Equation (105.17)*.*

This is proved as Theorem 100.10 by comparing the two orientations of the loop; the symmetry reading is one line. A closed fermion loop with n photons attached is a matrix element of an operator that is C-even (the vacuum is C-even, and so is the interaction), while the n-photon state has $C = (-1)^n$; for odd n the two disagree and the amplitude must vanish. The practical value is large: Furry's theorem removes whole classes of diagrams before they are computed, and it is the reason light-by-light scattering starts at four photons and not at three.

G-parity Charge conjugation is useless for a charged pion, since $\mathcal{C}\,|\pi^+\rangle = |\pi^-\rangle$. But the strong interaction of Chapter 102 conserves isospin, and a rotation by π about the 2-axis in isospin space maps $\pi^\pm$ into $\pi^\mp$ as well. Composing the two gives an operation under which the whole pion triplet is an eigenstate. Isospin is a dimensionless internal label and not an angular momentum, so its generators carry no $\hbar$ and the rotation angle appears alone in the exponent:

$$\mathcal{G} := \mathcal{C}\,\mathrm{e}^{\mathrm{i}\pi \hat{I}_2}\,. \tag{105.19}$$

For a neutral, non-strange multiplet of isospin I one has

$$\eta_G = \eta_C\,(-1)^I\,, \tag{105.20}$$

because the isospin rotation acts on the $I_3 = 0$ member of an isospin-I multiplet with eigenvalue $(-1)^I$, exactly as an ordinary rotation by π acts on Y_{I0}. For the pion, $\eta_C(\pi^0) = +1$ and $I = 1$, so

$$\eta_G(\pi) = -1\,, \qquad \eta_G(n\pi) = (-1)^n\,. \tag{105.21}$$

G-parity is conserved by the strong interaction, and the selection rule is immediately useful: the ρ meson ($I = 1$, $\eta_C = -1$, hence $\eta_G = +1$) decays strongly to two pions and not to three, while the ω ($I = 0$, $\eta_C = -1$, hence $\eta_G = -1$) decays to three and not to two. Both patterns are observed, and neither follows from energy or angular momentum. It fails, as it must, for the electromagnetic and weak interactions, which do not conserve isospin.

105.1.3 Time reversal

Why the operator must be antiunitary Classically, motion reversal sends $t \to -t$, leaves positions alone and reverses velocities; Newton's equation with a velocity-independent force is invariant. The quantum-mechanical implementation of the same idea is constrained far more tightly, and the constraint is Wigner's [710].

Theorem 105.10 (Time reversal is antiunitary)**.** *Let $\mathcal{T}$ implement motion reversal, in the sense that*

$$\mathcal{T}\hat{\boldsymbol{x}}\mathcal{T}^{-1} = +\hat{\boldsymbol{x}}\,, \qquad \mathcal{T}\hat{\boldsymbol{p}}\mathcal{T}^{-1} = -\hat{\boldsymbol{p}}\,, \tag{105.22}$$

on a system with a Hamiltonian bounded below. Then $\mathcal{T}$ is antiunitary.
Rests on Theorem 94.3.

Proof. *Derives Theorem 105.10.*

By Theorem 94.3 the operator is unitary and linear or antiunitary and antilinear; the argument is to exclude the first.

Two independent exclusions are available, and both are worth having. The first uses nothing beyond the canonical commutator of ordinary quantum mechanics, $[\hat{x}_j, \hat{p}_k] = \mathrm{i}\hbar\delta_{jk}\mathbb{1}$. Conjugating with $\mathcal{T}$ and using Equation (105.22),

$$\mathcal{T}\,[\hat{x}_j, \hat{p}_k]\,\mathcal{T}^{-1} = [\hat{x}_j, -\hat{p}_k] = -\mathrm{i}\hbar\delta_{jk}\mathbb{1}\,.$$

If $\mathcal{T}$ were linear the left-hand side would instead be $\mathcal{T}(\mathrm{i}\hbar\delta_{jk}\mathbb{1})\mathcal{T}^{-1} = +\mathrm{i}\hbar\delta_{jk}\mathbb{1}$, since a linear operator commutes with multiplication by a number. The two are inconsistent, so $\mathcal{T}$ is antilinear: it satisfies $\mathcal{T}(\mathrm{i}\hbar)\mathcal{T}^{-1} = -\mathrm{i}\hbar$, which reconciles them.

The second exclusion is the physically decisive one, because it invokes the spectrum rather than a commutator. Motion reversal must satisfy

$$\mathcal{T}\mathrm{e}^{-\mathrm{i}Ht/\hbar}\mathcal{T}^{-1} = \mathrm{e}^{+\mathrm{i}Ht/\hbar}\,, \tag{105.23}$$

since propagating forward for a time t and then reversing must equal reversing and then propagating backward. If $\mathcal{T}$ were linear, the left-hand side would be $\exp(-\mathrm{i}(\mathcal{T}H\mathcal{T}^{-1})t/\hbar)$

and Equation (105.23) would force $\mathcal{T}H\mathcal{T}^{-1} = -H$. Then for every eigenvalue E of H the value $-E$ would also be an eigenvalue, the spectrum would be unbounded below, and there would be no stable ground state. If instead $\mathcal{T}$ is antilinear, the i in the exponent changes sign as it is pulled through, the left-hand side is $\exp(+\mathrm{i}(\mathcal{T}H\mathcal{T}^{-1})t/\hbar)$, and Equation (105.23) gives

$$\mathcal{T}H\mathcal{T}^{-1} = +H\,, \tag{105.24}$$

with the spectrum intact. Antiunitarity then follows from Theorem 94.3. □

Remark 105.11 (Why P escapes). The same argument applied to space inversion produces no obstruction: $\mathcal{P}$ leaves t alone, so Equation (105.23) is replaced by $\mathcal{P}\mathrm{e}^{-\mathrm{i}Ht/\hbar}\mathcal{P}^{-1} = \mathrm{e}^{-\mathrm{i}Ht/\hbar}$, which a linear operator satisfies with $\mathcal{P}H\mathcal{P}^{-1} = H$. The asymmetry between P and T in this chapter — one has a conserved quantum number and selection rules, the other has neither — traces back to precisely this line, and not to any experimental fact.

Proposition 105.12 (No conserved quantum number for T)**.** *Let $\mathcal{T}$ be antiunitary with $\mathcal{T}\,|a\rangle = \eta_T\,|a\rangle$ for a nondegenerate state $|a\rangle$. Then η_T carries no physical information: it can be set to* 1 *by a choice of phase for $|a\rangle$.*

Rests on Theorem 105.10.

Proof. *Derives Proposition 105.12.*

Replace $|a\rangle$ by $|a'\rangle = \mathrm{e}^{\mathrm{i}\theta}\,|a\rangle$. Antilinearity gives $\mathcal{T}\,|a'\rangle = \mathrm{e}^{-\mathrm{i}\theta}\eta_T\,|a\rangle = \mathrm{e}^{-2\mathrm{i}\theta}\eta_T\,|a'\rangle$, so the eigenvalue is $\eta_T\mathrm{e}^{-2\mathrm{i}\theta}$ and $\theta = \frac{1}{2}\arg\eta_T$ makes it unity. □

There is a deeper way to say the same thing. A conserved quantum number comes from a Hermitian generator, by Noether's theorem or by the spectral theorem; an antiunitary operator has no Hermitian generator, because it is not part of any continuous one-parameter group. What T invariance gives instead is a relation between amplitudes, derived in Equation (105.25) below, and one genuine degeneracy theorem.

Kramers degeneracy

Theorem 105.13 (Kramers)**.** *Let H be T-invariant, $\mathcal{T}H\mathcal{T}^{-1} = H$, on a system of half-integer total angular momentum, for which $\mathcal{T}^2 = -\mathbb{1}$. Then every energy eigenvalue is at least twofold degenerate, and no perturbation that is itself T-invariant can lift the degeneracy [659].*

Rests on Theorem 105.10 and Equation (105.24)*.*

Proof. *Derives Theorem 105.13.*

On a system of angular momentum j the time-reversal operator may be written $\mathcal{T} = \mathrm{e}^{-\mathrm{i}\pi\hat{J}_y/\hbar}K$ with K complex conjugation in the $\hat{J}_z$ eigenbasis, since this is the unique antiunitary satisfying $\mathcal{T}\hat{\boldsymbol{J}}\mathcal{T}^{-1} = -\hat{\boldsymbol{J}}$. Squaring, the K's cancel and the rotation doubles: $\mathcal{T}^2 = \mathrm{e}^{-2\mathrm{i}\pi\hat{J}_y/\hbar} = (-1)^{2j}\mathbb{1}$, which is $-\mathbb{1}$ for half-integer j (Remark 94.6 is the same sign seen in a neutron interferometer).

Let $H\,|\psi\rangle = E\,|\psi\rangle$. Then $H\mathcal{T}\,|\psi\rangle = \mathcal{T}H\,|\psi\rangle = E\mathcal{T}\,|\psi\rangle$ — note that E is real, so antilinearity does not disturb it — and $\mathcal{T}\,|\psi\rangle$ has the same energy. It remains to show that it is a *different* state. Antiunitarity means $\langle\mathcal{T}\varphi|\mathcal{T}\psi\rangle = \langle\psi|\varphi\rangle$ for all φ, ψ. Take $\varphi = \mathcal{T}\psi$:

$$\big\langle \mathcal{T}^2\psi \big| \mathcal{T}\psi \big\rangle = \langle\psi|\mathcal{T}\psi\rangle\ .$$

The left side is $\langle -\psi|\mathcal{T}\psi\rangle = -\langle\psi|\mathcal{T}\psi\rangle$, so $\langle\psi|\mathcal{T}\psi\rangle = 0$: the two states are orthogonal. Finally, if V is a T-invariant perturbation, the same argument applies to $H + V$ at every order, so the pair stays degenerate. □

Kramers's own context was the paramagnetic rotation of light in crystals: an ion with an odd number of electrons has half-integer total angular momentum, and no electrostatic crystal field — however low its symmetry — can split its lowest level into singlets. Only a magnetic field can, because $\boldsymbol{B}$ is T-odd and breaks the hypothesis. The theorem thus explains, without any detailed calculation, why odd-electron and even-electron ions behave completely differently in a crystal, and it remains the standard tool for classifying magnetic ions.

Motion reversal is not the arrow of time It is worth stating plainly, because the confusion is common, that the T of this chapter has nothing to do with the second law of Chapter 54. The microscopic laws of electromagnetism, of the strong interaction and of gravity are exactly T-invariant, and the weak interaction violates T by a few parts in a thousand; yet entropy increases by factors of 10^{20} in ordinary processes. The second law is a statement about the overwhelming majority of *initial conditions* compatible with a given macroscopic description, together with the coarse-graining that defines that description. It would hold with the same force in a world where T were an exact symmetry, and it is not measurably strengthened by the 10^{-3} violation that the kaon exhibits. A theory of the arrow of time must explain a low-entropy initial condition; no amount of microscopic T violation does that.

Reciprocity, detailed balance, and why they prove less than they seem to

Proposition 105.14 (Reciprocity)**.** *If $\mathcal{T}H\mathcal{T}^{-1} = H$, the S-matrix satisfies*

$$\langle f|\, S\,|i\rangle = \langle \mathcal{T}i|\, S\,|\mathcal{T}f\rangle \ , \tag{105.25}$$

where $\mathcal{T}i$ denotes the state i with all momenta and spins reversed.

Rests on Theorem 105.10 and Equation (105.23).

Proof. *Derives Proposition 105.14.*

First an identity used repeatedly below. For an antiunitary $\mathcal{A}$ and any operator O,

$$\langle \mathcal{A}\varphi|\, O\,|\mathcal{A}\psi\rangle = \Big\langle \varphi\Big|\mathcal{A}^{-1}O\mathcal{A}\psi\Big\rangle^{*} = \langle\psi|\, \mathcal{A}^{-1}O^{\dagger}\mathcal{A}\,|\varphi\rangle^{*} \ , \tag{105.26}$$

which follows by writing $O\mathcal{A}\,|\psi\rangle = \mathcal{A}(\mathcal{A}^{-1}O\mathcal{A})\,|\psi\rangle$ and applying $\langle\mathcal{A}\varphi|\mathcal{A}\chi\rangle = \langle\chi|\varphi\rangle$.

Now T invariance of H implies, by Equation (105.23) extended to the interacting evolution operator, that $\mathcal{T}S\mathcal{T}^{-1} = S^{\dagger}$: reversing the motion turns the scattering operator into its inverse, which unitarity identifies with $S^{\dagger}$. Substituting $O = S$, $\mathcal{A} = \mathcal{T}$, $\varphi = i$, $\psi = f$ into Equation (105.26) and using $\mathcal{T}^{-1}S^{\dagger}\mathcal{T} = S$ gives $\langle \mathcal{T}i|\, S\,|\mathcal{T}f\rangle = \langle f|\, S\,|i\rangle^{**} = \langle f|\, S\,|i\rangle$. □

Equation (105.25) is the correct statement of "detailed balance": the amplitude for $i \to f$ equals that for the reversed process $\mathcal{T}f \to \mathcal{T}i$, with every momentum and every spin flipped. Two cautions attach to using it as a test.

First, at lowest order in a weak perturbation the test is empty. If $S = \mathbb{1} - \frac{\mathrm{i}}{\hbar}\int \mathrm{d}t\, H_w(t)$ to first order, then $|\langle f|\, S\,|i\rangle| = |\langle i|\, S\,|f\rangle|$ follows from the hermiticity of H_w alone, with no reference to T. A first-order process therefore cannot violate detailed balance whether or not T holds, and observing detailed balance in, say, a nuclear reaction and its inverse tests T only to the extent that higher orders matter.

Second, and more insidiously, a nonzero T-*odd* observable does not by itself establish T violation. The classic example is the triple correlation

$$D\,\frac{\langle \boldsymbol{J}\rangle}{J}\cdot\frac{\boldsymbol{p}_e\times\boldsymbol{p}_\nu}{E_eE_\nu/c^2} \tag{105.27}$$

in the decay of polarized neutrons, in which all three vectors reverse under motion reversal so that the product is odd. Yet a nonzero D can be generated by final-state interactions in a perfectly T-invariant theory. To see how, write an amplitude as a sum of two contributions carrying a CP-odd ("weak") phase ϕ and a CP-even ("strong", or Coulomb) rescattering phase δ,

$$\mathcal{M} = a_1 \mathrm{e}^{\mathrm{i}\delta_1}\mathrm{e}^{\mathrm{i}\phi_1} + a_2 \mathrm{e}^{\mathrm{i}\delta_2}\mathrm{e}^{\mathrm{i}\phi_2}\,, \tag{105.28}$$

with $a_{1,2}$ real. A T-odd correlation is proportional to $\mathrm{Im}(\mathcal{M}_1\mathcal{M}_2^*) = a_1 a_2 \sin(\delta_1-\delta_2+\phi_1-\phi_2)$, which is nonzero when $\delta_1 \neq \delta_2$ even if $\phi_1 = \phi_2$. In neutron decay the Coulomb interaction between the outgoing proton and electron supplies $\delta_1 - \delta_2 \sim \alpha m_e c/p_e$, faking $D \sim 10^{-5}$; the measured value, $D = (-1.2 \pm 2.0)\times 10^{-4}$ [2427], is not yet sensitive to the fake, let alone to anything beyond it. The lesson generalizes: a genuine T test must either compare a process with its motion reverse (Section 105.4) or measure a static property of a nondegenerate state (Section 105.4.3), where no final state exists to rescatter.

Time reversal on the fields On a Dirac field, antiunitarity plus the requirement that solutions map to solutions gives

$$\mathcal{T}\psi(ct,\boldsymbol{x})\mathcal{T}^{-1} = \eta_T\, \mathrm{T_D}\, \psi(-ct,\boldsymbol{x})\,, \qquad \mathrm{T_D^{-1}}\gamma^{\mu *}\mathrm{T_D} = \gamma_\mu\,, \tag{105.29}$$

with $\mathrm{T_D} = \gamma^1\gamma^3$ in the Dirac representation; the complex conjugate on γ^μ is the signature of the antilinearity, exactly as in Theorem 105.10. For the gauge field the assignment is fixed by the classical behaviour of the sources: reversing all motions reverses currents but not charges, so

$$\mathcal{T}A^0(ct,\boldsymbol{x})\mathcal{T}^{-1} = A^0(-ct,\boldsymbol{x})\,, \quad \mathcal{T}A^i(ct,\boldsymbol{x})\mathcal{T}^{-1} = -A^i(-ct,\boldsymbol{x})\,, \tag{105.30}$$

whence $\boldsymbol{E} \to +\boldsymbol{E}$ and $\boldsymbol{B} \to -\boldsymbol{B}$. That $\boldsymbol{B}$ is T-odd and $\boldsymbol{E}$ is T-even is used at three separate places below — in Kramers degeneracy, in the electric dipole moment argument, and in the sign of the magnetic moment under CPT — and it is worth remembering as a picture: a magnetic field is made by currents, and currents reverse.

Carrying Equation (105.29) through the bilinears as before completes Table 105.1.

Table 105.1: The five Dirac bilinears under the discrete operations. An entry x_P means the argument is $(ct,-\boldsymbol{x})$, x_T means $(-ct,\boldsymbol{x})$, and $-x$ means $(-ct,-\boldsymbol{x})$; a lowered index means the factor $(+1,-1,-1,-1)$ componentwise. The last column, the product of the other three, is the content of the CPT theorem at the level of a single bilinear: the sign is $(-1)^n$ with n the number of Lorentz indices, and *nothing else about the bilinear matters.*

bilinear	P	C	T	$\Theta = CPT$
$S = \bar\psi\psi$	$+S(x_P)$	$+S$	$+S(x_T)$	$+S(-x)$
$\tilde S = \bar\psi\gamma^5\psi$	$-\tilde S(x_P)$	$+\tilde S$	$-\tilde S(x_T)$	$+\tilde S(-x)$
$V^\mu = \bar\psi\gamma^\mu\psi$	$+V_\mu(x_P)$	$-V^\mu$	$+V_\mu(x_T)$	$-V^\mu(-x)$
$A^\mu = \bar\psi\gamma^\mu\gamma^5\psi$	$-A_\mu(x_P)$	$+A^\mu$	$+A_\mu(x_T)$	$-A^\mu(-x)$
$T^{\mu\nu} = \bar\psi\sigma^{\mu\nu}\psi$	$+T_{\mu\nu}(x_P)$	$-T^{\mu\nu}$	$-T_{\mu\nu}(x_T)$	$+T^{\mu\nu}(-x)$

The last column will be derived and used in Section 105.5.1; it is written here because it costs nothing once the other three exist, and because seeing it early makes the shape of the CPT theorem obvious. Under $\Theta = CPT$ a bilinear returns to itself, at the reflected point, with the sign $(-1)^n$ where n counts its Lorentz indices — and with no memory of whether it was scalar or pseudoscalar, vector or axial. That insensitivity is why the product survives when the factors do not.

105.2 Parity violation

105.2.1 The tau–theta puzzle and the Lee–Yang question

By 1955 two strange mesons were known that agreed in every property that could then be measured and disagreed in one that could not be explained. The θ^+ decayed to $\pi^+\pi^0$; the τ^+ decayed to $\pi^+\pi^+\pi^-$. Their masses agreed to better than a per cent and their lifetimes to better than the experimental errors, which for two unrelated particles would be an extraordinary coincidence. But the parities of the two final states are opposite.

Proposition 105.15 (Parity of an n-pion state)**.** *A state of n pions with total orbital angular momentum content L_{tot} (the sum of the relative orbital angular momenta) has parity*

$$\eta_P(n\pi) = (-1)^n\,(-1)^{L_{\mathrm{tot}}}\,. \tag{105.31}$$

Rests on Phenomenon 105.5 and Equation (105.4).

Proof. *Derives Proposition 105.15.*

Immediate from multiplicativity, the pseudoscalar assignment $\eta_P(\pi) = -1$ of Phenomenon 105.5, and Equation (105.4) applied to each relative coordinate. □

For the θ^+, the parent is spinless (its decay products are two spinless particles, and the angular distribution is isotropic), so $L = 0$ and $\eta_P = (-1)^2 = +1$. For the τ^+, if all relative angular momenta vanish then $\eta_P = (-1)^3 = -1$. Two particles of the same mass and the same lifetime were therefore decaying into final states of opposite parity — impossible if parity is conserved, unless the two are genuinely different particles that happen to be degenerate, or unless the τ decay carries $L_{\mathrm{tot}} = 1$ and the parent has spin ≥ 1.

The Dalitz plot It was Dalitz's construction that closed the second escape [1016]. For a three-body decay of a particle at rest into three pions of kinetic energies T_1, T_2, T_3 with $T_1 + T_2 + T_3 = Q$ fixed, plot the event as the point of an equilateral triangle of height Q whose perpendicular distances to the three sides are T_1, T_2, T_3. The kinematically allowed region is a closed curve inside the triangle, and — this is the content of the construction — the *density of points is uniform whenever the decay matrix element is constant.*

Derivation. The three-body phase space for a parent of mass M decaying at rest is

$$\mathrm{d}\Gamma \propto |\mathcal{M}|^2\;\frac{\mathrm{d}^3p_1}{E_1}\frac{\mathrm{d}^3p_2}{E_2}\frac{\mathrm{d}^3p_3}{E_3}\,\delta^3\!\left(\textstyle\sum_a \boldsymbol{p}_a\right)\delta\!\left(Mc^2 - \textstyle\sum_a E_a\right)\,.$$

Do the $\boldsymbol{p}_3$ integral with the momentum delta function, write $\mathrm{d}^3p_1 = p_1^2\mathrm{d}p_1\mathrm{d}\Omega_1$ and $\mathrm{d}^3p_2 = p_2^2\mathrm{d}p_2\,\mathrm{d}\cos\theta_{12}\,\mathrm{d}\varphi$, and use the energy delta function to fix $\cos\theta_{12}$. Since $E_3^2/c^2 = \boldsymbol{p}_1^2 + \boldsymbol{p}_2^2 + 2p_1p_2\cos\theta_{12} + m_3^2c^2$, one has $\mathrm{d}\cos\theta_{12} = E_3\mathrm{d}E_3/(c^2p_1p_2)$ at fixed p_1, p_2, and the three factors of E_a in the denominators cancel against $p_a^2\mathrm{d}p_a = E_a\mathrm{d}E_a p_a/c^2$. Collecting,

$$\mathrm{d}\Gamma \propto |\mathcal{M}|^2\;\mathrm{d}E_1\,\mathrm{d}E_2\,, \tag{105.32}$$

up to the overall angular factors that integrate to a constant. Since $T_a = E_a - m_\pi c^2$, the density in the (T_1, T_2) plane — and hence in the Dalitz triangle, which is an affine image of it — is $|\mathcal{M}|^2$ and nothing else. □

The plot therefore reads off the matrix element directly. If the decay proceeded with a unit of relative orbital angular momentum, the amplitude would vanish at the centre of the plot, where all three momenta are small, and the population there would be depleted; a $J = 1$ or $J = 2$ parent would leave a characteristic pattern of zeros along the boundary. The observed τ plot was flat. The three pions were in a state of no relative orbital angular momentum, the parent had $J = 0$, and its parity was -1: no escape.

Lee and Yang's question Lee and Yang responded not with a mechanism but with an audit [1091]. They asked what the experimental evidence for parity conservation actually was, and found that all of it came from strong and electromagnetic processes — Laporte's rule, nuclear level schemes, the selection rules of Proposition 105.1. In these sectors the evidence was overwhelming, bounding any parity-violating admixture in the nuclear force at the level of 10^{-7} in amplitude. In the weak interaction there was none at all: not one experiment had ever been designed to detect a parity-odd observable in a weak decay. The θ–τ puzzle was, on this reading, not a puzzle about two particles but a signal that the weak interaction does not conserve parity, and that θ and τ are one particle, the K^+, decaying two ways.

Lee and Yang then did the thing that makes the paper a model of method: they listed the experiments that would settle it. Each of their proposals is the expectation value of a pseudoscalar, since a pseudoscalar is precisely a quantity that must vanish if P is a symmetry:

- the correlation $\langle \boldsymbol{J} \rangle \cdot \boldsymbol{p}_e$ between the polarization of an oriented nucleus and the momentum of the emitted electron — an axial vector dotted into a polar vector;
- the longitudinal polarization $\langle \boldsymbol{s} \rangle \cdot \boldsymbol{p}$ of a lepton emitted in a weak decay — the same structure;
- in the chain $\pi \to \mu\nu$, $\mu \to e\nu\bar{\nu}$, the correlation between the muon direction and the electron direction, which is nonzero only if the muon is produced polarized.

All three were done within a year, and all three gave the maximum possible answer.

105.2.2 The experimental collapse of 1957

Phenomenon 105.16 (Beta emission from oriented nuclei is asymmetric)**.** Nuclei of ^{60}Co polarized in a magnetic field at a temperature of about 0.01 K emit beta electrons preferentially opposite to the nuclear spin. The angular distribution is

$$W(\theta) = 1 + A_\beta \, \frac{\langle \boldsymbol{J} \rangle}{J} \cdot \frac{\boldsymbol{v}_e}{c} = 1 + A_\beta \, \frac{\langle J \rangle}{J} \, \frac{v_e}{c} \cos\theta \,, \tag{105.33}$$

with A_β consistent with -1, and the asymmetry disappears as the sample warms and the polarization is lost [1133]. The apparatus is Chapter 110; see Phenomenon 110.10.

Rests on Equations (105.34) *and* (105.37).

Derivation. *Derives Phenomenon 105.16.*

That $A_\beta \neq 0$ is impossible in a parity-conserving theory needs one line. $\boldsymbol{J}$ is an axial vector and $\boldsymbol{v}_e$ a polar vector, so their scalar product is a pseudoscalar and changes sign under $\mathcal{P}$. If $[\mathcal{P}, H] = 0$ the rate is $\mathcal{P}$-invariant, so $W(\theta) = W(\pi - \theta)$ and the coefficient of $\cos\theta$ must vanish.

That the observed value is the *maximum* takes a little more. The transition $^{60}\mathrm{Co}(5^+) \to {}^{60}\mathrm{Ni}^*(4^+)$ is allowed and pure Gamow–Teller: the leptons carry off one unit of angular momentum in their spins and none in orbital motion. If the emitted antineutrino is purely right-handed and the electron predominantly left-handed — the content of the $V - A$ structure below — then conservation of angular momentum along the nuclear spin axis forces the electron to go backwards. Quantitatively, for a pure Gamow–Teller transition with $\Delta J = -1$ the $V - A$ prediction is $A_\beta = -1$ exactly, and the factor v_e/c in Equation (105.33) is the longitudinal polarization of an electron of speed v_e, derived at Equation (105.37). □

Wu, Ambler, Hayward, Hoppes and Hudson polarized their source by adiabatic demagnetization of a cerium magnesium nitrate crystal into which a thin layer of ^{60}Co had been

grown, and monitored the polarization independently through the anisotropy of the gamma rays from the daughter. The two signals — gamma anisotropy and beta asymmetry — decayed together over some six minutes as the crystal warmed, which is the control that makes the result an observation of parity violation rather than of an instrumental drift.

Two further experiments were reported in the same issue of the same journal. Garwin, Lederman and Weinrich [1110] and Friedman and Telegdi [1109] used the chain $\pi^+ \to \mu^+\nu_\mu$ followed by $\mu^+ \to e^+\nu_e\bar{\nu}_\mu$. If parity is conserved the muon from a spinless pion cannot be longitudinally polarized, since $\langle \boldsymbol{s}_\mu \rangle \cdot \boldsymbol{p}_\mu$ is a pseudoscalar; and even if it were, the electron from its decay could not remember the direction. Both correlations were found: the positrons came out preferentially along the muon spin, with an asymmetry that also measured the muon magnetic moment to be very close to the Dirac value $g = 2$ (the quantity refined to twelve digits in Chapter 112). Phenomenon 110.13 records the modern statement.

Phenomenon 105.17 (The neutrino is left-handed)**.** The neutrino emitted in the electron capture $^{152}\mathrm{Eu}^m + e^- \to {}^{152}\mathrm{Sm}^* + \nu_e$ has helicity -1 within the experimental accuracy [1148]. The apparatus, which used resonant scattering of the daughter gamma ray to select neutrinos emitted backwards, is Chapter 98; see Phenomenon 98.2.

Rests on Equations (105.34) *and* (105.37)*.*

Derivation. Derives Phenomenon 105.17.

The charged-current interaction Equation (105.34) contains the neutrino field only through $(1-\gamma^5)\,\psi_\nu = 2P_L\psi_\nu$, so a neutrino leaving a weak vertex is emitted in a state of definite *chirality* — the left-handed one — whatever the nucleus does. Chirality is not helicity for a massive particle, and the exact relation is Equation (105.37): the longitudinal polarization of a fermion emitted by a $V-A$ current is $-v/c$, so helicity -1 is predicted only up to

$$1 - \frac{v}{c} = 1 - \frac{|\boldsymbol{p}|\,c}{E} = 1 - \sqrt{1 - \frac{m_\nu^2 c^4}{E^2}} \approx \frac{m_\nu^2 c^4}{2E^2}\,.$$

The neutrino of this decay carries $E \approx 0.90\,\mathrm{MeV}$, and the laboratory bound on the neutrino mass is of order $1\,\mathrm{eV}/c^2$ (Chapter 103), so the correction is at most 6×10^{-13}. The prediction is therefore helicity -1 to twelve significant figures, and the measurement — which resolves the sign and roughly its first digit — tests the chiral structure of Equation (105.34) and nothing about the neutrino mass. What the experiment adds to the theory is the *sign*: $V-A$ with the opposite relative sign, $V+A$, would predict $+1$ just as sharply. □

Goldhaber, Grodzins and Sunyar's measurement is the sharpest of the three, because helicity is not merely a nonzero pseudoscalar but a *saturated* one: $\langle \boldsymbol{s} \rangle \cdot \hat{\boldsymbol{p}} = -\hbar/2$ is as large as it can be. Parity is not slightly violated in the weak interaction; it is violated maximally.

The V–A structure The theoretical form that accommodates all of this was proposed by Feynman and Gell-Mann [1144] and independently by Sudarshan and Marshak [1162], with the two-component neutrino of Landau [1118], Lee and Yang [1121] and Salam [1130] as its kinematic core. The charged-current weak interaction at low energy is a product of two currents, each of the form $V-A$:

$$\mathcal{H}_w = \frac{G_F}{\sqrt{2}}\left[\bar{\psi}_p\gamma^\mu\left(g_V - g_A\gamma^5\right)\psi_n\right]\left[\bar{\psi}_e\gamma_\mu\left(1-\gamma^5\right)\psi_\nu\right] + \mathrm{h.c.}\,, \tag{105.34}$$

with $g_A/g_V = -1.2754 \pm 0.0013$ for the nucleon [2427] — the deviation from -1 being a strong-interaction effect, since the quark-level current is pure $V-A$.

Remark 105.18 (The SI dimension of Fermi's constant)*.* Fermi's constant [738] is quoted throughout the literature in the natural-unit form

$$\frac{G_F}{(\hbar c)^3} = (1.166\,378\,8 \pm 0.000\,000\,6)\times10^{-5}/\mathrm{GeV}^2\,, \tag{105.35}$$

which is a statement about $G_F/(\hbar c)^3$ and not about G_F. The constant itself, as it appears in Equation (105.34) where the four fields carry $1/\mathrm{m}^6$ between them and $\mathcal{H}_w$ is an energy density, has the SI value

$$G_F = (1.435\,851\,2 \pm 0.000\,000\,7) \times 10^{-62}\,\mathrm{J\,m^3}\,, \tag{105.36}$$

obtained from Equation (105.35) by multiplying by $(\hbar c)^3 = (3.161\,526\,77 \times 10^{-26}\,\mathrm{J\,m})^3$ and converting GeV^2 to J^2. Both numbers describe the same physics; the first is the one to compare with a textbook and the second is the one that makes Equation (105.34) dimensionally correct. The smallness of the weak interaction at low energy is the smallness of Equation (105.36) measured against $\hbar c$ and a length: at a nuclear scale $r \sim 1\,\mathrm{fm}$ the ratio $G_F/(\hbar c\, r^2) \approx 10^{-6}$, which is the familiar suppression.

Proposition 105.19 ($V - A$ violates parity maximally)**.** *Under $\mathcal{P}$ the chiral projector $P_L = \frac{1}{2}(1-\gamma^5)$ is mapped to $P_R = \frac{1}{2}(1+\gamma^5)$, so the parity image of Equation (105.34) is an interaction with no overlap with the original. The parity-violating part of the leptonic current is as large as the parity-conserving part.*

Rests on Equations (105.7), (105.11) and (105.34).

Proof. *Derives Proposition 105.19.*

From Equation (105.7), $\mathcal{P}\left[\bar\psi_1\gamma^\mu P_L\psi_2\right]\mathcal{P}^{-1} = \bar\psi_1\gamma^0\gamma^\mu\frac{1}{2}(1-\gamma^5)\gamma^0\psi_2$ evaluated at x_P. Because γ^5 anticommutes with γ^0, moving the γ^0 through gives $\gamma^0\gamma^5\gamma^0 = -\gamma^5$ and hence $P_L \to P_R$; the index is lowered as in Equation (105.11). The transformed interaction couples right-handed leptons and is orthogonal, in the space of possible couplings, to the one we started from. Writing $V - A$ as V and A pieces of equal magnitude makes the same point: parity violation is proportional to the V–A interference term, which for coefficients of equal magnitude is as large as either term alone. □

The physical content of the projector is helicity.

Proposition 105.20 (Chirality and helicity)**.** *Let $u(p,s)$ solve the free Dirac equation with $\boldsymbol{p}$ along $\hat{\boldsymbol{z}}$. Then*

$$\langle P_L\rangle - \langle P_R\rangle = -\frac{|\boldsymbol{p}|\,c}{E} = -\frac{v}{c}\,, \tag{105.37}$$

so a $V - A$ current emits a fermion with longitudinal polarization $-v/c$ and an antifermion with $+v/c$; in the ultrarelativistic limit the helicity is ∓ 1 exactly, and for a strictly massless fermion chirality and helicity coincide.

Rests on Equations (100.5) and (105.34).

Proof. *Derives Proposition 105.20.*

The helicity operator is $h = \boldsymbol{\Sigma}\cdot\hat{\boldsymbol{p}}$ with $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma},\boldsymbol{\sigma})$. The Dirac equation $(\gamma^\mu p_\mu - mc)u = 0$ gives, in the Dirac representation with $\boldsymbol{p} = p\hat{\boldsymbol{z}}$, the lower two components in terms of the upper: $u = N(\chi,\ c\,\boldsymbol{\sigma}\cdot\boldsymbol{p}\,\chi/(E+mc^2))$ with $N^2 = (E+mc^2)/(2mc^2)$ fixing $\bar u u = 1$. Take χ the eigenvector of σ_z with eigenvalue $+1$. Then $\gamma^5 u$, which exchanges the upper and lower two-spinors, gives

$$u^\dagger\gamma^5 u = 2N^2\,\frac{cp}{E+mc^2} = \frac{cp}{mc^2}\,,$$

the factor 2 being the two equal cross terms between the upper and lower two-spinors. It is $u^\dagger\gamma^5 u$ and not the bilinear $\bar u\gamma^5\gamma^0 u$ that is wanted, the two differing by a sign: $\gamma^0\gamma^5\gamma^0 = -\gamma^5$, so $\bar u\gamma^5\gamma^0 u = -u^\dagger\gamma^5 u$. Likewise $u^\dagger u = N^2\,(1 + c^2p^2/(E+mc^2)^2) = E/(mc^2)$, the bracket collapsing to $2E/(E+mc^2)$ because $c^2p^2 = (E-mc^2)(E+mc^2)$. The ratio is $cp/E = v/c$, and since $\langle P_L\rangle - \langle P_R\rangle = -\langle\gamma^5\rangle$ normalized to $u^\dagger u$, Equation (105.37) follows. As $m \to 0$, $E \to cp$ and the ratio tends to 1: P_L then projects exactly onto helicity -1. □

Equation (105.37) is the factor v_e/c in Equation (105.33), so the shape of Wu's angular distribution, the polarization of the beta electron, and the helicity of the neutrino are three faces of a single structure. The neutrino, being (very nearly) massless, saturates it. The theory of the current itself — why $V-A$, and how it arises from an SU(2) gauge interaction with a left-handed doublet — belongs to Chapter 101 and Chapter 104.

105.2.3 Parity violation as a precision tool

Once parity violation is established it stops being a discovery and becomes an instrument, for a reason worth stating carefully. A weak amplitude is far smaller than an electromagnetic one at ordinary energies, so a weak *rate* is invisible under the electromagnetic background. But a parity-odd observable receives no contribution at all from the electromagnetic amplitude squared: its leading term is the *interference* between the weak and electromagnetic amplitudes, which is linear in the small quantity. Measuring a parity-violating asymmetry therefore buys a factor $\mathcal{M}_\gamma/\mathcal{M}_Z$ of sensitivity relative to measuring a weak rate.

Proposition 105.21 (Size of a parity-violating asymmetry)**.** *For electron scattering at squared momentum transfer Q^2, the ratio of the neutral-current to the photon-exchange amplitude is*

$$\frac{\mathcal{M}_Z}{\mathcal{M}_\gamma} \sim \frac{G_F Q^2 c^2}{4\sqrt{2}\,\pi\alpha\,(\hbar c)^3} \approx 9\times 10^{-5}\,\frac{Q^2c^2}{1\,\mathrm{GeV}^2}\,, \tag{105.38}$$

so the parity-violating asymmetry $A_{PV}=(\sigma_R-\sigma_L)/(\sigma_R+\sigma_L)$ is of order 10^{-4} at $Q^2c^2=1\,\mathrm{GeV}^2$ and of order 10^{-7} at the Q^2 of a low-energy fixed-target experiment.

Rests on Equations (105.34) and (105.35).

Proof. *Derives Proposition 105.21.*

The photon-exchange amplitude carries $4\pi\alpha\hbar c/Q^2$ from the propagator and coupling; the Z-exchange amplitude at $Q^2 \ll m_Z^2c^2$ is contact-like and carries $G_F/(\sqrt{2}(\hbar c)^3)$ in the same normalization. The ratio is Equation (105.38), and inserting $G_F/(\hbar c)^3 = 1.1664\times10^{-5}/\mathrm{GeV}^2$ together with $\alpha = 1/137.036$ gives $1.1664\times10^{-5}/(4\sqrt{2}\pi\times 7.2974\times10^{-3}) = 8.99\times10^{-5}$ per GeV^2. □

Deep inelastic scattering: Prescott 1978 Prescott and collaborators scattered longitudinally polarized electrons of 16.2–22.2 GeV from deuterium at SLAC and reversed the beam helicity at random, pulse to pulse [1625]. They measured

$$\frac{A_{PV}}{Q^2c^2} = (-9.5\pm1.6)\times10^{-5}\,1/\mathrm{GeV}^2\,, \tag{105.39}$$

in magnitude exactly the estimate of Equation (105.38). The electroweak analysis writes

$$\frac{A_{PV}}{Q^2c^2} = a_1 + a_2\,\frac{1-(1-y)^2}{1+(1-y)^2}\,, \qquad y=\frac{\nu}{E}\,, \tag{105.40}$$

with $a_1 \propto 1-\frac{20}{9}\sin^2\theta_W$ and $a_2 \propto 1-4\sin^2\theta_W$. The second coefficient nearly vanishes near $\sin^2\theta_W \approx 0.23$, so it is the *y-dependence* of the asymmetry — the fact that it is nearly flat — that fixes the weak mixing angle rather than its magnitude. Prescott obtained $\sin^2\theta_W = 0.20\pm0.03$, in agreement with the value inferred from neutrino scattering, and thereby confirmed the SU(2) × U(1) structure of Chapter 104 against the alternatives then in play. See Phenomenon 110.17.

Atomic parity violation In an atom, the exchange of a Z between an electron and the nucleus adds to the Hamiltonian a term that is parity-odd, mixing S and P states and giving a normally forbidden $E1$ transition a small amplitude. The nuclear-spin-independent part is governed by the weak charge

$$Q_W = Z\left(1 - 4\sin^2\theta_W\right) - N\,, \tag{105.41}$$

and the induced $E1$ amplitude grows roughly as Z^3: one power from $Q_W \approx -N \propto Z$, and two from the relativistic enhancement of the electron density at the nucleus. Heavy atoms are therefore enormously favoured, and caesium ($Z = 55$) is the practical optimum because its single valence electron makes the atomic structure calculable to a per cent.

Wood and collaborators measured the parity-violating $6S \to 7S$ amplitude in caesium and extracted [1946]

$$Q_W(^{133}\mathrm{Cs}) = -72.06 \pm 0.28 \pm 0.34\,, \tag{105.42}$$

the two errors being experimental and atomic-theoretical. This is a measurement of the weak interaction made in a tabletop apparatus at an energy of a few electronvolts, and it agrees with the value obtained at LEP at 91 GeV — the same weak mixing angle at two energies about ten decades apart, since 91 GeV is 3×10^{10} times a few electronvolts. The agreement is not automatic: $\sin^2\theta_W$ runs with the scale at which it is probed, and the two determinations agree only once that running is applied. See Phenomenon 110.16.

The proton's weak charge, and a nuclear skin Two modern measurements exploit the near-cancellation in Equation (105.41). For a proton, $Z = 1$ and $N = 0$, so $Q_W^p = 1 - 4\sin^2\theta_W \approx 0.0719$: a small number, and a fractional measurement of it is a much finer measurement of $\sin^2\theta_W$. Differentiating, $\delta(\sin^2\theta_W) = \tfrac{1}{4}Q_W^p\,\delta Q_W^p/Q_W^p$, so a 6 % determination of Q_W^p gives $\delta\sin^2\theta_W \approx 0.001$. The Qweak experiment measured an asymmetry of -226.5 ± 9.3 parts per billion at $Q^2c^2 = 0.0248\,\mathrm{GeV}^2$ and obtained [2301]

$$Q_W^p = 0.0719 \pm 0.0045\,, \tag{105.43}$$

in agreement with the Standard Model. See Phenomenon 110.19.

For a heavy nucleus the same cancellation acts the other way. Since $Q_W^p \approx 0.07$ while $Q_W^n = -1$, the Z couples almost exclusively to neutrons, and the parity-violating asymmetry in elastic electron scattering measures the *neutron* distribution while ordinary electron scattering measures the proton distribution. The difference of the two radii is the neutron skin. PREX-2 measured $A_{PV} = 550 \pm 16 \pm 8$ parts per billion on ^{208}Pb and obtained [2370]

$$R_n - R_p = (0.283 \pm 0.071)\,\mathrm{fm}\,. \tag{105.44}$$

The skin thickness is controlled by the density dependence of the nuclear symmetry energy, which is the same quantity that sets the radius of a neutron star, so a parity-violating asymmetry measured on a lead target in a Virginia accelerator hall constrains the equation of state used in Chapter 49. It is a fair example of what a symmetry-violating observable buys: an otherwise inaccessible distribution, extracted cleanly because the electromagnetic background cannot contribute to it.

105.3 CP violation

105.3.1 The neutral kaon system

Landau's proposal, made immediately after the collapse of parity, was that the true symmetry is the product CP [1118]. It is an attractive suggestion: C exchanges a left-handed neutrino

for a left-handed antineutrino, which does not exist; P exchanges it for a right-handed neutrino, which does not exist either; but CP exchanges it for a right-handed antineutrino, which does. Maximal violation of P and of C separately is exactly compatible with exact CP, and for seven years it appeared to be the truth.

The system in which it failed is the one Gell-Mann and Pais had analysed for a different purpose in 1955 [1067], and it remains the most instructive two-state problem in physics.

Two states, two bases The strong interaction produces states of definite strangeness. Write $|K^0\rangle$ for the state of strangeness $S = +1$ (quark content $d\bar{s}$) and $\left|\bar{K}^0\right\rangle$ for $S = -1$ ($\bar{d}s$); they are distinct particles, since strangeness is conserved by the strong interaction that makes them, and each is the antiparticle of the other. Fix phases by

$$\mathcal{CP}\left|K^0\right\rangle = \left|\bar{K}^0\right\rangle , \qquad \mathcal{CP}\left|\bar{K}^0\right\rangle = \left|K^0\right\rangle , \tag{105.45}$$

which is consistent because $(\mathcal{CP})^2 = \mathbb{1}$ on these states. (The opposite sign convention, $\mathcal{CP}\,|K^0\rangle = -\left|\bar{K}^0\right\rangle$, is equally common and changes nothing physical; it interchanges the roles of the two combinations below.) The CP eigenstates are then

$$|K_1\rangle = \frac{|K^0\rangle + \left|\bar{K}^0\right\rangle}{\sqrt{2}} , \quad |K_2\rangle = \frac{|K^0\rangle - \left|\bar{K}^0\right\rangle}{\sqrt{2}} , \tag{105.46}$$

with $CP = +1$ and $CP = -1$ respectively.

The weak interaction, which destroys them, does not conserve strangeness, so it is $|K_1\rangle$ and $|K_2\rangle$ — not $|K^0\rangle$ and $\left|\bar{K}^0\right\rangle$ — that have definite decay properties if CP holds. And the two available final states have opposite CP:

$$\eta_{CP}(\pi\pi) = +1 , \qquad \eta_{CP}(3\pi^0) = -1 . \tag{105.47}$$

For $\pi\pi$ from a spinless parent, $L = 0$, so $\eta_P = (-1)^2(-1)^0 = +1$ by Equation (105.31) and $\eta_C = +1$ for $\pi^0\pi^0$ or, by Equation (105.18) applied to the charge-conjugate pair, for $\pi^+\pi^-$ in an $L = 0$ state. For $3\pi^0$, $\eta_P = (-1)^3 = -1$ and $\eta_C = +1$. Hence if CP is exact, $K_1 \to \pi\pi$ is allowed and $K_2 \to \pi\pi$ is forbidden, and K_2 must decay to three pions or semileptonically.

That is a strong prediction, because three pions barely fit: with $m_K c^2 = (497.611 \pm 0.013)\,\mathrm{MeV}$ [2427] and $3m_\pi c^2 \approx 405\,\mathrm{MeV}$, the released kinetic energy is only about 90 MeV, and the resulting phase-space suppression must make K_2 long-lived. Gell-Mann and Pais predicted exactly this: a second neutral kaon, of the same mass to within the weak interaction's accuracy, with a lifetime orders of magnitude longer. Lande, Booth, Impeduglia, Lederman and Chinowsky found it at the Brookhaven Cosmotron [1090]. The measured lifetimes are

$$\tau_S = (8.954 \pm 0.004) \times 10^{-11}\,\mathrm{s} , \qquad \tau_L = (5.116 \pm 0.021) \times 10^{-8}\,\mathrm{s} , \tag{105.48}$$

a ratio of 571 [2427].

The effective Hamiltonian To do better than the symmetry argument one needs the dynamics of an unstable two-state system, and the only honest way to obtain it is to start from the full unitary problem and eliminate what is not being watched. This is the Wigner–Weisskopf construction, and it is worth carrying out because it is also the prototype of the open-system formalism of Chapter 87 and because the non-Hermitian matrix that comes out of it is otherwise mysterious.

Proposition 105.22 (The two-state effective generator)**.** *Let the Hilbert space be spanned by* $|K^0\rangle, \left|\bar{K}^0\right\rangle$ *and a continuum of decay states* $|n\rangle$ *of energy* E_n*, with a weak perturbation* H_w *connecting the two sectors. On time scales long compared with the memory time of the decay channels — of order* $\hbar$ *divided by the width of the accessible continuum — the amplitudes* $a(t), b(t)$ *of the two kaon states obey*

$$\mathrm{i}\hbar \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} a \\ b \end{pmatrix} = \mathbf{H}\begin{pmatrix} a \\ b \end{pmatrix}, \qquad \mathbf{H} = \mathbf{M} - \frac{\mathrm{i}\hbar}{2}\mathbf{\Gamma}, \tag{105.49}$$

where $\mathbf{M}$ *and* $\mathbf{\Gamma}$ *are Hermitian* 2×2 *matrices,*

$$M_{ij} = m_K c^2 \delta_{ij} + \langle i|\, H_w\, |j\rangle + \mathrm{P}\sum_n \frac{\langle i|\, H_w\, |n\rangle \langle n|\, H_w\, |j\rangle}{m_K c^2 - E_n}, \tag{105.50}$$

$$\Gamma_{ij} = \frac{2\pi}{\hbar}\sum_n \langle i|\, H_w\, |n\rangle \langle n|\, H_w\, |j\rangle\ \delta\!\left(m_K c^2 - E_n\right), \tag{105.51}$$

P *denoting the principal value.* $\mathbf{M}$ *has the SI dimension of energy and* $\mathbf{\Gamma}$ *of inverse time.*
Rests on Proposition 17.42.

Proof. *Derives Proposition 105.22.*

Write $|\psi(t)\rangle = a\,|K^0\rangle + b\left|\bar{K}^0\right\rangle + \sum_n c_n\,|n\rangle$ and project the Schrödinger equation $\mathrm{i}\hbar\partial_t\,|\psi\rangle = (H_0 + H_w)\,|\psi\rangle$ onto each sector. For the continuum,

$$\mathrm{i}\hbar \dot{c}_n = E_n c_n + \sum_j \langle n|\, H_w\, |j\rangle\ \psi_j(t),$$

with $\psi_1 = a$, $\psi_2 = b$. Integrate with $c_n(0) = 0$:

$$c_n(t) = -\frac{\mathrm{i}}{\hbar}\sum_j \langle n|\, H_w\, |j\rangle \int_0^t \mathrm{d}t'\, \mathrm{e}^{-\mathrm{i}E_n(t-t')/\hbar}\psi_j(t').$$

Substituting into the equation for the two-dimensional sector gives an exact integro-differential equation with a memory kernel. The Wigner–Weisskopf approximation is to note that the kernel is sharply peaked in $t - t'$ on the scale $\hbar/\Delta E$ set by the width of the continuum, which for a hadronic decay is far shorter than the kaon lifetime, so that $\psi_j(t')$ may be replaced by $\psi_j(t)\mathrm{e}^{\mathrm{i}m_K c^2 (t-t')/\hbar}$ and the upper limit sent to infinity. The remaining integral is

$$\int_0^\infty \mathrm{d}s\, \mathrm{e}^{\mathrm{i}\left(m_K c^2 - E_n + \mathrm{i}0\right)s/\hbar} = \mathrm{i}\hbar\,\frac{1}{m_K c^2 - E_n + \mathrm{i}0} = \mathrm{i}\hbar\left[\mathrm{P}\frac{1}{m_K c^2 - E_n} - \mathrm{i}\pi\delta\!\left(m_K c^2 - E_n\right)\right],$$

using the Sokhotski–Plemelj identity, Proposition 17.42. Collecting the real and imaginary parts reproduces Equations (105.50) and (105.51). That $\mathbf{M}$ and $\mathbf{\Gamma}$ are separately Hermitian is manifest: each is built from $\langle i|\, H_w\, |n\rangle\,\overline{\langle j|\, H_w\, |n\rangle}$ with a real weight, since H_w is Hermitian. □

Two structural consequences follow at once. First, $\mathbf{\Gamma}$ is positive semidefinite, being of the form $\sum_n v_n v_n^\dagger$ with $v_{n,i} = \langle i|\, H_w\, |n\rangle$ weighted by a positive delta function; hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(|a|^2 + |b|^2\right) = -\psi^\dagger \mathbf{\Gamma}\psi \le 0, \tag{105.52}$$

and probability leaks out of the two-dimensional space at exactly the rate at which decay products appear. Second, $\mathbf{H}$ is not Hermitian, so its eigenvectors need not be orthogonal — and in the presence of CP violation they are not. This is not a pathology; it is what an open subsystem looks like.

What the discrete symmetries impose

Proposition 105.23 (Symmetry constraints on the mass matrix)**.** *With the phase convention Equation* (105.45)*:*

1. *CPT invariance implies $H_{11} = H_{22}$, that is $M_{11} = M_{22}$ and $\Gamma_{11} = \Gamma_{22}$;*
2. *T invariance implies $\mathrm{Im}\,(M_{12}\hbar\Gamma_{12}^{*}) = 0$, equivalently $|H_{12}| = |H_{21}|$;*
3. *CP invariance implies both.*

Rests on Equations (105.26) *and* (105.45) *and Proposition 105.22.*

Proof. *Derives Proposition 105.23.*

(i) Let Θ be the antiunitary CPT operator, so that $\Theta\,|K^0\rangle = \left|\bar{K}^0\right\rangle$ up to a phase and $\Theta H_w \Theta^{-1} = H_w$, $\Theta H_0 \Theta^{-1} = H_0$. Apply Equation (105.26) to $M_{11} = \langle K^0|\,\mathbf{M}\,|K^0\rangle$: since $\mathbf{M}$ is built from H_w, H_0 and real weights, $\Theta\mathbf{M}\Theta^{-1} = \mathbf{M}$, and Equation (105.26) gives $\left\langle \bar{K}^0\right|\mathbf{M}\left|\bar{K}^0\right\rangle = \langle K^0|\,\mathbf{M}\,|K^0\rangle^{*}$. But $\mathbf{M}$ is Hermitian, so its diagonal elements are real, and $M_{22} = M_{11}$. The same argument on $\mathbf{\Gamma}$ gives $\Gamma_{22} = \Gamma_{11}$. Note that Θ maps the state to its *antiparticle* without reversing momentum (the kaon is at rest), which is exactly what makes this constraint a relation between the two diagonal entries.

(ii) T invariance means $\mathcal{T} H_w \mathcal{T}^{-1} = H_w$ with $\mathcal{T}$ antiunitary and, for kaons at rest, $\mathcal{T}\,|K^0\rangle = \mathrm{e}^{\mathrm{i}\alpha}\,|K^0\rangle$. Applying Equation (105.26) to the off-diagonal element gives $M_{12} = \mathrm{e}^{-2\mathrm{i}\alpha} M_{12}^{*}$ and likewise $\Gamma_{12} = \mathrm{e}^{-2\mathrm{i}\alpha}\Gamma_{12}^{*}$: both off-diagonal elements carry the *same* phase α, which the phase freedom of $|K^0\rangle$ can then remove. The convention-independent content is that their relative phase vanishes, $\mathrm{Im}(M_{12}\hbar\Gamma_{12}^{*}) = 0$, whereupon $H_{21} = H_{12}^{*}$ with H_{12} of a removable phase, so $|H_{12}| = |H_{21}|$.

(iii) $CP = T^{-1}\cdot CPT$ up to the phases already fixed, so it imposes the union of (i) and (ii). □

The three constraints are logically independent, and that is the architecture of the whole subject: the diagonal of $\mathbf{H}$ tests CPT, the moduli of its off-diagonal elements test T, and CP tests both at once. Experiment finds (i) satisfied to the accuracy of Section 105.5.3 and (ii) violated at the 10^{-3} level.

Diagonalization, and the definition of epsilon Assume CPT, so $H_{11} = H_{22} =: H_0$ and

$$\mathbf{H} = \begin{pmatrix} H_0 & H_{12} \\ H_{21} & H_0 \end{pmatrix}. \tag{105.53}$$

The characteristic equation gives eigenvalues $\lambda_{\pm} = H_0 \pm \sqrt{H_{12}H_{21}}$ with right eigenvectors proportional to $(\sqrt{H_{12}}, \pm\sqrt{H_{21}})$. Define the complex number ϵ by

$$\frac{1-\epsilon}{1+\epsilon} = \sqrt{\frac{H_{21}}{H_{12}}}, \qquad \epsilon = \frac{\sqrt{H_{12}} - \sqrt{H_{21}}}{\sqrt{H_{12}} + \sqrt{H_{21}}}, \tag{105.54}$$

so that the two physical states are

$$|K_S\rangle = \frac{(1+\epsilon)\,|K^0\rangle + (1-\epsilon)\left|\bar{K}^0\right\rangle}{\sqrt{2\left(1+|\epsilon|^2\right)}} = \frac{|K_1\rangle + \epsilon\,|K_2\rangle}{\sqrt{1+|\epsilon|^2}}, \tag{105.55}$$

$$|K_L\rangle = \frac{(1+\epsilon)\,|K^0\rangle - (1-\epsilon)\left|\bar{K}^0\right\rangle}{\sqrt{2\left(1+|\epsilon|^2\right)}} = \frac{|K_2\rangle + \epsilon\,|K_1\rangle}{\sqrt{1+|\epsilon|^2}}, \tag{105.56}$$

with eigenvalues written as

$$\lambda_{S,L} = m_{S,L}c^2 - \frac{\mathrm{i}\hbar}{2}\Gamma_{S,L}\,. \tag{105.57}$$

Equation (105.54) is the whole of CP violation in the mixing: $\epsilon = 0$ if and only if $H_{12} = H_{21}$, which by Proposition 105.23 holds if and only if T (and, given CPT, CP) is a symmetry of the mass matrix. It is also manifest that $|K_S\rangle$ and $|K_L\rangle$ are *not* orthogonal: $\langle K_S|K_L\rangle = 2\mathrm{Re}\,\epsilon/(1+|\epsilon|^2) \neq 0$.

To first order in the small asymmetry, write Equation (105.54) as $\epsilon = (H_{12}-H_{21})/[(H_{12}+H_{21}) + 2\sqrt{H_{12}H_{21}}]$ and use $H_{12}+H_{21} \approx 2\sqrt{H_{12}H_{21}}$, so that the whole denominator is $4\sqrt{H_{12}H_{21}} = 2(\lambda_S - \lambda_L)$. The sign of that difference is not free, and it is where a factor of -1 is most easily lost: Equation (105.57) gives $\lambda_S - \lambda_L = -\Delta m\,c^2 - \frac{\mathrm{i}\hbar}{2}(\Gamma_S - \Gamma_L)$, with *both* terms negative, because K_S is at once the lighter state and the broader one. With $H_{21} = M_{12}^* - \frac{\mathrm{i}\hbar}{2}\Gamma_{12}^*$ the numerator is $H_{12} - H_{21} = 2\mathrm{i}\,\mathrm{Im}\,M_{12} + \hbar\,\mathrm{Im}\,\Gamma_{12}$, whence

$$\epsilon \approx \frac{\mathrm{i}\,\mathrm{Im}\,M_{12} + \frac{1}{2}\hbar\,\mathrm{Im}\,\Gamma_{12}}{-\Delta m\,c^2 - \mathrm{i}\frac{\hbar}{2}\Gamma_S}\,, \qquad \Delta m := m_L - m_S\,, \tag{105.58}$$

where $\Gamma_L \ll \Gamma_S$ has been used. Both numerator terms vanish in a T-invariant theory, as they must.

Equation (105.58) makes a prediction that is one of the most satisfying numerical accidents in physics. The $\pi\pi$ channel dominates Γ_S and hence Γ_{12}, and it carries essentially the phase of M_{12}, so $\mathrm{Im}\,\Gamma_{12}$ is negligible next to $\mathrm{Im}\,M_{12}$. The phase of ϵ is then fixed by the denominator alone, up to the 180° carried by the sign of $\mathrm{Im}\,M_{12}$:

$$\arg\epsilon \approx \arctan\frac{2\,\Delta m\,c^2}{\hbar\Gamma_S} = \arctan\frac{2\,\Delta m\,c^2\,\tau_S}{\hbar}\,. \tag{105.59}$$

With the measured [2427]

$$\Delta m\,c^2 = (3.484 \pm 0.006)\times 10^{-6}\,\mathrm{eV}\,, \qquad \frac{\hbar}{2\tau_S} = 3.676\times 10^{-6}\,\mathrm{eV}\,, \tag{105.60}$$

this gives $\arg\epsilon = \arctan(0.948) = 43.5^\circ$, and the measured value is $(43.5 \pm 0.5)^\circ$ — which also settles the branch, since $+43.5^\circ$ rather than -136.5° requires $\mathrm{Im}\,M_{12} < 0$ in the phase convention of Equation (105.45). The agreement is a check on the whole construction, and the near-equality $\Delta m\,c^2 \approx \hbar\Gamma_S/2$ that produces it is, so far as anyone knows, a coincidence.

Strangeness oscillation Because $|K^0\rangle$ is not an eigenstate of $\mathbf{H}$, a beam produced as pure K^0 develops a $\bar{K}^0$ component. Neglecting ϵ (a 10^{-3} correction) and using $|K^0\rangle = (|K_S\rangle + |K_L\rangle)/\sqrt{2}$,

$$\left|K^0(t)\right\rangle = \frac{1}{\sqrt{2}}\left(\mathrm{e}^{-\mathrm{i}\lambda_S t/\hbar}\,|K_S\rangle + \mathrm{e}^{-\mathrm{i}\lambda_L t/\hbar}\,|K_L\rangle\right)\,,$$

so that, projecting back onto the strangeness basis,

$$P\left(K^0 \to \bar{K}^0\right)(t) = \frac{1}{4}\left[\mathrm{e}^{-\Gamma_S t} + \mathrm{e}^{-\Gamma_L t} - 2\mathrm{e}^{-(\Gamma_S+\Gamma_L)t/2}\cos\frac{\Delta m\,c^2 t}{\hbar}\right]\,, \tag{105.61}$$

$$P\left(K^0 \to K^0\right)(t) = \frac{1}{4}\left[\mathrm{e}^{-\Gamma_S t} + \mathrm{e}^{-\Gamma_L t} + 2\mathrm{e}^{-(\Gamma_S+\Gamma_L)t/2}\cos\frac{\Delta m\,c^2 t}{\hbar}\right]\,. \tag{105.62}$$

The oscillation frequency is $\Delta m\,c^2/\hbar = 5.29\times 10^9/\mathrm{s}$, comparable to $1/\tau_S$, so the interference term is observable over the first few τ_S and then dies, leaving equal K^0 and $\bar{K}^0$ content in the surviving K_L beam.

It is worth pausing on what Equation (105.61) makes measurable. A mass difference of $3.5\,\mu\mathrm{eV}/c^2$ is being read off a particle of mass $497.6\,\mathrm{MeV}/c^2$:

$$\frac{\Delta m}{m_K} = 7.0 \times 10^{-15}\,. \tag{105.63}$$

No spectrometer resolves that. The interferometer does, because the observable is a phase accumulated over a macroscopic flight path, and this is the same trick that makes the neutral-meson systems the sharpest CPT laboratories in Section 105.5.3.

Regeneration A K_L beam traversing matter emerges with a K_S component [1077]. The reason is that the strong interaction sees strangeness, not CP: the forward scattering amplitude f for K^0 on a nucleus differs from the amplitude $\bar{f}$ for $\bar{K}^0$, because $\bar{K}^0$ can convert a nucleon into a hyperon while K^0 cannot. Writing the transmitted state through a slab of N scatterers per unit volume and thickness ℓ, each strangeness component picks up its own index of refraction, and the regenerated amplitude is

$$\rho = \frac{\mathrm{i}\pi N\hbar^2\left(f-\bar{f}\right)}{p\,(\lambda_S-\lambda_L)/c}\left(1-\mathrm{e}^{\mathrm{i}(\lambda_S-\lambda_L)\ell/(\hbar v)}\right), \tag{105.64}$$

proportional to $f-\bar{f}$ and therefore vanishing if the medium cannot distinguish the two. Coherent regeneration is the standard tool for producing a K_S beam far from the target; it is also the background that any claim of $K_L \to \pi\pi$ must exclude, and excluding it was the principal experimental burden of Chapter 111.

105.3.2 The 1964 discovery

Christenson, Cronin, Fitch and Turlay put a two-arm spectrometer at the end of a 17 m helium-filled decay volume at the Brookhaven AGS, far enough downstream that every K_S had decayed, and looked for two-body decays of the surviving beam [1274]. They found 45 events with the reconstructed invariant mass of the kaon and with the reconstructed parent momentum collinear with the beam, over a background of about 12. The branching fraction was about 2×10^{-3}; the modern value is $(1.967\pm0.010)\times10^{-3}$ [2427]. The collinearity is what kills regeneration — a regenerated K_S from the residual gas would produce a non-collinear parent — and the substance of the apparatus and its controls is Chapter 111; see Phenomenon 111.8.

The immediate reading of the result is Equation (105.56): the long-lived state is not the CP eigenstate $|K_2\rangle$ but carries an admixture $\epsilon\,|K_1\rangle$, and the CP-even component decays to two pions. Quantitatively,

$$\eta_{+-} := \frac{\mathcal{A}\left(K_L\to\pi^+\pi^-\right)}{\mathcal{A}\left(K_S\to\pi^+\pi^-\right)}, \qquad \eta_{00} := \frac{\mathcal{A}\left(K_L\to\pi^0\pi^0\right)}{\mathcal{A}\left(K_S\to\pi^0\pi^0\right)}, \tag{105.65}$$

and if the only source of CP violation were the mixing, both would equal ϵ. The measured magnitude is

$$|\epsilon| = (2.228\pm0.011)\times10^{-3}, \qquad \arg\epsilon = (43.5\pm0.5)^\circ, \tag{105.66}$$

consistent with Equation (105.59).

Indirect and direct violation The distinction that took thirty-five years to settle is whether CP violation lives only in the mass matrix — in the *state* — or also in the decay amplitudes. Decompose the two-pion final state by isospin. Two pions in a relative S wave from a spinless parent must have a symmetric isospin wavefunction, because the spatial part is symmetric and pions are bosons; of the three combinations $I=0,1,2$ available to two

isospin-1 particles, $I = 1$ is antisymmetric and is therefore forbidden. Write the amplitudes for the two allowed channels as

$$\mathcal{A}\left(K^0 \to (\pi\pi)_I\right) = A_I\, \mathrm{e}^{\mathrm{i}\delta_I}\,, \qquad \mathcal{A}\left(\bar{K}^0 \to (\pi\pi)_I\right) = A_I^*\, \mathrm{e}^{\mathrm{i}\delta_I}\,, \tag{105.67}$$

where δ_I is the $\pi\pi$ strong rescattering phase — real and identical for the two, since the strong interaction conserves CP — and any CP violation in the decay resides in $\operatorname{Im} A_I$. The relative sign of the two conjugate amplitudes is fixed by Equation (105.45).

The observed states are recovered from the isospin eigenstates by the Clebsch–Gordan coefficients of $1 \otimes 1$ in the $I_3 = 0$ sector. Writing $|m_1, m_2\rangle$ for the product state of two pions of third components m_1, m_2, those coefficients are

$$|2\rangle = \tfrac{1}{\sqrt{6}}\,|+,-\rangle + \sqrt{\tfrac{2}{3}}\,|0,0\rangle + \tfrac{1}{\sqrt{6}}\,|-,+\rangle\,, \tag{105.68}$$

$$|0\rangle = \tfrac{1}{\sqrt{3}}\,|+,-\rangle - \tfrac{1}{\sqrt{3}}\,|0,0\rangle + \tfrac{1}{\sqrt{3}}\,|-,+\rangle\,, \tag{105.69}$$

the third combination $|1\rangle = \left(|+,-\rangle - |-,+\rangle\right)/\sqrt{2}$ being the antisymmetric one already excluded. Since $\left|\pi^+\pi^-\right\rangle$ in an S wave *is* the normalized symmetric combination $\left(|+,-\rangle + |-,+\rangle\right)/\sqrt{2}$, while $\left|\pi^0\pi^0\right\rangle = |0,0\rangle$, projecting gives

$$\left|\pi^+\pi^-\right\rangle = \sqrt{\tfrac{1}{3}}\,|2\rangle + \sqrt{\tfrac{2}{3}}\,|0\rangle\,, \qquad \left|\pi^0\pi^0\right\rangle = \sqrt{\tfrac{2}{3}}\,|2\rangle - \sqrt{\tfrac{1}{3}}\,|0\rangle\,. \tag{105.70}$$

The two are orthogonal and normalized, as they must be. The weight of A_2 relative to A_0 is $1/\sqrt{2}$ in the charged channel and $-\sqrt{2}$ in the neutral one, the second being -2 times the first, and that single ratio is the entire origin of the factor -2 below. Carrying Equation (105.70) through gives

$$\eta_{+-} = \epsilon + \epsilon'\,, \qquad \eta_{00} = \epsilon - 2\epsilon'\,, \tag{105.71}$$

with

$$\epsilon' = \frac{\mathrm{i}}{\sqrt{2}}\, \mathrm{e}^{\mathrm{i}(\delta_2 - \delta_0)}\, \frac{\operatorname{Re} A_2}{\operatorname{Re} A_0} \left[\frac{\operatorname{Im} A_2}{\operatorname{Re} A_2} - \frac{\operatorname{Im} A_0}{\operatorname{Re} A_0}\right]. \tag{105.72}$$

Everything in Equation (105.72) is a statement about the decay amplitudes and none of it about the mass matrix, so $\epsilon' \neq 0$ is CP violation *in the decay*: direct CP violation. Note the structure of the bracket — a *difference* of two phases. If the two isospin amplitudes carried a common CP-odd phase it could be rotated away, and ϵ' would vanish; direct CP violation requires two interfering amplitudes with different weak phases, a point proved in general at Proposition 105.39.

The observable is the double ratio.

Proposition 105.24 (The double ratio)**.**

$$\frac{\Gamma(K_L \to \pi^0\pi^0)\,\Gamma(K_S \to \pi^+\pi^-)}{\Gamma(K_L \to \pi^+\pi^-)\,\Gamma(K_S \to \pi^0\pi^0)} = \left|\frac{\eta_{00}}{\eta_{+-}}\right|^2 \approx 1 - 6\operatorname{Re}\frac{\epsilon'}{\epsilon}\,. \tag{105.73}$$

Rests on Equation (105.71).

Proof. *Derives Proposition 105.24.*

From Equation (105.71), $\eta_{00}/\eta_{+-} = (1 - 2\epsilon'/\epsilon)/(1 + \epsilon'/\epsilon)$. With $|\epsilon'/\epsilon| \sim 10^{-3}$, expand to first order: $\eta_{00}/\eta_{+-} \approx 1 - 3\epsilon'/\epsilon$, whose squared modulus is $1 - 6\mathrm{Re}(\epsilon'/\epsilon)$. □

The double ratio is designed so that every quantity that is common to the four channels — kaon flux, detector solid angle, most of the efficiency — cancels. NA48 at CERN [1973] and KTeV at Fermilab [1965] measured it with independent systematics, and the world average is [2427]

$$\operatorname{Re}\frac{\epsilon'}{\epsilon} = (1.66 \pm 0.23) \times 10^{-3}\,. \tag{105.74}$$

See Phenomenon 111.11. A hypothetical superweak $\Delta S = 2$ interaction acting only in the mass matrix reproduces ϵ exactly and predicts $\epsilon' = 0$; Equation (105.74) excludes it. That ϵ'/ϵ is itself of order 10^{-3} rather than order unity is largely the $\Delta I = \tfrac{1}{2}$ rule, the empirical fact that $|A_0/A_2| \approx 22$, which enhances ϵ and suppresses the $I = 2$ amplitude appearing in Equation (105.72).

An absolute definition of matter The semileptonic decays give the cleanest statement of what CP violation means. The rule $\Delta S = \Delta Q$, which holds because the strange quark decays to an up quark by emitting a W^-, allows $K^0 \to \pi^- e^+ \nu_e$ and forbids $K^0 \to \pi^+ e^- \bar{\nu}_e$; the conjugate statements hold for $\bar{K}^0$. From Equation (105.56) the K^0 content of K_L carries amplitude $(1+\epsilon)$ and the $\bar{K}^0$ content $-(1-\epsilon)$, so

$$\delta_L = \frac{\Gamma(\pi^- e^+ \nu_e) - \Gamma(\pi^+ e^- \bar{\nu}_e)}{\Gamma(\pi^- e^+ \nu_e) + \Gamma(\pi^+ e^- \bar{\nu}_e)} = \frac{|1+\epsilon|^2 - |1-\epsilon|^2}{|1+\epsilon|^2 + |1-\epsilon|^2} = \frac{2\mathrm{Re}\,\epsilon}{1+|\epsilon|^2}\,. \tag{105.75}$$

Numerically $2\,|\epsilon|\cos 43.5^\circ = 3.23 \times 10^{-3}$, against the measured $(3.32 \pm 0.06) \times 10^{-3}$ [2427]; see Phenomenon 111.9. The significance is not the number but its status. Instruct a correspondent, with whom no sample and no convention has been shared, to prepare a long-lived neutral kaon beam and count the charge of the emitted lepton: the commoner sign defines the positron, and therefore matter, absolutely. Before 1964 no such instruction existed — C and P violation alone leave the labelling conventional, since one can compensate a mirror reflection by an exchange of matter and antimatter. Only CP violation makes "matter" an objective term.

105.3.3 The CKM phase

Kobayashi and Maskawa asked, in 1973, what the minimal modification of the Standard Model would be that admits CP violation at all [1505]. Their answer is a counting argument, and it predicted a third generation of quarks from a single measured asymmetry of 2×10^{-3}.

Theorem 105.25 (Number of physical phases)**.** *Let there be N generations of quarks. After diagonalizing the mass matrices, the charged-current interaction is governed by a unitary $N \times N$ matrix V, of which*

$$n_{angles} = \frac{N(N-1)}{2}\,, \qquad n_{phases} = \frac{(N-1)(N-2)}{2} \tag{105.76}$$

parameters are physical. There is no CP-violating phase for $N \le 2$ and exactly one for $N = 3$.

Rests on Equation (104.46).

Proof. *Derives Theorem 105.25.*

The Yukawa couplings of Chapter 104 give arbitrary complex mass matrices M_u and M_d; each is brought to positive diagonal form by a bi-unitary transformation $M = U_L^\dagger M_{\mathrm{diag}} U_R$, and the charged current, which couples $\bar{u}_L \gamma^\mu d_L$, retains the combination $V = U_L^{u\dagger} U_L^d$, a general $N \times N$ unitary matrix.

A unitary $N \times N$ matrix depends on N^2 real parameters (the condition $V^\dagger V = \mathbb{1}$ imposes N^2 real conditions on $2N^2$ real entries). Of these, an orthogonal $N \times N$ matrix would use $N(N-1)/2$; the remainder, $N^2 - N(N-1)/2 = N(N+1)/2$, are phases.

Not all phases are physical. The quark fields may be rephased, $u_j \to \mathrm{e}^{\mathrm{i}\alpha_j} u_j$ and $d_k \to \mathrm{e}^{\mathrm{i}\beta_k} d_k$, which changes $V_{jk} \to \mathrm{e}^{-\mathrm{i}\alpha_j} V_{jk} \mathrm{e}^{\mathrm{i}\beta_k}$ without touching the diagonal mass terms. That is $2N$ phases, but the overall common phase $\alpha_j = \beta_k = \theta$ leaves V unchanged, so only $2N - 1$ are usable. Hence

$$n_{\mathrm{phases}} = \frac{N(N+1)}{2} - (2N-1) = \frac{N^2 - 3N + 2}{2} = \frac{(N-1)(N-2)}{2}\,.$$

For $N = 2$ this is zero: the two-generation mixing matrix is a real rotation through the Cabibbo angle [1248], and no phase survives. For $N = 3$ it is one. □

The argument's force is that it is a counting theorem, not a model. Given that CP violation exists and that the Standard Model gauge structure is right, three generations are the minimum. In 1973 three quarks were known, the fourth was conjectural, and the fifth and sixth were found in 1977 and 1995.

Definition 105.26 (The Jarlskog invariant)**.** For any $i \neq k$ and $j \neq l$,

$$\operatorname{Im}\left(V_{ij} V_{kl} V_{il}^{*} V_{kj}^{*}\right) = J \sum_{m,n} \varepsilon_{ikm} \varepsilon_{jln} \,, \tag{105.77}$$

defining a single real number J independent of which quartet is chosen [1756].

The quartet in Equation (105.77) is invariant under the rephasing $V_{jk} \to \mathrm{e}^{-\mathrm{i}\alpha_j} V_{jk} \mathrm{e}^{\mathrm{i}\beta_k}$, since each index appears once with a phase and once with its conjugate; J is therefore a physical observable and not a parametrization artefact. Unitarity makes all nine quartets equal up to sign, which is what Equation (105.77) records. The measured value is [2074, 2427]

$$J = \left(3.08^{+0.15}_{-0.13}\right) \times 10^{-5} \,, \tag{105.78}$$

a small number whose smallness is a fact about the mixing angles and not about the phase, which is of order unity.

Proposition 105.27 (When CP violation is possible)**.** *CP is violated in the quark sector if and only if*

$$\det\left[M_u M_u^{\dagger} c^4, M_d M_d^{\dagger} c^4\right] = 2\mathrm{i} J \prod_{i>j} \left[\left(m_{u_i}^2 - m_{u_j}^2\right) c^4\right] \prod_{i>j} \left[\left(m_{d_i}^2 - m_{d_j}^2\right) c^4\right] \tag{105.79}$$

is nonzero. In particular CP is conserved if any two quarks of the same charge are degenerate, or if $J = 0$.

Rests on Definition 105.26 and Theorem 105.25.

Derivation. *Derives Proposition 105.27.*

The left-hand side is invariant under the field redefinitions that leave the physics alone, and it vanishes identically if the two Hermitian matrices commute, which happens exactly when they are simultaneously diagonalizable — i.e. when V can be taken real. Evaluating the determinant in the basis where $M_u M_u^{\dagger}$ is diagonal, each entry of the commutator is $(m_{u_i}^2 - m_{u_j}^2)c^4$ times the corresponding entry of $V\ \mathrm{diag}(m_d^2 c^4)\, V^{\dagger}$, and expanding the 3×3 determinant collects the six mass differences together with the imaginary part of a quartet, which is J by Definition 105.26. That last step is where the sign on the right is fixed, and it is fixed by a convention worth stating: V is the matrix whose (u, d) entry is V_{ud}, so that in the up-diagonal basis $M_d M_d^{\dagger} c^4 = V\, \mathrm{diag}(m_d^2 c^4) V^{\dagger}$, and J is the quartet of Definition 105.26; with the opposite convention for V both J and the sign reverse together, and the statement is unchanged.

The factors of c^4 in Equation (105.79) are not decoration. $M_u M_u^{\dagger}$ is a squared mass, kg^2, so $M_u M_u^{\dagger} c^4$ is a squared energy; each entry of the commutator of two such matrices carries J^4 and the 3×3 determinant J^{12}, which is also what the right-hand side's six factors of J^2 come to. Written without them the left-hand side would be kg^{12} and the two sides would not have the same dimension at all. □

If two up-type quarks had equal mass, the freedom to rotate between them could be used to make V real; the phase would be unphysical. CP violation in the Standard Model therefore requires three *non-degenerate* generations, and the extreme hierarchy of

the observed masses is precisely what makes Equation (105.79) numerically tiny — the fact that returns, with consequences, in Section 105.6.2.

Unitarity of V gives six orthogonality relations, of which

$$V_{ud}V_{ub}^{*} + V_{cd}V_{cb}^{*} + V_{td}V_{tb}^{*} = 0 \tag{105.80}$$

has three terms of comparable magnitude and so draws a non-degenerate triangle in the complex plane. Its area is $J/2$: for any three complex numbers summing to zero, the triangle they close has area $\frac{1}{2}\,|\mathrm{Im}(z_1 z_2^{*})|$, and that imaginary part is a Jarlskog quartet. The angles are conventionally α, β, γ, and the programme of flavour physics is to measure the sides and the angles by independent means and ask whether they close. They do, at the few-per-cent level, which is the quantitative statement that one phase accounts for all observed CP violation in the quark sector; the detailed over-determination belongs to Chapter 103.

105.3.4 CP violation beyond the kaon

The B factories If one phase controls everything, the kaon's 10^{-3} effect and a possible order-unity effect elsewhere are the same physics. The Standard Model predicts the latter in $B^0 \to J/\psi\, K_S^0$, where the interference is between decay after mixing and decay without mixing, and no small ratio of amplitudes suppresses it.

Proposition 105.28 (Time-dependent asymmetry)**.** *For a final state f that is a CP eigenstate accessible to both B^0 and $\bar{B}^0$, define $\lambda = (q/p)\left(\bar{\mathcal{A}}_f/\mathcal{A}_f\right)$, with q/p the mixing parameter analogous to $(1-\epsilon)/(1+\epsilon)$. Then*

$$\frac{\Gamma(\bar{B}^0(t)\to f) - \Gamma(B^0(t)\to f)}{\Gamma(\bar{B}^0(t)\to f) + \Gamma(B^0(t)\to f)} = S_f \sin\frac{\Delta m_d c^2 t}{\hbar} - C_f \cos\frac{\Delta m_d c^2 t}{\hbar}\,, \tag{105.81}$$

with $S_f = 2\mathrm{Im}\,\lambda/(1+|\lambda|^2)$ and $C_f = (1-|\lambda|^2)/(1+|\lambda|^2)$.

Rests on Equations (105.53), (105.55) and (105.56).

Derivation. Derives Proposition 105.28.

The B system has $\Delta\Gamma \approx 0$, so the two mass eigenstates share a width and the time evolution of an initially pure B^0 is $|B^0(t)\rangle = g_+(t)\,|B^0\rangle + (q/p)g_-(t)\left|\bar{B}^0\right\rangle$ with $g_\pm(t) = \mathrm{e}^{-\Gamma t/2}\mathrm{e}^{-\mathrm{i}\bar{m}c^2 t/\hbar}$ times cos or $+\mathrm{i}\sin$ of $\Delta m_d c^2 t/(2\hbar)$ respectively, $\bar{m}$ being the mean of the two masses, obtained by diagonalizing Equation (105.53) exactly as in Equations (105.55) and (105.56).

That sign in g_- is not free, and it is the one place the derivation can silently go wrong. The eigenvector $p\,|B^0\rangle + q\left|\bar{B}^0\right\rangle$ is the *light* state, exactly as $|K_S\rangle$ is in Equation (105.55), so writing $\Delta m_d = m_H - m_L > 0$ gives $g_\pm = \frac{1}{2}\left(\mathrm{e}^{-\mathrm{i}\lambda_L t/\hbar} \pm \mathrm{e}^{-\mathrm{i}\lambda_H t/\hbar}\right)$ and hence $+\mathrm{i}\sin$ and not $-\mathrm{i}\sin$; taking the other sign reverses the coefficient of $\sin\Delta m_d c^2 t/\hbar$ in Equation (105.81), which is the whole measurement. Then $\mathcal{A}(B^0(t)\to f) = \mathcal{A}_f\,[g_+ + \lambda g_-]$ and $\mathcal{A}(\bar{B}^0(t)\to f) = (p/q)\mathcal{A}_f\,[g_- + \lambda g_+]$. Squaring, using $|g_\pm|^2 = \frac{1}{2}\mathrm{e}^{-\Gamma t}(1 \pm \cos\Delta m_d c^2 t/\hbar)$ and $g_+^{*}g_- = \frac{\mathrm{i}}{2}\mathrm{e}^{-\Gamma t}\sin\Delta m_d c^2 t/\hbar$, and forming the asymmetry gives Equation (105.81). □

For $f = J/\psi\, K_S^0$ the decay proceeds through a single dominant amplitude, so $|\lambda| = 1$, $C_f = 0$, and the phase of λ is the mixing phase $\mathrm{e}^{-2\mathrm{i}\beta}$ of Equation (105.80); hence

$$S_{J/\psi K_S} = \sin 2\beta\,. \tag{105.82}$$

The measurement requires knowing which meson was which at production, and the B factories solved this by producing the pair coherently from the $\Upsilon(4S)$: the two mesons are in a C-odd, $L = 1$ state, so until one decays there is exactly one B^0 and one $\bar{B}^0$,

and a flavour-specific decay of one tags the other. Belle [2005] and BaBar [2010] reported simultaneously, with $\sin 2\beta = 0.99 \pm 0.14 \pm 0.06$ and $0.59 \pm 0.14 \pm 0.05$ respectively — two values differing by about two standard deviations, which is a useful reminder that first measurements scatter. The world average has since settled at $\sin 2\beta = 0.708 \pm 0.011$ [2427], and $\Delta m_d c^2 = (3.334 \pm 0.013) \times 10^{-4}\,\text{eV}$. See Phenomenon 111.13.

The effect is of order unity. CP violation is therefore not intrinsically small: it is small in the kaon because the kaon system happens to sit where the interfering amplitudes are unequal, and large in this B channel because they are not.

Charm The up-type sector was the last to yield. LHCb measured the difference of the time-integrated CP asymmetries of two singly Cabibbo-suppressed charm decays [2323],

$$\Delta A_{CP} = A_{CP}(D^0 \to K^- K^+) - A_{CP}(D^0 \to \pi^- \pi^+) = (-15.4 \pm 2.9) \times 10^{-4}\,, \qquad (105.83)$$

nonzero at 5.3 standard deviations. The difference is taken because production and detection asymmetries cancel in it. The effect requires interference of tree and loop ("penguin") amplitudes with different strong phases, per Proposition 105.39, and the long-distance strong phases are not calculable, so the measurement establishes the phenomenon without sharply testing the CKM prediction. See Phenomenon 111.15.

Leptons The mixing matrix of the leptons admits the same counting as Theorem 105.25, so with three neutrino generations there is one Dirac phase δ_{CP}. T2K, comparing $\nu_\mu \to \nu_e$ with $\bar{\nu}_\mu \to \bar{\nu}_e$ over a 295 km baseline, finds the data preferring values of δ_{CP} near $-\pi/2$ and excluding CP conservation ($\delta_{CP} = 0$ or π) at the 95 % confidence level [2345]. This is a constraint and not a discovery: the exclusion is a two-sigma statement resting on a comparison of two event samples of a few tens of events each, and it depends on the mass ordering, which is not yet settled. The apparatus and the oscillation formalism are Chapter 114; the honest summary is that leptonic CP violation is likely and unproven.

105.4 Direct tests of time reversal

Everything in Section 105.3 tests CP. Given the CPT theorem, CP violation is T violation, and one might think nothing remains to be measured. But that inference *assumes* the theorem whose hypotheses are the deepest in the subject, and a test of T that does not assume CPT is a genuinely independent piece of information — indeed it is the only way to separate the three symmetries experimentally. This section describes the two measurements that compare a process directly with its motion reverse, and the static observable that needs no process at all.

105.4.1 The CPLEAR asymmetry

Kabir's observation is that the neutral kaon offers a transition and its exact reverse [1426]. The process $K^0 \to \bar{K}^0$ and the process $\bar{K}^0 \to K^0$ are related by exchanging initial and final state; both particles are at rest, so no momenta or spins need reversing, and motion reversal is nothing but the exchange. Define

$$A_T(t) = \frac{P\left(\bar{K}^0 \to K^0\right)(t) - P\left(K^0 \to \bar{K}^0\right)(t)}{P\left(\bar{K}^0 \to K^0\right)(t) + P\left(K^0 \to \bar{K}^0\right)(t)}\,. \qquad (105.84)$$

If T is a symmetry, Equation (105.25) forces $A_T = 0$ identically.

Proposition 105.29 (The Kabir asymmetry measures the real part of epsilon)**.** *To first order in ϵ,*

$$A_T = 4\,\mathrm{Re}\,\epsilon\,, \tag{105.85}$$

independent of t.

Rests on Equations (105.55)*,* (105.56) *and* (105.84)*.*

Proof. *Derives Proposition 105.29.*

Invert Equations (105.55) and (105.56) to express $|K^0\rangle$ and $\left|\bar{K}^0\right\rangle$ in terms of $|K_S\rangle\,, |K_L\rangle$. Propagating and projecting, the two transition amplitudes are

$$\mathcal{A}\left(K^0 \to \bar{K}^0\right) = \frac{1-\epsilon}{2(1+\epsilon)}\left(\mathrm{e}^{-\mathrm{i}\lambda_S t/\hbar} - \mathrm{e}^{-\mathrm{i}\lambda_L t/\hbar}\right)\,,$$

and the reverse with $(1-\epsilon)/(1+\epsilon)$ replaced by its inverse; the time-dependent bracket is common. Hence, writing $r = |(1+\epsilon)/(1-\epsilon)|^2 \approx 1 + 4\,\mathrm{Re}\,\epsilon$ and therefore $r^{-1} \approx 1 - 4\,\mathrm{Re}\,\epsilon$ to the same order, so that the numerator is $r - r^{-1} \approx 8\,\mathrm{Re}\,\epsilon$ while the denominator is $r + r^{-1} \approx 2$:

$$A_T = \frac{r - r^{-1}}{r + r^{-1}} \approx \frac{8\,\mathrm{Re}\,\epsilon}{2} = 4\,\mathrm{Re}\,\epsilon\,,$$

and the time dependence cancels between numerator and denominator. □

Numerically $4\,|\epsilon|\cos 43.5^\circ = 6.46 \times 10^{-3}$. The CPLEAR collaboration at CERN's Low Energy Antiproton Ring measured

$$A_T = (6.6 \pm 1.3 \pm 1.0) \times 10^{-3}\,, \tag{105.86}$$

the first direct observation of a time-reversal asymmetry [1948]. The method is worth stating because the tagging is the whole difficulty. CPLEAR produced neutral kaons in $p\bar{p} \to K^{\mp}\pi^{\pm}K^0/\bar{K}^0$ at rest; strangeness conservation in the strong production ties the sign of the accompanying charged kaon to the strangeness of the neutral one, so the initial state is tagged by a charge measured in the same event. The final strangeness is tagged by the charge of the lepton in the semileptonic decay, using $\Delta S = \Delta Q$ as in Equation (105.75).

The objection, and the answer The measurement has been criticized on the ground that it is not free of CPT assumptions, and the criticism is partly right. The final-state tag uses the semileptonic amplitudes, and if those amplitudes themselves violated CPT — or if $\Delta S = \Delta Q$ failed — the observed asymmetry could be produced without any T violation in the mixing. Two answers are available. First, the same data set constrains the offending parameters: the K_L and K_S semileptonic charge asymmetries and their time dependence bound the CPT-violating and $\Delta S \neq \Delta Q$ amplitudes at a level well below the measured A_T, so the interpretation survives *within* the standard phenomenological parametrization [2427]. Second — and this is the honest statement — no measurement of a decaying system can be completely free of assumptions about the decay, because the observation is made through the decay products. A T test with no such entanglement requires a stable system, which is what Section 105.4.3 provides at the cost of measuring something else.

105.4.2 Time reversal in the B system

The BaBar measurement is the cleanest of its kind, and it works by exploiting entanglement as a state-preparation device [2197].

The $\Upsilon(4S)$ decays to $B^0\bar{B}^0$ in a C-odd, $L = 1$ state, so before either decays the pair is in the antisymmetric combination

$$|\Psi\rangle = \frac{1}{\sqrt{2}}\left(\left|B^0\right\rangle_1\left|\bar{B}^0\right\rangle_2 - \left|\bar{B}^0\right\rangle_1\left|B^0\right\rangle_2\right) = \frac{1}{\sqrt{2}}\left(|B_+\rangle_1\,|B_-\rangle_2 - |B_-\rangle_1\,|B_+\rangle_2\right)\,, \tag{105.87}$$

the second form following because an antisymmetric combination of a two-dimensional space is basis-independent up to a phase; $B_\pm$ denote the states that do and do not decay to the CP eigenstate $J/\psi K^0_S$. This is the Einstein–Podolsky–Rosen correlation of Chapter 84, used here not to test locality but as a preparation: observing the decay of one meson at time t_1 to a flavour-specific state $\ell^\pm X$ projects the *other* onto a definite flavour at that instant, and observing a decay to $J/\psi K^0_S$ projects it onto a definite CP label.

The comparison is then available in both time orders. Events in which the flavour tag comes first and the CP tag second measure the transition $B^0 \to B_-$; events in the opposite order measure $B_- \to B^0$. These are a process and its motion reverse: the initial and final states are exchanged, and *nothing is conjugated*. That is what distinguishes the measurement from a CP test, which would compare $B^0 \to B_-$ with $\bar{B}^0 \to B_-$, exchanging particle for antiparticle rather than initial for final. BaBar extracted all three sets of parameters — T-odd, CP-odd and CPT-odd — from one fit to the same events, and found the T-violating parameters nonzero at 14 standard deviations, with

$$\Delta S_T^+ = -1.37 \pm 0.14 \pm 0.06\,, \qquad \Delta S_T^- = 1.17 \pm 0.18 \pm 0.11\,, \tag{105.88}$$

while the CPT-odd parameters were consistent with zero.

Two caveats keep the claim honest. The B mesons have a finite width, so the "reversed" transition is not an exact motion reverse of the original — the two occur over different intervals of proper time, and the equality of the rates is derived within the same effective two-state formalism as Proposition 105.22. And the projection interpretation of the tag is exact only if the tagging decay is flavour-specific, which is an assumption of the same family as $\Delta S = \Delta Q$ above. What the measurement does establish, and what no earlier one did, is that the T-odd and the CP-odd parameters are separately determined from a single data set and are separately nonzero.

105.4.3 Electric dipole moments as T tests

The observables above compare two processes. A permanent electric dipole moment compares nothing: it is a static property of a single state, and its existence would violate P and T at once. Purcell and Ramsey made the point in 1950, at a time when parity was universally assumed, with an argument that is really about method [966]: a symmetry that has never been tested in a given sector is an assumption, and the way to test it is to measure the quantity it forbids.

Theorem 105.30 (An EDM violates P and T)**.** *Let $|\psi\rangle$ be a nondegenerate state of a system with definite total angular momentum j, and let $\hat{\boldsymbol{d}} = \int \mathrm{d}^3x\, \boldsymbol{x}\, \rho(\boldsymbol{x})$ be the electric dipole operator. If $\left\langle \hat{\boldsymbol{d}} \right\rangle \neq 0$ then the Hamiltonian violates both P and T.*

Rests on Equations (105.6)*,* (105.17) *and* (105.30)*.*

Proof. *Derives Theorem 105.30.*

Only one fact about the state is needed, and it can be had without any general machinery. Write $|\psi\rangle = |j, m\rangle$, the quantization axis chosen along $\left\langle \hat{\boldsymbol{J}} \right\rangle$. The rotation $U(\varphi) = \mathrm{e}^{-\mathrm{i}\varphi \hat{J}_z/\hbar}$ about that axis multiplies $|\psi\rangle$ by the phase $\mathrm{e}^{-\mathrm{i}m\varphi}$ and so changes no expectation value whatever; but on a vector operator it acts as $U^\dagger \hat{V}_i U = R_{ij}(\varphi)\hat{V}_j$. Hence $\left\langle \hat{\boldsymbol{V}} \right\rangle = R(\varphi) \left\langle \hat{\boldsymbol{V}} \right\rangle$ for every φ, and a vector fixed by every rotation about $\hat{\boldsymbol{z}}$ has no transverse component: every vector expectation value in such a state points along $\left\langle \hat{\boldsymbol{J}} \right\rangle$. (The general statement, of which this is the diagonal case, is the projection theorem carried by the Wigner–Eckart theorem.) Fixing the constant so that d is the value of $\left\langle \hat{d}_z \right\rangle$ in the stretched state $m = j$,

$$\left\langle \hat{\boldsymbol{d}} \right\rangle = d\, \frac{\left\langle \hat{\boldsymbol{J}} \right\rangle}{j\hbar} \tag{105.89}$$

with d real because $\hat{\boldsymbol{d}}$ is Hermitian: the dipole, if it exists, must point along the spin, because the spin is the only vector the state possesses. Nondegeneracy is what licenses the first step — it is what gives the state a definite m with no admixture of opposite parity.

The interaction with a static electric field is $H_d = -\left\langle \hat{\boldsymbol{d}} \right\rangle \cdot \boldsymbol{E} \propto -d \left\langle \hat{\boldsymbol{J}} \right\rangle \cdot \boldsymbol{E}$. By Equation (105.6) $\boldsymbol{E}$ is polar and $\hat{\boldsymbol{J}}$ axial; by Equation (105.30) $\boldsymbol{E}$ is T-even and $\hat{\boldsymbol{J}}$ is T-odd; and by Equation (105.17) $\boldsymbol{E}$ is C-odd, while $\hat{\boldsymbol{J}}$, being an angular momentum, is C-even. So $\hat{\boldsymbol{J}} \cdot \boldsymbol{E}$ is odd under each of the three separately. The dipole operator supplies the missing sign under C: the charge density ρ reverses under charge conjugation, so $\hat{\boldsymbol{d}}$ is C-odd and a particle and its antiparticle carry opposite dipoles. The interaction $-\hat{\boldsymbol{d}} \cdot \boldsymbol{E}$ is therefore *C-even*, P-odd and T-odd. A nonzero d means the Hamiltonian contains a term with those properties, i.e. P and T are violated (and, by the CPT theorem, CP with them — the three signs multiply to CPT-even, as Theorem 105.34 requires of any term in a Hermitian local Lagrangian). □

Remark 105.31 (Degeneracy is the loophole, and it is not academic). The hypothesis of nondegeneracy is essential and is often omitted. The hydrogen atom in its $n = 2$ level, where $2S_{1/2}$ and $2P_{1/2}$ are degenerate up to the Lamb shift, exhibits a *linear* Stark effect: the field mixes states of opposite parity and the energy shifts linearly in $|\boldsymbol{E}|$, exactly as a permanent dipole would. No symmetry is violated, because the state that acquires the dipole is not an eigenstate of parity to begin with. The same mechanism is what makes polar molecules useful in EDM searches: a molecule such as ThO has closely spaced levels of opposite parity, which a modest laboratory field fully polarizes, producing an enormous *effective* internal field acting on the electron — in ThO of order $7.8 \times 10^{12}\,\mathrm{V/m}$, which is the 78 GV/cm of the ACME literature, and enormously larger than any field a laboratory can apply directly. The molecule is a lever, not a violation.

How the measurement is made The technique is Ramsey's method of separated oscillatory fields [968]. Neutrons in a storage cell see parallel magnetic and electric fields; the Larmor precession frequency is

$$h\nu_{\pm} = |2\mu_n B \pm 2 d_n E| \,, \tag{105.90}$$

the sign depending on whether $\boldsymbol{E}$ is parallel or antiparallel to $\boldsymbol{B}$. Reversing $\boldsymbol{E}$ and taking the difference isolates the dipole:

$$\Delta\nu = \frac{4 d_n E}{h} \,. \tag{105.91}$$

The statistical sensitivity of a Ramsey measurement with coherence time T on N neutrons is

$$\sigma(d_n) = \frac{\hbar}{2\,\alpha_v\, E\, T\, \sqrt{N}} \,, \tag{105.92}$$

with $\alpha_v \lesssim 1$ the fringe visibility. Inserting the PSI apparatus's parameters — $E \approx 1.1 \times 10^6\,\mathrm{V/m}$, $T = 180\,\mathrm{s}$, and of order 10^7 ultracold neutrons accumulated over the run — gives $\sigma(d_n) \sim 8 \times 10^{-47}\,\mathrm{C\,m}$, which is the order of the published limit and confirms that the experiment is statistics-limited rather than limited by any subtlety.

Phenomenon 105.32 (The neutron has no measurable electric dipole moment)**.** Seven decades of measurement find no permanent electric dipole moment of the neutron. The current bound, from ultracold neutrons at the Paul Scherrer Institute, is

$$|d_n| < 2.9 \times 10^{-47}\,\mathrm{C\,m} \tag{105.93}$$

at 90 % confidence [2346], improving on [2090]. The same number is customarily quoted as 1.8×10^{-26} in units of the elementary charge times the centimetre; the conversion is $1\,e\,\mathrm{cm} = 1.602\,176\,634 \times 10^{-21}\,\mathrm{C\,m}$. See Phenomenon 111.19.

Rests on Theorem 105.30 and Equations (105.91) *and* (105.92).

Derivation. *Derives Phenomenon 105.32.*

That the observable is forbidden at all is Theorem 105.30; what is derived here is the number, since a bound is a statement about a resolved frequency and Equation (105.91) is what converts the two. At the PSI storage field $E \approx 1.1 \times 10^{6}\,\mathrm{V/m}$, a dipole equal to Equation (105.93) would shift the Ramsey resonance by

$$\Delta\nu = \frac{4 d_n E}{h} = \frac{4 \times 2.9 \times 10^{-47}\,\mathrm{C\,m} \times 1.1 \times 10^{6}\,\mathrm{V/m}}{6.626\,070\,15 \times 10^{-34}\,\mathrm{J\,s}} = 1.9 \times 10^{-7}\,\mathrm{Hz}\,,$$

against a Larmor frequency $2\mu_n B/h \approx 29\,\mathrm{Hz}$ at the working field $B \approx 1\,\mu\mathrm{T}$: the experiment is a determination of a frequency ratio to 7×10^{-9}, made twice with $\boldsymbol{E}$ reversed. That such a resolution is available follows from Equation (105.92): with $T = 180\,\mathrm{s}$ and of order 10^{7} ultracold neutrons the statistical limit is $8.4 \times 10^{-47}\,\mathrm{C\,m}$, within a factor of three of the published 90 % bound, which is the quantitative content of the claim that the measurement is statistics-limited. The bound is therefore not a theoretical estimate at all: it is the frequency resolution of a Ramsey interferometer, divided by $4E/h$. □

The number is hard to feel, so it is worth converting once. A dipole of $2.9 \times 10^{-47}\,\mathrm{C\,m}$ corresponds to separating one elementary charge from its opposite by $1.8 \times 10^{-28}\,\mathrm{m}$ — about 10^{13} times smaller than the neutron itself, whose charge radius is of order 0.8 fm. Scaled up so that the neutron were the size of the Earth, the charge separation would be a few micrometres.

The electron is bounded even harder, in units of its own size, by experiments on polar molecules. ACME, using the metastable H state of ThO, obtained $|d_e| < 1.8 \times 10^{-50}\,\mathrm{C\,m}$ ($1.1 \times 10^{-29}\,e\,\mathrm{cm}$) at 90 % confidence [2300], and a trapped molecular-ion experiment on HfF^{+} has since reached $|d_e| < 6.6 \times 10^{-51}\,\mathrm{C\,m}$ ($4.1 \times 10^{-30}\,e\,\mathrm{cm}$) [2421].

What makes these null results powerful is the size of the Standard Model background. In the quark sector the CKM phase can generate a neutron EDM only at three loops, and the chiral suppression is severe: the estimate is of order $10^{-32}\,e\,\mathrm{cm}$, six orders below Equation (105.93). For the electron the CKM contribution requires four loops and is of order $10^{-38}\,e\,\mathrm{cm}$, nine orders below the current limit. Every improvement in these experiments therefore probes a range in which the Standard Model predicts nothing observable, so a nonzero result would be unambiguous — and a null result excludes new sources of CP violation over a wide range without any background to subtract. The one Standard Model contribution that is *not* negligible is the strong CP term of Section 105.7, which is why Equation (105.93) is simultaneously the sharpest constraint on $\bar{\theta}$.

105.5 The CPT theorem

105.5.1 Statement and hypotheses

Definition 105.33 (The CPT operator)**.** $\Theta := \mathcal{C}\mathcal{P}\mathcal{T}$, an antiunitary operator (the product of two unitaries and one antiunitary), acting on spacetime arguments by $x \mapsto -x$ and exchanging particles with antiparticles.

Theorem 105.34 (CPT)**.** *Let a quantum field theory satisfy*

1. ***Lorentz invariance**: the fields transform under a representation of the proper orthochronous group $L_{+}^{\uparrow}$ of Definition 39.5, and the dynamics is invariant under it and under translations;*

2. ***locality**: fields at spacelike separation commute (integer spin) or anticommute (half-integer spin);*

3. ***the spectrum condition****: the energy-momentum spectrum lies in the closed forward light cone, with a unique Poincaré-invariant vacuum;*

4. ***hermiticity****: the Hamiltonian is Hermitian, so the evolution is unitary.*

Then there exists an antiunitary Θ *with* $\Theta\,|0\rangle = |0\rangle$ *that commutes with the Hamiltonian: CPT is an exact symmetry [1049, 1055, 1079, 1115, 1123].*

Rests on Definition 39.5, Lemma 105.35, and Equation (100.2).

Two proofs exist and they are of quite different kinds. The first works at the level of a Lagrangian and is short enough to give in full; the second is the theorem proper, since it assumes nothing about Lagrangians, and its machinery is not yet available in this book.

The Lagrangian-level proof

Lemma 105.35 (Every Lorentz index costs a sign)**.** *Let* $O^{\mu_1\cdots\mu_n}(x)$ *be a local monomial built from Dirac bilinears, gauge fields and derivatives, carrying* n *uncontracted Lorentz indices. Then*

$$\Theta\, O^{\mu_1\cdots\mu_n}(x)\,\Theta^{-1} = (-1)^n\,[O^{\mu_1\cdots\mu_n}(-x)]^\dagger\,. \tag{105.94}$$

Rests on Equations (105.6), (105.17) and (105.30).

Proof. *Derives Lemma 105.35.*

For the five Dirac bilinears the rule is the last column of Table 105.1, obtained by multiplying the three preceding columns; the dagger appears because $\mathcal{T}$ is antiunitary and reverses the order of operators in a product. Inspection confirms the count: the scalar and pseudoscalar ($n=0$) are even, the vector and axial vector ($n=1$) are odd, the tensor ($n=2$) is even — and, which is the essential point, CPT does not distinguish scalar from pseudoscalar nor vector from axial, so only n enters.

For the gauge field, $\mathcal{P}A^\mu\mathcal{P}^{-1}$ and $\mathcal{T}A^\mu\mathcal{T}^{-1}$ each lower the index (Equations (105.6) and (105.30)), so the product returns A^μ with sign $+1$, while $\mathcal{C}$ contributes -1 by Equation (105.17): total $(-1)^1$, as Equation (105.94) requires for $n=1$.

For a derivative, let $O'(x) = \Theta O(x)\Theta^{-1} = s\,O^\dagger(-x)$ with s the sign already established. Differentiating, $\partial_\mu\left[O^\dagger(-x)\right] = -\,(\partial_\mu O)^\dagger(-x)$ by the chain rule, so ∂_μ contributes one extra factor -1 and one extra index, preserving Equation (105.94). A product of such objects multiplies the signs and adds the index counts, so the rule is closed under multiplication. □

Proof of Theorem 105.34, Lagrangian level. *Derives Theorem 105.34.*

A Lorentz-scalar Lagrangian density is a sum of monomials in which every index is contracted, either with the metric $\eta_{\mu\nu}$ or with the Levi-Civita symbol $\varepsilon^{\mu\nu\rho\sigma}$. In either case indices are used in pairs — η takes two, and ε takes four — so the total number of Lorentz indices appearing in any scalar monomial is even. By Lemma 105.35 the sign $(-1)^n$ is therefore $+1$, and

$$\Theta\,\mathcal{L}(x)\,\Theta^{-1} = \mathcal{L}^\dagger(-x) = \mathcal{L}(-x)\,, \tag{105.95}$$

the last step by hypothesis (iv), hermiticity of $\mathcal{L}$. The action Equation (100.2) is then invariant,

$$\Theta S\Theta^{-1} = \frac{1}{c}\int \mathrm{d}^4x\,\mathcal{L}(-x) = \frac{1}{c}\int \mathrm{d}^4x'\,\mathcal{L}(x') = S\,,$$

because the Jacobian of $x \to -x$ in four dimensions is $+1$. Hence $[\Theta, H] = 0$.

Two hypotheses have been used silently and must be flagged. First, reordering the fermion operators inside a normal-ordered monomial — which the dagger in Equation (105.94) requires — produces a sign that is $+1$ only if half-integer-spin fields anticommute; had a spinor field been quantized with commutators the argument would fail at exactly this step. This is hypothesis (ii), and its appearance here is the connection between the CPT theorem

and spin–statistics noticed by Pauli [854] and made precise by Lüders and Zumino [1151]. Second, the fields have been assumed to carry finite-dimensional representations built from vectors and spinors, which is what makes the index count exhaustive. □

The Lagrangian argument is complete and correct as far as it goes, and it is what one uses in practice: given a candidate interaction, count its indices. What it does not do is prove the theorem, because it presupposes a Lagrangian, a perturbative expansion, and a set of fields transforming in specified representations. The theorem proper assumes none of these.

The axiomatic proof, honestly Jost's proof [1115], and the canonical account in Streater and Wightman [1297], work directly with the vacuum expectation values

$$W_n(x_1,\dots,x_n) = \langle 0|\,\phi(x_1)\cdots\phi(x_n)\,|0\rangle\ , \tag{105.96}$$

the Wightman functions [1099]. The structure of the argument is the following, and it is worth seeing even though the machinery is reserved.

The spectrum condition (iii) says the energy-momentum of every state lies in the forward cone. Fourier-transforming Equation (105.96) in the differences $\xi_j = x_j - x_{j+1}$, the support of the transform is confined to the forward cone, which by the Paley–Wiener theorem — imported here, not proved in this book, and not among the results of Chapter 17 — makes W_n the boundary value of a function analytic in the *forward tube*, the domain in which each ξ_j has an imaginary part in the open forward cone. Lorentz covariance (i) then extends the domain: the *extended tube* is the union of the images of the forward tube under the *complex* proper Lorentz group $L_+(\mathbb{C})$, and W_n continues analytically to all of it.

The decisive fact is a statement about that group and nothing else. $L_+(\mathbb{C})$ is *connected*, and it contains the total inversion $-\mathbb{1}$, which corresponds to $x \mapsto -x$. Over the reals, $-\mathbb{1}$ lies in $L_+^{\downarrow}$, a different component from the identity (Equation (39.9)), and cannot be reached; over the complexes it can, by a continuous path through complex "rotations" by an imaginary angle. Analytic continuation along that path therefore gives

$$W_n(x_1,\dots,x_n) = W_n(-x_1,\dots,-x_n)\,, \tag{105.97}$$

and Jost's theorem states that Equation (105.97) is equivalent to *weak local commutativity* at the Jost points — the totally spacelike configurations — which is a weakened form of hypothesis (ii). Equation (105.97) is the CPT condition, and from it the operator Θ is reconstructed.

This is also the clean answer to the question a reader should be asking: why CPT and not CP, or T? Because $-\mathbb{1}$ is in $L_+(\mathbb{C})$ while Λ_P and Λ_T separately have determinant -1 and are in no component of $L_+(\mathbb{C})$ at all. The theorem protects exactly the combination that the complexified group can reach, and nothing else. No amount of ingenuity extends it to CP.

Derivation pending #160

The axiomatic CPT theorem: Jost's proof requires the Wightman axioms, the analyticity of the Wightman functions in the forward tube, the extended tube and the edge-of-the-wedge theorem. None of that machinery is developed here; the axiomatic-QFT chapter reserves it, and the proof belongs in Appendix A alongside it. What is proved in this chapter is the Lagrangian-level theorem, which is weaker in that it assumes a Lagrangian, a field content and a perturbative expansion.

What the theorem forbids The consequences are three, and each is a short argument from antiunitarity. All three use the identity Equation (105.26); the reader who has not read that line should read it now, because everything below is a corollary of it.

Theorem 105.36 (Equal masses)**.** *A particle and its antiparticle have equal masses.*
Rests on Definition 105.33, Theorem 105.34, and Equation (105.26).

Proof. *Derives Theorem 105.36.*

Let $|a\rangle$ be a one-particle state at rest with spin projection s, so $H|a\rangle = m_a c^2 |a\rangle$ and hence $\langle a| H |a\rangle = m_a c^2$, a real number. By Definition 105.33, $\Theta|a\rangle = |\bar{a}\rangle$ up to a phase (the momentum is zero and so is unaffected; the spin projection reverses, which does not matter since the rest energy does not depend on it). Apply Equation (105.26) with $\mathcal{A} = \Theta$, $O = H$, $\varphi = \psi = |a\rangle$, using $\Theta^{-1} H \Theta = H$ from Theorem 105.34:

$$m_{\bar{a}} c^2 = \langle \Theta a| H |\Theta a\rangle = \langle a| H |a\rangle^* = m_a c^2 .$$

The phase cancels between bra and ket. □

Theorem 105.37 (Opposite charges and opposite magnetic moments)**.** *A particle and its antiparticle carry electric charges of equal magnitude and opposite sign, and magnetic moments of equal magnitude and opposite sign.*
Rests on Lemma 105.35 and Equation (105.26).

Proof. *Derives Theorem 105.37.*

Both statements follow from the transformation of the current operator. By Lemma 105.35 with $n = 1$, $\Theta j^\mu(x)\Theta^{-1} = -j^\mu(-x)$, the dagger being irrelevant since j^μ is Hermitian.

For the charge, $Q = \frac{1}{c}\int \mathrm{d}^3x\, j^0(x)$, so $\Theta Q \Theta^{-1} = -Q$ (the spatial integration measure is invariant under $\boldsymbol{x} \to -\boldsymbol{x}$, and the time argument is a label). Then Equation (105.26) gives $q_{\bar{a}} = \langle \Theta a| Q |\Theta a\rangle = \langle a| \Theta^{-1} Q \Theta |a\rangle^* = -q_a$.

For the magnetic moment, the operator is $\hat{\boldsymbol{\mu}} = \frac{1}{2}\int \mathrm{d}^3x\, \boldsymbol{x} \times \boldsymbol{j}(x)$. Under Θ the current picks up a minus sign *and* its spatial argument is reflected; substituting $\boldsymbol{x} \to -\boldsymbol{x}$ in the integral, the two minus signs — one from $\boldsymbol{x}$ and one from $\boldsymbol{j}$ — cancel, so

$$\Theta \hat{\boldsymbol{\mu}} \Theta^{-1} = +\hat{\boldsymbol{\mu}} .$$

Define μ_a as the expectation of $\hat{\mu}_z$ in the state of maximal spin projection along z. Since $\Theta|a, s_z = +\rangle = |\bar{a}, s_z = -\rangle$ — $\mathcal{T}$ and $\mathcal{P}$ together reverse the spin, $\mathcal{P}$ leaving it alone and $\mathcal{T}$ reversing it — Equation (105.26) gives

$$\langle \bar{a}, -| \hat{\mu}_z |\bar{a}, -\rangle = \langle a, +| \hat{\mu}_z |a, +\rangle^* = \mu_a ,$$

and rotational invariance gives $\langle \bar{a}, -| \hat{\mu}_z |\bar{a}, -\rangle = -\mu_{\bar{a}}$. Hence $\mu_{\bar{a}} = -\mu_a$. □

Theorem 105.37 is exactly what the electron–positron comparison of Chapter 112 tests, and it discharges the prediction recorded there at Phenomenon 112.13: the two g factors must agree in magnitude and the moments must point oppositely relative to the spin. They agree to about two parts in 10^{12} [1798].

Theorem 105.38 (Equal total lifetimes)**.** *A particle and its antiparticle have equal total decay widths, and hence equal lifetimes — although their individual partial widths need not be equal.*
Rests on Theorem 105.34 and Equation (105.26).

Proof. *Derives Theorem 105.38.*

Write $S = \mathbb{1} + \mathrm{i}T$ with T the transition operator; unitarity $S^\dagger S = \mathbb{1}$ gives the optical relation

$$-\mathrm{i}\,(T_{ii} - T_{ii}^{*}) = \sum_{f} |T_{fi}|^{2}\,, \tag{105.98}$$

whose right-hand side is, up to kinematic factors, the total width Γ_i. The left-hand side is $2\,\mathrm{Im}\,T_{ii}$, since $T_{ii} - T_{ii}^{*} = 2\mathrm{i}\,\mathrm{Im}\,T_{ii}$, and it is *positive* — a sum of squares — which fixes the sign once and for all. Now CPT invariance means $\Theta S \Theta^{-1} = S$, and Equation (105.26) converts this into

$$T_{fi} = T_{(\Theta i)(\Theta f)}\,, \tag{105.99}$$

an equality between the amplitude for $i \to f$ and that for the CP-conjugated, motion-reversed process. Setting $f = i$ and summing Equation (105.98) over all channels,

$$\Gamma_i = 2\,\mathrm{Im}\,T_{ii} = 2\,\mathrm{Im}\,T_{(\Theta i)(\Theta i)} = \Gamma_{\Theta i}\,,$$

so the total width of the antiparticle equals that of the particle. The individual terms $|T_{fi}|^2$ in Equation (105.98) need not match their conjugates, because Equation (105.99) pairs $i \to f$ with $\bar{f} \to \bar{i}$ and not with $\bar{i} \to \bar{f}$; unitarity forces the differences to cancel in the sum. □

The last sentence is the whole reason direct CP violation is subtle, and it is worth extracting as a separate statement.

Proposition 105.39 (A direct CP asymmetry needs two amplitudes and two kinds of phase)**.** *Let $\mathcal{A} = a_1 \mathrm{e}^{\mathrm{i}\delta_1}\mathrm{e}^{\mathrm{i}\phi_1} + a_2 \mathrm{e}^{\mathrm{i}\delta_2}\mathrm{e}^{\mathrm{i}\phi_2}$ with a_k real, ϕ_k CP-odd (weak) phases and δ_k CP-even (rescattering) phases, so that the conjugate amplitude is $\bar{\mathcal{A}} = a_1 \mathrm{e}^{\mathrm{i}\delta_1}\mathrm{e}^{-\mathrm{i}\phi_1} + a_2 \mathrm{e}^{\mathrm{i}\delta_2}\mathrm{e}^{-\mathrm{i}\phi_2}$. Then*

$$|\mathcal{A}|^2 - \left|\bar{\mathcal{A}}\right|^2 = -4 a_1 a_2 \,\sin(\delta_1 - \delta_2)\sin(\phi_1 - \phi_2)\,. \tag{105.100}$$

A direct asymmetry therefore requires two interfering amplitudes, a difference of weak phases, and *a difference of strong phases.*

Rests on Equation (105.28) and Theorem 105.38.

Proof. *Derives Proposition 105.39.*

$|\mathcal{A}|^2 = a_1^2 + a_2^2 + 2a_1 a_2 \cos(\delta_1 - \delta_2 + \phi_1 - \phi_2)$, and $\left|\bar{\mathcal{A}}\right|^2$ is the same with $\phi_1 - \phi_2 \to -(\phi_1 - \phi_2)$. Subtracting and using $\cos(u+v) - \cos(u-v) = -2\sin u \sin v$ gives Equation (105.100). □

This explains a great deal at once: why ϵ' in Equation (105.72) is a *difference* of two phases; why the strong rescattering phases δ_0, δ_2 appear there at all; why the charm asymmetry of Equation (105.83) cannot be predicted, its strong phases being incalculable; and why an inclusive total rate can never show a CP asymmetry, since summing over final states washes out the strong phases and Theorem 105.38 takes over.

105.5.2 Consequences and loopholes

The companion theorem Hypothesis (ii) of Theorem 105.34 assigns commutators to integer spin and anticommutators to half-integer spin. That assignment is not a convention: it is forced by the same hypotheses, and the result is the spin–statistics theorem [854]. The two theorems are so closely linked that Lüders and Zumino proved a converse — CPT invariance, plus the usual axioms, implies the correct connection between spin and statistics [1151].

The essential computation is visible already for free fields. For a real scalar field the commutator at arbitrary separation is

$$[\phi(x),\phi(y)] = \mathrm{i}\hbar c^2\,\Delta(x-y)\,, \qquad \Delta(x) = -\mathrm{i}\int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k}\left(\mathrm{e}^{-\mathrm{i}k\cdot x} - \mathrm{e}^{+\mathrm{i}k\cdot x}\right), \tag{105.101}$$

the Pauli–Jordan function, in which $k^\mu = (\omega_k/c, \boldsymbol{k})$ is on shell, $\omega_k = c\sqrt{\boldsymbol{k}^2 + m^2c^2/\hbar^2}$, so that $k\cdot x = k_\mu x^\mu$ is a pure number and Δ carries $\mathrm{s/m^3}$. The powers of $\hbar$ and c in front are then forced: the canonical commutator $[\phi(t,\boldsymbol{x}),\pi(t,\boldsymbol{y})] = \mathrm{i}\hbar\,\delta^3(\boldsymbol{x}-\boldsymbol{y})\mathbb{1}$ with $\pi = c^{-2}\partial_t\phi$ makes ϕ^2 an energy per unit length, $\mathrm{J/m}$, and $\hbar c^2 \times \mathrm{s/m^3}$ is exactly that. With $\mathrm{i}\hbar c$ in place of $\mathrm{i}\hbar c^2$ the two sides would differ by a factor of c.

Δ is Lorentz invariant and odd, so it vanishes for spacelike separation — a spacelike interval can be reversed by a proper orthochronous Lorentz transformation, and an odd invariant function of a reversible argument is zero. Had the same field been quantized with anticommutators, the object appearing would be the *even* combination Δ_1, which is nonzero everywhere outside the light cone: microcausality would fail. For the Dirac field the situation is exactly reversed: the anticommutator $\{\psi(x),\bar\psi(y)\} \propto (\mathrm{i}\hbar\gamma^\mu\partial_\mu + mc)\Delta(x-y)$ vanishes at spacelike separation because Δ does, while the commutator would involve Δ_1 and would not. A second, independent strand of the argument is that a scalar field quantized with anticommutators has no ground state, its Hamiltonian being unbounded below, while a spinor field quantized with commutators produces negative-norm states. Either way the wrong choice destroys one of the hypotheses that Theorem 105.34 needs.

Derivation pending #161

The spin–statistics theorem for arbitrary spin: the free-field computation sketched here settles spin 0 and spin 1/2, but the general statement requires the analyticity of the two-point Wightman function and the same extended-tube machinery as the axiomatic CPT proof. It belongs in Appendix A next to it, and the identical-particle chapter of Part IX states the result it needs.

The physical content is the exclusion principle of Chapter 83, and it is worth remembering how much rests on it: the stability of matter, the periodic table, the degeneracy pressure of Chapter 49. That a statement about analytic continuation in complexified Lorentz transformations underwrites the structure of the periodic table is one of the stranger facts in physics, and it is not a coincidence — both follow from locality and positivity of the energy.

Parametrizing a violation If CPT were violated, how would one report the fact? The answer matters because experiments in wholly different systems — an antihydrogen atom, a kaon beam, a trapped antiproton — must be compared, and each measures something different. The standard device is an effective field theory in which every Lorentz- and CPT-violating operator that can be built from the observed fields is added to the Standard Model Lagrangian with an arbitrary coefficient [1952]. A term such as $b_\mu\bar\psi\gamma^\mu\gamma^5\psi$, with b_μ a fixed background four-vector rather than a field, is CPT-odd by Lemma 105.35 — it has one uncontracted index, since b_μ does not transform — and each experiment then bounds some combination of the coefficients. The tabulated bounds run to hundreds of entries [2176].

It is important to be clear about the status of this construction. It is not a theory for which there is evidence; it is a bookkeeping scheme for null results, and its value is precisely that it lets a limit from one system be compared with a limit from another. The coefficients are all consistent with zero.

The relation to Lorentz invariance

Proposition 105.40 (Greenberg)**.** *In an interacting local quantum field theory satisfying the usual axioms, violation of CPT implies violation of Lorentz invariance. The converse is false: there are Lorentz-violating theories that preserve CPT [2028].*

Rests on Theorem 105.34 and Lemma 105.35.

The asymmetry is easy to see in the parametrization above. An operator with an odd number of uncontracted indices, contracted with a fixed background vector, is both CPT-odd and Lorentz-violating; an operator with an even number, contracted with a fixed background tensor, is CPT-even but still Lorentz-violating. Hence CPT-violating $\subset$ Lorentz-violating, strictly. Two consequences follow for experiment. A CPT test is automatically a Lorentz test, and in some channels the sharpest one available. And a search for Lorentz violation cannot be replaced by a CPT test, because the CPT-even sector is invisible to it.

Why these are tests of the axioms The hypotheses of Theorem 105.34 are, essentially verbatim, the Wightman axioms of Chapter 109. That is what gives a CPT measurement its unusual standing. A discrepancy between the electron's and the positron's g factor would not be evidence for some new particle or some unmeasured coupling; it would mean that one of locality, Lorentz invariance, or the positivity of the energy spectrum is false. There are very few measurements in physics whose failure would be that structural, and it is the reason the null results below are worth the effort they cost.

105.5.3 Experimental tests

Table 105.2: Tests of CPT invariance. The quantities compared are predicted equal (or equal and opposite) by Theorems 105.36 to 105.38. The kaon entry is the customary one and is discussed in the text: it is a fraction of $m_K c^2$, and the underlying bound on the energy is far weaker than the exponent suggests. Every entry is consistent with exact CPT.

system	quantity compared	fractional limit	source
K^0 and $\bar{K}^0$	rest energy	6×10^{-19}	[2427]
antihydrogen and hydrogen	$1S$–$2S$ transition frequency	2×10^{-12}	[2299]
$\bar{p}$ and p	magnetic moment	1.5×10^{-9}	[2288]
$\bar{p}$ and p	charge-to-mass ratio	1.6×10^{-11}	[2400]
e^+ and e^-	g factor	2×10^{-12}	[1798]

The kaon bound, and how to read it The neutral kaon entry is the tightest number in Table 105.2 and the one most often misquoted. What is measured is a combination of the semileptonic charge asymmetries of K_L and K_S together with the phase of η_{+-}, which bounds the CPT-violating diagonal difference $M_{11} - M_{22}$ of Proposition 105.23. The bound on that *energy* is of order $3 \times 10^{-10}\,\mathrm{eV}$ — roughly 10^{-4} of the K_L–K_S mass difference Equation (105.60), which is itself a $3.5\,\mu\mathrm{eV}$ splitting. Dividing by $m_K c^2 = 497.6\,\mathrm{MeV}$ turns it into 6×10^{-19}. The exponent therefore reflects the largeness of the denominator as much as the sharpness of the measurement, and comparing it with the 10^{-12} of an atomic experiment is comparing two different things. Both statements are true; neither is "the best CPT test" without a choice of what to divide by.

Antihydrogen The ALPHA collaboration at CERN traps antihydrogen, formed by mixing antiprotons from the Antiproton Decelerator with positrons, in a magnetic minimum, and performs laser spectroscopy on the trapped sample. The two-photon $1S$–$2S$ transition, a

natural linewidth of about 1 Hz on a frequency of 2.466×10^{15} Hz, is the sharpest observable in hydrogen. ALPHA first observed the transition in antihydrogen [2280] and then measured its frequency, in zero field, in agreement with hydrogen at the level of 2×10^{-12} [2299]. The comparison tests Theorem 105.36 and Theorem 105.37 jointly, since the transition frequency depends on the electron and proton masses and charges in combination.

The antiproton Two BASE measurements bound the antiproton's properties directly. The magnetic moment, measured on a single trapped antiproton by the continuous Stern–Gerlach effect, is $\mu_{\bar{p}} = -2.7928473441(42)\mu_N$, agreeing with the proton's in magnitude to 1.5 parts in 10^9 and opposite in sign [2288] — exactly Theorem 105.37. The charge-to-mass ratio, measured by comparing the cyclotron frequencies of an antiproton and a negative hydrogen ion in the same trap, agrees with the proton's to 16 parts in 10^{12} [2400].

What CPT does not say The ALPHA-g experiment released antihydrogen from a vertical trap and found that it falls, with a measured acceleration $a_g = (0.75 \pm 0.13 \pm 0.16)\, g$ downward [2417]. This is a fine measurement and it is *not* a test of CPT. The theorem relates inertial masses, charges and moments; it says nothing about gravitational coupling, and a hypothetical antigravity would violate the equivalence principle of Chapter 42 without disturbing any hypothesis of Theorem 105.34. Reporting it as a CPT test, as several popular accounts did, confuses two independent structures. Every genuine CPT test performed to date is consistent with exact invariance.

105.6 The baryon asymmetry

105.6.1 Sakharov's conditions

The universe contains matter and essentially no antimatter. If it began in a state symmetric under C and CP, with equal numbers of baryons and antibaryons, then getting from there to here requires three things, identified by Sakharov in a two-page paper of 1967 [1358]. All three are necessary, and each necessity is a theorem.

Theorem 105.41 (Sakharov's conditions)**.** *Let a system begin with $\langle B \rangle = 0$, where B is baryon number. Then $\langle B \rangle = 0$ at all later times if any one of the following holds:*

1. *B is conserved, $[B, H] = 0$;*
2. *C and CP are symmetries of H and of the initial state;*
3. *the system is in thermal equilibrium.*

Generating a baryon asymmetry therefore requires baryon number violation, C and CP violation, and a departure from thermal equilibrium.

Rests on Theorem 105.34 and Equation (105.34).

Proof. *Derives Theorem 105.41.*

(i) is immediate: a conserved quantity that starts at zero stays there.

(ii) Let $\rho(t)$ be the density operator and suppose $\mathcal{C}H\mathcal{C}^{-1} = H$ and $\mathcal{C}\rho(0)\mathcal{C}^{-1} = \rho(0)$, whence $\mathcal{C}\rho(t)\mathcal{C}^{-1} = \rho(t)$ for all t since the evolution commutes with $\mathcal{C}$. Baryon number is C-odd, $\mathcal{C}B\mathcal{C}^{-1} = -B$. Then, using cyclicity of the trace and the unitarity of $\mathcal{C}$,

$$\langle B \rangle = \operatorname{tr}\left[\rho(t) B\right] = \operatorname{tr}\left[\mathcal{C}\rho(t)\mathcal{C}^{-1}\, \mathcal{C}B\mathcal{C}^{-1}\right] = -\operatorname{tr}\left[\rho(t) B\right] = 0\,.$$

The same computation with $\mathcal{CP}$ in place of $\mathcal{C}$ gives the same conclusion, and both are needed: a theory may violate C maximally and still conserve CP — the $V - A$ interaction of

Equation (105.34) is the standard example — in which case the CP-version of the argument still forces $\langle B\rangle = 0$. It is the *pair* of symmetries that must fail.

(iii) is the deepest of the three, and it is where CPT enters. In thermal equilibrium at temperature T the density operator is $\rho = \mathrm{e}^{-H/k_BT}/Z$; there is no chemical potential for B, because by hypothesis (i) has failed and B is not conserved, so it cannot label a conserved charge. Now $\langle B\rangle = \mathrm{tr}(\rho B)$ is real, being the trace of a product of two Hermitian operators. For an antiunitary Θ and any operator O, choosing the orthonormal basis $\{\Theta e_n\}$ to evaluate the trace gives

$$\mathrm{tr}\left(\Theta O \Theta^{-1}\right) = \sum_n \langle \Theta e_n |\, \Theta O\, | e_n\rangle = \sum_n \langle O e_n | e_n\rangle = (\mathrm{tr}\, O)^* \,. \tag{105.102}$$

Apply this with $O = \mathrm{e}^{-H/k_BT}B$. By Theorem 105.34 $\Theta H \Theta^{-1} = H$, and B is CPT-odd (it is C-odd and P- and T-even), so $\Theta\left(\mathrm{e}^{-H/k_BT}B\right)\Theta^{-1} = -\mathrm{e}^{-H/k_BT}B$. Hence

$$\mathrm{tr}\left(\mathrm{e}^{-H/k_BT}B\right) = -\left[\mathrm{tr}\left(\mathrm{e}^{-H/k_BT}B\right)\right]^* = -\,\mathrm{tr}\left(\mathrm{e}^{-H/k_BT}B\right),$$

the last step because the quantity is real; therefore it vanishes. □

Part (iii) deserves a comment, because it is the point at which the two halves of this chapter meet. The reason equilibrium forbids a baryon asymmetry is *that particles and antiparticles have the same mass*, which is Theorem 105.36: with equal masses and no conserved charge to bias them, the equilibrium populations are equal whatever the interactions. The CPT theorem, which guarantees the symmetry of the world, is thereby also the obstacle to explaining its asymmetry. Any mechanism must arrange for the universe to be out of equilibrium while the baryon-number-violating processes are active.

The Standard Model's baryon-number violation Condition (i) looks hopeless at first sight, since baryon number is an exact symmetry of the classical Standard Model Lagrangian. It is not a symmetry of the quantum theory. The baryon and lepton currents are anomalous: the divergence that vanishes classically acquires a contribution from the non-invariance of the fermion measure [1388, 1391, 1569], with the integrated statement

$$\Delta B = \Delta L = n_f\, \Delta n_{\mathrm{CS}}\,, \tag{105.103}$$

where $n_f = 3$ is the number of generations and $n_{\mathrm{CS}} \in \mathbb{Z}$ is the Chern–Simons number of the SU(2) gauge field, an integer for a transition between distinct vacua. Baryon number is violated; $B - L$ is not. The barrier between adjacent vacua is the sphaleron, a static unstable solution whose energy is set by the $\mathrm{SU}(2)_L$ coupling g and the W mass,

$$E_{\mathrm{sph}} = B\,\frac{8\pi}{g^2}\, m_W c^2 \approx 9\,\mathrm{TeV} \quad \text{at } T = 0\,. \tag{105.104}$$

Here $g^2 = 4\pi\alpha/\sin^2\theta_W = 0.397$, taking $\sin^2\theta_W = 0.231$ and $\alpha = 1/137.036$, so that $g = 0.630$ and $8\pi/g^2 = 63.4$; with $m_W c^2 = 80.4\,\mathrm{GeV}$ the bare combination $(8\pi/g^2)m_W c^2$ is only 5.1 TeV. The factor B is a dimensionless shape factor fixed numerically by the field equations: it depends on nothing but the ratio of the Higgs self-coupling to g^2, runs from about 1.5 to 2.7 across the whole range of that ratio, and is close to 1.8 at the observed Higgs mass. That is where the 9 TeV comes from, and quoting the formula without B — as is often done — understates the barrier by nearly a factor of two. At zero temperature the tunnelling rate then carries the factor $\exp(-4\pi/\alpha_W) \sim 10^{-173}$, with $\alpha_W = g^2/4\pi = 0.0316$, and nothing happens. Above the electroweak scale, $k_BT \gtrsim 100\,\mathrm{GeV}$, the barrier can be crossed thermally and the rate per unit volume is unsuppressed, $\Gamma/V \sim (\alpha_W k_B T)^4/(\hbar c)^3\hbar$

[1758]. The Standard Model therefore satisfies Sakharov's first condition, in the early universe and nowhere else.

Derivation pending #162

The triangle anomaly and the Chern–Simons current: the derivation that the baryon current's divergence is proportional to the topological density, whether by the triangle diagram or by the non-invariance of the path-integral measure, is not carried out in this chapter. It belongs to the path-integral chapter of this part, which reserves the anomaly, and to Appendix A. What is used here is only the integrated statement, that a change of Chern-Simons number by one unit changes baryon number by the number of generations.

105.6.2 Why the Standard Model fails

The number to be explained The asymmetry is quantified by the baryon-to-photon ratio.

Proposition 105.42 (Baryon-to-photon ratio from the CMB)**.**

$$\eta = \frac{n_B - n_{\bar{B}}}{n_\gamma} = 2.74 \times 10^{-8}\, \Omega_b h^2 = (6.12 \pm 0.04) \times 10^{-10}\,, \tag{105.105}$$

using $\Omega_b h^2 = 0.02237 \pm 0.00015$ *from Planck [2348].*
Rests on Equation (68.5) and Phenomenon 51.3.

Derivation. *Derives Proposition 105.42.*
The photon number density of a blackbody at temperature T is

$$n_\gamma = \frac{2\zeta(3)}{\pi^2}\left(\frac{k_B T}{\hbar c}\right)^3 = 4.107 \times 10^8/\mathrm{m}^3$$

at $T_0 = 2.7255\,\mathrm{K}$. The baryon number density is $\Omega_b \rho_c/\bar{m}$ with $\rho_c = 3H_0^2/(8\pi G) = 1.878\,34 \times 10^{-26}\,\mathrm{kg/m^3}\,h^2$ and $\bar{m}$ the mean mass per baryon, close to the proton mass $1.672\,62 \times 10^{-27}\,\mathrm{kg}$; hence $n_B = 11.23/\mathrm{m}^3\,\Omega_b h^2$. The ratio is $11.23/4.107 \times 10^8 = 2.74 \times 10^{-8}$ per unit of $\Omega_b h^2$. □

The same number follows independently from the light-element abundances produced in primordial nucleosynthesis, which depend on η through the neutron-to-proton ratio at freeze-out; the two determinations agree, which is what makes Equation (105.105) a measurement rather than a fit. No antimatter region has been found: a matter–antimatter boundary anywhere within the observable universe would radiate annihilation gamma rays at a level long since excluded, and the antinuclei searched for in the cosmic radiation (Chapter 115) are consistent with secondary production. See Phenomenon 111.17.

The CKM phase is far too small By Proposition 105.27, the strength of Standard Model CP violation is measured by the invariant Equation (105.79). Made dimensionless at the temperature $k_B T \approx 100\,\mathrm{GeV}$ where sphalerons are active, it is

$$d_{CP} = \frac{J \prod_{i>j} \left(m_{u_i}^2 - m_{u_j}^2\right) c^4 \prod_{i>j} \left(m_{d_i}^2 - m_{d_j}^2\right) c^4}{(k_B T)^{12}}\,. \tag{105.106}$$

Inserting the running quark masses at that scale — $m_t c^2 \approx 170\,\mathrm{GeV}$, $m_b c^2 \approx 2.9\,\mathrm{GeV}$, $m_c c^2 \approx 0.6\,\mathrm{GeV}$, $m_s c^2 \approx 50\,\mathrm{MeV}$, $m_d c^2 \approx 3\,\mathrm{MeV}$, $m_u c^2 \approx 1.5\,\mathrm{MeV}$ — and Equation (105.78) gives

$$d_{CP} \approx 2 \times 10^{-21}\,. \tag{105.107}$$

The observed η is 6×10^{-10}. Detailed calculations of what the Standard Model would actually produce, which involve more than this crude ratio, arrive at asymmetries of order 10^{-20} or smaller [1882]: the shortfall is at least eight orders of magnitude and on the standard estimates ten or more. The reason is structural and worth stating: CP violation in the Standard Model requires all six quark masses to be distinct (Proposition 105.27), and at a temperature of 100 GeV five of the six are negligible compared with the temperature, so the effect is suppressed by twelve powers of small mass ratios. The phase itself is of order unity; it is the masses that kill it.

And the third condition fails too Even a larger phase would not help, because the Standard Model does not depart from thermal equilibrium at the electroweak scale. A departure would require a first-order phase transition, proceeding by the nucleation of bubbles of the broken phase whose walls sweep through the plasma; and it would have to be *strongly* first order, since after the transition the sphaleron processes must switch off fast enough not to erase the asymmetry just created. The washout condition is

$$\frac{v(T_c)}{T_c} \gtrsim 1 \tag{105.108}$$

in the customary units, v being the Higgs expectation value at the critical temperature — a requirement that the broken-phase sphaleron energy Equation (105.104) exceed the thermal energy by enough that the Boltzmann factor quenches the rate.

Lattice studies settled the question. The electroweak transition is first order only for a light Higgs; the line of first-order transitions ends at a critical endpoint near $m_H c^2 \approx 72 - -80\,\mathrm{GeV}$, and beyond it the change from the symmetric to the broken phase is a smooth crossover with no bubbles, no walls and no departure from equilibrium [1925]. The measured Higgs mass is $m_H c^2 = (125.20 \pm 0.11)\,\mathrm{GeV}$ [2427], far above the endpoint (Chapter 104). Two of Sakharov's three conditions therefore fail in the Standard Model, one by ten orders of magnitude and one outright.

This is recorded as an open problem, not a solved one: see Section 133.5 in Chapter 133. The measurement — Equation (105.105) — is secure; the mechanism is unknown. It is worth being precise about what is and is not established, because the literature contains many proposals and no evidence for any of them. What is established: baryon number is not conserved in the Standard Model at high temperature; CP is violated by the CKM phase; the observed asymmetry exceeds what those two can produce by a very large factor; and the electroweak transition is a crossover. What is not established: anything about where the missing asymmetry came from.

105.7 The strong CP problem

105.7.1 The theta term

Section 105.1.1 showed that the quark–gluon coupling, being a vector current, conserves P, C and T. That argument is incomplete, because it enumerated only the terms one would think to write. There is one more term allowed by every symmetry that QCD is supposed to have — gauge invariance, Lorentz invariance, locality and renormalizability — and it violates P and T.

Definition 105.43 (The topological term)**.** Let $q(x)$ denote the topological charge density of the gluon field, normalized so that

$$\int \mathrm{d}t\, \mathrm{d}^3x\, q(x) = n \in \mathbb{Z} \tag{105.109}$$

for any finite-action field configuration, n being its winding number; q therefore carries the SI dimension $1/(\mathrm{m}^3\,\mathrm{s})$. The QCD Lagrangian density may contain

$$\mathcal{L}_\theta = \theta\,\hbar\,q(x)\,, \tag{105.110}$$

with θ a dimensionless number. Its contribution to the action is $S_\theta = \theta\hbar n$, so it enters the path-integral weight $\exp(\mathrm{i}S/\hbar)$ of Chapter 106 only through the phase $\mathrm{e}^{\mathrm{i}\theta n}$, and θ is periodic with period 2π.

In terms of the gluon field strengths, q is proportional to $\sum_a \boldsymbol{E}^a \cdot \boldsymbol{B}^a$, the colour analogue of the electromagnetic invariant $\boldsymbol{E}\cdot\boldsymbol{B}$; the constant of proportionality is fixed by Equation (105.109) and involves g_s^2 and $\hbar$ but nothing else. That representation is all that is needed for the symmetry properties.

Proposition 105.44 (The theta term is P-odd, T-odd and C-even)**.** *$\mathcal{L}_\theta$ is odd under P and under T, even under C, and therefore odd under CP and even under CPT.*
Rests on Equations (105.6), (105.17) and (105.30).

Proof. *Derives Proposition 105.44.*
Use the abelian analogue, which carries the same index structure. By Equation (105.6), $\boldsymbol{E}$ is a polar vector and $\boldsymbol{B}$ an axial vector, so $\boldsymbol{E}\cdot\boldsymbol{B}$ is a pseudoscalar: P-odd. By Equation (105.30), $\boldsymbol{E}$ is T-even and $\boldsymbol{B}$ is T-odd, so the product is T-odd. Under C, both fields reverse (Equation (105.17)), so the product is even. Hence CP-odd and, as Theorem 105.34 requires of any term in a Hermitian local Lagrangian, CPT-even. The same signs hold for the non-abelian $\sum_a \boldsymbol{E}^a \cdot \boldsymbol{B}^a$, since the colour index is inert under all three operations. □

Two facts then make the term a problem rather than a curiosity.

First, q is a total derivative: $q = \partial_\mu K^\mu$ for the Chern–Simons current K^μ. A total derivative contributes nothing in perturbation theory, which is why θ never appears in a Feynman diagram; but K^μ is not gauge invariant, and for field configurations with nontrivial topology at infinity — the instantons of [1569, 1573, 1582] — the surface term is the nonzero integer of Equation (105.109). The term is therefore invisible perturbatively and physical nonperturbatively, which is the worst combination for anyone hoping to argue it away.

Second, θ is not by itself observable. A chiral rotation of the quark fields, $\psi_j \to \mathrm{e}^{\mathrm{i}\alpha\gamma^5}\psi_j$, changes the phase of the quark mass matrix and, because the fermion measure is not invariant under it [1388, 1391], shifts θ by a compensating amount. Only the combination

$$\bar\theta = \theta + \arg\det M_q \tag{105.111}$$

is invariant, and it is $\bar\theta$ that is measurable. The strong and Yukawa sectors are thereby tied together: a quantity of the gluon sector can be cancelled against a quantity of the Higgs sector, and no symmetry of the Standard Model relates them.

There is exactly one way the problem could evaporate. If any quark were massless, $\det M_q = 0$, $\arg\det M_q$ would be undefined, and the chiral rotation could be used to set $\bar\theta = 0$ with no cost. That escape is closed by measurement: lattice determinations give $m_u c^2 = (2.16 \pm 0.07)\,\mathrm{MeV}$ at the customary renormalization point [2427], tens of standard deviations from zero.

Proposition 105.45 (The induced neutron electric dipole moment)**.** *A nonzero $\bar\theta$ generates a neutron electric dipole moment*

$$d_n \approx \bar\theta \times 1.6 \times 10^{-37} - -6.4 \times 10^{-37}\,\mathrm{C\,m}\,, \tag{105.112}$$

equivalently $(1\text{–}4) \times 10^{-16}\,\bar\theta\,e\,\mathrm{cm}$ *[1633, 1637].*
Rests on Proposition 105.44, Equation (105.111), and Theorem 105.30.

Derivation. *Derives Proposition 105.45.*

The parametric form follows from dimensional analysis plus chiral counting, and is worth doing because it explains the size. A neutron EDM requires the violation of P and T, which $\bar{\theta}$ supplies; it requires a charge, which supplies e; and it requires a length, for which the only candidates are the nucleon Compton wavelength $\hbar/(m_N c) = 0.21\,\mathrm{fm}$ and the pion's. Chiral symmetry supplies a further suppression: in the limit of vanishing quark masses the θ term is unobservable, so the answer must be proportional to the reduced quark mass $m_* = m_u m_d/(m_u + m_d) \approx 1.5\,\mathrm{MeV}/c^2$ measured against the nucleon mass. Hence

$$d_n \sim e\,\bar{\theta}\,\frac{\hbar}{m_N c}\,\frac{m_*}{m_N} \approx \bar{\theta} \times 5 \times 10^{-38}\,\mathrm{C\,m}\,,$$

that is $3 \times 10^{-17}\,\bar{\theta}\,e\,\mathrm{cm}$. The chiral-loop calculations of Baluni [1633] and of Crewther, Di Vecchia, Veneziano and Witten [1637] supply the coefficient, and it is between about three and thirteen times larger than this estimate — $1.6 \times 10^{-37}\,\mathrm{C\,m}/5 \times 10^{-38}\,\mathrm{C\,m} = 3.2$ at one end and 12.8 at the other — because the dominant contribution is a chiral logarithm from the pion cloud rather than a contact term; modern lattice determinations also fall inside Equation (105.112). That spread — a factor of a few in the coefficient — is the dominant uncertainty in the bound on $\bar{\theta}$ below, and it is why the bound is quoted as an order of magnitude and not as a number. □

105.7.2 The experimental bound and its status

Combining Equation (105.93) — the ultracold-neutron measurement at the Paul Scherrer Institute, whose Ramsey technique is described at Equation (105.91) — with Equation (105.112) gives

$$\left|\bar{\theta}\right| \lesssim \frac{2.9 \times 10^{-47}\,\mathrm{C\,m}}{1.6 \times 10^{-37}\,\mathrm{C\,m}} \approx 1.8 \times 10^{-10}\,, \tag{105.113}$$

so $\left|\bar{\theta}\right| \lesssim 10^{-10}$, with the factor-of-a-few uncertainty inherited from Proposition 105.45.

This is the strong CP problem, and it is worth stating exactly what kind of problem it is. $\bar{\theta}$ is a free dimensionless parameter of the Standard Model. It is periodic with period 2π, so there is no sense in which small values are "near a boundary" of its range; a value of order unity is the natural expectation for a number about which nothing else is known. It is measured to be smaller than 10^{-10}. No symmetry of the Standard Model requires it to vanish — the one candidate, a massless quark, is excluded by measurement — and no dynamical mechanism within the Standard Model relaxes it. The problem is therefore an *observation*: a parameter that could have been anything is measured to be extraordinarily small, and nothing in the accepted theory explains why. It is recorded as such in Section 133.1.

The Peccei–Quinn proposal, and its status The best-known response is to make $\bar{\theta}$ dynamical. Peccei and Quinn observed that if the Standard Model is extended by a global U(1) symmetry that is spontaneously broken and itself anomalous, then $\bar{\theta}$ is replaced by a field whose potential — generated by the same instanton effects that make θ physical — is minimized at zero [1609]. The spontaneously broken symmetry implies a light pseudo-Goldstone boson, the axion, as Weinberg [1628] and Wilczek [1630] pointed out at once. Its mass is tied to the symmetry-breaking scale f_a by

$$m_a c^2 \approx 5.7\,\mu\mathrm{eV} \times \frac{1 \times 10^{12}\,\mathrm{GeV}}{f_a c^2}\,, \tag{105.114}$$

so that a very weakly coupled axion is a very light one.

The chapter states the position plainly. No axion has been observed. The haloscope searches, of which ADMX is the most sensitive, look for the resonant conversion of a galactic-halo axion into a microwave photon in a strong magnetic field; the reported results are

exclusions, covering a band of couplings in a narrow mass window around 2.66–2.81 μeV [2355]. Helioscopes and light-shining-through-walls experiments likewise report null results. The Peccei–Quinn mechanism is a hypothesis with an attractive structure and no experimental support, and Section 133.7 lists it accordingly. What this chapter records as physics is the measurement, Equation (105.113): a dimensionless parameter of the Standard Model, free to be of order unity, is smaller than 10^{-10}, and no one knows why.

Chapter 106

Path-Integral Quantization

The canonical quantization of Chapter 99 promotes coordinates to operators and privileges a time coordinate; the path integral does neither. It computes an amplitude as a sum over every history weighted by $\exp(\mathrm{i}S/\hbar)$, with S the classical action of Chapter 16, so the Lorentz invariance of the Lagrangian is manifest at every step. The idea is Dirac's [718] and its execution Feynman's [911, 930]. This chapter develops it from non-relativistic quantum mechanics through the functional quantization of fields: generating functionals, the Euclidean continuation, the LSZ reduction [1073], Feynman rules and the one-particle-irreducible effective action. It therefore follows Chapter 64 and supplies the machinery used throughout Chapters 102 and 104 and Chapter 108.

Three things make the path integral more than a restatement. Gauge theories require it: the Faddeev–Popov construction [1347] and the BRST symmetry that organizes it [1564, 1571] have no simple canonical counterpart, and Gribov showed that the gauge-fixing is not globally well posed [1623]. Non-perturbative physics requires it: Wilson's lattice formulation [1537] turns the Euclidean integral into a finite-dimensional one that a computer can sample, which is how the hadron spectrum of Chapter 102 is computed from first principles [2122]. And it explains anomalies: a classical symmetry can fail to survive quantization because the functional measure is not invariant [1638] — a statement with a measured consequence, the decay rate of the neutral pion [1388, 1391, 2361]. The canonical monographs are [1306, 1751, 1933].

Conventions are those of Section 100.1.1 without exception. The metric is $\eta_{\mu\nu} = \mathrm{diag}(+1,-1,-1,-1)$; coordinates are $x^\mu = (ct, \boldsymbol{x})$, so d^4x has SI dimension m^4; four-momenta are $p^\mu = (E/c, \boldsymbol{p})$ of dimension $\mathrm{kg\,m/s}$, and the mass shell is $p^2 = m^2c^2$ as in Equation (100.1), so every propagator denominator reads $p^2 - m^2c^2$ and never $p^2 - m^2$. The action is $S = c^{-1}\int \mathrm{d}^4x\,\mathcal{L}$ as in Equation (100.2), with $\mathcal{L}$ an energy density; the Fourier convention is Equation (100.14). Every $\hbar$ and every c is written out, and here that is not a matter of taste: the expansion parameter of this whole chapter *is* $\hbar$. The loop expansion of Section 106.3.4 is an expansion in $\hbar$ and in nothing else, and the semiclassical limit of Section 106.1.3 is the statement that $\hbar$ is small compared with the classical action. A treatment that puts the reduced Planck constant equal to unity erases the very quantity whose smallness is being exploited. Readers coming from the standard references should use the translation rule of Remark 100.4, which is the only place in this part where that convention is discussed.

Five symbols carry more than one standard meaning in the material of this chapter, and the collisions are settled here rather than in passing. The *electric charge* of a particle or a field is written Q throughout, never q: the letter q is the mechanical coordinate of Section 106.1 and the topological charge density of Definition 106.79, and both meet a charge on the same page. The *BRST charge* of Definition 106.59 is therefore written Q_B. The *ghost* fields are c^a and $\bar{c}^a$, always carrying an adjoint index or an explicit generator, so

a bare c is the speed of light and nothing else. The lattice coupling β of Equation (106.81) is not the inverse temperature β of Equation (106.27); each name is fixed by its own literature, and they never enter the same equation here. And Γ is the effective action of Definition 106.41 everywhere except in Section 106.7.2, where it is a decay width, and at one step of Proposition 106.45, where it is the Euler function and is flagged as such on the spot.

106.1 The path integral in quantum mechanics

106.1.1 Dirac's Lagrangian formulation

Take one degree of freedom q with Hamiltonian $H = p^2/2m + V(q)$, both operators, and write the finite-time propagator

$$K(q'', t''; q', t') := \langle q'' | \, U(t'', t') \, | q' \rangle \,, \qquad U(t'', t') = \exp\!\left[-\frac{\mathrm{i}}{\hbar} H (t'' - t') \right] , \tag{106.1}$$

for a time-independent H. Since $\langle q'' | q' \rangle = \delta(q'' - q')$, the states $|q\rangle$ carry dimension $1/\mathrm{m}^{1/2}$ and K carries $1/\mathrm{m}$. Everything about the dynamics is in K: given $\psi(q', t')$, the wavefunction at any later time is $\psi(q'', t'') = \int \mathrm{d}q' \, K(q'', t''; q', t') \psi(q', t')$.

The classical object that propagates a mechanical system from (q', t') to (q'', t'') is Hamilton's principal function — the value of the action along the classical path,

$$S_{\mathrm{cl}}(q'', t''; q', t') = \int_{t'}^{t''} \mathrm{d}t \; L(q_{\mathrm{cl}}, \dot{q}_{\mathrm{cl}}, t) \,, \tag{106.2}$$

of dimension J s. It is the type-one generating function of the canonical transformation that maps the state at t' onto the state at t'' (Chapter 23), obeying

$$\frac{\partial S_{\mathrm{cl}}}{\partial q''} = p'' \,, \qquad \frac{\partial S_{\mathrm{cl}}}{\partial q'} = -p' \,, \qquad \frac{\partial S_{\mathrm{cl}}}{\partial t''} + H\!\left(q'', \frac{\partial S_{\mathrm{cl}}}{\partial q''} \right) = 0 \,. \tag{106.3}$$

Dirac's observation [718] was that the quantum object Equation (106.1) and the classical object Equation (106.2) are related by

$$\langle q'' | \, U(t'', t') \, | q' \rangle \quad \text{“corresponds to”} \quad \exp\!\left[\frac{\mathrm{i}}{\hbar} S_{\mathrm{cl}}(q'', t''; q', t') \right] . \tag{106.4}$$

The scare quotes are Dirac's own. He did not claim an equality, and he was right not to: the two sides do not even have the same dimension, the left-hand side carrying $1/\mathrm{m}$ and the right-hand side being a pure phase. What he had was the observation that a canonical transformation is implemented in quantum mechanics by a unitary operator, that the generating function of the classical transformation therefore ought to control the phase of the quantum kernel, and that in the classical limit Equation (106.3) is exactly the eikonal relation satisfied by the phase of a rapidly oscillating wave.

What Feynman added is that Equation (106.4) becomes an *equality* in the infinitesimal case, up to a normalization that carries the missing dimension, and that finite times then follow by composition. The infinitesimal statement is a short computation.

Lemma 106.1 (The short-time kernel)**.** *Let $H = p^2/2m + V(q)$ with V bounded and continuous. Then, for $\Delta t \to 0$,*

$$\langle q'' | \, \mathrm{e}^{-\mathrm{i} H \Delta t / \hbar} \, | q' \rangle = \sqrt{\frac{m}{2\pi \mathrm{i} \hbar \Delta t}} \, \exp\!\left\{ \frac{\mathrm{i} \Delta t}{\hbar} \left[\frac{m}{2} \left(\frac{q'' - q'}{\Delta t} \right)^2 - V\!\left(\frac{q'' + q'}{2} \right) \right] \right\} \left[1 + O((\Delta t)^2) \right] . \tag{106.5}$$

The prefactor carries $1/\mathrm{m}$, the exponent is dimensionless, and the bracket in the exponent is the Lagrangian $L = \tfrac{1}{2} m \dot{q}^2 - V$ evaluated on the straight segment joining the endpoints.

Rests on Lemma 106.2 and Corollary 8.15.

Derivation. *Derives Lemma 106.1.*

The two terms of H do not commute, but over a time Δt the error of separating them is second order: by the Baker–Campbell–Hausdorff expansion,

$$\mathrm{e}^{-\mathrm{i}(T+V)\Delta t/\hbar} = \mathrm{e}^{-\mathrm{i}T\Delta t/\hbar}\,\mathrm{e}^{-\mathrm{i}V\Delta t/\hbar}\exp\!\left[-\frac{(\Delta t)^2}{2\hbar^2}\,[T,V] + O((\Delta t)^3)\right],$$

with $T = p^2/2m$. Dropping the correction and inserting the momentum-space resolution of the identity $\mathbb{1} = \int \mathrm{d}p\;|p\rangle\langle p|$ with $\langle q|p\rangle = (2\pi\hbar)^{-1/2}\mathrm{e}^{\mathrm{i}pq/\hbar}$,

$$\begin{aligned}\langle q''|\,\mathrm{e}^{-\mathrm{i}T\Delta t/\hbar}\mathrm{e}^{-\mathrm{i}V\Delta t/\hbar}\,|q'\rangle &= \mathrm{e}^{-\mathrm{i}\Delta t V(q')/\hbar}\int\frac{\mathrm{d}p}{2\pi\hbar}\,\exp\!\left[\frac{\mathrm{i}}{\hbar}p(q''-q') - \frac{\mathrm{i}\Delta t p^2}{2m\hbar}\right]\\ &= \mathrm{e}^{-\mathrm{i}\Delta t V(q')/\hbar}\sqrt{\frac{m}{2\pi\mathrm{i}\hbar\Delta t}}\,\exp\!\left[\frac{\mathrm{i}m(q''-q')^2}{2\hbar\Delta t}\right], \qquad (106.6)\end{aligned}$$

the p integral being the Gaussian $\int \mathrm{d}p\,\mathrm{e}^{-ap^2/2+bp} = \sqrt{2\pi/a}\,\mathrm{e}^{b^2/2a}$ of Lemma 106.2 below, analytically continued to $a = \mathrm{i}\Delta t/(m\hbar)$ on the boundary of its domain of convergence — legitimate because the integrand is the boundary value of a function holomorphic in $\Re a > 0$ and the contour may be rotated by Corollary 8.15. The square root is the principal branch, $\sqrt{\mathrm{i}} = \mathrm{e}^{\mathrm{i}\pi/4}$.

Finally $V(q')$ may be replaced by $V\big((q''+q')/2\big)$: the two differ by $\tfrac12(q''-q')V' + O((q''-q')^2)$, and the prefactor confines $|q''-q'|$ to order $\sqrt{\hbar\Delta t/m}$, so the change in the exponent is of order $(\Delta t)^{3/2}$; it cancels against its mirror image on symmetrizing, leaving an error $O((\Delta t)^2)$. The midpoint form is not an aesthetic choice, and Section 106.1.3 shows where it becomes compulsory.

The prefactor's dimension is $[\,m/(\hbar\Delta t)\,]^{1/2} = [\mathrm{kg}/(\mathrm{J\,s}^2)]^{1/2} = 1/\mathrm{m}$, as K requires. □

So Dirac's correspondence is an equality once the normalization $(m/2\pi\mathrm{i}\hbar\Delta t)^{1/2}$ is supplied, and once the principal function is read as the action of the *straight* segment, which for infinitesimal times is the classical path to the accuracy retained. The whole of the next subsection is the observation that finite times are built out of infinitesimal ones, and that when they are, the straight segments become arbitrary paths.

106.1.2 Feynman's sum over histories

The Gaussian integral, once and for all Every explicit evaluation in this chapter reduces to one integral, so it is worth recording in the generality that will be needed.

Lemma 106.2 (Gaussian integrals)**.** *Let A be a real symmetric positive-definite $n\times n$ matrix and $b\in\mathbb{R}^n$. Then*

$$\int_{\mathbb{R}^n}\mathrm{d}^n x\;\exp\!\Big(-\tfrac12 x^{\mathsf T}Ax + b^{\mathsf T}x\Big) = \frac{(2\pi)^{n/2}}{\sqrt{\det A}}\;\exp\!\Big(\tfrac12 b^{\mathsf T}A^{-1}b\Big)\,. \qquad (106.7)$$

The identity extends by analytic continuation to every complex symmetric A with positive-definite real part, and to the boundary case $A = -\mathrm{i}M$ with M real symmetric and non-singular, where

$$\int_{\mathbb{R}^n}\mathrm{d}^n x\;\exp\!\Big(\tfrac{\mathrm{i}}{2}x^{\mathsf T}Mx\Big) = \frac{(2\pi)^{n/2}}{|\det M|^{1/2}}\;\mathrm{e}^{\mathrm{i}\pi\sigma(M)/4}\,, \qquad (106.8)$$

$\sigma(M)$ being the signature of M, i.e. the number of positive minus the number of negative eigenvalues.

Rests on Equation (17.42) and Lemma 8.9.

Derivation. *Derives Lemma 106.2.*

Diagonalize: $A = O^{\mathsf{T}} D O$ with O orthogonal and $D = \operatorname{diag}(a_1, \dots, a_n)$, $a_i > 0$, which is possible by the finite-dimensional spectral theorem for real symmetric matrices. That theorem is quoted here from textbook linear algebra and is not proved anywhere in this book: the subsection it belongs to is Section 12.4.2, and Remark 11.86 records the same import where it is made in the probability chapter. The substitution $y = Ox$ has unit Jacobian and factorizes the integral into n copies of the one-dimensional case. Completing the square in one dimension, $-\tfrac{1}{2}ay^2 + \beta y = -\tfrac{1}{2}a(y - \beta/a)^2 + \beta^2/(2a)$, and $\int \mathrm{d}u\, \mathrm{e}^{-au^2/2} = \sqrt{2\pi/a}$ — the Gauss integral, which is Equation (17.42) of Chapter 17 evaluated at zero frequency. Reassembling, $\prod_i a_i = \det A$ and $\sum_i \beta_i^2/a_i = b^{\mathsf{T}} A^{-1} b$, which is Equation (106.7).

Both sides of Equation (106.7) are holomorphic in the entries of A on the domain $\Re A > 0$ — the left-hand side because the integral converges absolutely and uniformly there, the right-hand side because $\det A \neq 0$ — so they agree throughout it by the identity theorem (Chapter 8). For the purely imaginary boundary case, diagonalize M and rotate each contour: for $a > 0$,

$$\int_{-\infty}^{\infty} \mathrm{d}u\, \mathrm{e}^{\mathrm{i}au^2/2} = \lim_{\delta \to 0^+} \int_{-\infty}^{\infty} \mathrm{d}u\, \mathrm{e}^{-(\delta - \mathrm{i}a)u^2/2} = \lim_{\delta \to 0^+} \sqrt{\frac{2\pi}{\delta - \mathrm{i}a}} = \sqrt{\frac{2\pi}{a}}\, \mathrm{e}^{\mathrm{i}\pi/4},$$

the limit being uniform on compacta by Cauchy's theorem applied to the wedge $0 \le \arg u \le \pi/4$, whose arc contribution vanishes by the estimation lemma Lemma 8.9. A negative eigenvalue rotates the other way and contributes $\mathrm{e}^{-\mathrm{i}\pi/4}$; collecting the phases gives the signature. □

The phase $\mathrm{e}^{\mathrm{i}\pi\sigma/4}$ in Equation (106.8) is not decoration. It is the origin of the Maslov index, which is what makes the semiclassical propagator of Section 106.1.3 agree with the turning-point counting of Remark 9.79.

Composition

Theorem 106.3 (Feynman's sum over histories)**.** *Let $H = p^2/2m + V(q)$ as above and let $t'' > t'$. Divide the interval into N steps of length $\Delta t = (t'' - t')/N$ and write $q_k = q(t' + k\Delta t)$, $q_0 = q'$, $q_N = q''$. Then*

$$K(q'', t''; q', t') = \lim_{N \to \infty} \left(\frac{m}{2\pi \mathrm{i}\hbar \Delta t} \right)^{N/2} \int \prod_{k=1}^{N-1} \mathrm{d}q_k\, \exp\left[\frac{\mathrm{i}}{\hbar} \sum_{k=1}^{N} \Delta t L_k \right], \tag{106.9}$$

with $L_k = \frac{m}{2}\big((q_k - q_{k-1})/\Delta t\big)^2 - V\big((q_k + q_{k-1})/2\big)$. Written compactly,

$$K(q'', t''; q', t') = \int_{q(t')=q'}^{q(t'')=q''} \mathcal{D}q\, \exp\left(\frac{\mathrm{i}}{\hbar} S[q] \right), \tag{106.10}$$

where $\mathcal{D}q$ denotes the limit in Equation (106.9) and nothing else.
Rests on Equation (106.1) and Lemma 106.1.

Derivation. *Derives Theorem 106.3.*

The evolution operator composes exactly, $U(t'', t') = U(t'', t_{N-1}) \cdots U(t_1, t')$, because H is time-independent. Insert $\mathbb{1} = \int \mathrm{d}q_k\, |q_k\rangle\langle q_k|$ between consecutive factors:

$$K = \int \prod_{k=1}^{N-1} \mathrm{d}q_k \prod_{k=1}^{N} \langle q_k |\, \mathrm{e}^{-\mathrm{i}H\Delta t/\hbar}\, | q_{k-1} \rangle ,$$

which is exact for every N. Now apply Lemma 106.1 to each factor. Each contributes the normalization $(m/2\pi\mathrm{i}\hbar\Delta t)^{1/2}$ — there are N of them, one per step, against $N - 1$

integrations — and the phase $\mathrm{e}^{\mathrm{i}\Delta t L_k/\hbar}$, and the individual errors $O((\Delta t)^2)$ accumulate to $N\cdot O((\Delta t)^2) = O(\Delta t)$, which vanishes with $N\to\infty$. The exponents add, giving the Riemann sum $\sum_k \Delta t L_k$ for the action $S[q]=\int_{t'}^{t''}\mathrm{d}t\,L$ of the piecewise-linear path through the sampled points. □

Three features of Equation (106.9) deserve emphasis, because they are what makes the formalism different from Chapter 99 rather than equivalent to it in disguise.

First, *every* path contributes, and every path contributes with the same magnitude. The integrand's modulus is one; only the phase $S[q]/\hbar$ distinguishes paths. There is no sense in which the classical path is weighted more heavily. Its distinction is that its neighbours share its phase — $\delta S=0$ means precisely that S is stationary — so that contributions near it add coherently while contributions elsewhere cancel. That is the content of Section 106.1.3.

Second, the paths summed over are not classical trajectories and are not even differentiable. The measure concentrates $|q_k-q_{k-1}|$ on $\sqrt{\hbar\Delta t/m}$, so the typical increment scales as $(\Delta t)^{1/2}$ and the typical "velocity" as $(\Delta t)^{-1/2}$, which diverges. The paths dominating Equation (106.9) are continuous and nowhere differentiable, the same class that carries Wiener measure in Section 106.2.2.

Third, the object $S[q]$ is the classical action — a Lorentz scalar in the relativistic case, a functional of the path and of nothing else. No operator ordering appears in Equation (106.10), no Hilbert space, no choice of time slicing beyond the one used to define the limit. That is why the formalism generalizes to gauge fields, where the operator formulation must first solve a constraint problem (Chapter 26) before it can begin.

Equivalence with the Schrödinger equation

Proposition 106.4 (From the kernel to the Schrödinger equation)**.** *Let ψ evolve by $\psi(q,t+\Delta t)=\int\mathrm{d}q'\,K(q,t+\Delta t;q',t)\psi(q',t)$ with K given by Equation (106.5). Then to first order in Δt,*

$$\mathrm{i}\hbar\,\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\,\frac{\partial^2\psi}{\partial q^2} + V(q)\psi\,. \tag{106.11}$$

Rests on Equations (106.5) and (106.8).

Derivation. *Derives Proposition 106.4.*

Write $q'=q+\eta$. Then

$$\psi(q,t+\Delta t)=\sqrt{\frac{m}{2\pi\mathrm{i}\hbar\Delta t}}\int\mathrm{d}\eta\;\exp\!\left[\frac{\mathrm{i}m\eta^2}{2\hbar\Delta t}-\frac{\mathrm{i}\Delta t}{\hbar}V\!\left(q+\tfrac{\eta}{2}\right)\right]\psi(q+\eta,t)\,.$$

The Gaussian factor suppresses all but $\eta=O(\sqrt{\hbar\Delta t/m})$, so expanding to the order that survives means keeping η^0,η^1,η^2 in ψ and $(\Delta t)^0,(\Delta t)^1$ elsewhere:

$$\mathrm{e}^{-\mathrm{i}\Delta t V/\hbar}\approx 1-\frac{\mathrm{i}\Delta t}{\hbar}V(q)\,,\qquad \psi(q+\eta)\approx\psi+\eta\,\frac{\partial\psi}{\partial q}+\frac{\eta^2}{2}\,\frac{\partial^2\psi}{\partial q^2}\,.$$

By Equation (106.8) with $M=m/(\hbar\Delta t)$,

$$\sqrt{\frac{m}{2\pi\mathrm{i}\hbar\Delta t}}\int\mathrm{d}\eta\,\mathrm{e}^{\mathrm{i}m\eta^2/2\hbar\Delta t}=1\,,\qquad \int\mathrm{d}\eta\,\eta\,\mathrm{e}^{\mathrm{i}m\eta^2/2\hbar\Delta t}=0\,,$$

by parity, and differentiating the first with respect to $m/(\hbar\Delta t)$ gives the second moment

$$\sqrt{\frac{m}{2\pi\mathrm{i}\hbar\Delta t}}\int\mathrm{d}\eta\,\eta^2\,\mathrm{e}^{\mathrm{i}m\eta^2/2\hbar\Delta t}=\frac{\mathrm{i}\hbar\Delta t}{m}\,.$$

Collecting,

$$\psi(q,t+\Delta t)=\psi-\frac{\mathrm{i}\Delta t}{\hbar}V\psi+\frac{\mathrm{i}\hbar\Delta t}{2m}\frac{\partial^2\psi}{\partial q^2}+O((\Delta t)^2)\,,$$

and $(\psi(q,t+\Delta t)-\psi(q,t))/\Delta t\to\partial\psi/\partial t$ gives Equation (106.11) after multiplication by $\mathrm{i}\hbar$. Note where the two derivatives came from: the potential from the first-order term of the exponential, the kinetic operator from the *second moment of the Gaussian*, which is why the dimensionless combination controlling the expansion is $\hbar\Delta t/m$ and why the paths have the roughness they do. □

Interference, and the evidence that the action is primitive The two-slit experiment is the elementary case of Theorem 106.3: the paths from source to detector fall into two homotopy classes, one through each slit, and within each class the contributions add coherently around the stationary path. Writing the two class sums as A_1 and A_2, the probability density on the screen is $|A_1+A_2|^2=|A_1|^2+|A_2|^2+2\Re(A_1^*A_2)$, and the phase of $A_1^*A_2$ is $(S_2-S_1)/\hbar$. For a free particle of momentum p the stationary paths are straight, $S=p\ell$ along a segment of length ℓ, so slits separated by d and a screen at distance $L\gg d$ give a path difference $\ell_2-\ell_1\approx dx/L$ and fringes wherever $p\,dx/(L\hbar)$ is a multiple of 2π: the spacing is $2\pi\hbar L/(pd)=hL/(pd)$, the de Broglie result of Chapter 71. That the pattern is built out of individual particles, each of which arrives at one point and none of which "goes through both slits", was shown directly by Tonomura and co-workers, who recorded the electron-by-electron accumulation of the interference pattern from a biprism [1826].

The sharpest evidence that the path integral's primitive quantity is the action, and not the force, is the Aharonov–Bohm effect.

Phenomenon 106.5 (The Aharonov–Bohm effect)**.** An electron beam split into two arms that enclose a magnetic flux Φ confined entirely to a region the beam never enters shows an interference pattern whose fringes shift with Φ, periodically with period

$$\Phi_0=\frac{h}{e}=4.135\,667\,696\times10^{-15}\,\mathrm{Wb}\,,\tag{106.12}$$

even though the electric and magnetic fields vanish identically along both arms [1164].

Rests on Theorem 106.3 and Equation (100.8)*.*

Derivation. *Derives Phenomenon 106.5.*

The Lagrangian of a particle of charge Q in an electromagnetic field is $L=\frac12 m\dot{\boldsymbol{x}}^2+Q\,\dot{\boldsymbol{x}}\cdot\boldsymbol{A}-Q\phi$, whose Euler–Lagrange equation is $m\ddot{\boldsymbol{x}}=Q(\boldsymbol{E}+\dot{\boldsymbol{x}}\times\boldsymbol{B})$, the Lorentz force; that derivation belongs to Chapter 64, with the covariant coupling $-A_\mu J^\mu$, and is not repeated here. Along any path γ the action therefore contains the term

$$S_{\mathrm{mag}}[\gamma]=Q\int_\gamma \mathrm{d}t\;\dot{\boldsymbol{x}}\cdot\boldsymbol{A}=Q\int_\gamma \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{x}\,,$$

which depends on the path but not on its parametrization. For two paths γ_1,γ_2 with the same endpoints, the difference of the two actions is the circulation around the closed loop $\gamma_2-\gamma_1$,

$$S_{\mathrm{mag}}[\gamma_2]-S_{\mathrm{mag}}[\gamma_1]=Q\oint\boldsymbol{A}\cdot\mathrm{d}\boldsymbol{x}=Q\int_\Sigma\boldsymbol{B}\cdot\mathrm{d}\boldsymbol{S}=Q\Phi\,,$$

by Stokes's theorem, Σ being any surface spanning the loop. The interference term in Theorem 106.3 therefore acquires the phase

$$\Delta\varphi=\frac{Q\Phi}{\hbar}\,,\tag{106.13}$$

and for the electron, $Q=-e$, the pattern repeats when $e\Phi/\hbar = 2\pi$, that is when Φ advances by h/e. The numerical value in Equation (106.12) is exact in the present SI, both h and e being defined constants [2431].

The point is the one Aharonov and Bohm made. The surface Σ necessarily passes through the solenoid, where $\boldsymbol{B}\neq 0$; the paths do not. In a formulation whose primitive quantities are the fields, the effect is impossible; in a formulation whose primitive quantity is $\int \boldsymbol{A}\cdot \mathrm{d}\boldsymbol{x}$, it is the first thing one would predict. What is measurable is not $\boldsymbol{A}$, which is gauge dependent, but the loop integral $\oint \boldsymbol{A}\cdot \mathrm{d}\boldsymbol{x}$, which is not — a gauge transformation Equation (100.8) changes it by $\oint \partial_\mu \Lambda \,\mathrm{d}x^\mu = 0$ for a single-valued Λ. The gauge-invariant content of the potential is exactly its holonomy, and Section 106.5.1 is built on that observation. □

Remark 106.6 (Two flux quanta, and which is which). The period Equation (106.12) is h/e because the interfering object carries charge e. The superconducting flux quantum is $h/2e = 2.067\,833\,848\times 10^{-15}\,\mathrm{Wb}$, half as large, because the interfering object there is a Cooper pair of charge $2e$. Neither number is a property of the electromagnetic field; both are properties of the charge of the thing whose phase is being compared.

106.1.3 Time slicing, the measure and the classical limit

Convergence of the product formula Theorem 106.3 took the splitting $\mathrm{e}^{-\mathrm{i}(T+V)\Delta t/\hbar}\approx \mathrm{e}^{-\mathrm{i}T\Delta t/\hbar}\mathrm{e}^{-\mathrm{i}V\Delta t/\hbar}$ on the strength of a formal expansion. That the product converges as $N\to\infty$ is a theorem about operator semigroups, not about exponentials of numbers, and it is due to Trotter [1175].

Theorem 106.7 (Trotter product formula)**.** *Let T and V be self-adjoint operators on a Hilbert space such that $T+V$ is essentially self-adjoint on the intersection of their domains. Then, in the strong operator topology,*

$$\mathrm{e}^{-\mathrm{i}(T+V)t/\hbar}=\lim_{N\to\infty}\left(\mathrm{e}^{-\mathrm{i}Tt/N\hbar}\,\mathrm{e}^{-\mathrm{i}Vt/N\hbar}\right)^N . \tag{106.14}$$

Rests on Equation (106.1).

Derivation, bounded case. *Derives Theorem 106.7.*

Let T and V be bounded operators, write $\Delta t=t/N$, $X_N=\mathrm{e}^{-\mathrm{i}(T+V)\Delta t/\hbar}$ and $Y_N=\mathrm{e}^{-\mathrm{i}T\Delta t/\hbar}\mathrm{e}^{-\mathrm{i}V\Delta t/\hbar}$. Both are contractions, $\|X_N\|\le 1$ and $\|Y_N\|\le 1$, and expanding both exponentials gives $\|X_N-Y_N\|\le K(\Delta t)^2$ with $K=\|[T,V]\|/2\hbar^2$, since the terms of order Δt^0 and Δt^1 agree. The telescoping identity

$$X_N^N-Y_N^N=\sum_{k=0}^{N-1}X_N^{N-1-k}(X_N-Y_N)Y_N^k$$

then bounds $\left\|X_N^N-Y_N^N\right\|\le N\,\|X_N-Y_N\|\le Kt^2/N\to 0$. Unbounded T and V — the physical case, since p^2 is unbounded — require the strong-topology argument of Trotter [1175], whose extension to operators bounded relative to T is Kato's [1335]; the mechanism is the same telescoping, applied to the resolvents. □

The hypothesis is met for $T=p^2/2m$ and any V that is Kato-small with respect to T, which includes every potential of physical interest here, the Coulomb potential among them [1335]. This matters more than it looks: without Theorem 106.7 the sliced expression Equation (106.9) would be a mnemonic, and its limit would have no reason to be the propagator of the theory one started from.

What $\mathcal{D}q$ is not

Remark 106.8 (There is no Feynman measure). The symbol $\mathcal{D}q$ in Equation (106.10) is not a measure. There is no countably additive complex measure of finite total variation on the space of continuous paths that reproduces Equation (106.9), and the obstruction is not technical: a measure with the required translation and scaling behaviour would have to assign every cylinder set a weight of unit modulus, and the total variation of such an object over the N-fold product diverges like $|2\pi\mathrm{i}\hbar\Delta t/m|^{-N/2}$. The precise statement and its consequences are Nelson's [1292]. The Euclidean continuation of Section 106.2 repairs exactly this: the weight there is real and positive, the total variation is finite, and the measure exists — it is Wiener's [507].

This is not a licence to be careless. It means that every manipulation of $\mathcal{D}q$ in this chapter is shorthand for a manipulation of the finite-dimensional integral in Equation (106.9), to be justified there. The two places where the difference bites are the ordering ambiguity below and the Jacobian of a change of variables, which is the subject of Section 106.7.3.

Ordering, and the midpoint rule Lemma 106.1 placed V at the midpoint and remarked that for a potential the choice is immaterial at the order retained. It is not immaterial when the Hamiltonian depends on momentum beyond p^2, and the case that matters physically is the magnetic one.

Proposition 106.9 (The midpoint rule is forced by gauge invariance)**.** *For* $H = (\boldsymbol{p} - Q\boldsymbol{A}(\boldsymbol{x}))^2/2m$ *the short-time kernel is*

$$\langle \boldsymbol{x}'' | \, \mathrm{e}^{-\mathrm{i}H\Delta t/\hbar} \, | \boldsymbol{x}' \rangle = \left(\frac{m}{2\pi\mathrm{i}\hbar\Delta t} \right)^{3/2} \exp\left\{ \frac{\mathrm{i}}{\hbar} \left[\frac{m(\boldsymbol{x}'' - \boldsymbol{x}')^2}{2\Delta t} + Q\,(\boldsymbol{x}'' - \boldsymbol{x}') \cdot \boldsymbol{A}\!\left(\frac{\boldsymbol{x}'' + \boldsymbol{x}'}{2} \right) \right] \right\}, \tag{106.15}$$

to the accuracy of Equation (106.5)*. Evaluating* $\boldsymbol{A}$ *at either endpoint instead gives an answer that differs at order* Δt *and does not reproduce the Schrödinger equation.*

Rests on Equations (106.5) *and* (106.6)*.*

Derivation. *Derives Proposition 106.9.*

Repeat the momentum-space computation Equation (106.6) with H quadratic in $\boldsymbol{p} - Q\boldsymbol{A}$. Writing $\boldsymbol{A}$ at a fixed point $\boldsymbol{x}_*$ inside the segment, the $\boldsymbol{p}$ integral is Gaussian with a shifted centre and gives Equation (106.15) with $\boldsymbol{A}(\boldsymbol{x}_*)$ in place of the midpoint value. To fix $\boldsymbol{x}_*$, demand that the kernel transform correctly under a gauge transformation $\boldsymbol{A} \to \boldsymbol{A} + \nabla\Lambda$, under which the wavefunction acquires the phase $\mathrm{e}^{\mathrm{i}Q\Lambda(\boldsymbol{x})/\hbar}$ and hence the kernel must acquire $\mathrm{e}^{\mathrm{i}Q[\Lambda(\boldsymbol{x}'') - \Lambda(\boldsymbol{x}')]/\hbar}$. The exponent in Equation (106.15) changes by $(\mathrm{i}Q/\hbar)(\boldsymbol{x}'' - \boldsymbol{x}') \cdot \nabla\Lambda(\boldsymbol{x}_*)$, and

$$\Lambda(\boldsymbol{x}'') - \Lambda(\boldsymbol{x}') = (\boldsymbol{x}'' - \boldsymbol{x}') \cdot \nabla\Lambda(\boldsymbol{x}_*) + O\big(|\boldsymbol{x}'' - \boldsymbol{x}'|^3\big)$$

holds *only* for the midpoint $\boldsymbol{x}_* = (\boldsymbol{x}'' + \boldsymbol{x}')/2$; at any other point the error is quadratic, i.e. of order $\hbar\Delta t/m$ in the exponent, which survives the limit. Endpoint prescriptions therefore quantize a different Hamiltonian, differing from H by a term proportional to $\nabla \cdot \boldsymbol{A}$. □

The moral generalizes: the time-sliced integral is a definition, and different discretizations of the same classical Lagrangian can define different quantum theories. What selects the right one is a symmetry that must survive — gauge invariance here, and in Section 106.7.3 chiral symmetry, where the answer turns out to be that no discretization preserves it and the failure is observable.

The classical limit

Theorem 106.10 (Stationary phase)**.** *Let $f:\mathbb{R}^n\to\mathbb{R}$ be smooth with a single non-degenerate critical point x_0, $\nabla f(x_0)=0$ and $\det f''(x_0)\neq 0$, and let g be smooth and compactly supported near x_0. Then as $\hbar\to 0^+$*

$$\int \mathrm{d}^n x\; g(x)\,\mathrm{e}^{\mathrm{i}f(x)/\hbar} = \frac{(2\pi\hbar)^{n/2}\,g(x_0)}{|\det f''(x_0)|^{1/2}}\,\mathrm{e}^{\mathrm{i}f(x_0)/\hbar}\,\mathrm{e}^{\mathrm{i}\pi\sigma/4}\big[1+O(\hbar)\big]\,, \tag{106.16}$$

with σ the signature of $f''(x_0)$.

Rests on Equation (106.8)*.*

Derivation. *Derives Theorem 106.10.*

Away from x_0 the integrand oscillates: writing $\mathrm{e}^{\mathrm{i}f/\hbar}=(\hbar/\mathrm{i})\nabla f\cdot\nabla \mathrm{e}^{\mathrm{i}f/\hbar}/|\nabla f|^2$ and integrating by parts k times produces a factor $\hbar^k$, so those regions contribute beyond all orders. Near x_0, substitute $x=x_0+\sqrt{\hbar}\,u$ and expand $f(x)=f(x_0)+\frac{\hbar}{2}u^{\mathsf{T}}f''(x_0)u+O(\hbar^{3/2})$. The measure supplies $\hbar^{n/2}$, the exponent becomes $\mathrm{i}f(x_0)/\hbar+\frac{\mathrm{i}}{2}u^{\mathsf{T}}f''u$, and Equation (106.8) evaluates the remaining integral. □

Apply this to Equation (106.9), whose integration variables are the $N-1$ intermediate positions and whose phase is the discretized action. The critical point of S is the path satisfying $\partial S/\partial q_k=0$ for every k, which is the discretized Euler–Lagrange equation, i.e. Hamilton's principle (Chapter 16); in the limit it is the classical trajectory. Therefore

$$K(q'',t'';q',t') \underset{\hbar\to 0}{\longrightarrow} \sqrt{\frac{1}{2\pi\mathrm{i}\hbar}\left(-\frac{\partial^2 S_{\mathrm{cl}}}{\partial q''\partial q'}\right)}\;\exp\!\left[\frac{\mathrm{i}}{\hbar}S_{\mathrm{cl}}(q'',t'';q',t')-\frac{\mathrm{i}\pi\nu}{2}\right], \tag{106.17}$$

the Van Vleck form, with ν the number of conjugate points passed — the Maslov index, the accumulated signature phase of Equation (106.8). Dirac's correspondence Equation (106.4) is recovered as the leading term, with the normalization now derived rather than guessed.

Proposition 106.11 (The semiclassical propagator is the WKB wavefunction)**.** *For a particle of energy E in one dimension, the Fourier transform of Equation (106.17) with respect to $t''-t'$ has stationary phase at the classical time of flight and gives, up to normalization,*

$$\psi(q)\propto\frac{1}{\sqrt{p(q)}}\,\exp\!\left[\pm\frac{\mathrm{i}}{\hbar}\int^{q}p(u)\,\mathrm{d}u\right], \qquad p(q)=\sqrt{2m\big(E-V(q)\big)}\,, \tag{106.18}$$

which is the semiclassical wave Equation (9.69) of Corollary 9.62, where the local momentum is written P and the amplitude factor $P^{-1/2}$ is derived from the transport equation rather than from a stationary phase.

Rests on Equations (106.3) *and* (106.17) *and Theorem 106.10.*

Derivation. *Derives Proposition 106.11.*

The energy eigenfunctions are obtained from K by $\int \mathrm{d}t\,\mathrm{e}^{\mathrm{i}Et/\hbar}K(q,t;q',0)$, a Fourier transform of the type of Chapter 17. Applying Theorem 106.10 in the single variable t, the stationary point is where $\partial S_{\mathrm{cl}}/\partial t=-E$ by Equation (106.3), i.e. at the classical time of flight at energy E, and the phase there becomes the Legendre transform $W(q,q'';E)=S_{\mathrm{cl}}+Et=\int p\,\mathrm{d}q$, Hamilton's characteristic function. The prefactor collects $(\partial^2 W/\partial q\partial q'')^{1/2}$, which for one degree of freedom is $(\partial p/\partial q'')^{1/2}\propto p(q)^{-1/2}$ after the reciprocity relation. This is Equation (106.18). □

Proposition 106.11 is the exact junction between this chapter and Section 9.6: the WKB ansatz $\psi=\exp(\mathrm{i}S/\hbar)$ of Equation (9.62) and the path-integral weight $\exp(\mathrm{i}S/\hbar)$ of Equation (106.10) are the same object seen from two sides, and the expansion of the exponent in powers of $\hbar$ in Equation (9.64) is the loop expansion of Section 106.3.4. In particular the $\hbar^1$ term of the WKB series, which produces the $p^{-1/2}$ amplitude, is the one-loop determinant, and the failure of the expansion at a turning point (Remark 9.66) is the failure of Theorem 106.10 at a degenerate critical point.

106.2 Euclidean formulation

106.2.1 Wick rotation

Remark 106.8 recorded that the oscillatory weight $\mathrm{e}^{\mathrm{i}S/\hbar}$ defines no measure. The remedy is to continue the time variable into the complex plane until the weight becomes real and decaying. The substitution is Wick's [1053].

Definition 106.12 (Euclidean time)**.** Set $t = -\mathrm{i}\tau$ with τ real, of SI dimension s. Equivalently $x^0 = ct \mapsto -\mathrm{i}x^4$ with $x^4 = c\tau$, and the metric $\eta_{\mu\nu} = \mathrm{diag}(+1,-1,-1,-1)$ becomes $-\delta_{\mu\nu}$: squared four-vectors change sign, $p^2 \mapsto -p_E^2$ with $p_E^2 = (p^4)^2 + \boldsymbol{p}^2 \geq 0$.

Proposition 106.13 (The Euclidean weight)**.** *Under Definition 106.12 the action of a particle in a potential becomes $S = \mathrm{i}S_E$ with*

$$S_E[q] = \int \mathrm{d}\tau \left[\frac{m}{2} \left(\frac{\mathrm{d}q}{\mathrm{d}\tau} \right)^2 + V(q) \right] , \tag{106.19}$$

of dimension J s*, so the path-integral weight is*

$$\exp\!\left(\frac{\mathrm{i}}{\hbar} S \right) \longmapsto \exp\!\left(-\frac{1}{\hbar} S_E \right) , \tag{106.20}$$

real, positive and bounded by one, since $S_E \geq 0$ for $V \geq 0$. The Euclidean action is the sum, not the difference, of kinetic and potential terms: the Euclidean motion takes place in the inverted potential $-V$.

Rests on Definition 106.12.

Derivation. *Derives Proposition 106.13.*

With $t = -\mathrm{i}\tau$, $\mathrm{d}t = -\mathrm{i}\,\mathrm{d}\tau$ and $\mathrm{d}q/\mathrm{d}t = \mathrm{i}\,\mathrm{d}q/\mathrm{d}\tau$, so $\frac{m}{2}(\mathrm{d}q/\mathrm{d}t)^2 = -\frac{m}{2}(\mathrm{d}q/\mathrm{d}\tau)^2$ and

$$S = \int \mathrm{d}t \left[\frac{m}{2} \left(\frac{\mathrm{d}q}{\mathrm{d}t} \right)^2 - V \right] = \int (-\mathrm{i}\,\mathrm{d}\tau) \left[-\frac{m}{2} \left(\frac{\mathrm{d}q}{\mathrm{d}\tau} \right)^2 - V \right] = \mathrm{i} \int \mathrm{d}\tau \left[\frac{m}{2} \left(\frac{\mathrm{d}q}{\mathrm{d}\tau} \right)^2 + V \right] ,$$

whence $\mathrm{i}S/\hbar = \mathrm{i} \cdot \mathrm{i}S_E/\hbar = -S_E/\hbar$. □

The Euclidean weight is a genuine probability density on path space once normalized — that is the subject of Section 106.2.2 — and it is what makes the numerical work of Section 106.5.2 possible at all. What must be established is that the continuation is legitimate: that the Minkowski amplitude may be recovered from the Euclidean one by rotating back.

The pole structure that licenses the rotation

Proposition 106.14 (The Feynman contour permits the rotation)**.** *Let $f(p^0)$ be a product of Feynman propagators Equation (100.23) and scalar propagators $\mathrm{i}\hbar^3 c/(p^2 - m^2c^2 + \mathrm{i}\epsilon)$ and let the integrand fall faster than $|p^0|^{-1}$ at infinity. Then*

$$\int_{-\infty}^{\infty} \mathrm{d}p^0 \, f(p^0) = \mathrm{i} \int_{-\infty}^{\infty} \mathrm{d}p^4 \, f(\mathrm{i}p^4) , \tag{106.21}$$

the contour being rotated counterclockwise from the real to the imaginary axis.

Rests on Equation (100.23), Theorem 8.12, and Lemma 8.9.

Derivation. *Derives Proposition 106.14.*

A propagator denominator $p^2 - m^2c^2 + \mathrm{i}\epsilon = (p^0)^2 - \boldsymbol{p}^2 - m^2c^2 + \mathrm{i}\epsilon$ vanishes at

$$p^0 = \pm\left(\sqrt{\boldsymbol{p}^2 + m^2c^2} - \mathrm{i}\epsilon'\right), \qquad \epsilon' > 0\,,$$

to first order in ϵ: the positive-energy pole sits just *below* the positive real axis and the negative-energy pole just *above* the negative real axis, i.e. in the fourth and second quadrants respectively. The first and third quadrants are therefore free of singularities, and the closed contour consisting of the real axis, the imaginary axis traversed downwards, and the two connecting arcs in those quadrants encloses no pole. By Cauchy's theorem (Theorem 8.12) its integral vanishes; the arcs contribute nothing by the estimation lemma Lemma 8.9 given the assumed fall-off. Rotating $p^0 = \mathrm{e}^{\mathrm{i}\theta}\,|p^0|$ from $\theta = 0$ to $\theta = \pi/2$ and writing $p^0 = \mathrm{i}p^4$ gives $\mathrm{d}p^0 = \mathrm{i}\,\mathrm{d}p^4$, which is Equation (106.21). Under it $p^2 = (p^0)^2 - \boldsymbol{p}^2 \mapsto -(p^4)^2 - \boldsymbol{p}^2 = -p_E^2$, so the denominator becomes $-(p_E^2 + m^2c^2)$, which never vanishes for real p_E: the Euclidean propagator has no poles at all, and no $\mathrm{i}\epsilon$ is needed. □

The $\mathrm{i}\epsilon$ prescription of Equation (100.23) is thus not a computational convenience but the statement of which boundary value of an analytic function the theory picks — and it is exactly the choice that puts the poles out of the way of the rotation. That is the same fact seen twice: the Feynman propagator is the one whose Euclidean continuation exists.

Remark 106.15 (What the rotation costs). Three things are lost or endangered.

Real-time information. The Euclidean correlator determines the Minkowski one only through an analytic continuation, and continuing numerical data from a finite set of Euclidean times back to real time is an ill-posed inverse problem: two spectral functions differing by an oscillation of high frequency give Euclidean correlators differing by less than the statistical error. This is why the lattice computes masses and matrix elements well and transport coefficients badly.

Positivity of the weight, at finite chemical potential. Introducing a chemical potential μ for a conserved charge adds μ to the time component of the covariant derivative in the fermion operator; after the rotation this term is imaginary relative to the rest, the determinant $\det(\mathrm{i}\hbar\gamma^\mu D_\mu - mc)$ of Section 106.3.3 becomes complex, and $\mathrm{e}^{-S_E/\hbar}$ is no longer a probability. This is the *sign problem.* It is not a difficulty of technique: no reweighting scheme has been found whose cost is not exponential in the volume, and the phase diagram of QCD at large baryon density is consequently not known from first principles. The honest position is that Section 106.5.2's successes are successes at zero or small chemical potential.

Nothing else. At zero chemical potential the rotation is exact, and Section 106.2.3 states the conditions under which the Euclidean theory contains the whole of the Minkowski one.

106.2.2 The Feynman–Kac formula

Theorem 106.16 (Feynman–Kac)**.** *Let $H = p^2/2m + V(q)$ with V bounded below and continuous. Then for $\tau > 0$*

$$\langle q''|\,\mathrm{e}^{-H\tau/\hbar}\,|q'\rangle = \int_{q(0)=q'}^{q(\tau)=q''} \mathcal{D}q\;\mathrm{e}^{-S_E[q]/\hbar} = \rho_\tau(q''-q')\;\mathbb{E}_{q'\to q''}\left[\exp\left(-\frac{1}{\hbar}\int_0^\tau V(q(s))\,\mathrm{d}s\right)\right], \tag{106.22}$$

where the expectation is over the Wiener measure of Brownian bridges from q' to q'' with diffusion constant

$$D = \frac{\hbar}{2m}\,, \qquad [D] = \mathrm{m}^2/\mathrm{s}\,, \tag{106.23}$$

and $\rho_\tau(x) = (4\pi D\tau)^{-1/2}\mathrm{e}^{-x^2/4D\tau}$ is the free heat kernel.

Rests on Theorem 106.3 and Proposition 106.13.

Derivation. *Derives Theorem 106.16.*

Repeat the slicing of Theorem 106.3 with $\mathrm{e}^{-H\tau/\hbar}$ in place of $\mathrm{e}^{-\mathrm{i}Ht/\hbar}$; every step goes through with $\mathrm{i}\varepsilon \mapsto \Delta\tau$, giving

$$\langle q''|\,\mathrm{e}^{-H\tau/\hbar}\,|q'\rangle = \lim_{N\to\infty}\left(\frac{m}{2\pi\hbar\Delta\tau}\right)^{N/2}\int\prod_{k=1}^{N-1}\mathrm{d}q_k\;\exp\!\left\{-\frac{1}{\hbar}\sum_{k=1}^{N}\Delta\tau\left[\frac{m}{2}\left(\frac{q_k-q_{k-1}}{\Delta\tau}\right)^2+V\right]\right\}.$$

Now the prefactor and the kinetic term together are $\prod_k \rho_{\Delta\tau}(q_k-q_{k-1})$ with ρ as stated, because $m/(2\hbar\Delta\tau) = 1/(4D\Delta\tau)$ with D of Equation (106.23). That product is precisely the finite-dimensional distribution of Brownian motion of diffusion constant D sampled at the times $k\Delta\tau$, and by Kolmogorov's extension theorem it defines a genuine countably additive probability measure on continuous paths — Wiener's measure [507]. What remains in the exponent is $-\frac{1}{\hbar}\sum\Delta\tau\,V$, the Riemann sum for $\frac{1}{\hbar}\int V$, and the identification of the limit with the expectation is Kac's theorem [935]. Conditioning on the endpoint $q(\tau)=q''$ produces the bridge and the explicit factor $\rho_\tau(q''-q')$. □

That $D=\hbar/2m$ is worth pausing over. The Euclidean quantum mechanics of a particle of mass m *is* a diffusion process whose diffusion constant is set by $\hbar$; the classical limit $\hbar\to 0$ is the limit of vanishing diffusion, in which the paths collapse onto the single deterministic solution. For an electron, $D=\hbar/2m_e = 5.79\times10^{-5}\,\mathrm{m^2/s}$. The roughness noted after Theorem 106.3 is now a theorem rather than an estimate: Wiener paths are almost surely continuous, almost surely nowhere differentiable, and almost surely Hölder continuous of every exponent below $1/2$.

Spectral consequences

Proposition 106.17 (Euclidean time projects onto the ground state)**.** *Let H have discrete spectrum $E_0<E_1\le\cdots$ with eigenfunctions ψ_n. Then*

$$\langle q''|\,\mathrm{e}^{-H\tau/\hbar}\,|q'\rangle = \sum_n \psi_n(q'')\psi_n^*(q')\,\mathrm{e}^{-E_n\tau/\hbar} \;\xrightarrow[\tau\to\infty]{}\; \psi_0(q'')\psi_0^*(q')\,\mathrm{e}^{-E_0\tau/\hbar}, \tag{106.24}$$

so the large-τ decay rate of a Euclidean correlator measures an energy, and the gap E_1-E_0 governs the leading correction.

Rests on Theorem 106.16.

Derivation. *Derives Proposition 106.17.*

Insert the resolution of the identity in energy eigenstates into $\mathrm{e}^{-H\tau/\hbar}$; each term carries $\mathrm{e}^{-E_n\tau/\hbar}$ and the smallest exponent dominates. The subleading term is suppressed by $\mathrm{e}^{-(E_1-E_0)\tau/\hbar}$. □

Equation (106.24) is the entire basis of hadron spectroscopy on the lattice. A field theory correlator of an operator carrying the quantum numbers of a particle of mass m decays in Euclidean time as $\mathrm{e}^{-mc^2\tau/\hbar}$, and fitting that exponential to Monte Carlo data is how the numbers of Section 106.5.2 are obtained. The correlation *length* in the fourth direction is

$$\xi = c\tau_{\mathrm{corr}} = \frac{\hbar}{mc}, \tag{106.25}$$

the Compton wavelength: a mass gap in the quantum theory is a finite correlation length in the statistical one, and a massless particle is a critical point.

Finite temperature

Theorem 106.18 (The partition function is a periodic path integral)**.** *The canonical partition function of a system with Hamiltonian H at temperature T is*

$$Z(T) = \operatorname{tr} \mathrm{e}^{-H/k_B T} = \oint_{q(0)=q(\beta\hbar)} \mathcal{D}q \, \mathrm{e}^{-S_E[q]/\hbar}, \qquad \beta = \frac{1}{k_B T}, \tag{106.26}$$

the integral running over paths periodic in Euclidean time with period

$$\tau_\beta = \beta\hbar = \frac{\hbar}{k_B T}, \qquad [\tau_\beta] = \mathrm{s}. \tag{106.27}$$

Rests on Equation (106.22).

Derivation. *Derives Theorem 106.18.*

Take Equation (106.22) at $\tau = \beta\hbar$, set $q'' = q' = q$ and integrate over q: the left-hand side becomes $\int \mathrm{d}q \, \langle q| \, \mathrm{e}^{-\beta H} \, |q\rangle = \operatorname{tr} \mathrm{e}^{-\beta H}$, and the right-hand side becomes the path integral restricted to closed paths. Fermionic degrees of freedom acquire antiperiodic boundary conditions instead, because the trace of a Grassmann bilinear picks up a sign; that computation is Section 106.3.3. □

The thermal circumference Equation (106.27) is a length in Euclidean spacetime, $c\tau_\beta = \hbar c/k_B T$. At room temperature, $T = 300\,\mathrm{K}$,

$$\tau_\beta = \frac{\hbar}{k_B T} = \frac{1.054\,571\,817 \times 10^{-34}\,\mathrm{J\,s}}{1.380\,649 \times 10^{-23}\,\mathrm{J/K} \times 300\,\mathrm{K}} = 2.55 \times 10^{-14}\,\mathrm{s}, \tag{106.28}$$

using the exactly defined Boltzmann constant [2431]; at the temperature $k_B T = 150\,\mathrm{MeV}$ characteristic of the quark–hadron transition, $\tau_\beta = 4.39 \times 10^{-24}\,\mathrm{s}$ and $c\tau_\beta = 1.32\,\mathrm{fm}$, comparable with the size of a hadron — which is why finite-temperature QCD is a strongly coupled problem and not a small correction.

Table 106.1: The Euclidean dictionary. A quantum field theory in $3+1$ spacetime dimensions, continued to imaginary time, is a classical statistical system in four Euclidean dimensions; a quantum system at temperature T is the same system on a slab of thickness $\hbar/k_B T$. Every entry is an identity between the two columns after Theorems 106.16 and 106.18, not an analogy.

Euclidean field theory	**Classical statistical mechanics**
Weight $\mathrm{e}^{-S_E/\hbar}$	Boltzmann factor $\mathrm{e}^{-\mathcal{H}/k_B T}$
$\hbar$	$k_B T$
Euclidean action S_E	energy $\mathcal{H}$ of a configuration
Generating functional $Z[0]$	partition function
Mass m of the lightest state	inverse correlation length, $\xi = \hbar/mc$
Massless particle	critical point, $\xi \to \infty$
Continuum limit $a \to 0$	second-order critical point
Vacuum expectation value	thermal average
Temperature T of the quantum system	circumference $\hbar/k_B T$ of the fourth direction

Table 106.1 is the reason this chapter sits next to Chapter 118 in the architecture of the book, and why the renormalization group of Chapter 108 is the same construction as the one applied to critical phenomena in Chapter 123. The identification $\hbar \leftrightarrow k_B T$ fixes the direction of the loop expansion of Section 106.3.4: it is an expansion in *powers* of $\hbar$,

hence in powers of $k_B T$, and so a *low*-temperature expansion of the statistical system — the saddle-point expansion about the configuration of least energy, with each loop one power of $k_B T$ smaller than the last. The classical limit $\hbar \to 0$ is correspondingly its zero-temperature limit, in which fluctuations are frozen and the configuration of least energy is the only one. A genuine high-temperature expansion of the statistical system is an expansion in $1/k_B T$ and is the strong-coupling expansion of Theorem 106.72, at the opposite end of the same axis.

106.2.3 Reconstruction of the real-time theory

The Euclidean theory is a computational device only if the Minkowski theory can be recovered from it. Osterwalder and Schrader gave necessary and sufficient conditions [1510, 1555].

Definition 106.19 (Schwinger functions)**.** The Euclidean correlation functions of a scalar field,

$$S_n(x_1, \dots, x_n) = \frac{1}{Z} \int \mathcal{D}\phi \; \phi(x_1) \cdots \phi(x_n) \, \mathrm{e}^{-S_E[\phi]/\hbar} \,, \tag{106.29}$$

with all x_i in Euclidean $\mathbb{R}^4$ and distinct.

Theorem 106.20 (Osterwalder–Schrader reconstruction)**.** *Let a family $\{S_n\}$ satisfy*

1. regularity*: each S_n is a tempered distribution, with growth in n bounded so that the reconstructed fields are operator-valued distributions;*
2. Euclidean invariance*: S_n is invariant under the rotations and translations of* $\mathrm{SO}(4) \ltimes \mathbb{R}^4$*;*
3. reflection positivity*: for the reflection $\theta : (x^4, \boldsymbol{x}) \mapsto (-x^4, \boldsymbol{x})$ and any finite family of test functions f_k supported in $x^4 > 0$,*
$$\sum_{j,k} \bar{c}_j c_k \, S(\theta \bar{f}_j \otimes f_k) \geq 0 \,, \tag{106.30}$$
4. symmetry*: S_n is invariant under permutations of its arguments;*
5. cluster decomposition*: $S_{n+m}(x_1, \dots, x_n, y_1 + a, \dots) \to S_n S_m$ as the Euclidean separation $|a| \to \infty$;*

then there is a Wightman quantum field theory in $3+1$ Minkowski dimensions — a Hilbert space with a positive-definite inner product, a unitary representation of the Poincaré group with spectrum in the forward cone, a unique vacuum, and local fields — whose vacuum correlators continue to the S_n. Conversely, the Schwinger functions of any Wightman theory satisfy (i)–(v). The Wightman axioms so reconstructed are the subject of Chapter 109.

Rests on Definition 106.19.

Of the five, reflection positivity is the one that carries the physics, and it is worth seeing why in the simplest case, where the whole content is visible.

Proposition 106.21 (Reflection positivity is the positivity of the Hilbert space)**.** *In Euclidean quantum mechanics, let F be any functional of the path depending only on $q(s)$ for $s > 0$, and let θF be its reflection, depending only on $s < 0$. Then*

$$\left\langle \overline{\theta F} \, F \right\rangle = \| \Psi_F \|^2 \geq 0 \,, \tag{106.31}$$

where Ψ_F is a vector in the Hilbert space of the theory built from F and the ground state $|\Omega\rangle$ by Theorem 106.16.

Rests on Theorem 106.16 and Equation (106.30).

Derivation. *Derives Proposition 106.21.*

Take $F = \prod_i f_i(q(s_i))$ with $0 < s_1 < \cdots < s_n$. By Theorem 106.16 the Euclidean expectation of a product of such factors is a time-ordered chain of operators separated by Euclidean evolutions,

$$\left\langle \overline{\theta F}\, F \right\rangle = \langle \Omega |\, \hat{f}_n^\dagger \mathrm{e}^{-H(s_n - s_{n-1})/\hbar} \cdots \hat{f}_1^\dagger\, \hat{f}_1 \cdots \mathrm{e}^{-H(s_n - s_{n-1})/\hbar} \hat{f}_n\, |\Omega\rangle \ ,$$

because reflecting $s \mapsto -s$ reverses the order of the factors and conjugates them, while the intervening evolution operators $\mathrm{e}^{-H\Delta s/\hbar}$ are positive and self-adjoint. The whole expression is therefore $\|\Psi\|^2$ with $\Psi = \hat{f}_1 \mathrm{e}^{-H(s_2 - s_1)/\hbar} \cdots |\Omega\rangle$, manifestly non-negative. Linear combinations give Equation (106.30).

Read backwards, this is the content of the axiom: the Euclidean functional integral supplies a bilinear form on functionals of the positive-time half-path, and reflection positivity says exactly that this form is a non-negative inner product. The physical Hilbert space is the completion of that space modulo null vectors, the operator $\mathrm{e}^{-H\tau/\hbar}$ is the translation in Euclidean time, and H is self-adjoint and bounded below because it generates a contraction semigroup. Without Equation (106.30) one obtains a vector space with an indefinite form and no probability interpretation — and no amount of Euclidean invariance repairs it. □

Remark 106.22 (Reflection positivity on the lattice). The lattice discretization of Section 106.5.1 must inherit Equation (106.30) or it is not the Euclidean form of any quantum theory, however well it reproduces the classical action. Wilson's plaquette action does inherit it, because it couples only nearest-neighbour time slices, so the transfer operator constructed above exists and is positive. Actions "improved" by adding couplings between next-to-nearest time slices generally do not: they can produce a transfer matrix with negative eigenvalues, whose signature is a correlator that oscillates in Euclidean time and which cannot be fitted by Equation (106.24) with any real spectrum. Improvement schemes are therefore constructed to act within a time slice, or their violation is arranged to vanish faster than the discretization errors they remove. This is a working constraint on the design of a lattice action, not a foundational nicety.

Derivation pending #163

Osterwalder–Schrader reconstruction: the derivation above establishes the easy half in quantum mechanics — that reflection positivity is the positivity of the reconstructed inner product — and states the field-theoretic theorem. The construction of the Wightman functions from the Schwinger functions by analytic continuation in the time arguments, and the verification of the spectrum condition and locality, are not carried out here; they belong in Appendix A, alongside the axiomatic material of the axiomatic quantum field theory chapter of this part.

106.3 Functional quantization of fields

106.3.1 Generating functionals

A field is a mechanical system with one degree of freedom per point of space, so Theorem 106.3 carries over with $q(t)$ replaced by $\phi(x)$ and the time slicing replaced by a slicing in x^0. The one thing that must be fixed first is dimensions, because the whole chapter turns on the exponent $S/\hbar$ being a pure number.

Notation 106.23 (The scalar field and its source). For a single real scalar field the Lagrangian density is

$$\mathcal{L} = \frac{1}{2}\partial_\mu\phi\, \partial^\mu\phi - \frac{1}{2}\left(\frac{mc}{\hbar}\right)^2 \phi^2 - \frac{g}{4!}\phi^4 \, , \tag{106.32}$$

of dimension $\mathrm{J/m^3}$ by Equation (100.2). Since ∂_μ carries $1/\mathrm{m}$, the field carries

$$[\phi] = \mathrm{J}^{1/2}/\mathrm{m}^{1/2} = \mathrm{kg}^{1/2}\,\mathrm{m}^{1/2}/\mathrm{s}\,, \tag{106.33}$$

the mass term carries $(mc/\hbar)^2$ of dimension $1/\mathrm{m}^2$ — the inverse squared Compton wavelength, so that the field equation is $(\Box + (mc/\hbar)^2)\phi = 0$ — and the quartic coupling carries

$$[g] = 1/(\mathrm{J\,m})\,, \qquad \hat{g} := g\hbar c \ \text{ dimensionless}\,. \tag{106.34}$$

The source J is coupled by adding $c^{-1}\int \mathrm{d}^4x\, J\phi$ to the action, so $[J] = \mathrm{J}^{1/2}/\mathrm{m}^{5/2}$.

That $\hat{g} = g\hbar c$ and not g is the dimensionless coupling is not bookkeeping. It says that the strength of a quartic self-interaction is measured against $\hbar c = 3.161\,53 \times 10^{-26}\,\mathrm{J\,m}$, the only quantity in the theory with the dimension of an energy times a length, exactly as the electromagnetic coupling is measured against it in Equation (100.3). A "dimensionless coupling" is always a ratio to $\hbar c$, and saying so is what makes the classical limit visible: as $\hbar \to 0$ at fixed g, every dimensionless coupling goes to zero and the theory becomes free and classical.

Definition 106.24 (The generating functional)**.**

$$Z[J] = \int \mathcal{D}\phi\, \exp\!\left\{\frac{\mathrm{i}}{\hbar}\left(S[\phi] + \frac{1}{c}\int \mathrm{d}^4x\; J(x)\phi(x)\right)\right\}, \tag{106.35}$$

with $S[\phi] = c^{-1}\int \mathrm{d}^4x\,\mathcal{L}$ and $\mathcal{D}\phi$ the slicing limit of Equation (106.9) applied to each mode. It is dimensionless, and $Z[0]$ is the vacuum persistence amplitude $\langle 0, +\infty | 0, -\infty\rangle$: $|Z[0]|^2$ is the probability that a system prepared in the vacuum is found in the vacuum, which is unity for a stable theory and less than unity when the external conditions can create particles.

Schwinger's insight, which is what makes Equation (106.35) the master object of the subject, is that all correlation functions are encoded in the response of the vacuum amplitude to an external source [989].

Theorem 106.25 (Correlators as functional derivatives)**.** *The time-ordered vacuum correlation functions are*

$$\langle 0|\,T\,\phi(x_1)\cdots\phi(x_n)\,|0\rangle = \frac{1}{Z[0]}\prod_{k=1}^{n}\left(\frac{\hbar c}{\mathrm{i}}\,\frac{\delta}{\delta J(x_k)}\right) Z[J]\Big|_{J=0}\,, \tag{106.36}$$

each functional derivative carrying the factor $\hbar c/\mathrm{i}$ so that the result has dimension $[\phi]^n$.
Rests on Equations (106.9) and (106.35).

Derivation. *Derives Theorem 106.25.*
Differentiating Equation (106.35),

$$\frac{\delta Z}{\delta J(x)} = \int \mathcal{D}\phi\, \frac{\mathrm{i}}{\hbar c}\,\phi(x)\,\mathrm{e}^{\mathrm{i}(S + c^{-1}\int J\phi)/\hbar}\,,$$

so multiplying by $\hbar c/\mathrm{i}$ inserts one factor of $\phi(x)$ into the integrand. Repeating gives n insertions; dividing by $Z[0]$ normalizes. That the insertions come out *time-ordered* is a property of the slicing: the fields in Equation (106.9) are c-numbers attached to time slices, and reassembling them into operators reinstates them in the order of their time arguments, whatever order they were written in. This is the technical advantage of the formalism — the object it computes naturally is the object that appears in the Dyson series Equation (100.25) — and also its limitation, since it does not directly give an unordered product. □

Proposition 106.26 (Connected correlators and the linked-cluster theorem)**.** *Define*

$$W[J] := -\mathrm{i}\hbar \ln Z[J]\,, \tag{106.37}$$

of dimension J s*. Then* W *generates the* connected *correlation functions: in particular*

$$\langle \phi(x) \rangle_J = c\,\frac{\delta W}{\delta J(x)}\,, \qquad \langle T\phi(x)\phi(y)\rangle_{\mathrm{conn}} = \frac{\hbar c^2}{\mathrm{i}}\,\frac{\delta^2 W}{\delta J(x)\,\delta J(y)}\,. \tag{106.38}$$

Moreover $Z[J] = \exp(\mathrm{i}W[J]/\hbar)$ *states that the sum of all diagrams is the exponential of the sum of the connected ones.*

Rests on Equations (106.35) *and* (106.36)*.*

Derivation. *Derives Proposition 106.26.*

From $Z = \mathrm{e}^{\mathrm{i}W/\hbar}$, $\delta Z/\delta J = (\mathrm{i}/\hbar)(\delta W/\delta J)Z$, and combining with Equation (106.36) for $n = 1$ gives $\langle\phi\rangle = (\hbar c/\mathrm{i})(\mathrm{i}/\hbar)\delta W/\delta J = c\,\delta W/\delta J$. Differentiating once more,

$$\frac{\delta^2 Z}{\delta J_1 \delta J_2} = \left[\frac{\mathrm{i}}{\hbar}\frac{\delta^2 W}{\delta J_1 \delta J_2} + \left(\frac{\mathrm{i}}{\hbar}\right)^2 \frac{\delta W}{\delta J_1}\frac{\delta W}{\delta J_2}\right] Z\,,$$

so that, multiplying by $(\hbar c/\mathrm{i})^2 = -\hbar^2 c^2$,

$$\langle T\phi_1\phi_2\rangle = -\mathrm{i}\hbar c^2 \frac{\delta^2 W}{\delta J_1 \delta J_2} + \langle\phi_1\rangle\,\langle\phi_2\rangle\ ,$$

and the first term is by definition the connected part.

For the general statement, argue combinatorially. Let the connected diagrams (with sources attached) be indexed by i with values C_i. A general diagram is a disjoint union of connected pieces; if it contains n_i copies of the connected diagram i, its symmetry factor includes $1/n_i!$ from the permutations of identical components, because the Wick contractions that produce it differ only by relabelling those components. Summing over all multiplicities independently,

$$Z = Z_{\mathrm{free}} \sum_{\{n_i\}} \prod_i \frac{C_i^{n_i}}{n_i!} = Z_{\mathrm{free}} \prod_i \mathrm{e}^{C_i} = Z_{\mathrm{free}}\,\exp\!\left(\sum_i C_i\right)\ ,$$

so $\ln Z$ is the sum of connected diagrams alone. That is the linked-cluster theorem, and it is why the vacuum diagrams — those with no external source — cancel between numerator and denominator in Equation (106.36): they exponentiate into an overall factor $Z[0]$. □

The linked-cluster theorem has a consequence worth stating separately, because it is the reason the formalism is usable at all. The connected correlators are extensive: $\ln Z$ is proportional to the spacetime volume, whereas Z itself is exponentially large in it. Any expansion organized around W rather than Z therefore has coefficients with a finite thermodynamic limit. In the Euclidean language of Table 106.1 this is the statement that the free energy is extensive while the partition function is not.

106.3.2 Free fields and the perturbative expansion

The Gaussian theory, solved exactly

Theorem 106.27 (The free generating functional and the Feynman propagator)**.** *For* $g = 0$ *in Equation* (106.32)*,*

$$Z_0[J] = Z_0[0]\,\exp\!\left\{-\frac{\mathrm{i}}{2\hbar c}\int \mathrm{d}^4x\,\mathrm{d}^4y\; J(x)\,G(x-y)\,J(y)\right\}\ , \tag{106.39}$$

where G is the Green function of the Klein–Gordon operator,

$$-\left(\Box_x + \left(\frac{mc}{\hbar}\right)^2\right) G(x-y) = \delta^4(x-y)\,, \qquad [G] = 1/\mathrm{m}^2\,. \tag{106.40}$$

The two-point function is

$$\Delta_F(x-y) := \langle 0|\, T\phi(x)\phi(y)\,|0\rangle = \mathrm{i}\hbar c\, G(x-y)\,, \qquad \widetilde{\Delta}_F(p) = \frac{\mathrm{i}\hbar^3 c}{p^2 - m^2c^2 + \mathrm{i}\epsilon}\,, \tag{106.41}$$

of dimension $\mathrm{J/m}$ *in position space and* $\mathrm{J\,m^3}$ *in momentum space.*
Rests on Equations (106.32) *and* (106.35) *and Lemma 106.2.*

Derivation. *Derives Theorem 106.27.*

Integrate the kinetic term by parts, discarding the surface term:

$$S_0[\phi] = \frac{1}{c}\int \mathrm{d}^4x\, \frac{1}{2}\left[\partial_\mu\phi\partial^\mu\phi - \left(\frac{mc}{\hbar}\right)^2\phi^2\right] = \frac{1}{2c}\int \mathrm{d}^4x\, \phi\,\mathcal{K}\phi\,, \qquad \mathcal{K} = -\left(\Box + \left(\frac{mc}{\hbar}\right)^2\right).$$

The exponent of Equation (106.35) is then $\frac{\mathrm{i}}{\hbar c}\left[\frac{1}{2}\int \phi\mathcal{K}\phi + \int J\phi\right]$, a quadratic form in ϕ, and Lemma 106.2 applies once the square is completed. Put $\phi = \phi_0 + \chi$ with $\phi_0(x) = -\int \mathrm{d}^4y\, G(x-y)J(y)$, so that $\mathcal{K}\phi_0 = -J$ by Equation (106.40). Then

$$\frac{1}{2}\phi\mathcal{K}\phi + J\phi = \frac{1}{2}\chi\mathcal{K}\chi + \chi(\mathcal{K}\phi_0 + J) + \frac{1}{2}\phi_0\mathcal{K}\phi_0 + J\phi_0 = \frac{1}{2}\chi\mathcal{K}\chi - \frac{1}{2}\iint J\,G\,J\,,$$

the cross term vanishing by construction. The shift $\phi \mapsto \chi$ is a translation of the integration variable and leaves $\mathcal{D}\phi$ invariant — this is a property of the sliced definition, each $\mathrm{d}\phi_k$ being a Lebesgue measure on $\mathbb{R}$ — so the χ integral is the J-independent constant $Z_0[0]$ and what remains is Equation (106.39).

By Equation (106.37), $W_0[J] = -\frac{1}{2c}\iint JGJ$, and Equation (106.38) gives $\langle T\phi_1\phi_2\rangle = (\hbar c^2/\mathrm{i})(-G/c) = \mathrm{i}\hbar c G$.

In momentum space, with the convention Equation (100.14), $\partial_\mu \mapsto -\mathrm{i}p_\mu/\hbar$ and $\Box \mapsto -p^2/\hbar^2$, so $\mathcal{K} \mapsto (p^2 - m^2c^2)/\hbar^2$ and $\widetilde{G}(p) = \hbar^2/(p^2 - m^2c^2)$. The pole on the real axis is resolved by the Feynman prescription $m^2c^2 \to m^2c^2 - \mathrm{i}\epsilon$, which is the choice singled out by Proposition 106.14. Multiplying by $\mathrm{i}\hbar c$ gives Equation (106.41). The result has exactly the structure of the photon propagator Equation (100.17), the factor $\hbar^3 c$ being common to both. □

Corollary 106.28 (Wick's theorem)**.** *In the free theory, correlators of an odd number of fields vanish, and*

$$\langle 0|\, T\phi(x_1)\cdots\phi(x_{2n})\,|0\rangle = \sum_{\text{pairings}}\ \prod_{\text{pairs }(i,j)} \Delta_F(x_i - x_j)\,, \tag{106.42}$$

the sum running over all $(2n-1)!!$ *ways of partitioning the arguments into pairs.*
Rests on Equations (106.36) *and* (106.39).

Derivation. *Derives Corollary 106.28.*

Apply Equation (106.36) to Equation (106.39). Each derivative $\hbar c\,\delta/(\mathrm{i}\delta J)$ acting on the exponential brings down one factor $-(\mathrm{i}/\hbar c)\int GJ$ times $\hbar c/\mathrm{i}$, i.e. $-\int GJ$, which is set to zero at $J = 0$ unless a later derivative acts on it. A non-vanishing term therefore requires the derivatives to be used in pairs: one to bring down the bilinear, one to remove the surviving J. The pair (i,j) contributes $(\hbar c/\mathrm{i})^2 \times (-\mathrm{i}/\hbar c)G = \mathrm{i}\hbar c G = \Delta_F$, and the number of ways of pairing $2n$ objects is $(2n-1)!!$. An odd number of derivatives always leaves one J unpaired. □

Corollary 106.28 is the same statement as the operator Wick theorem [974] invoked in Proposition 100.14. In the operator formalism it is a statement about normal-ordering products of creation and annihilation operators, and that chapter quotes it from Wick's paper rather than proving it; here it is a property of the Gaussian integral, derived above, and no operators appear. That is the second road promised in the introduction.

Perturbation theory and the Feynman rules

Proposition 106.29 (Feynman rules from the functional integral)**.** *Write $S = S_0 + S_{\mathrm{int}}$ with $S_{\mathrm{int}} = -\frac{1}{c}\int \mathrm{d}^4x\, \frac{g}{4!}\phi^4$. Then*

$$Z[J] = \exp\!\left\{ \frac{\mathrm{i}}{\hbar} S_{\mathrm{int}}\!\left[\frac{\hbar c}{\mathrm{i}} \frac{\delta}{\delta J} \right] \right\} Z_0[J]\,, \tag{106.43}$$

and expanding the outer exponential generates, order by order in $\hat{g} = g\hbar c$, the diagrammatic rules: a line for each propagator Equation (106.41), a four-point vertex carrying

$$-\frac{\mathrm{i}g}{\hbar c}\int \mathrm{d}^4x \quad \leftrightarrow \quad -\frac{\mathrm{i}g}{\hbar c}\,(2\pi\hbar)^4\,\delta^4\!\left(\textstyle\sum_k p_k\right) \quad \text{in momentum space,} \tag{106.44}$$

one integration $\int \mathrm{d}^4\ell/(2\pi\hbar)^4$ per loop, and the symmetry factor of the diagram.
Rests on Equation (106.35), Theorem 106.25, and Corollary 106.28.

Derivation. *Derives Proposition 106.29.*

Every occurrence of $\phi(x)$ in S_{int} may be replaced by $(\hbar c/\mathrm{i})\delta/\delta J(x)$ acting on the integrand of Equation (106.35), by the computation in Theorem 106.25; the resulting operator may then be pulled outside the ϕ integral, leaving $Z_0[J]$. That is Equation (106.43), which is exact. Expanding $\exp(\mathrm{i}S_{\mathrm{int}}/\hbar)$ in powers and applying Corollary 106.28 to each term produces sums of products of propagators with four lines meeting at each vertex position, each vertex carrying $(\mathrm{i}/\hbar)\times(-1/c)\times(g/4!)$ and an integration $\int \mathrm{d}^4x$; the 4! is cancelled by the number of ways of assigning the four lines to the four fields, except when a symmetry of the diagram leaves some of those assignments equivalent, which is the origin of the symmetry factor. Fourier transforming each propagator by Equation (100.14) turns the position integrals into the momentum-conserving delta functions of Equation (106.44), leaving one free integration per independent loop. □

Theorem 106.30 (The loop expansion is the expansion in $\hbar$)**.** *A connected diagram with V vertices, I internal lines and L independent loops carries the factor*

$$\hbar^{I-V} = \hbar^{L-1}\,, \qquad L = I - V + 1\,, \tag{106.45}$$

counting the vertices and the internal lines alone. A tree graph has $L = 0$ and therefore carries $\hbar^{-1}$; relative to it an L-loop graph is suppressed by $\hbar^{L}$. The two statements are the same one and must not be added: $\hbar^{L-1}$ is the absolute power, $\hbar^{L}$ the power relative to the classical contribution. The perturbative expansion in the number of loops is therefore an expansion in powers of $\hbar$, and in nothing else.
Rests on Proposition 106.29 and Equation (106.40).

Derivation. *Derives Theorem 106.30.*

By Proposition 106.29 each vertex carries an explicit $\hbar^{-1}$ (from $\mathrm{i}S_{\mathrm{int}}/\hbar$) and each internal line one factor of $\hbar$ (the propagator $\mathrm{i}\hbar cG$ is linear in $\hbar$, G being $\hbar$-independent by Equation (106.40) once $mc/\hbar$ is regarded as the fixed inverse Compton wavelength). A diagram therefore carries $\hbar^{I-V}$. The second relation of Equation (106.45) is Euler's relation for a connected graph: each internal line carries one momentum, the V vertices impose V delta functions of which one is the overall momentum conservation, leaving $I-(V-1)$ undetermined loop momenta. Hence $\hbar^{I-V} = \hbar^{L-1}$, and a tree, having $L = 0$, sets the reference power $\hbar^{-1}$ against which the loops are counted. □

Theorem 106.30 is the reason this chapter refuses to drop $\hbar$ from its formulae. In the standard presentation the loop counting is a combinatorial fact about graphs; here it is a physical statement — a one-loop correction is the first quantum correction, suppressed relative to the classical answer by one power of the quantity whose smallness defines the classical limit, and the tree diagrams alone reproduce classical field theory exactly. This will be made precise in Section 106.3.4, where the tree approximation to Γ is shown to be the classical action.

The series does not converge

Proposition 106.31 (Dyson's argument)**.** *The perturbation series of quantum electrodynamics in powers of e^2 has zero radius of convergence [998].*

Rests on Proposition 106.29.

Derivation. *Derives Proposition 106.31.*

Suppose the series converged in some disc $|e^2| < r$. Then it would converge for real negative e^2, that is for an imaginary charge, and by analyticity the sum would be the physical quantity of that theory. But with $e^2 < 0$ like charges attract: the vacuum is unstable against the spontaneous creation of a large number of electron–positron pairs, which segregate into two clouds of like charge whose negative potential energy exceeds their rest energy once enough pairs are made. No lowest-energy state exists, so the theory has no ground state, no S-matrix, and no analytic continuation of the physical quantities to compare with. Hence the radius of convergence is zero.

Dyson's argument is physical, not a theorem: it establishes that the theory at $e^2 < 0$ is pathological, and infers from this that the function of e^2 cannot be analytic at the origin. □

The concrete form of the divergence is visible in a single ordinary integral, which is the whole of the phenomenon with none of the field theory.

Example 106.32 (The zero-dimensional model)**.** Let $Z(\hat g) = \int_{-\infty}^{\infty} \mathrm{d}x\ \mathrm{e}^{-x^2/2 - \hat g x^4/4!}$, the $0+0$-dimensional version of Equation (106.32) after Wick rotation. Expanding the quartic term and using $\langle x^{4n}\rangle = (4n-1)!!$ from Lemma 106.2,

$$\frac{Z(\hat g)}{Z(0)} = \sum_{n=0}^{\infty} \frac{(-\hat g)^n}{n!\,(4!)^n}\,(4n-1)!! = 1 - \frac{\hat g}{8} + \frac{35\hat g^2}{384} - \frac{385\hat g^3}{3072} + \cdots . \tag{106.46}$$

The coefficients grow factorially. Writing $(4n-1)!! = (4n)!/(2^{2n}(2n)!)$ and applying Stirling's formula to each factorial gives

$$|a_n| = \frac{(4n-1)!!}{n!\,(4!)^n} \sim \frac{n!}{\sqrt{2}\,\pi n}\left(\frac{2}{3}\right)^n \qquad (n\to\infty)\,, \tag{106.47}$$

so the series diverges for every $\hat g \neq 0$ although the integral it came from is perfectly finite and analytic in $\hat g$ on the cut plane. The best one can do is truncate at the smallest term, $n \approx 3/(2\hat g)$, with an error of order $\mathrm{e}^{-3/(2\hat g)}$: an *asymptotic* series.

The same estimate applied to quantum electrodynamics with $\alpha = 1/137.036$ puts the optimal truncation near order $1/\alpha \approx 137$ and the irreducible error near $\mathrm{e}^{-1/\alpha} \approx 10^{-60}$ — some fifty orders of magnitude below the precision of the most accurate measurement in physics (Section 100.8.2). The non-convergence of the series is therefore a statement about the far tail of a computation nobody will ever do, not a defect of the predictions. What it does mean is that perturbation theory alone cannot define the theory: quantities of the size $\mathrm{e}^{-1/\alpha}$, or $\mathrm{e}^{-1/\hat g}$, are invisible to it, and those are exactly the instanton effects of Section 106.6.1 and the confining physics of Section 106.5. The Euclidean path integral, defined non-perturbatively on a lattice, is at present the only construction that reaches them.

106.3.3 Fermions and Grassmann integration

Why the classical fields must anticommute The construction of Theorem 106.3 inserted a resolution of the identity built from eigenstates of the coordinate operator. For a field theory the corresponding step inserts eigenstates of the field operator, $\hat{\phi}(\boldsymbol{x})\,|\phi\rangle = \phi(\boldsymbol{x})\,|\phi\rangle$ at fixed time, and the eigenvalues $\phi(\boldsymbol{x})$ are the integration variables. For a Dirac field this is impossible with ordinary numbers.

Proposition 106.33 (Field eigenvalues inherit the statistics)**.** *Let $\hat{\psi}_\alpha(\boldsymbol{x})$ satisfy the equal-time anticommutation relations $\{\hat{\psi}_\alpha(\boldsymbol{x}), \hat{\psi}_\beta(\boldsymbol{y})\} = 0$ required by the spin–statistics connection of Chapter 83, and let $|\psi\rangle$ be a simultaneous eigenstate, $\hat{\psi}_\alpha(\boldsymbol{x})\,|\psi\rangle = \psi_\alpha(\boldsymbol{x})\,|\psi\rangle$. Then the eigenvalues satisfy*

$$\psi_\alpha(\boldsymbol{x})\psi_\beta(\boldsymbol{y}) = -\psi_\beta(\boldsymbol{y})\psi_\alpha(\boldsymbol{x})\,, \tag{106.48}$$

and in particular $\psi_\alpha(\boldsymbol{x})^2 = 0$: they are Grassmann *numbers, generators of an exterior algebra, not elements of $\mathbb{R}$ or $\mathbb{C}$.*

Rests on Theorem 99.46.

Derivation. *Derives Proposition 106.33.*

Apply the two operators in both orders to the eigenstate: $\hat{\psi}_\alpha(\boldsymbol{x})\hat{\psi}_\beta(\boldsymbol{y})\,|\psi\rangle = \psi_\beta(\boldsymbol{y})\psi_\alpha(\boldsymbol{x})\,|\psi\rangle$ — reading the operators from right to left — while the anticommutator gives $\hat{\psi}_\alpha\hat{\psi}_\beta = -\hat{\psi}_\beta\hat{\psi}_\alpha$, whose eigenvalue is $-\psi_\alpha(\boldsymbol{x})\psi_\beta(\boldsymbol{y})$. Equating the two gives Equation (106.48). Setting $\alpha = \beta$, $\boldsymbol{x} = \boldsymbol{y}$ gives $\psi^2 = -\psi^2 = 0$. □

Nothing was assumed here beyond the anticommutation relations themselves. The Grassmann variables of the fermionic path integral are not a formal device chosen for convenience: they are what the eigenvalues of an anticommuting operator are. What remains is to say what it means to integrate over them, and Berezin's answer [1325] is not a postulate either — it is forced by two requirements.

Berezin's rules, derived

Definition 106.34 (Grassmann algebra)**.** Let $\theta_1, \dots, \theta_n$ generate an associative algebra over $\mathbb{C}$ with $\theta_i\theta_j = -\theta_j\theta_i$, hence $\theta_i^2 = 0$. Every element is a finite sum $f = \sum_S c_S \prod_{i\in S} \theta_i$ over subsets $S \subseteq \{1, \dots, n\}$; there are 2^n terms and the algebra is finite-dimensional. In one generator the general element is simply $f(\theta) = a + b\theta$.

Theorem 106.35 (The Berezin integral)**.** *Suppose a linear functional $f \mapsto \int \mathrm{d}\theta\, f(\theta)$ on the one-generator algebra is required to be*

1. linear*, and*
2. translation invariant*: $\int \mathrm{d}\theta\, f(\theta+\eta) = \int \mathrm{d}\theta\, f(\theta)$ for every Grassmann η.*

Then it is determined up to normalization, and with the normalization $\int \mathrm{d}\theta\, \theta = 1$,

$$\int \mathrm{d}\theta\; 1 = 0\,, \qquad \int \mathrm{d}\theta\; \theta = 1\,, \qquad \textit{hence} \qquad \int \mathrm{d}\theta\, f(\theta) = \frac{\partial f}{\partial \theta}\,. \tag{106.49}$$

Integration over a Grassmann variable is *differentiation with respect to it.*

Rests on Definition 106.34.

Derivation. *Derives Theorem 106.35.*

By linearity the functional is fixed by the two numbers $I_0 = \int \mathrm{d}\theta\, 1$ and $I_1 = \int \mathrm{d}\theta\, \theta$. Apply translation invariance to $f = a + b\theta$:

$$\int \mathrm{d}\theta\,(a + b\theta + b\eta) = aI_0 + bI_1 + b\eta I_0 \quad \text{must equal} \quad aI_0 + bI_1\,,$$

for every b and every η, which forces $I_0 = 0$. Normalizing $I_1 = 1$ then gives Equation (106.49), and since $\partial(a + b\theta)/\partial\theta = b = aI_0 + bI_1$, integration and differentiation coincide.

Translation invariance is not a technical convenience: it is exactly the property used in Theorem 106.27 to complete the square, and without it there would be no fermionic Gaussian integral and hence no fermionic path integral. It is also the property whose *failure* under a chiral rotation produces the anomaly of Section 106.7.3 — there the transformation is linear but not a translation, and its Jacobian is the whole story. □

Corollary 106.36 (Change of variables inverts)**.** *For an invertible linear change of Grassmann variables $\theta'_i = A_{ij}\theta_j$,*

$$\mathrm{d}^n\theta' = (\det A)^{-1}\mathrm{d}^n\theta\,, \tag{106.50}$$

the determinant appearing with the opposite *power to the ordinary case.*

Rests on Definition 106.34 and Theorem 106.35.

Derivation. *Derives Corollary 106.36.*

Only the top monomial survives an n-fold integration, so the rule is fixed by $\int \mathrm{d}^n\theta'\,\theta'_1\cdots\theta'_n = 1$. Expanding, $\theta'_1\cdots\theta'_n = A_{1j_1}\cdots A_{nj_n}\theta_{j_1}\cdots\theta_{j_n} = (\det A)\,\theta_1\cdots\theta_n$, since reordering the θ's into canonical order produces the sign of the permutation and the sum over permutations with signs is the Leibniz formula for the determinant. To keep the normalization, $\mathrm{d}^n\theta'$ must carry $(\det A)^{-1}$. □

Theorem 106.37 (The Grassmann Gaussian is a determinant)**.** *Let $\theta_i, \bar\theta_i$, $i = 1,\dots,n$, be independent Grassmann generators, with the convention $\int \mathrm{d}\bar\theta_i\,\mathrm{d}\theta_i\;\theta_i\bar\theta_i = 1$. Then for any $n\times n$ matrix A,*

$$\int \prod_{i=1}^{n} \mathrm{d}\bar\theta_i\,\mathrm{d}\theta_i\;\exp\!\Big(-\bar\theta_i A_{ij}\theta_j\Big) = \det A\,, \tag{106.51}$$

to be compared with the bosonic $\int \mathrm{d}^n x\,\mathrm{e}^{-x^{\mathsf T}Ax/2} = (2\pi)^{n/2}(\det A)^{-1/2}$ of Equation (106.7). With sources,

$$\int \prod_{i} \mathrm{d}\bar\theta_i\,\mathrm{d}\theta_i\;\exp\!\Big(-\bar\theta A\theta + \bar\eta\theta + \bar\theta\eta\Big) = \det A\;\exp\!\Big(\bar\eta A^{-1}\eta\Big)\,. \tag{106.52}$$

Rests on Equation (106.7) and Theorem 106.35.

Derivation. *Derives Theorem 106.37.*

Expand the exponential. Because each generator squares to zero, the only term that survives the $2n$-fold integration is the one containing each of $\theta_1,\dots,\theta_n$ and each of $\bar\theta_1,\dots,\bar\theta_n$ exactly once, namely

$$\frac{(-1)^n}{n!}\Big(\bar\theta_i A_{ij}\theta_j\Big)^n = (-1)^n\sum_{\sigma\in S_n} A_{1\sigma(1)}\cdots A_{n\sigma(n)}\,\bar\theta_1\theta_{\sigma(1)}\cdots\bar\theta_n\theta_{\sigma(n)}\,,$$

the $n!$ cancelling against the $n!$ orderings of the identical factors. Reordering each $\bar\theta_k\theta_{\sigma(k)}$ pair into the canonical order $\theta_k\bar\theta_k$ produces $\operatorname{sgn}(\sigma)$ together with $(-1)^n$, and the surviving sum is $\sum_\sigma \operatorname{sgn}(\sigma)\prod_k A_{k\sigma(k)} = \det A$. The sourced version follows by the shift $\theta \to \theta + A^{-1}\eta$, $\bar\theta \to \bar\theta + \bar\eta A^{-1}$, which is legitimate by Theorem 106.35(ii) and has unit Jacobian, being a translation. □

Equation (106.51) and Equation (106.7) differ in exactly one place — the sign of the exponent of the determinant — and every distinctive feature of fermions in the path integral follows from it.

The fermion determinant

Proposition 106.38 (Integrating out the fermions)**.** *For a Dirac field coupled to a gauge field A_μ with the Lagrangian density $\mathcal{L}_\psi = \bar\psi(\mathrm{i}\hbar c\,\gamma^\mu D_\mu - mc^2)\psi$ of Equation (100.5),*

$$\int \mathcal{D}\bar\psi\,\mathcal{D}\psi\ \exp\!\left\{\frac{\mathrm{i}}{\hbar c}\int \mathrm{d}^4x\ \bar\psi(\mathrm{i}\hbar c\gamma^\mu D_\mu - mc^2)\psi\right\}\ \propto\ \det(\mathrm{i}\hbar\gamma^\mu D_\mu - mc)\ , \tag{106.53}$$

the constant of proportionality being field-independent. Each entry of the operator carries the dimension of a momentum, so the determinant must be normalized against its free value to be a number; only ratios appear in physical quantities.

Rests on Equations (100.5) *and* (106.51)*.*

Derivation. *Derives Proposition 106.38.*

The exponent is $\mathrm{i}S/\hbar$ with $S = c^{-1}\int \mathrm{d}^4x\,\mathcal{L}_\psi$, which is a Grassmann bilinear $-\bar\psi\mathcal{A}\psi$ with $\mathcal{A} = -(\mathrm{i}/\hbar c)(\mathrm{i}\hbar c\gamma^\mu D_\mu - mc^2) = -(\mathrm{i}/\hbar)(\mathrm{i}\hbar\gamma^\mu D_\mu - mc)$ after extracting the common factor c. Apply Equation (106.51) mode by mode: the answer is $\det\mathcal{A}$, and the field-independent factors $-\mathrm{i}/\hbar$ per mode contribute an infinite but A-independent constant, absorbed into the normalization. □

Corollary 106.39 (The minus sign of a closed fermion loop)**.** *Write $\det\mathcal{A} = \exp\operatorname{tr}\ln\mathcal{A}$ and expand in powers of the gauge field. The resulting contribution to the effective action is $-\mathrm{i}\hbar\operatorname{tr}\ln\mathcal{A}$, whereas a boson of the same coupling gives $+\frac{\mathrm{i}\hbar}{2}\operatorname{tr}\ln$ by Equation (106.7). Hence every closed fermion loop carries a factor (-1) relative to the corresponding boson loop, together with a Dirac trace — which is rule five of Proposition 100.14, here derived rather than imposed.*

Rests on Proposition 106.38 and Equations (106.7) *and* (106.37)*.*

Derivation. *Derives Corollary 106.39.*

By Proposition 106.38, integrating out the fermions contributes $\det\mathcal{A}$ to Z, hence $\operatorname{tr}\ln\mathcal{A}$ to $\ln Z$ and $-\mathrm{i}\hbar\operatorname{tr}\ln\mathcal{A}$ to W by Equation (106.37). Expanding $\operatorname{tr}\ln(\mathcal{A}_0+\delta\mathcal{A}) = \operatorname{tr}\ln\mathcal{A}_0 + \sum_{k\ge1}\frac{(-1)^{k+1}}{k}\operatorname{tr}\big[(\mathcal{A}_0^{-1}\delta\mathcal{A})^k\big]$ produces, at order k, a single trace around a closed chain of k propagators and k vertices: a one-loop diagram with k external legs. For a real boson the same manipulation applied to $(\det)^{-1/2}$ gives $+\frac12\operatorname{tr}\ln$, so the fermion loop differs by the factor $-2\times\frac12 = -1$ per loop and by the appearance of a Dirac trace, the boson having no spinor index. The factor $1/2$ absent in the fermionic case is the symmetry factor of a loop of a real field, which a complex fermion loop does not have because its two orientations are distinct. □

Remark 106.40 (Quenching, and why it was abandoned)*.* The determinant Equation (106.53) is a non-local functional of the gauge field: changing a link far away changes it. Early lattice computations therefore replaced it by a constant — the *quenched* approximation, equivalent to omitting all internal quark loops — because the cost of evaluating it repeatedly exceeded the machines of the time by orders of magnitude. The approximation is not controlled: it has no small parameter, its error was estimated empirically at ten to twenty percent for typical hadronic quantities, and in some channels it is qualitatively wrong, since a quenched theory has no two-particle thresholds from sea quarks and hence a different analytic structure. It is no longer used for results that are quoted with error bars; the computation of Section 106.5.2 includes the determinant of two degenerate light flavours and one strange flavour [2122]. Squaring the determinant of the degenerate pair is what keeps the Euclidean weight positive, which is the practical form of Remark 106.15.

106.3.4 The effective action

Definition 106.41 (Effective action)**.** Let $\phi_c(x) := c\,\delta W/\delta J(x) = \langle\phi(x)\rangle_J$ be the field expectation value in the presence of the source. The effective action is the Legendre transform

$$\Gamma[\phi_c] := W[J] - \frac{1}{c}\int \mathrm{d}^4x\; J(x)\phi_c(x)\,, \tag{106.54}$$

with J eliminated in favour of ϕ_c. It has the dimension of an action, J s.

The construction is Jona-Lasinio's [1286], who introduced it precisely to have a functional whose stationary points locate the vacuum: a symmetry of the action that is not a symmetry of the minimising ϕ_c is a symmetry broken by the state and not by the dynamics, which is the situation Chapter 104 turns into masses.

Proposition 106.42 (The quantum equation of motion)**.**

$$\frac{\delta\Gamma}{\delta\phi_c(x)} = -\frac{1}{c}J(x)\,, \tag{106.55}$$

so that in the absence of a source the physical field configurations are the stationary points of Γ*, exactly as the classical ones are the stationary points of* S*.*

Rests on Equation (106.54)*.*

Derivation. *Derives Proposition 106.42.*

Vary Equation (106.54) with respect to ϕ_c, allowing J to vary with it:

$$\delta\Gamma = \int \mathrm{d}^4x\left[\frac{\delta W}{\delta J}\,\delta J - \frac{1}{c}\phi_c\,\delta J - \frac{1}{c}J\,\delta\phi_c\right] = -\frac{1}{c}\int \mathrm{d}^4x\; J\,\delta\phi_c\,,$$

the first two terms cancelling by the definition of ϕ_c. □

Theorem 106.43 (The effective action generates 1PI vertices)**.** *Write* $W_{x_1\cdots x_n}$ *and* $\Gamma_{x_1\cdots x_n}$ *for the functional derivatives of* W *with respect to* J *and of* Γ *with respect to* ϕ_c*. Then*

$$\int \mathrm{d}^4z\; W_{xz}\,\Gamma_{zy} = -\frac{1}{c^2}\,\delta^4(x-y)\,, \tag{106.56}$$

so Γ_{xy} *is, up to the factor* $-c^{-2}$*, the inverse of the full connected propagator; and*

$$W_{xyz} = c^3\int \mathrm{d}^4a\,\mathrm{d}^4b\,\mathrm{d}^4d\; W_{xa}W_{yb}W_{zd}\,\Gamma_{abd}\,, \tag{106.57}$$

so the connected three-point function is the third derivative of Γ *dressed with a full propagator on each leg. In general the* n*-th derivative of* Γ *is the amputated one-particle-irreducible* n*-point vertex.*

Rests on Definition 106.41 and Equation (106.55)*.*

Derivation. *Derives Theorem 106.43.*

By definition $\phi_c = c\,\delta W/\delta J$ and, by Equation (106.55), $J = -c\,\delta\Gamma/\delta\phi_c$. The two are inverse maps, so by the chain rule

$$\delta^4(x-y) = \int \mathrm{d}^4z\;\frac{\delta\phi_c(x)}{\delta J(z)}\,\frac{\delta J(z)}{\delta\phi_c(y)} = \int \mathrm{d}^4z\;(cW_{xz})(-c\Gamma_{zy})\,,$$

which is Equation (106.56). Differentiate it with respect to $J(w)$, using $\delta/\delta J(w) = c\int \mathrm{d}^4u\; W_{uw}\,\delta/\delta\phi_c(u)$:

$$\int \mathrm{d}^4z\; W_{xzw}\Gamma_{zy} = -c\int \mathrm{d}^4z\,\mathrm{d}^4u\; W_{xz}W_{uw}\Gamma_{zyu}\,.$$

Contract with $-c^2 W_{yb}$ and integrate over y; the left-hand side collapses by Equation (106.56) to W_{xbw}, and the right-hand side becomes $c^3 \int W_{xz} W_{by} W_{wu} \Gamma_{zyu}$, which is Equation (106.57) after renaming.

A connected three-point function with a full propagator removed from each external leg cannot be cut into two pieces by severing one internal line, since any such cut would isolate one external leg and the piece so isolated is precisely the propagator already amputated. Hence Γ_{abd} is one-particle irreducible. Differentiating Equation (106.57) once more produces

$$W_4 = c^4 \Big[WWWW\,\Gamma_4 + (\text{three channels}) WWWWW\,\Gamma_3 \Gamma_3 \Big] ,$$

in which the second group of terms is exactly the set of one-particle-*reducible* tree graphs built from two cubic vertices joined by one full propagator. The pattern repeats: at each order the new derivative of Γ supplies the irreducible part and the terms generated by differentiating the propagators supply all the reducible trees. Induction on n gives the general statement. □

Theorem 106.43 is what makes Γ the practical object. Since it generates the irreducible vertices and the exact propagator, knowing Γ is knowing the theory: every connected amplitude is assembled from it by tree graphs alone.

The loop expansion of Γ

Theorem 106.44 (One loop)**.**

$$\Gamma[\phi_c] = S[\phi_c] + \frac{\mathrm{i}\hbar}{2}\,\operatorname{tr}\ln\left(\frac{\delta^2 S}{\delta\phi\,\delta\phi}\Big|_{\phi=\phi_c} \right) + O(\hbar^2) , \tag{106.58}$$

so the classical action is the tree approximation to Γ *and the first quantum correction is one power of* $\hbar$ *smaller, in accord with Theorem 106.30.*

Rests on Equations (106.7)*,* (106.54) *and* (106.55)*.*

Derivation. Derives Theorem 106.44.

Combining Equations (106.35), (106.37) and (106.54),

$$\exp\!\left(\frac{\mathrm{i}}{\hbar}\Gamma[\phi_c]\right) = \int \mathcal{D}\phi\, \exp\!\left\{ \frac{\mathrm{i}}{\hbar}\left(S[\phi] - \frac{1}{c}\int \mathrm{d}^4x\, \frac{\delta\Gamma}{\delta\phi_c}\,(\phi-\phi_c) \right) \right\} ,$$

using Equation (106.55) to eliminate J. Substitute $\phi = \phi_c + \sqrt{\hbar}\,\chi$ and expand:

$$S[\phi_c + \sqrt{\hbar}\chi] = S[\phi_c] + \sqrt{\hbar}\int \frac{\delta S}{\delta\phi}\chi + \frac{\hbar}{2}\iint \chi\,\frac{\delta^2 S}{\delta\phi\delta\phi}\,\chi + O(\hbar^{3/2}) .$$

At order $\sqrt{\hbar}$ the term linear in χ cancels against the source term, because $\delta\Gamma/\delta\phi_c = \delta S/\delta\phi_c + O(\hbar)$; what remains is a Gaussian integral over χ, evaluated by Equation (106.7) in the form $(\det)^{-1/2} = \exp(-\tfrac{1}{2}\operatorname{tr}\ln)$. Multiplying by $\hbar/\mathrm{i}$ to extract Γ turns $-\tfrac{1}{2}\operatorname{tr}\ln$ into $+\tfrac{\mathrm{i}\hbar}{2}\operatorname{tr}\ln$, which is Equation (106.58). The substitution $\phi = \phi_c + \sqrt{\hbar}\chi$ is what makes the $\hbar$ counting automatic: each additional vertex in the expansion brings $\hbar^{1/2}$ per field beyond the second, and the Gaussian contractions pair them up. □

The effective potential For a constant ϕ_c the effective action reduces to a function of one variable. Writing $V_4 = \int \mathrm{d}^4x$,

$$\Gamma[\phi_c] = -\frac{V_4}{c}\,U(\phi_c) , \tag{106.59}$$

which defines the effective potential U, of dimension $\mathrm{J/m^3}$; at tree level U is the classical potential $U_0 = \tfrac{1}{2}(mc/\hbar)^2\phi_c^2 + \tfrac{g}{4!}\phi_c^4$.

Proposition 106.45 (One-loop effective potential)**.** *Define the field-dependent mass* $M(\phi_c)$ *by* $(Mc/\hbar)^2 = (mc/\hbar)^2 + \frac{g}{2}\phi_c^2$*. Then, in the minimal-subtraction scheme of Definition 100.54 at renormalization energy* $\mathcal{E}$*,*

$$U_1(\phi_c) = \frac{(Mc^2)^4}{64\pi^2(\hbar c)^3}\left[\ln\frac{(Mc^2)^2}{\mathcal{E}^2} - \frac{3}{2}\right], \tag{106.60}$$

of dimension $\mathrm{J/m^3}$ *as required, since* $(Mc^2)^4/(\hbar c)^3$ *is an energy divided by a volume.*
Rests on Equations (106.58) and (106.59) and Lemma 100.28.

Derivation. *Derives Proposition 106.45.*

By Equation (106.58) the one-loop term is $\frac{\mathrm{i}\hbar}{2}\operatorname{tr}\ln\mathcal{K}_\phi$ with $\mathcal{K}_\phi = -(\Box + M^2c^2/\hbar^2)$ read off from the second variation of Equation (106.32). For a translation-invariant kernel the trace is $V_4\int \mathrm{d}^4p\,(2\pi\hbar)^{-4}$ of the momentum-space symbol, so

$$\Gamma_1 = \frac{\mathrm{i}\hbar V_4}{2}\int\frac{\mathrm{d}^4p}{(2\pi\hbar)^4}\,\ln\frac{p^2 - M^2c^2 + \mathrm{i}\epsilon}{\hbar^2},$$

and Equation (106.59) gives $U_1 = -c\Gamma_1/V_4$. Rotate by Proposition 106.14, $\mathrm{d}^4p = \mathrm{i}\,\mathrm{d}^4p_E$ and $p^2 = -p_E^2$:

$$U_1 = \frac{\hbar c}{2}\int\frac{\mathrm{d}^4p_E}{(2\pi\hbar)^4}\,\ln\frac{p_E^2 + M^2c^2}{\hbar^2},$$

up to a field-independent constant. The integral is quartically divergent and is evaluated in $d = 4 - 2\varepsilon$ dimensions by Lemma 100.28. Writing $\Lambda^2 = M^2c^2$ for the mass-squared parameter, the standard d-dimensional result is

$$\int\frac{\mathrm{d}^dp_E}{(2\pi)^d}\,\ln(p_E^2 + \Lambda^2) = -\frac{1}{(4\pi)^{d/2}}\,\frac{(\Lambda^2)^{d/2}}{\Gamma(1)}\,\Gamma\!\left(-\frac{d}{2}\right),$$

Γ here denoting the Euler gamma function and not the effective action. Expanding its pole at $d = 4$ by Lemma 100.29 and discarding that pole together with the $-\gamma_{\mathrm{E}} + \ln 4\pi$ which accompanies it — the definition of minimal subtraction, Equation (100.75) — leaves $(\Lambda^2)^2[\ln(\Lambda^2/\mu^2) - 3/2]/(32\pi^2)$. Restoring the $\hbar^{-4}$ of the measure and the prefactor $\hbar c/2$ gives Equation (106.60) with $\mathcal{E} = \mu c$. □

The physical content of Equation (106.60) is Coleman and Weinberg's [1495]: a theory whose classical potential has its minimum at the origin can have its *effective* potential minimized away from it, so that a symmetry unbroken classically is broken by radiative corrections alone. The mechanism is real and is used in Chapter 104 as one of the two possible origins of a scalar expectation value; whether it is the origin of the electroweak one is settled by measurement, and the measured Higgs and top masses say that it is not. The one-loop potential also generates the vacuum-stability question of that chapter, since Equation (106.60) turns negative at large field when the fermion loops dominate.

Remark 106.46 (The exact effective potential is convex). The one-loop potential can have two minima; the exact one cannot have the shape between them that Equation (106.60) suggests. In the Euclidean formulation $W[J]$ is a convex functional of J, because

$$\frac{\delta^2 W_E}{\delta J(x)\delta J(y)} \propto \langle\phi(x)\phi(y)\rangle - \langle\phi(x)\rangle\,\langle\phi(y)\rangle$$

is a covariance matrix, hence positive semidefinite — and it is positive by reflection positivity Equation (106.30). The Legendre transform of a convex function is convex, so Γ, and with it U, is convex exactly. The apparent non-convex region between two degenerate minima is therefore replaced in the exact theory by the straight segment joining them —

the Maxwell equal-area construction, the same object that replaces the unphysical loop of the van der Waals isotherm in Chapter 123. Physically, a constrained expectation value ϕ_c intermediate between the two minima is realized not by a homogeneous state of that field value but by a mixture — domains of one vacuum and of the other, in the proportion that gives the average. The loop expansion does not see this because it expands about a single homogeneous configuration. The non-convex one-loop potential is nonetheless the right object for computing the physics of a single phase, which is what it is used for in Chapter 104; the convexity statement is about a question the loop expansion was not asked.

106.3.5 The LSZ reduction formula

The path integral computes time-ordered correlation functions of fields. Experiments measure the probabilities of scattering processes. Lehmann, Symanzik and Zimmermann supplied the bridge [1073], and its content is that the S-matrix lives in the poles of the correlators.

Proposition 106.47 (Källén–Lehmann representation)**.** *For an interacting scalar theory the exact two-point function is*

$$\widetilde{\Delta}(p) = \frac{\mathrm{i}\hbar^3 c Z}{p^2 - m^2c^2 + \mathrm{i}\epsilon} + \int_{(2mc)^2}^{\infty} \mathrm{d}s\, \rho(s)\, \frac{\mathrm{i}\hbar^3 c}{p^2 - s + \mathrm{i}\epsilon}\,, \tag{106.61}$$

where m is the physical mass of the one-particle state, s is a squared momentum, $\rho \geq 0$ is the spectral density of the multiparticle continuum, and the field-strength renormalization *$Z = |\langle 0|\,\phi(0)\,|p\rangle|^2 \,/\, |\langle 0|\,\phi_{\mathrm{free}}(0)\,|p\rangle|^2$ is a dimensionless number with $0 \leq Z \leq 1$.*

Rests on Theorem 106.20 and Equation (106.30).

Derivation. *Derives Proposition 106.47.*

Insert a complete set of eigenstates of the four-momentum operator between the two fields in $\langle 0|\,T\phi(x)\phi(y)\,|0\rangle$. Poincaré invariance and the spectrum condition (Theorem 105.34) organize them into the vacuum, which contributes only $|\langle 0|\,\phi\,|0\rangle|^2$ and is removed by normal-ordering the field; the one-particle states of mass m, whose contribution is fixed by Lorentz invariance to be that of a free field of mass m times a constant Z; and states whose invariant mass is at least $2m$, summarized by a density $\rho(s) \geq 0$ supported on $s \geq (2mc)^2$. The positivity of ρ is the positivity of the Hilbert-space norm, i.e. Equation (106.30) once more, and the sum of all contributions must reproduce the canonical equal-time commutator, which gives $Z + \int \rho = 1$ and hence $Z \leq 1$. □

Theorem 106.48 (LSZ reduction)**.** *Let $\widetilde{G}_n(p_1, \dots, p_n)$ be the momentum-space connected correlator of n fields, all momenta incoming. Then the S-matrix element for the process in which the momenta with $p_k^0 < 0$ are incoming particles and the rest outgoing is*

$$\mathrm{i}\mathcal{M}\,(2\pi\hbar)^4 \delta^4\!\left(\textstyle\sum_k p_k\right) = \lim_{p_k^2 \to m^2c^2} \prod_{k=1}^{n} \frac{p_k^2 - m^2c^2}{\mathrm{i}\hbar^3 c\sqrt{Z}}\; \widetilde{G}_n(p_1, \dots, p_n)\,, \tag{106.62}$$

that is: put each external leg on its mass shell, divide out the full propagator, and take one factor $\sqrt{Z}$ off each leg.

Rests on Proposition 106.47.

Derivation. *Derives Theorem 106.48.*

Consider the dependence of $\widetilde{G}_n$ on one momentum p_1. By Proposition 106.47 applied to the leg carrying p_1, the correlator has a simple pole at $p_1^2 = m^2c^2$ whose residue is $\mathrm{i}\hbar^3 c Z$ times the matrix element of the remaining fields between the vacuum and the one-particle state $|p_1\rangle$, divided by $\sqrt{Z}$ — one factor of $\sqrt{Z}$ being consumed by the overlap $\langle 0|\,\phi(0)\,|p_1\rangle$ and the other remaining in the residue. Multiplying by $(p_1^2 - m^2c^2)$ and taking the limit isolates that residue and discards the multiparticle continuum, which has only a branch cut

starting at $(2mc)^2$ and is regular at the pole. Repeating for each leg replaces every field by an asymptotic one-particle state; what remains between them is by definition the S-matrix element, which carries the overall momentum-conserving delta function.

The physical statement behind the algebra is that an asymptotic particle is a pole in a correlation function, and nothing else: the interacting field ϕ need not create a one-particle state cleanly, and generically it creates a superposition of one-particle and multiparticle states. The prescription discards the second because it is analytically distinguishable from the first. □

Corollary 106.49 (Any interpolating field will do)**.** *Let $\mathcal{O}(x)$ be any local operator with the quantum numbers of the particle and with $\langle 0|\,\mathcal{O}(0)\,|p\rangle \neq 0$. Then Equation (106.62) holds with ϕ replaced by $\mathcal{O}$ and $\sqrt{Z}$ replaced by $\langle 0|\,\mathcal{O}(0)\,|p\rangle$ appropriately normalized, and the S-matrix element obtained is the same.*

Rests on Theorem 106.48 and Equation (106.61).

Derivation. Derives Corollary 106.49.

The derivation above used only two properties of ϕ: that it is local, so the spectral decomposition Equation (106.61) applies to it; and that it has non-zero overlap with the one-particle state, so the pole is present. Any $\mathcal{O}$ with those properties has the same pole at $p^2 = m^2c^2$, differing only in its residue, and the prescription divides that residue out. The S-matrix element is a property of the states, not of the operator used to reach them. □

Corollary 106.49 is why lattice computations may use whatever operator is numerically convenient to create a proton — any gauge-invariant combination of three quark fields with the right quantum numbers — and still measure the same mass. It is also the reason the composite pion field and the quark bilinear $\bar{d}\gamma^5 u$ give the same $\pi \to \mu\nu$ amplitude, which the weak-interaction phenomenology of Chapter 101 takes for granted. The scattering formalism into which Equation (106.62) feeds is Chapter 86, and the analyticity assumptions under which the pole structure used here is guaranteed belong to Chapter 109.

106.4 Gauge fields

106.4.1 Faddeev–Popov ghosts

Notation for the non-abelian case

Notation 106.50 (Absorbing the coupling into the potential). Let T^a be Hermitian generators of a compact simple Lie group with $\left[T^a, T^b\right] = \mathrm{i} f^{abc} T^c$ and $\mathrm{tr}(T^a T^b) = \frac{1}{2}\delta^{ab}$ (Section 14.3). Write

$$\mathcal{A}_\mu := \frac{g}{\hbar} A^a_\mu T^a\,, \qquad [\mathcal{A}_\mu] = 1/\mathrm{m}\,, \tag{106.63}$$

so that the covariant derivative is $D_\mu = \partial_\mu + \mathrm{i}\mathcal{A}_\mu$ — which for one abelian generator is exactly Equation (100.9) with $\mathcal{A}_\mu = QA_\mu/\hbar$ for a field of charge Q — and the field strength is

$$\mathcal{F}_{\mu\nu} := \frac{1}{\mathrm{i}}\,[D_\mu, D_\nu] = \partial_\mu \mathcal{A}_\nu - \partial_\nu \mathcal{A}_\mu + \mathrm{i}\,[\mathcal{A}_\mu, \mathcal{A}_\nu]\,, \qquad [\mathcal{F}_{\mu\nu}] = 1/\mathrm{m}^2\,. \tag{106.64}$$

The Yang–Mills Lagrangian density is then

$$\mathcal{L}_{\mathrm{YM}} = -\frac{\hbar c}{2\hat{g}^2}\,\mathrm{tr}(\mathcal{F}_{\mu\nu}\mathcal{F}^{\mu\nu}) = -\frac{\hbar c}{4\hat{g}^2}\,\mathcal{F}^a_{\mu\nu}\mathcal{F}^{a\,\mu\nu}\,, \tag{106.65}$$

of dimension $\mathrm{J/m^3}$, since $\hbar c = 3.161\,53 \times 10^{-26}\,\mathrm{J\,m}$ and $\mathcal{F}^2$ carries $1/\mathrm{m}^4$. The number $\hat{g}^2$ is the *dimensionless* coupling,

$$\hat{g}^2 := 4\pi\alpha_s\,, \tag{106.66}$$

in exact parallel with Equation (100.3): for electromagnetism $\hat{g}^2 = \mu_0 c e^2/\hbar = 4\pi\alpha$ by Equation (100.4), $\mathcal{F}_{\mu\nu} = eF_{\mu\nu}/\hbar$, and Equation (106.65) collapses to $-F_{\mu\nu}F^{\mu\nu}/4\mu_0$, which is Equation (100.12). For the strong interaction α_s is the running coupling measured in Equation (102.47). Infinitesimal gauge transformations are

$$\delta\mathcal{A}_\mu = D_\mu\omega := \partial_\mu\omega + \mathrm{i}\,[\mathcal{A}_\mu, \omega]\ , \tag{106.67}$$

with $\omega = \omega^a T^a$ dimensionless ($\omega^a = g\Lambda^a/\hbar$ in the notation of Equation (100.8)).

Every formula of this section is therefore free of g and $\hbar$ except through the single dimensionless number $\hat{g}^2$, and the factor $\hbar c/\hat{g}^2$ multiplying the action. That is not a retreat into natural units: $\mathcal{A}$, $\mathcal{F}$ and ω have been given SI dimensions above, and the action $S_{\mathrm{YM}} = c^{-1}\int \mathrm{d}^4x\,\mathcal{L}_{\mathrm{YM}} = -\hbar\hat{g}^{-2}\int \mathrm{d}^4x\,\mathcal{F}^a_{\mu\nu}\mathcal{F}^{a\mu\nu}/4$ carries J s because $\hbar$ stands in front of a dimensionless integral. It is precisely this form that makes $S/\hbar$ visibly equal to a pure number divided by $\hat{g}^2$, which is the origin of the $\mathrm{e}^{-8\pi^2/\hat{g}^2}$ of Section 106.6.1.

The redundancy, and the divergence it causes Section 100.1.3 showed that the quadratic form of the photon action is singular: $\widetilde{M}_{\mu\nu}p^\nu = 0$ in Equation (100.15). In the functional integral the same fact appears as an infinite factor.

Proposition 106.51 (The naive integral diverges by the volume of the gauge group)**.** *The integrand of $\int \mathcal{D}\mathcal{A}\,\mathrm{e}^{\mathrm{i}S_{\mathrm{YM}}/\hbar}$ is constant along each gauge orbit $\{\mathcal{A}^\omega : \omega \in \mathcal{G}\}$, so the integral contains an overall factor equal to the volume of the group $\mathcal{G}$ of gauge transformations — an infinite constant, independent of the physics but not factorizable in the presence of sources, since the source term is not gauge invariant.*

Rests on Equations (106.65) and (106.67).

Derivation. Derives Proposition 106.51.

S_{YM} is invariant under Equation (106.67) by construction, and the measure $\mathcal{D}\mathcal{A}$ is invariant because Equation (106.67) is, at each point, a translation of $\mathcal{A}_\mu$ by $\partial_\mu\omega$ plus a rotation in the adjoint index, and both preserve the Lebesgue measure on the fibre. The integral therefore factorizes as $\left(\int_{\mathcal{G}} \mathcal{D}\omega\right) \times \left(\int_{\mathcal{A}/\mathcal{G}} \mathrm{e}^{\mathrm{i}S/\hbar}\right)$, the first factor being the divergent volume. □

Faddeev and Popov's device [1347] is to insert a resolution of unity that selects one representative from each orbit and supplies the Jacobian of that choice.

Theorem 106.52 (Faddeev–Popov)**.** *Let $G^a[\mathcal{A}](x)$ be a gauge-fixing function — for the covariant family, $G^a = \partial^\mu\mathcal{A}^a_\mu$ — such that $G^a[\mathcal{A}^\omega] = \sigma^a$ has exactly one solution ω on each orbit. Then*

$$\int \mathcal{D}\mathcal{A}\,\mathrm{e}^{\mathrm{i}S_{\mathrm{YM}}/\hbar} = \left(\int_{\mathcal{G}} \mathcal{D}\omega\right)\int \mathcal{D}\mathcal{A}\,\delta\big[G[\mathcal{A}] - \sigma\big]\,\det\mathcal{M}[\mathcal{A}]\,\mathrm{e}^{\mathrm{i}S_{\mathrm{YM}}/\hbar}\,, \tag{106.68}$$

with the Faddeev–Popov operator

$$\mathcal{M}^{ab}(x,y) = \left.\frac{\delta G^a[\mathcal{A}^\omega](x)}{\delta\omega^b(y)}\right|_{\omega=0} = \partial^\mu\left(D_\mu\right)^{ab}\delta^4(x-y)\,, \qquad \left(D_\mu\right)^{ab} = \delta^{ab}\partial_\mu + f^{abc}\mathcal{A}^c_\mu\,, \tag{106.69}$$

for the covariant gauge condition; the sign of the second term is fixed by $\left[T^a, T^b\right] = \mathrm{i}f^{abc}T^c$ together with $D_\mu = \partial_\mu + \mathrm{i}\mathcal{A}_\mu$ and is derived below. The divergent factor now stands outside and cancels in every normalized expectation value.

Rests on Proposition 106.51 and Equation (106.67).

Derivation. *Derives Theorem 106.52.*

For each $\mathcal{A}$ define $\Delta[\mathcal{A}]$ by

$$\Delta[\mathcal{A}]^{-1} := \int_{\mathcal{G}} \mathcal{D}\omega \; \delta\big[G[\mathcal{A}^{\omega}] - \sigma\big] \, .$$

By hypothesis the delta functional is supported at a single ω_*, and the one-dimensional identity $\int \mathrm{d}x \, \delta(f(x)) = |f'(x_*)|^{-1}$ generalizes to $\Delta[\mathcal{A}] = |\det(\delta G/\delta\omega)|_{\omega_*} = |\det \mathcal{M}|$. Also Δ is gauge invariant, since $\mathcal{D}\omega$ is a left- and right-invariant Haar measure on $\mathcal{G}$ and the defining integral is unchanged by relabelling the orbit representative. Insert $1 = \Delta[\mathcal{A}] \int \mathcal{D}\omega \, \delta[G[\mathcal{A}^{\omega}] - \sigma]$ into the left-hand side of Equation (106.68), exchange the order of integration, and change variables $\mathcal{A} \to \mathcal{A}^{-\omega}$; the action, the measure and Δ are all invariant, so the ω integral factorizes out as the group volume and what remains is the right-hand side.

For the covariant condition, $G^a[\mathcal{A}^{\omega}] = \partial^{\mu}(\mathcal{A}^a_{\mu} + (D_{\mu}\omega)^a)$ by Equation (106.67), whose derivative with respect to ω^b is $\partial^{\mu}(D_{\mu})^{ab}$, i.e. Equation (106.69). The component form of the adjoint covariant derivative follows from $\mathrm{i}\,[\mathcal{A}_{\mu}, \omega] = \mathrm{i}\mathcal{A}^c_{\mu}\omega^b \mathrm{i} f^{cba} T^a = -f^{cba}\mathcal{A}^c_{\mu}\omega^b T^a$, and $f^{cba} = -f^{bca} = -f^{abc}$ by total antisymmetry and cyclicity, so that $(D_{\mu}\omega)^a = \partial_{\mu}\omega^a + f^{abc}\mathcal{A}^c_{\mu}\omega^b$ — the plus sign of Equation (106.69). Equivalently, the adjoint generators are $(T^a_{\mathrm{adj}})^{bc} = -\mathrm{i}f^{abc}$, as they are for SU(3) in Proposition 14.55, and $\delta^{bc}\partial_{\mu} + \mathrm{i}\mathcal{A}^a_{\mu}(T^a_{\mathrm{adj}})^{bc} = \delta^{bc}\partial_{\mu} + f^{abc}\mathcal{A}^a_{\mu}$ is the same operator with the summed index moved by one cyclic step. □

Corollary 106.53 (Why electrodynamics has no ghosts)**.** *For an abelian group $f^{abc} = 0$, so $\mathcal{M} = \partial^{\mu}\partial_{\mu} = \Box$ is independent of $\mathcal{A}$; $\det \mathcal{M}$ is then a constant that cancels between numerator and denominator, and the Faddeev–Popov procedure reduces to the gauge-fixing term Equation (100.16) alone. This is the derivation of a fact that Section 100.1.3 could only assert.*

Rests on Theorem 106.52 and Equations (100.16) and (106.69).

Exponentiating the determinant: ghosts The delta functional is inconvenient and the determinant is not a local object. Both are repaired at once, and it is Section 106.3.3 that repairs them.

Theorem 106.54 (The gauge-fixed action)**.** *Averaging Equation (106.68) over the arbitrary function σ with the Gaussian weight $\exp\{-\frac{\mathrm{i}}{2\hat{g}^2\xi}\int \mathrm{d}^4x \, \sigma^a\sigma^a\}$ — dimensionless, since σ carries $1/\mathrm{m}^2$ — and writing the determinant as a Grassmann integral by Equation (106.51) gives*

$$Z = \int \mathcal{D}\mathcal{A}\,\mathcal{D}\bar{c}\,\mathcal{D}c \; \exp\!\Big\{\frac{\mathrm{i}}{\hbar}(S_{\mathrm{YM}} + S_{\mathrm{gf}})\Big\} \exp\!\Big\{-\int \mathrm{d}^4x \; \bar{c}^a (\partial^{\mu} D_{\mu})^{ab} c^b\Big\} \, , \tag{106.70}$$

with

$$S_{\mathrm{gf}} = -\frac{\hbar}{2\hat{g}^2\xi}\int \mathrm{d}^4x \; (\partial^{\mu}\mathcal{A}^a_{\mu})^2 \, , \tag{106.71}$$

the R_{ξ} family, ξ dimensionless. The ghost *fields $c^a, \bar{c}^a$ are anticommuting, carry dimension $1/\mathrm{m}$, are scalars under the Lorentz group, and transform in the adjoint representation of the gauge group.*

Rests on Equations (106.51) and (106.68) and Lemma 106.2.

Derivation. *Derives Theorem 106.54.*

The left-hand side of Equation (106.68) does not depend on σ, so it may be averaged over σ with any normalized weight; the Gaussian one converts $\delta[G - \sigma]$ into $\exp(\mathrm{i}S_{\mathrm{gf}}/\hbar)$ with S_{gf} as stated, by Lemma 106.2. The determinant is written as a Grassmann integral by reading Equation (106.51) from right to left with $A = \mathcal{M}$ — and it is worth noting that this is the one place in physics where the exponent of a determinant being $+1$ rather than $-\frac{1}{2}$ decides the statistics of a field: the Jacobian sits in the numerator, so the field

that represents it must be the one whose Gaussian integral gives det and not $(\det)^{-1/2}$, i.e. an anticommuting one, by Theorem 106.37. The ghosts are spin-zero fields obeying Fermi statistics, which is why they cannot be physical: they violate the spin–statistics connection of Chapter 83.

Written as a single exponent $\mathrm{i}S_{\mathrm{tot}}/\hbar$, the ghost term is $S_{\mathrm{gh}} = \mathrm{i}\hbar \int \mathrm{d}^4x\, \bar{c}^a (\partial^\mu D_\mu)^{ab} c^b$, whose explicit factor of i records that the ghost action is not real. Nothing is wrong: an action is required to be real only when its exponential must be a phase for physical configurations, and the ghosts have no physical configurations. The dimension assignment follows from requiring the exponent in Equation (106.70) to be a pure number: $\int \mathrm{d}^4x$ supplies m^4, $\partial^\mu D_\mu$ supplies $1/\mathrm{m}^2$, so $\bar{c}c$ must supply $1/\mathrm{m}^2$. □

Proposition 106.55 (Ghosts are required by unitarity)**.** *In a non-abelian theory the ghost loop is not optional. Cutting a one-loop gluon self-energy diagram with the covariant polarization sum $-\eta^{\mu\nu}$ counts four intermediate polarizations where the physical theory has two; the ghost loop, carrying (-1) by Corollary 106.39 and running over a pair $(c, \bar{c})$, subtracts exactly the two unphysical ones, and the optical theorem is restored.*

Rests on Theorem 106.54 and Corollary 106.39.

Derivation. *Derives Proposition 106.55.*

A massless vector of momentum k has two physical transverse polarizations, whose sum is $\sum_{\lambda=1,2} \varepsilon^\mu_{(\lambda)} \varepsilon^{*\nu}_{(\lambda)} = -\eta^{\mu\nu} + (k^\mu n^\nu + n^\mu k^\nu)/(k\cdot n)$ for a lightlike auxiliary n. In a covariant gauge the propagator supplies $-\eta^{\mu\nu}$, which is the same object plus the longitudinal and timelike pair; those two do not cancel of their own accord when the external current is not conserved, and in a non-abelian theory it is not conserved diagram by diagram, because the gluon carries colour and its own current is only covariantly conserved. The discrepancy is a two-state contribution of positive norm. The ghost loop supplies two scalar states (c and $\bar{c}$) with the sign (-1) of a closed anticommuting loop, i.e. -2 states, which cancels it identically. The all-orders statement — that the cancellation holds to every order and for every process — is the Kugo–Ojima quartet mechanism of Section 106.4.2, and it is a theorem about the BRST cohomology rather than a diagrammatic accident. □

The same construction can be reached from the Hamiltonian side. Gauge invariance makes the Yang–Mills Lagrangian singular in Dirac's sense: the momentum conjugate to $\mathcal{A}_0$ vanishes identically, so the system carries first-class constraints and its quantization requires the Dirac-bracket apparatus of Chapter 26. Carrying that programme through and then integrating out the non-dynamical components reproduces Equation (106.70), ghosts and all. The functional route is shorter and manifestly covariant, which is why it is the one used; the Hamiltonian route is what shows that the answer is not an artefact of the trick.

106.4.2 BRST symmetry

The gauge-fixed action Equation (106.70) is not gauge invariant — that was the point of fixing the gauge. Becchi, Rouet and Stora [1518, 1571], and independently Tyutin [1564], found that it retains an exact global symmetry whose parameter is anticommuting, and that this residual symmetry carries everything the lost gauge invariance was doing.

Definition 106.56 (The BRST operator)**.** Introduce the auxiliary Nakanishi–Lautrup field $B = B^a T^a$, of dimension $1/\mathrm{m}^2$, and define a Grassmann-odd derivation s by

$$s\mathcal{A}_\mu = D_\mu c\,, \qquad sc = -\mathrm{i}c^2\,, \qquad s\bar{c} = B\,, \qquad sB = 0\,, \tag{106.72}$$

where $c = c^a T^a$, $\bar{c} = \bar{c}^a T^a$, and $c^2 = \tfrac{1}{2}\{c, c\}$ is non-zero because the components anticommute while the matrices do not. On a matter field ψ transforming in a representation R, $s\psi = -\mathrm{i}c\psi$. The operator acts on products by $s(XY) = (sX)Y + (-1)^{|X|}X(sY)$.

Theorem 106.57 (Nilpotency)**.** $s^2 = 0$ *on every field of Equation* (106.72)*.*
Rests on Equation (106.72)*.*

Derivation. *Derives Theorem 106.57.*

On the ghost: $s^2 c = -\mathrm{i}\big[(sc)c - c(sc)\big] = -\mathrm{i}\big[(-\mathrm{i}c^2)c - c(-\mathrm{i}c^2)\big] = -c^3 + c^3 = 0$, the sign in the middle being that of s passing the odd c. Associativity of matrix multiplication has done here exactly the work that the Jacobi identity does in component notation, which is the reason for writing the algebra in matrix form.

On the connection, expand $s^2\mathcal{A}_\mu = s(\partial_\mu c + \mathrm{i}\,[\mathcal{A}_\mu, c])$. Since $\mathcal{A}$ is even and c odd, $s\,[\mathcal{A}_\mu, c] = \{s\mathcal{A}_\mu, c\} + [\mathcal{A}_\mu, sc]$, so

$$s^2\mathcal{A}_\mu = \partial_\mu(sc) + \mathrm{i}\,\{D_\mu c, c\} + \mathrm{i}\,[\mathcal{A}_\mu, sc]\ .$$

Insert $sc = -\mathrm{i}c^2$ and use $\partial_\mu(c^2) = \{\partial_\mu c, c\}$ and $\{[\mathcal{A}_\mu, c]\,, c\} = [\mathcal{A}_\mu, c^2]$:

$$s^2\mathcal{A}_\mu = -\mathrm{i}\,\{\partial_\mu c, c\} + \mathrm{i}\,\{\partial_\mu c, c\} - \Big[\mathcal{A}_\mu, c^2\Big] + \Big[\mathcal{A}_\mu, c^2\Big] = 0\,.$$

On $\bar c$ and B nilpotency is immediate from $sB = 0$. Since s is a derivation and vanishes on the square of every generator, $s^2 = 0$ on any polynomial in the fields. □

Theorem 106.58 (The gauge-fixing sector is BRST-exact)**.** *With the gauge fermion*

$$\Psi = -\frac{\hbar}{\hat g^2}\int \mathrm{d}^4x\ \bar c^a\left(\partial^\mu \mathcal{A}^a_\mu - \frac{\xi}{2}B^a\right)\,, \tag{106.73}$$

the total action is

$$S_{\mathrm{tot}} = S_{\mathrm{YM}} + s\Psi\,, \tag{106.74}$$

and $sS_{\mathrm{tot}} = 0$*, since* S_{YM} *is gauge invariant and* $s^2 = 0$*.*
Rests on Equations (106.71) *and* (106.72) *and Theorem 106.57.*

Derivation. *Derives Theorem 106.58.*

Applying s to Equation (106.73) and using Equation (106.72),

$$s\Psi = -\frac{\hbar}{\hat g^2}\int \mathrm{d}^4x\left[B^a\left(\partial^\mu\mathcal{A}^a_\mu - \frac{\xi}{2}B^a\right) - \bar c^a\,\partial^\mu(D_\mu c)^a\right]\,,$$

the sign of the second term coming from s passing the odd $\bar c$. The auxiliary field is algebraic and may be integrated out, its equation of motion being $B^a = \partial^\mu\mathcal{A}^a_\mu/\xi$; the first bracket then becomes $(\partial^\mu\mathcal{A}^a_\mu)^2/2\xi$ and, with the prefactor, is exactly S_{gf} of Equation (106.71). The second term is $+(\hbar/\hat g^2)\int \mathrm{d}^4x\,\bar c^a\partial^\mu(D_\mu c)^a$, which enters the weight $\exp(\mathrm{i}S_{\mathrm{tot}}/\hbar)$ as $\exp\big[(\mathrm{i}/\hat g^2)\int \bar c\,\partial^\mu D_\mu c\big]$, whereas Equation (106.70) carries $\exp\big[-\int \bar c\,\partial^\mu D_\mu c\big]$. The two differ by the constant factor $-\mathrm{i}/\hat g^2$ multiplying the ghost bilinear, and are brought into coincidence by rescaling the pair, $c \mapsto \kappa c$ *and* $\bar c \mapsto \kappa\bar c$ with $\kappa^2 = \mathrm{i}\hat g^2$. It must be the pair and not one against the other: the ghosts enter only through the product $\bar c c$, so a rescaling $c \mapsto \kappa c$, $\bar c \mapsto \bar c/\kappa$ leaves the ghost term exactly as it was and cannot fix a normalization. The rescaling used here multiplies the Grassmann measure by a field-independent constant, by Corollary 106.36, which cancels between numerator and denominator in every normalized expectation value — equivalently, $\det(\lambda\mathcal{M}) = \lambda^n \det\mathcal{M}$ and only ratios of determinants are ever used. Finally $sS_{\mathrm{YM}} = 0$ because s acts on $\mathcal{A}$ as a gauge transformation with parameter c, and $s(s\Psi) = s^2\Psi = 0$ by Theorem 106.57. □

Definition 106.59 (BRST charge and physical states)**.** Let Q_B be the conserved charge generating s through the graded commutator, $sX = \mathrm{i}\hbar^{-1}\,[Q_B, X]_\pm$; nilpotency of s makes $Q_B^2 = 0$. A state is *physical* when it is annihilated by Q_B, and two physical states differing by Q_B applied to something are identified:

$$\mathcal{H}_{\mathrm{phys}} := \frac{\ker Q_B}{\mathrm{im}\,Q_B}\,. \tag{106.75}$$

The construction is the standard one for any nilpotent operator: Q_B is a differential, $\ker Q_B / \operatorname{im} Q_B$ its cohomology, and the physical Hilbert space is a cohomology class rather than a subspace. The differential-geometric setting in which that pattern is usually met — connections and curvature on a principal bundle — belongs to Section 14.6, which at present fixes the notation used by Theorem 14.111 and leaves the underlying definitions owed. Two facts make the construction work, and both are checkable.

Proposition 106.60 (The quartet mechanism)**.** *The unphysical modes of a gauge field organize into quartets: for each colour and momentum, the longitudinal and timelike polarizations together with the ghost and antighost form a set of four states on which Q_B acts by mapping two into the other two. Every such quartet contributes zero to any matrix element between states of* $\ker Q_B$*, so unphysical states cancel in pairs [1641].*

Rests on Equation (106.72) *and Definition 106.59.*

Derivation. *Derives Proposition 106.60.*

From Equation (106.72) the free-field limit of the transformations is $s\mathcal{A}_\mu = \partial_\mu c$, $sc = 0$, $s\bar{c} = B$, $sB = 0$: the ghost c is Q_B-closed and maps the longitudinal polarization; the antighost $\bar{c}$ maps onto B, which is the timelike combination on shell. So among the four states $\{\mathcal{A}_L, \mathcal{A}_0, c, \bar{c}\}$, two are Q_B-exact and two are not Q_B-closed. A physical state, being Q_B-closed, has zero inner product with a Q_B-exact one (since $\langle \Phi | Q_B \Lambda \rangle = \langle Q_B \Phi | \Lambda \rangle = 0$ by the hermiticity of Q_B), and the states that are not Q_B-closed are excluded from $\mathcal{H}_{\text{phys}}$ by definition. Hence the quartet decouples entirely, leaving the two transverse polarizations, which is Proposition 106.55 at all orders. The inner product induced on $\mathcal{H}_{\text{phys}}$ is therefore positive definite and the S-matrix restricted to it is unitary [1641]. □

Theorem 106.61 (Gauge-parameter independence)**.** *Any expectation value of a Q_B-closed operator is independent of the gauge parameter ξ and of the choice of gauge-fixing function altogether.*

Rests on Equation (106.74) *and Theorem 106.58.*

Derivation. *Derives Theorem 106.61.*

By Equation (106.74) the entire dependence on the gauge-fixing choice is in Ψ. Let $\mathcal{O}$ satisfy $s\mathcal{O} = 0$ and vary $\Psi \to \Psi + \delta\Psi$. Then

$$\delta \langle \mathcal{O} \rangle = \frac{\mathrm{i}}{\hbar} \langle \mathcal{O}\, s\, \delta\Psi \rangle = \frac{\mathrm{i}}{\hbar} \langle s(\mathcal{O}\, \delta\Psi) \rangle = 0 \,,$$

the second step because $s\mathcal{O} = 0$ and the last because the expectation value of a BRST variation vanishes: the measure and the action are s-invariant, so $\langle sX \rangle = 0$ for every X, by the same argument that gives a Ward identity from any symmetry of the functional integral. In particular the ξ-dependence of Equation (106.71) cancels out of every physical quantity, which is the checkable statement quoted at Equation (100.18). □

That is what BRST buys, and it is worth listing plainly because the formalism can look like machinery for its own sake. It supplies (i) a definition of the physical state space in a manifestly covariant gauge, Equation (106.75); (ii) a proof that the unphysical states decouple to all orders, Proposition 106.60; (iii) a one-line proof of gauge-parameter independence, Theorem 106.61, replacing an order-by-order diagrammatic check; and (iv) the functional identities of Section 106.4.3, which are what makes renormalizability provable. The generalization to gauge algebras that close only on shell is the antifield construction of Batalin and Vilkovisky [1664], in which s is generated by an antibracket and nilpotency becomes a master equation; it is not needed for the Standard Model, whose gauge algebra closes off shell, and it is mentioned here only because Section 106.4.3 states its identity in that language.

106.4.3 Slavnov–Taylor identities

Theorem 106.62 (Slavnov–Taylor)**.** *For every functional X of the fields,*

$$\langle sX \rangle = 0\,, \tag{106.76}$$

and the resulting relations among Green functions are the Slavnov–Taylor identities [1458, 1484]. In the abelian case they reduce to the Ward–Takahashi identity Equation (100.71).

Rests on Equations (106.70) and (106.72) and Theorem 106.58.

Derivation. *Derives Theorem 106.62.*

Change variables in Equation (106.70) by an infinitesimal BRST transformation with a constant anticommuting parameter ζ: $\Phi \to \Phi + \zeta\, s\Phi$ for every field Φ. The action is invariant by Theorem 106.58. The measure is invariant too, and this requires proof rather than assertion: the Jacobian of the transformation is $\det(1 + \zeta\, \partial(s\Phi)/\partial\Phi) = 1 + \zeta\ \mathrm{tr}(\partial(s\Phi)/\partial\Phi)$ with a sign for each Grassmann direction, and the bosonic and fermionic traces cancel — the $\mathcal{A}$–c block against the c–c block — because s maps a boson to a ghost and a ghost to a ghost bilinear with matched coefficients. (This cancellation is exactly what fails for a chiral rotation of the fermion measure, and its failure is the anomaly of Section 106.7.3; a gauge anomaly is the statement that it fails here too, and Section 106.7.4 is the condition that it does not.) With action and measure invariant, the integral of a total BRST variation vanishes, which is Equation (106.76).

In the abelian case the ghosts decouple by Corollary 106.53 and Equation (106.72) collapses to $s\mathcal{A}_\mu = \partial_\mu c$ with c a free field, so that a BRST variation is a gauge transformation whose parameter is an anticommuting constant times that free field. Taking $X = \bar{c}(y)\,\psi(x_1)\bar{\psi}(x_2)$ then turns Equation (106.76) into the relation between the vertex function and the inverse propagator, $q_\mu \Gamma^\mu(p', p) = S^{-1}(p') - S^{-1}(p)$, whose derivation is carried out in full at Theorem 100.48. □

Corollary 106.63 (The Zinn-Justin equation)**.** *Coupling sources K^μ_a, L_a to the composite operators $s\mathcal{A}^a_\mu$ and sc^a and Legendre-transforming as in Definition 106.41, the whole family Equation (106.76) collapses into the single bilinear equation*

$$\int \mathrm{d}^4x \left[\frac{\delta\Gamma}{\delta K^\mu_a} \frac{\delta\Gamma}{\delta \mathcal{A}^a_\mu} + \frac{\delta\Gamma}{\delta L_a} \frac{\delta\Gamma}{\delta c^a} \right] = 0\,, \tag{106.77}$$

which is the statement that the effective action is BRST invariant with the quantum-corrected transformation law.

Rests on Definition 106.41, Equation (106.76), and Theorem 106.57.

Derivation. *Derives Corollary 106.63.*

Add to the action the terms $\int \mathrm{d}^4x\, (K^\mu_a\, s\mathcal{A}^a_\mu + L_a\, sc^a)$. They are BRST invariant by Theorem 106.57, since s applied to them gives $K\, s^2\mathcal{A} + L\, s^2 c = 0$, so the extended action is still s-invariant and Equation (106.76) still holds. Take $X = \Phi$ for each field in turn: the resulting identities say that W is annihilated by the operator $\int \mathrm{d}^4x\, [(\delta W/\delta K^\mu_a)(\delta/\delta j^a_\mu) + \cdots]$ acting on the sources. Legendre-transforming in the field variables only — K and L are carried along untouched, so that $\delta\Gamma/\delta K = \delta W/\delta K$ — converts the source derivatives into field derivatives of Γ and produces Equation (106.77). The equation is bilinear in Γ because the composite operators $s\mathcal{A}$ and sc are themselves products of two fields, so their sources cannot be eliminated linearly. □

The identities do three jobs. First, they constrain the divergences: the counterterms needed to renormalize a gauge theory must themselves satisfy Equation (106.77), which forces the gauge coupling appearing in the three-gluon, four-gluon, ghost and matter vertices to renormalize by the same factor — the non-abelian generalization of $Z_1 = Z_2$

(Corollary 100.49). Without that, a single coupling constant would not stay single, and the theory would not be renormalizable with finitely many parameters. This is the role the identities play in 't Hooft's proof [1448, 1449], completed with the dimensional-regularization technique that respects them [1463] — and the fact that dimensional regularization respects gauge invariance, where a momentum cutoff does not, is why Definition 100.30 is the scheme of choice throughout this part.

Second, they give the practical check of Theorem 106.61: a physical quantity computed in Feynman gauge $\xi = 1$ and in Landau gauge $\xi = 0$ must agree, while the individual propagators and vertices do not. Every serious calculation in this part is done twice for that reason.

Third, they are the object that a gauge anomaly destroys, which is the subject of Section 106.7.4.

106.4.4 The Gribov ambiguity

Theorem 106.52 assumed that $G[\mathcal{A}^\omega] = \sigma$ has *exactly one* solution on each orbit. Gribov showed that for the Coulomb and Landau conditions it does not [1623].

Proposition 106.64 (Gribov copies exist)**.** *Let $\mathcal{A}$ satisfy the Landau condition $\partial^\mu \mathcal{A}_\mu = 0$. A gauge-equivalent field $\mathcal{A}^\omega$ satisfies the same condition, to first order in ω, if and only if*

$$\mathcal{M}^{ab}\omega^b = \partial^\mu (D_\mu)^{ab}\omega^b = 0\,, \tag{106.78}$$

i.e. if and only if the Faddeev–Popov operator has a normalizable zero mode. Since $\mathcal{M} = \Box + O(\mathcal{A})$ is positive for small $\mathcal{A}$ and its lowest eigenvalue decreases continuously as $\mathcal{A}$ grows, there is a surface in field space — the Gribov horizon *— on which the lowest eigenvalue first vanishes, and beyond it copies exist.*

Rests on Equations (106.67) and (106.69).

Derivation. Derives Proposition 106.64.

Under Equation (106.67), $\partial^\mu \mathcal{A}^\omega_\mu = \partial^\mu \mathcal{A}_\mu + \partial^\mu (D_\mu \omega)$, so preserving the condition requires Equation (106.78). In the Euclidean theory the Landau condition $\partial^\mu \mathcal{A}_\mu = 0$ makes $-\mathcal{M}$ Hermitian on square-integrable functions, with $-\mathcal{M}^{ab} = -\delta^{ab}\partial^2 - f^{abc}\mathcal{A}^{c\mu}\partial_\mu$ by Equation (106.69), so its spectrum is real; at $\mathcal{A} = 0$ it is the positive operator $-\partial^2$ on non-constant modes. As the amplitude of $\mathcal{A}$ is increased along any ray in field space the eigenvalues move continuously, and Gribov exhibited explicit configurations for which the lowest one crosses zero, at which point Equation (106.78) has a solution and the orbit meets the gauge-fixing surface again. □

The failure is therefore not perturbative: it occurs only for fields of amplitude of order $1/\hat{g}$, which is invisible to any finite order in $\hat{g}^2$. Everything computed in Sections 100.4 and 106.4.3 is unaffected. What is affected is any statement about the infrared behaviour of the non-abelian theory, which is where the large-amplitude configurations live.

Theorem 106.65 (Singer)**.** *For a non-abelian structure group there is no continuous gauge-fixing prescription that works globally: no continuous section of the principal bundle of gauge fields over the space of orbits exists, for gauge fields on S^4 or on S^3 [1626].*

Rests on Proposition 106.64.

Singer's statement is not about the particular choice $\partial^\mu \mathcal{A}_\mu = 0$; it says that *every* choice fails. The obstruction is topological: a gauge-fixing prescription is a section of the principal bundle whose base is the space of orbits, a principal bundle admits a global section if and only if it is trivial, and this one is not. Both statements are quoted here from [1626]; the bundle language they are written in belongs to Section 14.6, which so far carries only the notation that Theorem 14.111 needs and reserves the definition of a section itself. The

Faddeev–Popov construction is therefore a local construction that has been used globally, and Theorem 106.52 is exactly true only in a neighbourhood of $\mathcal{A} = 0$.

Two responses exist, and the honest position is that neither is closed.

Restrict the integration region. The *Gribov region* Ω is the set of transverse fields on which $\mathcal{M}$ is positive; it is convex and bounded in every direction, and every orbit meets it. It still contains copies, and the *fundamental modular region* $\Lambda \subset \Omega$, defined by minimizing $\int \mathrm{d}^4x\, \mathcal{A}^a_\mu \mathcal{A}^{a\mu}$ along the orbit, contains exactly one representative of each orbit except on its boundary. Restricting the functional integral to Ω modifies the infrared behaviour of the gluon propagator, and the two candidate answers in the literature — a propagator vanishing at zero momentum and one approaching a finite constant — correspond to different treatments of the boundary. Which is right is not settled.

Do not fix a gauge at all. On the lattice (Section 106.5.1) the integration variables are group elements and the group volume per link is finite, so Proposition 106.51 does not arise: the integral over gauge orbits contributes a finite constant and is simply left in. No gauge fixing means no Faddeev–Popov determinant, no ghosts and no Gribov copies. This is the strongest practical argument for the lattice formulation, and it is why the results of Section 106.5.2 are free of the ambiguity entirely.

Derivation pending #164

The Gribov horizon and Singer's obstruction: the derivation above establishes the criterion — a copy exists exactly when the Faddeev–Popov operator acquires a normalizable zero mode — and states that configurations reaching it exist. Gribov's explicit spherically symmetric family, the proof that the Gribov region is convex and bounded, and Singer's theorem that the bundle of gauge fields over the orbit space admits no continuous global section are stated here and not proved; all three belong in Appendix A.

106.5 The lattice

106.5.1 Wilson's lattice gauge theory

Everything so far has been perturbative, and Proposition 106.31 showed that perturbation theory cannot be the definition of the theory. Wilson supplied a definition [1537]: replace Euclidean spacetime by a finite hypercubic lattice of spacing a, so that the functional integral becomes an ordinary integral of finite dimension, and choose the variables so that gauge invariance is exact at finite a rather than only in the limit.

Definition 106.66 (Link variables and the plaquette)**.** Let the sites be $x = an$, $n \in \mathbb{Z}^4$. The dynamical variable on the link from n to $n + \hat{\mu}$ is the group element

$$U_\mu(n) = \exp\!\left[\mathrm{i}a\, \mathcal{A}^a_\mu \left(n + \tfrac{1}{2}\hat{\mu}\right) T^a\right] \in G\,, \tag{106.79}$$

dimensionless because $\mathcal{A}$ carries $1/\mathrm{m}$ by Equation (106.63) — the same product $\mathcal{A} \cdot \mathrm{d}x$ that appeared as the Aharonov–Bohm phase in Equation (106.13). The elementary closed loop, the *plaquette*, is

$$U_p = U_\mu(n)\, U_\nu(n+\hat{\mu})\, U^\dagger_\mu(n+\hat{\nu})\, U^\dagger_\nu(n)\,. \tag{106.80}$$

Proposition 106.67 (Exact gauge invariance at finite spacing)**.** *Under* $U_\mu(n) \mapsto \Omega(n) U_\mu(n) \Omega^\dagger(n+\hat{\mu})$ *with* $\Omega(n) \in G$ *arbitrary at each site, the trace of any closed loop of links is invariant. In particular* $\operatorname{tr} U_p$ *is invariant.*

Rests on Definition 106.66 and Equation (106.80).

Derivation. *Derives Proposition 106.67.*

Each link carries Ω at its start and $\Omega^\dagger$ at its end, so in a product of links forming a path the factors at every interior site cancel, $\Omega^\dagger(m)\Omega(m) = 1$. For a closed path the two remaining factors sit at the same site and are removed by the cyclicity of the trace. □

This is the decisive structural point. The gauge field lives in the *group*, not in the algebra, and gauge invariance is therefore an exact symmetry of a finite-dimensional integral rather than a property that survives only in the continuum. There is no gauge fixing, hence no Faddeev–Popov determinant, no ghosts and — as Section 106.4.4 noted — no Gribov problem: the volume of the gauge group is $(\mathrm{vol}\, G)^{\text{sites}}$, a finite constant.

Theorem 106.68 (The Wilson action and its continuum limit)**.** *Let*

$$S_W = \beta \sum_p \left(1 - \frac{1}{N} \Re\, \mathrm{tr}\, U_p\right), \qquad \beta = \frac{2N}{\hat{g}^2}, \tag{106.81}$$

the sum running over all plaquettes of the lattice, N the dimension of the fundamental representation, and β dimensionless. Then as $a \to 0$

$$S_W \longrightarrow \frac{1}{4\hat{g}^2} \int \mathrm{d}^4 x_E \; \mathcal{F}^a_{\mu\nu} \mathcal{F}^a_{\mu\nu} = \frac{S_E}{\hbar}, \tag{106.82}$$

with S_E the Euclidean Yang–Mills action of Equation (106.65). For SU(3)*, $\beta = 6/\hat{g}^2 = 6/(4\pi\alpha_s)$. This β is the lattice coupling and not the inverse temperature of Equation (106.27); under the dictionary of Table 106.1 it does play the role of an inverse temperature for the four-dimensional statistical system, which is why strong coupling is that system's high-temperature regime.*

Rests on Equations (106.64), (106.65) and (106.80).

Derivation. *Derives Theorem 106.68.*

Multiplying the four exponentials of Equation (106.80) and applying the Baker–Campbell–Hausdorff formula, the leading term of the exponent is the discrete curl of $\mathcal{A}$, so $U_p = \exp\!\left[\mathrm{i} a^2 \mathcal{F}_{\mu\nu} + O(a^3)\right]$ with $\mathcal{F}_{\mu\nu} = \mathcal{F}^a_{\mu\nu} T^a$ evaluated at the centre of the plaquette — this is the finite version of Equation (106.64), and the commutator term of Equation (106.64) arises here from the non-commutativity of the four link matrices. Expanding,

$$1 - \frac{1}{N} \Re\, \mathrm{tr}\, U_p = \frac{1}{2N}\, \mathrm{tr}\Big(a^2 \mathcal{F}_{\mu\nu}\Big)^2 + O(a^6) = \frac{a^4}{4N}\, \mathcal{F}^a_{\mu\nu} \mathcal{F}^a_{\mu\nu} + O(a^6),$$

with no sum over μ, ν, using $\mathrm{tr}(T^a T^b) = \tfrac{1}{2}\delta^{ab}$. The sum over plaquettes is a sum over sites and over $\mu < \nu$, which becomes $a^{-4} \int \mathrm{d}^4 x_E \, \tfrac{1}{2} \sum_{\mu\nu}$; hence $S_W \to (\beta/8N) \int \mathrm{d}^4 x_E \, \mathcal{F}^a_{\mu\nu} \mathcal{F}^a_{\mu\nu}$, and matching to Equation (106.82) gives $\beta = 2N/\hat{g}^2$. □

Remark 106.69 (A regulator, and a critical continuum limit)*.* Momenta on the lattice are bounded, $|p_\mu| \le \pi\hbar/a$, so the theory is finite: the spacing is an ultraviolet cutoff and no divergence occurs at any order. Recovering continuum physics requires the physical correlation length to be large in lattice units, $\xi/a \to \infty$, and by Equation (106.25) $\xi = \hbar/mc$, so $a \to 0$ at fixed physical mass means $mac/\hbar \to 0$: in the language of Table 106.1, the continuum limit is a second-order critical point of the four-dimensional statistical system defined by Equation (106.81). Asymptotic freedom says where that point is — at $\hat{g} \to 0$, i.e. $\beta \to \infty$ — and the rate of approach is fixed by the beta function of Chapter 108, which is what makes the extrapolation controlled rather than a guess.

Confinement as an area law

Definition 106.70 (Wilson loop)**.** For a closed contour C of links, $W(C) = \frac{1}{N}\,\langle \operatorname{tr} \prod_{\ell \in C} U_\ell \rangle$, the continuum form being the trace of the path-ordered exponential $\frac{1}{N}\,\big\langle \operatorname{tr} P \exp\!\big(\mathrm{i} \oint_C \mathcal{A}_\mu \mathrm{d}x^\mu\big)\big\rangle$.

Proposition 106.71 (The Wilson loop measures the static potential)**.** *For a rectangular contour of spatial extent R and Euclidean time extent τ, with $\tau \gg R/c$,*

$$W(R,\tau) \;\sim\; \exp\!\left[-\frac{V(R)\,\tau}{\hbar}\right], \tag{106.83}$$

where $V(R)$ is the energy of a static quark–antiquark pair at separation R. An area law*, $W \sim \exp[-\sigma R c\tau/(\hbar c)]$, is therefore equivalent to a linearly rising potential $V(R) = \sigma R$, i.e. to confinement; a* perimeter law *corresponds to a potential approaching a constant, i.e. to a screened, deconfined phase.*

Rests on Definition 106.70 and Proposition 106.17.

Derivation. *Derives Proposition 106.71.*

The two timelike sides of the rectangle are the parallel transporters that carry a static colour source forward in Euclidean time; the two spacelike sides create and annihilate the pair. By Proposition 106.17 the correlator of a creation and an annihilation operator separated by Euclidean time τ decays as $\mathrm{e}^{-E_0\tau/\hbar}$, with E_0 the lowest energy in the sector created — here the energy of the static pair, $V(R)$, measured from the vacuum. Substituting $A = Rc\tau$ for the enclosed area converts an exponential in the area into an exponential in $V\tau$ with $V = \sigma R$. Dimensions: σ carries N, so $\sigma A/(\hbar c)$ is dimensionless. □

Theorem 106.72 (Confinement at strong coupling)**.** *For $\beta \ll 1$ the Wilson loop of a planar rectangle obeys an area law. For* SU(N) *with $N \geq 3$,*

$$W(C) = \left(\frac{\beta}{2N^2}\right)^{A/a^2} \big[1 + O(\beta)\big], \qquad \sigma = \frac{\hbar c}{a^2}\,\ln\!\frac{2N^2}{\beta}, \tag{106.84}$$

A being the minimal area spanned by C. For SU(2) *every plaquette contributes twice as much,*

$$W(C) = \left(\frac{\beta}{4}\right)^{A/a^2} \big[1 + O(\beta)\big], \qquad \sigma = \frac{\hbar c}{a^2}\,\ln\!\frac{4}{\beta}, \tag{106.85}$$

and not the $\beta/2N^2 = \beta/8$ that Equation (106.84) would give: the fundamental representation of SU(2) *is pseudo-real, so U and $U^\dagger$ are not independent and the two orientations of a plaquette supply the same term rather than two different ones. The distinction is not academic here —* SU(2) *is the group of the measurement quoted in Section 106.5.2.*

Rests on Equations (106.81) and (106.83) and Theorem 5.153.

Derivation. *Derives Theorem 106.72.*

Expand e^{-S_W} in powers of β: for each plaquette, $\exp\!\big[(\beta/N)\Re\operatorname{tr} U_p\big] = 1 + \frac{\beta}{2N}\big(\operatorname{tr} U_p + \operatorname{tr} U_p^\dagger\big) + O(\beta^2)$. The integration is over independent Haar measures on each link, and the two identities that are needed are

$$\int \mathrm{d}U\; U_{ij} = 0\,, \qquad \int \mathrm{d}U\; U_{ij}\big(U^\dagger\big)_{kl} = \frac{1}{N}\delta_{il}\delta_{jk}\,,$$

the first because the integrand transforms in a non-trivial representation and the Haar measure is invariant, the second by Schur's first lemma (Theorem 5.153) applied to the invariant tensor. Every link of the contour must therefore be paired with a link of opposite orientation supplied by some plaquette; the cheapest way to do this is to tile the minimal surface spanned by C with A/a^2 plaquettes, each bringing $\beta/2N$ from the expansion and

each internal link contributing $1/N$ from the second identity. The single-plaquette case is exact and does the counting transparently:

$$\left\langle \tfrac{1}{N}\operatorname{tr} U_p \right\rangle = \frac{\beta}{2N}\cdot\frac{1}{N}\int \mathrm{d}U\ \operatorname{tr} U\ \operatorname{tr} U^{\dagger} + O(\beta^2) = \frac{\beta}{2N^2} + O(\beta^2)\,,$$

since $\int \mathrm{d}U\ \operatorname{tr} U \operatorname{tr} U^{\dagger} = \sum_{ij}\delta_{ij}\delta_{ij}/N = 1$. Iterating over the tiling gives Equation (106.84), and comparing with Equation (106.83) gives σ.

That counting used $\int \mathrm{d}U\, U_{ij}U_{kl} = 0$, i.e. $\int \mathrm{d}U\,(\operatorname{tr} U)^2 = 0$, which holds only when $\boldsymbol{N}\otimes\boldsymbol{N}$ contains no singlet. For $N \geq 3$ it splits into the symmetric and antisymmetric two-index representations and neither is trivial, so the identity holds. For SU(2), $\mathbf{2}\otimes\mathbf{2} = \mathbf{3}\oplus\mathbf{1}$ does contain the singlet, $\operatorname{tr} U$ is real, and orthogonality of characters gives $\int \mathrm{d}U\,(\operatorname{tr} U)^2 = 1$. The plaquette factor is then $(\beta/2N)(\operatorname{tr} U_p + \operatorname{tr} U_p^{\dagger}) = \frac{\beta}{2}\operatorname{tr} U_p$ and

$$\left\langle \tfrac{1}{2}\operatorname{tr} U_p \right\rangle = \frac{1}{2}\cdot\frac{\beta}{2}\int \mathrm{d}U\ (\operatorname{tr} U)^2 + O(\beta^2) = \frac{\beta}{4} + O(\beta^2)\,,$$

which is Equation (106.85): twice the value the generic formula assigns, hence $\ln(4/\beta)$ in place of $\ln(8/\beta)$ in the string tension. □

Remark 106.73 (What the strong-coupling result does and does not prove). Theorem 106.72 is a genuine derivation of confinement — of an area law with a computable string tension — but for the lattice theory at large $\hat{g}$, not for the continuum theory. The continuum limit lives at $\beta \to \infty$ (Remark 106.69), which is the opposite end, and the expansion Equation (106.84) says nothing about whether the area law survives to it. Worse, the expansion converges for compact U(1) too, and there confinement is known *not* to survive: the theory has a phase transition at intermediate β and its weak-coupling phase is ordinary massless electrodynamics. Whether SU(3) has such a transition — and the evidence is that it does not — is a question that only a numerical computation covering the whole range of β can answer, which is Section 106.5.2. The confinement of quarks remains, in the strict sense, unproved; what exists is a definition of the theory in which the question is well posed, and overwhelming numerical evidence about the answer.

For orientation, the customary phenomenological value of the string tension, obtained by matching the linear potential to the slope of the observed Regge trajectories, is $\sqrt{\sigma\hbar c} \approx 0.44\,\mathrm{GeV}$, i.e.

$$\sigma = \frac{(0.44\,\mathrm{GeV})^2}{\hbar c} \approx 1.6\times 10^{5}\,\mathrm{N}\,, \tag{106.86}$$

about the weight of sixteen tonnes, independent of separation. That number is the reason no free quark has been seen: separating a pair by one femtometre costs $\sigma\times 1\,\mathrm{fm} = 1.6\times10^{-10}\,\mathrm{J} \approx 1\,\mathrm{GeV}$, and by one metre would cost 10^{15} times more — long before which it is cheaper to create a quark–antiquark pair out of the vacuum and break the string.

Fermions on a lattice

Proposition 106.74 (Naive discretization doubles the species sixteen times)**.** *Replacing $\gamma^{\mu}\partial_{\mu}$ by the symmetric difference $\gamma^{\mu}[\psi(n+\hat{\mu}) - \psi(n-\hat{\mu})]/2a$ gives a momentum-space operator whose massless part is*

$$\widetilde{D}(p) = \frac{\mathrm{i}}{a}\sum_{\mu}\gamma^{\mu}\sin\!\left(\frac{ap_{\mu}}{\hbar}\right)\,, \tag{106.87}$$

which vanishes not only at $p_{\mu} = 0$ but at every corner of the Brillouin zone with $ap_{\mu}/\hbar \in \{0,\pi\}$: $2^4 = 16$ species where one was intended.

Rests on Definition 106.66 and Equation (100.5).

Derivation. *Derives Proposition 106.74.*

Fourier transform the difference operator: the shift $\psi(n \pm \hat{\mu})$ becomes $\mathrm{e}^{\pm \mathrm{i} a p_\mu/\hbar}\psi$, so the symmetric difference becomes $\mathrm{i}\sin(a p_\mu/\hbar)/a$, giving Equation (106.87). The propagator is $\widetilde{D}^{-1}$, and a pole occurs wherever $\widetilde{D}$ vanishes; $\sin(a p_\mu/\hbar) = 0$ has two solutions in the Brillouin zone $-\pi < a p_\mu/\hbar \le \pi$ for each of the four directions, hence 2^4 poles, each of which is a physical particle in the continuum limit. □

Theorem 106.75 (Nielsen–Ninomiya)**.** *No lattice Dirac operator can simultaneously (i) be local, (ii) possess the continuum chiral symmetry $\{D, \gamma^5\} = 0$, (iii) have the correct continuum limit, and (iv) describe a single species. One of the four must be given up [1673].*

Rests on Proposition 106.74.

The choices actually made are: Wilson's, which adds $-\frac{r a}{2}\partial^2$ to the operator, lifting the fifteen doublers to masses of order $\hbar/(ac)$ at the price of breaking chiral symmetry explicitly at finite a — the symmetry is recovered in the continuum limit, but only after an additive renormalization of the quark mass; and the staggered construction, which distributes the spinor components over the corners of a hypercube, retaining a remnant of chiral symmetry and reducing the doubling from sixteen to four. The computation of Section 106.5.2 uses the Wilson operator with smeared links [2122]. That a theorem of this kind exists at all is the sharpest illustration of the point made after Proposition 106.9: a discretization is a definition, and a symmetry of the classical Lagrangian need not survive it. Section 106.7.3 shows that for the axial symmetry the failure is not an artefact of any discretization but a property of the continuum theory itself, with a measured consequence.

Derivation pending #165

The Nielsen–Ninomiya theorem: the fermion doubling of the naive operator is derived above, but the no-go theorem itself — that no local, chirally symmetric, doubler-free operator exists, proved by counting the zeros of the energy–momentum relation as a topological invariant of the Brillouin torus — is stated and not proved here. It belongs in Appendix A.

106.5.2 Monte Carlo evidence

Why sampling, and why it works A lattice of $64^3 \times 128$ sites carries $33\,554\,432$ sites, $134\,217\,728$ links and, for SU(3) with eight parameters per link, about 1.1×10^9 real integration variables. No deterministic quadrature can touch an integral of that dimension: a product rule with k points per axis costs k^d evaluations. Monte Carlo integration does not degrade with dimension at all.

Proposition 106.76 (Importance sampling)**.** *Let configurations $\{U^{(i)}\}_{i=1}^{M}$ be drawn with probability proportional to $\mathrm{e}^{-S_E[U]/\hbar}$. Then*

$$\langle \mathcal{O} \rangle = \frac{\int \mathcal{D}U\; \mathcal{O}[U]\, \mathrm{e}^{-S_E/\hbar}}{\int \mathcal{D}U\; \mathrm{e}^{-S_E/\hbar}} = \lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} \mathcal{O}\big[U^{(i)}\big]\,, \tag{106.88}$$

with a statistical error falling as $M^{-1/2}$ independently of the dimension of the integral, by the central limit theorem.

Rests on Proposition 106.13 and Theorem 106.68.

Derivation. *Derives Proposition 106.76.*

The estimator is the sample mean of a random variable whose expectation is $\langle \mathcal{O} \rangle$ by construction, so it is unbiased; its variance is $\mathrm{Var}(\mathcal{O})/M$, and the central limit theorem gives the distribution of the error. Nothing in either statement refers to the dimension. What

replaces the dimension in the cost is the *autocorrelation* of the algorithm that generates the sample, since successive configurations are not independent. □

Two facts make Equation (106.88) legitimate, and both were established earlier in this chapter rather than assumed. Proposition 106.13 makes the weight real and positive, so that "probability proportional to $\mathrm{e}^{-S_E/\hbar}$" means something; and Remark 106.40 explains why the fermion determinant is included as $\det^2$ for the degenerate light pair, which keeps it positive. Where positivity fails — at non-zero baryon chemical potential — the method fails with it (Remark 106.15).

Confinement and asymptotic scaling Creutz's computation [1651] was the first to show that the lattice theory and the continuum theory are the same theory. He extracted the string tension of SU(2) gauge theory from ratios of Wilson loops,

$$\chi(I,J) = -\ln \frac{W(I,J)\,W(I-1,J-1)}{W(I-1,J)\,W(I,J-1)}\,, \tag{106.89}$$

a combination in which the perimeter and corner contributions cancel and only the area term survives, so that $\chi \to \sigma a^2/(\hbar c)$ for large loops. Measuring χ over a range of β, he found that it followed the asymptotic-freedom scaling law

$$\sqrt{\chi} \;\propto\; \exp\!\left[-\frac{4\pi^2\beta}{N b_0}\right], \qquad b_0 = \frac{11N}{3} - \frac{2n_f}{3}\,, \tag{106.90}$$

in which b_0 is the one-loop beta-function coefficient in the convention $\mu\,\mathrm{d}\hat{g}/\mathrm{d}\mu = -b_0\hat{g}^3/16\pi^2$ — $b_0 = 22/3$ for the pure SU(2) theory Creutz simulated, so that the exponent is $-3\pi^2\beta/11$. The exponent follows from $a\Lambda = \exp[-8\pi^2/(b_0\hat{g}^2)]$ with $\hat{g}^2 = 2N/\beta$, and the scaling law itself is the prediction of the renormalization group of Chapter 108 — that is, that the dimensionless lattice quantity ran with β exactly as a physical quantity held fixed in units of a continuum scale must. Confinement at strong coupling was already known from Theorem 106.72; what Creutz showed is that it persists into the region where the continuum limit is taken.

The hadron spectrum from three inputs

Phenomenon 106.77 (The light hadron spectrum is a prediction of QCD)**.** With the three free parameters of the light-quark sector fixed by three measured masses — those of the pion, the kaon and one baryon, which set the light-quark mass, the strange-quark mass and the overall scale — the Euclidean lattice path integral reproduces the masses of the remaining light hadrons, including the nucleon, to a few percent [2122]. Nothing else is put in: no potential, no constituent masses, no fitted form factors.

Rests on Equation (106.24)*, Corollary 106.49, and Proposition 106.17.*

Derivation. *Derives Phenomenon 106.77.*

The procedure is Equation (106.24) applied to a lattice correlator. Choose an interpolating operator $\mathcal{O}$ with the quantum numbers of the hadron — for the nucleon, a colour-singlet product of three quark fields; Corollary 106.49 guarantees that the choice does not affect the answer. Measure

$$C(\tau) = \left\langle \mathcal{O}(\tau)\,\mathcal{O}^\dagger(0)\right\rangle \underset{\tau\to\infty}{\longrightarrow} |\langle 0|\,\mathcal{O}\,|h\rangle|^2\;\mathrm{e}^{-m_h c^2\tau/\hbar}\,,$$

by Proposition 106.17, and read the mass off the exponential decay rate. Repeat at several lattice spacings, volumes and quark masses, and extrapolate.

The counting of inputs is what makes this a prediction rather than a fit. The Lagrangian of the light sector contains exactly three parameters that are not fixed by symmetry: $\hat{g}^2$,

the degenerate up–down mass, and the strange mass. Measuring m_π, m_K and one baryon mass fixes all three (the coupling enters only through the overall scale, by dimensional transmutation — Chapter 108). Every other hadron mass is then output. The masses to be reproduced include, from the shipped Particle Data Group compilation [2427], $m_p c^2 = 938.272\,\mathrm{MeV}$, $m_n c^2 = 939.565\,\mathrm{MeV}$, $m_\Lambda c^2 = 1115.683\,\mathrm{MeV}$ and $m_\Delta c^2 = (1232.0 \pm 2.0)\,\mathrm{MeV}$, together with the Σ, Ξ, Σ^*, Ξ^* and Ω; the computed values agree with all of them within the combined statistical and systematic uncertainties, which are at the level of a few percent [2122]. □

This is the strongest single piece of evidence that the quantization scheme of this chapter describes Nature and not merely a formal expansion. The perturbative successes of Chapter 100 test the theory where the coupling is small and the series is asymptotic; here the coupling is of order unity, no series is involved, and the answer still comes out right. It is also the answer to Remark 106.73: no phase transition intervenes between the strong-coupling region and the continuum limit, because if one did the spectrum would not come out.

Remark 106.78 (The systematics, and how they are quantified). A lattice result is not a measurement and its error budget is not statistical alone. Three systematic effects dominate, and each is controlled by computing at several values of the corresponding parameter and extrapolating rather than by estimating.

Finite spacing. The discretization error vanishes as a power of a — $O(a)$ for the unimproved Wilson operator, $O(a^2)$ after improvement. Computations are repeated at three or more spacings, in the range $a \approx 0.065 - -0.125\,\mathrm{fm}$ for [2122], and extrapolated to $a = 0$.

Finite volume. A box of side L with periodic boundaries distorts a hadron by the exchange of the lightest particle around the box, an effect falling as $\mathrm{e}^{-m_\pi c L/\hbar}$. The working rule is $m_\pi c L/\hbar \gtrsim 4$, i.e. $L \gtrsim 4\hbar/(m_\pi c)$, four reduced Compton wavelengths of the pion. At the physical charged-pion mass $m_\pi c^2 = 139.570\,39\,\mathrm{MeV}$ [2427] that wavelength is $\hbar c/(m_\pi c^2) = 1.414\,\mathrm{fm}$, so the rule demands $L \gtrsim 5.7\,\mathrm{fm}$ — a demanding box, which is why the requirement is usually met at a pion mass above the physical one, where the same rule costs less: at $m_\pi c^2 = 190\,\mathrm{MeV}$ it asks only for $L \gtrsim 4.2\,\mathrm{fm}$. The dependence is checked by running at two volumes.

Quark mass. Simulating at the physical light-quark mass is expensive because the algorithms slow down as the pion becomes light, so computations are done at larger masses and extrapolated using the functional form supplied by chiral perturbation theory. This was historically the largest uncertainty, and it is the one that has shrunk most as machines have grown.

Note that these are all controlled extrapolations in parameters the computation itself can vary, which is what distinguishes them from the uncontrolled quenched approximation of Remark 106.40. It is that difference, and not raw computer power, that turned lattice QCD from an illustration into a source of numbers with error bars.

106.6 Topology of the gauge vacuum

106.6.1 Instantons

The Euclidean action of Equation (106.82) has non-trivial finite-action stationary points, found by Belavin, Polyakov, Schwartz and Tyupkin [1539]. They contribute to the path integral factors of the form $\mathrm{e}^{-S_E/\hbar}$ which are invisible to every order of perturbation theory, and they are the mechanism by which the vacuum acquires structure.

Definition 106.79 (Topological charge)**.** For a Euclidean gauge field of finite action define the dual $\widetilde{\mathcal{F}}^a_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma}\mathcal{F}^a_{\rho\sigma}$ and

$$\nu := \frac{1}{32\pi^2}\int \mathrm{d}^4x_E\, \mathcal{F}^a_{\mu\nu}\widetilde{\mathcal{F}}^a_{\mu\nu} = \frac{1}{32\pi^2}\int \mathrm{d}^4x_E\, \epsilon_{\mu\nu\rho\sigma}\,\mathrm{tr}(\mathcal{F}_{\mu\nu}\mathcal{F}_{\rho\sigma})\ , \tag{106.91}$$

dimensionless, since $\mathcal{F}$ carries $1/\mathrm{m}^2$ by Equation (106.64). This is the quantity denoted n in Equation (105.109), and the corresponding density $q(x) = (c/32\pi^2)\epsilon_{\mu\nu\rho\sigma}\,\mathrm{tr}(\mathcal{F}_{\mu\nu}\mathcal{F}_{\rho\sigma})$, of dimension $1/(\mathrm{m}^3\,\mathrm{s})$, is the q of Definition 105.43.

Theorem 106.80 (Bogomolny bound and self-duality)**.** *For any Euclidean configuration,*

$$S_E \geq \frac{8\pi^2\hbar}{\hat{g}^2}\,|\nu|\ , \tag{106.92}$$

with equality if and only if the field is self-dual ($\mathcal{F} = \widetilde{\mathcal{F}}$, for $\nu > 0$) or anti-self-dual ($\mathcal{F} = -\widetilde{\mathcal{F}}$, for $\nu < 0$). A configuration saturating the bound with $\nu = 1$ is an instanton*, and its contribution to the path integral is weighted by*

$$\mathrm{e}^{-S_E/\hbar} = \exp\!\left(-\frac{8\pi^2}{\hat{g}^2}\right) = \exp\!\left(-\frac{2\pi}{\alpha_s}\right)\ , \tag{106.93}$$

which at $\alpha_s = 0.1180$ (Equation (102.47)) is 8×10^{-24} — a number with no expansion in powers of α_s at all.

Rests on Equations (106.82) *and* (106.91)*.*

Derivation. *Derives Theorem 106.80.*

The Euclidean action is $S_E = (\hbar/4\hat{g}^2)\int \mathrm{d}^4x_E\,\mathcal{F}^a_{\mu\nu}\mathcal{F}^a_{\mu\nu}$ by Equation (106.82), and the integrand is a sum of squares, hence non-negative. Consider

$$0 \leq \int \mathrm{d}^4x_E\ \left(\mathcal{F}^a_{\mu\nu} \mp \widetilde{\mathcal{F}}^a_{\mu\nu}\right)^2 = 2\int \mathrm{d}^4x_E\,\mathcal{F}^a_{\mu\nu}\mathcal{F}^a_{\mu\nu} \mp 2\int \mathrm{d}^4x_E\,\mathcal{F}^a_{\mu\nu}\widetilde{\mathcal{F}}^a_{\mu\nu}\,,$$

using $\widetilde{\mathcal{F}}\widetilde{\mathcal{F}} = \mathcal{F}\mathcal{F}$, which follows from $\epsilon_{\mu\nu\rho\sigma}\epsilon_{\mu\nu\alpha\beta} = 2(\delta_{\rho\alpha}\delta_{\sigma\beta} - \delta_{\rho\beta}\delta_{\sigma\alpha})$ in four Euclidean dimensions. Hence $\int \mathcal{F}\mathcal{F} \geq \left|\int \mathcal{F}\widetilde{\mathcal{F}}\right| = 32\pi^2\,|\nu|$ by Equation (106.91), with equality iff $\mathcal{F} = \pm\widetilde{\mathcal{F}}$. Multiplying by $\hbar/4\hat{g}^2$ gives Equation (106.92). A self-dual field automatically solves the Euclidean field equations, since $D_\mu\mathcal{F}_{\mu\nu} = D_\mu\widetilde{\mathcal{F}}_{\mu\nu} = 0$ by the Bianchi identity; self-duality, a first-order condition, therefore replaces the second-order equations of motion.

Numerically, $2\pi/0.1180 = 53.24$ and $\mathrm{e}^{-53.24} = 8\times10^{-24}$. □

Proposition 106.81 (The topological charge is an integer)**.** *Finiteness of the action forces $\mathcal{F}\to 0$ at infinity, hence $\mathcal{A}_\mu \to -\mathrm{i}\,g^{-1}\partial_\mu g$ for some map $g : S^3 \to G$ on the sphere at infinity. Then*

$$\nu = \frac{1}{24\pi^2}\int_{S^3} \mathrm{d}^3S_\mu\ \epsilon_{\mu\nu\rho\sigma}\ \mathrm{tr}\!\left(g^{-1}\partial_\nu g\ g^{-1}\partial_\rho g\ g^{-1}\partial_\sigma g\right)\ , \tag{106.94}$$

which is the degree of g as a map $S^3\to S^3$ — for $G = \mathrm{SU}(2)$, whose manifold is S^3 — and is therefore an integer.

Rests on Equations (14.155) *and* (106.91) *and Theorem 14.111.*

Derivation. *Derives Proposition 106.81.*

The integrand of Equation (106.91) is a total divergence, $\epsilon\,\mathrm{tr}(\mathcal{F}\mathcal{F}) = \partial_\mu K_\mu$ with the Chern–Simons current $K_\mu = 2\epsilon_{\mu\nu\rho\sigma}\,\mathrm{tr}(\mathcal{A}_\nu\partial_\rho\mathcal{A}_\sigma + \frac{2\mathrm{i}}{3}\mathcal{A}_\nu\mathcal{A}_\rho\mathcal{A}_\sigma)$; this is the transgression form of the second Chern class, and it is Equation (14.155) of Theorem 14.111 specialized to $r = 2$. The volume integral is therefore a surface integral over the sphere at infinity, where $\mathcal{A}$ is pure gauge; substituting $\mathcal{A}_\mu = -\mathrm{i}g^{-1}\partial_\mu g$ makes the first term of K_μ combine with the second into Equation (106.94).

That the resulting number is an integer is the degree theorem. Let y be a regular value of the smooth map $g : S^3 \to S^3$; its preimage is a finite set of points, and Equation (106.94) evaluates to the sum over those points of the sign of the Jacobian determinant of g, i.e. to the number of times g covers the target counted with orientation. That count is an integer, and it is unchanged under continuous deformation of g, since the preimages can only be created or destroyed in pairs of opposite sign. The map $g(x) = (x_4 + \mathrm{i}\,\boldsymbol{x}\cdot\boldsymbol{\tau})/|x|$, with $\boldsymbol{\tau}$ the Pauli matrices, covers $\mathrm{SU}(2) \cong S^3$ exactly once and has $\nu = 1$: it is the identity map, whose Jacobian is everywhere positive. □

Remark 106.82 (The dilute-gas approximation and its failure). Instantons come in a family: the classical equations are scale invariant, so a solution of size ρ costs the same action for every ρ. Integrating over the collective coordinates — position, size and global colour orientation — requires the determinant of the fluctuation operator about the instanton, computed by 't Hooft [1568]. The result is a density of instantons per unit four-volume of the form

$$\mathrm{d}n \;\propto\; \frac{\mathrm{d}\rho}{\rho^5}\,\left(\frac{1}{\hat{g}^2(\rho)}\right)^{2N} \exp\!\left(-\frac{8\pi^2}{\hat{g}^2(\rho)}\right),$$

in which the running of $\hat{g}$ with the scale $1/\rho$ turns the exponential into a power, ρ^{b_0} with $b_0 = 11N/3 - 2n_f/3$ the one-loop beta-function coefficient in the same convention as Equation (106.90), so that $8\pi^2/\hat{g}^2(\rho) = b_0 \ln(1/\rho\Lambda)$. For $\mathrm{SU}(3)$ with three light flavours $b_0 = 9$, so the density behaves as $\rho^4\mathrm{d}\rho$: the integral is dominated by *large* instantons, where the coupling is strong and the semiclassical approximation that produced the formula has failed. The honest conclusion is that the dilute-gas approximation is not a controlled calculation in QCD. What survives is the qualitative structure — the vacuum has topological sectors, they are populated, and the population is a non-perturbative effect — together with the one place where the counting is protected by an index theorem, which is Section 106.7.3. The quantitative statements about the topological susceptibility of the QCD vacuum come from the lattice, not from instanton calculus.

The physical use of all this in Chapter 102 is the mass of the η'. The nine light pseudoscalars would be the Goldstone bosons of a spontaneously broken $\mathrm{U}(3)_L \times \mathrm{U}(3)_R$, and eight of them — the pions, kaons and η — are light for that reason. The ninth is not: the singlet axial symmetry is broken by the anomaly of Section 106.7.1, and the topological configurations of this subsection are what make the anomaly have consequences rather than being a total derivative one can discard. That is the resolution of what was called the $\mathrm{U}(1)_A$ problem, and it is the reason the η' is heavier than the nucleon-scale expectation for a Goldstone boson.

Derivation pending #166

The instanton fluctuation determinant: the Bogomolny bound, the integrality of the winding number and the exponential weight are derived above, but the explicit self-dual solution of Belavin, Polyakov, Schwartz and Tyupkin and the one-loop determinant about it, which supply the measure over collective coordinates quoted in the dilute-gas remark, are not computed here. They belong in Appendix A.

106.6.2 Theta vacua

Proposition 106.83 (The vacuum is a superposition of winding sectors)**.** *Let $|n\rangle$ denote the state built on the classical vacuum of Chern–Simons number $n \in \mathbb{Z}$, and let T be the operator implementing a large gauge transformation, $T\,|n\rangle = |n+1\rangle$. Since T commutes with the Hamiltonian and with all local observables, the energy eigenstates are simultaneous*

eigenstates of T:

$$|\theta\rangle = \sum_{n\in\mathbb{Z}} \mathrm{e}^{\mathrm{i}n\theta}\,|n\rangle\ , \qquad T\,|\theta\rangle = \mathrm{e}^{-\mathrm{i}\theta}\,|\theta\rangle\ , \tag{106.95}$$

with $\theta \in [0, 2\pi)$ a superselection parameter: no local operator connects different θ [1573, 1582].

Rests on Proposition 106.81.

Derivation. *Derives Proposition 106.83.*

Gauge transformations that tend to the identity at infinity are generated by the Gauss constraint and act trivially on physical states. Those that tend to a non-trivial element — the *large* transformations, classified by Proposition 106.81 — are symmetries of the theory but are not required to act trivially, exactly as a translation by a lattice vector in a crystal is a symmetry that acts non-trivially on Bloch states. T is unitary and commutes with H, so H and T may be diagonalized together; the eigenvalues of a unitary operator with $T\,|n\rangle = |n+1\rangle$ are phases $\mathrm{e}^{-\mathrm{i}\theta}$, and the corresponding eigenvector is Equation (106.95). Because a local operator changes n by a finite amount, its matrix elements between $|\theta\rangle$ and $|\theta'\rangle$ are proportional to $\delta(\theta-\theta')$: the θ sectors are superselected and θ is a parameter of the theory, not a state to be chosen. □

Corollary 106.84 (The theta term)**.** *Working in the θ vacuum is equivalent to adding to the Lagrangian density*

$$\mathcal{L}_\theta = \theta\hbar\, q(x) = \frac{\theta\hbar c}{32\pi^2}\,\epsilon^{\mu\nu\rho\sigma}\ \mathrm{tr}(\mathcal{F}_{\mu\nu}\mathcal{F}_{\rho\sigma})\ , \tag{106.96}$$

which is Equation (105.110), and which contributes the phase $\mathrm{e}^{\mathrm{i}\theta\nu}$ to the weight of a configuration of winding number ν.

Rests on Equation (106.95) and Definition 106.79.

Derivation. *Derives Corollary 106.84.*

Summing over sectors with the weights of Equation (106.95) inserts $\mathrm{e}^{\mathrm{i}n\theta}$ into the path integral for each topological class n. By Definition 106.79 that class is labelled by $\nu = \int \mathrm{d}t\,\mathrm{d}^3x\ q$, so the insertion is $\exp(\mathrm{i}\theta\int \mathrm{d}t\,\mathrm{d}^3x\,q) = \exp(\frac{\mathrm{i}}{\hbar}\cdot\frac{1}{c}\int \mathrm{d}^4x\ \theta\hbar q)$, which is the addition of Equation (106.96) to $\mathcal{L}$. That $\mathcal{L}_\theta$ has the dimension of an energy density follows from $[q] = 1/(\mathrm{m}^3\,\mathrm{s})$ and $[\hbar] = \mathrm{J\,s}$. □

Three consequences follow, and each is taken up elsewhere in this part rather than repeated here.

Baryon number is not conserved. The baryon and lepton currents of the Standard Model are anomalous in the sense of Section 106.7.1, with the topological density of Equation (106.91) of the $\mathrm{SU}(2)_L$ field on the right-hand side. Integrating gives $\Delta B = \Delta L = n_f\nu$ with $n_f = 3$ generations, which is Equation (105.103); $B-L$ is conserved. At zero temperature the transition rate carries the factor Equation (106.93) with α_W in place of α_s and is utterly negligible, but above the electroweak scale the barrier — the sphaleron of Equation (105.104) — can be crossed thermally and the rate is unsuppressed [1569, 1758]. This is how the Standard Model satisfies the first of Sakharov's conditions, Theorem 105.41.

The theta term violates P and CP. That Equation (106.96) is P-odd and T-odd is Proposition 105.44; that only the combination $\bar\theta = \theta + \arg\det M_q$ is physical, because a chiral rotation of the quark fields shifts θ through the non-invariance of the measure, is Equation (105.111) — and the mechanism of that shift is precisely Section 106.7.3. The experimental bound from the neutron electric dipole moment, $\left|\bar\theta\right| \lesssim 10^{-10}$, is Equation (105.113) [2090, 2346]. The strong CP problem is the observation that a parameter which could have been anything is measured to be this small, and it is recorded as an open question in Section 105.7. The best-known proposed resolution promotes $\bar\theta$ to a dynamical field [1609,

1630]; the particle it predicts has not been observed, and this book records it as a proposal, not as physics.

The θ dependence is invisible in perturbation theory. q is a total derivative, so Equation (106.96) contributes nothing to any Feynman diagram; its effects exist only because the boundary term is non-zero on configurations of non-trivial topology. A parameter that is invisible perturbatively and physical non-perturbatively is the worst possible combination for anyone hoping to argue it away, and that is the whole difficulty.

106.7 Anomalies

106.7.1 The Adler–Bell–Jackiw anomaly

A classical conservation law

Proposition 106.85 (Classical conservation of the axial current)**.** *For a Dirac field of mass m coupled to an external gauge field, the axial current*

$$j_5^\mu = c\,\bar\psi\gamma^\mu\gamma^5\psi\,, \qquad [j_5^\mu] = 1/(\mathrm{m}^2\,\mathrm{s})\,, \tag{106.97}$$

satisfies, by the classical field equations,

$$\partial_\mu j_5^\mu = \frac{2\mathrm{i}mc^2}{\hbar}\,\bar\psi\gamma^5\psi\,, \tag{106.98}$$

which vanishes identically for $m = 0$: the axial charge of a massless Dirac field is conserved.
Rests on Equation (100.5)*.*

Derivation. *Derives Proposition 106.85.*

The Dirac equation in an external field is $(\mathrm{i}\hbar\gamma^\mu D_\mu - mc)\psi = 0$, and its conjugate reads $\mathrm{i}\hbar\,(D_\mu\bar\psi)\gamma^\mu + mc\,\bar\psi = 0$, the derivative acting on $\bar\psi$. Then

$$\partial_\mu(\bar\psi\gamma^\mu\gamma^5\psi) = (D_\mu\bar\psi)\gamma^\mu\gamma^5\psi + \bar\psi\gamma^\mu\gamma^5(D_\mu\psi) = \frac{\mathrm{i}mc}{\hbar}\bar\psi\gamma^5\psi + \frac{\mathrm{i}mc}{\hbar}\bar\psi\gamma^5\psi\,,$$

using $\{\gamma^\mu,\gamma^5\} = 0$ to move γ^5 past γ^μ in the second term, which flips the sign of the mass contribution and makes the two add rather than cancel. Multiplying by c gives Equation (106.98). The gauge field drops out because the covariant derivatives combine into an ordinary one in a neutral bilinear. □

By Noether's theorem [456] this conservation law is the statement that the massless action is invariant under the global chiral rotation

$$\psi \longmapsto \mathrm{e}^{\mathrm{i}\alpha\gamma^5}\psi\,, \qquad \bar\psi \longmapsto \bar\psi\,\mathrm{e}^{\mathrm{i}\alpha\gamma^5}\,, \tag{106.99}$$

the second transformation carrying $\mathrm{e}^{+\mathrm{i}\alpha\gamma^5}$ rather than its inverse because γ^5 anticommutes with the γ^0 hidden in $\bar\psi$. Everything above is classical. What Adler [1388] and Bell and Jackiw [1391] found is that Equation (106.98) is false in the quantum theory even at $m = 0$.

Where the failure comes from The one-loop amplitude with one axial and two vector vertices — the triangle — is linearly divergent, and a linearly divergent integral is not invariant under a shift of its integration variable.

Lemma 106.86 (A shift of a divergent integral leaves a surface term)**.** *For a smooth $f:\mathbb{R}^4\to\mathbb{C}$ and constant a,*

$$\int \mathrm{d}^4\ell\,\big[f(\ell+a) - f(\ell)\big] = \lim_{R\to\infty}\oint_{|\ell|=R} a_\mu\, f(\ell)\,\mathrm{d}S^\mu\,, \tag{106.100}$$

which vanishes when f falls faster than $|\ell|^{-3}$ and is a finite non-zero constant, linear in a, when f falls exactly as $|\ell|^{-3}$ — that is, when the integral is linearly divergent.
Rests on Theorem 7.101.

Derivation. *Derives Lemma 106.86.*

Expand $f(\ell + a) - f(\ell) = a_\mu \partial^\mu f + O(a^2)$ — the higher terms give integrals of higher divergences of f and are handled the same way — and apply the divergence theorem to the ball $|\ell| \le R$. The surface element grows as R^3, so the limit is zero, finite, or infinite according as f falls faster than, exactly as, or slower than R^{-3}. □

Theorem 106.87 (The axial anomaly)**.** *In the quantum theory of a massless Dirac field of charge Q in an external electromagnetic field, the axial current is not conserved:*

$$\partial_\mu j_5^\mu = \frac{Q^2 c}{16\pi^2 \hbar^2}\, \epsilon^{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma} = \frac{\alpha}{4\pi \mu_0 \hbar}\, \epsilon^{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma}\,, \tag{106.101}$$

the second form for $Q = e$, with $\epsilon^{0123} = +1$. Both sides carry $1/(\mathrm{m}^3\,\mathrm{s})$. The vector current $j^\mu = Qc\bar\psi\gamma^\mu\psi$ of Equation (100.6) remains exactly conserved.

Rests on Proposition 106.85 and Equation (106.99).

The coefficient is derived in Section 106.7.3, where the computation is short and its topological character is manifest. What the triangle diagram contributes is the *reason*: the anomaly is not a choice but an obstruction.

Proposition 106.88 (The obstruction)**.** *The triangle amplitude $T^{\mu\nu\rho}$ with one axial and two vector vertices depends on how the loop momentum is routed, by Lemma 106.86. Fixing the routing so that the two vector currents are conserved, $k_{1\nu} T^{\mu\nu\rho} = k_{2\rho} T^{\mu\nu\rho} = 0$, forces $(k_1 + k_2)_\mu T^{\mu\nu\rho} \neq 0$, and conversely. No regulator preserves both.*

Rests on Lemma 106.86 and Theorem 100.17.

Derivation. *Derives Proposition 106.88.*

The integrand of the triangle falls as $|\ell|^{-3}$ at large loop momentum, since it is a product of three fermion propagators (one power of ℓ each in the denominator, one in the numerator from each $\gamma \cdot \ell$) — the superficial degree of divergence is $D = +1$ by the counting of Theorem 100.17 applied to three external boson lines and a closed fermion loop. By Lemma 106.86 a shift of the routing therefore changes the amplitude by a finite, non-zero, momentum-dependent constant. The three Ward identities — one per vertex — are each linear conditions on that constant, and the constant has fewer free parameters than there are conditions: their sum is fixed independently of the routing and is non-zero. The routing can distribute the failure among the three vertices but cannot remove it. Since the vector current is coupled to the photon and its non-conservation would destroy gauge invariance and unitarity (Section 106.7.4), the failure is assigned entirely to the axial vertex.

The same statement in the language of regulators: Pauli–Villars requires a heavy fermion of mass M, and a Dirac mass term $M\bar\psi\psi$ is not chirally invariant, so the regulator breaks the axial symmetry by construction and the breaking survives $M \to \infty$ by Equation (106.98). Dimensional regularization requires γ^5 in $d \neq 4$ dimensions, where a totally antisymmetric product of four gamma matrices is not available. A lattice regulator is defeated by Theorem 106.75. Every route fails in a different way, which is itself the evidence that the obstruction is physical rather than technical. □

Theorem 106.89 (Adler–Bardeen)**.** *The one-loop result Equation (106.101) is exact: the coefficient of the anomaly receives no corrections at any higher order in the coupling [1389].*

Rests on Equation (106.101).

The reason is visible in Section 106.7.3 and not in the diagrams: the integrated anomaly is an integer — the index of the Dirac operator — and an integer cannot depend continuously on a coupling constant. This is the second time in this chapter that a topological argument makes a statement that a perturbative one can only verify order by order, the first being Proposition 106.81.

Proposition 106.90 (Wess–Zumino consistency)**.** *Let $G_a(x)$ generate gauge transformations, satisfying the algebra $[G_a(x), G_b(y)] = \mathrm{i} f_{abc}\, \delta^4(x-y)\, G_c(x)$, and let $\mathsf{A}_a(x) := G_a(x)\, W$ be the anomalous variation of the effective action. Then*

$$G_a(x)\, \mathsf{A}_b(y) - G_b(y)\, \mathsf{A}_a(x) = \mathrm{i} f_{abc}\, \delta^4(x-y)\, \mathsf{A}_c(x)\,, \tag{106.102}$$

a strong constraint on the possible form of an anomaly [1459].

Rests on Equations (106.37) *and* (106.67)*.*

Derivation. *Derives Proposition 106.90.*

Apply the commutator of two gauge generators to the single functional W and use the algebra: $[G_a(x), G_b(y)]\, W = \mathrm{i} f_{abc} \delta^4(x-y) G_c(x) W$. Writing out the left-hand side as $G_a(x)(G_b(y)W) - G_b(y)(G_a(x)W)$ and substituting the definition of A gives Equation (106.102). The content is that the anomaly is a cocycle of the gauge group: it is not an arbitrary functional but one constrained by the group structure, which is why the possible anomalies in four dimensions form a one-parameter family and the coefficient is the only thing left to compute. □

106.7.2 Neutral pion decay as the measurement

The anomaly is not a formal curiosity. It fixes, with no free parameter beyond one measured decay constant, the rate of a decay that has been measured to about one and a half percent — and the answer depends on the number of colours.

Phenomenon 106.91 (The neutral pion decays to two photons at the anomaly rate)**.** The two-photon width of the neutral pion is fixed by the axial anomaly to be

$$\Gamma(\pi^0 \to \gamma\gamma) = \left(\frac{N_c}{3}\right)^2 \frac{\alpha^2\, (m_\pi c^2)^3}{64\pi^3 F_\pi^2}\,, \qquad \frac{1}{\tau_{\gamma\gamma}} = \frac{\Gamma(\pi^0 \to \gamma\gamma)}{\hbar}\,, \tag{106.103}$$

with F_π the pion decay constant. The first quantity is an energy — the width as the Particle Data Group quotes it — and $\hbar$ enters only in the second, which converts it to a rate; the right-hand side of the first carries $(\text{energy})^3/(\text{energy})^2$ and no $\hbar$ at all. Using $m_{\pi^0} c^2 = (134.9768 \pm 0.0005)\,\text{MeV}$ [2427], $F_\pi = f_\pi/\sqrt{2} = 92.07\,\text{MeV}$ from $f_\pi = (130.2 \pm 1.2)\,\text{MeV}$ [2427], and $\alpha = 7.297\,352\,564\,3 \times 10^{-3}$ [2431], the prediction for $N_c = 3$ is

$$\Gamma = (7.79 \pm 0.14)\,\text{eV}\,, \qquad \frac{1}{\tau_{\gamma\gamma}} = 1.18 \times 10^{16}/\text{s}\,, \tag{106.104}$$

the quoted uncertainty being that of f_π alone, which enters squared and so contributes twice its own 0.9 %. Against this stands the PrimEx-II measurement

$$\Gamma_{\text{exp}} = (7.802 \pm 0.117)\,\text{eV} \tag{106.105}$$

[2361]: the two agree to 0.2 %, which is 0.09 of the combined uncertainty — and the prediction is now the less precise of the two, its error being dominated by the measured decay constant rather than by anything in the anomaly. Were there one colour instead of three, the prediction would be nine times smaller.

Rests on Equation (106.101) *and Theorem 106.89.*

Derivation. *Derives Phenomenon 106.91.*

The anomaly Equation (106.101), applied to the third component of the axial isospin current of the light quarks and combined with the partial conservation of that current, produces an effective interaction between the neutral pion field and two photons,

$$\mathcal{L}_{\pi\gamma\gamma} = -\frac{g_{\pi\gamma\gamma}}{4}\, \pi^0\, F_{\mu\nu} \widetilde{F}^{\mu\nu}\,, \qquad g_{\pi\gamma\gamma} = \frac{N_c}{3} \cdot \frac{\alpha}{\pi F_\pi}\,, \tag{106.106}$$

with $\widetilde{F}^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma}F_{\rho\sigma}$ and $g_{\pi\gamma\gamma}$ of dimension $(\text{energy})^{-1}$ once the fields are normalized as in Notation 106.23. The colour factor is $N_c(e_u^2 - e_d^2) = N_c(\frac{4}{9} - \frac{1}{9}) = \frac{N_c}{3}$: each quark flavour runs around the triangle once for each colour, weighted by the square of its electric charge, and the two light flavours enter with opposite sign because the current is the third isospin component. That is where the sensitivity to N_c comes from, and it is the only place in this book where a decay rate counts colours directly.

Given Equation (106.106), the width follows from kinematics. The amplitude for $\pi^0(p) \to \gamma(k_1,\varepsilon_1)\gamma(k_2,\varepsilon_2)$ is $\mathcal{M} = g_{\pi\gamma\gamma}\,\epsilon^{\mu\nu\rho\sigma}k_{1\mu}\varepsilon^*_{1\nu}k_{2\rho}\varepsilon^*_{2\sigma}$. Sum over photon polarizations with $\sum \varepsilon^*_\mu\varepsilon_\nu \to -\eta_{\mu\nu}$ — legitimate because $\mathcal{M}$ vanishes when either ε is replaced by its momentum. The two metric factors enter with $(-1)^2 = +1$, and the two Levi-Civita symbols contract by

$$\epsilon^{\mu\nu\rho\sigma}\,\epsilon^{\alpha}{}_{\nu}{}^{\gamma}{}_{\sigma} = -2(\eta^{\mu\alpha}\eta^{\rho\gamma} - \eta^{\mu\gamma}\eta^{\rho\alpha})$$

in the signature $\mathrm{diag}(+1,-1,-1,-1)$ with $\epsilon^{0123} = +1$, so that

$$\sum_{\text{pol}} |\mathcal{M}|^2 = -2g_{\pi\gamma\gamma}^2\left[k_1^2k_2^2 - (k_1\cdot k_2)^2\right] = 2g_{\pi\gamma\gamma}^2\,(k_1\cdot k_2)^2\,,$$

the first term vanishing because both photons are massless. The coefficient is 2, not 8: the same number comes out of summing $|\mathcal{M}|^2$ over the two physical transverse polarizations of each photon directly, without the covariant substitution. On the mass shell $2k_1\cdot k_2 = p^2 = m_\pi^2c^2$. Measuring every momentum of this kinematic step as an energy, $k \mapsto ck$, that reads $k_1\cdot k_2 = (m_\pi c^2)^2/2$ and makes $g_{\pi\gamma\gamma}$ an inverse energy throughout, whence $\sum|\mathcal{M}|^2 = \frac{1}{2}g_{\pi\gamma\gamma}^2(m_\pi c^2)^4$. The two-body massless phase-space volume is $1/8\pi$, the flux factor for a decay at rest is $1/(2m_\pi c^2)$, and the two photons are identical so the result is halved:

$$\frac{1}{\tau_{\gamma\gamma}} = \frac{1}{2}\cdot\frac{1}{2m_\pi c^2}\cdot\frac{1}{8\pi}\cdot\frac{g_{\pi\gamma\gamma}^2(m_\pi c^2)^4}{2\hbar} = \frac{g_{\pi\gamma\gamma}^2(m_\pi c^2)^3}{64\pi\hbar}\,,$$

a rate, since $g_{\pi\gamma\gamma}^2(m_\pi c^2)^3$ is an energy and $\hbar$ divides it. The width is $\hbar$ times that rate, $\Gamma = g_{\pi\gamma\gamma}^2(m_\pi c^2)^3/64\pi$, and substituting $g_{\pi\gamma\gamma}$ from Equation (106.106) gives Equation (106.103). That is the only place $\hbar$ enters, and it is why the displayed width carries none.

Arithmetic, in MeV: $\alpha^2 = 5.325\,14\times10^{-5}$; $(m_\pi c^2)^3 = 2.4591\times10^6\,\text{MeV}^3$, the cube of $134.9768\,\text{MeV}$; $64\pi^3 = 1984.40$; and $F_\pi^2 = 8476.0\,\text{MeV}^2$, the square of $130.2\,\text{MeV}/\sqrt{2}$. Then

$$\Gamma = \frac{5.325\,14\times10^{-5}\times2.4591\times10^6}{1984.40\times8476.0}\,\text{MeV} = 7.786\times10^{-6}\,\text{MeV} = 7.79\,\text{eV}\,,$$

which is Equation (106.104), and $\Gamma/\hbar = 7.79\,\text{eV}/6.582\,119\,569\times10^{-16}\,\text{eV s} = 1.18\times10^{16}/\text{s}$. The prediction and the measurement differ by 0.2 % — well inside either uncertainty separately, and 0.09 of the two combined in quadrature. □

Remark 106.92 (Two lifetimes that are both right). The shipped Particle Data Group compilation gives the *total* width $\Gamma_{\text{tot}} = (7.81\pm0.12)\,\text{eV}$ [2427], whence a mean life

$$\tau = \frac{\hbar}{\Gamma_{\text{tot}}} = \frac{1.054\,571\,817\times10^{-34}\,\text{J s}}{7.81\,\text{eV}\times1.602\,176\,634\times10^{-19}\,\text{J/eV}} = 8.43\times10^{-17}\,\text{s}\,.$$

The PrimEx-II paper quotes $\tau = 8.34\times10^{-17}\,\text{s}$ [2361], which is a different number and not a discrepancy: what PrimEx-II measures is the two-photon partial width Equation (106.105), and converting it to a lifetime requires dividing by the branching fraction $B(\pi^0\to\gamma\gamma) = 0.988\,23$, giving $\Gamma_{\text{tot}} = 7.895\,\text{eV}$ and $\tau = \hbar/\Gamma_{\text{tot}} = 8.34\times10^{-17}\,\text{s}$. The two lifetimes differ by about one percent because the PrimEx-II partial width is about one percent above the world average it is now the dominant input to. Quoting either without saying which is which is how a one-percent inconsistency gets into a book.

This is the cleanest existing evidence that the number of colours is three, and it is independent of the other route — the ratio R of hadronic to muonic annihilation cross-sections at an electron–positron collider, Equation (102.14), which counts colours through a sum over quark charges rather than through an anomaly. The two are worth having both, because they fail in different ways: R requires the perturbative treatment of the final state and carries QCD corrections of order α_s/π, while Equation (106.103) is protected from corrections by Theorem 106.89 but requires the chiral treatment of the pion.

Derivation pending #167

The chiral matching for the pion two-photon coupling: the phase space, the polarization sum and the colour counting of the neutral-pion width are derived above, but the step from the anomalous divergence of the axial isospin current to the effective coupling $g_{\pi\gamma\gamma}$ — the partial conservation of the axial current, the pion pole dominance and the normalization by the decay constant — is quoted rather than derived. It belongs in Appendix A, alongside the chiral material of the QCD chapter of this part.

106.7.3 Fujikawa's derivation from the measure

Every derivation so far has located the anomaly in a divergent loop integral. Fujikawa showed that it is in the *measure* [1638]: the classical action of a massless fermion is chirally invariant, and so is the classical field equation, but $\mathcal{D}\bar\psi\,\mathcal{D}\psi$ is not.

Theorem 106.93 (Fujikawa)**.** *Under the local chiral rotation Equation (106.99) with $\alpha=\alpha(x)$, the Euclidean fermion measure transforms with the Jacobian*

$$\mathcal{D}\bar\psi'\,\mathcal{D}\psi' = \mathcal{D}\bar\psi\,\mathcal{D}\psi\;\exp\!\left[-\frac{\mathrm{i}}{16\pi^2}\int \mathrm{d}^4x\;\alpha(x)\,\epsilon_{\mu\nu\rho\sigma}\,\mathrm{tr}(\mathcal{F}_{\mu\nu}\mathcal{F}_{\rho\sigma})\right], \tag{106.107}$$

which for constant α is $\mathrm{e}^{-2\mathrm{i}\alpha\nu}$ with ν the topological charge of Equation (106.91). The anomaly Equation (106.101) follows.

Rests on Equations (106.91) and (106.99) and Corollary 106.36.

Derivation. *Derives Theorem 106.93.*

Expand the fields in eigenfunctions of the Hermitian Euclidean Dirac operator $\mathcal{D}=\gamma^\mu D_\mu$, $\mathcal{D}\varphi_n=\lambda_n\varphi_n$, with the φ_n orthonormal: $\psi(x)=\sum_n a_n\varphi_n(x)$, $\bar\psi(x)=\sum_n \bar b_n\varphi_n^\dagger(x)$, the coefficients $a_n,\bar b_n$ being Grassmann. The measure is $\prod_n \mathrm{d}a_n\,\mathrm{d}\bar b_n$.

Under Equation (106.99) the coefficients transform linearly, $a'_m=\sum_n C_{mn}a_n$ with

$$C_{mn}=\int \mathrm{d}^4x\;\varphi_m^\dagger(x)\,\mathrm{e}^{\mathrm{i}\alpha(x)\gamma^5}\varphi_n(x)=\delta_{mn}+\mathrm{i}\int \mathrm{d}^4x\;\alpha(x)\,\varphi_m^\dagger\gamma^5\varphi_n+O(\alpha^2),$$

and $\bar b$ transforms with the same matrix, because Equation (106.99) carries the same sign for $\bar\psi$. By Corollary 106.36 a Grassmann measure transforms with the *inverse* determinant, so

$$\prod_n \mathrm{d}a'_n\,\mathrm{d}\bar b'_n=(\det C)^{-2}\prod_n \mathrm{d}a_n\,\mathrm{d}\bar b_n\,,\qquad (\det C)^{-2}=\exp\!\left[-2\mathrm{i}\int \mathrm{d}^4x\;\alpha(x)\,\mathcal{J}(x)\right],$$

using $\det=\exp\mathrm{tr}\ln$ to first order in α, with $\mathcal{J}(x)=\sum_n\varphi_n^\dagger(x)\gamma^5\varphi_n(x)$.

The sum $\mathcal{J}$ is formally $\mathrm{tr}\,\gamma^5$ times a delta function — zero times infinity — and must be regularized. The regulator that respects gauge invariance is the operator's own spectrum:

$$\mathcal{J}(x)=\lim_{M\to\infty}\sum_n \varphi_n^\dagger(x)\gamma^5\mathrm{e}^{-\lambda_n^2/M^2}\varphi_n(x)=\lim_{M\to\infty}\mathrm{tr}\,\langle x|\,\gamma^5\mathrm{e}^{-\mathcal{D}^2/M^2}\,|x\rangle\,,$$

where the trace is over Dirac and gauge indices. Insert plane waves, $\langle x|\cdots|x\rangle = \int \mathrm{d}^4k\,(2\pi)^{-4}\,\mathrm{e}^{-\mathrm{i}k\cdot x}(\cdots)\mathrm{e}^{\mathrm{i}k\cdot x}$, and use

$$\mathcal{D}^2 = (\gamma^\mu D_\mu)^2 = D^2 + \frac{1}{2}\gamma^\mu\gamma^\nu\,[D_\mu, D_\nu] = D^2 + \frac{\mathrm{i}}{2}\gamma^\mu\gamma^\nu\mathcal{F}_{\mu\nu}\,,$$

by Equation (106.64). Expanding the exponential, the Dirac trace with γ^5 vanishes unless at least four gamma matrices are present, so the first surviving term is the second order in $\mathcal{F}$:

$$\mathcal{J} = \lim_{M\to\infty}\int \frac{\mathrm{d}^4k}{(2\pi)^4}\,\mathrm{e}^{-k^2/M^2}\,\frac{1}{2!}\,\frac{1}{M^4}\left(\frac{\mathrm{i}}{2}\right)^2 \mathrm{tr}\Big(\gamma^5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma\Big)\,\mathrm{tr}(\mathcal{F}_{\mu\nu}\mathcal{F}_{\rho\sigma})\,.$$

Terms with fewer than four gammas vanish on the Dirac trace; terms with more carry additional inverse powers of M and vanish in the limit. Now $\int \mathrm{d}^4k\,(2\pi)^{-4}\mathrm{e}^{-k^2/M^2} = M^4/(16\pi^2)$ by Lemma 106.2, which cancels the M^{-4} exactly — so the limit is finite and independent of the regulator scale — and $\mathrm{tr}(\gamma^5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma) = -4\epsilon^{\mu\nu\rho\sigma}$. Collecting the numbers, $\frac{1}{2}\times(-\frac{1}{4})\times(-4)\times\frac{1}{16\pi^2} = \frac{1}{32\pi^2}$, so

$$\mathcal{J}(x) = \frac{1}{32\pi^2}\,\epsilon_{\mu\nu\rho\sigma}\,\mathrm{tr}(\mathcal{F}_{\mu\nu}\mathcal{F}_{\rho\sigma})\,, \tag{106.108}$$

which is Equation (106.107). Comparison with Equation (106.91) gives $\int \mathrm{d}^4x\,\mathcal{J} = \nu$, so for constant α the Jacobian is $\mathrm{e}^{-2\mathrm{i}\alpha\nu}$.

Finally, the anomalous divergence. Invariance of Z under a change of integration variables requires that the variation of the action cancel the Jacobian. The action's variation under a local rotation produces $\int \alpha\,\partial_\mu j_5^\mu$ after integrating by parts, so $\partial_\mu j_5^\mu = 2c\,\mathcal{J}$ for a massless field, i.e. $\partial_\mu j_5^\mu = 2q(x)$ with the topological charge density q of Definition 106.79 — not to be confused with the electric charge, which is why the latter is written Q throughout this chapter. For the abelian case, $\mathcal{F}_{\mu\nu} = QF_{\mu\nu}/\hbar$ with no trace factor, and $2c/(32\pi^2)\times Q^2/\hbar^2 = Q^2c/(16\pi^2\hbar^2)$, which is Equation (106.101). □

Three things about this derivation are worth separating out.

It uses the Grassmann Jacobian, and only the Grassmann Jacobian. Corollary 106.36 — that a fermionic measure transforms with $(\det)^{-1}$ where a bosonic one transforms with $\det$ — was derived in Section 106.3.3 from Berezin's rules, which were themselves derived from translation invariance. That chain is the whole content of the anomaly: a symmetry of the action that fails to be a symmetry of the measure.

It is manifestly topological. By Equation (106.108) and Proposition 106.81, $\int \mathcal{J}$ is an integer. The Atiyah–Singer index theorem identifies it further: the eigenvalue sum $\sum_n \varphi_n^\dagger\gamma^5\varphi_n$ receives contributions only from zero modes, because $\mathcal{D}$ maps the $\gamma^5 = +1$ and $\gamma^5 = -1$ subspaces into each other and therefore pairs every non-zero eigenvalue with a partner of opposite chirality, whose contributions cancel. Hence

$$\nu = n_+ - n_-\,, \tag{106.109}$$

the difference between the numbers of right- and left-handed zero modes of the Dirac operator. A decay rate — the measured $(7.802 \pm 0.117)\,\mathrm{eV}$ of Equation (106.105) — is thereby tied to an index. It is also the reason for Theorem 106.89: a quantity that equals an integer cannot vary continuously with a coupling constant.

It explains the third failure. The chiral rotation that would remove the phase of the quark mass matrix does not remove it: it shifts θ by the Jacobian, which is why only $\bar\theta = \theta + \arg\det M_q$ is observable (Equation (105.111)). The mechanism promised there is Equation (106.107), exactly.

106.7.4 Anomaly cancellation

An anomaly in a *global* symmetry is a prediction, and Section 106.7.2 measured one. An anomaly in a *gauge* symmetry is fatal.

Theorem 106.94 (A gauge anomaly destroys the theory)**.** *If the currents coupled to gauge fields are anomalous, the Slavnov–Taylor identities Equation* (106.76) *fail. Then the unphysical polarizations no longer decouple (Proposition 106.60), the S-matrix is not unitary on $\mathcal{H}_{\text{phys}}$, and the divergences no longer organize into a renormalization of finitely many parameters [1475].*

Rests on Theorem 106.62, Proposition 106.60, and Equation (106.107)*.*

Derivation. *Derives Theorem 106.94.*

Theorem 106.62 derived Equation (106.76) from the invariance of the action *and of the measure* under BRST. An anomaly is precisely the statement that the measure is not invariant: by Equation (106.107) the Jacobian contributes an extra term to every identity, so $\langle sX\rangle = \langle X\,\mathsf{A}\rangle \neq 0$. The chain of consequences then runs backwards through Section 106.4.2: the decoupling of the quartet in Proposition 106.60 used $\langle\Phi|Q_B\Lambda\rangle = 0$, which requires Q_B to be conserved, and gauge-parameter independence Theorem 106.61 used $\langle sX\rangle = 0$ directly. Losing them means that a gauge-dependent quantity appears in a physical answer and that states of negative norm contribute to the optical theorem, which is a loss of probability, not a technical inconvenience. □

Theorem 106.95 (The cancellation condition)**.** *For chiral fermions in a representation with generators T^a, the gauge anomaly is proportional to*

$$\mathsf{A}^{abc} = \operatorname{tr}\Big(\big\{T^a,T^b\big\}\,T^c\Big)_L - \operatorname{tr}\Big(\big\{T^a,T^b\big\}\,T^c\Big)_R\,, \tag{106.110}$$

the traces running over the left- and right-handed multiplets separately, and the theory is consistent if and only if $\mathsf{A}^{abc}=0$ for every triple [1467].

Rests on Propositions 106.88 and 106.90.

Derivation. *Derives Theorem 106.95.*

The triangle of Proposition 106.88 now carries a group generator at each of its three vertices, and the two orientations of the fermion loop contribute the two orderings, so the group factor is the symmetrized trace $\operatorname{tr}(T^aT^bT^c + T^aT^cT^b) = \operatorname{tr}(\big\{T^b,T^c\big\}\,T^a)$. A right-handed fermion contributes with the opposite sign, because γ^5 has the opposite eigenvalue and the anomalous piece of the triangle is the one proportional to γ^5. Summing over all fermions in the theory gives Equation (106.110), and by Proposition 106.90 nothing else can appear. □

Proposition 106.96 (The Standard Model cancels, and only with three colours)**.** *Write each generation as left-handed Weyl fields with hypercharges Y normalized by $Q = T^3 + Y$: $Q_L = (3,2)_{1/6}$, $u_L^c = (\bar3,1)_{-2/3}$, $d_L^c = (\bar3,1)_{1/3}$, $L_L = (1,2)_{-1/2}$, $e_L^c = (1,1)_{1}$. Then every anomaly coefficient of $\mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1)$ vanishes, and two of them vanish only because the number of colours equals three.*

Rests on Theorem 106.95 and Equation (106.110)*.*

Derivation. *Derives Proposition 106.96.*

There are four independent conditions, since $\mathrm{SU}(3)^3$ vanishes identically (the quarks appear in a vector-like pair of triplet and antitriplet with respect to colour, once left and right are combined) and $\mathrm{SU}(2)^3$ vanishes because the symmetrized trace of three SU(2) generators is zero for any representation, SU(2) having only real representations.

(i) $\mathrm{SU}(3)^2\,\mathrm{U}(1)$. Sum Y over the colour-triplet Weyl fields, weighted by the number of SU(2) components:

$$2\times\tfrac{1}{6} + 1\times\left(-\tfrac{2}{3}\right) + 1\times\tfrac{1}{3} = \tfrac{1}{3} - \tfrac{2}{3} + \tfrac{1}{3} = 0\,.$$

(ii) $\mathrm{SU}(2)^2\,\mathrm{U}(1)$. Sum Y over the SU(2) doublets, weighted by their colour multiplicity:

$$N_c \times \tfrac{1}{6} + 1 \times \left(-\tfrac{1}{2}\right) = \frac{N_c}{6} - \frac{1}{2}\,,$$

which vanishes if and only if $N_c = 3$.

(iii) $\mathrm{U}(1)^3$. Sum Y^3 over all Weyl fields with their multiplicities $6, 3, 3, 2, 1$:

$$6\left(\tfrac{1}{6}\right)^3 + 3\left(-\tfrac{2}{3}\right)^3 + 3\left(\tfrac{1}{3}\right)^3 + 2\left(-\tfrac{1}{2}\right)^3 + 1\cdot 1^3 = \tfrac{1}{36} - \tfrac{32}{36} + \tfrac{4}{36} - \tfrac{9}{36} + \tfrac{36}{36} = 0\,.$$

Carrying N_c instead of 3 through the quark multiplicities gives $N_c\left[\frac{1}{108} - \frac{8}{27} + \frac{1}{27}\right] + \frac{3}{4} = \frac{3-N_c}{4}$, so this condition too holds only for $N_c = 3$.

(iv) Mixed gauge–gravitational. The same triangle with two insertions of the energy–momentum tensor requires $\sum Y = 0$:

$$6\left(\tfrac{1}{6}\right) + 3\left(-\tfrac{2}{3}\right) + 3\left(\tfrac{1}{3}\right) + 2\left(-\tfrac{1}{2}\right) + 1 = 1 - 2 + 1 - 1 + 1 = 0\,.$$

This one is a statement about the Standard Model in a fixed classical background metric — a one-loop consistency condition on the gauge currents, not a theory of quantum gravity.

In every case the quarks alone and the leptons alone fail; only the complete generation cancels, and only with three colours. This is the theoretical counterpart of the measurement in Section 106.7.2, and the reason Chapter 104 requires complete generations: a fourth lepton doublet without its quark partners would make the theory inconsistent, so the discovery of a new charged lepton is a prediction that its quarks exist. □

Remark 106.97 (The surviving non-perturbative constraint). Theorem 106.95 is a perturbative condition, derived from a triangle diagram. Witten found a further one that no diagram sees [1707]: because $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$, there are gauge transformations on S^4 not continuously connected to the identity, and the fermion determinant changes sign under them unless the number of SU(2) doublets is even. The functional integral then vanishes identically — the theory has no non-vanishing amplitudes at all, which is a more complete failure than a broken Ward identity.

The Standard Model passes. Each generation contributes $N_c = 3$ quark doublets and one lepton doublet, four in all, and three generations give twelve: even. Had the leptons been counted without the quarks, or the colours been two, the count would have been odd and the theory empty. That two independent consistency conditions — one perturbative and one topological — and one measurement, Equation (106.105), all point at the same integer is the kind of redundancy that makes the Standard Model believable.

Chapter 107

Nuclear Forces and Nuclear Structure

The nucleus is the one place in this book where a residual force does the work. Quantum chromodynamics (Chapter 102) confines quarks into colour-neutral hadrons, and what leaks out between those hadrons binds protons and neutrons into nuclei — the analogue of the van der Waals force between neutral atoms, and about as hard to derive from first principles. The chapter therefore begins where the evidence began: Rutherford's inference of a small massive nucleus from large-angle alpha scattering [379], Chadwick's neutron [693], and Heisenberg's two-nucleon picture of nuclear matter [699]. From there it develops the Yukawa exchange mechanism [765], the pion that confirmed it [902], and the two complementary models — the liquid drop [763, 768, 830] and the shell model [933, 940] — whose reconciliation is the collective model of Bohr and Mottelson [1015].

The chapter sits after the field theory of Chapters 64 and 102 because that is where its force comes from, and before the astrophysical applications of Chapter 52, where nuclear binding energies set the energy budget of every star. Its data are unusually good and unusually well tabulated: masses to sub-keV precision [2393], charge radii from electron scattering [1088, 2209], and gamma-ray resonances so narrow that Mössbauer's recoil-free absorption [1154] measured a gravitational redshift over 22.5 m [1189]. Every quantitative claim below is anchored to such a measurement.

Notation 107.1 (Units and constants used throughout). Units are SI, with the electronvolt admitted as in Section 100.1.1, and $\hbar$ and c are written out in every formula. Three combinations recur so often that they are fixed here once, from the CODATA 2022 adjustment [2431]:

$$\hbar c = 1.973\,269\,8 \times 10^{-7}\ \text{eV m} = 197.326\,98\ \text{MeV fm}\,, \tag{107.1}$$

$$\frac{e^2}{4\pi\varepsilon_0} = \alpha\,\hbar c = 1.439\,964\,5\ \text{MeV fm}\,, \tag{107.2}$$

$$u\,c^2 = 931.494\ \text{MeV}\,, \tag{107.3}$$

with $\alpha = e^2/(4\pi\varepsilon_0\hbar c) = 7.297\,352\,6 \times 10^{-3}$ the dimensionless fine-structure constant and u the unified atomic mass unit. One femtometre, 1×10^{-15} m, is the natural length here; one MeV the natural energy. A nucleus is written ${}^{A}_{Z}\mathrm{X}_{N}$ with Z protons, N neutrons and $A = Z + N$ nucleons; Z is redundant given the chemical symbol and N is redundant given A, so the short form ${}^{A}\mathrm{X}$ is used except where the neutron number is the point. Nuclides sharing Z are *isotopes*, sharing N *isotones*, sharing A *isobars*; a long-lived excited state of one nuclide is an *isomer*, marked ${}^{Am}\mathrm{X}$.

Two masses anchor everything that follows. From the 2024 Particle Data Group compilation [2427],

$$m_p c^2 = 938.272\,09\ \text{MeV}\,, \qquad m_n c^2 = 939.565\,42\ \text{MeV}\,, \tag{107.4}$$

equivalently $m_p = 1.672\,621\,93 \times 10^{-27}\,\mathrm{kg}$ and $m_n = 1.674\,927\,50 \times 10^{-27}\,\mathrm{kg}$ [2431]. Their difference,

$$(m_n - m_p)\,c^2 = 1.293\,33\,\mathrm{MeV}\,, \tag{107.5}$$

is larger than the electron rest energy 0.511 00 MeV and smaller than the binding energy per nucleon of any bound nucleus, and those two inequalities between them decide which nuclei are stable, which decay, and in which direction — a point taken up in Section 107.3.2.

107.1 Discovery of the nucleus

107.1.1 Radioactivity

Becquerel found in 1896 that uranium salts fog a photographic plate through opaque wrapping, and — decisively — that the effect does not require prior exposure to sunlight, so it is not phosphorescence [271]. The radiation is emitted continuously, at a rate independent of temperature, pressure and chemical combination, and it discharges an electroscope, which made it measurable rather than merely visible. Two years later the Curies, following the ionization current rather than the photographic plate, isolated from pitchblende two new elements far more active than uranium: polonium [278] and radium [279]; radium's activity per unit mass exceeds uranium's by more than six orders of magnitude [279], which is what made the subsequent experiments possible at all.

Rutherford separated the radiation by absorption [287]. One component is stopped by a few hundredths of a millimetre of aluminium and one requires some millimetres; he named them α and β. A third, far more penetrating and undeflected by a magnetic field, was identified later and named γ. The identifications that followed are: α is a helium nucleus, β an electron (or, after 1934, a positron [737]), γ a photon.

Phenomenon 107.2 (Radioactive decay is exponential and spontaneous)**.** A sample of N identical nuclei of one species disintegrates at a rate proportional to N alone,

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -\lambda N\,, \qquad N(t) = N(0)\,\mathrm{e}^{-\lambda t}\,, \tag{107.6}$$

with a decay constant λ of dimension 1/s that is a property of the species and of nothing else. It is unaffected by temperature over the range from liquid helium to a flame, by pressure, by the chemical state of the atom and by the presence of other nuclei — an insensitivity of many orders of magnitude, and the first evidence that the process happens inside a system decoupled from the chemistry. Rutherford and Soddy drew the conclusion that in a decay one element is transmuted into another [312].

Derivation. *Derives Phenomenon 107.2.*

Suppose only that each nucleus has, in any short interval $\mathrm{d}t$, the same probability $\lambda\,\mathrm{d}t$ of decaying, independently of its age and of what the other nuclei do. This is a statement about a collection of identical, memoryless, independent systems, and it is the whole physical input; everything else is arithmetic. The number of decays expected in $\mathrm{d}t$ is then $\lambda N\,\mathrm{d}t$, which is Equation (107.6), and the solution follows by separation of variables. The *half-life* is the time for N to halve,

$$T_{1/2} = \frac{\ln 2}{\lambda}\,, \tag{107.7}$$

and the *mean life* is $\tau = 1/\lambda = T_{1/2}/\ln 2$.

Two properties of this law are worth stating exactly, because both are routinely misread. First, memorylessness is not an extra assumption but a consequence: the probability that a nucleus surviving to time t survives a further s is $\mathrm{e}^{-\lambda s}$, independent of t, so a nucleus does not age. The exponential is the unique distribution with that property. Second, λ is a

statement about an ensemble and says nothing whatever about when a given nucleus will decay. The number $\mathcal{N}$ counted in a fixed interval is a Poisson variable of mean $\mu = \lambda N\,\Delta t$ and variance μ, so the fractional uncertainty on a measured activity is $\mathcal{N}^{-1/2}$; this is the statistics of Chapter 11, and it is why a counting experiment quotes $\sqrt{\mathcal{N}}$ as its error bar.

The *activity* $\mathcal{A} = \lambda N$ is measured in becquerels, $\mathrm{Bq} = 1/\mathrm{s}$, one disintegration per second. The older unit, the curie, was defined as the activity of one gram of ^{226}Ra, and the definition can be checked: with $T_{1/2} = 1600\,\mathrm{yr} = 5.05 \times 10^{10}\,\mathrm{s}$ [2382] and $N = N_A/226$ atoms per gram,

$$\mathcal{A} = \frac{\ln 2}{5.05 \times 10^{10}\,\mathrm{s}} \cdot \frac{6.022\,140\,76 \times 10^{23}/\mathrm{mol}}{226\,\mathrm{g/mol}} \times 1\,\mathrm{g} = 3.66 \times 10^{10}\,\mathrm{Bq}\,, \tag{107.8}$$

against the defined $3.7 \times 10^{10}\,\mathrm{Bq}$; the small discrepancy is the improvement in the measured half-life since 1910. □

Proposition 107.3 (Secular equilibrium in a decay chain)**.** *Let a parent of decay constant λ_1 feed a daughter of decay constant λ_2. Then*

$$N_2(t) = \frac{\lambda_1 N_1(0)}{\lambda_2 - \lambda_1}\left(\mathrm{e}^{-\lambda_1 t} - \mathrm{e}^{-\lambda_2 t}\right) + N_2(0)\,\mathrm{e}^{-\lambda_2 t}\,, \tag{107.9}$$

and if the parent is much the longer lived, $\lambda_1 \ll \lambda_2$, the daughter's activity approaches the parent's, $\lambda_2 N_2 \to \lambda_1 N_1$, after a time of order $1/\lambda_2$.

Rests on Phenomenon 107.2 and Equation (107.6).

Proof. *Derives Proposition 107.3.*

The daughter is created at the rate at which the parent decays and destroyed at its own rate, $\mathrm{d}N_2/\mathrm{d}t = \lambda_1 N_1 - \lambda_2 N_2$, a linear first-order equation whose integrating factor is $\mathrm{e}^{\lambda_2 t}$. With $N_1 = N_1(0)\mathrm{e}^{-\lambda_1 t}$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(N_2 \mathrm{e}^{\lambda_2 t}\right) = \lambda_1 N_1(0)\,\mathrm{e}^{(\lambda_2 - \lambda_1)t}\,,$$

which integrates to Equation (107.9). When $\lambda_1 \ll \lambda_2$ the second exponential dies first and $N_2 \to \lambda_1 N_1/\lambda_2$, so the two activities are equal. This is why a sealed uranium ore assays as though every member of the chain were a separate element of fixed abundance, and why the Curies could isolate radium at all: its abundance in pitchblende is set by Equation (107.9) at $\lambda_1/\lambda_2 \approx 3.6 \times 10^{-7}$ of the uranium by number, which is about one gram of radium in three tonnes of uranium. □

Remark 107.4 (What "spontaneous" does and does not mean). Nothing in Phenomenon 107.2 says a decay is uncaused; it says the cause is not in the laboratory. The insensitivity to chemistry follows because a chemical energy is of order 1 eV and a nuclear one of order 1 MeV: a perturbation six orders of magnitude below the scale of the states involved cannot shift a rate. The exceptions prove the rule and are measurable. Electron capture (Section 107.5.2) proceeds through the electron density at the nucleus, so the half-life of ^{7}Be differs by a fraction of a per cent between chemical compounds; and a fully ionized nucleus in a storage ring cannot capture at all.

107.1.2 Rutherford scattering

Geiger and Marsden, at Rutherford's suggestion, looked for alpha particles scattered backwards from a thin metal foil, and found them [354]. The Thomson atom, a diffuse positive charge with electrons embedded in it, cannot produce them: a single encounter can deflect an alpha particle by at most a small angle, so a large deflection requires many small ones to conspire, and the probability of that falls off exponentially. What was observed instead falls off as a power.

Phenomenon 107.5 (Alpha particles are occasionally turned back)**.** A beam of α particles crossing a thin metal foil is overwhelmingly transmitted with small deflection, but a small and countable fraction is scattered through more than a right angle [354]. The number counted per unit solid angle varies with the scattering angle as

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \left(\frac{Z_1 Z_2 e^2}{16\pi\varepsilon_0 E}\right)^2 \frac{1}{\sin^4(\theta/2)}\,, \tag{107.10}$$

with E the kinetic energy of the incident particle [379]. The charge and the mass of the atom are therefore concentrated in a volume small compared with the atom, and the energy at which the formula begins to fail is the first bound ever placed on the size of that volume.

Derivation. *Derives Phenomenon 107.5.*

The interaction is the repulsive Coulomb potential $V(r) = Z_1 Z_2 e^2/4\pi\varepsilon_0 r$. It is central, so the orbit is a hyperbola and the deflection of a particle incident with impact parameter b and kinetic energy E is fixed by

$$\tan\frac{\theta}{2} = \frac{Z_1 Z_2 e^2}{8\pi\varepsilon_0 E b} \tag{107.11}$$

[1585, 2027], the hyperbolic-orbit relation of the classical repulsive Kepler problem. Write $\kappa = Z_1 Z_2 e^2/8\pi\varepsilon_0 E$, so that $b = \kappa\cot(\theta/2)$. The deflection decreases monotonically with b, so the particles arriving in the annulus between b and $b + \mathrm{d}b$ are exactly those leaving between θ and $\theta - \mathrm{d}\theta$, and the cross-section is the ratio of the two areas,

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{b}{\sin\theta}\left|\frac{\mathrm{d}b}{\mathrm{d}\theta}\right|\,. \tag{107.12}$$

From $b = \kappa\cot(\theta/2)$ one has $|\mathrm{d}b/\mathrm{d}\theta| = \frac{1}{2}\kappa/\sin^2(\theta/2)$, and with $\sin\theta = 2\sin(\theta/2)\cos(\theta/2)$,

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{\kappa\cos(\theta/2)/\sin(\theta/2)}{2\sin(\theta/2)\cos(\theta/2)}\cdot\frac{\kappa}{2\sin^2(\theta/2)} = \frac{\kappa^2}{4\sin^4(\theta/2)}\,,$$

which is Equation (107.10). Two features carry the physics. The distribution diverges as $\theta \to 0$ because the Coulomb force has infinite range, so no total cross-section exists and only a differential rate is meaningful. And it falls only as $\sin^{-4}(\theta/2)$: backscattering is rare but not exponentially rare, which is precisely what distinguishes a single point charge from a charge spread over the whole atom, for which small deflections compound at random and large ones are suppressed exponentially. □

Corollary 107.6 (First bound on the size of the nucleus)**.** *An α particle of kinetic energy E fired head-on at a nucleus of charge $Z_2 e$ stops at the distance d where all its kinetic energy has become potential energy,*

$$d = \frac{Z_1 Z_2 e^2}{4\pi\varepsilon_0 E}\,. \tag{107.13}$$

So long as Equation (107.10) is obeyed at angle $\theta = \pi$, the nucleus is smaller than d.
Rests on Equations (107.2) and (107.10).

Proof. *Derives Corollary 107.6.*

Energy conservation in the head-on collision gives $E = Z_1 Z_2 e^2/4\pi\varepsilon_0 d$, which is Equation (107.13); and Equation (107.10) was derived assuming a pure Coulomb field down to the turning point, so its persistence at $\theta = \pi$ certifies that nothing else was met. For the 7.7 MeV α particles from ^{214}Po incident on gold, $Z_2 = 79$, and Equation (107.2) gives

$$d = \frac{2 \times 79 \times 1.4400\,\mathrm{MeV\,fm}}{7.7\,\mathrm{MeV}} = 30\,\mathrm{fm}\,,$$

about 2×10^{-4} of the atomic radius. The volume ratio is 10^{-11}: essentially all of the atom is empty, and that is the content of the experiment. Geiger and Marsden went on to verify every separate dependence of Equation (107.10) — on angle, on foil thickness, on the kinetic energy and on Z_2^2 — which is what converted a suggestive anomaly into a measurement [404]. □

Remark 107.7 (Why the classical formula is right). Equation (107.10) was obtained from Newtonian orbits, and there is no reason in advance for a classical calculation to describe the scattering of a quantum particle. It does, exactly: the first Born approximation gives the same formula, and so does the exact solution of the Schrödinger equation in parabolic coordinates. The coincidence is special to the inverse-square law, and its discussion belongs to Chapter 86; the exact treatment differs from the classical one only in a logarithmic distortion of the phase, which no rate measurement sees. At energies high enough for the projectile's spin and the target's finite size to matter, the formula is replaced by Mott's [633] and then by a form factor, and the departure from Equation (107.10) becomes the measurement of the charge distribution treated in Section 107.3.3.

107.1.3 The neutron and the constitution of nuclei

Beryllium bombarded by α particles emits a radiation of extraordinary penetrating power that carries no charge. Curie and Joliot found that it ejects fast protons from paraffin. Chadwick measured the recoils and showed that no photon can account for them [693].

Phenomenon 107.8 (A neutral particle of nucleonic mass)**.** The radiation from α-bombarded beryllium is uncharged, is not absorbed like a photon, and ejects from hydrogen protons of maximum speed $3.3 \times 10^7\,\mathrm{m/s}$ and from nitrogen recoils of maximum speed $4.7 \times 10^6\,\mathrm{m/s}$ [693]. Elastic collisions of a single neutral particle account for both numbers with one mass, close to the proton's; a photon interpretation requires two different and impossibly large energies.

Derivation. *Derives Phenomenon 107.8.*

Take a neutral particle of mass m and speed v striking a nucleus of mass M at rest. In a head-on elastic collision, non-relativistic kinematics gives the target the maximum speed

$$V = \frac{2m}{m+M}\,v\,, \tag{107.14}$$

so the ratio of the maximum recoil speeds of hydrogen and nitrogen is

$$\frac{V_{\mathrm{H}}}{V_{\mathrm{N}}} = \frac{m + M_{\mathrm{N}}}{m + M_{\mathrm{H}}}\,, \tag{107.15}$$

in which the unknown v has cancelled. Measured, the left-hand side is $3.3 \times 10^7\,\mathrm{m/s}/4.7 \times 10^6\,\mathrm{m/s} = 7.0$; with $M_{\mathrm{H}} = m_p$ and $M_{\mathrm{N}} \approx 14 m_p$,

$$\frac{m/m_p + 14}{m/m_p + 1} = 7.0 \quad\Longrightarrow\quad m \approx 1.2\,m_p\,,$$

a mass within the experimental error of the proton's. The modern value is $m_n c^2 = 939.565\,42\,\mathrm{MeV}$ [2427], larger than the proton's by Equation (107.5).

The alternative had to be excluded quantitatively. A photon of energy E_γ scattering elastically from a mass M at rest imparts at most

$$T_{\max} = \frac{2E_\gamma^2}{Mc^2 + 2E_\gamma}\,, \tag{107.16}$$

the Compton kinematics of Chapter 69. The observed proton speed corresponds to $T = 5.7\,\mathrm{MeV}$, which by Equation (107.16) needs $E_\gamma \approx 55\,\mathrm{MeV}$; the observed nitrogen recoil corresponds to $T = 1.6\,\mathrm{MeV}$ and needs $E_\gamma \approx 1.0 \times 10^2\,\mathrm{MeV}$. One radiation cannot have two energies, and neither energy is available: the reaction $^9\mathrm{Be} + \alpha$ releases about 10 MeV in total. The photon hypothesis fails by an order of magnitude and by an internal inconsistency, and the neutral-particle hypothesis fails by neither. □

Proposition 107.9 (No electrons in the nucleus)**.** *A nucleus cannot be built of protons and electrons.*

Rests on Phenomenon 107.8 and Equation (107.1)*.*

Proof. *Derives Proposition 107.9.*

Three independent arguments close the question, and each is quantitative.

Localization. Confining an electron to a region of size $\Delta x \approx 5\,\mathrm{fm}$ requires, by the uncertainty relation, a momentum spread $\Delta p \gtrsim \hbar/2\Delta x$, that is

$$\Delta p\, c \gtrsim \frac{197.3\,\mathrm{MeV\,fm}}{2 \times 5\,\mathrm{fm}} = 20\,\mathrm{MeV}\,,$$

so the electron would be ultrarelativistic with an energy of order 20 MeV. Nothing in the nucleus binds at that scale — the total binding energy per nucleon is 8 MeV — and the electrons actually observed emerging in β decay carry about 1 MeV. The same estimate for a nucleon, heavier by $m_p/m_e \approx 1836$, gives a non-relativistic kinetic energy $(\Delta p)^2/2m_p \approx 0.2\,\mathrm{MeV}$, comfortably below the binding: a nucleon fits in a nucleus and an electron does not.

Spin and statistics. The proton–electron model assigns $^{14}\mathrm{N}$ fourteen protons and seven electrons, twenty-one spin-$\frac{1}{2}$ particles, hence half-integer total spin and Fermi–Dirac statistics for the molecule $^{14}\mathrm{N}_2$. Band spectra show integer spin 1 and Bose–Einstein statistics. The proton–neutron model assigns seven protons and seven neutrons, fourteen fermions, integer spin: correct. The argument is decisive because the alternation of intensities in a band spectrum measures the statistics directly and admits of no adjustment; the general theory of identical particles belongs to Chapter 83.

Magnetic moment. A nuclear magnetic moment is measured in units of the nuclear magneton $\mu_N = e\hbar/2m_p = 5.050\,783\,7 \times 10^{-27}\,\mathrm{J/T}$ [2431], some 1836 times smaller than the Bohr magneton. An electron inside the nucleus would contribute of order a Bohr magneton. No nuclear moment of that size exists [652]. □

Heisenberg took the neutron as a constituent immediately, and with it introduced the formal device that has organized nuclear physics ever since [699].

Definition 107.10 (Isospin)**.** Treat the proton and the neutron as two states of one particle, the *nucleon*, distinguished by the eigenvalue of a formal spin-$\frac{1}{2}$ operator $\boldsymbol{T}$ acting in an internal two-dimensional space:

$$|p\rangle = \left|T = \tfrac{1}{2},\, T_3 = +\tfrac{1}{2}\right\rangle\,, \qquad |n\rangle = \left|T = \tfrac{1}{2},\, T_3 = -\tfrac{1}{2}\right\rangle\,. \tag{107.17}$$

For a nucleus, $T_3 = (Z - N)/2$. The algebra is that of SU(2) (Chapter 14), and $\boldsymbol{T}$ is dimensionless: it is not an angular momentum and carries no factor of $\hbar$.

Remark 107.11 (What isospin asserts, and how well)*.* The claim is that the nuclear force is invariant under rotations in this space, so that states differing only in T_3 have the same energy. It is a good approximate symmetry and not an exact one, and the two sources of breaking are known and separately measurable: the electromagnetic interaction, which sees Z and not A, and the up–down quark mass difference of Chapter 102. The evidence is collected in Section 107.2.4; the anticipation is that the SU(2) here is a subgroup of the flavour symmetry of the light quarks, so that the approximate symmetry of nuclear physics is inherited from an approximate degeneracy in the quark masses.

Phenomenon 107.12 (The free neutron lifetime is disputed)**.** A neutron bound in a stable nucleus lives forever; a free neutron decays, $n \to p + e^- + \bar{\nu}_e$, with mean life

$$\tau_n = (878.4 \pm 0.5)\,\mathrm{s} \tag{107.18}$$

[2427]. Two classes of experiment measure it. Bottle experiments store ultracold neutrons and count the survivors; beam experiments pass a neutron beam through a trap and count the protons produced. The bottle result, $(877.75 \pm 0.36)\,\mathrm{s}$ [2380], and the beam result, near 888 s, differ by about 10 s, some four standard deviations. The discrepancy is unresolved.

Rests on Phenomenon 107.8 and Equation (107.5).

Remark 107.13 (How to report an unresolved discrepancy)*.* The two methods do not measure quite the same thing. A bottle counts neutrons that disappear by any route; a beam counts protons that appear by one route. If some small branch of neutron decay produced no proton, the beam lifetime would exceed the bottle lifetime by exactly the observed sense — which is why searches for such a branch have been made, and have found nothing. The honest statement is that either one experiment has an unaccounted systematic effect or the neutron has an unobserved decay channel at the 10^{-2} level, and that as of the 2024 compilation neither has been established [2427]. What is not honest is to average the two and quote the result, which is what an uncritical combination does. The number in Equation (107.18) is the Particle Data Group's average over the bottle measurements, with the beam results excluded on exactly these grounds; the derivation of the lifetime from the weak coupling belongs to Chapter 101.

107.2 The nucleon–nucleon interaction

107.2.1 Yukawa's exchange mechanism

Two facts about the nuclear force were established before anyone knew what carried it. It is enormously stronger than the electrostatic repulsion it overcomes, and it has a range: nucleons a few femtometres apart do not feel it at all, or nuclear matter would not saturate (Section 107.3.1). Electromagnetism has no range because the photon is massless. Yukawa's proposal was that the nuclear force is carried by a quantum that is not [765].

Theorem 107.14 (The Yukawa potential)**.** *A neutral scalar field of mass m coupled to static point sources produces between two of them the potential energy*

$$V(r) = -\frac{g^2}{4\pi}\,\frac{\hbar c}{r}\,\mathrm{e}^{-r/\lambda}\,, \qquad \lambda = \frac{\hbar}{mc}\,, \tag{107.19}$$

where λ is the reduced Compton wavelength of the quantum, in m*, and g is dimensionless. Equivalently, in terms of the momentum $\boldsymbol{q}$ transferred between the sources,*

$$\tilde{V}(\boldsymbol{q}) = \int \mathrm{d}^3 r\; \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}/\hbar}\, V(r) = -\frac{g^2\hbar^3 c}{\boldsymbol{q}^2 + m^2c^2}\,, \tag{107.20}$$

of SI dimension $\mathrm{J\,m^3}$*: the exchanged quantum's propagator, evaluated at zero energy transfer.*

Rests on Equation (107.1) and Notation 100.1.

Derivation. *Derives Theorem 107.14.*

Let ϕ obey the Klein–Gordon equation of Chapter 92 with a source,

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2 + \frac{1}{\lambda^2}\right)\phi = \rho\,, \qquad \frac{1}{\lambda} = \frac{mc}{\hbar}\,, \tag{107.21}$$

in which $1/\lambda$ has dimension $1/\mathrm{m}$, as it must: it is the only combination of m, c and $\hbar$ with that dimension, and its appearance is the whole of Yukawa's idea. For a static source the time derivative drops and

$$\left(\nabla^2 - \frac{1}{\lambda^2}\right)\phi = -g\,\delta^3(\boldsymbol{r})\,, \tag{107.22}$$

with a strength g whose dimension is fixed below. Away from the origin the solution is spherically symmetric, and the substitution $\phi = u(r)/r$ turns $\nabla^2\phi$ into u''/r, so that $u'' = u/\lambda^2$ and $u \propto \mathrm{e}^{\pm r/\lambda}$. Only the decaying sign is admissible. The coefficient is fixed by integrating Equation (107.22) over a small ball: the $1/\lambda^2$ term contributes nothing as the radius shrinks, and the divergence theorem gives $\oint \nabla\phi \cdot \mathrm{d}\boldsymbol{S} = -g$, which for $\phi = C\mathrm{e}^{-r/\lambda}/r$ is $-4\pi C$. Hence $C = g/4\pi$ and

$$\phi(r) = \frac{g}{4\pi}\,\frac{\mathrm{e}^{-r/\lambda}}{r}\,. \tag{107.23}$$

As $\lambda \to \infty$ this is the Coulomb field, which is the check that the algebra is right.

A second source at distance r has energy $-g'\phi(r)$ in this field, and for two identical nucleons $g' = g$. Writing the product g^2 so that the result carries the dimension of energy fixes the normalization: with g dimensionless the natural coefficient is $g^2\hbar c/4\pi$, since $\hbar c$ has dimension $\mathrm{J\,m}$ by Equation (107.1) and dividing by r gives an energy. That is Equation (107.19). The convention is deliberately the electromagnetic one: comparing with $V_{\mathrm{Coulomb}} = e^2/4\pi\varepsilon_0 r = \alpha\hbar c/r$ shows that $g^2/4\pi$ is the exact analogue of the fine-structure constant, dimensionless and directly comparable with $\alpha = 7.3 \times 10^{-3}$.

For Equation (107.20), do the angular integral,

$$\int \mathrm{d}^3 r\; \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}/\hbar}\frac{\mathrm{e}^{-r/\lambda}}{r} = \frac{4\pi\hbar}{|\boldsymbol{q}|}\int_0^\infty \mathrm{d}r\;\sin\!\left(\frac{|\boldsymbol{q}|\,r}{\hbar}\right)\mathrm{e}^{-r/\lambda} = \frac{4\pi\hbar^2}{\boldsymbol{q}^2 + \hbar^2/\lambda^2}\,,$$

the last step by elementary integration (or by closing a contour, as in Chapter 8). Multiplying by $-g^2\hbar c/4\pi$ and using $\hbar/\lambda = mc$ gives Equation (107.20). The denominator $\boldsymbol{q}^2 + m^2c^2$ is a squared momentum, in $\mathrm{kg^2\,m^2/s^2}$, exactly as the mass-shell condition of Notation 100.1 requires; it is the propagator denominator $p^2 - m^2c^2$ continued to a purely spacelike momentum transfer, which is what a static force is. □

Corollary 107.15 (Range fixes mass)**.** *A force of observed range λ requires a quantum of mass $m = \hbar/\lambda c$. For $\lambda = 1.4\,\mathrm{fm}$,*

$$mc^2 = \frac{\hbar c}{\lambda} = \frac{197.33\,\mathrm{MeV\,fm}}{1.4\,\mathrm{fm}} = 141\,\mathrm{MeV}\,, \tag{107.24}$$

that is $m = 141\,\mathrm{MeV}/c^2$, some 270 electron masses and about a seventh of a nucleon mass.
Rests on Theorem 107.14 and Equation (107.1).

Proof. *Derives Corollary 107.15.*

Immediate from $\lambda = \hbar/mc$ and Equation (107.1). The inference runs both ways, and Yukawa ran it in this direction: the range was known from the size of light nuclei — the deuteron is bound and the diproton is not, which places the range within a factor of two of a femtometre — and the mass was the prediction. It is one of the few predictions in physics of a new particle's mass, made in 1935 and confirmed in 1947 [902] with the observed $m_{\pi^\pm}c^2 = 139.570\,39\,\mathrm{MeV}$ [2427]. The equivalent range is

$$\lambda_\pi = \frac{\hbar}{m_\pi c} = \frac{197.326\,98\,\mathrm{MeV\,fm}}{139.570\,39\,\mathrm{MeV}} = 1.4138\,\mathrm{fm}\,. \tag{107.25}$$

□

Remark 107.16 (The exchange picture and its limits). Proposition 101.65 makes the same statement for the weak interaction, where the exchanged quantum weighs $80.37\,\mathrm{GeV}/c^2$ [2427] and the range is $2.5\times10^{-3}\,\mathrm{fm}$ — so short that the interaction looks pointlike at every energy at which it was first studied. The two cases differ in one respect that decides how far each may be pushed. The electroweak and electromagnetic couplings are small, so a single exchange is the leading term of a controlled expansion. Here $g^2/4\pi$ is not small: with the pion–nucleon coupling $g_{\pi N}=13.05$ from the Nijmegen partial-wave analysis of the nucleon–nucleon scattering data [1874],

$$\frac{g_{\pi N}^2}{4\pi}=13.6\,, \tag{107.26}$$

against $\alpha=1/137$, a ratio of nearly 2000. A one-quantum exchange is therefore not the first term of a convergent series, and Equation (107.19) cannot be trusted at short distance, where two-pion and heavier exchanges enter with comparable weight. What survives is the *tail*: at r well beyond λ_π the single-pion term dominates because every other exchange is more heavily damped, and it is there that Equation (107.19) is a theorem rather than a model. This is the observation that Section 107.2.4 turns into a systematic expansion.

Remark 107.17 (The quantum is pseudoscalar, and the potential changes). Theorem 107.14 assumed a *scalar* field, which is the simplest case and gives the range correctly. The pion is pseudoscalar (Section 107.2.2), so it couples to the nucleon through γ^5, and the derivative structure of that vertex replaces the plain e^{-x}/x by a spin- and isospin-dependent form. The change is not cosmetic. The effective coupling that appears is not $g_{\pi N}$ but

$$f=\frac{m_\pi c}{2m_N c}\,g_{\pi N}\,,\qquad \frac{f^2}{4\pi}=\frac{g_{\pi N}^2}{4\pi}\left(\frac{m_\pi}{2m_N}\right)^2=0.075\,, \tag{107.27}$$

smaller by the square of the mass ratio, and the strength of the long-range potential is set by $(f^2/4\pi)\,m_\pi c^2/3=3.5\,\mathrm{MeV}$ rather than by $13.6\,\hbar c/r$. Evaluated at $r=2\,\mathrm{fm}$ the one-pion-exchange tail is of order 0.6 MeV, which is the right size for the outer part of a force that binds the deuteron by 2.2 MeV. The apparent paradox of Equation (107.26) — a coupling of 13.6 producing a potential of a fraction of a MeV — is entirely accounted for by Equation (107.27), and noticing that is the difference between a numerical estimate and a number pulled from the air.

107.2.2 The pion

Yukawa's quantum was looked for in cosmic rays and, in 1937, apparently found: a penetrating charged particle of roughly the right mass. It was the wrong particle, and the experiment that showed so is a model of how to kill an identification.

Phenomenon 107.18 (The cosmic-ray mesotron is not strongly interacting)**.** Negative mesotrons brought to rest in carbon decay rather than being absorbed by a nucleus [899]. A negatively charged particle stopped in matter cascades into a Bohr orbit about a nucleus; if it felt the nuclear force it would be absorbed from that orbit in a time of order the nuclear transit time. The observed absorption rate is smaller than that estimate by some thirteen orders of magnitude.

Derivation. *Derives Phenomenon 107.18.*

The mesonic Bohr radius scales as $a_0 m_e/m$, so a particle of mass $\approx 200\,m_e$ orbits carbon at $a\approx a_0 m_e/(200\times6)=4.4\times10^{-14}\,\mathrm{m}$, which is still large compared with the nucleus but small enough that the overlap of the orbital wave function with the nuclear volume is $|\psi(0)|^2\,V_{\mathrm{nuc}}\sim(R/a)^3\approx2\times10^{-4}$. A strong interaction absorbs on contact, in a time $R/c\approx1\times10^{-23}\,\mathrm{s}$, so the expected lifetime against nuclear capture is $\sim5\times10^{-20}\,\mathrm{s}$ — against

a measured decay time $2.2 \times 10^{-6}\,\mathrm{s}$ [2427] that is unchanged by the presence of the nucleus. Thirteen orders of magnitude is not a discrepancy to be argued about. The particle is the muon; its lifetime and its couplings belong to Chapter 101.

The resolution came from photographic emulsions exposed at altitude, where Lattes, Muirhead, Occhialini and Powell found tracks in which one particle comes to rest and a second, of definite and shorter range, sets off from the end point [902]: the chain $\pi \to \mu \to e$. Two particles of similar mass had been conflated; the heavier one is Yukawa's, is produced strongly and is absorbed strongly, and therefore never reaches sea level, while the lighter is the weakly interacting decay product that does. □

Table 107.1: The pion triplet, from the 2024 Particle Data Group compilation [2427]. The mean life is computed from the listed width as $\tau = \hbar/\Gamma$, and the range from $\lambda = \hbar/mc$. The neutral pion decays electromagnetically to two photons, which is why it is shorter lived by nine orders of magnitude than its charged partners, whose only open channel is weak.

	mc^2 (MeV)	λ (fm)	τ	principal decay
$\pi^{\pm}$	139.57039(18)	1.4138	$2.603 \times 10^{-8}\,\mathrm{s}$	$\mu^{\pm}\nu$
π^{0}	134.9768(5)	1.4619	$8.43 \times 10^{-17}\,\mathrm{s}$	$\gamma\gamma$

Proposition 107.19 (The pion has odd intrinsic parity)**.** *The reaction $\pi^- + d \to n + n$, observed to proceed from a captured s state [985], requires the intrinsic parity of the pion to be -1.*

Rests on Phenomena 83.1 and 105.5.

Proof. *Derives Proposition 107.19.*

A π^- stopped in deuterium cascades to the $1s$ level of pionic deuterium and is captured from there, so the initial orbital angular momentum is $\ell = 0$. The deuteron has spin 1 and the pion spin 0 (Section 107.2.3 and the observed decay to two spin-zero particles), so the initial total angular momentum is $J = 1$ and the initial parity is $P_\pi P_d (-1)^\ell = P_\pi$, the deuteron being even.

The final state is two identical fermions, whose total wave function must be antisymmetric. Enumerate the two-neutron states with $J = 1$: ${}^1P_1$ has $S = 0$ (antisymmetric spin) and $L = 1$ (antisymmetric space), hence a symmetric total — forbidden; ${}^3S_1$ and ${}^3D_1$ have $S = 1$ (symmetric) and even L (symmetric) — also forbidden. Only ${}^3P_1$, with $S = 1$ and $L = 1$, is antisymmetric, and its parity is $(-1)^L = -1$. Parity is conserved by the strong interaction (Chapter 105), so $P_\pi = -1$.

The pion is therefore a *pseudoscalar*: spin 0, parity -1, written $J^P = 0^-$. Nothing in the argument is model dependent; it uses only the Pauli principle, the observed spins, and the fact that the capture is from an s state — which is itself measured, from the intensity of the X-rays emitted in the atomic cascade. □

Remark 107.20 (Why the pion is light, and why that is not a coincidence)*.* A pion weighs a seventh of a nucleon although both are made of the same light quarks, and the whole shape of nuclear physics — saturation at $0.17/\mathrm{fm}^3$, a binding energy per nucleon of 8 MeV, a force of range 1.4 fm — follows from that one number. The explanation is that the pion is not an ordinary bound state but the almost-Goldstone boson of a chiral symmetry of the light quarks that is spontaneously broken, so that its mass vanishes in the limit of massless quarks and is small because the quark masses are small. That account is the business of Chapter 102, and the relation between the pion decay constant, the pion–nucleon coupling and the axial coupling of the nucleon is already derived in Proposition 101.42, where it is verified to 0.6 per cent. What belongs here is only the consequence: the range of the nuclear

force is a low-energy relic of a symmetry of the quark Lagrangian, and not an independent parameter of nuclear physics.

107.2.3 The deuteron

The one bound state of two nucleons is the hydrogen atom of nuclear physics: simple enough to solve and rich enough to constrain the force. Urey, Brickwedde and Murphy found it spectroscopically, as a faint set of lines displaced from the Balmer series by the isotope shift [708]; Chadwick and Goldhaber measured its binding directly, by breaking it with the 2.62 MeV gamma ray of thorium C″ and detecting the proton [734] — the first nuclear photodisintegration, and the first direct measurement of a nuclear binding energy. The modern value from the atomic-mass evaluation is

$$B_d = (m_p + m_n - m_d)\,c^2 = 2.2246\,\mathrm{MeV} \tag{107.28}$$

[2393].

Phenomenon 107.21 (The deuteron is not spherical)**.** The only bound state of two nucleons has spin 1, a binding energy of 2.2246 MeV [2393], no excited state, and — decisively — a small but non-vanishing electric quadrupole moment, measured by molecular-beam magnetic resonance [840]. There is moreover no bound state of two protons and none of two neutrons.

Derivation. *Derives Phenomenon 107.21.*

A state of orbital angular momentum $L = 0$ has a spherically symmetric spatial distribution and therefore exactly zero quadrupole moment, so the observed moment proves that the deuteron ground state is not a pure S wave. But a central potential $V(r)$ commutes with $\boldsymbol{L}$, so L would then be a good quantum number and no admixture of $L = 2$ into an $L = 0$ ground state would be possible. The nucleon–nucleon interaction therefore contains a non-central piece. Which piece is fixed by what is available: the interaction must be a scalar built from the separation $\hat{\boldsymbol{r}}$ and the two spins, and the only non-central such scalar that is even under parity, even under time reversal and independent of velocity is the tensor operator

$$S_{12} = 3\,(\boldsymbol{\sigma}_1 \cdot \hat{\boldsymbol{r}})\,(\boldsymbol{\sigma}_2 \cdot \hat{\boldsymbol{r}}) - \boldsymbol{\sigma}_1 \cdot \boldsymbol{\sigma}_2\,. \tag{107.29}$$

This conserves the total angular momentum J but not L, and since $L = 0$ and $L = 2$ have the same parity it mixes exactly the ${}^3S_1$ and ${}^3D_1$ states that share $J = 1$. The measured quadrupole moment is thus a direct measurement of the D-state admixture. The two null results complete the picture: that no diproton and no dineutron binds, while the deuteron does, says the force is spin dependent, since the ${}^1S_0$ configuration is the only one open to two identical nucleons at low energy and it is very nearly, but not quite, binding. □

Proposition 107.22 (The deuteron is barely bound)**.** *Model the ${}^3S_1$ interaction as a square well of depth V_0 and radius R. The bound state at energy $-B_d$ requires*

$$k \cot kR = -\kappa\,, \qquad k = \frac{\sqrt{2\mu\,(V_0 - B_d)}}{\hbar}\,, \qquad \kappa = \frac{\sqrt{2\mu B_d}}{\hbar}\,, \tag{107.30}$$

with $\mu = m_N/2$ the reduced mass. For $R = 2.0\,\mathrm{fm}$ this gives $V_0 = 36.6\,\mathrm{MeV}$, against a minimum depth of 25.6 MeV needed to bind at all and 2.3×10^2 MeV needed to bind a second s state.

Rests on Equations (107.1) and (107.28).

Proof. *Derives Proposition 107.22.*

Inside the well the regular s-wave radial function is $u \propto \sin kr$; outside it is $u \propto \mathrm{e}^{-\kappa r}$, since a bound state must decay. Continuity of u'/u at $r = R$ is Equation (107.30). Numerically,

$$\kappa = \frac{\sqrt{2 \times 469.46\,\mathrm{MeV} \times 2.2246\,\mathrm{MeV}}}{197.327\,\mathrm{MeV\,fm}} = 0.2316/\mathrm{fm}\,, \tag{107.31}$$

so $1/\kappa = 4.32\,\mathrm{fm}$. Solving Equation (107.30) at $R = 2.0\,\mathrm{fm}$ gives $k = 0.910/\mathrm{fm}$ and $V_0 = B_d + (\hbar c k)^2/2\mu c^2 = 36.6\,\mathrm{MeV}$. The threshold for binding is $kR = \pi/2$ with $\kappa \to 0$, that is $V_0 = \pi^2\hbar^2/8\mu R^2 = 25.6\,\mathrm{MeV}$; the second level needs $kR = 3\pi/2$, nine times as much.

Three conclusions follow, and each is a statement about the force rather than about the model. First, $B_d/V_0 = 0.06$: the well is only about forty per cent deeper than the minimum that binds anything at all, so the deuteron sits just below threshold and its properties are correspondingly sensitive to the force — which is what makes it a good measuring instrument. Second, there is no excited state because there is no room for one, and none is observed. Third, and most striking, $1/\kappa = 4.3\,\mathrm{fm}$ exceeds the range of the force by a factor of two or three: most of the deuteron's probability lies *outside* the region where the two nucleons interact. A bound state that spends most of its time out of range of the force that binds it is the signature of a system barely bound, and it is the reason the low-energy properties of the two-nucleon system can be described without knowing the force in detail — the subject of the next proposition. □

Proposition 107.23 (The deuteron and the triplet scattering length)**.** *The effective-range expansion, whose systematic derivation belongs to Section 86.3.3,*

$$k\cot\delta_0 = -\frac{1}{a} + \frac{1}{2}r_e k^2 + \cdots , \tag{107.32}$$

relates the triplet scattering length a_t to the deuteron binding through $1/a_t = \kappa - \frac{1}{2}r_t\kappa^2$. With $B_d = 2.2246\,\mathrm{MeV}$ and the measured triplet effective range $r_t = 1.75\,\mathrm{fm}$, this predicts $a_t = 5.42\,\mathrm{fm}$, the measured value.

Rests on Equation (107.31) and Phenomenon 86.1.

Proof. *Derives Proposition 107.23.*

A bound state is a pole of the s-wave scattering amplitude at $k = \mathrm{i}\kappa$, where $\cot\delta_0 = \mathrm{i}$; substituting $k = \mathrm{i}\kappa$ in Equation (107.32) gives $-\kappa = -1/a_t - \frac{1}{2}r_t\kappa^2$, that is $1/a_t = \kappa - \frac{1}{2}r_t\kappa^2$. With Equation (107.31),

$$\frac{1}{a_t} = 0.2316/\mathrm{fm} - \tfrac{1}{2}\times 1.75\,\mathrm{fm}\times(0.2316/\mathrm{fm})^2 = 0.1847/\mathrm{fm}\,,$$

so $a_t = 5.42\,\mathrm{fm}$, which is what neutron–proton scattering measures. Two quantities from unrelated experiments — a photodisintegration threshold and a scattering cross-section — are tied by two parameters, and the tie holds. Note that the deuteron enters nowhere in the derivation as a model: only the position of the pole. This is the content of Phenomenon 86.1, that low-energy scattering sees one length and does not care what produced it. □

Derivation pending #168

The effective-range expansion itself: the proof that $k\cot\delta_0$ is an even analytic function of the wave number at low energy, with the scattering length and the effective range as its first two coefficients. It is the premise of the proposition above; its derivation belongs to the low-energy section of the scattering-theory chapter, where it is not yet written.

Phenomenon 107.24 (The neutron–proton cross-section depends violently on spin)**.** The total cross-section for thermal neutrons on hydrogen is $20.5\times10^{-28}\,\mathrm{m}^2$, five and a half times what the triplet interaction alone would give. The singlet scattering length is large and *negative*, $a_s = -23.7\,\mathrm{fm}$, against $a_t = 5.42\,\mathrm{fm}$: the two spin channels of the same pair of particles behave completely differently.

Rests on Phenomenon 86.1 and Equation (107.32).

Derivation. *Derives Phenomenon 107.24.*

Of the four spin states of a neutron and a proton, three are triplet and one singlet, and an unpolarized beam populates them in that ratio. Since $\sigma = 4\pi a^2$ at zero energy (Phenomenon 86.1),

$$\sigma = 4\pi\left(\frac{3}{4}a_t^2 + \frac{1}{4}a_s^2\right) = 4\pi\left[\frac{3}{4}\,(5.42\,\mathrm{fm})^2 + \frac{1}{4}\,(23.7\,\mathrm{fm})^2\right] = 20.4\times10^{-28}\,\mathrm{m}^2\,, \tag{107.33}$$

against the measured $20.5\times10^{-28}\,\mathrm{m}^2$. Had the force been spin independent, $a_s = a_t$ would give $4\pi a_t^2 = 3.7\times10^{-28}\,\mathrm{m}^2$, less by a factor of five and a half. The cross-section was measured long before it was understood, and it is the singlet channel that supplies almost all of it.

The negative sign is the further information. By Equation (107.32) with $r_s = 2.73\,\mathrm{fm}$, solving $1/a_s = \kappa_s - \frac{1}{2}r_s\kappa_s^2$ gives $\kappa_s = -0.0400/\mathrm{fm}$ — the root of the wrong sign, so the pole lies on the unphysical sheet and the state is *virtual*, not bound. Its energy is

$$E_s = \frac{(\hbar c\kappa_s)^2}{2\mu c^2} = 66\,\mathrm{keV}\,, \tag{107.34}$$

just *above* threshold, where the triplet state lies 2.2 MeV below it. Fitting the square well of Proposition 107.22 to a_s at the same radius $R = 2.0\,\mathrm{fm}$ gives a singlet depth of 23.9 MeV against the 25.6 MeV that would just bind: the singlet channel fails by six per cent in the depth of the well, and by 66 keV in the energy. That near-miss is why there is no dineutron and no diproton: two identical nucleons at low energy are confined by the Pauli principle — whose general theory belongs to Chapter 83 — to ${}^1S_0$, which is precisely the channel that just fails to bind. Six per cent more attraction and the universe would contain bound dineutrons, and hydrogen burning in stars would proceed by a strong reaction rather than a weak one — in which case the Sun would have exhausted its fuel long before the Earth formed. □

Proposition 107.25 (The D-state admixture from the magnetic moment)**.** *If the deuteron ground state is* $|d\rangle = \sqrt{1-P_D}\,|{}^3S_1\rangle + \sqrt{P_D}\,|{}^3D_1\rangle$*, then*

$$\mu_d = \mu_p + \mu_n - \frac{3}{2}\left(\mu_p + \mu_n - \frac{1}{2}\mu_N\right)P_D\,, \tag{107.35}$$

and the measured moments give $P_D \approx 0.04$.

Rests on Equation (107.29) and Phenomenon 107.21.

Proof. *Derives Proposition 107.25.*

In a pure ${}^3S_1$ state the orbital motion contributes nothing and the moment is the sum of the intrinsic moments, $\mu_p + \mu_n = (2.79285 - 1.91304)\,\mu_N = 0.87981\,\mu_N$.

In a ${}^3D_1$ state the spins and the orbital motion are coupled to $J = 1$ with $L = 2$ and $S = 1$, and the moment must be projected. Within a manifold of fixed J only the component of a vector operator along $\boldsymbol{J}$ survives — the projection theorem, whose proof belongs to Chapter 79 — so with $\boldsymbol{\mu} = \mu_N\hbar^{-1}\,(g_L\boldsymbol{L} + g_S\boldsymbol{S})$ and

$$\langle\boldsymbol{L}\cdot\boldsymbol{J}\rangle = \tfrac{1}{2}\,[J(J+1) + L(L+1) - S(S+1)]\,\hbar^2 = 3\hbar^2\,,$$
$$\langle\boldsymbol{S}\cdot\boldsymbol{J}\rangle = \tfrac{1}{2}\,[J(J+1) + S(S+1) - L(L+1)]\,\hbar^2 = -\hbar^2\,,$$

the projections in the state of maximal J_z are $\langle L_z\rangle = \frac{3}{2}\hbar$ and $\langle S_z\rangle = -\frac{1}{2}\hbar$, since $J(J+1) = 2$. Two g factors are then needed. The orbital one is $g_L = \frac{1}{2}$, because only the proton carries charge and in the centre-of-mass frame it carries half of the relative orbital angular momentum. The spin one is fixed by the triplet coupling: with $\boldsymbol{\mu}_{\mathrm{spin}} = \mu_N\hbar^{-1}(g_p\boldsymbol{s}_p + g_n\boldsymbol{s}_n)$ and

$g_p = 2\mu_p/\mu_N$, $g_n = 2\mu_n/\mu_N$, projecting on $\boldsymbol{S} = \boldsymbol{s}_p + \boldsymbol{s}_n$ gives $g_S = \tfrac{1}{2}(g_p + g_n) = (\mu_p + \mu_n)/\mu_N$. Hence

$$\mu\left({}^3D_1\right) = \tfrac{1}{2}\cdot\tfrac{3}{2}\mu_N + \frac{\mu_p + \mu_n}{\mu_N}\left(-\tfrac{1}{2}\right)\mu_N = \tfrac{3}{4}\mu_N - \tfrac{1}{2}\left(\mu_p + \mu_n\right) = 0.310\,\mu_N\,,$$

markedly smaller than the S-state value, which is why a few per cent of D state suffices to explain the discrepancy. Weighting the two by $1 - P_D$ and P_D yields Equation (107.35). With $\mu_d = 0.85744\,\mu_N$,

$$P_D = \frac{0.87981 - 0.85744}{\tfrac{3}{2}\,(0.87981 - 0.5)} = 0.039\,,$$

some four per cent. The quadrupole moment, $Q_d = +2.86 \times 10^{-31}\,\mathrm{m}^2$ in units of the elementary charge, gives a compatible figure of four to six per cent depending on the assumed radial functions — and gives it with a sign, the positive value saying the deuteron is prolate, stretched along its spin axis, which is the shape the tensor force Equation (107.29) produces when the spins are aligned along the separation. Two independent static moments, measured by molecular-beam resonance [840], therefore agree on the size of a component of the wave function that no experiment observes directly. The interpretation is not free of assumptions — P_D is not an observable, because it depends on the off-shell convention for the potential — but the tensor force it diagnoses is. □

107.2.4 Phenomenological and effective-field-theory potentials

Phenomenon 107.26 (Charge symmetry and charge independence)**.** After the electromagnetic interaction has been removed, the ${}^1S_0$ scattering lengths of the three nucleon pairs are

$$a_{pp} = -17.3\,\mathrm{fm}\,, \qquad a_{nn} = -18.9\,\mathrm{fm}\,, \qquad a_{np} = -23.7\,\mathrm{fm}\,. \tag{107.36}$$

The first two agree to within one part in ten (*charge symmetry*: the nn force equals the pp force); all three agree to within about a quarter (*charge independence*: the force does not depend on the nucleons' identity at all). Neither is exact.

Rests on Definition 107.10 and Proposition 107.23.

Derivation. *Derives Phenomenon 107.26.*

Charge independence is precisely the statement that the interaction commutes with the isospin operators of Definition 107.10, so that the three $T = 1$ states pp, nn and the symmetric combination of np share one interaction. Charge symmetry is the weaker statement that it commutes with the single rotation by π about the 2-axis in isospin space, which exchanges p and n.

The apparent size of the violation is deceptive, and this is where Proposition 107.23 earns its keep. A scattering length close to a pole is not a linear measure of the force: with the ${}^1S_0$ state a mere 66 keV from threshold, a change of one per cent in the well depth moves a by tens of per cent. Converting Equation (107.36) back into an equivalent well depth through Equation (107.30) shows the underlying interactions differing by about 1 per cent between np and pp, and by a few tenths of a per cent between nn and pp. Two sources of that residual are known and are not adjustable: the electromagnetic interaction beyond the Coulomb term already removed, and the up–down quark mass difference of Chapter 102, which makes $\pi^{\pm}$ and π^0 differ in mass by 4.59 MeV [2427] and therefore in exchange range by 0.048 fm. □

The empirical picture that the scattering data enforce has four ingredients, and every one of them is required by a measurement already quoted.

Definition 107.27 (The phenomenological nucleon–nucleon potential)**.** The static two-nucleon interaction is written

$$V = V_C(r) + V_\sigma(r)\,\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_2 + V_T(r)\,S_{12} + V_{LS}(r)\,\boldsymbol{L}\cdot\boldsymbol{S} + \cdots , \tag{107.37}$$

each radial function carrying a further dependence on $\boldsymbol{\tau}_1\cdot\boldsymbol{\tau}_2$, and each of dimension J. The operators are the complete list of scalars that can be built from $\hat{\boldsymbol{r}}$, $\boldsymbol{\sigma}_1$, $\boldsymbol{\sigma}_2$ and $\boldsymbol{L}$ subject to invariance under rotations, parity and time reversal, up to terms quadratic in $\boldsymbol{L}$.

Remark 107.28 (What each term is doing). V_C is strongly repulsive inside about 0.5 fm — a *hard core*, without which nuclear matter would collapse rather than saturate, since an attraction alone gives an energy growing as A^2 and not as A (Section 107.3.1); the evidence is that the $^1S_0$ phase shift, positive at low energy, changes sign near 250 MeV of laboratory energy, which is what a repulsive core at short distance does. V_σ is demanded by Phenomenon 107.24. V_T is demanded by Phenomenon 107.21. V_{LS} is demanded by the polarization of nucleons scattered from nuclei, and independently by the shell structure of Section 107.4.2, whose magic numbers cannot be reproduced without it. Fitted to the several thousand measured phase shifts below 350 MeV, this form with about forty parameters — the Argonne v_{18} potential [1913] — reproduces the world scattering data with χ^2 per datum near unity. That is an excellent fit and a poor theory: forty parameters are forty things not understood, and the potential has no systematic error estimate because it has no expansion parameter.

Theorem 107.29 (Chiral power counting for the nuclear force)**.** *Let Q denote a momentum of order the pion mass or a nucleon momentum inside a nucleus, and $\Lambda \approx 1\,\mathrm{GeV}/c$ the scale at which the pionic description fails. Then a connected diagram with A nucleons, L loops, C separately connected pieces and vertices of type i carrying d_i derivatives and n_i nucleon fields contributes to the potential at order $(Q/\Lambda)^\nu$ with*

$$\nu = 2\,(L - C + 1) + \sum_i \left(d_i + \frac{n_i}{2} - 2 \right) , \tag{107.38}$$

a sum of non-negative terms [1840]. Consequently many-body forces are suppressed: the leading two-nucleon force enters at $\nu = 0$, the leading three-nucleon force at $\nu = 3$, and the leading four-nucleon force at $\nu = 4$.

Rests on Proposition 102.67 and Theorem 102.70.

Remark 107.30 (What the chiral power counting buys). The counting of Theorem 107.29 is effective-field-theory power counting — a subject that belongs to Chapter 108 — applied to the effective theory whose fields are nucleons and pions and whose symmetry is the spontaneously broken chiral symmetry of Chapter 102. Three things follow that no fit can supply. The long-range part of the force is *predicted*, since at leading order it is exactly the one-pion exchange of Remark 107.17 with the coupling already measured in πN scattering. The short-range part is parametrized by contact terms whose number at each order is finite and known in advance, so the theory has a controlled expansion rather than a free-form fit. And the hierarchy of many-body forces is derived rather than assumed: the three-nucleon force is down by $(Q/\Lambda)^3$, which is the order of a per cent in the amplitude, and this is why two-body forces alone describe light nuclei to a few hundred keV per nucleon and no better. Modern implementations carried to fourth and fifth order fit the same phase shifts as [1913] with comparable quality and, unlike it, with an error band [2177].

The three-nucleon force is not optional. Two-body forces fitted to scattering underbind ^{3}H by about 0.8 MeV out of 8.48 MeV and misplace the saturation point of nuclear matter badly; adding the leading chiral three-body terms, whose two parameters are fixed once from light nuclei, repairs both. A force between three particles that is not the sum of pairwise

forces is unfamiliar from electromagnetism, where it is absent to leading order, and it is unavoidable here for a plain reason: the nucleon is composite, and eliminating its internal degrees of freedom in favour of a potential leaves behind exactly such terms.

Derivation pending #169

Chiral power counting: the derivation of the counting formula for the exponent, by assembling the powers of Q from each propagator, each loop integration and each vertex of the effective Lagrangian, and the demonstration that every bracket in the sum is non-negative for a Lagrangian consistent with chiral symmetry. It belongs in Appendix A, since it needs the full vertex enumeration and runs to several pages; the statement and its consequences are used above without it.

107.3 Nuclear masses and binding

107.3.1 The binding-energy curve

Definition 107.31 (Binding energy and mass defect)**.** The binding energy of a nuclide with Z protons and N neutrons is the energy needed to take it apart into free nucleons at rest,

$$B(Z,N) = \left[Z m_p + N m_n - M(Z,N)\right] c^2 \,, \tag{107.39}$$

in J, with $M(Z,N)$ the mass of the bound nucleus. A bound system weighs less than its parts by B/c^2, the *mass defect*, which is the only place in this book where the relativistic mass–energy relation is used as a routine measuring instrument rather than tested: nuclear masses are measured to keV by Penning traps and mass spectrometry [2393], and the binding energy is read off.

Phenomenon 107.32 (Nuclear matter saturates, and iron is the most tightly bound)**.** The binding energy per nucleon, $B/A = \left[Z m_p + N m_n - M(Z,N)\right] c^2/A$, is nearly constant across the chart of nuclides at about 8 MeV, rising to a maximum near 8.79 MeV in the neighbourhood of ^{56}Fe and falling away on both sides [2393]. Charge radii measured by electron scattering follow $R \approx r_0 A^{1/3}$ with $r_0 \approx 1.2\,\mathrm{fm}$, so the density of nuclear matter is the same in a light nucleus as in a heavy one [1088, 2209]. Both regularities say the same thing — a nucleon binds only to its neighbours, not to every other nucleon — and because the curve has an interior maximum, energy is released both by fusing light nuclei and by splitting heavy ones.

Rests on Definition 107.31.

Derivation. *Derives Phenomenon 107.32.*

Suppose first that every nucleon attracted every other with the same strength. There are $A(A-1)/2$ pairs, so B would grow as A^2 and B/A as A; the binding energy per nucleon of uranium would be sixty times that of carbon. It is not: the two agree within fifteen per cent. The interaction must therefore *saturate*, each nucleon binding to a fixed number of neighbours regardless of how many nucleons the nucleus contains. Two independent facts follow at once and both are observed. The energy is proportional to the number of nucleons, $B \approx a_V A$, and the volume is proportional to the number of nucleons, so the density is constant and $R = r_0 A^{1/3}$.

Saturation is not automatic. A purely attractive short-range force does not saturate: nucleons would simply pile into the same region until the density diverged. Two mechanisms prevent that, and both are established independently. The Pauli principle keeps identical nucleons apart, since at fixed volume the kinetic energy rises as $\rho^{2/3}$ (the Fermi gas of

Proposition 107.36 below); and the short-range repulsive core of Remark 107.28 forbids close packing outright. The observed saturation density is

$$\rho_0 \approx 0.17/\mathrm{fm}^3 = 1.7 \times 10^{44}/\mathrm{m}^3 \,, \tag{107.40}$$

which in mass terms is $2.8 \times 10^{17}\,\mathrm{kg/m^3}$, some fourteen orders of magnitude above the density of water and the same density that supports a neutron star (Chapter 49).

That B/A has an interior maximum then follows from two competing corrections to $B = a_V A$, both of which reduce it. Nucleons near the surface have fewer neighbours, which costs an energy proportional to the surface area, $A^{2/3}$; being a smaller fraction of the whole in a large nucleus, this correction favours large A. The protons repel each other electrostatically, which costs $Z^2/A^{1/3} \sim A^{5/3}$; per nucleon this grows as $A^{2/3}$ and so disfavours large A. One term decreasing and one increasing with A produce a maximum in between, and Proposition 107.41 locates it at $A \approx 61$.

That the maximum exists is the whole energy economy of the chart. Combining two nuclei below it, or dividing one above it, moves nucleons to a more tightly bound configuration and releases the difference. The measured curve is tabulated in Table 107.2; the most tightly bound nuclide is in fact ^{62}Ni at 8.7945 MeV per nucleon, with ^{58}Fe at 8.7923 MeV and ^{56}Fe at 8.7903 MeV [2393] — a spread of two parts in a thousand, so the customary phrase "iron is the most tightly bound nucleus" is very nearly right and exactly wrong. It is ^{56}Fe that dominates the cosmic abundance peak, for reasons of nuclear statistical equilibrium and of the decay of ^{56}Ni that belong to Chapter 52, not because it is the minimum of the mass curve. □

Table 107.2: Binding energy per nucleon, measured [2393] and from the semi-empirical mass formula Equation (107.41) with the coefficients of Equation (107.43). The formula is not intended below $A \approx 20$, where the surface term removes more than the volume term supplies, and ^{2}H is included only to show how badly it fails there. Above $A \approx 20$ the residuals are below one per cent.

nuclide	A	B/A measured (MeV)	B/A formula (MeV)	difference (MeV)
^{2}H	2	1.112	−2.331	−3.443
^{4}He	4	7.074	5.710	−1.364
^{12}C	12	7.680	7.468	−0.212
^{16}O	16	7.976	7.873	−0.103
^{40}Ca	40	8.551	8.614	+0.063
^{56}Fe	56	8.790	8.846	+0.056
^{62}Ni	62	8.794	8.863	+0.069
^{90}Zr	90	8.710	8.749	+0.039
^{120}Sn	120	8.504	8.548	+0.044
^{208}Pb	208	7.867	7.857	−0.010
^{235}U	235	7.591	7.645	+0.054
^{238}U	238	7.570	7.625	+0.055

107.3.2 The semi-empirical mass formula

Von Weizsäcker wrote the nuclear mass as a sum of terms with the A and Z dependences that the liquid-drop picture and the Pauli principle between them dictate [763]; Bethe and Bacher systematized it and published the first fitted coefficients [768]. What follows derives

each term's dependence from the physics that produces it. Only the five coefficients are fitted.

Definition 107.33 (The semi-empirical mass formula)**.**

$$B(Z,N) = a_V A - a_S A^{2/3} - a_C \frac{Z(Z-1)}{A^{1/3}} - a_A \frac{(N-Z)^2}{A} + \delta(Z,N)\,, \tag{107.41}$$

with

$$\delta(Z,N) = \begin{cases} +a_P A^{-1/2} & Z,N \text{ both even}\,, \\ 0 & A \text{ odd}\,, \\ -a_P A^{-1/2} & Z,N \text{ both odd}\,, \end{cases} \tag{107.42}$$

and the coefficients, in MeV,

$$a_V = 15.75\,, \quad a_S = 17.8\,, \quad a_C = 0.711\,, \quad a_A = 23.7\,, \quad a_P = 11.18\,. \tag{107.43}$$

Proposition 107.34 (Origin of the volume and surface terms)**.** *Saturation gives $B = a_V A$; the deficit of neighbours within one force range of the surface gives a correction $-a_S A^{2/3}$ with*

$$a_S = 4\pi r_0^2 \sigma\,, \tag{107.44}$$

σ being a surface energy per unit area. The fitted value corresponds to $\sigma = 0.98\,\mathrm{MeV/fm^2} = 1.58 \times 10^{17}\,\mathrm{J/m^2}$.

Rests on Phenomenon 107.32 and Equation (107.41).

Proof. *Derives Proposition 107.34.*

If each nucleon binds to n neighbours with average energy ε, the total is $\frac{1}{2}nA\varepsilon$, linear in A; that is the volume term, and a_V is not predicted by the argument, only its A dependence. A nucleon within a distance of order the force range λ of the surface has fewer neighbours, and the number of such nucleons is $\rho_0 \times 4\pi R^2 \lambda \propto A^{2/3}$. Writing the deficit as an energy per unit area gives Equation (107.44) with $R = r_0 A^{1/3}$; from $a_S = 17.8\,\mathrm{MeV}$ and $r_0 = 1.2\,\mathrm{fm}$,

$$\sigma = \frac{17.8\,\mathrm{MeV}}{4\pi\,(1.2\,\mathrm{fm})^2} = 0.98\,\mathrm{MeV/fm^2}\,,$$

which converts to $1.58 \times 10^{17}\,\mathrm{J/m^2}$ — eighteen orders of magnitude above the $0.072\,\mathrm{J/m^2}$ of water, and the reason the liquid-drop analogy is quantitative rather than merely pictorial: the same two competing energies, surface tension and Coulomb repulsion, govern a charged raindrop and a fissioning nucleus (Section 107.6.1). □

Proposition 107.35 (The Coulomb term)**.** *The electrostatic energy of Z protons spread uniformly through a sphere of radius $R = r_0 A^{1/3}$ is*

$$E_C = \frac{3}{5}\frac{Z(Z-1)e^2}{4\pi\varepsilon_0 R} = a_C \frac{Z(Z-1)}{A^{1/3}}\,, \qquad a_C = \frac{3}{5}\frac{e^2}{4\pi\varepsilon_0 r_0}\,, \tag{107.45}$$

which for $r_0 = 1.2\,\mathrm{fm}$ evaluates to $0.720\,\mathrm{MeV}$, against the fitted $0.711\,\mathrm{MeV}$.

Rests on Equations (107.2) and (107.41).

Proof. *Derives Proposition 107.35.*

Assemble the sphere by bringing in successive shells from infinity. When the charge assembled is q and the radius is s, with uniform density $q = Q(s/R)^3$, the work to add $\mathrm{d}q$ at radius s is $q\,\mathrm{d}q/4\pi\varepsilon_0 s$. Writing $\mathrm{d}q = 3Qs^2\mathrm{d}s/R^3$,

$$E_C = \frac{1}{4\pi\varepsilon_0}\int_0^R \frac{Qs^3/R^3}{s}\,\frac{3Qs^2}{R^3}\,\mathrm{d}s = \frac{3Q^2}{4\pi\varepsilon_0 R^6}\int_0^R s^4\,\mathrm{d}s = \frac{3}{5}\frac{Q^2}{4\pi\varepsilon_0 R}\,.$$

Setting $Q = Ze$ counts each proton as repelling itself, which is wrong; the correction replaces Z^2 by $Z(Z-1)$, the number of distinct pairs counted twice. Numerically, using Equation (107.2),

$$a_C = \frac{3}{5} \times \frac{1.4400\,\text{MeV fm}}{1.2\,\text{fm}} = 0.720\,\text{MeV}\,,$$

within one and a half per cent of the fitted value. This is the one coefficient of Equation (107.41) that is genuinely *calculated*: it involves no nuclear physics at all, only electrostatics and the measured radius, and its agreement with the fit is a check that the radius entering the mass formula is the radius that electron scattering measures. □

Proposition 107.36 (The Fermi gas and the asymmetry term)**.** *Treat the nucleons as two independent Fermi gases, Z protons and N neutrons, filling a volume $V = A/\rho_0$. Then the Fermi momentum and energy of symmetric nuclear matter are*

$$k_F = \left(\frac{3\pi^2\rho_0}{2}\right)^{1/3} = 1.36/\text{fm}\,, \qquad E_F = \frac{(\hbar c k_F)^2}{2m_N c^2} = 38\,\text{MeV}\,, \tag{107.46}$$

and the excess kinetic energy of an asymmetric system is

$$\Delta E_{\text{kin}} = \frac{E_F}{3}\,\frac{(N-Z)^2}{A}\,, \qquad \frac{E_F}{3} = 12.8\,\text{MeV}\,. \tag{107.47}$$

Rests on Equations (107.1) and (107.40).

Proof. *Derives Proposition 107.36.*

Each species fills momentum states up to its own Fermi momentum. With two spin states per momentum state, the number of one species is $n = V k_{F,n}^3/3\pi^2$, so for symmetric matter $Z = N = A/2$ and $\rho_0 = A/V$ give $k_F = (3\pi^2\rho_0/2)^{1/3}$, which is Equation (107.46) for $\rho_0 = 0.17/\text{fm}^3$; the Fermi energy is $(\hbar k_F)^2/2m_N$, evaluated using Equation (107.1). The total kinetic energy of a filled Fermi sea is

$$E_{\text{kin}} = \frac{3}{5}\left(N\,\frac{\hbar^2 k_{F,n}^2}{2m_N} + Z\,\frac{\hbar^2 k_{F,p}^2}{2m_N}\right) = \frac{3}{5}\,\frac{C}{V^{2/3}}\left(N^{5/3} + Z^{5/3}\right)\,, \qquad C = \frac{\hbar^2}{2m_N}\left(3\pi^2\right)^{2/3}\,,$$

since $k_{F,i} \propto (n_i/V)^{1/3}$. Put $N = \frac{A}{2}(1+x)$ and $Z = \frac{A}{2}(1-x)$ with $x = (N-Z)/A$ and expand: $(1+x)^{5/3} + (1-x)^{5/3} = 2 + \frac{10}{9}x^2 + O(x^4)$. The constant part is the symmetric kinetic energy; the x^2 part is

$$\frac{3}{5}\cdot\frac{C}{V^{2/3}}\left(\frac{A}{2}\right)^{5/3}\cdot\frac{10}{9}x^2 = \frac{2}{3}\cdot\frac{A}{2}\cdot C\left(\frac{A}{2V}\right)^{2/3}x^2 = \frac{A}{3}E_F x^2\,,$$

using $E_F = C(\rho_0/2)^{2/3}$, which is Equation (107.47).

Comparison with the fitted $a_A = 23.7\,\text{MeV}$ is the point of the calculation. The Pauli principle alone accounts for 12.8 MeV, a little more than half, and the remainder must come from the interaction — specifically from the fact, already visible in Phenomenon 107.24, that the isospin-singlet *np* interaction is stronger than the isospin-triplet one, so that a nucleus does better with equal numbers of the two species than the counting of levels alone would suggest. Both halves of the asymmetry term are therefore consequences of the exclusion principle and of isospin, and neither has anything to do with the liquid drop, whose name the term is often filed under. □

Remark 107.37 (The depth of the nuclear well). Equation (107.46) fixes a number used repeatedly in Section 107.4.2. The most weakly bound nucleon sits at the top of the Fermi sea and its separation energy is about 8 MeV, so the single-particle well must have depth

$$V_0 \approx E_F + S_n \approx 38\,\text{MeV} + 8\,\text{MeV} \approx 46\,\text{MeV}\,, \tag{107.48}$$

which is where the conventional 50 MeV of the shell model comes from. Note what has and has not been derived: the *shape* of the asymmetry and the depth of the well follow from Proposition 107.36, but a_V does not. A Fermi gas has positive energy; binding requires the interaction, and the average potential energy per nucleon needed to produce $B/A = 15.75\,\mathrm{MeV}$ against a kinetic $\frac{3}{5}E_F = 23\,\mathrm{MeV}$ is about $-39\,\mathrm{MeV}$. That number is an output of the fit, not of the model.

Remark 107.38 (The pairing term is measured, not derived). δ in Equation (107.42) is the one term with no derivation offered here. Its existence is a direct observation: of the roughly 250 stable nuclides, some 150 have both Z and N even and only four have both odd (^{2}H, ^{6}Li, ^{10}B, ^{14}N). Its origin is that two identical nucleons in the same orbital, coupled to $J = 0$, have maximal spatial overlap and so gain the most from a short-range attraction — the same mechanism, and the same mathematics, as the formation of Cooper pairs in a superconductor, a subject that belongs to Chapter 130. The consequences recur throughout this chapter: it is why even-even nuclei all have $J^P = 0^+$ ground states, why the moment of inertia of a deformed nucleus falls below the rigid-body value (Section 107.4.3), and why ^{235}U fissions with thermal neutrons and ^{238}U does not (Section 107.6.1).

Proposition 107.39 (The valley of beta stability). *At fixed A, the mass Equation (107.41) is minimized at*

$$Z^*(A) = \frac{4a_A + (m_n - m_p)\,c^2 + a_C A^{-1/3}}{8a_A/A + 2a_C A^{-1/3}}\,, \tag{107.49}$$

which for small A tends to $A/2$ and for large A falls well below it.

Rests on Equations (107.5) and (107.41).

Proof. *Derives Proposition 107.39.*

Write $M(Z,A)c^2 = Zm_pc^2 + (A-Z)m_nc^2 - B$ and differentiate at fixed A, using $\partial B/\partial Z = -a_C(2Z-1)A^{-1/3} + 4a_A(A-2Z)/A$ from Equation (107.41) (the pairing term is not differentiable in Z and is treated by comparing the three parities separately). Setting $\partial M/\partial Z = 0$,

$$(m_p - m_n)\,c^2 + a_C\frac{2Z-1}{A^{1/3}} - \frac{4a_A(A-2Z)}{A} = 0\,,$$

which rearranges to Equation (107.49). For small A the $8a_A/A$ term dominates the denominator and $Z^* \to A/2$: the asymmetry term wants equal numbers, and light stable nuclei have them (^{4}He, ^{12}C, ^{16}O, ^{40}Ca). As A grows, the Coulomb term in the denominator takes over and Z^*/A falls: at $A = 56$, $Z^* = 25.6$, so the stable isobar is ^{56}Fe with $Z = 26$; at $A = 120$, $Z^* = 51.5$ against the observed ^{120}Sn with $Z = 50$; at $A = 208$, $Z^* = 83.6$ against ^{208}Pb with $Z = 82$. The systematic overshoot of one to one and a half units at large A is the formula's known weakness, and it is not a rounding accident: it is the absence of shell structure, since both ^{120}Sn ($Z = 50$) and ^{208}Pb ($Z = 82$) are closed-proton-shell nuclei bound more tightly than any smooth formula predicts. That the deviation appears exactly at the magic numbers is the first evidence for Section 107.4.2. □

Corollary 107.40 (Which decay a nuclide chooses). *A nuclide with $Z < Z^*$ has too many neutrons and reaches the valley by β^- decay, $n \to p + e^- + \bar{\nu}_e$; one with $Z > Z^*$ by β^+ decay or electron capture. For odd A the mass parabola is single and there is one stable isobar; for even A the pairing term splits it into two parabolas separated by $2a_PA^{-1/2}$, the even–even one lying lower, so that two or even three stable isobars can coexist and an odd–odd nuclide between them can be unstable in both directions at once.*

Rests on Proposition 107.39 and Equation (107.42).

Proof. *Derives Corollary 107.40.*

The direction follows from the sign of $\partial M/\partial Z$. For the parity structure, note that Equation (107.42) shifts the parabola by $\pm a_PA^{-1/2}$ according to whether Z and N are both

even or both odd, which for even A alternates with every unit step in Z. The even–even isobars therefore lie on a parabola below the odd–odd one by $2a_P A^{-1/2}$, which at $A = 100$ is 2.2 MeV. An odd–odd nuclide with even–even neighbours on both sides can then decay to either, which is exactly what ^{40}K does — β^- to ^{40}Ca in 89 per cent of decays and electron capture to ^{40}Ar in 11 per cent [2382] — and it is why $A = 40$ has two stable isobars while $A = 39$ has one. The four odd–odd stable nuclides of Remark 107.38 are the exceptions, all at $A \leq 14$ where the parabolas are too shallow for the argument to bite. □

Proposition 107.41 (Where the binding curve peaks)**.** *Along the valley Equation* (107.49)*, and dropping the pairing term, B/A from Equation* (107.41) *has its maximum at*

$$A_{\max} = 61\,, \qquad Z^* = 27.7\,, \qquad \left(\frac{B}{A}\right)_{\max} = 8.84\,\mathrm{MeV}\,, \tag{107.50}$$

against the measured maximum at ^{62}Ni, $A = 62$, $Z = 28$, $B/A = 8.7945$ MeV.

Rests on Equations (107.41) *and* (107.49)*.*

Proof. *Derives Proposition 107.41.*

Substitute $Z = Z^*(A)$ from Equation (107.49) into Equation (107.41), divide by A, and maximize numerically in A; the function is smooth and has a single interior maximum. The competition is transparent in the two dominant A-dependent terms per nucleon, $-a_S A^{-1/3}$, which rises towards zero as A grows, and $-a_C Z^{*2} A^{-4/3}$, which falls; the asymmetry term per nucleon, $-a_A(1 - 2Z^*/A)^2$, also falls with A because Z^*/A departs from $\frac{1}{2}$. The maximum is broad — B/A is within 0.01 MeV of its peak over the whole range $A = 54$ to $A = 69$ — which is why the identity of the most bound nuclide is decided by shell effects of a few tenths of a MeV that Equation (107.41) does not contain, and why it is ^{62}Ni, with $Z = 28$ magic, that wins. The formula gets the position of the peak right to one unit in A and its height right to half a per cent, which is all a five-parameter fit can be asked for. □

Remark 107.42 (Reading the residuals)*.* Table 107.2 shows the formula failing badly below $A \approx 20$ — for ^{2}H it predicts a negative binding energy, and ^{4}He is underbound by 5.5 MeV out of 28.3 MeV — and succeeding to better than one per cent above it. The failure at small A is not mysterious: the surface term is a correction of relative size $A^{-1/3}$, which at $A = 4$ is 0.63, so the expansion it belongs to has no small parameter there.

What matters is the structure of the residuals above $A = 20$. They are not noise. Plotted against N or Z they show sharp excursions at 2, 8, 20, 28, 50, 82 and 126, and the excursions have a consistent sign: nuclei at those numbers are bound more than the formula allows. That pattern was visible in the earliest fits, and it is the whole content of Phenomenon 107.51. Reading a regularity out of the residuals of a fit is how nuclear structure was found, and it is worth naming as a method: a formula that is right to one per cent is most useful for the one per cent it gets wrong.

107.3.3 Radii, form factors and the charge distribution

Everything above assumed $R = r_0 A^{1/3}$. That is a measurement, and the instrument is the electron: structureless, interacting only electromagnetically through a coupling small enough that one photon exchange suffices, and available at wavelengths short compared with a nucleus. Hofstadter turned it into a quantitative map of the nuclear charge density [1088].

Theorem 107.43 (Scattering measures the charge form factor)**.** *In the one-photon-exchange approximation, elastic scattering of electrons from a static charge distribution $\rho(\boldsymbol{r})$ normalized to $\int \rho\, \mathrm{d}^3 r = Ze$ gives*

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \left(\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}\right)_{\text{point}} |F(\boldsymbol{q})|^2\,, \qquad F(\boldsymbol{q}) = \frac{1}{Ze}\int \mathrm{d}^3 r\; \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}/\hbar}\rho(\boldsymbol{r})\,, \tag{107.51}$$

with $\boldsymbol{q}$ the momentum transferred, of dimension kg m/s. *For a spherical distribution*

$$F(q) = 1 - \frac{q^2 \langle r^2 \rangle}{6\hbar^2} + O(q^4)\,, \tag{107.52}$$

so the initial slope of the measured form factor is the mean square charge radius.
Rests on Equation (107.10).

Proof. *Derives Theorem 107.43.*

The Born amplitude for scattering off a potential V is proportional to its Fourier transform, and here $V(\boldsymbol{r}) = -e\phi(\boldsymbol{r})$ with ϕ the electrostatic potential of ρ. Since $\nabla^2\phi = -\rho/\varepsilon_0$, the transforms obey $\tilde{\phi}(\boldsymbol{q}) = \hbar^2\tilde{\rho}(\boldsymbol{q})/\varepsilon_0\boldsymbol{q}^2$, so the amplitude is the point-charge amplitude — which is what gives the Rutherford or Mott cross-section — times $\tilde{\rho}(\boldsymbol{q})/Ze$, and the cross-section is its modulus squared. That is Equation (107.51). For a spherical ρ, doing the angular integral,

$$F(q) = \frac{4\pi\hbar}{Zeq}\int_0^\infty r\rho(r)\sin\!\left(\frac{qr}{\hbar}\right)\mathrm{d}r\,,$$

and expanding $\sin u = u - u^3/6 + \cdots$ gives Equation (107.52) with $\langle r^2\rangle = \int r^2\rho\,\mathrm{d}^3r/Ze$. The Fourier inversion that recovers ρ from F is the transform pair of Chapter 17; in practice F is measured over a finite range of q, so the inversion is not unique and the data are fitted to a parametrized shape instead. □

Corollary 107.44 (Diffraction minima give the radius directly)**.** *For a uniformly charged sphere of radius R,*

$$F(q) = \frac{3}{x^3}\left(\sin x - x\cos x\right)\,, \qquad x = \frac{qR}{\hbar}\,, \tag{107.53}$$

whose first zero is at $x = 4.493$. A minimum in the measured angular distribution at momentum transfer q_1 therefore gives $R = 4.493\,\hbar/q_1$.
Rests on Theorem 107.43.

Proof. *Derives Corollary 107.44.*

Insert $\rho = 3Ze/4\pi R^3$ for $r < R$ in the radial integral above and integrate by parts twice. The zeros of $\sin x - x\cos x$, that is of $\tan x = x$, are at $x = 4.493, 7.725, \ldots$, and the diffraction pattern they produce is the same phenomenon as the Airy pattern of a circular aperture: a sharp-edged object has a form factor with zeros, a diffuse one has minima that do not reach zero. That the observed minima are filled in is the first evidence for a surface of finite thickness, quantified next. □

Definition 107.45 (The two-parameter Fermi distribution)**.**

$$\rho(r) = \frac{\rho_0}{1 + \exp\left[(r-c)/a\right]}\,, \tag{107.54}$$

with c the half-density radius and a the diffuseness. The *surface thickness*, the distance over which the density falls from 90 to 10 per cent of ρ_0, is $t = 4a\ln 3$.

Phenomenon 107.46 (Constant central density, constant surface)**.** Fits of Equation (107.54) to electron-scattering data across the chart give

$$c \approx 1.1\,\mathrm{fm} \times A^{1/3}\,, \qquad a \approx 0.54\,\mathrm{fm}\,, \qquad t \approx 2.4\,\mathrm{fm}\,, \tag{107.55}$$

with a central density ρ_0 near $0.17/\mathrm{fm}^3$ of nucleons that does not depend on A [1088]. The surface thickness is the same for carbon and for lead.
Rests on Equation (107.54) *and Theorem 107.43.*

Derivation. *Derives Phenomenon 107.46.*

Normalizing Equation (107.54) to A nucleons gives, to the accuracy of an expansion in a/c,

$$A = \rho_0 \frac{4\pi c^3}{3}\left[1 + \left(\frac{\pi a}{c}\right)^2\right],$$

so for ^{208}Pb, with $c = 6.52\,\text{fm}$ and $a = 0.54\,\text{fm}$, $\rho_0 = 0.168/\text{fm}^3$, which is Equation (107.40). The relation to the equivalent-uniform-sphere radius $R = r_0 A^{1/3}$ used in Equation (107.41) is the mean square radius: a uniform sphere has $\langle r^2\rangle = \frac{3}{5}R^2$, so

$$r_0 = \sqrt{\frac{5}{3}}\,\frac{\langle r^2\rangle^{1/2}}{A^{1/3}}, \tag{107.56}$$

and Table 107.3 evaluates it. That r_0 is 1.20 fm for lead but 1.39 fm for carbon is not an inconsistency: with a surface of fixed thickness 2.4 fm, a carbon nucleus of radius 2.5 fm is essentially all surface, and its equivalent uniform sphere is correspondingly larger than its half-density radius. The constant that is really constant is ρ_0, not r_0. □

Table 107.3: Root-mean-square charge radii from the compilation of [2209], with the equivalent-uniform-sphere parameter r_0 from Equation (107.56). The trend in r_0 is the effect of the finite surface thickness, not a failure of saturation.

nuclide	A	$\langle r^2\rangle^{1/2}$ (fm)	r_0 (fm)
^{12}C	12	2.4702	1.393
^{40}Ca	40	3.4776	1.313
^{56}Fe	56	3.7377	1.261
^{208}Pb	208	5.5012	1.199

Remark 107.47 (Two other rulers, and a discrepancy). Electron scattering is not the only measurement of a nuclear size. An atomic level is shifted by the finite size of the nucleus, by an amount proportional to $|\psi(0)|^2\,\langle r^2\rangle$, so optical isotope shifts measure *differences* of $\langle r^2\rangle$ along an isotopic chain with great precision. Replacing the electron by a muon multiplies $|\psi(0)|^2$ by $(m_\mu/m_e)^3 \approx 8.9\times10^6$ and turns the shift into the dominant term, which is what makes muonic-atom spectroscopy the most accurate absolute method.

Applied to the proton itself the two rulers disagreed. Muonic hydrogen gave a charge radius $(0.841\,84 \pm 0.000\,67)$ fm [2164, 2210], some five standard deviations below the then-accepted value near 0.8775 fm from ordinary hydrogen spectroscopy and electron scattering. The resolution was not new physics: later measurements of the ordinary-hydrogen Lamb shift [2331] and a re-analysis of low-q electron scattering [2344] both moved to the smaller value, and the CODATA 2022 recommendation is $(8.4075 \pm 0.0064)\times10^{-16}$ m [2431]. The episode is a worked case of systematic error rather than of discovery; its full experimental record belongs to Chapter 90.

107.4 Nuclear models

Two pictures of a nucleus are in flat contradiction, and both work. In one, the nucleons are so strongly coupled that no individual motion survives and the nucleus behaves as a drop of incompressible charged fluid; that is the picture behind Equation (107.41), and it accounts for masses, for fission and for the compound nucleus. In the other, each nucleon moves independently in an average field produced by all the others, with well-defined orbitals and

shell closures; that is what Phenomenon 107.51 demands. Reconciling them is the subject of Section 107.4.3, and the reconciliation turns on one point: the Pauli principle removes almost all of the collisions that the strong coupling would otherwise cause.

107.4.1 The liquid-drop model and the compound nucleus

The drop analogy is already quantitative in Proposition 107.34: an incompressible fluid of fixed density with a surface energy $4\pi R^2\sigma$ reproduces the first two terms of Equation (107.41) with a surface tension $1.58\times10^{17}\,\mathrm{J/m^2}$. What the analogy adds beyond the mass formula is dynamics — the drop can vibrate, deform and divide — and a picture of what happens when a projectile is absorbed.

Phenomenon 107.48 (Slow-neutron capture makes long-lived states)**.** A heavy nucleus absorbing a slow neutron shows, as a function of the neutron energy, a dense forest of sharp resonances: in ^{238}U the first lies at 6.67 eV, the spacing between neighbours is of order 20 eV, and the widths are of order 25 meV. A width that small corresponds to a lifetime of order $2.6\times10^{-14}\,\mathrm{s}$, some 10^8 times the time a nucleon takes to cross the nucleus. Moreover the products of the decay do not depend on how the compound system was made: the same excited nucleus reached by neutron capture and by a different reaction decays in the same proportions to the same channels [770].

Rests on Equation (107.46).

Derivation. *Derives Phenomenon 107.48.*

The two observations are one statement. A width Γ corresponds to a mean life $\tau=\hbar/\Gamma$; for $\Gamma=25\,\mathrm{meV}$,

$$\tau=\frac{6.582\times10^{-16}\,\mathrm{eV\,s}}{25\times10^{-3}\,\mathrm{eV}}=2.6\times10^{-14}\,\mathrm{s}\,,$$

whereas a nucleon moving at the Fermi velocity $v_F=\hbar k_F/m_N=0.29\,c$ from Equation (107.46) crosses a lead-sized nucleus in $2R/v_F=1.7\times10^{-22}\,\mathrm{s}$. The incident neutron therefore does not pass through: it is absorbed, its 6 MeV or so of binding energy is shared among all A nucleons, and the system wanders through an enormous number of configurations before, by a statistical fluctuation, enough energy is again concentrated on one nucleon for it to escape. That is Bohr's compound nucleus. Its signature consequence is factorization,

$$\sigma(a\to b)=\sigma_{\mathrm{form}}(a)\,\frac{\Gamma_b}{\Gamma}\,,\qquad \Gamma=\sum_c\Gamma_c\,, \tag{107.57}$$

the compound nucleus having forgotten how it was formed — which is exactly what the branching-ratio measurements show, and what the resonance formalism that belongs to Section 86.5.1 encodes as a pole whose residue splits into an entrance and an exit factor. □

Proposition 107.49 (The $1/v$ law for slow-neutron capture)**.** *Far below an isolated capture resonance, the cross-section for (n,γ) varies as the reciprocal of the neutron velocity.*

Rests on Phenomenon 107.48 and Equation (86.2).

Proof. *Derives Proposition 107.49.*

The Breit–Wigner cross-section for capture through a single resonance at E_0 is

$$\sigma_{n\gamma}(E)=\frac{\pi\hbar^2}{p^2}\,g\,\frac{\Gamma_n(E)\,\Gamma_\gamma}{(E-E_0)^2+\Gamma^2/4}\,, \tag{107.58}$$

with $p=\sqrt{2m_nE}$ the neutron momentum and g a statistical factor. Two energy dependences matter at small E. The prefactor $\pi\hbar^2/p^2$ goes as $1/E$. The neutron width is a transmission probability through the centrifugal and surface barriers and for s waves is proportional to the momentum, $\Gamma_n\propto p\propto\sqrt{E}$, which is the threshold law of Equation (86.2) in another

guise. The photon width Γ_γ and the denominator are both essentially constant for $E \ll E_0$. Hence

$$\sigma_{n\gamma} \propto \frac{1}{E} \cdot \sqrt{E} = \frac{1}{\sqrt{E}} \propto \frac{1}{v} .$$

The law holds over many decades below the first resonance and is what makes a moderated reactor work: slowing a neutron from 2 MeV to the thermal 0.025 eV raises its capture and fission cross-sections by a factor of $\sqrt{8 \times 10^7} \approx 9 \times 10^3$ (Section 107.6.1). It also means that a tabulated "thermal cross-section" is a convention, quoted at $v = 2200\,\mathrm{m/s}$, and not a property of the nucleus alone. □

Derivation pending #170

The Breit–Wigner formula: the derivation of the single-level resonance cross-section used above, with its partial widths and the statistical spin factor g, from the pole structure of the S matrix. It is taken here as a premise; its derivation belongs to the resonance section of the scattering-theory chapter, where it is not yet written.

Remark 107.50 (What the drop cannot do). The liquid drop has no single-particle levels, so it cannot produce a magic number, a ground-state spin, a magnetic moment or a low-lying excited state at a particular energy. It also predicts, through Equation (107.41), a smooth mass surface, and the residuals of Remark 107.42 show that the smooth surface is wrong by up to 10 MeV in total binding at the closed shells. The two pictures are not alternatives to be chosen between; they describe different degrees of freedom of the same object, and Section 107.4.3 couples them.

107.4.2 Magic numbers and the shell model

Phenomenon 107.51 (Certain nucleon numbers are anomalously stable)**.** At proton or neutron numbers $2, 8, 20, 28, 50, 82$ and 126 nuclei are more tightly bound than the smooth mass formula predicts, are more abundant in Nature [1098], have anomalously small neutron-capture cross-sections, carry their first excited state unusually high, and have vanishing electric quadrupole moments [933, 940]. The pattern is that of closed shells, as in the noble gases; but it is not the pattern of any simple one-body well, since a harmonic oscillator or a square well gives the first three numbers and then none of the rest.

Rests on Equations (107.41) *and* (107.48).

Derivation. *Derives Phenomenon 107.51.*

The derivation is in two steps: a mean field, which gets three of the seven numbers, and a spin–orbit term, which gets all seven.

Step one: the oscillator. Take the nucleons to move independently in a common potential. Its shape is bounded between two solvable cases — a harmonic oscillator, which is too soft at the edge, and an infinite square well, which is too hard — and the real Woods–Saxon well, of the shape Equation (107.54) with depth Equation (107.48), lies between them. For the isotropic oscillator the levels are

$$E_{\mathcal{N}} = \hbar\omega \left(\mathcal{N} + \frac{3}{2} \right) , \qquad \mathcal{N} = 2(n-1) + \ell \, , \tag{107.59}$$

so a given $\mathcal{N}$ collects $\ell = \mathcal{N}, \mathcal{N}-2, \ldots$, and, counting $2(2\ell+1)$ states per level for the two spin projections, has degeneracy $(\mathcal{N}+1)(\mathcal{N}+2)$. The cumulative occupancies are therefore

$$2,\ 8,\ 20,\ 40,\ 70,\ 112,\ 168\,, \tag{107.60}$$

which reproduces 2, 8 and 20 and then fails: the observed numbers are 28, 50, 82, 126. A square well gives $2, 8, 18, 20, 34, 40, \ldots$ and fails in a different way. No choice of radial shape repairs this, which is why the problem stood for fifteen years.

The oscillator constant itself follows from the size. Each nucleon species fills its own well, so for a symmetric nucleus filled up to shell $\mathcal{N}_F$ the cumulative degeneracy gives $A/2 \approx \mathcal{N}_F^3/3$, and the virial theorem gives $\langle r^2 \rangle = \frac{3}{4}\mathcal{N}_F\hbar/m_N\omega$; equating this to the $\frac{3}{5}r_0^2 A^{2/3}$ of a uniform sphere,

$$\hbar\omega = \frac{5}{4}\left(\frac{3}{2}\right)^{1/3} \frac{\hbar^2}{m_N r_0^2}\, A^{-1/3} \approx \frac{41\,\mathrm{MeV}}{A^{1/3}}\,, \tag{107.61}$$

the familiar empirical value, which for ^{208}Pb is 6.9 MeV and for ^{16}O is 16 MeV.

Step two: the spin–orbit term. Mayer [940] and, independently, Haxel, Jensen and Suess [933] added

$$V_{LS} = -C\,\boldsymbol{L}\cdot\boldsymbol{S}\,, \qquad C > 0\,, \tag{107.62}$$

with C carrying whatever dimension makes $C\hbar^2$ an energy, since $\boldsymbol{L}\cdot\boldsymbol{S}$ has dimension $\mathrm{J}^2\,\mathrm{s}^2$. Since $\boldsymbol{L}\cdot\boldsymbol{S} = \frac{1}{2}\left(\boldsymbol{J}^2 - \boldsymbol{L}^2 - \boldsymbol{S}^2\right)$,

$$\langle \boldsymbol{L}\cdot\boldsymbol{S}\rangle = \begin{cases} +\frac{1}{2}\ell\hbar^2\,, & j = \ell + \frac{1}{2}\,, \\ -\frac{1}{2}(\ell+1)\hbar^2\,, & j = \ell - \frac{1}{2}\,, \end{cases} \tag{107.63}$$

so the two members of a spin–orbit doublet are separated by

$$\Delta E_{LS} = \tfrac{1}{2}C\,(2\ell+1)\,\hbar^2\,, \tag{107.64}$$

with the level of *higher* j pushed *down*. Two features are essential and both are contrary to the atomic case: the sign, so that $j = \ell + \frac{1}{2}$ is the more bound, and the magnitude, which is comparable with $\hbar\omega$ rather than a fine-structure correction — a factor of order $(v/c)^{-2} \sim 10^4$ larger, because this spin–orbit force is a property of the strong interaction (Remark 107.28) and not a relativistic correction to an electromagnetic one.

Since Equation (107.64) grows with ℓ, the highest-ℓ level of each oscillator shell is depressed the most. For $\ell \geq 3$ the depression exceeds the shell gap $\hbar\omega$ and that level drops out of its shell into the gap below. Theorem 107.52 does the arithmetic. □

Theorem 107.52 (The magic numbers)**.** *Let the level of highest total angular momentum in oscillator shell $\mathcal{N}$, namely $j = \mathcal{N} + \frac{1}{2}$ with $\ell = \mathcal{N}$, be depressed below the shell gap. Then the closed-shell occupancies are*

$$M_{\mathcal{N}} = \frac{\mathcal{N}(\mathcal{N}+1)(\mathcal{N}+2)}{3} + 2\mathcal{N} + 2\,, \tag{107.65}$$

for $\mathcal{N} \geq 3$, and the unmodified oscillator closures $2, 8, 20$ for $\mathcal{N} \leq 2$. These are

$$2,\ 8,\ 20,\ 28,\ 50,\ 82,\ 126\,, \tag{107.66}$$

with 184 *predicted next.*

Rests on Equations (107.60) *and* (107.64).

Proof. *Derives Theorem 107.52.*

The first term of Equation (107.65) is the cumulative oscillator degeneracy through shell $\mathcal{N}-1$,

$$\sum_{k=0}^{\mathcal{N}-1} (k+1)(k+2) = \frac{\mathcal{N}(\mathcal{N}+1)(\mathcal{N}+2)}{3}\,,$$

verified by induction. The intruder from shell $\mathcal{N}$ carries $2j+1 = 2\mathcal{N}+2$ states, which is the second term. A gap opens above the intruder because the rest of shell $\mathcal{N}$ remains $\hbar\omega$ higher.

Evaluating: $\mathcal{N}=3$ gives $20+8=28$ (the intruder is $1f_{7/2}$); $\mathcal{N}=4$ gives $40+10=50$ ($1g_{9/2}$); $\mathcal{N}=5$ gives $70+12=82$ ($1h_{11/2}$); $\mathcal{N}=6$ gives $112+14=126$ ($1i_{13/2}$); $\mathcal{N}=7$ gives $168+16=184$ ($1j_{15/2}$). For $\mathcal{N}\le 2$ the formula would give 6 and 14; those are not observed, and correctly so, because Equation (107.64) for $\ell\le 2$ is smaller than $\hbar\omega$ and the intruder does not leave its shell. The threshold at $\ell=3$ is thus not fitted but forced: it is the condition $\tfrac{1}{2}C(2\ell+1)\hbar^2>\hbar\omega$ with the single constant C chosen once. Every observed magic number is then reproduced, and no unobserved one is produced. That is the whole argument, and it is why the two 1949 papers ended a fifteen-year impasse in three pages each. □

Table 107.4: The single-particle levels in order of increasing energy, grouped by the gaps that produce Equation (107.66). Each level $n\ell_j$ holds $2j+1$ nucleons. The intruder that closes each shell above 20 is the high-j level driven down from the oscillator shell above by Equation (107.62).

levels	states added	cumulative
$1s_{1/2}$	2	2
$1p_{3/2}$, $1p_{1/2}$	6	8
$1d_{5/2}$, $2s_{1/2}$, $1d_{3/2}$	12	20
$1f_{7/2}$	8	28
$2p_{3/2}$, $1f_{5/2}$, $2p_{1/2}$, $1g_{9/2}$	22	50
$1g_{7/2}$, $2d_{5/2}$, $2d_{3/2}$, $3s_{1/2}$, $1h_{11/2}$	32	82
$1h_{9/2}$, $2f_{7/2}$, $2f_{5/2}$, $3p_{3/2}$, $3p_{1/2}$, $1i_{13/2}$	44	126

Proposition 107.53 (Why independent-particle motion survives)**.** *A nucleon inside a nucleus has a mean free path against collision with its neighbours that is short compared with the nuclear radius if computed from the free nucleon–nucleon cross-section, and long compared with it once the Pauli principle is imposed.*

Rests on Equations (107.40) *and* (107.46).

Proof. *Derives Proposition 107.53.*

The free cross-section at a laboratory energy of a few tens of MeV is about $2\times10^{-29}\,\mathrm{m}^2$, so at the density Equation (107.40) the naive mean free path is

$$\ell=\frac{1}{\rho_0\sigma}=\frac{1}{1.7\times10^{44}/\mathrm{m}^3\times2\times10^{-29}\,\mathrm{m}^2}=2.9\times10^{-16}\,\mathrm{m}=0.29\,\mathrm{fm}\,,$$

a quarter of the internucleon spacing. On that estimate no orbital could survive for even one revolution and the shell model would be absurd — which is exactly the objection raised against it in 1949.

The estimate omits the exclusion principle. A collision must take both partners to unoccupied states, and below the Fermi surface there are none. Let a nucleon of energy E above E_F collide with one of the filled sea; energy conservation confines the two final energies to a shell of thickness E above E_F, and momentum conservation confines the initial partner similarly, so the fraction of collisions that is kinematically allowed is of order

$$\frac{\Gamma}{\Gamma_{\mathrm{free}}}\sim\left(\frac{E-E_F}{E_F}\right)^2\,, \tag{107.67}$$

the same phase-space counting that makes a Fermi liquid's quasiparticles long-lived, a subject that belongs to Chapter 126. For a nucleon near the top of the filled sea, $E-E_F$ is a few MeV against $E_F=38\,\mathrm{MeV}$, so the factor is of order 10^{-2} and the mean free path becomes tens of femtometres — larger than the nucleus. The paradox dissolves completely:

the nucleons that matter for the ground state are exactly those for which collisions are forbidden, and a strongly interacting system can nevertheless be an almost ideal gas of independent particles. Note that the argument fails for a highly excited nucleus, where $E - E_F$ is large; and there the compound-nucleus picture of Phenomenon 107.48, which assumes the opposite extreme, takes over. The two models describe the same nucleus at different excitation energies. □

Corollary 107.54 (Ground-state spins and parities)**.** *An even–even nucleus has $J^P = 0^+$; a nucleus with one nucleon outside closed shells, or one hole in them, has the j and the parity $(-1)^\ell$ of that level.*

Rests on Theorem 107.52 and Equation (107.42).

Proof. *Derives Corollary 107.54.*

Pairing (Remark 107.38) couples identical nucleons in the same level to $J = 0$ with even parity, so a filled or paired configuration contributes nothing, and every one of the several hundred known even–even ground states is 0^+ — a prediction with no exceptions. For an odd-A nucleus the unpaired nucleon carries the spin. Reading Table 107.4: ^{17}O has one neutron in $1d_{5/2}$, predicting $\frac{5}{2}^+$, observed $\frac{5}{2}^+$; ^{15}N has a proton hole in $1p_{1/2}$, predicting $\frac{1}{2}^-$, observed $\frac{1}{2}^-$; ^{41}Ca has a neutron in $1f_{7/2}$, predicting $\frac{7}{2}^-$, observed $\frac{7}{2}^-$; ^{207}Pb has a neutron hole in $3p_{1/2}$, predicting $\frac{1}{2}^-$, observed $\frac{1}{2}^-$; ^{209}Bi has a proton in $1h_{9/2}$, predicting $\frac{9}{2}^-$, observed $\frac{9}{2}^-$. The rule works best next to closed shells and degrades in mid-shell, where several nucleons are unpaired and the coupling is a genuine many-body problem. □

Proposition 107.55 (The Schmidt lines)**.** *For a single nucleon of orbital angular momentum ℓ and total j, with $g_\ell = 1$ for a proton and 0 for a neutron and $g_s = +5.586$ for a proton and -3.826 for a neutron, the magnetic moment is*

$$\frac{\mu}{\mu_N} = \begin{cases} \left(j - \tfrac{1}{2}\right) g_\ell + \tfrac{1}{2} g_s \,, & j = \ell + \tfrac{1}{2} \,, \\ \dfrac{j}{j+1}\left[\left(j + \tfrac{3}{2}\right) g_\ell - \tfrac{1}{2} g_s\right] \,, & j = \ell - \tfrac{1}{2} \,. \end{cases} \tag{107.68}$$

Measured moments of odd-A nuclei fall between *the two curves rather than on them.*

Rests on Proposition 107.25 and Corollary 107.54.

Proof. *Derives Proposition 107.55.*

Write $\boldsymbol{\mu} = \mu_N \hbar^{-1} (g_\ell \boldsymbol{L} + g_s \boldsymbol{S})$ and take the expectation in the state of maximal j_z. Since $\boldsymbol{L}$ and $\boldsymbol{S}$ are not separately conserved, project each on $\boldsymbol{J}$ by the same projection theorem used in Proposition 107.25, that within a manifold of fixed j only the component of a vector operator along $\boldsymbol{J}$ contributes: $\langle \boldsymbol{L} \rangle = \langle \boldsymbol{L} \cdot \boldsymbol{J} \rangle \, \boldsymbol{J} / j(j+1)\hbar^2$ and likewise for $\boldsymbol{S}$, with $\langle \boldsymbol{L} \cdot \boldsymbol{J} \rangle = \tfrac{1}{2}\left[j(j+1) + \ell(\ell+1) - s(s+1) \right] \hbar^2$. Substituting $\ell = j \mp \tfrac{1}{2}$ and simplifying gives Equation (107.68).

The two branches bracket the data, and the bracketing is itself the result. For a neutron with $j = \ell + \tfrac{1}{2}$ the prediction is $\tfrac{1}{2} g_s = -1.913$ regardless of ℓ; ^{17}O measures -1.894, within one per cent. But ^{209}Bi, a $1h_{9/2}$ proton with $j = \ell - \tfrac{1}{2}$, is predicted $+2.62$ and measures $+4.11$, which lies between that value and the $+6.79$ of the other branch. Schmidt's compilation [800] showed this pattern for every measured moment before the shell model existed. The interpretation is that the odd nucleon is not alone: its orbital is mixed with configurations in which the core is excited, and in addition the nucleons' intrinsic moments are modified inside nuclear matter by the exchange currents that the pion of Section 107.2.2 carries. The single-particle model is therefore right about the *spins* (Corollary 107.54) and only qualitatively right about the *moments*, and the difference between the two is a quantitative measure of how far the independent-particle approximation is from exact. □

107.4.3 The collective model

Phenomenon 107.56 (Deformed nuclei have rotational spectra)**.** Away from closed shells, even–even nuclei show a sequence of excited states with spins $0^+, 2^+, 4^+, 6^+, \ldots$ whose energies are proportional to $J(J+1)$. In ^{238}U the 2^+ lies at 44.9 keV and the 4^+ at 148.4 keV, a ratio of 3.30 against the rigid-rotor value $10/3 = 3.33$. Their electric quadrupole moments are one to two orders of magnitude larger than any single nucleon could produce, so many nucleons are moving together.

Derivation. *Derives Phenomenon 107.56.*

A rigid axially symmetric body has Hamiltonian $\mathcal{H} = \boldsymbol{J}^2/2\mathcal{I}$ about the axes perpendicular to its symmetry axis, so

$$E_J = \frac{\hbar^2}{2\mathcal{I}}\, J(J+1)\,, \tag{107.69}$$

in J, with $\mathcal{I}$ in $\mathrm{kg\,m^2}$. Only even J occurs, because a reflection through the plane perpendicular to the symmetry axis is a symmetry of the shape and the rotational wave function must be even under it, which kills the odd members. The ratio $E_4/E_2 = 20/6 = 3.33$ is then a parameter-free prediction, against the 2.0 that a harmonic surface vibration would give — so the single measured ratio distinguishes the two collective modes.

The moment of inertia is a measurement. From $E_2 = 44.9\,\mathrm{keV}$,

$$\frac{\hbar^2}{2\mathcal{I}} = \frac{E_2}{6} = 7.5\,\mathrm{keV}\,, \qquad \mathcal{I} = 4.6 \times 10^{-54}\,\mathrm{kg\,m^2}\,,$$

whereas a rigid sphere of mass $238\,u$ and radius $r_0 A^{1/3} = 7.4\,\mathrm{fm}$ has $\frac{2}{5}MR^2 = 8.7 \times 10^{-54}\,\mathrm{kg\,m^2}$. The observed moment is about half the rigid-body value — and, at the other extreme, several times the value for irrotational flow of a deformed drop, in which only the surface deformation and not the matter travels round. □

Remark 107.57 (Why the moment of inertia is between the two limits). That $\mathcal{I}$ falls between rigid and irrotational is the central quantitative fact of the collective model, and its explanation is pairing. A paired system has a gap: to change the motion of a single nucleon costs at least 2Δ, of order 1 MeV by Equation (107.42), so slow rotation cannot excite individual nucleons and part of the nuclear fluid simply does not rotate. This is the nuclear counterpart of the non-classical rotational inertia of a superfluid, and the same mathematics applies: the reduction factor is governed by the ratio of the rotational frequency to the pairing gap. Two independent consequences confirm it. The moment of inertia *increases* with angular momentum along a band and jumps sharply at a critical rotation — the pairing is being broken; and the odd-A neighbours of an even–even rotor have systematically larger moments, because the odd nucleon is already unpaired.

Remark 107.58 (Rainwater's argument and the unified model). Why should a nucleus be deformed at all, when a drop with surface tension is spherical? Rainwater's answer [967] is that the single-particle levels of Table 107.4 respond to deformation: a partly filled high-j level splits, and the nucleons occupy the lowered members, so the energy gained is linear in the deformation while the surface energy lost is quadratic. A linear gain always beats a quadratic cost for small enough deformation, so a nucleus with a partly filled shell is *always* deformed, and one with a closed shell — where there is nothing to lower — is not. That is exactly the observed pattern: quadrupole moments vanish at the magic numbers and are largest in mid-shell.

Bohr and Mottelson built this into a model in which the shape variables and the single-particle motion are dynamical partners [1015, 1396]: the nucleons move in a deformed well whose deformation is itself a degree of freedom, and the two are coupled. The quantitative test is the electric quadrupole transition strength $B(E2)$ between the members of a band,

which measures the square of the collective matrix element and comes out one to two orders of magnitude above the single-particle estimate of Equation (107.84) — the standard statement being that the transition is "a hundred Weisskopf units", which is a direct count of how many nucleons are moving coherently.

107.4.4 Where the models fail

Phenomenon 107.59 (Halo nuclei break the $A^{1/3}$ law)**.** The interaction cross-sections of light neutron-rich nuclei measured with radioactive beams give matter radii that follow $r_0 A^{1/3}$ for stable species and depart from it abruptly at the neutron drip line: ^{11}Li has a matter radius near 3.1 fm where the systematics predict 2.7 fm, comparable with that of ^{40}Ca although it has a quarter as many nucleons [1764].

Rests on Proposition 107.22 and Phenomenon 107.32.

Derivation. *Derives Phenomenon 107.59.*

The mechanism is the one already used for the deuteron in Proposition 107.22. A weakly bound s or p neutron has an exterior wave function $\mathrm{e}^{-\kappa r}$ with $\kappa = \sqrt{2\mu S_n}/\hbar$, so its spatial extent is $1/\kappa$; for ^{11}Li the two-neutron separation energy is only $S_{2n} = 0.37\,\mathrm{MeV}$, and with the reduced mass of a neutron pair against a ^{9}Li core,

$$\frac{1}{\kappa} = \frac{\hbar c}{\sqrt{2\mu c^2 S_{2n}}} = \frac{197.33\,\mathrm{MeV\,fm}}{\sqrt{2 \times 1536\,\mathrm{MeV} \times 0.37\,\mathrm{MeV}}} = 5.9\,\mathrm{fm}\,,$$

several times the radius of the core. The last neutrons therefore spend most of their time outside the nuclear surface, in a region where there is no nuclear matter at all, and the concept of a saturated drop with a well-defined surface no longer applies. That the phenomenon is confined to the drip line follows from the same formula: at a normal separation energy of 8 MeV, $1/\kappa = 1.6\,\mathrm{fm}$ and nothing unusual happens. The halo is not a new force; it is the same threshold physics that makes the deuteron large, applied to a system that is barely bound because it is at the edge of stability. □

Remark 107.60 (Magic numbers migrate)*.* Theorem 107.52 was derived for a well filled near the valley of Equation (107.49). Far from it the ordering of Table 107.4 changes, because the parts of the nucleon–nucleon interaction that depend on isospin — in particular the tensor term Equation (107.29) — shift proton and neutron levels differently when N/Z departs from its usual value. The consequences are measured. $N = 20$ and $N = 28$, magic near stability, are not magic in the neutron-rich magnesium and silicon isotopes; and a new closure appears at $N = 34$, established from the energy of the first 2^+ state of ^{54}Ca [2224]. A magic number is therefore not a property of the nuclear force but of a particular filling of a particular mean field, and the extrapolation of Equation (107.66) to the superheavy region (Section 107.6.3) has to be made with that in mind.

Remark 107.61 (What is and is not computable from the interaction)*.* The programme of Theorem 107.29 — solve the many-body Schrödinger equation with a two- and three-nucleon interaction derived from chiral effective field theory, with no further parameters — is complete for $A \lesssim 16$ and, with various controlled truncations, reaches the medium-mass region and the doubly magic ^{100}Sn and ^{132}Sn. Binding energies come out to a few tenths of a MeV per nucleon and spectra to a few hundred keV. What remains out of reach is the deformed mid-shell heavy nuclei — precisely the ones that Section 107.4.3 describes with two collective coordinates — because the number of configurations grows combinatorially and the collective correlations are exactly the ones that converge most slowly. The honest statement, as of the present writing, is that the light nuclei are computed, the heavy spherical ones are approached, and the heavy deformed ones are modelled.

107.5 Radioactive decay

107.5.1 Alpha decay

Phenomenon 107.62 (Alpha lifetimes span twenty orders of magnitude)**.** The half-lives of α emitters run from less than a microsecond to more than 10^{17} seconds, while the energies of the emitted α particles vary only between about 4 MeV and 9 MeV. Within a decay chain the logarithm of the decay constant is an almost exactly linear function of the reciprocal square root of that energy [372] — in the modern statement; Geiger and Nuttall's own relation was between $\log\lambda$ and the logarithm of the α particle's range in air, which is equivalent because the range grows as a power of the speed. A factor of two in energy therefore buys more than twenty orders of magnitude in lifetime — an extreme sensitivity of a rate to an energy, and one that no classical picture of escape over a barrier can produce.

Rests on Phenomenon 107.2.

The classical impossibility is worth stating precisely before the resolution. The α particle emerges with 4 MeV; the Coulomb barrier it must have crossed, evaluated at the nuclear surface $R = 9.3\,\mathrm{fm}$ of Equation (107.71), is

$$V(R) = \frac{2Z_d e^2}{4\pi\varepsilon_0 R} = \frac{2 \times 90 \times 1.4400\,\mathrm{MeV\,fm}}{9.3\,\mathrm{fm}} = 27.9\,\mathrm{MeV}\,, \tag{107.70}$$

nearly seven times higher. A classical particle inside the nucleus with 4 MeV can never be outside it, and one outside with 4 MeV can never get in — which is also why an α-particle beam of that energy shows pure Rutherford scattering. Gamow [600], and independently Gurney and Condon [603, 625], resolved it by tunnelling: the first application of quantum mechanics to the nucleus, made within two years of the Schrödinger equation.

Theorem 107.63 (Alpha decay rate)**.** *Let a nucleus of mass number A emit an α particle to a daughter of charge $Z_d e$ and mass number $A_d = A - 4$, with total energy release Q. Write μ for the reduced mass of the α–daughter pair,*

$$R = r_0\left(A_d^{1/3} + 4^{1/3}\right)\,, \qquad b = \frac{2Z_d e^2}{4\pi\varepsilon_0 Q}\,, \qquad \xi = \frac{R}{b}\,. \tag{107.71}$$

Then the decay constant is $\lambda = fT$ with T the barrier transmission of Theorem 9.76 and f an assault frequency, and the Gamow exponent is

$$2\gamma = \frac{2\sqrt{2\mu Q}\;b}{\hbar}\left[\arccos\sqrt{\xi} - \sqrt{\xi\,(1-\xi)}\right]\,. \tag{107.72}$$

Rests on Theorem 9.76 and Equation (9.93).

Derivation. Derives Theorem 107.63.

The mathematics is done, in SI, in Section 9.6.6: Theorem 9.76 gives the transmission through a barrier between turning points as $T \approx \mathrm{e}^{-2\gamma}$ with $\gamma = \hbar^{-1}\int\sqrt{2m(V-E)}\,\mathrm{d}x$, and Example 9.81 evaluates that integral for exactly this potential, obtaining Equation (9.93), which is Equation (107.72). Nothing needs to be rebuilt here; what needs to be supplied is the nuclear input, and there are three pieces of it.

The two-body reduction. The problem is not a particle in a fixed potential but two bodies, so the mass in the exponent is the reduced mass $\mu = m_\alpha M_d/(m_\alpha + M_d)$ and the energy is the total release Q, not the α particle's laboratory kinetic energy $E_\alpha = QM_d/(M_d + m_\alpha)$. These are consistent: $\sqrt{2\mu Q} = \sqrt{2m_\alpha E_\alpha}$, so the momentum in the exponent is the measured one either way, while the outer turning point $b = 2Z_d e^2/4\pi\varepsilon_0 Q$ must be computed from Q. For ^{238}U the difference between using Q and using E_α is two per cent in b and a factor of three in the half-life — which is the measure of how sensitive this calculation is.

The inner radius. Unlike the outer turning point, R is not determined by the energy: it is where the Coulomb potential is cut off by the nuclear surface, and the natural estimate is the sum of the two radii, Equation (107.71). At $r = R$ the potential is Equation (107.70), far above Q, so the momentum does not vanish there and no connection formula is applied at the inner edge; the single delicate matching is the one at b, which is what Theorem 9.76 performs.

The prefactor. The α particle must exist inside the nucleus and must present itself at the barrier. Writing $\lambda = fT$ with $f = v/2R$, $v = \sqrt{2Q/\mu}$ the relative speed, treats the α as rattling to and fro across the nuclear diameter. This is crude in two ways that pull in opposite directions: the speed inside the well is larger than v, because the well is 40 MeV deep, which raises f; and there is a *preformation* probability, the chance that four nucleons are found grouped as an α particle at the surface at all, which is of order 10^{-2} and lowers λ. Neither is supplied by the method — Remark 9.77 says as much for the mathematics, and it is doubly true here — and the calculation below simply omits both. What can be defended is the exponent. □

Table 107.5: Alpha decay computed from Equation (107.72) with $r_0 = 1.20\,\mathrm{fm}$ and $\lambda = (v/2R)\,\mathrm{e}^{-2\gamma}$, against the measured half-lives [2382]. The Q values are from the mass evaluation [2393]. All seven agree within a factor of five, over a range of 1.5×10^{24} in the half-life. No parameter is fitted: r_0 is the value measured independently by electron scattering (Table 107.3).

parent	Q (MeV)	2γ	T	$T_{1/2}$ calc. (s)	$T_{1/2}$ obs. (s)	ratio
^{238}U	4.270	88.0	6.1×10^{-39}	1.5×10^{17}	1.4×10^{17}	1.0
^{232}Th	4.082	89.1	2.0×10^{-39}	4.6×10^{17}	4.4×10^{17}	1.0
^{226}Ra	4.871	72.8	2.5×10^{-32}	3.3×10^{10}	5.0×10^{10}	0.66
^{222}Rn	5.590	60.8	3.8×10^{-27}	2.0×10^{5}	3.3×10^{5}	0.61
^{218}Po	6.115	53.0	9.4×10^{-24}	7.8×10^{1}	1.9×10^{2}	0.42
^{214}Po	7.833	39.0	1.2×10^{-17}	5.5×10^{-5}	1.6×10^{-4}	0.34
^{212}Po	8.954	32.3	9.5×10^{-15}	6.3×10^{-8}	2.9×10^{-7}	0.21

Corollary 107.64 (The Geiger–Nuttall relation)**.** *For $\xi \ll 1$ the exponent Equation (107.72) reduces to*

$$2\gamma \to 2\pi\eta = \frac{2\pi}{\hbar}\,\frac{2Z_d e^2}{4\pi\varepsilon_0}\sqrt{\frac{\mu}{2Q}}\,, \tag{107.73}$$

so that

$$\log_{10}(\lambda\,\mathrm{s}) = a(Z_d) - \frac{b(Z_d)}{\sqrt{Q}}\,, \tag{107.74}$$

which is the empirical relation of [372], in the modern energy variable rather than their original logarithm of the range (Phenomenon 107.62).

Rests on Equation (107.72) and Phenomenon 107.62.

Proof. *Derives Corollary 107.64.*

Setting $\xi \to 0$ in Equation (107.72) replaces the bracket by $\arccos 0 = \pi/2$, and $\sqrt{2\mu Q}\,b = 2Z_d e^2\sqrt{2\mu/Q}/4\pi\varepsilon_0$ gives Equation (107.73). Taking the logarithm of $\lambda = f\mathrm{e}^{-2\gamma}$ and absorbing the slowly varying f into a gives Equation (107.74) with

$$b(Z_d) = \frac{2\pi}{\hbar \ln 10}\,\frac{2Z_d e^2}{4\pi\varepsilon_0}\sqrt{\frac{\mu}{2}} = 1.70\,Z_d\ \mathrm{MeV}^{1/2}\,,$$

using Equations (107.1) and (107.2) and $\mu c^2 = 3657\,\mathrm{MeV}$ for the heavy emitters.

The finite radius R cuts the barrier off at $\xi \approx 0.15$ to 0.35 and roughly halves the exponent — for ^{238}U, $2\pi\eta = 171$ against the exact $2\gamma = 88$ — but it barely touches the *slope*, and the reason is worth having. Expanding the bracket of Equation (107.72) for small ξ gives $2\gamma \approx 2\pi\eta\,[1 - (4/\pi)\sqrt{\xi} + \cdots]$, and since $2\pi\eta \propto Q^{-1/2}$ while $\sqrt{\xi} \propto Q^{1/2}$, the correction term is nearly independent of Q: it shifts the intercept $a(Z_d)$, not the coefficient of $Q^{-1/2}$. The exact exponents of Table 107.5 confirm both statements. Take the three polonium isotopes, for which $Z_d = 82$ is fixed: plotting 2γ against $Q^{-1/2}$ gives successive slopes 298 and 290 — linear to three per cent, and within ten per cent of the point-nucleus $b \ln 10 = 322$ — while the differences $2\pi\eta - 2\gamma$ are 77, 76 and 75: the predicted constant. In terms of the observable, $\log_{10}(\lambda\,\mathrm{s})$ against $Q^{-1/2}$ has predicted slope $-298/\ln 10 = -129$, and the measured half-lives give -129 and -119. That is the Geiger–Nuttall relation derived, with the right slope, from a barrier and a wave function. □

Remark 107.65 (What the agreement does and does not prove). Table 107.5 spans 1.5×10^{24} in half-life and holds within a factor of five, which is a striking result for a calculation with no fitted parameter. It should not be over-read. The exponent is so large that everything depends on it and almost nothing on the prefactor: changing r_0 from 1.20 fm to 1.30 fm shortens the computed half-life of ^{238}U by a factor of 22, and to 1.10 fm lengthens it by a factor of 26. What the table establishes is that a single value of r_0, the one independently measured by electron scattering, reproduces the whole range; it does not establish the prefactor, and the agreement of f with the truth to order unity is partly a cancellation between the two errors identified in Theorem 107.63. The preformation factor is the model's honest weak point, and it is where the α-decay rate stops being a barrier problem and becomes a nuclear-structure problem: it is smallest for odd-A emitters, whose unpaired nucleon must be broken out of its orbital, which is why they are systematically slower than Table 107.5 would predict by a further factor of 10 to 1000 — the phenomenon called hindrance.

Proposition 107.66 (Why alpha and not proton or deuteron emission)**.** *Among the light fragments a heavy nucleus might emit, only the α particle generally has $Q > 0$.*

Rests on Phenomenon 107.32 and Definition 107.31.

Proof. *Derives Proposition 107.66.*

The energy release for emitting a fragment X is $Q = B(X) + B(\text{daughter}) - B(\text{parent})$, and the term that decides it is $B(X)$. From Table 107.2, $B(^4\text{He}) = 28.30\,\text{MeV}$, against $B(^2\text{H}) = 2.22\,\text{MeV}$, $B(^3\text{H}) = 8.48\,\text{MeV}$ and, for a single nucleon, zero. The rest of the expression — the loss of binding by the parent — is of order A times the local slope of the binding curve and is much the same in each case, roughly 25 MeV for the removal of four nucleons from a heavy nucleus. Only the α particle brings enough binding of its own to pay for it, because it is doubly magic and exceptionally tightly bound for its size. That proton emission and ^{12}C emission do occur, the first beyond the proton drip line and the second as a rare branch called cluster decay with branching ratios near 10^{-11}, is the exception that measures the rule. □

107.5.2 Beta decay

The dynamics of β decay — Fermi's four-fermion interaction [738], the spectrum shape, the Kurie plot, parity violation and the $V - A$ structure — is derived in Chapter 101, and the detection of the antineutrino that closed the argument is [1085]. What belongs here is what the nucleus contributes: which decays are energetically open, which are fast and which are slow, and what the systematics measure.

Proposition 107.67 (Energy conditions for the three modes)**.** *Let* $M_{\mathrm{at}}(Z,A)$ *denote* atomic *masses, which is what mass spectrometry reports. Then*

$$Q_{\beta^-} = [M_{\mathrm{at}}(Z,A) - M_{\mathrm{at}}(Z+1,A)]\,c^2\,, \tag{107.75}$$

$$Q_{\beta^+} = [M_{\mathrm{at}}(Z,A) - M_{\mathrm{at}}(Z-1,A)]\,c^2 - 2m_e c^2\,, \tag{107.76}$$

$$Q_{\mathrm{EC}} = [M_{\mathrm{at}}(Z,A) - M_{\mathrm{at}}(Z-1,A)]\,c^2 - B_e\,, \tag{107.77}$$

with B_e *the binding energy of the captured atomic electron. Electron capture is therefore open in a window of* $2m_ec^2 = 1.022\,\mathrm{MeV}$ *in which* β^+ *emission is forbidden.*

Rests on Equation (107.5).

Proof. *Derives Proposition 107.67.*

Work with nuclear masses first and convert. For β^-, $M_{\mathrm{nuc}}(Z) \to M_{\mathrm{nuc}}(Z+1) + m_e$, so $Q = [M_{\mathrm{nuc}}(Z) - M_{\mathrm{nuc}}(Z+1) - m_e]\,c^2$; substituting $M_{\mathrm{nuc}} = M_{\mathrm{at}} - Zm_e$ (neglecting electron binding) the Z electrons on the left and the $Z+1$ on the right differ by exactly the emitted electron, and it cancels, giving Equation (107.75). For β^+ the same substitution leaves *two* electron masses uncancelled — the positron created, and the atomic electron the daughter no longer needs — and gives Equation (107.76). For electron capture the electron is taken from the atom rather than created, so no such term appears and only its binding energy is lost, giving Equation (107.77).

The window follows. Between $Q_{\mathrm{EC}} = 0$ and $Q_{\mathrm{EC}} = 1.022\,\mathrm{MeV}$ a nuclide can capture but not emit a positron, which is why ^{7}Be ($Q_{\mathrm{EC}} = 0.862\,\mathrm{MeV}$) decays only by capture. The free neutron sits at the other extreme with

$$Q_n = (m_n - m_p - m_e)\,c^2 = 939.565\,42\,\mathrm{MeV} - 938.272\,09\,\mathrm{MeV} - 0.511\,00\,\mathrm{MeV} = 0.782\,33\,\mathrm{MeV}\,, \tag{107.78}$$

which is the endpoint of the neutron β spectrum and the number that Equation (107.5) made available. □

Remark 107.68 (Why a bound neutron does not decay)*.* Equation (107.78) is positive, so a free neutron decays; yet a neutron in ^{4}He does not, and the reason is entirely nuclear. Converting a neutron into a proton in a stable nucleus moves it up the parabola of Proposition 107.39, and the cost — the asymmetry term plus the Coulomb term, several MeV — exceeds 0.78 MeV. Conversely, in a neutron-rich nucleus the gain exceeds it and the decay proceeds. The neutron's fate is decided by the mass formula, not by the weak interaction, which only fixes how fast.

Definition 107.69 (Comparative half-life and forbiddenness)**.** The ft value of Definition 101.16 divides out the phase-space factor $f(Z,E_0)$, leaving a number proportional to the inverse square of the nuclear matrix element. Transitions are classified by the orbital angular momentum L carried off by the lepton pair: $L=0$ is *allowed*, $L=1$ *first forbidden*, and so on.

Proposition 107.70 (The forbiddenness ladder)**.** *Each unit of forbiddenness suppresses the rate by a factor of order*

$$\left(\frac{p_e R}{\hbar}\right)^2 \approx 6 \times 10^{-4}\,, \tag{107.79}$$

for $p_ec = 1\,\mathrm{MeV}$ *and* $R = 5\,\mathrm{fm}$*, and so adds between three and four to* $\log_{10}(ft/\mathrm{s})$.

Rests on Definition 107.69 and Equation (107.1).

Proof. *Derives Proposition 107.70.*

The lepton wave functions are plane waves over the nuclear volume, and the Rayleigh plane-wave expansion $e^{i\mathbf{p}\cdot\mathbf{r}/\hbar} = \sum_L i^L(2L+1)j_L(pr/\hbar)P_L(\cos\theta)$ has $j_L(x) \sim x^L/(2L+1)!!$ for small x. With $x = p_eR/\hbar$ evaluated at $p_ec = 1\,\mathrm{MeV}$ and $R = 5\,\mathrm{fm}$,

$$x = \frac{1\,\mathrm{MeV} \times 5\,\mathrm{fm}}{197.33\,\mathrm{MeV\,fm}} = 0.0253\,,$$

so the amplitude for $L = 1$ is smaller than for $L = 0$ by x and the rate by $x^2 = 6.4 \times 10^{-4}$, which is Equation (107.79): three and a fifth in the logarithm. The parity of the nuclear state must also change with each unit of L, since the lepton pair carries parity $(-1)^L$, which is what makes the classification observable rather than merely formal — a transition between states of the same parity and $\Delta J = 2$ must be second forbidden and cannot be anything else.

Measured $\log_{10}(ft/\mathrm{s})$ values confirm the ladder: about 3.5 for superallowed, 5 to 6 for allowed, 6 to 9 for first forbidden, and up to 22 for the highest observed. The extremes are instructive. ^{6}He decays in 0.81 s with $\log_{10} ft = 2.9$; ^{40}K, a third-forbidden unique transition from 4^- to 0^+, lives 1.25×10^9 yr [2382] with $\log_{10} ft \approx 18$. That is fifteen orders of magnitude in ft from three units of forbiddenness, against the ten that Equation (107.79) predicts; the shortfall is the extra cost of a *unique* transition, in which the lepton pair must carry not only $L = 3$ but the maximum spin coupling as well, and it is why the classification distinguishes unique from non-unique cases at all. □

Derivation pending #171

The Rayleigh plane-wave expansion: the expansion of a plane wave in Legendre polynomials and spherical Bessel functions, with the small-argument behaviour $j_L(x) \sim x^L/(2L+1)!!$ used above. Both are premises of the forbiddenness ladder; the expansion belongs to the chapter on Fourier and integral transforms, where it is not yet derived.

Remark 107.71 (What the superallowed decays measure). The Fermi operator of Proposition 101.15 is the isospin raising operator, whose matrix element between members of an isospin multiplet is fixed by the SU(2) algebra of Definition 107.10 and not by the nuclear wave function at all. The $0^+ \to 0^+$ decays within an isospin triplet therefore carry no nuclear-structure uncertainty, and the constancy of their $\mathcal{F}t$ across some fourteen cases spanning a factor of 10^4 in phase space (Phenomenon 101.18 and Equation (101.19)) is simultaneously the proof that the vector coupling is not renormalized by the strong interaction and the most precise determination of $|V_{ud}|$. This is the one place where nuclear physics supplies a number to the Standard Model rather than the reverse, and it works precisely because the nuclear physics cancels.

Phenomenon 107.72 (Double beta decay)**.** A dozen even–even nuclides for which single β decay is energetically forbidden decay by emitting two electrons and two antineutrinos at once, with half-lives between 10^{18} and 10^{24} years — the slowest processes ever measured. The neutrinoless mode, in which the two electrons carry the whole energy release and no neutrinos are emitted, has not been observed; limits on its half-life exceed 10^{26} years.

Rests on Corollary 107.40.

Derivation. *Derives Phenomenon 107.72.*

That the mode exists at all is a consequence of Corollary 107.40. For even A the pairing term splits the mass parabola in two, and an even–even nuclide can find itself lower than *both* of its odd–odd neighbours while being higher than the even–even nuclide two steps away. Single β decay is then closed and double β decay is open. The classic case is ^{76}Ge, which cannot reach ^{76}As but can reach ^{76}Se; the same arithmetic selects ^{82}Se, ^{100}Mo, ^{130}Te and ^{136}Xe.

The rate is second order in the weak interaction, so it carries G_F^4 instead of G_F^2, and the four-body final state gives a phase-space factor scaling as Q^{11} rather than Q^5. Those two together are the twenty orders of magnitude between an allowed β decay and a double one. The neutrinoless mode is a different matter: it requires the neutrino emitted at one vertex to be absorbed at the other, which is possible only if the neutrino is its own antiparticle and has mass, so its observation would establish lepton-number violation and the Majorana nature

of the neutrino. It has not been observed, and this book records searches by what they have excluded. The status belongs to Chapter 103; what belongs here is that the experimental signature is entirely nuclear — a peak at the Q value of the two-electron spectrum, sitting on the continuum of the two-neutrino mode — and that the conversion of a half-life limit into a mass limit requires a nuclear matrix element which different many-body methods disagree about by a factor of two or three. □

107.5.3 Gamma decay and the Mössbauer effect

Proposition 107.73 (Multipole selection rules)**.** *A gamma transition between nuclear states of spins and parities $J_i^{\pi_i}$ and $J_f^{\pi_f}$ proceeds by multipoles L with*

$$|J_i - J_f| \le L \le J_i + J_f\,, \qquad L \ge 1\,, \tag{107.80}$$

and the parity change fixes the type: $\pi_i \pi_f = (-1)^L$ for an electric multipole EL and $(-1)^{L+1}$ for a magnetic one ML. A transition $0 \to 0$ cannot emit a photon at all.

Rests on Phenomenon 79.5.

Proof. *Derives Proposition 107.73.*

The photon carries away angular momentum $L\hbar$ and the total is conserved, which with the triangle rule for adding $\boldsymbol{J}_i = \boldsymbol{J}_f + \boldsymbol{L}$ (Chapter 79) gives the range in Equation (107.80). The lower bound $L \ge 1$ is the statement that a free photon has no monopole mode: its field is transverse, so the $L = 0$ component of the multipole expansion of a radiation field vanishes identically. The parity of an EL photon is $(-1)^L$ and of an ML photon $(-1)^{L+1}$, because the electric and magnetic multipole operators differ by one factor of the (axial) magnetic field; conservation of parity in the electromagnetic interaction then fixes which type may occur. For $J_i = J_f = 0$ the range in Equation (107.80) is empty and single-photon emission is strictly forbidden. Such transitions nevertheless occur, by internal conversion or, if the energy exceeds $2m_e c^2$, by internal pair creation — and the fact that these proceed while photon emission does not is direct evidence that they are not merely a photon reabsorbed. □

Theorem 107.74 (Single-particle transition rates)**.** *For an electric multipole of order L,*

$$\lambda(EL) = \frac{8\pi\,(L+1)}{L\,[(2L+1)!!]^2}\,\frac{1}{4\pi\varepsilon_0\hbar}\left(\frac{E_\gamma}{\hbar c}\right)^{2L+1} B(EL)\,, \tag{107.81}$$

with $B(EL)$ the reduced transition probability of dimension $\mathrm{C}^2\,\mathrm{m}^{2L}$. Estimating $B(EL)$ for a single proton moving in a uniform sphere of radius $R = r_0 A^{1/3}$,

$$B(EL)_W = \frac{e^2}{4\pi}\left(\frac{3}{L+3}\right)^2 R^{2L}\,, \tag{107.82}$$

gives, with $r_0 = 1.2\,\mathrm{fm}$ and E_γ in MeV*,*

$$\lambda(E1) = 1.0 \times 10^{14}\; A^{2/3} E_\gamma^3\;\mathrm{1/s}\,, \tag{107.83}$$

$$\lambda(E2) = 7.3 \times 10^{7}\; A^{4/3} E_\gamma^5\;\mathrm{1/s}\,, \tag{107.84}$$

$$\lambda(E3) = 34\; A^{2} E_\gamma^7\;\mathrm{1/s}\,, \tag{107.85}$$

$$\lambda(E4) = 1.1 \times 10^{-5}\; A^{8/3} E_\gamma^9\;\mathrm{1/s}\,. \tag{107.86}$$

Rests on Propositions 107.70 and 107.73.

Derivation. *Derives Theorem 107.74.*

Equation (107.81) is the multipole expansion of the radiation field applied to Fermi's golden rule; its structure is what matters here and it is fixed by dimensional analysis together

with the small parameter of the problem. The photon field at the nucleus varies as $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}$ with $k = E_\gamma/\hbar c$, and $kR \ll 1$ — for $E_\gamma = 1\,\mathrm{MeV}$ and $R = 6\,\mathrm{fm}$, $kR = 0.030$ — so each successive term of the expansion of Proposition 107.70 costs one power of kR in the amplitude. Two powers of kR per unit of L in the rate, and one factor of E_γ from the density of photon states, give the $(E_\gamma/\hbar c)^{2L+1}$; the numerical coefficient comes from the normalization of the vector spherical harmonics and is quoted.

Substituting Equation (107.82) and using $e^2/4\pi\varepsilon_0\hbar = \alpha c$ collapses the prefactor to a pure number times αc:

$$\lambda(EL) = \frac{8\pi(L+1)}{L\,[(2L+1)!!]^2}\cdot\frac{1}{4\pi}\left(\frac{3}{L+3}\right)^2\cdot\alpha c\,R^{2L}\left(\frac{E_\gamma}{\hbar c}\right)^{2L+1}.$$

For $L = 1$ the two numerical factors are $16\pi/9$ and $(1/4\pi)(3/4)^2$, whose product is exactly $\frac{1}{4}$, so

$$\lambda(E1) = \frac{\alpha c}{4}\,R^2\left(\frac{E_\gamma}{\hbar c}\right)^3,$$

and inserting $\alpha = 7.2974\times10^{-3}$, $c = 3.0\times10^{23}\,\mathrm{fm/s}$, $R^2 = 1.44A^{2/3}\,\mathrm{fm}^2$ and $\hbar c = 197.33\,\mathrm{MeV\,fm}$ gives Equation (107.83). The same substitution at $L = 2, 3, 4$ gives the rest. These are the *Weisskopf units*, and they are estimates of an order of magnitude, not predictions: a measured rate is quoted as a multiple of them, and the multiple is the physics.

Magnetic multipoles are suppressed relative to electric ones of the same order by roughly

$$\frac{\lambda(ML)}{\lambda(EL)} \sim \left(\frac{\hbar}{m_p c R}\right)^2 = \left(\frac{0.210\,\mathrm{fm}}{R}\right)^2, \tag{107.87}$$

about 2×10^{-3} for a medium-mass nucleus, since the magnetic moment of a nucleon carries the extra factor $\hbar/m_p c$ that a charge does not. □

Derivation pending #172

The multipole normalization: the coefficient $8\pi(L+1)/L[(2L+1)!!]^2$ of the electric multipole rate is quoted, not derived. Its derivation — the multipole expansion of the quantized radiation field, the normalization of the vector spherical harmonics and Fermi's golden rule — belongs in Appendix A; the power counting in kR and the dimensional structure are derived in the text.

Corollary 107.75 (Isomers)**.** *A transition requiring a high multipole is slow by many orders of magnitude, and a nuclear state whose only decay is such a transition is metastable.*
Rests on Theorems 107.52 and 107.74.

Proof. *Derives Corollary 107.75.*

From Equations (107.83) and (107.86), for $A = 125$ and $E_\gamma = 1\,\mathrm{MeV}$ the rates are 2.6×10^{15}, 4.6×10^{10}, 5.3×10^{5} and $4.2\,1/\mathrm{s}$ for $L = 1, 2, 3, 4$: a factor near 10^5 per unit of L, which is $(kR)^2$ combined with the falling numerical coefficient. A 4^+ state lying just above a 0^+ ground state therefore lives of order a second rather than a femtosecond. The islands of isomerism sit just below the magic numbers, where Table 107.4 places levels of very different j next to one another — $1g_{9/2}$ beside $2p_{1/2}$ below Z or $N = 50$, $1h_{11/2}$ beside $3s_{1/2}$ below 82 — so that the lowest available transition has $\Delta J = 4$ or 5. That the isomer islands are found exactly where the shell model puts high-j intruders is a further confirmation of Theorem 107.52 that uses no binding energies at all. □

Remark 107.76 (Internal conversion). A nucleus may also give its energy directly to an atomic electron, which leaves with kinetic energy $E_\gamma - B_e$ — a *line*, not a continuum, and historically the first clue that the continuous β spectrum was a genuine puzzle rather than an instrumental artefact, since the two were seen side by side. The conversion coefficient $\alpha_{\mathrm{IC}} = \lambda_e/\lambda_\gamma$ grows as Z^3, grows with L, and falls steeply with E_γ; measuring it is a standard way of determining the multipolarity of a transition whose spins are unknown. It is not internal photoelectric absorption of an emitted photon: it happens even when Proposition 107.73 forbids photon emission outright.

Phenomenon 107.77 (Recoil-free resonant absorption of gamma rays)**.** A nucleus bound in a crystal can emit a gamma ray, and another such nucleus absorb it, with no energy lost to recoil, so that the line keeps the natural width set by the lifetime of the state [1154]. For the 14.4 keV transition of ^{57}Fe that width is a fraction of about 3×10^{-13} of the transition energy, and by measuring displacements of a fraction of the line the method resolved the gravitational redshift over a vertical drop of 22.5 m, a shift of a few parts in 10^{15} [1189]. Free nuclei show no such resonance at all.

Derivation. *Derives Phenomenon 107.77.*

A free nucleus of mass M emitting a photon of energy E_γ recoils with momentum E_γ/c and so carries away

$$E_R = \frac{E_\gamma^2}{2Mc^2}\,, \tag{107.88}$$

to first order in E_R/E_γ. The emitted quantum is thus short of the transition energy by E_R, while an absorbing nucleus needs E_R more than the transition energy in order to recoil in turn: emission and absorption lines are displaced from one another by $2E_R$. Resonance survives only if that displacement is small against the line width Γ. For the 14.4 keV line of ^{57}Fe, $Mc^2 = 53.0\,\mathrm{GeV}$ gives $E_R = 1.96 \times 10^{-3}\,\mathrm{eV}$, whereas the mean life of 141 ns gives $\Gamma = \hbar/\tau = 4.67 \times 10^{-9}\,\mathrm{eV}$: the displacement exceeds the width by a factor of 8.4×10^5, and no resonance between free nuclei is possible. Bound in a lattice the nucleus cannot recoil by itself. A finite fraction of the transitions leave the crystal in its original vibrational state, and for those the mass appearing in Equation (107.88) is the mass of the whole crystal, which makes E_R utterly negligible. The line is then as narrow as the lifetime permits, and it is that narrowness, not any large signal, which gives the method its resolution:

$$\frac{\Gamma}{E_\gamma} = \frac{4.67 \times 10^{-9}\,\mathrm{eV}}{14.4 \times 10^{3}\,\mathrm{eV}} = 3.2 \times 10^{-13}\,. \tag{107.89}$$

□

Proposition 107.78 (The recoil-free fraction)**.** *In the Debye model of a lattice at zero temperature, the fraction of emissions that leave the crystal in its original vibrational state is*

$$f = \exp\left(-\frac{3E_R}{2k_B\Theta_D}\right)\,, \tag{107.90}$$

which for ^{57}Fe *in iron,* $\Theta_D = 470\,\mathrm{K}$*, is* $f = 0.93$*.*

Rests on Phenomenon 107.77 and Equation (107.88)*.*

Proof. *Derives Proposition 107.78.*

The probability of a zero-phonon transition is $f = \exp(-k^2\,\langle x^2\rangle)$ with $k = E_\gamma/\hbar c$ the photon wave number and $\langle x^2\rangle$ the mean square displacement of the emitting nucleus along the photon direction — the same Debye–Waller factor that governs the intensity of a Bragg peak, a subject that belongs to Chapter 125. Both formulas are used here as stated. In the Debye model at $T = 0$, $\langle x^2\rangle = 3\hbar^2/(4Mk_B\Theta_D)$, so $k^2\,\langle x^2\rangle = 3E_\gamma^2/(4Mc^2k_B\Theta_D) = 3E_R/2k_B\Theta_D$, which is Equation (107.90). With $k_B\Theta_D = 40.5\,\mathrm{meV}$ and $E_R = 1.96\,\mathrm{meV}$ the exponent

is 0.073 and $f = 0.93$. Two conditions for a usable Mössbauer isotope follow at once: a low transition energy, since $E_R \propto E_\gamma^2$, and a stiff lattice, since f falls with Θ_D. That ^{57}Fe satisfies both is why essentially all Mössbauer spectroscopy uses it. □

Derivation pending #173

The Debye–Waller factor: the derivation of the zero-phonon probability $f = \exp(-k^2 \langle x^2 \rangle)$ for a harmonic lattice, and of the Debye-model zero-temperature mean square displacement $\langle x^2 \rangle = 3\hbar^2/(4Mk_B\Theta_D)$. Both are premises of the proposition above; they belong to the chapter on phonons and lattices, where they are not yet derived.

Corollary 107.79 (The velocity scale, and the Pound–Rebka measurement)**.** *A source moved at speed v shifts the emitted line by $\Delta E/E = v/c$, so the velocity needed to shift by one natural width is*

$$v_\Gamma = c\,\frac{\Gamma}{E_\gamma} = 9.7 \times 10^{-5}\,\mathrm{m/s}\,, \tag{107.91}$$

about a tenth of a millimetre per second. The gravitational redshift over a height H is gH/c^2, which for $H = 22.5\,\mathrm{m}$ is 2.46×10^{-15}, or 7.6×10^{-3} of the natural width.

Rests on Phenomenon 107.77 and Equation (107.89).

Proof. *Derives Corollary 107.79.*

The first-order Doppler shift gives Equation (107.91) directly from Equation (107.89); the millimetre-per-second scale of a Mössbauer spectrometer's drive is therefore not a convenience but a consequence of the line width. For the redshift,

$$\frac{gH}{c^2} = \frac{9.806\,65\,\mathrm{m/s^2} \times 22.5\,\mathrm{m}}{(299\,792\,458\,\mathrm{m/s})^2} = 2.46 \times 10^{-15}\,,$$

so Pound and Rebka had to locate the centre of a line to better than one per cent of its width, which they did by modulating the source velocity and measuring the asymmetry of the counting rate [1189]. The experiment is a test of the equivalence principle and belongs equally to Chapter 42; what belongs here is that the instrument is a nuclear energy level, and that its precision comes from a lifetime. □

107.6 Fission and fusion

107.6.1 Fission

Hahn and Strassmann, following up Fermi's irradiation of uranium with neutrons, established by radiochemistry that one of the products is barium — an element of half the atomic number, which no known nuclear process could produce [837]. Meitner and Frisch identified the process as the division of the nucleus and estimated its energy release from the mass defect before any measurement of it existed [842].

Phenomenon 107.80 (Uranium divides, and the energy released is enormous)**.** Uranium irradiated with neutrons yields barium among its products [837]: the nucleus divides into two fragments of comparable mass. The energy released per division is of order 200 MeV [842] — some eight orders of magnitude more than the few eV released when the same number of atoms rearrange chemically — and the division is accompanied by two or three prompt neutrons, which is what allows the process to sustain itself.

Rests on Phenomenon 107.32.

Derivation. *Derives Phenomenon 107.80.*

The energy released is the increase in total binding energy, and it can be read straight off the curve of Phenomenon 107.32. At $A \approx 236$ the binding is about 7.6 MeV per nucleon; near the fragment masses, $A \approx 100$ to 140, it is about 8.5 MeV per nucleon [2393]. The gain is therefore

$$\Delta E \approx 236\,(8.5\,\mathrm{MeV} - 7.6\,\mathrm{MeV}) \approx 210\,\mathrm{MeV}\,, \tag{107.92}$$

the observed order of magnitude, and — since two numbers were read off a curve — as much precision as the argument is entitled to. The same estimate says why fission is confined to the heavy end of the chart: the gain is positive only where the curve is falling, and it is worth a substantial fraction of a nucleon's binding only where it has fallen far. It says nothing about the *rate*, which is governed by the barrier and not by the energy release, and that is why ^{235}U and ^{238}U behave so differently under slow neutrons although their release is nearly the same. □

Table 107.6: Where the energy of one thermal-neutron fission of ^{235}U goes [1809]. The antineutrinos leave the reactor entirely, so the recoverable total is about 190 MeV; they are also what makes a reactor the brightest artificial antineutrino source available, which is how [1085] detected the neutrino at all.

channel	energy (MeV)
kinetic energy of the fragments	168
prompt neutrons	5
prompt gamma rays	7
beta particles from fragment decay	7
delayed gamma rays	6
antineutrinos (not recoverable)	9
total	202

Theorem 107.81 (The fission barrier and the fissility parameter)**.** *Deform a nucleus at constant volume into the spheroid $R(\theta) = R_0\,[1 + \alpha_2 P_2(\cos\theta)]$. To second order in α_2 the surface and Coulomb terms of Equation (107.41) become*

$$E_S = a_S A^{2/3}\left(1 + \tfrac{2}{5}\alpha_2^2\right)\,, \tag{107.93}$$

$$E_C = a_C \frac{Z^2}{A^{1/3}}\left(1 - \tfrac{1}{5}\alpha_2^2\right)\,, \tag{107.94}$$

so the energy change on deformation is

$$\Delta E = \frac{\alpha_2^2}{5}\left(2a_S A^{2/3} - a_C \frac{Z^2}{A^{1/3}}\right)\,, \tag{107.95}$$

and the sphere is unstable against spontaneous fission when

$$\frac{Z^2}{A} > \left(\frac{Z^2}{A}\right)_{\mathrm{crit}} = \frac{2a_S}{a_C} = 50.1\,. \tag{107.96}$$

Rests on Equations (107.41) and (107.44).

Proof. *Derives Theorem 107.81.*

Volume conservation fixes R_0 in terms of the equivalent spherical radius R: expanding $V = \frac{4\pi}{3}R_0^3\left(1 + \frac{3}{5}\alpha_2^2\right)$ to second order gives $R_0 = R\left(1 - \frac{1}{5}\alpha_2^2\right)$. The surface area of the

deformed shape, expanded to the same order, is $4\pi R^2\left(1+\frac{2}{5}\alpha_2^2\right)$, which with $E_S = 4\pi R^2\sigma$ from Equation (107.44) gives Equation (107.93); the electrostatic energy of a uniformly charged spheroid, expanded likewise, gives Equation (107.94). The first coefficient is positive and the second negative because stretching increases the surface but moves charge apart. Subtracting the spherical values leaves Equation (107.95), and the sphere is stable if and only if the bracket is positive, which rearranges to Equation (107.96) with the coefficients Equation (107.43).

The parameter is customarily normalized as $x = (Z^2/A)/(Z^2/A)_{\text{crit}}$, and for the actinides: ^{232}Th has $x = 0.70$, ^{238}U $x = 0.71$, ^{239}Pu $x = 0.74$. All are stable against small deformations and all can fission, because a stable sphere is only a *local* minimum: the ultimate configuration, two separated fragments, lies 200 MeV lower (Equation (107.92)), and between them stands a barrier that Equation (107.95) describes only near $\alpha_2 = 0$. The full barrier requires the whole deformation energy surface, which is the calculation of Bohr and Wheeler [830]; its height for the actinides is 5 MeV to 6 MeV, small compared with the release and comparable with the neutron separation energy — which is the coincidence on which the whole subject turns. □

Derivation pending #174

The fission barrier: the deformation energy surface of the charged liquid drop beyond second order, through the saddle point to the scission configuration, giving the barrier height as a function of the fissility x and its collapse to zero as x approaches one. It belongs in Appendix A, since it requires the multipole expansion of the electrostatic energy of an arbitrary axially symmetric shape. The second-order result and the critical fissility are derived in the text.

Phenomenon 107.82 (Only one uranium isotope fissions with slow neutrons)**.** The fission cross-section of ^{235}U for neutrons at 0.025 eV is 5.9×10^{-25} m^2; that of ^{238}U is immeasurably small below a neutron energy of about 1 MeV. The two nuclides differ by three neutrons and their fission energy releases agree within a few MeV.

Rests on Equation (107.42) and Theorem 107.81.

Derivation. Derives Phenomenon 107.82.

Capture of a neutron leaves a compound nucleus excited by exactly the neutron separation energy of the product, and it is that number, not the fission Q value, which must be compared with the barrier. From the mass evaluation [2393],

$$S_n\left(^{236}\text{U}\right) = 6.5\,\text{MeV}\,, \qquad S_n\left(^{239}\text{U}\right) = 4.8\,\text{MeV}\,, \tag{107.97}$$

against barriers of about 5.7 MeV and 6.2 MeV respectively. In the first case the compound nucleus is created *above* its barrier and fissions promptly, with no help from the neutron's kinetic energy; in the second it is created more than a MeV below, and the neutron must supply the difference.

The 1.7 MeV between the two separation energies is pairing. ^{235}U has $N = 143$, odd, so the captured neutron completes a pair and the system gains the pairing energy of Equation (107.42); ^{238}U has $N = 146$, even, so the captured neutron is unpaired and gains nothing. The predicted difference is $2a_P A^{-1/2} = 2\times 11.18/\sqrt{237} = 1.45$ MeV, against the 1.7 MeV observed. A term worth 1 MeV in a formula whose leading term is 3700 MeV therefore decides which isotope of uranium is a fuel, which is as sharp an illustration as the subject affords of why the small terms of Equation (107.41) were worth deriving. □

Remark 107.83 (The chain reaction). Fission of ^{235}U emits on average $\nu = 2.4$ prompt neutrons at a mean energy near 2 MeV. Three facts then decide whether a chain can be

sustained. Those neutrons are born fast, where the fission cross-section is small; slowing them to thermal energy raises it by the $1/v$ law of Proposition 107.49 by some three orders of magnitude, which is what a moderator is for. But natural uranium is 99.3 per cent ^{238}U, which absorbs neutrons in its capture resonances on the way down without fissioning, so the moderator must slow them in as few collisions as possible and away from the fuel — hence lumps of fuel in a lattice of graphite or heavy water, and hence enrichment. The multiplication factor k is the number of fissions in one generation per fission in the preceding one, and criticality is $k = 1$ exactly.

That last condition would be uncontrollable were it not for a detail of nuclear structure. A small fraction $\beta \approx 0.0065$ of the fission neutrons is not prompt: it is emitted seconds later by excited fragment nuclei whose neutron separation energy lies below the β-decay Q value of their parents, so that a β decay leaves a state that immediately emits a neutron. The mean generation time of the prompt neutrons is of order $1 \times 10^{-4}\,\mathrm{s}$, while these *delayed* neutrons arrive with the half-lives of their precursors, of order 10 s. A reactor held just below prompt criticality — $k = 1$ including the delayed fraction, $k < 1$ without it — therefore has an effective response time set by the precursors and not by the prompt neutrons, longer by five orders of magnitude, which is what makes mechanical control possible at all. The margin is β, six-tenths of one per cent, and it is a nuclear-structure accident.

Example 107.84 (Energy density)**.** Taking 200 MeV per fission, complete fission of one kilogram of ^{235}U releases

$$\frac{200\,\mathrm{MeV} \times N_A}{0.235\,\mathrm{kg/mol}} = \frac{3.204 \times 10^{-11}\,\mathrm{J} \times 6.022\,140\,76 \times 10^{23}/\mathrm{mol}}{0.235\,\mathrm{kg/mol}} = 8.2 \times 10^{13}\,\mathrm{J/kg}\,, \quad (107.98)$$

against about $3 \times 10^7\,\mathrm{J/kg}$ for the combustion of coal: a factor of 2.7×10^6. The ratio is not accidental — it is the ratio of a nuclear binding energy to a chemical one, of 1 MeV to 1 eV — and it is the entire practical significance of this chapter.

107.6.2 Fusion

Fusion faces the same Coulomb barrier as α decay, from the other side. Two protons at 2 fm have electrostatic energy

$$\frac{e^2}{4\pi\varepsilon_0 r} = \frac{1.4400\,\mathrm{MeV\,fm}}{2\,\mathrm{fm}} = 720\,\mathrm{keV}\,, \quad (107.99)$$

whereas the centre of the Sun is at $T = 1.57 \times 10^7\,\mathrm{K}$, that is $k_B T = 1.35\,\mathrm{keV}$ — smaller by a factor of 530. No classical gas at that temperature fuses at any rate whatever, and the resolution is again tunnelling, combined with the fact that a Maxwell–Boltzmann distribution has an exponential tail.

Theorem 107.85 (The Gamow peak)**.** *Let two nuclei of charges $Z_1 e$, $Z_2 e$ and reduced mass μ react with cross-section*

$$\sigma(E) = \frac{S(E)}{E}\,\exp\left(-\sqrt{\frac{E_G}{E}}\right)\,, \qquad E_G = 2\mu c^2\,(\pi\alpha Z_1 Z_2)^2\,, \quad (107.100)$$

with $S(E)$ slowly varying. In a thermal gas at temperature T the reaction rate per pair is dominated by a narrow band of energies centred on

$$E_0 = \left(\frac{E_G\,(k_B T)^2}{4}\right)^{1/3}\,, \quad (107.101)$$

of full width at $1/\mathrm{e}$ *of the maximum*

$$\Delta E = 4\sqrt{\frac{E_0 k_B T}{3}}\,. \tag{107.102}$$

Rests on Equation (107.73).

Derivation. *Derives Theorem 107.85.*

The exponent in Equation (107.100) is the point-Coulomb Gamow factor Equation (107.73) rewritten: with $\eta = Z_1 Z_2 e^2/(4\pi\varepsilon_0\hbar v)$ and $v = \sqrt{2E/\mu}$,

$$2\pi\eta = \frac{2\pi Z_1 Z_2 e^2}{4\pi\varepsilon_0\hbar}\sqrt{\frac{\mu}{2E}} = \pi\alpha Z_1 Z_2\sqrt{\frac{2\mu c^2}{E}} = \sqrt{\frac{E_G}{E}}\,,$$

using $e^2/4\pi\varepsilon_0\hbar = \alpha c$; the factor $1/E$ outside is the geometric area $\pi(\hbar/p)^2 \propto 1/E$, and everything not exponential and not geometric is collected into the *astrophysical S factor* $S(E)$, which is what a laboratory measurement extrapolates to stellar energies. Here the point-Coulomb limit is the right one — unlike in Corollary 107.64, where the finite nuclear radius mattered — because the classical turning point $b = Z_1 Z_2 e^2/4\pi\varepsilon_0 E$ at $E = 6\,\mathrm{keV}$ is $240\,\mathrm{fm}$, so $\xi = R/b \approx 0.01$ and the correction is a per cent.

The rate is the cross-section averaged over the Maxwell–Boltzmann distribution — whose derivation belongs to Chapter 55 — which contributes e^{-E/k_BT} and a factor E that varies slowly. So

$$\langle\sigma v\rangle \propto \int_0^\infty S(E)\,\exp\left[-\frac{E}{k_B T} - \sqrt{\frac{E_G}{E}}\right]\mathrm{d}E\,. \tag{107.103}$$

The two exponents pull in opposite directions — one kills large E, the other kills small E — so the integrand is sharply peaked. Setting the derivative of the bracket to zero,

$$-\frac{1}{k_B T} + \frac{1}{2}\frac{\sqrt{E_G}}{E^{3/2}} = 0 \quad\Longrightarrow\quad E_0 = \left(\frac{\sqrt{E_G}\,k_B T}{2}\right)^{2/3},$$

which is Equation (107.101). Expanding the bracket about E_0 to second order gives a Gaussian of variance $1/|f''(E_0)|$ with $f'' = -\frac{3}{4}\sqrt{E_G}E_0^{-5/2} = -3/(2E_0 k_B T)$, whence Equation (107.102). The construction is an evaluation by steepest descent, carried out on the real axis. □

Example 107.86 (Hydrogen burning in the Sun)**.** For two protons, $\mu c^2 = m_p c^2/2 = 469.14\,\mathrm{MeV}$ and $Z_1 Z_2 = 1$, so

$$E_G = 2 \times 469.14\,\mathrm{MeV} \times \left(\pi \times 7.2974 \times 10^{-3}\right)^2 = 493\,\mathrm{keV}\,,$$

and at the solar central temperature $k_B T = 1.353\,\mathrm{keV}$,

$$E_0 = \left(\frac{493\,\mathrm{keV} \times (1.353\,\mathrm{keV})^2}{4}\right)^{1/3} = 6.1\,\mathrm{keV}\,, \qquad \Delta E = 6.6\,\mathrm{keV}\,.$$

Fusion in the Sun therefore happens almost entirely between protons whose relative energy is between about $3\,\mathrm{keV}$ and $9\,\mathrm{keV}$ — four and a half times the mean thermal energy, in a part of the tail containing a small fraction of the protons, and still more than a hundred times below the barrier Equation (107.99). Both exponentials are essential: a purely thermal argument gives no reactions and a purely quantum one gives no temperature dependence. It is also why stellar burning is so violently temperature sensitive — the rate goes as a high power of T, four for the *pp* chain and near twenty for the CNO cycle — which is what Chapter 52 turns into the stability of a star.

Phenomenon 107.87 (Stars burn hydrogen to helium)**.** The net reaction $4p \to {}^4\mathrm{He} + 2e^+ + 2\nu_e$ releases

$$Q = 26.73\,\mathrm{MeV} \tag{107.104}$$

per helium nucleus, of which the neutrinos carry away on average about 0.6 MeV. Bethe identified the two competing routes, the *pp* chain and the CNO cycle [828], and the neutrinos predicted by the first have been detected at the rate the Sun's luminosity requires (Chapter 114).

Rests on Phenomenon 107.24 and Proposition 107.23.

Derivation. *Derives Phenomenon 107.87.*

The energy release is the mass difference. Using atomic masses, so that the two positrons annihilate with two of the electrons already counted,

$$Q = \left[4M_{\mathrm{at}}\left({}^1\mathrm{H}\right) - M_{\mathrm{at}}\left({}^4\mathrm{He}\right)\right]c^2 = 0.028\,697\,u\,c^2 = 0.028\,697 \times 931.494\,\mathrm{MeV} = 26.73\,\mathrm{MeV}\,,$$

equivalently $4m_pc^2 - M_{\mathrm{nuc}}({}^4\mathrm{He})c^2 = 25.71\,\mathrm{MeV}$ of nuclear release plus $4m_ec^2 = 2.04\,\mathrm{MeV}$ from the two annihilations, minus the $2m_ec^2$ carried into existence as positrons — the two bookkeepings agreeing, as they must.

The rate-limiting step is the first, $p + p \to d + e^+ + \nu_e$, which is a *weak* process: two protons cannot form a bound diproton (Phenomenon 107.24), so one of them must convert to a neutron during the collision. Its cross-section has never been measured and cannot be, being some twenty orders of magnitude below anything a laboratory can reach; it is computed from the weak coupling of Chapter 101 and the two-nucleon wave function of Proposition 107.23. That a weak interaction gates the whole chain is why the Sun burns for 10^{10} years instead of exploding, and it is a rare case in which the slowness of a rate is the reason there is anybody to measure it.

The flux at the Earth follows from the luminosity alone, with no solar model. Each helium nucleus releases Q and two neutrinos, so

$$\Phi_\nu = \frac{2L_\odot}{4\pi d^2\,Q} = \frac{2 \times 3.828 \times 10^{26}\,\mathrm{W}}{4\pi\,(1.496 \times 10^{11}\,\mathrm{m})^2 \times 4.28 \times 10^{-12}\,\mathrm{J}} = 6.4 \times 10^{14}/(\mathrm{m^2\,s})\,, \tag{107.105}$$

some sixty billion neutrinos through every square centimetre of the reader every second. That this number is right, and that the number of *electron* neutrinos observed was for thirty years only a third of it, is the subject of Chapter 114. □

Remark 107.88 (Carbon, and the resonance predicted from its existence)*.* Beyond helium the chart has a gap: there is no stable nuclide at $A = 5$ or $A = 8$. ${}^8\mathrm{Be}$ is unbound against two α particles by 92 keV and has a half-life of $8.2 \times 10^{-17}\,\mathrm{s}$. Carbon must therefore be assembled from three α particles, and a genuine three-body collision at stellar densities is far too rare. The route is two-step: the tiny equilibrium population of ${}^8\mathrm{Be}$ maintained by $\alpha + \alpha \rightleftharpoons {}^8\mathrm{Be}$ captures a third α, with total release

$$Q = B\left({}^{12}\mathrm{C}\right) - 3B\left({}^4\mathrm{He}\right) = 92.16\,\mathrm{MeV} - 84.89\,\mathrm{MeV} = 7.27\,\mathrm{MeV}\,.$$

Even so the rate is far too small unless the capture is resonant. Hoyle argued that since carbon exists in quantity, ${}^{12}\mathrm{C}$ must have a 0^+ state a few hundred keV above the three-α threshold [1044]; the state was then looked for and found, at an excitation energy of 7.65 MeV, that is 0.38 MeV above threshold. It remains the clearest case in physics of a nuclear level predicted from an astronomical abundance, and the systematic account of how the rest of the chart is built — the *s* and *r* neutron-capture processes, the α ladder, the iron peak — is [1105, 1106], developed in Chapter 52.

Example 107.89 (Deuterium–tritium fusion)**.** The reaction easiest to arrange on Earth is $d + t \to {}^4\mathrm{He} + n$, with

$$Q = B\left({}^4\mathrm{He}\right) - B\left({}^2\mathrm{H}\right) - B\left({}^3\mathrm{H}\right) = 28.30\,\mathrm{MeV} - 2.22\,\mathrm{MeV} - 8.48\,\mathrm{MeV} = 17.59\,\mathrm{MeV}\,,$$

shared by momentum conservation in the inverse ratio of the masses: 3.5 MeV to the α particle and 14.1 MeV to the neutron. Per unit mass of a one-to-one d–t mixture the release is 3.4×10^{14} J/kg, four times the fission value Equation (107.98). It is favoured not for that but for its S factor, which is enhanced by a broad resonance in ${}^5\mathrm{He}$ just above threshold, and the practical operating point is near $T = 1.5 \times 10^8$ K, where Equation (107.101) with $\mu c^2 = 1125$ MeV gives $E_G = 1.18$ MeV and a Gamow peak at $E_0 = 37$ keV. That the α particle is charged and the neutron is not divides the engineering in two: the first can be confined magnetically and heats the plasma (Chapter 66), the second cannot and must be absorbed in a blanket, where it also breeds the tritium.

107.6.3 The heaviest nuclei

Proposition 107.90 (The liquid-drop limit)**.** *Equation* (107.96) *places the end of the chart at* $Z^2/A = 50.1$. *Along the observed neutron-to-proton ratio of the heaviest nuclides,* $A \approx 2.5Z$, *this is* $Z \approx 125$*; but the barrier falls to below* 1 MeV *— comparable with the zero-point energy of the deformation mode itself — already near* $Z = 106$.

Rests on Equations (107.95) *and* (107.96).

Proof. *Derives Proposition 107.90.*

Setting $A = 2.5Z$ in Equation (107.96) gives $Z/2.5 = 50.1$, that is $Z = 125$: beyond that, a nucleus is unstable against an infinitesimal deformation and cannot exist even momentarily. Below it the barrier does not vanish but shrinks rapidly, since Equation (107.95) is the leading term of a surface that flattens as $x \to 1$; the height falls from 6 MeV at $x = 0.7$ (uranium) to a fraction of a MeV at $x \approx 0.9$ ($Z \approx 106$ at the same neutron-to-proton ratio). A barrier that low is transparent, and the liquid drop therefore predicts that no nuclide beyond about $Z = 106$ survives long enough to have a chemical identity. □

Phenomenon 107.91 (Elements exist that the liquid drop forbids)**.** Nuclides up to $Z = 118$ have been synthesized and identified, with half-lives from milliseconds to minutes — many orders of magnitude longer than Proposition 107.90 allows. Their existence is due entirely to shell corrections: the single-particle level density of Table 107.4 is low near a closed shell, which adds several MeV of extra binding to the spherical shape and so restores a barrier where the drop has none.

Rests on Proposition 107.90.

Derivation pending #175

The shell correction: the derivation that a low single-particle level density near a closed shell adds several MeV *of binding relative to the smooth liquid-drop average and so restores a fission barrier — the Strutinsky shell-correction method. It belongs in Appendix A; the phenomenon above uses the statement without it.*

Remark 107.92 (The island, and how far it has been approached). Extrapolating Equation (107.65) past 126 gives 184 as the next neutron closure, and various mean fields put the next proton closure at 114, 120 or 126 — the spread being itself a measure of how far the extrapolation is being pushed, and a reminder of Remark 107.60, that a magic number is a property of a filling and not of the force. Around such a doubly closed configuration the shell correction should be largest and the half-lives longest: the predicted *island of stability*.

The synthesis route is fusion of a heavy target with a beam, followed by evaporation of a few neutrons; the $Z = 117$ case [2162] used a ^{48}Ca beam on a ^{249}Bk target, and is the standard worked example of the identification chain. A recoil separator removes the beam and the transfer products; a surviving recoil is implanted in a position-sensitive silicon detector; and the claim rests on observing, at the same position, a time-correlated sequence of α decays with the right energies terminating in a known nuclide or in spontaneous fission. The identification is thus of a decay *chain*, not of an atom, and the statistics are correspondingly harsh — a handful of chains from weeks of irradiation.

The honest statement is that the island has not been reached. The most neutron-rich nuclides made carry $N = 177$, seven short of 184, and their half-lives are tens of milliseconds; every heavier isotope would require a neutron-rich projectile that no accelerator currently delivers. What has been established is that shell corrections are real and large — without them none of these nuclides would exist for a microsecond — and what has not is the size of the effect at the predicted centre. The standard textbook account of everything in this chapter is [1809].

Chapter 108

The Renormalization Group

A coupling constant is not constant. What the renormalization of Chapter 100 treats as a technical device for removing infinities is, read correctly, a statement about nature: the strength of an interaction depends on the scale at which it is probed, and the dependence is measurable. The idea began as an invariance of the subtraction procedure [1028, 1038], was put in differential form by Callan [1415] and Symanzik [1435], and was then transformed by Kadanoff [1334] and Wilson [1461, 1462] into something far more general — a flow in the space of all theories, generated by integrating out short-distance degrees of freedom.

That reformulation is why this chapter sits at the end of Part XI and faces in two directions. Towards particle physics it gives asymptotic freedom [1501, 1511], the running of α_s measured over more than two decades of energy (Chapter 102), and the effective-field-theory logic [1538, 1647] by which a low-energy theory can be predictive while remaining ignorant of what lies above it. Towards condensed matter it gives universality: the critical exponents of a fluid and of a uniaxial magnet agree because both flow to the same fixed point (Chapter 123), a prediction confirmed to better than a percent [2054]. The chapter also states the technical apparatus honestly — dimensional regularization [1463, 1465], the minimal subtraction schemes [1492, 1613] — and the limits that follow from it, including the triviality of the scalar sector [1662, 2371] that bears on Chapter 104. Standard treatments are [1536, 1918, 2042].

108.1 Conventions carried into this chapter

Everything below is written in SI, as the front matter requires and as Section 100.1.1 sets out in detail for this part. Four points of bookkeeping are used so often that they are collected here.

Notation 108.1 (The renormalization scale, and what it is). The subtraction scale introduced by dimensional continuation in Definition 100.30 is written $\bar{\mu}$ and carries the SI dimension of a *momentum*, $\mathrm{kg\,m/s}$; the corresponding energy is $\bar{\mu}c$, and the corresponding length is $\hbar/\bar{\mu}$. The literature calls both the momentum and the energy μ and does not distinguish them, for the reason recorded in Remark 108.5. Nothing in this chapter depends on the choice, since

$$\mathrm{d}\ln\bar{\mu} = \mathrm{d}\ln(\bar{\mu}c) = -\mathrm{d}\ln(\hbar/\bar{\mu})\ , \tag{108.1}$$

$\hbar$ and c being constants: a derivative with respect to the logarithm of the scale is the same operator whether the scale is measured as a momentum, as an energy or as an inverse length. Where an energy is wanted the symbol E is used, with $E = \bar{\mu}c$. The quantity Q always denotes a momentum transfer, $Q^2 = -q^2 \geq 0$ for a spacelike exchange in the metric of Notation 100.1, so that Qc is the energy scale a scattering experiment resolves and $\hbar c/(Qc) = \hbar/Q$ the length.

Notation 108.2 (Reduced fields and the dimension of an operator). Power counting requires every field to be a pure power of an inverse length, which in SI two of the three are already and the third is not. The Dirac field carries $1/\mathrm{m}^{3/2}$ by Definition 100.2; the geometric gauge field $\mathcal{A}_\mu = (g_s/\hbar)A^a_\mu T_a$ of Equation (102.18) carries $1/\mathrm{m}$; but the scalar field of Equation (106.33) carries $\mathrm{J}^{1/2}/\mathrm{m}^{1/2}$. Define therefore the *reduced* scalar

$$\hat{\phi} := \frac{\phi}{\sqrt{\hbar c}}\,, \qquad \left[\hat{\phi}\right] = 1/\mathrm{m}\,, \tag{108.2}$$

since $\hbar c = 3.161\,527 \times 10^{-26}\,\mathrm{J\,m}$. With this one substitution every field of the Standard Model is a power of an inverse length: $\hat{\phi} \sim \ell^{-1}$, $\psi \sim \ell^{-3/2}$, $\mathcal{A}_\mu \sim \ell^{-1}$, $\partial_\mu \sim \ell^{-1}$. A local operator O built from n_ϕ reduced scalars, n_ψ Dirac fields, n_A geometric gauge fields and n_∂ derivatives therefore carries

$$[O] = 1/\mathrm{m}^D\,, \qquad D = n_\phi + \frac{3}{2}n_\psi + n_A + n_\partial\,, \tag{108.3}$$

and D — an integer or half-integer, and a pure number — is what this chapter means by *the dimension of the operator*. It is the same integer the natural-units literature calls the mass dimension, obtained here without setting anything to one.

Proposition 108.3 (The SI dimension of a coupling)**.** *If the Lagrangian density contains a term $c_O O$ with O of dimension D in the sense of Equation (108.3), then*

$$[c_O] = \mathrm{J}/\mathrm{m}^3 \times \mathrm{m}^D = [\hbar c] \times \mathrm{m}^{D-4}\,, \tag{108.4}$$

so that c_O may always be written

$$c_O = \hat{c}_O\,\hbar c\,\ell_\star^{D-4} = \hat{c}_O\,\hbar c\left(\frac{\hbar c}{E_\star}\right)^{D-4} \tag{108.5}$$

with $\hat{c}_O$ a pure number, $\ell_\star$ a reference length and $E_\star = \hbar c/\ell_\star$ the corresponding energy. The dimensionless strength with which O acts on a process of energy E is

$$\hat{c}_O(E) = \hat{c}_O\left(\frac{E}{E_\star}\right)^{D-4}. \tag{108.6}$$

Rests on Equations (100.2) and (108.3).

Proof. *Derives Proposition 108.3.*

A Lagrangian density is an energy density, J/m^3, by Equation (100.2), and $[O] = 1/\mathrm{m}^D$, which gives the first equality in Equation (108.4). Since $[\hbar c] = \mathrm{J\,m}$, dividing by it leaves m^{D-4}, which is the second. Writing that residual length dependence as a power of a chosen $\ell_\star$ defines $\hat{c}_O$, and Equation (108.6) follows because the only length a process of energy E supplies is $\hbar c/E$, so the matrix element of $c_O O$ carries $c_O(E/\hbar c)^D$ divided by the energy density scale $E^4/(\hbar c)^3$ that any term of the Lagrangian is measured against. □

Example 108.4 (Three couplings of the Standard Model, checked)**.** Take $\ell_\star = \hbar c/E_\star$ with $E_\star$ whatever the context supplies.

- The scalar self-coupling. $\hat{\phi}^4$ has $D = 4$, so $[c_O] = [\hbar c]$ exactly, and $\phi^4 = (\hbar c)^2\,\hat{\phi}^4$ by Equation (108.2), so $-\frac{g}{4!}\phi^4 = -\frac{g\hbar c}{4!}\,\hbar c\,\hat{\phi}^4$, which gives $\hat{c}_O = g\hbar c = \hat{g}$ as Equation (106.34) defines it, the residual $\hbar c/4!$ belonging to the operator's normalization. A marginal coupling is a pure number and does not run with E at tree level.
- The Dirac mass term. $\bar{\psi}\psi$ has $D = 3$, so $[c_O] = [\hbar c] \times 1/\mathrm{m} = \mathrm{J\,m} \cdot 1/\mathrm{m}$, that is an energy; and indeed the term is $-mc^2\bar{\psi}\psi$. Its dimensionless strength at energy E is mc^2/E, which *falls* as E rises: a mass is a relevant coupling and matters at low energy.

- The Fermi interaction. $(\bar\psi\Gamma\psi)^2$ has $D=6$, so $[c_O]=\mathrm{J\,m^3}$, which is Equation (101.7) exactly. Its dimensionless strength is $G_F E^2/(\hbar c)^3$, which *grows* as E rises. That is the whole of Section 101.7.1 read off from one integer.

The length $\ell_F=[G_F/(\hbar c)]^{1/2}=6.7392\times10^{-19}\,\mathrm{m}$ built from Fermi's constant of Equation (101.8) corresponds to the energy $\hbar c/\ell_F=292.8\,\mathrm{GeV}$ of Equation (101.9): the four-fermion description carries in its own dimension the energy at which it must be replaced, and it was.

Remark 108.5 (Reading this chapter against the natural-units literature). Every reference cited below sets $\hbar=c=1$. The dictionary is the one given once in Remark 100.4; three entries are used so heavily here that they are repeated. A scalar field maps to its reduced form Equation (108.2); a coupling of a dimension-D operator maps to the hatted number of Equation (108.5), which is what those texts write with no hat; and the "mass dimension" of an operator is the integer D of Equation (108.3), which in this book is a power of metres. To go the other way, restore an $\hbar c$ for every unit by which the operator's dimension exceeds four. No derivation in this chapter is carried out with $\hbar$ or c set to one; this remark exists so that a reader can check the chapter against its sources, which is a different thing.

108.2 Origins in field theory

108.2.1 The arbitrariness of the subtraction point

Renormalization as Chapter 100 presents it is a procedure with a free choice in it. The divergent parts of the counterterms are forced; their finite parts are not, and Definition 100.54 fixes them by a convention — a mass shell, a subtraction momentum, a minimal pole. Stueckelberg and Petermann observed that the freedom is not a defect but a structure: the transformations relating one admissible choice of finite parts to another compose, are invertible, and therefore form a group, which they named the *groupe de normalisation* [1028]. Gell-Mann and Low turned the observation into a calculation by writing a functional equation for the electromagnetic coupling measured at short distance [1038], and Bogoliubov and Shirkov gave the first systematic account of the resulting apparatus [1166].

The physical content is best seen before any diagram is drawn.

Definition 108.6 (The effective coupling)**.** Let a scattering process be governed at momentum transfer Q by a single dimensionless coupling. Define the *effective coupling* $\bar\alpha(Q)$ to be the number that, inserted in the tree-level formula, reproduces the measured rate at that Q. It is by construction an observable: it is a rate, re-expressed. Its SI dimension is that of a pure number, and Q is a momentum.

Theorem 108.7 (The Gell-Mann–Low functional equation)**.** *There is no scale in a massless theory other than the one the experimenter brings, so $\bar\alpha(Q)$ cannot depend on Q alone. Fixing a reference momentum κ and writing $\alpha_\kappa:=\bar\alpha(\kappa)$, there is a function F of two pure numbers with*

$$\bar\alpha(Q)=F\!\left(\frac{Q}{\kappa},\alpha_\kappa\right),\qquad F(1,x)=x\,, \tag{108.7}$$

and F satisfies the group law

$$F\!\left(\frac{Q}{\kappa_1},\alpha_{\kappa_1}\right)=F\!\left(\frac{Q}{\kappa_2},F\!\left(\frac{\kappa_2}{\kappa_1},\alpha_{\kappa_1}\right)\right) \tag{108.8}$$

for every pair of reference momenta [1038]. Equivalently, in differential form,

$$\frac{\mathrm{d}\bar\alpha}{\mathrm{d}\ln Q}=\beta(\bar\alpha)\,,\qquad \beta(x):=\left.\frac{\partial F(t,x)}{\partial\ln t}\right|_{t=1}, \tag{108.9}$$

a first-order ordinary differential equation whose right-hand side does not contain Q.
Rests on Definition 108.6 and Notation 108.1.

Proof. *Derives Theorem 108.7.*

$\bar{\alpha}$ is a pure number and the only quantities available are Q, κ and α_κ; since $\hbar$ and c are constants they build no further independent dimensionless combination, and Q and κ can enter only through their ratio. That is Equation (108.7), and $F(1,x)=x$ is the definition of α_κ.

For Equation (108.8): the left side computes $\bar{\alpha}(Q)$ using κ_1 as reference. The right side computes it using κ_2, whose own effective coupling is $\alpha_{\kappa_2}=F(\kappa_2/\kappa_1,\alpha_{\kappa_1})$ by Equation (108.7) applied at $Q=\kappa_2$. Both are the same measured number, so they are equal. The reference momentum is a bookkeeping device, and Equation (108.8) is the statement that the answer does not remember which one was used. The transformations $T_\lambda : x \mapsto F(\lambda,x)$ therefore satisfy $T_{\lambda\lambda'}=T_\lambda\circ T_{\lambda'}$ and $T_1=\mathrm{id}$: a one-parameter group, which is what [1028] named.

For Equation (108.9), put $\kappa_2=\kappa_1(1+\epsilon)$ in Equation (108.8) and expand to first order in ϵ. Write $t=Q/\kappa_1$ and $x=\alpha_{\kappa_1}$. The left side is $F(t,x)$, independent of ϵ. On the right, $Q/\kappa_2=t(1-\epsilon)$ and $F(1+\epsilon,x)=x+\epsilon\beta(x)$ by the definition of β in Equation (108.9), so the right side is

$$F(t(1-\epsilon),x+\epsilon\beta(x))=F(t,x)-\epsilon\,\frac{\partial F}{\partial \ln t}+\epsilon\,\beta(x)\frac{\partial F}{\partial x}\,.$$

Equality for all ϵ forces $\partial F/\partial\ln t=\beta(x)\,\partial F/\partial x$. Now use the group law a second time, in the form $F(t,x)=F(t/t',F(t',x))$, differentiated with respect to $\ln t$ at $t=t'$: it gives $\partial F(t,x)/\partial\ln t=\beta(F(t,x))$, which is Equation (108.9) with $\bar{\alpha}=F(t,x)$. □

Corollary 108.8 (The flow is integrable in closed form)**.** *Wherever β does not vanish,*

$$\ln\frac{Q}{\kappa}=\int_{\alpha_\kappa}^{\bar{\alpha}(Q)}\frac{\mathrm{d}x}{\beta(x)}\,. \tag{108.10}$$

Rests on Theorem 108.7 and Equation (108.9).

Proof. *Derives Corollary 108.8.*

Separate the variables in Equation (108.9) and integrate from $Q=\kappa$, where $\bar{\alpha}=\alpha_\kappa$. □

Remark 108.9 (What has and has not been assumed). Theorem 108.7 uses no perturbation theory and no regulator. It uses one physical assumption — that the theory contains no mass with which Q could be compared — and one piece of dimensional analysis. Everything specific to a theory sits in the one function β, and computing β is what perturbation theory is for. Two consequences are worth stating at once, because they are the skeleton of the whole chapter.

First, a *zero* of β is a scale-invariant theory: by Equation (108.9) the coupling then does not move, and by Equation (108.10) the integral diverges as $\bar{\alpha}$ approaches the zero, so the zero is reached only as $\ln Q\to\pm\infty$. Such a point is a *fixed point* (Section 108.4.3).

Second, if β has no zero between α_κ and some value, Equation (108.10) converts a statement about a dimensionless coupling into a statement about an *energy*. Where the integral converges at the upper limit the theory acquires a finite scale out of nothing: this is dimensional transmutation, and it is why Λ_{QCD} exists (Definition 102.38) and why the one-loop Landau pole of Proposition 100.64 has a location at all.

108.2.2 The Callan–Symanzik equation

Theorem 108.7 is about an observable. The equation Callan [1415] and Symanzik [1435] obtained independently is about the Green functions of the theory, which are not observables, and it is more useful precisely because it says how the unobservable parts move too.

Definition 108.10 (Bare and renormalized Green functions)**.** Following Definition 100.45, let ϕ_0 be the bare field of a theory regulated as in Definition 100.30, and let

$$\phi_0 = Z_\phi^{1/2}(\bar{\mu})\,\phi\,, \tag{108.11}$$

define the renormalized field, Z_ϕ being a pure number depending on $\bar{\mu}$, on the regulator and on the dimensionless coupling $\hat{g}$. The bare and renormalized connected n-point functions

$$G_0^{(n)}(x_1,\dots,x_n) := \langle T\,\phi_0(x_1)\cdots\phi_0(x_n)\rangle\,, \qquad G^{(n)} = Z_\phi^{-n/2} G_0^{(n)}\,, \tag{108.12}$$

carry the SI dimension $[\phi]^n$ and are otherwise pure functions of the separations.

Theorem 108.11 (The Callan–Symanzik equation)**.** *The renormalized n-point functions of a theory with one dimensionless coupling $\hat{g}$ and one mass m satisfy*

$$\left[\frac{\partial}{\partial \ln\bar{\mu}} + \beta(\hat{g})\,\frac{\partial}{\partial \hat{g}} + \gamma_m(\hat{g})\,m\frac{\partial}{\partial m} + n\gamma_\phi(\hat{g})\right] G^{(n)} = 0\,, \tag{108.13}$$

where, at fixed bare parameters and fixed regulator,

$$\beta := \frac{\partial \hat{g}}{\partial \ln\bar{\mu}}\,, \qquad \gamma_\phi := \frac{1}{2}\frac{\partial \ln Z_\phi}{\partial \ln\bar{\mu}}\,, \qquad \gamma_m := \frac{\partial \ln m}{\partial \ln\bar{\mu}}\,. \tag{108.14}$$

All three are pure numbers, functions of $\hat{g}$ alone in a mass-independent scheme (Theorem 108.18) [1415, 1435].

Rests on Definition 108.10 and Equation (108.12).

Proof. *Derives Theorem 108.11.*

The bare field ϕ_0, the bare coupling and the bare mass are defined by the regulated Lagrangian; the scale $\bar{\mu}$ enters only through the *choice* of what to call renormalized, that is through Z_ϕ and through the finite parts of the counterterms. Hence

$$\left.\frac{\partial G_0^{(n)}}{\partial \ln\bar{\mu}}\right|_0 = 0\,, \tag{108.15}$$

the subscript recording that the bare quantities are held fixed. This is the whole physical input, and it is the exact counterpart of Equation (108.8): the arbitrary reference must drop out.

Substituting $G_0^{(n)} = Z_\phi^{n/2} G^{(n)}$ into Equation (108.15) and using the chain rule — $G^{(n)}$ depends on $\bar{\mu}$ explicitly, through $\hat{g}(\bar{\mu})$ and through $m(\bar{\mu})$ —

$$0 = \frac{\partial}{\partial \ln\bar{\mu}}\left[Z_\phi^{n/2} G^{(n)}\right] = Z_\phi^{n/2}\left[\frac{n}{2}\frac{\partial \ln Z_\phi}{\partial \ln\bar{\mu}} + \frac{\partial}{\partial \ln\bar{\mu}} + \frac{\partial \hat{g}}{\partial \ln\bar{\mu}}\frac{\partial}{\partial \hat{g}} + \frac{\partial m}{\partial \ln\bar{\mu}}\frac{\partial}{\partial m}\right] G^{(n)}\,.$$

Dividing by $Z_\phi^{n/2}$, which is non-zero, and inserting the definitions Equation (108.14) — with $\partial m/\partial \ln\bar{\mu} = \gamma_m m$ — gives Equation (108.13).

The three coefficients are pure numbers: $\bar{\mu}$ appears only inside a logarithm, so $\partial/\partial\ln\bar{\mu}$ is dimensionless; $\hat{g}$ and Z_ϕ are pure numbers by construction; and $m\,\partial/\partial m$ is dimensionless whatever the SI dimension of m. The equation is therefore a statement among pure numbers acting on a quantity of dimension $[\phi]^n$, and every term of it has that same dimension, as it must. □

Definition 108.12 (Beta function and anomalous dimensions)**.** β is the *beta function* of the coupling. γ_ϕ is the *anomalous dimension* of the field and γ_m that of the mass. The names are earned in Corollary 108.14: at a zero of β the two-point function scales with a power that differs from the one dimensional analysis alone would give, and the difference is $2\gamma_\phi$.

The equation is solved by the method of characteristics, which is the statement that moving the scale and changing the coupling are the same motion seen twice.

Proposition 108.13 (Solution by characteristics)**.** *Let $\Gamma^{(n)}(p_1,\dots,p_n;\hat g,m,\bar\mu)$ be the one-particle-irreducible n-point function in momentum space, whose SI dimension after the overall momentum-conserving factor is removed is an inverse length raised to some power d_n, that is $(1/\mathrm{m})^{d_n}$; and suppose all n momenta are scaled together, $p_i \mapsto \sigma p_i$ with σ a pure number. Then*

$$\Gamma^{(n)}(\sigma p_i;\hat g,m,\bar\mu) = \sigma^{d_n}\,\Gamma^{(n)}(p_i;\bar g(\sigma),\bar m(\sigma),\bar\mu)\exp\!\left[-n\int_1^\sigma \gamma_\phi(\bar g(s))\,\mathrm{d}\ln s\right], \tag{108.16}$$

where the running *coupling and mass solve*

$$\frac{\mathrm{d}\bar g}{\mathrm{d}\ln\sigma} = \beta(\bar g)\,, \qquad \frac{\mathrm{d}\ln\bar m}{\mathrm{d}\ln\sigma} = \gamma_m(\bar g) - 1\,, \tag{108.17}$$

with $\bar g(1)=\hat g$ and $\bar m(1)=m$.

Rests on Theorem 108.11 and Equation (108.12).

Proof. *Derives Proposition 108.13.*

Two facts are combined. The first is Equation (108.13), applied to $\Gamma^{(n)}$, for which the sign of the field term is reversed. The one-particle-irreducible function is obtained by amputating the external propagators, which inverts the field normalization: where Equation (108.12) has $G_0^{(n)} = Z_\phi^{n/2} G^{(n)}$, the corresponding relation is $\Gamma_0^{(n)} = Z_\phi^{-n/2}\Gamma^{(n)}$, as the case $n=2$ makes plain since $\Gamma^{(2)}$ is the reciprocal of $G^{(2)}$. Repeating the argument of Theorem 108.11 with that substitution gives

$$\left[\frac{\partial}{\partial\ln\bar\mu} + \beta\frac{\partial}{\partial\hat g} + \gamma_m m\frac{\partial}{\partial m} - n\gamma_\phi\right]\Gamma^{(n)} = 0\,. \tag{108.18}$$

The second is ordinary dimensional analysis. $\Gamma^{(n)}$ is a function of the momenta, of m, of $\bar\mu$ and of the pure number $\hat g$, and its SI dimension is fixed. Euler's theorem for a homogeneous function — every one of p_i, mc and $\bar\mu$ being a momentum — gives

$$\left[\sum_i p_i\cdot\frac{\partial}{\partial p_i} + m\frac{\partial}{\partial m} + \frac{\partial}{\partial\ln\bar\mu} - d_n\right]\Gamma^{(n)} = 0\,. \tag{108.19}$$

This is not an extra assumption: it says that multiplying every momentum, the mass and the subtraction scale by a common number multiplies $\Gamma^{(n)}$ by that number to the power d_n, which is what having an SI dimension means.

Subtracting Equation (108.19) from Equation (108.18) eliminates $\partial/\partial\ln\bar\mu$ and leaves

$$\left[\sum_i p_i\cdot\frac{\partial}{\partial p_i} - \beta\frac{\partial}{\partial\hat g} + (1-\gamma_m)\,m\frac{\partial}{\partial m} + n\gamma_\phi - d_n\right]\Gamma^{(n)} = 0\,. \tag{108.20}$$

For momenta scaled by σ the first operator is $\partial/\partial\ln\sigma$, and Equation (108.20) becomes a linear first-order partial differential equation in $(\ln\sigma,\hat g,\ln m)$. Its characteristics are exactly Equation (108.17), and along them the equation reduces to $\mathrm{d}\ln\Gamma^{(n)}/\mathrm{d}\ln\sigma = d_n - n\gamma_\phi$, which integrates to Equation (108.16). □

Corollary 108.14 (Anomalous scaling at a fixed point)**.** *Let $\hat{g}_\star$ be a zero of β and set $m = 0$. Then $\bar{g}(\sigma) \equiv \hat{g}_\star$, the exponential in Equation (108.16) becomes a power, and*

$$\Gamma^{(n)}(\sigma p_i) = \sigma^{\,d_n - n\gamma_\phi(\hat{g}_\star)}\,\Gamma^{(n)}(p_i)\ . \tag{108.21}$$

For $n = 2$ this says that the inverse propagator behaves as $p^2\,(p/\bar{\mu})^{-2\gamma_\phi}$ rather than as p^2, so that the field's correlation function in position space falls as $r^{-2}\,(\hbar/\bar{\mu}r)^{2\gamma_\phi}$ rather than as r^{-2}. The reference scale is written explicitly because a non-integer power of a dimensionful quantity is not defined without one: $p/\bar{\mu}$ and $\bar{\mu}r/\hbar$ are pure numbers by Notation 108.1, and the literature's $p^{2-2\gamma_\phi}$ and $r^{-2-2\gamma_\phi}$ are these expressions with $\bar{\mu}$ suppressed. The exponent is shifted from its classical value by $2\gamma_\phi$: hence the name.

Rests on Equation (108.16) and Notation 108.1.

Proof. *Derives Corollary 108.14.*

Immediate from Equation (108.16) with β vanishing, so that $\gamma_\phi(\bar{g}(s))$ is constant and the integral is $\ln\sigma$. The position-space statement follows by Fourier transform in $3+1$ dimensions (Chapter 17), a power p^{-a} in four dimensions transforming to r^{a-4} up to a constant. □

Remark 108.15 (The relation to the observable flow). Theorem 108.7 and Theorem 108.11 are the same statement addressed to different objects. Equation (108.9) governs a measured effective coupling and contains only β; Equation (108.13) governs Green functions, which depend on the arbitrary Z_ϕ, and pays for that generality with the extra term $n\gamma_\phi$. In a scheme where the renormalized coupling is *defined* as the effective coupling at $Q = \bar{\mu}$ the two beta functions coincide, which is how Proposition 100.63 was obtained in Chapter 100 without writing Equation (108.13) at all.

Example 108.16 (The one-loop running of the electromagnetic coupling)**.** For quantum electrodynamics Proposition 100.63 gives

$$\frac{\mathrm{d}\alpha}{\mathrm{d}\ln\bar{\mu}} = \frac{2\alpha^2}{3\pi}\sum_f N_c^{(f)}\left(\frac{q_f}{e}\right)^2 + O\!\left(\alpha^3\right)\ , \tag{108.22}$$

with $\alpha = e^2/(4\pi\varepsilon_0\hbar c)$ the dimensionless combination of Equation (100.3). Inserting this in Equation (108.10) with $\beta(x) = Bx^2$, $B := (2/3\pi)\sum_f N_c^{(f)}(q_f/e)^2$, gives

$$\frac{1}{\alpha(\bar{\mu})} = \frac{1}{\alpha(\kappa)} - B\ln\frac{\bar{\mu}}{\kappa}\,, \tag{108.23}$$

the closed form behind the measured six-per-cent growth of α between zero momentum transfer and the Z pole (Phenomenon 100.62). The sign is positive, so $1/\alpha$ decreases with $\bar{\mu}$ and the coupling grows: the vacuum screens (Remark 100.34). The measurement itself belongs to Chapter 112, where the value at zero momentum transfer is extracted from the electron anomaly.

108.2.3 Renormalization schemes and their arbitrariness

A renormalized coupling is not a measurement. It is a measurement processed through a convention, and the convention has to be quoted with it.

Definition 108.17 (Four schemes)**.** The prescriptions in use fix the finite parts of the counterterms in one of four ways.

- *On shell.* The parameters are defined by physical quantities at physical kinematics: the pole of the propagator is the mass, the residue is one, the coupling is the value at zero momentum transfer. This is Definition 100.46, and it is why $\alpha^{-1} = 137.035999177(21)$ [2431] is a number with no scale attached.

- *Momentum subtraction.* A Green function is required to take its tree-level value at a chosen spacelike point $Q^2 = \bar{\mu}^2$. This keeps the mass dependence of the counterterms and hence produces a beta function that depends on $m/\bar{\mu}$; it is the scheme in which decoupling is automatic (Theorem 108.64).

- *Minimal subtraction* MS. The counterterms remove the poles $1/\varepsilon$ of Definition 100.30 and nothing else [1492].

- *Modified minimal subtraction* $\overline{\text{MS}}$. The counterterms remove the whole recurring combination Equation (100.75), $\varepsilon^{-1} - \gamma_{\text{E}} + \ln 4\pi$, which by Lemma 100.28 accompanies every pole [1613]. This is the default in which α_s and the quark masses are quoted, including in Table 102.1 and in Equation (102.47).

The first two are *mass-dependent*, the last two *mass-independent*, in the sense made precise next.

Theorem 108.18 ('t Hooft's mass-independence)**.** *In minimal subtraction the beta function and the anomalous dimensions of Equation* (108.14) *are functions of the dimensionless couplings alone. No mass, and no ratio $m/\bar{\mu}$, appears in them [1492].*

Rests on Equation (108.14) *and Theorem 100.52.*

Derivation. *Derives Theorem 108.18.*

Two properties of the counterterms do the work, and both were established in Chapter 100. First, by Remark 100.53 a counterterm is a *polynomial* in the masses and external momenta, of degree the superficial divergence of the graph it cancels: the subtraction operator of Theorem 100.52 is a Taylor expansion, so nothing non-analytic in m can appear in it. Second, in minimal subtraction the counterterm is by definition a pure pole $a_k(\hat{g})/\varepsilon^k$ with no finite part attached.

Now count dimensions. The coefficient a_k multiplies an operator already present in the Lagrangian, so it is a pure number; a polynomial in the masses that is also a pure number and carries no $\bar{\mu}$ can only be a constant, because the only dimensionless combination available is $mc/\bar{\mu}$ and that is not polynomial in m unless it is absent. Hence $a_k = a_k(\hat{g})$. Since β and the γ's are computed from the a_k by differentiating with respect to $\ln\bar{\mu}$ at fixed bare coupling, they inherit the property.

The price is stated in Section 108.6.1: a heavy particle does not disappear from a mass-independent beta function when the scale drops below its mass, and must be removed by hand. □

Theorem 108.19 (The first two beta coefficients are scheme independent)**.** *Write the beta function of a dimensionless coupling a as*

$$\frac{\mathrm{d}a}{\mathrm{d}\ln\bar{\mu}} = -\beta_0 a^2 - \beta_1 a^3 - \beta_2 a^4 - \cdots . \tag{108.24}$$

Under any redefinition $a \mapsto a' = a + c_1 a^2 + c_2 a^3 + \cdots$ with finite c_i — which is what a change of scheme is, by Proposition 100.55 — the coefficients β_0 and β_1 are unchanged. Every coefficient from β_2 on is changed.

Rests on Proposition 100.55.

Proof. *Derives Theorem 108.19.*

Differentiate the redefinition:

$$\frac{\mathrm{d}a'}{\mathrm{d}\ln\bar{\mu}} = \left(1 + 2c_1 a + 3c_2 a^2 + \cdots\right)\left(-\beta_0 a^2 - \beta_1 a^3 - \beta_2 a^4\right) = -\beta_0 a^2 - (\beta_1 + 2c_1\beta_0)\, a^3 + O\left(a^4\right) .$$

Now express the right-hand side in a'. Inverting the redefinition, $a = a' - c_1 a'^2 + O(a'^3)$, so $a^2 = a'^2 - 2c_1 a'^3 + O(a'^4)$ and $a^3 = a'^3 + O(a'^4)$. Substituting,

$$\frac{\mathrm{d}a'}{\mathrm{d}\ln\bar{\mu}} = -\beta_0 a'^2 + 2c_1\beta_0 a'^3 - (\beta_1 + 2c_1\beta_0)\,a'^3 + O\Big(a'^4\Big) = -\beta_0 a'^2 - \beta_1 a'^3 + O\Big(a'^4\Big)\,.$$

The c_1 terms cancel identically. Carrying the same computation one order further — keeping c_2 in the derivative, inverting the redefinition to $a = a' - c_1 a'^2 + (2c_1^2 - c_2)a'^3 + O(a'^4)$ and collecting the coefficient of a'^4 — leaves

$$\beta_2' = \beta_2 - c_1^2\beta_0 + c_2\beta_0 - c_1\beta_1\,, \tag{108.25}$$

which depends on c_1 and c_2 and so is not invariant. □

Corollary 108.20 (What is physical in a beta function)**.** *The statement "$b_0 = 11 - \frac{2}{3}n_f$" of Theorem 102.35 is a property of* SU(3) *with n_f flavours and of nothing else; so is the two-loop coefficient. From three loops on, a quoted coefficient is a property of the theory* and *of the scheme, and the two cannot be separated. Correspondingly, a quoted* value *of a coupling is meaningless without both its scheme and its scale: $\alpha_s = 0.118$ is a statement about* $\overline{\text{MS}}$ *at $E = m_Z c^2 = (91.1880 \pm 0.0020)\,\text{GeV}$ [2427] and about nothing else, and the same holds for the top mass and for the three different weak mixing angles listed in Remark 104.34, whose values differ by about one per cent — hundreds of times the experimental errors.*

Rests on Theorems 102.35 and 108.19 and Equation (108.24)*.*

Proof. *Derives Corollary 108.20.*

The invariance of β_0 and β_1 is Theorem 108.19, and the non-invariance of β_2 is the last line of its proof. That a *value* of the coupling is scheme dependent follows from the redefinition $a \mapsto a + c_1 a^2 + \cdots$ itself, which changes the number at order a^2, and its scale dependence is Equation (108.24) with a non-vanishing right-hand side. □

Remark 108.21 (Why $\overline{\text{MS}}$ won). Minimal subtraction is not more physical than any other scheme; it is more *convenient*, for three reasons that are worth separating from any claim of naturalness. It is mass-independent, so the beta function is a polynomial in the couplings and multi-loop calculations are tractable. It respects gauge invariance, because dimensional regularization does (Section 108.3.1), so Ward–Takahashi identities hold order by order without extra counterterms. And the modification of [1613] — removing $-\gamma_{\text{E}} + \ln 4\pi$ along with the pole — removes a combination that would otherwise appear at every order in every process, so that residual coefficients are of order one rather than of order ten. That third point is bookkeeping and is the reason the modified scheme, not the minimal one, is the default for every number quoted in Chapter 102.

108.3 Regularization

108.3.1 Dimensional regularization

The machinery of dimensional continuation — the master integral, the measure, the poles — is Lemma 100.28 and Definition 100.30, and is not repeated. What belongs here is the reason the method *generates a scale*, since that is where the whole renormalization group comes from, and it is a purely dimensional argument that can be given in SI once and for all.

Proposition 108.22 (A coupling's dimension away from four)**.** *In d spacetime dimensions — one time and $d-1$ space, with $x^0 = ct$ so that $\mathrm{d}^d x$ carries m^d — a reduced scalar field carries*

$$\left[\hat{\phi}\right] = 1/\mathrm{m}^{(d-2)/2}\,, \tag{108.26}$$

and the coefficient of $\hat{\phi}^4$ carries

$$[c_4] = [\hbar c] \times \mathrm{m}^{d-4}\,. \tag{108.27}$$

The same is true of the squared gauge coupling. At $d = 4$ both are pure multiples of $\hbar c$; at $d = 4 - 2\varepsilon$ they are not, and a length must be supplied to restore a dimensionless coupling.
Rests on Equations (100.2) *and* (102.18).

Proof. *Derives Proposition 108.22.*

The exponent $S/\hbar$ of Chapter 106 is a pure number and $S = c^{-1}\int \mathrm{d}^d x\,\mathcal{L}$ by Equation (100.2), so $[\mathcal{L}] = [\hbar c] \times 1/\mathrm{m}^d$: in four dimensions that is $\mathrm{J\,m} \times 1/\mathrm{m}^4 = \mathrm{J/m}^3$, as it should be. The kinetic term is $\partial_\mu\hat{\phi}\,\partial^\mu\hat{\phi}$ up to a factor of $\hbar c$, carrying $1/\mathrm{m}^2[\hat{\phi}]^2$, and equating exponents gives Equation (108.26). Then $[\hat{\phi}^4] = 1/\mathrm{m}^{2(d-2)}$ and $[c_4] = [\hbar c]1/\mathrm{m}^d\mathrm{m}^{2(d-2)} = [\hbar c]\mathrm{m}^{d-4}$, which is Equation (108.27). For a gauge theory the geometric field $\mathcal{A}_\mu$ of Equation (102.18) must keep $1/\mathrm{m}$ in every dimension, since it appears added to ∂_μ; the Yang–Mills action then forces the squared coupling to carry the same m^{d-4}. □

Corollary 108.23 (Where $\bar{\mu}$ comes from, and why it must be arbitrary)**.** *Writing the length of Equation* (108.27) *as $\hbar/\bar{\mu}$ with $\bar{\mu}$ a momentum,*

$$c_4 = \hat{g}\,\hbar c \left(\frac{\bar{\mu}}{\hbar}\right)^{2\varepsilon}, \qquad d = 4 - 2\varepsilon\,, \tag{108.28}$$

with $\hat{g}$ a pure number; the factor is exactly the $(\bar{\mu}/\hbar)^{4-d}$ that Equation (100.48) *attaches to the loop measure. Nothing selects $\bar{\mu}$. The theory was specified by c_4, which is $\bar{\mu}$-independent; the split into a number $\hat{g}$ and a scale is a choice, and physical quantities cannot know about it. That is Equation* (108.15)*, and therefore Theorem 108.11: the running of couplings is the shadow of an arbitrary choice made when the regulator was installed.*
Rests on Equations (100.48) *and* (108.27).

Proof. *Derives Corollary 108.23.*

Equation (108.28) is Equation (108.27) with the residual length written as $\hbar/\bar{\mu}$ and $d - 4 = -2\varepsilon$, and the independence of c_4 from $\bar{\mu}$ is the statement that c_4 was fixed before the split was made. □

Remark 108.24 (Why this regulator and not another). Dimensional regularization was introduced independently by 't Hooft and Veltman [1463] and by Bollini and Giambiagi [1465], and it is the only regulator in common use that preserves gauge invariance and Lorentz invariance simultaneously. The reason is narrow and worth naming: it never introduces a preferred momentum into the integrand, so the shift $\ell \mapsto \ell + k$ of a loop momentum remains legitimate, and every Ward–Takahashi identity of Chapter 100 — each of which is proved by exactly such a shift — survives order by order. A momentum cutoff does introduce a preferred momentum, the shift is then illegal by a surface term, and the surface term is precisely a photon mass counterterm forbidden by gauge invariance. This is not an aesthetic preference. It is what made 't Hooft's proof that the spontaneously broken gauge theory of Chapter 104 is renormalizable possible at all, and the same point is recorded there in the discussion preceding Remark 104.50.

Remark 108.25 (The one defect, and the standing of d). Two honest qualifications.

The first is technical and specific to chiral theories: there is no definition of γ^5 away from four dimensions that both anticommutes with every γ^μ and keeps $\mathrm{tr}\,(\gamma^5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma)$ non-zero, because the two requirements are contradictory once the number of dimensions is not four. Every consistent prescription therefore breaks one of them and needs compensating finite counterterms; the point is taken up where it bites, in Remark 104.50. That the difficulty is unavoidable rather than a failure of ingenuity is the same fact that makes the

Adler–Bell–Jackiw anomaly of Theorem 106.87 real: a symmetry of the classical Lagrangian that no regulator can preserve is not a symmetry of the quantum theory.

The second is a matter of standing. The physical world of this book has three space dimensions and one of time, and every instantiated result below is at $d = 4$. The continuation in d is a device for defining an integral, exactly as the continuation of a factorial to the gamma function is a device for defining a product; the limit $\varepsilon \to 0$ is taken before any number is quoted. The same caution applies to the ε expansion of Section 108.4.4, where the continuation is in the number of *space* dimensions and the physical case is always the observed three.

108.3.2 Other regulators and the counterterm expansion

Every regulator breaks something. Knowing what each breaks is how one chooses, and it is also how one recognizes a spurious result.

- *Pauli–Villars.* Subtract from each propagator the same propagator with a large mass M, so that the difference falls two powers faster and the loop converges [945]. It preserves Lorentz invariance and abelian gauge invariance — a massive photon regulator does not spoil U(1) — and it fails for non-abelian theories, where a mass term for the gauge field is not gauge invariant and no set of regulator masses restores the symmetry.

- *Momentum cutoff.* Restrict $|\boldsymbol{\ell}| < \Lambda$. This is the regulator with the clearest physical meaning, and it is the one Wilson's construction uses (Section 108.4.2); it breaks both Lorentz invariance, by singling out a frame, and gauge invariance, by forbidding the shift of Remark 108.24.

- *The lattice.* Replace continuous space and Euclidean time by a hypercubic array of spacing a, as in Theorem 106.68. Gauge invariance is preserved *exactly*, at finite a, because the variables are group elements on links rather than algebra elements; rotational and Lorentz invariance are broken to a discrete subgroup and recovered only in the limit $a \to 0$. The lattice is the only regulator in this list that is non-perturbative, which is why it is the one that can address confinement (Section 102.7). Its own obstruction is exact and unavoidable: a lattice Dirac operator that is local, hermitian and chirally symmetric necessarily describes an equal number of left- and right-handed species, so the naive discretization of a single Weyl fermion produces sixteen [1673]. Like the γ^5 problem above, this is the anomaly asserting itself through whatever regulator is chosen.

The classification of theories by their divergences follows from Equation (108.3) alone.

Theorem 108.26 (Power counting in $3+1$)**.** *Consider a theory in four spacetime dimensions whose interaction vertices are local operators of dimensions D_i in the sense of Equation (108.3). A connected amputated diagram with E_B external boson lines, E_F external fermion lines and V_i vertices of type i has superficial degree of divergence*

$$D = 4 - E_B - \frac{3}{2} E_F + \sum_i V_i \left(D_i - 4 \right) , \tag{108.29}$$

meaning that the loop integrand behaves as $\ell^D \, \mathrm{d}^{4L}\ell / \ell^{4L}$ at large loop momentum.
Rests on Equations (100.1) and (108.3).

Proof. *Derives Theorem 108.26.*

Count powers of momentum. Each of the L loop integrations contributes $+4$; each internal boson propagator, going as $1/(\ell^2 - m^2 c^2)$ by Equation (100.1), contributes -2; each

internal fermion propagator, going as $1/(\gamma^\mu \ell_\mu - mc)$, contributes -1; and a vertex carrying $n_\partial^{(i)}$ derivatives contributes $+n_\partial^{(i)}$. Writing I_B, I_F for the numbers of internal lines,

$$D = 4L - 2I_B - I_F + \sum_i V_i n_\partial^{(i)} . \tag{108.30}$$

Three topological identities relate these. The number of loops is $L = I_B + I_F - V + 1$ with $V = \sum_i V_i$, since each internal line carries an integration and each vertex a delta function, one of which is the overall momentum conservation. Counting line ends at the vertices, with b_i and f_i the numbers of boson and fermion fields at a vertex of type i,

$$2I_B + E_B = \sum_i V_i b_i , \qquad 2I_F + E_F = \sum_i V_i f_i . \tag{108.31}$$

Substituting L into Equation (108.30),

$$D = 2I_B + 3I_F - 4V + 4 + \sum_i V_i n_\partial^{(i)} ,$$

and then eliminating I_B and I_F with Equation (108.31),

$$D = 4 - E_B - \frac{3}{2} E_F + \sum_i V_i \left(b_i + \frac{3}{2} f_i + n_\partial^{(i)} - 4 \right) .$$

The bracket is $D_i - 4$ by Equation (108.3), which is Equation (108.29). □

Corollary 108.27 (The three classes)**.** *The sign of $D_i - 4$ in Equation (108.29) sorts every theory into one of three kinds.*

- *If every $D_i < 4$ the theory is* super-renormalizable*: D decreases with the order, so only finitely many diagrams in the whole perturbation series diverge.*
- *If every $D_i = 4$ the theory is* renormalizable*: D is independent of the order, so the divergent amplitudes are those with few enough external legs — a finite list — and finitely many counterterms suffice at every order. Quantum electrodynamics is the case $D_i = 4$ for the single vertex $\bar\psi \gamma^\mu \psi \mathcal{A}_\mu$, and Equation (108.29) then reduces to Equation (100.26).*
- *If some $D_i > 4$ the theory is* non-renormalizable*: D grows without bound with the number of such vertices, so every amplitude diverges at sufficiently high order and no finite set of counterterms suffices. Fermi's theory has $D_i = 6$ and, with four external fermions, $D = 2V - 2$: this is Proposition 101.7 read as a statement about diagrams.*

Rests on Theorem 108.26 and Equation (108.29).

Proof. *Derives Corollary 108.27.*

Each case reads off the sign of the last term of Equation (108.29), which is the only place the order of perturbation theory enters. For the Fermi instance, put $E_B = 0$, $E_F = 4$ and $D_i = 6$ in Equation (108.29): $D = 4 - 6 + 2V = 2V - 2$. □

Remark 108.28 (Non-renormalizable is not the same as useless). It was regarded so for thirty years, and the reversal of that judgement is the single most consequential thing the renormalization group did to particle physics. Corollary 108.27 says that a theory with $D_i > 4$ cannot be extended to arbitrarily high energy with a finite number of parameters. It does not say the theory is wrong, and it does not say it is unpredictive: at energies below the scale $E_\star$ set by Equation (108.5), the operator's contribution is suppressed by $(E/E_\star)^{D-4}$, so a calculation to a stated accuracy needs only finitely many operators. Fermi's theory predicted the muon lifetime to a few per cent for forty years without knowing that the W existed, and that was not luck. The systematic version of the argument is Section 108.6.

Remark 108.29 (Superficial is not actual: what BPHZ adds). Theorem 108.26 controls the *overall* divergence of a diagram; a graph whose D is negative may still contain a divergent subgraph, and at two loops the electron self-energy already contains two vertex subgraphs that overlap (Definition 100.51). That the recursive subtraction terminates, that the forest formula solves it in closed form, and that the counterterms so generated are exactly those of the local Lagrangian, is the theorem of Bogoliubov and Parasiuk [1103], Hepp [1331] and Zimmermann [1411], stated as Theorem 100.52. It is what licenses every statement in this chapter about "the" beta function of a renormalizable theory: without it, a beta function would be an order-by-order accident rather than a property.

108.4 The Wilsonian picture

Everything above treats renormalization as a way of removing infinities and reads the group structure off the residual arbitrariness. Kadanoff [1334] and Wilson [1461, 1462] reversed the logic. The cutoff is kept, and taken seriously as a statement that the description is incomplete above some momentum; the transformation of interest is the one that *lowers* the cutoff by integrating out what lies between the old one and the new; and the object that flows is not a coupling but the whole Hamiltonian.

Notation 108.30 (Units in the statistical sections). The four subsections of Section 108.4 that follow this one treat classical statistical mechanics in three space dimensions, where the Boltzmann weight is e^{-H/k_BT} with H an energy in J, $k_B = 1.380\,649 \times 10^{-23}\,\mathrm{J/K}$ exact by the 2019 SI, and lengths in m. No $\hbar$ appears in a classical partition function and none is written; that is the correct SI statement, not an omission. The reduced temperature $t := (T - T_c)/T_c$ and the dimensionless field h conjugate to the order parameter are pure numbers, as is the block factor $b > 1$. Where the same equations are used for a field theory the correspondence is the Euclidean dictionary of Table 106.1, whose second row is the entry used here: k_BT takes the place of $\hbar$, the continuation to imaginary time itself being Definition 106.12.

108.4.1 Block spins and Kadanoff scaling

Definition 108.31 (The block transformation)**.** Let spins $s_{\boldsymbol{n}}$ sit on a cubic lattice of spacing a in d space dimensions, with $d = 3$ in the physical case, and let the Hamiltonian near the critical point be characterized by the reduced temperature t and the field h. Partition the lattice into cells of side ba containing b^d spins each, and assign to each cell a single block variable — the sign of the cell's total spin, or its normalized sum. Summing over the configurations of the original spins consistent with each block value defines a new Hamiltonian for the block variables, on a lattice of spacing ba [1334].

Theorem 108.32 (Kadanoff scaling and the exponent relations)**.** *Suppose that, sufficiently near the critical point and for b small compared with the correlation length measured in lattice units, the blocked Hamiltonian has the same form as the original with new parameters, and that these depend on the old ones through pure powers*

$$t' = b^{y_t}\, t\,, \qquad h' = b^{y_h}\, h\,, \tag{108.32}$$

for two numbers $y_t, y_h > 0$. Then the singular part of the free energy per site satisfies

$$f_s(t,h) = b^{-d} f_s(b^{y_t} t, b^{y_h} h)\ , \tag{108.33}$$

the correlation length satisfies $\xi(t) = b\,\xi(b^{y_t}t)$, and consequently

$$\nu = \frac{1}{y_t}\,, \quad \alpha = 2 - \frac{d}{y_t}\,, \quad \beta = \frac{d - y_h}{y_t}\,, \quad \gamma = \frac{2y_h - d}{y_t}\,, \quad \delta = \frac{y_h}{d - y_h}\,, \quad \eta = d + 2 - 2y_h\,. \tag{108.34}$$

Every exponent is determined by the two numbers y_t and y_h; the Rushbrooke and Griffiths inequalities of Phenomenon 123.4 hold as equalities; and in addition

$$2 - \alpha = d\nu \tag{108.35}$$

— hyperscaling*, which involves the dimension of space and which the thermodynamic arguments cannot produce.*

Rests on Definition 108.31, Phenomenon 123.4, and Equation (123.6).

Derivation. *Derives Theorem 108.32.*

The block transformation is a partial sum over configurations, so the partition function is unchanged; the total free energy is therefore unchanged, and only the number of sites has altered, from N to N/b^d. Hence $N f_s(t,h) = \left(N/b^d\right) f_s(t',h')$, which with Equation (108.32) is Equation (108.33). The correlation length is a physical distance, so it is the same before and after; measured in units of the new lattice spacing it is b times smaller, which is $\xi(t) = b\,\xi(t')$.

Now set $h = 0$ and choose $b = |t|^{-1/y_t}$, which sends $t' \to \pm 1$ and so removes all t dependence from the arguments. The correlation-length relation becomes $\xi(t) = |t|^{-1/y_t}\,\xi(\pm 1)$, and comparison with the definition $\xi \sim |t|^{-\nu}$ gives $\nu = 1/y_t$. The free-energy relation becomes

$$f_s(t,h) = |t|^{d/y_t}\;\Phi\!\left(\frac{h}{|t|^{y_h/y_t}}\right), \qquad \Phi(z) := f_s(\pm 1, z)\,, \tag{108.36}$$

which is exactly Widom's homogeneity assumption Equation (123.6) [1323] with

$$2 - \alpha = \frac{d}{y_t} = d\nu\,, \qquad \beta\delta = \frac{y_h}{y_t}\,. \tag{108.37}$$

The first of these is Equation (108.35) and the third entry of Equation (108.34). The remaining exponents follow by differentiating Equation (108.36), exactly as in the derivation accompanying Phenomenon 123.4: $m = -\partial f_s/\partial h \sim |t|^{(d-y_h)/y_t}$ gives β; $\chi = -\partial^2 f_s/\partial h^2 \sim |t|^{(d-2y_h)/y_t}$ gives γ; and setting $t = 0$ in Equation (108.33) with $b = h^{-1/y_h}$ gives $m \sim h^{(d-y_h)/y_h}$, that is $\delta = y_h/(d-y_h)$. For η, apply the same argument to the pair correlation function, which carries 2 powers of the field and so scales with $b^{-2(d-y_h)}$ at separations rescaled by b; matching to $G(r) \sim r^{-(d-2+\eta)}$ gives $\eta = d + 2 - 2y_h$.

The two classical relations are now identities. Rushbrooke:

$$\alpha + 2\beta + \gamma = 2 - \frac{d}{y_t} + \frac{2d - 2y_h}{y_t} + \frac{2y_h - d}{y_t} = 2\,.$$

Griffiths: $\beta(\delta - 1) = \frac{d-y_h}{y_t} \cdot \frac{y_h - (d-y_h)}{d-y_h} = \frac{2y_h - d}{y_t} = \gamma$. And Fisher's relation $\gamma = \nu(2-\eta)$ follows from $\nu(2-\eta) = (2 - d - 2 + 2y_h)/y_t = \gamma$. □

Remark 108.33 (What Kadanoff assumed, and what Wilson supplied). Theorem 108.32 derives every exponent relation from Equation (108.32), and derives nothing at all about the two numbers y_t and y_h. Worse, the assumption that the blocked Hamiltonian has *the same form* as the original is false as stated: blocking a nearest-neighbour Ising model generates couplings between next-nearest neighbours, four-spin couplings, and infinitely many others. Kadanoff's argument therefore explains why the exponents satisfy relations and why the same relations hold for a fluid and a magnet, but it cannot compute a single exponent and it rests on a premise known to be wrong.

Wilson's contribution was to accept the proliferation of couplings and to work in the infinite-dimensional space of Hamiltonians, where Equation (108.32) is not an assumption but the linearization of a flow about a fixed point, and where y_t and y_h are its eigenvalues, computable. That is Section 108.4.2 and Section 108.4.3. The three-dimensional critical exponents themselves, and the experimental verdict, are Section 108.4.4; the phenomenology they explain is collected in Section 123.5.

108.4.2 Integrating out momentum shells

Definition 108.34 (The Wilsonian transformation)**.** Let the theory be defined by an effective Hamiltonian $H_\Lambda[\phi]$ for field modes $\tilde{\phi}(\boldsymbol{k})$ with $|\boldsymbol{k}| < \Lambda$, Λ an inverse length. Split $\phi = \phi_< + \phi_>$, where $\phi_<$ contains the modes with $|\boldsymbol{k}| < \Lambda/b$ and $\phi_>$ those in the shell $\Lambda/b < |\boldsymbol{k}| < \Lambda$. Define the new Hamiltonian by

$$\exp\!\left(-\frac{H_{\Lambda/b}\left[\phi_<\right]}{k_B T}\right) := \int \mathcal{D}\phi_> \, \exp\!\left(-\frac{H_\Lambda\left[\phi_< + \phi_>\right]}{k_B T}\right), \tag{108.38}$$

and then restore the original cutoff by rescaling $\boldsymbol{k} \mapsto b\boldsymbol{k}$, equivalently $\boldsymbol{x} \mapsto \boldsymbol{x}/b$, together with a rescaling of the field chosen to keep the coefficient of the gradient term fixed. The composite operation $\mathcal{R}_b$ maps Hamiltonians to Hamiltonians on the same space and satisfies $\mathcal{R}_b \mathcal{R}_{b'} = \mathcal{R}_{bb'}$ [1461, 1462, 1536].

Remark 108.35 (Why it is a semigroup and not a group). Equation (108.38) integrates degrees of freedom *out*. Information is discarded and cannot be recovered: $\mathcal{R}_b$ has no inverse, and the "group" of the name is in fact a semigroup, $b \geq 1$ only. The name is historical, from [1028] where the transformations really did form a group because nothing was being discarded — only relabelled. The distinction is not pedantic: it is the reason the flow has a direction, which is the subject of Section 108.7.2.

The transformation acts on the infinite-dimensional space of local Hamiltonians. Its first, and purely dimensional, effect is already decisive.

Proposition 108.36 (Tree-level scaling of the couplings)**.** *Write the Hamiltonian in d space dimensions as*

$$\frac{H\left[\phi\right]}{k_B T} = \int \mathrm{d}^d x \left[\frac{1}{2}\left(\nabla\phi\right)^2 + \frac{1}{2} r \phi^2 + \frac{u}{4!}\phi^4 + \cdots\right], \tag{108.39}$$

the exponent being a pure number, so that $[\phi] = \mathrm{m}^{(2-d)/2}$, $[r] = 1/\mathrm{m}^2$ and

$$[u] = \mathrm{m}^{d-4}. \tag{108.40}$$

Under the rescaling step of Definition 108.34, and ignoring the shell integration itself, the couplings transform as

$$r \mapsto b^2 r, \qquad u \mapsto b^{4-d} u, \qquad u_6 \mapsto b^{6-2d} u_6, \tag{108.41}$$

u_6 being the coefficient of $\phi^6/6!$.

Rests on Definition 108.34.

Proof. *Derives Proposition 108.36.*

The exponent of a Boltzmann weight is dimensionless, so $\int \mathrm{d}^d x (\nabla\phi)^2$ is a pure number: $\mathrm{m}^d \times 1/\mathrm{m}^2 \times [\phi]^2 = 1$ gives $[\phi] = \mathrm{m}^{(2-d)/2}$. Then $\mathrm{m}^d [r][\phi]^2 = 1$ gives $[r] = 1/\mathrm{m}^2$, and $\mathrm{m}^d [u][\phi]^4 = 1$ gives $[u] = \mathrm{m}^{2d-4-d} = \mathrm{m}^{d-4}$, which is Equation (108.40). A coupling of dimension $1/\mathrm{m}^y$ measured against a length shrunk by b becomes b^y times itself, which gives Equation (108.41) with $y_r = 2$, $y_u = 4 - d$, $y_6 = 6 - 2d$. □

Corollary 108.37 (The upper critical dimension is four)**.** *u has SI dimension m^{d-4}, so it is dimensionless exactly at $d = 4$ and grows under coarse graining for $d < 4$. In the observed three space dimensions $[u] = 1/\mathrm{m}$, the quartic coupling is relevant, and mean-field theory — which is Equation (108.39) with the fluctuations dropped — must fail. For $d > 4$ it is irrelevant and mean-field exponents are exact. The sextic coupling has $y_6 = 6 - 2d$, which is negative for every $d > 3$ and* vanishes *at $d = 3$: in the observed three dimensions the sextic term is marginal at the Gaussian fixed point rather than relevant, so it opens no new relevant direction, and Equation (108.39) may be truncated at fourth order.*

Rests on Equations (108.39) to (108.41).

Proof. *Derives Corollary 108.37.*

Set $d = 4$ in Equation (108.40) to get m^0, and $d = 3$ to get $1/\mathrm{m}$; the eigenvalues are Equation (108.41), and $4 - d > 0$ for $d < 4$ while $6 - 2d \le 0$ for $d \ge 3$, with equality exactly at $d = 3$. That $d = 3$ is the borderline dimension for ϕ^6 in precisely the way $d = 4$ is for ϕ^4 is Equation (108.50) read at $n = 3$; the truncation is justified because a marginal coupling generates no new relevant scaling field, not because y_6 is negative there. Mean-field theory is Equation (108.39) evaluated at its saddle point, with the fluctuations that the shell integration would supply dropped; that same approximation, in its statistical-mechanics setting, belongs to Section 123.4. □

Remark 108.38 (This is the Ginzburg criterion, seen from above). The Ginzburg criterion [1181] compares the fluctuation of the order parameter within a correlation volume to its mean, and yields a computable window of reduced temperature inside which mean-field theory fails for $d < 4$; that estimate belongs to Section 123.4.4. Corollary 108.37 reaches the same conclusion from the SI dimension of one coupling and no estimate at all. The two are the same statement: a coupling with a positive length dimension is one whose dimensionless strength grows as the probe scale grows, and the correlation length near T_c is the probe scale that diverges.

The shell integration itself supplies the corrections to Equation (108.41), and one loop is enough to see the whole structure.

Theorem 108.39 (The one-loop flow of the quartic coupling)**.** *For the $O(N)$-symmetric version of Equation (108.39), with ϕ an N-component field and the quartic term $(u/4!)(\phi\cdot\phi)^2$, integrating an infinitesimal shell $b = 1 + \delta l$ and rescaling gives, near four dimensions with $\varepsilon = 4 - d$,*

$$\frac{\mathrm{d}u}{\mathrm{d}l} = \varepsilon u - \frac{N+8}{48\pi^2}\,u^2 + O\!\left(u^3\right) , \tag{108.42}$$

$$\frac{\mathrm{d}r}{\mathrm{d}l} = 2r + \frac{N+2}{48\pi^2}\,u\Lambda^2 - \frac{N+2}{48\pi^2}\,ur + O\!\left(u^2\right) , \tag{108.43}$$

where $l = \ln b$ increases towards the infrared [1489].

Rests on Equations (108.38), (108.39) and (108.41).

Derivation. *Derives Theorem 108.39.*

Write $\phi = \phi_< + \phi_>$ in Equation (108.38) and expand the quartic term. The Gaussian average over the shell modes has propagator $G(\boldsymbol{k}) = 1/(k^2 + r)$, supported only on $\Lambda/b < k < \Lambda$, and the cumulant expansion gives

$$\frac{H_{\Lambda/b}}{k_B T} = \frac{H_\Lambda\left[\phi_<\right]}{k_B T} + \langle \Delta H \rangle_> - \frac{1}{2}\left(\left\langle \Delta H^2 \right\rangle_> - \langle \Delta H \rangle_>^2\right) + \cdots , \tag{108.44}$$

ΔH collecting the terms containing $\phi_>$.

First order. The piece of $(\phi\cdot\phi)^2$ with two $\phi_<$ and two $\phi_>$ is, after averaging and using $\langle \phi_{>a}\phi_{>b}\rangle = \delta_{ab}\int_{\text{shell}} \mathrm{d}^d k\, G(k)/(2\pi)^d$, proportional to $\phi_<\cdot\phi_<$ and so renormalizes r. The combinatorial factor for $O(N)$ is $(N+2)/3$ relative to $u/4!$ times the 6 ways of choosing which two of four fields are hard, giving

$$\delta r = \frac{N+2}{6}\,u\int_{\text{shell}} \frac{\mathrm{d}^d k}{(2\pi)^d}\,\frac{1}{k^2 + r} . \tag{108.45}$$

For $N = 1$ the factor is $1/2$, which is the elementary count: of the $\binom{4}{2} = 6$ pairings, all contribute, and $6/4! = 1/4$, doubled by matching to $\frac{1}{2}\delta r\phi^2$.

Second order. The connected part of $\langle \Delta H^2 \rangle$ with four $\phi_<$ and two internal shell propagators renormalizes u. The same combinatorics with the $O(N)$ index contractions

gives the factor $(N+8)/6$, and the sign is negative because the term enters Equation (108.44) with $-1/2$:

$$\delta u = -\frac{N+8}{6}\,u^2 \int_{\text{shell}} \frac{\mathrm{d}^d k}{(2\pi)^d}\,\frac{1}{(k^2+r)^2}\,. \tag{108.46}$$

The shell integrals. At $d=4$, with $S_d = 2\pi^{d/2}/\Gamma(d/2)$ so that $S_4 = 2\pi^2$,

$$\int_{\text{shell}} \frac{\mathrm{d}^4 k}{(2\pi)^4} f(k) = \frac{2\pi^2}{16\pi^4}\int_{\Lambda(1-\delta l)}^{\Lambda} k^3 f(k)\,\mathrm{d}k = \frac{\Lambda^4 f(\Lambda)}{8\pi^2}\,\delta l\,, \tag{108.47}$$

for an infinitesimal shell. Hence $\int_{\text{shell}} G = \Lambda^2 \delta l/8\pi^2$ at $r=0$, with the correction $-r/\Lambda^2$ from expanding $1/(\Lambda^2+r)$, and $\int_{\text{shell}} G^2 = \delta l/8\pi^2$.

Assembling. Adding Equations (108.45) and (108.46) to the tree-level rescalings Equation (108.41), with $b^2 = 1+2\delta l$ and $b^\varepsilon = 1+\varepsilon\delta l$,

$$\frac{\mathrm{d}u}{\mathrm{d}l} = \varepsilon u - \frac{N+8}{6}\cdot\frac{u^2}{8\pi^2}\,, \qquad \frac{\mathrm{d}r}{\mathrm{d}l} = 2r + \frac{N+2}{6}\cdot\frac{u\,(\Lambda^2 - r)}{8\pi^2}\,,$$

which are Equations (108.42) and (108.43) since $6\times 8\pi^2 = 48\pi^2$. What is carried out in full above is the single-component count; the $O(N)$ index contractions giving $(N+2)/6$ in Equation (108.45) and $(N+8)/6$ in Equation (108.46) are quoted from [1489], and at $N=1$ they return $\frac{1}{2}$ and $\frac{3}{2}$, the first of which is the elementary count carried out above. □

Derivation pending #176

The $O(N)$ combinatorial factors of the one-loop shell integration: the explicit index contractions of $(\phi\cdot\phi)^2$ that produce $(N+2)/6$ in the mass renormalization and $(N+8)/6$ in the vertex renormalization, for general N. The $N=1$ case is counted in the derivation above and both general factors reduce to it; the general count belongs in Appendix A.

Remark 108.40 (Sign conventions, and translating to the scale). $l = \ln b$ increases towards *longer* distances, so $\mathrm{d}/\mathrm{d}l = -\mathrm{d}/\mathrm{d}\ln\bar{\mu}$. Translating Equation (108.42) accordingly, and using the dimensionless $\hat{g}$ of Equation (106.34) in place of u, the field-theoretic beta function of a single real scalar ($N=1$, $\varepsilon=0$) is

$$\frac{\mathrm{d}\hat{g}}{\mathrm{d}\ln\bar{\mu}} = \frac{3\hat{g}^2}{16\pi^2} + O\big(\hat{g}^3\big)\,, \tag{108.48}$$

positive, so a scalar self-coupling grows with energy exactly as the electromagnetic coupling does. That is the sign that will produce Section 108.5.3. A reader comparing with the sources should note that many of them write $\mathrm{d}/\mathrm{d}l$ and many others $\mathrm{d}/\mathrm{d}\ln\bar{\mu}$, with no warning, so an apparent sign disagreement between two texts is usually this.

Remark 108.41 (Exact formulations, and the method as a computational tool). Equation (108.38) can be written as an exact differential equation for the cutoff dependence of the effective Hamiltonian rather than as a perturbative expansion. Polchinski's form [1743] follows the flow of the Wilsonian action with a smooth cutoff and is the version usually used to *prove* things, including renormalizability; Wetterich's form [1876] flows the effective action of Definition 106.41 instead, with an added infrared regulator, and is the version usually used to *compute* in strongly coupled regimes, since truncating it does not require a small coupling. Both are exact statements about a functional and both require truncation to be evaluated; neither is free of that approximation, and the systematic error of a truncation is the honest weak point of the method.

That the Wilsonian construction computes rather than merely interprets was established by Wilson's numerical renormalization group, which solved the Kondo problem — a magnetic impurity in a metal, whose perturbation series diverges logarithmically at low temperature — by discretizing the conduction band logarithmically and diagonalizing shell by shell [1567]. It produced the impurity susceptibility and the Wilson ratio to a few per cent, in a regime where no expansion in the coupling exists at all.

108.4.3 Fixed points, relevant and irrelevant operators

Definition 108.42 (Fixed point and linearized flow)**.** Let $\{K_\alpha\}$ be coordinates on the space of Hamiltonians and $\mathrm{d}K_\alpha/\mathrm{d}l = \mathcal{B}_\alpha(K)$ the flow generated by Definition 108.34. A *fixed point* $K^\star$ is a zero of every $\mathcal{B}_\alpha$. Writing $K = K^\star + \delta K$ and linearizing,

$$\frac{\mathrm{d}\delta K_\alpha}{\mathrm{d}l} = \sum_\beta M_{\alpha\beta}\,\delta K_\beta\,, \qquad M_{\alpha\beta} := \left.\frac{\partial \mathcal{B}_\alpha}{\partial K_\beta}\right|_{K^\star}. \tag{108.49}$$

Let y_i be the eigenvalues of M and g_i the corresponding eigen-coordinates, the *scaling fields*. Then $g_i(l) = \mathrm{e}^{y_i l} g_i(0)$, and g_i is called *relevant* if $y_i > 0$, *irrelevant* if $y_i < 0$ and *marginal* if $y_i = 0$.

Theorem 108.43 (Criticality needs one tuning per relevant field)**.** *A Hamiltonian flows into the fixed point under repeated coarse graining if and only if every relevant scaling field vanishes. Hence the set of critical Hamiltonians is a surface of codimension n_r, n_r being the number of positive eigenvalues, and reaching criticality requires tuning exactly n_r parameters.*

Rests on Definition 108.42 and Equation (108.49)*.*

Proof. *Derives Theorem 108.43.*

By Equation (108.49) an irrelevant field decays as $\mathrm{e}^{-|y|l}$ and is forgotten; a relevant field grows as e^{yl} and eventually carries the system away from the fixed point, however small it was to begin with. The initial Hamiltonian therefore approaches $K^\star$ if and only if its relevant components vanish, which is n_r conditions on the parameters. The critical surface is the stable manifold of $K^\star$. □

Corollary 108.44 (Universality)**.** *Two microscopic Hamiltonians on the same critical surface flow to the same fixed point and therefore have the same singular behaviour, with exponents given by Equation* (108.34) *in terms of the eigenvalues of M at that fixed point. The exponents are properties of the fixed point; the microscopic Hamiltonians differ only in irrelevant directions, which decay.*

Rests on Theorems 108.32 and 108.43 and Equation (108.34)*.*

Proof. *Derives Corollary 108.44.*

Immediate from Theorem 108.43 and from Theorem 108.32, whose assumed relations Equation (108.32) are now *derived*: t and h are the two relevant scaling fields, and y_t, y_h their eigenvalues, since a scale change by b multiplies g_i by b^{y_i}. □

Example 108.45 (The Gaussian fixed point in three dimensions)**.** The free theory $r = u = 0$ is a fixed point of Equations (108.42) and (108.43). Its eigenvalues are read off from Equation (108.41): $y_r = 2$, $y_u = 4 - d$, $y_6 = 6 - 2d$, and in general the operator ϕ^{2n} carries

$$y_{2n} = d - n\,(d-2)\,. \tag{108.50}$$

In the observed $d = 3$ this gives $y_2 = +2$ (relevant, the temperature), $y_4 = +1$ (relevant, the quartic coupling), $y_6 = 0$ (marginal), $y_8 = -1$, and everything beyond irrelevant. The Gaussian fixed point therefore has two relevant directions in $d = 3$ and only one for $d > 4$, which is Corollary 108.37 again: above four dimensions u is irrelevant, the Gaussian fixed

point is the critical one, and the mean-field exponents $\nu = 1/y_2 = 1/2$, $\eta = 0$ and $\alpha = 0$ are exact. The last of these is *not* $2-d\nu = 2-d/2$, which is what hyperscaling Equation (108.35) would give: hyperscaling is exactly what fails above the upper critical dimension, for the reason recorded in Remark 108.51, and applying it at $d = 5$ would return $\alpha = -1/2$ for a theory whose specific heat has a finite jump, that is $\alpha = 0$. For $d < 4$ the second relevant direction means the Gaussian fixed point is unstable and cannot govern criticality: something else must, and Section 108.4.4 finds it.

Derivation of the Gaussian eigenvalues. *Derives Example 108.45.*

By Proposition 108.36, $[\phi] = \mathrm{m}^{(2-d)/2}$, so the coefficient of ϕ^{2n} in Equation (108.39) carries $1/\mathrm{m}^d \times \mathrm{m}^{n(d-2)}$, and a coupling of dimension $1/\mathrm{m}^y$ has eigenvalue y. Hence $y_{2n} = d - n(d-2)$. □

Remark 108.46 (Why an exponent is an eigenvalue, and what that explains). Three facts that had no explanation before Wilson follow at once, and Fisher's review sets them out in the order in which they were understood [1526]. That only two exponents are independent is Theorem 108.43 with $n_r = 2$. That a fluid and a uniaxial magnet share exponents is Corollary 108.44, the two Hamiltonians lying on the same critical surface because they have the same space dimension and the same order-parameter symmetry and differ only in irrelevant couplings. And that the exponents are not simple rational numbers is because an eigenvalue of a linearized flow on an infinite-dimensional space has no reason to be: the mean-field values $\frac{1}{2}$, 1, 3 are rational precisely because they come from naive dimensional counting at a Gaussian fixed point, and the true values do not.

108.4.4 Universality

Theorem 108.47 (The Wilson–Fisher fixed point)**.** *For $0 < \varepsilon = 4 - d$ small, Equation (108.42) has, besides the Gaussian fixed point $u = 0$, a second zero at*

$$u^\star = \frac{48\pi^2}{N+8}\,\varepsilon + O\!\left(\varepsilon^2\right), \tag{108.51}$$

which is stable in the u direction. Linearizing Equation (108.43) about it gives the thermal eigenvalue

$$y_t = 2 - \frac{N+2}{N+8}\,\varepsilon + O\!\left(\varepsilon^2\right), \qquad \nu = \frac{1}{y_t} = \frac{1}{2} + \frac{N+2}{4\,(N+8)}\,\varepsilon + O\!\left(\varepsilon^2\right), \tag{108.52}$$

and $\eta = O(\varepsilon^2)$, there being no field renormalization at one loop [1489].

Rests on Equations (108.42) and (108.43).

Proof. *Derives Theorem 108.47.*

Setting $\mathrm{d}u/\mathrm{d}l = 0$ in Equation (108.42) gives $u = 0$ or $u = 48\pi^2\varepsilon/(N+8)$, which is Equation (108.51). Its stability follows from $\mathrm{d}(\mathrm{d}u/\mathrm{d}l)/\mathrm{d}u = \varepsilon - 2(N+8)u^\star/48\pi^2 = -\varepsilon < 0$: the quartic direction, marginal at $d = 4$ and relevant at the Gaussian fixed point for $d < 4$, has become irrelevant here with eigenvalue $-\varepsilon$.

For the thermal eigenvalue, the term in Equation (108.43) proportional to Λ^2 is a constant shift of r, which relocates the critical temperature and is not universal; absorbing it into r and linearizing the remainder,

$$\frac{\mathrm{d}\delta r}{\mathrm{d}l} = \left(2 - \frac{N+2}{48\pi^2}u^\star\right)\delta r = \left(2 - \frac{N+2}{48\pi^2}\cdot\frac{48\pi^2\varepsilon}{N+8}\right)\delta r\,,$$

which is Equation (108.52). Expanding $1/(2-x) = \frac{1}{2}(1 + x/2) + O(x^2)$ gives the stated ν.

That η vanishes at this order is because γ_ϕ requires a correction to the coefficient of $(\nabla\phi)^2$, and the one-loop diagrams computed above carry no external momentum dependence: the tadpole Equation (108.45) is momentum independent and the vertex correction Equation (108.46) contributes to u, not to the gradient term. The first contribution to η is the two-loop sunset graph, of order $u^2 \sim \varepsilon^2$. □

Remark 108.48 (The ε expansion is a device). Theorem 108.47 is an expansion in a parameter that Nature does not supply. Its purpose is to give a controlled approximation scheme in which the fixed point is close to the Gaussian one and therefore calculable; the physical case is $\varepsilon = 1$, where the expansion parameter is not small and the series is in fact asymptotic rather than convergent. Setting $\varepsilon = 1$ in Equation (108.52) gives, for the Ising class $N = 1$, $y_t = 5/3$ and $\nu = 0.600$ against the true 0.6300, and for the XY class $N = 2$, $y_t = 8/5$ and $\nu = 0.625$ against the true 0.6717: the leading term captures roughly two thirds of the departure from the mean-field $\frac{1}{2}$ and no more. This is honest arithmetic about a one-loop truncation and not a defect of the framework; resummation of the series to high order, or methods that do not use the continuation at all, do far better, as Table 108.1 shows.

Table 108.1: Critical exponents of the three-dimensional Ising universality class ($N = 1$, $d = 3$), by three routes. The first column is Equation (108.52) at $\varepsilon = 1$; the second is the resummed field-theoretic calculation directly in three dimensions [1600]; the third is the conformal-bootstrap determination [2239], from which the remaining exponents are computed here through Equation (108.34) with $y_t = 1/\nu$ and $y_h = (d + 2 - \eta)/2$. The last row is a check, not a result: Rushbrooke must give exactly 2. In the first and third columns γ is $\nu(2 - \eta)$; the middle column is instead reproduced entry by entry as published, so its γ is [1600]'s own quoted value and not the 1.2402 that $\nu(2 - \eta)$ would give from the ν and η of the same column. The difference sits inside the quoted uncertainties, and it is a reminder that a resummed calculation determines each exponent separately rather than deriving one from another.

	one loop, $\varepsilon = 1$	resummed	bootstrap
ν	0.600	0.6300(15)	0.629971(4)
η	0	0.0315(35)	0.0362978(20)
$\alpha = 2 - 3\nu$	0.200	0.110(5)	0.110087
β	—	0.3250(15)	0.326419
γ	1.200	1.2410(20)	1.237075
δ	—	—	4.78984
$\alpha + 2\beta + \gamma$	—	—	2.000000

Phenomenon 108.49 (A critical exponent, and a disagreement)**.** The specific heat of liquid helium-4 at the superfluid transition can be followed to within nanokelvin of T_c only in free fall, because on the ground the hydrostatic pressure head across a sample of usable height smears the transition temperature over more than the resolution sought. Measured aboard an orbiting laboratory over more than five decades in reduced temperature, the singular specific heat gives

$$\alpha = -0.0127 \pm 0.0003 \tag{108.53}$$

[2054]: small, and *negative*, so that the specific heat approaches a finite cusp rather than diverging. Through hyperscaling Equation (108.35) in $d = 3$ this is

$$\nu = \frac{2 - \alpha}{3} = 0.670900 \pm 0.000100\,. \tag{108.54}$$

The best theoretical determination of the same exponent for the three-dimensional XY universality class is $\nu = 0.67155 \pm 0.00027$ [2035], that is $\alpha = -0.01465 \pm 0.00081$. The two do not agree: the difference is 0.00195 against a combined uncertainty of 0.00086, about two and a quarter standard deviations, and later and more precise theoretical determinations have moved away from the measurement rather than towards it. This is the most precisely measured critical exponent in physics, it is a quantitative test of a calculated number, and the test is currently not passed. This is the same measurement recorded as Phenomenon 123.6, not a second one; the analysis of it — the universality class, the hyperscaling step and the disagreement — is carried out here.

Rests on Corollary 108.44, Equation (108.35), and Phenomenon 123.6.

Derivation of what is being compared. *Derives Phenomenon 108.49.*

Helium-4's order parameter is a complex condensate wavefunction — the object whose theory belongs to Chapter 130 — which has two real components and a U(1) symmetry, so by Corollary 108.44 the transition belongs to the $N = 2$ class in $d = 3$: the same class as a planar magnet, and a different one from the Ising class of Table 108.1. The theoretical input is therefore ν for $N = 2$, and the measured input is α; the two are compared through Equation (108.35), $\alpha = 2 - 3\nu$, which is a consequence of Theorem 108.32 and holds below the upper critical dimension. Propagating the uncertainty, $\delta\alpha = 3\delta\nu = 0.00081$, and the discrepancy quoted follows by subtraction. Note what is *not* in doubt: the sign of α, the smallness of its magnitude, and the fact that both theory and experiment place it within two per cent of zero while mean-field theory predicts $\alpha = 0$ exactly and the Ising class gives $+0.110$. The framework is confirmed; the fourth decimal place is not. □

Remark 108.50 (Where the discrepancy might live). Three candidates are on the table and none has been eliminated. The measurement may carry an unmodelled systematic, most plausibly from the finite size of the sample cell, which cuts off the divergence of ξ at a few centimetres and enters the fit as a correction whose form must be assumed. The theory may be misestimating its own truncation error, the resummations involved being asymptotic. Or the comparison may be incomplete: corrections to scaling enter the measured specific heat with an exponent $\omega \approx 0.79$ and an amplitude that is not universal, and the two analyses do not treat them identically. The honest summary is that a two-standard-deviation disagreement between the sharpest measurement and the sharpest calculation in critical phenomena has stood for over twenty years without being resolved or explained, and that it is recorded here rather than smoothed over.

Remark 108.51 (What universality does and does not claim). It claims that the *exponents* and certain amplitude *ratios* depend only on the space dimension, the number of order-parameter components and the range of the interaction. It does not claim that T_c is universal — that is Equation (108.43)'s non-universal Λ^2 term, which depends on every microscopic detail — nor that amplitudes themselves are. And it fails where its premises do: for long-range interactions, for which the gradient term of Equation (108.39) is not the leading one; at a transition governed by a different fixed point, such as the Berezinskii–Kosterlitz–Thouless transition, which belongs to Section 123.6.2, where the singularity is essential rather than a power and the exponent language does not apply at all; and above the upper critical dimension, where hyperscaling Equation (108.35) breaks because the dangerous irrelevant variable u cannot be set to zero in the free energy.

108.5 Running couplings in the Standard Model

The preceding sections built an apparatus. This one is why the apparatus is believed: the flow it predicts is measured, over more than two decades of energy, in half a dozen unrelated processes.

108.5.1 Asymptotic freedom

The derivation of the negative beta function of $\mathrm{SU}(3)_c$ is Theorem 102.35, carried out in Chapter 102 by computing the magnetic permeability of the gauge vacuum, and it is not repeated here. What belongs here is what the result means for the flow, why the case is special, and what it generates.

Proposition 108.52 (Asymptotic freedom as a property of the flow)**.** *Let $\beta(a) = -\beta_0 a^2 + O(a^3)$ as in Equation (108.24). If $\beta_0 > 0$ then Equation (108.10) gives*

$$\frac{1}{a(\bar{\mu})} = \frac{1}{a(\kappa)} + \beta_0 \ln \frac{\bar{\mu}}{\kappa}\,, \tag{108.55}$$

so that $a(\bar{\mu}) \to 0$ as $\bar{\mu} \to \infty$, logarithmically, and the ultraviolet limit of the theory is the free theory: the Gaussian fixed point is reached from above, in infinite renormalization-group time. If $\beta_0 < 0$ the same formula sends a to infinity at a finite scale

$$\bar{\mu}_{\mathrm{L}} = \kappa \exp\!\left[\frac{1}{|\beta_0|\, a(\kappa)}\right], \tag{108.56}$$

and the perturbative description destroys itself there.

Rests on Equation (108.24) and Corollary 108.8.

Proof. *Derives Proposition 108.52.*

Separate variables in $\mathrm{d}a/\mathrm{d}\ln\bar{\mu} = -\beta_0 a^2$: $-\mathrm{d}(1/a) = -\beta_0 \mathrm{d}\ln\bar{\mu}$, which integrates to Equation (108.55). For $\beta_0 > 0$ the right-hand side grows without bound, so $a \to 0$; for $\beta_0 < 0$ it reaches zero at the scale Equation (108.56). The logarithmic approach is what makes the ultraviolet limit *asymptotically* free rather than free: at any finite energy the coupling is non-zero, and the corrections it produces fall only as $1/\ln$. □

Corollary 108.53 (The condition on the number of flavours)**.** *For $\mathrm{SU}(3)_c$, $\beta_0 = b_0/2\pi$ with $b_0 = 11 - \frac{2}{3}n_f$ by Equation (102.42), so asymptotic freedom requires $n_f < 33/2$, that is $n_f \le 16$. Nature supplies six. The eleven is $12 - 1$, the excess of the gluon's paramagnetic moment over its diamagnetic orbital response (Remark 102.36); the $-\frac{2}{3}n_f$ is the ordinary screening of a fermion loop, the same sign as in Proposition 100.63 and reproduced from the same master formula by Corollary 102.34.*

Rests on Proposition 108.52 and Equations (102.42) and (108.24).

Proof. *Derives Corollary 108.53.*

Comparing $\mathrm{d}\alpha_s/\mathrm{d}\ln E = -(b_0/2\pi)\alpha_s^2$ of Equation (102.42) with $-\beta_0\alpha_s^2$ of Equation (108.24) gives $\beta_0 = b_0/2\pi$, so $\beta_0 > 0$ is $b_0 > 0$, that is $11 > \frac{2}{3}n_f$ and $n_f < 33/2$. Since n_f is an integer this is $n_f \le 16$. □

Remark 108.54 (Non-abelian gauge theory is the only case). Among renormalizable theories in the observed $3+1$ dimensions — that is, among theories whose vertices all have $D_i = 4$ in the sense of Corollary 108.27 — a negative one-loop beta function requires a non-abelian gauge field. Scalar self-couplings contribute with the sign of Equation (108.48), Yukawa couplings and abelian gauge couplings with the sign of Equation (108.22), and only the gluon self-interaction contributes the other way, by Lemma 102.31. Adding matter can only reduce b_0, never increase it, which is Corollary 108.53. The theorem to this effect was proved by Coleman and Gross in the same year as [1501, 1511]; the treatise's bibliography does not carry that paper, and the statement is recorded here without a citation rather than attached to a source that does not contain it.

Remark 108.55 (Dimensional transmutation, and what it costs). Equation (108.55) can be read backwards. The pair $(a(\kappa), \kappa)$ — a pure number and an energy — can be traded for the single energy at which the right-hand side would vanish, which is Equation (102.45):

$$\Lambda_{\mathrm{QCD}} = E_0 \exp\!\left[-\frac{2\pi}{b_0 \alpha_s(E_0)}\right] ,$$

and Λ_{QCD} does not depend on which E_0 was used, as Definition 102.38 verifies by differentiation. A theory whose Lagrangian contains no dimensionful parameter at all has acquired one — Phenomenon 102.39 — and it is of the observed size, a few hundred MeV, corresponding to a hadronic length of about $1 \times 10^{-15}\,\mathrm{m}$.

What this costs is worth stating plainly, because it is easy to present dimensional transmutation as though a scale had been created from nothing. It has not. The number of free parameters is unchanged: one dimensionless coupling quoted at one energy has been re-expressed as one energy. What is genuinely explained is why that energy is far below any scale appearing in the Lagrangian, since the exponential in Equation (102.45) converts a modest coupling into an enormous ratio, and hence why 99 % of the proton's mass has no counterpart among the quark masses (Remark 102.40).

108.5.2 The measured running

Phenomenon 108.56 (The strong coupling runs as predicted)**.** The dimensionless coupling α_s of Equation (102.19), extracted independently from hadronic τ decays at $E = 1.777\,\mathrm{GeV}$, from the scaling violations of deep inelastic structure functions, from event shapes and jet rates in electron–positron annihilation, from heavy-quarkonium level spacings, from short-distance lattice quantities, and from the hadronic width of the Z at $E = (91.1880 \pm 0.0020)\,\mathrm{GeV}$, falls with energy along one curve whose shape contains no adjustable parameter [2142, 2427]. The curve is Equation (108.55) with β_0 fixed by Equation (102.42) from the gauge group and the particle content alone. The world average is

$$\alpha_s\!\left(m_Z c^2\right) = 0.1180 \pm 0.0009 \tag{108.57}$$

[2427], the 2009 compilation having given 0.1184 ± 0.0007 [2142]. The two determinations whose scales are quoted above differ by a factor of 51; the full set compiled in [2427] spans more than two decades in energy, which is the span Phenomenon 102.41 also reports. The agreement of all of them, after evolution by a formula with no free coefficient, is the measurement.

Rests on Equations (102.19) *and* (102.42) *and Proposition 108.52.*

Derivation of the predicted curve. *Derives Phenomenon 108.56.*

Integrate Equation (108.24) keeping two terms,

$$\frac{\mathrm{d}\alpha_s}{\mathrm{d}\ln E} = -\frac{b_0}{2\pi}\alpha_s^2 - \frac{b_1}{8\pi^2}\alpha_s^3 + O\!\left(\alpha_s^4\right) , \qquad b_0 = 11 - \frac{2}{3}n_f , \quad b_1 = 102 - \frac{38}{3}n_f , \tag{108.58}$$

the first coefficient being Equation (102.42) and the second the two-loop coefficient, quoted in the Particle Data Group's review of quantum chromodynamics [2427]. Both are scheme independent by Theorem 108.19, which is what makes the comparison below a test rather than a fit; the third coefficient is not, and is not needed at this accuracy.

Both b_0 and b_1 depend on the number of quark flavours light enough to circulate in a loop at the scale considered. The theory is therefore integrated piecewise: between thresholds n_f is constant, and at $E = m_Q c^2$ the coupling is matched continuously as the flavour is added or removed, the discontinuity in the matching being of order α_s^2 and negligible here. The slope of the reciprocal coupling in the one-loop approximation is $\mathrm{d}\,(1/\alpha_s)\,/\mathrm{d}\ln E = b_0/2\pi$, which takes the values 1.4324, 1.3263, 1.2202 and 1.1141 for $n_f = 3, 4, 5, 6$: the curve of

$1/\alpha_s$ against $\ln E$ is a polygon whose slope decreases by about 0.106 at each threshold, and the kinks are a prediction as sharp as the slopes.

Starting from Equation (108.57) at $E = m_Z c^2 = 91.1880\,\text{GeV}$ with $n_f = 5$, and integrating Equation (108.58) numerically with thresholds at the quark masses of Table 102.1, gives Table 108.2. Of the two coefficients, b_0 is derived in Chapter 102; b_1 is quoted, and every entry of the two-loop column depends on it. □

Derivation pending #177

The two-loop coefficient $b_1 = 102 - \frac{38}{3}n_f$ of the SU(3) beta function: the two-loop vacuum-polarization, vertex and self-energy graphs, their renormalization, and the cancellation that leaves a scheme-independent result. It is quoted from the Particle Data Group here and it fixes the whole two-loop column of the table above, so it is a number the book uses and does not yet derive; the derivation belongs in Appendix A.

Table 108.2: The strong coupling evolved from Equation (108.57) by Equation (108.58), with n_f changing at the quark masses of Table 102.1. The one-loop column reproduces Table 102.2, whose caption records honestly that the truncation drifts about ten per cent low at the τ mass. The second column shows where that drift goes: adding the single scheme-independent coefficient b_1 moves the prediction at $E = m_\tau c^2$ from 0.279 to 0.313, against a measured value of about 0.31 [2427]. The last column is the distance $\hbar c/E$ that energy resolves.

E	n_f	one loop	two loop	$\hbar c/E$
1.777 GeV $(m_\tau c^2)$	4	0.279	0.313	1.11×10^{-16} m
4.183 GeV $(m_b c^2)$	5	0.212	0.224	4.72×10^{-17} m
10 GeV	5	0.173	0.178	1.97×10^{-17} m
91.188 GeV $(m_Z c^2)$	5	0.1180	0.1180	2.16×10^{-18} m
1 TeV	6	0.089	0.088	1.97×10^{-19} m

Remark 108.57 (What the two-loop improvement demonstrates). The one-loop curve of Table 102.2 misses the τ-mass point by about ten per cent, and a reader is entitled to ask whether that is a failure of the theory or of the truncation. Table 108.2 answers it without any new input: the next term in a series whose coefficient is fixed by Equation (108.58) and is not adjustable moves the prediction by almost exactly the discrepancy, in the right direction, at the scale where the expansion parameter is largest and nowhere else. The correction at $m_Z c^2$ is nil by construction and at 1 TeV is one part in a hundred. That pattern — a truncation error that grows as the coupling grows and vanishes where it is small — is what an asymptotic series behaving properly looks like, and it is a stronger statement than agreement at any single point.

Remark 108.58 (The electromagnetic coupling, for contrast). The same apparatus applied to Equation (108.22) runs α the other way: from $\alpha^{-1} = 137.035999177(21)$ at zero momentum transfer [2431] to $\alpha^{-1} \approx 128.94$ at $E = m_Z c^2$ [2427], a rise of about six per cent in the coupling, of which roughly half is leptonic and computable from Equation (100.83) and roughly half hadronic and extracted from measured annihilation cross-sections through a dispersion relation. The derivation is Phenomenon 100.62. Two couplings of the same theory therefore run in opposite directions, by the same mechanism and with the same machinery, and both runs are measured. If the renormalization group were bookkeeping, that could not happen. The measurement at zero momentum transfer belongs to Chapter 112; the scaling

violations that measure α_s by a slope rather than a rate belong to Section 102.5.2 and to the experiment chapter Chapter 116.

108.5.3 Landau poles and triviality

A positive β_0 in Equation (108.24) is a promise about the ultraviolet; a negative one is a warning. This subsection states what the warning does and does not establish, for the two Standard Model sectors that carry it.

Proposition 108.59 (The hypercharge coupling and its pole)**.** *For the hypercharge group* $\mathrm{U}(1)_Y$ *with the Standard Model content of Table 104.1 — three generations of chiral fermions and one complex scalar doublet — the one-loop beta function of* $\alpha_Y := \hat{g}'^2/4\pi$ *is*

$$\frac{\mathrm{d}\alpha_Y}{\mathrm{d}\ln E} = \frac{b_Y}{2\pi}\alpha_Y^2\,, \qquad b_Y = \frac{41}{6}\,, \tag{108.59}$$

positive. Taking both inputs in the same scheme, as Corollary 108.20 requires — the effective on-shell coupling $\alpha^{-1}(m_Z c^2) = 128.94$ *of Remark 108.58 together with the on-shell mixing angle* $s_W^2 = 0.2232$ *of Remark 104.34, both from [2427] — gives* $\alpha_Y^{-1}(m_Z c^2) = 100.16$*, and the one-loop coupling diverges at*

$$\ln\frac{E_{\mathrm{L}}}{m_Z c^2} = \frac{2\pi}{b_Y}\cdot 100.16 = 92.10\,, \qquad E_{\mathrm{L}} \approx 9\times 10^{41}\ \mathrm{GeV}\,. \tag{108.60}$$

Rests on Corollary 108.20 and Equations (102.35) and (104.13).

Derivation. Derives Proposition 108.59.

The master formula Equation (102.35) of Chapter 102,

$$\frac{\mathrm{d}\alpha}{\mathrm{d}\ln E} = \frac{S}{12\pi}\alpha^2\,, \qquad S = \sum_i x_i^2\,(-1)^{2s_i}\sum_{s_z}\left(1-12s_z^2\right)\,,$$

sums over one-particle states i of charge x_i in units of the coupling, and is the general one-loop vacuum-polarization result, not a QCD-specific one. In the convention $\mathrm{d}\alpha/\mathrm{d}\ln E = (b/2\pi)\alpha^2$ used throughout this chapter, $b = S/6$. Evaluate the spin sum for each kind of state.

- A *Dirac* fermion of charge x: two one-particle states, the fermion and the antifermion, each with $s_z = \pm\frac{1}{2}$, so $\sum_{s_z}(1-12s_z^2) = 2(1-3) = -4$ each; the statistics factor is $(-1)^1 = -1$; hence $S = x^2(-1)(-8) = 8x^2$ and $b = \frac{4}{3}x^2$, which is Proposition 100.63 and Corollary 102.34.
- A *Weyl* fermion of charge x: one helicity per particle, so half the Dirac value, $S = 4x^2$ and $b = \frac{2}{3}x^2$.
- A *complex scalar* of charge x: two states, each with $s_z = 0$ and $\sum_{s_z}(1-0) = 1$; the statistics factor is $(-1)^0 = +1$; hence $S = 2x^2$ and $b = \frac{1}{3}x^2$.

The hypercharge coupling appears in the covariant derivative as $\hat{g}'Y/2$ by the assignment $Q = T^3 + Y/2$ of Equation (104.13), so $x = Y/2$ for each field. Summing $(Y/2)^2$ over one generation of Table 104.1, with a factor $N_c = 3$ for coloured fields and a factor 2 for each $\mathrm{SU}(2)_L$ doublet,

$$2\cdot 3\cdot\left(\tfrac{1}{6}\right)^2 + 3\cdot\left(\tfrac{2}{3}\right)^2 + 3\cdot\left(\tfrac{1}{3}\right)^2 + 2\cdot\left(\tfrac{1}{2}\right)^2 + 1\cdot 1^2 = \frac{1}{6}+\frac{4}{3}+\frac{1}{3}+\frac{1}{2}+1 = \frac{10}{3}\,,$$

the five terms being the quark doublet, the up singlet, the down singlet, the lepton doublet and the charged-lepton singlet. Three generations give 10, and the Weyl coefficient $\frac{2}{3}$ gives

$b_Y^{\text{fermion}} = \frac{20}{3}$. The scalar doublet has two complex components with $Y/2 = \frac{1}{2}$, contributing $\frac{1}{3} \cdot 2 \cdot \frac{1}{4} = \frac{1}{6}$. Adding, $b_Y = \frac{20}{3} + \frac{1}{6} = \frac{41}{6}$, which is Equation (108.59).

For the numbers: $\alpha_Y = \alpha/\cos^2\theta_W$, so $\alpha_Y^{-1} = 128.94 \times (1 - 0.2232) = 100.16$. Integrating Equation (108.59) as in Equation (108.55) with the sign reversed, α_Y^{-1} falls linearly in $\ln E$ with slope $b_Y/2\pi = 1.0876$ and reaches zero after $100.16/1.0876 = 92.10$ e-folds, that is at $\log_{10}(E/\text{GeV}) = 1.96 + 92.10/\ln 10 = 41.96$.

Both inputs are on-shell quantities, which is why they may be combined. Doing the same arithmetic consistently in $\overline{\text{MS}}$ instead — $\hat{s}_Z^2 = 0.23129$ of Remark 104.34 with the $\overline{\text{MS}}$ coupling $\hat{\alpha}^{-1}(m_Z c^2) = 127.95$ [2427] — gives $\alpha_Y^{-1} = 98.36$, 90.44 e-folds and $E_{\text{L}} \approx 2 \times 10^{41}\,\text{GeV}$. The two schemes move the pole by a factor of about five, which is nothing beside the forty decades of extrapolation and is exactly why Remark 108.60 follows. What is not permitted is to combine 128.94 with $\hat{s}_Z^2$: that pairs a coupling from one scheme with an angle from another and produces a number belonging to neither. □

Remark 108.60 (What a Landau pole does not establish). The four objections stated in Remark 100.65 for the electromagnetic case apply verbatim here and are not repeated; the essential one is that Equation (108.60) extrapolates a one-loop formula over forty decades in energy, into a regime where the expansion parameter it is built from is of order one. The location of the pole is not a prediction. Two things are worth adding.

First, the hypercharge pole sits at about $9 \times 10^{41}\,\text{GeV}$, which is some 7×10^{22} times the Planck energy $\sqrt{\hbar c^5/G} = 1.22 \times 10^{19}\,\text{GeV}$; the QED pole of Equation (100.85) sits far higher still. In neither case is the question reachable, and any statement about it is a statement about an extrapolation and not about an experiment.

Second, and this is the point that makes the discussion worth having at all: the *existence* of a scale at which the description fails, rather than its value, is a structural fact. A theory with a positive β_0 carries an ultraviolet expiry date in exactly the way Fermi's theory carried one through Equation (101.9) — and in that case the date was 292.8 GeV, the theory did fail there, and what replaced it was found.

Remark 108.61 (Triviality: what is proved and what is not). The perturbative pole is not the strongest available statement. For the scalar theory of Equation (108.39) there are rigorous non-perturbative results, and they say something sharper.

Above four dimensions, the continuum limit of the lattice ϕ^4 theory — and of the Ising model, which the cited work treats in the same framework and for the same reason — is a *free* field theory: the renormalized quartic coupling vanishes when the cutoff is removed [1662]. This is the rigorous counterpart of Corollary 108.37: above the upper critical dimension the quartic coupling is irrelevant, so it flows to zero, and nothing survives the limit but the Gaussian fixed point. Exactly at four dimensions the coupling is marginal and the question is delicate; the answer, obtained only in 2021, is that the scaling limits of the critical four-dimensional Ising and $\lambda\phi^4$ models are Gaussian as well, marginally so [2371]. There is therefore no non-trivial continuum ϕ^4 theory in the observed number of dimensions.

For abelian gauge theory the corresponding statement is not a theorem but a lattice result: numerical study of lattice quantum electrodynamics with charged matter finds the continuum limit to exist only at zero renormalized coupling [1955], consistent with triviality and not proving it. The strength of that conclusion depends on the lattice action used — compact and non-compact U(1) have different phase structures at strong coupling — so what is recorded here is the conclusion the cited work draws and not a general statement about lattice U(1).

Corollary 108.62 (Two Standard Model sectors are effective). *The scalar self-coupling runs with the sign of Equation* (108.48) *and the hypercharge coupling with that of Equation* (108.59)*: both grow with energy, and both belong to sectors for which the evidence points to triviality in the strict continuum limit. The Standard Model is therefore not a candidate*

for a complete theory valid to arbitrarily high energy, independently of any consideration of gravity. It is a theory with a cutoff, and what the cutoff is remains unmeasured.

Rests on Equations (108.48) *and* (108.59) *and Proposition 108.52.*

Proof. *Derives Corollary 108.62.*

The two signs are Equation (108.48) and Equation (108.59), both derived above; by Proposition 108.52 a coupling with $\beta_0 < 0$ in the convention of Equation (108.24) reaches infinity at a finite scale, so the perturbative description of each sector has an ultraviolet endpoint. That the endpoint is not an artefact of perturbation theory in the scalar case is Remark 108.61. The conclusion follows without invoking gravity because no gravitational quantity has entered any step. □

Remark 108.63 (The same conclusion from the other side). Corollary 108.62 is an ultraviolet statement about a coupling that grows. Section 104.8.1 reaches the same place from the opposite direction, and it does so at a far lower energy. The quartic coupling's beta function Equation (104.72) is dominated by the top Yukawa and is *negative* at the top mass, $\mathrm{d}\hat{\lambda}/\mathrm{d}\ln\mu = -0.0206$ by Proposition 104.66; integrating the coupled system sends $\hat{\lambda}$ through zero at about 10^{10}–10^{11} GeV [2193, 2215], where the potential turns over.

The growth that Equation (108.48) describes for a self-coupled scalar therefore never happens in the Standard Model, because the fermion loop wins first. Asking where it *would* have arrived requires care with the normalization, because Equation (108.48) is written for a single real field with the interaction $-(g/4!)\,\phi^4$, while $\hat{\lambda} = 0.126$ is the quartic coupling of a complex doublet in the normalization of Equation (104.27), where the interaction is $-\lambda\left(\Phi^\dagger\Phi\right)^2$. The beta function belonging to the second is the leading term of Equation (104.72), $16\pi^2\mathrm{d}\hat{\lambda}/\mathrm{d}\ln\mu = 24\hat{\lambda}^2$, and not the $3\hat{g}^2$ of the first: the two normalizations differ by a factor of eight in the coefficient, and the eight sits in an exponent. Dropping every other term of Equation (104.72) and integrating that one from the top mass,

$$\ln\frac{E_{\mathrm{L}}}{m_t c^2} = \frac{16\pi^2}{24\hat{\lambda}} = \frac{157.91}{3.024} = 52.2\,, \qquad E_{\mathrm{L}} \approx 8\times10^{24}\,\mathrm{GeV}\,, \tag{108.61}$$

with $m_t c^2 = 172.57\,\mathrm{GeV}$. Feeding $\hat{\lambda}$ into the single-real-scalar coefficient instead would give 417.8 e-folds and $10^{184}\,\mathrm{GeV}$, roughly a hundred and sixty decades higher: that is what a normalization mismatch costs in an exponentiated quantity, and it is why the normalization is stated rather than assumed. Either number is far beyond reach and neither is a prediction. The honest composite statement is that the scalar sector's extrapolation fails at $10^{10}\,\mathrm{GeV}$ for a reason having nothing to do with triviality, and would fail eventually for triviality if it survived; Remark 104.67 states how fragile the first of those numbers is, and none of it is evidence for any particular completion.

108.6 Effective field theory

Corollary 108.62 left the Standard Model as a description with a cutoff and no knowledge of what lies above it. That situation is not a predicament; it is the normal situation of every successful theory in the history of physics, and the renormalization group turns it into a method. The method has three parts: a statement that heavy things drop out of low-energy physics in a controlled way (Section 108.6.1); a recipe for writing down what is left (Section 108.6.2); and, applied to the Standard Model itself, a set of operators whose coefficients are measured or bounded (Section 108.6.3).

108.6.1 Decoupling

Theorem 108.64 (Appelquist–Carazzone)**.** *Let a renormalizable theory contain a field of mass M and let observables be computed at energies $E \ll Mc^2$, using a mass-dependent*

scheme in the sense of Definition 108.17. Then the effects of the heavy field are of two kinds and no others: a finite renormalization of the parameters of the light theory, and corrections suppressed by positive powers of E/Mc^2 [1538].

Rests on Definition 108.17, Proposition 100.33, and Equation (100.82).

Derivation of the one-loop case. Derives Theorem 108.64.

Take the heavy field to be a Dirac fermion of mass m and unit charge in quantum electrodynamics, and take the momentum-subtraction scheme in which the coupling at spacelike momentum transfer Q is $\alpha(Q)$ of Equation (100.82). The subtracted polarization in full is the Feynman-parameter integral obtained in the derivation of Proposition 100.33 — of which Equation (100.53) is only the leading term for $|q^2| \ll m^2c^2$, and Equation (100.83) only the leading term in the opposite limit. For $q^2 = -Q^2$ it reads

$$\bar{\Pi}\left(Q^2\right) = \frac{2\alpha}{\pi} \int_0^1 \mathrm{d}x \, x\,(1-x) \ln\left[1 + \frac{x\,(1-x)\,Q^2}{m^2c^2}\right], \tag{108.62}$$

an expression in which Q and mc are both momenta and the argument of the logarithm is a pure number, as it must be. The beta function of this scheme is obtained by differentiating $\alpha(Q) = \alpha/[1-\bar{\Pi}]$ with respect to $\ln Q$, which at leading order is $\alpha^2\,\mathrm{d}\bar{\Pi}/\mathrm{d}\ln Q$:

$$\frac{\mathrm{d}\alpha}{\mathrm{d}\ln Q} = \frac{2\alpha^2}{3\pi} f\left(\frac{Q}{mc}\right), \qquad f(\rho) := 6\int_0^1 \mathrm{d}x \, \frac{x^2\,(1-x)^2\,\rho^2}{1 + x\,(1-x)\,\rho^2}, \tag{108.63}$$

the factor 6 chosen so that $f(\infty) = 1$. Two limits settle the theorem.

High energy. As $\rho \to \infty$ the integrand tends to $x(1-x)$, and $\int_0^1 x(1-x)\mathrm{d}x = \frac{1}{6}$, so $f \to 1$ and Equation (108.63) reduces to Equation (100.84): the fermion contributes in full, as a massless particle would.

Low energy. As $\rho \to 0$ the denominator tends to 1 and

$$f(\rho) \to 6\rho^2 \int_0^1 x^2\,(1-x)^2\,\mathrm{d}x = \frac{6\rho^2}{30} = \frac{\rho^2}{5} = \frac{1}{5}\left(\frac{Q}{mc}\right)^2,$$

using $\int_0^1 x^2(1-x)^2\mathrm{d}x = B(3,3) = 2!\,2!/5! = 1/30$. The heavy fermion's contribution to the running is therefore suppressed by $(Q/mc)^2 = (E/mc^2)^2$: it does not merely become small, it becomes small at a computable rate, quadratically in the ratio of the probe energy to the heavy mass. Numerically $f = 0.945$ at $Q = 10mc$, $f = 0.165$ at $Q = mc$ and $f = 2.0 \times 10^{-3}$ at $Q = mc/10$.

The first kind of effect — the finite renormalization — is the constant that $\bar{\Pi}$ would have contributed had the subtraction been made elsewhere, and it is by construction absorbed into the definition of α at the low scale. This is the whole content of the theorem at one loop, in one theory, for one kind of heavy field. The statement of Theorem 108.64 is broader on all three counts, and the general proof is not given here. □

Derivation pending #178

Appelquist–Carazzone to all orders: the proof that in a mass-dependent scheme the effects of a heavy field on light-field observables are exhausted, at every order in the loop expansion and for a heavy field of any spin whose mass is an independent parameter, by a finite renormalization of the light theory's parameters plus corrections of positive power in the ratio of the probe energy to the heavy mass. What is derived in the text is the one-loop case for a heavy Dirac fermion in quantum electrodynamics, including the quadratic suppression rate; the general theorem belongs in Appendix A.

Remark 108.65 (The theorem fails in minimal subtraction). Theorem 108.64 presupposes a mass-dependent scheme. In minimal subtraction the beta function contains no mass at all (Theorem 108.18), so a heavy quark keeps contributing its full $-\frac{2}{3}$ to b_0 at every scale, including scales far below its own mass, and α_s computed that way at $E = 1\,\mathrm{GeV}$ would carry the top quark. That is not a defect of the scheme but the price of its convenience, and the remedy is a construction rather than a correction:

- Below the threshold one uses a *different theory*, the effective theory with $n_f - 1$ flavours, in which the heavy quark does not appear as a field at all.
- At a scale near the threshold, conventionally $\bar{\mu}c = m_Q c^2$, one *matches*: the two theories are required to give the same value for a set of low-energy observables, which fixes the couplings of the light theory in terms of those of the heavy one. At one loop the matching is continuity, $\alpha_s^{(n_f-1)} = \alpha_s^{(n_f)}$; the difference first appears at order α_s^2.
- Above and below, one runs with the appropriate b_0.

The observable consequence is the kink already used in the derivation of Table 108.2: the slope $\mathrm{d}(1/\alpha_s)/\mathrm{d}\ln E$ changes from 1.2202 to 1.3263 on crossing the b threshold downwards, a discontinuity of about 0.106 in a quantity that is measured. Decoupling in $\overline{\mathrm{MS}}$ is therefore not automatic; it is imposed, and the imposing is matching.

Remark 108.66 (Where decoupling genuinely fails). Two failures matter physically, and both are visible in the electroweak theory. Neither contradicts Theorem 108.64: each violates one of its hypotheses, and saying which is the useful part.

A mass that is not an independent parameter. The theorem takes $M \to \infty$ with the couplings of the theory fixed. In the Standard Model a fermion's mass is $m_f c^2 = \hat{y}_f E_v/\sqrt{2}$ by Section 104.4.3, so making the top heavy means making $\hat{y}_t$ large, and the limit is not the one the theorem contemplates. The consequence is measured: the correction $\Delta\rho$ of Equation (104.52) grows as m_t^2 rather than falling, so the top does *not* decouple from the electroweak fit — which is exactly why that fit located the top mass before the top was produced. A non-decoupling effect is not a pathology; here it was the measurement.

A relevant operator. Equation (108.6) says that an operator of dimension $D < 4$ has a dimensionless strength growing as $(E_\star/E)^{4-D}$ when the probe energy falls. The scalar mass term is such an operator, $D = 2$, and a heavy field of mass M coupled to Φ shifts $m_h^2 c^4$ by an amount of order $(Mc^2)^2$ times a loop factor: the heavy state's effect on the one dimensionful parameter of the light theory grows with its mass instead of decaying. The same statement applies to the vacuum energy, $D = 0$, more strongly still. What follows from this and what does not is set out in Section 104.8.2, and the honest summary is repeated here because it is a renormalization-group statement: it is an argument about the form a more fundamental theory would have to take, it is not an observation, and none of the programmes proposed to address it has observational support of the kind this book requires.

Chiral symmetry breaking is sometimes given as a third example. It is better described as the failure of a different assumption: the light theory below Λ_{QCD} is not the Standard Model Lagrangian with heavy fields removed but a theory of different degrees of freedom altogether — pions rather than quarks — and matching onto it is Section 108.6.2, not Theorem 108.64.

108.6.2 Phenomenological Lagrangians and power counting

Theorem 108.67 (Weinberg's principle). *Suppose one writes down the most general possible Lagrangian containing all terms consistent with the assumed symmetry principles and built from the fields that are light at the energy of interest, and computes with it, ordering the terms by their dimension D in the sense of Equation (108.3). Then the result is the most*

general S-matrix consistent with analyticity, perturbative unitarity, cluster decomposition and the assumed symmetries: it is not an approximation to a theory but a theory, and it is predictive to any stated accuracy with finitely many parameters [1647, 1867].

Rests on Equations (108.3) *and* (108.6) *and Theorem 108.26.*

Derivation of the predictivity claim. *Derives Theorem 108.67.*

What has to be shown is that a stated accuracy needs only finitely many operators. By Equation (108.6) an operator of dimension D contributes to a process of energy E relative to the leading, $D = 4$, term by a factor

$$\left(\frac{E}{E_\star}\right)^{D-4}, \tag{108.64}$$

$E_\star$ being the energy at which the description fails. Demanding a fractional accuracy δ therefore requires only those operators with

$$D - 4 \leq \frac{\ln \delta}{\ln (E/E_\star)}, \tag{108.65}$$

a finite number for any $\delta > 0$ and any $E < E_\star$, since the number of local operators of bounded dimension built from finitely many fields is finite. Each carries one unknown coefficient, so the theory has finitely many parameters at each order; once they are measured, everything else at that order is predicted.

The step that makes this consistent rather than merely convenient is that loops do not spoil it. A loop of the light fields produces a divergence which, being local (Remark 100.53), is a local operator; Theorem 108.26 fixes its dimension, and the counting shows it is always an operator of dimension already included at that order. The divergence is therefore absorbed by a coefficient already present, and the theory is renormalizable *order by order in the expansion* even though it is non-renormalizable in the sense of Corollary 108.27. That is the whole reversal recorded in Remark 108.28.

What is established here is predictivity: that a stated accuracy needs only finitely many operators, and that loops do not spoil the counting. The other half of Theorem 108.67 — that what results is the *most general* S-matrix consistent with analyticity, perturbative unitarity, cluster decomposition and the assumed symmetries — is a separate statement, and it is not derived here. □

Derivation pending #179

Weinberg's generality claim: the demonstration that a Lagrangian containing every term allowed by the assumed symmetries and built from the light fields generates the most general S-matrix consistent with analyticity, perturbative unitarity and cluster decomposition, that is, that no further S-matrix satisfying those constraints exists which the construction fails to reach. The predictivity half of the theorem is derived in the text above; the generality half belongs in Appendix A.

Remark 108.68 (The two scales, and the loop counting). An effective theory has two expansion parameters, not one. The first is $E/E_\star$, the operator expansion Equation (108.64). The second is the loop expansion, and for a theory whose interactions are generated by a single scale the natural loop factor is $(E/4\pi E_\star)^2$ — the 4π being the $1/(4\pi)^2$ that Lemma 100.28 attaches to every four-dimensional loop. The two expansions are therefore not independent: one loop costs as much as two extra dimensions of operator, up to $16\pi^2$, and a consistent calculation at a given order in $E/E_\star$ must include both. That correspondence is what makes a power counting formula — an integer attached to each diagram counting its total suppression — possible at all.

Example 108.69 (Chiral perturbation theory)**.** The light degrees of freedom below Λ_{QCD} are the pions, which are light not by accident but because they are the almost-Goldstone bosons of the spontaneously broken chiral symmetry of Chapter 102. Their interactions are constrained by that symmetry to vanish as the momenta go to zero, so an expansion in momenta and quark masses exists, and it is an effective field theory in exactly the sense of Theorem 108.67 [1647, 1737]. The scale $E_\star$ is set by the pion decay constant through the loop counting of Remark 108.68:

$$E_\star \approx 4\pi F_\pi = 4\pi \times 92.07\,\mathrm{MeV} = 1.16\,\mathrm{GeV}\,, \tag{108.66}$$

with F_π from Example 102.73, which is where the pionic description must fail, and which is where Theorem 107.29 places it. At $E = m_\pi c^2$ the expansion parameter is $(139.6/1160)^2 = 0.014$: a per-cent theory at leading order, which is what is observed (Remark 102.74).

Example 108.70 (The nuclear force)**.** Applying the same construction to nucleons and pions gives the counting formula Equation (107.38) of Theorem 107.29, in which a diagram with A nucleons, L loops and C connected pieces contributes at order $(Q/\Lambda)^\nu$ with ν a sum of manifestly non-negative terms [1840]. Two consequences follow that no fit to scattering data can supply, and Remark 107.30 draws them: the long-range force is *predicted*, being one-pion exchange with a coupling already measured elsewhere; and many-body forces are ordered, the three-nucleon force entering at $\nu = 3$ and the four-nucleon force at $\nu = 4$, so that the observed necessity of a three-body force is a prediction rather than a patch [2177]. The derivation of the counting formula itself is owed to Appendix A and is recorded as pending in Chapter 107; what this chapter supplies is the reason such a formula exists, which is Theorem 108.67 together with Remark 108.68.

Remark 108.71 (The operator product expansion). Wilson's operator product expansion [1410] asserts that the product of two local operators at short separation may be written

$$O_a(x)\,O_b(0) \;\longrightarrow\; \sum_i C^i_{ab}(x)\,O_i(0) \qquad \text{as } x \to 0\,, \tag{108.67}$$

with the whole singularity as $x \to 0$ carried by the c-number coefficients C^i_{ab} and none by the operators O_i. The degree of the singularity is fixed by dimensions: by Equation (108.3), $C^i_{ab} \sim |x|^{D_i - D_a - D_b}$ up to logarithms, so operators of higher dimension are suppressed by more powers of the separation. This is Equation (108.64) with the two scales exchanged: an effective field theory expands in the ratio of a low energy to a high one, and Equation (108.67) expands in the ratio of a short distance to a long one, but the organizing integer is the same D and the factorization of long from short is the same factorization. In an asymptotically free theory the coefficients are computable in perturbation theory while the matrix elements of the O_i are not, which is exactly the division of labour of Theorem 102.53, and the logarithms the coefficients carry are governed by the anomalous dimensions Equation (102.62) of the operators — the same γ that Definition 108.12 introduced.

108.6.3 The Standard Model as an effective theory

Take the Standard Model Lagrangian Equation (104.28) as the $D = 4$ term of an expansion, keep its gauge symmetry and its field content, and ask what else Theorem 108.67 permits. The answer is short at dimension five and long at dimension six, and the short answer is the more interesting one.

Proposition 108.72 (There is exactly one operator of dimension five)**.** *With the field content of Table 104.1 and the gauge group* $\mathrm{SU}(3)_c \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$*, the only gauge- and*

Lorentz-invariant local operator of dimension $D = 5$ is, up to flavour indices and hermitian conjugation,

$$\mathcal{L}_5 = \frac{\hat{c}_5 \, (\hbar c)^2}{E_\Lambda} \left(\overline{\hat{L}^c} \, \hat{\Phi} \right) \left(\hat{\Phi}^\mathsf{T} \hat{L} \right) + h.c. \,, \tag{108.68}$$

$\hat{L}$ being a lepton doublet, $\hat{\Phi}$ the reduced scalar doublet of Equation (108.2), $\hat{c}_5$ a pure number and E_Λ an energy, in accordance with Equation (108.5) for $D = 5$. It violates lepton number by two units [1646].

Rests on Equations (108.2), (108.3) and (108.5).

Derivation. Derives Proposition 108.72.

Write $D = n_\phi + \frac{3}{2} n_\psi + n_A + n_\partial = 5$ from Equation (108.3). Lorentz invariance requires n_ψ even, and $n_\psi \geq 4$ already gives $D \geq 6$; so $n_\psi \in \{0, 2\}$.

Case $n_\psi = 0$. A purely bosonic gauge-invariant operator is built from $\hat{\Phi}$, $\hat{\Phi}^\dagger$, field strengths and derivatives. A field strength contributes $n_A + n_\partial = 2$; a Lorentz-invariant contraction of leftover derivatives requires them in pairs, contributing 2 each. Hypercharge neutrality with a single doublet requires equal numbers of $\hat{\Phi}$ and $\hat{\Phi}^\dagger$, so n_ϕ is even. Hence D is even, and $D = 5$ is impossible.

Case $n_\psi = 2$. Then $n_\phi + n_A + n_\partial = 5 - 3 = 2$. The options are two scalars, one scalar and one derivative, two derivatives, or one field strength. The last three all reduce: an operator with $n_\partial \geq 1$ acting on a fermion bilinear can be integrated by parts and reduced by the equations of motion to operators of dimension four, and a two-fermion operator with a field strength — the dipole $\bar{\psi} \sigma^{\mu\nu} \psi F_{\mu\nu}$ — is not $\mathrm{SU}(2)_L$ invariant for the chiral fields of Table 104.1 unless a Higgs doublet is inserted to carry the weak index, which raises it to $D = 6$.

There remains $n_\phi = 2$: two fermion fields and two scalars. Both fermions must be left-handed, since a right-handed singlet cannot pair with a doublet and two scalars into a hypercharge singlet; so the bilinear is a Lorentz scalar formed from two left-handed Weyl fields by charge conjugation, which is why the operator violates fermion number. Colour neutrality excludes quark doublets, since $3 \otimes 3$ contains no singlet. The remaining candidate is two lepton doublets, $Y = -1$ each, with two scalar doublets, $Y = +1$ each: total hypercharge $-2+2 = 0$, and $2 \otimes 2 \otimes 2 \otimes 2$ contains a singlet. That is Equation (108.68), and it is unique up to the choice of $\mathrm{SU}(2)$ contraction, which differs from the one written by a term that vanishes identically, and up to the flavour indices carried by the two lepton doublets, which make $\hat{c}_5$ a symmetric matrix in generation space. The exhaustive enumeration behind the reductions asserted above is the systematic construction of [1770, 2159]; the conclusion is [1646]. □

Phenomenon 108.73 (Neutrino mass points above the electroweak scale)**.** The neutrinos of the Standard Model Lagrangian Equation (104.28) are massless, because Table 104.1 contains no right-handed singlet with which to build a Yukawa term. Oscillation experiments show that at least two of them are not (Chapters 103 and 114): the measured squared-mass splittings are $\Delta m^2_{21} c^4 = 7.53 \times 10^{-5} \, \mathrm{eV}^2$ and $|\Delta m^2_{32}| \, c^4 = 2.455 \times 10^{-3} \, \mathrm{eV}^2$ [2427], so the heaviest neutrino satisfies

$$m_3 c^2 \geq \sqrt{|\Delta m^2_{32}| \, c^4} = 0.0495 \, \mathrm{eV} \,. \tag{108.69}$$

Inserted into Equation (108.68) with a coefficient of order unity, this is a measurement of the scale E_Λ at which the Standard Model ceases to be the whole story:

$$E_\Lambda = \frac{\hat{c}_5 E_v^2}{m_3 c^2} = \hat{c}_5 \times \frac{(246.220 \, \mathrm{GeV})^2}{4.95 \times 10^{-11} \, \mathrm{GeV}} = \hat{c}_5 \times 1.2 \times 10^{15} \, \mathrm{GeV} \,. \tag{108.70}$$

Rests on Equations (104.28), (104.29) and (108.68).

Derivation. *Derives Phenomenon 108.73.*

Substitute the vacuum Equation (104.29) into Equation (108.68). The neutral component of $\hat{\Phi}$ acquires $\left\langle \hat{\Phi}^0 \right\rangle = v/\sqrt{2\hbar c} = E_v/\left(\sqrt{2}\,\hbar c\right)$, using $\hat{\Phi} = \Phi/\sqrt{\hbar c}$ from Equation (108.2) and $E_v = v\sqrt{\hbar c}$ from Equation (104.2). The operator becomes

$$\frac{\hat{c}_5\,(\hbar c)^2}{E_\Lambda} \cdot \frac{E_v^2}{2\,(\hbar c)^2}\,\overline{\hat{\nu}^c}\,\hat{\nu} = \frac{\hat{c}_5 E_v^2}{2E_\Lambda}\,\overline{\nu^c}\,\nu\,,$$

which is a Majorana mass term $-\frac{1}{2}m_\nu c^2\overline{\nu^c}\nu$ with

$$m_\nu c^2 = \frac{\hat{c}_5 E_v^2}{E_\Lambda}\,. \tag{108.71}$$

Every factor is dimensionally accounted for: $\hat{\nu}$ carries $1/\mathrm{m}^{3/2}$ so $\overline{\hat{\nu}^c}\hat{\nu}$ carries $1/\mathrm{m}^3$, and the prefactor is an energy, so the product is an energy density as required. Solving Equation (108.71) for E_Λ with $m_\nu c^2 = 0.0495\,\mathrm{eV} = 4.95 \times 10^{-11}\,\mathrm{GeV}$ and $E_v^2 = 60\,624.29\,\mathrm{GeV}^2$ from Equation (104.32) gives Equation (108.70), since $60624.29/4.95 \times 10^{-11} = 1.22 \times 10^{15}$. □

Remark 108.74 (What the neutrino scale does and does not say). Equation (108.70) says: *if* the neutrino masses come from the unique dimension-five operator, and *if* its coefficient is of order one, then the scale at which the Standard Model is superseded is about $10^{15}\,\mathrm{GeV}$. Both conditionals matter. The first is a parametrization of ignorance, not a mechanism: Equation (108.68) records that lepton number is violated somewhere above the electroweak scale and says nothing about what does the violating. The second is unconstrained, and E_Λ scales linearly with $\hat{c}_5$: a coefficient of 10^{-10} would place the scale at $10^5\,\mathrm{GeV}$, and the data cannot distinguish. Nor is the operator the only possibility — the neutrinos may be Dirac particles with Yukawa couplings smaller than the electron's by seven orders of magnitude, the ratio $0.0495\,\mathrm{eV}/0.511\,\mathrm{MeV} = 9.7 \times 10^{-8}$, which is equally unexplained and equally consistent. The alternatives, and the honest statement that no direct evidence selects among them, belong to Chapter 103 and to Chapter 133.

What is not conditional is the structural point, and it is considerable: the smallness of the neutrino masses is, within this framework, not a puzzle but a *measurement*. An irrelevant operator is suppressed by the scale of the physics that generated it, so a very small mass is the signature of a very large scale. This is the only place in the Standard Model where an observed number points above the electroweak scale, and the pointing is the content.

Remark 108.75 (Dimension six, and what the bounds establish). At $D = 6$ the list is no longer short. The first systematic construction [1770], corrected and completed into what is now the standard basis [2159], contains 59 independent baryon-number-conserving operators for a single fermion generation, up to hermitian conjugation; flavour indices multiply that by a large factor for three generations. Their coefficients are constrained by global fits to electroweak precision data, to Higgs couplings and to high-energy tails of distributions at the LHC, and the constraints are usually quoted as lower bounds on $E_\Lambda/\sqrt{\hat{c}}$ of a few TeV: at $E = 1\,\mathrm{TeV}$ the suppression Equation (108.64) is $(E/E_\Lambda)^2$, which is 0.11 for $E_\Lambda = 3\,\mathrm{TeV}$ and 0.01 for $10\,\mathrm{TeV}$, so a measurement at the per-cent level probes the ten-TeV range and no further.

A small subset of the dimension-six operators violates baryon number. They take the form of four-quark-and-lepton operators such as $(\hat{Q}\hat{Q})(\hat{Q}\hat{L})$, they conserve $B - L$, and they mediate proton decay.

Phenomenon 108.76 (The proton does not decay)**.** Super-Kamiokande, searching an enlarged fiducial volume of its water-Cherenkov detector over four running periods, finds no proton decay and sets, at ninety per cent confidence,

$$\frac{\tau}{B}\left(p \to e^+\pi^0\right) > 2.4 \times 10^{34}\,\mathrm{yr}\,, \qquad \frac{\tau}{B}\left(p \to \mu^+\pi^0\right) > 1.6 \times 10^{34}\,\mathrm{yr} \tag{108.72}$$

[2365]. Read through a dimension-six operator with a coefficient of order unity, the first of these requires

$$E_\Lambda \gtrsim 3 \times 10^{16}\,\mathrm{GeV}\,. \tag{108.73}$$

Rests on Equations (108.3) *and* (108.5)*.*

Derivation. *Derives Phenomenon 108.76.*

A dimension-six operator carries, by Equation (108.5), a coefficient $\hat{c}_6 \hbar c\,(\hbar c/E_\Lambda)^2$. The decay rate is quadratic in it, and the only other scale in the problem is the proton mass, so on dimensional grounds

$$\hbar\Gamma \sim \hat{c}_6^2\,\frac{\left(m_p c^2\right)^5}{E_\Lambda^4}\,, \tag{108.74}$$

both sides being energies: five powers of energy divided by four is one. Set $\hat{c}_6 = 1$ and insert the numbers. A year is $3.155\,76 \times 10^7$ s, so $\Gamma = 1/(2.4 \times 10^{34} \times 3.155\,76 \times 10^7\,\mathrm{s}) = 1.320 \times 10^{-42}/\mathrm{s}$, and with $\hbar = 6.582\,119\,569 \times 10^{-16}\,\mathrm{eV\,s}$,

$$\hbar\Gamma = 8.69 \times 10^{-58}\,\mathrm{eV} = 8.69 \times 10^{-67}\,\mathrm{GeV}\,.$$

With $m_p c^2 = 0.938\,272\,\mathrm{GeV}$, $(m_p c^2)^5 = 0.727\,18\,\mathrm{GeV}^5$, so $E_\Lambda^4 = 0.727\,18/8.69 \times 10^{-67} = 8.4 \times 10^{65}\,\mathrm{GeV}^4$ and $E_\Lambda = 3.0 \times 10^{16}\,\mathrm{GeV}$, which is Equation (108.73). The second limit in Equation (108.72) gives $2.7 \times 10^{16}\,\mathrm{GeV}$ by the same arithmetic; the two modes differ by less than the uncertainty in the hadronic matrix element, which is a factor of a few and which Equation (108.74) does not contain. □

Remark 108.77 (What these bounds do and do not establish)*.* Three statements, in decreasing order of security.

Secure. Baryon number is conserved to an extraordinary accuracy, and any physics that violates it does so through operators whose coefficient-to-scale ratio satisfies Equation (108.73). This is a bound on $\hat{c}_6/E_\Lambda^2$ and on nothing else. A theory with $\hat{c}_6 = 10^{-12}$ and $E_\Lambda = 3 \times 10^{13}\,\mathrm{GeV}$ is equally consistent with Equation (108.72), and there is no measurement that separates them.

Weaker. The near-coincidence of the neutrino scale Equation (108.70), at $10^{15}\,\mathrm{GeV}$, with the proton bound Equation (108.73), at $3 \times 10^{16}\,\mathrm{GeV}$, is suggestive of a common origin. It is one order of magnitude apart in quantities whose coefficients are unknown by many orders of magnitude, so it is suggestive and not more.

Not established at all. The programme that historically motivated the proton-decay searches — grand unification, in which the three gauge couplings are supposed to meet at a single scale — has no observational support. Its central prediction is proton decay, and the prediction has been tested for forty years with a null result that has excluded the simplest versions. By editorial rule 1 it is named here to record that fact and for no other reason. The renormalization-group statement that survives is the one that does not depend on it: an unobserved decay is a bound on a coefficient over a squared scale, and Equation (108.73) is that bound.

108.7 The renormalization group beyond perturbation theory

108.7.1 Scale invariance and its breaking

Definition 108.78 (Dilatations and the trace of the stress tensor)**.** A *dilatation* is the transformation $x^\mu \mapsto \lambda x^\mu$ with λ a pure number, each field transforming as $\Phi \mapsto \lambda^{-D_\Phi}\Phi$ with D_Φ its dimension from Equation (108.3). If the action is invariant, Noether's theorem supplies a conserved current

$$\mathcal{D}^\mu = x_\nu T^{\mu\nu}\,, \qquad \partial_\mu \mathcal{D}^\mu = T^\mu{}_\mu\,, \tag{108.75}$$

$T^{\mu\nu}$ being the (improved, symmetric) energy–momentum tensor, of SI dimension $\mathrm{J/m^3}$, whose general construction belongs to Chapter 64. Classical scale invariance is therefore the statement $T^\mu{}_\mu = 0$.

Proposition 108.79 (A classically massless theory is classically scale invariant)**.** *If every coupling in the Lagrangian multiplies an operator of dimension exactly four — that is, if every coupling is a pure multiple of $\hbar c$ by Equation (108.4) — then the classical action is dilatation invariant and $T^\mu{}_\mu = 0$.*

Rests on Definition 108.78 and Equation (108.4).

Proof. *Derives Proposition 108.79.*

Under $x \mapsto \lambda x$ the measure d^4x carries λ^4 and an operator of dimension D carries λ^{-D}, so a term $c_O \int \mathrm{d}^4x\, O$ is invariant if and only if $D = 4$ or $c_O = 0$. Massless Yang–Mills and massless ϕ^4 have only $D = 4$ terms; a mass term has $D = 2$ or 3 and breaks the invariance explicitly, by an amount proportional to the mass. Hence $T^\mu{}_\mu = \partial_\mu \mathcal{D}^\mu = 0$ for the massless theory. □

Theorem 108.80 (The trace anomaly: structure)**.** *In the quantum theory the trace does not vanish, even when every mass is set to zero. Its expectation value is proportional to the beta function times the operator conjugate to the coupling:*

$$T^\mu{}_\mu = \frac{\beta(\hat g)}{\hat g}\,\mathcal{O}_{\hat g} + \sum_f \left(1 + \gamma_{m_f}\right) m_f c^2\, \bar\psi_f \psi_f\,, \qquad \mathcal{O}_{\hat g} := \hat g\,\frac{\partial \mathcal{L}}{\partial \hat g}\,, \tag{108.76}$$

up to a numerical factor in the first term which is not derived here. In particular $T^\mu{}_\mu = 0$ in a massless theory if and only if $\beta = 0$: exact scale invariance survives quantization only at a fixed point.

Rests on Proposition 108.79, Corollary 108.23, and Equation (108.75).

Derivation of the structure. *Derives Theorem 108.80.*

Regulate the theory as in Definition 100.30. By Corollary 108.23 the whole dependence of the regulated Lagrangian on the arbitrary scale $\bar\mu$ resides in the coupling: changing $\bar\mu$ changes $\hat g$ by $\delta\hat g = \beta(\hat g)\,\delta \ln \bar\mu$ and changes nothing else. A dilatation by $\lambda = 1 + \delta \ln \lambda$ acts on the regulated theory exactly like a change of $\bar\mu$ by $\delta \ln \bar\mu = -\delta \ln \lambda$, since $\bar\mu$ is the only scale present and lengths and $\bar\mu^{-1}$ scale oppositely. Therefore

$$\delta S\big|_{\text{dilatation}} = -\frac{\partial S}{\partial \hat g}\,\beta(\hat g)\;\delta \ln \lambda\,,$$

while by Equation (108.75) the same variation is $\delta S = \delta \ln \lambda\, c^{-1} \int \mathrm{d}^4x\, T^\mu{}_\mu$. Comparing integrands, $T^\mu{}_\mu$ is proportional to $\beta(\hat g)\,\partial\mathcal{L}/\partial\hat g$, which is Equation (108.76) up to the constant. The mass term is the classical breaking of Proposition 108.79, corrected by the anomalous dimension γ_m of Equation (108.14) because the mass that appears is the running one. □

Derivation pending #180

The trace anomaly, complete: the operator equation for the trace of the improved energy–momentum tensor in a non-abelian gauge theory, with the numerical coefficient of the beta-function term obtained by explicit regularization rather than by the scaling argument given in the text, together with the demonstration that the result is independent of the regulator used and that the improvement terms do not contribute. The structure — proportionality to the beta function times the operator conjugate to the coupling, and vanishing of the trace exactly at a zero of the beta function — is derived above; the coefficient is not, and the full operator derivation belongs in Appendix A.

Corollary 108.81 (Almost all of the proton's mass is anomalous)**.** *For a hadron state of four-momentum p — the momentum symbol of Notation 100.1, the wave four-vector $k = p/\hbar$ playing no part here — Lorentz invariance fixes the forward matrix element of the energy–momentum tensor to be proportional to $p^\mu p^\nu$, and in the covariant normalization customary for it the constant is $2\mathcal{N}$:*

$$\langle p|T^{\mu\nu}|p\rangle = 2\mathcal{N}\, p^\mu p^\nu \,, \tag{108.77}$$

where $\mathcal{N}$ carries whatever SI dimension the chosen state normalization does. It is not a pure number: $[T^{\mu\nu}] = \mathrm{J/m^3}$ while $p^\mu p^\nu$ carries $\mathrm{kg^2\,m^2/s^2}$, so $[\mathcal{N}] = 1/(\mathrm{kg\,m^3})$ — and nothing below depends on its value, because it cancels from every ratio taken. Contracting Equation (108.77) with the metric,

$$\langle p|T^{\mu}{}_{\mu}|p\rangle = 2\mathcal{N}p^2 = 2\mathcal{N}m^2c^2 \,, \tag{108.78}$$

by the mass-shell condition Equation (100.1): the trace of the tensor measures the mass and nothing else. Since the quark mass terms in Equation (108.76) account for only a few MeV *of the proton's* 938.272 MeV*, the remainder — some* 99 % *— comes from the first term, that is from the beta function.*

Rests on Equations (100.1) and (108.76).

Proof. *Derives Corollary 108.81.*

Equation (108.78) is immediate from the stated normalization and Equation (100.1). The quark contribution is bounded by $(m_u + m_d + m_s)c^2$ times matrix elements of order unity, and Table 102.1 gives $m_u c^2 = 2.16\,\mathrm{MeV}$ and $m_d c^2 = 4.70\,\mathrm{MeV}$; the strange contribution is larger but still small. The rest is carried by $\mathcal{O}_{\hat{g}}$, whose coefficient vanishes with β. □

Remark 108.82 (Two ways of saying the same thing). Corollary 108.81 and Phenomenon 102.39 are one statement in two languages. In the language of the running coupling, a massless Lagrangian generates the scale Λ_{QCD} of Equation (102.45) by dimensional transmutation, and every hadron mass is a pure number times $\Lambda_{\mathrm{QCD}}/c^2$. In the language of the trace anomaly, the same fact says that the classical scale invariance of the massless theory is destroyed by quantization with a coefficient proportional to β, and the proton's mass is what the destruction produces. Neither is more fundamental; the second is the operator statement of the first, and both are consequences of the single fact that $\beta \neq 0$. Chapter 102 notes the identification in passing at Remark 102.40; this is where it is made precise.

Remark 108.83 (Conformal invariance at a fixed point). At a zero of β in a massless theory the trace vanishes and the theory is scale invariant. In every case relevant here it is more: Lorentz invariance, locality and scale invariance together promote the symmetry to the full conformal group, which in $3+1$ dimensions has fifteen parameters against the eleven of

Poincaré plus dilatations, the four extra being the special conformal transformations. The promotion is a theorem under technical hypotheses and not an identity — there are known counterexamples in other dimensions — so it is recorded here as what it is.

The consequences are not decorative. Conformal invariance constrains two- and three-point functions completely, up to constants, and it is those constraints, combined with unitarity and the associativity of the operator product expansion Equation (108.67), that the conformal bootstrap turns into numerical bounds. That is how the three-dimensional Ising exponents in the last column of Table 108.1 were obtained [2239]: not by computing a beta function at all, but by asking which sets of dimensions are consistent, and finding that consistency alone determines them to six figures. The critical point of a fluid or a uniaxial magnet is a conformal field theory, and the exponents measured in Section 123.5 are its operator dimensions.

108.7.2 Monotonicity of the flow

Remark 108.84 (The empirical fact). Coarse graining loses information — Remark 108.35 — and in every case that has been examined the loss shows up as a reduction in the number of degrees of freedom describing the theory. A ferromagnet above its Curie point is described by the full lattice of spins; at long wavelength it is described by one order parameter. Quantum chromodynamics at short distance has quarks and gluons, six flavours and eight colours; below Λ_{QCD} it has pions, and below the pion mass it has nothing at all in the strong sector. A heavy field integrated out by Theorem 108.64 is gone. The direction is always the same, and never the reverse.

Theorem 108.85 (The four-dimensional a-theorem)**.** *Let a unitary, Lorentz-invariant, four-dimensional quantum field theory flow from one conformal fixed point in the ultraviolet to another in the infrared. Each fixed point has a number a, defined as the coefficient of the Euler density in the trace anomaly of the theory placed in a background gravitational field. Then*

$$a_{\mathrm{UV}} > a_{\mathrm{IR}}\,, \tag{108.79}$$

so no such flow can be reversed [2175]. The two-dimensional predecessor of this statement, in which the corresponding quantity is the Virasoro central charge and the result is known as the c-theorem, was proved by Zamolodchikov in 1986; that paper is not in this book's bibliography and is named here without a citation rather than attached to a key that does not contain it.

Rests on Definition 108.78 and Theorem 108.80.

Derivation pending #181

The four-dimensional a-theorem: the construction of the dilaton effective action for a spontaneously broken conformal symmetry, the identification of the difference of the anomaly coefficients with the coefficient of the four-derivative dilaton self-interaction, and the positivity of that coefficient from the analyticity and unitarity of forward dilaton scattering, by an argument of the same shape as an unsubtracted dispersion relation. The statement is used here only qualitatively, to say that the flow has a direction; the proof is several pages and belongs in Appendix A.

Remark 108.86 (Why irreversibility would matter, and why it does not yet). If the flow is a gradient flow of a monotone function, then the space of possible ultraviolet completions of a given low-energy theory is constrained: a candidate completion must have a larger a than the theory it completes, which is a necessary condition that can be checked. That is the reason the theorem is of interest here rather than only in mathematical physics.

Three qualifications must be attached, and the third is the one usually left out.

First, Theorem 108.85 is a statement about flows *between conformal fixed points*. The Standard Model is not such a flow. Its infrared is not a conformal theory: it contains a massless photon and nearly massless neutrinos alongside a gapped strong sector, and its ultraviolet is not a fixed point at all — by Corollary 108.62 the scalar and hypercharge couplings grow and the description fails. The theorem does not apply to the theory this book is about.

Second, the quantity a is not measurable. It is a coefficient in the response of a theory to a background curvature, and no experiment has measured one.

Third, and consequently, no classification of the possible ultraviolet completions of the observed $3 + 1$ Standard Model exists, and none is in prospect. What exists is a theorem about a class of theories to which the Standard Model does not belong, together with the empirical observation of Remark 108.84 that in every examined case the flow has run one way. That is worth recording and it is not worth more than that. It is recorded here rather than in Chapter 133, which lists the measured discrepancies and the one missing existence proof and does not carry the absence of a classification of ultraviolet completions as an entry — correctly, since nothing is measured that such a classification would explain.

Remark 108.87 (What this chapter has and has not established). The measured content is small, sharp and repeatedly confirmed. Two dimensionless couplings of the Standard Model have been watched to change with the energy at which they are measured, in opposite directions and by amounts predicted with no free coefficient (Phenomena 100.62 and 108.56). A critical exponent has been measured to four decimal places against a calculation of comparable precision, and the two do not quite agree (Phenomenon 108.49). The universality of critical behaviour across chemically unrelated systems is explained, not merely described (Corollary 108.44). And an observed number — the neutrino mass — points to a scale fifteen orders of magnitude above the electroweak one, conditionally but for a stated reason (Phenomenon 108.73).

The unmeasured content is larger and should not be confused with the above. The location of any Landau pole is an extrapolation over dozens of decades. The triviality of the scalar sector is a theorem about ϕ^4 in isolation, not about the Standard Model. The cutoff of the Standard Model is unmeasured, and every statement about what lies above it is a statement about a bound on a coefficient divided by a power of a scale. The renormalization group tells one how to be predictive without knowing what lies above; it does not tell one what lies above, and the two must not be traded for one another.

Chapter 109

Axiomatic Quantum Field Theory

Everything the preceding chapters of Part XI compute is perturbative: a formal power series in a coupling, defined by rules whose objects — an interaction picture, a Fock space of free quanta, a product of fields at the same point — do not survive mathematical scrutiny. Haag's theorem [1072] states the sharpest form of the difficulty: no unitary transformation connects a free field to an interacting one in the same Hilbert space, so the interaction picture that generates the whole perturbative apparatus does not exist. Axiomatic quantum field theory is the response. It asks what a relativistic quantum field theory *is*, independently of any Lagrangian or expansion, writes that down as a short list of assumptions — fields as operator-valued distributions on a Hilbert space carrying a positive-energy unitary representation of the Poincaré group, with a unique vacuum and local commutativity [1099, 1297] — and derives consequences.

The chapter earns its place because the consequences are theorems the rest of the book uses as if they were facts of Nature: the PCT theorem [1049, 1115], the spin–statistics connection [854, 1139, 1151] that Chapters 83 and 120 assume throughout, the Källén–Lehmann spectral representation [1003, 1048] that underlies every renormalized propagator in Chapters 100 and 108, the LSZ reduction [1073] that connects fields to the cross-sections of Chapter 86, and the operator product expansion [1410] on which the short-distance analysis of Chapters 102 and 116 rests. It closes honestly: no interacting quantum field theory in $3+1$ dimensions has ever been constructed satisfying these axioms. The constructive successes are in spacetimes of lower dimension [1371, 1500], and the existence of four-dimensional Yang–Mills theory with a mass gap is an open problem carrying a Clay Millennium Prize [2097]. Standing monographs are [1297, 1312, 1792, 1856].

Three analytic facts are used repeatedly below and are *imported*: the Paley–Wiener characterization of the Fourier transforms of functions with support in a cone, the edge-of-the-wedge theorem, and the Bargmann–Hall–Wightman extension theorem [1113]. None is proved in this book, and none is among the results of Chapter 17 or of Chapter 8, whose function theory stops at the residue theorem. Where one of them is used it is named at the point of use, and everything else in the chapter is derived. That bookkeeping is the whole point of an axiomatic chapter: the reader is entitled to know exactly which statements are assumptions, which are imports, and which are consequences.

109.1 Why axioms

109.1.1 Fields as operator-valued distributions

The first axiom is forced by a computation, not chosen for elegance. A quantum field cannot be an operator at a point, because the vector it would produce out of the vacuum has infinite length.

Proposition 109.1 (The field at a point is not an operator)**.** *Let ϕ be a Hermitian scalar field whose two-point function is that of a free field of mass m,*

$$\langle 0|\,\phi(x)\phi(y)\,|0\rangle = \hbar c^2 \int \frac{\mathrm{d}^3k}{(2\pi)^3\,2\omega_{\boldsymbol{k}}}\,\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}\,, \qquad \omega_{\boldsymbol{k}} = c\sqrt{\boldsymbol{k}^2 + \frac{m^2c^2}{\hbar^2}}\,. \tag{109.1}$$

Then $\|\phi(x)\,|0\rangle\|^2 = +\infty$ for every x, while $\|\phi(f)\,|0\rangle\|^2$ is finite for every Schwartz test function f, where $\phi(f) = \int \mathrm{d}^4x\, f(x)\phi(x)$.

Rests on Notation 100.1 and Equation (105.101).

Derivation. Derives Proposition 109.1.

In Equation (109.1) the wave four-vector is $k^\mu = (\omega_{\boldsymbol{k}}/c, \boldsymbol{k})$, of dimension 1/m, so $k\cdot(x-y)$ is a pure number (Notation 100.1); the integral $\int \mathrm{d}^3k/(2\omega_{\boldsymbol{k}})$ carries $\mathrm{s/m^3}$ and the prefactor $\hbar c^2$ turns that into J/m, which is $[\phi]^2$ for a canonically normalized scalar in this book (Equation (105.101)). Setting $y = x$,

$$\|\phi(x)\,|0\rangle\|^2 = \hbar c^2 \int \frac{\mathrm{d}^3k}{(2\pi)^3\,2\omega_{\boldsymbol{k}}} = \frac{\hbar c^2}{(2\pi)^3}\,4\pi \int_0^\infty \frac{k^2\,\mathrm{d}k}{2\omega_k}\,, \tag{109.2}$$

and for $k \gg mc/\hbar$ the integrand is $k^2/(2ck) = k/(2c)$, so the integral diverges as Λ^2 with an ultraviolet cutoff Λ on $|\boldsymbol{k}|$. The divergence is quadratic and no subtraction of a constant removes it.

Smearing repairs it. With f real and $\tilde f(k) = \int \mathrm{d}^4x\, f(x)\mathrm{e}^{\mathrm{i}k\cdot x}$,

$$\|\phi(f)\,|0\rangle\|^2 = \hbar c^2 \int \frac{\mathrm{d}^3k}{(2\pi)^3\,2\omega_{\boldsymbol{k}}}\,\left|\tilde f(k)\right|^2\Big|_{k^0=\omega_{\boldsymbol{k}}/c}\,, \tag{109.3}$$

which converges because $\tilde f$ is Schwartz and therefore decays faster than any power. The physical content is that the divergence in Equation (109.2) is a statement about resolution: the number of modes available below a wavelength λ grows as λ^{-3}, each carries a zero-point amplitude, and a measurement confined to no region at all samples all of them. □

Definition 109.2 (Smeared field and test functions)**.** $\mathcal{S}(\mathbb{R}^4)$ is the Schwartz space of complex functions on Minkowski space, smooth and decaying with all derivatives faster than any power [971, 1276]; the treatment used here is Section 17.2.3. A test function $f \in \mathcal{S}$ is assigned the SI dimension $1/\mathrm{m}^4$, so that

$$\phi(f) := \int \mathrm{d}^4x\, f(x)\,\phi(x) \tag{109.4}$$

carries the dimension of the field itself. A *quantum field* is then a map $f \mapsto \phi(f)$ from $\mathcal{S}(\mathbb{R}^4)$ into (generally unbounded) operators on a Hilbert space, linear in f and continuous in the topology of $\mathcal{S}$: an operator-valued tempered distribution [1099, 1307].

The word *tempered* is not decoration either. A distribution is tempered exactly when its Fourier transform is a distribution of polynomial growth, and the spectrum condition of Axiom 109.10 will be a support statement about that Fourier transform. A field whose smeared correlation functions grew faster than polynomially in momentum would have no spectral condition to satisfy, and the entire analytic apparatus of Section 109.2.3 would be unavailable.

Bohr and Rosenfeld reached the same conclusion from the laboratory rather than from the mathematics [715]. A charged test body used to measure a field has finite extent and finite mass, so what it registers is an average of the field over the spacetime region it occupies during the exposure — exactly Equation (109.4) with f the indicator of that region — and the unavoidable uncertainties of such an apparatus turn out to be precisely those the commutators of the averaged fields demand, neither more nor less. That analysis is carried out for the electromagnetic field in Section 99.7.3; the point to carry forward is that the smearing is a physical prescription and not a mathematical convenience.

Remark 109.3 (What a common dense domain is for). An unbounded operator is not defined on all of a Hilbert space, and a product $\phi(f_1)\phi(f_2)$ makes no sense unless the range of the second lies in the domain of the first. The axioms therefore demand a single dense subspace $D \subset \mathcal{H}$, invariant under every $\phi(f)$ and containing the vacuum, on which all polynomials in the smeared fields are defined. Without it the Wightman functions of Section 109.2.2 would be symbols rather than numbers. The functional analysis of unbounded operators, domains and self-adjointness is standard [1481]; the chapter uses it and does not develop it.

109.1.2 Haag's theorem

Perturbation theory is built on the interaction picture, in which the interacting field at a fixed time is a unitary transform of a free field, so that both live in one Hilbert space and the dynamics is a unitary evolution of free-field states. Haag proved that this cannot happen [1072], and Hall and Wightman sharpened the hypotheses to the form used here [1113]. Theorem 99.53 states the result where it first bites; this is its derivation.

Theorem 109.4 (Haag)**.** *Let ϕ_1, π_1 and ϕ_2, π_2 be two irreducible sets of field and conjugate momentum at the common time $t = 0$, each acting on a Hilbert space carrying a continuous unitary representation of the Euclidean group of space — translations and rotations — with a unique invariant vacuum $|0_1\rangle$, $|0_2\rangle$. Suppose there is a unitary V with*

$$\phi_2(0,\boldsymbol{x}) = V\,\phi_1(0,\boldsymbol{x})\,V^{-1}\,, \qquad \pi_2(0,\boldsymbol{x}) = V\,\pi_1(0,\boldsymbol{x})\,V^{-1}\,, \tag{109.5}$$

intertwining the two representations of the Euclidean group. Then all equal-time vacuum expectation values of the two fields coincide. If in addition both theories are Poincaré covariant with spectrum in the forward cone and ϕ_1 is a free field of mass m, then ϕ_2 is a free field of the same mass.

Rests on Equation (109.1) and Theorems 109.21 and 109.26.

Derivation. *Derives Theorem 109.4.*

Step 1: the vacuum goes to the vacuum. Let E denote the Euclidean group of space and $U_i(R,\boldsymbol{a})$ its representation on $\mathcal{H}_i$. By hypothesis $VU_1V^{-1} = U_2$. Then $V\,|0_1\rangle$ is invariant under U_2, because $U_2V\,|0_1\rangle = VU_1\,|0_1\rangle = V\,|0_1\rangle$. Uniqueness of the invariant vector in theory 2 forces

$$V\,|0_1\rangle = \mathrm{e}^{\mathrm{i}\theta}\,|0_2\rangle\,, \qquad \theta \in \mathbb{R}\,. \tag{109.6}$$

Step 2: all equal-time correlators coincide. Let A_1 be any polynomial in the smeared $\phi_1(0,\cdot)$ and $\pi_1(0,\cdot)$, and A_2 the same polynomial in ϕ_2, π_2. Conjugating Equation (109.5) term by term gives $A_2 = VA_1V^{-1}$, and with Equation (109.6) the phases cancel between bra and ket:

$$\langle 0_2|\,A_2\,|0_2\rangle = \langle 0_1|\,V^{-1}VA_1V^{-1}V\,|0_1\rangle = \langle 0_1|\,A_1\,|0_1\rangle\,. \tag{109.7}$$

This holds to all orders and is the whole content of unitary equivalence: a unitary preserves every expectation value, so the two theories are indistinguishable by any equal-time measurement.

Step 3: the equal-time two-point function knows the mass. For a free field of mass m, set $x^0 = y^0$ in Equation (109.1):

$$\langle 0|\,\phi(0,\boldsymbol{x})\phi(0,\boldsymbol{y})\,|0\rangle = \hbar c^2 \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\frac{\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})}}{2c\sqrt{\boldsymbol{k}^2 + m^2c^2/\hbar^2}}\,. \tag{109.8}$$

The right-hand side is the Fourier transform of $\hbar c/(2\sqrt{\boldsymbol{k}^2 + m^2c^2/\hbar^2})$, and two different masses give two different functions of $\boldsymbol{k}$ — they already differ at $\boldsymbol{k} = 0$, where the value is $\hbar^2/(2m)$ — of dimension $\mathrm{J\,m^2}$, as it must be to produce J/m against the measure $\mathrm{d}^3k/(2\pi)^3$. Fourier inversion is injective, so Equation (109.8) determines m.

Step 4: from equal times to all times. By Step 2 the equal-time correlators of ϕ_2 are those of a free field of mass m, and by Step 3 that mass is m_1. A free field is Gaussian: all its correlators are sums of products of two-point functions, so *all* equal-time correlators of ϕ_2 agree with the free ones. The remaining step promotes this to all times. Poincaré covariance determines $\phi_2(x^0, \boldsymbol{x})$ from its restriction to $x^0 = 0$ together with π_2, because a boost mixes the time argument into the spatial ones and the Wightman functions are boundary values of functions analytic in the tube domain of Theorem 109.26, which are fixed by their values on a real neighbourhood. Hall and Wightman carried this out [1113]; the conclusion is that every W_n of theory 2 is the free one, and by the reconstruction theorem Theorem 109.21 the theory itself is free. □

Corollary 109.5 (The interaction picture does not exist)**.** *There is no unitary operator relating the fields of an interacting theory at a fixed time to those of a free theory in the same Hilbert space. Equivalently, the two representations of the canonical commutation relations are unitarily inequivalent, and the interacting vacuum is not a vector in the free theory's Fock space.*

Rests on Theorem 109.4.

Derivation. *Derives Corollary 109.5.*

Immediate from Theorem 109.4: if such a V existed the interacting theory would be free. The second sentence is the same statement in representation-theoretic language — Stone–von Neumann uniqueness holds for finitely many canonical pairs and fails for infinitely many, and quantum field theory has infinitely many. The mechanism is developed for the free field itself in Section 99.9.1, whose warning about the choice of Fock space this corollary makes precise. □

Remark 109.6 (What practitioners actually do, and what it costs)*.* Every one of Haag's hypotheses is violated in a real calculation, and each violation is the point of the corresponding device.

A momentum cutoff destroys Lorentz covariance and, with it, Step 4. *A finite volume* destroys translation invariance and with it Step 1, since a box has no unique translation-invariant state in the required sense. *Adiabatic switching*, in which the coupling is multiplied by a function $g(t)$ that vanishes as $t \to \pm\infty$, destroys time-translation invariance and hence the identification of an interacting vacuum at all. *Dimensional regularization* changes the dimension of spacetime, so the theorem as stated does not apply.

What is lost is the meaning of the limit. Each regularized theory is perfectly consistent and Haag's theorem says nothing about it; the difficulty is that the objects the regularization is supposed to approximate — an interacting field with a Poincaré-invariant vacuum in $3+1$ dimensions — are not known to exist. Perturbation theory is therefore an asymptotic construction: a recipe generating a formal power series whose terms are finite after renormalization (Chapter 100) and whose sum is believed to diverge [998]. That the first few terms of that series reproduce the electron's anomalous moment to about one part in 10^9 of a_e — the limit set by the independent determinations of α that the series must be fed, not by the measurement, as Chapter 112 is at pains to establish — is the strongest evidence available that the series is computing something real. It is not a proof that the something exists.

109.1.3 What axiomatization is for

An axiom system supports two quite different claims, and confusing them is the commonest error made about this subject.

The first is *conditional*: if a structure satisfying the axioms exists, then it has such and such properties. Every theorem of Section 109.3 is of this kind. *PCT* invariance,

spin–statistics, the Källén–Lehmann representation and the Reeh–Schlieder theorem are consequences of the axioms and therefore hold in *every* theory satisfying them, whatever its Lagrangian, whatever its particle content, to all orders and beyond all orders. That is a far stronger statement than any perturbative calculation can make, and it is why a measured violation of PCT would not be a discovery of a new particle but a refutation of locality, of Lorentz invariance, or of the positivity of the energy (Section 105.5.2).

The second is *existential*: a structure satisfying the axioms exists. Here axiomatic quantum field theory has almost nothing to offer in $3+1$ dimensions. No interacting example is known (Section 109.7), and for the simplest candidate — a single self-interacting scalar — the evidence points the other way (Section 109.7.2).

Remark 109.7 (The model-theoretic parallel, and its limits). The distinction is the one drawn in Section 3.8 between what a formal system proves and whether it has a model. A consistent first-order theory has a model, by completeness; but consistency itself is not provable from within (Theorem 3.88), and the existence question for a physical theory is harder still, because the Wightman axioms are not a first-order theory — they quantify over Hilbert spaces, operator domains and distributions, all of which are higher-order objects. Definition 3.82 does not apply to them, and no completeness theorem is available. The parallel is therefore an analogy about the *shape* of the two questions and not a transfer of results: nobody has shown that constructing a four-dimensional interacting field theory is undecidable, and nobody should say so.

The position taken in this book follows from the distinction. A theorem derived from the axioms is treated as a statement about all field theories and quoted as such throughout Part XI. A perturbative calculation is treated as a computation within a scheme, whose agreement with experiment is evidence for the scheme and not a proof of anything. And the fact that the Standard Model, the most accurately tested theory in physics, is not known to exist as a mathematical object is recorded in Section 109.7.4 as what it is: an open problem, not a technicality.

109.2 The Wightman axioms

109.2.1 Statement of the axioms

The axioms below are those of [1099], in the form given by Streater and Wightman [1297], with every dimensioned quantity carried in SI. They are stated for a single Hermitian scalar field; the extension to several fields and to fields carrying Lorentz indices is notational and is indicated in Remark 109.16.

Notation 109.8 (Conventions of this chapter). The metric, coordinates and momenta are those of Notation 100.1: $\eta_{\mu\nu}=\mathrm{diag}(+1,-1,-1,-1)$, $x^\mu=(ct,\boldsymbol{x})$ of dimension m, $p^\mu=(E/c,\boldsymbol{p})$ of dimension $\mathrm{kg\,m/s}$, and $k^\mu=p^\mu/\hbar$ of dimension $1/\mathrm{m}$. The closed forward cone in momentum space is

$$\overline{V}^{+}:=\left\{p\in\mathbb{R}^4:\ p\cdot p\geq 0 \text{ and } p^0\geq 0\right\}, \tag{109.9}$$

and V^+ is its interior. The corresponding cone in coordinate space, written $\overline{V}^{+}_{x}$, is the same set with m in place of $\mathrm{kg\,m/s}$. A scalar field is normalized so that $[\phi]^2=\mathrm{J/m}$, as in Chapter 99. $\mathcal{H}$ denotes the Hilbert space and $D\subset\mathcal{H}$ the common dense domain of Remark 109.3.

Axiom 109.9 (Relativistic states)**.** The states of the theory are the rays of a separable Hilbert space $\mathcal{H}$, on which there is a continuous unitary representation $U(a,\Lambda)$ of the universal covering group $\widetilde{\mathcal{P}}^{\uparrow}_{+}$ of the proper orthochronous Poincaré group (Definition 39.5), with a

a translation of dimension m. Translations are generated by four commuting self-adjoint operators P^μ of dimension kg m/s,

$$U(a,\mathbb{1}) = \exp\!\left(\frac{\mathrm{i}}{\hbar}\, a^\mu P_\mu\right) , \tag{109.10}$$

the exponent being dimensionless because $\hbar$ has the dimension of m × kg m/s.

Axiom 109.10 (Spectrum condition)**.** The joint spectrum of (P^0, P^1, P^2, P^3) is contained in $\overline{V}^+$. Physically: no state has negative energy in any inertial frame, and no state has spacelike four-momentum.

Axiom 109.11 (Unique vacuum)**.** There is a vector $|0\rangle \in D$, unique up to phase, invariant under every $U(a,\Lambda)$. It is the only eigenvector of P^μ with eigenvalue zero.

Axiom 109.12 (Fields)**.** There is a map $f \mapsto \phi(f)$ from $\mathcal{S}(\mathbb{R}^4)$ into operators defined on D, linear in f, with $\phi(f)D \subset D$ and $|0\rangle \in D$, such that $f \mapsto \langle\Psi|\,\phi(f)\,|\Psi'\rangle$ is a tempered distribution for every $\Psi, \Psi' \in D$, and $\phi(f)^\dagger \supset \phi(\bar f)$ on D.

Axiom 109.13 (Covariance)**.** For all (a,Λ) and all f,

$$U(a,\Lambda)\,\phi(f)\,U(a,\Lambda)^{-1} = \phi\Big(f_{(a,\Lambda)}\Big) , \qquad f_{(a,\Lambda)}(x) := f\Big(\Lambda^{-1}(x-a)\Big) , \tag{109.11}$$

which is the smeared form of $U\phi(x)U^{-1} = \phi(\Lambda x + a)$.

Axiom 109.14 (Local commutativity)**.** If the supports of f and g are spacelike separated — every point of one is at spacelike interval from every point of the other — then

$$[\phi(f), \phi(g)] = 0 \tag{109.12}$$

on D. For a field of half-integer spin the commutator is replaced by the anticommutator; that this replacement is forced and not chosen is Theorem 109.49.

Axiom 109.15 (Cyclicity of the vacuum)**.** The set of vectors $\phi(f_1)\cdots\phi(f_n)\,|0\rangle$, over all n and all test functions, spans a dense subspace of $\mathcal{H}$.

Remark 109.16 (Several fields, and fields with indices). For a finite collection ϕ_j carrying a finite-dimensional representation $S(\Lambda)$ of the covering group $\mathrm{SL}(2,\mathbb{C})$ of the Lorentz group, Equation (109.11) acquires the matrix $S(\Lambda)^{-1}_{jk}$ and Equation (109.12) is imposed on every pair, with the commutator for tensor fields and the anticommutator for spinor fields. Nothing else changes, and no theorem below depends on the number of fields. Two different pieces of representation theory are involved and it is worth keeping them apart. Chapter 94 classifies the *unitary* representations of the Poincaré group by Wigner's method — the Casimirs, the orbits, the little groups, massive spin and massless helicity — which is what labels the one-particle states; the general apparatus behind it is Chapters 5 and 14, and $\mathrm{SL}(2,\mathbb{C})$ appears there as the universal cover. What $S(\Lambda)$ carries is the other object: the *finite-dimensional*, necessarily non-unitary representations of $\mathrm{SL}(2,\mathbb{C})$, labelled by a pair of half-integers (A,B) and built as tensor products of the fundamental and its conjugate. That classification is not derived anywhere in this book; it is used below only through the index count of Theorem 109.49, where a field in the (A,B) representation carries $2A$ undotted and $2B$ dotted spinor indices.

Remark 109.17 (Reading this chapter against the literature). The monographs [1297, 1312, 1856] all work in units in which $\hbar$ and c are numerically one, and in that convention the exponent of Equation (109.10) is $\mathrm{i}a\cdot P$ and momenta, wave vectors and inverse lengths are the same quantity. The rule for reading their formulas here is the one already given in

Remark 100.4: a momentum p becomes $p/\hbar$ wherever it multiplies a coordinate, a mass m becomes mc wherever it stands beside a momentum and mc^2 wherever it stands beside an energy, and the spectral variable s of Section 109.3.3 is a squared momentum, $\mathrm{kg}^2\,\mathrm{m}^2/\mathrm{s}^2$, rather than a squared mass. No derivation in this chapter is carried out in those units; this remark exists so that the reader can check the chapter against its sources.

Remark 109.18 (What is deliberately absent). Three things a physicist expects are not axioms. There is no Lagrangian, no equation of motion and no coupling constant: the axioms describe a theory, not a way of constructing one. There is no requirement that the theory have particles — the existence of one-particle states is an additional assumption, made in Section 109.3.4 and known to fail for the electron in quantum electrodynamics, where the charged states are infraparticles [1684]. And there is no assumption of *asymptotic completeness*, that the scattering states span the whole Hilbert space; it is an extra hypothesis, not a theorem, and it has never been verified in an interacting four-dimensional model.

109.2.2 Wightman functions and reconstruction

The axioms are about operators, which are hard to handle. Their entire content can be repackaged as a sequence of ordinary distributions, and this is what makes the subject tractable.

Definition 109.19 (Wightman functions)**.** The *Wightman functions* of the theory are the vacuum expectation values

$$W_n(x_1,\dots,x_n) := \langle 0|\,\phi(x_1)\cdots\phi(x_n)\,|0\rangle\ , \tag{109.13}$$

understood as tempered distributions: the number attached to test functions $f_1,\dots,f_n$ is $\langle 0|\,\phi(f_1)\cdots\phi(f_n)\,|0\rangle$, with $W_0 = 1$. Each W_n carries the SI dimension $(\mathrm{J/m})^{n/2}$, that is $[\phi]^n$.

Proposition 109.20 (Properties of the Wightman functions)**.** *The Wightman functions of a theory satisfying Axioms 109.9 to 109.15 satisfy:*

1. ***Temperedness.*** *Each W_n is a tempered distribution on $\mathbb{R}^{4n}$.*
2. ***Covariance.*** $W_n(\Lambda x_1 + a,\dots,\Lambda x_n + a) = W_n(x_1,\dots,x_n)$.
3. ***Spectral support.*** *W_n depends only on the differences $\xi_j = x_j - x_{j+1}$, and the Fourier transform of W_n in those differences is supported in $\overline{V}^+ \times\cdots\times \overline{V}^+$ ($n-1$ factors).*
4. ***Locality.*** *If x_j and x_{j+1} are spacelike separated, W_n is unchanged by exchanging them.*
5. ***Positivity.*** *For any finite sequence of test functions $f_0 \in \mathbb{C}$, $f_1 \in \mathcal{S}(\mathbb{R}^4)$, $f_2 \in \mathcal{S}(\mathbb{R}^8),\dots,$*

$$\sum_{m,n\geq 0} \int \mathrm{d}^{4m}x\ \mathrm{d}^{4n}y\ \overline{f_m(x_1,\dots,x_m)}\, W_{m+n}(x_m,\dots,x_1,y_1,\dots,y_n)\, f_n(y_1,\dots,y_n)\ \geq\ 0\,. \tag{109.14}$$

6. ***Hermiticity.*** $\overline{W_n(x_1,\dots,x_n)} = W_n(x_n,\dots,x_1)$.

Rests on Axioms 109.9 to 109.14.

Derivation. *Derives Proposition 109.20.*

(i) is Axiom 109.12 applied to the vector $\phi(f_2)\cdots\phi(f_n)\,|0\rangle$.

(ii) Insert $U^{-1}U$ between every pair of fields in Equation (109.13), use Equation (109.11) n times, and use $U\,|0\rangle = |0\rangle$ from Axiom 109.11.

(iii) Translation invariance is the $\Lambda = \mathbb{1}$ case of (ii) and gives the dependence on differences alone. For the support statement, write $\phi(x_j) = U(x_j)\phi(0)U(x_j)^{-1}$ with $U(a) = \exp(\mathrm{i}a\cdot P/\hbar)$ from Equation (109.10); then

$$W_n = \langle 0|\, \phi(0)\, \mathrm{e}^{-\mathrm{i}\xi_1\cdot P/\hbar}\, \phi(0) \cdots \mathrm{e}^{-\mathrm{i}\xi_{n-1}\cdot P/\hbar}\, \phi(0)\, |0\rangle \ , \tag{109.15}$$

using $P^\mu\,|0\rangle = 0$ at both ends. Each factor $\mathrm{e}^{-\mathrm{i}\xi\cdot P/\hbar}$ contributes, by the spectral theorem for the commuting family P^μ, a Fourier transform supported where the spectral measure lives, and Axiom 109.10 confines that to $\overline{V}^+$. Explicitly, with $\tilde{W}_n$ the transform in the variables ξ_j,

$$\tilde{W}_n(q_1, \dots, q_{n-1}) = 0 \quad \text{unless} \quad q_j \in \overline{V}^+ \ \forall j \, . \tag{109.16}$$

(iv) is Axiom 109.14 sandwiched between the vacuum and the remaining fields.

(v) The left-hand side of Equation (109.14) is $\|\Psi\|^2$ for $\Psi = \sum_n \int \mathrm{d}^{4n}y\, f_n(y)\phi(y_1)\cdots\phi(y_n)\,|0\rangle$, a vector in $\mathcal{H}$ by Axiom 109.12. A squared norm is non-negative; that is the whole derivation, and it is worth stressing that *positivity of the Wightman functions is the positivity of the Hilbert-space metric and nothing else.* The same statement reappears in Euclidean dress as reflection positivity in Section 109.5.1.

(vi) follows from $\phi(f)^\dagger \supset \phi(\bar{f})$ and $\overline{\langle 0|\, A\,|0\rangle} = \langle 0|\, A^\dagger\,|0\rangle$. □

The remarkable fact is that these six properties are not merely necessary but sufficient. Nothing is lost in passing from operators to distributions.

Theorem 109.21 (Wightman reconstruction)**.** *Let $\{W_n\}_{n\ge 0}$ be a family of tempered distributions satisfying (i)–(vi) of Proposition 109.20. Then there exist a separable Hilbert space $\mathcal{H}$, a continuous unitary representation $U(a,\Lambda)$ with spectrum in $\overline{V}^+$, a vector $|0\rangle$ and a field ϕ satisfying all the axioms, whose Wightman functions are the given W_n; and the construction is unique up to unitary equivalence [1099].*

Rests on Definition 109.19 and Proposition 109.20.

Derivation. *Derives Theorem 109.21.*

The construction is the same one that will reappear as the GNS construction in Section 109.4.2, and it is worth doing once by hand.

The algebra. Let $\mathfrak{B}$ be the set of terminating sequences $\mathsf{f} = (f_0, f_1, f_2, \dots)$ with $f_0 \in \mathbb{C}$ and $f_n \in \mathcal{S}(\mathbb{R}^{4n})$, made into an associative algebra by

$$(\mathsf{f}\,\mathsf{g})_n(x_1, \dots, x_n) = \sum_{j+k=n} f_j(x_1, \dots, x_j)\, g_k(x_{j+1}, \dots, x_n) \, , \tag{109.17}$$

with the involution $(\mathsf{f}^*)_n(x_1, \dots, x_n) = \overline{f_n(x_n, \dots, x_1)}$. This is the Borchers tensor algebra; the unit is $(1, 0, 0, \dots)$.

The linear functional. Define $w(\mathsf{f}) := \sum_n \int \mathrm{d}^{4n}x\, W_n(x) f_n(x)$, which converges because the sequence terminates. Property (vi) says $w(\mathsf{f}^*) = \overline{w(\mathsf{f})}$, and property (v) says exactly

$$w(\mathsf{f}^*\mathsf{f}) \ge 0 \qquad \text{for every } \mathsf{f} \in \mathfrak{B} \, . \tag{109.18}$$

The Hilbert space. Put $\langle \mathsf{f}, \mathsf{g}\rangle := w(\mathsf{f}^*\mathsf{g})$. This is a sesquilinear form, non-negative by Equation (109.18), hence obeying the Cauchy–Schwarz inequality $|\langle \mathsf{f}, \mathsf{g}\rangle|^2 \le \langle \mathsf{f}, \mathsf{f}\rangle\, \langle \mathsf{g}, \mathsf{g}\rangle$. Therefore $\mathfrak{N} := \{\mathsf{f} : w(\mathsf{f}^*\mathsf{f}) = 0\}$ is a linear subspace, and by Cauchy–Schwarz it is a left ideal:

product, and D the image of $\mathfrak{B}$, dense by construction.

The field. For $f \in \mathcal{S}(\mathbb{R}^4)$ let $\mathsf{f}^{(1)} = (0, f, 0, \dots)$ and define $\phi(f)[\mathsf{g}] := [\mathsf{f}^{(1)}\mathsf{g}]$. This is well defined because $\mathfrak{N}$ is a left ideal, maps D into D, is linear in f, and satisfies $\langle \mathsf{g}, \phi(f)\mathsf{h}\rangle = \big\langle \phi(\bar{f})\mathsf{g}, \mathsf{h}\big\rangle$ by the definition of the involution — which is Axiom 109.12. The

vacuum is $|0\rangle := [(1,0,\dots)]$, and cyclicity Axiom 109.15 holds because $\mathfrak{B}$ is generated by the $\mathsf{f}^{(1)}$.

The symmetry. Property (ii) says the form $\langle\cdot,\cdot\rangle$ is invariant under the action of (a,Λ) on test functions given by Equation (109.11). An invariant form induces an isometry of the quotient, and an isometry of a dense subspace extends to a unitary $U(a,\Lambda)$ on $\mathcal{H}$; the group law is inherited from the action on test functions. Continuity follows from continuity of that action in $\mathcal{S}$ together with temperedness (i), and Stone's theorem then supplies the self-adjoint generators P^μ of Equation (109.10). Property (iii) says the spectral measure of P^μ is supported in $\overline{V}^+$, which is Axiom 109.10; property (iv) gives Axiom 109.14.

Uniqueness. If two constructions $(\mathcal{H},\phi,|0\rangle)$ and $(\mathcal{H}',\phi',|0\rangle')$ have the same W_n, the map $\phi(f_1)\cdots\phi(f_n)\,|0\rangle \mapsto \phi'(f_1)\cdots\phi'(f_n)\,|0\rangle'$ preserves inner products by hypothesis, is defined on a dense set by cyclicity, and therefore extends to a unitary intertwining the two theories. □

Remark 109.22 (What the reconstruction theorem is worth)*.* It converts an existence problem about operators into an existence problem about distributions, which is the form every constructive attempt takes: build a family of W_n — or, in Section 109.5, a family of Euclidean S_n — verify six properties, and the theory follows. It also explains why Theorem 109.4 bites so hard: the Wightman functions are the theory, so two theories with the same correlators are the same theory, and no amount of reinterpretation can make one interacting and the other free.

Proposition 109.23 (Uniqueness of the vacuum equals clustering)**.** *Let the axioms hold except possibly Axiom 109.11, and let $|0\rangle$ be a translation-invariant unit vector. Then $|0\rangle$ is the* unique *such vector if and only if the Wightman functions cluster: for any spacelike a and any m,n,*

$$\lim_{\lambda\to\infty} W_{m+n}(x_1,\dots,x_m,y_1+\lambda a,\dots,y_n+\lambda a) = W_m(x_1,\dots,x_m)\,W_n(y_1,\dots,y_n)\,. \tag{109.19}$$

Rests on Axioms 109.9 and 109.10 and Theorem 109.26.

Derivation. *Derives Proposition 109.23.*

Write $A=\phi(x_1)\cdots\phi(x_m)$ and $B=\phi(y_1)\cdots\phi(y_n)$, so the left-hand side of Equation (109.19) is $\langle 0|\,A\,U(\lambda a)\,B\,|0\rangle$ with $U(\lambda a)=\exp(\mathrm{i}\lambda a\cdot P/\hbar)$. Let $E(\mathrm{d}^4p)$ be the joint spectral measure of P^μ and let E_0 be the projector onto the eigenvalue $p=0$. Then

$$\langle 0|\,A\,U(\lambda a)\,B\,|0\rangle = \langle 0|\,AE_0B\,|0\rangle + \int_{\overline{V}^+\setminus\{0\}} \mathrm{e}^{\mathrm{i}\lambda a\cdot p/\hbar}\;\langle 0|\,A\,E(\mathrm{d}^4p)\,B\,|0\rangle\;. \tag{109.20}$$

Two things must now be established: that the only atom of the spectral measure sits at $p=0$, and that the remaining integral vanishes in the limit. The first is a derivation and is where the physics is.

The only eigenvalue of P^μ is zero. Suppose $P^\mu|\chi\rangle = p^\mu|\chi\rangle$ with $p\neq 0$ and $p\in\overline{V}^+$. For every Λ in the proper orthochronous group the vector $U(0,\Lambda)\,|\chi\rangle$ is again an eigenvector, with eigenvalue Λp, and two eigenvectors of a self-adjoint family with different eigenvalues are orthogonal. The orbit $\{\Lambda p\}$ of a non-zero $p\in\overline{V}^+$ is an entire mass hyperboloid or half of the light cone — in either case an uncountable set — so $\mathcal{H}$ would contain an uncountable orthonormal family and could not be separable, contradicting Axiom 109.9. Hence the spectral measure has no atom away from the origin, and E_0 projects onto the translation-invariant vectors.

The oscillatory term vanishes. The remaining integral is the Fourier transform of a finite complex measure supported in $\overline{V}^+\setminus\{0\}$, evaluated along the spacelike ray λa. That it tends to zero is the content of the cluster theorem of [1297]: the measure has been shown to be non-atomic, so a Riemann–Lebesgue argument gives the vanishing of its transform in

the mean, and analyticity in the tube of Theorem 109.26 upgrades the mean convergence to a genuine limit, exponential with rate $\Delta/(\hbar c)$ when the spectrum has a gap Δ above the vacuum and power-law otherwise. That upgrade is the imported half of this proposition; the atom analysis above is the half that carries the physical content.

Granting it, if the vacuum is unique then $E_0 = |0\rangle\langle 0|$ and the surviving term is $W_m W_n$, which is Equation (109.19). Conversely, if Equation (109.19) holds then $\langle 0|\, A E_0 B\, |0\rangle = \langle 0|\, A\, |0\rangle\, \langle 0|\, B\, |0\rangle$ for all local A, B; since the vacuum is cyclic these exhaust a dense set, so $E_0 = |0\rangle\langle 0|$, which says precisely that the only translation-invariant vector is $|0\rangle$. □

Remark 109.24 (Why clustering is the physically meaningful form). Equation (109.19) says that two experiments performed far apart in space are statistically independent, which is a precondition for doing physics at all: a laboratory result must not depend on the contents of a distant galaxy. That this is *equivalent* to uniqueness of the vacuum is the reason a theory with several vacua — a spontaneously broken theory, Section 104.3 — must be decomposed into superselection sectors before it can be interpreted, each sector having one vacuum and clustering within itself. The algebraic version of that decomposition is Section 109.4.2. The convergence is not uniform in m and n and the rate depends on the mass gap: with a gap Δ the approach is exponential, with correlation length $\hbar c/\Delta$, and in a massless theory it is only power-law.

109.2.3 Analyticity and the Wightman–Hall domain

Everything sharp in this subject comes from one observation: the spectrum condition is a statement about the support of a Fourier transform, and a support condition makes a function analytic. This is where the theorems of Section 109.3 come from.

Lemma 109.25 (Positivity of the cone pairing). *If $q \in \overline{V}^+$ and $\eta \in V^+$, then $q \cdot \eta \geq 0$, with equality only for $q = 0$. The statement uses only the shape of the cones and not the scale of either vector, so it holds verbatim with q in the momentum cone and η in the coordinate cone V_x^+ of Notation 109.8 — which is how Theorem 109.26 applies it, the product $q \cdot \eta$ then carrying* $\mathrm{J}\,\mathrm{s} = [\hbar]$.

Rests on Notation 109.8.

Derivation. *Derives Lemma 109.25.*

Since $\eta \in V^+$ has $\eta \cdot \eta > 0$ and $\eta^0 > 0$, choose the frame in which $\boldsymbol{\eta} = 0$, so $\eta = (\eta^0, \mathbf{0})$ with $\eta^0 > 0$; this is possible by a proper orthochronous transformation, which preserves the pairing. Then $q \cdot \eta = q^0 \eta^0$, and $q \in \overline{V}^+$ gives $q^0 \geq |\boldsymbol{q}| \geq 0$. Hence $q \cdot \eta \geq 0$, and it vanishes only if $q^0 = 0$, which with $q^0 \geq |\boldsymbol{q}|$ forces $q = 0$. □

Theorem 109.26 (Analyticity in the forward tube). *Let W_n satisfy Equation (109.16). Define the* forward tube

$$\mathcal{T}_{n-1} := \left\{ (\zeta_1, \dots, \zeta_{n-1}) \in \mathbb{C}^{4(n-1)} :\ \zeta_j = \xi_j - \mathrm{i}\eta_j,\ \eta_j \in V_x^+ \right\}. \tag{109.21}$$

Then W_n is the boundary value, as $\eta_j \to 0$ from inside the cone, of a function $\mathcal{W}_n$ holomorphic in $\mathcal{T}_{n-1}$.

Rests on Equation (109.16) and Lemma 109.25.

Derivation. *Derives Theorem 109.26.*

Write the inverse transform of Equation (109.16),

$$\mathcal{W}_n(\zeta_1, \dots, \zeta_{n-1}) = \int \prod_j \frac{\mathrm{d}^4 q_j}{(2\pi\hbar)^4}\, \tilde{W}_n(q_1, \dots, q_{n-1}) \exp\left(-\frac{\mathrm{i}}{\hbar} \sum_j q_j \cdot \zeta_j \right). \tag{109.22}$$

With $\zeta_j = \xi_j - \mathrm{i}\eta_j$ the exponent is

$$-\frac{\mathrm{i}}{\hbar}\sum_j q_j \cdot \xi_j - \frac{1}{\hbar}\sum_j q_j \cdot \eta_j \,,$$

whose real part is $-\hbar^{-1}\sum_j q_j\cdot\eta_j$. By Lemma 109.25 each term is non-negative on the support of $\tilde{W}_n$ and grows linearly with $|q_j|$ when η_j is in the open cone, so the exponential decays exponentially in $|q_j|$. Since $\tilde{W}_n$ is a tempered distribution — of at most polynomial growth — the integral converges absolutely, and it converges together with all its ζ-derivatives; the integrand is holomorphic in each ζ_j in the region where the decay holds, so Equation (109.22) defines a holomorphic function there. Letting $\eta_j \to 0$ recovers W_n in the sense of distributions.

Two remarks on rigour. That a tempered distribution supported in a cone has a Fourier–Laplace transform holomorphic in the dual cone is the Paley–Wiener–Schwartz theorem, imported here and not proved; what Lemma 109.25 supplies is the identification of the dual cone, and that is the geometric content. The factors of $\hbar$ are forced: $q\cdot\zeta$ has dimension $\mathrm{kg\,m^2/s} = \mathrm{J\,s}$, so it is $q\cdot\zeta/\hbar$ that is a phase, and each four-dimensional momentum measure carries $(2\pi\hbar)^{-4}$ exactly as in the loop integrals of Chapter 100. □

The tube is a domain of holomorphy in the difference variables, but it contains no real points: setting $\eta_j = 0$ lands on the boundary. Lorentz covariance enlarges it until it does.

Definition 109.27 (The complex Lorentz group)**.** $L(\mathbb{C})$ is the set of complex 4×4 matrices Λ with $\Lambda^{\mathsf{T}}\eta\Lambda = \eta$, and $L_+(\mathbb{C})$ the subgroup with $\det\Lambda = +1$. The *extended tube* $\mathcal{T}'_{n-1}$ is the union of the images $\Lambda\mathcal{T}_{n-1}$ over $\Lambda \in L_+(\mathbb{C})$, acting simultaneously on all $n-1$ difference variables.

Theorem 109.28 (Bargmann–Hall–Wightman; imported)**.** *The holomorphic function $\mathcal{W}_n$ of Theorem 109.26 extends to a single-valued holomorphic function on the extended tube $\mathcal{T}'_{n-1}$, and satisfies there*

$$\mathcal{W}_n(\Lambda\zeta_1,\dots,\Lambda\zeta_{n-1}) = \mathcal{W}_n(\zeta_1,\dots,\zeta_{n-1})\,, \qquad \Lambda \in L_+(\mathbb{C})\,, \tag{109.23}$$

for a scalar field; for a field carrying a finite-dimensional representation S of $\mathrm{SL}(2,\mathbb{C})$ the right-hand side is multiplied by $S(\Lambda)$ analytically continued in the group parameters [1113].

Rests on Theorem 109.26, Definition 109.27, and Proposition 109.20.

Status. *Derives Theorem 109.28.*

Not proved here. The content is that real Lorentz covariance, Proposition 109.20(ii), plus holomorphy in the tube force covariance under the complexified group, and that the extension is single-valued — the delicate point, since $L_+(\mathbb{C})$ is not simply connected and the naive continuation could be multiple-valued. The proof is in [1113] and [1297, Ch. 2]. □

Derivation pending #182

The Bargmann–Hall–Wightman extension theorem: that the holomorphic Wightman function on the forward tube extends single-valuedly to the extended tube and is covariant under the complex Lorentz group; together with Jost's characterization of the real points of that extended tube as the totally spacelike configurations, stated below. These are the substantial analytic imports of the chapter and belong in Appendix A, together with the Paley–Wiener–Schwartz theorem and the edge-of-the-wedge theorem they use. Everything the chapter derives from them is derived explicitly in the text, and the non-emptiness of the set of real points is derived below.

The gain is that the extended tube contains real points, and they are exactly the totally spacelike ones.

Definition 109.29 (Jost points)**.** A real configuration $(\xi_1, \dots, \xi_{n-1})$ is a *Jost point* if every convex combination $\sum_j \lambda_j \xi_j$ with $\lambda_j \geq 0$ and $\sum_j \lambda_j > 0$ is spacelike, $\left(\sum_j \lambda_j \xi_j\right)^2 < 0$.

Theorem 109.30 (Jost; imported)**.** *The real points of the extended tube $\mathcal{T}'_{n-1}$ are exactly the Jost points [1115, 1297]. In particular the set of Jost points is open and non-empty.*

Rests on Definitions 109.27 and 109.29.

Proposition 109.31 (Jost points exist)**.** *For every $n \geq 2$ the set of Jost points in the difference variables is a non-empty open subset of $\mathbb{R}^{4(n-1)}$.*

Rests on Definition 109.29.

Derivation. *Derives Proposition 109.31.*

Take all ξ_j equal to a fixed purely spatial vector, $\xi_j = (0, \boldsymbol{d})$ with $\boldsymbol{d} \neq 0$. Then $\sum_j \lambda_j \xi_j = (0, (\sum \lambda_j)\boldsymbol{d})$ is spacelike whenever $\sum \lambda_j > 0$, so the configuration is a Jost point. Openness: the condition in Definition 109.29 need only be checked on the compact simplex $\sum \lambda_j = 1$, $\lambda_j \geq 0$, where $(\sum \lambda_j \xi_j)^2$ is a continuous function of (λ, ξ) attaining a strictly negative maximum; a small perturbation of ξ keeps it negative. □

Remark 109.32 (Why Jost points are the right real points). The Jost condition says more than that each ξ_j is spacelike: it says that the whole configuration is spacelike *simultaneously*, so that all n points $x_1, \dots, x_n$ lie at mutually spacelike separation and can be put on one spacelike hyperplane. That is exactly where Axiom 109.14 lets fields be permuted freely. Locality is thus an input *on the real boundary*, analyticity carries it into the complex domain, and the identity theorem carries the consequence back out to configurations that are not spacelike at all. Every theorem in Section 109.3 has this shape, and it is the reason a statement about the periodic table can rest on a statement about complex rotations.

109.2.4 The Reeh–Schlieder theorem

Cyclicity of the vacuum, Axiom 109.15, is stated for fields smeared over all of spacetime. The following theorem says that an arbitrarily small open region suffices, which is far from obvious and has consequences out of all proportion to the length of its proof.

Definition 109.33 (Local polynomial algebra)**.** For an open region $\mathcal{O} \subset \mathbb{R}^4$, let $\mathcal{P}(\mathcal{O})$ be the algebra of polynomials in the smeared fields $\phi(f)$ with $\operatorname{supp} f \subset \mathcal{O}$.

Theorem 109.34 (Reeh–Schlieder)**.** *Let $\mathcal{O}$ be any non-empty open region of spacetime. Then*

1. *$\mathcal{P}(\mathcal{O})\,|0\rangle$ is dense in $\mathcal{H}$: the vacuum is* cyclic *for every local algebra;*
2. *if $\mathcal{O}$ has a non-empty spacelike complement and $A \in \mathcal{P}(\mathcal{O})$ satisfies $A\,|0\rangle = 0$, then $A = 0$: the vacuum is* separating *[1218].*

Rests on Theorem 109.26 and Axioms 109.14 and 109.15.

Derivation. *Derives Theorem 109.34.*

(i) Choose an open $\mathcal{O}_0$ whose closure lies inside $\mathcal{O}$, and $\epsilon > 0$ so small that translating $\mathcal{O}_0$ by any a with $|a^\mu| < \epsilon$ keeps it inside $\mathcal{O}$. Suppose $\Psi \in \mathcal{H}$ is orthogonal to $\mathcal{P}(\mathcal{O})\,|0\rangle$. Fix n and test functions $f_1, \dots, f_n$ supported in $\mathcal{O}_0$ and consider

$$F(a_1, \dots, a_n) := \langle \Psi |\, \phi(f_1^{a_1}) \cdots \phi(f_n^{a_n})\, |0\rangle\,, \qquad f^a(x) := f(x - a)\,. \tag{109.24}$$

For all real a_j with $\left|a_j^\mu\right| < \epsilon$ every translated support is still in $\mathcal{O}$, so F vanishes on that open real neighbourhood of the origin by hypothesis.

Now repeat the argument of Theorem 109.26 on F. Writing $\phi(f^a) = U(a)\phi(f)U(a)^{-1}$ and using $P^\mu|0\rangle = 0$ at the right end, the exponentials collapse pairwise and Equation (109.24) becomes

$$F = \langle\Psi|\, U(a_1)\,\phi(f_1)\,\mathrm{e}^{-\mathrm{i}b_1\cdot P/\hbar}\,\phi(f_2)\cdots\mathrm{e}^{-\mathrm{i}b_{n-1}\cdot P/\hbar}\,\phi(f_n)\,|0\rangle\ ,$$

in the differences $b_j = a_j - a_{j+1}$. One step more is needed than in Equation (109.15), which used $P^\mu|0\rangle = 0$ at *both* ends: Ψ is not translation invariant, so a residual $U(a_1)$ survives on the left. Absorb it, holding a_1 fixed and replacing Ψ by $\Psi_{a_1} := U(a_1)^{-1}\Psi$, which is again orthogonal to $\mathcal{P}(\mathcal{O})|0\rangle$ for the correspondingly translated region and is a legitimate state. What remains is of exactly the form of Equation (109.15) with $\langle\Psi_{a_1}|$ in place of $\langle 0|$ at the left end. The spectrum condition applies unchanged to the right of each field, so for each fixed a_1, F is the boundary value of a function holomorphic in the forward tube in the variables b_j.

A function holomorphic in a tube whose distributional boundary value vanishes on a non-empty open set of the real boundary vanishes identically. (This is the edge-of-the-wedge theorem, imported; in the one-variable case it is the Schwarz reflection principle, and it is the statement that a real open set is a set of uniqueness for such functions.) Hence $F \equiv 0$ for all real a_j, so Ψ is orthogonal to $\phi(g_1)\cdots\phi(g_n)|0\rangle$ for all test functions g_j whatever — translates of a fixed f span a dense set of $\mathcal{S}$ in the relevant sense, and n was arbitrary. By Axiom 109.15, $\Psi = 0$.

(ii) Let $\mathcal{O}'$ be an open region spacelike to $\mathcal{O}$, and let $B \in \mathcal{P}(\mathcal{O}')$. Since $A \in \mathcal{P}(\mathcal{O})$ and B commute by Axiom 109.14,

$$A\,B\,|0\rangle = B\,A\,|0\rangle = 0\,.$$

By part (i) applied to $\mathcal{O}'$, the vectors $B|0\rangle$ are dense in $\mathcal{H}$; an operator vanishing on a dense subset of its domain is zero. □

Corollary 109.35 (No local operator annihilates the vacuum)**.** *There is no non-zero local operator A with $A|0\rangle = 0$. In particular there is no local number operator, no local projector onto the vacuum, and no local operator that creates "nothing".*

Rests on Theorem 109.34.

Derivation. *Derives Corollary 109.35.*

Part (ii) of Theorem 109.34 with $\mathcal{O}$ any bounded region, whose spacelike complement is never empty. A local projector E with $E|0\rangle = 0$ would violate it, and so would a number operator, since a number operator annihilates the vacuum by definition. □

Remark 109.36 (Vacuum entanglement, and what it does not permit). Part (i) has a startling reading. Acting on the vacuum with operators confined to a laboratory, one can approximate *any* state of the whole universe — a state containing a hundred protons on the far side of the Galaxy — to arbitrary accuracy in norm. The reason is that the vacuum is not a product state across any spatial division: it is entangled between every region and its complement, and part (i) is one expression of how strongly. The entanglement entropy of the vacuum restricted to a region is in fact ultraviolet divergent, growing with the area of the boundary; that is the field-theoretic case flagged in Chapter 84, and it is what Section 109.4.3 organizes.

It does *not* permit signalling. Let A act in $\mathcal{O}$ and let B be any observable in a spacelike-separated region $\mathcal{O}'$. An operation performed in $\mathcal{O}$ takes a state ρ to $\sum_k A_k \rho A_k^\dagger$ with $\sum_k A_k^\dagger A_k = \mathbb{1}$ and every $A_k \in \mathcal{P}(\mathcal{O})$. The expectation of B afterwards is

$$\sum_k \mathrm{tr}\Big(A_k \rho A_k^\dagger B\Big) = \sum_k \mathrm{tr}\Big(\rho A_k^\dagger B A_k\Big) = \mathrm{tr}\left(\rho\, B \sum_k A_k^\dagger A_k\right) = \mathrm{tr}(\rho B)\,,$$

the middle step being cyclicity of the trace and the next Axiom 109.14, which lets B move through $A_k^\dagger$ and A_k. Nothing measurable in $\mathcal{O}'$ changes. The approximation in part (i) is

achieved by operators of enormous norm acting with vanishing probability; it is a statement about the closure of a subspace, not about what an experimenter can bring about. The distinction is the field-theoretic form of Phenomenon 84.4, whose derivation from trace preservation is the same computation in a smaller setting.

109.2.5 Borchers classes

Which field a theory is written in terms of turns out not to be part of its physical content.

Definition 109.37 (Relative locality)**.** Two fields ϕ and ψ on the same Hilbert space are *relatively local* if $[\phi(f), \psi(g)] = 0$ (or the anticommutator vanishes, for two half-integer-spin fields) whenever the supports of f and g are spacelike separated. The *Borchers class* of ϕ is the set of all fields relatively local to it.

Theorem 109.38 (Borchers)**.** *Two relatively local fields, each satisfying the Wightman axioms with the same vacuum and the same representation of the Poincaré group, have the same S matrix [1177].*

Rests on Definition 109.37 and Theorem 109.59.

Status and mechanism. Derives Theorem 109.38.

The full proof requires the asymptotic construction of Section 109.3.4 and is not carried out here. The mechanism is that the asymptotic fields are built from a field only through its matrix elements between the vacuum and the one-particle states, together with its commutation properties; two relatively local fields have the same one-particle overlaps up to normalization and the same commutation properties, so they generate the same in and out states, and the S matrix is the map between those. □

Derivation pending #183

Borchers' theorem: that two relatively local Wightman fields generate the same asymptotic states and hence the same S matrix. The argument runs through the Haag–Ruelle construction and belongs in Appendix A next to it. The consequence used in this book, that a field redefinition leaves scattering amplitudes unchanged, is stated with its mechanism in the text.

Corollary 109.39 (Interpolating fields are conventions)**.** *A local field is not an observable of the theory. Any local operator with non-vanishing overlap with a one-particle state may be used to compute that particle's scattering amplitudes, and the result is independent of the choice.*

Rests on Theorem 109.38.

This is assumed without comment wherever a field redefinition is made: in the counterterm rearrangements of Chapter 100, where $\phi \to Z^{1/2}\phi$ is treated as harmless; in the scattering formalism of Chapter 86; and in Corollary 106.49, whose statement that lattice practitioners may create a proton with any convenient operator is Theorem 109.38 in another dress. The independence is not a convenience: it is the reason the S matrix is well defined at all, since the fields themselves are not measurable and only the asymptotic particle states are.

109.3 Theorems from the axioms

109.3.1 The PCT theorem

Chapter 105 proves the PCT theorem at the level of a Lagrangian, by counting Lorentz indices, and records the measurements that test it. That proof assumes a Lagrangian, a

field content and a perturbative expansion, and the theorem assumes none of them. This subsection supplies the theorem proper.

Definition 109.40 (The *PCT* condition)**.** A family of Wightman functions of a Hermitian scalar field satisfies the *PCT condition* if for every n

$$W_n(x_1,\dots,x_n) = W_n(-x_n,\dots,-x_1)\,. \tag{109.25}$$

For fields carrying Lorentz indices the right-hand side is multiplied by $(-1)^J$, J being the total number of undotted spinor indices plus the number of vector indices — the same index count that Lemma 105.35 performs on a Lagrangian monomial.

Definition 109.41 (Weak local commutativity)**.** The theory satisfies *weak local commutativity* (WLC) if for every n and every Jost point of the difference variables

$$W_n(x_1,\dots,x_n) = W_n(x_n,\dots,x_1)\,, \tag{109.26}$$

again with $(-1)^J$ for indexed fields.

WLC is strictly weaker than Axiom 109.14: locality lets *any* two spacelike-separated arguments be exchanged and hence implies Equation (109.26) by repeated transposition, while Equation (109.26) demands only that the total reversal be a symmetry, and only at totally spacelike configurations.

The decisive input is a fact about a group, and it can be verified by writing down a path.

Lemma 109.42 (Total inversion is a complex Lorentz transformation)**.** *$L_+(\mathbb{C})$ is connected and contains $-\mathbb{1}$, the total inversion $x \mapsto -x$. By contrast, over the reals $-\mathbb{1}$ lies in the component $L_+^{\downarrow}$ and is not connected to the identity.*

Rests on Definitions 39.5 and 109.27.

Derivation. *Derives Lemma 109.42.*

Consider the one-parameter family, written in the block form $(x^0,x^1)\oplus(x^2,x^3)$,

$$\Lambda(s) = \begin{pmatrix} \cos(\pi s) & \mathrm{i}\sin(\pi s) & 0 & 0 \\ \mathrm{i}\sin(\pi s) & \cos(\pi s) & 0 & 0 \\ 0 & 0 & \cos(\pi s) & -\sin(\pi s) \\ 0 & 0 & \sin(\pi s) & \cos(\pi s) \end{pmatrix}, \qquad s\in[0,1]\,. \tag{109.27}$$

The upper block is a boost with rapidity $\chi = \mathrm{i}\pi s$, since $\cosh(\mathrm{i}\pi s) = \cos(\pi s)$ and $\sinh(\mathrm{i}\pi s) = \mathrm{i}\sin(\pi s)$; it is therefore *symmetric*, with both off-diagonal entries equal to $\mathrm{i}\sin(\pi s)$, as a boost matrix always is. The lower block is a rotation by πs in the (x^2,x^3) plane. Each block separately preserves the corresponding part of $\eta_{\mu\nu} = \mathrm{diag}(+1,-1,-1,-1)$ — the first because a boost does for any complex rapidity, the second because a rotation does — so $\Lambda(s)^{\mathsf{T}}\eta\Lambda(s) = \eta$ for every s, and $\det\Lambda(s) = 1$ because each block has unit determinant: $\cos^2(\pi s) - (\mathrm{i}\sin(\pi s))^2 = \cos^2(\pi s) + \sin^2(\pi s) = 1$ for the boost, and likewise for the rotation. Hence $\Lambda(s)\in L_+(\mathbb{C})$ throughout. At the endpoints

$$\Lambda(0) = \mathbb{1}\,, \qquad \Lambda(1) = \mathrm{diag}(-1,-1,-1,-1) = -\mathbb{1}\,,$$

so the path is continuous, lies in $L_+(\mathbb{C})$, and joins the identity to the total inversion.

Over the reals the same construction is unavailable: a real boost has $\cosh\chi \geq 1$, so the 00 entry of a real proper orthochronous transformation never reaches -1. Indeed $-\mathbb{1}$ has ${\Lambda^0}_0 = -1$, so it fails the condition ${\Lambda^0}_0 \geq 1$ that Definition 39.5 imposes and lies on the non-orthochronous branch; $\det(-\mathbb{1}) = +1$ in four dimensions, so $-\mathbb{1}\in L_+^{\downarrow}$. □

Corollary 109.43 (Total inversion is automatic)**.** *For a Hermitian scalar field the analytic Wightman function satisfies* $\mathcal{W}_n(-\zeta_1,\dots,-\zeta_{n-1}) = \mathcal{W}_n(\zeta_1,\dots,\zeta_{n-1})$ *everywhere on the extended tube, and hence*

$$W_n(x_1,\dots,x_n) = W_n(-x_1,\dots,-x_n) \tag{109.28}$$

at every Jost point. For a field carrying indices the right-hand side carries the factor $(-1)^J$*.*
Rests on Theorems 109.28 and 109.30 and Lemma 109.42.

Derivation. *Derives Corollary 109.43.*

Apply Equation (109.23) with $\Lambda = -\mathbb{1}$, which lies in $L_+(\mathbb{C})$ by Lemma 109.42; the extended tube is $L_+(\mathbb{C})$-invariant by Definition 109.27, so both sides are defined. Restricting to the real points, which are the Jost points by Theorem 109.30, gives Equation (109.28). □

Equation (109.28) is *not* the *PCT* condition: it reverses the signs of the arguments but leaves their order alone, whereas Equation (109.25) reverses both. The content of Jost's theorem is exactly the difference between the two.

Theorem 109.44 (Jost)**.** *A Wightman theory satisfies the PCT condition Equation (109.25) if and only if it satisfies weak local commutativity Equation (109.26) [1115].*
Rests on Definitions 109.40 and 109.41 and Corollary 109.43.

Derivation. *Derives Theorem 109.44.*

Take the scalar case; the index bookkeeping in the general case multiplies both statements by the same $(-1)^J$ and cancels.

($\Leftarrow$) Assume WLC. At a Jost point, combine Equation (109.26) with Equation (109.28):

$$W_n(-x_n,\dots,-x_1) \overset{\text{Equation (109.28)}}{=} W_n(x_n,\dots,x_1) \overset{\text{Equation (109.26)}}{=} W_n(x_1,\dots,x_n)\,,$$

the first step being legitimate because the reversed configuration $(x_n,\dots,x_1)$ is a Jost point whenever $(x_1,\dots,x_n)$ is — the Jost condition of Definition 109.29 is invariant under reversal, since reversing the order replaces each difference ξ_j by $-\xi_{n-j}$ and the defining condition is even. So Equation (109.25) holds on the open set of Jost points. Both sides of Equation (109.25) are boundary values of functions holomorphic on the extended tube, which is connected, and they agree on a real open subset of it; by the identity theorem for holomorphic functions of several variables they agree everywhere, so Equation (109.25) holds as an identity of distributions.

($\Rightarrow$) Assume the *PCT* condition. Reading the same chain backwards, at a Jost point $W_n(x_1,\dots,x_n) = W_n(-x_n,\dots,-x_1) = W_n(x_n,\dots,x_1)$, which is Equation (109.26). □

Corollary 109.45 (Locality implies *PCT*)**.** *Every theory satisfying the Wightman axioms satisfies the PCT condition.*
Rests on Axiom 109.14 and Theorem 109.44.

Derivation. *Derives Corollary 109.45.*

At a Jost point all n arguments are mutually spacelike, so Axiom 109.14 permits an arbitrary permutation of the fields inside Equation (109.13), in particular the total reversal. That is WLC, and Theorem 109.44 applies. □

The last step is to produce the operator.

Theorem 109.46 (Existence of the *PCT* operator)**.** *In a theory satisfying the PCT condition there is an antiunitary operator* Θ *on* $\mathcal{H}$ *with*

$$\Theta\,|0\rangle = |0\rangle\,, \qquad \Theta\,\phi(f)\,\Theta^{-1} = \phi\big(\bar{f}^{-}\big)\,, \qquad f^{-}(x) := f(-x)\,, \tag{109.29}$$

and $\Theta^2 = \mathbb{1}$ *for a Hermitian scalar field. In particular* $[\Theta, P^0] = 0$*, so PCT is an exact symmetry of the dynamics.*
Rests on Definition 109.40 and Axioms 109.9 and 109.13.

Derivation. *Derives Theorem 109.46.*

Define Θ on the dense set of vectors $\Psi = \phi(f_1)\cdots\phi(f_m)\,|0\rangle$ by

$$\Theta\,\phi(f_1)\cdots\phi(f_m)\,|0\rangle := \phi\big(\bar{f}_1^{-}\big)\cdots\phi\big(\bar{f}_m^{-}\big)\,|0\rangle\ , \tag{109.30}$$

extended antilinearly. Two things must be checked: that this preserves inner products in the antilinear sense, and that it is therefore well defined.

Let $\Phi = \phi(g_1)\cdots\phi(g_n)\,|0\rangle$. Then

$$\begin{aligned}\langle\Theta\Psi,\Theta\Phi\rangle &= \langle 0|\,\phi(f_m^{-})\cdots\phi\big(f_1^{-}\big)\,\phi\big(\bar{g}_1^{-}\big)\cdots\phi(\bar{g}_n^{-})\,|0\rangle \\ &= \int\prod \mathrm{d}^4x\,\mathrm{d}^4y\; W_{m+n}(-x_m,\dots,-x_1,-y_1,\dots,-y_n)\prod f_j(x_j)\prod\overline{g_k(y_k)}\end{aligned} \tag{109.31}$$

after the substitutions $x\to -x$, $y\to -y$, whose Jacobian is $+1$. Applying Equation (109.25) to the $(m+n)$-point function turns its argument list into $(y_n,\dots,y_1,x_1,\dots,x_m)$, and the result is exactly

$$\langle 0|\,\phi(\bar{g}_n)\cdots\phi(\bar{g}_1)\phi(f_1)\cdots\phi(f_m)\,|0\rangle = \langle\Phi,\Psi\rangle = \overline{\langle\Psi,\Phi\rangle}\,,$$

which is antiunitarity. Well-definedness follows: if a linear combination of the Ψ has zero norm, so does its image, because norms are preserved. A norm-preserving antilinear map on a dense domain extends to an antiunitary on $\mathcal{H}$.

The relation $\Theta\phi(f)\Theta^{-1} = \phi(\bar{f}^{-})$ is Equation (109.30) read on one factor, and $\Theta\,|0\rangle = |0\rangle$ is the case $m=0$. Applying Equation (109.30) twice returns f, so $\Theta^2 = \mathbb{1}$. Finally, from Equation (109.29) and Equation (109.11), $\Theta U(a,\mathbb{1})\Theta^{-1} = U(-a,\mathbb{1})$; writing both sides with Equation (109.10) and using antilinearity of Θ, which conjugates the i, gives $\Theta P^{\mu}\Theta^{-1} = P^{\mu}$, so Θ commutes with the energy. □

Remark 109.47 (Why PCT and not CP or T). The whole theorem rests on Lemma 109.42: the transformation that the complexified Lorentz group can reach is $-\mathbb{1}$ and nothing else. Space inversion $\Lambda_P = \mathrm{diag}(1,-1,-1,-1)$ and time reversal $\Lambda_T = \mathrm{diag}(-1,1,1,1)$ each have $\det = -1$ and therefore lie outside $L_+(\mathbb{C})$ altogether — not in a different component of it, but not in it at all, since $L_+(\mathbb{C})$ is defined by $\det\Lambda = +1$. No analytic continuation reaches them, and no sharpening of the hypotheses will produce a theorem asserting P, T or CP invariance. That is not a limitation of the proof: P, C and CP are all violated in Nature (Chapters 110 and 111), and a correct theorem cannot prove a false statement.

Remark 109.48 (What the theorem says about experiment). The observable consequences — equal masses, equal total lifetimes and opposite magnetic moments for a particle and its antiparticle — are derived from antiunitarity in Theorems 105.36 to 105.38 and measured in Section 105.5.3, where the sharpest entries are the neutral kaon rest-energy comparison at the level of 6×10^{-19} of m_Kc^2 [2427] and the antihydrogen $1S$–$2S$ frequency at 2×10^{-12}. The point worth carrying back here is that these are tests of Axioms 109.10 and 109.14 and Axiom 109.13 directly, since Corollary 109.45 used nothing else. A confirmed violation would not be a new particle; it would mean that one of the three axioms is false, and with it every theorem of this chapter.

109.3.2 Spin and statistics

This is the standing debt of the book. The connection between spin and statistics is assumed in Chapters 83 and 120, used to force the colour degree of freedom in Section 102.2, used again for the shell model of Chapter 107, and proved in Chapter 99 only for free fields of

spin 0 and spin 1/2 (Theorem 99.46). Here it is proved from the axioms, for any spin, at $3+1$.

Pauli's original argument [854] was that the wrong pairing is catastrophic: a half-integer-spin field quantized with commutators has an energy unbounded below, and an integer-spin field quantized with anticommutators loses microcausality. Those are the two computations carried out in Propositions 99.47 and 99.48 — the second showing that the anticommutator of the real scalar fails to vanish outside the light cone, and that the theory collapses besides, every even local polynomial in ϕ becoming a c-number. The axiomatic statement, due to Lüders and Zumino [1151] and to Burgoyne [1139], is sharper and stranger: a Wightman field with the wrong connection does not merely misbehave, it *vanishes*.

Theorem 109.49 (Spin and statistics)**.** *Let ϕ satisfy Axiom 109.9 to Axiom 109.15 except that Axiom 109.14 is imposed with the wrong bracket: the anticommutator $\{\phi(x),\phi^\dagger(y)\}$ vanishes at spacelike separation for a field of integer spin, or the commutator $[\phi(x),\phi^\dagger(y)]$ vanishes at spacelike separation for a field of half-integer spin. Then $\phi(f)=0$ for every test function f.*

Rests on Theorems 109.28 and 109.34 and Corollary 109.43.

Derivation, integer spin. Derives Theorem 109.49.

Take ϕ a complex scalar field and define the two two-point functions of the difference $\xi = x-y$,

$$F(\xi) := \langle 0|\,\phi(x)\phi^\dagger(y)\,|0\rangle\ , \qquad G(\xi) := \langle 0|\,\phi^\dagger(x)\phi(y)\,|0\rangle\ . \tag{109.32}$$

Note that in *both* the argument x stands on the left, so both are covered by Theorem 109.26 with the same tube: each is the boundary value of a function $\mathcal{F}$, $\mathcal{G}$ holomorphic on the forward tube $\mathcal{T}_1$ and, by Theorem 109.28, on the extended tube $\mathcal{T}_1'$.

Step 1: evenness. By Corollary 109.43 applied to each of $\mathcal{F}$ and $\mathcal{G}$ — these are two-point functions of a scalar field, so $J=0$ —

$$\mathcal{F}(-\zeta)=\mathcal{F}(\zeta)\,, \qquad \mathcal{G}(-\zeta)=\mathcal{G}(\zeta)\,, \qquad \zeta\in\mathcal{T}_1'\,. \tag{109.33}$$

Step 2: the wrong bracket at a Jost point. For ξ spacelike — which for $n=2$ is exactly the Jost condition — the assumed anticommutation gives $\phi(x)\phi^\dagger(y)=-\phi^\dagger(y)\phi(x)$, so

$$F(\xi) = -\,\langle 0|\,\phi^\dagger(y)\phi(x)\,|0\rangle = -G(-\xi) \overset{\text{Equation (109.33)}}{=} -G(\xi)\,, \tag{109.34}$$

the middle equality because G was defined with its arguments in the order (dagger, no dagger) and here they appear at y and x, i.e. with difference $y-x=-\xi$.

Step 3: analytic propagation. $\mathcal{F}+\mathcal{G}$ is holomorphic on the connected domain $\mathcal{T}_1'$ and, by Step 2, vanishes on the open set of real Jost points contained in it (Theorem 109.30 and Proposition 109.31). A holomorphic function vanishing on a real open subset of its domain vanishes identically, so $\mathcal{F}=-\mathcal{G}$ on all of $\mathcal{T}_1'$, and taking boundary values from the same tube,

$$F(\xi) = -G(\xi) \qquad \text{as distributions, for all } \xi\,. \tag{109.35}$$

Step 4: positivity kills it. Smear Equation (109.35) with $\overline{g(x)}\,g(y)$ for an arbitrary $g\in\mathcal{S}(\mathbb{R}^4)$. The left side gives $\langle 0|\,\phi(\bar g)\phi^\dagger(g)\,|0\rangle = \left\|\phi^\dagger(g)\,|0\rangle\right\|^2$, since $\phi(\bar g)=\left(\phi^\dagger(g)\right)^\dagger$; the right side gives $-\langle 0|\,\phi^\dagger(\bar g)\phi(g)\,|0\rangle = -\left\|\phi(g)\,|0\rangle\right\|^2$. Hence

$$\left\|\phi^\dagger(g)\,|0\rangle\right\|^2 + \left\|\phi(g)\,|0\rangle\right\|^2 = 0\,, \tag{109.36}$$

a sum of two non-negative numbers, so both vanish: $\phi(g)\,|0\rangle=0$ for every g.

Step 5: from the vacuum to the operator. Take g supported in a bounded open region, whose spacelike complement is non-empty. By Theorem 109.34(ii) the vacuum is separating for that region, so $\phi(g)=0$. Such g are dense in $\mathcal{S}$, and ϕ is continuous in g by Axiom 109.12, so $\phi=0$. □

Derivation, half-integer spin. *Derives Theorem 109.49.*

The only change is in Step 1. A field in the (A, B) representation of $\mathrm{SL}(2,\mathbb{C})$ carries $2A$ undotted and $2B$ dotted spinor indices, and Theorem 109.28 for such a field multiplies the right-hand side of Equation (109.23) by the analytic continuation of $S(\Lambda)$. Along the path Equation (109.27) that continuation returns, at $s = 1$, the factor $(-1)^{2A+2B} = (-1)^{2s_\phi}$ on the two-point function, where s_ϕ is the spin. For half-integer spin this is -1, so Equation (109.33) becomes $\mathcal{G}(-\zeta) = -\mathcal{G}(\zeta)$. The assumed *commutator* at spacelike separation gives, in place of Equation (109.34), $F(\xi) = +G(-\xi) = -G(\xi)$ — the same relation as before. Steps 3 to 5 are unchanged, and again $\phi = 0$. □

Corollary 109.50 (The connection between spin and statistics)**.** *A non-trivial Wightman field of integer spin commutes at spacelike separation and a non-trivial Wightman field of half-integer spin anticommutes there. Consequently the states of identical integer-spin quanta are symmetric under exchange and those of identical half-integer-spin quanta are antisymmetric.*

Rests on Theorem 109.49.

Derivation. *Derives Corollary 109.50.*

The first sentence is the contrapositive of Theorem 109.49. For the second, let $a^\dagger$ denote the creation part of the field in the sense of Section 109.3.4; the exchange symmetry of a two-quantum state $a^\dagger(f)a^\dagger(g)\,|0\rangle$ is the sign in the bracket relating $a^\dagger(f)$ and $a^\dagger(g)$, which is inherited from the field's spacelike bracket by the asymptotic construction. That is the symmetrization postulate assumed at Section 83.1.4, which is a postulate there because none of the hypotheses used here — Lorentz covariance, positivity of the energy, locality — is available in nonrelativistic quantum mechanics. □

Derivation pending #184

The analytic continuation of the finite-dimensional representation matrix $S([\Lambda])$ along the path joining the identity to the total inversion in the complex Lorentz group, which is what supplies the factor $(-1)^{2s}$ used in the half-integer case. The path itself and the scalar case are derived in full in the text; the index bookkeeping for arbitrary spin follows Burgoyne and belongs in Appendix A with the Bargmann–Hall–Wightman theorem it rests on.

Remark 109.51 (What has and has not been proved)*.* Four points, because this theorem is quoted loosely more often than any other in the subject.

It is a statement about which hypotheses are compatible. Nothing forbids constructing anticommuting scalars; the Faddeev–Popov ghosts of Chapters 102 and 106 are exactly that. They evade Theorem 109.49 by failing its hypotheses — they are not observables, they carry indefinite metric, and Axiom 109.9 with its positive-definite Hilbert space does not hold in the enlarged space where they live. The same qualification appears in Remark 99.49.

It is a $3+1$ theorem. Lemma 109.42 used four dimensions twice: the boost–rotation pairing needs two commuting planes, and $\det(-\mathbb{1}) = +1$ needs an even dimension. In $2+1$ dimensions neither holds, and the conclusion is false — anyons exist. This treatise does not treat $2+1$ physics (editorial rule on dimensions), and the observation is recorded only because it shows the theorem is not a formal triviality.

It needs a positive-definite metric. Step 4 is the whole proof, and it is nothing but the non-negativity of a norm. Every known counterexample is a theory in which that fails.

The consequences are measured. The exclusion principle is bounded directly: the Ramberg–Snow experiment and its successors limit the probability of a Pauli-violating atomic transition to below 1.7×10^{-26} (Section 99.7.2). That a statement about complex Lorentz transformations underwrites the periodic table, the chemistry of Chapter 88 and

the degeneracy pressure of Chapter 49 is one of the stranger facts in physics, and it is not a coincidence: both follow from locality and positivity of the energy.

109.3.3 The Källén–Lehmann representation

Every renormalized propagator in Chapters 100 and 106 rests on a decomposition of the exact two-point function that is usually quoted and rarely derived. It follows from the axioms in half a page, is exact rather than perturbative, and its positivity is again nothing but the positivity of the Hilbert-space metric.

Notation 109.52 (Spectral variable and state normalization). The spectral variable s is a *squared momentum*, of dimension $\mathrm{kg}^2\,\mathrm{m}^2/\mathrm{s}^2$, matching the mass-shell condition $p^2=m^2c^2$ of Equation (100.1). States of sharp four-momentum are labelled by the wave vector $\boldsymbol{k}=\boldsymbol{p}/\hbar$ and normalized covariantly,

$$\langle \lambda_{\boldsymbol{k}} | \lambda_{\boldsymbol{k}'} \rangle = (2\pi)^3 \, \frac{2\omega_{\boldsymbol{k}}}{c} \, \delta^3(\boldsymbol{k}-\boldsymbol{k}') \,, \tag{109.37}$$

so that the resolution of the identity on the sector λ is $\int \mathrm{d}^3k\, c\, |\lambda_{\boldsymbol{k}}\rangle\langle\lambda_{\boldsymbol{k}}| \,/\, \big[(2\pi)^3 2\omega_{\boldsymbol{k}}\big]$, in which the measure $\mathrm{d}^3k\,c/(2\omega_{\boldsymbol{k}})$ is Lorentz invariant.

Definition 109.53 (Free positive-frequency function)**.** For a squared momentum $s\geq 0$,

$$\Delta^+(\xi;s) := \hbar c^2 \int \frac{\mathrm{d}^3k}{(2\pi)^3\, 2\omega_{\boldsymbol{k}}} \, \mathrm{e}^{-\mathrm{i}k\cdot\xi} \,, \qquad \omega_{\boldsymbol{k}} = c\sqrt{\boldsymbol{k}^2 + \frac{s}{\hbar^2}} \,, \tag{109.38}$$

which is Equation (109.1) for a free field with $s=m^2c^2$ and carries $[\phi]^2=\mathrm{J/m}$.

Theorem 109.54 (Källén–Lehmann)**.** *Let ϕ be a Hermitian scalar Wightman field with $\langle 0|\,\phi(x)\,|0\rangle = 0$. Then there is a positive measure $\rho(s)\mathrm{d}s$ on $[0,\infty)$, of dimension $\mathrm{s}^2/(\mathrm{kg}^2\,\mathrm{m}^2)$, such that*

$$\langle 0|\,\phi(x)\phi(y)\,|0\rangle = \int_0^\infty \mathrm{d}s\; \rho(s)\, \Delta^+(x-y;s) \,, \qquad \rho(s)\geq 0 \,. \tag{109.39}$$

If in addition ϕ obeys the canonical equal-time commutation relation with $\pi = c^{-2}\partial_t\phi$, then

$$\int_0^\infty \mathrm{d}s\; \rho(s) = 1 \,. \tag{109.40}$$

Rests on Axioms 109.10 and 109.13 and Definition 109.53.

Derivation. *Derives Theorem 109.54.*

Insert a complete set of states between the fields. By Axiom 109.10 every state has four-momentum in $\overline{V}^+$; group the states by their sector label λ and their invariant mass m_λ, and use Notation 109.52. The vacuum contributes $|\langle 0|\,\phi(0)\,|0\rangle|^2=0$ by hypothesis, so

$$\langle 0|\,\phi(x)\phi(y)\,|0\rangle = \sum_{\lambda} \int \frac{\mathrm{d}^3k}{(2\pi)^3}\, \frac{c}{2\omega_{\boldsymbol{k}}} \; \langle 0|\,\phi(x)\,|\lambda_{\boldsymbol{k}}\rangle\, \langle \lambda_{\boldsymbol{k}}|\,\phi(y)\,|0\rangle \;. \tag{109.41}$$

Translation covariance, Equation (109.10) with $\phi(x)=U(x)\phi(0)U(x)^{-1}$ and $P^\mu|0\rangle=0$, gives $\langle 0|\,\phi(x)\,|\lambda_{\boldsymbol{k}}\rangle = \mathrm{e}^{-\mathrm{i}k\cdot x}\,\langle 0|\,\phi(0)\,|\lambda_{\boldsymbol{k}}\rangle$, so the x and y dependence collapses to $\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}$ and the matrix elements enter only through $|\langle 0|\,\phi(0)\,|\lambda_{\boldsymbol{k}}\rangle|^2$.

Lorentz covariance, Equation (109.11), says that this modulus squared is invariant: acting with $U(\Lambda)$ on both sides sends $|\lambda_{\boldsymbol{k}}\rangle$ to a state of the same sector with boosted wave vector, and Equation (109.37) is constructed to be invariant, so $|\langle 0|\,\phi(0)\,|\lambda_{\boldsymbol{k}}\rangle|^2$ depends on $\boldsymbol{k}$ only through the invariant $\hbar^2 k\cdot k = m_\lambda^2c^2$ — that is, not at all within a sector. Call it c_λ.

Now insert $1 = \int_0^\infty \mathrm{d}s\, \delta(s - m_\lambda^2 c^2)$ and define

$$\rho(s) := \frac{1}{\hbar c} \sum_\lambda c_\lambda\, \delta\Big(s - m_\lambda^2 c^2\Big)\,, \tag{109.42}$$

whose positivity is manifest, each c_λ being a squared modulus. Exchanging the order of the s and $\boldsymbol{k}$ integrations and comparing with Equation (109.38) gives Equation (109.39). The prefactor is the ratio of the two measures, not of the two prefactors: Equation (109.41) carries $\mathrm{d}^3k\, c/\left[(2\pi)^3 2\omega_{\boldsymbol{k}}\right]$ while Equation (109.38) carries $\hbar c^2\, \mathrm{d}^3k/\left[(2\pi)^3 2\omega_{\boldsymbol{k}}\right]$, and $c/(\hbar c^2) = 1/(\hbar c)$. The extra power of c is supplied by the Lorentz-invariant state measure of Notation 109.52, and it is what makes the bookkeeping close: that normalization gives $|\lambda_{\boldsymbol{k}}\rangle$ the dimension m, so $c_\lambda = |\langle 0|\, \phi(0)\, |\lambda_{\boldsymbol{k}}\rangle|^2$ carries $\mathrm{J\,m} = [\hbar c]$, the quotient $c_\lambda/(\hbar c)$ is a pure number, and ρ inherits the dimension $\mathrm{s}^2/(\mathrm{kg}^2\,\mathrm{m}^2)$ of the delta function alone — that is, $1/[s]$, so that Equation (109.40) is a statement between pure numbers. For a free field of mass m the single sector has $c_\lambda = \hbar c$ and $\rho(s) = \delta(s - m^2c^2)$, as it must.

For the sum rule, subtract Equation (109.39) from its $x \leftrightarrow y$ transpose:

$$[\phi(x), \phi(y)] = \int_0^\infty \mathrm{d}s\; \rho(s)\, \mathrm{i}\hbar c^2\, \Delta(x - y; s)\,, \tag{109.43}$$

with Δ the Pauli–Jordan function of Equation (105.101) for squared momentum s. Now $\Delta(\xi; s)$ vanishes at equal times and

$$\partial_{t_x} \Delta(x - y; s)\Big|_{t_x = t_y} = -\delta^3(\boldsymbol{x} - \boldsymbol{y})\,, \tag{109.44}$$

which is *independent of* s: differentiating Equation (105.101) brings down $\mp\mathrm{i}\omega_{\boldsymbol{k}}$ from the two exponentials, which cancels the $1/(2\omega_{\boldsymbol{k}})$ and leaves $-\int \mathrm{d}^3k\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})}/(2\pi)^3$ regardless of the mass. Therefore

$$[\phi(t, \boldsymbol{x}), \pi(t, \boldsymbol{y})] = \frac{1}{c^2} \partial_{t_y} [\phi(x), \phi(y)]\Big|_{t_x = t_y} = \mathrm{i}\hbar\, \delta^3(\boldsymbol{x} - \boldsymbol{y}) \int_0^\infty \mathrm{d}s\; \rho(s)\,,$$

using $\partial_{t_y} = -\partial_{t_x}$ on a function of $x - y$. Comparing with the canonical commutator $[\phi, \pi] = \mathrm{i}\hbar\, \delta^3 \mathbb{1}$ gives Equation (109.40). □

Corollary 109.55 (The field-strength constant is bounded)**.** *Suppose the spectrum contains an isolated one-particle state of mass m, so that $\rho(s) = Z\, \delta(s - m^2c^2) + \sigma(s)$ with $\sigma \geq 0$ supported on $s \geq s_{thr} > m^2c^2$. Then*

$$0 \leq Z = 1 - \int_{s_{thr}}^\infty \mathrm{d}s\; \sigma(s) \leq 1\,, \tag{109.45}$$

with $Z = 1$ if and only if $\sigma \equiv 0$, that is if and only if the field creates nothing but the one-particle state.

Rests on Equation (109.40)*.*

Derivation. Derives Corollary 109.55.

Substitute the decomposition into Equation (109.40). Both Z and $\int \sigma$ are non-negative because ρ is, and they sum to one, which gives Equation (109.45). For the equality case, $Z = 1$ forces $\int \sigma = 0$, and a non-negative measure of vanishing total mass is the zero measure, so $\sigma \equiv 0$; conversely $\sigma \equiv 0$ gives $Z = 1$ by the same sum rule. □

Remark 109.56 (From a free two-point function to a free field). It is tempting to end the corollary with "— a free field", and the conclusion is true, but it is a further theorem and not a restatement. What $\sigma \equiv 0$ says is that the *two-point* function of ϕ coincides with that of a free field of mass m; that all the higher Wightman functions are then the free ones as well, so that ϕ itself satisfies the Klein–Gordon equation $(\Box + m^2c^2/\hbar^2)\, \phi = 0$ and is a free field, is the Jost–Schroer theorem [1297, 1312]. It is proved from analyticity, not from the sum rule.

Derivation pending #185

The Jost–Schroer theorem: a Wightman field whose two-point function is that of a free field of mass m is a free field. The argument runs through the analyticity of the truncated Wightman functions in the extended tube and the vanishing of the truncated two-point function's support off the mass shell; it belongs in Appendix A beside the other analyticity results of this chapter.

This is the structural fact Remark 100.38 appeals to, in the case this chapter can prove it: a Hermitian scalar field, with the sum rule coming from the canonical equal-time *commutator*. The residue of the exact propagator at its pole is Z; the deficit $1 - Z$ is the probability weight the interacting field puts on multiparticle states; and the inequality $Z \leq 1$ is not a perturbative estimate but a consequence of the axioms. What Remark 100.38 owes is the *fermion* case — the same representation for the exact electron two-point function, with the sum rule taken instead from the canonical equal-time anticommutator, bounding Z_2 rather than Z. That derivation is not carried out here, and the pending entry standing beside Remark 100.38 remains live; nor would Corollary 109.55 transfer to it unchanged, for the reason given in Remark 109.57 below. In momentum space the same content reads as Equation (106.61), whose Feynman-ordered form follows from Equation (109.39) by the standard contour prescription; the branch point at $s = s_{\mathrm{thr}}$ visible there is the threshold of Corollary 109.55, and for a self-interacting scalar of mass m it sits at $(2mc)^2$.

Remark 109.57 (Why the representation is exact). Nothing in the derivation expanded in a coupling. Only three ingredients were used: completeness of the states, the spectrum condition Axiom 109.10, and Lorentz covariance Axiom 109.13. The result is therefore a constraint on *any* theory satisfying the axioms, and it is why $Z \leq 1$ can be quoted against a perturbative calculation that appears to violate it: such a calculation has an error, or the scheme in which it is performed is not one in which Z is the residue of a physical pole. The corresponding statement for the electron propagator, where the one-particle "pole" is not isolated because a charged state is accompanied by arbitrarily soft photons, is the infraparticle difficulty of Remark 109.18 and Section 100.6: there Z_2 is infrared divergent and Corollary 109.55 does not apply, because its hypothesis fails.

109.3.4 Asymptotic states and LSZ

Nothing so far mentions particles. The axioms describe fields and states; a particle is an extra structure, and the theorem that produces it from the axioms is the Haag–Ruelle construction.

Definition 109.58 (Isolated mass shell)**.** The theory has a *one-particle state of mass m* if the joint spectrum of P^μ contains an isolated hyperboloid $p^2 = m^2c^2$, $p^0 > 0$, separated from the rest of the spectrum by a gap: the continuum begins at $s_{\mathrm{thr}} > m^2c^2$.

Theorem 109.59 (Haag–Ruelle)**.** *Let a Wightman theory have an isolated mass shell of mass m and let ϕ have non-vanishing overlap with it, $\langle 0|\,\phi(0)\,|p\rangle \neq 0$. Choose f with $\tilde f$ supported in a neighbourhood of that shell meeting no other part of the spectrum, and set*

$$B_f(t) := \frac{\mathrm{i}}{\hbar c^2}\int \mathrm{d}^3x\,\left[\phi(t,\boldsymbol{x})\,\partial_t \overline{f(t,\boldsymbol{x})} - (\partial_t\phi(t,\boldsymbol{x}))\,\overline{f(t,\boldsymbol{x})}\right], \tag{109.46}$$

with f a positive-frequency solution of the Klein–Gordon equation of mass m. The prefactor $\mathrm{i}/(\hbar c^2)$ is the one that makes $\left[B_f, B_g^\dagger\right]$ the Klein–Gordon inner product of the two packets and a pure number, since the equal-time commutator is $[\phi, \partial_t\phi] = \mathrm{i}\hbar c^2\delta^3$; f is normalized accordingly. Then the limits

$$|f_1, \dots, f_n\rangle_{in}^{out} = \lim_{t\to\pm\infty} B_{f_1}(t)\cdots B_{f_n}(t)\,|0\rangle \tag{109.47}$$

exist in norm, are independent of the interpolating field within its Borchers class, and span Fock spaces of free particles of mass m on which the Poincaré group acts as it does on free states.

Rests on Definition 109.58 and Equation (109.19).

Status and mechanism. Derives Theorem 109.59.

The mechanism is a competition of rates. A single $B_f(t)$ applied to the vacuum is t-independent: its one-particle component is exact, because $\tilde{f}$ meets only the one-particle shell and Equation (109.46) is the Klein–Gordon inner product, which is conserved. For a product of n such operators the t-dependence comes from the overlap of the wave packets, and cluster decomposition Equation (109.19) makes connected contributions fall off faster than $t^{-3/2}$ per pair while the number of pairs is fixed; the Cauchy criterion then gives norm convergence. What must be supplied is the quantitative decay estimate, which uses the mass gap of Definition 109.58 and stationary phase. □

Derivation pending #186

The Haag–Ruelle convergence estimate: that the vectors of the displayed limit form a Cauchy family, with the rate supplied by cluster decomposition together with the stationary-phase decay of a wave packet built on an isolated mass shell. The construction and its mechanism are given in the text; the estimate is a page of analysis and belongs in Appendix A.

Theorem 109.60 (LSZ reduction)**.** *Under the hypotheses of Theorem 109.59, the S-matrix elements between the states Equation (109.47) are the residues of the time-ordered Wightman functions at their one-particle poles, in the form Equation (106.62) [1073].*

Rests on Theorems 109.54 and 109.59 and Corollary 109.55.

Derivation. Derives Theorem 109.60.

The pole structure is Theorem 109.54: a leg carrying momentum p has, by Corollary 109.55, an isolated pole of residue proportional to Z at $p^2 = m^2c^2$ and a branch cut starting at s_{thr}, and multiplying by $(p^2 - m^2c^2)$ before taking the limit isolates the first and discards the second. The identification of what survives with a matrix element between Haag–Ruelle states is Theorem 109.59. The algebra is carried out in Theorem 106.48; what this chapter adds is that its two ingredients — the analytic structure of the two-point function and the existence of asymptotic states — are theorems of the axioms rather than assumptions of the path integral. □

Remark 109.61 (Three honest qualifications)*. Asymptotic completeness is an assumption.* The Haag–Ruelle states span *a* subspace of $\mathcal{H}$; that they span all of it is an additional hypothesis. It is what allows the S matrix to be unitary, and hence what underwrites the optical theorem and every cross-section normalization of Chapter 86. It has never been proved in an interacting four-dimensional model, and it is false in some soluble lower-dimensional ones.

Massless particles break the construction. The gap in Definition 109.58 is used essentially, and for a massless particle there is no gap: the one-particle shell touches the two-particle continuum at the origin. This is not a technicality. In quantum electrodynamics a charged particle is always accompanied by an indefinite number of soft photons and is not an eigenvector of the mass operator at all — an *infraparticle*, in the analysis of Buchholz and Fredenhagen [1684], whose superselection structure is that of Section 109.4.2 rather than that of a Wigner particle. The infrared divergences cancelled by hand in Section 100.6 are the perturbative shadow of this.

Confined fields have no shell at all. The quark and gluon fields of Chapter 102 have no isolated one-particle hyperboloid in the physical spectrum, so Theorem 109.59 does not

apply to them. This is a precise statement of what confinement would mean in this language, and the reason Section 102.7.3 must argue from evidence rather than from theorem.

109.4 Algebraic quantum field theory

109.4.1 Haag–Kastler nets

Corollary 109.39 showed that the field is a convention. The algebraic formulation takes that seriously and drops the field from the list of primitive objects, keeping only what is measurable.

Definition 109.62 (Local net of observables)**.** A *Haag–Kastler net* assigns to every bounded open region $\mathcal{O} \subset \mathbb{R}^4$ a von Neumann algebra $\mathcal{A}(\mathcal{O})$ of bounded operators on a Hilbert space, subject to

1. **isotony**: $\mathcal{O}_1 \subset \mathcal{O}_2$ implies $\mathcal{A}(\mathcal{O}_1) \subset \mathcal{A}(\mathcal{O}_2)$;
2. **locality**: if $\mathcal{O}_1$ and $\mathcal{O}_2$ are spacelike separated, every element of $\mathcal{A}(\mathcal{O}_1)$ commutes with every element of $\mathcal{A}(\mathcal{O}_2)$;
3. **covariance**: there is a unitary representation $U(a,\Lambda)$ with $U(a,\Lambda)\mathcal{A}(\mathcal{O})U(a,\Lambda)^{-1} = \mathcal{A}(\Lambda\mathcal{O}+a)$, with spectrum in $\overline{V}^{+}$ and an invariant vacuum.

The *quasilocal algebra* $\mathfrak{A}$ is the norm closure of $\bigcup_{\mathcal{O}} \mathcal{A}(\mathcal{O})$, an inductive limit directed by isotony, and is a C^*-algebra [1280].

Remark 109.63 (Why bounded operators, and where the units went)*.* A Wightman field is unbounded and its domain problems are real (Remark 109.3); the algebraic formulation avoids them by taking as generators bounded functions of the smeared fields, for example the unitaries $\exp(\mathrm{i}\phi(f)/[\phi])$ with the exponent made dimensionless by the field's own normalization. All elements of $\mathfrak{A}$ are therefore *dimensionless* bounded operators, and every dimensioned quantity of this chapter — the J/m of $[\phi]^2$, the kg m/s of P^μ, the $\hbar$ in Equation (109.10) — enters only through the particular fields chosen to generate the net and through the representation of the translations. That is not a loss: it is the statement that the physical content is in the net and its symmetry, not in the units of whatever field was used to write it down.

The physical argument for the shift is short. A measurement occupies a bounded region of spacetime, so what it can determine is an element of $\mathcal{A}(\mathcal{O})$ for that region; the field is not such an element, it is merely a convenient set of generators, and by Theorem 109.38 a whole class of fields generates the same net. Everything physical — the particle content, the statistics, the charges, the S matrix — must therefore be recoverable from the net alone. Section 109.4.2 shows that it is.

Remark 109.64 (Curved spacetime)*.* The Wightman axioms of Section 109.2.1 cannot be transplanted to a curved spacetime, because a generic solution of the Einstein equations (Chapter 44) has no Poincaré group, no preferred vacuum and no global momentum operator, so Axioms 109.9 to 109.11 all fail to parse. Definition 109.62 survives, because isotony and locality are statements about regions and their causal relations, which a Lorentzian manifold still has. The replacement of covariance is *local covariance*: the assignment of an algebra to a region must be functorial under isometric embeddings of one spacetime into another, so that the same physics holds in every small enough neighbourhood [2013]. This is how a quantum field on a curved background is formulated when it is formulated carefully — including the Hawking process [1549], the prediction that a black hole radiates thermally. That prediction shares the status recorded in Remark 109.75 for the acceleration temperature, and for the

same arithmetic reason: it has never been observed, no astrophysical black hole is anywhere near hot enough against the microwave background, and the laboratory work on it measures analogue systems obeying the same equations rather than the physical claim. It is named here for completeness; no theorem of this chapter uses it, and nothing in this book rests on it.

109.4.2 States, GNS and superselection

Once the observables are an abstract algebra, a state is not a vector but a rule assigning expectations to observables — Segal's formulation of quantum mechanics [904] — and the Hilbert space is something one constructs from it.

Definition 109.65 (State on an algebra)**.** A *state* on a C^*-algebra $\mathfrak{A}$ with unit is a linear functional $\omega : \mathfrak{A} \to \mathbb{C}$ with $\omega(A^*A) \geq 0$ for all A and $\omega(\mathbb{1}) = 1$.

Theorem 109.66 (Gelfand–Naimark–Segal)**.** *For every state ω on a C^*-algebra $\mathfrak{A}$ there is a Hilbert space $\mathcal{H}_\omega$, a representation π_ω of $\mathfrak{A}$ by bounded operators on it, and a unit vector $|\omega\rangle \in \mathcal{H}_\omega$, cyclic for $\pi_\omega(\mathfrak{A})$, such that*

$$\omega(A) = \langle \omega | \, \pi_\omega(A) \, | \omega \rangle \qquad \text{for all } A \in \mathfrak{A} \,. \tag{109.48}$$

The triple is unique up to unitary equivalence [871, 904].
Rests on Definition 109.65.

Derivation. *Derives Theorem 109.66.*

Put $\langle A, B \rangle_\omega := \omega(A^*B)$ on $\mathfrak{A}$ regarded as a vector space. This is sesquilinear and non-negative, hence satisfies Cauchy–Schwarz, $|\omega(A^*B)|^2 \leq \omega(A^*A)\,\omega(B^*B)$.

Let $\mathfrak{N} := \{A : \omega(A^*A) = 0\}$. By Cauchy–Schwarz, $A \in \mathfrak{N}$ implies $\omega(B^*A) = 0$ for all B, so $\mathfrak{N}$ is a subspace; and for $A \in \mathfrak{N}$, $\omega((BA)^*(BA)) = \omega(A^*B^*BA) \leq \|B^*B\|\,\omega(A^*A) = 0$ using the C^* inequality $A^*CA \leq \|C\|\,A^*A$ for $C \geq 0$. So $\mathfrak{N}$ is a left ideal.

Let $\mathcal{H}_\omega$ be the completion of $\mathfrak{A}/\mathfrak{N}$ in the induced — now definite — inner product, write $[A]$ for the class of A, and set $\pi_\omega(B)[A] := [BA]$, well defined because $\mathfrak{N}$ is a left ideal and bounded because $\|[BA]\|^2 = \omega(A^*B^*BA) \leq \|B\|^2\,\|[A]\|^2$. That π_ω is a $*$-homomorphism is immediate from the definitions. Take $|\omega\rangle := [\mathbb{1}]$: then $\langle \omega | \, \pi_\omega(A) \, | \omega \rangle = \omega(\mathbb{1}^*A\mathbb{1}) = \omega(A)$, which is Equation (109.48), and cyclicity holds because $\pi_\omega(\mathfrak{A})\,|\omega\rangle = \mathfrak{A}/\mathfrak{N}$ is dense by construction. Uniqueness: two such triples are intertwined by $\pi_\omega(A)\,|\omega\rangle \mapsto \pi'_\omega(A)\,|\omega\rangle'$, which preserves inner products by Equation (109.48) and is densely defined by cyclicity. □

Remark 109.67 (The reconstruction theorem again). Theorem 109.66 and Theorem 109.21 are the same construction. In the Wightman case the algebra is the Borchers tensor algebra Equation (109.17), the state is the collection of Wightman functions, and property (v) of Proposition 109.20 is $\omega(A^*A) \geq 0$. This is why a positivity condition is the one axiom that cannot be relaxed in either formulation, and why it appears a third time as reflection positivity in Section 109.5.1. In the commutative case the same piece of mathematics is the classical moment problem — reconstructing a measure from its moments, the condition being positivity of the Hankel form built from them. That problem is not treated anywhere in this book, and is recorded here only to name the mathematics the reconstruction theorem generalizes.

Definition 109.68 (Superselection sector)**.** Two states on $\mathfrak{A}$ lie in the same *superselection sector* if their GNS representations are unitarily equivalent. Inequivalent representations describe states between which no observable has a matrix element, so no coherent superposition of them is preparable.

Remark 109.69 (Why sectors are unavoidable in field theory). For finitely many canonical pairs the Stone–von Neumann theorem makes all irreducible representations equivalent and the notion is empty. In field theory it is not: Corollary 109.5 is the statement that a free and an interacting theory are inequivalent representations of the same canonical relations, and the several vacua of a spontaneously broken theory (Section 104.3) are inequivalent representations too — which is the precise sense in which choosing a vacuum is not choosing a state within one theory but choosing a theory within one algebra. The remark in Chapter 25 that this is a feature and not a pathology is Definition 109.68.

Theorem 109.70 (Doplicher–Haag–Roberts; imported)**.** *Let the sectors of a net satisfying Definition 109.62 be restricted to those that look like the vacuum sector when observed outside any bounded region — the* localizable *charges. Then the set of such sectors carries the structure of the representation category of a unique compact group G, the global gauge group; each sector is labelled by an irreducible representation of G; a permutation statistics is attached to each sector, and in $3+1$ dimensions it is necessarily Bose or Fermi with a definite integer order (para-statistics of order d, reducible to ordinary statistics of d-component fields); and a field algebra generating the sectors from the vacuum exists, on which G acts with the observables as its fixed points [1445, 1525].*

Rests on Definitions 109.62 and 109.68.

Remark 109.71 (What the theorem accomplishes). It is the strongest structural result in the subject. The input is a net of local observables and nothing else: no fields, no gauge group, no charges, no statistics. The output contains all four. That the internal symmetry group of a theory is compact, that charges are labelled by its irreducible representations, that the exchange statistics is Bose or Fermi and not something else, and that a field algebra exists at all are *derived*, not postulated — whereas in Chapter 104 the gauge group is chosen to fit the data and the fields are chosen to carry it. The result is a theorem about the framework and carries no prediction that distinguishes one Standard Model from another; it says that the shape of the Standard Model is the only shape a local theory can have.

Two honest limitations. The analysis assumes charges localizable in a bounded region, which excludes the electric charge itself — Gauss's law makes it measurable at infinity, and charged states in quantum electrodynamics are the infraparticles of Remark 109.61. And the conclusion about Bose or Fermi statistics uses $3+1$ dimensions, through the connectedness of the spacelike complement of a bounded region; the same conclusion fails in lower dimension, which is the algebraic counterpart of the dimensional caveat in Remark 109.51.

Derivation pending #187

The Doplicher–Haag–Roberts reconstruction: that localizable superselection sectors of a local net form the representation category of a compact group, with Bose or Fermi statistics forced in $[3+1]$ dimensions, and that a field algebra carrying that group exists. The statement, its inputs and its two limitations are given in the text; the proof is long and belongs in Appendix A.

109.4.3 Modular theory

The Reeh–Schlieder theorem, Theorem 109.34, says the vacuum is cyclic and separating for a local algebra. That is exactly the hypothesis of a theorem of operator algebras which then hands back a canonical dynamics that nobody put in.

Theorem 109.72 (Tomita–Takesaki; imported)**.** *Let $\mathcal{M}$ be a von Neumann algebra with a cyclic and separating vector $|\Psi\rangle$. Define S on the dense set $\mathcal{M}\,|\Psi\rangle$ by $S\,A\,|\Psi\rangle := A^{*}\,|\Psi\rangle$,*

and let $S = J\Delta^{1/2}$ be its polar decomposition, with J antiunitary and $\Delta > 0$ self-adjoint. Then

$$\Delta^{\mathrm{i}\tau}\,\mathcal{M}\,\Delta^{-\mathrm{i}\tau} = \mathcal{M} \quad \text{for all } \tau \in \mathbb{R}\,, \qquad J\mathcal{M}J = \mathcal{M}'\,, \tag{109.49}$$

where $\mathcal{M}'$ is the commutant [1436].

Rests on Theorem 109.34.

The one-parameter group $\sigma_\tau(A) = \Delta^{\mathrm{i}\tau} A \Delta^{-\mathrm{i}\tau}$ is the *modular automorphism group*: an intrinsic time evolution of the algebra, determined by the algebra and the state alone. In general it has no geometric meaning. For one class of regions it does, and the fact is remarkable.

Theorem 109.73 (Bisognano–Wichmann; imported)**.** *Let $\mathcal{A}(W)$ be the algebra of the right wedge $W = \{x : x^1 > |x^0|\}$ in a Wightman theory, and $|0\rangle$ the vacuum. Then the modular group of $(\mathcal{A}(W), |0\rangle)$ is the one-parameter group of Lorentz boosts preserving W,*

$$\Delta^{\mathrm{i}\tau} = U(\Lambda_W(-2\pi\tau))\ , \tag{109.50}$$

$\Lambda_W(\eta)$ being the boost of rapidity η in the (x^0, x^1) plane; and the modular conjugation is $J = \Theta\, U(R_1(\pi))$, the PCT operator of Theorem 109.46 composed with a rotation by π about the x^1 axis [1540, 1572].

Rests on Theorems 109.46 and 109.72.

Proposition 109.74 (The vacuum is thermal in a wedge)**.** *Restricted to $\mathcal{A}(W)$, the vacuum satisfies the KMS condition with respect to the boost flow at inverse temperature 2π in the rapidity. Consequently, for an observer moving on a worldline of constant proper acceleration a inside W, whose proper time τ_p and rapidity are related by $\eta = a\tau_p/c$, the vacuum is a thermal state at*

$$T = \frac{\hbar a}{2\pi c\, k_B}\,. \tag{109.51}$$

Rests on Theorems 109.72 and 109.73 and Proposition 40.3.

Derivation. *Derives Proposition 109.74.*

The KMS condition characterizing a thermal state at temperature T with respect to a time evolution α_t is that $t \mapsto \omega(A\,\alpha_t(B))$ extends analytically to the strip $0 < \operatorname{im} t < \hbar/(k_B T)$ with $\omega\big(A\,\alpha_{t+\mathrm{i}\hbar/k_BT}(B)\big) = \omega(\alpha_t(B)\,A)$; equivalently the correlation functions are periodic in imaginary time with period $\hbar/(k_BT)$. Tomita–Takesaki delivers exactly this for σ_τ with period 1 in τ, which is a general property of the modular group; by Equation (109.50) the flow is the boost with rapidity $\eta = -2\pi\tau$, so the vacuum correlation functions are periodic in imaginary rapidity with period 2π.

Now convert to the proper time of the accelerated observer. A worldline of constant proper acceleration a in W is an orbit of the boost, and the relation $\eta = a\tau_p/c$ is not quoted from elsewhere but follows in two lines. Parametrize the orbit by the boost rapidity η of Definition 38.21, so that the four-velocity of Definition 40.2 is

$$u^\mu = c\,(\cosh\eta,\ \sinh\eta,\ 0,\ 0)\ ,$$

which has $u^2 = c^2$ as Proposition 40.3 requires. The four-acceleration is then $a^\mu = \mathrm{d}u^\mu/\mathrm{d}\tau_p = c\,(\sinh\eta, \cosh\eta, 0, 0)\,\mathrm{d}\eta/\mathrm{d}\tau_p$, whose invariant square is $a^\mu a_\mu = -c^2\,(\mathrm{d}\eta/\mathrm{d}\tau_p)^2$. The proper acceleration is the magnitude of this spacelike vector, $a = \sqrt{-a^\mu a_\mu} = c\,\mathrm{d}\eta/\mathrm{d}\tau_p$, so a constant integrates to $\eta = a\tau_p/c$, and c/a is the length scale of the resulting hyperbola. Periodicity in imaginary η with period 2π is therefore periodicity in imaginary τ_p with period $2\pi c/a$, and equating this to $\hbar/(k_BT)$ gives Equation (109.51). Dimensional check: $\hbar a$ has $\mathrm{J\,s} \times \mathrm{m/s^2} = \mathrm{J\,m/s}$, and dividing by c leaves J, so $\hbar a/(2\pi c k_B)$ is a temperature. □

Remark 109.75 (The honest status of the acceleration temperature). Equation (109.51) is the Unruh temperature [1590], and it has never been observed. The reason is arithmetic: an acceleration of $g = 9.806\,65\,\mathrm{m/s^2}$ gives $T = 3.98 \times 10^{-20}\,\mathrm{K}$, and reaching $T = 1\,\mathrm{K}$ requires $a = 2.47 \times 10^{20}\,\mathrm{m/s^2}$. No apparatus accelerates a detector at that rate for long enough to thermalize it.

Two things follow, and neither is a hedge. First, this chapter's theorems — Theorems 109.72 and 109.73 — are mathematics and stand independently of whether the temperature is ever measured; what is unobserved is a physical consequence, not the structure. Second, proposals to see an analogue of the effect in condensed-matter or optical systems [1675] measure an analogue: a laboratory system engineered to obey the same equations, whose agreement confirms the mathematics and not the physical claim about the quantum vacuum. Under the evidence rule of this treatise the distinction is not a technicality, and the entry stands as: theorem established, prediction untested.

Remark 109.76 (Modular theory and the entanglement of the vacuum). The operator Δ is the field-theoretic analogue of a reduced density matrix: for a finite quantum system with a bipartition, $\Delta = \rho_L \otimes \rho_R^{-1}$ and the modular flow is $\rho^{\mathrm{i}\tau} \cdot \rho^{-\mathrm{i}\tau}$. The reason field theory needs the modular operator rather than the density matrix is that $\mathcal{A}(\mathcal{O})$ is a von Neumann algebra of type III, in which no trace exists and no density matrix can be written — which is the structural reason the entanglement entropy of a region diverges, and the sense in which the divergence flagged in Chapter 84 is not an artefact of a cutoff but a property of the algebra. Theorem 109.34 is what makes $|0\rangle$ cyclic and separating and hence makes all of this available; the two theorems are the same fact told twice.

Derivation pending #188

Modular theory: the Tomita–Takesaki theorem, that the polar decomposition of the Tomita operator yields a modular automorphism group preserving the algebra and a conjugation exchanging it with its commutant; and the Bisognano–Wichmann identification of that group, for a wedge in a Wightman theory, with the boosts preserving the wedge. Both are stated with their hypotheses and used only through the KMS property, from which the acceleration temperature is derived in full in the text. The two proofs are long and belong in Appendix A.

109.5 The Euclidean route

109.5.1 Osterwalder–Schrader axioms

Nothing above suggests how to *construct* a theory. The one route that has ever produced an interacting example runs through imaginary time, where the oscillatory weight $\exp(\mathrm{i}S/\hbar)$ of Chapter 106 becomes the positive weight $\exp(-S_E/\hbar)$ of a probability measure, and the problem becomes one in classical statistical mechanics.

Definition 109.77 (Schwinger functions)**.** The Schwinger functions $S_n(x_1, \dots, x_n)$ are the moments of that measure, Equation (106.29), with the x_i distinct points of Euclidean $\mathbb{R}^4$ and the fourth coordinate $x^4 = \mathrm{i}ct$.

Theorem 109.78 (Osterwalder–Schrader)**.** *A family of Schwinger functions satisfying regularity, Euclidean covariance under* $\mathrm{SO}(4) \ltimes \mathbb{R}^4$*, reflection positivity, permutation symmetry and cluster decomposition is the analytic continuation of the Wightman functions of a unique Wightman theory in* $3+1$ *Minkowski dimensions, and conversely [1510, 1555]. The precise statement, with the five conditions written out, is Theorem 106.20.*

Rests on Definition 109.77 and Theorem 106.20.

The reconstruction of the Wightman functions from the Schwinger functions by analytic continuation in the time arguments, and the verification of the spectrum condition and of locality for the result, are not carried out here; they are owed in Appendix A, and the pending entry that records the debt is the one attached to Theorem 106.20, which is not duplicated here. What is derived below is the half that carries the physics.

Proposition 109.79 (Reflection positivity is Hilbert positivity)**.** *Under the correspondence of Theorem 109.78, the reflection positivity condition Equation* (106.30) *is the Euclidean form of property (v) of Proposition 109.20, and the operator* $\mathrm{e}^{-\tau H/\hbar}$ *reconstructed from Euclidean time translation is a self-adjoint contraction semigroup, so the Hamiltonian it generates is self-adjoint and bounded below.*

Rests on Theorem 109.78, Equation (106.30)*, and Proposition 109.20.*

Derivation. *Derives Proposition 109.79.*

Let $\mathcal{E}_+$ be the space of functionals of the field supported in $x^4 > 0$, and θ the reflection $x^4 \mapsto -x^4$. Reflection positivity says that $\langle F, G\rangle := \left\langle \overline{\theta F}\, G \right\rangle$ is a non-negative sesquilinear form on $\mathcal{E}_+$. Quotient by its null space and complete, exactly as in Theorem 109.66: the result is a Hilbert space $\mathcal{H}$ with a positive-definite inner product, which is property (v).

Euclidean time translation T_τ by $\tau > 0$ maps $\mathcal{E}_+$ into itself, and $\langle F, T_\tau G\rangle = \left\langle T_{\tau/2}F, T_{\tau/2}G \right\rangle$ by translation invariance and the reflection property, so T_τ descends to a self-adjoint operator on $\mathcal{H}$, positive by the same identity with $F = G$.

That T_τ is a *contraction* takes one step more than Cauchy–Schwarz alone. Write $g(\tau) := \|T_\tau F\|$, so that $g(0) = \|F\|$. The identity above with $G = F$ gives $g(\tau/2)^2 = \langle F, T_\tau F\rangle$, and Cauchy–Schwarz bounds the right-hand side by $\|F\|\,\|T_\tau F\|$, so

$$g(\tau/2)^2 \le g(0)\, g(\tau) \qquad \text{equivalently} \qquad g(\tau)^2 \le g(0)\, g(2\tau)\,.$$

This relates two *different* times and by itself bounds nothing; it must be iterated. Feeding it into itself n times,

$$g(\tau) \le g(0)^{1-2^{-n}}\, g(2^n\tau)^{2^{-n}}\,,$$

and if g is bounded uniformly in τ — which is what the regularity condition of Remark 109.80 is there to supply — then $g(2^n\tau)^{2^{-n}} \to 1$ as $n \to \infty$ and $g(\tau) \le g(0)$, that is $\|T_\tau\| \le 1$. A self-adjoint contraction semigroup is $\mathrm{e}^{-\tau H/\hbar}$ for a unique self-adjoint $H \ge 0$, and the factor $\hbar$ is fixed by the Euclidean weight $\mathrm{e}^{-S_E/\hbar}$ from which T_τ was built. The quantum-mechanical case of this argument, with the operator chain written out explicitly, is Proposition 106.21. □

Remark 109.80 (The linear-growth condition, and why it was added)*.* The first paper [1510] stated a regularity condition that proved too weak: it did not control the growth of S_n with n well enough to guarantee that the reconstructed objects are operator-valued *tempered* distributions rather than something wilder. The second paper [1555] replaced it by a linear-growth condition bounding $\|S_n\|$ by $n!\,(\text{const})^n$ times a Schwartz seminorm. The episode is worth recording as an example of an axiom system being corrected, and of why the correction matters: without it Theorem 109.78 would reconstruct a theory whose fields need not satisfy Axiom 109.12, and Theorem 109.26 — on which the whole of Section 109.3 rests — would be unavailable.

Remark 109.81 (What the Euclidean route buys)*.* Three things, and they are why every constructive success is Euclidean. The weight $\mathrm{e}^{-S_E/\hbar}$ is a positive measure, so the whole apparatus of probability — correlation inequalities, cluster expansions, the theory of Gibbs measures — becomes available. None of those three is developed in this book, whose probability chapter Chapter 11 covers the inferential statistics the experimental parts need rather than the theory of random fields; they are taken here from the constructive literature [1792]. The Euclidean symmetry group SO(4) is compact, unlike the Lorentz group, so its

representation theory is far better behaved. And a cutoff on a lattice respects reflection positivity exactly (Section 109.5.2), so the approximating theories are themselves quantum theories rather than formal expressions. What is bought is a construction problem in classical probability; what is paid is that the continuation back to real time must be established, and that is where Theorem 109.78 does the work.

109.5.2 Reflection positivity and the lattice

Proposition 109.82 (Wilson's action is reflection positive)**.** *The lattice gauge action of Equation* (106.81) *[1537], being a sum of plaquette terms each of which involves link variables on at most two adjacent time slices, satisfies Equation* (106.30) *for reflection in a plane midway between two slices. Consequently the lattice theory has a positive self-adjoint transfer operator* $T = \mathrm{e}^{-aH/(\hbar c)}$*, with* a *the lattice spacing in* m*, and hence a Hilbert space with a Hamiltonian bounded below.*

Rests on Equations (106.30) *and* (106.81) *and Proposition 106.21.*

Derivation. *Derives Proposition 109.82.*

Write the Euclidean action as a sum over time slices, $S_E = \sum_n [V(U_n) + K(U_n, U_{n+1})]$, in which V involves only the links within slice n and K only the links of two adjacent slices — this is exactly the plaquette structure, since a plaquette either lies in one slice or joins two neighbours. The functional integral then factorizes into a product of kernels, and reflection about the plane between slice 0 and slice 1 exchanges the two halves while conjugating, so that $\langle \overline{\theta F}\, F \rangle$ is the squared norm of the vector obtained by applying the half-chain to the ground state — the argument of Proposition 106.21, with the continuous evolution replaced by a finite product of kernels. Positivity of the transfer operator follows because the kernel e^{-K} is a positive-definite function of the pair of slice configurations, and self-adjointness because K is symmetric under exchange of the two slices. The exponent $aH/(\hbar c)$ is dimensionless because H is an energy and a/c a time.

The converse warning is Remark 106.22: an action coupling next-to-nearest time slices generally breaks the factorization, and the resulting transfer operator can have negative eigenvalues, whose signature is a correlator oscillating in Euclidean time. □

Proposition 109.83 (Mass gap and correlation length)**.** *In a reflection-positive Euclidean theory with a mass gap* $\Delta = Mc^2$ *above the vacuum, the connected correlation function of any local operator* O *decays as*

$$\langle O(0)\, O(r) \rangle_c \;\sim\; |\langle 0|\, O\, |1\rangle|^2 \, \mathrm{e}^{-r/\xi}, \qquad \xi = \frac{\hbar c}{\Delta} = \frac{\hbar}{Mc}, \tag{109.52}$$

for Euclidean separation r *in the time direction, where* $|1\rangle$ *is the lowest state with a non-zero overlap. The correlation length is the Compton wavelength of the lightest such state.*

Rests on Proposition 109.79.

Derivation. *Derives Proposition 109.83.*

By Proposition 109.79, Euclidean time translation is $\mathrm{e}^{-\tau H/\hbar}$. Inserting a complete set of energy eigenstates between the two operators and subtracting the vacuum contribution,

$$\langle O(0) O(\tau) \rangle_c = \sum_{n \neq 0} |\langle 0|\, O\, |n\rangle|^2 \, \mathrm{e}^{-(E_n - E_0)\tau/\hbar},$$

which for large τ is dominated by the smallest gap. Writing $\tau = r/c$ for the Euclidean distance r in metres and $E_1 - E_0 = \Delta$ gives Equation (109.52). With $\Delta = Mc^2$, $\xi = \hbar c/(Mc^2) = \hbar/(Mc)$. □

Example 109.84 (The pion sets the range of the strong force)**.** For quantum chromodynamics the lightest state carrying the quantum numbers of the *charged* component of the isovector pseudoscalar current is the charged pion, of rest energy $(139.570\,39 \pm 0.000\,18)\,\mathrm{MeV}$ [2427]. The neutral pion is lighter, $(134.9768 \pm 0.0005)\,\mathrm{MeV}$ [2427], and is equally an isovector pseudoscalar; the charged pion is the one taken here because it is the exchanged quantum in the nuclear-force estimate this example is written to match. By Equation (109.52) the corresponding correlation length is

$$\xi = \frac{\hbar c}{m_\pi c^2} = \frac{197.326\,980\,4\,\mathrm{MeV\,fm}}{139.570\,39\,\mathrm{MeV}} = 1.4138\,\mathrm{fm}\,, \tag{109.53}$$

which is the range of the nuclear force used throughout Chapter 107. That the range of a force and the decay rate of a Euclidean correlator are the same number is Proposition 109.83, and it is why a lattice computation measures masses by fitting exponentials.

Remark 109.85 (What the lattice does and does not establish)*.* Proposition 109.82 is what makes lattice gauge theory a candidate *constructive* route rather than merely a numerical method: each lattice theory is a bona fide quantum theory with a positive Hilbert space, so a continuum limit of them — if it exists and if the limit inherits reflection positivity, Euclidean invariance and the growth bound — would satisfy Theorem 109.78 and hence produce a Wightman theory. Neither the existence of the limit nor the inheritance has been proved.

What has been done is numerical, and it is impressive. With improved actions, physical quark masses, several spacings extrapolated to zero and several volumes extrapolated to infinity, the light hadron spectrum is reproduced at the percent level from the quark masses and the coupling alone [2122]; the details of the procedure are in Remark 102.61. The spectrum so computed begins at a non-zero mass, which is evidence that the gap exists. It is not a proof, for two reasons that must both be stated: the computation is performed at finite spacing and finite volume and extrapolated, with the extrapolation controlled by fits rather than by bounds; and a numerical demonstration that the lowest state has non-zero mass to within a stated error is not the statement $\Delta > 0$ that Section 109.7.3 requires.

109.6 The operator product expansion

Wilson conjectured that the singular behaviour of a product of local operators at short separation is carried entirely by c-number coefficient functions, the operator content being regular [1410]. It is the structural statement behind every short-distance calculation in the book.

Definition 109.86 (Operator product expansion)**.** Local operators A and B admit an *operator product expansion* if, as $\xi = x - y \to 0$,

$$A(x)\,B(y) \;\sim\; \sum_{n} C_n^{AB}(\xi)\,\mathcal{O}_n(y)\,, \tag{109.54}$$

asymptotically in the sense that truncating the sum after finitely many terms leaves a remainder vanishing faster than the last coefficient retained, when both sides are inserted into any correlation function. The C_n are c-number distributions carrying all the singularity; the $\mathcal{O}_n$ are local operators, ordered by increasing dimension.

Example 109.87 (The free scalar field)**.** For a free Hermitian scalar of mass m,

$$\phi(x)\phi(y) = \Delta^{+}(\xi; m^2c^2)\,\mathbb{1} + {:}\phi(x)\phi(y){:}\,, \tag{109.55}$$

by Wick's theorem (Theorem 99.41), and the second term is regular at $\xi = 0$, where it becomes ${:}\phi^2(y){:}$. Expanding it in ξ produces the higher operators $\xi^\mu\,{:}\phi\partial_\mu\phi{:}$ and so on. The whole

singularity therefore sits in the identity coefficient $\Delta^{+}(\xi; m^2c^2)$. Evaluating Equation (109.38) at equal times and spatial separation $r = |\boldsymbol{\xi}|$ small compared with the Compton wavelength $\hbar/(mc)$, the mass drops out of the integrand and $\int \mathrm{d}^3k\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\xi}}/\left[(2\pi)^3\,|\boldsymbol{k}|\right] = 1/(2\pi^2 r^2)$, so

$$\Delta^{+} \to \frac{\hbar c}{4\pi^2 r^2}\,, \tag{109.56}$$

an inverse square in the separation times $\hbar c$, of dimension J m, which delivers the J/m of $[\phi]^2$ as it must.

Proposition 109.88 (Dimensional counting fixes the leading power)**.** *If A, B and $\mathcal{O}_n$ have engineering mass dimensions d_A, d_B, d_n — defined so that an operator of dimension d scales as $(\text{length})^{-d}$ — and the theory is scale invariant at short distance, then*

$$C_n^{AB}(\xi) \;\propto\; |\xi|^{\,d_n - d_A - d_B + \gamma_n}\,, \tag{109.57}$$

where γ_n is the anomalous dimension of $\mathcal{O}_n$.

Rests on Definition 109.86.

Derivation. *Derives Proposition 109.88.*

Both sides of Equation (109.54) must scale the same way under $x \to \lambda x$. At a fixed point of the renormalization group the operator $\mathcal{O}_n$ scales with the power $d_n + \gamma_n$ rather than d_n, the correction being generated by the same loop effects that make the coupling run; matching powers gives Equation (109.57). The operator with the *smallest* $d_n + \gamma_n$ dominates as $\xi \to 0$, which is why the expansion is useful: an infinite sum is controlled by its first few terms at short distance, and the accuracy improves as the separation shrinks. The identity operator has $d = 0$ and no anomalous dimension, so it always leads, and the leading singularity is $|\xi|^{-d_A - d_B}$ as in Example 109.87, where $d_A = d_B = 1$ gives $|\xi|^{-2}$. □

Remark 109.89 (Status of the expansion). Within renormalized perturbation theory the expansion is established order by order: the technology is Zimmermann's convergence proof for momentum-space subtraction [1411], in which a normal-product expansion of a composite operator is defined and its short-distance behaviour controlled, and the expansion of Equation (109.54) follows for the theories treated there. Outside perturbation theory it remains a hypothesis in the Wightman framework, verified in soluble models and assumed elsewhere; the exception is the Euclidean setting, where it is a theorem in some constructed models. The anomalous dimensions appearing in Equation (109.57) are computed in practice by continuing the dimension of spacetime away from four and expanding in the deviation, the device Wilson and Fisher introduced for critical exponents [1489]; that is a calculational continuation of the same kind as the dimensional regularization of Chapter 100, not a claim about physics in a non-integer dimension.

An alternative to the whole regularization apparatus is the causal construction of Epstein and Glaser [1498], which builds the time-ordered products directly as distributions by extending them from the region of non-coincident points to the diagonal. There the only ambiguity is the finite freedom in that extension, and it is exactly the renormalization freedom; no divergence appears at any stage, because none was ever written down. It is the cleanest way to see that renormalization is a statement about products of distributions, not about infinities.

The applications are what make the expansion indispensable. The short-distance analysis of Chapter 102 uses it to separate a calculable coefficient from a non-perturbative matrix element, which is what makes a strongly coupled theory predictive at all; the deep-inelastic sum rules belong to Chapter 116; and the anomalous dimensions of Equation (109.57) are what the renormalization group of Chapter 108 organizes. In every case the logical structure is the same: Equation (109.54) converts a product of operators at nearly coincident points

— which is precisely the object Proposition 109.1 declared not to exist — into a sum of well-defined operators with singular numerical coefficients. The expansion is how the theory answers the objection that opened this chapter.

109.7 Constructive results and what is missing

109.7.1 Models below four dimensions

Remark 109.90 (Dimensional flag). This subsection names models in spacetimes of dimension two and three. They are cited as facts about the *axiomatic framework* — as evidence that the axioms of Section 109.2.1 are consistent and that non-trivial solutions of them exist at all — and never as physics of the observed world, which is $3+1$ and to which the dimensional rule of this treatise confines every physical claim. No result of this subsection is used anywhere else in the book, and none of the numbers below is a measurement of anything.

Glimm and Jaffe constructed the two-dimensional φ^4 theory without cutoffs [1371], verifying that the interacting Hamiltonian is self-adjoint and bounded below and that the removal of the ultraviolet and volume cutoffs leaves a theory satisfying the axioms. The three-dimensional case is harder because the coupling has positive mass dimension there, and the positivity of the Hamiltonian — the step that fails first — was established later [1500]. The programme and its methods are set out in [1792].

Example 109.91 (Which axioms are verified)**.** For the two-dimensional model, all of them: the Hilbert space and its positive metric, the unitary representation of the Poincaré group of that spacetime with spectrum in the forward cone, a unique vacuum, fields as operator-valued distributions on a common domain, covariance and local commutativity, and cyclicity. The scattering theory was constructed and the theory shown to be non-trivial, in the sense that the S matrix is not the identity. For the three-dimensional model the same list holds with more work; in both cases the route is Euclidean, through Theorem 109.78, so what is actually built is a measure and what is verified is reflection positivity.

Remark 109.92 (Why four dimensions is different, in one sentence). The number of ultraviolet divergences a theory has is governed by the mass dimension of its coupling, and for φ^4 that dimension falls from positive in two and three dimensions — where only finitely many diagrams diverge and a finite number of subtractions suffices — to exactly zero in four, where every order diverges and the subtractions must be organized to all orders. That is the whole difference, and it is the same marginality that makes the coupling run logarithmically (Chapter 108) and that Section 109.7.2 is about.

109.7.2 Triviality of the four-dimensional scalar

For the simplest candidate in four dimensions the evidence points not merely to difficulty but to impossibility.

Theorem 109.93 (Triviality of φ^4; imported)**.** *The continuum limit of the lattice φ^4 theory, taken along the critical line, is a free field: in more than four spacetime dimensions by the correlation-inequality proofs of Aizenman [1662] and, independently, Fröhlich [1688]; and in the marginal case of exactly four dimensions by Aizenman and Duminil-Copin [2371], who show that the scaling limits of the critical four-dimensional Ising and φ^4 models are Gaussian.*

Rests on Theorem 109.78.

Derivation pending #189

Triviality of the four-dimensional scalar: the correlation-inequality argument bounding the connected four-point function of the lattice model above dimension four, and the tree-diagram bound that settles the marginal four-dimensional case. Both are results in classical probability rather than in field theory and are quoted here with their citations; they belong in Appendix A. What is derived in the text is the consequence, the scale at which the isolated scalar coupling diverges.

Remark 109.94 (What "trivial" means and does not mean). It does not mean that a scalar field theory is useless, nor that lattice φ^4 at finite spacing is free — it is not, and its correlation functions are perfectly non-Gaussian. It means that the *limit* of removing the cutoff while holding the physical mass fixed drives the renormalized coupling to zero. A self-interacting scalar in $3+1$ dimensions can therefore exist only as an effective theory carrying a finite cutoff, with its interaction strength tied to how far away that cutoff sits.

Proposition 109.95 (Where the cutoff must sit for the Higgs)**.** *Take the Standard Model scalar sector in isolation, dropping the gauge and Yukawa couplings, so that Equation* (104.72) *reduces to* $16\pi^2 \mathrm{d}\hat{\lambda}/\mathrm{d}\ln\mu = 24\hat{\lambda}^2$*. Then* $\hat{\lambda}$ *diverges at*

$$\mu_L = \mu_0 \exp\!\left(\frac{2\pi^2}{3\hat{\lambda}(\mu_0)}\right), \tag{109.58}$$

and with the measured value $\hat{\lambda} = 0.126$ *at* $\mu_0 = m_t c^2 = 172.57\,\mathrm{GeV}$ *[2427] this gives* $\mu_L \approx 8 \times 10^{24}\,\mathrm{GeV}$*, some* 6.7×10^5 *times the Planck energy* $1.22 \times 10^{19}\,\mathrm{GeV}$*.*

Rests on Equation (104.72)*.*

Derivation. *Derives Proposition 109.95.*

Write $\beta = \mathrm{d}\hat{\lambda}/\mathrm{d}\ln\mu = 24\hat{\lambda}^2/(16\pi^2) = 3\hat{\lambda}^2/(2\pi^2)$. Separating variables, $\mathrm{d}\hat{\lambda}/\hat{\lambda}^2 = 3\,\mathrm{d}\ln\mu/(2\pi^2)$, so

$$\frac{1}{\hat{\lambda}(\mu_0)} - \frac{1}{\hat{\lambda}(\mu)} = \frac{3}{2\pi^2}\ln\frac{\mu}{\mu_0},$$

and the right-hand side reaches $1/\hat{\lambda}(\mu_0)$ — at which point $1/\hat{\lambda}(\mu) = 0$ — when $\ln(\mu/\mu_0) = 2\pi^2/\big(3\hat{\lambda}(\mu_0)\big)$, which is Equation (109.58). Numerically $2\pi^2/(3 \times 0.126) = 52.2201$ and $\mathrm{e}^{52.2201} = 4.7743 \times 10^{22}$, so $\mu_L = 172.57 \times 4.7743 \times 10^{22} = 8.24 \times 10^{24}\,\mathrm{GeV}$; the fourth figure of the exponential is carried because rounding it to 4.77×10^{22} would move the answer to 8.23×10^{24}. The Planck energy is $E_P = \sqrt{\hbar c^5/G} = 1.956 \times 10^9\,\mathrm{J} = 1.221 \times 10^{19}\,\mathrm{GeV}$ from $\hbar = 1.054\,571\,817 \times 10^{-34}\,\mathrm{J\,s}$, $c = 299\,792\,458\,\mathrm{m/s}$ and $G = 6.674\,30 \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$ [2431], giving the stated ratio. □

Remark 109.96 (How much this argument is worth). Less than the arithmetic suggests, and the honest accounting is short. The one-loop pure-scalar running used above is not the Standard Model running: the top Yukawa term $-6\hat{y}_t^4$ in Equation (104.72) outweighs $24\hat{\lambda}^2$ by a factor of order ten at m_t, so the actual quartic coupling runs *down* and crosses zero near $10^{10}\,\mathrm{GeV}$ (Remark 104.67) rather than blowing up at $10^{25}\,\mathrm{GeV}$. Equation (109.58) is therefore not a statement about the Higgs boson; it is a statement about the isolated scalar theory to which Theorem 109.93 applies, and it is quoted here to show that even in that theory the triviality obstruction is astronomically remote from any accessible scale.

What survives is the point of principle. A pure φ^4 theory in $3+1$ dimensions cannot be a fundamental theory: either it carries a cutoff, or it is free. The parallel obstruction on the other side of quantum electrodynamics is the Landau pole of the running fine structure constant, the same phenomenon in a theory that is also not asymptotically free; both belong to Chapter 108, which is where the running is organized. That the strong coupling of

Chapter 102 runs the other way — to zero at short distance — is exactly why Yang–Mills theory, and not the scalar, is the candidate for a four-dimensional construction.

109.7.3 The Yang–Mills mass gap

Definition 109.97 (The open problem, stated precisely)**.** Prove that for any compact simple gauge group G a non-trivial quantum Yang–Mills theory exists on $\mathbb{R}^4$ — that is, a family of Wightman functions satisfying Axiom 109.9 to Axiom 109.15, or equivalently a family of Schwinger functions satisfying the hypotheses of Theorem 109.78 — whose classical limit is the Yang–Mills action, and prove that it has a mass gap: a number $\Delta > 0$ such that every state other than the vacuum has energy at least Δ in the rest frame [2097].

Remark 109.98 (Why one end is tractable and the other is not)*.* The ultraviolet end is favourable. Asymptotic freedom (Section 102.4) makes the coupling vanish logarithmically at short distance, so the theory becomes weakly coupled exactly where the divergences are, and the continuum limit is approached along a trajectory on which perturbation theory is a good guide — the opposite of the situation in Section 109.7.2, where the coupling grows at short distance and the limit is forced to zero. This is the strongest structural reason to expect that a four-dimensional interacting theory exists at all.

The infrared end is where everything is hard, and the mass gap is an infrared statement. The Lagrangian contains no mass; the gap, if it exists, is generated dynamically and is a pure number times Λ_{QCD}, which Equation (102.45) defines as an energy — matching Δ, which Definition 109.97 likewise states as an energy. The corresponding mass is $\Delta/c^2 = M$ in the notation of Proposition 109.83. No expansion is available there — the coupling is large, the strong coupling expansion of the lattice is valid only at spacings far from the continuum, and no controlled interpolation between the two regimes has been constructed.

Remark 109.99 (The evidence, and why it is not a proof)*.* Three independent lines of evidence say the gap exists.

No massless hadron. The observed spectrum begins at the pion, of rest energy $(139.570\,39 \pm 0.000\,18)$ MeV [2427], and the pion is light for a reason the theory explains — it is the pseudo-Goldstone boson of chiral symmetry breaking (Chapter 102) — not because the spectrum approaches zero. Nothing massless has ever been seen carrying colour or built from gluons, and a massless strongly interacting particle would produce a long-range force that is conspicuously absent (Example 109.84).

The glueball spectrum. Lattice computations in the pure gauge theory, which has no quarks and hence no chiral mechanism, find a lowest state of non-zero mass. That is the cleanest numerical statement of the gap, since in that theory the gap is the whole question.

The light hadron spectrum from first principles. Taking as input only the quark masses and the coupling, [2122] reproduces the measured masses at the percent level (Phenomenon 102.64), which would be impossible if the theory did not have the spectrum it is being asked to have.

None of this is a proof, and the distinction is not pedantry. Every one of these is a computation at finite lattice spacing and finite volume, extrapolated; the extrapolations are controlled by fits with estimated errors, not by bounds; and the statement to be proved is a statement about a limit that has not been shown to exist. The treatise's own rules make the position sharp: the confinement of quarks and the mass gap are observed facts of Nature for which no derivation exists, and they are recorded as such in Section 102.7 and Section 133.6 rather than being quietly upgraded to results.

109.7.4 What axiomatics does and does not secure

The balance sheet, drawn honestly.

Secured. The *PCT* theorem (Theorems 109.44 and 109.46), from locality, covariance and the spectrum condition alone — so that the measurements of Section 105.5.3 test those three axioms and nothing else. The spin–statistics connection (Theorem 109.49), which discharges the debt owed by Chapters 83, 102, 107 and 120 and underwrites the exclusion principle, the periodic table and degeneracy pressure. The Källén–Lehmann representation and the bound $0 \le Z \le 1$ (Theorem 109.54 and Corollary 109.55), exact rather than perturbative, which every renormalized propagator in Chapter 100 presumes. The Reeh–Schlieder theorem and its two faces, vacuum entanglement and the impossibility of a local operator annihilating the vacuum (Theorem 109.34 and Corollary 109.35). Cluster decomposition as the equivalent of a unique vacuum (Proposition 109.23). The irrelevance of the choice of interpolating field (Theorem 109.38), assumed without comment wherever a field is redefined. And the structural results of Section 109.4: that superselection sectors carry the representation theory of a compact group and that the vacuum's modular flow in a wedge is a geometric boost.

Not secured. The existence of any interacting model satisfying the axioms in $3 + 1$ dimensions, the Standard Model included. Nothing in Part XI has been shown to be a theory in the sense of Section 109.2.1; what has been shown is that a particular calculational scheme, order by order in a coupling, reproduces measurement to extraordinary accuracy. Asymptotic completeness — that the scattering states exhaust the Hilbert space — is an assumption of Chapter 86 and not a theorem. The mass gap of Definition 109.97 is open. And for the simplest four-dimensional candidate the answer is known and is negative (Theorem 109.93).

Remark 109.100 (The position this book takes). The Standard Model is an extraordinarily well-tested effective theory whose mathematical existence is unproven. Both halves of that sentence are load-bearing. The first is why Part XI computes with it without apology: agreement with the electron's anomalous moment at about one part in 10^9 of a_e, which Chapter 112 shows is the precision of the input α rather than of the measurement, is evidence of a kind that no foundational anxiety outweighs. The second is why this chapter exists and why Section 133.6 lists the gap among the open problems rather than among the technicalities.

Neither half licenses the third position sometimes taken, that because the theory does not exist mathematically it should be replaced. There is no evidence-based replacement — the programmes usually named in this connection are excluded from this treatise for exactly that reason (Section 133.7) — and a theory that predicts to nine significant figures is not overthrown by the absence of a proof that it exists. The correct response is the one the constructive programme has been making for sixty years: build the object, and see.

A survey of the state of the subject as it stood after its first two decades, still the clearest short account of what the axioms deliver, is [1562].

Chapter 110

Experiment: Parity Violation

Tests Phenomena 101.26, 104.55 and 105.16.

Assuming Propositions 101.15, 101.27, 105.19 and 105.20.

Parity conservation was, until 1956, treated as a law of nature rather than as a hypothesis under test. This chapter records the experiment that ended that status: the cobalt-60 beta-decay asymmetry measured by Wu, Ambler, Hayward, Hoppes and Hudson [1133], prompted by the observation of Lee and Yang [1091] that not one existing measurement bore on parity in the weak interaction, and confirmed within days by the pion–muon–electron chain [1109, 1110]. The chapter sits immediately after the weak-interaction theory of Chapter 101 and the discrete-symmetry formalism of Chapter 105, because it is the measurement those chapters exist to explain; it is the direct experimental root of the two-component neutrino of Chapter 96, of the helicity measurement in Chapter 98, and of the chiral gauge structure that Chapter 104 builds on.

The division of labour with those chapters is worth stating at the outset, because it decides what is and is not repeated here. The formal apparatus — the space-inversion operator on states and on Dirac fields, the classification of bilinears, the $V - A$ current, the relation between chirality and helicity — is derived in Sections 101.4 and 105.2 and is used here, not rebuilt. What this chapter owes is the measurement: what the source physically was, how cold it had to be and why, what the counters saw, what numbers came out with what uncertainties, and by what controls a counting-rate difference of a few tens of per cent was shown to be a property of the weak interaction rather than of a photomultiplier in a magnetic field. Its second half follows the same pseudoscalar observable out of nuclear beta decay and into neutral-current physics — atomic parity violation [1519, 1946], polarized deep-inelastic scattering [1625, 1643] and the weak charge of the proton [2301] — where parity violation has become a precision tool for measuring the weak mixing angle rather than a surprise.

110.1 Historical context and the prediction under test

110.1.1 Parity as a conserved quantum number

Space inversion is the map

$$P : \boldsymbol{x} \longmapsto -\boldsymbol{x}\,, \tag{110.1}$$

implemented on quantum states by a unitary operator $\mathcal{P}$ with $\mathcal{P}\psi(\boldsymbol{x},t) = \eta_P\,\psi(-\boldsymbol{x},t)$. Applying it twice returns every position to itself, so $\mathcal{P}^2$ is a symmetry operation that does nothing; with the usual convention that it is the identity, $\mathcal{P}^2 = \mathbb{1}$ and the eigenvalues of $\mathcal{P}$ are ± 1. That is the whole of the kinematics, and it is why parity is a two-valued quantum number rather than a continuous one: there is no infinitesimal space inversion, so parity has no associated conserved current and no Noether charge, only a multiplicative selection rule.

Wigner introduced parity as a conserved quantum number of atomic states in 1927 [592], and the evidence for its conservation in the electromagnetic interaction was, and remains, overwhelming. It is carried chiefly by Laporte's rule: an electric dipole transition connects only atomic states of opposite parity, because the dipole operator $-e\boldsymbol{x}$ is itself parity-odd. The rule is derived as Proposition 105.1; here we need only its status as evidence. Every line in every atomic spectrum that obeys it — and every line whose absence it explains — is a measurement that the electromagnetic Hamiltonian commutes with $\mathcal{P}$, and the accumulated precision of atomic and nuclear spectroscopy by 1956 bounded any parity-violating admixture in the nuclear force at the level of 10^{-7} in amplitude [1091].

The observables of a physical theory divide, under Equation (110.1), into four classes according to how they transform. A *scalar* is unchanged; a *pseudoscalar* changes sign; a *vector* (or polar vector) reverses; an *axial vector* (or pseudovector) does not. The position $\boldsymbol{x}$, the momentum $\boldsymbol{p} = m\mathrm{d}\boldsymbol{x}/\mathrm{d}t$ and the electric field $\boldsymbol{E}$ are polar; the angular momentum $\boldsymbol{J} = \boldsymbol{x} \times \boldsymbol{p}$, the spin $\boldsymbol{S}$ and the magnetic field $\boldsymbol{B}$ are axial, because a cross product of two polar vectors acquires two sign changes and therefore none. Consequently $\boldsymbol{J} \cdot \boldsymbol{p}$, $\boldsymbol{S} \cdot \boldsymbol{p}$ and $\boldsymbol{E} \cdot \boldsymbol{B}$ are pseudoscalars. The complete table of Dirac bilinears and their behaviour is Table 105.1.

The entire experimental programme of this chapter rests on one statement about pseudoscalars, and it is worth isolating it because everything else is instrumentation.

Proposition 110.1 (A pseudoscalar mean vanishes if parity holds)**.** *Let Ω be a pseudoscalar observable, $\mathcal{P}\Omega\mathcal{P}^{-1} = -\Omega$, and let ρ be the state of a system prepared by an interaction that commutes with $\mathcal{P}$ from an initial state that is itself parity invariant, $\mathcal{P}\rho\mathcal{P}^{-1} = \rho$. Then*

$$\langle \Omega \rangle = \operatorname{tr}(\rho\,\Omega) = 0\,. \tag{110.2}$$

A measured $\langle\Omega\rangle \neq 0$ therefore proves that the interaction responsible does not commute with $\mathcal{P}$, with no dynamical input of any kind.

Rests on Equation (110.1)*.*

Proof. *Derives Proposition 110.1.*

Insert $\mathbb{1} = \mathcal{P}^{-1}\mathcal{P}$ twice under the trace and use its cyclic property:

$$\langle \Omega \rangle = \operatorname{tr}(\rho\,\Omega) = \operatorname{tr}\Big(\mathcal{P}\rho\mathcal{P}^{-1}\,\mathcal{P}\Omega\mathcal{P}^{-1}\Big) = \operatorname{tr}(\rho\,(-\Omega)) = -\,\langle \Omega \rangle\ ,$$

whence $\langle\Omega\rangle = 0$. The two hypotheses are exactly the two things an experiment must arrange or verify: that the state was prepared without reference to a handedness — a magnetic field is axial, so polarizing a sample with a solenoid does *not* introduce one — and that Ω is genuinely parity-odd. Nothing is assumed about the Hamiltonian beyond $[\mathcal{P}, H] = 0$, which is why one number settles a question that decades of spectroscopy had not thought to ask. □

Remark 110.2 (Why a magnetic field is a legitimate preparation)*.* The point deserves emphasis because it is the commonest misreading of Wu's experiment. Aligning nuclear spins with a solenoid does not smuggle a handedness into the apparatus: $\boldsymbol{B}$ is an axial vector, so the mirror image of the polarized source is a source polarized *in the same direction* with the current in the coil reversed — and reversing a current is not a physical change of the state prepared, only of the wiring. The mirror image of "spins up, electrons preferentially down" is "spins up, electrons preferentially up". Those two statements are about the same physical preparation and they disagree, which is the whole content of the result.

110.1.2 The tau–theta puzzle

The pressure to test parity came from two strange mesons. The θ^+ decayed to $\pi^+\pi^0$ and the τ^+ to $\pi^+\pi^+\pi^-$; by 1955 their masses agreed to better than a per cent and their

lifetimes to within the experimental errors, and the modern values — one particle, the K^+, with $m_K c^2 = (493.677 \pm 0.015)\,\mathrm{MeV} = (7.9095 \pm 0.0002) \times 10^{-11}\,\mathrm{J}$ and mean life $\tau_K = (1.2380 \pm 0.0020) \times 10^{-8}\,\mathrm{s}$ [2427] — say why the coincidence was intolerable. But the parities of the two final states are opposite: by Equation (105.31) a state of n pions with no relative orbital angular momentum has parity $(-1)^n$, so $\eta_P(\pi\pi) = +1$ while $\eta_P(\pi\pi\pi) = -1$.

Only two escapes existed. Either θ and τ were genuinely distinct particles with an unexplained mass and lifetime degeneracy, or the three-pion decay carried a unit of relative orbital angular momentum, in which case the parent would have spin ≥ 1 and the parities could agree. Dalitz closed the second by inventing the plot that bears his name [1016]: for a three-body decay at rest the density of events in the (T_1, T_2) plane of two of the kinetic energies is the squared matrix element and nothing else, so a uniform population is a direct measurement that the decay amplitude is constant — no relative orbital angular momentum, no spin. The observed τ plot was uniform. The construction and its derivation are given at Section 105.2.1, which also carries the phase-space calculation behind Equation (105.32); the point here is only that by 1956 the escape route through angular momentum had been closed experimentally, not assumed away.

110.1.3 Lee and Yang: parity untested in the weak interaction

Lee and Yang responded with an audit rather than a mechanism [1091]. They asked what the experimental evidence for parity conservation actually consisted of, and found that all of it came from strong and electromagnetic processes. In the weak interaction there was none: not one experiment had ever been designed to detect a parity-odd observable in a weak decay, and the reason was structural rather than accidental. Every weak measurement then existing was a lifetime, a branching ratio or an energy spectrum, and each of those is a scalar; by Proposition 110.1 a scalar is insensitive to parity violation at leading order, since the parity-odd part of the interaction contributes to a scalar only through its square. A theory in which parity was violated maximally would have reproduced every weak number known in 1956.

What makes the paper a model of method is its second half, which lists the experiments that would settle the question. Each proposal is the mean of a pseudoscalar:

- the correlation $\langle \boldsymbol{J} \rangle \cdot \boldsymbol{p}_e$ between the polarization of an oriented nucleus and the momentum of the emitted electron — an axial vector dotted into a polar one;
- the longitudinal polarization $\langle \boldsymbol{s} \rangle \cdot \boldsymbol{p}$ of the electron emitted in an unoriented beta decay, which requires no cryogenics but a polarimeter;
- in the chain $\pi^+ \to \mu^+ \nu_\mu$ followed by $\mu^+ \to e^+ \nu_e \bar{\nu}_\mu$, the correlation between the muon direction and the positron direction, which is nonzero only if the muon is produced longitudinally polarized.

All three were carried out within a year of the paper, by three independent groups with no common apparatus, and all three returned the largest answer the kinematics permits. The first is the subject of this chapter; the second was done by Frauenfelder and collaborators with Mott-scattering polarimetry on the electrons from ^{60}Co itself [1108] and by Schopper through the circular polarization of the accompanying gamma rays [1131]; the third is Section 110.6 below. Lee and Yang received the 1957 Nobel prize, in the same year as the measurements they proposed.

110.1.4 The observable and what a measurement must deliver

The observable of the cobalt experiment is the correlation between the nuclear polarization and the electron momentum. Write $\hat{\boldsymbol{n}}$ for the unit vector along the mean nuclear spin and $\hat{\boldsymbol{p}}_e$

for the electron direction. The most general angular distribution linear in that correlation is

$$W(\theta) \propto 1 + A\,\frac{v}{c}\,f\cos\theta\,, \qquad \cos\theta = \hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{p}}_e\,, \tag{110.3}$$

in which three factors appear that the experiment must supply separately, and the failure to supply any one of them makes the number uninterpretable:

1. the *asymmetry coefficient* A, the quantity of physical interest, dimensionless, and predicted by the theory of the interaction;
2. the *degree of nuclear polarization* $f = |\langle \boldsymbol{J}\rangle|\,/J$, a property of the source and its temperature, which must be measured independently and which decays to zero as the source warms;
3. the *electron speed* v/c, which enters because the chirality projected out by the interaction is converted into helicity only up to a factor v/c (Proposition 105.20), and which for a continuous beta spectrum means an average weighted by the counter's acceptance.

What a counter measures is not $W(\theta)$ but a rate, and rates depend on solid angles, detector efficiencies and source strength — all of which drift. The design of the experiment therefore forms the observable as a ratio in which those cancel.

Proposition 110.3 (The measured quantity)**.** *Let $N_\uparrow$ and $N_\downarrow$ be the counts recorded in one fixed counter during equal intervals with the polarizing field in the two opposite senses. Then*

$$\varepsilon \equiv \frac{N_\uparrow - N_\downarrow}{N_\uparrow + N_\downarrow} = A\,\frac{v}{c}\,f\;\langle\cos\theta\rangle\;, \tag{110.4}$$

where $\langle\cos\theta\rangle$ is the average of $\cos\theta$ over the counter's acceptance and equals 1 for a counter subtending a small solid angle on the polarization axis. Every geometric and efficiency factor cancels in the ratio, and the statistical uncertainty on ε is $\sigma_\varepsilon = 1/\sqrt{N}$ with $N = N_\uparrow + N_\downarrow$.
Rests on Equation (110.3).

Proof. *Derives Proposition 110.3.*
Reversing the solenoid current reverses $\hat{\boldsymbol{n}}$ and hence the sign of $\cos\theta$ at fixed counter position, so with a common normalization K containing the solid angle, the efficiency and the source activity,

$$N_\uparrow = K\big(1 + A\tfrac{v}{c}f\,\langle\cos\theta\rangle\big)\;, \qquad N_\downarrow = K\big(1 - A\tfrac{v}{c}f\,\langle\cos\theta\rangle\big)\;.$$

Forming the ratio Equation (110.4) cancels K identically. For the error, write $\varepsilon = (N_\uparrow - N_\downarrow)/N$ with N fixed; the numerator is a difference of Poisson counts of variance $N_\uparrow + N_\downarrow = N$, so $\sigma_\varepsilon^2 = N/N^2 = 1/N$ up to a correction of relative order ε^2, negligible here. □

The last statement fixes the scale of the whole enterprise. Reaching a statistical precision of $1\,\%$ on ε requires $N = 10^4$ counts, and — because the polarization survives only a few minutes — those counts must be accumulated in of order $60\,\mathrm{s}$. The counter must therefore see a rate of order $2\times10^2/\mathrm{s}$. With a solid-angle fraction of a few per cent and an escape-and-detection efficiency of order one half, the source activity must be at least

$$\mathcal{A} \gtrsim \frac{2\times10^2/\mathrm{s}}{0.05\times0.5} \approx 1\times10^4\,\mathrm{Bq}\,, \tag{110.5}$$

which is a modest laboratory source but a hard lower bound: a weaker one cannot deliver the statistics inside the warm-up window however long the experiment runs, because the signal is gone before the counts accumulate. Equation (110.5) is the reason the measurement had to wait for a source preparation technique and a cryostat that were compatible with each other.

110.2 Apparatus

110.2.1 The polarized cobalt-60 source

Table 110.1: Properties of the ^{60}Co source and of its decay, in SI units with the customary nuclear units alongside. Energies are given in joules and in electronvolts because the two appear throughout the experimental literature; the half-life and the gamma energies are standard nuclear data, the mass-difference input being the atomic-mass evaluation [2393]. The nuclear magnetic moment is quoted in units of the nuclear magneton $\mu_N = (5.050\,783\,739\,3 \pm 0.000\,000\,001\,6) \times 10^{-27}\,\mathrm{J/T}$ [2431].

Quantity	Value (SI)	Customary form
Parent spin and parity	—	5^+
Daughter level fed	—	4^+ of ^{60}Ni
Half-life	$1.6635 \times 10^{8}\,\mathrm{s}$	$(5.2714 \pm 0.0005)\,\mathrm{yr}$
Decay constant	$4.1667 \times 10^{-9}\,/\mathrm{s}$	—
Beta endpoint kinetic energy	$5.098 \times 10^{-14}\,\mathrm{J}$	$318.2\,\mathrm{keV}$
First cascade gamma	$1.8797 \times 10^{-13}\,\mathrm{J}$	$1173.23\,\mathrm{keV}$
Second cascade gamma	$2.1349 \times 10^{-13}\,\mathrm{J}$	$1332.49\,\mathrm{keV}$
Nuclear magnetic moment	$1.919 \times 10^{-26}\,\mathrm{J/T}$	$3.799\,\mu_N$
Minimum activity required, Equation (110.5)	$1 \times 10^{4}\,\mathrm{Bq}$	—
Corresponding number of ^{60}Co atoms	2.4×10^{12}	—

The choice of nucleus is not incidental; it is forced by four requirements that almost no other nuclide satisfies at once, and Table 110.1 collects the numbers.

First, the transition must be *allowed and pure Gamow–Teller*, so that the predicted asymmetry coefficient is the maximal $A = -1$ of Proposition 101.27 rather than a mixture of Fermi and Gamow–Teller amplitudes with an unknown ratio. The decay $^{60}\mathrm{Co}\,(5^+) \to {}^{60}\mathrm{Ni}^*\,(4^+)$ has $\Delta J = 1$ with no parity change, which by the allowed selection rules of Proposition 101.15 is pure Gamow–Teller: a Fermi transition cannot change J at all.

Second, the parent spin must be *large*. The polarization attainable at a given temperature grows with J, because the Zeeman ladder has $2J + 1$ rungs and the Boltzmann factor acts across all of them; $J = 5$ is among the largest available in a convenient beta emitter.

Third, the decay must be followed by a *gamma cascade*, because that cascade is the only practical thermometer for the nuclear orientation. The 4^+ daughter level de-excites through $4^+ \to 2^+ \to 0^+$, emitting photons of $1173.23\,\mathrm{keV}$ and $1332.49\,\mathrm{keV}$, and the angular distribution of those photons relative to the orientation axis is anisotropic whenever the parent nuclei are aligned. Without an independent measurement of f, Equation (110.4) determines only the product Af, and the experiment would establish that parity is violated without measuring by how much.

Fourth, the half-life must be long enough that the source does not decay away during preparation and short enough that a usable activity fits in a thin layer. At $(5.2714 \pm 0.0005)\,\mathrm{yr} = 1.6635 \times 10^{8}\,\mathrm{s}$ the decay constant is $\lambda = \ln 2 / t_{1/2} = 4.1667 \times 10^{-9}\,/\mathrm{s}$, so Equation (110.5) is met by $N = \mathcal{A}/\lambda \approx 2.4 \times 10^{12}$ atoms, a quantity of ^{60}Co of order $2.4 \times 10^{-10}\,\mathrm{g}$. The active material is therefore a trace, and the layer thickness is set by beta absorption rather than by chemistry.

Remark 110.4 (The endpoint energy quoted in 1957). The 1957 literature quotes a beta endpoint of $0.312\,\mathrm{MeV}$; the modern value, from the mass difference of [2393] less the

excitation energy of the daughter level, is 0.3182 MeV. The two per cent shift is irrelevant to the result — it moves v/c by about half a per cent — but it is recorded here so that a reader comparing this chapter with the original paper is not left wondering which number is wrong.

Why the active layer must be thin The source is a layer of ^{60}Co grown on the surface of the paramagnetic salt crystal, and its thickness is the leading systematic. An electron created inside the layer loses energy before it escapes, which lowers v/c and hence, by Equation (110.4), the observed asymmetry; a thick layer also scatters electrons, which randomizes $\cos\theta$ and dilutes the asymmetry a second time. Both effects reduce $|\varepsilon|$, so they cannot manufacture the result — but they do prevent the extraction of A.

Proposition 110.5 (Tolerable source thickness)**.** *For a beta electron of kinetic energy near* 150 keV*, which is representative of the counted part of the* ^{60}*Co spectrum, the practical range in a solid of density* $\rho \approx 2\times10^{3}\,\mathrm{kg/m^3}$ *is about* $1.3\times10^{-4}\,\mathrm{m}$*. A source layer of a few tens of micrometres therefore costs the average escaping electron of order* 10 keV *and a tenth of its range, which is tolerable; a layer of a hundred micrometres would not be.*

Rests on Equation (110.4)*.*

Derivation. *Derives Proposition 110.5.*

The empirical range–energy relation for electrons in the range 0.01–3 MeV, expressed as a mass thickness $R\rho$ in $\mathrm{kg/m^2}$ with the kinetic energy T in MeV, is

$$R\rho = 4.12\,\mathrm{kg/m^2}\;T^{\,1.265-0.0954\ln T}\,. \tag{110.6}$$

At $T=0.15$ the exponent is $1.265+0.181=1.446$ and $T^{1.446}=0.0644$, giving $R\rho = 0.265\,\mathrm{kg/m^2}$; at $\rho=2\times10^{3}\,\mathrm{kg/m^3}$ this is $R=1.33\times10^{-4}\,\mathrm{m}$. At the endpoint $T=0.318$ the same formula gives $R\rho=0.855\,\mathrm{kg/m^2}$, i.e. $R=4.3\times10^{-4}\,\mathrm{m}$. A layer of thickness $d=3\times10^{-5}\,\mathrm{m}$ presents an average path of $d/2$ to an isotropically emitted electron that escapes forwards, that is $0.03\,\mathrm{kg/m^2}$, about 11 % of the 150 keV range. Inverting Equation (110.6) for the energy that has $0.235\,\mathrm{kg/m^2}$ of range left gives $T=139.4$ keV, so the mean energy loss is of order 11 keV, and by $\mathrm{d}(v/c)/(v/c)=(T+m_ec^2)^{-1}(1-v^2/c^2)\,\mathrm{d}T\cdot(c/v)^2$ this shifts $v/c=0.63$ downwards by about 2.5 % — the same figure the exact ratio $\beta(139.4\,\mathrm{keV})/\beta(150\,\mathrm{keV})=0.6186/0.6343$ gives. That is the size of the correction the source thickness imposes on the extraction of A from Equation (110.4), and it is why the paper's conclusion is stated as "$|A|$ consistent with the maximum" rather than as a number with a small error bar. □

110.2.2 Adiabatic demagnetization and the cryostat

Nuclear orientation requires that the Zeeman splitting of the nuclear sublevels exceed the thermal energy, and that single requirement dictates the entire cryogenic apparatus.

Proposition 110.6 (The orientation condition)**.** *For a nucleus of spin* I *and magnetic moment* μ *in a field* B *at temperature* T*, the degree of polarization* $f=\langle m\rangle/I$ *is a function of the single dimensionless combination*

$$x=\frac{\mu B}{I\,k_B T}\,, \tag{110.7}$$

namely the Brillouin function $f=B_I(Ix)$*. Reaching* $f\approx0.6$ *for* ^{60}*Co, which has* $I=5$ *and* $\mu=3.799\,\mu_N=1.919\times10^{-26}\,\mathrm{J/T}$*, requires* $x=0.378$*, that is*

$$\frac{B}{T}=\frac{x\,I\,k_B}{\mu}\approx 1.4\times10^{3}\,\mathrm{T/K}\,. \tag{110.8}$$

Rests on Equation (79.5)*.*

Derivation. *Derives Proposition 110.6.*

The Hamiltonian of the nuclear moment in the field is $H = -\boldsymbol{\mu}\cdot\boldsymbol{B}$, with eigenvalues $E_m = -(\mu/I)\,B\,m$ for $m = -I, \dots, I$, so the Boltzmann weight of the mth sublevel is e^{xm} with x as in Equation (110.7). Hence

$$f = \frac{1}{I}\,\frac{\sum_{m=-I}^{I} m\,\mathrm{e}^{xm}}{\sum_{m=-I}^{I}\mathrm{e}^{xm}} = \frac{1}{I}\,\frac{\partial}{\partial x}\ln\sum_{m=-I}^{I}\mathrm{e}^{xm}\,, \tag{110.9}$$

which is the Brillouin function and depends on B and T only through their ratio. Evaluating the sum numerically for $I = 5$ gives $f = 0.196$ at $x = 0.1$, $f = 0.513$ at $x = 0.3$, $f = 0.600$ at $x = 0.378$ and $f = 0.837$ at $x = 0.8$. Solving $f = 0.6$ and inserting $I = 5$, $k_B = 1.380\,649 \times 10^{-23}\,\mathrm{J/K}$ [2431] and $\mu = 1.919 \times 10^{-26}\,\mathrm{J/T}$ gives Equation (110.8). □

Equation (110.8) is a severe constraint and it is worth reading it twice. A laboratory solenoid delivering $B = 1\,\mathrm{T}$ would need $T = 0.7\,\mathrm{mK}$, below what any refrigerator of the period could reach and below the ordering temperature of the salt. The way out, and the reason the experiment was possible at all, is that the field acting on the *nucleus* is not the applied field: in a paramagnetic salt containing the active ion, the hyperfine interaction with the ion's own electron shell produces an effective field at the nucleus of order $10\,\mathrm{T}$ once the electronic moments are themselves aligned, and the electronic moments are aligned by a field of a few millitesla because the electronic moment is three orders of magnitude larger than the nuclear one. The applied field steers the electrons; the electrons polarize the nuclei. With $B \sim 10\,\mathrm{T}$, Equation (110.8) demands only $T \lesssim 7\,\mathrm{mK}$.

Cooling by adiabatic demagnetization The millikelvin range was reached by the one technique available in 1957, and its principle is a single line of thermodynamics.

Proposition 110.7 (Adiabatic demagnetization)**.** *For a system of non-interacting magnetic moments the entropy is a function of B/T alone. An isentropic reduction of the field from B_i to a residual internal field b therefore cools the sample as*

$$T_f = T_i\,\frac{b}{\sqrt{B_i^2 + b^2}} \approx T_i\,\frac{b}{B_i} \qquad (b \ll B_i)\,. \tag{110.10}$$

Starting from $T_i = 1.2\,\mathrm{K}$ in $B_i = 1\,\mathrm{T}$ and ending in an internal field $b \approx 3\,\mathrm{mT}$ gives $T_f \approx 3.6\,\mathrm{mK}$.

Rests on Equation (110.7).

Derivation. *Derives Proposition 110.7.*

The single-moment partition function is $Z(B,T) = \sum_m \exp(\mu_e B m/(J_e k_B T))$, which depends on B and T only through B/T. Every thermodynamic quantity derived from $\ln Z$ without an explicit factor of T is therefore a function of B/T, and in particular the entropy $S = k_B(\ln Z - (B/T)\,\partial \ln Z/\partial(B/T))$ is $S = S(B/T)$. Holding S fixed holds B/T fixed, so $T_f/T_i = B_f/B_i$. The demagnetization cannot be carried to $B_f = 0$, because the moments interact: dipolar and exchange couplings inside the crystal act as a residual internal field b, which enters in quadrature since it is randomly oriented with respect to the applied one, giving Equation (110.10). The salt of choice, cerium magnesium nitrate, is used precisely because its internal field is exceptionally small — its magnetic ordering temperature is of order $2\,\mathrm{mK}$, and no other common paramagnetic salt orders so low, so none can be demagnetized so far. □

The sequence is thus: cool the salt to $1.2\,\mathrm{K}$ against a pumped liquid-helium bath, apply about $1\,\mathrm{T}$ and let the heat of magnetization flow out through helium exchange gas, pump the exchange gas away to isolate the salt thermally, then remove the field. The crystal arrives at a few millikelvin thermally disconnected from everything, which is both the point and the limitation.

The heat leak and the observation window Because the sample is isolated, its temperature rises under any residual heat leak, and the polarization decays with it. The measured decay time is about six minutes, and it fixes the leak.

Proposition 110.8 (Size of the residual heat leak)**.** *A warm-up from* 3–10 mK *in* 3.6×10^{2} s *for a crystal of order one gram corresponds to a heat leak of order* 1×10^{-7} W.
Rests on Proposition 110.7.

Derivation. *Derives Proposition 110.8.*

The heat capacity available in this range is magnetic: the entropy stored in the electronic spins is at most $R \ln 2$ per mole of paramagnetic ion for a doublet ground state, and cerium magnesium nitrate, $Ce_2Mg_3(NO_3)_{12} \cdot 24\,H_2O$, has a formula-unit molar mass of $2(140.116) + 3(24.305) + 12(62.005) + 24(18.015) = 1.5296\,\mathrm{kg/mol}$ carrying two cerium ions, so one gram carries $n = 2 \times 1 \times 10^{-3}\,\mathrm{kg}/1.5296\,\mathrm{kg/mol} = 1.31 \times 10^{-3}\,\mathrm{mol}$ of moments. Warming through $\Delta T = 5 \times 10^{-3}\,\mathrm{K}$ at $T \sim 5 \times 10^{-3}\,\mathrm{K}$ absorbs at most $Q \sim T\,\Delta S \sim \Delta T\, n\, R \ln 2 = 5 \times 10^{-3}\,\mathrm{K} \times 1.31 \times 10^{-3}\,\mathrm{mol} \times 5.76\,\mathrm{J/(mol\,K)} = 3.8 \times 10^{-5}\,\mathrm{J}$, and spreading that over 3.6×10^{2} s gives 1.0×10^{-7} W; taking the full 7 mK of the stated warm-up rather than the 5 mK used here raises it to 1.5×10^{-7} W, which is the width of the estimate. A leak of a fraction of a microwatt is what a well-made cryostat of the period achieved, and the agreement is the check that nothing pathological — eddy-current heating from the field reversal, say — dominated the warm-up. □

Finally, the polarizing solenoid. It is a small vertical coil, of a few millitesla, whose sense can be reversed; its only job is to align the electronic moments so that the hyperfine field aligns the nuclei, and reversing it interchanges the two counting directions of Equation (110.4) without moving a single piece of the apparatus. That is the design decision on which the whole measurement turns: the two configurations being compared differ in nothing except the direction of an axial vector.

110.2.3 Beta and gamma detection

The beta counter An anthracene crystal a few centimetres above the source, inside the vacuum space of the cryostat, absorbs the escaping electrons and converts their energy to scintillation light. Anthracene is the reference organic scintillator, with a light yield of about 1.7×10^{4} photons per MeV of electron energy deposited, so a 150 keV electron produces about 2.6×10^{3} photons. The crystal cannot be viewed directly, because a photomultiplier is useless in a magnetic field: its gain depends on the trajectories of low-energy electrons between dynodes, and a field of even 1 mT changes it by a measurable amount. That single fact is the reason for the most conspicuous feature of the apparatus — a lucite light pipe about a metre long, carrying the scintillation light out of the cryostat and out of the fringe field to a photomultiplier standing at room temperature. Even with a transmission of order 10 % through the pipe and a photocathode quantum efficiency of order 10 %, some tens of photoelectrons per event remain, comfortably above the single- photoelectron noise, so pulse-height discrimination against the photomultiplier's dark rate is possible.

The consequence for the physics is that the beta counter has a threshold rather than a spectrometer's resolution. It counts electrons above some minimum energy, which is why v/c in Equation (110.4) is an acceptance-weighted average and not a measured quantity per event, and why 0.6 rather than a number with three digits is the honest value to carry.

The gamma counters Two sodium-iodide scintillation counters view the source from outside the cryostat, one on the polar axis and one in the equatorial plane. They count the 1173.23 keV and 1332.49 keV cascade photons, whose penetration through the cryostat walls is what makes an external gamma counter possible at all where an external beta counter is

not. Their function is not to measure a gamma energy but to compare two rates: the ratio of the polar to the equatorial rate is an anisotropy, and by Proposition 110.9 below it is a thermometer for the nuclear orientation.

Both gamma counters sit at fixed geometry, unaffected by the reversal of the polarizing field, which is what allows the gamma anisotropy and the beta asymmetry to be recorded simultaneously and compared as functions of time through the same warm-up.

Table 110.2: Conditions attained in the cobalt-60 run of 1957. These are settings of the apparatus and states of the sample rather than measured results, and the 1957 paper quotes them as design values without error bars; they are given here to the precision the design supports and no further. The measured numbers, with their uncertainties and their years, are collected separately in Table 110.3. The constants used are CODATA 2022 values [2431].

Quantity	Value	Source
Bath temperature before demagnetization	$1.2\,\mathrm{K}$	apparatus
Magnetizing field	$\sim 1\,\mathrm{T}$	apparatus
Temperature after demagnetization	$\sim 3\,\mathrm{mK}$	Equation (110.10)
Effective field at the nucleus	$\sim 10\,\mathrm{T}$	hyperfine
Orientation parameter x	0.38	Equation (110.7)
Nuclear alignment f_2	≈ 0.26	gamma counters
Gamma anisotropy ε_γ	$\approx 20\,\%$	[1133]
Warm-up time constant	$\approx 3.6 \times 10^{2}\,\mathrm{s}$	[1133]
Residual heat leak	$\sim 1 \times 10^{-7}\,\mathrm{W}$	Proposition 110.8

110.3 Procedure

110.3.1 The run cycle

A run is a cycle of preparation, a few minutes of measurement, and a long recovery, and its steps are:

1. **Pre-cool.** The cryostat is filled and the helium bath pumped to about $1.2\,\mathrm{K}$. Helium exchange gas couples the salt crystal, with its cobalt layer, to the bath.

2. **Magnetize.** A field of order $1\,\mathrm{T}$ is applied along the horizontal axis. The entropy of the electronic moments falls and the heat of magnetization is carried to the bath by the exchange gas; the crystal must be allowed to return to $1.2\,\mathrm{K}$ before the next step, or the demagnetization starts warm and ends warm.

3. **Isolate.** The exchange gas is pumped out. The crystal is now thermally disconnected, and everything that follows happens under the heat leak of Proposition 110.8.

4. **Demagnetize.** The magnetizing field is reduced to zero. By Equation (110.10) the crystal cools to a few millikelvin.

5. **Polarize.** The small vertical solenoid is energized in one sense. The electronic moments align with it, the hyperfine field aligns the cobalt nuclei, and f rises to its maximum within a time short compared with the warm-up.

6. **Count.** Beta counts in the polar anthracene counter and gamma counts in both sodium-iodide counters are recorded in successive short intervals — of order tens of seconds — against a clock, for as long as any anisotropy survives.

7. **Reverse and repeat.** The solenoid sense is reversed and the counting repeated, the whole cycle being restarted from step 1 when the polarization is gone.

110.3.2 Interleaving, and why it is not optional

Proposition 110.3 cancels the geometric and efficiency factors only if they are the same in the two configurations being compared. Over a full cooling cycle they are not: the source decays (negligibly), the photomultiplier gain drifts, the discriminator threshold moves with temperature, and above all the polarization f is itself a function of time. If $N_{\uparrow}$ were taken in one cooling cycle and $N_{\downarrow}$ in the next, the difference would contain the difference of the two cycles' base temperatures — an effect of the same size as the signal.

The measurement is therefore constructed so that the two field senses are compared *within a common warm-up window*, and the comparison is made at equal times after demagnetization. The asymmetry is formed between the counting rate at time t in a run with the field up and the counting rate at the same t in a run with the field down, and the gamma anisotropy recorded in the same intervals certifies that the two runs had the same polarization at that t. Where they did not, the gamma data say so, and the beta data can be corrected by the ratio of the two anisotropies rather than discarded.

110.3.3 The controls

Three null tests separate a genuine asymmetry from an instrumental one, and their logic is worth setting out because it is the part of the experiment that makes the result believable rather than merely striking.

1. **The warm sample.** With the crystal warm — above $1\,\mathrm{K}$, where by Equation (110.7) $x < 10^{-5}$ and f is zero to five decimal places — the polarizing field is applied in both senses and the beta rates compared. Any difference found here is instrumental by construction, since there is nothing to orient. This is the direct test for a magnetic effect on the photomultiplier, on the light pipe, or on the electron trajectories between source and counter.

2. **The warm-up curve itself.** The polarized measurement is not a single number but a time series. If the beta asymmetry and the gamma anisotropy decay to zero on the same time constant, they are both tracking f; an instrumental asymmetry has no reason to depend on the temperature of a salt crystal at all, and a fortiori no reason to share a decay constant with a gamma-ray angular distribution measured by different counters outside the cryostat.

3. **Solenoid sense versus counter position.** Reversing the field and moving the counter to the opposite side of the source are two different ways of reversing $\cos\theta$. They must give the same asymmetry with the same sign. Any effect that depends on the field direction in the laboratory rather than on the angle between the spin and the electron momentum fails this test.

The second of these is the decisive one, and it is what elevates the result above a difference of two counting rates. It converts a comparison between two configurations into a comparison between two *functions of time* measured in the same apparatus by different detectors, and no plausible instrumental effect reproduces it.

110.4 Observations and data

110.4.1 Gamma anisotropy as the polarization thermometer

The gamma counters measure an angular distribution, and an angular distribution of photons from an oriented nucleus is even in $\cos\theta$: a photon distribution cannot distinguish $\hat{\boldsymbol{n}}$ from $-\hat{\boldsymbol{n}}$, because the electromagnetic transition conserves parity. The gammas therefore measure the *alignment* of the ensemble — its second moment — and not the *polarization* — its first moment — which is precisely the quantity Equation (110.4) needs. That the one nevertheless determines the other is a consequence of the Boltzmann distribution, and it is the step in the analysis that is most often passed over.

Proposition 110.9 (Alignment determines polarization)**.** *For nuclei of spin I distributed over the Zeeman sublevels with the Boltzmann weights e^{xm} of Equation (110.9), the polarization $f_1 = \langle m\rangle / I$ and the alignment*

$$f_2 = \frac{3\langle m^2\rangle - I(I+1)}{I(2I-1)} \tag{110.11}$$

are both functions of the single parameter x. A measurement of f_2 therefore fixes x and hence f_1. For $I = 5$, $f_2 = 0.263$ corresponds to $x = 0.378$ and to $f_1 = 0.600$.

Rests on Equation (110.9).

Derivation. *Derives Proposition 110.9.*

Both moments are derivatives of the same generating function: $\langle m\rangle = \partial_x \ln Z$ and $\langle m^2\rangle = \partial_x^2 \ln Z + \langle m\rangle^2$ with $Z(x) = \sum_m \mathrm{e}^{xm}$, so both are functions of x alone, and f_2 is normalized in Equation (110.11) so that a fully polarized ensemble, $m = I$ for every nucleus, gives $f_2 = 1$: then $3I^2 - I(I+1) = 2I^2 - I = I(2I-1)$. Evaluating for $I = 5$ over a grid of x gives the pairs $(x, f_1, f_2) = (0.1, 0.196, 0.025)$, $(0.2, 0.371, 0.093)$, $(0.3, 0.513, 0.186)$, $(0.378, 0.600, 0.263)$, $(0.5, 0.701, 0.380)$, from which the quoted correspondence is read off. The map $x \mapsto f_2$ is strictly increasing, so the inversion is unique. What the gamma counters deliver is the anisotropy

$$\varepsilon_\gamma = \frac{W_\gamma(0) - W_\gamma(\pi/2)}{W_\gamma(\pi/2)}\,, \tag{110.12}$$

which is proportional to f_2 with a coefficient fixed by the multipolarities and spins of the $5^+ \to 4^+ \to 2^+ \to 0^+$ chain — pure nuclear-structure input, independent of the weak interaction, and therefore not something the beta measurement can be accused of assuming. □

The observed behaviour is that ε_γ is of order 20 % immediately after demagnetization and decays to zero over about six minutes as the crystal warms [1133]. Both gamma counters see it, with opposite signs — the polar counter's rate falls while the equatorial counter's rises, or the reverse, according to the sign of the alignment — which is itself a check that the effect is an angular redistribution of a fixed number of photons and not a change in the source.

110.4.2 The beta asymmetry and its warm-up decay

Phenomenon 110.10 (Parity violation in nuclear beta decay)**.** Electrons from the allowed Gamow–Teller decay of ^{60}Co nuclei polarized along a unit vector $\hat{\boldsymbol{n}}$ are emitted preferentially *opposite* to $\hat{\boldsymbol{n}}$. For a fully oriented ensemble — the limit $f = 1$ of Equation (110.3) — their angular distribution has the form

$$W(\theta) \propto 1 + A\,\frac{v}{c}\cos\theta\,, \qquad \cos\theta = \hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{p}}_e\,, \tag{110.13}$$

and a partially oriented one carries the extra factor f there; what a counter sees is the ratio Equation (110.4), in which f appears explicitly. Here $A < 0$ and $|A|$ is close to the largest value the transition admits: the counting-rate difference between the two senses of the polarizing field reached of order 40 % at the polarization achieved, and decayed to zero on the same time constant with which the nuclear orientation itself was lost [1101, 1133].

Rests on Equations (110.3) *and* (110.4) *and Proposition 110.1.*

Derivation. *Derives Phenomenon 110.10.*

What has to be established is negative — that no parity-conserving interaction can produce Equation (110.13) with $A \neq 0$ — and it follows from transformation properties alone, with no nuclear physics whatever. Under the space inversion $P : \boldsymbol{x} \mapsto -\boldsymbol{x}$ a momentum is odd, $\boldsymbol{p}_e \mapsto -\boldsymbol{p}_e$, while an angular momentum is a pseudovector and is even, $\hat{\boldsymbol{n}} \mapsto \hat{\boldsymbol{n}}$; hence $\cos\theta \mapsto -\cos\theta$. If the interaction responsible for the decay commuted with P, the transition rate computed from it would be invariant under P, so the distribution would have to satisfy $W(\theta) = W(\pi - \theta)$ at every angle. Comparing with Equation (110.13),

$$1 + A\,\frac{v}{c}\cos\theta = 1 - A\,\frac{v}{c}\cos\theta \qquad \text{for all } \theta\,,$$

which forces $A = 0$. A measured $A \neq 0$ is therefore not a statement about the nucleus but a proof that the interaction does not commute with P. The correlation $\hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{p}}_e$ is a pseudoscalar, and the argument just given is Proposition 110.1 applied to the case in hand: find a pseudoscalar observable, show its mean is not zero. □

The arithmetic of the observed number closes as follows, and it does not close without the polarization factor. Equation (110.4) gives $\varepsilon = A\,(v/c)\,f$. With $A = -1$, the value Proposition 101.27 derives for a pure Gamow–Teller transition fed by an interaction that emits only left-chiral electrons, with $v/c \approx 0.63$ for the counted part of the spectrum, and with $f \approx 0.6$ from Proposition 110.9, the prediction is

$$|\varepsilon| = 1 \times 0.63 \times 0.6 = 0.38\,, \tag{110.14}$$

against an observed value of order 0.4. Dropping f would predict 0.63, half again too large; dropping both f and v/c would predict 1. The two dilution factors are not corrections, they are the reason the experiment had to measure the gamma anisotropy at all.

110.4.3 Numerical results

Table 110.3 collects what the measurements delivered. Three features of the cobalt entries are worth reading deliberately.

The first is that the headline number, $|\varepsilon| \approx 0.4$, is enormous by the standards of a symmetry test. A violation at the level of a per cent would have been a discovery; one at the level of 10^{-4} would have been a triumph of technique. Forty per cent means that the parity-odd and parity-even parts of the interaction are of the same size, which is a statement about structure and not about magnitude, and it is the empirical content of "maximal".

The second is that the uncertainty on A is systematic, not statistical. Proposition 110.3 shows that 10^4 counts suffice for a per-cent statistical error, and the source easily delivered them; but A is extracted by dividing ε by $(v/c)\,f$, and both divisors carry several per cent of systematic uncertainty — v/c through the source self-absorption of Proposition 110.5 and the counter threshold, f through the nuclear-structure coefficients relating ε_γ to f_2. Propagating the three in quadrature, $\delta A/A = [(0.01/0.4)^2 + (0.02/0.63)^2 + (0.03/0.60)^2]^{1/2} = 6.4\,\%$, gives the $A = -1.06 \pm 0.07$ of Table 110.3: the statistical term contributes a quarter of the variance and the polarization term more than half. This is why the 1957 result is correctly stated as "$|A|$ consistent with the maximum" rather than as a number, and why the follow-up work with improved polarimetry was needed [1101].

Table 110.3: The measured quantities of this chapter, with their uncertainties and their years. The 1957 cobalt paper reports the effect as large and unambiguous rather than as a number with a formal error bar, so the first five entries carry the error budget reconstructed here: the statistical uncertainty of Proposition 110.3 at $N = 10^4$ counts, the source-thickness uncertainty on v/c of Proposition 110.5, and a 5 % uncertainty on f standing for the several per cent carried by the nuclear-structure coefficients that relate the gamma anisotropy to the alignment. Those three propagate in quadrature to the asymmetry coefficient. The later entries are the published uncertainties of the works cited, and each is discussed where it is introduced. The conditions attained in the cobalt run are separated into Table 110.2; the constants used are CODATA 2022 values [2431].

Quantity	Value	Year	Source
Electron speed v/c, acceptance averaged	0.63 ± 0.02	1957	Proposition 110.5
Nuclear polarization f	0.60 ± 0.03	1957	Proposition 110.9
Beta counting asymmetry $\lvert\varepsilon\rvert$, cold	0.40 ± 0.01	1957	[1133]
Asymmetry coefficient A	-1.06 ± 0.07	1957	Equation (110.14)
Beta counting asymmetry, warm sample	0.00 ± 0.01	1957	[1133]
Muon g factor, pion chain	2.00 ± 0.10	1957	[1110]
Emulsion asymmetry $\lvert a\rvert$	0.174 ± 0.038	1957	[1109]
Caesium $\operatorname{Im} E1_{PNC}/\beta$, in V/m	$0.159\,35 \pm 0.000\,56$	1997	[1946]
E122 $(A/Q^2c^2) \times 10^5$, in $1/\mathrm{GeV}^2$	-9.5 ± 1.6	1978	[1625]
E158 $\sin^2\theta_W$	0.2397 ± 0.0013	2005	[2069]
Qweak asymmetry $A_{PV} \times 10^9$	-226.5 ± 9.3	2018	[2301]
Proton weak charge Q_W^p	0.0719 ± 0.0045	2018	[2301]
Neutron skin of ^{208}Pb, $R_n - R_p$, in fm	0.283 ± 0.071	2021	[2370]

The third is that the most important entry is the one consistent with zero: with the sample warm there is nothing to orient, and the asymmetry vanishes to the statistical precision of the cold run. An experiment that reports only a large effect has reported one number; an experiment that reports a large effect together with its disappearance under a control that removes the cause has reported a measurement.

110.4.4 Systematic checks

The controls of Section 110.3.3 were carried out and the following is what they established.

Field reversal and solenoid sense. The asymmetry reverses sign with the polarizing field and does not depend on which of the two solenoid windings produced the field. An effect of the field on the counters would not know about the winding; an effect of the field on the nuclei is indifferent to it.

The warm sample. With the crystal above 1 K the counting rates in the two field senses are equal within the counting statistics. This is the direct exclusion of a magnetic effect on the photomultiplier or on the electron trajectories, and it is why the light pipe of Section 110.2.3 was worth its optical losses: the photomultiplier is far enough away that the residual field at its photocathode is small, and the warm-sample test verifies that what remains does not matter.

Scattering. Electrons scattered in the source layer, in the crystal, or from the cryostat walls dilute the asymmetry by randomizing $\cos\theta$. Every such process reduces $|\varepsilon|$ towards zero, so none of them can generate the observed effect; they enter as a one-sided systematic on $|A|$ and are the reason the extracted $|A|$ is a lower bound as much as a measurement.

Simultaneous decay of the two signals. The beta asymmetry and the gamma anisotropy vanish together, on the same time constant, as the crystal warms. This is the control that no instrumental explanation survives, and it is the reason the experiment is remembered as decisive rather than suggestive.

Reproduction with better polarimetry. The whole measurement was repeated and extended by the same group with improved nuclear orientation and thinner sources [1101], which is what turned a single striking run into a reproducible number and which added the positron-emitting cases discussed in Section 110.5.2.

110.5 Interpretation

110.5.1 Maximal violation and the V–A structure

A measured $A \neq 0$ proves that the interaction has both a parity-even and a parity-odd part. The *size* of A measures their ratio, and that is the quantitative content of the cobalt result.

Write the beta-decay interaction as a sum of a parity-even piece and a parity-odd piece, with amplitudes a_+ and a_-. The total rate goes as $|a_+|^2 + |a_-|^2$, which is parity-even; the asymmetry goes as the interference $2\,\mathrm{Re}(a_+ a_-^*)$, which is parity-odd. The asymmetry coefficient is therefore

$$A \propto \frac{2\,\mathrm{Re}\left(a_+ a_-^*\right)}{|a_+|^2 + |a_-|^2}\,, \tag{110.15}$$

a ratio bounded by unity, attaining ± 1 if and only if $|a_+| = |a_-|$ and the two are relatively real. An observed $|A|$ of order one is thus not merely evidence of parity violation but the statement that the two parts are equal in magnitude, which is a very much stronger constraint. An interaction whose parity-even and parity-odd parts are equal couples to one chirality and not at all to the other.

That is the empirical origin of the $V - A$ form of the charged weak current, proposed by Feynman and Gell-Mann [1144] and independently by Sudarshan and Marshak [1162], in which the leptonic current appears only through the chiral projector $P_L = \tfrac{1}{2}(1 - \gamma^5)$.

The current is written and its consequences derived at Equation (105.34) and throughout Section 101.4; the proposition that it violates parity maximally in exactly the sense of Equation (110.15) is Proposition 105.19, and the factor v/c that appears in Equation (110.13), and which had to be divided out to reach Equation (110.14), is the chirality-to-helicity conversion of Proposition 105.20. This chapter does not repeat those derivations; what it contributes to them is the number they had to reproduce.

Remark 110.11 (Reading this chapter against the natural-units literature). Every standard account of the weak interaction is written with $\hbar = c = 1$, and a reader moving between that literature and this book needs one rule and one worked case. The rule is that the powers of $\hbar$ and c are recoverable uniquely from dimensions: an energy E and a mass m are related by $E = mc^2$, an energy and an inverse time by $E = \hbar\omega$, a momentum and an inverse length by $p = \hbar k$. The worked case is Fermi's constant, quoted everywhere as $G_F = (1.166\,378\,8 \pm 0.000\,000\,6) \times 10^{-5}/\mathrm{GeV}^2$ — which is a statement about $G_F/(\hbar c)^3$, not about G_F — and equal in SI to $(1.435\,851\,2 \pm 0.000\,000\,7) \times 10^{-62}\,\mathrm{J\,m}^3$. Both numbers appear in this chapter, always labelled, and the conversion between them is set out at Remark 105.18. No derivation in this book is carried out in natural units.

110.5.2 Charge conjugation, CP and CPT

The same data violate charge conjugation. The argument is direct: C maps a process into the process with every particle replaced by its antiparticle, leaving momenta and spins alone. Applied to the cobalt decay it maps electron emission from a polarized nucleus into positron emission from the polarized antinucleus, with the *same* angular distribution — so C invariance would require the two asymmetry coefficients to be equal. In the pion chain, where both charges are available, the prediction is testable directly, and the π^+ and π^- chains show asymmetries of opposite sign [1110]. In the nuclear case the corresponding statement is that a positron-emitting Gamow–Teller decay must show the opposite asymmetry to an electron-emitting one, which is what the follow-up measurements found [1101].

What survives is the product. Under CP a left-handed neutrino maps to a right-handed antineutrino, and both exist; under P alone it would map to a right-handed neutrino, and under C alone to a left-handed antineutrino, neither of which occurs. Landau proposed immediately that CP is the true symmetry of the weak interaction [1118]. That proposal held for seven years, until the neutral-kaon measurement of Chapter 111, whose theory is developed at Section 105.3.

CPT is a different kind of statement. It is not a symmetry that the weak interaction happens to respect but a theorem: any local, Lorentz-invariant quantum field theory with a Hermitian Hamiltonian and the normal spin–statistics connection is invariant under the product of the three operations [1123]. It is proved at Theorem 105.34 and it is the reason that the violation of P and of C carries no threat to the framework, while a violation of CPT would.

110.5.3 Neutrino helicity

The cobalt result says the interaction distinguishes handedness but does not by itself say which handedness the neutrino has. Landau, Lee and Yang, and Salam proposed within weeks that the neutrino is described by a two-component field — the two-component neutrino belongs to Chapter 96 — rather than by a four-component Dirac field with half of it inert [1118, 1121, 1130]. A two-component field has no mass term available to it and no partner of the opposite handedness, so the proposal predicts that the neutrino is completely longitudinally polarized, and it leaves the *sign* of that polarization open. Goldhaber, Grodzins and Sunyar measured it [1148].

Phenomenon 110.12 (The neutrino is longitudinally polarized)**.** Neutrinos emitted in orbital electron capture are longitudinally polarized with negative helicity, and within the uncertainty of the measurement completely so [1148]. Helicity is the pseudoscalar $h = 2\boldsymbol{S}\cdot\hat{\boldsymbol{p}}/\hbar$, so a non-zero mean helicity is by itself a violation of parity, and a mean helicity of magnitude one is a violation of the largest size a spin-$\frac{1}{2}$ particle admits.

Rests on Phenomenon 98.2.

Derivation. *Derives Phenomenon 110.12.*

The measurement is wholly indirect — no neutrino is detected — and what makes it work is that a chain of angular-momentum conservation laws transfers the neutrino's helicity to a photon that can be analysed. The chain is $^{152m}\mathrm{Eu}\,(0^-) + e^- \to {}^{152}\mathrm{Sm}^*\,(1^-) + \nu_e$, followed by $^{152}\mathrm{Sm}^*\,(1^-) \to {}^{152}\mathrm{Sm}\,(0^+) + \gamma$ with a photon of 961 keV. Take the axis $+\hat{\boldsymbol{z}}$ along the neutrino momentum. Angular momentum along that axis gives

$$m_{\mathrm{cap}} = \tfrac{1}{2}h_\nu + m_N\,, \tag{110.16}$$

where $m_{\mathrm{cap}} = \pm\frac{1}{2}$ is the captured electron's spin projection — written with a subscript that cannot be confused with the electron mass m_e used elsewhere in this chapter — $h_\nu = \pm 1$ the neutrino helicity and m_N the projection of the recoiling 1^- nucleus, which travels along $-\hat{\boldsymbol{z}}$. Solving Equation (110.16) for the nuclear projection gives $m_N = m_{\mathrm{cap}} - \frac{1}{2}h_\nu$, so a neutrino of helicity $h_\nu = -1$ leaves $m_N = 0$ when the captured electron had $m_{\mathrm{cap}} = -\frac{1}{2}$ and $m_N = +1$ when it had $m_{\mathrm{cap}} = +\frac{1}{2}$.

The nucleus then decays to a 0^+ ground state, so the photon must carry the whole of m_N; and the experiment selects, by resonant scattering on a samarium scatterer, only those photons emitted *antiparallel* to the neutrino, because only those are Doppler-shifted enough to compensate the nuclear recoil and reach resonance. A real photon has no angular-momentum projection zero along its own direction of motion, so the $m_N = 0$ branch contributes nothing to the detected sample and every counted photon comes from $m_N = +1$. A photon travelling along $-\hat{\boldsymbol{z}}$ with projection m_γ on $+\hat{\boldsymbol{z}}$ has helicity $h_\gamma = -m_\gamma$, so $h_\gamma = -1 = h_\nu$: the photon inherits the neutrino's helicity exactly, and the same algebra run with $h_\nu = +1$ gives $h_\gamma = +1$, so the two hypotheses differ in the sign of a measurable circular polarization and not in its magnitude. The circular polarization of the resonantly scattered photon, measured by transmission through magnetized iron, is therefore the neutrino helicity, and it was found negative. The full apparatus, the recoil kinematics that make the resonance condition selective, and the dilution introduced by the finite lifetime of the intermediate state belong to Chapter 98, whose Phenomenon 98.2 states the result and carries this bookkeeping in its complete form. □

Two qualifications belong with the result. The first is that helicity is a Lorentz-invariant label only for a massless particle: for a massive one an observer moving faster than the particle sees the helicity reversed, so what is measured at a given energy is the chirality of the current rather than an intrinsic property of the particle. Proposition 105.20 makes this quantitative — a $V-A$ current emits a fermion with longitudinal polarization $-v/c$, so a strictly massless neutrino has helicity exactly -1 and a massive one falls short by $m_\nu^2 c^4/(2E^2)$. The second is that neutrinos do have mass — the oscillation measurements that show it belong to Chapter 114 — but at the sub-eV masses involved and the MeV energies of the Goldhaber measurement, the correction is smaller than 10^{-12} and the conclusion is untouched.

110.6 Immediate confirmation: the pion–muon–electron chain

110.6.1 Garwin, Lederman and Weinrich

The third of Lee and Yang's proposals was carried out at the Nevis cyclotron over a weekend, with apparatus that already existed, and it was submitted and published alongside the cobalt paper [1110]. Its subject is the chain

$$\pi^+ \to \mu^+ + \nu_\mu\,, \qquad \mu^+ \to e^+ + \nu_e + \bar{\nu}_\mu\,, \tag{110.17}$$

and its logic is that the first decay *produces* a polarization and the second *analyses* it, so that a single counting-rate modulation demonstrates parity violation twice over.

The Nevis counter arrangement A beam of positive pions from the cyclotron passes through an absorber some tens of centimetres of carbon thick. The absorber is chosen to stop the pions — which interact strongly and have a short range at the beam momentum — while transmitting the muons produced by decay in flight, which have a longer range because they do not interact strongly. What emerges is therefore a muon beam, and because the transmitted muons are those emitted forwards in the pion rest frame, they carry the full longitudinal polarization that the two-body decay gives them. The muons are brought to rest in a carbon target surrounded by a magnetizing coil producing a small, uniform, reversible field perpendicular to the muon spin. A scintillation-counter telescope in a fixed direction records positrons from the muon decays, and the counting is gated in time: a first counter signals the arrival and stopping of a muon, and positrons are accepted only within a window of a few microseconds afterwards, comparable with the muon mean life $\tau_\mu = 2.1970 \times 10^{-6}\,\mathrm{s}$ obtained from the PDG width $\Gamma_\mu = (2.995\,983\,6 \pm 0.000\,003\,0) \times 10^{-19}\,\mathrm{GeV}$ through $\tau = \hbar/\Gamma$ [2427].

The precession scan The muon spin precesses about the applied field while the muon waits to decay, so the angle between the spin and the fixed counter is a function of the field strength and of the elapsed time. Holding the time gate fixed and varying the magnetizing current sweeps that angle, and the counting rate in the fixed telescope traces out a sinusoid. The measurement is thus a scan of counting rate against coil current, which is a robust thing to measure: the geometry never moves, and the quantity extracted is the *phase and amplitude* of a periodic modulation rather than the difference of two absolute rates.

Phenomenon 110.13 (Muon polarization in the pion decay chain)**.** Positive muons brought to rest from the decay $\pi^+ \to \mu^+\nu_\mu$ are longitudinally polarized, and the positrons from their subsequent decay $\mu^+ \to e^+\nu_e\bar{\nu}_\mu$ are emitted asymmetrically about the muon spin: precessing the stopped muons in a small magnetic field modulates the rate in a fixed counter sinusoidally with the precession phase [1109, 1110]. Parity is violated twice over — once in the pion decay that creates the polarization, once in the muon decay that reveals it.

Rests on Equation (110.17) *and Phenomenon 110.12.*

Derivation. *Derives Phenomenon 110.13.*

The pion has spin 0 and decays at rest into two spin-$\frac{1}{2}$ particles emitted back to back. Take the neutrino momentum along $+\hat{\boldsymbol{z}}$, so the muon momentum is along $-\hat{\boldsymbol{z}}$. Conservation of the z component of angular momentum gives $S^\mu_z + S^\nu_z = 0$. If the neutrino carries negative helicity, as Phenomenon 110.12 states, then $S^\nu_z = -\frac{1}{2}\hbar$ and therefore $S^\mu_z = +\frac{1}{2}\hbar$; since the muon moves along $-\hat{\boldsymbol{z}}$, its helicity is

$$h_\mu = \frac{2}{\hbar}\,\boldsymbol{S}^\mu \cdot \hat{\boldsymbol{p}}^\mu = \frac{2}{\hbar}\left(+\tfrac{1}{2}\hbar\right)(-1) = -1\,.$$

The muon is produced fully polarized with its spin antiparallel to its momentum, and this is forced by the two-body kinematics together with the neutrino helicity, whatever produced the pion. A completely polarized ensemble is what makes the second, muon-decay asymmetry observable at all; the frequency of the observed modulation measures the muon precession and hence g_μ, which is where Chapter 112 begins. □

The kinematics, in numbers The two-body decay of a pion at rest fixes the muon momentum completely. With the PDG masses $m_\pi c^2 = (139.570\,39 \pm 0.000\,18)\,\mathrm{MeV}$ and $m_\mu c^2 = (105.658\,375\,5 \pm 0.000\,002\,3)\,\mathrm{MeV}$ [2427],

$$p_\mu c = \frac{\left(m_\pi^2 - m_\mu^2\right) c^4}{2 m_\pi c^2} = 29.792\,\mathrm{MeV}\,, \tag{110.18}$$

so the muon carries a kinetic energy of $\sqrt{p_\mu^2 c^2 + m_\mu^2 c^4} - m_\mu c^2 = 4.120\,\mathrm{MeV} = 6.601 \times 10^{-13}\,\mathrm{J}$ and moves at $v/c = 0.2714$. The neutrino carries the same momentum, $29.792\,\mathrm{MeV}/c$, and hence an energy of $29.792\,\mathrm{MeV} = 4.773 \times 10^{-12}\,\mathrm{J}$. That the muon is non-relativistic is what makes it stoppable in a modest absorber, and that the decay is two-body is what makes the polarization complete.

The precession, in numbers

Proposition 110.14 (Muon spin precession in a transverse field)**.** *A muon at rest with gyromagnetic ratio g_μ in a uniform field $\boldsymbol{B}$ perpendicular to its spin precesses at the angular frequency*

$$\omega = \frac{g_\mu\, e\, B}{2 m_\mu}\,, \tag{110.19}$$

so that for $g_\mu = 2$ a field of $1\,\mathrm{mT}$ gives $\omega/2\pi = 1.354 \times 10^5\,\mathrm{Hz}$ and a spin rotation of 0.30 of a turn in one muon lifetime.

Rests on Equation (79.7).

Derivation. *Derives Proposition 110.14.*

The magnetic moment of a spin-$\frac{1}{2}$ particle of mass m and charge e is $\boldsymbol{\mu} = g\,(e/2m)\,\boldsymbol{S}$, and the torque equation $\mathrm{d}\boldsymbol{S}/\mathrm{d}t = \boldsymbol{\mu} \times \boldsymbol{B}$ gives $\mathrm{d}\boldsymbol{S}/\mathrm{d}t = (geB/2m)\,\hat{\boldsymbol{B}} \times \boldsymbol{S}$, a rotation about $\boldsymbol{B}$ at the rate Equation (110.19). With $e = 1.602\,176\,634 \times 10^{-19}\,\mathrm{C}$ (exact) and $m_\mu = (1.883\,531\,627 \pm 0.000\,000\,042) \times 10^{-28}\,\mathrm{kg}$ [2431], $g_\mu = 2$ and $B = 1 \times 10^{-3}\,\mathrm{T}$,

$$\frac{\omega}{2\pi} = \frac{eB}{2\pi m_\mu} = \frac{1.602\,176\,634 \times 10^{-19}\,\mathrm{C} \times 1 \times 10^{-3}\,\mathrm{T}}{2\pi \times 1.883\,531\,627 \times 10^{-28}\,\mathrm{kg}} = 1.354 \times 10^5\,\mathrm{Hz}\,,$$

and $1.354 \times 10^5\,\mathrm{Hz} \times 2.197 \times 10^{-6}\,\mathrm{s} = 0.297$ turns per lifetime. A field of a few millitesla therefore sweeps the spin through a full turn inside the observation window, which is the design requirement: too small a field gives no modulation to fit, and too large a one washes the modulation out inside the finite time gate. □

The counting rate in the fixed telescope is then

$$N(t) \propto \mathrm{e}^{-t/\tau_\mu}\left[1 + a\cos(\omega t + \varphi)\right]\,, \tag{110.20}$$

with a the product of the muon polarization and the intrinsic positron asymmetry of the muon decay. Garwin, Lederman and Weinrich found a modulation of amplitude $|a| \approx \frac{1}{3}$ and, from the frequency at fixed gate, a muon gyromagnetic ratio $g_\mu = 2.00 \pm 0.10$ [1110] — the first measurement of the muon magnetic moment, made as a by-product of a symmetry test, and the ancestor of the precision programme of Chapter 112. The intrinsic asymmetry

$\frac{1}{3}$ is itself a $V-A$ prediction: the positron spectrum and its angular correlation both follow from the four-fermion matrix element of the decay, whose energy dependence alone is parametrized at Equation (101.58) in the form Michel introduced before the structure was known [965]. Writing $x = 2E_e/(m_\mu c^2)$ for the positron energy fraction and θ for the angle between the positron momentum and the muon spin, the $V-A$ doubly differential rate is

$$\frac{\mathrm{d}^2\Gamma}{\mathrm{d}x\,\mathrm{d}\cos\theta} \propto x^2\left[(3-2x) + P_\mu(2x-1)\cos\theta\right], \tag{110.21}$$

and integrating the $\cos\theta$-odd and $\cos\theta$-even parts separately over $0 \le x \le 1$ gives, at full polarization $P_\mu = 1$,

$$\frac{\int_0^1 x^2(2x-1)\,\mathrm{d}x}{\int_0^1 x^2(3-2x)\,\mathrm{d}x} = \frac{\frac{1}{2}-\frac{1}{3}}{1-\frac{1}{2}} = \frac{1}{3}, \tag{110.22}$$

which is the number Garwin, Lederman and Weinrich measured. What this book does not yet contain is the step before it.

Derivation pending #190

Doubly differential muon-decay rate from the $V-A$ four-fermion matrix element, i.e. the angular factor $[(3-2x)+P_\mu(2x-1)\cos\theta]$ that is quoted rather than computed here; the energy spectrum alone is parametrized in the weak-interactions chapter, and the integration of the angular factor is carried out in this chapter. Belongs in the long-proofs appendix

110.6.2 Friedman and Telegdi

Friedman and Telegdi measured the same chain by an entirely different technique and with entirely different systematics [1109]. Pions were brought to rest inside a nuclear emulsion; the muon from the decay, with the 4.120 MeV of Equation (110.18), travels a short, visible track and stops; and the decay positron leaves a third track from the same point. Every event is therefore a complete, reconstructed geometry — pion direction, muon direction, positron direction — read off under a microscope, with no counters, no magnetic field and no timing.

The observable is the angle θ between the muon direction and the positron direction, and the distribution is $W(\theta) \propto 1 + a\cos\theta$. The measured coefficient was of magnitude 0.174 ± 0.038, diluted from the pion-chain maximum because a muon coming to rest in an emulsion depolarizes through hyperfine interactions with the atoms of the stopping medium — a systematic entirely absent from the counter experiment, and absent in the opposite direction, since depolarization can only reduce the asymmetry and never create it.

Proposition 110.15 (Event count implied by the quoted error)**.** *For a distribution $1+a\cos\theta$ with $|a| \ll 1$ observed over the full solid angle, the maximum-likelihood estimator $\hat{a} = 3\langle\cos\theta\rangle$ has variance $3/N$. An uncertainty of 0.038 therefore corresponds to $N = 3/0.038^2 \approx 2.1\times10^3$ reconstructed decay chains.*

Rests on Equation (110.21).

Derivation. *Derives Proposition 110.15.*

Normalize the distribution over $\cos\theta \in [-1,1]$: $p(\cos\theta) = \frac{1}{2}(1+a\cos\theta)$. Then $\langle\cos\theta\rangle = \frac{1}{2}\int_{-1}^{1} u(1+au)\,\mathrm{d}u = a/3$, so $\hat{a} = 3\langle\cos\theta\rangle$ is unbiased. Its variance is $9/N$ times the variance of $\cos\theta$, and to leading order in a that variance is $\langle\cos^2\theta\rangle - \langle\cos\theta\rangle^2 = \frac{1}{3} + O(a^2)$, giving $\mathrm{Var}(\hat{a}) = 3/N$. Setting $\sqrt{3/N} = 0.038$ gives $N \approx 2.1\times10^3$. The same arithmetic applied to the counter experiment's quoted precision on a of about 0.03 gives $N \approx 3.3\times10^3$ — both experiments rest on a few thousand events, which is worth stating plainly: the effect is

so large that a few thousand events settle it, and that is why two groups could do it in weeks. □

Table 110.4: The five founding parity-violation measurements — four published in 1957, the neutrino-helicity one in 1958 — and what each delivered. All five test the mean of a pseudoscalar, and all five find it as large as the kinematics permits. The last column is the pseudoscalar whose non-vanishing is the violation.

Measurement	Result	Pseudoscalar
^{60}Co beta asymmetry, 1957 [1133]	$\lvert\varepsilon\rvert = 0.40 \pm 0.01$; $A = -1.06 \pm 0.07$	$\langle \boldsymbol{J} \rangle \cdot \boldsymbol{p}_e$
$\pi \to \mu \to e$ counters, 1957 [1110]	$\lvert a \rvert \approx \frac{1}{3}$; $g_\mu = 2.00 \pm 0.10$	$\langle \boldsymbol{s}_\mu \rangle \cdot \boldsymbol{p}_\mu$ and $\langle \boldsymbol{s}_\mu \rangle \cdot \boldsymbol{p}_e$
$\pi \to \mu \to e$ emulsion, 1957 [1109]	$\lvert a \rvert = 0.174 \pm 0.038$, depolarization-diluted	$\boldsymbol{p}_\mu \cdot \boldsymbol{p}_e$
Beta polarization, 1957 [1108, 1131]	electrons longitudinally polarized, negative helicity	$\langle \boldsymbol{s}_e \rangle \cdot \boldsymbol{p}_e$
Neutrino helicity, 1958 [1148]	h_ν negative, consistent with -1	$\langle \boldsymbol{s}_\nu \rangle \cdot \boldsymbol{p}_\nu$

Table 110.4 is the summary of that short period. The five entries share no apparatus, no technique and no systematic; two are counting experiments at millikelvin temperatures, one is a microscope study of photographic emulsion, one is Mott-scattering polarimetry and one is resonant nuclear scattering through magnetized iron. What they share is the structure of Proposition 110.1, and that is why the conclusion was accepted within months.

110.7 Modern repetitions and precision

110.7.1 Neutral currents and the weak mixing angle

For fifteen years parity violation was a property of charged-current weak decays, in which the charge of the participating particles changes. The Gargamelle bubble chamber then observed neutrino interactions with no charged lepton in the final state [1502] — a weak *neutral* current, discussed at Section 104.6.1. Because that current couples to ordinary charged particles as well as to neutrinos, it makes parity violation observable in processes with no charge exchange at all: in the scattering of electrons from nuclei, and in atoms.

The electroweak model of Glashow, Weinberg and Salam [1206, 1362, 1381] fixes the strength of every one of those effects in terms of a single parameter, the weak mixing angle. The observable that all the measurements below deliver is $\sin^2\theta_W$, and the content of the comparison is not the value but the consistency: the same angle, running with momentum transfer according to a calculable prescription, must fit determinations spanning ten orders of magnitude in energy. The theory is Section 104.4.2 and the statement under test is Phenomenon 104.55.

110.7.2 Atomic parity violation

Phenomenon 110.16 (Parity violation in atoms)**.** The $6S \to 7S$ transition of atomic caesium, forbidden as an electric dipole transition by the parity selection rule, nevertheless proceeds with a small parity-nonconserving electric dipole amplitude whose sign reverses with the handedness of the applied static fields. The amplitude has been measured to 0.35 %, and the same measurement resolved its nuclear spin dependence, the first observation of an

anapole moment [1946]. The effect requires no charge exchange — it is a neutral-current effect — and it grows roughly as the cube of the nuclear charge, which is why heavy atoms were chosen [1519].

Rests on Proposition 105.1 and Phenomenon 104.55.

Derivation. *Derives Phenomenon 110.16.*

Two things need showing: that the neutral current produces a parity-nonconserving $E1$ amplitude at all, and that it grows as Z^3.

Exchange of a Z boson between an atomic electron and a nucleon is, at atomic momentum transfers, a contact interaction: the range $\hbar/(m_Z c) = 2.2 \times 10^{-18}\,\mathrm{m}$ is smaller than the nucleus by three orders of magnitude, so the nucleus acts as a point source. The piece of it that violates parity and does not vanish on averaging over nucleon spins couples the electron *axial* current to the nucleon *vector* charge, giving an effective Hamiltonian

$$H_{PNC} = \frac{G_F}{\sqrt{2}}\,\frac{Q_W}{2}\,\gamma^5\,\delta^3(\boldsymbol{r})\,, \tag{110.23}$$

in which Q_W is the weak charge of the nucleus and γ^5 is the parity-odd Dirac matrix. Since γ^5 is a pseudoscalar operator, first-order perturbation theory mixes into an S state a small admixture of the nearby P states,

$$\left|\tilde{S}\right\rangle = |S\rangle + \sum_P \frac{\langle P|\,H_{PNC}\,|S\rangle}{E_S - E_P}\,|P\rangle\ ,$$

and the ordinary electric dipole operator then connects $\left|6\tilde{S}\right\rangle$ to $\left|7\tilde{S}\right\rangle$ through those admixtures. The transition acquires an $E1$ amplitude proportional to Q_W, whose sign flips under reversal of the handedness of the applied electric and magnetic fields, which is the experimental signature.

For the Z^3 scaling, count the powers. The weak charge is, at tree level, $Q_W = Z(1 - 4\sin^2\theta_W) - N$, and since $\sin^2\theta_W \approx 0.231$ [2427] the first term nearly cancels and $Q_W \approx -N$, one power of the nuclear size. The delta function in Equation (110.23) requires the electron density at the nucleus, and for a valence electron in a heavy atom the relativistic enhancement of that density supplies two further powers of Z. Hence Z^3, which is why caesium ($Z = 55$) and thallium were the practical choices, and why hydrogen — where the atomic theory is exact — is hopeless. □

The Boulder measurement is the reference point [1946]. Its apparatus is a beam of caesium atoms crossing a region of static electric and magnetic fields inside an optical cavity, in which a laser drives the nominally forbidden $6S \to 7S$ transition; the parity-violating amplitude is detected through its interference with a Stark-induced amplitude. That is the whole trick: a rate proportional to $|\mathcal{M}_{\mathrm{Stark}} + \mathcal{M}_{PNC}|^2$ carries a cross term $2\,\mathrm{Re}(\mathcal{M}^*_{\mathrm{Stark}}\mathcal{M}_{PNC})$ that is *linear* in the small amplitude and changes sign with the applied field, whereas the parity-violating rate on its own would be quadratic in it and therefore smaller by one further power of the same small ratio. The order of magnitude of $\mathcal{M}_Z/\mathcal{M}_\gamma$ that makes $\mathcal{M}_{PNC}$ this small is estimated, for electron scattering, at Proposition 105.21. The measured quantity is the ratio of the parity-nonconserving amplitude to the Stark coefficient,

$$\frac{\mathrm{Im}\,E1_{PNC}}{\beta} = (0.159\,35 \pm 0.000\,56)\,\mathrm{V/m}\,, \tag{110.24}$$

customarily quoted as $(1.5935 \pm 0.0056)\,\mathrm{mV}$ per centimetre; the fractional precision is $0.35\,\%$. The difference between the two hyperfine components of the transition, $(0.0077 \pm 0.0011)\,\mathrm{V/m}$, is the nuclear-spin-dependent part and constitutes the first observation of an anapole moment — a parity-odd, time-reversal-even electromagnetic moment of the nucleus, generated by the

weak interaction between its own nucleons. Converting Equation (110.24) into the weak charge

$$Q_W(^{133}\mathrm{Cs}) = -72.06 \pm 0.28 \pm 0.34 \tag{110.25}$$

requires atomic many-body theory, and the second uncertainty in Equation (110.25) is that theory's, not the experiment's.

Derivation pending #191

Atomic many-body calculation converting the measured caesium amplitude into a nuclear weak charge: the relativistic coupled-cluster evaluation of the valence electron density at the nucleus and its error budget. Belongs in the long-proofs appendix; the mixing of opposite-parity states and the Z-cubed scaling are derived in this chapter

That the theoretical uncertainty exceeds the experimental one is the structural limitation of the method, and it is what motivates the isotope-ratio technique: measuring the parity-violating amplitude in several isotopes of the same element and forming ratios cancels the electronic structure, which is common to all of them, leaving the dependence on the neutron number. Antypas and collaborators carried this out in ytterbium, whose parity-violating amplitude is some two orders of magnitude larger than caesium's, and observed the variation of the amplitude across four isotopes [2329]. The atomic-structure input that the ratio does not cancel is the distribution of neutrons inside the nucleus, so the measurement trades one theoretical uncertainty for another — but for a nuclear one, which is measured independently by the technique of Section 110.7.4. The atomic-structure background is Chapter 88.

110.7.3 Polarized electron scattering

Phenomenon 110.17 (Parity violation in electron scattering)**.** The cross section for scattering longitudinally polarized electrons from an unpolarized target depends on the sign of the beam helicity. The fractional asymmetry

$$A = \frac{\sigma_+ - \sigma_-}{\sigma_+ + \sigma_-} \tag{110.26}$$

between the two beam helicities was found to be of order 10^{-4}, and to grow in proportion to Q^2, in deep inelastic scattering of 19 GeV electrons from deuterium [1625].

Rests on Proposition 110.1 and Phenomenon 104.55.

Derivation. *Derives Phenomenon 110.17.*

Beam helicity is a pseudoscalar, so a dependence of the rate on its sign violates parity for exactly the reason Equation (110.13) does, and the one-photon amplitude alone cannot produce it. The size follows from the interference of the two exchanges. At momentum transfers far below the Z mass the neutral-current amplitude is contact-like, of order $G_F/\sqrt{2}$, while the one-photon amplitude carries the propagator $1/Q^2$ together with the coupling $4\pi\alpha$; the asymmetry is the ratio of the interference term to the electromagnetic term, hence of order

$$A \sim \frac{G_F Q^2 c^2}{4\sqrt{2}\,\pi\alpha\,(\hbar c)^3}\,. \tag{110.27}$$

With $G_F/(\hbar c)^3 = (1.166\,378\,8 \pm 0.000\,000\,6) \times 10^{-5}/\mathrm{GeV}^2$, equivalently $G_F = (1.435\,851\,2 \pm 0.000\,000\,7) \times 10^{-62}\,\mathrm{J\,m}^3$, and $\alpha^{-1} = 137.035\,999\,177 \pm 0.000\,000\,021$ [2431], a momentum transfer $Q^2c^2 = 1\,\mathrm{GeV}^2$ gives $A \sim 9.0 \times 10^{-5}$, the observed order of magnitude; the coefficient suppressed in Equation (110.27) is of order unity and depends on the target and on $\sin^2\theta_W$. It is the linear growth with Q^2, not the magnitude, that identifies the effect as weak-neutral-current interference and not an instrumental correlation with the beam helicity, since no instrumental effect has any reason to scale that way. □

SLAC E122: apparatus The experiment that first saw this is SLAC E122 [1625, 1643]. Its apparatus is worth describing, because measuring an asymmetry of 10^{-4} against helicity-correlated instrumental effects is the whole difficulty, and every feature of the design addresses it.

The beam is produced by photoemission from a gallium arsenide photocathode illuminated by circularly polarized laser light. The handedness of the light is set by the sign of the voltage on a Pockels cell, and the electron polarization follows it: the crystal's band structure converts optical helicity into electron spin polarization, reaching about 37 % in practice against a theoretical maximum of 50 %. The essential property is that reversing the beam helicity requires nothing to move — only the sign of a voltage — so the reversal can be made *randomly, pulse to pulse*, and any slow drift in the apparatus averages out between neighbouring pulses rather than accumulating into a false asymmetry.

The beam energy is between 16.2–22.2 GeV, the target is liquid deuterium (and hydrogen), and the scattered electrons are detected in a magnetic spectrometer set to accept deep-inelastic kinematics near $Q^2c^2 \approx 1.6\,\mathrm{GeV}^2$. The flux is far too high to count events individually, so the spectrometer is instrumented as an *integrating* calorimeter: the total charge deposited by each beam pulse is measured, and the asymmetry is formed between pulses of opposite helicity. That removes counting statistics as a limitation and replaces it with the requirement that the beam intensity, position, angle and energy be identical for the two helicities — monitored pulse by pulse and corrected for.

The spin-precession scan: the control that makes it a measurement The most elegant feature of E122 is a null test that no instrumental effect can imitate, and it is worth deriving because it turns the electron's anomalous magnetic moment into a piece of experimental apparatus.

Proposition 110.18 (Spin precession in the beam transport)**.** *An ultrarelativistic electron of Lorentz factor γ bent through an angle θ_{bend} in a magnetic field has its spin rotated relative to its momentum by*

$$\phi = a_e\,\gamma\,\theta_{\mathrm{bend}}\,, \qquad a_e = \frac{g_e - 2}{2} = (1.159\,652\,180\,46 \pm 0.000\,000\,000\,18) \times 10^{-3}\,, \quad (110.28)$$

so the longitudinal polarization delivered to the target varies as $\cos\phi$ with the beam energy. For the SLAC beam line, whose bend to the experimental hall is $\theta_{\mathrm{bend}} = 24.5^\circ = 0.4276$, the four energies 16.2 GeV, 17.8 GeV, 19.4 GeV and 22.2 GeV give $\phi/\pi = 5.00$, 5.50, 5.99 and 6.86, and hence $\cos\phi = -1.00$, -0.01, 1.00 and -0.90.

Rests on Equations (79.7) and (100.89).

Derivation. Derives Proposition 110.18.

In a uniform magnetic field the momentum of a relativistic charged particle rotates at the cyclotron frequency $\omega_c = eB/(\gamma m_e)$, while its spin rotates at $\omega_s = geB/(2m_e) + (1-\gamma)\,eB/(\gamma m_e)$ — the Thomas–BMT result, in which the second term is the Thomas precession of the rest frame. Subtracting,

$$\omega_s - \omega_c = \frac{eB}{m_e}\left(\frac{g}{2} - 1\right) = a_e\,\frac{eB}{m_e} = a_e\,\gamma\,\omega_c\,,$$

so in the time the momentum turns through $\theta_{\mathrm{bend}} = \omega_c t$ the spin turns through $\theta_{\mathrm{bend}} + a_e\gamma\theta_{\mathrm{bend}}$ and the angle between them is Equation (110.28). Numerically, $\gamma = E/(m_ec^2)$ with $m_ec^2 = (0.510\,998\,950\,69 \pm 0.000\,000\,000\,16)$ MeV [2431] gives $\gamma = 3.80 \times 10^4$ at $E = 19.4$ GeV, and $a_e\gamma\theta_{\mathrm{bend}} = 1.159\,65 \times 10^{-3} \times 3.7965 \times 10^4 \times 0.4276 = 18.83 = 5.99\pi$. Repeating for the other three energies gives the quoted list. That the numbers land so close to integer multiples of π is not a coincidence: the energies were chosen so. □

The consequence is a test with no instrumental counterpart. The measured asymmetry must reverse sign between 16.2 GeV and 19.4 GeV, and must vanish at 17.8 GeV where the spin arrives transverse and the beam carries no longitudinal polarization at all — while the beam intensity, the target, the spectrometer and the electronics are unchanged. No helicity-correlated instrumental effect depends on the beam energy through $\cos(a_e \gamma \theta_{\text{bend}})$. The scan was performed and the asymmetry followed the prediction [1643], which is what converted the measurement from a small number into a result.

What E122 measured The asymmetry, normalized to the momentum transfer, is

$$\frac{A}{Q^2c^2} = -9.5 \pm 1.6 \times 10^{-5}\,1/\text{GeV}^2 = (-3.70 \pm 0.62) \times 10^{15}/\text{J}^2\,, \tag{110.29}$$

the second form being the SI one, since Q^2c^2 is an energy squared and A is dimensionless; the conversion is $1/\text{GeV}^2 = 3.8954 \times 10^{19}/\text{J}^2$. So $|A| \approx 1.5 \times 10^{-4}$ at the working point — in magnitude exactly the estimate of Equation (110.27). The electroweak interpretation is not read off the magnitude, however, but off the *shape*. Writing $y = \nu/E$ for the fractional energy transfer, the prediction is

$$\frac{A}{Q^2c^2} = a_1 + a_2\,\frac{1-(1-y)^2}{1+(1-y)^2}\,, \tag{110.30}$$

with $a_1 \propto 1 - \frac{20}{9}\sin^2\theta_W$ and $a_2 \propto 1 - 4\sin^2\theta_W$, both quoted here from the analysis of the experiment rather than computed.

Derivation pending #192

Coefficients of the deep-inelastic parity-violating asymmetry on an isoscalar target: the combinations $1 - \frac{20}{9}\sin^2\theta_W$ and $1 - 4\sin^2\theta_W$ follow from the electron and quark neutral-current couplings C_{1q}, C_{2q} contracted with the quark-parton structure functions of a deuteron, and are quoted rather than derived here. Belongs in the long-proofs appendix; the couplings themselves are written down in this chapter for the proton case, and the deep-inelastic kinematics live in the deep-inelastic chapter

Since $\sin^2\theta_W \approx 0.23$, the second coefficient nearly vanishes, so the asymmetry is nearly independent of y — and it is that flatness, not the magnitude, that discriminates. Several models then in play reproduced the Gargamelle neutrino data with a parity-conserving neutral current, and were killed by this single observation. E122 gave $\sin^2\theta_W = 0.20 \pm 0.03$ [1625], sharpened to 0.224 ± 0.020 by the full kinematic scan [1643]. The deep-inelastic kinematics and structure functions the analysis rests on belong to Chapter 116.

SLAC E158 The same technique applied to a purely leptonic target removes the hadronic-structure uncertainty altogether. SLAC E158 scattered longitudinally polarized electrons from the atomic electrons of a liquid-hydrogen target — polarized Møller scattering — at $Q^2c^2 = 0.026\,\text{GeV}^2$, and measured an asymmetry of order 10^{-7} — three orders of magnitude smaller than E122's $|A| \approx 1.5 \times 10^{-4}$, and correspondingly harder. The result, $\sin^2\theta_W = 0.2397 \pm 0.0013$ [2069], is the first observation of the *running* of the weak mixing angle to low momentum transfer: it differs from the Z-pole value by far more than its uncertainty, in the direction and by the amount the renormalization group predicts.

110.7.4 The weak charge of the proton

Phenomenon 110.19 (The weak charge of the proton is small)**.** Elastic scattering of longitudinally polarized electrons from protons at low momentum transfer shows a parity-violating asymmetry of about -226 parts per billion, from which the weak charge of the

proton is extracted as $Q_W^p = 0.0719 \pm 0.0045$ [2301]. The proton's weak charge is thus more than an order of magnitude smaller than its electric charge in the corresponding normalization, and it is this smallness that makes the measurement a sharp determination of the weak mixing angle at low momentum transfer.

Rests on Phenomena 104.55 and 110.17.

Derivation. *Derives Phenomenon 110.19.*

Write the parity-violating electron–quark interaction in contact form, with coefficients C_{1u} and C_{1d} multiplying the products of the axial electron current with the vector quark currents. The Z couplings of Chapter 104 give, at tree level,

$$C_{1u} = -\tfrac{1}{2} + \tfrac{4}{3}\sin^2\theta_W\,, \qquad C_{1d} = \tfrac{1}{2} - \tfrac{2}{3}\sin^2\theta_W\,.$$

A proton has valence content uud, so its weak charge is the corresponding sum,

$$Q_W^p = -2\,(2C_{1u} + C_{1d}) = 1 - 4\sin^2\theta_W\,. \tag{110.31}$$

With $\sin^2\theta_W \approx 0.231$ [2427] the right-hand side is $1 - 4(0.231) = 0.076$: the two terms very nearly cancel, because the measured mixing angle happens to sit close to the value $1/4$ at which Q_W^p would vanish identically. A small quantity written as the difference of two quantities of order unity is maximally sensitive to them, which is why an asymmetry of a few hundred parts per billion determines $\sin^2\theta_W$ competitively; differentiating Equation (110.31) gives $\delta(\sin^2\theta_W) = \tfrac{1}{4}Q_W^p\,\delta Q_W^p/Q_W^p$, so a 6 % determination of Q_W^p yields $\delta\sin^2\theta_W \approx 0.001$. □

Equation (110.31) is a tree-level relation and must not be applied to the measured numbers without saying so, because it fails on them by far more than their errors. Substituting the $\sin^2\theta_W = 0.2383 \pm 0.0011$ that the same measurement reports gives $1 - 4(0.2383) = 0.0468$, which is 5.6 standard deviations below the reported $Q_W^p = 0.0719 \pm 0.0045$. The running invoked above does not account for the gap and points the wrong way: driving $\sin^2\theta_W$ up from its Z-pole value towards 0.2386 at $Q \to 0$ makes $1 - 4\sin^2\theta_W$ *smaller*, not larger. The $+0.025$ that reconciles the two is supplied by the one-loop electroweak corrections proper — the ρ factor and the WW, ZZ and γZ box graphs — which add a constant to Q_W^p without changing $\mathrm{d}Q_W^p/\mathrm{d}\sin^2\theta_W = -4$. That is why the error propagation of the previous paragraph survives the corrections although the central-value relation does not: $0.0045/4 = 0.0011$ is exactly the uncertainty quoted on $\sin^2\theta_W$ [2301].

Derivation pending #193

One-loop electroweak corrections to the weak charge of the proton: the ρ factor, the charge-radius term, and the WW, ZZ and γZ box graphs that together shift the tree value $1 - 4\sin^2\theta_W = 0.047$ to the Standard Model prediction near 0.071. They are a third of the predicted value, so the comparison with experiment does not exist without them. Belongs in the long-proofs appendix

Qweak The measurement was made at Jefferson Lab with a beam of longitudinally polarized electrons on a liquid-hydrogen target, detecting elastically scattered electrons at $Q = 0.157\,\mathrm{GeV}/c$, that is $Q^2c^2 = 0.0248\,\mathrm{GeV}^2$. The measured asymmetry is

$$A_{PV} = -226.5 \pm 9.3 \times 10^{-9}\,, \tag{110.32}$$

in parts per billion, -226.5 ± 9.3, from which

$$Q_W^p = 0.0719 \pm 0.0045\,, \qquad \sin^2\theta_W = 0.2383 \pm 0.0011\,, \tag{110.33}$$

in agreement with the Standard Model [2301]. Measuring 2.3×10^{-7} requires the helicity-correlated differences in beam current, position, angle and energy to be controlled at the level of parts per billion in the asymmetry — achieved by the same random pulse-to-pulse helicity reversal as E122, supplemented by two slow reversals whose sign flips independently of the Pockels cell: inserting a half-wave plate in the laser beam, and reversing the spin with a Wien filter. Any false asymmetry tied to the electronics does not follow those reversals; the physics asymmetry does.

PREX and the neutron skin The near-cancellation in Equation (110.31) works the other way for a heavy nucleus. Since $Q_W^p \approx 0.07$ while the neutron's weak charge is $Q_W^n = -1$, the Z couples almost exclusively to neutrons; a parity-violating asymmetry in elastic electron scattering therefore measures the *neutron* distribution, where ordinary electron scattering measures the proton distribution. PREX-2 measured $A_{PV} = 550 \pm 16 \pm 8 \times 10^{-9}$ on ^{208}Pb and obtained a neutron skin thickness

$$R_n - R_p = (0.283 \pm 0.071)\,\mathrm{fm} = (2.83 \pm 0.71) \times 10^{-16}\,\mathrm{m} \tag{110.34}$$

[2370]. The skin thickness is fixed by the density dependence of the nuclear symmetry energy, which is the same quantity that sets the radius of a neutron star, so a parity-violating asymmetry measured on a lead target constrains the dense-matter equation of state, which belongs to Chapter 49. The nuclear-structure background is Section 107.3.3 and Section 107.4.2.

110.7.5 Consistency and what remains open

Table 110.5: Parity violation as a precision instrument: determinations of the weak mixing angle from parity-violating observables, over four decades of momentum transfer. The values are in the conventions of the Particle Data Group [2427]; the running of $\sin^2\theta_W$ with Q is part of the prediction, so the entries are not expected to agree numerically and their agreement *after* running is the test.

Measurement	Momentum transfer	$\sin^2\theta_W$	Reference
Caesium atomic parity violation, 1997	atomic, $Q \approx 0$	≈ 0.24	[1946]
Polarized Møller scattering, E158, 2005	$Q^2c^2 = 0.026\,\mathrm{GeV}^2$	0.2397 ± 0.0013	[2069]
Polarized ep elastic, Qweak, 2018	$Q = 0.157\,\mathrm{GeV}/c$	0.2383 ± 0.0011	[2301]
Polarized ed deep inelastic, E122, 1979	$Q^2c^2 \approx 1.6\,\mathrm{GeV}^2$	0.224 ± 0.020	[1643]
Z-pole asymmetries, LEP and SLD, 2006	$Q = m_Z c$	$0.231\,53 \pm 0.000\,16$	[2104]

Table 110.5 is the state of the art, and reading it correctly requires one caution and admits one honest complaint.

The caution is that the entries are not repeated measurements of a constant. The weak mixing angle runs with the momentum transfer at which it is probed, exactly as the fine-structure constant does, and the predicted variation between the atomic and the Z-pole entries is larger than several of the quoted uncertainties. The test is therefore not that the numbers agree but that they lie on the predicted curve; the scheme in which the angle is defined must also be stated, since different schemes differ by about a per cent, which is

hundreds of times some of these errors. Remark 104.34 makes that point in the form it needs.

The complaint is a genuine tension inside the most precise entry. The Z-pole value in Table 110.5 is an average over determinations that do not agree well among themselves: the left–right asymmetry measured with polarized beams at SLD gives $\sin^2\theta_W^{\mathrm{eff}} = 0.230\,98 \pm 0.000\,26$ while the forward–backward asymmetry of b quarks at LEP gives $0.232\,21 \pm 0.000\,29$, a discrepancy of about three standard deviations that has stood since the measurements were finished [2104]. It is stated here rather than averaged away because averaging a three-sigma disagreement and quoting the result to five digits is exactly the practice this book's editorial rules exist to prevent. No explanation has been established; the possibilities range from an underestimated systematic in one of the two to a statistical fluctuation, and the honest statement is that the world average carries an inflated error for a reason.

Two further things should be said about what parity violation is now for.

It is a *null test*, not a discovery channel. Every measurement in Table 110.5 is a comparison of a measured asymmetry with a Standard Model prediction, and the interest lies in the residual. A new neutral current, coupling to electrons and quarks with a strength suppressed by a mass scale Λ, contributes a contact term of strength $(\hbar c)^3/\Lambda^2$ where the weak interaction contributes G_F, so the relative shift of the asymmetries is of order

$$\frac{\delta A}{A} \sim \frac{(\hbar c)^3}{G_F \Lambda^2}\,. \tag{110.35}$$

The combination is dimensionless and that is the point of writing it out: G_F has the SI dimension $\mathrm{J\,m^3}$, so $G_F/(\hbar c)^3$ is an inverse energy squared, $4.5438 \times 10^{14}/\mathrm{J^2}$, and multiplying by Λ^2 leaves a pure number. Numerically $G_F\Lambda^2/(\hbar c)^3 = 11.7$ at $\Lambda = 1\,\mathrm{TeV}$ and 1166 at $\Lambda = 10\,\mathrm{TeV}$, so Equation (110.35) predicts relative shifts of 8.6×10^{-2} and 8.6×10^{-4} respectively. The present 6 % determination of Q_W^p therefore reaches $\Lambda \approx 1.2\,\mathrm{TeV}$ for a new coupling of weak strength, and correspondingly further for a stronger one — the reach is customarily quoted as a bound on Λ/g for that reason [2301]. Either way it is a TeV-scale statement obtained without producing anything, which is the characteristic economy of a symmetry test.

And parity is *not* violated in the strong and electromagnetic interactions, to the precision anyone has been able to reach. The bound is the one Lee and Yang already quoted from spectroscopy — about 10^{-7} in amplitude for a parity-violating admixture in the nuclear force [1091] — and it has since been sharpened, but the qualitative statement is unchanged and it is a fact about the Standard Model rather than an assumption of it: the photon and gluon couple vectorially, to left- and right-handed fields alike, and only the W and Z do not. The parity-violating nucleon–nucleon force that does exist is the weak interaction acting inside a nucleus, and its observation — of the order the bound allows — is a measurement of the weak interaction, not a failure of the strong one.

110.8 Primary references

The measurement itself is [1133]: Wu, Ambler, Hayward, Hoppes and Hudson, "Experimental Test of Parity Conservation in Beta Decay", *Physical Review* **105**, 1413 (1957). It is three pages long, contains two figures and no error analysis in the modern sense, and it is the primary source for the cobalt entries of Table 110.2 and for those rows of Table 110.3 attributed to the experiment rather than derived here. A reader consulting it should expect the conclusions to be argued from the coincidence of the beta and gamma warm-up curves rather than from a quoted uncertainty; that is the form the argument took, and it is why the follow-up [1101] — Ambler, Hayward, Hoppes, Hudson and Wu, "Further Experiments on β Decay of Polarized Nuclei", same journal, **106**, 1361 (1957) — matters: it repeated

the measurement with better polarimetry, extended it to other nuclei, and supplied the positron-emitting case that establishes the violation of charge conjugation.

The prediction under test is [1091]: Lee and Yang, "Question of Parity Conservation in Weak Interactions", *Physical Review* **104**, 254 (1956). It should be read for its method rather than its result — it contains no new measurement and proposes no mechanism, and its lasting contribution is the observation that a universally believed law had never been tested in the sector where it was about to fail, together with an explicit list of feasible tests. The tau–theta puzzle that motivated it is analysed in [1016], whose phase-space construction is what removed the angular-momentum escape route; the introduction of parity as a conserved quantum number is [592].

The two independent confirmations are [1110] — Garwin, Lederman and Weinrich, published back to back with Wu's paper — and [1109], Friedman and Telegdi, submitted earlier and published a fortnight later in the same volume. Their systematics have almost nothing in common, which is the point of citing both. The longitudinal polarization of the beta electrons themselves, the second of Lee and Yang's three proposals, is [1108] by Mott scattering and [1131] through the circular polarization of the accompanying gamma rays. The neutrino helicity measurement is [1148]; its apparatus and analysis belong to Chapter 98, where the angular-momentum chain is worked out in full.

For the theoretical response, [1118], [1121] and [1130] propose the two-component neutrino independently and within weeks of each other; [1144] and [1162] give the $V - A$ interaction; [965] supplies the parametrization of the muon decay spectrum that the pion-chain experiments were interpreted against, written eight years before the structure it parametrizes was known; and [1123] is the CPT theorem, the symmetry that survives.

For the modern programme: [1519] is the proposal that made atomic parity violation practical by identifying the Z^3 enhancement, and [1946] the caesium measurement that realized it to 0.35 % together with the first anapole moment; [2329] is the isotope-ratio method in ytterbium. [1625] is SLAC E122 and [1643] its kinematic scan — two distinct measurements, giving $\sin^2\theta_W = 0.20 \pm 0.03$ and 0.224 ± 0.020 respectively, and they should be cited separately. [2069] is E158, [2301] is Qweak and [2370] is PREX-2. The Z-pole compilation is [2104], which is also the source for the unresolved discrepancy between the two most precise determinations discussed in Section 110.7.5. The discovery of neutral currents that made all of this possible is [1502], and the model it confirmed is [1206, 1362, 1381].

Numerical inputs are taken from two evaluated compilations and from nowhere else: particle masses, widths and the weak mixing angle from the Particle Data Group [2427], and the fundamental constants — k_B, e, μ_N, m_μ, $m_e c^2$, a_e and α — from CODATA 2022 [2431]. Nuclear mass differences behind the ^{60}Co decay energetics are from the atomic mass evaluation [2393]. Where a number in this chapter is derived rather than quoted, the derivation is given in place and the inputs are named.

Chapter 111

Experiment: CP Violation

Tests Equations (105.46) *and* (105.47) *and Phenomenon 103.31.*
Assuming Theorems 105.34, 105.36 and 105.38.

After parity fell in 1957 (Chapter 110) the combined operation CP was expected to be exact, and the neutral kaon system was built around that expectation: the long-lived state was supposed to be CP-odd and therefore forbidden to decay to two pions. This chapter records the two-pion decay that Christenson, Cronin, Fitch and Turlay found at the Brookhaven alternating-gradient synchrotron in a branching ratio of about 2×10^{-3} [1274], and everything measured since: direct CP violation in kaon decay [1965, 1973], the large asymmetries of the B mesons [2005, 2010], and the small one in charm [2323]. It sits after the weak-interaction and flavour chapters (Chapters 101 and 103) because it is the measurement that forced a third quark generation [1505], and beside Chapter 105, whose CPT theorem is what makes CP violation and T violation the same statement.

The division of labour with those two chapters is worth stating at the outset, because it decides what is and is not repeated here. The two-state formalism of the neutral kaon, the Wigner–Weisskopf elimination of the decay continuum, the definition of ϵ and ϵ', the double ratio, the Jarlskog invariant and the Sakharov conditions are all *derived* in Sections 105.3 and 105.6; the flavour structure that ties them to the quark mixing matrix is Sections 103.4.3 and 103.5. What follows is the experimental side: what the beams and the detectors physically were, how each number was extracted, what background had to be excluded, and what the numbers are with their uncertainties in SI units. Where a result of the theory chapters is needed it is cited by its equation number rather than re-derived, and where the experimental argument requires a derivation of its own — the purity of the beam, the regeneration background, the geometry of the spectrometer, the extraction of $\sin 2\beta$ from a vertex separation — it is carried out here.

Its closing sections take the result out of particle physics and into cosmology: CP violation is one of Sakharov's three conditions for a baryon asymmetry [1358], and the honest arithmetic is that the Cabibbo–Kobayashi–Maskawa phase is many orders of magnitude too small to produce the asymmetry observed in the sky [1882] — which is stated here as an open problem, not papered over.

Remark 111.1 (Units in this chapter, and in its sources). Every equation below carries $\hbar$ and c explicitly and every number is quoted in SI, with the electronvolt — an accepted non-SI unit — used for energies and its joule equivalent given where the comparison matters. The primary literature of this subject does not do this: it works in units in which $\hbar$ and c are set to unity, so that a mass, a momentum and an energy are all quoted in MeV and a lifetime in $1/\mathrm{MeV}$. To map a formula from those papers onto one here, restore one factor of c for every unit of momentum written as an energy, one factor of c^2 for every mass written as an energy, and one factor of $\hbar$ for every inverse energy written as a time. Two conversions recur and are worth fixing once: $1\,\mathrm{MeV} = 1.602\,176\,634 \times 10^{-13}\,\mathrm{J}$ exactly, since the elementary charge

is exact in the present SI [2431], and $\hbar = 6.582\,119\,569 \times 10^{-16}\,\mathrm{eV\,s}$, so a width of 1 MeV is a lifetime of $6.58 \times 10^{-22}\,\mathrm{s}$. No derivation in this chapter is carried out in those units.

111.1 Historical context and the prediction under test

111.1.1 The neutral kaon system

The experiment of 1964 was possible only because an unusual physical system had been identified nine years earlier, and the identification was made on paper before any of it was seen.

Gell-Mann and Pais asked what happens to a neutral particle that is not its own antiparticle when the interaction that produces it conserves a quantum number that the interaction destroying it does not [1067]. The strong interaction makes the neutral kaon in a state of definite strangeness — $|K^0\rangle$ with $S = +1$, $|\bar{K}^0\rangle$ with $S = -1$ — and conserves that label, so the two are distinct particles. The weak interaction does not conserve it, and therefore connects them. The states of definite mass and lifetime are then not the strangeness eigenstates at all but the combinations that diagonalize the weak effective Hamiltonian, which under an exact CP symmetry are the CP eigenstates $|K_1\rangle$ and $|K_2\rangle$ of Equation (105.46).

The consequence Gell-Mann and Pais drew was a prediction of a new particle. Two pions have $CP = +1$ and three neutral pions $CP = -1$ (Equation (105.47)), so if CP is exact the CP-even combination may decay to $\pi\pi$ and the CP-odd one may not. Since three pions barely fit inside the kaon mass, the second must be long-lived. Numerically, with the masses of Table 111.5, the energy released into $3\pi^0$ is

$$Q_{3\pi^0} = m_{K^0}c^2 - 3m_{\pi^0}c^2 = 497.611\,\mathrm{MeV} - 404.930\,\mathrm{MeV} = 92.68\,\mathrm{MeV}\,, \tag{111.1}$$

against $Q_{\pi\pi} = 218.47\,\mathrm{MeV}$ for $\pi^+\pi^-$. The non-relativistic phase-space volume for n bodies sharing a kinetic energy release Q grows as $Q^{(3n-5)/2}$: the momentum delta function leaves $3(n-1)$ independent components, the energy constraint restricts them to a sphere of radius proportional to $\sqrt{Q}$ in that space, and the surface measure of that sphere divided by the gradient of the energy supplies $R^{3(n-1)-1}/R \propto Q^{(3n-5)/2}$. Two bodies therefore give $Q^{1/2}$ and three bodies Q^2, so the three-body channel is far more strongly throttled near its own threshold, and it has less than half the energy release to work with. That is what produces the observed factor of several hundred in lifetime. Lande, Booth, Impeduglia, Lederman and Chinowsky found the long-lived partner at the Brookhaven Cosmotron the following year [1090], with a lifetime some six hundred times that of the known short-lived V particle. The modern values,

$$\tau_S = (8.954 \pm 0.004) \times 10^{-11}\,\mathrm{s}\,, \qquad \tau_L = (5.116 \pm 0.021) \times 10^{-8}\,\mathrm{s}\,, \tag{111.2}$$

give a ratio $\tau_L/\tau_S = 571$ [2427].

The second consequence is the one that makes the system an instrument rather than a curiosity. Because the two mass eigenstates are superpositions of the two strangeness eigenstates, a beam prepared as K^0 evolves into a mixture whose strangeness content oscillates, and the oscillation frequency is the mass difference divided by $\hbar$. This is Equation (105.61). The measured splitting is

$$\Delta m_K c^2 = (3.484 \pm 0.006) \times 10^{-6}\,\mathrm{eV} = 5.582 \times 10^{-25}\,\mathrm{J}\,, \qquad \frac{\Delta m_K c^2}{\hbar} = 5.293 \times 10^{9}/\mathrm{s}\,, \tag{111.3}$$

corresponding to a mass difference of $\Delta m_K = 6.211 \times 10^{-42}\,\mathrm{kg}$, which is $\Delta m_K/m_K = 7.0 \times 10^{-15}$ of the kaon's own mass (Equations (105.60) and (105.63)). No spectrometer

resolves a fractional mass difference of 10^{-14}; an interferometer whose fringe is a phase accumulated over a flight path of metres does, and that is what the neutral kaon is. Everything in this chapter rests on that fact: the quantities measured are interference terms, not energies.

Remark 111.2 (Why the long-lived state was expected to be pure). It is worth being precise about the strength of the pre-1964 expectation, because it is what makes the measurement decisive rather than merely interesting. If CP commutes with the full Hamiltonian then $|K_2\rangle$ is an exact eigenstate of it and the amplitude for $K_2 \to \pi\pi$ is exactly zero — not small, not suppressed by a coupling, but forbidden by a superselection of the same kind that forbids a $0^+ \to 0^-$ electromagnetic transition. There was no parameter to adjust. A single observed event of the right kind, above background and above every mundane explanation, falsifies the hypothesis; and that is what Section 111.4.1 reports.

111.1.2 Regeneration

The mundane explanation that had to be excluded is regeneration, and because it is simultaneously the cleanest demonstration of quantum superposition in the subject and the principal background of the 1964 experiment, it is derived here rather than assumed.

Pais and Piccioni observed that a long-lived beam traversing matter emerges with a short-lived component [1077]. The reason is that the strong interaction sees strangeness and not CP: a $\bar{K}^0$ can convert a nucleon into a hyperon, a K^0 cannot, so the two forward scattering amplitudes differ. A medium that attenuates or phase-shifts the two strangeness components differently is, in the CP basis, a medium that rotates $|K_L\rangle$ into $|K_S\rangle$.

Proposition 111.3 (Coherent regeneration in a thin slab)**.** *Let a K_L beam of momentum p traverse a slab of thickness ℓ containing N scattering centres per unit volume, with complex forward scattering amplitudes f for K^0 and $\bar{f}$ for $\bar{K}^0$, both carrying the SI dimension of length. If ℓ is short compared with the K_S decay length, the emerging state contains a coherent K_S amplitude*

$$\rho = \mathrm{i}\tan\frac{\Delta\varphi}{2} \approx \frac{\mathrm{i}\Delta\varphi}{2} = \mathrm{i}\pi N\left(f-\bar{f}\right)\ell\,\frac{\hbar}{p}\,, \tag{111.4}$$

relative to the surviving K_L amplitude. It vanishes when the medium cannot distinguish the two strangeness states.

Rests on Equations (105.46) *and* (105.64).

Derivation. *Derives Proposition 111.3.*

A wave of reduced wavelength $\hbar/p$ propagating through a medium of N centres per unit volume with forward amplitude f acquires, per unit length, the extra phase $2\pi N f\hbar/p$: this is the optical relation $k^2(n-1) = 2\pi N f$ with $k = p/\hbar$, so that the phase per unit length is $k(n-1) = 2\pi N f/k = 2\pi N f\hbar/p$, and it is the only piece of scattering theory needed. Both sides of the optical relation carry the SI dimension $1/\mathrm{m}^2$, since $[N] = 1/\mathrm{m}^3$ and $[f] = \mathrm{m}$. Over the slab the two strangeness components therefore acquire

$$\varphi = 2\pi N f\ell\,\frac{\hbar}{p}\,, \qquad \bar{\varphi} = 2\pi N \bar{f}\ell\,\frac{\hbar}{p}\,, \qquad \Delta\varphi = \varphi - \bar{\varphi}\,.$$

Write the incident state, neglecting the 10^{-3} admixture ϵ, as $|K_L\rangle = \left(|K^0\rangle - \left|\bar{K}^0\right\rangle\right)/\sqrt{2}$

from Equation (105.46). Applying the two phases and factoring out the common part,

$$\begin{aligned}|\psi\rangle &= \frac{1}{\sqrt{2}}\left(\mathrm{e}^{\mathrm{i}\varphi}\left|K^0\right\rangle - \mathrm{e}^{\mathrm{i}\bar{\varphi}}\left|\bar{K}^0\right\rangle\right)\\ &= \frac{\mathrm{e}^{\mathrm{i}(\varphi+\bar{\varphi})/2}}{\sqrt{2}}\left(\mathrm{e}^{\mathrm{i}\Delta\varphi/2}\left|K^0\right\rangle - \mathrm{e}^{-\mathrm{i}\Delta\varphi/2}\left|\bar{K}^0\right\rangle\right)\\ &= \mathrm{e}^{\mathrm{i}(\varphi+\bar{\varphi})/2}\left[\cos\frac{\Delta\varphi}{2}\,|K_L\rangle + \mathrm{i}\sin\frac{\Delta\varphi}{2}\,|K_S\rangle\right],\end{aligned}$$

where the last line uses $|K_S\rangle = \left(|K^0\rangle + \left|\bar{K}^0\right\rangle\right)/\sqrt{2}$. The ratio of the two amplitudes is Equation (111.4), and it is proportional to $f - \bar{f}$, hence zero for a medium blind to strangeness. Retaining the K_S decay and the K_S–K_L phase difference over the slab replaces the linear ℓ by a bounded oscillating factor of the kind written in Equation (105.64); for the thin, dilute media of interest here that factor reduces to ℓ and the difference is immaterial. Only $|\rho|$ is used below, so the overall phase convention of the thick-slab formula does not enter. □

Two features of Equation (111.4) decide the design of the experiment. The first is that ρ is proportional to N, so a dilute medium regenerates weakly and a vacuum not at all. The second, less obvious, is that the regenerated amplitude is *coherent* with the incident beam: it is forward, it carries the beam's momentum direction, and its decay to $\pi^+\pi^-$ is therefore collinear with the beam — exactly like the signal. Coherent regeneration cannot be removed by an angular cut. Only its incoherent relatives can: diffractive regeneration off a single nucleus, and inelastic production, both leave the reconstructed parent momentum at a measurable angle to the beam, and those are what the collinearity requirement of Section 111.3 eliminates. The coherent part has to be made small, and made small by construction; that is what the helium bag is for, and Proposition 111.5 puts a number on it.

111.1.3 CP as the symmetry that was supposed to survive

The parity experiments of 1957 (Chapter 110) showed that the weak interaction distinguishes left from right maximally: in the beta decay of polarized cobalt the electrons come out preferentially opposite to the nuclear spin [1133], and in the pion–muon–electron chain the muon is fully polarized [1110]. Charge conjugation fares no better, and for the same reason — applied to the left-handed neutrino it produces a left-handed antineutrino, which does not exist.

Landau's response was to propose that the true symmetry is the product [1118]. It is an economical suggestion: CP maps the left-handed neutrino onto the right-handed antineutrino, which does exist, so maximal violation of C and of P separately is exactly compatible with an exact CP. Together with the CPT theorem of local Lorentz-invariant quantum field theory [1123] — proved as Theorem 105.34 — the proposal has three testable consequences, and it is worth listing them because the experiment tested the first and the other two survived:

1. $K_L \to \pi\pi$ is strictly forbidden, the long-lived state being the CP-odd combination and the two-pion state CP-even.
2. A particle and its antiparticle have exactly equal masses and exactly equal total lifetimes (Theorem 105.36, Theorem 105.38), whether or not CP holds; these follow from CPT alone.
3. Any CP violation is, given CPT, a T violation of the same size, so a direct comparison of a process with its motion reverse must show it too — which is the programme of Section 111.7.4.

111.1.4 The observable and what a measurement must deliver

The prediction under test is item (i), and stating it as an experimental requirement fixes the whole design.

A neutral kaon beam is prepared far enough downstream of its production target that the K_S component has decayed to an utterly negligible level. Any $\pi^+\pi^-$ pair observed to emerge from the surviving beam must then satisfy two independent conditions if it is a genuine two-body decay of a beam particle: the invariant mass of the pair, computed under the pion hypothesis, must equal the kaon mass; and the vector sum of the two pion momenta — the reconstructed parent momentum — must be collinear with the beam. A measurement must therefore deliver

- a beam whose K_S contamination is calculably negligible,
- a decay region containing so little material that coherent regeneration cannot fake the signal,
- a spectrometer whose mass resolution separates $\pi^+\pi^-$ from the three-body decays $\pi e\nu$, $\pi\mu\nu$ and $\pi^+\pi^-\pi^0$, which are 10^3 times commoner,
- an angular resolution fine enough that a genuine two-body decay piles up in one bin of $\cos\theta$ while three-body decays spread over many, and
- a normalization sample — the charged decays of the same beam in the same apparatus — so that the answer is a ratio in which the flux and most of the acceptance cancel.

Each of the five is met by a specific piece of the apparatus, and the next two sections take them in that order.

111.2 Apparatus

111.2.1 The Brookhaven beam and the decay volume

The source was the alternating-gradient synchrotron at Brookhaven, delivering protons of $30\,\mathrm{GeV}$ — a kinetic energy of $4.8\times10^{-9}\,\mathrm{J}$ per proton — onto an internal target. A neutral beam was taken off at 30° to the circulating protons, passed through a collimator, swept clear of charged particles by a magnet, and re-collimated; the detector viewed the beam $17.4\,\mathrm{m}$ (57 feet in the units of the original report) from the target [1274]. The momentum spectrum of the surviving neutral kaons peaked near $1.1\,\mathrm{GeV}/c$, which is the value used for every kinematic estimate below.

That distance is the first design requirement, and it is met with an enormous margin.

Proposition 111.4 (The beam is pure long-lived kaons)**.** *At $p = 1.1\,\mathrm{GeV}/c$ the K_S decay length is $5.93\times10^{-2}\,\mathrm{m}$ and the K_L decay length is $33.9\,\mathrm{m}$. A flight path of $17.4\,\mathrm{m}$ is 293 K_S decay lengths and 0.513 K_L decay lengths.*

Rests on Equation (111.2).

Derivation. *Derives Proposition 111.4.*

With $m_Kc^2 = 497.611\,\mathrm{MeV}$ and $pc = 1.1\,\mathrm{GeV}$, the total energy is

$$E = \sqrt{p^2c^2 + m_K^2c^4} = 1.2073\,\mathrm{GeV} = 1.9343\times10^{-10}\,\mathrm{J}\,,$$

so $\gamma = E/(m_Kc^2) = 2.426$, $\beta = pc/E = 0.9111$ and $\gamma\beta = pc/(m_Kc^2) = 2.211$. The laboratory decay length of a state of proper lifetime τ is $\gamma\beta c\tau$; with Equation (111.2),

$$\lambda_S = 2.211\times2.997\,924\,58\times10^{8}\,\mathrm{m/s}\times8.954\times10^{-11}\,\mathrm{s} = 5.934\times10^{-2}\,\mathrm{m}\,,$$
$$\lambda_L = 2.211\times2.997\,924\,58\times10^{8}\,\mathrm{m/s}\times5.116\times10^{-8}\,\mathrm{s} = 33.90\,\mathrm{m}\,.$$

Hence $17.4\,\mathrm{m}/\lambda_S = 293$ and $17.4\,\mathrm{m}/\lambda_L = 0.513$. The surviving K_S fraction is e^{-293}, a number smaller than 10^{-127} and therefore not a background of any kind; the surviving K_L fraction is $\mathrm{e}^{-0.513} = 0.599$, so more than half the long-lived beam is still available where the apparatus looks at it. The flight path is close to the optimum: shortening it would not increase the K_L flux much and would begin to admit K_S, lengthening it would cost K_L flux exponentially for no gain. □

The decay volume itself was a bag of helium at atmospheric pressure. The reason is Proposition 111.3: coherent regeneration scales as the number density of scatterers times the difference of their forward amplitudes, and both factors are minimized by a light, dilute gas. A true vacuum over a volume of several cubic metres was impractical in 1964; helium was the practical approximation to one, and the following estimate is what justifies the choice.

Proposition 111.5 (Helium suppresses the regeneration background)**.** *Over a decay path of* $2\,\mathrm{m}$ *of helium at atmospheric pressure, the coherently regenerated* K_S *amplitude is of order* 3×10^{-5}*, that is about one per cent of the signal amplitude* $|\eta_{+-}| = 2.2\times10^{-3}$*; in air it would have been several times larger.*

Rests on Equation (111.4)*.*

Derivation. Derives Proposition 111.5.

Insert into Equation (111.4). The reduced wavelength of the beam is

$$\frac{\hbar}{p} = \frac{1.054\,571\,817\times10^{-34}\,\mathrm{J\,s}}{5.879\times10^{-19}\,\mathrm{kg\,m/s}} = 1.794\times10^{-16}\,\mathrm{m}\,,$$

where $p = 1.1\,\mathrm{GeV}/c$ has been converted using $1\,\mathrm{eV} = 1.602\,176\,634\times10^{-19}\,\mathrm{J}$ [2431]. Helium at $101.3\,\mathrm{kPa}$ and room temperature, taken as $293\,\mathrm{K}$, has $N = P/(k_{\mathrm{B}}T) = 2.50\times10^{25}/\mathrm{m}^3$ atoms; the familiar $2.69\times10^{25}/\mathrm{m}^3$ is the same pressure at $273.15\,\mathrm{K}$. The amplitude difference $\left|f-\bar{f}\right|$ is not calculable from first principles here, but it cannot exceed the geometric scale of the nucleus, so taking $\left|f-\bar{f}\right| \sim 1\times10^{-15}\,\mathrm{m}$ is a generous bound. With $\ell = 2\,\mathrm{m}$,

$$|\rho| \sim \pi\times2.50\times10^{25}/\mathrm{m}^3\times1\times10^{-15}\,\mathrm{m}\times2\,\mathrm{m}\times1.794\times10^{-16}\,\mathrm{m} = 3\times10^{-5}\,.$$

Air is denser in nucleons by a factor $1.20\,\mathrm{kg/m^3}/0.166\,\mathrm{kg/m^3} = 7.2$ — both densities at $293\,\mathrm{K}$ and $101.3\,\mathrm{kPa}$, since setting a room-temperature air density against the tabulated $273.15\,\mathrm{K}$ helium value of $0.179\,\mathrm{kg/m^3}$ would understate the ratio as 6.7 — and its nuclei are heavier, so $\left|f-\bar{f}\right|$ is larger by roughly $(14/4)^{2/3}\approx2.3$ per atom while the atom density rises by about a factor of two; the product is an increase of four or five. The regenerated amplitude adds coherently to η_{+-}, so a regeneration amplitude that is one per cent of η_{+-} perturbs the measured rate by about two per cent, while one that is five per cent of it perturbs the rate by about ten per cent — the difference between a negligible correction and a systematic error comparable with the statistical one. This is an order-of-magnitude estimate and was treated as such: the experiment also measured the regeneration directly by inserting known material, which is the calibration described in Section 111.3.3. □

111.2.2 The two-arm spectrometer

The detector was two symmetric arms viewing the helium bag, each consisting of a bending magnet with a spark chamber before it and a spark chamber after it, followed by a water Cherenkov counter and a scintillation counter [1274]. Each arm measured one charged particle: its direction from the two spark chambers, its momentum from the deflection between them, and its identity, loosely, from the Cherenkov signal. The two arms together therefore delivered the two three-vectors $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ from which every quantity in the analysis is built.

The arms sat at about 22° to the beam on either side, and that angle is not arbitrary.

Proposition 111.6 (Why the arms sit near 22 degrees)**.** *A K^0 of momentum $1.1\,\mathrm{GeV}/c$ decaying to $\pi^+\pi^-$ sends both pions to the same laboratory angle 20.5°, on opposite sides of the beam, when they are emitted transversely to the beam in the kaon rest frame. That symmetric configuration is the one a pair of symmetric arms can accept with both arms at the same angle, and it is what fixes that angle. A single pion can reach as far as 41.9° from the beam, attained at $\cos\theta^* = -0.909$, but only when its partner has collapsed to 4.9°.*

Rests on Proposition 111.4.

Derivation. *Derives Proposition 111.6.*

In the kaon rest frame the two pions are back to back with momentum

$$p^*c = \sqrt{\frac{m_K^2c^4}{4} - m_\pi^2c^4} = \sqrt{(248.806\,\mathrm{MeV})^2 - (139.570\,\mathrm{MeV})^2} = 205.97\,\mathrm{MeV}\,,$$

that is $p^* = 1.101 \times 10^{-19}\,\mathrm{kg\,m/s}$, and each carries energy $E^* = m_Kc^2/2 = 248.806\,\mathrm{MeV}$. Boosting to the laboratory with $\gamma = 2.426$, $\beta = 0.9111$ from Proposition 111.4, a pion emitted at 90° in the rest frame keeps its transverse momentum $p_\perp = p^*$ and acquires

$$p_\parallel = \gamma\beta\frac{E^*}{c} = 2.211 \times 248.806\,\mathrm{MeV}/c = 550.1\,\mathrm{MeV}/c\,,$$

so its laboratory angle is $\arctan(205.97/550.1) = 20.5^\circ$. Its partner, emitted transversely in the opposite sense, arrives at the same 20.5° on the other side of the beam: this is the only configuration in which the two pions make equal angles with the beam, and it is therefore the configuration a symmetric two-arm spectrometer is built around.

The maximum laboratory angle of a *single* pion is a different number and is worth separating out. With $u = \cos\theta^*$ the laboratory angle satisfies

$$\tan\theta = \frac{p^*\sqrt{1-u^2}}{\gamma\,(p^*u + \beta E^*/c)} = \frac{1}{\gamma}\,\frac{\sqrt{1-u^2}}{u + \beta/\beta^*}\,, \qquad \beta^* = \frac{p^*c}{E^*} = 0.8278\,.$$

Differentiating the last fraction gives $-u(u+\beta/\beta^*) - (1-u^2) = 0$, that is $u = -\beta^*/\beta = -0.909$; a maximum exists precisely because $\beta = 0.9111$ exceeds β^*, so that the origin lies outside the ellipse of laboratory momenta and the tangent from it touches at $u = -0.909$ rather than at the transverse point. There $\tan\theta = 0.8968$ and $\theta = 41.9^\circ$, while the partner at $u = +0.909$ is thrown forward to $\tan\theta = 0.0857$, that is 4.9°. Two arms placed at 22° therefore sit on the symmetric configuration, accept a wide band of rest-frame angles about it at once, and keep clear of the beam; the wide-angle pairs are lost, but they are lost together with their forward partners and cost only acceptance, not bias. □

The two remaining pieces of the apparatus are there to reject the common decays. The Cherenkov counters used water, of refractive index $n = 1.33$, whose threshold velocity is $\beta_{\mathrm{th}} = 1/n = 0.752$; for a pion this is a momentum of

$$p_{\mathrm{th}} = \frac{m_\pi c\,\beta_{\mathrm{th}}}{\sqrt{1-\beta_{\mathrm{th}}^2}} = 159\,\mathrm{MeV}/c\,, \tag{111.5}$$

so essentially every pion in the accepted momentum band radiates while the counters remain useful in the trigger. The scintillators supplied the fast timing that defined the coincidence.

The resolutions matter more than any of this, because the entire argument of the experiment is a two-dimensional separation. The spectrometer reconstructed the invariant mass of the pair to about $4\,\mathrm{MeV}/c^2$ and the direction of the reconstructed parent to a few milliradians; the signal bin used in the analysis, $\cos\theta > 0.9999$, corresponds to $\theta < 14.1\,\mathrm{mrad}$ (Table 111.1).

Table 111.1: Parameters of the Christenson–Cronin–Fitch–Turlay apparatus and beam [1274], with the derived kinematic quantities of Propositions 111.4 and 111.6. Imperial figures in the original report are converted; the 17.4 m flight path is the 57 feet of the published layout. Quantities marked "derived" are computed in this chapter from the beam momentum and the particle masses of Table 111.5, not taken from the source.

Quantity	Value	Note
Proton energy (AGS)	$30\,\mathrm{GeV}$	$4.8 \times 10^{-9}\,\mathrm{J}$
Neutral-beam production angle	30°	to circulating protons
Target-to-detector flight path	$17.4\,\mathrm{m}$	57 feet as published
Beam momentum at the peak	$1.1\,\mathrm{GeV}/c$	$5.879 \times 10^{-19}\,\mathrm{kg\,m/s}$
Lorentz factor of the beam	$\gamma = 2.426$	derived
K_S decay length	$5.93 \times 10^{-2}\,\mathrm{m}$	derived
K_L decay length	$33.9\,\mathrm{m}$	derived
Flight path in K_S decay lengths	293	derived
Decay volume	helium, $101.3\,\mathrm{kPa}$	bag, not a vacuum vessel
Spectrometer arm angle	22°	each side of the beam
Symmetric-pair laboratory angle	20.5°	derived
Maximum single-pion angle	41.9°	derived, at $\cos\theta^* = -0.909$
Rest-frame pion momentum	$206\,\mathrm{MeV}/c$	derived
Cherenkov radiator	water, $n = 1.33$	pion threshold $159\,\mathrm{MeV}/c$
Invariant-mass resolution	$\approx 4\,\mathrm{MeV}/c^2$	pion hypothesis
Signal angular bin	$\cos\theta > 0.9999$	$\theta < 14.1\,\mathrm{mrad}$

111.3 Procedure

111.3.1 Trigger and event recording

An event was accepted when the scintillation and Cherenkov counters of both arms fired in coincidence, indicating one charged particle in each arm within the resolving time of the counters. The coincidence pulsed the spark chambers to a few tens of kilovolts, so that the ionization trails left by the two particles broke down into visible sparks, and triggered a camera which photographed the chambers [1274]. The recorded object was therefore a photograph, and the analysis a measurement of positions on film, followed by a reconstruction.

This is worth stating plainly because it bounds the sample size. The experiment could not record a large fraction of the beam; it recorded about twenty-two thousand charged decays in total, and the signal it found is forty-five events. Everything about the argument — the use of a two-dimensional discriminant rather than a single cut, the reliance on sidebands for the background, the insistence on a control sample — follows from having a sample of that size and no possibility of enlarging it quickly.

111.3.2 Reconstruction in the mass-angle plane

From the two measured momenta the analysis formed two numbers. The first is the invariant mass of the pair under the hypothesis that both particles are pions,

$$m_{\pi\pi}^2 c^4 = 2m_\pi^2 c^4 + 2\left(E_1 E_2 - \boldsymbol{p}_1 \cdot \boldsymbol{p}_2 c^2\right), \qquad E_i = \sqrt{p_i^2 c^2 + m_\pi^2 c^4}\,. \tag{111.6}$$

The second is the angle θ between the reconstructed parent momentum $\boldsymbol{p}_1 + \boldsymbol{p}_2$ and the beam direction, reported as $\cos\theta$.

The discriminating power lies in requiring both at once, and the reason is a kinematic identity rather than a statistical convention.

Proposition 111.7 (A two-body decay is a point in the mass-angle plane)**.** *If a beam particle of mass m_K decays to exactly two pions, then $m_{\pi\pi} = m_K$ and $\cos\theta = 1$ identically. If it decays to three or more bodies of which two are detected, then $m_{\pi\pi} < m_K$, and $\cos\theta < 1$ unless the undetected momentum happens to vanish.*

Rests on Equation (111.6)*.*

Derivation. Derives Proposition 111.7.

For a two-body decay, four-momentum conservation gives $p_K^\mu = p_1^\mu + p_2^\mu$, so the invariant mass of the pair is the parent mass and its three-momentum is the parent's, which points along the beam. Both statements are exact and both are independent of the decay angle, which is what makes the pair $(m_{\pi\pi}, \cos\theta)$ a point rather than a distribution.

For a decay $K \to \pi\pi X$ with X undetected of four-momentum q^μ and mass $m_X \geq 0$,

$$m_{\pi\pi}^2 c^4 = (p_K - q)^2 c^2 = m_K^2 c^4 + m_X^2 c^4 - 2p_K \cdot q\, c^2\,.$$

Here a four-momentum carries the SI dimension of momentum, so that $p^2 = m^2 c^2$ and every term above is an energy squared. In the kaon rest frame $p_K \cdot q\, c^2 = m_K c^2 E_X^*$ with $E_X^* \geq m_X c^2$, and E_X^* is bounded below by the kinematic minimum for the given final state, so $m_{\pi\pi} < m_K$ strictly. The missing three-momentum likewise tilts $\boldsymbol{p}_1 + \boldsymbol{p}_2$ away from the beam, by an angle of order $q_\perp/(p_1 + p_2)$; for the dominant K_L modes — $\pi^\pm e^\mp \nu$, $\pi^\pm \mu^\mp \nu$ and $\pi^+\pi^-\pi^0$ — the missing transverse momentum is of order $100\,\mathrm{MeV}/c$ against a total of order $1\,\mathrm{GeV}/c$, giving angles of order 100 mrad, seven times the signal bin. The two variables therefore fail together for background and hold together for signal, and it is their *joint* behaviour, not the size of the excess, that carries the argument. □

For the semileptonic modes there is a second, independent handicap. Because the electron or muon is assigned the pion mass in Equation (111.6), its energy is overestimated, and the reconstructed mass is pushed further from m_K rather than towards it — so the misassignment cannot manufacture a signal, only remove background from the peak.

111.3.3 Backgrounds and the control samples

Three backgrounds had to be dealt with, and each was measured rather than modelled.

Three-body decays under the peak. The overwhelming majority of the recorded events are $\pi e\nu$, $\pi\mu\nu$ and 3π, which populate the whole $(m_{\pi\pi}, \cos\theta)$ plane. Their density in the signal region was fixed empirically from mass sidebands: the $\cos\theta$ distribution was formed in the mass intervals immediately below and above the kaon mass, where no two-body signal can exist, and interpolated across. The interpolation is safe because the background varies smoothly in $m_{\pi\pi}$ on the scale of the mass resolution while the signal does not.

Incoherent regeneration. Diffractive and inelastic regeneration in the residual material produce a K_S whose momentum is not collinear with the beam, and are therefore removed by the same angular requirement that defines the signal, as Proposition 111.7 makes explicit.

Coherent regeneration. This one is collinear and cannot be cut away, so it was measured. Slabs of material were deliberately placed in the beam, producing a large and unmistakable coherent $K_S \to \pi^+\pi^-$ signal at $m_{\pi\pi} = m_K$ and $\cos\theta = 1$; that sample served two purposes at once. It calibrated the mass and angular resolution of the spectrometer on real two-body decays of the right mass and the right momentum spectrum — the ideal control, since it is the signal in every respect except its origin — and it fixed the constant of proportionality in Equation (111.4) for the materials involved, from which the residual regeneration in helium follows by scaling. The observed forward peak did not scale with the amount of material in the beam, which is the direct experimental statement that it is not regeneration.

Finally, the normalization. The denominator of the published ratio is the number of *all* charged decays of the same beam recorded in the same apparatus. Taking a ratio to that sample cancels the beam flux, the live time, the film-scanning efficiency and a large part of the geometric acceptance, which is why the result could be quoted to twenty per cent from forty-five events.

111.4 Observations and data

111.4.1 The forward peak at the kaon mass

In the mass interval containing the kaon mass, and only there, the $\cos\theta$ distribution shows a sharp excess in the single bin $\cos\theta > 0.9999$. In the two adjacent mass intervals, ten MeV below and above, the same bin shows nothing above the smooth background. The excess amounts to 45 ± 9 events, and the sample of all charged decays recorded in the same apparatus numbers 22 700 [1274].

Phenomenon 111.8 (The long-lived neutral kaon decays to two pions)**.** A neutral kaon beam allowed to propagate far enough that every short-lived component has decayed still produces two-charged-pion events whose reconstructed invariant mass is the kaon mass and whose reconstructed parent momentum is collinear with the beam: an excess of 45 ± 9 events over an estimated background of about twelve in the forward bin, corresponding to

$$\frac{\Gamma\left(K_L \to \pi^+\pi^-\right)}{\Gamma\left(K_L \to \text{all charged}\right)} = \frac{45 \pm 9}{22\,700} = (2.0 \pm 0.4) \times 10^{-3} \tag{111.7}$$

[1274]. The decay is forbidden if CP is a symmetry.

Rests on Equations (105.46) *and* (105.47)*.*

Derivation. *Derives Phenomenon 111.8.*

The two pions from a spinless parent are in a state of zero relative orbital angular momentum, $L = 0$. Space inversion multiplies such a state by the product of the two intrinsic parities and by $(-1)^L$; the pion is pseudoscalar, so $P\,|\pi^+\pi^-\rangle = (-1)^2(-1)^L\,|\pi^+\pi^-\rangle = +\,|\pi^+\pi^-\rangle$. Charge conjugation exchanges the two pions, which for relative angular momentum L contributes the same factor $(-1)^L = +1$. Hence

$$CP\left|\pi^+\pi^-\right\rangle = +\left|\pi^+\pi^-\right\rangle ,$$

the final state is CP-even. Before 1964 the long-lived kaon was identified with the CP-odd combination $|K_2\rangle = \left(|K^0\rangle - \left|\bar{K}^0\right\rangle\right)/\sqrt{2}$, for which $CP\,|K_2\rangle = -\,|K_2\rangle$, precisely because that assignment forbids the two-pion channel and so explains the observed factor of some six hundred between the two lifetimes. An interaction commuting with CP cannot connect states of opposite CP eigenvalue, so Equation (111.7) establishes that the weak interaction does not commute with CP. What the observation does not by itself decide is whether the violation lies in the mixing that defines the long-lived state or in the decay amplitude; that is the question of Section 111.5.1. □

The 1964 numbers and their modern successors are collected in Table 111.2. One comparison in that table deserves comment, because it is the only place where the original measurement can be checked against a modern one and the two are not quoted in the same variable. The published ratio has "all charged decays" in its denominator, whereas the modern branching fraction has all decays. The charged modes of the K_L — $\pi^\pm e^\mp\nu$, $\pi^\pm\mu^\mp\nu$, $\pi^+\pi^-\pi^0$ and $\pi^+\pi^-$ itself — account for 0.8033 ± 0.0014 of the total width [2427], the quoted uncertainty being the quadrature combination of the uncertainties on those four branching fractions. It is negligible beside the twenty per cent statistical error on the 1964 ratio, so

$$\mathcal{B}\left(K_L \to \pi^+\pi^-\right)_{1964} = (2.0 \pm 0.4) \times 10^{-3} \times 0.8033 = (1.6 \pm 0.3) \times 10^{-3} , \tag{111.8}$$

against the present $(1.967 \pm 0.010) \times 10^{-3}$ [2427]: agreement at 1.2 standard deviations. The original measurement was correct, and it was correct to twenty per cent on forty-five events.

Table 111.2: The 1964 measurement and its modern counterparts. The first block is what Christenson, Cronin, Fitch and Turlay reported [1274]; the second is the current world average [2427]. The conversion between the two normalizations is Equation (111.8). Uncertainties are as quoted by the sources; the 1964 error is dominated by the counting statistics of the forty-five-event excess.

Quantity	Value	Year
Charged decays recorded	22 700	1964
Excess in the forward mass peak	45 ± 9 events	1964
Estimated background in that bin	≈ 12 events	1964
$\Gamma(K_L \to \pi^+\pi^-)/\Gamma(K_L \to \text{charged})$	$(2.0 \pm 0.4) \times 10^{-3}$	1964
Converted branching fraction	$(1.6 \pm 0.3) \times 10^{-3}$	1964
$\mathcal{B}(K_L \to \pi^+\pi^-)$	$(1.967 \pm 0.010) \times 10^{-3}$	2024
$\mathcal{B}(K_L \to \pi^0\pi^0)$	$(8.64 \pm 0.06) \times 10^{-4}$	2024
Charged fraction of the K_L width	0.8033 ± 0.0014	2024

111.4.2 The branching ratio and the parameter epsilon

The branching fraction is not the natural variable. What the theory predicts is an *amplitude* ratio, because the two-pion final state is reached from the CP-even admixture in the

long-lived state, and the natural comparison is with the same final state reached from the short-lived one. The definitions are Equation (105.65):

$$|\eta_{+-}|^2 = \frac{\Gamma\left(K_L \to \pi^+\pi^-\right)}{\Gamma\left(K_S \to \pi^+\pi^-\right)} = \frac{\mathcal{B}\left(K_L \to \pi^+\pi^-\right)\tau_S}{\mathcal{B}\left(K_S \to \pi^+\pi^-\right)\tau_L}\,, \tag{111.9}$$

the second form following because a branching fraction is a partial width divided by a total width and the total widths are $\hbar/\tau$. Inserting $\mathcal{B}(K_L \to \pi^+\pi^-) = 1.967 \times 10^{-3}$, $\mathcal{B}(K_S \to \pi^+\pi^-) = 0.6920$ and the lifetimes of Equation (111.2) gives

$$|\eta_{+-}|^2 = \frac{1.967 \times 10^{-3} \times 8.954 \times 10^{-11}\,\mathrm{s}}{0.6920 \times 5.116 \times 10^{-8}\,\mathrm{s}} = 4.98 \times 10^{-6}\,,$$

that is $|\eta_{+-}| = 2.23 \times 10^{-3}$, which is the measured value in Table 111.3. The arithmetic is worth seeing once: the tiny branching fraction and the enormous lifetime ratio very nearly cancel, and what survives is a number of order 10^{-3} that has nothing to do with either.

The phase of η_{+-} is measured independently, by observing the interference between the K_S and K_L amplitudes in a beam containing both, and it comes out at 43.51°. That value is predicted, not fitted: Equation (105.59) gives $\arg\epsilon = \arctan\left[2\Delta m_K c^2 \tau_S/\hbar\right]$, and with Equation (111.3) and Equation (111.2),

$$\arctan\frac{2 \times 3.484 \times 10^{-6}\,\mathrm{eV}}{7.3510 \times 10^{-6}\,\mathrm{eV}} = 43.47^\circ\,, \tag{111.10}$$

where $\hbar/\tau_S = 7.3510 \times 10^{-6}\,\mathrm{eV}$ is the K_S full width. Equivalently, dividing numerator and denominator by two, $\arctan\left[\Delta m_K c^2/(\hbar/2\tau_S)\right]$ with the half width $3.6755 \times 10^{-6}\,\mathrm{eV}$: the factor of two in Equation (105.59) is the one that turns the half width into the full width, and must not be applied twice.

The agreement of a measured phase with a number built out of a mass difference and a lifetime is one of the sharper consistency checks in the subject, and it is what licenses the identification of η_{+-} with the mixing parameter ϵ of Equation (105.54) up to the small direct term of Section 111.6.

Table 111.3: Measured parameters of the neutral kaon system [2427], with SI equivalents where the customary unit is not SI. The mass values are those of the machine-readable Particle Data Group release; the mass difference, the lifetimes and the CP parameters are from the same review. The last column gives the year of the compilation, not of the first measurement; the years of first measurement are in Table 111.4.

Quantity	Value	SI equivalent	Year
$m_{K^0}c^2$	$(497.611 \pm 0.013)\,\mathrm{MeV}$	$7.9726 \times 10^{-11}\,\mathrm{J}$	2024
m_{K^0}	—	$8.8707 \times 10^{-28}\,\mathrm{kg}$	2024
τ_S	$(8.954 \pm 0.004) \times 10^{-11}\,\mathrm{s}$	—	2024
τ_L	$(5.116 \pm 0.021) \times 10^{-8}\,\mathrm{s}$	—	2024
$\Delta m_K c^2$	$(3.484 \pm 0.006) \times 10^{-6}\,\mathrm{eV}$	$5.582 \times 10^{-25}\,\mathrm{J}$	2024
$\Delta m_K c^2/\hbar$	—	$5.293 \times 10^{9}/\mathrm{s}$	2024
$\lvert\eta_{+-}\rvert$	$(2.232 \pm 0.011) \times 10^{-3}$	dimensionless	2024
$\arg\eta_{+-}$	43.51°	0.7594 rad	2024
$\lvert\epsilon\rvert$	$(2.228 \pm 0.011) \times 10^{-3}$	dimensionless	2024
δ_L	$(3.32 \pm 0.06) \times 10^{-3}$	dimensionless	2024
$\mathrm{Re}(\epsilon'/\epsilon)$	$(1.66 \pm 0.23) \times 10^{-3}$	dimensionless	2024

111.4.3 Charge asymmetry in semileptonic decay

The two-pion decay establishes that CP is violated; it does not by itself give a sign. The observable that does is the charge asymmetry of the semileptonic decays of the same long-lived state, and it deserves attention out of all proportion to its size.

Phenomenon 111.9 (Charge asymmetry in semileptonic kaon decay)**.** The long-lived kaon decays slightly more often to $\pi^- e^+ \nu_e$ than to $\pi^+ e^- \bar{\nu}_e$:

$$\delta_L = \frac{\Gamma(\pi^- e^+ \nu_e) - \Gamma(\pi^+ e^- \bar{\nu}_e)}{\Gamma(\pi^- e^+ \nu_e) + \Gamma(\pi^+ e^- \bar{\nu}_e)} = (3.32 \pm 0.06) \times 10^{-3} \tag{111.11}$$

[2427]. Small as it is, this is an absolute, convention-free distinction between matter and antimatter: which sign of lepton charge is the commoner one can be communicated to a correspondent with whom we share no sample and no convention.

Rests on Equation (105.56) *and Theorem 105.34.*

Derivation. *Derives Phenomenon 111.9.*

Let the long-lived state be $|K_L\rangle \propto (1+\epsilon)\,|K^0\rangle - (1-\epsilon)\,\left|\bar{K}^0\right\rangle$ with $|\epsilon| \ll 1$, which is Equation (105.56). The selection rule $\Delta S = \Delta Q$ allows $K^0 \to \pi^- e^+ \nu_e$ and $\bar{K}^0 \to \pi^+ e^- \bar{\nu}_e$ but not the charge-exchanged pair, and CPT makes the two allowed amplitudes equal in magnitude. The two rates are therefore proportional to $|1+\epsilon|^2$ and $|1-\epsilon|^2$, so

$$\delta_L = \frac{|1+\epsilon|^2 - |1-\epsilon|^2}{|1+\epsilon|^2 + |1-\epsilon|^2} = \frac{2\,\mathrm{Re}\,\epsilon}{1+|\epsilon|^2} \approx 2\,\mathrm{Re}\,\epsilon\,.$$

With $|\epsilon| = 2.228 \times 10^{-3}$ and its phase $43.5°$ [2427], $2\,\mathrm{Re}\,\epsilon = 2 \times 2.228 \times 10^{-3} \times \cos 43.5° = 3.23 \times 10^{-3}$, against the measured $(3.32 \pm 0.06) \times 10^{-3}$ of Equation (111.11). The asymmetry is thus the same ϵ that Phenomenon 111.8 exhibits, seen in an entirely different channel — which is what makes the mixing interpretation testable rather than merely available. □

The convention-free character of the statement is worth spelling out, since it is the one experimental result in this book that changes what can be said about the words used to describe the world. Before 1964 the labels "matter" and "antimatter" were conventional: C violation and P violation alone leave open the possibility that what we call matter is simply the mirror image of what a distant correspondent calls matter, because a mirror reflection can be compensated by an exchange of particles for antiparticles. Equation (111.11) closes that possibility. The instruction "prepare a long-lived neutral kaon beam, count the charge of the emitted lepton, and call the commoner sign positive" can be transmitted in words, needs no shared sample, and gives the same answer everywhere.

111.4.4 The measured CP-violating observables

The remaining sections of the chapter analyse one measurement at a time. Table 111.4 collects them all first, with the year and the source of each, so that the numbers can be read against one another; Table 111.5 gives the masses used throughout, and Table 111.6 the three mixing frequencies in SI. Every value is quoted as its source quotes it, with the uncertainties separated into statistical and systematic where the source separates them.

111.5 Interpretation

111.5.1 Indirect versus direct violation

The 1964 result establishes that the weak interaction does not commute with CP. It does not say where the violation sits, and there are two places available.

Table 111.4: Measured CP- and T-violating observables, in the order they are discussed. Where two uncertainties are given the first is statistical and the second systematic. All the quantities are dimensionless except the neutron electric dipole moment. The world-average entries are the compilations of the Particle Data Group [2427], which absorb later data than the discovery papers in the same block.

System	Observable	Value	Year	Source
K	$\Gamma_{\pi\pi}/\Gamma_{\text{charged}}$	$(2.0 \pm 0.4) \times 10^{-3}$	1964	[1274]
K	$\lvert\epsilon\rvert$	$(2.228 \pm 0.011) \times 10^{-3}$	2024	[2427]
K	δ_L	$(3.32 \pm 0.06) \times 10^{-3}$	2024	[2427]
K	$\mathrm{Re}(\epsilon'/\epsilon)$, NA31	$(3.3 \pm 1.1) \times 10^{-3}$	1988	[1806]
K	$\mathrm{Re}(\epsilon'/\epsilon)$, NA48	$(18.5 \pm 4.5 \pm 5.8) \times 10^{-4}$	1999	[1973]
K	$\mathrm{Re}(\epsilon'/\epsilon)$, KTeV	$(28.0 \pm 3.0 \pm 2.8) \times 10^{-4}$	1999	[1965]
K	$\mathrm{Re}(\epsilon'/\epsilon)$, average	$(1.66 \pm 0.23) \times 10^{-3}$	2024	[2427]
K	A_T, CPLEAR	$(6.6 \pm 1.3 \pm 1.0) \times 10^{-3}$	1998	[1948]
B	$\sin 2\beta$, BaBar	$0.59 \pm 0.14 \pm 0.05$	2001	[2010]
B	$\sin 2\beta$, Belle	$0.99 \pm 0.14 \pm 0.06$	2001	[2005]
B	$\sin 2\beta$, average	0.708 ± 0.011	2024	[2427]
B	ΔS_T^+, BaBar	$-1.37 \pm 0.14 \pm 0.06$	2012	[2197]
B	$A_{CP}(B^0 \to K^+\pi^-)$	$-0.080 \pm 0.007 \pm 0.003$	2013	[2202]
B_s	$A_{CP}(B_s^0 \to K^-\pi^+)$	$0.27 \pm 0.04 \pm 0.01$	2013	[2202]
D	ΔA_{CP}	$(-15.4 \pm 2.9) \times 10^{-4}$	2019	[2323]
n	$\lvert d_n\rvert$ (90 % CL)	$< 2.9 \times 10^{-47}\,\mathrm{C\,m}$	2020	[2346]
CKM	J	$\left(3.08^{+0.15}_{-0.13}\right) \times 10^{-5}$	2024	[2074, 2427]
Cosmos	η	$(6.12 \pm 0.04) \times 10^{-10}$	2020	[2348]

Table 111.5: Masses used in the kinematic derivations of this chapter, taken from the machine-readable Particle Data Group release [2427], with the SI conversion made using the exact elementary charge and speed of light of the present SI [2431]. The energy equivalent is the value the source quotes; the mass in kilograms is that value divided by c^2.

Particle	mc^2	mc^2 (J)	m (kg)
K^0	$(497.611 \pm 0.013)\,\mathrm{MeV}$	7.9726×10^{-11}	8.8707×10^{-28}
$K^\pm$	$(493.677 \pm 0.015)\,\mathrm{MeV}$	7.9096×10^{-11}	8.8006×10^{-28}
$\pi^\pm$	$(139.570\,39 \pm 0.000\,18)\,\mathrm{MeV}$	2.2362×10^{-11}	2.4881×10^{-28}
π^0	$(134.9768 \pm 0.0005)\,\mathrm{MeV}$	2.1626×10^{-11}	2.4062×10^{-28}
J/ψ	$(3.096\,900 \pm 0.000\,006)\,\mathrm{GeV}$	4.9618×10^{-10}	5.5207×10^{-27}

Table 111.6: Mixing frequencies of the three neutral-meson systems in which CP violation has been established, expressed as an energy, as a frequency and as a mass [2427]. The oscillation frequency is $\Delta m\,c^2/\hbar$ throughout, with $\hbar = 1.054\,571\,817 \times 10^{-34}\,\mathrm{J\,s}$ [2431]. The fractional column is $\Delta m/m$, and it is the reason the neutral meson is the most sensitive interferometer available.

System	$\Delta m\,c^2$	$\Delta m\,c^2$ (J)	$\Delta m\,c^2/\hbar$	$\Delta m/m$
K^0	$(3.484 \pm 0.006) \times 10^{-6}\,\mathrm{eV}$	5.582×10^{-25}	$5.293 \times 10^{9}/\mathrm{s}$	7.0×10^{-15}
B^0	$(3.334 \pm 0.013) \times 10^{-4}\,\mathrm{eV}$	5.341×10^{-23}	$5.065 \times 10^{11}/\mathrm{s}$	6.3×10^{-14}
B_s^0	$(1.1693 \pm 0.0004) \times 10^{-2}\,\mathrm{eV}$	1.8734×10^{-21}	$1.7765 \times 10^{13}/\mathrm{s}$	2.2×10^{-12}

It may sit in the *state*. If the long-lived eigenvector of the effective Hamiltonian is not the CP-odd combination $|K_2\rangle$ but carries a small admixture $\epsilon\,|K_1\rangle$ of the CP-even one, then the two-pion decay proceeds entirely from that admixture, using a decay amplitude that itself respects CP perfectly. This is CP violation *in the mixing*, and it is a single complex number ϵ for the whole system.

Or it may sit in the *decay*. If the amplitudes for $K^0 \to \pi\pi$ and $\bar{K}^0 \to \pi\pi$ differ in more than the strong rescattering phase they are obliged to share, then even an exact CP eigenstate would decay asymmetrically. This is CP violation *in the decay*, or direct CP violation, and by Equation (105.72) it is measured by a second complex number ϵ'.

The distinction is not academic, and for thirty-five years it was the open question of the subject, because Wolfenstein had immediately supplied a hypothesis in which the second number is exactly zero [1300]. Suppose there exists a new interaction, enormously weaker than the weak interaction, which changes strangeness by two units and violates CP maximally. Acting once, it contributes to the off-diagonal element M_{12} of the mass matrix (Equation (105.50)) and so to ϵ; it never contributes to a decay, because the decays are $\Delta S = 1$ transitions. Such a "superweak" interaction reproduces every kaon measurement then available with one free parameter and predicts

$$\epsilon' = 0\,, \qquad \eta_{+-} = \eta_{00} = \epsilon\,, \qquad \arg\epsilon = \arctan\frac{2\Delta m_K c^2 \tau_S}{\hbar}\,, \tag{111.12}$$

and the last of these is the phase relation verified numerically in Equation (111.10). That agreement is worth pausing on: the superweak hypothesis was not a desperate one, it fitted the phase correctly, and the experiment that killed it had to measure a difference between η_{00} and η_{+-} of a few parts in a thousand of a quantity that is itself a few parts in a thousand. That is a measurement at the 10^{-6} level in the ratio of rates, and it is why it took thirty-five years. It is Section 111.6.

111.5.2 The CKM matrix and the Jarlskog invariant

The alternative to a new interaction is that the violation is already present in the interaction we have, and this is what Kobayashi and Maskawa proposed [1505]. Their argument is arithmetic and is proved as Theorem 105.25: a unitary $N \times N$ quark mixing matrix, after all the rephasings of the quark fields have been used up, retains $N(N-1)/2$ real angles and $(N-1)(N-2)/2$ phases. For $N = 2$ — the Cabibbo case [1248], one angle and no phase — the matrix can be made real, and a real Lagrangian conserves CP. For $N = 3$ there is exactly one phase, and it cannot be rotated away.

The status of this as a prediction should not be softened. In 1973 three quarks were known. The measurement being explained was a branching ratio of 2×10^{-3} in a kaon. The inference was that three more quarks exist. The charm quark appeared the following year (Section 103.2.3), the bottom quark in 1977 and the top quark in 1995, and the Nobel award of 2008 to Kobayashi and Maskawa recognized an argument, not a discovery of a particle.

Two further pieces of apparatus, in the theoretical sense, make the phase measurable. The first is a parametrization in which the observed hierarchy of the mixing angles is manifest: Wolfenstein's expansion in $\lambda = \sin\theta_C$ [1733], whose four parameters, measured, are Equation (103.24). The second is a rephasing-invariant measure, since the phase δ itself depends on the convention chosen for the quark fields and is therefore not an observable. Jarlskog's quantity [1756], defined at Definition 105.26, is the imaginary part of a quartet of matrix elements and is invariant; its measured value is Equation (105.78),

$$J = \left(3.08^{+0.15}_{-0.13}\right) \times 10^{-5}\,. \tag{111.13}$$

Geometrically $J/2$ is the area of every unitarity triangle (Proposition 103.30), so the single statement "CP is violated" is the single statement "the triangle is not degenerate".

Numerically, J is small not because the phase is small — it is of order unity, $\delta = 65.70^\circ$ by Equation (103.25) — but because the mixing angles are. That distinction returns with consequences in Section 111.8.2.

111.6 Direct CP violation in the kaon system

111.6.1 NA31 and the first evidence

The observable that separates the two hypotheses of Section 111.5.1 is the double ratio of the four two-pion widths, and the reason it and not something simpler is that every quantity common to the four channels cancels in it: the kaon flux, the solid angle, most of the reconstruction efficiency. What does not cancel is the difference between reconstructing two charged tracks and reconstructing four photons, and that difference is the systematic error of every measurement in this section.

The first experiment to see a non-zero value was NA31 at the CERN Super Proton Synchrotron [1806]. Protons of 450 GeV struck a beryllium target; the resulting neutral beam was allowed to run down a long evacuated decay region for the K_L measurement, and for the K_S measurement a second target mounted on a movable train was brought close to the detector, so that the same apparatus saw K_S decays distributed over the same fiducial region. Alternating the two configurations, rather than running them at once, is what makes the acceptance cancel.

NA31 carried no magnetic spectrometer. Its detector was a liquid-argon and lead sampling calorimeter for the photons of $\pi^0\pi^0$, an iron–scintillator hadron calorimeter, and wire chambers for the charged mode; the charged pion momenta were never measured individually, and the charged decays were identified from the calorimetric energy and the transverse position. The result,

$$\mathrm{Re}\frac{\epsilon'}{\epsilon} = (3.3 \pm 1.1) \times 10^{-3}\,, \tag{111.14}$$

is three standard deviations from zero and was the first evidence for direct CP violation.

It was not accepted, and the reason is that the contemporaneous Fermilab experiment E731, using a different technique — two parallel K_L beams with a regenerator placed alternately in one of them, so that K_S and K_L decays were recorded simultaneously — obtained a value consistent with zero, of order 7×10^{-4} with an uncertainty of similar size. The two results were about two standard deviations apart, which is not a crisis in itself; what made the decade that followed uncomfortable is that they disagreed about a *qualitative* question, the existence or not of ϵ', and therefore about whether the superweak hypothesis had been excluded. Neither collaboration could settle it with the data in hand, and both built successors designed around the other's technique. That is the correct response to an irreconcilable pair of measurements, and it is worth recording as such.

Remark 111.10 (A source this book cannot cite). The E731 result quoted above is described but not cited, because the bibliography of this treatise contains no entry for it. The number is given to an order of magnitude and its role in the argument is historical rather than evidential — the argument turns on the two 1999 measurements, which are cited — but the reader should know that this one sentence rests on an uncited source and treat it accordingly.

111.6.2 NA48 and KTeV

Both successors reported in 1999, and both found a non-zero value.

NA48, at CERN, produced K_L and K_S *simultaneously* rather than alternately [1973]. The K_L beam came from a target 126 m upstream of the decay region. A small fraction of the protons that missed that target were deflected by a bent silicon crystal onto a second

target 6 m upstream, producing a K_S beam nearly collinear with the first. The two decay populations were then recorded in the same detector at the same time, in the same beam conditions, with the acceptance difference reduced to the small difference in the decay distributions. Distinguishing a K_S decay from a K_L decay was done by timing: an array of scintillator counters — the tagger — in the proton beam heading for the K_S target recorded the passage of each proton to about 200 ps, and a decay coincident with a tagged proton is a K_S decay.

The detector was a magnetic spectrometer of four drift chambers around a dipole magnet, for $\pi^+\pi^-$, together with a liquid-krypton electromagnetic calorimeter of about $10\,\mathrm{m}^3$ and some thirteen thousand cells, twenty-seven radiation lengths deep, for the four photons of $\pi^0\pi^0$. The krypton calorimeter is the critical instrument: the neutral mode has no tracks, the kaon's decay point must be reconstructed from the photon energies and positions alone by imposing the π^0 mass, and the energy resolution therefore propagates directly into the fiducial volume definition. The 1999 result was

$$\mathrm{Re}\frac{\epsilon'}{\epsilon} = (18.5 \pm 4.5 \pm 5.8) \times 10^{-4}\,, \tag{111.15}$$

the second uncertainty systematic and larger than the first.

KTeV, at Fermilab, was the successor of E731 and kept its technique [1965]. Protons of 800 GeV on a beryllium-oxide target produced two parallel neutral beams; a regenerator — a block of scintillator, itself instrumented so that inelastic interactions in it could be vetoed — sat in one of the two and was moved from one beam to the other about once a minute, so that the regenerated K_S sample and the K_L sample were recorded simultaneously and each beam served in turn as each. The detector was a drift-chamber spectrometer around an analysing magnet and an electromagnetic calorimeter of 3100 pure caesium-iodide crystals, whose energy resolution below one per cent for the photons of interest is what made the neutral mode competitive. The result was

$$\mathrm{Re}\frac{\epsilon'}{\epsilon} = (28.0 \pm 3.0 \pm 2.8) \times 10^{-4}\,. \tag{111.16}$$

The two 1999 numbers differ by 9.5×10^{-4} against a combined uncertainty of 8.4×10^{-4}, that is by 1.1 standard deviations, which is unremarkable. Their distances from zero are not alike: combining each experiment's two errors in quadrature, KTeV is 6.8 standard deviations from zero and NA48 is 2.5. It is KTeV that established the effect in 1999; NA48 confirmed it, at a significance that on its own would not have settled the question, and the two together left no room for zero. With the later data of both experiments the world average is Equation (105.74), $\mathrm{Re}(\epsilon'/\epsilon) = (1.66 \pm 0.23) \times 10^{-3}$.

Phenomenon 111.11 (Direct CP violation in kaon decay)**.** *CP* violation occurs not only in the mixing that defines the long-lived state but in the decay amplitudes themselves. The double ratio of the four two-pion widths,

$$\frac{\Gamma(K_L \to \pi^0\pi^0)\,\Gamma(K_S \to \pi^+\pi^-)}{\Gamma(K_L \to \pi^+\pi^-)\,\Gamma(K_S \to \pi^0\pi^0)} = 1 - 6\,\mathrm{Re}\,(\epsilon'/\epsilon)\ , \tag{111.17}$$

differs from unity. Two experiments with independent systematics found $\mathrm{Re}(\epsilon'/\epsilon)$ of order 10^{-3} and non-zero [1965, 1973], and the world average is $(1.66 \pm 0.23) \times 10^{-3}$ [2427]. A superweak $\Delta S = 2$ interaction, which reproduces ϵ exactly, predicts $\epsilon' = 0$ and is thereby excluded.

Rests on Equations (105.71) *and* (105.72) *and Proposition 105.39.*

Derivation. *Derives Phenomenon 111.11.*

Two steps: the double ratio in terms of ϵ'/ϵ, and the reason ϵ' requires two amplitudes with different strong phases.

The double ratio. Two pions in a relative S wave from a spinless parent have a symmetric isospin wavefunction, so only $I = 0$ and $I = 2$ are available; write the two amplitudes as $A_I \mathrm{e}^{\mathrm{i}\delta_I}$ with δ_I the real strong rescattering phase, as in Equation (105.67). Projecting the physical final states onto the isospin basis with the Clebsch–Gordan coefficients Equation (105.70) gives Equation (105.71),

$$\eta_{+-} = \epsilon + \epsilon' , \qquad \eta_{00} = \epsilon - 2\epsilon' ,$$

the factor -2 being nothing but the ratio of the weight of A_2 in $|\pi^0\pi^0\rangle$ to its weight in $|\pi^+\pi^-\rangle$. The four widths of Equation (111.17) are $|\eta_{00}|^2$ over $|\eta_{+-}|^2$, since each $\Gamma(K_L \to f)$ divided by $\Gamma(K_S \to f)$ is the squared modulus of the corresponding amplitude ratio and everything else cancels. Hence

$$\left|\frac{\eta_{00}}{\eta_{+-}}\right|^2 = \left|\frac{1 - 2\epsilon'/\epsilon}{1 + \epsilon'/\epsilon}\right|^2 = \left|1 - 3\frac{\epsilon'}{\epsilon} + O\Big((\epsilon'/\epsilon)^2\Big)\right|^2 = 1 - 6\,\mathrm{Re}\frac{\epsilon'}{\epsilon} + O\Big(10^{-6}\Big) ,$$

which is Equation (111.17). With the measured average the double ratio is $1 - 0.00996$, so the experiments are looking for a one per cent departure from unity in a ratio of four separately measured rates — and the factor of six in front, which triples the naive $-2\epsilon'/\epsilon$, is the only thing that makes it feasible at all.

Why a non-zero value needs two phases. From Equation (105.72),

$$\epsilon' = \frac{\mathrm{i}}{\sqrt{2}}\,\mathrm{e}^{\mathrm{i}(\delta_2 - \delta_0)}\,\frac{\mathrm{Re}\,A_2}{\mathrm{Re}\,A_0}\left[\frac{\mathrm{Im}\,A_2}{\mathrm{Re}\,A_2} - \frac{\mathrm{Im}\,A_0}{\mathrm{Re}\,A_0}\right] ,$$

and the bracket is a *difference*. A phase common to both isospin amplitudes is a rephasing of the kaon state and carries no physics; it cancels in the bracket, leaving $\epsilon' = 0$. So direct CP violation requires at least two amplitudes with different weak phases. It requires more than that: the prefactor carries $\mathrm{e}^{\mathrm{i}(\delta_2 - \delta_0)}$, and if the two *strong* phases were equal that factor would be real, ϵ' would be purely imaginary relative to ϵ, and $\mathrm{Re}(\epsilon'/\epsilon)$ — the observable of Equation (111.17) — would vanish even with $\epsilon' \neq 0$. Both a weak-phase difference and a strong-phase difference are necessary. This is the general statement proved at Proposition 105.39, and it is used again without change for the charm asymmetry in Section 111.7.3. □

Remark 111.12 (Why the measured value is so small). That ϵ'/ϵ is of order 10^{-3} rather than of order unity is not a statement about the CP-violating phase. The prefactor of ϵ' contains $\mathrm{Re}\,A_2/\mathrm{Re}\,A_0$, and the $\Delta I = \frac{1}{2}$ rule — the empirical observation that $|A_0/A_2| \approx 22$ in kaon decay, itself an unexplained strong-interaction fact — suppresses that ratio by a factor of twenty-two while simultaneously enhancing ϵ. Roughly five hundred of the three orders of magnitude come from that one number.

111.6.3 The Standard Model prediction and what it is worth

The measurement excludes the superweak hypothesis by itself, and that conclusion needs no theory. Comparing the measured value with the Cabibbo–Kobayashi–Maskawa prediction is a different and much weaker statement, and the reason should be stated plainly rather than left to a citation.

The short-distance part of the calculation is under control: the $\Delta S = 1$ effective Hamiltonian is a sum of local four-quark operators with coefficients computed perturbatively, and the phase enters only through the CKM combination $\mathrm{Im}\,(V_{td}V_{ts}^*)$. The long-distance part is not: what is needed is the matrix element of each operator between a kaon and a two-pion state at the physical kinematics, including the final-state rescattering that produces δ_0 and δ_2, and there is no expansion parameter in which that is small. Lattice calculations of these matrix elements now exist and give a value compatible with Equation (105.74) with

an uncertainty several times larger, dominated by the $I = 0$ channel where the rescattering is strongest; those calculations are not cited here because this treatise's bibliography carries no entry for them, and the reader should not take the agreement on the authority of this sentence.

What is established without any of that machinery is the qualitative result: $\epsilon' \neq 0$, hence CP is violated in a $\Delta S = 1$ decay amplitude, hence no interaction acting only in the mass matrix can account for the phenomenon. That is the content of Phenomenon 111.11 and it is secure.

Derivation pending #194

The short-distance box-diagram computation of the $[\Delta S = 2]$ *mixing amplitude, giving the mass difference and epsilon in terms of the CKM elements and the quark masses — the step that converts the measured epsilon into a constraint on the unitarity triangle. It is an appendix-scale calculation requiring the loop functions of the box with internal charm and top, and it belongs in Appendix A rather than in an experiment chapter.*

111.7 CP violation beyond the kaons

111.7.1 The B factories

If one phase in one matrix accounts for all of it, then the kaon's part in a thousand and an effect of order unity somewhere else are the same physics. The Standard Model says where to look: in $B^0 \to J/\psi\, K^0_S$, where the interference is between decay after mixing and decay without mixing, and no small ratio of amplitudes suppresses it.

The machines. Two asymmetric-energy electron–positron colliders were built for this measurement, PEP-II at SLAC and KEKB at KEK. Both run at a centre-of-mass energy of 10.58 GeV, the mass of the $\Upsilon(4S)$, which decays to a $B^0\bar{B}^0$ pair with almost no energy to spare — with $m_{B^0}c^2 = 5279.66\,\text{MeV}$ each B has a momentum $p = \sqrt{(m_\Upsilon c^2/2)^2 - (m_B c^2)^2}/c = 331\,\text{MeV}/c$ in the Υ frame. Asymmetric energies are the whole point. PEP-II collided 9.0 GeV electrons with 3.1 GeV positrons, KEKB 8.0 GeV with 3.5 GeV, so that the centre of mass moves in the laboratory with

$$\beta\gamma = \frac{E_- - E_+}{\sqrt{4E_-E_+}} = 0.56\ \text{(PEP-II)}\,, \qquad 0.43\ \text{(KEKB)}\,. \tag{111.18}$$

The one line behind this: for head-on beams whose energies far exceed the electron mass, the total four-momentum has energy $E_- + E_+$ and momentum $(E_- - E_+)/c$ along the electron direction, so $\beta = (E_- - E_+)/(E_- + E_+)$, while $(E_- + E_+)^2 - (E_- - E_+)^2 = 4E_-E_+$ is the square of the invariant centre-of-mass energy, so $\gamma = (E_- + E_+)/\sqrt{4E_-E_+}$. The product is Equation (111.18), and $\sqrt{4E_-E_+}$ is 10.56 GeV for PEP-II and 10.58 GeV for KEKB, as it must be.

The detectors. BaBar and Belle are conventional collider detectors of their generation, built around the interaction point inside a 1.5 T solenoid: a silicon strip vertex detector of five layers immediately around the beam pipe, a gas drift chamber for momentum measurement, a Cherenkov device for separating kaons from pions — internally reflecting quartz bars in BaBar, aerogel threshold counters and time-of-flight in Belle — a caesium-iodide crystal calorimeter, and an instrumented flux return for muons. What the measurement demands of them is unusual, however: not energy resolution but *position* resolution along the beam axis, to better than 100 µm per decay vertex.

Phenomenon 111.13 (A large CP asymmetry in B meson decay)**.** In $\Upsilon(4S)$ decays to $B^0\bar{B}^0$ pairs, the rate for $B^0 \to J/\psi\, K^0_S$ differs from that for $\bar{B}^0 \to J/\psi\, K^0_S$ by an amount

that oscillates with the proper-time difference between the two B decays, with an amplitude of order unity; two experiments reported it simultaneously [2005, 2010]. The parameter extracted is $\sin 2\beta = 0.708 \pm 0.011$ [2427]. CP violation is therefore not intrinsically a small effect: in the kaon system it is a part in a thousand, in this B channel it is maximal in order of magnitude.

Rests on Equations (105.82) *and* (105.87) *and Proposition 105.28.*

Derivation. *Derives Phenomenon 111.13.*

Three things have to be shown: that the observable is a *difference* of decay times and not a decay time, that its amplitude is $\sin 2\beta$, and that the boost of Equation (111.18) makes the difference measurable.

The observable. The $\Upsilon(4S)$ has $J^{PC} = 1^{--}$ and decays to two pseudoscalars in a P wave, so the pair is in the antisymmetric state Equation (105.87). Antisymmetry is preserved until one meson decays; therefore at every instant before the first decay there is exactly one B^0 and exactly one $\bar{B}^0$, and no oscillation is observable in the pair as a whole. The clock starts when one of them decays to a flavour-specific final state — a semileptonic decay, whose lepton charge tags the flavour, or a $D^{(*)}\pi$ state — because that decay projects the partner onto the opposite flavour at that instant. The partner then oscillates freely until it decays to $J/\psi K^0_S$, and the phase it has accumulated is $\Delta m_d c^2 \Delta t/\hbar$ with $\Delta t = t_{CP} - t_{\text{tag}}$, which may be of either sign. The asymmetry is therefore Equation (105.81),

$$\mathcal{A}(\Delta t) = S_f \sin \frac{\Delta m_d c^2 \Delta t}{\hbar} - C_f \cos \frac{\Delta m_d c^2 \Delta t}{\hbar} \,,$$

odd in Δt for the CP-violating term, which is what makes the sign of Δt — and hence the vertex ordering — essential rather than incidental.

The amplitude. With $\lambda = (q/p)\left(\bar{\mathcal{A}}_f/\mathcal{A}_f\right)$, the coefficients are $S_f = 2\mathrm{Im}\,\lambda/(1+|\lambda|^2)$ and $C_f = (1-|\lambda|^2)/(1+|\lambda|^2)$ (Proposition 105.28). For $f = J/\psi K^0_S$ the decay $b \to c\bar{c}s$ proceeds through one dominant amplitude, so $|\lambda| = 1$ and $C_f = 0$, and λ is a pure phase built from three CKM ratios: the B^0 mixing, dominated by the top quark in the box, contributes $q/p = V^*_{tb}V_{td}/\left(V_{tb}V^*_{td}\right)$; the decay amplitude ratio contributes $V_{cb}V^*_{cs}/\left(V^*_{cb}V_{cs}\right)$; and the neutral kaon in the final state must itself be projected onto K^0_S, which contributes $V_{cd}V^*_{cs}/\left(V^*_{cd}V_{cs}\right)$. The product, with the CP eigenvalue $\eta_f = -1$ of $J/\psi K^0_S$, is

$$\lambda = -\frac{V^*_{tb}V_{td}}{V_{tb}V^*_{td}}\,\frac{V_{cb}V^*_{cd}}{V^*_{cb}V_{cd}} = -\mathrm{e}^{-2\mathrm{i}\beta} \,, \qquad \beta = \arg\left(-\frac{V_{cd}V^*_{cb}}{V_{td}V^*_{tb}}\right) \,,$$

the factors of V_{cs} cancelling between the decay and the kaon. Since $|\lambda| = 1$, $S_f = \mathrm{Im}\,\lambda = \mathrm{Im}\left(-\mathrm{e}^{-2\mathrm{i}\beta}\right) = \sin 2\beta$, which is Equation (105.82). Note what has and has not been assumed: unitarity of the mixing matrix, top dominance of the box, and a single decay amplitude. Nothing about hadronic matrix elements enters, which is why this channel measures an angle of the unitarity triangle to a per-cent theoretical accuracy while the kaon's ϵ' does not.

The measurement of Δt. A B meson lives $\tau_B = 1.519 \times 10^{-12}\,\mathrm{s}$ and moves with the centre of mass, so the two decay vertices are separated along the beam axis by

$$\Delta z = \beta\gamma c\,\Delta t \,, \qquad \langle|\Delta z|\rangle \sim \beta\gamma c\tau_B = 0.56 \times 2.998 \times 10^8\,\mathrm{m/s} \times 1.519 \times 10^{-12}\,\mathrm{s} = 2.5 \times 10^{-4}\,\mathrm{m}$$

at PEP-II, and $1.9 \times 10^{-4}\,\mathrm{m}$ at KEKB. Without the boost the two mesons would be nearly at rest and Δz would be a few micrometres, far below any silicon detector's resolution; with it, Δz is a quarter of a millimetre and a vertex resolution of about $180\,\mu\mathrm{m}$ gives $\sigma(\Delta t) = \sigma(\Delta z)/(\beta\gamma c) \approx 1.1 \times 10^{-12}\,\mathrm{s}$, comparable to τ_B itself and to the oscillation period $2\pi\hbar/(\Delta m_d c^2) = 1.24 \times 10^{-11}\,\mathrm{s}$. The measurement is therefore possible but resolution-limited, which is why the fitted asymmetry must be convolved with a resolution function determined from the data themselves.

Dilution. The flavour tag is not perfect: a fraction w of tags is wrong, and each wrong tag contributes with the opposite sign, so the observed amplitude is $(1-2w)S_f$ and the statistical power of a sample of N events with tagging efficiency ε scales as $\varepsilon(1-2w)^2 N$. Both experiments used several tagging categories with different purities and fitted them simultaneously, reaching an effective efficiency $\varepsilon(1-2w)^2$ of about 0.3. This is the dominant reason the first measurements, on samples of about thirty million $B\bar{B}$ pairs each, carried statistical uncertainties of 0.14. □

The two 2001 results were $\sin 2\beta = 0.59 \pm 0.14 \pm 0.05$ from BaBar, on a sample of about thirty-two million $B\bar{B}$ pairs [2010], and $0.99 \pm 0.14 \pm 0.06$ from Belle, on about thirty-one million [2005]. They differ by roughly two standard deviations, which is a useful reminder that first measurements scatter and that the agreement of two simultaneous results is not to be expected as a matter of course. Both are far from zero, which is the discovery; the average settled at 0.708 ± 0.011 as the samples grew by a factor of ten [2427].

Remark 111.14 (The prediction this confirms). The value was predicted before it was measured, from quantities having nothing to do with CP asymmetries. Taking the apex of the unitarity triangle from $|V_{ub}/V_{cb}|$, from the B^0 and B_s^0 oscillation frequencies and from ϵ itself — Equation (103.24) — gives $\beta = 22.7°$ and hence $\sin 2\beta = 0.713$, against the measured 0.708 ± 0.011. Three loop processes and one tree process, none of them an asymmetry, anticipate an asymmetry of order unity to better than one per cent. That is the strongest single test the Cabibbo–Kobayashi–Maskawa picture has passed.

111.7.2 Neutral B_s mesons at the LHC

The B_s^0 system, in which the light quark is strange rather than down, oscillates thirty-five times faster than the B^0 (Table 111.6) and was for that reason the last of the neutral-meson systems to be brought under control.

The apparatus. LHCb is not a collider detector in the usual sense: it is a single-arm forward spectrometer, covering pseudorapidities from about 2 to 5, built on the observation that at the Large Hadron Collider the b and $\bar{b}$ of a produced pair both go forward at small angles. Its elements, in order from the interaction point, are a silicon strip vertex locator whose sensors approach to within about 8 mm of the beams, a tracking system around a warm dipole magnet of about 4 T m bending power, two ring-imaging Cherenkov detectors with gas radiators, which give kaon–pion separation over the momentum range from a few GeV/c to about 100 GeV/c, calorimeters, and muon chambers. The two numbers that matter for flavour physics are an impact-parameter resolution of about 20 µm and a decay-time resolution of about 45 fs — the latter being what makes a B_s^0 oscillation of period $2\pi\hbar/(\Delta m_s c^2) = 3.5 \times 10^{-13}$ s resolvable at all.

That a hadron collider can do this at all rests on two things. The production cross section for $b\bar{b}$ pairs is large, so statistics are never the limitation; and the b hadron flies a measurable distance — of order a centimetre at LHC momenta — before decaying, so a vertex detached from the primary interaction is by itself a powerful trigger against the enormous light-quark background.

The result. The first observation of CP violation in the decays of B_s^0 mesons was of the *direct* kind, in the decay $B_s^0 \to K^-\pi^+$ [2202]. The measured asymmetry,

$$A_{CP}\left(B_s^0 \to K^-\pi^+\right) = 0.27 \pm 0.04 \pm 0.01\,, \tag{111.19}$$

is non-zero at 6.5 standard deviations, and the same analysis measured $A_{CP}(B^0 \to K^+\pi^-) = -0.080 \pm 0.007 \pm 0.003$. Two features are worth noting. The asymmetry is of order a quarter — CP violation in a decay amplitude can be large when the interfering amplitudes are comparable, as tree and penguin are here. And the two asymmetries have opposite signs,

which is a prediction of the Cabibbo–Kobayashi–Maskawa structure and not an accident: the ratio of the two, weighted by the branching fractions and lifetimes, is predicted to be -1 by U-spin symmetry, the interchange of the d and s quarks. The reason in outline is that U-spin maps $B^0 \to K^+\pi^-$ onto $B_s^0 \to K^-\pi^+$ while leaving the strong phases and the moduli of the interfering tree and penguin amplitudes invariant and flipping the sign of the weak phase difference, so the two CP asymmetries have equal magnitude and opposite sign once each is normalized to its own width.

Derivation pending #195

The U-spin relation between the direct CP asymmetries of $[B^0 \to K^+\pi^-]$ and $[B_s^0 \to K^-\pi^+]$ is stated here in outline only. Owed: the explicit tree-plus-penguin decomposition, the action of the $d \leftrightarrow s$ interchange on each amplitude, and the width weighting that turns the amplitude statement into the ratio -1, together with the size of the U-spin-breaking corrections. This belongs in Appendix A. No key in the bibliography covers the U-spin literature, so the claim carries no citation.

The null test. The observable usually associated with the B_s^0 is not Equation (111.19) but the mixing phase ϕ_s, the analogue of 2β for the strange-beauty system, measured from the time-dependent asymmetry in $B_s^0 \to J/\psi\,\phi$ and related channels. Its Standard Model value is small and sharply predicted, because in the Wolfenstein expansion the relevant combination $V_{ts}V_{tb}^*$ is almost real; the world average, $\phi_s \approx -0.05$ rad with an uncertainty of about 0.02 rad [2427], agrees with it. A small, precisely predicted quantity is the ideal place to look for a new short-distance contribution, since anything new entering the box would shift it; nothing has been seen, and the measurement is therefore a constraint rather than a discovery.

111.7.3 Charm

The up-type quarks were the last to yield, twenty years after the B mesons and fifty-five after the kaon.

Phenomenon 111.15 (CP violation in charm decay)**.** The difference between the time-integrated CP asymmetries of $D^0 \to K^-K^+$ and $D^0 \to \pi^-\pi^+$ is not zero,

$$\Delta A_{CP} = A_{CP}\left(D^0 \to K^-K^+\right) - A_{CP}\left(D^0 \to \pi^-\pi^+\right) = (-15.4 \pm 2.9) \times 10^{-4}\,, \quad (111.20)$$

established at 5.3 standard deviations [2323]. With this the effect is observed in the strange, bottom and charm sectors alike; it has never been observed in any process that does not involve the weak interaction.

Rests on Proposition 105.39.

Derivation. *Derives Phenomenon 111.15.*

Why a difference is measured and not an asymmetry. A raw counting asymmetry between D^0 and $\bar{D}^0$ decays contains three contributions: the physics asymmetry, a production asymmetry, because a proton–proton collider does not produce charm and anticharm symmetrically, and a detection asymmetry, because the tagging pion of the $D^{*+} \to D^0\pi^+$ chain is reconstructed with a different efficiency according to its charge. The last two are common to the two final states K^-K^+ and $\pi^-\pi^+$, both of which are self-conjugate, so they cancel in the difference Equation (111.20) to first order. This is the same design principle as the double ratio of Equation (111.17): build the observable so that everything not being measured appears in both terms.

What generates the asymmetry. A direct asymmetry requires two interfering amplitudes with both a weak and a strong phase difference, by Proposition 105.39. Here they are the tree

amplitude $c \to s\bar{s}u$ (or $d\bar{d}u$) and the loop, or "penguin", amplitude in which the c converts to u via a W and a down-type quark. Write the total as $\mathcal{A} = A_1 \mathrm{e}^{\mathrm{i}(\varphi_1+\delta_1)} + A_2 \mathrm{e}^{\mathrm{i}(\varphi_2+\delta_2)}$, with φ_i the weak phases, which change sign under CP, and δ_i the strong phases, which do not; the conjugate amplitude $\bar{\mathcal{A}}$ is the same with $\varphi_i \to -\varphi_i$. Then, with $r = A_2/A_1$, $\Delta\varphi = \varphi_1 - \varphi_2$ and $\Delta\delta = \delta_1 - \delta_2$,

$$\left|\bar{\mathcal{A}}\right|^2 - |\mathcal{A}|^2 = 2A_1A_2\left[\cos\left(\Delta\delta - \Delta\varphi\right) - \cos\left(\Delta\delta + \Delta\varphi\right)\right] = 4A_1A_2 \sin\Delta\delta \, \sin\Delta\varphi \,,$$

$$A_{CP} = \frac{\left|\bar{\mathcal{A}}\right|^2 - |\mathcal{A}|^2}{\left|\bar{\mathcal{A}}\right|^2 + |\mathcal{A}|^2} = \frac{2r \sin\Delta\delta \, \sin\Delta\varphi}{1 + 2r\cos\Delta\delta \, \cos\Delta\varphi + r^2} \,.$$

Both sines must be non-zero. If the two amplitudes shared a weak phase it could be rotated away; if they shared a strong phase the two conjugate amplitudes would have equal moduli. For $r \ll 1$ this reduces to $A_{CP} \approx 2r \sin\Delta\delta \sin\Delta\varphi$, and with $r \sim 10^{-3}$ from the CKM factor $|V_{cb}V_{ub}^*/(V_{cs}V_{us}^*)| \approx 6 \times 10^{-4}$ and strong phases of order unity, an asymmetry of order 10^{-3} is what the Standard Model leads one to expect — which is what was found.

Why this does not test the theory sharply. The expression above contains r and $\Delta\delta$, and neither is calculable. The charm quark mass sits in the worst possible place: too heavy for chiral perturbation theory, too light for the heavy-quark expansion that makes B decays tractable, and the final states $K\bar{K}$ and $\pi\pi$ lie in a region dense with hadronic resonances that supply large and unpredictable strong phases. Estimates of the Standard Model value of Equation (111.20) span an order of magnitude and do not exclude the measured number at either end. The measurement therefore establishes the *phenomenon* in the up-type sector, which is a genuine addition to knowledge, and constrains rather than tests the Cabibbo–Kobayashi–Maskawa picture. □

The apparatus and the sample. The measurement is LHCb's, using proton–proton collisions at centre-of-mass energies of 7, 8 and 13 TeV and an integrated luminosity of about $9 \times 10^{43}/\mathrm{m}^2$. Two independent tagging methods were combined: the charge of the soft pion in $D^{*+} \to D^0\pi^+$, and the charge of the muon in a semileptonic b-hadron decay producing the D^0. The samples run to tens of millions of decays in each final state, and at that size the measurement is systematics-limited in practice — which is why the whole design is arranged so that the systematics cancel in a difference.

111.7.4 Direct T violation

Given the CPT theorem, every result above is also a statement about time reversal. But that inference *assumes* the theorem, whose hypotheses — locality, Lorentz invariance, the spin-statistics connection — are the deepest in the subject; a test of T that does not assume it is independent information. Two experiments have done it, and the difficulty in both is the same: a genuine T test compares a transition with its *reverse*, exchanging initial and final states, and conjugates nothing.

CPLEAR. The neutral kaon offers a transition and its exact reverse without any need to reverse momenta or spins, because both states are the same particle at rest: $K^0 \to \bar{K}^0$ and $\bar{K}^0 \to K^0$ are related by exchanging initial and final state, and nothing else [1426]. The Kabir asymmetry Equation (105.84) must vanish identically if T is a symmetry.

The apparatus was built around the tagging problem, which is the whole difficulty: one must know the strangeness at production *and* at decay, in the same event. CPLEAR used the Low Energy Antiproton Ring at CERN, stopping antiprotons of about 200 MeV/c in a gaseous hydrogen target at high pressure at a rate of order 10^6 per second. About two annihilations in a thousand give $\bar{p}p \to K^{\mp}\pi^{\pm}K^0/\bar{K}^0$, and strangeness conservation in the strong production ties the sign of the accompanying charged kaon to the strangeness of the

neutral one: the initial state is tagged by a charge measured in the same event. The final strangeness is tagged by the charge of the lepton in the semileptonic decay, using $\Delta S = \Delta Q$ exactly as in Equation (111.11). The detector was a cylindrical assembly inside a solenoid of about 0.44 T — tracking chambers, a threshold Cherenkov and scintillator sandwich for identifying the charged kaon, and a lead and gas-sampling electromagnetic calorimeter. The measured asymmetry Equation (105.86),

$$A_T = (6.6 \pm 1.3 \pm 1.0) \times 10^{-3}\,, \tag{111.21}$$

is the first direct observation of a time-reversal asymmetry [1948], and it agrees with the prediction $A_T = 4\,\mathrm{Re}\,\epsilon = 6.46 \times 10^{-3}$ of Proposition 105.29 evaluated with $|\epsilon| = 2.228 \times 10^{-3}$ and $\arg\epsilon = 43.5°$.

BaBar. The cleaner measurement uses entanglement as a state-preparation device [2197]. From the antisymmetric $B^0\bar{B}^0$ state Equation (105.87), observing one meson decay to a flavour-specific state projects the other onto a definite flavour at that instant, and observing a decay to $J/\psi K^0_S$ projects it onto a definite CP label. The comparison is then available in both time orders: events in which the flavour tag comes first measure $B^0 \to B_-$, events in the opposite order measure $B_- \to B^0$, and these are a process and its reverse with nothing conjugated. This is the entanglement machinery of Chapter 84 used not to test locality but as an instrument. On a sample of 468 million $B\bar{B}$ pairs BaBar found the T-odd parameters non-zero at fourteen standard deviations,

$$\Delta S_T^+ = -1.37 \pm 0.14 \pm 0.06\,, \qquad \Delta S_T^- = 1.17 \pm 0.18 \pm 0.11\,, \tag{111.22}$$

while the CPT-odd parameters extracted from the same fit were consistent with zero [2197].

The CPT tests obtained as a by-product are worth recording, because they are among the sharpest tests of anything. The same formalism that yields Equation (111.21) bounds the CPT-violating parameter of the kaon mass matrix, and the resulting limit on the fractional mass difference between K^0 and $\bar{K}^0$ is of order 10^{-18} — the most precise equality of a particle and antiparticle property in physics. The neutral-meson interferometer is what makes it possible, for the reason set out at Equation (105.63): the observable is an accumulated phase, not an energy. The details belong to Section 105.5.3.

Remark 111.16 (What these measurements do and do not assume). Neither experiment is entirely free of assumptions, and both have been criticized on that ground. CPLEAR's final-state tag uses the semileptonic amplitudes, so a CPT violation in those amplitudes, or a failure of $\Delta S = \Delta Q$, could in principle produce the observed asymmetry with no T violation in the mixing; the same data bound those parameters well below the measured A_T, but the argument is internal. BaBar's mesons have a finite width, so the reversed transition occupies a different interval of proper time and is not an exact motion reverse; and the projection interpretation of the tag requires the tagging decay to be strictly flavour-specific. What is established, and what no earlier measurement established, is that the T-odd and CP-odd parameters are separately determined from one data set and are separately non-zero.

111.7.5 The unitarity triangle

The individual measurements above are of very different quantities in very different apparatus. What ties them together is one closure condition, the orthogonality of the first and third columns of the mixing matrix,

$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0\,, \tag{111.23}$$

which is Equation (105.80). Three complex numbers summing to zero close a triangle in the complex plane; dividing by the middle term puts its base on the real axis from 0 to 1 and its apex at $(\bar{\rho}, \bar{\eta})$. Every measurement in this chapter is a constraint on that apex, and they come in two kinds:

- *Sides*, from decay rates that involve no CP violation at all: $|V_{ub}/V_{cb}|$ from semileptonic B decay fixes the distance from the origin, and the ratio of the B_s^0 to the B^0 oscillation frequency fixes the distance from $(1, 0)$.
- *Angles*, from the CP asymmetries themselves: $\sin 2\beta$ from $B^0 \to J/\psi K_S^0$, γ from the interference of $b \to c$ and $b \to u$ amplitudes in $B \to DK$, α from $B \to \pi\pi$ and $\rho\rho$. The kaon's ϵ is a constraint of its own, a hyperbola in the $(\bar{\rho}, \bar{\eta})$ plane.

The measured angles are Equation (103.32), $\alpha = 85°$, $\beta = 22.2°$ and $\gamma = 66°$; they sum to $173° \pm 6°$ against the identity $\alpha + \beta + \gamma = \pi$, which is a consistency test the three measurements had no way of enforcing on one another. The global fits [2074] combine all of it, and the constraints cross in one region of the plane a few per cent across. The full account, with the fitted apex and every input, is Section 103.4.3 and Phenomenon 103.31; what belongs here is the experimental summary and the honest qualifications.

Those qualifications are three, and they are reported as they stand rather than smoothed over. The angle α carries a $5°$ uncertainty dominated not by statistics but by the hadronic penguin amplitudes that must be subtracted from $B \to \pi\pi$ — that is, by theory. The side $|V_{ub}/V_{cb}|$ inherits an unresolved tension: the inclusive determinations, which sum over all final states and use an operator product expansion, sit two to three standard deviations above the exclusive ones, which measure a single channel and need a lattice form factor, for both $|V_{ub}|$ and $|V_{cb}|$; choosing one or the other moves the apex by about one standard deviation. And the most precisely measured unitarity statement of all is not a triangle at all but the first-row relation $|V_{ud}|^2 + |V_{us}|^2 + |V_{ub}|^2 = 1$, which at present does not come out right at the 10^{-3} level (Remark 103.32). The correct summary is that one phase accounts for all observed CP violation in the quark sector at the few-per-cent level, and that two specific tensions remain.

111.8 Cosmological relevance and open questions

111.8.1 The Sakharov conditions

The measurements of this chapter acquire a significance beyond particle physics because of an observation about the universe and an argument by Sakharov [1358].

Phenomenon 111.17 (The universe is made of matter)**.** No astrophysical concentration of antimatter has been found: neither the annihilation radiation that a matter–antimatter boundary would produce, nor antinuclei in the cosmic radiation beyond the rate expected from secondary production. The baryon-to-photon ratio inferred from the acoustic peaks of the cosmic microwave background is

$$\eta = \frac{n_B - n_{\bar{B}}}{n_\gamma} = (6.12 \pm 0.04) \times 10^{-10}\,, \tag{111.24}$$

in agreement with the value independently required by primordial nucleosynthesis [2348]. A universe that began symmetric under C and CP, in thermal equilibrium and with conserved baryon number, could not have arrived at Equation (111.24).

Rests on Theorems 105.36 and 105.41 and Equation (105.105)*.*

Derivation. *Derives Phenomenon 111.17.*

The three requirements are proved as Theorem 105.41; what is carried out here is the arithmetic that turns them into a number, and the identification of which laboratory measurement supplies each.

Baryon number violation. Without it the initial value of $n_B - n_{\bar{B}}$ is the final one, and a symmetric start stays symmetric. The Standard Model supplies it: the baryon current has

an Adler–Bell–Jackiw anomaly, so its divergence is a total derivative of a topological density [1569], and configurations interpolating between vacua of different winding number change B by three units per generation. At zero temperature the process is suppressed by a barrier factor of order $\mathrm{e}^{-4\pi/\alpha_W}$ and is entirely unobservable; above the electroweak scale the barrier is crossed thermally by the sphaleron of Equation (105.104), and the rate is unsuppressed [1758]. This condition is met.

C and CP violation. If C or CP were a symmetry, every baryon-number-violating process would be accompanied by its mirror at exactly the same rate and no net asymmetry could accumulate. This is the condition that the present chapter has measured, and it is met — that is the whole content of Phenomenon 111.8 and everything after it.

Departure from thermal equilibrium. In equilibrium the number density of a species depends only on its mass, chemical potential and the temperature, and CPT makes the masses of a particle and its antiparticle equal (Theorem 105.36); with B not conserved the corresponding chemical potential vanishes, so the two densities are equal whatever the microphysics. Any asymmetry generated is therefore erased as fast as it is made unless the expansion outruns the reactions.

The number. What must be explained is Equation (111.24), and that value is itself derived rather than observed directly: the acoustic peak structure of the microwave background measures $\Omega_b h^2 = 0.02237 \pm 0.00015$ [2348], and dividing the resulting baryon number density by the blackbody photon density at $T_0 = 2.7255\,\mathrm{K}$ gives $\eta = 2.74 \times 10^{-8}\,\Omega_b h^2$, which is Equation (105.105) and Equation (103.45). The independent determination from the light-element abundances (Chapter 48) agrees, which is what makes Equation (111.24) a measurement rather than a fit. The absence of antimatter concentrations is a separate observational statement: a matter–antimatter boundary anywhere within the observable universe would radiate annihilation gamma rays at a level long since excluded, and no search of the cosmic radiation has produced an antinucleus heavier than the antiproton, whose observed flux is consistent with secondary production on the interstellar medium. □

111.8.2 Insufficiency of Standard Model CP violation

All three Sakharov conditions are formally present in the Standard Model. Two of them fail quantitatively, and the failures are independent of each other.

The phase is far too small, and it is the masses that make it so. CP violation in the quark sector is controlled not by J alone but by the rephasing-invariant combination Equation (105.79), which multiplies J by the six squared-mass differences and therefore carries the SI dimension J^{12}. An asymmetry generated at temperature T is a pure number, so it can depend on that invariant only through the dimensionless ratio Equation (105.106), obtained by dividing by $(k_B T)^{12}$. Evaluated with the running quark masses at $k_B T \approx 100\,\mathrm{GeV}$, where the sphaleron processes are active, this gives Equation (105.107),

$$d_{CP} \approx 2 \times 10^{-21}\,, \tag{111.25}$$

against the measured $\eta = 6.12 \times 10^{-10}$. The ratio is

$$\frac{\eta}{d_{CP}} = \frac{6.12 \times 10^{-10}}{2 \times 10^{-21}} = 3 \times 10^{11}\,, \tag{111.26}$$

a shortfall of eleven orders of magnitude, and the detailed calculations [1882] do not improve it. The reason is structural rather than numerical. CP violation in this theory requires all six quark masses to be distinct (Proposition 105.27); at 100 GeV five of the six are negligible against the temperature, so the effect is suppressed by twelve powers of small mass ratios. The phase itself is of order unity — $\delta = 65.70^\circ$ — and the masses kill it. This is the same arithmetic as Equation (103.46), and it is quoted here in the same form so that the two chapters cannot drift apart.

The third condition fails outright. A departure from equilibrium at the electroweak scale requires the transition from the symmetric to the broken phase to be first order, proceeding by the nucleation of bubbles whose walls sweep the plasma — and strongly first order at that, since the sphalerons must switch off fast enough after the transition not to erase what was just made. Lattice computations locate the endpoint of the first-order line at a scalar mass of about $72\,\mathrm{GeV}/c^2$ [1925]; beyond it the change is a smooth crossover with no bubbles, no walls and no departure from equilibrium. The measured Higgs mass, Equation (104.66), is $(125.20 \pm 0.11)\,\mathrm{GeV}/c^2$ [2427], far above the endpoint (Remark 104.65). No amount of CP violation would help.

Remark 111.18 (Stated as a failure, not as a puzzle solved). Mechanisms that would generate Equation (111.24) exist in outline — heavy Majorana neutrinos decaying out of equilibrium, additional scalars driving a first-order transition — and not one of them has any experimental support. The shortfall itself is recorded in Section 133.5, which states that no confirmed mechanism supplies it, and Remark 103.45 makes the same point from the flavour side. What is *measured* is Equation (111.24); what is *calculated* is Equation (111.25); the eleven orders of magnitude between them are one of the few places where the Standard Model is known, from laboratory numbers plus one cosmological observation, to be incomplete.

111.8.3 Electric dipole moments and the strong CP problem

Every measurement so far in this chapter is of a CP violation that exists. The last one is of a CP violation that does not, and the null result is the more constraining of the two kinds.

Phenomenon 111.19 (The neutron has no electric dipole moment)**.** Repeated searches over seven decades find no permanent electric dipole moment of the neutron. The current bound is

$$|d_n| < 2.9 \times 10^{-47}\,\mathrm{C\,m} \tag{111.27}$$

at 90 % confidence [2346], improving on [2090]. The same limit is customarily written 1.8×10^{-26} in units of the elementary charge times the centimetre, the conversion being $1\,e\,\mathrm{cm} = 1.602\,176\,634 \times 10^{-21}\,\mathrm{C\,m}$. This is a null result about a symmetry, not about a nuclear property.

Rests on Theorem 105.30 and Equation (105.91).

Derivation. *Derives Phenomenon 111.19.*

A neutron in a state of definite spin has no vector at its disposal other than its spin: by the projection theorem carried by the Wigner–Eckart theorem, the expectation value of any vector operator in a nondegenerate state of given angular momentum is proportional to $\langle \boldsymbol{S} \rangle$. Hence $\langle \boldsymbol{d} \rangle = d_n \langle \boldsymbol{S} \rangle / |\langle \boldsymbol{S} \rangle|$ for some real number d_n, and the interaction with a static electric field is

$$H_d = -\langle \boldsymbol{d} \rangle \cdot \boldsymbol{E} \propto -d_n \, \langle \boldsymbol{S} \rangle \cdot \boldsymbol{E}\,.$$

Under space inversion $\boldsymbol{E} \mapsto -\boldsymbol{E}$ while $\boldsymbol{S} \mapsto \boldsymbol{S}$, so $H_d \mapsto -H_d$; under time reversal $\boldsymbol{E} \mapsto \boldsymbol{E}$ while $\boldsymbol{S} \mapsto -\boldsymbol{S}$, so again $H_d \mapsto -H_d$. A non-zero d_n therefore requires the violation of P and of T separately, and — given the CPT theorem of Chapter 105 — of CP. The full statement, with the C properties made explicit, is Theorem 105.30.

The nondegeneracy clause is what makes the statement non-trivial: a state with an exact degeneracy of opposite parity, such as the hydrogen $n = 2$ level, shows a linear Stark effect with no symmetry violation at all, because there the dipole matrix element is off-diagonal and the argument above does not apply (Remark 105.31).

From the bound to a frequency. The experiment measures a frequency, not a dipole. Ultracold neutrons are stored in a cell in parallel magnetic and electric fields and their Larmor precession is measured by Ramsey's method of separated oscillatory fields [968]; reversing $\boldsymbol{E}$ and differencing isolates $\Delta\nu = 4 d_n E / h$ (Equation (105.91)). At the Paul

Scherrer Institute storage field $E \approx 1.1 \times 10^{6}\,\mathrm{V/m}$ a dipole equal to Equation (111.27) would shift the resonance by

$$\Delta\nu = \frac{4 \times 2.9 \times 10^{-47}\,\mathrm{C\,m} \times 1.1 \times 10^{6}\,\mathrm{V/m}}{6.626\,070\,15 \times 10^{-34}\,\mathrm{J\,s}} = 1.9 \times 10^{-7}\,\mathrm{Hz}\,,$$

against a Larmor frequency of about 29 Hz at the working field of 1 µT: the bound is a frequency ratio measured to 7×10^{-9}, twice, with the field reversed. It is not a theoretical estimate of any kind. □

The number is worth converting once, because it is otherwise unimaginable. A dipole of $2.9 \times 10^{-47}\,\mathrm{C\,m}$ corresponds to separating one elementary charge from its opposite by $1.8 \times 10^{-28}\,\mathrm{m}$, about 10^{13} times smaller than the neutron itself, whose charge radius is of order 0.8 fm.

What the null result constrains. Two things, and they are of very different character. The first is new physics: the Cabibbo–Kobayashi–Maskawa phase generates a neutron dipole only at three loops with a severe chiral suppression, some six orders of magnitude below Equation (111.27).

Derivation pending #196

The three-loop, chirally suppressed Cabibbo–Kobayashi–Maskawa contribution to the neutron electric dipole moment is quoted here as an order of magnitude and is not derived. Owed: the argument that the one- and two-loop quark diagrams cancel by the Glashow–Iliopoulos–Maiani mechanism because the rephasing invariant needs all six quark masses, and the resulting estimate of order $3 \times 10^{-53}\,\mathrm{C\,m}$. *This belongs in Appendix A. The bibliography carries no key for the electric-dipole calculation, so the estimate is uncited.*

The entire experimental range is one in which the Standard Model predicts nothing observable. A non-zero result would be unambiguous, and the null result excludes new CP-violating sources over a wide range with no background to subtract.

The second is a parameter of the Standard Model itself. Quantum chromodynamics admits a term proportional to $\theta\ \mathrm{tr}\left(G_{\mu\nu}\tilde{G}^{\mu\nu}\right)$ which is a total derivative, invisible in perturbation theory, and physical nonperturbatively because the gauge field has topologically distinct sectors. Only the combination $\bar{\theta} = \theta + \arg\det M_q$ is invariant under the chiral rotations that move phase between the gluon and Yukawa sectors (Equation (105.111)), and $\bar{\theta}$ violates P and T, hence generates a neutron dipole. The chiral estimates [1633, 1637] give $d_n \approx \bar{\theta} \times 1.6 \times 10^{-37} - -6.4 \times 10^{-37}\,\mathrm{C\,m}$ (Equation (105.112)), so Equation (111.27) implies

$$\left|\bar{\theta}\right| \lesssim \frac{2.9 \times 10^{-47}\,\mathrm{C\,m}}{1.6 \times 10^{-37}\,\mathrm{C\,m}} \approx 1.8 \times 10^{-10}\,, \tag{111.28}$$

that is $\left|\bar{\theta}\right| \lesssim 10^{-10}$, with a factor-of-a-few uncertainty inherited from the chiral coefficient (Equation (105.113)).

This is the strong CP problem, and what makes it a problem is worth stating precisely, because it is unlike the other open questions in this chapter. $\bar{\theta}$ is a free dimensionless parameter, periodic with period 2π, so there is no sense in which a small value sits near a boundary of its range; a value of order unity is the natural expectation for a number about which nothing else is known. It is measured to be smaller than 10^{-10}. No symmetry of the Standard Model requires it to vanish — the one candidate, a massless up quark, is excluded by the lattice determination $m_u c^2 = (2.16 \pm 0.07)\,\mathrm{MeV}$ [2427] — and no dynamical mechanism within the theory relaxes it. The problem is an *observation*, and it is recorded as such in Section 133.1.

The best-known response is to make $\bar{\theta}$ dynamical: a global symmetry, spontaneously broken and itself anomalous, replaces it by a field whose instanton-generated potential is minimized at zero [1609], at the cost of a light pseudo-Goldstone boson, the axion [1628, 1630]. No axion has been observed. The haloscope, helioscope and light-shining-through-walls searches report exclusions and nothing else, and the mechanism is a hypothesis with an attractive structure and no experimental support; it is listed as such in Section 133.7 and named here only to say so. What this chapter records as evidence is Equation (111.28): a free parameter that could have been of order unity is smaller than 10^{-10}, and nothing explains why.

111.9 Primary references

The literature of CP violation divides into papers that predicted a system, papers that measured something in it, and papers that supplied the framework in which the measurement means something. It is worth separating them, because the reader who goes to the sources will find that several of the most-cited ones do not contain what they are cited for.

The system. Gell-Mann and Pais [1067] is the paper that invented the neutral kaon as a two-state problem and predicted the long-lived partner; it is a short paper and it is worth reading for the argument alone, which uses nothing beyond the superposition principle and a conservation law. Lande and collaborators [1090] reported the long-lived particle at the Brookhaven Cosmotron the following year. Pais and Piccioni [1077] predicted regeneration; the paper is the origin of the technique that produces K_S beams and of the background that Proposition 111.5 bounds.

The discovery. Christenson, Cronin, Fitch and Turlay [1274] is two and a half pages of *Physical Review Letters*, and every quantitative statement about the 1964 apparatus in Sections 111.2 and 111.3 comes from it. The paper is unusual in modern eyes for how much of its argument is carried by one two-dimensional plot and how carefully it enumerates the alternatives it must exclude. Note that the published ratio has "all charged decay modes" in its denominator, not the total width; the conversion is Equation (111.8), and a reader comparing the 1964 number directly with a modern branching fraction will find a spurious twenty per cent discrepancy.

The frameworks. Landau [1118] proposed CP as the surviving symmetry after the parity results of Wu [1133] and Garwin [1110]; Lüders [1123] proved the CPT theorem which is what makes CP violation and T violation the same statement. Wolfenstein [1300] supplied the superweak alternative in the same year as the discovery — a model that survived thirty-five years of measurement and was killed by ϵ'. Kobayashi and Maskawa [1505] extended Cabibbo's angle [1248] to three generations and showed a phase then becomes irremovable; the parametrization in which the hierarchy is manifest is Wolfenstein's second contribution [1733], and the convention-independent measure of the effect is Jarlskog's [1756]. Kabir [1426] is the source of the observable that CPLEAR measured.

The epsilon-prime programme. NA31 [1806] reported the first evidence; the contemporaneous Fermilab result is described in Section 111.6.1 but not cited, since this treatise's bibliography carries no entry for it, and Remark 111.10 says so. The two 1999 papers that settled the question are Fanti and collaborators for NA48 [1973] and Alavi-Harati and collaborators for KTeV [1965]. The world average is the Particle Data Group's [2427]; the lattice calculations to which it should be compared are not cited, for the same reason, and Section 111.6.3 says so.

Beyond the kaon. BaBar [2010] and Belle [2005] reported the $B^0 \to J/\psi K^0_S$ asymmetry simultaneously in 2001; the two papers are worth reading side by side, since they disagree by two standard deviations while both establishing the effect. LHCb's first observation of CP violation in B^0_s decays [2202] and in charm [2323] complete the set of quark flavours.

CPLEAR [1948] and BaBar [2197] are the two direct T tests. The global fit that combines all of it is CKMfitter [2074].

The cosmological side. Sakharov [1358] is a two-page note and states the three conditions in a form that has not needed amendment. 't Hooft [1569] showed that baryon number is anomalous in the electroweak theory; Kuzmin, Rubakov and Shaposhnikov [1758] showed that the resulting rate is unsuppressed at high temperature. Gavela and collaborators [1882] carried out the calculation that quantifies the shortfall, and Kajantie and collaborators [1925] showed by lattice computation that the electroweak transition is a crossover for any Higgs mass above about $72\,\mathrm{GeV}/c^2$, which removes the third Sakharov condition independently. The measured baryon-to-photon ratio is Planck's [2348].

The null results. The current neutron electric dipole moment bound is from ultracold neutrons at the Paul Scherrer Institute [2346], improving the Institut Laue-Langevin result [2090]; the technique is Ramsey's [968] and the argument for measuring the quantity at all was made by Purcell and Ramsey [966] at a time when parity was universally assumed. The chiral estimates relating the bound to $\bar{\theta}$ are Baluni [1633] and Crewther, Di Vecchia, Veneziano and Witten [1637]. Peccei and Quinn [1609] proposed the dynamical relaxation of $\bar{\theta}$ and Weinberg [1628] and Wilczek [1630] pointed out the axion that follows; none of it has been observed.

The data. Every world-average number in this chapter is from the Particle Data Group's 2024 Review [2427], and the masses in Table 111.5 are from its machine-readable release, stored with this treatise as evidence. The SI conversions — the elementary charge, the speed of light, the Planck and reduced Planck constants — are CODATA 2022 [2431], likewise stored; in the present SI the first three of those are exact, so the conversions in this chapter carry no uncertainty of their own.

Chapter 112

Experiment: The Electron Anomalous Magnetic Moment

Tests Phenomenon 93.3 and Equation (100.62).

Assuming Theorem 112.9 and Proposition 112.12.

The electron magnetic moment anomaly $a_e = (g-2)/2$ is the most precisely measured quantity in particle physics and the most precisely computed: the Harvard one-electron quantum cyclotron reports

$$a_e = 1.15965218059(13) \times 10^{-3}\,, \tag{112.1}$$

a fractional uncertainty of about 1.1×10^{-10} on the anomaly itself [2419], against a QED series evaluated through tenth order [2330]. That 2023 measurement is the experiment this chapter is built around, and it is the year registered for the chapter in the List of Experiments; the two earlier results from the same apparatus [2101, 2126] are reported alongside it because the systematic budget is cumulative and the three values are not independent. This chapter is the experimental counterpart of the renormalized perturbation theory of Chapter 100: it takes Dirac's $g = 2$ [596], the first measured departure from it [912, 916], and Schwinger's one-loop $a_e = \alpha/2\pi$ [919], and follows the comparison down to the digit at which it currently stops.

Where it stops is the point of the chapter, and it is not where the headline figures suggest. Extracting a_e from theory requires an independently measured fine-structure constant, and the two best determinations — caesium recoil [2316] and rubidium recoil [2363] — disagree with each other by several standard deviations, a discrepancy far larger than either quoted uncertainty. The comparison of measurement with theory is therefore verified only to the precision of the worse-behaved input, and the chapter states that bound explicitly rather than quoting agreement to twelve digits. The same discipline is applied to the muon anomaly, where a genuine measurement–theory tension [2091, 2368, 2416] is entangled with a disagreement between lattice and data-driven evaluations of the hadronic vacuum polarization [2375]. The chapter follows the house experiment pattern of editorial rule 4: apparatus, procedure, observations and data, interpretation, primary references, in that order, with the historical setting and the theoretical prediction placed before the apparatus so that the reader knows what is being tested before meeting the hardware.

Remark 112.1 (Units, and the literature's convention). Everything in this chapter is written in SI. Frequencies are in hertz, fields in tesla, energies in joules with the electronvolt given alongside, and $\hbar$ and c appear explicitly in every equation, including the intermediate steps. The anomaly itself is a pure number, so no unit convention can change it — which is precisely why it is the sharpest comparison available between a measurement and a calculation.

The QED literature computes a_e with $\hbar = c = 1$, in which convention masses, momenta and energies are all quoted in electronvolts and the fine-structure constant is written

$\alpha = e^2/4\pi$ with a rationalized Heaviside–Lorentz charge. To map a formula from that literature onto this chapter, restore one factor of $\hbar$ for each unit of action, one factor of c for each power of velocity, and read the coupling as

$$\alpha = \frac{e^2}{4\pi\varepsilon_0\hbar c} = (7.297\,352\,564\,3 \pm 0.000\,000\,001\,1) \times 10^{-3}\,, \tag{112.2}$$

which is dimensionless in SI as it is in any other system [2431]. No derivation in this chapter is carried out in the natural-units convention; this remark exists only so that a reader holding [2330] open beside the book can align the two.

112.1 Historical context and the prediction under test

112.1.1 Dirac's $g = 2$

A magnetic moment attached to an angular momentum is written

$$\boldsymbol{\mu} = -g\,\frac{\mu_{\mathrm{B}}}{\hbar}\,\boldsymbol{J}\,, \qquad \mu_{\mathrm{B}} = \frac{e\hbar}{2m_e} = (9.274\,010\,065\,7 \pm 0.000\,000\,002\,9) \times 10^{-24}\,\mathrm{J/T}\,, \tag{112.3}$$

the numerical value of the Bohr magneton being the CODATA 2022 recommended one [2431]. The dimensionless factor g is what is at issue. For a classical current loop, or for the orbital motion of a charge in an atom, $g = 1$: this follows from nothing more than the definition of the magnetic moment of a current distribution, $\boldsymbol{\mu} = \frac{1}{2}\int \mathrm{d}^3x\, \boldsymbol{x} \times \boldsymbol{j}$, applied to a charge $-e$ of mass m_e moving on a closed orbit, whose angular momentum is $\boldsymbol{L} = m_e \int \mathrm{d}^3x\, \boldsymbol{x} \times \boldsymbol{v}\, \rho/(-e)$ with ρ the charge density.

Spin is not orbital motion, and the factor is not 1. Uhlenbeck and Goudsmit had to postulate $g = 2$ for the electron's intrinsic angular momentum to make the alkali doublets and the anomalous Zeeman effect come out [532], and Pauli inserted the same factor by hand into the nonrelativistic wave equation [587]. What Dirac's relativistic equation supplies, without being asked, is exactly that factor [596]. The derivation is carried out in full elsewhere in this book — see Phenomenon 93.3, where squaring the minimally coupled Dirac operator produces the term $-(q\hbar/2m_e)\,\boldsymbol{\sigma}\cdot\boldsymbol{B}$ and the factor of two comes entirely from the algebra of $(\boldsymbol{\sigma}\cdot\boldsymbol{a})(\boldsymbol{\sigma}\cdot\boldsymbol{b})$ — and it is not repeated here.

Two features of that result are what make the present chapter possible. First, $g = 2$ is exact in the single-particle theory: it is not fitted, it contains no parameter, and it does not depend on the field strength, the electron's velocity, or anything else that could be adjusted. Second, it is a statement about a *free* electron, so a measurement on a free electron tests it without any intervening atomic structure. Any measured departure from 2 is therefore a measurement of something that the relativistic wave equation does not contain, and the only candidate in a theory built on it is the interaction of the electron with the quantized electromagnetic field. That is why the anomaly is diagnostic rather than merely a correction.

It is worth being explicit about the size of the effect to be found, and about a sign convention that is easy to count twice. In Equation (112.3) the orientation of the moment relative to the spin is carried by the explicit minus sign, so the g written there is a positive number, and the anomaly is defined on that magnitude:

$$a_e := \frac{|g_e| - 2}{2}\,, \qquad \mu_e = (1 + a_e)\,\mu_{\mathrm{B}}\,. \tag{112.4}$$

CODATA 2022 tabulates the g factor with its sign included, $g_e = -2.00231930436092(36)$, the sign recording that the moment is antiparallel to the spin for a negative charge [2431]. Every g written in this chapter is $|g_e|$; inserting the tabulated negative value into Equation (112.4) would apply the same sign twice and give -2.0012 instead of the anomaly. The excess over

the Dirac value is therefore about one part in 860 — large enough to have been found in 1948 with atomic beams, small enough that measuring its own tenth significant figure is a seventy-five-year project.

112.1.2 The hyperfine anomaly and the Kusch–Foley measurement

The excess was not found by measuring the electron moment against an absolute standard. It was found because a set of measurements that assumed $g_S = 2$ failed to be self-consistent.

The atomic beam. Kusch and Foley used the atomic-beam magnetic-resonance method that Rabi's group had developed a decade earlier [798, 819]. An oven produces a thermal beam of neutral atoms — gallium, indium and sodium in the experiments at issue — which passes in vacuum through three magnet regions in series. The first (the A magnet) has a strong transverse gradient and deflects the atoms according to the sign of their effective moment; the third (the B magnet) has a gradient of the opposite sense and refocuses onto a detector exactly those atoms whose effective moment is unchanged. Between them sits the C region: a strong, uniform field $\boldsymbol{B}$ with a small radio-frequency field applied transverse to it. If the radio frequency matches a Zeeman transition frequency of the atom in the C field, the atom changes state, the B magnet no longer refocuses it, and the detected beam intensity drops. The observable is a dip in a beam current as a frequency is swept.

Why a ratio evades the calibration problem. The resonance condition in the C field is

$$h\nu = g_F \, \mu_{\mathrm{B}} \, B \, \Delta m_F \, , \tag{112.5}$$

so an absolute determination of g_F requires knowing B to the same accuracy as ν — and in 1948 a magnetic field could not be measured to anything like a part in 10^4, whereas a radio frequency could be counted to a part in 10^6. The way out is to measure two resonances in the *same* field. The ratio of their frequencies is

$$\frac{\nu_1}{\nu_2} = \frac{g_{F,1} \, \Delta m_{F,1}}{g_{F,2} \, \Delta m_{F,2}} \, , \tag{112.6}$$

in which B, h and μ_{B} have all cancelled. The field need only be stable for the duration of the two sweeps, not known. This is the same structural trick — measure a ratio of frequencies in a common field — that the Penning-trap experiments of Section 112.3.2 carry to a part in 10^{13}, and it is the single most important idea in the whole subject.

The inconsistency. For an atom in Russell–Saunders coupling, the Landé factor of a level with quantum numbers L, S, J is a known function of g_L and g_S,

$$g_J = g_L \, \frac{J(J+1) - S(S+1) + L(L+1)}{2J(J+1)} + g_S \, \frac{J(J+1) + S(S+1) - L(L+1)}{2J(J+1)} \, , \tag{112.7}$$

with the orbital factor $g_L = 1$ up to the known reduced-mass correction.

Derivation. The total moment of the electron shell is the sum of an orbital and a spin part,

$$\boldsymbol{\mu} = -\frac{\mu_{\mathrm{B}}}{\hbar} \left(g_L \boldsymbol{L} + g_S \boldsymbol{S} \right) \, ,$$

which is *not* proportional to $\boldsymbol{J} = \boldsymbol{L} + \boldsymbol{S}$ unless $g_L = g_S$. Within one level of definite J, however, the Wigner–Eckart theorem makes every vector operator proportional to $\boldsymbol{J}$: for

any vector $\boldsymbol{V}$, $\langle \boldsymbol{V} \rangle = \langle \boldsymbol{V}\cdot\boldsymbol{J}\rangle \, \langle \boldsymbol{J}\rangle / \hbar^2 J(J+1)$. Applying this with $\boldsymbol{V} = g_L \boldsymbol{L} + g_S \boldsymbol{S}$ defines g_J through $\langle \boldsymbol{\mu}\rangle = -g_J(\mu_{\mathrm{B}}/\hbar)\,\langle \boldsymbol{J}\rangle$, so that

$$g_J = \frac{g_L \langle \boldsymbol{L}\cdot\boldsymbol{J}\rangle + g_S \langle \boldsymbol{S}\cdot\boldsymbol{J}\rangle}{\hbar^2 J(J+1)} \, .$$

Squaring $\boldsymbol{S} = \boldsymbol{J} - \boldsymbol{L}$ and $\boldsymbol{L} = \boldsymbol{J} - \boldsymbol{S}$ gives the two scalar products,

$$\boldsymbol{L}\cdot\boldsymbol{J} = \tfrac{1}{2}\left(\boldsymbol{J}^2 + \boldsymbol{L}^2 - \boldsymbol{S}^2\right) , \qquad \boldsymbol{S}\cdot\boldsymbol{J} = \tfrac{1}{2}\left(\boldsymbol{J}^2 + \boldsymbol{S}^2 - \boldsymbol{L}^2\right) ,$$

and inserting the eigenvalues $\hbar^2 J(J+1)$, $\hbar^2 L(L+1)$ and $\hbar^2 S(S+1)$ — the factors of $\hbar^2$ cancelling against the denominator, as they must for g_J to be dimensionless — gives Equation (112.7). □

Because the two coefficients depend on L, S and J, different levels weight g_S and g_L differently: the ${}^2P_{1/2}$ and ${}^2P_{3/2}$ levels of gallium and indium, and the ${}^2S_{1/2}$ ground state of sodium, give three independent combinations. Measuring the ratios of their g factors therefore over-determines the pair (g_L, g_S), and the system has no solution with $g_S = 2g_L$.

Kusch and Foley found that the discrepancy is removed by allowing the electron's spin g factor to exceed twice the orbital one by about 0.12 % [912, 916]. In the modern notation that is $a_e \approx 1.2 \times 10^{-3}$, and the reader should notice how coarse the measurement was by present standards and how decisive it was anyway: the quantity being tested was believed exact, so a departure of one part in 10^3, established at many times its own uncertainty, could not be absorbed into any parameter of the theory. The two 1948 papers are a matched pair — Foley and Kusch [912] carries the interpretation and Kusch and Foley [916] the fuller experimental account — and they are cited together throughout this chapter for that reason.

Phenomenon 112.2 (The electron magnetic moment is not the Dirac value)**.** The magnetic moment of the electron exceeds one Bohr magneton. The excess was first established from the internal inconsistency of atomic-beam magnetic-resonance measurements of hyperfine g-factor ratios in gallium, indium and sodium, which could not be reconciled with $g_S = 2$ and required an electron moment larger by about 0.12 % [912, 916]. The modern value of the anomaly $a_e = (|g_e| - 2)/2$ of Equation (112.4) is

$$a_e = 1.15965218059(13) \times 10^{-3} \, , \tag{112.8}$$

an uncertainty of 0.13 parts per trillion on $g/2$ [2419].

Rests on Equation (112.4) *and Theorem 100.41.*

Derivation. *Derives Phenomenon 112.2.*

The excess is not a fitted parameter but a computed one. At one loop the correction to the electron–photon vertex gives $a_e = \alpha/2\pi = 1.1614 \times 10^{-3}$ [919], which already reproduces Equation (112.8) to three significant figures and accounts for the whole of the 0.12 % excess that Kusch and Foley found; the loop integral is evaluated in full at Theorem 100.41 and is not repeated here. The remaining figures are supplied by the higher orders of the same series, carried to tenth order [2330] and set out term by term in Table 112.2. What limits the comparison is not the series but the input discussed in Section 112.2.3. □

112.1.3 Schwinger's $\alpha/2\pi$

The one-loop correction to the electron–photon vertex splits, by Lorentz invariance and current conservation, into two form factors: $F_1(q^2)$, which renormalizes the charge, and $F_2(q^2)$, which multiplies the tensor structure $\mathrm{i}\sigma^{\mu\nu}q_\nu/2m_e c$ and therefore contributes an extra magnetic coupling. The anomaly is $a_e = F_2(0)$, and at one loop

$$a_e^{(1)} = \frac{\alpha}{2\pi} = \frac{e^2}{8\pi^2 \varepsilon_0 \hbar c} = 1.161\,410 \times 10^{-3} \, . \tag{112.9}$$

This chapter does not repeat that calculation: it is Theorem 100.41, proved in full in Chapter 100, where the loop integral is regularized, the divergence is shown to sit entirely in F_1, and $F_2(0)$ is extracted as a convergent Feynman-parameter integral. What matters here is the epistemic shape of the result, and it is worth stating carefully because it is the reason renormalization became believable rather than merely tolerated.

The one-loop vertex is divergent. The divergence is absorbed into the definition of the electron's charge, which is not predicted by the theory and never was. What is left over, after that absorption, is a pure number times α — no cutoff, no regulator, no scheme dependence, and no adjustable parameter. Schwinger's number was published in 1948, in the same year and in the same journal volume as the Kusch–Foley measurement, and it agreed with it. A theory that had been widely regarded as a formal device for organizing divergences had produced a parameter-free number that matched an experiment nobody had computed it for. Everything in the rest of this chapter is a seventy-five-year continuation of that single comparison, carried from three significant figures to ten.

The second-order term is worth one further remark. Because Equation (112.9) contains no mass, the same number is the one-loop anomaly of the muon and of the tau. Lepton-universality of the leading term is what makes the electron and the muon anomalies complementary rather than redundant: they differ only in the corrections, and the corrections are what carry information about hadrons and about anything heavier (Section 112.2.2).

112.1.4 What a measurement must deliver

Phenomenon 112.3 (The anomaly is a ratio of two frequencies)**.** For a free electron in a uniform magnetic field $\boldsymbol{B}$, the spin precession frequency ν_s and the cyclotron frequency ν_c satisfy

$$a_e = \frac{\nu_s - \nu_c}{\nu_c} = \frac{\nu_a}{\nu_c}\,, \qquad \nu_a := \nu_s - \nu_c\,, \tag{112.10}$$

in which the magnetic field has cancelled identically. The anomaly can therefore be obtained from two frequency measurements in one field, with no absolute magnetometry anywhere in the chain.

Rests on Equations (112.3) *and* (112.4).

Derivation. *Derives Phenomenon 112.3.*

The cyclotron motion of a charge $-e$ of mass m_e in a uniform field $\boldsymbol{B} = B\hat{\boldsymbol{z}}$ is quantized into Landau levels of energy

$$E_n = \left(n + \tfrac{1}{2}\right)\hbar\omega_c\,, \qquad \omega_c = \frac{eB}{m_e}\,, \qquad \nu_c = \frac{\omega_c}{2\pi} = \frac{eB}{2\pi m_e}\,, \tag{112.11}$$

with $\hbar\omega_c$ carrying SI dimensions of joules, ω_c of rad/s and ν_c of hertz. The spin contributes the Zeeman energy $-\boldsymbol{\mu}\cdot\boldsymbol{B}$, which from Equation (112.3) splits the two spin projections $m_s = \pm\frac{1}{2}$ by

$$\Delta E_{\text{spin}} = g\,\mu_{\text{B}} B = g\,\frac{e\hbar B}{2m_e} = \frac{g}{2}\,\hbar\omega_c\,, \tag{112.12}$$

so the spin precession frequency is $\nu_s = \Delta E_{\text{spin}}/h = (g/2)\nu_c$. Hence

$$\frac{\nu_s - \nu_c}{\nu_c} = \frac{g}{2} - 1 = \frac{g-2}{2} = a_e\,,$$

which is Equation (112.10). Both B and the ratio e/m_e have cancelled: the field enters ν_s and ν_c in exactly the same way because the same μ_{B} and the same ω_c appear in both, and that identity is what Equations (112.11) and (112.12) exhibit. The anomaly is the fractional difference of two frequencies measured on one particle in one field. □

The consequences of this structure govern everything that follows, and they are worth setting out before any hardware is described.

First, the measurement is a frequency measurement, and frequency is the SI quantity that can be realized best: the second is defined by a caesium hyperfine transition and time intervals are routinely compared to a part in 10^{16}, far beyond what is needed here (Chapter 2). Nothing else in the experiment has to be that good.

Second, the difference $\nu_a = \nu_s - \nu_c = a_e \nu_c$ is smaller than either frequency by three orders of magnitude, so a fractional error ϵ in either ν_s or ν_c produces a fractional error of order $\epsilon / a_e \approx 860\,\epsilon$ in a_e unless it is common to both. This is the reason ν_a is measured *directly*, as the frequency of a transition that flips the spin and lowers the cyclotron quantum number by one, rather than as a difference of two separately measured large numbers. Measured that way, the requirement on ν_c is only that it be known to about a_e times the target precision.

Third, since B does not appear, magnetometry is replaced by a different set of systematic effects: whatever makes the trap's frequencies differ from the free-space ones. Those are the electrostatic confining potential itself, the special-relativistic mass increase, the image charges induced in the electrodes, and — dominantly — the coupling of the cyclotron motion to the modes of the conducting cavity that surrounds it. These four are the whole of the systematic problem, and Sections 112.3.2 and 112.3.3 exist to explain how each is controlled; the published line-by-line error budget is not reproduced in this book and must be read in [2419].

Fourth, and least obviously, the quantity delivered is dimensionless. No SI unit enters the final number at all, which is why a_e can be compared with a calculation without a single conversion factor and why its uncertainty is not limited by the realization of the kilogram, the ampere or the tesla. The uncertainty conventions used to quote it — standard uncertainty, coverage factor 1, correlations propagated explicitly — are those of Chapter 2.

112.2 The theoretical prediction

112.2.1 The QED series through tenth order

Because the one-loop vertex correction of Theorem 100.41 is α times a dimensionless integral, and because every additional loop adds one factor of the coupling and one dimensionless integral, the anomaly of a lepton is a power series in α with pure numbers as coefficients:

$$a_e = \sum_{n \geq 1} C_n \left(\frac{\alpha}{\pi} \right)^n + a_e^{\mu,\tau} + a_e^{\mathrm{had}} + a_e^{\mathrm{weak}} . \tag{112.13}$$

The expansion parameter is written α/π rather than α because a factor $1/\pi$ per loop is what the four-dimensional loop measure $\mathrm{d}^4\ell / (2\pi\hbar)^4$ actually supplies; with $\alpha/\pi = 2.3228 \times 10^{-3}$ the terms fall by a factor of a few hundred at each order, which is why ten significant figures need only five of them. Every C_n is a pure number, and every term in Equation (112.13) is dimensionless in SI, as it must be for the sum to be a g factor.

The coefficients are those adopted in Chapter 100, whose provenance is [2330]:

$$\begin{aligned} &C_1 = \tfrac{1}{2}\,, && C_2 = -0.328478965\ldots\,, && C_3 = +1.181241456\ldots\,, \\ &C_4 = -1.912245\ldots\,, && C_5 = +6.737(159)\,. \end{aligned} \tag{112.14}$$

C_1 is Schwinger's; C_2 and C_3 are known in closed form, in terms of $\zeta(3)$, $\ln 2$ and π^2; C_4 and C_5 are numerical. The counting of Feynman diagrams contributing to the mass-independent part at each order is 1, 7, 72, 891 and 12 672 [2190, 2302, 2330], which is the practical reason the series stops where it does: the tenth-order evaluation is a decade-long numerical project, not an algebraic one.

Two features of Equation (112.14) deserve to be flagged rather than passed over.

The tenth-order term is not negligible, and it is disputed. With the CODATA 2022 value of α the fifth-order term contributes $C_5(\alpha/\pi)^5 = 4.556 \times 10^{-13}$ to a_e, which is *three and a half times* the experimental uncertainty 1.3×10^{-13} of Equation (112.8). The quoted uncertainty on C_5 itself, ± 0.159, propagates to 1.1×10^{-14}, comfortably below the experiment. But that uncertainty is the error of a numerical integration, and an independent recomputation of the largest gauge-invariant subset of the tenth-order diagrams — those without closed lepton loops — does not agree with the published value within the quoted errors [2343]. This is recorded here, as it is in Chapter 100, because a single unverified numerical computation would otherwise be load-bearing in the most precisely tested prediction in physics. It does not currently change the conclusion of Section 112.7, because the α input is a larger limitation than any plausible revision of C_5; it would matter immediately if the α inputs were reconciled.

The mass-dependent terms enter at fourth order. The C_n of Equation (112.14) are the mass-independent coefficients: they are what remains when every internal fermion line is the same lepton as the external one. Diagrams in which a muon or tau loop is inserted into a photon propagator depend on the ratios m_e/m_μ and m_e/m_τ and are collected into $a_e^{\mu,\tau}$. Their leading behaviour is a clean piece of physics and is derived in Section 112.2.2: a loop of mass M decouples from the anomaly of a lepton of mass m_ℓ as $(m_\ell/M)^2$, so muon loops are suppressed in a_e by $(m_e/m_\mu)^2 \approx 2.34 \times 10^{-5}$. The full value from [2330] is

$$a_e^{\mu,\tau} = 2.7478 \times 10^{-12}\,, \tag{112.15}$$

which is twenty times the experimental uncertainty and therefore mandatory. The logarithms $\ln(m_\mu/m_e)$ that accompany such ratios are the mass singularities analysed by Kinoshita [1232]; they are finite here because the internal mass is the larger one.

112.2.2 Hadronic and electroweak contributions

The remaining two terms of Equation (112.13) are not QED. The hadronic term collects diagrams in which the virtual photon converts into quarks and gluons — hadronic vacuum polarization at leading order, and hadronic light-by-light scattering beyond it — and the electroweak term collects diagrams containing a W, a Z or a Higgs boson. For the electron [2330] gives

$$a_e^{\mathrm{had}} = 1.6927 \times 10^{-12}\,, \qquad a_e^{\mathrm{weak}} = 0.030\,53 \times 10^{-12}\,. \tag{112.16}$$

Neither can be computed from first principles in perturbation theory: the hadronic one because the strong coupling is large at the relevant scale (Chapter 102), and it is obtained instead either from measured $e^+e^- \to$ hadrons cross sections through a dispersion relation or from lattice QCD (Remark 102.61). The electroweak one is perturbative but requires the full Standard Model.

Phenomenon 112.4 (Hadronic sensitivity scales as the square of the lepton mass)**.** The hadronic contribution to a lepton anomaly is a small correction for the electron and the dominant theoretical uncertainty for the muon. Numerically the hadronic terms contribute of order 1.7×10^{-12} to a_e [2330] — already an order of magnitude above the experimental uncertainty of Equation (112.8), and therefore no longer negligible — and of order 7×10^{-8} to a_μ [2353].

Rests on Equations (112.8)*,* (112.13) *and* (112.16)*.*

Derivation. Derives Phenomenon 112.4.

The hadronic vacuum polarization enters the anomaly through a dispersion integral over the measured cross section for $e^+e^- \to$ hadrons, weighted by a kernel which, for centre-of-mass energies $\sqrt{s}$ well above the lepton mass, behaves as $K(s) \simeq m_\ell^2 c^4/(3s)$. The integral is

dominated by the hadronic resonances, whose scale is fixed by the strong interaction and is the same whichever lepton is being computed, so the whole contribution scales as m_ℓ^2. The ratio for the two leptons is therefore

$$\frac{a_\mu^{\mathrm{had}}}{a_e^{\mathrm{had}}} \simeq \left(\frac{m_\mu}{m_e}\right)^2 = (206.7683)^2 = 4.28 \times 10^4 \,, \tag{112.17}$$

using the mass ratio built from the CODATA 2022 electron and muon masses $m_e = (9.109\,383\,713\,9 \pm 0.000\,000\,002\,8) \times 10^{-31}\,\mathrm{kg}$ and $m_\mu = (1.883\,531\,627 \pm 0.000\,000\,042) \times 10^{-28}\,\mathrm{kg}$ [2431]; applied to the electron value quoted above this reproduces the muon one. The same m_ℓ^2 scaling governs the electroweak contribution and the sensitivity to any new heavy state, and it is the whole reason the muon is the better probe of physics beyond the Standard Model despite its far worse experimental precision — and the reason the electron anomaly, whose theory is almost purely QED, is the better test of QED itself. □

The same statement can be checked against a case where the heavy mass is known exactly, which is a useful consistency test of the scaling argument just made. The two-loop contribution to a_e from a vacuum polarization loop of a heavier lepton of mass M is

$$\delta a_e = \left(\frac{\alpha}{\pi}\right)^2 \left[\frac{1}{45}\left(\frac{m_e}{M}\right)^2 + O\!\left(\left(\frac{m_e}{M}\right)^4 \ln\frac{M}{m_e}\right)\right] , \tag{112.18}$$

which is dimensionless, as every term in Equation (112.13) must be.

Derivation. At this order the heavy lepton reaches the electron only by polarizing the internal photon of the one-loop vertex, so the dispersive representation of Equation (112.43) applies verbatim with the hadronic $R(s)$ replaced by the contribution of one lepton pair of mass M,

$$R_M(s) = \left(1 + \frac{2M^2c^4}{s}\right)\sqrt{1 - \frac{4M^2c^4}{s}} \,, \qquad s \geq 4M^2c^4 \,, \tag{112.19}$$

which is the usual threshold factor for producing a pair of mass M and is dimensionless, s carrying SI dimensions of J^2. Because $M \gg m_e$ the entire integration region satisfies $s \geq 4M^2c^4 \gg m_e^2c^4$, so the asymptotic kernel $K(s) \to m_e^2c^4/(3s)$ of Equation (112.43) may be used throughout, and

$$\delta a_e = \frac{1}{3}\left(\frac{\alpha}{\pi}\right)^2 \int_{4M^2c^4}^{\infty} \frac{\mathrm{d}s}{s}\, R_M(s)\, \frac{m_e^2c^4}{3s} = \frac{1}{9}\left(\frac{\alpha}{\pi}\right)^2 m_e^2c^4 \int_{4M^2c^4}^{\infty} \frac{R_M(s)}{s^2}\,\mathrm{d}s \,.$$

Substitute $u = 4M^2c^4/s$, which maps the range onto $0 < u \leq 1$ and turns $\mathrm{d}s/s^2$ into $-\mathrm{d}u/(4M^2c^4)$:

$$\int_{4M^2c^4}^{\infty} \frac{R_M(s)}{s^2}\,\mathrm{d}s = \frac{1}{4M^2c^4}\int_0^1 \left(1 + \frac{u}{2}\right)\sqrt{1-u}\,\mathrm{d}u = \frac{1}{4M^2c^4}\left(\frac{2}{3} + \frac{2}{15}\right) = \frac{1}{5M^2c^4} \,,$$

using $\int_0^1 \sqrt{1-u}\,\mathrm{d}u = \frac{2}{3}$ and $\int_0^1 u\sqrt{1-u}\,\mathrm{d}u = \frac{4}{15}$. Hence

$$\delta a_e = \frac{1}{9}\left(\frac{\alpha}{\pi}\right)^2 \frac{m_e^2c^4}{5M^2c^4} = \frac{1}{45}\left(\frac{\alpha}{\pi}\right)^2 \left(\frac{m_e}{M}\right)^2 ,$$

which is the leading term of Equation (112.18). What has been dropped is the correction to the asymptotic kernel, of relative order $m_e^2c^4/s \leq (m_e/2M)^2$, and that is the source of the $O((m_e/M)^4)$ remainder shown. Note that the factor $1/3$ inside the kernel is load-bearing: with $K \to m_e^2c^4/s$ the same integral would give $1/15$, three times too large. □

With $M = m_\mu$, so that $(m_e/m_\mu)^2 = 2.3390 \times 10^{-5}$, this gives $5.3955 \times 10^{-6} \times 5.198 \times 10^{-7} = 2.80 \times 10^{-12}$. The full mass-dependent term of Equation (112.15) is 2.7478×10^{-12}, so the leading $(m_e/M)^2$ piece accounts for it to two per cent — but note the direction: the leading piece *overshoots* the full value, and the residual $2.7478 \times 10^{-12} - 2.8046 \times 10^{-12} = -5.7 \times 10^{-14}$ is negative. The same formula with $M = m_\tau$ adds a tau loop of about $+1 \times 10^{-14}$, which pushes the other way, so the negative sign must come from the three-loop mass-dependent diagrams, which are the only contribution of the right sign at this size. The decoupling of a heavy state as the square of the mass ratio is therefore not an approximation invented for the hadronic estimate; it is a property of the anomaly, verified here against an independently known mass.

Remark 112.5 (What the anomaly can and cannot see). Equation (112.18) sets the reach of the measurement. A new state of mass M coupling to the lepton with roughly electromagnetic strength shifts the anomaly by of order $(\alpha/\pi)(m_\ell c^2/Mc^2)^2$, so an experimental sensitivity δa_ℓ probes

$$Mc^2 \gtrsim m_\ell c^2 \sqrt{\frac{\alpha}{\pi\, \delta a_\ell}}\,. \tag{112.20}$$

The lepton mass in the numerator is why the muon, 206.7683 times heavier, buys a factor 206.7683 in mass reach for the same fractional precision. The square root is why the electron's enormous advantage in precision does not simply transfer: the electron anomaly is measured about 1840 times more precisely than the muon's in fractional terms (1.1×10^{-10} against 2.1×10^{-7}, from Tables 112.1 and 112.4), and a factor 1840 in δa_ℓ buys only $\sqrt{1840} \approx 43$ in mass — less than the mass ratio it is competing against. The two experiments are answering different questions: the electron anomaly tests QED, the muon anomaly searches. Numbers are given in Section 112.7.

112.2.3 The fine-structure constant as an input

Equation (112.13) is a function of α. It predicts a number only when α is supplied, and if α is taken from the anomaly itself the exercise is a definition rather than a test. This subsection sets out where an independent α comes from and how well it is known; Section 112.7 draws the consequence.

The best independent determinations are atom-recoil measurements, and the route is worth deriving because it shows exactly which quantity carries the uncertainty. The Rydberg constant is, by definition,

$$R_\infty = \frac{m_e c\, \alpha^2}{2h} = (10\,973\,731.568\,157 \pm 0.000\,012)/\mathrm{m}\,, \tag{112.21}$$

the value being CODATA 2022 [2431]. Solving for α^2 and inserting the mass of an atom of species X in the obvious place,

$$\alpha^2 = \frac{2R_\infty h}{m_e c} = \frac{2R_\infty}{c}\,\frac{m_X}{m_e}\,\frac{h}{m_X}\,. \tag{112.22}$$

Each of the three factors on the right is separately measurable. The first is Equation (112.21), known to 1.1×10^{-12} fractionally from hydrogen spectroscopy and utterly negligible here. The second, the mass of the atom in units of the electron mass, is a ratio of cyclotron frequencies in a Penning trap — the same instrument as Section 112.3.2, used for a different purpose — and is known to better than 10^{-10}. The third, h/m_X, has the SI dimensions of J s/kg, that is m^2/s, and is what the recoil experiment measures: an atom that absorbs a photon of wavevector $k = 2\pi/\lambda$ recoils with velocity

$$v_{\mathrm{rec}} = \frac{\hbar k}{m_X} = \frac{h}{m_X \lambda}\,, \tag{112.23}$$

and the resulting Doppler shift, accumulated over many photon recoils in an atom interferometer, is counted as a frequency. Since Equation (112.22) gives α as the square root of h/m_X, a fractional uncertainty ϵ on h/m_X becomes $\epsilon/2$ on α: a determination of α at 2.0×10^{-10} requires h/m_X at 4.0×10^{-10}, which is the actual achievement of these experiments.

Phenomenon 112.6 (The two recoil determinations disagree)**.** The fine-structure constant can be obtained from a photon-recoil frequency together with the mass of the recoiling atom in units of the electron mass. Two such determinations,

$$\alpha^{-1} = 137.035999046(27)\ \text{(caesium)}\,, \qquad \alpha^{-1} = 137.035999206(11)\ \text{(rubidium)}\,, \tag{112.24}$$

differ by 1.60×10^{-7}, about five and a half times their combined standard uncertainty [2316, 2363]. The recommended value accordingly carries an inflated uncertainty [2431].

Rests on Equations (112.8) *and* (112.22).

Derivation. *Derives Phenomenon 112.6.*

The consequence for the comparison of Equation (112.8) with theory is arithmetic. To leading order $a_e = \alpha/2\pi$, so $\delta a_e = \delta\alpha/2\pi$ and, fractionally, $\delta a_e/a_e = \delta\alpha/\alpha$. The two inputs of Equation (112.24) differ fractionally by

$$\frac{\delta\alpha}{\alpha} = \frac{1.60 \times 10^{-7}}{137.036} = 1.17 \times 10^{-9}\,,$$

and with $\alpha = 7.2974 \times 10^{-3}$ this is $\delta\alpha = 8.5 \times 10^{-12}$, so that

$$\delta a_e = \frac{\delta\alpha}{2\pi} = 1.4 \times 10^{-12}\,. \tag{112.25}$$

The experimental uncertainty on a_e in Equation (112.8) is 1.3×10^{-13}. The two predictions therefore straddle the measurement by about ten times its own uncertainty, and the honest statement of the test is that quantum electrodynamics is confirmed to the precision of its input — roughly one part in 10^9 of a_e, about nine significant figures — and not to the twelve or thirteen that the quoted precision of a_e alone would suggest [2419]. The alternative reading, in which a_e together with the series is used to *define* α, is internally consistent but defers the test until the two recoil measurements agree. □

The two experiments are not variants of one technique in any way that would make a common systematic likely. Parker and collaborators [2316] used a caesium atom interferometer with Ramsey–Bordé interferometry and Bragg diffraction; Morel and collaborators [2363] used rubidium with Bloch oscillations in an accelerated optical lattice to accumulate hundreds of photon recoils. The atomic species differ, the mass ratios m_X/m_e come from different measurements, and the interferometric schemes differ. That is what makes the disagreement uncomfortable: it is not obviously attributable to one of them, and the CODATA 2022 adjustment therefore inflates the uncertainty of the recommended value $\alpha^{-1} = 137.035999177(21)$ rather than choosing between them [2431]. Note that this recommended value is not independent of a_e — the adjustment uses the anomaly together with the QED series as one of its inputs — which is exactly why it cannot be used to test the series. That distinction is the subject of Section 112.7 and is the single most important statement in this chapter.

112.3 Apparatus

112.3.1 Free-electron precession experiments

The first measurements to use Equation (112.10) directly, on free rather than bound electrons, were made at Michigan by Crane and collaborators over the 1950s and early 1960s, culminating in the precision experiment of Wilkinson and Crane [1268].

The Michigan experiment. Electrons of about 100 keV, that is 1.602×10^{-14} J, of kinetic energy were transversely polarized by Mott scattering from a thin gold foil — the spin–orbit coupling in the Coulomb field of a high-Z nucleus makes the differential cross section depend on the spin projection normal to the scattering plane, so scattering at a fixed angle selects a polarized beam. The polarized electrons were injected into a region of static, nearly uniform magnetic field of order 0.1 T, in which a weak axial gradient provided a shallow magnetic bottle that trapped them in a spiral orbit for a controlled dwell time. They were then ejected onto a second gold foil, and the left–right asymmetry of the Mott scattering measured the angle between the spin and the momentum at the moment of ejection.

Why the method works. This apparatus measures the anomaly by allowing it to accumulate.

Proposition 112.7 (The anomaly accumulates as a turning angle)**.** *An electron that completes N cyclotron orbits in a uniform magnetic field turns its spin, relative to its momentum, through the angle*

$$\theta = 2\pi a_e N \,. \tag{112.26}$$

Rests on Phenomenon 112.3.

Proof. *Derives Proposition 112.7.*

The momentum direction rotates at $\omega_c = eB/m_e$ and the spin precesses at $\omega_s = (g/2)\,\omega_c$, both about $\hat{\boldsymbol{z}}$ and both derived in Phenomenon 112.3. Their difference is

$$\omega_a = \omega_s - \omega_c = \left(\frac{g}{2} - 1\right)\omega_c = a_e\,\omega_c \,,$$

so after a time t the accumulated relative angle is $\theta = a_e \omega_c t$. In N orbits, $\omega_c t = 2\pi N$, which gives Equation (112.26). □

The point is the factor N. A single orbit turns the spin by $2\pi a_e = 7.29 \times 10^{-3}$ radians, far too small to measure against the analysing power of a Mott polarimeter; after $N = 1 \times 10^5$ orbits it is $\theta = 728.6$ rad, which is $a_e N = 116$ full turns — a hundred and more — and the asymmetry oscillates that many times as the dwell time is swept, so the *frequency* of that oscillation — which is ν_a — is what is fitted. The measurement has become a frequency measurement again, and the magnetic field appears only through ν_c, which is measured in the same apparatus. Wilkinson and Crane reached a fractional uncertainty of a few parts in 10^5 on a_e [1268] — their $a_e = 1159.622(27) \times 10^{-6}$ is 2.3×10^{-5} of itself, which on the customary $g/2$ scale is the 2.7×10^{-8} more often quoted. That is enough to test the two-loop coefficient C_2 of Equation (112.14), whose term is 1.5×10^{-3} of a_e and therefore about sixty-five times the uncertainty, and not enough to reach the three-loop term, which is 1.3×10^{-5} of a_e and sits half a standard deviation away.

What limited it. Three things, all structural. The electrons are unconfined radially over a large volume, so B in Equation (112.11) is an average over an orbit whose extent is not precisely known; the trapping bottle needed to hold them is itself a field inhomogeneity; and the polarimetry is destructive and statistically inefficient, so each electron yields one bit and the ensemble must be re-prepared. All three are removed at once by confining a *single* electron in a small, well-characterized volume and reading its state nondestructively — which is the Penning trap.

112.3.2 The Penning trap

A Penning trap confines a charged particle using a uniform magnetic field $\boldsymbol{B} = B\hat{\boldsymbol{z}}$ for radial confinement and a static quadrupole electrostatic potential for axial confinement. A magnetic field alone cannot confine along $\hat{\boldsymbol{z}}$, and by Earnshaw's theorem no static electric field can confine in all three directions; the combination works because the electric field need only confine in one direction while defeating the radial anti-confinement that Laplace's equation forces on it.

The fields. The electrode surfaces are shaped to produce, near the centre,

$$\Phi(\rho, z) = \frac{V_0}{2d^2}\left(z^2 - \frac{\rho^2}{2}\right), \qquad d^2 = \frac{1}{2}\left(z_0^2 + \frac{\rho_0^2}{2}\right), \tag{112.27}$$

where V_0 is the signed potential difference applied between the ring and the endcaps, z_0 is the axial half-spacing and ρ_0 the inner radius; d is the characteristic trap dimension, with SI units of metres. Equation (112.27) satisfies $\nabla^2\Phi = 0$, as it must, and the minus sign in front of $\rho^2/2$ is the price: the potential that confines axially pushes the particle out radially. For a particle of charge $-e$ the axial potential energy is $-e\Phi = -eV_0 z^2/2d^2$ on the axis, so trapping requires $V_0 < 0$; the trap depth is $|V_0|$ and only that magnitude appears in the mode frequencies below.

Theorem 112.8 (The three modes of a Penning trap)**.** *An electron of charge $-e$ and mass m_e in the fields of Equation (112.27) and $\boldsymbol{B} = B\hat{\boldsymbol{z}}$ executes three independent harmonic motions, of angular frequencies*

$$\omega_z = \sqrt{\frac{e\,|V_0|}{m_e d^2}}, \qquad \omega_\pm = \frac{\omega_c}{2} \pm \sqrt{\frac{\omega_c^2}{4} - \frac{\omega_z^2}{2}}, \qquad \omega_c = \frac{eB}{m_e}, \tag{112.28}$$

provided $\omega_c^2 > 2\omega_z^2$. They satisfy

$$\omega_+ + \omega_- = \omega_c, \qquad \omega_+\omega_- = \frac{\omega_z^2}{2}. \tag{112.29}$$

Rests on Equation (112.27).

Proof. *Derives Theorem 112.8.*

The equation of motion is $m_e\ddot{\boldsymbol{r}} = -e\,(-\nabla\Phi) - e\,\dot{\boldsymbol{r}} \times \boldsymbol{B}$, with $-e$ the electron's charge. The axial component decouples:

$$m_e\ddot{z} = e\,\partial_z\Phi = \frac{eV_0}{d^2}\,z = -\frac{e\,|V_0|}{d^2}\,z,$$

the last step using $V_0 < 0$, which is the polarity that traps a negative charge. This is $\ddot{z} = -\omega_z^2 z$ with $\omega_z^2 = e\,|V_0|\,/(m_e d^2)$, manifestly positive, and it is the first of Equation (112.28); its SI check is $[e\,|V_0|\,/m_e d^2] = \mathrm{J}/(\mathrm{kg\,m^2}) = 1/\mathrm{s^2}$, as required.

For the radial motion write $u = x + \mathrm{i}y$. The two transverse components combine into

$$\ddot{u} + \mathrm{i}\omega_c\dot{u} - \frac{\omega_z^2}{2}u = 0,$$

where the $\mathrm{i}\omega_c\dot{u}$ term is the magnetic force and the last term is the radial defocusing of Equation (112.27). Substituting $u = u_0 \mathrm{e}^{-\mathrm{i}\omega t}$ gives the characteristic equation $\omega^2 - \omega_c\omega + \omega_z^2/2 = 0$, whose roots are the $\omega_\pm$ of Equation (112.28). They are real — that is, the motion is bounded — exactly when the discriminant is non-negative, $\omega_c^2 > 2\omega_z^2$. Vieta's formulas for that quadratic are Equation (112.29). □

The fast root ω_+ is the *trap-modified cyclotron* frequency, the slow root ω_- the *magnetron* frequency; expanding Equation (112.28) for $\omega_z \ll \omega_c$,

$$\omega_- \simeq \frac{\omega_z^2}{2\omega_c}\,, \qquad \omega_+ \simeq \omega_c - \omega_-\,. \tag{112.30}$$

The magnetron motion is a slow $\boldsymbol{E}\times\boldsymbol{B}$ drift of the orbit centre around the trap axis. It is a metastable mode: its energy *decreases* as its radius grows, so damping it would expel the electron, and it must instead be driven inwards by sideband coupling to the axial motion. In the Harvard trap the three frequencies are separated by six orders of magnitude, which is what makes them independently addressable. With $B = 5.32\,\mathrm{T}$ and $\nu_z = 200\,\mathrm{MHz}$,

$$\nu_c = \frac{eB}{2\pi m_e} = 148.9\,\mathrm{GHz}\,, \qquad \nu_z = 200\,\mathrm{MHz}\,, \qquad \nu_- = \frac{\nu_z^2}{2\nu_c} = 134\,\mathrm{kHz}\,, \tag{112.31}$$

and the anomaly frequency to be measured is $\nu_a = a_e\nu_c = 172.7\,\mathrm{MHz}$, close to ν_z. That proximity is a design choice and not a coincidence, but the dependence runs one way only: $\nu_a = a_e eB/2\pi m_e$ is fixed by B alone, and V_0 is then chosen so that $\nu_z = \sqrt{e\,|V_0|\,/m_e d^2}/2\pi$ falls near the ν_a that the field has already determined. Only ν_z is set by V_0.

The invariance theorem. A real trap is not the ideal one of Equation (112.27): its magnetic and electric axes are not exactly aligned, and the quadrupole has a small ellipticity. Both perturb ω_+, ω_z and ω_- individually, at a level far above the target precision. What rescues the measurement is that one particular combination is untouched.

Theorem 112.9 (Invariance theorem)**.** *The free-space cyclotron frequency is recovered from the three measured trap frequencies as*

$$\omega_c^2 = \omega_+^2 + \omega_z^2 + \omega_-^2\,, \tag{112.32}$$

and this relation continues to hold, to the order at which it matters, when the trap axis is tilted with respect to $\boldsymbol{B}$ and when the electrostatic quadrupole is elliptical rather than axially symmetric [1769].

Rests on Equation (112.29)*.*

Proof. *Derives Theorem 112.9.*

For the ideal trap the identity is algebraic. By Equation (112.29), $\omega_z^2 = 2\omega_+\omega_-$, so

$$\omega_+^2 + \omega_z^2 + \omega_-^2 = \omega_+^2 + 2\omega_+\omega_- + \omega_-^2 = (\omega_+ + \omega_-)^2 = \omega_c^2\,,$$

the last step being the first of Equation (112.29). That the sum of squares survives a tilt and an ellipticity, whereas each frequency separately does not, is the content of Brown and Gabrielse's analysis, which diagonalizes the quadratic form of the perturbed trap and shows that the trace of the relevant matrix — which is what the sum of squares computes — is fixed by B alone. The derivation of that general case is not reproduced here. □

Derivation pending #197

The invariance theorem for a tilted and elliptical Penning trap: the general Brown–Gabrielse result, that the sum of squares of the three measured mode frequencies equals the squared free-space cyclotron frequency independently of the tilt angle and the ellipticity parameter. Belongs in Appendix A beside the ideal-trap identity proved in this chapter.

Numerically the theorem matters less than it might appear, because in the ideal trap $\omega_+\omega_- = \omega_z^2/2$ lets Equation (112.32) be rewritten as $\omega_c = \omega_+ + \omega_z^2/(2\omega_+)$ with no approximation at all; what the general theorem buys is that the same combination survives a tilt and an ellipticity. The correction term $\nu_z^2/(2\nu_+)$ is 134 kHz out of 148.9 GHz, a fractional shift of 9×10^{-7} — a million times the target precision, so it must be applied, but it depends only on ν_z, which is measured continuously.

The continuous Stern–Gerlach effect. The remaining problem is reading the spin. A single electron emits no light one can collect, and a Stern–Gerlach deflection of a charged particle is defeated by the Lorentz force. Dehmelt's solution is to make the *axial frequency* depend on the spin state, by superimposing on the uniform field a small magnetic bottle

$$\Delta\boldsymbol{B} = B_2\left[\left(z^2 - \frac{\rho^2}{2}\right)\hat{\boldsymbol{z}} - z\rho\,\hat{\boldsymbol{\rho}}\right], \tag{112.33}$$

produced by a ring of ferromagnetic material (nickel) in the trap electrodes. In the Harvard trap $B_2 \approx 1540\,\mathrm{T/m^2}$.

Proposition 112.10 (A spin flip shifts the axial frequency)**.** *In the field of Equation* (112.33)*, a change $\Delta\mu$ in the z-component of the electron's magnetic moment shifts the axial frequency by*

$$\Delta\nu_z = \frac{\Delta\mu\, B_2}{4\pi^2 m_e \nu_z}. \tag{112.34}$$

Both a spin flip and a one-quantum cyclotron excitation change the moment by $\Delta\mu = 2\mu_\mathrm{B}$, and both therefore produce the same shift, $\Delta\nu_z \approx 4\,\mathrm{Hz}$ out of 200 MHz.

Rests on Equations (112.11) *and* (112.33)*.*

Proof. *Derives Proposition 112.10.*

The energy of a magnetic moment μ in the bottle is $U = -\mu\,\Delta B_z = -\mu B_2 z^2$ on the axis, which adds to the axial potential energy $\tfrac12 m_e\omega_z^2 z^2$ a term quadratic in z. The effective spring constant is therefore $k = m_e\omega_z^2 - 2\mu B_2$, and since $\omega_z = \sqrt{k/m_e}$,

$$|\Delta\omega_z| = \left|\frac{\partial\omega_z}{\partial k}\,\Delta k\right| = \frac{1}{2m_e\omega_z}\,|2\,\Delta\mu\, B_2| = \frac{|\Delta\mu|\ B_2}{m_e\omega_z},$$

which is Equation (112.34) after $\omega_z = 2\pi\nu_z$. The sign is fixed by the same expression: a moment aligned so as to lower its energy at large $|z|$ softens the axial spring and lowers ν_z, and the two spin states therefore sit either side of the unperturbed frequency.

For the size of $\Delta\mu$: the cyclotron energy of Equation (112.11) is $E_n = (n+\tfrac12)\hbar eB/m_e$, so the magnetic moment carried by the cyclotron motion is

$$\mu_n = -\frac{\partial E_n}{\partial B} = -\left(n+\tfrac12\right)\frac{e\hbar}{m_e} = -\left(n+\tfrac12\right)2\mu_\mathrm{B},$$

so one cyclotron quantum is worth $2\mu_\mathrm{B}$; and a spin flip changes the spin moment by $g\mu_\mathrm{B} \approx 2\mu_\mathrm{B}$ by Equation (112.3). Inserting $\Delta\mu = 2\mu_\mathrm{B} = 1.8548\times10^{-23}\,\mathrm{J/T}$, $B_2 = 1540\,\mathrm{T/m^2}$, $m_e = 9.109\,383\,713\,9\times10^{-31}\,\mathrm{kg}$ and $\nu_z = 200\,\mathrm{MHz}$,

$$\Delta\nu_z = \frac{1.8548\times10^{-23}\,\mathrm{J/T}\times 1540\,\mathrm{T/m^2}}{4\pi^2\times 9.109\,383\,713\,9\times10^{-31}\,\mathrm{kg}\times 2\times10^{8}\,\mathrm{Hz}} = 3.97\,\mathrm{Hz}.$$

□

Four hertz in two hundred megahertz is two parts in 10^8, and the axial motion can be tracked to far better than that, so a single spin flip of a single electron is a resolved, nondestructive, repeatable observation. The electron is not destroyed by the measurement

and can be interrogated for months. That is what the phrase "quantum jump spectroscopy" in Section 112.4 means.

The same apparatus loads a positron as readily as an electron, and Van Dyck, Schwinberg and Dehmelt compared the two g factors in the same trap, in the same field, with the same readout [1798]. Because everything common to the two measurements cancels in the ratio, the bound on the *difference* is far sharper than the uncertainty on either value — the same structural argument as Equation (112.6). The result and its standing as a CPT test are in Section 112.5.2.

112.3.3 The cylindrical trap and the one-electron quantum cyclotron

The Harvard apparatus that produced Equation (112.8) is a Penning trap with three further properties, each of which removes one item that would otherwise dominate the error budget.

Cylindrical electrodes, so the cavity is calculable. The classical Penning trap uses hyperboloidal electrodes, which generate Equation (112.27) exactly but whose enclosed volume is a cavity whose electromagnetic modes cannot be written down. The Harvard trap instead uses a stack of right circular cylinders — a ring, two endcaps and two compensation electrodes, gold-plated, of inner radius about 4.6 mm and comparable total length — with the compensation electrodes tuned so that the leading anharmonic term of the potential vanishes and Equation (112.27) is recovered near the centre. The price is a slightly less perfect potential; the prize is that the enclosed volume is a right circular cylindrical cavity, whose TE and TM modes are Bessel functions with analytically known frequencies. The cavity shift of the cyclotron frequency can therefore be *calculated* and corrected, rather than merely bounded [2101].

That shift is not small on the scale of the measurement. The cyclotron motion is a harmonic oscillator coupled to the modes of its enclosure, and the coupling both damps it and shifts its frequency. The damping is easily estimated. In free space a cyclotron orbit radiates at the Larmor rate, giving a damping constant

$$\gamma_c = \frac{4}{3}\,\frac{r_e\,\omega_c^2}{c}\,, \qquad r_e = \frac{e^2}{4\pi\varepsilon_0 m_e c^2} = 2.8179\times10^{-15}\,\mathrm{m}\,, \tag{112.35}$$

where r_e is the classical electron radius, itself built from the same combination $e^2/4\pi\varepsilon_0$ that appears in $\alpha = e^2/(4\pi\varepsilon_0\hbar c)$.

Derivation. A nonrelativistic charge of acceleration a radiates at the Larmor power $P = e^2a^2/(6\pi\varepsilon_0 c^3)$, in watts. On a circular orbit of angular frequency ω_c and speed v the acceleration is $a = \omega_c v$, and the kinetic energy is $E = \tfrac{1}{2}m_e v^2$, so $v^2 = 2E/m_e$ and

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -P = -\frac{e^2\omega_c^2 v^2}{6\pi\varepsilon_0 c^3} = -\frac{e^2\omega_c^2}{3\pi\varepsilon_0 m_e c^3}\,E =: -\gamma_c E\,.$$

The energy therefore decays exponentially at the rate γ_c, and writing $e^2/4\pi\varepsilon_0 = r_e m_e c^2$ turns that rate into $\gamma_c = 4r_e m_e c^2\omega_c^2/(3m_e c^3) = \tfrac{4}{3}r_e\omega_c^2/c$, which is Equation (112.35). Its SI dimension is $\mathrm{m}\times 1/\mathrm{s}^2\times\mathrm{s/m} = 1/\mathrm{s}$, as a rate must be. The argument is classical, but the rate is also the spontaneous-emission rate of the $n = 1$ Landau level, because for a harmonic oscillator the classical energy-damping constant and the one-quantum decay rate coincide. □

With $\omega_c = 2\pi\times 148.9\,\mathrm{GHz}$ this gives $\gamma_c = 11.0/\mathrm{s}$, a free-space lifetime of 0.091 s for the first excited cyclotron state. Inside the cavity, when ν_c is detuned from every mode, the density of states into which the electron can radiate is suppressed and the lifetime is lengthened by an order of magnitude. This is inhibited spontaneous emission — the

counterpart of Purcell's cavity enhancement [893], observed directly for a Rydberg atom between conducting plates [1755] — and it is used here as an instrument rather than studied. Its general treatment belongs to the cavity-QED section of Chapter 122, which reserves the heading but does not yet carry the derivation. A cyclotron excited state that survives for seconds can be observed by watching the axial frequency, which is the whole measurement.

The frequency shift is the dispersive counterpart of the same coupling and is of order a part in 10^9 of ν_c, i.e. of order 10^{-6} of a_e: enormous compared with the target. Because the mode structure of a cylinder is known, the shift can be mapped by measuring the cyclotron damping rate as a function of B — the rate peaks where ν_c crosses a mode — and the measurement is then made at fields chosen to sit between modes. The cavity shift and its correction remain the largest single systematic of the experiment; this book does not reproduce the published line-by-line budget, which is in [2126, 2419].

A dilution refrigerator, so the cyclotron motion is in its ground state. The trap is operated at about 100 mK in a dilution refrigerator, with the electrodes, the vacuum enclosure and the detection amplifier all at that temperature.

Proposition 112.11 (The electron is in the cyclotron ground state)**.** *At* $B = 5.32\,\mathrm{T}$ *and* $T = 100\,\mathrm{mK}$ *the mean thermal occupation of the cyclotron oscillator is* $\bar{n} \approx 9 \times 10^{-32}$*.*

Rests on Equations (112.11) *and* (112.35)*.*

Proof. *Derives Proposition 112.11.*

The level spacing of Equation (112.11) is

$$\hbar\omega_c = h\nu_c = 6.626\,070\,15\times10^{-34}\,\mathrm{J\,s}\times1.4892\times10^{11}\,\mathrm{Hz} = 9.8675\times10^{-23}\,\mathrm{J} = 6.159\times10^{-4}\,\mathrm{eV}\,,$$

using the exact SI value of the Planck constant [2431]. In temperature units this is $h\nu_c/k_\mathrm{B} = 7.147\,\mathrm{K}$, with $k_\mathrm{B} = 1.380\,649\times10^{-23}\,\mathrm{J/K}$ exact. The Bose–Einstein occupation at $T = 0.100\,\mathrm{K}$ is

$$\bar{n} = \frac{1}{\mathrm{e}^{h\nu_c/k_\mathrm{B}T} - 1} \simeq \mathrm{e}^{-71.47} = 9\times10^{-32}\,.$$

□

The consequence is that the electron sits in $n = 0$ and stays there: a thermally excited cyclotron quantum appears roughly once in 10^{31} attempts, so any observed excitation was driven. This is what converts a classical oscillator into a two-level system that can be addressed one quantum at a time, and it is the difference between the 100 mK apparatus and the 4 K traps that preceded it. The same Bose–Einstein formula at $T = 4.0\,\mathrm{K}$ gives $\bar{n} = 1/(\mathrm{e}^{7.147/4.0} - 1) = 0.201$, and at $T = 4.2\,\mathrm{K}$ it gives 0.223: about one excitation in five, so the oscillator spends a fifth of its time out of the ground state and the cyclotron frequency is smeared by the thermally fluctuating relativistic mass shift. Two tenths is not a small number when the shift per quantum is $\delta = 179\,\mathrm{Hz}$, and it is the factor 2×10^{30} between 0.201 and 9×10^{-32} — the exponential $\mathrm{e}^{-7.147/T}$ read at the two temperatures — that turns the oscillator into a two-level system.

The same cryogenics supply the vacuum. The trap can is sealed at room temperature and cooled; every residual gas except helium freezes onto the walls, and the pressure falls below about $1\times10^{-14}\,\mathrm{Pa}$. At that pressure the collision rate with background gas is such that a single electron has been held and interrogated continuously for months, which is what makes an experiment consisting of hundreds of thousands of individual quantum jumps possible at all.

A self-excited oscillator, so the axial frequency is read continuously. The axial motion induces an image current in the endcap electrodes, which is amplified and, in the Harvard implementation, fed back to the electrode with a controlled phase and an amplitude-limiting gain. The electron's axial motion then self-oscillates at its own frequency, at a large and stable amplitude, and ν_z is read out as the frequency of a strong signal rather than inferred from a weak damped ring-down. The signal-to-noise gain is what allows the 4 Hz shift of Proposition 112.10 to be detected in a fraction of a second, and it is one of the refinements separating [2126] from [2101].

The 2023 measurement [2419] adds no new principle to this list. Its improvement by a factor of 2.2 over 2008 comes from better thermal isolation and control of the trap electrodes, longer averaging, a more thorough mapping of the cavity modes, and the accumulated understanding of an apparatus that had by then been operated for two decades. That is worth saying plainly: the most precise measurement in physics was improved not by a new idea but by twenty years of attention to the same one.

112.4 Procedure

Loading exactly one electron. Electrons are produced by a field-emission point held at the trap electrode potential, or by driving a sharp tungsten tip until it emits; a burst of them is captured by pulsing the endcap potential. The resulting cloud contains an unknown number N of electrons, and the next step is to reduce it to one. The image current driven in the detection circuit by the axial motion is proportional to N, so the amplitude of the axial resonance signal falls in *discrete, equal steps* as the trap depth is lowered and electrons are ejected one at a time. The last step before the signal vanishes identifies $N = 1$ unambiguously: this is a counting measurement, not an inference from a calibrated amplitude, and it is one of the reasons the experiment can claim to be about a single particle.

Cooling the three modes. Each mode is brought under control by a different mechanism, and all three must be cold because each couples back to the frequencies being measured through the magnetic bottle of Equation (112.33).

- *Cyclotron.* Damped radiatively into the cavity at the rate of Equation (112.35), as modified by the mode structure. Since the cavity walls are at 100 mK, the motion thermalizes to the ground state by Proposition 112.11, in a few seconds.

- *Axial.* Damped resistively by the image current it drives through the detection resistor, which is also at 100 mK; the damping time is of order 0.2 s. This is also the readout channel.

- *Magnetron.* Cannot be damped — see the discussion after Equation (112.30) — and is instead cooled by driving the axial sideband at $\nu_z + \nu_-$, which couples the two motions so that axial damping removes magnetron action, shrinking the magnetron orbit towards the trap axis.

Quantum jump spectroscopy. The electron sits in the ground state $\left|n = 0, m = -\tfrac{1}{2}\right\rangle$, with m labelling the two spin projections and the level scheme as in Equation (112.37). Two drives are applied and the axial frequency is watched for the 4 Hz step of Proposition 112.10:

- a microwave drive near $\bar{\nu}_+ = 148.9$ GHz, injected through a waveguide, which raises the cyclotron quantum number by one at fixed m;

- an anomaly drive near $\bar{\nu}_a = 172.7\,\mathrm{MHz}$, applied as an oscillating potential on a trap electrode, which drives the electron axially through the bottle field of Equation (112.33). In the electron's frame this is an oscillating transverse magnetic field at $\bar{\nu}_a$, and it drives the simultaneous transition $\left|1,-\tfrac{1}{2}\right\rangle \to \left|0,+\tfrac{1}{2}\right\rangle$ — a spin flip accompanied by the loss of one cyclotron quantum.

A spin flip is the signature, because it alone leaves the axial frequency permanently shifted while the cyclotron excitation decays back. The cyclotron line entering the denominator of Equation (112.36) is the one measured with the spin in the upper state, $\left|0,+\tfrac{1}{2}\right\rangle \to \left|1,+\tfrac{1}{2}\right\rangle$; that choice is what fixes the coefficient of the relativistic correction there at $\tfrac{3}{2}\delta$ rather than $\tfrac{1}{2}\delta$, as the proof of Proposition 112.12 shows. Each attempt yields one bit: jump or no jump. The drive frequency is stepped, several hundred attempts are made at each step, and the fraction that produce a jump traces the resonance line. Because the electron is not consumed, the same electron supplies every point of every line, which removes the particle-to-particle systematics that limited Section 112.3.1.

The lineshape, which is the dominant modelling input. The magnetic bottle that makes the readout possible also broadens the lines. The axial motion samples the bottle, so the field the cyclotron and spin motions experience is $B + B_2\langle z^2\rangle$, and $\langle z^2\rangle$ fluctuates with the axial energy. The result is a resonance whose width is set by the axial temperature and whose shape is *asymmetric*, with a tail on the high-frequency side, because $\langle z^2\rangle \geq 0$ can only raise the field. The line centre is therefore not the line maximum, and extracting it requires the lineshape model of [1769], in which the frequency executes a random walk driven by the axial energy's thermal fluctuations. Two limits of that model are used and their difference is taken as a systematic uncertainty. This is the single largest piece of theory entering the *experimental* number and the honest place to look for an unrecognized error.

Interleaving, so that field drift cancels. The superconducting solenoid drifts. Its field is regulated and the apparatus is temperature-stabilized, but a drift of a part in 10^9 over an hour would still swamp the target. The two drives are therefore interleaved on a timescale of minutes, so that $\bar{\nu}_a$ and $\bar{\nu}_+$ are measured in what is effectively the same field, and the ratio of Equation (112.10) is formed from contemporaneous data. A residual linear drift is fitted and removed; the check is that the extracted a_e is independent of the interleaving period.

The working formula. What is measured is not ν_s and ν_c but $\bar{\nu}_a$, $\bar{\nu}_+$ and $\bar{\nu}_z$, and the free-space ν_c has to be reconstructed from them.

Proposition 112.12 (The measured anomaly)**.** *In terms of the three measured trap frequencies,*

$$\frac{g}{2} = 1 + \frac{\bar{\nu}_a - \bar{\nu}_z^2/2\bar{\nu}_+}{\bar{\nu}_+ + \tfrac{3}{2}\delta + \bar{\nu}_z^2/2\bar{\nu}_+ + \Delta\nu_{\mathrm{cav}}}\,, \qquad \delta := \frac{h\nu_c}{m_e c^2}\,\nu_c\,, \tag{112.36}$$

where δ is the relativistic shift per cyclotron quantum and $\Delta\nu_{\mathrm{cav}}$ the cavity shift of Section 112.3.3.

Rests on Theorem 112.9 and Equations (112.10) and (112.29).

Proof. *Derives Proposition 112.12.*

Start from Equation (112.10), $g/2 = 1 + \nu_a/\nu_c$, and express both free-space frequencies through measured ones.

For the denominator, the invariance theorem Equation (112.32) together with $\nu_- \ll \nu_z \ll \nu_+$ gives

$$\nu_c = \sqrt{\bar{\nu}_+^2 + \bar{\nu}_z^2 + \bar{\nu}_-^2} = \bar{\nu}_+ + \frac{\bar{\nu}_z^2}{2\bar{\nu}_+}\,.$$

This is an identity, not a truncation, and it is worth saying why, because expanding the square root suggests otherwise. The second of Equation (112.29) gives $\bar{\nu}_- = \bar{\nu}_z^2/2\bar{\nu}_+$ exactly, so

$$\bar{\nu}_+^2 + \bar{\nu}_z^2 + \bar{\nu}_-^2 = \bar{\nu}_+^2 + 2\bar{\nu}_+\bar{\nu}_- + \bar{\nu}_-^2 = \left(\bar{\nu}_+ + \frac{\bar{\nu}_z^2}{2\bar{\nu}_+}\right)^2 ,$$

and the $-\frac{1}{8}\bar{\nu}_z^4/\bar{\nu}_+^3$ that the binomial expansion produces is cancelled term for term by $\bar{\nu}_-^2/2\bar{\nu}_+$; at the operating point of Equation (112.31) the two are $\mp 6.06 \times 10^{-2}\,\mathrm{Hz}$ and their sum is zero to machine precision. What is actually neglected is not a power of $\bar{\nu}_z/\bar{\nu}_+$ but the departure of the measured $\bar{\nu}_-$ from $\bar{\nu}_z^2/2\bar{\nu}_+$ in a tilted and elliptical trap, which is second order in the misalignment and is the content of Theorem 112.9.

Next the relativistic correction. The cyclotron levels are not exactly equally spaced, because the electron's mass grows with its energy. Expanding $E = \sqrt{(pc)^2 + (m_e c^2)^2}$ to first order beyond the nonrelativistic term and evaluating in the Landau basis gives the level scheme [1769]

$$\frac{E(n,m)}{h} = \left(n + \tfrac{1}{2}\right)\nu_c + \frac{g}{2}\,m\,\nu_c - \frac{\delta}{2}\left(n + \tfrac{1}{2} + m\right)^2 , \tag{112.37}$$

with $m = \pm\frac{1}{2}$ the spin projection and $\delta/\nu_c = h\nu_c/m_e c^2$ the ratio of one cyclotron quantum to the electron rest energy. Everything relativistic sits in the last term, which is quadratic in the total excitation $n + \frac{1}{2} + m$ because the leading correction to $E = \sqrt{(pc)^2 + (m_e c^2)^2}$ is $-p^4/8m_e^3c^2$, that is, minus the square of the nonrelativistic energy divided by $2m_e c^2$.

Evaluate it on the two transitions actually driven. The cyclotron line used for the denominator connects $(0, +\frac{1}{2}) \to (1, +\frac{1}{2})$, so the bracket $(n + \frac{1}{2} + m)^2$ runs from 1^2 to 2^2 and

$$\bar{\nu}_+\big|_{\mathrm{rel}} = \nu_c - \frac{\delta}{2}\left(2^2 - 1^2\right) = \nu_c - \frac{3}{2}\delta\,.$$

The anomaly drive connects $(1, -\frac{1}{2}) \to (0, +\frac{1}{2})$, for which the bracket takes the value 1^2 at *both* ends, so the relativistic term cancels identically and the anomaly transition is unshifted — which is the reason the correction appears in the denominator of Equation (112.36) and not in the numerator. Adding $\frac{3}{2}\delta$ back to $\bar{\nu}_+$ in the denominator and leaving the numerator alone therefore reconstructs ν_c and ν_a respectively, and finally $\nu_a = \bar{\nu}_a + \bar{\nu}_+ - \nu_c = \bar{\nu}_a - \bar{\nu}_z^2/2\bar{\nu}_+$ because the anomaly drive is referred to the trap cyclotron frequency, not the free-space one. Collecting the pieces and adding the cavity shift as an explicit correction to ν_c gives Equation (112.36). □

The sizes are worth recording, because they show which corrections matter. At $B = 5.32\,\mathrm{T}$, $h\nu_c/m_e c^2 = 1.21 \times 10^{-9}$, so $\delta = 179\,\mathrm{Hz}$ and $\frac{3}{2}\delta = 269\,\mathrm{Hz}$, against $\bar{\nu}_+ = 148.9\,\mathrm{GHz}$: a fractional correction of 1.8×10^{-9}, which is 1.6×10^{-6} of a_e and four orders of magnitude above the target. The magnetron term $\bar{\nu}_z^2/2\bar{\nu}_+ = 134\,\mathrm{kHz}$ is larger still. Neither is a source of uncertainty, because both are computed from measured quantities; the cavity shift is, because it is computed from a model of the cavity.

Blind analysis. An unknown offset, drawn from a distribution wide compared with the expected uncertainty and unknown to the analysts, is added in software to the extracted $\bar{\nu}_a$. All lineshape choices, cuts, drift models and systematic corrections are fixed with the offset in place; only then is it removed and the value revealed. This protects against the specific failure mode of a measurement whose expected answer is known to ten digits in advance, which is otherwise very hard to guard against and which the 2008 and 2023 analyses both took seriously [2126, 2419].

112.5 Observations and data

112.5.1 The measured electron anomaly

Table 112.1: Determinations of the electron magnetic moment anomaly $a_e = (g-2)/2$. The first two entries are historical and are quoted to the precision their methods supported. The last three come from the same Harvard apparatus and are *not* independent: each supersedes its predecessor, and the systematic analysis is cumulative. The final row is not a measurement but the CODATA 2022 least-squares adjustment, which combines the measurements with the QED series and with the recoil determinations of α. "Fractional" is the standard uncertainty divided by a_e itself; note that the experimental papers customarily quote instead the uncertainty on $g/2$, which is smaller by the factor $a_e/(1+a_e) \approx 1/863$.

year	method	$a_e \times 10^3$	fractional	source
1948	atomic beam resonance	≈ 1.2	$\sim 10^{-2}$	[912, 916]
1963	free-electron precession	1.159622(27)	2.3×10^{-5}	[1268]
2006	one-electron quantum cyclotron	1.15965218085(76)	6.6×10^{-10}	[2101]
2008	one-electron quantum cyclotron	1.15965218073(28)	2.4×10^{-10}	[2126]
2023	one-electron quantum cyclotron	1.15965218059(13)	1.1×10^{-10}	[2419]
2022	CODATA adjustment	1.15965218046(18)	1.6×10^{-10}	[2431]

Table 112.1 is the chapter's primary data, and three features of it need comment.

Two different fractional precisions are in circulation. The experimental papers quote parts per trillion *on* $g/2$; this book quotes the fractional uncertainty *on* a_e, because that is what is compared with a calculation. The two differ by the factor $(g/2)/a_e = (1+a_e)/a_e = 863$. So [2419]'s "0.13 parts per trillion" is 1.3×10^{-13} absolute, which is 1.1×10^{-10} of a_e — about 0.11 parts per billion. Likewise [2126]'s 0.28 parts per trillion is 0.24 parts per billion of a_e, and [2101]'s 0.76 parts per trillion is 0.66 parts per billion. Quoting the smaller number is not wrong, but a reader who compares it against the theoretical uncertainty — which is naturally expressed on a_e — without converting will be out by nearly three orders of magnitude.

The three Harvard values are consistent and correlated. The 2023 value lies 1.4×10^{-13} below the 2008 value, about half of the 2008 standard uncertainty, and 2.6×10^{-13} below the 2006 value, a third of its uncertainty. They are not independent measurements that happen to agree: the same trap, the same solenoid and much of the same systematic analysis are shared, so the correct statement is that the successive refinements did not move the number, which is a weaker but still meaningful check.

The CODATA value is not an independent measurement. The 2022 adjustment value $1.15965218046(18) \times 10^{-3}$ [2431] sits 1.3×10^{-13} below [2419], within the combined uncertainty. It must not be used to test the QED series, because the series is one of the inputs from which it was derived. This is the circularity that Section 112.7 is about.

Table 112.2 makes the structure of the prediction visible, and the count of figures each order buys is worth doing rather than guessing. Comparing each running sum with the measured $1.15965218059 \times 10^{-3}$: one loop is off by 1.5×10^{-3} fractionally, so it agrees to three significant figures; two loops by 1.3×10^{-5}, five figures; three loops by 4.4×10^{-8}, seven figures. Beyond that the terms themselves say where they land. The fourth-order term

Table 112.2: The theoretical value of a_e, term by term, evaluated at the CODATA 2022 fine-structure constant $\alpha^{-1} = 137.035999177(21)$ [2431] with the coefficients of Equation (112.14). The expansion parameter is $\alpha/\pi = 2.322\,819\times10^{-3}$. Every entry is dimensionless. The last column is the running sum, and the comparison with the measured $1.15965218059(13) \times 10^{-3}$ is made in Section 112.7.

order	term	value	running sum
α^1	$C_1(\alpha/\pi)$	$+1.1614097321 \times 10^{-3}$	$1.1614097321 \times 10^{-3}$
α^2	$C_2(\alpha/\pi)^2$	$-1.7723050605 \times 10^{-6}$	$1.1596374270 \times 10^{-3}$
α^3	$C_3(\alpha/\pi)^3$	$+1.4804203632 \times 10^{-8}$	$1.1596522312 \times 10^{-3}$
α^4	$C_4(\alpha/\pi)^4$	$-5.5667966898 \times 10^{-11}$	$1.1596521756 \times 10^{-3}$
α^5	$C_5(\alpha/\pi)^5$	$+4.5555816 \times 10^{-13}$	$1.1596521760 \times 10^{-3}$
—	$a_e^{\mu,\tau}$	$+2.7478 \times 10^{-12}$	$1.1596521788 \times 10^{-3}$
—	a_e^{had}	$+1.6927 \times 10^{-12}$	$1.1596521805 \times 10^{-3}$
—	a_e^{weak}	$+0.0305 \times 10^{-12}$	$1.1596521805 \times 10^{-3}$

is 4.8×10^{-8} of a_e and therefore occupies the eighth figure; the muon-loop and hadronic terms (2.4×10^{-9} and 1.5×10^{-9} of a_e) the ninth; the fifth-order term (3.9×10^{-10}) the tenth, which is where the experiment's own 1.1×10^{-10} now lives; and the electroweak term (2.6×10^{-11}) the eleventh, below it. Notice that $a_e^{\mu,\tau} + a_e^{\text{had}} = 4.44\times10^{-12}$ is *thirty-four times* the experimental uncertainty: the electron anomaly is no longer a pure QED observable, and a comparison that omitted the hadronic term would fail by an amount easily resolved.

Table 112.3: Independent determinations of the fine-structure constant, and the CODATA 2022 recommended value. The two recoil measurements use different atomic species, different interferometric schemes and different atom-to-electron mass ratios, and they disagree by 1.60×10^{-7} in α^{-1}, about 5.5 times their combined standard uncertainty. The last column is the resulting standard uncertainty on the *predicted* a_e, computed as $\delta a_e = \delta\alpha/2\pi$ from Equation (112.25), and is to be compared with the experimental 1.3×10^{-13}.

year	method	α^{-1}	δa_e	source
2018	caesium recoil, atom interferometer	137.035999046(27)	2.29×10^{-13}	[2316]
2020	rubidium recoil, Bloch oscillations	137.035999206(11)	9.32×10^{-14}	[2363]
2022	CODATA adjustment (not independent)	137.035999177(21)	1.78×10^{-13}	[2431]

112.5.2 Positrons and the CPT test

Phenomenon 112.13 (Electron and positron moments agree)**.** The g factors of the electron and of the positron, measured in the same Penning trap with the same field and the same readout, agree to about two parts in 10^{12} [1798]. No difference between the magnetic moment of a lepton and that of its antilepton has ever been observed.

Rests on Theorem 105.37 and Equation (112.36)*.*

Derivation. *Derives Phenomenon 112.13.*

That the two moments must have equal magnitude and opposite sign is not an assumption of this chapter but a theorem, proved in Theorem 105.37 from the transformation of the electromagnetic current operator under the antiunitary CPT operator: the magnetic moment operator $\hat{\boldsymbol{\mu}} = \frac{1}{2}\int \mathrm{d}^3x\, \boldsymbol{x} \times \boldsymbol{j}$ is CPT-even, because the sign the current picks up is cancelled

by the reflection of its spatial argument, while the state of maximal spin projection maps to the antiparticle state of *opposite* spin projection — so the moment per unit spin reverses. The companion statement that the masses are equal is Theorem 105.36. Both rest only on locality, Lorentz invariance and the positivity of the energy spectrum, which is what gives a null result here its unusual weight: a nonzero difference would falsify one of those three.

The experimental value is a ratio, and it is sharper than either absolute measurement for the reason already used at Equation (112.6). Writing g_- and g_+ for the electron and positron g factors measured in the same field B, each is extracted from Equation (112.36) using frequencies measured in that field; forming

$$\frac{g_- - g_+}{g_{\mathrm{avg}}}$$

cancels the cavity shift, the relativistic correction, the lineshape model and the field itself to the extent that they are common to the two species — and they are common, because the two particles differ only in the sign of the charge and therefore occupy the same trap with the same geometry. Van Dyck, Schwinberg and Dehmelt bounded this ratio at about 2×10^{-12} [1798], which is where the entry in Table 105.2 comes from. It remains the sharpest lepton CPT test of this type; the kaon rest-energy comparison of Chapter 111 is often quoted as sharper, but the two numbers are normalized by different quantities and are not directly comparable — a point made carefully in Section 105.5.3. □

112.6 The muon anomaly

The muon anomaly is included in this chapter as a contrast, not as a digression. It is the same observable computed from the same series, measured by an entirely different technique, and its status is the opposite of the electron's: the experiment is nearly two thousand times less precise in fractional terms (2.1×10^{-7} against 1.1×10^{-10}, from Tables 112.1 and 112.4), the theory is far less certain, and the comparison currently does not agree. Setting the two side by side is the clearest way to show what a precision test actually consists of.

112.6.1 CERN and Brookhaven

A muon lives for

$$\tau_\mu = \frac{\hbar}{\Gamma_\mu} = \frac{1.054\,571\,817 \times 10^{-34}\,\mathrm{J\,s}}{4.800 \times 10^{-29}\,\mathrm{J}} = 2.197 \times 10^{-6}\,\mathrm{s}\,, \tag{112.38}$$

the width $\Gamma_\mu = 2.995\,983\,6 \times 10^{-19}\,\mathrm{GeV} = 4.800 \times 10^{-29}\,\mathrm{J}$ being the PDG value [2427]. Two microseconds is not long enough to trap a muon and interrogate it for months, so the whole strategy of Section 112.3.2 is unavailable. Instead the muon is stored relativistically in a ring, and the anomaly is read out exactly as in Proposition 112.7: as an angle that accumulates between the spin and the momentum.

Storage rings at CERN. Three generations of muon storage ring were built at CERN between 1961 and 1977. The final one stored polarized muons of both signs at a momentum of about $3.1\,\mathrm{GeV}/c$ in a ring of 14 m diameter, and reported the anomaly to about seven parts per million — together with a direct measurement of relativistic time dilation, using the same stored muons, which is written up as an experiment in its own right at Section 41.4 [1592, 1632]. The readout exploits the parity violation of the weak interaction (Chapter 110): in the decay $\mu^+ \to e^+ \nu_e \bar{\nu}_\mu$ the highest-energy positrons are emitted preferentially along the muon spin, so counting decay positrons above an energy threshold as a function of time gives a decaying exponential with a superimposed oscillation, and the oscillation frequency is ν_a.

The magic momentum. A ring needs electrostatic quadrupoles for vertical focusing, and an electric field spoils Equation (112.10): a moving magnetic moment in an electric field precesses too. The Bargmann–Michel–Telegdi equation [1165] gives the precession of the spin relative to the momentum, for a particle of charge q and mass m moving with $\boldsymbol{\beta} = \boldsymbol{v}/c$ perpendicular to $\boldsymbol{B}$, as

$$\boldsymbol{\omega}_a = -\frac{q}{m}\left[a\,\boldsymbol{B} - \left(a - \frac{1}{\gamma^2 - 1}\right)\frac{\boldsymbol{\beta}\times\boldsymbol{E}}{c}\right], \tag{112.39}$$

with $\gamma = (1-\beta^2)^{-1/2}$ and a the anomaly of the particle in question. Both terms have SI dimensions of rad/s: qB/m is a cyclotron frequency, and $q\beta E/(mc)$ is the same thing with E/c playing the part of a magnetic field, which is the correct SI statement of the fact that $\boldsymbol{E}/c$ and $\boldsymbol{B}$ are components of one field tensor.

Proposition 112.14 (The magic momentum)**.** *The electric field drops out of Equation* (112.39) *at the Lorentz factor*

$$\gamma_{\text{magic}} = \sqrt{1 + \frac{1}{a_\mu}} = 29.3034\,, \tag{112.40}$$

corresponding to a muon momentum of $p = \gamma_{\text{magic}}\beta\, m_\mu c = 3.094\,\text{GeV}/c = 1.6537 \times 10^{-18}\,\text{kg m/s}$.

Rests on Equation (112.39).

Proof. *Derives Proposition 112.14.*

The bracket in Equation (112.39) loses its electric term when $a_\mu = 1/(\gamma^2-1)$, that is $\gamma^2 = 1 + 1/a_\mu$, which is Equation (112.40); with the CODATA 2022 muon anomaly $a_\mu = (1.165\,920\,62 \pm 0.000\,000\,41) \times 10^{-3}$ [2431] this gives $1 + 1/a_\mu = 858.69$ and $\gamma_{\text{magic}} = 29.3034$, hence $\beta = 0.9994175$. The momentum follows from $p = \gamma\beta m_\mu c$ with the PDG muon rest energy $m_\mu c^2 = (0.105\,658\,375\,5 \pm 0.000\,000\,002\,3)$ GeV [2427]:

$$pc = 29.3034 \times 0.9994175 \times 0.105\,658\,375\,5\,\text{GeV} = 3.0944\,\text{GeV}\,,$$

so $p = 3.0944\,\text{GeV}/c$, which in SI base units is $3.0944 \times 10^9\,\text{eV} \times 1.602\,176\,634 \times 10^{-19}\,\text{J/eV} / 2.997\,924\,58 \times 10^8\,\text{m/s} = 1.6537 \times 10^{-18}\,\text{kg m/s}$. □

At that momentum, with the ring's magnetic field and radius, the kinematics is fixed and can be checked against itself. Brookhaven's E821 — and, later, Fermilab's E989, which reuses the same magnet — has a storage radius $r = 7.112\,\text{m}$ and a field $B = 1.4513\,\text{T}$, and the momentum a ring of that geometry selects is

$$p = eBr = 1.602\,176\,634 \times 10^{-19}\,\text{C} \times 1.4513\,\text{T} \times 7.112\,\text{m} = 1.6537 \times 10^{-18}\,\text{kg m/s}\,, \tag{112.41}$$

which is Proposition 112.14 to five figures. The ring was built to the magic momentum, and the agreement of Equation (112.41) with Proposition 112.14 is a consistency check on the numbers quoted here rather than an independent fact.

Brookhaven's E821 measured a_μ to 0.54 parts per million and reported a value about 2.5 standard deviations above the Standard Model evaluation then current [2091]. That discrepancy is what motivated everything in Sections 112.6.2 and 112.6.3.

112.6.2 Fermilab E989

The E821 storage ring — a 14.2 m diameter superconducting magnet, too large to disassemble — was transported intact from Brookhaven to Fermilab in 2013 and rebuilt there with a far more intense and cleaner muon beam, a completely new detector system, and roughly a threefold improvement in field uniformity. The experimental scales are worth listing, since each is derived above:

- cyclotron period $T_c = 2\pi r/\beta c = 149\,\text{ns}$;
- anomalous precession period $1/\nu_a = 4.37\,\mu\text{s}$, from $\nu_a = a_\mu eB/2\pi m_\mu = 229\,\text{kHz}$, so the spin turns once relative to the momentum every 29 turns of the ring;
- dilated muon lifetime $\gamma\tau_\mu = 29.3034 \times 2.197\,\mu\text{s} = 64.4\,\mu\text{s}$, giving about 15 anomalous precession cycles per lifetime and a usable observation window of several hundred microseconds.

Two frequencies are measured. The first is ω_a, from the oscillation in the decay-positron time spectrum described above. The second is not B but ω_p, the Larmor precession frequency of a free proton in the same field, measured by pulsed nuclear magnetic resonance in water-based probes on a trolley that is driven around the storage region to map the field, cross-calibrated against a spherical-water standard. The anomaly is then extracted as

$$a_\mu = \frac{\omega_a}{\omega_p} \cdot \frac{\mu_p}{\mu_e} \cdot \frac{m_\mu}{m_e} \cdot \frac{|g_e|}{2}\,, \tag{112.42}$$

a product of one measured ratio with three externally known ones, each taken as a magnitude in the sense of Equation (112.4). Note what has happened: the field has again cancelled, and the price is that the muon anomaly depends on the proton-to-electron moment ratio, the muon-to-electron mass ratio and the electron g factor — the last of which is the subject of the rest of this chapter. The two measurements are not as independent as they look.

As at Harvard, the analysis is blind: a frequency offset unknown to the analysis teams is applied to the reference clock in hardware, and the several independent analysis groups agree their result before it is removed. The published values are 0.46 parts per million from the first data set [2368] and 0.20 parts per million from the second and third [2416]; they agree with each other and with Brookhaven.

Table 112.4: The muon anomaly: measurements, the recommended value, and the Standard Model evaluation. The four measurements are independent experiments and agree; the disagreement is between the measurements as a group and the prediction. The last two rows show that the prediction depends on which evaluation of the leading hadronic vacuum polarization is used — the data-driven dispersive one or the lattice one — and that the two do not agree with each other. See the text: the tension is currently between two theoretical evaluations of one hadronic quantity, and no claim of physics beyond the Standard Model follows from it.

year	source of the number	$a_\mu \times 10^3$	reference
1979	CERN muon storage ring	1.165924(9)	[1632]
2006	BNL E821 (0.54 ppm)	1.16592089(63)	[2091]
2021	FNAL E989, first data set	1.16592040(54)	[2368]
2023	FNAL E989, later data sets	1.16592055(24)	[2416]
2022	CODATA recommended	1.16592062(41)	[2431]
2020	Standard Model, data-driven HVP	1.16591810(43)	[2353]
2021	the same, with lattice HVP	$\approx$ 1.16591954(57)	[2375]

112.6.3 Hadronic vacuum polarization: lattice versus data-driven

By Phenomenon 112.4 the hadronic contribution to a_μ is some 4.3×10^4 times that to a_e, which puts it at about 7×10^{-8} — three hundred times the experimental uncertainty of Table 112.4. It is therefore the dominant theoretical uncertainty, and since it cannot be computed perturbatively (Chapter 102) it must be obtained some other way.

The dispersive route. The photon self-energy is an analytic function of $s = (pc)^2$ with a cut along the positive real axis, and the discontinuity across the cut is fixed by the optical theorem to be the total cross section for the photon to make hadrons — which is measured, in e^+e^- annihilation, as the ratio $R(s)$ of Phenomenon 102.15. A dispersion relation then converts the measured cross section into the contribution to the anomaly:

$$a_\ell^{\mathrm{HVP,LO}} = \frac{1}{3}\left(\frac{\alpha}{\pi}\right)^2 \int_{s_{\mathrm{th}}}^{\infty} \frac{\mathrm{d}s}{s}\, R(s)\, K(s)\,, \qquad K(s) \to \frac{m_\ell^2 c^4}{3s} \quad \text{for } s \gg m_\ell^2 c^4\,, \tag{112.43}$$

with $s_{\mathrm{th}} = 4m_\pi^2 c^4$ the two-pion threshold, R and K both dimensionless, and s carrying SI dimensions of J^2. The asymptotic form of the kernel is where the m_ℓ^2 scaling of Equation (112.17) comes from, and the $1/s$ weighting is why the integral is dominated by the lowest resonances: about 73 % of the muon integral comes from the ρ resonance region, that is from $e^+e^- \to \pi^+\pi^-$ below 1 GeV, equivalently 1.602×10^{-10} J. The factor $1/3$ in the asymptotic kernel is the same one that carried the derivation of Equation (112.18), where the known two-loop coefficient $1/45$ fixes it independently.

Derivation pending #198

The dispersive representation of the hadronic vacuum-polarization contribution to a lepton anomaly: analyticity of the photon self-energy, the optical theorem relating its discontinuity to the total hadronic cross section, and the derivation of the kernel $[K(s)]$ from the one-loop vertex with a dressed photon propagator. Belongs in Appendix A.

The two evaluations disagree. Carrying out Equation (112.43) over the compiled world data gives the leading-order hadronic term $a_\mu^{\mathrm{HVP,LO}} = 693.1(4.0) \times 10^{-10}$ [2353], which is the input to the data-driven prediction in Table 112.4. Computing the same quantity directly from the Euclidean correlator on the lattice, without any cross-section data, gives $707.5(5.5) \times 10^{-10}$ [2375]. The difference, $14.4(6.8) \times 10^{-10}$, is about 2.1 standard deviations between two calculations of one number, and it is nearly the whole of the measurement–prediction gap. Substituting the lattice value moves the prediction from the $1.16591810(43) \times 10^{-3}$ of Table 112.4 to about $1.16591954 \times 10^{-3}$, and against the Fermilab measurement $1.16592055(24) \times 10^{-3}$ of that same table [2416] the difference falls from 2.45×10^{-9}, which is 5.0 combined standard deviations, to 1.01×10^{-9}, which is 1.6. No world average is used here: both numbers are rows of Table 112.4, so the arithmetic can be checked against the table.

And the input data are themselves in conflict. The CMD-3 experiment at Novosibirsk measured the $e^+e^- \to \pi^+\pi^-$ cross section from threshold to 1.2 GeV, that is 1.923×10^{-10} J, and obtained a result several per cent above every previous measurement in the dominant ρ region [2426]. Used in Equation (112.43) it would raise the data-driven $a_\mu^{\mathrm{HVP,LO}}$ towards the lattice value and remove much of the discrepancy — but it is inconsistent with the earlier data sets, so it cannot simply be averaged in. The honest description of the present situation is therefore a chain of three disagreements: between two theoretical evaluations of the hadronic term, between two sets of hadronic cross-section measurements, and between the resulting prediction and the muon measurement. Only the last of these would be evidence of anything new, and it cannot be assessed until the first two are settled.

The subleading hadronic light-by-light term, of order 9×10^{-10}, carries an uncertainty of about 2×10^{-10} [2353]: smaller than the vacuum polarization discrepancy, but not negligible, and it is not constrained by any dispersion relation as directly.

112.6.4 Status

Three things would settle the question, and none of them is a new theory.

1. *Independent lattice computations of the same quantity.* The lattice result [2375] is one collaboration's. The sharpest way to test it is not to recompute the whole integral but to compare restricted "window" observables — the correlator integrated over a bounded range of Euclidean time — which are cleaner on the lattice and can also be extracted from the cross-section data. Several groups now do this, and it localizes the disagreement in Euclidean time rather than merely confirming it.

2. *A resolution of the two-pion cross-section conflict.* Either the CMD-3 measurement [2426] or the earlier body of data is wrong, and the discrepancy is large enough that it should be resolvable by further measurements with different systematics.

3. *The complete Fermilab data set.* The measurement in Table 112.4 uses part of the collected data; the final analysis roughly halves the experimental uncertainty again. This sharpens the comparison but cannot resolve it, because the limitation is theoretical.

A fourth route deserves mention because it is orthogonal to all three: measuring the hadronic contribution to the running of α directly in elastic muon–electron scattering, which would give $a_\mu^{\mathrm{HVP,LO}}$ from a single spacelike measurement with systematics unrelated to either the annihilation data or the lattice. A proposal of this kind exists at CERN; no primary reference for it is available in this book's bibliography, and the claim is therefore made without a citation rather than with an invented one.

What must not be said is that the muon anomaly shows physics beyond the Standard Model. Under this book's first editorial rule the tension is reported, not interpreted: the measurements agree with one another, the theoretical evaluations do not agree with one another, and the difference between the two theoretical evaluations is comparable to the difference between theory and experiment. The open problems chapter Chapter 133 collects the questions this book regards as unresolved; the muon anomaly is not among the entries currently written there, and on the reading given here it does not yet qualify — it is an unresolved *calculation*, not an unexplained observation.

112.7 Interpretation

The question this section answers is: how many digits are actually verified? The headline numbers invite the answer "twelve", and that answer is wrong. The comparison has to be run in both directions before it can be stated honestly.

112.7.1 Direction one: theory plus α against the measured a_e

Evaluate Equation (112.13), with the coefficients of Equation (112.14) and the non-QED terms of Equations (112.15) and (112.16), at each of the independently measured values of α in Table 112.3, and subtract the measured a_e of Equation (112.8).

Table 112.5 is the chapter's central result, and it should be read slowly. The same measurement, the same series and the same non-QED corrections give a discrepancy of 3.9 standard deviations with one input for α and 2.1 with the other. Neither is a comfortable agreement; they are on opposite sides of the measurement; and the difference between the two rows is entirely due to the disagreement between two atom-recoil experiments, which is a problem in atomic physics and has nothing whatever to do with QED.

Table 112.5: The test of quantum electrodynamics, run three ways. Each row evaluates the QED series at one determination of α and compares the result with the measured $a_e = 1.15965218059(13) \times 10^{-3}$ [2419]. The combined uncertainty adds in quadrature the experimental one, the propagated uncertainty of the α input from Table 112.3, and about 2×10^{-14} for the truncation of the series and the hadronic term. The verdict of the test depends on which α is used, which is the point of this section. The third row is shown for completeness and is *not* a test: the CODATA adjustment uses a_e and the QED series as inputs, so the comparison is circular.

α input	source	$a_e^{\mathrm{th}} \times 10^3$	$a_e^{\mathrm{exp}} - a_e^{\mathrm{th}}$	deviation
caesium 2018	[2316]	1.1596521816	-1.02×10^{-12}	3.9σ
rubidium 2020	[2363]	1.1596521803	$+3.36 \times 10^{-13}$	2.1σ
CODATA 2022	[2431]	1.1596521805	$+9.05 \times 10^{-14}$	0.4σ (circular)

Phenomenon 112.15 (The test is limited by its input, not by the measurement)**.** The comparison of the measured electron anomaly with its calculated value confirms quantum electrodynamics to a fractional precision of about 1.2×10^{-9} in a_e — roughly nine significant figures — and not to the eleven or twelve figures that the experimental precision alone would allow. The limitation is the independent determination of α [2316, 2363, 2419].

Rests on Equation (112.25).

Derivation. *Derives Phenomenon 112.15.*

The spread in the predicted a_e produced by the two α inputs is $\delta a_e = 1.4 \times 10^{-12}$ by Equation (112.25). Fractionally,

$$\frac{\delta a_e}{a_e} = \frac{1.4 \times 10^{-12}}{1.15965 \times 10^{-3}} = 1.2 \times 10^{-9}\,,$$

which is nine significant figures. The experimental uncertainty is 1.3×10^{-13}, or 1.1×10^{-10} fractionally — eleven figures — and the series truncation and hadronic uncertainty together contribute about 2×10^{-14}, or twelve figures. The largest of the three is the α input, by a factor of ten over the experiment, so it is the α input that sets the verdict. A test is only as sharp as its least well known ingredient, and quoting the sharpest ingredient instead is the specific error this chapter exists to avoid. □

112.7.2 Direction two: the measurement plus theory against α

The same three ingredients can be arranged differently. Take the QED series as correct, insert the measured a_e, and solve Equation (112.13) for α. Inverting the series numerically with the coefficients of Equation (112.14) at the measured value of Equation (112.8) gives

$$\alpha^{-1}\,(a_e) = 137.035999166(15)\,, \tag{112.44}$$

the uncertainty being the experimental 1.3×10^{-13} on a_e propagated back through $\delta\alpha = 2\pi\,\delta a_e$. That is a fractional uncertainty of 1.1×10^{-10}, and it is worth checking that figure against Table 112.3 rather than repeating the claim usually attached to it. The three fractional uncertainties on α^{-1} are $1.5 \times 10^{-8}/137.036 = 1.09 \times 10^{-10}$ here, $2.7 \times 10^{-8}/137.036 = 1.97 \times 10^{-10}$ for the caesium recoil [2316], and $1.1 \times 10^{-8}/137.036 = 8.03 \times 10^{-11}$ for the rubidium one [2363]. So the anomaly route is 1.8 times sharper than caesium and 1.4 times *blunter* than rubidium: it is competitive with the best recoil measurement, not better than it, and Table 112.3 says so already in its δa_e column, where the rubidium entry 9.32×10^{-14} sits below the experimental 1.3×10^{-13} that this paragraph propagates.

It is also not a test of anything. Equation (112.44) assumes the QED series in order to produce α; using it to check the QED series would be circular, and the same is true of the CODATA recommended value, which is an adjustment containing Equation (112.44) as one of its inputs. This is the distinction the chapter has been building towards, and it can now be stated flatly:

- If α is taken from the anomaly itself, the exercise is a *definition*. It produces the best available value of α and tests nothing.
- If α is taken from an independent measurement — and atom recoil is the only method currently competitive — the exercise is a *test*, and its sharpness is bounded by Phenomenon 112.15.

Notice, incidentally, that Equation (112.44) lies between the caesium and rubidium values and is closer to the rubidium one, differing from it by 4.0×10^{-8} and from the caesium one by 1.2×10^{-7}. That is a statement about which recoil experiment is more likely to be in error *if* QED is right; it is not evidence about QED.

112.7.3 What the agreement bounds

Whatever else is true, the electron anomaly agrees with a parameter-free calculation to about 10^{-12} in absolute terms, and that constrains anything not in the calculation.

Electron substructure. If the electron were a bound state of constituents held together at a scale Λ, the anomaly would acquire a contribution of order $(m_e c^2/\Lambda)^2$ from the compositeness. Requiring it to be smaller than the $\delta a_e = 1.4 \times 10^{-12}$ of Equation (112.25),

$$\Lambda \gtrsim \frac{m_e c^2}{\sqrt{\delta a_e}} = \frac{0.510\,998\,950\,69\,\mathrm{MeV}}{\sqrt{1.4 \times 10^{-12}}} = 432\,\mathrm{GeV} = 6.92 \times 10^{-8}\,\mathrm{J}\,, \tag{112.45}$$

using the CODATA electron rest energy [2431]. The electron is pointlike down to a distance $\hbar c/\Lambda \approx 4.6 \times 10^{-19}\,\mathrm{m}$, less than a thousandth of the proton charge radius $(8.4075 \pm 0.0064) \times 10^{-16}\,\mathrm{m}$ [2431].

New particles coupling at loop level. The weaker but more generic bound applies to a state that couples to the electron with roughly electromagnetic strength and therefore contributes only at one loop. Equation (112.20) then gives

$$Mc^2 \gtrsim m_e c^2 \sqrt{\frac{\alpha}{\pi\,\delta a_e}} = 20.8\,\mathrm{GeV} = 3.33 \times 10^{-9}\,\mathrm{J}\,,$$

using $\delta a_e = 1.4 \times 10^{-12}$ again. The muon's sensitivity in the same sense is read off Table 112.4 as the combined uncertainty of the comparison there, $\delta a_\mu = \sqrt{(2.4 \times 10^{-10})^2 + (4.3 \times 10^{-10})^2} = 4.9 \times 10^{-10}$, the two entries being the Fermilab measurement and the data-driven Standard Model evaluation. The same expression with the muon's mass then gives $Mc^2 \gtrsim 230\,\mathrm{GeV} = 3.69 \times 10^{-8}\,\mathrm{J}$. The muon reaches about eleven times higher while being some 350 times *less* sensitive in δa, two and a half orders of magnitude, because the reach grows only as $\delta a^{-1/2}$: the factor 350 costs $\sqrt{350} \approx 19$ in mass, and the mass ratio 206.7683 more than repays it. That is Remark 112.5 made numerical. The two anomalies are not competitors: the electron's is the test of QED, the muon's is the search.

What it does not bound. The agreement says nothing about physics that does not couple to the electron's magnetic moment, and nothing about the parts of the Standard Model that do not enter at this order — the electroweak contribution to a_e is 3×10^{-14}, below the experimental uncertainty, so the electron anomaly does not test the electroweak sector of Chapter 104 at all. It is a test of quantum electrodynamics specifically, and it is the sharpest one there is.

112.7.4 A closing accounting

Set out plainly, the state of the comparison is this. The measurement is secure to 1.3×10^{-13} and has been improved twice by the same group without moving. The QED series is computed to fifth order, contributes at the eleventh figure, and its fifth-order coefficient is disputed by an independent recomputation [2343] at a level that does not currently matter but would if the input improved. The hadronic contribution is thirteen times the experimental uncertainty and is taken from the muon programme's dispersive machinery, so the electron anomaly is no longer purely a QED observable. And the fine-structure constant, the one ingredient that must come from elsewhere, is known from two experiments that disagree by five and a half standard deviations.

The correct summary sentence is therefore not "QED has been verified to twelve significant figures". It is: *the measured and calculated electron anomalies agree to about nine significant figures, which is the precision of the best independent fine-structure constant, and the remaining three figures of experimental precision are waiting for atomic physics to resolve a discrepancy of its own.* That is a less impressive sentence and a true one.

112.8 Primary references

The chapter rests on four groups of sources, and they are of different kinds.

The prediction. Dirac's 1928 paper [596] is where $g = 2$ appears, as a consequence rather than an assumption; the derivation reproduced in this book is at Phenomenon 93.3. Schwinger's 1948 letter [919] is the one-loop anomaly, a page long, and it is the founding calculation of renormalized quantum field theory as a predictive rather than a formal enterprise; the full derivation is Theorem 100.41. For the modern series the standing reference is Aoyama, Kinoshita and Nio's review [2330], which supersedes the same group's earlier tenth-order papers [2190, 2302] and is the source of every coefficient in Equation (112.14) and of the mass-dependent, hadronic and electroweak terms. It should be read together with Volkov's independent five-loop computation [2343], which does not agree with it; citing the review alone would misrepresent the state of the calculation, which is why both appear wherever the fifth-order coefficient is used. Kinoshita's analysis of mass singularities [1232] is background for the logarithms in the mass-dependent terms.

The measurement. Foley and Kusch [912] and Kusch and Foley [916] are the discovery pair; they are cited together throughout because the interpretation and the experimental detail are split between them. The method they used is Rabi's [798, 819]. Wilkinson and Crane [1268] is the best of the free-precession experiments and the bridge to the trap era. For the Penning trap itself, Brown and Gabrielse's review [1769] is the standard reference and the source of the invariance theorem, the lineshape model and the relativistic level scheme of Equation (112.37); it is a review rather than a primary paper, and is cited as such. The Harvard sequence is Odom *et al.* [2101], Hanneke *et al.* [2126] and Fan *et al.* [2419]; the last is the source of Equation (112.8) and the experiment this chapter is registered under. Van Dyck, Schwinberg and Dehmelt [1798] is the electron–positron comparison of

Section 112.5.2. The cavity-QED effect exploited in Section 112.3.3 traces to Purcell [893] and was observed in its inhibitory form by Hulet, Hilfer and Kleppner [1755].

The input. Parker *et al.* [2316] and Morel *et al.* [2363] are the two atom-recoil determinations of α whose disagreement bounds this chapter's conclusion. They are primary papers in atomic physics, not in particle physics, and the reader who wants to judge which is more likely to be right must read them directly; this book takes no position beyond reporting the disagreement and its consequence.

The data sets. Two machine-readable sources carry every constant quoted here. CODATA 2022 [2431] supplies the Bohr magneton, the electron and muon masses, the Rydberg constant, the recommended α, and the recommended electron and muon anomalies; the values used are those in the repository copy of the NIST listing, and every number in Tables 112.1, 112.3 and 112.4 attributed to it has been checked against that file. The PDG 2024 compilation [2427] supplies the muon rest energy and the muon width used in Equation (112.38). Where a value appears in both, CODATA is used for constants of the electron and PDG for properties of the muon as a particle, and the two are consistent within their stated uncertainties.

The muon programme. Bailey *et al.* [1592, 1632] report the CERN storage ring, the second of these being the final CERN account of both the anomaly and the time-dilation measurement written up at Section 41.4. Bennett *et al.* [2091] is the Brookhaven E821 final report, and Abi *et al.* [2368] and Aguillard *et al.* [2416] the first two Fermilab results. The Standard Model evaluation is the Muon $g-2$ Theory Initiative's white paper [2353]; the competing lattice evaluation of the hadronic vacuum polarization is Borsanyi *et al.* [2375]; and the cross-section measurement that unsettles the dispersive input is CMD-3 [2426]. The spin-precession formula behind the magic momentum is Bargmann, Michel and Telegdi [1165].

Remark 112.16 (What is owed). Two derivations are marked pending in this chapter and both belong in Appendix A rather than in the running text: the general form of the invariance theorem for a tilted and elliptical Penning trap, of which only the ideal-trap identity is proved here; and the dispersive representation of the hadronic vacuum-polarization contribution, of which only the structure and the asymptotic kernel are given. Neither affects a number quoted in this chapter — the first because the trap corrections are computed from measured frequencies, the second because the hadronic term is taken from [2353] rather than recomputed — but both are results the book asserts and must therefore prove.

Chapter 113

Experiment: The Higgs Boson Discovery

Tests Phenomena 104.59 and 104.62.

Assuming Propositions 104.26 and 104.61 and Definition 11.70.

On 4 July 2012 the ATLAS [2186] and CMS [2192] collaborations independently reported a new boson near $125\,\mathrm{GeV}/c^2$, each at five standard deviations, in the diphoton and four-lepton channels. This chapter is the experimental closure of Chapter 104: the Brout–Englert–Higgs mechanism [1275, 1279, 1283] had been established indirectly for decades by the masses of the W and Z [1710, 1711] and by the precision electroweak fits [2104], but the scalar itself — the one particle of the Standard Model whose mass the theory does not predict — remained unobserved, its allowed mass window closed from below by LEP [2045] and eroded by the Tevatron [2187]. The chapter sits at the end of the Standard Model sequence because it completes its particle content; what it does not complete is the theory, and the final sections say so.

Two features of this experiment demand more space than usual. The first is statistics: a discovery claim built on a few hundred events over a large smooth background needs its test statistic, its look-elsewhere correction and its blinding protocol stated as carefully as its detector [2036, 2172]. All of that machinery is derived in Chapter 11, and Section 113.5.1 applies it to the published numbers rather than gesturing at "five sigma". The second is that the measurement did not stop in 2012 — spin-parity [2201, 2217], couplings against mass [2263, 2394, 2410] and the total width [2411] have all been measured since, while the Higgs self-coupling, which is what would actually test the shape of the potential, has not.

Notation 113.1 (Units and the quantities a collider quotes). Everything here is in SI, in the conventions of Section 104.1: $\hbar$ and c appear explicitly, a mass is written as an energy divided by c^2, and the electroweak scale is the energy $E_v = v\sqrt{\hbar c} = 246.220\,\mathrm{GeV}$ of Equation (104.2). Three collider habits need a translation that is given once and then used silently.

Energy. $1\,\mathrm{GeV}$ is $1.602\,176\,634 \times 10^{-10}\,\mathrm{J}$ exactly, since the elementary charge is exact in the 2019 SI [2431]. A mass of $125.20\,\mathrm{GeV}/c^2$ is therefore $2.2319 \times 10^{-25}\,\mathrm{kg}$.

Cross section. An area. The literature quotes barns ($1 \times 10^{-28}\,\mathrm{m}^2$) with SI prefixes, so a picobarn is $1 \times 10^{-40}\,\mathrm{m}^2$ and a femtobarn is $1 \times 10^{-43}\,\mathrm{m}^2$. Every cross section below is given in m^2 with the customary figure named in words.

Luminosity. The instantaneous luminosity L is defined by $\mathrm{d}N/\mathrm{d}t = L\sigma$ and carries $1/(\mathrm{m}^2\,\mathrm{s})$; its time integral carries $1/\mathrm{m}^2$. One inverse femtobarn is $1 \times 10^{43}/\mathrm{m}^2$, and the discovery datasets are a few of them.

The letter μ. Collider practice and field-theory practice collide on this symbol, and five distinct meanings appear in this chapter. To keep them apart: the mass parameter of the scalar potential is written μ only inside Equation (113.1) and the curvature statement that follows from it, always with the combination $\hbar\mu c$ carrying an energy; the *signal strength*, the

ratio of an observed rate to the Standard Model prediction, is written μ with a subscript naming the measurement where one is needed (μ_{on}, μ_{off}, $\bar{\mu}$), and the mean pile-up is $\langle\mu\rangle$; the renormalization scale is written μ_R throughout, never bare μ; and the muon appears only as a particle label, in $\mu^{\pm}$, m_μ and κ_μ. The SI prefix *micro* is a different glyph and appears only inside a unit, as in 18.8 µm.

Remark 113.2 (Reading this chapter against the collider literature). Every experimental paper cited below sets $\hbar = c = 1$, quotes masses in GeV with the c^2 suppressed, cross sections in barns and luminosities in inverse femtobarns. The dictionary is the one fixed in Remark 104.2 together with the three conversions of Notation 113.1. No derivation in this chapter is carried out with $\hbar$ or c suppressed; this remark exists so that a reader can put a published figure beside a formula here and see that they are the same statement.

113.1 Historical context and the prediction under test

113.1.1 The Brout–Englert–Higgs mechanism

The problem the mechanism solves is stated in Theorem 104.8: a mass term for a gauge field is not gauge invariant, so a gauge theory of the weak interaction predicts massless carriers, and the carriers are not massless. The obvious repair — write the mass term anyway — destroys renormalizability (Theorem 104.4). The mechanism is the non-obvious repair: leave the Lagrangian invariant and let the *vacuum* fail to be.

Three papers in 1964 said this independently. Englert and Brout [1275] computed the vector self-energy directly and found a pole at nonzero momentum; Higgs [1282, 1283] gave the abelian model and, in the second paper, noted explicitly that a massive scalar excitation survives; Guralnik, Hagen and Kibble [1279] gave the operator argument for why Goldstone's theorem does not apply. Higgs later wrote the argument out in full [1332], and Kibble extended it to a non-abelian group [1352], which is the case the Standard Model needs. The physical precedent was already in the literature: Anderson [1245] had pointed out that the plasmon in a superconductor is exactly a gauge field that has acquired a mass from a condensate, and that the would-be massless mode is absent there for the same reason. Section 104.3.1 develops the analogy through Nambu's treatment of superconductivity [1187].

What has to be evaded is Goldstone's theorem [1207], proved as Theorem 104.17: a spontaneously broken continuous *global* symmetry produces one massless scalar for every broken generator. Massless scalars are not observed. The evasion, derived as Theorem 104.19 and in the non-abelian case as Theorem 104.22, is that when the broken symmetry is *gauged* the would-be Goldstone modes are not physical states at all: a gauge transformation removes them from the scalar sector, and they reappear as the longitudinal polarizations that a massless vector does not have and a massive one does.

The bookkeeping is worth doing explicitly, because it is what makes the prediction sharp.

Proposition 113.3 (Degrees of freedom before and after)**.** *The electroweak scalar sector is one complex* SU(2) *doublet, four real fields. Before breaking, the field content* $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ *carries*

$$4 \text{ (scalar)} + 3 \times 2 \text{ (massless } W^a\text{)} + 1 \times 2 \text{ (massless } B\text{)} = 12$$

propagating polarizations. After breaking it carries

$$1 \text{ (scalar)} + 3 \times 3 \text{ (massive } W^{\pm}, Z\text{)} + 1 \times 2 \text{ (photon)} = 12\,.$$

Exactly one physical scalar remains, and its existence is the falsifiable content of the mechanism.

Rests on Proposition 26.34, Theorem 104.25, and Equation (104.23).

Derivation. *Derives Proposition 113.3.*

A massless vector field in four spacetime dimensions has two polarizations and a massive one has three; the constraint analysis that states those counts is Remark 26.33 and Proposition 26.34 — where the Proca count is stated and its constraint analysis is itself recorded as owed — and the Proca mass term itself is Equation (104.5). A complex doublet is four real scalar fields. The vacuum Equation (104.29) is invariant under the electric-charge generator and under nothing else in the four-parameter group, so three of the four generators are broken (Theorem 104.25). Goldstone's theorem in its global form would give three massless scalars; gauged, each is instead removed by a gauge choice — the unitary gauge of Equation (104.23) — and supplies the missing longitudinal polarization of one vector. Three vectors gain one polarization each, so the vector sector gains three; the scalar sector loses three. The total is unchanged, which it must be, because a gauge choice cannot create or destroy a physical state. What is left over is the radial excitation of the doublet about the minimum, one real field, and the theory says its mass is $\sqrt{2\hat{\lambda}}\,E_v$ by Equation (104.37).

The counting also fixes what the experiment must find. Not "a scalar": exactly one, electrically neutral, and — once its mass is measured — with every coupling and every branching fraction already determined. A second scalar, a charged scalar, or a scalar with the wrong couplings would each falsify the minimal mechanism, and none would be an adjustment of it. □

Remark 113.4 (What was already evidence before 2012). The mechanism was not an untested hypothesis in 2012. Three of its consequences had been measured. The W and Z exist with the masses the theory predicts from E_v and the mixing angle (Section 104.6.2); the photon is exactly massless, which in Theorem 104.22 is a consequence of the vacuum's residual symmetry and not an input; and the ρ parameter of Equation (113.6) is unity, which is a statement about the *representation* of the breaking sector and not merely about the fact of breaking. What was missing was the scalar itself. The distinction matters: a reader who takes 2012 as the year the electroweak theory was confirmed has the history backwards. The theory was confirmed in 1973 and 1983; 2012 completed its particle content.

113.1.2 The one free parameter

The scalar potential is

$$V(\Phi) = -\mu^2 |\Phi|^2 + \lambda |\Phi|^4 \,, \tag{113.1}$$

with Φ of dimension $\mathrm{J}^{1/2}/\mathrm{m}^{1/2}$, μ of dimension $1/\mathrm{m}$ and λ of dimension $1/(\mathrm{J\,m})$, so that V is an energy density (Definition 104.24). Two parameters appear, and one combination of them is already measured.

Proposition 113.5 (One number is unknown, and it is the mass)**.** *The vacuum expectation value is fixed by the Fermi constant,*

$$E_v = v\sqrt{\hbar c} = \left(\frac{(\hbar c)^3}{\sqrt{2}\, G_F} \right)^{1/2} = 246.220\,\mathrm{GeV}\,, \tag{113.2}$$

and the scalar mass is

$$m_H c^2 = \sqrt{2\hat{\lambda}}\, E_v \,, \qquad \hat{\lambda} = \lambda \hbar c \,. \tag{113.3}$$

Since $\hat{\lambda}$ is not predicted, m_H is the single unknown of the sector; every production cross section and every branching fraction is a function of it alone.

Rests on Propositions 104.26 and 104.61 and Corollary 104.29.

Derivation. *Derives Proposition 113.5.*

Equation (113.2) is Proposition 104.26: matching the low-energy limit of W exchange to Fermi's contact theory cancels the gauge coupling and leaves $E_v^2 = (\hbar c)^3/\sqrt{2}G_F$, with

$G_F/(\hbar c)^3 = (1.166\,378\,7 \pm 0.000\,000\,6) \times 10^{-5}/\mathrm{GeV}^2$ from the positive-muon lifetime [2225], that is $G_F = 1.435\,85 \times 10^{-62}\,\mathrm{J\,m^3}$. In SI, $E_v = 3.944\,88 \times 10^{-8}\,\mathrm{J}$ and the associated length is $\hbar c/E_v = 8.0143 \times 10^{-19}\,\mathrm{m}$. Equation (113.3) is Corollary 104.29, obtained by expanding Equation (113.1) about the minimum $|\Phi| = v/\sqrt{2}$ and reading off the coefficient of the quadratic term in the radial fluctuation.

The claim that everything else follows is the content of Proposition 104.61: the coupling of the scalar to any massive particle is that particle's mass term divided by v, so the couplings contain no parameter beyond E_v, which is already measured. Fixing m_H therefore fixes every partial width through Equation (104.67) and its gauge analogue, hence every branching fraction (Table 104.4); and it fixes every production cross section, because production is the same set of vertices read backwards, convolved with parton distributions measured elsewhere (Section 116.6.5). □

The consequence for the experiment is structural and it is what makes a null result meaningful. A search that scans m_H is testing a *one-parameter family of fully specified theories*, not fitting a shape to data. At each candidate mass the predicted rate in every channel is a number, so the data either accommodate it or exclude it; there is nothing to adjust. And when a signal does appear, the same rigidity makes the discovery over-determined: the mass measured in the diphoton channel must agree with the mass measured in the four-lepton channel, the ratio of the two rates must be the predicted one, and the production mechanisms must appear in the predicted proportion. Two experiments each measuring two channels give eight numbers where the theory has one.

The wider accounting — how many free parameters the Standard Model has in total, and which measurement fixes each — is tabulated in Appendix C. The scalar mass is one entry of that table, and Section 113.5.2 is the entry that this chapter fills in.

113.1.3 Indirect constraints

Before the scalar was produced it was constrained, because it circulates in loops. The relation between M_W, M_Z, α and G_F receives a radiative correction Δr, defined by Equation (104.43) and evaluated from the measured masses in Remark 104.33, which finds $\Delta r = 0.036$ and identifies the scalar-mass logarithm as one of the pieces making it up. What matters here is that dependence's *form*: it is logarithmic,

$$\frac{\partial\,\Delta r}{\partial \ln\left(m_H c^2\right)} = \frac{\hat{g}^2}{16\pi^2}\,c_H\,, \tag{113.4}$$

with $\hat{g}$ the dimensionless SU(2) coupling of Notation 104.1 and c_H a pure number of order unity that this book does not derive, whereas the dependence on the top mass is quadratic, $\Delta r|_t \propto m_t^2$ by Equation (104.52). The two sensitivities have very different consequences and the difference is Veltman's screening theorem [1612]: a heavy scalar decouples logarithmically, so precision data constrain it weakly, while a heavy top does not decouple at all.

Derivation pending #199

The coefficient of the scalar-mass logarithm in the radiative correction Δr, written parametrically in this subsection, together with the constant term that accompanies it — which is not a refinement at the measured mass but comparable to the logarithm itself. Obtaining both requires the one-loop W and Z self-energies with the scalar circulating, renormalized on shell. The electroweak chapter already records the same logarithm as owed, in the paragraph following its screening-theorem discussion; the two belong together in Appendix A. Nothing in this chapter is evaluated numerically from the coefficient.

Weak is not zero. The LEP and SLC programme measured M_Z, the Z partial widths and a dozen asymmetries to a part in 10^3 or better [2104]; combined with M_W, m_t, α and G_F the system is over-constrained, and the global fit before 2012 preferred a scalar mass of order $100\,\mathrm{GeV}/c^2$ with an upper bound in the low hundreds [2228]. The particle was found at $125\,\mathrm{GeV}/c^2$, inside that range.

What made the constraint credible in advance was that the same method had already worked once, on a harder case. The quadratic top sensitivity let the LEP data locate the top quark's mass before the Tevatron produced it, and the indirect value agreed with the direct measurement [1891, 1892]. A method that has predicted one unseen particle correctly is evidence about the next one; a method that has not is an extrapolation.

Phenomenon 113.6 (The weak bosons are massive and the photon is not)**.** The carriers of the charged and neutral weak currents are massive,

$$m_Z c^2 = (91.1880 \pm 0.0020)\,\mathrm{GeV}\,, \qquad m_W c^2 = (80.369 \pm 0.013)\,\mathrm{GeV}\,, \tag{113.5}$$

that is $m_Z = 1.625\,57 \times 10^{-25}\,\mathrm{kg}$ and $m_W = 1.432\,71 \times 10^{-25}\,\mathrm{kg}$ [2427], while the photon, which belongs to the same electroweak gauge group, has no measurable mass at all [2427]. The measured masses moreover satisfy

$$\rho = \frac{m_W^2}{m_Z^2 \cos^2\theta_W} = 1 \tag{113.6}$$

to within the radiative corrections — a coincidence that a general symmetry-breaking sector has no reason to produce.

Rests on Equation (104.36), Theorem 104.41, and Proposition 104.43.

Derivation. *Derives Phenomenon 113.6.*

Let the breaking be produced by a single complex scalar doublet with vacuum expectation value v (Chapter 104). Expanding the square of its covariant derivative about the vacuum generates

$$m_W c^2 = \tfrac{1}{2}\hat{g} E_v\,, \qquad m_Z c^2 = \tfrac{1}{2}\sqrt{\hat{g}^2 + \hat{g}'^2}\, E_v\,,$$

by Equation (104.36), with $\hat{g}$ and $\hat{g}'$ the dimensionless SU(2) and U(1) couplings of Notation 104.1, while the orthogonal combination of neutral gauge fields — the photon — gets no mass term at all, because the vacuum is left invariant by the electric-charge generator. The mixing angle is defined by $\cos\theta_W = \hat{g}/\sqrt{\hat{g}^2 + \hat{g}'^2}$, so

$$\frac{m_W c^2}{m_Z c^2 \cos\theta_W} = \frac{\tfrac{1}{2}\hat{g} E_v}{\tfrac{1}{2}\sqrt{\hat{g}^2 + \hat{g}'^2}\, E_v} \cdot \frac{\sqrt{\hat{g}^2 + \hat{g}'^2}}{\hat{g}} = 1\,,$$

which is Equation (113.6). The relation is a property of the *doublet* assignment and not of symmetry breaking in general: a scalar in a higher representation gives $\rho \neq 1$, as Theorem 104.41 makes precise. The measured ρ is therefore evidence about the representation content of the breaking sector, not merely evidence that something breaks the symmetry, and it is what made the search for a single neutral scalar, rather than for some richer sector, the right search to mount.

Numerically, the comparison can only be made in a scheme that defines the mixing angle independently of the mass ratio, and Remark 104.34 is where the schemes are separated. With the *on-shell* angle, $\sin^2\theta_W := 1 - M_W^2/M_Z^2$, Equation (113.6) is $\rho = 1$ identically by construction and there is nothing to test. Taking instead the $\overline{\mathrm{MS}}$ angle at the Z scale, $\hat{s}_Z^2 = 0.231\,29 \pm 0.000\,04$ [2427], and the masses of Equation (113.5),

$$\rho = \frac{(80.369\,\mathrm{GeV})^2}{(91.1880\,\mathrm{GeV})^2\,(1 - 0.231\,29)} = 1.0105\,,$$

a departure from unity of about one per cent — the size of the one-loop correction $\Delta\rho = 9.3 \times 10^{-3}$ computed in Proposition 104.43. The relation is not exact and is not supposed to be; what is significant is that the deviation is of the size a loop produces and not of the size a wrong representation would produce. □

113.1.4 LEP and Tevatron searches

Two machines searched before the LHC, in two entirely different ways, and between them they set the window the LHC had to cover.

LEP. An e^+e^- collider produces the scalar through $e^+e^- \to Z^* \to ZH$, the same HZZ vertex read as a production vertex. The event is fully constrained — the initial state has known energy — so the search is a bump hunt in the recoil mass against an identified Z, with essentially no combinatorial background. What limits it is not statistics but kinematics.

Proposition 113.7 (LEP's reach was set by its beam energy)**.** *Associated production $e^+e^- \to ZH$ at centre-of-mass energy $\sqrt{s}\,c$ requires*

$$m_H c^2 \le \sqrt{s}\,c - m_Z c^2 \,, \tag{113.7}$$

so LEP's final energy of $\sqrt{s}\,c = 209\,\mathrm{GeV}$ admits at most $m_H c^2 = 117.8\,\mathrm{GeV}$, and the useful reach is a few GeV *below that because the cross section vanishes at threshold.*

Rests on Equation (113.5).

Derivation. *Derives Proposition 113.7.*

In the centre-of-mass frame the total energy is $\sqrt{s}\,c$ and the final state is two particles at rest or better, so $\sqrt{s}\,c \ge m_Z c^2 + m_H c^2$, which is Equation (113.7). With $m_Z c^2 = (91.1880 \pm 0.0020)\,\mathrm{GeV}$ from Equation (113.5) and $\sqrt{s}\,c = 209\,\mathrm{GeV}$ the bound is 117.8 GeV. Near threshold the two-body phase space carries the momentum factor p^*, and for this process the cross section falls as p^* rather than as p^{*3} because the S-wave is allowed; either way it vanishes, so the last few GeV of the kinematic range are not usable. The published limit, $m_H c^2 > 114.4\,\mathrm{GeV}$ at 95 % confidence [2045], sits about 3.4 GeV below the kinematic edge, which is the size of that effect. □

The LEP combination of ALEPH, DELPHI, L3 and OPAL left a mild excess at the very edge of its sensitivity, near $115\,\mathrm{GeV}/c^2$, of less than two standard deviations in the four-experiment combination and driven largely by one experiment [2045]. It did not survive, and the way it was reported is worth recording: the collaborations quoted the limit and the excess together, with the excess described as compatible with a background fluctuation, rather than either suppressing it or promoting it. The combination used the modified frequentist CL_s construction [2036], whose purpose is exactly to prevent a downward background fluctuation from excluding a signal the experiment had no sensitivity to (Remark 11.95).

Tevatron. A $p\bar{p}$ collider at $\sqrt{s}\,c = 1.96\,\mathrm{TeV}$ cannot use the diphoton channel — the rate is too low for its integrated luminosity — so CDF and D0 attacked the largest branching fraction instead, $H \to b\bar{b}$, made usable by demanding an associated W or Z to suppress the enormous QCD production of b quarks. The combined analysis of the full Tevatron dataset reported a broad excess in that channel with a maximum local significance of about three standard deviations near $125\,\mathrm{GeV}/c^2$, falling to roughly two and a half after correcting for the mass range searched [2187]. The same combination excluded a Standard Model scalar between about $147\,\mathrm{GeV}/c^2$ and $179\,\mathrm{GeV}/c^2$.

The Tevatron result is the one place in this story where the associated-production channel with the largest branching ratio was competitive, and it is instructive that it produced evidence rather than a discovery. The excess was broad — the $b\bar{b}$ mass resolution is of order $15\,\mathrm{GeV}/c^2$, ten times worse than the diphoton channel — so it constrained a rate over a wide window rather than locating a resonance.

The window entering 2012. Combining the LEP lower bound, the Tevatron exclusion and the LHC's own 2011 exclusions, a Standard Model scalar was confined to roughly 115–130 GeV/c^2, with a smaller surviving region at high mass that the electroweak fit already disfavoured. Inside that window the branching fractions of Table 104.4 are changing rapidly: WW^* and ZZ^* are climbing as the thresholds approach, $b\bar{b}$ is falling, and $\gamma\gamma$ passes through its maximum. It is the one mass region in which the two high-resolution channels are simultaneously accessible, which is why the LHC was able to settle the question there and would have had a much harder time 20 GeV/c^2 higher.

113.2 Apparatus

Three instruments are involved: one accelerator and two detectors. The accelerator's job is to deliver, in a few years, enough proton–proton collisions that a process occurring once in 3×10^9 of them yields a few hundred events. The detectors' job is to recognise those events in real time, discard the rest before they can be written to disk, and measure the survivors well enough that a resonance 1 GeV/c^2 wide can be seen on a continuum background a hundred times larger.

ATLAS and CMS were built to different designs on purpose. Their magnets, their calorimeter technologies and their reconstruction strategies differ in almost every respect, so their systematic errors are largely independent. That is not redundancy: it is the only protection available against a mistake that a single collaboration could make and not detect, and it is why the phrase "two experiments independently" carries weight in Section 113.4.3 that "two analyses of the same data" would not.

113.2.1 The Large Hadron Collider

The machine [2123] occupies the tunnel dug for LEP, a ring of circumference 26 659 m lying between 45 m and 170 m below the surface and tilted by 1.4°. Reusing the tunnel fixed the circumference and therefore fixed everything else: to reach a beam energy of several TeV in a ring of that size the bending field must be far beyond what iron can supply, so the LHC is a superconducting machine, and because two counter-rotating proton beams need opposite bending fields it uses twin apertures inside one cryostat and one yoke rather than two separate rings.

Proposition 113.8 (The dipole field the ring demands)**.** *A relativistic proton of energy E on a circular arc of radius ρ requires a field*

$$B = \frac{p}{e\rho} \simeq \frac{E}{ec\rho}\,, \tag{113.8}$$

p being the momentum and e the elementary charge. The LHC's 1232 *main dipoles are each* 14.3 m *long, so the total bending length is* 17 618 m *and* $\rho = 2804$ m*. The three beam energies that matter here then need*

$$B\,(3.5\,\mathrm{TeV}) = 4.16\,\mathrm{T}\,, \quad B\,(4\,\mathrm{TeV}) = 4.76\,\mathrm{T}\,, \quad B\,(7\,\mathrm{TeV}) = 8.33\,\mathrm{T}\,. \tag{113.9}$$

Rests on Equations (40.4) *and* (40.15).

Derivation. *Derives Proposition 113.8.*

The Lorentz force supplies the centripetal acceleration, $evB = \gamma m_p v^2/\rho$, so $B = \gamma m_p v/(e\rho) = p/(e\rho)$ exactly. At 4 TeV the proton's Lorentz factor is $\gamma = E/(m_p c^2) = 4000/0.938\,272 = 4263$, so v/c differs from unity by $1/(2\gamma^2) = 2.8 \times 10^{-8}$ and $p = E/c$ to

that accuracy. The bending radius is the total magnetic length divided by 2π: $1232 \times 14.3\,\mathrm{m} = 17\,617.6\,\mathrm{m}$, and $17\,617.6\,\mathrm{m}/2\pi = 2803.9\,\mathrm{m}$. Then

$$p = \frac{4 \times 10^{12}\,\mathrm{eV} \times 1.602\,176\,634 \times 10^{-19}\,\mathrm{J/eV}}{2.997\,924\,58 \times 10^{8}\,\mathrm{m/s}} = 2.1377 \times 10^{-15}\,\mathrm{kg\,m/s}\,,$$

and dividing by $e\rho = 1.602\,176\,634 \times 10^{-19}\,\mathrm{C} \times 2803.9\,\mathrm{m} = 4.4923 \times 10^{-16}\,\mathrm{C\,m}$ gives $4.759\,\mathrm{T}$. The other two entries follow by proportionality. An iron-dominated warm dipole saturates near $2\,\mathrm{T}$; a $7\,\mathrm{TeV}$ beam in such a ring would need a bending radius above $11\,\mathrm{km}$ and a circumference near $100\,\mathrm{km}$, which is the whole argument for superconductivity. The LHC dipoles use niobium–titanium cable operated at $1.9\,\mathrm{K}$ in superfluid helium, below the $2.17\,\mathrm{K}$ lambda point, because the critical field of that alloy at $4.2\,\mathrm{K}$ would not reach $8\,\mathrm{T}$. □

The beams are bunched by $400\,\mathrm{MHz}$ superconducting radio-frequency cavities, eight per beam delivering about $2\,\mathrm{MV}$ each. A proton completes the ring at

$$f_{\mathrm{rev}} = \frac{c}{26\,659\,\mathrm{m}} = 11\,245\,\mathrm{Hz}\,, \tag{113.10}$$

to within the 2.8×10^{-8} by which v falls short of c. In the 2011 and 2012 runs the bunches were spaced by $50\,\mathrm{ns}$ — twice the design spacing — which puts consecutive bunches $15.0\,\mathrm{m}$ apart along the beam and gives a crossing rate of $20\,\mathrm{MHz}$ at each interaction point where bunches are present. With 1374 filled bunches in 2012 the actual crossing rate was $1374\,f_{\mathrm{rev}} = 15.4\,\mathrm{MHz}$.

Remark 113.9 (The stored energy is the reason for the caution). Each 2012 bunch carried about 1.6×10^{11} protons, so a beam held 2.20×10^{14} protons at $4\,\mathrm{TeV}$, a stored kinetic energy of

$$2.20 \times 10^{14} \times 4 \times 10^{12}\,\mathrm{eV} \times 1.602\,176\,634 \times 10^{-19}\,\mathrm{J/eV} = 1.41 \times 10^{8}\,\mathrm{J}\,,$$

$141\,\mathrm{MJ}$ per beam, comparable to the kinetic energy of a $4 \times 10^{5}\,\mathrm{kg}$ train at $28\,\mathrm{m/s}$, which is $\tfrac{1}{2} \times 4 \times 10^{5}\,\mathrm{kg} \times (28\,\mathrm{m/s})^2 = 1.6 \times 10^{8}\,\mathrm{J}$. The magnet system stores an order of magnitude more. This is not a detector concern but it dictated the 2011–2012 operating energy: after the 2008 interconnect failure the machine ran at half the design field while the splices were surveyed, which is why the discovery datasets are at $7\,\mathrm{TeV}$ and $8\,\mathrm{TeV}$ rather than $14\,\mathrm{TeV}$.

Proposition 113.10 (Luminosity and the pile-up it forces)**.** *For head-on collisions of n_b bunch pairs with N protons each, Gaussian transverse profiles of common width σ at the interaction point,*

$$L = \frac{f_{rev} n_b N^2}{4\pi\sigma^2}\,F\,, \qquad \sigma = \sqrt{\frac{\beta^{*}\epsilon_n}{\gamma}}\,, \tag{113.11}$$

with $F \le 1$ a geometric reduction from the crossing angle, β^{*} the optical amplitude function at the collision point and ϵ_n the normalized transverse emittance. The 2012 parameters give $\sigma = 18.8\,\mu\mathrm{m}$ and a peak luminosity of order $8 \times 10^{37}/(\mathrm{m^2\,s})$. The mean number of inelastic collisions per bunch crossing is

$$\langle \mu \rangle = \frac{L \sigma_{inel}}{f_{rev} n_b}\,, \tag{113.12}$$

which at that luminosity is about 36.

Rests on Notation 113.1.

Derivation. *Derives Proposition 113.10.*

The rate of any process is $\mathrm{d}N/\mathrm{d}t = L\sigma$ by the definition of luminosity, and for two Gaussian bunches crossing head-on the overlap integral of two normalized transverse densities of width σ is $1/(4\pi\sigma^2)$; multiplying by the number of protons in each bunch and by the

rate at which bunch pairs meet gives Equation (113.11). The transverse size follows from the definition of emittance: the geometric emittance is $\epsilon = \epsilon_n/\gamma$ because adiabatic damping shrinks the transverse phase-space area in proportion to the momentum, and $\sigma^2 = \beta^*\epsilon$ is the definition of the amplitude function. With $\epsilon_n = 2.5 \times 10^{-6}\,\mathrm{m}$, $\gamma = 4263$ and $\beta^* = 0.6\,\mathrm{m}$,

$$\sigma = \sqrt{0.6\,\mathrm{m} \times \frac{2.5 \times 10^{-6}\,\mathrm{m}}{4263}} = 1.876 \times 10^{-5}\,\mathrm{m}\,.$$

Then

$$\frac{f_{\mathrm{rev}} n_b N^2}{4\pi\sigma^2} = \frac{11\,245\,\mathrm{Hz} \times 1374 \times \left(1.6 \times 10^{11}\right)^2}{4\pi\left(1.876 \times 10^{-5}\,\mathrm{m}\right)^2} = 8.9 \times 10^{37}/(\mathrm{m}^2\,\mathrm{s})\,,$$

and the crossing angle of a few hundred microradians, needed to keep the two beams from colliding parasitically elsewhere in the common pipe, supplies $F \approx 0.85$; the measured 2012 peak was $7.7 \times 10^{37}/(\mathrm{m}^2\,\mathrm{s})$.

For Equation (113.12), the inelastic proton–proton cross section at $\sqrt{s}\,c = 8\,\mathrm{TeV}$ is about $7.3 \times 10^{-30}\,\mathrm{m}^2$, which the collider literature writes as 73 millibarns, so the inelastic rate at peak luminosity is $7.7 \times 10^{37}/(\mathrm{m}^2\,\mathrm{s}) \times 7.3 \times 10^{-30}\,\mathrm{m}^2 = 5.6 \times 10^{8}/\mathrm{s}$. Dividing by the crossing rate $1.54 \times 10^{7}/\mathrm{s}$ gives 36 collisions per crossing at the start of a fill. Luminosity decays through a fill as the bunches are consumed and heated, so the luminosity-weighted average over the whole 2012 run was about 21, against about 9 in 2011. □

Remark 113.11 (Why pile-up is the defining difficulty). Twenty simultaneous inelastic collisions is not twenty times more data. All of them occur within 50 ns — in fact within the same crossing — and within a luminous region whose longitudinal extent is $\sigma_z \approx 5\,\mathrm{cm}$, so they are not separated in time and barely in space. Three specific damages follow, and every element of the analyses in Section 113.3 is shaped by one of them.

First, the event of interest must be assigned to the right vertex out of twenty candidates, and for a final state of two photons there are no tracks from the hard process to point with. Proposition 113.14 shows what a wrong assignment costs.

Second, energy from the other collisions lands in the same calorimeter cells, so any isolation or energy measurement must be corrected for a pedestal that fluctuates event by event. This is *in-time* pile-up.

Third, the ionization signal in a liquid-argon calorimeter is triangular with a drift time near 450 ns, nine times the bunch spacing, so the shaped pulse also carries contributions from neighbouring crossings. This is *out-of-time* pile-up, and it is removed by bipolar shaping whose undershoot cancels the average of the preceding crossings — which works on the average and leaves a fluctuation.

The last machine property that matters is not a machine property at all but a property of the proton. The scalar is produced dominantly by gluon fusion, $gg \to H$, through a top-quark loop, and that dominance is a statement about parton densities rather than about couplings: at $\sqrt{s}\,c = 8\,\mathrm{TeV}$ a central resonance of mass $125\,\mathrm{GeV}/c^2$ is made from partons carrying momentum fractions near $x = m_H c^2/(\sqrt{s}\,c) = 0.0156$, and at that x the gluon distribution is an order of magnitude above any single quark distribution (Section 116.6.5). The distributions themselves were measured in electron–proton scattering (Chapter 116) and are usable here only because collinear factorization (Theorem 102.53) makes them universal — a theorem stated there at leading order, with its all-orders proof recorded as owed. Without it, and without the deep inelastic programme that measured its inputs, no LHC cross section could be predicted at all.

113.2.2 ATLAS

ATLAS [2119] is 44 m long and 25 m in diameter and weighs about 7×10^6 kg. Its distinguishing choices are a modest solenoid around a compact tracker, a sampling electromagnetic

calorimeter with unusually fine longitudinal and lateral segmentation, and a very large air-core toroid outside the calorimeters that measures muons independently of the tracker.

Tracking. A superconducting solenoid 5.3 m long with a 2.4 m bore produces an axial field of 2 T; it is deliberately thin, about 0.66 radiation lengths, because it sits in front of the electromagnetic calorimeter. Inside it are three subsystems covering $|\eta| < 2.5$: a silicon pixel detector of three barrel layers at radii 50.5 mm, 88.5 mm and 122.5 mm with 8.0×10^7 channels of 50 µm × 400 µm; a silicon microstrip tracker of four double barrel layers out to 514 mm with 80 µm pitch; and a transition-radiation tracker of about 3.5×10^5 straw tubes of 4 mm diameter, giving roughly 36 measurements per track and, from the transition radiation emitted in the interleaved radiator, an electron-versus-pion discriminant.

Electromagnetic calorimetry. Lead absorbers immersed in liquid argon at 87.3 K, folded into an accordion so that the readout can be brought out from the front and back with no cracks in azimuth. The barrel covers $|\eta| < 1.475$ and the two endcaps $1.375 < |\eta| < 3.2$; the total depth is at least 22 radiation lengths in the barrel. The segmentation is the point of the design and is worth stating in full:

- a presampler layer for $|\eta| < 1.8$, which measures the energy already lost upstream;
- a first sampling of about 4.3 radiation lengths divided into strips of $\Delta\eta = 0.0031$, eight times finer than the layers behind it;
- a second sampling of about 16 radiation lengths in cells of $\Delta\eta \times \Delta\phi = 0.025 \times 0.025$, which collects most of the shower;
- a third sampling of $\Delta\eta = 0.05$ to catch the tail.

The design energy resolution is

$$\frac{\sigma_E}{E} = \frac{10\,\%}{\sqrt{E/1\,\mathrm{GeV}}} \oplus 0.7\,\%\,, \tag{113.13}$$

the first term being the sampling fluctuation of a lead–argon stack and the second the constant term set by calibration and mechanical uniformity.

The fine first sampling does two things that the discovery depended on. A neutral pion of 50 GeV decays to two photons with a typical opening angle of $2m_\pi c^2/E \approx 5.4 \times 10^{-3}$, which at the calorimeter face is a separation of a few millimetres — resolvable in strips of $\Delta\eta = 0.0031$ and invisible in cells of 0.025. And because the first and second samplings measure η at two different radii, the photon's direction can be reconstructed from the calorimeter alone, giving its origin along the beam to about 15 mm with no help from the tracker. Proposition 113.14 explains why that number matters.

Hadronic calorimetry and muons. Steel and scintillating tile in the barrel, copper and liquid argon in the endcaps, and a copper–tungsten forward calorimeter reaching $|\eta| = 4.9$, which is what makes a missing-transverse-momentum measurement possible. Outside everything sits the muon spectrometer: eight air-core superconducting coils forming a barrel toroid 25.3 m long and 20.1 m in outer diameter, plus two endcap toroids, giving a bending integral $\int B\,\mathrm{d}l$ between 1.5 T m and 7.5 T m depending on η. Air-core means there is no iron to scatter the muon, so the spectrometer measures momentum on its own to about 10 % at 1 TeV; the price is a very large and mechanically demanding structure. Precision coordinates come from drift tubes of 30 mm diameter filled with argon and carbon dioxide at 3×10^5 Pa — three atmospheres — resolving about 80 µm per tube.

Trigger. The crossing rate is 20 MHz and the recording rate is a few hundred hertz, so the trigger must reject about 5×10^4 events for every one it keeps, in real time and without ever looking at an event twice. A hardware first level, using reduced-granularity calorimeter sums and dedicated muon chambers, made its decision within 2.5 µs and reduced the rate to

about 70 kHz; two software levels running on a farm of order 2×10^4 processor cores reduced it to about 400 Hz recorded in 2012. The diphoton trigger required two electromagnetic clusters above roughly 35 GeV and 25 GeV of transverse energy. Any event not selected within 2.5 µs is gone permanently, which is why the trigger menu is part of the physics analysis and not part of the plumbing.

113.2.3 CMS

CMS [2120] is 21.6 m long, 14.6 m in diameter and weighs about 1.25×10^7 kg — half the length of ATLAS and nearly twice the mass. Every one of its choices is the opposite of the corresponding ATLAS choice, and the discovery is stronger for it.

The magnet. A single superconducting solenoid of 6 m inner diameter and 12.5 m length, operated at 3.8 T with a stored energy of 2.6 GJ. Everything except the muon chambers sits inside it. The high field buys track momentum resolution in a small volume; the cost is that the coil is thick, so the electromagnetic calorimeter must be inside it, which in turn constrains its radius and therefore its cost per unit solid angle. The return flux is carried by an iron yoke of about 1.8 T in which the muon chambers are embedded, so the muon system reuses the same magnetic circuit rather than needing one of its own.

Tracking. All silicon, about 200 m^2 of it, the largest such device built at the time. Three pixel barrel layers at radii 4.4 cm, 7.3 cm and 10.2 cm carry 6.6×10^7 pixels of 100 µm × 150 µm; ten microstrip barrel layers reach $r = 1.1$ m with 9.3×10^6 strips. Coverage is $|\eta| < 2.5$.

Electromagnetic calorimetry. This is the subsystem CMS built for the diphoton channel specifically, and the reasoning is explicit in the design report: a homogeneous crystal calorimeter has no sampling fluctuation, so its stochastic term can be several times smaller than a sampling device's, and the diphoton mass resolution is dominated by that term. The calorimeter is 75 848 lead-tungstate ($PbWO_4$) scintillating crystals — 61 200 in the barrel for $|\eta| < 1.479$ and 7324 in each endcap. The material was chosen for density: 8280 $\mathrm{kg/m}^3$, a radiation length of 0.89 cm and a Molière radius of 2.2 cm, so a barrel crystal 230 mm long is 25.8 radiation lengths in a 22 mm × 22 mm front face. The design resolution measured in a test beam is

$$\frac{\sigma_E}{E} = \frac{2.8\,\%}{\sqrt{E/1\,\mathrm{GeV}}} \oplus \frac{12\,\%}{E/1\,\mathrm{GeV}} \oplus 0.30\,\%\,, \tag{113.14}$$

a stochastic term nearly four times smaller than Equation (113.13).

That advantage is bought at a price that had to be engineered away. Lead tungstate is a poor scintillator — about 30 photons per MeV reach the photodetector — so the barrel is read out with avalanche photodiodes of gain 50 and the endcaps with vacuum phototriodes, both able to operate inside 3.8 T where a photomultiplier cannot. Its light yield also falls by about 2 % per kelvin, so the whole calorimeter is held at 18 °C stabilised to 0.05 K; and the crystals darken under irradiation and recover between fills, which forces a continuous transparency monitoring with injected laser light. A preshower of two lead converters and two silicon strip planes, three radiation lengths in all with a strip pitch of 1.9 mm, sits in front of the endcaps and does there what ATLAS's fine strips do in the barrel: separate a π^0's two photons from one.

Reconstruction. CMS reconstructs events by *particle flow*: rather than reading each subdetector separately, the algorithm builds a list of individual particles by linking tracks to calorimeter clusters to muon segments, and then constructs jets, missing transverse momentum and isolation variables from that list. The method needs a tracker that measures nearly every charged particle and a calorimeter finely segmented enough to associate clusters with tracks, which is exactly what the 3.8 T solenoid and the crystal array provide. Its relevance here is isolation: a photon candidate's isolation can be computed from the charged particles associated with the chosen vertex alone, which suppresses the pile-up dependence that a purely calorimetric isolation would carry.

Trigger. The same two-stage structure with different numbers: a hardware level with a 3.2 µs latency reducing 20 MHz to at most 100 kHz, and a software high-level trigger on a commercial farm reducing that to a few hundred hertz. CMS's diphoton triggers used lower transverse-energy thresholds than ATLAS's, near 26 GeV and 18 GeV, compensated by isolation and shower-shape requirements applied online.

Table 113.1: Design and construction parameters of the two detectors and of the machine that fed them, as built and as operated in 2011–2012. These are settings and specifications, not measurements, so no uncertainty is attached to them; the measured quantities of this chapter are collected in Table 113.3 and Table 113.5. Energy resolutions are the design or test-beam parametrizations of Equation (113.13) and Equation (113.14), with E in GeV; the resolutions achieved in situ are worse and are discussed in Section 113.4.1.

Parameter	ATLAS	CMS
Reference	[2119]	[2120]
Length; diameter	44 m; 25 m	21.6 m; 14.6 m
Mass	7×10^6 kg	1.25×10^7 kg
Tracking field	2 T solenoid	3.8 T solenoid
Muon field	air-core toroid, $\int B\,\mathrm{d}l$ from 1.5–7.5 T m	1.8 T iron return yoke
Silicon pixels	8.0×10^7 of 50 µm×400 µm	6.6×10^7 of 100 µm×150 µm
Outer tracking	3.5×10^5 straw tubes, 4 mm diameter	9.3×10^6 silicon strips
Electromagnetic calorimeter	lead/liquid argon accordion, sampling, ≥ 22 radiation lengths	75 848 $PbWO_4$ crystals, homogeneous, 25.8 radiation lengths
Finest lateral cell	$\Delta\eta = 0.0031$ (first sampling)	22 mm crystal face; preshower pitch 1.9 mm
Design σ_E/E	$10\,\%/\sqrt{E} \oplus 0.7\,\%$	$2.8\,\%/\sqrt{E} \oplus 12\,\%/E \oplus 0.30\,\%$
Level-1 latency; output	2.5 µs; ~ 70 kHz	3.2 µs; ≤ 100 kHz
Recorded rate, 2012	~ 400 Hz	~ 400 Hz

113.3 Procedure

113.3.1 Production and decay channels

Four production mechanisms contribute at the LHC, and all four are the vertices of Equation (104.69) read as production rather than decay. In descending order of rate at $m_H c^2 = 125$ GeV and $\sqrt{s}\,c = 8$ TeV:

- *gluon fusion*, $gg \to H$, through a loop of virtual quarks dominated by the top because its Yukawa coupling is the largest, $\sigma \approx 1.93 \times 10^{-39}\,\mathrm{m}^2$ (19.3 picobarns), about 87 % of the total;

- *vector-boson fusion*, $qq \to qqH$, in which each incoming quark radiates a W or Z that fuse, $\sigma \approx 1.58 \times 10^{-40}\,\mathrm{m}^2$, and which is identified by two forward jets with a large rapidity gap;

Table 113.2: The machine as operated in the two discovery runs [2123]. These are settings, not measurements, and carry no uncertainty; the two derived rows are computed in Proposition 113.8 and Proposition 113.10 from the geometry and the beam parameters in the rows above them. The mean pile-up is a luminosity-weighted average over the run and is about 1.7 times smaller than the value at the start of a fill, because the luminosity decays as the bunches are consumed.

Machine parameter	2011 run	2012 run
Circumference	26 659 m	26 659 m
Main dipoles	1232×14.3 m, NbTi at 1.9 K	as 2011
Bending radius (derived)	2804 m	2804 m
Beam energy	3.5 TeV	4 TeV
Centre-of-mass energy $\sqrt{s}\,c$	7 TeV	8 TeV
Dipole field (derived)	4.16 T	4.76 T
Bunch spacing	50 ns = 15.0 m	as 2011
Filled bunches per beam	up to 1380	up to 1374
Protons per bunch	$\approx 1.45 \times 10^{11}$	$\approx 1.6 \times 10^{11}$
Stored energy per beam (derived)	1.1×10^{8} J	1.4×10^{8} J
Revolution frequency	11 245 Hz	11 245 Hz
Peak luminosity	$3.7 \times 10^{37}/(\mathrm{m}^2\,\mathrm{s})$	$7.7 \times 10^{37}/(\mathrm{m}^2\,\mathrm{s})$
Mean pile-up	≈ 9	≈ 21

- *associated production*, $q\bar{q} \to VH$ with $V = W$ or Z, $\sigma \approx 1.12 \times 10^{-40}\,\mathrm{m}^2$ summed over the two, the channel the Tevatron used;
- *top-quark associated production*, $gg \to t\bar{t}H$, $\sigma \approx 1.29 \times 10^{-41}\,\mathrm{m}^2$, which measures the top Yukawa coupling at tree level rather than through a loop.

The sum is $2.21 \times 10^{-39}\,\mathrm{m}^2$ at 8 TeV and $1.74 \times 10^{-39}\,\mathrm{m}^2$ at 7 TeV — in the customary units 22.1 and 17.4 picobarns [2186, 2192]. Set against the inelastic proton–proton cross section of $7.3 \times 10^{-30}\,\mathrm{m}^2$, the scalar is produced in 3.0×10^{-10} of all collisions: one in three thousand million.

Derivation pending #200

The Standard Model production cross sections quoted in this section. Obtaining them requires the partonic amplitudes — above all the top-quark triangle that drives gluon fusion, which has no tree-level counterpart — convolved with the measured parton distributions and corrected to next-to-next-to-leading order in the strong coupling, where the corrections are large (a K factor near two for gluon fusion). The loop amplitude and the resulting cross sections are already recorded as owed in the electroweak chapter's pending derivation of the loop-induced scalar vertices; the convolution and the perturbative corrections belong with them in Appendix A. What is derived in this chapter from those inputs is everything downstream of them: the produced event counts, the channel comparison, the mass resolution and the significance.

Proposition 113.12 (Sensitivity is set by ratio, not by rate)**.** *In the Gaussian limit the significance of a resonance search in a channel with s expected signal and b expected background events in the mass window is $s/\sqrt{b}$, not s. For the ATLAS discovery dataset,*

the three candidate channels give

$$\frac{s}{\sqrt{b}}\Big|_{b\bar{b}} \sim 7\times 10^{-2}\,, \qquad \frac{s}{\sqrt{b}}\Big|_{\gamma\gamma} \sim 2\,, \qquad \frac{s}{\sqrt{b}}\Big|_{4\ell} \sim 4\,, \tag{113.15}$$

although the produced event counts are in the ratio $1.2\times 10^5 : 4.9\times 10^2 : 25$. *The channel with the largest branching fraction is the least useful of the three.*

Rests on Propositions 11.93 and 113.5.

Derivation. *Derives Proposition 113.12.*

Take the ATLAS dataset: $4.8\times 10^{43}/\mathrm{m}^2$ at 7 TeV and $5.8\times 10^{43}/\mathrm{m}^2$ at 8 TeV, that is 4.8 and 5.8 inverse femtobarns. The number of scalars produced is the product of integrated luminosity and cross section, summed over the two runs:

$$N_H = 4.8\times 10^{43}/\mathrm{m}^2 \times 1.74\times 10^{-39}\,\mathrm{m}^2 + 5.8\times 10^{43}/\mathrm{m}^2 \times 2.21\times 10^{-39}\,\mathrm{m}^2 = 2.12\times 10^5\,.$$

Multiplying by the branching fractions of Table 104.4 gives 1.2×10^5 decays to $b\bar{b}$, 487 to $\gamma\gamma$, and — using $\mathcal{B}(H\to ZZ^*) = 0.026$ together with $\mathcal{B}(Z\to e^+e^-)+\mathcal{B}(Z\to\mu^+\mu^-) = 0.0673$, so that $\mathcal{B}(H\to ZZ^*\to 4\ell) = 0.026\times 0.0673^2 = 1.18\times 10^{-4}$ — 25 decays to four charged leptons of the first two generations.

Now the backgrounds.

Bottom pairs. The inclusive $b\bar{b}$ production cross section at these energies is of order $2.8\times 10^{-32}\,\mathrm{m}^2$, which the collider literature writes as 280 microbarns — seven orders of magnitude above the rate at which scalars are produced and decay to $b\bar{b}$, and four thousandths of the inelastic proton–proton cross section quoted above, which it must be. The same luminosity therefore yields $1.06\times 10^{44}/\mathrm{m}^2 \times 2.8\times 10^{-32}\,\mathrm{m}^2 = 3\times 10^{12}$ background events. Even at 100 % efficiency, $s/\sqrt{b} = 1.2\times 10^5/\sqrt{3\times 10^{12}} = 0.07$. Requiring an associated W or Z, as the Tevatron did, cuts the background by orders of magnitude but cuts the signal to the associated-production cross section, a twentieth of the total; that trade is what produced three standard deviations at the Tevatron and not five.

Two photons. Photon reconstruction and the selection of Section 113.3.2 retain roughly 40 % of the produced decays, so $s\approx 195$. The background is continuum production of two photons, plus jets misidentified as photons; in a window of a few GeV around 125 GeV it is of order 1×10^4 events in this dataset. Hence $s/\sqrt{b}\approx 195/100\approx 2$, and a signal fraction $s/b\approx 2\,\%$.

Four leptons. Acceptance and lepton efficiency retain roughly 35 %, so $s\approx 8.7$; the background in a window of a few GeV, dominated by the irreducible continuum $q\bar{q}\to ZZ^*$, is of order 5 events. Hence $s/\sqrt{b}\approx 4$ — and here $s/b>1$, so the "background" is not the thing being fought.

The three numbers are Equation (113.15). Two lessons follow, and both shaped the analyses. A large branching fraction is worth nothing if the final state is one the strong interaction produces copiously by itself, which disqualifies $b\bar{b}$, $c\bar{c}$ and gg — three of the four largest entries in Table 104.4. And the two surviving channels are useful for opposite reasons: $\gamma\gamma$ has many events and a poor signal fraction, 4ℓ has few events and an excellent one, so their systematic weaknesses do not overlap either. □

Remark 113.13 (The discovery channel has no tree-level vertex). The scalar is electrically neutral and colourless, so $H\to\gamma\gamma$ and $gg\to H$ do not exist at tree level. Both proceed through loops of the heaviest charged or coloured particles: gluon fusion through the top-quark triangle, and the diphoton decay through the top triangle *and* the W loop, which interfere destructively with the W dominating. The discovery therefore rested on a one-loop amplitude at both ends — the dominant production mechanism and the most significant decay channel are each loop-induced — which is an unusual situation and a strong one: any heavy charged or coloured particle coupling to the scalar would change these rates whether or not it can be produced directly. The amplitudes themselves are recorded as a pending derivation in Section 104.7.1 and are not repeated here.

113.3.2 Event selection and background estimation

Photons. A photon candidate is an electromagnetic cluster with no matched track (or, for a converted photon, a matched conversion vertex) whose lateral and longitudinal shower shape is consistent with a single electromagnetic shower rather than with two overlapping ones or with a jet fluctuating to a leading π^0. About a quarter of photons convert in the tracker material before reaching the calorimeter, which is why both experiments treat converted and unconverted candidates separately. Isolation — the transverse energy in a cone around the candidate, corrected for pile-up — rejects photons produced inside jets. ATLAS required two candidates with transverse energies above 40 GeV and 30 GeV within $|\eta| < 2.37$ excluding the barrel to endcap transition; CMS scaled its thresholds with the candidate mass, requiring $E_T > m_{\gamma\gamma}c^2/3$ and $m_{\gamma\gamma}c^2/4$, which keeps the selection from sculpting a peak into the background.

Leptons. Electrons are clusters with a matched track; muons are tracks matched between the inner detector and the muon system. The four-lepton selection requires two same-flavour opposite-charge pairs, with the pair closer to the Z mass required to be near it and the second allowed to be far off shell — which is the whole point, since at $m_H c^2 = 125\,\mathrm{GeV}$ the decay $H \to ZZ$ is below threshold by 57 GeV and one boson must be virtual. Lepton transverse-momentum thresholds are low, of order $7\,\mathrm{GeV}/c$ for the softest, because the four leptons share 125 GeV between them.

The measurement that everything else serves is the invariant mass, and its resolution is where the two detector designs cash out.

Proposition 113.14 (What limits the diphoton mass resolution)**.** *For two photons of energies E_1, E_2 separated by an angle θ,*

$$\left(m_{\gamma\gamma}c^2\right)^2 = 2E_1E_2\left(1-\cos\theta\right), \tag{113.16}$$

and the fractional resolution is

$$\frac{\sigma_m}{m} = \frac{1}{2}\left[\left(\frac{\sigma_{E_1}}{E_1}\right)^2 + \left(\frac{\sigma_{E_2}}{E_2}\right)^2 + \left(\frac{|\boldsymbol{p}_H|\,c}{m_H c^2}\right)^2 \sigma_\theta^2\right]^{1/2}, \tag{113.17}$$

the angular term being evaluated for the symmetric decay. It vanishes for a scalar produced at rest and grows with the scalar's momentum; with a misassigned collision vertex it dominates.
Rests on Equations (113.13) and (113.14).

Derivation. *Derives Proposition 113.14.*

For two massless particles $|\boldsymbol{p}_i|\,c = E_i$, so the squared invariant mass of the pair is

$$\left(mc^2\right)^2 = \left(E_1+E_2\right)^2 - \left(\boldsymbol{p}_1+\boldsymbol{p}_2\right)^2 c^2 = 2E_1E_2 - 2\boldsymbol{p}_1\cdot\boldsymbol{p}_2 c^2 = 2E_1E_2\left(1-\cos\theta\right),$$

which is Equation (113.16). Taking the logarithm and differentiating,

$$2\,\frac{\mathrm{d}m}{m} = \frac{\mathrm{d}E_1}{E_1} + \frac{\mathrm{d}E_2}{E_2} + \frac{\sin\theta}{1-\cos\theta}\mathrm{d}\theta = \frac{\mathrm{d}E_1}{E_1} + \frac{\mathrm{d}E_2}{E_2} + \cot\!\left(\frac{\theta}{2}\right)\mathrm{d}\theta,$$

using $\sin\theta/(1-\cos\theta) = \cot(\theta/2)$. Adding the three independent errors in quadrature gives Equation (113.17) once $\cot(\theta/2)$ is expressed kinematically. For the symmetric decay of a parent of energy E_H, $E_1 = E_2 = E_H/2$, so Equation (113.16) reads $(m_H c^2)^2 = E_H^2\sin^2(\theta/2)$, whence $\sin(\theta/2) = m_H c^2/E_H$ and

$$\cot\!\left(\frac{\theta}{2}\right) = \frac{\sqrt{E_H^2 - \left(m_H c^2\right)^2}}{m_H c^2} = \frac{|\boldsymbol{p}_H|\,c}{m_H c^2}.$$

So the angular sensitivity is exactly the parent's momentum in units of its mass: a scalar at rest decays back to back, $\theta = \pi$, and its mass is then $E_1 + E_2$ with no angular information required at all.

Now the numbers, at a typical photon energy of 60 GeV and a typical scalar energy of 200 GeV, for which $|\boldsymbol{p}_H|\,c = 156\,\mathrm{GeV}$ and $\cot(\theta/2) = 1.25$.

Energy term. ATLAS's parametrization Equation (113.13) gives $\sigma_E/E = 10\,\%/\sqrt{60} \oplus 0.7\,\% = 1.47\,\%$, so the two energy terms contribute $1.47\,\%/\sqrt{2} = 1.04\,\%$, that is $1.30\,\mathrm{GeV}/c^2$. CMS's Equation (113.14) gives $2.8\,\%/\sqrt{60} \oplus 12\,\%/60 \oplus 0.30\,\% = 0.51\,\%$ and hence $0.45\,\mathrm{GeV}/c^2$ — a factor of three better, which is the crystal calorimeter's entire justification.

Angular term with the wrong vertex. A photon leaves no track, so the collision vertex must be inferred. If it is taken at random from the twenty available, the error is the longitudinal spread of the luminous region, $\sigma_z \approx 5\,\mathrm{cm}$. At a calorimeter radius of 1.5 m this is an angular error $\sigma_\theta \approx 0.05\,\mathrm{m}/1.5\,\mathrm{m} = 0.033$, and $\tfrac{1}{2} \times 1.25 \times 0.033 = 2.1\,\%$, that is $2.6\,\mathrm{GeV}/c^2$. This is twice ATLAS's energy term and six times CMS's: with no vertex information the diphoton channel would not have worked in either detector.

Angular term with the right one. ATLAS's calorimeter pointing — the same shower measured in two samplings at different radii — locates the photon's origin to about 15 mm without any tracker information, giving $\sigma_\theta \approx 0.010$ and a contribution of $0.63\,\%$, or $0.79\,\mathrm{GeV}/c^2$: now subdominant. CMS instead identifies the vertex from the charged tracks recoiling against the diphoton system, and from the tracks of converted photons, reaching a comparable accuracy in the large majority of events.

Combining, the predicted resolutions are $1.5\,\mathrm{GeV}/c^2$ for ATLAS and $0.9\,\mathrm{GeV}/c^2$ for CMS; the values achieved in situ are $1.6\,\mathrm{GeV}/c^2$ to $1.7\,\mathrm{GeV}/c^2$ and $1.1\,\mathrm{GeV}/c^2$ to $1.5\,\mathrm{GeV}/c^2$ respectively [2186, 2192]. The excess over the test-bench numbers is upstream material, intercalibration between 7.6×10^4 channels, and the fact that a converted photon is measured worse than an unconverted one. The design calculation predicts the achieved resolution to within some tens of per cent, which is the level at which such an estimate can be trusted, and it identifies correctly which term dominates in which detector. □

Because the resolution and the signal fraction both vary strongly across the detector and across event topologies, neither experiment performed a single inclusive fit. ATLAS divided its diphoton candidates into ten categories in 2011 and fourteen in 2012, by conversion status, by pseudorapidity, and by the component of the diphoton transverse momentum orthogonal to the thrust axis; CMS used a multivariate classifier trained on the expected resolution and predicted signal fraction, plus a separate class tagged by two forward jets for vector-boson fusion. The gain from doing so is not a technicality.

Proposition 113.15 (Why categorizing helps, and by how much)**.** *Let a search be divided into independent categories i with expected signal s_i and background b_i. In the Gaussian limit the combined discovery significance is*

$$Z = \left(\sum_i \frac{s_i^2}{b_i} \right)^{1/2} \geq \frac{\sum_i s_i}{\left(\sum_i b_i\right)^{1/2}} = Z_{\mathrm{merged}}\,, \tag{113.18}$$

with equality if and only if s_i/b_i is the same in every category.

Rests on Equation (11.88)*, Definition 11.60, and Proposition 11.93.*

Derivation. *Derives Proposition 113.15.*

Let n_i be Poisson with mean $\mu s_i + b_i$ and let μ be the signal strength. The profile-likelihood statistic of Equation (11.88) in the Gaussian limit is $\tilde{q}_0 = \hat{\mu}^2/\sigma_\mu^2$, and the Fisher information of Definition 11.60 for μ at $\mu = 0$ is

$$\frac{1}{\sigma_\mu^2} = -\left\langle \frac{\partial^2 \ln L}{\partial \mu^2} \right\rangle \Bigg|_{\mu=0} = \sum_i \frac{s_i^2}{b_i}\,,$$

since $\ln L = \sum_i [n_i \ln(\mu s_i + b_i) - (\mu s_i + b_i)]$ up to constants, whose second derivative is $-\sum_i n_i s_i^2/(\mu s_i + b_i)^2$, with expectation $-\sum_i s_i^2/b_i$ at $\mu = 0$. For the median dataset under the signal hypothesis $\hat{\mu} = 1$, so $\tilde{q}_0 = \sum_i s_i^2/b_i$, and $Z = \sqrt{\tilde{q}_0}$ by Proposition 11.93. That is the first equality in Equation (113.18), and the merged case is the special case of a single category.

The inequality is Cauchy–Schwarz. Write $s_i = (s_i/\sqrt{b_i})\,\sqrt{b_i}$; then

$$\left(\sum_i s_i\right)^2 = \left(\sum_i \frac{s_i}{\sqrt{b_i}}\sqrt{b_i}\right)^2 \le \left(\sum_i \frac{s_i^2}{b_i}\right)\left(\sum_i b_i\right),$$

with equality exactly when the vectors $(s_i/\sqrt{b_i})$ and $(\sqrt{b_i})$ are proportional, that is when s_i/b_i is constant. Dividing by $\sum_i b_i$ gives Equation (113.18). □

Example 113.16 (The size of the effect)**.** Suppose 100 signal events sit on 10 000 background events, so the inclusive significance is $100/\sqrt{10\,000} = 1.0$. Suppose the events divide into a clean category holding 30 signal on 600 background and a dirty one holding 70 on 9400. Then

$$Z = \left(\frac{900}{600} + \frac{4900}{9400}\right)^{1/2} = \sqrt{2.02} = 1.42\,,$$

a gain of 42 % from partitioning the same events into two boxes. Nothing has been added; what has been used is the information that the two boxes have different signal purity, which a single fit throws away. This is why both collaborations spent as much effort on categorization as on the selection itself.

Background estimation. The two channels are treated in opposite ways, and the difference is instructive.

For $\gamma\gamma$ the background is large, smooth and not calculable: it comes from continuum diphoton production, from photon plus jet with the jet misidentified, and from dijets with two misidentifications, in proportions that no simulation predicts to the accuracy required. It is therefore taken from the data. An analytic function — an exponential, or a polynomial in a Bernstein basis — is fitted to the observed mass spectrum over a wide range simultaneously with a signal peak of fixed shape, so the background normalization and shape are determined by the sidebands and the fit returns the signal yield. The danger of the method is that a wrong functional form can absorb a real signal or manufacture a spurious one, and it is controlled by the *spurious-signal test*: the chosen function is fitted to a background-only sample far larger than the data, and the fitted "signal" yield it returns is required to be a small fraction of the statistical uncertainty on the real measurement. The residual bias is then carried as a systematic uncertainty. The whole procedure is fixed, function and categories and all, before the signal region is examined.

For 4ℓ the background is small and partly calculable. The irreducible part is continuum ZZ^* production, taken from simulation and normalized where the simulation can be checked — on the Z peak itself, in the same final state, where the rate is thousands of times higher and the physics is the same. The reducible part, Z plus jets and $t\bar{t}$ with jets faking leptons, is estimated from control regions in the data defined by inverting the lepton identification or isolation requirements. Both estimates are validated in sidebands before use.

113.3.3 Blinding and the statistical procedure

The statistical machinery is derived in Chapter 11 and is used here as derived. This subsection states which pieces are in force; Section 113.5.1 applies them to the published numbers.

The test statistic is the profile-likelihood ratio of Definition 11.70. The parameter of interest is a single signal strength μ, defined so that $\mu = 0$ is the background-only hypothesis

and $\mu = 1$ the Standard Model prediction at the mass being tested; every systematic uncertainty — energy scales, resolutions, luminosity, background shape, theoretical cross sections — enters as a nuisance parameter, constrained by an auxiliary measurement and profiled out. The one-sided discovery statistic is $\tilde{q}_0$ of Equation (11.88), which refuses to count a deficit as evidence for a signal, and the significance is $Z = \sqrt{\tilde{q}_0}$ by Proposition 11.93. That identity is what allows a significance to be quoted without simulating an ensemble of experiments, and it rests on Wilks' theorem (Theorem 11.88) applied to the profile statistic (Corollary 11.89); the asymptotic formulae in the form the collaborations use them are [2172].

Exclusion uses the same machinery reversed, with the modified frequentist CL_s criterion [2036] — the ratio of the two tail probabilities rather than the numerator alone — so that a downward background fluctuation cannot exclude a signal the experiment had no power to see (Remark 11.95). This is the construction behind the LEP limit of Proposition 113.7 and behind every exclusion plot in the 2012 papers.

Two conventions must be stated because mixing them shifts a significance by more than a tenth of a standard deviation (Remark 11.94). The p-values quoted below are *one-sided*, so that $Z = 5$ corresponds to $p = 2.87 \times 10^{-7}$; and the significances quoted by the collaborations as "local" are computed at a mass fixed in advance, which no search does. The correction is Section 113.5.1.

Blinding. None of this is worth anything if the analysis was adjusted after the data were seen. Both collaborations therefore fixed the entire chain — trigger, selection, categorization, background parametrization, systematic model and statistical procedure — using simulation and the mass sidebands, with the signal region hidden, and unblinded only afterwards. The reason is not procedural hygiene but arithmetic: the p-value of Definition 11.75 is the probability, under the null, of a result at least as extreme as the one observed *by a test specified in advance*. A cut adjusted after seeing the excess makes the number it produces uninterpretable, because the effective number of hypotheses tested is then unknown and unboundable. Blinding is what makes the five-sigma threshold mean the thing it appears to mean, and it is the reason the 2012 result was believed immediately rather than after a decade of argument.

113.4 Observations and data

113.4.1 The diphoton channel

What the data show is a narrow peak sitting on a steeply falling smooth continuum. The continuum falls by roughly an order of magnitude between $m_{\gamma\gamma}c^2 = 100\,\mathrm{GeV}$ and $160\,\mathrm{GeV}$ and has no structure in it; the peak is one resolution width across, of order $1.5\,\mathrm{GeV}/c^2$ by Proposition 113.14, and stands a few per cent above the continuum in the best categories and well under one per cent in the worst. Nothing about the picture is dramatic; what makes it a measurement is the arithmetic done to it.

ATLAS reported a local significance of 4.5 standard deviations near $m_H c^2 = 126.5\,\mathrm{GeV}$ in this channel alone, against 2.5 expected for a Standard Model scalar at that mass [2186]; CMS reported 4.1 against 2.8 expected near $125\,\mathrm{GeV}$ [2192]. Two things about those pairs of numbers deserve saying plainly, because they are usually skipped.

First, the expected significances — 2.5 and 2.8 — are close to the crude estimate $s/\sqrt{b} \approx 2$ obtained in Proposition 113.12 from cross section, branching fraction, efficiency and background. That agreement is a check on the whole chain: an experiment whose expected sensitivity disagreed by a factor of two with a back-of-envelope calculation from published inputs would have something wrong with it.

Second, both observed values exceed both expected values. The fitted diphoton signal strength was correspondingly high in both experiments, around 1.8 with an uncertainty near

0.5 in ATLAS and similarly above unity in CMS. That is an upward fluctuation of about 1.5 standard deviations, present in the channel that dominates the combination — exactly the configuration in which an unblinded, after-the-fact analysis would be untrustworthy, and exactly why the selection was frozen beforehand. The later, much larger datasets brought the diphoton signal strength back to unity, which is the expected behaviour of a fluctuation and confirms the 2012 diagnosis.

One structural fact is available from this channel with no further analysis: a resonance that decays to two real photons cannot have spin 1. That is the Landau–Yang argument [917, 975], derived in Theorem 104.60 and again, in the vector form, in the derivation accompanying Phenomenon 113.22. The discovery channel therefore excluded one of the four spin hypotheses on the day it was announced, without any angular measurement.

113.4.2 The four-lepton channel

$H \to ZZ^* \to 4\ell$ is the opposite kind of measurement. The final state is fully reconstructed — four charged leptons, all their momenta measured in the tracker — so the mass resolution is set by lepton momentum resolution alone and is of order $2\,\mathrm{GeV}/c^2$, and the background is small enough that the signal fraction exceeds unity in a narrow window. What it does not have is events: the arithmetic of Proposition 113.12 gives about twenty-five produced decays in the ATLAS dataset and fewer than ten surviving selection.

Both experiments observed roughly a dozen candidates in a window $10\,\mathrm{GeV}/c^2$ wide around $125\,\mathrm{GeV}/c^2$, against an expected background of about five, and reported local significances of 3.6 standard deviations (ATLAS, against 2.7 expected) and 3.2 (CMS, against 3.8 expected) [2186, 2192]. With numbers this small the Gaussian approximation is doing real work, and the collaborations used the asymptotic formulae [2172] whose accuracy at such yields is itself part of what Chapter 11 establishes; the CMS value fell below its own expectation, which is an ordinary downward fluctuation of a Poisson count.

The complementarity is the point of quoting both channels. A high-rate, high-background channel and a low-rate, low-background one fail in different ways: a mismodelled background shape would move the diphoton result and not the four-lepton one, and a mismeasured lepton efficiency would do the reverse. Their masses agreed to within the resolutions, and their signal strengths agreed with each other and with unity. Two supporting channels, $H \to WW^* \to \ell\nu\ell\nu$ and $H \to \tau^+\tau^-$, contributed sensitivity but no mass information at all: the neutrinos in the first and in the tau decays of the second make full reconstruction impossible, so those channels report an excess of events with the expected kinematics and not a resonance. They are evidence for a signal and not evidence for a particle, and both papers treat them as such.

113.4.3 Combined significance, 4 July 2012

Combining channels, ATLAS reported a local significance of 5.9 standard deviations at $m_H c^2 = 126.0\,\mathrm{GeV}$ against 4.9 expected, and CMS 5.0 at 125.5 GeV against 5.8 expected [2186, 2192]. The masses were

$$m_H c^2\big|_{\mathrm{ATLAS}} = 126.0\,\mathrm{GeV} \pm 0.4\,\mathrm{GeV} \pm 0.4\,\mathrm{GeV}\,, \quad m_H c^2\big|_{\mathrm{CMS}} = 125.3\,\mathrm{GeV} \pm 0.4\,\mathrm{GeV} \pm 0.5\,\mathrm{GeV}\,, \tag{113.19}$$

the first uncertainty statistical and the second systematic, and the signal strengths relative to the Standard Model prediction were $\mu = 1.4 \pm 0.3$ and $\mu = 0.87 \pm 0.23$ respectively.

The two collaborations were blind to each other's results until the seminar of 4 July 2012. That is not a detail of etiquette. Two measurements that agree are evidence in proportion to how independent they are, and these two shared only the accelerator: different magnets, different calorimeter technologies, different reconstruction software, different people, and no

exchange of intermediate results. An error common to both would have to originate in the beam or in the theory of the Standard Model, not in either apparatus.

Phenomenon 113.17 (A new boson near 125 GeV)**.** Two experiments at the Large Hadron Collider, blind to each other's results, reported on the same day a narrow resonance in the diphoton and four-lepton final states, with local significances of 5.9 and 5.0 standard deviations respectively [2186, 2192]. The combined mass from the two high-resolution channels is

$$m_H c^2 = (125.09 \pm 0.24)\,\mathrm{GeV}\,, \tag{113.20}$$

[2242], and the present world average is $(125.20 \pm 0.11)\,\mathrm{GeV}$, that is $m_H = 2.2319 \times 10^{-25}\,\mathrm{kg}$ [2427]. The particle is electrically neutral, decays to two photons and to two Z bosons, and is produced at a rate compatible with the single scalar that Phenomenon 113.6 requires.

Rests on Phenomenon 113.6, Equation (113.19)*, and Proposition 113.5.*

Derivation of the consistency checks. *Derives Phenomenon 113.17.*

Three separate agreements have to hold, and each is an arithmetic statement that could have failed.

The two experiments agree. Adding the errors of Equation (113.19) in quadrature gives $(126.0 \pm 0.6)\,\mathrm{GeV}$ for ATLAS and $(125.3 \pm 0.6)\,\mathrm{GeV}$ for CMS, where $\sqrt{0.4^2 + 0.4^2} = 0.57$ and $\sqrt{0.4^2 + 0.5^2} = 0.64$. The difference is $0.7\,\mathrm{GeV}$ with a combined uncertainty $\sqrt{0.57^2 + 0.64^2} = 0.86$, that is 0.8 standard deviations. Two independent instruments measuring the same mass agreed at once.

The 2012 numbers survive. The joint ATLAS and CMS combination of the full Run-1 data [2242] gives $125.09\,\mathrm{GeV} \pm 0.21\,\mathrm{GeV}\,(\mathrm{stat}) \pm 0.11\,\mathrm{GeV}\,(\mathrm{syst})$, and adding those in quadrature is $(125.09 \pm 0.24)\,\mathrm{GeV}$, which is Equation (113.20). The present world average is $(125.20 \pm 0.11)\,\mathrm{GeV}$ [2427], higher by $0.11\,\mathrm{GeV}$.

These two are successive best values, not two measurements. The world average contains the Run-1 combination, so the two are strongly correlated, their errors may not be added in quadrature, and the difference cannot be converted into a number of standard deviations without an estimate of the correlation that neither publication supplies. What can be said without one is the thing worth saying: twelve years and two orders of magnitude in integrated luminosity moved the central value by $0.11\,\mathrm{GeV}$, less than half the 2015 uncertainty, while more than halving that uncertainty. A shift of that size is what an unbiased measurement being refined looks like; a shift of several GeV would have been a retraction.

The rate is the predicted one. The signal strengths $\mu = 1.4 \pm 0.3$ and $\mu = 0.87 \pm 0.23$ are ratios of the observed rate to the Standard Model prediction computed from Proposition 113.5 with the mass now fixed. Their weighted mean is

$$\bar{\mu} = \frac{1.4/0.3^2 + 0.87/0.23^2}{1/0.3^2 + 1/0.23^2} = 1.06\,, \qquad \sigma_{\bar{\mu}} = \left(\frac{1}{0.3^2} + \frac{1}{0.23^2}\right)^{-1/2} = 0.18\,,$$

so $\bar{\mu} = 1.06 \pm 0.18$, consistent with unity. This is the check that a bump alone would not supply: a narrow resonance at $125\,\mathrm{GeV}/c^2$ produced at a hundredth of the predicted rate, or a hundred times it, would be a new particle and would not be this one.

What the combination does *not* establish is treated in Section 113.5.1: the quoted significances are local, and the mass was not nominated in advance. □

113.4.4 The measured numbers

Table 113.3 collects the quantities of the discovery itself and Table 113.4 what has been measured since; Tables 113.5 and 113.6 collect the coupling test, which is the measurement that distinguishes this scalar from any other. All four sit here, under this rubric, rather than beside the sections that use them, so that a reader looking for the chapter's numbers

finds them in one place. Rows that record a run setting carry no uncertainty because they define the conditions rather than measure anything; every row that reports a measured or derived quantity carries one, and a derived row names the equation that produced it.

Remark 113.18 (Where each number came from, and where it did not). The masses in Table 113.5 and the world-average scalar mass and width in Table 113.4 are taken from the Particle Data Group's own machine-readable 2024 release, stored with this treatise as evidence [2427]; the conversions to kg and J use the exact 2019 SI values of e and c [2431]. The 2012 significances, signal strengths and per-channel numbers are the collaborations' published values [2186, 2192] and are not recomputed here. The rows marked derived are computed in this chapter from the rows above them, with the arithmetic shown; the produced-scalar count inherits the roughly ten per cent uncertainty of the predicted production cross sections, which is why it is quoted to three figures and not to five. The predicted reduced couplings are exact functions of the masses and E_v and carry no theoretical uncertainty at tree level.

113.5 Interpretation

113.5.1 The look-elsewhere effect

Every significance in Table 113.3 labelled *local* is the answer to a question nobody asked: what is the probability of a fluctuation this large *at a mass named in advance*? No search names a mass in advance. Both experiments scanned a range and reported the largest excess found in it, and the probability that *some* point of a range fluctuates upward is larger — here by one to two orders of magnitude — than the probability that a nominated point does. The correction is derived in Section 11.6 and is applied here to the published numbers.

Proposition 113.19 (The trials factors of the 2012 result)**.** *With the one-sided convention of Proposition 11.93, the ATLAS local significance $Z_{loc} = 5.9$ corresponds to $p_{loc} = 1.82 \times 10^{-9}$, and the two global significances the collaboration quotes correspond to trials factors*

$$N_{tr}\big|_{110--150\,\mathrm{GeV}} = 31.9\,, \qquad N_{tr}\big|_{110--600\,\mathrm{GeV}} = 93.4\,. \tag{113.21}$$

The first is reproduced to within one per cent by counting resolution elements; the second is not, and the reason is physical.

Rests on Definitions 11.92 and 11.96 and Proposition 113.14.

Derivation. *Derives Proposition 113.19.*

By Definition 11.92, $p = 1 - \Phi(Z)$, so $p_{\mathrm{loc}} = 1 - \Phi(5.9) = 1.8175 \times 10^{-9}$, $1 - \Phi(5.3) = 5.7901 \times 10^{-8}$ and $1 - \Phi(5.1) = 1.6983 \times 10^{-7}$. The trials factor is the ratio $N_{\mathrm{tr}} = p_{\mathrm{glob}}/p_{\mathrm{loc}}$ of Definition 11.96, giving $5.7901 \times 10^{-8}/1.8175 \times 10^{-9} = 31.86$ for the narrow range and $1.6983 \times 10^{-7}/1.8175 \times 10^{-9} = 93.44$ for the wide one, which is Equation (113.21).

The narrow range, from resolution elements. The naive estimate of Equation (11.93) treats the scan as $N = \Delta m/\delta m$ independent tests. With $\Delta mc^2 = 40\,\mathrm{GeV}$ and the diphoton resolution $\delta mc^2 = 1.7\,\mathrm{GeV}$ of Proposition 113.14 — the channel that dominates the combination — this is $N = 23.5$, whence $p_{\mathrm{glob}} \simeq 23.5 \times 1.8175 \times 10^{-9} = 4.27 \times 10^{-8}$ and $Z_{\mathrm{glob}} = \Phi^{-1}(1 - p_{\mathrm{glob}}) = 5.36$. ATLAS quotes 5.3. The agreement is closer than the method deserves and should not be over-read — Remark 11.99 says why the independence assumption has no general justification — but it does show that the published correction is the size a reader can reconstruct from the resolution and the window.

The wide range, where the same estimate fails. Repeating it over $110 - -600\,\mathrm{GeV}$ gives $N = 490\,\mathrm{GeV}/1.7\,\mathrm{GeV} = 288$ and $Z_{\mathrm{glob}} = 4.88$, whereas ATLAS quotes 5.1, corresponding to $N_{\mathrm{tr}} = 93$. The naive count is too large by a factor of three, and the reason is that $\delta m = 1.7\,\mathrm{GeV}/c^2$ is not the correlation length of the test statistic above about $180\,\mathrm{GeV}/c^2$.

Table 113.3: The discovery itself: the datasets it rests on, the prior constraints it had to clear, and the significances reported in 2012. Integrated luminosities are given in $1/\mathrm{m}^2$ with the customary inverse-femtobarn figure alongside (Notation 113.1); one inverse femtobarn is $1 \times 10^{43}/\mathrm{m}^2$. Significances are pure numbers. The global significances are the collaboration's own look-elsewhere corrections over the mass range named; Section 113.5.1 reconstructs the trials factors they imply. What was measured afterwards is Table 113.4.

Quantity	Value	Year	Source
ATLAS integrated luminosity, 7 TeV (setting)	$4.8 \times 10^{43}/\mathrm{m}^2$, i.e. 4.8 inverse femtobarns	2011	[2186]
ATLAS integrated luminosity, 8 TeV (setting)	$5.8 \times 10^{43}/\mathrm{m}^2$	2012	[2186]
CMS integrated luminosity, 7 TeV (setting)	$5.1 \times 10^{43}/\mathrm{m}^2$	2011	[2192]
CMS integrated luminosity, 8 TeV (setting)	$5.3 \times 10^{43}/\mathrm{m}^2$	2012	[2192]
Scalars produced in the ATLAS dataset	2.12×10^5 (derived from the two rows above and the cross sections of Section 113.3.1; good to the 10 % accuracy of those predictions)	2012	Proposition 113.12
LEP lower limit, 95 % confidence	$m_H c^2 > 114.4\,\mathrm{GeV}$	2003	[2045]
Tevatron excess in VH, $H \to b\bar{b}$	local significance ≈ 3, global ≈ 2.5	2012	[2187]
ATLAS local significance, $\gamma\gamma$	4.5 observed, 2.5 expected	2012	[2186]
ATLAS local significance, $ZZ^* \to 4\ell$	3.6 observed, 2.7 expected	2012	[2186]
ATLAS combined local significance	5.9 observed, 4.9 expected	2012	[2186]
ATLAS global significance, 110–150 GeV range	5.3	2012	[2186]
ATLAS global significance, 110–600 GeV range	5.1	2012	[2186]
CMS local significance, $\gamma\gamma$	4.1 observed, 2.8 expected	2012	[2192]
CMS local significance, $ZZ^* \to 4\ell$	3.2 observed, 3.8 expected	2012	[2192]
CMS combined local significance	5.0 observed, 5.8 expected	2012	[2192]

Table 113.4: What has been measured since. The 2012 masses carry a statistical and a systematic error, quoted separately by the collaborations and combined in quadrature here, with the arithmetic shown in the derivation accompanying Phenomenon 113.17. The two 2012 masses and the later combinations are successive best values of one quantity and not independent measurements to be averaged — the world average already contains the earlier data. Signal strengths are pure numbers. The width, the lifetime and the electroweak scale are derived quantities, and each is given in SI beside its customary form.

Quantity	Value	Year	Source
ATLAS mass	$126.0\,\mathrm{GeV}/c^2$, ± 0.4 (stat) ± 0.4 (syst); $(126.0 \pm 0.6)\,\mathrm{GeV}$ combined	2012	[2186]
CMS mass	$125.3\,\mathrm{GeV}/c^2$, ± 0.4 (stat) ± 0.5 (syst); $(125.3 \pm 0.6)\,\mathrm{GeV}$ combined	2012	[2192]
ATLAS signal strength μ at $126\,\mathrm{GeV}/c^2$	1.4 ± 0.3	2012	[2186]
CMS signal strength μ at $125.5\,\mathrm{GeV}/c^2$	0.87 ± 0.23	2012	[2192]
Weighted mean signal strength	1.06 ± 0.18 (derived)	2012	Phenomenon 113.17
ATLAS and CMS combined mass	$(125.09 \pm 0.24)\,\mathrm{GeV}/c^2$	2015	[2242]
World-average mass	$(125.20 \pm 0.11)\,\mathrm{GeV}/c^2$, i.e. $(2.2319 \pm 0.0020) \times 10^{-25}\,\mathrm{kg}$	2024	[2427]
Total width, as an energy $\hbar\Gamma_H$	$3.7\,\mathrm{MeV}$, $+1.9$, -1.4; in SI $5.93 \times 10^{-13}\,\mathrm{J}$	2024	[2427]
Mean lifetime $\tau = \hbar/(\hbar\Gamma_H)$	$1.78 \times 10^{-22}\,\mathrm{s}$ (derived), with $+1.08$ and -0.60 on the same power of ten; decay length $c\tau = 5.3 \times 10^{-14}\,\mathrm{m}$, with $+3.2$ and -1.8 likewise; both errors are the width's $+51\,\%$, $-38\,\%$ propagated	2024	Phenomenon 113.26
Electroweak scale E_v	$246.220\,\mathrm{GeV}$, i.e. $3.944\,88 \times 10^{-8}\,\mathrm{J}$ (derived from the muon lifetime; the fractional uncertainty, 3×10^{-7}, is beyond the digits shown)	2013	[2225]

Table 113.5: The coupling test. The predicted reduced coupling is mc^2/E_v for a fermion and $M_V c^2/E_v$ for a weak boson, by Equation (104.70); it is a pure number, it contains no free parameter, and its uncertainty here is propagated from the measured mass alone, since E_v is known to three parts in 10^7. Masses are the PDG 2024 values, which are also the masses in kg shown alongside; each kilogram value carries the same fractional uncertainty as the GeV value beside it, written in brackets on the last digits quoted. The fourth column records what the ten-year coupling maps report: every measured coupling modifier κ_i, defined so that $\kappa_i = 1$ reproduces the predicted value, is consistent with unity, and the entry gives the approximate precision, not a central value — the papers cited should be consulted for those and for their correlations. The last column carries the year of each of the two measured entries in the row, the mass first. The partners whose couplings are not yet measured are collected separately in Table 113.6.

Partner	Mass [2427]	Predicted reduced coupling	Measured κ [2394, 2410]	Year
W	$(80.369 \pm 0.013)\ \mathrm{GeV}/c^2$; $(1.432\,708 \pm 0.000\,023) \times 10^{-25}$ kg	$0.326\,411 \pm 0.000\,053$	1 to about 10 %	2024; 2022
Z	$(91.1880 \pm 0.0020)\ \mathrm{GeV}/c^2$; $(1.625\,573\,8 \pm 0.000\,003\,6) \times 10^{-25}$ kg	$0.370\,351\,7 \pm 0.000\,008\,1$	1 to about 10 %	2024; 2022
t	$(172.57 \pm 0.29)\ \mathrm{GeV}/c^2$; $(3.076\,34 \pm 0.000\,52) \times 10^{-25}$ kg	$0.700\,88 \pm 0.001\,18$	1 to about 10 %; $t\bar{t}H$ observed [2320]	2024; 2022
b	$(4.183 \pm 0.007)\ \mathrm{GeV}/c^2$; $(7.4569 \pm 0.0012) \times 10^{-27}$ kg	$0.016\,989 \pm 0.000\,028$	1 to about 10 %	2024; 2022
τ	$(1.776\,93 \pm 0.000\,09)\ \mathrm{GeV}/c^2$; $(3.167\,665\,4 \pm 0.000\,001\,6) \times 10^{-27}$ kg	$0.007\,216\,8 \pm 0.000\,000\,4$	1 to about 10 %	2024; 2022
μ	$(0.105\,658\,375\,5 \pm 0.000\,000\,002\,3)\ \mathrm{GeV}/c^2$; $(1.883\,531\,63 \pm 0.000\,000\,04) \times 10^{-28}$ kg	$4.291\,218 \times 10^{-4}$	1 to about 20 %; decay observed at three standard deviations [2390]	2024; 2021

Table 113.6: What the coupling test does *not* yet reach. The predicted reduced coupling is defined as in Table 113.5; the difference is that no entry here has a measured κ. The charm coupling is bounded but not observed, the first-generation and strange couplings are far below present sensitivity, and the self-coupling — the one number that probes the shape of the potential rather than its minimum — is constrained only to a wide interval. The last column carries the year of the mass and of the bound respectively.

Partner	Mass [2427]	Predicted reduced coupling	Status [2394, 2410]	Year
c	$(1.273 \pm 0.005)\,\mathrm{GeV}/c^2$; $(2.2693 \pm 0.0089) \times 10^{-27}\,\mathrm{kg}$	$0.005\,170 \pm 0.000\,020$	not measured; only bounded	2024; —
e, u, d, s	below $0.1\,\mathrm{GeV}/c^2$	below 4×10^{-4}	not measured	2024; —
Self-coupling κ_λ	—	1 by Proposition 104.63	allowed anywhere from about -1 to $+7$	—; 2022

Two effects lengthen it: the channels that carry the sensitivity at high mass ($ZZ \to 4\ell$ with degrading resolution, and WW, which has no mass peak at all) are much coarser than the diphoton channel; and the predicted width of a Standard Model scalar grows rapidly once the WW and ZZ thresholds open, reaching tens of GeV by $500\,\mathrm{GeV}/c^2$, so neighbouring hypotheses in that region are nearly the same hypothesis. Fewer effectively independent tests, hence a smaller trials factor.

The upcrossing form. The honest machinery is Theorem 11.101 with the level dependence of Theorem 11.103, which for large Z_{loc} gives Equation (11.98). That statement is an *upper bound* — it descends from Davies' bound — and reads $N_{\mathrm{tr}} \le 1 + \sqrt{2\pi}\,\langle N_{u_0}\rangle\,\mathrm{e}^{u_0/2} Z_{\mathrm{loc}}$ with $u_0 = 1$, so inverting it about $\langle N_{u_0}\rangle$ turns it into a *lower* bound on the upcrossing count and not into a value. Since $\sqrt{2\pi}\,\mathrm{e}^{1/2} = 4.133$ and $Z_{\mathrm{loc}} = 5.9$, the published trials factors imply mean numbers of one-sigma upcrossings of at least

$$\langle N_1\rangle\,\big|_{110--150\,\mathrm{GeV}} \ge \frac{31.86 - 1}{4.133 \times 5.9} = 1.27\,, \qquad \langle N_1\rangle\,\big|_{110--600\,\mathrm{GeV}} \ge 3.79\,.$$

Both floors are of order unity, which is exactly the regime in which the calibrate-and-extrapolate procedure works: a mean upcrossing count near one can be measured from a few hundred simulated background-only experiments, whereas the tail probability it is used to compute, 5×10^{-8}, could not be sampled by any feasible number of them. And the two floors stand in the ratio 2.99, far below the ratio of the two mass ranges, 12.25 — which is the quantitative form of the statement that the correlation length grows with mass, provided the bound is comparably tight in both ranges, as it is, both being evaluated at the same Z_{loc}. □

Remark 113.20 (What survives the correction, honestly). Applied to ATLAS the correction moves 5.9 to 5.3 over the range where a Standard Model scalar was still allowed, and to 5.1 over the full range scanned. Both remain above the conventional threshold. Applied to CMS it does not: repeating the resolution-element estimate with a window $110 - -145\,\mathrm{GeV}$ and a diphoton resolution near $1.4\,\mathrm{GeV}/c^2$ gives $N = 25$, hence $p_{\mathrm{glob}} \simeq 25 \times 2.87 \times 10^{-7} = 7.2 \times 10^{-6}$ and $Z_{\mathrm{glob}} \simeq 4.3$ — an estimate made here, not a published number, and one that puts the CMS result below five standard deviations globally.

That is the correct state of affairs to record, and it is worth being blunt about it. The 2012 discovery does not rest on either experiment individually crossing an arbitrary threshold. It rests on two independent experiments finding an excess of the same size at the same mass in the same two channels on the same day, with the rate the theory predicts — the three checks in the derivation accompanying Phenomenon 113.17. A joint probability is not the product of two marginal ones when the alternatives are correlated, so no single combined number is quoted here; what can be said is that the agreement of two blind measurements is evidence of a kind that no amount of significance in one of them supplies.

Three further cautions, all of them from Chapter 11 and all of them routinely mishandled in secondary accounts. A p-value is the probability of the data given the null, never the probability of the null given the data (Remark 11.77); "a one-in-3.5-million chance that it is a fluctuation" is a mistranslation of $Z = 5$, not a paraphrase of it. The one-sided and two-sided conventions differ by 0.14 in Z at $q = 25$ (Remark 11.94), which is enough to move a result across the threshold, so the convention must be stated — it is one-sided throughout this chapter. And a significance quoted without the range searched is a local number wearing a global name (Remark 11.108).

113.5.2 The mass

The measured mass is now

$$m_H c^2 = (125.20 \pm 0.11)\,\mathrm{GeV}\,, \qquad m_H = (2.2319 \pm 0.0020) \times 10^{-25}\,\mathrm{kg}\,, \tag{113.22}$$

a fractional precision of 8.8×10^{-4} [2427]. Almost all of it comes from the two high-resolution channels, and the systematic error is dominated by one thing: the absolute energy scale of the electromagnetic calorimeters.

That scale is not known a priori. A crystal's light yield, a liquid-argon gap's response and the material in front of both are known to a per cent at best, which would give a mass uncertainty of over $1\,\mathrm{GeV}/c^2$ — ten times the quoted error. What fixes it is a calibration source inside the same detector: the $Z \to e^+e^-$ peak, tens of millions of events at $(91.1880 \pm 0.0020)\,\mathrm{GeV}/c^2$ whose mass is known from LEP to two parts in 10^5 [2104, 2427]. Reconstructing it with the experiment's own electrons and demanding that the reconstructed peak sit at the LEP value calibrates the energy scale *in situ*, cell by cell and category by category. The residual uncertainty is then the extrapolation from the electron scale at 45 GeV to the photon scale at 60 GeV — a small step, and it is small because the calibration source and the measurement are in the same instrument and differ by less than a factor of two in energy.

Remark 113.21 (Fixing the mass converts the rest into tests). m_H was the last free parameter of the Standard Model to be measured. Its value fixes the quartic coupling by Equation (113.3),

$$\hat{\lambda} = \frac{1}{2}\left(\frac{m_H c^2}{E_v}\right)^2 = \frac{1}{2}\left(\frac{125.20}{246.220}\right)^2 = \frac{1}{2}\,(0.508\,49)^2 = 0.129\,28\,,$$

in agreement with Equation (104.38), and with it the curvature parameter $\hbar\mu c = m_H c^2/\sqrt{2} = 88.53\,\mathrm{GeV}$. From that moment every other statement the theory makes about the scalar — every branching fraction, every production rate, every self-coupling — is a prediction with no adjustable parameter, and each subsequent measurement is a test rather than a determination. The rest of this section, and the whole of Section 113.6, is the score.

113.5.3 Spin and parity

Four hypotheses were on the table in 2012: $J^P = 0^+$, the prediction; 0^-, a pseudoscalar; $1^\pm$; and 2^+, a graviton-like tensor. One of them was excluded by the discovery channel itself.

Phenomenon 113.22 (The new boson has spin zero and even parity)**.** The angular distributions of the four-lepton and diphoton final states are those of a $J^P = 0^+$ resonance; the 0^-, $1^\pm$ and 2^+ alternatives are excluded at high confidence by both experiments [2201, 2217]. The exclusion of spin 1 requires no angular analysis at all: it follows from the mere existence of the diphoton decay.

Rests on Theorem 104.60.

Derivation. *Derives Phenomenon 113.22.*

This is the Landau–Yang argument [917, 975], given in the vector form here and in the amplitude form as Theorem 104.60. Work in the rest frame of the parent, where the two photons carry momenta $\pm\boldsymbol{k}$ and transverse polarization vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, with $\boldsymbol{e}_1 \cdot \boldsymbol{k} = \boldsymbol{e}_2 \cdot \boldsymbol{k} = 0$. A parent of spin 1 is described by a polarization vector $\boldsymbol{J}$, and the decay amplitude must be a rotational scalar, linear in each of $\boldsymbol{e}_1$, $\boldsymbol{e}_2$ and $\boldsymbol{J}$. The only such scalars available are

$$(\boldsymbol{e}_1 \cdot \boldsymbol{e}_2)\,(\boldsymbol{J} \cdot \boldsymbol{k}) \;, \quad \boldsymbol{J} \cdot (\boldsymbol{e}_1 \times \boldsymbol{e}_2) \;, \quad (\boldsymbol{J} \cdot \boldsymbol{e}_1)\,(\boldsymbol{k} \cdot \boldsymbol{e}_2) + (\boldsymbol{J} \cdot \boldsymbol{e}_2)\,(\boldsymbol{k} \cdot \boldsymbol{e}_1) \;.$$

The list is complete, and transversality is what makes it so. Since $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are both perpendicular to $\boldsymbol{k}$, the vector $\boldsymbol{e}_1 \times \boldsymbol{e}_2$ is parallel to $\boldsymbol{k}$; a further candidate such as $(\boldsymbol{J} \cdot \boldsymbol{k})\,[\boldsymbol{k} \cdot (\boldsymbol{e}_1 \times \boldsymbol{e}_2)]$ is therefore the second structure above multiplied by $|\boldsymbol{k}|^2$, and not an independent one.

The photons are identical bosons, so the amplitude must be unchanged under the exchange $(\boldsymbol{e}_1, \boldsymbol{k}) \leftrightarrow (\boldsymbol{e}_2, -\boldsymbol{k})$. The first structure changes sign, because $\boldsymbol{J} \cdot \boldsymbol{k}$ does; the second changes sign, because the cross product does; the third vanishes identically by transversality. No amplitude survives. A spin-1 particle therefore cannot decay into two real photons, whatever its parity, and the observed diphoton channel excludes $J = 1$ on its own.

The argument uses rotational invariance, Bose statistics and transversality and nothing else. It does not use parity, so it kills 1^+ and 1^- together; and it fails for two *virtual* photons or for a massive vector in the final state, both of which have a longitudinal polarization and therefore a third structure. The angular analysis is needed only to separate 0^+ from 0^- and from the spin-2 hypotheses. □

That separation uses the four-lepton final state, which is fully reconstructed and therefore carries its complete angular information: five angles — the production polar angle, the two polar angles of the lepton pairs in their parent frames, and two azimuths, one of them the angle Φ between the planes of the two lepton pairs — plus the two lepton-pair masses. The discriminating variable is Φ. A scalar couples to the two Z bosons through the structure $\varepsilon_1^* \cdot \varepsilon_2^*$ of Equation (104.69), which produces a distribution containing a $\cos\Phi$ term; a pseudoscalar can only couple through the antisymmetric structure $\epsilon_{\mu\nu\rho\sigma}\varepsilon_1^{*\mu}\varepsilon_2^{*\nu}k_1^\rho k_2^\sigma$, which has no $\cos\Phi$ term and a $\cos 2\Phi$ term of the opposite sign; and a spin-2 state produces polar distributions that neither can. Both collaborations built a likelihood in the seven variables and reported the alternatives disfavoured at high confidence, CMS excluding the pseudoscalar at above 99 % and ATLAS reaching a comparable conclusion [2201, 2217]. The scalar hypothesis was not merely preferred: the alternatives were rejected.

The three structures named in the previous paragraph are quoted, not derived. Landau–Yang is derived above and disposes of $J = 1$ from the diphoton channel alone; the separation of 0^+ from 0^- and from spin 2 is the load-bearing half of Phenomenon 113.22 and this chapter does not carry it.

Derivation pending #201

The azimuthal distributions that separate 0^+ *from* 0^- *and from the spin-2 hypotheses in* $H \to ZZ^* \to 4\ell$*. What is owed is the decay amplitude for each hypothesis — the scalar contraction of the two boson polarizations, the Levi-Civita contraction that a pseudoscalar is restricted to, and the tensor couplings — projected onto the two lepton pairs and integrated to give the distribution in the angle between their decay planes, exhibiting the* $\cos\Phi$ *term the scalar produces, its absence for the pseudoscalar, and the* $\cos 2\Phi$ *term of opposite sign that the pseudoscalar carries. It belongs in Appendix A beside the vector-boson vertices of the electroweak chapter, whose loop-induced counterparts are already recorded as owed there.*

Remark 113.23 (A CP-odd admixture is bounded, not excluded). Excluding a pure 0^- state is not the same as establishing that the state is purely 0^+. What the data constrain is the fraction of the rate arising from a CP-odd contribution, and the constraint is weak in a specific and instructive way: in the HZZ channel any CP-odd coupling would have to be generated by a loop, so it is suppressed by construction and the channel's apparent sensitivity overstates the test. The sharper probe is $H \to \tau^+\tau^-$, where a CP-odd Yukawa coupling enters at tree level and shows up as a phase in the correlation between the tau decay planes; there the mixing angle is currently bounded to a few tens of degrees.

The question is not academic. The Standard Model's CP violation is quantitatively insufficient to generate the observed baryon asymmetry by many orders of magnitude (Section 111.8.2), and the electroweak transition it predicts is a crossover, which supplies no departure from equilibrium either (Remark 104.65). A CP-odd component in the scalar sector is one of the few places where new CP violation could hide and be found with existing machines. It has not been found; it has been bounded.

113.5.4 Couplings against mass

Everything so far would be satisfied by a scalar that had nothing to do with mass generation. The measurement that closes the argument is the pattern of couplings.

Phenomenon 113.24 (Couplings proportional to mass)**.** The measured couplings of the new boson to the W, the Z, the top, the bottom and the tau, together with the observed $H \to \mu^+\mu^-$ decay [2390] and the observed $t\bar{t}H$ production [2320], lie on the single line that the mechanism predicts when they are plotted against the mass of the partner: coupling proportional to mass for fermions and to mass squared for gauge bosons, over more than three decades in mass [2263, 2394, 2410].

Rests on Equation (104.70) *and Proposition 104.61.*

Derivation. *Derives Phenomenon 113.24.*

In the mechanism a fermion mass comes from a Yukawa term evaluated on the vacuum, $m_f c^2 = \hat{y}_f E_v/\sqrt{2}$, while the coupling of the physical scalar to that same fermion is the same $\hat{y}_f$; hence

$$g_{Hff} = \frac{\hat{y}_f}{\sqrt{2}} = \frac{m_f c^2}{E_v}\,, \tag{113.23}$$

a pure number. A gauge-boson mass comes instead from the square of the covariant derivative, which on the substitution $v \to v + H$ multiplies the mass term by $(1 + H/v)^2$ and therefore gives a linear coupling twice the mass term over v,

$$g_{HVV} = \frac{2\,(M_V c^2)^2}{E_v}\,, \tag{113.24}$$

an energy rather than a pure number. Both are Equation (104.70), derived in Proposition 104.61. Neither contains a free parameter beyond $E_v = 246.220\,\mathrm{GeV}$, which is fixed by the muon lifetime.

The variable that puts the two families on one axis is the *reduced* coupling: g_{Hff} itself for a fermion and $\sqrt{g_{HVV}/2E_v} = M_V c^2/E_v$ for a vector. In that variable every Standard Model particle lies on one straight line through the origin of slope $1/E_v$, and Table 113.5 evaluates it: from 4.29×10^{-4} for the muon to 0.701 for the top, a range of 1630, each value known to the precision of the corresponding mass and carrying no theoretical uncertainty at tree level.

What is measured is not a coupling but a rate — a product of a production cross section and a branching fraction — so a global fit over many final states is required to disentangle them. The experiments report that fit in terms of modifiers κ_i defined so that $\kappa_i = 1$ reproduces Equation (113.23) and Equation (113.24), with rates scaling as κ^2. Every κ_i comes out consistent with unity, to roughly a tenth for W, Z, t, b and τ and roughly a fifth for μ [2394, 2410].

Two features make this a sharper test than a fit to six numbers suggests. First, there is nothing to adjust: a deviation anywhere along the line would falsify the mechanism rather than shift a parameter, because the parameter that would have to shift, E_v, is independently fixed by the muon lifetime to seven significant figures and enters every point of the line at once. Second, the $\gamma\gamma$ and gg vertices exist only at one loop (Remark 113.13), so their rates test the same couplings in a different combination — the diphoton amplitude is the W loop interfering destructively with the top loop — which constrains the *relative sign* of κ_W and κ_t, something no tree-level rate reaches. □

Remark 113.25 (Three decades, and where they stop). The verified range runs from the muon to the top. It does not include the electron, the up, the down or the strange quark, whose reduced couplings are all below 4×10^{-4} and none of which has been observed to couple to the scalar at all. The claim that one mechanism gives mass to all the charged fermions is therefore established for the third generation, established at three standard deviations for one second-generation lepton [2390], and *assumed* for the first generation and for the light quarks. That is a real gap and Section 113.6.2 states what would close it.

113.5.5 The total width

Phenomenon 113.26 (The resonance is narrow)**.** The total decay width of the boson lies far below the mass resolution of either detector, so the width of the observed peak is entirely instrumental and no direct line-shape measurement is possible. Comparing ZZ production far above the resonance with production on it constrains the width to a value consistent with the predicted few megaelectronvolts [2411], and the world average is $\hbar\Gamma_H = 3.7\,\mathrm{MeV}$ with an asymmetric uncertainty of $+1.9$ and -1.4 [2427].

Rests on Proposition 113.14 and Equation (104.67)*.*

Derivation. *Derives Phenomenon 113.26.*

Two statements are needed: that the direct measurement is impossible, and that the indirect one works.

The direct measurement. The Standard Model prediction obtained by summing the partial widths of Table 104.4 is $\hbar\Gamma_H = 4.1\,\mathrm{MeV}$, that is $6.57 \times 10^{-13}\,\mathrm{J}$, a fraction 3.3×10^{-5} of the mass. The best mass resolution achieved is $1.1\,\mathrm{GeV}/c^2$ (Proposition 113.14), larger by a factor of 270. The observed peak is therefore a pure instrumental response function and carries no information about Γ_H beyond a bound of order 1 GeV. Equivalently, the mean lifetime is

$$\tau = \frac{\hbar}{\hbar\Gamma_H} = \frac{1.054\,572 \times 10^{-34}\,\mathrm{J\,s}}{5.93 \times 10^{-13}\,\mathrm{J}} = 1.78 \times 10^{-22}\,\mathrm{s}\,,$$

and the decay length $c\tau = 5.3 \times 10^{-14}\,\mathrm{m}$, a twentieth of a picometre: there is no displaced vertex to find either. Both inherit the width's asymmetric error, $+51\,\%$ and $-38\,\%$, inverted: the width's upper end 5.6 MeV gives $\tau = 1.18 \times 10^{-22}\,\mathrm{s}$ and its lower end 2.3 MeV gives $2.86 \times 10^{-22}\,\mathrm{s}$, so τ is known only to a factor of about two either way. Nothing in the argument depends on which end is taken.

The off-shell method. Let W be the invariant-mass energy of the ZZ system produced through the scalar in gluon fusion. The differential cross section carries a relativistic Breit–Wigner,

$$\frac{\mathrm{d}\sigma}{\mathrm{d}W^2} \propto \frac{\hbar\Gamma_{gg}\,\hbar\Gamma_{ZZ}}{\left[W^2 - (m_H c^2)^2\right]^2 + (m_H c^2)^2\,(\hbar\Gamma_H)^2}\,, \tag{113.25}$$

with $\hbar\Gamma_{gg}$ and $\hbar\Gamma_{ZZ}$ the partial widths into the initial and final states, written as energies like every other width in this chapter, so that numerator and denominator carry J^2 and J^4 respectively. Write each coupling as a modifier times its Standard Model value, so $\hbar\Gamma_{gg}\,\hbar\Gamma_{ZZ} = \kappa_g^2\kappa_Z^2\,\hbar\Gamma_{gg}^{\mathrm{SM}}\,\hbar\Gamma_{ZZ}^{\mathrm{SM}}$, and consider two regions.

On the peak, $\hbar\Gamma_H \ll m_H c^2$ lets the numerator be taken constant across the resonance, and

$$\int \frac{\mathrm{d}W^2}{\left[W^2 - (m_H c^2)^2\right]^2 + (m_H c^2)^2\,(\hbar\Gamma_H)^2} = \frac{\pi}{m_H c^2\,\hbar\Gamma_H}\,,$$

the standard Cauchy integral. Hence

$$\mu_{\mathrm{on}} := \frac{\sigma_{\mathrm{on}}}{\sigma_{\mathrm{on}}^{\mathrm{SM}}} = \frac{\kappa_g^2\kappa_Z^2}{\hbar\Gamma_H/\hbar\Gamma_H^{\mathrm{SM}}}\,. \tag{113.26}$$

Far above the peak, $W \gg 2m_Z c^2 \gg m_H c^2$, the denominator of Equation (113.25) is W^4 up to corrections of order $(m_H c^2/W)^2$ and $(\hbar\Gamma_H/W)^2$, both utterly negligible: the width has dropped out. Hence

$$\mu_{\mathrm{off}} := \frac{\sigma_{\mathrm{off}}}{\sigma_{\mathrm{off}}^{\mathrm{SM}}} = \kappa_g^2\kappa_Z^2\,, \tag{113.27}$$

and dividing Equation (113.27) by Equation (113.26),

$$\frac{\hbar\Gamma_H}{\hbar\Gamma_H^{\mathrm{SM}}} = \frac{\mu_{\mathrm{off}}}{\mu_{\mathrm{on}}}\,. \tag{113.28}$$

The couplings cancel and the width is measured by a ratio of two rates, neither of which resolves anything.

The method is usable only because the off-shell rate is not negligible: about a tenth of all $gg \to H^* \to ZZ$ production occurs above $W = 2m_Z c^2$, because the threshold opening enhances the rate and the top-quark loop amplitude keeps growing until $W = 2m_t c^2$. ATLAS reported evidence for that off-shell component and used Equation (113.28) to constrain the width to a value consistent with the prediction [2411]. □

Remark 113.27 (The assumptions the width bound rests on). Equation (113.28) is exact given Equation (113.25), and three of its inputs are assumptions rather than measurements. They are stated here because the resulting constraint is often quoted as if it were a direct measurement of a lifetime, and it is not.

First, κ_g and κ_Z are taken to be the same at $W = 125\,\mathrm{GeV}$ and at $W \gtrsim 400\,\mathrm{GeV}$. Any new particle in the gluon-fusion loop, or any new contribution to the high-mass ZZ amplitude, breaks that and the extracted width is wrong — and precisely such a particle is what an anomalous width would be evidence for, so the method assumes the absence of the thing it would be used to detect.

Second, the continuum $gg \to ZZ$ background proceeds through a box diagram with the same initial and final state and therefore *interferes* with the signal. The interference is

destructive and large, and it must be modelled; its higher-order strong corrections are not known to the order the signal's are, and are assumed to be the same.

Third, the narrow-width approximation used in Equation (113.26) is safe here only because $\hbar\Gamma_H/m_Hc^2 = 3.3\times10^{-5}$; that part of the derivation is not in doubt.

The honest summary is that the width is constrained, under stated model assumptions, to a value consistent with 4.1 MeV, and that no assumption-free measurement of it exists or is planned at a hadron collider.

113.5.6 What the discovery does and does not establish

It is worth separating the four claims, because they are established to very different degrees and are routinely run together.

A neutral scalar exists at $(125.20\pm0.11)\,\mathrm{GeV}/c^2$. Established, by two independent instruments, in two independent channels each, with the mass measured consistently and the rate as predicted. This is as secure as any result in particle physics.

Its quantum numbers are $J^P = 0^+$. Established: spin 1 by Landau–Yang from the existence of the diphoton decay, spin 0 over spin 2 and even over odd parity by the four-lepton angular analysis [2201, 2217]. A small CP-odd admixture is bounded, not excluded (Remark 113.23).

Its couplings are proportional to mass. Established over the range from the muon to the top — a factor of 1630 — to a precision of roughly ten per cent, with the first generation and the light quarks untested (Remark 113.25).

The potential is the quartic one of Equation (113.1). **Not established.** Everything above probes the theory at the minimum of the potential. Its shape away from the minimum enters only through the self-couplings, and those are bounded to within a factor of several of the predicted value and not measured (Section 113.6.1). The central postulate of the mechanism — that the vacuum sits at the minimum of a quartic — remains an assumption supported by the mass and the couplings, and not a measurement.

113.6 What is not yet measured

113.6.1 The self-coupling and the shape of the potential

Expanding Equation (113.1) in unitary gauge gives, by Proposition 104.63,

$$V(h) = \frac{1}{2}\left(\frac{m_Hc}{\hbar}\right)^2 h^2 + \frac{(m_Hc)^2}{2\hbar^2 v}h^3 + \frac{(m_Hc)^2}{8\hbar^2v^2}h^4\,, \tag{113.29}$$

so the trilinear and quartic self-couplings are fixed by m_H and E_v, both measured. Writing the trilinear coupling as κ_λ times its Standard Model value, the mechanism predicts $\kappa_\lambda = 1$ with nothing to adjust. Nobody has measured it.

The reason is the size of the only channel that carries it. Producing two scalars at once, $gg\to HH$, proceeds through two interfering amplitudes: a *triangle* in which a top loop makes one off-shell scalar that then splits, carrying κ_λ at the splitting vertex, and a *box* in which two scalars are emitted from the same top loop, carrying no self-coupling at all. The two interfere destructively — a low-energy theorem forces them to cancel at threshold in the limit of a heavy top — so the cross section is a quadratic in κ_λ with a minimum inside the range of interest rather than a monotone function of it. Three consequences follow, and all three are unhelpful.

First, the rate is tiny: some three orders of magnitude below single production, tens of femtobarns at LHC energies against the tens of picobarns of Section 113.3.1. Second, because the dependence is quadratic with a minimum inside the interesting range, a measured rate maps onto two values of κ_λ, so the constraint is naturally an interval with a possible second

solution rather than a value. Third, the destructive interference means the observable loses sensitivity in the neighbourhood of the predicted value rather than gaining it there.

The present constraints [2394, 2410] allow κ_λ over a range of order -1 to $+7$: consistent with the predicted 1, and far too loose to be called a test.

Derivation pending #202

The double-production cross section $\sigma(gg \to HH)$ as a function of κ_λ: the interference of the top-loop triangle carrying the trilinear vertex with the top-loop box that does not, giving the quadratic whose minimum locates the least sensitive value of κ_λ, and the normalization that puts the rate three orders of magnitude below single production. The location of the minimum and the absolute rate are stated in this subsection on the authority of the experimental compilations and are not derived here; the two loop amplitudes belong in Appendix A with the single-production ones already recorded as owed. The published constraint on κ_λ quoted below is a measurement and needs no derivation.

The high-luminosity programme is expected to reduce that to a band of order tens of per cent around the prediction, which would be the first real measurement of the shape of the potential; that is a projection, not a result, and it is recorded here as one.

Remark 113.28 (What is actually being assumed). It is worth being precise about what is unmeasured, because the phrase "the Higgs potential" is used for two different things. That the electroweak vacuum breaks the gauge symmetry and sits at some nonzero field value is established — it is what the boson masses of Phenomenon 113.6 measure, and what E_v is. That the function whose minimum it sits at is $-\mu^2 |\Phi|^2 + \lambda |\Phi|^4$, with the single coupling λ generating the mass and both self-interactions, is the minimal hypothesis and is unmeasured. Any potential with a minimum in the right place and the right curvature reproduces every result in Section 113.5; only the self-couplings distinguish them. Loop corrections make the point sharper: the object whose minimum determines the vacuum is not the tree-level potential but the effective potential of Coleman and Weinberg [1495], whose shape differs (Remark 104.64).

113.6.2 First-generation couplings and rare decays

The scalar has been seen to couple to the top, the bottom, the tau and — at three standard deviations — the muon [2390]. It has not been seen to couple to the electron, the up quark, the down quark or the strange quark, which between them account for essentially all the mass of ordinary matter that is not binding energy. The claim that one mechanism gives mass to all the fermions is therefore verified for the third generation, for one second-generation lepton, and nowhere else.

Proposition 113.29 (Why the electron coupling is out of reach)**.** *The Standard Model branching fraction to electrons is*

$$\mathcal{B}\left(H \to e^+e^-\right) = \mathcal{B}\left(H \to \mu^+\mu^-\right)\left(\frac{m_e}{m_\mu}\right)^2 = 2.2 \times 10^{-4} \times 2.34 \times 10^{-5} = 5.1 \times 10^{-9}\,, \tag{113.30}$$

so even the full high-luminosity dataset would contain of order one such decay, against a Drell–Yan background of order 10^9 events in the same mass window. The measurement is not difficult; it is impossible at this machine.

Rests on Equations (104.67) and (113.23).

Derivation. *Derives Proposition 113.29.*

By Equation (104.67) the partial width into a fermion pair is proportional to $\hat{y}_f^2$ times a phase-space factor which is unity to one part in 10^5 for both leptons, and $\hat{y}_f \propto m_f$ by Equation (113.23). Hence the ratio of branching fractions is $(m_e/m_\mu)^2$. With $m_e c^2 = 0.510\,998\,950\,69\,\mathrm{MeV}$ and $m_\mu c^2 = 105.658\,375\,5\,\mathrm{MeV}$ [2427, 2431] the ratio is $(4.8363 \times 10^{-3})^2 = 2.339 \times 10^{-5}$, and $2.2 \times 10^{-4} \times 2.339 \times 10^{-5} = 5.1 \times 10^{-9}$, which is Equation (113.30).

The high-luminosity programme aims at $3 \times 10^{46}/\mathrm{m}^2$, three thousand inverse femtobarns per experiment. At $\sqrt{s}\,c = 13\,\mathrm{TeV}$ the total production cross section is about $5.5 \times 10^{-39}\,\mathrm{m}^2$, so $N_H \approx 3 \times 10^{46}/\mathrm{m}^2 \times 5.5 \times 10^{-39}\,\mathrm{m}^2 = 1.7 \times 10^8$ scalars, of which $1.7 \times 10^8 \times 5.1 \times 10^{-9} = 0.87$ decay to an electron pair. One event, with an acceptance below unity, on a background of Drell–Yan electron pairs whose cross section in the same mass window exceeds the signal by nine orders of magnitude. No selection recovers that. The light quarks are worse still, because there is no way to tag a light-quark jet as up rather than down, so even the final state cannot be identified.

What is achievable is a *bound*. An enhanced electron Yukawa coupling, hundreds of times the Standard Model value, would produce a visible $H \to e^+e^-$ peak, and searches set limits at that level. A bound three orders of magnitude above the prediction is worth recording as a bound and is not a test of the prediction. □

Three other classes of measurement are open and are being pursued.

Loop-level rare decays. $H \to Z\gamma$ has a predicted branching fraction of 1.5×10^{-3} (Table 104.4) and, like $H \to \gamma\gamma$, no tree-level vertex: it proceeds through W and top loops in a combination different from the diphoton one, so measuring both constrains the loop content in two independent ways. Evidence for it has been reported by the two collaborations combined, at a significance near the evidence threshold; the treatise's bibliography does not currently hold that publication, and the claim is therefore stated here without a citation rather than attached to the wrong one.

Lepton-flavour-violating decays. $H \to \mu\tau$ and $H \to e\tau$ are forbidden in the Standard Model, where the Yukawa matrix and the mass matrix are diagonalized by the same transformation (Theorem 104.38) so that no flavour-changing neutral scalar current exists at tree level. Searches have found nothing and bound the branching fractions at the level of a part in a thousand. This is a null result and it is a meaningful one: it says the scalar's couplings are flavour-diagonal to that accuracy, which the mechanism requires and a general two-doublet sector would not.

Invisible decays. The Standard Model predicts $\mathcal{B}(H \to ZZ^* \to 4\nu) \approx 1.1 \times 10^{-3}$, which is one line of Table 104.4: the branching fraction to ZZ^* is 0.026 and each Z goes to neutrinos with probability 0.20, so $0.026 \times 0.20^2 = 1.04 \times 10^{-3}$, the small remainder coming from the off-shell boson's distorted phase space. Any additional invisible width would signal decays to particles the detector cannot see, so the bound on it — currently of order ten per cent of the total width — constrains a light unseen sector coupling to the scalar. Collecting those searches, together with the direct and indirect ones, belongs to Chapter 134, whose collider subsection is reserved for exactly this material and does not yet contain it.

113.6.3 Vacuum stability

The measured mass has one consequence that reaches far beyond the collider, and it is the most-quoted and least-secure statement in this chapter.

λ is a coupling, so it runs (Chapter 108). Its one-loop beta function Equation (104.72) is dominated by the top Yukawa through the term $-6\hat{y}_t^4$, negative because it descends from a closed fermion loop, and numerically fifty times larger than the bosonic terms. Proposition 104.66 evaluates it at a renormalization scale $\mu_R = m_t c^2$ — the scale of Notation 113.1, not the potential's mass parameter and not a signal strength — and finds $\mathrm{d}\hat{\lambda}/\mathrm{d}\ln\mu_R = -0.0206$: the quartic coupling falls with scale. Integrating the coupled system

upward, with the measured $m_H c^2 = (125.20 \pm 0.11)\,\mathrm{GeV}$ and $m_t c^2 = (172.57 \pm 0.29)\,\mathrm{GeV}$ as inputs, $\hat{\lambda}$ crosses zero somewhere near $1 \times 10^{10} - -1 \times 10^{11}\,\mathrm{GeV}$ [2193, 2215]. Above that scale the potential turns over and a deeper minimum exists at large field values: the electroweak vacuum is not the global minimum but a metastable one, with a computed tunnelling lifetime exceeding the age of the universe by hundreds of orders of magnitude.

This chapter records the result and, more importantly, records what it is worth. Remark 104.67 states the three caveats in full and none may be dropped: the conclusion sits on a knife edge in the top mass, where a shift of about 1 GeV moves the answer between absolute stability, metastability and instability, and the relation between what a hadron collider reconstructs and the short-distance mass entering the beta function carries a theoretical uncertainty of that size; the extrapolation assumes that nothing couples to the scalar anywhere between $10^3\,\mathrm{GeV}$ and the $10^{10}\,\mathrm{GeV}$ at which the quartic changes sign — seven decades of energy in which nothing has been looked at, and the extrapolation is only worth as much as that assumption; and the running is a statement about a coupling in a renormalization scheme, not about an observable.

That interval is seven decades wide and not sixteen: $\log_{10}(10^{10}/10^3) = 7$. Remark 104.67 names the same two endpoints and calls them sixteen, which is an arithmetic slip there; the count used here is the one the endpoints give.

The experimental content of this subsection is therefore narrow and should be stated as such: this chapter measured m_H, and Chapter 116 and its successors measured m_t and α_s, and those measurements are the inputs to a calculation whose output is interesting and whose error bar is dominated by an ambiguity in one of its inputs. It is a striking near-coincidence that the Standard Model, extrapolated with no new physics, lands so close to the boundary. It is not evidence for anything.

The related question — why $E_v = 246.220\,\mathrm{GeV}$ at all, when nothing in the theory protects a fundamental scalar's mass from corrections scaling with whatever completes it at high energy — is not a question this chapter can address, and it is recorded among the open problems in Section 133.2. The discovery made it sharper rather than softer: before 2012 one could hope the electroweak sector was not built from an elementary scalar; it is.

113.7 Primary references

The experiment is two papers, submitted on the same day to the same journal and published back to back. ATLAS [2186] and CMS [2192] each report the diphoton and four-lepton analyses in full, with the supporting WW^*, $\tau\tau$ and $b\bar{b}$ channels, the categorization, the background models and the statistical treatment. Both are readable end to end and both should be read in preference to any account of them, including this one; the numbers in Table 113.3 attributed to 2012 come from them and are not recomputed here.

The instruments have their own literature and it is where the apparatus section comes from. The ATLAS [2119] and CMS [2120] technical papers in the 2008 LHC machine special issue describe the detectors as built, subsystem by subsystem, with the design resolutions quoted in Equation (113.13) and Equation (113.14); Evans and Bryant [2123] in the same issue describe the accelerator, and it is the source for every machine parameter in Table 113.2 other than the two derived rows.

The statistical machinery has a small canonical literature and this chapter uses all of it. Cowan, Cranmer, Gross and Vitells [2172] fix the asymptotic formulae, the one-sided discovery statistic and the Asimov construction that the collaborations use to quote expected sensitivities; Read [2036] is the CL_s construction behind every exclusion limit quoted here, including LEP's; Wilks [825] is the theorem that makes $Z = \sqrt{q}$ possible and Gross and Vitells [2158] the trials-factor practice applied in Proposition 113.19, resting in turn on Davies' bound [1594, 1791] and on Rice's level-crossing formula [877]. All of it is derived in

Chapter 11 rather than quoted.

The theory the experiment tested was in print half a century before the data. Englert and Brout [1275], Higgs [1282, 1283, 1332] and Guralnik, Hagen and Kibble [1279] give the mechanism in 1964, with Kibble's non-abelian extension [1352] supplying the case the Standard Model needs and Anderson's plasmon argument [1245] the physical precedent; Goldstone [1207] and Nambu [1187] give the theorem that has to be evaded and the superconducting analogy that shows how. Glashow [1206], Weinberg [1362] and Salam [1381] assemble the electroweak theory around it, and Chapter 104 derives the whole of it; Weinberg's second volume [1933] is the standard reference treatment.

The searches that preceded the discovery are the LEP combination of ALEPH, DELPHI, L3 and OPAL [2045], which set the lower limit and reported the excess at the edge of its sensitivity, and the Tevatron combination of CDF and D0 [2187], which reported the associated-production excess in $b\bar{b}$. The indirect constraints are the LEP and SLD precision electroweak measurements [2104] and the global fits built on them [2228], together with Veltman's screening theorem [1612], which is why the constraint on the scalar mass was logarithmic and therefore weak.

The measurements made since 2012 divide by what they test. The spin-parity analyses are ATLAS [2201] and CMS [2217]; the first joint mass combination is [2242] and the first joint coupling fit [2263]; the ten-year coupling maps of the full Run-2 dataset are ATLAS [2394] and CMS [2410], which are the source for every κ statement in Tables 113.5 and 113.6 and for the self-coupling range in Section 113.6.1. Note that [2410] is the CMS ten-year portrait and not a discovery paper, and it is cited here only for couplings. Individual channel observations are $t\bar{t}H$ production [2320] and the $H \to \mu^+\mu^-$ evidence [2390]; the off-shell width constraint of Section 113.5.5 is [2411]. Vacuum stability is Degrassi *et al.* [2193] and its sequel [2215].

Two sources are used for numbers rather than for claims. The Particle Data Group's 2024 review [2427] supplies every world-average mass and width in this chapter, taken from its own machine-readable release and stored with the treatise; CODATA 2022 [2431] supplies the exact defining constants used in every conversion between electronvolts, joules and kilograms.

Finally, one gap should be named rather than papered over. The Standard Model production cross sections quoted in Section 113.3.1 are the compilations of the LHC Higgs Cross Section Working Group, which the collaborations cite and which this treatise's bibliography does not hold; they are attributed above to the two discovery papers, which quote and use them. The same applies to the ATLAS and CMS combination reporting evidence for $H \to Z\gamma$, referred to in Section 113.6.2 without a citation for the same reason.

Chapter 114

Experiment: Neutrino Oscillations

Tests Phenomenon 103.50.

Assuming Definition 103.46 and Theorems 103.54 and 103.57.

Neutrino oscillation is the only laboratory phenomenon that unambiguously falsifies the Standard Model as originally written: it requires neutrino masses, which the massless left-handed neutrino of Chapter 96 does not have. The evidence accumulated over thirty years, beginning as an anomaly rather than a discovery — the solar electron-neutrino deficit found at Homestake [1368, 1950] and confirmed by Kamiokande [1822] and the gallium detectors [1974, 2138] — and closing with two measurements that admit no astrophysical escape: the zenith-angle dependence of atmospheric muon neutrinos in Super-Kamiokande [1954] and the flavour-blind neutral-current flux measured at Sudbury [2021], which showed the missing solar neutrinos had changed flavour rather than gone away. The Nobel Prize of 2015 recognized those two. This chapter sits after the flavour and neutrino theory of Chapters 96 and 103 and supplies the parameters they use.

The prediction under test is that a neutrino changes flavour in flight. That is the single statement every instrument described below was built to confirm or to kill, and it is worth being exact about what a confirmation requires. It is not enough to find fewer neutrinos of a given flavour than expected: a deficit is what a wrong source model, a wrong cross-section or a wrong efficiency also produces. Three things, and only these three, distinguish flavour change from every mundane alternative, and the chapter is organized around them.

1. *The missing flux reappears in another flavour.* A detector that measures the electron-flavour rate and the flavour-summed rate in the same target, with the same exposure, can subtract one from the other; the difference is model-independent. This is the Sudbury measurement.

2. *The deficit depends on L/E and on nothing else.* Path length and energy enter the interference phase only in that combination. No absorption law, no decay law and no production spectrum reproduces a dependence on the ratio alone. This is the Super-Kamiokande zenith-angle measurement, and, in its sharpest form, the KamLAND spectrum.

3. *The deficit is not monotone in L/E but oscillates.* A normalization error can only scale a rate; only interference makes it come back. This is what a controlled terrestrial source, at a baseline chosen so that the phase is of order a radian across the detected spectrum, can show and an astrophysical source cannot.

The chapter is unusual among the experiment chapters in covering a family of instruments rather than one — radiochemical tanks, water Cherenkov detectors, heavy water, liquid scintillator, and long-baseline accelerator beams — because no single detector establishes the effect, and the argument is carried by the pattern across baselines and energies, sharpened

by the matter effect of Wolfenstein, Mikheyev and Smirnov [1631, 1779]. It closes on what is still unknown: the mass ordering, the CP phase, the absolute mass scale, and whether the neutrino is its own antiparticle.

Remark 114.1 (Units and the conventions of the experimental literature). Every experimental paper cited below writes its formulae with $\hbar = c = 1$, and this book does not: editorial rule 3 requires SI throughout, so $\hbar$ and c are carried explicitly in every equation and every intermediate step here. The bridge is Equation (103.1), $\hbar c = 1.973\,269\,8 \times 10^{-7}$ eV m, and the general dictionary is stated once in Remark 103.1. No derivation in this chapter is carried out inside this remark.

Four conventions of the experimental literature recur below and are translated here rather than in the running text.

Squared-mass splittings. A splitting is universally quoted as "Δm^2 in eV^2" when the quantity meant is $\Delta m^2 c^4$, an energy squared. This chapter writes the c^4.

The solar neutrino unit. A radiochemical capture rate is quoted in solar neutrino units, and one SNU is defined as 10^{-36} captures per target atom per second — in SI, 1×10^{-36}/s per target atom. It is a rate per nucleus, not a flux, and converting it to atoms per day requires the number of target atoms, which is why that number is given for every radiochemical detector below.

Fluxes. Solar neutrino fluxes are quoted in the literature per square centimetre per second. The SI value is larger by 10^4: a flux of 10^6 1/(cm^2 s) is 1×10^{10}/(m^2 s), and only the SI form appears below.

Source strengths. Calibration sources are quoted in curies. One curie is exactly 3.7×10^{10} Bq, so a 1.7×10^6 Bq source — the units in which this chapter writes it — is not what the papers call a megacurie; a 1.7-megacurie source is 6.3×10^{16} Bq, and that is the number written here.

114.1 Historical context and the prediction under test

114.1.1 Neutrinos: detection and flavour

The neutrino entered physics in 1930 as a bookkeeping device and became an object of experiment twenty-six years later. Cowan, Reines and their collaborators put a detector beside a fission reactor at Savannah River and looked for the inverse beta reaction Equation (101.3), $\bar{\nu}_e + p \to n + e^+$, whose threshold Equation (101.4) is 1.806 MeV, that is 2.894×10^{-13} J. Their target was two tanks each holding about 200 L of water loaded with cadmium chloride, sandwiched between three tanks of about 1400 L of liquid scintillator; the detector sat some 12 m below ground level and about 11 m from the reactor core. The signature was a delayed coincidence: two 511 keV annihilation photons from the positron, followed some microseconds later by the several megaelectronvolts of gamma rays released when the neutron, having thermalized, is captured on cadmium. The rate with the reactor on exceeded the rate with it off by about three counts an hour [1085, 1097], at a cross-section of order 10^{-47} m^2 — the number derived in Equation (101.5) and recorded as Phenomenon 101.5.

Everything in this chapter descends from that apparatus. The delayed-coincidence tag on a free-proton target is still the technique of KamLAND and Daya Bay sixty years later, scaled from a few hundred litres to a kilotonne; what changed is the radiopurity of the scintillator and the photocathode coverage, not the idea.

The second ingredient is flavour. Firing a beam of neutrinos from pion decay at a ten-tonne spark chamber behind 13.5 m of iron, Danby and collaborators at the Brookhaven alternating-gradient synchrotron found the charged particles produced to be muons and never electrons [1223]: thirty-four single-muon events and no electron showers, where the electron channel is kinematically the more open of the two. The null result establishes a

quantum number distinguishing the two families. The third family's neutrino was observed directly at Fermilab in 2000 by the DONUT collaboration, which found the short kinked track of a tau in nuclear emulsion exposed to a beam from 800 GeV protons on a tungsten target: four candidates on an expected background of 0.34 [2016].

The counting is closed from the other end by the invisible width of the Z boson, which fixes the number of light neutrino species that couple to it at

$$N_\nu = 2.9840 \pm 0.0082 \tag{114.1}$$

[2104]; this is Phenomenon 103.22, and its derivation from the measured widths is Equation (103.15). Three flavours, conserved separately in the Standard Model as written, and therefore three quantum numbers for an oscillation experiment to violate: that is the state of knowledge against which every measurement below is read.

114.1.2 Mixing and oscillation

Pontecorvo proposed in 1957 that a neutrino might transform into its antiparticle by analogy with the neutral kaon, whose strangeness oscillates because the states of definite strangeness are not the states of definite mass [1127], in the paper whose English translation is [1159]. Maki, Nakagawa and Sakata gave the idea its modern form in 1962: the states produced with definite flavour are superpositions of states of definite mass, and the mismatch is a unitary matrix [1237]. That matrix is defined in Definition 103.46 and its measured magnitudes are Equation (103.49).

The formalism is derived in Chapter 103 and is not rebuilt here. Three results from there are used throughout this chapter and are quoted now so that the experimental sections can refer to them without interruption. For two flavours with mixing angle θ and squared-mass splitting $\Delta m^2 := m_2^2 - m_1^2$, the probability that a neutrino born as ν_α is found as ν_β after a flight of length L at energy E is Equation (103.52),

$$P_{\alpha\to\beta} = \sin^2 2\theta \, \sin^2 \varphi \,, \qquad P_{\alpha\to\alpha} = 1 - \sin^2 2\theta \, \sin^2 \varphi \,, \qquad \varphi = \frac{\Delta m^2 c^3 L}{4\hbar E} \,. \tag{114.2}$$

Every symbol is SI: $\Delta m^2 c^4$ is an energy squared, $\hbar c E$ is an energy squared times a length, and φ is a pure number. The practical form of the phase, derived as Proposition 103.52, is

$$\varphi = \frac{(\Delta m^2 c^4)\, L}{4\hbar c\, E} = 1.26693 \, \frac{\left(\Delta m^2 c^4 / \mathrm{eV}^2\right) (L/\mathrm{km})}{(E/\mathrm{GeV})} \,, \tag{114.3}$$

and the oscillation length — the distance over which φ advances by π — is Equation (103.55),

$$L_{\mathrm{osc}} = \frac{4\pi\hbar c E}{\Delta m^2 c^4} = 2.4797\,1 \,\mathrm{km} \, \frac{(E/\mathrm{GeV})}{\left(\Delta m^2 c^4 / \mathrm{eV}^2\right)} \,. \tag{114.4}$$

The number 1.266 93 is the one every experimental paper writes as "1.27" and quotes rather than derives; it is $(4\hbar c)^{-1}$ expressed in the units the field happens to use, and nothing else.

Two consequences organize everything below. Oscillation is an interference between two mass components, so it measures a *difference* of squared masses and can never measure a mass; and it depends on L and E only through their ratio, so an experiment is characterized by L/E and not by either separately. A detector placed where $\varphi \ll 1$ sees nothing; one placed where $\varphi \gg 1$ sees a constant suppression by the averaged factor $1 - \frac{1}{2}\sin^2 2\theta$; only a detector placed where φ is of order one across the detected spectrum sees the oscillation itself. Equation (114.4) is therefore the design equation of every experiment in this chapter, and Table 114.1 is the resulting map.

Table 114.1: Why each experiment has the baseline it has. The phase φ is Equation (114.3) evaluated at the stated L and E with the splitting the experiment is sensitive to, $\Delta m^2_{21}c^4 = 7.53 \times 10^{-5}\,\mathrm{eV}^2$ or $\left|\Delta m^2_{32}\right| c^4 = 2.455 \times 10^{-3}\,\mathrm{eV}^2$ [2427]. An experiment is informative only where φ is of order unity; $\varphi \ll 1$ gives no effect and $\varphi \gg 1$ gives a constant averaged suppression that carries no information about the splitting.

Experiment	L	E	Splitting	φ
Daya Bay near	0.47 km	3.5 MeV	Δm^2_{32}	0.42
Daya Bay far	1.65 km	3.5 MeV	Δm^2_{32}	1.47
KamLAND	180 km	4 MeV	Δm^2_{21}	4.3
K2K	250 km	1.3 GeV	Δm^2_{32}	0.60
T2K	295 km	0.6 GeV	Δm^2_{32}	1.5
MINOS	735 km	3 GeV	Δm^2_{32}	0.76
NOvA	810 km	2 GeV	Δm^2_{32}	1.26
Atmospheric, downward	15 km	1 GeV	Δm^2_{32}	0.047
Atmospheric, upward	12 800 km	1 GeV	Δm^2_{32}	40
Sun to Earth	1.5×10^8 km	1 MeV	Δm^2_{21}	1.4×10^7

The last row is the reason the solar measurements are treated separately from all the others. At a phase of 10^7 radians the oscillating factor has averaged out completely long before the neutrino reaches the Earth, and a vacuum calculation would give a featureless suppression carrying no information; what the solar experiments actually measure is not vacuum oscillation at all but the matter-modified propagation of Theorem 103.57, whose signature is a dependence on *energy* rather than on baseline. That is taken up in Section 114.5.2.

114.1.3 The standard solar model prediction

The Sun burns hydrogen to helium by the pp chains, with a small contribution from the CNO cycle, and the net reaction of either is

$$4\,{}^1\mathrm{H} \longrightarrow {}^4\mathrm{He} + 2e^+ + 2\nu_e\,, \tag{114.5}$$

releasing $Q = 26.73\,\mathrm{MeV} = 4.283 \times 10^{-12}\,\mathrm{J}$ per helium nucleus and producing exactly two electron neutrinos. The chains themselves belong to Sections 52.2.2 and 52.2.3, and the detailed model that turns them into a predicted spectrum — opacities, equation of state, composition, calibrated against the observed luminosity and radius — is that of Bahcall, Serenelli and Basu [2072]. Its predictions are Table 114.2.

Two features of that table decided thirty years of argument. The first is that the total is not really a prediction at all. The luminosity constraint — derived in Equation (52.36) from nothing but the observed solar constant, the energy release of Equation (114.5) and the Earth–Sun distance — fixes

$$\Phi_\nu = \frac{2L_\odot}{Q\,4\pi d^2} \approx 6.4 \times 10^{14}/(\mathrm{m}^2\,\mathrm{s}) \tag{114.6}$$

independently of every uncertain input of the model, and the *pp* branch carries about 94 % of it. A deficit measured in the *pp* neutrinos therefore cannot be blamed on the model.

The second is that the ^{8}B flux is the least certain quantity in the table by a wide margin, because the reaction ${}^7\mathrm{Be} + p \to {}^8\mathrm{B} + \gamma$ sits far out on the Gamow peak and its rate varies as a very high power of the central temperature. For three decades that was the escape route: Homestake and Kamiokande measured only the ^{8}B branch, and a central temperature lower by a few per cent than the model's would have removed the discrepancy without any new

Table 114.2: Standard solar model neutrino fluxes at the Earth, from the BS05(OP) model of [2072], converted to SI: the source quotes them per square centimetre per second and the SI value is larger by 10^4. Endpoint energies are given in electronvolts and in joules. The ^{7}Be and *pep* neutrinos are lines rather than continua because the reaction is electron capture and the final state is two-body. The quoted uncertainty is the model's own, and its size is the point: the *pp* flux is fixed to one per cent by the luminosity constraint, while the ^{8}B flux — the one Homestake, Kamiokande and SNO are sensitive to — is uncertain by 16 %.

Branch	Flux $(1/(\mathrm{m}^2\,\mathrm{s}))$	Endpoint or line	Uncertainty
pp	5.99×10^{14}	420 keV, 6.73×10^{-14} J	1 %
^{7}Be	4.84×10^{13}	862 keV, 1.38×10^{-13} J	11.5 %
pep	1.42×10^{12}	1.44 MeV, 2.31×10^{-13} J	2 %
^{8}B	5.69×10^{10}	15 MeV, 2.40×10^{-12} J	16 %
^{13}N	3.07×10^{12}	1.20 MeV, 1.92×10^{-13} J	31 %
^{15}O	2.33×10^{12}	1.73 MeV, 2.77×10^{-13} J	34 %

physics. Closing that route — by measuring a branch the model cannot get wrong, and then by measuring the flavour content directly — is what the rest of this chapter is about.

That sensitivity is now derived where it belongs, in the stellar-structure chapter: the ^{8}B flux responds with the *sum* of the Gamow exponents of the reactions feeding and draining ^{7}Be, giving $\Phi(^8\mathrm{B}) \propto T_c^{\sim 29}$ at fixed density and composition — self-consistent solar-model perturbation studies soften the exponent toward the customary T^{24} without changing the argument — so a central temperature a few per cent low reproduces the whole measured deficit. The derivation, and the closing of that escape route by the luminosity-pinned *pp* flux and by helioseismology, are carried out under Phenomenon 52.22, with the exponents tabulated in Table 52.1.

114.1.4 The observable and what a measurement must deliver

An oscillation experiment reports one of four observables, and it is worth naming them before any apparatus is described, because the sources of error are different in each and the tables of Section 114.4 are organized by them.

A rate. The number of interactions of a given flavour per unit time in a known target, compared with a prediction. This is the weakest observable, because it inherits the full uncertainty of the source model, the cross-section and the detection efficiency. Homestake delivered a rate, and that is why its result was arguable for thirty years.

A ratio of rates in one detector. Two channels with different flavour sensitivity, measured in the same target during the same exposure. The source normalization cancels exactly. This is Sudbury's charged-current to neutral-current ratio, and it is the strongest observable in the chapter.

A ratio of rates in two detectors. Identical detectors at two baselines from one source. Source normalization, cross-section and efficiency all cancel to the extent that the detectors are identical. This is the Daya Bay and RENO design.

A shape. The dependence of the survival probability on L/E within one dataset. Nothing that scales a rate can produce a shape, so this is the observable that distinguishes oscillation from every rival hypothesis without appeal to any external prediction. Super-Kamiokande delivered it in the zenith angle, KamLAND in the reactor energy spectrum.

The chapter's argument is the accumulation of these four, in that order of strength, over thirty-five years.

114.2 Apparatus

Every instrument in this chapter solves the same two problems. A neutrino cross-section of order $10^{-47}\,\mathrm{m}^2$ (Phenomenon 101.5) means that a useful counting rate demands a target of 10^{30} nuclei or more, that is hundreds of tonnes; and a rate of a few events per day in such a target means that any background above a few events per day destroys the measurement. Mass and quiet are therefore the two axes along which these detectors are designed, and everything else — the choice of target nucleus, the depth, the purification chemistry, the photomultiplier coverage — follows from one or the other.

Table 114.3 collects the physical parameters. The subsections that follow give, for each class, the target, the reaction it exploits, the signal it produces and the background it must beat.

Table 114.3: The instruments of this chapter. Depth is given both as overburden of rock and, where the collaborations quote it, as the equivalent depth of water that would give the same muon attenuation, which is the comparable number because rock densities differ. Target mass is the sensitive mass, not the total mass of the installation: Super-Kamiokande holds 50 kt of water of which 22.5 kt is fiducial for atmospheric neutrinos, and Sudbury's 1 kt of heavy water sits inside 7 kt of ordinary water that is shielding and not target.

Detector	Target	Sensitive mass	Depth
Homestake	C_2Cl_4	615 t	1478 m, 4200 m w.e.
GALLEX/GNO	$GaCl_3$ solution	30.3 t Ga	3300 m w.e.
SAGE	metallic Ga	50 t	4700 m w.e.
Kamiokande-II	H_2O	680 t fid.	1000 m, 2700 m w.e.
Super-Kamiokande	H_2O	22.5 kt fid.	1000 m, 2700 m w.e.
SNO	D_2O	1 kt	2092 m, 6010 m w.e.
KamLAND	liquid scintillator	1 kt	1000 m, 2700 m w.e.
Borexino	liquid scintillator	278 t	3800 m w.e.
Daya Bay	Gd-loaded scintillator	20 t per module	250 m to 860 m w.e.

114.2.1 Radiochemical detectors

A radiochemical detector counts atoms, not events. The neutrino converts a stable nucleus into a radioactive one by inverse beta decay; the product is a different chemical element, so it can be separated from a target of hundreds of tonnes by ordinary chemistry and counted afterwards by its own decay. The method is uniquely insensitive to background *during* exposure — nothing that is not a neutrino makes the target element into the product element — and it pays for that with the total loss of every other piece of information. There is no energy, no direction, no time of arrival: only a number of atoms integrated over a run of weeks.

The two properties that matter are the threshold, fixed by the mass difference of the two nuclei, and the half-life of the product, which must be long enough to accumulate a countable population and short enough that counting does not take forever. Both detectors described here sit near the optimum: a product half-life of a few weeks against an exposure of a few weeks.

114.2.2 The Homestake chlorine tank

Davis's detector was a horizontal steel tank 6.1 m in diameter and 14.6 m long, holding 615 t — about 3.8×10^5 L — of tetrachloroethylene, C_2Cl_4, an ordinary dry-cleaning fluid chosen because it is cheap, dense in chlorine and chemically inert. It was installed 1478 m below the surface in the Homestake gold mine at Lead, South Dakota, an overburden of about 4200 m of water equivalent [1368, 1950].

The reaction is

$$\nu_e + {}^{37}\mathrm{Cl} \longrightarrow {}^{37}\mathrm{Ar} + e^- \,, \tag{114.7}$$

with a threshold of 814 keV, that is 1.304×10^{-13} J. The threshold is what makes the detector a ^{8}B instrument: by Table 114.2 it is above the *pp* endpoint of 420 keV altogether, so the branch carrying 94 % of the flux is invisible to it, and about three quarters of the predicted rate comes from ^{8}B with most of the remainder from ^{7}Be.

The number of target atoms is worth computing, because every rate in Section 114.4 is quoted per target atom and the conversion to atoms per day runs through it. Tetrachloroethylene has molar mass 165.83 g/mol, so 615 t is 3.71×10^6 moles, containing $4 \times 3.71 \times 10^6$ moles of chlorine and hence 8.93×10^{30} chlorine atoms; the isotopic abundance of ^{37}Cl is 24.23 %, giving

$$N\left({}^{37}\mathrm{Cl}\right) = 2.16 \times 10^{30} \,. \tag{114.8}$$

At the measured capture rate of 2.56×10^{-36}/s per atom (Section 114.4.1) this is a production rate of 5.54×10^{-6}/s, that is 0.48 atoms of argon per day in 615 t of liquid. The saturated population, reached when production balances decay of the 3.024×10^6 s half-life, is $5.54 \times 10^{-6}\ 1/\mathrm{s} / (\ln 2 / 3.024 \times 10^6\ \mathrm{s}) \approx 24$ atoms. Those two dozen atoms, in 2.2×10^{30} target nuclei, are the entire signal.

The product is counted by its own decay. ^{37}Ar decays by electron capture back to ^{37}Cl with a half-life of 3.024×10^6 s, that is 35.0 days, and the K-shell capture leaves an Auger cascade of 2.82 keV, that is 4.52×10^{-16} J, deposited in a few nanoseconds within a fraction of a millimetre. A gas proportional counter of about 0.5 cm^3 internal volume — deliberately tiny, because the background scales with the volume and the signal does not — registers this as a fast-rising pulse of characteristic energy. The counters were made of low-background materials, filled with the extracted argon mixed with tritium-free methane, and operated inside a well of sodium iodide within a lead castle.

The backgrounds that had to be excluded before the deficit could be believed were of two kinds. Cosmic-ray muons surviving to 4200 m of water equivalent produce ^{37}Ar by spallation on chlorine at an estimated 0.047 ± 0.013 atoms per day, some 10 % of the signal; this was measured by exposing the same target at shallower depths and extrapolating. Fast neutrons from the rock, and alpha emitters in the tank walls acting through $^{37}\mathrm{Cl}(\alpha, n)$-type reactions, contribute less. The counting background — a proportional counter firing when no argon atom decayed — was suppressed by rise-time discrimination, since a genuine 2.82 keV Auger cascade is far more localized than a Compton electron traversing the same gas, and by requiring the decay times of a run's candidate events to be consistent with a 3.024×10^6 s half-life.

114.2.3 The gallium detectors

The gallium experiments exist because of one number: the threshold of

$$\nu_e + {}^{71}\mathrm{Ga} \longrightarrow {}^{71}\mathrm{Ge} + e^- \tag{114.9}$$

is 233 keV, that is 3.74×10^{-14} J, comfortably below the *pp* endpoint of 420 keV. A gallium detector therefore sees the branch whose flux is fixed by Equation (114.6) to one per cent, and a deficit there cannot be attributed to the solar model. That is the whole argument

for spending a substantial fraction of the world's annual gallium production on a physics experiment.

GALLEX, and its successor GNO, operated at the Gran Sasso laboratory under about 3300 m of water equivalent, with 30.3 t of gallium held as about 101 t of gallium trichloride in concentrated hydrochloric acid, in a tank of some 54 m^3 [1974]. SAGE, at the Baksan Neutrino Observatory under about 4700 m of water equivalent, used 50 t of gallium in the metallic form — gallium melts at 29.8 °C, so the target is a liquid metal held in seven reactors of about 7 t each [2138].

The target counts are, by the same arithmetic as Equation (114.8) with a molar mass of 69.72 g/mol and an isotopic abundance of 39.89 % for ^{71}Ga,

$$N_{\mathrm{GALLEX}} = 1.04 \times 10^{29}\,, \qquad N_{\mathrm{SAGE}} = 1.72 \times 10^{29}\,. \tag{114.10}$$

At the measured rates of Table 114.4 these correspond to 0.63 and 0.97 atoms of ^{71}Ge per day, and to saturated populations of about ten and sixteen atoms respectively, the ^{71}Ge half-life being 9.875×10^5 s, that is 11.43 days.

The chemistry differs from Homestake's in one important respect: germanium tetrachloride is volatile from the acid solution, so extraction is again a gas purge, but the counting gas is germane, $\mathrm{GeH_4}$, synthesized from the extracted germanium. The decay is again electron capture, with K-shell and L-shell peaks at 10.37 keV and 1.17 keV, that is 1.66×10^{-15} J and 1.88×10^{-16} J. Having two peaks is a real advantage over the single Auger line of ^{37}Ar: a background source that mimics one rarely mimics both in the right ratio.

The gallium detectors are also the only ones in this chapter that were calibrated with an artificial neutrino source of known strength. A ^{51}Cr source of about 6.3×10^{16} Bq — 1.7 megacuries, produced by neutron irradiation of enriched ^{50}Cr in a reactor — emits electron neutrinos by electron capture at 750 keV, above the gallium threshold and below the chlorine one. Lowered into the tank, it turns the detector into an experiment with a known input. SAGE additionally used an ^{37}Ar source. What those calibrations returned is discussed in Section 114.6.4, because the answer was not quite one and the discrepancy has not gone away.

114.2.4 Water Cherenkov detectors

A water Cherenkov detector gives up the radiochemical method's freedom from background and buys, in exchange, everything the radiochemical method throws away: the time of each event to a few nanoseconds, its position to tens of centimetres, its energy, its direction, and — decisively for this chapter — its flavour.

The mechanism is the Cherenkov cone of Section 115.4.4. A charged particle faster than the phase velocity of light in the medium radiates on a cone of half-angle given by Equation (115.36),

$$\cos\theta_c = \frac{1}{n\beta}\,, \tag{114.11}$$

which for water, $n = 1.34$, and an ultrarelativistic particle gives $\theta_c = 41.7°$. The threshold follows from requiring $\beta > 1/n$: with $\gamma_{\mathrm{th}} = \left(1 - n^{-2}\right)^{-1/2} = 1.502$, a particle of rest mass m must have total energy above $1.502\,mc^2$, that is 768 keV for an electron and 159 MeV for a muon. The projection of the cone on the detector wall is a ring, and the number of photons per unit path length is given by Equation (115.38) — some 340 photons per centimetre in the visible band, of which a detector with 40 % photocathode coverage records a few per cent.

Kamiokande-II was the first instrument to detect solar neutrinos in real time. It was a cylindrical tank holding 3 kt of purified water 1000 m underground in the Kamioka mine, viewed by 948 photomultipliers of 50 cm diameter giving about 20 % photocathode coverage,

with a fiducial volume of 680 t for the solar analysis [1822]. It detected neutrino–electron elastic scattering,

$$\nu_x + e^- \longrightarrow \nu_x + e^- \,, \tag{114.12}$$

which has no threshold of its own but is limited from below by radioactive background: the analysis threshold was 9.3 MeV initially and 7.5 MeV later, so that Kamiokande, like Homestake, saw only the ^{8}B branch. Its decisive advantage was that the recoil electron remembers the neutrino direction to within about 25° at these energies, so the signal appears as a peak in the distribution of event directions relative to the Sun sitting on a flat background. The same detector recorded the neutrino burst from SN 1987A [1793].

Super-Kamiokande is the same idea built seventeen times larger. The tank is 39 m in diameter and 42 m high and holds 50 kt of water; the inner detector is viewed by 11 146 photomultipliers of 50 cm diameter for a photocathode coverage of 40 %, and is optically separated from an outer shell instrumented with 1885 smaller tubes that serves as an anticoincidence veto against entering cosmic-ray muons. The fiducial volume for atmospheric neutrinos is 22.5 kt. It sits under 1000 m of rock, about 2700 m of water equivalent, in the Mozumi mine at Kamioka, where the residual cosmic-ray muon rate through the detector is a few per second — a rate that would swamp everything were the outer veto not there [1954]. The water is continuously recirculated through filtration, reverse osmosis, ion exchange and ultraviolet sterilization, reaching a resistivity near $1.8 \times 10^{5}\,\Omega\,\mathrm{m}$ and a light attenuation length beyond 70 m; radon-reduced air is kept above the water surface, because ^{222}Rn dissolved in the water is the limiting background at low energy.

The property that makes Super-Kamiokande an oscillation instrument rather than merely a large one is flavour tagging by ring shape. A charged-current interaction of a ν_μ produces a muon, which is heavy, loses energy slowly and travels in a straight line: its Cherenkov ring has a sharp outer edge. A charged-current interaction of a ν_e produces an electron, which scatters and radiates, initiating an electromagnetic shower whose many low-energy tracks smear the ring into a fuzzy disc. The two are separated event-by-event with a misidentification probability of a few per cent. Nothing else in this chapter measures the two flavours of a single natural source side by side in one detector, and that is precisely why the atmospheric measurement is decisive.

114.2.5 Heavy water and the neutral-current channel

The Sudbury Neutrino Observatory was designed around one property of the deuteron: it is the lightest nucleus that a neutrino can break up by the neutral current, so a heavy-water target gives access, in one detector, to a reaction that counts electron neutrinos and a reaction that counts all three flavours equally.

The apparatus was 1 kt of heavy water of 99.92 % isotopic purity — on loan from the Canadian strategic reserve, the loan being the reason the experiment was possible at all — held in a spherical acrylic vessel 12 m in diameter with a wall about 5 cm thick. The vessel was surrounded by 7 kt of ultrapure ordinary water acting as shielding against radioactivity from the rock and the photomultipliers, and viewed by 9456 photomultipliers of 20 cm diameter mounted on a geodesic sphere of 17.8 m diameter. The whole assembly sat in a cavity 2092 m below the surface in the Creighton nickel mine near Sudbury, Ontario, an overburden of about 6010 m of water equivalent — the greatest of any detector in this chapter, and necessary because a cosmic-ray muon passing through the heavy water would produce spallation neutrons indistinguishable from the neutral-current signal. At that depth the muon rate through the detector is about three per hour [2008, 2021].

Three reactions were used.

$$\nu_e + d \longrightarrow p + p + e^- \qquad \text{(charged current, CC)}\,, \tag{114.13}$$

$$\nu_x + d \longrightarrow p + n + \nu_x \qquad \text{(neutral current, NC)}\,, \tag{114.14}$$

$$\nu_x + e^- \longrightarrow \nu_x + e^- \qquad \text{(elastic scattering, ES)}\,. \tag{114.15}$$

The charged-current reaction has a threshold of $1.442\,\mathrm{MeV}$ ($2.31 \times 10^{-13}\,\mathrm{J}$) and, because only the electron flavour has a charged-current partner light enough to be produced at solar energies, counts ν_e alone. The neutral-current reaction is the breakup of the deuteron, so its threshold is the deuteron binding energy, $2.224\,\mathrm{MeV}$ or $3.56 \times 10^{-13}\,\mathrm{J}$; the Z couples identically to all three flavours, so its rate is proportional to the total active flux with no flavour weighting whatever. Elastic scattering is sensitive to all three but not equally: ν_e can scatter through both the charged and the neutral current while ν_μ and ν_τ have only the neutral one, so the effective flux it measures is

$$\Phi_{\mathrm{ES}} = \Phi_e + 0.154\,\Phi_{\mu\tau}\,. \tag{114.16}$$

The weight is a ratio of cross-sections and is worth deriving, since it is used quantitatively twice below.

Derivation of the elastic-scattering weight. For $E_\nu \gg m_e c^2$ the differential cross-section for $\nu_x + e^- \to \nu_x + e^-$ in the recoil kinetic energy T is

$$\frac{\mathrm{d}\sigma}{\mathrm{d}T} = \frac{2 G_{\mathrm{F}}^2 m_e c^2}{\pi \hbar^4 c^4}\left[g_L^2 + g_R^2\left(1 - \frac{T}{E_\nu}\right)^2 - g_L g_R \frac{m_e c^2 T}{E_\nu^2}\right], \tag{114.17}$$

with $G_{\mathrm{F}}/(\hbar c)^3 = 1.1664 \times 10^{-5}/\mathrm{GeV}^2$, that is $G_{\mathrm{F}} = 1.4359 \times 10^{-62}\,\mathrm{J\,m^3}$ in SI. The prefactor is then an area per energy: $\mathrm{J^2\,m^6} \times \mathrm{J}$ divided by $\mathrm{J^4\,m^4}$ is $\mathrm{m^2/J}$. Evaluated for ν_e at $E_\nu = 10\,\mathrm{MeV}$ it gives a total cross-section of $9.5 \times 10^{-48}\,\mathrm{m^2}$, about a hundredth of the inverse beta decay cross-section at the same energy. The chiral couplings differ between the flavours only through the charged-current contribution, which the electron flavour alone has:

$$g_L^{(e)} = \tfrac{1}{2} + \sin^2\theta_{\mathrm{W}}\,, \qquad g_L^{(\mu,\tau)} = -\tfrac{1}{2} + \sin^2\theta_{\mathrm{W}}\,, \qquad g_R^{(e)} = g_R^{(\mu,\tau)} = \sin^2\theta_{\mathrm{W}}\,. \tag{114.18}$$

Integrating Equation (114.17) over all T and dropping the interference term, which is smaller by $m_e c^2/E_\nu$, gives $\sigma \propto g_L^2 + \tfrac{1}{3} g_R^2$. With $\sin^2\theta_{\mathrm{W}} = 0.2312$ [2427] this is 0.5525 for ν_e and 0.0901 for ν_μ and ν_τ, a ratio of 0.163. Imposing the analysis threshold lowers it: for $E_\nu = 10\,\mathrm{MeV}$ and $T > 5\,\mathrm{MeV}$ the same integral run from the threshold gives 0.142, because the flat g_L^2 term survives the cut in the same proportion for both flavours while the g_R^2 term, which is relatively far more important for ν_μ and ν_τ, is concentrated at low recoil. The coefficient in Equation (114.16) is this ratio averaged over the $^8\mathrm{B}$ spectrum and the detector acceptance and corrected radiatively, which the Sudbury analysis evaluates as 0.154 [2021]; the two bracketing estimates above show where that number comes from and how much of it is threshold-dependent. □

The experimental difficulty is entirely in the neutral-current channel, whose only observable is a free neutron. Three successive configurations were run, deliberately chosen so that the systematic errors of the neutron measurement would be unrelated.

Phase I, pure heavy water. The neutron captures on a deuteron, $n + d \to {}^3\mathrm{H} + \gamma$, emitting a single photon of $6.25\,\mathrm{MeV}$ ($1.00 \times 10^{-12}\,\mathrm{J}$) which Compton-scatters an electron above the Cherenkov threshold. The capture cross-section of deuterium is small, so the overall neutron detection efficiency was only about $14\,\%$, and the resulting Cherenkov light is similar in amount and in isotropy to a charged-current event: the separation between the two channels rests on a statistical fit to the distributions of energy, radius and direction.

Phase II, salt. Two tonnes of sodium chloride were dissolved in the heavy water. Neutrons then capture predominantly on ^{35}Cl, which has a far larger capture cross-section and de-excites through a cascade of photons totalling 8.6 MeV (1.38×10^{-12} J). Both the efficiency and the discrimination improve: efficiency rises to about 40 %, and a multi-photon cascade produces a markedly more isotropic pattern of hit tubes than the single electron of a charged-current event, so the two channels separate on event topology rather than on energy alone.

Phase III, discrete counters. The salt was removed and an array of ^{3}He proportional counters — some forty vertical strings, about 400 m of counter in total — was deployed in the heavy water. A neutron captured on ^{3}He gives $n + {}^{3}\mathrm{He} \to p + {}^{3}\mathrm{H}$ with 764 keV released, detected as an ionization pulse in the counter and *not* in the photomultipliers at all. The neutral-current measurement is then completely decoupled from the Cherenkov measurement of the charged current, so the statistical anticorrelation between the two channels that limited the first two phases disappears.

114.2.6 Liquid scintillator detectors

An organic liquid scintillator emits some ten thousand photons per megaelectronvolt of deposited energy, against the few hundred of a Cherenkov radiator, and it emits them isotropically and with no velocity threshold. The trade is exact: the scintillator sees far lower energies, and it loses all directional information. For a reactor experiment, where the source direction is known and the energies are a few megaelectronvolts, that is the right trade.

The reaction is the one Reines and Cowan used, inverse beta decay Equation (101.3) on the free protons of the hydrocarbon, with the threshold of Equation (101.4), 1.806 MeV. The signature is the delayed coincidence described in Section 101.1.3: a prompt pulse from the positron carrying $E_{\bar{\nu}} - 0.78$ MeV plus its two annihilation photons, followed by a delayed pulse when the neutron, having thermalized, is captured. That the reaction is charged-current on a proton means it detects $\bar{\nu}_e$ and nothing else, which is what makes a reactor experiment a pure disappearance measurement.

KamLAND placed 1 kt of liquid scintillator — about 80 % dodecane and 20 % pseudocumene by volume, with 1.5 g/L of the fluor PPO — inside a balloon of nylon and ethylene-vinyl-alcohol copolymer 13 m in diameter, itself suspended in 1800 t of non-scintillating mineral oil inside a steel sphere 18 m in diameter carrying 1879 photomultipliers, for a photocathode coverage of about 34 %. A water Cherenkov outer detector of 3.2 kt vetoes cosmic-ray muons. It occupies the cavern that held Kamiokande, under the same 2700 m of water equivalent [2049]. Neutrons capture on the free protons of the scintillator itself, giving a 2.22 MeV photon (3.56×10^{-13} J) after a mean capture time near 200 µs.

What makes KamLAND a solar-parameter experiment rather than a reactor experiment is its position. Japan's fifty-odd power reactors, some 70 GW of thermal power in total, are distributed at distances from 140 km to several hundred kilometres, with a flux-weighted mean baseline near 180 km. A thermal power P produces antineutrinos at a rate of about

$$\frac{\mathrm{d}N}{\mathrm{d}t} \approx \frac{6P}{200\,\mathrm{MeV}} = 1.9 \times 10^{20}/\mathrm{s} \times \frac{P}{1\,\mathrm{GW}}\,, \tag{114.19}$$

since fission releases about 200 MeV and the resulting fragments beta-decay about six times each on the way to stability, so the Japanese reactor complex is a source of order 1.3×10^{22}/s. At 180 km that is a flux of order $1 \times 10^{10}/(\mathrm{m}^2\,\mathrm{s})$ at the detector — and, above the prompt-energy threshold actually analysed, about three hundred detected events per year per kilotonne of scintillator. Only part of the kilotonne is usable: the analysis is restricted to a fiducial volume of 408 t, well inside the balloon, so that the radioactivity of

the balloon film and of the surrounding mineral oil cannot reach it. That is why the first result rested on 54 events.

Daya Bay and RENO use the same reaction at a hundredth of the baseline and with a completely different logic. Daya Bay observes six reactors of 2.9 GW thermal each, 17.4 GW in total, from three underground halls: two near halls at flux-weighted baselines of about 470 m and 576 m under 250 m and 265 m of water equivalent, and a far hall at about 1648 m under 860 m [2189]. Each hall holds one or more functionally identical antineutrino detector modules. A module is a nest of three volumes: an inner acrylic vessel 3 m in diameter holding 20 t of liquid scintillator doped with 0.1 % gadolinium by mass; an outer acrylic vessel holding a further 20 t of undoped scintillator, which catches photons escaping from the inner volume; and 40 t of mineral oil, optically transparent and non-scintillating, separating both from the 192 photomultipliers of 20 cm diameter that view the assembly. The whole module is immersed in a pool of purified water 2.5 m deep instrumented as a Cherenkov muon veto, itself covered by resistive plate chambers.

Gadolinium is the essential ingredient. Its thermal-neutron capture cross-section is some five orders of magnitude larger than hydrogen's, so at 0.1 % doping the great majority of neutrons capture on gadolinium rather than on protons; the capture releases a gamma cascade totalling about 8 MeV (1.28×10^{-12} J) after a mean time near 28 µs. A delayed pulse at 8 MeV is far above natural radioactivity, whereas the 2.22 MeV of capture on hydrogen sits in the middle of it; and the shorter capture time narrows the coincidence window, cutting accidental backgrounds by a further order of magnitude. RENO is built on the same plan with two identical detectors of 16 t of gadolinium-loaded scintillator at about 294 m and 1383 m from the centre of a six-reactor array of 16.5 GW thermal [2188].

Borexino is the extreme case of the same technology, optimized not for reactor antineutrinos but for the lowest-energy solar neutrinos. It holds 278 t of pseudocumene with 1.5 g/L of PPO in a nylon vessel of 8.5 m diameter, inside a stainless steel sphere of 13.7 m carrying 2212 photomultipliers, inside 2.1 kt of water serving as shield and muon veto, at Gran Sasso under about 3800 m of water equivalent [2229]. It detects elastic scattering Equation (114.12) on electrons, which has no threshold, so the limit is set entirely by radiopurity. The achieved contamination is the lowest ever reached in a bulk material: below 1×10^{-19} g/g for ^{238}U and below 6×10^{-19} g/g for ^{232}Th, with an isotopic ratio $^{14}\mathrm{C}/^{12}\mathrm{C}$ near 3×10^{-18}. The last of these is irreducible — ^{14}C is chemically identical to ^{12}C and cannot be purified away — and its beta spectrum, with endpoint 156 keV, is what sets the lower edge of the *pp* measurement.

114.2.7 Accelerator beams

An accelerator gives what no natural source can: a beam whose flavour, energy spectrum, direction, time structure and intensity are all under the experimenter's control, so that the same detector can be compared with itself with the beam on and off, and a near detector can measure the unoscillated flux directly.

The production chain is the same one that makes atmospheric neutrinos (Section 115.7.2), run deliberately. Protons of tens of gigaelectronvolts strike a thin target of graphite or beryllium; the secondary pions and kaons are focused by magnetic horns — pulsed conductors carrying currents of order 200 kA whose toroidal field sign-selects and collimates one charge — and are allowed to decay in an evacuated or helium-filled tunnel a few hundred metres long. Charged pion decay is two-body, $\pi^+ \to \mu^+ \nu_\mu$, so the beam is almost pure ν_μ; the residual ν_e contamination, from three-body kaon decay and from muon decay in the tunnel, is at the level of 1 % and is the irreducible background of every appearance measurement. A beam dump and hundreds of metres of rock absorb the surviving hadrons and muons.

K2K sent a beam from the 12 GeV proton synchrotron at KEK, with a mean neutrino energy of 1.3 GeV, over 250 km to Super-Kamiokande, using the same detector that had

made the atmospheric measurement [2089].

MINOS used the NuMI beam from Fermilab's 120 GeV main injector over 735 km to a detector in the Soudan mine: 5.4 kt of alternating steel and plastic scintillator planes, magnetized to about 1.3 T so that the muon charge, and hence whether the parent was a neutrino or an antineutrino, can be read from the track curvature. A functionally similar near detector of 1 kt sat 1 km from the target [2100].

T2K and NOvA introduced the off-axis technique, which is the main advance in beam design of the last two decades. Because pion decay is two-body, the neutrino energy in the pion rest frame is fixed by the two masses alone,

$$E_\nu^* = \frac{\left(m_\pi^2 - m_\mu^2\right) c^4}{2 m_\pi c^2} = 29.79\,\mathrm{MeV} = 4.773 \times 10^{-12}\,\mathrm{J}\,, \tag{114.20}$$

and the Lorentz transformation to the laboratory at a fixed angle θ to the pion direction gives

$$E_\nu = \frac{E_\nu^*}{\gamma_\pi \left(1 - \beta_\pi \cos\theta\right)} = \frac{\left(m_\pi^2 - m_\mu^2\right) c^4}{2\left(E_\pi - p_\pi c \cos\theta\right)}\,, \tag{114.21}$$

where the second form follows from $\gamma_\pi = E_\pi / m_\pi c^2$ and $\gamma_\pi \beta_\pi = p_\pi c / m_\pi c^2$, so that $\gamma_\pi \left(1 - \beta_\pi \cos\theta\right) = \left(E_\pi - p_\pi c \cos\theta\right) / m_\pi c^2$. On axis ($\theta = 0$) and for $E_\pi \gg m_\pi c^2$ the denominator is $m_\pi^2 c^4 / E_\pi$ and the result is $E_\nu \approx 0.427\, E_\pi$: the neutrino spectrum simply copies the pion spectrum, which is broad. Off axis it does not. At fixed θ the right-hand side rises with E_π, turns over at $E_\pi \approx m_\pi c^2 / \theta$ and falls again, so across a wide band of parent energies E_ν is almost independent of E_π. A detector placed slightly off the beam axis therefore sees a narrow band of energies rather than the broad spectrum on axis, which is exactly what an oscillation measurement wants: all the flux is concentrated where φ is near $\pi/2$, and the high-energy tail that would otherwise feed the ν_e background is suppressed. T2K sends a beam from the 30 GeV J-PARC main ring 295 km to Super-Kamiokande at 2.5° (43.6 mrad) off axis, peaking at 0.6 GeV [2168]; Equation (114.21) at that angle gives 0.61 GeV at $E_\pi = 2$ GeV, 0.68 GeV at the stationary point $E_\pi = 3.2$ GeV and 0.62 GeV at $E_\pi = 5$ GeV. NOvA sends the NuMI beam 810 km to a 14 kt segmented liquid scintillator detector at Ash River, 14.6 mrad off axis, peaking near 2 GeV [2398]; there the same expression gives 1.84 GeV, 2.04 GeV and 1.85 GeV at $E_\pi = 6$ GeV, 9.6 GeV and 15 GeV.

OPERA solved the hardest problem of the field: detecting the appearance of ν_τ, whose charged-current interaction produces a tau that decays after travelling of order a millimetre. Nothing with electronic readout resolves that. OPERA used 1.25 kt of target in the form of about 150 000 bricks, each a stack of fifty-seven lead plates 1 mm thick interleaved with nuclear emulsion films — lead to supply the mass, emulsion to supply a position resolution of under a micrometre. Electronic trackers identified which brick had been hit; the brick was then extracted, developed and scanned automatically. The CNGS beam from CERN, of mean energy 17 GeV, travelled 730 km through the Earth to Gran Sasso [2247].

114.2.8 What all of them have in common

Three requirements recur in every one of these instruments, and it is worth stating them once.

Depth. The rate of cosmic-ray muons falls roughly as the inverse cube of the water-equivalent overburden over the range covered by Table 114.3. The depth an experiment needs is set by its signal rate and by whether muons produce a background that mimics its signal. Homestake, with a signal of 0.5 atoms a day and a muon-induced spallation background in the same channel, needed 4200 m of water equivalent; SNO, whose neutral-current signature is a free neutron and for which every muon-induced spallation neutron is an irreducible

fake, needed 6010 m; Daya Bay, whose delayed-coincidence tag rejects almost everything and whose signal rate is hundreds of events a day, manages with 250 m.

Radiopurity. Every material within sight of the target contributes. The dominant chains are ^{238}U and ^{232}Th, whose daughters emit betas and gammas up to a few megaelectronvolts, and ^{40}K. Above about 5 MeV natural radioactivity essentially stops, which is why every Cherenkov measurement in this chapter has an analysis threshold in the range 5–9 MeV and why reaching below that — Borexino's achievement — requires purity at the level of 1×10^{-19} g/g.

Calibration with a known source. A rate measurement is only as good as the efficiency it is divided by. The instruments in this chapter are calibrated with deployed radioactive sources of known strength: gamma sources of a few megaelectronvolts to fix the energy scale, neutron sources such as ^{252}Cf or Am–Be to fix the neutron detection efficiency that the SNO and reactor measurements rest on, and, in the gallium case, an artificial neutrino source that calibrates the whole apparatus end to end.

114.3 Procedure

114.3.1 The radiochemical run cycle

A radiochemical run is a single measurement lasting weeks, and it proceeds in five steps that have not changed since 1968.

Reset. The tank is purged, removing all product atoms accumulated so far, and a measured quantity of a stable isotopic carrier is added: about 0.1 cm^3 at standard temperature and pressure of ^{36}Ar or ^{38}Ar for Homestake, a comparable trace of stable germanium for the gallium detectors. The carrier is the key to the whole procedure: it is chemically identical to the product, it is present in a quantity that can be measured by ordinary means, and whatever fraction of it is recovered at the end is the fraction of the signal atoms recovered too. Chemical efficiency is thereby measured in every run rather than estimated once.

Exposure. The tank stands undisturbed for a period comparable with the product half-life — some 3×10^6–9×10^6 s at Homestake, that is between one and three months, and about 1.8×10^6 s for the gallium runs. The number of product atoms present at time t after a reset obeys

$$\frac{\mathrm{d}N}{\mathrm{d}t} = R - \lambda N\,, \qquad N(t) = \frac{R}{\lambda}\left(1 - \mathrm{e}^{-\lambda t}\right)\,, \tag{114.22}$$

with R the production rate and $\lambda = \ln 2 / t_{1/2}$ the decay constant. Exposing for one half-life recovers half the saturated population; exposing for three recovers seven eighths. There is no gain beyond about two half-lives, which is why the exposure time is what it is.

Extraction. A large volume of carrier gas is circulated through the target. At Homestake, several hundred thousand litres of helium were bubbled through the 615 t of perchloroethylene, sweeping the dissolved argon — a noble gas, chemically bound to nothing — into a charcoal trap at liquid-nitrogen temperature. At GALLEX, nitrogen was swept through the acid solution, carrying volatile $\mathrm{GeCl_4}$ with it. The extraction is repeated and the recovered carrier measured, giving efficiencies of about 95 %.

Counting. The purified product is loaded into a proportional counter of about 0.5 cm^3 together with a counting gas, and the counter is observed for several months — long enough that the decay curve itself can be fitted. Each pulse is recorded with its energy and its rise time.

Analysis. A candidate event must satisfy three independent conditions: its energy must match the K- or L-capture peak, its rise time must be that of a localized Auger cascade rather than of an extended ionizing track, and — the decisive one — the times of the run's candidate events must be distributed as an exponential of the known half-life on a flat

background. A maximum-likelihood fit to the time distribution separates the product decays from the constant counter background, and it is this fit, not a raw count, that yields the number quoted. In practice, of order twenty candidate pulses per run.

The result of one run is a production rate in atoms per day, converted to a capture rate per target atom by dividing by the target count of Equation (114.8) or Equation (114.10). The published number is the average over more than a hundred such runs.

114.3.2 Reconstructing a Cherenkov event

A water Cherenkov detector records, for every photomultiplier, the arrival time of the first photon and the integrated charge. Everything else is inference, and the inference proceeds in a fixed order.

Vertex. Cherenkov light from a point of origin arrives at the tubes at times $t_i = t_0 + n\,|\boldsymbol{x}_i - \boldsymbol{x}_0|\,/c$, with n the refractive index. Minimizing the spread of the implied t_0 over all hit tubes locates the vertex. With a timing resolution of a few nanoseconds per tube and several thousand tubes hit, the vertex resolution is a few tens of centimetres. This is what defines the fiducial volume: events reconstructed within about 2 m of the wall are discarded, because the wall and the tubes themselves are the dominant source of radioactivity.

Direction and ring. With the vertex fixed, the pattern of hit tubes is fitted to a cone of half-angle θ_c from Equation (114.11). The cone axis is the particle direction; for a multi-GeV muon the angular resolution is a degree or two.

Flavour. The fitted ring is tested for sharpness. The figure of merit is essentially the goodness of fit of a sharp-edged cone against that of a diffuse disc, and the classification is calibrated on beams of electrons and muons of known momentum injected into the detector, and cross-checked on the sample of cosmic-ray muons that stop in the detector and decay, which supplies a muon and, microseconds later, an electron from the same vertex.

Energy. The total collected charge, corrected for water attenuation, tube-by-tube gain and geometrical acceptance, is proportional to the track length and hence to the energy. The proportionality constant is fixed by deployed sources and by the end-point of the electron spectrum from stopping-muon decay, which is a known 52.8 MeV.

Zenith angle to baseline. The step that makes the atmospheric measurement possible is the last one. The reconstructed direction of the outgoing lepton approximates the direction of the incoming neutrino, to within the scattering angle of the interaction — of order 15° to 20° at 1 GeV and better at higher energy. The zenith angle θ_z of the neutrino then fixes the distance it travelled from its production point at an altitude $h \approx 15\,\mathrm{km}$,

$$L\,(\theta_z) = \sqrt{(R_\oplus + h)^2 - R_\oplus^2 \sin^2\theta_z} - R_\oplus \cos\theta_z\,, \tag{114.23}$$

which runs from 15 km directly overhead ($\cos\theta_z = 1$) through 437 km at the horizon ($\theta_z = 90°$, where the expression reduces to $\sqrt{2R_\oplus h + h^2}$) to 12 800 km directly below. A single detector therefore scans three orders of magnitude in L at fixed E, with the same target, the same efficiency and the same analysis — the cleanest L/E scan available anywhere in physics, and it costs nothing but the angular resolution.

114.3.3 Separating the three Sudbury channels

SNO's three reactions all end as Cherenkov light in the same detector — Equations (114.13) to (114.15) — and separating them is the entire experimental problem. The separation is statistical, not event by event, and rests on three distributions.

Energy. The charged-current electron carries essentially the full neutrino energy less 1.442 MeV, so its spectrum follows the ^{8}B spectrum. The neutral-current signal is monoenergetic in the sense that every neutron capture gives the same photon energy — 6.25 MeV on deuterium, 8.6 MeV on chlorine — so it appears as a bump.

Radial distribution. Charged-current and elastic events occur uniformly in the heavy water, so their density in the cube of the radius is flat out to the acrylic vessel. Neutrons diffuse before capturing, so the neutral-current events are depleted near the vessel wall.

Direction. Elastic scattering is strongly forward-peaked, so its events cluster at $\cos\theta_\odot \to 1$ with respect to the direction from the Sun. Charged-current events on deuterium have a mild backward asymmetry, roughly $1 - \frac{1}{3}\cos\theta_\odot$. Neutron captures are isotropic, having lost all memory of the incident direction during thermalization.

An extended maximum-likelihood fit to the joint distribution in energy, radius and direction returns the three normalizations at once. The three are anticorrelated — the fit can trade charged-current events against neutral-current events to some extent — and it is that anticorrelation, not the statistics, which dominated the uncertainty in the first phase. Adding salt sharpens the separation by adding a fourth variable, the isotropy of the hit pattern, which distinguishes a multi-photon cascade from a single electron; and the ^{3}He counters of the third phase remove the correlation entirely by moving the neutral-current measurement into a separate instrument.

The neutron detection efficiency, which multiplies the whole neutral-current result, is measured by deploying a calibrated ^{252}Cf neutron source at many positions in the vessel and mapping the response.

114.3.4 The delayed coincidence and the reactor flux model

A reactor experiment selects an antineutrino by requiring two pulses in sequence.

The prompt pulse is the positron: its kinetic energy plus the two 511 keV annihilation photons, so that the visible energy is $E_{\bar\nu} - 0.78\,\mathrm{MeV}$. The prompt spectrum is therefore a shifted copy of the antineutrino spectrum, which is what makes a reactor experiment a spectral measurement and not merely a counting one.

The delayed pulse is the neutron capture: 2.22 MeV on hydrogen after about 200 µs, or about 8 MeV on gadolinium after about 28 µs.

The coincidence conditions are a time window a few capture times wide, a spatial separation below a metre or so, and energy windows on both pulses. Their product rejects everything uncorrelated. The residual backgrounds are of three kinds, and all three are measured rather than modelled: accidental coincidences of two uncorrelated radioactive decays, whose rate is obtained by opening an off-time window; correlated events from fast neutrons produced by cosmic-ray muons, which give a proton recoil followed by the same capture and are measured from the muon-tagged sample; and the (α, n) reactions of ^{210}Po on ^{13}C in the scintillator, which are estimated from the measured polonium activity.

The flux prediction is the one place where a reactor experiment depends on something it does not measure. The antineutrino emission of a reactor core is computed from four ingredients: the thermal power, known to a few tenths of a per cent from the coolant flow and temperature rise; the fission fractions of ^{235}U, ^{238}U, ^{239}Pu and ^{241}Pu, which drift over a fuel cycle and are tracked by the reactor operator's own core simulation; the energy released per fission of each isotope; and the antineutrino spectrum per fission of each. The last is the weak link, carrying an uncertainty of a few per cent in normalization and having twice been revised. For KamLAND, where the observable is the survival probability itself, that uncertainty enters the result directly and is the second-largest systematic. For Daya Bay it very nearly cancels, and the next subsection is why.

114.3.5 The near-to-far ratio

The design of Daya Bay and RENO is the reason θ_{13} could be measured at the per-cent level within months of turning on, and the argument is short enough to give in full.

Proposition 114.2 (What a near-to-far ratio cancels)**.** *Let a reactor emit $S(E)\,\mathrm{d}E$ antineutrinos per unit time in $\mathrm{d}E$, and let detector d, at baseline L_d, contain N_d free protons and detect an interaction with efficiency $\varepsilon_d(E)$. Its counting rate is*

$$R_d = \frac{N_d}{4\pi L_d^2} \int \mathrm{d}E\, S(E)\, \sigma(E)\, \varepsilon_d(E)\, P\left(E, L_d\right) \,, \tag{114.24}$$

with σ the inverse-beta cross-section Equation (101.5) and P the survival probability. If two detectors are built identically, so that $\varepsilon_n(E) = \varepsilon_f(E)$ for all E, then the geometrically corrected ratio

$$\mathcal{R} := \frac{R_f}{R_n} \cdot \frac{N_n L_f^2}{N_f L_n^2} = \frac{\int \mathrm{d}E\, S\sigma\varepsilon\, P\left(E, L_f\right)}{\int \mathrm{d}E\, S\sigma\varepsilon\, P\left(E, L_n\right)} \tag{114.25}$$

is independent of the absolute normalization of S, of the absolute normalization of σ, and of the absolute normalization of ε. It depends on their shapes *in E only through the difference between $P(E, L_f)$ and $P(E, L_n)$, and it equals unity identically if P does not depend on L.*
Rests on Equation (101.5).

Proof. *Derives Proposition 114.2.*
Multiply the numerator and denominator of R_f/R_n by $4\pi L_f^2/N_f$ and $4\pi L_n^2/N_n$ respectively, which is what the geometric factor in Equation (114.25) does; the prefactors of Equation (114.24) cancel exactly. Write $S = S_0 s(E)$, $\sigma = \sigma_0 \hat{\sigma}(E)$ and $\varepsilon = \varepsilon_0 \hat{\varepsilon}(E)$ with the hatted functions normalized however one likes. The constants $S_0\sigma_0\varepsilon_0$ multiply both integrals and cancel, which is the first statement. If P is independent of L it comes out of both integrals as the same function and the two integrals are equal, which is the second. What survives is the ratio of two weighted averages of P over the same weight $s\hat{\sigma}\hat{\varepsilon}$, so only the difference of P between the two baselines is observable. □

The proposition converts a hard absolute measurement into an easy relative one, and it explains every feature of the apparatus. The detector modules are built to the same drawings, filled from the same batch of scintillator so that the proton density is common, and their masses measured to better than a part in a thousand; at Daya Bay two modules were placed in the same near hall specifically so that their measured rates could be compared, giving a direct experimental bound of a few tenths of a per cent on the residual difference in efficiency. The far hall is placed where Equation (114.4) makes the phase near maximal, the near halls where it is small, so that the numerator of Equation (114.25) is suppressed and the denominator is not.

The one thing the ratio does not cancel is the relative contribution of the six reactors to each hall, since the halls are at different distances from each core; that is why the baselines quoted in Section 114.2.6 are flux-weighted, and why the reactors' individual thermal powers must still be tracked.

114.3.6 Calibration, blindness and the treatment of systematics

Two procedural points deserve recording because they are what makes the numbers in Section 114.4 usable.

Calibration is end-to-end where possible. The gallium experiments are the only ones that could be calibrated with a neutrino source of known activity, and they were: a ^{51}Cr source of 6.3×10^{16} Bq lowered into the target tests the capture cross-section, the extraction chemistry, the counter efficiency and the analysis in a single measurement. Every other instrument is calibrated in pieces — an optical calibration for the light transport, a gamma source for the energy scale, a neutron source for the capture efficiency — and the composition of those pieces is itself a systematic.

The analyses were blind. SNO, KamLAND, Daya Bay and T2K each fixed their selection cuts, their efficiency corrections and their background estimates on a subset of the data,

or on data with an unknown offset applied, before looking at the number the experiment was built to measure. This matters more here than in most fields because the expected answer was known in advance: an experiment looking for a factor of three that everyone expects to find will find it if the cuts are chosen after the fact. The published SNO result was unblinded once, and the value quoted is the value that appeared.

114.4 Observations and data

114.4.1 The solar neutrino deficit

Table 114.4 gives the radiochemical results. Every entry is a rate per target atom, in SI, together with the prediction of the standard solar model of Table 114.2 and the ratio of the two.

Table 114.4: Radiochemical solar neutrino capture rates. The literature quotes these in solar neutrino units, one SNU being 10^{-36} captures per target atom per second; the SI column is that number written as a rate. Statistical and systematic uncertainties are combined in quadrature where the source gives them separately. Target counts are Equation (114.8) and Equation (114.10). Predictions are the BS05(OP) model of [2072]. The last column is what mattered: chlorine, which sees only ^{8}B, and gallium, which sees the *pp* flux fixed by Equation (114.6), are both short, and by different amounts.

Detector	Years	Measured (1/s/atom)	Predicted	Ratio
Homestake	1970–1994	$(2.56 \pm 0.23) \times 10^{-36}$	$(8.1 \pm 1.2) \times 10^{-36}$	0.32 ± 0.05
GALLEX/GNO	1991–2003	$(6.93 \pm 0.55) \times 10^{-35}$	$(1.26 \pm 0.08) \times 10^{-34}$	0.55 ± 0.05
SAGE	1990–2007	$(6.54 \pm 0.41) \times 10^{-35}$	$(1.26 \pm 0.08) \times 10^{-34}$	0.52 ± 0.04

The Homestake number, $(2.56 \pm 0.23) \times 10^{-36}$ captures per target atom per second [1950], is the average over one hundred and eight extractions between 1970 and 1994. Multiplied by the 2.16×10^{30} target atoms of Equation (114.8) it is a production rate of

$$R = 2.16 \times 10^{30} \times 2.56 \times 10^{-36}/\mathrm{s} = 5.54 \times 10^{-6}/\mathrm{s} = 0.478 \text{ atoms per day}\,, \qquad (114.26)$$

against a cosmic-ray-induced background in the same channel of 0.047 ± 0.013 atoms per day and a predicted solar rate near 1.5 atoms per day. Twenty-five years of running, in a target of six hundred tonnes, to establish a shortfall of one atom of argon per day.

The gallium results [1974, 2138] are the ones that closed the astrophysical escape route, and the reason is arithmetic rather than statistical. Their threshold of 233 keV admits the *pp* neutrinos, whose flux is fixed by Equation (114.6) to about one per cent from the observed solar luminosity alone. Of the predicted 1.26×10^{-34} captures per second per atom, some 55 % comes from the *pp* branch; even setting the entire ^{8}B and ^{7}Be contribution to zero and moving the whole luminosity-constrained flux into *pp* — which no solar model allows, but which brackets the astrophysical uncertainty — leaves a predicted rate of 7.4×10^{-35} per second per atom, against a measured 6.5×10^{-35} to 6.9×10^{-35}. The deficit is reduced by that extreme assumption but not removed, and the assumption itself is contradicted by the direct measurements of the two branches it sets to zero.

Phenomenon 114.3 (The solar electron-neutrino deficit)**.** The rate at which solar neutrinos convert ^{37}Cl into ^{37}Ar is about one third of the rate predicted by the standard solar model, and the deficit persisted through twenty-five years of running [1368, 1950]. It reappears in the real-time, directional elastic-scattering measurement, which also confirms that the

neutrinos come from the Sun [1822], and — decisively — at the far lower threshold of the gallium detectors, which reach the *pp* neutrinos whose flux is fixed by the solar luminosity itself and is therefore almost independent of the uncertain inputs of the model [1974, 2138]. The predicted fluxes are those of [2072].

Rests on Equations (114.5), (114.6) *and* (114.9).

Derivation. *Derives Phenomenon 114.3.*

What has to be shown is that the deficit cannot be absorbed into the solar model, and the argument is a bound, not a fit. Write the predicted capture rate of a gallium detector as a sum over branches,

$$\mathcal{C} = \sum_{b} \Phi_b \, \langle \sigma \rangle_b \, ,$$

with $\langle \sigma \rangle_b$ the cross-section for Equation (114.9) averaged over branch b. The branch fluxes Φ_b are model outputs and carry the uncertainties of Table 114.2, but they are not independent: they satisfy the luminosity constraint Equation (114.6), which the model does not supply and cannot violate, since it follows from the observed solar constant and the energy release of Equation (114.5) alone. Because the *pp* branch produces two neutrinos per completion of Equation (114.5) and the higher branches produce them at a lower rate per unit energy released, maximizing the astrophysical freedom means transferring as much flux as possible *out* of the branches gallium sees best. The extreme case sets $\Phi(^8\mathrm{B}) = \Phi(^7\mathrm{Be}) = 0$ and puts everything in *pp*, which *raises* the total flux slightly while lowering the mean energy. Fix $\langle \sigma \rangle_{pp}$ from the model's own bookkeeping: the *pp* branch supplies 55 % of the predicted 1.26×10^{-34}, that is 6.93×10^{-35}, at the *pp* flux $5.99 \times 10^{14}/(\mathrm{m^2\,s})$ of Table 114.2, so $\langle \sigma \rangle_{pp} = 1.157 \times 10^{-49}\,\mathrm{m^2}$. Scaling that contribution up to the whole luminosity-constrained flux $6.4 \times 10^{14}/(\mathrm{m^2\,s})$ of Equation (114.6) gives

$$\mathcal{C} \geq 6.4 \times 10^{14}/(\mathrm{m^2\,s}) \times 1.157 \times 10^{-49}\,\mathrm{m^2} = 7.40 \times 10^{-35}/\mathrm{s\ per\ atom}\,. \tag{114.27}$$

The measurements are $(6.93 \pm 0.55) \times 10^{-35}$ and $(6.54 \pm 0.41) \times 10^{-35}$, whose inverse-variance combination is $(6.68 \pm 0.33) \times 10^{-35}$: below the bound Equation (114.27) by 2.2 standard deviations. The luminosity constraint alone therefore leaves a real but not decisive tension — and it is a bound obtained by assuming something now known to be false, since the two branches it zeroes have since been measured. Sudbury's flavour-blind channel gives the total ^{8}B production flux $(5.09 \pm 0.64) \times 10^{10}/(\mathrm{m^2\,s})$ with no oscillation input at all (Table 114.6), and Borexino gives the ^{7}Be production flux $(4.99 \pm 0.13) \times 10^{13}/(\mathrm{m^2\,s})$ (Table 114.11) — the latter obtained from a measured ν_e rate by dividing out the survival probability, so it presupposes oscillation and is quoted here only for the branch normalization, not as independent evidence for it. Both agree with Table 114.2. With the branch fluxes fixed by observation rather than left free, the prediction returns to 1.26×10^{-34} and the measured rate is 0.53 of it: a factor of two, far outside the uncertainty of either the measurement or the model. No redistribution of the solar fluxes consistent with the Sun's observed luminosity, and none consistent with the branch fluxes now measured directly, reproduces the gallium rate.

Two escapes remain and both were closed. Either the capture cross-section of Equation (114.9) is wrong, or the detector efficiency is; and the ^{51}Cr source calibrations of Section 114.3.6 test the product of the two directly against a known activity, returning a ratio consistent with unity to within about a tenth (Section 114.6.4 records the residual discrepancy honestly, and it is far too small to account for a factor of two). What is left is that fewer electron neutrinos arrive than leave, and the deficit is real. □

114.4.2 The directional and real-time solar measurements

Kamiokande-II was the first detector to see solar neutrinos in real time and the first to see that they come from the Sun. Its signal is the elastic scattering Equation (114.12), whose

recoil electron retains the neutrino direction well enough that the distribution of events in $\cos\theta_\odot$, the cosine of the angle between the reconstructed electron direction and the direction away from the Sun, shows a clear forward peak on a flat background [1822]. The measured ^{8}B flux was

$$\Phi\left(^8\mathrm{B}\right)_{\mathrm{ES}} = (2.7 \pm 0.5) \times 10^{10}/(\mathrm{m}^2\,\mathrm{s})\,, \tag{114.28}$$

for an analysis threshold of 7.5 MeV in recoil electron energy. Against the ^{8}B entry of Table 114.2, $(5.69 \pm 0.91) \times 10^{10}/(\mathrm{m}^2\,\mathrm{s})$, that is a ratio of 0.48 ± 0.12. The ratio quoted in the source itself is 0.46 ± 0.15 [1822], taken against the solar model current in 1989; the two are not the same comparison, and this chapter keeps them apart for the reason given in Section 114.8.

Two things follow that the radiochemical experiments could not establish. The neutrinos are solar, not terrestrial or cosmic — the angular distribution says so directly. And the deficit is about a factor of two rather than the factor of three seen in chlorine, at the same ^{8}B branch. That difference is not an inconsistency: elastic scattering retains sensitivity to the other flavours through the neutral current, with the weight of Equation (114.16), so a detector that measures $\Phi_e + 0.154\,\Phi_{\mu\tau}$ must report a smaller deficit than one measuring Φ_e alone — if and only if $\Phi_{\mu\tau}$ is not zero. The comparison of the two numbers was in fact the first quantitative hint of flavour change, well before SNO measured it in a single detector; but it required comparing two experiments with different targets, thresholds and systematics, and no one was prepared to draw the conclusion from it.

114.4.3 Atmospheric zenith-angle dependence

Super-Kamiokande's atmospheric measurement is the first in this chapter that needs no external prediction at all.

Two observables carry it. The first is the *double ratio*

$$R := \frac{(N_\mu/N_e)_{\mathrm{data}}}{(N_\mu/N_e)_{\mathrm{expected}}}\,, \tag{114.29}$$

in which the absolute atmospheric flux — uncertain by some 20 % — cancels, leaving only the flavour ratio, which by Equation (115.61) is fixed at two by the decay chain $\pi \to \mu\nu_\mu$, $\mu \to e\nu_e\bar{\nu}_\mu$ and is reliable to a few per cent. The second is the *up-down asymmetry*

$$A := \frac{U - D}{U + D}\,, \tag{114.30}$$

with U and D the numbers of events with $-1 < \cos\theta_z < -0.2$ and $0.2 < \cos\theta_z < 1$ respectively, in which the flux normalization, the cross-section and the detector efficiency all cancel, because the same detector counts both samples. The atmospheric flux is up-down symmetric above a few gigaelectronvolts to better than one per cent — the geomagnetic field distorts it at lower energies, which is why the asymmetry is quoted for the multi-GeV sample.

Phenomenon 114.4 (Zenith-angle dependence of atmospheric muon neutrinos)**.** Muon neutrinos produced in cosmic-ray air showers reach an underground detector from every direction. Those arriving from above have travelled some 15 km; those arriving from below have crossed the Earth, some 12 800 km. At the same energy the upward-going muon-neutrino rate is about half the downward-going one, while the electron-neutrino rates show no such deficit; and the size of the deficit is a function of the ratio L/E of path length to energy, not of L or E separately [1954, 2057].

Rests on Equations (114.2)*,* (114.3)*,* (114.23) *and* (114.30)*.*

Table 114.5: The Super-Kamiokande atmospheric neutrino result, from 33 kt yr of exposure — 535 days of the fiducial 22.5 kt [1954]. The first uncertainty is statistical, the second systematic. R is Equation (114.29) and A is Equation (114.30). A value $R = 1$ and $A = 0$ is what no oscillation would give. The electron-like asymmetry is consistent with zero and the muon-like asymmetry is not, which is the whole result in two numbers.

Sample	Quantity	Value
Sub-GeV	R	$0.63 \pm 0.03 \pm 0.05$
Multi-GeV	R	$0.65 \pm 0.05 \pm 0.08$
Multi-GeV, μ-like	A	$-0.296 \pm 0.048 \pm 0.01$
Multi-GeV, e-like	A	$-0.036 \pm 0.067 \pm 0.02$

Derivation. *Derives Phenomenon 114.4.*

The formalism is Equation (114.2), derived in Chapter 103; what is derived here is the number. Take the atmospheric splitting $|\Delta m^2_{32}|\,c^4 = 2.455 \times 10^{-3}\,\mathrm{eV}^2$ and $E = 1\,\mathrm{GeV}$, and evaluate the phase Equation (114.3) at the two extreme baselines of Equation (114.23):

$$\varphi_{\mathrm{down}} = 1.26693 \times 2.455 \times 10^{-3} \times \frac{15}{1} = 0.047\,, \tag{114.31}$$

$$\varphi_{\mathrm{up}} = 1.26693 \times 2.455 \times 10^{-3} \times \frac{12800}{1} = 40\,. \tag{114.32}$$

For the downward path, $\sin^2\varphi = 2.2 \times 10^{-3}$ and Equation (114.2) gives a survival probability of $1 - 2.2 \times 10^{-3} \sin^2 2\theta_{23}$: no observable deficit. For the upward path the phase is forty radians and sweeps rapidly across the finite spread of L within a zenith bin and of E within an energy bin, so the oscillating factor averages to $\frac{1}{2}$ and

$$\langle P \rangle = 1 - \tfrac{1}{2}\sin^2 2\theta_{23} = 1 - \tfrac{1}{2} \times 0.9915 = 0.504\,, \tag{114.33}$$

using $\sin^2\theta_{23} = 0.546$ from Table 103.3, whence $\sin^2 2\theta_{23} = 4\sin^2\theta_{23}\cos^2\theta_{23} = 0.9915$. The predicted up-down asymmetry Equation (114.30) follows immediately:

$$A = \frac{\langle P \rangle - 1}{\langle P \rangle + 1} = \frac{0.504 - 1}{0.504 + 1} = -0.330\,, \tag{114.34}$$

against the measured $-0.296 \pm 0.048 \pm 0.01$ of Table 114.5 — agreement to within seven tenths of a standard deviation, from a calculation containing no free parameter once θ_{23} and Δm^2_{32} are taken from the global fit. The electron-like asymmetry is predicted to vanish at this splitting because θ_{13} is small (Phenomenon 114.8), so the transition here is predominantly $\nu_\mu \to \nu_\tau$; the measured -0.036 ± 0.067 is consistent with zero.

That the two extreme baselines differ is not by itself decisive — an absorption law would also deplete the upward sample. What excludes absorption is that the depletion of the intermediate zenith bins follows $\sin^2$ of Equation (114.3) as a function of $L(\theta_z)/E$, and that the *electron* sample crossing the same rock shows no depletion at all. Matter cannot absorb one flavour and not another at these energies, because the neutral current is flavour-blind and the charged-current cross-sections differ by less than a per cent above a gigaelectronvolt. □

114.4.4 Sudbury and the flavour-blind flux

Phenomenon 114.5 (The missing solar neutrinos arrive as other flavours)**.** Heavy water gives access to reactions of different flavour sensitivity in one detector. The charged-current rate on deuterium, sensitive to ν_e alone, corresponds to a flux about one third of the

Table 114.6: The Sudbury fluxes, for an analysis threshold of 5 MeV in effective electron kinetic energy [2021]. The source quotes them per square centimetre per second; these are the SI values. Statistical and systematic uncertainties are combined. $\Phi_{\mu\tau}$ is the derived quantity Equation (114.35), and the model prediction is the ^{8}B entry of Table 114.2. The point of the table is the first two rows: a detector that counts only electron neutrinos and a detector that counts all three, in the same tonne of heavy water during the same exposure, disagree by a factor of three.

Channel	Flavour sensitivity	Flux $(1/(\mathrm{m}^2\,\mathrm{s}))$
Charged current	ν_e only	$(1.76 \pm 0.11) \times 10^{10}$
Elastic scattering	$\Phi_e + 0.154\,\Phi_{\mu\tau}$	$(2.39 \pm 0.27) \times 10^{10}$
Neutral current	all three equally	$(5.09 \pm 0.64) \times 10^{10}$
$\Phi_{\mu\tau}$ (derived)	—	$(3.41 \pm 0.66) \times 10^{10}$
Model ^{8}B	—	$(5.69 \pm 0.91) \times 10^{10}$

standard solar model prediction; the neutral-current deuteron breakup, equally sensitive to all three active flavours, gives a total flux of about $5.1 \times 10^{10}/(\mathrm{m}^2\,\mathrm{s})$, in agreement with that prediction [2008, 2021].

Rests on Equations (114.13) *and* (114.14).

Derivation. *Derives Phenomenon 114.5.*

Write ϕ_e for the solar ν_e flux above threshold and $\phi_{\mu\tau}$ for the combined ν_μ and ν_τ flux. The charged-current reaction measures ϕ_e; the neutral-current reaction has the same cross section for all three active flavours and therefore measures $\phi_e + \phi_{\mu\tau}$. Subtracting,

$$\phi_{\mu\tau} = \phi_{\mathrm{NC}} - \phi_{\mathrm{CC}} = 5.09 \times 10^{10}/(\mathrm{m}^2\,\mathrm{s}) - 1.76 \times 10^{10}/(\mathrm{m}^2\,\mathrm{s}) = 3.33 \times 10^{10}/(\mathrm{m}^2\,\mathrm{s})\,, \quad (114.35)$$

which the published likelihood fit, propagating the correlations between the two channels, gives as $(3.41 \pm 0.66) \times 10^{10}/(\mathrm{m}^2\,\mathrm{s})$ — positive by more than five standard deviations. No solar-model input enters Equation (114.35) at all: the conclusion that neutrinos leave the Sun as ν_e and arrive partly as ν_μ and ν_τ survives whatever the predicted flux, which the neutral-current number then independently confirms.

Three consequences follow in order.

First, the flavour changed in flight, so the flavour states are not stationary states of the free Hamiltonian, hence not states of definite mass.

Second, by Equation (114.2) a non-zero transition probability requires $\Delta m^2 \neq 0$, so at least one neutrino mass is non-zero. The strictly massless neutrino of Chapter 96 is thereby excluded, and with it the Standard Model as originally written.

Third — and this is the part the subtraction alone does not give — the ratio

$$\frac{\phi_{\mathrm{CC}}}{\phi_{\mathrm{NC}}} = \frac{1.76}{5.09} = 0.35 \pm 0.05 \quad (114.36)$$

is the electron-neutrino survival probability at ^{8}B energies, and it is a measurement of an oscillation parameter and not merely of a deficit. Vacuum oscillation over 1.5×10^{11} m at these energies averages to $1 - \frac{1}{2}\sin^2 2\theta_{12} = 0.57$, which Equation (114.36) excludes. The matter-modified value predicted by Example 103.58 is $\sin^2\theta_{12} = 0.31$, or 0.29 once the third angle is restored, and that is what is measured. The Sudbury result is therefore not only the proof that flavour changes but the first quantitative confirmation of the Mikheyev–Smirnov–Wolfenstein mechanism. □

The internal consistency of Table 114.6 is worth one more line, because it is a check the experiment supplies on itself. Given ϕ_e and $\phi_{\mu\tau}$ from the first and third rows, Equation (114.16) predicts an elastic-scattering flux of $1.76 + 0.154 \times 3.41 = 2.29$, in units of

$1 \times 10^{10}/(\mathrm{m}^2\,\mathrm{s})$, against the measured 2.39 ± 0.27. Three channels, two unknowns, one consistency condition satisfied.

114.4.5 KamLAND and reactor antineutrinos

Phenomenon 114.6 (Reactor antineutrinos disappear, and the deficit oscillates)**.** Electron antineutrinos from power reactors at an average baseline near 180 km arrive with a survival probability of about 0.6, and that probability, measured as a function of L/E, is not a monotone suppression but an oscillatory function of it [2049, 2071]. The parameters extracted from this terrestrial, controlled-source measurement agree with those extracted from the Sun.

Rests on Equations (114.2) *and* (114.3)*.*

The first KamLAND exposure was 162 t yr — 145.1 days of live time in a fiducial mass of 408 t inside the kilotonne target, 145.1 days being 0.397 of a year and $0.397 \times 408\,\mathrm{t} = 162\,\mathrm{t\,yr}$ — and it yielded 54 antineutrino candidates above a prompt-energy threshold of 2.6 MeV, against 86.8 ± 5.6 expected in the absence of oscillation from the reactor bookkeeping of Section 114.3.4, on an estimated background of 1.0 ± 1.0 events [2049]. The ratio is

$$\frac{N_{\mathrm{obs}} - N_{\mathrm{bg}}}{N_{\mathrm{exp}}} = 0.611 \pm 0.085 \pm 0.041\,, \tag{114.37}$$

the first uncertainty statistical, the second systematic. A ratio of one is excluded at better than 99.95 %.

Derivation. *Derives Phenomenon 114.6.*

The distinction that carries the argument is between a deficit and an oscillation: any error in the flux normalization produces the first, only interference produces the second. Evaluate the phase Equation (114.3) at the reactor scale, with $\Delta m^2_{21} c^4 = 7.53 \times 10^{-5}\,\mathrm{eV}^2$, $L = 180\,\mathrm{km}$ and the detected antineutrino energies E between 3.4 MeV and 8 MeV:

$$\varphi = 1.26693 \times 7.53 \times 10^{-5} \times \frac{180}{E/\mathrm{GeV}} = \frac{1.72 \times 10^{-2}}{E/\mathrm{GeV}}\,, \tag{114.38}$$

which runs from $\varphi = 5.05$ at the low end of the detected range to $\varphi = 2.15$ at the high end. That is nearly one complete oscillation *within a single detector's energy acceptance* — neither negligible, as at the reactor baselines of a few hundred metres, nor so large that it averages away, as for upward-going atmospheric neutrinos. In that regime two things are true at once. The survival probability averaged over the spectrum approaches the asymptotic value

$$\langle P \rangle \to 1 - \tfrac{1}{2}\sin^2 2\theta_{12} = 1 - \tfrac{1}{2} \times 0.851 = 0.57\,,$$

which Equation (114.37) matches within its uncertainty. And the *shape* of the variation across the prompt spectrum is measurable, because φ changes by 2.90 radians across it. The period of $\sin^2\varphi$ is π, so that swing is 0.92 of one full oscillation, and it straddles both a maximum of $\sin^2\varphi$ (at $\varphi = 3\pi/2 = 4.71$) and a minimum (at $\varphi = \pi$): the survival probability rises and falls again inside the accepted range rather than drifting monotonically. Observing the shape is what converts a missing flux into a determination of Δm^2_{21}: the position of the first minimum in φ fixes $\Delta m^2_{21} L/E$, and L and E are both known.

It is worth being explicit about why no rival hypothesis survives. Neutrino decay would give a survival probability falling monotonically as $\exp(-\alpha L/E)$; decoherence would give a monotone approach to a constant; an error in the reactor spectrum would scale the prediction smoothly. All three are monotone in L/E and none returns. The measured spectrum does return, and that is the signature of interference and of nothing else. □

Remark 114.7 (Why the coherence condition is not violated here). KamLAND is the one experiment in this chapter where the wave-packet caveats of Remark 103.51 could have mattered, since its baseline is long and its energy low. The coherence length Equation (103.53) at $E = 3\,\mathrm{MeV}$ and $\Delta m^2_{21}c^4 = 7.53 \times 10^{-5}\,\mathrm{eV}^2$ is $6.8 \times 10^3\,\mathrm{km}$ if the relevant localization is the atomic scale of the decaying fission fragment, against 180 km of baseline — a margin of thirty-eight — and 0.68 km if it were the nuclear scale, in which case coherence would be lost hundreds of times over and no oscillatory pattern could appear at all. KamLAND sees one. The observation is therefore also a measurement of the production localization, and it comes out atomic.

114.4.6 The third mixing angle

Table 114.7: The short-baseline reactor measurements of θ_{13}, both published in 2012. $\mathcal{R}$ is the geometrically corrected far-to-near ratio Equation (114.25); a value of one would mean no oscillation. Uncertainties are statistical then systematic. The two experiments use the same technique at different sites and reactor complexes, and their systematic errors are largely independent.

Experiment	Exposure	$\mathcal{R}$	$\sin^2 2\theta_{13}$
Daya Bay [2189]	55 days	$0.940 \pm 0.011 \pm 0.004$	$0.092 \pm 0.016 \pm 0.005$
RENO [2188]	229 days	$0.920 \pm 0.009 \pm 0.014$	$0.113 \pm 0.013 \pm 0.019$

Phenomenon 114.8 (The third mixing angle is small but not zero)**.** Comparing near and far identical detectors at reactor complexes reveals a deficit corresponding to $\sin^2 2\theta_{13}$ of order 0.09, measured independently and consistently by two experiments [2188, 2189]. It is the smallest of the three mixing angles, and many had expected it to vanish; that it does not is the precondition for any observation of leptonic CP violation.

Rests on Theorem 103.54, Equation (114.25)*, and Proposition 114.2.*

Derivation. *Derives Phenomenon 114.8.*

Two things have to be shown: that the three-flavour survival probability at a reactor baseline of a kilometre or two reduces to a two-flavour form governed by θ_{13} alone, and that the ratio Equation (114.25) then reproduces the measured number.

The survival probability. The exact three-flavour expression for $\bar{\nu}_e$ disappearance, obtained from Theorem 103.54 with the CP-odd term absent because the survival channel is CP-even, is

$$P_{\bar{e}\bar{e}} = 1 - \sin^2 2\theta_{13}\left(\cos^2\theta_{12}\sin^2\varphi_{31} + \sin^2\theta_{12}\sin^2\varphi_{32}\right) - \cos^4\theta_{13}\sin^2 2\theta_{12}\sin^2\varphi_{21}\,, \tag{114.39}$$

with φ_{jk} the phase Equation (114.3) built from Δm^2_{jk}. At $L = 1.65\,\mathrm{km}$ and $E = 3.5\,\mathrm{MeV}$ the solar phase is

$$\varphi_{21} = 1.26693 \times 7.53 \times 10^{-5} \times \frac{1.65}{3.5 \times 10^{-3}} = 0.0450\,,$$

so $\sin^2\varphi_{21} = 2.0 \times 10^{-3}$ and the last term of Equation (114.39), with $\cos^4\theta_{13}\sin^2 2\theta_{12} = 0.81$, contributes 1.6×10^{-3} — a sixth of a per cent, below the experimental resolution. Meanwhile Δm^2_{31} and Δm^2_{32} differ by three per cent, so φ_{31} and φ_{32} may be identified, and the bracket collapses to $\sin^2\varphi_{31}$ since $\cos^2\theta_{12} + \sin^2\theta_{12} = 1$. Hence

$$P_{\bar{e}\bar{e}} \approx 1 - \sin^2 2\theta_{13}\sin^2\left(\frac{(\Delta m^2_{31}c^4)\,L}{4\hbar c\,E}\right)\,, \tag{114.40}$$

a two-flavour form in which θ_{13} appears alone. This is the reason the experiment can be done at all: there is a window in L/E in which one, and only one, of the three angles is visible.

The ratio. Evaluate Equation (114.40) at the Daya Bay baselines with $\Delta m^2_{31} c^4 = 2.53 \times 10^{-3}\,\mathrm{eV}^2$, $E = 3.5\,\mathrm{MeV}$ and $\sin^2 2\theta_{13} = 4 \sin^2\theta_{13} \cos^2\theta_{13} = 0.086$ from Table 114.9. The phases are $\varphi_{\mathrm{far}} = 1.511$ and $\varphi_{\mathrm{near}} = 0.430$, giving $\sin^2\varphi_{\mathrm{far}} = 0.996$ and $\sin^2\varphi_{\mathrm{near}} = 0.174$, so

$$P_{\mathrm{far}} = 1 - 0.086 \times 0.996 = 0.914\,, \qquad P_{\mathrm{near}} = 1 - 0.086 \times 0.174 = 0.985\,,$$

and by Proposition 114.2

$$\mathcal{R} = \frac{P_{\mathrm{far}}}{P_{\mathrm{near}}} = 0.928\,, \tag{114.41}$$

against the measured $0.940 \pm 0.011 \pm 0.004$. The one per cent residual is what the full analysis recovers by averaging over the six reactor cores, the two near halls at different baselines and the whole prompt spectrum rather than evaluating at a single representative energy; the point of the estimate is that the effect is of the size measured and arises from θ_{13} alone.

Why the ratio is trustworthy. By Proposition 114.2 the reactor flux normalization, the inverse-beta cross-section and the absolute detector efficiency all cancel in $\mathcal{R}$; what remains is the relative efficiency of two detectors built to the same drawings and filled from the same batch of scintillator, bounded experimentally at Daya Bay by comparing two modules in the same hall. A measurement whose dominant systematic is the difference between two nominally identical objects is a very different thing from one whose dominant systematic is a predicted flux, and that is why an angle previously bounded only from above was resolved at five standard deviations in fifty-five days. □

114.4.7 Accelerator beams: disappearance and appearance

The accelerator programme does three things the natural sources cannot: it confirms the atmospheric parameters with a source whose spectrum and composition are known independently, it observes the *appearance* of a flavour rather than the disappearance of one, and it provides the only access to the CP phase.

Table 114.8: Long-baseline accelerator results. The disappearance experiments count ν_μ charged-current events at the far detector against the number predicted from the near detector; the appearance experiments count events of a flavour the beam did not contain. Uncertainties are as quoted by the sources, statistical and systematic combined where they are given separately.

Experiment	L	Observable	Result
K2K [2089]	250 km	ν_μ disappearance	112 events, 158 ± 9 expected
MINOS [2100]	735 km	ν_μ disappearance	215 events, 336 ± 14 expected
T2K [2168]	295 km	ν_e appearance	6 events, 1.5 ± 0.3 expected
NOvA [2398]	810 km	$\lvert\Delta m^2_{32}\rvert\, c^4$	$(2.41 \pm 0.07) \times 10^{-3}\,\mathrm{eV}^2$
OPERA [2247]	730 km	ν_τ appearance	5 events, background 0.25

K2K's deficit corresponds to $\lvert\Delta m^2_{32}\rvert\, c^4$ near $2.8 \times 10^{-3}\,\mathrm{eV}^2$ at maximal mixing, and excludes the no-oscillation hypothesis at about four standard deviations [2089]; MINOS, with ten times the statistics and a magnetized detector, gives $\lvert\Delta m^2_{32}\rvert\, c^4 = 2.74 \times 10^{-3}\,\mathrm{eV}^2$ with an uncertainty of about 15 % and $\sin^2 2\theta_{23} > 0.87$ [2100]. Both agree with the atmospheric determination from a completely different source, which is the point of doing them.

The appearance measurements are the harder and the more informative. T2K's six electron-neutrino candidates on an expected background of 1.5 ± 0.3 were a two-and-a-half standard deviation indication in 2011 [2168], published months before the reactor measurements of Table 114.7 and consistent with them: $\nu_\mu \to \nu_e$ appearance requires $\theta_{13} \neq 0$ by the same three-flavour expression that gives Equation (114.40), so the two are measurements of the same parameter in different channels. OPERA's five tau candidates on a background of 0.25 close the three-flavour picture from the other side [2247]: the flavour that the atmospheric and accelerator disappearance measurements infer must be there, because the muon neutrinos go somewhere, is observed directly.

114.4.8 The global fit

All of the above, together with the solar and reactor spectra in full, is combined into a fit for six parameters — two independent squared-mass splittings, three mixing angles and one phase. Table 114.9 is the result. It repeats Table 103.3 deliberately, because an experiment chapter must state the numbers its measurements produce, and because the third column here is different: it records which measurement dominates each parameter.

Table 114.9: The measured oscillation parameters, from the global fit of [2356, 2427], with the measurement that dominates each. The splittings are quoted as $\Delta m^2 c^4$, an energy squared, per Remark 114.1. The sign of Δm^2_{32} is not measured and the value listed assumes the normal ordering; the inverted ordering gives $(-2.529 \pm 0.029) \times 10^{-3}\,\mathrm{eV}^2$. The phase δ_{CP} is constrained, not measured, and is recorded as such.

Parameter	Value	Dominant measurement
$\Delta m^2_{21} c^4$	$(7.53 \pm 0.18) \times 10^{-5}\,\mathrm{eV}^2$	KamLAND reactor spectrum shape
$\lvert\Delta m^2_{32}\rvert\, c^4$	$(2.455 \pm 0.028) \times 10^{-3}\,\mathrm{eV}^2$	accelerator ν_μ disappearance; atmospheric
$\sin^2\theta_{12}$	0.307 ± 0.013	Sudbury charged-current to neutral-current ratio; Borexino
$\sin^2\theta_{23}$	0.546 ± 0.021	atmospheric and accelerator ν_μ disappearance
$\sin^2\theta_{13}$	0.0220 ± 0.0007	short-baseline reactor near-to-far ratio
δ_{CP}	$\approx 1.2\pi$, poorly bounded	$\nu_\mu \to \nu_e$ against $\bar\nu_\mu \to \bar\nu_e$

Two properties of that table are physics rather than bookkeeping. The two splittings differ by a factor of about thirty-three, which is why the solar and atmospheric sectors decouple and why each could be measured before the other was understood. And two of the three angles are large — $\sin^2 2\theta_{23} = 0.99$ and $\sin^2 2\theta_{12} = 0.85$ — in flat contradiction to the near-diagonal quark mixing matrix of Section 103.4.2, whose largest off-diagonal magnitude is about 0.22. Two mixing matrices generated by the same mechanism in the same theory, one nearly the identity and one nearly democratic: that is a fact with no explanation. It belongs to the flavour puzzle of Section 133.1, which catalogues the Yukawa couplings the Standard Model accommodates without predicting, though the contrast between the two matrices is not itemized there and is recorded here.

114.5 Interpretation

114.5.1 What the data force

The interpretation of Section 114.4 is a chain of four steps, and it is worth setting it out as a chain, because each link excludes something the previous one did not.

Step one: the deficit is real. Three radiochemical detectors with different thresholds and different chemistry, one water Cherenkov detector with direction, and one heavy-water detector all find fewer solar electron neutrinos than the model predicts (Tables 114.4 and 114.6). The gallium threshold reaches a branch whose flux is fixed by Equation (114.6) independently of the model, so the deficit is not an artefact of the astrophysics.

Step two: the missing neutrinos are still there. Sudbury's neutral-current channel measures the flavour-summed flux and finds it in agreement with the model, while the charged-current channel finds a third of it (Equation (114.35)). The subtraction is model-free. Whatever happened to the electron neutrinos, they were not absorbed and they did not decay into something invisible.

Step three: the change depends on L/E and oscillates. Super-Kamiokande's zenith-angle distribution is a scan in L at fixed E over three orders of magnitude with one detector (Equation (114.23)), and the depletion follows $\sin^2$ of Equation (114.3). KamLAND's prompt-energy spectrum is a scan in E at fixed L, and the survival probability returns rather than falling monotonically (Equation (114.38)). Absorption, decay and decoherence are all monotone in L/E; interference is not.

Step four: therefore the neutrino has mass. By Equation (114.2) the phase is proportional to $\Delta m^2 c^4$, so a non-zero oscillation amplitude requires $m_2^2 \neq m_1^2$ and hence at least one non-zero mass. The left-handed neutrino of Chapter 96 has none, and the Standard Model as originally written is therefore false. That is the only place where a laboratory measurement contradicts it.

Three technical points belong here rather than in the observational sections.

Plane waves. The derivation of Equation (114.2) gives the two mass components the same momentum and the same flight time, both of which are approximations. The conditions under which they are good — production and detection regions small compared with the oscillation length, and wave packets still overlapping on arrival — are stated and evaluated in Remark 103.51, and the tightest case in this chapter is KamLAND, examined in Remark 114.7.

Three flavours. The two-flavour formula is used throughout the observational sections because the two splittings differ by a factor of thirty-three and each experiment sits in a window where one dominates. The exact three-flavour expression is Theorem 103.54, and Equation (114.39) shows explicitly how it reduces at a short reactor baseline, with the neglected term quantified at 1.6×10^{-3}. Wherever the two-flavour statement is used here, the correction is of order $\cos^4 \theta_{13} = 0.957$ or smaller.

What oscillation does not measure. The probabilities depend on Δm^2, on the mixing angles and on one phase. They do not depend on the absolute masses, on whether the mass term is of Dirac or of Majorana type (Proposition 103.47), or on the mechanism that generates the masses. Everything in Table 114.9 is silent on all three.

114.5.2 Matter effects: the MSW mechanism

The solar measurements are not measurements of vacuum oscillation, and treating them as such gives the wrong answer by a factor of two. The theory is Theorem 103.57; what belongs here is the evidence for it.

The mechanism in one sentence: matter contains electrons and contains no muons or taus, so ν_e alone can forward-scatter coherently through the charged current and acquires

the extra potential energy Equation (103.57), $V_e = \sqrt{2} G_F n_e$, which at the centre of the Sun is 1.22×10^{-30} J, that is 7.6×10^{-12} eV. Because the shift is flavour-dependent it changes the mixing angle, and by Equation (103.61) the effective mixing becomes maximal at the energy Equation (103.62), $E_{\text{res}} = 1.9$ MeV. That single number splits the solar neutrino spectrum in two.

Table 114.10: The measured electron-neutrino survival probability of solar neutrinos as a function of energy, which is the observable signature of the matter effect. The Borexino values are from the comprehensive pp-chain measurement [2297]; the Sudbury value is Equation (114.36). The predictions are those of Example 103.58: below the resonance energy of 1.9 MeV the propagation is essentially vacuum-like and $P_{ee} \to 1 - \frac{1}{2}\sin^2 2\theta_{12}$, above it the neutrino leaves the Sun as the mass eigenstate ν_2 and $P_{ee} \to \sin^2\theta_{12}$, both multiplied by $\cos^4\theta_{13}$. There is no free parameter: θ_{12} and θ_{13} come from Table 114.9. The two asymptotes are sharp predictions; the ^{7}Be and *pep* entries lie in the transition region and are quoted as a smooth interpolation between them, which is where the prediction is least sharp and where the data are weakest.

Branch	Energy	Measured P_{ee}	Predicted
pp	0.1–0.4 MeV	0.57 ± 0.09	0.55
^{7}Be	0.862 MeV	0.53 ± 0.05	0.54
pep	1.44 MeV	0.43 ± 0.11	0.50
^{8}B	> 5 MeV	0.37 ± 0.08	0.29
^{8}B (SNO)	> 5 MeV	0.35 ± 0.05	0.29

Table 114.10 is the test, and it is worth saying exactly what makes it one. A survival probability that is *constant* in energy is what vacuum oscillation over an averaged phase gives; the measured values are not constant, and they fall with energy in the direction and by the amount the matter effect predicts. Two energy decades, one curve, no adjustable parameter. The intermediate points — *pep* at 1.44 MeV, sitting almost exactly at the resonance — are where the transition happens, and they are also where the measurement is weakest, which is an honest statement about what has and has not been established: the two asymptotes are solid, the shape of the transition between them is not yet resolved.

Two further consequences of Theorem 103.57 are the subject of current work, and Remark 103.60 names them both.

Day–night asymmetry. Solar neutrinos arriving at night have crossed the Earth, whose electron density partially regenerates the ν_e component, so the night-time rate should exceed the daytime rate by a few per cent with no free parameter beyond those already measured. The predicted asymmetry for the ^{8}B flux is a few per cent and the measurements sit at about the level of two standard deviations: an indication, not an observation, and this treatise records it as one. No source for the day–night measurement is cited in this treatise; the size of the effect is stated here without one, and nothing later in the chapter rests on it.

The sign of the splitting. This is the deepest consequence. Vacuum oscillation depends on Δm^2 only through $\sin^2\varphi$, which is even in Δm^2 and therefore blind to its sign; the matter term is not, because $V_e > 0$ for matter and $V_e < 0$ for antimatter, so the resonance condition Equation (103.61) can be satisfied for neutrinos or for antineutrinos but not both. The solar data therefore determine the sign of Δm^2_{21} — ν_2 is heavier than ν_1 — which no vacuum measurement could. The same argument applied to Δm^2_{32} with the Earth's density in place of the Sun's is the route to the mass ordering, and it is why the long-baseline experiments of Table 114.8 matter beyond their statistics.

114.5.3 The PMNS matrix and the measured parameters

Collecting Table 114.9 into the mixing matrix Definition 103.46 gives magnitudes Equation (103.49): every entry of the first two rows is of order one half, and only $|U_{e3}| = \sin\theta_{13} = 0.148$ is small. This is Phenomenon 103.48.

Three observations about the parameter set, each of which is a statement about the evidence rather than about the theory.

The measurements are over-determined and they agree. The solar splitting is fixed by KamLAND's terrestrial reactor spectrum and, independently, by the energy dependence of the solar survival probability in Table 114.10; the atmospheric splitting by Super-Kamiokande's zenith distribution and, independently, by MINOS and NOvA with a controlled beam; θ_{13} by two reactor experiments at different sites and, independently, by T2K's appearance channel. In each case the sources are astrophysical against terrestrial, or disappearance against appearance, with unrelated systematics. That is the strongest form of confirmation an experimental result can have.

The parameter count is the honest statement of ignorance. Three neutrino masses, three mixing angles and one phase — seven new numbers, and nine if the neutrino is of Majorana type — are added to the Standard Model by the results of this chapter, and not one of them is predicted by anything. The free-parameter bookkeeping is in the appendix on the fundamental parameters of physics, which sets out how the count changes according to the mass mechanism assumed.

The contrast with quark mixing is unexplained. The quark mixing matrix of Section 103.4.2 is nearly the identity, its largest off-diagonal magnitude being 0.22 and its smallest 0.0037; the lepton mixing matrix is nearly democratic. Both arise, in the Standard Model extended by neutrino masses, as the same kind of misalignment between two bases. There is at present no evidence-based account of why the two should look so different, and it belongs among the mass and mixing puzzles of Section 133.1.

114.5.4 What oscillation cannot say

Phenomenon 114.9 (Neutrino masses are non-zero and very small)**.** Oscillation fixes two squared-mass splittings, $\Delta m^2_{21}c^4 \approx 7.5\times10^{-5}\,\mathrm{eV}^2$ and $|\Delta m^2_{31}|\,c^4 \approx 2.5\times10^{-3}\,\mathrm{eV}^2$ [2356, 2427]. The model-independent kinematic measurement of the tritium beta-decay endpoint bounds the effective electron-neutrino mass below about $0.8\,\mathrm{eV}/c^2$ [2399], while the cosmological bound on the sum of the three masses is about $0.12\,\mathrm{eV}/c^2$ — tighter, but dependent on the cosmological model [2348].

Rests on Equation (114.2) and Definition 103.61.

Derivation. *Derives Phenomenon 114.9.*

A splitting bounds the larger mass from below. From $m_3^2 - m_1^2 = \Delta m^2_{31}$ with $m_1^2 \geq 0$,

$$m_3c^2 \geq \sqrt{|\Delta m^2_{31}|\,c^4} = \sqrt{2.530\times10^{-3}\,\mathrm{eV}^2} = 0.0503\,\mathrm{eV}\,, \tag{114.42}$$

that is $m_3 \geq 8.97\times10^{-38}\,\mathrm{kg}$, some seven orders of magnitude below the electron mass of $9.109\times10^{-31}\,\mathrm{kg}$ and definitely not zero. The splitting substituted here is $\Delta m^2_{31}c^4 = 2.530\times10^{-3}\,\mathrm{eV}^2$, the value used in the reactor analysis of Section 114.4.6, and not the $|\Delta m^2_{32}|\,c^4 = 2.455\times10^{-3}\,\mathrm{eV}^2$ of Table 114.9; the two differ by Δm^2_{21} and the bound is stated for the pair 31. Summing over the three states in the normal ordering, with $m_1 \to 0$ and $m_2c^2 = \sqrt{\Delta m^2_{21}c^4} = 0.0087\,\mathrm{eV}$, gives the minimum total $\sum m_ic^2 \geq 0.059\,\mathrm{eV}$, which is within a factor of two of the cosmological bound and is the reason that bound is interesting.

Oscillation says nothing further. Equation (114.2) depends only on differences of squared masses, so the lightest mass is entirely unconstrained by it, and Equation (114.42) is a lower bound on one mass, not on the scale. The three quantities bounded by experiment — the

effective mass in beta decay, the effective Majorana mass in double beta decay, and the cosmological sum — are three different combinations of the same three masses, defined in Definition 103.61 as Equations (103.64) to (103.66), and cannot be merged into one number. Table 114.12 reports them separately for that reason. □

Remark 114.10 (The seesaw is a hypothesis and is named as one). The smallness of the masses bounded above — a factor of 10^7 below the electron — has an economical explanation in the seesaw mechanism, in which a heavy right-chiral singlet of mass M gives a light eigenvalue $m_\nu \approx m_D^2/M$. This treatise names it and stops there. There is no experimental evidence for the heavy state, none for the scale, and none distinguishing the seesaw from any other way of generating the same dimension-five operator; the discussion, with the honest verdict, is Remark 103.71. The same applies to leptogenesis, which Remark 103.72 records as untested and, at the relevant scale, untestable. Editorial rule 1 permits naming these programmes only to say that they lack evidence, and that is what is said.

114.6 Modern repetitions and precision

114.6.1 Borexino and the solar fusion chains

Borexino did what no earlier solar detector could: it measured the individual branches of the pp chain in real time, one at a time, at energies below a megaelectronvolt.

The measurement is a spectral decomposition. Elastic scattering Equation (114.12) has no threshold, so a monoenergetic neutrino line produces a recoil-electron spectrum with a sharp Compton-like edge at $T_{\max} = 2E^2/(m_e c^2 + 2E)$ — 665 keV for the 862 keV ^{7}Be line, 261 keV for the 420 keV *pp* endpoint. The observed spectrum is fitted as a sum of those shapes plus the known beta and alpha spectra of the residual contaminants, and the branch fluxes come out of the fit.

Table 114.11: Borexino interaction rates, in counts per day per 100 t of scintillator, and the neutrino fluxes inferred from them given the survival probabilities of Table 114.10. Fluxes are in SI; the sources quote them per square centimetre per second. The CNO measurement [2349] is the first detection of the cycle that powers stars heavier than the Sun, and it is a five-standard-deviation result on a rate of seven counts a day.

Branch	Rate (per day per 100 t)	Flux ($1/(\mathrm{m}^2\,\mathrm{s})$)	Source
pp	144 ± 16	$(6.6 \pm 0.7) \times 10^{14}$	[2229]
^{7}Be	48.3 ± 1.3	$(4.99 \pm 0.13) \times 10^{13}$	[2297]
pep	2.43 ± 0.42	$(1.27 \pm 0.19) \times 10^{12}$	[2297]
^{8}B	0.223 ± 0.017	$(5.68 \pm 0.39) \times 10^{10}$	[2297]
CNO	7.2 ± 3.0	$(7.0 \pm 3.0) \times 10^{12}$	[2349]

Three results follow from Table 114.11.

The *pp* flux of $(6.6 \pm 0.7) \times 10^{14}/(\mathrm{m}^2\,\mathrm{s})$ [2229] agrees with the luminosity constraint Equation (114.6) and with the model entry of Table 114.2. This is a direct check that the Sun is powered by the reaction Equation (114.5) at the rate its luminosity requires — and, because neutrinos leave the core immediately while photons take of order 10^5 years to diffuse out, that the solar output has been steady over that interval. This is the content of Phenomenon 52.23.

The energy dependence of the survival probability across two decades, Table 114.10, is the direct confirmation of the matter effect discussed in Section 114.5.2.

The CNO detection [2349] establishes that the cycle operates in the Sun at all, supplying about one per cent of its power. Its flux depends on the abundance of carbon, nitrogen and

oxygen in the solar core, so it is in principle an independent handle on the solar metallicity, on which the spectroscopic determination and the helioseismic one — the latter belonging to Section 52.6 — have disagreed for two decades. At the present uncertainty of about 40 % the measurement does not yet discriminate between them, and saying so is part of reporting it.

114.6.2 Supernova and astrophysical neutrinos

SN 1987A. On 23 February 1987 the collapse of a blue supergiant in the Large Magellanic Cloud, at about 5×10^4 pc, produced the only neutrino burst ever detected from a stellar collapse. Kamiokande-II recorded eleven events within 13 s with energies between about 7 MeV and 36 MeV [1793], and the Irvine–Michigan–Brookhaven detector recorded eight in 6 s [1789]. Nineteen events, and they remain the entire observational database of core-collapse neutrino emission.

The bound they give on the neutrino mass is kinematic and follows from the arrival-time spread. A neutrino of mass m and energy E travels at a group velocity $v/c = \left[1 - m^2c^4/E^2\right]^{1/2} \approx 1 - m^2c^4/2E^2$, so two neutrinos emitted together at energies E_1 and E_2 arrive separated by

$$\Delta t = \frac{D}{2c}\, m^2 c^4 \left(\frac{1}{E_1^2} - \frac{1}{E_2^2} \right) , \tag{114.43}$$

with D the distance. Requiring the observed spread to be no larger than the plausible emission duration bounds mc^2 below 17 eV, the value obtained in Equation (115.67) — weak by comparison with the laboratory limit of Table 114.12, but obtained over a baseline 10^{15} times longer than any terrestrial one, and hence a bound of a completely different kind. The burst itself is Phenomenon 115.48.

Neutrinos in ice. A cubic kilometre of Antarctic ice instrumented with photomultipliers detects the same atmospheric neutrinos as Super-Kamiokande, at energies from tens of gigaelectronvolts upwards, and its densely instrumented core has been used to measure the atmospheric oscillation parameters in that range [2292] — an independent measurement in a medium, a geometry and an energy range unrelated to the water Cherenkov detectors of Section 114.2.4. The same instrument established a diffuse astrophysical neutrino flux of much higher energy [2203] and associated one event with a flaring blazar [2293]. Those results belong to Section 115.7.4 and are named here only because the astrophysical flux arrives with a flavour composition that is itself an oscillation measurement over cosmological baselines: whatever the production ratio, the phases average completely and the observed composition should be near one to one to one.

114.6.3 Absolute mass: kinematics, cosmology, double beta decay

Three experiments bound three different combinations of the three masses, defined in Definition 103.61. They must be reported separately, and Table 114.12 does so.

The kinematic measurement. KATRIN measures the shape of the tritium beta spectrum within a few electronvolts of its 18.6 keV endpoint. A non-zero neutrino mass removes the last $m_\nu c^2$ of the spectrum and distorts what remains, by Proposition 101.20; the observable is m_β^2, an incoherent sum $\sum_i |U_{ei}|^2\, m_i^2$, and the measurement is of that squared quantity, so the fitted value may come out negative without anything being wrong. The current result is $m_\beta^2 c^4 = 0.26 \pm 0.34\,\mathrm{eV}^2$, which is consistent with zero and yields the limit of Table 114.12 [2399]. This is Phenomenon 101.21 and Phenomenon 103.62.

The cosmological bound. Massive neutrinos are relativistic when they decouple and non-relativistic today; they free-stream out of small-scale density perturbations and suppress the growth of structure below a characteristic scale. Fitting the microwave background together with the distribution of galaxies bounds $\sum_i m_i c^2 < 0.12\,\mathrm{eV}$ [2348]. It is the tightest

Table 114.12: The three bounds on the absolute neutrino mass scale, each on a different combination of the same three masses, and each with a different model dependence. The KATRIN limit is the only one that depends on nothing but kinematics. The cosmological limit is tighter by nearly an order of magnitude and rests on the whole cosmological model. The double beta decay limit exists only if the neutrino is its own antiparticle, and its conversion from a half-life to a mass carries a nuclear matrix element uncertain by a factor of two to three, which is why the mass bound is quoted as a range.

Observable	Bound	Source
m_β (tritium endpoint)	$< 0.8\,\mathrm{eV}/c^2$, 90 % CL	[2399]
$\sum_i m_i$ (cosmology)	$< 0.12\,\mathrm{eV}/c^2$, 95 % CL	[2348]
$\lvert\langle m_{\beta\beta}\rangle\rvert$ (double beta)	$< 36 - -156\,\mathrm{meV}/c^2$	[2414]

bound available and the most model-dependent one: it assumes the standard cosmological model, three species with standard thermal history, and a particular treatment of the nonlinear scales. Compare it with the minimum from Equation (114.42), $\sum_i m_i c^2 \geq 0.059\,\mathrm{eV}$ in the normal ordering, and the window is a factor of two. This is Phenomenon 103.68.

Neutrinoless double beta decay. If the neutrino is its own antiparticle then a nucleus may undergo 2β decay emitting two electrons and no neutrinos, violating lepton number by two units. It has not been observed. KamLAND-Zen, using enriched ^{136}Xe dissolved in the KamLAND scintillator, bounds the half-life above $2.3 \times 10^{26}\,\mathrm{yr}$ and the effective Majorana mass in the range of Table 114.12 [2414]. The range, rather than a number, is the nuclear matrix element: converting a half-life limit into a mass limit requires a many-body calculation whose results differ by a factor of two to three between methods (Remark 103.65, and the nuclear structure of Chapter 107). This is Phenomenon 103.64.

The three are logically independent, and a treatise of evidence must not collapse them. A positive double beta signal would establish lepton number violation and hence a Majorana mass term whatever the mechanism; a positive KATRIN signal would fix the scale model-independently; a cosmological detection would fix the sum within one cosmology. None has happened.

114.6.4 Anomalies and sterile searches

A treatise of evidence must state which of its anomalies have not gone away. Four are outstanding, and none of them is established.

LSND. A stopped-pion source at Los Alamos produced $\bar{\nu}_\mu$ of a few tens of megaelectronvolts, and a liquid-scintillator detector 30 m away found an excess of $\bar{\nu}_e$-like events: 87.9 ± 23.2 events above background, an apparent transition probability of $(2.64 \pm 0.81) \times 10^{-3}$, at about 3.8 standard deviations [2007]. At that L/E the required splitting would be $\Delta m^2 c^4$ of order $1\,\mathrm{eV}^2$ — two orders of magnitude above $\lvert\Delta m^2_{32}\rvert\, c^4$, and therefore a fourth mass state. By Equation (114.1) a fourth state cannot couple to the Z, so it would have to be sterile.

MiniBooNE. A different beam, a different baseline and a different detector technology found an excess of low-energy electron-like events consistent in size with LSND [2298]. Whether the excess is electrons or photons — which a Cherenkov detector cannot always distinguish, since a converted photon makes an electromagnetic shower too — was the central question, and what follows is the measurement that tests it.

MicroBooNE. A liquid-argon time projection chamber of 85 t in the same beam images the interaction vertex with millimetre resolution and separates a single electron from a converted photon by the ionization density at the start of the track. It finds no excess in the electron-neutrino channel across several final-state topologies [2397]. That does not by

itself close the question — it constrains the electron interpretation, not every interpretation of the MiniBooNE excess — but it removes the simplest one.

The gallium anomaly. The ^{51}Cr and ^{37}Ar source calibrations of Section 114.3.6, which exist to verify the detectors, returned rates below expectation: the ratio of measured to predicted is about 0.87 ± 0.05 when the four exposures are combined [1974, 2138]. A deficit at a baseline of a few metres would require a splitting $\Delta m^2 c^4$ of order $1\,\mathrm{eV}^2$, which is the LSND scale; alternatively the predicted cross-section for Equation (114.9) into the excited states of ^{71}Ge is wrong. The discrepancy is small enough not to affect any conclusion of Section 114.4.1 — it would move the gallium ratio from 0.55 to about 0.63, still far below one — and it has not been resolved.

Reactor rate anomaly. Measured reactor antineutrino rates at short baselines sit a few per cent below the predicted flux. Part of this was resolved by recalculating the fission spectra, and part remains; a spectral bump near 5 MeV in the measured prompt spectrum, seen by all the modern experiments and absent from every prediction, is a defect of the flux model and not of the neutrinos, since it appears identically at near and far detectors and therefore cancels in Equation (114.25). No source for the rate anomaly or for the spectral bump is cited in this treatise; both are stated here without one, and the near–far cancellation is the reason neither affects any oscillation result quoted above.

No sterile neutrino is established. The anomalies do not form a consistent picture with one another: the parameter regions preferred by LSND, by the reactor rates and by the gallium sources overlap only partially, and the disappearance searches that a sterile state would also require have found nothing. This chapter records them as open.

114.7 Open questions

114.7.1 The mass ordering

The sign of Δm^2_{32} is not measured. Normal ordering means $m_1 < m_2 < m_3$, the close pair at the bottom; inverted means $m_3 < m_1 < m_2$, the close pair at the top. Global fits prefer the normal ordering at a level between two and three standard deviations that depends on which datasets are combined [2356, 2427], which is a preference and not a measurement.

Two routes will settle it, and they are independent.

The matter effect in a long-baseline beam. By the sign argument of Section 114.5.2, the Earth's matter enhances $\nu_\mu \to \nu_e$ and suppresses $\bar{\nu}_\mu \to \bar{\nu}_e$ if the ordering is normal, and does the reverse if it is inverted. The effect grows with baseline and with energy, which is why NOvA's 810 km is more sensitive to it than T2K's 295 km, and why a longer baseline still would settle the question outright. The difficulty is that the same comparison is also the measurement of δ_{CP}, and the two enter with a partial degeneracy.

The vacuum interference pattern at a medium baseline. A reactor experiment at some 50 km, where the solar and atmospheric phases are comparable, sees the fast atmospheric oscillation modulating the slow solar one; the *phase* of the fast modulation relative to the slow envelope differs between the two orderings. This requires no matter effect and no beam, only an energy resolution of about three per cent — which is the hard part, since it is a factor of two better than any large scintillator has achieved.

The answer matters beyond bookkeeping. In the inverted ordering the effective Majorana mass of Table 114.12 has a lower bound of about $15\,\mathrm{meV}/c^2$, within reach of the next generation of double beta decay experiments, so a null result there would exclude the inverted ordering for a Majorana neutrino; in the normal ordering the effective mass can vanish by cancellation and no null result excludes anything. The cosmological sum likewise has different minima in the two orderings (Equation (103.75)).

114.7.2 The CP phase

The phase δ_{CP} is the one parameter of Table 114.9 that is constrained rather than measured. It is observable only through a difference between $\nu_\mu \to \nu_e$ and $\bar{\nu}_\mu \to \bar{\nu}_e$, because by Theorem 103.54 the CP-odd term is proportional to the leptonic Jarlskog invariant

$$J_\ell = \tfrac{1}{8} \sin 2\theta_{12} \sin 2\theta_{23} \sin 2\theta_{13} \cos\theta_{13} \sin\delta_{CP} \,, \tag{114.44}$$

which vanishes if any angle vanishes. That is why Phenomenon 114.8 reorganized the field: before 2012 it was possible that $\theta_{13} = 0$ and that leptonic CP violation was therefore unobservable in principle.

T2K's data disfavour CP conservation at about the 95 % level and prefer a value near $-\pi/2$ [2345]; NOvA's are compatible with CP conservation [2398]. Combining them gives a preference, not a measurement, and this treatise records it as such. Three degeneracies stand in the way and all three must be broken together: with the mass ordering, through the matter effect just described; with the octant of θ_{23}, since $|U_{\mu 3}|^2 = \sin^2\theta_{23}$ enters the appearance channel unsquared while disappearance sees only $\sin^2 2\theta_{23}$, which is symmetric about $\pi/4$; and with the overall normalization of the beam, which is what the near detector is for.

Leptogenesis is often named as the motivation. It is worth stating plainly, as Remark 103.72 does, that a measured δ_{CP} would not establish it: the phases entering the decays of the hypothetical heavy states are not the phases of Definition 103.46, and the connection is a model assumption and not a theorem.

114.7.3 Dirac or Majorana

Whether the neutrino is its own antiparticle is not answerable by any experiment in this chapter. Oscillation cannot answer it, by Proposition 103.47: the extra phases of a Majorana mass matrix cancel from every oscillation probability. The kinematic and cosmological observables cannot, being phase-independent. The only practical discriminator is a lepton-number-violating process, and the searches [2414] have found nothing — with the interpretation of any half-life limit further limited by the nuclear matrix elements of Remark 103.65 and Chapter 107.

The neutrino sector is therefore the one place where the Standard Model is known to be incomplete from laboratory data alone. Every other experimental anomaly in this book is either a measurement that agrees with the theory to remarkable precision, or an astrophysical observation whose interpretation requires assumptions the laboratory cannot check. Here the laboratory itself says the theory is wrong, and says almost nothing about what replaces it: not the mass scale, not the ordering, not the mechanism, not whether lepton number is conserved. Of that list Section 133.1 records the absolute mass scale and the Dirac–Majorana question; the ordering and the mechanism are recorded here and nowhere else in this book.

114.8 Primary references

The literature of this chapter divides into five groups, and the division matters because the groups differ in what they can be relied on for.

The theoretical origin. Pontecorvo's paper [1127] — the Russian original, whose English translation [1159] carries a later year and is the same work, so the bibliography lists it twice for one paper — proposes oscillation by analogy with the neutral kaon and is about neutrino–antineutrino transitions, not flavour transitions; the flavour-mixing formulation that this chapter actually tests is that of Maki, Nakagawa and Sakata [1237]. It is a common misattribution to credit the second idea to the first papers, and the distinction

is worth preserving. The matter effect is Wolfenstein's [1631], with the resonance and its consequences for solar neutrinos due to Mikheyev and Smirnov [1779]; note that the latter, though universally dated 1985, appeared in a 1986 issue, and the bibliography entry records the journal's own date. Giunti and Kim [2106] is the standard monograph and the place to go for the wave-packet treatment that Remark 103.51 summarizes.

The instrument papers. Davis, Harmer and Hoffman [1368] describes the chlorine detector and reports the first upper limit; the twenty-five-year result is Cleveland and collaborators [1950], which remains the authority for the Homestake apparatus, its extraction efficiencies and its backgrounds. For the gallium detectors, [1974] covers GALLEX and [2138] the third and final SAGE dataset, each with its own source calibrations. Hirata and collaborators [1822] is the Kamiokande-II solar measurement and [1793] the supernova burst from the same detector. The Super-Kamiokande atmospheric result [1954] is the paper that made the case, and it should be read for its treatment of the up-down asymmetry rather than for its best-fit parameters, which are superseded; the later L/E analysis of the same atmospheric sample is [2057]. Sudbury published the charged-current rate first [2008] and the neutral-current comparison a year later [2021]; only the second contains the flavour-blind measurement, and citing the first for it is an error that appears in the secondary literature. The salt phase, with its enhanced neutral-current sensitivity, is [2056], and the combined analysis of all three phases is [2208].

The reactor papers. KamLAND's first result [2049] established the deficit on fifty-four events; the demonstration that the survival probability oscillates as a function of L/E rather than falling monotonically came with later exposures, and is the spectral distortion of [2071], which is the source for the oscillatory dependence stated in Phenomenon 114.6. The final reactor analysis, taken across the shutdown of the Japanese reactors, is [2218]. Daya Bay [2189] and RENO [2188] are the θ_{13} measurements; the first is the more detailed on the near-to-far cancellation.

The accelerator papers. K2K [2089] and MINOS [2100] are the disappearance confirmations; T2K [2168] is the first appearance indication and [2345] its δ_{CP} constraint; NOvA [2398] is the current precision measurement of $|\Delta m^2_{32}|$ and θ_{23}; OPERA [2247] is the direct ν_τ appearance. The detection of the tau neutrino itself, which these presuppose, is [2016], and the two-neutrino experiment on which the whole notion of lepton flavour rests is [1223]. The original detection of the neutrino is [1085, 1097], and the counting of the light species is [2104].

The compilations and the model. The oscillation parameters of Table 114.9 are the Particle Data Group's [2427] and the global fit of Esteban and collaborators [2356]; the two use overlapping data and are not independent, and where they differ this chapter follows the former. The dataset shipped with this treatise for [2427] carries the particle masses used in the kinematic derivations; the constants are CODATA 2022 [2431]. The standard solar model is Bahcall, Serenelli and Basu [2072], whose BS05(OP) fluxes are Table 114.2; the much earlier prediction against which the original deficit was measured is [1365], and quoting the modern fluxes as though they were the ones Davis was compared with in 1968 misrepresents the history. Borexino's pp-chain and CNO results are [2229, 2297, 2349]. The absolute mass bounds are KATRIN [2399], Planck [2348] and KamLAND-Zen [2414]. The anomalies are LSND [2007] and MicroBooNE [2397], and the astrophysical neutrino flux is [2203, 2293].

Chapter 115

Cosmic Rays and Astroparticle Physics

Before there were accelerators there was the sky. Every charged particle species known in 1950 — the positron, the muon, the pion and the strange mesons — was first seen in the cosmic radiation, and the highest laboratory-confirmed particle energies still come from it and not from any machine: the atmosphere is struck by nuclei above $1 \times 10^{20}\,\mathrm{eV} = 16\,\mathrm{J}$, some seven orders of magnitude beyond the beam energy of the collider of Chapter 113. This chapter closes Part XI by treating that natural beam as what it is — a measured flux, with a spectrum, a composition, an arrival-time structure and an angular distribution, all quoted here with their uncertainties. It covers the discovery of the radiation and the geomagnetic experiments that showed the primaries are charged and predominantly positive; the particles found in it, which feed Chapters 41 and 95; the primary spectrum with its knee and ankle; extensive air showers and the detection techniques built on them; the astrophysical acceleration mechanisms that are consistent with the data; the Greisen–Zatsepin–Kuz'min cutoff and its observed counterpart; and the neutral messengers — gamma rays and neutrinos — that point back to their sources where charged nuclei cannot.

Its neighbours are the Standard Model chapters it uses as input (Chapters 102, 103 and 107) and the astrophysics it feeds (Chapters 48 and 52). The standing monograph is Gaisser, Engel and Resconi [2271]; current flux and composition numbers are taken from the Particle Data Group review [2427]. Nothing in this chapter is inferred from a mechanism that has not been observed: where the origin of a component is unknown — as it is above the ankle — the chapter says so.

Units and conventions are those fixed once for the whole part in Section 100.1.1: the metric $\eta_{\mu\nu} = \mathrm{diag}(+1,-1,-1,-1)$ of Notation 100.1, the mass shell $p^2 = m^2c^2$ of Equation (100.1), and $\hbar$ and c written out in every formula and every intermediate step. Nothing below is computed in units where they are set to one; Remark 100.4 gives the dictionary for reading this chapter against the literature, which does. Two conventions of the field need saying at the outset because they are not SI and are used on every page of it.

Notation 115.1 (Rigidity and atmospheric depth). The *rigidity* of a particle of momentum p and charge Ze is

$$R := \frac{pc}{Ze}\,, \tag{115.1}$$

an SI voltage, measured in volts. It is the natural variable for magnetic problems because a magnetic field bends two particles of the same rigidity along the same path whatever their mass or charge: the Larmor radius of Equation (115.8) depends on p and Ze only through R. Rigidities in this chapter are quoted in gigavolts and in volts.

The *atmospheric depth* X at a point is the mass of air per unit area lying above it along

the direction of travel,

$$X(h) := \int_h^\infty \rho\,(h')\,\mathrm{d}h'\,, \tag{115.2}$$

an SI areal mass density in $\mathrm{kg/m^2}$. The literature universally quotes it in $\mathrm{g/cm^2}$, and $1\,\mathrm{g/cm^2} = 10\,\mathrm{kg/m^2}$ exactly; both are given here. Its vertical value at sea level follows from hydrostatic equilibrium alone: the whole weight of the column is carried by the surface pressure, so

$$X_{\mathrm{atm}} = \frac{p_0}{g_0} = \frac{101\,325\,\mathrm{Pa}}{9.806\,65\,\mathrm{m/s^2}} = 1.0332 \times 10^4\,\mathrm{kg/m^2} = 1033\,\mathrm{g/cm^2}\,. \tag{115.3}$$

Every shower in Section 115.4 develops in that calorimeter and in no other.

115.1 Discovery of the cosmic radiation

115.1.1 The penetrating radiation

The instrument that found cosmic rays was the gold-leaf electroscope, and understanding what it measures is the whole of the argument. A charged conductor inside a sealed vessel of gas loses its charge because the gas is not a perfect insulator: ionising radiation creates free electron–ion pairs, the field between the conductor and the vessel wall sweeps them apart before they recombine, and the resulting current discharges the leaves. If the field is strong enough that essentially every pair created is collected — the *saturation* regime — the current no longer depends on the field, and counts ionisation events directly:

$$I = e\,q\,V\,, \tag{115.4}$$

with q the number of ion pairs produced per unit volume per unit time, V the volume of the vessel and e the elementary charge. The observable is q, an SI rate density in $1/(\mathrm{m^3\,s})$, though the period quoted it as ion pairs per cubic centimetre per second and this chapter gives both.

The numbers involved show why the measurement was difficult and why it was nevertheless decisive. A sealed litre vessel at sea level, with every known source of contamination removed, still shows $q \approx 1 \times 10^7/(\mathrm{m^3\,s})$, that is about ten ion pairs per cubic centimetre per second, and by Equation (115.4) this is a current

$$I = 1.602\,176\,634 \times 10^{-19}\,\mathrm{C} \times 1 \times 10^7/(\mathrm{m^3\,s}) \times 1 \times 10^{-3}\,\mathrm{m^3} \approx 1.6 \times 10^{-15}\,\mathrm{A}\,,$$

a femtoampere. No galvanometer of 1910 could see such a current; an electroscope, which integrates it on a capacitance of a few picofarads, turns it into a leaf deflection of a few volts per hour and sees it easily. That is the instrumental fact behind the whole discovery.

By 1900 the residual conductivity of sealed air was known to Wilson, to Elster and Geitel, and to Rutherford, and the standing explanation was gamma radiation from radioactive material in the ground and in the walls of the apparatus. The explanation is testable, because a gamma ray is attenuated exponentially in the mass it traverses. For photons of the few MeV that the hardest radium gamma rays carry, the mass attenuation coefficient of air is $\mu/\rho \approx 4 \times 10^{-3}\,\mathrm{m^2/kg}$, so at the ground-level density $\rho \approx 1.2\,\mathrm{kg/m^3}$ the linear coefficient is $\mu \approx 5 \times 10^{-3}/\mathrm{m}$ and the attenuation length is

$$\lambda_{\mathrm{air}} = \frac{1}{\mu} \approx 200\,\mathrm{m}\,.$$

Wulf therefore carried an electroscope of his own design to the top of the Eiffel Tower, 300 m above the ground, where a purely terrestrial source demanded a reduction by

$$\exp\!\left(-\frac{300\,\mathrm{m}}{200\,\mathrm{m}}\right) = \mathrm{e}^{-1.5} = 0.22\,,$$

that is to about a fifth of the ground value. What he measured was a reduction to roughly sixty percent [366] — nearly three times the ionisation the terrestrial hypothesis allows. A factor of three is a large discrepancy, but one tower, one instrument and an attenuation coefficient known only to a factor of two could not settle the question, and Wulf did not claim that it had.

Hess settled it by going higher. In seven balloon ascents in 1911 and 1912, the last reaching 5350 m, he carried three sealed electroscopes of different wall thickness and found that the ionisation first *fell* with altitude, through a shallow minimum near 1500 m, and then *rose*, reaching about twice its sea-level value near 5000 m [387]. He drew the only available conclusion: a radiation of very great penetrating power enters the atmosphere from above. He also excluded the obvious candidate for a source above, flying at night and during the solar eclipse of 17 April 1912 and finding no reduction — so the Sun is not the source of the penetrating component. Kolhörster extended the measurement to 9200 m and found the ionisation about ten times the ground value [405], which fixed the effect beyond argument.

Phenomenon 115.2 (Ionisation increases with altitude)**.** The rate at which a sealed electroscope loses its charge, after every known terrestrial contamination has been excluded, first falls with height above the ground and then rises. In the balloon ascents of 1911–1912 the ionisation passed through a minimum near 1500 m and by about 5000 m was roughly twice its sea-level value [387]; at 9200 m it was some ten times the ground value [405]. Flights by night and during a solar eclipse gave the same result, which excludes the Sun as the source [366, 387].

Rests on Equation (115.2).

Derivation. *Derives Phenomenon 115.2.*

Let the ionising agent be attenuated exponentially in the mass of absorber traversed, $I = I_0 \exp(-X/\lambda)$, with X the column density of Equation (115.2) and λ an attenuation length; this is the behaviour of any penetrating radiation with a fixed interaction cross section per unit mass. Two sources are in play and they see opposite overburdens.

For a source in the ground the absorber between source and instrument is the air *below* the instrument, whose column density *increases* with height; write it $\ell(h)$ with $\ell(0) = 0$ and $\mathrm{d}\ell/\mathrm{d}h = \rho(h) > 0$. For a source above the atmosphere the absorber is the depth $X(h)$ of Equation (115.2), which *decreases* with height, $\mathrm{d}X/\mathrm{d}h = -\rho(h) < 0$. The measured rate is the sum,

$$q(h) = q_g \exp\!\left(-\frac{\ell(h)}{\lambda_g}\right) + q_c \exp\!\left(-\frac{X(h)}{\lambda_c}\right) , \tag{115.5}$$

and differentiating,

$$\frac{\mathrm{d}q}{\mathrm{d}h} = \rho(h)\left[-\frac{q_g}{\lambda_g}\mathrm{e}^{-\ell/\lambda_g} + \frac{q_c}{\lambda_c}\mathrm{e}^{-X/\lambda_c}\right] .$$

The bracket is negative at $h = 0$ whenever the terrestrial term dominates there, and it grows monotonically with h because the first exponential decreases and the second increases. It therefore has exactly one zero, at the altitude where

$$\frac{q_g}{\lambda_g}\mathrm{e}^{-\ell(h)/\lambda_g} = \frac{q_c}{\lambda_c}\mathrm{e}^{-X(h)/\lambda_c} , \tag{115.6}$$

and that zero is a minimum. So the observed shape — fall, minimum, rise — is not merely consistent with an external source; it is what an external source superposed on a terrestrial one *must* produce, and a purely terrestrial source cannot produce it at all, since for $q_c = 0$ the bracket is negative everywhere and q falls monotonically. The two attenuation lengths are then read off the two limbs of the curve, and the upper limb gives a λ_c far larger than any then-known gamma radiation possessed, which is the sense in which the radiation is "penetrating". □

Remark 115.3 (What the altitude curve does not say). Phenomenon 115.2 establishes that something enters the atmosphere from outside. It says nothing about what that something *is*: an exponential absorption curve is produced by a photon beam, an electron beam and a proton beam alike, and the only parameter it yields is λ_c. This is why the question of the nature of the primaries stayed open for two decades after 1912 and was settled not by absorption but by geomagnetism (Sections 115.1.3 and 115.1.4). It is also worth recording that the rise does not continue indefinitely: above about 20 km the ionisation passes through a maximum, the Pfotzer maximum, and falls again, because what the electroscope records at any depth is not the primary but the shower it has made — the subject of Section 115.4.

115.1.2 Naming and the extraterrestrial origin

Millikan and Cameron settled the extraterrestrial origin by a different route and, in doing so, named the subject. Their idea was to use water as the absorber and mountain lakes as the vessel: an instrument lowered to depth d in a snow-fed lake at altitude h sits under a total column density

$$X_{\mathrm{tot}} = X_{\mathrm{atm}}(h) + \rho_{\mathrm{w}} d\,, \qquad \rho_{\mathrm{w}} = 1.0 \times 10^{3}\,\mathrm{kg/m^{3}}\,,$$

so that one metre of water is worth $1.0 \times 10^{3}\,\mathrm{kg/m^{2}} = 100\,\mathrm{g/cm^{2}}$, about a tenth of the whole atmosphere by Equation (115.3). Measuring in Muir Lake, at 3540 m, and in Arrowhead Lake, at 1530 m, they found that the two ionisation curves, plotted against X_{tot} rather than against depth in water, fell on one another [552]. That is exactly what Equation (115.5) predicts for the extraterrestrial term and it is what no terrestrial source can produce, because a terrestrial source has no reason to care about the air above the lake. They called the agent *cosmic rays*, and the name stuck.

Millikan drew a further conclusion that was wrong, and the way it was wrong is instructive. He read the attenuation curve as that of a gamma-ray beam and identified the photon energies with the binding energies released in the synthesis of helium and heavier elements from hydrogen — the "birth cries of atoms". The inference from an absorption curve to a photon energy requires a cross section, the cross section assumed was Compton scattering (Chapter 65), and the whole chain collapses if the primaries are not photons. Compton argued from the outset that they were charged particles. The dispute could not be settled by absorption measurements of any precision, for the reason given in Remark 115.3: attenuation curves measure λ, and λ alone does not identify a species. What settled it was an argument in which the charge appears explicitly, and the Earth supplies the apparatus.

115.1.3 The latitude effect

Clay, making ionisation measurements on the voyages between Java and the Netherlands, found the intensity systematically lower near the equator than at high latitude, by some 10 % [569]. Compton organised a worldwide survey with a single standardised ionisation chamber design at sixty-nine stations on every continent, and showed that the variation is ordered not by geographic latitude but by *geomagnetic* latitude λ_m, with the intensity flat above about $|\lambda_m| = 40^{\circ}$ and falling towards the geomagnetic equator [716]. A magnetic ordering can only mean one thing.

Phenomenon 115.4 (The primary intensity depends on geomagnetic latitude)**.** The cosmic-ray intensity measured at sea level falls by of order 10 % between geomagnetic latitude 40° and the geomagnetic equator, and the ordering parameter is the geomagnetic and not the geographic latitude [569, 716]. The primaries are therefore electrically charged, and the fall sets the scale of their momenta.

Rests on Equation (115.1).

Derivation. *Derives Phenomenon 115.4.*

A photon is undeflected by a magnetic field, so a photon primary produces no latitude effect of any size; the observation of one excludes photons as the dominant component. For a charged primary, the Earth's field acts as a momentum filter, and the scale of the filter follows from a single comparison of lengths.

Model the geomagnetic field as a dipole of moment m, so that in the equatorial plane at distance r from the centre

$$B(r) = \frac{\mu_0 m}{4\pi r^3} =: \frac{M}{r^3}\,, \qquad M = \frac{\mu_0 m}{4\pi} \approx 8.1 \times 10^{15}\,\mathrm{T\,m^3}\,, \tag{115.7}$$

the value of M being fixed by the measured surface field at the magnetic equator, $B\,(R_\oplus) \approx 3.1 \times 10^{-5}\,\mathrm{T}$ with $R_\oplus = 6.371 \times 10^6\,\mathrm{m}$. A particle of momentum p and charge Ze moving perpendicular to a field B turns on a circle of radius

$$r_L = \frac{p}{ZeB} = \frac{R}{cB}\,, \tag{115.8}$$

with R the rigidity of Equation (115.1); the second form shows that r_L depends on the particle only through R, which is the whole reason rigidity is used. A particle can reach the Earth from outside only if its trajectory is not turned round before it arrives, which requires the turning radius in the field it crosses to be at least of order the distance over which that field acts. Setting $r_L = r$ with $B = M/r^3$ gives the rigidity scale

$$R_{\mathrm{sc}}(r) = \frac{cM}{r^2}\,, \qquad R_{\mathrm{sc}}\,(R_\oplus) = \frac{2.998 \times 10^8\,\mathrm{m/s} \times 8.1 \times 10^{15}\,\mathrm{T\,m^3}}{(6.371 \times 10^6\,\mathrm{m})^2} = 6.0 \times 10^{10}\,\mathrm{V}\,, \tag{115.9}$$

that is 60 GV. The exact treatment is Störmer's: the motion of a charge in a dipole field has a first integral, and it partitions directions at each point into an allowed cone, from which particles of a given rigidity can arrive, and a forbidden one. The resulting *cutoff rigidity* for vertical incidence at geomagnetic latitude λ_m is [723]

$$R_c\,(\lambda_m) = \frac{cM}{4r^2}\cos^4\lambda_m = 14.9\,\mathrm{GV} \times \cos^4\lambda_m\,, \tag{115.10}$$

which is Equation (115.9) with the numerical coefficient $\frac{1}{4}$ that the exact integration supplies. No particle of rigidity below R_c arrives vertically at that latitude.

Derivation pending #203

The Störmer theory itself: the first integral of the motion of a charge in a magnetic dipole field, the partition of directions at each point into an allowed and a forbidden cone that follows from it, and the integration supplying the coefficient one quarter in the vertical cutoff together with the full azimuthal form used in the east–west subsection below. Belongs in Appendix A; both formulae are quoted here from the source cited beside them, and only their consequences are derived

The latitude effect follows at once. At $\lambda_m = 50°$ the cutoff is $14.9\,\mathrm{GV} \times \cos^4 50° = 2.5\,\mathrm{GV}$, while at the geomagnetic equator it is the full 14.9 GV: a detector at high latitude sees the primary spectrum down to a rigidity six times lower than one at the equator, and therefore sees more particles. The *size* of the effect is then a measurement of the spectrum. If the integral intensity above rigidity R behaves as $R^{-\Gamma}$, the ratio of equatorial to polar rates is $(14.9/2.5)^{-\Gamma}$, so a 10 % deficit requires $\Gamma \approx 0.06$ for the ionisation-chamber response — that is, the ionisation is dominated by secondaries from primaries well above both cutoffs, which is why the effect is a ten-percent effect and not a factor of six. The qualitative conclusion is independent of all of this: only a charged primary can produce a latitude effect at all. □

115.1.4 The east–west asymmetry

The latitude effect shows the primaries are charged; it does not show the sign of the charge, because Equation (115.10) depends on Z only through the rigidity and is the same for positive and negative particles arriving vertically. Rossi saw that the sign appears as soon as one looks away from the vertical: for a given latitude the Störmer cutoff depends on the azimuth of arrival, and it does so with opposite senses for the two signs [668]. The full form of the cutoff, for arrival at zenith angle θ and azimuth ξ measured clockwise from magnetic north — so that arrival from the east is $\xi = 90°$ and arrival from the west is $\xi = 270°$ — is [723]

$$R_c\left(\lambda_m, \theta, \xi\right) = \frac{cM\cos^4\lambda_m}{r^2\left[1 + \left(1 - \operatorname{sgn}(Z)\sin\xi\sin\theta\cos^3\lambda_m\right)^{1/2}\right]^2}\,. \tag{115.11}$$

At the geomagnetic equator and on the horizon, $\cos\lambda_m = 1$ and $\sin\theta = 1$, so the bracket carries $\sin\xi = -1$ for arrival from the west and $\sin\xi = +1$ for arrival from the east; at the zenith $\sin\theta = 0$ and the bracket is $[1+1]^2 = 4$, which returns Equation (115.10). With $cM/r^2 = 59.8\,\mathrm{GV}$ from Equation (115.9), this gives for a positively charged primary $R_c = 10.2\,\mathrm{GV}$ from the west and $59.6\,\mathrm{GV}$ from the east, against $14.9\,\mathrm{GV}$ from the zenith; for a negative primary the two horizontal values are exchanged. Since the spectrum falls with rigidity, the lower cutoff is the more intense direction, and the sign of the excess is the sign of the charge.

Phenomenon 115.5 (The primaries arrive preferentially from the west)**.** Near the geomagnetic equator the cosmic-ray intensity measured with a counter telescope at a fixed zenith angle is greater when the telescope points west than when it points east [712, 722], as had been predicted for a flux of charged particles [668]. The primaries are therefore charged, and predominantly positively charged.

Rests on Equation (115.11).

Derivation. *Derives Phenomenon 115.5.*

The sign can be got without Equation (115.11), from the Lorentz force alone. Set up axes at the geomagnetic equator: $\hat{\boldsymbol{z}}$ upward, $\hat{\boldsymbol{x}}$ east, $\hat{\boldsymbol{y}}$ north. There the geomagnetic field is horizontal and points north, $\boldsymbol{B} = B\hat{\boldsymbol{y}}$ with B of order $3\times10^{-5}\,\mathrm{T}$. A primary of charge q arriving vertically has velocity $\boldsymbol{v} = -v\hat{\boldsymbol{z}}$, and feels

$$\boldsymbol{F} = q\,\boldsymbol{v}\times\boldsymbol{B} = -qvB\,(\hat{\boldsymbol{z}}\times\hat{\boldsymbol{y}}) = +qvB\,\hat{\boldsymbol{x}}\,,$$

since $\hat{\boldsymbol{z}}\times\hat{\boldsymbol{y}} = -\hat{\boldsymbol{x}}$. For $q > 0$ the force is eastward: as the particle descends, its velocity is deflected towards the east, so that it reaches the ground travelling east of vertical — that is, arriving *from* the west. Positive primaries therefore produce a western excess and negative primaries an eastern one, and the sign of the observed excess fixes the sign of the charge. A neutral primary, which was still the alternative hypothesis in 1933, produces no azimuthal asymmetry whatever. The measurement thus decides between the two hypotheses without any assumption about the energy spectrum or the composition, which is exactly what the absorption curves of Section 115.1.2 could not do.

The size of the asymmetry follows from the cutoffs. Writing the integral intensity above rigidity R as $J(>R) \propto R^{-\Gamma}$, the west-to-east ratio at the horizon is

$$\frac{J\left(>R_c^W\right)}{J\left(>R_c^E\right)} = \left(\frac{10.2\,\mathrm{GV}}{59.6\,\mathrm{GV}}\right)^{-\Gamma} = (5.84)^{\Gamma}\,,$$

so that even a modest Γ gives a large effect: the observed excess at Mexico City was of order 10 % at moderate zenith angle [722], growing towards the horizon, and it was found independently and with the same sign by Alvarez and Compton [712]. The primaries are positive. They were later identified as protons and heavier nuclei, the composition measurement being the subject of Section 115.3.4. □

115.1.5 Solar modulation and geomagnetic transients

The geomagnetic field is not the only magnetic filter between a Galactic source and a detector on Earth. The solar wind carries the solar magnetic field outward through the heliosphere, drawn into an Archimedean spiral by the Sun's rotation, and cosmic rays entering against that outflow must diffuse upstream. The consequence is observed directly: the intensity of low-energy cosmic rays is *anticorrelated* with the eleven-year sunspot cycle, being highest at solar minimum, and it drops sharply — by a few percent within hours — after a solar disturbance, recovering over the following days. Forbush established the transient decreases from continuously running ionisation chambers [791], and they carry his name; the eleven-year anticorrelation he drew out of the same network over the two decades that followed [1037].

The quantitative description is a transport equation for the phase-space density $f(\boldsymbol{r}, p, t)$ of cosmic rays, combining outward convection with the wind at speed u, spatial diffusion, and the adiabatic energy loss a particle suffers in an expanding flow. Its spherically symmetric steady state, when diffusion and convection balance, has a first integral of great practical use: the intensity observed at 1 au is that of the unmodulated interstellar spectrum evaluated at a higher energy, the particle having paid a toll

$$\Delta E = Ze\Phi \tag{115.12}$$

on the way in, with Φ a *modulation potential* in volts that absorbs all the transport physics into one number. Measured values of Φ range from about 400 MV at solar minimum to above 1 GV at solar maximum.

Derivation of the force-field intensity. Liouville's theorem, applied along the characteristics of the transport equation, states that the phase-space density is unchanged, so $f_{\oplus}(p) = f_{\mathrm{LIS}}(p_{\mathrm{LIS}})$ where the subscript denotes the local interstellar value and the two momenta are connected by Equation (115.12). The differential intensity per unit kinetic energy T is related to f by the phase-space volume element: the number of particles per unit volume with momentum in d^3p is $f\,\mathrm{d}^3p = 4\pi p^2 f\,\mathrm{d}p$, and $J(T) := p^2 f$ up to constants, since $\mathrm{d}T = v\,\mathrm{d}p$. Hence

$$J_{\oplus}(T) = J_{\mathrm{LIS}}\left(T + Ze\Phi\right) \frac{T\left(T + 2mc^2\right)}{\left(T + Ze\Phi\right)\left(T + Ze\Phi + 2mc^2\right)}\,, \tag{115.13}$$

the ratio being p^2/p_{LIS}^2 written out with the relativistic relation $p^2c^2 = T\left(T + 2mc^2\right)$ of Section 40.2. For a proton of $T = 100\,\mathrm{MeV}$ and $\Phi = 500\,\mathrm{MV}$ the factor is severe — the observed intensity is a small fraction of the interstellar one — while for $T \gg Ze\Phi$ it tends to unity. This is why the spectrum below a few GeV per nucleon measured at Earth is *not* a source spectrum and must not be read as one, and why the acceleration arguments of Section 115.5 are tested above rather than below that energy. □

Remark 115.6 (Where the heliospheric physics belongs). The magnetohydrodynamics that produces the spiral field and the turbulent scattering that produces the diffusion coefficient belong to Chapter 66, whose headings are reserved and which does not yet carry either; nothing in the present chapter derives them either, and no statement below rests on them. What is used here is only the empirical fact that the modulation exists, that it is measured by neutron monitors as a function of time, and that it is negligible above about 10 GeV — which is where every spectral statement below begins.

115.2 Cosmic rays as the first particle accelerator

115.2.1 Track detectors

Everything in this section is a statement about a photograph. The instruments that made the photographs are two, and each measures a different pair of quantities.

Wilson's expansion cloud chamber [394] makes a track visible by supersaturating a vapour and letting it condense on the ions left along a charged particle's path. Placed between the poles of a magnet, it measures momentum and sign at once: the projected track is an arc of a circle, and by Equation (115.8) its radius of curvature ρ gives

$$p_\perp = ZeB\rho\,, \tag{115.14}$$

in SI a momentum in kg m/s when B is in tesla and ρ in metres. The product $B\rho$, an SI quantity of dimension T m, is the historical observable and is quoted in the papers of the period in gauss-centimetres. One gauss-centimetre is $1\times10^{-4}\,\mathrm{T}\times1\times10^{-2}\,\mathrm{m}=1\times10^{-6}\,\mathrm{T\,m}$, so 10^5 of them make 0.1 T m; Anderson's 2.1×10^5 gauss-centimetres below is 0.21 T m. The sense of the curvature gives the sign of Ze once the direction of travel is known, and fixing that direction is the hard part of every one of the discoveries below.

The second instrument is the nuclear emulsion: a thick photographic layer, exposed for weeks at mountain altitude and developed, in which the track is a chain of silver grains. It measures ionisation and range rather than momentum. The grain density along the track follows the specific ionisation, which for a fast particle behaves as

$$-\frac{\mathrm{d}E}{\mathrm{d}x}\propto\frac{Z^2}{\beta^2}\,,\qquad \beta=\frac{v}{c}\,, \tag{115.15}$$

so a heavily ionising track is either slow or multiply charged, and the two cases are separated by following the track to its end: the *range* of a particle that stops in the emulsion is a single-valued function of its initial energy for each species, and the combination of range and grain density therefore determines mass and charge. The derivation of Equation (115.15) from the electromagnetic interaction of a charge with atomic electrons belongs to Chapter 65; it is used here as an experimental calibration curve, which is how the emulsion groups used it.

Remark 115.7 (What these instruments can and cannot claim). A single cloud-chamber photograph is one event with one measurement uncertainty, and the uncertainty on a curvature is not small: the fractional momentum error from a sagitta measurement grows linearly with momentum, since a stiffer track is a straighter one. Every discovery in this section was therefore contested at the time and settled by accumulating events, not by the first photograph. Where a number below carries an uncertainty it is the one the discoverers quoted; where it does not, the modern value from [2427] is given alongside so that the reader can see how good the original measurement was.

115.2.2 The positron

Anderson operated a vertical cloud chamber in a field of $B=1.5\,\mathrm{T}$, divided across its middle by a lead plate 6 mm thick, and photographed a track of positive curvature crossing the plate [713]. The plate is what makes the photograph an argument rather than an ambiguity. A particle loses energy in crossing it, so it is more sharply curved after the crossing than before; the observed curvatures were $B\rho=0.21\,\mathrm{T\,m}$ on one side and $0.077\,\mathrm{T\,m}$ on the other, which fixes the direction of travel as being from the side of smaller curvature to the side of larger — upward, in this event. With the direction known, the sense of the curvature gives the sign, and the sign is positive.

Phenomenon 115.8 (A positive particle of electronic mass)**.** Cosmic-ray tracks exist that are positively charged and whose specific ionisation is that of a singly charged particle moving at nearly the speed of light, at a momentum at which a proton would be slow and would ionise hundreds of times more heavily [713]. Their mass is therefore of the order of the electron mass and not of the proton mass. The particle is the positron, the antiparticle of the electron.

Rests on Equations (115.14) *and* (115.15)*.*

Derivation. *Derives Phenomenon 115.8.*

Take the segment with $B\rho = 0.21\,\mathrm{T\,m}$. By Equation (115.14) with $Z = 1$,

$$p = eB\rho = 1.602\,176\,634 \times 10^{-19}\,\mathrm{C} \times 0.21\,\mathrm{T\,m} = 3.36 \times 10^{-20}\,\mathrm{kg\,m/s}\,,$$

so that

$$pc = 1.01 \times 10^{-11}\,\mathrm{J} = 63\,\mathrm{MeV}\,.$$

Now test the two candidate species against the observed grain density, which was that of a minimum-ionising particle, i.e. one with $\beta \approx 1$ in Equation (115.15).

If the particle is a proton, $mc^2 = 938.272\,\mathrm{MeV}$ [2427], then

$$E = \sqrt{(pc)^2 + (mc^2)^2} = 940.4\,\mathrm{MeV}\,, \qquad \beta = \frac{pc}{E} = \frac{63}{940.4} = 0.067\,,$$

and Equation (115.15) then demands an ionisation $1/\beta^2 = 223$ times the minimum. A track ionising two hundred times more heavily than the observed one is not a subtle discrepancy; it is a different photograph. If instead the particle has the electron mass, $mc^2 = 0.511\,\mathrm{MeV}$, then $\beta = 1 - 3.3 \times 10^{-5}$ and the track is minimum ionising, which is what was seen. Anderson's stated bound from the ionisation was a mass below twenty electron masses; the modern value is exactly one, since the positron and electron masses are equal by the CPT theorem of Chapter 105.

A second, independent, check comes from the range. The track was followed for several centimetres of chamber gas. A proton of $pc = 63\,\mathrm{MeV}$ has kinetic energy $T = E - mc^2 = 2.1\,\mathrm{MeV}$ and, by Equation (115.15), stops within a few centimetres of air; a 63 MeV positron crosses the whole chamber and the lead plate as well. □

The full apparatus, procedure and uncertainty budget of the discovery, together with the Blackett–Occhialini confirmation and the first pair-production photographs, belong to Chapter 95, whose headings for them are reserved and carry no content yet; the field-theoretic account of the antiparticle is that of Chapter 100. What matters here is that the first antiparticle ever seen was seen in the cosmic radiation, four years after Dirac's equation and with no accelerator involved.

115.2.3 The muon

By the mid-1930s the cosmic radiation at sea level was known to have two components, distinguished by how much absorber they survive: a *soft* component absorbed by some 10 cm of lead, and a *hard* or penetrating component that crosses a metre of it. Neddermeyer and Anderson showed that the penetrating particles cannot be electrons [797], and the argument is a comparison of two energy-loss laws.

Derivation: the penetrating particles are not electrons. An electron of energy E crossing matter loses energy chiefly by bremsstrahlung, at a rate proportional to E itself,

$$-\frac{\mathrm{d}E}{\mathrm{d}X} = \frac{E}{X_0}\,, \qquad E(X) = E_0 \mathrm{e}^{-X/X_0}\,, \tag{115.16}$$

where X is the traversed depth of Equation (115.2) and X_0 is the *radiation length* of the material, an SI areal mass density: $X_0 = 365\,\mathrm{kg/m^2} = 36.5\,\mathrm{g/cm^2}$ for air and $63.7\,\mathrm{kg/m^2} = 6.37\,\mathrm{g/cm^2}$ for lead. The exponential is the content of the Bethe–Heitler cross-section [732], whose derivation belongs to the bremsstrahlung section of Chapter 65, where the heading is reserved and the calculation is not yet written; what matters here is that the loss is proportional to E and is therefore catastrophic at high energy.

Derivation pending #204

The Bethe–Heitler bremsstrahlung cross section, the resulting exponential energy-loss law, and the radiation length of air and of lead used throughout this chapter. Belongs in Appendix A or in the bremsstrahlung section of the radiation and scattering chapter, which currently reserves the heading

A metre of lead has depth $X = 11.35 \times 10^3\,\mathrm{kg/m^3} \times 1\,\mathrm{m} = 1.135 \times 10^4\,\mathrm{kg/m^2}$, i.e. $X/X_0 = 178$ radiation lengths. An electron crossing it would emerge with e^{-178} of its energy: nothing survives, at any incident energy. The observed penetrating particles cross such absorbers losing energy at the far slower, nearly energy-independent rate characteristic of ionisation alone, $-\mathrm{d}E/\mathrm{d}X \approx 0.2\,\mathrm{MeV\,m^2/kg}$, i.e. about 2 MeV per $\mathrm{g/cm^2}$. Since the radiative rate carries a factor $1/m^2$ from the recoil of the radiating particle against the nucleus, a particle heavier than the electron radiates less, and *much* less: the penetrating component is a particle of mass well above m_e and below m_p, because a proton of the observed momenta would ionise far too heavily by Equation (115.15). □

Street and Stevenson closed the argument by measuring the mass directly, from the curvature and the range of a particle stopping in a cloud chamber, obtaining about 130 electron masses [801]. The modern value is $m_\mu c^2 = (105.658\,375\,5 \pm 0.000\,002\,3)\,\mathrm{MeV}$, that is $m_\mu = (1.883\,531\,627 \pm 0.000\,000\,042) \times 10^{-28}\,\mathrm{kg}$ and $206.77\,m_e$ [2427, 2431], so the first measurement was low by a factor 1.6 — an honest reflection of what a single stopping track can do.

What the muon is *not* took another decade. Yukawa's meson, the quantum of the nuclear force of Section 107.2.1, has a mass in the right range, and the identification was universally assumed. It is wrong, and Conversi, Pancini and Piccioni proved it wrong by stopping negative mesotrons in carbon and finding that they decay with their ordinary lifetime instead of being captured by the nucleus [899].

Derivation: the mesotron is not strongly interacting. A negative particle stopped in matter cascades into a Bohr orbit about a nucleus. The mesic Bohr radius scales as the inverse mass and the inverse nuclear charge,

$$a_\mu = \frac{a_0}{Z}\frac{m_e}{m_\mu} = \frac{5.29 \times 10^{-11}\,\mathrm{m}}{6} \times \frac{1}{206.8} = 4.3 \times 10^{-14}\,\mathrm{m}$$

for carbon, $Z = 6$. The nuclear radius is $R = r_0 A^{1/3} = 1.2 \times 10^{-15}\,\mathrm{m} \times 12^{1/3} = 2.7 \times 10^{-15}\,\mathrm{m}$, so the probability of finding the particle inside the nucleus is of order

$$P \sim \left(\frac{R}{a_\mu}\right)^3 = (0.064)^3 = 2.6 \times 10^{-4}\,.$$

If the particle interacts strongly, the rate at which it is absorbed once inside is of order the only rate the strong interaction has, $c/R = 1.1 \times 10^{23}/\mathrm{s}$. The capture rate is then

$$\Gamma_{\mathrm{cap}} \sim P\,\frac{c}{R} \approx 3 \times 10^{19}/\mathrm{s}\,,$$

a capture time of order $3 \times 10^{-20}\,\mathrm{s}$, against the observed decay rate $1/\tau_\mu = 4.6 \times 10^5/\mathrm{s}$. The two differ by fourteen orders of magnitude: a strongly interacting particle stopped in carbon would never be seen to decay. It was seen to decay. Hence the mesotron does not interact strongly and is not Yukawa's meson — the same conclusion reached, with the nuclear-physics apparatus, in Phenomenon 107.18. □

The muon's other role is as a clock. Rossi and Hall measured the momentum dependence of the surviving fraction between altitudes [863], Frisch and Smith compared the flux on Mount Washington with the flux at sea level [1253], and the CERN storage ring measured the dilated lifetime at $\gamma = 29.3$ to a few parts in a thousand [1592]. All three are treated as experiments in Chapter 41 (Sections 41.2 to 41.4); the cosmic-ray side of the argument is Section 115.8.3 below.

115.2.4 The pion

The particle Yukawa predicted was found in 1947, in emulsion stacks exposed at the Pic du Midi and at Chacaltaya in the Andes, by Lattes, Occhialini and Powell [902]. The signature is a track that ends and gives birth to a second track: a heavily ionising particle comes to rest, and from its end point a second particle sets off, itself coming to rest after a short and — this is the point — always the same range, about 600 µm of emulsion. A third, much lighter track then leaves the second end point. The chain is $\pi \to \mu \to e$.

Phenomenon 115.9 (A constant secondary range means a two-body decay)**.** In emulsion the secondary of a stopping cosmic-ray meson has a fixed range, corresponding to a kinetic energy of about 4 MeV, whatever the history of the primary [902]. The parent therefore decays into exactly two bodies, one of which is unseen, and the parent and the charged secondary are two *different* particles of different mass.

Rests on Equation (40.11).

Derivation. *Derives Phenomenon 115.9.*

Let a particle of mass m_π decay at rest into a charged particle of mass m_μ and n unseen particles. Conservation of four-momentum, Equation (40.11), gives for the charged secondary

$$m_\pi c^2 = E_\mu + E_X\,, \qquad \boldsymbol{p}_\mu + \boldsymbol{p}_X = \boldsymbol{0}\,,$$

with X the unseen system. Squaring the four-momentum of X,

$$m_X^2 c^4 = \left(m_\pi c^2 - E_\mu\right)^2 - p_\mu^2 c^2\,,$$

so that

$$E_\mu = \frac{\left(m_\pi^2 + m_\mu^2 - m_X^2\right)c^2}{2m_\pi}\,, \qquad p_\mu c = \frac{c^2}{2m_\pi}\left[\left(m_\pi^2 - (m_\mu + m_X)^2\right)\left(m_\pi^2 - (m_\mu - m_X)^2\right)\right]^{1/2}. \tag{115.17}$$

If $n = 1$ then m_X is a fixed mass and E_μ is a single number: the secondary is monoenergetic and its range is constant. If $n \geq 2$ then m_X ranges continuously from $\sum m_i$ upwards, E_μ is spread over a continuum, and no fixed range can occur. A constant range therefore establishes a two-body decay — which is the same logic that identifies the beta spectrum as three-body in Section 101.1.1, read in the opposite direction.

Putting $m_X = 0$ and the modern masses $m_{\pi^+}c^2 = 139.570\,39\,\mathrm{MeV}$, $m_\mu c^2 = 105.658\,375\,5\,\mathrm{MeV}$ [2427] into Equation (115.17),

$$p_\mu c = \frac{\left(m_\pi^2 - m_\mu^2\right)c^4}{2m_\pi c^2} = 29.79\,\mathrm{MeV}\,, \qquad T_\mu = \sqrt{(p_\mu c)^2 + (m_\mu c^2)^2} - m_\mu c^2 = 4.12\,\mathrm{MeV}\,,$$

in SI $T_\mu = 6.60 \times 10^{-13}\,\mathrm{J}$, and the range of a 4.12 MeV muon in emulsion is indeed about 600 µm. The unseen partner is the muon neutrino, whose own story is Chapter 103. □

The pion is the strongly interacting particle the nuclear force demands, and its role as the lightest quantum of that force is developed in Section 107.2.2; that it is light for a reason — being the pseudo-Goldstone boson of the broken chiral symmetry of Section 102.8 — is the QCD side of the same fact. For the present chapter its importance is different and entirely practical: pion production is the dominant channel by which a primary nucleus converts its energy into a shower, and Section 115.4.2 is built on the branching ratios of pion decay and nothing else.

115.2.5 Strange particles

Rochester and Butler, operating a counter-controlled cloud chamber, photographed tracks that fork in mid-flight with no incoming track at the vertex [903]. A neutral particle had decayed into two charged ones. The mass of the parent follows from the two secondaries by the invariant-mass construction of Equation (40.12),

$$M^2c^4 = (E_1 + E_2)^2 - (\boldsymbol{p}_1 + \boldsymbol{p}_2)^2\, c^2 \,, \tag{115.18}$$

each E_i and $\boldsymbol{p}_i$ coming from the curvature and ionisation of one track, and it came out near $5 \times 10^8\,\mathrm{eV}/c^2$ — about half the proton mass, and a particle nobody had ordered. The modern value for the neutral kaon is $m_{K^0}c^2 = (497.611 \pm 0.013)\,\mathrm{MeV}$ [2427].

What made these particles "strange" is a rate, not a mass.

Phenomenon 115.10 (Copious production, slow decay)**.** The V particles are produced with cross sections characteristic of the strong interaction, and yet they live of order $1 \times 10^{-10}\,\mathrm{s}$ before decaying into ordinary hadrons [903] — some thirteen orders of magnitude longer than the strong-interaction time scale. A quantum number conserved in production and violated in decay is therefore required.

Derivation. *Derives Phenomenon 115.10.*

The only time scale the strong interaction possesses is the light transit time of a hadron. With a hadronic radius $r_h \approx 1 \times 10^{-15}\,\mathrm{m}$,

$$\tau_{\mathrm{strong}} \sim \frac{r_h}{c} = \frac{1 \times 10^{-15}\,\mathrm{m}}{2.998 \times 10^8\,\mathrm{m/s}} = 3 \times 10^{-24}\,\mathrm{s}\,,$$

and any decay proceeding through the strong interaction into kinematically allowed hadronic channels must take a time of that order. The observed lifetime is $1 \times 10^{-10}\,\mathrm{s}$, larger by 3×10^{13}. Since the production is strong — the cross sections are millibarns, not microbarns — the initial and final states are connected by the strong interaction in one direction and not in the other, which is impossible for a single interaction. The resolution is an additive quantum number S, conserved by the strong interaction and violated by the weak one: strange particles are then produced only in pairs of opposite S, so that production conserves S, and each must then wait for a weak decay, whose rate is smaller by the square of the Fermi constant. The classification of the resulting spectrum by SU(3) is Section 102.1.1, and the identification of S as a quark flavour is Section 102.1.3; the weak decay itself is quark mixing, Chapter 101. □

After 1953 this physics moved to accelerators, where a controlled beam and a known incident energy do in an afternoon what a mountain laboratory did in a year. What cosmic rays retained is what no accelerator can supply: energies above the beam energy of any machine, which is Section 115.8, and an isotropic natural flux over the whole sky, which is what makes them astronomy.

115.3 The primary spectrum and composition

115.3.1 The all-particle spectrum

The primary observable is the *differential intensity* $J(E)$: the number of particles per unit area, per unit time, per unit solid angle and per unit energy crossing a surface normal to their arrival direction. Its SI unit is $1/(\mathrm{m^2\,s\,sr\,J})$; the literature quotes it per GeV, and $1\,\mathrm{GeV} = 1.602\,176\,634 \times 10^{-10}\,\mathrm{J}$ converts between them.

Over the range in which direct measurement above the atmosphere is possible, the primary nucleon intensity is described by a single power law [2427],

$$J_N(E) \approx 1.8 \times 10^{4} \left(\frac{E}{1\,\mathrm{GeV}}\right)^{-2.7} 1/(\mathrm{m^2\,s\,sr\,GeV}) = 1.12 \times 10^{14} \left(\frac{E}{1\,\mathrm{GeV}}\right)^{-2.7} 1/(\mathrm{m^2\,s\,sr\,J})\,, \tag{115.19}$$

valid from a few GeV — below which the solar modulation of Equation (115.13) takes over — up to about $1 \times 10^{14}\,\mathrm{eV}$. Above that the flux is too small for a satellite and is measured through air showers, and the power law continues, with two changes of index, to the highest energies observed.

Phenomenon 115.11 (A power-law spectrum over eleven decades)**.** The differential intensity of primary cosmic rays is, to a good approximation, a featureless power law

$$\frac{\mathrm{d}N}{\mathrm{d}E} \propto E^{-\gamma}\,, \qquad \gamma \approx 2.7\,, \tag{115.20}$$

from about 1 GeV to above $1 \times 10^{20}\,\mathrm{eV}$, across which the flux falls by some thirty-two orders of magnitude. Three integral rates fix the scale: about one particle per square centimetre per second per steradian above 1 GeV, about four per square metre per second per steradian above 100 GeV, and of order one per square kilometre per year above $1 \times 10^{19}\,\mathrm{eV}$ [2271, 2427]. No terrestrial accelerator produces anything remotely like this range.

Rests on Equation (115.19).

Derivation of the integral rates. *Derives Phenomenon 115.11.*

For a power law $J(E) = J_0\,(E/E_0)^{-\gamma}$ with $\gamma > 1$ the integral intensity above E converges and is

$$J(>E) = \int_E^{\infty} J\,(E')\,\mathrm{d}E' = \frac{J_0 E_0}{\gamma - 1}\left(\frac{E}{E_0}\right)^{-(\gamma-1)}\,. \tag{115.21}$$

Inserting Equation (115.19) with $E_0 = 1\,\mathrm{GeV}$, $J_0 = 1.8 \times 10^{4}/(\mathrm{m^2\,s\,sr\,GeV})$ and $\gamma = 2.7$,

$$J\,(>1\,\mathrm{GeV}) = \frac{1.8 \times 10^{4}}{1.7} = 1.06 \times 10^{4}/(\mathrm{m^2\,s\,sr}) = 1.06/(\mathrm{cm^2\,s\,sr})\,,$$

which is the first anchor, and

$$J\,(>100\,\mathrm{GeV}) = 1.06 \times 10^{4} \times 100^{-1.7} = 4.2/(\mathrm{m^2\,s\,sr})\,,$$

which is the second. For the third the index must be changed at the knee of Section 115.3.2. Matching a second power law $\gamma_2 = 3.1$ continuously at $E_k = 3 \times 10^{15}\,\mathrm{eV}$ gives, for $E > E_k$,

$$J(>E) = \frac{J_0 E_0}{\gamma_2 - 1}\left(\frac{E_k}{E_0}\right)^{\gamma_2-\gamma_1}\left(\frac{E}{E_0}\right)^{-(\gamma_2-1)}\,,$$

and at $E = 1 \times 10^{19}\,\mathrm{eV} = 1 \times 10^{10}\,\mathrm{GeV}$ this evaluates to $3.3 \times 10^{-15}/(\mathrm{m^2\,s\,sr})$. Multiplying by $1 \times 10^{6}\,\mathrm{m^2/km^2}$ and by $3.156 \times 10^{7}\,\mathrm{s/yr}$,

$$J\left(>1 \times 10^{19}\,\mathrm{eV}\right) \approx 0.1/(\mathrm{km^2\,yr\,sr})\,,$$

so that an array accepting some π steradians records of order one event per square kilometre per year — the third anchor, and the reason the observatories of Section 115.4.3 are the size of a small country. The total dynamic range follows in the same way: $J\left(1\times10^{20}\,\mathrm{eV}\right)/J\left(1\,\mathrm{GeV}\right)=3.1\times10^{-32}$, thirty-one and a half orders of magnitude. □

Remark 115.12 (Why the spectrum is plotted multiplied by E^3). A quantity falling over thirty-two decades cannot be examined on a linear scale, and even on a logarithmic one a change of index from 2.7 to 3.1 is invisible to the eye. Every published spectrum is therefore plotted as $E^3J(E)$, or as $E^{2.7}J(E)$, which turns a pure power law of that index into a horizontal line and turns the knee and ankle into visible kinks. The multiplication is cosmetic and carries a trap: the vertical scale then has SI units $\mathrm{J}^2/(\mathrm{m}^2\,\mathrm{s}\,\mathrm{sr})$, and a systematic shift of the energy scale by a factor $1+\epsilon$ moves every plotted point diagonally — right by ϵ and up by 2ϵ, since the abscissa carries one power of the energy and the ordinate the remaining two. What that displacement does to the *curve* is what matters, and at a fixed abscissa the plotted ordinate is then in error by $(\gamma-1)\,\epsilon$: the common factor E^3 cancels between the true and the reconstructed curve, leaving exactly the error Equation (115.40) derives for the integral intensity. Section 115.4.6 treats that effect quantitatively, because it is the dominant uncertainty in every spectral measurement above $1\times10^{18}\,\mathrm{eV}$.

A featureless power law over eleven decades is the strongest single constraint on the acceleration mechanism. No process with a characteristic energy can produce one: a thermal source produces a Maxwellian with a temperature, a resonance produces a line, and a mechanism with a preferred scale produces a break at that scale. What produces a power law is a process in which the fractional energy gain per step and the escape probability per step are both independent of energy, and Section 115.5.3 shows that a strong shock is exactly such a process.

115.3.2 The knee

Phenomenon 115.13 (The knee)**.** Near $3\times10^{15}\,\mathrm{eV}$ the spectrum of Equation (115.20) steepens, the index changing from about 2.7 to about 3.1. The break was first seen in the shower-size spectrum [1171], and was later resolved into successive rigidity-ordered cutoffs of the individual elemental groups, the light component breaking first [2070]. Which of the two standing readings — a limit of the Galactic accelerators, or a change in Galactic confinement — is correct is not settled by the data.

Rests on Equation (115.20).

Kulikov and Khristiansen found the break not in the energy spectrum but in the *shower-size* spectrum: the number of showers containing more than N particles at the ground falls as a power of N, and the power changes near $N\approx1\times10^6$ [1171]. Since $N\propto E$ for an electromagnetic cascade (Equation (115.29)), a break in N is a break in E, and the calibration puts it at $3\times10^{15}\,\mathrm{eV}=4.8\times10^{-4}\,\mathrm{J}$.

Derivation: any magnetic origin of the knee is ordered by rigidity. *Derives Phenomenon 115.13.*

Suppose the break is produced by a magnetic process, either at the accelerator (a maximum energy the source can reach) or in transport (a maximum energy the Galaxy can retain). Both are governed by the equation of motion of a charge in a magnetic field, and that equation, written for the momentum,

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}=Ze\,\boldsymbol{v}\times\boldsymbol{B}\,,$$

depends on the species only through Ze. Dividing by Ze and using Equation (115.1),

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\boldsymbol{p}c}{Ze}\right)=c\,\boldsymbol{v}\times\boldsymbol{B}\quad\Longrightarrow\quad\frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}t}=c\,\boldsymbol{v}\times\boldsymbol{B}\,,$$

so two particles of the same rigidity and the same velocity follow identical trajectories in any magnetic field whatever, regardless of their charge and mass. Any limit imposed by magnetic fields is therefore a limit on rigidity, $R \le R_{\max}$, and for an ultrarelativistic nucleus of charge Z this is a limit on energy

$$E_{\max}(Z) = ZeR_{\max} = Z\,E_{\max}(1)\,. \tag{115.22}$$

The prediction is sharp: the knee of the proton component sits at E_k, that of helium at $2E_k$, and that of iron at $26E_k$, so that the all-particle spectrum steepens gradually as one species after another drops out, and the composition becomes progressively heavier across the knee region. With $E_k = 3 \times 10^{15}\,\mathrm{eV}$ the iron knee is predicted at $7.8 \times 10^{16}\,\mathrm{eV}$. This is what KASCADE found by unfolding the measured shower-size and muon-number distributions into elemental groups: the light component breaks first, the heavy components later, in the rigidity-ordered sequence Equation (115.22) demands [2070]. □

Remark 115.14 (What the rigidity ordering does not decide). The derivation above is the reason the knee is *not* yet understood. Both candidate explanations are magnetic and both therefore predict the same rigidity ordering, so the observation that confirmed the ordering cannot discriminate between them. An acceleration limit predicts that the sources stop supplying particles above $ZeR_{\max}$; a confinement limit predicts that the Galaxy stops retaining them. Distinguishing the two requires an observable that is not a spectrum: the anisotropy expected from a leaking population, or a direct measurement of a source spectrum in gamma rays extending past the knee — which is what the LHAASO sources of Section 115.7.1 begin to supply. It is reported here as open because it is open.

115.3.3 The ankle and the highest energies

Linsley, operating the Volcano Ranch array in New Mexico, recorded in 1962 a shower whose reconstructed primary energy was about $1 \times 10^{20}\,\mathrm{eV} = 16\,\mathrm{J}$ — the kinetic energy of a well-struck tennis ball, carried by a single nucleus [1261]. It remains, sixty years later, within a factor of a few of the highest energy ever measured.

Between the knee and that regime the spectrum steepens once more and then *flattens* again, near $4 \times 10^{18}\,\mathrm{eV}$, the index falling from about 3.3 back to about 2.6. The feature is called the ankle and was measured with high statistics at the Pierre Auger Observatory [2116]. The standard reading is that it marks the crossing of two populations: a Galactic component dying out and an extragalactic component taking over. The argument that the highest-energy particles cannot be Galactic is a comparison of two lengths and is worth doing in SI.

Phenomenon 115.15 (The highest-energy primaries are not confined by the Galaxy)**.** A proton of $1 \times 10^{19}\,\mathrm{eV}$ in the Galactic magnetic field has a gyroradius of several kiloparsecs, larger than the thickness of the Galactic disc by more than an order of magnitude and comparable to the distance from the Sun to the Galactic centre. Such particles cannot be stored in the Galaxy, and a Galactic source population cannot supply an isotropic flux of them.

Rests on Equation (115.8).

Derivation. *Derives Phenomenon 115.15.*

The Galactic magnetic field measured through Faraday rotation and synchrotron polarisation has a regular component of order $B = 3 \times 10^{-10}\,\mathrm{T}$, that is 0.3 nT or three microgauss in the older literature. An ultrarelativistic proton of energy E has $p = E/c$, so Equation (115.8) gives

$$r_L = \frac{p}{ZeB} = \frac{E}{ZecB}\,. \tag{115.23}$$

For $E = 1 \times 10^{19}\,\mathrm{eV} = 1.602\,\mathrm{J}$ and $Z = 1$,

$$r_L = \frac{1.602\,\mathrm{J}}{1.602\,176\,634 \times 10^{-19}\,\mathrm{C} \times 2.998 \times 10^{8}\,\mathrm{m/s} \times 3 \times 10^{-10}\,\mathrm{T}} = 1.11 \times 10^{20}\,\mathrm{m} = 3.6\,\mathrm{kpc}\,,$$

using $1\,\mathrm{pc} = 3.0857 \times 10^{16}\,\mathrm{m}$. The half-thickness of the Galactic disc is about 0.15 kpc and the distance to the Galactic centre about 8 kpc, so the orbit is twenty-four times the disc thickness and about half the size of the whole Galaxy. At $1 \times 10^{20}\,\mathrm{eV}$ the same formula gives $r_L = 36\,\mathrm{kpc}$, larger than the Galaxy itself: such a particle crosses the disc essentially in a straight line. A population confined by the Galactic field therefore cannot extend to these energies, and conversely a Galactic *source* of them would produce a strong anisotropy towards the Galactic plane, which Section 115.6.3 shows is not observed. The particles are extragalactic. □

115.3.4 Elemental and leptonic composition

Above the modulation region and below the knee the primaries are measured species by species with magnetic spectrometers and calorimeters above the atmosphere. By number they are roughly 90 % protons, 9 % helium and 1 % heavier nuclei, with an electron component of order 1 % of the protons at the same energy. The elemental pattern tracks solar-system abundances closely — with two conspicuous exceptions, and the exceptions are the measurement.

Phenomenon 115.16 (Light secondary nuclei are enormously overabundant)**.** Lithium, beryllium and boron are present in the cosmic radiation at a level of order 25 % of the carbon–nitrogen–oxygen group, whereas in solar-system material their abundance relative to that group is of order 10^{-6}. The same holds, less dramatically, for scandium through manganese relative to iron. These nuclei are not made in stars in such quantities; they are made in flight, by the fragmentation of heavier cosmic-ray nuclei on interstellar gas, and their abundance measures the amount of matter traversed.

Rests on Equation (115.2).

Derivation of the traversed grammage. *Derives Phenomenon 115.16.*

Let N_C be the number of carbon nuclei in a beam and N_B the number of boron nuclei produced from them by spallation, and let X be the traversed grammage of Equation (115.2), now measured through interstellar gas rather than air. Over a path element the number of fragmentations is the number of target nuclei per unit area times the cross section,

$$\mathrm{d}N_B = N_C\,\frac{\sigma_{C\to B}}{\bar{m}}\,\mathrm{d}X\,,$$

with $\bar{m}$ the mean target mass, essentially the proton mass $1.673 \times 10^{-27}\,\mathrm{kg}$ for interstellar hydrogen. As long as the secondary fraction stays small this integrates to

$$\frac{N_B}{N_C} \approx \frac{\sigma_{C\to B}}{\bar{m}}\,X\,, \qquad \text{so} \qquad X \approx \frac{\bar{m}}{\sigma_{C\to B}}\,\frac{N_B}{N_C}\,. \tag{115.24}$$

The measured boron-to-carbon ratio at a few GeV per nucleon is $N_B/N_C \approx 0.3$, and the fragmentation cross section is $\sigma_{C\to B} \approx 6 \times 10^{-30}\,\mathrm{m}^2$, which is 60 millibarn in the customary non-SI unit (one barn is $1 \times 10^{-28}\,\mathrm{m}^2$ exactly), whence

$$X \approx \frac{1.673 \times 10^{-27}\,\mathrm{kg}}{6 \times 10^{-30}\,\mathrm{m}^2} \times 0.3 = 84\,\mathrm{kg/m^2} = 8.4\,\mathrm{g/cm^2}\,.$$

That is the whole of the matter a cosmic ray meets between its source and the Earth. It is a small quantity — less than one percent of the atmosphere of Equation (115.3) — and it is

nevertheless enough to make the secondary nuclei, because their solar-system abundance is essentially zero. The corresponding path length in an interstellar medium of number density $n \approx 3 \times 10^5/\mathrm{m}^3$ is $L = X/(n\bar{m}) = 1.7 \times 10^{23}\,\mathrm{m}$, some 5 Mpc of path — folded, of course, into a volume a few kiloparsecs across, since the trajectory is a random walk and not a straight line. Since X is by Equations (115.2) and (115.24) the matter traversed *along* the path, $L/c = 5.6 \times 10^{14}\,\mathrm{s}$, that is $1.8 \times 10^7\,\mathrm{yr}$, is already the residence time; it is recomputed directly as Equation (115.53) and confirmed there by the independent radioactive clock of Section 115.5.5. □

The ratio is not constant: AMS-02 measured the boron-to-carbon ratio above a rigidity of 65 GV and found it to follow a single power law in rigidity,

$$\frac{N_B}{N_C} \propto R^{-\delta}\,, \qquad \delta = 0.333 \pm 0.014_{\mathrm{fit}} \pm 0.005_{\mathrm{syst}}\,, \tag{115.25}$$

with no break [2266]. By Equation (115.24) this says the grammage — and hence the residence time — falls as $R^{-1/3}$: faster particles leave sooner. That single exponent is the input to every propagation calculation in Section 115.5.5, and the value 1/3 is what Kolmogorov turbulence in the Galactic magnetic field would give.

The leptonic composition carries an anomaly that is still unresolved. Secondary positrons are made in the same collisions that make boron — $p + \mathrm{gas} \to \pi^+ + \ldots$, $\pi^+ \to \mu^+ \to e^+$ — so their intensity is that of a secondary and must fall relative to the primary electrons as $R^{-\delta}$, exactly as the boron does. The positron *fraction* $e^+/(e^+ + e^-)$ must therefore decrease monotonically with energy. PAMELA found it *rising* above about 10 GeV [2139], and AMS-02 measured the rise precisely over the band 0.5 GeV to 350 GeV, finding a fraction that increases with energy while its slope decreases [2207]. Later AMS-02 data extend the measurement to 1 TeV and show the rise flattening and turning over [2327].

Remark 115.17 (The positron excess is unexplained). A rising positron fraction requires a source of primary positrons. Three classes of candidate are discussed: nearby pulsar wind nebulae, which produce electron–positron pairs and are known to exist; a nearby supernova remnant with an unusually hard secondary production; and the annihilation of a dark-matter particle, which is the subject of Chapter 134. The measurement does not distinguish them. The spectral shape alone cannot: a turnover is produced by a nearby source of finite age and by an annihilation cutoff equally well. What would distinguish them is a measured anisotropy towards a nearby pulsar, and none has been detected at the sensitivity so far reached. The honest statement is that a primary positron source exists and is unidentified; asserting a dark-matter interpretation would be asserting a mechanism that has not been observed, which Chapter 134 declines to do and so does this chapter.

115.4 Extensive air showers

115.4.1 Auger's coincidence experiment

Above about $1 \times 10^{14}\,\mathrm{eV}$ the primary flux is too small for any detector that can be flown, and the atmosphere must be used as the detector. That this is possible at all was discovered by Auger and collaborators, who placed Geiger counters at increasing separations on the roof of a laboratory, at the Jungfraujoch and at the Pic du Midi, and recorded coincidences [827]. Coincidences persisted at separations of 300 m.

Phenomenon 115.18 (A single primary makes a wide particle front)**.** Counters separated by distances up to 300 m fire simultaneously far more often than uncorrelated counting can produce, and the coincidence rate falls smoothly with separation instead of being independent of it [827]. A single primary particle therefore produces a coherent front of

secondaries tens of thousands strong spread over a large area at the ground, and the primary energies required are of order $1 \times 10^{15}\,\mathrm{eV}$ — six orders of magnitude above anything then measured.

Derivation. *Derives Phenomenon 115.18.*

Two counters with singles rates r_1 and r_2 and a coincidence resolving time τ register accidental coincidences at the rate

$$r_{\mathrm{ch}} = 2\tau r_1 r_2\,, \tag{115.26}$$

since counter 2 must fire within $\pm\tau$ of counter 1. For counters of area $100\,\mathrm{cm}^2$ at sea level, where the vertical muon intensity is about $70/(\mathrm{m}^2\,\mathrm{s\,sr})$, the singles rate is $r \approx 1.7/\mathrm{s}$; with the $\tau \approx 10\,\mu\mathrm{s}$ of the period, Equation (115.26) gives $r_{\mathrm{ch}} = 5.8 \times 10^{-5}/\mathrm{s}$, about five accidental coincidences a day.

The decisive point is not the size of this number but its *independence of the separation*: Equation (115.26) contains no distance at all. The measured coincidence rate, by contrast, falls steeply and smoothly as the counters are drawn apart, and remains above the accidental rate out to 300 m. A distance-dependent excess over a distance-independent background can only be a correlation, and a correlation over 300 m at the speed of light means a single common ancestor.

The energy follows from the particle number. If the front contains N particles at the ground and each was produced by dividing the primary energy down to the point where production stops — the critical energy $E_c \approx 85\,\mathrm{MeV}$ of Section 115.4.2 — then $E_0 \gtrsim N E_c$. Auger's estimate of $N \approx 10^6$ particles gives $E_0 \gtrsim 8.5 \times 10^{13}\,\mathrm{eV}$, and correcting for the fact that the ground lies well past the shower maximum, where most of the particles have already been absorbed, raises this to of order $1 \times 10^{15}\,\mathrm{eV} = 1.6 \times 10^{-4}\,\mathrm{J}$. □

The shower is thus an amplifier: one particle arriving once per square metre per century is converted by the atmosphere into a signal spread over square kilometres, which a sparse array of detectors covering a 10^{-5} fraction of that area can still sample. Every measurement above $1 \times 10^{15}\,\mathrm{eV}$ in this chapter rests on it.

115.4.2 Cascade development

Take the electromagnetic part of the cascade first, since the hadronic part feeds it and cannot be described without it.

A high-energy photon in matter converts to an e^+e^- pair; a high-energy electron radiates a bremsstrahlung photon. Both processes have the same characteristic depth, the radiation length X_0 of Equation (115.16), which for air is

$$X_0 = 365\,\mathrm{kg/m}^2 = 36.5\,\mathrm{g/cm}^2\,, \tag{115.27}$$

so that the vertical atmosphere of Equation (115.3) is $1.0332 \times 10^4\,\mathrm{kg/m}^2 / 365\,\mathrm{kg/m}^2 = 28$ radiation lengths deep — a calorimeter as thick as any built. Below a *critical energy* E_c the electron loses energy faster by ionisation than by radiation and the multiplication stops; for air $E_c \approx 85\,\mathrm{MeV} = 1.36 \times 10^{-11}\,\mathrm{J}$ [2080]. Heitler's model replaces the two processes by one [1043]: every particle splits into two, each carrying half the energy, after a fixed *splitting length* $d = X_0 \ln 2$.

Theorem 115.19 (Heitler's cascade)**.** *In the splitting model the number of particles at depth X and the energy per particle are*

$$N(X) = 2^{X/d}\,, \qquad E(X) = \frac{E_0}{N(X)} = E_0 2^{-X/d}\,, \qquad d = X_0 \ln 2\,, \tag{115.28}$$

the cascade stops at the depth where $E(X) = E_c$, and therefore

$$N_{\max} = \frac{E_0}{E_c}\,, \qquad X_{\max} = X_0 \ln\!\left(\frac{E_0}{E_c}\right). \tag{115.29}$$

Two consequences follow, and both are what is actually measured: the particle number at maximum is linear *in the primary energy, and the depth of maximum grows only* logarithmically *with it, at the rate*

$$D_{10} := \frac{\mathrm{d}X_{\max}}{\mathrm{d}\log_{10} E_0} = X_0 \ln 10 = 84\,\mathrm{g/cm^2} \tag{115.30}$$

per decade of primary energy.

Rests on Equations (115.2) *and* (115.16).

Proof. *Derives Theorem 115.19.*

After n splittings the depth is $X = nd$ and the number of particles is 2^n, which is Equation (115.28) with $n = X/d$. Energy is conserved and shared equally, so $E(X)N(X) = E_0$. Multiplication continues while $E(X) > E_c$ and stops when equality is reached; at that depth $N = E_0/E_c$, which is the first of Equation (115.29), and $2^{X_{\max}/d} = E_0/E_c$ gives

$$X_{\max} = d\,\frac{\ln\left(E_0/E_c\right)}{\ln 2} = X_0 \ln\!\left(\frac{E_0}{E_c}\right),$$

which is the second. Differentiating with respect to $\log_{10} E_0 = \ln E_0 / \ln 10$ gives Equation (115.30). □

The numbers are informative. For $E_0 = 1 \times 10^{15}\,\mathrm{eV}$, Equation (115.29) gives $N_{\max} = 1.2 \times 10^7$ particles and $X_{\max} = 594\,\mathrm{g/cm^2}$; for $E_0 = 1 \times 10^{20}\,\mathrm{eV}$, $N_{\max} = 1.2 \times 10^{12}$ and $X_{\max} = 1014\,\mathrm{g/cm^2}$, which is just about the depth of the whole vertical atmosphere — so the highest-energy showers reach maximum near the ground and a detector on a high plateau sees them close to their peak. That is why the observatories of Section 115.4.3 sit at altitude.

Remark 115.20 (What the toy model gets wrong, and why it is still used). Heitler's model is not a calculation of a cascade; it is a book-keeping device that gets the two scaling laws right and the absolute numbers wrong by a fixed factor. Real cascades do not split in two: an electron radiates a photon of any energy fraction with the Bethe–Heitler spectrum, so the particle number at maximum is smaller than E_0/E_c by a factor of order 10, and the sub-threshold particles are not immediately absorbed. The full treatment is the coupled pair of integro-differential cascade equations, solved by Mellin transform in the depth variable — the technique of Chapter 17, applied to a physically different problem from the one it is introduced for. What survives unchanged from Theorem 115.19 is $N_{\max} \propto E_0$ and $X_{\max} \propto \ln E_0$, and those two are the entire basis of air-shower energy and composition measurement.

A primary nucleus does not start an electromagnetic cascade; it starts a hadronic one, and the electromagnetic component is a by-product. The interaction length follows from the cross section as $\lambda_I = \bar{m}_{\mathrm{air}}/\sigma_{p\text{-air}}$ with $\bar{m}_{\mathrm{air}} = 14.5 \times 1.6605 \times 10^{-27}\,\mathrm{kg} = 2.41 \times 10^{-26}\,\mathrm{kg}$ the mean mass of an air nucleus; the inelastic proton–air cross section grows slowly with energy, and near $1 \times 10^{15}\,\mathrm{eV}$ it is about $4.5 \times 10^{-29}\,\mathrm{m^2}$, that is 450 millibarn in the customary unit of Equation (115.24), so

$$\lambda_I = \frac{2.41 \times 10^{-26}\,\mathrm{kg}}{4.5 \times 10^{-29}\,\mathrm{m^2}} \approx 550\,\mathrm{kg/m^2} = 55\,\mathrm{g/cm^2}\,, \tag{115.31}$$

so that by Equation (115.3) the atmosphere is about nineteen interaction lengths deep and a primary interacts near the top with certainty [2271]. Each interaction produces of order

N_{ch} charged pions and $N_{\mathrm{ch}}/2$ neutral ones, the isospin content of Chapter 102 making the three charge states roughly equally populated. The neutral pions decay to two photons in $8.4 \times 10^{-17}\,\mathrm{s}$, far faster than they can interact, and feed the electromagnetic channel; the charged pions either interact again or decay to muons.

Derivation of the pion critical energy. A charged pion of Lorentz factor γ travels a mean distance $\gamma c\tau_\pi$ before decaying, with $\tau_\pi = 2.6033 \times 10^{-8}\,\mathrm{s}$ and $c\tau_\pi = 7.80\,\mathrm{m}$. It travels λ_I/ρ before interacting. At an altitude of 10 km, where most of the shower develops, the air density is $\rho \approx 0.41\,\mathrm{kg/m^3}$ and the interaction length is

$$\frac{\lambda_I}{\rho} = \frac{550\,\mathrm{kg/m^2}}{0.41\,\mathrm{kg/m^3}} = 1.34 \times 10^3\,\mathrm{m}\,.$$

Equating the two lengths gives the Lorentz factor at which the two fates are equally likely,

$$\gamma_c = \frac{1.34 \times 10^3\,\mathrm{m}}{7.80\,\mathrm{m}} = 172\,, \qquad E_c^\pi = \gamma_c m_\pi c^2 = 172 \times 139.57\,\mathrm{MeV} = 24\,\mathrm{GeV}\,,$$

in SI $3.8 \times 10^{-9}\,\mathrm{J}$. Pions above this energy interact and keep the hadronic cascade going; pions below it decay and become muons. The value depends on altitude through ρ and on energy through λ_I, and is conventionally taken as $E_c^\pi \approx 20\,\mathrm{GeV}$ for the shower as a whole [2080]; the derivation above shows where that number comes from and how much it can be trusted. Note how sensitive it is to the cross section: the older tabulated interaction length for air, $90\,\mathrm{g/cm^2}$, which is a few-GeV value, would give 39 GeV instead and turn a twenty-percent agreement into a factor-of-two puzzle. □

Theorem 115.21 (Muon number and its scaling)**.** *In the hadronic extension of Theorem 115.19, after n generations the energy remaining in the hadronic channel is $E_0\,(2/3)^n$, shared among N_{ch}^n charged pions; the cascade ends when the pion energy falls to E_c^π, and every surviving charged pion becomes a muon. Hence*

$$N_\mu = \left(\frac{E_0}{E_c^\pi}\right)^{\beta}\,, \qquad \beta = \frac{\ln N_{\mathrm{ch}}}{\ln\left(\frac{3}{2}N_{\mathrm{ch}}\right)} \approx 0.85\,, \tag{115.32}$$

so the muon number grows more slowly than linearly with primary energy, while the electromagnetic particle number grows linearly. The ratio of the two is therefore an energy-independent handle on the primary mass.

Rests on Theorem 115.19.

Proof. *Derives Theorem 115.21.*

Of the $\frac{3}{2}N_{\mathrm{ch}}$ pions made in one interaction, N_{ch} are charged and carry 2/3 of the energy onward; after n generations the hadronic energy is $E_0\,(2/3)^n$ spread over N_{ch}^n pions, so the energy per pion is $E_0\left(\frac{3}{2}N_{\mathrm{ch}}\right)^{-n}$. Setting this equal to E_c^π determines the number of generations,

$$n = \frac{\ln\left(E_0/E_c^\pi\right)}{\ln\left(\frac{3}{2}N_{\mathrm{ch}}\right)}\,,$$

and $N_\mu = N_{\mathrm{ch}}^n = \exp\left(n \ln N_{\mathrm{ch}}\right)$, which is Equation (115.32). With $N_{\mathrm{ch}} = 10$, $\beta = \ln 10/\ln 15 = 0.850$; the exponent is insensitive to N_{ch}, moving only from 0.80 to 0.88 as N_{ch} runs from 5 to 20, which is why it is quoted as a number rather than as a function. □

Corollary 115.22 (Superposition: more muons, shallower maximum)**.** *A nucleus of mass number A and total energy E_0 behaves, to a first approximation, as A independent nucleons each of energy E_0/A. Then*

$$N_\mu^A = A\left(\frac{E_0/A}{E_c^\pi}\right)^\beta = A^{1-\beta}N_\mu^p\,, \qquad \langle X_{\max}\rangle^A = \langle X_{\max}\rangle^p - D_{10}\log_{10}A\,. \tag{115.33}$$

An iron primary ($A=56$) therefore produces $56^{0.15}=1.83$ times as many muons as a proton of the same total energy, and its shower maximum lies about $D_{10}\log_{10}56 \approx 100\,\mathrm{g/cm^2}$ higher in the atmosphere, using the measured hadronic elongation rate $D_{10}\approx 58\,\mathrm{g/cm^2}$ per decade rather than the purely electromagnetic Equation (115.30)*.*

Rests on Equations (115.30) *and* (115.32)*.*

Proof. *Derives Corollary 115.22.*

Each of the A nucleons initiates a sub-shower of energy E_0/A; the muon numbers add, giving the first of Equation (115.33) directly from Equation (115.32), and $1-\beta=0.15$. The depths of maximum do not add: the shower maximum of the superposition is that of a single sub-shower, so $\langle X_{\max}\rangle$ is evaluated at E_0/A instead of at E_0, and by the definition Equation (115.30) of the elongation rate this shifts it by $-D_{10}\log_{10}A$. □

Remark 115.23 (The muon puzzle). Equation (115.32) is the point at which air-shower physics depends on hadronic interaction models extrapolated an order of magnitude in centre-of-mass energy beyond any collider measurement, and it is where the extrapolation currently fails. Measured muon numbers in showers above $1\times10^{19}\,\mathrm{eV}$ exceed the predictions of every hadronic model, by a factor of order 1.3 to 1.6, with the primary mass constrained independently by $X_{\max}$. The discrepancy is reported here as an unresolved defect of the modelling rather than absorbed into a fitted parameter, and it is one of the reasons a composition inference from muon number alone is not currently trustworthy; Section 115.8.1 returns to it.

115.4.3 Detection techniques

Two techniques dominate, and they measure different things.

A *surface array* samples the particle density of the shower front where it meets the ground, using either plastic scintillators or sealed tanks of water read by photomultipliers, the latter responding to Cherenkov light (Section 115.4.4) from the traversing particles and so being sensitive to the muonic as well as the electromagnetic component. From the sampled densities a lateral distribution function is fitted, and its value at a fixed distance from the axis — 1000 m for a kilometre-spaced array — is used as the energy estimator, because at that distance the fitted density is least sensitive to the shape assumed. The arrival direction comes from the relative timing.

Derivation of the angular resolution of an array. Treat the shower front as a plane advancing at speed c along the unit vector $\hat{\boldsymbol{n}}$. Two detectors separated by the baseline $\boldsymbol{d}$ are struck at times differing by

$$\Delta t = -\frac{\boldsymbol{d}\cdot\hat{\boldsymbol{n}}}{c} = \frac{d\sin\theta}{c}\,, \tag{115.34}$$

for a baseline in the plane of incidence, so that measuring Δt measures θ. Differentiating, $\sigma_\theta = c\,\sigma_t/\left(d\cos\theta\right)$. With a timing precision $\sigma_t = 10\,\mathrm{ns}$, achievable with photomultipliers and satellite time transfer, and a baseline $d=1500\,\mathrm{m}$,

$$\sigma_\theta = \frac{2.998\times10^{8}\,\mathrm{m/s}\times 1\times10^{-8}\,\mathrm{s}}{1500\,\mathrm{m}} = 2.0\times10^{-3}\,\mathrm{rad} = 0.11^\circ$$

at the zenith. The realised resolution of a large array is about 1°, an order of magnitude worse, because the front is not a plane but a curved and thick shell whose sampled thickness is tens of nanoseconds; Equation (115.34) sets the floor, not the answer. □

An *air-fluorescence detector* does something different: it photographs the shower in the dark. Nitrogen molecules excited by the shower's charged particles fluoresce isotropically in the near ultraviolet, at a yield of a few photons per metre per charged particle, and the light is proportional to the energy deposited. A telescope watching a dark sky therefore records the shower's longitudinal profile $N(X)$ directly, and the primary energy follows by integration,

$$E_{\mathrm{cal}} = \int \left(\frac{\mathrm{d}E}{\mathrm{d}X}\right)_{\mathrm{ion}} N(X)\,\mathrm{d}X\,, \tag{115.35}$$

which is calorimetry with no hadronic model in it at all. That is the technique's great advantage. It has two costs: the light must traverse kilometres of atmosphere whose transmission must be measured continuously, and Equation (115.35) misses the energy carried into the ground by muons and neutrinos, some 10 % to 15 %, which must be added back with a model-dependent correction. The technique was realised in the Fly's Eye detector [1748].

Their weaknesses are complementary. Fluorescence works only on clear moonless nights, a duty cycle of order 13 %; an array runs continuously. An array's energy scale depends on shower simulations; a fluorescence detector's does not. The Pierre Auger Observatory therefore does both, covering 3000 km^2 of Argentine plateau with 1600 water-Cherenkov tanks on a 1500 m triangular grid, overlooked by fluorescence telescopes at four sites, and using the subset of events seen by both — the *hybrid* events — to calibrate the array's estimator against the calorimetric energy [2241]. The energy scale of the whole observatory is then that of the fluorescence measurement, with a systematic uncertainty of about 14 %, and Section 115.4.6 shows what that costs.

115.4.4 The Cherenkov technique

A charged particle moving through a medium faster than the phase velocity of light in that medium radiates, at a fixed angle to its track. The phenomenon and the cone condition are established in Phenomenon 65.7: with n the refractive index and $\beta = v/c$,

$$\cos\theta_c = \frac{1}{n\beta}\,, \tag{115.36}$$

which has a solution only if $\beta > 1/n$ [789]. The intensity is Frank and Tamm's [792]; in SI, for a particle of charge Ze,

$$\frac{\mathrm{d}^2E}{\mathrm{d}x\,\mathrm{d}\omega} = \frac{(Ze)^2}{4\pi\varepsilon_0 c^2}\,\omega\left(1 - \frac{1}{\beta^2 n^2(\omega)}\right)\,, \tag{115.37}$$

an energy per unit path length per unit angular frequency, of SI dimension J s/m. Dividing by $\hbar\omega$ converts it to a photon count, and the factor of ω cancels:

$$\frac{\mathrm{d}^2N}{\mathrm{d}x\,\mathrm{d}\omega} = \frac{(Ze)^2}{4\pi\varepsilon_0\hbar c^2}\left(1 - \frac{1}{\beta^2 n^2}\right) = \frac{\alpha Z^2}{c}\left(1 - \frac{1}{\beta^2 n^2}\right)\,, \tag{115.38}$$

with $\alpha = e^2/(4\pi\varepsilon_0\hbar c)$ the dimensionless fine-structure constant of Section 100.1.1. The photon number per unit path is thus *flat in frequency*, which is why Cherenkov light looks blue: equal numbers per unit ω means many more per unit wavelength at the short end. The cone condition Equation (115.36) is derived, by the wavelet construction, in Phenomenon 65.7; the intensity Equation (115.37) is not derived anywhere in this book yet.

Derivation pending #205

The Frank–Tamm intensity formula in SI, obtained from the retarded fields of a uniformly moving charge in a dispersive medium by Fourier transform in space and time. Belongs in Appendix A; this chapter quotes the result and uses only its numerical consequences

Example 115.24 (Numbers for the three media that matter)**.** Rewriting Equation (115.38) per unit wavelength using $\omega = 2\pi c/\lambda$,

$$\frac{\mathrm{d}^2 N}{\mathrm{d}x\,\mathrm{d}\lambda} = \frac{2\pi\alpha Z^2}{\lambda^2}\left(1 - \frac{1}{\beta^2 n^2}\right),$$

and integrating over the photomultiplier band 300 nm to 600 nm for $\beta = 1$, $Z = 1$:

- *Water*, $n = 1.333$: $\theta_c = 41.4°$, and $\mathrm{d}N/\mathrm{d}x = 3.34 \times 10^4/\mathrm{m}$, that is 334 photons per centimetre of track. This is the tank of a surface array and the volume of Chapter 114.
- *Ice*, $n = 1.32$: $\theta_c = 40.7°$ and $\mathrm{d}N/\mathrm{d}x = 3.26 \times 10^4/\mathrm{m}$, essentially the same. This is IceCube (Section 115.7.4).
- *Air* at sea level, $n = 1.000293$: the cone collapses to $\theta_c = 1.39°$ and the yield to 45/m, but the track lengths are kilometres. The threshold is severe: $\beta > 1/n$ requires

  $$\gamma > \gamma_{\mathrm{th}} = \frac{1}{\sqrt{1 - n^{-2}}} = 41.3\,,$$

  so an electron needs 21 MeV and a muon 4.4 GeV. A shower's Cherenkov light is therefore emitted almost entirely by its electromagnetic component and is beamed into a narrow forward cone that illuminates a pool some 125 m in radius at the ground.

That narrow cone is the basis of gamma-ray astronomy from the ground. An *imaging atmospheric Cherenkov telescope* focuses the pool of light onto a camera of photomultipliers and forms an image of the shower; the image of a shower initiated by a gamma ray is a compact, regular ellipse pointing at the source position in the field, while a hadronic shower, which contains many sub-showers with large transverse momenta, gives a broad and irregular image. Hillas showed that a few moments of the light distribution — the ellipse's length, width, orientation and total intensity — separate the two classes with a rejection of better than 99 percent [1754]. The Whipple telescope used exactly this to detect TeV gamma rays from the Crab Nebula at 9σ, the first source ever found this way [1827], and every ground-based gamma-ray instrument since is a development of it.

115.4.5 Depth of maximum and mass composition

The observable that carries composition information with the least model dependence is $X_{\max}$, together with its shower-to-shower fluctuation. Both are predicted by Corollary 115.22 and both are measured directly by a fluorescence detector.

Phenomenon 115.25 (Composition inferred from the depth of maximum)**.** The mean depth of shower maximum increases logarithmically with primary energy, and at fixed energy is about $100\,\mathrm{g/cm^2} = 1000\,\mathrm{kg/m^2}$ smaller for an iron primary than for a proton. Its shower-to-shower spread is likewise smaller for heavy primaries. Auger measurements show $\langle X_{\max}\rangle$ and its spread consistent with a light, largely protonic composition near the ankle, becoming progressively heavier with increasing energy above it [2226].

Rests on Equation (115.31) and Corollary 115.22.

Derivation of the fluctuation scaling. *Derives Phenomenon 115.25.*

Two independent sources of fluctuation enter. The first is the depth of the first interaction, which is exponentially distributed with mean λ_I, so it contributes a standard deviation λ_I to $X_{\max}$. The relevant λ_I is the one at the energy of the measurement, and by Equation (115.31) it shrinks as the cross section grows: near $1 \times 10^{19}\,\mathrm{eV}$ the inelastic proton–air cross section is about $5.0 \times 10^{-29}\,\mathrm{m}^2$, that is 500 millibarn, so

$$\lambda_I = \frac{2.41 \times 10^{-26}\,\mathrm{kg}}{5.0 \times 10^{-29}\,\mathrm{m}^2} \approx 480\,\mathrm{kg/m^2} = 48\,\mathrm{g/cm^2}\,.$$

The second source is the fluctuation in the subsequent cascade, of comparable size; the two add in quadrature, and $\left(48^2 + 36^2\right)^{1/2} = 60$ reproduces the measured proton width of some $60\,\mathrm{g/cm^2}$.

For a nucleus of mass A, superposition (Corollary 115.22) replaces one shower by A sub-showers of energy E_0/A, each with its own first-interaction depth. The shower maximum of the sum is essentially the average of the A sub-maxima, and the variance of an average of A independent variables is $1/A$ of the individual variance:

$$\sigma\left(X_{\max}\right)\Big|_A \approx \frac{\sigma\left(X_{\max}\right)\Big|_p}{\sqrt{A}}\,. \tag{115.39}$$

For iron $\sqrt{56} = 7.5$, so the rule predicts a collapse from the proton value of some $60\,\mathrm{g/cm^2}$ to $8\,\mathrm{g/cm^2}$. The measured iron width is about $20\,\mathrm{g/cm^2}$: the superposition rule gets the direction and the order of the effect right and underpredicts the width by a factor of about two and a half, which is the same crudeness Remark 115.20 concedes for the model it rests on. The qualitative conclusion — heavy primaries fluctuate much less than light ones — is robust; the number is not. *Mean* and *width* are thus two independent composition observables, and they must agree: a mean lying between the proton and iron expectations can be produced either by an intermediate species, which would have a narrow distribution, or by a mixture of protons and iron, which would have a very wide one. Auger sees the mean move heavier above the ankle *and* the width narrow, which favours a genuinely intermediate composition over a mixture [2226]. □

The consequence for Section 115.6.2 is direct and uncomfortable. If the highest-energy primaries are not protons, the suppression at the end of the spectrum cannot be read straightforwardly as photopion production on the microwave background, because Equation (115.22) allows the same break to be the maximum rigidity of the sources. That objection is stated where it belongs, in the derivation of Phenomenon 115.42.

115.4.6 Energy reconstruction and its uncertainty

Every spectrum in this chapter is a histogram of reconstructed energies, and two distinct errors act on it. Neither is small, and the statistical machinery for both is that of Chapter 11.

Counting. In an energy bin containing N events the estimator of the rate is the maximum-likelihood estimator of a Poisson mean, which by Example 11.58 is N itself with standard deviation $\sqrt{N}$. At the highest energies N is single digits: a bin with four events carries a 50 % statistical uncertainty, and no amount of care in the reconstruction improves it. This is why the exposure — area times time times solid angle, in $\mathrm{km^2\,yr\,sr}$ — is the figure of merit quoted for every observatory.

Energy scale and resolution. These act in opposite directions and must not be confused.

Proposition 115.26 (A steep spectrum converts an energy error into a flux error)**.** *Let the true differential intensity be* $J(E) = AE^{-\gamma}$.

1. *If the energy scale carries a systematic multiplicative bias, $E_{\mathrm{rec}} = (1+\epsilon)\,E$, the reconstructed integral intensity above a fixed energy is in error by the fraction*

$$\frac{\Delta J}{J} = (\gamma - 1)\,\epsilon\,, \tag{115.40}$$

   to first order in ϵ.

2. *If instead the energy is unbiased but smeared, with $\ln E_{\mathrm{rec}} = \ln E + u$ and u Gaussian of zero mean and standard deviation σ, the reconstructed differential intensity is the true one multiplied by*

$$\frac{J_{\mathrm{rec}}(E)}{J(E)} = \exp\left[\frac{(\gamma-1)^2\,\sigma^2}{2}\right] > 1\,, \tag{115.41}$$

   the same power law with an inflated normalisation. Resolution alone therefore raises *a measured flux; it never lowers it.*

Rests on Equation (115.21).

Proof. *Derives Proposition 115.26.*

For (i), $J(>E) = AE^{-(\gamma-1)}/(\gamma-1)$ by Equation (115.21). A scale bias means the flux reported at E is the true flux at $E/(1+\epsilon)$, so

$$\frac{J_{\mathrm{rec}}(>E)}{J(>E)} = (1+\epsilon)^{\gamma-1} = 1 + (\gamma-1)\,\epsilon + O\left(\epsilon^2\right)\,.$$

For (ii), change variable to $x = \ln E$. The number of events per unit x is $J(E)E = Ae^{-(\gamma-1)x}$, and smearing convolves this with a Gaussian of width σ:

$$f_{\mathrm{rec}}(x_r) = \int A\mathrm{e}^{-(\gamma-1)x}\frac{\mathrm{e}^{-(x_r-x)^2/2\sigma^2}}{\sqrt{2\pi}\sigma}\mathrm{d}x = A\mathrm{e}^{-(\gamma-1)x_r}\mathrm{e}^{(\gamma-1)^2\sigma^2/2}\,,$$

by completing the square — this is the moment generating function of a Gaussian evaluated at $\gamma - 1$. Dividing by E_r to return to a differential intensity gives Equation (115.41). The result is exact, not perturbative, and holds for any σ. □

Example 115.27 (What this costs at the end of the spectrum)**.** Above the suppression the local index is $\gamma \approx 4.5$. An energy-scale uncertainty of $\epsilon = 14\,\%$ — the Auger systematic [2241] — gives by Equation (115.40) a flux uncertainty of $3.5 \times 0.14 = 49\,\%$: a half-order-of-magnitude band on the vertical axis of every published spectrum, and much larger than the statistical error in the same bins. An energy resolution of $\sigma = 0.20$ in the logarithm gives by Equation (115.41) a flux inflation of $\exp\left(3.5^2 \times 0.04/2\right) = 1.28$, a $28\,\%$ overestimate that must be unfolded away before the spectrum is quoted. Both effects grow with γ, which is precisely why they matter most where the spectrum is steepest and the physics is most interesting.

Remark 115.28 (Stating a break honestly). The significance quoted for a spectral break — the 5σ and 6σ of Phenomenon 115.42 — is a likelihood-ratio statistic: the maximised likelihood of a broken power law against that of a single one, converted to a significance by Wilks's theorem (Theorem 11.88) and Proposition 11.93. Two cautions apply and are respected in the published analyses. First, the break energy is a free parameter that is *not* identified under the null hypothesis, which is the failure mode of Wilks's theorem discussed in Remark 11.91; the correct calibration is by simulation, and it lowers the significance. Second, the systematic of Equation (115.40) is common to all bins and so shifts the whole spectrum rather than scattering it, which means it cannot create a break — a genuinely useful fact, and the reason the suppression survives a $14\,\%$ scale uncertainty intact.

115.5 Acceleration and origin

115.5.1 The supernova-remnant hypothesis

Baade and Zwicky, in the short paper in which they proposed the neutron star, suggested that supernovae are the source of the cosmic radiation [730]; the term *super-nova* itself was introduced in the companion paper printed immediately before it in the same volume. The proposal was made before any acceleration mechanism was known, on energetic grounds alone, and the energetics remain the strongest argument for it.

Phenomenon 115.29 (Supernovae can supply the Galactic cosmic-ray power)**.** The energy density of cosmic rays in the Galaxy is about $1\,\mathrm{eV/cm^3} = 1.60 \times 10^{-13}\,\mathrm{J/m^3}$, and they escape on a time scale of about $1 \times 10^7\,\mathrm{yr}$. Sustaining that density requires a power of order $6 \times 10^{33}\,\mathrm{W}$, which is about 10 % of the kinetic power of Galactic supernovae. No other Galactic population comes within an order of magnitude of supplying it.

Rests on Equation (115.24).

Derivation. *Derives Phenomenon 115.29.*

Take the confinement volume to be a disc of radius $R = 15\,\mathrm{kpc} = 4.63 \times 10^{20}\,\mathrm{m}$ and half-thickness $h = 300\,\mathrm{pc} = 9.26 \times 10^{18}\,\mathrm{m}$:

$$V = \pi R^2\,(2h) = 1.25 \times 10^{61}\,\mathrm{m^3}\,,$$

so the stored energy is

$$U = u_{\mathrm{CR}} V = 1.602 \times 10^{-13}\,\mathrm{J/m^3} \times 1.25 \times 10^{61}\,\mathrm{m^3} = 2.0 \times 10^{48}\,\mathrm{J}\,.$$

In a steady state the injected power equals the stored energy divided by the residence time $\tau_{\mathrm{esc}} \approx 1 \times 10^7\,\mathrm{yr} = 3.16 \times 10^{14}\,\mathrm{s}$ of Section 115.5.5:

$$P_{\mathrm{CR}} = \frac{U}{\tau_{\mathrm{esc}}} = \frac{2.0 \times 10^{48}\,\mathrm{J}}{3.16 \times 10^{14}\,\mathrm{s}} = 6.3 \times 10^{33}\,\mathrm{W}\,. \tag{115.42}$$

The available power is that of the supernovae. A core-collapse supernova delivers of order $1 \times 10^{44}\,\mathrm{J}$ to the kinetic energy of its ejecta — three tenths of one percent of the $3 \times 10^{46}\,\mathrm{J}$ of gravitational binding energy released, the remaining ninety-nine and more percent leaving as the neutrino burst of Section 115.7.3 — and the Galactic rate is about two per century:

$$P_{\mathrm{SN}} = 1 \times 10^{44}\,\mathrm{J} \times \frac{2}{100\,\mathrm{yr}} = 1 \times 10^{44}\,\mathrm{J} \times 6.34 \times 10^{-10}/\mathrm{s} = 6.3 \times 10^{34}\,\mathrm{W}\,.$$

The required efficiency is therefore

$$\eta = \frac{P_{\mathrm{CR}}}{P_{\mathrm{SN}}} = \frac{6.3 \times 10^{33}\,\mathrm{W}}{6.3 \times 10^{34}\,\mathrm{W}} \approx 0.10\,, \tag{115.43}$$

ten percent of the mechanical energy of the blast wave. That is demanding but not absurd, and it is exactly what the shock acceleration of Section 115.5.3 predicts when the accelerated particles are dynamically important. Every alternative Galactic reservoir — stellar winds, pulsar rotational energy, the interstellar magnetic field — falls short of P_{CR} by an order of magnitude or more, which is why the supernova hypothesis has survived ninety years without a rival. □

An energy budget is not a detection, and for most of that ninety years the argument stopped there: supernova remnants were known to be bright in radio synchrotron emission, which proves they accelerate *electrons*, but the cosmic radiation is overwhelmingly protons and nuclei, and an electron accelerator is not evidence of a proton accelerator. The proof arrived in 2013.

Phenomenon 115.30 (Protons are accelerated in supernova remnants)**.** The gamma-ray spectra of the remnants IC 443 and W44, measured between 60 MeV and 2 GeV, show a characteristic break and low-energy turnover that is the kinematic signature of neutral-pion decay, and that no electron process reproduces [2205]. Protons and nuclei are therefore accelerated in these objects, not only electrons.

Rests on Phenomenon 115.29.

Derivation of the pion-decay signature. Derives Phenomenon 115.30.

A neutral pion at rest decays to two photons each of energy $m_{\pi^0}c^2/2 = 67.5\,\mathrm{MeV}$. Boost it to Lorentz factor γ_π: by the transformation of a null four-momentum, the photon energy in the laboratory is

$$E_\gamma = \gamma_\pi \frac{m_\pi c^2}{2}\left(1 + \beta_\pi \cos\theta^*\right) ,$$

with θ^* the emission angle in the pion frame. Because the decay is isotropic in that frame, $\cos\theta^*$ is uniform on $[-1, 1]$, so E_γ is uniformly distributed between $\gamma_\pi\left(1-\beta_\pi\right)m_\pi c^2/2$ and $\gamma_\pi\left(1+\beta_\pi\right)m_\pi c^2/2$. The distribution is flat, and in $\log E_\gamma$ it is *symmetric about* $\log\left(m_\pi c^2/2\right)$, since the two limits multiply to $\left(m_\pi c^2/2\right)^2$ whatever γ_π is. Summing over any pion spectrum preserves that symmetry point. The resulting photon spectrum therefore rises to a maximum at $E_\gamma = 67.5\,\mathrm{MeV}$ and falls below it, whatever the parent proton spectrum — the "pion bump".

No leptonic process produces such a feature. Bremsstrahlung from electrons follows the electron spectrum with no scale in it; inverse-Compton scattering likewise. The observed turnover below 200 MeV in IC 443 and W44 is thus a direct measurement of a hadronic population, and it is the observation that converts Phenomenon 115.29 from a budget into evidence [2205]. □

115.5.2 Fermi acceleration

Fermi proposed that cosmic rays gain energy by scattering repeatedly off moving magnetised clouds in the interstellar medium [929]. A magnetic field does no work, so all the energy comes from the motion of the scattering centre, and the sign of the exchange depends on whether the encounter is head-on or overtaking.

Theorem 115.31 (Second-order Fermi acceleration)**.** *A relativistic particle scattering elastically off a randomly moving cloud of speed $V \ll c$, isotropised in the cloud frame, gains on average the fraction*

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{4}{3}\left(\frac{V}{c}\right)^2 \tag{115.44}$$

of its energy per encounter. The gain is second *order in V/c, because head-on and overtaking encounters cancel at first order and only the excess of head-on encounters survives.*

Rests on Definition 40.1 and Equation (40.3).

Proof. Derives Theorem 115.31.

Write $\beta_V = V/c$ and $\gamma_V = \left(1-\beta_V^2\right)^{-1/2}$, and let θ be the angle in the laboratory between the particle's direction and the cloud's velocity. Transforming the particle's energy into the cloud frame (a boost, as in Section 40.1),

$$E' = \gamma_V E\left(1 - \beta_V \cos\theta\right) ,$$

for an ultrarelativistic particle, $pc = E$. The scattering is elastic in the cloud frame — the field does no work there — so $E'' = E'$ with a new direction θ'', and transforming back,

$$E_f = \gamma_V E''\left(1 + \beta_V \cos\theta''\right) = \gamma_V^2 E\left(1 - \beta_V\cos\theta\right)\left(1 + \beta_V\cos\theta''\right) .$$

Isotropisation in the cloud frame gives $\langle \cos\theta'' \rangle = 0$, so

$$\left\langle \frac{\Delta E}{E} \right\rangle = \gamma_V^2 \left(1 - \beta_V \langle \cos\theta \rangle\right) - 1 \,.$$

The laboratory angles are not uniformly sampled: the rate of encounters with a cloud approaching at angle θ is proportional to the relative closing speed, $\propto (1 - \beta_V \cos\theta)$, so

$$\langle \cos\theta \rangle = \frac{\int_{-1}^{1} \mu \,(1 - \beta_V \mu)\, \mathrm{d}\mu}{\int_{-1}^{1} (1 - \beta_V \mu)\, \mathrm{d}\mu} = \frac{-\frac{2}{3}\beta_V}{2} = -\frac{\beta_V}{3} \,.$$

Head-on encounters, which gain energy, are commoner than overtaking ones, which lose it. Hence

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{1 + \frac{1}{3}\beta_V^2}{1 - \beta_V^2} - 1 = \frac{4}{3}\beta_V^2 + O\left(\beta_V^4\right) \,,$$

which is Equation (115.44). □

Remark 115.32 (Why the second-order mechanism is not the answer). Interstellar clouds move at $V \approx 30\,\mathrm{km/s}$, so $\beta_V = 1 \times 10^{-4}$ and Equation (115.44) gives 1.3×10^{-8} per encounter. With a scattering mean free path of order a parsec the encounter interval is $1\,\mathrm{pc}/c = 3.3\,\mathrm{yr}$, so the energy e-folding time is

$$t_{\mathrm{acc}} \approx \frac{3.3\,\mathrm{yr}}{1.3 \times 10^{-8}} = 2.4 \times 10^{8}\,\mathrm{yr} \,,$$

some twenty-five times longer than the residence time $\tau_{\mathrm{esc}} \approx 1 \times 10^{7}\,\mathrm{yr}$ of Section 115.5.5. The particles leave the Galaxy long before they are accelerated. Worse, the predicted spectral index depends on the ratio $t_{\mathrm{acc}}/\tau_{\mathrm{esc}}$, which is a free parameter with no reason to take the same value in different regions, so the mechanism does not explain why the spectrum is a power law of one universal index. Both objections are removed at a stroke if the scattering centres approach the particle systematically rather than randomly, which is what a shock provides.

115.5.3 Diffusive shock acceleration

A supernova blast wave is a strong shock: a discontinuity across which the fluid speed drops and the density rises, propagating into the interstellar medium at $u_1 \approx 1 \times 10^{7}\,\mathrm{m/s}$. Magnetic irregularities are frozen into the fluid on both sides, so a particle crossing the shock finds the scattering centres on the far side approaching it — *whichever way it crosses.* Every crossing is head-on, and the first-order cancellation of Theorem 115.31 is gone.

Theorem 115.33 (The spectrum from a strong shock)**.** *Let a shock have upstream speed u_1 and downstream speed u_2 in the shock frame, with compression ratio $r = u_1/u_2$. A particle that repeatedly crosses and recrosses gains, per complete cycle,*

$$\xi := \left\langle \frac{\Delta E}{E} \right\rangle = \frac{4}{3}\,\frac{u_1 - u_2}{c} \,, \tag{115.45}$$

first order in the speeds, and escapes downstream with probability

$$P_{\mathrm{esc}} = \frac{4u_2}{c} \tag{115.46}$$

per cycle. The resulting integral spectrum is a power law

$$N(>E) \propto E^{-3/(r-1)} \,, \qquad \frac{\mathrm{d}N}{\mathrm{d}E} \propto E^{-\gamma_s} \,, \qquad \gamma_s = 1 + \frac{3}{r-1} = \frac{r+2}{r-1} \,, \tag{115.47}$$

and for a strong shock in a monatomic gas, $r = 4$ and $\gamma_s = 2$ — independently of the shock speed, the magnetic field and every other detail.

Rests on Theorem 115.31.

Proof. *Derives Theorem 115.33.*

The gain. Transform to the frame of the upstream fluid. The downstream fluid approaches it at speed $u_1 - u_2$, so a particle that crosses into the downstream region and is isotropised there undergoes exactly the calculation of Theorem 115.31 with $\beta_V = (u_1 - u_2)/c$, but with one difference: the crossing is always head-on, so the average over $\cos\theta$ is taken over the *flux* of particles crossing a plane, whose distribution is $\propto \mu\,\mathrm{d}\mu$ on $[0,1]$, giving $\langle\mu\rangle = \frac{2}{3}$. Each of the two crossings that make a cycle contributes $\beta_V\langle\mu\rangle = \frac{2}{3}\beta_V$, so $\xi = \frac{4}{3}\beta_V$, which is Equation (115.45). The result is first order because the approach is systematic: there are no overtaking encounters to cancel it.

The escape. Downstream the particles are isotropic in the fluid frame and their flux onto the shock from behind is $nc/4$, the standard one-sided flux of an isotropic gas. They are simultaneously carried away from the shock by the downstream flow at the rate nu_2. The fraction lost per cycle is the ratio,

$$P_{\mathrm{esc}} = \frac{nu_2}{nc/4} = \frac{4u_2}{c}\,,$$

which is Equation (115.46). Note that neither ξ nor P_{esc} depends on the particle's energy. That is the whole reason the spectrum is a power law.

The spectrum. After k cycles a particle has energy $E_k = E_0(1+\xi)^k$ and the number surviving is $N_k = N_0(1-P_{\mathrm{esc}})^k$. Eliminating k,

$$\frac{\ln(N_k/N_0)}{\ln(E_k/E_0)} = \frac{\ln(1-P_{\mathrm{esc}})}{\ln(1+\xi)} \approx -\frac{P_{\mathrm{esc}}}{\xi}\,,$$

so $N(>E) \propto E^{-P_{\mathrm{esc}}/\xi}$ with

$$\frac{P_{\mathrm{esc}}}{\xi} = \frac{4u_2/c}{\frac{4}{3}(u_1-u_2)/c} = \frac{3u_2}{u_1-u_2} = \frac{3}{r-1}\,,$$

which is Equation (115.47); the differential index is one greater.

The compression ratio. Work in the shock frame, where the flow is steady, and impose conservation of mass, momentum and energy across the discontinuity — the Rankine–Hugoniot conditions, which belong to the fluid mechanics of Section 31.10.2:

$$\rho_1 u_1 = \rho_2 u_2 =: j\,, \qquad P_1 + \rho_1 u_1^2 = P_2 + \rho_2 u_2^2\,, \qquad h_1 + \tfrac{1}{2}u_1^2 = h_2 + \tfrac{1}{2}u_2^2\,,$$

with the specific enthalpy of an ideal gas $h = \gamma_{\mathrm{ad}}P/[(\gamma_{\mathrm{ad}}-1)\rho]$. A *strong* shock is one whose upstream pressure is negligible, $P_1 \to 0$ and hence $h_1 \to 0$; the momentum condition then gives $P_2 = j(u_1-u_2)$, and substituting this and $\rho_2 = j/u_2$ into the energy condition,

$$\frac{u_1^2 - u_2^2}{2} = \frac{\gamma_{\mathrm{ad}}P_2u_2}{(\gamma_{\mathrm{ad}}-1)j} = \frac{\gamma_{\mathrm{ad}}(u_1-u_2)u_2}{\gamma_{\mathrm{ad}}-1}\,.$$

Cancelling the common factor $u_1 - u_2$, which is non-zero for a genuine shock, leaves $u_1 + u_2 = 2\gamma_{\mathrm{ad}}u_2/(\gamma_{\mathrm{ad}}-1)$ and therefore

$$r = \frac{u_1}{u_2} = \frac{\gamma_{\mathrm{ad}}+1}{\gamma_{\mathrm{ad}}-1} = \frac{5/3+1}{5/3-1} = 4 \tag{115.48}$$

for a monatomic ionised gas, whence $\gamma_s = 2$. □

The independence of Equation (115.47) from every parameter except r is what makes the mechanism credible, and it is why the result was found independently and almost simultaneously by several groups [1614, 1616]. It also answers Phenomenon 115.11: a featureless power law over eleven decades requires an energy-independent gain and an energy-independent escape, and a strong shock supplies both.

Proposition 115.34 (Propagation steepens the source spectrum)**.** *If sources inject a spectrum $Q(E) \propto E^{-\gamma_s}$ into a Galaxy from which particles escape on a rigidity-dependent time scale $\tau_{\rm esc} \propto R^{-\delta}$, the steady-state spectrum observed is*

$$J(E) \propto Q(E)\tau_{\rm esc}(E) \propto E^{-(\gamma_s+\delta)}\,. \tag{115.49}$$

Rests on Equations (115.1) *and* (115.25)*.*

Proof. *Derives Proposition 115.34.*

In a steady state the number density at energy E satisfies injection equals escape: $Q(E) = n(E)/\tau_{\rm esc}(E)$, so $n(E) = Q(E)\tau_{\rm esc}(E)$, and the intensity is $J = nc/4\pi$ for an isotropic distribution. Since the particles are ultrarelativistic, $R = E/(Ze)$ and the rigidity dependence is an energy dependence at fixed Z. □

Remark 115.35 (The arithmetic does not close, and saying so is the point). Put the two results together. Theorem 115.33 gives $\gamma_s = 2$; the boron-to-carbon measurement Equation (115.25) gives $\delta = 0.333 \pm 0.014$ [2266]; Equation (115.49) therefore predicts an observed index of 2.33 ± 0.02. The measured index below the knee is 2.7. The discrepancy is 0.37 in the exponent, which over the four decades between 10 GeV and 1×10^{14} eV is a factor of $10^{4\times 0.37} = 30$ in flux — far outside any uncertainty in either input.

Something in the chain is wrong, and the candidates are known: the test-particle assumption behind Theorem 115.33 fails when the accelerated particles carry a tenth of the shock's energy, as Equation (115.43) requires them to, and their pressure modifies the shock structure so that r and hence γ_s become energy dependent; and the escape time inferred from secondary nuclei is an average over the Galaxy that need not describe the region where the sources are. This chapter does not adjudicate between them. It records that diffusive shock acceleration explains the *existence* and the *universality* of the power law, and does not yet explain its *index*.

Derivation of the maximum energy. *Derives Equation* (115.51)*.*

The mechanism has a limit, and it is the reason the knee exists in the first reading of Remark 115.14. The time to complete one cycle is set by how far the particle diffuses from the shock before returning. Consider the upstream side. The flow sweeps the particles onto the shock at speed u_1 while diffusion carries them away from it, and the balance $u_1 n = D_1\,{\rm d}n/{\rm d}x$ gives a precursor of scale length D_1/u_1, hence a column nD_1/u_1 of particles per unit area waiting to recross. They recross at the one-sided isotropic flux $nc/4$ already used in Equation (115.46), so the mean time spent upstream per crossing is $(nD_1/u_1)\,/\,(nc/4) = 4D_1/(u_1 c)$, and the same argument downstream gives $4D_2/(u_2 c)$. One complete cycle therefore takes

$$t_{\rm cyc} = \frac{4}{c}\left(\frac{D_1}{u_1} + \frac{D_2}{u_2}\right),$$

and since each cycle multiplies the energy by $1+\xi$ with the ξ of Equation (115.45), the time to raise the energy by one e-folding is $t_{\rm acc} = t_{\rm cyc}/\xi$, that is

$$t_{\rm acc} = \frac{3}{u_1 - u_2}\left(\frac{D_1}{u_1} + \frac{D_2}{u_2}\right).$$

Taking the same diffusion coefficient D on both sides and $r = 4$, so that $u_2 = u_1/4$ and $u_1 - u_2 = 3u_1/4$, the bracket is $5D/u_1$ and

$$t_{\rm acc} = \frac{20\,D}{u_1^2}\,, \qquad D = \frac{r_L c}{3} = \frac{E}{3ZeB}\,, \tag{115.50}$$

the second expression being the smallest diffusion coefficient a magnetic field permits, the Bohm value, obtained when the scattering mean free path equals the gyroradius of Equation (115.8). Setting t_{acc} equal to the lifetime T of the shock,

$$E_{\max} = \frac{3ZeB\,u_1^2 T}{20}\,. \tag{115.51}$$

With $Z = 1$, $u_1 = 1 \times 10^7\,\mathrm{m/s}$, $T = 1000\,\mathrm{yr} = 3.16 \times 10^{10}\,\mathrm{s}$ and the ambient interstellar field $B = 3 \times 10^{-10}\,\mathrm{T}$,

$$E_{\max} = 2.3 \times 10^{-5}\,\mathrm{J} = 1.4 \times 10^{14}\,\mathrm{eV}\,,$$

a factor of twenty below the knee at $3 \times 10^{15}\,\mathrm{eV}$. The mechanism as stated therefore fails to reach the knee. It succeeds if the field at the shock is amplified above the ambient value by a factor of order 30: with $B = 1 \times 10^{-8}\,\mathrm{T}$, Equation (115.51) gives $E_{\max} = 4.7 \times 10^{15}\,\mathrm{eV}$, which does reach it. Such amplification is observed — the X-ray synchrotron filaments at young remnant shocks are thin, and their thinness measures a downstream field of order $1 \times 10^{-8}\,\mathrm{T}$ — and the physical mechanism proposed for it is a streaming instability driven by the accelerated particles themselves. □

Derivation pending #206

The non-resonant streaming instability by which the accelerated particles amplify the magnetic field at the shock, and the saturation level that fixes $[B]$ in the maximum-energy estimate. Belongs in Appendix A; the estimate in this chapter takes the amplified field from the X-ray filament observations instead of deriving it

115.5.4 The Hillas criterion and candidate sources

Whatever the mechanism, an accelerator cannot confine a particle whose gyroradius exceeds its own size. That single observation, due to Hillas [1741], excludes most of the astrophysical zoo at the highest energies and it does so with no assumption about how the acceleration works.

Theorem 115.36 (The Hillas criterion)**.** *An accelerator of characteristic size L containing a magnetic field B, whose scattering centres move with speed βc, cannot accelerate a nucleus of charge Ze beyond*

$$E_{\max} = Ze\,\beta\,B\,L\,c\,, \tag{115.52}$$

an SI energy in joules when B is in tesla and L in metres. Equivalently the maximum rigidity of Equation (115.1) is $R_{\max} = \beta BLc$, in volts.

Rests on Equations (115.1), (115.8) and (115.45).

Proof. *Derives Theorem 115.36.*

By Equation (115.8) a particle of rigidity R has a gyroradius $r_L = R/(cB)$. It remains in the accelerating region only while $r_L \le L$, i.e. while $R \le BLc$; beyond that it leaves in less than one gyration and cannot be turned back. The factor β enters because the energy gain per crossing is first order in the speed of the scattering centres (Equation (115.45)), so a slow shock requires many crossings and loses a corresponding factor. Converting rigidity to energy with $E = ZeR$ gives Equation (115.52). Dimensionally, $[Ze]\,[B]\,[L]\,[c] = \mathrm{C \cdot T \cdot m \cdot m/s} = \mathrm{J}$, as required. □

Example 115.37 (The criterion applied)**.** Evaluating Equation (115.52) for $Z = 1$:

Object	B (T)	L (m)	β	$E_{\max}$ (eV)
Neutron star	1×10^{8}	1×10^{4}	1	3×10^{20}
Supernova remnant shock	1×10^{-8}	3.1×10^{17}	0.03	2.8×10^{16}
Active-galaxy jet (parsec scale)	1×10^{-7}	3.1×10^{16}	1	9×10^{17}
Radio-galaxy lobe	1×10^{-9}	3.1×10^{21}	1	9×10^{20}
Cluster accretion shock	1×10^{-10}	3.1×10^{22}	0.003	2.8×10^{18}
Galactic disc (for comparison)	3×10^{-10}	4.6×10^{20}	1	4×10^{19}

Only two entries reach 1×10^{20} eV for a proton, and each has a difficulty. The neutron-star value assumes the whole magnetospheric potential is available to one particle, which radiative losses in that field forbid. The radio-galaxy lobe is a plausible site but the nearest such object is tens of megaparsecs away, which the horizon of Section 115.6.1 makes marginal. Note also that Equation (115.52) scales with Z, so an iron nucleus reaches 26 times the plotted energy, which is why the composition result of Phenomenon 115.25 changes the shortlist.

Remark 115.38 (No source of the highest-energy rays is identified). This is the honest state of the subject. The Hillas criterion tells us which objects are *not* excluded; it identifies nothing. The anisotropy of Section 115.6.3 establishes that the highest-energy particles come from outside the Galaxy and no more. Correlations of arrival directions with catalogues of nearby active galaxies and starburst galaxies have been reported at significances between 3σ and 4σ and have neither strengthened decisively nor gone away with additional exposure. This chapter therefore names candidates and states that none is established.

115.5.5 Galactic propagation and confinement

Between the source and the detector the particles diffuse through the Galactic magnetic field. The simplest description that fits the data is the *leaky box*: the Galaxy is treated as a container of uniform density from which particles escape with a constant probability per unit time, $1/\tau_{\mathrm{esc}}$. It has one free function, $\tau_{\mathrm{esc}}(R)$, and two independent measurements of it.

The first is the grammage of Equation (115.24): the traversed matter is $X = \bar{m} n \, c \, \tau_{\mathrm{esc}}$ for a particle moving at essentially c through gas of number density n, so

$$\tau_{\mathrm{esc}} = \frac{X}{\bar{m} n c} = \frac{84\,\mathrm{kg/m^2}}{1.673 \times 10^{-27}\,\mathrm{kg} \times 3 \times 10^{5}/\mathrm{m^3} \times 2.998 \times 10^{8}\,\mathrm{m/s}} = 5.6{\times}10^{14}\,\mathrm{s} = 1.8{\times}10^{7}\,\mathrm{yr}\,, \tag{115.53}$$

where the effective density $n \approx 3 \times 10^{5}/\mathrm{m^3}$ (that is $0.3/\mathrm{cm^3}$) is the disc density averaged over a trajectory that spends most of its time in the low-density halo. That average is the weak point of the estimate, and it is why a second, independent measurement matters.

The second is a clock. Beryllium-10 is produced by the same spallation that produces boron, and it is radioactive with a half-life of 1.39×10^{6} yr, that is a mean life $\tau_{1/2}/\ln 2 = 2.0 \times 10^{6}$ yr. If the residence time were much shorter than this, all the ^{10}Be produced would still be present and the measured $^{10}\mathrm{Be}/^{9}\mathrm{Be}$ ratio would equal the production ratio; if it were much longer, essentially none would survive. The observed ratio is suppressed by a factor of several relative to production, which requires

$$\tau_{\mathrm{esc}} \approx 1.5 \times 10^{7}\,\mathrm{yr} = 4.7 \times 10^{14}\,\mathrm{s}\,,$$

in agreement with Equation (115.53) and independent of the assumed gas density — which is what makes the agreement informative rather than circular. The corresponding diffusion

coefficient, for a halo of half-height $H \approx 3\,\mathrm{kpc} = 9.3 \times 10^{19}\,\mathrm{m}$, is

$$D \approx \frac{H^2}{2\tau_{\mathrm{esc}}} = \frac{(9.3 \times 10^{19}\,\mathrm{m})^2}{2 \times 4.7 \times 10^{14}\,\mathrm{s}} \approx 1 \times 10^{25}\,\mathrm{m^2/s}\,, \tag{115.54}$$

some thirty orders of magnitude above the self-diffusion coefficient of a gas at room temperature and atmospheric pressure, which is of order $2 \times 10^{-5}\,\mathrm{m^2/s}$, and rising as R^{δ} with the $\delta = 0.333 \pm 0.014$ of Equation (115.25) [2266]. Standard treatments of the whole transport problem, of which the leaky box is the crudest useful limit, are in [2271].

Remark 115.39 (Why charged cosmic rays are not astronomy). Equation (115.54) is also the end of any hope of pointing with them. A particle that diffuses for $1.5 \times 10^{7}\,\mathrm{yr}$ before leaving the Galaxy has had its direction randomised completely: by Equation (115.23) a $1 \times 10^{15}\,\mathrm{eV}$ proton in $3 \times 10^{-10}\,\mathrm{T}$ has a gyroradius of 0.36 pc, utterly negligible against the kiloparsec scales of the Galaxy, so its arrival direction retains no memory whatever of its source. Only above $1 \times 10^{19}\,\mathrm{eV}$, where Phenomenon 115.15 gives gyroradii of kiloparsecs, does any directional information survive — and there the sources are extragalactic and the intervening extragalactic fields are poorly known. This is the reason the subject needs neutral messengers, and Section 115.7 is about them.

115.6 The end of the spectrum

115.6.1 The Greisen–Zatsepin–Kuz'min prediction

Within two years of the discovery of the cosmic microwave background, Greisen [1330] and, independently, Zatsepin and Kuz'min [1339] pointed out that it makes intergalactic space opaque to protons above a definite energy. The argument is pure relativistic kinematics applied to a measured photon bath, and it is one of the sharpest predictions in astrophysics: it uses no unknown source, no unknown field, and no free parameter.

Phenomenon 115.40 (Intergalactic space is opaque to the highest-energy protons)**.** A proton propagating through the 2.725 48 K microwave background of Phenomenon 51.3 loses energy by photopion production, $p + \gamma \to \Delta^{+} \to p + \pi^{0}$ or $n + \pi^{+}$, once its energy exceeds a threshold of order $1 \times 10^{20}\,\mathrm{eV}$ against a typical background photon and a few times $1 \times 10^{19}\,\mathrm{eV}$ against the high-energy tail of the bath. The energy-loss length above the threshold is of order tens of megaparsecs, far smaller than the Hubble radius, so the sources of any observed particle above the threshold must lie within that distance [1330, 1339, 1385].

Rests on Phenomenon 51.3 and Equation (40.12).

Derivation of the threshold. *Derives Phenomenon 115.40.*

Let a proton of energy E and momentum $p = \sqrt{E^2/c^2 - m_p^2 c^2}$ meet a photon of energy E_γ head-on. The invariant mass of the pair is given by Equation (40.12):

$$s = (p_p + p_\gamma)^2\, c^2 = m_p^2 c^4 + 2\,(E E_\gamma + p c E_\gamma) \longrightarrow m_p^2 c^4 + 4 E E_\gamma\,, \tag{115.55}$$

the limit being the ultrarelativistic one, $pc \to E$; for a general collision angle θ the last term is $2EE_\gamma\,(1 - \cos\theta)$, maximal head-on. Photopion production $p + \gamma \to N + \pi$ is kinematically possible once $\sqrt{s} \geq (m_N + m_\pi)\,c^2$, and the lowest such threshold is the $p\pi^0$ channel. Solving Equation (115.55),

$$E \geq E_{\mathrm{th}} = \frac{m_\pi c^2\,(2 m_p c^2 + m_\pi c^2)}{4 E_\gamma}\,. \tag{115.56}$$

The photon bath is measured, not assumed. Its temperature is $T_0 = (2.725\,48 \pm 0.000\,57)\,\mathrm{K}$ [2144], so

$$k_B T_0 = 1.380\,649 \times 10^{-23}\,\mathrm{J/K} \times 2.725\,48\,\mathrm{K} = 3.7629 \times 10^{-23}\,\mathrm{J} = 2.348\,69 \times 10^{-4}\,\mathrm{eV}\,,$$

and the mean energy of a Planck photon is

$$\langle E_\gamma \rangle = \frac{\pi^4}{30\zeta(3)} k_B T_0 = 2.70118\, k_B T_0 = 6.344 \times 10^{-4}\,\mathrm{eV} = 1.0164 \times 10^{-22}\,\mathrm{J}\,, \tag{115.57}$$

while their number density is

$$n_\gamma = 16\pi\zeta(3) \left(\frac{k_B T_0}{hc}\right)^3 = 4.107 \times 10^8/\mathrm{m}^3 = 411/\mathrm{cm}^3\,. \tag{115.58}$$

With $m_p c^2 = 938.272\,\mathrm{MeV}$ and $m_{\pi^0} c^2 = 134.977\,\mathrm{MeV}$ [2427], Equation (115.56) gives

$$E_{\mathrm{th}} = \frac{1.349\,77 \times 10^8\,\mathrm{eV} \times 2.011\,52 \times 10^9\,\mathrm{eV}}{4 \times 6.344 \times 10^{-4}\,\mathrm{eV}} = 1.07 \times 10^{20}\,\mathrm{eV} = 17.1\,\mathrm{J}\,,$$

and the same calculation with the $\Delta(1232)$ mass in place of $(m_p + m_\pi)$ puts the resonance *peak* at $2.5 \times 10^{20}\,\mathrm{eV}$. □

Derivation of the onset against the Planck tail. A threshold computed with the mean photon energy is not the energy at which the effect begins, because the Planck distribution has a tail. At proton energy E the reaction proceeds on any photon with $E_\gamma \geq \varepsilon_{\mathrm{th}}(E)$, where by Equation (115.56) $\varepsilon_{\mathrm{th}} = m_\pi c^2 \left(2 m_p c^2 + m_\pi c^2\right)/(4E)$. Writing $x = \varepsilon_{\mathrm{th}}/\left(k_B T_0\right)$, the fraction of the bath above that energy is

$$\frac{n_\gamma(>x)}{n_\gamma} = \frac{1}{2\zeta(3)} \int_x^\infty \frac{u^2\,\mathrm{d}u}{\mathrm{e}^u - 1} \approx \frac{\mathrm{e}^{-x}\left(x^2 + 2x + 2\right)}{2\zeta(3)}\,, \tag{115.59}$$

the approximation being the leading term of the expansion of $(\mathrm{e}^u - 1)^{-1}$, excellent for $x \gtrsim 3$. At $E = 5 \times 10^{19}\,\mathrm{eV}$, $\varepsilon_{\mathrm{th}} = 1.36 \times 10^{-3}\,\mathrm{eV}$ and $x = 5.78$, so Equation (115.59) gives 6.0 %: six percent of the bath, that is $2.5 \times 10^7/\mathrm{m}^3$ photons, are already above threshold. At $E = 1 \times 10^{20}\,\mathrm{eV}$, $x = 2.89$ and the fraction is 37 %. The onset of the suppression is therefore expected in the range $4 \times 10^{19}\,\mathrm{eV}$ to $1 \times 10^{20}\,\mathrm{eV}$, and not at the nominal $1.07 \times 10^{20}\,\mathrm{eV}$ — which is where the observed break in Phenomenon 115.42 lies. □

Order-of-magnitude derivation of the horizon. The mean free path is $\ell = 1/\left(n_{\mathrm{eff}}\sigma\right)$ with n_{eff} the density of above-threshold photons and σ the photopion cross section, which near the Δ peak is about $5 \times 10^{-32}\,\mathrm{m}^2$, that is 500 microbarn. At $E = 1 \times 10^{20}\,\mathrm{eV}$ the fraction above threshold is 37 %, so $n_{\mathrm{eff}} = 1.5 \times 10^8/\mathrm{m}^3$ and

$$\ell = \frac{1}{1.5 \times 10^8/\mathrm{m}^3 \times 5 \times 10^{-32}\,\mathrm{m}^2} = 1.3 \times 10^{23}\,\mathrm{m} = 4\,\mathrm{Mpc}\,.$$

Each interaction removes only the fraction $\kappa \approx 0.2$ of the proton's energy — the inelasticity, fixed by the kinematics of Equation (115.17) applied to the Δ decay — so the *energy-loss* length is $\ell/\kappa \approx 20\,\mathrm{Mpc}$. This estimate is crude in two ways that both push it upward: it uses the resonance-peak cross section for every above-threshold photon, and it treats every collision as head-on rather than averaging over angle. The proper calculation, integrating the measured cross section over the Planck spectrum and over angle, gives an energy-loss length of order 100 Mpc at $1 \times 10^{20}\,\mathrm{eV}$, falling to a few tens of megaparsecs above it [1385].

Either number is small in the only comparison that matters. The Hubble radius is

$$\frac{c}{H_0} = \frac{2.998 \times 10^5\,\mathrm{km/s}}{67.4\,\mathrm{km/s/Mpc}} = 4.4 \times 10^3\,\mathrm{Mpc}\,,$$

with H_0 from [2348]. A particle above the threshold therefore reaches us from at most one or two percent of the radius of the observable universe: the sky at these energies is a local sky, and if a source is identified it will be a nearby one. □

Derivation pending #207

The photopion energy-loss length as a proper integral of the measured $p\gamma$ cross section over the Planck spectrum and over collision angle, replacing the order-of-magnitude estimate given here. Belongs in Appendix A

Remark 115.41 (Nuclei are attenuated too, but differently). For a nucleus the relevant process is not photopion production but photodisintegration through the giant dipole resonance, which sits at a photon energy of about 20 MeV *in the nucleus rest frame*. A photon of laboratory energy E_γ appears in that frame boosted by up to 2γ, so the threshold is

$$\gamma \gtrsim \frac{20\,\mathrm{MeV}}{2 \times 6.34 \times 10^{-4}\,\mathrm{eV}} = 1.6 \times 10^{10}\,,$$

that is 1.5×10^{19} eV per nucleon, and hence 8×10^{20} eV for an iron nucleus of $A = 56$. A heavy primary therefore survives to a *higher total energy* than a proton, and its attenuation, when it sets in, proceeds by shedding nucleons rather than by losing a fifth of its energy at a stroke. The shape of the predicted suppression consequently depends on the composition, which is exactly the loophole Section 115.4.5 opened and Phenomenon 115.42 must respect.

115.6.2 The observed suppression

Phenomenon 115.42 (Suppression at the highest energies)**.** The flux falls below the extrapolation of Equation (115.20) above a few times 1×10^{19} eV: a break at about 5.6×10^{19} eV with a significance near 5σ [2115], and a suppression above about 4×10^{19} eV at more than 6σ [2116], measured by two observatories with different techniques. Individual events well above the break exist, including one of $(3.2 \pm 0.9) \times 10^{20}$ eV [1895].

Rests on Equations (115.20) *and* (115.59) *and Phenomenon 115.40.*

Derivation and its limits. *Derives Phenomenon 115.42.*

The predicted position is Equation (115.56) softened by the Planck tail, Equation (115.59): an onset between 4×10^{19} eV and 1×10^{20} eV. The two measurements bracket that range. The HiRes fluorescence detectors found the spectrum to break at 5.6×10^{19} eV $= 9.0$ J, the break being significant at about five standard deviations against a continued power law [2115]; the Pierre Auger Observatory, with a larger exposure and the hybrid calibration of Section 115.4.3, found the flux above 4×10^{19} eV $= 6.4$ J to be suppressed relative to the extrapolation at more than six standard deviations [2116]. The two techniques share no systematic, which is why the joint statement is stronger than either.

That events exist above the break is not a contradiction; it is required. Attenuation is exponential in distance, not a wall, so a source within 20 Mpc can deliver a particle of any energy its accelerator reaches. The Fly's Eye event of $(3.2 \pm 0.9) \times 10^{20}$ eV $= (51 \pm 14)$ J [1895] is such a particle, and its existence constrains its own source to lie within the horizon of Phenomenon 115.40 — which is a strong statement, since no candidate accelerator is known within that distance.

The inference is nevertheless not closed, and the reason is Remark 115.41 together with Phenomenon 115.25. If the primaries at these energies are nuclei rather than protons, the same break is produced, equally well, by the maximum rigidity of the sources through Equation (115.22): an accelerator limited to $R_{\max} \approx 5 \times 10^{18}$ V delivers protons to 5×10^{18} eV and iron to 1.3×10^{20} eV, and the superposition of the elemental cutoffs mimics a propagation suppression. The measurement therefore establishes that the spectrum is suppressed; it does not by itself establish that the mechanism is the one Greisen, Zatsepin and Kuz'min predicted. Distinguishing them requires the composition above the break, which is where the current data are weakest, and the appearance of the cosmogenic neutrinos that photopion production must also produce, which have not been detected. □

Remark 115.43 (The AGASA disagreement, and what settled it). For a decade the field held two incompatible measurements: AGASA, a scintillator array in Japan, reported a spectrum *continuing* unbroken past 1×10^{20} eV with eleven events there, while HiRes reported the break. Both could not be right. The disagreement was resolved not by a new argument but by the energy-scale analysis of Proposition 115.26: a scintillator array infers energy from a simulation of the shower's particle content at the ground, a fluorescence detector measures it calorimetrically by Equation (115.35), and the two scales differed by some 20 %. By Equation (115.40) with $\gamma \approx 4.5$, a 20 % scale difference moves the flux by 70 % and moves events across a threshold; the Auger hybrid design, which measures both ways on the same showers, was built to remove exactly this ambiguity, and it did.

115.6.3 Anisotropy at the highest energies

Phenomenon 115.44 (A large-scale dipole above 8×10^{18} electronvolts)**.** The arrival directions of cosmic rays above 8×10^{18} eV are not isotropic: they carry a dipole of amplitude about 6.5 % whose direction lies some 125° away from the Galactic centre, with a chance probability below 10^{-6} [2276]. A dipole pointing away from the Galactic centre is not what a Galactic source population produces; the origin of these particles is extragalactic.

Rests on Theorem 11.46 and Definition 11.80.

Derivation: why a few percent is a detection. *Derives Phenomenon 115.44.*

Let N events arrive with right ascensions φ_i, and form the first-harmonic amplitudes

$$a = \frac{2}{N}\sum_{i=1}^{N}\cos\varphi_i\,, \qquad b = \frac{2}{N}\sum_{i=1}^{N}\sin\varphi_i\,, \qquad r = \sqrt{a^2+b^2}\,.$$

Under isotropy each φ_i is uniform, so $\sum\cos\varphi_i$ and $\sum\sin\varphi_i$ are sums of N independent zero-mean terms of variance $\frac{1}{2}$ each; by the central limit theorem (Theorem 11.46) they are Gaussian of variance $N/2$, and independent. Then $Na^2/2$ and $Nb^2/2$ are independent χ^2 variables of one degree of freedom (Definition 11.80), so $Nr^2/2$ is χ^2 with two, i.e. exponential, and

$$P\,(>r) = \exp\!\left(-\frac{Nr^2}{4}\right)\,. \tag{115.60}$$

This is the Rayleigh probability. With the $N \approx 3.2 \times 10^4$ events recorded above 8×10^{18} eV and $r = 0.065$, Equation (115.60) gives $\exp(-33.8)$, a number so small that it is not the quoted result: the published chance probability is far more conservative, because a three-dimensional dipole must be reconstructed from an exposure that does not cover the whole sky, and the declination component is measured much less well than the right-ascension component. Equation (115.60) shows *why* a percent-level anisotropy is measurable at all with tens of thousands of events — the sensitivity goes as $N^{-1/2}$, so 3×10^4 events probe amplitudes near 1 % — and the published 10^{-6} is the honest number after the exposure is accounted for [2276]. □

Remark 115.45 (What survives magnetic deflection). It is not an accident that a dipole and not a set of point sources is what is seen. By Equation (115.23) a proton of 8×10^{18} eV crossing the Galaxy in a 3×10^{-10} T field has a gyroradius of 2.9 kpc, and the deflection over a path of a few kiloparsecs is therefore tens of degrees; for an iron nucleus, $Z = 26$, it is larger by that factor and the direction is lost entirely. The extragalactic fields add more, and are less well known. The result is that a genuinely anisotropic extragalactic source distribution is smeared into precisely the lowest multipole that survives — a dipole. The absence of a confirmed point source is therefore expected, not anomalous, and no honest reading of the present data identifies one (Remark 115.38).

115.7 Neutral messengers

115.7.1 Gamma-ray astronomy

A photon travels in a straight line. Above the atmosphere, a gamma ray is detected by letting it convert to an e^+e^- pair in a thin foil, tracking the pair to reconstruct the incoming direction, and stopping it in a calorimeter to measure the energy. That is the Large Area Telescope on the Fermi observatory: a tungsten-and-silicon pair-conversion tracker over a caesium-iodide calorimeter, covering 20 MeV to above 300 GeV with a field of view of 2.4 sr, so that it surveys the entire sky every two orbits [2141]. Above about 100 GeV the flux is too small for any instrument that can be flown, and the ground-based Cherenkov technique of Section 115.4.4 takes over, using the atmosphere as the converter and the calorimeter both.

Two results define the current reach. The MAGIC telescopes detected teraelectronvolt emission from the gamma-ray burst GRB 190114C, establishing that the most violent transients accelerate particles to those energies [2326]. And LHAASO, a kilometre-scale array on the Tibetan plateau, detected photons up to $1.4\,\mathrm{PeV} = 2.24 \times 10^{-4}\,\mathrm{J}$ from twelve sources within the Galaxy [2376].

Phenomenon 115.46 (PeV accelerators exist in the Galaxy)**.** Photons of energy up to 1.4 PeV arrive from twelve identified Galactic sources [2376]. Whatever population produces them accelerates parent particles to at least that energy, and in the hadronic case to more than ten times it — which places accelerators reaching the knee of Phenomenon 115.13 inside the Galaxy.

Rests on Phenomenon 115.13 and Equation (115.57).

Derivation of the parent energy, and of the ambiguity. Derives Phenomenon 115.46.

Suppose first the emission is hadronic. A proton of energy E_p striking interstellar gas produces pions carrying, on average, some 20 % of its energy each, and a π^0 decays to two photons sharing its energy, so

$$E_\gamma \approx \tfrac{1}{2} \times 0.2\, E_p = 0.1\, E_p\,,$$

and a 1.4 PeV photon requires a proton of about 1.4×10^{16} eV — above the knee. Suppose instead the emission is leptonic, by inverse-Compton scattering of electrons on the microwave background. In the Thomson regime the scattered photon energy is $E_\gamma \approx \frac{4}{3}\gamma_e^2 \left\langle E_\gamma^{\mathrm{CMB}} \right\rangle$, which for $E_\gamma = 1.4\,\mathrm{PeV}$ and Equation (115.57) would need $\gamma_e = 1.3 \times 10^9$, i.e. $E_e = 0.66\,\mathrm{PeV}$. But at that Lorentz factor the scattering is deep in the Klein–Nishina regime, since $\gamma_e \left\langle E_\gamma^{\mathrm{CMB}} \right\rangle = 0.8\,\mathrm{MeV}$ exceeds $m_e c^2 = 0.511\,\mathrm{MeV}$; there the electron transfers nearly all its energy in one scattering and the cross section is suppressed, so the required electron energy rises to $E_e \gtrsim 1.4\,\mathrm{PeV}$ and the emission becomes inefficient.

Either way the source accelerates particles to at least a petaelectronvolt. Which species it accelerates the gamma-ray spectrum alone cannot say, because both mechanisms produce power-law photon spectra of similar index over a limited band. This is the standing *hadronic–leptonic ambiguity*, and it is why the detection of neutrinos from a source would be decisive: Section 115.7.4 shows that neutrinos at these energies have no leptonic production channel at all. □

115.7.2 Atmospheric neutrinos

The showers of Section 115.4.2 are prolific neutrino sources. Every charged pion below the critical energy E_c^π decays, $\pi^+ \to \mu^+ \nu_\mu$, and every muon that decays before reaching the ground gives $\mu^+ \to e^+ \nu_e \bar{\nu}_\mu$. Counting neutrinos in the full chain,

$$\frac{\nu_\mu + \bar{\nu}_\mu}{\nu_e + \bar{\nu}_e} \longrightarrow 2\,, \tag{115.61}$$

two muon-flavour neutrinos for every electron-flavour one, a prediction that involves nothing but the decay chain and is therefore reliable to a few percent at low energy. It fails at high energy for a reason already derived: a muon of energy E has a laboratory decay length $\gamma c\tau_\mu = (E/m_\mu c^2) \times 659\,\mathrm{m}$, which reaches the 15 km of overlying atmosphere at $E = 2.4\,\mathrm{GeV}$; above a few GeV most muons reach the ground and are absorbed before decaying, so the electron-flavour component is suppressed and the ratio rises above two.

Phenomenon 115.47 (Upward-going muon neutrinos are missing)**.** A large underground water-Cherenkov detector records fewer muon-flavour atmospheric neutrino interactions from below — from neutrinos that have crossed the Earth — than from above, while the electron-flavour rate shows no such zenith dependence. The deficit depends on the ratio of path length to neutrino energy in the manner required by flavour oscillation and in no other [1954].

Rests on Equation (115.61).

Derivation of the oscillation phase in SI units. *Derives Phenomenon 115.47.*

Take two neutrino mass eigenstates of masses m_1, m_2 and write the flavour states as $|\nu_\alpha\rangle = \cos\theta\,|\nu_1\rangle + \sin\theta\,|\nu_2\rangle$, $|\nu_\beta\rangle = -\sin\theta\,|\nu_1\rangle + \cos\theta\,|\nu_2\rangle$; the unitarity of the mixing matrix is licensed by Chapter 5 and the physics of the mixing itself belongs to Chapter 103. A state created with definite flavour and definite momentum p evolves with the phase $\exp(-\mathrm{i}E_i t/\hbar)$, where

$$E_i = \sqrt{p^2c^2 + m_i^2c^4} \approx pc + \frac{m_i^2 c^4}{2pc}\,,$$

the expansion being valid because $m_i c^2 \ll pc$ for every neutrino ever detected. After a flight time $t = L/c$ over a baseline L, the accumulated phase difference is

$$\Delta\Phi = \frac{(E_2 - E_1)\,t}{\hbar} = \frac{(m_2^2 - m_1^2)\,c^4}{2pc} \cdot \frac{L}{\hbar c} = \frac{\Delta m^2 c^4\, L}{2\hbar c E}\,, \tag{115.62}$$

writing $E \approx pc$ for the neutrino energy and $\Delta m^2 := m_2^2 - m_1^2$. Every symbol is SI: $\hbar c$ has dimension J m, $\Delta m^2 c^4$ has dimension J^2, and L/E has dimension m/J, so $\Delta\Phi$ is a pure number as it must be.

The transition probability follows from the amplitude $A_{\alpha\alpha} = \cos^2\theta\,\mathrm{e}^{-\mathrm{i}\Phi_1} + \sin^2\theta\,\mathrm{e}^{-\mathrm{i}\Phi_2}$:

$$|A_{\alpha\alpha}|^2 = 1 - 2\cos^2\theta\sin^2\theta\,(1 - \cos\Delta\Phi) = 1 - \sin^2 2\theta\,\sin^2\frac{\Delta\Phi}{2}\,,$$

so that, defining

$$\Phi := \frac{\Delta\Phi}{2} = \frac{\Delta m^2 c^4\, L}{4\hbar c E}\,, \tag{115.63}$$

the survival and appearance probabilities are

$$P(\nu_\alpha \to \nu_\beta) = \sin^2 2\theta\,\sin^2\Phi\,, \qquad P(\nu_\alpha \to \nu_\alpha) = 1 - \sin^2 2\theta\,\sin^2\Phi\,. \tag{115.64}$$

Equation (115.63) is the formula the literature quotes as $1.267\,\Delta m^2 L/E$, and the number is derived and not conventional. Put $\Delta m^2 c^4$ in eV^2, L in kilometres and E in GeV, and use $\hbar c = 1.973\,269\,8 \times 10^{-7}\,\mathrm{eV\,m}$:

$$\Phi = \frac{\left(\Delta m^2 c^4/\mathrm{eV}^2\right) \times (L/\mathrm{km}) \times 1 \times 10^3\,\mathrm{m}}{4 \times 1.973\,269\,8 \times 10^{-7}\,\mathrm{eV\,m} \times (E/\mathrm{GeV}) \times 1 \times 10^9\,\mathrm{eV}} = 1.2669\,\frac{\Delta m^2 c^4\,\left[\mathrm{eV}^2\right]\,L\,[\mathrm{km}]}{E\,[\mathrm{GeV}]}\,.$$

□

Derivation of the zenith-angle deficit. Atmospheric neutrinos are produced at an altitude of about 15 km over the whole Earth. One arriving from directly overhead has travelled $L \approx 15\,\mathrm{km}$; one arriving from directly below has crossed a diameter, $L \approx 2R_\oplus = 1.27\times10^4\,\mathrm{km}$. The path length varies by three orders of magnitude with zenith angle, and the neutrino energy does not.

Take the atmospheric mass splitting $\Delta m^2c^4 \approx 2.5\times10^{-3}\,\mathrm{eV}^2$ and $E = 1\,\mathrm{GeV}$. For the downward-going neutrino,

$$\Phi = 1.267 \times 2.5 \times 10^{-3} \times 15/1 = 0.048\,, \qquad \sin^2\Phi = 2.3\times10^{-3}\,,$$

so no oscillation is observable. For the upward-going one,

$$\Phi = 1.267 \times 2.5 \times 10^{-3} \times 1.27 \times 10^4/1 = 40\,,$$

which is many complete oscillations; averaged over the spread in L and E within any experimental bin, $\langle\sin^2\Phi\rangle = \frac{1}{2}$ and

$$P\,(\nu_\mu \to \nu_\mu) = 1 - \tfrac{1}{2}\sin^2 2\theta\,,$$

which for maximal mixing is one half. The prediction is therefore specific and asymmetric: the upward-going muon-neutrino rate is halved while the downward-going rate is untouched, and the electron-flavour rate is unaffected at either zenith angle because the corresponding splitting is thirty times smaller and the phase is correspondingly negligible over these baselines. That is exactly the pattern Super-Kamiokande measured [1954], and it is the first compelling evidence that neutrinos have mass. □

The apparatus, the data and the uncertainty budget of that measurement belong to Chapter 114, and the theory of mixing — three flavours, the PMNS matrix, matter effects — to Chapter 103; this chapter supplies the source. It is worth recording that the *solar* neutrino deficit, measured by Davis's chlorine experiment from 1968 [1368, 1950] and resolved by the SNO measurement of the neutral-current rate [2021], is the independent half of the same story; solar neutrinos are not cosmic rays and are not treated here, but the two deficits are explained by one mechanism and neither alone would have been believed.

For everything that follows, the atmospheric flux changes role: it stops being the signal and becomes the irreducible background to Section 115.7.4.

115.7.3 Supernova 1987A

On 23 February 1987 a star exploded in the Large Magellanic Cloud, at a distance of about $50\,\mathrm{kpc} = 1.54\times10^{21}\,\mathrm{m}$. Some three hours before the optical brightening, two underground water-Cherenkov detectors built to look for proton decay recorded bursts of events: eleven in Kamiokande-II over 13 s [1793] and eight in the IMB detector over 6 s [1789], with energies between about 7 MeV and 40 MeV. Nineteen events remain, four decades later, the entire observational sample of extragalactic neutrino astronomy below the teraelectronvolt scale, and they carry a remarkable amount of physics.

Phenomenon 115.48 (The neutrino burst of SN 1987A)**.** A core-collapse supernova at 50 kpc produced a neutrino burst of duration of order 10 s carrying a total energy of order $3\times10^{46}\,\mathrm{J}$, detected in coincidence at two independent detectors on opposite sides of the Earth [1789, 1793]. The energy matches the gravitational binding energy of a neutron star, and the arrival timing bounds both the neutrino speed and the neutrino mass.

Derivation: the energy is the binding energy of a neutron star. *Derives Phenomenon 115.48.*

If a stellar core of mass M collapses to a uniform sphere of radius R, the gravitational energy released is

$$|U| = \frac{3}{5}\frac{GM^2}{R}\,,$$

the standard self-energy integral. With $M = 1.4 M_\odot = 2.78 \times 10^{30}\,\mathrm{kg}$ and $R = 10\,\mathrm{km}$,

$$|U| = \frac{3}{5} \times \frac{6.674 \times 10^{-11}\,\mathrm{N\,m^2/kg^2} \times (2.78 \times 10^{30}\,\mathrm{kg})^2}{1.0 \times 10^4\,\mathrm{m}} = 3.1 \times 10^{46}\,\mathrm{J}\,,$$

which is 12 % of the rest energy $Mc^2 = 2.50 \times 10^{47}\,\mathrm{J}$ of the collapsing core. Essentially none of it emerges as light: the total optical output of the supernova is smaller by three orders of magnitude, and the kinetic energy of the ejecta by two — the $1 \times 10^{44}\,\mathrm{J}$ used in Phenomenon 115.29. It emerges as neutrinos, because at the densities of the collapsing core nothing else escapes. The total energy inferred from the nineteen detected events, given the detector masses, the cross section for $\bar{\nu}_e p \to n e^+$ and the assumption of equipartition among the six neutrino species, is of order $3 \times 10^{46}\,\mathrm{J}$. The agreement is the confirmation, and it is the only direct observation ever made of the mechanism of a core-collapse supernova. □

Derivation: the bound on the neutrino speed. The neutrinos preceded the first optical light by about $\Delta t = 3\,\mathrm{h} = 1.1 \times 10^4\,\mathrm{s}$, and that lead is astrophysical — the shock takes hours to reach the stellar surface after the core collapses, whereas neutrinos leave immediately. Taking the lead as an upper bound on any difference in travel time over the flight time

$$\frac{D}{c} = \frac{1.54 \times 10^{21}\,\mathrm{m}}{2.998 \times 10^8\,\mathrm{m/s}} = 5.15 \times 10^{12}\,\mathrm{s} = 1.63 \times 10^5\,\mathrm{yr}\,,$$

gives

$$\frac{|v_\nu - c|}{c} \le \frac{\Delta t}{D/c} = \frac{1.1 \times 10^4\,\mathrm{s}}{5.15 \times 10^{12}\,\mathrm{s}} = 2.1 \times 10^{-9}\,. \tag{115.65}$$

The bound is weak as a fractional velocity but it is obtained over a baseline of a hundred and sixty thousand light years, which no terrestrial measurement can approach, and it is the reason a claim of superluminal neutrino propagation is checked against SN 1987A first. □

Derivation: the bound on the neutrino mass. A massive neutrino is dispersive. From $E = \sqrt{p^2c^2 + m^2c^4}$ the group velocity is

$$\frac{v}{c} = \frac{pc^2}{Ec} = \sqrt{1 - \frac{m^2c^4}{E^2}} \approx 1 - \frac{m^2c^4}{2E^2}\,,$$

so a neutrino of energy E arrives later than a massless one by

$$\delta t(E) = \frac{D}{c}\frac{m^2c^4}{2E^2}\,. \tag{115.66}$$

Two neutrinos of energies $E_1 < E_2$ emitted simultaneously therefore arrive separated by

$$\Delta t_{\mathrm{disp}} = \frac{D}{2c}\, m^2c^4 \left(\frac{1}{E_1^2} - \frac{1}{E_2^2}\right)\,.$$

The observed burst spanned $\Delta t \approx 13\,\mathrm{s}$ with energies from about 7.5 MeV to 35 MeV. Requiring the dispersion not to exceed the observed spread,

$$m^2c^4 \le \frac{2c\,\Delta t}{D}\left(\frac{1}{E_1^2} - \frac{1}{E_2^2}\right)^{-1} = \frac{2 \times 13\,\mathrm{s}}{5.15 \times 10^{12}\,\mathrm{s}} \times \frac{1}{1.70 \times 10^{-14}/\mathrm{eV}^2} = 300\,\mathrm{eV}^2\,,$$

whence

$$m_\nu c^2 \lesssim 17\,\mathrm{eV} = 2.7 \times 10^{-18}\,\mathrm{J}\,, \qquad m_\nu \lesssim 3.0 \times 10^{-35}\,\mathrm{kg}\,. \tag{115.67}$$

The bound is honest only under an assumption that cannot be checked: that the burst was emitted in a time short compared with 13 s. Since the emission is expected to last of order that long by itself, the true limit is weaker, and careful analyses that model the emission profile obtain a few electronvolts rather than seventeen. It is quoted here because it is the only kinematic neutrino mass bound ever obtained over an astronomical baseline, and because the method — dispersion over a known distance — is the one used again for photons in Section 115.8.2. □

Remark 115.49 (Where the neutrino mass is actually measured). Equation (115.67) has been superseded twice over. The kinematic endpoint of tritium beta decay gives $m_\nu c^2 < 0.8\,\mathrm{eV}$ at 90 % confidence [2399], and the cosmological sum of neutrino masses is bounded well below that [2348]. Oscillation experiments — Equation (115.64) — measure only Δm^2 and say nothing about the absolute scale, which is why the direct bounds remain necessary. SN 1987A's value today is not its number but its method, and the nineteen events remain the only extragalactic neutrino source ever identified below the teraelectronvolt scale.

115.7.4 High-energy astrophysical neutrinos

A neutrino is neutral, so it points; it is weakly interacting, so it escapes from regions opaque to photons and is not attenuated by the microwave background. The same weakness is the difficulty: a kilometre-scale target is needed. IceCube is a cubic kilometre of Antarctic ice — $1 \times 10^9\,\mathrm{m}^3$, a mass of $9.2 \times 10^{11}\,\mathrm{kg}$ — instrumented with 5160 digital optical modules on 86 vertical strings between 1450 m and 2450 m depth, each module a photomultiplier reading the Cherenkov light of Example 115.24 from charged secondaries. The detector concept, and the performance of the nine-string configuration with which it began operating, are described in [2088]; the figures above are those of the completed array.

Phenomenon 115.50 (An extraterrestrial high-energy neutrino flux)**.** A cubic kilometre of instrumented Antarctic ice records neutrino events in excess of the atmospheric background: 28 events with deposited energies between 30 TeV and 1.2 PeV against an expectation of about 10, a 4σ excess [2203]. Neutrinos at these energies are produced only in hadronic collisions, so their detection is direct evidence that protons and nuclei, and not only electrons, are accelerated to such energies somewhere in the universe.

Rests on Equation (115.47).

Derivation: why the background falls away above 100 *teraelectronvolts.* *Derives Phenomenon 115.50.*

The atmospheric neutrino flux of Section 115.7.2 is not simply a copy of the primary spectrum. A pion of energy above the critical energy $E_c^\pi \approx 24\,\mathrm{GeV}$ derived in Section 115.4.2, conventionally rounded to 20 GeV, interacts before it decays, and the probability that it decays instead is the ratio of the two lengths,

$$P_{\mathrm{dec}} \approx \frac{\lambda_I/\rho}{\gamma c \tau_\pi} = \frac{E_c^\pi}{E_\pi}\,,$$

falling as $1/E$. The neutrino flux is therefore the parent spectrum $E^{-2.7}$ multiplied by E^{-1}:

$$J_\nu^{\mathrm{atm}}(E) \propto E^{-3.7}\,, \tag{115.68}$$

one power steeper than the cosmic-ray spectrum that makes it. An astrophysical flux produced by shock acceleration, by contrast, carries the source index of Equation (115.47), near E^{-2} and certainly harder than E^{-3}, and suffers no such suppression because its pions

decay in vacuum. Two power laws differing by more than one unit of index must cross, and they cross near 100 TeV: below it the sky is atmospheric, above it the atmospheric flux has fallen away and an astrophysical component, if it exists, dominates. That is why the IceCube analysis selects events above 30 TeV and why the excess is at the high-energy end of the selection.

A second discriminant is available and is used. An atmospheric neutrino is made in an air shower, and the same shower makes muons that arrive at the detector at the same time. Selecting only events whose interaction vertex is inside the instrumented volume, with no entering track, therefore vetoes atmospheric neutrinos arriving from above against their own accompanying muons — the *self-veto* — while an astrophysical neutrino, which has no accompanying shower, passes. The veto is what reduces the expected background of the Phenomenon 115.50 sample to about ten events. □

Derivation: a neutrino detection is evidence of hadronic acceleration. Neutrinos at these energies have exactly one production channel: the weak decay of pions and kaons, which are made only in hadronic collisions. Electrons produce photons — by synchrotron radiation, by bremsstrahlung, by inverse-Compton scattering — and produce no neutrinos at all, because none of those processes has a weak vertex. Detecting a teraelectronvolt neutrino from a direction therefore establishes that protons or nuclei are accelerated there, and resolves the hadronic–leptonic ambiguity of Phenomenon 115.46 in one measurement.

The same chain fixes the expected ratio of neutrinos to gamma rays. In a pp collision the three pion charge states are produced in roughly equal numbers. The π^0 gives two photons carrying all its energy; each $\pi^\pm$ gives, through $\pi^+ \to \mu^+\nu_\mu$ and $\mu^+ \to e^+\nu_e\bar{\nu}_\mu$, three neutrinos carrying about three quarters of its energy. Hence

$$\frac{L_\nu}{L_\gamma} \approx \frac{2 \times \frac{3}{4}}{1} = \frac{3}{2}\,, \tag{115.69}$$

a source's all-flavour neutrino luminosity is comparable to its hadronic gamma-ray luminosity, which is what makes a joint detection feasible at all. The flavour composition at the source is $\nu_e : \nu_\mu : \nu_\tau = 1 : 2 : 0$ from the same counting, and oscillation over astronomical baselines — where Equation (115.63) gives an enormous Φ and every $\sin^2\Phi$ averages to one half — converts this to approximately $1 : 1 : 1$ at Earth. The measured flavour ratio is consistent with that, which is a non-trivial check that the sources are pion sources. □

115.7.5 Multimessenger astronomy

On 22 September 2017 IceCube recorded a muon-neutrino event of reconstructed energy $290\,\text{TeV} = 4.6 \times 10^{-5}\,\text{J}$ and issued an automated alert within a minute. Follow-up by some twenty instruments from radio to gamma rays found the blazar TXS 0506+056 — an active galaxy whose jet points nearly at us — in a gamma-ray flaring state, coincident in direction with the neutrino. The chance-coincidence probability, allowing for the number of blazars and the number of alerts, is at the level of 3σ [2306]. A search of the archival IceCube data then found, independently, an excess of 13 ± 5 events from the same direction during 2014–2015 — a period in which the source was not flaring in gamma rays — at 3.5σ [2293].

Remark 115.51 (Two three-sigma results are evidence, not a discovery). The two observations are statistically independent: different years, different event samples, different analyses, one a directional coincidence with a flare and the other a time-integrated point-source search. Taken together they are the strongest candidate identification of a source of high-energy cosmic rays that exists. Taken separately neither reaches the threshold this book would require to call something observed, and the combination of two 3σ results is not a 6σ result — significances add in quadrature only for the statistic, not for the trials factor of Section 11.6, and the archival search was made in the direction the first result selected, which is precisely

the situation the look-elsewhere correction exists to handle. The honest statement is that TXS 0506+056 is a candidate, that no other source has been identified since, and that the diffuse flux of Phenomenon 115.50 remains unresolved into sources.

The method, however, is established, and its clearest success is elsewhere: the coincidence of a gravitational-wave signal with a short gamma-ray burst, treated in Chapter 47, in which two independent messengers from one event were combined to identify the source and to bound the propagation speed of gravity. Multimessenger astronomy is that method, and cosmic rays supply two of its four channels.

115.8 Cosmic rays as a physics laboratory

115.8.1 Centre-of-mass energies beyond the colliders

Proposition 115.52 (The natural beam exceeds the collider by a factor of thirty)**.** *A proton of* $1 \times 10^{20}\,\mathrm{eV}$ *striking a nucleon at rest in the atmosphere collides at a centre-of-mass energy* $\sqrt{s} = 433\,\mathrm{TeV}$*, some thirty-two times the* $13.6\,\mathrm{TeV}$ *at which the Large Hadron Collider now runs and fifty-four times the* $8\,\mathrm{TeV}$ *of the Higgs discovery datasets of Chapter 113.*

Rests on Equation (40.12)*.*

Proof. *Derives Proposition 115.52.*

By Equation (40.12) for a projectile of energy E on a target of mass m_p at rest,

$$s = 2Em_pc^2 + 2m_p^2c^4 \approx 2Em_pc^2\,,$$

so

$$\sqrt{s} = \sqrt{2 \times 1 \times 10^{20}\,\mathrm{eV} \times 9.382\,72 \times 10^{8}\,\mathrm{eV}} = 4.33 \times 10^{14}\,\mathrm{eV} = 433\,\mathrm{TeV}\,,$$

in SI $6.9 \times 10^{-5}\,\mathrm{J}$. The square root is the whole story: a fixed-target collision converts energy to centre-of-mass energy only as $\sqrt{E}$, which is why colliders were built (Example 40.13), and why even a beam seven orders of magnitude above the LHC's buys only a factor of thirty in $\sqrt{s}$. □

That factor is nevertheless real, and two measurements exploit it.

The first is the proton–air cross section, extracted from the *tail* of the $X_{\max}$ distribution. A shower whose first interaction happens deep in the atmosphere has a deep maximum, and since the first-interaction depth is exponentially distributed with mean $\lambda_I = \bar{m}_{\mathrm{air}}/\sigma_{p\text{-air}}$, the deep tail of the measured $X_{\max}$ distribution falls exponentially with a slope that measures λ_I and hence $\sigma_{p\text{-air}}$ directly. Converting to a proton–proton cross section requires a nuclear-overlap calculation, and the results agree with the extrapolation of accelerator measurements [2427] — a consistency check of the strong interaction three decades in $\sqrt{s}$ beyond where it was calibrated.

The second is the muon content, and it does *not* agree.

Remark 115.53 (The muon excess is a real discrepancy and is left as one)*.* The muon number predicted by Equation (115.32), with N_{ch} and the inelasticity taken from every current hadronic interaction model tuned to LHC data, falls below the measured number in showers above $1 \times 10^{19}\,\mathrm{eV}$ by a factor between about 1.3 and 1.6, with the primary mass fixed independently by $X_{\max}$ through Phenomenon 115.25. Since N_μ depends on the mass only through the mild $A^{1-\beta}$ of Equation (115.33), the discrepancy cannot be absorbed into the composition without contradicting the depth measurement. It is therefore a failure of the hadronic modelling at energies no collider reaches, and it is the one place in this chapter where an extrapolated Standard Model calculation is measurably wrong. It is reported as such.

115.8.2 Tests of Lorentz invariance

The invariance postulates of Chapter 39 assert that every species has the same limiting speed. Suppose instead that species a has maximal attainable velocity $c_a = c\,(1+\delta_a)$ with $|\delta_a| \ll 1$, so that its dispersion relation reads $E^2 = p^2c_a^2 + m^2c_a^4$. Cosmic rays test this at a sensitivity no laboratory can approach, for a simple reason: the effect of δ grows with energy while every terrestrial measurement is made at low energy.

Derivation of the bound from the photopion threshold. Redo the threshold calculation of Equation (115.56) with species-dependent limiting speeds. Writing $\delta := \delta_\pi - \delta_p$, the invariant that must reach $(m_p + m_\pi)^2\,c^4$ acquires an extra term of order δE^2, so the threshold condition becomes

$$4EE_\gamma - 2\delta E^2 \geq m_\pi c^2 \left(2m_p c^2 + m_\pi c^2\right) .$$

For $\delta > 0$ the reaction is forbidden above a maximum energy rather than allowed above a minimum one, and the photopion suppression would not occur at all. Requiring the second term not to dominate the first at the energy where the suppression is observed, $E \approx 5 \times 10^{19}\,\mathrm{eV}$, gives

$$|\delta| \lesssim \frac{2E_\gamma}{E} = \frac{2 \times 6.3 \times 10^{-4}\,\mathrm{eV}}{5 \times 10^{19}\,\mathrm{eV}} = 2.5 \times 10^{-23} , \tag{115.70}$$

a bound on a fractional velocity difference at the level of 10^{-23} [1968]. For comparison, the best terrestrial Michelson–Morley-type experiments — rotating optical cavities, the modern descendants of [1635] — bound anisotropies of the speed of light at the level of 10^{-17} [2147]; the cosmic-ray bound is better by six orders of magnitude, and it is obtained not by building anything but by observing that a predicted suppression occurs. □

A second and independent channel is the propagation of high-energy photons over cosmological distances. If the photon dispersion relation acquires an energy-dependent correction, photons of different energy emitted simultaneously arrive at different times, and the delay accumulates over the flight path exactly as in Equation (115.66). Gamma-ray bursts, which are both distant and variable on sub-second time scales, are the natural clock, and the observed simultaneity of low- and high-energy photons from them bounds the effect stringently [2137].

Remark 115.54 (No violation has been observed). Both channels give null results, and the resulting bounds are among the strongest constraints on any modification of special relativity. This chapter records that fact and nothing more: no Lorentz-violating framework is developed here, because none has evidence, and the programmes that predict such violations lie outside the scope stated in the front matter. What is in scope is the measurement, and the measurement says the postulates of Chapter 39 hold to Equation (115.70).

115.8.3 Muon time dilation as a standing test

Phenomenon 115.55 (Muons reach sea level)**.** Muons produced by cosmic-ray interactions near 15 km altitude are detected in quantity at sea level — about $70/(\mathrm{m^2\,s\,sr})$ from the vertical, or roughly one per square centimetre per minute on a horizontal surface — and the surviving fraction depends on their momentum [863, 1253]. The muon lifetime at rest is 2.197 µs [2427], and muons circulating in a storage ring at $\gamma \approx 29.3$ are observed to decay with that same proper lifetime, dilated in the laboratory by exactly the factor γ [1592].

Rests on Equation (115.3).

Derivation. *Derives Phenomenon 115.55.*

The PDG quotes the muon width as $\Gamma_\mu = (2.995\,983\,6 \pm 0.000\,003\,0) \times 10^{-19}\,\mathrm{GeV}$ [2427], which is a lifetime

$$\tau = \frac{\hbar}{\Gamma_\mu} = \frac{1.054\,571\,817 \times 10^{-34}\,\mathrm{J\,s}}{4.8000 \times 10^{-29}\,\mathrm{J}} = 2.1970 \times 10^{-6}\,\mathrm{s}\,,$$

and the distance a muon covers in one lifetime at a speed arbitrarily close to c is $c\tau = 658.6\,\mathrm{m}$. Were the laboratory lifetime the same as the proper one, the surviving fraction over 15 km of atmosphere would be

$$\exp\!\left(-\frac{15\,\mathrm{km}}{658.6\,\mathrm{m}}\right) = \mathrm{e}^{-22.8} = 1.3 \times 10^{-10}\,,$$

and no muons would reach the ground in any detectable number. What the flight consumes is instead the muon's *proper* time, shorter than the laboratory time by the factor γ (Chapter 41), so the laboratory decay length is $\gamma\beta c\tau$; for a typical sea-level muon of a few gigaelectronvolts, $\gamma \approx 20$ — that is $E = 20\,m_\mu c^2 = 2.1\,\mathrm{GeV}$, which is also about the energy a muon needs to survive the 2 MeV per g/cm^2 ionisation loss across the atmosphere of Equation (115.3) — and

$$\exp\!\left(-\frac{15\,\mathrm{km}}{20 \times 658.6\,\mathrm{m}}\right) = \mathrm{e}^{-1.14} = 0.32\,.$$

The observed flux is of this order. The nine and a half orders of magnitude between the two estimates are what make the sea-level muon flux a continuously running test of relativistic kinematics rather than a curiosity, and the momentum dependence of the surviving fraction is the form in which the test was first made quantitative [863]. □

The three measurements that turn this into a quantitative test are sections of Chapter 41: Rossi and Hall's momentum dependence (Section 41.2), the Frisch–Smith mountain-to-sea-level comparison (Section 41.3) and the CERN storage-ring lifetime (Section 41.4). What this chapter adds is the source: the muons are not supplied by an apparatus, they are supplied by the atmosphere, at a rate fixed by Equation (115.19) and a chain of decays that nothing on Earth controls.

115.9 What is settled and what is not

It is worth separating, at the close of the chapter, the statements that rest on measurement from those that do not.

Settled by observation: cosmic rays arrive from outside the atmosphere (Phenomenon 115.2) and are charged and predominantly positive (Phenomena 115.4 and 115.5); the spectrum is a power law over eleven decades with a knee, an ankle and a high-energy suppression (Phenomena 115.11 and 115.13); the flux is suppressed at the very top (Phenomenon 115.42); a single primary makes an extensive air shower whose particle number measures its energy and whose depth of maximum measures its mass (Phenomena 115.18 and 115.25); supernova remnants accelerate protons (Phenomenon 115.30) and can supply the Galactic energy budget (Phenomenon 115.29); petaelectronvolt accelerators exist in the Galaxy (Phenomenon 115.46); the arrival directions above $8 \times 10^{18}\,\mathrm{eV}$ carry a dipole of extragalactic origin (Phenomenon 115.44); an extraterrestrial high-energy neutrino flux exists (Phenomenon 115.50); and atmospheric neutrinos oscillate (Phenomenon 115.47).

Not settled, and stated as such throughout:

- Which of the two magnetic readings of the knee is correct. Both predict the rigidity ordering that is observed, and nothing in the present data separates them (Remark 115.14).

- The composition above 1×10^{19} eV, and therefore whether the observed suppression is the propagation effect of Phenomenon 115.40 or the maximum rigidity of the sources (Phenomenon 115.42).
- The identity of any source of the highest-energy particles. No point source has been confirmed, and the Hillas criterion excludes rather than identifies (Remark 115.38).
- The origin of the rising positron fraction (Remark 115.17).
- Why the observed spectral index is 2.7 when shock acceleration and the measured escape time predict 2.33 (Remark 115.35).
- Why showers contain more muons than every hadronic model predicts (Remark 115.53).

Six open problems in a subject a century old is not a poor showing, and it is a fair description of where the field stands. The instrument that will settle most of them is exposure: the flux at the top of the spectrum is one particle per square kilometre per century, and every statement about it is limited by how much sky has been watched for how long.

Chapter 116

Experiment: Deep Inelastic Scattering

Tests Phenomena 102.45 and 102.47 and Equations (102.56) *and* (102.59).
Assuming Definitions 102.5 and 102.44 and Proposition 102.6.

Rutherford found the nucleus by observing that a few alpha particles came back; the SLAC–MIT collaboration found the quarks the same way, by observing that electrons scattered from hydrogen at large angles far more often than a soft, extended proton would allow. This chapter is the experimental counterpart of Chapter 102: it reports the electron-scattering measurements carried out in End Station A at the Stanford Linear Accelerator Center between 1967 and 1972, the Bjorken scaling they exhibited, the Callan–Gross relation that identified the constituents as spin-$\tfrac{1}{2}$ objects, and the missing momentum fraction that required a neutral, non-scattering carrier — the gluon. It then follows the same observable forward: the logarithmic scaling violations measured at high statistics by muon and neutrino beams and at HERA are the most direct quantitative confirmation of quantum chromodynamics that exists, because the theory predicts not the structure functions themselves but exactly how they must change with Q^2.

The prediction under test is a single sentence: *the proton contains point-like charged constituents.* It is falsifiable in the strongest sense, because a proton whose charge is smoothly spread over its measured size predicts an inelastic rate at large momentum transfer smaller than the observed one by orders of magnitude. The chapter is laid out in the five rubrics that editorial rule 4 requires of an experiment chapter, and that Chapter 41 exhibits — apparatus, procedure, observations with numbers and uncertainties in SI units, interpretation, and primary references — with the historical setting first and the later high-precision programme placed between interpretation and references, because it is the same measurement continued. It sits at the end of Part XI with Chapter 115; its theoretical neighbours are Chapters 102 and 108 and, for the leptonic probe, Chapters 100 and 101. The standard textbook account is [1739]; current structure function and parton-distribution data are collected in [2427].

Remark 116.1 (Units, and how to read this chapter against the literature). Every equation here carries $\hbar$ and c explicitly, as Chapter 102 does. The deep inelastic literature universally sets $\hbar = c = 1$ and writes Q^2 in GeV^2; to map that literature onto this chapter, restore one factor of c for every power of momentum and one factor of $\hbar c$ for every power of length. Concretely: a four-momentum q carries $\mathrm{kg\,m/s}$, so the invariant $Q^2 := -q^2$ carries $\mathrm{kg^2\,m^2/s^2}$ and the literature's "Q^2 in GeV^2" is Q^2c^2 here. The conversion factors used throughout are

$$1\,\mathrm{GeV} = 1.602\,176\,634 \times 10^{-10}\,\mathrm{J}\,, \qquad 1\,\mathrm{GeV}^2/c^2 = 2.8561 \times 10^{-37}\,\mathrm{kg^2\,m^2/s^2}\,, \tag{116.1}$$

the first exact because the 2019 SI fixes the elementary charge, and the second following from it, together with $\hbar c = 3.161\,526\,77 \times 10^{-26}\,\mathrm{J\,m}$. Areas quoted in the older literature in

cm^2 are given here in m^2, with $1\,\mathrm{cm}^2 = 1\times10^{-4}\,\mathrm{m}^2$. No derivation in this chapter is carried out in natural units.

116.1 Historical context and the prediction under test

116.1.1 Elastic scattering and the proton form factors

Before the proton could be found to have constituents it had to be measured as a whole, and the instrument for that was elastic electron scattering, $e + p \to e + p$, in which the proton survives.

Definition 116.2 (The Rosenbluth cross section)**.** For one-photon exchange from a proton with an anomalous magnetic moment, Lorentz invariance, current conservation and parity allow exactly two form factors, the electric $G_E(Q^2)$ and the magnetic $G_M(Q^2)$, both dimensionless functions of the invariant momentum transfer alone. The elastic cross section is then

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \sigma_{\mathrm{Mott}}\,\frac{E'}{E}\left[\frac{G_E^2 + \tau G_M^2}{1+\tau} + 2\tau G_M^2 \tan^2\frac{\vartheta}{2}\right], \qquad \tau := \frac{Q^2}{4M^2c^2}\,, \tag{116.2}$$

with σ_{Mott} the point-charge cross section Equation (102.50), M the proton mass, ϑ the electron scattering angle, E and E' the incident and scattered electron energies, and the recoil factor

$$\frac{E'}{E} = \left[1 + \frac{2E}{Mc^2}\sin^2\frac{\vartheta}{2}\right]^{-1}. \tag{116.3}$$

Both τ and E'/E are dimensionless: Q^2 carries $\mathrm{kg}^2\,\mathrm{m}^2/\mathrm{s}^2$ and so does M^2c^2 [970].

Remark 116.3 (Why the formula is written this way)*.* The bracket in Equation (116.2) is linear in $\tan^2(\vartheta/2)$ at fixed Q^2. Measuring at several angles and several beam energies chosen to keep Q^2 fixed therefore gives a straight line whose slope is $2\tau G_M^2$ and whose intercept is $(G_E^2 + \tau G_M^2)/(1+\tau)$: the two form factors are separated with no model of the proton whatever. This is the *Rosenbluth separation*, and exactly the same device, applied to the inelastic cross section, separates the two inelastic structure functions in Section 116.3.2. The experimental technique of the deep inelastic programme is in this sense a direct continuation of the elastic one.

Phenomenon 116.4 (The proton is soft and extended)**.** The elastic form factors of the proton fall steeply with momentum transfer, and over the range measured in the 1950s they are described to within about 10 % by the dipole

$$G_E(Q^2) = \frac{G_M(Q^2)}{\mu_p} = \left[1 + \frac{Q^2c^2}{\Lambda^2}\right]^{-2}, \qquad \Lambda^2 = 0.71\,\mathrm{GeV}^2\,, \tag{116.4}$$

corresponding to a root mean square charge radius near 0.8 fm [1088]. The proton is therefore not a point: it has a size, and at momentum transfers large compared with Λ/c the probability of leaving it intact collapses.

Rests on Definition 116.2 and Equations (102.50) *and* (116.3)*.*

Derivation. *Derives Phenomenon 116.4.*

For a static charge distribution $\rho(\boldsymbol{r})$ normalized to unit total charge, the Born amplitude carries the Fourier transform Equation (116.23) below, and expanding the exponential for small momentum transfer gives, after angular averaging,

$$G_E(Q^2) = 1 - \frac{Q^2\,\langle r^2\rangle}{6\hbar^2} + O(Q^4)\,, \tag{116.5}$$

which is dimensionally consistent because $Q^2\,\langle r^2\rangle$ carries $\mathrm{kg}^2\,\mathrm{m}^4/\mathrm{s}^2$, exactly $\hbar^2$. Expanding the dipole Equation (116.4) to the same order gives $G_E = 1 - 2Q^2c^2/\Lambda^2$, so

$$\left\langle r^2\right\rangle = \frac{12\hbar^2c^2}{\Lambda^2} \quad\Longrightarrow\quad \sqrt{\langle r^2\rangle} = \frac{\sqrt{12}\,\hbar c}{\Lambda} = \frac{3.4641 \times 0.197\,33\,\mathrm{GeV\,fm}}{0.842\,62\,\mathrm{GeV}} = 0.811\,\mathrm{fm}\,. \quad (116.6)$$

The modern value from the CODATA adjustment is $(8.4075 \pm 0.0064) \times 10^{-16}\,\mathrm{m}$ [2431], so the dipole radius lies about 3.5 % *below* it. That gap is the honest statement that the dipole is a convenient parametrisation of a decade of data and not a law, and that a radius extracted from it is not a radius extracted from the slope at $Q^2 = 0$.

The consequence that matters for this chapter is the falloff, and it is worth accounting for term by term rather than quoting a single power. At large Q^2 the bracket of Equation (116.2) is dominated by its magnetic term $2\tau G_M^2 \tan^2(\vartheta/2)$; since $G_M^2 \propto Q^{-8}$ while $\tau \propto Q^2$, the bracket falls as Q^{-6}. At fixed angle and asymptotically large energy the recoil factor Equation (116.3) contributes a further Q^{-2} and σ_{Mott} a further Q^{-4}, giving the familiar Q^{-12}. That law is asymptotic, however, and the first survey did not reach it: evaluating Equation (116.2) with Equation (116.4) and the proton magnetic moment $\mu_p = 2.7928$ nuclear magnetons [2427] gives a local exponent running only from about Q^{-6} at $Q^2c^2 = 1\,\mathrm{GeV}^2$ to about Q^{-9} at $8\,\mathrm{GeV}^2$. The honest number is therefore the one the formula itself returns: raising Q^2c^2 from $1\,\mathrm{GeV}^2$ to $8\,\mathrm{GeV}^2$ at $\vartheta = 10^\circ$ suppresses the elastic cross section by a factor 3.2×10^3, and suppresses its ratio to σ_{Mott} — the like-for-like comparison, since that ratio is what the inelastic measurement reports — by a factor 3.4×10^2. The corresponding factors at 6° and 25° are 2.9×10^3 and 4.5×10^3 raw, 3.2×10^2 and 3.7×10^2 relative to Mott, so the conclusion does not depend on the angle chosen. That is what made the inelastic result of Section 116.4.1 a surprise: the same beam, the same target and a modest change of final state gave a rate that hardly fell at all. □

116.1.2 The quark hypothesis

By 1964 the proliferating hadrons had been organized into representations of a flavour SU(3) — the eightfold way of Section 102.1.1, whose representation theory is Chapter 5. Gell-Mann [1277] and, independently, Zweig [1301] observed that every observed multiplet is built from the fundamental triplet and its conjugate, and proposed that the triplet corresponds to physical constituents: three of them, with electric charges $+\frac{2}{3}e$, $-\frac{1}{3}e$ and $-\frac{1}{3}e$, from which mesons are made as $q\bar{q}$ and baryons as qqq (Definition 102.5 and Proposition 102.6).

The proposal was widely, and reasonably, not believed as a statement about matter. Three objections carried weight.

- *No free quark had ever been seen.* Searches in cosmic rays, in accelerator beams, in sea water and in terrestrial minerals had set limits on the abundance of fractionally charged particles far below anything a genuine constituent should permit; the modern statement of that null result is Phenomenon 102.65. A constituent that cannot be liberated was, in 1964, a contradiction in terms.

- *The masses did not add up.* The constituent masses needed to reproduce the baryon spectrum are roughly a third of the nucleon mass each, but the binding energy implied is then comparable with the masses themselves, so the non-relativistic picture that made the counting work had no right to succeed.

- *The statistics were wrong.* The Δ^{++} is uuu with all three spins aligned in a spatially symmetric ground state, which is symmetric under exchange of identical fermions (Section 102.2.1). The resolution — colour — was not established until later.

Quarks were therefore treated by most of the community as a bookkeeping device: a mnemonic for SU(3) multiplets, not a claim about what is inside a proton. The distinction is exactly what deep inelastic scattering could settle, because a mnemonic has no form factor and a constituent does.

116.1.3 Bjorken scaling

Bjorken's prediction preceded the data and did not depend on the quark model. It came from current algebra: the equal-time commutators of the electromagnetic currents, which encode the charges of whatever the currents are built from, imply sum rules for the structure functions, and those sum rules are finite only if the structure functions have a particular limiting behaviour [1393].

Definition 116.5 (The Bjorken limit)**.** The *Bjorken limit* is $Q^2 \to \infty$ and $\nu \to \infty$ with

$$x = \frac{Q^2}{2M\nu} \tag{116.7}$$

held fixed, where $\nu := P \cdot q/M$ is the energy transferred to the target in its rest frame and M the proton mass. Both limits are taken together: it is not a high-Q^2 limit at fixed ν, which would be the elastic region, nor a high-ν limit at fixed Q^2.

Phenomenon 116.6 (Bjorken's prediction)**.** In the limit of Definition 116.5 the two structure functions of Definition 102.44 become functions of the single dimensionless variable x:

$$Mc^2 W_1\left(x, Q^2\right) \longrightarrow F_1(x)\,, \qquad \nu W_2\left(x, Q^2\right) \longrightarrow F_2(x)\,. \tag{116.8}$$

This was published in 1969, before the SLAC data were analysed [1393], and the operator-product machinery that makes it a statement about short-distance behaviour rather than an assumption is Wilson's [1410].

Rests on Definitions 102.44 and 116.5.

Derivation. *Derives Phenomenon 116.6.*

The content is dimensional, and the dimensional analysis is the whole argument. The functions W_1 and W_2 of Definition 102.44 each carry 1/J, and the products in Equation (116.8) are therefore dimensionless. A dimensionless function of the two invariants can depend on them only through dimensionless combinations, and there are exactly three available: x, the ratio Q^2c^2/M^2c^4, and ν/Mc^2. If the target has no internal length scale beyond its own Compton wavelength — if, that is, the current is absorbed by a structureless object — then in the limit where Q^2c^2 and νMc^2 both dwarf M^2c^4 the last two ratios drop out of any finite answer and only x survives.

The converse is the informative half. Suppose the constituent had a size a. Then $\hbar/a$ is a momentum, $Q^2a^2/\hbar^2$ is a fourth dimensionless combination that does *not* disappear in the Bjorken limit, and the structure functions would depend on Q^2 separately through it — falling once $Q \gtrsim \hbar/a$, in exactly the way the elastic form factors of Phenomenon 116.4 do. Scaling is thus the absence of a constituent length, and a measured violation of scaling is either a constituent size or an interaction that supplies a scale of its own. Which of the two the observed violations are is settled in Section 116.6: they are logarithmic, not power-like, and a size gives a power. □

116.1.4 The parton model

Feynman supplied the picture that makes the limit intuitive [1401]. View the hadron in a frame in which its momentum P is very large. Time dilation slows every internal process by the Lorentz factor, so during the brief moment in which the virtual photon is absorbed the

constituents — Feynman's *partons* — neither interact with one another nor change their momenta. Each carries a fraction ξ of the hadron's longitudinal momentum, and the photon scatters elastically from one of them while the rest are spectators. This is the impulse approximation of Phenomenon 102.45, and Equation (102.53) identifies ξ with the measured x.

Bjorken and Paschos turned the picture into a formula [1394]. Writing $q_i(x)\mathrm{d}x$ for the number of partons of species i and charge e_i carrying momentum fraction in $\mathrm{d}x$, the incoherent sum over point-like elastic scatterings gives Equation (102.54), which in the notation of this chapter reads

$$F_2(x) = \sum_i \left(\frac{e_i}{e}\right)^2 x\, q_i(x)\,, \tag{116.9}$$

a function of x alone — which is Bjorken scaling. The parton distributions q_i carry no dimension at all: x is dimensionless and $q_i(x)\mathrm{d}x$ is a number.

Remark 116.7 (What the parton model assumes and does not explain). The model presumes that the constituents are free during the absorption. Nothing in it explains why constituents bound tightly enough never to escape should behave as free particles when probed hard, and in 1969 this was the model's central embarrassment rather than its central insight: it was a hypothesis with no dynamics behind it. The debt is discharged by asymptotic freedom (Theorem 102.35), which is the subject of Section 116.6.1 — and the sequence matters historically, because the parton model was believed on experimental grounds for four years before a field theory was found that could support it.

116.1.5 The Callan–Gross relation

The parton model as stated says nothing about the constituents' spin. Callan and Gross showed that the two structure functions are not independent if the constituents carry spin $\frac{1}{2}$, and that the relation between them is opposite for spin 0 [1398].

Definition 116.8 (The longitudinal structure function and R)**.** Define

$$F_L\left(x, Q^2\right) := \left(1 + \frac{Q^2c^2}{\nu^2}\right) F_2\left(x, Q^2\right) - 2xF_1\left(x, Q^2\right)\,, \tag{116.10}$$

and the ratio of longitudinal to transverse virtual-photon absorption cross sections

$$R\left(x, Q^2\right) := \frac{\sigma_L}{\sigma_T} = \frac{F_L}{2xF_1}\,. \tag{116.11}$$

The ratio Q^2c^2/ν^2 equals $4M^2x^2c^4/Q^2c^2$ and vanishes in the Bjorken limit at fixed x, so $F_L \to F_2 - 2xF_1$ there. All three quantities are dimensionless.

Phenomenon 116.9 (Callan–Gross)**.** For constituents of spin $\frac{1}{2}$ and negligible mass, $F_L = 0$, equivalently $F_2(x) = 2xF_1(x)$ and $R = 0$; for constituents of spin 0, $F_1 = 0$ and $R \to \infty$ [1398]. Any measured value of R therefore discriminates between the two hypotheses without needing an absolute normalisation, a knowledge of the parton distributions, or a model of the proton.

Rests on Definition 116.8, Equation (116.9)*, and Phenomenon 102.45.*

The derivation is the helicity argument given in Phenomenon 102.45 and repeated, in the form in which it is used against the data, in Section 116.4.3. What must be recorded here is why this is the sharpest of the early tests. Every other comparison between the parton model and the data involves either an absolute cross section, which carries the beam normalisation and the target thickness, or a distribution $q_i(x)$, which is not predicted by anything. R is a ratio of two cross sections measured with the same beam on the same target, and the two hypotheses do not predict different numbers for it — they predict opposite ends of its range.

116.2 Apparatus

116.2.1 The Stanford linear accelerator

The instrument is a 3.05 km disc-loaded copper waveguide buried under the hills west of Stanford, operating as a travelling-wave structure at 2856 MHz [1363]. Electrons injected at a few tens of MeV ride the phase velocity of the wave and gain energy continuously along the machine, reaching about $20\,\mathrm{GeV} = 3.2 \times 10^{-9}\,\mathrm{J}$ at the end of the machine. Radiofrequency power is supplied by some 240 klystrons distributed along the tunnel, each delivering tens of megawatts of peak power into its own group of four 3.05 m accelerating sections: $240 \times 4 \times 3.05\,\mathrm{m} = 2.93\,\mathrm{km}$ of structure, which with the drift and injection regions is the length of the machine.

Three features of the beam decided the experiment.

- *Energy and its calibration.* The maximum energy sets the kinematic reach: the largest accessible Q^2 at a given angle grows as E^2. The energy was defined by a magnetic analysing system in the beam switchyard and known to a few parts in 10^3, which is what makes an absolute measurement of the invariants possible at all; every kinematic variable of Section 116.3.1 is built from E, E' and ϑ and from nothing else.

- *Pulse structure.* The machine ran at up to 360/s with a pulse length of order 1.6 µs, a duty factor of about 6×10^{-4}. A short, intense pulse is a mixed blessing: it delivers the charge needed for a cross section falling by six orders of magnitude across the kinematic range, but it forces every counter to resolve several particles inside the pulse, which is the reason the detector stack of Section 116.2.2 is built from many small scintillator elements rather than a few large ones.

- *Current.* Average beam currents in the microampere range were typical. At 1 µA the beam delivers $1\,\mu\mathrm{C/s} / 1.602\,176\,634 \times 10^{-19}\,\mathrm{C} = 6.24 \times 10^{12}/\mathrm{s}$ electrons; against the areal target density of Section 116.2.3 this is the luminosity that makes a cross section of order $1 \times 10^{-38}\,\mathrm{m^2/sr}$ measurable in hours.

Remark 116.10 (Why an electron beam). A hadron beam probes a hadron with the strong interaction, whose short-distance form was precisely what was unknown; any measured rate then mixes the structure of the target with the structure of the probe and with the unknown dynamics between them. The electron interacts through a single virtual photon whose coupling is the fine-structure constant $\alpha = e^2/(4\pi\varepsilon_0\hbar c) = (7.297\,352\,564\,3 \pm 0.000\,000\,001\,1) \times 10^{-3}$ [2431], small enough that one-photon exchange dominates and calculable exactly (Chapter 100). The whole of the unknown physics is then confined to the two structure functions of Definition 102.44, and the experiment measures them with no theoretical input beyond quantum electrodynamics. The price is rate: $\alpha^2 \approx 5.3 \times 10^{-5}$ against a strong cross section of order unity, which is why the beam power and the target thickness matter as much as they do.

116.2.2 The End Station A spectrometers

End Station A is a shielded experimental hall at the end of the switchyard, large enough to hold three magnetic spectrometers mounted on carriages that pivot about the target position on a circular rail. They are named by the momentum they can analyse: $20\,\mathrm{GeV}/c$, $8\,\mathrm{GeV}/c$ and $1.6\,\mathrm{GeV}/c$. The first two carried the deep inelastic programme; the third served the largest angles.

Proposition 116.11 (Why the spectrometer weighs thousands of tonnes)**.** *Analysing the momentum of a* $20\,\mathrm{GeV}/c$ *electron by bending it through an angle* θ_B *requires a field integral*

$$\int B\,\mathrm{d}l = \frac{p\,\theta_B}{e}\,, \tag{116.12}$$

which for $\theta_B = 10^\circ$ *is about* $12\,\mathrm{T\,m}$*. With conventional iron-cored dipoles limited by saturation to fields of order* $1.5\,\mathrm{T}$*, this is eight metres of magnet; with the drift spaces needed to convert an angle into a displacement large enough to measure, the instrument is tens of metres long, and its yoke is the mass that must be rotated to change the scattering angle.*

Rests on Equations (40.15) *and* (113.8)*.*

Proof. *Derives Proposition 116.11.*

A charge e moving with momentum p perpendicular to a uniform field B follows a circle of radius $r = p/(eB)$, so a path length $\mathrm{d}l$ turns the trajectory through $\mathrm{d}\theta = \mathrm{d}l/r = eB\,\mathrm{d}l/p$; integrating gives Equation (116.12), which is exact for any field profile provided the bend stays in one plane. Numerically,

$$p = 20\,\mathrm{GeV}/c = \frac{20\,\mathrm{GeV} \times 1.602\,176\,634 \times 10^{-10}\,\mathrm{J/GeV}}{2.997\,924\,58 \times 10^{8}\,\mathrm{m/s}} = 1.0689 \times 10^{-17}\,\mathrm{kg\,m/s}\,,$$

and with $\theta_B = 0.174\,53\,\mathrm{rad}$,

$$\int B\,\mathrm{d}l = \frac{1.0689 \times 10^{-17}\,\mathrm{kg\,m/s} \times 0.174\,53}{1.602\,176\,634 \times 10^{-19}\,\mathrm{C}} = 11.6\,\mathrm{T\,m}\,.$$

The momentum resolution follows from the same relation: a fractional error $\delta p/p$ shows up as an angular error $\delta\theta_B = \theta_B\,\delta p/p$, so resolving $\delta p/p \sim 1 \times 10^{-3}$ at $\theta_B = 10^\circ$ means measuring the exit angle to $1.7 \times 10^{-4}\,\mathrm{rad}$ — a fraction of a millimetre over a metre of lever arm, which is what the hodoscope granularity of the detector stack has to deliver. □

The optics of both spectrometers were designed to be *point-to-point* in momentum and independent of the production angle within the acceptance, so that a hit position in one hodoscope plane measures E' and a hit position in another measures ϑ, with the two decoupled. The acceptances are small: of order a tenth of a millisteradian in solid angle and a few percent in momentum for the $20\,\mathrm{GeV}/c$ instrument, roughly an order of magnitude more solid angle for the $8\,\mathrm{GeV}/c$ one, which is why the latter, reaching larger angles where the cross section is smaller, was the instrument for the Rosenbluth separations. The $20\,\mathrm{GeV}/c$ spectrometer covered angles from a few degrees to about 25°; the $8\,\mathrm{GeV}/c$ one reached 90°.

The detector stack behind the magnets had one job: identify an electron of measured momentum inside a burst that also contains pions from hadronic final states, at a pion-to-electron ratio that can reach 10^3 or more at small E'. It did so redundantly.

- *Scintillation hodoscopes.* Planes of narrow plastic scintillator strips, read by photomultipliers, giving the position in the dispersive and non-dispersive directions and hence E' and ϑ, together with the fast coincidence that forms the trigger.

- *A threshold gas Cherenkov counter.* A tank of gas at about atmospheric pressure — carbon dioxide or nitrogen, refractive index $n - 1 \approx 4.5 \times 10^{-4}$ — in which a particle radiates only if $\beta > 1/n$. The threshold Lorentz factor is $\gamma_{\mathrm{th}} = (1 - n^{-2})^{-1/2} \approx [2(n-1)]^{-1/2} = 33$, so the threshold momenta are

  $$p_{\mathrm{th}}(e) \approx \gamma_{\mathrm{th}} m_e c = 17\,\mathrm{MeV}/c\,, \qquad p_{\mathrm{th}}(\pi) \approx \gamma_{\mathrm{th}} m_\pi c = 4.7\,\mathrm{GeV}/c\,,$$

  using $m_e c^2 = (0.510\,998\,950\,00 \pm 0.000\,000\,000\,15)\,\mathrm{MeV}$ and $m_{\pi^\pm} c^2 = (139.570\,39 \pm 0.000\,18)\,\mathrm{MeV}$ [2427]. Every electron in the acceptance radiates; no pion below $4.7\,\mathrm{GeV}/c$ does.

- *A total-absorption shower counter.* Lead glass, some fifteen radiation lengths deep, in which an electron deposits its entire energy in an electromagnetic cascade while a pion deposits, on average, the small energy of a minimum-ionizing track plus whatever a nuclear interaction happens to give. Requiring the deposited energy to match the momentum measured magnetically is an independent rejection.

Together the two particle-identification devices suppress pions by three to four orders of magnitude with an electron efficiency near unity, and — because they are independent — each can be used to measure the other's efficiency directly from the data. First-hand accounts of the instrument and its use are the Nobel lectures of Taylor [1849] and Kendall [1848].

116.2.3 Targets, monitoring and normalisation

The target was liquid hydrogen, and for the neutron structure function liquid deuterium, held in thin-walled cells in a cryostat on the pivot axis.

Example 116.12 (The target as a number of scattering centres)**.** Liquid hydrogen at its boiling point, 20.3 K at atmospheric pressure, has density $\rho = 70.8\,\mathrm{kg/m^3}$. A cell of length $L = 0.30\,\mathrm{m}$ therefore presents an areal number density of protons

$$n_A = \frac{\rho L N_A}{M_{\mathrm{H}}} = \frac{70.8\,\mathrm{kg/m^3} \times 0.30\,\mathrm{m} \times 6.022\,140\,76 \times 10^{23}/\mathrm{mol}}{1.008 \times 10^{-3}\,\mathrm{kg/mol}} = 1.27 \times 10^{28}/\mathrm{m^2}\,, \tag{116.13}$$

with N_A the Avogadro constant, exact by the 2019 SI [2431]. Combined with the $6.24 \times 10^{12}/\mathrm{s}$ electrons of a 1 µA beam, the luminosity is $7.9 \times 10^{40}/(\mathrm{m^2\,s})$, so a differential cross section of $1 \times 10^{-38}\,\mathrm{m^2/sr}$ into the $1 \times 10^{-4}\,\mathrm{sr}$ acceptance of Section 116.2.2 yields $7.9 \times 10^{40}/(\mathrm{m^2\,s}) \times 1 \times 10^{-38}\,\mathrm{m^2/sr} \times 1 \times 10^{-4}\,\mathrm{sr} = 0.079/\mathrm{s}$, that is 285 counts an hour, which is the scale of the experiment.

Three systematic problems follow directly from the target and are dealt with in the design.

- *Boiling.* The beam deposits energy in the liquid and produces vapour bubbles, so the density in the beam path is lower than the static density and depends on the beam current. The cells were therefore built with forced circulation, and the density correction was measured *in situ* by comparing rates at different currents and extrapolating to zero.

- *Walls.* The cell windows scatter too. An identical empty cell was measured under identical conditions and subtracted; the subtraction is largest at small E', where the wall material's own inelastic yield is comparable with the hydrogen signal.

- *Radiation length.* The hydrogen itself is a radiator. With $X_0 = 630.4\,\mathrm{kg/m^2}$ for hydrogen, the 0.30 m cell is $\rho L/X_0 = 21.2\,\mathrm{kg/m^2}/630.4\,\mathrm{kg/m^2} = 0.034$ radiation lengths, and a 0.07 m cell about 0.008. That number enters the external radiative correction of Section 116.3.3 directly, and is the reason the short cell was used wherever the rate allowed.

The incident charge was measured by toroidal current monitors surrounding the beam pipe — non-intercepting, reading the pulse's own magnetic field — and by secondary-emission monitors, the two cross-calibrated against a Faraday cup that collects the beam absolutely. The absolute scale of the cross sections was in addition tied to elastic electron–proton scattering measured with the same apparatus, where Equation (116.2) and the

known form factors supply the answer: any error in luminosity, acceptance or efficiency shows up as a discrepancy in a cross section that is independently known. The residual normalisation uncertainty quoted for the programme is a few percent. The whole apparatus, the calibrations and the analysis are reviewed by Friedman and Kendall [1472].

116.3 Procedure

116.3.1 Kinematics of the measurement

The reaction is inclusive: $e + p \to e' + X$, with only the scattered electron detected and the hadronic system X left unobserved. That is not a limitation but the point — summing over every final state is what makes the cross section an object with a short-distance interpretation, and it is why a single-arm spectrometer suffices.

Notation 116.13 (The measured quantities and the invariants). Three numbers are measured for each event: the incident energy E, the scattered energy E' and the scattering angle ϑ. From them, neglecting the electron mass against the energies,

$$\nu = E - E' \,, \tag{116.14}$$

$$Q^2 = \frac{4EE'}{c^2} \sin^2 \frac{\vartheta}{2} \,, \tag{116.15}$$

$$x = \frac{Q^2}{2M\nu} = \frac{2EE' \sin^2(\vartheta/2)}{Mc^2\,(E - E')} \,, \tag{116.16}$$

$$y = \frac{\nu}{E} \,, \tag{116.17}$$

$$W^2 c^4 = M^2 c^4 + 2Mc^2 \nu - Q^2 c^2 \,, \tag{116.18}$$

where M is the proton mass, $Mc^2 = (0.938\,272\,088\,16 \pm 0.000\,000\,000\,29)\,\mathrm{GeV}$ [2427], and W is the invariant mass of the unobserved hadronic system. The variables x and y are dimensionless and lie in $[0, 1]$; ν is an energy; Q^2 carries $\mathrm{kg^2\,m^2/s^2}$; W is a mass. Elastic scattering is $W = M$, that is $x = 1$. This is Notation 102.43 written in laboratory variables.

Derivation of Equation (116.15). *Derives Equation* (116.15).

With $k = (E/c, \boldsymbol{k})$ and $k' = (E'/c, \boldsymbol{k}')$ the incident and scattered electron four-momenta and $q = k - k'$,

$$Q^2 = -q^2 = -\left(k - k'\right)^2 = -k^2 - k'^2 + 2k \cdot k' = 2\left(\frac{EE'}{c^2} - |\boldsymbol{k}|\,|\boldsymbol{k}'| \cos\vartheta\right)$$

using $k^2 = k'^2 = m_e^2 c^2 \approx 0$ at these energies. With $|\boldsymbol{k}| = E/c$ and $|\boldsymbol{k}'| = E'/c$ to the same accuracy,

$$Q^2 = \frac{2EE'}{c^2}\,(1 - \cos\vartheta) = \frac{4EE'}{c^2} \sin^2 \frac{\vartheta}{2} \,.$$

The neglect of m_e is quantitatively safe: at $E = 10\,\mathrm{GeV}$ the ratio $m_e^2 c^4 / E^2$ is 2.6×10^{-9}. □

The (E', ϑ) plane at fixed E divides into three regions, and the experiment is the traverse across them.

- *The elastic peak*, $W = M$, a line in the plane given by $E' = E/[1 + 2E \sin^2(\vartheta/2)/Mc^2]$. Its cross section is Equation (116.2) and it falls steeply, asymptotically as Q^{-12} and by a factor of a few thousand across the range of this survey (Phenomenon 116.4).
- *The resonance region*, W between M and about $2\,\mathrm{GeV}/c^2$, where the $\Delta(1232)$ and the higher nucleon resonances appear as bumps. Each resonance has its own transition form factor and each falls steeply with Q^2.

- *The deep inelastic region*, W above about $2\,\mathrm{GeV}/c^2$, where no individual state can be resolved and the cross section is smooth. This is where scaling is tested, and the $W > 2\,\mathrm{GeV}/c^2$ cut is the operational definition of "deep" used throughout the SLAC analysis.

The double-differential cross section in the measured variables is Equation (102.49),

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E'} = \sigma_{\mathrm{Mott}}\left[W_2\left(x,Q^2\right) + 2W_1\left(x,Q^2\right)\tan^2\frac{\vartheta}{2}\right], \tag{116.19}$$

with σ_{Mott} given by Equation (102.50) and carrying m^2, so that the left-hand side carries $\mathrm{m}^2/(\mathrm{sr\,J})$ and W_1, W_2 each carry $1/\mathrm{J}$, as Definition 102.44 states.

116.3.2 The measurement programme

- *1967.* Commissioning runs with the $20\,\mathrm{GeV}/c$ spectrometer, elastic scattering measured first because it is the absolute calibration.
- *1968.* The inelastic survey at $\vartheta = 6^\circ$ and 10°, with incident energies from about $4.5\,\mathrm{GeV}$ to $17.7\,\mathrm{GeV}$, covering Q^2c^2 from roughly $0.7\,\mathrm{GeV}^2$ to $8\,\mathrm{GeV}^2$ and W up to about $5\,\mathrm{GeV}/c^2$. The results were reported at the Vienna conference in September 1968 [1379] and published as [1395].
- *1969.* The scaling analysis, in which νW_2 measured at different E, E' and ϑ is plotted against a single variable and found to collapse [1397].
- *1970–1972.* Extension to larger angles with the $8\,\mathrm{GeV}/c$ spectrometer, which is what makes the separation of W_1 from W_2 — and hence the measurement of R — possible; and deuterium running, which gives the neutron structure function by subtraction and hence the ratio F_2^{en}/F_2^{ep} of Section 116.4.4. Reviewed in [1472].

Proposition 116.14 (How two angles separate the two structure functions)**.** *Write the measured cross section Equation* (116.19) *in terms of the virtual-photon polarisation parameter*

$$\varepsilon = \left[1 + 2\left(1 + \frac{\nu^2}{Q^2c^2}\right)\tan^2\frac{\vartheta}{2}\right]^{-1}, \tag{116.20}$$

which is dimensionless and runs over $(0,1]$*. Then*

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E'} = \Gamma_T\left[\sigma_T + \varepsilon\,\sigma_L\right], \tag{116.21}$$

with Γ_T *a purely kinematic flux factor. At fixed* (x,Q^2)*, plotting the measured cross section divided by* Γ_T *against* ε *gives a straight line whose intercept is* σ_T *and whose slope is* σ_L*; their ratio is* R *of Equation* (116.11)*.*

Rests on Equations (116.11) *and* (116.19) *and Notation 116.13.*

Proof. *Derives Proposition 116.14.*

Both ν and Q^2 are fixed once (x,Q^2) is fixed, by Equation (116.7), so the only freedom left in Equation (116.20) is ϑ; and ϑ can be varied at fixed (ν,Q^2) only by varying the beam energy, from Equations (116.14) and (116.15). Since $\tan^2(\vartheta/2)$ appears linearly in Equation (116.19) and ε is a Möbius function of it, the cross section is linear in ε, which is Equation (116.21); matching the two forms identifies the combinations $\sigma_T \propto W_1$ and $\sigma_L \propto W_2\left(1 + \nu^2/Q^2c^2\right) - W_1$, whose ratio is Equation (116.11).

The practical difficulty is visible in the algebra. A precise slope needs a long lever arm in ε, which needs a large range of ϑ at fixed (x,Q^2), which needs a large range of E — and each

beam energy carries its own normalisation. The uncertainty on R is therefore dominated by the *relative* normalisation between data sets taken months apart at different energies, not by counting statistics, and this is why R was quoted with an uncertainty comparable to its own value while νW_2 was known to better than 10 %. Section 116.4.3 returns to why the physics conclusion survives that. □

116.3.3 Radiative corrections and the systematic budget

The largest correction in the experiment is not instrumental. An electron that radiates a photon has a different energy from the one recorded, so the cross section extracted at a nominal (E, E') is contaminated by events that occurred at other kinematics — above all by elastic events in which a hard photon was emitted, which migrate into the inelastic region and can dominate it. Mo and Tsai gave the treatment used [1404].

Definition 116.15 (The three radiative processes)**.** *Internal* corrections are the $O(\alpha)$ QED corrections to the scattering itself: the vertex correction, the electron self-energies, vacuum polarisation of the exchanged photon (Chapter 100), and real bremsstrahlung from the incoming or outgoing electron at the scattering vertex. The virtual and real pieces are separately infrared divergent and their sum is finite, which is the Bloch–Nordsieck cancellation. *External* corrections are bremsstrahlung in the bulk material of the target and its windows, before and after the scattering, governed by the number of radiation lengths traversed. *Ionization* loss shifts the energy by a smaller, calculable amount.

Proposition 116.16 (The equivalent radiator)**.** *In the peaking approximation — in which bremsstrahlung photons are taken to be emitted exactly along the incident or the scattered electron direction, which is accurate because the characteristic emission angle is $m_e c^2/E \approx 5\times10^{-5}$ at $E = 10\,\mathrm{GeV}$ — the internal radiative correction is equivalent to placing a radiator of thickness*

$$t_{\mathrm{eq}} = \frac{\alpha}{\pi}\left[\ln\frac{Q^2c^2}{m_e^2c^4} - 1\right] \tag{116.22}$$

radiation lengths before and after the target. At $Q^2c^2 = 4\,\mathrm{GeV}^2$ and $m_e c^2 = 0.510\,998\,95\,\mathrm{MeV}$ [2427] this is

$$t_{\mathrm{eq}} = \frac{7.2974\times10^{-3}}{\pi}\left[\ln\frac{4}{2.611\times10^{-7}} - 1\right] = 2.323\times10^{-3}\times 15.54 = 0.036\,,$$

so the radiative tail is generated by an effective 3.6 % of a radiation length, to be compared with the 0.008 to 0.034 radiation lengths of real target material computed in Section 116.2.3.

Rests on Definition 116.15 and Proposition 102.48.

Proof. *Derives Proposition 116.16.*

The bremsstrahlung spectrum from a thin radiator of thickness t is $\mathrm{d}N/\mathrm{d}k = t/k$ per unit photon energy k to leading logarithmic accuracy, so the only thing a radiator does at this order is supply the coefficient of $\mathrm{d}k/k$. The one-loop QED correction to the elastic vertex supplies exactly the same $\mathrm{d}k/k$ spectrum with coefficient $(\alpha/\pi)\,[\ln(Q^2c^2/m_e^2c^4) - 1]$, the logarithm being the collinear (mass) singularity regulated by the electron mass and the -1 the constant left by the infrared cancellation — the same structure as Proposition 102.48, with the photon in place of the gluon. Equating the two coefficients gives Equation (116.22). The approximation is a leading-logarithm one; the exact treatment of [1404] keeps the non-peaking contributions and the exact elastic tail, and it is the difference between the two that sets the uncertainty quoted below. □

The unfolding is then a deconvolution: the measured yield at (E, E') is an integral of the true cross section over all (E_s, E_p) with $E_s \le E$ and $E_p \ge E'$ weighted by the radiator

kernel, plus the elastic tail computed from the known form factors. The equation is solved iteratively, starting from an assumed cross section and re-radiating until the prediction reproduces the data. In the SLAC analysis the correction reached several tens of percent at the smallest E', where the elastic tail is largest, and was typically 10–30 % in the deep inelastic region.

The systematic budget of the experiment is therefore dominated by theory, not by hardware:

- radiative corrections, a few percent of the cross section, being of order 10 % of a correction of order 25 %;
- absolute normalisation — beam charge, target density, solid angle and efficiency — a few percent, tied to elastic scattering as described in Section 116.2.3;
- relative normalisation between run periods, which is what limits R (Proposition 116.14);
- pion contamination, reduced below the percent level by the Cherenkov and shower counters and measured from the data;
- target-wall subtraction and acceptance modelling, each at the percent level.

The quoted uncertainty on νW_2 in the deep inelastic region is a few percent point-to-point with an overall scale uncertainty of a few percent on top [1472].

116.4 Observations and data

116.4.1 The weak fall-off of the inelastic cross section

Phenomenon 116.17 (The inelastic rate falls only weakly with momentum transfer)**.** The cross section for $e + p \to e' + X$ in the deep inelastic region falls far more slowly with Q^2 than the elastic cross section from the same target: measured relative to the point-charge (Mott) cross section, the inelastic rate is approximately independent of Q^2, whereas the elastic rate falls as the fourth power of the dipole form factor [1395].

Rests on Phenomenon 116.4 and Equation (116.19)*.*

Derivation. Derives Phenomenon 116.17.

For scattering from a static charge distribution $\rho(\boldsymbol{r})$ the Born amplitude is the Fourier transform

$$F(\boldsymbol{q}) = \int \rho(\boldsymbol{r})\, \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}/\hbar}\, \mathrm{d}^3 r\,, \tag{116.23}$$

normalized so that $F(0) = 1$, and the cross section is the point-charge one multiplied by $|F|^2$. If ρ is smooth on a scale R then F varies on the scale $\hbar/R$ in q and falls rapidly beyond it, so a distribution of root mean square radius near 0.8 fm [1088] suppresses the elastic rate steeply once $Q \gtrsim \hbar/R$ — which is the measured dipole behaviour of Phenomenon 116.4. The one way to obtain a rate that does *not* fall is a scattering centre with $F \equiv 1$, that is, one with no structure at the resolution probed. A cross section flat in Q^2 therefore says the electron is rebounding from something point-like inside the proton, exactly as Rutherford's large-angle alpha particles said the atom contained a small hard core; the soft extended proton of Section 116.1.1 cannot produce it. □

The size of the contrast is worth stating numerically, because it is the whole result, and the comparison must be made between like quantities. Over the range Q^2c^2 from 1 GeV^2 to 8 GeV^2, the elastic cross section *divided by* σ_{Mott} falls by a factor of about 3.4×10^2 (Phenomenon 116.4), while the ratio of the inelastic cross section to σ_{Mott} at fixed W above

$2\,\mathrm{GeV}/c^2$ changes by less than a factor of two [1395]. Between two and three orders of magnitude therefore separate the two behaviours, and no experimental systematic of the kind catalogued in Section 116.3.3 is of that size. Comparing the raw cross sections instead of their ratios to σ_{Mott} makes the contrast larger still, by the further factor of about ten that the Mott fall-off itself supplies.

116.4.2 Scaling of the structure functions

Phenomenon 116.18 (Bjorken scaling)**.** Plotted against the single variable $x = Q^2/2M\nu$, the measured structure function νW_2 collapses onto one curve: data taken at different beam energies, scattering angles and momentum transfers, covering Q^2 from about $1\,\mathrm{GeV}^2/c^2$ to $8\,\mathrm{GeV}^2/c^2$, lie on a common function of x alone, with no separate dependence on Q^2 within the errors [1397] — as had been predicted from current algebra [1393].

Rests on Phenomenon 116.6, Notation 116.13, and Equation (116.9)*.*

Derivation. *Derives Phenomenon 116.18.*

Regard the proton, in a frame where its momentum P is large, as a collection of free constituents, one of which carries a fraction ξ of its four-momentum [1394, 1401]. The electron scatters elastically off that one, so after absorbing the virtual photon of four-momentum q the struck constituent must again be on shell. Neglecting its mass against Q,

$$(\xi P + q)^2 = (\xi P)^2 \quad \Longrightarrow \quad 2\xi P \cdot q + q^2 = 0\,,$$

and with $q^2 = -Q^2$ and $P \cdot q = M\nu$ in the proton rest frame,

$$\xi = \frac{Q^2}{2M\nu} = x\,. \tag{116.24}$$

The measured x is therefore the momentum fraction of whatever was struck, and the inclusive cross section is the incoherent sum over constituents of a point-like elastic cross section, weighted by the probability of finding one at that fraction. That probability is a property of the proton and not of the probe, so the structure functions depend on x and not separately on Q^2: which is the observed collapse. The step that must eventually fail is the assumption that the constituents are free, and its failure is exactly the logarithmic Q^2 dependence of Section 116.6. □

Two things about the 1969 presentation deserve recording, because both are lessons about how such a result is established.

The original plots were made against $\omega = 1/x = 2M\nu/Q^2$ rather than against x, which spreads out the small-x region where the data then were; the modern convention is x, and the two are trivially related. And the collapse was demonstrated not by superposing curves but by plotting νW_2 at fixed ω against Q^2 and showing the result to be flat: a null test, in which the quantity that would carry the signature of a constituent size is displayed as a function of the variable it would depend on. Within the errors then available — a few percent point to point — no dependence was seen over the factor of eight in Q^2 covered [1379, 1397]. That the dependence is in fact present, and logarithmic, was established only when the lever arm in Q^2 grew by two further orders of magnitude (Sections 116.6.3 and 116.6.4).

116.4.3 The longitudinal-to-transverse ratio

This is the subsection to which Section 102.5 defers the experimental side of the Callan–Gross relation.

Phenomenon 116.19 (The longitudinal cross section is small)**.** The ratio $R = \sigma_L/\sigma_T$ of longitudinal to transverse virtual-photon absorption is measured to be small and consistent with zero, in the range 0.1 to 0.2 with comparable uncertainty, throughout the deep inelastic region [1397, 1472]. Equivalently the Callan–Gross relation $F_2(x) = 2xF_1(x)$ holds [1398].

Rests on Phenomenon 116.9, Proposition 116.14, and Equation (116.11)*.*

Derivation. Derives Phenomenon 116.19.

Work in the Breit frame, where the virtual photon transfers no energy and the struck constituent simply reverses its momentum, from $-p\hat{\boldsymbol{z}}$ to $+p\hat{\boldsymbol{z}}$. For a constituent of spin $\frac{1}{2}$ and negligible mass the vector coupling to the photon does not connect states of opposite chirality, so helicity is conserved; since the momentum reverses while the helicity does not, the projection of the spin on the fixed axis $\hat{\boldsymbol{z}}$ must reverse, and the constituent's angular momentum along that axis changes by ± 1 in units of $\hbar$. A transverse photon carries exactly $J_z = \pm 1$ and can supply it; a longitudinal photon carries $J_z = 0$ and cannot. Hence $\sigma_L = 0$ and $R = 0$ for spin-$\frac{1}{2}$ constituents. For spin-0 constituents the argument runs the other way: there is no spin to reverse, so J_z must be unchanged, only the longitudinal photon can be absorbed, and $R \to \infty$. The two hypotheses therefore predict opposite extremes of the same measured ratio, and the observed small R selects spin $\frac{1}{2}$. This is the sharpest of the early discriminators precisely because it is a ratio: it needs no absolute normalization and no knowledge of any structure function. □

Remark 116.20 (Why the conclusion survives the systematic budget)*.* Proposition 116.14 showed that R is the quantity most exposed to the relative normalisation between run periods, and the quoted uncertainty — of the same size as the central value — says so honestly. The conclusion is nevertheless not in doubt, and the reason is the structure of Phenomenon 116.9. The two hypotheses do not differ by a factor: spin $\frac{1}{2}$ gives $R = 0$ and spin 0 gives $R = \infty$, so what the measurement has to establish is only that R is finite and small. A systematic error large enough to move a measured $R = 0.18 \pm 0.10$ to a value consistent with a scalar constituent would have to be not a few percent but several hundred percent, which is excluded by the elastic normalisation of Section 116.2.3. The measurement is imprecise and the inference is not; distinguishing the two is the point of quoting the uncertainty.

The small non-zero value has since been understood and is itself a QCD prediction: gluon radiation gives a struck quark transverse momentum, which breaks the helicity argument at order α_s and generates $F_L \neq 0$. That R came out small rather than exactly zero is therefore, in retrospect, evidence for the interacting theory rather than against the parton picture.

116.4.4 Quark charges from the structure functions

Running the same apparatus on liquid deuterium gives, after correction for Fermi motion and the subtraction of the proton, the neutron structure function; and the ratio of the two is a direct measurement of the constituents' charges.

Proposition 116.21 (What the neutron-to-proton ratio measures)**.** *In the parton model with u and d quarks and isospin symmetry relating the neutron to the proton by $u \leftrightarrow d$,*

$$\frac{F_2^{en}(x)}{F_2^{ep}(x)} = \frac{4d(x) + u(x)}{4u(x) + d(x)}\,, \tag{116.25}$$

where u and d are the proton's distributions and the sea has been neglected. The ratio is bounded by $\frac{1}{4}$ and 4 for any *distributions, the extremes corresponding to pure u and pure d; the measured values lie inside those bounds and approach $\frac{1}{4}$ as $x \to 1$ [1394, 1472].*

Rests on Equation (116.9) *and Definition 102.5.*

Proof. *Derives Proposition 116.21.*

Substituting the charges $e_u = \frac{2}{3}e$ and $e_d = -\frac{1}{3}e$ into Equation (116.9) gives $F_2^{ep} = x[\frac{4}{9}u + \frac{1}{9}d]$, and isospin exchanges the two for the neutron, $F_2^{en} = x[\frac{4}{9}d + \frac{1}{9}u]$; dividing gives Equation (116.25), in which the overall factor $\frac{1}{9}$ and the common x cancel. The bounds follow by setting $d = 0$ or $u = 0$. The content is that the bounds depend only on the *charges*: constituents of charge ± 1 in units of e would give a ratio bounded by different numbers, and integer-charged constituents in any assignment reproducing the proton and neutron charges cannot produce a ratio approaching $\frac{1}{4}$ at large x. This is a measurement of e_u/e_d that never requires a free quark, and it is why the deuterium running mattered as much as the hydrogen running. □

Remark 116.22 (The Gottfried sum and the asymmetric sea). The same two structure functions give a sum rule. From the charges,

$$S_G := \int_0^1 \frac{F_2^{ep}(x) - F_2^{en}(x)}{x}\,\mathrm{d}x = \frac{1}{3} + \frac{2}{3}\int_0^1 \left[\bar{u}(x) - \bar{d}(x)\right]\mathrm{d}x\,, \tag{116.26}$$

because the valence content of the proton is uud and of the neutron udd, so that $\int_0^1 u_v = 2$ and $\int_0^1 d_v = 1$ and the valence part of the difference integrates to

$$(2-1)\times\tfrac{4}{9} + (1-2)\times\tfrac{1}{9} = \tfrac{4}{9} - \tfrac{1}{9} = \tfrac{1}{3}\,,$$

the charge weights being those of Proposition 116.21; the sea survives only in the combination shown. A symmetric sea therefore predicts $S_G = \frac{1}{3}$. The measured value is significantly smaller, near 0.24, which is direct evidence that the light-quark sea is flavour-asymmetric with $\bar{d} > \bar{u}$ — an asymmetry no perturbative gluon splitting can generate, since gluons produce flavours democratically. The measurement is the New Muon Collaboration's reevaluation of the Gottfried sum from muon scattering on hydrogen and deuterium [1879]; the same asymmetry is seen directly, in a different process, by the Drell–Yan measurement of Section 116.7.3.

116.4.5 The measured numbers

Table 116.1 collects the headline quantities of the electron-scattering programme and the constants used to interpret them; Table 116.2 does the same for the neutrino measurements analysed in Section 116.5.3, and Table 116.3 for the later high-precision programme. All three sit here, under this rubric, rather than beside the sections that use them: a reader looking for the chapter's numbers should find them in one place. Every value is given with the uncertainty its source quotes, and a value inferred here from cited inputs is labelled derived, with the arithmetic shown in the text and its error propagated.

116.5 Interpretation

116.5.1 Point-like constituents

What the data establish, stated as narrowly as possible: over the range of momentum transfer covered, the charged objects from which the virtual photon scatters exhibit no form factor. Equivalently, they have no measurable size down to the resolution the experiment reaches.

Proposition 116.23 (The resolution actually achieved)**.** *A virtual photon of momentum transfer Q resolves transverse distances of order*

$$\ell = \frac{\hbar}{Q} = \frac{\hbar c}{Qc}\,, \tag{116.27}$$

Table 116.1: Headline numbers of the SLAC–MIT deep inelastic programme and the constants used to interpret them. Energies in GeV carry their joule equivalent through Equation (116.1). The value of R is the one the collaboration used in extracting νW_2 and quotes as its measured average over the deep inelastic region; its uncertainty is dominated by relative normalisation, as Proposition 116.14 explains. The structure-function integral is quoted to two figures because that is the accuracy of the review it comes from, and the uncertainty attached to it here is that implied precision and not a quoted error. Rows that record a beam setting, an angular position or an analysis cut carry no uncertainty because they define the run conditions rather than measure them; every row that reports a measured or derived quantity carries one.

Quantity	Value	Year	Source
Beam energy range (setting; calibrated to a few parts in 10^3)	4.5–20 GeV, i.e. 7.2×10^{-10}–3.2×10^{-9} J	1968	[1363, 1395]
Scattering angles, first survey (setting)	6° and 10°	1968	[1395]
Q^2c^2 covered (kinematic reach)	0.7–8 GeV^2; in SI, Q^2 from 2.0×10^{-37} to 2.3×10^{-36} kg^2 m^2/s^2	1968	[1395]
Deep inelastic cut on W (analysis cut)	$W > 2\,\mathrm{GeV}/c^2$	1969	[1397]
Elastic suppression across that Q^2 range, relative to σ_{Mott} at 10°	factor $\approx 6.3 \times 10^2$ (raw cross section $\approx 8.6 \times 10^3$); derived from the dipole, whose own 10 % accuracy makes these factors good to a few tens of percent	1956	[1088]
Inelastic-to-Mott ratio across the same range	changes by less than a factor 2	1969	[1395]
$R = \sigma_L/\sigma_T$	0.18 ± 0.10	1969–72	[1397, 1472]
$\int_0^1 F_2^{eN}(x)\,\mathrm{d}x$ (isoscalar)	0.14 ± 0.01, the error being the precision to which the review quotes it	1972	[1472]
Resolution $\hbar c/(Qc)$ at $Q^2c^2 = 8\,\mathrm{GeV}^2$	7.0×10^{-17} m (derived; exact given Q, to which the 10^{-3} beam calibration propagates)	1968	Equation (116.27)
Proton mass	$(1.672\,621\,925\,95 \pm 0.000\,000\,000\,52) \times 10^{-27}$ kg	2022	[2431]
Proton mass energy	$(0.938\,272\,088\,16 \pm 0.000\,000\,000\,29)$ GeV	2024	[2427]
Proton rms charge radius	$(8.4075 \pm 0.0064) \times 10^{-16}$ m	2022	[2431]
Fine-structure constant	$(7.297\,352\,564\,3 \pm 0.000\,000\,001\,1) \times 10^{-3}$	2022	[2431]

Table 116.2: Neutrino deep inelastic scattering: the measured total cross sections, the ratio that tests the valence hypothesis, and the momentum fractions derived from them in Proposition 116.28. Uncertainties on the derived rows are propagated from the quoted ones only and exclude the flux normalisation, which is the larger error and which the text discusses. The mean squared charge compares the neutrino integral with the electron one of Table 116.1.

Quantity	Value	Year	Source
$\sigma^{\nu N}/E_\nu$	$(7.4\pm0.2)\times10^{-43}\,\mathrm{m^2/GeV}$	1973	[1497]
$\sigma^{\bar\nu N}/E_{\bar\nu}$	$(2.8\pm0.1)\times10^{-43}\,\mathrm{m^2/GeV}$	1973	[1497]
Ratio $\sigma^{\bar\nu}/\sigma^{\nu}$	0.378 ± 0.017; $\frac{1}{3}$ for pure quarks	1973	[1497]
$C = 2Mc^2G_F^2/\pi(\hbar c)^4$	$3.164 \times 10^{-42}\,\mathrm{m^2/GeV}$ (derived; exact to the constants it is built from)	1973	Equation (116.34)
Quark momentum fraction Q	0.230 ± 0.007 (derived)	1973	Equation (116.35)
Antiquark momentum fraction $\bar Q$	0.0119 ± 0.0043 (derived)	1973	Equation (116.35)
$\int_0^1 F_2^{\nu N}\mathrm{d}x$	0.484 ± 0.011 (derived)	1973	Equation (116.35)
Mean squared charge $\int F_2^{eN}/\int F_2^{\nu N}$	0.29 ± 0.02 (derived); $5/18 = 0.278$	1972–73	[1472, 1497]
Nucleon momentum in charged constituents	0.50 ± 0.04 (derived from $\int F_2^{eN}\mathrm{d}x$ and 5/18); reported as about one half on iron	1972–79	[1472, 1639]

Table 116.3: The later high-precision programme. Rows giving a beam setting or a published kinematic coverage carry no uncertainty because they state run conditions rather than measurements. A value read from a published curve rather than from a table is marked as such and is quoted with its reading precision, which is not the collaboration's error and is always the larger of the two. The strong coupling is the world average, to which deep inelastic determinations contribute.

Quantity	Value	Year	Source
BCDMS muon beam energies (setting)	100–280 GeV	1989	[1815]
BCDMS coverage (kinematic reach)	$0.06 < x < 0.8$; $7\,\mathrm{GeV}^2 < Q^2c^2 < 260\,\mathrm{GeV}^2$	1989	[1815]
HERA beam energies (setting)	27.5 GeV on 920 GeV, $\sqrt{s}\,c = 318\,\mathrm{GeV}$	2010–15	[2152, 2244]
Equivalent fixed-target energy	54 TeV (derived; exact given the beam energies)	2010–15	Equation (116.39)
HERA combined coverage (kinematic reach)	$6 \times 10^{-7} < x < 0.65$; $0.045\,\mathrm{GeV}^2 < Q^2c^2 < 5 \times 10^4\,\mathrm{GeV}^2$	2015	[2244]
Small-x exponent λ, read from the published curves	0.1 at $Q^2c^2 = 1\,\mathrm{GeV}^2$ and 0.3 at $100\,\mathrm{GeV}^2$, each to the reading precision ± 0.05	2015	[2244]
Gluon density $[xg](2 \times 10^{-3})$ at $Q^2c^2 = 20\,\mathrm{GeV}^2$	8 ± 2 (derived from a slope read off the data; the leading-order relation is itself good only to that accuracy)	2015	Equation (116.40)
$\alpha_s(m_Z c^2)$	0.1180 ± 0.0009	2024	[2427]
$F_2^{\mathrm{Fe}}/F_2^{\mathrm{D}}$ at $x = 0.65$, read from the published ratio curve	0.89 ± 0.01 reading precision	1983	[1712]
$F_2^{\mathrm{Fe}}/F_2^{\mathrm{D}}$ at $x = 0.05$, read from the published ratio curve	1.15 ± 0.01 reading precision	1983	[1712]
$\Gamma_1^p = \int_0^1 g_1^p \mathrm{d}x$ at $\langle Q^2 \rangle\, c^2 = 10.7\,\mathrm{GeV}^2$	$0.114 \pm 0.012 \pm 0.026$; Ellis–Jaffe expectation ≈ 0.19	1988	[1804]

so at the largest momentum transfer of the original survey, $Q^2c^2 = 8\,\mathrm{GeV}^2$ *and hence* $Qc = 2.83\,\mathrm{GeV}$*,*

$$\ell = \frac{3.161\,526\,77 \times 10^{-26}\,\mathrm{J\,m}}{2.83\,\mathrm{GeV} \times 1.602\,176\,634 \times 10^{-10}\,\mathrm{J/GeV}} = 7.0 \times 10^{-17}\,\mathrm{m}\,,$$

about one twelfth of the proton's charge radius $(8.4075 \pm 0.0064) \times 10^{-16}\,\mathrm{m}$ *[2431]. The absence of a form factor at that resolution is the whole of the experimental claim: it is an upper limit on a constituent size of about* $1 \times 10^{-16}\,\mathrm{m}$*, not a demonstration that the size is zero.*

Rests on Equations (116.1) *and* (116.23)*.*

Proof. *Derives Proposition 116.23.*

The Fourier transform Equation (116.23) relates a distribution of extent ℓ to a form factor varying on the momentum scale $\hbar/\ell$; a structure smaller than $\hbar/Q$ produces a form factor still equal to one at the Q probed, and is therefore invisible. The numerical evaluation uses only Equation (116.1) and the definition of $\hbar c$. □

Remark 116.24 (What the data do not say). Three limitations are worth stating explicitly, because the result was sometimes over-read at the time and is sometimes over-read now.

First, scaling says nothing about confinement. The experiment establishes that something point-like inside the proton absorbs the photon; it says nothing whatever about whether that something can be removed from the proton, and indeed the same experiment sees only hadrons in the final state. The two facts sat uncomfortably together until asymptotic freedom explained how a theory can be free at short distance and strong at long distance (Section 116.6.1).

Second, scaling does not by itself identify the constituents as the quarks of the spectroscopy. It says there are point-like charged objects. That they carry spin $\frac{1}{2}$ comes from Section 116.4.3, that they carry fractional charge from Section 116.4.4 and Section 116.5.3, and that their number and quantum numbers match the quark model from the combination.

Third, the alternatives were serious and were killed by the data, not by taste. Vector-meson dominance, in which the virtual photon converts into a ρ meson which then interacts hadronically, predicts a cross section falling with Q^2 like the square of a vector-meson propagator and cannot give a flat ratio to the Mott cross section. Regge-based and diffractive pictures predicted specific ν and Q^2 dependences that the scaling plot excludes. It is the null test of Section 116.4.2 — flat in Q^2 at fixed ω — that discriminates, and that is why it was presented that way.

Friedman's Nobel lecture [1844] is the best first-hand account of how the interpretation was arrived at, including how long it took: the 1968 data were presented as a puzzle, and the parton reading became the consensus only over the following three or four years, as the Callan–Gross test and the neutrino comparison came in.

116.5.2 The missing momentum fraction

This is the subsection to which Phenomenon 102.47 defers the details of the extraction.

Phenomenon 116.25 (Half the proton momentum is invisible to the photon)**.** The momentum fraction carried by the electrically charged constituents, obtained by integrating the measured structure function, is about one half rather than one [1472]. Comparing neutrino with electron scattering on an isoscalar target confirms the deficit independently and returns a mean squared constituent charge close to 5/18 in units of the elementary charge squared [1497, 1639].

Rests on Equation (116.9) *and Proposition 102.46.*

Derivation. *Derives Phenomenon 116.25.*

In the parton model $F_2(x) = \sum_i e_i^2\, x\, q_i(x)$, where $q_i(x)\mathrm{d}x$ is the number of constituents of charge e_i with momentum fraction in $\mathrm{d}x$. The proton's momentum is shared among all its constituents, so

$$\sum_{\text{all } i} \int_0^1 x\, q_i(x)\, \mathrm{d}x = 1\,, \tag{116.28}$$

the sum running over *every* constituent, charged or not. What the electron measures is the charge-weighted sum: dividing $\int F_2\, \mathrm{d}x$ by the mean squared charge gives the momentum carried by the charged constituents alone, and that comes out near one half. The remainder of Equation (116.28) is carried by constituents that do not couple to the photon. The mean squared charge itself is not assumed but measured, by comparing the electron and neutrino structure functions on an isoscalar target: the weak charged current couples with the same strength to every quark flavour, so the ratio of the two integrals is the mean squared electric charge, and for equal numbers of u and d quarks

$$\left\langle e^2 \right\rangle = \tfrac{1}{2}\left[\left(\tfrac{2}{3}\right)^2 + \left(\tfrac{1}{3}\right)^2\right] = \frac{5}{18}\,,$$

which is what is found [1639]. Numerically, the neutrino integral extracted in Section 116.5.3 is $\int F_2^{\nu N}\mathrm{d}x = 0.484 \pm 0.011$ and the electron one is 0.14 ± 0.01 [1472], the second error being the precision to which the review quotes the number rather than an error it states; their ratio is 0.29 ± 0.02, to be compared with $5/18 = 0.278$. Dividing the electron integral by $5/18$ instead gives the momentum fraction in charged constituents directly, $0.14 \pm 0.01/0.278 = 0.50 \pm 0.04$ — the "about one half" of the statement above, now with the error it carries. The neutral, non-scattering carrier of the missing momentum is the gluon of Chapter 102; the inference here is from a deficit, and the direct confirmation came later, from three-jet events in electron–positron annihilation (Section 102.6.2). □

The deficit is not a defect of the measurement but a prediction of the theory, and quantum chromodynamics says where the missing momentum must eventually settle.

Proposition 116.26 (The asymptotic sharing of momentum)**.** *Under the evolution equations of Theorem 102.49, the momentum fractions carried by quarks and by gluons approach fixed values determined by n_f alone:*

$$\langle x \rangle_{q+\bar{q}} \longrightarrow \frac{3n_f}{16 + 3n_f}\,, \qquad \langle x \rangle_{g} \longrightarrow \frac{16}{16 + 3n_f}\,, \tag{116.29}$$

which for $n_f = 4$ is $\frac{12}{28} = 0.43$ and $\frac{16}{28} = 0.57$. The observed near-equal sharing is therefore not a coincidence: it is where the evolution is heading.

Rests on Theorem 102.49, Corollary 102.50, and Equation (102.62).

Proof. *Derives Proposition 116.26.*

Take the momentum moment $N = 1$ of the coupled singlet equations, using the moment notation $\gamma_N = \int_0^1 z^N P(z)\, \mathrm{d}z$ of Corollary 102.50. Write $\Sigma := \int_0^1 x \sum_q [q(x) + \bar{q}(x)]\mathrm{d}x$ and $G := \int_0^1 x\, g(x)\, \mathrm{d}x$. Four coefficients are needed:

$$A_{qq} = \int_0^1 z\, P_{qq}(z)\, \mathrm{d}z = -\tfrac{4}{3}C_F = -\tfrac{16}{9}\,, \qquad A_{qg} = 2n_f \int_0^1 z\, P_{qg}(z)\, \mathrm{d}z = \frac{n_f}{3}\,,$$

the first from Equation (102.62) and the second from $P_{qg}(z) = T_F[z^2 + (1-z)^2]$ with $T_F = \frac{1}{2}$, since $\int_0^1 z[z^2 + (1-z)^2]\mathrm{d}z = \frac{1}{3}$ and there are $2n_f$ quark and antiquark species a gluon can convert into. The remaining two follow from momentum conservation without further

computation: the total momentum $\Sigma + G$ cannot change with Q^2, so $A_{qq} + A_{gq} = 0$ and $A_{qg} + A_{gg} = 0$, giving $A_{gq} = +\frac{16}{9}$ and $A_{gg} = -n_f/3$. The evolution is then

$$\frac{\mathrm{d}}{\mathrm{d}\ln Q^2}\begin{pmatrix}\Sigma\\ G\end{pmatrix} = \frac{\alpha_s\left(Q^2\right)}{2\pi}\begin{pmatrix}-\frac{16}{9} & \frac{n_f}{3}\\ \frac{16}{9} & -\frac{n_f}{3}\end{pmatrix}\begin{pmatrix}\Sigma\\ G\end{pmatrix}. \tag{116.30}$$

The matrix has eigenvalue 0, with eigenvector $(\Sigma, G) \propto (n_f/3,\, 16/9)$, and eigenvalue $-(\frac{16}{9} + \frac{n_f}{3})$, whose eigenvector decays. Since α_s falls only logarithmically the decay is slow, but its direction is fixed: the surviving combination is the null eigenvector, normalized by $\Sigma + G = 1$, which gives Equation (116.29) after dividing through by $\frac{1}{3}$. Note that the answer does not involve α_s, Λ_{QCD} or the initial distributions — only the group theory in C_F and T_F and the number of active flavours.

Measured values at Q^2c^2 near $10\,\mathrm{GeV}^2$ put the gluon fraction near 0.45 [2427], below the asymptotic 0.57 for $n_f = 4$, and rising slowly with Q^2 as the second eigenvector dies away — which is exactly the pattern the global fits of Section 116.6.5 find. □

116.5.3 Neutrino deep inelastic scattering

The electron probes the quarks with weights e_i^2. The charged weak current probes them with weights that are independent of electric charge and that distinguish quarks from antiquarks, so the two probes together separate the charge weights from the distributions. The weak-current theory is Chapter 101; the neutrino beams and the Gargamelle chamber are described in Chapter 104, where the same exposures gave the neutral current.

Phenomenon 116.27 (The neutrino cross section rises linearly with energy)**.** The total charged-current cross sections for $\nu_\mu + N \to \mu^- + X$ and $\bar{\nu}_\mu + N \to \mu^+ + X$ on an approximately isoscalar target rise in proportion to the neutrino energy, with

$$\frac{\sigma^{\nu N}}{E_\nu} = (7.4 \pm 0.2) \times 10^{-43}\,\mathrm{m}^2/\mathrm{GeV} = (4.62 \pm 0.12) \times 10^{-33}\,\mathrm{m}^2/\mathrm{J}\,, \tag{116.31}$$

$$\frac{\sigma^{\bar{\nu} N}}{E_{\bar{\nu}}} = (2.8 \pm 0.1) \times 10^{-43}\,\mathrm{m}^2/\mathrm{GeV} = (1.75 \pm 0.06) \times 10^{-33}\,\mathrm{m}^2/\mathrm{J}\,, \tag{116.32}$$

measured in the Gargamelle bubble chamber at CERN [1497]. A linear rise is what point-like constituents require and what a target with a form factor forbids.

Rests on Equations (104.3) *and* (116.9).

Derivation. *Derives Phenomenon 116.27.*

The four-fermion charged-current interaction of Chapter 101 has coupling $G_F = (1.435\,851\,2 \pm 0.000\,000\,7) \times 10^{-62}\,\mathrm{J\,m}^3$, equivalently $G_F/(\hbar c)^3 = (1.166\,378\,7 \pm 0.000\,000\,6) \times 10^{-5}/\mathrm{GeV}^2$ (Equation (104.3)), the uncertainty being the relative 5.1×10^{-7} of the second form carried over; it is a dimensionful coupling, and that is the whole argument. For the point process $\nu + q \to \mu^- + q'$ the amplitude is proportional to G_F and carries no propagator, so the squared amplitude is proportional to G_F^2 times the only invariant available, and the cross section is

$$\hat{\sigma}\left(\nu q\right) = \frac{G_F^2\,\hat{s}c^2}{\pi\left(\hbar c\right)^4}\,, \qquad \hat{\sigma}\left(\nu \bar{q}\right) = \frac{G_F^2\,\hat{s}c^2}{3\pi\left(\hbar c\right)^4}\,, \tag{116.33}$$

where $\hat{s}$ is the squared centre-of-mass four-momentum of the neutrino–parton system. It is written here with the same dimension as Q^2, namely $\mathrm{kg}^2\,\mathrm{m}^2/\mathrm{s}^2$, so that $\hat{s}c^2$ is an energy squared and the convention is the one used for s at HERA in Proposition 116.33 and for the Drell–Yan variable in Equation (116.43); this chapter never gives s the dimension of an energy squared. The dimensions check: $G_F^2\hat{s}c^2$ carries $\mathrm{J}^4\,\mathrm{m}^6$ and $(\hbar c)^4$ carries $\mathrm{J}^4\,\mathrm{m}^4$, leaving m^2. The factor $\frac{1}{3}$ for antiquarks is the angular-momentum suppression: ν and q are both left-handed, so their total angular momentum along the beam vanishes and the scattering is

isotropic, while ν and $\bar{q}$ have $J = 1$ aligned and the $\mathrm{d}\sigma/\mathrm{d}y \propto (1-y)^2$ distribution integrates to a third.

With the struck parton carrying fraction x of the nucleon momentum, $\hat{s} = x\,s$ and $sc^2 = M^2c^4 + 2Mc^2E_\nu \approx 2Mc^2E_\nu$ for a stationary target, so integrating over the parton content gives a cross section *proportional to* E_ν:

$$\frac{\sigma^{\nu N}}{E_\nu} = C\left[Q + \frac{\bar{Q}}{3}\right], \qquad \frac{\sigma^{\bar{\nu}N}}{E_{\bar{\nu}}} = C\left[\frac{Q}{3} + \bar{Q}\right], \qquad C := \frac{2Mc^2G_F^2}{\pi\,(\hbar c)^4}, \tag{116.34}$$

where Q and $\bar{Q}$ are the momentum fractions carried by the quarks and antiquarks that the current can reach on an isoscalar nucleon. Evaluating the constant in SI,

$$\begin{aligned} C &= \frac{2 \times 1.503\,277 \times 10^{-10}\,\mathrm{J} \times \left(1.435\,851 \times 10^{-62}\,\mathrm{J\,m^3}\right)^2}{\pi\,(3.161\,526\,77 \times 10^{-26}\,\mathrm{J\,m})^4} \\ &= 1.975 \times 10^{-32}\,\mathrm{m^2/J} = 3.164 \times 10^{-42}\,\mathrm{m^2/GeV}\,. \end{aligned}$$

The linearity is thus a direct consequence of G_F having dimensions of $\mathrm{J\,m^3}$ together with the absence of any other scale in the target: a constituent with a size a would supply a form factor cutting the rise off at $E_\nu \sim \hbar^2/(2Ma^2)$, and no such cut-off is seen. □

Proposition 116.28 (The momentum fractions from the two cross sections)**.** *Solving Equation* (116.34) *for* Q *and* $\bar{Q}$ *with the measured values Equations* (116.31) *and* (116.32) *gives*

$$Q = 0.230 \pm 0.007\,, \qquad \bar{Q} = 0.0119 \pm 0.0043\,, \qquad \int_0^1 F_2^{\nu N}(x)\,\mathrm{d}x = 2\left(Q + \bar{Q}\right) = 0.484 \pm 0.011\,. \tag{116.35}$$

About half the nucleon's momentum is therefore carried by objects the weak current can reach, and the antiquark "sea" accounts for $(4.9 \pm 1.8)\,\%$ *of that at leading order — a fraction whose error is more than a third of itself, which is why no more than its first figure may be used.*

Rests on Equations (116.31), (116.32) *and* (116.34).

Proof. *Derives Proposition 116.28.*

Write $a := \sigma^{\nu N}/(CE_\nu) = Q + \bar{Q}/3$ and $b := \sigma^{\bar{\nu}N}/(CE_{\bar{\nu}}) = Q/3 + \bar{Q}$. From the measured values and the constant C just evaluated,

$$a = \frac{7.4 \times 10^{-43}}{3.164 \times 10^{-42}} = 0.2339\,, \qquad b = \frac{2.8 \times 10^{-43}}{3.164 \times 10^{-42}} = 0.0885\,.$$

Adding, $a + b = \frac{4}{3}(Q + \bar{Q})$, so $Q + \bar{Q} = \frac{3}{4}(a + b) = 0.2418$; and subtracting a from $3b$ gives $\frac{8}{3}\bar{Q} = 3b - a = 0.0316$, whence $\bar{Q} = 0.0119$ and $Q = 0.2299$. On an isoscalar nucleon the neutrino structure function is $F_2^{\nu N}(x) = x\sum_q [q(x) + \bar{q}(x)]$ with unit weights, and the sum over both nucleon types doubles the per-current fractions, so $\int F_2^{\nu N}\mathrm{d}x = 2(Q + \bar{Q}) = 0.484$.

The errors propagate from $\delta a = 0.006\,32$ and $\delta b = 0.003\,16$, which are the quoted relative errors of Equations (116.31) and (116.32) applied to a and b. Since $Q = \frac{9}{8}a - \frac{3}{8}b$ and $\bar{Q} = \frac{3}{8}(3b - a)$,

$$\delta Q = \sqrt{\left(\tfrac{9}{8}\delta a\right)^2 + \left(\tfrac{3}{8}\delta b\right)^2} = 0.0072\,, \qquad \delta\bar{Q} = \tfrac{3}{8}\sqrt{(3\delta b)^2 + (\delta a)^2} = 0.0043\,,$$

and $\delta\int F_2^{\nu N}\mathrm{d}x = \frac{3}{2}\sqrt{(\delta a)^2 + (\delta b)^2} = 0.011$. Note how much worse the antiquark fraction is determined than the sum: the difference $3b - a$ subtracts two numbers of comparable size, so its error is essentially the full error of a.

Three caveats belong with the number, and each is a real limitation. The uncertainties propagated are the collaboration's quoted ones and do not include the neutrino flux normalisation, which was uncertain at the 10 % level and would widen the result to about 0.48 ± 0.05 — which is why the value usually quoted from this era, 0.49 ± 0.07, carries a much larger error than the arithmetic above. The Gargamelle target was freon, not an exactly isoscalar nucleus, and the neutron excess was corrected for. And Equation (116.33) is leading order in α_s, so the split between Q and $\bar{Q}$ in particular is not to be taken past its first figure: higher-order and strange-quark contributions move the sea fraction upwards. □

The numbers extracted here are collected, with the measured cross sections they come from, in Table 116.2.

The higher-statistics continuation was CDHS, which exposed a magnetized-iron calorimeter and spectrometer to the CERN SPS narrow-band neutrino beam and measured F_2 and xF_3 over a wide range of x and Q^2 [1639]. Two of its results matter here. It confirmed the mean squared charge of $5/18$ by direct comparison of its own $F_2^{\nu N}$ with the charged-lepton measurements, which is the content of Proposition 102.46; and it separated the antiquark content, whose momentum fraction is small and concentrated at small x, confirming that the momentum deficit of Phenomenon 116.25 cannot be repaired by a larger sea. The parity-violating structure function xF_3, which exists only for a weak probe, counts valence quarks.

Proposition 116.29 (The Gross–Llewellyn Smith sum rule in the parton model)**.** *On an isoscalar nucleon the third charged-current structure function integrates to the number of valence quarks,*

$$\int_0^1 F_3^{\nu N}(x)\,\mathrm{d}x = 3\,, \tag{116.36}$$

a pure number, and its measurement is a count of the valence quarks in a nucleon obtained without ever isolating one.

Rests on Equation (116.33) *and Proposition 102.6.*

Derivation. *Derives Proposition 116.29.*

The charged current is $V - A$: it couples to left-handed quarks and to right-handed antiquarks, and it is the interference of its vector with its axial part that produces a third structure function at all. That interference term changes sign between a quark and an antiquark — which is why F_3 exists only for a parity-violating probe and vanishes identically for the photon of Equation (116.19) — so at leading order in the parton model

$$F_3^{\nu N}(x) = \sum_q \left[q(x) - \bar{q}(x)\right]\,,$$

with unit weights, since the current is blind to electric charge. The sea contributes a quark and an antiquark at the same x and therefore cancels in the difference, leaving only the valence densities: $\int_0^1 [q - \bar{q}]\mathrm{d}x = \int_0^1 q_v \mathrm{d}x$. Averaged over the proton (uud) and the neutron (udd) of an isoscalar target the valence content is the same three quarks either way, so

$$\int_0^1 F_3^{\nu N}\,\mathrm{d}x = \int_0^1 \left[u_v(x) + d_v(x)\right]\mathrm{d}x = 2 + 1 = 3\,,$$

which is Equation (116.36). Note what makes this a stronger statement than the momentum sum rule of Equation (116.28): no unknown weight enters, so the prediction is an integer and not a fraction to be compared with a measured one. The quantity is dimensionless, F_3 being a function of x alone. □

The measured value is smaller than three, and the deficit is itself calculable: gluon radiation off the struck quark reduces the integral by a factor $1 - \alpha_s(Q^2)/\pi - \ldots$, which

turns the sum rule into a determination of the strong coupling at the Q^2 of the measurement. That correction is what remains owed here.

Derivation pending #208

The perturbative correction to the Gross–Llewellyn Smith sum rule. The parton-model value of 3 is derived in the text above; what is still owed is the calculation of the coefficient function that multiplies it by $1 - \alpha_s/\pi - \dots$, order by order, and the demonstration that the same coefficient function is the one appearing in the Bjorken sum rule for polarized scattering. That is what makes a measurement of this integral a competitive determination of the strong coupling, and it belongs in Appendix A alongside the other sum rules.

116.5.4 The 1990 Nobel Prize

The 1990 Nobel Prize in Physics was awarded to Jerome I. Friedman, Henry W. Kendall and Richard E. Taylor for their pioneering investigations concerning deep inelastic scattering of electrons on protons and bound neutrons, of essential importance for the development of the quark model in particle physics.

The three lectures are complementary and all three are worth reading as primary material rather than as summaries. Taylor's [1849] is the account of the accelerator and the spectrometers — how End Station A was built and what it could and could not do. Kendall's [1848] is the account of the measurements themselves and of the analysis that turned raw yields into νW_2, including the radiative corrections. Friedman's [1844] is the account of the interpretation: which alternatives were on the table, what killed each of them, and how long the parton reading took to become the consensus. Taken together they are an unusually candid record of an experiment whose result was understood only some years after it was obtained.

116.6 Scaling violations and the confirmation of QCD

116.6.1 Asymptotic freedom

Scaling is the statement that the constituents are free. A field theory in which they are genuinely free has no interactions and no bound states; the parton model was therefore, as Remark 116.7 says, a hypothesis without dynamics. What was needed was a theory whose coupling *vanishes* at short distance, and in 1973 Gross and Wilczek [1501] and, independently, Politzer [1511] showed that a non-abelian gauge theory is such a theory and, within the class of renormalizable field theories in four dimensions, essentially the only one.

The result is Theorem 102.35: the one-loop beta function of an SU(3) gauge theory with n_f quark flavours has the coefficient $b_0 = 11 - \tfrac{2}{3}n_f$, positive for $n_f \le 16$, so the coupling runs to zero logarithmically at large momentum transfer. The measured curve is Phenomenon 102.41, with $\alpha_s(m_Z c^2) = 0.1180 \pm 0.0009$ [2427] and the values at lower scales tabulated in Table 102.2.

Two consequences for this chapter follow immediately, and they pull in opposite directions, which is what makes the resulting prediction sharp.

- *Scaling must hold approximately.* At large Q^2 the coupling is small, the struck quark is nearly free during the absorption, and the parton model is the leading term of a systematic expansion rather than a guess.
- *Scaling must fail logarithmically.* The coupling is not zero, so a quark may radiate a gluon before or after being struck. Raising Q^2 resolves emissions that a coarser probe

could not separate, so the number of partons seen rises and each carries less momentum. Structure functions must therefore fall with $\ln Q^2$ at large x, where partons are lost to radiation, and rise at small x, where the radiated partons accumulate.

The second is a prediction with no free parameter beyond α_s itself, and it is what the rest of this section reports as measured.

116.6.2 The evolution equations

The equations that make the second statement quantitative were derived by Gribov and Lipatov [1474], by Altarelli and Parisi [1591] and by Dokshitzer [1595], and are Theorem 102.49; the splitting functions come from the collinear emission probability Equation (102.58), and the Mellin diagonalisation that turns the integro-differential system into ordinary differential equations for the moments is Corollary 102.50. None of that is rebuilt here. What belongs here is the experimental use to which it is put.

Proposition 116.30 (How the strong coupling is read off a slope)**.** *For a non-singlet combination of parton distributions, the moment $M_N = \int_0^1 f(x) x^N \mathrm{d}x$ obeys Equation (102.60), so*

$$\alpha_s\left(Q^2\right) = \frac{2\pi}{\gamma_N} \frac{\mathrm{d}\ln M_N}{\mathrm{d}\ln Q^2}\,, \tag{116.37}$$

with γ_N the rational number Equation (102.62). The strong coupling is therefore measured by a rate of change *of the data, not by a rate; no absolute normalisation and no knowledge of the parton distributions enters.*

Rests on Equations (102.60) and (102.62).

Proof. *Derives Proposition 116.30.*

Equation (102.60) reads $\mathrm{d}M_N/\mathrm{d}\ln Q^2 = (\alpha_s/2\pi)\gamma_N M_N$; dividing by M_N and solving for α_s gives Equation (116.37). For the second moment, with $C_F = \frac{4}{3}$ and $H_m = \sum_{k\le m} 1/k$,

$$\gamma_2 = C_F\left[\tfrac{3}{2} - H_2 - H_4\right] = \tfrac{4}{3}\left[1.5 - 1.5 - 2.0833\right] = -2.778\,,$$

so at $\alpha_s = 0.20$ the moment falls at $\mathrm{d}\ln M_2/\mathrm{d}\ln Q^2 = -0.088$, that is by about 20 % per decade in Q^2. Measuring that slope to 10 % determines α_s to 10 %.

The parameter-free version of the test is the ratio of two slopes. With $\gamma_1 = -\frac{16}{9} = -1.778$ and $\gamma_3 = C_F[\frac{3}{2} - H_3 - H_5] = -3.489$,

$$\frac{\mathrm{d}\ln M_3/\mathrm{d}\ln Q^2}{\mathrm{d}\ln M_1/\mathrm{d}\ln Q^2} = \frac{\gamma_3}{\gamma_1} = 1.96\,, \tag{116.38}$$

a pure number containing neither α_s nor Λ_{QCD} nor the initial distributions — the content of Remark 102.52. Plotting one measured moment against another on logarithmic axes must give a straight line of that slope. □

Remark 116.31 (What the equations do not predict). The evolution equations say how the distributions change, not what they are. The distributions at some reference scale Q_0 are non-perturbative objects and must be measured; QCD supplies no prediction for them, and any claim that it does is a claim about a model and not about the theory. This is not a weakness of the test but what makes it stringent: the initial distributions are fitted at one scale, and then several thousand data points at other scales must be described with no further freedom beyond the single number α_s.

116.6.3 Muon-beam measurements

The lever arm in Q^2 needed to see a logarithm requires beam energies well beyond SLAC's. Muon beams supplied them: a muon is a charged lepton, so the process and its interpretation are identical to electron scattering, and a muon beam produced from pion decay can be made at several hundred GeV where an electron beam of the same energy cannot.

The Bologna–CERN–Dubna–Munich–Saclay collaboration (BCDMS) exposed a long toroidal-iron spectrometer with internal hydrogen, deuterium and carbon targets to the CERN SPS muon beam at energies of order 100–280 GeV, and measured $F_2(x, Q^2)$ and R with high statistics over x from about 0.06 to 0.8 and Q^2c^2 from about $7\,\mathrm{GeV}^2$ to $260\,\mathrm{GeV}^2$ [1815] — a factor of nearly forty in Q^2, against the factor of eight available in 1969. The European Muon Collaboration made companion measurements over an overlapping range, and its data on nuclear targets produced the separate result of Section 116.7.1.

Phenomenon 116.32 (Scaling is violated logarithmically)**.** Measured over a far wider range of Q^2 than the original experiment, the structure function $F_2(x, Q^2)$ is not exactly a function of x alone: at large x it decreases with $\log Q^2$ and at small x it increases, the two movements being tied to each other [1815]. At the electron–proton collider the rise at small x is dramatic, and F_2 continues to grow down to x below 10^{-4} [2152, 2244]. This pattern, and not scaling itself, is the quantitative confirmation of quantum chromodynamics, because the theory predicts not the structure functions but exactly their rate of change.

Rests on Phenomenon 116.18, Proposition 102.48, and Equation (102.59)*.*

Derivation. Derives Phenomenon 116.32.

The mechanism is Proposition 102.48: a quark of fraction y radiates a gluon and is left with $x = zy < y$, with probability $(\alpha_s/2\pi)P_{qq}(z)\mathrm{d}z\,\mathrm{d}k_T^2/k_T^2$ per unit transverse momentum squared. Integrating the last factor up to the resolution Q^2 produces $\ln Q^2$ and nothing steeper, which is why the violation is logarithmic and not a power — the distinguishing signature against a constituent size, which by Equation (116.23) would give a power. Assembling the gain and loss terms gives Equation (102.59), whose sign structure is the statement in the phenomenon: the plus prescription in Equation (102.58) makes $\int_0^1 P_{qq} = 0$, so whatever is lost at large x reappears at smaller x, and the two movements are not independent but are the same probability counted twice. The crossing point, where $\mathrm{d}F_2/\mathrm{d}\ln Q^2$ changes sign, is near $x \approx 0.15$ in the data and is itself predicted.

The quantitative comparison is Proposition 116.30 applied to the measured moments; the collaboration's own QCD analysis of its non-singlet data returned a strong coupling at the low end of the range now accepted, and the modern world average from all methods is $\alpha_s(m_Zc^2) = 0.1180 \pm 0.0009$ [2427], with deep inelastic determinations among the inputs. □

116.6.4 HERA

This is the subsection to which Remark 102.52 defers the experimental side of the DGLAP test.

HERA, at DESY in Hamburg, was the only electron–proton collider ever built: a ring of about 6.3 km circumference in which 27.5 GeV electrons or positrons collided with protons of 920 GeV (820 GeV in the earlier running). Two general-purpose detectors, H1 and ZEUS, recorded the collisions.

Proposition 116.33 (What colliding changes)**.** *The centre-of-mass energy of the HERA collisions is*

$$\sqrt{s}\,c \approx 2\sqrt{E_eE_p} = 2\sqrt{27.5\,\mathrm{GeV} \times 920\,\mathrm{GeV}} = 318\,\mathrm{GeV}\,,$$

neglecting the particle masses. A fixed-target experiment would need a beam energy

$$E_{\mathrm{lab}} = \frac{sc^2}{2Mc^2} = \frac{(318\,\mathrm{GeV})^2}{2 \times 0.938\,27\,\mathrm{GeV}} = 54\,\mathrm{TeV} \tag{116.39}$$

to reach the same invariant mass. Since $Q^2c^2 \le sc^2$ and $x \ge Q^2c^2/(sc^2)$ at $y \le 1$, the accessible region extends to Q^2c^2 of order $1 \times 10^5\,\mathrm{GeV}^2$ and to x some four orders of magnitude below anything a fixed target can reach.

Rests on Notation 116.13.

Proof. *Derives Proposition 116.33.*

For head-on beams of energies E_e and E_p and negligible masses, the total four-momentum has energy $E_e + E_p$ and momentum $(E_p - E_e)/c$, so $sc^2 = (E_e + E_p)^2 - (E_p - E_e)^2 = 4E_eE_p$, giving the first line; s carries $\mathrm{kg}^2\,\mathrm{m}^2/\mathrm{s}^2$, the dimension of Q^2. For a stationary target of mass M, $sc^2 = M^2c^4 + 2Mc^2E_{\mathrm{lab}}$, which for $E_{\mathrm{lab}} \gg Mc^2$ inverts to Equation (116.39). The kinematic bound follows from $Q^2 = xys$ with $x, y \in [0, 1]$ — both sides carrying $\mathrm{kg}^2\,\mathrm{m}^2/\mathrm{s}^2$ — so $Q^2c^2 \le sc^2$ and, at fixed Q^2, small x requires large s — which is the whole reason a collider was built for this measurement. □

The combination of the two experiments' inclusive cross sections is the definitive data set: about 2900 measured points covering Q^2c^2 from about $0.045\,\mathrm{GeV}^2$ to $5 \times 10^4\,\mathrm{GeV}^2$ and x from about 6×10^{-7} to 0.65, with total uncertainties reaching the percent level in the bulk of the range [2152, 2244]. Combining the two experiments before fitting is itself part of the result: the cross-calibration reduces correlated systematics that neither experiment could remove alone.

Phenomenon 116.34 (The structure function rises steeply at small x)**.** At fixed Q^2, $F_2(x, Q^2)$ grows as x decreases, with an effective power $F_2 \sim x^{-\lambda}$ whose exponent itself grows with Q^2: λ is near 0.1 at $Q^2c^2 \approx 1\,\mathrm{GeV}^2$ and near 0.3 at $Q^2c^2 \approx 100\,\mathrm{GeV}^2$. The growth continues down to the smallest x measured, about 6×10^{-7}, with no sign of saturation [2152, 2244].

Rests on Equation (102.59) and Proposition 116.33.

Derivation. *Derives Phenomenon 116.34.*

At small x the quark distributions are dominated by gluon splitting, so the driving term of Equation (102.59) is $P_{qg} \otimes g$, and each unit of $\ln Q^2$ feeds gluons into quarks at small x while each step down in x collects the contributions of all larger fractions. Iterating the two effects generates a distribution rising as a power of $1/x$ with an exponent that grows with the evolution length, hence with $\ln Q^2$; that is the observed pattern. Quantitatively the rise is not an input to the fit but an output: the gluon distribution is fixed by the Q^2 slopes at moderate x and the small-x rise then follows. □

Proposition 116.35 (The gluon density is measured, not assumed)**.** *At small x the slope of the structure function measures the gluon distribution directly:*

$$\frac{\partial F_2\left(x, Q^2\right)}{\partial \ln Q^2} \approx \frac{10\,\alpha_s\left(Q^2\right)}{27\pi}\,[xg]\left(2x, Q^2\right)\,, \tag{116.40}$$

for $n_f = 4$ active flavours. The gluon density, which couples to neither the photon nor the weak current and is therefore invisible to every direct measurement in this chapter, is thus read off the rate of change of something that is measured.

Rests on Equations (102.59) and (116.9).

Proof. *Derives Proposition 116.35.*

Keep only the gluon term of Equation (102.59) and weight it by the squared charges, since $F_2 = \sum_q (e_q/e)^2 x[q + \bar{q}]$ and a gluon splits into q and $\bar{q}$ alike:

$$\frac{\partial F_2}{\partial \ln Q^2} = \frac{\alpha_s}{2\pi} \cdot 2 \sum_q \left(\frac{e_q}{e}\right)^2 x \int_x^1 \frac{\mathrm{d}y}{y}\, P_{qg}\left(\frac{x}{y}\right) g\left(y\right)\,,$$

with $\sum_q (e_q/e)^2 = \frac{4}{9} + \frac{1}{9} + \frac{1}{9} + \frac{4}{9} = \frac{10}{9}$ for $n_f = 4$. Substitute $z = x/y$, so $\mathrm{d}y/y = -\mathrm{d}z/z$ and $y = x/z$; then

$$x \int_x^1 \frac{\mathrm{d}y}{y} P_{qg}\!\left(\frac{x}{y}\right) g\,(y) = \int_x^1 \mathrm{d}z\, P_{qg}(z)\; [yg]\!\left(\frac{x}{z}\right) ,$$

because $x\,g(x/z)/z = [yg](x/z)$. At small x the combination $yg(y)$ varies slowly over the support of P_{qg}, so replace it by its value at the mean splitting fraction

$$\langle z \rangle = \frac{\int_0^1 z\, P_{qg}(z)\,\mathrm{d}z}{\int_0^1 P_{qg}(z)\,\mathrm{d}z} = \frac{1/6}{1/3} = \frac{1}{2} ,$$

that is, at $y = x/\langle z \rangle = 2x$. What remains is $\int_0^1 P_{qg} = \frac{1}{3}$, so

$$\frac{\partial F_2}{\partial \ln Q^2} \approx \frac{\alpha_s}{2\pi} \times 2 \times \frac{10}{9} \times \frac{1}{3} \times [xg](2x) = \frac{10\alpha_s}{27\pi}\,[xg](2x) ,$$

which is Equation (116.40). Numerically: at $x = 1 \times 10^{-3}$ and $Q^2c^2 = 20\,\mathrm{GeV}^2$, so $Qc = 4.5\,\mathrm{GeV}$ and $\alpha_s \approx 0.21$ from Table 102.2, a measured slope of about 0.2 gives

$$[xg]\!\left(2 \times 10^{-3}\right) \approx \frac{27\pi \times 0.2}{10 \times 0.21} \approx 8 ,$$

which is the size the global fits of Section 116.6.5 return at that fraction and scale. The relation is leading order and the expansion about $\langle z \rangle$ is crude; a real extraction fits the full evolution to all the data at once. What it exhibits, and what a fit conceals, is that the gluon enters the measurement only through a derivative. □

The numbers of the whole later programme, HERA's included, are collected in Table 116.3.

116.6.5 Parton distributions today

The measurements of this chapter are not archived as measurements. They are absorbed, together with fixed-target, Drell–Yan and collider data, into global fits that determine the parton distributions $f_i(x, \mu_F^2)$ with uncertainty bands, and it is those objects that the rest of particle physics uses [2244, 2427].

The property that makes them usable is collinear factorization, Theorem 102.53: the cross section for any process with a large scale splits into a calculable partonic part and a set of distributions that depend on the hadron and not on the process. The distributions are therefore *universal*, and the whole enterprise rests on that. Measured in electron–proton scattering at $Qc \sim 10\,\mathrm{GeV}$, evolved by Equation (102.59) to $Qc \sim 1\,\mathrm{TeV}$, and used to predict proton–proton cross sections at the LHC, they work — including for the Higgs production rates, which belong to Chapter 113 and which are unpredictable without them. Universality across two different initial states and three decades of scale is what makes factorization a theorem with evidence rather than a modelling assumption.

Three cautions belong with any quoted parton distribution, and Remark 102.54 states the first of them: a distribution is defined only together with a factorization scheme and a scale, and quoting one without both is meaningless. The second is that the uncertainty bands are not measurement errors in the ordinary sense; they combine experimental errors with the freedom in the functional form assumed at the starting scale, and different collaborations' bands are not independent. The third is that the fits are least well constrained exactly where they are most used at a hadron collider: the gluon at large x, and every distribution at x below the HERA reach.

116.7 Extensions and modern precision

116.7.1 The EMC effect

Phenomenon 116.36 (The EMC effect)**.** The structure function per nucleon measured on iron differs from the one measured on deuterium by up to about 15 %, with a characteristic dependence on x [1712]. Quark distributions are therefore modified by the nuclear medium itself, and not merely smeared by the Fermi motion of nucleons that would otherwise be unchanged. No agreed explanation exists.

Rests on Definition 102.44 and Equation (116.9).

The measurement was not the one the collaboration set out to make: iron was the target material of a spectrometer built for other purposes, and the deuterium comparison was a control. The observed ratio $F_2^{\mathrm{Fe}}/F_2^{\mathrm{D}}$ per nucleon falls below unity through the valence region, reaching about 0.89 near $x = 0.65$, and rises above unity at small x, reaching about 1.15 near $x = 0.05$ [1712]. Later measurements on many nuclei established that the shape is universal and the magnitude grows slowly with mass number, and that the ratio turns upward again above $x \approx 0.8$, where Fermi motion allows a bound nucleon to carry more than the whole nucleon momentum fraction.

Why this is a result rather than a nuisance: before 1983 it was generally supposed that a nucleus is a collection of nucleons whose internal structure is what it is in free space, so that nuclear effects on F_2 would be confined to the smearing produced by the nucleons' Fermi motion — an effect calculable from nuclear physics (Chapter 107) and confined to large x. The observed modification extends over the whole valence region and has the wrong sign at moderate x for Fermi smearing. Whatever the mechanism, the quark momentum distribution inside a bound nucleon is not the one inside a free nucleon.

Derivation pending #209

The nuclear modification of parton distributions, stated as the open problem it is: no derivation from quantum chromodynamics currently reproduces the observed dependence on the momentum fraction. The candidate mechanisms — nucleon swelling, pion-cloud enhancement, multi-quark clusters, and a dependence on the local nuclear density rather than on the mass number — each account for part of the shape and none for all of it, and they are not mutually exclusive. What is owed in Appendix A is a statement of what each predicts and of which measurements discriminate among them.

Remark 116.37 (An honest limit on the sources here). The bibliography of this book carries the original EMC measurement and nothing later on nuclear parton distributions. The statements above about the universality of the shape across nuclei, about its growth with mass number and about the local-density correlation are therefore made without a citation that supports them, and a reader who wants to check them must go outside this book's reference list. They are recorded here because omitting them would leave the 1983 result looking more isolated than it is.

116.7.2 The proton spin

Scattering polarized electrons or muons from a polarized target gives access to a third structure function, $g_1(x, Q^2)$, which in the parton model counts the quark spins:

$$g_1(x) = \frac{1}{2}\sum_q \left(\frac{e_q}{e}\right)^2 \left[\Delta q(x) + \Delta\bar{q}(x)\right] , \tag{116.41}$$

with $\Delta q(x)$ the difference between the densities of quarks with spin along and against the proton's spin. The integral $\Gamma_1^p = \int_0^1 g_1^p(x)\,\mathrm{d}x$ is therefore a charge-weighted count of how much of the proton's spin the quark spins carry.

Phenomenon 116.38 (The quark spins do not account for the proton spin)**.** The European Muon Collaboration measured $\Gamma_1^p = 0.114 \pm 0.012 \pm 0.026$ at $\langle Q^2 \rangle\, c^2 = 10.7\,\mathrm{GeV}^2$, where the first uncertainty is statistical and the second systematic [1804]. The naive expectation, obtained by combining the octet axial charges measured in hyperon beta decay with the assumption that the strange sea is unpolarized — the Ellis–Jaffe sum rule — is about 0.19; that number is quoted here without derivation and without a source in this book's bibliography, and the debt is recorded immediately below. The quark spins therefore carry a small and possibly vanishing fraction of the proton's angular momentum, in flat contradiction with the constituent quark model, in which they carry all of it.

Rests on Equation (116.41) *and Proposition 102.6.*

Derivation pending #210

The Ellis–Jaffe sum rule, the theoretical number the EMC measurement contradicts. What is owed is the derivation of the expectation for the first moment of the polarized proton structure function from the octet axial charges: the isovector combination fixed by the neutron beta-decay ratio, the octet combination fixed by the hyperon semileptonic decays under the flavour symmetry of the baryon octet, and the assumption of an unpolarized strange sea that sets the singlet piece equal to the octet one — together with the leading perturbative correction, which lowers the parton-model value. Only that chain shows why the expected number is near 0.19 *and which of its assumptions the measurement actually falsifies. It belongs in Appendix A beside the other sum rules; this book's bibliography carries no entry for the original Ellis–Jaffe paper.*

The resolution, so far as there is one, is that the proton's spin decomposes as

$$\frac{1}{2} = \frac{1}{2}\Delta\Sigma + \Delta G + L_q + L_g\,, \tag{116.42}$$

in units of $\hbar$, where $\Delta\Sigma$ is the quark spin contribution measured above, ΔG the gluon spin, and L_q, L_g the orbital angular momenta of quarks and gluons (Chapter 79). Nothing requires the first term to dominate. Subsequent polarized experiments have measured $\Delta\Sigma$ to be roughly 0.3 and have constrained ΔG to be small over the measured range of momentum fraction; the orbital terms are essentially unmeasured, and are among the principal targets of the programme in Section 116.7.4.

Derivation pending #211

The decomposition of the proton spin into quark spin, gluon spin and orbital contributions, and the reason the split is ambiguous. What is owed is the demonstration that the gauge-invariant decomposition of the angular momentum of a gauge theory into a "spin" and an "orbital" part is not unique — different decompositions differ by terms that vanish for a free field and do not vanish here — so that a measured ΔG has meaning only relative to a stated convention and factorization scheme. It belongs in Appendix A with the operator definitions written out.

Remark 116.39 (Sources for the modern spin measurements). This book's bibliography carries the original EMC measurement and no later polarized-scattering result. The values quoted above for $\Delta\Sigma$ and for the constraint on ΔG are therefore uncited, and are flagged as such.

116.7.3 The Drell–Yan process and the sea

Deep inelastic scattering is not the only way to measure a parton distribution, and the second way is the test of whether the distributions mean anything. Drell and Yan proposed that in a hadron collision a quark from one hadron and an antiquark from the other annihilate into a virtual photon, which materializes as a lepton pair of large invariant mass [1421].

Proposition 116.40 (The Drell–Yan cross section scales)**.** *For the annihilation of a quark of fraction x_1 from one hadron with an antiquark of fraction x_2 from the other, the lepton pair has invariant mass $M_{\ell\ell}c^2$ — written with a subscript because M without one is the proton mass throughout this chapter (Notation 116.13) — with*

$$\tau := \frac{M_{\ell\ell}^2 c^4}{s c^2} = x_1 x_2 \,, \tag{116.43}$$

and the cross section takes the form

$$M_{\ell\ell}^3 c^6 \, \frac{\mathrm{d}\sigma}{\mathrm{d}\left(M_{\ell\ell}c^2\right)} = F\left(\tau\right) \,, \tag{116.44}$$

a function of the dimensionless ratio τ alone. Here s carries $\mathrm{kg}^2\,\mathrm{m}^2/\mathrm{s}^2$ as in Proposition 116.33, so τ is dimensionless, and F carries $\mathrm{J}^2\,\mathrm{m}^2$, since $M_{\ell\ell}^3c^6$ is a cubed energy and $\mathrm{d}\sigma/\mathrm{d}(M_{\ell\ell}c^2)$ carries m^2/J. This is Bjorken scaling in a different process, with the same variables and the same distributions.

Rests on Theorem 102.53 and Equation (116.9)*.*

Proof. *Derives Proposition 116.40.*

The partonic process $q\bar{q} \to \ell^+\ell^-$ is the crossed channel of $\ell^+\ell^- \to q\bar{q}$ and its cross section is the point-like $\hat{\sigma} = 4\pi\alpha^2(\hbar c)^2(e_q/e)^2/(3M_{\ell\ell}^2c^4) \times \frac{1}{3}$, the last factor because the quark and antiquark must match in colour, one chance in three. It contains no scale but $M_{\ell\ell}$. The invariant mass of a pair made from fractions x_1 and x_2 of two hadrons whose squared centre-of-mass four-momentum is s is $M_{\ell\ell}^2c^4 = x_1x_2sc^2$, which is Equation (116.43). Folding $\hat{\sigma}$ with the two distributions and changing variables from (x_1, x_2) to (τ, x_1) produces a factor $1/M_{\ell\ell}^3$ from the partonic cross section and the Jacobian, leaving a function of τ alone, which is Equation (116.44). The measured cross section obeys it, and — the point of the exercise — the distributions extracted from it are the same ones extracted from deep inelastic scattering, which is the experimental content of Theorem 102.53. □

The process is also the sharpest probe of the antiquark sea, because it requires an antiquark in the initial state, and comparing hydrogen with deuterium targets isolates $\bar{d}/\bar{u}$. Measurements of that ratio find it significantly greater than one over a range of x, in agreement with the Gottfried sum of Remark 116.22 and in contradiction with any picture in which the sea comes from perturbative gluon splitting, which cannot distinguish flavours. The measurement is the Fermilab E866/NuSea experiment [2020], which compares Drell–Yan yields from hydrogen and deuterium targets; it and the Gottfried-sum determination [1879] are independent determinations of the same asymmetry, one in lepton–nucleon scattering and one in hadron–hadron collisions.

116.7.4 Continuing programmes

Deep inelastic scattering is not a closed subject, and it is worth being precise about which of the things still being measured would count as a new result and which as a refinement.

Refinements. Extending F_2 and g_1 to smaller x and to larger Q^2; reducing the uncertainty on the gluon at large x, which currently limits several LHC predictions; mapping the nuclear modification of Section 116.7.1 across the periodic table; improving the determination of

α_s from scaling violations. All of these tighten numbers inside a framework that is not in question.

New results. Three outcomes would be different in kind.

- *A measured violation of factorization.* If the distributions extracted from one process failed to describe another at a level exceeding the calculable higher-order corrections, Theorem 102.53 would be false as applied, and the entire apparatus by which hadron-collider rates are predicted would lose its foundation.

- *A gluon density inconsistent with the evolution equations.* The rise of Phenomenon 116.34 cannot continue indefinitely — at some small enough x the density of gluons must be high enough that they recombine as well as split, and the linear evolution of Section 116.6.2 must fail. Observing that non-linear regime would be the first measurement of a genuinely new dynamical regime of QCD rather than a refinement of the known one. Whether HERA reached its onset is disputed.

- *The orbital angular momentum of the proton's constituents.* Equation (116.42) has two terms that have never been measured. Determining them requires correlating a parton's momentum fraction with its transverse position or transverse momentum, which inclusive scattering cannot do.

The instrument designed for the last two is the Electron–Ion Collider now under construction at Brookhaven, which will collide polarized electrons with polarized protons and light ions and with heavy nuclei at centre-of-mass energies from about 20–140 GeV and luminosities of order $1 \times 10^{38}/(\mathrm{m}^2\,\mathrm{s})$, some three orders of magnitude above HERA's [2265]. Its observables are the exclusive and semi-inclusive ones — deeply virtual Compton scattering, tagged final-state hadrons — from which a transverse spatial or transverse momentum image of the parton distribution can be built, and the nuclear targets that would show saturation if it is there. It is a continuation of the measurement reported in this chapter with the same beam, the same probe and the same variables, on an instrument fifty years newer.

116.8 Primary references

The experiment itself is contained in two Physical Review Letters of 1969 submitted back to back. Bloom *et al.* [1395] report the inelastic cross sections at 6° and 10° and the observation that the ratio to the Mott cross section falls far more slowly with momentum transfer than the elastic form factors do; Breidenbach *et al.* [1397] report the structure function νW_2 and its collapse onto a single curve. Neither is long, and both should be read in preference to any account of them. The public announcement preceded both, at the Vienna conference of 1968, in Panofsky's rapporteur talk [1379]. The most useful single secondary source on the apparatus, the corrections and the whole 1967–1972 data set is Friedman and Kendall's review [1472], which is where the numbers in Table 116.1 that are not from the two Letters come from.

The instrument has its own literature. Neal's edited volume on the Stanford two-mile accelerator [1363] is the engineering account of the machine. The three Nobel lectures [1844, 1848, 1849] divide the story between apparatus, measurement and interpretation, and are the best first-hand record of what the collaboration believed at each stage. Mo and Tsai [1404] is the radiative-correction calculation on which every number in the experiment depends; it is a piece of applied quantum electrodynamics rather than of hadron physics, and it is the largest single theoretical input to the measured cross sections.

The theory the experiment tested was in print before the data. Rosenbluth [970] gives the elastic cross section that defines the form factors; Hofstadter [1088] is the review of the

elastic measurements that established the proton's size. Gell-Mann [1277] and Zweig [1301] propose the quarks; Zweig's is an unpublished CERN preprint and is cited as such. Bjorken [1393] predicts scaling from current algebra and Wilson [1410] supplies the operator-product framework; Feynman [1401] gives the parton picture and Bjorken and Paschos [1394] the parton-distribution formulation used throughout this chapter. Callan and Gross [1398] give the spin discriminator of Section 116.4.3.

For the neutrino comparison, the Gargamelle total cross sections are Eichten *et al.* [1497], from which the momentum fractions of Proposition 116.28 are derived here; the same exposures gave the weak neutral current in the leptonic channel [1503] and the hadronic channel [1502], reported in Chapter 104. The high-statistics neutrino structure functions are the CDHS measurement in iron [1639].

For the scaling violations, the theory is Gross and Wilczek [1501] and Politzer [1511] for asymptotic freedom, and Gribov and Lipatov [1474], Altarelli and Parisi [1591] and Dokshitzer [1595] for the evolution equations. The Gribov–Lipatov and Dokshitzer papers are translations from the Russian and neither carries a DOI. The Dokshitzer entry records that fact explicitly, together with the Russian original it translates; the Gribov–Lipatov entry carries only the journal reference of the translation, with no note, so a reader who wants the original must find it from the volume and pages alone. The measurements are BCDMS [1815] for the muon beam and the two HERA combinations [2152, 2244] for the collider; the second supersedes the first and is the data set in current use. Collins, Soper and Sterman [1817] is the factorization proof that licenses using the extracted distributions elsewhere. Current values of α_s, of the parton distributions and of the particle masses used here are from the Review of Particle Physics [2427], and the fundamental constants from the CODATA 2022 adjustment [2431]; the shipped data files behind those two citations are the source of every constant in Tables 116.1 and 116.3.

The extensions are Aubert *et al.* [1712] for the EMC effect, Ashman *et al.* [1804] for the proton spin, Drell and Yan [1421] for the lepton-pair process, and the Electron–Ion Collider white paper [2265] for the programme now being built. The flavour asymmetry of the light-quark sea rests on two measurements in different processes: Arneodo *et al.* [1879] for the New Muon Collaboration's reevaluation of the Gottfried sum, and Towell *et al.* [2020] for the Fermilab E866/NuSea Drell–Yan determination of $\bar{d}/\bar{u}$. Halzen and Martin [1739] is the textbook treatment closest to the level of this chapter.

Finally, what this reference list does *not* contain, since a reader is entitled to know where the chapter's statements outrun its sources. There is no entry for a dedicated modern measurement of $R = \sigma_L/\sigma_T$, so the value in Table 116.1 is the SLAC one and the modern refinement is not reported. There is no entry for nuclear parton distributions after 1983 (Remark 116.37) and none for polarized measurements after 1988 (Remark 116.39). There is no entry for Ellis and Jaffe, so the expectation of 0.19 against which Phenomenon 116.38 reports a contradiction is both uncited and underived, and the derivation is owed as a pending proof. Each of these is a gap in the bibliography, not a gap in the physics, and each is flagged where it occurs in the text rather than only here.

Chapter 117

Experiment: Artificial Transmutation and the Instruments of Nuclear Physics

Two questions had to be answered before particle physics could exist as a quantitative science. Can one nucleus be turned into another at will, and can the products be counted one at a time? This chapter reports the experiments that answered them. The first half follows the reaction: the transmutation of nitrogen by natural alpha particles, and then the disintegration of lithium by protons from a machine — the first nuclear reaction produced entirely by human means and, more importantly for this book, the first quantitative test of the mass–energy equivalence of Chapter 40 in a nuclear reaction. The second half follows the instruments, from Wilson's cloud chamber through Glaser's bubble chamber to Charpak's multiwire proportional chamber, which replaced the photograph with an electronic address and made the modern event rate possible.

The chapter is the experimental floor of Part XI. Its nuclear reactions belong to Chapter 107, where the binding-energy systematics they sample are developed and where Chadwick's discovery of the neutron — made in 1932 with the same scintillation technique used here, in the same laboratory, in the same year as the lithium disintegration — is treated in full [693]. Its instruments are the ancestors of every detector in Chapter 115, Chapter 116 and Chapter 113. Five of the six experiments below were recognised by a Nobel Prize in Physics: Wilson in 1927, Lawrence in 1939, Cockcroft and Walton in 1951, Glaser in 1960 and Charpak in 1992.

117.1 Conventions and constants

Three conventions are used throughout, and they are worth stating before any number appears, because every one of the primary papers below uses different ones.

Nuclear masses are the evaluated atomic masses of the AME 2020 tables [2393], and are converted to energy with the CODATA 2022 value of the unified atomic mass constant,

$$u\,c^2 = 931.494\,\mathrm{MeV}\,, \qquad u = 1.660\,539\,07 \times 10^{-27}\,\mathrm{kg}\,, \tag{117.1}$$

to the precision needed here [2431]. Because atomic rather than nuclear masses are used, the electrons balance on the two sides of every reaction written below, and no electron-mass correction is required.

Energies are quoted in electronvolts and in joules. The conversion is exact, since the elementary charge is a defining constant of the SI [2431]:

$$1\,\mathrm{eV} = 1.602\,176\,634 \times 10^{-19}\,\mathrm{J}\,, \qquad 1\,\mathrm{MeV} = 1.602\,176\,634 \times 10^{-13}\,\mathrm{J}\,. \tag{117.2}$$

Two further combinations recur, and are taken from Equations (107.1) and (107.2) of Chapter 107, where they are fixed once from the same adjustment: $\hbar c = 197.326\,98\,\mathrm{MeV\,fm}$ and $e^2/(4\pi\varepsilon_0) = 1.439\,964\,5\,\mathrm{MeV\,fm}$. Both are ordinary SI quantities written in convenient units — the first is $3.1615 \times 10^{-26}\,\mathrm{J\,m}$, the second $2.3071 \times 10^{-28}\,\mathrm{J\,m}$ — and $\hbar$ and c appear explicitly in every equation below.

Ranges are the third convention and the one most easily misread. The experimenters of 1912–1932 quoted the penetrating power of a particle as the thickness of *air at atmospheric pressure* that would stop it, and inserted solid absorbers whose stopping power had been calibrated against air. A length is not the physical quantity: what stops a particle is the number of electrons it must pass, so the invariant statement is an areal mass density. At 288.15 K and 101.325 kPa the ideal-gas law with the mean molar mass of dry air, $M_{\mathrm{air}} = 28.9647 \times 10^{-3}\,\mathrm{kg/mol}$ [2427], gives $\rho_{\mathrm{air}} = pM_{\mathrm{air}}/(RT) = 1.2250\,\mathrm{kg/m^3}$, so

$$1 \times 10^{-2}\,\mathrm{m\ of\ air} = 1.2250 \times 10^{-2}\,\mathrm{kg/m^2}\,, \tag{117.3}$$

and every range quoted below is given both ways. Imperial units, in which two of the primary papers are partly written, are converted on sight and not reproduced.

Remark 117.1 (Reading this chapter against the natural-units literature). The nuclear and detector literature almost universally sets $\hbar = c = 1$, so that an energy, a momentum and a mass are all quoted in MeV and a length in 1/MeV. This book does not: $\hbar$ and c are carried explicitly, a momentum is written $2.5 \times 10^8\,\mathrm{eV}/c$ and a mass $7\,\mathrm{GeV}/c^2$. To map a formula from that literature onto one here, restore the unique powers of $\hbar$ and c that make the two sides dimensionally equal in SI; the combinations that do the work are the two quoted above, and the statement "a length of 1 fm corresponds to an energy of 197.3 MeV" means $\hbar c/1\,\mathrm{fm}$ and nothing more. No derivation in this chapter is carried out in those units.

117.2 Rutherford: the first artificial transmutation (1919)

Tests Phenomenon 117.2.

117.2.1 Apparatus

The apparatus is a scintillation counter with a gas cell in front of it, and nothing else. A rectangular brass box about 0.18 m long, with an internal cross-section of roughly $0.06\,\mathrm{m} \times 0.02\,\mathrm{m}$, was closed at one end by a silver foil whose stopping power had been calibrated as equivalent to about $6 \times 10^{-2}\,\mathrm{m}$ of air, that is $7.4 \times 10^{-2}\,\mathrm{kg/m^2}$ by Equation (117.3). Immediately beyond the foil sat a screen of zinc sulfide, viewed through a microscope of low power by an observer who had dark-adapted for half an hour in a light-tight room.

Inside the box, on a slide that could be moved along its length without breaking the seal, a metal disc carried an active deposit collected from radon gas. The alpha emitter in that deposit is radium C′, known today as ^{214}Po: a nuclide of half-life $1.64 \times 10^{-4}\,\mathrm{s}$ which is fed continuously by the longer-lived members of the chain above it and which therefore behaves as a steady source. Its alpha particle is among the most energetic available from any natural source in useful quantity — thorium C′ (^{212}Po), reached through the thorium active deposit and worked with at the Cavendish in the same period, is more energetic still, at 8.785 MeV by the same mass balance [2393] — and its energy is not an experimental input but a consequence of a mass difference. The alpha decay energy from the mass evaluation is $Q_\alpha = 7.8335\,\mathrm{MeV}$ [2393], and the recoiling ^{210}Pb takes a share fixed by momentum conservation, so the alpha particle carries

$$E_\alpha = Q_\alpha \,\frac{m(^{210}\mathrm{Pb})}{m(^{210}\mathrm{Pb}) + m(^{4}\mathrm{He})} = 7.687\,\mathrm{MeV} = 1.2316 \times 10^{-12}\,\mathrm{J}\,, \tag{117.4}$$

which is the number quoted in every table of the period and the number used throughout this section. Its range in air at 288.15 K and 101.325 kPa is 7.0×10^{-2} m, that is 8.6×10^{-2} kg/m^2.

The box could be filled with a dry gas of choice at a controlled pressure — oxygen, carbon dioxide, dry air, or nitrogen — and evacuated and refilled between runs. Gas purity is not a refinement here but the entire experiment: the effect looked for is the appearance of hydrogen nuclei, and hydrogen is the commonest contaminant in any vacuum system. The gas was therefore dried over phosphorus pentoxide, and the comparison between gases, rather than the absolute rate in any one of them, carries the argument.

The detector was the dark-adapted eye of the observer. A zinc sulfide screen converts the energy a stopping ion deposits into a flash of visible photons, and the flash from a single particle is at the edge of what the eye can see; the technique works only at rates of a few counts per minute, only in complete darkness, and only for observers who alternate in short shifts. Its virtue is that it counts single particles, and in 1919 nothing else did.

117.2.2 Procedure

Absorbing foils were interposed between source and screen and their stopping power expressed as an equivalent thickness of air, so that a count could be plotted against the distance the emerging particles were able to travel. Enough absorber was always present to stop every primary alpha particle: with the silver foil alone accounting for 6×10^{-2} m of air equivalent and the alpha range being 7.0×10^{-2} m, a short additional path in the gas suffices, and whatever produced a scintillation on the screen therefore had a longer range than the alpha particles that made it.

The experiment is then a comparison between gases, made under four controls, each of which excludes a specific alternative.

First, the choice of filling gas. Oxygen and carbon dioxide contain no nitrogen and no hydrogen, and are chemically comparable stopping media; they are the null runs. Dry air is one fifth oxygen and four fifths nitrogen; pure nitrogen is the signal run. If the effect is proportional to the amount of nitrogen present it cannot be a property of the source, of the box, or of the screen.

Second, the source-to-screen distance and the gas pressure were varied independently. Both change the amount of gas traversed, and a genuine product of the gas must respond to their product and not to either separately.

Third, the absorber thickness was scanned. A particle with a definite energy has a definite range, so the count as a function of absorber falls to zero at a characteristic thickness; reading that end point is how an energy was measured before magnetic analysis was available in this energy range.

Fourth, the residual effect with the box evacuated was measured and subtracted. Hydrogen occluded in the source mount itself produces long-range particles by exactly the elastic mechanism the experiment must exclude, and that background is not small.

117.2.3 Observations and data

With oxygen or carbon dioxide in the box, no scintillations survived beyond the alpha range. With dry air, and more strongly with pure nitrogen, scintillations persisted at absorber thicknesses far beyond it — out to an air equivalent of about 0.40 m, that is 0.49 kg/m^2 — and beyond the greatest range attainable by a hydrogen nucleus knocked on elastically by an alpha particle. The effect was small: a handful of scintillations per minute against an enormous number of incident alphas, which is why it took the full apparatus of scintillation counting to see it at all, and why the comparison between fillings rather than any absolute rate is the published result.

Table 117.1 collects the quantities the argument turns on. The historical rows are the observations; the computed rows are evaluated here from the mass evaluation [2393] and the kinematics derived below, and are what the observations have to be tested against. The counting uncertainties on the original scintillation tallies are Poissonian on tallies of order ten per run, that is of order 30 % per point; the ranges quoted from the paper are end points read off an absorption curve and are good to perhaps 10 %. Neither is quoted to more precision than that here, and the original run-by-run tallies are not reproduced: what the paper establishes, and what is reproduced, is the ordering of the end points.

Table 117.1: The range and energy comparison that carries Rutherford's argument. Rows marked *obs.* are the experimental values, quoted from the primary paper; rows marked *calc.* are computed in this chapter from the evaluated atomic masses and the elastic and two-body kinematics derived below, and involve no fitted quantity. Air equivalents are converted to areal density by Equation (117.3). The elastic ceiling is the largest energy an alpha particle can give a free proton in any collision whatever; the forward reaction proton exceeds it, and that is the whole observation. Bracketed uncertainties on the observed rows are the 10 % range-end-point reading error and the Poisson error on the fork count discussed in the text; the primary papers quote none, so these are estimates from the stated technique and are labelled as such. The computed rows inherit only the mass-evaluation uncertainty, which is of order 10 eV and therefore falls six orders of magnitude below the last digit shown.

Quantity	Value (eV)	Value (SI)	Kind, source
Alpha energy of radium C′ (^{214}Po)	7.687 MeV	1.2316×10^{-12} J	calc., [2393]
Range of those alphas in air	—	$(8.6 \pm 0.9) \times 10^{-2}$ kg/m^2	obs. 1919, [461]
Longest range seen in nitrogen	—	(0.49 ± 0.05) kg/m^2	obs. 1919, [461]
Elastic ceiling for a struck proton	4.941 MeV	7.916×10^{-13} J	calc., Equation (117.6)
Reaction energy Q of ^{14}N$(\alpha, p)^{17}$O	−1.1919 MeV	-1.9096×10^{-13} J	calc., [2393]
Laboratory threshold	1.5326 MeV	2.4554×10^{-13} J	calc., Equation (117.11)
Proton energy at 0° to the beam	5.929 MeV	9.499×10^{-13} J	calc., Equation (117.12)
Proton energy at 90°	4.423 MeV	7.086×10^{-13} J	calc., Equation (117.12)
Proton energy at 180°	3.299 MeV	5.286×10^{-13} J	calc., Equation (117.12)
Two-branch forks among alpha tracks	—	8 ± 3 in 415 000	obs. 1925, [518]

The uncertainty on the calculated rows deserves a word, because it is negligible and it is worth saying why. The AME 2020 evaluation quotes the masses of ^{1}H, ^{4}He, ^{14}N and ^{17}O to eleven significant figures, so the propagated uncertainty on the Q value of Equation (117.10) is at the level of electronvolts — six orders of magnitude below the 0.1 MeV with which a range end point can be read. In this comparison the theory is exact and the measurement is not, which is the reverse of the usual situation and is what makes a nineteen-nineteen range measurement still worth reporting.

117.2.4 Interpretation

The long-range particles are hydrogen nuclei, and they cannot have come from hydrogen in the gas, for two independent reasons. Their range exceeds what an elastic collision can give — the ceiling in Table 117.1 is 4.941 MeV and particles of longer range than that were counted — and the yield tracks the nitrogen content of the filling rather than the total amount of gas. Rutherford drew the conclusion in the plainest possible terms: the nitrogen atom is disintegrated.

What he could not establish was the fate of the alpha particle. Two possibilities remained open in 1919. The alpha might knock a proton out and continue, in which case four objects leave the collision point (alpha, proton, residual ^{13}C, and nothing else) and a photograph of the event would show three tracks emerging. Or the alpha might be captured, in which case only two objects leave and a photograph shows two. That was settled six years later by Blackett, who photographed the collisions in a Wilson chamber (Section 117.5) and found, among some 415 000 alpha tracks recorded on 23 000 stereoscopic photographs, eight forks with exactly two branches and no third. A two-branch fork means the alpha particle did not survive the encounter: it was captured, and the reaction is

$$^{14}\mathrm{N} + {}^{4}\mathrm{He} \longrightarrow {}^{17}\mathrm{O} + {}^{1}\mathrm{H}\,. \tag{117.5}$$

The counting is itself the measurement: eight events in 415 000 tracks, each track traversing a few centimetres of air, is a yield of order 1×10^{-5} per alpha, and a Poisson uncertainty of $\sqrt{8} \approx 3$ on eight events is why Blackett's number is a statement about the mechanism and not about the cross-section.

That the two branches also close the momentum triangle is what excludes the remaining alternative, that the alpha was captured and a neutral object emitted alongside the proton. In a stereoscopic pair the directions of both branches are measured, and the vector sum of the two final momenta must equal the incident momentum. Blackett's forks close; there is no missing momentum to assign to an unseen particle. The same argument run in reverse, on a reaction where the triangle does *not* close, is what Chadwick used in 1932 to establish the neutron [693], and it is the ancestor of every missing-momentum measurement in Chapter 113.

This is the first transmutation of one element into another by deliberate human action, and it was done with a naturally occurring projectile whose energy the experimenter could not change. Removing that restriction is the content of Section 117.3.

Phenomenon 117.2 (Nitrogen is transmuted by alpha bombardment)**.** Bombarding nitrogen with alpha particles from a natural source ejects hydrogen nuclei whose range exceeds that of any proton an alpha particle can set in motion by an elastic collision [461]. The alpha particle is absorbed in the process, so that the reaction is the capture Equation (117.5) and not a scattering [518].

Rests on Equation (117.5).

Derivation. *Derives Phenomenon 117.2.*

Three independent statements have to be checked: that an elastic collision cannot produce the observed particles, that the capture reaction is energetically open to alphas of 7.687 MeV, and that when it is open the protons it produces do exceed the elastic ceiling.

The elastic ceiling. Take a projectile of mass m_1 and kinetic energy E striking a free target of mass m_2 at rest, non-relativistically — and here $E/m_1c^2 = 7.687\,\mathrm{MeV}/3727.4\,\mathrm{MeV} = 2.06\times10^{-3}$, with the alpha-particle mass energy from [2431], so the non-relativistic treatment is ample. Conservation of momentum and of kinetic energy in the head-on case give the target the energy

$$E_2^{\max} = \frac{4m_1m_2}{(m_1+m_2)^2}\,E\,, \tag{117.6}$$

which is the largest transfer any elastic collision can achieve, at any angle, for any impact parameter, and independently of the force law: only the two masses enter. For an alpha particle on a proton the mass numbers alone would give $16/25 = 0.64$, but the ceiling is a number the argument leans on, so it is evaluated from the same atomic masses as everything else: $4m_1m_2/(m_1+m_2)^2 = 0.642\,74$ and

$$E_2^{\max} = 0.642\,74 \times 7.687\,\mathrm{MeV} = 4.941\,\mathrm{MeV} = 7.916 \times 10^{-13}\,\mathrm{J}\,. \tag{117.7}$$

Protons of longer range than this were seen, and no elastic process can make them. The same formula applied to an alpha on a nitrogen nucleus gives a factor 0.6915, so elastic recoils of nitrogen carry up to 5.32 MeV — energetic, but a nitrogen nucleus of that energy has a range in air of a few millimetres and cannot reach the screen.

The reaction energy. Form the mass balance of Equation (117.5) from the evaluated atomic masses [2393]. Summing the two sides,

$$m\left({}^{14}\mathrm{N}\right) + m\left({}^{4}\mathrm{He}\right) = 18.005\,677\,u\,, \tag{117.8}$$

$$m\left({}^{17}\mathrm{O}\right) + m\left({}^{1}\mathrm{H}\right) = 18.006\,957\,u\,, \tag{117.9}$$

so the products are *heavier* than the reactants by $0.001\,280\,u = 2.1247 \times 10^{-30}\,\mathrm{kg}$ and the reaction energy is negative,

$$Q = \left[m\left({}^{14}\mathrm{N}\right) + m\left({}^{4}\mathrm{He}\right) - m\left({}^{17}\mathrm{O}\right) - m\left({}^{1}\mathrm{H}\right)\right]c^2 = -1.1919\,\mathrm{MeV} = -1.9096 \times 10^{-13}\,\mathrm{J}\,. \tag{117.10}$$

An endothermic reaction needs more than $|Q|$ in the laboratory, because the centre of mass must keep moving and its kinetic energy is unavailable; demanding that the invariant available energy suffice to create the products at rest in their own centre of mass gives the threshold

$$E_{\mathrm{th}} = |Q|\left(1 + \frac{m({}^{4}\mathrm{He})}{m({}^{14}\mathrm{N})}\right) = 1.5326\,\mathrm{MeV} = 2.4554 \times 10^{-13}\,\mathrm{J}\,. \tag{117.11}$$

Radium C′ supplies 7.687 MeV, five times the threshold, so the reaction is open.

The proton energy. Being open is not enough; the surplus has to end up in the proton in the observable amount. For a two-body reaction $1 + 2 \to 3 + 4$ with target 2 at rest, non-relativistic throughout, conservation of energy and of momentum give the energy of product 3 at laboratory angle θ as

$$\sqrt{E_3} = v + \sqrt{v^2 + w}\,, \qquad v = \frac{\sqrt{m_1 m_3 E_1}}{m_3 + m_4}\cos\theta\,, \qquad w = \frac{m_4 Q + (m_4 - m_1)\,E_1}{m_3 + m_4}\,, \tag{117.12}$$

where E_1 is the projectile energy and the masses may be taken in any consistent unit, the ratios being all that enter. Both v^2 and w carry the SI dimension of energy, so $\sqrt{E_3}$ is an energy to the half power on both sides and the relation is dimensionally homogeneous.

Equation (117.12) is used six times in this chapter, so it is derived rather than quoted. Let p_1, p_3, p_4 be the laboratory momenta and θ the angle of product 3 to the beam. Momentum conservation gives $\boldsymbol{p}_4 = \boldsymbol{p}_1 - \boldsymbol{p}_3$, hence

$$p_4^2 = p_1^2 + p_3^2 - 2p_1p_3\cos\theta\,, \tag{117.13}$$

and energy conservation, with the reaction energy Q released into kinetic energy, gives $E_4 = E_1 + Q - E_3$. Non-relativistically $p^2 = 2mE$ throughout — legitimate here because $E/mc^2 \le 2.1 \times 10^{-3}$ for every particle involved — so Equation (117.13) becomes

$$2m_4\,(E_1 + Q - E_3) = 2m_1E_1 + 2m_3E_3 - 2\sqrt{4m_1m_3E_1E_3}\,\cos\theta\,. \tag{117.14}$$

Dividing by $2\,(m_3 + m_4)$ and collecting the terms in E_3 and in $\sqrt{E_3}$ turns this into $E_3 - 2v\sqrt{E_3} - w = 0$ with exactly the v and w of Equation (117.12); solving the quadratic in $\sqrt{E_3}$ and discarding the negative root, which would make E_3 negative whenever $w > 0$, gives $\sqrt{E_3} = v + \sqrt{v^2 + w}$. With $m_1 = m(^4\mathrm{He})$, $m_2 = m(^{14}\mathrm{N})$, $m_3 = m(^1\mathrm{H})$, $m_4 = m(^{17}\mathrm{O})$, $E_1 = 7.687\,\mathrm{MeV}$ and Q from Equation (117.10), the bracket w is $4.423\,\mathrm{MeV}$ — which is $E_p(90°)$ itself, since v vanishes there — and

$$E_p(0°) = 5.929\,\mathrm{MeV}\,, \qquad E_p(90°) = 4.423\,\mathrm{MeV}\,, \qquad E_p(180°) = 3.299\,\mathrm{MeV}\,. \tag{117.15}$$

The forward protons exceed the elastic ceiling of $4.941\,\mathrm{MeV}$ by $0.988\,\mathrm{MeV}$, a margin of 20 % in energy and considerably more in range, since a proton's range grows faster than its energy. Backward protons do not exceed it, which is consistent: the effect appears in the forward hemisphere, is a small fraction of the total yield, and is exactly what a scintillation screen placed beyond an absorber selects.

All three statements come from the same mass table and from elementary kinematics, and none of them is available to a hydrogen-contamination explanation, which is bounded by Equation (117.6) whatever the amount of hydrogen present. □

Remark 117.3 (What this experiment does not measure). It is worth being exact about the limits, because the experiment is often described as though it had measured a cross-section. It did not. The observable is a count of scintillations above a range threshold, with an efficiency set by a human eye, a solid angle known to perhaps a factor of two, and a source strength known from an ionisation measurement. What is established at high confidence is a qualitative statement of a kind that no amount of precision could improve on — that particles of a range no elastic collision can produce are emitted from nitrogen and from no other gas tried — and the identification of the reaction had to wait six years for a different instrument. The quantitative rows of Table 117.1 are the modern calculation of what should have been seen, not a modern measurement of what was.

117.2.5 Primary references

The observation is Rutherford's [461], the fourth and last paper of a series in the *Philosophical Magazine* on the collision of alpha particles with light atoms; the first three establish the scintillation technique and the elastic behaviour of hydrogen and of the other light gases, which is what makes the fourth readable as a controlled comparison rather than as an isolated claim. It is worth reading for its caution: Rutherford states the anomaly, the controls that exclude contamination, and explicitly declines to say what became of the alpha particle.

Blackett's paper [518] is the identification, and it is a different kind of document — an exhaustive scan of a photographic record, with the eight forks reproduced and measured. The two together are the pattern this chapter follows throughout: a counting experiment establishes that something happens, and an imaging experiment establishes what. Wilson's instrument [394], reported in Section 117.5, is what made the second half possible, and it is seven years older than the anomaly it resolved.

117.3 Cockcroft and Walton: lithium split by accelerated protons (1932)

Tests Phenomena 40.6, 117.4 and 117.5.
Assuming Theorem 9.76.

117.3.1 Apparatus

The apparatus divides into a high-voltage generator, an accelerating tube, and a scintillation counter of the kind already described in Section 117.2.

The generator is the part that carries the names. A transformer supplied an alternating potential of peak value V_0 of order 100 kV, and a cascade of capacitors and rectifiers converted it to a steady direct potential several times larger. The principle is worth stating because it is the reason the machine existed. In a cascade of n stages, the alternating column of capacitors is charged through the rectifiers on alternate half cycles, each stage of the direct column ends up at $2V_0$ above the one below it, and the output is

$$V_{\text{out}} = 2nV_0 \,. \tag{117.16}$$

With four stages and $V_0 = 100\,\text{kV}$ this is 800 kV, which is what the Cavendish machine reached; in practice 100 kV to 700 kV was used. No component ever holds more than $2V_0$ across it, which is what makes the arrangement buildable at all, and the same circuit — now called a Cockcroft–Walton generator — is still the first stage of many accelerator chains.

The direct potential was applied across a sealed accelerating tube built of glass cylinders about a metre long and 0.3 m in diameter, with intermediate electrodes so that the drop was shared and no single gap had to hold the whole voltage. That constraint is physical, not merely engineering: the dielectric strength of air at atmospheric pressure is about $3 \times 10^6\,\text{V/m}$, so 800 kV requires a clearance of at least 0.27 m in air even between ideal electrodes, and several times that between real ones with sharp edges and dust. It is the reason the electrostatic route to high energies has a ceiling of a few megavolts, and hence the reason for Section 117.4.

Protons were drawn from a hydrogen discharge at the top of the tube, a canal-ray source running at a few kilovolts, delivering a beam current of order $1 \times 10^{-5}\,\text{A}$ at the target with the tube continuously pumped to a residual pressure of order $1 \times 10^{-3}\,\text{Pa}$. The target was lithium metal, deposited on a backing and set at an angle to the beam so that the products could leave sideways. They left through a mica window whose stopping power had been calibrated against air, and were counted as scintillations on a zinc-sulfide screen under a microscope, exactly as in Section 117.2; further mica sheets placed in front of the screen measured their range.

117.3.2 Procedure

The measurement has three parts, and only the third identifies the reaction.

The accelerating potential was set and measured — by a sphere gap and a resistance chain — and the scintillation rate recorded as a function of it, so that the yield curve could be followed downwards towards the point where the effect disappeared into the background. The importance of pushing the curve downwards rather than upwards is that the interesting statement is a lower limit on where the reaction is still detectable, which is a statement about the barrier and not about the apparatus.

The range of the emitted particles was then measured by absorption in mica, converted to an equivalent range in air by the calibration, and hence to an energy through the range–energy relation for helium nuclei established with natural alpha sources. The particles' behaviour in absorbers — the shape of the absorption curve near its end point, and the specific ionisation, which follows Equation (115.15) — identified them as helium nuclei rather than protons or electrons.

Finally, two screens were placed on opposite sides of the target and their scintillations examined for coincidence, with an observer at each. This is the step that turns a detection into an identification. If the particles are emitted singly, the two observers see uncorrelated flashes at a rate set by chance; if they are emitted in pairs travelling in opposite directions,

the coincidence rate exceeds the accidental rate by a large factor. The kinematics of the proposed reaction makes that test sharp, as Equation (117.24) below shows.

117.3.3 Observations and data

Scintillations appeared with accelerating potentials as low as about 125 kV, and the yield rose rapidly as the potential was raised. The particles had a range in air of about 8.4×10^{-2} m, that is $0.103\,\mathrm{kg/m^2}$, which for a helium nucleus corresponds to a kinetic energy of about 8.6 MeV, that is 1.38×10^{-12} J; their behaviour in absorbers was that of ordinary alpha particles. The coincidence test showed them emitted in pairs, in opposite directions.

Table 117.2 sets the measured quantities beside the predictions computed here. Their accuracy is limited on the experimental side by the reading of a range end point, good to a few per cent, and by the calibration of the range–energy relation for helium nuclei; on the calculated side it is limited by nothing that matters, the mass-evaluation uncertainty on the Q value being of order 10 eV.

Table 117.2: Cockcroft and Walton's lithium disintegration: measured against calculated. The calculated alpha energies come from the two-body kinematics of Equation (117.12) applied to Equation (117.17) at a proton energy of 300 keV, which is representative of the potentials at which the range measurement was made; the mass difference is from the evaluated atomic masses. The last row is the reason the coincidence test works: at these energies the two alpha particles are back to back in the laboratory to within a few degrees. Bracketed uncertainties on the observed rows are estimates from the stated technique — a few per cent on a range end point and on its conversion to energy, and the sphere-gap calibration on the potential — since the primary papers quote none; the onset potential is in addition a limit of detection set by the scintillation background, not a physical threshold. The calculated rows carry the mass-evaluation uncertainty of order 10 eV, far below the last digit shown.

Quantity	Value (eV)	Value (SI)	Kind, source
Lowest potential showing the effect	—	(125 ± 5) kV	obs. 1932, [694]
Range of the products in air	—	$(0.103 \pm 0.004)\,\mathrm{kg/m^2}$	obs. 1932, [695]
Energy inferred from that range	(8.6 ± 0.3) MeV	$(1.38 \pm 0.05) \times 10^{-12}$ J	obs. 1932, [695]
Mass lost, $^7\mathrm{Li} + {}^1\mathrm{H} \to 2\,{}^4\mathrm{He}$	—	3.0922×10^{-29} kg	calc., [2393]
Energy released, $Q = \Delta m\, c^2$	17.3462 MeV	2.7792×10^{-12} J	calc., Equation (117.21)
Alpha energy at 90° to the beam	8.785 MeV	1.4076×10^{-12} J	calc., Equation (117.12)
Alpha energy at 0°	9.639 MeV	1.5443×10^{-12} J	calc., Equation (117.12)
Alpha energy at 180°	8.008 MeV	1.2830×10^{-12} J	calc., Equation (117.12)
Departure of the pair from collinearity	—	5.3°	calc., Equation (117.24)

The second table is the sub-barrier behaviour, and it is computed rather than observed. Table 117.3 evaluates the semiclassical barrier transmission for a proton incident on a lithium nucleus at each of the accelerating potentials used, from the Gamow exponent of Equation (9.93). The point of tabulating it is that the numbers are not small in the way one

expects of a tunnelling problem: the barrier here is low and thin, the exponent is of order unity to five, and a transmission of 6×10^{-3} at the lowest potential is entirely consistent with a countable rate. It is the *classical* prediction — exactly zero at every potential in the table — that the observation contradicts.

Table 117.3: Barrier transmission for a proton on ^{7}Li, computed here in SI from Equation (9.93) with a contact radius $R = 4.08\,\mathrm{fm}$ and the reduced mass $\mu = 1.4626 \times 10^{-27}\,\mathrm{kg}$. V is the accelerating potential, E the energy in the centre of mass, b the classical turning point, 2γ the Gamow exponent and $T = \mathrm{e}^{-2\gamma}$ the transmission. The last column is the relative thick-target yield of Equation (117.31), normalised to the lowest potential; it is the quantity a yield curve actually measures, and it rises far more steeply than T alone. The classical value of T is zero in every row, since the barrier height is 1.06 MeV. Nothing here is a measurement, so no row carries an experimental uncertainty; what the numbers inherit is the 4 % uncertainty of the contact radius R, which moves 2γ in the first row by about 0.05.

V	E (centre of mass)	b	2γ	T	Y/Y_0
125 kV	109.3 keV	39.5 fm	5.04	6.5×10^{-3}	1.00
150 kV	131.2 keV	32.9 fm	4.32	1.33×10^{-2}	2.24
200 kV	174.9 keV	24.7 fm	3.31	3.64×10^{-2}	7.02
250 kV	218.6 keV	19.8 fm	2.63	7.18×10^{-2}	15.4
300 kV	262.3 keV	16.5 fm	2.14	1.18×10^{-1}	27.5
400 kV	349.8 keV	12.4 fm	1.46	2.31×10^{-1}	62.4
500 kV	437.2 keV	9.9 fm	1.02	3.60×10^{-1}	109
700 kV	612.1 keV	7.1 fm	0.48	6.17×10^{-1}	225

117.3.4 Interpretation

Two alpha particles emitted back to back, with nothing else, identifies the reaction uniquely as

$$^{7}\mathrm{Li} + {}^{1}\mathrm{H} \longrightarrow {}^{4}\mathrm{He} + {}^{4}\mathrm{He}\,. \tag{117.17}$$

This is the first nuclear reaction driven by artificially accelerated particles: unlike Section 117.2, the projectile energy is now a knob on the apparatus, and the yield curve is a measurement rather than an accident of which radioactive source was to hand.

Its importance for this book, however, is quantitative. The energy released was measured directly, from the range of the products, and compared with the energy predicted by the difference of the masses — a prediction that has nothing to do with nuclear physics and everything to do with Phenomenon 40.6. The two agree. Cockcroft and Walton made the comparison in 1932 with the mass-spectrograph values then available, which gave a mass difference of about $0.018\,u$ and hence an energy release close to 17 MeV, against the $2 \times 8.6\,\mathrm{MeV} = 17.2\,\mathrm{MeV}$ they measured; with the modern evaluation the prediction is 17.3462 MeV (Equation (117.21)) and the kinematically correct comparison at 90° is 8.785 MeV per alpha against a measured 8.6 MeV, a discrepancy of 2 % and entirely within the accuracy of a range measurement. This is the first quantitative confirmation of mass–energy equivalence in a nuclear reaction, and it was obtained from a table of atomic weights and a ruler laid against an absorption curve.

The comparison is worth appreciating for how little it assumes. The left-hand side of Equation (117.21) is measured by a mass spectrograph, an instrument that deflects ions in electric and magnetic fields and knows nothing of nuclear forces. The right-hand side is measured by the distance an alpha particle travels in mica. The only bridge between

them is $E = mc^2$. Had the two disagreed, no nuclear model would have been at fault; the relativistic energy–momentum relation of Chapter 40 would have been.

A second reading of the same data was equally consequential. The reaction was observed at proton energies an order of magnitude below the height of the Coulomb barrier between a proton and a lithium nucleus, which classical mechanics forbids outright. The experiment is therefore also a direct confirmation of barrier penetration [600, 603, 625], whose other early evidence — alpha decay — is developed in Theorem 107.63 of Chapter 107. It is a cleaner confirmation than alpha decay in one respect and a dirtier one in another. Cleaner, because the energy of the incident particle is set by the experimenter and can be varied, so the exponential dependence predicted by Equation (9.87) can be traced out rather than inferred from a correlation across nuclides. Dirtier, because the absolute yield involves the nuclear matrix element and the thick-target energy loss as well as the barrier, so only the shape of the curve is a clean test.

Phenomenon 117.4 (Lithium disintegrates into two alpha particles)**.** Lithium bombarded with protons accelerated through a few hundred kilovolts emits helium nuclei in pairs, in nearly opposite directions, each of kinetic energy near 8.6 MeV [694, 695]. The total energy released is equal, within the accuracy of the measurement, to the mass lost by the system, converted at the rate $E = mc^2$.

Rests on Equations (40.6)*,* (117.12) *and* (117.17)*.*

Derivation. *Derives Phenomenon 117.4.*

Write the mass balance of Equation (117.17) with the evaluated atomic masses [2393]:

$$m\left({}^{7}\mathrm{Li}\right) + m\left({}^{1}\mathrm{H}\right) = 7.016\,003\,4\,u + 1.007\,825\,0\,u = 8.023\,828\,4\,u\,, \tag{117.18}$$

$$2\,m\left({}^{4}\mathrm{He}\right) = 2 \times 4.002\,603\,3\,u = 8.005\,206\,5\,u\,. \tag{117.19}$$

The system is lighter after the reaction than before by

$$\Delta m = 0.018\,622\,0\,u = 3.0922 \times 10^{-29}\,\mathrm{kg}\,, \tag{117.20}$$

and by Equation (117.1) the energy released is

$$Q = \Delta m\,c^2 = 3.0922{\times}10^{-29}\,\mathrm{kg}{\times}\left(2.997\,924\,58 \times 10^{8}\,\mathrm{m/s}\right)^2 = 2.7792{\times}10^{-12}\,\mathrm{J} = 17.3462\,\mathrm{MeV}\,. \tag{117.21}$$

Note that the electron count balances — three electrons on each side, since lithium is $Z = 3$ and each helium is $Z = 2$ — so atomic masses may be used without correction, and that Δm is 0.23 % of the total mass, which is why a mass spectrograph of the period could see it.

Now share it. Because the two products are identical, the reaction is symmetric in the centre of mass and each alpha particle takes half the available energy there; in the laboratory the pair is boosted along the beam. Applying Equation (117.12) with $m_1 = m({}^{1}\mathrm{H})$, $m_2 = m({}^{7}\mathrm{Li})$, $m_3 = m_4 = m({}^{4}\mathrm{He})$, Q from Equation (117.21) and $E_1 = 300\,\mathrm{keV}$:

$$E_\alpha(0^\circ) = 9.639\,\mathrm{MeV}\,, \quad E_\alpha(90^\circ) = 8.785\,\mathrm{MeV}\,, \quad E_\alpha(180^\circ) = 8.008\,\mathrm{MeV}\,, \tag{117.22}$$

the three summing correctly in pairs to $Q + E_1 = 17.646\,\mathrm{MeV}$ for the forward–backward pair. The screens in the coincidence experiment sat transverse to the beam, where the prediction is 8.785 MeV per particle, and the measured 8.6 MeV agrees to 2 %.

Two checks complete the argument. First, that the crude statement “each alpha takes half” is a good approximation and why: with $E_1 \ll Q$ the incident momentum is small beside the momenta of the products. An alpha particle of 8.785 MeV has

$$p_\alpha c = \sqrt{E_\alpha^2 + 2E_\alpha m_\alpha c^2} = 256.1\,\mathrm{MeV}\,, \qquad p_\alpha = 1.368 \times 10^{-19}\,\mathrm{kg\,m/s}\,, \tag{117.23}$$

using $m_\alpha c^2 = 3727.4\,\mathrm{MeV}$ [2431], while a proton of 300 keV has $p_p c = 23.7\,\mathrm{MeV}$, that is $1.27 \times 10^{-20}\,\mathrm{kg\,m/s}$. Second, the angular consequence: if one alpha leaves at 90°, the other must carry the whole incident longitudinal momentum, so the pair departs from collinearity by

$$\delta = \arctan\frac{p_p}{p_\alpha} = \arctan\frac{23.7\,\mathrm{MeV}/c}{256.1\,\mathrm{MeV}/c} = 5.3^\circ\,. \tag{117.24}$$

Five degrees is small enough that two screens set at 180° both intercept the pair, which is why the coincidence test worked, and large enough to have been detectable had the apparatus been built to look for it.

Nothing in this argument is nuclear. The only physical input beyond bookkeeping is $E = mc^2$ (Equation (40.6)), and the experiment measures the two sides of Equation (117.21) independently — the left from a mass spectrograph, the right from an absorption curve. □

Phenomenon 117.5 (The reaction proceeds far below the Coulomb barrier)**.** Disintegration is observed with protons accelerated through as little as about 125 kV [694], an order of magnitude below the electrostatic barrier that a proton must surmount to reach a lithium nucleus. A classical particle of that energy never arrives; the observed yield is the transmission of the barrier by tunnelling [600, 603].

Rests on Theorem 9.76 and Equations (9.93) *and* (107.2).

Derivation. *Derives Phenomenon 117.5.*

The barrier. Two nuclei of charges $Z_1 e$ and $Z_2 e$ touch when their centres are separated by $R \approx r_0\left(A_1^{1/3} + A_2^{1/3}\right)$ with $r_0 \approx 1.4\,\mathrm{fm}$, and up to that separation the only force acting is the Coulomb repulsion, so the barrier height is

$$B = \frac{Z_1 Z_2 e^2}{4\pi\varepsilon_0 R}\,. \tag{117.25}$$

With $e^2/(4\pi\varepsilon_0) = 1.4400\,\mathrm{MeV\,fm}$ from Equation (107.2), and for a proton on lithium, $Z_1 Z_2 = 3$ and $R = 1.4\,\mathrm{fm} \times (1 + 7^{1/3}) = 4.08\,\mathrm{fm}$, giving

$$B = 1.059\,\mathrm{MeV} = 1.697 \times 10^{-13}\,\mathrm{J}\,. \tag{117.26}$$

The classical statement. The relevant energy is the one available in the centre of mass, not the laboratory energy: with a proton of 125 keV on a lithium nucleus,

$$E = E_{\mathrm{lab}}\frac{m_{\mathrm{Li}}}{m_{\mathrm{Li}} + m_p} = 109.3\,\mathrm{keV} = 1.751 \times 10^{-14}\,\mathrm{J}\,, \tag{117.27}$$

a factor 9.7 below Equation (117.26). A classical proton of that energy turns around at $b = Z_1 Z_2 e^2/(4\pi\varepsilon_0 E) = 39.5\,\mathrm{fm}$, nearly ten times the radius at which the nuclear force acts, and the yield would be exactly zero rather than merely small. There is no classical parameter to adjust: the observation is not of a rate lower than expected but of a rate where none was possible.

The quantum statement. The transmission of a Coulomb barrier between a contact radius R and a turning point b is a solved problem in this book and is not rebuilt here. Theorem 9.76 gives the semiclassical transmission as $T \approx \mathrm{e}^{-2\gamma}$ with $\gamma = \hbar^{-1}\int\sqrt{2\mu(V-E)}\,\mathrm{d}r$ taken between the turning points (Equation (9.87)), and Example 9.81 evaluates that integral for exactly this potential, in SI, obtaining

$$\gamma = \frac{\sqrt{2\mu E}\,b}{\hbar}\left[\arccos\sqrt{\xi} - \sqrt{\xi\,(1-\xi)}\right], \qquad \xi = \frac{R}{b}\,, \tag{117.28}$$

which is Equation (9.93). Only three inputs are needed to use it. The reduced mass of the proton–lithium system is

$$\mu = \frac{m_p m_{\mathrm{Li}}}{m_p + m_{\mathrm{Li}}} = 1.4626 \times 10^{-27}\,\mathrm{kg}\,, \qquad \mu c^2 = 820.5\,\mathrm{MeV}\,, \tag{117.29}$$

built from the nuclear masses of [2393] with the electrons removed, since it is the nuclei that move. The turning point b and the ratio ξ follow from E as above. And $\hbar$ is $1.054\,572 \times 10^{-34}\,\mathrm{J\,s}$, which is what makes γ dimensionless: the numerator of Equation (117.28) is a momentum times a length, in $\mathrm{kg\,m/s} \times \mathrm{m}$, and dividing by $\mathrm{J\,s}$ leaves a pure number.

At $E = 109.3\,\mathrm{keV}$ the momentum scale is $\sqrt{2\mu E} = 7.157 \times 10^{-21}\,\mathrm{kg\,m/s}$, $b = 39.5\,\mathrm{fm}$, $\xi = 0.1032$, the bracket is 0.9396, and

$$2\gamma = 5.04\,, \qquad T = \mathrm{e}^{-5.04} = 6.5 \times 10^{-3}\,. \tag{117.30}$$

The whole of Table 117.3 is this calculation repeated.

The shape of the yield curve. T itself is not what a yield curve measures. A proton entering a thick lithium target loses energy continuously to atomic electrons and has a chance of reacting at every energy from its incident value down to zero, so the yield per incident proton is

$$Y(E_0) = n \int_0^{E_0} \frac{\sigma(E)}{|\mathrm{d}E/\mathrm{d}x|}\,\mathrm{d}E\,, \qquad \sigma(E) = \frac{S(E)}{E}\,\mathrm{e}^{-2\gamma(E)}\,, \tag{117.31}$$

with n the number density of lithium nuclei and $S(E)$ the astrophysical S factor of Equation (107.100), which is slowly varying because it is defined to absorb both the barrier and the geometrical $1/E$. Evaluating Equation (117.31) with S constant and $|\mathrm{d}E/\mathrm{d}x|$ taken as constant over this narrow range gives the relative yields 1.00, 2.24, 7.02, 15.4, 27.5, 62.4, 109 and 225 at the eight accelerating potentials of Table 117.3, which run from 125 kV to 700 kV; they are its last column. The measured curve does rise by more than two orders of magnitude across that range, and the reason it rises faster than T does is the integral, not the barrier. □

Remark 117.6 (What the barrier factor does and does not explain). The transmission in Table 117.3 is of order 1×10^{-2}, while the observed yield per incident proton is of order 1×10^{-8}. The six orders of magnitude between them are not a defect of the calculation and are not tunnelling. A proton of 125 keV in lithium metal has a range of order $1 \times 10^{-6}\,\mathrm{m}$ and makes of order 1×10^{5} collisions with atomic electrons before stopping, in the course of which it passes within nuclear distances of essentially no nucleus at all; the geometrical probability of an encounter, and the nuclear matrix element once an encounter occurs, supply the rest of the suppression. What the barrier factor explains is the *shape* of the curve and the fact that there is any yield whatever at 125 kV. Reading it as an absolute prediction would be exactly the error Remark 9.77 warns against for the semiclassical method in general: the exponent is the reliable output, the prefactor is not.

Remark 117.7 (The exponent is small here, and that is unusual). Barrier problems in this book are usually deep in the semiclassical regime: the alpha decay of Example 9.81 has $2\gamma = 67$ and a transmission of 7×10^{-30}. Here 2γ runs from 5.0 down to 0.5, so the thick-barrier form Equation (9.87) is being used near the edge of its validity, and the two turning points are not far apart on the Airy scale at the upper end of the table. Neither refinement of the prefactor is exact — Remark 9.77 is explicit that Equation (9.86) is not, and offers the uniform form $T = \left(1 + \mathrm{e}^{2\gamma}\right)^{-1}$ as the alternative — but both move in the same direction, *downwards*, and by how much is the point. At the lowest row, $2\gamma = 5.04$: $\mathrm{e}^{-2\gamma} = 6.47 \times 10^{-3}$, Equation (9.86) gives 6.45×10^{-3} and the uniform form 6.43×10^{-3}, so the correction is below 1 % where the physical argument is made. At the last row, $2\gamma = 0.48$, the same three read 0.617, 0.463 and 0.381: a correction of 25 % and 38 % respectively, in a row where the proton is nearly over the barrier and the notion of tunnelling has stopped being the right description. The rows that carry the physical argument are the first three, and there the approximation is sound.

117.3.5 Primary references

Three papers report this work and they should be read in order. The *Nature* letter [694] is the announcement, four weeks after the observation, and it is where the 125 kV figure and the identification of the products as alpha particles first appear. The first *Proceedings* paper [696] is about the machine and not about the physics: it is the reference for the cascade generator, the accelerating tube and the beam, and it is the document that made the technique reproducible elsewhere. The second [695] is the full report of the disintegration, with the yield curve, the absorption measurements and the coincidence test.

The tunnelling interpretation rests on Gamow [600] and, independently, Gurney and Condon [603, 625], all published four years earlier for alpha decay; the Cockcroft–Walton experiment was in fact undertaken because Gamow's calculation, applied in reverse, showed that a few hundred kilovolts might suffice. That is worth noticing: this is one of the few experiments of the period designed from a quantum-mechanical prediction rather than interpreted by one afterwards.

117.4 Lawrence and Livingston: the cyclotron (1932)

Tests Phenomenon 117.8.

117.4.1 Apparatus

A flat evacuated brass chamber, about 0.28 m across — eleven inches, in the units of the paper — placed between the circular pole faces of an electromagnet whose field is perpendicular to its plane and uniform to a fraction of a per cent over the working region. Inside it, two hollow D-shaped electrodes, the "dees", face each other across a narrow diametral gap and are connected to a radiofrequency oscillator. The interior of each dee is a field-free region, being a closed conductor: the whole accelerating action takes place in the gap, and the dees exist to shield the ion from the alternating field during the half revolution in which that field would decelerate it.

An ion source sits at the centre, a filament and a gas feed producing hydrogen ions in the middle of the gap. A deflecting electrode at the periphery, held at a few kilovolts, pulls the beam out of its last orbit and onto a collector. The chamber was pumped to a residual pressure of order 1×10^{-4} Pa, which is set not by any technical convenience but by the physics: the ion travels a total path length of order

$$L = \sum_{k=1}^{N} 2\pi r_k = 2\pi R \sum_{k=1}^{N} \sqrt{\frac{k}{N}} \approx \frac{4\pi}{3} N R \approx 73\,\mathrm{m}\,, \tag{117.32}$$

inside the machine — using $r_k = R\sqrt{k/N}$, since the energy grows linearly with the crossing count and the radius as its square root — and must not scatter out of resonance on the way. Put the requirement in areal density, which is what stops a particle. A proton of 1.22 MeV has a range in air of about 3.3×10^{-2} kg/m^2, while 73 m of air at 1 Pa and 288.15 K — a density of 1.21×10^{-5} kg/m^3 by Equation (117.3) scaled with pressure — presents only 8.8×10^{-4} kg/m^2. So 1 Pa does not stop the beam outright; the pressure that would is about 40 Pa, and the residual pressure has to be far below even that, because a single elastic scatter through a small angle already takes the ion out of resonance. At 1×10^{-4} Pa the gas-kinetic mean free path $k_B T/(\sqrt{2}\pi d^2 p)$, with a molecular diameter $d \approx 3 \times 10^{-10}$ m and $T = 300$ K, is 1.0×10^{2} m — longer than the 73 m orbit, and for a 1 MeV proton, whose scattering cross-section is far below gas-kinetic, longer still. That is the margin the pump had to buy.

The magnet is the expensive component and the one that sets the energy. Everything else in the machine handles a few kilovolts.

117.4.2 Procedure

Ions released at the centre are bent by the magnetic field into a circle and cross the gap twice per revolution. If the oscillator frequency is matched to the revolution frequency, each crossing finds the gap field reversed and therefore accelerating, so the ion gains qV_{gap} at every crossing while its orbit grows as the square root of its energy.

The experiment consists of three measurements. First, tuning: the magnetic field is swept at fixed oscillator frequency and the collected current recorded, which produces a resonance curve whose peak locates the condition $\omega_c = \omega_{\mathrm{osc}}$ and whose width measures how far off resonance the machine will still accelerate. Second, the current at resonance, which measures how much beam survives the 73 m of orbit of Equation (117.32). Third, the energy of the extracted beam, measured by absorption in foils against the known range–energy relation for protons — the same technique as Section 117.3, and the same one that provides the only independent check that the ions really did make a hundred and fifty turns.

The resonance curve is the crucial one, because it is what distinguishes this machine from a peculiar high-voltage device. A current that appears only in a narrow band of magnetic field, centred on the value that makes Equation (117.34) equal the oscillator frequency, cannot be explained by any static potential in the apparatus.

117.4.3 Observations and data

Protons of kinetic energy above 1.2 MeV were obtained from a peak voltage on the dees of a few kilovolts, with beam currents of order 1×10^{-9} A at the collector. The ratio of the two voltages is the whole point: at about 4 kV per gap crossing and two crossings per turn, the final energy corresponds to something like a hundred and fifty revolutions, and the accelerating voltage that has to be insulated anywhere in the machine is smaller than the beam energy in electronvolts by more than two orders of magnitude.

Table 117.4 collects the measured quantities and the machine parameters that follow from them. The field and frequency are not independent measurements but consequences of the resonance condition and the extraction radius, and are computed here; they are included because they are what the resonance curve actually determines.

117.4.4 Interpretation

The cyclotron removed the ceiling that high-voltage technique imposes. Cockcroft and Walton's method needs the full energy to exist as a potential difference somewhere in the laboratory, and the dielectric strength of air sets a limit of a few million volts; the cyclotron needs only the gap voltage, and buys energy with turns instead of volts. The exchange is not free — it is paid for in orbit length, in vacuum, and in magnetic-field uniformity — but all three are technical problems with technical solutions, whereas insulation is a material property.

Everything downstream in this treatise depends on that trade: the beams of Chapter 116, the colliding beams of Chapter 113, and the reactor and accelerator neutrino sources of Chapter 114 are all descendants of it. So is a large part of nuclear astrophysics, since the reaction cross-sections that Chapter 52 needs — the cross-sections whose Gamow-peak energies are computed in Theorem 107.85 — are measured on machines of this lineage, at energies where the direct astrophysical rate is unmeasurably small.

The classical cyclotron has one limit, and it is relativistic. As the ion's energy grows its revolution frequency falls, the resonance detunes, and acceleration stops. The cure is to modulate the frequency (the synchrocyclotron) or to raise the field in step with the energy at fixed radius (the synchrotron); both are corrections demanded by Chapter 40 and by nothing else, and they are the reason no accelerator built after the 1940s is a cyclotron in

Table 117.4: The 1932 cyclotron: what was measured and what follows. The proton energy and the beam current are the reported observations; the magnetic field, the revolution frequency and the turn count are computed here from Equation (117.36) and Equation (117.34) with the quoted chamber size, and are the parameters the resonance condition fixes. The last two rows are the relativistic limit of the design, and are the reason no later machine was built this way. Bracketed uncertainties on the observed rows are estimates from the stated technique, the paper quoting none: 5 % on an energy read from a foil absorption curve, 4 % on a radius set by the position of the deflector, and an order of magnitude on the collected current. The computed rows inherit them — the momentum to 2.5 %, the field and the frequency to 5 %, the turn count, the orbit length and the phase slip to about 14 % — and are quoted with the propagated figure in brackets.

Quantity	Value	Kind, source
Chamber diameter	$(0.28 \pm 0.01)\,\mathrm{m}$	obs. 1932, [703]
Extraction radius	$(0.114 \pm 0.005)\,\mathrm{m}$	obs. 1932, [703]
Proton kinetic energy	$(1.22 \pm 0.06)\,\mathrm{MeV}$, i.e. $(1.955 \pm 0.096) \times 10^{-13}\,\mathrm{J}$	obs. 1932, [703]
Beam current at the collector	order $1 \times 10^{-9}\,\mathrm{A}$	obs. 1932, [703]
Peak dee voltage	$(4.0 \pm 0.5)\,\mathrm{kV}$	obs. 1932, [703]
Proton momentum at extraction	$(2.557 \pm 0.063) \times 10^{-20}\,\mathrm{kg\,m/s}$	calc.
Magnetic field required	$(1.40 \pm 0.07)\,\mathrm{T}$	calc., Equation (117.33)
Revolution frequency	$(21.3 \pm 1.1)\,\mathrm{MHz}$	calc., Equation (117.34)
Number of revolutions	153 ± 20	calc.
Total orbit length	$(73 \pm 10)\,\mathrm{m}$	calc., Equation (117.32)
Relativistic frequency shift at extraction	$(1.30 \pm 0.06) \times 10^{-3}$	calc.
Accumulated phase slip over the acceleration	$(0.62 \pm 0.09)\,\mathrm{rad}$, i.e. $(36 \pm 5)^\circ$	calc., Equation (117.38)

the original sense. The 1932 machine sits exactly at the boundary, which is what makes it worth computing the phase slip below rather than merely asserting the limitation.

Phenomenon 117.8 (Cyclotron resonance)**.** A non-relativistic ion of charge q and mass m moving in a uniform magnetic field of magnitude B circulates with an angular frequency that does not depend on its speed or on the radius of its orbit. A fixed-frequency alternating voltage across a single gap therefore stays in step with the ion for hundreds of revolutions and accelerates it at every crossing, so that the final energy is set by the magnet and not by the applied voltage [703].

Derivation. *Derives Phenomenon 117.8.*

For circular motion in a uniform field perpendicular to the orbit, the magnetic force supplies the centripetal acceleration,

$$qvB = \frac{mv^2}{r} \quad \Longrightarrow \quad r = \frac{mv}{qB} = \frac{p}{qB}\,, \tag{117.33}$$

in SI a radius in metres when p is in kg m/s, q in coulombs and B in tesla. The angular frequency is therefore

$$\omega_c = \frac{v}{r} = \frac{qB}{m}\,, \tag{117.34}$$

independent of v and of r: the ion takes the same time to complete a small slow circle as a large fast one, which is exactly the property that lets one fixed frequency serve the whole acceleration. It is a peculiarity of the Lorentz force and of no other force law that this is so; a linear restoring force would give the same property, and an inverse-square attraction would not.

Numbers. At the extraction radius $R = 0.114\,\mathrm{m}$ a proton of $1.22\,\mathrm{MeV}$ has momentum $p = \sqrt{2m_pT} = 2.557 \times 10^{-20}\,\mathrm{kg\,m/s}$, using $m_p = 1.6726 \times 10^{-27}\,\mathrm{kg}$ [2431], so Equation (117.33) requires

$$B = \frac{p}{eR} = \frac{2.557 \times 10^{-20}\,\mathrm{kg\,m/s}}{1.602\,177 \times 10^{-19}\,\mathrm{C} \times 0.114\,\mathrm{m}} = 1.400\,\mathrm{T}\,, \tag{117.35}$$

an ordinary iron-cored electromagnet, and Equation (117.34) then gives $\omega_c/2\pi = eB/2\pi m_p = 21.3\,\mathrm{MHz}$, a free-space wavelength of 14 m and an entirely standard radiofrequency in 1932.

Energy from the magnet, not the voltage. Eliminating v between Equation (117.33) and the kinetic energy gives the extraction energy at the outer radius R,

$$T = \tfrac{1}{2}mv^2 = \frac{q^2B^2R^2}{2m}\,, \tag{117.36}$$

which contains the gap voltage nowhere. Checking: $e^2B^2R^2/2m_p = 1.955 \times 10^{-13}\,\mathrm{J} = 1.22\,\mathrm{MeV}$, as observed. The number of crossings needed is $N_{\mathrm{cross}} = T/qV_{\mathrm{gap}} = 1.22\,\mathrm{MeV}/4\,\mathrm{keV} = 305$, that is 153 revolutions, and the gap voltage merely fixes how long the ion must stay in the machine. Note the scaling: $T \propto B^2R^2$, so the energy is quadratic in the size of the magnet, which is why the next machines were bigger rather than cleverer.

The limit. Relativistically the momentum is γmv with $\gamma = (1 - v^2/c^2)^{-1/2}$, so Equation (117.34) becomes $\omega = qB/\gamma m$: the frequency falls as the ion becomes energetic, and the resonance that Equation (117.34) guarantees is lost. For protons at $1.22\,\mathrm{MeV}$ the shift is

$$\gamma - 1 = \frac{T}{m_pc^2} = \frac{1.22\,\mathrm{MeV}}{938.27\,\mathrm{MeV}} = 1.30 \times 10^{-3}\,, \tag{117.37}$$

using the proton mass energy of [2427]. The consequence is cumulative and that is what has to be estimated. Suppose the energy grows linearly with turn number n, so that

$\gamma_n - 1 \approx (n/N)(\gamma_N - 1)$; each turn the ion arrives at the gap late by a fraction $\gamma_n - 1$ of a period, and after N turns the accumulated phase error is

$$\Delta\phi = 2\pi \sum_{n=1}^{N} (\gamma_n - 1) \approx \pi N \, (\gamma_N - 1) = \pi \times 153 \times 1.30 \times 10^{-3} = 0.62\,\mathrm{rad}\,, \qquad (117.38)$$

that is 36°. Acceleration continues as long as the ion arrives within a quarter period of the correct phase, so 36° is tolerable and the machine works. At ten times the energy the same estimate gives $\Delta\phi \approx 62\,\mathrm{rad}$, since both N and $\gamma_N - 1$ grow with T and the slip therefore grows as T^2; the ion would be decelerated for most of its life and never reach the rim. The classical cyclotron works precisely in the regime where Chapter 40 is not yet needed, and it stops working exactly where it is. □

Remark 117.9 (Why the resonance is not fragile). It is easy to read Equation (117.38) as saying that the machine requires the field and the frequency to be matched to a part in 10^3, and that is not quite right. What matters is the accumulated phase, so a static mismatch $\Delta\omega/\omega$ produces a slip $2\pi N \Delta\omega/\omega$, and tolerating a quarter period over 153 turns means $|\Delta\omega/\omega| < 1.6 \times 10^{-3}$ — which is a demanding but attainable tolerance on an iron magnet, and is precisely the width of the resonance curve that the experiment measured. The measured width is therefore not an instrumental artefact but a direct reading of the turn count, and it is the strongest evidence in the paper that the ions really were accelerated many times rather than once.

117.4.5 Primary references

The single reference is Lawrence and Livingston's *Physical Review* paper [703], whose title — the production of high-speed light ions without the use of high voltages — states the whole result. It should be read alongside [696], published the same year, since the two describe opposite solutions to the same problem and each is explicit about the limitation of the other's approach. Neither paper reports a nuclear reaction; the cyclotron's first disintegrations followed within months, using the lithium target of Section 117.3 as the obvious first check that the beam was real.

117.5 Wilson: the expansion cloud chamber (1912)

Tests Phenomenon 117.10.

117.5.1 Apparatus

A shallow cylindrical chamber with glass walls and roof, of order 0.16 m in diameter and a few centimetres deep, containing dust-free air saturated with water vapour. Its floor is a piston whose sudden withdrawal expands the gas; the expansion is completed in a time of order $1 \times 10^{-2}\,\mathrm{s}$, short compared with the time in which heat can reach the gas from the walls, so it is adiabatic and therefore cools the gas, leaving the vapour supersaturated.

Three further elements make it an instrument rather than a demonstration. The expansion is stopped mechanically at a *preset* ratio, because the ratio is the control variable and must be reproducible to better than a per cent. The chamber is illuminated from the side by a bright flash — the illumination must be transverse, since the droplets are seen by scattered light against a dark background — and photographed, in the developed form stereoscopically, so that the track can be reconstructed in three dimensions. And a clearing field of order $1 \times 10^3\,\mathrm{V/m}$ between electrodes above and below the gas sweeps out the ions left by previous events between expansions, without which the chamber accumulates a fog of old tracks.

Dust is the enemy and its removal is the reason the instrument took fifteen years to develop. Every dust grain is a condensation nucleus, and a chamber with dust in it fogs at any expansion. Wilson removed it by repeated expansions with the condensate allowed to settle, each cycle sweeping the remaining particles down with the droplets they nucleated.

117.5.2 Procedure

The expansion ratio is the control. Below a threshold no condensation occurs at all; above a first threshold droplets form on ions; above a second, higher one the vapour condenses on everything and the chamber fogs uniformly. Working between the two thresholds makes ions, and only ions, into condensation nuclei — and the existence of a window between the thresholds, rather than a single onset, is the physical fact on which the instrument rests. It is derived below.

The operating cycle is then fixed by the physics of droplet growth. The chamber is expanded in synchronism with the passage of the radiation under study — in the developed form, triggered by a counter telescope that has already registered the particle. The illumination is flashed a few hundredths of a second later: long enough for the droplets to grow by condensation to a size that scatters visible light efficiently, which requires a radius of order $1 \times 10^{-6}\,\mathrm{m}$, and short enough that diffusion has not yet blurred the trail. A water molecule in air at $270\,\mathrm{K}$ has a diffusion coefficient of order $2 \times 10^{-5}\,\mathrm{m^2/s}$, so in $2 \times 10^{-2}\,\mathrm{s}$ an ion wanders $\sqrt{2Dt} \approx 9 \times 10^{-4}\,\mathrm{m}$ — under a millimetre, which is why the track is still sharp at that delay and would not be at one second.

After the photograph the chamber is recompressed, the clearing field applied, and the cycle repeated, at a rate of a few per minute. That duty cycle — seconds of dead time for each millisecond of sensitivity — is the instrument's defining weakness and is what Section 117.7 eventually removed.

117.5.3 Observations and data

The photographs show the paths of individual ionising particles as sharp lines of droplets. Alpha particles give dense, straight, sharply terminated trails, occasionally with a sudden deflection near the end where a nuclear encounter has occurred; fast beta particles give thin, tortuous trails; X-rays give no trail at all but a scatter of short electron tracks starting at points along the beam.

That contrast is quantitative and is the first thing the instrument measures. The specific ionisation of a charged particle behaves as Equation (115.15), so a slow or multiply charged particle ionises far more densely than a fast singly charged one. For an alpha particle of $7.687\,\mathrm{MeV}$ depositing its whole energy in $7.0 \times 10^{-2}\,\mathrm{m}$ of air, the mean deposit is $1.10 \times 10^{8}\,\mathrm{eV/m}$, that is $1.76 \times 10^{-11}\,\mathrm{J/m}$; at a mean expenditure of about $34\,\mathrm{eV}$ — $5.4 \times 10^{-18}\,\mathrm{J}$ — per ion pair in air [2427], that is roughly $3.2 \times 10^{6}/\mathrm{m}$ of track. A minimum-ionising electron in the same gas loses about $2.2 \times 10^{5}\,\mathrm{eV/m}$, that is $3.5 \times 10^{-14}\,\mathrm{J/m}$, and makes about $6.5 \times 10^{3}/\mathrm{m}$. The ratio, some five hundred, is exactly the difference between a dense unbroken trail and a sparse dotted one, and it is why the two species are distinguishable by eye in a photograph.

The thresholds themselves are the chamber's calibration, and are collected in Table 117.5. The expansion ratios are Wilson's; the temperatures and supersaturations are computed here from them by Equations (117.39) and (117.40), with no free parameter — the initial temperature is the only input and the saturation vapour pressure cancels.

The tabulated droplet densities of the original paper are not transcribed here. What is reproduced is the calibration that makes the instrument an instrument: the two thresholds, their ratio, and the prediction of both from the thermodynamics of a charged drop.

Table 117.5: The condensation window of a Wilson chamber. The expansion ratio $r = V_2/V_1$ is the experimental control [394]; T_2 and the supersaturation S are computed here from Equations (117.39) and (117.40) for air initially saturated at $T_1 = 293.15\,\mathrm{K}$, using $\gamma = 1.40$ and the Clausius–Clapeyron relation with a latent heat of $2.50 \times 10^6\,\mathrm{J/kg}$. The last column is the critical supersaturation predicted by Equation (117.46) for a singly charged nucleus, or by the neutral nucleation barrier Equation (117.43) where marked; the agreement between the observed thresholds and the predictions is the content of Phenomenon 117.10. The expansion ratio is set mechanically and is reproducible to better than 0.5 %, which is the bracketed uncertainty; T_2 and S carry what that propagates to through Equations (117.39) and (117.40). The barrier in the last row is quoted with the constant $\sigma = 0.0765\,\mathrm{N/m}$ used throughout the derivation, and would fall to $36\,k_B T$ with the surface tension of supercooled water at 257.7 K; the difference is the honest size of the model's grip on the third figure.

$r = V_2/V_1$	T_2	S	What is observed, and what is predicted
1.250 ± 0.005	$(268.1 \pm 0.4)\,\mathrm{K}$	4.11 ± 0.11	droplets on negative ions; charged-nucleus threshold predicted at $S = 4.47$
1.310 ± 0.005	$(263.1 \pm 0.4)\,\mathrm{K}$	5.64 ± 0.15	droplets on positive ions as well; same prediction, and the sign asymmetry is outside the model
1.380 ± 0.005	$(257.7 \pm 0.4)\,\mathrm{K}$	8.08 ± 0.23	dense fog on no nucleus at all; neutral barrier here is $34\,k_B T$, the onset of macroscopic homogeneous nucleation

117.5.4 Interpretation

Before this apparatus, radiation was studied through the aggregate effects of very many particles — an ionisation current, a photographic blackening, a scintillation rate. The cloud chamber makes a single particle individually visible and individually measurable, and with a magnetic field applied it measures momentum and sign of charge from the curvature, through $p_\perp = ZeB\rho$ (Equation (115.14)).

Everything that follows in this book from the inspection of individual events begins here: the two-branch forks that identified Equation (117.5) [518], the electron–positron pairs photographed by Blackett and Occhialini [714] and treated in Chapter 95, and the cosmic-ray showers and strange particles of Chapter 115. In each case the decisive information is not a rate but a picture: how many branches, which way they curve, whether the momenta balance.

Its weaknesses defined the next two instruments. The gas is thin — a cubic metre of air holds 1.2 kg of matter and a cubic metre of liquid holds a tonne — so interactions are rare and the chamber is a poor target; and the vapour must be re-expanded and allowed to settle, so the chamber is dead for most of every minute. Section 117.6 fixes the first, Section 117.7 the second.

Phenomenon 117.10 (Ion tracks are made visible by condensation)**.** In a gas supersaturated with vapour, droplets condense preferentially on the ions left by the passage of a charged particle, and on nothing else, provided the supersaturation lies between the threshold for condensation on ions and the threshold for condensation on neutral molecules. The trail of droplets marks the particle's path and can be photographed [394].

Rests on Equation (115.15).

Derivation. *Derives Phenomenon 117.10.*

The claim has two halves — that an expansion produces a calculable supersaturation, and that a charge lowers the supersaturation at which a droplet can grow — and the instrument exists because the two thresholds so produced are different.

The supersaturation from the expansion. Air is expanded from volume V_1 to V_2 fast enough to be adiabatic and slowly enough to be reversible, so pV^γ is constant with $\gamma = c_p/c_V = 1.40$ for a diatomic gas, and $TV^{\gamma-1}$ is constant. Writing $r = V_2/V_1$,

$$T_2 = T_1\, r^{1-\gamma}\,. \tag{117.39}$$

The water vapour is a minor component and is carried along: its partial pressure falls both because the volume grew and because the temperature fell, by the ideal-gas law at fixed particle number, $p_{v2} = p_{v1}\,(V_1/V_2)\,(T_2/T_1) = p_{v1} r^{-\gamma}$. The saturation vapour pressure falls far faster, by the Clausius–Clapeyron relation integrated with a constant latent heat L of vaporisation per unit mass and a specific gas constant $R_v = R/M_{\mathrm{H_2O}} = 461.5\,\mathrm{J/(kg\,K)}$: $p_{\mathrm{sat}}(T_2) = p_{\mathrm{sat}}(T_1)\exp\left[-\,(L/R_v)\,(1/T_2 - 1/T_1)\right]$. Since the chamber starts saturated, $p_{v1} = p_{\mathrm{sat}}(T_1)$ and that quantity cancels entirely: the supersaturation $S = p_{v2}/p_{\mathrm{sat}}(T_2)$ is

$$S = r^{-\gamma}\exp\left[\frac{L}{R_v}\left(\frac{1}{T_2} - \frac{1}{T_1}\right)\right], \tag{117.40}$$

a function of the expansion ratio and the initial temperature alone. It requires no vapour-pressure table, which is what makes it a prediction rather than a fit; the one measured property it does need is the latent heat, and that enters only through the exponential. With $L = 2.50 \times 10^6\,\mathrm{J/kg}$ and $T_1 = 293.15\,\mathrm{K}$, Equation (117.39) gives $T_2 = 268.1\,\mathrm{K}$ at $r = 1.25$ and Equation (117.40) gives $S = 4.11$; the three rows of Table 117.5 are this calculation repeated.

Why a charge helps. Consider forming a spherical droplet of radius ρ from vapour at supersaturation S, the droplet carrying a charge q at its centre. Three terms contribute to

the free energy of formation. The surface costs $4\pi\rho^2\sigma$ with σ the surface tension. The bulk gains, because the vapour is supersaturated: each molecule transferred from vapour to liquid lowers the free energy by $k_B T \ln S$, and a droplet of radius ρ holds $\frac{4}{3}\pi\rho^3 n_l$ of them, with the number density of liquid water — mass density over molar mass, times the Avogadro constant, so that it is not an independent input either —

$$n_l = \frac{1.000 \times 10^3\,\mathrm{kg/m^3} \times 6.022\,140\,76 \times 10^{23}/\mathrm{mol}}{18.0153 \times 10^{-3}\,\mathrm{kg/mol}} = 3.343 \times 10^{28}/\mathrm{m^3}\,. \tag{117.41}$$

And the electrostatic self-energy of the charge falls as the droplet grows around it, because the charge's field is being moved out of vacuum and into a medium of relative permittivity ε_r. The geometry matters and is easy to get backwards: the *droplet* is the dielectric, so it is the region $a < r < \rho$ between the bare ion of radius a and the droplet surface whose permittivity changes, while the field beyond ρ is unaltered. The energy stored there falls from $(q^2/8\pi\varepsilon_0)\,(1/a - 1/\rho)$ to $(q^2/8\pi\varepsilon_0\varepsilon_r)\,(1/a - 1/\rho)$, so relative to the bare ion the change is $-\,(q^2/8\pi\varepsilon_0)\,(1 - 1/\varepsilon_r)\,(1/a - 1/\rho)$, whose ρ-independent piece is a constant of the ion and drops out. Collecting the terms that depend on ρ,

$$\Delta G(\rho) = 4\pi\rho^2\sigma - \tfrac{4}{3}\pi\rho^3 n_l k_B T \ln S + \frac{q^2}{8\pi\varepsilon_0\rho}\left(1 - \frac{1}{\varepsilon_r}\right), \tag{117.42}$$

which is Thomson's expression [332], in SI, with every term in joules. Setting $q = 0$ recovers the classical nucleation problem: the surface term wins at small ρ and the bulk term at large, so there is a barrier at $\rho_c = 2\sigma/\left(n_l k_B T \ln S\right)$ of height

$$W_c = \frac{16\pi\sigma^3}{3\left(n_l k_B T \ln S\right)^2}\,, \tag{117.43}$$

and nothing condenses unless a thermal fluctuation supplies W_c.

The two thresholds. With $q \neq 0$ the third term dominates at small ρ, so ΔG rises to $+\infty$ as the droplet shrinks back onto the bare ion — but it *falls* as ρ grows, with a downhill slope $q^2\,(1 - 1/\varepsilon_r)\,/8\pi\varepsilon_0\rho^2$ that diverges as $\rho \to 0$, while the surface term's uphill slope $8\pi\rho\sigma$ vanishes there. Near the ion the free energy is therefore downhill throughout, and the barrier, if one survives at all, is pushed out to larger ρ and lowered. Above some supersaturation it disappears altogether, and a droplet grows on the ion without any fluctuation at all. Setting $\mathrm{d}\Delta G/\mathrm{d}\rho = 0$ and asking when the resulting equation has no root gives the condition directly. Dividing $\mathrm{d}\Delta G/\mathrm{d}\rho = 0$ through by $4\pi\rho^2$, the stationarity condition reads

$$n_l k_B T \ln S = \frac{2\sigma}{\rho} - \frac{q^2\,(1 - 1/\varepsilon_r)}{32\pi^2\varepsilon_0\rho^4}\,. \tag{117.44}$$

The right-hand side rises from $-\infty$, peaks, and falls as $2\sigma/\rho$; its maximum is at

$$\rho_* = \left[\frac{q^2\,(1 - 1/\varepsilon_r)}{16\pi^2\varepsilon_0\sigma}\right]^{1/3}, \tag{117.45}$$

where the second term is one quarter of the first, so the peak value is $\frac{3}{4} \times 2\sigma/\rho_*$. If the left-hand side exceeds that peak there is no stationary point at all, ΔG decreases monotonically, and the droplet grows freely. The threshold is therefore

$$\ln S_{\mathrm{crit}} = \frac{3\sigma}{2\,n_l k_B T\,\rho_*}\,. \tag{117.46}$$

Numbers. At $T = 268.1\,\mathrm{K}$, with $\sigma = 0.0765\,\mathrm{N/m}$, $\varepsilon_r = 88$ for water at that temperature, and $q = e = 1.602\,177 \times 10^{-19}\,\mathrm{C}$, Equation (117.45) gives $\rho_* = 6.19 \times 10^{-10}\,\mathrm{m}$ — about

33 water molecules — and Equation (117.46) gives $S_{\mathrm{crit}} = 4.47$. The observed threshold for condensation on ions is $S = 4.11$, from the first row of Table 117.5. Meanwhile at that same supersaturation the neutral barrier Equation (117.43) is $2.45 \times 10^{-19}\,\mathrm{J} = 66\,k_BT$, so the probability of a spontaneous fluctuation forming a droplet is of order e^{-66} and nothing whatever condenses on neutral molecules. At the third row, $S = 8.08$ and $T = 257.7\,\mathrm{K}$, the same barrier has fallen to $1.21 \times 10^{-19}\,\mathrm{J} = 34\,k_BT$. The prefactor here is the number density of *vapour* molecules and not the liquid density n_l: the partial pressure of the water vapour after the expansion is of order $1 \times 10^3\,\mathrm{Pa}$, so p/k_BT is of order $1 \times 10^{23}/\mathrm{m}^3$. With molecular attempt frequencies of order $1 \times 10^{12}/\mathrm{s}$ and a Boltzmann factor $\mathrm{e}^{-34} = 1.7 \times 10^{-15}$, the nucleation rate is of order $1 \times 10^{20}/(\mathrm{m}^3\,\mathrm{s})$ — that is 1×10^{14} droplets per cubic centimetre per second, a dense fog immediately. That is the second threshold. The conclusion is insensitive to the prefactor by a wide margin: getting it wrong by five orders of magnitude, which is the size of the gap between the vapour and liquid densities, is worth a change of 12 in W_c/k_BT — and W_c/k_BT moves from 66 to 34 across the two rows.

The window between $S = 4.5$ and $S = 8$ is therefore not an empirical convenience but a consequence of Equation (117.42): the charge removes the barrier while the neutral vapour still has one. Operating inside it makes ions, and only ions, into condensation nuclei, which is the statement of Phenomenon 117.10. □

Remark 117.11 (Where this derivation is honest and where it is not). Two things in the calculation above deserve flagging rather than burying. First, $\rho_* = 0.62\,\mathrm{nm}$ contains about thirty molecules, and applying a macroscopic surface tension and a bulk permittivity to an object that size is not defensible in detail; that the answer comes out within 10 % of the measurement is better than the method deserves, and the right claim is that the mechanism and the order of magnitude are established, not the third figure. Second, Equation (117.42) depends on the charge only through q^2 and therefore predicts the same threshold for positive and negative ions, whereas Wilson measured $r = 1.25$ for negative and $r = 1.31$ for positive, a difference of 37 % in S. The asymmetry is real and is outside this model entirely: it involves the orientation of the permanent dipole of the water molecule at the droplet surface, which no theory built from a scalar ε_r can see. Reporting the model as though it accounted for both thresholds would misstate what it does.

Third, the thermophysical inputs are not cited, because this bibliography carries no source for them: the latent heat of vaporisation $2.50 \times 10^6\,\mathrm{J/kg}$, the surface tension $0.0765\,\mathrm{N/m}$ and the static permittivity $\varepsilon_r = 88$ are standard tabulated properties of water taken at the stated temperature, and the reader should treat them as handbook values rather than as claims this book supports. They matter unequally: S depends on L through an exponential, ρ_* on $\sigma^{-1/3}$ and only logarithmically on ε_r, and W_c on σ^3. The surface tension is the one that carries the third figure, and it is also the one that is least defensible at this droplet size, as the first point above says. Holding σ fixed at its 268.1 K value across the table, rather than letting it run with temperature, is what separates the $34\,k_BT$ quoted in the last row of Table 117.5 from the $36\,k_BT$ the temperature-dependent value would give.

117.5.5 Primary references

Wilson's 1912 paper [394] is the instrument paper and reads like one: most of it is about dust, valves, timing and the photography, and the tracks appear at the end as the demonstration that the apparatus works. The physics of the condensation window is older — it is Wilson's own work of the eighteen-nineties on the two expansion thresholds — and the theory of the charged drop is Thomson's [332], given there in the chapter on the condensation of vapour on ions.

The papers that made the instrument famous are not Wilson's. Blackett's identification of the nitrogen reaction [518], Blackett and Occhialini's photographs of electron–positron

pairs [714], and the cosmic-ray work of Chapter 115 are all cloud-chamber results, and each is a case where the answer was visible in a photograph and inaccessible to any counting experiment then available.

117.6 Glaser: the bubble chamber (1952)

Tests Phenomenon 117.12.

117.6.1 Apparatus

A small, scrupulously clean glass bulb of a few cubic centimetres, filled with diethyl ether and held under pressure at a temperature above its normal boiling point of 307.7 K. In the working condition the liquid sat near 413 K, where its saturated vapour pressure is about 1.48 MPa, roughly fifteen atmospheres; the bulb was held above that pressure so that the liquid stayed liquid. Releasing the pressure to about 0.10 MPa leaves the liquid superheated by $\Delta p \approx 1.38$ MPa: it is far above its boiling point at the prevailing pressure and yet, in the absence of nucleation sites, does not boil.

A radioactive source could be brought up to the bulb or removed, or interposed behind lead, and the onset of boiling was registered — in the first experiments by timing, later photographically.

Cleanliness is the apparatus. A superheated liquid is metastable against boiling everywhere, and the barrier that protects it, Equation (117.50) below, is enormous only for a *homogeneous* liquid. A dissolved gas bubble, a dust grain, or a scratch on the glass provides a surface of large radius of curvature at which the barrier is absent, so the vessel boils at once and no measurement is possible. Glass was chosen for its smoothness, the bulbs were sealed after prolonged degassing, and the ether was purified; the experimental difficulty of the whole subject is in that sentence.

117.6.2 Procedure

The bulb is taken through a cycle: compress, heat, expand into the superheated state, wait. The quantity measured is the delay between the expansion and the onset of boiling, recorded many times over with the source present and with it absent, so that a shift in the distribution of delays is a statistical statement rather than an anecdote.

Two features of that protocol matter. The delay is a random variable, because nucleation is a stochastic process — a barrier crossing has an exponentially distributed waiting time — so the observable is a distribution and not a number, and the comparison must be between distributions. And the null run is not “no source” but “source shielded”, which controls for everything about the thermal cycle while changing only the radiation reaching the liquid.

The degree of superheat was varied, since the barrier Equation (117.50) depends on it as $(\Delta p)^{-2}$ and therefore very steeply: too little superheat and even a radiation track cannot nucleate, too much and the liquid boils spontaneously and the chamber is useless. Finding the window is the same experimental problem as finding Wilson’s condensation window in Section 117.5, and has the same two-sided structure.

117.6.3 Observations and data

Superheated ether that stayed liquid for a long time when shielded boiled promptly when an ionising source was brought near, and the distribution of delays shifted with the source strength. The nucleation is therefore caused by the radiation. Photographs of bubbles strung out along the paths of the ionising particles — the bubble chamber proper — followed shortly after.

Table 117.6 gives the thermodynamic state of the liquid in the working condition, with the critical bubble it implies. The temperature and the pressures are the operating point of the experiment; the surface tension, critical radius and barrier are computed here, and the surface tension in particular is an estimate rather than a measurement, obtained by scaling the room-temperature value with $(1 - T/T_c)^{1.26}$ to the working temperature, which is the corresponding-states form [879]. It is flagged as such because the barrier depends on its cube.

Table 117.6: The superheated state in Glaser's bulbs, and the critical bubble it implies. Rows marked *est.* are computed here from corresponding-states scaling and the Clausius–Clapeyron relation and are order-of-magnitude statements, not measurements; the barrier depends on σ^3 and inherits three times the relative uncertainty of σ. The last two rows are the comparison that decides the mechanism: the barrier is far too high for a thermal fluctuation and far too low to stop a charged particle's energy deposit. The two temperatures and the released pressure are the operating point and are quoted with the uncertainty an operating point of this kind carries; the rows marked *est.* have no meaningful bracket, since their uncertainty is the validity of the scaling and not a measurement error, and the caveat above is the honest statement of it. The two rows marked *tabulated* are handbook properties of diethyl ether for which this bibliography carries no source.

Quantity	Value	Kind
Normal boiling point of diethyl ether	307.7 K	tabulated, uncited
Critical point	466.7 K, 3.64 MPa	tabulated, uncited
Working temperature	(413 ± 5) K	obs. 1952, [1000]
Saturated vapour pressure there	1.48 MPa	est.
Pressure after release	(0.10 ± 0.01) MPa	obs. 1952, [1000]
Superheat Δp	1.38 MPa	est.
Surface tension at 413 K	3.9×10^{-3} N/m	est., [879]
Critical bubble radius r_c	5.7×10^{-9} m	calc., Equation (117.48)
Nucleation barrier W_c	5.2×10^{-19} J, i.e. 3.3 eV	calc., Equation (117.50)
$W_c / k_B T$	91.5	calc.
Energy a minimum-ionising particle deposits over $2r_c$	1.6 eV, i.e. 2.6×10^{-19} J	est.
Energy a 5 MeV alpha deposits over $2r_c$	1.6×10^{3} eV, i.e. 2.6×10^{-16} J	est.

The measured delay distributions are not transcribed. What the table establishes is the quantitative frame in which they are read, and the frame is one-sided. A barrier of $91.5\,k_B T$ is essentially never crossed by a thermal fluctuation — the derivation below puts homogeneous nucleation at about one event per clean cubic centimetre per week — while it is crossed at once where a heavily ionising particle has passed. The delays of seconds to minutes actually observed in the shielded runs are therefore *not* homogeneous nucleation: they are set by residual heterogeneous sites and by the ambient background radiation, which is why the cleanliness of the bulb is the apparatus and why the comparison is made between distributions rather than against a computed spontaneous rate.

117.6.4 Interpretation

The bubble chamber is the cloud chamber with the phase transition run backwards, and the exchange is decisive: the sensitive medium is now a liquid, roughly a thousand times denser than a gas, so it is a target as well as a detector. Air at 288.15 K has a density of $1.225\,\mathrm{kg/m^3}$ and diethyl ether about $714\,\mathrm{kg/m^3}$, a factor 580; liquid hydrogen, at $70.8\,\mathrm{kg/m^3}$ [2427], is still a factor 58. Interactions that a cloud chamber would see once a month happen continuously, and the whole event — production point, decay point, and every charged secondary — is recorded in one picture.

Filled with liquid hydrogen the chamber becomes a pure proton target, which is what made it the workhorse of hadron physics; the Gargamelle heavy-liquid chamber's photographs of neutrino interactions with no outgoing muon or electron [1502] are the discovery of the weak neutral current, and hence part of the evidential base of Chapter 116 and of the electroweak theory. That result is a good illustration of the instrument's character: the signature is the *absence* of a track, which is a statement no counting experiment could make and only a complete picture of the final state can support.

Its limitation is the one Charpak removed. A bubble chamber cannot be triggered: the liquid is sensitive only during the few milliseconds after the expansion, so it must be expanded on a schedule synchronised with the beam and photographed whether or not anything interesting happened, and the film must then be scanned and measured by human beings. The rate is set by mechanics and by scanning tables, not by physics.

Phenomenon 117.12 (Ionising radiation nucleates bubbles in a superheated liquid)**.** A liquid held above its boiling point without boiling begins to boil promptly when ionising radiation passes through it, and the bubbles form along the path of the ionising particle [1000].

Rests on Equation (115.15).

Derivation. *Derives Phenomenon 117.12.*

Mechanical equilibrium. A spherical bubble of radius r in a liquid holds its vapour at a pressure exceeding the liquid's by the Laplace term: balancing the force on a hemisphere, the pressure difference across the surface times the cross-section πr^2 equals the surface tension σ times the circumference $2\pi r$, so

$$p_{\mathrm{vap}} - p_{\mathrm{liq}} = \frac{2\sigma}{r}\,, \tag{117.47}$$

in SI a pressure in pascals when σ is in N/m and r in metres. In a liquid superheated by $\Delta p = p_{\mathrm{vap}} - p_{\mathrm{liq}} > 0$, Equation (117.47) is satisfied at exactly one radius,

$$r_c = \frac{2\sigma}{\Delta p} = \frac{2 \times 3.9 \times 10^{-3}\,\mathrm{N/m}}{1.38 \times 10^{6}\,\mathrm{Pa}} = 5.7 \times 10^{-9}\,\mathrm{m}\,, \tag{117.48}$$

and that equilibrium is unstable: a bubble smaller than r_c has too much curvature, is crushed and collapses, while one larger than r_c finds the vapour pressure winning and grows without limit.

The barrier. Nothing happens spontaneously because reaching r_c costs work. The free energy of a bubble of radius r is the surface term $4\pi r^2\sigma$ less the work $\frac{4}{3}\pi r^3 \Delta p$ done against the pressure difference,

$$W(r) = 4\pi r^2\sigma - \tfrac{4}{3}\pi r^3 \Delta p\,, \tag{117.49}$$

whose derivative $8\pi r\sigma - 4\pi r^2 \Delta p$ vanishes at $r = r_c$, confirming Equation (117.48), and whose value there is the barrier

$$W_c = \frac{16\pi\sigma^3}{3\,(\Delta p)^2} = 5.2 \times 10^{-19}\,\mathrm{J} = 3.3\,\mathrm{eV}\,. \tag{117.50}$$

Compared with $k_B T = 5.70 \times 10^{-21}\,\mathrm{J}$ at 413 K this is $91.5\,k_B T$, so the Boltzmann factor is $\mathrm{e}^{-91.5} = 1.8 \times 10^{-40}$: with of order 1×10^{28} molecular sites per cubic metre and attempt frequencies of order $1 \times 10^{12}/\mathrm{s}$, the homogeneous nucleation rate is of order $1.8/(\mathrm{m}^3\,\mathrm{s})$, that is $1.8 \times 10^{-6}/\mathrm{s}$ in a clean cubic centimetre — one event a week, against a cycle time of seconds. That is why the liquid can be kept superheated at all. It is also why the observed delays cannot be homogeneous nucleation: the calculation is not a prediction of the shielded-run delay distribution but a demonstration that homogeneous nucleation contributes nothing to it.

Why a track nucleates. A charged particle passing through deposits energy in a narrow cylinder along its path, in localised amounts far larger than thermal. Take a minimum-ionising particle in ether: the mass stopping power is about $0.2\,\mathrm{MeV\,m^2/kg}$, so at a liquid density of $714\,\mathrm{kg/m^3}$ — a handbook value for which this bibliography carries no source — it loses about $1.43 \times 10^8\,\mathrm{eV/m}$, that is $2.29 \times 10^{-11}\,\mathrm{J/m}$, and over the critical diameter $2r_c = 11.3\,\mathrm{nm}$ it deposits about 1.6 eV, that is $2.6 \times 10^{-19}\,\mathrm{J}$ — which is *below* W_c. A 5 MeV alpha particle, with a range in ether of some tens of micrometres, deposits about 1×10^3 times more per unit length and delivers of order 1.6 keV, that is $2.6 \times 10^{-16}\,\mathrm{J}$, over the same distance, far above W_c.

Two consequences follow and both are observed: the boiling delay collapses when a sufficiently ionising source is present, and the bubbles appear *along the track*, because that is where the energy was deposited. Note also from Equation (117.50) that the barrier falls as $(\Delta p)^{-2}$, steeply with increasing superheat, which is why the chamber must be expanded into a narrow window and photographed at once — the operational counterpart of the two thresholds in Phenomenon 117.10. □

Remark 117.13 (The energy budget is not as simple as the comparison above). The comparison at the end of the derivation is the standard one and it is too generous. Not all the energy a particle deposits within a critical diameter is available to the nucleus: most of it goes into electronic excitation and into heating a region much larger than r_c, and only the fraction thermalised inside the critical volume, within the time the bubble needs to reach r_c, can pay for the barrier. The efficiency is of order a few per cent, which is why real chambers require a deposit one to two orders of magnitude above W_c, and why the bubble density along a track is a usable measure of $\mathrm{d}E/\mathrm{d}x$ — Equation (115.15) — rather than a threshold effect. The correct treatment follows the local temperature field produced by the track and asks when a supercritical bubble appears in it; it is a heat-conduction problem with a source, and it is what fixes the operating superheat of a real chamber.

Derivation pending #212

The thermal-spike calculation of bubble nucleation: the temperature field around an ionising track from the heat equation with a line source, the fraction of the deposited energy thermalised within the critical radius before the bubble can grow, and the resulting threshold in $[\mathrm{d}E/\mathrm{d}x]$ *as a function of the superheat. It belongs in Appendix A; the derivation above establishes the barrier and the mechanism but treats the energy delivery as instantaneous and perfectly local.*

117.6.5 Primary references

Glaser's paper [1000] is a two-page letter, and it is about the delay statistics of boiling in a superheated liquid; the track photographs that give the instrument its name came afterwards. It is worth reading as an example of an experiment whose published result is a shift in a distribution — there is no picture of a bubble in it — and whose importance is entirely in what the shift implies.

The instrument's scientific yield is documented elsewhere in this book rather than in Glaser's paper. The Gargamelle neutral-current photographs [1502] are the case discussed above; the hydrogen-chamber results on hadron spectroscopy underlie Chapter 116, and the technique's decline is recorded in Section 117.7.

117.7 Charpak: the multiwire proportional chamber (1968)

Tests Phenomenon 117.14.

117.7.1 Apparatus

A plane of fine parallel anode wires, of the order of 20 µm in diameter and spaced 2 mm apart, stretched midway between two parallel cathode planes about 8 mm away on either side, the whole in a gas mixture of a noble gas with a polyatomic quenching additive. The wires are held at a positive potential of a few kilovolts with respect to the cathodes.

Each of those dimensions is doing work. The wire must be thin because the field at its surface, and hence the avalanche gain, depends on its radius — Equation (117.54) — and a thick wire gives no multiplication at usable voltages. The spacing must be small because it sets the position resolution directly. The cathode gap must be several times the spacing, or the cells stop being independent and the field near each wire is distorted by its neighbours. And the quenching additive is not optional: a pure noble gas de-excites by emitting ultraviolet photons that liberate photoelectrons from the cathode metres away, so the avalanche propagates and the counter goes into permanent discharge. A polyatomic molecule absorbs those photons into vibrational modes and dissipates them, confining the avalanche to the wire that started it.

The decisive feature, however, is not the mechanics but the readout: every wire has its own amplifier and its own discriminator, so the chamber emits a pattern of logic pulses rather than an image. In 1968 that was newly affordable, and it is the reason the paper was written when it was.

117.7.2 Procedure

A charged particle crossing the gas leaves a track of electron–ion pairs — of order 100 of them in 8 mm of argon at atmospheric pressure, at a mean expenditure of about 26 eV, that is 4.2×10^{-18} J, per pair [2427]. The electrons drift towards the nearest anode wire through the low-field region between the cathodes, and in the last few tens of micrometres, where the field rises as $1/r$ and exceeds the threshold for ionisation by collision, each initiates an avalanche whose gain is set by the wire voltage and the gas.

The measurements reported are those any new counter owes. The gas gain was measured against anode voltage, by comparing the collected charge with the known primary ionisation. The detection efficiency was measured against voltage with an external trigger — a scintillator telescope defining a particle's passage — which produces the characteristic plateau, a range of voltage over which efficiency is near unity and insensitive to the setting, and finding that plateau is what makes a counter usable. The time resolution was measured against the same trigger. And, the result that was not obvious in advance, the pulses on wires adjacent to the struck one were recorded, to establish that a chamber with wires two millimetres apart really does behave as a row of independent counters and not as one large one.

117.7.3 Observations and data

Each wire behaved as an independent proportional counter, with adjacent wires showing negligible cross-talk despite being only millimetres apart. Detection efficiency was high over

a plateau of several hundred volts, the position was given to within the wire spacing, and both the time resolution — of the order of the maximum electron drift time computed in Table 117.7, which is what bounds it — and the rate a single wire could sustain were better than any photographic technique by orders of magnitude.

Uncertainties are owed on all of this and the paper quotes none, so what Table 117.7 carries are estimates from the mechanical tolerance of a wound wire plane and from the stability of a high-voltage supply, each flagged in the table. The efficiency and the time resolution are reported here as the paper reports them, in words and without a figure, because assigning them a numerical uncertainty would mean inventing the measurement they belong to; what the argument below needs from them is only that the plateau exists and that the resolution is drift-limited, and both are qualitative statements.

Table 117.7 gives the geometry, the reported performance, and the electrostatic quantities those imply, computed here from Equations (117.53) and (117.54) and from Equations (117.56) and (117.57). Table 117.8 is the avalanche gain against anode voltage, computed from Equation (117.55) with representative Townsend coefficients for argon; it is included because the steepness of that dependence is the whole reason a proportional chamber has an operating plateau at all, and because it is a closed-form consequence of the $1/r$ field.

117.7.4 Interpretation

This is the change of era. A cloud or bubble chamber produces a photograph, and a photograph must be developed, scanned and measured by a person; the experiment's rate is therefore set by human labour. A bubble chamber cycles at a few hertz at best, and a scanner works through pictures at a rate of order one per minute, so a large experiment of the nineteen-sixties accumulated of order 1×10^6 pictures over years and employed rooms full of people to measure them. A multiwire chamber produces a number, immediately, on a wire, and a computer can read it; a single wire sustains of order 1×10^5/s and the limit is the electronics.

Triggering becomes possible with it — the apparatus can be told to record only events satisfying a condition, evaluated in the time between one particle and the next — and with it the whole strategy of searching for rare processes in an overwhelming background, which is how every result in Chapter 113 was obtained. Modern detectors record of order 1×10^9 interactions per second and select a handful for storage, and the reason they can is that the sensitive element gives an electronic address rather than an image.

The device also founded a family. Measuring the drift time of the electrons, rather than merely which wire fired, turns the wire spacing into a far finer coordinate: an electron drifting at 5×10^4 m/s covers 50 µm in a nanosecond, so timing to a nanosecond gives a resolution an order of magnitude better than Equation (117.56) — that is the drift chamber. Extending the drift volume to metres and reading the arrival pattern on an endplate gives the time projection chamber; replacing wires with etched electrodes on a substrate gives the micropattern detectors of current practice. All of them are read out as numbers, and all of them descend from this experiment.

Phenomenon 117.14 (Each anode wire is an independent counter)**.** In a plane of parallel anode wires only millimetres apart between two cathode planes, the avalanche produced by an ionising particle develops on a single wire and induces a signal there, with negligible response on its neighbours. The chamber therefore reports position as the address of the wire that fired, at a rate limited by electronics rather than by photography [1367].

Rests on Theorem 58.10.

Derivation. *Derives Phenomenon 117.14.*

Table 117.7: The multiwire proportional chamber: geometry, performance and the electrostatics that follow. Geometry and performance are the reported values [1367]; the capacitance, line charge, surface field, multiplication radius, position resolution and drift and ion times are computed here in the derivation below. The signal charge assumes a gain of 1×10^4. Bracketed uncertainties on the observed geometry are estimates from the mechanical tolerance of a wound wire plane, the paper quoting none; the computed rows are exact consequences of that geometry and inherit it, the capacitance and everything built on it to about 1.5 %. The efficiency row carries no bracket for the reason given in the text, and the primary ion-pair count is an order-of-magnitude estimate while the gain is a chosen reference value.

Quantity	Value	Kind, source
Anode wire diameter	$(20 \pm 1)\,\mu\text{m}$	obs. 1968, [1367]
Wire spacing s	$(2.00 \pm 0.05)\,\text{mm}$	obs. 1968, [1367]
Anode–cathode gap L	$(8.0 \pm 0.1)\,\text{mm}$	obs. 1968, [1367]
Anode potential	3 kV to 4 kV, stable to 10 V	obs. 1968, [1367]
Detection efficiency, over a plateau several hundred volts wide	near unity	obs. 1968, [1367]
Capacitance per unit length of one wire	3.47 pF/m	calc., Equation (117.53)
Line charge at 3 kV	$1.04 \times 10^{-8}\,\text{C/m}$	calc.
Field at the wire surface	$1.87 \times 10^{7}\,\text{V/m}$	calc., Equation (117.54)
Radius within which multiplication occurs	62 µm	calc.
Primary ion pairs in the gap	about 100	est.
Signal charge at gain 1×10^4	$1.6 \times 10^{-13}\,\text{C}$	calc.
Position resolution	0.58 mm	calc., Equation (117.56)
Maximum electron drift time	160 ns	calc.
Ion signal time constant t_0	1.8 ns	calc., Equation (117.57)
Total ion collection time	1.14 ms	calc., Equation (117.58)

Table 117.8: Avalanche gain against anode potential, computed here from Equation (117.55) for the geometry of Table 117.7 in argon at 101.3 kPa, with Townsend coefficients $A = 9.0/(\text{m Pa})$ and $B = 135\,\text{V}/(\text{m Pa})$ representative of argon, and a multiplication threshold $E^* = 3 \times 10^6\,\text{V/m}$. The gain rises by nearly four orders of magnitude for a factor 2 in voltage, which is why the operating point is chosen on an efficiency plateau and not by tuning the gain.

Anode potential	Surface field	Multiplication radius	Gain M
2.0 kV	$1.25 \times 10^{7}\,\text{V/m}$	42 µm	1.5×10^{1}
2.5 kV	$1.56 \times 10^{7}\,\text{V/m}$	52 µm	6.8×10^{1}
3.0 kV	$1.87 \times 10^{7}\,\text{V/m}$	62 µm	3.6×10^{2}
3.5 kV	$2.18 \times 10^{7}\,\text{V/m}$	73 µm	2.1×10^{3}
4.0 kV	$2.50 \times 10^{7}\,\text{V/m}$	83 µm	1.3×10^{4}
4.5 kV	$2.81 \times 10^{7}\,\text{V/m}$	94 µm	8.1×10^{4}

The field. Model one wire as a long straight conductor of radius a carrying charge λ per unit length. Gauss's theorem over a coaxial cylinder gives

$$E(r) = \frac{\lambda}{2\pi\varepsilon_0 r}\,, \tag{117.51}$$

in V/m when λ is in C/m, and hence a potential varying only logarithmically, $V \propto \ln(r/a)$. The periodic geometry has to be summed, and the sum is worth doing because it fixes λ and hence every electrostatic number in this section. Put the wires along y at $y = ns$, $n \in \mathbb{Z}$, each carrying λ, and let x measure distance from the wire plane. Superposing the logarithmic potentials of the whole row and using the product formula $\sin(\pi w)/(\pi w) = \prod_{n\geq 1}(1 - w^2/n^2)$, applied to the complex combination $w = (y + \mathrm{i}x)/s$, gives the closed form

$$V(x,y) = -\frac{\lambda}{4\pi\varepsilon_0}\ln\left[2\left(\cosh\frac{2\pi x}{s} - \cos\frac{2\pi y}{s}\right)\right] + \text{const}\,. \tag{117.52}$$

Two limits of Equation (117.52) are all that is needed. Close to one wire, $r = \sqrt{x^2 + (y - ns)^2} \ll s$, expanding both hyperbolic and circular functions to second order gives $2(\cosh - \cos) \to (2\pi r/s)^2$ and hence $V \to -(\lambda/2\pi\varepsilon_0)\ln(2\pi r/s)$, which recovers Equation (117.51) and shows that the row looks like a single wire from close up. Far from the plane, $|x| \gg s$, the cosine is negligible against $\cosh(2\pi x/s) \to \tfrac{1}{2}\mathrm{e}^{2\pi|x|/s}$, and the potential becomes linear, $V \to -(\lambda/2\varepsilon_0)\,|x|/s$: the wire plane looks like a charged sheet, as it must. That second limit is the reason the cathode gap has to be several times the wire spacing: at the $L/s = 4$ of this chamber the terms dropped are smaller than the one kept by a factor 2×10^{-11}, while at $L/s = 1$ they would not be negligible and the cells would stop being independent. Subtracting the second limit at the cathode, $|x| = L$, from the first at the wire surface, $r = a$, gives the working potential $V = (\lambda/2\pi\varepsilon_0)\,[\pi L/s - \ln(2\pi a/s)]$, so that the capacitance per unit length of one wire is

$$C = \frac{\lambda}{V} = \frac{2\pi\varepsilon_0}{\dfrac{\pi L}{s} - \ln\dfrac{2\pi a}{s}} = 3.47 \times 10^{-12}\,\mathrm{F/m} \tag{117.53}$$

for $a = 10\,\mu\mathrm{m}$, $s = 2\,\mathrm{mm}$ and $L = 8\,\mathrm{mm}$, so at $V = 3\,\mathrm{kV}$ the line charge is $\lambda = CV = 1.04 \times 10^{-8}\,\mathrm{C/m}$ and the field at the wire surface is

$$E(a) = \frac{\lambda}{2\pi\varepsilon_0 a} = 1.87 \times 10^{7}\,\mathrm{V/m}\,. \tag{117.54}$$

Two facts follow from the $1/r$ dependence, and between them they are the whole device.

First: the multiplication is local. Avalanche multiplication requires an electron to gain the ionisation energy of the gas between collisions, that is a field above some threshold E^* of order $3 \times 10^6\,\mathrm{V/m}$ in argon at atmospheric pressure; by Equation (117.51) the region where $E > E^*$ is a cylinder of radius $\lambda/2\pi\varepsilon_0 E^* = 62\,\mu\mathrm{m}$, a few wire radii. The multiplication is therefore confined to the immediate neighbourhood of one wire, thirty times closer to it than to its neighbours, and the gain does not depend on where in the gas volume the original ionisation occurred — which is what makes the signal proportional to the deposited charge, and what keeps the wires independent.

The gain itself follows in closed form. The first Townsend coefficient — the number of ionising collisions an electron makes per metre — is well represented by $\alpha = Ap\exp(-Bp/E)$ with p the pressure, and the multiplication is $M = \exp\int \alpha\,\mathrm{d}r$ taken inward from $E = E^*$ to $E = E(a)$. Changing variable from r to E using Equation (117.51), so that $r = k/E$ with $k = \lambda/2\pi\varepsilon_0$ and $\mathrm{d}r = -k\,\mathrm{d}E/E^2$, the integral is elementary:

$$\ln M = Apk\int_{E^*}^{E(a)} \frac{\mathrm{e}^{-Bp/E}}{E^2}\,\mathrm{d}E = \frac{Ak}{B}\left(\mathrm{e}^{-Bp/E(a)} - \mathrm{e}^{-Bp/E^*}\right)\,. \tag{117.55}$$

With $A = 9.0/(\mathrm{m\,Pa})$ and $B = 135\,\mathrm{V}/(\mathrm{m\,Pa})$ for argon, $p = 1.013 \times 10^5\,\mathrm{Pa}$ and $k = \lambda/2\pi\varepsilon_0$ proportional to V through Equation (117.53), this generates Table 117.8: a gain of 15 at 2 kV and 1.3×10^4 at 4 kV. The extreme steepness is structural — $\ln M$ is roughly linear in V, so M is exponential in it — and it is why the operating point of a proportional chamber is chosen by finding the efficiency plateau rather than by setting a gain.

Second: the cells are defined by the field, not by walls. The drift region between the cathodes is a low-field region in which electrons simply move to the nearest wire along the field lines, and the division of the plane into cells belongs to the field geometry: no walls are needed between adjacent counters, which is why the wires can be packed at millimetre pitch. A particle's coordinate is then the address of the wire that fired, uniformly distributed over the cell of width s, so the resolution is the standard deviation of a uniform distribution,

$$\sigma_x = \frac{s}{\sqrt{12}} = \frac{2\,\mathrm{mm}}{3.464} = 0.58\,\mathrm{mm}\,. \tag{117.56}$$

Third: the signal is fast because it is induced by the ions. The electrons produced in the avalanche are collected in a fraction of a nanosecond and induce almost nothing, having been created within a few micrometres of the wire; the signal is the induced charge from the positive ions drifting *away* from the wire through the whole logarithmic field. An ion of mobility μ obeys $\mathrm{d}r/\mathrm{d}t = \mu E(r) = \mu k/r$, so $r^2(t) = a^2 + 2\mu k t$ and the induced charge grows as $\ln(1 + t/t_0)$ with

$$t_0 = \frac{a^2}{2\mu k} = \frac{\pi\varepsilon_0 a^2}{\mu\lambda} = 1.8\,\mathrm{ns}\,, \tag{117.57}$$

for $\mu = 1.5 \times 10^{-4}\,\mathrm{m^2/(V\,s)}$ [2427]. The signal therefore rises on a nanosecond scale even though full ion collection is slow. The same $r^2(t)$ run out to the cathode, at $r = L$ and not $L/2$ — L is the whole anode-to-cathode gap of Table 117.7, which is also the distance the electron drift time of that table is computed over — gives

$$t_{\max} = \frac{L^2 - a^2}{2\mu k} \approx \frac{\pi\varepsilon_0 L^2}{\mu\lambda} = 1.14\,\mathrm{ms}\,, \tag{117.58}$$

six hundred thousand times t_0, since $t_{\max}/t_0 = (L/a)^2$. Differentiating the signal with a short time constant keeps the fast leading edge and discards the long tail, which is what makes the coordinate available promptly enough to be used in a trigger. The residual limit on rate is the space charge of those slow ions, which is what fixes the 1×10^5/s per wire quoted above. □

Remark 117.15 (What is representative and what is measured here). The Townsend coefficients used in Equation (117.55) and the ion mobility used in Equation (117.57) are representative values for argon at atmospheric pressure, not values measured in Charpak's chamber, which ran on a gas mixture whose composition changes both by factors of order unity. They are used to show that the closed form reproduces the observed orders of magnitude — a gain of 1×10^4 at a few kilovolts, a multiplication region of tens of micrometres, a nanosecond leading edge — and the argument does not depend on their precise values. The measurement of electron transport in argon on which such coefficients ultimately rest is Townsend and Bailey's [495].

117.7.5 Primary references

The paper is Charpak and collaborators' [1367], in *Nuclear Instruments and Methods*, and it is short. Its significance is easy to miss on a first reading because it claims very little: that a plane of closely spaced wires can be operated with each wire as an independent proportional counter, with the gains, efficiencies and time resolutions tabulated. The claim

that was not obvious — and that the paper's referees would have been right to doubt — is the independence of the wires at millimetre spacing, which Equation (117.54) makes plausible and the measurements established.

What followed is not in this paper and is not any single paper. The technique displaced the bubble chamber over the following decade, and the detectors of Chapter 116, Chapter 114 and Chapter 113 are all, in their tracking elements, its descendants.

117.8 Summary of the evidence

Table 117.9: The reactions and the instruments reported in this chapter. The first two rows are nuclear reactions, the third is the machine that freed the projectile energy from the insulation limit, and the last three are the detectors that made reactions countable one at a time. The three detectors are best read by what each removed: the invisibility of the single particle, the thinness of the target, and the photograph.

Experiment	Year	Result
Rutherford	1919	first artificial transmutation; nitrogen yields protons of up to 5.93 MeV against an elastic ceiling of 4.94 MeV
Cockcroft–Walton	1932	first accelerator-driven nuclear reaction, and the first quantitative nuclear test of $E = mc^2$: alpha particles of about 8.6 MeV measured against 8.79 MeV predicted from a mass difference of 3.09×10^{-29} kg
Lawrence–Livingston	1932	protons of 1.22 MeV from 4 kV per gap, in 153 turns at 1.40 T
Wilson	1912	individual charged particles made visible as droplet trails, in a supersaturation window between $S = 4.1$ and $S = 8.1$
Glaser	1952	the sensitive medium becomes a liquid 580 times denser than air, hence target and detector at once
Charpak	1968	position read as an electronic address to 0.58 mm, at 1×10^5/s per wire instead of a photograph

Table 117.9 contains no fitted parameter and no theoretical input beyond mass–energy equivalence, elementary electrostatics, the semiclassical barrier transmission of Section 9.6, and classical nucleation theory. The first two rows establish that nuclei can be rearranged at will and that the energy bookkeeping of Chapter 40 holds when they are; the third makes the projectile energy an adjustable quantity rather than a gift of nature; the last three establish that the rearrangements can be counted, one at a time, at rates that rose by roughly six orders of magnitude across the half-century the table spans. Neither half is of any use without the other, which is why they are reported together here rather than split between a nuclear chapter and a detector appendix.

Two threads run through all six and are worth naming, because they recur in every later experiment in this part. The first is that the decisive measurement is almost never the one the experiment appears to be making. Rutherford's result is a comparison between

gases, not a count; Cockcroft and Walton's is a coincidence rate, not a yield; Lawrence's is the *width* of a resonance curve, which is what proves the turn count; Glaser's is a shift in a distribution of delays. In each case the number that carries the physics is a ratio or a correlation, chosen because it is insensitive to the quantities the experimenter cannot control.

The second is that every one of the six instruments is a threshold device operated inside a deliberately chosen window. The cloud chamber must sit between two supersaturations, the bubble chamber between two superheats, the proportional chamber between the onset of multiplication and the onset of discharge, the scintillation counter between the eye's sensitivity and its fatigue, the cyclotron between resonance and relativistic detuning, the cascade generator between usable voltage and breakdown. Locating those windows is what the *Procedure* of each experiment above actually describes, and it is the part of experimental physics that no theory supplies.

Part XII

Statistical Mechanics

Chapter 118

Statistical Mechanics

Thermodynamics, developed in Part VI, states relations between measured macroscopic quantities without ever asking what matter is made of; statistical mechanics derives those same relations from the mechanics of the constituents, and in doing so predicts things thermodynamics cannot — the magnitude of fluctuations, the absolute value of the entropy, and the point at which the classical description must fail. This chapter opens Part XII by building that bridge: phase space and Liouville's theorem inherited from Chapter 22, the Boltzmann and Gibbs definitions of entropy, the microcanonical, canonical and grand canonical ensembles with their partition functions, the thermodynamic limit in which they agree, and the classical ideal gas as the case where every step can be carried through in closed form. It states the ergodic hypothesis explicitly and reports its status honestly — proved for a small class of systems, assumed for the rest — rather than passing over the gap in the foundations.

The chapter is deliberately classical: the counting is corrected for indistinguishability where the measurements demand it (the Gibbs paradox and the Sackur–Tetrode entropy), but the quantum statistics that explain why are deferred to Chapter 120, the kinetic and transport theory to Chapters 55 and 124, and phase transitions to Chapter 123. Its direct experimental sequel is Chapter 119, where the fluctuations predicted here are measured and Avogadro's constant extracted from them. Founding sources are Boltzmann [196] and Gibbs [304]; the modern treatments used throughout are [1656, 2180], with the thermodynamic side in [1749].

Derivation pending #213

Statistical Mechanics: all derivations of this chapter are pending.

118.1 From mechanics to thermodynamics

118.1.1 Why a statistical description

[Reserved: the impossibility in practice and the irrelevance in principle of following $\sim 10^{23}$ trajectories; Maxwell's introduction of a distribution of molecular velocities as the object of study rather than the trajectories themselves [154], and Boltzmann's transport equation for its evolution [183]; the division of labour with Chapter 55 (dynamics of the distribution) and Chapter 54 (relations among the averages); what "equilibrium" means operationally.]

118.1.2 Phase space and Liouville's theorem

[Reserved: the $6N$-dimensional phase space of Chapter 22, the ensemble density $\rho(q,p,t)$, and Liouville's theorem $\partial_t \rho + \{\rho, \mathcal{H}\} = 0$ as the statement that the flow preserves phase-space

volume, with the symplectic geometry of Chapter 24 behind it; stationarity of any ρ that depends on the phase point only through conserved quantities; Gibbs's formulation of the whole subject in these terms [304].]

118.1.3 Macrostates, microstates and coarse graining

[Reserved: the distinction between a microstate and the macrostate specified by a few extensive variables; the multiplicity W of a macrostate; the coarse-grained density and why the fine-grained Gibbs entropy is constant under Liouville evolution while the observed entropy is not — the Ehrenfests' analysis of the issue [370]; the cell size and its cancellation from all entropy differences, with the absolute value fixed only by h (Chapter 120).]

118.2 Entropy and the Boltzmann principle

118.2.1 The combinatorial definition

[Reserved: Boltzmann's 1877 counting of complexions — the distribution of a fixed energy among discrete cells, the multiplicity of each distribution, and the identification of the most probable distribution with equilibrium [196]; the Stirling limit; the logarithm as the only function turning multiplicative multiplicities into an additive quantity; the recovery of the Maxwell–Boltzmann distribution as the maximiser.]

118.2.2 The relation between entropy and multiplicity

[Reserved: $S = k \log W$, written in that form by Planck when he introduced the constant k in the radiation analysis [300]; the constant as a conversion factor between energy and temperature, now exact by the 2019 redefinition of the kelvin [2333] with the value quoted in Chapter 2 and [2392]; the sense in which the relation is a definition and the sense in which it is a testable claim.]

118.2.3 The H-theorem and the reversibility objections

[Reserved: the functional $H = \int f \log f \, \mathrm{d}^3 v$ and Boltzmann's proof that it decreases under the collision term [183]; Loschmidt's reversibility objection — the reversed microstate has increasing H [195]; the recurrence objection built by Zermelo [273] on Poincaré's recurrence theorem [247], and the estimated recurrence time; where the arrow enters — the molecular-chaos assumption — stated as an assumption and not concealed; the initial-condition question left open and referred to Chapter 48.]

118.2.4 The Gibbs entropy

[Reserved: $S = -k \int \rho \log \rho \, \mathrm{d}\Gamma$ and its discrete form $-k \sum_i p_i \log p_i$ [304]; agreement with the Boltzmann expression in the thermodynamic limit and disagreement away from it; the concavity and additivity properties; its identification with the Shannon functional of Section 118.8.1.]

118.3 Ensembles

118.3.1 The microcanonical ensemble

[Reserved: the postulate of equal a priori probabilities on the energy shell, the density of states $\Omega(E, V, N)$, and $S = k \log \Omega$ [304]; temperature, pressure and chemical potential as

derivatives of S; the shell thickness and its irrelevance; the two-subsystem derivation of the zeroth law and of the equality of temperatures at equilibrium.]

118.3.2 The canonical ensemble

[Reserved: a system in contact with a reservoir, the Boltzmann factor $\mathrm{e}^{-\beta E_i}$, and the partition function $Z = \sum_i \mathrm{e}^{-\beta E_i}$ [304]; the Helmholtz free energy $F = -kT \log Z$ as the generating function of every thermodynamic quantity; the energy distribution and its sharpness; the classical phase-space integral with the $1/(N!h^{3N})$ prefactor justified in Section 118.4.3; standard development in [1656].]

118.3.3 The grand canonical ensemble

[Reserved: exchange of particles as well as energy, the fugacity $z = \mathrm{e}^{\beta\mu}$, and the grand partition function $\Xi = \sum_N z^N Z_N$ [304]; the grand potential $\Phi = -kT \log \Xi = -pV$; particle-number fluctuations and their relation to the isothermal compressibility; why this is the natural ensemble for the quantum gases of Chapter 120.]

118.3.4 Equivalence of ensembles

[Reserved: the coincidence of the three ensembles for all intensive quantities as $N \to \infty$, and the $O(N^{-1/2})$ relative size of the differences; the Legendre-transform structure relating S, F and Φ; the conditions under which equivalence fails — long-range interactions (Section 118.5.3) and at a first-order transition; rigorous statements in [1407].]

118.3.5 Partition functions in practice

[Reserved: factorisation for non-interacting degrees of freedom; translational, rotational and vibrational partition functions of a molecule and the resulting heat capacities, compared with measured values for diatomic gases; the classical virial expansion for weak interactions; the systematic use of $\log Z$ as a cumulant generator, linking to Chapter 11; thermodynamic identities collected in [1749].]

118.4 The classical ideal gas

118.4.1 The Maxwell–Boltzmann distribution

[Reserved: the speed distribution $f(v) \propto v^2 \mathrm{e}^{-mv^2/2kT}$ [154] and Boltzmann's generalisation to an arbitrary external potential [171]; the mean, root mean square and most probable speeds; the direct measurement of the distribution in a potassium and thallium atomic beam by Miller and Kusch, agreeing with the predicted form to about 1 % [1076]; the barometric formula as the same law in gravity, which is the observable used in Chapter 119.]

118.4.2 Equipartition and its failure

[Reserved: $\tfrac{1}{2}kT$ per quadratic degree of freedom [171]; the Dulong–Petit law for solids [64] as its most successful instance; the failures Maxwell singled out as the greatest difficulty facing the molecular theory — the missing vibrational and rotational contributions to diatomic heat capacities, and the fall of solid heat capacities at low temperature [191]; the resolution by level quantisation, carried out in Chapters 68 and 125 and Chapter 120.]

118.4.3 The Gibbs paradox

[Reserved: the spurious entropy of mixing between two samples of the same gas, and Gibbs's own statement and discussion of it [193]; the $1/N!$ correction and the resulting extensivity of S; the fact that classical mechanics gives no reason for the factor, so that the paradox is evidence for indistinguishability before quantum mechanics supplies it (Chapters 83 and 120); the operational reading in terms of which mixings can be undone.]

118.4.4 The Sackur–Tetrode entropy

[Reserved: the absolute entropy of a monatomic ideal gas, $S = Nk[\log(V/N\lambda^3) + \frac{5}{2}]$ with the thermal wavelength $\lambda = h/\sqrt{2\pi mkT}$, obtained independently by Sackur [380] and Tetrode [393]; the appearance of Planck's constant in a classical result and what that signals; the vapour-pressure test — Tetrode's comparison with mercury vapour data, and the chemical-constant measurements that followed — as the first quantitative confirmation; the Sackur–Tetrode constant among the CODATA quantities [2392].]

118.5 The thermodynamic limit

118.5.1 Existence and extensivity

[Reserved: the limit $N \to \infty$, $V \to \infty$ at fixed N/V; van Hove's proof that the free energy per particle converges for interactions with a hard core and a sufficiently rapid decay [934], and the systematic treatment in [1407]; stability and tempering as the hypotheses; why the limit is what makes thermodynamics sharp rather than approximate.]

118.5.2 Analyticity and its failure

[Reserved: the partition function of a finite system as an analytic function of β, so that no finite system has a phase transition; the Yang–Lee account of how non-analyticity appears in the limit through the pinching of the real axis by zeros of the partition function [1014]; the transfer of this to Chapter 123.]

118.5.3 Long-range forces and non-extensive systems

[Reserved: systems in which the energy is not extensive — self-gravitating matter above all; negative heat capacity and the gravothermal catastrophe [1375]; the inequivalence of ensembles that follows, and why a self-gravitating system has no maximum-entropy equilibrium; the treatment of stellar and compact matter in Chapters 49 and 52, where these caveats are load-bearing.]

118.6 The ergodic hypothesis: status

118.6.1 What the hypothesis is asked to do

[Reserved: the identification of time averages along a trajectory with ensemble averages over the energy shell; the Ehrenfests' distinction between the original ergodic hypothesis (the trajectory passes through every point), which is false for dimensional reasons, and the quasi-ergodic hypothesis [370]; the role the assumption actually plays — justifying the microcanonical postulate of Section 118.3.1 — and the alternatives to it.]

118.6.2 The ergodic theorems

[Reserved: Birkhoff's pointwise ergodic theorem, giving almost everywhere convergence of the time average [674], and von Neumann's mean ergodic theorem in L^2 [705]; metric transitivity as the condition under which the limit is the ensemble average; Khinchin's argument that only sum functions of many degrees of freedom need behave well, which weakens what must be proved [936].]

118.6.3 Systems proved ergodic

[Reserved: Sinai's proof of ergodicity and mixing for a gas of hard elastic spheres in a box [1433]; hyperbolicity and sensitive dependence as the mechanism (Chapter 32); the smallness of the class — no realistic interatomic potential has been shown ergodic — stated plainly.]

118.6.4 Honest status of the foundations

[Reserved: the Fermi–Pasta–Ulam–Tsingou numerical experiment, in which a weakly nonlinear chain failed to equipartition and instead recurred [1065], and the KAM tori that explain it; the resulting position — statistical mechanics works far better than its proofs justify — recorded as an open problem rather than argued away; the pragmatic defence by timescales; forward reference to Chapter 133.]

118.6.5 Thermalisation in isolated quantum systems

[Reserved: the eigenstate thermalisation hypothesis of Deutsch [1842] and Srednicki [1887], in which individual energy eigenstates already encode thermal expectation values; the measured approach to thermal equilibrium of an isolated lattice of ultracold bosons through the growth of entanglement entropy [2272]; the relation to decoherence and open systems in Chapter 87.]

118.7 Fluctuations

118.7.1 Energy and particle-number fluctuations

[Reserved: $\langle \Delta E^2 \rangle = kT^2 C_V$ from the canonical ensemble and the analogous compressibility relation in the grand canonical ensemble [304]; the relative size $O(N^{-1/2})$, which is why thermodynamics appears deterministic; the inversion of the argument — a measured fluctuation determines k and hence Avogadro's constant — which is the strategy of Chapter 119; Einstein's general fluctuation formula [361].]

118.7.2 Critical opalescence

[Reserved: the divergence of the compressibility near a critical point, the resulting long-wavelength density fluctuations, and the strong scattering of light that makes a critical fluid turn milky; Smoluchowski's kinetic account [349] and Einstein's quantitative theory of the scattered intensity [361]; the observation as a direct visual demonstration that fluctuations are real; critical exponents deferred to Chapter 123.]

118.7.3 Thermal noise

[Reserved: Johnson's measurement of a voltage noise across a resistor proportional to resistance and to absolute temperature and independent of material [608], and Nyquist's derivation of $\langle V^2 \rangle = 4kTR\,\Delta f$ from equilibrium statistical mechanics alone [615]; the measurement as a second independent determination of k; the general fluctuation–dissipation statement developed in Chapter 124.]

118.8 Entropy, information and inference

118.8.1 The Shannon functional

[Reserved: Shannon's uniqueness derivation of $-\sum p_i \log p_i$ from continuity, monotonicity and the composition rule [921]; its formal identity with the Gibbs entropy of Section 118.2.4; the care needed in reading that identity physically — the physical entropy carries units of J/K *and refers to a specified macroscopic description.]*

118.8.2 Maximum entropy as inference

[Reserved: Jaynes's derivation of the canonical and grand canonical distributions by maximising the Shannon entropy subject to the known expectation values, with the Lagrange multipliers becoming β and $\beta\mu$ [1114]; what this reformulation buys — a uniform derivation and a clear account of which assumptions are doing the work — and what it does not settle, namely why the resulting distributions describe the system after a finite time; Lagrange multipliers as in Chapter 16.]

118.8.3 Information and thermodynamic cost

[Reserved: Szilard's one-molecule engine and the entropy cost of a measurement [638]; Landauer's principle, that erasing one bit dissipates at least $kT \log 2$ [1213]; the experimental verification with a colloidal particle in a double-well optical trap, confirming the bound at the level of a few percent of kT [2191]; the exorcism of Maxwell's demon that follows, and its relation to the second law as stated in Chapter 54.]

118.8.4 Negative absolute temperature

[Reserved: a bounded-energy spectrum admitting population inversion and hence $\partial S/\partial E < 0$; the nuclear spin system of lithium fluoride prepared by Purcell and Pound, which relaxed through an inverted state with a measurable negative spin temperature [986]; Ramsey's demonstration that such states are consistent thermodynamic equilibria hotter than any positive temperature [1096]; the motional degrees of freedom of an ultracold lattice gas prepared at negative temperature [2214]; why this does not violate the second law.]

118.8.5 Beyond equilibrium

[Reserved: the Jarzynski equality relating the free-energy difference to an exponential average of the work done in arbitrarily fast processes [1938], and its verification by mechanically unfolding a single RNA molecule [2033]; the second law recovered as an inequality from an equality, with rare trajectories of negative entropy production observed; the systematic development in Chapter 124.]

Chapter 119

Experiment: Brownian Motion and Avogadro's Number

At the start of the twentieth century the existence of atoms was still a live question: the energetic school held that thermodynamics needed no molecules, and no measurement had yet counted them. Brownian motion settled it. The theory of Chapter 118 predicts that a particle large enough to be watched under a microscope, but small enough for the bombardment it receives to be unbalanced, must wander, and that the wandering must obey a quantitative law containing Avogadro's constant and nothing else that is unknown. Perrin measured that law four independent ways — from sedimentation equilibrium, from translational displacements, from rotational displacements, and from diffusion — and each gave the same number to within his stated errors. That agreement, not any single value, is what ended the argument.

This chapter reports the measurement as an experiment: the emulsions and optics Perrin used, the counting procedure, the observed numbers with their uncertainties in SI units, the interpretation, and the modern repetitions that reach the regime Einstein himself declared unobservable — the instantaneous velocity of a Brownian particle. Written out it will use the structured `experiment` environment of the treatise, in the order of the sections below. Its theoretical neighbours are Chapter 118 for the fluctuation formulae, Chapter 55 for the molecular picture, Chapter 124 for the general fluctuation–dissipation relation, and Chapter 11 for the stochastic processes the episode created. Primary sources are Brown [82], Einstein [318], Smoluchowski [331], Langevin [344] and Perrin [358, 409].

Derivation pending #214

Experiment: Brownian Motion and Avogadro's Number: all derivations of this chapter are pending.

119.1 Historical context and the prediction under test

119.1.1 Brown's observation

[Reserved: the ceaseless irregular motion of pollen-derived particles suspended in water, and Brown's decisive control — the same motion in particles ground from window glass, from rock and from a fragment of the Sphinx, which excluded any vital cause [82]; the seventy years of failed explanations (convection, evaporation, capillarity, electricity) and the observations that excluded each; the qualitative facts any theory had to reproduce — the motion never stops, is faster for smaller particles, and is faster in less viscous liquids.]

Phenomenon 119.1 (Ceaseless irregular motion of suspended grains)**.** Particles about a micrometre across, suspended in a liquid and watched under a microscope, execute a perpetual irregular motion that never ceases, is more vigorous the smaller the particle and the less viscous the liquid, and does not depend on the chemical nature or the origin of the particle: Brown found the same agitation in pollen-derived grains, in particles ground from window glass and from rock, and in a fragment of the Sphinx, which excludes any vital cause [82]. No external agency — convection, evaporation, illumination or capillarity — has been found that switches it off, and the motion persists indefinitely in a sealed, thermally equilibrated cell.

Derivation pending #215

Molecular origin of the agitation: the unbalanced momentum delivered by molecular impacts on a grain of radius a, and the reason the relative size of the resulting fluctuation grows as the grain shrinks

119.1.2 The atomic hypothesis under attack

[Reserved: the positivist objection that atoms are unobservable and therefore not physics, argued from the economy of thought [225]; the energetic programme's claim that thermodynamics is complete without them; what would count as an answer — not a plausibility argument but a measurement of how many molecules there are; Perrin's own framing of the question [409], and the connection to the epistemology of Chapter 1.]

119.1.3 Einstein's prediction

[Reserved: the 1905 paper deriving, from the osmotic pressure of the suspended particles and Stokes drag, the diffusion coefficient $D = RT/(6\pi\eta a N_A)$ and the mean square displacement $\langle x^2 \rangle = 2Dt$ [318]; the crucial point that the displacement grows as $\sqrt{t}$, not as t, so that a velocity measured over any accessible interval is meaningless — the statement Einstein made explicitly and that Section 119.7.3 eventually overturned; the second paper extending the treatment and adding rotational diffusion $\langle \theta^2 \rangle = RTt/(4\pi\eta a^3 N_A)$ [326].]

119.1.4 Smoluchowski and Sutherland

[Reserved: Smoluchowski's independent kinetic derivation by direct collision counting, differing from Einstein's coefficient by a numerical factor and reaching the same scaling [331]; Sutherland's independent statement of the same diffusion relation, obtained while estimating molecular masses from solution data [324]; the resulting name Stokes–Einstein–Sutherland relation, and the honest attribution.]

119.1.5 The Langevin equation

[Reserved: the phenomenological equation of motion $m\ddot{x} = -6\pi\eta a\dot{x} + F(t)$ with a random force of zero mean and vanishing correlation time, and Langevin's derivation of $\langle x^2 \rangle = 2Dt$ from it in a page — what he called an infinitely simpler demonstration [344]; the momentum relaxation time $\tau_p = m/(6\pi\eta a)$ that the overdamped limit discards, and which reappears as the observable of Section 119.7.3; the equation as the first stochastic differential equation.]

119.2 Apparatus

119.2.1 The gamboge and mastic emulsions

[Reserved: gamboge resin and mastic emulsified in water and purified by repeated fractional centrifuging to obtain grains uniform in radius to a few percent; typical radii from $0.2\,\mu\mathrm{m}$ *to* $0.5\,\mu\mathrm{m}$ *and densities near* $1.2\times10^{3}\,\mathrm{kg/m^{3}}$*; the months of preparation required for a usable fraction, and why monodispersity is the controlling systematic of the whole measurement [358, 409].]*

119.2.2 Microscope, cell and recording

[Reserved: the shallow observation cell of depth about $100\,\mu\mathrm{m}$ *sealed against evaporation and convection; the short depth of focus used to select a horizontal layer; the calibrated eyepiece grid and the camera lucida for recording successive positions at fixed time intervals of* $30\,\mathrm{s}$*; temperature control and its uncertainty; the counting of grains per layer by direct enumeration [358].]*

119.2.3 Calibration of the grain radius

[Reserved: three independent determinations of the grain radius — from Stokes-law sedimentation velocity, from direct microscopic measurement of a chain of grains, and from weighing a counted number of grains — and their mutual agreement as the internal check on the most sensitive input; the viscosity of water at the working temperature and its uncertainty; the propagation of these into the final result [358], in the manner of Chapter 2 and [1945].]

119.3 Procedure

119.3.1 Sedimentation equilibrium

[Reserved: the vertical distribution of grains in equilibrium as a barometric law with the molecular mass replaced by the buoyant mass of a grain, so that the scale height is measurable under a microscope; the counting of grains at four levels separated by tens of micrometres; the exponential fit and its slope giving N_A *directly; the demonstration that a visible suspension obeys the same statistics as a gas [358].]*

Phenomenon 119.2 (Sedimentation equilibrium of a visible suspension)**.** A dilute suspension of uniform grains, left undisturbed, settles into a stationary vertical distribution in which the number density falls off exponentially with height,

$$n(z) = n(0)\,\mathrm{e}^{-m^{*}gz/k_{B}T}\,, \qquad m^{*} = \tfrac{4}{3}\pi a^{3}\,(\rho-\rho_{f})\,, \tag{119.1}$$

where m^{*} is the buoyant mass of a grain of radius a and density ρ suspended in a fluid of density ρ_{f}. Perrin measured the scale height by counting grains at four levels separated by tens of micrometres and found the law obeyed over a fourfold change of concentration [358]. A suspension one can see therefore obeys the same barometric statistics as a gas one cannot.

Derivation. Treat the grains as a dilute solute, so rare that they do not interact with one another; their osmotic pressure is then that of an ideal gas, $p = nk_{B}T$. In mechanical equilibrium the gradient of that pressure must support the net downward force per unit volume, which is the weight of the grains less the buoyancy of the fluid they displace,

$$\frac{\mathrm{d}p}{\mathrm{d}z} = -nm^{*}g\,. \tag{119.2}$$

Substituting $p = nk_BT$ at uniform temperature gives $k_BT\,\mathrm{d}n/\mathrm{d}z = -nm^*g$, a first-order linear equation whose solution is Equation (119.1). Equivalently, the grains are distributed over the potential energy $U = m^*gz$ with the Boltzmann weight e^{-U/k_BT}; the two readings agree, which is precisely the point at issue, since the second treats a visible object as a statistical-mechanical degree of freedom. Because m^* is known from the measured radius and densities, the slope of $\ln n$ against z delivers $k_B = R/N_A$, and hence N_A, from a photograph. □

119.3.2 Displacement statistics

[Reserved: the recording of several thousand successive horizontal displacements over fixed intervals; the histogram of displacements tested against the predicted Gaussian; the mean square displacement per unit time compared with $\langle x^2\rangle = 2Dt$ and hence with Section 119.1.3 [318, 358]; the explicit test that the increments are independent, by checking $\langle x^2\rangle$ against interval length.]

Phenomenon 119.3 (The displacement law)**.** Measured along any fixed horizontal axis, the displacement of a suspended grain over an interval t has zero mean and a variance proportional to the interval itself, not to its square,

$$\left\langle x^2(t)\right\rangle = 2Dt\,, \qquad D = \frac{k_BT}{6\pi\eta a} = \frac{RT}{6\pi\eta a N_A}\,, \tag{119.3}$$

for a sphere of radius a in a fluid of viscosity η, and the histogram of displacements is Gaussian. The root mean square displacement therefore grows as $\sqrt{t}$, so the ratio of an observed displacement to the interval over which it was recorded diverges as that interval is shortened: a Brownian "velocity" read off a finite interval is a property of the stopwatch and not of the particle. Equation (119.3) was predicted by Einstein [318], obtained independently by Sutherland [324] and by Smoluchowski [331], and confirmed on emulsions of gamboge and mastic by Perrin [358].

Derivation. Follow Langevin [344]. Resolve the force on the grain into the systematic Stokes drag it would feel if the fluid were a continuum and a residual random force $F(t)$ of zero mean, uncorrelated with the position of the grain,

$$m\ddot{x} = -6\pi\eta a\,\dot{x} + F(t)\,, \qquad \langle F(t)\rangle = 0\,, \qquad \langle xF\rangle = 0\,. \tag{119.4}$$

Multiply by x and use $x\ddot{x} = \mathrm{d}(x\dot{x})/\mathrm{d}t - \dot{x}^2$:

$$m\frac{\mathrm{d}}{\mathrm{d}t}\langle x\dot{x}\rangle - m\left\langle \dot{x}^2\right\rangle = -6\pi\eta a\,\langle x\dot{x}\rangle\ . \tag{119.5}$$

The grain is one mechanical degree of freedom in thermal equilibrium with the liquid, so equipartition fixes $m\langle\dot{x}^2\rangle = k_BT$. Writing $z = \langle x\dot{x}\rangle = \frac{1}{2}\mathrm{d}\langle x^2\rangle/\mathrm{d}t$ turns Equation (119.5) into the linear equation

$$m\frac{\mathrm{d}z}{\mathrm{d}t} + 6\pi\eta a\,z = k_BT\,, \tag{119.6}$$

whose solution relaxes to the constant $z = k_BT/(6\pi\eta a)$ over the momentum relaxation time $\tau_p = m/(6\pi\eta a)$, which for Perrin's sub-micrometre resin grains in water is a small fraction of a microsecond and so utterly unresolvable at his 30 s sampling interval. For $t \gg \tau_p$, then, $\mathrm{d}\langle x^2\rangle/\mathrm{d}t = 2k_BT/(6\pi\eta a)$, and integrating from $x(0) = 0$ gives Equation (119.3) with $D = k_BT/(6\pi\eta a)$. Substituting $k_B = R/N_A$ puts Avogadro's constant into a quantity measurable with a microscope and a clock; every other symbol in Equation (119.3) is independently known. The Gaussian shape follows because the displacement over an interval long compared with τ_p is a sum of many independent increments, which is the central limit theorem of Chapter 11. □

119.3.3 Rotational Brownian motion

[Reserved: larger grains of about $13\,\mu\mathrm{m}$ *carrying visible internal inclusions used as angular markers; the measurement of the mean square angular displacement against Einstein's rotational prediction [326]; the very different power of the radius in the two laws, which makes their agreement a strong test rather than a repetition [358].]*

119.4 Observations

119.4.1 Measured displacements and distributions

[Reserved: tables of the observed mean square displacements with their standard errors, the fitted diffusion coefficients in m^2/s*, and the observed Gaussian form of the displacement histogram; the sedimentation scale heights; the rotational data; all quantities in SI with uncertainties, as required by the treatise convention [358].]*

119.4.2 Values of Avogadro's constant

[Reserved: the four independent routes giving values clustered near $6.8 \times 10^{23}/\mathrm{mol}$*, within roughly* $15\,\%$ *of the modern figure; Perrin's tabulation of the same constant obtained from thirteen unrelated phenomena — viscosity of gases, blackbody radiation, radioactivity, opalescence, blue of the sky, electric charge — agreeing to within their errors [409]; why the concordance of independent routes carries the evidential weight, an argument revisited in Chapter 1.]*

Phenomenon 119.4 (Concordance of Avogadro's constant)**.** Four measurements made on the same emulsions but resting on different physics — sedimentation equilibrium (Equation (119.1)), translational displacements (Equation (119.3)), rotational displacements, and diffusion of the grains themselves — return one and the same value of N_A to within their stated errors, and that value agrees to within roughly $15\,\%$ with values obtained from a dozen entirely unrelated phenomena, among them the viscosity of gases, blackbody radiation, the counting of radioactive decays, critical opalescence and the blue of the sky [358, 409]. The constant is today fixed by definition at $N_A = 6.022\,140\,76 \times 10^{23}/\mathrm{mol}$ [2333]. It is the concordance, and not the accuracy of any single route, that closed the question of whether molecules are real.

Derivation pending #216

The rotational displacement law for a sphere, in which the mean square angular displacement grows linearly in time with a coefficient carrying the inverse cube of the grain radius rather than its first power — the different power being what makes agreement between the two routes an independent check and not a repetition

119.4.3 Uncertainty budget

[Reserved: the dominant contributions — grain radius, polydispersity, viscosity, temperature drift, finite observation interval, and the two-dimensional projection of a three-dimensional walk; the systematic from tracking a grain that leaves the focal layer; comparison with the modern uncertainty conventions of Chapter 2 and [1945].]

119.5 Interpretation

119.5.1 The fluctuation–dissipation content

[Reserved: the relation $D = kT/(6\pi\eta a)$ read as a link between a fluctuation (D) and a dissipation (η) at one temperature [318, 324]; the same structure in the thermal noise of Chapter 118 and in the general theorem of Chapter 124; the physical content — the same molecular collisions that damp a motion also drive it, so the two coefficients cannot be independent.]

119.5.2 Why this settled the atomic hypothesis

[Reserved: the difference between explaining a phenomenon by atoms and counting the atoms; Perrin's argument that a definite N_A obtained concordantly from unrelated phenomena leaves no economical alternative to their reality [409]; Ostwald's and the energeticists' public change of position; the modern reading of this as the standard example of the argument from consilience.]

119.5.3 The second law as a statistical statement

[Reserved: the visible violation of the second law in a suspended grain, which rises against gravity for part of every second; Smoluchowski's analysis of the fluctuation-driven limits on any perpetual-motion device built from them [331], and its relation to the Landauer analysis in Chapter 118; the entropy statement of Chapter 54 restated as a statement about overwhelming probability, not impossibility.]

119.5.4 Recognition

[Reserved: the 1926 Nobel Prize in Physics awarded to Perrin for his work on the discontinuous structure of matter, and especially for his discovery of sedimentation equilibrium; Perrin's Nobel lecture as his own summary of the evidence [554]; the delayed and partial recognition of Smoluchowski, who died in 1917.]

119.6 Mathematical aftermath

119.6.1 The Wiener process

[Reserved: Wiener's construction of a measure on continuous paths realising Einstein's law, and the proof that almost every path is continuous and nowhere differentiable [507]; why nowhere-differentiability is the mathematical form of Einstein's warning about velocity in Section 119.1.3; the process as the basic object of Chapter 11.]

119.6.2 Ornstein–Uhlenbeck and the velocity correlation

[Reserved: the Ornstein–Uhlenbeck solution of the Langevin equation, giving the exponentially decaying velocity autocorrelation $\langle v(0)v(t)\rangle = (kT/m)\mathrm{e}^{-t/\tau_p}$ and recovering $\langle x^2\rangle = 2Dt$ only for $t \gg \tau_p$ [672]; the ballistic regime $\langle x^2\rangle = (kT/m)t^2$ at short times, and the equipartition value of the velocity variance — the prediction tested in Section 119.7.3.]

119.6.3 The Fokker–Planck description

[Reserved: the evolution equation for the probability density in position and velocity, its equivalence to the Langevin description, and the Maxwell–Boltzmann stationary solution;

Chandrasekhar's review as the standard synthesis of the whole subject [869]; forward reference to Chapter 124 for transport coefficients and to Chapter 66 for the same equation in a plasma.]

119.7 Modern repetitions and precision

119.7.1 Kappler's torsion balance

[Reserved: a small mirror on a quartz fibre in a low-pressure gas, whose angular fluctuations obey $\langle\theta^2\rangle = kT/\kappa$ by equipartition alone, independent of pressure over four decades; Kappler's determination of N_A to about 1 % from these fluctuations [682]; the experiment as the cleanest instance of the fluctuation formulae of Chapter 118 for a single degree of freedom.]

119.7.2 Optical tweezers

[Reserved: the single-beam gradient-force trap [1767] as the modern apparatus — a micrometre bead held in a harmonic potential of known stiffness, with position read out interferometrically at nanometre and microsecond resolution; the power spectrum of the trapped bead as a direct measurement of both D and τ_p; the same instrument used for the single-molecule work cited in Chapter 118.]

119.7.3 The instantaneous velocity

[Reserved: Li, Kheifets, Medellin and Raizen's measurement of the instantaneous velocity of a silica bead in air, resolving the ballistic regime and recovering the equipartition value $\langle v^2\rangle = kT/m$ [2160]; the extension to liquids, where hydrodynamic memory modifies the correlation function and the simple Ornstein–Uhlenbeck form fails [2234]; the explicit overturning of Einstein's judgement that this velocity could never be measured, and what made it possible — trap stiffness, detection bandwidth and low pressure.]

Phenomenon 119.5 (The ballistic regime and equipartition)**.** Resolved on time scales short compared with the momentum relaxation time τ_p of Equation (119.6), the motion of a Brownian particle is not diffusive but ballistic: the mean square displacement grows as t^2 rather than as t, the trajectory acquires a well-defined instantaneous velocity, and the variance of that velocity takes the equipartition value

$$\left\langle v^2 \right\rangle = \frac{k_B T}{m} \,. \tag{119.7}$$

This was measured for a silica bead of micrometre scale held in an optical trap in air, with the position read out at nanometre and sub-microsecond resolution [2160]; in a liquid the same measurement finds the simple exponential velocity correlation replaced by a slower, hydrodynamic-memory form, so that Equation (119.7) is approached but the Ornstein–Uhlenbeck description of Section 119.6.2 fails in detail [2234]. The measurement overturns Einstein's explicit judgement [318] that this velocity could never be observed.

Derivation pending #217

The velocity autocorrelation of the Ornstein–Uhlenbeck process obtained from the Langevin equation, its short-time ballistic limit, and the crossover to the diffusive law at the momentum relaxation time; and separately the hydrodynamic-memory kernel that replaces it for a sphere in a dense fluid

119.7.4 Avogadro's constant today

[Reserved: the silicon-sphere X-ray crystal density route and the watt-balance route that superseded Brownian determinations; the 2019 redefinition fixing N_A exactly at $6.022\,140\,76 \times 10^{23}/\mathrm{mol}$ by convention, so that the quantity Perrin measured is now a defined constant [2333]; the current values of the related constants [2392]; what a modern Brownian measurement now tests instead — the Boltzmann constant, the viscosity, or the apparatus itself; the SI framework of Chapter 2.]

Chapter 120

Quantum Statistics

Classical statistical mechanics counts states as though the particles carried labels, and the answer is wrong: the entropy is not extensive until a factor $1/N!$ is inserted by hand (Chapter 118), the heat capacities of gases and solids collapse at low temperature, and a metal's conduction electrons contribute almost nothing to its specific heat while contributing everything to its conductivity. This chapter replaces the counting. Identical particles are indistinguishable in principle, the states of N of them are either totally symmetric or totally antisymmetric under exchange, and the two cases give the Bose–Einstein and Fermi–Dirac distributions — from which the Planck spectrum, the degeneracy pressure that holds up a white dwarf, the linear electronic heat capacity of a metal, Pauli paramagnetism, Landau diamagnetism and Bose–Einstein condensation all follow. Each is stated below with the measurement that established it.

The chapter takes the exchange symmetry itself from Chapter 83 and the spin–statistics theorem from Chapter 109, and hands its results on to Chapter 121 (the atomic condensate as a laboratory system), Chapters 125 and 126 (solids), Chapter 130 (paired and condensed phases) and Chapter 49 (degenerate stellar matter). Founding sources are Bose [509], Einstein [512, 522], Fermi [539] and Dirac [537]; the standard modern treatments are [1656, 2180].

Derivation pending #218

Quantum Statistics: all derivations of this chapter are pending.

120.1 Indistinguishability

120.1.1 Counting states for identical particles

[Reserved: the difference between Boltzmann counting, in which exchanging two particles gives a new microstate, and the counting used by Bose, in which it does not [509]; occupation-number representation and the specification of a state by $\{n_i\}$ alone; the resulting combinatorics for unrestricted and restricted occupancy; the symmetrisation postulate and the exchange operator, developed in Chapter 83.]

120.1.2 The exclusion principle

[Reserved: Pauli's formulation — no two electrons in an atom share all four quantum numbers — introduced to explain the closing of electron shells and the anomalous Zeeman multiplet structure [529]; the periodic table and atomic spectra as the evidence base (Chapters 70 and 88); the restatement as antisymmetry of the many-electron wavefunction and the Slater determinant.]

Phenomenon 120.1 (Exclusion)**.** The electrons of a many-electron atom do not accumulate in the lowest orbital. The chemical periods close at atomic numbers $Z = 2, 10, 18, \ldots$, the first ionization energy peaks at exactly those values and drops abruptly at the next, and the alkali spectra are those of a single valence electron outside an inert core: no two electrons of an atom share all four quantum numbers [529]. Nor is this a rule that softens at high precision. Passing a fresh current through a copper strip, so that the conduction electrons arriving at a given atom have never been there before, and searching the X-ray spectrum for a transition into an already-filled K shell, bounds the probability of a Pauli-violating state below about 1.7×10^{-26} [1837].

Derivation pending #219

Exclusion derived rather than postulated: the symmetrisation of the state of identical particles, the vanishing of a totally antisymmetric state with two coincident single-particle labels, and the spin–statistics theorem that assigns antisymmetry to half-integer spin

120.1.3 The spin–statistics theorem

[Reserved: Pauli's proof that, in a relativistic quantum field theory with a positive-definite energy and local commutativity, half-integer spin fields must anticommute and integer spin fields must commute [854]; the failure of the alternative assignment — negative energies or negative norms — rather than mere inelegance; the axiomatic form of the theorem in Chapter 109 and its relation to the CPT theorem of Chapter 105.]

120.1.4 Composite particles

[Reserved: Ehrenfest and Oppenheimer's rule that a bound system of an even number of fermions behaves as a boson and an odd number as a fermion, with the validity conditions — the exchange must not resolve the internal structure [679]; the worked cases ^{4}He *(boson) against* ^{3}He *(fermion), whose radically different low-temperature behaviour is the direct experimental demonstration; Cooper pairs and excitons as the same mechanism (Chapter 130).]*

120.1.5 Experimental limits on violations

[Reserved: the Ramberg–Snow experiment, in which a fresh current is passed through a copper strip and the X-ray spectrum searched for transitions to an already-filled K shell, bounding the probability of a Pauli-violating state below about 1.7×10^{-26} *[1837]; later and stronger bounds from nuclear and electron stability; why an exact conservation law deserves a direct test rather than an assumption.]*

120.2 The two distributions

120.2.1 Bose's derivation of the Planck law

[Reserved: Bose's counting of photon states in phase-space cells of volume h^3*, which derives the Planck spectrum without invoking classical electrodynamics at any point [509]; Einstein's recognition and translation of the paper; the relation to the blackbody measurements of Chapter 68 and to the cosmic microwave spectrum of Chapter 51, the most precise blackbody ever measured; the photon chemical potential vanishing because photon number is not conserved.]*

120.2.2 The Bose–Einstein distribution

[Reserved: Einstein's extension of Bose's counting to a gas of material particles, giving $\langle n_i \rangle = [\mathrm{e}^{\beta(\epsilon_i - \mu)} - 1]^{-1}$ [512]; the constraint $\mu \leq \epsilon_0$; the enhancement of multiple occupancy relative to classical counting, which is the statistical statement behind stimulated emission (Chapter 122).]

120.2.3 Fermi–Dirac statistics

[Reserved: Fermi's quantisation of the ideal monatomic gas with the exclusion principle imposed [539] and Dirac's independent derivation from the antisymmetry of the wavefunction, with his identification of which particles obey which statistics [537]; the distribution $\langle n_i \rangle = [\mathrm{e}^{\beta(\epsilon_i - \mu)} + 1]^{-1}$; the step function at $T = 0$ and the thermal smearing of width $\sim kT$ about the chemical potential.]

120.2.4 The classical limit

[Reserved: both distributions reducing to $\mathrm{e}^{-\beta(\epsilon - \mu)}$ when the occupancies are small, i.e. when the thermal de Broglie wavelength is much less than the interparticle spacing; the degeneracy parameter $n\lambda^3$ and its value for air at room temperature against a metal's conduction electrons; the recovery of the Sackur–Tetrode entropy of Chapter 118 with the $1/N!$ now derived rather than inserted [380, 393].]

120.2.5 Grand canonical treatment

[Reserved: why the grand canonical ensemble is the natural one — the occupation numbers are then independent and $\log \Xi$ factorises over modes; the fugacity, the Bose and Fermi functions $g_\nu(z)$ and $f_\nu(z)$, and the resulting equations of state; the quantum virial expansion and the statistical "exchange pressure" that is repulsive for fermions and attractive for bosons in the absence of any interaction [1656].]

120.3 The ideal Bose gas

120.3.1 Einstein's condensation

[Reserved: the saturation of the excited-state population, so that below a critical temperature a macroscopic fraction of the particles must occupy the single-particle ground state [522]; Einstein's own statement that the phenomenon occurs without any interaction, purely from the statistics; the initial scepticism and the delay before it was taken to be physical.]

120.3.2 Critical temperature and condensate fraction

[Reserved: $kT_c = 2\pi\hbar^2 n^{2/3}/[m\,\zeta(3/2)^{2/3}]$ in three dimensions, the condensate fraction $1 - (T/T_c)^{3/2}$, and the $T^{3/2}$ heat capacity with its cusp at T_c [522, 1656]; the dimensional statement — no condensation of the ideal gas in a homogeneous system of fewer than three space dimensions — and the modification by a trapping potential; numerical values for helium and for a dilute alkali vapour.]

120.3.3 Superfluid helium-4

[Reserved: the abrupt loss of viscosity below $2.17\,\mathrm{K}$ measured independently by Kapitza [817] and by Allen and Misener [809]; London's identification of the lambda transition with Bose–Einstein condensation, noting that the ideal-gas T_c at the density of liquid helium is

3.1 K, *close to the observed* 2.17 K *[818]; the strong interactions that make the identification qualitative, the condensate fraction of only about* 10 %*, and the full account in Chapter 130.]*

120.3.4 The weakly interacting Bose gas

[Reserved: Bogoliubov's diagonalisation of the weakly interacting Bose Hamiltonian, giving a phonon-like linear excitation spectrum at long wavelength and a depletion of the condensate [898]; Landau's criterion for superfluidity, that a critical velocity exists whenever the excitation spectrum has no low-lying free-particle branch [858]; why the ideal Bose gas, despite condensing, is not a superfluid.]

120.3.5 Dilute atomic condensates

[Reserved: the observation of condensation in rubidium-87 at 170 nK *by Anderson and coworkers [1893] and in sodium at higher atom number by Davis and coworkers [1898]; the bimodal momentum distribution as the signature; laser and evaporative cooling, apparatus and uncertainties treated in Chapter 121; the condensate as a nearly ideal realisation of Section 120.3.1, seventy years after the prediction.]*

Phenomenon 120.2 (Condensation without interaction)**.** A gas of identical bosons cooled at fixed density undergoes a phase transition at a temperature set by the density and the mass alone, below which a macroscopic fraction of the atoms occupies one and the same single-particle state. The transition is observed in a dilute vapour of rubidium-87 at 170 nK and in sodium at higher atom number [1893, 1898], where the interactions are so weak that the ideal-gas prediction [522] applies with per-cent corrections; and it is observed in liquid helium-4, which loses its viscosity abruptly below 2.17 K [809, 817] at a temperature close to the ideal-gas value of 3.1 K computed at the density of the liquid [818]. No attraction between the particles is involved: what condenses them is the counting rule.

Derivation pending #220

Saturation of the excited-state population of the ideal Bose gas: that the sum over excited modes of the Bose occupation is bounded above at fixed temperature and volume when the chemical potential reaches the ground-state energy, so that any further particles must enter the ground state; the resulting critical temperature, condensate fraction and heat-capacity cusp in three dimensions

120.4 The ideal Fermi gas

120.4.1 Fermi energy and degeneracy pressure

[Reserved: the filled Fermi sphere at $T = 0$*, the Fermi wavevector* $k_F = (3\pi^2 n)^{1/3}$*, the Fermi energy and Fermi temperature, and the ground-state pressure* $P = \frac{2}{5} n \epsilon_F$ *which survives at zero temperature [539]; numerical values for the conduction electrons of copper (*$\epsilon_F \approx 7$ eV*,* $T_F \approx 8 \times 10^4$ K*), showing that a metal at room temperature is a strongly degenerate gas; the Fermi surface as an observable object.]*

120.4.2 The Sommerfeld expansion

[Reserved: Sommerfeld's application of Fermi statistics to the electron theory of metals [620], resolving the long-standing failure of the Drude model to explain why the electrons contribute a full Wiedemann–Franz conductivity but almost no heat capacity; the expansion

in kT/ϵ_F and the linear electronic heat capacity $C = \gamma T$; the measured low-temperature form $C = \gamma T + \beta T^3$ of a metal, whose two terms separate the electrons from the phonons of Chapter 125; band-structure corrections in Chapter 126.]

Phenomenon 120.3 (The linear electronic heat capacity of a metal)**.** The conduction electrons of a metal carry the whole of its electrical and thermal current, as a free gas of one carrier per atom would, and yet contribute to its heat capacity about two orders of magnitude less than the $\frac{3}{2}k_B$ per electron that classical equipartition demands. Between about 1 K and 5 K the heat capacity of a normal metal is measured to be

$$C = \gamma T + \beta T^3 \,, \tag{120.1}$$

so that C/T plotted against T^2 is a straight line whose intercept γ is electronic and whose slope β is the lattice contribution of Chapter 125; the separation was carried out for copper, silver and gold [1061]. A term linear in T, and a coefficient this small, are what Fermi statistics predict and what Maxwell–Boltzmann statistics cannot produce [620].

Derivation. Let $g(\epsilon)$ be the density of one-electron states per unit volume, counting both spins. At temperature T the occupation $f(\epsilon) = [\mathrm{e}^{\beta(\epsilon-\mu)} + 1]^{-1}$ differs from the zero-temperature step only within a shell of width of order k_BT about μ. Hence only about $g(\epsilon_F)k_BT$ electrons per unit volume are thermally active, and each carries an excess energy of order k_BT, so the thermal energy density rises above its ground-state value by an amount of order $g(\epsilon_F)(k_BT)^2$, and the heat capacity is of order $k_B^2 T g(\epsilon_F)$ — linear in T, not constant. The coefficient follows from the Sommerfeld expansion of $\int H(\epsilon) f(\epsilon) \mathrm{d}\epsilon$ in powers of k_BT/μ, whose leading correction is $(\pi^2/6)(k_BT)^2 H'(\mu)$; applied to $H = \epsilon g(\epsilon)$ and differentiated with respect to T it gives

$$c_V = \frac{\pi^2}{3} k_B^2 \, T \, g(\epsilon_F) \,, \tag{120.2}$$

which is Equation (120.1) with $\gamma = \frac{1}{3}\pi^2 k_B^2 g(\epsilon_F)$. For the free electron gas $g(\epsilon_F) = 3n/(2\epsilon_F)$ and Equation (120.2) becomes $c_V = \frac{1}{2}\pi^2 n k_B (T/T_F)$ with $k_BT_F = \epsilon_F$, smaller than the classical $\frac{3}{2}nk_B$ by the factor $(\pi^2/3)(T/T_F)$. With $T_F \approx 8 \times 10^4$ K for copper this factor is about 10^{-2} at room temperature, which is the observed suppression. The measured γ is thus a direct measurement of the density of states at the Fermi level. □

120.4.3 Pauli paramagnetism

[Reserved: Pauli's calculation of the spin susceptibility of a degenerate electron gas, temperature-independent and smaller than the Curie law by a factor $\sim T/T_F$ [586]; the resolution of the observed weak, temperature-independent paramagnetism of the alkali metals; the susceptibility as a measurement of the density of states at the Fermi level; the wider magnetism of Chapter 129.]

Phenomenon 120.4 (Pauli paramagnetism)**.** The alkali metals are weakly paramagnetic with a susceptibility that is essentially independent of temperature, in flat contradiction with the Curie law $\chi \propto 1/T$ obeyed by the same number of localized moments, and smaller than the Curie value at room temperature by about two orders of magnitude [586].

Derivation. In a field B the two spin populations acquire energies shifted by $\mp\mu_B B$. They share one chemical potential, so the shift is absorbed by transferring electrons from the minority to the majority band until the two Fermi levels coincide; to first order in the field the transferred density is $\frac{1}{2}g(\epsilon_F)\mu_B B$ from each band, giving a magnetization

$$M = \mu_B^2 g(\epsilon_F) B \,, \qquad \chi = \frac{\mu_0 M}{B} = \mu_0 \mu_B^2 g(\epsilon_F) \,, \tag{120.3}$$

in which the temperature appears nowhere: only the electrons within k_BT of the Fermi surface can flip, and their number is proportional to T while the polarization each contributes is proportional to $1/T$, so the two cancel. Dividing Equation (120.3) by the Curie susceptibility $\chi_C = n\mu_0\mu_B^2/k_BT$ of n free spins one half, and using $g(\epsilon_F) = 3n/(2\epsilon_F)$, gives $\chi/\chi_C = \frac{3}{2}(T/T_F)$, of order 10^{-2} at room temperature — the observed suppression, and the same degeneracy factor that suppresses the heat capacity in Phenomenon 120.3. The measured susceptibility is therefore a second, independent determination of $g(\epsilon_F)$. □

120.4.4 Landau diamagnetism and quantum oscillations

[Reserved: Landau's demonstration that orbital quantisation in a magnetic field gives a diamagnetic contribution of exactly minus one third of the Pauli susceptibility for free electrons [660]; the Landau levels and their degeneracy; the de Haas–van Alphen oscillations of the susceptibility periodic in $1/B$, discovered in bismuth [655], which measure extremal Fermi-surface cross sections directly; the same levels underlying the quantum Hall measurements of Chapter 128.]

120.4.5 Degenerate Fermi gases in the laboratory

[Reserved: DeMarco and Jin's cooling of potassium-40 to $T/T_F \approx 0.5$, with the onset of degeneracy seen as a deviation of the released energy from the classical value [1971]; the condensation of fermionic atom pairs across a Feshbach resonance and the crossover between a paired condensate and a Bardeen–Cooper–Schrieffer superfluid [2065]; the connection to Chapter 130.]

120.5 Degenerate matter in astrophysics

120.5.1 White dwarfs

[Reserved: Fowler's identification of the electron degeneracy pressure of Section 120.4.1 as what supports a white dwarf against gravity, the first astrophysical application of the new statistics [541]; Chandrasekhar's relativistic treatment and the limiting mass near 1.4 solar masses, above which no degenerate configuration exists [676]; the observed white-dwarf mass distribution and the type Ia supernova connection, treated in Chapters 49 and 52.]

Phenomenon 120.5 (Degeneracy pressure holds up a dead star)**.** A white dwarf is a star of roughly solar mass compressed to roughly the radius of the Earth. It has no nuclear source, radiates freely, and is cooling — and yet it does not contract. The pressure that supports it does not vanish with the temperature, and is supplied by the exclusion principle acting on the electrons [541]. The support is not unlimited: since the electrons become relativistic as they are compressed, a degenerate configuration exists only below a limiting mass near 1.4 solar masses [676], and the observed white-dwarf masses respect it. The same argument applied to neutrons is bounded by the measured pulsar mass of 1.97 ± 0.04 solar masses, obtained from the Shapiro delay in a binary and sufficient by itself to exclude the softest nuclear equations of state [2155].

Derivation pending #221

The ground-state pressure of the ideal Fermi gas, its persistence at zero temperature, the softening of the equation of state when the Fermi momentum becomes relativistic, and the resulting limiting mass for a self-gravitating degenerate sphere

120.5.2 Neutron stars

[Reserved: neutron degeneracy pressure and the nuclear-matter corrections that dominate it, so that the maximum mass is set by the equation of state rather than by the ideal-gas result; the measurement of a pulsar of 1.97 ± 0.04 *solar masses through the Shapiro delay, which excludes the softest equations of state [2155]; the nuclear physics in Chapter 107 and the relativistic structure equations in Chapter 49.]*

120.6 Correlations as direct evidence

120.6.1 Photon bunching

[Reserved: the Hanbury Brown and Twiss measurement of a positive correlation between the photocurrents of two detectors illuminated by the same thermal source [1083]; the initial disbelief and the resolution — the effect is a consequence of Bose statistics, not of any interaction; intensity interferometry as a technique for stellar angular diameters; the quantum-optical treatment of $g^{(2)}(0)$ *in Chapter 122.]*

120.6.2 Bunching and antibunching compared

[Reserved: the direct side-by-side experiment of Jeltes and coworkers, measuring the two-particle correlation of clouds of metastable ^{4}He *(bunching) and* ^{3}He *(antibunching) released from the same apparatus [2109]; the identical preparation of the two isotopes, which makes the opposite sign of the correlation an isolated demonstration of the statistics; the correlation length as the source coherence length.]*

Phenomenon 120.6 (Bunching and antibunching)**.** Identical particles arriving at two detectors from a chaotic source are correlated even though nothing couples them. Hanbury Brown and Twiss found the photocurrents of two detectors illuminated by one thermal light source to be positively correlated at small detector separation, an effect with no classical-particle explanation and none in any interaction [1083]. The sign of the correlation is fixed by the statistics alone, as the decisive control shows: clouds of metastable helium-4 and of helium-3, prepared and released from the same apparatus under the same conditions, give a two-particle correlation that is a peak for the boson and a dip for the fermion [2109].

Derivation pending #222

The two-particle detection amplitude for identical particles from a chaotic source, the interference of the direct and exchanged terms, and the resulting second-order correlation of two at zero separation for bosons against zero for fermions, with the source coherence length as the correlation scale

120.6.3 Exchange effects in scattering

[Reserved: the interference of direct and exchange amplitudes in the scattering of identical particles, and Mott's cross section for two electrons with the resulting oscillatory term [663]; the factor of two enhancement at 90° *for identical bosons and the corresponding suppression for identical fermions; the measured alpha–alpha scattering pattern as the classic demonstration; scattering formalism in Chapter 86.]*

120.7 Beyond the two statistics

120.7.1 Why three space dimensions admit only two

[Reserved: the topology of the configuration space of identical particles in three space dimensions, whose exchange paths form the permutation group, so that only the two one-dimensional representations — symmetric and antisymmetric — are available [1603]; the spin–statistics theorem of Section 120.1.3 fixing which particles take which; the absence of any observed violation, per Section 120.1.5.]

120.7.2 Quasiparticle statistics in planar systems

[Reserved: electrons confined to a plane within a three-dimensional sample, for which the exchange paths form the braid group and the emergent quasiparticles may carry fractional exchange phases [1603]; the strict statement of scope — these are collective excitations of a physical three-dimensional material, not fundamental particles, and the treatise makes no claim about physics in fewer than three space dimensions; the fractional quantum Hall measurements that bear on it are reported in Chapter 128.]

Chapter 121

Experiment: Bose–Einstein Condensation

Bose–Einstein condensation is the one phase transition that requires no interaction whatsoever: it follows from the counting rule for identical bosons alone, which is why Chapter 120 can derive it for the ideal gas and why Chapter 83 is its real premise. This chapter is the experimental counterpart. It presents the 1995 JILA and MIT condensates in full phenomenological detail [1893, 1898] — apparatus, cooling sequence, measured numbers with their uncertainties — together with the seventy years of technique that separate Einstein's prediction [522] from its realization in a dilute alkali vapour, and with the modern repetitions that have turned a once-heroic measurement into a routine laboratory tool. The delay was not conceptual: a condensate of the kind Einstein described needs a gas held at a temperature six orders of magnitude below anything a cryostat can reach, in a container with no walls, at densities low enough that three-body recombination does not convert it to a solid first. Every element of the apparatus in Section 121.2 exists to meet one of those three constraints.

The chapter sits between the statistics that predict the effect (Chapter 120) and the interacting condensed phases of Chapter 130, where the same macroscopic wavefunction appears in a strongly correlated liquid; its coherence measurements are the atomic counterpart of the optical ones in Chapter 122, and its matter-wave interference closes the argument opened in Chapter 71. When written out, the chapter will use the structured `experiment` environment throughout, with apparatus, procedure, observations, interpretation and primary references as separate fields, and with all quantities in SI.

Derivation pending #223

Experiment: Bose–Einstein Condensation: all derivations of this chapter are pending.

121.1 The prediction under test

121.1.1 Bose statistics and Einstein's condensation

[Reserved: Bose's re-derivation of Planck's law by counting cells in phase space rather than by assuming a classical field [509]; Einstein's extension of the counting to a gas of massive particles [512] and, in the second paper, the conclusion that below a finite temperature a macroscopic fraction of the particles occupies the single-particle ground state [522]; why the effect is purely statistical, arising from the symmetry of the many-body state (Chapter 83) and not from any attraction; the ideal-gas thermodynamics derived in Chapter 120.]

121.1.2 The condensation criterion

[Reserved: the thermal de Broglie wavelength $\lambda_{\mathrm{dB}} = h/\sqrt{2\pi m k_B T}$ and the criterion $n\lambda_{\mathrm{dB}}^3 \geq \zeta(3/2) \approx 2.612$ for the homogeneous ideal gas [522]; the modified criterion $k_B T_c = \hbar\bar{\omega}(N/\zeta(3))^{1/3}$ for the harmonic trap that every experiment actually uses [1970]; the numerical consequence that a vapour at $n \sim 1 \times 10^{19}/\mathrm{m}^3$ condenses only below about 1 µK, six decades colder than liquid helium; the standard monograph treatment [2133].]

121.1.3 Seventy years of near misses

[Reserved: London's identification of the helium-4 lambda transition as a Bose–Einstein condensation distorted by strong interactions [818], and why a liquid with a condensate fraction near 10 % cannot test Einstein's ideal-gas prediction (the interacting case is Chapter 130); the spin-polarized atomic hydrogen programme as the deliberate attempt at a weakly interacting Bose gas [1776, 1811], its recombination-limited densities, and the switch to alkali vapours that could be laser-cooled.]

121.2 Apparatus

121.2.1 Laser cooling

[Reserved: radiation pressure as a mechanical force on neutral particles [1412]; the proposal to cool a gas with red-detuned light [1547]; optical molasses — six counter-propagating beams producing a viscous damping force and temperatures near the Doppler limit $k_B T_D = \hbar\Gamma/2$ [1750]; the magneto-optical trap that adds a quadrupole field to make the force position-dependent [1796]; Zeeman slowing of the atomic beam that loads it; the level structures involved belong to Chapter 88.]

121.2.2 Sub-Doppler cooling

[Reserved: the measurement that found sodium at (43 ± 20) µK, well below the 240 µK Doppler limit, and so falsified the accepted theory of molasses [1810]; the polarization-gradient (Sisyphus) mechanism that explains it, in which an atom climbs a light shift and is optically pumped to the bottom of the next well [1818]; the recoil limit $k_B T_R = \hbar^2 k^2/m$ as the residual floor; why sub-Doppler temperatures still leave phase-space density a factor $\sim 10^6$ short of condensation.]

Phenomenon 121.1 (Cooling below the accepted limit)**.** Laser-cooled sodium in optical molasses is measured at (43 ± 20) µK, well below the 240 µK Doppler limit $k_B T_D = \hbar\Gamma/2$ that the accepted theory of molasses predicted and that had been treated as a floor [1810]. The measurement was a falsification, not a refinement: it forced the recognition of a second cooling mechanism that the two-level model does not contain, and the temperature it reaches instead is set by the photon recoil.

Derivation pending #224

The polarization-gradient (Sisyphus) mechanism: the spatially modulated light shifts of the ground-state sublattice in a standing wave of varying polarization, the optical pumping that preferentially transfers an atom from a hill to the bottom of the next well, the resulting friction coefficient, and the recoil energy as the residual floor

121.2.3 Magnetic traps

[Reserved: the Zeeman potential $U = \mu_B g_F m_F |B|$ and the first magnetic confinement of neutral atoms [1761]; Majorana spin flips at the zero of a pure quadrupole field, and the two cures — the JILA time-orbiting potential trap that rotates the zero faster than the atoms can follow [1908] and the MIT optical plug that repels atoms from it [1898]; why only weak-field-seeking states can be trapped, and why Earnshaw's theorem forbids a static maximum.]

121.2.4 Evaporative cooling

[Reserved: forced evaporation — selectively removing the highest-energy atoms with a radio-frequency knife and letting elastic collisions rethermalize the remainder — proposed and demonstrated for spin-polarized hydrogen [1776, 1811]; the runaway condition relating the ratio of good (elastic) to bad (inelastic) collisions; the price paid in atom number, typically a factor 10^3 lost for a factor 10^6 gained in phase-space density; why this last stage, and not laser cooling, is what actually reaches degeneracy.]

121.2.5 Imaging and thermometry

[Reserved: destructive absorption imaging after ballistic expansion, in which the released cloud maps momentum onto position and the optical depth is fitted to a Bose-distributed column density; non-destructive dispersive (phase-contrast) imaging, which permits repeated observation of one condensate [1914]; how temperature is extracted from the thermal wings rather than from any thermometer in contact with the gas; systematic uncertainties in magnification, atom number and the resonant cross-section, which dominate the error budget quoted in Section 121.4.2.]

121.3 Procedure

121.3.1 The JILA rubidium-87 sequence

[Reserved: the full 1995 cycle — vapour-cell magneto-optical trap, polarization-gradient molasses, optical pumping into $|F = 2, m_F = 2\rangle$, transfer to the time-orbiting potential trap, and radio-frequency evaporation over about $70\,\mathrm{s}$ [1893, 1908]; the endpoint of roughly 2×10^3 condensed atoms of rubidium-87 at $170\,\mathrm{nK}$ and a peak density near $2.5 \times 10^{18}\,\mathrm{m}^{-3}$; the anisotropic expansion after release as the signature that survived every check.]

121.3.2 The MIT sodium-23 sequence

[Reserved: the contrasting design — a Zeeman-slowed sodium beam, a dark-spot magneto-optical trap, and a plugged quadrupole trap — reaching some 5×10^5 condensed atoms, two orders of magnitude more than JILA, at a transition temperature near $2\,\mu\mathrm{K}$ [1898]; why the larger condensate is what made the coherence experiments of Section 121.4.4 possible; the independence of the two apparatus designs as the strongest argument against an artefact.]

121.3.3 Lithium-7 and attractive interactions

[Reserved: the third 1995 result, in lithium-7, whose negative scattering length makes the condensate mechanically unstable above a critical atom number [1896]; the observed limit of order 10^3 atoms and its later refinement; why this case tests the interaction-dependent part of the theory that rubidium and sodium, both repulsive, cannot.]

121.4 Observations

121.4.1 The bimodal velocity distribution

[Reserved: the primary datum — a narrow, anisotropic peak emerging from a broad isotropic thermal background as the evaporation endpoint is lowered through T_c [1893, 1898]; why the anisotropy is decisive: a thermal cloud expands isotropically whatever the trap shape, while a condensate inherits the inverted aspect ratio of the trapping potential through the uncertainty relation; the three control observations that excluded a cold thermal cloud, an imaging artefact and a mechanical compression.]

Phenomenon 121.2 (The condensate appears as a second component)**.** As the endpoint of the evaporation is lowered through a critical temperature, the velocity distribution of the released cloud ceases to be a single broad isotropic Gaussian and becomes bimodal: a narrow peak grows out of the broad thermal background and takes over as the temperature is lowered further. The narrow component is *anisotropic*, and expands fastest along the axis in which the trap was tightest — the inverse of the shape it had in the trap — while the thermal background expands isotropically whatever the trap geometry [1893, 1898].

Derivation. The anisotropy is the discriminating observation, and it follows from the uncertainty relation alone. A thermal cloud in equilibrium at temperature T has, whatever the shape of the confining potential, an isotropic Maxwell–Boltzmann momentum distribution of width $\sqrt{mk_BT}$ in every direction; after release its size grows as $\sqrt{k_BT/m}\,t$ equally in all three directions, so a long expansion always drives it towards a sphere. A condensate, by contrast, occupies one single-particle state, the ground state of a harmonic trap of frequencies ω_i, whose widths and momentum widths are

$$\sigma_i = \sqrt{\frac{\hbar}{m\omega_i}}\,, \qquad \Delta p_i = \frac{\hbar}{2\sigma_i} = \tfrac{1}{2}\sqrt{m\hbar\omega_i}\,. \tag{121.1}$$

The in-trap aspect ratio is therefore $\sigma_1/\sigma_2 = \sqrt{\omega_2/\omega_1}$, while after a free expansion long enough that the initial size is negligible the observed size is $\Delta p_i t/m \propto \sqrt{\hbar\omega_i/m}\,t$ and the aspect ratio is $\sqrt{\omega_1/\omega_2}$ — exactly inverted. The tightly confined direction is the one with the largest momentum spread and so expands fastest. No thermal cloud, at any temperature, can produce that inversion, which is why the anisotropy and not the narrowness of the peak is what identifies the condensate. □

121.4.2 Condensate fraction and transition temperature

[Reserved: the quantitative test — condensate fraction $N_0/N = 1 - (T/T_c)^3$ for an ideal gas in a harmonic trap, measured against T/T_c together with the released energy [1921]; the measured transition temperature and the several-per-cent negative shift from the ideal-gas value attributable to finite N and to mean-field repulsion; the tabulated SI values, uncertainties and systematic error budget that will constitute the dataset of this chapter; comparison with the trapped-gas theory [1970].]

Phenomenon 121.3 (Condensate fraction and transition temperature)**.** The condensed fraction of a harmonically trapped Bose gas rises from zero at a sharp transition temperature according to

$$\frac{N_0}{N} = 1 - \left(\frac{T}{T_c}\right)^3, \qquad k_BT_c = \hbar\bar{\omega}\left(\frac{N}{\zeta(3)}\right)^{1/3}, \tag{121.2}$$

with $\bar{\omega}$ the geometric mean trap frequency. Both the fraction and the released energy were measured against T/T_c and found to follow Equation (121.2) [1921]; the measured T_c lies a few per cent *below* the ideal-gas value, by an amount attributable to the finite atom number

and to the mean-field repulsion between the atoms [1970]. A prediction made for a gas that does not interact thus survives contact with one that does, to within the corrections that interaction requires.

Derivation. For a three-dimensional harmonic trap with frequencies ω_i and $\bar{\omega} = (\omega_1\omega_2\omega_3)^{1/3}$, the number of single-particle states with energy below E is $E^3/6(\hbar\bar{\omega})^3$ for $E \gg \hbar\bar{\omega}$, so the density of states is $g(E) = E^2/2(\hbar\bar{\omega})^3$. At and below the transition the chemical potential has reached the ground-state energy, and the number of atoms the excited states can hold is

$$N_{\mathrm{ex}} = \int_0^\infty \frac{g(E)\,\mathrm{d}E}{\mathrm{e}^{E/k_BT} - 1} = \frac{1}{2}\left(\frac{k_BT}{\hbar\bar{\omega}}\right)^3 \int_0^\infty \frac{x^2\,\mathrm{d}x}{\mathrm{e}^x - 1} = \zeta(3)\left(\frac{k_BT}{\hbar\bar{\omega}}\right)^3, \tag{121.3}$$

using $\int_0^\infty x^2(\mathrm{e}^x - 1)^{-1}\mathrm{d}x = 2\zeta(3)$. The right-hand side of Equation (121.3) is finite and depends only on the temperature: the excited states *saturate*. Any atoms beyond that number have nowhere to go but the ground state, which is the whole content of the effect. Setting $N_{\mathrm{ex}} = N$ defines T_c as in Equation (121.2); below T_c the excited population is still given by Equation (121.3), that is $N_{\mathrm{ex}} = N(T/T_c)^3$, and the remainder $N_0 = N - N_{\mathrm{ex}}$ gives the condensate fraction. Note that the exponent is 3 for the harmonic trap, not the 3/2 of the homogeneous gas; the density of states, not the statistics, is what differs. □

121.4.3 Collective excitations

[Reserved: the low-lying shape oscillations of the trapped condensate, measured to the percent level and compared with the hydrodynamic predictions of the Gross–Pitaevskii equation [1924]; the frequency ratios as a parameter-free test, since they depend on the trap geometry alone; the damping rates and their temperature dependence as the first quantitative probe of the condensate–thermal-cloud coupling.]

121.4.4 Interference of two condensates

[Reserved: two independently prepared sodium condensates, released and overlapped, producing high-contrast straight fringes of period about 15 µm *[1934]; the fringes as the direct evidence of long-range phase coherence, and the subtle point that the relative phase is established by the measurement rather than pre-existing; the matter-wave interpretation continuous with Chapter 71 and Chapter 67; the output coupler that turned the condensate into a coherent atom beam [1940].]*

Phenomenon 121.4 (Long-range phase coherence)**.** Two condensates prepared independently in separate wells of one trap, then released and allowed to overlap, produce high-contrast straight interference fringes of period about 15 µm [1934]. The fringes are matter-wave interference on a macroscopic scale: the phase of the condensate is well defined across the whole cloud, over distances some five orders of magnitude larger than the interparticle spacing. The subtlety the data also establish is that the two clouds never shared a phase — the relative phase is established by the act of detection, and is random from shot to shot while the fringe contrast is not.

Derivation pending #225

Interference of two overlapping expanding condensates: the fringe spacing set by the relative momentum of the two sources, and the emergence of a definite relative phase from repeated measurement on a state of fixed total atom number, in which no relative phase exists before the first detection

121.5 Interpretation

121.5.1 The order parameter and the Gross–Pitaevskii equation

[Reserved: the condensate wavefunction $\psi(\boldsymbol{r},t)=\sqrt{n_0}\,\mathrm{e}^{\mathrm{i}\theta}$ as the order parameter of a broken U(1) *symmetry, and the nonlinear Schrödinger equation governing it, derived independently for the vortex problem [1208, 1217]; the coupling constant $g=4\pi\hbar^2 a/m$ fixed by the measured s-wave scattering length; the Thomas–Fermi profile and the healing length $\xi=(8\pi n a)^{-1/2}$, both confirmed by the expansion data [1970].]*

121.5.2 Bogoliubov excitations and superfluidity

[Reserved: Bogoliubov's diagonalization of the weakly interacting Bose gas [898], giving the excitation spectrum $\varepsilon(k)=\sqrt{(\hbar^2k^2/2m)(\hbar^2k^2/2m+2gn)}$, phonon-like at long wavelength with sound speed $c=\sqrt{gn/m}$; the Landau criterion and the measured critical velocity for a stirred laser beam [1982]; why condensation and superfluidity are distinct properties, related in Chapter 130.]

121.5.3 What was and was not shown

[Reserved: the honest ledger — what the 1995 data establish (a phase transition at the predicted temperature, with the predicted condensate fraction and the predicted anisotropic expansion), what they do not (nothing about the interacting many-body ground state beyond mean field), and which later measurements closed each gap; statistical versus systematic uncertainty in atom number, the dominant error throughout; the relation to the general theory of phase transitions in Chapter 123.]

121.6 Repetitions, extensions and precision

121.6.1 Other species, and hydrogen

[Reserved: condensation of atomic hydrogen, the species the programme began with, reached three years after the alkalis and with 10^9 atoms [1953]; the subsequent extension to metastable helium, chromium, and the lanthanides; two-photon spectroscopy of the hydrogen condensate as the link to Chapter 90; the present status of species and atom numbers, tabulated with sources.]

121.6.2 Vortices and quantized circulation

[Reserved: the first vortex, imprinted by a coherent two-component coupling and imaged interferometrically [1979]; vortex nucleation by mechanical stirring above a critical rotation frequency [1998]; the Abrikosov lattice of up to 130 vortices observed in a rotating sodium condensate [2006]; the circulation quantum h/m measured directly, and the parallel with flux quantization in Chapter 130.]

Phenomenon 121.5 (Quantized circulation and the vortex lattice)**.** A condensate does not rotate as a fluid body. Set into rotation it remains at rest until a critical angular frequency is exceeded, and then admits angular momentum only in the form of discrete vortex lines, each carrying one quantum of circulation h/m and a density node at its core [1979, 1998]. Above threshold the lines multiply and arrange themselves in a regular triangular array, up to about 130 of them counted in one rotating sodium condensate [2006]. The array is the direct image of a quantity that classical hydrodynamics allows to take any value whatever.

Derivation. The condensate is described by a single complex field $\psi = \sqrt{n_0}\,\mathrm{e}^{\mathrm{i}\theta}$, and the current it carries identifies the superfluid velocity as $\boldsymbol{v} = (\hbar/m)\nabla\theta$. Let C be any closed contour along which ψ is non-zero. Since ψ is a single-valued function of position, θ can change round C only by an integer multiple of 2π, and therefore

$$\oint_C \boldsymbol{v}\cdot \mathrm{d}\boldsymbol{l} = \frac{\hbar}{m}\oint_C \nabla\theta\cdot \mathrm{d}\boldsymbol{l} = \frac{\hbar}{m}\,2\pi n = n\,\frac{h}{m}\,, \qquad n\in\mathbb{Z}\,. \tag{121.4}$$

The circulation is thus quantized in units of h/m, with a value that depends on nothing — not the contour, not the density, not the strength of the interaction — but the integer n. Two consequences follow at once. Since the flow is irrotational wherever $\psi \neq 0$, non-zero circulation requires ψ to vanish somewhere inside C: a vortex has a core. And since n cannot vary continuously, a rotating condensate cannot take up angular momentum smoothly and must instead nucleate whole vortices one at a time above a threshold rotation rate, which is what the observations show. □

121.6.3 Optical lattices

[Reserved: the condensate loaded into a standing-wave potential as a defect-free realization of the lattice models of Chapter 125 and Chapter 126; the observed superfluid-to-Mott-insulator transition, a quantum phase transition driven by the ratio of tunnelling to on-site repulsion [2029]; the loss and revival of interference fringes as its order parameter; the review of the field [2118].]

Phenomenon 121.6 (A transition driven by a ratio, not a temperature)**.** A condensate loaded into a three-dimensional optical standing wave loses its interference pattern abruptly as the lattice depth is increased, and recovers it on lowering the depth again. The loss is reversible, occurs at a definite lattice depth, and is not caused by heating: it marks a transition from a superfluid, in which the atoms are delocalized and the phase is shared, to an insulating state with a fixed integer number of atoms per lattice site and a gap for adding one more [2029]. The control parameter is the ratio of the tunnelling amplitude between neighbouring sites to the energy cost of double occupancy, a purely quantum-mechanical competition that survives at zero temperature.

Derivation pending #226

The Bose–Hubbard description of atoms in a deep optical lattice, the tunnelling and on-site interaction parameters expressed through the lattice depth and the scattering length, and the mean-field estimate of the ratio at which the compressible superfluid gives way to the gapped, integer-filled insulator

121.6.4 Degenerate Fermi gases and the BEC–BCS crossover

[Reserved: the fermionic counterpart — onset of Fermi degeneracy in trapped potassium-40 [1971]; molecular condensates formed from fermion pairs on a Feshbach resonance [2051]; condensation of fermionic atom pairs on the BCS side of the resonance [2065]; vortex lattices as the unambiguous proof of superfluidity in the strongly interacting regime [2087]; the crossover as a controlled interpolation between Bose condensation and BCS pairing (Chapter 130).]

121.6.5 Recognition

[Reserved: the 1997 award for laser cooling and trapping, with its three Nobel lectures as authoritative first-hand accounts of the apparatus [1949, 1951, 1959]; the 2001 award for the

condensates themselves [2025, 2030]; a note on what the prize citations record about priority among the JILA, MIT and Rice groups, and on the hydrogen programme that supplied the evaporation technique without reaching the transition first.]

Chapter 122

Quantum Optics and the Photon

Light is the cleanest quantum-statistical laboratory there is. A mode of the electromagnetic field is exactly a harmonic oscillator (Chapter 28), its excitations are bosons (Chapter 120), and detectors count them one at a time; so the statistical questions that are awkward to pose for a gas of atoms — what is the distribution of occupation numbers, what are the correlations between successive detections, what states have no classical description at all — can here be asked and answered with photodetectors and coincidence counters. This chapter takes photon statistics, coherence, antibunching, squeezing and cavity electrodynamics in that order, and in each case names the measurement that settled the matter. Its running theme is the one question that took sixty years to close: which optical phenomena actually require the field to be quantized, and which are reproduced by a classical wave falling on quantized matter.

It sits deliberately inside the statistical part rather than in Part IX: the objects of study are not single-atom amplitudes but the correlation functions of a many-boson field, and the Bose statistics of Chapter 120 is what makes photons bunch. The field quantization it uses is set up in Chapter 64 and treated dynamically in Chapter 100; the entanglement of photon pairs is the material of Chapter 84 and Chapter 89, and the coherence measurements here are the optical counterpart of the atomic ones in Chapter 121. Standard monographs are [1904, 1996].

Derivation pending #227

Quantum Optics and the Photon: all derivations of this chapter are pending.

122.1 The quantized radiation field

122.1.1 Modes, Fock states and the vacuum

[Reserved: Dirac's quantization of the free field as a set of independent oscillators, and the first derivation of the emission and absorption rates from it [572]; annihilation and creation operators with $\left[a, a^\dagger\right] = 1$*, number states* $|n\rangle$*, and the vacuum energy* $\frac{1}{2}\hbar\omega$ *per mode; the quadrature operators and their commutator as the origin of every uncertainty statement in this chapter; the field-theoretic setting of Chapter 64.]*

122.1.2 Einstein A and B coefficients

[Reserved: Einstein's statistical argument that detailed balance between matter and blackbody radiation requires stimulated emission in addition to absorption and spontaneous emission, with $A/B = 8\pi h\nu^3/c^3$ *[446]; the momentum transfer* $h\nu/c$ *in the same paper, later the*

basis of laser cooling (Chapter 121); the relation to the Planck spectrum of Chapter 68; why stimulated emission makes the laser of Section 122.7 possible.]

122.1.3 Is the photon necessary?

[Reserved: the honest accounting — the photoelectric effect (Chapter 69) and blackbody spectrum admit semiclassical explanations with quantized matter and a classical field, so neither proves field quantization; the experiments that do: the photoelectric coincidence test that violates a classical inequality [1521], the beam-splitter anticorrelation with a single-photon input [1774], and the antibunching of Section 122.4; Compton scattering as the momentum-space argument.]

122.2 Photon statistics

122.2.1 Photodetection and photocount distributions

[Reserved: the photodetection formalism and the normally ordered counting formula [1256]; the Poissonian distribution $P(n) = \mathrm{e}^{-\bar{n}}\bar{n}^n/n!$ for a stable classical wave against the Bose–Einstein distribution $P(n) = \bar{n}^n/(1+\bar{n})^{n+1}$ for a single thermal mode, derived in Chapter 120; the measurement that exhibited both distributions from one apparatus by driving a laser above and below threshold [1302]; dead time and quantum efficiency as the systematic limits.]

122.2.2 Poissonian, super- and sub-Poissonian light

[Reserved: the variance of the photon number as a classifier; the Mandel parameter $Q = (\langle(\Delta n)^2\rangle - \bar{n})/\bar{n}$, negative only for states with no positive-definite classical distribution [1904]; the first observation of sub-Poissonian light, in the fluorescence of a single atom [1728]; photon bunching and antibunching as the temporal statement of the same fact.]

122.2.3 The Hanbury Brown–Twiss experiment

[Reserved: intensity interferometry as a way of measuring stellar angular diameters with baselines immune to atmospheric phase noise, first demonstrated at radio wavelengths [1031]; the laboratory experiment with a mercury arc, showing a positive correlation between the currents of two photodetectors on the two arms of a beam splitter [1083]; the measurement of the angular diameter of Sirius, 3.1×10^{-8} radians — about 6.3(5) milliarcseconds [1086]; the objection that photons could not correlate at all, and Purcell's resolution in terms of Bose statistics [1095], which makes this the founding measurement of the chapter.]

122.3 Coherence

122.3.1 Correlation functions and the Glauber hierarchy

[Reserved: Glauber's normally ordered correlation functions $G^{(n)}$ and the definition of coherence to all orders, replacing the first-order (interferometric) notion inherited from classical optics [1256]; the normalized second-order function $g^{(2)}(\tau)$, with $g^{(2)}(0) = 2$ for thermal light, 1 for a coherent state and 0 for a single photon; the classical inequalities $g^{(2)}(0) \geq 1$ and $g^{(2)}(0) \geq g^{(2)}(\tau)$ whose violation is the operational definition of nonclassical light.]

122.3.2 Coherent states

[Reserved: the eigenstates of the annihilation operator, $a\,|\alpha\rangle = \alpha\,|\alpha\rangle$, their Poissonian number distribution, minimum-uncertainty quadratures and non-orthogonal overcompleteness [1255]; why they are the quantum state closest to a classical field and the state a laser well above threshold actually produces; the displacement operator and the coherent state as a displaced vacuum, connecting to the harmonic oscillator of Chapter 78.]

122.3.3 The optical equivalence theorem

[Reserved: the diagonal coherent-state representation of an arbitrary density operator, introduced independently by Sudarshan [1267] and Glauber [1255]; the theorem that every normally ordered expectation is a classical average over the weight $P(\alpha)$; the precise sense in which a state is nonclassical — P fails to be a probability density — and why that criterion, not "smallness" of the field, is the right one; the standing dispute over naming, recorded honestly.]

122.3.4 Nonclassical light

[Reserved: the catalogue of states with no classical analogue and the signature by which each is recognized — number states (sub-Poissonian statistics), antibunched fluorescence (Section 122.4), squeezed states (Section 122.5) and entangled photon pairs (Chapter 84); quantum state tomography by balanced homodyne detection as the modern way to exhibit a negative quasiprobability; the reconstruction of Fock states and their Wigner functions [2017].]

122.4 Antibunching and single photons

122.4.1 Resonance fluorescence and photon antibunching

[Reserved: the prediction that light scattered by a single atom must show $g^{(2)}(0) < g^{(2)}(\tau)$, because the atom cannot emit again until it has been re-excited [1574]; the measurement on a dilute sodium beam that observed it [1598], the first phenomenon with no classical wave description whatever; the Mollow triplet and Rabi oscillations visible in the correlation function; residual multi-atom events as the systematic that had to be extrapolated away.]

122.4.2 Single-photon sources

[Reserved: the progression from attenuated lasers (which remain Poissonian and are not single-photon sources) to true antibunched emitters — single molecules at room temperature [1997], nitrogen-vacancy centres in diamond [1995], semiconductor quantum dots in a turnstile device [2000], and deterministic emission from a single atom in a cavity [2032]; the figures of merit — purity $g^{(2)}(0)$, indistinguishability, and collection efficiency — with their currently measured values.]

122.4.3 Two-photon interference

[Reserved: the Hong–Ou–Mandel effect, in which two indistinguishable photons entering opposite ports of a balanced beam splitter always leave together, producing a coincidence dip whose width measures their wavepacket overlap to subpicosecond precision [1794]; the dip as an interference of amplitudes, not of intensities, and its use as the standard test of photon indistinguishability; the connection to boson exchange symmetry in Chapter 83.]

122.5 Squeezed light

122.5.1 Quadratures and the uncertainty ellipse

[Reserved: the quadratures X_1, X_2 with $\langle \Delta X_1^2 \rangle \langle \Delta X_2^2 \rangle \geq 1/16$; squeezing as redistribution of vacuum fluctuations between them at fixed product; the review that set the programme [1730]; Caves's demonstration that the shot-noise limit of an interferometer is set by vacuum fluctuations entering the unused port, and can be beaten by injecting squeezed vacuum there [1667] — the theoretical origin of Section 122.5.3.]

122.5.2 Generation and detection

[Reserved: the first observation of squeezed light, by four-wave mixing in a sodium-vapour cavity, with a noise reduction of about 7 % below vacuum [1763]; parametric down-conversion in an optical parametric oscillator, which reached more than 60 % and remains the standard method [1785]; balanced homodyne detection with a local oscillator as the phase-sensitive measurement; loss as the enemy, since every inefficiency mixes vacuum back in.]

122.5.3 Squeezed light in gravitational-wave detectors

[Reserved: the application that made squeezing an engineering tool — injection of squeezed vacuum into the dark port of the kilometre-scale interferometers of Advanced LIGO [2243], first demonstrated in LIGO at about 2 decibels of shot-noise reduction [2204] and now routine at roughly 3 decibels across the observing band [2342]; the resulting increase in detection rate; radiation-pressure back-action and frequency-dependent squeezing as the response; links to Chapter 47.]

122.6 Cavity quantum electrodynamics

122.6.1 The Jaynes–Cummings model

[Reserved: one two-level atom coupled to one field mode under the rotating-wave approximation, and the comparison with the semiclassical theory it was written to test [1258]; the dressed-state ladder with splitting $2g\sqrt{n+1}$; collapse and revival of the Rabi oscillation as a signature that the field is quantized, observed in the one-atom maser [1797] and resolved into individual Fock components in the microwave regime [1917].]

122.6.2 Modified spontaneous emission

[Reserved: spontaneous emission as a property of the mode structure rather than of the atom alone; Purcell's enhancement factor $F = 3Q\lambda^3/4\pi^2 V$ for an atom in a resonant cavity [893]; the observed enhancement for a Rydberg atom between mirrors [1719] and the complementary inhibition when the cavity is detuned [1755]; the practical consequence for the single-photon sources of Section 122.4.2.]

122.6.3 Strong coupling and vacuum Rabi splitting

[Reserved: the strong-coupling condition $g > \kappa, \gamma$ in which a single quantum is exchanged coherently before it is lost; the observation of normal-mode splitting for one atom in an optical cavity [1859], the optical counterpart of the microwave work; the extension to solid-state systems; why the splitting scales as $\sqrt{N}$ with atom number, and what a single-atom measurement therefore has to control.]

122.6.4 Photon counting without absorption

[Reserved: quantum non-demolition measurement of photon number by the phase shift imprinted on a Rydberg atom crossing the cavity; the progressive collapse of a coherent field onto a Fock state, observed photon by photon, and the subsequent quantum jumps as photons decay [2107]; the same apparatus used to watch a mesoscopic superposition decohere at a rate proportional to its size [1916], an experimental entry point to Chapter 87; the first-hand account [2220].]

122.7 The laser

122.7.1 Stimulated emission, inversion and the maser

[Reserved: population inversion as the condition for net gain, and why it is forbidden in thermal equilibrium (Chapter 118); the ammonia beam maser, the first device to oscillate on stimulated emission [1039]; three- and four-level pumping schemes; the inverted medium as a negative-temperature system, and the care that statement requires.]

122.7.2 The optical maser and its realization

[Reserved: the proposal to extend maser action to optical frequencies using a Fabry–Pérot resonator to select modes, with the threshold condition and the predicted linewidth [1160]; the first working laser, a flash-pumped ruby rod emitting at 694.3 nm *[1186]; the semiclassical theory of the laser and its mode competition [1287]; the helium–neon and semiconductor lasers that made the technique general.]*

122.7.3 Laser linewidth and coherence

[Reserved: the Schawlow–Townes limit $\Delta\nu = \pi h\nu(\Delta\nu_c)^2/P$*, spontaneous emission into the lasing mode as the irreducible source of phase diffusion [1160]; the output as a coherent state with a slowly diffusing phase, and its measured* $g^{(2)}(0) = 1$ *[1302]; the laser transition as a nonequilibrium phase transition with the field amplitude as order parameter (Chapter 123); modern stabilized lasers and the frequency comb, whose metrology belongs to Chapter 90.]*

122.8 Zewail: femtosecond probing of the transition state (1987)

The laser of Section 122.7 enters here not as the object of study but as a shutter. Mode locking compresses its output into packets shorter than one vibration of a chemical bond, and two such packets — one to start a reaction, one to photograph it — resolve the motion of the nuclei themselves. What that buys is the configuration a reacting molecule occupies on its way from reactants to products: the activated complex of Eyring's rate theory [754], which for half a century had been an inference drawn from measured rate constants and never an observation, because it endures only as long as it takes two atoms to move apart by about their own diameter. Femtochemistry is the demonstration that this interval is long enough to be timed [1790, 1813].

The section sits in this chapter as the ultrafast limit of the material around it. The pulse doing the timing is a coherent state (Section 122.3.2) whose duration and spectral width are conjugate through the Fourier relation of Chapter 17, so a shutter short enough to resolve nuclear motion is necessarily broad enough to span the vibrational structure it interrogates; that structure is the molecular spectroscopy of Chapter 88.

122.8.1 Apparatus

A colliding-pulse mode-locked ring dye laser [1669], the source that first delivered pulses shorter than 100 fs, amplified and frequency-converted into two synchronized beams: an ultraviolet *pump* lying inside the dissociative absorption continuum of the parent molecule, and a tunable *probe* resonant with an electronic transition of one of the fragments. The two travel separate arms of a Michelson-type splitter, the probe arm carrying a retroreflector on a computer-driven translation stage. Sample: a collision-free beam of the parent molecule, so that what is timed is one molecule's internal motion and not an encounter with a neighbour. Detection: laser-induced fluorescence excited by the probe, collected by a photomultiplier and gated so that only light following the probe is counted.

The translation stage is the clock, and its arithmetic is why the technique is possible at all. Light travels 0.3 µm in 1 fs, so a delay step of 10 fs is 1.5 µm of stage travel in a double-pass arm — an ordinary mechanical tolerance. No electronic gate resolves 100 fs; a screw thread does.

122.8.2 Procedure

The pump promotes the molecule to a dissociative surface and thereby fixes the zero of time, to within the pulse duration. The probe arrives a delay τ later and lifts whatever the system has become to a fluorescing state; the fluorescence yield, recorded against τ, is the signal. Two probe settings are used, and they answer different questions. Tuned to a transition of the *free* fragment, the probe counts finished products, and the curve is a rise to a plateau whose half-rise time is the appearance time of those products. Detuned from that transition, so that it is resonant only while the departing fragments still perturb each other's levels, the probe counts molecules caught in transit, and the curve is a transient that rises and falls. The delay is scanned in steps, many shots are averaged at each step, and the pump–probe cross-correlation, measured in a nonlinear crystal, fixes both the time resolution and the origin.

[Reserved: the systematic checks — pump and probe intensity dependence, confirming one photon of each; beam density low enough to exclude collisions inside the observation window; calibration of zero delay and of the instrument response function; the deconvolution by which a transient narrower than the cross-correlation is recovered.]

122.8.3 Observations and data

Phenomenon 122.1 (A chemical bond carries its own femtosecond clock)**.** The vibrational fundamentals of molecules, read off infrared absorption and Raman scattering, occupy a band of wavenumbers running from about 100/cm for the bending of heavy atoms to about 4400/cm for the stretch of the hydrogen molecule [961]. The period of the corresponding nuclear motion,

$$T = \frac{1}{\nu} = \frac{1}{c\tilde{\nu}}\,, \tag{122.1}$$

therefore lies between about 330 fs and 8 fs. A configuration the nuclei pass through once and do not return to — the transition state of a dissociation — endures for a time of this same order, and is invisible to any probe slower than it.

Derivation. A wavenumber is a frequency divided by the speed of light, so $\nu = c\tilde{\nu}$ and the period is $T = 1/(c\tilde{\nu})$, which is Equation (122.1). A typical stretching fundamental at 1000/cm, that is 1×10^5/m, gives a frequency of 3.0×10^{13} Hz and a period of 33 fs; the two ends of the observed band give 330 fs and about 8 fs. A second and independent estimate uses only kinematics: the fragments of a dissociation stop interacting once they have separated by a distance of order the bond length itself, about 0.1 nm, and they separate

at the recoil speed set by the energy released, of order 1 km/s, so the transit time d/v is of order 100 fs. Both routes give the same order of magnitude, and the order is what matters: a probe of duration Δt cannot resolve structure in time shorter than Δt, so an apparatus that is to watch nuclei move must carry a shutter of a few tens of femtoseconds. Nothing electronic is that fast. The shutter therefore has to be the light pulse itself, and the delay has to be generated as a difference of optical path. □

Phenomenon 122.2 (The transition state lasts long enough to be timed)**.** Photodissociation of a triatomic — the worked case is ICN → I + CN — started by an ultraviolet pump and interrogated by a delayed probe yields two distinguishable signals. With the probe tuned to a transition of the *free* CN fragment, the fluorescence rises from zero and levels off: the products appear after a delay rather than at once. With the probe detuned from that transition, so that it is resonant only while the departing fragments still perturb each other's levels, the signal instead rises and falls again — a transient that exists only during the passage through the intermediate configuration. Both the delay and the width of the transient are of order 100 fs [1790, 1813].

Derivation pending #228

The shape of the pump–probe signal for two fragments separating on a repulsive potential energy surface: the classical clocking relation that maps probe detuning onto internuclear separation, the wave-packet computation of the fluorescence yield as a function of delay, and the extraction of the free-fragment appearance time from the saturating on-resonance curve

Phenomenon 122.3 (A dissociating molecule can vibrate before it comes apart)**.** When the surface reached by the pump is not simply repulsive but is held together at large separation by an avoided crossing with a second surface, the same measurement returns a train of transients instead of one. In sodium iodide, whose covalent and ionic curves cross at several times the equilibrium bond length, the nuclear wave packet prepared by the pump oscillates in the resulting well and returns to the same internuclear separation once per vibrational period, a fixed fraction of it escaping to free atoms at each pass through the crossing. The probe accordingly records an oscillation whose period is that of the trapped motion and whose envelope decays, while the signal from free sodium rises in steps synchronized with it [1812].

Derivation pending #229

The two-state model of alkali-halide dissociation: the diabatic covalent and ionic curves and the adiabatic well their crossing produces, the Landau–Zener probability of remaining on the adiabatic branch at each pass, and the resulting periodic recurrences with a geometrically decaying envelope and a step-wise product signal

122.8.4 Interpretation

Rate theory describes a reaction through an activated complex in equilibrium with the reactants and reports one number, the rate constant, which is an average over an enormous number of uncorrelated barrier crossings [754]. The pump–probe measurement removes that average: preparing every molecule in the sample at a common instant makes the subsequent evolution of the ensemble a function of time, and it is that evolution, not its integral, which is recorded. Three consequences follow. The transition state moves from the class of inferred quantities to the class of observed ones, since it now has a duration and the duration is what

appears on the chart. The reaction coordinate is followed rather than postulated, because the detuning of the probe selects the internuclear separation at which the system is counted. And the assumption on which rate theory rests, that the barrier region is crossed once, becomes testable — in the alkali halide of Phenomenon 122.3 it visibly fails, the system crossing and recrossing many times before it commits.

Two limits should be stated with the result. The measured interval is a transit time, not the lifetime of a stationary state, so it has no meaning independent of the observable used to define it; and the zero of time and the resolution are both fixed by the pump–probe cross-correlation, so every quoted duration is only as good as the characterization of that correlation.

[Reserved: the extension of the method to bimolecular reactions initiated inside a van der Waals complex, to the condensed phase where the solvent turns the problem into one of Chapter 87, and to ultrafast electron and X-ray diffraction, which replace the spectroscopic probe by one that reads geometry directly; coherent control as the inverse problem, in which the pulse is shaped to steer the outcome rather than to observe it.]

122.8.5 Primary references

[1790, 1812, 1813]. The light source that made the timescale accessible is [1669]; the transition-state concept under test is [754]; Zewail's own survey of the field, covering both measurements above and much besides, is [2004].

122.9 Which-path information and the quantum eraser

122.9.1 Complementarity and the duality relation

[Reserved: the quantitative duality relation $D^2 + V^2 \leq 1$ between which-path distinguishability and fringe visibility [1920], replacing the qualitative slogan; Wheeler's delayed-choice configuration realized with a single-photon source and a randomly switched output beam splitter [2108]; why momentum transfer is not the mechanism, and what remains of the uncertainty-principle account.]

122.9.2 The quantum eraser

[Reserved: the proposal to mark the path with an internal atomic state and then erase the mark, restoring interference in coincidence [1702]; the delayed-choice realization with down-converted photon pairs, in which the decision to erase is taken after the signal photon has been detected [1994]; the essential and often-omitted point that no fringes appear in the unconditioned signal, so no signalling is possible; the interpretive questions deferred to Chapter 85.]

Chapter 123

Phase Transitions and Critical Phenomena

A phase transition is a singularity in a thermodynamic potential, and therefore something the finite-volume statistical mechanics of Chapter 118 cannot produce: partition functions of finitely many degrees of freedom are analytic, so every transition is a statement about the thermodynamic limit. That is the organizing fact of this chapter. It begins with phase diagrams, latent heats and the liquid–gas critical point, whose discovery [174] and first theory [187] set the vocabulary; it introduces the Ising paradigm [525] and its exact two-dimensional solution [875], the one non-trivial case in which the singularity can be watched forming; and it then treats order parameters, critical exponents and universality — the discovery that a fluid near its critical point and a uniaxial magnet near its Curie point obey the same power laws with the same exponents, so that microscopic detail is irrelevant to the singular behaviour. The measured exponents that vindicate that picture are collected here, against the renormalization-group explanation developed in Chapter 108.

The chapter sits after the equilibrium ensembles of Chapter 118 and the quantum statistics of Chapter 120, both of which supply the free energies it differentiates, and before the specific ordered phases of the rest of the part: the magnetic transitions of Chapter 129, the superfluid and superconducting transitions of Chapter 130, and the condensation of Chapter 121. Its last section treats liquid crystals, whose mesophases are the cleanest laboratory realization of an orientational order parameter. Standard monographs are [1456, 1855].

Derivation pending #230

Phase Transitions and Critical Phenomena: all derivations of this chapter are pending.

123.1 Phases and phase diagrams

123.1.1 Coexistence, triple and critical points

[Reserved: the phase diagram in the pressure–temperature plane, its coexistence curves, triple point and critical point; Gibbs's phase rule $f = c - p + 2$ derived from equality of the chemical potentials [193]; the triple point of water as a former defining fixture of the kelvin, now superseded (Chapter 2); metastability, superheating and nucleation as the reason a first-order transition can be delayed but not avoided.]

123.1.2 Latent heat and the Clausius–Clapeyron relation

[Reserved: latent heat as the discontinuity in entropy across a first-order line, and the slope relation $\mathrm{d}P/\mathrm{d}T = L/(T\,\Delta v)$ [95], whose thermodynamic derivation belongs to Chapter 54; measured latent heats of fusion and vaporization in SI, with the anomalous negative slope of the ice–water line and its consequences; the vanishing of L at the critical point as the definition of second order.]

Phenomenon 123.1 (Latent heat and the slope of a coexistence line)**.** Along a first-order phase boundary two phases coexist at one pressure and one temperature, and converting a mole of the first into the second absorbs a finite latent heat L at constant temperature while the molar volume jumps discontinuously by Δv. The boundary in the pressure–temperature plane is then not free: its slope is fixed by those two measured discontinuities,

$$\frac{\mathrm{d}P}{\mathrm{d}T} = \frac{L}{T\,\Delta v}\,, \tag{123.1}$$

a relation of Clapeyron [95] that has held wherever it has been tested. Its most conspicuous instance is an anomaly: water contracts on melting, so $\Delta v < 0$ for the ice–water line, whose slope is therefore negative — pressure lowers the melting point of ice, which is true of very few substances.

Derivation. Two phases in equilibrium at (T, P) have equal molar Gibbs functions, $\mu_1(T,P) = \mu_2(T,P)$ [193]. Displace along the coexistence curve by $(\mathrm{d}T, \mathrm{d}P)$; the equality must persist, so $\mathrm{d}\mu_1 = \mathrm{d}\mu_2$. The Gibbs–Duhem relation for each phase gives $\mathrm{d}\mu = -s\,\mathrm{d}T + v\,\mathrm{d}P$ with s and v the molar entropy and volume, whence

$$-s_1\mathrm{d}T + v_1\mathrm{d}P = -s_2\mathrm{d}T + v_2\mathrm{d}P \quad\Longrightarrow\quad \frac{\mathrm{d}P}{\mathrm{d}T} = \frac{s_2 - s_1}{v_2 - v_1} = \frac{\Delta s}{\Delta v}\,. \tag{123.2}$$

The conversion is reversible and isothermal, so the heat absorbed is $L = T\Delta s$; substituting gives Equation (123.1). Nothing in the argument uses a model of either phase, which is exactly what makes Equation (123.1) a test rather than a fit: L, Δv and the slope are three independently measurable quantities constrained by one identity. The sign of the slope is the sign of Δv, since $L > 0$ on melting and on boiling; the ice anomaly is therefore a statement about the density of ice and nothing else. □

123.1.3 The Ehrenfest classification and its limits

[Reserved: Ehrenfest's ordering of transitions by the lowest discontinuous derivative of the free energy [719]; why the scheme fails in practice — the specific heat at the helium lambda point and at the Curie point diverges rather than jumping, so there is no "second derivative discontinuity" to classify; the modern two-way division into first-order (latent heat, coexistence) and continuous (diverging correlation length) transitions, adopted for the rest of the chapter.]

123.2 The liquid–gas transition

123.2.1 Andrews and the critical point

[Reserved: Andrews's isotherms of carbon dioxide, showing the coexistence region shrinking to a point at $T_c = 304.1\,\mathrm{K}$ and $P_c = 7.38\,\mathrm{MPa}$, above which liquid and gas are not distinguishable [174]; the continuity of state that follows — one can pass from liquid to vapour without ever crossing a phase boundary; the flat critical isotherm and the diverging compressibility as the first measured critical exponents.]

Phenomenon 123.2 (The critical point and the continuity of state)**.** Compressing a fluid along isotherms of rising temperature, the flat portion of the isotherm — the interval over which liquid and vapour coexist, the volume changes and the pressure does not — grows shorter and finally closes at a single point. Andrews located that point for carbon dioxide at $T_c = 304.1\,\mathrm{K}$ and $P_c = 7.38\,\mathrm{MPa}$ [174]. Above T_c no compression produces a meniscus, and one may pass from a dense liquid to a dilute vapour along a path that circles the critical point without ever crossing a phase boundary: liquid and gas are not two kinds of matter but two regions of one surface. At the critical point the isotherm is flat, so the isothermal compressibility diverges, and the fluid becomes strongly scattering (Section 123.2.3).

Derivation. That such a point must exist in any equation of state combining a short-range repulsion with a long-range attraction is van der Waals' result [187]. Take

$$P = \frac{RT}{v - b} - \frac{a}{v^2}\,, \tag{123.3}$$

in which b is the excluded molar volume and a measures the mean attraction. A critical point is an inflection of the isotherm with horizontal tangent, so

$$\left(\frac{\partial P}{\partial v}\right)_T = -\frac{RT}{(v-b)^2} + \frac{2a}{v^3} = 0\,, \qquad \left(\frac{\partial^2 P}{\partial v^2}\right)_T = \frac{2RT}{(v-b)^3} - \frac{6a}{v^4} = 0\,. \tag{123.4}$$

Dividing the first of Equation (123.4) by the second eliminates RT and gives $v - b = \tfrac{2}{3}v$, hence $v_c = 3b$; substituting back yields

$$v_c = 3b\,, \qquad T_c = \frac{8a}{27Rb}\,, \qquad P_c = \frac{a}{27b^2}\,. \tag{123.5}$$

The critical point exists for every $a > 0$, $b > 0$, and the divergence of the compressibility there is immediate from the first of Equation (123.4). Eliminating a and b from Equation (123.5) gives the dimensionless combination $P_c v_c / RT_c = 3/8$ for every fluid, and rewriting Equation (123.3) in the reduced variables P/P_c, v/v_c, T/T_c removes a and b altogether: this is the law of corresponding states, the first statement of universality in physics. Two caveats belong with the result and are taken up later in this chapter. Below T_c, Equation (123.3) produces a loop along which $(\partial P/\partial v)_T > 0$, which is mechanically unstable and must be replaced by the horizontal tie line fixed by Maxwell's equal-area rule [191]. And the exponents that Equation (123.4) implies near the critical point are the mean-field ones, which measurement contradicts (Section 123.5.2). □

123.2.2 The van der Waals equation and the Maxwell construction

[Reserved: the equation $(P + a/v^2)(v - b) = RT$ derived from excluded volume and a mean attraction [187], the first theory with a critical point built in; the unphysical loop below T_c and Maxwell's equal-area rule that replaces it with the coexistence tie line [191]; the law of corresponding states in reduced variables; the mean-field exponents $\beta = 1/2$, $\gamma = 1$, $\delta = 3$ it predicts, and the measurements that contradict them (Section 123.5.2).]

123.2.3 Critical opalescence

[Reserved: the milky scattering of a fluid at its critical point, explained as light scattering from density fluctuations whose size grows with the correlation length [361]; the Ornstein–Zernike analysis of the pair correlation function and its Lorentzian structure factor [419]; opalescence as the first direct evidence that a diverging correlation length, not a diverging response alone, is what characterizes criticality.]

Phenomenon 123.3 (Critical opalescence)**.** A transparent fluid held within a fraction of a kelvin of its critical point turns milky white and scatters light so strongly that a centimetre of it is opaque; on warming or cooling by a degree the turbidity vanishes again. The scattered intensity is strongly peaked in the forward direction and its angular width narrows as the critical point is approached. The effect is not a mist of droplets — there is one phase present — but scattering from density fluctuations whose spatial extent has grown to the wavelength of visible light [361, 419]. It is the first direct evidence that what diverges at a critical point is a *length*, and that the divergence of the compressibility is a consequence rather than the essence.

Derivation pending #231

The scattered intensity as the Fourier transform of the density–density correlation function; the Ornstein–Zernike form with its Lorentzian structure factor and correlation length; and the fluctuation–response identity that ties the forward scattering to the isothermal compressibility, so that a diverging compressibility and a diverging correlation length are one statement

123.3 The Ising model

123.3.1 Lenz, Ising and the one-dimensional failure

[Reserved: the model of spins on a lattice with nearest-neighbour coupling, proposed by Lenz [468] and solved in one dimension by Ising, who found no transition at any positive temperature and wrongly concluded that none occurs in higher dimensions either [525]; the transfer-matrix solution and the domain-wall (entropy versus energy) argument that explains the one-dimensional result; the honest note that the model's name records its first solver, not its inventor.]

123.3.2 Existence of a transition in two dimensions

[Reserved: Peierls's contour argument, the first proof that the two-dimensional Ising model has spontaneous magnetization below some finite temperature [782]; the counting of domain boundaries that makes the same entropy–energy comparison come out the other way in two dimensions; the general lesson that the lower critical dimension depends on the symmetry of the order parameter, taken up in Section 123.6.1.]

123.3.3 Onsager's exact solution

[Reserved: the exact free energy of the two-dimensional Ising model on a square lattice in zero field, with $\sinh(2J/k_BT_c) = 1$ *and a logarithmically divergent specific heat [875]; the spontaneous magnetization* $m = (1 - \sinh^{-4} 2\beta J)^{1/8}$*, hence* $\beta = 1/8$*, announced by Onsager and first published in full by Yang [1013]; why an exponent of* $1/8$ *rather than* $1/2$ *was the decisive evidence against mean-field theory; the absence of any exact three-dimensional solution.]*

123.3.4 Lattice gas and the Yang–Lee theory

[Reserved: the exact mapping of the Ising magnet onto a lattice gas, which makes the ferromagnetic and liquid–gas transitions the same problem [1004]; the theory of condensation locating the singularity in the accumulation of zeros of the grand partition function in complex fugacity [1014], and — in the companion paper, which is where the result actually stands —

the circle theorem placing those zeros on the unit circle for the ferromagnetic Ising model [1004]; why this is the sharpest available statement that a phase transition exists only in the thermodynamic limit.]

123.4 Mean-field theory and order parameters

123.4.1 The Weiss molecular field

[Reserved: the Curie law $\chi \propto 1/T$ measured across a wide temperature range [264]; Weiss's hypothesis of an internal field proportional to the magnetization, giving the Curie–Weiss law $\chi \propto 1/(T - T_c)$ and a spontaneous moment below T_c [339]; the same approximation for alloy ordering [733]; the self-consistency equation $m = \tanh(T_c m/T)$ and its universality across mean-field models; magnetic instances in Chapter 129.]

123.4.2 Landau's expansion of the free energy

[Reserved: Landau's postulate that the free energy is analytic in the order parameter near T_c and constrained only by symmetry, $f = f_0 + a(T - T_c)m^2 + bm^4$ [795]; the classification of transitions by which terms symmetry permits, with a cubic term forcing first order; the resulting exponents $\alpha = 0$, $\beta = 1/2$, $\gamma = 1$, $\delta = 3$, identical to van der Waals and to Weiss; the gradient term that turns it into Landau–Ginzburg theory, used for superconductors in Chapter 130.]

123.4.3 Symmetry breaking and the order parameter

[Reserved: the order parameter as the quantity that is zero in the symmetric phase and non-zero below the transition, with its symmetry group and number of components as the only data that matter for universality; the catalogue — magnetization ($\mathbb{Z}_2$ or $\mathrm{SO}(3)$), density difference, condensate wavefunction ($\mathrm{U}(1)$), nematic director; spontaneous symmetry breaking as the selection of one of degenerate minima, the classical counterpart of the mechanism in Chapter 104; Goldstone modes for continuous symmetries.]

123.4.4 Where mean field fails

[Reserved: the Ginzburg criterion comparing the fluctuation of the order parameter within a correlation volume to its mean value, and the resulting estimate of the reduced-temperature window in which mean-field theory is wrong [1181]; the upper critical dimension $d_c = 4$ above which mean field is exact; why superconductors have a Ginzburg window too narrow to observe while fluids and magnets do not — the quantitative reason mean field survived so long.]

123.5 Critical exponents, scaling and universality

123.5.1 The exponents and the inequalities they obey

[Reserved: definitions of $\alpha, \beta, \gamma, \delta, \nu, \eta$ in terms of the reduced temperature $t = (T - T_c)/T_c$; the thermodynamic inequality $\alpha + 2\beta + \gamma \geq 2$ [1264] and its companion $\gamma \geq \beta(\delta - 1)$ [1308], proved from convexity alone; the empirical fact that both are satisfied as equalities, which is the clue scaling explains; the review that organized the field [1348].]

Phenomenon 123.4 (The exponent inequalities are satisfied as equalities)**.** Thermodynamics alone, using only the convexity of the free energy, bounds the measured critical exponents by $\alpha + 2\beta + \gamma \geq 2$ [1264] and $\gamma \geq \beta(\delta - 1)$ [1308]. What is observed, in fluids and

in magnets alike and across every universality class measured, is that both hold as *equalities* to within the experimental uncertainties [1348, 1782]. Of the six exponents $\alpha, \beta, \gamma, \delta, \nu, \eta$ only two are independent. That is a fact about nature, not about thermodynamics, and it is the clue from which the scaling picture was read.

Derivation. Widom's hypothesis [1323] is that the singular part of the free energy per unit volume is a generalized homogeneous function of the reduced temperature $t = (T - T_c)/T_c$ and the ordering field h,

$$f_s(t, h) = |t|^{2-\alpha}\, \Phi\!\left(\frac{h}{|t|^{\beta\delta}} \right) , \tag{123.6}$$

with Φ a scaling function finite and non-zero at the origin. Two differentiations extract the exponents. The specific heat is $c \sim -T\partial^2 f_s/\partial T^2 \sim |t|^{-\alpha}$ at $h = 0$, which is what fixes the prefactor in Equation (123.6). The order parameter is $m = -\partial f_s/\partial h \sim |t|^{2-\alpha-\beta\delta}$ at $h = 0$, and identifying this with $|t|^{\beta}$ gives

$$2 - \alpha = \beta + \beta\delta = \beta(1 + \delta)\,. \tag{123.7}$$

The susceptibility is $\chi = -\partial^2 f_s/\partial h^2 \sim |t|^{2-\alpha-2\beta\delta}$, and identifying it with $|t|^{-\gamma}$ gives, using Equation (123.7),

$$\gamma = 2\beta\delta - (2 - \alpha) = 2\beta\delta - \beta(1 + \delta) = \beta(\delta - 1)\,, \tag{123.8}$$

which is the Griffiths relation as an equality. Adding $\alpha + 2\beta$ to Equation (123.8) and using Equation (123.7) once more,

$$\alpha + 2\beta + \gamma = \alpha + 2\beta + \beta\delta - \beta = \alpha + \beta(1 + \delta) = \alpha + (2 - \alpha) = 2\,, \tag{123.9}$$

the Rushbrooke relation as an equality. Both are consequences of the single homogeneity assumption Equation (123.6), which also leaves exactly two free exponents — the scale $2-\alpha$ and the ratio $\beta\delta$ — as observed. The hypothesis itself is not derived here; it is explained by the renormalization group of Chapter 108, in which the homogeneity is the statement that the coarse-graining flow has a fixed point. □

123.5.2 Corresponding states and universality

[Reserved: Guggenheim's collapse of the coexistence curves of eight different fluids onto a single cubic curve with $\beta \approx 1/3$, not the mean-field $1/2$ [879]; the parallel measurement of the same exponent in uniaxial magnets; universality classes labelled by dimensionality and order-parameter symmetry alone; the modern precision values for the three-dimensional Ising and XY classes [2035] and the fluid data [1782].]

Phenomenon 123.5 (Universality)**.** Plotted in reduced variables, the liquid–vapour coexistence curves of eight chemically unrelated substances — among them neon, argon, krypton, xenon, nitrogen, oxygen, carbon monoxide and methane — fall on one and the same curve, and that curve is a cube root rather than the square root of mean-field theory: $\beta \approx 1/3$, not $\beta = 1/2$ [879]. The same exponent, within uncertainty, is measured for the spontaneous magnetization of a uniaxial ferromagnet near its Curie point, a system with nothing microscopically in common with a fluid [1348]. Modern determinations for the three-dimensional Ising class [2035] and for fluids [1782] agree to several decimal places. What survives at a critical point is therefore only the dimensionality of space and the symmetry and number of components of the order parameter; the chemistry drops out entirely.

Derivation pending #232

Universality as the basin of attraction of a fixed point of the coarse-graining flow: that the couplings distinguishing one microscopic system from another are irrelevant in the technical sense, so that the singular behaviour is governed by the fixed point alone; and the accompanying calculation of the exponents, developed in the renormalization-group chapter

123.5.3 Scaling hypotheses

[Reserved: Widom's homogeneity assumption for the singular part of the free energy, from which every exponent relation follows and only two exponents remain independent [1323]; Kadanoff's block-spin construction, which explains the homogeneity by coarse graining and introduces the flow of couplings under a change of scale [1334]; the data collapse of the equation of state onto a single scaling function as the experimental form of the hypothesis; hyperscaling $2-\alpha = d\nu$ and where it fails.]

123.5.4 Correlation functions and the correlation length

[Reserved: the pair correlation $G(r) \sim r^{-(d-2+\eta)}\mathrm{e}^{-r/\xi}$ and the divergence $\xi \sim |t|^{-\nu}$ as the physical content of criticality [419, 1348]; scale invariance at T_c, where ξ is infinite and the system looks the same at every magnification; measurement of ξ and η by neutron and light scattering; the fluctuation–response identity linking G to the susceptibility, an instance of Chapter 124.]

123.5.5 The renormalization group and measured exponents

[Reserved: Wilson's realization that Kadanoff's flow can be made a calculation, with fixed points, relevant and irrelevant operators, and universality as the basin of attraction of a fixed point [1461] — developed in full in Chapter 108; the epsilon expansion about four dimensions and the first controlled exponent calculation [1489]; the sharpest test on record, the specific-heat exponent of the helium lambda transition measured in microgravity as $\alpha = -0.0127(3)$ [2054], against theory at the same precision.]

Phenomenon 123.6 (A critical exponent measured to four decimal places)**.** The specific heat of liquid helium-4 at its lambda transition can be followed to within nanokelvin of T_c only in free fall, because on the ground the hydrostatic pressure head across a sample of any usable height smears the transition temperature. Measured aboard an orbiting laboratory, the singular specific heat yields $\alpha = -0.0127(3)$ [2054]: the exponent is small and *negative*, so the specific heat approaches a finite cusp rather than diverging. This is the most precisely measured critical exponent in physics, and it is a quantitative test of a calculated number rather than a qualitative confirmation of a picture.

Derivation pending #233

The superfluid transition as the three-dimensional XY universality class, and the calculation of its specific-heat exponent to a precision comparable with the measurement, by renormalization-group and high-order perturbative methods; together with the hyperscaling relation that ties the negative sign of the exponent to the correlation-length exponent in three dimensions

123.6 Lower critical dimension and topological order

123.6.1 The Mermin–Wagner–Hohenberg theorem

[Reserved: the proof that a continuous symmetry cannot break spontaneously in one or two dimensions at any positive temperature, because Goldstone fluctuations diverge logarithmically [1337], extended to superfluid and crystalline order [1350]; the contrast with the discrete Ising symmetry, which does order in two dimensions (Section 123.3.2); why the theorem does not forbid a transition of a different kind, which is Section 123.6.2.]

123.6.2 The Berezinskii–Kosterlitz–Thouless transition

[Reserved: the two-dimensional XY model, whose low-temperature phase has no long-range order but algebraically decaying correlations; the transition as the unbinding of vortex–antivortex pairs when the entropy of a free vortex overcomes its logarithmic energy [1442, 1506]; an essential rather than power-law singularity, so the exponent language of Section 123.5 does not apply; the predicted universal jump in superfluid density, $\rho_s/T_c = 2mk_B^2/\pi\hbar^2$ [1607].]

123.6.3 Observation in helium films

[Reserved: the torsional-oscillator measurement of superfluid density in adsorbed helium-4 films of varying thickness, which found the onset temperature proportional to the areal superfluid density with the predicted universal coefficient [1615]; why this is one of the very few parameter-free confirmations in the whole of critical phenomena; later realizations in two-dimensional atomic gases (Chapter 121) and in Josephson junction arrays.]

123.7 Liquid crystals

123.7.1 Discovery and the mesomorphic states

[Reserved: Reinitzer's observation that cholesteryl benzoate melts twice, at 418.6 K to a turbid fluid and at 451.6 K to a clear one [240]; Lehmann's polarizing-microscope identification of the intermediate phase as a flowing crystal [244]; Friedel's classification into nematic, smectic and cholesteric mesophases and his demonstration that they are true thermodynamic states, not emulsions [489].]

Phenomenon 123.7 (Matter that melts twice)**.** Cholesteryl benzoate has two melting points. Heated, the crystal becomes a turbid but freely flowing fluid at 418.6 K, and that fluid becomes an ordinary clear liquid at 451.6 K [240]. The intermediate state is birefringent under crossed polarizers, so it is optically anisotropic like a crystal while flowing like a liquid [244]. It is a genuine thermodynamic phase and not an emulsion or a mixture: its transitions are reversible and occur at reproducible temperatures, and its several varieties — nematic, smectic, cholesteric — are distinguished by their own textures and transitions [489]. Positional and orientational order are therefore separately destroyable, and a phase may possess one without the other.

Derivation pending #234

The orientational order parameter as a traceless symmetric second-rank tensor, forced by the equivalence of the director and its negative; the cubic invariant that this symmetry permits and that therefore makes the isotropic–nematic transition weakly first order, with a small latent heat and a jump in the order parameter; and the purely entropic hard-rod mechanism that produces the same transition with no attraction at all

123.7.2 The nematic transition

[Reserved: the nematic order parameter as a traceless symmetric tensor, since the director $\boldsymbol{n}$ and $-\boldsymbol{n}$ are equivalent, and the cubic invariant that symmetry therefore permits — forcing the isotropic–nematic transition to be weakly first order [1444]; Onsager's purely entropic transition for hard rods, driven by excluded volume alone [944]; the Maier–Saupe mean-field theory with attractive coupling [1153]; the measured latent heat and order-parameter jump, both small.]

123.7.3 Defects, elasticity and textures

[Reserved: the Frank free energy with its splay, twist and bend elastic constants [1146]; disclinations of half-integer strength, whose existence is a direct consequence of the director's headless symmetry and whose classification is the homotopy theory of Chapter 6; the schlieren textures by which they are counted under crossed polarizers; the monograph [1866]; electro-optic switching as the application that made the phase ubiquitous.]

Chapter 124

Nonequilibrium Thermodynamics and Transport

Equilibrium statistical mechanics describes states in which nothing happens. This chapter is about everything else: a bar with a temperature difference across it, a wire carrying current, a colloidal particle jostled by a fluid, a molecule pulled apart by an optical trap. Its central claim is that near equilibrium the whole of this behaviour is fixed by equilibrium fluctuations — Johnson–Nyquist noise [608, 615], the Einstein relation between diffusion and mobility, the Green–Kubo formulae [1041, 1117] and the fluctuation–dissipation theorem [979] are four statements of one fact — and that far from equilibrium the second law survives in a sharpened, quantitative form: the fluctuation theorems [1863, 1899] and the work relations [1938, 1969], which have been tested directly by pulling single RNA molecules [2033, 2075]. Onsager's reciprocal relations [684, 685], derived from microscopic time-reversal invariance, tie the cross coefficients together and are the reason a chapter on heat flow must also treat thermoelectricity.

The chapter builds on the kinetic theory of Chapter 55 in Part VI, where the Boltzmann equation and the H-theorem are set up, and on the ensembles of Chapter 118, whose correlation functions supply every transport coefficient computed here. It supplies in turn the scattering-limited conductivities used in Chapter 126 and Chapter 127, the critical dynamics that accompanies the static criticality of Chapter 123, and the classical limit of the open-system dynamics of Chapter 87. Standard monographs are [1228, 1356].

Derivation pending #235

Nonequilibrium Thermodynamics and Transport: all derivations of this chapter are pending.

124.1 The kinetic route

124.1.1 The Boltzmann equation and the H-theorem

[Reserved: a summary, not a re-derivation, of the material of Chapter 55: the integro-differential equation for the one-particle distribution with its collision integral, and the H-theorem showing $\mathrm{d}H/\mathrm{d}t \le 0$ *[183]; the assumption of molecular chaos as the exact point at which time-reversal symmetry is given up, and the Loschmidt and Zermelo objections; why the chapter takes the H-theorem as its starting point and asks instead what survives when the system is driven.]*

124.1.2 Chapman–Enskog and the transport coefficients

[Reserved: the systematic expansion of the Boltzmann equation in gradients about local equilibrium [434], which yields the Navier–Stokes equations of Chapter 31 together with first-principles expressions for viscosity and thermal conductivity; the predicted independence of the viscosity of a dilute gas from its density, and the measurements confirming it; the Prandtl number as a parameter-free test; the standard monograph [1416].]

124.1.3 The phenomenological transport laws

[Reserved: the four linear constitutive laws that the microscopic theory must reproduce — Fourier's law of heat conduction [72], Fick's law of diffusion [143], Ohm's law [81] and Newton's law of viscous shear [14]; their common structure, flux proportional to gradient; measured coefficients in SI for representative solids, liquids and gases; the range of validity, and the ballistic and hydrodynamic regimes in which each fails.]

124.2 Linear irreversible thermodynamics

124.2.1 Entropy production and thermodynamic forces

[Reserved: local equilibrium as the working hypothesis, and the entropy balance $\partial_t s + \nabla\cdot\boldsymbol{J}_s = \sigma$ with $\sigma = \sum_i J_i X_i \geq 0$ [1228]; the canonical pairing of fluxes with forces — heat flux with $\nabla(1/T)$, electric current with $\nabla(-\mu/T)$ — and why that pairing, not another, makes the coefficient matrix symmetric; measured entropy production rates in simple conduction experiments.]

124.2.2 Onsager reciprocal relations

[Reserved: the theorem $L_{ik} = L_{ki}$ for the linear coefficients relating fluxes to forces, derived from microscopic reversibility and the regression hypothesis for equilibrium fluctuations [684, 685]; the status of the regression hypothesis as the one non-trivial assumption; the independent experimental evidence, principally the Thomson relations of Section 124.3.2 and thermal diffusion in mixtures; why the relations count as a genuine prediction and not a definition.]

124.2.3 Reciprocity with a magnetic field

[Reserved: the extension $L_{ik}(\boldsymbol{B}) = L_{ki}(-\boldsymbol{B})$ required when the microscopic dynamics is reversible only under simultaneous reversal of the field [878]; the consequence for the Hall and Nernst–Ettingshausen coefficients, and the fact that reciprocity does not forbid an antisymmetric part; the same structure in rotating frames (Chapter 29); the quantized case in Chapter 128.]

124.2.4 Steady states and minimum entropy production

[Reserved: the theorem that in the linear regime a stationary state with fixed boundary constraints minimizes the entropy production [1356], with an explicit statement of its narrow conditions of validity; the honest caveat that no such variational principle is known beyond the linear regime, and that claims to the contrary are not supported by evidence; dissipative structures and pattern formation as the far-from-equilibrium alternative, whose dynamics belongs to Chapter 32.]

124.3 Thermoelectricity

124.3.1 The Seebeck, Peltier and Thomson effects

[Reserved: Seebeck's discovery that a circuit of two metals with junctions at different temperatures deflects a compass needle, and its correct reading as a thermally driven electromotive force [79]; Peltier's converse effect, heat absorbed or released at a junction carrying current [98]; Thomson's prediction and observation of the third, bulk effect in a single conductor with a temperature gradient [148]; measured Seebeck coefficients in $\mu\mathrm{V/K}$*.]*

124.3.2 The Thomson relations as a test of reciprocity

[Reserved: the two relations $\Pi = ST$ *and* $\mu_T = T\,\mathrm{d}S/\mathrm{d}T$ *connecting the three effects, obtained by Thomson from a thermodynamic argument he knew to be incomplete [148] and derived properly only from reciprocity [685]; the measured agreement, quoted with uncertainties, as the classic experimental confirmation of the Onsager relations; the thermocouple as a practical thermometer (Chapter 2).]*

124.3.3 Conduction in metals and the figure of merit

[Reserved: the Wiedemann–Franz law, the near-constancy of $\kappa/\sigma T$ *across metals [140]; the Drude model's accidental success and its failure on the specific heat [289], resolved only by Fermi statistics in Chapter 126; the Lorenz number* $L_0 = \pi^2 k_B^2/3e^2 = 2.44 \times 10^{-8}\,\mathrm{W\,\Omega/K^2}$ *and the conditions under which it is violated; the thermoelectric figure of merit* ZT *and the materials strategy of decoupling* κ *from* σ *[2136].]*

124.4 Brownian motion and stochastic dynamics

124.4.1 The Einstein–Smoluchowski relation

[Reserved: the derivation of the diffusion equation from a random-walk continuity argument and the identification $D = k_B T/6\pi\eta a$ *[318], obtained independently by Sutherland [324] and by a different route by Smoluchowski [331]; the general Einstein relation* $D = \mu k_B T$ *between diffusion constant and mobility as the first fluctuation–dissipation statement in physics; Perrin's measurement of Avogadro's number from displacement statistics [358], treated in full in Chapter 119.]*

124.4.2 The Langevin equation

[Reserved: the stochastic equation $m\ddot{x} = -\gamma\dot{x} + \xi(t)$ *with white noise satisfying* $\langle \xi(t)\xi(t')\rangle = 2\gamma k_B T\delta(t-t')$ *[344]; the second fluctuation–dissipation relation built into that variance, without which the particle would not thermalize; the exact velocity autocorrelation and mean square displacement, crossing from ballistic to diffusive at the momentum relaxation time [672]; the observation of the ballistic regime in an optically trapped bead [2160].]*

124.4.3 The Fokker–Planck equation

[Reserved: the evolution equation for the probability density in phase space, obtained for a rotating dipole in a radiation field [415] and given its general statistical-dynamical form by Planck [448]; drift and diffusion coefficients as the first two Kramers–Moyal moments; the stationary solution reproducing the Boltzmann distribution as the consistency condition; Kramers' escape rate over a potential barrier, with its low- and high-friction limits [853], and the Arrhenius factor it explains.]

124.4.4 Master equations

[Reserved: the gain–loss balance equation for the occupation probabilities of discrete states, and detailed balance as the condition for an equilibrium stationary solution; Pauli's derivation of an H-theorem for the quantum transition rates [617]; the quantum generalization to completely positive dynamics [1586], whose physical content is developed in Chapter 87; birth–death processes and chemical kinetics as classical instances.]

124.5 Linear response and fluctuation–dissipation

124.5.1 Johnson–Nyquist noise

[Reserved: Johnson's measurement of a temperature-dependent voltage fluctuation across any resistor, independent of its material and composition [608]; Nyquist's derivation of $\langle V^2 \rangle = 4 k_B T R \, \Delta f$ from equipartition among the modes of a transmission line, in the same issue [615]; its use as a primary thermometer and as the noise floor of every amplifier; shot noise as the distinct, non-thermal fluctuation, and the measured crossover between them.]

124.5.2 The fluctuation–dissipation theorem

[Reserved: the general theorem relating the imaginary part of a generalized susceptibility to the equilibrium spectral density of the conjugate variable, with the quantum form carrying a factor $\coth(\hbar\omega / 2 k_B T)$ [979]; Nyquist's formula and the Einstein relation as the two special cases already met; the review that established the modern statement [1336]; violation of the theorem as an operational definition of "out of equilibrium", and the effective temperatures inferred from it in glasses.]

124.5.3 Kubo formulae and Green–Kubo relations

[Reserved: transport coefficients as time integrals of equilibrium current autocorrelation functions — conductivity from the current–current correlator [1117], viscosity and thermal conductivity from the corresponding fluxes [1041]; the remarkable content of these formulae, that a dissipative quantity is computed entirely within the equilibrium ensemble; long-time tails and the divergence of the two-dimensional transport integrals; molecular-dynamics evaluation as the standard numerical route.]

124.5.4 Projection operators and memory kernels

[Reserved: the exact reduction of the full dynamics onto a chosen set of slow variables, producing a generalized Langevin equation with a memory kernel and a residual force [1197, 1316]; the second fluctuation–dissipation theorem identifying the kernel with the residual force correlation; Markovian approximation as the limit of separated timescales, and what is lost when it fails; the formal link to coarse graining in Chapter 108.]

124.6 Fluctuation theorems

124.6.1 The steady-state fluctuation theorem

[Reserved: the observation in shearing molecular dynamics that trajectories with negative entropy production occur with a probability ratio $\mathrm{e}^{-\Sigma t / k_B}$ against their conjugates [1863]; the proof for chaotic dynamical systems obeying the chaotic hypothesis [1899]; the reading of the second law as an inequality on an average that individual small systems may violate for short times; the smooth recovery of the thermodynamic statement as system size grows.]

124.6.2 Work relations far from equilibrium

[Reserved: Jarzynski's equality $\langle \mathrm{e}^{-W/k_BT} \rangle = \mathrm{e}^{-\Delta F/k_BT}$, valid for arbitrarily fast driving and yielding an equilibrium free-energy difference from irreversible measurements [1938]; Crooks's stronger relation between the forward and reverse work distributions, whose crossing point locates ΔF directly [1969]; recovery of $\langle W \rangle \geq \Delta F$ by Jensen's inequality, so the second law is a corollary; the exponential average's practical dominance by rare trajectories.]

124.6.3 Experimental verification

[Reserved: the mechanically unfolded RNA hairpin, whose measured work distributions recovered the equilibrium folding free energy to within a few k_BT and so verified the Jarzynski equality [2033]; the Crooks relation tested on RNA with and without a bound magnesium ion, including the far-from-equilibrium regime where the Jarzynski estimator is biased [2075]; the colloidal bead in a moving optical trap that exhibited transient second-law violations [2041]; the SI energy scale of these experiments, set by $k_BT \approx 4.14 \times 10^{-21}$ J at room temperature.]

124.6.4 Stochastic thermodynamics and information

[Reserved: heat, work and entropy defined along a single stochastic trajectory, and the resulting integral and detailed fluctuation theorems [2200]; the generalized second law including an information term, and its experimental demonstration by a feedback-controlled Brownian ratchet converting information into work [2167]; Maxwell's demon and Landauer's principle as the resolution — the cost lies in erasure, not in measurement; molecular motors as the biological instance.]

Chapter 125

Phonons and Lattice Dynamics

A crystal is the physical realization of the loaded string of Chapter 28: a periodic array of masses coupled by springs, whose normal modes are travelling waves and whose quanta — phonons — are bosons obeying the statistics of Chapter 120. Almost every thermal property of a solid follows from that identification. This chapter establishes the lattice itself as an observed structure, by X-ray diffraction [385, 399] rather than by assumption; derives its normal modes [381]; quantizes them; and then confronts the resulting heat capacity with calorimetric data, where the classical Dulong–Petit law [64] fails at low temperature and the quantum models of Einstein [333] and Debye [382] succeed. The dispersion relations that the theory predicts are not inferred but measured, branch by branch, by inelastic neutron scattering [1056, 1137] — the cleanest available demonstration that the phonon is a real excitation with a definite energy and momentum. The chapter closes on anharmonicity, without which a crystal would neither expand when heated [386] nor conduct heat at a finite rate [634].

It sits between the general theory of phase transitions (Chapter 123), which supplies the notion of a broken translation symmetry, and the electronic chapters Chapter 126 and Chapter 127, which inherit its reciprocal lattice and Brillouin zones wholesale; its transport section uses the machinery of Chapter 124, and the electron–phonon coupling it introduces is what pairs electrons in Chapter 130. Standard monographs are [1030, 1570, 2078].

Derivation pending #236

Phonons and Lattice Dynamics: all derivations of this chapter are pending.

125.1 The crystal lattice

125.1.1 Bravais lattices and the crystallographic restriction

[Reserved: the enumeration of the fourteen distinct lattices of points invariant under a discrete translation group in three dimensions [126]; primitive vectors, unit cells and the Wigner–Seitz construction; the crystallographic restriction theorem — only two-, three-, four- and six-fold rotation axes are compatible with translational periodicity — proved from the trace of the rotation in the lattice basis, an application of Chapter 5.]

125.1.2 Point groups, space groups and the basis

[Reserved: the 32 crystallographic point groups and the 230 space groups obtained by adjoining screw axes and glide planes, enumerated independently by Fedorov and by Schoenflies [251]; lattice plus basis as the description of a real crystal; Neumann's principle relating the

symmetry of a physical tensor to that of the crystal, which is what makes a cubic crystal optically isotropic but elastically anisotropic (Chapter 30).]

125.1.3 The reciprocal lattice and Brillouin zones

[Reserved: the reciprocal lattice as the Fourier dual of the direct lattice, with $\mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{R}} = 1$*, and the Ewald construction that makes diffraction geometry visible [474]; the first Brillouin zone as the Wigner–Seitz cell of the reciprocal lattice and as the complete set of inequivalent wave vectors [644]; Bloch's theorem stated here for lattice waves and reused for electrons in Chapter 126; zone boundaries as the locus of Bragg reflection.]*

125.1.4 Quasicrystals: the exception that fixed the rule

[Reserved: the observation of sharp diffraction spots with icosahedral symmetry in a rapidly cooled aluminium–manganese alloy, which is forbidden for any periodic lattice [1744]; the resolution — long-range order without periodicity — and the consequent redefinition of "crystal" by diffraction rather than by translational symmetry; what survives of the phonon description, and the additional phason modes that do not.]

Phenomenon 125.1 (Sharp diffraction with a forbidden symmetry)**.** A rapidly solidified aluminium–manganese alloy gives an electron diffraction pattern with spots as sharp as those of a good crystal, and with icosahedral point symmetry — including five-fold axes, which the crystallographic restriction forbids to any lattice periodic in three dimensions [1744]. The two observations are individually routine and jointly contradictory under the definition of a crystal then in force. The resolution is that sharp diffraction requires long-range *order* and not periodicity, and the definition was changed accordingly.

Derivation pending #237

The crystallographic restriction, proved from the integrality of the matrix of a lattice rotation in the lattice basis and hence of its trace, which admits only two-, three-, four- and six-fold axes; and the demonstration that a non-periodic but long-range-ordered point set can nevertheless have a Fourier transform consisting of sharp discrete peaks, so that sharp diffraction does not entail a periodic lattice

125.2 Experiment: the lattice made visible

125.2.1 Laue diffraction of X-rays

[Reserved: the 1912 Munich experiment in which a collimated X-ray beam through a copper sulphate crystal produced a symmetric pattern of spots on a photographic plate [385], settling two questions at once — that X-rays are waves, and that crystals are periodic on the $1 \times 10^{-10}\,\mathrm{m}$ *scale; von Laue's diffraction conditions and their equivalence to the reciprocal-lattice condition [389]; the apparatus, exposure times and the control observations.]*

125.2.2 Bragg's law and structure determination

[Reserved: the reformulation of the Laue conditions as reflection from lattice planes, $n\lambda = 2d\sin\theta$ *[399]; the ionization spectrometer and the first structures solved, sodium chloride and diamond, which showed that the rock-salt crystal contains no NaCl molecule; measured lattice constants in SI with their modern uncertainties, and the silicon lattice parameter as a route to the Avogadro constant (Chapter 2).]*

Phenomenon 125.2 (Crystals diffract X-rays into discrete spots)**.** A collimated X-ray beam passed through a single crystal emerges not as a diffuse halo but as a pattern of sharp, discrete spots arranged with the point symmetry of the crystal [385]. The pattern is the joint signature of two facts, neither previously established: that X-rays are waves, and that a crystal is a lattice periodic on the scale of $1 \times 10^{-10}\,\mathrm{m}$. Read as reflection from families of parallel lattice planes of spacing d, the observed directions obey

$$n\lambda = 2d\sin\theta\,, \qquad n \in \mathbb{Z}\,, \tag{125.1}$$

with θ the glancing angle measured from the planes and not from their normal [399]. Equation (125.1) turns the diffraction pattern into a measuring instrument: it converts measured angles into lattice spacings, and it is what showed that the rock-salt crystal contains no NaCl molecule at all, only alternating ions.

Derivation. Consider a family of parallel lattice planes of spacing d and a plane wave of wavelength λ incident at glancing angle θ. Two rays reflected specularly from adjacent planes travel paths differing by the two segments $d\sin\theta$ dropped from the upper plane to the lower and back, that is by $2d\sin\theta$. Specular reflection makes the phase difference between rays scattered within one plane vanish identically, so the whole family adds in phase if and only if the inter-plane path difference is a whole number of wavelengths, which is Equation (125.1). Two features of the observation follow at once. The spots are *discrete* because the condition selects isolated angles for each d; a non-periodic scatterer has no family of equally spaced planes and returns a continuous distribution. And Equation (125.1) has no solution unless $\lambda \le 2d$, which is why visible light shows no such pattern and why the measurement had to wait for X-rays. The plane picture is the crystallographers' form of the von Laue conditions; both say that the scattering vector must be a reciprocal-lattice vector (Section 125.1.3). □

125.2.3 Structure factor and systematic absences

[Reserved: the geometric structure factor as the Fourier transform of the basis, and the atomic form factor as that of the electron density; systematic absences as the fingerprint of centring and of screw axes, and how a space group is read off a diffraction pattern; the phase problem and why intensities alone do not determine a structure; the Fourier machinery of Chapter 17.]

125.3 Heat capacity of solids

125.3.1 The Dulong–Petit law and its failure

[Reserved: the empirical constancy of the molar heat capacity of the solid elements at $3R = 24.94\,\mathrm{J/(mol\,K)}$ [64], and its derivation from equipartition over $3N$ classical oscillators (Chapter 118); the anomalies known by 1900 — diamond, boron and silicon far below the value at room temperature — and the low-temperature calorimetry showing every solid's heat capacity falling to zero [378], which no classical theory can produce.]

Phenomenon 125.3 (The Dulong–Petit value, and its collapse)**.** Near room temperature the molar heat capacity of the solid elements is very nearly the same for all of them, close to $3R = 24.94\,\mathrm{J/(mol\,K)}$, irrespective of the mass of the atom, the stiffness of the bonding or the crystal structure [64]. The regularity fails in two directions. Already at room temperature diamond, boron and silicon sit far below the value; and calorimetry at low temperature shows the heat capacity of *every* solid falling towards zero as the temperature does [378]. A universal constant that is right at high temperature and wrong at low temperature is the signature of a classical limit.

Derivation. A crystal of N atoms has $3N$ vibrational degrees of freedom, and near equilibrium each is a harmonic normal mode (Section 125.4.1). A classical harmonic oscillator in equilibrium at temperature T has, by equipartition, mean kinetic energy $\frac{1}{2}k_BT$ and mean potential energy $\frac{1}{2}k_BT$, so $\langle E\rangle = k_BT$ per mode, independent of its frequency. Hence

$$U = 3Nk_BT\,, \qquad C = \left(\frac{\partial U}{\partial T}\right)_V = 3Nk_B\,, \tag{125.2}$$

which for one mole is $3R$. Every material property has cancelled: the masses and the force constants enter only through the frequencies, and the frequencies have dropped out. That is precisely why the result cannot be repaired. A classical theory containing no energy scale of its own has nothing with which to compare k_BT, so it cannot produce a temperature below which the heat capacity falls. The observed collapse therefore requires a scale that the classical theory does not possess, and the only candidate is $\hbar\omega$: modes with $\hbar\omega \gg k_BT$ are frozen out, which is why the light, stiffly bound solids — diamond, boron, beryllium — are the ones that deviate first (Section 125.3.2). □

125.3.2 The Einstein model

[Reserved: the first application of quantization to matter rather than radiation — $3N$ independent oscillators of a single frequency, with $x = \theta_E/T$ giving $C = 3Nk_Bx^2\mathrm{e}^x/(\mathrm{e}^x-1)^2$ [333]; the quantitative account of the diamond anomaly with one fitted frequency; the exponential low-temperature falloff, which is too fast, and why: the model has no low-frequency modes; the historical weight of this paper as evidence for quanta outside Chapter 68.]

125.3.3 The Debye model and the cubic law

[Reserved: replacement of the single frequency by an elastic continuum with a linear dispersion cut off at ω_D so that the mode count is $3N$ [382]; the interpolation formula and the limiting law $C = \frac{12\pi^4}{5}Nk_B(T/\theta_D)^3$; the Debye temperature as the single material parameter, from 105 K for lead to 2230 K for diamond; the exact parallel with the photon gas of Chapter 68, differing only in the cutoff and the three polarizations.]

125.3.4 Calorimetric tests at low temperature

[Reserved: the measurement that separates the two contributions — C/T plotted against T^2 gives a straight line whose intercept is the electronic term γT and whose slope is the lattice T^3, verified for copper, silver and gold between 1 K and 5 K [1061]; the extracted Debye temperatures compared with those from elastic constants, a consistency test between two unrelated experiments; where the model fails at intermediate T, and why.]

Phenomenon 125.4 (The cubic law)**.** Well below its Debye temperature the lattice contribution to the heat capacity of a solid is not exponentially small but follows a power law,

$$C_{\text{lattice}} = \frac{12\pi^4}{5}\,Nk_B\left(\frac{T}{\theta_D}\right)^3\,, \tag{125.3}$$

with a single material parameter θ_D ranging from about 105 K for lead to about 2230 K for diamond [382]. In a metal the lattice term is accompanied by the electronic term of Chapter 120, and the two are separated experimentally by plotting C/T against T^2: the result is a straight line whose intercept is electronic and whose slope is the lattice term, as verified for copper, silver and gold between 1 K and 5 K [1061]. The θ_D so obtained agrees with the value computed from independently measured elastic constants, which is a consistency test between two unrelated experiments and not a fit.

Derivation. Debye's approximation replaces the true dispersion by an isotropic elastic continuum, $\omega = v_s k$ on three polarizations, and truncates the spectrum at a frequency ω_D chosen so that the total number of modes is $3N$. Counting modes in a box, the number with wavevector below k is $3 \cdot V k^3/6\pi^2$, so the density of modes in frequency is $g(\omega) = 3V\omega^2/2\pi^2 v_s^3$, and imposing $\int_0^{\omega_D} g\,\mathrm{d}\omega = 3N$ lets it be written without reference to v_s as

$$g(\omega) = \frac{9N\omega^2}{\omega_D^3}\,, \qquad 0 \le \omega \le \omega_D\,. \tag{125.4}$$

Each mode is a quantized oscillator whose thermal occupation is the Bose function of Chapter 120 with zero chemical potential, since phonon number is not conserved. Dropping the temperature-independent zero-point term,

$$U = \int_0^{\omega_D} \frac{\hbar\omega\, g(\omega)\,\mathrm{d}\omega}{\mathrm{e}^{\hbar\omega/k_B T} - 1} = 9Nk_BT\left(\frac{T}{\theta_D}\right)^3 \int_0^{\theta_D/T} \frac{x^3\,\mathrm{d}x}{\mathrm{e}^x - 1}\,, \tag{125.5}$$

with $x = \hbar\omega/k_BT$ and $k_B\theta_D = \hbar\omega_D$. For $T \ll \theta_D$ the upper limit may be sent to infinity with exponentially small error, and $\int_0^\infty x^3(\mathrm{e}^x-1)^{-1}\mathrm{d}x = \pi^4/15$, so $U = (3\pi^4/5)Nk_BT^4/\theta_D^3$ and differentiation gives Equation (125.3). The physical content is in the low-frequency end of Equation (125.4): because a crystal has acoustic modes of arbitrarily low frequency, there are always modes with $\hbar\omega < k_BT$ available to excite, and the number of them scales as T^3 in three dimensions. This is exactly the photon-gas calculation of Chapter 68 with two changes, three polarizations instead of two and a finite cutoff instead of none; the cutoff is what the photon gas lacks and is why radiation has no Dulong–Petit regime. Einstein's earlier model, which gives every mode the same frequency, has no low-frequency modes at all and therefore predicts an exponential falloff, too fast to match the data [333]. □

125.4 Lattice dynamics

125.4.1 The Born–von Kármán theory

[Reserved: the harmonic expansion of the crystal potential in atomic displacements, the dynamical matrix $D(\boldsymbol{k})$ and the eigenvalue problem $\omega^2(\boldsymbol{k})\boldsymbol{e} = D(\boldsymbol{k})\boldsymbol{e}$ [381]; periodic boundary conditions and the resulting discrete, finite set of allowed wave vectors in the Brillouin zone; the monatomic chain and its dispersion $\omega = 2\sqrt{K/M}\,|\sin(ka/2)|$ as the worked case; the long-wavelength limit reproducing the elastic waves of Chapter 30; the monograph [1030].]

125.4.2 Acoustic and optical branches

[Reserved: the diatomic chain, whose $3p$ branches for p atoms per cell split into three acoustic branches with $\omega \to 0$ as $k \to 0$ and $3p-3$ optical branches with a finite gap [381]; the physical distinction — neighbouring atoms move in phase in an acoustic mode and against each other in an optical one; why only optical modes couple to light at $k \approx 0$; group velocity vanishing at the zone boundary, connecting to Chapter 28.]

125.4.3 Quantization: the phonon

[Reserved: each normal mode quantized as a harmonic oscillator, giving energies $(n_{\boldsymbol{k}s} + \tfrac{1}{2})\hbar\omega_{\boldsymbol{k}s}$ and an occupation obeying the Bose–Einstein distribution of Chapter 120 with zero chemical potential, since phonon number is not conserved; crystal momentum $\hbar\boldsymbol{k}$ defined only modulo a reciprocal-lattice vector, and why it is not true momentum; zero-point motion and its measured consequences, including the failure of helium to solidify under its own vapour pressure (Chapter 130).]

125.4.4 Polar crystals and the Lyddane–Sachs–Teller relation

[Reserved: the long-range Coulomb field of a longitudinal optical mode in an ionic crystal, splitting it from the transverse mode, and the relation $\omega_{LO}^2/\omega_{TO}^2 = \epsilon(0)/\epsilon(\infty)$ [859]; the reststrahlen band of total reflection between the two frequencies as its direct observation; polaritons as the mixed phonon–photon excitations, and their measured dispersion; the link to dielectric response in Chapter 62.]

125.5 Measured dispersion relations

125.5.1 Inelastic neutron scattering

[Reserved: why the thermal neutron is the right probe — its de Broglie wavelength matches interatomic spacings while its energy, about 25 meV, matches phonon energies, so energy and momentum transfer are measured simultaneously; the triple-axis spectrometer and the first phonon measurement, in a single crystal of aluminium [1056]; the complete branch-by-branch dispersion of germanium [1137], compared with Born–von Kármán force-constant fits; conservation laws with crystal momentum, including umklapp processes.]

Phenomenon 125.5 (The dispersion relation is measured, not assumed)**.** A neutron scattered by a crystal exchanges a definite energy and a definite wavevector with it, and the pairs so measured do not fill the plane: they lie on curves. Scanning energy transfer against wavevector transfer in a single crystal of aluminium [1056] and, branch by branch through the Brillouin zone, in germanium [1137], one measures $\omega(\boldsymbol{k})$ directly. Three features of the measured curves are the substance of lattice dynamics. The dispersion is linear only at long wavelength; it flattens towards the zone boundary, where the group velocity vanishes, so a crystal is not an elastic continuum. It is periodic in the reciprocal lattice, so $\boldsymbol{k}$ is defined only modulo a reciprocal-lattice vector. And a crystal with p atoms in its primitive cell shows $3p$ branches, of which three are acoustic, vanishing at $\boldsymbol{k} = 0$, and the remaining $3p-3$ are optical, with a finite frequency there. The existence of a sharp ω for each $\boldsymbol{k}$ is what entitles one to call the excitation a particle.

Derivation. The simplest case exhibits every feature. Take a chain of identical masses M at spacing a, coupled to nearest neighbours by springs of constant K; let u_n be the displacement of the n-th mass. In the harmonic approximation

$$M\ddot{u}_n = K\left(u_{n+1} + u_{n-1} - 2u_n\right) . \tag{125.6}$$

The equations are invariant under $n \to n+1$, so their normal modes may be labelled by the eigenvalue of that translation: try $u_n = A\,\mathrm{e}^{\mathrm{i}(kna-\omega t)}$. Substituting into Equation (125.6) gives $-M\omega^2 = K(\mathrm{e}^{\mathrm{i}ka} + \mathrm{e}^{-\mathrm{i}ka} - 2) = -4K\sin^2(ka/2)$, that is

$$\omega(k) = 2\sqrt{\frac{K}{M}}\left|\sin\frac{ka}{2}\right| . \tag{125.7}$$

Equation (125.7) is periodic under $k \to k + 2\pi/a$, so nothing outside the first Brillouin zone $|k| \le \pi/a$ is new: a wave shorter than twice the spacing is indistinguishable from a longer one, because only the discrete u_n exist. For $ka \ll 1$ it reduces to $\omega \approx a\sqrt{K/M}\,|k|$, the acoustic branch, whose slope is the sound velocity of Chapter 30. At the zone boundary $\mathrm{d}\omega/\mathrm{d}k = 0$: the group velocity vanishes and the mode is a standing wave, which is the lattice-dynamical counterpart of Bragg reflection at the same wavevector. Repeating the calculation with two different masses per cell splits the single branch into an acoustic one that still vanishes at $k = 0$, in which neighbours move together, and an optical one with a finite gap there, in which they move against each other [381]; the general statement of

$3p$ branches follows from diagonalizing the $3p \times 3p$ dynamical matrix. Quantizing each mode as an oscillator makes $\hbar\omega(\boldsymbol{k})$ the energy of one phonon, which is what the neutron measures. □

125.5.2 Inelastic X-ray and light scattering

[Reserved: Raman scattering as the optical route to zone-centre optical phonons, from its discovery in liquids [618] to its routine crystallographic use; Brillouin scattering for acoustic phonons and the sound velocities it yields; high-resolution inelastic X-ray scattering at synchrotron sources, which reaches phonons in samples too small or too absorbing for neutrons [1985]; the complementary selection rules of the three techniques.]

125.5.3 The Debye–Waller factor

[Reserved: the attenuation e^{-2W} of Bragg intensities by thermal displacement, with $2W = \langle(\boldsymbol{Q}\cdot\boldsymbol{u})^2\rangle$, derived by Debye [402] and corrected for the quantum zero-point term by Waller [506]; the crucial fact that thermal motion weakens the spots but does not broaden them, the intensity moving into a diffuse background; recoilless nuclear resonance absorption as the same factor in the gamma-ray regime [1154], and the mean square displacements it measures.]

125.6 Anharmonicity

125.6.1 Thermal expansion and the Grüneisen parameter

[Reserved: the theorem that a strictly harmonic crystal does not expand, so thermal expansion measures the cubic term directly; the Grüneisen parameter $\gamma_G = -\mathrm{d}\ln\omega/\mathrm{d}\ln V$ and the relation $\alpha = \gamma_G\kappa_T C_V/V$ [386]; the observed near-constancy of $\gamma_G \approx 2$ across many solids and the parallel temperature dependence of α and C_V; negative thermal expansion in silicon and in framework structures as the instructive exception.]

Phenomenon 125.6 (Solids expand when heated, and do so in step with their heat capacity)**.** The volume thermal expansion coefficient α of a solid is not an independent material function of temperature: measured over a wide range it is proportional to the heat capacity C_V of the same sample, falling towards zero at low temperature in the same way and saturating at the same temperature, so that the dimensionless ratio

$$\gamma_G = \frac{\alpha V}{\kappa_T C_V} \tag{125.8}$$

is nearly independent of temperature and takes a value of order two across a wide range of solids [386]. A strictly harmonic crystal would not expand at all — the mean position of a harmonic oscillator is independent of its energy — so the whole of the effect measures the anharmonic part of the interatomic potential, and Equation (125.8) says that one number suffices to describe it. The exceptions are as instructive as the rule: silicon and several open framework structures contract on heating over some temperature ranges.

Derivation pending #238

The Grüneisen relation: that a phonon frequency depends on volume, quantified by the logarithmic derivative of frequency with respect to volume; the resulting thermal pressure of the phonon gas; and the equality of that pressure gradient with the elastic restoring stress, which yields the expansion coefficient as the product of the compressibility, the heat capacity and the mean logarithmic derivative, together with the sign reversal produced by transverse modes that stiffen under expansion

125.6.2 Phonon–phonon scattering and umklapp

[Reserved: three-phonon processes as the leading anharmonic interaction; Peierls's decisive observation that normal processes conserve total crystal momentum and therefore cannot degrade a heat current, so only umklapp processes — in which the summed wave vector lands outside the Brillouin zone and is returned by a reciprocal-lattice vector — give a finite thermal resistance [634]; the resulting $\mathrm{e}^{\theta_D/bT}$ freezing-out of umklapp at low temperature; phonon lifetimes measured as neutron linewidths.]

125.6.3 Thermal conductivity of insulators

[Reserved: the kinetic-theory expression $\kappa = \frac{1}{3} C v \ell$ and the characteristic peak in $\kappa(T)$ where umklapp scattering gives way to boundary scattering [1196]; isotope scattering as an independently controllable term, demonstrated by the roughly 50 % rise in the conductivity of diamond on isotopic enrichment [1830]; measured conductivities in $\mathrm{W/(m\,K)}$ for representative insulators; the transport formalism of Chapter 124.]

Phenomenon 125.7 (Thermal conductivity of an insulating crystal)**.** A pure dielectric crystal conducts heat at a finite rate, and its conductivity is a strongly non-monotonic function of temperature: it rises steeply on cooling, passes through a pronounced maximum at a temperature of order a few per cent of the Debye temperature, and then falls again towards zero. Both flanks are informative. The high-side fall is exponential in θ_D/T; the low-side fall follows the heat capacity, and its magnitude depends on the size and surface of the specimen, which no bulk material property should. That the conductivity is limited by the crystal's own imperfections is shown decisively by isotopes: enriching diamond in carbon-12 raises its room-temperature thermal conductivity by roughly 50 %, without changing its chemistry, its structure or its phonon spectrum to any comparable degree [1830].

Derivation pending #239

Why a harmonic crystal has infinite thermal conductivity, and why anharmonicity alone is not enough: that three-phonon processes conserving the total crystal momentum cannot degrade a heat current, so that a finite thermal resistance requires umklapp processes in which the summed wavevector leaves the Brillouin zone and is returned by a reciprocal-lattice vector; the exponential freezing-out of those processes at low temperature; and the kinetic-theory expression for the conductivity in terms of heat capacity, group velocity and mean free path, whose crossover to boundary-limited and isotope-limited scattering produces the observed maximum

Chapter 126

Electrons in Solids: Band Theory

Why one solid conducts electricity and the next does not is the oldest quantitative question in the physics of matter, and the answer — energy bands — is the most consequential single application of quantum mechanics to anything larger than a molecule. This chapter builds it in the order in which the evidence forced it. Drude modelled a metal as a classical gas of electrons scattering off the ions [289, 290], recovered Ohm's law and the Wiedemann–Franz ratio measured half a century earlier [140], and failed by two orders of magnitude on the electronic specific heat. Sommerfeld replaced Maxwell–Boltzmann by the Fermi–Dirac statistics of Chapter 120 and repaired the thermal properties at a stroke [620]. Bloch then put the electrons into the periodic potential of the lattice of Chapter 125 and proved the theorem [623] from which bands, gaps and Wilson's criterion separating metals from insulators [688, 689] all follow.

The chapter's second half is about evidence, because a band structure is a falsifiable curve $E_n(\boldsymbol{k})$ and not a picture. The Fermi surface is measured by the de Haas–van Alphen oscillations [655] read through Onsager's relation [1008], by cyclotron resonance [1063], and today directly by angle-resolved photoemission [2048]. The chapter closes honestly on where the one-electron picture breaks: on the Mott insulators that band theory declares to be metals [941], and on disorder [1134]. Its neighbours are the lattice dynamics of Chapter 125 before it and, after it, the doped band insulators of Chapter 127, the quantized magnetotransport of Chapter 128, the ordered moments of Chapter 129 and the paired condensate of Chapter 130. The standard modern treatment is [1570].

Derivation pending #240

Electrons in Solids: Band Theory: all derivations of this chapter are pending.

126.1 The free-electron metal

126.1.1 Drude's kinetic model

[Reserved: the metal as a classical gas of conduction electrons scattering off fixed ions, introduced by Drude [289]; the relaxation-time equation of motion, the direct-current conductivity $\sigma = ne^2\tau/m$ recovering Ohm's law, the frequency-dependent Drude response and the plasma frequency; the galvanomagnetic and thermomagnetic sequel [290], in which the same model delivers the Hall coefficient $R_H = -1/ne$ for the effect Hall had discovered [208]; relaxation times of order 1×10^{-14} s at room temperature, inferred rather than assumed; the kinetic-theory machinery imported from Chapter 55.]

Phenomenon 126.1 (Ohm's law and the Hall coefficient)**.** Two facts about metallic conduction are simple enough to look uninformative and are not. First, the current density

in a metal is strictly proportional to the applied electric field over many decades of field strength, with a conductivity that depends on temperature and purity but not on the field: a metal has no memory of how hard it is being driven. Second, a current j_x crossed with a magnetic field B along z develops a steady transverse electric field E_y, in the ratio

$$R_H = \frac{E_y}{j_x B}\,, \tag{126.1}$$

which Hall found to be independent of the field, of the current and of the sample dimensions [208]. For the alkali metals R_H is negative and its magnitude corresponds to about one mobile carrier of charge $-e$ per atom [1570]. That a solid should hand back its carrier density to a voltmeter is the first quantitative measurement of the electron gas.

Derivation. Drude's model treats the conduction electrons as free particles of charge $-e$ and mass m, accelerated by the applied fields and randomized by collisions at a mean rate $1/\tau$, so that on average the collisions remove momentum at the rate $m\boldsymbol{v}/\tau$ [289]. The steady-state equation of motion is

$$0 = -e\,(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}) - \frac{m\boldsymbol{v}}{\tau}\,. \tag{126.2}$$

With $\boldsymbol{B} = 0$ this gives $\boldsymbol{v} = -e\tau\boldsymbol{E}/m$ and hence

$$\boldsymbol{j} = -ne\boldsymbol{v} = \frac{ne^2\tau}{m}\boldsymbol{E} \equiv \sigma\boldsymbol{E}\,, \tag{126.3}$$

which is Ohm's law: the response is linear because the drag term in Equation (126.2) is linear, and the conductivity contains no reference to $\boldsymbol{E}$. Now take $\boldsymbol{B} = B\hat{\boldsymbol{z}}$ and a sample whose sides are insulating, so that in the steady state $v_y = 0$. The y component of Equation (126.2) then reads $-e(E_y - v_x B) = 0$, so $E_y = v_x B$; and $j_x = -nev_x$ gives $v_x = -j_x/ne$. Substituting into Equation (126.1),

$$R_H = -\frac{1}{ne}\,, \tag{126.4}$$

independent of τ, of m and of the field — which is why the measurement is a carrier count and not a fit. Two things should be noted about what has *not* been derived. The sign in Equation (126.4) is fixed by the assumption that the carriers are electrons, and it is contradicted in several metals (Section 126.1.3); and τ has been postulated, not computed, so Equation (126.3) explains the linearity of Ohm's law but not the magnitude or the temperature dependence of σ (Section 126.6). □

126.1.2 The Wiedemann–Franz law

[Reserved: the empirical proportionality of thermal to electrical conductivity at fixed temperature, measured across the metals by Wiedemann and Franz [140]; the Lorenz ratio $L = \kappa/\sigma T$ and its near-universal value close to $2.44 \times 10^{-8}\,\mathrm{W\,\Omega/K^2}$; why Drude obtained it almost exactly right through two compensating errors of order one hundred, in the electronic heat capacity and in the mean square speed [289, 1570]; the regimes in which the law is observed to fail, and what inelastic scattering has to do with it.]

Phenomenon 126.2 (The Wiedemann–Franz law)**.** At ordinary temperatures the thermal conductivity κ and the electrical conductivity σ of a metal are proportional, and the constant of proportionality is very nearly the same for every metal: the Lorenz ratio

$$L = \frac{\kappa}{\sigma T} \tag{126.5}$$

clusters about $2.44 \times 10^{-8}\,\mathrm{W\,\Omega/K^2}$ across metals differing by orders of magnitude in σ itself [140, 1570]. A ratio that survives when its two members do not is a strong statement: the same carriers, scattered by the same mechanism, must be responsible for both currents. The law fails, in a way that is itself diagnostic, at intermediate temperatures, where the scattering that degrades a heat current and the scattering that degrades a charge current cease to be the same.

Derivation. Kinetic theory gives the thermal conductivity of a gas of carriers as $\kappa = \tfrac{1}{3} c_V v^2 \tau$, with c_V the heat capacity per unit volume of the carriers and v their typical speed. For a degenerate electron gas both quantities are those of the Fermi surface, not of a classical gas: from Chapter 120, $c_V = \tfrac{1}{2}\pi^2 n k_B (T/T_F)$, and the carriers move at the Fermi speed, $\tfrac{1}{2} m v_F^2 = k_B T_F$. Hence

$$\kappa = \frac{1}{3} \cdot \frac{\pi^2 n k_B T}{2 T_F} \cdot \frac{2 k_B T_F}{m} \cdot \tau = \frac{\pi^2}{3} \frac{n k_B^2 T \tau}{m} \,. \tag{126.6}$$

Dividing by σT with $\sigma = n e^2 \tau / m$ from Equation (126.3), the carrier density, the mass and — the essential point — the relaxation time all cancel:

$$L = \frac{\pi^2}{3} \left(\frac{k_B}{e} \right)^2 = 2.44 \times 10^{-8}\,\mathrm{W\,\Omega/K^2} \,, \tag{126.7}$$

a combination of fundamental constants alone, which is why it is the same for every metal. The cancellation of τ is what makes Equation (126.7) robust, and its failure is what makes the observed deviations informative: the derivation assumed one relaxation time for both currents, whereas a small-angle inelastic collision with a phonon can relax an energy current while barely relaxing a charge current. Drude obtained almost this number classically because his two errors were reciprocal: the classical heat capacity is too large by about $T_F/T \sim 10^2$ and the classical mean square speed too small by about the same factor, so their product survived while neither factor did [289, 1570]. □

126.1.3 What the classical electron gas cannot do

[Reserved: the missing electronic specific heat — classically $\tfrac{3}{2} k_B$ per electron, observed near room temperature to be smaller by about two orders of magnitude; the sign of the Hall coefficient, positive in aluminium, zinc and beryllium where [290] demands it be negative; the complete absence of any account of why a solid should be an insulator at all; each of these a symptom of applying Maxwell–Boltzmann statistics to a degenerate fermion gas, and each resolved in Chapter 120.]

Phenomenon 126.3 (Hall coefficients of the wrong sign)**.** In aluminium, zinc, beryllium and cadmium the measured Hall coefficient Equation (126.1) is *positive*, and in several metals its magnitude does not correspond to any plausible carrier density [208, 1570]. Equation (126.4) admits no such value: within the free-electron picture the sign of R_H is the sign of the carrier charge, and the carriers are electrons. The observation is therefore not a quantitative discrepancy but a qualitative refutation, and the fact that it occurs in ordinary, well-characterized, chemically simple metals rules out an explanation by impurities or by a second species of ion.

Derivation pending #241

The dynamics of a nearly filled band described through its vacancies: that the total current of a filled band vanishes, so that the current of an almost-filled band equals minus the current of the missing states; that a wave packet near a band maximum has negative band curvature and therefore responds as a particle of positive charge and positive mass; and the resulting positive Hall coefficient, with the multi-band and open-orbit corrections that govern its magnitude

126.1.4 The Sommerfeld model

[Reserved: Sommerfeld's substitution of the Fermi–Dirac distribution into Drude's kinetic picture [620]; the Fermi sphere, the Fermi energy $E_F = \hbar^2 k_F^2/2m$ of a few eV *and the Fermi temperature of order 1×10^4* K *that makes every metal a degenerate gas at room temperature; the Sommerfeld expansion, the linear electronic heat capacity $c = \gamma T$ and the measured coefficients γ of the alkali metals; Pauli paramagnetism [586] and Landau diamagnetism [660] as the magnetic counterparts, taken up in Chapter 129; exactly what of Drude survives.]*

126.2 Electrons in a periodic potential

126.2.1 Bloch's theorem

[Reserved: the theorem that eigenfunctions of a Hamiltonian invariant under the lattice translations of Chapter 125 take the form $\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_{n\boldsymbol{k}}(\boldsymbol{r})$ with $u_{n\boldsymbol{k}}$ lattice periodic, proved by Bloch [623]; crystal momentum $\hbar\boldsymbol{k}$ as a label of an irreducible representation of the discrete translation group and not a true momentum; the band index n; Born–von Karman boundary conditions and the counting of allowed $\boldsymbol{k}$ that gives N states per band per spin; the group-theoretic reading via Chapter 5.]

126.2.2 Reciprocal lattice and Brillouin zones

[Reserved: the reciprocal lattice and the equivalence $\boldsymbol{k} \sim \boldsymbol{k} + \boldsymbol{G}$ that makes every band a periodic function on a torus; the first Brillouin zone as the Wigner–Seitz cell of the reciprocal lattice, introduced together with the reading of zone boundaries as Bragg planes by Brillouin [644]; the consequent link between a band gap and the diffraction condition measured in Chapter 125; reduced, extended and repeated zone schemes; high-symmetry points and the conventional paths along which band structures are plotted.]

126.2.3 Nearly free electrons

[Reserved: degenerate perturbation theory (Chapter 82) applied to a weak periodic potential; the gap $2\,|V_{\boldsymbol{G}}|$ that opens at each Bragg plane, the distortion of the free-electron parabola and the origin of the very idea of a forbidden band [623, 1570]; why the weak-potential limit describes real sp-bonded metals far better than the strength of the ionic potential suggests, which is the pseudopotential argument of Section 126.3.3 [1172].]

126.2.4 Tight binding and Wannier functions

[Reserved: the opposite limit of weakly overlapping atomic orbitals; the band $E(\boldsymbol{k}) = \epsilon - 2t\sum_i \cos(k_i a)$ of a simple cubic lattice, with bandwidth set by the hopping integral t and centre by the on-site energy ϵ; the Slater–Koster parametrization of the two-centre integrals that made the method practical [1051]; Wannier functions as the localized unitary dual of

the Bloch states [805]; why two limits that could hardly differ more agree on the qualitative band picture.]

126.2.5 The Kronig–Penney model

[Reserved: the exactly solvable periodic square-well chain of Kronig and Penney [683]; the transfer-matrix condition $\cos(ka) = \cos(\alpha a) + P\sin(\alpha a)/(\alpha a)$*, whose failure to admit a real* k *over finite energy intervals is a band gap in closed form; the delta-comb limit and the width of the bands as a function of barrier strength; what the model proves — that periodicity alone, and no special property of atoms, produces gaps.]*

126.3 Band structure and the classification of solids

126.3.1 Filling and Wilson's criterion

[Reserved: the count of $2N$ *states per band including spin, and the resulting rule that a crystal with an odd number of electrons per primitive cell must be a metal; Wilson's identification of an insulator as a solid whose bands are exactly filled, and of a semiconductor as an insulator whose gap is small enough to be bridged thermally [688, 689]; the divalent metals, saved by band overlap; the criterion's known counterexamples, deferred honestly to Section 126.7.2.]*

Phenomenon 126.4 (Metals and insulators are qualitatively, not quantitatively, different)**.** The room-temperature electrical resistivity of solids spans more than twenty orders of magnitude, from about $1 \times 10^{-8}\,\Omega\,\mathrm{m}$ in a good metal to above $1 \times 10^{16}\,\Omega\,\mathrm{m}$ in a good insulator, with no bunching in between that would mark a natural boundary. What does mark one is the *sign* of the temperature dependence: on cooling, the resistivity of a metal falls and tends to a finite residual value, while that of an insulator or semiconductor rises without limit. The same distinction appears in the low-temperature heat capacity, which carries the electronic term γT of Chapter 120 in a metal and does not in an insulator. Empirically the class a crystal belongs to correlates with the number of electrons in its primitive cell rather than with any property of the individual atoms: an odd number of electrons per cell has never yielded an insulator [688, 689].

Derivation pending #242

Bloch's theorem for a periodic potential and the counting that follows from it: that the allowed wavevectors in one Brillouin zone number the primitive cells of the crystal, so that each band holds exactly two electrons per cell; that a crystal whose bands are either completely full or completely empty has no states available at infinitesimal excitation energy and therefore carries no current at zero temperature; and the consequent identification of a semiconductor as an insulator whose gap is comparable with the thermal energy

126.3.2 Semimetals, semiconductors and the gap scale

[Reserved: the measured gaps that set the scale of all electronics — about $1.12\,\mathrm{eV}$ *in silicon,* $0.67\,\mathrm{eV}$ *in germanium,* $1.42\,\mathrm{eV}$ *in gallium arsenide, against about* $5.5\,\mathrm{eV}$ *in diamond [2061]; direct against indirect gaps and the phonon that must be supplied in the second case; the semimetals bismuth, antimony and graphite, with band overlaps of* meV *and carrier densities four orders below a normal metal; handover to Chapter 127.]*

126.3.3 Computing band structures

[Reserved: the cellular method and the first quantitative band calculation for a real solid, metallic sodium, by Wigner and Seitz [728]; orthogonalized plane waves [851]; the pseudopotential construction that explains why nearly-free-electron bands work [1172]; the augmented-plane-wave and $k \cdot p$ families [1570]; what "ab initio" does and does not mean here, and the discipline that every computed band is judged against the measurements of Section 126.5.]

126.3.4 Density functional theory

[Reserved: the Hohenberg–Kohn theorems — the ground-state density fixes the external potential up to a constant, and the energy is a variational functional of that density [1284]; the Kohn–Sham construction trading the interacting problem for non-interacting particles in a self-consistent effective potential [1314]; the local-density and generalized-gradient approximations [1927]; and the honest caveats — that Kohn–Sham eigenvalues are not excitation energies, that computed gaps come out low by tens of percent, and that the many-body GW self-energy [1310] is what brings them into agreement with the photoemission of Section 126.5.4.]

126.4 Semiclassical dynamics of Bloch electrons

126.4.1 Group velocity and effective mass

[Reserved: the wave-packet velocity $\boldsymbol{v}_n = \hbar^{-1}\nabla_{\boldsymbol{k}} E_n(\boldsymbol{k})$ and the semiclassical force law $\hbar\dot{\boldsymbol{k}} = -e(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$, in which the lattice force is absorbed entirely into the band; the inverse effective-mass tensor $(m^*)^{-1}_{ij} = \hbar^{-2}\partial^2 E/\partial k_i \partial k_j$, introduced with the theory of galvanomagnetic effects by Peierls [635]; measured values spanning $0.067\,m_e$ in gallium arsenide to tens of m_e in heavy-fermion metals; the interband conditions under which the semiclassical description is legitimate.]

126.4.2 Holes

[Reserved: an almost-full band described by its vacancies; the hole's positive charge, positive effective mass near a band maximum, and hence the positive Hall coefficient observed in aluminium, zinc and beryllium — precisely the anomaly of Section 126.1.3; Peierls' introduction of the concept [635]; holes as the majority carriers of p-type material in Chapter 127, and the care needed not to confuse this collective vacancy with the antiparticle of Chapter 93.]

126.4.3 Bloch oscillations and Zener breakdown

[Reserved: a static field sweeps $\boldsymbol{k}$ uniformly across the zone, so a perfect crystal should oscillate at $eEa/\hbar$ rather than conduct [623]; why scattering destroys the effect in bulk metals and why it is nevertheless observed in semiconductor superlattices, whose small zone raises the frequency above the scattering rate [1853]; interband tunnelling at large field, computed by Zener [750], which returns as the tunnel diode of Chapter 127.]

126.4.4 Magnetotransport and the Hall coefficient

[Reserved: Hall's discovery that a field transverse to a current produces a transverse voltage [208]; the free-electron result $R_H = -1/ne$ and its everyday use as a carrier-density measurement; magnetoresistance, closed and open orbits, and the saturation criterion that reads Fermi-surface topology directly off a magnetoresistance curve [1570]; the same geometry

taken to two dimensions, high field and low temperature becomes the quantized effect of Chapter 128.]

126.5 The Fermi surface and how it is measured

126.5.1 Landau levels and quantum oscillations

[Reserved: the quantization of cyclotron orbits into Landau levels [660], with degeneracy eB/h per unit area, and the passage of those levels through the Fermi energy periodically in $1/B$; the Lifshitz–Kosevich amplitude formula with its temperature and Dingle damping factors [1092]; why oscillations demand $\omega_c\tau \gg 1$, hence high-purity crystals, fields of order $10\,\mathrm{T}$ and temperatures of order $1\,\mathrm{K}$.]

126.5.2 The de Haas–van Alphen effect

[Reserved: the oscillation of the magnetic susceptibility of bismuth periodic in $1/B$, discovered by de Haas and van Alphen [655]; Onsager's relation $\Delta(1/B) = 2\pi e/\hbar A_{\mathrm{ext}}$, which ties the observed period to an extremal cross-sectional area of the Fermi surface and thereby converted a curiosity into the primary instrument of the field [1008]; the reconstructed Fermi surfaces of copper, gold and aluminium, with their necks touching the zone boundary [1745]; Pippard's independent determination for copper from the anomalous skin effect [1126].]

Phenomenon 126.5 (Quantum oscillations measure the Fermi surface)**.** The magnetic susceptibility of a pure crystal of bismuth at low temperature does not vary smoothly with the applied field but oscillates, and the oscillations are periodic not in B but in $1/B$ [655]. The same periodicity appears in the resistivity [670] and, as instrumentation improved, in every clean metal examined. The observed frequencies are properties of the material alone; there may be several of them, and they change in a characteristic way as the crystal is rotated in the field. Read through Onsager's relation [1008], each frequency measures an extremal cross-sectional area of the Fermi surface normal to the field, which converts the effect into a tomograph: the Fermi surfaces of copper, gold and aluminium, complete with the necks by which they touch the zone boundary, were reconstructed this way [1745] and confirmed independently for copper from the anomalous skin effect [1126]. The requirement is a cyclotron period short compared with the scattering time, hence high purity, fields of order $10\,\mathrm{T}$ and temperatures of order $1\,\mathrm{K}$.

Derivation pending #243

Semiclassical quantization of a closed cyclotron orbit, giving orbits of quantized area in reciprocal space evenly spaced in the magnetic field; the resulting oscillation of the free energy as successive quantized tubes cross the Fermi surface, periodic in the inverse field with a frequency proportional to the extremal cross-sectional area; and the Lifshitz–Kosevich amplitude with its temperature and scattering damping factors

126.5.3 Shubnikov–de Haas and cyclotron resonance

[Reserved: the same quantization seen in the resistivity rather than the susceptibility [670], the transport analogue that returns in Chapter 128; cyclotron resonance, in which microwave power is absorbed at $\omega = eB/m^*$, used by Dresselhaus, Kip and Kittel to measure the effective-mass tensors of silicon and germanium directly [1063] — the measurement that established the many-valley conduction bands underlying Chapter 127.]

126.5.4 Photoemission

[Reserved: angle-resolved photoemission as the only probe that maps $E_n(\boldsymbol{k})$ directly, using the photoelectric effect of Chapter 69 together with energy and in-plane momentum conservation at the surface; what is actually measured — the one-particle spectral function, hence self-energies, lifetimes and mass renormalizations — and the surface sensitivity and final-state uncertainty that limit it [2048]; present resolutions of a few meV *and the direct imaging of Fermi surfaces, kinks and gaps [2391].]*

126.6 Transport and scattering

126.6.1 Residual resistivity and Matthiessen's rule

[Reserved: the striking consequence of Bloch's theorem that a perfect periodic crystal has infinite conductivity, so that all resistance is a defect of periodicity; impurity and defect scattering, the temperature-independent residual resistivity, and the residual resistance ratio as the standard purity figure of merit; the additivity of scattering rates measured by Matthiessen and Vogt in alloys [160] and the conditions under which it fails; the Boltzmann transport equation that derives it, developed in Chapter 124.]

Phenomenon 126.6 (Resistivity separates into an impurity term and a temperature term)**.** The resistivity of a metal is, to a good approximation, a sum of two independent contributions,

$$\rho(T) = \rho_0 + \rho_{\mathrm{i}}(T)\,, \tag{126.8}$$

in which ρ_0 is temperature-independent and depends only on the concentration of impurities and defects, while $\rho_{\mathrm{i}}(T)$ depends only on the temperature and is the same function for every sample of the metal, however pure — the rule Matthiessen and Vogt established on alloy series [160]. Two consequences are measured. As $T \to 0$ the resistivity saturates at ρ_0 instead of vanishing, so the residual resistance ratio is the standard figure of merit for crystal purity; and ρ_{i} is linear in T well above the Debye temperature and falls as a high power of T well below it, following the form Bloch derived and Grüneisen fitted across the metals [642, 720]. The striking part is ρ_0: a perfectly periodic crystal would have none, so every ohm a metal presents is a defect of its periodicity.

Derivation pending #244

That a Bloch state is a stationary state of the perfect periodic potential and therefore propagates without attenuation, so that resistance requires a departure from periodicity; the Boltzmann equation in the relaxation-time approximation and the additivity of scattering rates from independent channels that yields Matthiessen's rule, together with the conditions under which the additivity fails; and the electron–phonon calculation giving the linear high-temperature and the steep low-temperature resistivity

126.6.2 Phonon scattering and the Bloch–Grüneisen law

[Reserved: electron–phonon scattering off the lattice modes of Chapter 125; the linear high-temperature resistivity and the T^5 low-temperature law derived by Bloch [642] and fitted across the metals by Grüneisen [720]; normal against umklapp processes and why only the latter degrade a current [1570]; the electron–phonon coupling constant that reappears as the pairing interaction of Chapter 130; thermopower, the Peltier effect and thermal transport deferred to Chapter 124.]

126.7 Where the one-electron picture fails

126.7.1 Fermi-liquid theory

[Reserved: the question the whole chapter has been begging — why a dense, strongly interacting electron liquid behaves like a gas of independent particles at all; Landau's answer in terms of quasiparticles adiabatically connected to the free states, carrying a renormalized mass and a lifetime diverging as $(E - E_F)^{-2}$ at the Fermi surface [1119]; the experimental signatures — enhanced but still linear specific heat, T^2 resistivity, and the sharp quasiparticle peak seen in Section 126.5.4; the measured metals that violate all three.]

126.7.2 Mott insulators and the Hubbard model

[Reserved: the solids band theory declares to be metals and that are measured to be insulators — nickel oxide, cobalt oxide, the transition-metal monoxides — and Mott's argument that on-site Coulomb repulsion rather than band filling localizes their electrons [941, 1376]; the Hubbard Hamiltonian, hopping t against on-site repulsion U, and the transition near $U \sim W$ [1257]; the antiferromagnetic order that usually accompanies it, in Chapter 129; recorded as an open problem of Chapter 133 rather than as a solved one.]

Phenomenon 126.7 (Insulators that band theory says are metals)**.** Nickel oxide, cobalt oxide and the other transition-metal monoxides have an odd number of d electrons per primitive cell and a partially filled band, and by the counting of Phenomenon 126.4 must therefore be metals. They are measured to be insulators, and good ones, with activated conduction and resistivities many orders of magnitude above any metal [941, 1376]. The failure is not marginal and cannot be repaired by a better band calculation, because it is not a failure of approximation but of the premise: the one-electron picture treats the Coulomb repulsion between electrons as an average potential, and in a narrow band the cost of putting two electrons on the same site exceeds the energy gained by letting them move.

Derivation pending #245

The Hubbard Hamiltonian, in which a hopping amplitude competes with an on-site repulsion, and the localization of one electron per site when the repulsion exceeds the bandwidth — a transition to an insulating state that occurs at fixed, partial band filling and therefore has no description in band theory at all; noting that no controlled solution of this model in three dimensions exists, so the derivation available is a limiting-case argument and not a theory

126.7.3 Disorder and localization

[Reserved: Anderson's demonstration that sufficiently strong random disorder localizes the single-particle eigenstates outright, so that diffusion stops and conduction ceases with no gap anywhere in the spectrum [1134]; the mobility edge separating localized from extended states; weak localization and its characteristic negative magnetoresistance; and the strongest quantitative evidence for the whole picture — localization is what gives the quantum Hall plateaux of Chapter 128 a finite width.]

Chapter 127

Semiconductors

A semiconductor is nothing more exotic than a band insulator whose gap is small enough for the thermal energy $k_B T$, or a deliberately introduced impurity, to put a controllable number of carriers into an otherwise empty band — the identification is Wilson's [688, 689], and everything in this chapter is a consequence of it and of the band theory of Chapter 126. What makes the subject worth a chapter of its own is that the controllability is quantitative over some twenty orders of magnitude in carrier density, that the resulting devices are the most exhaustively verified artefacts in physics, and that the chain from a measured band gap to a working transistor is short enough to be followed in full. The Nobel prizes of 1956 (transistor) and 2000 (heterostructures and the integrated circuit) bracket that chain.

The chapter therefore runs from the band structure of silicon, germanium and gallium arsenide and the optical measurement of their gaps [1152], through carrier statistics, doping [946] and the measurement of mobility and lifetime [983], to rectification at a barrier [811, 843, 847], the p–n junction [948], the transistor [907, 990], and the optoelectronic devices that run the junction backwards — solar cell [1032], light-emitting diode [337, 1230] and injection laser [1229]. It ends in the low-dimensional structures [1422, 1524] that supply the two-dimensional electron gas of Chapter 128. The standard device reference is [2113]; the founding monograph is Shockley's [972].

Derivation pending #246

Semiconductors: all derivations of this chapter are pending.

127.1 What a semiconductor is

127.1.1 Wilson's band picture

[Reserved: Wilson's two papers, which identified a semiconductor as a filled-band insulator with a small gap and showed that its conductivity must rise with temperature rather than fall [688, 689]; the resulting exponential $\sigma \propto \mathrm{e}^{-E_g/2k_BT}$ and the resistivity range from $1\times10^{-5}\,\Omega\,\mathrm{m}$ to $1\times10^{5}\,\Omega\,\mathrm{m}$ that separates semiconductors from metals and insulators; the older empirical facts this explained — Faraday's observation that the resistance of silver sulfide falls *on heating, the opposite of every metal [93], and the erratic behaviour of "impure" samples now known to be doping.]*

Phenomenon 127.1 (Conduction that improves on heating)**.** The electrical resistance of silver sulfide *falls* as it is heated — the opposite of the behaviour of every metal, and the first recorded observation of semiconduction [93]. In a pure semiconductor the effect is not a weak drift but an exponential: over several decades of conductivity, and to within a slowly

varying prefactor,

$$\sigma(T) \propto \mathrm{e}^{-E_g/2k_BT}\,, \tag{127.1}$$

so that a plot of $\ln\sigma$ against $1/T$ is a straight line whose slope measures a single energy E_g characteristic of the material [688, 689]. The same substances have room-temperature resistivities anywhere between about $1\times10^{-5}\,\Omega\,\mathrm{m}$ and $1\times10^{5}\,\Omega\,\mathrm{m}$, and the value in a given specimen is exquisitely sensitive to impurities present at parts per million — which is why the early measurements were irreproducible and were long dismissed as dirt.

Derivation. Let the conduction band edge be at E_c and the valence band edge at E_v, with $E_g = E_c - E_v$, and let the chemical potential lie several k_BT inside the gap, so that the Fermi–Dirac occupation of Chapter 120 may be replaced by its Boltzmann tail at both edges. Integrating the band-edge density of states against that tail gives the standard forms

$$n = N_c\,\mathrm{e}^{-(E_c-\mu)/k_BT}\,, \qquad p = N_v\,\mathrm{e}^{-(\mu-E_v)/k_BT}\,, \tag{127.2}$$

where N_c and N_v are effective densities of states varying only as $T^{3/2}$. Multiplying the two expressions in Equation (127.2) eliminates the chemical potential entirely:

$$np = N_cN_v\,\mathrm{e}^{-E_g/k_BT} \equiv n_i^2\,, \tag{127.3}$$

the law of mass action, which holds whatever the doping and is the single most useful relation in the subject. In an undoped crystal every conduction electron leaves a hole behind, so $n = p = n_i$, and the conductivity $\sigma = e(n\mu_n + p\mu_p)$ inherits the square root of Equation (127.3), which is Equation (127.1). The factor of two in the exponent is the whole content of the measurement: it is there because the carriers are created in pairs across the gap, so the slope of the Arrhenius plot returns $E_g/2$ and not E_g. A metal has no gap and hence no such factor; its resistivity is governed instead by the scattering of Chapter 126, which grows with temperature, and the two temperature dependences therefore run in opposite directions. □

127.1.2 The band structures of silicon, germanium and gallium arsenide

[Reserved: the diamond and zincblende lattices and their Brillouin zones; the indirect gap of silicon and germanium, with conduction minima away from the zone centre — six valleys along Δ in silicon, eight half-valleys at L in germanium — against the direct gap of gallium arsenide at Γ; the anisotropic effective-mass tensors measured by cyclotron resonance [1063]; the light-, heavy- and split-off hole bands; $k\cdot p$ theory as the standard analytic parametrization near a band edge [1116]; tabulated parameters [2061].]

127.1.3 Measuring the gap optically

[Reserved: the absorption edge as the direct measurement of E_g; the square-root edge of a direct gap against the phonon-assisted structure of an indirect one, resolved into separate phonon-emission and phonon-absorption thresholds by Macfarlane and co-workers in silicon [1152] — a measurement of the band structure and the lattice modes of Chapter 125 at once; the shrinking of the gap with temperature and the empirical Varshni form [1360]; photoluminescence and photoconductivity as complementary determinations.]

Phenomenon 127.2 (The absorption edge, and the two thresholds of an indirect gap)**.** A semiconductor is transparent to photons below a threshold energy and absorbs strongly above it, and the threshold measures the band gap: about 1.12 eV in silicon, 0.67 eV in germanium and 1.42 eV in gallium arsenide, against about 5.5 eV in diamond [2061]. In silicon the edge is not one threshold but two, separated by a fixed energy of order tens

of meV, and the lower of the two weakens as the crystal is cooled while the upper does not; the separation matches twice the energy of a lattice mode of Chapter 125 at a definite wavevector [1152]. One optical measurement therefore returns the electronic gap, the fact that the gap is indirect, and the energy of the phonon that bridges it.

Derivation. A photon of energy of order 1 eV has wavevector ω/c of order $1 \times 10^{7}/\mathrm{m}$, three orders of magnitude smaller than the width $1 \times 10^{10}/\mathrm{m}$ of a Brillouin zone. On the scale of Chapter 126 the photon therefore carries energy but no crystal momentum, and an optical transition is vertical: it connects states of the same $\boldsymbol{k}$. If the valence maximum and the conduction minimum lie at the same $\boldsymbol{k}$, as in gallium arsenide, absorption begins as soon as $\hbar\omega \geq E_g$ and there is a single threshold. If they lie at different $\boldsymbol{k}$, as in silicon and germanium, no vertical transition of energy E_g exists, and the transition can proceed only with the simultaneous emission or absorption of a phonon of wavevector equal to the offset. Conserving energy in the two cases,

$$\hbar\omega = E_g + \hbar\Omega \quad \text{(phonon emitted)}\,, \qquad \hbar\omega = E_g - \hbar\Omega \quad \text{(phonon absorbed)}\,, \tag{127.4}$$

which are two thresholds separated by $2\hbar\Omega$ and placed symmetrically about E_g. The absorption branch requires a phonon to be waiting in the crystal, so its strength carries the Bose occupation factor of that mode and vanishes exponentially as $T \to 0$; the emission branch requires only that a phonon can be created, and survives. That asymmetry is what identifies which threshold is which, and it converts Equation (127.4) into a measurement of E_g and $\hbar\Omega$ separately. It also explains why an indirect-gap material is a poor light emitter and a good solar absorber only when thick: the transition needs a third particle, and is correspondingly weak. □

127.2 Carriers and their statistics

127.2.1 Intrinsic conduction

[Reserved: electrons thermally promoted across the gap leaving holes behind, so that $n = p = n_i$; the effective density of states, the law of mass action $np = n_i^2$ independent of doping, and the intrinsic densities that follow — of order $1 \times 10^{16}/\mathrm{m}^3$ in silicon at 300 K against $1 \times 10^{29}/\mathrm{m}^3$ of conduction electrons in copper; the chemical potential near mid-gap and its temperature drift; the Fermi–Dirac statistics of Chapter 120 and the Boltzmann limit in which they are usually applied.]

127.2.2 Doping and extrinsic conduction

[Reserved: substitutional donors and acceptors; the hydrogenic impurity level, binding energy scaled by m^*/m and $1/\epsilon_r^2$ to a few tens of meV*, and the consequent freeze-out at low temperature; the systematic measurements of boron- and phosphorus-doped silicon by Pearson and Bardeen that established the picture quantitatively [946]; compensation, degenerate doping and the Mott criterion for the impurity-band transition [1376]; zone refining and the purity — better than one part in 10^{10} — without which none of this is reproducible.]*

Phenomenon 127.3 (Doping controls the carrier density and its sign)**.** Substituting a group-V atom for a silicon atom at the level of parts per million raises the electron density by many orders of magnitude and leaves the hole density correspondingly depressed, in accordance with Equation (127.3); substituting a group-III atom does the same for holes. Over a wide intermediate range of temperature the carrier density equals the dopant density and is flat in temperature, the impurities being fully ionized; below that range it falls exponentially as the carriers freeze back onto their donors, with an activation energy of a few

tens of meV, and above it the intrinsic pairs of Equation (127.1) take over. The systematic measurements of boron- and phosphorus-doped silicon that established this picture are those of Pearson and Bardeen [946]. The Hall coefficient of Chapter 126 [208] reverses sign between the two doping types, which is the direct evidence that the positive carriers are real; near the intrinsic regime, where both species conduct, it can reverse sign again with temperature in one sample.

Derivation pending #247

The hydrogenic impurity level: a donor treated as a charge screened by the dielectric constant of the host and binding an electron of the band effective mass, giving a binding energy reduced from the hydrogenic value by the effective-mass ratio and by the square of the relative permittivity, hence tens of meV*; and the charge-neutrality condition combined with the law of mass action, which yields the freeze-out, saturation and intrinsic regimes as limits of one expression*

127.2.3 The Hall effect as a carrier measurement

[Reserved: the Hall coefficient [208] used in a semiconductor for what it does best — returning the sign of the majority carrier and, with the conductivity, both the density and the mobility; the sign reversal on passing from n-type to p-type material as the direct evidence for holes; two-carrier conduction near intrinsic, where the Hall coefficient can change sign with temperature; van der Pauw's theorem, which extracts resistivity and Hall coefficient from a sample of arbitrary shape with four contacts on its edge [1157], as the working laboratory method.]

127.2.4 Drift, diffusion and mobility

[Reserved: the drift–diffusion equations, the Einstein relation $D/\mu = k_B T/e$ linking them (the same relation proved for Brownian motion in Chapter 119); measured room-temperature mobilities of about $0.15\,\mathrm{m^2/(V\,s)}$ for electrons in silicon and eight times that in gallium arsenide; scattering by ionized impurities, by acoustic and polar optical phonons and their differing temperature laws; velocity saturation and the transport formalism of Chapter 124.]

127.2.5 Recombination and carrier lifetime

[Reserved: radiative band-to-band recombination, Auger recombination, and the trap-assisted Shockley–Read–Hall mechanism that dominates in indirect-gap material [1001, 1012]; minority-carrier lifetimes from nanoseconds to milliseconds and the diffusion length that follows; surface recombination and passivation; why the lifetime, not the mobility, is what limits solar cells and bipolar transistors alike.]

127.3 Junctions and rectification

127.3.1 Metal–semiconductor contacts

[Reserved: asymmetric conduction at a metal contact, known empirically since the cat's-whisker detector; the barrier theories advanced independently by Davydov [811], Mott [843] and Schottky [847], and the space-charge layer they predict; thermionic emission over the barrier against diffusion through it; Bardeen's demonstration that surface states pin the Fermi level and largely decouple the measured barrier height from the metal work function [896] — the result that redirected the effort toward the junction transistor; ohmic contacts.]

127.3.2 The p–n junction

[Reserved: the depletion region, built-in potential and internal field of a junction between p- and n-type material; Shockley's theory and the ideal diode law $I = I_0(\mathrm{e}^{eV/k_BT} - 1)$ [948]; capacitance–voltage profiling as the measurement of the doping profile; deviations from ideality — recombination current, high-injection, series resistance; avalanche and Zener breakdown [750]; the junction as the single component from which every device in Section 127.4 and Section 127.5 is assembled.]

Phenomenon 127.4 (Rectification at a junction)**.** A single crystal doped p-type on one side and n-type on the other conducts asymmetrically: forward biased it passes a current that rises exponentially with the applied voltage, with a logarithmic slope of e/k_BT per volt — about one decade of current per 60 mV at room temperature — while reverse biased it passes a small current that saturates and is almost independent of voltage. Over many decades the characteristic follows

$$I = I_0 \left(\mathrm{e}^{eV/k_BT} - 1 \right) , \tag{127.5}$$

with I_0 fixed by the doping and the minority-carrier properties and not adjustable [948]. The asymmetry arises inside a homogeneous single crystal with no interface between different substances, and the exponent contains the temperature rather than any device parameter, which is what marks Equation (127.5) as a statement about carrier statistics.

Derivation pending #248

The p–n junction in equilibrium: the diffusion of carriers across the metallurgical boundary, the depletion region and built-in potential that arrest it, and the equality of drift and diffusion currents that fixes that potential through the law of mass action; then, under bias, the Boltzmann factor by which the barrier lowering multiplies the minority-carrier density at the depletion edge, the diffusion of those excess carriers into the neutral regions, and the resulting saturation current — together with the recombination, high-injection and series-resistance terms by which real diodes depart from the ideal law

127.3.3 The Haynes–Shockley experiment

[Reserved: the experiment that made minority carriers visible — a pulse of holes injected into an n-type germanium bar, drifted by a field and detected downstream, giving drift velocity, diffusion constant and lifetime from the arrival time, the width and the area of one measured pulse [983]; the direct test of the Einstein relation it provides; why it settled the question of whether injected minority carriers exist at all, which underpins Section 127.4.1.]

Phenomenon 127.5 (Minority carriers drift, diffuse and recombine)**.** A pulse of holes injected at one point of an n-type germanium bar, where holes are vastly outnumbered, does not vanish into the electron sea. It travels down the bar at a velocity proportional to the applied field, arrives at a downstream contact after a measurable delay, broadens as it goes by an amount that grows as the square root of the transit time, and loses area exponentially with a lifetime of its own [983]. One oscilloscope trace therefore yields the drift mobility from the delay, the diffusion constant from the width and the lifetime from the area. The two transport coefficients so obtained are not independent: they satisfy

$$\frac{D}{\mu} = \frac{k_BT}{e} , \tag{127.6}$$

about 25.7 mV at 298 K, the same relation between fluctuation and dissipation that governs Brownian motion in Chapter 119.

Derivation. Consider the carriers in thermal equilibrium in a static potential $\phi(x)$, with no current flowing anywhere — equilibrium, not a steady state, so that drift and diffusion must cancel locally and not merely on average. For electrons of charge $-e$ the potential energy is $U = -e\phi$, so the equilibrium density is the Boltzmann distribution

$$n(x) \propto \mathrm{e}^{-U/k_B T} = \mathrm{e}^{e\phi/k_B T} \quad \Longrightarrow \quad \frac{\mathrm{d}n}{\mathrm{d}x} = \frac{e}{k_B T}\, n\, \frac{\mathrm{d}\phi}{\mathrm{d}x} = -\frac{e}{k_B T}\, n\, E\,. \tag{127.7}$$

An electron of mobility μ drifts at velocity $-\mu E$, against the field, so the particle flux is $\Gamma = -n\mu E - D\,\mathrm{d}n/\mathrm{d}x$. Setting it to zero and substituting Equation (127.7),

$$0 = -n\mu E + \frac{eD}{k_B T}\, nE = nE\left(\frac{eD}{k_B T} - \mu\right), \tag{127.8}$$

which must hold for every E and every n, and therefore holds if and only if $D/\mu = k_B T/e$, which is Equation (127.6). The content of the argument is that a single microscopic process — the same collisions — limits the drift and produces the random walk, so the coefficient describing the response to a force and the coefficient describing the spontaneous spreading cannot be chosen independently. Since μ and D are measured from different features of the *same* recorded pulse, the Haynes–Shockley experiment tests Equation (127.6) rather than assuming it. □

127.3.4 Tunnelling and the Esaki diode

[Reserved: a junction doped degenerately on both sides, so thin that carriers tunnel through rather than over the barrier; Esaki's observation of negative differential resistance in narrow germanium junctions, the first clear demonstration of tunnelling in a solid [1141]; the peak-to-valley ratio and its interpretation; phonon-assisted structure in the tunnelling current as spectroscopy; the line from here to superconducting tunnelling in Chapter 131.]

127.4 Devices

127.4.1 The transistor

[Reserved: the point-contact transistor of Bardeen and Brattain, with power gain demonstrated in a germanium slab in December 1947 [907]; Shockley's junction transistor, in which the minority-carrier injection of Section 127.3.3 replaces the fragile point contacts [972, 990]; current gain, base transit time and the Early effect [2113]; the 1956 Nobel prize; why the device is a physics result and not only an engineering one — it is a working measurement of minority-carrier transport.]

Phenomenon 127.6 (Power gain in a solid)**.** A germanium slab carrying two closely spaced point contacts, one forward biased at low impedance and one reverse biased at high impedance, delivers to a load at the second contact more signal power than is supplied to the first: a small current modulation at the input controls a larger power at the output, with gain measured at audio frequencies in December 1947 [907]. Gain, not rectification, is the observation — rectifying contacts had been in use for decades — and it requires that carriers injected at one contact survive the journey to the other, which is the minority-carrier transport of Phenomenon 127.5. Replacing the fragile point contacts by two p–n junctions sharing a thin base gives the same effect in a robust and quantitatively predictable form [972, 990].

Derivation pending #249

Transistor action from the junction physics already in hand: injection of minority carriers across a forward-biased emitter junction, their diffusion across a base thin compared with the diffusion length so that few recombine, and their collection by the reverse-biased collector junction; the resulting current gain as the fraction surviving the base, and the power gain as the product of that fraction with the ratio of the collector to the emitter impedance — the transfer of current from a low-resistance to a high-resistance circuit that gives the device its name

127.4.2 The MOSFET and the inversion layer

[Reserved: the field-effect idea, older than the point-contact transistor and blocked for two decades by the surface states of [896], unblocked by thermally grown silicon dioxide; the metal–oxide–semiconductor structure, the inversion layer and the threshold voltage [1583]; the inversion layer as a genuine two-dimensional electron gas, with quantized subbands [1678] — the same system in which quantized magnetoresistance [1329] and then the quantum Hall effect of Chapter 128 were found.]

127.4.3 The integrated circuit

[Reserved: Kilby's demonstration that every element of a circuit can be made in one block of semiconductor [1584], and the planar process that made it manufacturable; scaling, Moore's empirical doubling law and its status as an economic observation rather than a physical one [1315]; the physical limits that eventually bind — gate oxide tunnelling, dopant number fluctuations, power density; the 2000 Nobel prize shared with the heterostructure work of Section 127.5.3.]

127.5 Light and semiconductors

127.5.1 Photoconductivity and the solar cell

[Reserved: photogeneration of electron–hole pairs above the absorption edge and the photocurrent that results; the silicon p–n junction photocell of Chapin, Fuller and Pearson, the first solar cell of usable efficiency at about six percent [1032]; the open-circuit voltage, short-circuit current and fill factor actually measured; the Shockley–Queisser detailed-balance limit, about 33 % for a single gap under the solar spectrum [1219], derived from the thermodynamics of Chapter 54; measured cell efficiencies against it.]

127.5.2 Light-emitting diodes

[Reserved: Round's note recording light from a silicon carbide contact, the first observation of electroluminescence in a semiconductor [337]; the recombination of injected carriers across a direct gap as its explanation, and the resulting relation between emitted photon energy and E_g; the first visible practical device, the red gallium arsenide phosphide diode of Holonyak and Bevacqua [1230]; the long-delayed blue diode built on gallium nitride [1885] and the 2014 Nobel prize; efficiency measurements and droop.]

Phenomenon 127.7 (Electroluminescence)**.** A current passed through certain semiconducting contacts produces light. Round recorded a yellowish glow from a silicon carbide detector under bias, the first observation of the effect in a semiconductor [337]. In a forward-biased junction of a direct-gap material the emission is not thermal: the spectrum is a narrow band whose photon energy is close to the band gap of the material and shifts with it as the

material composition is changed, which is how the first practical visible emitter was tuned to red in gallium arsenide phosphide [1230] and, thirty years later, to blue in gallium nitride [1885]. The device is the junction of Phenomenon 127.4 run in reverse in the energetic sense: instead of light generating carriers, carriers generate light.

Derivation pending #250

Radiative recombination of injected minority carriers across a direct gap: the emission spectrum as the joint density of states weighted by the carrier occupations, giving a peak just above the gap with a width of order the thermal energy; the selection rule that suppresses the process in an indirect-gap material, where a phonon must participate, and the competing non-radiative Auger and trap-assisted channels that fix the internal quantum efficiency

127.5.3 The semiconductor laser and heterostructures

[Reserved: population inversion in a degenerately doped junction and the coherent emission observed from gallium arsenide diodes by Hall and co-workers [1229]; the threshold current and its temperature dependence; Kroemer's double-heterostructure proposal, confining carriers and photons in the same layer [1260], and the room-temperature continuous-wave operation it eventually delivered [2009]; quantum-well active regions [1524]; the stimulated-emission physics itself belongs to Chapter 122.]

127.5.4 Excitons

[Reserved: the bound electron–hole pair, tightly bound and localized in Frenkel's molecular limit [680] and weakly bound over many lattice constants in Wannier's [805]; hydrogenic binding energies of a few meV *in gallium arsenide, and the sharp absorption lines below the gap that measure them; why excitons are visible at room temperature in wide-gap and two-dimensional material and not in bulk silicon.]*

127.6 Low-dimensional semiconductors

127.6.1 Quantum wells and superlattices

[Reserved: epitaxial growth with monolayer control; the square well of Chapter 78 realized physically, with confined subbands observed as steps in absorption and as discrete emission lines [1524]; the Esaki–Tsu superlattice, its minibands and predicted negative differential conductivity [1422], in which the Bloch oscillations of Chapter 126 become observable; band-gap engineering as an experimental programme.]

127.6.2 The two-dimensional electron gas

[Reserved: modulation doping, which separates the donors from the carriers and so removes the dominant scattering channel [1618]; mobilities improved by more than three orders of magnitude, up to $1 \times 10^{3}\,\mathrm{m^{2}/(V\,s)}$ *in the best gallium arsenide heterostructures; the electronic properties of such two-dimensional systems reviewed comprehensively in [1678]; this is the sample that Chapter 128 takes as its starting point, and the reason the fractional effect was found at all.]*

Chapter 128

Experiment: The Quantum Hall Effect

On the night of 4–5 February 1980, measuring the Hall voltage of a silicon field-effect transistor at 1.5 K in about 15 T, von Klitzing, Dorda and Pepper found that the Hall resistance did not rise linearly with field but sat on flat plateaux at the values $h/\nu e^2$ with ν an integer, while the longitudinal resistance vanished [1655]. The plateau values depend on nothing — not on the material, the geometry, the mobility or the disorder — and have since been reproduced to parts in 10^{10} in different substances [1846]. That a macroscopic, irreducibly dirty solid should hand back a combination of fundamental constants exactly is the most striking single measurement in condensed matter physics, and it is why the ohm has been realized from this effect since 1990 and why $R_K = h/e^2$ is exact in the present SI [2333, 2431].

This chapter treats that measurement as an experiment, with the structured `experiment` environment (apparatus, procedure, observations with uncertainties in SI, interpretation, primary references) to be filled in when the content is written. It rests on the band theory and Landau quantization of Chapter 126 and on the heterostructure two-dimensional electron gas of Chapter 127; it carries forward into the topological language of Chapter 6 and Chapter 14, and into the metrology of Chapter 2. Two years after the integer effect, the same apparatus applied to a cleaner sample produced plateaux at *fractional* ν [1705], whose explanation [1725] required a new correlated ground state and excitations carrying a fraction of the electron charge — since measured directly [1943, 1944]. The standard collection is [1836]; a forty-year retrospective is [2360].

Derivation pending #251

Experiment: The Quantum Hall Effect: all derivations of this chapter are pending.

128.1 Historical context and the prediction under test

128.1.1 The classical Hall effect

[Reserved: Hall's 1879 measurement of a transverse voltage across a current-carrying gold leaf in a magnetic field [208]; the classical result $R_H = B/ne$ for a three-dimensional conductor and $R_{xy} = B/n_s e$ for a sheet of areal density n_s, linear in B with no structure whatever; the Hall bar geometry and the four-terminal measurement that separates R_{xx} from R_{xy}; this is the prediction the experiment of Section 128.4 contradicts.]

128.1.2 Two-dimensional electron gases

[Reserved: the inversion layer of a silicon metal–oxide–semiconductor transistor as the first true two-dimensional electron system, with electric quantum limit and subbands [1678]; the gate voltage as a direct and calibrated control of the areal carrier density; the modulation-doped gallium arsenide heterostructure [1618], which trades tunability for mobility; why two dimensions is essential — the Landau spectrum is then fully discrete, with no dispersion along the field.]

128.1.3 Landau quantization and what was expected

[Reserved: Landau levels at $E_n = (n + \tfrac{1}{2})\hbar\omega_c$ with degeneracy eB/h per unit area [660], and the filling factor $\nu = n_s h/eB$; the earlier observation of magneto-oscillatory conductance in silicon inversion layers by Fowler, Fang, Howard and Stiles [1329], the two-dimensional Shubnikov–de Haas effect; the expectation before 1980 — oscillations in R_{xx} and a Hall resistance with wiggles about the classical line; von Klitzing's actual purpose, a high-accuracy determination of the fine-structure constant [1655].]

128.2 Apparatus

128.2.1 The silicon MOSFET Hall bar

[Reserved: the original sample — a silicon metal–oxide–semiconductor field-effect transistor with a Hall-bar channel, source and drain, two pairs of potential probes and a gate covering the channel [1655]; typical dimensions and mobilities of order $1\,\mathrm{m^2/(V\,s)}$; the gate voltage sweeping n_s through several Landau levels at fixed field, which is the axis along which the original data were taken.]

128.2.2 Heterostructure samples

[Reserved: the modulation-doped gallium arsenide and aluminium gallium arsenide heterojunction [1618], in which the ionized donors sit in a barrier layer set back from the channel; mobilities exceeding $1 \times 10^3\,\mathrm{m^2/(V\,s)}$ and mean free paths of hundreds of μm; the sample quality without which the fractional plateaux of Section 128.4.3 do not appear at all — the single clearest case in the book of a phenomenon gated by materials purity.]

128.2.3 Cryogenics, magnets and electrical metrology

[Reserved: pumped helium-4, helium-3 and dilution refrigeration down to a few mK; superconducting solenoids to about 20 T and resistive or hybrid magnets beyond (Chapter 130); low-noise four-terminal measurement with lock-in detection at currents small enough to avoid electron heating, typically nA to μA; the cryogenic current comparator that transfers the quantized resistance to a calibrated standard resistor with an uncertainty of parts in 10^9 [2014].]

128.3 Procedure

[Reserved: the two equivalent sweeps — gate voltage at fixed field, or field at fixed density — and the measurement of R_{xx} and R_{xy} simultaneously; contact and lead-resistance checks; current reversal and field reversal to remove thermoelectric offsets and to symmetrize the Hall signal; the tests that must be done before any exactness claim is credible — current dependence, temperature dependence, contact placement, and comparison of samples of different geometry and material [1655, 2014].]

128.4 Observations

128.4.1 The integer plateaux

[Reserved: the primary data — R_{xy} flat over a finite range of density or field at $h/\nu e^2 = 25\,812.807\,\Omega/\nu$, with R_{xx} falling to an unmeasurably small value on each plateau and peaking sharply between them [1655]; the numbers to be tabulated with their uncertainties, together with the activation energy extracted from the temperature dependence of R_{xx}; plateau width as a function of disorder; the dissipationless conduction implied by $R_{xx} \to 0$ at finite current.]

Phenomenon 128.1 (Quantized Hall plateaux)**.** In a two-dimensional electron system at low temperature and high magnetic field, the Hall resistance does not rise linearly with the field as Section 128.1.1 requires. It rises in steps: over a finite range of carrier density, or equivalently of field, it sits on a flat plateau at

$$R_{xy} = \frac{h}{\nu e^2}\,, \qquad \nu = 1, 2, 3, \dots\,, \tag{128.1}$$

and jumps between plateaux in between [1655]. Across each plateau the longitudinal resistance R_{xx} falls to a value too small to measure and peaks sharply at the transitions, so on a plateau a finite current flows along the sample with no measurable voltage drop along it — dissipationless transport in a disordered solid at a temperature where the same solid is an ordinary conductor in zero field. The plateau value Equation (128.1) contains no material parameter whatever.

Derivation. The value on the plateau, though not the existence of the plateau, is elementary once the Landau degeneracy is counted. Take a sample $L_x \times L_y$ in a perpendicular field B and work in the gauge $\boldsymbol{A} = (0, Bx, 0)$. The Hamiltonian then commutes with translations along y, so eigenstates may be taken proportional to $\mathrm{e}^{\mathrm{i}k_y y}$, and completing the square shows that each is a one-dimensional oscillator of frequency $\omega_c = eB/m$ centred on

$$x_0 = \frac{\hbar k_y}{eB}\,, \tag{128.2}$$

with energy $(n + \tfrac{1}{2})\hbar\omega_c$ independent of k_y. Periodic boundary conditions in y quantize $k_y = 2\pi j/L_y$ with j an integer, and requiring the centre Equation (128.2) to lie inside the sample, $0 \le x_0 \le L_x$, bounds j by eBL_xL_y/h. Each Landau level therefore holds

$$\frac{N}{A} = \frac{eB}{h} \tag{128.3}$$

states per unit area — one per quantum of magnetic flux h/e threading the sample, which is the way to remember it. If exactly ν levels are filled, the areal carrier density is $n_s = \nu eB/h$. Now apply the classical two-dimensional Hall result of Section 128.1.1, $R_{xy} = B/n_s e$, which follows from force balance alone and holds whenever the carriers move without dissipation:

$$R_{xy} = \frac{B}{n_s e} = \frac{B}{e}\cdot\frac{h}{\nu eB} = \frac{h}{\nu e^2}\,, \tag{128.4}$$

which is Equation (128.1). Note what has been shown and what has not. The field B cancels, and with it every scale in the problem, which is why Equation (128.4) contains no material parameter. But the argument holds only at exactly integer filling, a set of measure zero in the density; it therefore explains the value of the plateau and gives no reason for there to be a plateau at all. That is the subject of Section 128.5.1 and Section 128.5.2, and it is where the disorder enters. □

128.4.2 Exactness and universality

[Reserved: the property that makes the effect extraordinary — the plateau value is independent of material, geometry, mobility and carrier density; the direct comparison of quantized resistances in silicon and in gallium arsenide, agreeing to about one part in 10^{10} [1846]; the reproducibility between national laboratories [2014]; the honest statement of what is measured — a ratio, not an absolute value — and of the systematic effects (current, temperature, contacts) that limit it.]

Phenomenon 128.2 (Exactness and universality of the plateau value)**.** The plateau value Equation (128.1) depends on nothing about the sample. Quantized Hall resistances measured in a silicon metal–oxide–semiconductor device and in a gallium arsenide heterostructure — different host materials, different carrier densities, different mobilities, different geometries, different disorder — agree with one another to about one part in 10^{10} [1846], and independent national laboratories reproduce the value at comparable uncertainty [2014]. This is not a small correction to a material-dependent answer; the material dependence is absent to the accuracy of the best resistance metrology in existence, in samples that are macroscopic, irreducibly disordered and individually unreproducible. The effect accordingly realizes the ohm, and the von Klitzing constant $R_K = h/e^2$ is exact in the present SI, $R_K = 25\,812.807\,45\,\Omega$ [2333, 2431].

Derivation pending #252

Laughlin's gauge argument: that adiabatically threading one flux quantum through a Hall cylinder is a gauge transformation and must map the spectrum to itself, so that its only effect is to transfer an integer number of electrons from one edge to the other; the resulting Hall conductance as that integer times the square of the elementary charge over Planck's constant, with no reference to the Hamiltonian; and the requirement of a mobility gap at the Fermi level that makes the process adiabatic — together with the topological restatement in which the integer is the Chern number of the occupied states, which cannot change under any smooth deformation of the sample

128.4.3 The fractional plateaux

[Reserved: Tsui, Störmer and Gossard's observation, in the extreme quantum limit of a high-mobility heterostructure, of a plateau and a vanishing R_{xx} at $\nu = 1/3$ [1705]; the subsequent hierarchy at $\nu = p/q$ with q odd, and the even-denominator state at $\nu = 5/2$ [1800]; the energy gaps, of order K, measured from activated transport; why no single-particle argument can produce a fractional plateau, which is the point of Section 128.5.5.]

Phenomenon 128.3 (Fractional plateaux)**.** In a heterostructure of much higher mobility, driven into the extreme quantum limit where the lowest Landau level is only partly filled, the same measurement finds a plateau in R_{xy} and a vanishing R_{xx} at the fractional filling $\nu = 1/3$ [1705]. Further plateaux appear as samples improve, at a hierarchy of fillings $\nu = p/q$ with q odd, and — against the pattern — at the even-denominator value $\nu = 5/2$ [1800]. The temperature dependence of R_{xx} on each plateau is activated, so each fraction carries a genuine energy gap, of order a kelvin in temperature units. The observation is not a refinement of Phenomenon 128.1: the derivation given there required an integer number of filled levels, and within a partly filled Landau level all single-particle states are degenerate, so no one-electron argument can produce a gap at all. The gap must be made by the interactions between the electrons, and the fact that the cleanest samples are the ones that show it confirms that disorder is not its source.

Derivation pending #253

Laughlin's correlated wavefunction for filling one over an odd integer, an incompressible liquid rather than a filled single-particle band; the plasma analogy that fixes its density and its excitation gap; the antisymmetry requirement that forces the exponent to be odd; and the hierarchy or composite-fermion construction that accounts for the remaining observed fractions

128.5 Interpretation

128.5.1 Laughlin's gauge argument

[Reserved: the argument that fixes the plateau value with almost no dynamical input — adiabatic insertion of one flux quantum h/e through a Hall cylinder must return the system to itself, so an integer number of electrons is transferred from one edge to the other and $\sigma_{xy} = \nu e^2/h$ exactly [1671]; gauge invariance and the role of a mobility gap; why this explains exactness without explaining plateau width, which needs Section 128.5.2.]

128.5.2 Disorder, localization and plateau width

[Reserved: the apparent paradox that the effect is more perfectly quantized in a dirtier sample; Anderson localization [1134] broadening each Landau level into a narrow band of extended states flanked by localized ones, so that adding carriers to localized states changes nothing measurable and the Hall resistance stays pinned [1689]; the mobility edge and the plateau-to-plateau transition as a quantum critical point with measured scaling exponents; the honest statement that disorder is not a nuisance here but a necessary ingredient.]

128.5.3 Edge states

[Reserved: the confining potential bending Landau levels through the Fermi energy at the sample boundary, producing one-dimensional chiral channels that cannot backscatter because the counter-propagating channel is on the far side of the sample; Büttiker's multiterminal transport formulation, in which each edge channel contributes exactly e^2/h [1807]; the experimental signatures — contact independence, selective population and detection of individual channels; edge states as the model for every later topological boundary mode.]

128.5.4 Topology: TKNN and the Chern number

[Reserved: Thouless, Kohmoto, Nightingale and den Nijs' result that σ_{xy} for a filled band is e^2/h times a topological invariant — the integral of the Berry curvature of the Bloch states over the Brillouin zone, an integer Chern number [1704]; why an integer cannot change under smooth deformation, which is the deepest available answer to why the plateau is exact; the fibre-bundle language of Chapter 14 and the homotopy of Chapter 6; Haldane's lattice model showing that no net field is needed, only broken time reversal [1808].]

128.5.5 The Laughlin state and fractional charge

[Reserved: Laughlin's variational wavefunction $\prod_{i<j}(z_i - z_j)^m \mathrm{e}^{-\sum |z_i|^2/4\ell^2}$ for $\nu = 1/m$ with m odd, an incompressible correlated liquid rather than a filled single-particle band [1725]; its quasiholes, created by adiabatic flux insertion and carrying charge e/m; the plasma analogy that fixes the charge and the fractional exchange statistics; the antisymmetry requirement of Chapter 83 that forces m odd.]

128.6 Later measurements and precision

128.6.1 Fractional charge and anyon statistics measured

[Reserved: shot noise at a quantum point contact in the weak backscattering limit, whose Poisson noise $2q\bar{I}$ measures the carrier charge directly; the independent 1997 determinations of $q = e/3$ at $\nu = 1/3$ [1943, 1944]; and the later interference and collision experiments reporting the fractional exchange phase $2\pi/3$ itself [2354, 2364], the first direct evidence for statistics that are neither Bose nor Fermi — with a clear statement of what the data do and do not establish.]

Phenomenon 128.4 (The current is carried in fractions of an electron charge)**.** Shot noise measures the charge of the carrier that crosses a barrier, because the noise power of a stream of independent discrete carriers is proportional to the product of the carrier charge and the mean current and to nothing else. Applied to weak backscattering at a quantum point contact in the $\nu = 1/3$ state, the measurement returns a carrier charge of $e/3$, in two independent determinations published within weeks of one another [1943, 1944]. Nothing in the sample carries a fraction of the electron charge as a constituent — the electrons are whole — so the object being counted is a collective excitation of the correlated liquid of Phenomenon 128.3. Later interference and collision experiments in the same states report the further, and far more delicate, result that exchanging two such excitations multiplies the state by a phase which is neither $+1$ nor -1 [2354, 2364]; these are excitations of a three-dimensional semiconductor confined to a plane, and the claim is about them and not about any elementary particle.

Derivation pending #254

The quasihole of the correlated ground state and the determination of its charge from the deficit of density produced by adiabatically threading one flux quantum, together with the plasma analogy that makes the deficit exact; and the Schottky formula for the noise of a Poisson stream of independent carriers, whose proportionality to charge times current is what converts a noise measurement into a charge measurement

128.6.2 Composite fermions

[Reserved: Jain's construction attaching an even number of flux quanta to each electron, mapping the fractional effect at $\nu = p/(2p\pm1)$ onto the integer effect of composite fermions in a reduced effective field [1823]; the prediction that at $\nu = 1/2$ the effective field vanishes and a Fermi sea of composite fermions forms; the measurements that support it — effective-field Shubnikov–de Haas oscillations and semiclassical geometric resonances [1861]; the hierarchy of observed fractions as the strongest test.]

128.6.3 Graphene and the relativistic quantum Hall effect

[Reserved: the isolation of single-layer graphene and the demonstration of its field effect [2063]; its linear Dirac dispersion at the zone corners, so that Landau levels go as $\sqrt{nB}$ and a zero-energy level exists; the half-integer sequence $\sigma_{xy} = 4e^2(n + \frac{1}{2})/h$ and the Berry phase π measured simultaneously by two groups [2081, 2085]; observation of the effect at room temperature; the connection to the Dirac equation of Chapter 93, and the care needed — these are quasiparticles of a lattice, not elementary fermions.]

Phenomenon 128.5 (A shifted quantization sequence in graphene)**.** In a single atomic layer of graphite the quantized Hall sequence is not that of Equation (128.1). The plateaux

are found at half-integer multiples of the fourfold-degenerate conductance quantum, $\sigma_{xy} = 4(n + \frac{1}{2})e^2/h$, measured simultaneously and independently by two groups [2081, 2085]; the spacing of the Landau levels goes as the square root of the field rather than linearly in it, and a level sits at zero energy for every field. The same measurements extract a Berry phase of π from the phase of the Shubnikov–de Haas oscillations. Because the level spacing at accessible fields is then far larger than in a semiconductor heterostructure, the quantization survives to room temperature. The carriers responsible are low-energy excitations of a two-dimensional crystal of carbon atoms, whose dispersion near the corners of its Brillouin zone is linear; the resemblance to the equation of Chapter 93 is a property of that band structure and not a claim about elementary particles.

Derivation pending #255

Landau quantization of a linear rather than quadratic dispersion, which gives level energies proportional to the square root of the product of level index and field and places one level exactly at zero energy; the shared occupation of that zero level by electrons and holes, which shifts the filling sequence by one half; the fourfold spin and valley degeneracy of the observed steps; and the Berry phase of the cyclotron orbit that appears as the offset in the Shubnikov–de Haas phase

128.6.4 Quantum spin Hall and topological insulators

[Reserved: Kane and Mele's proposal that spin–orbit coupling alone, with time-reversal symmetry intact and no magnetic field, yields a $\mathbb{Z}_2$-classified insulator with helical edge states [2077]; the mercury telluride quantum-well prediction [2092] and its confirmation in transport, with a conductance plateau at $2e^2/h$ independent of sample width [2110]; the first three-dimensional topological insulator, its surface Dirac cone imaged by the photoemission of Chapter 126 [2128].]

128.6.5 The quantum anomalous Hall effect

[Reserved: Haldane's model realized in a material — a magnetically doped topological insulator thin film showing a quantized Hall resistance h/e^2 at zero applied field [2216]; the required conditions (ferromagnetic order, chemical potential in the gap, temperatures of tens of mK*) and the honest note that the quantization there is far less precise than in Section 128.4.2; the link to the magnetism of Chapter 129.]*

Phenomenon 128.6 (Quantization without an applied field)**.** A thin film of a magnetically doped topological insulator, cooled to tens of mK and with its chemical potential tuned into the gap, shows a Hall resistance equal to h/e^2 and a vanishing longitudinal resistance at *zero* applied magnetic field [2216]. Every ingredient that Phenomenon 128.1 used — the Landau levels, their degeneracy, the field itself — is absent; what remains is broken time reversal, supplied here by the ferromagnetic order of the dopants rather than by an external field. The quantization is real but is far less accurate than that of Phenomenon 128.2, and it survives only at temperatures well below the magnetic ordering temperature, which is the honest measure of how much remains to be understood about it.

Derivation pending #256

That a quantized Hall conductance requires only a gapped band structure with broken time-reversal symmetry and a non-zero topological invariant of the occupied bands, not a uniform magnetic field; the lattice model exhibiting this with zero net flux through the unit cell; and the mechanism by which magnetic dopants in a strong-spin–orbit thin film open the required gap in the surface states

128.7 Metrology: the ohm and the SI

[Reserved: the von Klitzing constant $R_K = h/e^2$, its conventional 1990 value and its status since the 2019 redefinition, in which h and e are exact and therefore $R_K = 25\,812.807\,45\,\Omega$ exactly [2333, 2431]; the practical realization of the ohm from a quantized Hall sample [2014], and of the volt from the Josephson effect of Chapter 131; the quantum metrology triangle closing volt, ohm and ampere; the original motivation — a determination of the fine-structure constant [1655] — and its present standing against the electron $g-2$ route of Chapter 112; graphene resistance standards operating at relaxed field and temperature [2360].]

Chapter 129

Magnetism in Matter

Magnetism in matter is the cleanest demonstration in the book that ordinary condensed matter is a quantum system. The Bohr–van Leeuwen theorem [368, 484] shows that a classical system of charges in thermal equilibrium has exactly zero magnetization at any field and temperature; every effect in this chapter — Faraday's diamagnetism [119], Curie's paramagnetic law [264], the spontaneous magnetization of iron — therefore has no classical explanation whatever. What supplies it is spin and the Pauli principle of Chapter 83, acting through the exchange interaction [537, 605]: an electrostatic energy, of order eV, masquerading as a coupling between spins and hence producing ordering temperatures of order 1000 K where the magnetic dipole–dipole interaction could manage about 1 K.

The chapter builds from the response of isolated moments to the collective phenomena: Weiss's molecular field [339] and its microscopic replacement by exchange; Néel's antiferromagnetism [781, 918], invisible to a magnetometer and made visible by neutron diffraction [949]; domains [459, 675] and hysteresis; spin waves [643] counted directly by inelastic neutron scattering [1104]; itinerant ferromagnetism in the band language of Chapter 126 [820]; and magnetic resonance [885, 887, 894] as the spectroscopy that measures all of it. The critical behaviour near the ordering temperature is the worked example of Chapter 123 and is not repeated here. The standard references are [1570, 2011], with Van Vleck's monograph [709] as the founding treatment.

Derivation pending #257

Magnetism in Matter: all derivations of this chapter are pending.

129.1 The magnetic response of matter

129.1.1 Diamagnetism

[Reserved: Faraday's discovery that every substance responds to a magnetic field, most of them by being repelled [119]; the Larmor precession argument and the Langevin diamagnetic susceptibility $\chi \propto -\langle r^2 \rangle$ per atom [323]; measured values of order -10^{-5} in SI volume susceptibility, and their near-independence of temperature; the definitions and unit conventions for $\boldsymbol{B}$, $\boldsymbol{H}$, $\boldsymbol{M}$ and χ fixed in Chapter 62; perfect diamagnetism as a distinct phenomenon, belonging to Chapter 130.]

Phenomenon 129.1 (Every substance responds to a magnetic field)**.** There is no magnetically inert matter. Suspended in a strong field gradient, some substances are drawn towards the strong-field region and the majority are pushed away from it; the repulsion is universal, weak, and — unlike the attraction — essentially independent of temperature [119]. In modern units the volume susceptibility of a diamagnetic solid or liquid is negative and

of order 10^{-5}. The universality is the informative part: a property possessed by bismuth, water, glass and gold alike cannot depend on permanent atomic moments, since most of these substances have none.

Derivation pending #258

Diamagnetism as the induced response of the bound electrons: the Larmor precession an applied field imposes on an atomic charge distribution, the induced moment opposing the field by Lenz's rule, and the resulting susceptibility proportional to the mean square radius of the electronic charge distribution and to the number of electrons per unit volume, with no temperature dependence because no level populations change

129.1.2 Paramagnetism and the Curie law

[Reserved: Curie's systematic measurements establishing $\chi = C/T$ for paramagnets and the sharp distinction between the three classes of response [264]; Langevin's classical statistical derivation, with the Langevin function $L(x) = \coth x - 1/x$ [323]; its quantum replacement, the Brillouin function for a moment of angular momentum J [566], and the saturation curves of chrome, iron and gadolinium salts that confirmed the quantized rather than the classical form; the effective moment p_{eff} extracted from a measured Curie constant.]

Phenomenon 129.2 (The Curie law)**.** Substances containing permanent atomic moments are attracted into a field, with a susceptibility inversely proportional to the absolute temperature,

$$\chi = \frac{C}{T}\,, \tag{129.1}$$

over a wide temperature range and to good accuracy — one of the three distinct classes of magnetic behaviour that Curie's systematic survey separated [264]. At low temperature and high field the magnetization departs from Equation (129.1) and saturates, and the measured saturation curves of paramagnetic salts follow the stepped, quantized form appropriate to a moment of definite angular momentum [566] rather than the classical curve for a freely orientable vector [323, 2011]. The Curie constant C therefore measures the size of the atomic moment, and the shape of the saturation curve measures its angular momentum.

Derivation. Take the simplest case, n independent moments per unit volume, each of magnitude μ and each with only two accessible orientations along the field, of energies $\mp\mu B$. Their thermal average is

$$\langle \mu_z \rangle = \mu\,\frac{\mathrm{e}^{\mu B/k_B T} - \mathrm{e}^{-\mu B/k_B T}}{\mathrm{e}^{\mu B/k_B T} + \mathrm{e}^{-\mu B/k_B T}} = \mu \tanh \frac{\mu B}{k_B T}\,, \tag{129.2}$$

so that $M = n\mu \tanh(\mu B/k_B T)$. Two limits follow at once. For $\mu B \gg k_B T$ the hyperbolic tangent saturates at unity and $M \to n\mu$: the magnetization has an upper bound, which is the saturation observed. For $\mu B \ll k_B T$, expanding $\tanh x \approx x$ gives $M = n\mu^2 B/k_B T$ and hence

$$\chi = \frac{\mu_0 M}{B} = \frac{n\mu_0 \mu^2}{k_B T}\,, \tag{129.3}$$

which is Equation (129.1) with $C = n\mu_0\mu^2/k_B$. The inverse temperature is the competition of alignment energy against thermal disorder, and nothing else; it is the same $1/T$ that governs any two-level population. For a moment of general angular momentum J and Landé factor g the sum over $2J+1$ orientations replaces Equation (129.2) by the Brillouin function [566], whose small-argument limit gives Equation (129.1) with the effective moment $\mu_{\text{eff}} = g\sqrt{J(J+1)}\,\mu_B$ in place of μ; the saturation value is $ngJ\mu_B$, so the two limits

measure different combinations of g and J and between them determine both. Note that Equation (129.2) already assumes that the orientations are discrete, which no classical theory permits (Phenomenon 129.3). □

129.1.3 The Bohr–van Leeuwen theorem

[Reserved: the theorem that in classical statistical mechanics (Chapter 118) the partition function is independent of the vector potential, because the field shifts the momentum integration variable, so the equilibrium magnetization vanishes identically; proved independently by Bohr [368] and van Leeuwen [484]; the consequence that magnetism is evidence for quantum mechanics in bulk matter, and the specific ingredients (quantized levels, spin, the exclusion principle) that evade it.]

Phenomenon 129.3 (Magnetism has no classical explanation)**.** Matter in thermal equilibrium is measurably magnetic: diamagnetic (Phenomenon 129.1), paramagnetic (Phenomenon 129.2) or, in a few elements, spontaneously magnetized. Classical statistical mechanics, applied to any collection of charges in thermal equilibrium at any temperature and in any applied field, gives exactly zero magnetization [368, 484]. The disagreement is not one of magnitude but of existence, so every measurement in this chapter, including the most banal, is a measurement of quantum mechanics operating in bulk matter at room temperature.

Derivation. Consider N classical charges q_i of masses m_i in an arbitrary potential $U(\boldsymbol{r}_1, \dots, \boldsymbol{r}_N)$ and an external magnetic field described by a vector potential $\boldsymbol{A}$. The field enters the Hamiltonian only through the minimal coupling,

$$\mathcal{H} = \sum_i \frac{\left[\boldsymbol{p}_i - q_i \boldsymbol{A}(\boldsymbol{r}_i)\right]^2}{2m_i} + U\,, \tag{129.4}$$

and the classical canonical partition function is the integral of $\mathrm{e}^{-\beta\mathcal{H}}$ over all positions and all momenta. Hold the positions fixed and change momentum variables to $\boldsymbol{\pi}_i = \boldsymbol{p}_i - q_i\boldsymbol{A}(\boldsymbol{r}_i)$. The shift is a translation of each momentum by a constant, so its Jacobian is unity, and — this is the whole of the argument — each momentum ranges over all of $\mathbb{R}^3$ both before and after, so the domain of integration is unchanged. Hence

$$Z = \int \prod_i \mathrm{d}^3 r_i\, \mathrm{d}^3 p_i\; \mathrm{e}^{-\beta\mathcal{H}} = \int \prod_i \mathrm{d}^3 r_i\, \mathrm{d}^3 \pi_i\; \mathrm{e}^{-\beta\left(\sum_i \pi_i^2/2m_i + U\right)}\,, \tag{129.5}$$

in which $\boldsymbol{A}$ has disappeared entirely. The free energy $F = -k_B T \ln Z$ is therefore independent of the applied field, and the equilibrium magnetization $M = -V^{-1}\partial F/\partial B$ vanishes identically, for every temperature, every field and every interaction U. The result is stronger than a null prediction for some model: it says no classical model can succeed. What evades it is the discreteness of the momentum-space integration in quantum mechanics — the levels are a countable set that the field genuinely rearranges rather than a continuum that a shift maps onto itself — together with spin, which is not of the form Equation (129.4) at all, and the exclusion principle of Chapter 83, which controls how the levels are filled. □

129.1.4 Magnetism of the conduction electrons

[Reserved: Pauli paramagnetism — a temperature-independent susceptibility $\chi = \mu_0\mu_B^2 g(E_F)$ because only electrons within k_BT of the Fermi surface can flip [586], a direct measurement of the density of states of Chapter 126; Landau diamagnetism from the orbital quantization of the same electrons, exactly $-1/3$ of the Pauli term for free electrons [660]; why the two nearly cancel and what the residual measures; the de Haas–van Alphen oscillations as the same physics at high field.]

129.2 Where the moments come from

129.2.1 Local moments and the effective moment

[Reserved: partially filled shells, Hund's rules and the Landé g-factor as established in Chapter 88; the measured effective moments of the rare-earth ions, in excellent agreement with the free-ion J; the transition-metal ions, in agreement only with the spin-only value, and crystal-field quenching of the orbital moment as the explanation [709, 2011]; the Bohr magneton $\mu_B = e\hbar/2m_e$ as the natural unit and its value from [2431].]

129.2.2 The gyromagnetic ratio measured

[Reserved: the Einstein–de Haas experiment — a suspended iron cylinder magnetized along its axis acquires a measurable rotation, because angular momentum is conserved and the magnetization carries some [426]; Barnett's converse, magnetization by rotation [424]; and the point that makes the pair worth recounting — Einstein and de Haas reported a gyromagnetic ratio agreeing with the orbital value $g = 1$ that their own theory expected, whereas the correct value is $g \approx 2$, as Barnett's measurement [424] and every later repetition found; the founding experiment of the subject therefore erred in the direction of its authors' prediction, which makes it a case study for the error analysis of Chapter 2; the moment is spin, not orbital, with the modern precision value from Chapter 112 and the formalism of Chapter 79.]

Phenomenon 129.4 (Magnetization carries angular momentum)**.** An iron cylinder hung on a fine fibre and magnetized along its axis by a surrounding solenoid begins to rotate about that axis, and reverses its rotation when the magnetization is reversed [426]. The converse also holds: an unmagnetized ferromagnetic rod spun about its axis acquires a magnetization proportional to its angular velocity [424]. Magnetization is therefore not merely accompanied by angular momentum, it *is* angular momentum, in a fixed ratio

$$\gamma = \frac{m_{\mathrm{tot}}}{L_{\mathrm{e}}} = \frac{g\,e}{2m_e}\,, \tag{129.6}$$

with m_{tot} and L_{e} the total magnetic moment and the total electronic angular momentum. The measured value is $g \approx 2$, not the $g = 1$ that orbital motion would give: the moment of a ferromagnet is carried by spin. The measurement is also a standing lesson in experimental method. Einstein and de Haas reported a ratio consistent with the orbital value $g = 1$ that their own theory expected, whereas Barnett's independent measurement [424] and every later repetition give approximately 2; the founding experiment of the subject erred in the direction of its authors' prediction.

Derivation. The suspended cylinder is mechanically isolated about its axis, so the component of total angular momentum along that axis is conserved. Split it into the mechanical angular momentum L_{mech} of the body as a rigid rotor and the internal electronic angular momentum L_{e} that the aligned moments carry. Conservation gives $\Delta L_{\mathrm{mech}} = -\Delta L_{\mathrm{e}}$, and Equation (129.6) converts the second into a change of magnetic moment:

$$\Delta L_{\mathrm{mech}} = -\frac{\Delta m_{\mathrm{tot}}}{\gamma} = -\frac{2m_e}{g\,e}\,\Delta m_{\mathrm{tot}}\,. \tag{129.7}$$

Every quantity in Equation (129.7) except g is directly measurable: the change in magnetic moment from the fluxmeter or from the saturation magnetization, and the change in mechanical angular momentum from the amplitude of the torsional oscillation the magnetization reversal excites, given the known moment of inertia and fibre constant. The experiment therefore determines g with no adjustable parameter. Because the two candidate answers,

1 for orbital and 2 for spin, differ by a factor of two rather than by a percentage, the measurement discriminates decisively — provided it is not biased, which is exactly the point of the historical note above. □

129.3 Ferromagnetism

129.3.1 Weiss's molecular field

[Reserved: Weiss's postulate of an internal field proportional to the magnetization, $B_{\mathrm{mol}} = \lambda M$, and the self-consistent equation it yields [339]; spontaneous magnetization below T_C, the Curie–Weiss law $\chi = C/(T-\theta)$ above it, and the measured Curie temperatures — 1043 K for iron, 1394 K for cobalt, 631 K for nickel [2011]; the fatal quantitative problem, that a molecular field of the required size corresponds to 1×10^3 T and cannot be magnetic in origin; mean-field theory's exponents and their disagreement with experiment, treated in Chapter 123.]

Phenomenon 129.5 (Spontaneous magnetization and the Curie–Weiss law)**.** A few elements are magnetized in zero applied field. Below a sharp temperature T_C — 1043 K for iron, 1394 K for cobalt, 631 K for nickel [2011] — a specimen carries a magnetization that grows continuously from zero as the temperature falls, with no applied field to define its direction. Above T_C the spontaneous moment is gone and the susceptibility follows not the Curie law Equation (129.1) but the Curie–Weiss form

$$\chi = \frac{C}{T-\theta}\,, \tag{129.8}$$

with θ positive and close to T_C [339]. The ordering temperature is the informative number. It corresponds to an internal field of order 1×10^3 T, whereas the magnetic field one aligned atomic moment produces at its neighbour is of order 0.1 T: whatever aligns the moments, it is four orders of magnitude too strong to be their magnetic interaction.

Derivation. Weiss's hypothesis is that each moment experiences, in addition to any applied field B, an internal field proportional to the magnetization already present, $B_{\mathrm{mol}} = \lambda M$ [339]. Substituting $B \to B + \lambda M$ into the two-level result Equation (129.2) makes the magnetization implicit in itself,

$$M = n\mu \tanh \frac{\mu(B+\lambda M)}{k_B T}\,, \tag{129.9}$$

and everything follows from that one equation. Put $B = 0$ and write $m = M/n\mu$ and $k_B T_C = n\mu^2\lambda$; then Equation (129.9) becomes

$$m = \tanh \frac{T_C m}{T}\,. \tag{129.10}$$

The right-hand side of Equation (129.10) is an odd, monotonically increasing, concave function of $m > 0$ with slope T_C/T at the origin and saturating at unity. It therefore meets the line m at a non-zero point if and only if that initial slope exceeds one, that is if and only if $T < T_C$: below T_C a spontaneous magnetization exists, above it only $m = 0$ solves the equation. For $T > T_C$ and small applied field, linearize Equation (129.9) in B and M, $M = n\mu^2(B+\lambda M)/k_B T$, and solve for the susceptibility:

$$\chi = \frac{\mu_0 M}{B} = \frac{n\mu_0\mu^2/k_B}{T-T_C} = \frac{C}{T-T_C}\,, \tag{129.11}$$

which is Equation (129.8) with the same Curie constant as Equation (129.3): a paramagnet with a positive feedback term diverges at T_C instead of at zero. What the derivation does *not*

supply is λ. Setting $k_B T_C = n\mu^2\lambda$ with the observed T_C gives λM of order 1×10^3 T, which no arrangement of magnetic dipoles can produce; the molecular field is a bookkeeping device for an interaction that must be electrostatic in origin (Section 129.3.2). The mean-field treatment also predicts critical exponents that measurement contradicts, which is the subject of Chapter 123. □

129.3.2 Exchange

[Reserved: the resolution — the antisymmetry of the many-electron wavefunction (Chapter 83) correlates spin orientation with spatial overlap, so the Coulomb energy itself depends on the relative spin; Dirac's formulation of the exchange energy in terms of a spin permutation operator [537]; Heisenberg's application to ferromagnetism [605]; the Heitler–London calculation for two atoms as the transparent worked case [577]; magnitudes of order eV*, which is what makes room-temperature magnets possible.]*

129.3.3 The Heisenberg model

[Reserved: the Hamiltonian $H = -\sum_{ij} J_{ij} \boldsymbol{S}_i \cdot \boldsymbol{S}_j$ [605], the sign convention for ferromagnetic and antiferromagnetic J, and its status as an effective model derived from the electronic problem rather than postulated; the Ising and XY limits and the role of anisotropy; the Mermin–Wagner theorem forbidding spontaneous breaking of a continuous spin symmetry in one and two dimensions at finite temperature [1337]; exchange constants extracted from measured spin-wave dispersions in Section 129.6.2; the connection to the lattice models of Chapter 123.]

129.3.4 Itinerant ferromagnetism

[Reserved: the difficulty that iron, cobalt and nickel have non-integer moments per atom — 2.22, 1.72 and 0.606 Bohr magnetons — which no localized-moment model can produce; Stoner's band theory of collective-electron ferromagnetism, with exchange splitting the spin-up and spin-down bands and the criterion $Ig(E_F) > 1$ for spontaneous polarization [820]; the spin-split bands as measured by spin-resolved photoemission; the failure of Stoner theory near T_C and what replaces it; half-metals, band-insulating in one spin channel and metallic in the other [1721], and their relevance to Section 129.9.2.]

129.4 Antiferromagnetism and ferrimagnetism

129.4.1 Néel's theory

[Reserved: Néel's two-sublattice molecular-field theory, predicting a state of zero net moment with a susceptibility that peaks at the ordering temperature and then falls [781]; the Néel temperature and the Curie–Weiss law with negative θ above it; ferrimagnetism as the unequal-sublattice case, which explains magnetite and the ferrites [918] — the material class behind ferrite cores and the 1970 Nobel prize; the honest point that until 1949 this was a hypothesis about something no instrument could see.]

129.4.2 Neutron diffraction as the decisive evidence

[Reserved: the neutron's magnetic moment scattering off ordered electron spins, so that magnetic order produces extra Bragg peaks; Shull and Smart's observation of the superlattice reflections of manganese oxide, doubling the chemical unit cell and confirming Néel's two-sublattice picture directly [949]; the systematic determination of magnetic structures

that followed [991]; the diffraction formalism of Chapter 125 and the neutron sources of Chapter 107; the 1994 Nobel prize.]

Phenomenon 129.6 (Order with no net moment)**.** Some compounds order magnetically without becoming magnets. Manganese oxide has a susceptibility that rises on cooling, passes through a maximum at a definite temperature and then *falls* — the signature Néel predicted for two sublattices magnetized oppositely [781] — and above that temperature obeys Equation (129.8) with θ negative. No magnetometer can see the order itself, since the net moment is zero at every temperature. Neutron diffraction can: below the transition, manganese oxide shows extra Bragg reflections corresponding to a magnetic unit cell twice the chemical one, exactly as two alternating sublattices require, and those reflections disappear on warming through the transition [949, 991]. The related case of unequal sublattices, ferrimagnetism, accounts for magnetite and the ferrites [918]. Until 1949 this was a hypothesis about something no instrument could detect; the neutron made it an observation.

Derivation pending #259

The two-sublattice molecular-field theory, giving the Néel temperature, the susceptibility maximum at it and the negative Curie–Weiss intercept above it; and the magnetic neutron cross section, in which the neutron's own moment couples to the ordered electronic moments so that the scattering amplitude alternates between sublattices, producing reflections at the reciprocal-lattice vectors of the doubled magnetic cell that are absent from the nuclear pattern

129.4.3 Superexchange and RKKY coupling

[Reserved: how manganese ions separated by an intervening oxygen couple antiferromagnetically at all — Kramers' proposal that the coupling is mediated by virtual excitations of the intervening ion [744], made quantitative by Anderson [952], with the Goodenough–Kanamori rules relating the sign of the coupling to the geometry and occupancy of the bond [1069, 1170]; the RKKY interaction, oscillating in sign with distance as conduction electrons mediate it [1050], which is the mechanism behind the interlayer coupling of Section 129.9.1; double exchange in the mixed-valence manganites [993].]

129.5 Domains and hysteresis

129.5.1 Domain structure and the Barkhausen effect

[Reserved: why a ferromagnet below T_C can have zero net moment — the dipolar field energy is lowered by breaking into domains of differing orientation, at the cost of the wall energy; Barkhausen's observation of discrete clicks in a coil wound on a slowly magnetized iron rod, the first direct evidence that magnetization proceeds in irreversible jumps [459]; jump-size statistics and their power law; the hysteresis loop, coercivity and remanence as the measured quantities, and hard against soft materials.]

129.5.2 Imaging domains

[Reserved: Bitter's technique of decorating a polished surface with a colloidal suspension of fine magnetic particles, which collect at the stray field of the walls and make the domain pattern visible under a microscope [675]; the magneto-optic Kerr effect, the rotation of the polarization of light reflected from a magnetized surface [199], together with Lorentz

microscopy and magnetic force microscopy, as the modern successors [2011]; what each measures and its resolution; the agreement between imaged wall widths and the theory of Section 129.5.3.]

129.5.3 Domain walls and anisotropy

[Reserved: the Bloch wall as the compromise between exchange, which wants a wide wall, and magnetocrystalline anisotropy, which wants a narrow one, giving widths of tens of nm *and the scaling $\delta \propto \sqrt{A/K}$ [756]; the Landau–Lifshitz equation of motion for the magnetization [756]; magnetostriction and shape anisotropy [2011]; wall pinning by defects as the microscopic origin of coercivity and of the Barkhausen jumps of Section 129.5.1.]*

129.6 Spin waves

129.6.1 Magnons and the three-halves law

[Reserved: the low-lying excitations of an ordered ferromagnet are not single flipped spins but collective precessional waves, with $\hbar\omega = Dk^2$ at long wavelength, found by Bloch [643]; their quantization as magnons, bosons whose thermal population gives the observed $M(0) - M(T) \propto T^{3/2}$ law — qualitatively different from the exponential a gapped excitation would give, and thus a test of the picture; the Holstein–Primakoff representation that makes the spin operators bosonic [852]; the antiferromagnetic case with its linear dispersion.]

Phenomenon 129.7 (The three-halves law)**.** The spontaneous magnetization of a ferromagnet falls below its zero-temperature value, at temperatures far below T_C, as a power of the temperature and not exponentially:

$$M(0) - M(T) \propto T^{3/2}\,, \tag{129.12}$$

the law Bloch obtained by treating the excitations as collective spin waves rather than as individually flipped spins [643]. The distinction is testable, because a flipped spin costs a finite energy and would give an exponentially small reduction. The excitations themselves are seen directly: inelastic neutron scattering from magnetite resolves a magnon branch and measures its dispersion [1104], which is quadratic, $\hbar\omega = Dk^2$, at long wavelength — and it is the vanishing of ω as $k \to 0$, guaranteed by the rotational symmetry the ordered state breaks, that supplies the low-energy excitations the exponent requires.

Derivation. Take the measured dispersion $\hbar\omega = Dk^2$ as given, and take the magnons to be independent bosons, which is the leading approximation at low temperature since their density is then small. Each magnon reduces the total spin projection by one unit, so the magnetization deficit is proportional to the magnon number density,

$$n_{\mathrm{m}} = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\frac{1}{\mathrm{e}^{Dk^2/k_BT} - 1}\,, \tag{129.13}$$

with no chemical potential, magnon number not being conserved. Substitute $u = Dk^2/k_BT$, so that $k = (k_BTu/D)^{1/2}$ and $k^2\mathrm{d}k = \tfrac{1}{2}(k_BT/D)^{3/2}u^{1/2}\mathrm{d}u$; the angular integral gives 4π and

$$n_{\mathrm{m}} = \frac{1}{4\pi^2}\left(\frac{k_BT}{D}\right)^{3/2}\int_0^\infty \frac{u^{1/2}\,\mathrm{d}u}{\mathrm{e}^u - 1} = \frac{\Gamma(3/2)\,\zeta(3/2)}{4\pi^2}\left(\frac{k_BT}{D}\right)^{3/2}\,, \tag{129.14}$$

which is Equation (129.12). The exponent is entirely a statement about the low-energy end of the spectrum: it is 3/2 because three-dimensional phase space contributes $k^2\mathrm{d}k$ while the energy goes as k^2, so the number of thermally accessible modes scales as $T^{3/2}$. Had the excitation spectrum a gap Δ — as a single flipped spin, held in place by its neighbours,

would have — the Bose factor would supply $\mathrm{e}^{-\Delta/k_BT}$ instead, and the measured power law excludes that picture. The fitted stiffness D from the neutron dispersion and the fitted coefficient of $T^{3/2}$ from the magnetization must agree, which closes the loop between two independent experiments. □

129.6.2 Magnons measured by neutron scattering

[Reserved: inelastic neutron scattering as the direct measurement of $\omega(\boldsymbol{k})$ for magnons, exactly as for the phonons of Chapter 125; Brockhouse's observation of spin-wave scattering in magnetite, the first such measurement [1104]; the exchange constants J read off the fitted stiffness D, closing the loop with Section 129.3.3; magnon gaps from anisotropy, and the scattering theory of Chapter 86.]

129.7 Magnetic phase transitions

[Reserved: the ferromagnetic and antiferromagnetic transitions as worked examples of the general theory in Chapter 123 — order parameter, spontaneous symmetry breaking, universality classes, and the measured critical exponents of nickel and of the model magnets, which disagree with the mean-field values implied by [339]; the divergence of the susceptibility and of the correlation length; the specific-heat anomaly; a cross-reference rather than a repetition, since the renormalization group belongs to Chapter 108.]

129.8 Magnetic resonance

129.8.1 Electron spin resonance

[Reserved: Zavoisky's observation of resonant microwave absorption by paramagnetic salts at $\hbar\omega = g\mu_B B$, the first magnetic resonance experiment [885]; the measured g-tensor as a probe of the local crystal field and of orbital quenching; hyperfine structure and its use in identifying defects and radicals; ferromagnetic resonance and its relation to the Landau–Lifshitz dynamics [756].]

129.8.2 Nuclear magnetic resonance

[Reserved: the independent 1946 detections of nuclear magnetic resonance in condensed matter, by absorption in paraffin [894] and by induction in water [887]; the Bloch equations, T_1 and T_2; the chemical shift and the Knight shift, the latter measuring the conduction-electron spin susceptibility of Section 129.1.4; the nuclear moments themselves belonging to Chapter 107; the 1952 Nobel prize, and magnetic resonance imaging as the applied consequence.]

129.9 Magnetotransport and spintronics

129.9.1 Giant magnetoresistance

[Reserved: the independent discovery, in iron–chromium multilayers, that the resistance changes by tens of percent when an applied field aligns the antiferromagnetically coupled ferromagnetic layers [1805, 1816] — an effect an order of magnitude larger than any previously known magnetoresistance, hence the name; spin-dependent scattering and the two-current model as its explanation; the RKKY oscillation [1050] that sets the interlayer coupling; the 2007 Nobel prize and the read heads that followed.]

129.9.2 Tunnelling magnetoresistance and spin transport

[Reserved: Jullière's magnetic tunnel junction, in which the tunnelling conductance between two ferromagnets depends on their relative magnetization through the spin polarizations of their densities of states [1551], and the large room-temperature ratios later obtained with crystalline barriers; spin-transfer torque as the reciprocal effect, a spin-polarized current switching a magnet [1930]; the honest boundary of this chapter — device physics is only included where it constitutes a measurement of the magnetism itself.]

Chapter 130

Superconductivity and Superfluidity

Superconductivity and superfluidity are the two places where quantum mechanics stops being microscopic. A single complex order parameter acquires a definite phase over a laboratory-sized sample, and the consequences are visible to the naked eye: a current that circulates for years without measurable decay, a magnet floating above a cooled disc, helium creeping up the wall of its own vessel. This chapter develops the phenomenological electrodynamics of London and of Ginzburg and Landau, the microscopic pairing theory of Bardeen, Cooper and Schrieffer, the energy gap and its spectroscopic signatures, the Josephson effects, and the superfluid phases of helium-4 and helium-3 — macroscopic quantum phenomena and the measured signatures that establish them.

It sits at the end of Part XII because it needs everything before it. The statistics of Chapter 120 and the indistinguishability of Chapter 83 supply Bose–Einstein condensation; the electron gas and Fermi surface of Chapter 126 supply the normal state that becomes unstable; the lattice vibrations of Chapter 125 supply the attraction that pairs the electrons; and the symmetry-breaking and critical-exponent machinery of Chapter 123 supplies the language for the transition itself. Looking forward, the same broken gauge symmetry gives the photon a mass inside a superconductor and is the condensed matter ancestor of the Higgs mechanism of Chapter 104; neutron-star interiors are believed to be superfluid (Chapter 49); and the flux quantum and the Josephson constant now define the practical volt in the SI of Chapter 2. The experimental record — from the mercury run of 1911 to modern tunnelling and metrology — is collected in Chapter 131. The standard modern monographs are Tinkham for superconductivity [1932], Khalatnikov for helium-4 [1313] and Vollhardt and Wölfle for helium-3 [1839].

Derivation pending #260

Superconductivity and Superfluidity: all derivations of this chapter are pending.

130.1 Discovery and phenomenology

130.1.1 Vanishing resistance

[Reserved: Kamerlingh Onnes' observation that the resistance of mercury falls discontinuously to immeasurably small values near 4.2 K *[374]; the distinction between a small resistance and a zero one, settled only by persistent-current decay bounds; the resulting upper limit on the resistivity, some twenty orders of magnitude below copper, and why a purely resistive account already fails — the full data are in Chapter 131.]*

Phenomenon 130.1 (Resistance vanishes at a sharp temperature)**.** Cooled through a temperature characteristic of the material — about 4.2 K for mercury — the electrical resistance of certain metals does not fall smoothly towards a residual value, as Chapter 126 would have it, but drops by more than four orders of magnitude within a few mK to a value below the resolution of the measurement [374]. The statement that the resistance is exactly zero is not read off that measurement, which can only set an upper limit; it is inferred from the persistence of a current circulating in a closed superconducting loop, whose decay time is bounded below at order 1×10^5 yr [1252] — a resistivity limit some seventeen orders of magnitude below that of copper at the same temperature. The transition is sharp, reversible, and occurs at a temperature that depends on the material but not on the measuring current so long as that current is small.

Derivation pending #261

Why the normal state cannot produce this: the temperature-independent residual resistivity that impurity scattering guarantees in any metal, so that a discontinuous drop to zero requires a new state rather than a limit of the old one; and the conversion of the persistent-current decay bound into a resistivity bound through the inductance of the loop

130.1.2 The Meissner–Ochsenfeld effect

[Reserved: the expulsion of magnetic flux from the interior of a superconductor cooled through its transition in an applied field, observed by Meissner and Ochsenfeld [724]; why this is not a consequence of perfect conductivity — a perfect conductor freezes the flux it had, a superconductor expels it, so the superconducting state is a thermodynamic state and not a history; $B = 0$ in the bulk as the defining property, perfect diamagnetism $\chi = -1$, and the thermodynamic critical field $H_c(T)$ obtained from the condensation energy.]

Phenomenon 130.2 (Flux is expelled, not merely excluded)**.** A specimen of tin or lead placed in a modest magnetic field and then cooled through its transition temperature *expels* the field from its interior: the field outside the sample rearranges as the flux is pushed out, and inside the bulk $B = 0$ [724]. The final state does not depend on the order of operations — cooling first and then applying the field gives the same interior field, namely none — so the superconducting state is a thermodynamic state of the metal and not a record of its history. This is the observation that a perfect conductor cannot reproduce. Perfect conductivity implies only that the flux already threading the sample cannot change; a perfect conductor cooled in a field would trap that field forever. Above a critical field $H_c(T)$, which vanishes at T_c and grows roughly parabolically below it, superconductivity is destroyed and the flux returns.

Derivation. The distinction is captured by replacing Ohm's law with a constitutive relation between the supercurrent and the field itself. Fritz and Heinz London postulated [757]

$$\nabla \times \boldsymbol{j}_s = -\frac{n_s e^2}{m}\,\boldsymbol{B}\,, \tag{130.1}$$

whereas the acceleration of free charges in a perfect conductor gives the same relation with $\boldsymbol{B}$ replaced by $\partial \boldsymbol{B}/\partial t$ — which is the whole of the difference. Take the curl of the magnetostatic Ampère law $\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{j}_s$ of Chapter 60, use $\nabla \times (\nabla \times \boldsymbol{B}) = \nabla(\nabla \cdot \boldsymbol{B}) - \nabla^2 \boldsymbol{B}$ together with $\nabla \cdot \boldsymbol{B} = 0$, and substitute Equation (130.1):

$$\nabla^2 \boldsymbol{B} = \frac{\boldsymbol{B}}{\lambda_L^2}\,, \qquad \lambda_L = \sqrt{\frac{m}{\mu_0 n_s e^2}}\,. \tag{130.2}$$

The only solution of Equation (130.2) bounded in the interior of a half-space is $B(x) = B(0)\mathrm{e}^{-x/\lambda_L}$: the field survives only within a penetration depth of the surface, typically $\lambda_L \approx 50\,\mathrm{nm}$, and vanishes in the bulk. Because Equation (130.2) constrains $\boldsymbol{B}$ and not $\partial\boldsymbol{B}/\partial t$, it admits no field-cooled solution with trapped flux, and the Meissner effect becomes a theorem rather than an extra assumption. What has been assumed is Equation (130.1), whose physical content is the rigidity of the many-body wavefunction against the perturbation of a vector potential; that rigidity is what the microscopic theory of Section 130.3.4 has to supply. □

130.1.3 Critical temperature, field and current

[Reserved: the critical surface in the (T, H, J) space; the empirical parabolic law $H_c(T) \approx H_c(0)[1 - (T/T_c)^2]$; Silsbee's rule relating the critical current to the field it generates at the surface [1932]; the tabulated values for the elemental superconductors [1589] and the reason no elemental T_c exceeds about $9.3\,\mathrm{K}$ at ambient pressure.]

130.1.4 The two-fluid picture

[Reserved: the Gorter–Casimir two-fluid model [739], in which a temperature-dependent fraction of the electrons is condensed and carries current without entropy; the consequent thermodynamics of the transition — a second-order transition in zero field with a jump, not a latent heat, in the specific heat; the model as the classical scaffolding that London and then BCS replaced with a microscopic wavefunction.]

130.2 Phenomenological electrodynamics

130.2.1 The London equations

[Reserved: Fritz and Heinz London's replacement of Ohm's law by a constitutive relation between the supercurrent and the vector potential [757]; the two London equations, the second of which, combined with Ampère's law from Chapter 60, yields $\nabla^2\boldsymbol{B} = \boldsymbol{B}/\lambda_L^2$ and hence exponential screening over the penetration depth $\lambda_L = \sqrt{m/\mu_0 n_s e^2}$ — the Meissner effect as a theorem rather than an observation; typical $\lambda_L \approx 50\,\mathrm{nm}$; the London rigidity of the wavefunction as the physical content, and the London moment in a rotating superconductor.]

130.2.2 Non-local response and the coherence length

[Reserved: Pippard's microwave measurements showing that the measured penetration depth in alloyed tin does not follow the local London prediction, and his non-local generalization with a coherence length ξ_0 over which the current responds to the field [1025]; the analogy to the anomalous skin effect; $\xi_0 \approx \hbar v_F/\pi\Delta$ as later derived from BCS, and the ratio $\kappa = \lambda/\xi$ that decides the type of the superconductor.]

130.2.3 Ginzburg–Landau theory

[Reserved: the complex order parameter ψ and the free energy expansion of Ginzburg and Landau [957], an application of the Landau theory of second-order transitions of Chapter 123 with a gauge-covariant gradient term; the two Ginzburg–Landau equations; the identification $|\psi|^2 = n_s$; surface energy between normal and superconducting regions and its sign change at $\kappa = 1/\sqrt{2}$; Gor'kov's derivation of the equations from BCS near T_c, which fixed the charge in the gradient term at $2e$ [1168].]

130.2.4 Type II superconductors and the Abrikosov lattice

[Reserved: Abrikosov's prediction that for $\kappa > 1/\sqrt{2}$ flux enters as a triangular lattice of quantized vortices between H_{c1} and H_{c2} rather than destroying the state at H_c [1100]; the vortex core, the circulating supercurrent and the single flux quantum each vortex carries; the direct observation of the lattice by Bitter decoration [1346]; flux pinning, critical currents and the practical superconducting magnet.]

130.3 Microscopic theory

130.3.1 The isotope effect

[Reserved: the simultaneous discovery by Maxwell [964] and by Reynolds, Serin, Wright and Nesbitt [969] that T_c of separated mercury isotopes scales as $M^{-\alpha}$ with $\alpha \approx 1/2$; why an ionic mass in an electronic transition temperature is decisive evidence that the lattice mediates the interaction, since $M^{-1/2}$ is the scaling of the phonon frequencies of Chapter 125; departures from $\alpha = 1/2$ in transition metals and their interpretation.]

Phenomenon 130.3 (The transition temperature depends on the mass of the ions)**.** Separated isotopes of mercury, chemically identical and differing only in the mass of the nucleus, have different superconducting transition temperatures, scaling as $T_c \propto M^{-\alpha}$ with α close to $1/2$ — found simultaneously and independently by two groups [964, 969]. The result is decisive out of proportion to its size. Superconductivity is a property of the conduction electrons, and the ionic mass enters no electronic energy scale; but it is exactly the mass that sets the phonon frequencies of Chapter 125, which scale as $M^{-1/2}$. The lattice must therefore mediate the interaction that produces the state. Departures from $\alpha = 1/2$ are measured in several transition metals and are informative in their turn.

Derivation pending #262

Why an electron–phonon-mediated pairing energy carries the phonon frequency as its prefactor, so that the transition temperature inherits the ionic mass dependence; and the corrections — from the screened Coulomb repulsion between the electrons, and from strong coupling — that reduce the exponent below one half without removing the effect

130.3.2 The energy gap

[Reserved: the exponentially small electronic specific heat at low temperature as the first sign of a gap [1034]; the threshold in far-infrared absorption measured by Glover and Tinkham, which located the gap at $2\Delta \approx 3.5\,k_B T_c$ [1112]; the same gap seen directly in the tunnelling current–voltage characteristic [1180]; the gap as an excitation gap for quasiparticles, not a gap in the single-particle band structure of Chapter 126.]

Phenomenon 130.4 (An energy gap in the excitation spectrum)**.** Three unrelated measurements agree that a superconductor cannot be excited at arbitrarily small energy. Its electronic heat capacity falls below T_c not as the linear γT of a normal metal (Chapter 120) but exponentially, as $\mathrm{e}^{-\Delta/k_B T}$ [1034]. Far-infrared radiation is transmitted without absorption below a threshold photon energy and absorbed above it [1112]. And the current–voltage characteristic of a tunnel junction between a superconductor and a normal metal is flat until a threshold voltage, above which it turns sharply upward, so that $\mathrm{d}I/\mathrm{d}V$ maps the quasiparticle density of states directly [1180]. All three return the same gap, close to

$$2\Delta(0) \approx 3.5\,k_B T_c\,, \tag{130.3}$$

a ratio that is roughly the same across the elemental superconductors. The gap is an excitation gap in a metal whose band structure has none: the states are there, and it is the correlated state that is expensive to disturb.

Derivation pending #263

The gap as the energy to break one pair, obtained from the self-consistent gap equation of the pairing theory; the quasiparticle dispersion, whose minimum is the gap and whose density of states diverges at the gap edge, giving the observed tunnelling threshold and infrared edge; and the exponential activation of the heat capacity, together with the derivation of the measured ratio of twice the gap to the transition temperature as a parameter-free consequence of weak coupling

130.3.3 Cooper pairing

[Reserved: Cooper's theorem — two electrons added above a filled Fermi sea form a bound state for arbitrarily weak attraction, because the density of states at the Fermi surface is finite [1084]; the logarithmic singularity and the non-perturbative binding energy $\sim \hbar\omega_D \exp(-1/N(0)V)$, which no order of perturbation theory reproduces; the Fröhlich phonon-exchange interaction as the source of the attraction [956]; the pair as an object of size 100 nm*, overlapping many others, and so not a molecule.]*

130.3.4 The BCS ground state

[Reserved: the variational pair-condensate wavefunction of Bardeen, Cooper and Schrieffer [1102]; the gap equation and its self-consistent solution; the Bogoliubov quasiparticle spectrum $E_k = \sqrt{\epsilon_k^2 + \Delta^2}$; the parameter-free predictions — $2\Delta(0) = 3.53\, k_B T_c$, the specific-heat jump $\Delta C/\gamma T_c = 1.43$, the isotope exponent $1/2$ — and the coherence factors that make ultrasonic attenuation fall while nuclear spin relaxation first rises [1124, 1169], a sign difference that is among the sharpest confirmations of the theory.]

130.3.5 Broken gauge symmetry

[Reserved: the superconducting state as spontaneously broken U(1) *symmetry; Nambu's demonstration that gauge invariance is compatible with the gap and the collective modes it requires [1187]; Anderson's observation that the would-be Goldstone mode is absorbed by the electromagnetic field, leaving a massive photon — the Meissner effect read as a photon mass $m_\gamma = \hbar/\lambda_L c$ [1245] — and the direct line from there to the Higgs mechanism of Chapter 104.]*

130.4 Macroscopic quantum coherence

130.4.1 Off-diagonal long-range order

[Reserved: Penrose and Onsager's criterion for condensation in an interacting system — a macroscopic eigenvalue of the one-particle density matrix [1094]; Yang's extension to the two-particle density matrix, which is what condenses in a superconductor, and his identification of off-diagonal long-range order as the single property shared by superfluid helium and every superconductor [1244]; the order parameter as an expectation value of a field operator, connecting to Chapter 87.]

130.4.2 Flux quantization

[Reserved: single-valuedness of the condensate phase around a ring forces the enclosed flux to be an integer multiple of $\Phi_0 = h/2e = 2.067\,833\,848 \times 10^{-15}$ Wb; the independent measurements of Deaver and Fairbank [1201] and of Doll and Näbauer [1202] in 1961, whose value of Φ_0 was half the naive h/e and thereby measured the carrier charge as $2e$ — pairing established without any appeal to BCS; the Little–Parks oscillations of T_c with enclosed flux as the same quantization seen in a transition temperature [1236].]

Phenomenon 130.5 (Flux quantization, and the carrier charge it measures)**.** The magnetic flux trapped by a hollow superconducting cylinder is not continuous: it takes only integer multiples of a fixed quantum, and the measured quantum is $\Phi_0 = 2.067\,833\,848 \times 10^{-15}$ Wb [1201, 1202, 2431]. The value is half the naive h/e. Since the quantization argument fixes the quantum as h divided by the charge of the entity whose phase must be single-valued, the measurement determines that charge to be $2e$ — pairing, established in 1961 by two independent groups, from a measurement that makes no appeal to any microscopic theory. The same periodicity in the enclosed flux appears in the transition temperature of a thin-walled cylinder [1236].

Derivation. Let the superconducting state be described by a single complex field $\psi = |\psi|\, \mathrm{e}^{\mathrm{i}\theta}$ for carriers of charge q and mass m^*. The gauge-invariant current is

$$\boldsymbol{j}_s = \frac{q\,|\psi|^2}{m^*}\,(\hbar\nabla\theta - q\boldsymbol{A})\ . \tag{130.4}$$

Take a contour C running round the hole of the cylinder but lying deep inside the superconducting material, at a depth of many penetration lengths, where by Equation (130.2) the field and hence the current vanish. Setting $\boldsymbol{j}_s = 0$ in Equation (130.4) gives $\hbar\nabla\theta = q\boldsymbol{A}$ along C, and integrating round it,

$$\hbar \oint_C \nabla\theta \cdot \mathrm{d}\boldsymbol{l} = q \oint_C \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{l} = q\Phi\,, \tag{130.5}$$

where Φ is the total flux enclosed. Single-valuedness of ψ forces the left-hand side to be $2\pi\hbar n$ with n an integer, so $\Phi = nh/q$. Nothing else enters: not the geometry, not the material, not the temperature. Reading the measurement backwards, $q = h/\Phi_0 = 3.204\,353\,268 \times 10^{-19}$ C, which is $2e$ to the accuracy of the experiment. The argument is the same one that quantizes circulation in a neutral superfluid (Phenomenon 130.8); the charge is what turns a circulation into a flux. □

130.4.3 The Josephson effects

[Reserved: Josephson's prediction that a supercurrent $I = I_c \sin\varphi$ flows across a thin insulating barrier at zero voltage, and that a constant voltage drives the phase at $\mathrm{d}\varphi/\mathrm{d}t = 2eV/\hbar$, giving an alternating current of frequency $2eV/h$ [1231]; the observation of the dc effect by Anderson and Rowell [1246]; Shapiro's microwave-induced constant-voltage steps as the observation of the ac effect [1266]; the Fraunhofer dependence of I_c on applied flux.]

Phenomenon 130.6 (Supercurrent through an insulator)**.** Two superconductors separated by an insulating barrier a few nanometres thick pass a current at *zero* voltage, up to a critical value I_c [1246]. That the barrier is not simply shorted is shown by the behaviour of I_c in a magnetic field: it does not fall monotonically but oscillates, vanishing whenever the flux threading the junction is an integer multiple of the quantum of Phenomenon 130.5, in the pattern of a single-slit diffraction envelope. Held at a constant voltage V instead, the junction carries an alternating current of frequency $2eV/h$; irradiated with microwaves of

frequency f, its characteristic develops steps of constant voltage spaced by $hf/2e$ [1266], which is how the alternating current is observed. Both effects were predicted before they were seen [1231]. A frequency and a voltage are thereby locked together by fundamental constants alone, which is the basis of the practical volt.

Derivation pending #264

Two weakly coupled condensates described by their amplitudes and phases: the coupled equations of motion for the pair, in which the current across the barrier is proportional to the sine of the phase difference and the phase difference evolves at a rate proportional to the electrochemical potential difference; the resulting zero-voltage supercurrent and the alternating current at the Josephson frequency; and the spatial variation of the phase difference across a junction in a field, whose integral gives the diffraction-like dependence of the critical current on the enclosed flux

130.4.4 SQUIDs and quantum metrology

[Reserved: two junctions in a ring interfere, and the critical current is periodic in the enclosed flux with period Φ_0 — the quantum interference observed by Jaklevic, Lambe, Silver and Mercereau [1285]; the superconducting quantum interference device as the most sensitive magnetometer built, with a noise floor of a few fT *in a one-hertz band [2059]; the Josephson constant $K_J = 2e/h = 483\,597.8484\,\mathrm{GHz/V}$ as an exact number in the revised SI [2333, 2431], and hence the practical realization of the volt.]*

130.5 Superconductors beyond the BCS regime

130.5.1 The cuprates

[Reserved: Bednorz and Müller's report of a superconducting onset near 35 K *in the Ba–La–Cu–O system [1768], beyond any then-known T_c; the rapid confirmation and the jump above the boiling point of nitrogen at* 93 K *in Y–Ba–Cu–O [1802]; the layered copper-oxide structure, the doped Mott insulator as the parent compound, and the phase diagram with its pseudogap and strange-metal regions.]*

130.5.2 Pairing symmetry in the cuprates

[Reserved: the phase-sensitive experiments that established $d_{x^2-y^2}$ pairing — the corner SQUID of Wollman and co-workers [1877] and the tricrystal ring experiment of Tsuei and co-workers [1889], which trapped a half-integer flux quantum where the order parameter changes sign; nodal quasiparticles and the resulting power-law rather than exponential low-temperature thermodynamics.]

130.5.3 Other families

[Reserved: magnesium diboride at 39 K*, a two-gap phonon-mediated superconductor [2018]; the iron-based layered pnictides [2130]; the high-pressure hydrides, where conventional phonon-mediated pairing reaches* 203 K *in sulfur hydride near* 155 GPa *[2252] and higher still in lanthanum superhydride [2340]; heavy-fermion and organic superconductors; the honest caveat that claims of near-ambient superconductivity have repeatedly failed replication.]*

130.5.4 What remains unexplained

[Reserved: an explicit statement of what is established — that the cuprates are superconductors, that their carriers are pairs of charge 2e, that the order parameter has d-wave symmetry — and what is not: no mechanism for cuprate pairing commands consensus, and no theory predicts T_c for a given compound. The problem is listed among the open ones in Chapter 133; the renormalization group of Chapter 108 explains why a strongly correlated normal state resists the weak-coupling treatment that worked for lead.]

130.6 Superfluid helium-4

130.6.1 The lambda transition and the discovery of superflow

[Reserved: the specific-heat anomaly at 2.17 K that gives the transition its name and separates He I from He II; the simultaneous reports of frictionless flow through narrow channels by Kapitza [817] and by Allen and Misener [809], published back to back; the fountain effect and the thermomechanical connection between heat and flow [808]; film creep; the modern microgravity measurement of the specific-heat exponent at the lambda point [2054], one of the most precise critical exponents ever obtained, feeding Chapter 123.]

Phenomenon 130.7 (Superflow in liquid helium-4)**.** Liquid helium-4 cooled below 2.17 K passes through channels and packed powders so narrow that the ordinary liquid would scarcely move at all, with no measurable pressure drop and hence no measurable viscosity — reported simultaneously by Kapitza and by Allen and Misener [809, 817]. The same liquid, measured by the drag on a rotating body, still exhibits a viscosity of the ordinary order, so the two experiments do not agree about what the viscosity is; the transition is marked independently by a specific-heat anomaly of a shape that gives it its name. Two further observations rule out any mundane explanation: helium-4 flows as a film up and over the wall of its container until the levels inside and outside are equal, and heating one side of a superleak drives a flow towards the heat rather than away from it, so that a jet of liquid can be raised by a lamp.

Derivation pending #265

The two-fluid description in which the liquid behaves as an interpenetrating mixture of a normal component carrying all the entropy and all the viscosity and a superfluid component carrying neither, and the reconciliation of the capillary and rotating-body measurements that it provides; the thermomechanical effect as the statement that a temperature difference and a pressure difference are equivalent driving forces for the superfluid component; and Landau's criterion, in which an excitation spectrum with no low-lying free-particle branch forbids dissipation below a critical velocity — which is why an ideal Bose gas would condense without being a superfluid

130.6.2 Bose–Einstein condensation as the mechanism

[Reserved: London's identification of the lambda transition with the Bose–Einstein condensation of Chapter 120 [818], and the near-coincidence of the ideal-gas condensation temperature with the observed 2.17 K; why the identification is nevertheless not an explanation — the interacting liquid has a condensate fraction of only about 7 % at $T = 0$, so the superfluid fraction and the condensate fraction are different quantities; the contrast with the dilute atomic gases of Chapter 121, where they nearly coincide; helium-4 as a boson, helium-3 as a fermion (Chapter 83).]

130.6.3 The two-fluid model

[Reserved: Tisza's proposal that He II behaves as an interpenetrating mixture of a normal viscous component and a superfluid component with zero entropy and zero viscosity [822]; Landau's reformulation in terms of the excitation spectrum, which supplies the normal density without invoking a condensate population [858]; Andronikashvili's rotating stack of discs, which measured $\rho_n(T)$ directly and confirmed the split [886]; second sound as a counter-oscillation of the two components, predicted by Landau and observed by Peshkov [876].]

130.6.4 The excitation spectrum and rotons

[Reserved: Landau's spectrum [858] — a linear phonon branch at small momentum and a minimum, the roton, at $p_0/\hbar \approx 1.9 \times 10^{10}/\mathrm{m}$ with gap $\Delta/k_B \approx 8.6\,\mathrm{K}$; the Landau criterion $v_c = \min_p \varepsilon(p)/p$ for the breakdown of superflow, which is why an ideal Bose gas would not be a superfluid; Feynman's microscopic account of the spectrum and of vortex lines [1066]; the spectrum measured directly by inelastic neutron scattering [1210].]

130.6.5 Quantized circulation and vortices

[Reserved: Onsager's proposal that circulation in a superfluid is quantized in units of $h/m_4 \approx 9.97 \times 10^{-8}\,\mathrm{m^2/s}$ [943], and Feynman's development of the vortex-line picture [1066]; Vinen's measurement of single quanta of circulation using the vibrating-wire method [1220]; the rotating-bucket experiment and the vortex array; quantum turbulence as a tangle of such lines.]

Phenomenon 130.8 (Circulation in a superfluid is quantized)**.** The circulation trapped round a fine wire immersed in superfluid helium-4, measured through the splitting it induces between the two circular modes of the vibrating wire, does not take arbitrary values. It is found in integer multiples of

$$\frac{h}{m_4} \approx 9.97 \times 10^{-8}\,\mathrm{m^2/s}\,, \tag{130.6}$$

with m_4 the mass of the helium-4 atom, and it changes between these values by whole steps and not continuously [943, 1066, 1220]. Rotating a vessel of superfluid therefore does not set it turning as a body: the angular momentum is taken up by an array of discrete vortex lines. This is the neutral counterpart of Phenomenon 130.5, and the appearance of the bare atomic mass m_4, rather than twice it, records that the condensing entity here is a single atom and not a pair.

Derivation. The argument is that of Phenomenon 130.5 with the charge set to zero. Writing the condensate field as $\psi = |\psi|\,\mathrm{e}^{\mathrm{i}\theta}$, the superfluid velocity is $\boldsymbol{v}_s = (\hbar/m_4)\nabla\theta$, so the flow is irrotational wherever $\psi \neq 0$. Round any closed contour on which ψ does not vanish, single-valuedness of ψ requires θ to advance by $2\pi n$, hence

$$\oint \boldsymbol{v}_s \cdot \mathrm{d}\boldsymbol{l} = \frac{\hbar}{m_4}\, 2\pi n = n\,\frac{h}{m_4}\,, \qquad n \in \mathbb{Z}\,. \tag{130.7}$$

The quantum in Equation (130.7) depends on the atomic mass and on Planck's constant and on nothing else, which is why the numerical value in Equation (130.6) can be quoted without reference to the temperature, the pressure or the apparatus. Two consequences are directly observed: circulation cannot decay smoothly, so a persistent flow is metastable against small perturbations; and non-zero circulation requires ψ to vanish along a line inside the contour, which is the vortex core. □

130.7 Superfluid helium-3

130.7.1 Discovery of the superfluid phases

[Reserved: Osheroff, Richardson and Lee's observation of two anomalies on the melting curve of helium-3 near 2.7 mK *and* 2 mK*, initially attributed to the solid [1480]; Leggett's identification of them as transitions of the liquid into anisotropic superfluid phases with spin-triplet, p-wave Cooper pairs [1478]; the A and B phases and the phase diagram in temperature, pressure and field.]*

130.7.2 The order parameter and its consequences

[Reserved: a 3×3 *complex matrix order parameter and the broken relative spin–orbit symmetry, developed in [1839]; the resulting textures, half-quantum vortices and nuclear-magnetic-resonance frequency shifts as the experimental handles; helium-3 as the laboratory in which unconventional pairing was first understood, and the template for the p-wave and d-wave superconductors of Section 130.5.2; Fermi-liquid theory as the necessary normal-state description.]*

130.8 Where the phenomena reappear

[Reserved: a short closing survey. Neutron-star interiors as neutron superfluids and proton superconductors, invoked to explain pulsar glitches [1390] and taken up in Chapter 49; the superfluidity of dilute trapped atomic gases (Chapter 121); the quantum Hall states of Chapter 128 as a different macroscopic quantum order; and, in high-energy physics, the same broken-symmetry mathematics in Chapter 104 and in the chiral condensate of Chapter 102. The common thread throughout is off-diagonal long-range order [1244], not any particular pairing mechanism.]

Chapter 131

Experiment: Superconductivity

Zero resistance and flux expulsion as measured. This chapter takes the phenomena of Chapter 130 and records the measurements that establish them: the Leiden helium liquefaction that made the temperature range accessible [343], the 1911 mercury resistance run [374], the persistent-current bounds that convert "very small resistance" into a quantitative upper limit [1252], the flux-expulsion experiment of Meissner and Ochsenfeld [724], the isotope effect that identified the phonon as the glue [964, 969], the tunnelling spectroscopy that measured the gap [1180], the flux quantum that measured the carrier charge [1201], the Josephson verification [1246], and the 1986 discovery run that broke the temperature ceiling [1768].

Each of these will be written up in the structured `experiment` environment of the house style, with its apparatus, procedure, observations in SI units with uncertainties, interpretation and primary references, so that a reader can see exactly which number carries which inference. The chapter is placed immediately after the theory it tests and cross-references it section by section; the low-temperature techniques it depends on belong to Chapter 118 and Chapter 54, and the metrological use of the results feeds Chapter 2.

Derivation pending #266

Experiment: Superconductivity: all derivations and data tables of this chapter are pending.

131.1 Historical context and the claim under test

131.1.1 The race to liquefy the permanent gases

[Reserved: Dewar's liquefaction of hydrogen and the vacuum flask [280]; Kamerlingh Onnes' liquefaction of helium in 1908, reaching 4.2 K *at atmospheric pressure and about* 1.5 K *by pumping [343]; why the new temperature decade, and not any new idea about metals, is what made the discovery possible.]*

131.1.2 Three rival predictions for the low-temperature resistance

[Reserved: the state of the question in 1911 — Matthiessen's empirical rule, that impurity and thermal contributions to the resistivity simply add, predicted saturation at a finite residual value [160]; a rival expectation held that conduction electrons would condense onto their ions and the resistance rise without bound; a third, that it would fall smoothly to zero. The experiment discriminates, and none of the three anticipated a discontinuity.]

131.1.3 Mercury as the test material

[Reserved: why mercury — repeated distillation gives the purest available specimen, and freezing it in a capillary yields a thin, high-resistance thread with well-defined geometry; the sample geometry and the resistance at 273 K *as the normalization used throughout the 1911 papers [374].]*

131.2 Apparatus

131.2.1 The Leiden cryostat and the helium cycle

[Reserved: the cascade liquefier, the Hampson–Linde regenerative cycle, the helium bath and its vapour-pressure thermometry [343]; the temperature scale in use and its uncertainty, which is the dominant systematic in the 1911 result; modern replacements — pumped helium-3, dilution refrigerators and adiabatic nuclear demagnetization — and the temperature ranges each reaches.]

131.2.2 Resistance measurement

[Reserved: the four-terminal method, separating current leads from voltage leads so that lead and contact resistance do not enter; galvanometer sensitivity and the resulting resolution floor, which is what sets the upper limit quoted rather than any property of the sample [374]; the modern equivalent using a superconducting quantum interference device as a null detector [2059].]

131.2.3 Magnetometry for flux expulsion

[Reserved: the pick-up coils and ballistic galvanometer with which Meissner and Ochsenfeld mapped the field distribution outside tin and lead cylinders [724]; the modern flux transformer coupled to a SQUID [2059]; specimen geometry and the demagnetizing factor, without which the intermediate state makes the measured susceptibility depend on shape.]

131.2.4 Tunnel junctions

[Reserved: the evaporated aluminium–oxide–lead sandwich, with an oxide barrier of order 2 nm*, used by Giaever to measure the gap [1180]; the same structure at lower barrier resistance as the Josephson junction of Anderson and Rowell [1246]; microwave coupling for the ac effect [1266]; the ring geometries used for flux quantization [1201, 1202] and quantum interference [1285].]*

131.3 Procedure

131.3.1 The 1911 mercury run

[Reserved: cooling in steps with the resistance read at each plateau; the sequence of communications from Leiden through 1911 in which the resistance was first reported to fall faster than expected, then to vanish, then to vanish discontinuously at a sharply defined temperature [374]; the short-circuit hypothesis that had to be excluded, and the control measurements that excluded it.]

131.3.2 Field-cooled and zero-field-cooled protocols

[Reserved: the protocol that makes the Meissner experiment decisive — apply the field first, then cool through T_c, and watch the flux leave [724]; contrast with cooling first and then applying the field, which a perfect conductor also passes; the reversibility check on cycling, which is the evidence that the superconducting state is a thermodynamic phase.]

Phenomenon 131.1 (The field-cooled protocol)**.** Meissner and Ochsenfeld mapped the field outside tin and lead cylinders with pick-up coils and a ballistic galvanometer, under two protocols that differ only in their order of operations. Cooled through T_c in zero field and then exposed to a field, the specimen excludes it. Exposed to the field first and then cooled through T_c, the specimen *expels* the flux it already contained, and the exterior field redistributes accordingly [724]. The two protocols end in the same state, and cycling the temperature or the field reproduces it. Only the second protocol discriminates, and it is the whole content of the experiment.

Derivation. Suppose the only property of the low-temperature state were infinite conductivity. A finite current density in a medium of zero resistivity requires $\boldsymbol{E} = 0$ inside it, and Faraday's law of Chapter 60 then gives

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} = 0\,, \tag{131.1}$$

so the interior flux is frozen at whatever value it had at the instant the resistance vanished. Equation (131.1) is a constraint on the time derivative alone, and it therefore makes the final state depend on the path taken to reach it: cooled in zero field the specimen would keep $B = 0$, cooled in a field it would keep $B = B_{\text{applied}}$ indefinitely. The observation is that the field-cooled specimen ends with $B = 0$ in its interior, exactly as the zero-field-cooled one does. That excludes Equation (131.1) as a complete description and requires a constitutive relation constraining $\boldsymbol{B}$ itself, which is the London relation of Chapter 130. The wider consequence is thermodynamic rather than electromagnetic: a state whose properties are independent of the history by which it was prepared is a phase, so the free energies of the normal and superconducting states may be compared, and the critical field measures their difference. □

131.3.3 Isotope separation and the mercury series

[Reserved: preparation of mercury samples of mean mass number from about 198 to 204 by isotope separation; measurement of T_c for each under otherwise identical conditions [964, 969]; the systematic errors that dominate — purity, strain and thermometry — and why two independent groups publishing back to back matters for a claim this consequential.]

131.4 Observations

131.4.1 The mercury transition

[Reserved: a table of resistance against temperature through the transition, in ohms and kelvin with uncertainties, normalized to the 273 K *value; the transition width of a few* mK*; the observed T_c of about* 4.2 K *and the drop by more than four orders of magnitude across it [374]; the resistance after the transition quoted as an upper limit set by the galvanometer, not as a measurement.]*

Phenomenon 131.2 (The 1911 mercury run)**.** Cooling a frozen mercury thread in the Leiden helium cryostat and reading its resistance at each plateau, Onnes found the resistance to fall, near 4.2 K, by more than four orders of magnitude within a temperature interval

of a few mK, to a value his galvanometer could not distinguish from zero [374]. The three features that make the run decisive are the sharpness, the size of the drop, and the fact that the resistance above the transition behaves normally. What the measurement establishes on its own is an *upper limit* on the resistance below T_c, set by the sensitivity of the detector and not by any property of the specimen — the "zero" of the textbooks is an inference, and the honest reading of the run is that the mercury sample became a conductor whose resistance the apparatus could not detect.

Derivation pending #267

The inference chain from the raw run: how a four-terminal measurement removes lead and contact resistance so that the reading belongs to the specimen; how the galvanometer sensitivity and the measuring current together convert into the quoted upper limit on the resistance; and the control observations that exclude the obvious alternative, a short circuit developing on cooling, by their dependence on measuring current and on applied magnetic field

131.4.2 Bounds on the residual resistivity

[Reserved: the persistent-current experiment — a superconducting solenoid short-circuited on itself, its field monitored for months; the decay-time lower bound of order 1×10^5 yr *reported by File and Mills [1252], which converts to a resistivity upper limit some seventeen orders of magnitude below that of copper at the same temperature; the honest statement that "zero" is an inference from an upper limit plus the theory of Chapter 130, not a direct measurement.]*

Phenomenon 131.3 (A current that does not decay)**.** A superconducting solenoid carrying a current and then short-circuited on itself maintains that current, monitored through its magnetic field, with no detectable decay over months of observation. The bound so obtained on the decay time is of order 1×10^5 yr [1252], which is the strongest available statement that the resistance is not merely small. Converted into a resistivity, the limit lies some seventeen orders of magnitude below that of copper at the same temperature. Note the structure of the claim: this is an experiment that reports how much of nothing happened, and it is therefore the persistence, not the sensitivity of any voltmeter, that carries the inference.

Derivation. A closed loop of self-inductance L and series resistance R carrying current I with no applied electromotive force obeys

$$L\frac{\mathrm{d}I}{\mathrm{d}t} + RI = 0 \quad \Longrightarrow \quad I(t) = I(0)\,\mathrm{e}^{-t/\tau}\,, \qquad \tau = \frac{L}{R}\,. \tag{131.2}$$

Observing over a time t that the current has not changed by more than a fraction ε bounds the decay by $t/\tau < -\ln(1-\varepsilon) \approx \varepsilon$, that is $\tau > t/\varepsilon$, and therefore

$$R < \frac{\varepsilon L}{t}\,. \tag{131.3}$$

The measurement is powerful because both factors on the right of Equation (131.3) work in its favour: ε is set by the sensitivity of a magnetometer to a *fractional* change, which is far better than any absolute voltage measurement, and t can be made arbitrarily long simply by waiting. Dividing R by the geometric factor of the winding converts Equation (131.3) into a resistivity, which is the seventeen-order-of-magnitude figure quoted. What remains true is that no experiment can measure a resistance to be exactly zero; the statement that it is exactly zero comes from the theory of Chapter 130, in which a gapped condensate has no state into which to scatter, and the experiment bounds the alternative. □

131.4.3 Critical temperatures and critical fields

[Reserved: a booktabs table of T_c and $\mu_0 H_c(0)$ for the elemental superconductors and the common alloys — aluminium, tin, lead, niobium, niobium–titanium, niobium–tin — in kelvin and tesla with uncertainties, from the critically evaluated compilation [1589]; the upper critical field H_{c2} for the type II entries [1100]; the parabolic $H_c(T)$ fit and its residuals.]

131.4.4 The isotope exponent

[Reserved: T_c against mean isotopic mass for mercury on logarithmic axes, and the fitted exponent $\alpha = 0.50 \pm 0.03$ in $T_c \propto M^{-\alpha}$ [964, 969]; tabulated exponents for other elements, including the anomalously small values in ruthenium, osmium and zirconium, and what those depress the naive reading to; the inference that the lattice sets the energy scale.]

Phenomenon 131.4 (The measured isotope exponent)**.** Mercury samples of mean mass number from about 198 to 204, prepared by isotope separation and measured under otherwise identical conditions, have transition temperatures falling with mass. Plotted logarithmically, $\log T_c$ against $\log M$ is a straight line of slope $-\alpha$ with

$$T_c \propto M^{-\alpha}, \qquad \alpha = 0.50 \pm 0.03, \tag{131.4}$$

reported back to back by two groups working independently [964, 969]. The independence matters for a claim of this consequence, and so does the smallness of the effect: a three per cent change in mass produces a change in T_c of about one and a half per cent, against systematic errors from sample purity, strain and thermometry of comparable size. Several transition metals give exponents well below one half.

Derivation pending #268

The inference from the measured isotope exponent: that a lattice frequency scales as the inverse square root of the ionic mass with the force constants, which are electronic, held fixed, so that an exponent of one half identifies the phonon energy as the prefactor of the pairing scale; and the corrections from the screened Coulomb repulsion, which enter the exponent and depress it below one half without altering the identification

131.4.5 The gap

[Reserved: three independent determinations to be tabulated together in reduced units $2\Delta(0)/k_B T_c$ — the exponential low-temperature electronic specific heat of vanadium [1034], the far-infrared absorption edge [1112], and the tunnelling conductance [1180], whose $\mathrm{d}I/\mathrm{d}V$ maps the quasiparticle density of states directly; the BCS value 3.53 and the measured spread; the Hebel–Slichter peak in nuclear spin relaxation [1169] against the fall in ultrasonic attenuation [1124] as the coherence-factor test.]

Phenomenon 131.5 (Three probes, one gap)**.** The gap of a superconductor is measured by three techniques that share no apparatus and no systematic errors: the exponential fall of the electronic specific heat at low temperature [1034], the threshold in far-infrared absorption [1112], and the threshold voltage in the tunnelling characteristic of a thin oxide junction, whose differential conductance maps the quasiparticle density of states [1180]. Expressed in the reduced form $2\Delta(0)/k_B T_c$ the three agree with one another, and cluster about the value 3.5 across the weak-coupling elemental superconductors. A related pair of measurements tests the theory more sharply than the gap does: on cooling through T_c the ultrasonic attenuation falls immediately [1124] while the nuclear spin relaxation rate first *rises* to a peak before falling [1169]. Two processes governed by the same excitation spectrum move in opposite directions, and the sign difference is predicted.

Derivation pending #269

The coherence factors: that the matrix element for a process between quasiparticle states is a sum of two amplitudes whose relative sign depends on the behaviour of the perturbation under time reversal, so that the two combinations cancel and reinforce respectively; and the resulting opposite temperature dependences of ultrasonic attenuation and nuclear spin relaxation just below the transition, against the divergent density of states at the gap edge that both share

131.4.6 The flux quantum

[Reserved: trapped flux against applied field for a superconducting cylinder, showing steps of constant height; the measured quantum $\Phi_0 = 2.07 \times 10^{-15}\,\mathrm{Wb}$ to within a few percent in both 1961 experiments [1201, 1202]; the arithmetic that turns this into a carrier charge of $2e$, and why that is direct evidence of pairing independent of any microscopic theory; the Little–Parks periodicity as a cross-check [1236].]

Phenomenon 131.6 (The measured flux quantum is half the naive value)**.** Trapped flux measured against applied field for a small superconducting cylinder does not follow the applied field but rises in steps of constant height. Two independent 1961 experiments, using different geometries and different detection schemes, measured the step to be $\Phi_0 = 2.07 \times 10^{-15}\,\mathrm{Wb}$ to within a few per cent [1201, 1202]. The number is half of h/e, and the factor of two is the point of the measurement.

Derivation. Single-valuedness of the condensate phase round the cylinder quantizes the enclosed flux in units of h/q, where q is the charge of the entity whose phase is in question — the argument is given in Chapter 130 and uses no microscopic theory. Reading it backwards, the measurement determines q:

$$q = \frac{h}{\Phi_0} = \frac{6.626\,070\,15 \times 10^{-34}\,\mathrm{J\,s}}{2.067\,833\,848 \times 10^{-15}\,\mathrm{Wb}} = 3.204\,353\,268 \times 10^{-19}\,\mathrm{C} = 2e\,, \tag{131.5}$$

using the exact 2019 values of h and e [2333, 2431]. At the several-per-cent accuracy of the 1961 experiments this discriminates cleanly between $q = e$ and $q = 2e$, which differ by a factor of two, and it selects the pair. That inference is worth isolating: it establishes that the charge carriers of a superconductor have twice the electronic charge, from a measurement of a flux and a ring, four years after the pairing theory was proposed and independently of it. The periodicity of the transition temperature of a thin-walled cylinder in the enclosed flux, with the same period, is an independent check on the same quantity [1236]. □

131.4.7 Josephson currents and Shapiro steps

[Reserved: the zero-voltage supercurrent through an oxide barrier and its Fraunhofer-pattern suppression by an applied field, which is what distinguished the effect from a metallic short [1246]; the constant-voltage steps induced by microwave irradiation, spaced by $hf/2e$ [1266]; frequency-to-voltage conversion at $483.6\,\mathrm{MHz/\mu V}$ and the agreement of $2e/h$ so obtained with other determinations.]

Phenomenon 131.7 (Constant-voltage steps locked to a frequency)**.** A tunnel junction irradiated with microwaves of frequency f develops, in its current–voltage characteristic, a series of vertical steps at which the voltage does not change while the current does. The steps sit at

$$V_n = n\,\frac{hf}{2e}\,, \qquad n \in \mathbb{Z}\,, \tag{131.6}$$

and their spacing depends on the microwave frequency and on nothing else — not the material, the junction area, the temperature or the barrier thickness [1266]. That the zero-voltage current seen without irradiation is a genuine supercurrent and not a metallic short is established separately, by its suppression in a magnetic field in a diffraction-like pattern [1246]. A frequency, which is the most accurately measurable quantity in physics, is thereby converted into a voltage by fundamental constants alone.

Derivation. The second Josephson relation, $\mathrm{d}\varphi/\mathrm{d}t = 2eV/\hbar$, makes the phase difference across the junction advance at a rate set by the voltage, so a constant voltage V produces a supercurrent $I_c \sin\varphi$ alternating at the Josephson frequency $f_J = 2eV/h$. Under irradiation the junction is driven at f as well, and the average current is non-zero only when the two are commensurate, $f_J = nf$; the junction then phase-locks and holds that average voltage over a finite range of bias current, which is the step. Setting $2eV_n/h = nf$ gives Equation (131.6). The conversion factor is the Josephson constant

$$K_J = \frac{2e}{h} = 483\,597.8484\,\mathrm{GHz/V} = 483.597\,848\,4\,\mathrm{MHz/\mu V}\,, \tag{131.7}$$

exact in the present SI because h and e are exact [2333, 2431]. Two properties make Equation (131.6) a metrological standard rather than a curiosity. The step voltage is a *ratio* of a frequency to a constant, so it inherits the accuracy of the frequency reference; and nothing about the junction appears in it, so thousands of junctions may be placed in series to reach a usable voltage without introducing a single sample-dependent parameter. □

131.5 Interpretation

131.5.1 What each measurement establishes

[Reserved: an explicit inference table — resistance run [374] plus persistent currents [1252] give dissipationless flow; flux expulsion [724] gives a thermodynamic phase, not a memory of history; the isotope effect [964, 969] gives a phonon-mediated interaction; gap spectroscopy [1112, 1180] gives a gapped excitation spectrum; flux quantization [1201] gives carriers of charge 2e; the Josephson effects [1246, 1266] give a single coherent phase across a macroscopic sample. Together these are the observational content of the BCS theory [1102].]

131.5.2 What the 1986 discovery did not settle

[Reserved: the Bednorz–Müller run in Ba–La–Cu–O, its resistive onset near 35 K and the cautious "possible" in its title [1768]; the confirmation by Meissner measurement, which is the step that turns a resistive onset into a superconductivity claim; the yttrium compound above the boiling point of nitrogen at 93 K [1802]; the honest statement that these measurements establish the phenomenon and its d-wave order parameter [1877, 1889] while leaving the mechanism open — see Chapter 133.]

Phenomenon 131.8 (Superconductivity above the boiling point of nitrogen)**.** A ceramic oxide of barium, lanthanum and copper — not a metal by any of the criteria that had guided the search — shows a resistive onset near 35 K, above any transition temperature then known [1768]; within months the related yttrium compound reached 93 K, above the boiling point of liquid nitrogen [1802]. The methodological point is contained in the cautious "possible" of the 1986 title: a resistive drop alone is not a superconductivity claim, and what converted the observation into one was the subsequent Meissner measurement, the flux expulsion of Phenomenon 131.1 being the property a percolating filament or a measurement artefact cannot imitate. Phase-sensitive experiments later established that the order parameter changes sign between crystallographic directions [1877, 1889], so the pairing is not of the

simple form that Equation (131.4) describes in mercury. What remains unexplained is the mechanism: no theory predicts the transition temperature of any of these compounds.

Derivation pending #270

What can and cannot be derived here: the demonstration that flux expulsion on field cooling requires a bulk, thermodynamic superconducting phase and excludes a percolating filamentary path; and the reading of the half-integer flux quantum trapped at a tricrystal ring as a sign change of the order parameter between directions. The mechanism itself is not derived anywhere in this treatise and is recorded as an open problem

131.6 Modern repetitions and precision

131.6.1 Metrology: the Josephson volt

[Reserved: the programmable Josephson voltage standard, in which thousands of junctions in series give a voltage fixed by a microwave frequency and the exact constant $K_J = 2e/h = 483\,597.8484\,\mathrm{GHz/V}$ [2333]; reproducibility at the part in 10^{10} level; consistency with the quantum Hall resistance of Chapter 128 through the CODATA adjustment [2392, 2431]; how this closed the loop from a 1911 curiosity to a base of the SI [2333].]

131.6.2 SQUID magnetometry and the London moment

[Reserved: modern SQUID sensitivity and its applications [2059]; the London moment — the magnetic moment of a rotating superconductor, proportional to the angular velocity — used as the gyroscope readout in Gravity Probe B, whose measurement of geodetic and frame-dragging precession is reported in Chapter 44 [2173]; a superconducting instrument as the enabling technology for a test of general relativity.]

131.6.3 High-pressure hydrides and the standard of proof

[Reserved: the diamond-anvil cell measurement of sulfur hydride with $T_c = 203\,\mathrm{K}$ near $155\,\mathrm{GPa}$, including its isotope shift on deuteration, which is the evidence that the pairing is phonon-mediated [2252]; lanthanum superhydride above $250\,\mathrm{K}$ [2340]; the methodological lesson — in a sample of picolitre volume a resistive drop alone is not sufficient, and the community now requires a magnetic-susceptibility signature and raw, unprocessed data, after several near-ambient claims failed replication and were retracted.]

131.6.4 Datasets

[Reserved: the digitized resistance-versus-temperature points of the 1911 mercury run [374], the isotope series [964, 969] and the critical-parameter compilation [1589] will be stored in the evidence library as comma-separated files keyed to their bibliography entries, so that every figure and fitted exponent in this chapter can be recomputed from the numbers as published.]

Chapter 132

Experiment: Neutron Scattering and the Structure of Solids

Chapter 125 builds the thermal physics of a crystal on a function it never measures. The heat capacity, the thermal expansion and the thermal conductivity all follow from the dispersion relation $\omega(\boldsymbol{k})$ of the lattice waves, and the classical route to that function — fitting a Debye temperature to a measured heat capacity — cannot recover it, because the heat capacity depends only on the density of modes $g(\omega)$ and many different dispersion relations share one $g(\omega)$. Chapter 129 is in the same position twice over. Néel's antiferromagnet is an arrangement of moments that cancel exactly, so that no magnetometer can see the order at all [781]; Bloch's spin waves are collective excitations whose dispersion fixes the exponent of the $T^{3/2}$ law but which no magnetization curve displays directly [643]. Three predictions, all of them about the interior of a solid at the scale of the interatomic spacing, and none of them within reach of any instrument before about 1945.

What closed all three was a particle found for quite other reasons [692] and made available in quantity, a decade later, by the fission reactors of Chapter 107. A neutron that has come to thermal equilibrium with a moderator at room temperature has, by an accident of its mass, a de Broglie wavelength of the order of an interatomic spacing *and* a kinetic energy of the order of a lattice vibration. No other probe has both at once. This chapter reports four experiments that exploited the coincidence: Shull and Smart's diffraction pattern of manganese oxide, which made a magnetic structure visible for the first time; Brockhouse and Stewart's scattering of neutrons by the phonons of an aluminium crystal, the measurement that turned lattice dynamics from a theory into a data set; Brockhouse's spin-wave scattering in magnetite, which did the same for magnons; and the branch-by-branch dispersion of germanium, taken with the triple-axis spectrometer in its mature form. It then shows that the same patterns deliver the Debye–Waller factor of Chapter 125 as a by-product, and closes with what the technique does and does not settle. The Nobel prize of 1994 went to Brockhouse [1897] and to Shull [1910] for the two halves of this work.

132.1 Why the neutron, and not the X-ray

The case for the neutron rests on four of its properties, and the decisive one is arithmetic.

A neutron of kinetic energy E has, non-relativistically, $p = \sqrt{2m_n E}$ and therefore

$$\lambda = \frac{h}{\sqrt{2m_n E}}\,. \tag{132.1}$$

Put $E = 25\,\mathrm{meV}$ into Equation (132.1), the mean kinetic energy scale of a gas of neutrons in equilibrium with a moderator near room temperature: the speed is about $2.2\,\mathrm{km/s}$ and the wavelength about $0.18\,\mathrm{nm}$. The wavelength resolves the interatomic spacing, so the

neutron diffracts from a crystal exactly as an X-ray does. The energy is of the same order as the quanta the crystal has to offer, which run from about 1 meV to about 100 meV: a phonon or a magnon can take a large fraction of the neutron's whole energy, and the change is measurable with an ordinary crystal analyser.

For any other probe of the same wavelength the second condition fails catastrophically. Table 132.1 makes the comparison. An X-ray photon of wavelength 0.18 nm carries about 6.9 keV, so a 25 meV energy transfer is a part in 3×10^5 of it — and it must be resolved twice over, once in the incident beam and once in the scattered one. That measurement is possible, but it needed a synchrotron and a generation of crystal-optics development [1985]; it was not available in 1955 and is still not the routine instrument.

Table 132.1: Three probes at the wavelength that resolves an interatomic spacing, $\lambda = 0.18$ nm. Only for the neutron is a typical excitation of the solid — from about 1 meV to about 100 meV — a sizeable fraction of the energy carried in. The last column takes 25 meV as the transfer to be measured. Energies follow from Equation (132.1) for the massive probes and from $E = hc/\lambda$ for the photon.

Probe	Energy at $\lambda = 0.18$ nm	25 meV as a fraction
Neutron	25 meV	1
Electron	46 eV	5×10^{-4}
X-ray photon	6.9 keV	4×10^{-6}

Three further properties matter, and each of them is used by one of the experiments below.

The neutron is uncharged. It feels no Coulomb barrier and no electron cloud, so it penetrates centimetres of most materials rather than micrometres: the sample can sit inside a cryostat, a furnace or a pressure cell, and what is measured is the bulk rather than the surface. The electron of Table 132.1 fails on this count as well as on the energy count.

The neutron scatters from nuclei, not from electron clouds. A nucleus is of order 10^{-15}–10^{-14} m across, four to five orders of magnitude smaller than the wavelength, so it is a point scatterer: the nuclear scattering length b is a single number, independent of the scattering angle. There is no form factor to fall away at large momentum transfer, which is what makes the high-order reflections strong enough for the Debye–Waller measurement of Section 132.6. And b varies erratically from element to element and even between isotopes of one element, instead of growing smoothly with atomic number as the X-ray amplitude does. Hydrogen is therefore visible next to uranium, and neighbouring elements in the periodic table can be told apart [1627].

The neutron carries a magnetic moment while carrying no charge. Its dipole couples to the moments of unpaired electrons, with an amplitude of the same order as the nuclear scattering length, so magnetic order produces diffraction peaks of ordinary strength [838]. This is the property that made Section 132.2 possible, and it is the one no other structural probe has.

132.2 Shull and Smart: antiferromagnetic order in MnO (1949)

Until 1949, antiferromagnetism was a hypothesis about something no instrument could detect. Néel's two-sublattice molecular field [781] predicted a state whose sublattice moments cancel, so that the susceptibility rises on cooling, peaks at the ordering temperature and then falls, and so that the substance is magnetically almost inert below it. Every one of

those signatures is indirect: they are consistent with the two-sublattice picture but do not exhibit it. The neutron exhibits it.

132.2.1 Apparatus

A beam from the graphite reactor at Oak Ridge, monochromated by Bragg reflection from a single crystal so that a narrow band of wavelengths near 0.1 nm enters the specimen, with Soller collimators and cadmium and boron shielding defining the beam. The specimen is a powder of manganese oxide in a thin-walled container mounted in a cryostat that can be held at liquid-nitrogen temperature or allowed to warm through the transition. The detector is a boron trifluoride proportional counter carried on an arm that rotates about the specimen axis, recording counts against scattering angle [923, 949, 991].

132.2.2 Procedure

The powder pattern is recorded twice, once with the specimen near 80 K and once at room temperature, and the two are subtracted. The logic of the comparison is what makes it decisive, and it rests on Section 132.6: the nuclear scattering is temperature-dependent only through the Debye–Waller factor, which *weakens* existing reflections and creates none. Any line that is present in the cold pattern and absent from the warm one therefore cannot be nuclear. The transition temperature is then located by following the intensity of the new lines as the specimen warms, and compared with the temperature of the susceptibility maximum measured on the same material by ordinary magnetometry.

132.2.3 Observations and data

At room temperature the pattern indexes completely on the face-centred-cubic rock-salt cell of manganese oxide, with the reflection conditions that cell requires. On cooling below the transition, additional reflections appear at scattering angles lying between the nuclear lines. They index on a cubic cell of twice the edge — equivalently, at half-integral indices of the chemical cell — and their intensities fall away with increasing scattering angle markedly faster than the nuclear lines do. They vanish on warming through a transition near 120 K, where the susceptibility has its maximum [949, 991].

Phenomenon 132.1 (Magnetic order makes Bragg peaks of its own)**.** A substance whose magnetic moments order with zero net moment scatters neutrons into reflections that are absent above the ordering temperature and absent from its X-ray pattern at every temperature. In manganese oxide these extra reflections index on a magnetic unit cell of twice the chemical edge, appear below about 120 K, and fall off with momentum transfer faster than the nuclear reflections [949, 991]. Their positions record the periodicity of the magnetic arrangement, their intensities its moment size, and their disappearance the ordering temperature — three quantities no magnetometer can supply, because the net moment they describe is zero.

Derivation. Let the magnetic ions occupy a Bravais lattice with primitive translations $\boldsymbol{a}_i$, and let the moment at the site $\boldsymbol{R}$ be $\sigma(\boldsymbol{R})\,\boldsymbol{m}$ with $\sigma = \pm 1$, reversing under each primitive translation. The neutron's dipole couples to the moment, so that site scatters with amplitude $p\,\sigma(\boldsymbol{R})$, where p carries the magnetic form factor and the geometrical factor but not the sign; the amplitude scattered by the crystal into momentum transfer $\boldsymbol{Q}$ is

$$F_m(\boldsymbol{Q}) = p \sum_{\boldsymbol{R}} \sigma(\boldsymbol{R})\, \mathrm{e}^{\mathrm{i}\boldsymbol{Q}\cdot\boldsymbol{R}} . \tag{132.2}$$

Any such sign function can be written $\sigma(\boldsymbol{R}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}_M\cdot\boldsymbol{R}}$ for a *propagation vector* $\boldsymbol{k}_M$ obeying $\boldsymbol{k}_M \cdot \boldsymbol{a}_i = \pi$ modulo 2π for every i; that is, $2\boldsymbol{k}_M$ is a reciprocal-lattice vector of the chemical

lattice while $\boldsymbol{k}_M$ itself is not. Substituting into Equation (132.2),

$$F_m(\boldsymbol{Q}) = p \sum_{\boldsymbol{R}} \mathrm{e}^{\mathrm{i}(\boldsymbol{Q}+\boldsymbol{k}_M)\cdot\boldsymbol{R}}, \tag{132.3}$$

which is the ordinary lattice sum and vanishes unless $\boldsymbol{Q}+\boldsymbol{k}_M = \boldsymbol{G}$ for some reciprocal-lattice vector $\boldsymbol{G}$ of the chemical lattice. Magnetic intensity therefore appears at $\boldsymbol{Q} = \boldsymbol{G} - \boldsymbol{k}_M$ and nowhere else: displaced from every nuclear reflection by $\boldsymbol{k}_M$, hence at half-integral indices, hence on the reciprocal lattice of a real-space cell of twice the edge. The two things the observation reports are one thing: the extra reflections at half-integral indices *are* the doubling of the magnetic period, and their absence from the nuclear pattern is guaranteed because $\boldsymbol{k}_M$ is not itself a reciprocal-lattice vector. In manganese oxide the magnetic ions form a face-centred cubic lattice and $\boldsymbol{k}_M = (\pi/a)(1,1,1)$, which reverses the sign across each successive (111) sheet: the moments lie ferromagnetically aligned within a sheet and antiparallel between neighbouring sheets, which is exactly Néel's two sublattices. The faster falloff of the magnetic intensities with angle is the form factor inside p, and is the second, independent signature that the reflections are magnetic: the nuclear amplitude has no form factor at all, as Section 132.1 explains, while the magnetic amplitude is the Fourier transform of an electron distribution of atomic size. □

132.2.4 Interpretation

The reflections are the two-sublattice structure itself, seen directly. A hypothesis proposed to account for a susceptibility curve [781] became a measured atomic arrangement, and the systematic determination of magnetic structures — moment directions, moment magnitudes, propagation vectors — began at once [991]. It should be said what the measurement does not fix. A powder pattern from a cubic crystal averages over the orientations of the magnetic domains, and the moment *direction* within the ordered sheets is not determined uniquely by the intensities alone; the propagation vector and the moment magnitude are. The technique also says nothing about *why* the manganese ions couple antiparallel through the intervening oxygen, which is the separate question of superexchange treated in Chapter 129.

132.2.5 Primary references

[949], with the systematic follow-up [991]. The prediction under test is [781].

132.3 Brockhouse and Stewart: phonons in aluminium (1955)

Diffraction alone locates atoms; it says nothing about how they move. The step from the elastic experiment above to a measurement of $\omega(\boldsymbol{k})$ requires the energy of the scattered neutron to be measured as well as its direction, and it was first taken in a single crystal of aluminium.

132.3.1 Apparatus

A beam from the Chalk River reactor, monochromated by Bragg reflection from a single crystal so that the neutrons entering the specimen have a known energy E_i and wavevector $\boldsymbol{k}_i$. The specimen is a single crystal of aluminium — face-centred cubic, one atom in the primitive cell — mounted on a goniometer that sets its orientation relative to the incident beam. The scattered beam is analysed in energy by a second Bragg reflection from a crystal before reaching the counter, so that the instrument records the number of neutrons arriving with a chosen final energy E_f in a chosen direction. This arrangement, monochromator–specimen–analyser, is the triple-axis spectrometer in its first form; Section 132.5 uses it in its mature one.

132.3.2 Procedure

Fix the incident energy and the scattering angle and record the distribution of final energies. Each pair of readings converts, by the kinematics of Equation (132.4) below, into a momentum transfer $\boldsymbol{Q} = \boldsymbol{k}_i - \boldsymbol{k}_f$ and an energy transfer $\hbar\omega = E_i - E_f$. Repeating over a range of angles and crystal orientations samples a region of the $(\boldsymbol{Q}, \omega)$ plane. The single crystal is essential: a powder averages over orientations and destroys the direction of $\boldsymbol{Q}$, which is exactly the information a dispersion relation is about.

132.3.3 Observations and data

The distribution of scattered energies at a fixed geometry is not smooth. It carries resolved peaks standing above a background, and the $(\boldsymbol{q}, \omega)$ pairs deduced from the peak positions do not fill the plane: they fall on curves. In aluminium the curves are three in number, they rise linearly out of the origin with slopes that agree with the sound velocities computed from the elastic constants of the same metal, and they flatten towards the zone boundary, so the group velocity $\mathrm{d}\omega/\mathrm{d}q$ falls to zero there [1056]. The measured frequencies of the primary paper are not reproduced here; what is quoted is the shape of the result, which is what the theory predicts.

Phenomenon 132.2 (At a fixed momentum transfer the neutron loses only certain energies)**.** Neutrons scattered by a single crystal into a fixed momentum transfer $\boldsymbol{Q}$ do not emerge with a continuum of energies. The spectrum carries sharp peaks at a small number of energy transfers, and the pairs $(\boldsymbol{q}, \omega)$ they define lie on a small number of continuous branches that are periodic in the reciprocal lattice [1056, 1137]. The excitations of a crystal are therefore quanta with a definite energy for each wavevector, and the dispersion relation is read off the instrument rather than postulated.

Derivation. A crystal is invariant under the discrete translations of its lattice and not under continuous ones, so what the collision conserves is the neutron momentum modulo $\hbar\boldsymbol{G}$, with $\boldsymbol{G}$ any reciprocal-lattice vector. Writing $\boldsymbol{Q} = \boldsymbol{k}_i - \boldsymbol{k}_f$ and $\hbar\omega = E_i - E_f$, a process that creates one quantum of the branch s at wavevector $\boldsymbol{q}$ obeys

$$\boldsymbol{Q} = \boldsymbol{q} + \boldsymbol{G}\,, \qquad \hbar\omega = \hbar\omega_s(\boldsymbol{q})\,. \tag{132.4}$$

The first of these fixes $\boldsymbol{q}$ once $\boldsymbol{Q}$ is chosen, since $\boldsymbol{q}$ is defined only modulo $\boldsymbol{G}$; the second then leaves for the energy transfer only the $3p$ values $\hbar\omega_s(\boldsymbol{q})$, one for each branch. A discrete set of allowed energy transfers at fixed $\boldsymbol{Q}$ is a set of peaks in the spectrum of scattered neutrons, which is the observation. Conversely one peak fixes one pair $(\boldsymbol{q}, \omega)$, so a scan is a set of points on $\omega_s(\boldsymbol{q})$. Two features of the data follow without further work. The branches are periodic in the reciprocal lattice because $\boldsymbol{q}$ enters Equation (132.4) only modulo $\boldsymbol{G}$, so nothing outside the first Brillouin zone is new; and the smooth background beneath the peaks is not instrumental error but multiphonon scattering, in which two or more quanta share the transfer, so that Equation (132.4) constrains only their sum and the allowed ω fill an interval. □

Derivation pending #271

The one-phonon scattering cross section itself: the factor that weights each peak by the square of the projection of the polarization vector on the momentum transfer, by the Bose population of the mode, and by the Debye–Waller factor, so that the measured intensities constrain the eigenvectors of the dynamical matrix and not only its eigenvalues

132.3.4 Interpretation

Before this measurement the vibrational spectrum of a solid was known only through quantities that integrate over it. A heat capacity determines the density of modes $g(\omega)$, and infinitely many dispersion relations share one $g(\omega)$; a Debye temperature is a one-parameter caricature of the whole function. Diffuse X-ray scattering carries the information in principle but requires an assumed force model to extract it. The neutron measurement is direct in the strict sense: one peak, one point of $\omega(\boldsymbol{q})$, no model. That is what entitles Chapter 125 to speak of the phonon as a particle — the sharpness of the peak is the statement that a wavevector has a definite energy — and it is what makes the Born–von Kármán force constants a fitted, falsifiable object rather than a parametrization.

132.3.5 Primary references

[1056]. The theoretical frame is the lattice dynamics of [381, 1030].

132.4 Brockhouse: spin waves in magnetite (1957)

The same instrument, pointed at a magnetically ordered crystal, does for Chapter 129 what it had just done for Chapter 125. Bloch's account of the fall of the magnetization with temperature rests on the claim that the cheap excitations of an ordered magnet are long-wavelength precessional waves rather than individually reversed spins [643]. The two pictures differ qualitatively — a reversed spin costs a finite energy and would give an exponential law, a spin wave costs arbitrarily little and gives a power law — so the claim is testable by the magnetization curve alone. Seeing the excitation itself is another matter.

132.4.1 Apparatus

The triple-axis arrangement of Section 132.3, with a single crystal of magnetite, Fe_3O_4, as the specimen, at room temperature — far below its ordering temperature, so the ordered state is well developed and the magnon population is small enough for the excitations to be sharp.

132.4.2 Procedure

Scans are taken at small wavevector transfer relative to a reciprocal-lattice point, where a magnon branch that starts from zero is well separated from the phonon branches. Three independent signatures distinguish a magnetic peak from a lattice one, and the identification uses all three. The intensity of magnetic scattering falls with $|\boldsymbol{Q}|$ through the magnetic form factor, whereas one-phonon intensity *grows* as $|\boldsymbol{Q}|^2$, so repeating a scan near a high-order reciprocal-lattice point separates them. The dispersion of a ferromagnetic magnon is quadratic at long wavelength while an acoustic phonon's is linear. And the magnon intensity and dispersion respond to temperature and to an applied field as the lattice modes do not.

132.4.3 Observations and data

A branch is resolved whose energy rises from zero quadratically with wavevector at long wavelength, distinct from the phonon branches of the same crystal, and whose intensity falls with momentum transfer in the manner of a magnetic form factor [1104]. The fitted coefficient of q^2 is the spin-wave stiffness D, the quantity that fixes the coefficient of the $T^{3/2}$ law in Chapter 129 and, through the derivation below, the exchange constant. The value obtained is not quoted here.

Magnetite is a ferrimagnet rather than a simple ferromagnet: it has two inequivalent magnetic sublattices, so its spin-wave spectrum has an acoustic branch, which starts at zero and is quadratic at long wavelength like a ferromagnet's, together with optical branches at higher energy. The branch measured at small wavevector is the acoustic one, and it is the acoustic branch alone that the $T^{3/2}$ law needs.

Phenomenon 132.3 (Spin waves are quanta with a quadratic dispersion)**.** An ordered magnet carries collective excitations that inelastic neutron scattering resolves as sharp peaks, on a branch whose energy vanishes as the wavevector goes to zero and rises as $\hbar\omega = Dq^2$ at long wavelength [1104]. The excitation spectrum of a ferromagnet therefore has no gap, which is what supplies the arbitrarily cheap excitations that the $T^{3/2}$ law of [643] requires, and the stiffness D measured here must equal the one fitted to the magnetization curve of the same material.

Derivation. Take the Heisenberg model of Chapter 129 on a simple cubic lattice of spacing a, written with each bond counted once,

$$H = -J \sum_{\text{bonds}} \boldsymbol{S}_n \cdot \boldsymbol{S}_m \,, \qquad J > 0 \,, \tag{132.5}$$

and treat the spins as classical vectors of length S — legitimate for the long-wavelength modes, in which neighbouring spins are almost parallel and the quantum corrections are of relative order $1/S$. Each spin precesses in the field of its neighbours,

$$\hbar \frac{\mathrm{d}\boldsymbol{S}_n}{\mathrm{d}t} = J\, \boldsymbol{S}_n \times \sum_{\boldsymbol{\delta}} \boldsymbol{S}_{n+\boldsymbol{\delta}} \,, \tag{132.6}$$

the sum running over the $z = 6$ nearest neighbours. Linearize about the ordered state by writing $\boldsymbol{S}_n = (S_n^x, S_n^y, S)$ with the transverse components small, and keep terms of first order in them:

$$\hbar \frac{\mathrm{d}S_n^x}{\mathrm{d}t} = JS \left(z S_n^y - \sum_{\boldsymbol{\delta}} S_{n+\boldsymbol{\delta}}^y \right) , \tag{132.7}$$

$$\hbar \frac{\mathrm{d}S_n^y}{\mathrm{d}t} = -JS \left(z S_n^x - \sum_{\boldsymbol{\delta}} S_{n+\boldsymbol{\delta}}^x \right) . \tag{132.8}$$

The pair decouples in the combination $S^+ = S^x + \mathrm{i}S^y$, giving $\hbar\,\mathrm{d}S_n^+/\mathrm{d}t = -\mathrm{i}JS\left(zS_n^+ - \sum_{\boldsymbol{\delta}} S_{n+\boldsymbol{\delta}}^+\right)$, and the lattice translation invariance of that equation makes plane waves $S_n^+ \propto \mathrm{e}^{\mathrm{i}(\boldsymbol{q}\cdot\boldsymbol{R}_n - \omega t)}$ its normal modes. Substituting them,

$$\hbar\omega(\boldsymbol{q}) = JS \sum_{\boldsymbol{\delta}} \left(1 - \cos \boldsymbol{q}\cdot\boldsymbol{\delta}\right) . \tag{132.9}$$

Two consequences are the content of the experiment. First, $\omega \to 0$ as $\boldsymbol{q} \to 0$: the uniform mode is a rigid rotation of every spin together, which costs nothing because Equation (132.5) depends only on relative orientations. The gaplessness is thus a consequence of the rotational symmetry that the ordered state breaks, not of any detail of J, and it is what makes the low-temperature magnetization deficit a power of T rather than an exponential. Second, expanding Equation (132.9) for $|\boldsymbol{q}|\,a \ll 1$ on the cubic lattice, where the six neighbours contribute in pairs along the three axes,

$$\hbar\omega \simeq JSa^2q^2 \,, \qquad D = JSa^2 \,, \tag{132.10}$$

which is the measured shape and which converts the measured stiffness into the exchange constant. The numerical factor in D depends on the convention for Equation (132.5): writing the Hamiltonian as an unrestricted double sum over pairs, as Chapter 129 does, counts each bond twice and doubles D at fixed J. What is convention-independent is the shape $\hbar\omega \propto q^2$ and the vanishing at $\boldsymbol{q} = 0$, and those are what the scan measures. □

132.4.4 Interpretation

Two independent measurements on the same material must now agree: the stiffness D read off the neutron dispersion and the stiffness implied by the coefficient of $T^{3/2}$ in the magnetization curve. That is a closed loop of the kind an evidence-based treatise should insist on, because either measurement alone can be accommodated by adjusting a parameter and the pair cannot. The experiment also settles the question Bloch's calculation raised and could not answer from thermodynamics alone: the excitations are collective waves with a gapless spectrum, and not reversed individual spins, which would have shown a finite minimum energy in exactly the scan performed here.

132.4.5 Primary references

[1104]. The prediction under test is [643], with the operator formulation in [852].

132.5 Brockhouse and Iyengar: the normal modes of germanium (1958)

The aluminium measurement showed that a point of $\omega(\boldsymbol{q})$ can be taken. Germanium showed that a whole dispersion relation can be taken — every branch, along the symmetry directions, in a crystal with more than one atom per cell — and it is the measurement against which lattice-dynamical models have been tested ever since.

132.5.1 Apparatus

The triple-axis crystal spectrometer at Chalk River, in its mature form: three axes of rotation, one at the monochromating crystal, which selects $\boldsymbol{k}_i$ out of the reactor's broad spectrum; one at the specimen, which sets its orientation and hence the direction of $\boldsymbol{Q}$ in the crystal; and one at the analysing crystal, which selects $\boldsymbol{k}_f$ before the counter. The specimen is a single crystal of germanium, whose diamond structure has $p = 2$ atoms in the primitive cell.

132.5.2 Procedure

The instrument's three angles and two crystal settings give more freedom than the measurement needs, and Brockhouse's use of that freedom is the methodological content of the experiment. In the *constant-$\boldsymbol{Q}$* method the incident and final energies are stepped together, with the arm angles moved to compensate, so that the momentum transfer stays fixed while the energy transfer is scanned. Each scan is then a cut through the $(\boldsymbol{Q}, \omega)$ plane along one axis, and a branch appears as a peak at a definite energy rather than as something to be reconstructed from a locus of intersections. Scans are repeated along the [100], [110] and [111] directions, and the peak positions are collected into $\omega_s(\boldsymbol{q})$ branch by branch.

132.5.3 Observations and data

Germanium shows six branches, as $3p = 6$ requires. Three are acoustic: they vanish at the zone centre, rise linearly there with slopes that agree with the sound velocities computed from the elastic constants of the same crystal, and flatten towards the zone boundary. Three are optical, with a finite frequency at $\boldsymbol{q} = 0$; branches degenerate along a symmetry direction separate away from it, as the symmetry of the direction requires. The complete set was compared with Born–von Kármán force-constant fits, and the fits require interactions extending beyond nearest neighbours [1137].

Phenomenon 132.4 (A crystal with p atoms per primitive cell has exactly $3p$ branches, three of them acoustic)**.** The measured dispersion relation of a crystal consists of $3p$ branches $\omega_s(\boldsymbol{q})$, where p is the number of atoms in the primitive cell; three of them vanish as $\boldsymbol{q} \to 0$ and the remaining $3p-3$ do not. Germanium, with $p=2$, shows six branches, three acoustic and three optical [1137]; aluminium, with $p=1$, shows three and they are all acoustic [1056].

Derivation. Let $u_\alpha(\boldsymbol{R}\kappa)$ be the Cartesian displacement of the atom κ of mass M_κ in the cell at $\boldsymbol{R}$. In the harmonic approximation the equations of motion are linear with coefficients depending on $\boldsymbol{R}-\boldsymbol{R}'$ alone, so they are diagonalized by the Bloch form

$$u_\alpha(\boldsymbol{R}\kappa) = M_\kappa^{-1/2}\, e_\alpha(\kappa)\, \mathrm{e}^{\mathrm{i}(\boldsymbol{q}\cdot\boldsymbol{R}-\omega t)}\,, \tag{132.11}$$

and reduce, for each $\boldsymbol{q}$ separately, to the eigenvalue problem

$$\omega^2 e_\alpha(\kappa) = \sum_{\beta\kappa'} D_{\alpha\beta}(\kappa\kappa';\boldsymbol{q})\, e_\beta(\kappa') \tag{132.12}$$

for the dynamical matrix D, which is Hermitian because the force constants are symmetric, and which is $3p\times 3p$ because the indices run over three Cartesian directions and p atoms. A Hermitian $3p\times 3p$ matrix has $3p$ real eigenvalues, so there are $3p$ branches; and D is built from $\mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{R}}$ with $\boldsymbol{R}$ on the lattice, so the branches are periodic in the reciprocal lattice.

Three of them vanish at $\boldsymbol{q}=0$. A rigid translation moves every atom of every cell by the same vector $\boldsymbol{c}$; it leaves every interatomic distance unchanged, so it costs no potential energy. It is of the form Equation (132.11) with $\boldsymbol{q}=0$ and $e_\alpha(\kappa)=\sqrt{M_\kappa}\,c_\alpha$, hence an eigenvector of $D(\boldsymbol{0})$ with eigenvalue zero; there are three independent choices of $\boldsymbol{c}$, so three branches with $\omega\to 0$ as $\boldsymbol{q}\to 0$. The remaining $3p-3$ modes move atoms within a cell against one another and have no reason to be free, so they start at a finite frequency. Nothing in the argument uses the values of the force constants: the count is a consequence of translational invariance and of the size of the cell, which is why it is a test of the framework rather than a fit to the data. Germanium must show six branches with three acoustic, and does. □

132.5.4 Interpretation

Two things are established at once, and they are of different kinds. The first is structural and survives any change of model: the number of branches and the three zero modes at the zone centre, which is the content of Phenomenon 132.4. The second is quantitative and model-dependent: the shapes of the branches determine the force constants, and in germanium they cannot be fitted with nearest-neighbour forces alone. That is a genuine result about the bonding of a covalent semiconductor, obtained from a mechanical measurement, and it is the sort of thing that only a full dispersion relation can deliver — a density of states, and therefore a heat capacity, is insensitive to it. With this experiment the function on which Chapter 125 rests stopped being an assumption.

132.5.5 Primary references

[1137], following [1056]; the model it is fitted against is that of [381, 1030].

132.6 The Debye–Waller factor, measured on the same patterns

The diffraction patterns of Section 132.2 carry, without any extra apparatus, the measurement that Chapter 125 needs to close its account of thermal motion. It is set down here because the procedure of that experiment leans on it, and because the neutron measures it more

cleanly than the X-ray does: the nuclear scattering length has no form factor, so the whole falloff of intensity with momentum transfer at fixed temperature is thermal, with nothing to disentangle.

Phenomenon 132.5 (Thermal motion weakens the Bragg peaks without broadening them, and does not stop at absolute zero)**.** The intensity of a Bragg reflection falls as the crystal is heated, by a factor that grows with the momentum transfer; the reflection stays at the same angle and keeps its width. The intensity lost from the peaks reappears as a diffuse background rather than disappearing. And the attenuation does not extrapolate to unity at zero temperature: a harmonic crystal at absolute zero still attenuates its own reflections [402, 506]. The same factor governs the recoilless fraction in nuclear resonance absorption [1154].

Derivation. Let the nucleus n sit at $\boldsymbol{R}_n + \boldsymbol{u}_n$, with $\boldsymbol{R}_n$ its equilibrium position and $\boldsymbol{u}_n$ its thermal displacement, and let b_n be its scattering length. The Bragg peaks come from the part of the scattering that is coherent with the average lattice, that is from the thermal average of the amplitude:

$$\langle F(\boldsymbol{Q}) \rangle = \sum_n b_n \, \mathrm{e}^{\mathrm{i}\boldsymbol{Q}\cdot\boldsymbol{R}_n} \left\langle \mathrm{e}^{\mathrm{i}\boldsymbol{Q}\cdot\boldsymbol{u}_n} \right\rangle . \tag{132.13}$$

In a harmonic crystal $\boldsymbol{u}_n$ is a linear combination of the normal-mode coordinates, each of which is Gaussian in thermal equilibrium, so $x = \boldsymbol{Q}\cdot\boldsymbol{u}_n$ is a Gaussian variable of zero mean. For such a variable $\langle \mathrm{e}^{\mathrm{i}x} \rangle = \mathrm{e}^{-\langle x^2 \rangle/2}$, whence

$$\langle F(\boldsymbol{Q}) \rangle = \mathrm{e}^{-W} \sum_n b_n \, \mathrm{e}^{\mathrm{i}\boldsymbol{Q}\cdot\boldsymbol{R}_n} , \qquad 2W = \left\langle (\boldsymbol{Q}\cdot\boldsymbol{u})^2 \right\rangle , \tag{132.14}$$

and the measured intensity, being the squared modulus of the amplitude, carries the factor e^{-2W}.

Each of the three observed features is now a line of the calculation. The factor e^{-W} multiplies the lattice sum in Equation (132.14) without altering it, so the peak positions and widths are untouched: the *average* lattice is still perfectly periodic, and it is the average lattice that the Bragg peaks see. Nothing is destroyed, because the total scattering from a set of nuclei is fixed by their scattering lengths and cannot depend on temperature; what leaves the peaks reappears in the one- and multi-phonon scattering that fills the space between them, which is the same background beneath the peaks of Phenomenon 132.2. And $\langle u^2 \rangle$ does not vanish at $T = 0$: each normal mode retains its zero-point amplitude, so $W > 0$ at absolute zero. That last term is Waller's correction [506] to Debye's classical calculation [402], and it is a direct consequence of the quantization that Chapter 125 imposes on the modes. At high temperature, by contrast, equipartition makes $\langle u^2 \rangle$ proportional to T, so $\ln I$ is linear in temperature with a slope proportional to $|\boldsymbol{Q}|^2$ — the form in which the factor is fitted, and from which the mean square displacement of an atom in a crystal is extracted. □

Derivation pending #272

The mode sum for the mean square displacement, giving $2W$ in terms of the measured dispersion relation, its evaluation in the Debye model with the zero-point term separated from the thermal one, and the resulting temperature dependence of the attenuation at fixed momentum transfer

Table 132.2: The four measurements of this chapter, with the prediction each tests and the chapter that makes it. In every case the quantity measured is one that no other instrument of the period could reach — which is the argument for the technique, and the reason the two theoretical chapters had to be written ahead of their evidence.

Experiment	Tests	Result
Shull–Smart 1949	Chapter 129	extra Bragg reflections below 120 K on a cell of twice the chemical edge: Néel's two sublattices seen directly
Brockhouse–Stewart 1955	Chapter 125	sharp energy transfers at fixed momentum transfer; three branches in aluminium, linear at long wavelength and flat at the zone boundary
Brockhouse 1957	Chapter 129	a gapless branch with $\hbar\omega = Dq^2$ in magnetite: magnons resolved, and D measured independently of the magnetization curve
Brockhouse–Iyengar 1958	Chapter 125	all six branches of germanium along three symmetry directions; force constants beyond nearest neighbours required

132.7 What the neutron settled

Table 132.2 closes four gaps at once, and the 1994 Nobel prize recognized the two halves of the technique: the diffraction method that shows where the atoms and the moments are, and the spectrometry that shows how they move. It is worth being explicit about what the closure does and does not amount to.

What is measured directly is a scattering law [1052], and its interpretation as a dispersion relation uses the one-phonon — or one-magnon — approximation, whose corrections are the multiphonon background of Phenomenon 132.2. Where the excitations are not sharp, the peaks broaden and the notion of a dispersion curve itself weakens; the measurement then reports a lifetime as well as an energy, which is more information rather than less, but it is no longer the clean reading of a function that Chapter 125 assumes. The neutron is also a flux-limited probe: sources are few, samples must be large by the standards of modern materials work, and weakly scattering or strongly absorbing elements are difficult. Inelastic X-ray scattering at synchrotron sources now reaches phonons in specimens far too small for a neutron beam [1985], at the price of the extreme energy resolution that Table 132.1 shows it needs.

Neither reservation touches the conclusions of this chapter, because all four are qualitative statements that a quantitative correction cannot reverse: the magnetic cell of manganese oxide is twice the chemical one; the energy transfer at fixed momentum transfer is discrete; germanium has six branches and three of them are acoustic; the magnon spectrum has no gap. Each is what its theory chapter predicted, and each was, before 1949, a claim about something nobody had seen.

132.8 Hodgkin, Hauptman and Karle, and Klug: the phase problem (1934–1969)

One step has been taken for granted in every diffraction measurement of this chapter, and it is not innocent. A detector counts quanta, and its reading at a reciprocal-lattice point is proportional to $|F(\boldsymbol{G})|^2$, the squared modulus of the scattered amplitude. A modulus is half of a complex number: the phase $\arg F(\boldsymbol{G})$ leaves no mark on any counter, and the Fourier synthesis that turns a diffraction pattern into a picture of the scatterers needs both halves. Bragg's law [399], in the form of Equation (125.1), fixes where the peaks are and therefore the lattice; what sits inside the cell is carried by the intensities, and the intensities do not contain enough to recover it. This is the *phase problem*, named as an open question in Section 125.2.3 and taken up here.

It did not obstruct the four measurements above, and the reason is worth stating, because it is the exception and not the rule: each of them had few enough unknowns to be *fitted* rather than inverted in the first place. The chemical structure of manganese oxide was already known from X-rays when Section 132.2 began, so the unknowns there were a propagation vector, a moment magnitude and a direction — a handful of numbers, which a comparison of computed with observed intensities settles without any Fourier synthesis at all. The inelastic measurements use peak positions and barely use intensities. A molecule of a hundred atoms in the cell offers three hundred coordinates, and there fitting has nothing to start from: a model must be guessed before it can be compared, and nobody can guess three hundred numbers.

Three Nobel prizes in Chemistry mark the three ways out. Dorothy Crowfoot Hodgkin's, in 1964, was for carrying X-ray analysis to molecules whose constitution chemistry had been unable to settle, using a heavy atom that the molecule contains, or can be made to contain, as the source of the phases [1311]. Herbert Hauptman and Jerome Karle's, in 1985, was for showing that the phases are not in fact independent of the amplitudes: a scattering density is non-negative and is made of atoms, and those two facts alone constrain the phases enough to determine them [1775, 1777]. Aaron Klug's, in 1982, was for taking the phases from a different instrument altogether — an electron microscope forms an image, and an image has its phases in it [1722].

132.8.1 Apparatus

For the X-ray work: a sealed tube with a copper or molybdenum target, whose characteristic $K\alpha$ line supplies a wavelength near 0.1 nm, filtered or monochromated by a crystal; a single crystal of the substance, a few tenths of a millimetre across, on a goniometer head that can be set to any orientation; and a recorder — photographic film in a Weissenberg or precession camera in the early work, a counter diffractometer later — that measures the integrated intensity of each reflection. A protein crystal is largely solvent and must stay wet, so it is sealed in a capillary with its mother liquor. Where heavy-atom phasing is used the apparatus includes the chemistry: crystals soaked in solutions of mercury, platinum or uranium salts, which have to bind at a few definite sites without disturbing the lattice.

The second instrument is arithmetic, and here it is apparatus in the strict sense. A three-dimensional Fourier synthesis over several thousand reflections is not a calculation anyone finishes by hand in a working lifetime, and the vitamin B_{12} analysis was among the first structural problems carried out on the electronic computers then becoming available. What can be solved is set by what can be computed, which is why the reach of the subject tracks the history of computing so closely.

For the electron work: a transmission electron microscope with magnetic lenses, at an accelerating voltage of order 100 kV; a specimen thin enough to be transparent to the beam, supported on a film and imaged either negatively stained or unstained; a tilting stage; a

photographic plate, and a densitometer to turn the developed plate into numbers a computer can transform.

132.8.2 Procedure

The X-ray procedure has three variants and one goal: obtain a first estimate of the phases good enough that a Fourier synthesis computed with the *measured* moduli and the *estimated* phases shows recognizable atoms. Once it does, the atoms found give better phases, a better map follows, and the cycle converges — the moduli being right throughout, the phases are the only thing being improved.

The first variant computes the synthesis whose coefficients are the measured intensities themselves, all phases set to zero, and reads the position of a heavy atom out of the resulting vector map. Phases computed from that one atom are then good enough to start the cycle, because a heavy atom contributes a large and known term to every structure factor. The molecule may supply the heavy atom itself — the cobalt at the centre of vitamin B_{12} — or one may be introduced, by crystallizing a salt of the substance with a heavy cation or by attaching a heavy group chemically.

The second variant is for proteins, where no atom of the molecule dominates. Crystals are soaked with a heavy-metal salt so as to add a few atoms per molecule without changing the cell, the intensities of native and derivative are measured separately, and their differences locate the added atoms; the derivative then plays the part of the known term. In the general case two independent derivatives are needed, for the reason given in the interpretation below.

The third variant uses no heavy atom and no chemistry. The measured moduli are scaled and normalized to what point atoms at rest would give; the strongest few hundred are selected; an origin is fixed by assigning arbitrary phases to a small permitted set of reflections; and phases are propagated through the relations of Equation (132.20) below, many trial sets being carried in parallel and ranked by internal consistency. The map is computed from the surviving set and inspected for a chemically sensible arrangement of atoms.

The electron procedure records micrographs of the specimen at a series of orientations — or exploits the specimen's own symmetry, which presents several orientations at once — digitises each plate, transforms it, places the result as a central plane of the three-dimensional transform, interpolates onto a grid and transforms back. The dose that would give one good image destroys the specimen, so each micrograph is recorded with far too few electrons and many images of identical particles are averaged.

132.8.3 Observations and data

The synthesis computed with intensities as coefficients and every phase set to zero is not noise. It is a map with a peak at the end of every interatomic vector of the structure, weighted by the product of the two scattering powers, and with a large peak at the origin [747]. That map is centrosymmetric whether the crystal is or not, and the measured intensities themselves obey $I(\boldsymbol{G}) = I(-\boldsymbol{G})$ to the precision of the measurement — with one reproducible exception, in crystals containing an element whose absorption edge lies near the wavelength in use, where the two members of the pair differ by a small but definite amount [977].

Penicillin was determined while its constitution was still disputed, and the analysis showed the strained four-membered β-lactam ring fused to a five-membered ring, which the chemistry had been unable to establish [926]. Vitamin B_{12}, a molecule with of order a hundred atoms heavier than hydrogen and no known constitution at all, was determined from the crystals with the cobalt atom at its centre as the heavy atom [1087]; insulin, a protein, followed by isomorphous replacement and a far larger computation [1387].

In structures already known, the phase sums built from triples of strong reflections are found not to be spread uniformly round the circle: they cluster about zero, and the clustering tightens as the product of the three normalized amplitudes grows. For a centrosymmetric structure, where a phase can only be 0 or π, the same statement reads that the product of the three signs is $+1$ more often than not, the fraction rising towards unity with that same product [962, 1060]. Structures with of order a hundred independent non-hydrogen atoms are now solved from the measured intensities alone, with no chemical input beyond the composition [1775, 1777].

An electron micrograph behaves differently from a diffraction pattern in exactly the way that matters. Its two-dimensional Fourier transform, computed from the digitised plate, reproduces the three-dimensional transform of the specimen on the plane through the origin perpendicular to the beam — with phases. A set of micrographs at different orientations therefore fills the transform, and back transforming yields a density in three dimensions. This was carried out for the tail of bacteriophage T4, whose helical symmetry presents every azimuth in a single image [1369], and set out in general form immediately afterwards [1418].

Phenomenon 132.6 (A diffraction measurement determines the autocorrelation of the scatterers, not the scatterers)**.** The Fourier synthesis whose coefficients are the measured intensities, with all phases set to zero, is the autocorrelation of the scattering density: a map carrying a peak at the end of every interatomic vector [747]. It is centrosymmetric even when the structure is not, and correspondingly the intensities obey Friedel's law $I(\boldsymbol{G}) = I(-\boldsymbol{G})$ wherever the scattering is non-absorbing, so that no diffraction pattern can distinguish a structure from its mirror image; near an absorption edge the law fails measurably, and the failure fixes the absolute configuration [977]. What the experiment delivers is thus exact and is not the structure.

Derivation. Write the scattering density of one cell as $\rho(\boldsymbol{r})$ — the electron density for X-rays, the nuclear scattering-length density for the neutrons of Section 132.1 — and keep the sign convention of Equation (132.13),

$$F(\boldsymbol{G}) = \int_V \rho(\boldsymbol{r})\, \mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{r}}\, \mathrm{d}^3 r\,, \qquad \rho(\boldsymbol{r}) = \frac{1}{V}\sum_{\boldsymbol{G}} F(\boldsymbol{G})\, \mathrm{e}^{-\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{r}}\,. \tag{132.15}$$

The second of these is the synthesis that would produce the structure, and it needs F and not $|F|$. Form instead the synthesis whose coefficients are the numbers a counter delivers,

$$P(\boldsymbol{u}) = \frac{1}{V}\sum_{\boldsymbol{G}} |F(\boldsymbol{G})|^2\, \mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{u}}\,. \tag{132.16}$$

Substituting Equation (132.15) into the autocorrelation $\int_V \rho(\boldsymbol{r})\rho(\boldsymbol{r}+\boldsymbol{u})\, \mathrm{d}^3 r$ gives a double sum over $\boldsymbol{G}$ and $\boldsymbol{G}'$ whose $\boldsymbol{r}$ integral is $V\delta_{\boldsymbol{G}',-\boldsymbol{G}}$; the surviving terms carry $F(\boldsymbol{G})F(-\boldsymbol{G}) = |F(\boldsymbol{G})|^2$, since ρ is real, and what is left is exactly Equation (132.16). The intensities therefore determine the autocorrelation completely and without approximation.

Three properties of that object are the three parts of the observation. Relabelling $\boldsymbol{G} \to -\boldsymbol{G}$ in Equation (132.16) leaves $|F|^2$ alone and flips the exponent, so $P(-\boldsymbol{u}) = P(\boldsymbol{u})$: the map is centrosymmetric whatever ρ may be. For a structure of well-separated atoms of strengths Z_j at positions $\boldsymbol{r}_j$ the autocorrelation is $\sum_{j,k} Z_j Z_k\, \Gamma(\boldsymbol{u} - (\boldsymbol{r}_k - \boldsymbol{r}_j))$ with Γ the autocorrelation of one atom, that is a peak of weight $Z_j Z_k$ at the end of each interatomic vector, with the N diagonal terms piled at the origin. And ρ real gives $F(-\boldsymbol{G}) = F(\boldsymbol{G})^*$, hence Friedel's law; a structure and its inversion $\rho(-\boldsymbol{r})$ have structure factors that are complex conjugates of one another and so identical intensities, which is the statement that handedness is invisible. Absorption is what breaks it: near an edge the effective scattering density acquires an imaginary part, ρ is no longer real, and the Friedel pair separates.

The same lines say why the vector map is not a solution. It carries $N(N-1)$ off-origin peaks in a cell that holds N atoms, so the peak density grows as N^2 while the space available does not: at $N \sim 10$ the map can be read, at $N \sim 100$ it is a continuum. The loss is not only of resolution — distinct structures can share an autocorrelation exactly — but overcrowding is what defeats the method in practice. One case escapes, and it is the case Hodgkin exploited: if one atom is much heavier than the rest, the peaks involving it carry weight Z_{heavy}^2 or $Z_{\text{heavy}} Z_j$ and stand above the rest, so the heavy atom alone can be located in a map that shows nothing else. □

Phenomenon 132.7 (Non-negativity and atomicity tie the phases to the amplitudes)**.** For a structure of resolved atoms the phases are not free given the moduli. The sum $\varphi(\boldsymbol{H}) + \varphi(\boldsymbol{K}) - \varphi(\boldsymbol{H}+\boldsymbol{K})$ over a triple of strong reflections is observed to cluster about zero rather than to be distributed uniformly, and the clustering tightens as the product of the three normalized amplitudes grows; in a centrosymmetric structure the corresponding product of signs is $+1$ with a probability rising towards unity with the same product [962, 1011, 1060]. Structures of the size of a small organic molecule are therefore solved from the intensities alone, with no heavy atom and no chemical assumption beyond the composition [1020, 1089, 1775, 1777].

Derivation. Take N identical atoms, resolved from one another, at positions $\boldsymbol{r}_j$, each of the same shape $s(\boldsymbol{r})$, so that $\rho(\boldsymbol{r}) = \sum_j s(\boldsymbol{r} - \boldsymbol{r}_j)$ and, by Equation (132.15),

$$F(\boldsymbol{G}) = f(\boldsymbol{G})\, S(\boldsymbol{G})\,, \qquad S(\boldsymbol{G}) = \sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{r}_j}\,, \tag{132.17}$$

with f the transform of s. Now square the density. The atoms do not overlap, so the cross terms vanish identically and $\rho^2(\boldsymbol{r}) = \sum_j s^2(\boldsymbol{r} - \boldsymbol{r}_j)$: squaring a set of separated equal blobs returns a set of separated equal blobs at the same places. Its structure factor is therefore $g(\boldsymbol{G})S(\boldsymbol{G})$, with g the transform of s^2, whence

$$F^{\text{sq}}(\boldsymbol{G}) = \frac{g(\boldsymbol{G})}{f(\boldsymbol{G})}\, F(\boldsymbol{G})\,. \tag{132.18}$$

On the other hand the transform of a product is the convolution of the transforms: substituting the synthesis of Equation (132.15) twice into $\int_V \rho^2 \mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{r}} \mathrm{d}^3 r$ and doing the $\boldsymbol{r}$ integral gives $F^{\text{sq}}(\boldsymbol{G}) = V^{-1} \sum_{\boldsymbol{H}} F(\boldsymbol{H}) F(\boldsymbol{G}-\boldsymbol{H})$. Equating the two expressions,

$$F(\boldsymbol{G}) = \frac{\theta(\boldsymbol{G})}{V} \sum_{\boldsymbol{H}} F(\boldsymbol{H})\, F(\boldsymbol{G}-\boldsymbol{H})\,, \qquad \theta = \frac{f}{g}\,, \tag{132.19}$$

which is Sayre's equation [1011]. It is exact, it contains no unknown but the structure factors themselves, and it is *nonlinear* — which is what lets it constrain phases, as no linear relation among the F could.

The phase relation is its leading term. For a real, spherically symmetric atomic shape both f and g are real, and over the range of momentum transfer in use both are positive, so θ is a positive number and Equation (132.19) makes the phase of $F(\boldsymbol{G})$ the phase of the sum on the right. A sum of many complex terms reaches a large modulus only if the terms dominating it are nearly in phase with one another; so wherever $|F(\boldsymbol{G})|$ is large, the dominant products $F(\boldsymbol{H})F(\boldsymbol{G}-\boldsymbol{H})$ — those whose two factors are both large — must carry, near enough, the phase of $F(\boldsymbol{G})$. Writing $F = |F|\,\mathrm{e}^{\mathrm{i}\varphi}$ and $\boldsymbol{K} = \boldsymbol{G} - \boldsymbol{H}$,

$$\Phi(\boldsymbol{H}, \boldsymbol{K}) = \varphi(\boldsymbol{H}) + \varphi(\boldsymbol{K}) - \varphi(\boldsymbol{H}+\boldsymbol{K}) \approx 0 \tag{132.20}$$

whenever all three reflections are strong.

That this particular combination is the right unknown is forced, and the argument is two lines. An intensity is unchanged by a shift of origin, so nothing an intensity measurement can ever determine may depend on where the origin was put. Moving it by $\boldsymbol{t}$ replaces $\rho(\boldsymbol{r})$ by $\rho(\boldsymbol{r}-\boldsymbol{t})$ and multiplies $F(\boldsymbol{G})$ by $\mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{t}}$, so every individual phase moves by $\boldsymbol{G}\cdot\boldsymbol{t}$; the combination in Equation (132.20) moves by $\boldsymbol{H}\cdot\boldsymbol{t}+\boldsymbol{K}\cdot\boldsymbol{t}-(\boldsymbol{H}+\boldsymbol{K})\cdot\boldsymbol{t}=0$ and does not move at all. Individual phases are thus not determinable even in principle — part of each is a choice — while triplet sums are. That is why a structure determination begins by *fixing* an origin, assigning arbitrary phases to a permitted handful of reflections, and then propagates through relations of the form Equation (132.20).

In a centrosymmetric structure with the origin at a centre, $\rho(-\boldsymbol{r})=\rho(\boldsymbol{r})$ makes every $F(\boldsymbol{G})$ real, so $\varphi\in\{0,\pi\}$ and $F=\pm|F|$; Equation (132.20) collapses to a statement about signs,

$$s(\boldsymbol{H})\,s(\boldsymbol{K})\,s(\boldsymbol{H}+\boldsymbol{K})=+1\,, \tag{132.21}$$

probably rather than certainly. How probably is the quantitative content of the method. With the moduli normalized to what point atoms at rest would give,

$$|E(\boldsymbol{G})|^2=\frac{|F(\boldsymbol{G})|^2}{\sum_j f_j(\boldsymbol{G})^2} \tag{132.22}$$

— up to a factor accounting for the symmetry of the reflection, and dividing out both the form factor and the Debye–Waller factor of Section 132.6, so that the E are the structure factors of the fictitious structure Sayre's argument assumed — the probability that Equation (132.21) holds is

$$P_+=\frac{1}{2}+\frac{1}{2}\tanh\left(\frac{|E(\boldsymbol{H})\,E(\boldsymbol{K})\,E(\boldsymbol{H}+\boldsymbol{K})|}{\sqrt{N}}\right) \tag{132.23}$$

for N equal atoms in the cell [1060], the non-centrosymmetric analogue distributing Φ about zero with a concentration $2\,|EEE|\,/\sqrt{N}$ [1059]. Two features of Equation (132.23) fix what the method can do. Only the strongest reflections are useful, since $|E|$ of order unity leaves P_+ barely above one half; and the argument falls as $N^{-1/2}$, so the relations weaken as the structure grows. Direct methods therefore took the small molecule and left the protein to the heavy atom. □

Derivation pending #273

The probability distribution of the triplet phase sum: the random-atom model in which the atomic positions are independent and uniform over the cell, the central-limit argument that makes the joint distribution of three normalized structure factors Gaussian, and the conditional distribution of the phase sum that follows from it, giving the hyperbolic tangent in the centrosymmetric case and the von Mises form in general

Phenomenon 132.8 (An image carries the phases a diffraction pattern loses)**.** The Fourier transform of a micrograph of a thin specimen reproduces the three-dimensional transform of that specimen on the central plane perpendicular to the beam, phases included, because a lens performs the back-transform physically. Micrographs at a sufficient set of orientations therefore determine the density in three dimensions with no phase problem to solve, as was first done for the tail of bacteriophage T4 [1369, 1418]. The number of distinct views required grows as the ratio of the particle diameter to the resolution sought, which is why symmetric particles were done first.

Derivation. Let $\rho(\boldsymbol{r})$ be the density and let the beam run along z. For a specimen thin enough that an electron scatters at most once, and a lens that collects the scattered and

unscattered waves and brings them back together, the image contrast is a linear functional of the projection

$$p(x,y) = \int \rho(x,y,z)\,\mathrm{d}z\,. \tag{132.24}$$

Compare the two-dimensional transform of the projection with the three-dimensional transform $\tilde{\rho}$ of the density:

$$\int p(x,y)\,\mathrm{e}^{\mathrm{i}(q_x x+q_y y)}\,\mathrm{d}x\,\mathrm{d}y = \int \rho(\boldsymbol{r})\,\mathrm{e}^{\mathrm{i}(q_x x+q_y y)}\,\mathrm{d}^3 r = \tilde{\rho}\,(q_x,q_y,0)\ , \tag{132.25}$$

the last step because the exponent is $\boldsymbol{q}\cdot\boldsymbol{r}$ with $q_z = 0$. The integration over z that forms a projection *is* the evaluation of the transform on $q_z = 0$. So one micrograph gives a central plane section of the object's transform, amplitude and phase together — the contrast with a diffraction pattern being that a micrograph is an image, formed in the plane conjugate to the object, and not a squared modulus recorded in the far field. The lens has already performed the back-transform the crystallographer cannot perform.

How many sections are needed follows from two facts. The object is bounded, of diameter D, so its transform is determined by samples at intervals $1/D$; and resolving detail of size d requires the transform out to radius $1/d$. Sections through the origin at angular spacing $\Delta\theta$ are separated at that radius by $\Delta\theta/d$, which must not exceed $1/D$, so $\Delta\theta \le d/D$ and the number of views over the available range π is

$$N_{\text{views}} \simeq \frac{\pi D}{d}\,. \tag{132.26}$$

This count is why symmetry mattered so much in the early work. A helical particle presents every azimuth at once, so a single micrograph of one particle already carries the whole set of sections, which is how the T4 tail was reconstructed from one image [1369]. □

132.8.4 Interpretation

The three routes share one feature: none of them extracts a phase from the intensities of a single diffraction pattern, because the phase is not in them. Each brings information from outside — a known sub-structure, a general property of scattering densities, or a second instrument.

The contrast that makes this sharpest is holography, which belongs with the interference experiments of Chapter 67. Let O be the wave scattered by the object, and superpose on it, before the detector, a reference wave R of known amplitude and phase and coherent with it. What the detector records is

$$|O+R|^2 = |O|^2 + |R|^2 + OR^* + O^*R\,, \tag{132.27}$$

and the last two terms are *linear* in O: they carry its phase, written into the positions of the interference fringes, which a detector of intensity reads as easily as it reads anything else. Illuminating the developed record with R again returns a transmitted field containing the term $|R|^2\,O$, which is the object wave reconstructed, alongside a conjugate image and a background [913]. Nothing has been recovered that was lost; the phase was never lost, because interference with a known wave converts a phase difference into an intensity difference. The crystallographer has no such wave — there was no coherent, splittable source at 0.1 nm in 1950, and a reference beam cannot be introduced inside a crystal — so direct methods are what one does instead: a computational substitute for a reference wave, paid for with an assumption about the object rather than an addition to the apparatus.

Heavy-atom phasing sits between the two, and reading it as a reference wave explains its one awkward feature. The heavy atom contributes to $F(\boldsymbol{G})$ a term known in amplitude

and phase once its position is known, exactly as R is known; but it travels with the object rather than beside it, so what the measurement gives is $|F|$ for the native crystal, $|F|$ for the derivative, and the known added term, and those three fix the phase of the native structure factor only up to a reflection about the phase of the added term. In the general case the ambiguity is twofold and irreducible from one derivative; a second, independent derivative gives a second pair, and the member the two pairs share is the answer. Where the projection is centrosymmetric the phase can only be 0 or π, the two branches coincide, and a single derivative settles the sign outright — which is the form in which the method entered protein crystallography [1040].

What none of this settles should be said plainly. The triplet relations are probabilistic and have no convergence proof; their reliability falls as $N^{-1/2}$ by Equation (132.23), and the atomicity that Sayre's equation assumes requires data extending to a resolution at which individual atoms are separated — which most protein crystals do not give, so proteins are still phased experimentally rather than directly. The electron route has its own reservation. An unstained thin biological specimen barely absorbs; it is a phase object, and an in-focus image of it has almost no contrast, so contrast is created by defocusing, which converts phase to amplitude through a transfer function that oscillates in spatial frequency and vanishes at a series of values. The phases obtained are the object's phases modulated by that instrument function, which has to be measured and inverted; and since the dose that would give one good image destroys the specimen, what is transformed is an average over many copies assumed identical.

Finally, the problem is not peculiar to X-rays, and this chapter has been living with it throughout. The magnetic structures that followed Section 132.2 are determined by refining a model against the observed intensities and not by inverting them, for the same reason the original measurement could be: a magnetic structure carries few parameters. Where that ceases to be true — incommensurate order, large magnetic cells, disordered moments — neutron diffraction meets the phase problem in the form the crystallographers met it, and borrows their machinery to deal with it.

132.8.5 Primary references

[747] for the vector map and [977] for the breakdown of Friedel's law. Direct methods: [1011], [962], [1020], [1089] and [1059, 1060], with the Nobel lectures [1775, 1777]. Hodgkin's structures: [926] for penicillin, [1087] for vitamin B_{12} and [1387] for insulin, surveyed in [1311]; the isomorphous-replacement method is [1040]. Three-dimensional reconstruction from micrographs: [1369] and [1418], surveyed in [1722]. Holography: [913].

Derivation pending #274

The contrast transfer function of a defocused electron microscope: the weak-phase-object approximation, the phase shift introduced by defocus and by spherical aberration, and the resulting oscillating transfer function whose zeros must be filled in by combining images taken at different defocus

Part XIII

Frontiers and Open Problems

Chapter 133

What We Observe but Do Not Understand

Every chapter of this treatise has drawn a hard line: physics enters only once Nature has voted, through a reproducible measurement. This closing chapter collects what stands on the *far* side of that line — established observational facts with no confirmed theoretical explanation — and, separately, names the theoretical programs sometimes proposed to explain them that have not yet crossed it. Every quantity below carries its measured value in Appendix C or the relevant experiment chapter; nothing here restates a number already given precisely elsewhere.

133.1 The mass and mixing puzzles

The Standard Model successfully accommodates, but does not explain, three tiers of numbers cataloged in Appendix C:

- The charged-lepton and quark masses span more than five orders of magnitude between the electron and the top quark, with the generation-to-generation ratios (Appendix C.3) fixed by Yukawa couplings the theory does not predict — the *flavor puzzle*.
- The strong CP problem (Appendix C.2): the QCD Lagrangian permits a CP-violating vacuum angle $\bar{\theta}_{\mathrm{QCD}}$ of order unity, yet the neutron electric dipole moment bounds it to $\lesssim 10^{-10}$. Nothing in the confirmed Standard Model forces this suppression.
- Neutrino oscillations establish that neutrinos are massive and mix (Appendix C.4), but oscillation data alone fix only mass-squared *differences*, never the absolute mass scale, and whether the neutrino mass term is Dirac or Majorana — decidable only by a direct observation of neutrinoless double beta decay, not yet seen — remains open (Appendix C.4).

133.2 The hierarchy problem

The Planck scale, fixed purely by dimensional analysis from $\hbar$, c, G (Proposition B.1), sits some seventeen to twenty-two orders of magnitude above every confirmed particle-physics mass scale (Appendix C.6, entries on m_e/M_{P}, m_p/M_{P}, v/M_{P}). For the Higgs vacuum expectation value specifically, this separation is a genuine puzzle rather than a mere curiosity: unlike a fermion mass, a fundamental scalar's mass is not protected by any confirmed symmetry from radiative corrections that scale quadratically with whatever ultraviolet physics completes the Standard Model. If that completion sits near the Planck scale, those corrections are technically expected to dwarf the observed electroweak scale by

many orders of magnitude, demanding an unexplained, extremely precise cancellation to keep v as small as it is measured to be. No confirmed dynamical mechanism protecting the electroweak scale in this way has been observed.

133.3 Dark matter

The gravitational evidence for non-baryonic matter is overwhelming and multiply cross-checked: galactic rotation curves, gravitational lensing by galaxy clusters, and the acoustic peak structure of the cosmic microwave background all independently require a cold dark matter density $\Omega_c h^2 = 0.1200 \pm 0.0012$ (Appendix C.5), roughly five times the baryon density. What dark matter *is* — a new particle, and if so with what mass and couplings — is not established by any confirmed direct-detection, collider, or indirect-detection signal to date. This treatise records the gravitational evidence as settled and the particle identification as open.

133.4 Dark energy and the Hubble tension

The accelerating expansion of the universe is established from Type Ia supernova distances, baryon acoustic oscillations, and the CMB angular-diameter distance, all converging on $\Omega_\Lambda = 0.6847 \pm 0.0073$ (Appendix C.5). Its microphysical origin — a true cosmological constant, vacuum energy, or a dynamical field — is unidentified. A second, independent tension sharpens the puzzle: the Hubble constant inferred from the CMB, $H_0 = (67.36 \pm 0.54)\,\mathrm{km/(s\,Mpc)}$, disagrees with late-universe distance-ladder measurements, $H_0 \approx (73.0 \pm 1.0)\,\mathrm{km/(s\,Mpc)}$, at the 4–5$\sigma$ level (Appendix C.5). Whether this Hubble tension signals unrecognized systematics or a genuine incompleteness of the base ΛCDM model is not yet resolved.

133.5 The baryon asymmetry

The observed universe is made overwhelmingly of matter, with essentially no primordial antimatter, quantified by the baryon-to-photon ratio $\eta = (6.12 \pm 0.04) \times 10^{-10}$ (Appendix C.5) — independently confirmed by Big Bang nucleosynthesis and the CMB. Generating any matter-antimatter asymmetry from an initially symmetric state requires baryon-number violation, C and CP violation, and a departure from thermal equilibrium (the Sakharov conditions). The Standard Model contains CP violation (Appendix C.3), but its magnitude, via the Jarlskog invariant (Appendix C.3), is many orders of magnitude too small to account for the observed η. No confirmed mechanism supplies the shortfall.

133.6 Confinement and the Yang–Mills mass gap

No isolated quark or gluon has ever been observed, in accelerator debris, in matter, or in cosmic rays, although the $SU(3)$ gauge Lagrangian of Chapter 102 contains only those fields; and the observed hadron spectrum begins at the pion mass rather than at zero, although the gluons of that Lagrangian are massless. Both statements are experimentally secure and neither has been derived from the theory. Lattice computation reproduces the hadron spectrum and exhibits the confining area law at strong coupling (Chapter 102), but a proof from the continuum theory — that four-dimensional Yang–Mills theory exists as a mathematically well-defined quantum field theory and that its spectrum has a mass gap $\Delta > 0$ — is missing. The statement is one of the recognized open problems of mathematical physics, posed as a Clay Millennium Prize problem [2097]. What is open here differs in

kind from the other entries of this chapter: the physics is not in doubt and no new particle or parameter is called for; what is absent is the derivation that the editorial rules of this treatise would demand of any other observed phenomenon, which is why the QCD chapter carries it as a named, pending derivation rather than as a result.

133.7 What lies outside this treatise

The puzzles above are observational facts awaiting explanation; they are evidence-based by construction, since each is a measured discrepancy or absence rather than a theoretical assertion. Several theoretical programs have been proposed, at various times, to address one or more of them: supersymmetry (a protective symmetry for the hierarchy problem), the Peccei–Quinn mechanism and its axion (for the strong CP problem), grand unified theories (for the apparent convergence of gauge couplings at high energy), string theory and its variants (for a unification including gravity), and speculative approaches to quantum gravity such as loop quantum gravity or asymptotic safety. As of this writing, **none of these programs has direct experimental confirmation**: no supersymmetric partner has been observed at any collider to date, no axion has been detected by any haloscope or helioscope search, proton decay has not been observed at the rates generic GUTs predict, and string theory makes no confirmed, falsified, or falsifiable prediction yet tested against data. Per the scope rule stated in Chapter 1, this treatise accordingly excludes them from its physics: they are named here only so that the boundary of current evidence is stated honestly, not silently. Should any of them — or some other mechanism — acquire direct experimental support, it belongs in a future edition of this chapter, exactly where the puzzle it resolves is recorded above.

A second category needs distinguishing from the first, because the two are easily confused and the scope rule treats them differently. The programs named above are candidate *physical theories* without evidence. Alongside them runs a large body of work that is *mathematics* about structures physics uses — classifying the symmetry algebras a spacetime could carry, extending them, and building gauge theories from the extensions. That work is not excluded from this treatise: its results are theorems, they are true whether or not nature uses them, and the treatise states them as such. The classification of kinematical algebras (Theorem 14.90) is in the book for exactly this reason, together with an explicit column recording which of the eleven possibilities a measurement has ever probed (Table 14.1) — three of them.

What the treatise declines to do is let the second category be read as the first. An algebra that admits a consistent extension, a Lagrangian that can be written for it, and a gauge theory that can be constructed from it are three mathematical facts and zero physical ones. Current examples include the Maxwell and non-Lorentzian extensions of the kinematical algebras and the three-dimensional Chern–Simons gravities built from them [2432]; gravity in three spacetime dimensions has no propagating degrees of freedom, no such theory has an observational test, and the papers concerned claim none. They are cited where their mathematics is used and nowhere else. The rule this implements is the one of Chapter 1: a derivation is not evidence, and an existence proof for a structure is not a measurement of it.

Chapter 134

The Dark Sector: Evidence Without Explanation

Rotation curves, gravitational lensing, cluster dynamics, the microwave-background budget and the supernova acceleration all say the same thing: gravitating matter exceeds luminous matter by about a factor of five, and the expansion of the universe is speeding up. This chapter records what is actually measured, and states honestly that no laboratory detection of a dark-matter particle exists and no mechanism for the acceleration has been identified. It is the largest quantitative discrepancy in physics between a well-tested theory — general relativity plus the Standard Model — and observation, and it is the reason the treatise has a Frontiers part at all.

The chapter is organized observation-first, on purpose. Candidate theories are named only where a search has been carried out and returned a null result or a bound, so that the reader can see which statements rest on data and which rest on nothing. It depends on the cosmological model of Chapter 48, the microwave-background measurements of Chapter 51 and the lensing formalism of Chapter 44; its particle-physics side draws on Chapter 115 and on the neutrino limits of Chapter 114. What remains unresolved is carried forward to Chapter 133; the closely related question of whether gravity itself needs modifying at the quantum level is separate and belongs to Chapter 135. A standard review of the particle side is [2073].

Derivation pending #275

The Dark Sector: all derivations of this chapter are pending.

134.1 The dynamical evidence

134.1.1 Cluster velocity dispersions

[Reserved: Zwicky's application of the virial theorem to the Coma cluster, finding a velocity dispersion far larger than the luminous mass supports and coining dunkle Materie *[729]; the more careful English restatement four years later, with the mass-to-light ratio and its assumptions made explicit [807]; the virial theorem itself from Chapter 27; why a factor of order a hundred in mass-to-light ratio was disbelieved for four decades, and what later reduced it to a factor of a few hundred in solar units.]*

134.1.2 Galaxy rotation curves

[Reserved: Rubin and Ford's spectroscopic rotation curve of Andromeda, flat far beyond the optical disc [1432]; the survey of twenty-one spirals that made the result general and

unavoidable [1658]; the Keplerian expectation $v \propto r^{-1/2}$ against the observed $v \approx$ const, and the implied enclosed mass $M(r) \propto r$; the mass discrepancy as a function of radius rather than a single global number.]

134.1.3 Neutral-hydrogen curves beyond the optical disc

[Reserved: the 21 cm hyperfine line as a kinematic tracer where there are no stars; Bosma's radio survey extending rotation curves to several optical radii [1665]; the detailed mass decomposition of NGC 3198 into disc and halo, which showed that no plausible disc mass-to-light ratio removes the need for a halo [1747]; the maximum-disc degeneracy and what breaks it.]

134.1.4 Galaxy stability and halo masses

[Reserved: the argument of Ostriker, Peebles and Yahil that cold self-gravitating discs are unstable to bar formation unless embedded in a massive halo, and their compilation of dynamical masses growing with the radius sampled [1529]; satellite kinematics and the timing argument for the Local Group; why an independent line of reasoning arriving at the same halo was what turned the discrepancy from an anomaly into a research programme.]

134.1.5 X-ray hydrostatic cluster masses

[Reserved: the hot intracluster medium in hydrostatic equilibrium, whose temperature and density profiles from Chandra give a total mass profile independent of galaxy kinematics [2105]; the resulting gas fraction of about 12 %, consistent with the cosmological baryon fraction and inconsistent with the cluster being made of what it is made of; the baryon budget compared against Section 134.3.2.]

134.2 Gravitational lensing

134.2.1 Strong lensing

[Reserved: Einstein's short note on the lens-like action of a star [774], with the deflection angle from Chapter 44; the discovery of the twin quasar 0957+561 as the first observed gravitational lens [1645]; Einstein rings and giant arcs in clusters; lensing masses as a purely geometric probe requiring no assumption of dynamical equilibrium, which is why they matter here.]

134.2.2 Weak lensing and shear surveys

[Reserved: the first detection of coherent image alignments behind clusters [1838]; the statistical measurement of cosmic shear over wide fields, mapping the projected mass distribution directly; modern survey constraints from the Dark Energy Survey [2396] and from KiDS [2374]; the shear two-point function and the parameter combination $S_8 = \sigma_8 \sqrt{\Omega_m / 0.3}$ it measures, whose mild disagreement with the microwave background is taken up in Section 134.7.2.]

134.2.3 The Bullet Cluster

[Reserved: the merging cluster 1E 0657-56, in which the X-ray emitting gas — most of the baryonic mass — is displaced from the lensing mass peaks that track the galaxies [2093]; why this is the single cleanest argument that the missing mass is not a failure of the gravitational force law, since here the mass and the baryons are spatially separated; the accompanying

bound on the dark-matter self-interaction cross-section per unit mass, of order $0.1\,\mathrm{m}^2/\mathrm{kg}$ *[2062], which constrains self-interacting candidates.]*

134.2.4 Microlensing and the compact-object bound

[Reserved: Paczyński's proposal to detect massive compact halo objects by the transient magnification of background stars [1781]; the EROS-2 survey of the Magellanic Clouds, which excluded such objects as the dominant halo component over roughly 10^{-7} *to* 10 *solar masses [2114]; the conclusion that the missing mass is not ordinary matter in dark lumps, which together with Section 134.3.2 closes off the astrophysically conservative options.]*

134.3 Cosmological evidence

134.3.1 The microwave-background budget

[Reserved: the acoustic peak structure of the temperature and polarization power spectra, and the separate roles of baryons and of pressureless dark matter in setting the peak heights — $\Omega_b h^2 = 0.0224$ against $\Omega_c h^2 = 0.120$, a factor above five [2348]; the measurement itself in Chapter 51; why a component that does not couple to photons is required by the ratio of odd to even peaks, not merely allowed by it.]

134.3.2 Nucleosynthesis and the baryon count

[Reserved: primordial abundances of deuterium, helium-3, helium-4 and lithium-7 as a one-parameter function of the baryon-to-photon ratio, and the resulting independent baryon density [2270]; its agreement with the microwave-background value to a few percent, which is a genuine concordance across an epoch ratio of about 10^5*; the lithium-7 discrepancy stated honestly as unresolved; the nuclear reaction network from Chapter 107 and Chapter 52.]*

134.3.3 Baryon acoustic oscillations

[Reserved: the same sound horizon imprinted on the galaxy distribution, first detected in the correlation function of luminous red galaxies [2076]; its use as a standard ruler and the modern spectroscopic measurement [2430] — customarily dated to 2024 by its report number, though the journal issue is 2025; the consistency of the ruler with the microwave-background calibration as a further constraint on the composition.]

134.3.4 Structure formation

[Reserved: why structure cannot grow from the observed microwave-background fluctuations without a component that begins clustering before recombination [1699]; the cold dark matter picture of hierarchical structure formation [1735]; the universal halo density profile found in simulations [1942]; large simulations reproducing the observed galaxy clustering [2083]; the honest counterweight — the core-cusp, missing-satellite and too-big-to-fail discrepancies on sub-galactic scales, which remain open and may be baryonic physics rather than a failure of the model.]

134.4 Candidates and the searches that found nothing

134.4.1 What a candidate must satisfy

[Reserved: the observational requirements — non-baryonic, cold or at least non-relativistic at decoupling, dissipationless, stable over the age of the universe, and electrically neutral;

the thermal relic calculation, in which the observed abundance follows from a weak-scale annihilation cross-section [1602], and why that coincidence motivated a generation of searches; the exclusion of Standard Model neutrinos as the dominant component by the free-streaming argument and by the mass bound of Chapter 114; the candidate catalogue and its status [2073].]

134.4.2 Direct detection: null results

[Reserved: the nuclear-recoil signature in an ultra-low background target, and the two current leading liquid-xenon experiments — LUX-ZEPLIN, with a spin-independent cross-section limit near $9 \times 10^{-52}\,\mathrm{m}^2$ at a mass of about $36\,\mathrm{GeV}/c^2$ [2413], and XENONnT with a comparable limit [2418]; four decades of improvement and no signal; the approaching neutrino fog, at which coherent solar and atmospheric neutrino scattering becomes an irreducible background; the flat statement that there is no detection.]

134.4.3 The DAMA claim and its non-confirmation

[Reserved: the annual modulation reported by DAMA/LIBRA in sodium iodide, at high statistical significance and with the phase expected from the Earth's motion through the halo [2117]; why it is inconsistent with the null results of Section 134.4.2 under standard halo and interaction assumptions; the same-target replications, which do not reproduce it [2296]; treated here as the worked example of how a positive claim is adjudicated rather than as evidence.]

134.4.4 Axions

[Reserved: the Peccei–Quinn mechanism proposed to solve the strong CP problem of Chapter 102 [1609], and the light pseudoscalar it implies [1628, 1630]; why an axion is simultaneously a cold dark matter candidate; the microwave-cavity haloscope and the ADMX exclusion over the mass range near $2.7\,\mu\mathrm{eV}$ to $3.3\,\mu\mathrm{eV}$ at Dine–Fischler–Srednicki–Zhitnitsky coupling [2309]; helioscope and light-shining-through-walls searches; again, no detection.]

134.4.5 Collider and indirect searches

[Reserved: the mono-jet plus missing transverse momentum signature at the Large Hadron Collider and its null result [2367]; the compilation of exclusion limits, including those on supersymmetric candidates [2427]; indirect searches for annihilation products in gamma rays, positrons and neutrinos, and the astrophysical backgrounds that limit them; the invisible-width bound on light candidates from Chapter 104.]

134.4.6 Primordial black holes

[Reserved: black holes formed from early-universe density perturbations as a candidate requiring no new particle [1520]; the windows closed by microlensing [2114], by microwave-background accretion constraints and by the gravitational-wave merger rate of Chapter 47; the asteroid-mass window that remains open; a fair statement of what fraction of the halo is still allowed.]

134.5 Modified dynamics as the rival hypothesis

134.5.1 MOND and the radial-acceleration relation

[Reserved: Milgrom's proposal that the discrepancy appears below an acceleration scale $a_0 \approx 1.2 \times 10^{-10}\,\mathrm{m/s}^2$ rather than beyond a length scale, and the resulting flat rotation

curves and baryonic Tully–Fisher relation [1726]; the tight empirical radial-acceleration relation between observed and baryonic centripetal acceleration across 153 galaxies [2273], which is a real regularity that any dark-matter model must also explain; the review that states the case at its strongest [2037].]

134.5.2 Where modified dynamics fails

[Reserved: the residual mass discrepancy in clusters, roughly a factor of two, which MOND does not remove; the Bullet Cluster separation of lensing mass from baryons [2093]; the relativistic completion TeVeS [2058] and its prediction of a gravitational-wave propagation speed differing from c, excluded to a part in 10^{15} by GW170817 and its electromagnetic counterpart [2278]; the microwave-background peak structure, which no modified-inertia theory has reproduced; the even-handed conclusion that the acceleration-scale regularity is unexplained by either side.]

134.6 Dark energy

134.6.1 The supernova evidence for acceleration

[Reserved: type Ia supernovae as standardizable candles after light-curve shape correction; the two independent teams' finding that distant supernovae are fainter than a decelerating universe allows [1962, 1981]; the modern compilation with about 1550 supernovae and its systematic error budget [2402]; the deceleration parameter and the transition redshift near $z \approx 0.6$; the honest caveat about progenitor evolution and dust.]

134.6.2 The cosmological constant and its problem

[Reserved: Λ as the one term general relativity permits without new physics, introduced for a static universe [445] and now measured as $\Omega_\Lambda \approx 0.685$ [2348]; the discrepancy between the observed value and any naive quantum-field-theoretic vacuum energy estimate, of order 10^{120} [1828]; anthropic and multiverse arguments recorded as what they are — not testable by any experiment now conceivable — with the bound that started them [1799]; the connection to Chapter 135.]

134.6.3 The equation of state

[Reserved: the parametrization $w = p/\rho c^2$ and its constraint $w = -1.03 \pm 0.03$ from the combination of supernovae, baryon acoustic oscillations and the microwave background [2348]; the recent spectroscopic hint of evolving w [2430], which is to be reported at its stated significance as a hint and not as a detection, since the preference depends on the supernova compilation combined with it; what a confirmed $w \neq -1$ would and would not imply.]

134.7 Tensions and the honest status

134.7.1 The Hubble tension

[Reserved: the local distance-ladder value $H_0 = (73.04 \pm 1.04)\,\mathrm{km/(s\,Mpc)}$ [2407] against the microwave-background inference $(67.36 \pm 0.54)\,\mathrm{km/(s\,Mpc)}$ [2348], a disagreement of about five standard deviations; the tip-of-the-red-giant-branch calibration, which lands between them and weakens the significance [2379]; the possibilities — an unidentified systematic on either side, or new physics before recombination — stated without preference.]

134.7.2 The S_8 tension

[Reserved: weak-lensing surveys measuring slightly less clustering than the microwave background extrapolates [2374, 2396], at the two-to-three standard deviation level; the sensitivity of the comparison to baryonic feedback in the modelling and to intrinsic alignments; why a tension at this significance is reported and not yet believed.]

134.7.3 What is and is not known

[Reserved: the closing ledger. Established beyond reasonable doubt — that gravitating mass exceeds luminous mass by about five to one on every scale from galaxies to the observable universe, that most of it is non-baryonic [2270, 2348], that it is spatially separable from the baryons [2093], and that the expansion is accelerating [1962, 1981]. Not established — the identity of the dark matter, the nature of the acceleration, and whether either requires physics beyond general relativity. Forwarded to Chapter 133.]

Chapter 135

Quantum Gravity: The Honest Status

General relativity and quantum field theory are each confirmed to extraordinary precision, and they are formulated in incompatible terms. This chapter sets out the theoretical arguments that motivate quantizing gravity, distinguishes carefully between what has been calculated and what has been measured, and states plainly that no candidate theory of quantum gravity has experimental support. The distinction is the entire point of the chapter: gravity as an effective field theory is established physics with computed predictions, black-hole thermodynamics is a theoretical result of great internal coherence and no direct observation, and the various ultraviolet completion programmes have, to date, no evidence whatsoever.

That last sentence is the operative editorial constraint. Under the scope rule of the preface, string theory, supersymmetry, loop quantum gravity and asymptotic safety are outside this treatise; they are named in Section 135.6 for exactly one purpose — to record honestly that they lack the observational support that would admit them, and to say what a measurement admitting them would have to look like. The chapter builds on the field equations of Chapter 44, the black-hole solutions of Chapter 45, the renormalization theory of Chapter 108 and Chapter 100, and the measurements of Chapter 50 and Chapter 47. What it cannot settle is carried to Chapter 133. Standard references for the established material are [1683, 1890].

Derivation pending #276

Quantum Gravity: The Honest Status: all derivations of this chapter are pending.

135.1 Why the question arises

135.1.1 Two frameworks with incompatible foundations

[Reserved: the field equations take a classical stress-energy tensor as source [1746], while matter is described by operator-valued fields [1912]; the semiclassical equation $G_{\mu\nu} = 8\pi G\, c^{-4}\langle T_{\mu\nu}\rangle$ as the only obvious patch, and its known defects — non-linearity in the state, and superluminal signalling if the expectation value is taken literally through a measurement; singularity theorems predicting the breakdown of the classical description in the interior of black holes and at the cosmological initial condition (Chapter 45).]

135.1.2 The Planck scale

[Reserved: the unique combination of $\hbar$, c and G giving a length, a time, a mass and an energy, constructed by Planck in 1899 as a system of natural units [285]; the values $\ell_P = 1.616\,255 \times 10^{-35}\,\mathrm{m}$ and $E_P = 1.22 \times 10^{28}\,\mathrm{eV}$ [2431]; the dimensional argument that the graviton loop expansion is controlled by E^2/E_P^2; the honest caveat that a scale constructed from constants is not by itself evidence that anything happens there, and the gap of some sixteen orders of magnitude to the highest collision energy reached.]

135.1.3 Must gravity be quantized at all?

[Reserved: the theoretical argument that a classical field coupled to quantum matter violates the uncertainty principle through a which-path measurement; the Page–Geilker experiment, in which a Cavendish balance was driven by a genuinely random quantum event, excluding the naive expectation-value semiclassical theory [1674]; the Wheeler–DeWitt canonical quantization as the conservative response [1344] and its problem of time; the question left genuinely open, with the proposed decisive tests in Section 135.5.1.]

135.2 What is established

135.2.1 Gravity as an effective field theory

[Reserved: the Einstein–Hilbert action treated as the leading term of a derivative expansion, with the graviton as a massless spin-2 field of Chapter 97; Deser's demonstration that a self-coupled massless spin-2 field on flat space is forced to general relativity [1420]; Donoghue's computation of the leading quantum correction to the Newtonian potential, finite and unambiguous because it comes from the long-distance non-analytic part of the loop [1880]; the size of that correction — of order $(\ell_P/r)^2$, some forty orders of magnitude below anything measurable; the correct statement that quantum gravity is calculable at low energies and that the problem is ultraviolet, not conceptual.]

135.2.2 Perturbative non-renormalizability

[Reserved: 't Hooft and Veltman's one-loop calculation, finding pure gravity finite on shell but gravity coupled to matter divergent already at one loop [1514]; Goroff and Sagnotti's two-loop computation, which produced a non-vanishing counterterm cubic in the Weyl tensor and settled the question for pure gravity [1773]; what non-renormalizability does and does not mean in the effective-field-theory language of Chapter 108 — a limited domain of validity, not an inconsistency.]

135.2.3 Quantum field theory in curved spacetime

[Reserved: quantum fields on a fixed classical background as the regime where both theories apply and agree; the absence of a preferred vacuum and hence the observer dependence of particle number; Parker's cosmological particle creation [1380]; the Unruh effect — a uniformly accelerated detector in the Minkowski vacuum registers a thermal bath at $T = \hbar a/2\pi c k_B$ [1590] — unobserved, since a temperature of $1\,\mathrm{K}$ requires an acceleration of order $1 \times 10^{20}\,\mathrm{m/s^2}$; the standard treatments [1683, 1890].]

135.2.4 Black-hole thermodynamics as theory

[Reserved: Bekenstein's identification of horizon area with entropy, from the requirement that the second law survive dropping matter into a hole [1494]; the four laws of black-hole

mechanics derived classically [1493]; Hawking's calculation that a black hole radiates with a temperature $\hbar c^3/8\pi GMk_B$, which fixed the proportionality constant at $S = k_B Ac^3/4G\hbar$ [1549]; the emphasis that every step here is theory — there is no measurement of a black-hole temperature or entropy, and the treatise records it as a derivation of great internal coherence rather than as an observed phenomenon.]

135.2.5 The information problem

[Reserved: Hawking's argument that evaporation maps a pure state to a thermal density matrix and so violates unitarity [1580]; Page's entanglement-entropy criterion for what unitary evaporation would look like [1870]; the statement that the problem is a conflict between calculations, none of which has been tested, and that its resolution is not known; why it is nevertheless the sharpest internal inconsistency available and the main reason the subject is pursued.]

135.3 What has actually been measured

135.3.1 Quantum matter in a classical gravitational field

[Reserved: the Colella–Overhauser–Werner neutron interferometer, which observed the gravitationally induced quantum phase shift and so confirmed that gravity enters the Schrödinger equation as expected [1542]; the quantization of neutron energy levels in the Earth's field [2034]; the gravitational Aharonov–Bohm phase measured with atom interferometry [2406]. These test quantum mechanics in a classical gravitational field — they are real results and they are not tests of quantum gravity, a distinction often blurred.]

135.3.2 Newton's law at short range

[Reserved: torsion-balance tests of the inverse-square law down to separations of order $50\,\mu\mathrm{m}$*, finding no deviation [2362]; what such a null result constrains and what it does not; the Cavendish lineage from Chapter 33, and the equivalence-principle tests of Chapter 42 as the companion null results.]*

135.3.3 Graviton and photon propagation speed

[Reserved: the arrival of GW170817 and its gamma-ray counterpart within about $1.7\,\mathrm{s}$ *after some* $1.3 \times 10^8\,\mathrm{yr}$ *of travel, bounding the fractional difference between the gravitational-wave and light speeds at a few parts in* 10^{15} *[2278]; the corresponding bound on the graviton mass; which classes of modified-gravity theory this excluded, tying back to Chapter 134.]*

135.3.4 Lorentz invariance at the highest energies

[Reserved: the proposal that a discrete or fluctuating spacetime would produce an energy-dependent photon speed accumulating over cosmological baselines [1947]; the Fermi observation of GRB 090510, whose gamma rays arrived with a dispersion small enough to exclude a linear Planck-scale effect [2137]; the systematic tabulation of Lorentz- and CPT-violation bounds across all sectors [2176], with the discrete symmetries themselves in Chapter 105; the honest reading — these are the only quantum-gravity-motivated predictions so far tested, and they came out negative.]

135.3.5 Classical black-hole predictions confirmed

[Reserved: the observation of binary black-hole coalescence and its ringdown [2264]; the test of the classical area theorem using the pre- and post-merger horizon areas of GW150914 [2381]; horizon-scale imaging, reported in Chapter 50; the point being that the classical sector is confirmed exactly where quantum gravity was hoped to show itself, which sharpens rather than relieves the problem.]

135.4 What has not been measured

135.4.1 Hawking radiation

[Reserved: the explicit statement that Hawking radiation [1549] has never been observed; the numbers that explain why — a solar-mass hole has a temperature of about $60\,\mathrm{nK}$*, far below the microwave background of Chapter 51, so astrophysical holes absorb far more than they emit, and an evaporation time exceeding* 10^{67} *years; the absence of any detected primordial hole of low enough mass, tying to Chapter 134; the Unruh effect [1590] likewise unobserved.]*

135.4.2 Analogue systems are analogues

[Reserved: Unruh's demonstration that a transsonic fluid flow carries an acoustic horizon obeying the same wave equation [1675]; stimulated Hawking emission measured on surface waves in a flume [2185]; the correlated emission and thermal spectrum measured in a Bose–Einstein condensate [2274]. These confirm the kinematics of the Hawking calculation in a system where the wave equation is known to hold — which is a genuine and useful result — and they say nothing about gravity, because no gravitational field is involved. The chapter will label them as analogue results wherever they appear.]

135.4.3 Gravitational decoherence and collapse models

[Reserved: the proposal that gravity induces objective state reduction, with a rate set by the gravitational self-energy difference between the superposed configurations [1926]; the underground search for the spontaneous radiation such models predict, which excluded the parameter-free version [2377]; a rare case of a quantum-gravity-adjacent hypothesis being genuinely falsified by experiment, and therefore worth recording carefully; the decoherence machinery from Chapter 87.]

135.4.4 Planck-scale interferometry

[Reserved: the conjecture of a Planck-scale transverse positional jitter accessible to correlated interferometers, and the Holometer null result excluding the proposed effect at more than 4σ *[2269]; the general lesson that the few Planck-scale conjectures sharp enough to test have been tested and excluded, while the programmes named in Section 135.6 make no prediction sharp enough to test at all.]*

135.5 Proposed experiments

135.5.1 Tabletop gravitational entanglement

[Reserved: the proposal that two masses in spatial superposition, interacting only gravitationally, become entangled if and only if the gravitational field itself has quantum degrees of freedom — the spin-entanglement witness of Bose and co-workers [2282] and the information-theoretic argument of Marletto and Vedral [2284]; the requirement that a local

classical channel cannot create entanglement, from Chapter 84; the experimental parameters needed — masses of order $1 \times 10^{-14}\,\mathrm{kg}$*, separations of order* $100\,\mu\mathrm{m}$*, and coherence times of seconds — and the plain statement that this is future work, not a result.]*

135.5.2 Other proposals

[Reserved: primordial gravitational waves as evidence of a quantized graviton field in the early universe, and the current non-detection of B-mode polarization (Chapter 51); improved Lorentz-violation searches with higher-energy transient sources [2176]; searches for a stochastic background of Planckian origin with pulsar timing; each listed with the sensitivity gain it would require.]

135.6 Programmes without evidence

135.6.1 The editorial position

[Reserved: a statement of why the following are named but not developed — the treatise admits physics with observational or experimental support, and none of these has any. Naming them without developing them is itself the honest record: a reader is entitled to know what large research programmes exist, what they would predict if true, and that no measurement to date distinguishes them from each other or from their absence.]

135.6.2 String theory

[Reserved: what it claims — a perturbatively finite graviton amplitude from extended objects, requiring extra spatial dimensions and supersymmetry [1960]; what would count as evidence — observation of supersymmetric partners, of extra dimensions in short-range gravity [2362], or of a characteristic signature in high-energy scattering; the record that none has been found, and that the extra-dimensional content also conflicts with the treatise's 3+1 dimensional scope.]

135.6.3 Supersymmetry and the collider nulls

[Reserved: the supersymmetry algebra as the unique extension of the Poincaré algebra of Chapter 94 consistent with a non-trivial S-matrix [1535]; the predicted superpartners; the exclusion limits from the Large Hadron Collider, pushing squark and gluino masses above roughly $2\,\mathrm{TeV}/c^2$ *[2427], which removes the naturalness motivation that made the weak scale the expected place to look; the honest verdict of no evidence, and the same for its local version, supergravity.]*

135.6.4 Loop quantum gravity

[Reserved: what it claims — a background-independent canonical quantization on a spin-network Hilbert space, with discrete area and volume spectra [2066]; the Lorentz-violating photon dispersion sometimes advanced as its signature, and the exclusion of the linear form [2137]; the unresolved recovery of a classical limit; no evidence.]

135.6.5 Asymptotic safety

[Reserved: Weinberg's suggestion that gravity may be non-perturbatively renormalizable at an ultraviolet fixed point of the renormalization group of Chapter 108 [1648]; the functional-renormalization evidence for such a fixed point within truncations [1961], and the truncation dependence that keeps it from being a proof; the honest status — a mathematical conjecture about a flow, with no observational consequence yet derived.]

135.7 Status

[Reserved: the closing ledger, in the same form as the dark sector chapter. Established — gravity is an effective quantum field theory with computed low-energy corrections [1880], non-renormalizable in perturbation theory [1514, 1773]; quantum fields on curved backgrounds are well understood [1890]; the classical sector is confirmed to high precision [2264, 2278, 2381]. Derived but unobserved — Hawking radiation and black-hole entropy [1494, 1549]. Neither established nor excluded — whether the gravitational field is quantum at all, with a decisive experiment proposed but not performed [2282, 2284]. Without evidence — every proposed ultraviolet completion. Forwarded to Chapter 133.]

Chapter 136

Experiment: Dark-Matter Searches and Their Null Results

Chapter 134 records the evidence that gravitating matter exceeds luminous matter by about a factor of five on every scale from galaxies to the observable universe, and names the candidates that might account for it. This chapter is the experimental face of that one, and its result is stated at the outset because it is the whole point: *there is no laboratory detection of a dark-matter particle.* Four decades of dedicated searching — deep underground detectors, resonant microwave cavities, collider missing-energy analyses and gamma-ray telescopes — have returned exclusion bounds and nothing else. One positive claim exists, the annual modulation of DAMA/LIBRA [2117], and it has not been reproduced by any other experiment, including experiments using the same target material [2296]. That discrepancy is reported here as what it is: unresolved.

A null result is evidence, and it is worth being precise about what kind. An exclusion curve is a statement about a two-dimensional parameter space — candidate mass against interaction strength — and it says that if the candidate existed with those parameters, this apparatus, with this exposure and this background, would have seen recoils that it did not see. It is therefore conditional on a halo model, on an assumed interaction, and on a background estimate, and every one of those conditions has to be stated with the bound. The searches reported below are chosen because between them they cover four different couplings of a hypothetical dark sector to the Standard Model of Chapter 104: nuclear recoil, photon conversion in a magnetic field, production at a collider, and annihilation in an astrophysical halo. They fail independently, which is what gives the combined null result its weight.

Two boundaries of scope are worth marking. This chapter does not adjudicate between a particle explanation and the modified-dynamics alternative of Chapter 134; it reports what the particle searches found. And it does not treat the neutrino, whose mass bounds from Chapter 114 and whose free-streaming length exclude it as the dominant component. What survives unexplained is carried to Chapter 133.

Derivation pending #277

Experiment: Dark-Matter Searches: the standard halo model and the nuclear form factor that together convert a measured recoil spectrum into a bound on a cross-section, and the statistical construction of a 90 per cent confidence exclusion contour in the presence of an estimated background, are pending. The kinematic and rate arguments of this chapter are derived inline.

136.1 What is being searched for, and what a null result bounds

136.1.1 The halo the searches assume

Every direct-detection bound is quoted against a common set of astrophysical assumptions, so that different experiments can be compared on one plot. The conventional choice is an isothermal sphere: a local mass density of about $5 \times 10^{-22}\,\mathrm{kg/m^3}$, a Maxwellian speed distribution in the Galactic rest frame with a dispersion corresponding to a solar motion through the halo of roughly $2.3 \times 10^{5}\,\mathrm{m/s}$, and a truncation at a local escape speed of order $5 \times 10^{5}\,\mathrm{m/s}$. The density is uncertain by a factor of about two and the velocity structure by more, so a published exclusion curve carries an astrophysical systematic that no detector improvement removes.

[Reserved: the standard halo model written out, with the velocity integral $\eta(v_{\min})$ that enters the rate; the dependence of a quoted limit on the local density, which is linear and therefore trivially rescalable, against the dependence on the high-speed tail, which is not; substructure — streams and debris flow — and why it matters more for an annual-modulation measurement than for a total-rate measurement; the Galactic rotation curve constraints of Chapter 134 as the input.]

136.1.2 Elastic recoil kinematics

The signature sought underground is a single nuclear recoil [1753]. A halo particle of mass M and speed v scattering elastically off a nucleus of mass m_N deposits, for centre-of-mass scattering angle θ, the recoil energy

$$E_R = \frac{\mu^2 v^2}{m_N}\,(1-\cos\theta)\,, \qquad \mu = \frac{M\,m_N}{M+m_N}\,, \tag{136.1}$$

so that E_R ranges from zero to

$$E_R^{\max} = \frac{2\mu^2 v^2}{m_N}\,. \tag{136.2}$$

Two features of Equation (136.2) govern the entire enterprise. First, v/c is of order 10^{-3}, so the deposited energy is smaller than the rest energy of the projectile by six orders of magnitude: a candidate of mass $100\,\mathrm{GeV}/c^2$ striking a xenon nucleus produces recoils of order tens of kilo-electronvolts, which is why every one of these detectors is built around the lowest energy threshold it can reach. Second, the reduced mass saturates. For $M \gg m_N$ one has $\mu \to m_N$ and $E_R^{\max} \to 2 m_N v^2$, independent of M; for $M \ll m_N$ one has $\mu \to M$ and $E_R^{\max} \propto M^2$, which collapses below any fixed threshold. A heavy target therefore buys signal at high mass and loses it entirely at low mass — the reason the xenon experiments of Section 136.2 and Section 136.3 and the cryogenic detectors of Section 136.4 are complementary rather than competing.

136.1.3 How to read an exclusion curve

The shape of every published spin-independent exclusion curve follows from Equation (136.2) and one rate argument, and it is worth deriving because the shape is often mistaken for a property of the apparatus. Let ρ_{loc} be the local mass density of the candidate. Its number density is $n = \rho_{\mathrm{loc}}/M$, and the event rate per unit target mass is

$$R \;\sim\; \frac{n\,\sigma\,\langle v\rangle}{m_N} \;=\; \frac{\rho_{\mathrm{loc}}\,\sigma\,\langle v\rangle}{M\,m_N}\,. \tag{136.3}$$

An experiment that observes no excess sets an upper limit on R. At high mass the kinematics of Equation (136.2) are saturated and nothing else in Equation (136.3) depends on M, so the excluded cross-section scales as

$$\sigma_{\mathrm{limit}} \;\propto\; M\,, \tag{136.4}$$

which is the straight line of unit slope that terminates every curve on a log–log plot. It is not a loss of sensitivity: it is the statement that at fixed mass density a heavier candidate is proportionately rarer. At low mass the curve instead turns upward almost vertically, because $E_R^{\max} \propto M^2$ falls below the detector threshold and the number of *detectable* recoils, not the number of recoils, goes to zero. The minimum between the two regimes sits where the reduced mass is comparable to the target mass, which is why a xenon experiment is most sensitive near a candidate mass of $30\,\mathrm{GeV}/c^2$ and a light-target experiment is most sensitive far below that.

136.2 LUX-ZEPLIN: the largest xenon target (2023)

136.2.1 Apparatus

A two-phase xenon time-projection chamber holding about seven tonnes of active liquid xenon, operated about 1.5 km underground at the Sanford Underground Research Facility so that the cosmic-ray muon flux of Chapter 115 is suppressed by several orders of magnitude. Two arrays of photomultipliers view the target from above and below; the liquid is overlaid by a gas phase across which an extraction field is applied, and the whole chamber sits inside an instrumented liquid-xenon skin and an outer gadolinium-loaded liquid-scintillator veto for neutrons.

[Reserved: the detector dimensions, the drift and extraction fields, the photomultiplier count and quantum efficiency at the xenon scintillation wavelength of 175 nm*; the radioassay campaign on every construction material, which is what sets the achievable background and is the real content of the apparatus section for an experiment of this class; the rock overburden in metres of water equivalent and the residual muon rate.]*

136.2.2 Procedure

A nuclear recoil in the liquid produces prompt scintillation, recorded as a first pulse, and a cloud of ionization electrons that the drift field carries to the surface and the extraction field pulls into the gas, where they produce a second, proportional electroluminescence pulse. The delay between the two pulses gives the depth of the event and the light pattern of the second gives its transverse position, so that every event is located in three dimensions and an inner fiducial volume can be defined in which the outer layers of xenon have absorbed the external radioactivity. The ratio of the two pulse areas discriminates nuclear recoils, produced by neutrons and by the sought signal, from electron recoils produced by gamma rays and beta decays.

[Reserved: the calibration programme — neutron sources for the nuclear-recoil band and injected radioactive sources dissolved in the xenon for the electron-recoil band; the fiducial mass and live time of the first science run; the blinding and salting protocol adopted before unblinding, which is the procedural guarantee that the analysis was not tuned to the data.]

136.2.3 Observations and data

[Reserved: the observed event distribution in the plane of ionization against scintillation, with the nuclear-recoil band drawn; the expected background from accidental coincidences, from residual radon daughters, from atmospheric and solar neutrinos and from detector

neutrons, each with its uncertainty; the number of observed events against the number expected, and the profile-likelihood test statistic from which the limit is constructed.]

Phenomenon 136.1 (No nuclear-recoil excess in seven tonnes of xenon)**.** In an exposure of about 0.9 t yr, the LUX-ZEPLIN experiment observed a distribution of events in its nuclear-recoil signal region consistent with the estimated background, and no excess attributable to halo particles scattering off xenon nuclei. Converted under the standard halo assumptions of Section 136.1.1 into a bound on a spin-independent cross-section per nucleon, the result excludes cross-sections above about $9 \times 10^{-52}\,\mathrm{m}^2$ at a candidate mass near $36\,\mathrm{GeV}/c^2$, with the bound weakening on either side in the manner derived in Section 136.1.3 [2413]. This was, on publication, the most stringent such bound over most of the mass range, and it is the strongest single statement in this chapter of what is *not* there.

Derivation. The bound quoted is a cross-section per nucleon, not per nucleus, and the conversion is the one place where an assumed interaction enters visibly. If the coupling is scalar and identical for protons and neutrons, the amplitudes of the A nucleons add coherently in the forward direction, so the nuclear cross-section carries a factor A^2 relative to the nucleonic one, modified by the reduced-mass ratio and by a form factor $F(q)$ that accounts for the loss of coherence once the momentum transfer q resolves the nuclear radius:

$$\sigma_N(q) = \sigma_n \,\frac{\mu_N^2}{\mu_n^2}\, A^2 \,\left|F(q)\right|^2 \,, \tag{136.5}$$

with μ_N and μ_n the candidate–nucleus and candidate–nucleon reduced masses. The factor A^2 is the reason xenon, with $A \approx 131$, is used at all: it buys more than four orders of magnitude in rate over a single nucleon, and it is what allows a bound near $10^{-51}\,\mathrm{m}^2$ — more than twenty orders of magnitude below a typical nuclear cross-section — to be reached with a tonne-scale target in a year. It is also the assumption that makes the bound conditional. A candidate coupling to nuclear spin rather than to mass gets no coherent enhancement, because the paired nucleons cancel and only the odd nucleon contributes; bounds on spin-dependent couplings are accordingly weaker by several orders of magnitude, and are quoted separately. □

Derivation pending #278

The form factor and the recoil spectrum: derivation of the Helm form factor as the Fourier transform of a uniform nuclear density convolved with a surface term, its first zero in relation to the recoil energies of interest for xenon, and the resulting distortion of the differential rate away from the pure exponential that the maximum-recoil relation alone would give; together with the detector response function that converts a recoil energy into an expected number of scintillation photons and ionization electrons, which is what actually sets the threshold

136.2.4 Interpretation

[Reserved: what the bound excludes and what it does not; the mass range over which the weak-scale thermal relic parameter space of Chapter 134 is now closed for the simplest scalar coupling, and the regions — low mass, spin-dependent coupling, inelastic and momentum-suppressed couplings — that survive untouched; the honest statement that a null result excludes a model, never a phenomenon, and that the gravitational evidence of Chapter 134 is undisturbed by it.]

136.2.5 Primary references

[2413].

136.3 XENONnT: the same signature in an independent detector (2023)

136.3.1 Apparatus

A two-phase xenon time-projection chamber of about six tonnes of active target, operated under the Gran Sasso massif in Italy, with a water Cherenkov muon veto and — the distinguishing feature of this apparatus — a surrounding gadolinium-loaded water neutron veto and an online cryogenic distillation column that continuously removes radon from the circulating xenon. The radon daughters dissolved in the liquid are the dominant electron-recoil background of a xenon experiment, and removing them from the target rather than rejecting them statistically is what allowed the electron-recoil rate to be pushed to a new low.

[Reserved: the active and fiducial masses, the drift field and electron lifetime, the photomultiplier arrays; the radon and krypton distillation columns and the residual specific activities achieved; the neutron-veto tagging efficiency, which converts the dominant radiogenic neutron background into a measurable and subtractable one.]

136.3.2 Procedure

Identical in principle to Section 136.2: prompt scintillation and delayed proportional ionization signals, three-dimensional event reconstruction, fiducialization, and discrimination of nuclear from electron recoils by the ratio of the two signals. The scientific value of running the two experiments in parallel lies precisely in the places where they are not identical — different rock, different veto technology, different radon-mitigation strategy, different analysis teams and different unblinding protocols — so that a common systematic is unlikely to survive in both.

[Reserved: the calibration sources and the neutron-veto efficiency measurement; the exposure and live time; the blinding scheme; the statistical treatment by which a nuclear-recoil signal is tested against a background model with several components.]

136.3.3 Observations and data

[Reserved: the observed events in the signal region against the background expectation, the achieved electron-recoil rate quoted per tonne per year per kilo-electronvolt, and the resulting exclusion curve with its minimum located in mass and cross-section; the comparison plot against Section 136.2 and the earlier generation of xenon experiments, which is where four decades of improvement becomes visible as a stack of curves descending by more than six orders of magnitude.]

Phenomenon 136.2 (Two independent xenon targets, one null result)**.** In an exposure of about 1.1 t yr XENONnT observed no excess of nuclear recoils over its background model, and set a spin-independent exclusion bound of the same order as Phenomenon 136.1 over the same mass range [2418]. The two experiments differ in overburden, in veto technology, in radon mitigation and in analysis team, and they agree. The agreement is the observation: a single null result bounds a model, but two null results obtained with independent systematics bound the model and simultaneously test each other's background estimate.

Derivation. The reason the next decade of this measurement is not simply a matter of building a larger tank is that an irreducible background is approaching, and its position in the mass–cross-section plane can be located by kinematics alone. Neutrinos scatter coherently off nuclei through the neutral current, producing exactly the signature sought: a single, unaccompanied nuclear recoil of a few kilo-electronvolts, carrying no feature by which it can be told from the sought signal event by event. For a massless projectile of energy E_ν on a nucleus at rest, energy and momentum conservation give the maximum recoil

$$E_R^{\max} = \frac{2E_\nu^2}{m_N c^2 + 2E_\nu} \;\simeq\; \frac{2E_\nu^2}{m_N c^2}\,, \tag{136.6}$$

since $E_\nu \ll m_N c^2$ for every relevant source. Setting Equation (136.6) equal to the halo-particle maximum Equation (136.2), written as $2\mu^2 c^4\,(v/c)^2\,/\,(m_N c^2)$, the nuclear mass cancels and leaves the mimicking relation

$$\mu c^2 \;\simeq\; E_\nu \,\frac{c}{v} \;\sim\; 10^3 E_\nu\,. \tag{136.7}$$

Equation (136.7) is the whole structure of the so-called neutrino fog in one line. Solar boron-8 neutrinos, of energies up to about 16 MeV, imitate candidates of mass of order $10\,\mathrm{GeV}/c^2$; atmospheric neutrinos, of energies from hundreds of MeV upwards, imitate candidates one to two orders of magnitude heavier, in precisely the region where Phenomenon 136.1 is strongest. Because this background scales with exposure exactly as the signal does, the statistical advantage of a longer run collapses once it dominates: the reach ceases to improve in proportion to the exposure and improves at best as its square root, and then more slowly still [2230]. It is a fog rather than a floor for the same reason — it degrades sensitivity progressively rather than forbidding progress [2388] — and directional detection, which uses the different arrival directions of solar neutrinos and halo particles, is the only known way through it. □

136.3.4 Interpretation

[Reserved: the combined status of the two xenon experiments and the parameter space they have jointly closed; the distance in cross-section still separating the current bounds from the atmospheric-neutrino component of Equation (136.7), which is what determines whether a next-generation experiment is worth building; the first measurements of coherent neutrino-nucleus scattering from solar boron-8 neutrinos in xenon detectors, reported in 2024, which demonstrate that the fog is not a theoretical projection but an observed background — to be cited when the primary references enter the bibliography.]

136.3.5 Primary references

[2418].

136.4 Cryogenic bolometers: the low-mass frontier

The xenon experiments lose sensitivity below a candidate mass of about $3\,\mathrm{GeV}/c^2$ for the kinematic reason derived in Section 136.1.2: a light projectile on a heavy nucleus deposits an energy falling as M^2, which passes below any threshold. The complementary technology uses light targets and thresholds several orders of magnitude lower, at the cost of target mass. A cryogenic bolometer is a single crystal held at a few tens of mK, in which a recoil is measured as a temperature rise of order microkelvin sensed by a superconducting thermometer at its transition edge. Because the phonon signal counts the deposited energy directly, rather than the small fraction of it that appears as light or charge, the threshold is set by thermal noise

rather than by a quantum efficiency, and thresholds in the range of tens of electronvolts have been achieved.

Two long-running programmes define this line. The Cryogenic Dark Matter Search operated germanium and silicon crystals reading out ionization and athermal phonons simultaneously, the ratio of the two providing event-by-event discrimination of nuclear from electron recoils in the same way that the ratio of scintillation to ionization does in xenon [2206]; its successor operates at deeper overburden with detectors optimized for masses below $1\,\mathrm{GeV}/c^2$. CRESST operates scintillating calcium tungstate crystals with simultaneous phonon and light readout, and its low-threshold detectors currently reach the lowest candidate masses of any direct-detection experiment [2267, 2325]. Neither has reported a detection.

[Reserved: this section is deliberately not written up as an `experiment`*, because the primary references for the CDMS, SuperCDMS and CRESST results are not yet in the bibliography of this treatise and the house rule forbids an experiment section without its primary source. To be written when they are entered: the detector masses, operating temperatures and thresholds; the transition-edge sensor and neutron-transmutation-doped thermistor readouts; the CDMS silicon result of 2013, in which three candidate events were observed against a smaller expected background and which was correctly reported as not constituting a detection; the CRESST low-mass exclusion curve; and the surface-event and low-energy-excess backgrounds that limit both. The compiled exclusion curves for these experiments are collected in the Particle Data Group review [2427].]*

136.5 DAMA/LIBRA and the annual-modulation claim (2008–2018)

136.5.1 Apparatus

An array of highly radiopure thallium-doped sodium iodide scintillator crystals, of total mass of order 250 kg, viewed by photomultipliers and operated under the Gran Sasso massif with passive shielding and a controlled, continuously monitored environment — temperature, radon concentration and nitrogen flux — since the claimed signal is a modulation of about one per cent with a period of one year, and any seasonal drift of the apparatus itself would imitate it. The replication experiments use the same target compound, at different sites and with active liquid-scintillator vetoes surrounding the crystals.

[Reserved: the crystal growth and radiopurity programme, which is the central technical claim of the experiment; the light yield in photoelectrons per kilo-electronvolt, which sets the analysis threshold; the shielding stack; the corresponding parameters of the replication arrays and the respects in which they differ.]

136.5.2 Procedure

The measured quantity is not a rate but the time dependence of a rate. Single-hit events — energy deposits in one crystal only, multiple hits being rejected as neutron or gamma-ray induced — are binned in energy and in time over many annual cycles, and the low-energy rate is fitted to a constant plus a cosine of period one year. Three parameters are tested against prediction: the amplitude, the period and the phase. The phase is the discriminating one, because the expected date of maximum is fixed by celestial mechanics and is not adjustable.

[Reserved: the single-hit selection and its efficiency; the energy calibration and the stability of the threshold over two decades; the treatment of the known seasonal modulation of the cosmic-ray muon flux, whose phase differs from the expected halo phase by several weeks and which is the leading conventional explanation that has been proposed and contested.]

136.5.3 Observations and data

[Reserved: the residual single-hit rate against time in the lowest energy bins, over the accumulated exposure of DAMA/NaI and DAMA/LIBRA; the fitted amplitude, period and phase with their uncertainties; the energy dependence of the modulation amplitude and its disappearance above a few kilo-electronvolts; the corresponding null fits from the replication arrays, quoted as the amplitude they exclude.]

Phenomenon 136.3 (A modulated rate that no other experiment sees)**.** The DAMA/LIBRA experiment reports, in the single-hit low-energy event rate of its sodium iodide crystals, a sinusoidal modulation with a period compatible with one year and a phase compatible with a maximum in late May or early June, sustained over many annual cycles and of combined statistical significance in excess of eight standard deviations [2117]. The modulation is a genuine feature of the data and has never been shown to be an artefact. It is nevertheless not accepted as a dark-matter detection, for two independent reasons. Interpreted as elastic spin-independent scattering under the halo model of Section 136.1.1, the implied cross-section is excluded by Phenomenon 136.1 and Phenomenon 136.2 by many orders of magnitude. And experiments using the same target compound, which are therefore not subject to the objection that a different target might respond differently, do not reproduce the modulation [2296, 2373]. Neither a dark-matter interpretation nor a conventional one is presently established: this is an unresolved experimental discrepancy, and it is reported here as one.

Derivation. The phase is the part of the claim that can be derived with certainty, and it is what makes the measurement interesting at all. The Sun moves through the Galactic halo at a speed $v_\odot$ of about $2.3 \times 10^5\,\mathrm{m/s}$; the Earth orbits the Sun at $v_\oplus = 2.98 \times 10^4\,\mathrm{m/s}$ in a plane inclined to the direction of solar motion by an angle γ of about $60°$. The Earth's speed through the halo is therefore

$$v_E(t) = v_\odot + v_\oplus \cos\gamma\, \cos[\omega\,(t - t_0)]\ , \qquad \omega = \frac{2\pi}{1\,\mathrm{yr}}\ , \tag{136.8}$$

with t_0 the date on which the orbital velocity is most nearly aligned with the solar motion, which falls in early June. The amplitude of the speed modulation is $v_\oplus \cos\gamma \approx 1.5 \times 10^4\,\mathrm{m/s}$, about 6 % of $v_\odot$. Because a detector counts only recoils above its threshold, and by Equation (136.2) only particles with v above a threshold-dependent minimum can produce them, the counted rate samples the steeply falling tail of the speed distribution; a 6 % modulation of the speed therefore produces a rate modulation of order a per cent, with the sign depending on where the threshold sits relative to the distribution. Three predictions follow with no free parameters: the period is one year, the phase is fixed by the Earth's orbit and not by the experiment, and the amplitude is of order a per cent of the rate [1771]. That all three are satisfied is why the claim could not simply be dismissed; that the implied cross-section contradicts other experiments is why it could not be accepted. □

Derivation pending #279

The quantitative incompatibility: the calculation that converts a modulation amplitude measured in sodium iodide into an allowed region in the mass–cross-section plane, including the quenching factors for sodium and for iodine recoils, which are measured quantities with their own uncertainties and which set the recoil-energy scale of the claimed region; and the comparison of that region against the exclusion curves of the xenon experiments, together with an enumeration of the non-standard halo and interaction assumptions — streams, inelastic scattering, isospin-violating couplings — that have been proposed to reconcile them and the reasons each is now disfavoured

136.5.4 Interpretation

[Reserved: how a positive claim is adjudicated when it is not reproduced — the distinction between an experiment being wrong, an experiment being right about an effect that is not dark matter, and a model-dependence in the comparison; the conventional explanations advanced, including the seasonal muon flux and seasonal detector effects, and the status of each; the methodological conclusion, which is that same-target replication is the only decisive test and that it has now been performed [2296].]

136.5.5 Primary references

[2117, 2296].

136.6 ADMX: the axion haloscope (2018–2020)

136.6.1 Apparatus

A tunable cylindrical microwave cavity of high quality factor, immersed in the bore of a superconducting solenoid producing a field of order 8 T, cooled to a few hundred mK, and read out through a near-quantum-limited amplifier. The resonant frequency is swept by moving dielectric or metallic tuning rods within the cavity, covering a band from roughly 650 MHz to 800 MHz across the two data runs reported here.

[Reserved: the cavity dimensions and unloaded quality factor; the tuning-rod mechanism and the mode map, including the mode crossings at which sensitivity is lost; the dilution refrigerator and the physical against noise temperature of the receiver; the superconducting quantum interference device and Josephson parametric amplifier readouts that took the system noise near the standard quantum limit, which is the advance that made a Dine–Fischler–Srednicki–Zhitnitsky coupling reachable.]

136.6.2 Procedure

The cavity is tuned to a frequency, power is integrated for a dwell time, the resulting spectrum is searched for a narrow excess above the thermal noise, and the cavity is retuned by a fraction of its bandwidth. A candidate excess is rescanned, and persistence under rescan, together with the expected line shape and the expected disappearance when the magnetic field is lowered, distinguishes a signal from a noise fluctuation or an environmental radio-frequency leak.

[Reserved: the dwell time and scan rate, and the trade between them that fixes the total mass range covered in a run; the candidate rescan protocol and the number of candidates that survived it, which was none; the synthetic-signal injections used to verify the analysis end to end.]

136.6.3 Observations and data

[Reserved: the excluded axion–photon coupling as a function of mass over the scanned band, against the two benchmark model bands; the achieved system noise temperature; the fraction of the band lost to mode crossings.]

Phenomenon 136.4 (No axion signal in the scanned band)**.** Scanning a tuned microwave cavity in a strong magnetic field, the Axion Dark Matter Experiment found no persistent narrow excess of microwave power over the thermal noise, and excluded halo axions with the coupling of the Dine–Fischler–Srednicki–Zhitnitsky model over the mass range covered by its two reported runs, together spanning roughly 2.7–3.3 µeV [2309, 2355]. This is the

first exclusion of that benchmark coupling by a haloscope, and it is worth stating what a narrow band it is: the theoretically motivated mass range for an axion that solves the strong CP problem of Chapter 102 and also constitutes the dark matter spans several orders of magnitude, and this experiment has closed a fraction of one of them.

Derivation. The design of the apparatus follows from two facts about the sought signal, both of which are certain given the hypothesis. First, a halo axion is non-relativistic, so its total energy is $E_a = m_a c^2 \left(1 + v^2/2c^2 + \dots\right)$; converting in a static magnetic field into a single photon [1729], it deposits that energy and no other, so the signal is a line at

$$\nu = \frac{m_a c^2}{h} \,, \tag{136.9}$$

and a cavity of fixed resonant frequency is sensitive to one axion mass only. The search must therefore be a scan, and the total time required is the time per mass step multiplied by the number of steps — which is why haloscope results cover narrow bands and why the figure of merit of such an experiment is a scan rate rather than an exposure. Second, the same expansion fixes the linewidth. The velocity dispersion of the halo gives $v/c \sim 10^{-3}$, so

$$\frac{\Delta\nu}{\nu} \simeq \frac{v^2}{2c^2} \sim 5 \times 10^{-7} \,, \tag{136.10}$$

that is, a signal quality factor of order 10^6. Two consequences follow. The cavity quality factor should be of the same order, since a narrower cavity resonance collects no more signal while a much broader one admits more noise; and the analysis must resolve the line, because a signal spread over 10^{-6} of the carrier frequency is overwhelmingly easier to find in a spectrum binned at that resolution than in the integrated power. Equation (136.10) is thus simultaneously the reason the search is feasible and the reason it is slow. □

136.6.4 Interpretation

[Reserved: what a haloscope null result excludes — a coupling at a mass, conditional on the axion constituting the whole of the local halo density, which is an assumption of a different character from those of the nuclear-recoil experiments and rescales the bound linearly if relaxed; the relation to the strong CP problem of Chapter 102, and the fact that an axion remains motivated independently of cosmology; the complementary helioscope and light-shining-through-walls searches, which do not assume the axion is the dark matter and are correspondingly weaker.]

136.6.5 Primary references

[2309, 2355].

136.7 ATLAS: missing transverse momentum at the Large Hadron Collider (2021)

136.7.1 Apparatus

The ATLAS detector at the Large Hadron Collider, in the configuration described in Chapter 104: an inner tracker in a solenoidal field, electromagnetic and hadronic calorimeters of nearly complete solid-angle coverage, and a muon spectrometer. For this search the calorimeter hermeticity is the critical property, since the observable is a momentum imbalance and any crack through which energy escapes unmeasured produces a false one.

[Reserved: the calorimeter coverage in pseudorapidity and the resulting resolution of the missing transverse momentum; the trigger thresholds; the integrated luminosity of the full Run 2 data set at a proton–proton centre-of-mass energy of 13 TeV*, and its uncertainty.]*

136.7.2 Procedure

Events are selected containing at least one energetic jet and a large imbalance of transverse momentum, with events containing charged leptons or photons vetoed. If a pair of invisible particles is produced in association with a jet radiated from the initial state, they carry away momentum that the detector cannot see, and the visible recoil against them is the signature. The dominant irreducible background is the production of a Z boson decaying to neutrinos in association with a jet, which yields exactly the same final state; it is estimated not from simulation alone but from control regions in the data — the same process with the Z decaying to charged leptons, and W production with the lepton lost — which is what makes the background estimate testable.

[Reserved: the signal region definition and its binning in missing transverse momentum; the control regions and the simultaneous fit that transfers their normalization to the signal region; the systematic uncertainties, dominated by the jet energy scale and the control-region transfer factors.]

136.7.3 Observations and data

[Reserved: observed against expected event yields in each bin of missing transverse momentum, with uncertainties; the resulting limits on the parameters of simplified mediator models, translated into the mass–cross-section plane of the direct-detection experiments for comparison; the compiled collider limits, including those on supersymmetric candidates, from the Particle Data Group review [2427].]

Phenomenon 136.5 (No excess of jets recoiling against nothing)**.** In the full Run 2 proton–proton data set, ATLAS observed event yields in the energetic-jet plus missing-transverse-momentum signal region consistent with the Standard Model expectation, dominated by $Z \to \nu\bar{\nu}$ production in association with a jet, and no excess attributable to the pair production of invisible particles [2367]. Limits on the mediator and candidate masses of simplified dark-matter models follow, and the compiled collider constraints, including those on supersymmetric candidates, are tabulated in the Particle Data Group review [2427].

Derivation. The observable is missing *transverse* momentum and not missing energy, and the reason is a kinematic fact about hadron colliders that determines the whole design of the search. The colliding objects are partons carrying unknown fractions x_1 and x_2 of their protons' longitudinal momenta, so the initial state has longitudinal momentum $(x_1 - x_2)\,p$, which is not measured event by event, and total energy $\sqrt{x_1 x_2}\,\sqrt{s}$, likewise unmeasured. Neither the longitudinal momentum balance nor the energy balance is therefore available. In the transverse plane, however, the initial-state momentum is zero up to the intrinsic transverse momentum of the partons, which is of order the confinement scale of Chapter 102 and negligible against the momenta of interest. Momentum conservation transverse to the beam then gives, for the visible particles i,

$$\boldsymbol{p}_T^{\text{miss}} = -\sum_{i \text{ visible}} \boldsymbol{p}_{T,i}\,, \tag{136.11}$$

which is a measurement, not an inference, provided the calorimeter is hermetic. Equation (136.11) also explains the requirement of a recoiling jet: a pair of invisible particles produced alone would leave a detector with nothing in it at all, which cannot be triggered on and is indistinguishable from no collision, so the search must rely on a jet radiated from the initial state and look for the imbalance against it. The essential limitation of the method is not statistical. Equation (136.11) establishes only that something left the detector without interacting — that is, a particle stable on the time scale of a detector crossing, some tens of nanoseconds. Dark matter must be stable over the age of the universe, a factor of order

10^{25} longer. A collider can therefore never, by itself, detect dark matter; it can produce a candidate, and the identification must come from elsewhere. □

136.7.4 Interpretation

[Reserved: the complementarity argument — colliders bound light candidates and light mediators where nuclear recoil is kinematically suppressed, direct detection bounds heavy candidates, and neither statement translates into the other without a model of the mediator; the model-dependence this introduces, which is of a different and stronger kind than in Section 136.2; the invisible-width constraint on candidates lighter than half the Z or Higgs boson mass from Chapter 104; and the honest note that the parameter space of the thermal relic scenario has been substantially reduced from both sides without being closed.]

136.7.5 Primary references

[2367, 2427].

136.8 Fermi-LAT: gamma rays from dwarf spheroidal galaxies (2009–)

136.8.1 Apparatus

The Large Area Telescope aboard the Fermi Gamma-ray Space Telescope: a pair-conversion instrument in which an incident gamma ray converts in one of a stack of tungsten foils interleaved with silicon-strip tracking planes, the electron–positron pair is tracked to reconstruct the arrival direction, and the energy is measured in a segmented caesium iodide calorimeter below. A segmented anticoincidence detector surrounding the tracker vetoes charged cosmic rays, which outnumber the gamma rays by several orders of magnitude. The instrument covers photon energies from about 20 MeV to above 300 GeV [2141].

[Reserved: the effective area, point-spread function and energy resolution as functions of energy and incidence angle; the survey observing mode, in which the whole sky is covered every two orbits; the anticoincidence segmentation, introduced to avoid the self-veto by backsplash that limited the previous generation of instruments.]

136.8.2 Procedure

Dwarf spheroidal satellites of the Milky Way are chosen as targets because they have the largest known mass-to-light ratios, are nearby, contain little gas and no recent star formation, and are therefore not expected to emit gamma rays by any conventional astrophysical process. The photon counts within a region around each dwarf are fitted jointly with the diffuse and point-source backgrounds, and a likelihood analysis stacking many dwarfs sets a limit on the velocity-averaged annihilation cross-section for an assumed annihilation channel and candidate mass.

[Reserved: the target selection and the kinematically determined astrophysical factor for each dwarf, with its uncertainty, which dominates the error budget; the joint-likelihood construction over targets, energies and channels; the background model for the Galactic and isotropic diffuse emission.]

136.8.3 Observations and data

[Reserved: photon counts against background expectation for the stacked sample; the resulting exclusion curves on the velocity-averaged annihilation cross-section as a function of candidate mass for the quark–antiquark and tau-pair channels, with the thermal relic value drawn

for comparison; the primary references for the stacked dwarf analyses to be added when they enter the bibliography of this treatise, the instrument paper alone being cited below. Separately: the excess of gamma rays from the Galactic centre, whose spectrum resembles annihilation but which is equally consistent with a population of unresolved millisecond pulsars, to be reported as an ambiguity and not as a signal.]

Phenomenon 136.6 (No annihilation signal from the darkest galaxies)**.** Gamma-ray observations of the dwarf spheroidal satellites of the Milky Way with the Large Area Telescope [2141] show no emission above the modelled diffuse and point-source backgrounds. Stacked likelihood analyses of these targets bound the velocity-averaged annihilation cross-section below the canonical thermal relic value of about $3 \times 10^{-32}\,\mathrm{m^3/s}$ for candidate masses up to of order $100\,\mathrm{GeV}/c^2$ in the standard annihilation channels [2245]; the compiled limits are collected in the Particle Data Group review [2427]. The significance of crossing that particular value is that it is the cross-section a thermal relic must have had in the early universe to leave the observed abundance [1602], so this is the first class of search to reach the theoretically preferred value rather than merely to approach it.

Derivation. Why dwarf spheroidals, and why the astrophysical factor is the dominant uncertainty, both follow from one line. If the candidate is its own antiparticle, the annihilation rate per unit volume is $\tfrac{1}{2}n^2\,\langle\sigma v\rangle$ with $n = \rho/M$, so the photon flux received from a direction within a solid angle $\Delta\Omega$ is

$$\Phi = \frac{\langle\sigma v\rangle}{8\pi M^2}\,\frac{\mathrm{d}N_\gamma}{\mathrm{d}E}\int_{\Delta\Omega}\int_{\text{l.o.s.}} \rho^2\,\mathrm{d}\ell\,\mathrm{d}\Omega\,, \tag{136.12}$$

the last factor being the astrophysical factor, conventionally called the J-factor, and the first two the particle-physics factor. The structure of Equation (136.12) dictates the observing strategy. The flux is *quadratic* in the density, unlike every gravitational probe in Chapter 134, which is linear; so the best targets are those that are dense and nearby rather than merely massive, and so the halo profile enters squared and its uncertainty is amplified. Dwarf spheroidals are selected because their J-factors can be estimated from stellar kinematics and because, having negligible gas and no star formation, they have no competing gamma-ray emission — the background is instrumental and diffuse, not intrinsic. The Galactic centre has a J-factor larger by orders of magnitude and is therefore a stronger target by Equation (136.12), but it is crowded with conventional gamma-ray sources, which is exactly why the excess observed there is an ambiguity rather than a detection. Finally, the M^{-2} prefactor means that a fixed flux sensitivity translates into a bound on $\langle\sigma v\rangle$ that weakens as M^2, which is why these limits cross the thermal value at low masses and rise above it at high ones. □

Derivation pending #280

The thermal relic cross-section: solution of the Boltzmann equation for a species annihilating in an expanding universe, freeze-out when the annihilation rate falls below the expansion rate, and the resulting inverse proportionality of the relic abundance to the annihilation cross-section, which is what fixes the benchmark value quoted above from the abundance measured in the dark sector chapter; and the reason the required value falls near a weak-scale cross-section, which is the coincidence that motivated a generation of searches

136.8.4 Interpretation

[Reserved: what an indirect null result excludes and how it differs from a direct one — it constrains the annihilation cross-section now rather than the scattering cross-section, so a

candidate that is not its own antiparticle, or that annihilates through a velocity-suppressed channel, escapes it entirely; the antiproton and positron channels and the astrophysical backgrounds that limit them; the Galactic-centre excess and the millisecond-pulsar explanation, presented with the evidence for each side; the neutrino channel from the solar core, bounded by neutrino telescopes.]

136.8.5 Primary references

[2141, 2427]. The stacked dwarf-spheroidal likelihood analyses on which the quoted bound rests are not yet entered in the bibliography of this treatise; only the instrument paper and the compilation are cited, and the bound is stated here at second hand.

136.9 The state of the search

Table 136.1: The searches reported in this chapter, by the coupling each would need in order to succeed. Every entry is a null result. The last column names the assumption that the bound is conditional on, so that no line of the table can be read as a bound on the phenomenon of Chapter 134 rather than on a model of it.

Search	Coupling tested	Conditional on
LUX-ZEPLIN and XENONnT	elastic scattering off a nucleus	the local halo density and speed distribution; a coherent spin-independent coupling; the nuclear form factor
Cryogenic bolometers	the same, at low candidate mass	the same, plus the detector response near threshold
DAMA/LIBRA	the same, measured as an annual modulation	the same, plus the quenching factors of sodium and iodine
ADMX	conversion to a photon in a magnetic field	the axion constituting the whole local halo density
ATLAS	production in a proton–proton collision	a mediator model relating production to scattering; stability only over a detector crossing
Fermi-LAT	annihilation to gamma rays in a halo	a self-conjugate candidate; the astrophysical factor of each dwarf; an assumed annihilation channel

The ledger of Table 136.1 should be read in two directions. Read as a bound on models, it is substantial: the simplest thermal relic scenario, in which a weak-scale particle with a weak-scale annihilation cross-section leaves the observed abundance, has been squeezed from three independent sides — scattering, production and annihilation — over most of the mass range where it was expected, and the region that motivated the whole programme is largely closed. Read as a bound on the phenomenon, it is nothing at all. The gravitational evidence assembled in Chapter 134 is untouched by every line of the table, because none of these experiments tests it: they test hypotheses about what the gravitating mass is made of, and the mass remains where it was.

That distinction is the reason this chapter exists, and it is worth stating in the form the treatise uses elsewhere. An experiment excludes a model; it does not exclude an

observation. The honest summary of four decades of searching is therefore the conjunction of two statements that are often heard separately and belong together: the gravitational discrepancy of Chapter 134 is one of the best-established facts in observational astronomy, and there is no laboratory detection of any particle that would account for it. The searches continue, and the next decade of direct detection will run into the neutrino background of Equation (136.7), at which point the method as presently practised reaches its limit and either directional detection or a different idea is required. What is carried forward to Chapter 133 is not a preferred candidate but the ledger itself.

[Reserved: a consolidated figure of the exclusion curves reported in this chapter, drawn in one mass–cross-section plane with the neutrino fog and the DAMA/LIBRA allowed region marked, to be authored as TikZ from the published tabulations; and a dataset entry in the evidence library holding the digitized curves, keyed to [2309, 2367, 2413, 2418], so that every statement in Table 136.1 can be recomputed from the numbers as published.]

Appendices

Appendix A

Long Proofs

This appendix collects the derivations that are too long to interrupt the running text. Each section opens by naming the statement it proves and closes by returning the reader to the chapter where that statement is used. The derivations are carried out in full: every algebraic step that a careful reader would want to check is displayed, and all quantities carry their SI dimensions throughout — in particular the constants G, c and $\hbar$ are kept explicit everywhere, in obedience to the SI axiom of the front matter.

A word on what *in full* can and cannot mean. Some of the results collected here descend, at exactly one step, from a theorem this treatise does not build — most often a convergence theorem of Lebesgue integration, of which Chapter 7 constructs only the Riemann predecessor, and in four cases a substantial theorem of another subject altogether. Where that happens the section does three things rather than pass over it: it states the imported theorem precisely, as a displayed environment of its own with its citation; it derives everything else from that statement; and it carries a remark headed *What is quoted here* naming the import and saying why it is not proved. A derivation that hides such a step would be worth less than the pending-derivation box it replaces. Fourteen of the thirty-two sections written for Part II quote nothing at all.

A.1 The Quadrupole Formula for Gravitational Radiation

This appendix proves Phenomenon 46.1 (Chapter 46): the field equations of general relativity, linearized about flat spacetime, admit transverse wave solutions propagating at speed c with two polarizations; these waves carry energy; and a slowly moving, self-gravitating source radiates them with a power fixed by the third time derivative of its mass quadrupole moment. The chain of results proved below — wave equation, transverse–traceless gauge, geodesic-deviation observable, Isaacson energy flux, quadrupole generation formula, and the binary-inspiral application — is exactly the chain that connects Einstein's field equations to the two great experimental confirmations: the orbital decay of the binary pulsar PSR B1913+16 (Section 47.1) and the direct detection of GW150914 by the LIGO interferometers (Section 47.3).

The history deserves one sentence. Einstein obtained the linearized wave solutions in 1916 [436], and in 1918 published the quadrupole radiation formula in its correct form, repairing a computational slip of the earlier paper [453]; the modern treatments we follow — and against which every step below can be checked — are those of Misner, Thorne and Wheeler [1509], Wald [1746], and Maggiore [2132].

Throughout this section the metric signature is $(-,+,+,+)$, coordinates are $x^{\mu} = (x^0, x^i) = (ct, \boldsymbol{x})$, Greek indices run over $0,1,2,3$, Latin indices over the spatial values $1,2,3$,

and the flat d'Alembertian is

$$\Box \equiv \eta^{\mu\nu}\partial_\mu\partial_\nu = -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \nabla^2 \,. \tag{A.1}$$

A.1.1 Setup: linearized geometry

We consider spacetimes that deviate weakly from Minkowski space, so that global coordinates exist in which

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \,, \qquad |h_{\mu\nu}| \ll 1 \,, \tag{A.2}$$

with $\eta_{\mu\nu} = \mathrm{diag}(-1,+1,+1,+1)$. All indices are raised and lowered with η, an operation that is consistent to the linear order in h at which we work. The inverse metric is

$$g^{\mu\nu} = \eta^{\mu\nu} - h^{\mu\nu} + O(h^2) \,, \tag{A.3}$$

as one checks by direct multiplication: $(\eta_{\mu\rho} + h_{\mu\rho})(\eta^{\rho\nu} - h^{\rho\nu}) = \delta_\mu{}^\nu + h_\mu{}^\nu - h_\mu{}^\nu + O(h^2) = \delta_\mu{}^\nu + O(h^2)$.

Christoffel symbols. Inserting Equations (A.2) and (A.3) into the definition of the Levi-Civita connection and discarding $O(h^2)$,

$$\begin{aligned}
\Gamma^\lambda{}_{\mu\nu} &= \frac{1}{2} g^{\lambda\rho}\left(\partial_\mu g_{\rho\nu} + \partial_\nu g_{\rho\mu} - \partial_\rho g_{\mu\nu}\right) \\
&= \frac{1}{2}\left(\eta^{\lambda\rho} - h^{\lambda\rho}\right)\left(\partial_\mu h_{\rho\nu} + \partial_\nu h_{\rho\mu} - \partial_\rho h_{\mu\nu}\right) \\
&= \frac{1}{2}\eta^{\lambda\rho}\left(\partial_\mu h_{\rho\nu} + \partial_\nu h_{\rho\mu} - \partial_\rho h_{\mu\nu}\right) + O(h^2) \,.
\end{aligned} \tag{A.4}$$

The connection is thus itself of first order in h.

Riemann tensor. In the curvature tensor $R^\rho{}_{\sigma\mu\nu} = \partial_\mu\Gamma^\rho{}_{\nu\sigma} - \partial_\nu\Gamma^\rho{}_{\mu\sigma} + \Gamma^\rho{}_{\mu\lambda}\Gamma^\lambda{}_{\nu\sigma} - \Gamma^\rho{}_{\nu\lambda}\Gamma^\lambda{}_{\mu\sigma}$ the $\Gamma\Gamma$ terms are $O(h^2)$ and drop. Substituting Equation (A.4),

$$\begin{aligned}
R^\rho{}_{\sigma\mu\nu} &= \partial_\mu\Gamma^\rho{}_{\nu\sigma} - \partial_\nu\Gamma^\rho{}_{\mu\sigma} + O(h^2) \\
&= \frac{1}{2}\eta^{\rho\lambda}\left(\partial_\mu\partial_\nu h_{\lambda\sigma} + \partial_\mu\partial_\sigma h_{\lambda\nu} - \partial_\mu\partial_\lambda h_{\nu\sigma} - \partial_\nu\partial_\mu h_{\lambda\sigma} - \partial_\nu\partial_\sigma h_{\lambda\mu} + \partial_\nu\partial_\lambda h_{\mu\sigma}\right) \\
&= \frac{1}{2}\eta^{\rho\lambda}\left(\partial_\mu\partial_\sigma h_{\lambda\nu} + \partial_\nu\partial_\lambda h_{\mu\sigma} - \partial_\mu\partial_\lambda h_{\nu\sigma} - \partial_\nu\partial_\sigma h_{\lambda\mu}\right) \,,
\end{aligned} \tag{A.5}$$

where the two $\partial_\mu\partial_\nu h_{\lambda\sigma}$ terms cancelled. Lowering the first index gives the fully covariant linearized Riemann tensor,

$$R_{\rho\sigma\mu\nu} = \frac{1}{2}\left(\partial_\mu\partial_\sigma h_{\rho\nu} + \partial_\nu\partial_\rho h_{\mu\sigma} - \partial_\mu\partial_\rho h_{\nu\sigma} - \partial_\nu\partial_\sigma h_{\rho\mu}\right) \,. \tag{A.6}$$

Ricci tensor, Ricci scalar, Einstein tensor. Contracting the first and third indices of Equation (A.5) and writing $h \equiv \eta^{\mu\nu}h_{\mu\nu}$ for the trace,

$$R_{\mu\nu} = R^\rho{}_{\mu\rho\nu} = \frac{1}{2}\left(\partial_\rho\partial_\mu h^\rho{}_\nu + \partial_\rho\partial_\nu h^\rho{}_\mu - \Box h_{\mu\nu} - \partial_\mu\partial_\nu h\right) \,, \tag{A.7}$$

and contracting once more,

$$R = \eta^{\mu\nu}R_{\mu\nu} = \partial_\mu\partial_\nu h^{\mu\nu} - \Box h \,. \tag{A.8}$$

The linearized Einstein tensor is therefore

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}R = \frac{1}{2}\left(\partial_\rho\partial_\mu h^\rho{}_\nu + \partial_\rho\partial_\nu h^\rho{}_\mu - \Box h_{\mu\nu} - \partial_\mu\partial_\nu h - \eta_{\mu\nu}\partial_\rho\partial_\sigma h^{\rho\sigma} + \eta_{\mu\nu}\Box h\right) \,. \tag{A.9}$$

Trace reversal. The expression Equation (A.9) simplifies decisively in terms of the *trace-reversed* perturbation

$$\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h\,, \qquad \bar{h} = \eta^{\mu\nu}\bar{h}_{\mu\nu} = h - 2h = -h\,, \qquad h_{\mu\nu} = \bar{h}_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}\bar{h}\,. \tag{A.10}$$

(The map is an involution: trace-reversing twice returns $h_{\mu\nu}$.) Substituting the third relation of Equation (A.10) into Equation (A.7) term by term,

$$\partial_\rho\partial_\mu h^\rho{}_\nu = \partial_\rho\partial_\mu \bar{h}^\rho{}_\nu - \frac{1}{2}\partial_\mu\partial_\nu\bar{h}\,, \quad \Box h_{\mu\nu} = \Box\bar{h}_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}\Box\bar{h}\,, \quad \partial_\mu\partial_\nu h = -\partial_\mu\partial_\nu\bar{h}\,, \tag{A.11}$$

so that

$$\begin{aligned} R_{\mu\nu} &= \frac{1}{2}\left(\partial_\rho\partial_\mu\bar{h}^\rho{}_\nu + \partial_\rho\partial_\nu\bar{h}^\rho{}_\mu - \partial_\mu\partial_\nu\bar{h} - \Box\bar{h}_{\mu\nu} + \frac{1}{2}\eta_{\mu\nu}\Box\bar{h} + \partial_\mu\partial_\nu\bar{h}\right) \\ &= \frac{1}{2}\left(\partial_\rho\partial_\mu\bar{h}^\rho{}_\nu + \partial_\rho\partial_\nu\bar{h}^\rho{}_\mu - \Box\bar{h}_{\mu\nu} + \frac{1}{2}\eta_{\mu\nu}\Box\bar{h}\right), \qquad R = \partial_\rho\partial_\sigma\bar{h}^{\rho\sigma} + \frac{1}{2}\Box\bar{h}\,, \end{aligned} \tag{A.12}$$

and the Einstein tensor assembles into

$$G_{\mu\nu} = -\frac{1}{2}\left(\Box\bar{h}_{\mu\nu} + \eta_{\mu\nu}\partial^\rho\partial^\sigma\bar{h}_{\rho\sigma} - \partial^\rho\partial_\mu\bar{h}_{\rho\nu} - \partial^\rho\partial_\nu\bar{h}_{\rho\mu}\right) + O(h^2)\,, \tag{A.13}$$

the two $\frac{1}{4}\eta_{\mu\nu}\Box\bar{h}$ contributions having cancelled between $R_{\mu\nu}$ and $-\frac{1}{2}\eta_{\mu\nu}R$. Every term now contains $\bar{h}_{\mu\nu}$ either under a d'Alembertian or under a divergence $\partial^\rho\bar{h}_{\rho\nu}$ — which is the invitation to fix a gauge.

A.1.2 Gauge freedom and the Lorenz gauge

The split Equation (A.2) is not unique: an infinitesimal coordinate change

$$x'^\mu = x^\mu + \xi^\mu(x)\,, \qquad |\partial_\mu\xi^\nu| \ll 1\,, \tag{A.14}$$

preserves the form $g = \eta + h$ with a different, equally small h. Indeed, from the tensor transformation law and $\partial x^\rho/\partial x'^\mu = \delta^\rho{}_\mu - \partial_\mu\xi^\rho + O(\xi^2)$,

$$g'_{\mu\nu} = \frac{\partial x^\rho}{\partial x'^\mu}\frac{\partial x^\sigma}{\partial x'^\nu}g_{\rho\sigma} = \eta_{\mu\nu} + h_{\mu\nu} - \partial_\mu\xi_\nu - \partial_\nu\xi_\mu + O(h\xi, \xi^2)\,, \tag{A.15}$$

so the perturbation transforms as

$$h'_{\mu\nu} = h_{\mu\nu} - \partial_\mu\xi_\nu - \partial_\nu\xi_\mu\,. \tag{A.16}$$

This is the gauge freedom of linearized gravity, the precise analogue of $A_\mu \to A_\mu - \partial_\mu\chi$ in electrodynamics [1976]. The linearized Riemann tensor Equation (A.6) is gauge *invariant*: substituting Equation (A.16),

$$\begin{aligned} \delta R_{\rho\sigma\mu\nu} = -\frac{1}{2}(&\partial_\mu\partial_\sigma\partial_\rho\xi_\nu + \partial_\mu\partial_\sigma\partial_\nu\xi_\rho + \partial_\nu\partial_\rho\partial_\mu\xi_\sigma + \partial_\nu\partial_\rho\partial_\sigma\xi_\mu \\ &- \partial_\mu\partial_\rho\partial_\nu\xi_\sigma - \partial_\mu\partial_\rho\partial_\sigma\xi_\nu - \partial_\nu\partial_\sigma\partial_\rho\xi_\mu - \partial_\nu\partial_\sigma\partial_\mu\xi_\rho) = 0\,, \end{aligned} \tag{A.17}$$

because partial derivatives commute and the eight triple-derivative terms cancel in pairs. Curvature — hence everything observable — is unaffected by the choice of ξ^μ; we may therefore spend that choice on simplifying the field equations.

Taking the trace of Equation (A.16) gives $h' = h - 2\partial_\rho\xi^\rho$, so the trace-reversed perturbation transforms as

$$\bar{h}'_{\mu\nu} = \bar{h}_{\mu\nu} - \partial_\mu\xi_\nu - \partial_\nu\xi_\mu + \eta_{\mu\nu}\partial_\rho\xi^\rho\,, \tag{A.18}$$

and its divergence as

$$\partial^\mu \bar{h}'_{\mu\nu} = \partial^\mu \bar{h}_{\mu\nu} - \Box\xi_\nu - \partial_\nu \partial^\mu \xi_\mu + \partial_\nu \partial_\rho \xi^\rho = \partial^\mu \bar{h}_{\mu\nu} - \Box\xi_\nu \,. \tag{A.19}$$

Given any $\bar{h}_{\mu\nu}$, choose ξ_ν to solve the inhomogeneous wave equation $\Box\xi_\nu = \partial^\mu \bar{h}_{\mu\nu}$ — a solution always exists, for instance by the retarded Green function of $\Box$ constructed in Appendix A.1.6 below. In the new coordinates the *Lorenz gauge* (also called harmonic or de Donder gauge) holds:

$$\partial^\mu \bar{h}_{\mu\nu} = 0 \,. \tag{A.20}$$

In this gauge the last three terms of Equation (A.13) vanish, and the full Einstein equations $G_{\mu\nu} = (8\pi G/c^4)\, T_{\mu\nu}$, linearized, collapse to a flat-space wave equation with source:

$$\boxed{\Box \bar{h}_{\mu\nu} = -\frac{16\pi G}{c^4}\, T_{\mu\nu}} \tag{A.21}$$

Taking ∂^μ of both sides and using Equation (A.20) shows $\partial^\mu T_{\mu\nu} = 0$: at linear order the source moves on the flat background, its energy–momentum conserved in the special-relativistic sense. This is the self-consistency of the linear approximation [1509, 1746].

A.1.3 Vacuum solutions: plane waves and the transverse–traceless gauge

In vacuum ($T_{\mu\nu} = 0$) equation Equation (A.21) reads $\Box \bar{h}_{\mu\nu} = 0$: each component obeys the massless wave equation, so disturbances of the metric propagate at the speed c — the first assertion of Phenomenon 46.1. Insert the plane-wave ansatz

$$\bar{h}_{\mu\nu} = A_{\mu\nu}\, \mathrm{e}^{\mathrm{i} k_\alpha x^\alpha} \,, \qquad A_{\mu\nu} = A_{\nu\mu} \text{ constant} \,, \tag{A.22}$$

(real part understood). Then $\Box \bar{h}_{\mu\nu} = -k_\alpha k^\alpha\, \bar{h}_{\mu\nu} = 0$ forces the wave vector to be null,

$$k_\alpha k^\alpha = 0 \qquad \Longleftrightarrow \qquad \omega = c\,|\boldsymbol{k}| \,, \quad k^\mu = (\omega/c, \boldsymbol{k}) \,, \tag{A.23}$$

and the Lorenz condition Equation (A.20) becomes the transversality constraint

$$k^\mu A_{\mu\nu} = 0 \,. \tag{A.24}$$

Counting the physical degrees of freedom. A symmetric $A_{\mu\nu}$ has 10 independent components; Equation (A.24) removes 4. But the gauge is not yet exhausted: any further ξ^μ with $\Box\xi^\mu = 0$ preserves Equation (A.20), by Equation (A.19). Take $\xi^\mu = B^\mu \mathrm{e}^{\mathrm{i} k_\alpha x^\alpha}$ with the same null k; then Equation (A.18) shifts the amplitude by

$$A'_{\mu\nu} = A_{\mu\nu} - \mathrm{i}\,(k_\mu B_\nu + k_\nu B_\mu - \eta_{\mu\nu} k_\rho B^\rho) \,. \tag{A.25}$$

The four free constants B_μ remove four more components. Explicitly, let the wave run along z, $k^\mu = (k, 0, 0, k)$, $k_\mu = (-k, 0, 0, k)$, and write $\beta \equiv k_\rho B^\rho = k(B_0 + B_3)$. From Equation (A.25),

$$\delta A^\mu{}_\mu = 2\mathrm{i}\beta \,, \qquad \delta A_{01} = \mathrm{i}k B_1 \,, \qquad \delta A_{02} = \mathrm{i}k B_2 \,, \qquad \delta A_{03} = \mathrm{i}k\,(B_3 - B_0) \,, \tag{A.26}$$

and the four quantities $(B_0 + B_3,\, B_1,\, B_2,\, B_3 - B_0)$ are independent linear combinations of the B_μ: the linear system is invertible. We may therefore impose

$$A^\mu{}_\mu = 0 \,, \qquad A_{0i} = 0 \quad (i = 1, 2, 3) \,. \tag{A.27}$$

The remaining component A_{00} then vanishes automatically: transversality Equation (A.24) with $\nu = 0$ reads $k^0 A_{00} + k^i A_{i0} = 0$, and $A_{i0} = 0$ with $k^0 = \omega/c \neq 0$ forces $A_{00} = 0$. The count of physical polarizations is thus

$$10 \text{ (symmetric } A_{\mu\nu}) \;-\; 4 \text{ (Lorenz)} \;-\; 4 \text{ (residual gauge)} \;=\; 2\,. \tag{A.28}$$

Because the trace vanishes, $\bar{h}_{\mu\nu} = h_{\mu\nu}$ in this gauge: bar and no-bar coincide. The conditions

$$h^{\mathrm{TT}}_{0\mu} = 0\,, \qquad h^{\mathrm{TT}\,i}{}_{i} = 0\,, \qquad \partial^j h^{\mathrm{TT}}_{ij} = 0 \tag{A.29}$$

define the *transverse–traceless* (TT) gauge. For the wave along z, transversality further gives $k(A_{0\nu} + A_{3\nu}) = 0$, hence $A_{3\nu} = -A_{0\nu} = 0$: only $A_{11} = -A_{22}$ and $A_{12} = A_{21}$ survive. Writing $A_{11} \equiv h_+$ and $A_{12} \equiv h_\times$,

$$h^{\mathrm{TT}}_{ij}(t,z) = \begin{pmatrix} h_+ & h_\times & 0 \\ h_\times & -h_+ & 0 \\ 0 & 0 & 0 \end{pmatrix} \cos[\omega\,(t - z/c)]\;, \tag{A.30}$$

the two physical polarizations of Phenomenon 46.1 [1509, 2132]. They are spin-2 objects: a rotation by ψ about the propagation axis mixes them as $h_+ \to h_+ \cos 2\psi + h_\times \sin 2\psi$, so the pattern returns to itself after half a turn.

A.1.4 Physical effect: geodesic deviation and the strain

A single free test mass tells us nothing: in TT coordinates a particle initially at rest *stays at fixed coordinates*. Its geodesic equation with initial four-velocity $u^\mu = (c,0,0,0)$ gives

$$\left.\frac{\mathrm{d}^2 x^i}{\mathrm{d}\tau^2}\right|_{\tau=0} = -\Gamma^i{}_{00}\,c^2 = -\frac{c^2}{2}\,\eta^{ij}\,(2\partial_0 h_{j0} - \partial_j h_{00}) = 0\,, \tag{A.31}$$

since $h_{0\mu} = 0$ in the TT gauge, Equation (A.29). The coordinates simply ride along with the wave. What oscillates is the *proper distance* between two such masses. Let both sit on the x axis with coordinate separation L_{c}, the wave Equation (A.30) passing along z. The proper distance at time t is

$$\begin{aligned} s(t) &= \int_0^{L_{\mathrm{c}}} \sqrt{g_{xx}}\,\mathrm{d}x = \sqrt{1 + h_+ \cos\omega t}\;L_{\mathrm{c}} \\ &= \left[1 + \frac{1}{2} h_+ \cos\omega t + O(h^2)\right] L_{\mathrm{c}}\,, \end{aligned} \tag{A.32}$$

so the fractional length change — the *strain* — is

$$\boxed{\frac{\Delta L}{L} = \frac{1}{2}\,h} \tag{A.33}$$

for a wave of amplitude h optimally oriented with respect to the separation.

The same result follows covariantly from geodesic deviation, which also exhibits the wave as a tidal force. For two neighbouring geodesics with separation vector ξ^μ and common four-velocity u^μ [1509, 1746],

$$\frac{D^2 \xi^\mu}{D\tau^2} = -R^\mu{}_{\nu\rho\sigma}\,u^\nu \xi^\rho u^\sigma\,. \tag{A.34}$$

For slowly moving masses $u^\mu \simeq (c,0,0,0)$, $\tau \simeq t$, and the only components of Riemann we need follow from Equation (A.6) with the TT conditions $h_{0\mu} = 0$:

$$R_{i0j0} = \frac{1}{2}\,(\partial_0\partial_j h_{i0} + \partial_i\partial_0 h_{0j} - \partial_i\partial_j h_{00} - \partial_0\partial_0 h_{ij}) = -\frac{1}{2c^2}\,\partial_t^2 h^{\mathrm{TT}}_{ij}\,. \tag{A.35}$$

Then Equation (A.34) gives, in the local proper frame of the first mass,

$$\ddot{\xi}^i = -c^2 R^i{}_{0j0}\,\xi^j = \frac{1}{2}\,\ddot{h}^{\mathrm{TT}}_{ij}\,\xi^j\,, \tag{A.36}$$

which integrates (for ξ nearly constant) to $\delta\xi^i = \frac{1}{2}h^{\mathrm{TT}}_{ij}\xi^j$, reproducing Equation (A.33). Note that the observable is curvature, Equation (A.35), which we proved gauge invariant: the strain is physics, not coordinates.

Applying $\delta\xi^i = \frac{1}{2}h^{\mathrm{TT}}_{ij}\xi^j$ to a ring of free masses of radius L in the xy plane, $\xi^j = L(\cos\varphi, \sin\varphi, 0)$: a pure h_+ wave displaces $(\delta x, \delta y) = \frac{1}{2}h_+ L\,(\cos\varphi, -\sin\varphi)$, squeezing the ring into an ellipse along x while stretching it along y and vice versa each half period; a pure $h_\times$ wave produces the same ellipse rotated by 45°. This quadrupolar breathing pattern is precisely what a Michelson interferometer with perpendicular arms measures: the wave lengthens one arm while it shortens the other, and the differential strain $\Delta L_x - \Delta L_y = h_+ L$ appears as an optical phase shift at the dark port. With $L = 4\,\mathrm{km}$ arms and strain sensitivity near 10^{-22} per root hertz in the band around 100 Hz, the Advanced LIGO instruments [2243] resolve arm-length changes of order $10^{-18}\,\mathrm{m}$ — the observable exploited in Section 47.3 [2264].

A.1.5 Energy carried by the wave: the Isaacson tensor

That gravitational waves carry energy was disputed into the 1950s; the resolution is that the energy is not localizable at a point (the equivalence principle forbids it) but is perfectly well defined once averaged over a few wavelengths. We follow the short-wave (Isaacson) analysis [1509, 2132].

Expand the Ricci tensor of $g = \eta + h$ to second order in h,

$$R_{\mu\nu} = R^{(1)}_{\mu\nu}[h] + R^{(2)}_{\mu\nu}[h,h] + O(h^3)\,, \tag{A.37}$$

where $R^{(1)}$ is the linear expression Equation (A.7). In vacuum, at second order, the equation $R_{\mu\nu} = 0$ says that the small residual background curvature generated at $O(h^2)$ is sourced by $-R^{(2)}_{\mu\nu}$: the wave gravitates. Averaging over a spacetime region several wavelengths across (denoted $\langle\cdot\rangle$) and defining

$$t_{\mu\nu} \equiv -\frac{c^4}{8\pi G}\left\langle R^{(2)}_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}R^{(2)}\right\rangle, \tag{A.38}$$

the coarse-grained field equations take the form $G^{(1)}_{\mu\nu}[\text{background}] = (8\pi G/c^4)\,t_{\mu\nu}$: the averaged quadratic terms act as an effective stress–energy tensor of the radiation.

The straightforward if lengthy expansion of the Ricci tensor to second order gives [2132]

$$\begin{aligned} R^{(2)}_{\mu\nu} = \frac{1}{2}\Big[&\frac{1}{2}\partial_\mu h_{\alpha\beta}\,\partial_\nu h^{\alpha\beta} + h^{\alpha\beta}\partial_\mu\partial_\nu h_{\alpha\beta} - h^{\alpha\beta}\partial_\nu\partial_\beta h_{\alpha\mu} - h^{\alpha\beta}\partial_\mu\partial_\beta h_{\alpha\nu} + h^{\alpha\beta}\partial_\alpha\partial_\beta h_{\mu\nu} \\ &+ \partial^\beta h^\alpha{}_\nu\,\partial_\beta h_{\alpha\mu} - \partial^\beta h^\alpha{}_\nu\,\partial_\alpha h_{\beta\mu} - \partial_\beta h^{\alpha\beta}\,\partial_\nu h_{\alpha\mu} - \partial_\beta h^{\alpha\beta}\,\partial_\mu h_{\alpha\nu} + \partial_\beta h^{\alpha\beta}\,\partial_\alpha h_{\mu\nu} \\ &- \frac{1}{2}\partial^\alpha h\,\partial_\alpha h_{\mu\nu} + \frac{1}{2}\partial^\alpha h\,\partial_\nu h_{\alpha\mu} + \frac{1}{2}\partial^\alpha h\,\partial_\mu h_{\alpha\nu}\Big]\,. \end{aligned} \tag{A.39}$$

Work in the TT gauge, where $h = 0$ and $\partial^\beta h_{\alpha\beta} = 0$: the last six terms of Equation (A.39) vanish identically. Under the average, total derivatives are suppressed by (wavelength)/(averaging scale) and may be dropped, which licenses integration by parts

inside $\langle\cdot\rangle$; and the field equations $\Box h_{\alpha\beta}=0$ hold. Then, term by term,

$$\left\langle h^{\alpha\beta}\partial_\mu\partial_\nu h_{\alpha\beta}\right\rangle = -\left\langle \partial_\mu h^{\alpha\beta}\,\partial_\nu h_{\alpha\beta}\right\rangle , \tag{A.40}$$

$$\left\langle h^{\alpha\beta}\partial_\nu\partial_\beta h_{\alpha\mu}\right\rangle = -\left\langle \partial_\beta h^{\alpha\beta}\,\partial_\nu h_{\alpha\mu}\right\rangle = 0 , \qquad \text{(and likewise with } \mu\leftrightarrow\nu\text{)} \tag{A.41}$$

$$\left\langle h^{\alpha\beta}\partial_\alpha\partial_\beta h_{\mu\nu}\right\rangle = -\left\langle \partial_\alpha h^{\alpha\beta}\,\partial_\beta h_{\mu\nu}\right\rangle = 0 , \tag{A.42}$$

$$\left\langle \partial^\beta h^\alpha{}_\nu\,\partial_\beta h_{\alpha\mu}\right\rangle = -\left\langle h^\alpha{}_\nu\,\Box h_{\alpha\mu}\right\rangle = 0 , \tag{A.43}$$

$$\left\langle \partial^\beta h^\alpha{}_\nu\,\partial_\alpha h_{\beta\mu}\right\rangle = -\left\langle h^\alpha{}_\nu\,\partial_\alpha\partial^\beta h_{\beta\mu}\right\rangle = 0 . \tag{A.44}$$

Only the first two terms of Equation (A.39) survive, combining to

$$\left\langle R^{(2)}_{\mu\nu}\right\rangle = \frac{1}{2}\left(\frac{1}{2}-1\right)\left\langle \partial_\mu h_{\alpha\beta}\,\partial_\nu h^{\alpha\beta}\right\rangle = -\frac{1}{4}\left\langle \partial_\mu h^{\mathrm{TT}}_{\alpha\beta}\,\partial_\nu h^{\mathrm{TT}\,\alpha\beta}\right\rangle . \tag{A.45}$$

Its trace, $\eta^{\mu\nu}\left\langle R^{(2)}_{\mu\nu}\right\rangle = \frac{1}{4}\left\langle h_{\alpha\beta}\Box h^{\alpha\beta}\right\rangle = 0$ after one more integration by parts, vanishes on shell, so Equation (A.38) yields the *Isaacson stress–energy tensor*. In the TT gauge only spatial components of h are nonzero, and

$$\boxed{\,t_{\mu\nu} = \frac{c^4}{32\pi G}\left\langle \partial_\mu h^{\mathrm{TT}}_{ij}\,\partial_\nu h^{\mathrm{TT}}_{ij}\right\rangle\,} \tag{A.46}$$

(spatial indices summed; their up/down position is immaterial in the flat background). Although derived in a particular gauge, the averaged tensor is gauge invariant to the order considered [1509, 2132].

For the plane wave Equation (A.30) travelling along z, every component is a function of $t-z/c$, so $\partial_z=-c^{-1}\partial_t$ acting on h, and the energy density and energy flux are

$$t^{00} = \frac{c^2}{32\pi G}\left\langle \dot h^{\mathrm{TT}}_{ij}\dot h^{\mathrm{TT}}_{ij}\right\rangle , \qquad F = c\,t^{0z} = \frac{c^3}{32\pi G}\left\langle \dot h^{\mathrm{TT}}_{ij}\dot h^{\mathrm{TT}}_{ij}\right\rangle = \frac{c^3}{16\pi G}\left\langle \dot h_+^2+\dot h_\times^2\right\rangle , \tag{A.47}$$

using $h^{\mathrm{TT}}_{ij}h^{\mathrm{TT}}_{ij}=2(h_+^2+h_\times^2)$ from Equation (A.30). The prefactor $c^3/16\pi G$ is enormous, $\approx 8.0\times10^{33}\,\mathrm{W/m^2}$ per unit $\left\langle \dot h^2\right\rangle/1/\mathrm{s}^2$: even a strain of 10^{-21} oscillating at 100 Hz carries a flux of order $10^{-3}\,\mathrm{W/m^2}$ — comparable to moonlight, delivered by a distortion of geometry a thousandth of a proton radius across a kilometre.

A.1.6 Generation: the quadrupole formula

We now solve the sourced equation Equation (A.21).

Retarded Green function. The retarded solution of $\Box f=-S$, with $\Box$ as in Equation (A.1), is

$$f(t,\boldsymbol{x}) = \frac{1}{4\pi}\int \frac{S(t-|\boldsymbol{x}-\boldsymbol{x}'|/c,\ \boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,\mathrm{d}^3x' . \tag{A.48}$$

To verify, consider a single spherical wave $f=g(t-r/c)/r$ centred on $\boldsymbol{x}'$, $r=|\boldsymbol{x}-\boldsymbol{x}'|$. For $r\neq0$, the radial Laplacian gives $\nabla^2 f = r^{-1}\partial_r^2(rf) = g''(t-r/c)/(c^2r) = c^{-2}\partial_t^2 f$, so the homogeneous wave equation holds; near $r=0$, $f\to g(t)/r$ and $\nabla^2(1/r)=-4\pi\delta^3(\boldsymbol{x}-\boldsymbol{x}')$ supplies a point source of strength $4\pi g(t)$. Superposing such waves with $g=S\,\mathrm{d}^3x'/4\pi$ yields Equation (A.48); causality selects the retarded rather than advanced root [1976]. Applying this to Equation (A.21) with $S=(16\pi G/c^4)T_{\mu\nu}$,

$$\bar h_{\mu\nu}(t,\boldsymbol{x}) = \frac{4G}{c^4}\int \frac{T_{\mu\nu}(t-|\boldsymbol{x}-\boldsymbol{x}'|/c,\ \boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,\mathrm{d}^3x' . \tag{A.49}$$

Far zone and slow motion. Let the source occupy a region of size d about the origin and let $r = |\boldsymbol{x}| \gg d$, with unit line of sight $\boldsymbol{n} = \boldsymbol{x}/r$. Then $|\boldsymbol{x} - \boldsymbol{x}'| = r - \boldsymbol{n}\cdot\boldsymbol{x}' + O(d^2/r)$, and keeping only the leading $1/r$,

$$\bar{h}_{\mu\nu}(t,\boldsymbol{x}) = \frac{4G}{c^4 r}\int T_{\mu\nu}\!\left(t - \frac{r}{c} + \frac{\boldsymbol{n}\cdot\boldsymbol{x}'}{c},\ \boldsymbol{x}'\right)\mathrm{d}^3x' \,. \tag{A.50}$$

If moreover the internal motions are slow, $v \ll c$ — equivalently the emitted wavelength $\lambda \sim c\,d/v$ far exceeds the source size — the retardation spread $\boldsymbol{n}\cdot\boldsymbol{x}'/c$ across the source may be dropped, and every component of $T_{\mu\nu}$ is evaluated at the single retarded time $t_r = t - r/c$.

From stress to quadrupole: the conservation identities. The spatial integral of T^{ij} is not an independent datum: flat-space conservation $\partial_\mu T^{\mu\nu} = 0$, guaranteed at this order by Equation (A.21), trades it for time derivatives of the energy density. With $x^0 = ct$, the two components of the conservation law read

$$\frac{1}{c}\partial_t T^{00} + \partial_i T^{i0} = 0\,, \qquad \frac{1}{c}\partial_t T^{0j} + \partial_i T^{ij} = 0\,. \tag{A.51}$$

Multiply the first by $x^i x^j$ and integrate over all space (the source has compact support, so boundary terms vanish under integration by parts):

$$\frac{\mathrm{d}}{\mathrm{d}t}\int T^{00}x^i x^j\,\mathrm{d}^3x = -c\int \partial_k T^{k0}\,x^i x^j\,\mathrm{d}^3x = c\int\left(T^{i0}x^j + T^{j0}x^i\right)\mathrm{d}^3x\,. \tag{A.52}$$

Multiply the second of Equation (A.51) by x^j, integrate, and symmetrize:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\left(T^{0i}x^j + T^{0j}x^i\right)\mathrm{d}^3x = -c\int\left(\partial_k T^{ki}\,x^j + \partial_k T^{kj}\,x^i\right)\mathrm{d}^3x = 2c\int T^{ij}\,\mathrm{d}^3x\,. \tag{A.53}$$

Chaining Equations (A.52) and (A.53),

$$\int T^{ij}\,\mathrm{d}^3x = \frac{1}{2c^2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\int T^{00}\,x^i x^j\,\mathrm{d}^3x = \frac{1}{2}\,\ddot{Q}_{ij}\,, \tag{A.54}$$

where, with the mass density $\rho \equiv T^{00}/c^2$ (the dominant part of T^{00} for slow motion), the *mass quadrupole moment* is

$$Q_{ij}(t) \equiv \int \rho(t,\boldsymbol{x})\,x_i x_j\,\mathrm{d}^3x\,. \tag{A.55}$$

Substituting Equation (A.54) into Equation (A.50),

$$\boxed{\bar{h}_{ij}(t,\boldsymbol{x}) = \frac{2G}{c^4 r}\,\ddot{Q}_{ij}(t - r/c)} \tag{A.56}$$

— the quadrupole formula for the field [453, 1509]. Dimensionally: $G/c^4 \sim \mathrm{s}^2/(\mathrm{kg\,m})$ and $\ddot{Q} \sim \mathrm{kg\,m}^2/\mathrm{s}^2$, so h is a pure number divided by r — a strain falling off as $1/r$, as a radiation field must. The time components need not be computed separately: the Lorenz condition Equation (A.20) determines $\bar{h}^{0\mu}$ from $\bar{h}^{ij}$ up to stationary pieces. Physically, the monopole moment $\int \rho\,\mathrm{d}^3x$ is the conserved total mass and the dipole moment $\int \rho\,x^i\,\mathrm{d}^3x$ moves uniformly by conservation of momentum, so neither can oscillate: mass conservation kills monopole radiation, momentum conservation kills dipole radiation, and the quadrupole is the first moment free to radiate [1509]. This is the deep reason gravitational radiation is so feeble compared with the electric-dipole radiation of electromagnetism.

TT projection and the reduced quadrupole. The physical waveform is the TT part of Equation (A.56). Define the projector transverse to the line of sight, $P_{ij}(\boldsymbol{n}) = \delta_{ij} - n_i n_j$ (idempotent: $P_{ik}P_{kj} = P_{ij}$), and the TT projector

$$\Lambda_{ij,kl}(\boldsymbol{n}) = P_{ik}P_{jl} - \frac{1}{2}P_{ij}P_{kl}\,, \qquad \Lambda_{ij,kl}\Lambda_{kl,mn} = \Lambda_{ij,mn}\,. \tag{A.57}$$

Because Λ annihilates any term proportional to δ_{kl}, we may freely replace Q_{ij} by its trace-free part. We adopt, here and in the luminosity below, the *reduced quadrupole moment*

$$\mathcal{I}_{ij}(t) \equiv \int \rho(t,\boldsymbol{x})\left(x_i x_j - \frac{1}{3}\,r'^2\,\delta_{ij}\right)\mathrm{d}^3x = Q_{ij} - \frac{1}{3}\,\delta_{ij}\,Q_{kk}\,, \tag{A.58}$$

(with $r'^2 = x_k x_k$ the squared distance from the origin), in terms of which

$$h_{ij}^{\mathrm{TT}}(t,\boldsymbol{x}) = \frac{2G}{c^4 r}\,\Lambda_{ij,kl}(\boldsymbol{n})\,\ddot{\mathcal{I}}_{kl}(t - r/c)\,. \tag{A.59}$$

Radiated power. Insert Equation (A.59) into the flux Equation (A.47) and integrate over a large sphere. For an outgoing wave $h \propto r^{-1} \times$ (function of $t - r/c$) one has $\partial_r h = -c^{-1}\dot{h} + O(1/r^2)$, so the radial energy flux is $c\,t^{0r} = (c^3/32\pi G)\left\langle \dot{h}_{ij}^{\mathrm{TT}}\dot{h}_{ij}^{\mathrm{TT}}\right\rangle$ and

$$P = \frac{\mathrm{d}E}{\mathrm{d}t} = \frac{c^3}{32\pi G}\,r^2\oint\left\langle \dot{h}_{ij}^{\mathrm{TT}}\dot{h}_{ij}^{\mathrm{TT}}\right\rangle \mathrm{d}\Omega = \frac{c^3}{32\pi G}\cdot\frac{4G^2}{c^8}\oint \Lambda_{kl,mn}(\boldsymbol{n})\,\langle\dddot{\mathcal{I}}_{kl}\,\dddot{\mathcal{I}}_{mn}\rangle\,\mathrm{d}\Omega\,, \tag{A.60}$$

where the projector identity of Equation (A.57) collapsed the two Λ's into one. The angular integral needs only the elementary moments of the unit vector,

$$\oint n_i n_j\,\mathrm{d}\Omega = \frac{4\pi}{3}\,\delta_{ij}\,, \qquad \oint n_i n_j n_k n_l\,\mathrm{d}\Omega = \frac{4\pi}{15}\left(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)\,, \tag{A.61}$$

fixed by isotropy and by contracting both sides. For any constant symmetric trace-free A_{ij}, expanding Equation (A.57) gives $\Lambda_{ij,kl}A_{ij}A_{kl} = A_{ij}A_{ij} - 2A_{ij}A_{ik}n_j n_k + \frac{1}{2}(n_i n_j A_{ij})^2$, and with Equation (A.61),

$$\oint \Lambda_{ij,kl}\,A_{ij}A_{kl}\,\mathrm{d}\Omega = 4\pi\left(1 - \frac{2}{3} + \frac{1}{15}\right)A_{ij}A_{ij} = \frac{8\pi}{5}\,A_{ij}A_{ij}\,. \tag{A.62}$$

Substituting into Equation (A.60),

$$\boxed{P = \frac{G}{5c^5}\,\langle\dddot{\mathcal{I}}_{ij}\,\dddot{\mathcal{I}}_{ij}\rangle} \tag{A.63}$$

— Einstein's quadrupole luminosity [453, 2132], with the reduced moment Equation (A.58) and the third time derivative evaluated at retarded time. The prefactor $G/c^5 \approx 2.75\times10^{-53}\,\mathrm{W}^{-1}$ explains at a glance why laboratory sources are hopeless and only relativistic astrophysical masses radiate detectably; its inverse $c^5/G \approx 3.6\times10^{52}\,\mathrm{W}$ sets the natural luminosity scale of strong-field gravity.

A.1.7 Application: the circular compact binary

The cleanest astrophysical source is two point masses m_1, m_2 in a circular orbit — the configuration realized, to excellent approximation, by compact binaries in their late inspiral. Work in the centre-of-mass frame with total and reduced masses

$$M = m_1 + m_2\,, \qquad \mu = \frac{m_1 m_2}{M}\,, \tag{A.64}$$

separation a, and Kepler angular frequency

$$\Omega^2 = \frac{GM}{a^3}\,. \tag{A.65}$$

With the centre of mass at the origin, $\boldsymbol{x}_1 = (m_2/M)\,\boldsymbol{x}$ and $\boldsymbol{x}_2 = -(m_1/M)\,\boldsymbol{x}$ in terms of the relative coordinate $\boldsymbol{x}(t) = a(\cos\Omega t, \sin\Omega t, 0)$, so the quadrupole of the two δ-function masses reduces to that of a single mass μ on the relative orbit:

$$Q_{ij} = m_1 x_{1i} x_{1j} + m_2 x_{2i} x_{2j} = \left(\frac{m_1 m_2^2 + m_2 m_1^2}{M^2}\right) x_i x_j = \mu\, x_i x_j\,. \tag{A.66}$$

Using $\cos^2\theta = \tfrac{1}{2}(1+\cos 2\theta)$, $\sin^2\theta = \tfrac{1}{2}(1-\cos 2\theta)$, $\sin\theta\cos\theta = \tfrac{1}{2}\sin 2\theta$, the reduced moment Equation (A.58) has time-dependent part

$$\mathcal{I}_{xx} = \frac{\mu a^2}{2}\cos 2\Omega t + \text{const}\,, \quad \mathcal{I}_{yy} = -\frac{\mu a^2}{2}\cos 2\Omega t + \text{const}\,, \quad \mathcal{I}_{xy} = \frac{\mu a^2}{2}\sin 2\Omega t\,, \tag{A.67}$$

with $\mathcal{I}_{zz}$ constant. Everything oscillates at 2Ω: the mass distribution returns to itself after half an orbit, so

$$f_{\mathrm{GW}} = 2 f_{\mathrm{orb}} = \frac{\Omega}{\pi}\,. \tag{A.68}$$

Differentiating Equation (A.67) three times,

$$\dddot{\mathcal{I}}_{xx} = 4\mu a^2\Omega^3 \sin 2\Omega t\,, \qquad \dddot{\mathcal{I}}_{yy} = -4\mu a^2\Omega^3 \sin 2\Omega t\,, \qquad \dddot{\mathcal{I}}_{xy} = -4\mu a^2\Omega^3 \cos 2\Omega t\,, \tag{A.69}$$

whence, counting $\mathcal{I}_{xy} = \mathcal{I}_{yx}$ twice,

$$\dddot{\mathcal{I}}_{ij}\,\dddot{\mathcal{I}}_{ij} = 16\mu^2 a^4\Omega^6\left(2\sin^2 2\Omega t + 2\cos^2 2\Omega t\right) = 32\,\mu^2 a^4\Omega^6\,, \tag{A.70}$$

constant in time, so the average in Equation (A.63) is trivial:

$$P = \frac{32}{5}\frac{G}{c^5}\,\mu^2 a^4\Omega^6 = \boxed{\frac{32}{5}\frac{G^4}{c^5}\,\frac{m_1^2 m_2^2\,(m_1+m_2)}{a^5}} \tag{A.71}$$

after eliminating Ω with Equation (A.65).

Orbital decay. The Newtonian orbital energy is, by the virial relation $K = \tfrac{1}{2}\mu a^2\Omega^2 = \tfrac{1}{2}\,Gm_1m_2/a = -\tfrac{1}{2}U$,

$$E = K + U = -\frac{Gm_1m_2}{2a}\,. \tag{A.72}$$

Radiation drains it, $\dot{E} = -P$; differentiating Equation (A.72) and inserting Equation (A.71),

$$\frac{Gm_1m_2}{2a^2}\,\dot{a} = -P \qquad \Longrightarrow \qquad \dot{a} = -\frac{64}{5}\,\frac{G^3}{c^5}\,\frac{m_1m_2\,(m_1+m_2)}{a^3}\,. \tag{A.73}$$

The orbit shrinks, Ω grows, $P \propto a^{-5}$ grows faster: a runaway — the *chirp*.

Chirp mass and frequency evolution. Express the evolution in the observable $f \equiv f_{\mathrm{GW}}$. From Equations (A.65) and (A.68), $\pi f = \Omega = (GM)^{1/2}a^{-3/2}$, i.e. $a = (GM)^{1/3}(\pi f)^{-2/3}$. Substituting into Equations (A.71) and (A.72) and introducing the *chirp mass*

$$\mathcal{M} \equiv \frac{(m_1m_2)^{3/5}}{(m_1+m_2)^{1/5}} = \mu^{3/5}M^{2/5}\,, \qquad \mathcal{M}^{5/3} = \mu\,M^{2/3} = m_1m_2\,M^{-1/3}\,, \tag{A.74}$$

both energy and power depend on the masses only through $\mathcal{M}$:

$$E(f) = -\frac{1}{2}\,\mathcal{M}^{5/3} G^{2/3} (\pi f)^{2/3}\,, \qquad P(f) = \frac{32}{5}\,\frac{c^5}{G}\left(\frac{\pi G \mathcal{M} f}{c^3}\right)^{10/3}. \tag{A.75}$$

Energy balance $\dot{f}\,(\mathrm{d}E/\mathrm{d}f) = -P$ with $\mathrm{d}E/\mathrm{d}f = -\frac{1}{3}\mathcal{M}^{5/3} G^{2/3} \pi^{2/3} f^{-1/3}$ then gives

$$\boxed{\dot{f} = \frac{96}{5}\,\pi^{8/3}\left(\frac{G\mathcal{M}}{c^3}\right)^{5/3} f^{11/3}} \tag{A.76}$$

— the announced law $\dot{f} \propto \mathcal{M}^{5/3} f^{11/3}$. Measuring f and $\dot{f}$ at any moment of the inspiral therefore yields the chirp mass directly, with no knowledge of the distance; the distance then follows from the amplitude. That amplitude is read off Equations (A.59) and (A.67): for the optimal (face-on) orientation the two polarizations have common magnitude

$$h = \frac{4 G \mu a^2 \Omega^2}{c^4 r} = \frac{4}{r}\,\frac{(G\mathcal{M})^{5/3}}{c^4}\left(\frac{\pi f}{c}\right)^{2/3} \cdot c^{2/3} = \frac{4}{r}\left(\frac{G\mathcal{M}}{c^2}\right)^{5/3}\left(\frac{\pi f}{c}\right)^{2/3}, \tag{A.77}$$

using $a^2\Omega^2 = GM/a$ and Equation (A.74).

Numerical estimate (i): a GW150914-like binary. Take two black holes of $30\,M_\odot$ each at luminosity distance $r = 410\,\mathrm{Mpc}$, the parameters of the first LIGO detection to within its uncertainties (the measured values were $36\,M_\odot$ and $29\,M_\odot$ at $410^{+160}_{-180}\,\mathrm{Mpc}$ [2264]). We use $G = (6.674\,30 \pm 0.000\,15) \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$ [2392], the exact $c = 299\,792\,458\,\mathrm{m/s}$ of the SI [2333], and $M_\odot = 1.989 \times 10^{30}\,\mathrm{kg}$. Then $m_1 = m_2 = 30\,M_\odot = 5.97 \times 10^{31}\,\mathrm{kg}$, $M = 60\,M_\odot = 1.193 \times 10^{32}\,\mathrm{kg}$, $\mu = 15\,M_\odot$, and

$$\mathcal{M} = \mu^{3/5} M^{2/5} = (15)^{3/5}(60)^{2/5}\,M_\odot \approx 26.1\,M_\odot = 5.19 \times 10^{31}\,\mathrm{kg}\,, \tag{A.78}$$

so that

$$\frac{G\mathcal{M}}{c^2} = \frac{6.674 \times 10^{-11} \times 5.19 \times 10^{31}}{8.988 \times 10^{16}}\,\mathrm{m} = 3.85 \times 10^4\,\mathrm{m}\,, \qquad \frac{G\mathcal{M}}{c^3} = 1.29 \times 10^{-4}\,\mathrm{s}\,. \tag{A.79}$$

Orbital geometry across the band. With $GM = 7.97 \times 10^{21}\,\mathrm{m^3/s^2}$, the separation at gravitational-wave frequency f is $a = (GM)^{1/3}(\pi f)^{-2/3}$. At the low edge of the detector band, $f = 35\,\mathrm{Hz}$: $a = 2.0 \times 10^7\,\mathrm{m/s}^{0.666\,666\,666\,666\,666\,7} \times (110\,1/\mathrm{s})^{-2/3} \approx 8.7 \times 10^5\,\mathrm{m}$ — two thirty-solar-mass black holes, each of Schwarzschild radius $2Gm/c^2 \approx 89\,\mathrm{km}$, orbiting 870 km apart at an orbital frequency of 17.5 Hz. Near merger the separation approaches a few hundred kilometres: at $a = 2.5 \times 10^5\,\mathrm{m}$,

$$\Omega = \sqrt{\frac{7.97 \times 10^{21}}{(2.5 \times 10^5)^3}}\,\mathrm{rad/s} = 714\,\mathrm{rad/s}\,, \qquad f_{\mathrm{GW}} = \frac{\Omega}{\pi} \approx 230\,\mathrm{Hz}\,, \tag{A.80}$$

reproducing the observed sweep of GW150914 from 35 Hz to about 250 Hz [2264]. *Strain.* At $f = 150\,\mathrm{Hz}$ and $r = 410\,\mathrm{Mpc} = 1.27 \times 10^{25}\,\mathrm{m}$, Equation (A.77) gives, with $(G\mathcal{M}/c^2)^{5/3} = (3.85 \times 10^4)^{5/3}\,\mathrm{m}^{5/3} = 4.4 \times 10^7\,\mathrm{m}^{5/3}$ and $(\pi f/c)^{2/3} = (1.57 \times 10^{-6})^{2/3}\,\mathrm{m}^{-2/3} = 1.35 \times 10^{-4}\,\mathrm{m}^{-2/3}$,

$$h = \frac{4 \times 4.4 \times 10^7 \times 1.35 \times 10^{-4}}{1.27 \times 10^{25}} \approx 1.9 \times 10^{-21}\,, \tag{A.81}$$

for optimal orientation; averaging over source inclination and detector antenna pattern reduces this by a factor of a few, in agreement with the observed peak strain of 1.0×10^{-21}

[2264]. By Equation (A.33) the LIGO arms of $L = 4\,\mathrm{km}$ then change length by $\Delta L = \frac{1}{2}hL \approx 2 \times 10^{-18}\,\mathrm{m}$ [2243]. *Chirp rate.* At $f = 35\,\mathrm{Hz}$, Equation (A.76) gives

$$\dot f = \frac{96}{5}\,\pi^{8/3}\left(1.29\times10^{-4}\right)^{5/3}(35)^{11/3}\,\mathrm{Hz/s} \approx 61\,\mathrm{Hz/s}\,, \tag{A.82}$$

so the residual time to coalescence, $\tau \approx \frac{3}{8} f/\dot f \approx 0.2\,\mathrm{s}$, matches the observed duration of the signal in band [2264]. All three numbers — band, amplitude, chirp — flow from Equations (A.76) and (A.77) with one chirp mass; that a single $\mathcal{M}$ fits the entire waveform is the quantitative content of the statement that GW150914 was a compact binary coalescence.

Numerical estimate (ii): orbital decay of PSR B1913+16. For a binary of period P_b, Kepler's law $P_b \propto a^{3/2}$ gives $\dot P_b/P_b = \frac{3}{2}\,\dot a/a$, and inserting Equation (A.73) with $a = (GM)^{1/3}(P_b/2\pi)^{2/3}$ yields the decay of the period. The derivation above assumed a circular orbit; for an eccentric orbit the same energy-balance computation, carried through the Keplerian ellipse harmonics, multiplies the circular result by the enhancement factor $f(e) = \left(1 + \frac{73}{24}e^2 + \frac{37}{96}e^4\right)(1-e^2)^{-7/2}$, which we quote from the standard treatment [2132]. Thus

$$\dot P_b = -\frac{192\pi}{5}\,\frac{G^{5/3}}{c^5}\left(\frac{P_b}{2\pi}\right)^{-5/3}\frac{m_p m_c}{(m_p+m_c)^{1/3}}\,f(e)\,. \tag{A.83}$$

The Hulse–Taylor pulsar PSR B1913+16, discovered in 1974 [1550], has [2275]

$$P_b = 27\,906.98\,\mathrm{s}\,,\quad e = 0.6171\,,\quad m_p = 1.438\,M_\odot\,,\quad m_c = 1.390\,M_\odot\,, \tag{A.84}$$

(masses themselves determined from two relativistic timing observables, so the test uses no free parameters). The pieces of Equation (A.83) are

$$f(e) = \frac{1 + \frac{73}{24}(0.3809) + \frac{37}{96}(0.1450)}{(0.6191)^{7/2}} = \frac{2.214}{0.1867} = 11.86\,, \tag{A.85}$$

$$\left(\frac{P_b}{2\pi}\right)^{5/3} = (4441.5)^{5/3}\,\mathrm{s}^{5/3} = 1.200\times10^{6}\,\mathrm{s}^{5/3}\,, \tag{A.86}$$

$$\frac{m_p m_c}{M^{1/3}} = \frac{7.908\times10^{60}}{1.778\times10^{10}}\,\mathrm{kg}^{5/3} = 4.447\times10^{50}\,\mathrm{kg}^{5/3}\,, \tag{A.87}$$

and assembling them with $G^{5/3} = 1.098\times10^{-17}$ and $c^5 = 2.422\times10^{42}$ (SI),

$$\dot P_b = -\frac{192\pi}{5}\times\frac{1.098\times10^{-17}\times8.33\times10^{-7}\times4.447\times10^{50}}{2.422\times10^{42}}\times11.86 \approx -2.40\times10^{-12}\,, \tag{A.88}$$

a dimensionless s/s: the orbit of period near eight hours shortens by about 76 µs per year. The full general-relativistic prediction with the precise timing parameters is $\dot P_b^{\mathrm{GR}} = (-2.402\,63 \pm 0.000\,05)\times10^{-12}$, and thirty-five years of timing give an observed intrinsic $\dot P_b = (-2.398 \pm 0.004)\times10^{-12}$ — a ratio of observed to predicted of 0.9983 ± 0.0016 [1703, 2275]. The same parameters give a present gravitational-wave luminosity $P \approx 7.8\times10^{24}\,\mathrm{W}$ from Equation (A.71) (times $f(e)$) at separation $a \approx 1.95\times10^{9}\,\mathrm{m}$ — about 2×10^{-2} of the Sun's electromagnetic output, radiated by two neutron stars orbiting within what would fit inside the Sun. Integrating Equation (A.73), the circular estimate gives coalescence after $a^4/(4\beta) \approx 5.2\times10^{16}\,\mathrm{s}$ (about 1.7 billion years, shortened to roughly 0.3 billion by the eccentricity); the binary pulsar is a young GW150914 caught three hundred million years before its chirp.

The two estimates close the loop of this proof: the strain, frequency sweep and chirp rate derived here are the quantities measured directly by the interferometers for GW150914, and the period decay Equation (A.83) is the quantity measured by pulsar timing for PSR B1913+16, matching Section 47.1. The experimental record of both confrontations — apparatus, data and uncertainties — is presented in Chapter 47.

A.2 Oppenheimer–Snyder Collapse: the Interior Matching

This appendix proves Theorem 45.23 (Chapter 45): the collapse of a uniform, pressureless, momentarily static ball of dust is an exact solution of the field equations, consisting of a contracting closed Friedmann–Lemaître–Robertson–Walker (FLRW) interior matched smoothly — induced metric and extrinsic curvature both continuous — to a Schwarzschild exterior across the falling surface [844]. The exterior half of the argument is already in the chapter: Birkhoff's theorem (Theorem 45.4) forces the outside geometry to be Equation (45.6), and the radial-plunge cycloid (Proposition 45.21) is the surface's worldline seen from outside. What remains — the interior solution, the identification of the boundary, and the junction — follows, in the treatment of [1509].

A.2.1 The interior: a contracting closed dust universe

Inside the ball the matter is homogeneous and isotropic about every comoving point, so the interior metric is of FLRW form; the momentarily static initial state will force the closed (positively curved) case, and we write it, in comoving coordinates $(\tau, \chi, \theta, \varphi)$ with $x^0 := c\tau$,

$$\mathrm{d}s^2 = -c^2\mathrm{d}\tau^2 + a^2(\tau)\left[\mathrm{d}\chi^2 + \sin^2\chi\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2\right)\right], \tag{A.89}$$

which is Equation (45.34); τ is proper time on every comoving worldline. Write $g_{ij} = a^2\gamma_{ij}$ with γ_{ij} the metric of the unit 3-sphere and primes for $\mathrm{d}/\mathrm{d}x^0$. The Christoffel symbols of Equation (A.89) are, by Equation (13.299),

$$\Gamma^0{}_{ij} = a\,a'\,\gamma_{ij}\,, \qquad \Gamma^i{}_{0j} = \frac{a'}{a}\,\delta^i{}_j\,, \qquad \Gamma^i{}_{jk} = \tilde{\Gamma}^i{}_{jk}\,, \tag{A.90}$$

with $\tilde{\Gamma}$ the Christoffel symbols of γ_{ij} alone. Contracting the Riemann tensor Equation (13.306) and using the fact that the unit 3-sphere is maximally symmetric with Ricci tensor $\tilde{R}_{ij} = 2\gamma_{ij}$,

$$R_{00} = -3\,\frac{a''}{a}\,, \qquad R_{ij} = \left(a\,a'' + 2a'^2 + 2\right)\gamma_{ij}\,. \tag{A.91}$$

Pressureless dust comoving with these coordinates has $T_{\mu\nu} = \rho\,u_\mu u_\nu$ with $u_\mu = (-c, 0, 0, 0)$, hence $T_{00} = \rho c^2$, $T_{ij} = 0$ and trace $T = -\rho c^2$. Two facts follow before any field equation is solved. First, since $\Gamma^\mu{}_{00} = 0$ in Equation (A.90), the comoving worldlines are geodesics — as they must be: $\nabla_\mu T^{\mu\nu} = 0$ for dust is exactly the statement that each grain free-falls, and it is what makes the boundary of the truncated ball a geodesic of both geometries below. Second, in the Ricci form $R_{\mu\nu} = (8\pi G/c^4)\left(T_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}T\right)$ (Chapter 44) the two components of Equation (A.91) give, restoring τ via $\dot{a} = c\,a'$,

$$\ddot{a} = -\frac{4\pi G}{3}\,\rho\,a\,, \tag{A.92}$$

$$\dot{a}^2 + c^2 = \frac{8\pi G}{3}\,\rho\,a^2\,, \tag{A.93}$$

the second obtained by eliminating a'' between the 00 and ij equations. (These are the Friedmann equations of Chapter 48 for dust and $k = +1$; the derivation is kept here so that this proof is self-contained.) Differentiating Equation (A.93) and inserting Equation (A.92) yields $\dot{\rho}/\rho = -3\dot{a}/a$, i.e.

$$\rho\,a^3 = \rho_{\mathrm{m}}\,a_{\mathrm{m}}^3 = \text{const}\,, \tag{A.94}$$

mass conservation in a contracting volume.

"Momentarily static" means $\dot{a} = 0$ at $\tau = 0$; call the initial values a_{m}, ρ_{m}. Then Equation (A.93) at $\tau = 0$ reads

$$\frac{8\pi G}{3}\,\rho_{\mathrm{m}} = \frac{c^2}{a_{\mathrm{m}}^2}\,, \tag{A.95}$$

which is why the curvature had to be positive: a momentarily static homogeneous dust ball has nowhere to go but down its own cycloid. With Equations (A.94) and (A.95), Equation (A.93) becomes $\dot{a}^2 = c^2\,(a_\mathrm{m}/a - 1)$, which the cycloid solves: substituting

$$a = \frac{a_\mathrm{m}}{2}\,(1+\cos\eta)\;, \qquad c\tau = \frac{a_\mathrm{m}}{2}\,(\eta+\sin\eta) \tag{A.96}$$

gives, exactly as in the proof of Proposition 45.21, $\dot{a}^2 = c^2(1-\cos\eta)/(1+\cos\eta) = c^2\,(a_\mathrm{m}/a-1)$. The scale factor reaches $a=0$ at $\eta=\pi$, i.e. at proper time

$$\tau_\mathrm{collapse} = \frac{\pi a_\mathrm{m}}{2c}\,, \tag{A.97}$$

the same on every comoving worldline: by homogeneity, all the dust — centre and surface alike — arrives at the singularity simultaneously in comoving time. This is part (iii) of Theorem 45.23.

A.2.2 The boundary, and the matching conditions

Truncate the interior at $\chi=\chi_0$ (with $\chi_0<\pi/2$ for a ball less than half the 3-sphere). The boundary consists of comoving dust worldlines, which are geodesics of the interior geometry by the first fact above; its areal radius — read off the angular part of Equation (A.89) — is

$$R(\tau) = a(\tau)\sin\chi_0\,, \qquad R_0 := R(0) = a_\mathrm{m}\sin\chi_0\,. \tag{A.98}$$

Outside, spherical symmetry and Theorem 45.4 force the Schwarzschild geometry Equation (45.6) with some mass parameter M; the boundary, being made of freely falling dust with no pressure acting on it from either side, must follow a radial timelike geodesic of that exterior, released from rest at R_0. By Proposition 45.21 its areal radius is the cycloid Equation (45.31),

$$R = \frac{R_0}{2}\,(1+\cos\eta)\;, \qquad \tau = \sqrt{\frac{R_0^3}{8GM}}\;(\eta+\sin\eta)\;. \tag{A.99}$$

The interior assigns the same boundary the areal radius Equation (A.98) with $a(\tau)$ the cycloid Equation (A.96): the same shape $R=(R_0/2)(1+\cos\eta)$, with proper time $\tau = (a_\mathrm{m}/2c)\,(\eta+\sin\eta)$. The two descriptions of one worldline agree for all η iff

$$\sqrt{\frac{R_0^3}{8GM}} = \frac{a_\mathrm{m}}{2c} \qquad\Longleftrightarrow\qquad r_\mathrm{s} = \frac{2GM}{c^2} = \frac{R_0^3}{a_\mathrm{m}^2} = a_\mathrm{m}\sin^3\chi_0\,, \tag{A.100}$$

which is Equation (45.35). The mass identity follows at once: by Equation (A.95),

$$\frac{4\pi}{3}\,\rho_\mathrm{m} R_0^3 = \frac{c^2}{2G\,a_\mathrm{m}^2}\,a_\mathrm{m}^3\sin^3\chi_0 = \frac{c^2}{2G}\,a_\mathrm{m}\sin^3\chi_0 = \frac{c^2 r_\mathrm{s}}{2G} = M\,. \tag{A.101}$$

The Schwarzschild mass is the initial density times $\frac{4}{3}\pi R_0^3$ — the *Euclidean* volume of areal radius R_0, not the (larger) proper volume of the curved interior. The deficit is not an error: Mc^2 is the total energy including the (negative) gravitational binding energy, and the curved-volume excess of rest mass is exactly what binding has removed.

It remains to show that agreement of the boundary worldline is the *whole* of the junction condition — that the two geometries fit with no surface layer. Two conditions must hold on the matching hypersurface: the induced metrics must agree, and the extrinsic curvatures must agree.

Induced metric. Coordinatize the boundary by (τ,θ,φ). From inside, Equation (A.89) restricted to $\chi=\chi_0$ gives $-c^2\mathrm{d}\tau^2 + a^2\sin^2\chi_0\,\mathrm{d}\Omega^2 = -c^2\mathrm{d}\tau^2 + R(\tau)^2\mathrm{d}\Omega^2$. From outside,

along a radial worldline parametrized by its proper time at areal radius $R(\tau)$, the induced metric is likewise $-c^2\mathrm{d}\tau^2 + R(\tau)^2\mathrm{d}\Omega^2$. Given Equation (A.100) the two functions $R(\tau)$ coincide, so the induced metrics agree identically.

Extrinsic curvature. By spherical symmetry the extrinsic curvature has two independent components, $K^{\tau}{}_{\tau}$ and $K^{\theta}{}_{\theta} = K^{\varphi}{}_{\varphi}$. The $\tau\tau$ component measures the normal component of the boundary worldlines' acceleration, and vanishes on *both* sides because the boundary is geodesic in both geometries — comoving dust inside, radial free fall outside. For the angular component, from inside the outward unit normal is $n = a^{-1}\partial_\chi$; for a hypersurface of constant χ in a metric with no χ-cross terms, $K_{ij} = \frac{1}{2}\, n^{\chi}\partial_\chi g_{ij}$, and with $g_{\theta\theta} = a^2\sin^2\chi$,

$$K^{\theta}{}_{\theta} = \frac{K_{\theta\theta}}{g_{\theta\theta}} = \frac{a\sin\chi_0\cos\chi_0}{a^2\sin^2\chi_0} = \frac{\cos\chi_0}{a\sin\chi_0} = \frac{\cos\chi_0}{R}\,. \tag{A.102}$$

From outside, write $f := 1 - r_{\mathrm{s}}/R$ and let $u^\mu = (u^t, u^r)$ be the falling boundary's four-velocity, with $u^t = \mathcal{E}/(fc^2)$ and $c^2(u^r)^2 = \mathcal{E}^2 - fc^4$ (Proposition 45.21). The outward unit normal $n^\mu = (n^t, n^r)$ is fixed by $n_\mu u^\mu = 0$ and $n_\mu n^\mu = 1$: orthogonality gives $n^t = u^r n^r/(f^2c^2u^t)$, and substituting into the normalization $-fc^2(n^t)^2 + f^{-1}(n^r)^2 = 1$,

$$(n^r)^2\left[\frac{1}{f} - \frac{(u^r)^2}{f^3c^2\,(u^t)^2}\right] = \frac{(n^r)^2}{f}\,\frac{\mathcal{E}^2 - c^2\,(u^r)^2}{\mathcal{E}^2} = (n^r)^2\,\frac{c^4}{\mathcal{E}^2} = 1\,,$$

using $f^3c^2(u^t)^2 = f\,\mathcal{E}^2/c^2$ in the first step and $\mathcal{E}^2 - c^2(u^r)^2 = fc^4$ in the second; so $n^r = \mathcal{E}/c^2$. Since the only angular Christoffel symbol involved is $\Gamma^{\theta}{}_{r\theta} = 1/R$ (Equation (45.2)),

$$K^{\theta}{}_{\theta} = n^r\,\Gamma^{\theta}{}_{r\theta} = \frac{\mathcal{E}}{c^2R} = \frac{\sqrt{1 - r_{\mathrm{s}}/R_0}}{R}\,. \tag{A.103}$$

The two expressions Equations (A.102) and (A.103) agree iff $\cos\chi_0 = \sqrt{1 - r_{\mathrm{s}}/R_0}$, i.e. iff $r_{\mathrm{s}} = R_0\sin^2\chi_0 = a_{\mathrm{m}}\sin^3\chi_0$ — precisely the condition Equation (A.100) already imposed by the worldline. The junction is therefore smooth, with no surface shell, exactly when Equation (45.35) holds; parts (i) and (ii) of Theorem 45.23 are proved.

A.2.3 The global picture

The matched spacetime settles the questions the vacuum extension of the chapter left open. The surface crosses $r = r_{\mathrm{s}}$ at the finite comoving time given by η_{h} with $\cos\eta_{\mathrm{h}} = 2r_{\mathrm{s}}/R_0 - 1$; from that moment the exterior region contains trapped surfaces (Remark 45.26) and the vacuum part of the spacetime is the regions I and II of the Kruskal diagram (Remark 45.20) — the white-hole region III and the second universe IV are excised, their place taken by the matter-filled interior, which is why they are artefacts of vacuum eternity rather than predictions about collapse. A distant observer, meanwhile, receives the exponentially redshifted last light of Proposition 45.22: the "frozen star" outside, the finite-time singularity inside, with no contradiction between them [844, 1509].

A.3 The Cantor–Schröder–Bernstein Theorem

This appendix proves Theorem 3.72 of Chapter 3: two sets that inject into each other are equipotent. It is what makes the comparison of cardinals a genuine order relation rather than a pair of unrelated inequalities, and it is used immediately in Corollary A.3 to identify the cardinal of the continuum with that of the power set of $\mathbb{N}$. The proof is elementary but not obvious — neither injection need be surjective anywhere, and the bijection must be manufactured out of the two of them.

Theorem A.1 (Cantor–Schröder–Bernstein)**.** *Let X and Y be sets and let $f : X \longrightarrow Y$ and $g : Y \longrightarrow X$ be injective. Then there exists a bijection $h : X \longrightarrow Y$.*
Rests on Definition 3.45 and Proposition 3.48.

Proof. *Derives Theorem A.1.*
The obstruction to using g^{-1} as the bijection is the set of points of X outside the image of g, where g^{-1} is undefined. Track where those points propagate under repeated application of $g \circ f$, and use f there instead.

Define a sequence of subsets of X by

$$C_0 := X \setminus g(Y)\,, \qquad C_{n+1} := g(f(C_n)) \quad (n \geq 0)\,, \tag{A.104}$$

and let

$$C := \bigcup_{n \geq 0} C_n\,. \tag{A.105}$$

Now define $h : X \longrightarrow Y$ by

$$h(x) := \begin{cases} f(x) & \text{if } x \in C\,, \\ g^{-1}(x) & \text{if } x \notin C\,. \end{cases} \tag{A.106}$$

h is well defined. The only question is the second branch. If $x \notin C$ then in particular $x \notin C_0 = X \setminus g(Y)$, so $x \in g(Y)$; and since g is injective there is exactly one $y \in Y$ with $g(y) = x$. Write $g^{-1}(x)$ for it.

h is injective. Each branch is injective on its own domain: f by hypothesis, and g^{-1} because g is a map. It remains to check that the two branches never collide, i.e. that

$$f(C) \cap g^{-1}(X \setminus C) = \varnothing\,. \tag{A.107}$$

Suppose not, and let $f(c) = g^{-1}(x)$ with $c \in C$ and $x \notin C$. Applying g to both sides gives $x = g(f(c))$. Since $c \in C$, there is an n with $c \in C_n$, whence

$$x = g(f(c)) \in g(f(C_n)) = C_{n+1} \subset C\,,$$

contradicting $x \notin C$. This proves Equation (A.107), and with it the injectivity of h.

h is surjective. Let $y \in Y$ and consider the point $g(y) \in X$. There are two cases, and they are exhaustive by Equation (3.25).

If $g(y) \notin C$, then by the second branch of Equation (A.106),

$$h(g(y)) = g^{-1}(g(y)) = y\,,$$

so y is attained.

If $g(y) \in C$, then $g(y) \in C_n$ for some n. That n cannot be 0, because $g(y)$ lies in $g(Y)$ while $C_0 = X \setminus g(Y)$ does not meet $g(Y)$. So $n = m + 1$ for some $m \geq 0$ and $g(y) \in C_{m+1} = g(f(C_m))$, i.e. $g(y) = g(f(c))$ for some $c \in C_m$. Injectivity of g gives $y = f(c)$, and since $c \in C_m \subset C$ the first branch of Equation (A.106) yields $h(c) = f(c) = y$. Again y is attained.

Being injective and surjective, h is a bijection by Proposition 3.48. □

Remark A.2 (No choice is used). The construction Equations (A.104) to (A.106) is explicit: given f and g, the bijection h is determined, with no appeal to Axiom 3.78. This matters, because the companion statement that *any* two cardinals are comparable — that at least one of the two injections always exists — is equivalent to the axiom of choice. CSB says that comparability, where it holds, is antisymmetric; it does not say that any two sets are comparable.

Corollary A.3 (The continuum is the power set of the naturals)**.** $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$, *that is* $\mathfrak{c} = 2^{\aleph_0}$.

Rests on Theorem A.1 and Corollary 3.68.

Proof. *Derives Corollary A.3.*

We exhibit injections in both directions and invoke Theorem A.1.

An injection $\mathcal{P}(\mathbb{N}) \longrightarrow \mathbb{R}$. Send $A \subset \mathbb{N}$ to the real number whose base-3 expansion has digit 1 at place n when $n \in A$ and digit 0 otherwise,

$$\Phi(A) := \sum_{n \in A} 3^{-n} \,. \tag{A.108}$$

If $A \neq B$, let k be the least element of their symmetric difference, say $k \in A \setminus B$. Then

$$\Phi(A) - \Phi(B) \geq 3^{-k} - \sum_{n>k} 3^{-n} = 3^{-k} - \frac{3^{-k}}{2} = \frac{3^{-k}}{2} > 0 \,,$$

since the digits omitted from B beyond place k contribute at most the full geometric tail. Hence $\Phi(A) \neq \Phi(B)$ and Φ is injective. Using digits 0 and 1 in base 3 — rather than base 2 — is exactly what makes the tail strictly smaller than the leading term, so that the two expansions cannot coincide.

An injection $\mathbb{R} \longrightarrow \mathcal{P}(\mathbb{N})$. Send $x \in \mathbb{R}$ to its Dedekind cut $\Psi(x) := \{q \in \mathbb{Q} \mid q < x\}$, injective because $\mathbb{Q}$ is dense in $\mathbb{R}$: if $x < y$ there is a rational strictly between them, lying in $\Psi(y)$ but not in $\Psi(x)$. This lands in $\mathcal{P}(\mathbb{Q})$ rather than $\mathcal{P}(\mathbb{N})$, but $|\mathbb{Q}| = \aleph_0$ by Corollary 3.68, and a bijection $\mathbb{N} \longrightarrow \mathbb{Q}$ induces a bijection $\mathcal{P}(\mathbb{Q}) \longrightarrow \mathcal{P}(\mathbb{N})$ by taking preimages. Composing gives the required injection. □

Appendix A.3 discharges the proof obligation of Theorem 3.72, and Corollary A.3 supplies the identification $\mathfrak{c} = 2^{\aleph_0}$ used in Section 3.5.3 when the continuum hypothesis is stated. The next appendix treats the other half of the foundational material of that chapter.

A.4 The Completeness Theorem

This appendix proves Theorem 3.84 of Chapter 3: for first-order logic, semantic consequence and syntactic derivability coincide,

$$T \vDash \varphi \iff T \vdash \varphi \,. \tag{A.109}$$

The left-to-right direction is the substantive one and is equivalent to a statement about existence: *every consistent theory has a model.* That is what the Henkin construction builds, and it builds it out of nothing but the syntax — the elements of the model are the closed terms of the language itself. The theorem is Gödel's doctoral work [654]; the proof given here is Henkin's later and simpler one.

Throughout, the logic is first-order logic with equality, so the equality axioms — reflexivity $t = t$ and the congruence schemas asserting that equals may be substituted in any function or relation symbol — are part of the logical apparatus and available in every derivation. The language is assumed countable, which every language in this book is; the uncountable case needs only Zorn's lemma (Axiom 3.78) in place of the enumeration in Lemma A.6.

A.4.1 Soundness

Theorem A.4 (Soundness)**.** *If* $T \vdash \varphi$ *then* $T \vDash \varphi$.

Rests on Axiom 3.34, Definition 3.14, and Equation (3.10).

Proof. *Derives Theorem A.4.*

Induction on the length of the derivation, using Axiom 3.34 at the meta-level. Let M be any structure with $M \vDash T$, and let $\psi_1, \dots, \psi_k$ be a derivation of φ from T. We show every ψ_i is true in M.

If ψ_i is a member of T, it is true in M by assumption. If it is a logical axiom, it is true in every structure whatever: the propositional axioms are tautologies in the sense of Definition 3.14, and their truth tables were computed once and for all in Section 3.1.2; the quantifier axiom $\forall x\, \psi(x) \to \psi(t)$ holds because if ψ is satisfied by every element it is satisfied by the one denoted by t; and the equality axioms hold because equality is interpreted as genuine identity on the domain.

If ψ_i follows by modus ponens from ψ_j and $\psi_j \to \psi_i$ with $j < i$, then both are true in M by the induction hypothesis, and row two of Equation (3.10) — the only row in which a conditional is false — is thereby excluded, so ψ_i is true. If ψ_i is $\forall x\, \psi$ obtained by generalization from ψ, where x does not occur free in T, then ψ is true in M under every assignment to x, which is what $\forall x\, \psi$ asserts.

Every line is therefore true in M, in particular the last, so $M \vDash \varphi$. As M was an arbitrary model of T, $T \vDash \varphi$. □

Soundness is the direction that makes derivations worth performing; it is also the direction used, in Theorem A.30, to conclude that Q does not prove a sentence false in $\mathbb{N}$. The remaining direction occupies the rest of this section.

A.4.2 Maximal consistent sets

Definition A.5 (Maximal consistent)**.** A set S of sentences is *maximal consistent* if it is consistent in the sense of Definition 3.82 and no proper superset of it is consistent.

Rests on Definition 3.82.

Lemma A.6 (Lindenbaum)**.** *Every consistent set of sentences extends to a maximal consistent set.*

Rests on Definition A.5 and Proposition 3.67.

Proof. *Derives Lemma A.6.*

The language is countable, so its sentences can be listed $\varphi_0, \varphi_1, \varphi_2, \dots$ — there are countably many finite strings over a countable alphabet, by Proposition 3.67. Given a consistent S, define an increasing chain by $S_0 := S$ and

$$S_{n+1} := \begin{cases} S_n \cup \{\varphi_n\} & \text{if this is consistent}, \\ S_n \cup \{\neg\varphi_n\} & \text{otherwise}. \end{cases} \tag{A.110}$$

Each S_n is consistent, by induction on n. The base case is the hypothesis on S. For the step, suppose S_n is consistent but both $S_n \cup \{\varphi_n\}$ and $S_n \cup \{\neg\varphi_n\}$ are not. Inconsistency of the first gives $S_n \vdash \neg\varphi_n$ and of the second $S_n \vdash \varphi_n$ — in each case by discharging the assumption, which is the deduction theorem — so S_n proves both a sentence and its negation, contradicting its consistency. Hence at least one of the two branches of Equation (A.110) is consistent, and S_{n+1} is.

Let $S^* := \bigcup_{n \geq 0} S_n$. It is consistent: a derivation of a contradiction is a finite object and mentions only finitely many members of S^*, all of which lie in some single S_n — the chain is increasing — and that S_n was shown consistent. It is maximal: every sentence of the language is some φ_n, and Equation (A.110) placed either it or its negation in S_{n+1}, so no consistent sentence can be added. □

Lemma A.7 (Behaviour of a maximal consistent set)**.** *Let S^* be maximal consistent. Then for all sentences φ, ψ:*

1. *exactly one of* φ *and* $\neg\varphi$ *belongs to* S^*;

2. S^* *is deductively closed: if* $S^* \vdash \varphi$ *then* $\varphi \in S^*$;

3. $\varphi \wedge \psi \in S^*$ *if and only if both* $\varphi \in S^*$ *and* $\psi \in S^*$;

4. $\varphi \vee \psi \in S^*$ *if and only if* $\varphi \in S^*$ *or* $\psi \in S^*$;

5. $\varphi \to \psi \in S^*$ *if and only if* $\varphi \notin S^*$ *or* $\psi \in S^*$.

Rests on Definition A.5 and Proposition 3.22.

Proof. *Derives Lemma A.7.*

(1) Not both, by consistency. At least one: if neither is in S^*, then by maximality $S^* \cup \{\varphi\}$ is inconsistent, giving $S^* \vdash \neg\varphi$, and likewise $S^* \vdash \varphi$; that contradicts consistency.

(2) If $S^* \vdash \varphi$ then $S^* \cup \{\varphi\}$ is consistent — adding what is already derivable cannot produce a new contradiction — so maximality forces $\varphi \in S^*$.

(3)–(5) Each follows from (1) and (2) together with the propositional tautologies of Proposition 3.22. For (5): if $\varphi \notin S^*$ then $\neg\varphi \in S^*$ by (1), and $\neg\varphi \vdash \varphi \to \psi$, so $\varphi \to \psi \in S^*$ by (2); if $\psi \in S^*$ the same conclusion follows from $\psi \vdash \varphi \to \psi$. Conversely if $\varphi \to \psi \in S^*$ and $\varphi \in S^*$, modus ponens and (2) give $\psi \in S^*$. The other two are proved the same way. □

A.4.3 Henkin witnesses

A maximal consistent set decides every sentence, but it may assert $\exists x\, \varphi(x)$ without containing any instance $\varphi(t)$ — and then no model can be read off it, because there is no term available to serve as the witness. The remedy is to put the witnesses into the language by hand.

Lemma A.8 (Adding witnesses preserves consistency)**.** *Let* S *be a consistent set of* L*-sentences. Let* L' *extend* L *by a new constant* c_φ *for each* L*-formula* $\varphi(x)$ *with exactly one free variable, and let*

$$H_\varphi : \quad \exists x\, \varphi(x) \to \varphi(c_\varphi) \tag{A.111}$$

be the corresponding Henkin axiom. *Then* $S \cup \{H_\varphi \mid \varphi\}$ *is consistent as a set of* L'*-sentences.*
Rests on Definition 3.82.

Proof. *Derives Lemma A.8.*

A derivation is finite, so if the extended set were inconsistent, finitely many Henkin axioms would already suffice; it is enough to show that adding one at a time preserves consistency. So suppose S' is consistent, S' does not mention the constant $c := c_\varphi$, and $S' \cup \{H_\varphi\}$ is inconsistent. Then

$$S' \vdash \neg H_\varphi\,, \qquad \text{that is} \qquad S' \vdash \exists x\, \varphi(x) \wedge \neg\varphi(c)\,.$$

In particular $S' \vdash \neg\varphi(c)$.

Now the crucial step. The constant c occurs nowhere in S', so it plays no role in the derivation beyond being a name; replacing every occurrence of c throughout the derivation by a variable y that appears nowhere in it yields again a correct derivation, and therefore $S' \vdash \neg\varphi(y)$ with y free and unconstrained. Generalization is then legitimate — its side condition is exactly that the variable is not free in the premises — and gives

$$S' \vdash \forall y\, \neg\varphi(y)\,, \qquad \text{i.e.} \qquad S' \vdash \neg\exists x\, \varphi(x)\,.$$

But $S' \vdash \exists x\, \varphi(x)$ was established above, so S' is inconsistent, contrary to hypothesis. Hence $S' \cup \{H_\varphi\}$ is consistent. □

Adding constants creates new formulas, which need witnesses of their own, so one application does not suffice. Iterate: set $L_0 := L$ and $S_0 := S$, let L_{n+1} and S_{n+1} be the result of applying Lemma A.8 to L_n and S_n, and put $L_\infty := \bigcup_n L_n$ and $S_\infty := \bigcup_n S_n$. Every L_∞-formula lies in some L_n and so has its witness constant in L_{n+1}; and S_∞ is consistent because any derivation of a contradiction from it is finite and lives inside some consistent S_n. Finally apply Lemma A.6 to S_∞ inside L_∞, giving a set

$$T^* \supset S\,, \qquad \text{maximal consistent and containing every } H_\varphi\,. \tag{A.112}$$

Extending to a maximal set adds sentences but removes none, so the Henkin axioms survive.

A.4.4 The term model

Everything needed to build a structure is now inside T^*. The domain will be the closed terms of L_∞ — expressions such as 0, $f(c_3)$, $g(c_1, f(c_2))$ — with terms identified exactly when T^* says they are equal.

Definition A.9 (The term structure)**.** Let CT be the set of closed L_∞-terms, and define

$$t \sim s \quad :\Longleftrightarrow \quad (t = s) \in T^*\,. \tag{A.113}$$

The structure M has domain CT$/\!\sim$, the set of classes $[t]$, and interprets

$$c^M := [c] \quad \text{for each constant } c\,, \tag{A.114}$$

$$f^M([t_1], \dots, [t_n]) := [f(t_1, \dots, t_n)]\,, \tag{A.115}$$

$$([t_1], \dots, [t_n]) \in R^M \quad :\Longleftrightarrow \quad R(t_1, \dots, t_n) \in T^*\,. \tag{A.116}$$

Rests on Definition A.5 and Equation (A.112)*.*

Lemma A.10. *$\sim$ is an equivalence relation,* CT *is non-empty, and Equations* (A.115) *and* (A.116) *do not depend on the representatives chosen.*
Rests on Definitions 3.58 and A.9, Lemma A.7, and Theorem 3.60.

Proof. *Derives Lemma A.10.*
Reflexivity: $t = t$ is a logical axiom, so it lies in T^* by Lemma A.7(2). Symmetry and transitivity: the equality congruence schemas give $\vdash (t = s) \to (s = t)$ and $\vdash ((t = s) \wedge (s = u)) \to (t = u)$, and deductive closure transfers these to membership in T^*. So $\sim$ satisfies the three clauses of Definition 3.58 and, by Theorem 3.60, partitions CT into classes.
CT is non-empty because L_∞ contains the Henkin constants, of which there is at least one; a structure must have a non-empty domain, and this is where that is secured.
Well-definedness: suppose $t_i \sim s_i$ for each i, i.e. $(t_i = s_i) \in T^*$. The congruence schema for the function symbol f gives

$$\vdash \Big(\bigwedge_i t_i = s_i\Big) \to f(t_1, \dots, t_n) = f(s_1, \dots, s_n)\,,$$

so by Lemma A.7(2),(3) the conclusion lies in T^* and $[f(\boldsymbol{t})] = [f(\boldsymbol{s})]$. The congruence schema for the relation symbol R gives the same conclusion for Equation (A.116). □

A.4.5 The truth lemma

Lemma A.11 (Truth lemma)**.** *For every L_∞-sentence φ,*

$$M \vDash \varphi \iff \varphi \in T^*\,. \tag{A.117}$$

Rests on Definition A.9 and Lemmas A.7, A.8 and A.10.

Proof. *Derives Lemma A.11.*

Induction on the number of connectives and quantifiers in φ. The statement being proved is about all sentences of that complexity at once, which is what makes the quantifier case go through.

Atomic. A closed atomic sentence is $R(t_1, \dots, t_n)$ or $t = s$. For the first, Equation (A.116) says $M \vDash R(t_1, \dots, t_n)$ precisely when $R(t_1, \dots, t_n) \in T^*$. For the second, $M \vDash t = s$ means $[t] = [s]$, which by Equation (A.113) means $(t = s) \in T^*$. Note that the interpretation of a closed term t in M is $[t]$ itself, by induction on the term using Equations (A.114) and (A.115).

Negation. $M \vDash \neg\varphi$ iff $M \nvDash \varphi$, iff $\varphi \notin T^*$ by the induction hypothesis, iff $\neg\varphi \in T^*$ by Lemma A.7(1).

Conditional. $M \vDash \varphi \to \psi$ iff $M \nvDash \varphi$ or $M \vDash \psi$, which by the induction hypothesis is $\varphi \notin T^*$ or $\psi \in T^*$, which by Lemma A.7(5) is $\varphi \to \psi \in T^*$. The cases of $\wedge$ and $\vee$ are identical, using clauses (3) and (4).

Existential. Every element of the domain is $[t]$ for some closed term t, so

$$M \vDash \exists x\, \varphi(x) \quad \text{iff} \quad M \vDash \varphi(t) \text{ for some closed term } t\,.$$

Each $\varphi(t)$ has fewer quantifiers than $\exists x\, \varphi(x)$, so the induction hypothesis applies to it and the right-hand side is equivalent to $\varphi(t) \in T^*$ for some closed t.

If $\exists x\, \varphi(x) \in T^*$, the Henkin axiom Equation (A.111) for φ is in T^* by Equation (A.112), and modus ponens with Lemma A.7(2) puts $\varphi(c_\varphi) \in T^*$; that is the required witness. Conversely if $\varphi(t) \in T^*$ for some closed t, then since $\vdash \varphi(t) \to \exists x\, \varphi(x)$ is a logical axiom, deductive closure gives $\exists x\, \varphi(x) \in T^*$.

Universal. $\forall x\, \varphi$ is treated as $\neg\exists x\, \neg\varphi$, already covered. □

A.4.6 Completeness, and compactness

Theorem A.12 (Model existence)**.** *Every consistent set of sentences has a model.*
Rests on Lemma A.11, Definition A.9, and Equation (A.112)*.*

Proof. *Derives Theorem A.12.*

Given consistent S, build T^* as in Equation (A.112) and let M be the term structure of Definition A.9. By Lemma A.11, M satisfies exactly the sentences of T^*, and $S \subset T^*$, so $M \vDash S$. The model is a structure for L_∞; forgetting the interpretations of the added constants leaves a model of S in the original language. □

Theorem A.13 (Completeness)**.** *$T \vDash \varphi$ implies $T \vdash \varphi$. With Theorem A.4, this establishes Equation* (A.109)*.*
Rests on Theorems A.4 and A.12.

Proof. *Derives Theorem A.13.*

Contrapositive. Suppose $T \nvdash \varphi$. Then $T \cup \{\neg\varphi\}$ is consistent — an inconsistency would give $T \vdash \varphi$ by reductio — so by Theorem A.12 it has a model M. That M satisfies T and falsifies φ, so $T \nvDash \varphi$. □

Corollary A.14 (Compactness)**.** *If every finite subset of T has a model, then T has a model.*
Rests on Theorems A.4 and A.12.

Proof. *Derives Corollary A.14.*

If T had no model it would, by Theorem A.12, be inconsistent, so some derivation would produce a contradiction from it. A derivation is a finite object and cites only finitely many members of T; those form a finite inconsistent subset, which by Theorem A.4 has no model. This contradicts the hypothesis. □

Compactness is the corollary a physicist is most likely to meet, because it manufactures models nobody intended. Take the theory of the ordered field $\mathbb{R}$ (Chapter 7), adjoin a new constant ϵ and the infinitely many axioms $0 < \epsilon < 1/\overline{n}$, one for each natural number n. Any *finite* subset mentions only finitely many of them and is satisfied in $\mathbb{R}$ itself, by choosing ϵ small enough. Compactness then hands over a model of the entire set: an ordered field elementarily equivalent to $\mathbb{R}$ — satisfying every first-order sentence that $\mathbb{R}$ satisfies — but containing an element smaller than every positive rational. That is a rigorous infinitesimal, and the non-standard analysis built on it is a legitimate alternative foundation for the calculus. The same argument shows no first-order theory can pin down $\mathbb{N}$ up to isomorphism, which is the model-theoretic shadow of Theorem 3.87.

Appendix A.4 discharges the proof obligation of Theorem 3.84, and it is worth restating why this sits beside Theorem 3.87 without tension. Completeness is a statement about the *logic*: the calculus derives everything that follows from the premises, whatever the premises are. Incompleteness is a statement about a particular *theory*: the axioms of arithmetic do not, among themselves, settle every arithmetical sentence. A complete calculus applied to incomplete axioms yields exactly what Equation (3.88) describes, a sentence that neither follows nor fails to follow — and by Theorem A.12 there are then models of the axioms in which it holds and models in which it does not.

A.5 Arithmetization and the Incompleteness Theorems

This appendix supplies the technical machinery behind Sections 3.8 and 3.9 of Chapter 3: the coding of syntax into arithmetic, the diagonal lemma (Lemma 3.86), Rosser's strengthening of the first incompleteness theorem, the derivability conditions from which the second theorem follows, and Church's theorem on the undecidability of first-order validity (Theorem 3.98). Throughout, T is an effectively axiomatized first-order theory in the language of arithmetic $\{0, S, +, \cdot\}$ of Definition 3.83 that contains *Robinson arithmetic* Q: the successor, addition and multiplication axioms of that definition, *without* the induction schema, together with the one axiom that induction would otherwise supply,

$$\forall x \, (x \neq 0 \to \exists y \, (x = Sy)) \, . \tag{A.118}$$

PA proves Equation (A.118) by induction, so once the schema is dropped it must be assumed outright; without it the bounded-quantifier lemmas used below all fail. Q is therefore *finitely* axiomatized, which Theorem A.30 depends on, and it is already strong enough for everything except the derivability conditions of Appendix A.5.5, which need PA itself. The order is the defined relation

$$x \leq y \;:\equiv\; \exists z \, (z + x = y) \, , \qquad x < y \;:\equiv\; Sx \leq y \, . \tag{A.119}$$

Everything below applies verbatim to PA and, through the standard translation of arithmetic into set theory, to ZFC.

A.5.1 Gödel numbering

Fix an injective assignment of positive integers to the finitely many symbols of the language, and code a finite string $\sigma_1 \sigma_2 \cdots \sigma_k$ of symbols by

$$\ulcorner \sigma_1 \cdots \sigma_k \urcorner := p_1^{c(\sigma_1)} p_2^{c(\sigma_2)} \cdots p_k^{c(\sigma_k)} \, , \tag{A.120}$$

where p_i is the i-th prime and $c(\sigma)$ the code of the symbol σ. Unique factorization makes Equation (A.120) injective and makes decoding a matter of computing prime exponents. A finite *sequence* of strings — a derivation — is coded the same way, one prime per member. We write $\overline{n}$ for the *numeral* of n, the closed term $S^n(0)$ of the object language.

The choice of scheme is immaterial, as Remark 3.39 warned: any injective coding whose encoding and decoding operations are computable serves, and all the results below are invariant under changing it.

Proposition A.15 (Syntax is computable)**.** *Under Equation* (A.120) *the following relations on natural numbers are decidable in the sense of Definition 3.93: "n codes a formula", "n codes a sentence", "n codes an axiom of T", "m codes a T-derivation of the formula coded by n", and the function* $\mathrm{sub}(m, n)$ *returning the code of the result of substituting the numeral* $\overline{n}$ *for the free variable of the formula coded by* m*.*

Rests on Equation (A.120) *and Definitions 3.81 and 3.93.*

Derivation. Derives Proposition A.15.

Each is settled by a bounded search through the prime factorization of the argument, composed with the syntactic tests of Definition 3.80, which are finite string comparisons. Axiomhood is decidable precisely because T was assumed effectively axiomatized (Definition 3.81); this is the only hypothesis of the theorem that is not automatic, and it is where the assumption earns its place. Being a derivation is the conjunction of finitely many such tests, one per line of the coded sequence, each of which asks whether a line is an axiom or follows from earlier lines by one of the finitely many rules. □

A.5.2 Representability

Computability of a relation is a statement in the metatheory; what the incompleteness proofs need is that T can *talk about* the relation and prove the right instances.

Definition A.16 (Representability)**.** A relation $R \subset \mathbb{N}^k$ is *representable* in T if there is a formula $\rho(x_1, \dots, x_k)$ such that for all $n_1, \dots, n_k \in \mathbb{N}$,

$$R(n_1, \dots, n_k) \implies T \vdash \rho(\overline{n_1}, \dots, \overline{n_k})\,, \tag{A.121}$$

$$\neg R(n_1, \dots, n_k) \implies T \vdash \neg\rho(\overline{n_1}, \dots, \overline{n_k})\,. \tag{A.122}$$

A function is representable if its graph is, together with a provable uniqueness clause.

Rests on Definition 3.80.

Theorem A.17 (Representability)**.** *Every decidable relation and every computable total function is representable in Q, and hence in every theory containing Q. Moreover Q is Σ_1-complete: every true sentence of the form $\exists x_1 \cdots \exists x_k\, \delta$, with δ containing only bounded quantifiers, is provable in Q.*

Rests on Definitions 3.93 and A.16, Theorem A.25, and Lemma A.20.

The proof occupies the rest of this subsection. It is the one genuinely long construction in the chapter, and it is worth saying at the outset why it cannot be shortened: Q has no induction, so nothing may be proved inside Q by induction on a variable. Every induction below is performed in the metatheory, on a concrete numeral, and what it produces is not one theorem of Q but a recipe yielding a separate Q-derivation for each numeral. That is exactly the strength the incompleteness proofs need, and no more.

Bounded and existential formulas

Definition A.18 (Δ_0 and Σ_1)**.** A *bounded quantifier* is one of the forms $\forall x\,(x \le t \to \cdots)$ or $\exists x\,(x \le t \wedge \cdots)$, abbreviated $\forall x \le t$ and $\exists x \le t$, where t is a term not containing x. A formula is Δ_0 if all its quantifiers are bounded, and Σ_1 if it has the form $\exists x_1 \cdots \exists x_k\, \delta$ with δ of class Δ_0.

Rests on Equation (A.119).

The point of the restriction is that a Δ_0 sentence can be checked by a finite computation: every quantifier ranges over an explicitly bounded set of numbers.

Lemma A.19 (Numerals behave)**.** *For all $m, n \in \mathbb{N}$:*

1. *if $m + n = p$ then $Q \vdash \overline{m} + \overline{n} = \overline{p}$, and similarly $Q \vdash \overline{m} \cdot \overline{n} = \overline{q}$ when $mn = q$;*
2. *if $m \neq n$ then $Q \vdash \overline{m} \neq \overline{n}$;*
3. *if $m \leq n$ then $Q \vdash \overline{m} \leq \overline{n}$, and if $m > n$ then $Q \vdash \neg(\overline{m} \leq \overline{n})$;*
4. *Q proves the* bounded expansion

$$\forall z \left(z \leq \overline{m} \to z = \overline{0} \vee z = \overline{1} \vee \cdots \vee z = \overline{m}\right). \tag{A.123}$$

Rests on Definition 3.83 and Equations (A.118) and (A.119).

Proof. *Derives Lemma A.19.*

(1) Induction in the metatheory on n. For $n = 0$ the axiom $x + 0 = x$ gives $\overline{m} + \overline{0} = \overline{m}$. If $Q \vdash \overline{m} + \overline{n} = \overline{m+n}$, then the axiom $x + Sy = S(x + y)$ gives $\overline{m} + \overline{n+1} = S(\overline{m} + \overline{n}) = S\overline{m+n} = \overline{m+n+1}$. Multiplication is the same argument using $x \cdot 0 = 0$ and $x \cdot Sy = x \cdot y + x$. Note what happened: for each fixed pair (m, n) this is a finite chain of equalities, hence a genuine Q-derivation; the induction that generated the chain was ours, not Q's.

(2) Suppose $m \neq n$, say $m < n$. Metatheoretic induction on m. If $m = 0$ then $\overline{n}$ is $S\overline{n-1}$ and the axiom $Sx \neq 0$ applies directly. If $m > 0$, both numerals are successors, and the axiom $Sx = Sy \to x = y$ contraposed reduces the claim for (m, n) to the claim for $(m - 1, n - 1)$, which the induction hypothesis supplies.

(3) If $m \leq n$, take $z := \overline{n-m}$; by (1), $Q \vdash \overline{n-m} + \overline{m} = \overline{n}$, so $Q \vdash \exists z\,(z + \overline{m} = \overline{n})$, which is $\overline{m} \leq \overline{n}$ by Equation (A.119). If $m > n$, we must refute $\exists z\,(z + \overline{m} = \overline{n})$. By (4), applied with $\overline{n}$, any such z would satisfy $z = \overline{j}$ for some $j \leq n$; but then $z + \overline{m} = \overline{j+m}$ by (1), and $j + m \geq m > n$, so $\overline{j+m} \neq \overline{n}$ by (2). As j ranges over finitely many values, this is a finite case analysis inside Q.

(4) Metatheoretic induction on m. For $m = 0$: suppose $z \leq \overline{0}$, i.e. $w + z = \overline{0}$ for some w. If $z \neq \overline{0}$ then Equation (A.118) gives $z = Sv$ for some v, whence $w + Sv = S(w + v)$ by the addition axiom, so $\overline{0}$ is a successor, contradicting $Sx \neq 0$. Hence $z = \overline{0}$. This is the step that fails without Equation (A.118): nothing else in Q rules out an element that is neither 0 nor a successor.

For the step, assume Equation (A.123) for m and suppose $z \leq \overline{m+1}$, say $w + z = \overline{m+1} = S\overline{m}$. If $z = \overline{0}$ we are done. Otherwise $z = Sv$ by Equation (A.118), so $S(w + v) = S\overline{m}$, and injectivity of S gives $w + v = \overline{m}$, i.e. $v \leq \overline{m}$. The induction hypothesis then yields $v = \overline{j}$ for some $j \leq m$, so $z = Sv = \overline{j+1}$ with $j + 1 \leq m + 1$. □

Lemma A.20 (Σ_1-completeness)**.** *Every true Σ_1 sentence is provable in Q.*
Rests on Definition A.18, Lemma A.19, and Proposition 3.22.

Proof. *Derives Lemma A.20.*

First the Δ_0 case, by induction in the metatheory on the structure of the formula. Atomic sentences are equations and inequations between closed terms; evaluating the terms and applying Lemma A.19(1)–(3) settles them, and a true one is proved while a false one has its negation proved. The connectives are immediate: if the parts are decided, Lemma A.7-style propositional reasoning decides the compound, using only the tautologies of Proposition 3.22.

For a bounded quantifier, let $\forall z \leq \overline{m}\,\psi(z)$ be true. Then $\psi(\overline{j})$ is true for every $j \leq m$, and each is proved by the induction hypothesis; Equation (A.123) converts the finitely many instances into the universally quantified statement, since any z below $\overline{m}$ is provably one of the $\overline{j}$. If instead $\exists z \leq \overline{m}\,\psi(z)$ is true, some witness $\overline{j}$ with $j \leq m$ has $\psi(\overline{j})$ true and provable, and Lemma A.19(3) supplies $\overline{j} \leq \overline{m}$.

Now the general case. Let $\exists x_1 \cdots \exists x_k\, \delta$ be true, and let $n_1, \dots, n_k$ be witnesses. Then $\delta(\overline{n_1}, \dots, \overline{n_k})$ is a true Δ_0 sentence, hence provable, and existential generalization gives the Σ_1 sentence. $\square$

Coding sequences: the β-function

Representing a recursively defined function requires quantifying over the sequence of intermediate values, and arithmetic has no sequences. Gödel's device turns a sequence of arbitrary length into a pair of numbers.

Lemma A.21 (Chinese remainder theorem)**.** *Let $m_0, \dots, m_n$ be pairwise coprime positive integers and let $r_0, \dots, r_n$ satisfy $0 \le r_i < m_i$. Then some a has $a \equiv r_i \pmod{m_i}$ for every i.*

Proof. *Derives Lemma A.21.*

Put $P := m_0 m_1 \cdots m_n$ and $P_i := P/m_i$. Since the moduli are pairwise coprime, P_i is coprime to m_i, so there is u_i with $P_i u_i \equiv 1 \pmod{m_i}$. Set $a := \sum_i r_i P_i u_i$. Modulo m_i every term with $j \neq i$ vanishes, because m_i divides P_j, and the remaining term is $r_i P_i u_i \equiv r_i$. $\square$

Definition A.22 (Gödel's β)**.**

$$\beta(a, b, i) := \operatorname{rem}\big(a,\ 1 + (i+1)b\big)\,, \tag{A.124}$$

the remainder of a on division by $1 + (i+1)b$.

Lemma A.23 (Sequence lemma)**.** *For every finite sequence $s_0, s_1, \dots, s_n$ of natural numbers there exist a, b with $\beta(a, b, i) = s_i$ for all $i \le n$.*

Rests on Definition A.22 and Lemma A.21.

Proof. *Derives Lemma A.23.*

Let $c := \max\{n, s_0, \dots, s_n\}$ and put $b := c!$. Write $m_i := 1 + (i+1)b$ for $i = 0, \dots, n$.

The moduli are pairwise coprime. Suppose a prime p divides both m_i and m_j with $i < j \le n$. Then p divides their difference,

$$m_j - m_i = (j - i)\, b\,,$$

so p divides $(j - i)$ or p divides b. If p divided b then, since p also divides $m_i = 1 + (i+1)b$, it would divide the difference $m_i - (i+1)b = 1$, which is impossible. Hence p divides $j - i$. But $0 < j - i \le n \le c$, so $j - i$ is one of the factors of $c! = b$, giving $p \mid b$ — the case just excluded. No such prime exists, so $\gcd(m_i, m_j) = 1$.

The residues fit. Each $s_i \le c \le c! = b < 1 + (i+1)b = m_i$, so s_i is a legitimate remainder modulo m_i.

Lemma A.21 now supplies a with $a \equiv s_i \pmod{m_i}$ for every $i \le n$, and since $0 \le s_i < m_i$ that congruence says exactly $\operatorname{rem}(a, m_i) = s_i$, i.e. $\beta(a, b, i) = s_i$. $\square$

Lemma A.24. *The relation $\beta(a, b, i) = v$ is defined by the Δ_0 formula*

$$B(a, b, i, v): \quad \exists q \le a\ \big(a = q \cdot (1 + (i+1) \cdot b) + v\ \wedge\ v < 1 + (i+1) \cdot b\big)\,, \tag{A.125}$$

and Q proves that the value is unique: for each numeral triple, at most one v satisfies it.

Rests on Definitions A.18 and A.22 and Lemma A.19.

Proof. *Derives Lemma A.24.*

Equation (A.125) is the division algorithm written out: v is the remainder exactly when $a = q \cdot m + v$ with $v < m$. The quotient q is at most a, so bounding its quantifier by a loses nothing, and every quantifier in Equation (A.125) is therefore bounded. For uniqueness at numeral arguments: if v, v' both satisfied it with $v < v'$, subtracting gives $(q - q')m = v' - v$ with $0 < v' - v < m$, impossible for a multiple of m; and by Lemma A.19 this finite arithmetic is carried out inside Q for each numeral instance. $\square$

Representing the computable functions

Theorem A.25. *Every primitive recursive function is representable in Q.*
Rests on Definition A.16 and Lemmas A.19, A.20, A.23 and A.24.

Proof. *Derives Theorem A.25.*

Induction in the metatheory on the construction of the function.

Initial functions. The zero function is represented by $y = \overline{0}$; the successor by $y = Sx$; the projection π_i^n by $y = x_i$. In each case Equation (A.121) holds because the defining equation is provable for numerals by Lemma A.19(1), and Equation (A.122) because a wrong value is refuted by Lemma A.19(2).

Composition. Let $f(\boldsymbol{x}) = h(g_1(\boldsymbol{x}), \dots, g_k(\boldsymbol{x}))$ with $h, g_1, \dots, g_k$ represented by $\eta, \gamma_1, \dots, \gamma_k$. Take

$$\varphi(\boldsymbol{x}, y) : \quad \exists z_1 \cdots \exists z_k \, (\gamma_1(\boldsymbol{x}, z_1) \wedge \cdots \wedge \gamma_k(\boldsymbol{x}, z_k) \wedge \eta(z_1, \dots, z_k, y)) \, . \tag{A.126}$$

At numeral arguments the z_i are provably the values of the g_i, by the uniqueness clause of Definition A.16, and η then pins y; the negative clause follows likewise.

Primitive recursion. Let

$$f(\boldsymbol{x}, 0) = g(\boldsymbol{x}) \, , \qquad f(\boldsymbol{x}, n+1) = h(\boldsymbol{x}, n, f(\boldsymbol{x}, n)) \, ,$$

with g, h represented by γ, η. The obstruction is that f's value at n depends on the whole computation $f(\boldsymbol{x}, 0), \dots, f(\boldsymbol{x}, n)$, which is a sequence. Encode it with β: say that y is the value if there is a pair (a, b) coding a sequence whose 0th entry is $g(\boldsymbol{x})$, whose entries step according to h, and whose nth entry is y. Formally,

$$\begin{aligned} \varphi(\boldsymbol{x}, n, y) : \quad \exists a \, \exists b \Big[\, & \exists u \, (B(a, b, \overline{0}, u) \wedge \gamma(\boldsymbol{x}, u)) \;\wedge\; B(a, b, n, y) \\ & \wedge \; \forall i < n \; \forall u \, \forall v \, (B(a, b, i, u) \wedge B(a, b, Si, v) \\ & \quad \to \eta(\boldsymbol{x}, i, u, v)) \, \Big] \, . \end{aligned} \tag{A.127}$$

For numeral arguments $\boldsymbol{m}, n$ with true value p, Lemma A.23 supplies a, b coding the actual computation sequence, every conjunct of Equation (A.127) is then a true Δ_0 or Σ_1 statement about numerals, and Lemma A.20 makes it provable — giving Equation (A.121). For Equation (A.122), if $y = \overline{p'}$ with $p' \neq p$, then $B(a, b, \overline{n}, y)$ together with the uniqueness of Lemma A.24 and the induction hypothesis on η forces $\overline{p'} = \overline{p}$, refuted by Lemma A.19(2). □

Proof of Theorem A.17. *Derives Theorem A.17.*

A decidable relation has a computable indicator function (Definition 3.93), so it suffices to treat total computable functions.

By Kleene's normal form, every total computable f can be written $f(\boldsymbol{x}) = V(\mu z \, [\, C(\boldsymbol{x}, z) = 0 \,])$ with V and C primitive recursive, where μz denotes the least z making the bracket true — C tests whether z codes a halting computation of the machine for f on $\boldsymbol{x}$, which by Proposition A.15 is a primitive recursive test, and V reads the output off it. Since f is total, such a z exists for every $\boldsymbol{x}$. Let χ and ν represent C and V, which Theorem A.25 provides, and set

$$\varphi(\boldsymbol{x}, y) : \quad \exists z \, \Big(\chi(\boldsymbol{x}, z, \overline{0}) \;\wedge\; \forall w < z \; \neg \chi(\boldsymbol{x}, w, \overline{0}) \;\wedge\; \nu(z, y) \Big) \, . \tag{A.128}$$

At numeral arguments, let z_0 be the true least witness. Then $\chi(\overline{\boldsymbol{m}}, \overline{z_0}, \overline{0})$ is provable, each of the finitely many $\neg\chi(\overline{\boldsymbol{m}}, \overline{w}, \overline{0})$ with $w < z_0$ is provable, and Equation (A.123) assembles them into the bounded universal clause. So Equation (A.121) holds. If y is given a wrong numeral value, the uniqueness clauses for χ and ν refute it, giving Equation (A.122).

The Σ_1-completeness half of the theorem is Lemma A.20. Every step used only the seven axioms of Q; no induction inside the theory was required, because each claim was proved separately for each numeral by an induction carried out here rather than there.

Axiom 3.94 entered nowhere as a step. It is needed only to identify the informal notion of "computable procedure" with the formal one in the statement of the theorem; the proof itself concerns Turing machines and primitive recursion throughout. □

This discharges the last outstanding obligation of the chapter. It is what licenses the substitution formula of Theorem A.26, the provability predicate Equation (A.129), and the Σ_1 step in both Theorem 3.87 and Theorem A.30.

By Proposition A.15 and Theorem A.17 there is a formula $\mathrm{Proof}_T(y, x)$ representing "y codes a T-derivation of the formula coded by x", and we set

$$\mathrm{Prov}_T(x) := \exists y\, \mathrm{Proof}_T(y, x)\,, \tag{A.129}$$

which is the formula promised in Equation (3.86). Note that Prov_T is Σ_1 but not decidable: the existential quantifier is unbounded, and searching for a proof is the semi-decidable procedure of Section 3.9.3.

A.5.3 The diagonal lemma

Theorem A.26 (Diagonal lemma)**.** *Let $\psi(x)$ be any formula with one free variable. Then there is a sentence σ with*

$$T \vdash\ \sigma \leftrightarrow \psi(\ulcorner \sigma \urcorner)\,. \tag{A.130}$$

Rests on Proposition A.15 and Theorem A.17.

Proof. *Derives Theorem A.26.*

By Proposition A.15 the substitution function sub is computable, so by Theorem A.17 it is representable: there is a formula $\mathrm{Sub}(x, y, z)$ such that for all m, n,

$$T \vdash\ \forall z \left(\mathrm{Sub}(\overline{m}, \overline{n}, z) \leftrightarrow z = \overline{\mathrm{sub}(m, n)}\right). \tag{A.131}$$

Given ψ, define the auxiliary formula in the single free variable x

$$\theta(x) := \forall z \left(\mathrm{Sub}(x, x, z) \rightarrow \psi(z)\right), \tag{A.132}$$

let $k := \ulcorner \theta \urcorner$ be its own Gödel number, and put

$$\sigma := \theta(\overline{k}) = \forall z \left(\mathrm{Sub}(\overline{k}, \overline{k}, z) \rightarrow \psi(z)\right). \tag{A.133}$$

The point of the construction is the arithmetical identity

$$\mathrm{sub}(k, k) = \ulcorner \theta(\overline{k}) \urcorner = \ulcorner \sigma \urcorner\,, \tag{A.134}$$

which holds because substituting the numeral $\overline{k}$ into the formula coded by k — namely θ — produces exactly σ.

Now instantiate Equation (A.131) at $m = n = k$ and use Equation (A.134):

$$T \vdash\ \forall z \left(\mathrm{Sub}(\overline{k}, \overline{k}, z) \leftrightarrow z = \ulcorner \sigma \urcorner\right).$$

Substituting this equivalence into Equation (A.133) gives

$$T \vdash\ \sigma \leftrightarrow \forall z \left(z = \ulcorner \sigma \urcorner \rightarrow \psi(z)\right),$$

and the right-hand side is provably equivalent to $\psi(\ulcorner \sigma \urcorner)$, since a universally quantified variable constrained to a single value may be replaced by that value. This is Equation (A.130). □

No circularity is involved. The sentence σ was written down explicitly in Equation (A.133); it mentions only the numeral $\overline{k}$, a number computed before σ existed. Self-reference is achieved without self-mention, and this is the whole trick.

A.5.4 Rosser's form of the first theorem

Theorem 3.87 was proved in Chapter 3 using soundness to rule out $T \vdash \neg G_T$. Gödel's own proof replaced soundness by ω-consistency — no formula φ has $T \vdash \exists x\, \varphi(x)$ together with $T \vdash \neg\varphi(\overline{n})$ for every n — and Rosser removed even that [786].

Theorem A.27 (Rosser)**.** *Let $T \supset Q$ be effectively axiomatized and merely* consistent*. Then there is a sentence R with $T \nvdash R$ and $T \nvdash \neg R$.*

Rests on Theorems A.17 and A.26 and Lemma A.19.

Proof. *Derives Theorem A.27.*

Let $\mathrm{neg}(x)$ be the computable function taking the code of a formula to the code of its negation, represented as in Theorem A.17. Apply Theorem A.26 to

$$\psi(x) := \forall y \left(\mathrm{Proof}_T(y, x) \to \exists z \leq y\ \mathrm{Proof}_T(z, \mathrm{neg}(x)) \right), \tag{A.135}$$

obtaining R with $T \vdash R \leftrightarrow \psi(\ulcorner R \urcorner)$. In words, R says: *for every proof of me there is a no-longer proof of my negation.*

$T \nvdash R$. Suppose $T \vdash R$, with a derivation coded by m. Then $\mathrm{Proof}_T(\overline{m}, \ulcorner R \urcorner)$ is true, hence provable by Equation (A.121). Combining with R itself, T proves $\exists z \leq \overline{m}\ \mathrm{Proof}_T(z, \ulcorner \neg R \urcorner)$. But T is consistent, so $T \nvdash \neg R$, so for each $j \leq m$ the statement $\neg\mathrm{Proof}_T(\overline{j}, \ulcorner \neg R \urcorner)$ is true and hence provable by Equation (A.122). Since Q proves the bounded expansion $\forall z\, (z \leq \overline{m} \to z = \overline{0} \vee \cdots \vee z = \overline{m})$, the theory proves the negation of the displayed existential. So T is inconsistent — contradiction.

$T \nvdash \neg R$. Suppose $T \vdash \neg R$, with derivation coded by m; then $T \vdash \mathrm{Proof}_T(\overline{m}, \ulcorner \neg R \urcorner)$, and therefore

$$T \vdash\ \forall y\, (\overline{m} \leq y \to \exists z \leq y\ \mathrm{Proof}_T(z, \ulcorner \neg R \urcorner)), \tag{A.136}$$

taking $z = \overline{m}$ as the witness. On the other hand $\neg R$ is provably equivalent to $\exists y\, (\mathrm{Proof}_T(y, \ulcorner R \urcorner) \wedge \neg\exists z \leq y\, \mathrm{Proof}_T(z, \ulcorner \neg R \urcorner))$, so with Equation (A.136) and the provable trichotomy of $\leq$, the theory proves $\exists y < \overline{m}\ \mathrm{Proof}_T(y, \ulcorner R \urcorner)$. But consistency gives $T \nvdash R$, so each $\neg\mathrm{Proof}_T(\overline{j}, \ulcorner R \urcorner)$ with $j < m$ is true and provable, and the bounded expansion again refutes the existential. So T is inconsistent — contradiction. □

Bare consistency therefore suffices, and R is undecided by T exactly as Equation (3.88) asserts.

A.5.5 The derivability conditions and the second theorem

Proposition A.28. *For $T \supset PA$ with the coding of Appendix A.5.1, the three conditions Equations (3.93) to (3.95) hold.*

Rests on Theorem A.17 and Equation (A.129).

Derivation. *Derives Proposition A.28.*

Equation (3.93) is Σ_1-completeness (Theorem A.17): a derivation of φ is a concrete finite object, so $\mathrm{Prov}_T(\ulcorner \varphi \urcorner)$ is a true Σ_1 sentence and is therefore provable. Equation (3.94) formalizes the observation that two derivations, one of $\varphi \to \chi$ and one of φ, can be concatenated and extended by one application of modus ponens to give a derivation of χ; the concatenation is a computable operation on codes and PA proves that it does what it does, by induction on the length of the derivations. Equation (3.95) is Equation (3.93) formalized inside PA, and is the laborious one: it requires proving in PA the general Σ_1-completeness statement rather than using it instance by instance, which is done by induction on the structure of Σ_1 formulas. □

Theorem A.29 (Second incompleteness theorem)**.** *If $T \supset PA$ is effectively axiomatized and consistent, then $T \nvdash \mathrm{Con}_T$.*

Rests on Proposition A.28, Theorem 3.87, and Equation (3.89).

Proof. *Derives Theorem A.29.*

Let G be the Gödel sentence of Equation (3.89),

$$T \vdash\ G \leftrightarrow \neg \mathrm{Prov}_T(\ulcorner G \urcorner)\,. \tag{A.137}$$

We derive Equation (3.92) from the conditions.

From Equation (A.137), $T \vdash G \to \neg\mathrm{Prov}_T(\ulcorner G \urcorner)$, equivalently $T \vdash \mathrm{Prov}_T(\ulcorner G \urcorner) \to \neg G$. Applying Equation (3.93) to this theorem and then Equation (3.94),

$$T \vdash\ \mathrm{Prov}_T\big(\ulcorner \mathrm{Prov}_T(\ulcorner G \urcorner) \urcorner\big) \to \mathrm{Prov}_T(\ulcorner \neg G \urcorner)\,. \tag{A.138}$$

By Equation (3.95),

$$T \vdash\ \mathrm{Prov}_T(\ulcorner G \urcorner) \to \mathrm{Prov}_T\big(\ulcorner \mathrm{Prov}_T(\ulcorner G \urcorner) \urcorner\big)\,, \tag{A.139}$$

and chaining Equation (A.139) with Equation (A.138),

$$T \vdash\ \mathrm{Prov}_T(\ulcorner G \urcorner) \to \mathrm{Prov}_T(\ulcorner \neg G \urcorner)\,. \tag{A.140}$$

A theory proving both a sentence and its negation proves everything, and this fact is itself formalizable, giving $T \vdash \big(\mathrm{Prov}_T(\ulcorner G \urcorner) \wedge \mathrm{Prov}_T(\ulcorner \neg G \urcorner)\big) \to \mathrm{Prov}_T(\ulcorner 0 = 1 \urcorner)$. With Equation (A.140),

$$T \vdash\ \mathrm{Prov}_T(\ulcorner G \urcorner) \to \mathrm{Prov}_T(\ulcorner 0 = 1 \urcorner)\,, \tag{A.141}$$

whose contrapositive, recalling $\mathrm{Con}_T = \neg\mathrm{Prov}_T(\ulcorner 0 = 1 \urcorner)$ from Equation (3.90), reads

$$T \vdash\ \mathrm{Con}_T \to \neg \mathrm{Prov}_T(\ulcorner G \urcorner)\,. \tag{A.142}$$

Combining Equation (A.142) with Equation (A.137) gives $T \vdash \mathrm{Con}_T \to G$, which is Equation (3.92).

Finally, suppose $T \vdash \mathrm{Con}_T$. Modus ponens would yield $T \vdash G$, contradicting the first half of Theorem 3.87, which requires only consistency. Hence $T \nvdash \mathrm{Con}_T$. □

A.5.6 Church's theorem

Theorem A.30 (Church, Turing)**.** *The set of valid sentences of first-order logic is undecidable [772, 803].*

Rests on Theorems 3.84, 3.96 and A.17.

Proof. *Derives Theorem A.30.*

The proof has two moves: reduce validity to provability in a finitely axiomatized theory, then reduce halting to that.

Validity to provability in Q. Robinson arithmetic Q has finitely many axioms; let χ be their conjunction, a single sentence. By the deduction theorem, for every sentence φ,

$$Q \vdash \varphi \iff \vdash (\chi \to \varphi) \iff \vDash (\chi \to \varphi)\,, \tag{A.143}$$

the last step by Theorem 3.84. A decision procedure for validity would therefore decide provability in Q. This is where finite axiomatizability is essential; the argument does not run for PA, whose induction schema is infinite.

Halting to provability in Q. The halting set

$$H := \{(\ulcorner M \urcorner, w) \mid M \text{ halts on } w\}$$

is semi-decidable: simulate M on w with Theorem 3.95 and halt when it halts. Every semi-decidable set is definable by a Σ_1 formula — the existential quantifier ranging over codes of terminating computations, whose verification is decidable by Proposition A.15 — so fix $\eta(x, y)$ of that form with

$$(m, w) \in H \iff \eta(\overline{m}, \overline{w}) \text{ is true in } \mathbb{N}\,.$$

Then:

- If $(m, w) \in H$, the sentence $\eta(\overline{m}, \overline{w})$ is a true Σ_1 sentence, so $Q \vdash \eta(\overline{m}, \overline{w})$ by Σ_1-completeness (Theorem A.17).
- If $(m, w) \notin H$, the sentence is false in $\mathbb{N}$. Since $\mathbb{N}$ is a model of Q, soundness (Theorem 3.84) gives $Q \nvdash \eta(\overline{m}, \overline{w})$.

Hence $Q \vdash \eta(\overline{m}, \overline{w})$ if and only if M halts on w. A decision procedure for Q-provability would decide the halting problem, contradicting Theorem 3.96; by Equation (A.143) a decision procedure for validity would supply one. Therefore no decision procedure for validity exists. □

Appendix A.5 discharges the proof obligations of Lemma 3.86, Theorem 3.88 and Theorem 3.98, and supplies the sharpened consistency-only form of Theorem 3.87. The one technical input those arguments rest on, the representability theorem Theorem A.17, is standard and long — and is proved here in full rather than cited, in obedience to the same editorial rule that governs every physical derivation in this book, so no link in the chain stands on an unproved result.

A.6 The Universal Turing Machine

This appendix proves Theorem 3.95 of Chapter 3: there is a single Turing machine U which, given the code of any machine M together with an input w, reproduces exactly what M does on w. The theorem is the reason a computer is a general-purpose device rather than a machine built for one task, and it is the technical core of [803]. It is also what makes Theorem 3.96 possible: a machine that can run any machine can be asked to run itself.

The machine of Definition 3.92 has one tape. Building U directly in that format is possible but obscures the argument in bookkeeping, so we proceed in two steps: first show that extra tapes buy no extra power (Lemma A.32), then construct U with three of them.

A.6.1 Coding a machine as a string

Fix a machine $M = (Q, \Gamma, \sqcup, \Sigma, \delta, q_0)$. Number its states $q_0, q_1, \dots, q_r$ and its tape symbols $a_0 = \sqcup, a_1, \dots, a_s$, and write $\mathrm{bin}(i)$ for the binary numeral of i. Each transition $\delta(q_i, a_j) = (q_k, a_l, D)$ is coded by the block

$$\mathrm{bin}(i)\ \texttt{,}\ \mathrm{bin}(j)\ \texttt{,}\ \mathrm{bin}(k)\ \texttt{,}\ \mathrm{bin}(l)\ \texttt{,}\ d\,, \qquad d \in \{\texttt{L}, \texttt{R}\}\,, \tag{A.144}$$

and the code of the whole machine is the concatenation of the blocks for all pairs on which δ is defined, separated by the marker `;` and enclosed by `#`:

$$\ulcorner M \urcorner := \texttt{\#}\, B_1\ \texttt{;}\ B_2\ \texttt{;}\ \cdots\ \texttt{;}\ B_m\, \texttt{\#}\,. \tag{A.145}$$

The alphabet of the code is the fixed finite set $\{\texttt{0}, \texttt{1}, \texttt{,}, \texttt{;}, \texttt{\#}, \texttt{L}, \texttt{R}\}$, the same for every M however large Q and Γ are — which is what allows one machine U, with one fixed alphabet, to read them all.

Two properties are immediate from Equation (A.145) and are used below. First, whether a string is a well-formed code is decidable in the sense of Definition 3.93: a machine scans left to right and checks that the punctuation alternates as the format requires and that every block has five fields. Second, a machine can *look up* a transition: given $\mathrm{bin}(i)$ and $\mathrm{bin}(j)$ written somewhere on its tape, it walks the blocks in order, comparing the first two fields of each with the target pair symbol by symbol, and either finds the matching block or reaches the closing `#`. Both are bounded searches through a finite string.

A.6.2 Extra tapes buy no extra power

Definition A.31 (k-tape machine)**.** A k-tape Turing machine has k two-way unbounded tapes with independent heads, and a transition function

$$\delta : Q \times \Gamma^k \longrightarrow Q \times \Gamma^k \times \{L, R, S\}^k \ , \tag{A.146}$$

reading the k scanned symbols at once and writing, and moving each head left, right or not at all. Input is on tape 1, output read from tape 1; undefined δ means halt.

Rests on Definition 3.92.

Lemma A.32 (Tape reduction)**.** *For every k-tape machine N there is a single-tape machine N', in the sense of Definition 3.92, computing the same partial function. If N halts on w in n steps, N' halts on w in $O(n^2)$ steps.*

Rests on Definitions 3.92 and A.31.

Proof. *Derives Lemma A.32.*

The construction is to lay the k tapes out as $2k$ parallel tracks of one tape. A tape whose cells each carry a $2k$-tuple of symbols is nothing more than a tape over the alphabet

$$\Gamma' := (\Gamma \cup \{\sqcup\})^k \times \{0, 1\}^k \ , \tag{A.147}$$

which is a finite set — it has $(|\Gamma|)^k 2^k$ elements — so N' is a legitimate single-tape machine over a legitimate finite alphabet. Of each pair of tracks, the first holds the contents of simulated tape i and the second holds a single 1 marking where head i stands, all other cells of that track carrying 0.

N' keeps the simulated state of N in its own finite control, which is possible because Q is finite. One simulated step is performed by two sweeps.

Gathering sweep. N' walks right from the left end of the used region to its right end, remembering in its control which of the k head markers it has passed and which symbol stood beneath each. There are k markers and each carries one of $|\Gamma|$ symbols, so the information to be remembered is one of $(|\Gamma|)^k$ possibilities — a finite number, hence storable in the control. At the end of the sweep N' knows the full k-tuple of scanned symbols, which is exactly the argument δ of Equation (A.146) requires.

Writing sweep. N' consults δ — a fixed finite table built into its control — obtaining the new state, the k symbols to write and the k head movements. It then walks back left, and at each marker it passes it writes the new symbol on that track and moves the marker one cell left, right, or not at all as prescribed. It records the new state in its control. If δ was undefined on the tuple, N' halts instead.

The used region grows by at most one cell per simulated step in each direction, so after n simulated steps it has length $O(n)$ and each sweep costs $O(n)$ moves; n simulated steps therefore cost $O(n^2)$ moves of N'. Finally, the input arrives on a single ordinary tape, so N' begins with a preparatory pass converting it to the track format, and ends with one converting track 1 back; each is $O(n)$.

By construction the track contents after the n-th simulated step encode the tapes of N after its n-th step, and N' halts exactly when N does, so the two compute the same partial function. □

A.6.3 Construction of U

Theorem A.33. *There is a Turing machine U such that for every machine M and every input w, U on input $(\ulcorner M \urcorner, w)$ halts if and only if M halts on w, and then leaves the same output.*

Rests on Equation (A.145), Lemma A.32, and Definition 3.92.

Proof. *Derives Theorem A.33.*

We build U with three tapes and appeal to Lemma A.32 at the end.

Layout. Tape 1 holds $\ulcorner M \urcorner$ and is never altered. Tape 2 holds the simulated tape of M, with each symbol a_j of Γ represented by the block $\mathrm{bin}(j)$ and blocks separated by `,`; the simulated head position is marked by a dot on the first symbol of one block, which costs only a doubling of the tape alphabet. Tape 3 holds $\mathrm{bin}(i)$ for the current simulated state q_i.

Initialization. U first verifies that tape 1 carries a well-formed code, as described in Appendix A.6.1, and halts rejecting if not. It then converts the input $w = w_1 \cdots w_p$ into block form on tape 2, marks the first block, and writes $\mathrm{bin}(0)$ on tape 3. Tape 2 now encodes the initial configuration of M on w.

The step loop. U repeats the following.

1. Read the marked block on tape 2 into the control — it is $\mathrm{bin}(j)$ for the currently scanned symbol a_j — and read $\mathrm{bin}(i)$ from tape 3. Neither can be held in the finite control as a whole, since $|Q|$ and $|\Gamma|$ are unbounded across machines M; instead U keeps its head positions and compares symbol by symbol, which is why the lookup is a scan rather than a table jump.

2. Walk tape 1 from the left `#`, comparing the first two fields of each block with $\mathrm{bin}(i)$ and $\mathrm{bin}(j)$. Comparison is performed by moving the two heads in step and matching one binary digit at a time.

3. If the closing `#` is reached with no match, then δ is undefined at (q_i, a_j); U erases its scaffolding, converts tape 2 back to plain symbols, copies it to the output tape and halts.

4. If a block matches, read its remaining three fields $\mathrm{bin}(k)$, $\mathrm{bin}(l)$, d. Copy $\mathrm{bin}(k)$ over tape 3. Overwrite the marked block of tape 2 with $\mathrm{bin}(l)$, shifting the rest of the tape if the two blocks differ in length. Move the mark to the neighbouring block in the direction d, appending a fresh block $\mathrm{bin}(0)$ for $\sqcup$ if the mark would leave the written region.

Faithfulness. We claim that for every n, if M has not halted within n steps then after n passes through the loop, tape 2 encodes the tape of M after n steps with the mark on the scanned cell, and tape 3 holds the numeral of M's state after n steps. The proof is induction on n (Axiom 3.34).

For $n = 0$ this is what initialization established. Assume it for n. Then at step 1 of pass $n + 1$, U reads exactly the state q_i and scanned symbol a_j that M has after n steps. Step 2 locates the block Equation (A.144) for that pair, and by Equation (A.145) there is at most one such block, since δ is a function. If none exists, M halts at step $n + 1$ by Definition 3.92, and step 3 makes U halt with tape 2 decoding to M's final tape. If one exists, step 4 writes a_l in place of a_j, moves the mark one cell in the direction D and records q_k — which is precisely the action Definition 3.92 prescribes for step $n + 1$. So the claim holds for $n + 1$.

Consequently U performs a pass for each step of M and halts exactly when M halts, with the same output; and if M never halts, neither branch of step 3 is ever taken and U loops forever as well. Finally, Lemma A.32 converts this three-tape U into a single-tape machine with the same behaviour, which is the machine promised by Theorem 3.95. □

Remark A.34 (What the construction costs). Each pass of the loop scans a code of fixed length $|\ulcorner M \urcorner|$ and a simulated tape of length $O(n)$ after n steps, so U simulates n steps of M in $O(n^2 \, |\ulcorner M \urcorner|)$ of its own, and the reduction to one tape squares that again. The overhead is polynomial, and that is the point: universality costs a slowdown, never a loss of what can be computed at all. This is the precise sense in which Axiom 3.94 is unaffected by the choice of machine model.

Appendix A.6 discharges the proof obligation of Theorem 3.95. Its consequence is the one exploited throughout Section 3.9: a program is a string, a string is data, and data can be fed to a program — including to the program that produced it. Theorem 3.96 is that observation turned against itself, and Theorem 3.97 generalizes the damage.

A.7 Completeness of the Real Numbers

This appendix proves the statement on which all of real and complex analysis rests (Chapter 7): *there exists a complete ordered field*, unique up to isomorphism — the real numbers. It is the deepest proof of Parts I and II: every convergence theorem of analysis (monotone convergence, Bolzano–Weierstrass, the Cauchy criterion, the intermediate and extreme value theorems, and through them the fundamental theorem of calculus and Cauchy's integral theorem) descends from the single property proven here. The construction is Dedekind's; the rationals $\mathbb{Q}$, with their arithmetic and order, are taken as already constructed (Chapters 3 and 4).

A.7.1 Statement

Theorem A.35 (Completeness of $\mathbb{R}$)**.** *There exists an ordered field $\mathbb{R}$ containing $\mathbb{Q}$ as an ordered subfield, with the* least-upper-bound property*: every nonempty subset $S \subset \mathbb{R}$ that is bounded above has a least upper bound $\sup S \in \mathbb{R}$. Moreover $\mathbb{R}$ is Archimedean, $\mathbb{Q}$ is dense in it, and every Cauchy sequence in $\mathbb{R}$ converges.*

Rests on Definition 4.30.

The proof occupies the rest of this section: the construction (Appendix A.7.2), the order and the least-upper-bound property (Appendix A.7.3) — proven *before* the arithmetic, because it is the property everything else serves — the field operations (Appendix A.7.4), and the consequences (Appendix A.7.5).

A.7.2 The construction: Dedekind cuts

The idea is geometric: a real number is identified with the set of all rationals to its left. What must be checked is that the totality of such left rays carries an order, an arithmetic, and no gaps.

Definition A.36 (Cut)**.** A *cut* is a set $\alpha \subset \mathbb{Q}$ such that

1. $\alpha \neq \varnothing$ and $\alpha \neq \mathbb{Q}$ (*nontriviality*);
2. if $p \in \alpha$ and $q < p$ then $q \in \alpha$ (*downward closure*);
3. α has no largest element: for every $p \in \alpha$ there is $r \in \alpha$ with $p < r$ (*openness*).

The set of all cuts is denoted $\mathbb{R}$. Every rational q embeds as the cut $q^* = \{p \in \mathbb{Q} \mid p < q\}$; properties (1)–(3) are immediate for q^*, and $p < q \iff p^* \subsetneq q^*$, so the embedding preserves order.

Lemma A.37 (Elementary properties of cuts)**.** *Let α be a cut, $p \in \alpha$, $q \notin \alpha$. Then $p < q$; and every rational $r > q$ satisfies $r \notin \alpha$.*

Rests on Definition A.36.

Proof. *Derives Lemma A.37.*

If $q \leq p$ then downward closure ($q < p$) or membership itself ($q = p$) would put $q \in \alpha$, a contradiction; so $p < q$. If $r > q$ and $r \in \alpha$, downward closure would give $q \in \alpha$ again — contradiction. □

A.7.3 Order and the least-upper-bound property

Definition A.38 (Order on $\mathbb{R}$)**.** For cuts α, β: $\alpha < \beta$ if and only if $\alpha \subsetneq \beta$; and $\alpha \leq \beta$ iff $\alpha \subseteq \beta$.

Rests on Definition A.36.

Lemma A.39 (The order is total)**.** *For any cuts α, β, exactly one of $\alpha < \beta$, $\alpha = \beta$, $\beta < \alpha$ holds.*

Rests on Definition A.38 and Lemma A.37.

Proof. *Derives Lemma A.39.*

At most one holds, since the three are mutually exclusive by construction. Suppose $\alpha \not< \beta$ and $\alpha \neq \beta$; we must show $\beta < \alpha$, i.e. $\beta \subsetneq \alpha$. Since $\alpha \not\subseteq \beta$, pick $p \in \alpha \setminus \beta$. Let $q \in \beta$ be arbitrary. Applying Lemma A.37 to the cut β (with $q \in \beta$, $p \notin \beta$) gives $q < p$; downward closure of α then gives $q \in \alpha$. Hence $\beta \subseteq \alpha$, and the inclusion is strict because $p \notin \beta$. □

Theorem A.40 (Least-upper-bound property)**.** *Let $A \subset \mathbb{R}$ be nonempty and bounded above. Then $\gamma = \bigcup_{\alpha \in A} \alpha$ is a cut, and $\gamma = \sup A$.*

Rests on Definitions A.36 and A.38.

Proof. *Derives Theorem A.40.*

γ is a cut. Nontriviality: $\gamma \supseteq \alpha_0 \neq \varnothing$ for any $\alpha_0 \in A$; and if β is an upper bound of A then every $\alpha \in A$ satisfies $\alpha \subseteq \beta$, so $\gamma \subseteq \beta \subsetneq \mathbb{Q}$. Downward closure: if $p \in \gamma$ then $p \in \alpha$ for some $\alpha \in A$, and any $q < p$ lies in $\alpha \subseteq \gamma$. Openness: the same α contains some $r > p$, and $r \in \gamma$.

γ is an upper bound. $\alpha \subseteq \gamma$ for every $\alpha \in A$ by construction, i.e. $\alpha \leq \gamma$.

γ is the least one. Let δ be any upper bound of A. Then $\alpha \subseteq \delta$ for all $\alpha \in A$, hence $\gamma = \bigcup_{\alpha \in A} \alpha \subseteq \delta$, i.e. $\gamma \leq \delta$. □

Note how little the proof costs once the construction is right: the supremum is literally the union. The gaps of $\mathbb{Q}$ — the cut $\{p \in \mathbb{Q} \mid p \leq 0 \text{ or } p^2 < 2\}$ has no rational supremum — are filled by fiat, because the union of cuts is again a cut.

A.7.4 The field operations

Definition A.41 (Addition)**.** $\alpha + \beta = \{p + q \mid p \in \alpha,\ q \in \beta\}$, with neutral element $0^* $ and $-\alpha = \{p \in \mathbb{Q} \mid \exists\, r > 0 :\ -p - r \notin \alpha\}$.

Rests on Definition A.36.

Lemma A.42 ($(\mathbb{R}, +)$ is an ordered abelian group)**.** *$\alpha + \beta$ and $-\alpha$ are cuts; addition is associative and commutative with $\alpha + 0^* = \alpha$ and $\alpha + (-\alpha) = 0^*$; and $\alpha \leq \beta$ implies $\alpha + \gamma \leq \beta + \gamma$.*

Rests on Definitions A.38 and A.41 and Lemma A.37.

Proof. *Derives Lemma A.42.*

$\alpha + \beta$ is a cut. It is nonempty; it omits $p' + q'$ whenever $p' \notin \alpha$, $q' \notin \beta$ (by Lemma A.37 any element $p + q < p' + q'$); it is downward closed, since $s < p + q$ writes as $s = (s - q) + q$ with $s - q < p$, so $s - q \in \alpha$; and it has no largest element, since p may be enlarged inside α. Associativity and commutativity are inherited from $\mathbb{Q}$ elementwise, and $\alpha + 0^* = \alpha$ follows from downward closure and openness: $p + n < p$ for $n < 0$ gives $\alpha + 0^* \subseteq \alpha$, while any $p \in \alpha$ has $r \in \alpha$, $r > p$, whence $p = r + (p - r) \in \alpha + 0^*$.

Inverse. $-\alpha$ is a cut (checks as above). For $\alpha + (-\alpha) \subseteq 0^*$: if $p \in \alpha$ and $q \in -\alpha$ with witness $r > 0$ (so $-q - r \notin \alpha$), then $p < -q - r$ by Lemma A.37, so $p + q < -r < 0$. Conversely let $s < 0$; put $t = -s/2 > 0$. By the Archimedean property of $\mathbb{Q}$ there is an integer n with $nt \in \alpha$ and $(n + 1)t \notin \alpha$ (walk upward in steps of t from inside α; the walk exits because $\alpha \neq \mathbb{Q}$). Take $p = nt \in \alpha$ and $q = s - nt$. Then $-q - t = -s + nt - t = nt + 2t - t = (n+1)t \notin \alpha$, so $q \in -\alpha$ with witness t, and $p + q = s$. Monotonicity of $+$ is elementwise. □

Definition A.43 (Multiplication)**.** For $\alpha, \beta > 0^*$:

$$\alpha\beta = \{s \in \mathbb{Q} \mid s \le 0\} \cup \{pq \mid p \in \alpha,\ q \in \beta,\ p > 0,\ q > 0\}\ , \tag{A.148}$$

extended to all signs by $\alpha\, 0^* = 0^*$ and the sign rules $\alpha\beta = -\left((-\alpha)\beta\right)$ etc., with unit 1^* and, for $\alpha > 0^*$, the reciprocal $\alpha^{-1} = \{p \in \mathbb{Q} \mid p \le 0\} \cup \{p > 0 \mid \exists\, r > 0 :\ (p^{-1} - r)^{-1} \notin \alpha\}$.

Rests on Definitions A.36 and A.41.

Lemma A.44 ($\mathbb{R}$ is an ordered field)**.** *With Definitions A.41 and A.43, $\mathbb{R}$ satisfies the ordered-field axioms (Chapter 4), and the embedding $q \mapsto q^*$ is a field homomorphism.*

Rests on Definitions 4.30, A.41 and A.43 and Lemma A.42.

Proof. *Derives Lemma A.44.*

The verifications are of exactly the same kind as in Lemma A.42 — elementwise inheritance from $\mathbb{Q}$ on the positive cone, then bookkeeping of the sign cases for associativity, commutativity, distributivity, and $\alpha\alpha^{-1} = 1^*$ (the reciprocal argument mirrors the additive-inverse argument with the multiplicative Archimedean walk $p, 2p, 4p, \dots$). They are long but routine, and no idea beyond those already displayed enters; we omit the repetition. Compatibility of order and product — $\alpha, \beta > 0^* \implies \alpha\beta > 0^*$ — is immediate from Equation (A.148). □

A.7.5 Consequences

Corollary A.45 (Archimedean property and density of $\mathbb{Q}$)**.** *For $\alpha, \beta \in \mathbb{R}$ with $\alpha > 0^*$ there is $n \in \mathbb{N}$ with $n\alpha > \beta$; and between any two reals lies a rational.*

Rests on Theorem A.40 and Definition A.36.

Proof. *Derives Corollary A.45.*

If $n\alpha \le \beta$ for all n, the set $A = \{n\alpha \mid n \in \mathbb{N}\}$ is bounded above, so $\gamma = \sup A$ exists (Theorem A.40). Then $\gamma - \alpha < \gamma$ is not an upper bound: some $n\alpha > \gamma - \alpha$, whence $(n+1)\alpha > \gamma$ — contradicting that γ bounds A. Density: if $\alpha < \beta$, pick $q \in \beta \setminus \alpha$; openness of β gives $r \in \beta$, $r > q$, and then $\alpha \le q^* < r^* \le \beta$, so the embedded rationals q^*, r^* separate α from β. □

Corollary A.46 (Monotone convergence)**.** *A nondecreasing sequence (a_n) in $\mathbb{R}$ that is bounded above converges to $\sup_n a_n$.*

Rests on Theorem A.40.

Proof. *Derives Corollary A.46.*

Let $s = \sup_n a_n$, which exists by Theorem A.40. Given $\varepsilon > 0$, $s - \varepsilon$ is not an upper bound, so $a_N > s - \varepsilon$ for some N; monotonicity gives $s - \varepsilon < a_n \le s$ for all $n \ge N$, i.e. $|a_n - s| < \varepsilon$. □

Corollary A.47 (Cauchy completeness)**.** *Every Cauchy sequence (a_n) in $\mathbb{R}$ converges in $\mathbb{R}$.*

Rests on Corollary A.46 and Theorem A.40.

Proof. *Derives Corollary A.47.*

A Cauchy sequence is bounded (take $\varepsilon = 1$: beyond some N all terms lie within 1 of a_N, and only finitely many terms remain). Define $b_n = \inf_{k \ge n} a_k$, which exists because $\{a_k \mid k \ge n\}$ is bounded below and Theorem A.40 applied to the negatives provides infima. The sequence (b_n) is nondecreasing and bounded above, so by Corollary A.46 it converges to $L = \sup_n b_n = \liminf_n a_n$. Given $\varepsilon > 0$ choose N with $|a_m - a_n| < \varepsilon/2$ for $m, n \ge N$. Then for $n \ge N$ every a_k with $k \ge n$ lies in $(a_N - \varepsilon/2,\ a_N + \varepsilon/2)$, so $b_n \in [a_N - \varepsilon/2,\ a_N + \varepsilon/2]$ and hence $|L - a_N| \le \varepsilon/2$; combining, $|a_n - L| \le |a_n - a_N| + |a_N - L| < \varepsilon$ for $n \ge N$. □

Proof of Theorem A.35. *Derives Theorem A.35.*

Assembling the section. The set of cuts of Definition A.36, ordered by inclusion, is a totally ordered set containing an order-embedded copy of $\mathbb{Q}$ (Lemmas A.37 and A.39); it has the least-upper-bound property by Theorem A.40; addition and multiplication of Definitions A.41 and A.43 make it an ordered field by Lemmas A.42 and A.44. The three remaining assertions are Corollary A.45 (Archimedean, and $\mathbb{Q}$ dense) and Corollary A.47 (Cauchy sequences converge). Existence is therefore established by construction, which is what the theorem claims. □

Remark A.48 (Uniqueness). Any two complete ordered fields are isomorphic as ordered fields: the isomorphism sends each element to the supremum of the embedded rationals below it, and Corollary A.45 (which followed from completeness alone) makes the map well defined, bijective, and operation-preserving. Thus *the* real numbers are characterized by the axioms, not by the construction; Cantor's alternative construction by Cauchy sequences of rationals yields the same field. This is why Chapter 7 may take Theorem A.35 as its single axiom for $\mathbb{R}$ and never mention cuts again.

A.8 Equivalence of the Einstein–Hilbert Action in Metric and Vielbein Variables

This appendix proves the equivalence asserted in Remark 13.159 (Chapter 13, Equation (13.317)) and used to set up the action of Chapter 43: in $D = p + q$ dimensions, the Einstein–Hilbert Lagrangian density built from the vielbein and the Cartan curvature two-form coincides, up to a fixed combinatorial constant, with $\sqrt{|g|}\,R$. All conventions (frame indices $a, b, \ldots$, the structure equations, the definition $R^a{}_b = \frac{1}{2} R^a{}_{b\mu\nu}\,\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$) are those of Sections 13.1, 13.8 and 13.12.

A.8.1 Statement

Theorem A.49. *Let (M, g) be a D-dimensional pseudo-Riemannian manifold with vielbein e^a, torsion-free spin connection $\omega^a{}_b$, and curvature two-form $R^a{}_b$. Then, as D-forms,*

$$\epsilon_{a_1 a_2 a_3 \ldots a_D}\, R^{a_1 a_2} \wedge e^{a_3} \wedge \cdots \wedge e^{a_D} = (D-2)!\, R\, \sqrt{|g|}\, \mathrm{d}^D x\,, \tag{A.149}$$

with $\epsilon_{a_1 \ldots a_D}$ the numerical (frame-index) Levi-Civita symbol, $\epsilon_{1\ldots D} = +1$, and R the Ricci scalar of Definition 13.153. Equivalently, defining $S_{\mathrm{EH}} = \dfrac{1}{2\kappa}\displaystyle\int \mathrm{d}^D x\, \sqrt{|g|}\, R$ (Chapter 43),

$$S_{\mathrm{EH}} = \frac{1}{2\kappa\,(D-2)!} \int \epsilon_{a_1 \ldots a_D}\, R^{a_1 a_2} \wedge e^{a_3} \wedge \cdots \wedge e^{a_D}\,. \tag{A.150}$$

Rests on Lemma A.50, Definition 13.153, Equations (13.264) and (13.310), and Proposition 13.123.

The proof needs one combinatorial lemma about the Levi-Civita symbol, proven in Appendix A.8.2, and a component computation carried out at a conveniently chosen point of M, in Appendix A.8.3.

A.8.2 A contraction lemma for the Levi-Civita symbol

Lemma A.50. *Let $\epsilon_{i_1 \ldots i_D}$ be the totally antisymmetric numerical symbol in D indices ranging over $1, \ldots, D$, normalized by $\epsilon_{1\ldots D} = +1$, and write $\epsilon^{i_1 \ldots i_D}$ for the same numerical values with the indices held upstairs. Then*

$$\epsilon^{i\, j\, k_3 \ldots k_D}\, \epsilon_{l\, m\, k_3 \ldots k_D} = (D-2)!\, \left(\delta^i_l \delta^j_m - \delta^i_m \delta^j_l\right)\,, \tag{A.151}$$

summed over $k_3, \ldots, k_D = 1, \ldots, D$.

Proof. *Derives Lemma A.50.*

Both sides of Equation (A.151) are, as functions of the four free indices i, j, l, m, antisymmetric under $i \leftrightarrow j$ and under $l \leftrightarrow m$ separately: for the right side this is immediate; for the left side, swapping $i \leftrightarrow j$ swaps the first two slots of $\epsilon^{ijk_3\cdots}$, which flips its sign by total antisymmetry, and the swap $l \leftrightarrow m$ acts the same way on the second factor. Hence both sides vanish whenever $i = j$ or $l = m$, and, by the double antisymmetry, it suffices to verify Equation (A.151) for the case $i = l$, $j = m$, $i \neq j$ (every other nonvanishing assignment of $\{i, j\} = \{l, m\}$ then follows by the antisymmetries just established, and every assignment with $\{i, j\} \neq \{l, m\}$ gives $0 = 0$: the left side vanishes because no choice of $k_3, \ldots, k_D$ can make *both* ϵ factors nonzero when $\{i, j\} \neq \{l, m\}$, since each factor forces its own two free indices to be distinct from the shared $k_3, \ldots, k_D$ and from each other, and a nonzero product further forces the sets $\{i, j\}$ and $\{l, m\}$ to be identical, as both equal the complement of $\{k_3, \ldots, k_D\}$ in $\{1, \ldots, D\}$).

At $i = l$, $j = m$, $i \neq j$: the right side of Equation (A.151) is $(D-2)!\,(1-0) = (D-2)!$. For the left side, $\epsilon^{ijk_3\ldots k_D}\epsilon_{ijk_3\ldots k_D}$ (no sum yet on i, j) is nonzero only when $(k_3, \ldots, k_D)$ is a permutation of the $D-2$ elements of $\{1, \ldots, D\} \setminus \{i, j\}$, and each such term contributes $\left(\epsilon^{ijk_3\ldots k_D}\right)^2 = 1$; there are exactly $(D-2)!$ such permutations, so the sum equals $(D-2)!$. The two sides agree. □

A.8.3 Proof of the theorem

Proof of Theorem A.49. *Derives Theorem A.49.*

Fix a point $P \in M$. Expand the curvature two-form and the vielbeins in the coordinate basis,

$$R^{a_1a_2} = \tfrac{1}{2}\, R^{a_1a_2}{}_{\mu\nu}\, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu\,, \qquad e^{a_i} = e^{a_i}{}_{\lambda_i}\, \mathrm{d}x^{\lambda_i}\,,$$

so that, writing Ω for the left side of Equation (A.149),

$$\Omega = \tfrac{1}{2}\, \epsilon_{a_1\ldots a_D}\, R^{a_1a_2}{}_{\mu\nu}\, e^{a_3}{}_{\lambda_3} \cdots e^{a_D}{}_{\lambda_D}\, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \wedge \mathrm{d}x^{\lambda_3} \wedge \cdots \wedge \mathrm{d}x^{\lambda_D}\,.$$

Choice of point-adapted frame. Both sides of Equation (A.149) are built — with no further differentiation — from the algebraic data $(g_{\mu\nu}(P),\, R^a{}_{b\mu\nu}(P),\, e^a{}_\mu(P))$ and the numerical symbols ϵ, η; consequently, once verified in *one* admissible frame and coordinate system at P, the identity holds in every frame and coordinate system at P, by the tensorial (respectively, local-Lorentz-covariant) transformation laws of g, $R^a{}_{b\mu\nu}$ and $e^a{}_\mu$ established in Sections 13.5.2 and 13.8 and Theorem 13.156 — both sides of Equation (A.149) transform as the components of a D-form, i.e. by the same Jacobian determinant, so an equality of components in one frame is an equality of components in every frame. It therefore suffices to prove Equation (A.149) at P in a coordinate system and frame chosen so that

$$e^a{}_\mu(P) = \delta^a_\mu\,, \qquad g_{\mu\nu}(P) = \eta_{\mu\nu}\,. \tag{A.152}$$

Such a choice exists: pick any coordinates first, so that $g_{\mu\nu}(P)$ is some fixed nondegenerate symmetric matrix of signature (p, q); a constant linear change of coordinates diagonalizes it to $\eta_{\mu\nu}$ at P (Sylvester's law of inertia, Chapter 4), after which the vielbein $e^a{}_\mu = \delta^a_\mu$ satisfies $\eta_{ab}\,\delta^a_\mu\delta^b_\nu = \eta_{\mu\nu} = g_{\mu\nu}(P)$ (Equation (13.264)), and every other vielbein compatible with $g_{\mu\nu}(P)$ differs from this one by a local Lorentz transformation (Proposition 13.123), so this choice is always reachable.

Component computation at P. Under Equation (A.152), $e^{a_i}{}_{\lambda_i}(P) = \delta^{a_i}_{\lambda_i}$ collapses each frame sum $a_3, \ldots, a_D$ against the epsilon symbol into the corresponding coordinate index, and

$R^{a_1a_2}{}_{\mu\nu}(P)$ becomes numerically the doubly-raised Riemann tensor $g^{a_1\rho}(P)g^{a_2\sigma}(P)R_{\rho\sigma\mu\nu}(P)$, so that at P,

$$\Omega(P) = \tfrac{1}{2}\,\epsilon_{\rho\sigma\lambda_3\ldots\lambda_D}\,R^{\rho\sigma}{}_{\mu\nu}(P)\,\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu\wedge\mathrm{d}x^{\lambda_3}\wedge\cdots\wedge\mathrm{d}x^{\lambda_D}\,,$$

where the ρ,σ sum has simply been relabelled from a_1,a_2. Reordering the wedge of coordinate differentials (Section 13.6.3), $\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu\wedge\mathrm{d}x^{\lambda_3}\wedge\cdots\wedge\mathrm{d}x^{\lambda_D} = \varepsilon^{\mu\nu\lambda_3\ldots\lambda_D}\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^D$, so

$$\Omega(P) = \tfrac{1}{2}\,R^{\rho\sigma}{}_{\mu\nu}(P)\,\left(\varepsilon^{\mu\nu\lambda_3\ldots\lambda_D}\epsilon_{\rho\sigma\lambda_3\ldots\lambda_D}\right)\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^D\,.$$

By Lemma A.50 (with $D-2$ contracted indices $\lambda_3,\ldots,\lambda_D$), the bracket equals $(D-2)!\left(\delta^\mu_\rho\delta^\nu_\sigma-\delta^\mu_\sigma\delta^\nu_\rho\right)$, giving

$$\Omega(P) = \frac{(D-2)!}{2}\,[R^{\rho\sigma}{}_{\rho\sigma}(P) - R^{\rho\sigma}{}_{\sigma\rho}(P)]\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^D\,.$$

Renaming the summed indices $\rho\leftrightarrow\sigma$ in the second term and using the antisymmetry of the curvature two-form in its upper pair, $R^{\rho\sigma}{}_{\mu\nu} = -R^{\sigma\rho}{}_{\mu\nu}$ (inherited from $\omega^{ab} = -\omega^{ba}$, Equation (13.310)),

$$R^{\rho\sigma}{}_{\sigma\rho}(P) = R^{\sigma\rho}{}_{\rho\sigma}(P) = -R^{\rho\sigma}{}_{\rho\sigma}(P)\,,$$

so the bracket is $2\,R^{\rho\sigma}{}_{\rho\sigma}(P)$ and

$$\Omega(P) = (D-2)!\;R^{\rho\sigma}{}_{\rho\sigma}(P)\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^D\,. \tag{A.153}$$

Identification with the Ricci scalar. It remains to recognize $R^{\rho\sigma}{}_{\rho\sigma}(P)$. Lowering both upper indices with $g^{-1}(P)$,

$$R^{\rho\sigma}{}_{\rho\sigma}(P) = g^{\rho\alpha}(P)\,g^{\sigma\beta}(P)\,R_{\alpha\beta\rho\sigma}(P)\,.$$

Contract the inner index pair first: by Definition 13.153, $R_{\mu\nu} = R^\lambda{}_{\mu\lambda\nu} = g^{\lambda\kappa}R_{\kappa\mu\lambda\nu}$, i.e. contracting the first index of the (fully lowered) Riemann tensor with the third, via g^{-1}, produces the Ricci tensor in its second and fourth slots. Applying this with $\kappa\to\rho$, $\mu\to\beta$, $\lambda\to\alpha$, $\nu\to\sigma$ gives $g^{\rho\alpha}R_{\alpha\beta\rho\sigma} = R_{\beta\sigma}$, so

$$R^{\rho\sigma}{}_{\rho\sigma}(P) = g^{\sigma\beta}(P)\,R_{\beta\sigma}(P) = g^{\mu\nu}(P)\,R_{\mu\nu}(P) = R(P)\,,$$

the last equality being Equation (13.307). Hence $R^{\rho\sigma}{}_{\rho\sigma}(P) = R(P)$, and Equation (A.153) reads

$$\Omega(P) = (D-2)!\;R(P)\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^D\,.$$

Finally, under Equation (A.152), $g(P) = \det\eta_{\mu\nu} = (-1)^q$, so $\sqrt{|g(P)|} = 1$ and $\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^D = \sqrt{|g(P)|}\,\mathrm{d}^Dx$ at P. Substituting,

$$\Omega(P) = (D-2)!\;R(P)\,\sqrt{|g(P)|}\,\mathrm{d}^Dx\,,$$

which is Equation (A.149) at P. Since $P\in M$ was arbitrary and, by the covariance argument above, the identity transported to any frame at P, Equation (A.149) holds identically on M. Dividing by $2\kappa\,(D-2)!$ and integrating gives Equation (A.150). □

Remark A.51 (Independent check: $D=2$). For $D=2$ there are no vielbein factors left in Equation (A.149) ($(D-2)! = 0! = 1$, and the product $e^{a_3}\wedge\cdots\wedge e^{a_D}$ is empty), so the claim reduces to $\epsilon_{ab}R^{ab} = R\sqrt{|g|}\,\mathrm{d}^2x$. Using $R^{12} = -R^{21}$ and $\epsilon_{12} = -\epsilon_{21} = 1$, the left side is $2R^1{}_2$, and $R^1{}_2 = K\,e^1\wedge e^2$ with K the Gaussian curvature of Section 13.3.3; together with $R = 2K$ in $D=2$ (Definition 13.153) this gives $2R^1{}_2 = 2K\,e^1\wedge e^2 = R\sqrt{|g|}\,\mathrm{d}^2x$, matching the theorem with the correct constant $(D-2)! = 1$ and independently confirming Lemma A.50 in its simplest nontrivial case — see Section 13.3.3 for the surface-theory computation this reproduces.

Appendix A.8 closes the derivation anticipated in Chapter 13; the action Equation (A.150) is put to use, and its variation carried out, in Chapters 43 and 44.

A.9 Wigner's Theorem on Quantum Symmetries

This appendix proves Theorem 94.3 (Chapter 94): every bijection of the unit rays of a Hilbert space that preserves transition probabilities is implemented by an operator that is either unitary or antiunitary, unique up to a phase. The theorem is stated in Wigner's 1931 monograph [687]; the proof below is Bargmann's [1269], streamlined but complete. Nothing in it assumes the Hilbert space separable: the basis index k may run over an arbitrary set.

Throughout, $\mathcal{H}$ is a complex Hilbert space with $\dim \mathcal{H} \geq 2$, the *unit ray* of a normalized vector is $[\psi] = \{e^{i\alpha}\psi \mid \alpha \in \mathbb{R}\}$, and T is a bijection of the set of unit rays with

$$|\langle \varphi' | \psi' \rangle| = |\langle \varphi | \psi \rangle| \qquad \text{for } [\psi'] = T[\psi],\ [\varphi'] = T[\varphi] \tag{A.154}$$

(representatives may be chosen arbitrarily: the modulus does not depend on the choice).

A.9.1 Orthonormal bases map to orthonormal bases

Lemma A.52. *Let $\{e_k\}$ be an orthonormal basis of $\mathcal{H}$ and choose any unit representatives $\psi_k \in T[e_k]$. Then $\{\psi_k\}$ is an orthonormal basis.*

Rests on Equation (A.154).

Proof. *Derives Lemma A.52.*

Orthonormality is Equation (A.154): $|\langle \psi_j | \psi_k \rangle| = |\langle e_j | e_k \rangle| = \delta_{jk}$. Completeness: let the unit vector χ be orthogonal to every ψ_k. Since T is surjective, $[\chi] = T[\varphi]$ for some unit φ, and then $|\langle \varphi | e_k \rangle| = |\langle \chi | \psi_k \rangle| = 0$ for every k, contradicting completeness of $\{e_k\}$ and $\|\varphi\| = 1$. □

A.9.2 Phase alignment along a basis

Fix the orthonormal basis $\{e_k\}$, single out one index, written $k = 1$, and abbreviate the normalized two- and three-component test vectors

$$u_k = \frac{e_1 + e_k}{\sqrt{2}} \quad (k \geq 2)\,, \qquad u_{jk} = \frac{e_1 + e_j + e_k}{\sqrt{3}} \quad (j \neq k\,,\ j,k \geq 2)\,. \tag{A.155}$$

Lemma A.53 (Aligned representatives)**.** *The representatives ψ_k of Lemma A.52 can be rephased, $f_1 = \psi_1$ and $f_k = e^{i\gamma_k}\psi_k$, so that the first relation below holds for every $k \geq 2$ and the remaining two for all $j \neq k$ with $j, k \geq 2$:*

$$T[u_k] = \left[\frac{f_1 + f_k}{\sqrt{2}}\right], \qquad T[u_{jk}] = \left[\frac{f_1 + f_j + f_k}{\sqrt{3}}\right], \qquad T\left[\frac{e_j + e_k}{\sqrt{2}}\right] = \left[\frac{f_j + f_k}{\sqrt{2}}\right]. \tag{A.156}$$

Rests on Lemma A.52 and Equations (A.154) and (A.155).

Proof. *Derives Lemma A.53.*

Expand a unit representative of $T[u_k]$ in the basis $\{\psi_j\}$: by Equation (A.154) tested against each $[e_j]$, its coefficients have moduli $1/\sqrt{2}$ at $j \in \{1, k\}$ and 0 elsewhere, so the representative is $(e^{i\alpha}\psi_1 + e^{i\beta}\psi_k)/\sqrt{2}$; multiplying by $e^{-i\alpha}$ (same ray) makes it $(\psi_1 + e^{i\gamma_k}\psi_k)/\sqrt{2}$ with a uniquely determined γ_k. Define $f_k = e^{i\gamma_k}\psi_k$; the first relation of Equation (A.156) holds by construction, and $f_k \in T[e_k]$ still. For the second: a representative of $T[u_{jk}]$ likewise reads $(f_1 + c_j f_j + c_k f_k)/\sqrt{3}$ with $|c_j| = |c_k| = 1$ after normalizing the f_1-coefficient; testing against $T[u_j]$ and $T[u_k]$ with Equation (A.154), $|1 + c_j| = |1 + 1| = 2$ forces $c_j = 1$, and similarly $c_k = 1$. For the third: a representative of $T[(e_j + e_k)/\sqrt{2}]$ is $(e^{i\mu} f_j + e^{i\nu} f_k)/\sqrt{2}$; testing against $T[u_{jk}]$, $|e^{-i\mu} + e^{-i\nu}| = 2$ forces $\mu = \nu$, and the common phase is dropped. □

A.9.3 The induced scalar map

Lemma A.54 (The map χ)**.** *For every $k \geq 2$ and $\alpha \in \mathbb{C}$ there is a unique $\chi_k(\alpha) \in \mathbb{C}$ with*

$$T\left[\frac{e_1 + \alpha e_k}{\sqrt{1+|\alpha|^2}}\right] = \left[\frac{f_1 + \chi_k(\alpha)\, f_k}{\sqrt{1+|\alpha|^2}}\right], \tag{A.157}$$

and $\chi_k(\alpha) \in \{\alpha, \bar{\alpha}\}$. Moreover a single alternative holds for all α and all k: either $\chi_k(\alpha) = \alpha$ for every k and α, or $\chi_k(\alpha) = \bar{\alpha}$ for every k and α. Write χ for the common map.

Rests on Lemma A.53 and Equation (A.154).

Proof. *Derives Lemma A.54.*

As in Lemma A.53, a representative of the left side has components only along f_1, f_k with moduli $1/\sqrt{1+|\alpha|^2}$ and $|\alpha|/\sqrt{1+|\alpha|^2}$; normalizing the f_1-component to be real positive determines the representative and hence $\chi_k(\alpha)$ uniquely, with $|\chi_k(\alpha)| = |\alpha|$. Testing against $T[u_k]$ via Equation (A.154) gives $|1 + \chi_k(\alpha)| = |1 + \alpha|$; together with equal moduli this forces $\Re\chi_k(\alpha) = \Re\alpha$, hence $\chi_k(\alpha) \in \{\alpha, \bar{\alpha}\}$.

No mixing within one k: suppose $\chi_k(\alpha) = \alpha$ and $\chi_k(\beta) = \bar{\beta}$ with $\Im\alpha \neq 0 \neq \Im\beta$. Test the rays Equation (A.157) for α and β against each other: Equation (A.154) demands $|1 + \bar{\alpha}\beta| = \left|1 + \bar{\alpha}\bar{\beta}\right|$, i.e. $\Re\left(\bar{\alpha}\beta\right) = \Re\left(\bar{\alpha}\bar{\beta}\right)$, i.e. $\Im\alpha\,\Im\beta = 0$ — a contradiction. Real α satisfy both alternatives, so χ_k is the identity on all of $\mathbb{C}$ or conjugation on all of $\mathbb{C}$.

No mixing across k: let $j \neq k$ and suppose $\chi_j = \mathrm{id}$ while χ_k is conjugation, and pick $\alpha = \beta = \mathrm{i}$. A representative of $T[(e_1 + \alpha e_j + \beta e_k)/\sqrt{3}]$, normalized so that its f_1-coefficient is real positive, reads $(f_1 + c_j f_j + c_k f_k)/\sqrt{3}$ with $|c_j| = |\alpha|$ and $|c_k| = |\beta|$ (tests against each $[e_m]$). Testing against the family Equation (A.157) with target j, at the two values $\gamma = 1$ and $\gamma = \mathrm{i}$, gives $\left|1 + \overline{\chi_j(\gamma)}\, c_j\right| = |1 + \bar{\gamma}\,\alpha|$, which pins the real and the imaginary part of c_j and forces $c_j = \chi_j(\alpha)$ — the within-target dichotomy just proved fixes χ_j globally, so $\chi_j(\gamma)$ is known at both test values — and likewise $c_k = \chi_k(\beta)$. Testing the resulting representative against $T[(e_j + e_k)/\sqrt{2}]$ — the third relation of Equation (A.156) — yields $|\chi_j(\alpha) + \chi_k(\beta)| = |\alpha + \beta| = 2$, whereas $\chi_j(\mathrm{i}) + \chi_k(\mathrm{i}) = \mathrm{i} + \bar{\mathrm{i}} = 0$: a contradiction. (When $\dim \mathcal{H} = 2$ there is a single k and this step is vacuous.) □

A.9.4 Proof of the theorem

Define, on finite linear combinations and extended by continuity,

$$U\left(\sum_k a_k e_k\right) = \sum_k \chi(a_k)\, f_k\,. \tag{A.158}$$

If $\chi = \mathrm{id}$, U is linear and maps the orthonormal basis $\{e_k\}$ to the orthonormal basis $\{f_k\}$ (Lemma A.52): it is unitary. If χ is conjugation, U is antilinear with $\langle U\varphi | U\psi\rangle = \overline{\langle\varphi|\psi\rangle}$: antiunitary.

Proof that U implements T. *Derives Theorem 94.3.*

Let $\psi = \sum_k a_k e_k$ be an arbitrary unit vector and $\varphi = \sum_k b_k f_k$ a unit representative of $T[\psi]$. Testing against each $[e_k]$ gives $|b_k| = |a_k|$. If only one $a_n \neq 0$, then $[\psi] = [e_n]$ and $T[\psi] = [f_n] = [U\psi]$ directly. Otherwise fix an index n with $a_n \neq 0$, taking $n = 1$ whenever $a_1 \neq 0$, and rephase φ so that $b_n = \chi(a_n)$ (possible since $|b_n| = |a_n|$). Components with $a_k = 0$ are settled at once: $|b_k| = |a_k| = 0$ gives $b_k = 0 = \chi(a_k)$; in particular, if $n \neq 1$ then $a_1 = 0$ by the choice of n, and only targets $k \geq 2$, $k \neq n$, remain to be pinned. For every such target and every $\gamma \in \mathbb{C}$ we claim

$$T\left[\frac{e_n + \gamma e_k}{\sqrt{1+|\gamma|^2}}\right] = \left[\frac{f_n + \chi(\gamma)\, f_k}{\sqrt{1+|\gamma|^2}}\right] \tag{A.159}$$

— with the *same* global χ. For $n = 1$ this is Lemma A.54 itself. For $n \geq 2$, the first two paragraphs of the proof of Lemma A.54, run with anchor n in place of 1 — the needed alignment inputs, the pair $(f_n + f_k)/\sqrt{2}$ and the triples based at n, are supplied by Equation (A.156) — yield Equation (A.159) with some uniform alternative $\chi_k^{(n)} \in \{\mathrm{id}, \text{conjugation}\}$ in place of χ. To identify the alternative, test the ray $[(e_n + \mathrm{i}e_k)/\sqrt{2}]$ against $T[(e_1 + e_n + \mathrm{i}e_k)/\sqrt{3}] = [(f_1 + f_n + \chi(\mathrm{i})\, f_k)/\sqrt{3}]$ — the latter expansion established exactly as in the no-mixing step of Lemma A.54 (anchor 1, targets n and k): Equation (A.154) demands

$$\left|1 + \overline{\chi_k^{(n)}(\mathrm{i})}\, \chi(\mathrm{i})\right| = |\langle e_n + \mathrm{i}e_k | e_1 + e_n + \mathrm{i}e_k\rangle| = 2\,,$$

and since $\chi_k^{(n)}(\mathrm{i}) = \pm\mathrm{i}$ and $\chi(\mathrm{i}) = \pm\mathrm{i}$, the left side equals 2 when $\chi_k^{(n)}(\mathrm{i}) = \chi(\mathrm{i})$ and 0 otherwise: hence $\chi_k^{(n)}(\mathrm{i}) = \chi(\mathrm{i})$, and the uniform dichotomy forces $\chi_k^{(n)} = \chi$, proving the claim. Applying Equation (A.154) to the ray Equation (A.159) against $[\psi]$:

$$|a_n + \bar{\gamma}\, a_k| = \left|b_n + \overline{\chi(\gamma)}\, b_k\right| \qquad \text{for all } \gamma \in \mathbb{C}\,.$$

Square both sides and cancel the equal terms $|a_n|^2 + |\gamma|^2\, |a_k|^2$:

$$\Re\,(\bar{a}_n\, \bar{\gamma}\, a_k) = \Re\left(\bar{b}_n\, \overline{\chi(\gamma)}\, b_k\right) = \Re\left(\overline{\chi(a_n)}\, \overline{\chi(\gamma)}\, b_k\right)\,.$$

For $\chi = \mathrm{id}$, running γ over 1 and i recovers both real and imaginary parts of $\bar{a}_n a_k$ and $\bar{a}_n b_k$ and forces $b_k = a_k = \chi(a_k)$; for χ conjugation, the same two tests force $b_k = \bar{a}_k = \chi(a_k)$. Hence $\varphi = \sum_k \chi(a_k) f_k = U\psi$, so $T[\psi] = [U\psi]$. □

Uniqueness up to phase, and exclusivity. *Derives Theorem 94.3.*

Let U and V both implement T and set $M = U^{-1}V$, an operator preserving every unit ray: $[M\psi] = [\psi]$. On the basis, $Me_k = \lambda_k e_k$ with $|\lambda_k| = 1$.

If U and V have the same linearity character, M is linear: $M(e_1 + e_k) = \lambda_1 e_1 + \lambda_k e_k$ must be proportional to $e_1 + e_k$, so $\lambda_k = \lambda_1$ for all k and $M = \lambda_1 \mathbb{1}$: $V = \lambda_1 U$.

If they had opposite character, M would be antilinear. Then $M(e_1 + e_k) = \lambda_1 e_1 + \lambda_k e_k \propto e_1 + e_k$ gives $\lambda_k = \lambda_1$, while $M(e_1 + \mathrm{i}e_k) = \lambda_1 e_1 - \mathrm{i}\lambda_k e_k \propto e_1 + \mathrm{i}e_k$ gives $\lambda_k = -\lambda_1$ — impossible. Hence for $\dim \mathcal{H} \geq 2$ a symmetry transformation is implemented *either* by unitaries *or* by antiunitaries, never both, and within its class the implementation is unique up to an overall phase. □

Remark A.55 (One dimension). If $\dim \mathcal{H} = 1$ there is a single unit ray, T is the identity, and it is implemented both by $\mathbb{1}$ (unitary) and by complex conjugation in any basis (antiunitary): the dichotomy of Theorem 94.3 genuinely requires $\dim \mathcal{H} \geq 2$.

Appendix A.9 discharges the proof obligation of Theorem 94.3; it is put to work in Corollary 94.4 — symmetries of a connected group act by honest unitaries — and through it in the whole classification of Chapter 94. The antiunitary branch is not an idle alternative: time reversal must take it, as Part XI will use.

A.10 The Classification of Kinematical Algebras

This appendix proves Theorem 14.90 (Chapter 14): under isotropy, the two discrete automorphisms, and noncompactness of the boosts, there are exactly eleven kinematical Lie algebras, organized as displayed in Table 14.1. The result is Bacry and Lévy-Leblond's [1364]; the derivation below is self-contained, and it is written out here because what it yields is not only a count but the parametrization that makes the whole family intelligible — three constants, two Jacobi constraints, and a lattice of limits that turns out to be exactly the contraction lattice.

A.10.1 The generators and the hypotheses

A *kinematical* algebra $\mathfrak{k}$ is a real ten-dimensional Lie algebra spanned by

$$J_i \text{ (rotations)}, \quad K_i \text{ (boosts)}, \quad P_i \text{ (space translations)}, \quad H \text{ (time translation)}, \tag{A.160}$$

with $i = 1, 2, 3$, subject to the three hypotheses of Theorem 14.90. Written out, they are the following.

(H1) Isotropy. The J_i close into $\mathfrak{so}(3)$, under which K_i and P_i are vectors and H is a scalar:

$$[J_i, J_j] = \epsilon_{ijk} J_k\,, \quad [J_i, K_j] = \epsilon_{ijk} K_k\,, \quad [J_i, P_j] = \epsilon_{ijk} P_k\,, \quad [J_i, H] = 0\,. \tag{A.161}$$

(H2) Parity and time reversal are automorphisms. The linear maps

$$\Pi: \quad (J_i, K_i, P_i, H) \longmapsto (J_i, -K_i, -P_i, H)\,, \tag{A.162}$$

$$\Theta: \quad (J_i, K_i, P_i, H) \longmapsto (J_i, -K_i, P_i, -H) \tag{A.163}$$

are automorphisms of $\mathfrak{k}$. These are the algebra-level statements of $\boldsymbol{x} \mapsto -\boldsymbol{x}$ and $t \mapsto -t$: a boost velocity reverses under either, a position reverses only under the first, and the generator of translation in time reverses only under the second.

(H3) Noncompact boosts. Each K_i generates a one-parameter subgroup isomorphic to $\mathbb{R}$ and not to a circle. Only the sign of one structure constant will turn out to be at stake, and (H3) is used exactly twice, in the last two cases of Appendix A.10.5.

A.10.2 Step one: the hypotheses reduce the algebra to five constants

Lemma A.56 (The general kinematical bracket)**.** *Hypotheses (H1) and (H2) force the brackets not already fixed by Equation* (A.161) *to take the form*

$$[K_i, K_j] = \alpha\, \epsilon_{ijk} J_k\,, \qquad [P_i, P_j] = \beta\, \epsilon_{ijk} J_k\,, \tag{A.164}$$

$$[K_i, P_j] = \gamma\, \delta_{ij} H\,, \qquad [K_i, H] = \mu\, P_i\,, \tag{A.165}$$

$$[P_i, H] = \nu\, K_i \tag{A.166}$$

for five real constants $\alpha, \beta, \gamma, \mu, \nu$.
Rests on Equations (A.160) *to* (A.163).

Proof. *Derives Lemma A.56.*

Each unknown bracket is an $\mathfrak{so}(3)$-covariant expression in the generators, so it may be expanded on the covariant objects available: the scalar H, the vectors J_k, K_k, P_k, and nothing else. Under (H2) each generator carries a definite pair of signs, and a bracket carries the product of the signs of its two entries:

	J_i	K_i	P_i	H
parity Π	$+$	$-$	$-$	$+$
time reversal Θ	$+$	$-$	$+$	$-$

The four cases are then immediate.

The bracket $[K_i, K_j]$. It is antisymmetric in ij, hence a vector, hence of the form $\epsilon_{ijk} V_k$ with $V \in \{J, K, P\}$. Its Π-sign is $(-)(-) = +$, which excludes K and P; only J survives. Its Θ-sign is $(-)(-) = +$, and J is Θ-even, so no further restriction arises. This gives the

first of Equation (A.164). The bracket $[P_i, P_j]$ is identical in structure, with Π-sign + and Θ-sign $(+)(+) = +$, giving the second.

The bracket $[K_i, P_j]$. As a rank-two tensor it decomposes into a trace, an antisymmetric part and a symmetric traceless part, of spins 0, 1 and 2. There is no spin-2 generator in Equation (A.160), so the symmetric traceless part vanishes identically. The trace part is $\gamma\,\delta_{ij}H$ and the antisymmetric part is $\delta\,\epsilon_{ijk}V_k$. The Π-sign is $(-)(-) = +$, which leaves H and J; the Θ-sign is $(-)(+) = -$, and H is Θ-odd while J is Θ-even, so the J term is forbidden and $\delta = 0$. This is the first of Equation (A.165), and it is worth pausing on: the reason a kinematical algebra cannot have $[K_i, P_j] \propto \epsilon_{ijk}J_k$ is time reversal alone.

The brackets with H. $[K_i, H]$ is a vector of Π-sign $(-)(+) = -$, excluding J, and of Θ-sign $(-)(-) = +$, excluding K (which is Θ-odd); only P survives. $[P_i, H]$ is a vector of Π-sign $-$, again excluding J, and of Θ-sign $(+)(-) = -$, excluding P; only K survives. These are the second of Equation (A.165) and Equation (A.166). □

A.10.3 Step two: the Jacobi identity eliminates two of the five

Lemma A.57 (The Jacobi constraints)**.** *The brackets of Lemma A.56 satisfy the Jacobi identity if and only if*

$$\boxed{\alpha = -\gamma\mu\,, \qquad \beta = \gamma\nu}\,. \tag{A.167}$$

A kinematical algebra is therefore determined by the three constants (γ, μ, ν) *alone.*

Rests on Lemma A.56 and Equation (A.161).

Proof. *Derives Lemma A.57.*

Every triple containing a J is satisfied automatically, because Equations (A.164) to (A.166) were constructed to be $\mathfrak{so}(3)$-covariant. Of the remaining triples, those built from three copies of one vector generator are also automatic: for (K_i, K_j, K_l) the cyclic sum is

$$\alpha\,(\epsilon_{ijk}\epsilon_{klm} + \epsilon_{jlk}\epsilon_{kim} + \epsilon_{lik}\epsilon_{kjm})\,K_m = 0$$

by three applications of $\epsilon_{kab}\epsilon_{kcd} = \delta_{ac}\delta_{bd} - \delta_{ad}\delta_{bc}$, the six resulting Kronecker terms cancelling in pairs; the triple (P_i, P_j, P_l) is the same computation with $\alpha \to \beta$ and $K \to P$. Four independent triples remain.

The triple (K_i, K_j, P_l). Using $[J_k, P_l] = \epsilon_{klm}P_m$ and then the same ϵ-identity,

$$\begin{aligned}
[[K_i, K_j]\,, P_l] &= \alpha\,\epsilon_{ijk}\epsilon_{klm}P_m = \alpha\,(\delta_{il}P_j - \delta_{jl}P_i)\ , \\
[[K_j, P_l]\,, K_i] &= \gamma\,\delta_{jl}\,[H, K_i] = -\gamma\mu\,\delta_{jl}P_i\,, \\
[[P_l, K_i]\,, K_j] &= -\gamma\,\delta_{il}\,[H, K_j] = +\gamma\mu\,\delta_{il}P_j\,.
\end{aligned}$$

The sum is $(\alpha + \gamma\mu)\,(\delta_{il}P_j - \delta_{jl}P_i)$, and the two tensors $\delta_{il}P_j$ and $\delta_{jl}P_i$ are linearly independent, so $\alpha = -\gamma\mu$.

The triple (P_i, P_j, K_l). Identically,

$$\begin{aligned}
[[P_i, P_j]\,, K_l] &= \beta\,(\delta_{il}K_j - \delta_{jl}K_i)\ , \\
[[P_j, K_l]\,, P_i] &= -\gamma\,\delta_{jl}\,[H, P_i] = +\gamma\nu\,\delta_{jl}K_i\,, \\
[[K_l, P_i]\,, P_j] &= \gamma\,\delta_{il}\,[H, P_j] = -\gamma\nu\,\delta_{il}K_j\,,
\end{aligned}$$

whose sum is $(\beta - \gamma\nu)\,(\delta_{il}K_j - \delta_{jl}K_i)$, so $\beta = \gamma\nu$.

The triples containing H. For (K_i, K_j, H) the three terms are $\alpha\epsilon_{ijk}\,[J_k, H] = 0$ by Equation (A.161), together with $\mu\,[P_j, K_i] - \mu\,[P_i, K_j] = \mu\gamma\delta_{ij}H - \mu\gamma\delta_{ij}H = 0$; the identity holds with no condition. The triple (P_i, P_j, H) is the same with $\mu \to \nu$. For (K_i, P_j, H),

$$\gamma\,\delta_{ij}\,[H, H] + \nu\,[K_j, K_i] - \mu\,[P_i, P_j] = -\,(\alpha\nu + \beta\mu)\,\epsilon_{ijk}J_k\,,$$

so $\alpha\nu + \beta\mu = 0$ — which is implied by Equation (A.167), since $-\gamma\mu\nu + \gamma\nu\mu = 0$. No further condition arises, and the two constraints of Equation (A.167) are the whole content of the Jacobi identity. □

A.10.4 Step three: rescaling reduces the constants to their signs

Two kinematical algebras describe the same kinematics if they differ only by a choice of units for boosts, lengths and times. The corresponding transformations

$$K_i \longmapsto \lambda K_i\,, \qquad P_i \longmapsto \rho P_i\,, \qquad H \longmapsto \sigma H\,, \qquad \lambda\rho\sigma \neq 0\,, \tag{A.168}$$

leave Equation (A.161) and both automorphisms of (H2) intact, and they exhaust that freedom: J is fixed by the normalization of Equation (A.161).

Lemma A.58 (Normal form)**.** *Under Equation* (A.168) *the three constants transform as*

$$\gamma \longmapsto \frac{\lambda\rho}{\sigma}\,\gamma\,, \qquad \mu \longmapsto \frac{\lambda\sigma}{\rho}\,\mu\,, \qquad \nu \longmapsto \frac{\rho\sigma}{\lambda}\,\nu\,. \tag{A.169}$$

Consequently:

1. *which of γ, μ, ν vanish is invariant;*
2. *every nonvanishing constant can be scaled to ± 1, and all of them simultaneously; and*
3. *the signs of the surviving constants are determined only up to a* simultaneous *flip of all three, so that the invariants are the pairwise products* $\operatorname{sgn}(\gamma\mu)$, $\operatorname{sgn}(\gamma\nu)$ *and* $\operatorname{sgn}(\mu\nu)$, *of which two are independent.*

Rests on Lemma A.57 and Equation (A.168)*.*

Proof. *Derives Lemma A.58.*

Equation (A.169) follows by substituting Equation (A.168) into Equations (A.165) and (A.166) and re-expressing the right-hand sides in the new generators; claim (1) is immediate from it, each constant being multiplied by a nonzero number. For (2), take logarithms of the absolute values: the exponents of (λ, ρ, σ) form the matrix

$$\begin{pmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \end{pmatrix}\,, \qquad \text{of determinant } -4 \neq 0\,,$$

so the system can be solved for $\log|\lambda|, \log|\rho|, \log|\sigma|$ whatever shifts are required; when only some constants are nonzero the corresponding subsystem is solvable a fortiori, the matrix having full rank on any subset of its rows. For (3), write $s_\lambda, s_\rho, s_\sigma$ for the signs of the scale factors. Each of the three coefficients in Equation (A.169) has the *same* sign, namely $s = s_\lambda s_\rho s_\sigma$, because each is a product of all three scale factors with one of them inverted and inversion does not change a sign. So all three constants are multiplied by one common $s \in \{\pm 1\}$, which may be chosen freely. The pairwise products are unchanged by that flip, and their own product is $(\gamma\mu\nu)^2 > 0$, so only two of them are independent. □

A.10.5 Step four: the enumeration

By Lemma A.58(1) the classification splits into the eight cases labelled by which of (γ, μ, ν) vanish; within each case, Lemma A.58(2) and (3) reduce the surviving constants to a sign pattern modulo one simultaneous flip. Writing $*$ for a nonzero constant, the count is as follows.

$(0,0,0)$ — **one algebra.** Every bracket among K, P, H vanishes, and with them α and β. This is the *static* algebra.

$(*,0,0)$ — **one algebra.** $\alpha = \beta = 0$, and the single surviving constant has a sign that the flip removes. Here $[K_i, P_j] = \delta_{ij} H$ with abelian boosts: the *Carroll* algebra.

$(0,*,0)$ — **one algebra.** Likewise one algebra, $[K_i, H] = P_i$ with every other new bracket zero: the *Galilei* algebra. Note that $\gamma = 0$ here, so $[K_i, P_j] = 0$ — the mass of Equation (25.15) is *not* in this algebra, which is exactly why it can only enter as a central extension.

$(0,0,*)$ — **one algebra.** The *para-Galilei* algebra, $[P_i, H] = K_i$.

$(0,*,*)$ — **two algebras.** $\alpha = \beta = 0$ and the invariant is $\operatorname{sgn}(\mu\nu)$: the *Newton–Hooke* algebras NH_+ and NH_-, whose difference is exponential against oscillatory behaviour of the free motion.

$(*,0,*)$ — **two algebras.** $\alpha = 0$ but $\beta = \gamma\nu \neq 0$, and the invariant is $\operatorname{sgn}\beta$. The set $\{J_i, P_i\}$ closes into $\mathfrak{so}(3,1)$ for $\beta < 0$ and into $\mathfrak{so}(4)$ for $\beta > 0$, while $\{K_i, H\}$ is an abelian ideal in both cases. These are *para-Poincaré* and the *inhomogeneous* $\mathrm{SO}(4)$. The boosts are abelian here, so (H3) is satisfied and neither is excluded.

$(*,*,0)$ — **one algebra.** $\beta = 0$ but $\alpha = -\gamma\mu \neq 0$, and the invariant is $\operatorname{sgn}\alpha$; now $\{J_i, K_i\}$ closes into $\mathfrak{so}(3,1)$ for $\alpha < 0$ and into $\mathfrak{so}(4)$ for $\alpha > 0$. In the second case the boosts generate a compact subgroup, so (H3) excludes it and only $\alpha < 0$ survives: the *Poincaré* algebra $\mathfrak{t}(4) \mathbin{\bar{\oplus}} \mathfrak{so}(3,1)$ of Equation (14.99).

$(*,*,*)$ — **two algebras.** Two independent sign invariants give four candidates; (H3) again forces $\alpha = -\gamma\mu < 0$, halving them to two, distinguished by $\operatorname{sgn}(\gamma\nu)$. These are the two simple algebras of Equations (14.109) and (14.110), $\mathfrak{so}(4,1)$ and $\mathfrak{so}(3,2)$: *de Sitter* and *anti-de Sitter*.

Adding the cases gives

$$1 + 1 + 1 + 1 + 2 + 2 + 1 + 2 = 11\,, \tag{A.170}$$

which proves Theorem 14.90. The assignment of the traditional names to the classes is Bacry and Lévy-Leblond's [1364]; what is established above is the count, the parametrization, and the structure of each class.

A.10.6 Two corollaries the enumeration makes visible

Corollary A.59 (The eight vanishing patterns are the contraction cube)**.** *Each of the three constants can be sent to zero by an Inönü–Wigner contraction, in the sense of Definition 14.83, that leaves the other two fixed:*

$$\nu \to 0: \quad P_i \mapsto \varepsilon P_i\,,\ H \mapsto \varepsilon H \qquad \textit{(the flat-space limit)}\,, \tag{A.171}$$

$$\gamma \to 0: \quad K_i \mapsto \varepsilon K_i\,,\ P_i \mapsto \varepsilon P_i \qquad \textit{(the nonrelativistic limit } c \to \infty)\,, \tag{A.172}$$

$$\mu \to 0: \quad K_i \mapsto \varepsilon K_i\,,\ H \mapsto \varepsilon H \qquad \textit{(the ultrarelativistic limit } c \to 0)\,, \tag{A.173}$$

in each case with $\varepsilon \to 0$. The three contractions commute, so from an algebra with all three constants nonzero they generate exactly the eight vanishing patterns of Appendix A.10.5. The kinematical cube *is therefore the lattice of subsets of $\{\gamma, \mu, \nu\}$: its eight vertices are the eight* cases*, not eight of the eleven algebras. Three of the eleven — one de Sitter, one Newton–Hooke, and the inhomogeneous* $\mathrm{SO}(4)$ *— are opposite-sign partners of a vertex rather than vertices of their own.*

Rests on Lemma A.58 and Definition 14.83.

Proof. *Derives Corollary A.59.*

Apply Equation (A.169) to each substitution. For Equation (A.171), $\lambda = 1$ and $\rho = \sigma = \varepsilon$ give $\gamma \mapsto \gamma$, $\mu \mapsto \mu$, $\nu \mapsto \varepsilon^2\nu$. For Equation (A.172), $\lambda = \rho = \varepsilon$ and $\sigma = 1$ give $\gamma \mapsto \varepsilon^2\gamma$

with μ and ν fixed. For Equation (A.173), $\lambda = \sigma = \varepsilon$ and $\rho = 1$ give $\mu \mapsto \varepsilon^2\mu$ with γ and ν fixed. In each case no structure constant diverges, so the limit exists and is a contraction. Commutativity is clear, the three substitutions acting as commuting diagonal matrices on (λ, ρ, σ). □

Corollary A.60 (The "para" algebras are exchanges, not new algebras)**.** *The linear map* Σ *exchanging boosts and space translations,*

$$\Sigma: \quad K_i \longmapsto P_i\,, \qquad P_i \longmapsto K_i\,, \qquad J_i \longmapsto J_i\,, \qquad H \longmapsto H\,, \tag{A.174}$$

carries the algebra with constants (γ, μ, ν) *to the algebra with constants* $(-\gamma, \nu, \mu)$*, and so exchanges the cases* $(*, *, 0)$ *and* $(*, 0, *)$*, and the cases* $(0, *, 0)$ *and* $(0, 0, *)$*. It therefore identifies,* as abstract Lie algebras*,*

$$\textit{Poincaré} \cong \textit{para-Poincaré}\,, \qquad \textit{Galilei} \cong \textit{para-Galilei}\,, \tag{A.175}$$

and identifies the inhomogeneous SO(4) *with the compact-boost algebra that (H3) excluded from the case* $(*, *, 0)$*.*

Rests on Lemmas A.56 and A.57.

Proof. *Derives Corollary A.60.*

Under Σ the bracket $[K_i, P_j] = \gamma\delta_{ij}H$ becomes $[P_i, K_j] = \gamma\delta_{ij}H$, that is $[K_j, P_i] = -\gamma\delta_{ij}H$, so $\gamma \mapsto -\gamma$; and $[K_i, H] = \mu P_i$ becomes $[P_i, H] = \mu K_i$, so $\nu \mapsto \mu$ and, symmetrically, $\mu \mapsto \nu$. The constrained constants follow from Equation (A.167): $\alpha' = -\gamma'\mu' = \gamma\nu = \beta$ and $\beta' = \gamma'\nu' = -\gamma\mu = \alpha$, so Σ exchanges the two quadratic constants as well. Applying this to Poincaré, which is $(\gamma, \mu, 0)$ with $\alpha < 0$, gives $(-\gamma, 0, \mu)$ with $\beta' = \alpha < 0$ — the case $(*, 0, *)$ with negative β, which is para-Poincaré. The same substitution applied to $(0, \mu, 0)$ gives $(0, 0, \mu)$, and applied to the excluded $\alpha > 0$ member of $(*, *, 0)$ gives the $\beta > 0$ member of $(*, 0, *)$, the inhomogeneous SO(4). □

Remark A.61 (Why the count is eleven, and what it is a count of). Corollary A.60 is the reason the eleven of Theorem 14.90 must be counted as eleven *kinematics* and not as eleven isomorphism classes of Lie algebras. The map Σ is an isomorphism of Lie algebras, but it is not an equivalence of kinematics: it fails to commute with the time reversal Equation (A.163), since it sends a Θ-odd generator to a Θ-even one, and it exchanges the physical roles of a boost and a translation. Two of the eleven are therefore abstractly the same algebra as another member of the list while describing a different physics, and Table 14.1 lists them separately for that reason. A statement of the theorem phrased as "eleven algebras up to isomorphism" is wrong, and the error is an easy one to make.

Appendix A.10 discharges the proof obligation of Theorem 14.90. Its yield for the rest of the book is Equation (A.167) and the three contractions of Corollary A.59. The second of them, Equation (A.172), is the $c \to \infty$ limit carrying the Poincaré algebra to the Galilei algebra, and the fact that it sets $\gamma = 0$ — so that $[K_i, P_j]$ vanishes identically in the contracted algebra — is precisely what leaves room for the mass to reappear there as the central charge of Proposition 25.14.

A.11 Invariant Bilinear Forms on an Expanded Algebra

This appendix proves Theorem 15.54 (Chapter 15): on $\mathfrak{g}_A = A \otimes \mathfrak{g}$ every invariant symmetric *bilinear* form has the factorised shape Equation (15.61), so for $k = 2$ the family constructed there is already the whole of it. This is the case the book uses — it is what makes Theorem 15.58 a statement about all invariant pairings and not merely about the ones written down. It is also the only case in which the statement is true: Proposition 15.55

exhibits an invariant symmetric 4-linear form that is not factorised, and Remark A.66 at the end says precisely which step of the argument below fails for $k \geq 3$ and why no repair is possible.

The hypothesis on $\mathfrak{g}$ needs care, and the care is not academic: the Lorentz algebra of the observed 3+1 dimensions is exactly the case where the naive hypothesis fails.

Definition A.62 (Absolutely simple)**.** A real Lie algebra $\mathfrak{g}$ is *absolutely simple* if its complexification $\mathfrak{g} \otimes_{\mathbb{R}} \mathbb{C}$ is a simple complex Lie algebra. Every absolutely simple algebra is simple; the converse fails, the standard counterexample being a complex simple algebra regarded as a real one.

Lemma A.63 (The invariant bilinear forms of an absolutely simple algebra form a line)**.** *Let $\mathfrak{g}$ be absolutely simple with Killing form $\kappa_{\mathfrak{g}}$. Then every invariant bilinear form on $\mathfrak{g}$ is a real multiple of $\kappa_{\mathfrak{g}}$; in particular the space of such forms is one-dimensional and every one of them is symmetric.*

Rests on Definition A.62.

Derivation. *Derives Lemma A.63.*

Let B be an invariant bilinear form, that is

$$B\left([Z,X],Y\right) + B\left(X,[Z,Y]\right) = 0 \qquad \text{for all } X,Y,Z \in \mathfrak{g}\,. \tag{A.176}$$

Define $\beta : \mathfrak{g} \to \mathfrak{g}^*$ by $\beta(X) = B(X,\cdot)$. Equation (A.176) says exactly that β intertwines the adjoint representation on $\mathfrak{g}$ with the coadjoint representation on $\mathfrak{g}^*$. Since $\mathfrak{g}$ is simple it is semisimple, so $\kappa_{\mathfrak{g}}$ is nondegenerate by Cartan's criterion and the induced map $\kappa_{\mathfrak{g}}^{\flat} : \mathfrak{g} \to \mathfrak{g}^*$ is an isomorphism of representations. Hence

$$T := \left(\kappa_{\mathfrak{g}}^{\flat}\right)^{-1} \circ \beta \in \mathrm{End}_{\mathfrak{g}}(\mathfrak{g})\,,$$

the algebra of endomorphisms commuting with the adjoint action, and $B = \kappa_{\mathfrak{g}}(T\,\cdot,\cdot)$.

It remains to show $\mathrm{End}_{\mathfrak{g}}(\mathfrak{g}) = \mathbb{R}$. Extending scalars,

$$\mathrm{End}_{\mathfrak{g}}(\mathfrak{g}) \otimes_{\mathbb{R}} \mathbb{C} \cong \mathrm{End}_{\mathfrak{g}_{\mathbb{C}}}\left(\mathfrak{g}_{\mathbb{C}}\right)\,, \qquad \mathfrak{g}_{\mathbb{C}} := \mathfrak{g} \otimes_{\mathbb{R}} \mathbb{C}\,,$$

since forming the commutant of a set of operators commutes with extension of scalars. By hypothesis $\mathfrak{g}_{\mathbb{C}}$ is simple, so its adjoint representation is irreducible over $\mathbb{C}$, and Schur's lemma over an algebraically closed field gives $\mathrm{End}_{\mathfrak{g}_{\mathbb{C}}}(\mathfrak{g}_{\mathbb{C}}) = \mathbb{C}$. Hence $\dim_{\mathbb{R}} \mathrm{End}_{\mathfrak{g}}(\mathfrak{g}) = 1$, so $T = \lambda\,\mathrm{id}$ and $B = \lambda\,\kappa_{\mathfrak{g}}$, which is symmetric. □

Theorem A.64 (Completeness of the factorised bilinear forms)**.** *Let $\mathfrak{g}$ be absolutely simple and let A be a finite-dimensional commutative associative unital $\mathbb{R}$-algebra. Then for every invariant symmetric bilinear form Ω on $\mathfrak{g}_A$ there is a* unique *linear functional $\phi : A \to \mathbb{R}$ with*

$$\boxed{\Omega\left(a \otimes X,\ b \otimes Y\right) = \phi(ab)\,\kappa_{\mathfrak{g}}(X,Y)} \qquad \textit{for all } a,b \in A,\ X,Y \in \mathfrak{g}\,. \tag{A.177}$$

Equivalently $\Omega = \langle\cdot,\cdot\rangle_{\kappa_{\mathfrak{g}},\phi}$ in the notation of Equation (15.61), and $\phi \mapsto \langle\cdot,\cdot\rangle_{\kappa_{\mathfrak{g}},\phi}$ is a linear isomorphism from A^* onto the space of invariant symmetric bilinear forms on $\mathfrak{g}_A$, which therefore has dimension $\dim A$.

Rests on Definition A.62, Lemmas 15.50 and A.63, Theorem 15.51, and Equation (15.2).

Derivation. *Derives Theorem A.64.*

Step 1: for fixed a,b the form is a multiple of the Killing form. Fix $a,b \in A$ and define the bilinear form

$$\omega_{a,b}(X,Y) := \Omega\left(a \otimes X,\ b \otimes Y\right) \tag{A.178}$$

on $\mathfrak{g}$. Invariance of Ω, applied with the element $1_A \otimes Z$ and using $[1_A \otimes Z, a \otimes X] = a \otimes [Z, X]$ from Equation (15.2), gives

$$\Omega\left(a \otimes [Z, X]\,,\ b \otimes Y\right) + \Omega\left(a \otimes X,\ b \otimes [Z, Y]\right) = 0\,,$$

that is, $\omega_{a,b}$ satisfies Equation (A.176). By Lemma A.63 there is a unique real number $\Psi(a, b)$ with

$$\omega_{a,b} = \Psi(a, b)\,\kappa_{\mathfrak{g}}\,. \tag{A.179}$$

Since Ω is bilinear in its two arguments and $\kappa_{\mathfrak{g}} \neq 0$, the function $\Psi : A \times A \to \mathbb{R}$ is bilinear; since Ω and $\kappa_{\mathfrak{g}}$ are symmetric, Ψ is symmetric.

Step 2: Ψ is a 2-trace. Now use invariance of Ω with a *general* element $c \otimes Z$. By Equation (15.2), $[c \otimes Z, a \otimes X] = (ca) \otimes [Z, X]$, so

$$\Omega\left((ca) \otimes [Z, X]\,,\ b \otimes Y\right) + \Omega\left(a \otimes X,\ (cb) \otimes [Z, Y]\right) = 0\,,$$

which by Equation (A.179) reads

$$\Psi(ca, b)\,\kappa_{\mathfrak{g}}\left([Z, X]\,, Y\right) + \Psi(a, cb)\,\kappa_{\mathfrak{g}}\left(X, [Z, Y]\right) = 0\,. \tag{A.180}$$

The Killing form is itself invariant, so $\kappa_{\mathfrak{g}}(X, [Z, Y]) = -\kappa_{\mathfrak{g}}([Z, X]\,, Y)$, and Equation (A.180) collapses to

$$\Big(\Psi(ca, b) - \Psi(a, cb)\Big)\,\kappa_{\mathfrak{g}}\left([Z, X]\,, Y\right) = 0 \qquad \text{for all } X, Y, Z \in \mathfrak{g}\,. \tag{A.181}$$

The Killing factor cannot vanish identically: a simple algebra is perfect, so the elements $[Z, X]$ span $\mathfrak{g}$, and $\kappa_{\mathfrak{g}}$ is nondegenerate, so some choice of X, Y, Z makes $\kappa_{\mathfrak{g}}([Z, X]\,, Y) \neq 0$. Hence

$$\Psi(ca, b) = \Psi(a, cb) \qquad \text{for all } a, b, c \in A\,, \tag{A.182}$$

which is Equation (15.58) at $k = 2$: Ψ is a 2-trace on A.

Step 3: a 2-trace is a linear functional of the product. By Lemma 15.50 with $k = 2$ there is a unique linear $\phi : A \to \mathbb{R}$, namely $\phi(x) = \Psi(x, 1_A)$, with $\Psi(a, b) = \phi(ab)$. Substituting into Equation (A.179) gives Equation (A.177).

Step 4: uniqueness and bijectivity. If ϕ and ϕ' both satisfy Equation (A.177) then $(\phi(ab) - \phi'(ab))\,\kappa_{\mathfrak{g}}(X, Y) = 0$; choosing X, Y with $\kappa_{\mathfrak{g}}(X, Y) \neq 0$ and $b = 1_A$ gives $\phi = \phi'$. Conversely every $\phi \in A^*$ produces an invariant symmetric bilinear form by the first part of Theorem 15.51, and the assignment is visibly linear in ϕ, while Equation (A.177) says it is onto. The two spaces are therefore isomorphic. □

Remark A.65 (The hypothesis is sharp, and 3+1 is where it bites). Lemma A.63 needs $\mathfrak{g}$ absolutely simple, not merely simple, and the difference is not a technicality in this book. The Lorentz algebra of four-dimensional spacetime, $\mathfrak{so}(3, 1)$, is simple as a real Lie algebra but *not* absolutely simple: its complexification is $\mathfrak{sl}_2(\mathbb{C}) \oplus \mathfrak{sl}_2(\mathbb{C})$, which is not simple — this is the self-dual/anti-self-dual splitting. Accordingly it carries a two-dimensional space of invariant symmetric bilinear forms, spanned by

$$\langle J_{\mu\nu}, J_{\rho\sigma}\rangle_1 = \eta_{\mu\rho}\eta_{\nu\sigma} - \eta_{\mu\sigma}\eta_{\nu\rho}\,, \qquad \langle J_{\mu\nu}, J_{\rho\sigma}\rangle_2 = \epsilon_{\mu\nu\rho\sigma}\,, \tag{A.183}$$

the first proportional to the Killing form and the second available only because $D = 4$ supplies a rank-four invariant alternating symbol. Both are invariant, and the two Chern–Weil 4-forms they generate are the two topological densities of four-dimensional gravity: the Killing form gives $R^{\mu\nu} \wedge R_{\mu\nu}$, the Pontryagin density, and $\epsilon_{\mu\nu\rho\sigma}$ gives $\epsilon_{\mu\nu\rho\sigma} R^{\mu\nu} \wedge R^{\rho\sigma}$, the Euler density of the Gauss–Bonnet term. That a four-dimensional gauge theory of the Lorentz group admits two such terms where a general dimension admits one is exactly this

failure of absolute simplicity. For such a $\mathfrak{g}$ the conclusion of Theorem A.64 must be replaced by a sum over a basis of the invariant forms of $\mathfrak{g}$, one functional ϕ for each. The algebras to which Theorem 15.29 attaches the three signs of the cosmological constant — $\mathfrak{so}(3,2)$ and $\mathfrak{so}(4,1)$ in 3+1 — *are* absolutely simple, their complexifications being $\mathfrak{sp}(4,\mathbb{C})$ and $\mathfrak{so}(5,\mathbb{C})$, so Theorem A.64 applies to them verbatim.

Remark A.66 (Why the argument stops at $k=2$, and why that is not a gap). It is worth being explicit about where the proof above uses $k=2$, because the manuscript this chapter follows [2434] states the converse for every k, and that statement is *false*.

Step 1 generalises without difficulty: evaluating invariance on $1_A \otimes Z$ shows, for any k, that $(X_1,\dots,X_k) \mapsto \Omega(a_1 \otimes X_1,\dots,a_k \otimes X_k)$ is an invariant k-linear form on $\mathfrak{g}$. Step 2 does not. It works here because the invariance relation for a bilinear form has exactly *two* terms, and the invariance of $\kappa_{\mathfrak{g}}$ itself collapses them into the single factor Equation (A.181); for $k \geq 3$ the same manipulation leaves a sum of $k-1$ terms with no reason to vanish separately. That is not a defect of the argument: by Proposition 15.55 there is an invariant symmetric 4-linear form on $\mathfrak{so}(2,1)\langle 1\rangle = \mathfrak{iso}(2,1)$ — the square of the standard pairing $\langle J_a, P_b\rangle = \eta_{ab}$ — which is not factorised at all, so no argument could have succeeded.

The structural reason is Theorem 15.53: the factorised forms are exactly the *balanced* ones, and a product of two invariant bilinear forms is not balanced, because $\phi(a_1a_2)\phi(a_3a_4)$ is not a function of $a_1a_2a_3a_4$. At $k=2$ no product is available, since $\mathrm{Inv}^1(\mathfrak{g}) = (\mathfrak{g}^*)^{\mathfrak{g}}$ vanishes for perfect $\mathfrak{g}$ — which is exactly what Theorem A.64 exploits without saying so, and is the whole of the difference between the two cases.

Appendix A.11 discharges the proof obligation of Theorem 15.54. Its yield for the rest of the book is that Theorem 15.58 classifies *all* invariant pairings on a resonant subalgebra and not merely the constructed ones, which is what makes the statement that the Galilei algebra admits none, while the Bargmann algebra does, a theorem rather than a failure to find one. Its other yield is negative and equally worth having: together with Proposition 15.55 it fixes the exact reach of the factorisation, which is $k=2$ and no further.

A.12 Darboux's Theorem: Local Canonical Coordinates

This appendix proves the geometric statement announced in Remark 5.116 of Chapter 5: a closed non-degenerate two-form on a manifold can be brought, in a neighbourhood of any point, to the constant normal form $\sum_i \mathrm{d}q^i \wedge \mathrm{d}p_i$. Symplectic geometry therefore has no local invariants whatever — no analogue of the curvature that distinguishes one Riemannian metric from another near a point — so that the absence of a signature established pointwise in Proposition 5.115 propagates to the manifold.

What is assumed as already available is exactly the differential-form machinery of Chapter 13: k-forms (Definition 13.98), the wedge product (Definition 13.102), the exterior derivative with its nilpotency and Leibniz rule (Definition 13.103 and Equations (13.245) and (13.246)), closed and exact forms (Definitions 13.105 and 13.106), the Lie derivative (Equation (13.267) and Proposition 13.127) and Cartan's magic formula (Proposition 13.128); from analysis, the mean value theorem (Theorem 7.35), the Cauchy criterion (Theorem 7.8), uniform continuity on a compact set (Theorem 7.25) and the Picard–Lindelöf theorem with its explicit contraction estimate (Theorem 9.8 and Corollary 9.9). Everything else — the pullback of a k-form, the smoothness of the flow of a time-dependent vector field, and the local converse of the Poincaré lemma — is built here, because the treatise does not carry it. The pointwise input is Proposition 5.115, and the argument that converts it into a statement about a varying form is the homotopy argument known as *Moser's trick*.

A.12.1 Statement

Definition A.67 (Symplectic manifold)**.** A *symplectic manifold* is a pair (M, ω) with M a smooth manifold (Definition 13.48) and ω a smooth two-form on M (Definition 13.98) which is

1. *closed*, $\mathrm{d}\omega = 0$ (Definition 13.105); and
2. *non-degenerate* at every point: for each $P \in M$ the matrix $\omega_{ij}(P)$ of components in any chart is invertible, equivalently $\omega_P(u, v) = 0$ for all $v \in T_M(P)$ forces $u = 0$.

Rests on Definitions 13.48, 13.98 and 13.105.

Theorem A.68 (Darboux)**.** *Let (M, ω) be a symplectic manifold and $P \in M$. Then $\dim M = 2m$ is even, and there is a chart $(u^1, \dots, u^{2m})$ of M defined on a neighbourhood of P, with $u(P) = 0$, in which*

$$\omega = \sum_{i=1}^{m} \mathrm{d}q^i \wedge \mathrm{d}p_i \,, \qquad q^i = u^i \,, \quad p_i = u^{m+i} \,. \tag{A.184}$$

Equivalently, the components of ω in this chart are the constant *block array Equation (5.144).*

Rests on Definition A.67, Proposition 5.115, and Equation (5.144).

Corollary A.69 (No local invariants)**.** *Any two symplectic manifolds of the same dimension are locally symplectomorphic: given $P \in M$ and $P' \in M'$ there are neighbourhoods $U \ni P$, $U' \ni P'$ and a diffeomorphism $f : U \longrightarrow U'$ with $f^*\omega' = \omega$. In particular no function of ω and its derivatives at a point can be a symplectic invariant.*

Rests on Theorem A.68.

Proof. *Derives Corollary A.69.*

Let u and u' be canonical charts around P and P' furnished by Theorem A.68, both with image containing a ball $B \subset \mathbb{R}^{2m}$ about the origin. Put $f = (u')^{-1} \circ u$ on $U = u^{-1}(B)$. Both ω and ω' read as the same constant form Equation (A.184) in their own charts, and the pullback of a form is computed chartwise (Definition A.70 below), so $f^*\omega' = \omega$. Any local invariant would have to take the same value at P and P', and ω, ω' were arbitrary. □

The proof of Theorem A.68 occupies the rest of this section. Appendix A.12.2 builds the pullback of a k-form and its two structural properties; Appendix A.12.3 proves that the flow of a time-dependent vector field exists and is smooth, and differentiates a pullback along it; Appendix A.12.4 proves the local converse of the Poincaré lemma on a star-shaped set; and Appendix A.12.5 assembles the three into Moser's homotopy argument.

Throughout, $|\cdot|$ is the Euclidean length on $\mathbb{R}^n$ (Equation (6.12)) and, for a real $n \times n$ matrix A,

$$\|A\| = \left(\sum_{i,j} (A_{ij})^2 \right)^{1/2} , \qquad \text{so that} \qquad |Av| \le \|A\|\,|v| \,, \quad \|AB\| \le \|A\|\,\|B\| \,, \tag{A.185}$$

the two inequalities following from the Cauchy–Schwarz inequality of Chapter 5 applied row by row.

A.12.2 The pullback of a k-form

Chapter 13 defines the pullback of a covector; the extension to a k-form is forced by the requirement that it act on k pushed-forward vectors.

Definition A.70 (Pullback)**.** Let $f : U \longrightarrow V$ be a smooth map between open subsets of $\mathbb{R}^m$ and $\mathbb{R}^n$, with components f^j, and let α be a k-form on V with components $\alpha_{j_1 \dots j_k}$. The *pullback* $f^*\alpha$ is the k-form on U with components

$$(f^*\alpha)_{i_1 \dots i_k}(x) = \alpha_{j_1 \dots j_k}(f(x))\, \partial_{i_1} f^{j_1}(x) \cdots \partial_{i_k} f^{j_k}(x) \,. \tag{A.186}$$

For $k = 1$ this is the covector pullback of Section 13.5.2, and for $k = 0$ it is $f^*g = g \circ f$. The right-hand side is totally antisymmetric in $i_1, \dots, i_k$ because α is antisymmetric in $j_1, \dots, j_k$ and the two index groups are contracted in order, so $f^*\alpha$ is again a k-form (Definition 13.98).

Rests on Definitions 7.65, 13.83 and 13.98.

Lemma A.71 (The pullback is an algebra map commuting with d)**.** *With f as above, and α, β forms on V of degrees k and l:*

$$f^*(\alpha \wedge \beta) = (f^*\alpha) \wedge (f^*\beta) \,, \tag{A.187}$$

$$f^*(\mathrm{d}\alpha) = \mathrm{d}(f^*\alpha) \,, \tag{A.188}$$

$$(g \circ f)^* \alpha = f^*(g^*\alpha) \,, \tag{A.189}$$

the last for a further smooth g defined on the target of the form.

Rests on Definition A.70 and Equations (13.245) and (13.246).

Proof. *Derives Lemma A.71.*

Composition. Equation (A.189) is the chain rule Equation (7.52) inserted k times into Equation (A.186): $\partial_i (g \circ f)^j = \partial_l g^j(f(x))\, \partial_i f^l(x)$.

Scalars and their differentials. For $k = 0$, $f^*g = g \circ f$ and $\partial_i(g \circ f) = \partial_j g(f(x))\, \partial_i f^j$, which is Equation (A.186) for the one-form $\mathrm{d}g$; hence

$$f^*(\mathrm{d}g) = \mathrm{d}(g \circ f) \,. \tag{A.190}$$

Wedge. It suffices to check Equation (A.187) on the components. Both sides are multilinear in α and β, so it is enough to take $\alpha = \mathrm{d}y^{a_1} \wedge \cdots \wedge \mathrm{d}y^{a_k}$ and $\beta = \mathrm{d}y^{b_1} \wedge \cdots \wedge \mathrm{d}y^{b_l}$ with the y the coordinates on V. By Definition 13.102 and Equation (A.186), the pullback of a wedge of coordinate differentials is the wedge of their pullbacks: for two of them,

$$\begin{aligned} \left(f^*\left(\mathrm{d}y^a \wedge \mathrm{d}y^b\right)\right)_{cd} &= \left(\delta^a{}_j \delta^b{}_k - \delta^a{}_k \delta^b{}_j\right) \partial_c f^j \, \partial_d f^k \\ &= \partial_c f^a \, \partial_d f^b - \partial_d f^a \, \partial_c f^b = \left(\mathrm{d}f^a \wedge \mathrm{d}f^b\right)_{cd} \,, \end{aligned}$$

and the same cancellation of index pairs, term by term over the permutations, gives $f^*(\mathrm{d}y^{a_1} \wedge \cdots \wedge \mathrm{d}y^{a_k}) = \mathrm{d}f^{a_1} \wedge \cdots \wedge \mathrm{d}f^{a_k}$, where $\mathrm{d}f^a = f^*\mathrm{d}y^a$ by Equation (A.190). Equation (A.187) follows because both sides then read $\mathrm{d}f^{a_1} \wedge \cdots \wedge \mathrm{d}f^{a_k} \wedge \mathrm{d}f^{b_1} \wedge \cdots \wedge \mathrm{d}f^{b_l}$.

Exterior derivative. Write $\alpha = \frac{1}{k!} \alpha_{a_1 \dots a_k}\, \mathrm{d}y^{a_1} \wedge \cdots \wedge \mathrm{d}y^{a_k}$. By the two paragraphs just proved,

$$f^*\alpha = \frac{1}{k!} (\alpha_{a_1 \dots a_k} \circ f)\, \mathrm{d}f^{a_1} \wedge \cdots \wedge \mathrm{d}f^{a_k} \,.$$

Applying d and using the Leibniz rule Equation (13.246) repeatedly, every term in which d falls on some $\mathrm{d}f^a$ vanishes by nilpotency Equation (13.245), so

$$\mathrm{d}(f^*\alpha) = \frac{1}{k!}\, \mathrm{d}(\alpha_{a_1 \dots a_k} \circ f) \wedge \mathrm{d}f^{a_1} \wedge \cdots \wedge \mathrm{d}f^{a_k} = f^*(\mathrm{d}\alpha) \,,$$

the last equality by Equation (A.190) applied to the component functions and one more use of the two paragraphs above. □

A.12.3 Flows of time-dependent vector fields

Moser's argument needs a family of diffeomorphisms generated by a vector field that changes with the parameter, and needs it to be smooth in the point as well as in the parameter. Chapter 9 supplies existence and uniqueness (Theorem 9.8); the dependence on the initial point is proved here.

Lemma A.72 (Iterated integral inequality)**.** *Let $u : [0,T] \longrightarrow [0,\infty)$ be continuous and suppose*

$$u(t) \le c + L\int_0^t u(s)\,\mathrm{d}s\,, \qquad 0 \le t \le T\,, \tag{A.191}$$

with constants $c \ge 0$, $L \ge 0$. Then $u(t) \le c\,\mathrm{e}^{Lt}$.

Rests on Theorem 7.43 and Definition 7.53.

Proof. *Derives Lemma A.72.*

Let $S = \max_{[0,T]} u$, finite by Theorem 7.24. We show by induction that for every $N \ge 0$

$$u(t) \le c\sum_{j=0}^{N} \frac{(Lt)^j}{j!} + S\,\frac{(Lt)^{N+1}}{(N+1)!}\,. \tag{A.192}$$

For $N = 0$, Equation (A.191) and $u \le S$ give $u(t) \le c + LSt$. Assuming Equation (A.192) at N and substituting it into Equation (A.191), term-by-term integration gives

$$u(t) \le c + L\int_0^t \left[c\sum_{j=0}^{N}\frac{(Ls)^j}{j!} + S\frac{(Ls)^{N+1}}{(N+1)!}\right]\mathrm{d}s = c\sum_{j=0}^{N+1}\frac{(Lt)^j}{j!} + S\frac{(Lt)^{N+2}}{(N+2)!}\,,$$

which is the statement at $N+1$. The last term of Equation (A.192) tends to 0 as $N \to \infty$, because the exponential series converges (Definition 7.53), and the sum tends to e^{Lt}. □

Lemma A.73 (Differentiation under the integral sign)**.** *Let $K \subset \mathbb{R}^n$ be open, $F : [0,1] \times K \longrightarrow \mathbb{R}$ continuous with $\partial F/\partial x^i$ existing and continuous on $[0,1] \times K$ for each i. Then $G(x) = \int_0^1 F(t,x)\,\mathrm{d}t$ has continuous partial derivatives on K and*

$$\partial_i G(x) = \int_0^1 \partial_i F(t,x)\,\mathrm{d}t\,. \tag{A.193}$$

Rests on Theorems 7.25 and 7.35 and Definition 7.65.

Proof. *Derives Lemma A.73.*

Fix $x \in K$ and $r > 0$ with $\overline{B}_r(x) \subset K$. For $0 < |s| \le r$ the mean value theorem Theorem 7.35 applied to $\lambda \mapsto F(t, x + \lambda \boldsymbol{e}_i)$ gives $\theta = \theta(t,s) \in (0,1)$ with

$$\frac{F(t, x + s\boldsymbol{e}_i) - F(t,x)}{s} = \partial_i F(t, x + \theta s\,\boldsymbol{e}_i)\,.$$

The set $[0,1] \times \overline{B}_r(x)$ is compact, so $\partial_i F$ is uniformly continuous there (Theorem 7.25): given $\varepsilon > 0$ there is $\delta > 0$ with $|\partial_i F(t,y) - \partial_i F(t,x)| < \varepsilon$ whenever $|y - x| < \delta$, uniformly in t. Hence for $|s| < \delta$ the difference quotient of G differs from $\int_0^1 \partial_i F(t,x)\mathrm{d}t$ by at most ε, which is Equation (A.193). Continuity of the right-hand side in x follows from the same uniform continuity. □

Theorem A.74 (Flow of a time-dependent vector field)**.** *Let $U \subseteq \mathbb{R}^n$ be open, let $X : [0,1] \times U \longrightarrow \mathbb{R}^n$ be smooth — all partial derivatives in (t,x) of every order exist and are continuous — and let $p \in U$ satisfy $X_t(p) = 0$ for every $t \in [0,1]$. Then there is an open ball $B \ni p$ with $\overline{B} \subset U$ and a smooth map $\psi : [0,1] \times B \longrightarrow U$ with*

$$\partial_t \psi(t,x) = X_t(\psi(t,x))\,, \qquad \psi(0,x) = x\,, \qquad \psi(t,p) = p\,, \tag{A.194}$$

and each $\psi_t = \psi(t,\cdot)$ is a diffeomorphism of B onto an open subset of U.

Rests on Theorem 9.8, Corollary 9.9, and Lemma A.72.

Proof. *Derives Theorem A.74.*

A Lipschitz box. Choose $r > 0$ with $\overline{B}_{2r}(p) \subset U$ and put $Q = [0,1] \times \overline{B}_{2r}(p)$, a compact set. On Q the continuous functions X and $\partial_j X^i$ are bounded (Theorem 7.24); set $M = \max_Q |X|$ and

$$L = \sqrt{n}\, \max_Q \|D_x X\| \,, \tag{A.195}$$

with $D_x X$ the matrix $\partial_j X^i$ and $\|\cdot\|$ as in Equation (A.185). Since $\overline{B}_{2r}(p)$ is convex, Theorem 7.35 applied componentwise along the segment from z to y bounds each component of the difference by $|\nabla_x X^i|\,|y-z| \le \|D_x X\|\,|y-z|$, and summing the n squares gives

$$|X_t(y) - X_t(z)| \le L\,|y-z| \,, \qquad y, z \in \overline{B}_{2r}(p)\,, \quad t \in [0,1]\,. \tag{A.196}$$

Existence on all of $[0,1]$. The constant curve $y \equiv p$ solves Equation (A.194) with $y(0) = p$, because $X_t(p) = 0$; by uniqueness in Theorem 9.8 it is *the* solution through p, which is the third assertion of Equation (A.194). Let now $x \in B := B_\delta(p)$ with $\delta = r\,\mathrm{e}^{-L}$, and let $\psi_t(x)$ be the maximal solution. So long as it remains in $\overline{B}_{2r}(p)$, the integral form of the equation and Equation (A.196) give

$$|\psi_t(x) - p| = \left| x - p + \int_0^t [X_s(\psi_s(x)) - X_s(p)]\,\mathrm{d}s \right| \le |x-p| + L\int_0^t |\psi_s(x) - p|\,\mathrm{d}s\,,$$

so Lemma A.72 yields

$$|\psi_t(x) - p| \le \mathrm{e}^L\,|x-p| < r\,. \tag{A.197}$$

The solution therefore never approaches the boundary of $\overline{B}_{2r}(p)$. Since the Picard step $h = \min\{1,\ r/M\}$ of Theorem 9.8 is the same at every starting time in $[0,1]$ and every starting point of $\overline{B}_r(p)$, the solution can be continued in steps of that fixed length and reaches $t = 1$ after finitely many of them.

Lipschitz dependence. For $x, y \in B$ the same computation with p replaced by y, using Equation (A.196) (both curves stay in $\overline{B}_r(p)$ by Equation (A.197)) and Lemma A.72, gives

$$|\psi_t(x) - \psi_t(y)| \le \mathrm{e}^L\,|x-y| \,, \qquad t \in [0,1]\,. \tag{A.198}$$

The variational equation. Put $A(t,x) = D_x X(t, \psi_t(x))$, continuous on $[0,1] \times B$ with $\|A\| \le L$, and let $\Phi(\cdot, x)$ be the unique solution of the linear system

$$\partial_t \Phi(t,x) = A(t,x)\,\Phi(t,x)\,, \qquad \Phi(0,x) = \mathbb{1}\,, \tag{A.199}$$

which exists on the whole of $[0,1]$ by Corollary 9.9. Its integral form and Lemma A.72 give $\|\Phi(t,x)\| \le \sqrt{n}\,\mathrm{e}^L$.

Φ *is continuous in* x. Subtracting the integral forms of Equation (A.199) at x and at y,

$$\Phi(t,x) - \Phi(t,y) = \int_0^t A(s,x)\,[\Phi(s,x) - \Phi(s,y)]\,\mathrm{d}s + \int_0^t [A(s,x) - A(s,y)]\,\Phi(s,y)\,\mathrm{d}s\,.$$

The second integral is bounded by $\sqrt{n}\,\mathrm{e}^L\,\eta(x,y)$ with $\eta(x,y) = \max_s \|A(s,x) - A(s,y)\|$, and $\eta(x,y) \to 0$ as $y \to x$: $D_x X$ is uniformly continuous on the compact Q (Theorem 7.25) and $|\psi_s(x) - \psi_s(y)| \le \mathrm{e}^L\,|x-y|$ by Equation (A.198). Lemma A.72 then gives $\|\Phi(t,x) - \Phi(t,y)\| \le \sqrt{n}\,\mathrm{e}^{2L}\eta(x,y)$.

Φ *is the derivative.* Fix $x \in B$ and let h be small enough that $x + h \in B$. Put $\Delta(t) = \psi_t(x+h) - \psi_t(x)$ and $R(t) = \Delta(t) - \Phi(t,x)h$, so $R(0) = 0$ and

$$\partial_t R(t) = X_t(\psi_t(x+h)) - X_t(\psi_t(x)) - A(t,x)\Phi(t,x)h = A(t,x)R(t) + \rho(t)\,,$$

where $\rho(t) = X_t(z+\Delta) - X_t(z) - D_x X(t,z)\Delta$ with $z = \psi_t(x)$. Applying Theorem 7.35 to $\lambda \mapsto X_t^i(z + \lambda\Delta)$ gives $\theta_i \in (0,1)$ with $\rho^i(t) = [\partial_j X^i(t, z + \theta_i\Delta) - \partial_j X^i(t,z)]\,\Delta^j$, whence

$|\rho(t)| \le w(|\Delta|)\,|\Delta|$, where w is $\sqrt{n}$ times a modulus of uniform continuity of D_xX on the compact set Q (Theorem 7.25), so that $w(\lambda) \to 0$ as $\lambda \to 0$. With $|\Delta| \le \mathrm{e}^{L}\,|h|$ from Equation (A.198), the integral form of the equation for R and Lemma A.72 give

$$|R(t)| \le \mathrm{e}^{L}\, w\Big(\mathrm{e}^{L}\,|h|\Big)\,\mathrm{e}^{L}\,|h|\ ,$$

so $|R(t)|\,/\,|h| \to 0$ as $h \to 0$. By Definition 7.67, ψ_t is differentiable at x with differential $\Phi(t,x)$; the previous paragraph makes the partial derivatives continuous, so ψ_t is C^1 (Definition 7.66).

Derivatives of every order. We show by induction on $k \ge 1$: if X is C^k then ψ is C^k in x. The case $k = 1$ is what was just proved. Let X be C^{k+1} and consider the enlarged system on $U \times \mathbb{R}^{n\times n}$,

$$\partial_t \begin{pmatrix} y \\ \Phi \end{pmatrix} = \begin{pmatrix} X_t(y) \\ D_xX(t,y)\,\Phi \end{pmatrix}\ , \qquad (y,\Phi)(0) = (x,\mathbb{1})\,, \tag{A.200}$$

whose right-hand side is C^k in (y,Φ) because X is C^{k+1}. Its solution is exactly $(\psi(t,x),\Phi(t,x))$, and the induction hypothesis applied to Equation (A.200) makes it C^k in the initial data, hence in x. So $\partial\psi/\partial x = \Phi$ is C^k in x, i.e. ψ is C^{k+1} in x. Since moreover $\partial_t\psi = X_t(\psi)$, every derivative in t is a polynomial expression in derivatives of X and of ψ in x; hence all mixed partial derivatives of ψ in (t,x) of every order exist and are continuous, and ψ is smooth.

Diffeomorphism. Fix $t \in [0,1]$ and let η_s be the flow built by the same construction for the field $Y_s(y) = -X_{t-s}(y)$, which also vanishes at p; let B_2 be its ball. That ball is the *same* for every t: the construction used only r, M and L, and those depend on X and the box Q alone, not on t, while $|Y|$ and $\|D_xY\|$ have the same maxima over Q as $|X|$ and $\|D_xX\|$. Shrinking B once and for all so that $\mathrm{e}^{L}\delta$ is smaller than the radius of B_2, Equation (A.197) puts $\psi_t(B)$ inside B_2 for every t. For $x \in B$ the curve $s \mapsto \psi_{t-s}(x)$ satisfies $\mathrm{d}/\mathrm{d}s = -X_{t-s}(\psi_{t-s}(x)) = Y_s(\cdot)$ and starts at $\psi_t(x)$, so uniqueness (Theorem 9.8) gives $\eta_t(\psi_t(x)) = \psi_0(x) = x$. Put $W = \{y \in B_2 \mid \eta_t(y) \in B\}$, open because η_t is continuous. Then $\psi_t(B) \subseteq W$; and for $y \in W$ the same uniqueness argument run backwards gives $\psi_t(\eta_t(y)) = y$, so $W \subseteq \psi_t(B)$. Hence $\psi_t : B \longrightarrow W$ is a bijection onto an open set, with inverse η_t, and both are smooth. □

Lemma A.75 (Differentiating a pullback along a flow)**.** *Let $\varphi : I \times U \longrightarrow V$ be smooth, $I \subseteq \mathbb{R}$ an interval, $U, V \subseteq \mathbb{R}^n$ open, and let X_t be a smooth time-dependent vector field on V with*

$$\partial_t\varphi_t(x) = X_t(\varphi_t(x))\,. \tag{A.201}$$

Let α_t be a smooth time-dependent k-form on V. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\,(\varphi_t^*\alpha_t) = \varphi_t^*\,(\mathcal{L}_{X_t}\alpha_t + \partial_t\alpha_t)\ , \tag{A.202}$$

where the Lie derivative of a k-form has the components

$$(\mathcal{L}_X\alpha)_{j_1\ldots j_k} = X^l\partial_l\alpha_{j_1\ldots j_k} + \sum_{a=1}^{k}\alpha_{j_1\ldots l\ldots j_k}\,\partial_{j_a}X^l\,, \tag{A.203}$$

the index l standing in the a-th slot.

Rests on Definition A.70 and Propositions 7.73 and 13.127.

Proof. *Derives Lemma A.75.*

By Equation (A.186),

$$(\varphi_t^*\alpha_t)_{i_1\ldots i_k}(x) = \alpha_{t,\,j_1\ldots j_k}(\varphi_t(x))\,\partial_{i_1}\varphi_t^{j_1}(x)\cdots\partial_{i_k}\varphi_t^{j_k}(x)\,.$$

Differentiate in t by the product rule. The first factor contributes, by the chain rule Equation (7.52) and Equation (A.201),

$$\left[\partial_t \alpha_{t,\, j_1 \dots j_k} + X_t^l \, \partial_l \alpha_{t,\, j_1 \dots j_k}\right] (\varphi_t(x)) \, \partial_{i_1} \varphi_t^{j_1} \cdots \partial_{i_k} \varphi_t^{j_k} \, .$$

For the remaining k factors, the mixed partial derivatives of φ may be exchanged (Proposition 7.73), so

$$\partial_t \, \partial_i \varphi_t^j(x) = \partial_i \left(X_t^j(\varphi_t(x)) \right) = \partial_l X_t^j(\varphi_t(x)) \, \partial_i \varphi_t^l(x) \, ,$$

again by Equation (7.52). The a-th such term is therefore

$$\alpha_{t,\, j_1 \dots j_a \dots j_k}(\varphi_t) \, \partial_l X_t^{j_a}(\varphi_t) \, \partial_{i_a} \varphi_t^l \prod_{b \neq a} \partial_{i_b} \varphi_t^{j_b} \, ,$$

which, after renaming the summation indices $j_a \leftrightarrow l$, is the pullback of the a-th drag term of Equation (A.203). Collecting the $k+1$ contributions gives Equation (A.202). Finally, taking $\alpha_t = \alpha$ independent of t, $\varphi_0 = \mathrm{id}$ and $X_t = X$, the identity at $t = 0$ reads $\mathrm{d}\left(\varphi_t^* \alpha\right) / \mathrm{d}t \big|_0 = \mathcal{L}_X \alpha$ with the right-hand side Equation (A.203): this is the definition Equation (13.267) of the Lie derivative, so Equation (A.203) is indeed the k-form case of Proposition 13.127 — one transport term and one drag term per index — and no separate verification is needed. □

A.12.4 The local converse of the Poincaré lemma

Lemma 13.107 states one half of the correspondence: an exact form is closed. The half needed here is the converse, which is false in general (it fails on a punctured plane) and true on a star-shaped set. Chapter 13 announces the dependence on topology but does not prove the positive statement; it is proved here.

Definition A.76 (Star-shaped set)**.** An open set $B \subseteq \mathbb{R}^n$ is *star-shaped about* $x_0 \in B$ if $x_0 + t\,(x - x_0) \in B$ for every $x \in B$ and every $t \in [0, 1]$. Every open ball is star-shaped about each of its points.

Rests on Definition 6.26.

Lemma A.77 (Poincaré lemma, converse form)**.** *Let $B \subseteq \mathbb{R}^n$ be open and star-shaped about x_0, and let α be a smooth closed k-form on B with $k \geq 1$. Then $\alpha = \mathrm{d}\sigma$, where σ is the $(k-1)$-form with components*

$$\sigma_{i_2 \dots i_k}(x) = \int_0^1 t^{k-1} \, (x - x_0)^j \, \alpha_{j\, i_2 \dots i_k}\big(x_0 + t(x - x_0)\big) \, \mathrm{d}t \, . \tag{A.204}$$

In particular $\sigma(x_0) = 0$, and for $k = 2$ the one-form σ vanishes at x_0.

Rests on Definition A.76, Lemmas A.73 and A.75, and Proposition 13.128.

Proof. *Derives Lemma A.77.*

Translating, take $x_0 = 0$. Put $\varphi_t(x) = tx$ for $t \in (0, 1]$, which maps B into B by Definition A.76, and

$$X_t(y) = \frac{y}{t} \, , \qquad \text{so that} \qquad \partial_t \varphi_t(x) = x = \frac{\varphi_t(x)}{t} = X_t(\varphi_t(x)) \, ,$$

which is Equation (A.201). Both φ and X are smooth on $(0, 1] \times B$. By Lemma A.75 with $\alpha_t = \alpha$ fixed, and by Cartan's magic formula Equation (13.276) together with $\mathrm{d}\alpha = 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\varphi_t^* \alpha\right) = \varphi_t^* \left(\mathcal{L}_{X_t} \alpha\right) = \varphi_t^* \left(\mathrm{d}\, i_{X_t} \alpha\right) = \mathrm{d}\left(\varphi_t^* \, i_{X_t} \alpha\right) \, , \tag{A.205}$$

the last step by Equation (A.188). The components of the form being differentiated are

$$(\varphi_t^*\, i_{X_t}\alpha)_{i_2\ldots i_k}(x) = t^{k-1}\, x^j\, \alpha_{j\, i_2\ldots i_k}(tx)\,, \tag{A.206}$$

since $(i_{X_t}\alpha)_{i_2\ldots i_k}(y) = t^{-1}y^j\alpha_{j\,i_2\ldots i_k}(y)$ by the interior-product convention of Proposition 13.128, each of the $k-1$ pullback factors $\partial_i\varphi_t^j = t\,\delta^j{}_i$ contributes a t, and $y = tx$ supplies $y^j/t = x^j$. The right-hand side of Equation (A.206) extends continuously to $t = 0$ because $k \geq 1$, and so do its partial derivatives in x, so σ of Equation (A.204) is a well-defined smooth $(k-1)$-form and, by Lemma A.73, d may be taken under the integral sign. Integrating Equation (A.205) from ε to 1,

$$\varphi_1^*\alpha - \varphi_\varepsilon^*\alpha = \mathrm{d}\int_\varepsilon^1 \varphi_t^*\, i_{X_t}\alpha\; \mathrm{d}t\,. \tag{A.207}$$

Now $\varphi_1 = \mathrm{id}$, so the first term is α; and $(\varphi_\varepsilon^*\alpha)_{i_1\ldots i_k}(x) = \varepsilon^k\alpha_{i_1\ldots i_k}(\varepsilon x)$ tends to 0 as $\varepsilon \to 0$, again because $k \geq 1$ and α is continuous at the origin. Letting $\varepsilon \to 0$ in Equation (A.207), with the interchange of d and the integral justified by Lemma A.73, gives $\alpha = \mathrm{d}\sigma$ with σ as in Equation (A.204). Setting $x = x_0$ there makes the factor $(x - x_0)^j$ vanish, so $\sigma(x_0) = 0$. □

A.12.5 Moser's homotopy argument

Proof of Theorem A.68. *Derives Theorem A.68.*

Step 1: the pointwise normal form. Let $n = \dim M$. The bilinear form ω_P on the n-dimensional real vector space $T_M(P)$ is antisymmetric and non-degenerate (Definition A.67), so Proposition 5.115 applies: $n = 2m$ is even and there is a basis $\{e_1, \ldots, e_m, f_1, \ldots, f_m\}$ of $T_M(P)$ in which Equation (5.143) holds, that is, in which the matrix of ω_P is Equation (5.144).

Step 2: a chart adapted at the point. Choose any chart $x = (x^1, \ldots, x^{2m})$ around P with $x(P) = 0$ and with the coordinate vectors ∂_i at P equal to that basis — possible because the coordinate frame of a chart may be composed with any constant invertible linear map (Definition 13.2), and the change of basis carrying the coordinate frame to $\{e_i, f_i\}$ is invertible. Let B_0 be an open ball in the image of the chart, centred at the origin. Define on B_0 the *constant-coefficient* two-form

$$\omega_0 = \sum_{i=1}^{m} \mathrm{d}x^i \wedge \mathrm{d}x^{m+i}\,, \qquad \text{components} \qquad (\omega_0)_{i,\,m+i} = +1 = -\,(\omega_0)_{m+i,\,i}\;, \tag{A.208}$$

all other components vanishing. Its matrix is exactly Equation (5.144), so by Step 1

$$\omega_0|_P = \omega|_P\;, \tag{A.209}$$

and $\mathrm{d}\omega_0 = 0$ because its components are constants (Definition 13.103).

Step 3: the interpolation. Put $\tau = \omega - \omega_0$ and

$$\omega_t = \omega_0 + t\,\tau = (1-t)\,\omega_0 + t\,\omega\,, \qquad t \in [0,1]\,. \tag{A.210}$$

Each ω_t is closed, since ω and ω_0 are; and by Equation (A.209), $\omega_t|_P = \omega|_P$ for *every* t, which is non-degenerate. The function

$$D(t,x) = \det\left[(\omega_t)_{ij}\,(x)\right]$$

is continuous on $[0,1] \times B_0$ — a polynomial in the components, which are continuous — and $D(t,0) \neq 0$ for every t. For each $t \in [0,1]$ continuity therefore supplies an open interval $I_t \ni t$ and a ball $B^{(t)} \ni 0$ on which D does not vanish; the intervals I_t cover the compact $[0,1]$

(Theorem 6.11), finitely many of them suffice, and the intersection B_1 of the corresponding finitely many balls is an open ball about the origin on which

$$\omega_t \text{ is non-degenerate for every } t \in [0,1]\,. \tag{A.211}$$

Step 4: the primitive. B_1 is star-shaped about the origin (Definition A.76) and τ is a closed two-form on it, so Lemma A.77 gives a smooth one-form σ on B_1 with

$$\mathrm{d}\sigma = \tau\,, \qquad \sigma(0) = 0\,. \tag{A.212}$$

The vanishing of σ at the origin — which is Equation (A.209) feeding through the explicit homotopy formula — is what will pin the point P and is the reason the argument is local rather than global.

Step 5: the vector field. Let $W^{ij}(t,x)$ be the inverse of the matrix $(\omega_t)_{ij}(x)$, which exists on $[0,1]\times B_1$ by Equation (A.211) and whose entries are smooth there, being rational functions of the components with non-vanishing denominator D. Define the time-dependent vector field

$$X_t^k(x) = -\,W^{ik}(t,x)\,\sigma_i(x)\,, \qquad \text{equivalently} \qquad i_{X_t}\omega_t = -\,\sigma\,, \tag{A.213}$$

the equivalence because $(i_X\omega_t)_i = X^j\,(\omega_t)_{ji}$, and contracting $X^j(\omega_t)_{ji} = -\sigma_i$ with W^{ik} returns Equation (A.213). It is smooth in (t,x), and by Equation (A.212)

$$X_t(0) = 0 \qquad \text{for every } t \in [0,1]\,. \tag{A.214}$$

Step 6: the flow. By Equation (A.214), Theorem A.74 applies: there is a ball $B_2 \ni 0$ and a smooth family $\psi_t : B_2 \longrightarrow B_1$, $t \in [0,1]$, of diffeomorphisms onto open sets, with $\psi_0 = \mathrm{id}$, $\partial_t\psi_t = X_t \circ \psi_t$ and $\psi_t(0) = 0$.

Step 7: the homotopy equation. By Lemma A.75, Cartan's magic formula Equation (13.276), $\mathrm{d}\omega_t = 0$, $\partial_t\omega_t = \tau$ and Equation (A.213),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\,(\psi_t^*\omega_t) &= \psi_t^*\,(\mathcal{L}_{X_t}\omega_t + \partial_t\omega_t)\\ &= \psi_t^*\,(\mathrm{d}\,i_{X_t}\omega_t + i_{X_t}\mathrm{d}\omega_t + \tau)\\ &= \psi_t^*\,(\mathrm{d}(-\sigma) + 0 + \mathrm{d}\sigma) = 0\,,\end{aligned} \tag{A.215}$$

where the last line used Equation (A.212). Hence $\psi_t^*\omega_t$ is independent of t on B_2, and at $t=0$ it equals $\psi_0^*\omega_0 = \omega_0$. Taking $t = 1$,

$$\psi_1^*\omega = \omega_0 \qquad \text{on } B_2\,. \tag{A.216}$$

Step 8: reading off the chart. Let $W = \psi_1(B_2)$, an open neighbourhood of 0, and define on it

$$u^a = x^a \circ \psi_1^{-1}\,, \qquad a = 1,\dots,2m\,, \tag{A.217}$$

a chart of M around P with $u(P) = 0$, since ψ_1 is a diffeomorphism and $\psi_1(0) = 0$. Applying $\left(\psi_1^{-1}\right)^*$ to Equation (A.216) and using Equation (A.189), $\omega = \left(\psi_1^{-1}\right)^*\omega_0$ on W. By Equation (A.190), $\left(\psi_1^{-1}\right)^*\mathrm{d}x^a = \mathrm{d}\left(x^a\circ\psi_1^{-1}\right) = \mathrm{d}u^a$, and by Equation (A.187) the pullback distributes over the wedge, so Equation (A.208) becomes

$$\omega = \sum_{i=1}^{m} \mathrm{d}u^i \wedge \mathrm{d}u^{m+i}\,,$$

which is Equation (A.184) with $q^i = u^i$ and $p_i = u^{m+i}$. □

A.12.6 Reading the result

Remark A.78 (Where each hypothesis is used). Both hypotheses of Definition A.67 are load-bearing and they enter at different places. Non-degeneracy is used twice: at *one point* in Step 1, to get the linear normal form from Proposition 5.115, and on a *neighbourhood* in Step 5, to solve Equation (A.213) for X_t — without it the homotopy equation cannot be solved even though it is a linear algebraic equation. Closedness is used twice as well: in Step 4, where $\mathrm{d}\tau = 0$ is precisely the hypothesis of Lemma A.77, and in Step 7, where $i_{X_t}\mathrm{d}\omega_t$ is dropped from Cartan's formula. A non-degenerate two-form that is not closed has no canonical chart, and the obstruction is exactly the three-form $\mathrm{d}\omega$, which is a genuine local invariant.

Remark A.79 (The SI units of a symplectic form). Remark 5.116 records that on the phase space of a mechanical system the value $\omega(u, v)$ carries the SI unit J s, that of action and of $\hbar$. Nothing in the proof above disturbs this: the chart of Step 2 is obtained by a *constant* linear change of frame, which distributes the unit between the two halves of the canonical pair, and every later step is an identity between two-forms and so is unit-homogeneous. In the canonical chart of Equation (A.184) the product $q^i p_i$ carries J s while the individual factors do not have to: for a point particle of mass m moving in three-dimensional space, q^i is a Cartesian position in m and p_i the conjugate momentum in kg m/s, so each summand $\mathrm{d}q^i \wedge \mathrm{d}p_i$ carries $\mathrm{kg\,m^2/s} = \mathrm{J\,s}$, as it must.

Example A.80 (The phase space of one particle in space)**.** Take $M = T^*\mathbb{R}^3$, the phase space of a single particle moving in three-dimensional space: $m = 3$, $\dim M = 6$, global coordinates $(q^1, q^2, q^3, p_1, p_2, p_3)$ and $\omega = \sum_{i=1}^{3} \mathrm{d}q^i \wedge \mathrm{d}p_i$. Here the canonical chart is global and Theorem A.68 says nothing new. Its force appears when the same phase space is written in another coordinate system — spherical coordinates and their conjugate momenta, or the action–angle variables of an integrable system: the theorem guarantees that the transformed ω can always be brought back to Equation (A.184) near any point, so that no computation in one such system can disagree with a computation in another about a local question.

Rests on Theorem A.68 and Remark A.79.

Remark A.81 (Attribution). The theorem is Darboux's; the proof given here is not his, but the homotopy argument that interpolates between two forms and integrates a time-dependent vector field, introduced by Moser for volume forms [1317] and adapted to the symplectic case shortly afterwards. Darboux's own memoir is not in this treatise's bibliography, so that half of the attribution is made in words; nothing in Appendices A.12.2 to A.12.5 rests on either, every step above having been carried out from the material of Chapter 5, Chapter 7, Chapter 9 and Chapter 13.

Remark A.82. Appendix A.12 discharges the derivation owed at Remark 5.116 of Chapter 5, where Proposition 5.115 establishes the pointwise half of the statement — a non-degenerate antisymmetric form on a single vector space has no invariant but its dimension — and Darboux's theorem is named as the geometric counterpart that does *not* follow from it. The gap between the two is exactly Steps 3 to 7 above: the linear normal form can be achieved at every point separately by a basis that varies from point to point, and it is closedness of ω, through the Poincaré lemma and the homotopy equation, that lets those bases be chosen to fit together into a chart. The three tools built along the way — the pullback of a k-form (Definition A.70), the smoothness of a flow (Theorem A.74) and the converse Poincaré lemma (Lemma A.77) — belong with the differential-form machinery of Sections 13.6, 13.6.5 and 13.9 and are used again in Appendix A.29.

A.13 The Korn–Lichtenstein Theorem and the Elliptic Canonical Form

This appendix discharges the obligation recorded in Remark 10.14 of Chapter 10: the elliptic branch of Proposition 10.13, the reduction of a second-order operator in two independent variables to the canonical form $u_{\xi\xi}+u_{\eta\eta}$, under coefficients that are merely C^1 and therefore only locally Hölder. The chapter's argument constructs the required complex first integral by complexifying the characteristic equation, which needs the coefficients to be real analytic; nothing in real ordinary differential equation theory replaces that step, because what is really being asked for is not a first integral of an ordinary equation at all but a solution of a first-order elliptic *system* in two variables — a Beltrami equation.

Four things are done below. First the reduction is made exact: the canonical form exists at a point if and only if a Beltrami equation with a coefficient of modulus bounded away from 1 has a local solution with non-vanishing Jacobian, and each of the equivalent formulations is derived from the previous one with nothing left implicit (Appendix A.13.2). Then the problem is normalized so that the Beltrami coefficient is as small as one pleases, in the Hölder norm and not merely in the supremum norm (Appendix A.13.3); this is where the Hölder hypothesis is spent, and it is what turns an existence problem into a convergent series. Then the singular-integral machinery is set out and the one deep estimate is stated precisely as a quoted theorem, with a remark saying exactly what is assumed and why this treatise does not prove it (Appendix A.13.4). Finally the solution is constructed as an explicitly convergent Neumann series and the Jacobian is bounded below by hand (Appendix A.13.5), which completes the proof.

Throughout, x and y are the two *independent variables* of the partial differential equation — not two spatial dimensions — and $z=x+\mathrm{i}y$, $\bar z=x-\mathrm{i}y$, with

$$\partial_z=\tfrac12(\partial_x-\mathrm{i}\,\partial_y)\,,\qquad \partial_{\bar z}=\tfrac12(\partial_x+\mathrm{i}\,\partial_y)\,, \tag{A.218}$$

so that $\partial_x=\partial_z+\partial_{\bar z}$ and $\partial_y=\mathrm{i}(\partial_z-\partial_{\bar z})$, and $\nabla^2=4\partial_z\partial_{\bar z}$. We write $D_\rho=\{z\in\mathbb{C}\mid |z|<\rho\}$, and for $0<\alpha<1$

$$[f]_\alpha=\sup_{z\neq w}\frac{|f(z)-f(w)|}{|z-w|^{\alpha}}\,, \tag{A.219}$$

C^α denoting the functions with $[f]_\alpha<\infty$ and $C^{k,\alpha}$ those whose derivatives up to order k lie in C^α.

A.13.1 Statement

Theorem A.83 (Local canonical form for an elliptic operator with Hölder coefficients)**.** *Let $Lu=a\,u_{xx}+2b\,u_{xy}+c\,u_{yy}+\cdots$ be elliptic on a neighbourhood of P_0, that is $\Delta=b^2-ac<0$ there, with a,b,c of class C^α for some $\alpha\in(0,1)$ — in particular whenever they are C^1. Then there are a neighbourhood V of P_0 and a map $(\xi,\eta):V\longrightarrow\mathbb{R}^2$ of class $C^{1,\alpha}$ whose Jacobian $J_0=\xi_x\eta_y-\xi_y\eta_x$ is bounded away from 0 on V, such that the transformed principal coefficients of Equation (10.12) satisfy*

$$\tilde a=\tilde c=\delta\,J_0>0\,,\qquad \tilde b=0\,,\qquad \delta=\sqrt{ac-b^2}\,, \tag{A.220}$$

so that in the coordinates (ξ,η) the principal part of L is $\delta J_0(\partial_\xi^2+\partial_\eta^2)$, a positive multiple of the Laplacian. If in addition $a,b,c\in C^{1,\alpha}$, the map is of class $C^{2,\alpha}$ and the reduction is classical: the transformed operator is a genuine second-order operator with continuous coefficients, of the third form in Equation (10.10).

Rests on Proposition 10.13, Definition 10.6, and Equation (10.6).

Two remarks fix what Theorem A.83 does and does not claim, and the second is the reason the theorem is stated in two halves.

Remark A.84 (Why the principal part is the whole claim at C^1). The congruence Equation (10.6) that carries A to $\tilde{A}$ involves the Jacobian J and nothing else, so it is meaningful for any C^1 change of coordinates and is a pointwise statement of linear algebra. The *extra* term of Equation (10.7), by contrast, carries the second derivatives $\partial_i \partial_j y^a$ of the change; a $C^{1,\alpha}$ change does not have them, so for coefficients that are merely C^α the transformed equation is not literally an equation with continuous first-order coefficients — the first-order remainder exists only as a distribution. Nothing is wrong with the reduction; what is wrong is the naive reading of what "the equation becomes $u_{\xi\xi} + u_{\eta\eta} + \cdots$" asserts. This is why the theorem separates the two statements, and why the second half buys the classical form with one extra derivative on the coefficients. The chapter's sentence that the conclusion of Proposition 10.13 "does hold at C^1" is correct in the first sense and should be read in it.

Remark A.85. There is no loss in assuming $a > 0$. Ellipticity gives $ac > b^2 \geq 0$, so a and c are nonzero and of the same sign; replacing L by $-L$, which changes no solution of $Lu = f$ except by the sign of f, makes both positive. The matrix A of Equation (10.11) is then positive definite, and $\delta = \sqrt{ac - b^2} = \sqrt{\det A} > 0$ is continuous and bounded away from 0 near P_0.

A.13.2 From the canonical form to a Beltrami equation

Proposition A.86 (The complex characteristic equation)**.** *Let ξ, η be real C^1 functions near P_0 and set $\zeta = \xi + \mathrm{i}\eta$. Then, with $\tilde{a}, \tilde{b}, \tilde{c}$ the transformed principal coefficients of Equation (10.12),*

$$a\,\zeta_x^2 + 2b\,\zeta_x\zeta_y + c\,\zeta_y^2 = (\tilde{a} - \tilde{c}) + 2\mathrm{i}\,\tilde{b}\,. \tag{A.221}$$

Hence $\tilde{a} = \tilde{c}$ and $\tilde{b} = 0$ hold at a point if and only if ζ satisfies the characteristic equation Equation (10.8) there, with the complex $\nabla\zeta$ in place of $\nabla\phi$.

Rests on Equations (10.8) and (10.12).

Proof. *Derives Proposition A.86.*

Expand, using $\zeta_x = \xi_x + \mathrm{i}\eta_x$ and $\zeta_y = \xi_y + \mathrm{i}\eta_y$:

$$a\zeta_x^2 = a(\xi_x^2 - \eta_x^2) + 2\mathrm{i}a\,\xi_x\eta_x\,,$$

$$2b\,\zeta_x\zeta_y = 2b(\xi_x\xi_y - \eta_x\eta_y) + 2\mathrm{i}b(\xi_x\eta_y + \eta_x\xi_y)\,,$$

$$c\,\zeta_y^2 = c(\xi_y^2 - \eta_y^2) + 2\mathrm{i}c\,\xi_y\eta_y\,.$$

The real parts sum to $(a\xi_x^2 + 2b\xi_x\xi_y + c\xi_y^2) - (a\eta_x^2 + 2b\eta_x\eta_y + c\eta_y^2) = \tilde{a} - \tilde{c}$ by Equation (10.12), and the imaginary parts sum to twice the polarisation $a\xi_x\eta_x + b(\xi_x\eta_y + \eta_x\xi_y) + c\xi_y\eta_y = \tilde{b}$. □

Proposition A.87 (The isothermal system)**.** *Assume $a > 0$ and $\Delta < 0$, and write $\delta = \sqrt{ac - b^2} > 0$. A C^1 pair (ξ, η) satisfies the equation of Proposition A.86 with the branch $a\zeta_x + (b + \mathrm{i}\delta)\zeta_y = 0$ if and only if*

$$\eta_x = -\frac{b\,\xi_x + c\,\xi_y}{\delta}\,, \qquad \eta_y = \frac{a\,\xi_x + b\,\xi_y}{\delta}\,. \tag{A.222}$$

For such a pair,

$$J_0 = \xi_x\eta_y - \xi_y\eta_x = \frac{a\xi_x^2 + 2b\xi_x\xi_y + c\xi_y^2}{\delta} = \frac{\tilde{a}}{\delta}\,, \tag{A.223}$$

so $J_0 > 0$ exactly where $\nabla\xi \neq 0$, and Equation (A.220) holds there.

Rests on Proposition A.86 and Equation (10.11).

Proof. *Derives Proposition A.87.*

The quadratic $at^2 + 2bt + c$ has the two roots $t_\pm = (-b \pm \mathrm{i}\delta)/a$, so $a\zeta_x^2 + 2b\zeta_x\zeta_y + c\zeta_y^2 = a(\zeta_x - t_+\zeta_y)(\zeta_x - t_-\zeta_y)$ and the equation of Proposition A.86 holds if and only if one of the two factors vanishes. Take the factor $\zeta_x - t_-\zeta_y = 0$, i.e. $a\zeta_x + (b + \mathrm{i}\delta)\zeta_y = 0$ (the other branch is the complex conjugate statement and produces $\bar{\zeta}$). Separating real and imaginary parts of $a(\xi_x + \mathrm{i}\eta_x) + (b + \mathrm{i}\delta)(\xi_y + \mathrm{i}\eta_y) = 0$,

$$a\xi_x + b\xi_y - \delta\eta_y = 0\,, \qquad a\eta_x + b\eta_y + \delta\xi_y = 0\,. \tag{A.224}$$

The first is the second half of Equation (A.222). Substituting it into the second and using $\delta^2 = ac - b^2$, so that $b^2 + \delta^2 = ac$,

$$a\eta_x = -\delta\xi_y - \frac{b(a\xi_x + b\xi_y)}{\delta} = -\frac{ab\,\xi_x + (b^2 + \delta^2)\xi_y}{\delta} = -\frac{a(b\,\xi_x + c\,\xi_y)}{\delta}\,,$$

and dividing by $a \neq 0$ gives the first half. The steps are reversible, which is the "only if".

For Equation (A.223), substitute Equation (A.222):

$$\xi_x\eta_y - \xi_y\eta_x = \frac{\xi_x(a\xi_x + b\xi_y) + \xi_y(b\xi_x + c\xi_y)}{\delta}\,,$$

which is the displayed quadratic form. That form is $\boldsymbol{v}^{\mathsf{T}} A\,\boldsymbol{v}/\delta$ with $\boldsymbol{v} = (\xi_x, \xi_y)$ and A the positive definite matrix of Equation (10.11) (Remark A.85), so it is positive whenever $\boldsymbol{v} \neq 0$ and zero otherwise. Finally Proposition A.86 gives $\tilde{a} = \tilde{c}$ and $\tilde{b} = 0$, and Equation (A.223) identifies $\tilde{a} = \delta J_0$, which is Equation (A.220). □

Remark A.88 (Why this is not an ordinary differential equation). Eliminating η from Equation (A.222) by the equality of mixed partials $\partial_y\eta_x = \partial_x\eta_y$ gives, for ξ alone, the second-order equation in divergence form

$$\partial_x\!\left(\frac{a\,\xi_x + b\,\xi_y}{\delta}\right) + \partial_y\!\left(\frac{b\,\xi_x + c\,\xi_y}{\delta}\right) = 0\,, \tag{A.225}$$

whose principal part is again L's. That is the exact sense in which the elliptic branch of Proposition 10.13 is circular if attacked head-on: putting an elliptic operator in canonical form is equivalent to solving an elliptic equation with the same principal part. The hyperbolic and parabolic branches escape this because there the two characteristic directions are *real*, so the corresponding first integrals come from Theorem 9.8 applied to a real ordinary differential equation, which needs no more than continuity and a Lipschitz condition. The complex route below breaks the circle by solving Equation (A.225) not by elliptic theory but by inverting a constant-coefficient operator, $\partial_{\bar{z}}$, and treating everything else as a perturbation.

Proposition A.89 (Beltrami form). *With $a > 0$, $\Delta < 0$ and $\delta = \sqrt{ac - b^2}$, the branch $a\zeta_x + (b + \mathrm{i}\delta)\zeta_y = 0$ of Proposition A.87 is exactly the* Beltrami equation

$$\partial_{\bar{z}}\zeta = \mu\,\partial_z\zeta\,, \qquad \mu = -\frac{a - \delta + \mathrm{i}b}{a + \delta - \mathrm{i}b}\,, \tag{A.226}$$

and its coefficient satisfies

$$|\mu|^2 = \frac{(a - \delta)^2 + b^2}{(a + \delta)^2 + b^2} < 1\,, \tag{A.227}$$

with $\mu = 0$ exactly where A is a positive multiple of the identity. Moreover the Jacobian of $z \longmapsto \zeta$ is

$$J_0 = |\partial_z\zeta|^2 - |\partial_{\bar{z}}\zeta|^2 = \left(1 - |\mu|^2\right)\,|\partial_z\zeta|^2\,. \tag{A.228}$$

Rests on Proposition A.87 and Equation (A.218).

Proof. *Derives Proposition A.89.*

Substitute $\zeta_x = \partial_z\zeta + \partial_{\bar z}\zeta$ and $\zeta_y = \mathrm{i}(\partial_z\zeta - \partial_{\bar z}\zeta)$ from Equation (A.218) into $a\zeta_x + (b + \mathrm{i}\delta)\zeta_y = 0$ and use $(b+\mathrm{i}\delta)\mathrm{i} = -\delta + \mathrm{i}b$:

$$(a - \delta + \mathrm{i}b)\partial_z\zeta + (a + \delta - \mathrm{i}b)\partial_{\bar z}\zeta = 0\,.$$

The coefficient of $\partial_{\bar z}\zeta$ has modulus squared $(a+\delta)^2 + b^2 > 0$ and never vanishes, so the equation may be solved for $\partial_{\bar z}\zeta$, giving Equation (A.226). Its modulus is Equation (A.227), and $(a-\delta)^2 < (a+\delta)^2$ because a and δ are both positive; equality of the numerator with 0 requires $b = 0$ and $a = \delta = \sqrt{ac}$, i.e. $a = c$ and $b = 0$.

For Equation (A.228), write $A_1 = \zeta_x$ and $B_1 = \mathrm{i}\zeta_y$, so that $\partial_z\zeta = \frac{1}{2}(A_1 - B_1)$ and $\partial_{\bar z}\zeta = \frac{1}{2}(A_1 + B_1)$. Then

$$|\partial_z\zeta|^2 - |\partial_{\bar z}\zeta|^2 = \tfrac{1}{4}\left(|A_1 - B_1|^2 - |A_1 + B_1|^2\right) = -\Re(\bar A_1 B_1) = -\Re(\mathrm{i}\,\bar\zeta_x\zeta_y) = \Im(\bar\zeta_x\zeta_y)\,,$$

and $\bar\zeta_x\zeta_y = (\xi_x - \mathrm{i}\eta_x)(\xi_y + \mathrm{i}\eta_y)$ has imaginary part $\xi_x\eta_y - \eta_x\xi_y = J_0$. Inserting Equation (A.226) gives the second equality. □

Proposition A.89 reduces Theorem A.83 to a single question: does Equation (A.226) have a local $C^{1,\alpha}$ solution with $\partial_z\zeta \neq 0$? Note that Equation (A.228) has already disposed of the Jacobian: because $|\mu|$ is bounded away from 1, a solution whose $\partial_z\zeta$ does not vanish automatically has J_0 bounded away from 0, and no separate argument is needed.

A.13.3 Normalizing the Beltrami coefficient

Lemma A.90 (Affine normalization)**.** *There is an invertible real linear change of the independent variables after which $A(P_0) = \mathbb{1}$, $\delta(P_0) = 1$ and $\mu(P_0) = 0$, the transformed coefficients being again of class C^α and again elliptic.*

Rests on Proposition A.89 and Theorem 10.8.

Proof. *Derives Lemma A.90.*

$A(P_0)$ is symmetric and positive definite (Remark A.85), so it factors as $A(P_0) = S\,S^{\mathsf T}$ with S real and invertible — take $S = O\Lambda^{1/2}$ with $A(P_0) = O\Lambda O^{\mathsf T}$ its orthogonal diagonalization, the eigenvalues in Λ being positive. Change variables by $\hat{\boldsymbol{x}} = S^{-1}\boldsymbol{x}$, whose Jacobian is the constant matrix $J = S^{-1}$; by Equation (10.6) the new principal matrix is

$$\hat A = S^{-1}A\,(S^{-1})^{\mathsf T}\,, \qquad \hat A(P_0) = S^{-1}S\,S^{\mathsf T}(S^{\mathsf T})^{-1} = \mathbb{1}\,. \tag{A.229}$$

Congruence with a constant matrix preserves Hölder continuity and, by Theorem 10.8, ellipticity. At P_0 one then has $\hat a = \hat c = 1$, $\hat b = 0$, $\hat\delta = 1$, and Equation (A.226) gives $\mu(P_0) = -(1-1+0)/(1+1-0) = 0$. Since a composition of two changes of coordinates with non-vanishing Jacobians is again one, proving Theorem A.83 for the normalized problem proves it in general. □

Lemma A.91 (Smallness after rescaling)**.** *Assume the normalization of Lemma A.90 and identify P_0 with $0 \in \mathbb{C}$. Fix once and for all a cutoff $\chi \in C^\infty(\mathbb{C})$ with $0 \le \chi \le 1$, $\chi = 1$ on $\overline{D_1}$ and $\chi = 0$ off D_2. For $r > 0$ put*

$$\nu_r(z) = \chi(z)\,\mu(rz)\,. \tag{A.230}$$

Then $\nu_r \in C^\alpha(\mathbb{C})$ vanishes off $\overline{D_2}$, coincides with $z \longmapsto \mu(rz)$ on $\overline{D_1}$, satisfies $\|\nu_r\|_\infty \le \sup|\mu| < 1$, and

$$\|\nu_r\|_\infty + [\nu_r]_\alpha \le K\,r^\alpha\,[\mu]_\alpha \longrightarrow 0 \qquad (r \to 0^+)\,, \tag{A.231}$$

with K depending only on χ and α. If moreover $a, b, c \in C^{1,\alpha}$, the same holds with the $C^{1,\alpha}$ norm in place of the C^α norm and r in place of r^α.

Rests on Lemma A.90 and Equation (A.219).

Proof. *Derives Lemma A.91.*

μ is a rational function of a, b, δ with non-vanishing denominator, hence of class C^α on a neighbourhood of 0; shrink that neighbourhood to a disc D_ρ and take $r < \rho/2$ so that $rz \in D_\rho$ for $z \in D_2$. Write $\mu^{(r)}(z) = \mu(rz)$. Since $\mu(0) = 0$ by Lemma A.90,

$$\left|\mu^{(r)}(z)\right| = |\mu(rz) - \mu(0)| \le [\mu]_\alpha \, |rz|^\alpha \le [\mu]_\alpha \, (2r)^\alpha \qquad (z \in D_2)\,, \tag{A.232}$$

and directly from Equation (A.219), $[\mu^{(r)}]_{\alpha, D_2} = r^\alpha [\mu]_\alpha$. For the product with the cutoff, use $|\chi f(z) - \chi f(w)| \le |\chi(z)|\,|f(z) - f(w)| + |f(w)|\,|\chi(z) - \chi(w)|$, so that

$$[\chi f]_\alpha \le \|\chi\|_\infty \, [f]_\alpha + \|f\|_\infty \, [\chi]_\alpha\,, \tag{A.233}$$

which with Equation (A.232) gives $[\nu_r]_\alpha \le r^\alpha [\mu]_\alpha + (2r)^\alpha [\mu]_\alpha [\chi]_\alpha$. Adding $\|\nu_r\|_\infty \le (2r)^\alpha [\mu]_\alpha$ gives Equation (A.231) with $K = 1 + 2^\alpha (1 + [\chi]_\alpha)$. The bound $\|\nu_r\|_\infty \le \sup|\mu| < 1$ is Equation (A.227) and $0 \le \chi \le 1$. Under the stronger hypothesis $a, b, c \in C^{1,\alpha}$ one has $\mu \in C^{1,\alpha}$ and $\partial \mu^{(r)} = r\,(\partial\mu)(rz)$, so every seminorm entering the $C^{1,\alpha}$ norm of $\mu^{(r)}$ carries at least one factor r, and the same product estimate applies. □

Remark A.92 (Where the Hölder hypothesis is spent). Lemma A.91 is the whole use made of the Hölder continuity of the coefficients, and it is worth seeing why mere continuity would not do. Continuity of μ at P_0 makes $\|\nu_r\|_\infty$ small, which is enough for an L^p theory but not for a Hölder one; what the exponent α buys is the factor r^α in Equation (A.231), which makes the *seminorm* small as well and thereby turns the Neumann series of Appendix A.13.5 into a convergent series in the very space where the solution's regularity is measured. A zooming argument of this kind is available only because the Beltrami equation is invariant under $z \longmapsto rz$: if $\hat\zeta$ solves $\partial_{\bar z}\hat\zeta = \mu^{(r)}\,\partial_z\hat\zeta$ on D_1, then $\zeta(w) = \hat\zeta(w/r)$ solves Equation (A.226) on D_r, both derivatives picking up the same factor $1/r$, and the Jacobian is multiplied by the positive number r^{-2}.

A.13.4 The Cauchy and Beurling transforms

Definition A.93 (Cauchy and Beurling transforms)**.** For f continuous and vanishing outside a compact set put

$$Pf(z) = \frac{1}{\pi}\int_{\mathbb{C}} \frac{f(w)}{z - w}\,\mathrm{d}A(w)\,, \qquad Tf(z) = -\frac{1}{\pi}\,\mathrm{p.v.}\!\int_{\mathbb{C}} \frac{f(w)}{(z-w)^2}\,\mathrm{d}A(w)\,, \tag{A.234}$$

$\mathrm{d}A$ being the area element of the plane and the principal value meaning the limit of the integral over $|z - w| > \varepsilon$ as $\varepsilon \to 0^+$. P is the *Cauchy transform* and T the *Beurling transform*.

Rests on Equation (A.218).

The integral defining Pf converges absolutely, the kernel $|z-w|^{-1}$ being integrable in the plane; the one defining Tf does not, and the principal value is not a notational convenience but the entire difficulty. The next lemma explains why P is the right object, and is the one part of the analytic package this treatise can prove outright, because it is the two-dimensional twin of Proposition 10.67.

Lemma A.94 (Fundamental solution of $\partial_{\bar z}$)**.** *In the sense of distributions on* $\mathbb{C} \cong \mathbb{R}^2$,

$$\nabla^2\!\left(\frac{1}{2\pi}\ln|z|\right) = \delta\,, \qquad \textit{and hence} \qquad \partial_{\bar z}\!\left(\frac{1}{\pi z}\right) = \delta\,. \tag{A.235}$$

Consequently $Pf = (\pi z)^{-1} * f$ *satisfies* $\partial_{\bar z} Pf = f$ *distributionally.*

Rests on Proposition 10.67 and Equation (A.218).

Proof. *Derives Lemma A.94.*

Let $G(z) = (2\pi)^{-1} \ln|z|$, which is harmonic on $\mathbb{C}\backslash\{0\}$ because $\nabla^2 \ln r = r^{-1}\partial_r(r\,\partial_r \ln r) = r^{-1}\partial_r(1) = 0$ in plane polar coordinates. Let φ be a test function. Since $\ln|z|$ is integrable near the origin, $\int G\nabla^2\varphi \,\mathrm{d}A = \lim_{\varepsilon\to 0^+} \int_{r>\varepsilon} G\nabla^2\varphi \,\mathrm{d}A$. On $r > \varepsilon$ apply Green's second identity Equation (10.41) with $p = 1$, $u = \varphi$, $v = G$; the term $\varphi\nabla^2 G$ vanishes, and the only boundary is the circle $r = \varepsilon$, whose outward normal for the region $r > \varepsilon$ points at the origin, so $\partial_\nu = -\partial_r$ there. Hence

$$-\int_{r>\varepsilon} G\,\nabla^2\varphi \,\mathrm{d}A = \oint_{r=\varepsilon} (-\varphi\,\partial_r G + G\,\partial_r\varphi)\,\mathrm{d}s = -\frac{1}{2\pi\varepsilon}\oint_{r=\varepsilon} \varphi\,\mathrm{d}s + \frac{\ln\varepsilon}{2\pi}\oint_{r=\varepsilon} \partial_r\varphi\,\mathrm{d}s\,,$$

using $\partial_r G = (2\pi r)^{-1}$. The circle has length $2\pi\varepsilon$, so the first term tends to $-\varphi(0)$ by continuity of φ and the second is bounded by $\varepsilon\,|\ln\varepsilon|\sup|\nabla\varphi| \to 0$. Thus $\int G\nabla^2\varphi\,\mathrm{d}A = \varphi(0)$, which is the first half of Equation (A.235). For the second, note $\nabla^2 = 4\partial_z\partial_{\bar z}$ and $\partial_z \ln|z| = \partial_z \frac{1}{2}\ln(z\bar z) = 1/(2z)$, so $\delta = 4\partial_{\bar z}\partial_z G = 4\partial_{\bar z}(1/(4\pi z)) = \partial_{\bar z}(1/(\pi z))$. Finally Equation (A.234) is the convolution of f with $1/(\pi z)$, and differentiating a convolution moves the derivative onto either factor. □

Theorem A.95 (Quoted: the Korn–Lichtenstein estimate for the Beurling transform)**.** *Let $0 < \alpha < 1$ and let $f \in C^\alpha(\mathbb{C})$ vanish outside $\overline{D_2}$. Then:*

1. *the principal value defining Tf in Equation (A.234) exists at every point of $\mathbb{C}$;*
2. *Pf is continuously differentiable on $\mathbb{C}$, with*

$$\partial_{\bar z} Pf = f\,, \qquad \partial_z Pf = Tf \tag{A.236}$$

   holding classically;

3. *$Tf \in C^\alpha(\mathbb{C})$ and*

$$\|Tf\|_\infty + [Tf]_\alpha \le C_\alpha\,[f]_\alpha\,, \tag{A.237}$$

   with C_α depending only on α;

4. *if in addition $f \in C^{k,\alpha}$ for an integer $k \ge 1$, then $Tf \in C^{k,\alpha}$ and $\|Tf\|_{C^{k,\alpha}} \le C_{k,\alpha}\,\|f\|_{C^{k,\alpha}}$.*

Remark A.96 (What is quoted here). Theorem A.95 is the only input of this section that is not proved, and it is worth being exact about its three parts. Part (ii) is half proved above: $\partial_{\bar z} Pf = f$ is Lemma A.94, and the Hölder hypothesis upgrades it from a distributional to a classical identity. Parts (i) and the second half of (ii) — that the singular integral converges and that it is the z-derivative of Pf — are the statement that differentiating the Newtonian-type potential Pf a second time is legitimate in the principal-value sense; the same question arises for the Newtonian potential of Proposition 10.67, where the chapter needed only one derivative and could avoid it. Part (iii) is the substantial content: it says the Beurling transform, a singular integral operator whose kernel $(z-w)^{-2}$ is exactly at the critical homogeneity for the plane, is bounded on Hölder spaces. That is the *Korn–Lichtenstein inequality*, proved for such kernels by Korn in 1914 and Lichtenstein in 1916 and placed in its general setting by the Calderón–Zygmund theory of singular integrals [997]; the standard modern account of the Beltrami equation built on it is Vekua's. The two original memoirs and Vekua's monograph are not entries of this bibliography, so those attributions are made in words; what the treatise does contain and does cite for the classical, real-analytic route is Courant and Hilbert [1222].

Why the estimate is not proved here is not a matter of length. Its proof requires the Calderón–Zygmund decomposition of a function relative to a cube, or an equivalent covering argument, and both rest on Lebesgue measure and on the maximal function; Chapter 7

develops the Riemann integral only, and the Lebesgue theory these arguments belong to is not built anywhere in this book. Everything else in this section — the reduction of Appendix A.13.2, the normalization of Appendix A.13.3, the convergent series and the Jacobian bound of Appendix A.13.5 — is elementary once Equation (A.237) is granted, and is carried out in full.

Remark A.97 (The shape of the estimate). Two features of Equation (A.237) are used below and are worth isolating. First, the right-hand side involves only the seminorm $[f]_\alpha$, which is legitimate because f vanishes off $\overline{D_2}$: every point of $\overline{D_2}$ lies within distance 4 of a point where f vanishes, so $\|f\|_\infty \le 4^\alpha [f]_\alpha$, and $[\,\cdot\,]_\alpha$ is therefore a genuine norm on the space of such f, complete because a $[\,\cdot\,]_\alpha$-Cauchy sequence converges uniformly and its limit inherits the seminorm bound. Second, the left-hand side controls $\|Tf\|_\infty$ as well as $[Tf]_\alpha$, and the supremum bound is what will be needed to keep $\partial_z \zeta$ away from zero.

A.13.5 The construction and the Jacobian

Proposition A.98 (A convergent Neumann series). *Let $\nu \in C^\alpha(\mathbb{C})$ vanish off $\overline{D_2}$ and set*

$$\varepsilon_\nu = \left(\|\nu\|_\infty + [\nu]_\alpha\right) C_\alpha \,, \tag{A.238}$$

with C_α the constant of Equation (A.237). If $\varepsilon_\nu < 1$ then the series

$$h = \sum_{n=0}^{\infty} h_n \,, \qquad h_0 = \nu \,, \qquad h_{n+1} = \nu\, T h_n \,, \tag{A.239}$$

converges in the space X of C^α functions vanishing off $\overline{D_2}$, and its sum is the unique element of X with

$$h = \nu(1 + Th) \,, \qquad [h]_\alpha \le \frac{[\nu]_\alpha}{1 - \varepsilon_\nu} \,. \tag{A.240}$$

Rests on Theorem A.95 and Equation (A.219).

Proof. *Derives Proposition A.98.*

Each h_n lies in X: it is a product one of whose factors is ν, which vanishes off $\overline{D_2}$, and it is Hölder because Th_{n-1} is, by Equation (A.237). By the product estimate Equation (A.233) applied to $\nu\, Tg$ and then Equation (A.237),

$$[\nu\, Tg]_\alpha \le \|\nu\|_\infty \,[Tg]_\alpha + [\nu]_\alpha \,\|Tg\|_\infty \le \left(\|\nu\|_\infty + [\nu]_\alpha\right)\left(\|Tg\|_\infty + [Tg]_\alpha\right) \le \varepsilon_\nu \,[g]_\alpha \tag{A.241}$$

for every $g \in X$. Hence $[h_n]_\alpha \le \varepsilon_\nu^n [\nu]_\alpha$ by induction, and the series Equation (A.239) is dominated term by term by a geometric series of ratio $\varepsilon_\nu < 1$. Since $[\,\cdot\,]_\alpha$ is a complete norm on X (Remark A.97), the series converges there, with the bound stated in Equation (A.240); the same domination and Lemma 9.6 give uniform convergence, so the sum may be rearranged and T applied term by term, T being bounded. Then

$$\nu(1 + Th) = \nu + \nu\, T\left(\sum_{n\ge 0} h_n\right) = h_0 + \sum_{n\ge 0} h_{n+1} = h \,,$$

which is Equation (A.240). Uniqueness: if $h, h' \in X$ both satisfy it, then $[h - h']_\alpha = [\nu T(h - h')]_\alpha \le \varepsilon_\nu [h - h']_\alpha$ by Equation (A.241), and $\varepsilon_\nu < 1$ forces $[h - h']_\alpha = 0$, hence $h = h'$ since both vanish off a compact set. □

Proposition A.99 (The solution and its Jacobian). *In the situation of Proposition A.98, suppose in addition*

$$\frac{C_\alpha [\nu]_\alpha}{1 - \varepsilon_\nu} \le \frac{1}{2} \,. \tag{A.242}$$

Then $\zeta(z) = z + Ph(z)$ is of class $C^{1,\alpha}$ on $\mathbb{C}$, solves $\partial_{\bar z}\zeta = \nu\,\partial_z\zeta$ there, and satisfies

$$|\partial_z\zeta| \geq \tfrac{1}{2}\,, \qquad J_0 = \left(1 - |\nu|^2\right)|\partial_z\zeta|^2 \geq \tfrac{1}{4}\left(1 - \|\nu\|_\infty^2\right) > 0\,. \tag{A.243}$$

Rests on Proposition A.98, Theorem A.95, and Equation (A.228).

Proof. *Derives Proposition A.99.*

By part (ii) of Theorem A.95 applied to $h \in X$,

$$\partial_{\bar z}\zeta = \partial_{\bar z} z + \partial_{\bar z} Ph = h\,, \qquad \partial_z\zeta = 1 + Th\,, \tag{A.244}$$

both continuous and Hölder by part (iii), so $\zeta \in C^{1,\alpha}$. The Beltrami equation $\partial_{\bar z}\zeta = \nu\,\partial_z\zeta$ reads, after Equation (A.244), precisely $h = \nu(1 + Th)$, which is Equation (A.240). For the Jacobian, Equation (A.237) and Equation (A.240) give

$$\|Th\|_\infty \leq C_\alpha [h]_\alpha \leq \frac{C_\alpha [\nu]_\alpha}{1 - \varepsilon_\nu} \leq \frac{1}{2}$$

by Equation (A.242), so $|\partial_z\zeta| = |1 + Th| \geq 1 - \frac{1}{2} = \frac{1}{2}$ everywhere. Equation (A.228) then yields the second half of Equation (A.243), and $\|\nu\|_\infty < 1$ makes it strictly positive. □

Proof of Theorem A.83. *Derives Theorem A.83.*

By Remark A.85 assume $a > 0$, and by Lemma A.90 assume further that $A(P_0) = \mathbb{1}$ and $\mu(P_0) = 0$, P_0 being the origin of $\mathbb{C}$; both reductions are linear changes of the independent variables with constant non-vanishing Jacobian, and composing them with the map constructed below gives the map of the theorem.

Let ν_r be the cutoff coefficient of Lemma A.91. By Equation (A.231) both ε_{ν_r} of Equation (A.238) and the left-hand side of Equation (A.242) are bounded by $C_\alpha K r^\alpha [\mu]_\alpha$ times a constant, so choosing r small enough makes $\varepsilon_{\nu_r} \leq \frac{1}{2}$ and Equation (A.242) hold. Fix such an r. Propositions A.98 and A.99 then produce $\hat\zeta \in C^{1,\alpha}(\mathbb{C})$ with $\partial_{\bar z}\hat\zeta = \nu_r \partial_z \hat\zeta$ and Jacobian bounded below by a positive constant.

On $\overline{D_1}$ one has $\nu_r(z) = \mu(rz)$, so $\zeta(w) = \hat\zeta(w/r)$ satisfies Equation (A.226) on D_r by the scaling identity of Remark A.92, with Jacobian multiplied by $r^{-2} > 0$ and hence still bounded away from 0. Set $V = D_r$ and $(\xi, \eta) = (\Re\zeta, \Im\zeta)$, which is $C^{1,\alpha}$. By Proposition A.89 the pair satisfies the branch $a\zeta_x + (b + \mathrm{i}\delta)\zeta_y = 0$ of the complex characteristic equation, so by Propositions A.86 and A.87 the transformed coefficients obey $\tilde b = 0$ and $\tilde a = \tilde c = \delta J_0 > 0$, which is Equation (A.220). The transformed principal part is therefore $\tilde a\, u_{\xi\xi} + \tilde c\, u_{\eta\eta} = \delta J_0 (u_{\xi\xi} + u_{\eta\eta})$, as claimed.

For the last sentence of the theorem, suppose $a, b, c \in C^{1,\alpha}$. Then $\mu \in C^{1,\alpha}$, and by the last clause of Lemma A.91 the quantity $\|\nu_r\|_{C^{1,\alpha}}$ is $O(r)$. Running Proposition A.98 in the space of $C^{1,\alpha}$ functions vanishing off $\overline{D_2}$ — legitimate because part (iv) of Theorem A.95 bounds T there, and because the product estimate Equation (A.233) has an evident $C^{1,\alpha}$ counterpart obtained by applying it to f and to each ∂f — gives $h \in C^{1,\alpha}$, hence $\zeta \in C^{2,\alpha}$ by Equation (A.244). The change of coordinates then has continuous second derivatives, the extra term of Equation (10.7) is an honest continuous first-order coefficient, and the transformed equation is the third canonical form of Equation (10.10) in the classical sense. □

Remark A.100 (What has been gained, and what the classical argument gives). For real-analytic coefficients the chapter's own construction is available and is much cheaper: complexify y, integrate the characteristic ordinary differential equation Equation (10.13) along the complex branch, and take real and imaginary parts. That argument produces an analytic ζ and needs nothing beyond existence for an analytic ordinary differential equation — which is the Cauchy–Kovalevskaya theorem Theorem 10.22 in its simplest instance. What it cannot do is survive the loss of analyticity, because a complexified non-analytic coefficient has no

meaning at all. The route taken here never complexifies the coefficients: it complexifies only the *unknown*, which turns one real second-order equation into one complex first-order equation, and then solves that by inverting the constant-coefficient operator $\partial_{\bar z}$ and treating the variable-coefficient part as a small perturbation. The price is Theorem A.95, and the gain is that C^1 coefficients suffice — which matters because the coefficients of a physical equation are material properties, measured and interpolated, and are never analytic for any reason a physicist could give.

Remark A.101. Appendix A.13 discharges the obligation stated in Remark 10.14 and completes the elliptic branch of Proposition 10.13 in Chapter 10. Its consequence for the chapter is that the trichotomy of Equation (10.10) is a statement about all elliptic operators with C^1 coefficients and not only about the real-analytic ones, so that the invariance of type proved in Theorem 10.8 is matched by an equally general normal form. One honest limitation survives and is stated in Remark A.84: at C^α coefficients the normal form is a statement about the principal part alone, and the classical form with continuous lower-order coefficients costs one more derivative. The single quoted input is Theorem A.95, delimited in Remark A.96.

A.14 The Cauchy–Kovalevskaya Theorem

This appendix proves Theorem 10.22 of Chapter 10: a Cauchy problem in normal form, with real-analytic coefficients and real-analytic data prescribed on a non-characteristic surface, has a real-analytic solution near the point in question, and it is the only analytic one. The method is Kovalevskaya's [190] and is the model for every convergence proof of this type: write the solution as a formal power series, observe that the equation and the data determine its coefficients by a recursion whose coefficients are *non-negative*, replace the data by larger data for which the same recursion can be summed in closed form, and conclude that the original series converges because it is dominated term by term by one that does.

The argument is arranged in four steps. The formal series and its recursion come first (Appendix A.14.1), together with the reduction of a normal problem of order k to a first-order quasilinear system with vanishing data — carried out at the level of formal series, which is all that is needed and avoids the compatibility questions a reduction of the differential problem would raise. The notion of a majorant and the standard majorizing function then follow (Appendix A.14.2), the explicit solution of the majorant problem (Appendix A.14.3), and the assembly (Appendix A.14.4). A closing remark exhibits Lewy's equation, which shows that the analyticity of the data cannot be weakened even to smoothness [1122].

Notation. $t \in \mathbb{R}$ and $x = (x_1, \dots, x_n) \in \mathbb{R}^n$; $\beta \in \mathbb{N}^n$ and $\gamma \in \mathbb{N}^m$ are multi-indices in the sense of Equation (10.2), with $\mathbb{N}$ containing 0; $x^\beta = x_1^{\beta_1} \cdots x_n^{\beta_n}$; e_k is the k-th coordinate multi-index; and a multi-index in the $1 + n$ variables (t, x) is written $\alpha = (\alpha_0, \alpha')$ with $\partial^\alpha = \partial_t^{\alpha_0} \partial_x^{\alpha'}$. A function is *real analytic* near a point if on some polydisc about it the function is the sum of an *absolutely* convergent power series; that is the only property of analyticity used below.

A.14.1 Formal series, the recursion, and the reduction

Definition A.102 (Normal Cauchy problem of order k)**.** A *normal Cauchy problem of order k* for the scalar unknown w is

$$\partial_t^k w = G\big(t, x, \{\partial^\gamma w\}_{|\gamma| \le k,\ \gamma_0 < k}\big)\,, \qquad \partial_t^p w(0, x) = \varphi_p(x) \quad (p = 0, \dots, k-1)\,, \tag{A.245}$$

with G and the φ_p real analytic near the relevant origins.

Rests on Definition 10.1 and Proposition 10.11.

Definition A.103 (First-order quasilinear system with zero data)**.** For the unknown $u = (u_1, \dots, u_m)$,

$$\partial_t u_i = \sum_{j=1}^{m} \sum_{k=1}^{n} a_{ij}^{k}(x,u)\, \partial_k u_j + b_i(x,u)\,, \qquad u_i(0,x) = 0\,, \tag{A.246}$$

with a_{ij}^{k} and b_i real analytic on a neighbourhood of $(x,u) = (0,0)$ in $\mathbb{R}^{n+m}$.

Rests on Definitions 10.2 and A.102.

Lemma A.104 (The recursion)**.** *Write the coefficients of Equation (A.246) as convergent series $a_{ij}^{k} = \sum_{\beta,\gamma} a_{ij,\beta\gamma}^{k}\, x^{\beta} u^{\gamma}$ and $b_i = \sum_{\beta,\gamma} b_{i,\beta\gamma}\, x^{\beta} u^{\gamma}$, and seek a formal solution $u_i = \sum_{\beta,l} u_{i,\beta,l}\, x^{\beta} t^{l}$. Then the coefficients $u_{i,\beta,l}$ are uniquely determined; they vanish for $l = 0$; and for every $l \geq 1$*

$$u_{i,\beta,l} = R_{i,\beta,l}\big(\big\{a_{ij,\beta'\gamma}^{k}\big\}, \{b_{i,\beta'\gamma}\}\big)\,, \tag{A.247}$$

where $R_{i,\beta,l}$ is a polynomial with non-negative rational coefficients, the same polynomial for every problem of the shape Equation (A.246) with the same m and n.

Rests on Definition A.103 and Equation (10.2).

Proof. *Derives Lemma A.104.*

The data give $u_{i,\beta,0} = 0$ for every i, β. Substitute the series into Equation (A.246). On the left,

$$\partial_t u_i = \sum_{\beta,l} (l+1)\, u_{i,\beta,l+1}\, x^{\beta} t^{l}\,. \tag{A.248}$$

On the right, $\partial_k u_j = \sum_{\beta,l} (\beta_k + 1) u_{j,\beta+e_k,l}\, x^{\beta} t^{l}$, whose coefficients are non-negative integers times coefficients of u; and $a_{ij}^{k}(x,u)$, being a power series in x and in u composed with the series for u, has coefficients that are polynomials with non-negative integer coefficients in the $a_{ij,\beta'\gamma}^{k}$ and the $u_{j',\beta',l'}$, since composition and multiplication of formal series involve only sums of products. The same holds for b_i. Multiplying the two series and collecting the coefficient of $x^{\beta} t^{l}$, the right-hand side is such a polynomial in which every u-coefficient carries a time index $l' \leq l$: a factor $t^{l'}$ from one series can only be accompanied by factors $t^{l''}$ with $l' + l'' \leq l$, and no negative powers occur.

Comparing with Equation (A.248),

$$u_{i,\beta,l+1} = \frac{1}{l+1}\, Q_{i,\beta,l}\big(\{a\}, \{b\}, \{u_{j,\beta',l'}\}_{l' \leq l}\big)\,, \tag{A.249}$$

with Q a polynomial with non-negative integer coefficients. Induction on l therefore determines every coefficient uniquely from the vanishing ones at $l = 0$, and substituting the previously obtained expressions into Equation (A.249) expresses $u_{i,\beta,l}$ as a polynomial in the coefficients of a and b alone. Non-negativity survives every substitution — the only new factors introduced are the positive rationals $1/(l+1)$ — which is Equation (A.247). That R does not depend on the particular a, b is clear from the construction: only m, n and the indices entered. □

Lemma A.105 (Formal reduction of a normal problem)**.** *Every normal Cauchy problem Equation (A.245) has a unique formal power series solution $\hat{w}$ about the origin. Moreover there is a system Equation (A.246), with m and with coefficients a_{ij}^{k}, b_i built analytically from G and the φ_p, whose unique formal solution $\hat{u}$ has one distinguished component equal to*

$$\hat{u}_{(0,0)} = \hat{w} - \sum_{p=0}^{k-1} \varphi_p(x)\, \frac{t^{p}}{p!}\,. \tag{A.250}$$

Rests on Lemma A.104 and Definition A.102.

Proof. *Derives Lemma A.105.*

Homogenization. Put $v = w - \sum_{p<k} \varphi_p(x) t^p/p!$. Since $\partial_t^p(\varphi_q t^q/q!)(0,x) = \delta_{pq}\varphi_q(x)$ for $p, q < k$, the new unknown satisfies $\partial_t^p v(0,x) = 0$ for $p < k$, and Equation (A.245) becomes an equation of the same shape with a new right-hand side that is G evaluated on the shifted arguments — again analytic near the origin, because it is the composition of G with polynomials in t whose coefficients are derivatives of the analytic φ_p.

The system. Take as unknowns the quantities u_α indexed by the multi-indices α in (t,x) with $|\alpha| \le k-1$, intended to be $\partial^\alpha v$, together with one extra unknown u_* intended to be t. Their equations are

1. $\partial_t u_* = 1$;
2. $\partial_t u_\alpha = u_{\alpha+e_t}$ whenever $|\alpha| \le k-2$, so that $|\alpha + e_t| \le k-1$;
3. for $|\alpha| = k-1$ with $\alpha_0 < k-1$ — so that α carries at least one x-derivative, say $\alpha \ge e_{x_j}$ — $\partial_t u_\alpha = \partial_{x_j} u_{\alpha+e_t-e_{x_j}}$, the index on the right having length $k-1$;
4. for $\alpha = (k-1)e_t$, $\partial_t u_\alpha = G\big(u_*, x, \{\cdot\}\big)$, where each argument $\partial^\gamma v$ with $|\gamma| \le k-1$ is the unknown u_γ and each argument with $|\gamma| = k$, $\gamma_0 < k$, carries an x-derivative and is written $\partial_{x_j} u_{\gamma-e_{x_j}}$.

Every right-hand side is analytic in (x,u) near the origin — the t-slot of G has been replaced by the unknown u_* — and depends linearly on the first x-derivatives of the unknowns, which is the shape of Equation (A.246). All data vanish: $u_*(0,x) = 0$, and $\partial^\alpha v(0,x) = \partial_x^{\alpha'}\big[\partial_t^{\alpha_0} v(0,x)\big] = 0$ because $\alpha_0 \le k-1$.

The two formal solutions agree. The recursion of Equation (A.245) determines a unique formal $\hat v$: rewriting the equation as $\partial_t^k v = \widetilde G$ and matching coefficients of $x^\beta t^l$ expresses l-th and higher t-coefficients in terms of lower ones exactly as in Lemma A.104, the data supplying the k lowest. Now form the formal derivatives $\partial^\alpha \hat v$ and $\hat u_* = t$. Formal differentiation is a ring homomorphism compatible with formal composition, so applying ∂^α to the formal identity $\partial_t^k \hat v = \widetilde G(\cdots)$ and to the trivial identities $\partial_t \partial^\alpha \hat v = \partial^{\alpha+e_t}\hat v$ shows that the family $(\partial^\alpha \hat v,\ t)$ satisfies every equation (i)–(iv) formally, with the correct vanishing data. By the uniqueness half of Lemma A.104 it *is* $\hat u$, and in particular $\hat u_{(0,0)} = \hat v$, which is Equation (A.250). □

Remark A.106. The reduction has been made at the level of formal series on purpose. A reduction of the *differential* problems would require showing that a solution of the system has components that really are the derivatives of its first one, which is an extra argument; here nothing of the kind is needed, because all the reduction is asked to do is to transport *convergence* from the system to the original problem, and convergence is a property of a formal series.

A.14.2 Majorants

Definition A.107 (Majorant)**.** For formal power series $f = \sum_\sigma f_\sigma w^\sigma$ and $F = \sum_\sigma F_\sigma w^\sigma$ in the same variables, write $f \ll F$ — "F majorizes f" — if $|f_\sigma| \le F_\sigma$ for every multi-index σ. In particular every coefficient of F is then non-negative.

Rests on Equation (10.2).

Lemma A.108 (The standard majorant)**.** *Let $g = \sum_\sigma g_\sigma w^\sigma$ be real analytic near the origin of $\mathbb{R}^N$, and let $r > 0$ be such that $C = \sum_\sigma |g_\sigma|\, r^{|\sigma|} < \infty$. Then $|g_\sigma| \le C\, r^{-|\sigma|}$ and*

$$g \ll \frac{C\, r}{r - (w_1 + \cdots + w_N)}\,. \tag{A.251}$$

Rests on Definition A.107.

Proof. *Derives Lemma A.108.*

The bound $|g_\sigma|\, r^{|\sigma|} \le C$ is immediate, every term of the defining sum being non-negative. For the majorant, expand the geometric series and then the powers by the multinomial theorem:

$$\frac{Cr}{r-(w_1+\cdots+w_N)} = C\sum_{q\ge 0}\frac{(w_1+\cdots+w_N)^q}{r^q} = C\sum_{\sigma}\frac{|\sigma|!}{\sigma!\, r^{|\sigma|}}\, w^\sigma\,, \tag{A.252}$$

where $\sigma! = \sigma_1!\cdots\sigma_N!$. The multinomial coefficient $|\sigma|!/\sigma!$ is a positive integer, hence at least 1, so the coefficient of w^σ in Equation (A.252) is at least $Cr^{-|\sigma|} \ge |g_\sigma|$. □

Lemma A.109 (Domination)**.** *Let Equation (A.246) have coefficients a, b and let A^k_{ij}, B_i be power series with $a^k_{ij} \ll A^k_{ij}$ and $b_i \ll B_i$. Let $U_{i,\beta,l}$ be the coefficients of the formal solution of the* majorant problem *— the system Equation (A.246) with A, B in place of a, b and the same vanishing data. Then*

$$|u_{i,\beta,l}| \le U_{i,\beta,l} \qquad \text{for every } i,\beta,l\,. \tag{A.253}$$

Rests on Lemma A.104 and Definition A.107.

Proof. *Derives Lemma A.109.*

By Lemma A.104 both sets of coefficients are given by the *same* polynomials $R_{i,\beta,l}$, evaluated at the coefficients of a, b and of A, B respectively. A polynomial with non-negative coefficients satisfies $|R(\boldsymbol{s})| \le R(|\boldsymbol{s}|)$ by the triangle inequality applied monomial by monomial, and is non-decreasing in each argument on the non-negative orthant; hence

$$|u_{i,\beta,l}| = |R_{i,\beta,l}(a,b)| \le R_{i,\beta,l}(|a|\,,|b|) \le R_{i,\beta,l}(A,B) = U_{i,\beta,l}\,,$$

the middle inequality using $|a| \le A$ and $|b| \le B$ coefficientwise. □

A.14.3 The majorant problem has a closed-form solution

Proposition A.110 (Explicit solution of the majorant problem)**.** *Fix $C, r > 0$ and take, in the notation of Lemma A.108 with the $N = n + m$ variables (x, u),*

$$A^k_{ij}(x,u) = B_i(x,u) = \frac{Cr}{r - s - (u_1+\cdots+u_m)}\,, \qquad s = x_1+\cdots+x_n\,. \tag{A.254}$$

Put $K = m(n+1)$. Then the functions $U_i = V$, $i = 1,\dots,m$, with

$$V(t,x) = \frac{1}{K}\left[(r-s) - \sqrt{(r-s)^2 - 2KCr\,t}\,\right], \tag{A.255}$$

solve the majorant problem, are real analytic on the polydisc

$$|x_k| < \frac{r}{4n} \quad (k=1,\dots,n)\,, \qquad |t| < \frac{9\,r}{32\,K\,C}\,, \tag{A.256}$$

and all their Taylor coefficients at the origin are non-negative.

Rests on Lemma A.108 and Definition A.103.

Proof. *Derives Proposition A.110.*

Write $\sigma = r - s$ and $R = \sqrt{\sigma^2 - 2KCrt}$, so $V = (\sigma - R)/K$. Since $\partial_t R = -KCr/R$ and $\partial_\sigma R = \sigma/R$,

$$\partial_t V = \frac{Cr}{R}\,, \qquad \partial_s V = -\partial_\sigma V = -\frac{1}{K}\left(1 - \frac{\sigma}{R}\right). \tag{A.257}$$

With $U_i = V$ for every i one has $U_1 + \cdots + U_m = mV$ and $\sum_{j,k} \partial_k U_j = mn\,\partial_s V$, so the majorant problem reads

$$\partial_t V = \frac{Cr}{\sigma - mV}\left(mn\,\partial_s V + 1\right). \tag{A.258}$$

Substituting Equation (A.257), the requirement Equation (A.258) is

$$\frac{Cr}{R} = \frac{Cr\left[1 - \dfrac{mn}{K} + \dfrac{mn\,\sigma}{K\,R}\right]}{\sigma\left(1 - \dfrac{m}{K}\right) + \dfrac{m}{K}R},$$

that is, after cross-multiplying by R and by the denominator,

$$\sigma\left(1 - \frac{m}{K}\right) + \frac{m}{K}R = R\left(1 - \frac{mn}{K}\right) + \frac{mn}{K}\,\sigma\,. \tag{A.259}$$

Because σ and R are functionally independent, Equation (A.259) holds identically if and only if the coefficients match separately: $1 - m/K = mn/K$ and $m/K = 1 - mn/K$. Both give $K = m + mn = m(n+1)$, which is the choice made. At $t = 0$, $R = |\sigma| = \sigma$ for $s < r$, so $V(0,x) = 0$: the data hold.

Analyticity and positivity. Factor

$$V = \frac{\sigma}{K}\left[1 - \sqrt{1 - \frac{2KCr\,t}{\sigma^2}}\right] = \frac{1}{K}\sum_{p\geq 1} c_p\,(2KCr)^p\,t^p\,\sigma^{1-2p}\,, \tag{A.260}$$

where $1 - \sqrt{1-y} = \sum_{p\geq 1} c_p y^p$ is the binomial series, whose coefficients $c_p = \dfrac{(2p-2)!}{2^{2p-1}\,p!\,(p-1)!}$ are strictly positive, and which converges absolutely for $|y| < 1$. Each factor $\sigma^{1-2p} = r^{1-2p}(1 - s/r)^{-(2p-1)}$ expands, for $|s| < r$, as $r^{1-2p}\sum_{d\geq 0}\binom{2p-2+d}{d}(s/r)^d$, again with positive coefficients, and $s^d = (x_1 + \cdots + x_n)^d$ expands with non-negative integer coefficients in the x^β. Hence every Taylor coefficient of V is non-negative, and the multiple series converges absolutely wherever $|s| < r$ and $2KCr\,|t| < (r - |s|)^2$. On the polydisc Equation (A.256) one has $|s| \leq \sum_k |x_k| < r/4$, so $(r - |s|)^2 > 9r^2/16$ and the second condition follows from $|t| < 9r/(32KC)$. □

Remark A.111. The choice Equation (A.254) is the whole trick, and its two features are independent. Making *all* the coefficient functions the same single function of s and of $u_1 + \cdots + u_m$ collapses the m-component system to one scalar equation in two variables; and making that function the geometric kernel $Cr/(r - s - \sum u_i)$ makes the resulting scalar equation quasilinear of first order with the closed-form solution Equation (A.255). Notice that Equation (A.258) is exactly an equation of the type treated by Proposition 10.15: written as $(\sigma - mV)\partial_t V - Crmn\,\partial_s V = Cr$ it is quasilinear of first order in the two variables (t,s), with characteristic system $\dot t = \sigma - mV$, $\dot s = -Crmn$, $\dot V = Cr$. It is because that system can be integrated — or, as done above, because the resulting closed form can be guessed and then verified — that the majorant problem is solvable at all; the square root in Equation (A.255) is the mark of the quadratic relation between V and t that the last two characteristic equations impose. That the exponent works out to $K = m(n+1)$ and not to something depending on the data is what makes the radius of convergence in Equation (A.256) explicit.

A.14.4 Proof of the theorem

Theorem A.112 (Cauchy–Kovalevskaya, first-order form)**.** *The problem Equation (A.246) has a real-analytic solution on a polydisc about the origin of $\mathbb{R}^{1+n}$, and it is the only analytic solution there.*

Rests on Lemmas A.108 and A.109 and Proposition A.110.

Proof. *Derives Theorem A.112.*

By hypothesis the a^k_{ij} and b_i are analytic near $(x,u)=(0,0)$; choose $r>0$ so small that all of them are represented by absolutely convergent series on the polydisc of radius r in the $n+m$ variables, and let C be the largest of the finitely many sums $\sum_{\beta\gamma}\left|a^k_{ij,\beta\gamma}\right| r^{|\beta|+|\gamma|}$ and $\sum_{\beta\gamma}|b_{i,\beta\gamma}|\,r^{|\beta|+|\gamma|}$. By Lemma A.108 every coefficient is then majorized by the function Equation (A.254).

Let $U_{i,\beta,l}$ be the formal solution of the majorant problem. By Proposition A.110 the function V of Equation (A.255) solves that problem and is analytic on the polydisc Equation (A.256); its Taylor coefficients therefore satisfy the recursion of Lemma A.104 for the majorant data, and by the uniqueness half of that lemma they *are* the $U_{i,\beta,l}$. Combining with Lemma A.109,

$$\sum_{\beta,l}|u_{i,\beta,l}|\;|x|^{\beta}\,|t|^{l}\le\sum_{\beta,l}U_{i,\beta,l}\;|x|^{\beta}\,|t|^{l}=V\big(|t|\,,|x|\big)<\infty \tag{A.261}$$

on Equation (A.256), where $|x|^{\beta}$ means $|x_1|^{\beta_1}\cdots$; the middle equality holds because all the U are non-negative, so the series is its own absolute value. Hence the formal series for each u_i converges absolutely on that polydisc and defines a function there.

The sum solves the problem. On any closed subpolydisc of ratio $\theta<1$ the differentiated series are dominated term by term by $\theta^{-1}\sup_{q\ge0}\big[(q+1)\theta^{q}\big]$ times the original one, hence converge uniformly by Lemma 9.6, so term-by-term differentiation in t and in each x_k is legitimate, one variable at a time, by Theorem 7.51. The coefficients satisfy Equation (A.249) by construction, so the analytic function $\partial_t u_i-\sum_{j,k}a^k_{ij}(x,u)\partial_k u_j-b_i(x,u)$ has every Taylor coefficient equal to zero and therefore vanishes identically; and $u_i(0,x)=0$ because $u_{i,\beta,0}=0$.

Uniqueness. An analytic solution on any polydisc about the origin has a Taylor series which, by the computation of Lemma A.104, must satisfy the same recursion with the same initial values; the coefficients are therefore those just constructed, and two analytic functions with the same Taylor series at a point agree near it. □

Proof of Theorem 10.22. *Derives Theorem 10.22.*

Let the problem be in the normal form Equation (A.245), which is what *normal Cauchy problem* means, with the initial surface already the slice $\{t=0\}$. By Lemma A.105 its unique formal solution $\hat w$ is recovered by Equation (A.250) from the distinguished component of the formal solution $\hat u$ of an associated system Equation (A.246). By Theorem A.112 that formal solution converges absolutely on a polydisc about the origin; hence so does $\hat u_{(0,0)}$, and hence so does $\hat w$, the finitely many extra terms $\varphi_p(x)t^p/p!$ being analytic. Its sum w is analytic, and its Taylor coefficients satisfy the recursion of the normal problem, so — by the argument already used in Theorem A.112 — the analytic function $\partial_t^k w-G(\cdots)$ has all Taylor coefficients zero and vanishes, while $\partial_t^p w(0,x)=\varphi_p(x)$ for $p<k$ by construction. Any other analytic solution has the same Taylor coefficients, by the same recursion, and therefore coincides with w near the origin. □

Remark A.113 (What is quoted here). The proof above is complete for a problem presented in the normal form Equation (A.245), which is the form Theorem 10.22 names. Bringing a general non-characteristic Cauchy problem to that form uses two facts that this treatise does not prove:

1. *Flattening the initial surface.* If the analytic hypersurface S is $\{\phi=0\}$ with $\nabla\phi\neq0$, the map $x\longmapsto(\phi(x),\psi_2(x),\dots)$ is an analytic change of coordinates near the point, by the analytic inverse function theorem.

2. *Solving for the highest normal derivative.* For a fully nonlinear equation $F(x,\{\partial^{\alpha}u\})=0$, non-characteristic means $\partial F/\partial(\partial_t^k u)\neq0$ at the point (Proposition 10.11 is the

linear instance of the same computation), and the analytic implicit function theorem then supplies an analytic G with $\partial_t^k u = G(\cdots)$.

Both are the analytic versions of theorems whose smooth versions belong to Chapter 7; both are themselves usually proved by the majorant method of this section, so nothing circular is being borrowed, but neither is derived here and neither is stated in the treatise. For a *linear* or quasilinear equation — which is every equation this book actually solves — step (ii) is not needed at all: being non-characteristic then means that the coefficient of $\partial_t^k u$ is nonzero, by Equation (10.8) and the computation in the proof of Proposition 10.11, and one divides by it. Nothing else in this section is quoted: the recursion, the reduction, the majorant, the closed-form solution and the convergence estimate are all carried out above.

Remark A.114 (Lewy's equation: analyticity cannot be dropped). Every hypothesis of Theorem 10.22 was used, and the analyticity of the *data* — not merely of the coefficients — is the one whose necessity is least obvious and most complete. Consider on $\mathbb{R}^3$, with coordinates (x_1, x_2, x_3), the first-order operator

$$Lu = \partial_1 u + \mathrm{i}\,\partial_2 u - 2\mathrm{i}(x_1 + \mathrm{i}x_2)\,\partial_3 u\,. \tag{A.262}$$

Its coefficients are polynomials, hence analytic, and it is elementary to check that it is non-characteristic in the x_3-direction at points where $x_1 + \mathrm{i}x_2 \neq 0$. Lewy proved that there exist $f \in C^\infty(\mathbb{R}^3)$ — real-valued, and depending on x_3 alone — for which $Lu = f$ has *no* solution of class C^1 on any neighbourhood of any point whatever [1122]. The failure is therefore not of uniqueness or of regularity but of existence, it is local and universal, and it is produced by an operator whose coefficients satisfy every hypothesis of Cauchy–Kovalevskaya. What is lost when f ceases to be analytic is exactly what the proof above consumed: the formal series still exists in the analytic case because the recursion Equation (A.249) always runs, and it is the majorant that turns it into a function. With a merely smooth right-hand side there is no series to majorize, and Equation (A.262) shows that nothing takes its place.

The reader should also keep in view the second limitation, which the chapter states in Remark 10.23 and which is independent of this one: even where the theorem applies it establishes existence and not well-posedness. Analytic data cannot be varied independently in disjoint regions, so continuous dependence in the sense of Definition 10.24 is not even being asked for, and the Cauchy problem Cauchy–Kovalevskaya solves for the Laplace equation is the ill-posed one of Remark 10.26.

Remark A.115. Appendix A.14 discharges the proof obligation of Theorem 10.22 in Chapter 10. The result is used in the treatise wherever an analytic construction is needed and nothing weaker will do: it is the existence statement behind the classical, real-analytic branch of Proposition 10.13 — the complex first integral constructed there by complexifying Equation (10.13) is supplied by this theorem, and Appendix A.13 is what replaces it when the coefficients are only C^1 — and it is the reason a formal power-series ansatz may be trusted at all when the data of a problem are analytic. Its limitations are the substance of Remark A.114 and Remark 10.23, and they are the reason the rest of Chapter 10 proceeds by energy estimates, transforms and Green's functions rather than by series.

A.15 The Stäckel Conditions and the Eleven Separable Systems

This appendix proves Theorem 10.34 of Chapter 10: the Helmholtz equation Equation (10.25) separates in an orthogonal coordinate system of three-dimensional Euclidean space if and only if the reciprocal squared scale factors are the first column of the inverse of a *Stäckel matrix* — a matrix each of whose rows depends on one coordinate alone — and the volume

element satisfies the accompanying *Robertson condition*. Both conditions are derived here, not postulated: they are read off by substituting the product ansatz into the Laplacian and demanding that the three separation constants enter the separated ordinary differential equations linearly. The three systems the treatise uses (Example 10.31, Example 10.32, Example 10.33) are then verified against the criterion, and so is the generic member of the family, the ellipsoidal system, from which all the others descend by confluence. The section closes with the enumeration of the eleven systems and with an exact statement of the one input that is quoted rather than proved.

Everything is set in the observed three-dimensional space, signature $3+0$, with the flat Euclidean metric; the classification is a statement about $\mathbb{R}^3$ and has no higher-dimensional counterpart in this treatise. Latin indices i, j, l run over $1, 2, 3$ and are *not* summed unless a summation sign is written, because the sums that appear below are not all over the same range.

A.15.1 Orthogonal coordinates and the Laplacian

Definition A.116 (Orthogonal curvilinear coordinates; scale factors)**.** Let $\boldsymbol{x} = \boldsymbol{x}(q^1, q^2, q^3)$ be a C^2 diffeomorphism of an open set of $\mathbb{R}^3$ onto another, with

$$\frac{\partial \boldsymbol{x}}{\partial q^i} \cdot \frac{\partial \boldsymbol{x}}{\partial q^j} = 0 \qquad (i \neq j)\,, \qquad h_i = \left| \frac{\partial \boldsymbol{x}}{\partial q^i} \right| > 0\,. \tag{A.263}$$

The h_i are the *scale factors*, the line element is $\mathrm{d}\ell^2 = \sum_i h_i^2\,(\mathrm{d}q^i)^2$, and the volume element is $\mathrm{d}^3x = g\,\mathrm{d}q^1\mathrm{d}q^2\mathrm{d}q^3$ with $g = h_1 h_2 h_3$.

Rests on Definition 7.66 and Equation (10.25).

Lemma A.117 (Laplacian in orthogonal coordinates)**.** *For $\psi \in C^2$,*

$$\nabla^2 \psi = \frac{1}{g} \sum_{i=1}^{3} \partial_i \left(\frac{g}{h_i^2}\, \partial_i \psi \right)\,, \qquad \partial_i = \frac{\partial}{\partial q^i}\,. \tag{A.264}$$

Rests on Definition A.116 and Theorem 7.101.

Proof. *Derives Lemma A.117.*

Let $\hat{\boldsymbol{e}}_i = h_i^{-1} \partial \boldsymbol{x}/\partial q^i$; by Equation (A.263) these form an orthonormal triad at each point. The gradient is characterized by $\mathrm{d}\psi = \nabla\psi \cdot \mathrm{d}\boldsymbol{x}$; writing $\mathrm{d}\boldsymbol{x} = \sum_i h_i\,\mathrm{d}q^i\,\hat{\boldsymbol{e}}_i$ and $\mathrm{d}\psi = \sum_i \partial_i\psi\,\mathrm{d}q^i$ and matching coefficients of the independent increments $\mathrm{d}q^i$,

$$\nabla\psi = \sum_i \frac{1}{h_i}\, \partial_i\psi\; \hat{\boldsymbol{e}}_i\,. \tag{A.265}$$

For the divergence of $\boldsymbol{V} = \sum_i V_i \hat{\boldsymbol{e}}_i$ apply the divergence theorem Equation (7.123) to the coordinate box $[q^1, q^1 + \mathrm{d}q^1] \times \cdots$. Its face at fixed q^1 has outward normal $-\hat{\boldsymbol{e}}_1$ and area $h_2 h_3\,\mathrm{d}q^2\mathrm{d}q^3 = (g/h_1)\,\mathrm{d}q^2\mathrm{d}q^3$, and the opposite face has the same area evaluated at $q^1 + \mathrm{d}q^1$; the net flux through the pair is therefore $\partial_1\big((g/h_1)V_1\big)\,\mathrm{d}q^1\mathrm{d}q^2\mathrm{d}q^3$ to first order. Summing over the three pairs and dividing by the volume $g\,\mathrm{d}q^1\mathrm{d}q^2\mathrm{d}q^3$,

$$\operatorname{div}\boldsymbol{V} = \frac{1}{g} \sum_i \partial_i \left(\frac{g}{h_i}\, V_i \right)\,. \tag{A.266}$$

Composing Equation (A.266) with Equation (A.265), so that $V_i = h_i^{-1}\partial_i\psi$, gives Equation (A.264). □

A.15.2 What separation means, and what it forces

Definition A.118 (Simple separation of the Helmholtz equation)**.** The Helmholtz equation Equation (10.25) *separates simply* in the orthogonal system (q^i) if there exist nowhere-vanishing functions $f_i(q^i)$ and functions $\phi_{ij}(q^i)$ $(i,j=1,2,3)$, each row index i carrying dependence on q^i alone, with

$$S=\det(\phi_{ij})\neq 0\,, \tag{A.267}$$

such that for every triple of constants $(\lambda_1,\lambda_2,\lambda_3)$ with $\lambda_1=k^2$, any solutions X_i of the three ordinary differential equations

$$\frac{1}{f_i}\frac{\mathrm{d}}{\mathrm{d}q^i}\left(f_i\frac{\mathrm{d}X_i}{\mathrm{d}q^i}\right)+\left(\sum_{j=1}^{3}\lambda_j\,\phi_{ij}(q^i)\right)X_i=0\qquad(i=1,2,3) \tag{A.268}$$

compose into a solution $\psi=X_1X_2X_3$ of Equation (10.25). The matrix $\Phi=(\phi_{ij})$ is the *Stäckel matrix* and S its *Stäckel determinant*.

Rests on Definition A.116 and Lemma 10.28.

Three remarks fix the content of Definition A.118 before it is used. First, the form Equation (A.268) is not a restriction but the general second-order linear equation in q^i with no first-derivative term beyond the one a weight can absorb: any equation $X''+pX'+\sigma X=0$ is brought to it by $f=\exp\int p$. Second, the requirement that the constants λ_j enter *linearly* is the definition of separation in Stäckel's sense, and it is the requirement that carries the physics: each λ_j is an eigenvalue of a commuting operator, and the separated problem is a Sturm–Liouville problem in each variable (Chapter 9) exactly because Equation (A.268) is linear in them. Third, $\lambda_1=k^2$ singles out the first column of Φ; the remaining two constants are free, and $S\neq 0$ is what makes them genuinely independent.

Lemma A.119 (Cofactor identities)**.** *Let $\Phi=(\phi_{ij})$ be a 3×3 matrix with $S=\det\Phi\neq 0$, and let M_{ij} be the cofactor of ϕ_{ij}, that is $(-1)^{i+j}$ times the determinant of the 2×2 matrix obtained by deleting row i and column j. Then*

$$\sum_{i=1}^{3}\phi_{ij}\,M_{i1}=S\,\delta_{j1}\,,\qquad(\Phi^{-1})_{1i}=\frac{M_{i1}}{S}\,, \tag{A.269}$$

and, if the i-th row of Φ depends on q^i alone, then M_{i1} does not depend on q^i at all.

Rests on Equation (A.267).

Proof. *Derives Lemma A.119.*

For $j=1$ the sum in Equation (A.269) is the Laplace expansion of $\det\Phi$ along the first column. For $j\neq 1$ it is the Laplace expansion along the first column of the matrix obtained from Φ by overwriting its first column with its j-th, since the cofactors M_{i1} do not involve the first column at all; that matrix has two equal columns and hence vanishing determinant. Dividing Equation (A.269) by S identifies M_{i1}/S as the $(1,i)$ entry of the inverse. The last assertion is immediate: the minor belonging to M_{i1} is built from the two rows other than the i-th, which by hypothesis carry no dependence on q^i. □

Theorem A.120 (Stäckel and Robertson conditions)**.** *The Helmholtz equation separates simply in the orthogonal system (q^i), in the sense of Definition A.118, if and only if there is a Stäckel matrix $\Phi=(\phi_{ij}(q^i))$ with $S=\det\Phi\neq 0$ such that*

$$\frac{1}{h_i^2}=\frac{M_{i1}}{S}\qquad(i=1,2,3)\,, \tag{A.270}$$

and the volume element satisfies the Robertson condition

$$\frac{h_1h_2h_3}{S}=f_1(q^1)\,f_2(q^2)\,f_3(q^3) \tag{A.271}$$

for some functions f_i of one variable each — the same f_i that weight the separated equations Equation (A.268).

Rests on Definition A.118 and Lemmas A.117 and A.119.

Sufficiency. Derives Theorem A.120.

Assume Equation (A.270) and Equation (A.271), and let the X_i solve Equation (A.268). Put $\psi = X_1X_2X_3$. By Equation (A.270) and $g = Sf_1f_2f_3$,

$$\frac{g}{h_i^2} = \frac{g}{S}\,M_{i1} = f_1f_2f_3\,M_{i1}\,, \tag{A.272}$$

and by the last clause of Lemma A.119 both M_{i1} and the factors f_l with $l \neq i$ are independent of q^i. Hence

$$\partial_i\left(\frac{g}{h_i^2}\,\partial_i\psi\right) = M_{i1}\left(\prod_{l\neq i} f_lX_l\right)\frac{\mathrm{d}}{\mathrm{d}q^i}\left(f_i\frac{\mathrm{d}X_i}{\mathrm{d}q^i}\right)\,, \tag{A.273}$$

because everything except f_iX_i' passes through ∂_i untouched. Insert Equation (A.273) into Equation (A.264) and divide by $\psi = X_1X_2X_3$, using $g = Sf_1f_2f_3$ once more:

$$\frac{\nabla^2\psi}{\psi} = \frac{1}{Sf_1f_2f_3}\sum_i M_{i1}\frac{\prod_{l\neq i} f_lX_l}{X_1X_2X_3}\left(f_iX_i'\right)' = \frac{1}{S}\sum_i M_{i1}\,\frac{1}{f_iX_i}\left(f_iX_i'\right)'\,. \tag{A.274}$$

The separated equations Equation (A.268) say that the last factor equals $-\sum_j \lambda_j\phi_{ij}$, so by the cofactor identity Equation (A.269)

$$\frac{\nabla^2\psi}{\psi} = -\frac{1}{S}\sum_j \lambda_j\sum_i \phi_{ij}M_{i1} = -\frac{1}{S}\sum_j \lambda_j\,S\,\delta_{j1} = -\lambda_1 = -k^2\,, \tag{A.275}$$

that is $\nabla^2\psi + k^2\psi = 0$. Note where each hypothesis acted: Equation (A.271) made Equation (A.273) a *one-variable* derivative, and Equation (A.270) collapsed the sum in Equation (A.275) onto the single constant λ_1, killing the two free constants exactly as separation requires. □

Necessity. Derives Theorem A.120.

Assume the equation separates simply and substitute $\psi = X_1X_2X_3$ into Equation (10.25) through Equation (A.264), dividing by ψ:

$$\sum_i \frac{1}{h_i^2}\left[\frac{X_i''}{X_i} + \frac{X_i'}{X_i}\,\partial_i\ln\frac{g}{h_i^2}\right] + k^2 = 0\,, \tag{A.276}$$

obtained by expanding $g^{-1}\partial_i\big((g/h_i^2)\partial_i\psi\big)$ and cancelling the $l \neq i$ factors of ψ.

Step 1: the weights. The i-th bracket in Equation (A.276) must be expressible through q^i and X_i alone — that is what makes the i-th separated equation an ordinary differential equation in q^i — and X_i is an arbitrary solution of one, so X_i''/X_i and X_i'/X_i are functionally independent quantities that cannot cancel each other's coefficients. Hence $\partial_i\ln(g/h_i^2)$ is a function of q^i alone. Call it $(\ln f_i)'(q^i)$, which defines $f_i > 0$ up to a constant; the bracket is then $\left(f_iX_i'\right)'/(f_iX_i)$, and Equation (A.276) reads

$$\sum_i \frac{E_i}{h_i^2} + k^2 = 0\,, \qquad E_i = \frac{1}{f_iX_i}\left(f_iX_i'\right)'\,, \tag{A.277}$$

each E_i a function of q^i alone.

Step 2: linearity in the constants. By hypothesis the separated equations carry the three constants linearly, so $E_i = -\sum_j \lambda_j\phi_{ij}(q^i)$ with $\lambda_1 = k^2$ and λ_2, λ_3 free. Substituting into Equation (A.277),

$$\sum_j \lambda_j\left[\sum_i \frac{\phi_{ij}}{h_i^2} - \delta_{j1}\right] = 0\,, \tag{A.278}$$

and since this must hold identically in the three independent constants λ_j, every bracket vanishes: $\sum_i \phi_{ij}/h_i^2 = \delta_{j1}$. In matrix language the row vector v with $v_i = 1/h_i^2$ satisfies $v\,\Phi = e_1^{\mathsf{T}}$, whence Φ is invertible — so $S \neq 0$, which is the independence of the constants — and $v_i = (\Phi^{-1})_{1i} = M_{i1}/S$ by Lemma A.119. That is Equation (A.270).

Step 3: the Robertson condition. Step 1 gave $\partial_i \ln(g/h_i^2) = (\ln f_i)'(q^i)$. Insert Equation (A.270): $g/h_i^2 = (g/S)\,M_{i1}$, and M_{i1} is independent of q^i by Lemma A.119, so its logarithmic q^i-derivative vanishes and

$$\partial_i \ln \frac{g}{S} = (\ln f_i)'(q^i) \qquad (i = 1, 2, 3)\,. \tag{A.279}$$

Write $\Lambda = \ln|g/S|$. Then $\partial_1\Lambda$ depends on q^1 alone, so $\Lambda = \ln|f_1(q^1)| + \Lambda_1(q^2, q^3)$; applying Equation (A.279) for $i = 2$ to this expression gives $\partial_2\Lambda_1 = (\ln f_2)'(q^2)$, hence $\Lambda_1 = \ln|f_2(q^2)| + \Lambda_2(q^3)$, and one more step gives $\Lambda_2 = \ln|f_3(q^3)|$ up to an additive constant, which is absorbed into any one f_i. Exponentiating, $|g/S| = |f_1 f_2 f_3|$; the sign is constant on a connected domain and is absorbed likewise, giving Equation (A.271). □

Remark A.121 (What the two conditions say). Equation (A.270) is a condition on the *metric alone* and is the substantive one: it says the three functions $1/h_i^2$, which in general are functions of all three coordinates, are the first column of the inverse of a matrix assembled out of nine functions of *one* coordinate each. That is an enormous restriction, and it is what makes the list of admissible systems finite. Equation (A.271) is a condition on the volume element relative to S, and is the one that distinguishes the Helmholtz operator from the Hamilton–Jacobi equation of Chapter 23: the latter involves no second derivatives of ψ and so imposes Equation (A.270) alone. Every system in the list below satisfies both, but the logical order matters — Stäckel's condition governs separation of the classical problem, Robertson's is the extra price of the wave operator, which is why the two subjects are *almost* but not exactly the same, as Remark 10.35 records.

A.15.3 The three systems the treatise uses

In each case we exhibit Φ, S, the cofactors M_{i1} and the weights f_i, and check Equation (A.270) and Equation (A.271) outright. The separation constants are named to match Section 10.2.1.

Example A.122 (Cartesian)**.** $h_1 = h_2 = h_3 = 1$, $g = 1$. With $\lambda_1 = k^2$, $\lambda_2 = \beta^2$, $\lambda_3 = \gamma^2$ the separated equations of Example 10.31 are $X'' + (k^2 - \beta^2 - \gamma^2)X = 0$, $Y'' + \beta^2 Y = 0$, $Z'' + \gamma^2 Z = 0$, so $f_i = 1$ and

$$\Phi = \begin{pmatrix} 1 & -1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad S = 1\,. \tag{A.280}$$

The cofactors of the first column are $M_{11} = +\det\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 1$, $M_{21} = -\det\begin{pmatrix} -1 & -1 \\ 0 & 1 \end{pmatrix} = 1$ and $M_{31} = +\det\begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix} = 1$, all equal to $1/h_i^2$; and $g/S = 1 = f_1 f_2 f_3$.

Rests on Theorem A.120 and Example 10.31.

Example A.123 (Circular cylindrical)**.** $(q^1, q^2, q^3) = (\rho, \varphi, z)$ with $h_1 = 1$, $h_2 = \rho$, $h_3 = 1$ and $g = \rho$. Set $\lambda_1 = k^2$, $\lambda_2 = m^2$, $\lambda_3 = \gamma^2$; the separated equations of Example 10.32 are Equation (10.32), $\Psi'' + m^2\Psi = 0$ and $Z'' + \gamma^2 Z = 0$, so $f_1 = \rho$, $f_2 = f_3 = 1$ and

$$\Phi = \begin{pmatrix} 1 & -\rho^{-2} & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad S = 1\,. \tag{A.281}$$

Then $M_{11} = 1 = 1/h_1^2$, $M_{21} = -\det\begin{pmatrix} -\rho^{-2} & -1 \\ 0 & 1 \end{pmatrix} = \rho^{-2} = 1/h_2^2$ and $M_{31} = +\det\begin{pmatrix} -\rho^{-2} & -1 \\ 1 & 0 \end{pmatrix} = 1 = 1/h_3^2$; and $g/S = \rho = f_1 f_2 f_3$.

Rests on Theorem A.120 and Example 10.32.

Example A.124 (Spherical)**.** $(q^1, q^2, q^3) = (r, \theta, \varphi)$ with $h_1 = 1$, $h_2 = r$, $h_3 = r\sin\theta$ and $g = r^2 \sin\theta$. Set $\lambda_1 = k^2$, $\lambda_2 = l(l+1)$, $\lambda_3 = m^2$; the separated equations of Example 10.33 are Equation (10.34), the associated Legendre equation and $\Psi'' + m^2\Psi = 0$, so $f_1 = r^2$, $f_2 = \sin\theta$, $f_3 = 1$ and

$$\Phi = \begin{pmatrix} 1 & -r^{-2} & 0 \\ 0 & 1 & -(\sin\theta)^{-2} \\ 0 & 0 & 1 \end{pmatrix}, \qquad S = 1\,. \tag{A.282}$$

The first-column cofactors are $M_{11} = 1 = 1/h_1^2$, $M_{21} = -\det\begin{pmatrix} -r^{-2} & 0 \\ 0 & 1 \end{pmatrix} = r^{-2} = 1/h_2^2$ and $M_{31} = +\det\begin{pmatrix} -r^{-2} & 0 \\ 1 & -(\sin\theta)^{-2} \end{pmatrix} = r^{-2}(\sin\theta)^{-2} = 1/h_3^2$; and $g/S = r^2\sin\theta = f_1 f_2 f_3$. The three checks reproduce, in one line each, the structure the chapter obtained by hand: the centrifugal term $-l(l+1)/r^2$ of Equation (10.34) is the entry ϕ_{12}, and the term $-m^2/\sin^2\theta$ of the Legendre equation is ϕ_{23}.

Rests on Theorem A.120 and Example 10.33.

A.15.4 The generic system: confocal quadrics

The three systems just checked are degenerate members of one family. The generic member is the ellipsoidal system, and it is worth carrying out in full because everything else follows from it by letting parameters coincide or run to infinity.

Definition A.125 (Ellipsoidal coordinates)**.** Fix $a_1^2 > a_2^2 > a_3^2 > 0$ and write x_1, x_2, x_3 for Cartesian coordinates and

$$P(\theta) = (\theta + a_1^2)(\theta + a_2^2)(\theta + a_3^2)\,. \tag{A.283}$$

The *confocal family* is

$$\sum_{p=1}^{3} \frac{x_p^2}{\theta + a_p^2} = 1\,. \tag{A.284}$$

Through a generic point pass exactly three members of the family, with parameters $\xi_1 > \xi_2 > \xi_3$ interlacing the poles,

$$\xi_1 > -a_3^2 > \xi_2 > -a_2^2 > \xi_3 > -a_1^2\,, \tag{A.285}$$

an ellipsoid, a hyperboloid of one sheet and a hyperboloid of two sheets respectively. These are the *ellipsoidal coordinates*.

Rests on Definition A.116.

Lemma A.126 (Cartesian coordinates and scale factors)**.** *With the notation of Definition A.125,*

$$x_p^2 = \frac{\prod_{i=1}^{3}(\xi_i + a_p^2)}{\prod_{q\neq p}(a_p^2 - a_q^2)}\,, \tag{A.286}$$

the system is orthogonal, and

$$h_i^2 = \frac{\prod_{l\neq i}(\xi_i - \xi_l)}{4\,P(\xi_i)} > 0\,. \tag{A.287}$$

Rests on Definition A.125 and Equation (A.263).

Proof. *Derives Lemma A.126.*

Multiply Equation (A.284), in the form $1-\sum_p x_p^2/(\theta+a_p^2)=0$, by $P(\theta)$. The result,

$$P(\theta)-\sum_p x_p^2\prod_{q\neq p}(\theta+a_q^2)=\prod_i(\theta-\xi_i)\,, \tag{A.288}$$

is a monic cubic in θ whose roots are by definition ξ_1,ξ_2,ξ_3, which identifies it with the right-hand side. Evaluating Equation (A.288) at $\theta=-a_p^2$ kills P and every term of the sum except the p-th, giving $-x_p^2\prod_{q\neq p}(a_q^2-a_p^2)=-\prod_i(\xi_i+a_p^2)$, which is Equation (A.286) after the two sign changes in the denominator.

Differentiating Equation (A.286) with respect to ξ_i removes one factor from the numerator: $2x_p\,\partial_i x_p=x_p^2/(\xi_i+a_p^2)$, hence

$$\partial_i x_p=\frac{x_p}{2(\xi_i+a_p^2)}\,. \tag{A.289}$$

Divide Equation (A.288) by $P(\theta)$ and rearrange:

$$G(\theta)=\sum_p\frac{x_p^2}{\theta+a_p^2}=1-\frac{N(\theta)}{P(\theta)}\,,\qquad N(\theta)=\prod_i(\theta-\xi_i)\,. \tag{A.290}$$

For $i\neq j$ the inner product of the two coordinate tangent vectors is, by Equation (A.289) and the partial fraction $\big[(\xi_i+a)(\xi_j+a)\big]^{-1}=(\xi_j-\xi_i)^{-1}\big[(\xi_i+a)^{-1}-(\xi_j+a)^{-1}\big]$,

$$\sum_p\partial_i x_p\,\partial_j x_p=\frac14\sum_p\frac{x_p^2}{(\xi_i+a_p^2)(\xi_j+a_p^2)}=\frac{G(\xi_i)-G(\xi_j)}{4(\xi_j-\xi_i)}=0\,, \tag{A.291}$$

because $N(\xi_i)=N(\xi_j)=0$ makes $G(\xi_i)=G(\xi_j)=1$ by Equation (A.290): the system is orthogonal. For the scale factors, Equation (A.289) gives $h_i^2=\frac14\sum_p x_p^2(\xi_i+a_p^2)^{-2}=-\frac14 G'(\xi_i)$, and differentiating Equation (A.290),

$$G'(\theta)=-\frac{N'(\theta)P(\theta)-N(\theta)P'(\theta)}{P(\theta)^2}\,,\qquad G'(\xi_i)=-\frac{N'(\xi_i)}{P(\xi_i)}=-\frac{\prod_{l\neq i}(\xi_i-\xi_l)}{P(\xi_i)}\,, \tag{A.292}$$

using $N(\xi_i)=0$. That is Equation (A.287). Positivity follows from Equation (A.285): $P(\xi_1)>0$ and $\prod_{l\neq1}(\xi_1-\xi_l)>0$; $P(\xi_2)<0$ and $\prod_{l\neq2}(\xi_2-\xi_l)<0$; $P(\xi_3)>0$ and $\prod_{l\neq3}(\xi_3-\xi_l)>0$. □

Lemma A.127 (Lagrange's interpolation identity)**.** *For distinct ξ_1,ξ_2,ξ_3 and an integer $0\le p\le 2$,*

$$\sum_{i=1}^{3}\frac{\xi_i^{\,p}}{\prod_{l\neq i}(\xi_i-\xi_l)}=\delta_{p2}\,. \tag{A.293}$$

Rests on Equation (A.287).

Proof. *Derives Lemma A.127.*

The polynomial $L(\theta)=\sum_i\xi_i^{\,p}\prod_{l\neq i}\frac{\theta-\xi_l}{\xi_i-\xi_l}$ has degree at most 2 and agrees with θ^p at the three distinct points ξ_1,ξ_2,ξ_3. Since θ^p also has degree at most 2, their difference is a polynomial of degree at most 2 with three roots, hence identically zero: $L(\theta)=\theta^p$. Comparing the coefficients of θ^2 on the two sides gives Equation (A.293), the left-hand coefficient being the sum displayed and the right-hand one δ_{p2}. □

Proposition A.128 (The ellipsoidal system separates)**.** *The ellipsoidal system of Definition A.125 satisfies Equation (A.270) and Equation (A.271) with*

$$\phi_{ij}=\frac{\xi_i^{\,3-j}}{4\,P(\xi_i)}\,,\qquad S=\frac{W}{64\,P(\xi_1)P(\xi_2)P(\xi_3)}\,,\qquad W=\prod_{i<l}(\xi_i-\xi_l)\,, \tag{A.294}$$

and, with $\epsilon_1 = \epsilon_3 = +1$ *and* $\epsilon_2 = -1$ *the signs of* P *on the three intervals of Equation* (A.285),

$$f_1 = -2\sqrt{P(\xi_1)}\,, \qquad f_2 = 2\sqrt{-P(\xi_2)}\,, \qquad f_3 = 2\sqrt{P(\xi_3)}\,. \tag{A.295}$$

Only the product of the three signs is fixed, by Equation (A.271)*; the sign of an individual* f_i *is immaterial because* f_i *occurs twice in Equation* (A.268)*. The separated equations Equation* (A.268) *become the* Lamé wave equations

$$\frac{1}{\sqrt{\epsilon_i P}}\,\frac{\mathrm{d}}{\mathrm{d}\xi_i}\left(\sqrt{\epsilon_i P}\,\frac{\mathrm{d}X_i}{\mathrm{d}\xi_i}\right) + \frac{k^2\xi_i^2 + \lambda_2\xi_i + \lambda_3}{4\,P(\xi_i)}\,X_i = 0\,, \tag{A.296}$$

with P *evaluated at* ξ_i*.*

Rests on Lemmas A.126 and A.127 and Theorem A.120.

Proof. *Derives Proposition A.128.*

Each row of Φ in Equation (A.294) depends on ξ_i alone, as a Stäckel matrix must. By Equation (A.287), $1/h_i^2 = 4P(\xi_i)/\prod_{l\neq i}(\xi_i - \xi_l)$, so

$$\sum_i \frac{\phi_{ij}}{h_i^2} = \sum_i \frac{\xi_i^{3-j}}{\prod_{l\neq i}(\xi_i - \xi_l)} = \delta_{3-j,\,2} = \delta_{j1} \tag{A.297}$$

by Lemma A.127. By Step 2 of the proof of Theorem A.120 this identity is equivalent to Equation (A.270), and it also shows $S \neq 0$.

For the determinant, factor $1/(4P(\xi_i))$ out of the i-th row: $S = \left[64\prod_i P(\xi_i)\right]^{-1}\det(\xi_i^{3-j})$. Reversing the column order — one transposition, of columns 1 and 3 — turns $\det(\xi_i^{3-j})$ into $-\det(\xi_i^{j-1})$, and the latter is the Vandermonde determinant $\prod_{i<l}(\xi_l - \xi_i) = -W$; hence $\det(\xi_i^{3-j}) = W$ and S is as stated.

For Robertson, take the product of Equation (A.287) over i. The numerator is $\prod_i\prod_{l\neq i}(\xi_i - \xi_l) = \prod_{i<l}\left[-(\xi_i - \xi_l)^2\right] = -W^2$, so $g^2 = -W^2\left[64\prod_i P(\xi_i)\right]^{-1}$; the sign is consistent because $\prod_i P(\xi_i) < 0$ by Equation (A.285), and therefore

$$g = \frac{W}{8\prod_i \sqrt{\epsilon_i P(\xi_i)}}\,, \qquad S = \frac{-W}{64\prod_i \epsilon_i P(\xi_i)}\,, \tag{A.298}$$

the second because $\prod_i \epsilon_i = -1$. Dividing,

$$\frac{g}{S} = -8\,\frac{\prod_i \epsilon_i P(\xi_i)}{\prod_i \sqrt{\epsilon_i P(\xi_i)}} = -8\prod_i \sqrt{\epsilon_i P(\xi_i)} = \prod_i f_i(\xi_i)\,, \tag{A.299}$$

the last step by Equation (A.295), whose three signs multiply to -1: a product of three functions of one variable each, which is exactly Equation (A.271). Substituting f_i and $\sum_j \lambda_j\phi_{ij} = (k^2\xi_i^2 + \lambda_2\xi_i + \lambda_3)/(4P(\xi_i))$ into Equation (A.268) gives Equation (A.296), the constant factor 2 in f_i cancelling between the two occurrences. □

Remark A.129. For $k = 0$, Equation (A.296) is Lamé's equation and its polynomial solutions are the ellipsoidal harmonics, the analogue for a triaxial body of the spherical harmonics Equation (9.163). The chapter's three examples are visible in Equation (A.296) as limits: when two of the a_p^2 coincide one of the three coordinates becomes an angle and the corresponding Lamé equation degenerates to the associated Legendre equation Equation (9.157); when all three coincide the remaining two degenerate to the Legendre equation and to $\Psi'' + m^2\Psi = 0$, and Example A.124 is recovered.

A.15.5 The enumeration

What remains is to count the systems. The count has two halves, and only one of them is proved here.

Theorem A.130 (Quoted: separable systems are confocal-quadric systems)**.** *Let (q^i) be an orthogonal coordinate system of $\mathbb{R}^3$ in which the Helmholtz equation separates simply. Then, up to a Euclidean motion and a relabelling and reparametrization of the coordinates, the coordinate surfaces are the members of a confocal family of quadrics Equation* (A.284) *or of one of its degenerate limits.*

Remark A.131 (What is quoted here)*.* Theorem A.130 is the only statement in this section that is not proved. It is the substantial half of the classification: the assertion that a metric of flat space admitting a Stäckel matrix must have confocal quadrics for its coordinate surfaces. Its proof is a long analysis of the second-order system that Equation (A.270) imposes on the nine functions ϕ_{ij} together with the vanishing of the Riemann tensor of the Euclidean metric written in the coordinates q^i — a computation belonging to classical differential geometry rather than to the theory of partial differential equations, and one this treatise does not carry out anywhere. The result and the resulting tables of separated equations, system by system, are those of Morse and Feshbach [1023]; the differential-geometric analysis behind it is Eisenhart's, whose paper is not among the sources catalogued in this treatise's bibliography, so the attribution here is uncited and the reader who wishes to check it must go to the secondary account just named. Everything else in this section — the two conditions, their necessity and sufficiency, the ellipsoidal verification and the degeneration analysis below — is derived in full.

Granting Theorem A.130, the enumeration becomes a finite case analysis over the ways the confocal family Equation (A.284) can degenerate, and *that* we carry out. The family is fixed by the cubic P of Equation (A.283), i.e. by the three parameters $a_1^2 > a_2^2 > a_3^2$; a common shift of all three is a reparametrization $\theta \to \theta + c$ and changes nothing, so only the two differences $a_1^2 - a_2^2$ and $a_2^2 - a_3^2$ matter, and a common rescaling fixes the unit of length. The degenerations are then exhausted by three independent binary choices:

1. *Coincidence.* Either both differences are nonzero (three distinct foci), or exactly one vanishes (an axis of rotational symmetry appears, and there are two inequivalent ways for this to happen — the focal set degenerates to a segment when $a_2^2 = a_3^2$ and to a disc when $a_1^2 = a_2^2$), or both vanish (full rotational symmetry: concentric spheres).

2. *A focus at infinity.* One may let one root of P recede, $a_1^2 \to \infty$, while rescaling; the ellipsoids and one family of hyperboloids open up into paraboloids. This can be done to the generic family and to the rotationally symmetric one, but not to the spherical one, where there is nothing left to send away.

3. *Translation.* One may instead let the whole configuration become independent of one Cartesian direction, so that the quadrics become cylinders over a two-dimensional confocal family of conics. The two-dimensional families are themselves classified by the same first two choices, one dimension down: confocal conics with two distinct foci, confocal parabolas, concentric circles, or the degenerate family of parallel lines.

Reading the choices off gives the list, in which the first seven are the genuinely three-dimensional families and the last four the cylindrical ones.

Proposition A.132 (The count)**.** *The degeneration scheme just described yields exactly the eleven systems of Table A.1, and no others.*

Rests on Proposition A.128 and Theorem A.130.

Table A.1: The eleven orthogonal systems in which the three-dimensional Helmholtz equation separates simply, organized by the degeneration of the confocal quadric family that produces them. That each entry does separate is checked by exhibiting its Stäckel matrix, as done in the text for the Cartesian, cylindrical, spherical and ellipsoidal systems; that the list is complete rests on the one quoted classification theorem of this section.

System	Coordinate surfaces	Degeneration
Ellipsoidal	three confocal quadrics	none: $a_1^2 > a_2^2 > a_3^2$
Paraboloidal	elliptic and hyperbolic paraboloids	one focus to infinity
Prolate spheroidal	prolate spheroids, two-sheeted hyperboloids, half-planes	$a_2^2 = a_3^2$ (focal segment)
Oblate spheroidal	oblate spheroids, one-sheeted hyperboloids, half-planes	$a_1^2 = a_2^2$ (focal disc)
Parabolic	two families of paraboloids of revolution, half-planes	rotational, one focus to infinity
Spherical	spheres, cones of revolution, half-planes	$a_1^2 = a_2^2 = a_3^2$
Conical	spheres and two families of elliptic cones	spheres with the focal cone retained
Elliptic cylindrical	confocal elliptic and hyperbolic cylinders, planes	translational, two foci
Parabolic cylindrical	confocal parabolic cylinders, planes	translational, one focus at infinity
Circular cylindrical	circular cylinders, half-planes, planes	translational, coincident foci
Cartesian	three families of parallel planes	translational, all foci at infinity

Proof. *Derives Proposition A.132.*

Consider first the non-translational families. The coincidence pattern of the three parameters has four cases: all distinct; $a_2^2 = a_3^2$; $a_1^2 = a_2^2$; all equal. In the first three cases one may additionally send a focus to infinity or not, which doubles them — but the two rotationally symmetric cases have the *same* paraboloidal limit, because sending the distinct root to infinity destroys the distinction between a focal segment and a focal disc, leaving the single parabolic system. That gives $1 + 1$ (ellipsoidal, paraboloidal) plus $2 + 1$ (prolate, oblate, parabolic), i.e. five. The fully symmetric case admits no focus at infinity, but it does admit two inequivalent coordinate systems on the sphere: the one whose second and third coordinates are the polar angle and the azimuth — spherical — and the one in which the sphere is cut by two families of elliptic cones sharing a vertex, which is the conical system and is the residue of the ellipsoidal system when the quadrics have collapsed but their asymptotic cones have not. That gives seven.

The translational families are cylinders over a two-dimensional confocal family of conics, and the same analysis one dimension down applies to the plane cross-section: two distinct foci give confocal ellipses and hyperbolae (elliptic cylindrical); sending one focus to infinity gives confocal parabolae (parabolic cylindrical); coincident foci give concentric circles and their radii (circular cylindrical); and sending both to infinity gives two families of parallel lines (Cartesian). No further case arises, because a family of conics in the plane is fixed by the two foci alone, and their configuration — distinct, coincident, one at infinity, both at infinity — has exactly these four types. That gives four, and $7 + 4 = 11$. □

Proof of Theorem 10.34. *Derives Theorem 10.34.*

The criterion asserted by the theorem is Theorem A.120: separation holds if and only if the reciprocal squared scale factors are expressible through a matrix each of whose rows depends on one coordinate alone — explicitly, if and only if $1/h_i^2 = M_{i1}/S$ in Equation (A.270), together with the Robertson condition Equation (A.271) on the volume element, which the chapter's statement subsumes under "the Stäckel conditions". The list of eleven systems is Table A.1: that each of them separates is verified by exhibiting its Stäckel matrix, done in Examples A.122 to A.124 for the three the treatise uses and in Proposition A.128 for the generic one, from which the remaining entries follow by the confluences of Proposition A.132; and that there are no others is Proposition A.132 together with the quoted Theorem A.130. □

Remark A.133. Appendix A.15 discharges the proof obligation of Theorem 10.34 in Chapter 10. Two consequences are used elsewhere in the treatise. The first is negative and is the practical one recorded in Remark 10.35: a boundary that is not a coordinate surface of one of the eleven systems admits no separation, which is why the boundary chooses the coordinates and not the operator. The second is that Equation (A.270) alone — without Robertson's Equation (A.271) — is the condition for the Hamilton–Jacobi equation of Chapter 23 to separate, so the separation constants λ_2, λ_3 of Equation (A.268) are the quantum counterparts of the classical integrals of motion of an integrable system; the spherical case (Example A.124), where they are $l(l+1)$ and m^2, is the one worked out physically in Chapter 81.

A.16 The Mean-Value Property Characterizes Harmonic Functions

This appendix proves the continuous form of the converse half of Theorem 10.69 (Chapter 10): a function that is merely *continuous* on an open set and equals its own spherical average over every ball whose closure the set contains is automatically infinitely differentiable there, and harmonic. The chapter proves the converse only for $u \in C^2$, because the argument it uses

decides the sign of a Laplacian whose existence is exactly what is in question (Remark 10.70). The gap is closed here by constructing the smoothing device the chapter does not have — a radial mollifier — and showing that convolution against it returns the function unchanged. Once that is done the chapter's own sign argument applies verbatim.

Nothing is quoted. Everything below is built from the Riemann integral of a continuous function over a compact region, the mean value theorem of one-variable calculus, and the polar decomposition of a volume integral in $\mathbb{R}^3$ already used in Lemma 10.58 and in the proof of Theorem 10.69; in particular the Lebesgue theory that Chapter 7 deliberately does not develop is nowhere needed, because every integrand that appears is continuous with compact support.

Throughout, $\Omega \subset \mathbb{R}^3$ is open and nonempty, $B(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathbb{R}^3 \mid |\boldsymbol{y} - \boldsymbol{x}| < r\}$, and $M_u(\boldsymbol{x}, r)$ is the spherical mean Equation (10.54). The space is the observed three-dimensional one; nothing in the argument depends on that choice beyond the numerical factor 4π.

A.16.1 Statement

Definition A.134 (Mean-value property)**.** A function $u \in C^0(\Omega)$ has the *mean-value property* on Ω if and only if

$$u(\boldsymbol{x}) = M_u(\boldsymbol{x}, r) = \frac{1}{4\pi} \int_{S^2} u(\boldsymbol{x} + r\boldsymbol{n}) \, \mathrm{d}\Omega \tag{A.300}$$

for every $\boldsymbol{x} \in \Omega$ and every $r > 0$ with $\overline{B(\boldsymbol{x}, r)} \subset \Omega$.

Rests on Equation (10.54) *and Definition 6.2.*

Theorem A.135 (Converse of the mean-value property)**.** *Let $u \in C^0(\Omega)$ have the mean-value property of Definition A.134. Then $u \in C^\infty(\Omega)$ and $\nabla^2 u = 0$ on Ω; that is, u is harmonic in the sense of Definition 10.66. Consequently the two conditions* harmonic *and* continuous with the mean-value property *define the same class of functions.*

Rests on Definitions 10.66 and A.134 and Theorem 10.69.

The proof occupies the rest of the section: an elementary lemma on differentiating an integral with respect to a parameter (Appendix A.16.2), the construction of the mollifier and the smoothness of the convolution (Appendix A.16.3), the reproduction identity $u * \rho_\varepsilon = u$ (Appendix A.16.4), and the assembly (Appendix A.16.5).

A.16.2 Differentiation under the integral sign

The one analytic tool the argument needs is stated and proved here, in the only form it will be used: the integrand is continuous, the domain of integration is a fixed compact set, and the parameter runs over an open set.

Lemma A.136 (Differentiation under the integral sign)**.** *Let $U \subset \mathbb{R}^3$ be open, let $Q \subset \mathbb{R}^3$ be a compact box, and let $F : U \times Q \longrightarrow \mathbb{R}$ be such that F and the partial derivative $\partial F / \partial x^i$ (taken in the first argument) are continuous on $U \times Q$ for one fixed i. Then*

$$g(\boldsymbol{x}) = \int_Q F(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}^3 y \tag{A.301}$$

has a continuous partial derivative $\partial_i g$ on U, and

$$\partial_i g(\boldsymbol{x}) = \int_Q \frac{\partial F}{\partial x^i}(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}^3 y \, . \tag{A.302}$$

Rests on Definition 7.66 and Theorem 6.11.

Proof. *Derives Lemma A.136.*

Fix $\boldsymbol{x} \in U$ and choose $r > 0$ with $\overline{B(\boldsymbol{x}, r)} \subset U$. The set $\overline{B(\boldsymbol{x}, r)} \times Q$ is compact, so $\partial F / \partial x^i$ is uniformly continuous on it: given $\varepsilon > 0$ there is $\delta \in (0, r)$ such that

$$|\boldsymbol{x}' - \boldsymbol{x}''| < \delta \quad \Longrightarrow \quad \left| \frac{\partial F}{\partial x^i}(\boldsymbol{x}', \boldsymbol{y}) - \frac{\partial F}{\partial x^i}(\boldsymbol{x}'', \boldsymbol{y}) \right| < \varepsilon \tag{A.303}$$

for every $\boldsymbol{y} \in Q$, the bound being uniform in $\boldsymbol{y}$ — which is the whole content of *uniform* continuity on the product and the reason the compactness of Q is needed.

Let $\boldsymbol{e}_i$ be the i-th coordinate vector and let $0 < |h| < \delta$. For each fixed $\boldsymbol{y}$ the mean value theorem applied to the function $s \longmapsto F(\boldsymbol{x} + s\boldsymbol{e}_i, \boldsymbol{y})$ on the interval between 0 and h supplies $\theta = \theta(h, \boldsymbol{y}) \in (0, 1)$ with

$$\frac{F(\boldsymbol{x} + h\boldsymbol{e}_i, \boldsymbol{y}) - F(\boldsymbol{x}, \boldsymbol{y})}{h} = \frac{\partial F}{\partial x^i}(\boldsymbol{x} + \theta h \boldsymbol{e}_i, \boldsymbol{y}) \,. \tag{A.304}$$

Both sides of Equation (A.304) are continuous in $\boldsymbol{y}$ — the left side manifestly, and hence the right side as well, whatever the measurability of θ — so both are Riemann integrable over Q. Subtracting Equation (A.302) and using Equation (A.303) with $|\theta h| < \delta$,

$$\left| \frac{g(\boldsymbol{x} + h\boldsymbol{e}_i) - g(\boldsymbol{x})}{h} - \int_Q \frac{\partial F}{\partial x^i}(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}^3 y \right| \leq \int_Q \varepsilon \, \mathrm{d}^3 y = \varepsilon \, |Q| \,, \tag{A.305}$$

with $|Q|$ the volume of the box. Since ε was arbitrary the difference quotient converges, which is Equation (A.302). Continuity of $\partial_i g$ follows from the same estimate with $\boldsymbol{x}'' = \boldsymbol{x}$ and $\boldsymbol{x}'$ a nearby point, the integral of a quantity bounded by ε being bounded by $\varepsilon \, |Q|$. □

Remark A.137. Lemma A.136 iterates: if F and all its partial derivatives in $\boldsymbol{x}$ of every order are continuous on $U \times Q$, then $g \in C^\infty(U)$ and every derivative may be taken under the integral sign, because the conclusion of the lemma is again a function of the same form with $\partial F / \partial x^i$ in place of F.

A.16.3 A radial mollifier

The device the chapter lacks is a smooth, radial, non-negative function supported in a ball and of unit integral. Its existence is not obvious: a function that is analytic cannot vanish on an open set without vanishing identically, so the construction must use a function that is smooth and *not* analytic. There is exactly one standard source of such a function.

Lemma A.138 (The flat exponential)**.** *Define* $f : \mathbb{R} \longrightarrow \mathbb{R}$ *by*

$$f(t) = \begin{cases} \mathrm{e}^{-1/t} \,, & t > 0 \,, \\ 0 \,, & t \leq 0 \,. \end{cases} \tag{A.306}$$

Then $f \in C^\infty(\mathbb{R})$, $f^{(n)}(0) = 0$ *for every* n, *and* $f(t) > 0$ *exactly for* $t > 0$.
Rests on Definition 7.66 and Theorem 7.38.

Proof. *Derives Lemma A.138.*

Derivatives for $t > 0$. We claim that for every $n \geq 0$ there is a polynomial p_n with

$$f^{(n)}(t) = p_n\!\left(\frac{1}{t}\right) \mathrm{e}^{-1/t} \,, \qquad t > 0 \,. \tag{A.307}$$

For $n = 0$ take $p_0 = 1$. Differentiating Equation (A.307) and using $\mathrm{d}(1/t)/\mathrm{d}t = -1/t^2$,

$$f^{(n+1)}(t) = \left[-\frac{1}{t^2} \, p_n'\!\left(\frac{1}{t}\right) + \frac{1}{t^2} \, p_n\!\left(\frac{1}{t}\right) \right] \mathrm{e}^{-1/t} \,, \tag{A.308}$$

so Equation (A.307) holds at $n+1$ with $p_{n+1}(s) = s^2\big(p_n(s) - p_n'(s)\big)$.

Decay at the origin. For $s > 0$ and any integer $m \geq 0$ the exponential series gives $\mathrm{e}^s \geq s^{m+1}/(m+1)!$, hence $s^m \mathrm{e}^{-s} \leq (m+1)!/s \longrightarrow 0$ as $s \to +\infty$. Consequently, for any polynomial p,

$$\lim_{t \to 0^+} p\left(\frac{1}{t}\right) \mathrm{e}^{-1/t} = \lim_{s \to +\infty} p(s)\, \mathrm{e}^{-s} = 0\,. \tag{A.309}$$

Derivatives at the origin. We show by induction that $f^{(n)}$ exists everywhere, is continuous, and $f^{(n)}(0) = 0$. For $n = 0$ this is Equation (A.309) with $p = 1$. Assume it at n. For $t < 0$ the difference quotient of $f^{(n)}$ at 0 vanishes identically; for $t > 0$ it is

$$\frac{f^{(n)}(t) - f^{(n)}(0)}{t} = \frac{1}{t}\, p_n\left(\frac{1}{t}\right) \mathrm{e}^{-1/t} = \big[s\, p_n(s)\big]\, \mathrm{e}^{-s}\,, \qquad s = \frac{1}{t}\,, \tag{A.310}$$

which tends to 0 by Equation (A.309) applied to the polynomial $s\, p_n(s)$. Hence $f^{(n+1)}(0)$ exists and is 0. It is continuous at 0 because $f^{(n+1)}(t) = p_{n+1}(1/t)\mathrm{e}^{-1/t} \to 0$ as $t \to 0^+$ by Equation (A.309) again, and vanishes for $t < 0$. Positivity for $t > 0$ is immediate. □

Definition A.139 (The standard radial mollifier)**.** With f of Equation (A.306) set

$$\rho(\boldsymbol{x}) = c_0\, f\Big(1 - |\boldsymbol{x}|^2\Big)\,, \qquad c_0^{-1} = \int_{\mathbb{R}^3} f\Big(1 - |\boldsymbol{x}|^2\Big)\, \mathrm{d}^3x\,, \tag{A.311}$$

and, for $\varepsilon > 0$, $\rho_\varepsilon(\boldsymbol{x}) = \varepsilon^{-3}\rho(\boldsymbol{x}/\varepsilon)$.

Rests on Lemma A.138.

Lemma A.140 (Properties of the mollifier)**.** *$\rho_\varepsilon \in C^\infty(\mathbb{R}^3)$; it is non-negative, vanishes outside $\overline{B(\mathbf{0}, \varepsilon)}$, depends on $\boldsymbol{x}$ only through $|\boldsymbol{x}|$, and $\int_{\mathbb{R}^3} \rho_\varepsilon\, \mathrm{d}^3x = 1$.*

Rests on Definition A.139 and Lemma A.138.

Proof. *Derives Lemma A.140.*

$|\boldsymbol{x}|^2 = x^2 + y^2 + z^2$ is a polynomial, hence C^∞, and the composition of a C^∞ function of one variable with a C^∞ function of three is C^∞ by the chain rule (Proposition 7.72); this is where Lemma A.138 is used, and it is the only place where smoothness that is not analyticity is needed. The argument $1 - |\boldsymbol{x}|^2$ is positive exactly on the open unit ball and non-positive outside it, so $\rho > 0$ there and $\rho = 0$ elsewhere; in particular ρ vanishes on a whole neighbourhood of every point of $|\boldsymbol{x}| = 1$ together with all its derivatives. The normalizing integral in Equation (A.311) is the integral of a continuous non-negative function that is positive on an open set, hence finite and strictly positive, so c_0 is well defined. Radial dependence is manifest. Finally the substitution $\boldsymbol{x} = \varepsilon \boldsymbol{z}$, whose Jacobian is ε^3, gives $\int \rho_\varepsilon(\boldsymbol{x}) \mathrm{d}^3x = \int \rho(\boldsymbol{z}) \mathrm{d}^3z = 1$, and rescales the support to $|\boldsymbol{x}| \leq \varepsilon$. □

Definition A.141 (Mollification)**.** For $u \in C^0(\Omega)$ and $\varepsilon > 0$ put

$$\Omega_\varepsilon = \Big\{ \boldsymbol{x} \in \Omega \mid \overline{B(\boldsymbol{x}, \varepsilon)} \subset \Omega \Big\}\,, \tag{A.312}$$

an open subset of Ω whose union over $\varepsilon > 0$ is Ω, and define on it

$$u_\varepsilon(\boldsymbol{x}) = \int_{B(\mathbf{0},\varepsilon)} \rho_\varepsilon(\boldsymbol{z})\, u(\boldsymbol{x} - \boldsymbol{z})\, \mathrm{d}^3z = \int_{B(\boldsymbol{x},\varepsilon)} \rho_\varepsilon(\boldsymbol{x} - \boldsymbol{y})\, u(\boldsymbol{y})\, \mathrm{d}^3y\,. \tag{A.313}$$

Rests on Definitions 6.2 and A.139.

The two integrals in Equation (A.313) agree by the substitution $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{z}$, of unit Jacobian; both integrands are continuous on a compact set, so both are ordinary Riemann integrals. That Ω_ε is open, and that every point of Ω lies in some Ω_ε, is immediate from Ω being open.

Proposition A.142 (The mollification is smooth)**.** *$u_\varepsilon \in C^\infty(\Omega_\varepsilon)$, and every partial derivative may be taken under the integral sign in the second form of Equation* (A.313)*.*

Rests on Lemmas A.136 and A.140.

Proof. *Derives Proposition A.142.*

Fix $\boldsymbol{x}_0 \in \Omega_\varepsilon$ and $\sigma > 0$ so small that $\overline{B(\boldsymbol{x}_0, \sigma+\varepsilon)} \subset \Omega$; let $U = B(\boldsymbol{x}_0, \sigma)$ and let Q be a closed box containing $\overline{B(\boldsymbol{x}_0, \sigma+\varepsilon)}$. Extend u from $\overline{B(\boldsymbol{x}_0, \sigma+\varepsilon)}$ to Q by any continuous function — the extension is irrelevant, because for $\boldsymbol{x} \in U$ the factor $\rho_\varepsilon(\boldsymbol{x}-\boldsymbol{y})$ already vanishes for $\boldsymbol{y}$ outside $\overline{B(\boldsymbol{x}_0, \sigma+\varepsilon)}$, so

$$u_\varepsilon(\boldsymbol{x}) = \int_Q F(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}^3 y\,, \qquad F(\boldsymbol{x}, \boldsymbol{y}) = \rho_\varepsilon(\boldsymbol{x}-\boldsymbol{y})\, u(\boldsymbol{y})\,, \tag{A.314}$$

for every $\boldsymbol{x} \in U$. In Equation (A.314) the whole dependence on the parameter $\boldsymbol{x}$ sits in the factor ρ_ε, which by Lemma A.140 is C^∞; hence F and all its $\boldsymbol{x}$-derivatives of every order,

$$\partial_{\boldsymbol{x}}^\alpha F(\boldsymbol{x}, \boldsymbol{y}) = (\partial^\alpha \rho_\varepsilon)(\boldsymbol{x}-\boldsymbol{y})\, u(\boldsymbol{y})\,, \tag{A.315}$$

are continuous on $U \times Q$, being products of continuous functions. Lemma A.136 and Remark A.137 therefore apply and give $u_\varepsilon \in C^\infty(U)$ with $\partial^\alpha u_\varepsilon(\boldsymbol{x}) = \int_Q (\partial^\alpha \rho_\varepsilon)(\boldsymbol{x}-\boldsymbol{y})\, u(\boldsymbol{y})\, \mathrm{d}^3 y$. Since $\boldsymbol{x}_0 \in \Omega_\varepsilon$ was arbitrary, the conclusion holds on all of Ω_ε. □

A.16.4 The mollification reproduces the function

This is the step at which the mean-value hypothesis enters, and it is the reason the mollifier was required to be *radial*.

Proposition A.143 (Reproduction identity)**.** *Let $u \in C^0(\Omega)$ have the mean-value property. Then*

$$u_\varepsilon(\boldsymbol{x}) = u(\boldsymbol{x}) \qquad \text{for every } \boldsymbol{x} \in \Omega_\varepsilon \text{ and every } \varepsilon > 0\,. \tag{A.316}$$

Rests on Definitions A.134 and A.141 and Lemma A.140.

Proof. *Derives Proposition A.143.*

Write $\rho_\varepsilon(\boldsymbol{z}) = \varepsilon^{-3}\tilde{\rho}(|\boldsymbol{z}|/\varepsilon)$, where $\tilde{\rho}(s) = c_0 f(1-s^2)$ is the radial profile supplied by Lemma A.140. Decompose the first integral of Equation (A.313) into spheres, $\mathrm{d}^3 z = s^2\, \mathrm{d}s\, \mathrm{d}\Omega$ — the same decomposition already used in Lemma 10.58 to pass between a solid and a spherical integral, and legitimate here because the integrand is continuous on the compact ball:

$$u_\varepsilon(\boldsymbol{x}) = \int_0^\varepsilon \varepsilon^{-3}\tilde{\rho}\!\left(\frac{s}{\varepsilon}\right) s^2 \left[\int_{S^2} u(\boldsymbol{x} - s\boldsymbol{n})\, \mathrm{d}\Omega\right] \mathrm{d}s\,. \tag{A.317}$$

The inner integral is unchanged under $\boldsymbol{n} \longmapsto -\boldsymbol{n}$, which is a symmetry of the unit sphere and of its solid-angle element, so it equals $\int_{S^2} u(\boldsymbol{x}+s\boldsymbol{n})\, \mathrm{d}\Omega = 4\pi\, M_u(\boldsymbol{x}, s)$. For $\boldsymbol{x} \in \Omega_\varepsilon$ and $0 < s \le \varepsilon$ the closed ball $\overline{B(\boldsymbol{x}, s)}$ lies in Ω, so the mean-value property Equation (A.300) gives $M_u(\boldsymbol{x}, s) = u(\boldsymbol{x})$ — a constant, which comes out of the s-integral:

$$u_\varepsilon(\boldsymbol{x}) = u(\boldsymbol{x}) \int_0^\varepsilon 4\pi s^2\, \varepsilon^{-3}\tilde{\rho}\!\left(\frac{s}{\varepsilon}\right) \mathrm{d}s = u(\boldsymbol{x}) \int_{\mathbb{R}^3} \rho_\varepsilon(\boldsymbol{z})\, \mathrm{d}^3 z = u(\boldsymbol{x})\,, \tag{A.318}$$

the middle step being Equation (A.317) read backwards with $u \equiv 1$, and the last being the normalization of Lemma A.140. □

Remark A.144 (Why radial, and why the normalization matters). Both hypotheses on ρ are used and neither can be dropped. Radial dependence is what lets the weight ρ_ε be pulled outside the *angular* integral in Equation (A.317), leaving exactly a spherical mean for the hypothesis to act on; a non-radial bump would leave an angular weight and the mean-value property would say nothing about the result. Unit mass is what makes the surviving factor

equal to 1 rather than to some other constant. The identity Equation (A.316) is thus not an approximation statement — it is not $u_\varepsilon \to u$, which holds for every continuous u — but an exact equality, valid for each ε separately, and that exactness is the whole content of the argument.

A.16.5 Proof of the theorem

Proof of Theorem A.135. *Derives Theorem A.135.*

Smoothness. Let $\boldsymbol{x}_0 \in \Omega$. Since Ω is open there is $\varepsilon > 0$ with $\overline{B(\boldsymbol{x}_0, 2\varepsilon)} \subset \Omega$, and then $B(\boldsymbol{x}_0, \varepsilon) \subset \Omega_\varepsilon$. On Ω_ε we have $u = u_\varepsilon$ by Proposition A.143, and $u_\varepsilon \in C^\infty(\Omega_\varepsilon)$ by Proposition A.142. Hence u is C^∞ on a neighbourhood of $\boldsymbol{x}_0$. Smoothness being a local property and $\boldsymbol{x}_0$ arbitrary, $u \in C^\infty(\Omega)$. Note that the smoothing parameter ε is allowed to depend on the point, as it must: no single ε works near $\partial\Omega$.

Harmonicity. Now that $u \in C^2(\Omega)$ is established, the argument of Theorem 10.69 applies without change, and we repeat it for completeness. Suppose $\nabla^2 u(\boldsymbol{x}_0) > 0$ for some $\boldsymbol{x}_0 \in \Omega$. By continuity of $\nabla^2 u$ there is $r > 0$ with $\overline{B(\boldsymbol{x}_0, r)} \subset \Omega$ and $\nabla^2 u > 0$ throughout that ball. Darboux's identity Equation (10.55) then gives, for $0 < s \le r$,

$$\partial_s M_u(\boldsymbol{x}_0, s) = \frac{1}{4\pi s^2} \int_{B(\boldsymbol{x}_0, s)} \nabla^2 u \, \mathrm{d}^3 y > 0 \,, \tag{A.319}$$

so $s \longmapsto M_u(\boldsymbol{x}_0, s)$ is strictly increasing on $(0, r]$. But the mean-value property makes it constant, equal to $u(\boldsymbol{x}_0)$, on that whole interval — a contradiction. The case $\nabla^2 u(\boldsymbol{x}_0) < 0$ is the same argument with the inequalities reversed. Hence $\nabla^2 u \equiv 0$ on Ω, which is Definition 10.66.

The two classes coincide. A harmonic function is C^2 by definition and has the mean-value property by the direct half of Theorem 10.69; conversely a continuous function with the mean-value property is harmonic by what has just been proved. □

Corollary A.145 (Harmonic functions are smooth)**.** *Every harmonic function on Ω is of class $C^\infty(\Omega)$, although its definition (Definition 10.66) demands only C^2.*

Rests on Theorems 10.69 and A.135.

Proof. *Derives Corollary A.145.*

A harmonic u is continuous and has the mean-value property (Theorem 10.69); Theorem A.135 then returns $u \in C^\infty(\Omega)$. □

Corollary A.146 (Locally uniform limits of harmonic functions)**.** *If u_n are harmonic on Ω and $u_n \to u$ uniformly on every compact subset of Ω, then u is harmonic.*

Rests on Theorem A.135 and Definition A.134.

Proof. *Derives Corollary A.146.*

u is continuous, being a locally uniform limit of continuous functions. Fix $\boldsymbol{x}$ and r with $\overline{B(\boldsymbol{x}, r)} \subset \Omega$; the sphere $|\boldsymbol{y} - \boldsymbol{x}| = r$ is compact, so $u_n \to u$ uniformly on it and the averages converge: $M_u(\boldsymbol{x}, r) = \lim_n M_{u_n}(\boldsymbol{x}, r) = \lim_n u_n(\boldsymbol{x}) = u(\boldsymbol{x})$. Thus u is continuous with the mean-value property, and Theorem A.135 applies. The corresponding statement with C^2 convergence would be trivial; the point is that mere uniform convergence suffices, and it does so only because the characterization proved here does not mention derivatives at all. □

Remark A.147 (What is quoted here)*.* Nothing. The three inputs are the mean value theorem of one-variable calculus, uniform continuity of a continuous function on a compact set (Theorem 6.11), and the polar decomposition $\mathrm{d}^3 z = s^2 \, \mathrm{d}s \, \mathrm{d}\Omega$ of a volume integral, all of them already in force in Chapter 7 and used by the chapter itself in Lemma 10.58. Every integrand appearing above is continuous on a compact region, so the Riemann integral of Chapter 7 suffices throughout and no appeal to Lebesgue's theory — which this treatise does

not build — is made or needed. In particular Lemma A.136 is proved, not cited: the usual statement of differentiation under the integral sign is a dominated-convergence argument, and the compactness of the domain of integration replaces the domination here.

Remark A.148 (The scope of the improvement). The gain over Theorem 10.69 is not cosmetic. The C^2 converse cannot be applied to a function one has only constructed as a limit, an average or a supremum, because such a function is typically known to be continuous and nothing more; the continuous converse can, and Corollary A.146 is the first instance. It is also the cleanest statement of the rigidity of the Laplace equation: the second-order differential condition $\nabla^2 u = 0$ is equivalent to a condition — Equation (A.300) — that mentions no derivative whatever, and *that* is why solutions of it are automatically smooth while solutions of, say, the wave equation are not.

Remark A.149. Appendix A.16 discharges the proof obligation left open in Remark 10.70 and completes Theorem 10.69 of Chapter 10. The chapter uses only the C^2 form in what follows — the strong maximum principle Theorem 10.77 is proved from it — so nothing there depended on the present section; what the section adds is the equivalence that makes the mean-value property a *definition* of harmonicity rather than a consequence of one, and with it the stability under uniform limits recorded in Corollary A.146. The mollifier constructed in Appendix A.16.3 is the same device that underlies the smooth test functions of Chapter 17 and the approximation arguments of Section 10.6.2.

A.17 Kruzhkov's Doubling of Variables and the Entropy Solution

This appendix proves Theorem 10.103 of Chapter 10: a scalar conservation law Equation (10.81) with bounded measurable datum has at most one entropy solution in the sense of Definition 10.101, two entropy solutions contract in the mean, and at least one exists. The uniqueness half is Kruzhkov's argument by *doubling of variables*, and it is the whole content of the section; the existence half is the vanishing-viscosity limit already announced in Remark 10.100, whose architecture is set out in Appendix A.17.6 with its two compactness inputs named.

Three features of the argument are worth flagging before it begins, because each is what makes it work. The first is the choice of entropy: the family $\eta(u) = |u - k|$, indexed by a real constant k, is not C^2 and so is not admitted by Definition 10.101 directly; it is reached by approximation in Appendix A.17.3, and it is the family for which the entropy flux of two solutions can be compared. The second is that the comparison is carried out in *four* variables — u is evaluated at (t, x) and v at (s, y) — with each solution's own entropy inequality tested against the same function and the constant k taken to be the *other* solution. The third is the cancellation that makes the whole thing finite: with a test function of the form $\chi(\frac{t+s}{2}, \frac{x+y}{2})$ times a mollifier in the differences, the two entropy inequalities add so that the mollifier is differentiated in neither of them.

Throughout, $F \in C^1(\mathbb{R})$, u and v are entropy solutions of Equation (10.81) with bounded measurable data u_0, v_0,

$$|u|, |v| \le M_0 \quad \text{almost everywhere}, \qquad M = \max_{|w| \le M_0} |F'(w)| , \tag{A.320}$$

and φ denotes a non-negative test function of the kind appearing in Equation (10.92). Convexity of F, assumed in Theorem 10.103, is nowhere used below: Kruzhkov's theorem needs only that F be C^1, which is a strictly stronger result and is recorded in Remark A.165.

A.17.1 Statement

Definition A.150 (Kruzhkov entropy pair)**.** For $k \in \mathbb{R}$ put

$$\eta_k(w) = |w - k| \,, \qquad q_k(w) = \operatorname{sgn}(w-k)\big(F(w) - F(k)\big) \,. \tag{A.321}$$

Rests on Definitions 10.93 and 10.101.

Theorem A.151 (L^1 contraction on a cone)**.** *Let u, v be entropy solutions in the sense of Definition 10.101 obeying Equation (A.320). Then for almost every $0 < \tau_1 < \tau_2$ and every $R > 0$*

$$\int_{|x| \le R} |u(\tau_2, x) - v(\tau_2, x)| \, \mathrm{d}x \le \int_{|x| \le R + M(\tau_2 - \tau_1)} |u(\tau_1, x) - v(\tau_1, x)| \, \mathrm{d}x \,, \tag{A.322}$$

and, using Lemma A.154, the same inequality holds with $\tau_1 = 0$ and u, v replaced there by u_0, v_0. If in addition $u_0 - v_0$ is integrable on $\mathbb{R}$, then for almost every $t > 0$

$$\int_{\mathbb{R}} |u(t, x) - v(t, x)| \, \mathrm{d}x \le \int_{\mathbb{R}} |u_0(x) - v_0(x)| \, \mathrm{d}x \,. \tag{A.323}$$

Rests on Definitions 10.95, 10.101 and A.150.

Corollary A.152 (Uniqueness)**.** *An entropy solution with a given bounded measurable datum is unique up to a null set, and the solution map $u_0 \longmapsto u(t, \cdot)$ is a contraction for the L^1 distance — which is condition (iii) of Definition 10.24 with the mean as the norm.*
Rests on Theorem A.151 and Definition 10.24.

Proof. *Derives Corollary A.152.*

Let $u_0 = v_0$ almost everywhere. Then the right-hand side of Equation (A.322) with $\tau_1 = 0$ vanishes for every R, so $u(t, \cdot) = v(t, \cdot)$ almost everywhere on every bounded interval and hence on $\mathbb{R}$, for almost every t. Continuous dependence is Equation (A.323) itself, the map being 1-Lipschitz from L^1 to L^1. □

A.17.2 The inputs this treatise does not build

Remark A.153 (What is quoted here). Unlike the rest of Chapter 10, the subject of this section is not formulated in terms of the Riemann integral and cannot be. A weak solution in the sense of Definition 10.95 is a *bounded measurable* function, its defining identity is an integral over a region of the plane against an arbitrary test function, and the conclusions are statements valid *almost everywhere*; every one of these words belongs to Lebesgue's theory, which Chapter 7 does not develop — it builds the Riemann integral only, as Chapter 11 also records where it needs dominated convergence and declines to use it. The following facts are therefore assumed, and they are assumed for the definitions as much as for the proofs:

1. *The Lebesgue integral* of a bounded measurable function over a measurable subset of $\mathbb{R}^d$, its linearity and monotonicity, and the fact that a function vanishing almost everywhere has zero integral.

2. *Fubini's theorem* for a bounded measurable function on a product, used in Proposition A.159 to integrate one solution's entropy inequality over the other solution's variables, and again to change variables from (t, x, s, y) to sum and difference coordinates.

3. *Dominated convergence*, used at each passage to a limit.

4. *Continuity of translation in L^1*: for w bounded and measurable and K compact, $\int_K |w(\cdot + h) - w| \longrightarrow 0$ as $h \to 0$. Equivalently, almost every point is a Lebesgue point. This is the one substantial analytic fact of Proposition A.162, and it is exactly what replaces the continuity that a classical solution would have had.

5. *Lebesgue's differentiation theorem* in one variable, used in Appendix A.17.5 to pass from an average over a short time interval to a value at almost every instant.

6. *Kruzhkov's initial-layer lemma* (Lemma A.154): an entropy solution attains its datum in the local mean. This is the only *theorem* of the subject that is quoted rather than proved, and it is isolated as such below.

7. For the existence half only (Appendix A.17.6): classical solvability of the viscous problem, and a compactness theorem for families of uniformly bounded functions of uniformly bounded variation.

Everything else — the passage from the C^2 entropies of Definition 10.101 to the Kruzhkov family, the doubling, the cancellation, the limit, the cone estimate, and the derivation of the entropy inequality from the viscous equation — is carried out in full. The primary source is [1427], where the entropy class, the doubling of variables and the L^1 contraction first appear together.

Lemma A.154 (Quoted: the initial layer)**.** *Let u be an entropy solution in the sense of Definition 10.101 with datum u_0. Then for every compact $K \subset \mathbb{R}$*

$$\lim_{t \to 0^+} \int_K |u(t,x) - u_0(x)| \, \mathrm{d}x = 0 \,, \tag{A.324}$$

the limit being taken through the instants at which $u(t,\cdot)$ is defined.

Remark A.155. Equation (A.324) is stronger than what the definition plainly gives, and the gap is worth naming. Testing Equation (10.85) with $\varphi(x,t) = \rho(x)\alpha(t)$ shows that $t \longmapsto \int u(t,x)\rho(x)\,\mathrm{d}x$ agrees almost everywhere with an absolutely continuous function taking the value $\int u_0\rho$ at $t = 0$, so the datum is attained *weakly*; that much is a two-line computation from the definition. Weak attainment is not enough for Theorem A.151, because the functional $w \longmapsto \int |w|$ is only lower semicontinuous under weak convergence and the inequality needed runs the other way. Kruzhkov obtains Equation (A.324) from the entropy inequality itself applied with constant k and a test function concentrating at $t = 0$, which yields an upper bound on $\limsup_{t\to 0} \int |u(t,x) - k|\,\rho$ by $\int |u_0 - k|\,\rho$; a covering argument over a countable dense set of values k, together with approximation of u_0 by step functions, upgrades that to the strong statement. The argument is measure-theoretic throughout and is not reproduced here.

A.17.3 From the admissible entropies to Kruzhkov's family

Proposition A.156 (The Kruzhkov inequality)**.** *Let u be an entropy solution in the sense of Definition 10.101. Then for every $k \in \mathbb{R}$ and every non-negative test function φ,*

$$\int_0^\infty \!\! \int_{\mathbb{R}} \left[|u-k| \, \partial_t\varphi + q_k(u) \, \partial_x\varphi \right] \mathrm{d}x \, \mathrm{d}t + \int_{\mathbb{R}} |u_0(x) - k| \, \varphi(x,0) \, \mathrm{d}x \geq 0 \,, \tag{A.325}$$

with q_k of Equation (A.321).

Rests on Definitions 10.101 and A.150 and Equation (10.92).

Proof. *Derives Proposition A.156.*

For $\delta > 0$ set

$$\eta_\delta(w) = \sqrt{(w-k)^2 + \delta^2} - \delta \,, \qquad q_\delta(w) = \int_k^w \eta_\delta'(\lambda)\, F'(\lambda)\, \mathrm{d}\lambda \,. \tag{A.326}$$

Then η_δ is C^∞, with

$$\eta_\delta'(w) = \frac{w-k}{\sqrt{(w-k)^2 + \delta^2}} \,, \qquad \eta_\delta''(w) = \frac{\delta^2}{\left[(w-k)^2 + \delta^2\right]^{3/2}} > 0 \,, \tag{A.327}$$

so it is convex; q_δ is C^1 with $q'_\delta = \eta'_\delta F'$ by the fundamental theorem of calculus. Thus (η_δ, q_δ) is an admissible entropy pair in the sense of Definition 10.101, and Equation (10.92) holds for it.

Uniform convergence of the entropy. From $|w-k| \le \sqrt{(w-k)^2+\delta^2} \le |w-k| + \delta$,

$$|\eta_\delta(w) - |w-k|| \le \delta \qquad \text{for every } w\,. \tag{A.328}$$

Uniform convergence of the flux on the range. Let $|w| \le M_0$ and abbreviate $r_\delta(\lambda) = \operatorname{sgn}(\lambda-k) - \eta'_\delta(\lambda)$, so that $|r_\delta| \le 2$ and, for $|\lambda - k| \ge \varepsilon$,

$$|r_\delta(\lambda)| = 1 - \frac{|\lambda-k|}{\sqrt{(\lambda-k)^2+\delta^2}} \le 1 - \frac{\varepsilon}{\sqrt{\varepsilon^2+\delta^2}}\,. \tag{A.329}$$

Since $\operatorname{sgn}(\lambda-k)$ is constant and equal to $\operatorname{sgn}(w-k)$ for λ strictly between k and w, $\int_k^w \operatorname{sgn}(\lambda-k)F'(\lambda)\,\mathrm{d}\lambda = \operatorname{sgn}(w-k)\big(F(w)-F(k)\big) = q_k(w)$, whence

$$|q_k(w) - q_\delta(w)| \le \left| \int_k^w |r_\delta(\lambda)|\;|F'(\lambda)|\;\mathrm{d}\lambda \right| \le M\left[2\varepsilon + 2M_0\left(1 - \frac{\varepsilon}{\sqrt{\varepsilon^2+\delta^2}}\right)\right], \tag{A.330}$$

the interval of integration having been split at $|\lambda-k| = \varepsilon$ and its length bounded by $2M_0$ — here $|k| \le M_0$ may be assumed, since for $|k| > M_0$ the sign of $u-k$ is constant and Equation (A.325) reduces to Equation (10.85) with a sign. Choosing first ε and then δ, the right-hand side of Equation (A.330) is made arbitrarily small uniformly in w: $q_\delta \longrightarrow q_k$ uniformly on $[-M_0, M_0]$.

Passage to the limit. Apply Equation (10.92) to (η_δ, q_δ) and subtract the corresponding expression built from (η_k, q_k). Each of the three differences is bounded in modulus by the uniform bounds Equation (A.328) and Equation (A.330) times the integral of $|\partial_t\varphi|$, $|\partial_x\varphi|$ or $|\varphi(\cdot,0)|$ over the compact support of φ, a fixed finite number. Letting $\delta \to 0$ therefore gives Equation (A.325). No convergence theorem is needed at this step: the convergence is uniform on a set of finite measure. □

Lemma A.157 (Two elementary identities)**.** *For all $a, b \in [-M_0, M_0]$, with $\Phi(a,b) = \operatorname{sgn}(a-b)\big(F(a)-F(b)\big)$,*

$$\Phi(a,b) = F\big(\max(a,b)\big) - F\big(\min(a,b)\big)\,, \tag{A.331}$$

so that Φ is symmetric, continuous, and Lipschitz:

$$|\Phi(a,b)| \le M\;|a-b|\,, \qquad |\Phi(a,b) - \Phi(a',b')| \le M\big(|a-a'| + |b-b'|\big)\,. \tag{A.332}$$

Rests on Definition A.150 and Equation (A.320).

Proof. *Derives Lemma A.157.*

If $a > b$ then $\operatorname{sgn}(a-b) = 1$, $\max = a$, $\min = b$ and both sides of Equation (A.331) equal $F(a) - F(b)$; if $a < b$ then $\operatorname{sgn}(a-b) = -1$, $\max = b$, $\min = a$ and both sides equal $F(b) - F(a)$; if $a = b$ both vanish. Symmetry and continuity are then manifest, max and min being continuous. For the bounds, the mean value theorem gives $|F(\max) - F(\min)| \le M\,|\max - \min| = M\,|a-b|$, which is the first; and since $a \longmapsto \max(a,b)$ and $a \longmapsto \min(a,b)$ are 1-Lipschitz in each argument, so is $F \circ \max$ and $F \circ \min$ up to the factor M, giving the second. □

Remark A.158 (Why $|u-k|$ and not a general convex entropy)*.* Equation (A.331) is the reason Kruzhkov's family is the right one. What the doubling will produce is the pair $\big(|u-v|\,,\,\Phi(u,v)\big)$ built from two *solutions*, and for this to be usable it must be a genuine entropy pair in each variable separately with the other frozen — which is exactly what Definition A.150 provides, the frozen solution playing the role of the constant k. No single convex η of Definition 10.101 has that property, because a general η knows nothing about the second solution. The price is that $|u-k|$ is not C^2, and Proposition A.156 is what pays it.

A.17.4 Doubling the variables

Proposition A.159 (The doubled inequality)**.** *Let $\psi = \psi(t,x,s,y) \geq 0$ be smooth with compact support in $\{t>0\}\times\mathbb{R}\times\{s>0\}\times\mathbb{R}$. Then*

$$\iiiint \Big[|u(t,x)-v(s,y)|\,(\partial_t+\partial_s)\psi + \Phi\big(u(t,x),v(s,y)\big)\,(\partial_x+\partial_y)\psi \Big] \geq 0\,, \tag{A.333}$$

the integral being over (t,x,s,y) with $t,s>0$.

Rests on Proposition A.156 and Lemma A.157.

Proof. *Derives Proposition A.159.*

Fix (s,y) with $s>0$ and apply Equation (A.325) to the solution u, with the constant $k=v(s,y)$ — legitimate for almost every (s,y), since $|v(s,y)|\leq M_0$ is then a genuine real number — and with the non-negative test function $(t,x)\longmapsto\psi(t,x,s,y)$, which is smooth with compact support in $\{t>0\}\times\mathbb{R}$. The initial term is absent because ψ vanishes for t near 0. Recalling $q_{v(s,y)}\big(u(t,x)\big)=\Phi\big(u(t,x),v(s,y)\big)$ from Equation (A.321),

$$\int_0^\infty\!\!\int_{\mathbb{R}} \Big[|u-v(s,y)|\,\partial_t\psi + \Phi\big(u,v(s,y)\big)\,\partial_x\psi \Big]\mathrm{d}x\,\mathrm{d}t \geq 0\,. \tag{A.334}$$

The left-hand side of Equation (A.334) is a bounded measurable function of (s,y) — bounded because the integrand is bounded by $2M_0\,|\partial_t\psi| + 2MM_0\,|\partial_x\psi|$ using Equation (A.332) — so it may be integrated over $\{s>0\}\times\mathbb{R}$, and Fubini's theorem rearranges the result into a single integral over the four variables. This gives Equation (A.333) with only the ∂_t and ∂_x terms.

Now exchange the roles. Fix (t,x) with $t>0$ and apply Equation (A.325) to v, with $k=u(t,x)$ and the test function $(s,y)\longmapsto\psi(t,x,s,y)$; by the symmetry of Φ recorded in Lemma A.157, $q_{u(t,x)}\big(v(s,y)\big)=\Phi\big(v(s,y),u(t,x)\big)=\Phi\big(u(t,x),v(s,y)\big)$, and $|v-u|=|u-v|$. Integrating over (t,x) gives Equation (A.333) with the ∂_s and ∂_y terms. Adding the two inequalities gives Equation (A.333). □

Proposition A.160 (The cancellation)**.** *Let $\chi\geq 0$ be smooth with compact support in $(0,\infty)\times\mathbb{R}$, let $\theta\in C^\infty(\mathbb{R})$ be even and non-negative with support in $(-1,1)$ and $\int\theta=1$, and put $\theta_\delta(\sigma)=\delta^{-1}\theta(\sigma/\delta)$ and*

$$\psi_\delta(t,x,s,y)=\chi\Big(\frac{t+s}{2},\frac{x+y}{2}\Big)\,\theta_\delta\Big(\frac{t-s}{2}\Big)\,\theta_\delta\Big(\frac{x-y}{2}\Big)\,. \tag{A.335}$$

For δ small enough ψ_δ is admissible in Proposition A.159, and in the sum-and-difference coordinates

$$\tau=\frac{t+s}{2}\,,\quad \zeta=\frac{x+y}{2}\,,\quad \sigma=\frac{t-s}{2}\,,\quad \omega=\frac{x-y}{2}\,, \tag{A.336}$$

one has

$$(\partial_t+\partial_s)\psi_\delta=(\partial_\tau\chi)\,\theta_\delta(\sigma)\theta_\delta(\omega)\,,\qquad (\partial_x+\partial_y)\psi_\delta=(\partial_\zeta\chi)\,\theta_\delta(\sigma)\theta_\delta(\omega)\,. \tag{A.337}$$

Rests on Proposition A.159 and Equation (A.335).

Proof. *Derives Proposition A.160.*

By the chain rule applied to Equation (A.336), $\partial_t=\frac12\partial_\tau+\frac12\partial_\sigma$ and $\partial_s=\frac12\partial_\tau-\frac12\partial_\sigma$, so $\partial_t+\partial_s=\partial_\tau$; likewise $\partial_x+\partial_y=\partial_\zeta$. In Equation (A.335) the factors θ_δ depend only on σ and ω and are therefore annihilated by ∂_τ and ∂_ζ, while χ depends only on (τ,ζ); that is Equation (A.337). Admissibility: the support of ψ_δ lies within distance 2δ of the diagonal $t=s$, $x=y$ over the support of χ, so if 2δ is smaller than the distance from the support of χ to $\{\tau=0\}$ then $t,s>0$ throughout. □

Remark A.161. This is the step the method is named for and the only one that could fail. Had the mollifier been differentiated, each derivative would have produced a factor δ^{-1}, and the limit $\delta \to 0$ would have been meaningless. It survives because the two entropy inequalities are added, not subtracted, and because the sum of the two time derivatives — one for each copy of the equation — is a derivative along the diagonal, in whose direction the mollifier is constant. In effect the argument tests the difference of two solutions against a function that is smooth along the diagonal and singular across it, and only the diagonal direction is ever differentiated.

A.17.5 The limit and the cone estimate

Proposition A.162 (The limit $\delta \to 0$)**.** *For every non-negative χ smooth with compact support in $(0,\infty)\times\mathbb{R}$,*

$$\int_0^\infty\!\!\int_{\mathbb{R}} \Big[|u-v|\;\partial_t\chi + \Phi(u,v)\,\partial_x\chi \Big]\,\mathrm{d}x\,\mathrm{d}t \geq 0\,, \tag{A.338}$$

that is, $\partial_t|u-v| + \partial_x\Phi(u,v) \leq 0$ in the sense of distributions on $(0,\infty)\times\mathbb{R}$.
Rests on Proposition A.160 and Lemma A.157.

Proof. *Derives Proposition A.162.*

Insert Equation (A.337) into Equation (A.333) and change variables to Equation (A.336). The map $(t,x,s,y) \longmapsto (\tau,\zeta,\sigma,\omega)$ is linear with $\mathrm{d}t\,\mathrm{d}s = 2\,\mathrm{d}\tau\,\mathrm{d}\sigma$ and $\mathrm{d}x\,\mathrm{d}y = 2\,\mathrm{d}\zeta\,\mathrm{d}\omega$, so

$$\iint \left[\iint \Big(G_{\sigma\omega}\,\partial_\tau\chi + H_{\sigma\omega}\,\partial_\zeta\chi \Big)\theta_\delta(\sigma)\theta_\delta(\omega)\,\mathrm{d}\sigma\,\mathrm{d}\omega \right] \mathrm{d}\zeta\,\mathrm{d}\tau \geq 0\,, \tag{A.339}$$

after dividing by the positive constant 4, where

$$G_{\sigma\omega}(\tau,\zeta) = |u(\tau+\sigma,\zeta+\omega) - v(\tau-\sigma,\zeta-\omega)|\,, \quad H_{\sigma\omega}(\tau,\zeta) = \Phi\big(u(\tau{+}\sigma,\zeta{+}\omega), v(\tau{-}\sigma,\zeta{-}\omega)\big)\,. \tag{A.340}$$

Let K be a compact neighbourhood of the support of χ in $(0,\infty)\times\mathbb{R}$. By continuity of translation in L^1 (item 4 of Remark A.153) both $(\tau,\zeta) \longmapsto u(\tau{+}\sigma,\zeta{+}\omega)$ and $(\tau,\zeta) \longmapsto v(\tau{-}\sigma,\zeta{-}\omega)$ converge to u and v in $L^1(K)$ as $(\sigma,\omega)\to(0,0)$; hence, the maps $(a,b)\longmapsto|a-b|$ and $(a,b)\longmapsto\Phi(a,b)$ being Lipschitz by Equation (A.332),

$$\int_K \big| G_{\sigma\omega} - |u-v| \big| \longrightarrow 0\,, \qquad \int_K |H_{\sigma\omega} - \Phi(u,v)| \longrightarrow 0\,. \tag{A.341}$$

Denote by $\Xi(\sigma,\omega)$ the inner bracket of Equation (A.339) integrated against χ's derivatives, i.e. the quantity $\iint (G_{\sigma\omega}\partial_\tau\chi + H_{\sigma\omega}\partial_\zeta\chi)\mathrm{d}\zeta\,\mathrm{d}\tau$, and by Ξ_0 the left-hand side of Equation (A.338). Then $|\Xi(\sigma,\omega) - \Xi_0| \leq \|\nabla\chi\|_\infty \int_K (|G_{\sigma\omega} - |u-v|| + |H_{\sigma\omega} - \Phi(u,v)|)$, which by Equation (A.341) tends to 0 as $(\sigma,\omega)\to 0$. But Equation (A.339) says exactly $\iint \Xi(\sigma,\omega)\theta_\delta(\sigma)\theta_\delta(\omega) \geq 0$, an average of Ξ over the square $|\sigma|,|\omega| < \delta$ with a non-negative weight of total mass 1; as $\delta\to 0$ that average converges to Ξ_0. Hence $\Xi_0 \geq 0$, which is Equation (A.338). □

Proof of Theorem A.151. *Derives Theorem A.151.*

Write $w = |u-v|$ and $\Psi = \Phi(u,v)$, so that Equation (A.338) holds and $|\Psi| \leq M\,w$ by Equation (A.332).

The cone cutoff. Fix $R > 0$ and $T > 0$. For $\epsilon > 0$ let $\Lambda_\epsilon \in C^\infty(\mathbb{R})$ be non-decreasing with $\Lambda_\epsilon = 0$ on $(-\infty, 0]$, $\Lambda_\epsilon = 1$ on $[\epsilon,\infty)$ and $0 \leq \Lambda_\epsilon' \leq 2/\epsilon$; and let $n_\epsilon \in C^\infty(\mathbb{R})$ be even with $|n_\epsilon'| \leq 1$ and $|n_\epsilon(x) - |x|| \leq \epsilon$ — for instance $n_\epsilon = \sqrt{x^2+\epsilon^2}$. Put

$$\beta(t,x) = \Lambda_\epsilon\big(R + M(T-t) - n_\epsilon(x)\big)\,, \tag{A.342}$$

a smooth function with values in $[0,1]$, equal to 1 well inside the backward cone $\{n_\epsilon(x) \le R + M(T-t)\}$ and to 0 outside it. Let $\alpha \ge 0$ be smooth with compact support in $(0,T]$ and take $\chi = \alpha\beta$, which is admissible in Equation (A.338) after multiplication by a fixed spatial cutoff equal to 1 on the (bounded) support of β — the cone is bounded for $t \ge 0$ — so that χ has compact support.

The lateral term has a sign. Since $\partial_t\chi = \alpha'\beta + \alpha\,\partial_t\beta$ and $\partial_x\chi = \alpha\,\partial_x\beta$, with

$$\partial_t\beta = -M\,\Lambda_\epsilon'\,, \qquad \partial_x\beta = -\Lambda_\epsilon'\,n_\epsilon'(x)\,, \tag{A.343}$$

the contribution of the second group to Equation (A.338) is

$$\iint \alpha\,[w\,\partial_t\beta + \Psi\,\partial_x\beta] = -\iint \alpha\,\Lambda_\epsilon'\,[M\,w + \Psi\,n_\epsilon'] \le 0\,, \tag{A.344}$$

because $\alpha \ge 0$, $\Lambda_\epsilon' \ge 0$, and $Mw + \Psi n_\epsilon' \ge Mw - |\Psi| \ge 0$ by $|n_\epsilon'| \le 1$ and $|\Psi| \le Mw$. This is the point of choosing the slope of the cone to be M: the flux can never carry mass out through the lateral boundary faster than the boundary retreats. Subtracting Equation (A.344) from Equation (A.338) leaves

$$\int_0^\infty \alpha'(t)\,\left[\int_{\mathbb{R}} \beta(t,x)\,w(t,x)\,\mathrm{d}x\right]\mathrm{d}t \ge 0\,. \tag{A.345}$$

From an average to a value. Let $0 < \tau_1 < \tau_2 \le T$ and let $\alpha = \alpha_h$ rise smoothly from 0 to 1 on $[\tau_1, \tau_1 + h]$ and fall back to 0 on $[\tau_2, \tau_2 + h]$, with α_h' bounded by $2/h$ and of one sign on each interval. Then Equation (A.345) reads, in the limit of the standard smooth ramps,

$$\frac{1}{h}\int_{\tau_1}^{\tau_1+h} \Theta(t)\,\mathrm{d}t \;\ge\; \frac{1}{h}\int_{\tau_2}^{\tau_2+h} \Theta(t)\,\mathrm{d}t\,, \qquad \Theta(t) = \int_{\mathbb{R}} \beta(t,x)\,w(t,x)\,\mathrm{d}x\,, \tag{A.346}$$

up to an error tending to 0 with the width of the ramps. Θ is bounded and measurable, so by Lebesgue's differentiation theorem (item 5 of Remark A.153) the two averages converge, as $h \to 0$, to $\Theta(\tau_1)$ and $\Theta(\tau_2)$ for almost every τ_1, τ_2. Hence $\Theta(\tau_2) \le \Theta(\tau_1)$ for almost all $\tau_1 < \tau_2$.

Removing the smoothing. As $\epsilon \to 0$, $\beta(t,x)$ tends to 1 at every x with $|x| < R + M(T-t)$ and to 0 at every x with $|x| > R + M(T-t)$, and $0 \le \beta \le 1$ throughout; by dominated convergence, applied with the integrable majorant $2M_0\,\mathbf{1}_{\{|x| \le R+MT\}}$,

$$\int_{|x| \le R+M(T-\tau_2)} w(\tau_2,x)\,\mathrm{d}x \;\le\; \int_{|x| \le R+M(T-\tau_1)} w(\tau_1,x)\,\mathrm{d}x\,. \tag{A.347}$$

Choosing $T = \tau_2$ turns Equation (A.347) into Equation (A.322).

Down to $t = 0$. By Lemma A.154 applied to u and to v on the compact set $|x| \le R + M\tau_2$, $w(\tau_1,\cdot) \longrightarrow |u_0 - v_0|$ in the mean over that set as $\tau_1 \to 0^+$ — the triangle inequality gives $\big|\,|u(\tau_1) - v(\tau_1)| - |u_0 - v_0|\,\big| \le |u(\tau_1) - u_0| + |v(\tau_1) - v_0|$ — so the right-hand side of Equation (A.322) converges to $\int_{|x| \le R+M\tau_2} |u_0 - v_0|$ along a sequence of admissible τ_1, which proves the version with $\tau_1 = 0$.

The global statement. If $u_0 - v_0$ is integrable, let $R \to \infty$ in that version: the left-hand side increases to $\int_{\mathbb{R}} w(t,x)\,\mathrm{d}x$ by monotone passage on an increasing family of sets, and the right-hand side increases to $\int_{\mathbb{R}} |u_0 - v_0|$, giving Equation (A.323). □

A.17.6 Existence by vanishing viscosity

The uniqueness half is now complete. Existence is the limit $\varepsilon \to 0^+$ of the viscous problem, and the one step that is genuinely about entropy — and the only one that would be invisible in a purely functional-analytic account — is carried out here in full.

Proposition A.163 (The viscous problem produces the entropy inequality)**.** *Let u^ε be a bounded classical solution of*

$$\partial_t u^\varepsilon + \partial_x F(u^\varepsilon) = \varepsilon\, \partial_x^2 u^\varepsilon\,, \qquad u^\varepsilon(0,\cdot) = u_0^\varepsilon\,, \tag{A.348}$$

with $|u^\varepsilon| \le M_0$, and let (η, q) be an entropy pair in the sense of Definition 10.101. Then for every non-negative test function φ

$$\int_0^\infty \int_{\mathbb{R}} [\eta(u^\varepsilon)\partial_t\varphi + q(u^\varepsilon)\partial_x\varphi]\,\mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}} \eta(u_0^\varepsilon)\varphi(x,0)\,\mathrm{d}x \;\ge\; -\varepsilon\, C_\varphi\,, \tag{A.349}$$

with $C_\varphi = \max_{|w|\le M_0} |\eta(w)| \cdot \iint |\partial_x^2\varphi|$, a constant independent of ε.
Rests on Definition 10.101, Equation (10.81), and Theorem 10.80.

Proof. *Derives Proposition A.163.*

Multiply Equation (A.348) by $\eta'(u^\varepsilon)$. On the left, the chain rule and $q' = \eta' F'$ give $\eta'(u^\varepsilon)\partial_t u^\varepsilon = \partial_t \eta(u^\varepsilon)$ and $\eta'(u^\varepsilon)\partial_x F(u^\varepsilon) = \eta'(u^\varepsilon)F'(u^\varepsilon)\partial_x u^\varepsilon = \partial_x q(u^\varepsilon)$. On the right,

$$\varepsilon\,\eta'(u^\varepsilon)\,\partial_x^2 u^\varepsilon = \varepsilon\,\partial_x^2\eta(u^\varepsilon) - \varepsilon\,\eta''(u^\varepsilon)\big(\partial_x u^\varepsilon\big)^2 \;\le\; \varepsilon\,\partial_x^2\eta(u^\varepsilon)\,, \tag{A.350}$$

the identity being the chain rule applied twice and the inequality being $\eta'' \ge 0$, which is the convexity demanded by Definition 10.101 and the *only* place it is used. Hence

$$\partial_t \eta(u^\varepsilon) + \partial_x q(u^\varepsilon) \le \varepsilon\,\partial_x^2\eta(u^\varepsilon) \tag{A.351}$$

pointwise. Multiply Equation (A.351) by $\varphi \ge 0$, integrate over $\{t>0\}\times\mathbb{R}$ and integrate by parts, all boundary terms at spatial infinity vanishing with φ and the one at $t=0$ producing $-\int \eta(u_0^\varepsilon)\varphi(x,0)\,\mathrm{d}x$:

$$-\iint [\eta\,\partial_t\varphi + q\,\partial_x\varphi] - \int \eta(u_0^\varepsilon)\varphi(x,0)\,\mathrm{d}x \;\le\; \varepsilon \iint \eta(u^\varepsilon)\,\partial_x^2\varphi\,,$$

and the right-hand side is bounded in modulus by εC_φ, which is Equation (A.349). The two integrations by parts on the viscous term are the reason the estimate is $O(\varepsilon)$ and not $O(1)$: all the differentiation has been moved onto the test function, and no derivative of u^ε survives. □

Remark A.164 (The remaining inputs for existence)*.* Proposition A.163 is the whole of the entropy argument; what it does not supply is a limit to take it to. Three further steps are needed, of which the first is elementary in this treatise and the other two are the quoted items (7) of Remark A.153.

1. *The supremum bound.* $|u^\varepsilon| \le \|u_0^\varepsilon\|_\infty \le \|u_0\|_\infty$, by the parabolic maximum principle Theorem 10.80 applied to Equation (A.348) on large space-time rectangles: the equation is of the form treated there with the extra first-order term $F'(u^\varepsilon)\partial_x u^\varepsilon$, which vanishes at an interior extremum and so does not disturb the argument. This is why M_0 and hence M in Equation (A.320) can be fixed independently of ε.

2. *Classical solvability of Equation* (A.348) for smooth bounded data $u_0^\varepsilon = u_0$ mollified, on $\{t>0\}\times\mathbb{R}$. This is parabolic theory and is quoted.

3. *Compactness.* The spatial derivative $p = \partial_x u^\varepsilon$ satisfies the linear parabolic equation $\partial_t p + \partial_x\big(F'(u^\varepsilon)p\big) = \varepsilon\partial_x^2 p$, from which the total variation $\int |p(t,x)|\,\mathrm{d}x$ is non-increasing in t; together with (i) this gives a family uniformly bounded and of uniformly bounded variation, and a compactness theorem for such families produces a sequence $\varepsilon_j \to 0$ with $u^{\varepsilon_j} \to u$ in L^1 on compact sets and almost everywhere. The compactness theorem is quoted.

Granting these, pass to the limit in Equation (A.349): η and q are continuous and $u^{\varepsilon_j} \to u$ almost everywhere with values in a fixed bounded interval, so dominated convergence carries both integrals, the data converge because $u_0^{\varepsilon} \to u_0$ in L^1 on compact sets, and the right-hand side tends to 0. The limit is exactly Equation (10.92). Taking $\eta(w) = w$ and $q = F$ — a legitimate, if degenerate, member of Definition 10.101, for which Equation (A.350) is an equality — gives the same statement with equality, which is Equation (10.85): the limit is a weak solution.

Proof of Theorem 10.103. *Derives Theorem 10.103.*

Existence is Remark A.164: the vanishing-viscosity limit is a weak solution in the sense of Definition 10.95 satisfying Equation (10.92) for every entropy pair, hence an entropy solution in the sense of Definition 10.101. *Uniqueness* and *continuous dependence in the mean* are Corollary A.152, which rests on Theorem A.151. The last clause of the chapter's statement — that where the solution is piecewise C^1 across a C^1 curve its jumps obey Equation (10.91) — is Proposition 10.102, proved in the chapter, and is not re-derived here. □

Remark A.165 (Convexity of the flux is not needed). Theorem 10.103 assumes F convex, and the chapter has good reason to: convexity is what makes Proposition 10.102 equate the entropy inequality with the geometric condition Equation (10.91), and it is what makes the shock picture of Section 10.6.3 the whole story. But the proof given above never used it. Only $F \in C^1$ entered, through the bound M of Equation (A.320) and the Lipschitz estimate Equation (A.332). What is lost without convexity is not uniqueness but the *description* of the admissible discontinuities: for a non-convex flux the entropy inequality still selects one solution, but the admissible jumps are described by a chord condition — the chord joining the two states must lie on one side of the graph of F between them — rather than by the simple inequality $u_- \geq u_+$. The function G of Equation (10.94) is exactly that chord difference, and the chapter's proof of Proposition 10.102 is where convexity is spent: it is used to pass from $G(u_-) = G(u_+) = 0$ to $G \leq 0$ between the two states, and that implication is what fails when F is not convex.

Remark A.166 (Why the entropy inequality and not the jump condition). The chapter already observes, in the paragraph following Definition 10.101, that Equation (10.91) presupposes one-sided limits that a bounded measurable weak solution need not have. The proof above shows what is gained by not presupposing them: at no point did the argument speak of a discontinuity, a curve, or a one-sided limit. It compared two solutions through their entropy inequalities alone, and the only regularity used — continuity of translation in the mean — is a property every bounded measurable function has. Example 10.98 is settled by the same stroke without any examination of the two candidate solutions: whichever of them satisfies Equation (10.92) is the only one that does.

Remark A.167. Appendix A.17 discharges the proof obligation of Theorem 10.103 in Chapter 10, and with it closes the pattern that Remark 10.104 describes: a nonlinear equation destroys its own classical solution in finite time (Proposition 10.94), weak solutions restore existence at the cost of uniqueness (Example 10.98), and the entropy inequality Equation (10.92) restores uniqueness, giving back all three of Hadamard's conditions (Definition 10.24) with the mean as the norm. The physical reading is Remark 10.100, and the gas-dynamical form of the admissible jump is Chapter 31. Six inputs are assumed and are listed in Remark A.153; five of them are the Lebesgue theory that this treatise deliberately does not build, and the sixth is Lemma A.154.

A.18 Lévy's Continuity Theorem

This appendix proves Theorem 11.45 of Chapter 11: if the characteristic functions of a sequence of random variables converge pointwise to a limit that is continuous at the origin, then that limit is itself a characteristic function and the variables converge in distribution. It is the bridge the central limit theorem crosses — Theorem 11.46 produces $\varphi_{Z_n}(t) \to \mathrm{e}^{-t^2/2}$ and needs to conclude $Z_n \to \mathcal{N}(0,1)$ — and it is used again in Appendix A.19 and Appendix A.20.

The route has four stages, each of which is a theorem in its own right: Helly's selection theorem, which extracts a limiting monotone function from any sequence of distribution functions (Appendix A.18.4); the Fourier inversion formula, which recovers the distribution from the transform and thereby shows the transform determines it (Appendix A.18.5); the tightness estimate, which is the single place continuity of the limit at the origin is used (Appendix A.18.6); and the assembly (Appendix A.18.7). Everything rests on one classical integral, Dirichlet's, which is evaluated first because the inversion formula is built out of it.

Throughout, $F_X(x) = \Pr(X \le x)$ is the distribution function of Definition 11.21 — non-decreasing, right-continuous, with limits 0 and 1 at $\mp\infty$, and *not* assumed continuous — and $\varphi_X(t) = \left\langle \mathrm{e}^{\mathrm{i}tX} \right\rangle$ the characteristic function of Definition 11.42. Convergence in distribution means $F_{X_n}(x) \to F_X(x)$ at every continuity point x of F_X. All the quantities here are pure numbers: t carries the reciprocal of the SI unit of X, so that tX, and with it every exponent below, is dimensionless.

A.18.1 Statement

Theorem A.168 (Lévy's continuity theorem)**.** *Let $X_1, X_2, \dots$ be real random variables with characteristic functions φ_n. Suppose $\varphi_n(t) \to \varphi(t)$ for every $t \in \mathbb{R}$, and suppose the limit function φ is continuous at $t = 0$. Then φ is the characteristic function of a random variable X, and $X_n \to X$ in distribution.*

Rests on Definitions 11.21 and 11.42.

The converse is elementary and is recorded at the end (Corollary A.181); it is the direction the theorem is *not* usually needed in.

A.18.2 The two facts imported from Lebesgue theory

Chapter 7 builds the Riemann–Darboux integral (Definition 7.39) and no more. Two theorems of the Lebesgue theory are used below and are not proved anywhere in this treatise; they are stated here in exactly the form in which they are applied, so that the reader can see precisely how much is being assumed.

Theorem A.169 (Dominated convergence)**.** *Let $Z, Z_1, Z_2, \dots$ be random variables with $Z_m \to Z$ pointwise (or in probability), and suppose there is a random variable D with $\langle D \rangle < \infty$ and $|Z_m| \le D$ for every m. Then $\langle Z_m \rangle \to \langle Z \rangle$. The same statement holds for a family indexed by a continuous parameter. In particular, if $|Z_m| \le c$ for a constant c, then $\langle Z_m \rangle \to \langle Z \rangle$ [1894].*

Rests on Definition 11.4.

Theorem A.170 (Fubini's theorem, bounded case)**.** *Let $h(t, \omega)$ be jointly measurable, let $|h| \le c$ for a constant c, and let $T < \infty$. Then*

$$\int_{-T}^{T} \langle h(t, \cdot) \rangle \, \mathrm{d}t = \left\langle \int_{-T}^{T} h(t, \cdot) \, \mathrm{d}t \right\rangle , \tag{A.352}$$

both sides being finite [1894].

Rests on Definition 11.4.

Remark A.171 (What is quoted here). Theorems A.169 and A.170 are the *only* statements in this appendix that are assumed rather than derived. Both belong to measure theory: dominated convergence is the theorem that makes the Lebesgue integral complete under limits — the property the Riemann integral lacks, which is why it cannot be proved from Chapter 7 — and Fubini's theorem is the corresponding statement for a product measure. This treatise develops neither, by the editorial choice recorded at Lemma 11.44, where the same gap was closed by an explicit truncation instead. No such elementary substitute is available here: the interchange Equation (A.352) is applied to the expectation of an oscillatory integral whose integrand depends on the random variable through $\mathrm{e}^{\mathrm{i}tX}$, and there is nothing to truncate. Each use below is flagged by name at the point where it occurs; there are four in all (two of each). Everything else — Helly's theorem, the Helly–Bray lemma, the Dirichlet integral, the inversion formula, the tightness estimate — is derived here from the real analysis of Chapter 7.

A.18.3 The Dirichlet integral

Lemma A.172 (Dirichlet)**.** *Write $S(Y)=\int_0^Y (\sin s)/s\,\mathrm{d}s$ for $Y>0$, the integrand being extended by the value 1 at $s=0$, where it is continuous. Then*

$$\lim_{Y\to\infty} S(Y)=\frac{\pi}{2}\,,\qquad \text{and}\qquad C_0\equiv\sup_{Y>0}|S(Y)|\;\le\;4\,. \tag{A.353}$$

Consequently, for every $u\in\mathbb{R}$ and $T>0$, $\int_0^T(\sin(tu))/t\,\mathrm{d}t=S(Tu)$ for $u>0$, $=-S(T|u|)$ for $u<0$ and $=0$ for $u=0$; so the integral is bounded by C_0 uniformly in T and u and tends to $\frac{\pi}{2}\operatorname{sgn}(u)$ as $T\to\infty$.

Rests on Theorem 7.43 and Equation (7.28).

Proof. *Derives Lemma A.172.*

The uniform bound. On $[0,1]$ the integrand is bounded by 1, so $\left|\int_0^1\right|\le 1$. For $Y\ge 1$ integrate by parts (Equation (7.28)) with $\sin s\,\mathrm{d}s=-\mathrm{d}\cos s$:

$$\int_1^Y\frac{\sin s}{s}\,\mathrm{d}s=\left[-\frac{\cos s}{s}\right]_1^Y-\int_1^Y\frac{\cos s}{s^2}\,\mathrm{d}s\,, \tag{A.354}$$

whose first term is at most $1/Y+\cos 1\le 2$ in modulus and whose second is at most $\int_1^\infty s^{-2}\mathrm{d}s=1$. Hence $|S(Y)|\le 1+2+1=4$ for $Y\ge 1$, and $|S(Y)|\le 1$ for $Y\le 1$. The same estimate applied on $[Y,Y']$ with $1\le Y<Y'$ gives

$$\left|\int_Y^{Y'}\frac{\sin s}{s}\,\mathrm{d}s\right|\;\le\;\frac{2}{Y}\,, \tag{A.355}$$

so $S(Y)$ satisfies the Cauchy criterion (Theorem 7.8) and the limit $D=\lim_{Y\to\infty}S(Y)$ exists.

The value. For $\lambda>0$ put $I(\lambda)=\int_0^\infty \mathrm{e}^{-\lambda s}(\sin s)/s\,\mathrm{d}s$, which converges absolutely since the integrand is bounded by $\mathrm{e}^{-\lambda s}$. Its λ-derivative is $-\mathrm{e}^{-\lambda s}\sin s$ in modulus at most $\mathrm{e}^{-\lambda_0 s}$ for $\lambda\ge\lambda_0>0$, an integrable bound independent of λ, so the difference quotients converge uniformly on $[\lambda_0,\infty)$ and differentiation under the integral sign is legitimate there:

$$I'(\lambda)=-\int_0^\infty \mathrm{e}^{-\lambda s}\sin s\,\mathrm{d}s=-\Im\int_0^\infty \mathrm{e}^{-(\lambda-\mathrm{i})s}\,\mathrm{d}s=-\Im\frac{1}{\lambda-\mathrm{i}}=-\frac{1}{1+\lambda^2}\,. \tag{A.356}$$

Also $|I(\lambda)|\le\int_0^\infty \mathrm{e}^{-\lambda s}\mathrm{d}s=1/\lambda\to 0$. Integrating Equation (A.356) from λ to ∞ therefore gives $I(\lambda)=\frac{\pi}{2}-\arctan\lambda$.

The two agree in the limit. Fix $Y\ge 1$ and estimate $I(\lambda)-D$ in three pieces. On $[0,Y]$, using $\left|\mathrm{e}^{-\lambda s}-1\right|\le\lambda s$ and $|\sin s|\le 1$,

$$\left|\int_0^Y\left(\mathrm{e}^{-\lambda s}-1\right)\frac{\sin s}{s}\,\mathrm{d}s\right|\;\le\;\int_0^Y\lambda\,\mathrm{d}s=\lambda Y\,.$$

The tail of D is at most $2/Y$ by Equation (A.355). For the tail of I, integrate by parts with $u = \mathrm{e}^{-\lambda s}/s$:

$$\left| \int_Y^\infty \mathrm{e}^{-\lambda s} \frac{\sin s}{s} \, \mathrm{d}s \right| \;\le\; \frac{\mathrm{e}^{-\lambda Y}}{Y} + \int_Y^\infty \mathrm{e}^{-\lambda s} \left(\frac{\lambda}{s} + \frac{1}{s^2} \right) \mathrm{d}s \;\le\; \frac{1}{Y} + \frac{1}{Y} + \frac{1}{Y} = \frac{3}{Y} \,,$$

the middle step bounding $1/s \le 1/Y$ in the λ-term and $\mathrm{e}^{-\lambda s} \le 1$ in the other. Hence $|I(\lambda) - D| \le \lambda Y + 5/Y$ for every $Y \ge 1$ and $\lambda > 0$. Given $\varepsilon > 0$ choose $Y = 10/\varepsilon$ and then $\lambda < \varepsilon/(2Y)$: the right side is below ε. So $D = \lim_{\lambda \downarrow 0} I(\lambda) = \frac{\pi}{2} - \arctan 0 = \frac{\pi}{2}$.

The final statement is the substitution $s = tu$ (Equation (7.27)), under which $(\sin(tu))/t \, \mathrm{d}t = (\sin s)/s \, \mathrm{d}s$ for $u > 0$; oddness of sin in u supplies the case $u < 0$. □

A.18.4 Helly's selection theorem

Theorem A.173 (Helly's selection theorem)**.** *Let $F_1, F_2, \dots$ be distribution functions. Then there are a subsequence $F_{n_1}, F_{n_2}, \dots$ and a non-decreasing, right-continuous $F : \mathbb{R} \to [0,1]$ such that $F_{n_k}(x) \to F(x)$ at every continuity point x of F.*

Rests on Definition 11.21 and Theorem 7.7.

The limit F need not be a distribution function: mass can escape to infinity, as it does for $F_n = F_{X+n}$, where $F \equiv 0$. Excluding that is precisely what tightness will do.

Proof. *Derives Theorem A.173.*

The diagonal argument. Enumerate the rationals, $\mathbb{Q} = \{q_1, q_2, \dots\}$. The numbers $F_n(q_1)$ lie in the bounded set $[0,1]$, so by Bolzano–Weierstrass (Theorem 7.7) some subsequence, indexed by an infinite set $N_1 \subseteq \mathbb{N}$, has $F_n(q_1)$ convergent along N_1; call the limit $G(q_1)$. Having chosen $N_1 \supseteq N_2 \supseteq \cdots \supseteq N_m$ with $F_n(q_i) \to G(q_i)$ along N_m for every $i \le m$, apply Theorem 7.7 again to the bounded sequence $(F_n(q_{m+1}))_{n \in N_m}$ to get $N_{m+1} \subseteq N_m$ and a limit $G(q_{m+1})$. Let n_k be the kth element of N_k, in increasing order and chosen larger than n_{k-1}, which is possible because each N_k is infinite. For each fixed i, all but the first $i-1$ terms of (n_k) lie in N_i, so

$$F_{n_k}(q) \;\longrightarrow\; G(q) \qquad \text{for every } q \in \mathbb{Q} \,. \tag{A.357}$$

Each F_n is non-decreasing with values in $[0,1]$, so G inherits both properties on $\mathbb{Q}$.

The limit function. Define $F(x) = \inf \{G(q) \mid q \in \mathbb{Q},\ q > x\}$. It is non-decreasing and $[0,1]$-valued because G is. It is right-continuous: given x and $\varepsilon > 0$, pick a rational $q > x$ with $G(q) < F(x) + \varepsilon$; every $y \in (x, q)$ then has $F(y) \le G(q) < F(x) + \varepsilon$, and monotonicity gives $F(y) \downarrow F(x)$ as $y \downarrow x$.

Two inequalities are used repeatedly and follow at once from the definition and the monotonicity of G: for rationals $r < s$,

$$F(r) \;\le\; G(s) \qquad \text{and} \qquad G(r) \;\le\; F(s) \,. \tag{A.358}$$

The first holds because $s > r$ makes $G(s)$ one of the numbers whose infimum is $F(r)$; the second because every rational $q > s$ has $G(q) \ge G(r)$, so the infimum defining $F(s)$ is at least $G(r)$.

Convergence at continuity points. Let x be a continuity point of F and let $\varepsilon > 0$. By left-continuity of F at x there is $y < x$ with $F(y) > F(x) - \varepsilon$; choose rationals r, s with $y < r < s < x$. Then $F_{n_k}(x) \ge F_{n_k}(s) \to G(s)$, and by Equation (A.358) twice, $G(s) \ge G(r) \ge F(y)$, whence

$$\liminf_k F_{n_k}(x) \;\ge\; F(y) \;>\; F(x) - \varepsilon \,.$$

For the other side, continuity gives $z > x$ with $F(z) < F(x) + \varepsilon$; choose a rational $q \in (x, z)$. Then $F_{n_k}(x) \le F_{n_k}(q) \to G(q)$, and $G(q) \le F(z)$ by the second inequality

of Equation (A.358), so $\limsup_k F_{n_k}(x) \le F(z) < F(x) + \varepsilon$. As $\varepsilon > 0$ was arbitrary, $F_{n_k}(x) \to F(x)$. □

Definition A.174 (Tightness)**.** A family $\{X_n\}$ of random variables is *tight* when for every $\varepsilon > 0$ there is $K < \infty$ with $\Pr(|X_n| > K) \le \varepsilon$ for every n.
Rests on Definition 11.21.

Lemma A.175 (Tightness makes the Helly limit proper)**.** *If in Theorem A.173 the underlying variables are tight, then the limit F satisfies $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to +\infty$: it is a distribution function.*
Rests on Theorem A.173 and Definition A.174.

Proof. *Derives Lemma A.175.*

Given $\varepsilon > 0$ take K as in Definition A.174. A monotone function has at most countably many discontinuities — the open intervals $(F(x^-), F(x^+))$ belonging to distinct jumps are disjoint and each contains a rational — so continuity points are dense and we may pick continuity points $x > K$ and $x' < -K$. Then $F(x) = \lim_k F_{n_k}(x) \ge \liminf_k \Pr(|X_{n_k}| \le K) \ge 1 - \varepsilon$ and $F(x') = \lim_k F_{n_k}(x') \le \varepsilon$. Monotonicity extends both to all larger, respectively smaller, arguments. □

Lemma A.176 (Helly–Bray)**.** *Let F_n, F be distribution functions with $F_n(x) \to F(x)$ at every continuity point of F, and let $g : \mathbb{R} \to \mathbb{R}$ be bounded and continuous. Then*

$$\langle g(X_n) \rangle \longrightarrow \langle g(X) \rangle \,, \tag{A.359}$$

where X_n, X have distribution functions F_n, F.
Rests on Theorems 7.25 and A.173.

Proof. *Derives Lemma A.176.*

Write $\|g\|_\infty = \sup |g|$ and let $\varepsilon > 0$. Since F is a distribution function and its continuity points are dense, choose continuity points $a < b$ with $F(a) < \varepsilon$ and $1 - F(b) < \varepsilon$; then for n large $F_n(a) < 2\varepsilon$ and $1 - F_n(b) < 2\varepsilon$. The contributions to $\langle g(X_n) \rangle$ and $\langle g(X) \rangle$ from outside $(a, b]$ are therefore at most $2\varepsilon \, \|g\|_\infty$ and $\varepsilon \, \|g\|_\infty$ in modulus.

On the compact interval $[a, b]$ the function g is uniformly continuous (Theorem 7.25), so there is $\delta > 0$ with $|g(x) - g(y)| < \varepsilon$ whenever $|x - y| < \delta$ inside $[a, b]$. Choose continuity points $a = x_0 < x_1 < \cdots < x_m = b$ of F with $x_j - x_{j-1} < \delta$ — possible because continuity points are dense. Then for any distribution function H,

$$\left| \int_{(a,b]} g \, \mathrm{d}H - \sum_{j=1}^{m} g(x_j) \, [H(x_j) - H(x_{j-1})] \right| \le \varepsilon \,,$$

because on each $(x_{j-1}, x_j]$ the integrand differs from $g(x_j)$ by less than ε and the increments of H sum to at most 1. Applying this to $H = F_n$ and to $H = F$, and using $F_n(x_j) \to F(x_j)$ at each of the finitely many partition points, gives

$$\limsup_n \, |\langle g(X_n) \rangle - \langle g(X) \rangle| \le 2\varepsilon + 3\varepsilon \, \|g\|_\infty \,.$$

Since $\varepsilon > 0$ was arbitrary, the limit is 0. □

A.18.5 The inversion formula

Theorem A.177 (Fourier inversion for a distribution function)**.** *Let X have distribution function F and characteristic function φ, and let $a < b$. Then*

$$\lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t} \, \varphi(t) \, \mathrm{d}t = \frac{F(b) + F(b^-)}{2} - \frac{F(a) + F(a^-)}{2} \,, \tag{A.360}$$

the integrand being extended by its limit $b-a$ at $t=0$. In particular, if a and b are continuity points of F the right-hand side is $F(b)-F(a)$.

Rests on Definition 11.42 and Lemma A.172.

Proof. *Derives Theorem A.177.*

Write J_T for the integral on the left. For real u, v,

$$\left|\frac{\mathrm{e}^{\mathrm{i}tu}-\mathrm{e}^{\mathrm{i}tv}}{\mathrm{i}t}\right| = \left|\int_v^u \mathrm{e}^{\mathrm{i}ts}\,\mathrm{d}s\right| \;\le\; |u-v|\,, \tag{A.361}$$

so with $u = X-a$, $v = X-b$ the integrand of J_T, written as $\left(\mathrm{e}^{\mathrm{i}t(X-a)}-\mathrm{e}^{\mathrm{i}t(X-b)}\right)/(\mathrm{i}t)$ under the expectation, is bounded by the constant $b-a$. Fubini's theorem (Theorem A.170) — the first of the two quoted uses — therefore applies:

$$J_T = \frac{1}{2\pi}\left\langle \int_{-T}^{T} \frac{\mathrm{e}^{\mathrm{i}t(X-a)}-\mathrm{e}^{\mathrm{i}t(X-b)}}{\mathrm{i}t}\,\mathrm{d}t \right\rangle . \tag{A.362}$$

Now $\mathrm{e}^{\mathrm{i}tu}/(\mathrm{i}t) = (\sin tu)/t - \mathrm{i}\,(\cos tu)/t$, and the cosine terms are odd in t, so they cancel over the symmetric interval. Hence the inner integral is

$$2\int_0^T \frac{\sin\left(t\,(X-a)\right)-\sin\left(t\,(X-b)\right)}{t}\,\mathrm{d}t\,, \tag{A.363}$$

which by Lemma A.172 is bounded by $4C_0 \le 16$ uniformly in T and in X, and converges as $T\to\infty$ to $\pi\left[\operatorname{sgn}(X-a)-\operatorname{sgn}(X-b)\right]$ pointwise. Dominated convergence (Theorem A.169) — the second quoted use — gives

$$\lim_{T\to\infty} J_T = \tfrac{1}{2}\left\langle \operatorname{sgn}(X-a)-\operatorname{sgn}(X-b)\right\rangle . \tag{A.364}$$

The bracket takes the value 0 when $X<a$ or $X>b$, the value 2 when $a<X<b$, and the value 1 when $X=a$ or $X=b$. So

$$\lim_{T\to\infty} J_T = \Pr\left(a<X<b\right) + \tfrac{1}{2}\Pr\left(X=a\right) + \tfrac{1}{2}\Pr\left(X=b\right)\,,$$

which, using $F(x^-)=\Pr(X<x)$ and $F(x)=\Pr(X\le x)$, is exactly the right-hand side of Equation (A.360). □

Corollary A.178 (The characteristic function determines the distribution)**.** *If two random variables have the same characteristic function, they have the same distribution function.*

Rests on Theorem A.177.

Proof. *Derives Corollary A.178.*

Let F_1, F_2 have the common characteristic function φ. The left-hand side of Equation (A.360) depends only on φ, so $F_1(b)-F_1(a) = F_2(b)-F_2(a)$ whenever a, b are continuity points of both. Continuity points of either are dense (the argument of Lemma A.175), so the continuity points common to both are dense as well, their complement being a union of two countable sets. Letting $a\to-\infty$ through such points gives $F_1(b)=F_2(b)$ at every common continuity point b; and every $x\in\mathbb{R}$ is the limit from the right of such points, so right-continuity of both functions gives $F_1=F_2$ everywhere. □

A.18.6 Tightness from continuity at the origin

Lemma A.179 (Truncation estimate)**.** *For any random variable X with characteristic function φ_X and any $u>0$,*

$$\Pr\left(|X|\ge\frac{2}{u}\right) \;\le\; \frac{1}{u}\int_{-u}^{u}\left(1-\Re\varphi_X(t)\right)\mathrm{d}t\,. \tag{A.365}$$

Rests on Definition 11.42.

Proof. *Derives Lemma A.179.*

The integrand $1-\cos(tX)$ is bounded by 2, so Fubini's theorem (Theorem A.170) — the third quoted use — allows the expectation and the t-integral to be exchanged:

$$\frac{1}{u}\int_{-u}^{u}\left(1-\Re\varphi_X(t)\right)\mathrm{d}t=\left\langle\frac{1}{u}\int_{-u}^{u}\left(1-\cos(tX)\right)\mathrm{d}t\right\rangle=\left\langle 2-\frac{2\sin\left(uX\right)}{uX}\right\rangle,$$

the inner integral being elementary and the quotient being read as 2 when $X=0$. Put $y=uX$. The function $1-(\sin y)/y$ is non-negative for every real y, since $|\sin y|\le|y|$; and when $|y|\ge 2$ the bound $|\sin y|\le 1$ gives $1-(\sin y)/y\ge 1-1/|y|\ge\frac{1}{2}$, so that $2\left(1-(\sin y)/y\right)\ge 1$ there. Hence the integrand of the expectation dominates $\mathbf{1}_{\{|uX|\ge 2\}}$ pointwise, and taking expectations gives Equation (A.365). □

Lemma A.180 (Tightness)**.** *Let $\varphi_n\to\varphi$ pointwise on $\mathbb{R}$ with φ continuous at 0. Then the corresponding family $\{X_n\}$ is tight.*

Rests on Lemma A.179 and Definition A.174.

Proof. *Derives Lemma A.180.*

First, $\varphi(0)=\lim_n\varphi_n(0)=1$ by Proposition 11.43(i). Let $\varepsilon>0$. Continuity of φ at 0 gives $u>0$ with $|1-\varphi(t)|<\varepsilon/4$ for $|t|\le u$, whence

$$\left|\frac{1}{u}\int_{-u}^{u}\left(1-\Re\varphi(t)\right)\mathrm{d}t\right|\;\le\;\frac{1}{u}\cdot 2u\cdot\frac{\varepsilon}{4}=\frac{\varepsilon}{2}.$$

The functions $1-\Re\varphi_n$ are bounded by 2 on $[-u,u]$ and converge pointwise to $1-\Re\varphi$, so dominated convergence (Theorem A.169) — the fourth and last quoted use, here applied to the uniform measure on $[-u,u]$ — gives

$$\frac{1}{u}\int_{-u}^{u}\left(1-\Re\varphi_n(t)\right)\mathrm{d}t\;\longrightarrow\;\frac{1}{u}\int_{-u}^{u}\left(1-\Re\varphi(t)\right)\mathrm{d}t,$$

so there is N with the left side below ε for $n\ge N$. By Lemma A.179, $\Pr(|X_n|\ge 2/u)\le\varepsilon$ for $n\ge N$. Each of the finitely many remaining variables $X_1,\dots,X_{N-1}$ is individually tight, its distribution function tending to 0 and 1 at the two ends, so there is K_0 with $\Pr(|X_n|>K_0)\le\varepsilon$ for $n<N$. Take $K=\max(2/u,K_0)$. □

This lemma is the whole reason the hypothesis of Theorem A.168 mentions the origin, and it shows the hypothesis cannot be dropped: the Cauchy-type variables $X_n=nY$ with Y standard Cauchy have $\varphi_n(t)=\mathrm{e}^{-n|t|}$, which converges pointwise to the function equal to 1 at $t=0$ and 0 elsewhere — not continuous at the origin — and the mass escapes to infinity, the distributions converging to nothing.

A.18.7 Proof of the theorem

Proof of Theorem A.168. *Derives Theorem A.168.*

By Lemma A.180 the family $\{X_n\}$ is tight.

Every subsequence has a convergent further subsequence with the same limit. Let (n_j) be any subsequence. By Helly's theorem (Theorem A.173) it has a further subsequence (n_{j_k}) along which $F_{n_{j_k}}\to F$ at every continuity point of some non-decreasing right-continuous $F:\mathbb{R}\to[0,1]$; by Lemma A.175 and tightness, F is a genuine distribution function. Let X be a random variable with distribution function F. Applying the Helly–Bray lemma (Lemma A.176) to the bounded continuous functions $x\mapsto\cos(tx)$ and $x\mapsto\sin(tx)$, for each fixed t, gives

$$\varphi_{n_{j_k}}(t)=\left\langle\cos\left(tX_{n_{j_k}}\right)\right\rangle+\mathrm{i}\left\langle\sin\left(tX_{n_{j_k}}\right)\right\rangle\;\longrightarrow\;\varphi_X(t).\tag{A.366}$$

But by hypothesis $\varphi_{n_{j_k}}(t)\to\varphi(t)$. Limits in $\mathbb{C}$ are unique, so $\varphi=\varphi_X$: the limit function *is* a characteristic function, which is the first assertion of the theorem. Moreover, by

Corollary A.178 the distribution function F is determined by φ alone, so *every* subsequential Helly limit obtained this way is the same function F.

Hence the whole sequence converges. Let x be a continuity point of F and suppose $F_n(x)$ does not converge to $F(x)$. Then there are $\eta > 0$ and a subsequence (n_j) with $\left|F_{n_j}(x) - F(x)\right| \geq \eta$ for every j. By the previous paragraph applied to (n_j) there is a further subsequence along which the distribution functions converge, at every continuity point of F and in particular at x, to F — contradicting the standing inequality. Therefore $F_n(x) \to F(x)$ at every continuity point of F, which is $X_n \to X$ in distribution. □

Corollary A.181 (The converse)**.** *If $X_n \to X$ in distribution then $\varphi_n(t) \to \varphi_X(t)$ for every t, and the limit is continuous everywhere.*

Rests on Lemma A.176 and Theorem A.168.

Proof. *Derives Corollary A.181.*

The first claim is Equation (A.366) read for the full sequence, the Helly–Bray lemma applying directly since $F_n \to F_X$ at continuity points by hypothesis. For continuity of φ_X: Equation (A.361) read with $u = X$, $v = 0$ and $t = h$ gives $\left|\mathrm{e}^{\mathrm{i}hX} - 1\right| \leq |h|\,|X|$, while the modulus of a difference of two unit complex numbers is at most 2, so $\left|\mathrm{e}^{\mathrm{i}(t+h)X} - \mathrm{e}^{\mathrm{i}tX}\right| = \left|\mathrm{e}^{\mathrm{i}hX} - 1\right| \leq \min(|hX|\,, 2)$ and hence $|\varphi_X(t+h) - \varphi_X(t)| \leq \langle \min(|hX|\,, 2)\rangle$, a bound independent of t that tends to 0 with h — by the same truncation as in Lemma 11.44: split the expectation at $|X| = K$, bound the inner part by $|h|\,K$ and the outer by $2\Pr(|X| > K)$, and choose K then h. So φ_X is in fact uniformly continuous. □

Remark A.182. Appendix A.18 discharges the proof obligation of Theorem 11.45, the one analytic input that Section 11.3.1 of Chapter 11 had to quote. With it, the derivation of the central limit theorem Theorem 11.46 in Section 11.3.2 is complete from the axioms of the chapter: Lemma 11.44 supplies the second-order expansion, Equation (11.36) the telescoping estimate, and this appendix the passage back from transforms to distributions. The same passage is the last step of Appendix A.19, and Theorem A.177 is the starting point of Appendix A.20. Two theorems of Lebesgue theory are assumed and are named in Remark A.171; nothing else is.

A.19 The Lindeberg–Feller Central Limit Theorem

This appendix proves Theorem 11.48 of Chapter 11: a triangular array of independent, individually negligible summands whose tails satisfy Lindeberg's condition has an asymptotically Gaussian normalised sum. It is the form of the central limit theorem the error theory of Chapter 2 actually needs, because the disturbances that perturb a measurement — a thermal drift, a vibration, a quantisation step, a reading error — are independent but certainly not identically distributed, so the Lindeberg–Lévy theorem Theorem 11.46 does not apply to them.

The argument is the characteristic-function argument of Section 11.3.2 carried out uniformly in the summand index. Three of its four ingredients are already in the chapter: the elementary properties of characteristic functions (Proposition 11.43), the exponential remainder bound Equation (11.31), and the telescoping product estimate Equation (11.36). The fourth is Lévy's continuity theorem, proved in Appendix A.18. Nothing else is assumed, and in particular the Lindeberg condition is shown here to imply the uniform asymptotic negligibility that the expansion needs — that implication is not an extra hypothesis.

A.19.1 Statement and notation

Theorem A.183 (Lindeberg–Feller)**.** *For each n let $X_{n,1}, \dots, X_{n,k_n}$ be independent random variables with $\langle X_{n,k} \rangle = 0$ and finite variances $\sigma_{n,k}^2 = \operatorname{var} X_{n,k}$, and put $s_n^2 = \sum_{k=1}^{k_n} \sigma_{n,k}^2$, assumed strictly positive. If for every $\varepsilon > 0$*

$$L_n(\varepsilon) \equiv \frac{1}{s_n^2} \sum_{k=1}^{k_n} \left\langle X_{n,k}^2 \, \mathbf{1}_{\{|X_{n,k}| > \varepsilon s_n\}} \right\rangle \longrightarrow 0 \qquad (n \to \infty)\,, \tag{A.367}$$

then $s_n^{-1} \sum_{k=1}^{k_n} X_{n,k} \to \mathcal{N}(0,1)$ in distribution.

Rests on Definitions 11.28 and 11.42 and Proposition 11.43.

It is convenient to normalise once and for all. Put

$$Y_{n,k} = \frac{X_{n,k}}{s_n}\,, \qquad \tau_{n,k}^2 = \left\langle Y_{n,k}^2 \right\rangle = \frac{\sigma_{n,k}^2}{s_n^2}\,, \qquad \sum_{k=1}^{k_n} \tau_{n,k}^2 = 1\,, \tag{A.368}$$

and write $\varphi_{n,k}$ for the characteristic function of $Y_{n,k}$. The event $\{|X_{n,k}| > \varepsilon s_n\}$ is the event $\{|Y_{n,k}| > \varepsilon\}$, so Equation (A.367) reads

$$L_n(\varepsilon) = \sum_{k=1}^{k_n} \left\langle Y_{n,k}^2 \, \mathbf{1}_{\{|Y_{n,k}| > \varepsilon\}} \right\rangle \longrightarrow 0 \qquad \text{for every } \varepsilon > 0\,. \tag{A.369}$$

Every quantity in Equation (A.368) is a pure number: the $X_{n,k}$ carry whatever SI unit the measurement error carries, and s_n carries the same one, so the ratio is dimensionless and so is the argument t of the characteristic functions below.

A.19.2 Lindeberg's condition implies uniform asymptotic negligibility

Lemma A.184 (Uniform asymptotic negligibility)**.** *Under Equation (A.369),*

$$\max_{1 \le k \le k_n} \tau_{n,k}^2 = \max_{1 \le k \le k_n} \frac{\operatorname{var} X_{n,k}}{s_n^2} \longrightarrow 0\,. \tag{A.370}$$

Rests on Equation (A.367).

Proof. *Derives Lemma A.184.*

Fix $\varepsilon > 0$ and k. Splitting the expectation at $|Y_{n,k}| = \varepsilon$ and bounding $Y_{n,k}^2$ by ε^2 on the inner piece,

$$\tau_{n,k}^2 = \left\langle Y_{n,k}^2 \, \mathbf{1}_{\{|Y_{n,k}| \le \varepsilon\}} \right\rangle + \left\langle Y_{n,k}^2 \, \mathbf{1}_{\{|Y_{n,k}| > \varepsilon\}} \right\rangle \le \varepsilon^2 + \left\langle Y_{n,k}^2 \, \mathbf{1}_{\{|Y_{n,k}| > \varepsilon\}} \right\rangle . \tag{A.371}$$

The last term is one summand of the non-negative sum $L_n(\varepsilon)$, hence at most $L_n(\varepsilon)$, and the bound is the same for every k:

$$\max_k \tau_{n,k}^2 \le \varepsilon^2 + L_n(\varepsilon)\,.$$

Letting $n \to \infty$ gives $\limsup_n \max_k \tau_{n,k}^2 \le \varepsilon^2$, and $\varepsilon > 0$ was arbitrary. □

This is Equation (11.39) of Chapter 11, proved. It is what licenses treating each factor of the product below as close to 1, and it is used twice: to know $1 - \tfrac{1}{2} t^2 \tau_{n,k}^2$ lies in $[0,1]$, and to control the sum of squares in Lemma A.187.

A.19.3 The uniform second-order expansion

Lemma 11.44 expands one characteristic function to second order, with a remainder $o(t^2)$ that depends on the distribution. Here there is a different distribution for every (n,k) and the remainders must be summed, so what is needed is a bound on the *total* remainder, uniform in the array. The Lindeberg condition is exactly what supplies it.

Lemma A.185 (Total remainder)**.** *For every $t \in \mathbb{R}$,*

$$\Delta_n(t) \equiv \sum_{k=1}^{k_n} \left| \varphi_{n,k}(t) - \left(1 - \tfrac{1}{2}t^2\tau_{n,k}^2\right) \right| \;\longrightarrow\; 0\,. \tag{A.372}$$

Rests on Equations (11.31) and (A.367).

Proof. *Derives Lemma A.185.*

Apply Equation (11.31) with $n = 2$ and $x = tY_{n,k}$:

$$\left| \mathrm{e}^{\mathrm{i}tY_{n,k}} - 1 - \mathrm{i}tY_{n,k} + \tfrac{1}{2}t^2Y_{n,k}^2 \right| \;\le\; \min\!\left(\tfrac{1}{6}\,|t|^3\,|Y_{n,k}|^3\,,\; t^2Y_{n,k}^2 \right).$$

Take expectations and use $|\langle Z\rangle| \le \langle |Z|\rangle$; since $\langle Y_{n,k}\rangle = 0$ and $\left\langle Y_{n,k}^2 \right\rangle = \tau_{n,k}^2$ the left side becomes the kth term of $\Delta_n(t)$, so

$$\left| \varphi_{n,k}(t) - 1 + \tfrac{1}{2}t^2\tau_{n,k}^2 \right| \;\le\; \left\langle \min\!\left(\tfrac{1}{6}\,|t|^3\,|Y_{n,k}|^3\,,\; t^2Y_{n,k}^2 \right) \right\rangle . \tag{A.373}$$

Fix $\varepsilon > 0$ and split the expectation at $|Y_{n,k}| = \varepsilon$, using a different branch of the minimum on each piece. Where $|Y_{n,k}| \le \varepsilon$ the first branch gives $\tfrac{1}{6}\,|t|^3\,|Y_{n,k}|^3 \le \tfrac{1}{6}\,|t|^3\,\varepsilon\,Y_{n,k}^2$; where $|Y_{n,k}| > \varepsilon$ the second gives $t^2Y_{n,k}^2$. Hence

$$\left| \varphi_{n,k}(t) - 1 + \tfrac{1}{2}t^2\tau_{n,k}^2 \right| \;\le\; \frac{\varepsilon\,|t|^3}{6}\,\tau_{n,k}^2 + t^2\left\langle Y_{n,k}^2\,\mathbf{1}_{\{|Y_{n,k}|>\varepsilon\}} \right\rangle . \tag{A.374}$$

From Equation (A.373) (split the expectation at $|Y_{n,k}| = \varepsilon$ and use one branch of the minimum on each piece).

Summing over k and using $\sum_k \tau_{n,k}^2 = 1$ from Equation (A.368) and the definition Equation (A.369),

$$\Delta_n(t) \;\le\; \frac{\varepsilon\,|t|^3}{6} + t^2 L_n(\varepsilon)\,. \tag{A.375}$$

For fixed t and ε the second term tends to 0, so $\limsup_n \Delta_n(t) \le \varepsilon\,|t|^3/6$; and $\varepsilon > 0$ was arbitrary. □

The contrast with Lemma 11.44 is worth stating plainly. There the remainder was $o(t^2)$ as $t \to 0$ for a *single* distribution, and the identically distributed case could rescale one such estimate n times. Here t is fixed and n grows, and the estimate Equation (A.375) is a statement about the array as a whole: the cubic branch of the minimum is used on the bulk of each summand and the quadratic branch on its tail, and Equation (A.367) is precisely the hypothesis that the tails contribute nothing in total.

A.19.4 Two product estimates

Lemma A.186 (Replacement)**.** *For every $t \in \mathbb{R}$ there is $N(t)$ such that for $n \ge N(t)$*

$$\left| \prod_{k=1}^{k_n} \varphi_{n,k}(t) - \prod_{k=1}^{k_n} \left(1 - \tfrac{1}{2}t^2\tau_{n,k}^2\right) \right| \;\le\; \Delta_n(t) \;\longrightarrow\; 0\,. \tag{A.376}$$

Rests on Equation (11.36) and Lemmas A.184 and A.185.

Proof. *Derives Lemma A.186.*

Set $a_k = \varphi_{n,k}(t)$ and $b_k = 1 - \frac{1}{2}t^2\tau_{n,k}^2$. By Proposition 11.43(i), $|a_k| \le 1$. For the b_k: by Lemma A.184 there is $N(t)$ with $\frac{1}{2}t^2 \max_k \tau_{n,k}^2 \le 1$ for $n \ge N(t)$, and then every b_k lies in $[0,1]$. The telescoping inequality Equation (11.36) applies to such factors and gives $|\prod a_k - \prod b_k| \le \sum_k |a_k - b_k|$, which is $\Delta_n(t)$ by Equation (A.372). □

Lemma A.187 (Exponentiation)**.** *For every $t \in \mathbb{R}$,*

$$\prod_{k=1}^{k_n} \left(1 - \tfrac{1}{2}t^2\tau_{n,k}^2\right) \longrightarrow \mathrm{e}^{-t^2/2}. \tag{A.377}$$

Rests on Equation (11.36), Lemma A.184, and Theorem 7.38.

Proof. *Derives Lemma A.187.*

Write $z_k = \frac{1}{2}t^2\tau_{n,k}^2 \ge 0$, so that $\sum_k z_k = \frac{1}{2}t^2$ by Equation (A.368) and $\max_k z_k \to 0$ by Lemma A.184. Take n large enough that $\max_k z_k \le 1$; then both $1 - z_k$ and e^{-z_k} lie in $[0,1]$ and Equation (11.36) applies:

$$\left|\prod_k (1 - z_k) - \prod_k \mathrm{e}^{-z_k}\right| \le \sum_k \left|1 - z_k - \mathrm{e}^{-z_k}\right|. \tag{A.378}$$

Taylor's theorem with Lagrange remainder (Theorem 7.38) applied to $z \mapsto \mathrm{e}^{-z}$ gives $\mathrm{e}^{-z} = 1 - z + \frac{1}{2}z^2\mathrm{e}^{-\xi}$ with ξ between 0 and z, so for $z \ge 0$ one has $|\mathrm{e}^{-z} - 1 + z| \le \frac{1}{2}z^2$. Hence

$$\sum_k \left|1 - z_k - \mathrm{e}^{-z_k}\right| \le \frac{1}{2}\sum_k z_k^2 \le \frac{1}{2}\left(\max_k z_k\right)\sum_k z_k = \frac{t^2}{4}\max_k z_k \longrightarrow 0,$$

the middle step bounding each term by $z_k^2/2$, factoring out the largest z_k and using $\sum_k z_k = t^2/2$; while $\prod_k \mathrm{e}^{-z_k} = \mathrm{e}^{-\sum_k z_k} = \mathrm{e}^{-t^2/2}$ exactly. □

A.19.5 Proof of the theorem

Proof of Theorem A.183. *Derives Theorem A.183.*

Write $T_n = s_n^{-1}\sum_k X_{n,k} = \sum_k Y_{n,k}$. The $Y_{n,k}$, $k = 1, \dots, k_n$, are independent, so Proposition 11.43(iii), extended from two factors to k_n by induction, gives

$$\varphi_{T_n}(t) = \prod_{k=1}^{k_n} \varphi_{n,k}(t). \tag{A.379}$$

Fix $t \in \mathbb{R}$. By the triangle inequality,

$$\left|\varphi_{T_n}(t) - \mathrm{e}^{-t^2/2}\right| \le \left|\prod_k \varphi_{n,k}(t) - \prod_k \left(1 - \tfrac{1}{2}t^2\tau_{n,k}^2\right)\right| + \left|\prod_k \left(1 - \tfrac{1}{2}t^2\tau_{n,k}^2\right) - \mathrm{e}^{-t^2/2}\right|,$$

whose first term tends to 0 by Lemma A.186 and whose second tends to 0 by Lemma A.187. Hence

$$\varphi_{T_n}(t) \longrightarrow \mathrm{e}^{-t^2/2} \qquad \text{for every } t \in \mathbb{R}. \tag{A.380}$$

From Equations (A.376), (A.377) and (A.379) (replace each factor by its quadratic approximation, then the product of those by the exponential of the sum).

The limit is continuous at $t = 0$, and it is the characteristic function of the standard Gaussian: as shown in the derivation of Theorem 11.46, the function $g(t) = \int \mathrm{e}^{itu}\mathrm{e}^{-u^2/2}\mathrm{d}u/\sqrt{2\pi}$ satisfies $g'(t) = -t\,g(t)$ with $g(0) = 1$, whence $g(t) = \mathrm{e}^{-t^2/2}$. Theorem A.168 therefore gives $T_n \to \mathcal{N}(0,1)$ in distribution. □

Corollary A.188 (Lindeberg–Lévy is the special case)**.** *Theorem 11.46 follows from Theorem A.183.*

Rests on Theorem A.183.

Proof. *Derives Corollary A.188.*

Let $X_1, X_2, \dots$ be independent and identically distributed with mean μ and variance $\sigma^2 \in (0, \infty)$, and set $k_n = n$, $X_{n,k} = X_k - \mu$. Then $\sigma_{n,k}^2 = \sigma^2$ and $s_n^2 = n\sigma^2$, so $Z_n = s_n^{-1} \sum_k X_{n,k}$ is the statistic of Equation (11.33). The Lindeberg sum is

$$L_n(\varepsilon) = \frac{n}{n\sigma^2} \left\langle (X-\mu)^2 \mathbf{1}_{\{|X-\mu| > \varepsilon\sigma\sqrt{n}\}} \right\rangle = \frac{1}{\sigma^2} \left\langle (X-\mu)^2 \mathbf{1}_{\{|X-\mu| > \varepsilon\sigma\sqrt{n}\}} \right\rangle ,$$

which tends to 0 as $n \to \infty$ because $\langle (X-\mu)^2 \rangle = \sigma^2$ is finite, a finite integral being one whose tail contribution vanishes — the same elementary fact used at Equation (11.32). The threshold $\varepsilon\sigma\sqrt{n}$ grows without bound, so the truncation level recedes and the tail expectation goes to zero for each fixed ε. □

Corollary A.189 (Lyapunov's condition)**.** *If there is $\delta > 0$ with*

$$\frac{1}{s_n^{2+\delta}} \sum_{k=1}^{k_n} \left\langle |X_{n,k}|^{2+\delta} \right\rangle \longrightarrow 0 , \tag{A.381}$$

then Equation (A.367) holds, and hence so does the conclusion of Theorem A.183.

Rests on Theorem A.183 and Equation (A.367).

Proof. *Derives Corollary A.189.*

On the event $|Y_{n,k}| > \varepsilon$ one has $1 < |Y_{n,k}|^\delta / \varepsilon^\delta$, so $Y_{n,k}^2 \mathbf{1}_{\{|Y_{n,k}| > \varepsilon\}} \le \varepsilon^{-\delta} |Y_{n,k}|^{2+\delta}$ pointwise. Taking expectations and summing over k,

$$L_n(\varepsilon) \le \frac{1}{\varepsilon^\delta} \sum_{k=1}^{k_n} \left\langle |Y_{n,k}|^{2+\delta} \right\rangle = \frac{1}{\varepsilon^\delta} \cdot \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{k_n} \left\langle |X_{n,k}|^{2+\delta} \right\rangle ,$$

which tends to 0 for each fixed $\varepsilon > 0$ by Equation (A.381). □

Remark A.190 (What the theorem still does not say). The condition is sufficient and not necessary, and Chapter 11 exhibits both failures of the converse: summands equal to $\pm\sqrt{n}$ with probability $1/(2n)$ have vanishing variance shares yet Lindeberg sum 1 for every n, and adjoining one $\mathcal{N}(0, n)$ summand to n standard Gaussians breaks Equation (A.367) while leaving the normalised sum exactly $\mathcal{N}(0, 1)$. What Lemma A.184 establishes is the one implication the error theory uses — no single disturbance may carry a fixed fraction of the total variance — and it is an implication, not a converse: a dominant source violates Equation (A.367), but the absence of a dominant source does not by itself deliver a Gaussian. Feller's companion theorem, which closes the loop inside the negligible class, is not proved here [1446].

Remark A.191. Appendix A.19 discharges the proof obligation of Theorem 11.48, stated in Section 11.3.2 of Chapter 11 and used there to justify treating the aggregate of many small, non-identically distributed measurement disturbances as Gaussian — the licence on which the standard uncertainty of Chapter 2 rests. Two by-products are worth noting: Equation (11.39) of the chapter is Lemma A.184 here, proved rather than asserted, and Lyapunov's condition, which the chapter quotes, is Corollary A.189. The only input this appendix does not derive is Lévy's continuity theorem, which is Appendix A.18.

A.20 The Berry–Esseen Inequality

This appendix proves the uniform rate Equation (11.40) quoted in Remark 11.49 of Chapter 11: the Kolmogorov distance between the distribution of a standardised sum of n independent identically distributed variables and the standard Gaussian is at most a constant times $\rho/(\sigma^3\sqrt{n})$, where ρ is the third absolute central moment. The central limit theorem Theorem 11.46 is a limit statement and an uncertainty budget needs a rate; this is the rate, and it is what converts "the errors are asymptotically Gaussian" into a number a reader can check.

The proof has three parts, in the order Esseen gave them. First a *smoothing inequality*, which bounds the largest vertical gap between a distribution function and a smooth comparison function by an integral of the difference of their Fourier–Stieltjes transforms over a finite interval, plus a penalty for truncating the interval (Appendix A.20.3). Then an estimate of that difference for the case at hand, valid on $|t| \lesssim \sqrt{n}\,\sigma^3/\rho$ and carrying a Gaussian decay factor (Appendix A.20.4). Then the integration, which balances the two terms and fixes an admissible constant (Appendix A.20.5).

Notation. Throughout this appendix $\varphi(x) = \mathrm{e}^{-x^2/2}/\sqrt{2\pi}$ is the standard Gaussian density and Φ its distribution function, as in Definition 11.28; characteristic functions are written ψ, to keep the two apart. For a distribution function F we write $\psi_F(t) = \int \mathrm{e}^{\mathrm{i}tx}\,\mathrm{d}F(x)$, which is $\left\langle \mathrm{e}^{\mathrm{i}tX} \right\rangle$ for a variable X with that distribution function (Definition 11.42). The one analytic input is the inversion formula Theorem A.177 of Appendix A.18, and through it the two Lebesgue-theory statements named in Remark A.171; everything else below is derived.

A.20.1 Statement

Theorem A.192 (Berry–Esseen)**.** *Let $X_1, X_2, \dots$ be independent and identically distributed with mean μ, variance $\sigma^2 \in (0,\infty)$ and finite third absolute central moment $\rho = \left\langle |X-\mu|^3 \right\rangle$. Put $Z_n = (\sigma\sqrt{n})^{-1}\sum_{i=1}^{n}(X_i-\mu)$ and let F_n be its distribution function. Then for every $n \geq 1$*

$$\Delta_n \equiv \sup_{z\in\mathbb{R}} |F_n(z) - \Phi(z)| \;\leq\; \frac{C\rho}{\sigma^3\sqrt{n}}\,, \qquad C = 7.23\,. \tag{A.382}$$

Rests on Definitions 11.28 and 11.42 and Equation (11.31).

The constant produced here is honest but crude; Remark A.201 says what is known about the smallest admissible one and why this proof does not reach it. The quantity ρ/σ^3 is a pure number — both moments carry the cube of the SI unit of X — as it must be, since the left-hand side is a difference of probabilities.

A.20.2 The Fejér kernel

Lemma A.193 (A cosine integral)**.** *For every $a \in \mathbb{R}$,*

$$\int_{-\infty}^{\infty} \frac{1-\cos(ax)}{x^2}\,\mathrm{d}x = \pi\,|a|\,. \tag{A.383}$$

Rests on Lemma A.172 and Equation (7.28).

Proof. *Derives Lemma A.193.*

For $a = 0$ both sides vanish. For $a > 0$ the substitution $u = ax$ (Equation (7.27)) turns the left side into $a\int_{-\infty}^{\infty}(1-\cos u)\,u^{-2}\mathrm{d}u$, and the integrand is even, so it suffices to show $\int_0^{\infty}(1-\cos u)\,u^{-2}\mathrm{d}u = \pi/2$. Integrate by parts (Equation (7.28)) with $u^{-2}\mathrm{d}u = -\mathrm{d}\,(u^{-1})$:

$$\int_0^{\infty}\frac{1-\cos u}{u^2}\,\mathrm{d}u = \left[-\frac{1-\cos u}{u}\right]_0^{\infty} + \int_0^{\infty}\frac{\sin u}{u}\,\mathrm{d}u = 0 + \frac{\pi}{2}\,,$$

by Lemma A.172. Both boundary terms vanish: at infinity because the numerator is bounded by 2, and at zero because $1-\cos u \le u^2/2$. Oddness of $\cos(ax)$ in a — it is even — gives the case $a<0$ from the case $-a>0$. □

Lemma A.194 (The Fejér kernel and its transform)**.** *For $T>0$ define*

$$K_T(x)=\frac{1-\cos(Tx)}{\pi T x^2}\,, \tag{A.384}$$

extended by $K_T(0)=T/(2\pi)$. Then $K_T\ge 0$, K_T is even, $\int_{\mathbb{R}} K_T=1$, and

$$\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx} K_T(x)\,\mathrm{d}x=\begin{cases}1-\dfrac{|t|}{T} & |t|\le T\,,\\ 0 & |t|>T\,.\end{cases} \tag{A.385}$$

Moreover $\int_{|x|>\eta} K_T(x)\,\mathrm{d}x \le 4/(\pi T\eta)$ for every $\eta>0$.

Rests on Lemma A.193.

Proof. *Derives Lemma A.194.*

Non-negativity and evenness are immediate, and Equation (A.383) with $a=T$ gives $\int K_T=\pi T/(\pi T)=1$. For the tail bound use $1-\cos\le 2$, so that $K_T(x)\le 2/(\pi T x^2)$ and

$$\int_{|x|>\eta} K_T(x)\,\mathrm{d}x \;\le\; 2\int_{\eta}^{\infty}\frac{2}{\pi T x^2}\,\mathrm{d}x=\frac{4}{\pi T\eta}\,.$$

For Equation (A.385): the sine part of $\mathrm{e}^{\mathrm{i}tx}$ pairs with an even function and integrates to zero, so the transform is $\int\cos(tx)K_T(x)\mathrm{d}x$, which is real. Write the numerator of the integrand, $(1-\cos(Tx))\cos(tx)$, using the product formula $\cos A\cos B=\frac12[\cos(A-B)+\cos(A+B)]$:

$$(1-\cos(Tx))\cos(tx)=-(1-\cos(tx))+\tfrac12(1-\cos((T+t)x))+\tfrac12(1-\cos((T-t)x))\,, \tag{A.386}$$

the constants cancelling because $-1+\frac12+\frac12=0$. Dividing by πTx^2, integrating, and applying Equation (A.383) three times,

$$\int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}tx}K_T(x)\,\mathrm{d}x=\frac{1}{\pi T}\left[-\pi|t|+\frac{\pi}{2}|T+t|+\frac{\pi}{2}|T-t|\right]. \tag{A.387}$$

From Equations (A.383) and (A.386) (divide by πTx^2 and integrate each of the three pieces).

For $|t|\le T$ the two moduli on the right are $T+t$ and $T-t$, so the bracket is $\pi(T-|t|)$ and the value is $1-|t|/T$. For $t>T$ they are $T+t$ and $t-T$, whose half-sum is $\pi t=\pi|t|$, and the bracket vanishes; the case $t<-T$ is the mirror image. □

A.20.3 Esseen's smoothing inequality

The inversion formula of Appendix A.18 recovers a distribution function from its transform. What is wanted here is the *difference* of two of them, at a single point, and in a form that needs no limit.

Lemma A.195 (Riemann–Lebesgue, continuous compactly supported case)**.** *Let $h:\mathbb{R}\to\mathbb{C}$ be continuous and vanish outside a bounded interval. Then $\int_{\mathbb{R}} h(t)\mathrm{e}^{-\mathrm{i}ta}\,\mathrm{d}t\to 0$ as $|a|\to\infty$.*

Rests on Theorem 7.25.

Proof. *Derives Lemma A.195.*

Let h vanish outside $[-T,T]$ and put $A(a)=\int_{\mathbb{R}} h(t)\mathrm{e}^{-\mathrm{i}ta}\mathrm{d}t$. Since $\mathrm{e}^{-\mathrm{i}(t+\pi/a)a}=-\mathrm{e}^{-\mathrm{i}ta}$, the substitution $t\mapsto t+\pi/a$ (Equation (7.27)) gives $A(a)=-\int_{\mathbb{R}} h(t-\pi/a)\mathrm{e}^{-\mathrm{i}ta}\mathrm{d}t$, so

$$2\,|A(a)|=\left|\int_{\mathbb{R}}\left[h(t)-h\Big(t-\frac{\pi}{a}\Big)\right]\mathrm{e}^{-\mathrm{i}ta}\,\mathrm{d}t\right| \;\le\; \int_{-T-\pi/|a|}^{T+\pi/|a|}\left|h(t)-h\Big(t-\frac{\pi}{a}\Big)\right|\mathrm{d}t\,.$$

The function h is uniformly continuous on $\mathbb{R}$ (Theorem 7.25 on a compact interval containing the support with room to spare, and $h \equiv 0$ outside), so the integrand tends to 0 uniformly and the interval length stays bounded. □

Lemma A.196 (Inversion for a difference)**.** *Let F_1, F_2 be distribution functions with transforms ψ_1, ψ_2, and suppose $h(t) = (\psi_1(t) - \psi_2(t))/(\mathrm{i}t)$, extended by its limit at $t = 0$, is continuous on $\mathbb{R}$ and vanishes outside $[-T, T]$. Then at every point x at which both F_1 and F_2 are continuous,*

$$F_1(x) - F_2(x) = -\frac{1}{2\pi} \int_{-T}^{T} \frac{\psi_1(t) - \psi_2(t)}{\mathrm{i}t} \, \mathrm{e}^{-\mathrm{i}tx} \, \mathrm{d}t \,. \tag{A.388}$$

Rests on Theorem A.177 and Lemma A.195.

Proof. *Derives Lemma A.196.*

Let $a < x$ be a common continuity point. Applying Equation (A.360) to F_1 and to F_2 and subtracting — both limits exist, so their difference is the limit of the difference — gives, for every $T' \geq T$,

$$[F_1(x) - F_1(a)] - [F_2(x) - F_2(a)] = \frac{1}{2\pi} \int_{-T'}^{T'} \left(\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tx} \right) h(t) \, \mathrm{d}t \,,$$

and since h vanishes outside $[-T, T]$ the integral does not depend on T' and no limit is needed. Split it into its two terms. By Lemma A.195 the term carrying $\mathrm{e}^{-\mathrm{i}ta}$ tends to 0 as $a \to -\infty$ through continuity points, while $F_1(a) \to 0$ and $F_2(a) \to 0$. What survives is Equation (A.388). □

Lemma A.197 (The transform is differentiable at the origin)**.** *If X has $\langle |X| \rangle < \infty$ and characteristic function ψ_X, then $\psi_X(t) = 1 + \mathrm{i}t \, \langle X \rangle + o(|t|)$ as $t \to 0$.*

Rests on Equation (11.31) and Definition 11.42.

Proof. *Derives Lemma A.197.*

Equation (11.31) at order one gives $\left| \mathrm{e}^{\mathrm{i}x} - 1 - \mathrm{i}x \right| \leq \min\left(x^2/2, \, 2\,|x|\right)$. Put $x = tX$, take expectations, use $|\langle Z \rangle| \leq \langle |Z| \rangle$ and divide by $|t|$:

$$\frac{|\psi_X(t) - 1 - \mathrm{i}t \, \langle X \rangle|}{|t|} \;\leq\; \left\langle \min\left(\tfrac{1}{2}\,|t|\, X^2, \; 2\,|X| \right) \right\rangle \;\leq\; \tfrac{1}{2}\,|t|\, K^2 + 2 \left\langle |X| \; \mathbf{1}_{\{|X| > K\}} \right\rangle ,$$

the last step splitting the expectation at $|X| = K$ and using one branch of the minimum on each piece, exactly as at Equation (11.32). Given $\varepsilon > 0$ choose K with the second term below $\varepsilon/2$ — possible because $\langle |X| \rangle$ is finite — and then $|t| < \varepsilon / K^2$. □

Theorem A.198 (Esseen's smoothing inequality)**.** *Let F be a distribution function and G a non-decreasing differentiable function with $G(-\infty) = 0$, $G(+\infty) = 1$ and $\sup_x G'(x) \leq m < \infty$; suppose both have a finite first absolute moment. Write ψ_F, ψ_G for their transforms and $\Delta = \sup_x |F(x) - G(x)|$. Then for every $T > 0$*

$$\Delta \;\leq\; \frac{1}{\pi} \int_{-T}^{T} \left| \frac{\psi_F(t) - \psi_G(t)}{t} \right| \mathrm{d}t + \frac{24\, m}{\pi T} \,. \tag{A.389}$$

Rests on Lemmas A.194 and A.196.

Proof. *Derives Theorem A.198.*

Write $H = F - G$, so $|H| \leq 1$ and $\Delta \leq 1$ is finite; if $\Delta = 0$ there is nothing to prove, and if the integral on the right is infinite the inequality is trivial, so assume both finite and $\Delta > 0$. Let W have density K_T, independently of everything else, and let F_T, G_T be the distribution functions of $U + W$ and $V + W$, where U has distribution function F and V

has G. Both are continuous, being convolutions with a density, and by Proposition 11.43(iii) and Equation (A.385) their transforms are $\psi_F \hat{K}_T$ and $\psi_G \hat{K}_T$ with $\hat{K}_T(t) = (1 - |t|/T)$ on $[-T, T]$ and 0 outside. Since K_T is even,

$$H_T(x) \equiv F_T(x) - G_T(x) = \int_{-\infty}^{\infty} H(x+s)\, K_T(s)\, \mathrm{d}s\,. \tag{A.390}$$

The transform side. The function $(\psi_F - \psi_G)\,\hat{K}_T/(\mathrm{i}t)$ is continuous — at $t = 0$ because Lemma A.197, applicable since both first absolute moments are finite, gives $\psi_F(t) - \psi_G(t) = \mathrm{i}t\,(\langle U\rangle - \langle V\rangle) + o(|t|)$, and at $t = \pm T$ because $\hat{K}_T$ vanishes there — and is supported in $[-T, T]$. So Lemma A.196 applies, and since $\left|\hat{K}_T\right| \le 1$,

$$\sup_x |H_T(x)| \;\le\; \frac{1}{2\pi} \int_{-T}^{T} \left| \frac{\psi_F(t) - \psi_G(t)}{t} \right| \mathrm{d}t\,. \tag{A.391}$$

The kernel side. Put $\eta = \Delta/(2m)$ and $\kappa = \int_{|s|>\eta} K_T(s)\,\mathrm{d}s$, so that $\kappa \le 4/(\pi T\eta) = 8m/(\pi T\Delta)$ by Lemma A.194. Let $\varepsilon \in (0, \Delta/2)$. Two cases, according to which sign of H realises the supremum.

Suppose first $\sup_x H(x) = \Delta$ and pick b with $H(b) > \Delta - \varepsilon$. For $y \ge b$ the mean value theorem (Theorem 7.35) gives $G(y) \le G(b) + m(y - b)$, while $F(y) \ge F(b)$; hence $H(y) \ge H(b) - m(y - b)$. Put $x_* = b + \eta$. For $|s| \le \eta$ the point $x_* + s$ lies at distance $\eta + s \in [0, 2\eta]$ to the right of b, so

$$H(x_* + s) \;\ge\; \Delta - \varepsilon - m\,(\eta + s) = \frac{\Delta}{2} - \varepsilon - ms \qquad (|s| \le \eta)\,. \tag{A.392}$$

Insert this in Equation (A.390). The term $-ms$ integrates to zero against the even kernel over the symmetric set $|s| \le \eta$; outside it, $H \ge -\Delta$. Therefore

$$H_T(x_*) \;\ge\; \left(\frac{\Delta}{2} - \varepsilon\right)(1 - \kappa) - \Delta\kappa \;\ge\; \frac{\Delta}{2} - \varepsilon - \frac{3\Delta}{2}\kappa \;\ge\; \frac{\Delta}{2} - \varepsilon - \frac{12\,m}{\pi T}\,. \tag{A.393}$$

From Equations (A.390) and (A.392) (split the convolution at $|s| = \eta$, use the wedge bound inside and $H \ge -\Delta$ outside, then insert $\kappa \le 8m/(\pi T\Delta)$).

If instead $\sup_x (-H(x)) = \Delta$, pick b with $H(b) < -\Delta + \varepsilon$ and put $x_* = b - \eta$. For $y \le b$ the same mean value theorem gives $G(y) \ge G(b) - m(b - y)$ and $F(y) \le F(b)$, so $H(y) \le H(b) + m(b - y)$; for $|s| \le \eta$ the point $x_* + s$ lies at distance $\eta - s \in [0, 2\eta]$ to the left of b and $H(x_* + s) \le -\Delta/2 + \varepsilon - ms$. The same two steps give $H_T(x_*) \le -\Delta/2 + \varepsilon + 12m/(\pi T)$.

In both cases $|H_T(x_*)| \ge \Delta/2 - \varepsilon - 12m/(\pi T)$. Combining with Equation (A.391), letting $\varepsilon \downarrow 0$ and multiplying by 2 gives Equation (A.389). □

A.20.4 The characteristic function of the standardised sum

Lemma A.199 (The standardised third moment is at least one)**.** *Let $\langle Y\rangle = 0$, $\langle Y^2\rangle = 1$ and $\beta = \left\langle |Y|^3 \right\rangle < \infty$. Then $\beta \ge 1$.*

Rests on Lemma 11.63.

Proof. *Derives Lemma A.199.*

The discriminant argument of Lemma 11.63, applied to the nowhere-negative quadratic $s \mapsto \langle (sV + W)^2\rangle$ without centring — which is all that proof uses, as Proposition 11.24 already records — gives $\langle VW\rangle^2 \le \langle V^2\rangle\,\langle W^2\rangle$ for any V, W of finite second moment. Take $V = |Y|^{1/2}$ and $W = |Y|^{3/2}$: $1 = \langle Y^2\rangle^2 \le \langle |Y|\rangle \left\langle |Y|^3 \right\rangle = \langle |Y|\rangle\,\beta$. Take $V = |Y|$, $W = 1$: $\langle |Y|\rangle^2 \le \langle Y^2\rangle = 1$. Combining, $1 \le \beta$. □

Lemma A.200 (Transform estimate)**.** *Let $Y, Y_1, Y_2, \dots$ be independent and identically distributed with $\langle Y\rangle = 0$, $\langle Y^2\rangle = 1$ and $\beta = \left\langle |Y|^3 \right\rangle < \infty$, and let $Z_n = n^{-1/2}\sum_{i=1}^{n} Y_i$. Then*

$$\left|\psi_{Z_n}(t) - \mathrm{e}^{-t^2/2}\right| \;\le\; \frac{\beta\,|t|^3}{2\sqrt{n}}\,\mathrm{e}^{-t^2/4} \qquad \text{for } |t| \le T_n = \frac{\sqrt{n}}{2\beta}\,. \tag{A.394}$$

Rests on Equation (11.31), Lemma A.199, and Proposition 11.43.

Proof. *Derives Lemma A.200.*

Write ψ for the common characteristic function of the Y_i and $s = t/\sqrt{n}$, so that $\psi_{Z_n}(t) = \psi(s)^n$ by Proposition 11.43(ii) and (iii). Fix t with $|t| \le T_n$; then $\beta\,|s| \le \frac{1}{2}$, and since $\beta \ge 1$ by Lemma A.199 also $|s| \le \frac{1}{2}$.

Step 1: the quadratic expansion. Equation (11.31) at order two, with $x = sY$, expectations taken and $\langle Y\rangle = 0$, $\langle Y^2\rangle = 1$ used, gives

$$\psi(s) = 1 + \omega\,, \qquad \omega = -\frac{s^2}{2} + r\,, \qquad |r| \le \frac{\beta\,|s|^3}{6}\,. \tag{A.395}$$

Since $\beta\,|s| \le \frac{1}{2}$ one has $\beta\,|s|^3/6 \le s^2/12$, so

$$|\omega| \;\le\; \frac{s^2}{2} + \frac{s^2}{12} = \frac{7s^2}{12} \;\le\; \frac{7}{48} \;<\; \frac{1}{2}\,. \tag{A.396}$$

In particular $\psi(s) \neq 0$, and $\psi(s)^n = \mathrm{e}^{n\ln\psi(s)}$ with the principal logarithm — for an integer power the branch is immaterial.

Step 2: the logarithm. For $|w| \le \frac{1}{2}$ the principal logarithm satisfies

$$|\ln(1+w) - w| = \left|\sum_{j\ge 2} \frac{(-1)^{j-1}w^j}{j}\right| \;\le\; \frac{|w|^2}{2}\sum_{j\ge 0}|w|^j = \frac{|w|^2}{2\,(1-|w|)} \;\le\; |w|^2\,. \tag{A.397}$$

Now $n\omega + \frac{1}{2}t^2 = nr$, because $ns^2 = t^2$, and $|nr| \le \beta\,|t|^3/(6\sqrt{n})$ by Equation (A.395). Also, using Equation (A.396) and then $|t|/n \le 1/(2\beta\sqrt{n}) \le \beta/(2\sqrt{n})$ — the last step being $1/\beta \le \beta$, valid since $\beta \ge 1$ —

$$n\,|\omega|^2 \;\le\; \frac{49}{144}\,\frac{t^4}{n} = \frac{49}{144}\,|t|^3\,\frac{|t|}{n} \;\le\; \frac{49}{288}\,\frac{\beta\,|t|^3}{\sqrt{n}}\,.$$

Adding, and using $\frac{1}{6} + \frac{49}{288} = \frac{97}{288} < \frac{1}{2}$,

$$\left|n\ln\psi(s) + \frac{t^2}{2}\right| \;\le\; \left|n\omega + \frac{t^2}{2}\right| + n\,|\ln(1+\omega) - \omega| \;\le\; \frac{\beta\,|t|^3}{2\sqrt{n}}\,. \tag{A.398}$$

From Equations (A.395) to (A.397) (add the cubic remainder of the expansion to the quadratic error of the logarithm, using $\beta\,|t| \le \sqrt{n}/2$).

Step 3: exponentiate. On $|t| \le T_n$,

$$\frac{\beta\,|t|^3}{2\sqrt{n}} = \frac{t^2}{2}\cdot\frac{\beta\,|t|}{\sqrt{n}} \;\le\; \frac{t^2}{2}\cdot\frac{1}{2} = \frac{t^2}{4}\,, \tag{A.399}$$

so by Equation (A.398) the real part of $n\ln\psi(s)$ is at most $-t^2/2 + t^2/4 = -t^2/4$, and the same bound holds trivially for $\Re\,(-t^2/2)$. For complex a, b,

$$\left|\mathrm{e}^{a} - \mathrm{e}^{b}\right| = \left|(a-b)\int_0^1 \mathrm{e}^{b+u(a-b)}\,\mathrm{d}u\right| \;\le\; |a-b|\;\mathrm{e}^{\max(\Re a,\,\Re b)}\,,$$

since $\Re\,(b + u(a-b)) = (1-u)\Re b + u\Re a$ is a convex combination. Taking $a = n\ln\psi(s)$ and $b = -t^2/2$ and inserting Equation (A.398) gives Equation (A.394). □

A.20.5 Integration, and the constant

Proof of Theorem A.192. *Derives Theorem A.192.*

Put $Y_i = (X_i - \mu)/\sigma$, so the Y_i are independent and identically distributed with mean 0, variance 1 and $\beta = \left\langle |Y|^3 \right\rangle = \rho/\sigma^3$, and $Z_n = n^{-1/2}\sum_i Y_i$ is the statistic of the theorem. Apply Theorem A.198 with $F = F_n$, $G = \Phi$ and $T = T_n = \sqrt{n}/(2\beta)$. The comparison function qualifies: Φ is non-decreasing and differentiable with $\Phi' = \varphi \le 1/\sqrt{2\pi}$, so $m = 1/\sqrt{2\pi}$; and both distributions have finite first moments, namely zero. Then, by Lemma A.200,

$$\Delta_n \;\le\; \frac{1}{\pi}\int_{-T_n}^{T_n} \frac{\beta t^2}{2\sqrt{n}}\, \mathrm{e}^{-t^2/4}\,\mathrm{d}t + \frac{24}{\pi\sqrt{2\pi}}\cdot\frac{2\beta}{\sqrt{n}}\,. \tag{A.400}$$

From Equations (A.389) and (A.394) (insert the transform estimate into the smoothing inequality at $T = T_n = \sqrt{n}/(2\beta)$).

Extend the first integral to all of $\mathbb{R}$, which only increases it, and evaluate it by the substitution $t = 2u$ together with the Gaussian moment $\int u^2\mathrm{e}^{-u^2}\mathrm{d}u = \sqrt{\pi}/2$ (Proposition 11.29, rescaled):

$$\int_{-\infty}^{\infty} t^2\mathrm{e}^{-t^2/4}\,\mathrm{d}t = 8\int_{-\infty}^{\infty} u^2\mathrm{e}^{-u^2}\,\mathrm{d}u = 4\sqrt{\pi}\,.$$

The first term of Equation (A.400) is therefore at most $(4\sqrt{\pi}/\pi)\,\beta/(2\sqrt{n}) = 2\beta/\sqrt{\pi n}$, and

$$\Delta_n \;\le\; \left(\frac{2}{\sqrt{\pi}} + \frac{48}{\pi\sqrt{2\pi}}\right)\frac{\beta}{\sqrt{n}} \;\le\; \frac{7.23\,\beta}{\sqrt{n}} = \frac{7.23\,\rho}{\sigma^3\sqrt{n}}\,, \tag{A.401}$$

since $2/\sqrt{\pi} = 1.1284$ and $48/\left(\pi\sqrt{2\pi}\right) = 6.0954$. That is Equation (A.382). Note that no step required n to be large: the inequality holds for every $n \ge 1$, being vacuous — because $\Delta_n \le 1$ always — until n exceeds about $52\beta^2$. □

Remark A.201 (On the constant). The value $C = 7.23$ derived here is admissible and is not sharp, and it is worth saying where the loss is. Almost all of it is in the smoothing inequality: of the two terms in Equation (A.401), the truncation penalty $24m/(\pi T)$ contributes 6.10 and the transform integral only 1.13. The factor 24 in Equation (A.389) comes from the crude estimate $1 - \cos \le 2$ used for the tail of the Fejér kernel and from the wedge construction's factor $\frac{3}{2}$; sharpening either changes the arithmetic but not the structure. The interval cannot simply be lengthened, because Equation (A.394) is proved only for $|t| \le \sqrt{n}/(2\beta)$ — beyond that the logarithm of $\psi(t/\sqrt{n})$ need not even be defined — so $T = T_n$ is the best this argument allows.

Determining the smallest admissible C is a separate literature and is *not* attempted here. Berry's paper [856] establishes finiteness, which is the qualitative content and all that Remark 11.49 needs; Esseen obtained the same result independently in the same period; and the sharpest published value for identically distributed summands is $C \le 0.4690$ [2238]. Since $\rho/\sigma^3 \ge 1$ always (Lemma A.199), no value of C below that can make the bound say anything at small n, which is the practical point Remark 11.49 draws.

Remark A.202. Appendix A.20 discharges the proof obligation of Equation (11.40), stated in Remark 11.49 of Chapter 11 as the rate that turns the central limit theorem into a usable statement about a finite sample. The consequences drawn there stand unchanged: the bound is uniform in z, so its *relative* accuracy is worst in the tails, exactly where a coverage factor $k = 3$ operates; and the $n^{-1/2}$ decay is slow enough that several hundred contributions are needed before the inequality itself certifies a few per cent. What this appendix adds is that both statements are now consequences of something proved, with an explicit constant, rather than of something quoted.

A.21 Wilks' Theorem in the Multi-Parameter Case

This appendix completes the derivation of Theorem 11.88 of Chapter 11 in k parameters with r constraints. The chapter's argument is correct at the level of the quadratic approximation and two of its steps are asserted rather than proved:

(i) the remainder in the k-dimensional Taylor expansion of the log-likelihood must be shown *uniform* on a neighbourhood of the true parameter that shrinks at rate $n^{-1/2}$, since the expansion is evaluated not at a fixed point but at two estimators that move with the data;

(ii) both estimators — the unconstrained one and the one constrained to Θ_0 — must be shown to lie in that neighbourhood with probability tending to one, since that is what licenses replacing the constraint surface Θ_0 by its tangent space at θ_0, the step written "$\Theta_0 \simeq \theta_0 + E$" at Equation (11.85).

Both are proved below, from exactly the regularity conditions (R1)–(R5) of Definition 11.59 and nothing stronger, and the chi-squared limit with r degrees of freedom is then assembled from them. The one-parameter case with a simple null needs none of this — there Θ_0 is a point, there is no surface to replace — which is why Chapter 11 carries that derivation in full and defers only this one.

A.21.1 Setting and notation

Let $X_1, \dots, X_n$ be independent and identically distributed with density $f(\,\cdot\,;\theta_0)$, let $\Theta \subseteq \mathbb{R}^k$ and let (R1)–(R5) hold at θ_0, read in k parameters: (R4) asks that $\theta \mapsto f(x;\theta)$ be three times continuously differentiable near θ_0 with the first two derivatives available under the integral, and (R5) that the information matrix $I = I_1(\theta_0)$ of Definition 11.60 be finite and positive definite and that a single integrable $M(x)$ dominate *every* third partial derivative $\partial^3 \ln f(x;\theta)/\partial\theta^a\partial\theta^b\partial\theta^c$ throughout a ball $\bar{B}(\theta_0, a_0) \subseteq \Theta$. Write $\lambda > 0$ for the smallest eigenvalue of I and $\|A\| = \sup_{\|\boldsymbol{v}\|=1} \|A\boldsymbol{v}\|$ for the operator norm of a matrix.

Put $\ell_n(\theta) = \sum_{i=1}^n \ln f(X_i;\theta)$ and

$$\boldsymbol{Z}_n = \frac{1}{\sqrt{n}}\,\nabla\ell_n(\theta_0)\,, \qquad J_n = \frac{1}{n}\,\nabla^2\ell_n(\theta_0)\,, \qquad \bar{M}_n = \frac{1}{n}\sum_{i=1}^n M(X_i)\,. \tag{A.402}$$

All three are averages of independent identically distributed terms: $\boldsymbol{Z}_n$ of the scores, whose common mean is zero and whose covariance is I (Lemma 11.61); J_n of the observation Hessians, whose common mean is $-I$ by Equation (11.48); and $\bar{M}_n$ of M, whose common mean $\bar{M} = \langle M \rangle$ is finite by (R5). The statistic is a pure number, the SI dimensions of θ cancelling in the likelihood ratio of Equation (11.80), and so is every quantity above: $\nabla\ell_n$ carries the reciprocal of the unit of θ and I its square, so the combinations formed below — $\boldsymbol{u}^{\mathsf{T}} I \boldsymbol{u}$ with $\boldsymbol{u} = \sqrt{n}(\theta - \theta_0)$, and $\boldsymbol{Z}_n^{\mathsf{T}} I^{-1} \boldsymbol{Z}_n$ — are dimensionless.

The two estimators are the local ones, defined exactly as Chapter 11 defines its root of the score equation. Fix $a_1 \in (0, a_0]$, to be chosen in Lemma A.211, and let

$$\hat{\theta}_n \in \arg\max_{\theta \in \bar{B}(\theta_0, a_1)} \ell_n(\theta)\,, \qquad \hat{\theta}_{0,n} \in \arg\max_{\theta \in \Theta_0 \cap \bar{B}(\theta_0, a_1)} \ell_n(\theta)\,. \tag{A.403}$$

Both maxima are attained: ℓ_n is continuous and both sets are closed and bounded, hence compact, so the extreme value theorem (Theorem 7.24) applies. Both sets contain θ_0, so

$$\ell_n(\hat{\theta}_n) \geq \ell_n(\theta_0) \qquad \text{and} \qquad \ell_n(\hat{\theta}_{0,n}) \geq \ell_n(\theta_0)\,, \tag{A.404}$$

the second because $\theta_0 \in \Theta_0$ by hypothesis. The statistic under study is $q = -2\left[\ell_n(\hat{\theta}_{0,n}) - \ell_n(\hat{\theta}_n)\right] \geq 0$.

The null set is a C^2 submanifold near θ_0, which is used in exactly one form and is therefore stated in that form.

Definition A.203 (The constraint surface)**.** Θ_0 is a C^2 *submanifold of dimension* $k-r$ *through* θ_0 *with tangent space* E when there are a linear subspace $E \subseteq \mathbb{R}^k$ with $\dim E = k-r$, a neighbourhood W of θ_0, an open $V \ni \mathbf{0}$ in E, and a twice continuously differentiable map $h : V \to E^\perp$ with $h(\mathbf{0}) = \mathbf{0}$ and $Dh(\mathbf{0}) = 0$, such that

$$\Theta_0 \cap W = \{\theta_0 + \boldsymbol{e} + h(\boldsymbol{e}) \mid \boldsymbol{e} \in V\} \,. \tag{A.405}$$

Rests on Definition 11.87.

Remark A.204 (Why the graph form, and not the constraint form). Theorem 11.88 states the hypothesis as "H_0 imposes r smooth, functionally independent constraints", i.e. Θ_0 is the level set $\{g(\theta) = \mathbf{0}\}$ of a C^2 map $g : \Theta \to \mathbb{R}^r$ with $Dg(\theta_0)$ of rank r. The equivalence of that description with Equation (A.405) is the implicit function theorem in several variables, which Chapter 13 records as owed rather than proved. This appendix therefore takes Definition A.203 as the hypothesis: it is the form the proof uses, it is one of the standard equivalent definitions of a submanifold, and adopting it keeps the argument free of a theorem the treatise has not yet built. A reader who has the implicit function theorem may read the two hypotheses as the same one.

A.21.2 What is quoted

Theorem A.205 (Weak law under integrability alone)**.** *Let* $W_1, W_2, \ldots$ *be independent and identically distributed with* $\langle |W| \rangle < \infty$*. Then* $n^{-1} \sum_{i=1}^n W_i \to \langle W \rangle$ *in probability [1894].*

Rests on Corollary 11.51.

Theorem A.206 (Cramér–Wold, and continuous mapping)**.** *Let* $\boldsymbol{W}_n, \boldsymbol{W}$ *be random vectors in* $\mathbb{R}^k$*. Then* $\boldsymbol{W}_n \to \boldsymbol{W}$ *in distribution if and only if* $\boldsymbol{a}^\mathsf{T}\boldsymbol{W}_n \to \boldsymbol{a}^\mathsf{T}\boldsymbol{W}$ *in distribution for every* $\boldsymbol{a} \in \mathbb{R}^k$*. Moreover, if* $\boldsymbol{W}_n \to \boldsymbol{W}$ *in distribution and* $g : \mathbb{R}^k \to \mathbb{R}$ *is continuous, then* $g(\boldsymbol{W}_n) \to g(\boldsymbol{W})$ *in distribution [1894].*

Rests on Definition 11.21 and Lemma 11.52.

Remark A.207 (What is quoted here). Three imports, all of them already relied on by Chapter 11 itself, and no more. Theorem A.205 is the weak law in the form that assumes only integrability; Corollary 11.51 is the Chebyshev form and needs a finite variance, which (R5) does not supply for the Hessian summands or for M — the chapter names this gap explicitly in the derivation of Theorem 11.67 and quotes the same source. Theorem A.206 is what promotes the scalar central limit theorem Theorem 11.46 to the vector statement $\boldsymbol{Z}_n \to \mathcal{N}(0, I)$ and lets a continuous function be applied under a limit in distribution; the chapter uses both silently in the k-parameter clause of Theorem 11.67 and proves the one-dimensional special case of the second at Equation (11.44). The third import is the implicit function theorem, and it is *avoided* rather than used, by taking Definition A.203 as the hypothesis (Remark A.204). Everything else below — the uniform remainder, the localisation of both estimators, the tangent-space replacement and the degrees of freedom — is derived here.

Lemma A.208 (The three averages)**.** $\|J_n + I\| \to 0$ *and* $\bar{M}_n \to \bar{M}$ *in probability, and* $\boldsymbol{Z}_n \to \mathcal{N}(0, I)$ *in distribution. In particular each component of* $\boldsymbol{Z}_n$*, and hence* $\|\boldsymbol{Z}_n\|$*, is bounded in probability.*

Rests on Lemma 11.61 and Theorems A.205 and A.206.

Proof. *Derives Lemma A.208.*

Each entry of J_n is the average of n independent copies of $\partial^2 \ln f(X;\theta_0)/\partial\theta^a\partial\theta^b$, integrable by (R4) with mean $-I_{ab}$ by Equation (11.48), so Theorem A.205 gives $J_n \to -I$

entrywise in probability; a matrix norm is a continuous function of finitely many entries, so $\|J_n + I\| \to 0$ in probability. The same theorem applied to M, integrable by (R5), gives the second claim.

For the third, fix $\boldsymbol{a} \in \mathbb{R}^k$, $\boldsymbol{a} \neq \mathbf{0}$. Then $\boldsymbol{a}^\mathsf{T} \boldsymbol{Z}_n = n^{-1/2} \sum_i \boldsymbol{a}^\mathsf{T} \boldsymbol{U}(\theta_0; X_i)$ with $\boldsymbol{U} = \nabla \ln f$ the score, a normalised sum of independent identically distributed variables with mean zero (Lemma 11.61) and variance $\boldsymbol{a}^\mathsf{T} I \boldsymbol{a}$, finite and strictly positive by (R5). Theorem 11.46 gives $\boldsymbol{a}^\mathsf{T} \boldsymbol{Z}_n \to \mathcal{N}(0, \boldsymbol{a}^\mathsf{T} I \boldsymbol{a})$, and Theorem A.206 converts this into $\boldsymbol{Z}_n \to \mathcal{N}(0, I)$. Boundedness in probability of each component is Equation (11.43), and $\|\boldsymbol{Z}_n\| \leq \sum_a |Z_n^a|$ is a finite sum of variables each bounded in probability. □

A.21.3 The uniform quadratic expansion

This is item (i). The content is not that a remainder exists — Taylor gives that at every fixed θ — but that a *single* random variable, not depending on θ, bounds it throughout the ball.

Lemma A.209 (Uniform third-order remainder)**.** *For every $\theta \in \bar{B}(\theta_0, a_0)$, writing $\delta = \theta - \theta_0$,*

$$\ell_n(\theta) = \ell_n(\theta_0) + \sqrt{n}\, \boldsymbol{Z}_n^\mathsf{T} \delta + \frac{n}{2}\, \delta^\mathsf{T} J_n \delta + R_n(\theta)\,, \qquad |R_n(\theta)| \;\leq\; \frac{n\, k^3}{6}\, \bar{M}_n\, \|\delta\|^3\,. \tag{A.406}$$

The bound is uniform in θ: apart from the explicit factor $\|\delta\|^3$ it involves only $\bar{M}_n$, which does not depend on θ.

Rests on Definition 11.59 and Theorem 7.38.

Proof. *Derives Lemma A.209.*

Fix θ and set $g(u) = \ell_n(\theta_0 + u\delta)$ for $u \in [0,1]$; the segment lies in $\bar{B}(\theta_0, a_0)$ by convexity of the ball, so g is three times continuously differentiable by (R4). Taylor's theorem with Lagrange remainder (Theorem 7.38) gives $g(1) = g(0) + g'(0) + \tfrac{1}{2} g''(0) + \tfrac{1}{6} g'''(\xi)$ for some $\xi \in (0,1)$. By the chain rule (Proposition 7.72),

$$g'(0) = \delta^\mathsf{T} \nabla \ell_n(\theta_0) = \sqrt{n}\, \boldsymbol{Z}_n^\mathsf{T} \delta\,, \qquad g''(0) = \delta^\mathsf{T} \nabla^2 \ell_n(\theta_0)\, \delta = n\, \delta^\mathsf{T} J_n \delta\,,$$

which are the first two displayed terms, and

$$g'''(\xi) = \sum_{a,b,c=1}^{k} \delta^a \delta^b \delta^c\, \frac{\partial^3 \ell_n}{\partial\theta^a \partial\theta^b \partial\theta^c}\, (\theta_0 + \xi\delta)\,. \tag{A.407}$$

Each third partial derivative of ℓ_n is a sum of n terms, one per observation, and (R5) bounds each of them by $M(X_i)$ throughout $\bar{B}(\theta_0, a_0)$ — in particular at the unknown intermediate point $\theta_0 + \xi\delta$, which is what makes the estimate independent of ξ and hence of θ. Since $|\delta^a| \leq \|\delta\|$ for each component and the sum Equation (A.407) has k^3 terms,

$$|g'''(\xi)| \;\leq\; k^3\, \|\delta\|^3 \sum_{i=1}^{n} M(X_i) = n\, k^3\, \bar{M}_n\, \|\delta\|^3\,,$$

and $R_n(\theta) = \tfrac{1}{6} g'''(\xi)$. □

Corollary A.210 (Uniform quadratic approximation on the $n^{-1/2}$ scale)**.** *Define, for $\boldsymbol{u} \in \mathbb{R}^k$,*

$$\Lambda_n(\boldsymbol{u}) = -2\, \boldsymbol{Z}_n^\mathsf{T} \boldsymbol{u} + \boldsymbol{u}^\mathsf{T} I\, \boldsymbol{u}\,. \tag{A.408}$$

Then for every $K > 0$ and every n with $K/\sqrt{n} \leq a_0$,

$$\sup_{\|\boldsymbol{u}\| \leq K} \left| -2 \left[\ell_n\!\left(\theta_0 + \frac{\boldsymbol{u}}{\sqrt{n}}\right) - \ell_n(\theta_0) \right] - \Lambda_n(\boldsymbol{u}) \right| \;\leq\; \epsilon_n(K)\,, \tag{A.409}$$

where $\epsilon_n(K) = K^2\, \|J_n + I\| + \tfrac{1}{3} k^3 K^3 \bar{M}_n / \sqrt{n}$ tends to 0 in probability.

Rests on Lemmas A.208 and A.209.

Proof. *Derives Corollary A.210.*

Put $\delta = \boldsymbol{u}/\sqrt{n}$ in Equation (A.406) and multiply by -2:

$$-2\left[\ell_n(\theta) - \ell_n(\theta_0)\right] = -2\boldsymbol{Z}_n^{\mathsf{T}}\boldsymbol{u} - \boldsymbol{u}^{\mathsf{T}} J_n \boldsymbol{u} - 2R_n(\theta)\,,$$

because $n\delta^{\mathsf{T}} J_n \delta = \boldsymbol{u}^{\mathsf{T}} J_n \boldsymbol{u}$ and $\sqrt{n}\boldsymbol{Z}_n^{\mathsf{T}}\delta = \boldsymbol{Z}_n^{\mathsf{T}}\boldsymbol{u}$. Subtracting Equation (A.408) leaves $-\boldsymbol{u}^{\mathsf{T}}(J_n + I)\,\boldsymbol{u} - 2R_n$, whose first part is at most $\|J_n + I\|\,\|\boldsymbol{u}\|^2$ in modulus and whose second is at most $\frac{1}{3}k^3\bar{M}_n\,\|\boldsymbol{u}\|^3/\sqrt{n}$ by Equation (A.406) with $\|\delta\| = \|\boldsymbol{u}\|/\sqrt{n}$. Both are increasing in $\|\boldsymbol{u}\|$, so the supremum over the ball is attained at the bound. That $\epsilon_n(K) \to 0$ in probability is Lemma A.208: the first term is K^2 times something tending to 0, and the second is $\bar{M}_n$, bounded in probability, times $\frac{1}{3}k^3K^3n^{-1/2} \to 0$. □

A.21.4 Localisation of both estimators

This is item (ii). Note that it is proved for *any* closed subset of the ball containing θ_0, so that the unconstrained and the constrained estimators are covered by one argument; nothing about the geometry of Θ_0 is used here.

Lemma A.211 (Root-n localisation)**.** *Choose $a_1 \in (0, a_0]$ with $\frac{1}{6}k^3\left(\bar{M}+1\right)a_1 \le \lambda/8$, and let $A \subseteq \bar{B}(\theta_0, a_1)$ be any closed set with $\theta_0 \in A$. Let $\hat{\theta}$ maximise ℓ_n over A. Then*

$$\Pr\!\left(\sqrt{n}\,\left\|\hat{\theta} - \theta_0\right\| \;\le\; \frac{4}{\lambda}\,\|\boldsymbol{Z}_n\|\right) \;\longrightarrow\; 1\,, \tag{A.410}$$

and consequently $\sqrt{n}\left\|\hat{\theta} - \theta_0\right\|$ is bounded in probability and $\hat{\theta} \to \theta_0$ in probability. In particular both estimators Equation (A.403) satisfy this.

Rests on Lemmas A.208 and A.209.

Proof. *Derives Lemma A.211.*

Write $\hat{\delta} = \hat{\theta} - \theta_0$ and $d = \left\|\hat{\delta}\right\| \le a_1$. Since $\theta_0 \in A$ we have $\ell_n(\hat{\theta}) \ge \ell_n(\theta_0)$, so Equation (A.406) gives

$$0 \;\le\; \sqrt{n}\,\boldsymbol{Z}_n^{\mathsf{T}}\hat{\delta} + \frac{n}{2}\,\hat{\delta}^{\mathsf{T}} J_n \hat{\delta} + R_n(\hat{\theta})\,. \tag{A.411}$$

Bound the three terms. The first is at most $\sqrt{n}\,\|\boldsymbol{Z}_n\|\,d$ by the Cauchy–Schwarz inequality for vectors. For the second, write $\hat{\delta}^{\mathsf{T}} J_n \hat{\delta} = -\hat{\delta}^{\mathsf{T}} I \hat{\delta} + \hat{\delta}^{\mathsf{T}}(J_n + I)\,\hat{\delta}$; the first piece is at most $-\lambda d^2$, since λ is the smallest eigenvalue of the positive-definite I, and the second is at most $\|J_n + I\|\,d^2$. The third is at most $\frac{1}{6}nk^3\bar{M}_n d^3$. Hence

$$\frac{n\lambda}{2}d^2 \;\le\; \sqrt{n}\,\|\boldsymbol{Z}_n\|\,d + \frac{n}{2}\,\|J_n + I\|\,d^2 + \frac{n\,k^3}{6}\bar{M}_n d^3\,. \tag{A.412}$$

From Equations (A.406) and (A.411) (bound the linear term by Cauchy–Schwarz, split the Hessian term into $-I$ plus its deviation, and insert the uniform remainder bound).

Let E_n be the event on which both $\|J_n + I\| \le \lambda/4$ and $\bar{M}_n \le \bar{M} + 1$; by Lemma A.208, $\Pr(E_n) \to 1$. On E_n, and using $d \le a_1$ with the choice of a_1,

$$\frac{n}{2}\,\|J_n + I\|\,d^2 \le \frac{n\lambda}{8}d^2\,, \qquad \frac{n\,k^3}{6}\bar{M}_n d^3 \le \left[\frac{k^3\left(\bar{M}+1\right)a_1}{6}\right] n\,d^2 \le \frac{n\lambda}{8}d^2\,.$$

Substituting in Equation (A.412) and cancelling $n\lambda d^2/4$ from both sides leaves $\frac{1}{4}n\lambda d^2 \le \sqrt{n}\,\|\boldsymbol{Z}_n\|\,d$; dividing by $\frac{1}{4}n\lambda d$ when $d > 0$ — and noting the conclusion is trivial when $d = 0$ — gives $\sqrt{n}\,d \le 4\,\|\boldsymbol{Z}_n\|/\lambda$, which is Equation (A.410). Boundedness in probability follows because $\|\boldsymbol{Z}_n\|$ is bounded in probability (Lemma A.208), and consistency because $\sqrt{n}d = O_P(1)$ forces $d \to 0$ in probability. □

A.21.5 Replacing the surface by its tangent space

Lemma A.212 (Second-order flatness)**.** *Under Definition A.203 there are $c_0 < \infty$ and $a_2 \in (0, a_1]$ such that*

1. *if $\theta \in \Theta_0$ and $\|\theta - \theta_0\| \le a_2$, then the orthogonal projection $\boldsymbol{w}$ of $\theta - \theta_0$ onto E satisfies $\|(\theta - \theta_0) - \boldsymbol{w}\| \le c_0 \|\theta - \theta_0\|^2$;*
2. *if $\boldsymbol{e} \in E$ and $\|\boldsymbol{e}\| \le a_2$, there is $\theta \in \Theta_0$ with $\|(\theta - \theta_0) - \boldsymbol{e}\| \le c_0 \|\boldsymbol{e}\|^2$.*

Rests on Definition A.203 and Theorem 7.38.

Proof. *Derives Lemma A.212.*

Choose $a_2 \in (0, a_1]$ small enough that the closed ball $\bar{B}(\theta_0, 2a_2)$ lies in W and that $\{\boldsymbol{e} \in E \mid \|\boldsymbol{e}\| \le 2a_2\} \subseteq V$; let $c_0 = \frac{1}{2} \sup \|D^2 h\|$ over that compact set, finite because h is twice *continuously* differentiable and the supremum of a continuous function on a compact set is attained (Theorem 7.24). Taylor's theorem applied to $u \mapsto h(u\boldsymbol{e})$ on $[0,1]$, with $h(\boldsymbol{0}) = \boldsymbol{0}$ and $Dh(\boldsymbol{0}) = 0$ killing the first two terms, gives

$$\|h(\boldsymbol{e})\| \;\le\; c_0 \|\boldsymbol{e}\|^2 \qquad \text{for } \|\boldsymbol{e}\| \le 2a_2 \,. \tag{A.413}$$

(ii) Take $\theta = \theta_0 + \boldsymbol{e} + h(\boldsymbol{e})$, which lies in Θ_0 by Equation (A.405); then $(\theta - \theta_0) - \boldsymbol{e} = h(\boldsymbol{e})$ and Equation (A.413) applies.

(i) Such a θ lies in W, so $\theta - \theta_0 = \boldsymbol{e} + h(\boldsymbol{e})$ for some $\boldsymbol{e} \in V$; the two summands are orthogonal, so $\boldsymbol{w} = \boldsymbol{e}$ and $\|\boldsymbol{e}\| \le \|\theta - \theta_0\| \le a_2$. Then $(\theta - \theta_0) - \boldsymbol{w} = h(\boldsymbol{e})$, and Equation (A.413) together with $\|\boldsymbol{e}\| \le \|\theta - \theta_0\|$ gives the bound. □

Lemma A.213 (The three minima)**.** *Write $\hat{\boldsymbol{u}} = \sqrt{n}(\hat{\theta}_n - \theta_0)$ and $\hat{\boldsymbol{u}}_0 = \sqrt{n}(\hat{\theta}_{0,n} - \theta_0)$. Then*

$$\Lambda_n(\hat{\boldsymbol{u}}) - \min_{\boldsymbol{u} \in \mathbb{R}^k} \Lambda_n(\boldsymbol{u}) \;\longrightarrow\; 0 \,, \qquad \Lambda_n(\hat{\boldsymbol{u}}_0) - \min_{\boldsymbol{e} \in E} \Lambda_n(\boldsymbol{e}) \;\longrightarrow\; 0 \,, \tag{A.414}$$

both in probability.

Rests on Corollary A.210 and Lemmas A.211 and A.212.

Proof. *Derives Lemma A.213.*

Since I is positive definite, Λ_n is a strictly convex quadratic; completing the square with $\boldsymbol{m} = I^{-1}\boldsymbol{Z}_n$,

$$\Lambda_n(\boldsymbol{u}) = (\boldsymbol{u} - \boldsymbol{m})^{\mathsf{T}} I \,(\boldsymbol{u} - \boldsymbol{m}) - \boldsymbol{Z}_n^{\mathsf{T}} I^{-1} \boldsymbol{Z}_n \,, \tag{A.415}$$

so the unconstrained minimum is attained at $\boldsymbol{m}$ and the minimum over the subspace E at the I-orthogonal projection $\boldsymbol{p}$ of $\boldsymbol{m}$ onto E. Both $\|\boldsymbol{m}\|$ and $\|\boldsymbol{p}\|$ are bounded in probability, being bounded linear images of $\boldsymbol{Z}_n$. Also Λ_n is Lipschitz on each ball: for $\|\boldsymbol{u}\|, \|\boldsymbol{u}'\| \le K$,

$$|\Lambda_n(\boldsymbol{u}) - \Lambda_n(\boldsymbol{u}')| \;\le\; L_n(K) \|\boldsymbol{u} - \boldsymbol{u}'\| \,, \qquad L_n(K) = 2\|\boldsymbol{Z}_n\| + 2K\|I\| \,, \tag{A.416}$$

which is bounded in probability for fixed K.

Fix $\varepsilon > 0$. By Lemma A.211 and the boundedness just noted, there is K such that with probability at least $1 - \varepsilon$ for all large n the four vectors $\hat{\boldsymbol{u}}, \hat{\boldsymbol{u}}_0, \boldsymbol{m}, \boldsymbol{p}$ all lie in $\bar{B}(\boldsymbol{0}, K)$; call that event A_n and intersect it with the event that $\epsilon_n(K+1) \le \varepsilon$, which also has probability approaching 1 by Corollary A.210. Work on the intersection, and abbreviate $D_n(\theta) = -2\,[\ell_n(\theta) - \ell_n(\theta_0)]$, so that Equation (A.409) reads $|D_n(\theta_0 + \boldsymbol{u}/\sqrt{n}) - \Lambda_n(\boldsymbol{u})| \le \varepsilon$ for every $\|\boldsymbol{u}\| \le K + 1$. The extra unit of radius is there because one auxiliary point below, the image in Θ_0 of the tangential minimiser, misses $\bar{B}(\boldsymbol{0}, K)$ by $O(n^{-1/2})$; all other points used lie in $\bar{B}(\boldsymbol{0}, K)$. Write $L_n = L_n(K+1)$ for the Lipschitz constant Equation (A.416) at that radius.

First limit. That $\Lambda_n(\hat{\boldsymbol{u}}) \geq \min_{\mathbb{R}^k} \Lambda_n$ is immediate. For the reverse, put $\theta_{\mathrm{m}} = \theta_0 + \boldsymbol{m}/\sqrt{n}$, which lies in $\bar{B}(\theta_0, a_1)$ once $K/\sqrt{n} \leq a_1$, so $\ell_n(\hat{\theta}_n) \geq \ell_n(\theta_{\mathrm{m}})$ and hence $D_n(\hat{\theta}_n) \leq D_n(\theta_{\mathrm{m}})$. Therefore

$$\Lambda_n(\hat{\boldsymbol{u}}) - \varepsilon \;\leq\; D_n(\hat{\theta}_n) \;\leq\; D_n(\theta_{\mathrm{m}}) \;\leq\; \Lambda_n(\boldsymbol{m}) + \varepsilon = \min_{\boldsymbol{u} \in \mathbb{R}^k} \Lambda_n(\boldsymbol{u}) + \varepsilon\,,$$

so the first difference in Equation (A.414) is at most 2ε on an event of probability approaching $1 - \varepsilon$.

Second limit. For the lower bound let $\boldsymbol{w}$ be the orthogonal projection of $\hat{\boldsymbol{u}}_0$ onto E. By Lemma A.212(i) applied to $\theta = \hat{\theta}_{0,n} \in \Theta_0$, whose distance to θ_0 is at most $K/\sqrt{n} \leq a_2$ for large n,

$$\|\hat{\boldsymbol{u}}_0 - \boldsymbol{w}\| = \sqrt{n}\,\left\|\left(\hat{\theta}_{0,n} - \theta_0\right) - \frac{\boldsymbol{w}}{\sqrt{n}}\right\| \;\leq\; \sqrt{n}\,c_0\left(\frac{K}{\sqrt{n}}\right)^2 = \frac{c_0 K^2}{\sqrt{n}}\,, \tag{A.417}$$

and $\|\boldsymbol{w}\| \leq \|\hat{\boldsymbol{u}}_0\| \leq K$, so Equation (A.416) gives $\Lambda_n(\hat{\boldsymbol{u}}_0) \geq \Lambda_n(\boldsymbol{w}) - L_n c_0 K^2/\sqrt{n} \geq \min_E \Lambda_n - L_n c_0 K^2/\sqrt{n}$.

For the upper bound apply Lemma A.212(ii) to $\boldsymbol{p}/\sqrt{n}$, of norm at most $K/\sqrt{n} \leq a_2$: there is $\theta_{\mathrm{p}} \in \Theta_0$ with $\|(\theta_{\mathrm{p}} - \theta_0) - \boldsymbol{p}/\sqrt{n}\| \leq c_0 K^2/n$, hence $\|\sqrt{n}\,(\theta_{\mathrm{p}} - \theta_0) - \boldsymbol{p}\| \leq c_0 K^2/\sqrt{n}$ and $\theta_{\mathrm{p}} \in \Theta_0 \cap \bar{B}(\theta_0, a_1)$ for large n. Since $\hat{\theta}_{0,n}$ maximises ℓ_n over that set, $D_n(\hat{\theta}_{0,n}) \leq D_n(\theta_{\mathrm{p}})$, and running the same three-step chain as before, then moving from $\sqrt{n}(\theta_{\mathrm{p}} - \theta_0)$ to $\boldsymbol{p}$ by Equation (A.416),

$$\Lambda_n(\hat{\boldsymbol{u}}_0) \;\leq\; \Lambda_n(\boldsymbol{p}) + 2\varepsilon + \frac{L_n c_0 K^2}{\sqrt{n}} = \min_{\boldsymbol{e} \in E} \Lambda_n(\boldsymbol{e}) + 2\varepsilon + \frac{L_n c_0 K^2}{\sqrt{n}}\,.$$

Both correction terms tend to 0 in probability, L_n being bounded in probability, and $\varepsilon > 0$ was arbitrary. □

A.21.6 The chi-squared limit

Theorem A.214 (Wilks, k parameters and r constraints)**.** *Let (R1)–(R5) hold at θ_0 in k parameters, let the sample be independent and identically distributed of size n, and let $\Theta_0 \ni \theta_0$ satisfy Definition A.203 with $\dim E = k - r$. Then the local likelihood-ratio statistic $q = -2\left[\ell_n(\hat{\theta}_{0,n}) - \ell_n(\hat{\theta}_n)\right]$ built from the estimators Equation (A.403) converges in distribution to χ^2_r.*

Rests on Lemmas 11.83 and A.213 and Theorem A.206.

Proof. *Derives Theorem A.214.*

Write D_n as in Lemma A.213, so that $q = D_n(\hat{\theta}_{0,n}) - D_n(\hat{\theta}_n)$ — the common reference value $\ell_n(\theta_0)$ cancels. By Corollary A.210 each D_n equals the corresponding Λ_n up to $\epsilon_n(K)$, and by Lemma A.213 each Λ_n equals the corresponding minimum up to a quantity tending to zero in probability. Hence

$$q = \min_{\boldsymbol{e} \in E} \Lambda_n(\boldsymbol{e}) - \min_{\boldsymbol{u} \in \mathbb{R}^k} \Lambda_n(\boldsymbol{u}) + o_P(1) = \min_{\boldsymbol{e} \in E} (\boldsymbol{e} - \boldsymbol{m})^{\mathsf{T}}\, I\, (\boldsymbol{e} - \boldsymbol{m}) + o_P(1)\,, \tag{A.418}$$

From Equations (A.414) and (A.415) (replace each statistic by the minimum of the quadratic form and cancel the common constant).

The second equality is Equation (A.415), whose additive constant $-\boldsymbol{Z}_n^{\mathsf{T}} I^{-1} \boldsymbol{Z}_n$ is the same in both minima and cancels.

Now substitute $\boldsymbol{V}_n = I^{1/2}\boldsymbol{m} = I^{-1/2}\boldsymbol{Z}_n$ and $S = I^{1/2}E$, where $I^{1/2}$ is any invertible matrix with $I^{1/2}\left(I^{1/2}\right)^{\mathsf{T}} = I$, symmetric here; the existence of one is the weak form of the

spectral theorem recorded in Remark 11.86. Since $I^{1/2}$ is invertible, $\dim S = \dim E = k - r$, and

$$\min_{\boldsymbol{e}\in E} (\boldsymbol{e}-\boldsymbol{m})^{\mathsf{T}}\, I\, (\boldsymbol{e}-\boldsymbol{m}) = \min_{\boldsymbol{s}\in S} \|\boldsymbol{s}-\boldsymbol{V}_n\|^2 = \|P\boldsymbol{V}_n\|^2 \;, \tag{A.419}$$

with P the orthogonal projector onto $S^{\perp}$, of dimension r: the nearest point of a subspace to a given vector is its orthogonal projection, and what is left over is the projection onto the orthogonal complement.

By Lemma A.208, $\boldsymbol{Z}_n \to \mathcal{N}(0, I)$ in distribution; applying Theorem A.206 to the linear map $\boldsymbol{z} \mapsto I^{-1/2}\boldsymbol{z}$ — for each $\boldsymbol{a}$, $\boldsymbol{a}^{\mathsf{T}} I^{-1/2}\boldsymbol{Z}_n \to \mathcal{N}(0, \boldsymbol{a}^{\mathsf{T}} I^{-1/2} I I^{-1/2}\boldsymbol{a}) = \mathcal{N}(0, \|\boldsymbol{a}\|^2)$ — gives $\boldsymbol{V}_n \to \boldsymbol{V} \sim \mathcal{N}(0, \mathbb{1})$ in distribution. The map $\boldsymbol{v} \mapsto \|P\boldsymbol{v}\|^2$ is continuous, so the continuous-mapping clause of Theorem A.206 gives $\|P\boldsymbol{V}_n\|^2 \to \|P\boldsymbol{V}\|^2$ in distribution, and $\|P\boldsymbol{V}\|^2 \sim \chi^2_r$ by Lemma 11.83. Finally, adding an $o_P(1)$ term does not change a limit in distribution, by the sum clause of Lemma 11.52. Hence $q \to \chi^2_r$. □

Remark A.215 (What "the" likelihood-ratio statistic means). Equation (11.80) defines q through the *global* suprema of the likelihood over Θ and over Θ_0, while Theorem A.214 is proved for the local maximisers Equation (A.403). The gap is the same one Remark 11.69 records for Theorem 11.67: identifying a consistent local maximiser with the global one is Wald's theorem [950], under conditions this treatise does not assume, and a likelihood with several local maxima can have a consistent root far from its highest peak at any finite n. Stating the theorem for the local statistic is therefore not a weakening introduced here — it is the same statement the chapter's one-parameter derivation proves, made explicit. The number r is unaffected: it is $\operatorname{codim} E$, a property of the constraint alone.

Remark A.216 (Where the two owed steps entered). It is worth marking the two places, because they are the whole content of this appendix. Uniformity — item (i) — is Equation (A.406): the remainder is bounded by $\frac{1}{6} n k^3 \bar{M}_n \|\delta\|^3$ with a coefficient that does not depend on θ, which is what (R5)'s single dominating function M buys and what a pointwise Taylor expansion would not. It is used at two moving points, $\hat{\theta}_n$ and $\hat{\theta}_{0,n}$, neither of which is known in advance; a θ-dependent remainder would say nothing about either. Localisation — item (ii) — is Equation (A.410): both estimators sit within $O_P(n^{-1/2})$ of θ_0, so the surface is only ever probed at distance $O(n^{-1/2})$, where Lemma A.212 bounds its departure from the tangent space by $O(n^{-1})$ — one order smaller, which after the $\sqrt{n}$ rescaling of Equation (A.417) is the $O(n^{-1/2})$ that vanishes. That is the honest content of writing $\Theta_0 \simeq \theta_0 + E$, and it is why second-order smoothness of the constraint, and not merely differentiability, is assumed.

Remark A.217. Appendix A.21 discharges the proof obligation left at Section 11.5.4 of Chapter 11. With it, Corollary 11.89 — one parameter of interest, any number of nuisance parameters, one degree of freedom — is a corollary of a proved theorem rather than of an argued one, and so is the practice it licenses: a collider search reads its significance off a one-degree-of-freedom distribution however many nuisance parameters describe the background and the detector [2172]. The two failure modes remain exactly as Remark 11.90 and Remark 11.91 describe them, and both are visible in the proof above: a parameter on the boundary violates (R3), so Equation (A.403) is no longer an interior maximisation and Lemma A.211 loses the two-sided estimate; a parameter unidentified under the null makes I singular, so $\lambda = 0$ and every step from Equation (A.412) onward fails at once.

A.22 Rice's Formula for the Expected Number of Upcrossings

This appendix proves Theorem 11.102 of Chapter 11: for a smooth Gaussian field of unit variance the expected number of upcrossings of a level z factorises into $\mathrm{e}^{-z^2/2}/(2\pi)$ times an integral of the standard deviation of the derivative, so that the level enters only

through a Gaussian factor. That factorisation is what the whole look-elsewhere machinery of Section 11.6 rests on: it is what makes Theorem 11.103 true, and hence what allows a search to count upcrossings at a low threshold in a few hundred simulated experiments and extrapolate to a threshold at which nothing could ever be simulated. The formula is section 3.3 of the second instalment of Rice's memoir on random noise [883]; the first instalment [877] carries the shot effect and the power spectra and not this.

The proof is in three moves. A purely deterministic counting identity expresses the number of upcrossings of a C^1 function as the limit of a normalised integral over the set where the function is near the level — Kac's device (Appendix A.22.3). The expectation of that integral factorises, because a field of constant variance is uncorrelated with its own derivative at each point, and for jointly Gaussian variables uncorrelated means independent (Appendix A.22.2). And the two elementary Gaussian integrals that remain are evaluated (Appendix A.22.5). The passage from the integral identity to its expectation is where Lebesgue theory enters, and it is the only thing quoted; it is stated precisely in Remark A.224.

A.22.1 Hypotheses and statement

Definition A.218 (The hypotheses)**.** Let $M = [m_1, m_2]$ be a compact interval and X a real-valued process on M satisfying:

(H1) *Gaussian, standardised*: every finite linear combination $\sum_j a_j X(n_j)$, $n_j \in M$, is Gaussian with mean zero; and the covariance function $r(n, n') = \langle X(n)X(n')\rangle$ satisfies $r(m, m) = 1$ for every $m \in M$.

(H2) *Smooth covariance*: r is twice continuously differentiable on $M \times M$.

(H3) *Smooth paths*: almost every sample path of X is continuously differentiable on M, and for each m the difference quotients $(X(m+h) - X(m))/h$ converge to $X'(m)$ in mean square as $h \to 0$.

(H4) *Non-degeneracy and non-tangency*: writing $\sigma_1(m)^2 = \operatorname{var} X'(m)$, the function σ_1 is continuous and strictly positive on M; and for the level z under consideration, almost surely there is no $m \in M$ with $X(m) = z$ and $X'(m) = 0$.

Rests on Definitions 11.28 and 11.100.

Theorem A.219 (Rice's formula)**.** *Under Definition A.218, the expected number of upcrossings of the level z by X on M, in the sense of Definition 11.100, is*

$$\langle N_z(X)\rangle = \frac{\mathrm{e}^{-z^2/2}}{2\pi} \int_{m_1}^{m_2} \sigma_1(m)\,\mathrm{d}m\,. \tag{A.420}$$

Rests on Definitions 11.100 and A.218.

The two sides carry the same SI dimension, namely none. The field is standardised, so X and the level z are pure numbers; the parameter m carries whatever unit the scanned quantity carries — in the application of Section 11.6 it is a mass — and X', being a derivative with respect to it, carries the reciprocal of that unit. So $\sigma_1(m)\,\mathrm{d}m$ is dimensionless, as a count must be. This is also the quickest check that no factor has been lost: the level enters only through $\mathrm{e}^{-z^2/2}$, which could carry no unit, and the geometry only through $\int \sigma_1\,\mathrm{d}m$, which is the total number of "independent looks" the scan contains.

A.22.2 The joint law of the field and its derivative

Lemma A.220 (At each point, field and derivative are independent Gaussians)**.** *Under (H1)–(H4), for each fixed $m \in M$ the pair $(X(m), X'(m))$ is jointly Gaussian with mean zero and covariance matrix* $\operatorname{diag}(1, \sigma_1(m)^2)$*. The two components are therefore independent, with joint density*

$$p_m(v, w) = \frac{\mathrm{e}^{-v^2/2}}{\sqrt{2\pi}} \cdot \frac{\mathrm{e}^{-w^2/(2\sigma_1(m)^2)}}{\sqrt{2\pi}\,\sigma_1(m)}\,. \tag{A.421}$$

Rests on Definition A.218 and Corollary A.178.

Proof. *Derives Lemma A.220.*

Joint Gaussianity. Fix $a, b \in \mathbb{R}$ and write $D_h = (X(m+h) - X(m))/h$, a linear combination of two values of X; by (H1) the variable $S_h = aX(m) + bD_h$ is Gaussian with mean zero, say $S_h \sim \mathcal{N}(0, v_h)$. Put $S = aX(m) + bX'(m)$. By (H3), $D_h \to X'(m)$ in mean square, so $S_h \to S$ in mean square and hence in probability, and therefore in distribution. Mean-square convergence also forces $v_h = \langle S_h^2 \rangle \to \langle S^2 \rangle = v$, since $\left| \langle S_h^2 \rangle^{1/2} - \langle S^2 \rangle^{1/2} \right| \le \langle (S_h - S)^2 \rangle^{1/2}$ by the triangle inequality for the second-moment norm. The characteristic function of S_h is $\mathrm{e}^{-v_h t^2/2}$ (Proposition 11.29 and the computation of the Gaussian transform in the derivation of Theorem 11.46), which converges to $\mathrm{e}^{-vt^2/2}$ for every t. By Corollary A.181 the characteristic function of S is that limit, and by Corollary A.178 S is therefore $\mathcal{N}(0, v)$. As a and b were arbitrary, the pair is jointly Gaussian.

The covariance vanishes. By the Cauchy–Schwarz inequality (Lemma 11.63, applied without centring, both variables having mean zero), $|\langle X(m)\,(D_h - X'(m))\rangle| \le \langle X(m)^2 \rangle^{1/2} \left\langle (D_h - X'(m))^2 \right\rangle^{1/2} \to 0$, so

$$\langle X(m) X'(m) \rangle = \lim_{h \to 0} \langle X(m) D_h \rangle = \lim_{h \to 0} \frac{r(m, m+h) - r(m, m)}{h} = \partial_2 r(m, m)\,, \tag{A.422}$$

where ∂_2 is the derivative in the second slot, which exists and is continuous by (H2). Now $r(m, m) = 1$ for every m by (H1); differentiating this identity in m by the chain rule (Proposition 7.72) gives $\partial_1 r(m, m) + \partial_2 r(m, m) = 0$, while r is symmetric, $r(n, n') = r(n', n)$, so $\partial_1 r(m, m) = \partial_2 r(m, m)$. Hence both vanish, and $\langle X(m) X'(m) \rangle = 0$ by Equation (A.422). Together with $\operatorname{var} X(m) = r(m, m) = 1$ and $\operatorname{var} X'(m) = \sigma_1(m)^2 > 0$ from (H4), the covariance matrix is $\operatorname{diag}(1, \sigma_1^2)$, which is invertible.

Independence. A jointly Gaussian pair with mean zero and invertible covariance Σ has density $(2\pi)^{-1} (\det \Sigma)^{-1/2} \exp\left(-\tfrac{1}{2} \boldsymbol{y}^{\mathsf{T}} \Sigma^{-1} \boldsymbol{y}\right)$; with Σ diagonal the quadratic form splits and the density factorises as Equation (A.421), which by Definition 11.25 is independence. □

The lemma is the structural heart of Equation (A.420) and is worth stating in words, because it is what Theorem 11.103 exploits: *constancy of the variance* is what makes field and derivative independent, and once they are, the level z can influence only the first factor of Equation (A.421) while the derivative supplies the same average slope at every level. Had the field not been standardised, the two would be correlated wherever $\mathrm{d} \operatorname{var} X / \mathrm{d}m \neq 0$ and no such factorisation would exist.

A.22.3 Kac's counting integral

The next lemma contains no probability at all. It is the observation that a transversal upcrossing contributes exactly 2δ to $\int \mathbf{1}_{\{|g-z|<\delta\}} (g')^+$, whatever δ is, while a downcrossing contributes nothing.

Lemma A.221 (Counting identity)**.** *Let $g : M \to \mathbb{R}$ be continuously differentiable and let $z \in \mathbb{R}$ satisfy $g(m_1) \neq z$, $g(m_2) \neq z$, and $g'(m) \neq 0$ at every m with $g(m) = z$. Then g*

has finitely many upcrossings of z, their number $N_z(g)$ in the sense of Definition 11.100 is finite, and there is $\delta_0(g) > 0$ with

$$\frac{1}{2\delta}\int_{m_1}^{m_2} \mathbf{1}_{\{|g(m)-z|<\delta\}}\,\big(g'(m)\big)^+\,\mathrm{d}m = N_z(g) \qquad \text{for every } 0<\delta<\delta_0(g)\,, \tag{A.423}$$

where $y^+ = \max(y,0)$. In particular the left side converges to $N_z(g)$ as $\delta \downarrow 0$.
Rests on Definition 11.100 and Theorems 7.24 and 7.43.

Proof. *Derives Lemma A.221.*

Finiteness. Let $Z = \{m \in M \mid g(m) = z\}$, a closed subset of the compact M and hence compact; by hypothesis it misses both endpoints. If $p \in Z$ then $g'(p) \neq 0$, and g' is continuous, so g is strictly monotone on a neighbourhood of p and takes the value z there only at p: every point of Z is isolated. A compact set all of whose points are isolated is finite — the singletons form an open cover with no proper subcover. Write $Z = \{p_1 < \cdots < p_J\}$, all interior to M.

A separating radius. For each j choose $\eta > 0$, the same for all j, small enough that the closed intervals $U_j = [p_j - \eta, p_j + \eta]$ are pairwise disjoint, contained in (m_1, m_2), and such that g' keeps the sign of $g'(p_j)$ throughout U_j; continuity of g' and finiteness of Z make this possible. On $M \setminus \bigcup_j (p_j - \eta, p_j + \eta)$ — a compact set on which $g - z$ does not vanish — the continuous function $|g - z|$ attains a strictly positive minimum μ (Theorem 7.24). Let $\delta_0 = \min\big(\mu,\ \min_j |g(p_j - \eta) - z|\,,\ \min_j |g(p_j + \eta) - z|\big)$, which is positive because g is strictly monotone on each U_j with $g(p_j) = z$.

The integral, for $\delta < \delta_0$. The set $A_\delta = \{m \in M \mid |g(m) - z| < \delta\}$ is contained in $\bigcup_j U_j$, since $\delta < \mu$. Fix j. On U_j the function g is strictly monotone and continuous, so $A_\delta \cap U_j$ is the open interval (α_j, β_j) whose endpoints are the two solutions of $g = z \mp \delta$ inside U_j; these exist and lie in the interior of U_j by the intermediate value theorem (Theorem 7.23) and the choice $\delta < \delta_0$.

If $g' > 0$ on U_j — an upcrossing at p_j, by Definition 11.100, since g passes from below z to above it there — then $(g')^+ = g'$ on U_j and, by the second fundamental theorem of calculus (Theorem 7.43),

$$\int_{\alpha_j}^{\beta_j} \big(g'(m)\big)^+\,\mathrm{d}m = g(\beta_j) - g(\alpha_j) = (z+\delta) - (z-\delta) = 2\delta\,. \tag{A.424}$$

If instead $g' < 0$ on U_j, then $(g')^+ = 0$ there and the contribution is 0; that is a downcrossing and is correctly not counted. Summing Equation (A.424) over the finitely many cells, the integral in Equation (A.423) equals 2δ times the number of j with $g'(p_j) > 0$, which is $N_z(g)$; dividing by 2δ gives the identity, with a value independent of δ. □

Remark A.222 (Why the hypotheses of the lemma are the hypotheses (H4)). The three conditions imposed on g are exactly what (H4) and the continuity of the one-dimensional Gaussian law provide for almost every path. That $X(m_i) \neq z$ almost surely is immediate: $X(m_i)$ is standard Gaussian and $\Pr(X(m_i) = z) = 0$ by Proposition 11.22. That no crossing is tangential is the second half of (H4), and it is assumed rather than derived — see Remark A.224.

A.22.4 Taking the expectation

Theorem A.223 (Tonelli and dominated convergence, as used here)**.** *Let $h(m,\omega) \geq 0$ be jointly measurable on $M \times \Omega$. Then $\omega \mapsto \int_M h(m,\omega)\mathrm{d}m$ and $m \mapsto \langle h(m,\cdot)\rangle$ are measurable and*

$$\Big\langle \int_M h(m,\cdot)\,\mathrm{d}m \Big\rangle = \int_M \langle h(m,\cdot)\rangle\,\mathrm{d}m\,, \tag{A.425}$$

both sides possibly infinite. And if $Y_\delta \to Y$ almost surely as $\delta \downarrow 0$ with $|Y_\delta| \le D$ for a single D with $\langle D\rangle < \infty$, then $\langle Y_\delta\rangle \to \langle Y\rangle$ [1894].
Rests on Definition 11.4.

Remark A.224 (What is quoted here). Two things, and they are the only ones.

First, Theorem A.223. Chapter 7 builds the Riemann–Darboux integral and no more, so neither Tonelli's theorem nor dominated convergence is available from anything earlier in this treatise; both belong to Lebesgue theory, and the same two statements are quoted for the same reason in Remark A.171. The interchange Equation (A.425) is unavoidable here: the object being averaged is an integral over M of a quantity that depends on the whole path, and there is no elementary truncation — of the kind used at Equation (11.32) to avoid dominated convergence in Lemma 11.44 — that reaches it. The dominated-convergence step is the passage from Equation (A.423), which holds path by path only for $\delta < \delta_0(g)$ with a threshold that depends on the path and has no positive lower bound over all paths, to a single limit under the expectation. The dominating variable is $D = \sup_{0<\delta<1} (2\delta)^{-1} \int_M \mathbf{1}_{\{|X-z|<\delta\}} (X')^+ \,\mathrm{d}m$, whose integrability is part of what is being assumed.

Second, the non-tangency half of (H4). It is stated as a hypothesis rather than derived. It is known to follow from $\sigma_1^2 > 0$ together with the smoothness already assumed, by an argument (Bulinskaya's lemma) that bounds the probability of the field being near z while its derivative is near 0 and requires the same Lebesgue theory; this treatise does not build it, and no elementary substitute is offered. It is not a vacuous assumption: a field with a genuine tangency at level z would touch the level without crossing it, and Lemma A.221 would fail on that path — the set $\{|g - z| < \delta\}$ would not be a finite union of monotone cells.

Everything else — the joint law Lemma A.220, the counting identity Lemma A.221, the two Gaussian integrals of Appendix A.22.5, and the assembly — is derived here.

A.22.5 Evaluation

Lemma A.225 (The mean positive part). *If $W \sim \mathcal{N}(0, \sigma^2)$ with $\sigma > 0$, then $\langle W^+\rangle = \sigma/\sqrt{2\pi}$.*
Rests on Definition 11.28 and Equation (7.27).

Proof. *Derives Lemma A.225.*
By Equation (11.20) and the substitution $w = \sigma y$ (Equation (7.27)),

$$\left\langle W^+ \right\rangle = \int_0^\infty w \, \frac{\mathrm{e}^{-w^2/(2\sigma^2)}}{\sqrt{2\pi}\,\sigma} \,\mathrm{d}w = \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty y \, \mathrm{e}^{-y^2/2} \,\mathrm{d}y = \frac{\sigma}{\sqrt{2\pi}} \left[-\mathrm{e}^{-y^2/2}\right]_0^\infty = \frac{\sigma}{\sqrt{2\pi}} \,.$$

□

Lemma A.226 (The level factor). *Write $\varphi(v) = \mathrm{e}^{-v^2/2}/\sqrt{2\pi}$ for the standard Gaussian density. Then for every $z \in \mathbb{R}$ and $\delta > 0$ there is $\zeta_\delta \in [z - \delta, z + \delta]$ with*

$$\frac{1}{2\delta} \Pr(|X(m) - z| < \delta) = \varphi(\zeta_\delta) \,, \qquad |\varphi(\zeta_\delta) - \varphi(z)| \;\le\; \frac{\delta}{\sqrt{2\pi\mathrm{e}}} \,, \tag{A.426}$$

and the bound does not depend on m.
Rests on Definition 11.28 and Theorem 7.35.

Proof. *Derives Lemma A.226.*
By (H1) the variable $X(m)$ is standard Gaussian for every m, so $\Pr(|X(m) - z| < \delta) = \int_{z-\delta}^{z+\delta} \varphi(v)\mathrm{d}v$, independently of m. Since φ is continuous, the mean value theorem for integrals — apply Theorem 7.35 to the antiderivative Φ, whose derivative is φ by Proposition 11.22 —

gives $\int_{z-\delta}^{z+\delta}\varphi = 2\delta\,\varphi(\zeta_\delta)$ for some ζ_δ in the interval. Applying Theorem 7.35 once more, this time to φ itself, $|\varphi(\zeta_\delta)-\varphi(z)| \le \delta \sup_v |\varphi'(v)|$, and $\varphi'(v) = -v\varphi(v)$ has modulus maximised at $v = \pm 1$, where it equals $\mathrm{e}^{-1/2}/\sqrt{2\pi}$. □

Proof of Theorem A.219. *Derives Theorem A.219.*

By (H3) and Remark A.222, almost every sample path satisfies the hypotheses of Lemma A.221, so almost surely

$$N_z(X) = \lim_{\delta\downarrow 0} Y_\delta\,, \qquad Y_\delta = \frac{1}{2\delta}\int_{m_1}^{m_2} \mathbf{1}_{\{|X(m)-z|<\delta\}}\,\big(X'(m)\big)^+\,\mathrm{d}m\,. \tag{A.427}$$

Take expectations. The dominated-convergence clause of Theorem A.223 gives $\langle N_z(X)\rangle = \lim_{\delta\downarrow 0}\langle Y_\delta\rangle$, and the Tonelli clause — the integrand is non-negative — gives, for each fixed δ,

$$\langle Y_\delta\rangle = \frac{1}{2\delta}\int_{m_1}^{m_2} \Big\langle \mathbf{1}_{\{|X(m)-z|<\delta\}}\,\big(X'(m)\big)^+\Big\rangle\,\mathrm{d}m\,. \tag{A.428}$$

At each fixed m the indicator is a function of $X(m)$ alone and the positive part a function of $X'(m)$ alone, and those two are independent by Lemma A.220; so the expectation factorises, and by Lemmas A.225 and A.226,

$$\frac{1}{2\delta}\Big\langle \mathbf{1}_{\{|X(m)-z|<\delta\}}\,\big(X'(m)\big)^+\Big\rangle = \varphi(\zeta_\delta)\,\frac{\sigma_1(m)}{\sqrt{2\pi}}\,. \tag{A.429}$$

From Equations (A.421) and (A.428) (factorise the expectation at each m by independence, then evaluate the level factor and the mean positive part).

The point ζ_δ does not depend on m (Lemma A.226), so it comes outside the integral:

$$\langle Y_\delta\rangle = \frac{\varphi(\zeta_\delta)}{\sqrt{2\pi}}\int_{m_1}^{m_2}\sigma_1(m)\,\mathrm{d}m\,, \tag{A.430}$$

the integral being finite because σ_1 is continuous on the compact M (H4) and therefore bounded and Riemann integrable (Theorem 7.40). Letting $\delta\downarrow 0$ and using the uniform estimate of Equation (A.426),

$$\langle N_z(X)\rangle = \frac{\varphi(z)}{\sqrt{2\pi}}\int_{m_1}^{m_2}\sigma_1(m)\,\mathrm{d}m = \frac{\mathrm{e}^{-z^2/2}}{2\pi}\int_{m_1}^{m_2}\sigma_1(m)\,\mathrm{d}m\,,$$

since $\varphi(z)/\sqrt{2\pi} = \mathrm{e}^{-z^2/2}/(2\pi)$. That is Equation (A.420). □

Corollary A.227 (The stationary case)**.** *If in addition X is stationary, so that $r(n,n')$ depends only on $n-n'$, then σ_1 is the constant $\sqrt{\lambda_2}$ with $\lambda_2 = -r''(0)$, and*

$$\langle N_z(X)\rangle = \frac{L}{2\pi}\sqrt{\lambda_2}\,\mathrm{e}^{-z^2/2}\,, \qquad L = m_2 - m_1\,. \tag{A.431}$$

Rests on Theorem A.219.

Proof. *Derives Corollary A.227.*

With D_h as in Lemma A.220, the second-moment norm is continuous under mean-square convergence, so $\langle D_h D_{h'}\rangle \to \langle X'(m)^2\rangle$ as $h,h'\to 0$. On the other hand

$$\langle D_h D_{h'}\rangle = \frac{r(m+h,m+h') - r(m+h,m) - r(m,m+h') + r(m,m)}{h\,h'}\,, \tag{A.432}$$

a second-order difference quotient of the twice continuously differentiable r, which converges to $\partial_1\partial_2 r(m,m)$. Hence $\sigma_1(m)^2 = \partial_1\partial_2 r(m,m)$ in general. Now write $r(n,n') = \varrho(n-n')$ with ϱ twice continuously differentiable by (H2) and $\varrho(0)=1$; then $\partial_1 r = \varrho'(n-n')$ and $\partial_2\partial_1 r = -\varrho''(n-n')$, so $\sigma_1(m)^2 = -\varrho''(0) = \lambda_2$, independent of m, and the integral in Equation (A.420) is $L\sqrt{\lambda_2}$. □

Remark A.228. Appendix A.22 discharges the proof obligation of Theorem 11.102 in Section 11.6.2 of Chapter 11, the last of the results that section had to import. With it, the chain the look-elsewhere effect rests on is complete: Davies' bound Theorem 11.101 was already derived there from the intermediate value theorem and Markov's inequality; Theorem 11.103 — the exponential level dependence $\langle N_u \rangle = \langle N_{u_0} \rangle \, \mathrm{e}^{-(u-u_0)/2}$, which is what lets a search estimate a trials factor from upcrossings counted at a one-sigma threshold — follows from Equation (A.420) at $z = \sqrt{u}$; and Corollary 11.105 follows from those two. What is assumed and not proved is named in Remark A.224: two theorems of Lebesgue theory, and the absence of tangential crossings. The extension of Theorem 11.103 to a scan over several parameters remains, as Remark 11.104 says, reported and not derived — and nothing in this treatise uses it.

A.23 The Hilbert–Schmidt Theorem for Compact Self-Adjoint Operators

This appendix proves Theorem 12.44 of Chapter 12: a compact self-adjoint operator on a Hilbert space has an orthonormal system of eigenvectors, with real eigenvalues accumulating only at 0, in terms of which the operator is a norm- convergent sum of rank-one pieces. It is the one infinite-dimensional theorem that reproduces the finite-dimensional spectral theorem Theorem 5.75 without weakening any of its conclusions, and it is the theorem Hilbert proved for symmetric integral kernels [388] and Schmidt recast in the language used here [338].

Everything below rests on material already available: the numerical-radius formula $\|A\| = \sup_{\|x\|=1} |\langle x|Ax\rangle|$ (Proposition 12.43), Bessel's inequality (Proposition 12.27), the projection theorem (Theorem 12.18), and the equivalence of compactness and sequential compactness in a metric space (Theorem 6.31). No weak topology is used; the closing Remark A.236 says precisely what the standard weak-compactness argument would have supplied and why the elementary route replaces it.

A.23.1 Statement

Theorem A.229 (Hilbert–Schmidt)**.** *Let $\mathcal{H} \neq \{0\}$ be a Hilbert space and let $A \in \mathcal{B}(\mathcal{H})$ be compact and self-adjoint (Definition 12.41). Then there is a finite or countable orthonormal system $\{e_n\}$ in $\mathcal{H}$ and real numbers $\lambda_n \neq 0$ with $|\lambda_1| \geq |\lambda_2| \geq \cdots$ such that*

$$Ax = \sum_n \lambda_n \, \langle e_n|x\rangle \; e_n \,, \qquad \forall\, x \in \mathcal{H} \,, \tag{A.433}$$

the series converging in the norm of $\mathcal{H}$. If the system is infinite then $\lambda_n \longrightarrow 0$. Every eigenspace belonging to a non-zero eigenvalue is finite dimensional, $Ae_n = \lambda_n e_n$, and $\|A\| = \max_n |\lambda_n|$ when $A \neq 0$. Finally

$$\mathcal{H} = \ker A \oplus \overline{\mathit{span}\,\{e_n\}} \,, \tag{A.434}$$

so that adjoining any orthonormal basis of $\ker A$ *to $\{e_n\}$ produces an orthonormal basis of $\mathcal{H}$ consisting of eigenvectors of A.*

Rests on Definition 12.41, Proposition 12.43, and Theorem 12.30.

The proof occupies the rest of the section: the sequential form of compactness (Appendix A.23.2), the attainment of the supremum and the first eigenvector (Appendix A.23.3), the recursion (Appendix A.23.4), and the passage to the limit (Appendix A.23.5).

A.23.2 Compactness in sequential form

Lemma A.230 (Sequential characterisation)**.** *$A \in \mathcal{B}(\mathcal{H})$ is compact if and only if every bounded sequence (x_n) in $\mathcal{H}$ admits a subsequence (x_{n_k}) for which (Ax_{n_k}) converges in $\mathcal{H}$.*
Rests on Definition 12.41 and Theorem 6.31.

Proof. *Derives Lemma A.230.*

Suppose A is compact and let $\|x_n\| \le c$ for all n; we may assume $c > 0$. The vectors Ax_n/c lie in the image of the closed unit ball, hence in the set $K = \overline{A\{x \mid \|x\| \le 1\}}$, which is compact by hypothesis. A compact subset of the metric space $\mathcal{H}$ is sequentially compact (Theorem 6.31), so some subsequence Ax_{n_k}/c converges in K, and therefore Ax_{n_k} converges.

Conversely, assume the sequential property and let (y_m) be a sequence in K. Each y_m is a limit of points Ax with $\|x\| \le 1$, so we may choose x_m with $\|x_m\| \le 1$ and $\|y_m - Ax_m\| < 1/m$. By hypothesis some Ax_{m_k} converges, and then y_{m_k} converges to the same limit, which lies in the closed set K. Hence K is sequentially compact, hence compact by Theorem 6.31. □

Lemma A.231 (Restriction to an invariant closed subspace)**.** *Let A be compact and self-adjoint and let $M \subseteq \mathcal{H}$ be a closed subspace with $AM \subseteq M$. Then M is a Hilbert space, and the restriction $A|_M$ is a compact self-adjoint operator on M with $\|A|_M\| \le \|A\|$. Moreover $AM^\perp \subseteq M^\perp$.*
Rests on Lemma A.230, Theorem 12.18, and Definition 12.41.

Proof. *Derives Lemma A.231.*

A closed subspace of a complete metric space is complete, so M is a Hilbert space with the inherited inner product. Symmetry of $A|_M$ is inherited, and by Proposition 12.42 a symmetric operator defined on all of the Hilbert space M is self-adjoint there. Compactness is Lemma A.230: a bounded sequence in M is a bounded sequence in $\mathcal{H}$, so (Ax_n) has a convergent subsequence, whose limit lies in the closed set M. The norm bound is immediate, the supremum being taken over a smaller set. For the last claim, let $y \in M^\perp$ and $z \in M$; then $Az \in M$, so $\langle z|Ay\rangle = \langle Az|y\rangle = 0$, and $Ay \in M^\perp$. □

A.23.3 The supremum is attained: the first eigenvector

This is the step Chapter 12 left open. The numerical-radius formula Equation (12.26) exhibits $\|A\|$ as a supremum; for a compact operator the supremum is a maximum, and the maximising vector is an eigenvector.

Lemma A.232 (Attainment)**.** *Let $A \in \mathcal{B}(\mathcal{H})$ be compact and self-adjoint with $A \neq 0$. Then there are $e \in \mathcal{H}$ with $\|e\| = 1$ and $\lambda \in \mathbb{R}$ with $|\lambda| = \|A\|$ such that $Ae = \lambda e$.*
Rests on Propositions 12.42 and 12.43 and Lemma A.230.

Proof. *Derives Lemma A.232.*

By Proposition 12.43 there are unit vectors x_n with

$$|\langle x_n|Ax_n\rangle| \longrightarrow \|A\| \,. \tag{A.435}$$

Each number $\langle x_n|Ax_n\rangle$ is real, because A is self-adjoint (Proposition 12.42). A real sequence whose absolute values converge to $\|A\|$ has a subsequence converging to $+\|A\|$ or to $-\|A\|$: infinitely many of its terms are non-negative or infinitely many are negative, and along the corresponding indices the absolute value carries the sign. Relabel, and let

$$\langle x_n|Ax_n\rangle \longrightarrow \lambda\,, \qquad \lambda \in \mathbb{R}\,, \qquad |\lambda| = \|A\| \neq 0\,. \tag{A.436}$$

Expand, using λ real, $\|x_n\| = 1$, and $\langle Ax_n|x_n\rangle = \langle x_n|Ax_n\rangle$:

$$\|Ax_n - \lambda x_n\|^2 = \|Ax_n\|^2 - 2\lambda\,\langle x_n|Ax_n\rangle + \lambda^2\,. \tag{A.437}$$

Now $\|Ax_n\| \le \|A\| = |\lambda|$, so the right-hand side of Equation (A.437) is at most $2\lambda^2 - 2\lambda\,\langle x_n|Ax_n\rangle$, which tends to 0 by Equation (A.436). Hence

$$Ax_n - \lambda x_n \longrightarrow 0\,. \tag{A.438}$$

This much uses only self-adjointness; it says that λ is an *approximate* eigenvalue. Compactness converts the approximation into an eigenvector.

The sequence (x_n) is bounded, so by Lemma A.230 there is a subsequence with $Ax_{n_k} \longrightarrow y$ for some $y \in \mathcal{H}$. Since $\lambda \neq 0$ we may write

$$x_{n_k} = \frac{1}{\lambda}\Big(Ax_{n_k} - \big(Ax_{n_k} - \lambda x_{n_k}\big)\Big) \longrightarrow \frac{y}{\lambda} =: e\,,$$

by Equation (A.438). The norm is continuous (Corollary 12.5), so $\|e\| = \lim \|x_{n_k}\| = 1$; and A is continuous (Proposition 12.36), so $Ax_{n_k} \longrightarrow Ae$. But $Ax_{n_k} \longrightarrow y = \lambda e$. Limits in a metric space are unique, so $Ae = \lambda e$. □

A.23.4 The recursion

Lemma A.233 (Construction of the system)**.** *Let A be compact and self-adjoint. There are a finite or countable orthonormal system $\{e_n\}$ and real numbers $\lambda_n \neq 0$ with $Ae_n = \lambda_n e_n$ and $|\lambda_1| \ge |\lambda_2| \ge \cdots$, such that, writing $\mathcal{H}_1 = \mathcal{H}$ and*

$$\mathcal{H}_{N+1} = \{e_1, \dots, e_N\}^{\perp}\,, \qquad A_{N+1} = A|_{\mathcal{H}_{N+1}}\,, \tag{A.439}$$

one has $\|A_{N+1}\| = |\lambda_{N+1}|$ at every stage at which the construction continues, and the construction stops at stage N exactly when $A_N = 0$, i.e. when $\mathcal{H}_N \subseteq \ker A$.

Rests on Lemmas A.231 and A.232 and Theorem 12.18.

Proof. *Derives Lemma A.233.*

Set $\mathcal{H}_1 = \mathcal{H}$ and $A_1 = A$. Suppose $\mathcal{H}_N$ and A_N have been defined, with $\mathcal{H}_N$ a closed A-invariant subspace and $A_N = A|_{\mathcal{H}_N}$ compact and self-adjoint on it (Lemma A.231). If $A_N = 0$ the construction stops, and then every $x \in \mathcal{H}_N$ has $Ax = 0$. If $A_N \neq 0$, apply Lemma A.232 inside the Hilbert space $\mathcal{H}_N$: it produces a unit vector $e_N \in \mathcal{H}_N$ and a real λ_N with $|\lambda_N| = \|A_N\| \neq 0$ and $Ae_N = A_N e_N = \lambda_N e_N$.

The vector e_N lies in $\mathcal{H}_N$, hence is orthogonal to $e_1, \dots, e_{N-1}$, so the system stays orthonormal. Put $\mathcal{H}_{N+1} = \mathcal{H}_N \cap \{e_N\}^{\perp} = \{e_1, \dots, e_N\}^{\perp}$, which is closed (Proposition 12.17). It is A-invariant: for $x \in \mathcal{H}_{N+1}$ and $j \le N$,

$$\langle e_j|Ax\rangle = \langle Ae_j|x\rangle = \lambda_j^{*}\,\langle e_j|x\rangle = \lambda_j\,\langle e_j|x\rangle = 0\,, \tag{A.440}$$

using self-adjointness and $\lambda_j \in \mathbb{R}$; and $Ax \in \mathcal{H}_N$ because $\mathcal{H}_N$ is invariant. Finally $\|A_{N+1}\| \le \|A_N\|$ because A_{N+1} is a restriction of A_N, whence $|\lambda_{N+1}| \le |\lambda_N|$; the construction therefore delivers a non-increasing sequence of moduli, and it is exactly the statement $\|A_N\| = |\lambda_N|$ that Lemma A.232 supplies at each stage. □

Lemma A.234 (The eigenvalues tend to zero, with finite multiplicity)**.** *If the construction of Lemma A.233 does not stop, then $\lambda_n \longrightarrow 0$. Moreover, for every $\lambda \neq 0$ the eigenspace $\ker(A - \lambda\mathbb{1})$ is finite dimensional.*

Rests on Lemmas A.230 and A.233.

Proof. *Derives Lemma A.234.*

Both statements follow from one computation. Let (f_n) be an orthonormal sequence with $Af_n = \nu_n f_n$. For $n \neq m$, orthogonality gives

$$\|Af_n - Af_m\|^2 = \|\nu_n f_n - \nu_m f_m\|^2 = |\nu_n|^2 + |\nu_m|^2\,. \tag{A.441}$$

Eigenvalues. Apply this to $f_n = e_n$, $\nu_n = \lambda_n$. If (λ_n) does not tend to 0 there are $\varepsilon > 0$ and infinitely many indices with $|\lambda_n| \geq \varepsilon$; along them Equation (A.441) gives $\|Ae_n - Ae_m\|^2 \geq 2\varepsilon^2$, so (Ae_n) has no Cauchy subsequence, hence no convergent subsequence. Since (e_n) is bounded, this contradicts Lemma A.230. As the moduli are non-increasing, $\lambda_n \longrightarrow 0$.

Multiplicity. If $\ker(A - \lambda\mathbb{1})$ were infinite dimensional for some $\lambda \neq 0$, Gram–Schmidt (Proposition 12.23) would produce an infinite orthonormal sequence (f_n) inside it, all with $\nu_n = \lambda$, and Equation (A.441) would give $\|Af_n - Af_m\|^2 = 2\,|\lambda|^2 > 0$: the same contradiction. □

A.23.5 Passage to the limit

Proof of Theorem A.229. *Derives Theorem A.229.*

Let $\{e_n\}$, $\{\lambda_n\}$ be the system of Lemma A.233. Fix $x \in \mathcal{H}$ and put

$$x_N = x - \sum_{n=1}^{N} \langle e_n | x \rangle \; e_n \,. \tag{A.442}$$

Then $\langle e_j | x_N \rangle = \langle e_j | x \rangle - \langle e_j | x \rangle = 0$ for $j \leq N$, so $x_N \in \mathcal{H}_{N+1}$, and Bessel's inequality (Proposition 12.27) gives

$$\|x_N\|^2 = \|x\|^2 - \sum_{n=1}^{N} |\langle e_n | x \rangle|^2 \leq \|x\|^2 \,. \tag{A.443}$$

Applying A to Equation (A.442) and using $Ae_n = \lambda_n e_n$,

$$Ax - \sum_{n=1}^{N} \lambda_n \, \langle e_n | x \rangle \; e_n = Ax_N = A_{N+1} x_N \,, \tag{A.444}$$

the last equality because $x_N \in \mathcal{H}_{N+1}$. Therefore, by Lemma A.233 and Equation (A.443),

$$\left\| Ax - \sum_{n=1}^{N} \lambda_n \, \langle e_n | x \rangle \; e_n \right\| \leq \|A_{N+1}\| \; \|x_N\| \leq |\lambda_{N+1}| \; \|x\| \,. \tag{A.445}$$

If the construction stops at stage N then $A_N = 0$ and Equation (A.444) already gives Equation (A.433) as a finite sum. If it does not stop, $|\lambda_{N+1}| \longrightarrow 0$ by Lemma A.234, so the right-hand side of Equation (A.445) tends to 0 and Equation (A.433) holds with convergence in norm. The eigenvalue relation $Ae_n = \lambda_n e_n$, the ordering of the moduli and the finiteness of the non-zero eigenspaces are Lemmas A.233 and A.234. For $A \neq 0$, $\|A\| = \|A_1\| = |\lambda_1| = \max_n |\lambda_n|$, the maximum being attained because the moduli decrease.

It remains to prove Equation (A.434). Write $M = \overline{\operatorname{span}\{e_n\}}$, a closed subspace, so that $\mathcal{H} = M \oplus M^\perp$ by Theorem 12.18. If $x \in M^\perp$ then $\langle e_n | x \rangle = 0$ for every n, and Equation (A.433) gives $Ax = 0$: thus $M^\perp \subseteq \ker A$. Conversely if $Ax = 0$ then for every n,

$$0 = \langle e_n | Ax \rangle = \langle Ae_n | x \rangle = \lambda_n \, \langle e_n | x \rangle \;,$$

and $\lambda_n \neq 0$ forces $\langle e_n | x \rangle = 0$, so $x \in M^\perp$; hence $\ker A = M^\perp$, which is Equation (A.434). Every vector of $\ker A$ is an eigenvector with eigenvalue 0, so an orthonormal basis of $\ker A$ adjoined to $\{e_n\}$ is an orthonormal system of eigenvectors whose closed span is $M \oplus \ker A = \mathcal{H}$: an orthonormal basis of $\mathcal{H}$ in the sense of Definition 12.29. □

Remark A.235 (Where the orthonormal basis of the kernel comes from). The last sentence of Theorem A.229 needs an orthonormal basis of the closed subspace $\ker A$, and that is the only point at which the theorem is not constructive. If $\mathcal{H}$ is separable — the case of every

physical application, and the standing assumption of Theorem 12.33 — so is $\ker A$, and Gram–Schmidt applied to a countable dense set produces the basis (Corollary 12.24). For a non-separable $\mathcal{H}$ the existence of an orthonormal basis of $\ker A$ is a maximality statement proved with Zorn's lemma (Chapter 3); the system $\{e_n\}$ itself, on which the whole content of the theorem rests, is always countable and always constructed, never chosen.

Remark A.236 (What the weak topology would have bought, and why it is not needed). The route usually taken to Lemma A.232 [1481] passes through two statements about the weak topology, neither of which this treatise develops: that the closed unit ball of a Hilbert space is weakly sequentially compact, so that a maximising sequence has a weakly convergent subsequence $x_n \rightharpoonup e$; and that a compact operator carries weakly convergent sequences to norm convergent ones, so that $Ax_n \longrightarrow Ae$ and the supremum is attained at e. Both are true and both are proved in [1481], chapter VI.

The proof given above needs neither. The reason is Equation (A.438): self-adjointness alone forces the maximising sequence to satisfy $Ax_n - \lambda x_n \longrightarrow 0$ in *norm*, and once that is known the sequential form of compactness, Lemma A.230, is enough to extract a genuine eigenvector — the vector e is recovered from Ax_{n_k} by dividing by $\lambda \neq 0$, not by any weak limit. The argument therefore quotes nothing beyond Proposition 12.43 and Theorem 6.31, and Chapter 12 may drop the reservation with which Theorem 12.44 was first stated.

Remark A.237. Appendix A.23 discharges the proof obligation of Theorem 12.44 (Section 12.3.1). The theorem is what makes the integral operators Equation (12.28) of Chapter 10 — the Green functions of a boundary-value problem — carry a discrete spectrum of modes with a complete set of eigenfunctions, and it is the exact infinite-dimensional counterpart of Equation (5.105). It is also the last point at which the counterpart is exact: Example 12.56 exhibits a bounded self-adjoint operator, not compact, with no eigenvector at all, and the repair for that case is the projection-valued measure of Appendix A.24.

A.24 The Spectral Theorem for a Bounded Self-Adjoint Operator

This appendix proves Theorem 12.59 of Chapter 12, in both of the forms stated there: the existence and uniqueness of a projection-valued measure E on $\mathbb{R}$, supported on $\sigma(A)$, with $A = \int_{\sigma(A)} \lambda \, \mathrm{d}E(\lambda)$; and the unitary equivalence of A, on a separable Hilbert space, with multiplication by a bounded real function on a finite measure space. It is the theorem three of the quantum postulates of Chapter 77 are statements about (Remark 12.63), and it is the replacement, forced by Example 12.56, for the eigenbasis that Theorem 5.75 supplies in finite dimension.

The route is the classical one of von Neumann [664, 706] in the arrangement of Reed and Simon [1481]: first the polynomial calculus and the identity $\|p(A)\| = \sup_{\sigma(A)} |p|$, which is what makes the construction possible at all; then the continuous calculus, obtained by completing the polynomials in the supremum norm; then, for each pair of vectors, a complex measure on $\sigma(A)$; then the extension of the calculus to bounded Borel functions, whose values on indicator functions are the projections $E(\Omega)$; and finally the multiplication form. Two classical theorems are used and are not proved in this treatise — Stone–Weierstrass and Riesz–Markov. Each is stated precisely where it is used, and Remark A.251 collects them.

Throughout, $A \in \mathcal{B}(\mathcal{H})$ is self-adjoint, $\mathcal{H} \neq \{0\}$, and $\sigma(A) \subseteq \mathbb{R}$ is compact and non-empty (Theorems 12.54 and 12.55). We write $C(\sigma(A))$ for the continuous complex functions on $\sigma(A)$ with the supremum norm $\|f\|_\infty = \sup_{\lambda \in \sigma(A)} |f(\lambda)|$, and $B(\sigma(A))$ for the bounded Borel functions with the same norm.

A.24.1 Statement

Theorem A.238 (Spectral theorem, both forms)**.** *Let $A \in \mathcal{B}(\mathcal{H})$ be self-adjoint. Then:*

1. *there is exactly one projection-valued measure E on $\mathbb{R}$ (Definition 12.58) with $E(\mathbb{R} \setminus \sigma(A)) = 0$ and*
$$\langle x | Ay \rangle = \int_{\sigma(A)} \lambda \, \langle x | \mathrm{d}E(\lambda) y \rangle \,, \qquad \forall\, x, y \in \mathcal{H} \,, \tag{A.446}$$
and every $B \in \mathcal{B}(\mathcal{H})$ commuting with A commutes with every $E(\Omega)$;
2. *if $\mathcal{H}$ is separable there are a measure space (M, μ) with $\mu(M) < \infty$, a bounded real measurable function a on M, and a unitary $U : \mathcal{H} \longrightarrow L^2(M, \mu)$ with*
$$(UAU^{-1} f)(m) = a(m)\, f(m) \,, \qquad \forall\, f \in L^2(M, \mu) \,. \tag{A.447}$$

Rests on Definition 12.58, Theorem 12.55, and Proposition 12.43.

A.24.2 The polynomial calculus

For a polynomial $p(z) = \sum_{k=0}^{d} c_k z^k$ with complex coefficients write $p(A) = \sum_k c_k A^k \in \mathcal{B}(\mathcal{H})$, with $A^0 = \mathbb{1}$, and let $\bar{p}(z) = \sum_k c_k^* z^k$. Since $A^\dagger = A$, Equation (12.24) gives $p(A)^\dagger = \bar{p}(A)$.

Lemma A.239 (The norm of a self-adjoint operator lies in its spectrum)**.** *Let $B \in \mathcal{B}(\mathcal{H})$ be self-adjoint. Then $\|B\| \in \sigma(B)$ or $-\|B\| \in \sigma(B)$; consequently*
$$\|B\| = \sup_{\lambda \in \sigma(B)} |\lambda| \,. \tag{A.448}$$

Rests on Propositions 12.43 and 12.52 and Definition 12.49.

Proof. *Derives Lemma A.239.*

If $B = 0$ then $\sigma(B) = \{0\}$ by Theorem 12.54 and Proposition 12.52 and the claim is trivial, so let $B \neq 0$. Exactly as in Equation (A.436), choose unit vectors x_n with $\langle x_n | B x_n \rangle \longrightarrow \lambda$ where $\lambda \in \mathbb{R}$ and $|\lambda| = \|B\|$; this is possible by Proposition 12.43 and because $\langle x_n | B x_n \rangle$ is real (Proposition 12.42). Then, as in Equation (A.437),
$$\|(B - \lambda \mathbb{1}) x_n\|^2 = \|B x_n\|^2 - 2\lambda \, \langle x_n | B x_n \rangle + \lambda^2 \leq 2\lambda^2 - 2\lambda \, \langle x_n | B x_n \rangle \longrightarrow 0 \,. \tag{A.449}$$
If $B - \lambda \mathbb{1}$ had a bounded inverse R we would get $1 = \|x_n\| \leq \|R\| \, \|(B - \lambda \mathbb{1}) x_n\| \longrightarrow 0$, which is false; so $\lambda \in \sigma(B)$ with $|\lambda| = \|B\|$. Since $\sigma(B)$ is contained in the disc of radius $\|B\|$ (Proposition 12.52), Equation (A.448) follows. □

Lemma A.240 (Spectral mapping for polynomials)**.** *For every polynomial p and every $A \in \mathcal{B}(\mathcal{H})$,*
$$\sigma(p(A)) = p(\sigma(A)) = \{ p(\lambda) \mid \lambda \in \sigma(A) \} \,. \tag{A.450}$$

Rests on Definition 12.49 and Proposition 12.37.

Proof. *Derives Lemma A.240.*

If p is constant, $p \equiv c$, then $p(A) = c\mathbb{1}$ and both sides are $\{c\}$ (using $\sigma(A) \neq \varnothing$, Theorem 12.54). Otherwise let $\mu \in \mathbb{C}$ and factor the non-constant polynomial $p(z) - \mu$ over $\mathbb{C}$ (Theorem 8.19):
$$p(z) - \mu = c \prod_{j=1}^{d} (z - \alpha_j) \,, \qquad c \neq 0 \,, \tag{A.451}$$
so that, substituting A and using that polynomials in A commute,
$$p(A) - \mu \mathbb{1} = c \prod_{j=1}^{d} (A - \alpha_j \mathbb{1}) \,. \tag{A.452}$$

If every factor is invertible, so is the product, and $\mu \in \rho(p(A))$. Conversely suppose the product $T = \prod_j (A - \alpha_j \mathbb{1})$ is invertible and fix j. Its inverse T^{-1} commutes with A — from $TA = AT$ one gets $AT^{-1} = T^{-1}A$ on multiplying by T^{-1} on both sides — hence with every factor. Writing $S_j = \prod_{k \neq j} (A - \alpha_k \mathbb{1})$ we have $T = (A - \alpha_j \mathbb{1}) S_j = S_j (A - \alpha_j \mathbb{1})$, so

$$(A - \alpha_j \mathbb{1})\, S_j T^{-1} = \mathbb{1} = T^{-1} S_j\, (A - \alpha_j \mathbb{1})\,,$$

and $A - \alpha_j \mathbb{1}$ is invertible with inverse $S_j T^{-1} = T^{-1} S_j$. Therefore $\mu \in \sigma(p(A))$ if and only if some $\alpha_j \in \sigma(A)$; and by Equation (A.451) the α_j are exactly the roots of $p - \mu$, so this happens exactly when $\mu = p(\lambda)$ for some $\lambda \in \sigma(A)$. □

Proposition A.241 (The polynomial calculus is isometric)**.** *For every polynomial p and every self-adjoint $A \in \mathcal{B}(\mathcal{H})$,*

$$\|p(A)\| = \sup_{\lambda \in \sigma(A)} |p(\lambda)| = \|p\|_\infty\,. \tag{A.453}$$

Rests on Lemmas A.239 and A.240 and Proposition 12.39.

Proof. *Derives Proposition A.241.*

The operator $p(A)^\dagger p(A) = (\bar{p}\, p)(A)$ is self-adjoint, and $\bar{p}\, p$ is a polynomial. By the C^* identity Equation (12.25), then Lemma A.239 applied to it, then Lemma A.240,

$$\|p(A)\|^2 = \|(\bar{p}\, p)(A)\| = \sup_{\mu \in \sigma((\bar{p}\, p)(A))} |\mu| = \sup_{\lambda \in \sigma(A)} |\bar{p}(\lambda)\, p(\lambda)| = \sup_{\lambda \in \sigma(A)} |p(\lambda)|^2\,, \tag{A.454}$$

the last step because $\sigma(A) \subseteq \mathbb{R}$ (Theorem 12.55), so that $\bar{p}(\lambda) = p(\lambda)^*$ there. □

A.24.3 The continuous calculus

Theorem A.242 (Stone–Weierstrass; quoted)**.** *Let K be a compact Hausdorff space and let $\mathcal{A} \subseteq C(K)$ be a subalgebra that contains the constants, separates the points of K, and is closed under complex conjugation. Then $\mathcal{A}$ is dense in $C(K)$ for the supremum norm. For $K \subseteq \mathbb{R}$ compact the polynomials form such a subalgebra, and the statement reduces to the classical Weierstrass approximation theorem quoted in Remark 12.1.*

Rests on Definition 6.9.

Proposition A.243 (Continuous functional calculus)**.** *There is exactly one map $\Phi : C(\sigma(A)) \longrightarrow \mathcal{B}(\mathcal{H})$ that is linear, multiplicative, sends $1 \longmapsto \mathbb{1}$ and $\lambda \longmapsto A$, and is continuous for $\|\cdot\|_\infty$ and the operator norm. It satisfies, writing $f(A) = \Phi(f)$,*

$$\|f(A)\| = \|f\|_\infty\,, \qquad f(A)^\dagger = \bar{f}(A)\,, \qquad f \geq 0 \implies f(A) \geq 0\,, \tag{A.455}$$

and $f(A)$ commutes with every bounded operator commuting with A.

Rests on Propositions 12.61 and A.241 and Theorem A.242.

Proof. *Derives Proposition A.243.*

Well defined on polynomials. If two polynomials agree on $\sigma(A)$ their difference q has $\|q\|_\infty = 0$, so $q(A) = 0$ by Equation (A.453): the map $p|_{\sigma(A)} \longmapsto p(A)$ depends only on the restriction, and it is isometric.

Extension. The polynomials restricted to the compact set $\sigma(A) \subseteq \mathbb{R}$ are dense in $C(\sigma(A))$ by Theorem A.242. An isometric linear map defined on a dense subspace of a normed space, with values in the complete space $\mathcal{B}(\mathcal{H})$ (Proposition 12.37), extends uniquely to an isometry of the closure: given f, choose polynomials p_n with $\|p_n - f\|_\infty \longrightarrow 0$; then $\|p_n(A) - p_m(A)\| = \|p_n - p_m\|_\infty$ makes $(p_n(A))$ Cauchy, and its limit $\Phi(f)$ is independent of the approximating sequence by the same identity. Linearity and multiplicativity pass to

the limit because addition and multiplication are continuous in a Banach algebra (Proposition 12.37), and the first identity of Equation (A.455) is the limit of Equation (A.453). Taking adjoints term by term — legitimate because the adjoint is isometric (Proposition 12.39) — gives $\Phi(f)^\dagger = \Phi(\bar{f})$.

Positivity. If $f \geq 0$ on $\sigma(A)$ then $g = \sqrt{f}$ is continuous and real, so $\Phi(f) = \Phi(g)^2 = \Phi(g)^\dagger \Phi(g)$ is positive by Proposition 12.42(iv).

Commutation. If $BA = AB$ then B commutes with every polynomial in A, hence with the norm limits $\Phi(f)$.

Uniqueness is Proposition 12.61, already proved in the chapter. □

A.24.4 The spectral measures

Theorem A.244 (Riesz–Markov; quoted)**.** *Let K be a compact metric space. For every bounded linear functional ℓ on $C(K)$ there is exactly one regular complex Borel measure ν on K with*

$$\ell(f) = \int_K f \, \mathrm{d}\nu \,, \qquad \forall\, f \in C(K) \,, \tag{A.456}$$

and $\|\ell\|$ equals the total variation $|\nu|(K)$. If $\ell(f) \geq 0$ whenever $f \geq 0$, then ν is a positive measure and $\|\ell\| = \nu(K) = \ell(1)$. Regularity carries with it that $C(K)$ is dense in $L^2(K, \nu)$ for every positive finite Borel measure ν.

Rests on Definition 6.9.

Proposition A.245 (The measures $\mu_{x,y}$)**.** *For $x, y \in \mathcal{H}$ there is a unique regular complex Borel measure $\mu_{x,y}$ on $\sigma(A)$ with*

$$\langle x | f(A)\, y \rangle = \int_{\sigma(A)} f \, \mathrm{d}\mu_{x,y} \,, \qquad \forall\, f \in C(\sigma(A)) \,. \tag{A.457}$$

The assignment is linear in y, antilinear in x, obeys $\mu_{y,x} = \mu_{x,y}^*$ and $|\mu_{x,y}|(\sigma(A)) \leq \|x\|\,\|y\|$; and $\mu_x := \mu_{x,x}$ is a positive measure with $\mu_x(\sigma(A)) = \|x\|^2$.

Rests on Proposition A.243 and Theorems 12.46 and A.244.

Proof. *Derives Proposition A.245.*

The map $f \longmapsto \langle x | f(A) y \rangle$ is linear on $C(\sigma(A))$ and bounded, since by Equation (12.1) and Equation (A.455)

$$|\langle x | f(A) y \rangle| \leq \|x\|\, \|f(A)\|\, \|y\| = \|f\|_\infty \, \|x\|\, \|y\| \,. \tag{A.458}$$

Theorem A.244 supplies a unique regular complex Borel measure $\mu_{x,y}$ satisfying Equation (A.457), with $|\mu_{x,y}|(\sigma(A)) \leq \|x\|\,\|y\|$ by the norm statement there. Linearity in y and antilinearity in x hold for the functionals and therefore, by the uniqueness in Theorem A.244, for the measures. For the conjugation rule,

$$\left(\int f \, \mathrm{d}\mu_{x,y} \right)^* = \langle f(A) y | x \rangle = \left\langle y \middle| f(A)^\dagger x \right\rangle = \left\langle y \middle| \bar{f}(A) x \right\rangle = \int \bar{f} \, \mathrm{d}\mu_{y,x} \,,$$

and since $f \longmapsto \bar{f}$ exhausts $C(\sigma(A))$, uniqueness gives $\mu_{y,x} = \mu_{x,y}^*$. Finally, if $f \geq 0$ then $\langle x | f(A) x \rangle \geq 0$ by the positivity in Equation (A.455), so μ_x is a positive measure, and taking $f = 1$ in Equation (A.457) gives $\mu_x(\sigma(A)) = \langle x | x \rangle = \|x\|^2$. □

A.24.5 The Borel calculus

Proposition A.246 (Bounded Borel functional calculus)**.** *There is a unique map $g \longmapsto g(A)$ from $B(\sigma(A))$ to $\mathcal{B}(\mathcal{H})$ extending Φ and satisfying*

$$\langle x | g(A)\, y \rangle = \int_{\sigma(A)} g \, \mathrm{d}\mu_{x,y} \,, \qquad \forall\, x, y \in \mathcal{H} \,. \tag{A.459}$$

It is linear, multiplicative, satisfies $g(A)^\dagger = \bar{g}(A)$ and $\|g(A)\| \le \|g\|_\infty$, and $g(A)$ commutes with every bounded operator commuting with A.

Rests on Propositions A.243 and A.245 and Theorem 12.46.

Proof. *Derives Proposition A.246.*

Existence of the operator. Fix $g \in B(\sigma(A))$ and consider

$$b_g(x,y) = \int_{\sigma(A)} g \, \mathrm{d}\mu_{x,y} \,, \qquad |b_g(x,y)| \le \|g\|_\infty \, |\mu_{x,y}|\,(\sigma(A)) \le \|g\|_\infty \, \|x\| \, \|y\| \,,$$

which by Proposition A.245 is antilinear in x and linear in y. For fixed x the map $y \longmapsto b_g(x,y)$ is a bounded linear functional, so by Theorem 12.46 there is a unique vector $w(x)$ with $b_g(x,y) = \langle w(x)|y\rangle$; the map $x \longmapsto w(x)$ is linear — antilinearity in the first slot on both sides — and bounded by $\|g\|_\infty$. Its adjoint in the sense of Theorem 12.38 is the operator we call $g(A)$, so that Equation (A.459) holds with $\|g(A)\| \le \|g\|_\infty$. For continuous g the formula reproduces Equation (A.457), so the map extends Φ, and it is unique because Equation (A.459) determines all matrix elements. Linearity in g and $g(A)^\dagger = \bar{g}(A)$ follow from Equation (A.459) and $\mu_{y,x} = \mu_{x,y}^*$.

Multiplicativity, first step. Let $f \in C(\sigma(A))$ and $x, y \in \mathcal{H}$. For every $h \in C(\sigma(A))$,

$$\int h \, \mathrm{d}\mu_{x,f(A)y} = \langle x|h(A)f(A)y\rangle = \langle x|(hf)(A)y\rangle = \int hf \, \mathrm{d}\mu_{x,y} \,,$$

using multiplicativity of the continuous calculus. The two regular measures $\mu_{x,f(A)y}$ and $f\,\mu_{x,y}$ therefore integrate every continuous function alike, so they are equal by the uniqueness in Theorem A.244. Consequently, for $g \in B(\sigma(A))$,

$$\langle x|g(A)f(A)y\rangle = \int g \, \mathrm{d}\mu_{x,f(A)y} = \int gf \, \mathrm{d}\mu_{x,y} = \langle x|(gf)(A)y\rangle \,, \tag{A.460}$$

i.e. $g(A)f(A) = (gf)(A)$ for g Borel and f continuous. Taking adjoints in Equation (A.460) and replacing (f,g) by $(\bar{f},\bar{g})$ gives $f(A)g(A) = (fg)(A)$ as well.

Multiplicativity, second step. Now let $g \in B(\sigma(A))$ be fixed. For every $h \in C(\sigma(A))$, by the step just proved,

$$\int h \, \mathrm{d}\mu_{x,g(A)y} = \langle x|h(A)g(A)y\rangle = \langle x|(hg)(A)y\rangle = \int hg \, \mathrm{d}\mu_{x,y} \,,$$

so $\mu_{x,g(A)y} = g\,\mu_{x,y}$ by uniqueness again. Hence for any two bounded Borel g, g',

$$\langle x|g'(A)g(A)y\rangle = \int g' \, \mathrm{d}\mu_{x,g(A)y} = \int g'g \, \mathrm{d}\mu_{x,y} = \langle x|(g'g)(A)y\rangle \,. \tag{A.461}$$

Commutation. Let $BA = AB$. By Proposition A.243, B commutes with $f(A)$ for every continuous f, so for such f

$$\int f \, \mathrm{d}\mu_{B^\dagger x,y} = \left\langle B^\dagger x \middle| f(A)y \right\rangle = \langle x|Bf(A)y\rangle = \langle x|f(A)By\rangle = \int f \, \mathrm{d}\mu_{x,By} \,,$$

whence $\mu_{B^\dagger x,y} = \mu_{x,By}$, and evaluating Equation (A.459) on both sides gives $\langle x|Bg(A)y\rangle = \langle x|g(A)By\rangle$ for every Borel g. □

A.24.6 The projection-valued measure

Proof of Theorem A.238, part (1): existence. *Derives Theorem A.238.*

For a Borel set $\Omega \subseteq \mathbb{R}$ let χ_Ω be its indicator function, restricted to $\sigma(A)$, and put

$$E(\Omega) = \chi_\Omega(A) \,. \tag{A.462}$$

This is a bounded operator by Proposition A.246.

Projections. $\chi_\Omega^2 = \chi_\Omega$ and $\bar{\chi}_\Omega = \chi_\Omega$, so $E(\Omega)^2 = E(\Omega) = E(\Omega)^\dagger$: each $E(\Omega)$ is a projection in the sense of Proposition 12.21.

Normalisation and multiplicativity. $\chi_\varnothing = 0$ and $\chi_{\mathbb{R}} = 1$ on $\sigma(A)$ give $E(\varnothing) = 0$ and $E(\mathbb{R}) = \mathbb{1}$; also $E(\mathbb{R} \setminus \sigma(A)) = 0$, since that indicator vanishes identically on $\sigma(A)$, which is the support statement. And $\chi_{\Omega_1 \cap \Omega_2} = \chi_{\Omega_1}\chi_{\Omega_2}$ gives $E(\Omega_1 \cap \Omega_2) = E(\Omega_1)E(\Omega_2)$ by Equation (A.461).

Countable additivity. Finite additivity is linearity of the calculus: for disjoint Ω_1, Ω_2, $\chi_{\Omega_1 \cup \Omega_2} = \chi_{\Omega_1} + \chi_{\Omega_2}$. For a disjoint sequence $\{\Omega_n\}$ with union Ω, write $S_N = \sum_{n \le N} E(\Omega_n) = E(\bigcup_{n \le N} \Omega_n)$, so that $E(\Omega) - S_N = E(\Omega_{>N})$ with $\Omega_{>N} = \bigcup_{n>N} \Omega_n$. Since $E(\Omega_{>N})$ is a projection,

$$\|E(\Omega)x - S_N x\|^2 = \langle x | E(\Omega_{>N}) x \rangle = \mu_x(\Omega_{>N}) = \sum_{n>N} \mu_x(\Omega_n)\,, \tag{A.463}$$

using Equation (A.459) with $g = \chi_{\Omega_{>N}}$ and the countable additivity of the ordinary positive measure μ_x. The series $\sum_n \mu_x(\Omega_n) = \mu_x(\Omega) \le \|x\|^2$ converges, so its tail tends to 0, and Equation (12.36) holds. Thus E is a projection-valued measure in the sense of Definition 12.58.

It represents A. Taking $g = \chi_\Omega$ in Equation (A.459) identifies the complex measure of Definition 12.58 with the measure of Proposition A.245: $\langle x | E(\Omega) y \rangle = \mu_{x,y}(\Omega)$. Hence, taking $f(\lambda) = \lambda$ in Equation (A.457) — the continuous calculus assigns A to that function — we obtain Equation (A.446). The commutation statement is the last part of Proposition A.246. □

Lemma A.247 (Integration against a projection-valued measure)**.** *Let F be any projection-valued measure on $\mathbb{R}$ supported on a compact set K, and for $g \in B(K)$ define $\int g \,\mathrm{d}F$ by $\langle x | (\int g \,\mathrm{d}F) y \rangle = \int g \,\mathrm{d}\langle x | F y \rangle$. Then $g \longmapsto \int g \,\mathrm{d}F$ is linear and multiplicative.*

Rests on Definition 12.58 and Theorem 12.46.

Proof. *Derives Lemma A.247.*

Exactly as in Proposition A.246, the sesquilinear form is bounded by $\|g\|_\infty \|x\| \|y\|$ — the total variation of $\langle x | F(\cdot) y \rangle$ is at most $\|x\| \|y\|$, by the Cauchy–Schwarz inequality applied to $\langle x | F(\Omega) y \rangle = \langle F(\Omega) x | F(\Omega) y \rangle$ over a partition — so the operator exists, and linearity is clear. For multiplicativity, note first that for simple functions $g = \sum_i c_i \chi_{\Omega_i}$ and $g' = \sum_j c'_j \chi_{\Omega'_j}$ the property $F(\Omega)F(\Omega') = F(\Omega \cap \Omega')$ gives

$$\left(\int g \,\mathrm{d}F\right)\left(\int g' \,\mathrm{d}F\right) = \sum_{i,j} c_i c'_j \, F(\Omega_i \cap \Omega'_j) = \int g g' \,\mathrm{d}F\,.$$

A bounded Borel function is a uniform limit of simple Borel functions — partition the disc of radius $\|g\|_\infty$ into finitely many Borel pieces of diameter $< \varepsilon$ and take preimages — and $\|\int g \,\mathrm{d}F\| \le \|g\|_\infty$, so both sides pass to the uniform limit. □

Proof of Theorem A.238, part (1): uniqueness. *Derives Theorem A.238.*

Let E and E' both satisfy Equation (A.446) and be supported on $\sigma(A)$, and write $\nu_{x,y}, \nu'_{x,y}$ for the associated complex measures. By Lemma A.247, $\int \lambda^n \,\mathrm{d}E = (\int \lambda \,\mathrm{d}E)^n = A^n$ for every $n \ge 0$, and likewise for E'; hence

$$\int_{\sigma(A)} p \,\mathrm{d}\nu_{x,y} = \langle x | p(A) y \rangle = \int_{\sigma(A)} p \,\mathrm{d}\nu'_{x,y}$$

for every polynomial p. The polynomials are dense in $C(\sigma(A))$ (Theorem A.242) and both measures are finite, so the two integrate every continuous function alike; by the uniqueness in Theorem A.244, $\nu_{x,y} = \nu'_{x,y}$. Taking Ω Borel and x, y arbitrary gives $\langle x | E(\Omega) y \rangle = \langle x | E'(\Omega) y \rangle$, i.e. $E = E'$. □

A.24.7 The multiplication form

Definition A.248 (Cyclic vector and cyclic subspace)**.** For $x \in \mathcal{H}$ the *cyclic subspace* generated by x is

$$\mathcal{H}_x = \overline{\{f(A)\,x \mid f \in C(\sigma(A))\}}\,, \tag{A.464}$$

a closed A-invariant subspace containing $x = 1(A)x$. The vector x is *cyclic* for A if $\mathcal{H}_x = \mathcal{H}$.

Rests on Proposition A.243 and Theorem 12.18.

Lemma A.249 (The cyclic case)**.** *Let $x \neq 0$ be cyclic for A. Then there is a unitary $U : \mathcal{H} \longrightarrow L^2(\sigma(A), \mu_x)$ with $Uf(A)x = f$ for every $f \in C(\sigma(A))$, and*

$$(UAU^{-1}h)(\lambda) = \lambda\, h(\lambda)\,, \qquad \forall\, h \in L^2(\sigma(A), \mu_x)\,. \tag{A.465}$$

Rests on Definition A.248, Proposition A.245, and Theorem A.244.

Proof. *Derives Lemma A.249.*

For $f \in C(\sigma(A))$, using Equation (A.455) and Equation (A.457),

$$\|f(A)x\|^2 = \Big\langle x \Big| \bar{f}(A)f(A)x \Big\rangle = \Big\langle x \Big| (|f|^2)(A)x \Big\rangle = \int_{\sigma(A)} |f|^2 \,\mathrm{d}\mu_x = \|f\|^2_{L^2(\mu_x)}\,, \tag{A.466}$$

so $f(A)x \longmapsto f$ is a well-defined linear isometry from the dense subspace $\{f(A)x\}$ of $\mathcal{H}$ onto the subspace $C(\sigma(A))$ of $L^2(\sigma(A), \mu_x)$ — well defined because $f(A)x = g(A)x$ forces $\|f - g\|_{L^2(\mu_x)} = 0$ by Equation (A.466). Its target is dense by the regularity statement of Theorem A.244, and both spaces are complete, so the isometry extends to a unitary U (a densely defined isometry with dense range extends to a surjective isometry by taking limits, and Equation (12.3) then gives preservation of the inner product). Finally, for f continuous, $UAf(A)x = U(\lambda f)(A)x = \lambda f = \lambda \cdot (Uf(A)x)$, which is Equation (A.465) on a dense subspace; both sides are bounded, so it holds everywhere. □

Lemma A.250 (Decomposition into cyclic subspaces)**.** *Let $\mathcal{H}$ be separable and A self-adjoint. Then there are finitely or countably many non-zero vectors x_n such that the cyclic subspaces $\mathcal{H}_{x_n}$ are mutually orthogonal and*

$$\mathcal{H} = \bigoplus_n \mathcal{H}_{x_n}\,. \tag{A.467}$$

Rests on Definitions 12.32 and A.248 and Proposition 12.87.

Proof. *Derives Lemma A.250.*

Let $\{y_j\}_{j\in\mathbb{N}}$ be dense in $\mathcal{H}$ (Definition 12.32). Construct the x_n recursively. Put $K_0 = \mathcal{H}$. Given mutually orthogonal cyclic subspaces $\mathcal{H}_{x_1}, \dots, \mathcal{H}_{x_{n-1}}$, let K_{n-1} be the orthogonal complement of their sum. Each $\mathcal{H}_{x_i}$ is A-invariant, and A is self-adjoint, so K_{n-1} is A-invariant too (Lemma A.231), and it is closed. If $K_{n-1} = \{0\}$, stop. Otherwise let $j(n)$ be the least index with $P_{K_{n-1}} y_j \neq 0$ — such an index exists, since otherwise every y_j would lie in $K_{n-1}^{\perp}$ and, by density and closedness, so would every vector, forcing $K_{n-1} = \{0\}$ — and set $x_n = P_{K_{n-1}} y_{j(n)}$, $\mathcal{H}_{x_n} \subseteq K_{n-1}$.

By construction $j(1) < j(2) < \cdots$, because $P_{K_{n-1}} y_j = 0$ for every $j < j(n)$ not already used means $y_j \in \mathcal{H}_{x_1} \oplus \cdots \oplus \mathcal{H}_{x_{n-1}}$, while $y_{j(n)} =$ (its part in that sum) $+ x_n$ also lies in $\mathcal{H}_{x_1} \oplus \cdots \oplus \mathcal{H}_{x_n}$. Hence every y_j lies in $\bigoplus_n \mathcal{H}_{x_n}$, whose closure is therefore all of $\mathcal{H}$; that closure is the internal direct sum of Proposition 12.87, which is Equation (A.467). If instead the recursion stops because $K_{n-1} = \{0\}$, the sum is already $\mathcal{H}$. □

Proof of Theorem A.238, part (2). *Derives Theorem A.238.*

Take the decomposition Equation (A.467) and write A_n for the restriction of A to $\mathcal{H}_{x_n}$, a bounded self-adjoint operator on that Hilbert space for which x_n is cyclic, with $\sigma(A_n) \subseteq \sigma(A)$

and with the same functional calculus (the calculus of A leaves $\mathcal{H}_{x_n}$ invariant). Lemma A.249 gives unitaries $U_n : \mathcal{H}_{x_n} \longrightarrow L^2(\sigma(A), \mu_n)$, $\mu_n = \mu_{x_n}$, carrying A into multiplication by λ.

Let M be the disjoint union of countably many copies of $\sigma(A)$, the nth copy written $\sigma(A) \times \{n\}$, and define on it the measure

$$\mu(\Omega \times \{n\}) = c_n \, \mu_n(\Omega) \,, \qquad c_n = \frac{2^{-n}}{\|x_n\|^2} \,, \tag{A.468}$$

so that, by Proposition A.245, $\mu(M) = \sum_n c_n \, \|x_n\|^2 = \sum_n 2^{-n} \leq 1 < \infty$. Define $a(\lambda, n) = \lambda$, a bounded real measurable function on M because $\sigma(A)$ is compact and real. The map

$$U\Big(\sum_n v_n\Big)\Big|_{\sigma(A)\times\{n\}} = c_n^{-1/2} \, U_n v_n \,, \qquad v_n \in \mathcal{H}_{x_n} \,, \tag{A.469}$$

is unitary from $\mathcal{H}$ onto $L^2(M, \mu)$: on the nth block it is an isometry from $\mathcal{H}_{x_n}$ onto $L^2(\sigma(A), c_n \mu_n)$, because $\int \left| c_n^{-1/2} h \right|^2 \, c_n \mathrm{d}\mu_n = \int |h|^2 \, \mathrm{d}\mu_n$, and $L^2(M, \mu)$ is by construction the orthogonal direct sum of those blocks (Proposition 12.85). Since multiplication by λ is unaffected by the constant $c_n^{-1/2}$, Equation (A.465) gives Equation (A.447), which completes the proof of Theorem A.238 and, with it, of Theorem 12.59. □

Remark A.251 (What is quoted here). Two classical theorems enter the proof and are not derived in this treatise; every other step above is carried out from Chapter 12 and its predecessors.

- Theorem A.242 (Stone–Weierstrass) is used twice: to extend the polynomial calculus to $C(\sigma(A))$, and in the uniqueness argument. Only the special case of a compact subset of $\mathbb{R}$, where it is the Weierstrass approximation theorem, is ever needed; that theorem is already declared as quoted in Remark 12.1 and is used there for Proposition 12.61.

- Theorem A.244 (Riesz–Markov) is the representation of a bounded functional on $C(K)$ by a regular complex Borel measure, together with the uniqueness of that measure and the density of $C(K)$ in $L^2(K, \nu)$. It is a theorem of measure theory, and this treatise develops no measure theory (Remark 12.1); it is what converts the operator inequality Equation (A.458) into a *measure*, and there is no route to a projection-valued measure that avoids it. Statement and proof: [1481], chapter IV.

Nothing else is assumed. In particular the polynomial identity Equation (A.453) — the step that makes the whole construction possible, and the one a reader should check first — rests only on the C^* identity Equation (12.25) and the numerical-radius formula Equation (12.26), both proved in the chapter.

Remark A.252. Appendix A.24 discharges the proof obligation of Theorem 12.59 (Section 12.4.2). The projection-valued measure it constructs is what Definition 12.60 integrates against, so the functional calculus $f(A)$ of the chapter — unique by Proposition 12.61 — now exists as well as being unique; Proposition A.246 is that existence statement for bounded Borel f. The theorem is used again in Appendix A.25, where the unbounded case is derived from it and the exponential $\exp(-\mathrm{i}Ht/\hbar)$ is defined by Equation (A.459), and in Appendix A.27, whose direct-integral decomposition is Lemma A.250 read fibrewise. Physically it is Remark 12.63: the support of E is the set of possible measured values, and $\|E(\Omega)\psi\|^2$ is the Born probability.

A.25 Stone's Theorem on One-Parameter Unitary Groups

This appendix proves the two halves of Theorem 12.66 (Chapter 12) that Proposition 12.67 left open: that the domain

$$D(H)=\left\{x\in\mathcal{H}\;\middle|\;\lim_{t\to0}\frac{U(t)x-x}{t}\ \text{exists in }\mathcal{H}\right\},\qquad Hx=\mathrm{i}\hbar\lim_{t\to0}\frac{U(t)x-x}{t},\tag{A.470}$$

is dense and that H is self-adjoint on it, not merely symmetric; and the converse, that every self-adjoint operator — bounded or not — exponentiates to a strongly continuous one-parameter unitary group. Together with Proposition 12.67 this establishes the bijection asserted by Theorem 12.66, which is what makes the two halves of the abstract Schrödinger equation Equation (12.48) equivalent statements [707].

Units are carried explicitly, as everywhere in this treatise. The group parameter t has the unit s, the generator H the unit J, and $\hbar$ the unit J s, so that $Ht/\hbar$ is a pure number; the energy scale $\varepsilon>0$ introduced in Appendix A.25.5 likewise has the unit J, and the spectral variable λ is an energy. The smoothing function φ of Appendix A.25.3 has the unit 1/s, so that $\int\varphi(t)\,\mathrm{d}t$ is dimensionless.

A.25.1 Statement

Theorem A.253 (Stone)**.** *Let $\{U(t)\}_{t\in\mathbb{R}}$ be a strongly continuous one-parameter unitary group on a Hilbert space $\mathcal{H}$ (Definition 12.64). Then the domain Equation (A.470) is dense and H is self-adjoint on it. Conversely, if H is any self-adjoint operator on $\mathcal{H}$ (Definition 12.72), then $U(t)=\exp(-\mathrm{i}Ht/\hbar)$, defined by the functional calculus of Proposition A.246 applied to the projection-valued measure of Proposition A.262, is a strongly continuous one-parameter unitary group whose generator in the sense of Equation (A.470) is exactly H. The two constructions are mutually inverse.*

Rests on Definition 12.64, Proposition 12.67, and Theorem A.238.

A.25.2 Vector-valued integrals

The smoothing argument needs to integrate a continuous $\mathcal{H}$-valued function. Nothing more than the Riemann integral is required, and it is built exactly as on the line.

Lemma A.254 (Riemann integral of a continuous curve)**.** *Let $G\colon[a,b]\longrightarrow\mathcal{H}$ be continuous. Then the Riemann sums of G over tagged partitions converge, as the mesh tends to 0, to a vector $\int_a^b G(t)\,\mathrm{d}t\in\mathcal{H}$, and*

$$\left\langle w\middle|\int_a^b G(t)\,\mathrm{d}t\right\rangle=\int_a^b\langle w|G(t)\rangle\,\mathrm{d}t\,,\qquad\left\|\int_a^b G(t)\,\mathrm{d}t\right\|\le\int_a^b\|G(t)\|\,\mathrm{d}t\,,\tag{A.471}$$

for every $w\in\mathcal{H}$. If $B\in\mathcal{B}(\mathcal{H})$ then $B\int_a^b G=\int_a^b BG$.

Rests on Definition 12.2 and Proposition 12.4.

Proof. *Derives Lemma A.254.*

G is uniformly continuous on the compact interval $[a,b]$: the Heine–Cantor argument of Theorem 7.25 uses only that the target is a metric space, and $\mathcal{H}$ is one. So given $\eta>0$ there is $\delta>0$ with $\|G(t)-G(s)\|<\eta$ whenever $|t-s|<\delta$. Two tagged partitions of mesh $<\delta$ have a common refinement, and comparing each with the refinement bounds the difference of their Riemann sums by $2\eta(b-a)$; the sums therefore form a Cauchy net, which converges because $\mathcal{H}$ is complete (Definition 12.2). Both identities in Equation (A.471) hold for Riemann sums — the first by linearity of the inner product in its second slot, the second by the triangle inequality — and pass to the limit, the first by continuity of the inner product (Proposition 12.4). The same argument gives $B\int G=\int BG$ for a bounded B. □

A.25.3 The domain is dense

Lemma A.255 (Smoothed vectors lie in the domain)**.** *Let U be a strongly continuous one-parameter unitary group, let $\varphi\colon \mathbb{R} \longrightarrow \mathbb{R}$ be continuously differentiable with support in a compact interval, and for $x \in \mathcal{H}$ put*

$$x_\varphi = \int_{\mathbb{R}} \varphi(t)\, U(t)x \,\mathrm{d}t\,. \tag{A.472}$$

Then $x_\varphi \in D(H)$ and

$$H\, x_\varphi = -\mathrm{i}\hbar \int_{\mathbb{R}} \varphi'(t)\, U(t)x \,\mathrm{d}t = -\mathrm{i}\hbar\, x_{\varphi'}\,. \tag{A.473}$$

Rests on Lemma A.254, Definition 12.64, and Equation (A.470).

Proof. *Derives Lemma A.255.*

The integrand is continuous, by strong continuity of U and continuity of φ, and vanishes outside a compact interval, so Lemma A.254 defines x_φ. Using $U(s)U(t) = U(t+s)$ and the substitution $t \longmapsto t - s$ — a translation, which the Riemann integral of Lemma A.254 respects because it acts on tagged partitions by translation —

$$U(s)\, x_\varphi = \int_{\mathbb{R}} \varphi(t)\, U(t+s)x \,\mathrm{d}t = \int_{\mathbb{R}} \varphi(t-s)\, U(t)x \,\mathrm{d}t\,. \tag{A.474}$$

Hence, for $s \neq 0$,

$$\frac{U(s)x_\varphi - x_\varphi}{s} = \int_{\mathbb{R}} \frac{\varphi(t-s) - \varphi(t)}{s}\, U(t)x \,\mathrm{d}t\,. \tag{A.475}$$

By the mean value theorem, $[\varphi(t-s)-\varphi(t)]/s = -\varphi'(t-\theta s)$ for some $\theta = \theta(t,s) \in (0,1)$, and φ' is continuous with compact support, hence uniformly continuous; so $[\varphi(\cdot - s) - \varphi]/s \longrightarrow -\varphi'$ uniformly as $s \to 0$, with all supports inside one fixed compact interval $[a,b]$ of length ℓ once $|s| \leq 1$. The second estimate of Equation (A.471) and $\|U(t)x\| = \|x\|$ then give

$$\left\| \frac{U(s)x_\varphi - x_\varphi}{s} + \int_{\mathbb{R}} \varphi'(t) U(t)x \,\mathrm{d}t \right\| \leq \ell\, \|x\| \sup_t \left| \frac{\varphi(t-s) - \varphi(t)}{s} + \varphi'(t) \right| \longrightarrow 0\,.$$

So the limit in Equation (A.470) exists and equals $-x_{\varphi'}$; multiplying by $\mathrm{i}\hbar$ gives Equation (A.473). □

Proposition A.256 (Density)**.** *$D(H)$ is dense in $\mathcal{H}$.*

Rests on Lemma A.255 and Definition 12.64.

Proof. *Derives Proposition A.256.*

For $n \in \mathbb{N}$ define

$$\varphi_n(t) = \frac{15\, n}{16} \left(1 - n^2 t^2\right)^2 \quad \text{for } |t| \leq \frac{1}{n}\,, \qquad \varphi_n(t) = 0 \ \text{ otherwise}\,. \tag{A.476}$$

Each φ_n is continuously differentiable — at $t = \pm 1/n$ both φ_n and φ_n' vanish, so the pieces match — non-negative, supported in $[-1/n, 1/n]$, and normalised: substituting $u = nt$,

$$\int_{\mathbb{R}} \varphi_n(t) \,\mathrm{d}t = \frac{15}{16} \int_{-1}^{1} \left(1 - u^2\right)^2 \mathrm{d}u = \frac{15}{16} \cdot \frac{16}{15} = 1\,,$$

a pure number, as it must be: φ_n carries the unit $1/\mathrm{s}$ and $\mathrm{d}t$ the unit s.

Fix $x \in \mathcal{H}$. Since $\int \varphi_n = 1$,

$$\|x_{\varphi_n} - x\| = \left\| \int_{\mathbb{R}} \varphi_n(t) \big(U(t)x - x\big) \mathrm{d}t \right\| \leq \sup_{|t| \leq 1/n} \|U(t)x - x\|\,,$$

by Equation (A.471) and non-negativity of φ_n. The right-hand side tends to 0 as $n \to \infty$, by strong continuity of U at $t = 0$ (Definition 12.64). Each x_{φ_n} lies in $D(H)$ by Lemma A.255, so x is a limit of vectors of $D(H)$. □

A.25.4 The generator is closed and self-adjoint

Lemma A.257 (Integrated form of the equation of motion)**.** *For $x \in D(H)$ and every t,*

$$U(t)x - x = -\frac{\mathrm{i}}{\hbar}\int_0^t U(r)\,Hx\,\mathrm{d}r\,. \tag{A.477}$$

Rests on Proposition 12.67 and Lemma A.254.

Proof. *Derives Lemma A.257.*

Fix $w \in \mathcal{H}$ and let $g(r) = \langle w|U(r)x\rangle$. By Equation (12.46) the curve $r \longmapsto U(r)x$ is differentiable in norm with derivative $-\mathrm{i}\hbar^{-1}U(r)Hx$, so g is differentiable with $g'(r) = -\mathrm{i}\hbar^{-1}\langle w|U(r)Hx\rangle$, which is continuous in r. The fundamental theorem of calculus (Theorem 7.43) and the first identity of Equation (A.471) give

$$\langle w|U(t)x - x\rangle = g(t) - g(0) = \int_0^t g'(r)\,\mathrm{d}r = \left\langle w\middle| -\frac{\mathrm{i}}{\hbar}\int_0^t U(r)Hx\,\mathrm{d}r\right\rangle .$$

As w is arbitrary, Equation (5.31) gives Equation (A.477). □

Proposition A.258 (The generator is closed)**.** *H is a closed operator (Definition 12.70).*
Rests on Lemma A.257 and Equation (A.470).

Proof. *Derives Proposition A.258.*

Let $x_n \in D(H)$ with $x_n \longrightarrow x$ and $Hx_n \longrightarrow z$. Each term of Equation (A.477) converges: the left side to $U(t)x - x$, since $U(t)$ is bounded; and the right side to $-\mathrm{i}\hbar^{-1}\int_0^t U(r)z\,\mathrm{d}r$, because

$$\left\|\int_0^t U(r)(Hx_n - z)\mathrm{d}r\right\| \le |t|\;\|Hx_n - z\| \longrightarrow 0\,.$$

Hence $U(t)x - x = -\mathrm{i}\hbar^{-1}\int_0^t U(r)z\,\mathrm{d}r$ for every t. Dividing by $t \neq 0$ and using continuity of $r \longmapsto U(r)z$ at $r = 0$,

$$\left\|\frac{1}{t}\int_0^t U(r)z\,\mathrm{d}r - z\right\| \le \sup_{|r|\le|t|}\|U(r)z - z\| \longrightarrow 0\,,$$

so the limit Equation (A.470) exists for x and equals $-\mathrm{i}\hbar^{-1}z$ before multiplication by $\mathrm{i}\hbar$. That is, $x \in D(H)$ and $Hx = z$: the graph is closed. □

Proposition A.259 (The generator is self-adjoint)**.** *H is self-adjoint on $D(H)$.*
Rests on Propositions 12.67, A.256 and A.258.

Proof. *Derives Proposition A.259.*

H is densely defined (Proposition A.256) and symmetric (Equation (12.45)), so $H \subseteq H^\dagger$ and the adjoint exists. Let $\varepsilon > 0$ be an energy.

Step 1: $\ker(H^\dagger \mp \mathrm{i}\varepsilon) = \{0\}$. Let $y \in D(H^\dagger)$ with $H^\dagger y = \mathrm{i}\varepsilon y$, fix $x \in D(H)$, and put $f(t) = \langle y|U(t)x\rangle$. By Proposition 12.67, $U(t)x \in D(H)$ and the curve is differentiable, so

$$f'(t) = \left\langle y\middle| -\frac{\mathrm{i}}{\hbar}H\,U(t)x\right\rangle = -\frac{\mathrm{i}}{\hbar}\left\langle H^\dagger y\middle|U(t)x\right\rangle = -\frac{\mathrm{i}}{\hbar}\langle \mathrm{i}\varepsilon y|U(t)x\rangle = -\frac{\varepsilon}{\hbar}f(t)\,, \tag{A.478}$$

where the second equality is Equation (12.50) and the third uses the antilinearity Equation (5.36), which turns $\mathrm{i}\varepsilon$ in the first slot into $-\mathrm{i}\varepsilon$ and cancels one factor of i against the prefactor. Hence $f(t) = f(0)\,\mathrm{e}^{-\varepsilon t/\hbar}$. But $|f(t)| \le \|y\|\,\|U(t)x\| = \|y\|\,\|x\|$ for all $t \in \mathbb{R}$, while $\mathrm{e}^{-\varepsilon t/\hbar} \longrightarrow \infty$ as $t \longrightarrow -\infty$; so $f(0) = \langle y|x\rangle = 0$. As x ranges over the dense set $D(H)$, $y = 0$. The case $H^\dagger y = -\mathrm{i}\varepsilon y$ is the same with $t \to +\infty$.

Step 2: $\mathrm{im}(H \pm \mathrm{i}\varepsilon) = \mathcal{H}$. For $x \in D(H)$, symmetry makes the cross terms cancel:

$$\|(H \pm \mathrm{i}\varepsilon)x\|^2 = \|Hx\|^2 + \varepsilon^2\,\|x\|^2 \ge \varepsilon^2\,\|x\|^2\,. \tag{A.479}$$

If $(H+\mathrm{i}\varepsilon)x_n \longrightarrow u$ then, by Equation (A.479) applied to differences, both (x_n) and (Hx_n) are Cauchy; H is closed (Proposition A.258), so $x_n \longrightarrow x \in D(H)$ with $(H+\mathrm{i}\varepsilon)x = u$. The image is therefore closed. It is also dense: exactly as in Equation (12.53), $y \perp \mathrm{im}(H+\mathrm{i}\varepsilon)$ means $\langle y|Hx\rangle = \langle \mathrm{i}\varepsilon y|x\rangle$ for all $x \in D(H)$, i.e. $y \in \ker(H^\dagger - \mathrm{i}\varepsilon) = \{0\}$ by Step 1. A dense closed subspace is everything, and likewise for $H - \mathrm{i}\varepsilon$.

Step 3. Let $y \in D(H^\dagger)$. By Step 2 there is $x \in D(H)$ with $(H-\mathrm{i}\varepsilon)x = (H^\dagger - \mathrm{i}\varepsilon)y$. Since $H \subseteq H^\dagger$, this reads $(H^\dagger - \mathrm{i}\varepsilon)(y-x) = 0$, so $y - x = 0$ by Step 1 and $y = x \in D(H)$. Hence $D(H^\dagger) = D(H)$ and $H = H^\dagger$. □

A.25.5 The converse: exponentiating a self-adjoint operator

The converse needs a projection-valued measure for an operator that is in general unbounded, and Theorem A.238 supplies one only for bounded operators. The gap is closed by the Cayley transform: the unbounded H is traded for a *unitary* operator V, for which the construction of Appendix A.24 goes through with three changes displayed below, and the measure is then transported back along an explicit Möbius map.

Lemma A.260 (Cayley transform of a self-adjoint operator)**.** *Let H be self-adjoint and let $\varepsilon > 0$ be an energy. Then $H \pm \mathrm{i}\varepsilon$ maps $D(H)$ bijectively onto $\mathcal{H}$, the operator*

$$V = (H - \mathrm{i}\varepsilon)(H + \mathrm{i}\varepsilon)^{-1} \tag{A.480}$$

is unitary, $\ker(\mathbb{1} - V) = \{0\}$*, and*

$$(H+\mathrm{i}\varepsilon)^{-1} = \frac{\mathbb{1} - V}{2\mathrm{i}\varepsilon}\,, \qquad D(H) = \mathrm{im}\,(\mathbb{1} - V)\ . \tag{A.481}$$

Rests on Propositions 12.73 and A.259 and Definition 12.72.

Proof. *Derives Lemma A.260.*

The identity Equation (A.479) holds for any symmetric H, so $H \pm \mathrm{i}\varepsilon$ is injective with $\|(H \pm \mathrm{i}\varepsilon)x\| \geq \varepsilon\,\|x\|$; and H is closed (Proposition 12.73), so, exactly as in Step 2 of Proposition A.259, the image is closed. It is dense because $\big(\mathrm{im}(H \pm \mathrm{i}\varepsilon)\big)^\perp = \ker(H^\dagger \mp \mathrm{i}\varepsilon) = \ker(H \mp \mathrm{i}\varepsilon) = \{0\}$, the middle equality being $H = H^\dagger$ and the last one Equation (A.479) again. Hence $\mathrm{im}(H \pm \mathrm{i}\varepsilon) = \mathcal{H}$ and $H \pm \mathrm{i}\varepsilon$ is a bijection of $D(H)$ onto $\mathcal{H}$. Thus V is defined on all of $\mathcal{H}$, and by Equation (A.479)

$$\|V(H+\mathrm{i}\varepsilon)x\| = \|(H-\mathrm{i}\varepsilon)x\| = \|(H+\mathrm{i}\varepsilon)x\|\ ,$$

so V is isometric; its image is $\mathrm{im}(H - \mathrm{i}\varepsilon) = \mathcal{H}$, so it is unitary. Writing $u = (H+\mathrm{i}\varepsilon)x$,

$$(\mathbb{1} - V)u = (H+\mathrm{i}\varepsilon)x - (H-\mathrm{i}\varepsilon)x = 2\mathrm{i}\varepsilon\, x\,, \tag{A.482}$$

which is Equation (A.481) — both the formula for the inverse and, since x runs over $D(H)$ as u runs over $\mathcal{H}$, the identification of $\mathrm{im}(\mathbb{1} - V)$ with $D(H)$. If $Vu = u$ then Equation (A.482) gives $x = 0$, hence $u = (H+\mathrm{i}\varepsilon)x = 0$. □

Proposition A.261 (Spectral theorem for a unitary operator)**.** *Let $V \in \mathcal{B}(\mathcal{H})$ be unitary and let $\mathbb{T} = \{z \in \mathbb{C} \mid |z| = 1\}$. Then $\sigma(V) \subseteq \mathbb{T}$, and there is a unique projection-valued measure F on $\mathbb{T}$, supported on $\sigma(V)$, with $V = \int z\,\mathrm{d}F(z)$; the associated bounded Borel calculus $g \longmapsto g(V)$ has all the properties of Proposition A.246.*

Rests on Theorem A.238 and Propositions 12.52 and A.246.

Proof. *Derives Proposition A.261.*

The construction of Appendix A.24 is repeated with Laurent polynomials $q(z) = \sum_{k=-m}^{m} c_k z^k$ in place of polynomials, $q(V) = \sum_k c_k V^k$ with $V^{-1} = V^\dagger$. Three steps need a

new argument; everything after them — Propositions A.245 and A.246 and the construction and uniqueness of the projection-valued measure in Appendix A.24.6 — uses only that Φ is an isometric, multiplicative, conjugation-preserving map on $C(K)$ for a compact $K \subseteq \mathbb{C}$, and applies word for word.

(a) The spectrum lies on the circle. $\|V\| = 1$, so $\sigma(V)$ lies in the closed unit disc (Proposition 12.52). For $|z| < 1$ write $V - z\mathbb{1} = V(\mathbb{1} - zV^{-1})$; since $\|zV^{-1}\| = |z| < 1$ the bracket is invertible by the geometric series of Proposition 12.52 and V is invertible, so $z \in \rho(V)$.

(b) Spectral mapping. Fix $\nu \in \mathbb{C}$ and put $r(z) = z^m(q(z) - \nu)$, an ordinary polynomial. Then $r(V) = V^m(q(V) - \nu\mathbb{1})$ with V^m invertible, so $q(V) - \nu\mathbb{1}$ is invertible exactly when $r(V)$ is, i.e. exactly when $0 \notin \sigma(r(V)) = r(\sigma(V))$ (Lemma A.240). Since $z \neq 0$ on $\sigma(V) \subseteq \mathbb{T}$, $r(z) = 0$ there means $q(z) = \nu$; hence $\sigma(q(V)) = q(\sigma(V))$.

(c) The norm identity. With $\bar{q}(z) = \sum_k c_k^* z^{-k}$ one has $q(V)^\dagger = \bar{q}(V)$, and on $\mathbb{T}$, where $z^{-1} = z^*$, $\bar{q}(z) = q(z)^*$. So $(\bar{q}\,q)(V)$ is self-adjoint and, by Equation (12.25), Lemma A.239 and (b),

$$\|q(V)\|^2 = \|(\bar{q}\,q)(V)\| = \sup_{z \in \sigma(V)} |q(z)|^2 \,. \tag{A.483}$$

Laurent polynomials restricted to the compact set $\sigma(V)$ form a subalgebra of $C(\sigma(V))$ that contains the constants, is closed under conjugation (by (c)) and separates points (the function z does), so Theorem A.242 makes it dense; with Equation (A.483) this yields the isometric continuous calculus, and the rest is Appendix A.24 unchanged. □

Proposition A.262 (Spectral theorem for an unbounded self-adjoint operator)**.** *Let H be self-adjoint. There is a projection-valued measure E on $\mathbb{R}$ such that, writing $\mu_x(\Omega) = \langle x|E(\Omega)x\rangle$,*

$$D(H) = \left\{ x \in \mathcal{H} \;\middle|\; \int_{\mathbb{R}} \lambda^2 \,\mathrm{d}\mu_x(\lambda) < \infty \right\}, \qquad \langle y|Hx\rangle = \int_{\mathbb{R}} \lambda \,\langle y|\mathrm{d}E(\lambda)x\rangle \tag{A.484}$$

for $x \in D(H)$, $y \in \mathcal{H}$. For bounded Borel g the operator $g(H) = \int g\,\mathrm{d}E$ is bounded with $\|g(H)\| \leq \|g\|_\infty$, and $g \longmapsto g(H)$ is linear, multiplicative and $*$-preserving.

Rests on Lemma A.260 and Propositions A.246 and A.261.

Proof. *Derives Proposition A.262.*

Let V be the Cayley transform Equation (A.480) and F its projection-valued measure (Proposition A.261). Write

$$w(z) = \frac{1-z}{2\mathrm{i}\varepsilon}, \qquad h(z) = \mathrm{i}\varepsilon\,\frac{1+z}{1-z} \quad (z \neq 1), \tag{A.485}$$

so that $(H + \mathrm{i}\varepsilon)^{-1} = w(V)$ by Equation (A.481).

h is real, and $F(\{1\}) = 0$. For $|z| = 1$, $z \neq 1$, multiply numerator and denominator by $1 - z^*$:

$$\frac{1+z}{1-z} = \frac{(1+z)(1-z^*)}{|1-z|^2} = \frac{z - z^*}{|1-z|^2} = \frac{2\mathrm{i}\,\mathrm{Im}\,z}{|1-z|^2}, \tag{A.486}$$

so $h(z) = -2\varepsilon\,\mathrm{Im}\,z/\,|1-z|^2 \in \mathbb{R}$, with the unit J. Since $w\chi_{\{1\}} = 0$ identically, multiplicativity gives $w(V)F(\{1\}) = 0$; but $w(V)$ is injective, so $F(\{1\}) = 0$.

The measure. Define $E(\Omega) = F(h^{-1}(\Omega))$ for Borel $\Omega \subseteq \mathbb{R}$. Preimages respect complements, intersections and countable unions, so E inherits every clause of Definition 12.58 from F, with $E(\mathbb{R}) = F(\mathbb{T} \setminus \{1\}) = \mathbb{1}$. By construction μ_x is the image of μ_x^F under h, so $\int (g \circ h)\,\mathrm{d}\mu_x^F = \int g\,\mathrm{d}\mu_x$ for every Borel $g \geq 0$ or bounded, and $g(H) := (g \circ h)(V)$ is the calculus claimed.

The domain. A direct computation on $\mathbb{T}$, using $|1 \pm z|^2 = 2 \pm 2\operatorname{Re} z$, gives the identity that makes the transport work:

$$|w(z)|^{-2} = \frac{4\varepsilon^2}{2 - 2\operatorname{Re} z} = \varepsilon^2 \frac{2 + 2\operatorname{Re} z}{2 - 2\operatorname{Re} z} + \varepsilon^2 = |h(z)|^2 + \varepsilon^2 \,. \tag{A.487}$$

Now $D(H) = \operatorname{im} w(V)$ by Equation (A.481), and

$$\operatorname{im} w(V) = \left\{ x \;\middle|\; \int_{\mathbb{T}} |w|^{-2} \, \mathrm{d}\mu_x^F < \infty \right\} . \tag{A.488}$$

Indeed, if $x = w(V)u$ then $\mu_x^F = |w|^2 \, \mu_u^F$ — because $\langle x | F(\Omega) x \rangle = \langle u | (\bar{w} \chi_\Omega w)(V) u \rangle$ — so $\int |w|^{-2} \, \mathrm{d}\mu_x^F = \|u\|^2 < \infty$. Conversely, if the integral is finite put $g_n = w^{-1} \chi_{\{|w| \geq 1/n\}}$, a bounded Borel function, and $u_n = g_n(V)x$; then $\|u_n - u_m\|^2 = \int |g_n - g_m|^2 \, \mathrm{d}\mu_x^F \longrightarrow 0$ by dominated convergence (Remark 12.1), the dominating function being $|w|^{-2}$, and $w(V)u_n = \chi_{\{|w| \geq 1/n\}}(V)x \longrightarrow x$ because $F(\{w = 0\}) = F(\{1\}) = 0$. Hence $x \in \operatorname{im} w(V)$. Combining Equations (A.487) and (A.488) with $\mu_x^F(\mathbb{T}) = \|x\|^2$,

$$\int_{\mathbb{T}} |w|^{-2} \, \mathrm{d}\mu_x^F = \int_{\mathbb{R}} \lambda^2 \, \mathrm{d}\mu_x(\lambda) + \varepsilon^2 \, \|x\|^2 \,,$$

so the two finiteness conditions agree and the first half of Equation (A.484) holds.

The action. Let $x = w(V)u \in D(H)$. Then, by Equation (A.481),

$$Hx = (H + \mathrm{i}\varepsilon)x - \mathrm{i}\varepsilon x = u - \mathrm{i}\varepsilon \, \frac{\mathbb{1} - V}{2\mathrm{i}\varepsilon} \, u = \frac{\mathbb{1} + V}{2} \, u \,. \tag{A.489}$$

On the other side, put $h_n = h \, \chi_{\{|h| \leq n\}}$, a bounded Borel function, so that $h_n(H) = \int_{|\lambda| \leq n} \lambda \, \mathrm{d}E$. Since $h_n w \longrightarrow hw$ pointwise on $\mathbb{T} \setminus \{1\}$ with $|h_n w| \leq |hw| = |1 + z| \, / 2 \leq 1$, dominated convergence gives

$$h_n(H) \, x = (h_n w)(V) \, u \longrightarrow (hw)(V) \, u = \frac{\mathbb{1} + V}{2} \, u = Hx \,,$$

because $h(z)w(z) = (1 + z)/2$ by Equation (A.485). Pairing with y and passing to the limit gives the second half of Equation (A.484); the integral converges absolutely because $\int |\lambda| \; \mathrm{d} \, |\mu_{y,x}| \leq \|y\| \, \big(\int \lambda^2 \mathrm{d}\mu_x \big)^{1/2}$, by Cauchy–Schwarz on the partition sums. □

Proof of Theorem A.253, converse half. *Derives Theorem A.253.*

Let H be self-adjoint, E its projection-valued measure (Proposition A.262), and for $t \in \mathbb{R}$ put

$$U(t) = \exp\left(-\frac{\mathrm{i}Ht}{\hbar} \right) := f_t(H) \,, \qquad f_t(\lambda) = \mathrm{e}^{-\mathrm{i}\lambda t/\hbar} \,, \tag{A.490}$$

which is legitimate because $|f_t| = 1$: the exponent $\lambda t/\hbar$ is dimensionless, λ being an energy.

Unitary, and a group. $f_t \bar{f}_t = 1$ and $f_t f_s = f_{t+s}$, $f_0 = 1$, so multiplicativity and the $*$-property of the calculus give $U(t)^\dagger U(t) = U(t) U(t)^\dagger = \mathbb{1}$ and Equation (12.41).

Strong continuity. For $x \in \mathcal{H}$,

$$\|U(t)x - x\|^2 = \int_{\mathbb{R}} \left| \mathrm{e}^{-\mathrm{i}\lambda t/\hbar} - 1 \right|^2 \mathrm{d}\mu_x(\lambda) \longrightarrow 0 \qquad (t \to 0) \,, \tag{A.491}$$

by dominated convergence (Remark 12.1): the integrand tends to 0 pointwise and is bounded by the constant 4, which is μ_x-integrable because $\mu_x(\mathbb{R}) = \|x\|^2$.

The generator is H. Let $x \in D(H)$. Then

$$\left\| \frac{U(t)x - x}{t} + \frac{\mathrm{i}}{\hbar} Hx \right\|^2 = \int_{\mathbb{R}} \left| \frac{\mathrm{e}^{-\mathrm{i}\lambda t/\hbar} - 1}{t} + \frac{\mathrm{i}\lambda}{\hbar} \right|^2 \mathrm{d}\mu_x(\lambda) \,, \tag{A.492}$$

whose integrand tends to 0 pointwise and is bounded by $4\lambda^2/\hbar^2$, since $\left|e^{i\theta}-1\right| \le |\theta|$ for real θ; that bound is μ_x-integrable exactly because $x \in D(H)$ (Equation (A.484)). Dominated convergence makes Equation (A.492) tend to 0, so x lies in the domain Equation (A.470) of the generator G of U, with $Gx = Hx$. Thus $H \subseteq G$. But G is self-adjoint by Proposition A.259 and H is self-adjoint by hypothesis, so

$$G = G^\dagger \subseteq H^\dagger = H \subseteq G\,,$$

the middle inclusion because taking adjoints reverses inclusions. Hence $G = H$: a self-adjoint operator has no proper symmetric extension. □

Proposition A.263 (The two constructions are inverse)**.** *If U and W are strongly continuous one-parameter unitary groups with the same generator H, then $U = W$. Consequently $H \longmapsto \exp(-iHt/\hbar)$ and $U \longmapsto H$ are mutually inverse bijections between the self-adjoint operators on $\mathcal{H}$ and the strongly continuous one-parameter unitary groups on $\mathcal{H}$.*

Rests on Propositions 12.67 and A.259 and Theorem A.253.

Proof. *Derives Proposition A.263.*

Fix $x \in D(H)$ and t, and set $g(s) = W(t-s)U(s)x$ for $s \in [0,t]$. By Proposition 12.67, $U(s)x \in D(H)$ and both curves are differentiable, so the product rule — legitimate because each factor is differentiable in norm and W is isometric — gives

$$\frac{dg}{ds} = W(t-s)\left(-\frac{i}{\hbar}H\right)U(s)x + \left(+\frac{i}{\hbar}H\right)W(t-s)\,U(s)x = 0\,,$$

since H commutes with $W(t-s)$ on $D(H)$ (again Proposition 12.67). Hence g is constant: $W(t)x = g(0) = g(t) = U(t)x$. As $D(H)$ is dense (Proposition A.256) and both operators are bounded, $U(t) = W(t)$.

For the bijection: Propositions A.256 and A.259 show that every group has a self-adjoint generator, and Theorem A.253 that every self-adjoint operator is the generator of the group Equation (A.490); the paragraph just proved shows that no group is the exponential of two different self-adjoint operators, since the generator is recovered from the group by Equation (A.470). □

Remark A.264 (What is quoted here)*.* The direct half — Propositions A.256, A.258 and A.259 — quotes nothing at all: it uses the Riemann integral of Lemma A.254, built here, and the scalar fundamental theorem of calculus (Theorem 7.43).

The converse inherits the two quoted inputs of Appendix A.24 — Stone–Weierstrass (Theorem A.242) and Riesz–Markov (Theorem A.244) — through Proposition A.261, and in addition uses the dominated convergence theorem of Lebesgue integration three times: in Equation (A.488), in Equation (A.491) and in Equation (A.492). That theorem is already declared as quoted in Remark 12.1, where the same declaration covers the completeness of L^2. The passage from the bounded spectral theorem to the unbounded one is *not* quoted: it is carried out in Lemma A.260 and Propositions A.261 and A.262, and the algebra that makes it work — $(\mathbb{1}-V)/(2i\varepsilon) = (H+i\varepsilon)^{-1}$ and $|w|^{-2} = |h|^2 + \varepsilon^2$ — is displayed in full.

Remark A.265*.* Appendix A.25 discharges the proof obligation of Theorem 12.66 (Section 12.4.3), of which Proposition 12.67 had proved the elementary part. What the theorem buys the physics is stated in Remark 12.68: the equivalence of the differential and the integrated Schrödinger equation Equation (12.48), and the reason a Hamiltonian must be self-adjoint rather than merely symmetric — symmetry gives no unitary group, hence no conserved probability. That is not a technicality, and Example 12.82 is the proof: there the generator has no self-adjoint extension at all, so by Proposition A.263 there is no unitary dynamics to be had, a point settled in general by Theorem 12.80 and Appendix A.26, where the Cayley transform used above is developed for an arbitrary closed symmetric operator.

A.26 Von Neumann's Criterion for Self-Adjoint Extensions

This appendix proves Theorem 12.80 of Chapter 12: a closed symmetric operator is self-adjoint exactly when both deficiency indices vanish; it admits self-adjoint extensions exactly when the two indices are equal; and the extensions are then in bijective correspondence with the unitary maps from one deficiency subspace to the other. The proof is von Neumann's [664], through the Cayley transform, which converts a closed symmetric operator into an isometry between closed subspaces and a self-adjoint operator into a unitary — so that the extension problem for operators becomes the extension problem for isometries, where it is a matter of counting dimensions. The two examples the criterion was made for, momentum on an interval and momentum on a half-line, are recovered in Appendix A.26.6 as corollaries, with the one-parameter family of boundary conditions Equation (12.55) produced explicitly from the unitary parameter.

Throughout, A is a densely defined closed symmetric operator (Definitions 12.70 and 12.72) and $\mu > 0$ is a fixed number carrying the same SI dimension as A, so that $A \pm \mathrm{i}\mu$ is dimensionally consistent, as in Definition 12.79; for the momentum operator of Appendix A.26.6, μ is a momentum in $\mathrm{kg\,m/s}$. The deficiency subspaces and indices are those of Equation (12.53),

$$K_{\pm} = \ker\left(A^{\dagger} \mp \mathrm{i}\mu\right) = \left(\mathrm{im}(A \pm \mathrm{i}\mu)\right)^{\perp}, \qquad n_{\pm} = \dim K_{\pm}\,, \tag{A.493}$$

where dim means the cardinality of an orthonormal basis.

A.26.1 Statement

Theorem A.266 (von Neumann)**.** *Let A be a closed symmetric operator with deficiency indices (n_+, n_-). Then*

1. *A is self-adjoint if and only if $n_+ = n_- = 0$;*
2. *the self-adjoint extensions of A are in bijective correspondence with the unitary maps $W : K_+ \longrightarrow K_-$, the extension A_W attached to W being*

$$D(A_W) = D(A) \oplus \{k - Wk \mid k \in K_+\}\,, \qquad A_W(x + k - Wk) = Ax + \mathrm{i}\mu\,(k + Wk)\,, \tag{A.494}$$

   the sum of subspaces being direct; such maps exist if and only if $n_+ = n_-$, and when $n_+ = n_- = n < \infty$ they form a family parametrised by $\mathrm{U}(n)$;
3. *if $n_+ \neq n_-$ there is no self-adjoint extension.*

Rests on Definitions 12.72, 12.78 and 12.79.

A.26.2 The basic identity

Lemma A.267 (Isometry of $A \pm \mathrm{i}\mu$, and closed range)**.** *For every $x \in D(A)$,*

$$\|(A \pm \mathrm{i}\mu)x\|^2 = \|Ax\|^2 + \mu^2\,\|x\|^2\,, \tag{A.495}$$

so $A \pm \mathrm{i}\mu$ is injective with $\|(A \pm \mathrm{i}\mu)x\| \geq \mu\,\|x\|$, and $\|(A + \mathrm{i}\mu)x\| = \|(A - \mathrm{i}\mu)x\|$. If A is closed, the subspaces $\mathrm{im}(A \pm \mathrm{i}\mu)$ are closed, so that

$$\mathcal{H} = \mathrm{im}(A + \mathrm{i}\mu) \oplus K_+ = \mathrm{im}(A - \mathrm{i}\mu) \oplus K_-\,. \tag{A.496}$$

Rests on Definitions 12.72 and 12.79 and Theorem 12.18.

Proof. *Derives Lemma A.267.*

Expanding and using $\langle Ax|x\rangle = \langle x|Ax\rangle$, which is symmetry, the two cross terms cancel:

$$\|(A \pm \mathrm{i}\mu)x\|^2 = \|Ax\|^2 \pm \mathrm{i}\mu \langle Ax|x\rangle \mp \mathrm{i}\mu \langle x|Ax\rangle + \mu^2 \|x\|^2 = \|Ax\|^2 + \mu^2 \|x\|^2 ,$$

the sign of the third term coming from the antilinearity Equation (5.36) of the first slot. Both consequences are immediate. For closedness of the image, let $(A + \mathrm{i}\mu)x_n \longrightarrow u$. Applying Equation (A.495) to differences shows that both (x_n) and (Ax_n) are Cauchy, hence convergent, say $x_n \longrightarrow x$ and $Ax_n \longrightarrow y$; since A is closed, $x \in D(A)$ and $Ax = y$, so $u = (A + \mathrm{i}\mu)x$ lies in the image. Equation (A.496) is then the projection theorem Theorem 12.18 together with Equation (A.493). □

Lemma A.268 (The indices do not depend on μ)**.** *For a closed symmetric A the numbers $n_\pm$ are the same for every $\mu > 0$.*

Rests on Lemma A.267 and Definition 12.79.

Proof. *Derives Lemma A.268.*

Write $K_+(\mu) = \ker(A^\dagger - \mathrm{i}\mu)$ and suppose $\dim K_+(\nu) < \dim K_+(\mu)$ for two positive μ, ν of the same dimension. Then there is a non-zero $u \in K_+(\mu) \cap K_+(\nu)^\perp$: otherwise the orthogonal projection onto $K_+(\nu)$ would be injective on $K_+(\mu)$, and its adjoint — the projection onto $K_+(\mu)$ restricted to $K_+(\nu)$ — would have dense range in $K_+(\mu)$, forcing $\dim K_+(\mu) \leq \dim K_+(\nu)$.

By Equations (A.493) and (A.496), $K_+(\nu)^\perp = \mathrm{im}(A + \mathrm{i}\nu)$, so $u = (A + \mathrm{i}\nu)x$ for some $x \in D(A)$; and $A^\dagger u = \mathrm{i}\mu u$. Hence, using Equation (12.50) and Equation (5.36),

$$\|u\|^2 = \langle u|(A + \mathrm{i}\nu)x\rangle = \left\langle A^\dagger u \middle| x \right\rangle + \mathrm{i}\nu \langle u|x\rangle = \mathrm{i}\,(\nu - \mu)\,\langle u|x\rangle \ . \tag{A.497}$$

But Equation (A.495) gives $\|u\| \geq \nu \|x\|$, so $|\langle u|x\rangle| \leq \|u\| \|x\| \leq \|u\|^2 / \nu$ by Equation (12.1), and Equation (A.497) forces $\|u\|^2 \leq |\nu - \mu| \ \|u\|^2 / \nu$, i.e. $|\nu - \mu| \geq \nu$. So whenever $|\nu - \mu| < \nu$ no such u exists and $\dim K_+(\mu) \leq \dim K_+(\nu)$; exchanging the roles of μ and ν, the two dimensions are equal whenever $|\nu - \mu| < \min(\mu, \nu)$. The function $\mu \longmapsto n_+(\mu)$ is therefore locally constant on the connected set $(0, \infty)$, hence constant (Theorem 6.14). The argument for K_- is the same with i replaced by $-\mathrm{i}$. □

A.26.3 The Cayley transform

Definition A.269 (Cayley transform)**.** The *Cayley transform* of a closed symmetric operator A is the map

$$V_A \colon \mathrm{im}(A + \mathrm{i}\mu) \longrightarrow \mathrm{im}(A - \mathrm{i}\mu)\,, \qquad V_A\,(A + \mathrm{i}\mu)x = (A - \mathrm{i}\mu)x\,, \quad x \in D(A)\,. \tag{A.498}$$

Rests on Lemma A.267 and Definition 12.79.

Proposition A.270 (Properties of the transform)**.** *V_A is a well-defined linear isometry of the closed subspace $K_+^\perp = \mathrm{im}(A + \mathrm{i}\mu)$ onto the closed subspace $K_-^\perp = \mathrm{im}(A - \mathrm{i}\mu)$. Moreover $\mathbb{1} - V_A$ is injective on $K_+^\perp$, its image is $D(A)$ up to the factor $2\mathrm{i}\mu$,*

$$(\mathbb{1} - V_A)\,(A + \mathrm{i}\mu)x = 2\mathrm{i}\mu\, x\,, \qquad (\mathbb{1} + V_A)\,(A + \mathrm{i}\mu)x = 2\,Ax\,, \tag{A.499}$$

and consequently A is recovered from V_A by

$$D(A) = \mathrm{im}\,(\mathbb{1} - V_A)\ , \qquad A = \mathrm{i}\mu\,(\mathbb{1} + V_A)\,(\mathbb{1} - V_A)^{-1}\ . \tag{A.500}$$

Rests on Definition A.269 and Lemma A.267.

Proof. *Derives Proposition A.270.*

$A + \mathrm{i}\mu$ is injective on $D(A)$ (Lemma A.267), so every $u \in \mathrm{im}(A + \mathrm{i}\mu)$ is $(A + \mathrm{i}\mu)x$ for exactly one x and Equation (A.498) defines $V_A u$ unambiguously; linearity is inherited from that of A. Isometry is the second consequence in Lemma A.267, $\|V_A u\| = \|(A - \mathrm{i}\mu)x\| = \|(A + \mathrm{i}\mu)x\| = \|u\|$, and the image is $\mathrm{im}(A - \mathrm{i}\mu)$ by construction. Both subspaces are closed and are the orthogonal complements of $K_\pm$, by Equation (A.496).

The two identities Equation (A.499) are immediate: $(A + \mathrm{i}\mu)x - (A - \mathrm{i}\mu)x = 2\mathrm{i}\mu x$ and $(A + \mathrm{i}\mu)x + (A - \mathrm{i}\mu)x = 2Ax$. The first shows that $(\mathbb{1} - V_A)u = 0$ forces $x = 0$, hence $u = 0$: injectivity; that $\mathrm{im}(\mathbb{1} - V_A) = D(A)$, since x runs over $D(A)$ as u runs over $K_+^\perp$; and, dividing the second identity by the first, Equation (A.500). □

A.26.4 From an isometry back to an operator

Lemma A.271 (Injectivity of $\mathbb{1} - V$ for any isometric extension)**.** *Let V be an isometry of a closed subspace of $\mathcal{H}$ into $\mathcal{H}$ extending V_A. Then* $\ker(\mathbb{1} - V) = \{0\}$.

Rests on Proposition A.270 and Definition 12.72.

Proof. *Derives Lemma A.271.*

An isometry preserves inner products, by the polarization identity Equation (12.3) applied on its domain. Let $Vz = z$ and let $u \in K_+^\perp$ be arbitrary. Then

$$\langle z|(\mathbb{1} - V)\, u\rangle = \langle z|u\rangle - \langle z|Vu\rangle = \langle z|u\rangle - \langle Vz|Vu\rangle = 0\,,$$

using $Vz = z$ in the third step. By Equation (A.499), $(\mathbb{1} - V)u = (\mathbb{1} - V_A)u$ ranges over $2\mathrm{i}\mu\, D(A)$, which is dense because A is densely defined. Hence $z \perp \mathcal{H}$ and $z = 0$. □

Lemma A.272 (The operator attached to an isometry)**.** *Let V be an isometry of a closed subspace $D(V)$ onto a closed subspace, with* $\ker(\mathbb{1} - V) = \{0\}$ *and* $\mathrm{im}(\mathbb{1} - V)$ *dense. Then*

$$D(B) = \mathrm{im}\,(\mathbb{1} - V)\,, \qquad B\,(\mathbb{1} - V)\,z = \mathrm{i}\mu\,(\mathbb{1} + V)\,z \tag{A.501}$$

defines a densely defined symmetric operator B whose Cayley transform is V. If V extends V_A then B extends A.

Rests on Proposition A.270, Lemma A.271, and Definition 12.72.

Proof. *Derives Lemma A.272.*

B is well defined and single valued because $\mathbb{1} - V$ is injective, and linear because V is; its domain is dense by hypothesis.

Symmetry. Let $z, z' \in D(V)$ and put $x = (\mathbb{1} - V)z$, $y = (\mathbb{1} - V)z'$, so $Bx = \mathrm{i}\mu(\mathbb{1} + V)z$ and $By = \mathrm{i}\mu(\mathbb{1} + V)z'$. Using $\langle Vz|Vz'\rangle = \langle z|z'\rangle$ and the antilinearity of the first slot, which turns the prefactor $\mathrm{i}\mu$ into $-\mathrm{i}\mu$,

$$\langle Bx|y\rangle = -\mathrm{i}\mu\Big[\langle z|z'\rangle - \langle z|Vz'\rangle + \langle Vz|z'\rangle - \langle Vz|Vz'\rangle\Big] = -\mathrm{i}\mu\Big[\langle Vz|z'\rangle - \langle z|Vz'\rangle\Big]\,, \tag{A.502}$$

while, the prefactor now sitting in the second slot,

$$\langle x|By\rangle = +\mathrm{i}\mu\Big[\langle z|z'\rangle + \langle z|Vz'\rangle - \langle Vz|z'\rangle - \langle Vz|Vz'\rangle\Big] = -\mathrm{i}\mu\Big[\langle Vz|z'\rangle - \langle z|Vz'\rangle\Big]\,. \tag{A.503}$$

So $\langle Bx|y\rangle = \langle x|By\rangle$ for all $x, y \in D(B)$.

Its Cayley transform is V. From Equation (A.501),

$$(B + \mathrm{i}\mu)x = \mathrm{i}\mu\,(\mathbb{1} + V)\,z + \mathrm{i}\mu\,(\mathbb{1} - V)\,z = 2\mathrm{i}\mu\,z\,, \qquad (B - \mathrm{i}\mu)x = 2\mathrm{i}\mu\,Vz\,,$$

so $\mathrm{im}(B + \mathrm{i}\mu) = D(V)$, $\mathrm{im}(B - \mathrm{i}\mu) = \mathrm{im}\,V$, and V_B sends $2\mathrm{i}\mu z$ to $2\mathrm{i}\mu Vz$: that is, $V_B = V$.

Extension. If $V \supseteq V_A$ then, by Equation (A.499), $D(B) = \mathrm{im}(\mathbb{1} - V) \supseteq \mathrm{im}(\mathbb{1} - V_A) = D(A)$, and on $D(A)$ the prescription Equation (A.501) is Equation (A.499) itself, so B agrees with A there. □

Proposition A.273 (Self-adjoint means unitary)**.** *A closed symmetric operator A is self-adjoint if and only if its Cayley transform is unitary on all of $\mathcal{H}$, that is, if and only if $n_+ = n_- = 0$.*

Rests on Proposition A.270, Lemma A.267, and Definition 12.72.

Proof. *Derives Proposition A.273.*

Suppose $A = A^\dagger$. Then $K_\pm = \ker(A^\dagger \mp \mathrm{i}\mu) = \ker(A \mp \mathrm{i}\mu) = \{0\}$ by Equation (A.495), so $n_\pm = 0$ and, by Equation (A.496), V_A is an isometry of $\mathcal{H}$ onto $\mathcal{H}$: unitary.

Conversely suppose $n_+ = n_- = 0$, i.e. $\mathrm{im}(A \pm \mathrm{i}\mu) = \mathcal{H}$. Let $y \in D(A^\dagger)$. There is $x \in D(A)$ with $(A - \mathrm{i}\mu)x = (A^\dagger - \mathrm{i}\mu)y$, and since $A \subseteq A^\dagger$ this reads $(A^\dagger - \mathrm{i}\mu)(y - x) = 0$, i.e. $y - x \in K_+ = \{0\}$. Hence $y = x \in D(A)$, so $D(A^\dagger) = D(A)$ and A is self-adjoint. This is part (1) of Theorem A.266. □

A.26.5 Proof of the criterion

Proof of Theorem A.266, parts (2) and (3). *Derives Theorem A.266.*

From an extension to a unitary. Let $B \supseteq A$ be self-adjoint. By Proposition A.273 its Cayley transform V_B is unitary on $\mathcal{H}$, and it extends V_A: indeed $\mathrm{im}(A + \mathrm{i}\mu) \subseteq \mathrm{im}(B + \mathrm{i}\mu)$ and, on that subspace, $V_B(A + \mathrm{i}\mu)x = (B - \mathrm{i}\mu)x = (A - \mathrm{i}\mu)x$. A unitary carrying the closed subspace $K_+^\perp$ onto $K_-^\perp$ carries the orthogonal complement onto the orthogonal complement, so $W = V_B|_{K_+}$ is a unitary map $K_+ \longrightarrow K_-$.

From a unitary to an extension. Conversely let $W : K_+ \longrightarrow K_-$ be unitary and define, using the decompositions Equation (A.496),

$$V_W\,(u + k) = V_A u + W k\,, \qquad u \in K_+^\perp\,,\ k \in K_+\,. \tag{A.504}$$

Both summands on the right are orthogonal — $V_A u \in K_-^\perp$ and $Wk \in K_-$ — so $\|V_W(u+k)\|^2 = \|u\|^2 + \|k\|^2$ and V_W is an isometry of $\mathcal{H}$; its image is $K_-^\perp \oplus K_- = \mathcal{H}$, so V_W is unitary. By Lemma A.271, $\mathbb{1} - V_W$ is injective, and its image is dense because

$$\mathrm{im}\,(\mathbb{1} - V_W)^\perp = \ker\left(\mathbb{1} - V_W^\dagger\right) = \ker\,(\mathbb{1} - V_W) = \{0\}\ ,$$

where the middle equality holds because for a unitary V_W the relations $V_W z = z$ and $V_W^\dagger z = z$ are equivalent. So Lemma A.272 applies: it produces a densely defined symmetric $A_W \supseteq A$ whose Cayley transform is the unitary V_W, and A_W is therefore self-adjoint by Proposition A.273.

The formula Equation (A.494). Write $z = u + k$ with $u = (A + \mathrm{i}\mu)x$, $x \in D(A)$. Then, by Equations (A.499) and (A.504),

$$(\mathbb{1} - V_W)\,z = 2\mathrm{i}\mu\,x + k - Wk\,, \qquad \mathrm{i}\mu\,(\mathbb{1} + V_W)\,z = \mathrm{i}\mu\,(2Ax + k + Wk)\ ,$$

so, dividing by $2\mathrm{i}\mu$ and renaming $k/(2\mathrm{i}\mu)$ as k — which is a bijection of the subspace K_+, and W is linear — Equation (A.501) becomes exactly Equation (A.494). The sum there is direct: if $x + k - Wk = 0$ then applying $\mathbb{1} - V_W$ backwards, i.e. using injectivity on $2\mathrm{i}\mu x = (\mathbb{1} - V_W)u$ and $k - Wk = (\mathbb{1} - V_W)k$, gives $u + k = 0$ with $u \perp k$, hence $u = k = 0$.

The correspondence is bijective. The two constructions are inverse to each other: from A_W one recovers $V_{A_W} = V_W$ (Lemma A.272) and hence $W = V_W|_{K_+}$; and from a self-adjoint $B \supseteq A$ one gets $W = V_B|_{K_+}$, whose associated operator has Cayley transform V_B and therefore equals B, since B is determined by V_B through Equation (A.500).

Counting. A unitary map $K_+ \longrightarrow K_-$ exists if and only if the two spaces have orthonormal bases of the same cardinality — given such bases, the induced map is unitary; conversely a unitary carries a basis to a basis — that is, if and only if $n_+ = n_-$. If $n_+ \neq n_-$ there is none, and by the first paragraph a self-adjoint extension would produce one: this is part (3). When $n_+ = n_- = n < \infty$, fixing orthonormal bases of $K_\pm$ identifies the unitary maps with the matrix group $\mathrm{U}(n)$, a transitive and free action of $\mathrm{U}(n)$ on the set of extensions. □

A.26.6 The two worked cases

Both examples of Chapter 12 use the same differential expression $P = -\mathrm{i}\hbar\,\mathrm{d}/\mathrm{d}x$ of Equation (12.54), and both use the boundary form Equation (12.57), which for absolutely continuous φ, ψ with $\varphi', \psi' \in L^2$ reads

$$\left\langle \varphi \middle| P^\dagger \psi \right\rangle - \left\langle P^\dagger \varphi \middle| \psi \right\rangle = -\mathrm{i}\hbar\,(\varphi(L)^*\psi(L) - \varphi(0)^*\psi(0))\,, \tag{A.505}$$

an integration by parts legitimate for absolutely continuous functions (Theorem 7.43); on $[0,\infty)$ the term at L is replaced by a limit at infinity, which vanishes for L^2 functions with L^2 derivative. As Chapter 12 records, $P^\dagger$ is the same differential expression on the maximal domain — absolutely continuous ψ with $\psi' \in \mathcal{H}$ and no boundary condition [1481].

Corollary A.274 (Momentum on $[0,L]$: the circle of extensions)**.** *Let $\mathcal{H} = L^2([0,L])$ and let P be Equation (12.54) on the domain D_0 of continuously differentiable functions vanishing at both endpoints. Then $n_+ = n_- = 1$; every P_θ of Equation (12.55) is self-adjoint; these are all the self-adjoint extensions; and the unitary parameter $Wk_+ = \mathrm{e}^{\mathrm{i}\alpha}k_-$ of Theorem A.266 corresponds to the boundary phase*

$$\mathrm{e}^{\mathrm{i}\theta} = \frac{q - \mathrm{e}^{\mathrm{i}\alpha}}{1 - q\,\mathrm{e}^{\mathrm{i}\alpha}}\,, \qquad q = \mathrm{e}^{-\mu L/\hbar} \in (0,1)\,, \tag{A.506}$$

a bijection of the circle onto the circle.

Rests on Theorem A.266, Example 12.81, and Equation (12.55).

Proof. *Derives Corollary A.274.*

Indices. This is the computation of Example 12.81: the equations $P^\dagger\psi = \pm\mathrm{i}\mu\psi$ have the one-dimensional solution spaces spanned by

$$k_+(x) = N_+\mathrm{e}^{-\mu x/\hbar}\,, \qquad k_-(x) = N_-\mathrm{e}^{+\mu x/\hbar}\,, \tag{A.507}$$

both square integrable on the compact interval, so $n_+ = n_- = 1$ and Theorem A.266 gives a family of extensions parametrised by U(1). Normalising, $N_+^{-2} = \int_0^L \mathrm{e}^{-2\mu x/\hbar}\mathrm{d}x = (\hbar/2\mu)\,(1-q^2)$ and $N_-^{-2} = (\hbar/2\mu)\,(q^{-2}-1)$, whence

$$N_- = q\,N_+\,. \tag{A.508}$$

Each P_θ is self-adjoint. Symmetry on D_θ is Equation (A.505): for $\varphi, \psi \in D_\theta$, $\varphi(L)^*\psi(L) = \mathrm{e}^{-\mathrm{i}\theta}\mathrm{e}^{\mathrm{i}\theta}\varphi(0)^*\psi(0) = \varphi(0)^*\psi(0)$ and the right-hand side vanishes. For the converse inclusion let $y \in D(P_\theta^\dagger)$. Since $D_0 \subseteq D_\theta$, y belongs to the maximal domain $D(P^\dagger)$ and $P_\theta^\dagger y = P^\dagger y$; so Equation (A.505) with $\varphi = y$ must vanish for every $\psi \in D_\theta$:

$$0 = y(L)^*\psi(L) - y(0)^*\psi(0) = \psi(0)\left(y(L)^*\mathrm{e}^{\mathrm{i}\theta} - y(0)^*\right)\,.$$

Choosing $\psi(x) = 1 + (\mathrm{e}^{\mathrm{i}\theta} - 1)x/L$, which lies in D_θ and has $\psi(0) = 1$, gives $y(L)^*\mathrm{e}^{\mathrm{i}\theta} = y(0)^*$, i.e. $y(L) = \mathrm{e}^{\mathrm{i}\theta}y(0)$ and $y \in D_\theta$. Hence $D(P_\theta^\dagger) = D_\theta$ and P_θ is self-adjoint.

There are no others. Let B be a self-adjoint extension of P. Then $P \subseteq B = B^\dagger \subseteq P^\dagger$, so $D(B)$ lies in the maximal domain and Equation (A.505) vanishes on $D(B) \times D(B)$. Taking $\varphi = \psi$ gives $|\psi(L)|^2 = |\psi(0)|^2$ for every $\psi \in D(B)$: the image $S \subseteq \mathbb{C}^2$ of the boundary map $\psi \longmapsto (\psi(0), \psi(L))$ is a subspace isotropic for the hermitian form $|b|^2 - |a|^2$, which has signature $(1,1)$, so $\dim S \leq 1$. If $\dim S = 1$, say $S = \mathbb{C}\,(1, \mathrm{e}^{\mathrm{i}\theta})$, then $D(B) \subseteq D_\theta$, i.e. $B \subseteq P_\theta$; taking adjoints reverses the inclusion, so $P_\theta = P_\theta^\dagger \subseteq B^\dagger = B$ and $B = P_\theta$. If $S = \{0\}$ then $D(B)$ is contained in D_θ for *every* θ, so $B \subseteq P_\theta$ and the same argument gives $B = P_\theta$ — contradicting $S = \{0\}$, since the domain of P_θ contains $\psi(x) = 1 + (\mathrm{e}^{\mathrm{i}\theta} - 1)x/L$, whose boundary values are not both zero. So every self-adjoint extension is one of the P_θ.

Matching the parameters. Let $Wk_+ = \mathrm{e}^{\mathrm{i}\alpha}k_-$ and let A be the closure of P, a closed symmetric operator with the same adjoint and the same deficiency subspaces — the defining condition Equation (12.50) for $A^\dagger$ extends from D_0 to its graph closure by continuity of the inner product. By Equation (A.494), every element of $D(A_W)$ is $\psi = \psi_0 + c\,(k_+ - \mathrm{e}^{\mathrm{i}\alpha}k_-)$ with $\psi_0 \in D(A)$ and $c \in \mathbb{C}$. Every $\psi_0 \in D(A)$ vanishes at both endpoints: if $\psi_n \in D_0$ converges to ψ_0 in the graph norm then, for absolutely continuous ϕ on $[0, L]$,

$$|\phi(x)| \le \frac{1}{L}\int_0^L |\phi| + \int_0^L |\phi'| \le L^{-1/2}\,\|\phi\| + L^{1/2}\,\|\phi'\|\,, \tag{A.509}$$

— average $\phi(x) = \phi(y) + \int_y^x \phi'$ over y and use Equation (12.1) — so graph convergence implies uniform convergence and $\psi_0(0) = \lim \psi_n(0) = 0$, likewise at L. Therefore the boundary values of ψ come from the second term alone: by Equations (A.507) and (A.508),

$$\psi(0) = c\,N_+\left(1 - \mathrm{e}^{\mathrm{i}\alpha}q\right)\,, \qquad \psi(L) = c\,N_+\left(q - \mathrm{e}^{\mathrm{i}\alpha}\right)\,,$$

using $k_+(L) = N_+ q$ and $k_-(L) = N_- q^{-1} = N_+$. The denominator $1 - \mathrm{e}^{\mathrm{i}\alpha}q$ never vanishes because $q < 1$, and

$$\left|q - \mathrm{e}^{\mathrm{i}\alpha}\right|^2 = q^2 - 2q\cos\alpha + 1 = \left|1 - q\,\mathrm{e}^{\mathrm{i}\alpha}\right|^2\,, \tag{A.510}$$

so the ratio $\psi(L)/\psi(0)$ is a phase, which is Equation (A.506). Hence $D(A_W) \subseteq D_\theta$ with that θ, and since both operators are self-adjoint, $A_W = P_\theta$ by the adjoint argument used above. Finally $z \longmapsto (q - z)/(1 - qz)$ is a Möbius transformation with $|q| < 1$; Equation (A.510) shows it maps the unit circle into itself, and it is injective, so it is a bijection of the circle: as α runs once around, so does θ, and the U(1) of Theorem A.266 is exactly the circle of boundary conditions Equation (12.55). □

Corollary A.275 (Momentum on $[0, \infty)$: no extension)**.** *Let $\mathcal{H} = L^2([0, \infty))$ and let P be Equation (12.54) on the continuously differentiable functions of compact support in $(0, \infty)$. Then $(n_+, n_-) = (1, 0)$ and P has no self-adjoint extension.*

Rests on Theorem A.266 and Example 12.82.

Proof. *Derives Corollary A.275.*

The solutions of $P^\dagger\psi = \pm\mathrm{i}\mu\psi$ are again $\psi(x) = C\mathrm{e}^{\mp\mu x/\hbar}$; on the half-line the decaying one is square integrable, with $\|\psi\|^2 = |C|^2\,\hbar/(2\mu)$, and the growing one is not, so $n_+ = 1$ and $n_- = 0$ (Example 12.82). By part (3) of Theorem A.266 — there is no unitary map from a one-dimensional space onto $\{0\}$ — there is no self-adjoint extension. By Proposition A.263 there is therefore no strongly continuous unitary group whose generator is a momentum on the half-line, which is the statement that probability cannot be conserved under translations that push it off the end of the half-line (Remark 12.83). □

Remark A.276 (What is quoted here)*.* The criterion itself — Lemma A.267 through Theorem A.266 — quotes nothing: every step is carried out from the definitions of Section 12.5 and the projection theorem. Two facts are imported in the worked cases of Appendix A.26.6, both of them classical real analysis rather than operator theory. First, that the adjoint of P on either interval is the same differential expression on the maximal domain of absolutely continuous functions with square-integrable derivative; this is stated without proof in Example 12.81 as well, with the reference [1481], and identifying it requires the du Bois-Reymond lemma on weak derivatives. Second, that integration by parts Equation (A.505) is valid for absolutely continuous functions, which is Theorem 7.43 applied to the product. Neither is needed for Theorem A.266.

Remark A.277*.* Appendix A.26 discharges the proof obligation of Theorem 12.80 (Section 12.5.2), and with Corollaries A.274 and A.275 it completes the two examples that section is built around — in particular the assertion of Example 12.81 that D_θ exhausts the

possibilities, which is proved above by the isotropic-subspace count and not merely asserted. Read with Appendix A.25, the criterion is the precise statement of what Remark 12.77 calls the physics of a domain: the deficiency indices decide whether a symmetric differential expression admits a unitary dynamics at all, and when they are equal and non-zero it is the boundary condition — the ring, or the ring threaded by a flux — that selects one, the mathematics offering the whole $\mathrm{U}(n)$ and choosing nothing.

A.27 The Nuclear Spectral Theorem of Gelfand and Maurin

This appendix proves Theorem 12.107 of Chapter 12 — that a self-adjoint operator acting continuously on the test space of a Gelfand triple possesses a complete set of generalized eigenvectors — and says exactly what the proof assumes. It is the theorem that licenses the completeness relation $\int |\lambda\rangle\langle\lambda| \, \mathrm{d}\mu(\lambda) = \mathbb{1}$ of Remark 12.108, and with it Dirac's formalism [646] in the form physics actually uses it. The result is due to Gelfand and his collaborators [1276], where it appears in volume 4 of *Generalized Functions*; Maurin obtained it independently at about the same time, and the theorem carries both names.

The architecture has three stages, and only the second imports anything. *First*, the spectral theorem of Appendices A.24 and A.25 is read as a *direct-integral* decomposition: $\mathcal{H}$ is realised as a space of square-integrable fields $\lambda \longmapsto \hat{x}(\lambda)$ over the spectrum, in which A acts by multiplication by λ. At that stage the assignment $x \longmapsto \hat{x}(\lambda)$ is defined only up to a μ-null set, and only as a map of equivalence classes: for a *fixed* λ it is not a map on $\mathcal{H}$ at all, so there are no eigenvectors yet, generalized or otherwise. *Second*, nuclearity of the test space is used — and this is the only import — to produce, for μ-almost every λ, a genuine *continuous* map $\Phi \longrightarrow \mathcal{H}_\lambda$ that is a version of the fibre map. *Third*, each vector in the fibre is paired with that map, giving a continuous linear functional on Φ, and the multiplication form of A makes it a generalized eigenvector of eigenvalue λ; the isometry of the first stage, read fibrewise, is the completeness relation.

A.27.1 Hypotheses and statement

The standing hypotheses of Definition 12.103 are made explicit here, because the proof uses each of them.

Definition A.278 (Countably Hilbert nuclear space)**.** A *countably Hilbert space* is a vector space Φ whose topology is given by an increasing sequence of norms $\|\cdot\|_1 \le \|\cdot\|_2 \le \cdots$, each coming from an inner product, and which is complete for the resulting metric; Φ_p denotes the Hilbert space completion of Φ in $\|\cdot\|_p$. It is *nuclear* if for every p there is $q \ge p$ such that the canonical map $\Phi_q \longrightarrow \Phi_p$ is nuclear, that is, of the form $u \longmapsto \sum_k s_k \, \langle a_k|u\rangle \, b_k$ with $\sum_k |s_k| < \infty$ and $\{a_k\}$, $\{b_k\}$ orthonormal.
Rests on Definitions 12.2 and 12.103.

Theorem A.279 (Gelfand–Maurin)**.** *Let $\Phi \subseteq \mathcal{H} \subseteq \Phi'$ be a Gelfand triple (Definition 12.103) in which $\mathcal{H}$ is separable and Φ is a separable countably Hilbert nuclear space, densely and continuously embedded in $\mathcal{H}$. Let A be self-adjoint on $\mathcal{H}$ and map Φ continuously into itself. Then there exist a finite Borel measure μ on $\sigma(A)$, separable Hilbert spaces $\mathcal{H}_\lambda$ and, for μ-almost every λ, a continuous linear map $\Lambda_\lambda \colon \Phi \longrightarrow \mathcal{H}_\lambda$ such that*

1. *$\Lambda_\lambda(A\varphi) = \lambda \, \Lambda_\lambda(\varphi)$ for every $\varphi \in \Phi$ and μ-almost every λ;*
2. *for every $\xi \in \mathcal{H}_\lambda$ the functional $F_{\lambda,\xi}(\varphi) = \langle \xi | \Lambda_\lambda \varphi \rangle_{\mathcal{H}_\lambda}$ belongs to Φ' and is a generalized eigenvector of A with eigenvalue λ (Definition 12.105);*
3. *writing $\{\xi_n\}$ for an orthonormal basis of $\mathcal{H}_\lambda$ and $F_{\lambda,n} = F_{\lambda,\xi_n}$,*

$$\langle \varphi | \psi \rangle = \int_{\sigma(A)} \sum_n F_{\lambda,n}(\varphi)^* \, F_{\lambda,n}(\psi) \, \mathrm{d}\mu(\lambda) \,, \qquad \forall \, \varphi, \psi \in \Phi \,. \tag{A.511}$$

When the spectrum of A is simple — when $\mathcal{H}$ has a single cyclic vector — every $\mathcal{H}_\lambda$ is one dimensional, the sum in Equation (A.511) has one term, and the identity is Equation (12.69) exactly as stated in Chapter 12.

Rests on Definitions 12.103 and 12.105 and Theorem A.238.

A.27.2 The direct integral

Proposition A.280 (Direct-integral form of the spectral theorem)**.** *Let A be self-adjoint on a separable $\mathcal{H}$, with projection-valued measure E (Theorem A.238 and Proposition A.262). There are a finite Borel measure μ on $\mathbb{R}$, supported on $\sigma(A)$, and an isometry*

$$J\colon \mathcal{H} \longrightarrow \Big\{ \textit{fields } \lambda \longmapsto \hat{x}(\lambda) \in \ell^2(\mathbb{N}) \Big\}\,, \qquad \|x\|^2 = \int_{\mathbb{R}} \|\hat{x}(\lambda)\|_{\ell^2}^2 \,\mathrm{d}\mu(\lambda)\,, \tag{A.512}$$

such that for $x \in D(A)$ one has $\widehat{Ax}(\lambda) = \lambda\,\hat{x}(\lambda)$ for μ-almost every λ.

Rests on Theorem A.238, Lemma A.250, and Proposition A.262.

Proof. *Derives Proposition A.280.*

Cyclic decomposition. The construction of Lemma A.250 applies verbatim with the bounded Borel calculus of Proposition A.246 — or, for unbounded A, of Proposition A.262 — in place of the continuous one: it used only that the cyclic subspaces $\mathcal{H}_n = \overline{\{g(A)x_n \mid g \text{ bounded Borel}\}}$ are A-invariant with A-invariant orthogonal complements, and that $\mathcal{H}$ is separable. So $\mathcal{H} = \bigoplus_n \mathcal{H}_n$ with each x_n cyclic in $\mathcal{H}_n$.

Each block. As in Lemma A.249, $\|g(A)x_n\|^2 = \int |g|^2 \,\mathrm{d}\mu_n$ with $\mu_n(\Omega) = \langle x_n | E(\Omega) x_n\rangle$, so $g(A)x_n \longmapsto g$ extends to a unitary $U_n\colon \mathcal{H}_n \longrightarrow L^2(\mathbb{R},\mu_n)$ — the bounded Borel functions are dense in $L^2(\mu_n)$, since the simple functions already are — and $U_n A U_n^{-1}$ is multiplication by λ.

One measure. Put $\mu = \sum_n 2^{-n}\,\|x_n\|^{-2}\,\mu_n$, a Borel measure with $\mu(\mathbb{R}) \le 1$, supported on $\sigma(A)$ because every μ_n is. Each μ_n is absolutely continuous with respect to μ, so the Radon–Nikodym theorem (Remark A.284) provides densities $\rho_n = \mathrm{d}\mu_n/\mathrm{d}\mu \ge 0$.

The isometry. For $x = \sum_n v_n$ with $v_n \in \mathcal{H}_n$ write $h_n = U_n v_n \in L^2(\mu_n)$ and define

$$\hat{x}(\lambda) = \Big(\sqrt{\rho_n(\lambda)}\; h_n(\lambda) \Big)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N})\,. \tag{A.513}$$

Then, exchanging sum and integral by monotone convergence (Remark 12.1),

$$\int_{\mathbb{R}} \|\hat{x}(\lambda)\|_{\ell^2}^2 \,\mathrm{d}\mu = \sum_n \int_{\mathbb{R}} |h_n|^2\,\rho_n \,\mathrm{d}\mu = \sum_n \int_{\mathbb{R}} |h_n|^2 \,\mathrm{d}\mu_n = \sum_n \|v_n\|^2 = \|x\|^2\,,$$

which is Equation (A.512); in particular $\hat{x}(\lambda)$ lies in $\ell^2(\mathbb{N})$ for μ-almost every λ. Since A acts on each block as multiplication by λ, and multiplication commutes with the fibrewise rescaling by $\sqrt{\rho_n}$, $\widehat{Ax}(\lambda) = \lambda\hat{x}(\lambda)$ almost everywhere for $x \in D(A)$. □

The map J is the whole content of the spectral theorem written fibrewise, and it is also the whole of the difficulty: $\hat{x}$ is an equivalence class of fields modulo μ-null sets, so the expression $\hat{x}(\lambda)$ for one fixed λ has no meaning, and there is nothing yet that could be evaluated on a test function.

A.27.3 What nuclearity buys

Theorem A.281 (Nuclear spaces embed by Hilbert–Schmidt maps; quoted)**.** *Let Φ be a countably Hilbert nuclear space (Definition A.278) continuously embedded in a Hilbert space $\mathcal{H}$. Then there is an index q such that the embedding extends to a* Hilbert–Schmidt *map $J_q\colon \Phi_q \longrightarrow \mathcal{H}$, that is, one for which*

$$\sum_k \|e_k\|_{\mathcal{H}}^2 < \infty \tag{A.514}$$

for one, hence for every, orthonormal basis $\{e_k\}$ *of* Φ_q.
Rests on Definition A.278.

The Schwartz triple Equation (12.66) is the case to keep in mind: $\mathcal{S}(\mathbb{R})$ is countably Hilbert for the norms built from the harmonic-oscillator Hermite expansion, and Equation (A.514) holds there because the Hermite coefficients of a Schwartz function decay faster than any power.

Proposition A.282 (The fibre maps are continuous on Φ)**.** *Under the hypotheses of Theorem A.279 there are, for* μ*-almost every* λ*, a constant* $C(\lambda) < \infty$ *and a linear map* $\Lambda_\lambda \colon \Phi_q \longrightarrow \ell^2(\mathbb{N})$ *with*

$$\|\Lambda_\lambda \varphi\|_{\ell^2} \le C(\lambda)\, \|\varphi\|_q\,, \qquad \forall\, \varphi \in \Phi_q\,, \tag{A.515}$$

such that for each fixed $\varphi \in \Phi$ *one has* $\Lambda_\lambda \varphi = \hat{\varphi}(\lambda)$ *for* μ*-almost every* λ*. In particular* Λ_λ *is continuous for the topology of* Φ.
Rests on Theorem A.281 and Proposition A.280.

Proof. *Derives Proposition A.282.*
Let q and $\{e_k\}$ be as in Theorem A.281. Applying Equation (A.512) to each e_k and summing,

$$\int_{\mathbb{R}} \left(\sum_k \|\hat{e}_k(\lambda)\|_{\ell^2}^2 \right) \mathrm{d}\mu(\lambda) = \sum_k \|e_k\|_{\mathcal{H}}^2 < \infty\,, \tag{A.516}$$

the exchange of sum and integral being monotone convergence for a series of non-negative terms (Remark 12.1). A non-negative function with finite integral is finite almost everywhere, so

$$C(\lambda)^2 = \sum_k \|\hat{e}_k(\lambda)\|_{\ell^2}^2 < \infty \qquad \text{for } \mu\text{-almost every } \lambda\,. \tag{A.517}$$

This is the entire role of nuclearity: without Equation (A.514) the left-hand side of Equation (A.516) is a divergent series and Equation (A.517) says nothing.

Fix such a λ. For $\varphi \in \Phi_q$ expand $\varphi = \sum_k c_k e_k$ with $c_k = \langle e_k | \varphi \rangle_q$ and $\sum_k |c_k|^2 = \|\varphi\|_q^2$ (Theorem 12.30), and put

$$\Lambda_\lambda \varphi = \sum_k c_k\, \hat{e}_k(\lambda)\,. \tag{A.518}$$

The series converges absolutely in $\ell^2(\mathbb{N})$, since by Cauchy–Schwarz

$$\sum_k |c_k|\, \|\hat{e}_k(\lambda)\| \le \Big(\sum_k |c_k|^2 \Big)^{1/2} \Big(\sum_k \|\hat{e}_k(\lambda)\|^2 \Big)^{1/2} = C(\lambda)\, \|\varphi\|_q\,,$$

which is also Equation (A.515). Linearity is clear.

It remains to identify $\Lambda_\lambda \varphi$ with the fibre of φ. Fix $\varphi \in \Phi$ and let $S_N = \sum_{k \le N} c_k e_k$. Then $S_N \longrightarrow \varphi$ in Φ_q, hence in $\mathcal{H}$ because the embedding is bounded, and J is isometric, so

$$\int_{\mathbb{R}} \big\| \hat{S}_N(\lambda) - \hat{\varphi}(\lambda) \big\|^2 \, \mathrm{d}\mu(\lambda) = \|S_N - \varphi\|_{\mathcal{H}}^2 \longrightarrow 0\,.$$

A sequence converging in $L^2(\mu)$ has a subsequence converging μ-almost everywhere (Remark A.284), so along it $\hat{S}_N(\lambda) \longrightarrow \hat{\varphi}(\lambda)$ for almost every λ. But $\hat{S}_N(\lambda)$ is the Nth partial sum of Equation (A.518), which converges to $\Lambda_\lambda \varphi$ for every λ satisfying Equation (A.517). Limits are unique, so $\Lambda_\lambda \varphi = \hat{\varphi}(\lambda)$ almost everywhere. Finally $\|\cdot\|_q$ is one of the norms defining the topology of Φ, so Equation (A.515) is continuity on Φ. □

A.27.4 Generalized eigenvectors, and completeness

Proof of Theorem A.279. *Derives Theorem A.279.*

Take μ, $\mathcal{H}_\lambda = \ell^2(\mathbb{N})$ and Λ_λ from Propositions A.280 and A.282.

(1) The eigenvalue relation. Since A maps Φ into Φ, in particular $\Phi \subseteq D(A)$, and Proposition A.280 gives, for each fixed $\varphi \in \Phi$,

$$\Lambda_\lambda(A\varphi) = \widehat{A\varphi}(\lambda) = \lambda\,\hat{\varphi}(\lambda) = \lambda\,\Lambda_\lambda\varphi \qquad \text{for } \mu\text{-almost every } \lambda\,, \tag{A.519}$$

the exceptional null set depending on φ. Let $\{\varphi_j\}$ be dense in Φ, which exists because Φ is separable, and let Z be the union of the countably many exceptional sets attached to the φ_j together with the null set outside which Equation (A.517) holds; Z is μ-null. For $\lambda \notin Z$ the two maps $\varphi \longmapsto \Lambda_\lambda(A\varphi)$ and $\varphi \longmapsto \lambda\Lambda_\lambda\varphi$ are continuous on Φ — the first because A is continuous from Φ to Φ and Λ_λ is continuous on Φ, the second by Equation (A.515) — and they agree on the dense set $\{\varphi_j\}$, hence everywhere. This is (1), now with one null set for all φ at once.

(2) The functionals. For $\lambda \notin Z$ and $\xi \in \mathcal{H}_\lambda$ put $F_{\lambda,\xi}(\varphi) = \langle \xi | \Lambda_\lambda\varphi\rangle$. It is linear in φ, because Λ_λ is linear and the inner product is linear in its second slot (Equation (5.34)), and continuous, since $|F_{\lambda,\xi}(\varphi)| \le \|\xi\|\, C(\lambda)\, \|\varphi\|_q$ by Equation (12.1) and Equation (A.515); so $F_{\lambda,\xi} \in \Phi'$ (Definition 12.103). By (1) and the reality of λ,

$$F_{\lambda,\xi}(A\varphi) = \langle \xi | \Lambda_\lambda(A\varphi)\rangle = \langle \xi | \lambda\Lambda_\lambda\varphi\rangle = \lambda\, F_{\lambda,\xi}(\varphi)\,,$$

which is Equation (12.68): $F_{\lambda,\xi}$ is a generalized eigenvector of A with eigenvalue λ.

(3) Completeness. Polarising the isometry Equation (A.512) — by Equation (12.3), an isometry of Hilbert spaces preserves inner products — gives, for $\varphi, \psi \in \Phi$,

$$\langle \varphi | \psi \rangle_{\mathcal{H}} = \int_{\mathbb{R}} \Big\langle \hat{\varphi}(\lambda) \Big| \hat{\psi}(\lambda) \Big\rangle_{\ell^2}\, \mathrm{d}\mu(\lambda)\,. \tag{A.520}$$

Let $\{\xi_n\}$ be the standard orthonormal basis of $\ell^2(\mathbb{N})$, the same for every λ, and $F_{\lambda,n} = F_{\lambda,\xi_n}$. Parseval's identity Equation (12.20) in the fibre reads

$$\Big\langle \hat{\varphi}(\lambda) \Big| \hat{\psi}(\lambda) \Big\rangle = \sum_n \langle \hat{\varphi}(\lambda) | \xi_n \rangle \Big\langle \xi_n \Big| \hat{\psi}(\lambda) \Big\rangle = \sum_n F_{\lambda,n}(\varphi)^*\, F_{\lambda,n}(\psi)\,,$$

using $\Lambda_\lambda\varphi = \hat{\varphi}(\lambda)$ almost everywhere and $\langle \hat{\varphi} | \xi_n\rangle = \langle \xi_n | \hat{\varphi}\rangle^* = F_{\lambda,n}(\varphi)^*$. Substituting into Equation (A.520) gives Equation (A.511). Since μ is supported on $\sigma(A)$, the integral may be written over $\sigma(A)$.

Simple spectrum. If a single cyclic vector generates $\mathcal{H}$ the decomposition of Proposition A.280 has one block, $\ell^2(\mathbb{N})$ is replaced by $\mathbb{C}$, and Equation (A.511) collapses to $\langle \varphi | \psi \rangle = \int F_\lambda(\varphi)^* F_\lambda(\psi)\, \mathrm{d}\mu$, which is Equation (12.69). □

Example A.283 (Momentum on the line)**.** For the Schwartz triple Equation (12.66) and $A = P = -\mathrm{i}\hbar\,\mathrm{d}/\mathrm{d}x$, the spectrum is simple, μ may be taken to be Lebesgue measure on $\mathbb{R}$ in the variable p, and $\Lambda_p\varphi$ is the Fourier transform Equation (12.67), so that $F_p = F_p$ of Proposition 12.106. Statement (2) of Theorem A.279 is then the computation already carried out there, and Equation (A.511) is the Plancherel identity of Chapter 17. The plane waves are a complete set of generalized eigenvectors of the momentum, and no one of them is a state.

Rests on Theorem A.279, Proposition 12.106, and Example 12.104.

Remark A.284 (What is quoted here). This is the one long proof of Chapter 12 that rests on a theory the treatise does not build, and the reader is owed a precise account of where.

- Theorem A.281 — that a countably Hilbert nuclear space embeds in $\mathcal{H}$ through some Hilbert–Schmidt map — is *assumed*. It is the definition of nuclearity (Definition A.278) plus the elementary fact that the composition of a nuclear map with a bounded one is Hilbert–Schmidt, but the theory of nuclear spaces itself, including the proof that $\mathcal{S}(\mathbb{R})$ is nuclear, is not developed anywhere in this treatise. Reference of record: [1276], volume 4, chapter I. Everything that nuclearity is used for is Equations (A.516) and (A.517) and nothing else — which is why the architecture above is worth having even though the input is quoted: the reader can see exactly how much is being bought, and how little.

- Three standard theorems of measure theory are used: the Radon–Nikodym theorem, in Proposition A.280; the monotone convergence theorem, twice; and the fact that a sequence convergent in L^2 has an almost-everywhere convergent subsequence. The treatise develops no measure theory and says so in Remark 12.1, where the convergence theorems are already declared.

- The spectral theorem is *not* quoted: the direct integral of Proposition A.280 is built from Appendix A.24 for bounded A and from Proposition A.262 for unbounded A — the momentum operator of Example A.283 is the unbounded case.

- Two hypotheses of Theorem A.279 are suppressed in the chapter's statement Theorem 12.107 and are used here: separability of $\mathcal{H}$, without which the cyclic decomposition need not be countable, and separability of Φ, without which the single null set of Equation (A.519) cannot be assembled. Both hold for the Schwartz triple. A third point of honesty: for a spectrum of multiplicity greater than one the completeness relation carries the sum over n displayed in Equation (A.511), which Equation (12.69) suppresses; the two agree exactly when the spectrum is simple.

Nothing above claims more. In particular the theorem does not say that the $F_{\lambda,n}$ are unique, nor that every generalized eigenvector arises this way, nor that μ is canonical — only its measure class is — and no use made of the theorem in this treatise requires any of those.

Remark A.285. Appendix A.27 discharges the proof obligation of Theorem 12.107 (Section 12.6.3). What it licenses is the working rule of Remark 12.108: a ket belonging to a point of continuous spectrum is an element of Φ', the pairing $\langle\lambda|\varphi\rangle$ is the number $F_\lambda(\varphi)$ for a test function φ, and the "normalisation" $\langle\lambda|\lambda'\rangle = \delta(\lambda-\lambda')$ is shorthand for Equation (A.511) — an identity between functionals, never between numbers. Proposition 12.102 is the proof that no other reading is available, and Example A.283 is the case every scattering calculation of Chapter 86 silently uses.

A.28 The Implicit Function Theorem

This appendix proves the one input that Theorem 13.59 of Chapter 13 takes from multivariable calculus and that Chapter 7 does not state: where the derivative of a system of equations with respect to one block of variables is invertible, the system can be solved locally for that block as a differentiable function of the remaining variables, and the solution is as smooth as the system. It is the theorem that turns a level set into a graph, and through the regular value theorem it is what makes a sphere, a hyperboloid, a constraint surface or a mass shell a manifold at all.

What is assumed as already available is exactly the several-variable differential calculus of Section 7.10.2: partial derivatives and the gradient (Definition 7.65), the class C^1 (Definition 7.66), differentiability at a point with the uniqueness of the differential (Definitions 7.67 and 7.70), the theorem that a C^1 map is differentiable (Theorem 7.68),

the chain rule (Proposition 7.72), the one-variable mean value theorem (Theorem 7.35), the Cauchy criterion (Theorem 7.8) and the geometric series (Proposition 7.46). *No* fixed-point theorem is imported: the contraction is carried out explicitly, in the manner of the Picard iteration of Theorem 9.8, with the geometric estimate written out. Nothing else is quoted — see Remark A.295.

Throughout, $|\cdot|$ is the Euclidean length (Equation (6.12)) and, for a real matrix A,

$$\|A\| = \left(\sum_{i,j} (A_{ij})^2 \right)^{1/2} , \qquad |Av| \le \|A\|\,|v| , \tag{A.521}$$

by the Cauchy–Schwarz inequality of Chapter 5 applied row by row. Points of $\mathbb{R}^m \times \mathbb{R}^n$ are written (x, y) with $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$; for a map F into $\mathbb{R}^n$ we write

$$(D_x F)^i{}_j = \frac{\partial F^i}{\partial x^j} , \qquad (D_y F)^i{}_j = \frac{\partial F^i}{\partial y^j} , \tag{A.522}$$

the first an $n \times m$ matrix, the second an $n \times n$ matrix.

A.28.1 Statement

Theorem A.286 (Implicit function theorem)**.** *Let $A \subseteq \mathbb{R}^m \times \mathbb{R}^n$ be open, let $F : A \longrightarrow \mathbb{R}^n$ be of class C^k with $k \ge 1$, and let $(a, b) \in A$ satisfy*

$$F(a, b) = 0 , \qquad \det\left[D_y F(a, b) \right] \neq 0 . \tag{A.523}$$

Then there are open sets $W \subseteq \mathbb{R}^m$ with $a \in W$ and $V \subseteq \mathbb{R}^n$ with $b \in V$, satisfying $W \times V \subseteq A$, and a map $h : W \longrightarrow V$ such that

$$\{ (x, y) \in W \times V \mid F(x, y) = 0 \} = \{ (x, h(x)) \mid x \in W \} , \qquad h(a) = b . \tag{A.524}$$

The map h is unique: it is the only map $W \longrightarrow V$ whose graph is the zero set. It is of class C^k, and

$$Dh(x) = - \left[D_y F(x, h(x)) \right]^{-1} D_x F(x, h(x)) , \qquad x \in W . \tag{A.525}$$

If F is smooth, so is h.

Rests on Theorem 7.68, Proposition 7.72, and Definition 7.67.

The proof is in four movements: the equation is recast as a fixed-point problem whose iteration contracts (Appendix A.28.2); the iteration is run and its limit shown to be the unique solution (Appendix A.28.3); the solution is shown Lipschitz and then differentiable, with Equation (A.525) (Appendix A.28.4); and the smoothness is bootstrapped to C^k there as well. Appendix A.28.5 draws the two corollaries the treatise uses.

A.28.2 The contraction

Lemma A.287 (The Newton map contracts)**.** *Under the hypotheses of Theorem A.286, put*

$$C = \left[D_y F(a, b) \right]^{-1} , \qquad T(x, y) = y - C\, F(x, y) . \tag{A.526}$$

Then there are $r > 0$ and $\rho \in (0, r]$ such that, writing $\overline{B}_r$ for closed balls, $\overline{B}_\rho(a) \times \overline{B}_r(b) \subseteq A$ and:

1. *for every $y \in \overline{B}_r(b)$ and $x \in \overline{B}_\rho(a)$, $F(x, y) = 0$ if and only if $T(x, y) = y$;*
2. *$|T(x, y) - T(x, y')| \le \frac{1}{2}\,|y - y'|$ for all $y, y' \in \overline{B}_r(b)$ and $x \in \overline{B}_\rho(a)$;*

3. $|T(x,b)-b| \le r/4$ *for every* $x \in \overline{B}_\rho(a)$;

4. $T(x,\cdot)$ *maps* $\overline{B}_r(b)$ *into itself; and*

5. $\det[D_y F(x,y)] \neq 0$ *throughout* $\overline{B}_\rho(a) \times \overline{B}_r(b)$.

Rests on Equation (A.523), Theorem 7.35, and Definition 7.66.

Proof. *Derives Lemma A.287.*

C exists by Equation (A.523). Assertion (1) is immediate from Equation (A.526): $T(x,y) = y$ says $C\,F(x,y) = 0$, and C is invertible, so this says $F(x,y) = 0$.

Differentiating Equation (A.526) in y,

$$D_y T(x,y) = \mathbb{1} - C\,D_y F(x,y)\,, \tag{A.527}$$

which vanishes at (a,b) by the definition of C. The entries of $D_y F$ are continuous, because F is C^1 (Definition 7.66), hence so are those of $D_y T$ and so is $\det[D_y F]$. Choose $r > 0$ with $\overline{B}_r(a) \times \overline{B}_r(b) \subseteq A$ and, by that continuity together with $D_y T(a,b) = 0$ and $\det[D_y F(a,b)] \neq 0$, small enough that

$$\|D_y T(x,y)\| \le \frac{1}{2\sqrt{n}}\,, \qquad \det[D_y F(x,y)] \neq 0\,, \tag{A.528}$$

on the whole of $\overline{B}_r(a) \times \overline{B}_r(b)$. That is assertion (5).

For (2), fix x and y, y' in the convex set $\overline{B}_r(b)$. Applying Theorem 7.35 to $\lambda \longmapsto T^i\,(x, y' + \lambda(y - y'))$ on $[0,1]$ gives $\theta_i \in (0,1)$ with

$$T^i(x,y) - T^i(x,y') = \nabla_y T^i(x,\xi_i) \cdot (y - y')\,, \qquad \xi_i = y' + \theta_i(y - y') \in \overline{B}_r(b)\,,$$

so that $|T^i(x,y) - T^i(x,y')| \le \|D_y T(x,\xi_i)\|\ |y - y'| \le |y - y'|\,/(2\sqrt{n})$, the row length being at most the whole matrix norm Equation (A.521). Squaring and summing the n components gives (2).

For (3): $T(a,b) = b$ because $F(a,b) = 0$, and $x \longmapsto T(x,b)$ is continuous, so there is $\rho \in (0,r]$ with $|T(x,b) - b| \le r/4$ for $|x - a| \le \rho$.

For (4), combine (2) and (3): for $y \in \overline{B}_r(b)$ and $|x - a| \le \rho$,

$$|T(x,y) - b| \le |T(x,y) - T(x,b)| + |T(x,b) - b| \le \frac{|y-b|}{2} + \frac{r}{4} \le \frac{3r}{4} < r\,. \tag{A.529}$$

□

A.28.3 Existence and uniqueness of the implicit map

Lemma A.288 (The iteration converges)**.** *With* r, ρ *as in Lemma A.287, fix* $x \in \overline{B}_\rho(a)$ *and define*

$$y_0 = b\,, \qquad y_{j+1} = T(x, y_j)\,. \tag{A.530}$$

Then every y_j *lies in* $\overline{B}_r(b)$*, the sequence converges to a limit* $h(x)$ *with*

$$|h(x) - b| \le \frac{r}{2}\,, \tag{A.531}$$

and $h(x)$ *is the unique point of* $\overline{B}_r(b)$ *with* $F(x, h(x)) = 0$.

Rests on Lemma A.287, Theorem 7.8, and Proposition 7.46.

Proof. *Derives Lemma A.288.*

The iterates stay in $\overline{B}_r(b)$ by Equation (A.529) and induction. By assertion (2) of Lemma A.287,

$$|y_{j+1} - y_j| = |T(x, y_j) - T(x, y_{j-1})| \le \frac{1}{2} |y_j - y_{j-1}| \le \cdots \le 2^{-j} |y_1 - y_0| \le 2^{-j} \frac{r}{4}, \tag{A.532}$$

the last step by assertion (3). For $l > j$ the triangle inequality (Proposition 7.3) and the geometric series (Proposition 7.46) give

$$|y_l - y_j| \le \sum_{s=j}^{l-1} |y_{s+1} - y_s| \le \frac{r}{4} \sum_{s \ge j} 2^{-s} = \frac{r}{2^{j+1}},$$

so each of the n coordinate sequences is a Cauchy sequence of reals and converges (Theorem 7.8); call the limit $h(x)$, which lies in the closed set $\overline{B}_r(b)$. Letting $l \to \infty$ with $j = 0$ in the display gives $|h(x) - b| \le r/2$, which is Equation (A.531). Since $T(x, \cdot)$ is continuous (indeed Lipschitz, by assertion (2)), passing to the limit in Equation (A.530) gives $T(x, h(x)) = h(x)$, i.e. $F(x, h(x)) = 0$ by assertion (1).

Uniqueness in $\overline{B}_r(b)$: if y and y' both satisfy $F(x, \cdot) = 0$ there, both are fixed points of $T(x, \cdot)$, and assertion (2) gives $|y - y'| \le \frac{1}{2} |y - y'|$, hence $y = y'$. □

Corollary A.289 (The zero set is a graph)**.** *Put $W = B_\rho(a)$ and $V = B_r(b)$, the* open *balls. Then $W \times V \subseteq A$, $h(W) \subseteq V$, $h(a) = b$, and Equation (A.524) holds; moreover h is the only map $W \longrightarrow V$ whose graph is the zero set of F in $W \times V$.*

Rests on Lemma A.288.

Proof. *Derives Corollary A.289.*

$h(x) \in V$ by Equation (A.531), since $r/2 < r$; and $h(a) = b$ because $y \equiv b$ is then a fixed point, unique by Lemma A.288. If $(x, y) \in W \times V$ has $F(x, y) = 0$ then $y \in \overline{B}_r(b)$, so $y = h(x)$ by the uniqueness clause; conversely $F(x, h(x)) = 0$ for every $x \in W$. That is Equation (A.524). Any other map $g : W \longrightarrow V$ with the same graph would satisfy $F(x, g(x)) = 0$ with $g(x) \in \overline{B}_r(b)$, hence $g = h$. □

A.28.4 Continuity, differentiability and the derivative formula

Lemma A.290 (h is Lipschitz)**.** *There is $K \ge 0$ with*

$$|h(x) - h(x')| \le K \, |x - x'|, \qquad x, x' \in W. \tag{A.533}$$

In particular h is continuous.

Rests on Lemma A.288 and Theorem 7.35.

Proof. *Derives Lemma A.290.*

The entries of $D_x F$ are continuous, hence bounded on the compact set $\overline{B}_\rho(a) \times \overline{B}_r(b)$ (Theorem 7.24); let μ bound $\|D_x F\|$ there. Exactly as in step (2) of Lemma A.287, the mean value theorem along the segment from x' to x — which lies in the convex ball $\overline{B}_\rho(a)$ — gives

$$|F(x, y) - F(x', y)| \le \sqrt{n}\, \mu \, |x - x'|, \qquad y \in \overline{B}_r(b). \tag{A.534}$$

Now, using that $h(x)$ and $h(x')$ are fixed points of $T(x, \cdot)$ and $T(x', \cdot)$,

$$\begin{aligned} |h(x) - h(x')| &= |T(x, h(x)) - T(x', h(x'))| \\ &\le |T(x, h(x)) - T(x, h(x'))| + |T(x, h(x')) - T(x', h(x'))| \\ &\le \frac{1}{2} |h(x) - h(x')| + \|C\| \, |F(x, h(x')) - F(x', h(x'))|, \end{aligned}$$

the last term because $T(x, y) - T(x', y) = -C\,[F(x, y) - F(x', y)]$. Absorbing the first term on the left and inserting Equation (A.534) gives Equation (A.533) with $K = 2\sqrt{n}\, \mu \, \|C\|$. □

Lemma A.291 (h is differentiable, with the stated derivative)**.** *h is differentiable at every $x \in W$ and its differential is Equation (A.525); the right-hand side of Equation (A.525) is continuous on W, so h is C^1.*

Rests on Lemma A.290, Theorem 7.68, and Definition 7.67.

Proof. *Derives Lemma A.291.*

Fix $x \in W$, write $y = h(x)$, and let $s \in \mathbb{R}^m$ be small enough that $x + s \in W$. Put

$$\Delta = h(x+s) - h(x)\,, \qquad \text{so} \qquad |\Delta| \le K\,|s|$$

by Equation (A.533). Since F is C^1 it is differentiable at (x, y) (Theorem 7.68), and its differential is the pair of blocks Equation (A.522) (Definition 7.70); so by Equation (7.45) there is a function ε with $|\varepsilon(s,\Delta)| \,/\, (|s| + |\Delta|) \longrightarrow 0$ and

$$0 = F(x+s,\, y+\Delta) - F(x,y) = D_xF(x,y)\,s + D_yF(x,y)\,\Delta + \varepsilon(s,\Delta)\,, \tag{A.535}$$

both values of F vanishing by Equation (A.524). Because $|\Delta| \le K\,|s|$, the remainder obeys $|\varepsilon| \le (1+K)\,|s| \cdot |\varepsilon| \,/ (|s| + |\Delta|)$, so it is $o(|s|)$. The matrix $D_yF(x,y)$ is invertible by assertion (5) of Lemma A.287, so solving Equation (A.535) for Δ,

$$\Delta = -\,[D_yF(x,y)]^{-1}\,D_xF(x,y)\,s - [D_yF(x,y)]^{-1}\,\varepsilon\,. \tag{A.536}$$

The second term is $o(|s|)$ as well, the matrix being a fixed one. Hence $h(x+s) - h(x) - \left(-\,[D_yF]^{-1}\,D_xF\right) s = o(|s|)$, which by Definition 7.67 says that h is differentiable at x with the differential Equation (A.525); the differential being unique, no other candidate exists.

Continuity: the entries of $[D_yF]^{-1}$ are, by Cramer's rule (Chapter 5), quotients of polynomials in the entries of D_yF by the determinant, which does not vanish on $\overline{B}_\rho(a) \times \overline{B}_r(b)$; those entries are continuous, and so is h (Lemma A.290), so the composite $x \longmapsto -\,[D_yF(x,h(x))]^{-1}\,D_xF(x,h(x))$ is continuous. By Definition 7.66, h is C^1. □

Proof of Theorem A.286. *Derives Theorem A.286.*

Corollary A.289 supplies W, V, h, Equation (A.524) and the uniqueness; Lemma A.291 supplies Equation (A.525) and the class C^1. It remains to raise the smoothness. Write

$$Dh(x) = G(x,h(x))\,, \qquad G(x,y) = -\,[D_yF(x,y)]^{-1}\,D_xF(x,y)\,. \tag{A.537}$$

If F is C^k then D_xF and D_yF are C^{k-1}, and so is $[D_yF]^{-1}$, its entries being quotients of polynomials in C^{k-1} functions with non-vanishing denominator; hence G is C^{k-1}. We prove by induction on j that h is C^j for $1 \le j \le k$. The case $j = 1$ is Lemma A.291. Suppose h is C^j with $j < k$. Then $x \longmapsto (x, h(x))$ is C^j, and G is C^{k-1} with $k - 1 \ge j$; a composition of a C^j map with a C^j map is C^j, because by the chain rule Equation (7.52) each first partial derivative of the composite is a sum of products of first partial derivatives of the two factors, and an induction on j carries this to order j. So Dh is C^j by Equation (A.537), that is, h is C^{j+1}. After $k-1$ steps h is C^k. If F is smooth the argument applies for every k, so h is smooth. □

A.28.5 The two consequences the treatise uses

Corollary A.292 (Inverse function theorem)**.** *Let $U \subseteq \mathbb{R}^n$ be open, $f : U \longrightarrow \mathbb{R}^n$ of class C^k with $k \ge 1$, and $\alpha \in U$ with $\det [Df(\alpha)] \neq 0$. Then there are open sets $U_0 \subseteq U$ with $\alpha \in U_0$ and $V_0 \subseteq \mathbb{R}^n$ with $f(\alpha) \in V_0$ such that f maps U_0 bijectively onto V_0, the inverse $g : V_0 \longrightarrow U_0$ is C^k, and*

$$Dg(v) = \left[Df(g(v))\right]^{-1}\,. \tag{A.538}$$

Rests on Theorem A.286 and Proposition 7.72.

Proof. *Derives Corollary A.292.*

Apply Theorem A.286 to

$$\Phi : \mathbb{R}^n \times U \longrightarrow \mathbb{R}^n , \qquad \Phi(v,u) = v - f(u) ,$$

at the point $(f(\alpha),\alpha)$, where $\Phi = 0$ and $D_u\Phi = -Df(\alpha)$ is invertible. It supplies open sets $V_0 \ni f(\alpha)$ and $V \ni \alpha$ and a C^k map $g : V_0 \longrightarrow V$ whose graph is the zero set of Φ in $V_0 \times V$; that is,

$$\text{for } (v,u) \in V_0 \times V : \qquad f(u) = v \iff u = g(v) . \tag{A.539}$$

Put $U_0 = V \cap f^{-1}(V_0)$, open because f is continuous, and note $\alpha \in U_0$. If $u \in U_0$ then $(f(u),u) \in V_0 \times V$, so $u = g(f(u))$ by Equation (A.539): f is injective on U_0. If $v \in V_0$ then $g(v) \in V$ and $f(g(v)) = v \in V_0$, so $g(v) \in U_0$ and $v \in f(U_0)$: f maps U_0 onto V_0, with g as its two-sided inverse. Finally, differentiating $g \circ f = \mathrm{id}$ on U_0 by the chain rule Equation (7.52) gives $Dg(f(u))Df(u) = \mathbb{1}$, which is Equation (A.538) at $v = f(u)$. □

Corollary A.293 (Solving one scalar equation for one coordinate)**.** *Let $A \subseteq \mathbb{R}^n$ be open, $F : A \longrightarrow \mathbb{R}$ of class C^k with $k \geq 1$, and let $P \in A$ satisfy $F(P) = c$ and $\partial F/\partial x^n(P) \neq 0$. Write $P = (P', P^n)$ with $P' \in \mathbb{R}^{n-1}$. Then there are an open set $W \subseteq \mathbb{R}^{n-1}$ containing P' and an open interval I containing P^n, with $W \times I \subseteq A$, and a C^k function $\varsigma : W \longrightarrow I$ with $\varsigma(P') = P^n$ such that, inside $W \times I$,*

$$F\left(x^1,\dots,x^n\right) = c \quad \iff \quad x^n = \varsigma\left(x^1,\dots,x^{n-1}\right) , \tag{A.540}$$

and

$$\frac{\partial \varsigma}{\partial x^i}\left(x'\right) = -\left.\frac{\partial F/\partial x^i}{\partial F/\partial x^n}\right|_{(x',\varsigma(x'))} , \qquad i = 1,\dots,n-1 . \tag{A.541}$$

Rests on Theorem A.286.

Proof. *Derives Corollary A.293.*

Apply Theorem A.286 with $m = n-1$, target dimension 1, and $G(x',x^n) = F(x',x^n) - c$ at the point (P',P^n): $G = 0$ there, and D_yG is the 1×1 matrix $\partial F/\partial x^n(P) \neq 0$, whose determinant is itself. The theorem returns W, an open ball $V \subseteq \mathbb{R}$ — that is, an open interval I — and ς of class C^k with Equation (A.524), which is Equation (A.540). Equation (A.525) reads, for 1×1 blocks, $D\varsigma = -\left(\partial F/\partial x^n\right)^{-1}\left(\partial F/\partial x^1,\dots,\partial F/\partial x^{n-1}\right)$, which is Equation (A.541). □

Example A.294 (The sphere, made explicit)**.** Take $n = 3$ and $F(x,y,z) = x^2 + y^2 + z^2$, $c = R^2 > 0$, the level set of Example 13.61. At a point of the sphere with $z \neq 0$ one has $\partial F/\partial z = 2z \neq 0$, so Corollary A.293 applies with the whole of $\mathbb{R}^3$ as A; the function it produces is the one the chapter writes down by hand, $\varsigma(x,y) = +\sqrt{R^2 - x^2 - y^2}$ on the northern hemisphere, and Equation (A.541) returns $\partial\varsigma/\partial x = -x/z$ and $\partial\varsigma/\partial y = -y/z$, which is what differentiating the explicit square root gives. The point of the corollary is that the same conclusion holds when no explicit solution can be written: at a point of the sphere with $z = 0$ one of the other two partial derivatives is non-zero, and the corollary — applied after relabelling the coordinates — produces a graph there too. If all three partial derivatives could vanish on the level set, no such chart would be guaranteed; that is exactly the regularity hypothesis of Theorem 13.59, and on the sphere it fails only at the origin, which is not on the level set.

Rests on Corollary A.293 and Example 13.61.

A.28.6 Reading the result

Remark A.295 (Nothing is quoted here). This section imports no theorem that the treatise does not prove. In particular it does not use a fixed-point theorem: the Banach contraction principle is not in this book, and rather than quote it the iteration Equation (A.530) is run explicitly and its convergence read off from the geometric estimate Equation (A.532) and the Cauchy criterion Theorem 7.8 — the same device, and the same estimate, by which Theorem 9.8 constructs the solution of an initial value problem. The only other analytic inputs are the mean value theorem Theorem 7.35, used twice to convert a bound on a derivative into a Lipschitz bound; the theorem that a C^1 map is differentiable (Theorem 7.68), used once to expand F to first order in Equation (A.535); and the chain rule Proposition 7.72. From linear algebra, only the invertibility criterion $\det \neq 0$ and Cramer's rule are used.

Remark A.296 (Both hypotheses are needed, and the conclusion is local). The invertibility of $D_y F(a,b)$ cannot be dropped. On $\mathbb{R}\times\mathbb{R}$, $F(x,y)=y^2-x$ has $F(0,0)=0$ and $\partial F/\partial y(0,0)=0$; the zero set near the origin is the parabola $x=y^2$, which is not the graph of any function of x — two solutions for $x>0$, none for $x<0$. The conclusion is also irreducibly local: $F(x,y)=x^2+y^2-1$ on $\mathbb{R}\times\mathbb{R}$ satisfies the hypotheses at $(0,1)$, and the implicit function $\varsigma(x)=\sqrt{1-x^2}$ exists only for $|x|<1$ and cannot be continued past $x=\pm1$, where $\partial F/\partial y$ vanishes. Finally, the smoothness of h can be no better than that of F: taking $m=n=1$ and $F(x,y)=y-\phi(x)$ with ϕ of class C^k but not C^{k+1} makes $h=\phi$.
Rests on Theorem A.286.

Remark A.297. Appendix A.28 discharges the derivation owed at Remark 13.60 of Chapter 13. The statement used there is Corollary A.293: in the proof of the regular value theorem Theorem 13.59 a chart is chosen in which $\partial F/\partial x^n \neq 0$ at the point, and Equation (A.540) is exactly the assertion Equation (13.146) that the level set meets the coordinate box in the graph of a smooth h — from which the chapter reads off the chart $u \longmapsto (u, h(u))$ of the hypersurface. The regular value theorem in turn underlies Example 13.61 and Definition 13.58 and every later construction of a submanifold by one equation. The inverse function theorem Corollary A.292, proved here as a companion, is what licenses a change of chart whenever a Jacobian determinant is non-zero, and is used in that role in Appendix A.29.

A.29 The General Stokes Theorem for Differential Forms

This appendix proves the theorem that Equation (13.231) of Chapter 13 is an instance of: for a compact oriented n-manifold M with boundary and a smooth $(n-1)$-form ω on it,

$$\int_M \mathrm{d}\omega = \int_{\partial M} \omega\,. \tag{A.542}$$

It is the single statement of which Green's theorem (Theorem 7.99), the classical Stokes theorem (Theorem 7.100), Gauss's theorem (Theorem 7.101) and the antisymmetric-tensor form Equation (13.231) are the low-dimensional faces, and Remark 7.102 names it as the formulation the treatise was still owed.

There is no circularity to create. Chapter 7 proves its three classical theorems directly from the fundamental theorem of calculus over simple regions and does not lean on the present section; the present section does not use them. What it does use from Chapter 7 is the multiple integral (Definition 7.93), the three properties of it collected as declared inputs in Remark 7.96 — integrability of a continuous function, additivity, and reduction to an iterated integral — and the fundamental theorem of calculus (Theorem 7.43). From Chapter 13 it uses the whole exterior calculus: Definitions 13.98, 13.102 and 13.103, Proposition 13.100, and Equation (13.246). From this appendix it uses the pullback of a k-form (Definition A.70 and Lemma A.71) and the inverse function theorem (Corollary A.292). One input is *quoted*,

the change-of-variables formula for multiple integrals; it is stated precisely as Theorem A.303 and discussed in Remark A.316.

A.29.1 Manifolds with boundary

Definition A.298 (Half-space; smoothness on it)**.** The *closed half-space* is

$$\mathbb{H}^n = \left\{ x = (x^1, \dots, x^n) \in \mathbb{R}^n \mid x^n \le 0 \right\}, \qquad \partial\mathbb{H}^n = \{ x \in \mathbb{H}^n \mid x^n = 0 \}, \tag{A.543}$$

carrying the subspace topology of $\mathbb{R}^n$ (Definition 6.1). A map defined on a subset $S \subseteq \mathbb{H}^n$ that is open in $\mathbb{H}^n$ is called *smooth* if it extends to a smooth map on some open subset of $\mathbb{R}^n$ containing S; all partial derivatives are then defined at points of $\partial\mathbb{H}^n$ as well, and take there the values of the one-sided limits, so that every identity of the differential calculus of Section 7.10.2 continues to hold on S.

Rests on Definitions 6.1 and 7.66.

Definition A.299 (Manifold with boundary)**.** An *n-manifold with boundary* is defined exactly as in Definition 13.48, with one change: the charts $\varphi : U \longrightarrow \varphi(U)$ of the atlas (Definition 13.44) are homeomorphisms onto subsets of $\mathbb{H}^n$ open in $\mathbb{H}^n$, rather than onto open subsets of $\mathbb{R}^n$, and the transition maps are smooth in the sense of Definition A.298. A point $P \in M$ is a *boundary point* if $\varphi(P) \in \partial\mathbb{H}^n$ for some chart around it, and an *interior point* otherwise; the set of boundary points is written ∂M.

Rests on Definitions 13.44, 13.48 and A.298.

Lemma A.300 (The boundary is well defined, and is a manifold)**.** *No point is a boundary point in one chart and an interior point in another. Moreover ∂M, with the charts obtained by dropping the last coordinate of a boundary chart, is an $(n-1)$-manifold without boundary.*

Rests on Definition A.299 and Corollary A.292.

Proof. *Derives Lemma A.300.*

Suppose P had charts φ on U and ψ on V with $z_0 = \varphi(P)$ satisfying $z_0^n < 0$ and $\psi(P) \in \partial\mathbb{H}^n$. Because $z_0^n < 0$, the set $\varphi(U \cap V)$ contains an open subset $O \subseteq \mathbb{R}^n$ around z_0. Write $\tau = \psi \circ \varphi^{-1}$ on O and let ς be a smooth extension to an open subset of $\mathbb{R}^n$, as in Definition A.298, of the transition map $\varphi \circ \psi^{-1}$ near $\psi(P)$. Shrinking O so that $\tau(O)$ lies in the domain of $\varphi \circ \psi^{-1}$, we have $\varsigma \circ \tau = \mathrm{id}$ on O, so the chain rule Equation (7.52) gives $D\varsigma(\tau(z_0))D\tau(z_0) = \mathbb{1}$ and $\det[D\tau(z_0)] \neq 0$. By the inverse function theorem Corollary A.292, τ maps some neighbourhood of z_0 onto an *open* subset of $\mathbb{R}^n$ containing $\tau(z_0) = \psi(P) \in \partial\mathbb{H}^n$. Every open subset of $\mathbb{R}^n$ containing a point of $\partial\mathbb{H}^n$ contains points with $x^n > 0$, which do not lie in $\mathbb{H}^n$ — contradicting that ψ takes values in $\mathbb{H}^n$.

For the second claim, let $\varphi = (x^1, \dots, x^n)$ be a chart with $\varphi(U) \cap \partial\mathbb{H}^n \neq \varnothing$. By the first claim, $\varphi(U \cap \partial M) = \varphi(U) \cap \partial\mathbb{H}^n$, which is an open subset of $\mathbb{R}^{n-1} = \partial\mathbb{H}^n$, and $\varphi' = (x^1, \dots, x^{n-1})$ is a homeomorphism of $U \cap \partial M$ onto it. Two such charts are smoothly related, because a transition map of M carries $\partial\mathbb{H}^n$ into $\partial\mathbb{H}^n$ — by the first claim again — so its first $n-1$ components restricted to $x^n = 0$ are the transition map of the boundary charts, and they are smooth. Hence Definition 13.44 is met with model space $\mathbb{R}^{n-1}$. □

A.29.2 Orientation and the integral of an n-form

Lemma A.301 (Transformation of the top form)**.** *Let $\varphi = (x^1, \dots, x^n)$ and $\psi = (y^1, \dots, y^n)$ be two charts on a common domain. Then*

$$\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n = \det\left[\frac{\partial x^i}{\partial y^j}\right] \mathrm{d}y^1 \wedge \cdots \wedge \mathrm{d}y^n . \tag{A.544}$$

Consequently, if an n-form reads $f\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n$ and $g\,\mathrm{d}y^1 \wedge \cdots \wedge \mathrm{d}y^n$ in the two charts, then $g = f \det[\partial x/\partial y]$.

Rests on Definition 13.102, Equation (13.241), and Proposition 13.100.

Proof. *Derives Lemma A.301.*

By the chain rule, $\mathrm{d}x^i = (\partial x^i/\partial y^j)\,\mathrm{d}y^j$. Substituting into the left-hand side and expanding the wedge, every term in which two of the y-indices coincide vanishes (Equation (13.241)), so only the $n!$ terms indexed by a permutation ρ of $(1,\dots,n)$ survive, and reordering $\mathrm{d}y^{\rho(1)}\wedge\cdots\wedge\mathrm{d}y^{\rho(n)}$ into increasing order costs $\operatorname{sgn}\rho$:

$$\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n=\left(\sum_{\rho}\operatorname{sgn}(\rho)\,\frac{\partial x^1}{\partial y^{\rho(1)}}\cdots\frac{\partial x^n}{\partial y^{\rho(n)}}\right)\mathrm{d}y^1\wedge\cdots\wedge\mathrm{d}y^n\,,$$

and the bracket is the Leibniz expansion of the determinant (Chapter 5). The last sentence follows because Λ^n is one-dimensional (Equation (13.235)). □

Definition A.302 (Orientation)**.** An atlas of M is *oriented* if every transition map between two of its charts has everywhere positive Jacobian determinant, $\det[\partial x/\partial y]>0$. M is *orientable* if it admits such an atlas, and an *oriented manifold* is a manifold together with a maximal oriented atlas; a chart belongs to it, and is called *positively oriented*, if its transitions with every chart of the atlas have positive Jacobian determinant.

Rests on Definition 13.44 and Lemma A.301.

Theorem A.303 (Change of variables for multiple integrals; quoted)**.** *Let $\tau:O\longrightarrow O'$ be a diffeomorphism between open subsets of $\mathbb{R}^n$ and let u be continuous with compact support in O'. Then $u\circ\tau\,|\det D\tau|$ has compact support in O and*

$$\int_{O'}u(z)\,\mathrm{d}z^1\cdots\mathrm{d}z^n=\int_{O}u(\tau(w))\,|\det D\tau(w)|\,\mathrm{d}w^1\cdots\mathrm{d}w^n\,,\tag{A.545}$$

both integrals being multiple integrals in the sense of Definition 7.93.

Rests on Definition 7.93 and Remark 7.96.

Definition A.304 (Integral over a chart)**.** Let M be an oriented n-manifold with boundary, let $\varphi=(x^1,\dots,x^n)$ be a positively oriented chart on U, and let α be a continuous n-form on M whose support is a compact subset of U. Writing $\alpha=f\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n$ on U, set

$$\int_M\alpha=\int_{\varphi(U)}\left(f\circ\varphi^{-1}\right)(x)\,\mathrm{d}x^1\cdots\mathrm{d}x^n\,.\tag{A.546}$$

Rests on Definitions 7.93 and A.302 and Proposition 13.100.

Lemma A.305 (The chart integral is well defined)**.** *The value Equation (A.546) does not depend on which positively oriented chart containing the support is used.*

Rests on Definition A.304, Theorem A.303, and Lemma A.301.

Proof. *Derives Lemma A.305.*

Let $\psi=(y^1,\dots,y^n)$ be a second positively oriented chart containing $\operatorname{supp}\alpha$, and let $\tau=\varphi\circ\psi^{-1}$, a bijection between the images of the common domain.

Reduction to open sets. By Lemma A.300, τ carries interior points to interior points and boundary points to boundary points, so it restricts to a diffeomorphism between the two *open* subsets of $\mathbb{R}^n$ obtained by deleting the face $x^n=0$; and deleting that face changes neither multiple integral, because a bounded portion of a hyperplane contributes nothing to an upper–lower Darboux gap — the argument of Remark 7.96 for the graph of a continuous function, applied to the constant function 0. Both sides of the claimed equality may therefore be computed over the open sets, where Theorem A.303 applies as stated.

By Lemma A.301 the component functions are related by $g=f\,\det[\partial x/\partial y]=f\,\det D\tau$, and $\det D\tau>0$ by Definition A.302, so $|\det D\tau|=\det D\tau$. Then Equation (A.545) with $u=f\circ\varphi^{-1}$ gives

$$\int_{\varphi(U)}\left(f\circ\varphi^{-1}\right)\mathrm{d}x^1\cdots\mathrm{d}x^n=\int_{\psi(U)}\left(f\circ\varphi^{-1}\circ\tau\right)\det D\tau\,\mathrm{d}y^1\cdots\mathrm{d}y^n=\int_{\psi(U)}\left(g\circ\psi^{-1}\right)\mathrm{d}y^1\cdots\mathrm{d}y^n\,,$$

since $f\circ\varphi^{-1}\circ\tau=f\circ\psi^{-1}$. □

A.29.3 Bump functions and partitions of unity

The construction below is elementary and is carried out in full, because the treatise does not have it: nothing here is quoted.

Lemma A.306 (The standard smooth bump)**.** *The function*

$$\lambda(t) = \mathrm{e}^{-1/t} \quad (t > 0)\,, \qquad \lambda(t) = 0 \quad (t \le 0) \tag{A.547}$$

is smooth on $\mathbb{R}$*. Consequently, for* $0 < a < b$*,*

$$\beta(t) = \frac{\lambda(b-t)}{\lambda(b-t) + \lambda(t-a)} \tag{A.548}$$

is smooth, takes values in $[0,1]$*, equals* 1 *for* $t \le a$ *and* 0 *for* $t \ge b$*; and for* $0 < r_1 < r_2$ *the function*

$$\chi(x) = \beta\left(|x|^2\right), \qquad a = r_1^2\,, \quad b = r_2^2\,, \tag{A.549}$$

is smooth on $\mathbb{R}^n$*, equals* 1 *on* $\overline{B}_{r_1}(0)$*, and vanishes off* $B_{r_2}(0)$*.*

Rests on Definition 7.53, Theorem 7.37, and Proposition 7.72.

Proof. *Derives Lemma A.306.*

Smoothness of λ*.* On $t > 0$ an induction gives $\lambda^{(k)}(t) = P_k(1/t)\,\mathrm{e}^{-1/t}$ with P_k a polynomial: this holds for $k = 0$ with $P_0 = 1$, and differentiating $P_k(1/t)\mathrm{e}^{-1/t}$ gives $\left[-t^{-2}P_k'(1/t) + t^{-2}P_k(1/t)\right]\mathrm{e}^{-1/t}$, of the same shape with $P_{k+1}(s) = s^2\left(P_k(s) - P_k'(s)\right)$. On $t < 0$ every derivative is 0. At $t = 0$ we show by induction that $\lambda^{(k)}(0)$ exists and is 0. Assume it for k; the left difference quotient is 0, and the right one is $P_k(1/t)\mathrm{e}^{-1/t}/t$. Put $s = 1/t \to +\infty$; the quotient is $s\,P_k(s)\,\mathrm{e}^{-s}$, and for every N the exponential series (Equation (7.31)) gives $\mathrm{e}^s > s^N/N!$, i.e. $\mathrm{e}^{-s} < N!\,s^{-N}$, so choosing N larger than $1 + \deg P_k$ makes $s\,P_k(s)\mathrm{e}^{-s} \to 0$. Hence $\lambda^{(k+1)}(0) = 0$. Each $\lambda^{(k)}$ is then continuous at 0 by the same estimate, so λ is smooth.

β. The denominator of Equation (A.548) never vanishes: for $t < b$ the first term is positive, and for $t \ge b > a$ the second is. So β is smooth, and it lies in $[0,1]$ because both terms are non-negative. For $t \le a$ the second term vanishes and $\beta = 1$; for $t \ge b$ the first vanishes and $\beta = 0$.

χ. $x \longmapsto |x|^2$ is a polynomial, hence smooth, so χ is smooth by the chain rule Proposition 7.72; $|x| \le r_1$ gives $|x|^2 \le a$ and $\chi = 1$, and $|x| \ge r_2$ gives $\chi = 0$. □

Lemma A.307 (Partition of unity on a compact manifold)**.** *Let* M *be a compact* n*-manifold with boundary and* $\{U_a\}$ *an open cover of* M*. Then there are finitely many smooth functions* $\rho_1, \dots, \rho_N : M \longrightarrow [0,1]$ *such that each* $\operatorname{supp}\rho_i$ *is a compact set contained in the domain of a single chart which is itself contained in some* U_a*, and*

$$\sum_{i=1}^{N} \rho_i(P) = 1 \qquad \text{for every } P \in M\,. \tag{A.550}$$

Rests on Lemma A.306 and Definitions 6.9 and A.299.

Proof. *Derives Lemma A.307.*

Let $P \in M$. Choose a chart φ_P on a domain $V_P \ni P$ with V_P contained in some U_a and $\varphi_P(P) = 0$ (translate the chart). Since $\varphi_P(V_P)$ is open in $\mathbb{H}^n$ there is $r_P > 0$ with $\overline{B}_{r_P}(0) \cap \mathbb{H}^n \subseteq \varphi_P(V_P)$. Let χ_P be the function Equation (A.549) with $r_1 = r_P/2$ and $r_2 = r_P$, and set

$$\widetilde{\chi}_P = \chi_P \circ \varphi_P \text{ on } V_P\,, \qquad \widetilde{\chi}_P = 0 \text{ on } M \setminus V_P\,.$$

This is smooth on M: it is smooth on V_P, and it vanishes identically on the open set $M \setminus \varphi_P^{-1}\left(\overline{B}_{r_P}(0) \cap \mathbb{H}^n\right)$, whose complement in M is the compact set $\varphi_P^{-1}\left(\overline{B}_{r_P}(0) \cap \mathbb{H}^n\right) \subseteq$

V_P; the two open sets cover M and the definitions agree on the overlap. The same compact set contains $\operatorname{supp}\widetilde{\chi}_P$.

Put $W_P = \varphi_P^{-1}\left(B_{r_P/2}(0)\cap\mathbb{H}^n\right)$, an open neighbourhood of P on which $\widetilde{\chi}_P = 1$. The family $\{W_P\}$ is an open cover of the compact M (Definition 6.9), so finitely many $W_{P_1},\dots,W_{P_N}$ cover it. Then $S=\sum_i \widetilde{\chi}_{P_i}$ is smooth and satisfies $S\geq 1$ everywhere, since each point lies in some W_{P_i} where the corresponding term is 1 and no term is negative. Setting $\rho_i = \widetilde{\chi}_{P_i}/S$ gives smooth functions with the stated supports and Equation (A.550). □

Definition A.308 (Integral over the manifold)**.** Let M be a compact oriented n-manifold with boundary and α a continuous n-form on M. Choose a partition of unity $\{\rho_i\}$ subordinate to a cover by domains of positively oriented charts (Lemma A.307) and set

$$\int_M \alpha = \sum_{i=1}^{N}\int_M \rho_i\alpha\,, \tag{A.551}$$

each summand being a chart integral (Definition A.304). The value does not depend on the partition: if $\{\sigma_j\}$ is another one, then by linearity of the multiple integral and Lemma A.305 — each $\rho_i\sigma_j\alpha$ being supported in one chart —

$$\sum_i\int_M \rho_i\alpha = \sum_{i,j}\int_M \rho_i\sigma_j\alpha = \sum_j\int_M \sigma_j\alpha\,.$$

Rests on Lemmas A.305 and A.307.

Definition A.309 (The induced orientation of the boundary)**.** Let M be oriented and $\varphi=(x^1,\dots,x^n)$ a positively oriented boundary chart. The *induced orientation* of ∂M is the one in which the restricted chart $\varphi'=(x^1,\dots,x^{n-1})$ of Lemma A.300 is positively oriented when n is odd and negatively oriented when n is even; equivalently, the chart $((-1)^{n-1}x^1,x^2,\dots,x^{n-1})$ is positively oriented. This is consistent: the transition map between two boundary charts has, at $x^n=0$, Jacobian determinant equal to $(\partial x^n/\partial y^n)$ times that of the restricted transition, and $\partial x^n/\partial y^n>0$ because both charts map into $\mathbb{H}^n$ and both send ∂M to $x^n=0$; hence the restricted transitions inherit positive determinants, and multiplying every one of them by the same sign $(-1)^{n-1}$ leaves them positive.

Rests on Definition A.302 and Lemma A.300.

Remark A.310 (Outward normal first). The sign $(-1)^{n-1}$ is not a convention chosen to make Equation (A.542) come out right; it is the *outward-normal-first* rule counted honestly. In a boundary chart the outward direction is that of increasing x^n, since $\mathbb{H}^n=\{x^n\leq 0\}$, so the rule declares $(v_1,\dots,v_{n-1})$ positively oriented in $T_{\partial M}$ exactly when $(\partial_n,v_1,\dots,v_{n-1})$ is positively oriented in T_M. Moving ∂_n from the front of $(\partial_n,\partial_1,\dots,\partial_{n-1})$ to the back costs $n-1$ transpositions, so $(\partial_1,\dots,\partial_{n-1})$ is positively oriented precisely when $(-1)^{n-1}=+1$, which is Definition A.309. For $n=3$ — a surface in space, the case of Equation (13.231) — the sign is $+1$ and the rule reduces to the familiar right-hand relation between the normal of S and the sense of circulation on ∂S.

A.29.4 Statement and the half-space computation

Theorem A.311 (General Stokes theorem)**.** *Let M be a compact oriented n-manifold with boundary, $n\geq 1$, let ∂M carry the induced orientation, and let ω be a smooth $(n-1)$-form on M. Then*

$$\int_M \mathrm{d}\omega = \int_{\partial M}\iota^*\omega\,, \tag{A.552}$$

where $\iota:\partial M\longrightarrow M$ is the inclusion. If $\partial M=\varnothing$ the right-hand side is 0.

Rests on Definitions 13.103, A.308 and A.309.

Lemma A.312 (The computation in one chart)**.** *Let ω be a smooth $(n-1)$-form on $\mathbb{H}^n$ with compact support. Write*

$$\omega = \sum_{i=1}^{n} (-1)^{i-1} f_i \, \mathrm{d}x^1 \wedge \cdots \wedge \widehat{\mathrm{d}x^i} \wedge \cdots \wedge \mathrm{d}x^n \,, \tag{A.553}$$

the hat marking the omitted factor; this is possible, with smooth f_i, because those n products form a basis of Λ^{n-1} (Equation (13.235)). Then

$$\int_{\mathbb{H}^n} \mathrm{d}\omega = \int_{\mathbb{R}^{n-1}} f_n\left(x^1, \ldots, x^{n-1}, 0\right) \mathrm{d}x^1 \cdots \mathrm{d}x^{n-1} = \int_{\partial \mathbb{H}^n} \iota^* \omega \,, \tag{A.554}$$

the last integral taken with the induced orientation Definition A.309. If the support of ω lies in the open half-space $\{x^n < 0\}$, both sides vanish.

Rests on Definition 13.103, Theorem 7.43, and Remark 7.96.

Proof. *Derives Lemma A.312.*

The exterior derivative. Write $\theta_i = \mathrm{d}x^1 \wedge \cdots \wedge \widehat{\mathrm{d}x^i} \wedge \cdots \wedge \mathrm{d}x^n$. Carrying $\mathrm{d}x^i$ leftwards past the $i-1$ factors $\mathrm{d}x^1, \ldots, \mathrm{d}x^{i-1}$ using Equation (13.241),

$$\mathrm{d}x^i \wedge \theta_i = (-1)^{i-1} \, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \,. \tag{A.555}$$

By the Leibniz rule Equation (13.246) and $\mathrm{d}(\mathrm{d}x^j) = 0$,

$$\mathrm{d}\omega = \sum_i (-1)^{i-1} \, \mathrm{d}f_i \wedge \theta_i = \sum_{i,j} (-1)^{i-1} \partial_j f_i \, \mathrm{d}x^j \wedge \theta_i \,,$$

and $\mathrm{d}x^j \wedge \theta_i = 0$ unless $j = i$, since otherwise $\mathrm{d}x^j$ is repeated. With Equation (A.555),

$$\mathrm{d}\omega = \left(\sum_{i=1}^{n} \partial_i f_i \right) \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \,. \tag{A.556}$$

The integral. Choose $R > 0$ so large that $\operatorname{supp}\omega$ is contained in the box $Q = [-R, R]^{n-1} \times [-R, 0]$, and note that each f_i vanishes identically outside it. By Equation (A.556), Definition A.304 and additivity, the left-hand side of Equation (A.554) is $\sum_i \int_Q \partial_i f_i$. Reduce each term to an iterated integral, integrating first in x^i (Remark 7.96) and applying the fundamental theorem of calculus Theorem 7.43. For $i < n$,

$$\int_{-R}^{R} \partial_i f_i \, \mathrm{d}x^i = f_i\big|_{x^i=-R}^{x^i=+R} = 0 \,,$$

both endpoint values lying outside Q; hence those $n-1$ terms vanish. For $i = n$ the x^n-integration runs only up to 0:

$$\int_{-R}^{0} \partial_n f_n \, \mathrm{d}x^n = f_n\left(x^1, \ldots, x^{n-1}, 0\right) - f_n\left(x^1, \ldots, x^{n-1}, -R\right) = f_n\left(x^1, \ldots, x^{n-1}, 0\right) \,,$$

which gives the first equality in Equation (A.554). If $\operatorname{supp}\omega \subseteq \{x^n < 0\}$ then $f_n(\cdot, 0) = 0$ and the expression vanishes.

The boundary integral. The inclusion is $\iota(x^1, \ldots, x^{n-1}) = (x^1, \ldots, x^{n-1}, 0)$, so by Equation (A.190) $\iota^* \mathrm{d}x^j = \mathrm{d}x^j$ for $j < n$ while $\iota^* \mathrm{d}x^n = \mathrm{d}\,(x^n \circ \iota) = 0$. By Equation (A.187) the pullback distributes over the wedge, so every θ_i with $i < n$ — each of which still contains the factor $\mathrm{d}x^n$ — pulls back to zero, and

$$\iota^* \omega = (-1)^{n-1} \, f_n\left(x^1, \ldots, x^{n-1}, 0\right) \, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^{n-1} \,. \tag{A.557}$$

The induced orientation is $(-1)^{n-1}$ times the one in which $(x^1, \dots, x^{n-1})$ is positively oriented (Definition A.309), and reversing an orientation reverses the sign of the integral Equation (A.546): replacing x^1 by $-x^1$ multiplies the component function by -1 (Lemma A.301, the Jacobian determinant of the reflection being -1), while the multiple integral over the reflected region is unchanged by Equation (A.545), whose factor $|\det D\tau|$ is 1 for a reflection. Hence

$$\int_{\partial \mathbb{H}^n} \iota^* \omega = (-1)^{n-1} \int_{\mathbb{R}^{n-1}} (-1)^{n-1} f_n\left(x', 0\right) \mathrm{d}x^1 \cdots \mathrm{d}x^{n-1} = \int_{\mathbb{R}^{n-1}} f_n\left(x', 0\right) \mathrm{d}x^1 \cdots \mathrm{d}x^{n-1} ,$$

which is the second equality in Equation (A.554). □

A.29.5 Proof of the theorem

Proof of Theorem A.311. *Derives Theorem A.311.*

Cover M by domains of positively oriented charts and let $\{\rho_1, \dots, \rho_N\}$ be a subordinate partition of unity (Lemma A.307). Since $\sum_i \rho_i = 1$ is constant, $\sum_i \mathrm{d}\rho_i = \mathrm{d}(1) = 0$, so by the Leibniz rule Equation (13.246)

$$\sum_{i=1}^{N} \mathrm{d}\left(\rho_i \omega\right) = \sum_i \mathrm{d}\rho_i \wedge \omega + \left(\sum_i \rho_i\right) \mathrm{d}\omega = \mathrm{d}\omega \,. \tag{A.558}$$

Each $\rho_i \omega$ is a smooth $(n-1)$-form whose support is a compact subset of the domain U_i of a positively oriented chart φ_i; transporting it by φ_i and extending by zero gives a smooth compactly supported $(n-1)$-form on the whole of $\mathbb{H}^n$, smooth because the two definitions — the transported one on the open set $\varphi_i(U_i)$, and zero on the complement of the compact support — agree on the overlap of two open sets covering $\mathbb{H}^n$. The transport carries the two sides of Equation (A.559) to the two sides of Equation (A.554): the chart integral Equation (A.546) is by definition the multiple integral of the chart representative, and pullback commutes with d (Equation (A.188)), so the chart representative of $\mathrm{d}(\rho_i\omega)$ is the exterior derivative of the chart representative of $\rho_i\omega$. Hence Lemma A.312 applies and gives

$$\int_M \mathrm{d}\left(\rho_i \omega\right) = \int_{\partial M} \iota^* \left(\rho_i \omega\right) \,, \tag{A.559}$$

where the right-hand side is 0 when U_i misses ∂M, that being the last clause of Lemma A.312. Summing Equation (A.559) over i, using Equation (A.558) on the left and, on the right, that $\{\iota^*\rho_i\}$ restricts to a partition of unity on ∂M — $\sum_i \rho_i = 1$ there too, and each $\iota^*\rho_i$ is supported in a boundary chart — gives Equation (A.552) by Definition A.308. If $\partial M = \varnothing$, every chart misses the boundary and every term on the right vanishes. □

A.29.6 The instance stated in the chapter

Corollary A.313 (Stokes' theorem for an antisymmetric tensor field)**.** *Let $S \subset E_3$ be a compact oriented two-dimensional submanifold with boundary ∂S, carrying the induced orientation, and let H_i be a smooth covector field defined near S, with rotor tensor*

$$H_{ij} = \partial_i H_j - \partial_j H_i \,. \tag{A.560}$$

Then, with $\mathrm{d}f^{ij}$ the antisymmetric surface element of S,

$$\frac{1}{2} \int_S H_{ij} \, \mathrm{d}f^{ij} = \oint_{\partial S} H_i \, \mathrm{d}x^i \,, \tag{A.561}$$

which is Equation (13.231).

Rests on Theorem A.311, Equation (13.232), and Definition 7.94.

Proof. *Derives Corollary A.313.*

Let $H = H_i \, \mathrm{d}x^i$ be the one-form with the given components. By Definition 13.103,

$$\mathrm{d}H = \partial_j H_i \, \mathrm{d}x^j \wedge \mathrm{d}x^i = \frac{1}{2} \left(\partial_i H_j - \partial_j H_i \right) \mathrm{d}x^i \wedge \mathrm{d}x^j = \frac{1}{2} H_{ij} \, \mathrm{d}x^i \wedge \mathrm{d}x^j \,, \tag{A.562}$$

the middle step by antisymmetrizing the coefficient against $\mathrm{d}x^j \wedge \mathrm{d}x^i = -\mathrm{d}x^i \wedge \mathrm{d}x^j$ (Equation (13.241)). Theorem A.311 applied to $M = S$ and $\omega = \iota_S^* H$, the restriction of H to S, gives

$$\int_S \iota_S^* \left(\mathrm{d}H \right) = \oint_{\partial S} \iota_{\partial S}^* H \,, \tag{A.563}$$

where pullback and d were exchanged by Equation (A.188).

It remains to recognize the two sides. Let $\boldsymbol{r} : D \longrightarrow S$ be a positively oriented parametrization of a piece of S, with parameters (u, v). The two-form $\mathrm{d}x^i \wedge \mathrm{d}x^j$ has components $\delta^i{}_a \delta^j{}_b - \delta^i{}_b \delta^j{}_a$ (Definition 13.102), so Equation (A.186) gives

$$\iota_S^* \left(\mathrm{d}x^i \wedge \mathrm{d}x^j \right) = \left(\partial_u r^i \, \partial_v r^j - \partial_v r^i \, \partial_u r^j \right) \mathrm{d}u \wedge \mathrm{d}v \,. \tag{A.564}$$

That bracket is the antisymmetric surface element. Indeed, Remark 13.97 writes $\mathrm{d}f^{ij} = \varepsilon^{ijk} \mathrm{d}f_k$ with $\mathrm{d}f_k$ the oriented surface element Equation (7.108) of Definition 7.94, whose components are $\mathrm{d}f_k = \varepsilon_{klm} \, \partial_u r^l \, \partial_v r^m \, \mathrm{d}u \, \mathrm{d}v$; contracting the two Levi-Civita symbols (Chapter 5),

$$\mathrm{d}f^{ij} = \varepsilon^{ijk} \varepsilon_{klm} \, \partial_u r^l \, \partial_v r^m \, \mathrm{d}u \, \mathrm{d}v = \left(\delta^i{}_l \delta^j{}_m - \delta^i{}_m \delta^j{}_l \right) \partial_u r^l \, \partial_v r^m \, \mathrm{d}u \, \mathrm{d}v \,, \tag{A.565}$$

which is the bracket of Equation (A.564) times $\mathrm{d}u \, \mathrm{d}v$. Hence, by Definition A.304 and Equation (A.562),

$$\int_S \iota_S^* \left(\mathrm{d}H \right) = \frac{1}{2} \iint_D H_{ij} \left(\boldsymbol{r}(u, v) \right) \left(\partial_u r^i \partial_v r^j - \partial_v r^i \partial_u r^j \right) \mathrm{d}u \, \mathrm{d}v = \frac{1}{2} \int_S H_{ij} \, \mathrm{d}f^{ij} \,,$$

and, parametrizing ∂S by arc parameter t, $\iota_{\partial S}^* H = H_i(\boldsymbol{r}(t)) \, \frac{\mathrm{d}r^i}{\mathrm{d}t} \, \mathrm{d}t$, whose integral is the line integral Equation (7.106) of $H_i \mathrm{d}x^i$ over ∂S. Substituting both into Equation (A.563) gives Equation (A.561). □

Remark A.314 (The reading under which the identity is false). Remark 13.97 records that Equation (13.231) holds for the rotor tensor Equation (13.232) and is *false* if H_i is instead read as the dual pseudo-vector $\frac{1}{2} \varepsilon_{ijk} H_{jk}$ of Equation (13.224). The derivation above respects that, and shows where the hypothesis enters: the only step in which the relation between H_{ij} and H_i is used is Equation (A.562), which asserts that the two-form $\frac{1}{2} H_{ij} \mathrm{d}x^i \wedge \mathrm{d}x^j$ is *exact*, with the one-form $H = H_i \mathrm{d}x^i$ as its primitive. Under the dual reading there is no such relation: $\frac{1}{2} H_{ij} \mathrm{d}x^i \wedge \mathrm{d}x^j$ would be an arbitrary two-form and $H_i \mathrm{d}x^i$ an unrelated one-form, so Theorem A.311 says nothing connecting their integrals, and the left-hand side of Equation (13.231) would be a flux and the right-hand side a circulation of the same vector field — two quantities with no reason to agree. Writing $\mathrm{d}f^{ij} = \varepsilon^{ijk} \mathrm{d}f_k$ turns Equation (A.561) into $\int_S (\nabla \times \boldsymbol{H}) \cdot \mathrm{d}\boldsymbol{f} = \oint_{\partial S} \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l}$, the classical statement Theorem 7.100 — which Chapter 7 proves directly and independently, so the two derivations check each other rather than one resting on the other.

Remark A.315 (The SI reading of the three-dimensional instance). The theorem is an identity between integrals of forms and is unit-homogeneous by construction: if H_i carries an SI unit $[H]$ then $H_i \mathrm{d}x^i$ carries $[H] \cdot \mathrm{m}$, H_{ij} carries $[H] / \mathrm{m}$, and $\mathrm{d}f^{ij}$ carries m^2, so both sides of Equation (A.561) carry $[H] \cdot \mathrm{m}$. The physical instance the treatise meets first is the magnetic vector potential: with $H_i = A_i$ in $\mathrm{T\,m}$ and $H_{ij} = \partial_i A_j - \partial_j A_i$ in T, the identity reads

$$\int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{f} = \oint_{\partial S} \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{l} \,, \tag{A.566}$$

both sides in $\mathrm{Wb} = \mathrm{T\,m^2}$: the magnetic flux through a surface equals the circulation of the potential round its edge.

A.29.7 What is quoted, and what is not

Remark A.316 (What is quoted here). Exactly one theorem is assumed without proof: Theorem A.303, the change-of-variables formula for multiple integrals. It is used once, in Lemma A.305, to show that the chart integral Equation (A.546) does not depend on the chart; every later step is built on that one fact about the integral. This treatise does not prove it, and the omission is not new here: Remark 7.102 of Chapter 7 already records the same formula as a quoted input, needed there for the independence of a flux integral from its parametrization, alongside the three properties of the multiple integral collected in Remark 7.96. A citation for it is owed and no entry of the bibliography currently carries it, so it is stated here rather than referenced. Nothing else is quoted: the partition of unity, which is the input a reader would most expect to be imported, is constructed outright in Lemmas A.306 and A.307, and the topology of the boundary is settled in Lemma A.300 from the inverse function theorem proved in Appendix A.28.

Remark A.317 (Where compactness and orientability are used). Compactness enters once, in Lemma A.307, to make the partition of unity *finite* — so that the sum Equation (A.558) has finitely many terms and no convergence question arises. It may be traded for the weaker hypothesis that ω have compact support, at the cost of building a locally finite partition of unity. Orientability enters twice, both times essentially: in Lemma A.305, where $\det D\tau > 0$ is what removes the absolute value from Equation (A.545) and lets the chart integrals agree rather than differ in sign; and in Definition A.309, where the induced orientation is what makes the sign $(-1)^{n-1}$ produced by the half-space computation cancel against the sign in Equation (A.557). On a non-orientable manifold no consistent $\int_M$ of an n-form exists at all, and Equation (A.552) has no meaning.

Remark A.318. Appendix A.29 discharges the derivation owed at Equation (13.231) of Chapter 13, and with it the general formulation that Remark 7.102 of Chapter 7 names as the common source of Theorems 7.99 to 7.101. The chapter's statement is recovered as Corollary A.313 under the hypothesis that Remark 13.97 makes explicit, and the general theorem Theorem A.311 is what Section 13.6 was building towards: the exterior derivative and the boundary operator are adjoint, so that $\mathrm{d}^2 = 0$ (Equation (13.245)) and $\partial\partial = \varnothing$ are the same statement seen from the two sides of Equation (A.552).

A.30 The Poincaré–Birkhoff–Witt Theorem

This appendix proves Theorem 14.7 of Chapter 14: the ordered monomials in a basis of a Lie algebra $\mathfrak{g}$ form a vector-space basis of its universal enveloping algebra $U(\mathfrak{g})$ of Definition 14.5. Only the defining relations Equation (14.12) and the Jacobi identity are assumed. That the ordered monomials *span* $U(\mathfrak{g})$ is a reordering argument and is done first (Appendix A.30.2); the content of the theorem is their linear *independence*, and the whole of it lies in the construction of an action of $U(\mathfrak{g})$ on the polynomial algebra in $\dim \mathfrak{g}$ commuting variables (Appendix A.30.3). The Jacobi identity is exactly what makes that action well defined: it is the statement that the two ways of reordering a triple product agree. The consequences drawn in Appendix A.30.5 — the embedding of $\mathfrak{g}$ in $U(\mathfrak{g})$, and the symmetrization isomorphism between $U(\mathfrak{g})$ and the polynomial algebra on $\mathfrak{g}$ — are what Appendix A.32 uses to count the Casimir operators.

Throughout, $\mathbb{K}$ is the field of scalars ($\mathbb{R}$ or $\mathbb{C}$ in every application in this book; the symbol k is reserved for the invariant tensors of Definition 14.14), $\mathfrak{g}$ is a finite-dimensional

Lie algebra over $\mathbb{K}$ with basis $T_1, \dots, T_M$, $M = \dim \mathfrak{g}$, and structure constants fixed as in Equation (14.11),

$$[T_a, T_b] = C^c{}_{ab}\, T_c\,, \qquad a, b, c = 1, \dots, M\,, \tag{A.567}$$

with the summation convention in force. Nothing below uses the characteristic of $\mathbb{K}$ except where the symmetrization map is introduced, which needs $\mathbb{Q} \subseteq \mathbb{K}$ and is flagged there.

A.30.1 Statement

Theorem A.319 (Poincaré–Birkhoff–Witt)**.** *The ordered monomials*

$$T_1^{n_1} T_2^{n_2} \cdots T_M^{n_M}\,, \qquad n_a \in \mathbb{N} \cup \{0\}\,, \tag{A.568}$$

form a basis of $U(\mathfrak{g})$ as a $\mathbb{K}$-vector space. In particular the linear map $\mathfrak{g} \to U(\mathfrak{g})$ is injective, so that $\mathfrak{g}$ may be regarded as a subspace of $U(\mathfrak{g})$, and no relation holds among the generators beyond those forced by Equation (14.12).

Rests on Definition 14.5 and Equation (14.12).

A.30.2 The filtration, and spanning

Definition A.320 (Filtration by degree)**.** Let $U_n \subseteq U(\mathfrak{g})$ be the span of the products $T_{a_1} T_{a_2} \cdots T_{a_m}$ with $m \le n$, the empty product being the unit. Then

$$\mathbb{K} = U_0 \subseteq U_1 \subseteq U_2 \subseteq \cdots\,, \qquad U_m U_n \subseteq U_{m+n}\,, \qquad \bigcup_n U_n = U(\mathfrak{g})\,, \tag{A.569}$$

the last equality because the T_a generate $U(\mathfrak{g})$ as an algebra with unit.

Rests on Definition 14.5.

Lemma A.321 (The ordered monomials span)**.** *U_n is spanned by the ordered monomials Equation (A.568) of total degree $n_1 + \cdots + n_M \le n$. Consequently the ordered monomials span $U(\mathfrak{g})$.*

Rests on Definition A.320 and Equation (14.12).

Proof. *Derives Lemma A.321.*

Induct on n. For $n = 0$ there is nothing to prove. Let $n \ge 1$ and assume the claim for $n - 1$. It suffices to treat a single word $w = T_{a_1} T_{a_2} \cdots T_{a_n}$ of length exactly n, and we induct on its number of *inversions*, the number of pairs $i < j$ with $a_i > a_j$. If w has no inversion its indices are already nondecreasing and w is an ordered monomial. If it has at least one, some *adjacent* pair is out of order, $a_i > a_{i+1}$; otherwise the sequence would be nondecreasing. Apply Equation (14.12) to that pair alone:

$$w = T_{a_1} \cdots T_{a_{i+1}} T_{a_i} \cdots T_{a_n} + C^c{}_{a_i a_{i+1}}\, T_{a_1} \cdots T_{a_{i-1}}\, T_c\, T_{a_{i+2}} \cdots T_{a_n}\,. \tag{A.570}$$

The first word on the right has length n and exactly one inversion fewer, so it is handled by the inner induction; the second has length $n - 1$ and lies in U_{n-1}, handled by the outer induction. Hence w is a combination of ordered monomials of degree at most n. □

Everything that follows exists to show that this spanning set is *free*. The danger is real and must be stated plainly: nothing seen so far excludes the possibility that repeated use of Equation (A.570) along two different routes returns two different combinations of ordered monomials, which would force a relation among them and could in principle collapse $U(\mathfrak{g})$ — in the extreme case to zero. What rules this out is the existence of *one* representation in which the ordered monomials visibly act independently.

A.30.3 The action on the polynomial algebra

Notation A.322 (Ordered multi-indices). Let $P = \mathbb{K}[x_1, \dots, x_M]$ be the polynomial algebra in M commuting variables. A *multi-index* is a nondecreasing finite sequence $A = (a_1 \le a_2 \le \cdots \le a_m)$ of indices from $\{1, \dots, M\}$; we write $x_A = x_{a_1} x_{a_2} \cdots x_{a_m}$, $|A| = m$, and $x_A = 1$ for the empty sequence. The monomials x_A are a basis of P. For an index a we write

$$a \le A \quad \text{if and only if} \quad a \le a_1 \,, \tag{A.571}$$

that is, if a does not exceed any index occurring in A; the condition is vacuous, hence true, for the empty multi-index. Finally P_n denotes the span of the x_A with $|A| \le n$, so that $P_0 = \mathbb{K}$ and $P_n P_m \subseteq P_{n+m}$.

Proposition A.323 (The polynomial representation)**.** *There exist linear maps $\theta_a : P \to P$, $a = 1, \dots, M$, such that for every multi-index A and all indices a, b*

$$\theta_a \, x_A = x_a x_A \qquad \text{whenever } a \le A \,, \tag{A.572}$$

$$\theta_a \, x_A - x_a x_A \in P_{|A|} \,, \tag{A.573}$$

$$\theta_a \theta_b - \theta_b \theta_a = C^c{}_{ab} \, \theta_c \,. \tag{A.574}$$

The assignment $T_a \mapsto \theta_a$ is therefore a representation of $\mathfrak{g}$ on P.
Rests on Equation (A.567) and Notation A.322.

Proof. *Derives Proposition A.323.*

The maps are constructed on P_n by induction on n. Write (A_n), (B_n), (C_n) for the three conditions restricted as follows: (A_n) and (B_n) are Equations (A.572) and (A.573) for $|A| \le n$, and (C_n) is Equation (A.574) applied to x_A with $|A| \le n-1$ — the largest range in which both sides are defined once θ is known on P_n, since $\theta_b x_A \in P_{|A|+1} \subseteq P_n$ by (B_{n-1}).

Base. $P_0 = \mathbb{K}$; set $\theta_a 1 = x_a$. Then (A_0) holds because $a \le A$ is vacuous for the empty multi-index, (B_0) holds because the difference is 0, and (C_0) is empty.

Extension to P_n. Assume θ defined on P_{n-1} with (A_{n-1}), (B_{n-1}), (C_{n-1}). Let $|A| = n$.

1. If $a \le A$, set $\theta_a x_A = x_a x_A$, as Equation (A.572) demands.

2. If not, let b be the smallest index occurring in A, so that $b < a$, and write $A = (b, B)$ with $b \le B$ and $|B| = n-1$. Then $x_A = x_b x_B = \theta_b x_B$ by (A_{n-1}). Using (B_{n-1}) write

$$\theta_a x_B = x_a x_B + w \,, \qquad w \in P_{n-1} \,, \tag{A.575}$$

and *define*

$$\theta_a x_A := x_a x_A + \theta_b w + C^c{}_{ab} \, \theta_c x_B \,. \tag{A.576}$$

Every term on the right is already defined: w and x_B lie in P_{n-1}.

Condition (A_n) holds by construction. For (B_n): in case 1 the difference is 0; in case 2 it is $\theta_b w + C^c{}_{ab} \theta_c x_B$, and (B_{n-1}) gives $\theta_b w \in P_n$ and $\theta_c x_B \in P_n$.

Before proving (C_n), record what Equation (A.576) says. With $b < a$ and $b \le B$, the monomial $x_a x_B$ is $x_{A'}$ for the multi-index A' obtained by inserting a into B; its smallest index is still b, so $b \le A'$ and $|A'| = n$, and case 1 gives $\theta_b(x_a x_B) = x_b x_a x_B = x_a x_A$, the variables commuting. Hence, by Equation (A.575), $\theta_b \theta_a x_B = x_a x_A + \theta_b w$, and Equation (A.576) reads

$$\theta_a \theta_b x_B = \theta_b \theta_a x_B + C^c{}_{ab} \, \theta_c x_B \qquad \text{whenever } b \le B, \ b < a, \ |B| = n-1 \,, \tag{A.577}$$

using $\theta_b x_B = x_A$ once more. So the commutation relation holds *by fiat* in the ordered case; the work is to show it then holds in general.

Proof of (C_n). Let $|A| = n-1$; the cases $|A| < n-1$ are (C_{n-1}). The claim is antisymmetric in a, b and trivial for $a = b$, since $C^c{}_{aa} = 0$, so assume $a > b$.

Case 1: $b \le A$. This is Equation (A.577) with $B = A$.

Case 2: the relation $b \le A$ *fails.* Let c_0 be the smallest index occurring in A, so $c_0 < b < a$, and write $A = (c_0, B)$ with $c_0 \le B$ and $|B| = n-2$. We first establish the auxiliary relation

$$\theta_a \theta_{c_0} y = \theta_{c_0} \theta_a y + C^f{}_{ac_0}\, \theta_f y \qquad \text{for } y = \theta_b x_B\,. \tag{A.578}$$

Indeed, by (B_{n-1}) write $y = x_b x_B + w'$ with $w' \in P_{n-2}$. The monomial $x_b x_B$ is $x_{A''}$ for the multi-index A'' obtained by inserting b into B; its smallest index is c_0, because $c_0 \le B$ and $c_0 < b$, and $|A''| = n-1$. So Case 1 — just proved, for multi-indices of length $n-1$ — applies to $x_{A''}$ with the pair (a, c_0), and (C_{n-1}) applies to $w' \in P_{n-2}$; adding the two gives Equation (A.578).

Now compute, using $x_A = \theta_{c_0} x_B$:

$$\begin{aligned} \theta_a \theta_b x_A &= \theta_a \theta_b \theta_{c_0} x_B \\ &= \theta_a \left(\theta_{c_0} \theta_b x_B + C^f{}_{bc_0} \theta_f x_B \right) \\ &= \theta_{c_0} \theta_a \theta_b x_B + C^f{}_{ac_0}\, \theta_f \theta_b x_B + C^f{}_{bc_0}\, \theta_a \theta_f x_B\,, \end{aligned} \tag{A.579}$$

the second line by (C_{n-1}) applied to x_B, $|B| = n-2$, and the third by Equation (A.578) applied to $y = \theta_b x_B$. Exchange a and b in Equation (A.579) and subtract:

$$\begin{aligned} \theta_a \theta_b x_A - \theta_b \theta_a x_A = \theta_{c_0} \left(\theta_a \theta_b - \theta_b \theta_a\right) x_B \\ + C^f{}_{ac_0} \left(\theta_f \theta_b - \theta_b \theta_f\right) x_B + C^f{}_{bc_0} \left(\theta_a \theta_f - \theta_f \theta_a\right) x_B\,, \end{aligned} \tag{A.580}$$

where the four mixed terms have been paired so that each parenthesis is a commutator. Every commutator in Equation (A.580) acts on x_B with $|B| = n-2$, so (C_{n-1}) evaluates all three:

$$\theta_a \theta_b x_A - \theta_b \theta_a x_A = C^g{}_{ab}\, \theta_{c_0} \theta_g x_B + C^f{}_{ac_0} C^g{}_{fb}\, \theta_g x_B + C^f{}_{bc_0} C^g{}_{af}\, \theta_g x_B\,. \tag{A.581}$$

The quantity to be matched is, again by (C_{n-1}) on x_B,

$$C^g{}_{ab}\, \theta_g x_A = C^g{}_{ab}\, \theta_g \theta_{c_0} x_B = C^g{}_{ab}\, \theta_{c_0} \theta_g x_B + C^g{}_{ab} C^h{}_{gc_0}\, \theta_h x_B\,. \tag{A.582}$$

The first terms of Equations (A.581) and (A.582) agree, so (C_n) holds if and only if

$$C^f{}_{ac_0} C^g{}_{fb} + C^f{}_{bc_0} C^g{}_{af} = C^f{}_{ab} C^g{}_{fc_0}\,, \tag{A.583}$$

after renaming the summation index in Equation (A.582). Written with brackets, Equation (A.583) is

$$\left[\left[T_a, T_{c_0}\right], T_b\right] + \left[T_a, \left[T_b, T_{c_0}\right]\right] = \left[\left[T_a, T_b\right], T_{c_0}\right]\,, \tag{A.584}$$

which is the Jacobi identity: expanding the right-hand side, $\left[\left[T_a, T_b\right], T_{c_0}\right] = \left[T_a, \left[T_b, T_{c_0}\right]\right] - \left[T_b, \left[T_a, T_{c_0}\right]\right]$, and the second term equals $+\left[\left[T_a, T_{c_0}\right], T_b\right]$ by antisymmetry. This completes the induction, and with it the construction of θ on all of $P = \bigcup_n P_n$. □

Remark A.324 (Where the Jacobi identity was needed, and where it was not). The bookkeeping of Proposition A.323 is elaborate, but the mathematical content sits entirely in Equation (A.584). Conditions Equation (A.572) and Equation (A.573) merely say that θ_a is "multiplication by x_a, up to lower degree", and the definition Equation (A.576) is forced on us: it is the only value compatible with Equation (A.574). Whether that forced value is *consistent* — whether reordering a triple by two different routes gives the same answer — is precisely Equation (A.583). An antisymmetric bracket violating Jacobi would therefore not merely fail to be a Lie algebra: its enveloping algebra would have fewer ordered monomials than expected, because the two routes would impose a relation between them.

A.30.4 Independence, and proof of the theorem

Proof of Theorem A.319. Derives Theorem A.319.

By Proposition A.323 the map $T_a \mapsto \theta_a$ is a representation of $\mathfrak{g}$ on the vector space P, and by Remark 14.6 it extends uniquely to a homomorphism of associative algebras with unit

$$\Phi : U(\mathfrak{g}) \longrightarrow \mathrm{End}(P)\,, \qquad \Phi(T_a) = \theta_a\,. \tag{A.585}$$

(The extension of Remark 14.6 is stated there for a representation on a vector space; that P is infinite-dimensional plays no part, since no topology or trace is involved.)

Evaluate Φ of an ordered monomial on the constant polynomial 1. Reading Equation (A.568) from right to left, the letter T_M acts n_M times on multi-indices all of whose entries are M, then T_{M-1} acts on multi-indices whose entries are all $\geq M-1$, and so on: at every step the acting index is $\leq$ every index already present, so Equation (A.572) applies at every step and

$$\Phi\left(T_1^{n_1} T_2^{n_2} \cdots T_M^{n_M}\right) 1 = x_1^{n_1} x_2^{n_2} \cdots x_M^{n_M}\,. \tag{A.586}$$

Distinct ordered monomials of $U(\mathfrak{g})$ thus have distinct monomials of P as images, and the monomials of a polynomial algebra are linearly independent. Hence a vanishing linear combination of ordered monomials of $U(\mathfrak{g})$ maps to a vanishing linear combination of distinct monomials of P, forcing every coefficient to vanish: the ordered monomials are linearly independent. With Lemma A.321 they are a basis.

The two consequences follow at once. The monomials of degree one are the T_a themselves, part of a basis, so they are linearly independent in $U(\mathfrak{g})$ and the map $\mathfrak{g} \to U(\mathfrak{g})$ is injective. And if some element of the free associative algebra on the symbols T_a were annihilated in $U(\mathfrak{g})$ beyond what Equation (14.12) forces, its reduction to ordered form — which uses those relations only — would be a nontrivial vanishing combination of basis elements. □

A.30.5 Consequences: the symbol and the symmetrization map

The next two results are what Appendix A.32 needs. From here on $\mathbb{Q} \subseteq \mathbb{K}$, so that $n!$ may be inverted.

Definition A.325 (Symmetric algebra and symbol)**.** Let $S(\mathfrak{g}) = \mathbb{K}[T_1, \ldots, T_M]$ be the polynomial algebra on the same generators taken as *commuting* symbols, graded by degree, $S(\mathfrak{g}) = \bigoplus_{n \geq 0} S^n(\mathfrak{g})$; invariantly, $S^n(\mathfrak{g})$ is the n-th symmetric power of the vector space $\mathfrak{g}$. By Theorem A.319 the ordered monomials of degree exactly n are a basis of a complement of U_{n-1} in U_n, so

$$\sigma_n : U_n / U_{n-1} \longrightarrow S^n(\mathfrak{g})\,, \qquad T_1^{n_1} \cdots T_M^{n_M} + U_{n-1} \longmapsto T_1^{n_1} \cdots T_M^{n_M}\,, \tag{A.587}$$

$n_1 + \cdots + n_M = n$, is a linear isomorphism. The element $\sigma_n(u + U_{n-1})$ is the *symbol* of $u \in U_n$, and $\sigma = \bigoplus_n \sigma_n$ is an isomorphism of graded algebras from $\mathrm{gr}\, U(\mathfrak{g}) = \bigoplus_n U_n/U_{n-1}$ onto $S(\mathfrak{g})$: it is multiplicative because two ordered monomials multiply, modulo U_{n+m-1}, by concatenation and reordering, and Equation (A.570) shows the reordering costs only terms of lower degree.

Rests on Theorem A.319 and Definition A.320.

Definition A.326 (Symmetrization)**.** The *symmetrization map* is the linear map $\omega : S(\mathfrak{g}) \to U(\mathfrak{g})$ determined by

$$\boxed{\omega\left(X_1 X_2 \cdots X_n\right) = \frac{1}{n!} \sum_{\pi} X_{\pi(1)} X_{\pi(2)} \cdots X_{\pi(n)}}\,, \qquad X_i \in \mathfrak{g}\,, \tag{A.588}$$

the sum running over the $n!$ permutations π of $\{1, \ldots, n\}$, with $\omega(1) = 1$. The right-hand side is symmetric in $X_1, \ldots, X_n$ and multilinear, so it descends from $\mathfrak{g}^n$ to $S^n(\mathfrak{g})$ and ω is well defined.

Rests on Definitions 14.5 and A.325.

Proposition A.327 (Symmetrization is an isomorphism of $\mathfrak{g}$-modules)**.** *Let $\mathbb{Q} \subseteq \mathbb{K}$. Then:*

1. *ω maps $S^n(\mathfrak{g})$ into U_n and $\sigma_n(\omega(p) + U_{n-1}) = p$ for $p \in S^n(\mathfrak{g})$; consequently ω is a linear isomorphism of $S(\mathfrak{g})$ onto $U(\mathfrak{g})$.*
2. *Let $\mathfrak{g}$ act on $U(\mathfrak{g})$ by $\mathrm{ad}_X\, u = Xu - uX$ and on $S(\mathfrak{g})$ by the derivation extending ad_X on $\mathfrak{g}$. Then ω intertwines the two actions, $\omega \circ \mathrm{ad}_X = \mathrm{ad}_X \circ \omega$.*

Rests on Definitions A.325 and A.326 and Theorem A.319.

Proof. *Derives Proposition A.327.*

(1) Each word on the right of Equation (A.588) is a product of n elements of $\mathfrak{g}$, hence lies in U_n. Take $p = T_1^{n_1} \cdots T_M^{n_M}$ of degree n. Every one of the $n!$ words in Equation (A.588) is a rearrangement of the same letters, so by Equation (A.570) each equals the ordered monomial $T_1^{n_1} \cdots T_M^{n_M}$ of $U(\mathfrak{g})$ modulo U_{n-1}; averaging, $\omega(p) - T_1^{n_1} \cdots T_M^{n_M} \in U_{n-1}$, which is the assertion about σ_n. Thus ω maps a basis of $S(\mathfrak{g})$ to a family of elements that is "triangular" with respect to the filtration, with the ordered monomials as leading terms. Such a family is a basis: a nontrivial vanishing combination would have a highest degree n occurring in it, and its image under σ_n would be a nontrivial vanishing combination of distinct monomials of $S^n(\mathfrak{g})$. Surjectivity follows because the leading terms exhaust the basis of Theorem A.319.

(2) On $U(\mathfrak{g})$ the map ad_X is a derivation of the associative product, $\mathrm{ad}_X(uv) = (\mathrm{ad}_X\, u)v + u(\mathrm{ad}_X\, v)$, as one checks by expanding $X(uv) - (uv)X$ and inserting $\pm uXv$. Hence

$$\mathrm{ad}_X \left(X_{\pi(1)} \cdots X_{\pi(n)} \right) = \sum_{j=1}^{n} X_{\pi(1)} \cdots \left[X, X_{\pi(j)} \right] \cdots X_{\pi(n)} \,, \tag{A.589}$$

using $\mathrm{ad}_X\, Y = [X, Y]$ for $Y \in \mathfrak{g}$. Average Equation (A.589) over π and group the resulting $n!\, n$ words by the value $i = \pi(j)$ of the slot that was hit: for fixed i, the pairs (π, j) with $\pi(j) = i$ put $[X, X_i]$ in position j and let the remaining letters occupy the remaining positions in every possible order, each exactly once — which is the definition of $\omega\left(X_1 \cdots [X, X_i] \cdots X_n\right)$ up to the common factor $1/n!$. Therefore

$$\mathrm{ad}_X\, \omega\left(X_1 \cdots X_n\right) = \sum_{i=1}^{n} \omega\left(X_1 \cdots [X, X_i] \cdots X_n\right) = \omega\left(\mathrm{ad}_X\left(X_1 \cdots X_n\right)\right) \,, \tag{A.590}$$

the last equality because ad_X acts on $S(\mathfrak{g})$ as the derivation extending ad_X, which is exactly the Leibniz sum on the right. □

Corollary A.328 (The Casimir elements are not accidentally zero)**.** *Let $k^{a_1 \cdots a_r}$ be a nonzero totally symmetric array. Then the element $C_r = k^{a_1 \cdots a_r} T_{a_1} \cdots T_{a_r}$ of Equation (14.19) is nonzero in $U(\mathfrak{g})$, and its symbol is the nonzero polynomial $k^{a_1 \cdots a_r} T_{a_1} \cdots T_{a_r} \in S^r(\mathfrak{g})$. In particular the quadratic Casimir Equation (14.20) of a semisimple algebra is a nonzero element of $U(\mathfrak{g})$.*

Rests on Theorems 14.16 and A.319 and Definition A.325.

Proof. *Derives Corollary A.328.*

Write $p = k^{a_1 \cdots a_r} T_{a_1} \cdots T_{a_r}$ for the corresponding element of $S^r(\mathfrak{g})$, the same expression read in commuting symbols. Then $C_r = \omega(p)$: by linearity $\omega(p) = k^{a_1 \cdots a_r} \frac{1}{r!} \sum_\pi T_{a_{\pi(1)}} \cdots T_{a_{\pi(r)}}$, and renaming the summation indices in each of the $r!$ terms turns it into $k^{a_1 \cdots a_r} T_{a_1} \cdots T_{a_r}$, because $k^{a_1 \cdots a_r}$ is totally symmetric. Now $p \neq 0$: the coefficient of the commuting monomial $T_1^{n_1} \cdots T_M^{n_M}$ in p is the multinomial coefficient $r!/(n_1! \cdots n_M!)$ times the corresponding component of $k^{a_1 \cdots a_r}$, and in characteristic zero that factor cannot vanish. Since ω is injective by Proposition A.327, $C_r = \omega(p) \neq 0$. For Equation (14.20) the array is κ^{ab}, which is invertible, hence nonzero. □

Remark A.329 (What is proved and what is not). Theorem A.319 is a statement about $U(\mathfrak{g})$ as a *vector space*: the ordered monomials are a basis. It is not a statement that $U(\mathfrak{g})$ is commutative, and ω of Proposition A.327 is emphatically *not* an algebra homomorphism — $\omega(XY) - \omega(X)\omega(Y) = -\frac{1}{2}[X, Y]$ for $X, Y \in \mathfrak{g}$, by Equation (A.588) and Equation (14.12). What survives at the level of algebras is the statement of Definition A.325, that the associated graded algebra of $U(\mathfrak{g})$ is the commutative algebra $S(\mathfrak{g})$; noncommutativity is entirely a phenomenon of lower order in the filtration. That is the exact sense in which $U(\mathfrak{g})$ is a deformation of the polynomial algebra on $\mathfrak{g}$, and it is the form in which the theorem is used in Appendix A.32.

Appendix A.30 discharges the proof obligation of Theorem 14.7. Its immediate consequence is the one Section 14.1 needs: the generators satisfy no hidden relations, so a Casimir element built as in Theorem 14.16 is a genuine, nonzero element of $U(\mathfrak{g})$ and not an elaborate way of writing zero (Corollary A.328). The deeper consequence, Proposition A.327, is the first link of the chain that counts those Casimirs in Appendix A.32.

A.31 Cartan's Criterion for Semisimplicity

This appendix proves Theorem 14.13 of Chapter 14: a finite-dimensional Lie algebra over a field of characteristic zero contains no nonzero abelian ideal if and only if its Killing form Equation (14.14) is nondegenerate. The chapter uses the criterion twice — to know when κ_{ab} may be inverted, which is what makes the quadratic Casimir Equation (14.20) available, and in Proposition 14.77 on central extensions, whose proof in Appendix A.37 rests on this one throughout.

Nothing is quoted. The easy half is a two-line index computation (Appendix A.31.3). The converse needs Cartan's criterion for *solvability*, which in turn needs the Jordan decomposition of an endomorphism and Engel's theorem; both are proved here (Appendices A.31.4 and A.31.5), the trace argument that is the heart of the matter is Appendix A.31.6, and the pieces are assembled in Appendix A.31.7. Appendix A.31.8 draws the structural corollaries that Appendix A.37 needs and verifies the criterion on the family this treatise actually uses, the algebras $\mathfrak{so}(p, q)$.

Throughout, $\mathbb{K}$ is a field of characteristic zero with algebraic closure $\overline{\mathbb{K}}$, and $\mathfrak{g}$ is a finite-dimensional Lie algebra over $\mathbb{K}$ with basis $\{T_a\}$ and structure constants $[T_a, T_b] = C^c{}_{ab} T_c$ as in Equation (14.5). For a finite-dimensional vector space $\mathbb{V}$ we write $\mathfrak{gl}(\mathbb{V})$ for the Lie algebra of all linear maps of $\mathbb{V}$ with the commutator bracket. Recall that $\kappa(X, Y) = \operatorname{tr}(\operatorname{ad}_X \operatorname{ad}_Y)$ and that κ is invariant, Equation (14.16).

A.31.1 Statement

Theorem A.330 (Cartan's criterion for semisimplicity)**.** *Let $\mathfrak{g}$ be a finite-dimensional Lie algebra over a field of characteristic zero. Then $\mathfrak{g}$ contains no nonzero abelian ideal if and only if its Killing form is nondegenerate, that is, if and only if*

$$\mathfrak{g}^{\perp} := \{X \in \mathfrak{g} \mid \kappa(X, Y) = 0 \text{ for all } Y \in \mathfrak{g}\} = 0\,. \tag{A.591}$$

Rests on Definition 14.10 and Lemma 14.11.

A.31.2 Solvability, and the language of the proof

Definition A.331 (Derived series, solvable, ideal)**.** For subspaces $\mathfrak{a}, \mathfrak{b} \subseteq \mathfrak{g}$ let $[\mathfrak{a}, \mathfrak{b}]$ denote the span of all $[X, Y]$ with $X \in \mathfrak{a}$, $Y \in \mathfrak{b}$. A subspace $\mathfrak{a}$ is a *subalgebra* if $[\mathfrak{a}, \mathfrak{a}] \subseteq \mathfrak{a}$ and an *ideal* if $[\mathfrak{g}, \mathfrak{a}] \subseteq \mathfrak{a}$. The *derived series* of a subalgebra $\mathfrak{a}$ is

$$\mathfrak{a}^{(0)} = \mathfrak{a}\,, \qquad \mathfrak{a}^{(j+1)} = \left[\mathfrak{a}^{(j)}, \mathfrak{a}^{(j)}\right]\,, \tag{A.592}$$

a decreasing chain of subalgebras, and $\mathfrak{a}$ is *solvable* if $\mathfrak{a}^{(m)} = 0$ for some m. An abelian subalgebra is solvable with $m = 1$.

Lemma A.332 (Elementary properties of solvability)**.** *Let $\mathfrak{a}$ be a subalgebra of $\mathfrak{g}$.*

1. *Subalgebras and homomorphic images of a solvable algebra are solvable.*
2. *If $\mathfrak{b} \subseteq \mathfrak{a}$ is an ideal of $\mathfrak{a}$ with $\mathfrak{b}$ and $\mathfrak{a}/\mathfrak{b}$ solvable, then $\mathfrak{a}$ is solvable.*
3. *If $\mathfrak{a}$ is an ideal of $\mathfrak{g}$, then every $\mathfrak{a}^{(j)}$ is an ideal of $\mathfrak{g}$.*
4. *A nonzero solvable ideal of $\mathfrak{g}$ contains a nonzero* abelian *ideal of $\mathfrak{g}$.*

Rests on Definition A.331.

Proof. *Derives Lemma A.332.*

(1) If $\mathfrak{b} \subseteq \mathfrak{a}$ then $\mathfrak{b}^{(j)} \subseteq \mathfrak{a}^{(j)}$ by induction; if ϕ is a homomorphism then $\phi\left(\mathfrak{a}^{(j)}\right) = \phi(\mathfrak{a})^{(j)}$, again by induction, since ϕ preserves brackets.

(2) If $(\mathfrak{a}/\mathfrak{b})^{(m)} = 0$ then $\mathfrak{a}^{(m)} \subseteq \mathfrak{b}$ by (1) applied to the quotient map, and if $\mathfrak{b}^{(l)} = 0$ then $\mathfrak{a}^{(m+l)} \subseteq \mathfrak{b}^{(l)} = 0$.

(3) By induction on j; the case $j = 0$ is the hypothesis. Let $\mathfrak{b} = \mathfrak{a}^{(j)}$ be an ideal. For $Z \in \mathfrak{g}$ and $X, Y \in \mathfrak{b}$ the Jacobi identity gives

$$[Z,[X,Y]] = [[Z,X],Y] + [X,[Z,Y]]\,, \tag{A.593}$$

and both terms on the right lie in $[\mathfrak{b},\mathfrak{b}] = \mathfrak{a}^{(j+1)}$ because $[Z,X]$ and $[Z,Y]$ lie in $\mathfrak{b}$.

(4) Let $\mathfrak{s} \neq 0$ be a solvable ideal and let m be largest with $\mathfrak{s}^{(m)} \neq 0$. Then $\left[\mathfrak{s}^{(m)},\mathfrak{s}^{(m)}\right] = \mathfrak{s}^{(m+1)} = 0$, so $\mathfrak{s}^{(m)}$ is abelian, and it is an ideal of $\mathfrak{g}$ by (3). □

Remark A.333 (Two definitions of semisimple, and why they agree here). Theorem 14.13 defines *semisimple* as "contains no nonzero abelian ideal". The literature usually says "contains no nonzero solvable ideal", equivalently "the radical — the largest solvable ideal — is zero". Part (4) of Lemma A.332 is exactly the statement that the two conditions coincide: an abelian ideal is solvable, and a nonzero solvable ideal produces a nonzero abelian one. We use whichever is convenient and say "semisimple" for both.

A.31.3 The easy half: an abelian ideal lies in the radical of the Killing form

Proposition A.334 (Nondegeneracy implies no abelian ideal)**.** *Let $\mathfrak{a} \subseteq \mathfrak{g}$ be an abelian ideal. Then $\mathfrak{a} \subseteq \mathfrak{g}^{\perp}$. Consequently, if κ is nondegenerate then $\mathfrak{g}$ has no nonzero abelian ideal.*

Rests on Definitions 14.10 and A.331.

Proof. *Derives Proposition A.334.*

Let $X \in \mathfrak{a}$ and $Y \in \mathfrak{g}$, and put $N = \mathrm{ad}_X \mathrm{ad}_Y$. Since $\mathfrak{a}$ is an ideal, $\mathrm{ad}_X \mathfrak{g} = [X,\mathfrak{g}] \subseteq \mathfrak{a}$, so $N\mathfrak{g} \subseteq \mathfrak{a}$; and since $\mathfrak{a}$ is an ideal, $\mathrm{ad}_Y \mathfrak{a} \subseteq \mathfrak{a}$, while $\mathrm{ad}_X \mathfrak{a} = [\mathfrak{a},\mathfrak{a}] = 0$ because $\mathfrak{a}$ is abelian. Hence

$$N^2\mathfrak{g} \subseteq N\mathfrak{a} = \mathrm{ad}_X \mathrm{ad}_Y \mathfrak{a} \subseteq \mathrm{ad}_X \mathfrak{a} = 0\,, \tag{A.594}$$

so $N^2 = 0$. A nilpotent endomorphism has all eigenvalues zero and therefore vanishing trace, so $\kappa(X,Y) = \operatorname{tr} N = 0$ for every Y, that is, $X \in \mathfrak{g}^{\perp}$.

The same computation in components, with the index conventions of Equation (14.15), is worth displaying because it is the form in which the statement is usually met. Choose a basis adapted to $\mathfrak{a}$: indices i,j,k label a basis of $\mathfrak{a}$ and indices μ,ν complete it. That $\mathfrak{a}$ is an ideal says $C^{\mu}{}_{ai} = 0$ for every a; that it is abelian says $C^{c}{}_{ij} = 0$. Then, from Equation (14.15),

$$\kappa_{ib} = C^{c}{}_{id}\, C^{d}{}_{bc} = C^{j}{}_{id}\, C^{d}{}_{bj} = C^{j}{}_{ik}\, C^{k}{}_{bj} = 0\,, \tag{A.595}$$

the second equality because $C^{c}{}_{id}$ vanishes unless the upper index lies in $\mathfrak{a}$, the third because $C^{d}{}_{bj}$ vanishes unless d does, and the last because $C^{j}{}_{ik} = 0$. □

A.31.4 The Jordan decomposition

Theorem A.335 (Jordan decomposition)**.** *Let $\mathbb{V}$ be a finite-dimensional vector space over an algebraically closed field $\mathbb{F}$ and let $x \in \mathrm{End}(\mathbb{V})$. There are unique $x_s, x_n \in \mathrm{End}(\mathbb{V})$ with*

$$x = x_s + x_n\,, \qquad x_s \text{ diagonalizable}\,, \quad x_n \text{ nilpotent}\,, \quad [x_s, x_n] = 0\,. \tag{A.596}$$

Moreover x_s and x_n are polynomials in x with zero constant term.

Rests on Definition 5.35.

Proof. *Derives Theorem A.335.*

Let $a_1, \dots, a_r$ be the distinct eigenvalues of x and $\prod_i (t - a_i)^{m_i}$ its characteristic polynomial, which splits because the field is algebraically closed. Put $\mathbb{V}_i = \ker (x - a_i)^{m_i}$. The polynomials $(t - a_i)^{m_i}$ are pairwise coprime, so by the Chinese remainder theorem in $\mathbb{F}[t]$ — applied modulo the characteristic polynomial, which annihilates x by Cayley–Hamilton — the $\mathbb{V}_i$ are the images of projections that are polynomials in x, and $\mathbb{V} = \bigoplus_i \mathbb{V}_i$.

By the same theorem choose $p \in \mathbb{F}[t]$ with

$$p \equiv a_i \pmod{(t - a_i)^{m_i}} \quad (i = 1, \dots, r)\,, \qquad p \equiv 0 \pmod{t}\,, \tag{A.597}$$

the last congruence being imposed only when 0 is not among the a_i — when it is, say $a_1 = 0$, the congruence $p \equiv 0 \pmod{t^{m_1}}$ already forces $p(0) = 0$, and the moduli stay pairwise coprime in either case. Set $x_s = p(x)$ and $x_n = x - x_s$. On $\mathbb{V}_i$ the first congruence gives $x_s = a_i \mathbb{1}$, so x_s is diagonalizable, and $x_n = x - a_i \mathbb{1}$ there, which is nilpotent on $\mathbb{V}_i$ by definition of $\mathbb{V}_i$; being nilpotent on each summand of a finite direct sum it is nilpotent. Both are polynomials in x with $p(0) = 0$, hence commute with x and with each other.

Uniqueness. Let $x = s' + n'$ be another such decomposition. Since s' and n' commute with each other they commute with x, hence with every polynomial in x, hence with x_s and x_n. Then $x_s - s' = n' - x_n$ is at once diagonalizable — a difference of two commuting diagonalizable maps, which are simultaneously diagonalizable — and nilpotent, being a difference of two commuting nilpotent maps. A map that is both is zero. □

Lemma A.336 (The adjoint of a Jordan decomposition)**.** *With the notation of Theorem A.335, viewing $\mathrm{ad}_x(y) = [x, y]$ as an endomorphism of $\mathrm{End}(\mathbb{V})$,*

$$(\mathrm{ad}_x)_s = \mathrm{ad}_{x_s}\,, \qquad (\mathrm{ad}_x)_n = \mathrm{ad}_{x_n}\,. \tag{A.598}$$

In particular ad_{x_s} is a polynomial in ad_x with zero constant term.

Rests on Theorem A.335.

Proof. *Derives Lemma A.336.*

Let $e_1, \dots, e_m$ be a basis of eigenvectors of x_s, $x_s e_i = a_i e_i$, and let E_{ij} be the corresponding matrix units. Then $\mathrm{ad}_{x_s} E_{ij} = (a_i - a_j) E_{ij}$, so ad_{x_s} is diagonalizable. Writing $L(y) = x_n y$ and $R(y) = y x_n$, the maps L and R commute and are nilpotent, so $\mathrm{ad}_{x_n} = L - R$ is nilpotent. And $[\mathrm{ad}_{x_s}, \mathrm{ad}_{x_n}] = \mathrm{ad}_{[x_s, x_n]} = 0$. Hence $\mathrm{ad}_x = \mathrm{ad}_{x_s} + \mathrm{ad}_{x_n}$ satisfies the three conditions of Equation (A.596), and uniqueness identifies the two summands. The last sentence is the final statement of Theorem A.335 applied to the endomorphism ad_x of $\mathrm{End}(\mathbb{V})$. □

A.31.5 Engel's theorem

Lemma A.337 (A nilpotent map has nilpotent adjoint)**.** *If $x \in \mathrm{End}(\mathbb{V})$ is nilpotent then ad_x is nilpotent.*

Rests on Definition 5.35.

Proof. *Derives Lemma A.337.*

With $L(y) = xy$ and $R(y) = yx$ as above, L and R commute and each is nilpotent, since $x^N = 0$ gives $L^N = R^N = 0$. The binomial theorem for commuting maps gives $(L - R)^{2N} = 0$. □

Theorem A.338 (Engel)**.** *Let $\mathbb{V} \neq 0$ be finite-dimensional and let $\mathfrak{h} \subseteq \mathfrak{gl}(\mathbb{V})$ be a Lie subalgebra all of whose elements are nilpotent endomorphisms. Then there is a nonzero $v \in \mathbb{V}$ with $\mathfrak{h}v = 0$; consequently there is a basis of $\mathbb{V}$ in which every element of $\mathfrak{h}$ is strictly upper triangular, and $\mathfrak{h}$ is solvable.*

Rests on Lemma A.337 and Definition A.331.

Proof. *Derives Theorem A.338.*

Induct on $\dim \mathfrak{h}$. If $\mathfrak{h} = 0$ any nonzero v will do. Let $\dim \mathfrak{h} \geq 1$ and choose a maximal proper subalgebra $\mathfrak{m} \subsetneq \mathfrak{h}$, which exists because 0 is a proper subalgebra and dimensions are finite.

$\mathfrak{m}$ is an ideal of codimension one. The adjoint action of $\mathfrak{m}$ on $\mathfrak{h}$ preserves $\mathfrak{m}$, so it descends to an action on $\mathfrak{h}/\mathfrak{m} \neq 0$; each ad_X, $X \in \mathfrak{m}$, is nilpotent by Lemma A.337, hence so is the induced map. The image of $\mathfrak{m}$ in $\mathfrak{gl}(\mathfrak{h}/\mathfrak{m})$ is a subalgebra of dimension at most $\dim \mathfrak{m} < \dim \mathfrak{h}$ consisting of nilpotent maps, so the induction hypothesis supplies a nonzero $z + \mathfrak{m}$ annihilated by it: there is $z \in \mathfrak{h} \setminus \mathfrak{m}$ with $[\mathfrak{m}, z] \subseteq \mathfrak{m}$. Then $\mathfrak{m} + \mathbb{K}z$ is a subalgebra strictly containing $\mathfrak{m}$, so it is $\mathfrak{h}$ by maximality, and $\mathfrak{m}$ is an ideal of codimension one.

The common kernel. Let $\mathbb{W} = \{v \in \mathbb{V} \mid \mathfrak{m}v = 0\}$, which is nonzero by the induction hypothesis applied to $\mathfrak{m}$. It is stable under $\mathfrak{h}$: for $X \in \mathfrak{m}$, $Y \in \mathfrak{h}$ and $v \in \mathbb{W}$,

$$X(Yv) = Y(Xv) + [X, Y]\, v = 0 + 0 = 0\,, \tag{A.599}$$

since $[X, Y] \in \mathfrak{m}$. In particular z maps $\mathbb{W}$ to itself, and z is nilpotent, so it has a nonzero kernel vector $v \in \mathbb{W}$. That v is annihilated by $\mathfrak{m}$ and by z, hence by $\mathfrak{h} = \mathfrak{m} + \mathbb{K}z$.

The flag. Induct on $\dim \mathbb{V}$: take v as above as the first basis vector and apply the statement to the induced action on $\mathbb{V}/\mathbb{K}v$, whose elements are again nilpotent. In the resulting basis every $X \in \mathfrak{h}$ is strictly upper triangular.

Solvability. Let $\mathfrak{n}_r$ be the space of matrices whose entries vanish except at least r places above the diagonal, so that $\mathfrak{h} \subseteq \mathfrak{n}_1$ and $[\mathfrak{n}_r, \mathfrak{n}_s] \subseteq \mathfrak{n}_{r+s}$ by matrix multiplication. Then $\mathfrak{h}^{(j)} \subseteq \mathfrak{n}_{2^j}$ by induction, and $\mathfrak{n}_r = 0$ once $r \geq \dim \mathbb{V}$. □

A.31.6 The trace lemma

Everything so far has been preparation. The following lemma is the mechanism of Cartan's criterion, and the only place where characteristic zero is used in an essential way — through the rational numbers, of all things, inside a field that may be as large as $\mathbb{C}$.

Lemma A.339 (Trace lemma)**.** *Let $\mathbb{V}$ be finite-dimensional over an algebraically closed field $\mathbb{F}$ of characteristic zero, let $A \subseteq B$ be subspaces of $\mathfrak{gl}(\mathbb{V})$, and put*

$$M = \{x \in \mathfrak{gl}(\mathbb{V}) \mid [x, B] \subseteq A\}\ . \tag{A.600}$$

If $x \in M$ satisfies $\mathrm{tr}(xy) = 0$ for every $y \in M$, then x is nilpotent.

Rests on Theorem A.335 and Lemma A.336.

Proof. *Derives Lemma A.339.*

Let $x = x_s + x_n$ be the Jordan decomposition (Theorem A.335), let $e_1, \dots, e_m$ be a basis of eigenvectors of x_s with $x_s e_i = a_i e_i$, and let $E \subseteq \mathbb{F}$ be the $\mathbb{Q}$-linear span of $a_1, \dots, a_m$, a finite-dimensional vector space over $\mathbb{Q}$. The claim is that $E = 0$; then every $a_i = 0$, so $x_s = 0$ and $x = x_n$ is nilpotent.

Since E is finite-dimensional over $\mathbb{Q}$, it is enough to show that every $\mathbb{Q}$-linear functional $f : E \to \mathbb{Q}$ vanishes. Fix such an f and let $y \in \mathrm{End}(\mathbb{V})$ be defined by $y e_i = f(a_i) e_i$.

Step 1: ad_y *is a polynomial in* ad_x *with zero constant term.* In the basis of matrix units E_{ij} built from the e_i,

$$\mathrm{ad}_{x_s} E_{ij} = (a_i - a_j)\, E_{ij}\,, \qquad \mathrm{ad}_y E_{ij} = (f(a_i) - f(a_j))\, E_{ij}\,. \tag{A.601}$$

Choose $r \in \mathbb{F}[t]$ with $r(0) = 0$ and $r(a_i - a_j) = f(a_i) - f(a_j)$ for all i, j. This is a consistent finite set of interpolation conditions: if $a_i - a_j = a_p - a_q$ then, applying the $\mathbb{Q}$-linear f, $f(a_i) - f(a_j) = f(a_p) - f(a_q)$; and if $a_i - a_j = 0$ the prescribed value is 0, which is what $r(0) = 0$ demands. Lagrange interpolation over the finitely many distinct values then produces r. By Equation (A.601), $\mathrm{ad}_y = r(\mathrm{ad}_{x_s})$, and by Lemma A.336 $\mathrm{ad}_{x_s} = q(\mathrm{ad}_x)$ for some q with $q(0) = 0$. Hence $\mathrm{ad}_y = r(q(\mathrm{ad}_x))$, a polynomial in ad_x without constant term.

Step 2: $y \in M$. Since $x \in M$ we have $\mathrm{ad}_x(B) \subseteq A \subseteq B$, so $(\mathrm{ad}_x)^j(B) \subseteq A$ for every $j \geq 1$. A polynomial in ad_x with zero constant term is a combination of such powers, so $\mathrm{ad}_y(B) \subseteq A$, that is, $y \in M$.

Step 3: the trace. By hypothesis $\mathrm{tr}(xy) = 0$. Compute it. Since x_n commutes with x_s it preserves each eigenspace of x_s, and inside each eigenspace a basis may be chosen making the nilpotent x_n strictly upper triangular; taking the e_i so adapted, x is upper triangular with diagonal entries $a_1, \ldots, a_m$ while y is diagonal with entries $f(a_1), \ldots, f(a_m)$. Then xy is upper triangular with diagonal entries $a_i f(a_i)$, so

$$\sum_{i=1}^{m} a_i\, f(a_i) = \mathrm{tr}(xy) = 0\,. \tag{A.602}$$

The left-hand side lies in E, since each $f(a_i) \in \mathbb{Q}$ and E is a $\mathbb{Q}$-subspace containing the a_i. Apply f to Equation (A.602) and use $\mathbb{Q}$-linearity:

$$\sum_{i=1}^{m} f(a_i)^2 = 0\,, \qquad f(a_i) \in \mathbb{Q}\,. \tag{A.603}$$

A sum of squares of rational numbers vanishes only if every term does, so $f(a_i) = 0$ for all i; as the a_i span E over $\mathbb{Q}$, $f = 0$. Every $\mathbb{Q}$-linear functional on E vanishes, so $E = 0$. □

Theorem A.340 (Cartan's criterion for solvability)**.** *Let* $\mathbb{V}$ *be finite-dimensional over a field* $\mathbb{K}$ *of characteristic zero and let* $\mathfrak{h} \subseteq \mathfrak{gl}(\mathbb{V})$ *be a Lie subalgebra with*

$$\mathrm{tr}(xy) = 0 \qquad \textit{for all } x \in [\mathfrak{h}, \mathfrak{h}]\ ,\ y \in \mathfrak{h}\,. \tag{A.604}$$

Then $\mathfrak{h}$ *is solvable.*

Rests on Lemma A.339 and Theorem A.338.

Proof. *Derives Theorem A.340.*

Assume first that $\mathbb{K}$ is algebraically closed. Apply Lemma A.339 with $A = [\mathfrak{h}, \mathfrak{h}]$, $B = \mathfrak{h}$ and M as in Equation (A.600); note $\mathfrak{h} \subseteq M$. Let $x \in [\mathfrak{h}, \mathfrak{h}]$ and $y \in M$. By bilinearity it suffices to treat $x = [u, v]$ with $u, v \in \mathfrak{h}$, and then

$$\mathrm{tr}\left([u, v]\, y\right) = \mathrm{tr}(uvy) - \mathrm{tr}(vuy) = \mathrm{tr}(uvy) - \mathrm{tr}(uyv) = \mathrm{tr}\left(u\,[v, y]\right) = \mathrm{tr}\left([v, y]\, u\right)\ , \tag{A.605}$$

using cyclicity of the trace twice. Now $y \in M$ gives $[v, y] \in [\mathfrak{h}, \mathfrak{h}]$, and $u \in \mathfrak{h}$, so the right-hand side vanishes by Equation (A.604). Lemma A.339 therefore makes every element of $[\mathfrak{h}, \mathfrak{h}]$ a nilpotent endomorphism, and Theorem A.338 makes $[\mathfrak{h}, \mathfrak{h}]$ solvable. Since $\mathfrak{h}/[\mathfrak{h}, \mathfrak{h}]$ is abelian, Lemma A.332(2) makes $\mathfrak{h}$ solvable.

For general $\mathbb{K}$, extend scalars to $\overline{\mathbb{K}}$. The hypothesis Equation (A.604) is bilinear and $[\mathfrak{h}, \mathfrak{h}]$ spans $[\mathfrak{h}_{\overline{\mathbb{K}}}, \mathfrak{h}_{\overline{\mathbb{K}}}]$, so the hypothesis holds for $\mathfrak{h}_{\overline{\mathbb{K}}} \subseteq \mathfrak{gl}\left(\mathbb{V}_{\overline{\mathbb{K}}}\right)$, which is therefore solvable. Since $\left(\mathfrak{h}^{(j)}\right)_{\overline{\mathbb{K}}} = \left(\mathfrak{h}_{\overline{\mathbb{K}}}\right)^{(j)}$ — brackets of spanning sets span — and a subspace vanishes if and only if its extension does, $\mathfrak{h}$ is solvable. □

A.31.7 Proof of the criterion

Proof of Theorem A.330. *Derives Theorem A.330.*

One direction is Proposition A.334. For the other, assume $\mathfrak{g}$ has no nonzero abelian ideal and let $\mathfrak{s} = \mathfrak{g}^{\perp}$ be the radical of the Killing form, Equation (A.591).

$\mathfrak{s}$ *is an ideal.* Let $X \in \mathfrak{s}$ and $Z, Y \in \mathfrak{g}$. By the invariance of the Killing form, Equation (14.16),

$$\kappa\left([Z,X],Y\right) = -\kappa\left(X,[Z,Y]\right) = 0\,, \tag{A.606}$$

so $[Z,X] \in \mathfrak{s}$.

$\mathrm{ad}(\mathfrak{s})$ *is solvable.* Consider $\mathrm{ad}(\mathfrak{s}) \subseteq \mathfrak{gl}(\mathfrak{g})$, a Lie subalgebra because ad is a homomorphism. Its derived algebra is $\mathrm{ad}\left([\mathfrak{s},\mathfrak{s}]\right)$. Take $x = \mathrm{ad}_W$ with $W \in [\mathfrak{s},\mathfrak{s}] \subseteq \mathfrak{s}$ and $y = \mathrm{ad}_Y$ with $Y \in \mathfrak{s}$; then

$$\mathrm{tr}(xy) = \mathrm{tr}\left(\mathrm{ad}_W \mathrm{ad}_Y\right) = \kappa(W,Y) = 0 \tag{A.607}$$

because $W \in \mathfrak{s} = \mathfrak{g}^{\perp}$. So Theorem A.340 applies and $\mathrm{ad}(\mathfrak{s})$ is solvable.

$\mathfrak{s}$ *is solvable.* The kernel of ad restricted to $\mathfrak{s}$ is $\mathfrak{s} \cap Z(\mathfrak{g})$, where $Z(\mathfrak{g}) = \{X \in \mathfrak{g} \mid [X,\mathfrak{g}] = 0\}$ is the centre; it is abelian, hence solvable; the image $\mathrm{ad}(\mathfrak{s})$ is solvable; so Lemma A.332(2) applies to the ideal $\ker\left(\mathrm{ad}\,|_{\mathfrak{s}}\right)$ of $\mathfrak{s}$.

Conclusion. If $\mathfrak{s} \neq 0$ then, being a nonzero solvable ideal of $\mathfrak{g}$, it would contain a nonzero abelian ideal of $\mathfrak{g}$ by Lemma A.332(4), contrary to hypothesis. Hence $\mathfrak{s} = 0$ and κ is nondegenerate. □

Remark A.341 (Scalar extension). Semisimplicity is unchanged by enlarging the field, and the criterion is what makes this obvious. In a basis of $\mathfrak{g}$ over $\mathbb{K}$ the Killing form has a Gram matrix κ_{ab} of scalars, and the same basis is a basis of $\mathfrak{g}_K = \mathfrak{g} \otimes_{\mathbb{K}} K$ over any extension K with the same structure constants, hence the same κ_{ab} by Equation (14.15). Nondegeneracy is the nonvanishing of $\det\left(\kappa_{ab}\right)$, a condition on that matrix alone. So $\mathfrak{g}$ is semisimple if and only if $\mathfrak{g}_K$ is — a statement that is *not* obvious from the definition by ideals, since $\mathfrak{g}_K$ has many more subspaces than $\mathfrak{g}$.

A.31.8 Consequences, and the case of $\mathfrak{so}(p,q)$

Corollary A.342 (Orthogonal splitting of ideals)**.** *Let $\mathfrak{g}$ be semisimple and $\mathfrak{a} \subseteq \mathfrak{g}$ an ideal, and let $\mathfrak{a}^{\perp} = \{X \mid \kappa(X,\mathfrak{a}) = 0\}$. Then*

$$\mathfrak{g} = \mathfrak{a} \oplus \mathfrak{a}^{\perp}\,, \qquad \left[\mathfrak{a},\mathfrak{a}^{\perp}\right] = 0\,, \tag{A.608}$$

both summands are ideals, both are semisimple, and $\mathfrak{g}/\mathfrak{a} \cong \mathfrak{a}^{\perp}$. Consequently $\mathfrak{g}$ is a direct sum of simple ideals, $Z(\mathfrak{g}) = 0$, and $[\mathfrak{g},\mathfrak{g}] = \mathfrak{g}$.

Rests on Theorems A.330 and A.340.

Proof. *Derives Corollary A.342.*

$\mathfrak{a}^{\perp}$ is an ideal by the computation Equation (A.606). Put $\mathfrak{b} = \mathfrak{a} \cap \mathfrak{a}^{\perp}$, an ideal. For $W \in [\mathfrak{b},\mathfrak{b}] \subseteq \mathfrak{a}$ and $Y \in \mathfrak{b} \subseteq \mathfrak{a}^{\perp}$ we have $\mathrm{tr}\left(\mathrm{ad}_W \mathrm{ad}_Y\right) = \kappa(W,Y) = 0$, so Theorem A.340 makes $\mathrm{ad}(\mathfrak{b})$ solvable and, exactly as in the proof of Theorem A.330, $\mathfrak{b}$ solvable. A semisimple algebra has no nonzero solvable ideal (Remark A.333), so $\mathfrak{b} = 0$. Nondegeneracy of κ gives $\dim \mathfrak{a}^{\perp} = \dim \mathfrak{g} - \dim \mathfrak{a}$, whence the direct sum. Then $\left[\mathfrak{a},\mathfrak{a}^{\perp}\right] \subseteq \mathfrak{a} \cap \mathfrak{a}^{\perp} = 0$, both being ideals.

An abelian ideal of $\mathfrak{a}$ is an ideal of $\mathfrak{g}$, since $\mathfrak{a}^{\perp}$ brackets to zero with it; so it vanishes, and $\mathfrak{a}$ is semisimple, as is $\mathfrak{a}^{\perp}$. The projection $\mathfrak{g} \to \mathfrak{a}^{\perp}$ along $\mathfrak{a}$ is a homomorphism with kernel $\mathfrak{a}$.

Iterating the splitting on a proper nonzero ideal, and stopping when none exists, writes $\mathfrak{g} = \mathfrak{g}_1 \oplus \cdots \oplus \mathfrak{g}_s$ with each $\mathfrak{g}_j$ having no proper nonzero ideal; each is nonabelian, since an abelian $\mathfrak{g}_j$ would be an abelian ideal of $\mathfrak{g}$, so each is simple. The centre of $\mathfrak{g}$ is an abelian ideal, hence zero. Finally $[\mathfrak{g}_j,\mathfrak{g}_j]$ is a nonzero ideal of the simple $\mathfrak{g}_j$ — nonzero because $\mathfrak{g}_j$ is not abelian — so it is all of $\mathfrak{g}_j$, and summing over j gives $[\mathfrak{g},\mathfrak{g}] = \mathfrak{g}$. □

Corollary A.343 (The Killing form of a simple ideal)**.** *In the decomposition* $\mathfrak{g} = \bigoplus_j \mathfrak{g}_j$ *of Corollary A.342,* $\kappa(\mathfrak{g}_i, \mathfrak{g}_j) = 0$ *for* $i \neq j$*, and the restriction of* κ *to* $\mathfrak{g}_j$ *is the Killing form of* $\mathfrak{g}_j$ *computed in* $\mathfrak{g}_j$ *itself.*

Rests on Corollary A.342 and Definition 14.10.

Proof. *Derives Corollary A.343.*

Let $X \in \mathfrak{g}_i$, $Y \in \mathfrak{g}_j$, $i \neq j$. Then ad_Y annihilates every summand but $\mathfrak{g}_j$ and maps $\mathfrak{g}_j$ into itself, while ad_X annihilates $\mathfrak{g}_j$; so $\mathrm{ad}_X \mathrm{ad}_Y = 0$ and $\kappa(X, Y) = 0$. For $X, Y \in \mathfrak{g}_j$ the map $\mathrm{ad}_X \mathrm{ad}_Y$ annihilates the other summands and preserves $\mathfrak{g}_j$, so its trace over $\mathfrak{g}$ equals its trace over $\mathfrak{g}_j$. □

Corollary A.344 (The algebras $\mathfrak{so}(p, q)$ are semisimple)**.** *Let* $D = p + q \geq 3$ *and let* $\mathfrak{so}(p, q)$ *have the generators* J_{AB} *and brackets Equation (14.90). In the basis* $\{J_{AB}\}_{A<B}$ *the Killing form is diagonal,*

$$\boxed{\kappa(J_{AB}, J_{CD}) = 2(D-2)\left(\eta_{AD}\eta_{BC} - \eta_{AC}\eta_{BD}\right)}, \tag{A.609}$$

with $\kappa(J_{AB}, J_{AB}) = -2(D-2)\,\eta_{AA}\eta_{BB} \neq 0$ *for* $A \neq B$ *and no summation. Hence* κ *is nondegenerate and* $\mathfrak{so}(p, q)$ *is semisimple for every signature with* $D \geq 3$*; in particular the Lorentz algebra* $\mathfrak{so}(D-1, 1)$ *is, and* $\mathfrak{so}(2)$ *— where Equation (A.609) vanishes identically — is not, being abelian.*

Rests on Theorem A.330, Equation (14.90), and Proposition 14.59.

Proof. *Derives Corollary A.344.*

That the J_{AB} with $A < B$ are a basis is part (ii) of Proposition 14.59, and η is diagonal by Notation 14.1.

Off-diagonal entries vanish. Fix an index A_0 and let ς be the reflection $x^{A_0} \mapsto -x^{A_0}$ of $\mathbb{R}^D$, the other coordinates unchanged. It preserves η, so its pushforward on vector fields maps Killing fields to Killing fields and, being induced by a diffeomorphism, preserves Lie brackets: it is an automorphism of $\mathfrak{so}(p, q)$. On the generators Equation (14.86) it acts by $\varsigma_* J_{AB} = \epsilon_A \epsilon_B J_{AB}$, where $\epsilon_A = -1$ if $A = A_0$ and $+1$ otherwise, because both x_A and ∂_B pick up their own sign. The Killing form is invariant under any automorphism ϕ, since $\mathrm{ad}_{\phi X} = \phi\, \mathrm{ad}_X\, \phi^{-1}$ and the trace is conjugation invariant; hence

$$\kappa(J_{AB}, J_{CD}) = \epsilon_A \epsilon_B \epsilon_C \epsilon_D\, \kappa(J_{AB}, J_{CD})\ . \tag{A.610}$$

If $\{A, B\} \neq \{C, D\}$, choose A_0 in the symmetric difference of the two pairs; then exactly one of the four factors is -1 — the pairs have distinct entries — so the sign is -1 and $\kappa(J_{AB}, J_{CD}) = 0$.

Diagonal entries. Take $H = J_{12}$; the general case is the same computation with the indices renamed. From Equation (14.90), $[J_{12}, J_{CD}] = \eta_{1C} J_{D2} + \eta_{2D} J_{C1} + \eta_{1D} J_{2C} + \eta_{2C} J_{1D}$, and since η is diagonal only the terms whose two indices coincide survive. For $k \geq 3$,

$$\mathrm{ad}_H\, J_{1k} = \eta_{11} J_{k2} = -\eta_{11} J_{2k}\,, \qquad \mathrm{ad}_H\, J_{2k} = \eta_{22} J_{1k}\,, \tag{A.611}$$

while $\mathrm{ad}_H\, J_{12} = 0$ and $\mathrm{ad}_H\, J_{kl} = 0$ for $k, l \geq 3$. Hence ad_H^2 is diagonal in the basis, acting as $-\eta_{11}\eta_{22}$ on each of the $2(D-2)$ generators J_{1k}, J_{2k} with $k \geq 3$, and as zero on the rest, so

$$\kappa(J_{12}, J_{12}) = \mathrm{tr}\left(\mathrm{ad}_H^2\right) = -2(D-2)\,\eta_{11}\eta_{22}\,, \tag{A.612}$$

which is Equation (A.609) for $A = C = 1$, $B = D = 2$; the general case of Equation (A.609) follows because both sides vanish off the diagonal and both are antisymmetric in AB and in CD. Since $\eta_{AA}\eta_{BB} = \pm 1$, the diagonal entries are nonzero exactly when $D \neq 2$, and Theorem A.330 converts nondegeneracy into semisimplicity. □

Remark A.345 (What the criterion buys, in the chapter's own terms). Three things follow immediately and are used without further comment. First, κ_{ab} may be inverted for a semisimple algebra, which is what Corollary 14.17 needs to write $C_2 = \kappa^{ab}T_aT_b$ and what Remark 14.15 needs to move indices between the two forms of invariance. Second, by Corollary A.342 a semisimple algebra equals its own derived algebra; this is why a semisimple algebra has no nonzero homomorphism to an abelian one, a fact used repeatedly in Appendix A.37. Third, Remark 14.23 is now precise rather than plausible: the Poincaré algebra Equation (14.99) contains the translations as an abelian ideal, so Proposition A.334 puts every translation generator in the radical of its Killing form, the form is degenerate, and no statement proved for semisimple algebras may be applied to it.

Appendix A.31 discharges the proof obligation of Theorem 14.13. It is used at Corollary 14.17 to invert the Killing form, at Proposition 14.22 and Theorem 14.21 wherever semisimplicity is assumed, and — through Corollaries A.342 and A.344 — as the foundation of Appendix A.37, which proves Proposition 14.77 on central extensions.

A.32 Racah's Theorem on the Number of Casimir Operators

This appendix proves Theorem 14.21 of Chapter 14: for a finite-dimensional semisimple Lie algebra of rank ℓ (Definition 14.20), the centre $Z(U(\mathfrak{g}))$ of the universal enveloping algebra is a polynomial algebra on ℓ algebraically independent Casimir elements, so that an irreducible representation carries exactly ℓ invariant labels. The chapter needs the count in two places: to know that $\mathfrak{su}(2)$ has one Casimir and $\mathfrak{su}(3)$ two (Remark 14.19), and to convert the rank of $\mathfrak{so}(p,q)$ computed in Proposition 14.22 into the number of Lorentz invariants.

The proof is a chain of three identifications:

$$\begin{aligned} Z(U(\mathfrak{g})) &\cong \text{the invariant polynomials on } \mathfrak{g} \\ &\cong \text{the } W\text{-invariant polynomials on } \mathfrak{h} \\ &\cong \text{a free polynomial algebra on } \ell \text{ generators}, \end{aligned} \tag{A.613}$$

where $\mathfrak{h}$ is a Cartan subalgebra, W its Weyl group, and "invariant" means annihilated by every ad_X. The first link is proved here in full (Appendix A.32.2), and rests on the Poincaré–Birkhoff–Witt theorem of Appendix A.30. The second is the Chevalley restriction theorem and the third is Chevalley's theorem on finite reflection groups; both belong to the structure theory of semisimple Lie algebras, which this treatise does not build, and both are stated precisely as quoted inputs in Appendix A.32.3 and named again in Remark A.353. The assembly and the count are Appendix A.32.4, and the answer is checked against the chapter's own worked algebras in Appendix A.32.5.

Throughout, $\mathfrak{g}$ is a finite-dimensional semisimple Lie algebra over a field $\mathbb{K}$ of characteristic zero, with basis $\{T_a\}$, Killing form κ — nondegenerate by Theorem A.330 — and enveloping algebra $U(\mathfrak{g})$ of Definition 14.5, filtered by U_n as in Definition A.320. We write $S(\mathfrak{g})$ for the symmetric algebra of Definition A.325 and ω for the symmetrization map Equation (A.588).

A.32.1 Statement

Theorem A.346 (Racah)**.** *Let $\mathfrak{g}$ be a finite-dimensional semisimple Lie algebra of rank ℓ over an algebraically closed field of characteristic zero. Then there are ℓ Casimir elements $C^{(1)},\dots,C^{(\ell)} \in Z(U(\mathfrak{g}))$, algebraically independent, such that every element of the centre is a polynomial in them with scalar coefficients,*

$$\boxed{Z(U(\mathfrak{g})) = \mathbb{K}\left[C^{(1)},\dots,C^{(\ell)}\right]}, \tag{A.614}$$

and no set of fewer than ℓ elements of the centre generates it. By Corollary 14.18 each $C^{(i)}$ acts on a finite-dimensional irreducible representation as a scalar, so such a representation carries ℓ invariant labels and no more; that these ℓ numbers also separate *the finite-dimensional irreducible representations is the quoted Theorem A.354.*

Rests on Definitions 14.8 and 14.20 and Theorem A.330.

A.32.2 The centre of the enveloping algebra and the invariant polynomials

Lemma A.347 (Central is the same as invariant)**.** *For $u \in U(\mathfrak{g})$ the following are equivalent: u is central; $[u, X] = 0$ for every $X \in \mathfrak{g}$; and u is invariant under the adjoint action $\mathrm{ad}_X\, u = Xu - uX$ of $\mathfrak{g}$ on $U(\mathfrak{g})$. In particular the Casimir elements of Definition 14.8 are exactly the* ad*-invariant elements of $U(\mathfrak{g})$.*

Rests on Definitions 14.5 and 14.8.

Proof. *Derives Lemma A.347.*

The second and third conditions are the same statement written twice. A central element certainly commutes with every $X \in \mathfrak{g}$. Conversely, the set of elements commuting with a fixed u is a subalgebra of $U(\mathfrak{g})$ containing the unit — it is closed under sums and, by the Leibniz rule $[vw, u] = v\,[w, u] + [v, u]\,w$, under products. If it contains $\mathfrak{g}$ it therefore contains the algebra generated by $\mathfrak{g}$ and the unit, which is $U(\mathfrak{g})$ by Definition 14.5. □

Proposition A.348 (Symmetrization identifies the invariants)**.** *Let $S(\mathfrak{g})^{\mathfrak{g}}$ denote the* ad*-invariant elements of $S(\mathfrak{g})$. Then:*

1. *ω restricts to a linear isomorphism $S(\mathfrak{g})^{\mathfrak{g}} \to Z\,(U(\mathfrak{g}))$;*
2. *$S(\mathfrak{g})^{\mathfrak{g}}$ is a graded subalgebra of $S(\mathfrak{g})$, and the symbol map Equation (A.587) identifies the associated graded algebra of the centre with it,*

$$\bigoplus_{n\geq 0} \frac{(Z\,(U(\mathfrak{g})) \cap U_n) + U_{n-1}}{U_{n-1}} \;\cong\; S(\mathfrak{g})^{\mathfrak{g}} \tag{A.615}$$

as graded algebras.

Rests on Proposition A.327, Definition A.325, and Lemma A.347.

Proof. *Derives Proposition A.348.*

(1) By Proposition A.327 the map ω is a linear isomorphism intertwining the two adjoint actions, so it carries the invariants of $S(\mathfrak{g})$ bijectively onto the invariants of $U(\mathfrak{g})$, and the latter are the centre by Lemma A.347.

(2) The action of ad_X on $S(\mathfrak{g})$ is by a derivation that preserves degree, so an invariant decomposes into invariant homogeneous pieces and $S(\mathfrak{g})^{\mathfrak{g}}$ is graded; it is a subalgebra because a derivation annihilating two elements annihilates their product.

For Equation (A.615), note first that ad_X maps U_n into U_n — by the Leibniz rule it replaces one letter of a word of length n by $[X, \cdot]$ of it, which is again a letter — and that the induced map on $U_n/U_{n-1} \cong S^n(\mathfrak{g})$ is exactly the derivation ad_X of $S^n(\mathfrak{g})$. Hence the symbol of a central element of U_n is an invariant of $S^n(\mathfrak{g})$, which gives the inclusion "$\subseteq$" in Equation (A.615). Conversely, if $p \in S^n(\mathfrak{g})^{\mathfrak{g}}$ then $\omega(p)$ lies in U_n, is central by part (1), and has symbol p by Proposition A.327(1); so the inclusion is an equality. The identification is multiplicative because the symbol map is (Definition A.325). □

Lemma A.349 (Lifting generators through the filtration)**.** *Suppose $S(\mathfrak{g})^{\mathfrak{g}}$ is a free polynomial algebra on homogeneous elements $p_1, \dots, p_\ell$ of degrees $d_1, \dots, d_\ell$. Put $C^{(i)} = \omega(p_i) \in Z\,(U(\mathfrak{g})) \cap U_{d_i}$. Then the $C^{(i)}$ are algebraically independent and generate $Z\,(U(\mathfrak{g}))$ as an algebra, so Equation (A.614) holds.*

Rests on Proposition A.348 and Definition A.325.

Proof. *Derives Lemma A.349.*

Give the monomial $\left(C^{(1)}\right)^{\alpha_1}\cdots\left(C^{(\ell)}\right)^{\alpha_\ell}$ the *weight* $\sum_i \alpha_i d_i$; by Equation (A.569) a monomial of weight N lies in U_N, and by Proposition A.348 its symbol in $S^N(\mathfrak{g})$ is the corresponding monomial $p^\alpha = p_1^{\alpha_1}\cdots p_\ell^{\alpha_\ell}$.

Independence. Let F be a nonzero polynomial in ℓ variables and let N be the largest weight carried by a monomial occurring in it. Then $F\left(C^{(1)},\dots,C^{(\ell)}\right)\in U_N$ and its symbol is $\sum_{\text{weight }\alpha=N}\lambda_\alpha p^\alpha$, a nonzero element of $S^N(\mathfrak{g})$ because the p_i are algebraically independent. A nonzero symbol means the element is not in U_{N-1}; in particular it is not zero.

Generation. Let z be central and induct on the smallest n with $z\in U_n$. Its symbol is a homogeneous invariant of degree n, hence a weighted-homogeneous polynomial Q of weight n in the p_i; then $z - Q\left(C^{(1)},\dots,C^{(\ell)}\right)$ is central and lies in U_{n-1}, and the induction hypothesis applies. The base $n=0$ is $z\in\mathbb{K}$. □

Lemma A.350 (Invariant polynomials on $\mathfrak{g}$)**.** *The Killing form, being nondegenerate and invariant, is an isomorphism $\mathfrak{g}\to\mathfrak{g}^*$ of $\mathfrak{g}$-modules, and hence induces an isomorphism of graded algebras*

$$S(\mathfrak{g})^{\mathfrak{g}} \;\cong\; S\left(\mathfrak{g}^*\right)^{\mathfrak{g}}\,, \tag{A.616}$$

the right-hand side being the algebra of polynomial functions on $\mathfrak{g}$ annihilated by the natural action of every ad_X. In components, an element of $S^r(\mathfrak{g})^{\mathfrak{g}}$ is exactly a totally symmetric invariant tensor $k^{a_1\cdots a_r}$ in the sense of Definition 14.14, and the corresponding Casimir element is Equation (14.19).

Rests on Definitions 14.10 and 14.14 and Theorem A.330.

Proof. *Derives Lemma A.350.*

The map $X\mapsto\kappa(X,\cdot)$ is injective because κ is nondegenerate (Theorem A.330) and hence bijective by dimension; it intertwines the adjoint and coadjoint actions precisely because κ is invariant, Equation (14.16). An isomorphism of modules induces one of symmetric algebras, and it carries invariants to invariants. The component statement is Remark 14.15: writing an element of $S^r(\mathfrak{g})$ as $k^{a_1\cdots a_r}T_{a_1}\cdots T_{a_r}$ with k totally symmetric, the vanishing of ad_{T_c} on it is Equation (14.18) term by term, and Corollary A.328 identifies ω of that element with the Casimir Equation (14.19). □

So the first link of Equation (A.613) is established, in the strong form of Lemma A.349: *the enveloping algebra contributes nothing of its own to the count.* Counting Casimir operators is counting the algebraically independent invariant polynomials on $\mathfrak{g}$, and that is a question about the adjoint action alone.

A.32.3 The two quoted theorems

Theorem A.351 (Chevalley restriction theorem, quoted)**.** *Let $\mathfrak{g}$ be a semisimple Lie algebra over an algebraically closed field of characteristic zero, $\mathfrak{h}\subseteq\mathfrak{g}$ a Cartan subalgebra (Definition 14.20), and W the Weyl group of $\mathfrak{g}$ acting on $\mathfrak{h}$. Then restriction of functions from $\mathfrak{g}$ to $\mathfrak{h}$,*

$$S\left(\mathfrak{g}^*\right)^{\mathfrak{g}} \longrightarrow S\left(\mathfrak{h}^*\right)^{W}\,, \qquad P\longmapsto P|_{\mathfrak{h}}\,, \tag{A.617}$$

is an isomorphism of graded algebras.

Rests on Definition 14.20 and Lemma A.350.

Theorem A.352 (Chevalley's theorem on finite reflection groups, quoted)**.** *Let $\mathbb{V}$ be a vector space of dimension ℓ over a field of characteristic zero and let W be a finite group of linear transformations of $\mathbb{V}$ generated by reflections — elements fixing a hyperplane pointwise. Then the algebra of W-invariant polynomial functions on $\mathbb{V}$ is a free polynomial algebra,*

$$S\left(\mathbb{V}^*\right)^{W} = \mathbb{K}\left[f_1,\dots,f_\ell\right]\,, \tag{A.618}$$

on exactly ℓ algebraically independent homogeneous generators $f_1, \dots, f_\ell$, whose degrees are determined by W. Moreover, for $\mathfrak{g}$ semisimple with Cartan subalgebra $\mathfrak{h}$ of dimension ℓ, the Weyl group W is a finite group acting faithfully on $\mathfrak{h}$ and generated by the reflections in the root hyperplanes, so Equation (A.618) applies to it with $\mathbb{V} = \mathfrak{h}$.

Rests on Definition 14.20.

The free-polynomial statement is [1058]. That the Weyl group is a finite reflection group of rank ℓ acting faithfully on $\mathfrak{h}$ is standard structure theory and is quoted with it; no entry of this bibliography carries that half.

Remark A.353 (What is quoted here). Two theorems are assumed and not proved, and it is worth being exact about what they are and why they are not proved here. Theorem A.351 and the last sentence of Theorem A.352 both belong to the root-space theory of semisimple Lie algebras: the decomposition $\mathfrak{g} = \mathfrak{h} \oplus \bigoplus_\alpha \mathfrak{g}_\alpha$ into eigenspaces of $\mathfrak{h}$, the root system it produces, the conjugacy of Cartan subalgebras, and the Weyl group generated by root reflections. That apparatus is the content of several chapters of a book on Lie algebras and this treatise does not build it: Chapter 14 constructs the algebras it needs by hand and computes their ranks directly, as Proposition 14.22 does for $\mathfrak{so}(p, q)$. Equation (A.618) itself is a theorem of invariant theory, not of Lie theory, and is likewise quoted.

What is *not* quoted is the first link of Equation (A.613), which is the part that could plausibly hide a circularity — it is proved in Appendix A.32.2 from the Poincaré–Birkhoff–Witt theorem of Appendix A.30, itself proved from nothing but the Jacobi identity; and it is the part that fails without semisimplicity, which is why Remark 14.23 is careful to exclude the Poincaré algebra. Nor is the arithmetic quoted: the count in Appendix A.32.4 and the check in Appendix A.32.5 are carried out here.

A.32.4 The count

Proof of Theorem A.346. *Derives Theorem A.346.*

Chain the identifications. By Lemma A.350, $S(\mathfrak{g})^{\mathfrak{g}} \cong S(\mathfrak{g}^*)^{\mathfrak{g}}$ as graded algebras; by Theorem A.351 the latter is $S(\mathfrak{h}^*)^W$; and by Theorem A.352 that is a free polynomial algebra on $\ell = \dim \mathfrak{h}$ homogeneous generators $f_1, \dots, f_\ell$, where ℓ is the rank of $\mathfrak{g}$ by Definition 14.20. Transporting the f_i back gives homogeneous $p_1, \dots, p_\ell \in S(\mathfrak{g})^{\mathfrak{g}}$ that are algebraically independent and generate, and Lemma A.349 converts them into Casimir elements $C^{(i)} = \omega(p_i)$ with the properties claimed in Equation (A.614).

No fewer than ℓ. A commutative algebra generated by m elements has transcendence degree at most m over $\mathbb{K}$, since every element is a polynomial in the generators and any $m + 1$ elements of such an algebra are algebraically dependent. By Equation (A.614) the centre has transcendence degree exactly ℓ, the $C^{(i)}$ being algebraically independent. Hence no $m < \ell$ elements of the centre generate it, and no $m < \ell$ Casimir eigenvalues can carry the information that all of them do. □

Theorem A.354 (Harish-Chandra, quoted: the labels separate)**.** *Let $\mathfrak{g}$ be semisimple over an algebraically closed field of characteristic zero. Two finite-dimensional irreducible representations of $\mathfrak{g}$ on which every element of $Z(U(\mathfrak{g}))$ takes the same scalar value are equivalent. Equivalently, the ℓ numbers*

$$c^{(i)} = \text{the eigenvalue of } D\left(C^{(i)}\right), \qquad i = 1, \dots, \ell, \tag{A.619}$$

determine the equivalence class of a finite-dimensional irreducible representation.

Rests on Definition 14.8 and Corollary 14.18.

Remark A.355 (Why the separation statement is quoted separately). Theorem A.346 says how many independent invariants there are; Theorem A.354 says that their values are enough to tell two irreducible representations apart. The two are different assertions and

the second does not follow from the first, which is why the chapter's phrase "a complete set of invariant labels" has been split here. Remark 14.19 makes the same point from the other side: *one* Casimir separates the irreducible representations of $\mathfrak{su}(2)$ and does not separate those of $\mathfrak{su}(3)$, and the reason is not a failure of Theorem A.346 but the fact that $\mathfrak{su}(3)$ has rank 2 — one label is simply not the whole set.

Remark A.356 (Real forms). Theorem A.346 is stated over an algebraically closed field, but the algebras of Chapter 14 are real. The transition is harmless and worth stating once. For a real semisimple $\mathfrak{g}$ one has $U(\mathfrak{g}) \otimes_{\mathbb{R}} \mathbb{C} = U(\mathfrak{g} \otimes_{\mathbb{R}} \mathbb{C})$, because both sides are generated by $\mathfrak{g} \otimes \mathbb{C}$ with the same relations Equation (14.12), and taking the centre commutes with this extension: an element is central if and only if its real and imaginary parts are. So the number of independent Casimir elements of a real form equals that of its complexification, and the rank meant throughout is that of the complexification. This is what Proposition 14.22 in fact computes: its proof passes to $\mathfrak{h}_{\mathbb{C}}$ at once and counts the eigenvalue pairs $\pm\alpha_i$ there, which is the only sensible reading, since the adjoint action of J_{12} on the compact algebra $\mathfrak{so}(3)$ has eigenvalues $\pm\mathrm{i}$ and is not diagonalizable over $\mathbb{R}$ at all.

A.32.5 The count checked against the chapter's algebras

Example A.357 (Rank one: $\mathfrak{su}(2)$)**.** Any single generator spans a maximal abelian subalgebra of $\mathfrak{su}(2) \cong \mathfrak{so}(3)$, whose complexification is $\mathfrak{sl}(2,\mathbb{C})$; the rank is $\ell = 1$, in agreement with Equation (14.22) at $D = 3$. Theorem A.346 then predicts a single Casimir generator, and the invariant found in Definition 14.40 is $\boldsymbol{J}^2$, of degree two — the free generator being f_1 of degree 2. That it generates everything is visible in Theorem 14.43: the eigenvalue $j(j+1)$ is injective in $j \geq 0$, so by Theorem A.354 nothing further is needed, and by Theorem A.346 nothing further exists.

Rests on Theorems 14.43 and A.346.

Example A.358 (Rank two: $\mathfrak{su}(3)$)**.** The diagonal traceless matrices give a maximal abelian subalgebra of dimension 2 (Proposition 14.50), so $\ell = 2$ and Theorem A.346 predicts exactly two independent Casimir elements. They are found in Proposition 14.54: one quadratic, built from κ^{ab}, and one cubic, built from the totally symmetric tensor d^{abc} of Equation (14.70) — degrees 2 and 3, which are the degrees Chevalley's theorem assigns to the invariants of the Weyl group of $\mathfrak{su}(3)$. The count is not idle: by Proposition 14.55 the fundamental and antifundamental representations share the value of the quadratic invariant and are told apart only by the cubic one. One label would have failed; three would have been one too many.

Rests on Theorem A.346 and Propositions 14.54 and 14.55.

Example A.359 (The Lorentz algebra)**.** By Corollary A.344 the algebra $\mathfrak{so}(3,1)$ is semisimple, and by Equation (14.22) its rank is $\lfloor 4/2 \rfloor = 2$. Theorem A.346 therefore gives exactly two independent Casimir operators for the Lorentz algebra of the observed spacetime, and they are the two familiar quadratics,

$$C^{(1)} = \tfrac{1}{2}\, J_{AB} J^{AB}\,, \qquad C^{(2)} = \tfrac{1}{8}\, \epsilon^{ABCD} J_{AB} J_{CD}\,, \tag{A.620}$$

the second existing only because $D = 4$ makes the Levi-Civita symbol carry exactly four indices — the same accident of dimension that Remark 14.67 records for the Pauli–Lubanski vector. Both are invariant by Lemma 14.62 and Corollary 14.12, and by Theorem A.346 there is no third.

Rests on Theorem A.346, Corollary A.344, and Equation (14.22).

Remark A.360 (The Poincaré algebra is outside the theorem, and the agreement is a coincidence). Theorem 14.65 finds two invariants for the Poincaré algebra in $3+1$ dimensions, P^2 and W^2, and it is tempting to read that as Theorem A.346 applied to a rank-two algebra.

It is not, and the temptation must be resisted for a reason that is now provable rather than asserted: the translations form an abelian ideal, so by Proposition A.334 the Killing form of the Poincaré algebra is degenerate, the algebra is not semisimple, and every step of Appendix A.32.2 — which inverts κ in Lemma A.350 — fails at once. That the two counts agree at $D = 4$ is the coincidence recorded in Remark 14.23, and Equation (14.107) shows the agreement breaking in odd dimension, where $\lceil D/2 \rceil \neq \lfloor D/2 \rfloor$. The Galilei algebra of Example 14.80 makes the point sharper still: its invariants include the central mass, which is not an invariant of any semisimple algebra at all.

Where Theorem A.346 *is* used for an inhomogeneous algebra, it is applied to a semisimple subalgebra and not to the whole: the count Equation (14.107) applies it to the little algebra $\mathfrak{so}(D-1)$ of Lemma 14.68, which is semisimple for $D - 1 \geq 3$ by Corollary A.344, and adds the mass by hand.

Appendix A.32 discharges the proof obligation of Theorem 14.21, modulo the two theorems of structure theory named in Remark A.353 and the separation theorem Theorem A.354. It is used at Proposition 14.22 to turn the rank $\lfloor D/2 \rfloor$ into a number of Lorentz invariants, at Remark 14.67 to count the labels of a massive representation in general dimension, and throughout Section 14.2 whenever a representation is labelled by its Casimir eigenvalues. Its physical content is stated in Remark 14.24: the labels are pure numbers, and the measured quantity is the label multiplied by the power of $\hbar$ that carries its SI dimension.

A.33 Covering Spaces and the Fundamental Group of the Rotation Group

This appendix proves Corollary 14.38 of Chapter 14: that $\pi_1(\mathrm{SO}(3,\mathbb{R})) \cong \mathbb{Z}_2$, and more generally that when a simply connected topological group covers another one, the fundamental group of the base is the kernel of the covering homomorphism. Theorem 14.37 already supplies everything algebraic — that $\Phi : \mathrm{SU}(2) \to \mathrm{SO}(3,\mathbb{R})$ is a surjective homomorphism with kernel $\{\mathbb{1}, -\mathbb{1}\}$ — and Proposition 14.33 supplies the topology of the total space, that $\mathrm{SU}(2) \cong S^3$ is compact, connected and simply connected. What is missing, and is built here, is the bridge between the two: the covering-space machinery that converts a discrete kernel upstairs into a fundamental group downstairs.

Only the material of Chapter 6 is assumed: open sets, continuity, connectedness, compactness (Theorem 6.11), and simple connectedness (Definition 6.17). That chapter already carries the technique in its simplest instance. The continuous argument along a path of Lemma 6.19 is precisely the unique lifting of a path through the covering $\mathbb{R} \to \mathbb{C} \setminus \{0\}$, $\theta \mapsto \mathrm{e}^{\mathrm{i}\theta}$ composed with a radius, and the local-constancy argument of Proposition 6.23 is the homotopy invariance of the lift in that instance. The two lemmas below are those two arguments written for an arbitrary covering; the reader who has followed Section 6.1.5 has seen the whole idea already, and what is added here is generality, not a new device.

A.33.1 Statement

Definition A.361 (Covering map)**.** A continuous surjection $p : E \longrightarrow B$ of topological spaces is a *covering map* if every $b \in B$ has an open neighbourhood V that is *evenly covered*: $p^{-1}(V)$ is a union of pairwise disjoint open sets $\{W_\alpha\}$ — the *sheets* over V — each of which p maps homeomorphically onto V. The set $p^{-1}(b)$ is the *fibre* over b; a *lift* of a continuous map $g : X \longrightarrow B$ is a continuous $\tilde{g} : X \longrightarrow E$ with $p \circ \tilde{g} = g$.

Rests on Definitions 6.2 and 6.6.

Definition A.362 (Path homotopy and the fundamental group)**.** Let X be a topological space and $x_0 \in X$. A *loop at* x_0 is a continuous $\gamma : [0,1] \longrightarrow X$ with $\gamma(0) = \gamma(1) = x_0$.

Two loops γ_0, γ_1 at x_0 are *path homotopic*, written $\gamma_0 \simeq \gamma_1$, if there is a continuous $H : [0,1] \times [0,1] \longrightarrow X$ with

$$H(s,0) = \gamma_0(s)\,, \quad H(s,1) = \gamma_1(s)\,, \quad H(0,t) = H(1,t) = x_0\,, \qquad s,t \in [0,1]\,. \tag{A.621}$$

The concatenation $\gamma \cdot \delta$ is the loop equal to $\gamma(2s)$ for $s \le 1/2$ and to $\delta(2s-1)$ for $s \ge 1/2$, and the reverse is $\bar\gamma(s) = \gamma(1-s)$. The set of path-homotopy classes of loops at x_0 with the product $[\gamma][\delta] = [\gamma \cdot \delta]$ is the *fundamental group* $\pi_1(X, x_0)$. A path-connected X is simply connected in the sense of Definition 6.17 exactly when $\pi_1(X, x_0)$ is trivial.

Rests on Definitions 6.15 and 6.17.

Theorem A.363 (Monodromy)**.** *Let $p : E \longrightarrow B$ be a covering map with E path-connected and simply connected, B path-connected, and fix $e_0 \in E$, $b_0 = p(e_0)$. Then the* monodromy map

$$\Psi : \pi_1(B, b_0) \longrightarrow p^{-1}(b_0)\,, \qquad \Psi\left([\gamma]\right) = \tilde\gamma(1)\,, \tag{A.622}$$

where $\tilde\gamma$ is the unique lift of γ with $\tilde\gamma(0) = e_0$, is a well-defined bijection.

Rests on Definitions A.361 and A.362.

Theorem A.364 (Covering homomorphisms)**.** *Let E and B be topological groups, E path-connected and simply connected, and let $p : E \longrightarrow B$ be a surjective continuous homomorphism that is also a covering map, with kernel $N = p^{-1}(1_B)$. Then N is a discrete central subgroup of E, the monodromy map Equation (A.622) based at $e_0 = 1_E$ is a group isomorphism*

$$\pi_1(B, 1_B) \cong N\,, \tag{A.623}$$

and the group of deck transformations of p — the homeomorphisms φ of E with $p \circ \varphi = p$ — consists exactly of the left translations by elements of N, so that it too is isomorphic to N and hence to $\pi_1(B, 1_B)$.

Rests on Theorem A.363 and Definition A.361.

The proof occupies the rest of this section: two lifting lemmas (Appendix A.33.2), the monodromy theorem (Appendix A.33.3), the group case (Appendix A.33.4), and the application to $\mathrm{SU}(2) \to \mathrm{SO}(3, \mathbb{R})$ (Appendix A.33.5).

A.33.2 Two lifting lemmas

Every lifting argument in covering-space theory is the same argument: chop the domain into pieces small enough to sit inside an evenly covered neighbourhood, lift piece by piece using the sheet that contains the value already fixed, and check that the pieces agree. The chopping is supplied by the following standard consequence of compactness, which Lemma 6.19 used in the disguised form of a partition of mesh smaller than a modulus of uniform continuity.

Lemma A.365 (Lebesgue number)**.** *Let $K \subset \mathbb{R}^N$ be closed and bounded and let $\{U_i\}$ be an open cover of K. There is $\delta > 0$ such that every subset of K of diameter smaller than δ is contained in a single U_i.*

Rests on Theorem 6.11 and Definition 6.2.

Proof. *Derives Lemma A.365.*

For each $x \in K$ choose $i(x)$ with $x \in U_{i(x)}$ and then $r(x) > 0$ with $B_{2r(x)}(x) \subseteq U_{i(x)}$, which is possible because $U_{i(x)}$ is open. The balls $B_{r(x)}(x)$ cover K, which is compact by Theorem 6.11, so finitely many of them suffice, say those centred at $x_1, \dots, x_k$; put $\delta = \min_j r(x_j) > 0$. Let $S \subseteq K$ have diameter smaller than δ and pick $y \in S$. Then $y \in B_{r(x_j)}(x_j)$ for some j, and every $z \in S$ satisfies

$$|z - x_j| \le |z - y| + |y - x_j| < \delta + r(x_j) \le 2r(x_j)\,,$$

so $S \subseteq B_{2r(x_j)}(x_j) \subseteq U_{i(x_j)}$. □

Lemma A.366 (Unique path lifting)**.** *Let $p : E \longrightarrow B$ be a covering map, let $\gamma : [0,1] \longrightarrow B$ be continuous, and let $e \in E$ satisfy $p(e) = \gamma(0)$. There is exactly one continuous $\tilde{\gamma} : [0,1] \longrightarrow E$ with $p \circ \tilde{\gamma} = \gamma$ and $\tilde{\gamma}(0) = e$.*

Rests on Definition A.361 and Lemma A.365.

Proof. *Derives Lemma A.366.*

Existence. Cover B by evenly covered open sets V and pull them back: the sets $\gamma^{-1}(V)$ form an open cover of $[0,1]$, which is closed and bounded, so Lemma A.365 supplies a $\delta > 0$ and with it a partition $0 = t_0 < t_1 < \cdots < t_M = 1$ of mesh smaller than δ such that each $\gamma([t_{j-1}, t_j])$ lies in a single evenly covered V_j.

Define $\tilde{\gamma}$ inductively. Set $\tilde{\gamma}(t_0) = e$. Suppose $\tilde{\gamma}$ has been defined and is continuous on $[0, t_{j-1}]$ with $p \circ \tilde{\gamma} = \gamma$ there. The point $\tilde{\gamma}(t_{j-1})$ lies in $p^{-1}(V_j)$, hence in exactly one sheet W over V_j, and $p|_W : W \longrightarrow V_j$ is a homeomorphism. Put

$$\tilde{\gamma}(t) = (p|_W)^{-1}(\gamma(t)) \,, \qquad t \in [t_{j-1}, t_j] \,. \tag{A.624}$$

This is continuous on $[t_{j-1}, t_j]$, satisfies $p \circ \tilde{\gamma} = \gamma$ there, and agrees at t_{j-1} with the value already assigned, because $\tilde{\gamma}(t_{j-1}) \in W$ and $p|_W$ is injective. Two continuous functions on the closed intervals $[0, t_{j-1}]$ and $[t_{j-1}, t_j]$ agreeing at the common endpoint patch to a continuous function on the union, so the induction proceeds and terminates at $t_M = 1$.

Uniqueness. Let $\tilde{\gamma}$ and $\tilde{\gamma}'$ be two lifts with the same initial value, and let $A = \{t \in [0,1] \mid \tilde{\gamma}(t) = \tilde{\gamma}'(t)\}$, which contains 0. Fix $t \in [0,1]$, let V be an evenly covered neighbourhood of $\gamma(t)$, and let W, W' be the sheets over V containing $\tilde{\gamma}(t)$ and $\tilde{\gamma}'(t)$. The set $O = \tilde{\gamma}^{-1}(W) \cap (\tilde{\gamma}')^{-1}(W')$ is an open neighbourhood of t in $[0,1]$. If $t \in A$ then $W = W'$, the sheets being disjoint, and on O both lifts equal $(p|_W)^{-1} \circ \gamma$, so $O \subseteq A$: the set A is open. If $t \notin A$ then $W \neq W'$, since a common sheet would force $\tilde{\gamma}(t) = \tilde{\gamma}'(t)$ by injectivity of $p|_W$; the sheets being disjoint, O misses A entirely, so the complement of A is open too. Thus A is a nonempty subset of $[0,1]$ that is both open and closed, and $[0,1]$ is connected (Theorem 6.14), so $A = [0,1]$. □

Lemma A.367 (Homotopy lifting)**.** *Let $p : E \longrightarrow B$ be a covering map, let $H : [0,1] \times [0,1] \longrightarrow B$ be continuous, and let $e \in E$ satisfy $p(e) = H(0,0)$. There is exactly one continuous $\tilde{H} : [0,1] \times [0,1] \longrightarrow E$ with $p \circ \tilde{H} = H$ and $\tilde{H}(0,0) = e$. If moreover H is a path homotopy in the sense of Equation (A.621) — so that $H(0,t)$ and $H(1,t)$ are constant in t — then $\tilde{H}(0,t)$ and $\tilde{H}(1,t)$ are constant in t as well.*

Rests on Lemmas A.365 and A.366.

Proof. *Derives Lemma A.367.*

Existence. As before, the sets $H^{-1}(V)$ with V evenly covered form an open cover of the square $[0,1]^2$, which is closed and bounded in $\mathbb{R}^2$; Lemma A.365 supplies $\delta > 0$, and a grid $0 = s_0 < \cdots < s_M = 1$, $0 = t_0 < \cdots < t_M = 1$ of mesh smaller than $\delta/\sqrt{2}$ makes each closed cell $R_{jk} = [s_{j-1}, s_j] \times [t_{k-1}, t_k]$ have diameter smaller than δ, hence $H(R_{jk})$ contained in a single evenly covered V_{jk}.

Order the cells lexicographically, bottom row left to right, then the next row, and lift them one at a time. At each stage the part of the boundary of the current cell R on which $\tilde{H}$ has already been defined — call it C — is the union of at most the left edge and the bottom edge of R, together with the corner joining them, and is therefore connected. (For the very first cell C is the single point (s_0, t_0), where $\tilde{H} = e$ is prescribed.) The set $\tilde{H}(C)$ is connected and lies in $p^{-1}(V)$ for the evenly covered $V \supseteq H(R)$, which is the disjoint union of open sheets; a connected subset of a disjoint union of open sets lies in one of them, so $\tilde{H}(C) \subseteq W$ for a single sheet W. Define $\tilde{H} = (p|_W)^{-1} \circ H$ on R. This is continuous on R, projects to H, and agrees on C with what was there — again because $p|_W$ is injective and both values lie in W. The extended function is continuous, being continuous on each of

finitely many closed cells and consistent on the overlaps. After the last cell $\tilde{H}$ is defined on the whole square.

Uniqueness. Verbatim the argument of Lemma A.366: the agreement set of two lifts is open and closed, and the square is connected, being path-connected (Proposition 6.16).

The vertical edges. Suppose $H(0,t) = b_0$ for every t. Then $t \mapsto \tilde{H}(0,t)$ is a lift of the constant path at b_0; so is the constant map $t \mapsto \tilde{H}(0,0)$; the two agree at $t = 0$, so by the uniqueness clause of Lemma A.366 they agree throughout. The same applies at $s = 1$. □

A.33.3 The monodromy theorem

The elementary bookkeeping behind Definition A.362 is recorded first, because it is what makes π_1 a group at all and because the injectivity of Ψ uses it.

Lemma A.368 (The concatenation rules)**.** *Let α, β, η be paths in a space X with $\alpha(1) = \beta(0)$ and $\beta(1) = \eta(0)$, let c_x denote the constant path at x, and write $\simeq$ for homotopy with endpoints held fixed. Then*

$$(\alpha \cdot \beta) \cdot \eta \simeq \alpha \cdot (\beta \cdot \eta)\,, \qquad c_{\alpha(0)} \cdot \alpha \simeq \alpha \simeq \alpha \cdot c_{\alpha(1)}\,, \qquad \alpha \cdot \bar{\alpha} \simeq c_{\alpha(0)}\,. \tag{A.625}$$

Consequently $\pi_1(X, x_0)$ of Definition A.362 is a group, with $[\gamma]^{-1} = [\bar{\gamma}]$; and a continuous $p : E \longrightarrow B$ with $p(e_0) = b_0$ induces a group homomorphism $p_* : \pi_1(E, e_0) \longrightarrow \pi_1(B, b_0)$, $p_*[\sigma] = [p \circ \sigma]$.

Rests on Definition A.362.

Proof. *Derives Lemma A.368.*

All three relations of Equation (A.625) are reparametrizations. If $\phi : [0,1] \to [0,1]$ is continuous with $\phi(0) = 0$ and $\phi(1) = 1$, then for any path μ the map $H(s,t) = \mu\left((1-t)\phi(s) + ts\right)$ is continuous, fixes the endpoints, and joins $\mu \circ \phi$ to μ; so $\mu \circ \phi \simeq \mu$. The two sides of the first relation are the same path $\alpha \cdot (\beta \cdot \eta)$ composed with such a ϕ — the piecewise-linear map sending $1/4$ to $1/2$ and $1/2$ to $3/4$ — and the two sides of the second are likewise related by the piecewise-linear ϕ sending $1/2$ to 0, respectively to 1. For the third, put

$$H(s,t) = \begin{cases} \alpha\left(2s(1-t)\right)\,, & s \le \frac{1}{2}\,, \\ \alpha\left(2(1-s)(1-t)\right)\,, & s \ge \frac{1}{2}\,, \end{cases}$$

which is continuous (the two formulas agree at $s = 1/2$), equals $\alpha \cdot \bar{\alpha}$ at $t = 0$ and $c_{\alpha(0)}$ at $t = 1$, and is constant on $s = 0$ and on $s = 1$. That the product on classes is well defined — homotopies of the two factors concatenate to a homotopy of the product — is immediate from the definition of concatenation. Finally $p \circ (\alpha \cdot \beta) = (p \circ \alpha) \cdot (p \circ \beta)$ by inspection, and $p \circ H$ is a path homotopy whenever H is, so p_* is a well-defined homomorphism. □

Proof of Theorem A.363. *Derives Theorem A.363.*

Ψ is well defined. The lift $\tilde{\gamma}$ exists and is unique by Lemma A.366, and $p\left(\tilde{\gamma}(1)\right) = \gamma(1) = b_0$, so $\tilde{\gamma}(1) \in p^{-1}(b_0)$. That it depends only on the class $[\gamma]$ is the content of Lemma A.367: let H be a path homotopy from γ_0 to γ_1 and $\tilde{H}$ its lift with $\tilde{H}(0,0) = e_0$. The left edge $t \mapsto \tilde{H}(0,t)$ is constant, so $\tilde{H}(\cdot, t)$ is for each t the lift of $H(\cdot, t)$ starting at e_0; the right edge $t \mapsto \tilde{H}(1,t)$ is constant too, so the common endpoint $\tilde{H}(1,0) = \tilde{H}(1,1)$, that is $\tilde{\gamma}_0(1) = \tilde{\gamma}_1(1)$.

Ψ is surjective. Let $e \in p^{-1}(b_0)$. Since E is path-connected there is a path σ in E from e_0 to e; then $\gamma = p \circ \sigma$ is a loop at b_0, and σ is its lift starting at e_0 by uniqueness, so $\Psi\left([\gamma]\right) = \sigma(1) = e$.

Ψ is injective. Suppose $\tilde{\gamma}_0(1) = \tilde{\gamma}_1(1)$ for loops γ_0, γ_1 at b_0. Both lifts start at e_0, so $\tilde{\gamma}_0$ and $\tilde{\gamma}_1$ are paths in E with the same two endpoints, and $\sigma = \tilde{\gamma}_0 \cdot \overline{\tilde{\gamma}_1}$ is a loop in E at e_0.

Since E is simply connected, $[\sigma]$ is the identity of $\pi_1(E, e_0)$. Apply the homomorphism p_* of Lemma A.368:

$$[\gamma_0]\,[\gamma_1]^{-1} = [\gamma_0 \cdot \bar{\gamma}_1] = [p \circ \sigma] = p_*\,[\sigma] = [c_{b_0}]\ , \tag{A.626}$$

using $p \circ \tilde{\gamma}_k = \gamma_k$ and $[\bar{\gamma}_1] = [\gamma_1]^{-1}$. Hence $[\gamma_0] = [\gamma_1]$. □

A.33.4 Covering homomorphisms and deck transformations

Proof of Theorem A.364. *Derives Theorem A.364.*

N is discrete and central. Let V be an evenly covered neighbourhood of 1_B and W the sheet over it containing 1_E. Then $W \cap N = \{1_E\}$, because p is injective on W and already sends 1_E to 1_B; translating, $nW \cap N = \{n\}$ for every $n \in N$, so N carries the discrete topology. For centrality, fix $n \in N$ and consider $f : E \longrightarrow E$, $f(x) = xnx^{-1}$. It is continuous, takes values in N — because $p\,(xnx^{-1}) = p(x)\,1_B\,p(x)^{-1} = 1_B$ — and E is connected, so its image is a connected subset of a discrete space, hence the single point $f(1_E) = n$. Thus $xn = nx$ for all x.

Ψ is a homomorphism. Let γ, δ be loops at 1_B with lifts $\tilde{\gamma}, \tilde{\delta}$ starting at 1_E, and put $n = \tilde{\gamma}(1) \in N$. The path $t \mapsto n\,\tilde{\delta}(t)$ is continuous, starts at $n\,1_E = n = \tilde{\gamma}(1)$, and projects to $p(n)\,\delta(t) = \delta(t)$. Hence the concatenation $\tilde{\gamma} \cdot \left(n\tilde{\delta}\right)$ is a lift of $\gamma \cdot \delta$ starting at 1_E, and by uniqueness (Lemma A.366) it is *the* lift. Its endpoint is $n\,\tilde{\delta}(1)$, so

$$\Psi\,([\gamma][\delta]) = \Psi\,([\gamma])\ \Psi\,([\delta])\ . \tag{A.627}$$

Being a bijection by Theorem A.363, Ψ is an isomorphism, which is Equation (A.623).

Deck transformations. If $n \in N$ then left translation $L_n(x) = nx$ is a homeomorphism of E with $p\,(L_n(x)) = p(n)p(x) = p(x)$, so L_n is a deck transformation, and $n \mapsto L_n$ is an injective homomorphism ($L_n = L_{n'}$ forces $n = n'$ at $x = 1_E$). Conversely let φ be any deck transformation and set $n = \varphi(1_E)$, which lies in N because $p(n) = p(1_E) = 1_B$. Both φ and L_n are lifts of the map $p : E \longrightarrow B$ through p itself, and they agree at 1_E. The agreement set of two lifts of one map on a connected domain is open and closed — the argument is word for word the uniqueness half of Lemma A.366, with $[0,1]$ replaced by E — and E is connected, so $\varphi = L_n$. □

A.33.5 The rotation group

It remains to check that Φ of Equation (14.46) is a covering map. That is the one place where the group is used concretely, and it is short.

Lemma A.369 (Φ is a two-sheeted covering)**.** *The homomorphism $\Phi : \mathrm{SU}(2) \longrightarrow \mathrm{SO}(3,\mathbb{R})$ of Theorem 14.37 is an open map and a covering map, each fibre having exactly two points.*
Rests on Theorem 14.37, Proposition 14.33, and Definition A.361.

Proof. *Derives Lemma A.369.*

Write $K = \{\mathbb{1}, -\mathbb{1}\} = \ker\Phi$ (Equation (14.47)). Because Φ is a homomorphism with kernel K, its fibres are exactly the pairs $\{U, -U\}$, which have two elements since $U \neq -U$; and Φ is surjective by Theorem 14.37.

Φ is open. First, Φ is a closed map: $\mathrm{SU}(2)$ is compact (Proposition 14.33) and $\mathrm{SO}(3,\mathbb{R})$ is Hausdorff, being a subspace of the matrices $\mathbb{R}^{3\times3}$, so a closed subset of $\mathrm{SU}(2)$ is compact, its continuous image is compact, and a compact subset of a Hausdorff space is closed. A continuous closed surjection is a quotient map: a set $S \subseteq \mathrm{SO}(3,\mathbb{R})$ with $\Phi^{-1}(S)$ open has open complement's preimage closed, hence $\mathrm{SO}(3,\mathbb{R}) \setminus S = \Phi\,(\mathrm{SU}(2) \setminus \Phi^{-1}(S))$ closed — the equality holding because Φ is surjective and $\Phi^{-1}(S)$ is a union of fibres — so S is open. Now let $O \subseteq \mathrm{SU}(2)$ be open. Then $-O$ is open, multiplication by $-\mathbb{1}$ being a homeomorphism, and $\Phi^{-1}\,(\Phi(O)) = O \cup (-O)$ because the fibre through U is $\{U, -U\}$. That union is open, so $\Phi(O)$ is open.

Even covering. Let $V_0 = \{U \in \mathrm{SU}(2) \mid \|U - \mathbb{1}\| < 1\}$, the norm being any norm on the 2×2 matrices with $\|2\mathbb{1}\| = 2$; for instance $\|M\|^2 = \frac{1}{2}\operatorname{tr}\left(M^\dagger M\right)$, for which $\|\mathbb{1}\| = 1$. Then $V_0 \cap (-V_0) = \varnothing$: if U and $-U$ both lay within distance 1 of $\mathbb{1}$ then

$$2 = \|2\mathbb{1}\| = \|(\mathbb{1} - U) + (\mathbb{1} + U)\| \leq \|\mathbb{1} - U\| + \|\mathbb{1} + U\| < 2\,,$$

which is absurd. Fix $U_0 \in \mathrm{SU}(2)$ and put $W = U_0 V_0$, $V = \Phi(W)$, which is open by the previous paragraph. The two sheets are W and $-W$: they are disjoint, since $U_0 v = -U_0 v'$ would give $v = -v'$ with $v, v' \in V_0$; their union is $\Phi^{-1}(V)$, because a point of $\Phi^{-1}(V)$ is $\pm w$ for some $w \in W$; and Φ restricted to either is a continuous open bijection onto V, hence a homeomorphism. Every point of $\mathrm{SO}(3,\mathbb{R})$ lies in such a V, namely for U_0 any preimage of it. □

Proof of Corollary 14.38. *Derives Corollary 14.38.*

By Theorem 14.37 the map Φ is a surjective homomorphism with kernel $\{\mathbb{1}, -\mathbb{1}\}$, so the induced map $\mathrm{SU}(2)/\{\mathbb{1}, -\mathbb{1}\} \longrightarrow \mathrm{SO}(3,\mathbb{R})$ is a group isomorphism, and it is a homeomorphism because Φ is continuous, open and surjective (Lemma A.369). By Proposition 14.33 the group $\mathrm{SU}(2)$ is path-connected and simply connected, and by Lemma A.369 the map Φ is a covering map; $\mathrm{SO}(3,\mathbb{R})$ is path-connected, being a continuous image of a path-connected space. Theorem A.364 therefore applies and gives

$$\pi_1\left(\mathrm{SO}(3,\mathbb{R}), \mathbb{1}\right) \cong \ker \Phi = \{\mathbb{1}, -\mathbb{1}\} \cong \mathbb{Z}_2\,, \tag{A.628}$$

which is Equation (14.48), and identifies the deck-transformation group of the covering with the same $\mathbb{Z}_2$, generated by $U \mapsto -U$.

The concrete description in Corollary 14.38 is now read off the isomorphism. The path $\phi \mapsto U(\hat{n}, \phi)$, $0 \leq \phi \leq 2\pi$, runs in $\mathrm{SU}(2)$ from $\mathbb{1}$ to $-\mathbb{1}$ by Equation (14.45); it is therefore the lift, starting at $\mathbb{1}$, of the loop $\phi \mapsto \Phi\left(U(\hat{n}, \phi)\right) = R(\hat{n}, \phi)$ in $\mathrm{SO}(3,\mathbb{R})$, whose endpoint $R(\hat{n}, 2\pi) = \mathbb{1}$ makes it a loop indeed. Its monodromy image is $-\mathbb{1} \neq \mathbb{1}$, so its class in π_1 is the nontrivial element: the loop is not contractible. Doubling the parameter range to $0 \leq \phi \leq 4\pi$ doubles the class, and $(-\mathbb{1})^2 = \mathbb{1}$, so that loop is contractible. □

Remark A.370 (What this costs, and what it buys). Nothing above is special to $\mathrm{SU}(2)$ until Lemma A.369: the two lifting lemmas and Theorems A.363 and A.364 hold for any covering, and the reader who compares them with Lemmas 6.19 and 6.22 and Proposition 6.23 will find the same three steps — lift along a partition, patch, and use connectedness of the parameter interval to rule out ambiguity — carried out there for the single covering $\theta \mapsto \mathrm{e}^{\mathrm{i}\theta}$ of the unit circle. The winding number of Definition 6.20 is the monodromy map Equation (A.622) for that covering, whose total space $\mathbb{R}$ is simply connected and whose kernel is $2\pi\mathbb{Z}$; that is why Equation (6.5) is an integer.

The purchase is Equation (14.48), and through it Remark 14.39: the rotation group is not simply connected, its universal cover is $\mathrm{SU}(2)$, and a quantum system may therefore carry a representation of $\mathrm{SU}(2)$ that is only a projective representation of $\mathrm{SO}(3,\mathbb{R})$. Theorem 14.43 and Proposition 14.45 say which ones those are — exactly the half-integer j — so the existence of half-integer angular momentum is a corollary of Equation (A.628) and not an independent postulate.

A.34 The Structure Constants of $\mathfrak{su}(3)$

This appendix completes Proposition 14.52 of Chapter 14: it computes every nonvanishing component of the totally antisymmetric array f_{abc} and the totally symmetric array d_{abc} of Equation (14.67) in the Gell-Mann basis of Definition 14.51, and exhibits the two complete

tables. The chapter carries two entries as specimens, f_{458} and d_{247}; the remaining twenty-three are obtained here, and — more to the point — so is the statement that there are no others. Nothing beyond Definition 14.51 and Equation (14.66) is used.

The naive route is $8^3 = 512$ traces, or $\binom{8+2}{3} = 120$ after the symmetries are imposed. That is not what is done below. A single observation about which products of three matrix units have a nonzero trace cuts the work to three short families, and the classification of those families is what makes the tables *exhaustive* rather than merely checked.

A.34.1 One trace for both arrays

Lemma A.371 (The generating trace)**.** *With $T_a = \frac{1}{2}\lambda_a$ as in Definition 14.51,*

$$\boxed{\operatorname{tr}(\lambda_a\lambda_b\lambda_c) = 2\,(d_{abc} + \mathrm{i}\, f_{abc})}\,, \tag{A.629}$$

so that $d_{abc} = \frac{1}{2}\Re\operatorname{tr}(\lambda_a\lambda_b\lambda_c)$ and $f_{abc} = \frac{1}{2}\Im\operatorname{tr}(\lambda_a\lambda_b\lambda_c)$; in particular both arrays are real.
Rests on Equations (14.66) and (14.68) and Definition 14.51.

Proof. *Derives Lemma A.371.*

Substituting $T_a = \frac{1}{2}\lambda_a$ into Equation (14.68), each of the three factors contributes a $\frac{1}{2}$, so

$$f_{abc} = -\frac{\mathrm{i}}{4}\operatorname{tr}\left([\lambda_a, \lambda_b]\,\lambda_c\right), \qquad d_{abc} = \frac{1}{4}\operatorname{tr}\left(\{\lambda_a, \lambda_b\}\,\lambda_c\right). \tag{A.630}$$

Now $\lambda_a\lambda_b = \frac{1}{2}\{\lambda_a, \lambda_b\} + \frac{1}{2}[\lambda_a, \lambda_b]$, so

$$\operatorname{tr}(\lambda_a\lambda_b\lambda_c) = \tfrac{1}{2}\operatorname{tr}\left(\{\lambda_a, \lambda_b\}\,\lambda_c\right) + \tfrac{1}{2}\operatorname{tr}\left([\lambda_a, \lambda_b]\,\lambda_c\right) = \tfrac{1}{2}\,(4d_{abc}) + \tfrac{1}{2}\,(4\mathrm{i}f_{abc}),$$

which is Equation (A.629). Reality: the λ_a are Hermitian, so $\overline{\operatorname{tr}(\lambda_a\lambda_b\lambda_c)} = \operatorname{tr}\left((\lambda_a\lambda_b\lambda_c)^\dagger\right) = \operatorname{tr}(\lambda_c\lambda_b\lambda_a)$, and reversing the order of three factors under the trace exchanges a and c in Equation (A.629); by the total symmetry of d and total antisymmetry of f established in Proposition 14.52, that conjugates the right-hand side. Hence d and f are the real and imaginary parts of one complex number, and both are real numbers. □

Remark A.372 (The symmetry conventions, and how many components there are)*.* Proposition 14.52 proves that f_{abc} is totally antisymmetric and d_{abc} totally symmetric in (a, b, c). Those two rules fix every component from the tables below, and it is worth saying exactly how many components each independent entry generates.

A nonzero f_{abc} necessarily has three *distinct* indices — $f_{aab} = -f_{aab} = 0$ — and its six permutations give six nonzero components, three equal to $+f_{abc}$ (the cyclic ones) and three equal to $-f_{abc}$. With nine independent entries that is $9 \times 6 = 54$ nonvanishing components of f.

For d the count depends on the repetition pattern. An entry with three distinct indices generates six equal components; an entry of the shape d_{aab} with $a \neq b$ generates three; and d_{aaa} generates one. Of the sixteen independent entries below, four have distinct indices, eleven have exactly one repetition, and one is d_{888}; so d has $4 \times 6 + 11 \times 3 + 1 = 58$ nonvanishing components.

Ordering within an entry is therefore free, and the tables list each independent entry once, in the index order used in Equations (14.69) and (14.70).

A.34.2 Which triples can be nonzero at all

Write E_{jk} for the 3×3 matrix whose only nonzero entry is a 1 in row j and column k, so that $E_{jk}E_{lm} = \delta_{kl}E_{jm}$ and $\operatorname{tr} E_{jk} = \delta_{jk}$. Reading off Equations (14.63) to (14.65), each

Gell-Mann matrix is supported on one of four index sets:

$$\begin{aligned} \lambda_1 &= E_{12} + E_{21}\,, & \lambda_2 &= -\mathrm{i}E_{12} + \mathrm{i}E_{21}\,, \\ \lambda_4 &= E_{13} + E_{31}\,, & \lambda_5 &= -\mathrm{i}E_{13} + \mathrm{i}E_{31}\,, \\ \lambda_6 &= E_{23} + E_{32}\,, & \lambda_7 &= -\mathrm{i}E_{23} + \mathrm{i}E_{32}\,, \\ \lambda_3 &= E_{11} - E_{22}\,, & \lambda_8 &= \tfrac{1}{\sqrt{3}}\,(E_{11} + E_{22} - 2E_{33})\ . \end{aligned} \tag{A.631}$$

Read as a graph on the vertex set $\{1,2,3\}$, the first row lives on the edge 1–2, the second on 1–3, the third on 2–3, and the fourth — λ_3 and λ_8, the diagonal ones — on loops. Call $\mathrm{A} = \{1,2\}$, $\mathrm{B} = \{4,5\}$, $\mathrm{C} = \{6,7\}$ the three *off-diagonal pairs* of generator labels and $\mathrm{D} = \{3,8\}$ the *diagonal pair*.

Lemma A.373 (Selection rule)**.** $\operatorname{tr}(\lambda_a\lambda_b\lambda_c) = 0$ *unless the unordered triple* $\{a,b,c\}$ *falls into one of the three families*

1. *all three indices in* D*;*
2. *exactly one index in* D *and the other two in a single off-diagonal pair;*
3. *one index in each of* A, B *and* C*.*

Rests on Equations (14.63) *to* (14.65)*.*

Proof. *Derives Lemma A.373.*

Expanding each factor in the matrix units of Equation (A.631),

$$\operatorname{tr}(\lambda_a\lambda_b\lambda_c) = \sum_{i,j,k} (\lambda_a)_{ij}\,(\lambda_b)_{jk}\,(\lambda_c)_{ki}\ , \tag{A.632}$$

so a nonzero contribution needs a closed walk $i \to j \to k \to i$ of length three in the index set $\{1,2,3\}$, whose three steps are supplied by λ_a, λ_b, λ_c in that order. A λ with index in D supplies only *loops* $i \to i$; a λ in an off-diagonal pair supplies only the two steps along its own edge, and along no other.

Count the loops among the three steps. If all three steps are loops, all three indices lie in D: family (1). If exactly two are loops, the third is an edge step, which cannot close a walk that has returned to its starting vertex twice — a walk $i \to i \to i \to j$ ends at $j \neq i$ — so this case contributes nothing. If exactly one step is a loop, the other two are edge steps $u \to v$ and $v \to u$ on one and the same edge, since the walk must return; hence one index in D and two in a single pair, which is family (2). If no step is a loop, the walk visits three vertices along three distinct edges of the triangle on $\{1,2,3\}$ — two of the three steps on one edge would force the third to be a loop — so one index comes from each of A, B, C: family (3). □

Everything is now a finite enumeration inside three small families, and the products needed are the ten diagonal matrices

$$\begin{aligned} \lambda_1\lambda_1 &= \lambda_2\lambda_2 = \operatorname{diag}(1,1,0)\,, & \lambda_1\lambda_2 &= \operatorname{diag}(\mathrm{i},-\mathrm{i},0)\,, \\ \lambda_4\lambda_4 &= \lambda_5\lambda_5 = \operatorname{diag}(1,0,1)\,, & \lambda_4\lambda_5 &= \operatorname{diag}(\mathrm{i},0,-\mathrm{i})\,, \\ \lambda_6\lambda_6 &= \lambda_7\lambda_7 = \operatorname{diag}(0,1,1)\,, & \lambda_6\lambda_7 &= \operatorname{diag}(0,\mathrm{i},-\mathrm{i})\,, \end{aligned} \tag{A.633}$$

each read straight off Equation (A.631): for instance $\lambda_6\lambda_7 = (E_{23} + E_{32})\,(-\mathrm{i}E_{23} + \mathrm{i}E_{32}) = \mathrm{i}E_{23}E_{32} - \mathrm{i}E_{32}E_{23} = \mathrm{i}E_{22} - \mathrm{i}E_{33}$.

A.34.3 Family (1): three diagonal indices

Here λ_3 and λ_8 commute, so every commutator vanishes and $f_{abc} = 0$ throughout the family. For d the four independent triples are 333, 338, 388, 888, and Equation (A.629) needs only a trace of a product of diagonal matrices:

$$\begin{aligned}
\operatorname{tr}\left(\lambda_3^3\right) &= \operatorname{tr}\operatorname{diag}(1,-1,0) = 0\,,\\
\operatorname{tr}\left(\lambda_3^2\lambda_8\right) &= \operatorname{tr}\left[\operatorname{diag}(1,1,0)\,\tfrac{1}{\sqrt{3}}\operatorname{diag}(1,1,-2)\right] = \tfrac{1}{\sqrt{3}}\,(1+1+0) = \tfrac{2}{\sqrt{3}}\,,\\
\operatorname{tr}\left(\lambda_3\lambda_8^2\right) &= \operatorname{tr}\left[\tfrac{1}{3}\operatorname{diag}(1,-1,0)\operatorname{diag}(1,1,4)\right] = \tfrac{1}{3}\,(1-1+0) = 0\,,\\
\operatorname{tr}\left(\lambda_8^3\right) &= \tfrac{1}{3\sqrt{3}}\operatorname{tr}\operatorname{diag}(1,1,-8) = \tfrac{-6}{3\sqrt{3}} = -\tfrac{2}{\sqrt{3}}\,.
\end{aligned}$$

Halving each, the family contributes

$$d_{333} = d_{388} = 0\,, \qquad d_{338} = \frac{1}{\sqrt{3}}\,, \qquad d_{888} = -\frac{1}{\sqrt{3}}\,. \tag{A.634}$$

A.34.4 Family (2): one diagonal index and one off-diagonal pair

Put the diagonal generator last, so that the trace is that of $(\lambda_a\lambda_b)\,\lambda_\delta$ with a, b in one pair and $\delta \in \{3, 8\}$; the first factor is one of the six diagonal matrices Equation (A.633). There are $3 \times 3 \times 2 = 18$ independent triples — three pairs, three choices of (a, b) up to the symmetries, two diagonal generators — and each is one line. Writing $\lambda_3 = \operatorname{diag}(1,-1,0)$ and $\lambda_8 = \frac{1}{\sqrt{3}}\operatorname{diag}(1,1,-2)$:

Pair $\mathrm{A} = \{1,2\}$*, on* $\operatorname{diag}(\cdot,\cdot,0)$.

$$\begin{aligned}
&\operatorname{tr}(\lambda_1\lambda_1\lambda_3) = \operatorname{tr}\operatorname{diag}(1,-1,0) = 0 && \Rightarrow \quad d_{113} = d_{223} = 0\,,\\
&\operatorname{tr}(\lambda_1\lambda_1\lambda_8) = \tfrac{1}{\sqrt{3}}\operatorname{tr}\operatorname{diag}(1,1,0) = \tfrac{2}{\sqrt{3}} && \Rightarrow \quad d_{118} = d_{228} = \tfrac{1}{\sqrt{3}}\,,\\
&\operatorname{tr}(\lambda_1\lambda_2\lambda_3) = \operatorname{tr}\operatorname{diag}(\mathrm{i},\mathrm{i},0) = 2\mathrm{i} && \Rightarrow \quad d_{123} = 0\,,\ f_{123} = 1\,,\\
&\operatorname{tr}(\lambda_1\lambda_2\lambda_8) = \tfrac{1}{\sqrt{3}}\operatorname{tr}\operatorname{diag}(\mathrm{i},-\mathrm{i},0) = 0 && \Rightarrow \quad d_{128} = f_{128} = 0\,.
\end{aligned}$$

The second line covers $\lambda_2\lambda_2$ as well, since $\lambda_2^2 = \lambda_1^2$ by Equation (A.633).

Pair $\mathrm{B} = \{4,5\}$*, on* $\operatorname{diag}(\cdot,0,\cdot)$.

$$\begin{aligned}
&\operatorname{tr}(\lambda_4\lambda_4\lambda_3) = \operatorname{tr}\operatorname{diag}(1,0,0) = 1 && \Rightarrow \quad d_{344} = d_{355} = \tfrac{1}{2}\,,\\
&\operatorname{tr}(\lambda_4\lambda_4\lambda_8) = \tfrac{1}{\sqrt{3}}\operatorname{tr}\operatorname{diag}(1,0,-2) = -\tfrac{1}{\sqrt{3}} && \Rightarrow \quad d_{448} = d_{558} = -\tfrac{1}{2\sqrt{3}}\,,\\
&\operatorname{tr}(\lambda_4\lambda_5\lambda_3) = \operatorname{tr}\operatorname{diag}(\mathrm{i},0,0) = \mathrm{i} && \Rightarrow \quad d_{345} = 0\,,\ f_{345} = \tfrac{1}{2}\,,\\
&\operatorname{tr}(\lambda_4\lambda_5\lambda_8) = \tfrac{1}{\sqrt{3}}\operatorname{tr}\operatorname{diag}(\mathrm{i},0,2\mathrm{i}) = \mathrm{i}\sqrt{3} && \Rightarrow \quad d_{458} = 0\,,\ f_{458} = \tfrac{\sqrt{3}}{2}\,.
\end{aligned}$$

The last line reproduces the specimen computed in Proposition 14.52.

Pair $\mathrm{C} = \{6,7\}$*, on* $\operatorname{diag}(0,\cdot,\cdot)$.

$$\begin{aligned}
&\operatorname{tr}(\lambda_6\lambda_6\lambda_3) = \operatorname{tr}\operatorname{diag}(0,-1,0) = -1 && \Rightarrow \quad d_{366} = d_{377} = -\tfrac{1}{2}\,,\\
&\operatorname{tr}(\lambda_6\lambda_6\lambda_8) = \tfrac{1}{\sqrt{3}}\operatorname{tr}\operatorname{diag}(0,1,-2) = -\tfrac{1}{\sqrt{3}} && \Rightarrow \quad d_{668} = d_{778} = -\tfrac{1}{2\sqrt{3}}\,,\\
&\operatorname{tr}(\lambda_6\lambda_7\lambda_3) = \operatorname{tr}\operatorname{diag}(0,-\mathrm{i},0) = -\mathrm{i} && \Rightarrow \quad d_{367} = 0\,,\ f_{367} = -\tfrac{1}{2}\,,\\
&\operatorname{tr}(\lambda_6\lambda_7\lambda_8) = \tfrac{1}{\sqrt{3}}\operatorname{tr}\operatorname{diag}(0,\mathrm{i},2\mathrm{i}) = \mathrm{i}\sqrt{3} && \Rightarrow \quad d_{678} = 0\,,\ f_{678} = \tfrac{\sqrt{3}}{2}\,.
\end{aligned}$$

Every entry follows from Equation (A.629) by halving the real part for d and the imaginary part for f; the index order has been rearranged into the one used in Equations (14.69) and (14.70), which is legitimate by Remark A.372 — note that $f_{367} = f_{673}$ and $f_{345} = f_{453}$, both cyclic and therefore sign-preserving.

A.34.5 Family (3): one index from each pair

This family is eight triples (a,b,c) with $a\in\{1,2\}$, $b\in\{4,5\}$, $c\in\{6,7\}$, and all eight fall out of a single observation: in Equation (A.632) only *one* closed walk survives. The first factor lies on the edge 1–2, the second on 1–3, the third on 2–3; a walk $i\to j\to k\to i$ must therefore start with a step of the first edge, and the only choice compatible with the second step lying on 1–3 is $j=1$, hence $i=2$, $k=3$. So

$$\operatorname{tr}(\lambda_a\lambda_b\lambda_c)=(\lambda_a)_{21}\,(\lambda_b)_{13}\,(\lambda_c)_{32}\,, \tag{A.635}$$

and Equation (A.631) supplies the three factors:

$$(\lambda_1)_{21}=1\,,\quad (\lambda_2)_{21}=\mathrm{i}\,,\qquad (\lambda_4)_{13}=1\,,\quad (\lambda_5)_{13}=-\mathrm{i}\,,\qquad (\lambda_6)_{32}=1\,,\quad (\lambda_7)_{32}=\mathrm{i}\,. \tag{A.636}$$

Multiplying out the eight products of Equation (A.636) and halving real and imaginary parts as in Equation (A.629):

$$\begin{aligned}
&\operatorname{tr}(\lambda_1\lambda_4\lambda_6)=1 &&\Rightarrow\ d_{146}=\tfrac12\,,\ f_{146}=0\,, &&\operatorname{tr}(\lambda_2\lambda_4\lambda_6)=\mathrm{i} &&\Rightarrow\ f_{246}=\tfrac12\,,\\
&\operatorname{tr}(\lambda_1\lambda_4\lambda_7)=\mathrm{i} &&\Rightarrow\ f_{147}=\tfrac12\,, &&\operatorname{tr}(\lambda_2\lambda_4\lambda_7)=-1 &&\Rightarrow\ d_{247}=-\tfrac12\,,\\
&\operatorname{tr}(\lambda_1\lambda_5\lambda_6)=-\mathrm{i} &&\Rightarrow\ f_{156}=-\tfrac12\,, &&\operatorname{tr}(\lambda_2\lambda_5\lambda_6)=1 &&\Rightarrow\ d_{256}=\tfrac12\,,\\
&\operatorname{tr}(\lambda_1\lambda_5\lambda_7)=1 &&\Rightarrow\ d_{157}=\tfrac12\,, &&\operatorname{tr}(\lambda_2\lambda_5\lambda_7)=\mathrm{i} &&\Rightarrow\ f_{257}=\tfrac12\,,
\end{aligned} \tag{A.637}$$

the unlisted partner of each entry vanishing. The value $d_{247}=-\frac12$ reproduces the second specimen of Proposition 14.52.

A.34.6 The tables

Collecting Equation (A.634), the three blocks of Appendix A.34.4 and Equation (A.637) gives Tables A.2 and A.3. By Lemma A.373 no triple outside the three families can contribute, and inside them the enumeration was exhaustive; the tables are therefore complete, and reproduce Equations (14.69) and (14.70).

Table A.2: The nine independent nonvanishing components of the antisymmetric structure constants f_{abc} of $\mathfrak{su}(3)$ in the Gell-Mann basis of Definition 14.51, with the family of Lemma A.373 each belongs to and the trace from which it was read. Total antisymmetry generates six nonzero components from each entry, fifty-four in all; every other component vanishes.

abc	f_{abc}	family	generating trace
123	1	(2), pair $\{1,2\}$	$\operatorname{tr}(\lambda_1\lambda_2\lambda_3)=2\mathrm{i}$
147	$\frac12$	(3)	$\operatorname{tr}(\lambda_1\lambda_4\lambda_7)=\mathrm{i}$
156	$-\frac12$	(3)	$\operatorname{tr}(\lambda_1\lambda_5\lambda_6)=-\mathrm{i}$
246	$\frac12$	(3)	$\operatorname{tr}(\lambda_2\lambda_4\lambda_6)=\mathrm{i}$
257	$\frac12$	(3)	$\operatorname{tr}(\lambda_2\lambda_5\lambda_7)=\mathrm{i}$
345	$\frac12$	(2), pair $\{4,5\}$	$\operatorname{tr}(\lambda_4\lambda_5\lambda_3)=\mathrm{i}$
367	$-\frac12$	(2), pair $\{6,7\}$	$\operatorname{tr}(\lambda_6\lambda_7\lambda_3)=-\mathrm{i}$
458	$\frac{\sqrt3}{2}$	(2), pair $\{4,5\}$	$\operatorname{tr}(\lambda_4\lambda_5\lambda_8)=\mathrm{i}\sqrt3$
678	$\frac{\sqrt3}{2}$	(2), pair $\{6,7\}$	$\operatorname{tr}(\lambda_6\lambda_7\lambda_8)=\mathrm{i}\sqrt3$

Two entries of Equation (14.70) deserve a word, because the chapter lists them in an order that hides where they come from. $d_{338}=\frac{1}{\sqrt3}$ is a family-(1) entry, computed from three diagonal matrices, while d_{118} and d_{228} are family-(2) entries computed from $\lambda_1^2=\lambda_2^2=\operatorname{diag}(1,1,0)$; the three come out equal because λ_3^2 equals that same matrix. That coincidence is the reason the chapter can write $d_{118}=d_{228}=d_{338}$ in one line.

Table A.3: The sixteen independent nonvanishing components of the symmetric tensor d_{abc} of $\mathfrak{su}(3)$, grouped by the family of Lemma A.373. Total symmetry generates six equal components from each of the four entries with distinct indices, three from each of the eleven with one repetition, and one from d_{888}: fifty-eight nonvanishing components in all.

abc	d_{abc}	abc	d_{abc}
Family (1): indices in $\{3,8\}$			
338	$\frac{1}{\sqrt{3}}$	888	$-\frac{1}{\sqrt{3}}$
Family (2): one index in $\{3,8\}$*, two in one pair*			
118	$\frac{1}{\sqrt{3}}$	448	$-\frac{1}{2\sqrt{3}}$
228	$\frac{1}{\sqrt{3}}$	558	$-\frac{1}{2\sqrt{3}}$
344	$\frac{1}{2}$	668	$-\frac{1}{2\sqrt{3}}$
355	$\frac{1}{2}$	778	$-\frac{1}{2\sqrt{3}}$
366	$-\frac{1}{2}$	377	$-\frac{1}{2}$
Family (3): one index from each pair			
146	$\frac{1}{2}$	247	$-\frac{1}{2}$
157	$\frac{1}{2}$	256	$\frac{1}{2}$

A.34.7 A Jacobi check

The tables are a computation, and a computation deserves an independent test. The Jacobi identity for the brackets Equation (14.67) is such a test, and it is nontrivial on a triple that involves several different entries at once. Take (T_1, T_4, T_5):

$$[T_1,[T_4,T_5]] + [T_4,[T_5,T_1]] + [T_5,[T_1,T_4]] = 0\,. \tag{A.638}$$

The first inner bracket has *two* nonzero components, which is the trap: $[T_4,T_5] = \mathrm{i} f_{45c} T_c = \mathrm{i}\left(\frac{1}{2}T_3 + \frac{\sqrt{3}}{2}T_8\right)$, by $f_{453} = f_{345} = \frac{1}{2}$ and $f_{458} = \frac{\sqrt{3}}{2}$ from Table A.2. Since $f_{18c} = 0$ for every c — no entry of Table A.2 contains both 1 and 8 — the T_8 piece drops, and with $f_{132} = -f_{123} = -1$,

$$[T_1,[T_4,T_5]] = \frac{\mathrm{i}}{2}\,[T_1,T_3] = \frac{\mathrm{i}}{2}\,(\mathrm{i} f_{132} T_2) = \frac{\mathrm{i}}{2}\,(-\mathrm{i} T_2) = \frac{1}{2}\,T_2\,.$$

For the second term, $f_{516} = -f_{156} = \frac{1}{2}$ gives $[T_5,T_1] = \frac{\mathrm{i}}{2}T_6$, and $f_{462} = f_{246} = \frac{1}{2}$ gives

$$[T_4,[T_5,T_1]] = \frac{\mathrm{i}}{2}\,[T_4,T_6] = \frac{\mathrm{i}}{2}\left(\frac{\mathrm{i}}{2}T_2\right) = -\frac{1}{4}\,T_2\,.$$

For the third, $f_{147} = \frac{1}{2}$ gives $[T_1,T_4] = \frac{\mathrm{i}}{2}T_7$, and $f_{572} = f_{257} = \frac{1}{2}$ gives

$$[T_5,[T_1,T_4]] = \frac{\mathrm{i}}{2}\,[T_5,T_7] = \frac{\mathrm{i}}{2}\left(\frac{\mathrm{i}}{2}T_2\right) = -\frac{1}{4}\,T_2\,.$$

The three add to $\left(\frac{1}{2} - \frac{1}{4} - \frac{1}{4}\right)T_2 = 0$, so Equation (A.638) holds. The check consumes f_{123}, f_{147}, f_{156}, f_{246}, f_{257}, f_{345} and f_{458} — seven of the nine independent entries, drawn from both families that carry a nonzero f — and it fails if any one of them is altered. Had the f_{453} contribution to $[T_4,T_5]$ been overlooked, leaving only the f_{458} piece that the chapter's specimen computes, the identity would have returned $-\frac{1}{2}T_2$ instead of zero.

Remark A.374. Tables A.2 and A.3 discharge the tabulation asserted in Proposition 14.52 and used throughout Chapter 14: in Proposition 14.53, where the entry $\kappa_{33} = 3$ is checked against

the f_{3cd} read off Table A.2; in Proposition 14.54, whose cubic Casimir $C_3 = d_{abc}T_aT_bT_c$ is built from Table A.3 and whose invariance identity Equation (14.76) couples the two tables; and in Proposition 14.55, where the adjoint generators are $\left(T_a^{\mathrm{ad}}\right)_{bc} = -\mathrm{i}f_{abc}$. The completeness established by Lemma A.373 is what licenses the sums over c and d in those places to be evaluated by listing the entries of the tables and stopping.

A.35 Simplicity of $\mathfrak{su}(3)$

This appendix proves what Chapter 14 quotes in two places: that $\mathfrak{su}(3)$ is a simple Lie algebra. The statement is used in Proposition 14.53, where the nondegeneracy of the Killing form is announced as the semisimplicity demanded by Theorem 14.13, and — more substantially — in Theorem 14.57, whose proof identifies the eight-dimensional summand of $\mathbf{3} \otimes \bar{\mathbf{3}}$ with the adjoint representation and calls it irreducible *because an invariant subspace would be an ideal of a simple Lie algebra.* That last inference is what is discharged here.

The proof is the standard one and is short: complexify, split a putative ideal into eigenspaces of the two-dimensional Cartan subalgebra, and then walk around the root hexagon of Equation (14.80), using the fact that any two roots at 120 degrees add to a third. A single root vector inside the ideal therefore drags in all six, and with them the Cartan directions.

Two remarks on what is *not* used, because both matter for circularity. First, nothing below quotes the classification of simple Lie algebras; only Definition 14.51 and the multiplication table of the 3×3 matrix units enter. Second, nothing below uses Proposition 14.53. That is deliberate: the chapter computes the Killing form of $\mathfrak{su}(3)$ *directly*, precisely so as not to lean on the uniqueness of an invariant form on a simple algebra, and the present appendix repays the compliment by not leaning on the Killing form.

A.35.1 Statement

Definition A.375 (Simple Lie algebra)**.** A Lie algebra $\mathfrak{g}$ over a field is *simple* if it is not abelian and its only ideals are $\{0\}$ and $\mathfrak{g}$ itself; an *ideal* is a subspace $\mathfrak{i}$ with $[\mathfrak{g}, \mathfrak{i}] \subseteq \mathfrak{i}$.
Rests on Definition 14.49.

Theorem A.376 ($\mathfrak{su}(3)$ is simple)**.** *The real Lie algebra $\mathfrak{su}(3)$ of Definition 14.49 is simple, and so is its complexification $\mathfrak{sl}(3,\mathbb{C})$, the algebra of traceless complex 3×3 matrices.*
Rests on Definitions 14.49 and A.375 and Proposition 14.50.

Corollary A.377 (The adjoint representation is irreducible)**.** *The adjoint representation of $\mathfrak{su}(3)$ on itself, and equally the representation $M \mapsto UMU^\dagger$ of $\mathrm{SU}(3)$ on the traceless complex 3×3 matrices, has no invariant subspace other than $\{0\}$ and the whole space.*
Rests on Theorem A.376 and Proposition 14.55.

A.35.2 Complexification

Lemma A.378 (Real ideals and complex ideals)**.** *As a real vector space,*

$$\mathfrak{sl}(3,\mathbb{C}) = \mathfrak{su}(3) \oplus \mathrm{i}\,\mathfrak{su}(3)\,, \tag{A.639}$$

the sum being direct. If $\mathfrak{i} \subseteq \mathfrak{su}(3)$ is an ideal of the real algebra, then $\mathfrak{i}_{\mathbb{C}} = \mathfrak{i} + \mathrm{i}\,\mathfrak{i}$ is a complex ideal of $\mathfrak{sl}(3,\mathbb{C})$ with $\dim_{\mathbb{C}} \mathfrak{i}_{\mathbb{C}} = \dim_{\mathbb{R}} \mathfrak{i}$.
Rests on Definition 14.49 and Proposition 14.50.

Proof. *Derives Lemma A.378.*

The splitting. Let X be traceless. Put $X_- = \frac{1}{2}\left(X - X^\dagger\right)$ and $X_+ = \frac{1}{2}\left(X + X^\dagger\right)$, so that $X = X_- + X_+$. Both are traceless, since $\operatorname{tr} X^\dagger = \overline{\operatorname{tr} X} = 0$; X_- is anti-Hermitian,

hence in $\mathfrak{su}(3)$ by Definition 14.49; and X_+ is Hermitian, so $-\mathrm{i}X_+$ is anti-Hermitian and $X_+ = \mathrm{i}\,(-\mathrm{i}X_+) \in \mathrm{i}\,\mathfrak{su}(3)$. The sum is direct: if $Y \in \mathfrak{su}(3) \cap \mathrm{i}\,\mathfrak{su}(3)$, write $Y = \mathrm{i}Z$ with Z anti-Hermitian; then $Y^\dagger = -\mathrm{i}Z^\dagger = \mathrm{i}Z = Y$, so Y is Hermitian as well as anti-Hermitian, and $Y = 0$.

Ideals. The bracket of $\mathfrak{sl}(3,\mathbb{C})$ is $\mathbb{C}$-bilinear, so for $X = A + \mathrm{i}B$ with $A, B \in \mathfrak{su}(3)$ and $Y = P + \mathrm{i}Q$ with $P, Q \in \mathfrak{i}$,

$$[X,Y] = ([A,P] - [B,Q]) + \mathrm{i}\,([A,Q] + [B,P])\ ,$$

and all four brackets lie in $\mathfrak{i}$ because $\mathfrak{i}$ is an ideal of $\mathfrak{su}(3)$. So $[X,Y] \in \mathfrak{i}_{\mathbb{C}}$, and $\mathfrak{i}_{\mathbb{C}}$ is closed under multiplication by i hence a complex subspace. Finally $\mathfrak{i} \cap \mathrm{i}\,\mathfrak{i} = 0$, being contained in $\mathfrak{su}(3) \cap \mathrm{i}\,\mathfrak{su}(3) = 0$, so $\mathfrak{i}_{\mathbb{C}} = \mathfrak{i} \oplus \mathrm{i}\,\mathfrak{i}$ has real dimension $2\dim_{\mathbb{R}}\mathfrak{i}$ and complex dimension $\dim_{\mathbb{R}}\mathfrak{i}$. □

A.35.3 The root decomposition

Write E_{jk} for the matrix unit with a single 1 in row j and column k, so that $E_{jk}E_{lm} = \delta_{kl}E_{jm}$. Let

$$\mathfrak{h} = \{\mathrm{diag}(a_1,a_2,a_3) \mid a_1 + a_2 + a_3 = 0\}\ , \tag{A.640}$$

the traceless diagonal matrices, which is two-dimensional and is the complex span of the T_3 and T_8 of Definition 14.51; it is the Cartan subalgebra whose existence Proposition 14.50 established, of rank $\ell = 2$. For $H = \mathrm{diag}(a_1,a_2,a_3)$ a one-line computation with the matrix units gives

$$[H, E_{jk}] = (a_j - a_k)\,E_{jk}\,, \qquad j \neq k\,, \tag{A.641}$$

so that $\mathfrak{sl}(3,\mathbb{C})$ decomposes as

$$\mathfrak{sl}(3,\mathbb{C}) = \mathfrak{h} \oplus \bigoplus_{j\neq k} \mathbb{C}\,E_{jk}\,, \tag{A.642}$$

a direct sum of $2 + 6 = 8$ pieces, each an eigenspace of every ad_H, and each of the six off-diagonal pieces one-dimensional. The linear functional $\alpha_{jk} \in \mathfrak{h}^*$, $\alpha_{jk}(H) = a_j - a_k$, is the *root* carried by E_{jk}.

Remark A.379 (These are the six vectors of the hexagon). Evaluating α_{jk} on the two Cartan generators of Definition 14.51 gives the coordinates used in Remark 14.56. With $T_3 = \frac{1}{2}\,\mathrm{diag}(1,-1,0)$ and $T_8 = \frac{1}{2\sqrt{3}}\,\mathrm{diag}(1,1,-2)$,

$$E_{12} : (\alpha(T_3),\alpha(T_8)) = (1,0)\,,\quad E_{13} : \left(\tfrac{1}{2},\tfrac{\sqrt{3}}{2}\right)\,,\quad E_{23} : \left(-\tfrac{1}{2},\tfrac{\sqrt{3}}{2}\right)\,, \tag{A.643}$$

and E_{21}, E_{31}, E_{32} carry the negatives. These are exactly the six vectors of Equation (14.80): unit vectors at 60 degrees to one another, the regular hexagon. What the proof below uses is the one geometric property that hexagon has, namely that the sum of two roots at 120 degrees is again a root — $|\gamma| = |\delta| = 1$ and $\gamma\cdot\delta = -\frac{1}{2}$ give $|\gamma+\delta|^2 = 1$, and the sum bisects the angle. In matrix units that property reads $[E_{jk}, E_{kl}] = E_{jl}$ for j,k,l distinct, which is the form in which it will be applied.

Lemma A.380 (An ideal is a sum of eigenspaces)**.** *Let $\mathfrak{i}$ be an ideal of $\mathfrak{sl}(3,\mathbb{C})$ and let $x = h + \sum_{j\neq k} c_{jk}E_{jk} \in \mathfrak{i}$ be the decomposition of an element according to Equation (A.642). Then $h \in \mathfrak{i}$ and $c_{jk}E_{jk} \in \mathfrak{i}$ for every pair; in particular each E_{jk} with $c_{jk} \neq 0$ lies in $\mathfrak{i}$.*

Rests on Equation (A.642) and Definition A.375.

Proof. *Derives Lemma A.380.*

Take the particular Cartan element $H_0 = \mathrm{diag}(3,-1,-2) \in \mathfrak{h}$. Its six root values are

$$\alpha_{12}(H_0) = 4\,,\quad \alpha_{13}(H_0) = 5\,,\quad \alpha_{23}(H_0) = 1\,, \tag{A.644}$$

together with $-4, -5, -1$: six *distinct and nonzero* numbers, which is the only property of H_0 that is used. Enumerate the six ordered pairs as $\mu = 1, \dots, 6$, write E_μ for the corresponding matrix unit and $\lambda_\mu \in \{\pm 1, \pm 4, \pm 5\}$ for its root value, so that $x = h + \sum_\mu c_\mu E_\mu$. Because $\mathfrak{i}$ is an ideal and $[H_0, h] = 0$, all six elements

$$\operatorname{ad}_{H_0}^n x = \sum_\mu \lambda_\mu^n \, c_\mu E_\mu \in \mathfrak{i}\,, \qquad n = 1, \dots, 6\,,$$

lie in $\mathfrak{i}$, by Equation (A.641). The 6×6 coefficient matrix $\left(\lambda_\mu^n\right)$ has determinant $\left(\prod_\mu \lambda_\mu\right)$ times the Vandermonde determinant $\prod_{\mu<\nu}(\lambda_\nu - \lambda_\mu)$, which is nonzero because the λ_μ are nonzero and pairwise distinct. Inverting it expresses each $c_\mu E_\mu$ as a linear combination of the $\operatorname{ad}_{H_0}^n x$, so $c_\mu E_\mu \in \mathfrak{i}$; and then $h = x - \sum_\mu c_\mu E_\mu \in \mathfrak{i}$. Since $\mathbb{C}E_\mu$ is one-dimensional, $c_\mu \neq 0$ forces $E_\mu \in \mathfrak{i}$. □

A.35.4 Walking round the hexagon

Lemma A.381 (One root vector generates everything)**.** *Let $\mathfrak{i}$ be an ideal of $\mathfrak{sl}(3, \mathbb{C})$ containing E_{jk} for one pair $j \neq k$. Then $\mathfrak{i} = \mathfrak{sl}(3, \mathbb{C})$.*
Rests on Equations (A.641) and (A.642).

Proof. *Derives Lemma A.381.*

Let l be the remaining index, so that $\{j, k, l\} = \{1, 2, 3\}$. Every bracket below is a bracket of an element of $\mathfrak{sl}(3, \mathbb{C})$ with an element of $\mathfrak{i}$, and therefore lies in $\mathfrak{i}$.

The opposite root vector. From $E_{kj}E_{jk} = E_{kk}$ and $E_{jk}E_{kj} = E_{jj}$,

$$[E_{jk}, E_{kj}] = E_{jj} - E_{kk} =: H_{jk} \in \mathfrak{i}\,, \tag{A.645}$$

and H_{jk} is the diagonal matrix with entries $+1$ at j, -1 at k and 0 at l, so Equation (A.641) gives $[H_{jk}, E_{kj}] = (-1-1)E_{kj} = -2E_{kj}$. Hence $E_{kj} \in \mathfrak{i}$. In the language of Remark A.379: the root γ and the coroot direction it generates carry the ideal to $-\gamma$, the opposite vertex of the hexagon.

The two neighbouring roots. Because $k \neq l$ and $l \neq j$,

$$[E_{lj}, E_{jk}] = E_{lk} \in \mathfrak{i}\,, \qquad [E_{jk}, E_{kl}] = E_{jl} \in \mathfrak{i}\,, \tag{A.646}$$

the second term of each commutator vanishing — $E_{jk}E_{lj} = 0$ since $k \neq l$, and $E_{kl}E_{jk} = 0$ since $l \neq j$. These are the two roots at 60 degrees from the starting one, reached by adding the roots of E_{lj} and of E_{kl}, each at 120 degrees to it.

Closing. Apply the first step to E_{lk} and to E_{jl}: it returns $E_{kl}, H_{lk} \in \mathfrak{i}$ and $E_{lj}, H_{jl} \in \mathfrak{i}$. All six matrix units are now in $\mathfrak{i}$, which is the whole hexagon. The diagonal part follows from Equation (A.645): $H_{jk} = E_{jj} - E_{kk}$ and $H_{lk} = E_{ll} - E_{kk}$ are linearly independent elements of the two-dimensional $\mathfrak{h}$ of Equation (A.640), so they span it. By Equation (A.642), $\mathfrak{i} = \mathfrak{sl}(3, \mathbb{C})$. □

A.35.5 Proof of the theorem

Proof of Theorem A.376. *Derives Theorem A.376.*

The complex algebra. $\mathfrak{sl}(3, \mathbb{C})$ is not abelian, since $[E_{12}, E_{21}] = E_{11} - E_{22} \neq 0$. Let $\mathfrak{i} \neq \{0\}$ be an ideal and pick $x \in \mathfrak{i}$ nonzero. By Lemma A.380 either some E_{jk} lies in $\mathfrak{i}$ — in which case $\mathfrak{i} = \mathfrak{sl}(3, \mathbb{C})$ by Lemma A.381 — or every c_{jk} vanishes for every element of $\mathfrak{i}$, so that $\mathfrak{i} \subseteq \mathfrak{h}$ and $x = h \neq 0$ is a nonzero traceless diagonal matrix. In that second case Equation (A.641) gives $[h, E_{jk}] = \alpha_{jk}(h)\, E_{jk} \in \mathfrak{i}$; the right-hand side lies in $\mathbb{C}E_{jk}$, which meets $\mathfrak{h}$ only in 0, so $\alpha_{jk}(h) = 0$ for all six pairs. Writing $h = \operatorname{diag}(a_1, a_2, a_3)$ this reads $a_1 = a_2 = a_3$, which with $\sum a_i = 0$ gives $h = 0$ — contradicting the choice of x. (Equivalently: the roots

Equation (A.643) span $\mathfrak{h}^*$, since $(1,0)$ and $\left(-\frac{1}{2},\frac{\sqrt{3}}{2}\right)$ are independent, so a Cartan element annihilated by all of them is zero.) The second case is therefore empty, and $\mathfrak{sl}(3,\mathbb{C})$ is simple.

The real algebra. $\mathfrak{su}(3)$ is not abelian: the generators T_1, T_2, T_3 of Definition 14.51 satisfy $[T_1, T_2] = \mathrm{i}T_3 \neq 0$ by Equation (14.69), and $\mathrm{i}T_a$ is the anti-Hermitian element of $\mathfrak{su}(3)$ attached to the Hermitian generator T_a. Let $\mathfrak{i} \neq \{0\}$ be an ideal of $\mathfrak{su}(3)$. By Lemma A.378, $\mathfrak{i}_{\mathbb{C}}$ is a nonzero complex ideal of $\mathfrak{sl}(3,\mathbb{C})$, hence all of it by the previous paragraph, so

$$\dim_{\mathbb{R}} \mathfrak{i} = \dim_{\mathbb{C}} \mathfrak{i}_{\mathbb{C}} = \dim_{\mathbb{C}} \mathfrak{sl}(3,\mathbb{C}) = 8 = \dim_{\mathbb{R}} \mathfrak{su}(3)\,,$$

the last equality being Proposition 14.50. A subspace of full dimension is the whole space, so $\mathfrak{i} = \mathfrak{su}(3)$. □

Proof of Corollary A.377. *Derives Corollary A.377.*

An invariant subspace V of the adjoint representation of a Lie algebra $\mathfrak{g}$ on itself satisfies $[\mathfrak{g}, V] \subseteq V$ by definition, which is exactly the statement that V is an ideal (Definition A.375); so Theorem A.376 leaves only $\{0\}$ and $\mathfrak{su}(3)$.

For the group statement, note first that the space of traceless complex 3×3 matrices is $\mathfrak{sl}(3,\mathbb{C})$ and that $M \mapsto UMU^\dagger = UMU^{-1}$ is the adjoint action of SU(3) on it, as Proposition 14.55 and the proof of Theorem 14.57 record. A subspace V invariant under the group is invariant under the algebra: for $X \in \mathfrak{su}(3)$ and $M \in V$, the curve $t \mapsto \mathrm{e}^{tX} M \mathrm{e}^{-tX}$ lies in V, which is a closed subspace of a finite-dimensional space, and its derivative at $t = 0$ is $[X, M]$, so $[X, M] \in V$. Complex-linearity extends this from $\mathfrak{su}(3)$ to $\mathfrak{sl}(3,\mathbb{C})$ by Equation (A.639), and the first paragraph applies. □

Remark A.382. Theorem A.376 discharges the two quotations named at the head of this section. In Proposition 14.53 it justifies the phrase "which is simple and hence semisimple": the Killing form $\kappa_{ab} = 3\delta_{ab}$ computed there is nondegenerate, as Theorem 14.13 requires of a semisimple algebra, and the implication now runs in the direction the chapter states rather than resting on an unproved assertion. In Theorem 14.57 it supplies the irreducibility of the eight-dimensional summand of $\mathbf{3} \otimes \bar{\mathbf{3}}$, so that the decomposition $3 \times 3 = 1 + 8$ of Equation (14.81) is proved and not merely counted; Corollary A.377 is the form in which it is used there.

One limitation is worth stating plainly. What is proved is the simplicity of this one algebra, by an argument that inspects its root system directly. The general theorem — that a complex simple Lie algebra is determined by its root system, and that the possible root systems are the classified Dynkin diagrams — is not proved here and is not needed: the whole of Chapter 14 uses $\mathfrak{su}(2)$ and $\mathfrak{su}(3)$ and nothing else, and both are handled by hand.

A.36 From the Little Algebra to the Labels of a Massive Representation

This appendix supplies the step left owed after Lemma 14.68 of Chapter 14: the passage from the *algebra* that fixes a timelike momentum to the *count of labels* of a massive irreducible representation of the inhomogeneous algebra, and with it the arithmetic of Equation (14.107),

$$1 + \left\lfloor \frac{D-1}{2} \right\rfloor = \left\lceil \frac{D}{2} \right\rceil . \tag{A.647}$$

What Lemma 14.68 establishes is that the subalgebra annihilating a timelike p is $\mathfrak{so}(D-1)$, of rank $\lfloor (D-1)/2 \rfloor$. What is owed is the representation theory that converts a rank into a count: that a massive irreducible representation is determined by the mass *together with* an irreducible representation of that subalgebra, and that no further invariant of the enveloping algebra escapes those labels.

The construction is Wigner's method of induced representations [849]. Chapter 94 carries it out in the observed four dimensions — Theorem 94.27 is the construction and Theorem 94.29 the result — and this appendix does the same in general D, with two differences of emphasis: the geometric ingredients (transitivity of the Lorentz group on the mass shell, the standard boost, the invariant measure) are proved here rather than quoted, and the arithmetic of the label count is separated cleanly from the classification theorem it rests on. That classification theorem is the one genuine import, and Remark A.393 says exactly what it is.

Under rule 7 of this treatise the general-D statement is the natural mathematical one and is stated as such; every worked instantiation below is $3+1$.

A.36.1 Conventions, and what is assumed

Indices $A, B, \dots$ run from 1 to D and $\eta_{AB} = \mathrm{diag}(+1, \dots, +1, -1)$ is the flat metric of signature $(D-1, 1)$ in the ordering of Notation 14.1, so that the time direction is the last one and $x^D = ct$, as in Remark 14.88. A vector p is *timelike* when $\eta_{AB} p^A p^B < 0$, which is the convention of Lemma 14.68.

The symmetry group is the connected inhomogeneous group

$$G = \mathbb{R}^D \rtimes \mathrm{SO}(D-1,1)^{\uparrow}, \qquad (\Lambda', a')(\Lambda, a) = (\Lambda'\Lambda,\ a' + \Lambda' a), \tag{A.648}$$

whose Lie algebra is $\mathfrak{iso}(D-1,1)$ of Equation (14.99), with brackets Equations (14.84), (14.90) and (14.97); the translation a carries the SI unit m. A quantum symmetry is a projective unitary representation, and by Theorem 94.3 — proved in Appendix A.9 — each symmetry is implemented by a unitary or antiunitary operator unique up to a phase, while Corollary 94.4 rules out the antiunitary branch for a connected group. Bargmann's analysis then replaces the projective representations of G by the ordinary unitary representations of its universal cover $\tilde{G}$ [1029], whose Lorentz factor is $\mathrm{Spin}(D-1,1)$. Everything below is written for G; for the half-integer labels every Λ is read as its lift to $\mathrm{Spin}(D-1,1)$ and every little-group element as an element of $\mathrm{Spin}(D-1)$, exactly as Theorem 94.27 does in four dimensions.

The generators are normalized as in Remark 14.66: with $P_A = \partial_A$ from Equation (14.82) the momentum operator is $\hat{P}_A = -\mathrm{i}\hbar P_A$, of SI unit kg m/s, and with J_{AB} from Equation (14.86) the angular-momentum operator is $\hat{J}_{AB} = -\mathrm{i}\hbar J_{AB}$, of unit J s.

Remark A.383 (A sign of signature). In the mostly-plus ordering fixed above, a timelike momentum satisfies $\eta_{AB} p^A p^B = -m^2c^2$ and therefore $\hat{P}^2 = -m^2c^2$ on a momentum eigenstate. Equation (14.105) records the same content with the opposite overall sign, being written in the mostly-minus convention customary in Chapter 94; the two differ by $\eta \mapsto -\eta$ and nothing else. Only $\left|\hat{P}^2\right| = m^2c^2$ is used below, and the ratio Equation (14.106) is unaffected by the choice, both invariants changing sign together.

A.36.2 Statement

Definition A.384 (The mass shell and the standard momentum)**.** For $m > 0$ put

$$\mathcal{M}_m = \left\{ p \in \mathbb{R}^D \mid \eta_{AB} p^A p^B = -m^2c^2,\ p^D > 0 \right\}, \qquad k^A = mc\,\delta^A_D, \tag{A.649}$$

so that $k \in \mathcal{M}_m$ is the rest momentum, of components $\boldsymbol{p} = 0$ and $p^D = mc$. The *little group* is the stabilizer $G_k = \left\{ \Lambda \in \mathrm{SO}(D-1,1)^{\uparrow} \mid \Lambda k = k \right\}$.

Rests on Notation 14.1 and Lemma 14.68.

Theorem A.385 (Labels of a massive representation)**.** *Let $D \geq 4$ and $m > 0$. The unitary equivalence classes of irreducible, strongly continuous, positive-energy unitary representations*

U of $\tilde{G}$ with $\hat{P}^2 = -m^2c^2\,\mathbb{1}$ are in bijection with the unitary equivalence classes of irreducible unitary representations σ of $\mathrm{Spin}(D-1)$*, the bijection being $U \cong U^{(m,\sigma)}$ with $U^{(m,\sigma)}$ the induced representation Equation (A.654) below. Consequently the class of U is fixed by the mass m together with the $\lfloor (D-1)/2 \rfloor$ invariant labels of σ furnished by Theorem 14.21, and by no fewer numbers; a massive irreducible representation therefore carries exactly $\lceil D/2 \rceil$ labels, which is Equation (A.647). Every central element of the enveloping algebra acts on U as a scalar that is a function of those labels.*

Rests on Lemma 14.68, Theorem 14.21, and Definition A.384.

The bijection itself — the classification — is the imported ingredient; the geometry that makes it meaningful, the construction of $U^{(m,\sigma)}$, and the arithmetic of the count are proved here.

A.36.3 The geometry of the mass shell

Lemma A.386 (The little group)**.** *$G_k = \mathrm{SO}(D-1)$, acting on the first $D-1$ coordinates, and its Lie algebra is the $\mathfrak{so}(D-1)$ of Lemma 14.68. It is compact.*

Rests on Definition A.384, Equation (14.23), and Lemma 14.68.

Proof. *Derives Lemma A.386.*

Let $\Lambda k = k$, that is $\Lambda e_D = e_D$ where e_D is the unit vector in the time direction. Since Λ preserves η, it preserves the orthogonal complement $e_D^{\perp} = \mathbb{R}^{D-1}$, on which η restricts to the Euclidean metric; so the restriction lies in $\mathrm{O}(D-1)$. Its determinant equals $\det \Lambda = 1$ because Λ acts as the identity on the remaining direction, whence $\Lambda|_{\mathbb{R}^{D-1}} \in \mathrm{SO}(D-1)$; conversely every such rotation, extended by $e_D \mapsto e_D$, preserves η, has determinant 1 and is orthochronous. Compactness is that of a closed bounded subset of $\mathbb{R}^{(D-1)\times(D-1)}$: the rows of an orthogonal matrix are unit vectors, and orthogonality is a closed condition. Differentiating a curve through the identity gives exactly the subalgebra of Lemma 14.68, namely the elements of $\mathfrak{so}(D-1,1)$ annihilating k. □

Lemma A.387 (Standard boost and transitivity)**.** *For $p \in \mathcal{M}_m$ write $\boldsymbol{p} = (p^1, \dots, p^{D-1})$ and $|\boldsymbol{p}|$ for its Euclidean length, so that $p^D = \sqrt{|\boldsymbol{p}|^2 + m^2c^2}$. The matrix*

$$L(p)^i{}_j = \delta^i{}_j + \frac{p^i p^j}{|\boldsymbol{p}|^2}\left(\frac{p^D}{mc} - 1\right), \qquad L(p)^i{}_D = L(p)^D{}_i = \frac{p^i}{mc}, \qquad L(p)^D{}_D = \frac{p^D}{mc}, \tag{A.650}$$

with $i, j = 1, \dots, D-1$, is a proper orthochronous Lorentz transformation, depends continuously on p with $L(k) = \mathbb{1}$, and satisfies $L(p)k = p$. In particular $\mathrm{SO}(D-1,1)^{\uparrow}$ acts transitively on $\mathcal{M}_m$, and $\mathcal{M}_m$ is a single orbit.

Rests on Definition A.384 and Equation (14.23).

Proof. *Derives Lemma A.387.*

Put $\hat{n}^i = p^i/|\boldsymbol{p}|$, $\cosh\zeta = p^D/(mc)$ and $\sinh\zeta = |\boldsymbol{p}|/(mc)$; these are consistent because $\cosh^2\zeta - \sinh^2\zeta = \left((p^D)^2 - |\boldsymbol{p}|^2\right)/(mc)^2 = 1$ by Equation (A.649), and $\zeta \geq 0$ is determined. In these terms Equation (A.650) reads

$$L(p) = \mathbb{1} + (\cosh\zeta - 1)\left(\hat{n}\hat{n}^{\mathsf{T}} + e_D\, e_D^{\mathsf{T}}\right) + \sinh\zeta\left(\hat{n}\, e_D^{\mathsf{T}} + e_D\, \hat{n}^{\mathsf{T}}\right), \tag{A.651}$$

$\hat{n}$ being regarded as a vector in $\mathbb{R}^D$ with vanishing last component. Reading off the entries confirms that this is Equation (A.650): the $\hat{n}\hat{n}^{\mathsf{T}}$ term gives $L^i{}_j$, the two mixed terms give $L^i{}_D = L^D{}_i = \sinh\zeta\,\hat{n}^i = p^i/(mc)$, and the $e_D e_D^{\mathsf{T}}$ term raises $L^D{}_D$ from 1 to $\cosh\zeta = p^D/(mc)$. It is the identity on the $(D-2)$-dimensional subspace orthogonal to both $\hat{n}$ and e_D, and on the plane they span it is $\begin{pmatrix} \cosh\zeta & \sinh\zeta \\ \sinh\zeta & \cosh\zeta \end{pmatrix}$ in the basis $(\hat{n}, e_D)$. Restricted

to that plane η is $\mathrm{diag}(+1,-1)$, and $\cosh^2\zeta-\sinh^2\zeta=1$ is exactly the statement that the matrix preserves it; on the complement $L(p)$ is the identity and the two blocks are η-orthogonal. So Equation (14.23) holds at the group level and $L(p)\in \mathrm{O}(D-1,1)$. Its determinant is $\cosh^2\zeta-\sinh^2\zeta=1$ and its (D,D) entry is $\cosh\zeta\geq 1>0$, so it is proper and orthochronous.

It maps k to p. From Equation (A.650), $(L(p)k)^i=L(p)^i{}_D\,mc=p^i$ and $(L(p)k)^D=L(p)^D{}_D\,mc=p^D$.

Continuity at $\boldsymbol{p}=0$. The only term in Equation (A.650) whose form is singular there is $p^ip^j\left(p^D/(mc)-1\right)/|\boldsymbol{p}|^2$, and

$$\frac{p^D}{mc}-1=\frac{\sqrt{|\boldsymbol{p}|^2+m^2c^2}-mc}{mc}=\frac{|\boldsymbol{p}|^2}{mc\left(\sqrt{|\boldsymbol{p}|^2+m^2c^2}+mc\right)},$$

so the quotient by $|\boldsymbol{p}|^2$ is bounded and tends to 0; the whole term vanishes with $\boldsymbol{p}$, leaving $L(k)=\mathbb{1}$. The remaining entries are manifestly continuous.

Transitivity. For $p,q\in\mathcal{M}_m$ the element $L(q)L(p)^{-1}$ is in $\mathrm{SO}(D-1,1)^\uparrow$ and carries p to q. Conversely a Lorentz transformation preserves $\eta_{AB}p^Ap^B$, and an orthochronous one preserves the sign of p^D on timelike vectors, so it maps $\mathcal{M}_m$ into itself. □

Lemma A.388 (The invariant measure)**.** *The measure*

$$\mathrm{d}\mu(p)=\frac{\mathrm{d}^{D-1}\boldsymbol{p}}{p^D},\qquad p^D=\sqrt{|\boldsymbol{p}|^2+m^2c^2},\tag{A.652}$$

on $\mathcal{M}_m$ is invariant under $\mathrm{SO}(D-1,1)^\uparrow$.

Rests on Definition A.384 and Lemma A.387.

Proof. *Derives Lemma A.388.*

Consider on $\mathbb{R}^D$ the measure

$$\mathrm{d}\nu(p)=\mathrm{d}^Dp\;\delta\left(\eta_{AB}p^Ap^B+m^2c^2\right)\theta\left(p^D\right),$$

with θ the unit step. Every factor is invariant: d^Dp because $|\det\Lambda|=1$, the delta because its argument is a Lorentz scalar, and the step because Λ is orthochronous. Now carry out the p^D integration. Writing $f(p^D)=|\boldsymbol{p}|^2-(p^D)^2+m^2c^2$, the only zero with $p^D>0$ is $p^D=\sqrt{|\boldsymbol{p}|^2+m^2c^2}$, at which $|f'|=2p^D$; so $\delta(f)\,\theta(p^D)\,\mathrm{d}p^D$ integrates to $1/\left(2p^D\right)$ and $\mathrm{d}\nu=\mathrm{d}^{D-1}\boldsymbol{p}/(2p^D)=\frac{1}{2}\mathrm{d}\mu$. A constant multiple of an invariant measure is invariant. □

Remark A.389 (The SI dimension of the measure). Each p^i carries the unit $\mathrm{kg\,m/s}$, so $\mathrm{d}\mu$ of Equation (A.652) carries $(\mathrm{kg\,m/s})^{D-2}$; in the observed $D=4$ that is $\mathrm{kg^2\,m^2/s^2}$. A wave function on the mass shell therefore carries the reciprocal of the square root of that unit, so that $\int\mathrm{d}\mu(p)\,|\psi(p)|^2$ is a pure number, as a probability must be. One may of course divide $\mathrm{d}\mu$ by $(mc)^{D-2}$ to obtain a dimensionless measure; nothing below depends on the choice, and the unnormalized form is kept because it is the one in which Lemma A.388 is transparent.

A.36.4 The induced representation

Proposition A.390 (Construction)**.** *Let σ be a unitary representation of $G_k=\mathrm{SO}(D-1)$ — or of $\mathrm{Spin}(D-1)$, for the covering group — on a finite-dimensional Hilbert space V_σ. On*

$$\mathcal{H}_\sigma=L^2(\mathcal{M}_m,\mathrm{d}\mu;V_\sigma)\tag{A.653}$$

define, for $a \in \mathbb{R}^D$ *and* $\Lambda \in \mathrm{SO}(D-1,1)^{\uparrow}$,

$$\left(U^{(m,\sigma)}(\mathbb{1},a)\psi\right)(p) = \mathrm{e}^{-\mathrm{i}\,\eta_{AB}p^A a^B/\hbar}\,\psi(p)\,, \qquad \left(U^{(m,\sigma)}(\Lambda,0)\psi\right)(p) = \sigma\Big(W(\Lambda,\Lambda^{-1}p)\Big)\,\psi\Big(\Lambda^{-1}p\Big)\,, \tag{A.654}$$

with the Wigner rotation

$$W(\Lambda,q) = L(\Lambda q)^{-1}\,\Lambda\,L(q)\,, \qquad q \in \mathcal{M}_m\,, \tag{A.655}$$

and $U^{(m,\sigma)}(\Lambda,a) = U^{(m,\sigma)}(\mathbb{1},a)\,U^{(m,\sigma)}(\Lambda,0)$. *Then* $W(\Lambda,q) \in G_k$, *and* $U^{(m,\sigma)}$ *is a strongly continuous unitary representation of* G *with*

$$\hat{P}_A\psi(p) = p_A\,\psi(p)\,, \qquad \hat{P}^2 = -m^2c^2\,\mathbb{1}\,, \tag{A.656}$$

and positive energy, $p^D > 0$ *throughout the spectrum.*

Rests on Lemmas A.386 to A.388.

Proof. *Derives Proposition A.390.*

The Wigner rotation lies in the little group. Using $L(q)k = q$ from Lemma A.387,

$$W(\Lambda,q)\,k = L(\Lambda q)^{-1}\Lambda L(q)k = L(\Lambda q)^{-1}\,(\Lambda q) = k\,,$$

so $W(\Lambda,q) \in G_k$ by Definition A.384, and it depends continuously on (Λ,q) because L does.

The composition law. Put $q = \Lambda^{-1}\Lambda'^{-1}p$, so that $\Lambda q = \Lambda'^{-1}p$ and $\Lambda'\Lambda q = p$. Then

$$\begin{aligned} W(\Lambda',\Lambda q)\,W(\Lambda,q) &= \Big[L(\Lambda'\Lambda q)^{-1}\Lambda' L(\Lambda q)\Big]\,\Big[L(\Lambda q)^{-1}\Lambda L(q)\Big] \\ &= L(\Lambda'\Lambda q)^{-1}\,\Lambda'\Lambda L(q) = W(\Lambda'\Lambda,q)\ , \end{aligned} \tag{A.657}$$

the two middle factors cancelling. Applying Equation (A.654) twice and using Equation (A.657) together with the homomorphism property of σ,

$$\left(U(\Lambda',0)U(\Lambda,0)\psi\right)(p) = \sigma\big(W(\Lambda',\Lambda q)\big)\,\sigma\big(W(\Lambda,q)\big)\,\psi(q) = \sigma\big(W(\Lambda'\Lambda,q)\big)\,\psi(q)\,,$$

which is $\left(U(\Lambda'\Lambda,0)\psi\right)(p)$. The translations compose by inspection, $\mathrm{e}^{-\mathrm{i}p\cdot a/\hbar}\mathrm{e}^{-\mathrm{i}p\cdot a'/\hbar} = \mathrm{e}^{-\mathrm{i}p\cdot(a+a')/\hbar}$, and the mixed relation required by Equation (A.648) is

$$\left(U(\Lambda,0)U(\mathbb{1},a)\psi\right)(p) = \mathrm{e}^{-\mathrm{i}\,(\Lambda^{-1}p)\cdot a/\hbar}\sigma(W)\psi(\Lambda^{-1}p) = \mathrm{e}^{-\mathrm{i}\,p\cdot(\Lambda a)/\hbar}\,\left(U(\Lambda,0)\psi\right)(p)\,,$$

because η is Λ-invariant: $(\Lambda^{-1}p)\cdot a = p\cdot(\Lambda a)$. That is $U(\Lambda,0)U(\mathbb{1},a) = U(\mathbb{1},\Lambda a)U(\Lambda,0)$, the semidirect product law.

Unitarity. For the translations the factor is a phase. For the Lorentz part,

$$\|U(\Lambda,0)\psi\|^2 = \int_{\mathcal{M}_m}\mathrm{d}\mu(p)\,\Big\|\sigma\Big(W(\Lambda,\Lambda^{-1}p)\Big)\,\psi(\Lambda^{-1}p)\Big\|_{V_\sigma}^2 = \int_{\mathcal{M}_m}\mathrm{d}\mu(p)\,\Big\|\psi(\Lambda^{-1}p)\Big\|_{V_\sigma}^2\,,$$

since σ is unitary on V_σ, and the last integral equals $\|\psi\|^2$ by the invariance of $\mathrm{d}\mu$ (Lemma A.388). Surjectivity is $U(\Lambda,0)^{-1} = U(\Lambda^{-1},0)$. Strong continuity follows from the continuity of L, W and σ together with dominated convergence on the dense set of compactly supported continuous ψ.

The momentum spectrum. The one-parameter groups of translations are $U(\mathbb{1},a) = \exp\left(-\mathrm{i}\,\eta_{AB}\hat{P}^A a^B/\hbar\right)$, which is the normalisation of Remark 14.66; comparing with the first line of Equation (A.654), whose phase is $-\mathrm{i}\,\eta_{AB}p^A a^B/\hbar$ for every a, the generator $\hat{P}^A$ acts as multiplication by the number p^A, which is Equation (A.656); then $\hat{P}^2 = \eta^{AB}p_A p_B = -m^2c^2$ everywhere on $\mathcal{M}_m$ by Equation (A.649), and $p^D > 0$ there by construction. □

Proposition A.391 (Reducible little-group data give reducible representations)**.** *If V_σ has a proper nonzero σ-invariant subspace V', then $L^2(\mathcal{M}_m, \mathrm{d}\mu; V')$ is a proper nonzero $U^{(m,\sigma)}$-invariant subspace of $\mathcal{H}_\sigma$. Hence irreducibility of $U^{(m,\sigma)}$ requires irreducibility of σ.*
Rests on Proposition A.390.

Proof. *Derives Proposition A.391.*

Both operations in Equation (A.654) act on the value $\psi(q) \in V_\sigma$ either by a scalar phase or by $\sigma(W)$ with $W \in G_k$, and both preserve V'; the argument q is merely relabelled. So the subspace of functions with values in V' is invariant, and it is proper and nonzero because V' is. □

A.36.5 The classification, and the count

The converse of Proposition A.391, and the statement that the list Equation (A.654) is exhaustive, is the imported ingredient.

Theorem A.392 (Wigner–Mackey classification; quoted)**.** *Let $D \geq 4$, $m > 0$, and let U be a strongly continuous unitary representation of $\tilde{G}$ that is irreducible, has positive energy and satisfies $\hat{P}^2 = -m^2c^2\mathbb{1}$. Then U is unitarily equivalent to $U^{(m,\sigma)}$ of Equation (A.654) for some irreducible unitary representation σ of $\mathrm{Spin}(D-1)$; conversely $U^{(m,\sigma)}$ is irreducible whenever σ is; and $U^{(m,\sigma)}$ and $U^{(m,\sigma')}$ are equivalent if and only if σ and σ' are [849, 1912].*
Rests on Proposition A.390 and Definition A.384.

Remark A.393 (What is quoted here)*.* Theorem A.392 is the one statement of this appendix that is not proved in this treatise, and it is worth being exact about what it costs.

Its content is Mackey's imprimitivity theorem [1005], specialized by Wigner [849] to the inhomogeneous Lorentz group. Three analytic inputs go into it, none of which this treatise develops: the spectral theorem for a strongly continuous unitary representation of the translation group $\mathbb{R}^D$ (the Stone–Naimark–Ambrose–Godement theorem, which supplies the projection-valued measure whose support is the momentum spectrum — Stone's one-parameter case is [707] and the general statement is standard functional analysis [1481]); the identification of the commutant of that measure's multiplication algebra with the decomposable operators, which is what turns an operator commuting with all translations into a measurable field $p \mapsto A(p)$ of matrices; and the disintegration of the Hilbert space over the orbit, which converts the field into a representation of the stabilizer. Given those three, the classification is the argument whose geometric half is proved above: the spectrum is a Λ-invariant subset of the hyperboloid Equation (A.649), which is one orbit by Lemma A.387, so the spectrum is the whole mass shell; the explicit continuous section L of Equation (A.650) then trivializes the bundle of fibres, and the residual freedom is exactly a representation of G_k.

Two things follow from that accounting. First, the imported theorem is functional analysis, not group theory: the Lie-theoretic content — which subalgebra fixes a timelike momentum, what its rank is, and how many labels its representations carry — is proved, in Lemma 14.68 and in Theorem 14.21. Second, the existence half of the classification, which is what a physical application actually uses, is proved here without any import: Proposition A.390 constructs $U^{(m,\sigma)}$ and computes its invariants, and Proposition A.391 supplies one direction of the irreducibility criterion. What is quoted is the *exhaustiveness* of the list, together with the converse irreducibility statement.

One further standard fact is used silently and is named here: a unitary representation of a compact group decomposes into finite-dimensional irreducible ones (the Peter–Weyl theorem [1912]), which is why V_σ may be taken finite-dimensional in Proposition A.390. Lemma 94.28 records the same point for SU(2) in four dimensions.

Proof of Theorem A.385. *Derives Theorem A.385.*

The bijection between equivalence classes is Theorem A.392. It remains to count.

The little algebra and its rank. By Lemma A.386 the little group is $\mathrm{SO}(D-1)$, whose Lie algebra is the $\mathfrak{so}(D-1)$ of Lemma 14.68; and since $D \geq 4$, that algebra has $D-1 \geq 3$ and is semisimple, of rank $\lfloor (D-1)/2 \rfloor$ by Proposition 14.22. Passing to $\mathrm{Spin}(D-1)$ changes neither the algebra nor its rank.

How many labels σ carries. Theorem 14.21 applied to $\mathfrak{so}(D-1)$ says that the centre of its enveloping algebra is a polynomial algebra in $\lfloor (D-1)/2 \rfloor$ algebraically independent Casimir elements, and that a complete set of invariant labels for its finite-dimensional irreducible representations consists of that many numbers *and of no fewer*. The representation σ is finite-dimensional by the Peter–Weyl theorem (Remark A.393), so exactly $\lfloor (D-1)/2 \rfloor$ numbers fix its class.

Adding the mass. The number m is an invariant of U independent of those: it is read off $\hat{P}^2$, which by Equation (A.656) is $-m^2c^2$ and involves the translation generators alone, whereas the little-group Casimirs are built from $\hat{J}$. Distinct masses therefore give inequivalent representations for every σ, and distinct σ give inequivalent representations for every mass, by Theorem A.392. The complete label set is thus m together with the $\lfloor (D-1)/2 \rfloor$ labels of σ, and by Equation (14.108) of Lemma 14.68 their number is

$$1 + \left\lfloor \frac{D-1}{2} \right\rfloor = \left\lceil \frac{D}{2} \right\rceil ,$$

which is Equation (A.647) and Equation (14.107).

No further invariant survives. Let C be any central element of the enveloping algebra $U(\mathfrak{iso}(D-1,1))$. On the irreducible U it acts as a scalar by Corollary 14.18, and that scalar depends only on the unitary equivalence class of U — an intertwiner conjugates C to itself. By the bijection the class is the pair $(m, [\sigma])$, so the scalar is a function of m and of the $\lfloor (D-1)/2 \rfloor$ labels of σ: no invariant of the enveloping algebra can separate two representations that those numbers already identify. □

A.36.6 The observed case

Example A.394 ($D = 4$)**.** Take $D = 4$, the observed spacetime. Then $\lfloor (D-1)/2 \rfloor = \lfloor 3/2 \rfloor = 1$ and $\lceil D/2 \rceil = 2$: a massive particle carries exactly two labels.

The little group is $\mathrm{SO}(3)$ by Lemma A.386, its cover is $\mathrm{SU}(2)$, and its algebra has rank one (Proposition 14.22 with $D-1=3$), so Theorem 14.21 allows a single Casimir — the $\boldsymbol{J}^2$ of Definition 14.40. Its irreducible representations are labelled by s with $2s \in \mathbb{N} \cup \{0\}$ (Theorem 14.43), of dimension $2s+1$, and the Casimir takes the value

$$\hat{\mathcal{J}}^2 = \hbar^2\, s(s+1)\, \mathbb{1}\,, \qquad \text{of SI unit } \mathrm{J}^2\,\mathrm{s}^2\,, \tag{A.658}$$

s itself being dimensionless. The two labels of Theorem A.385 are therefore m, of unit kg, and s.

In four dimensions — and, of the cases treated in Chapter 14, only there — both labels are realized by genuine Casimir elements of the inhomogeneous algebra itself: $\hat{P}^2$ and the Pauli–Lubanski square $\hat{W}^2$ of Theorem 14.65, whose values are Equation (14.105). That $\hat{W}^2$ measures the rest-frame angular momentum, so that $\hat{W}^2 = -m^2c^2\hat{\mathcal{J}}^2$ on the fibre over k, is the computation carried out in Remark 14.66 and again in Theorem 94.29; combined with Equation (A.658) it gives the second line of Equation (14.105), and the ratio Equation (14.106) isolates the $\hbar^2 s(s+1)$ that survives. Theorem 94.27 is Proposition A.390 written out for this case, with $\mathrm{d}\mu$ of Equation (A.652) replaced by the equivalent normalization used there.

Rests on Theorems 14.43, 14.65 and A.385.

Remark A.395 (Why the higher invariants are not constructed here). Remark 14.67 states that in general D the invariants beyond $\hat{P}^2$ are built from antisymmetrized products of

$\hat{J}$ with $\hat{P}$, the Pauli–Lubanski construction Equation (14.103) being available only at $D = 4$ because it contracts a D-index Levi-Civita symbol with one J and one P. That construction is not carried out in this appendix, and it is not needed for the theorem: the count Equation (A.647) follows from the classification and from Theorem 14.21 applied to the little algebra, neither of which requires the invariants to be exhibited as elements of the enveloping algebra. What has been proved is that a massive irreducible representation is specified by $\lceil D/2 \rceil$ numbers and by no fewer, and that any central element of the enveloping algebra is a function of them. The explicit general-D construction of the higher invariants would say which polynomial in $\hat{J}$ and $\hat{P}$ realizes each label; only the $D = 4$ case is used anywhere in this book, and it is Theorem 14.65.

Remark A.396. Theorem A.385 discharges the obligation recorded after Lemma 14.68: the lemma supplies the little algebra and its rank, and the theorem converts that rank into the label count Equation (14.107) quoted in Remark 14.67 — two labels at $D = 4$ and three at both $D = 5$ and $D = 6$, the ceiling and not the floor. The $3 + 1$ instantiation, Example A.394, is the mass and the spin, and it is the only case the physical parts of this treatise use; its detailed development, including the massless representations that this appendix does not treat, is Chapter 94.

A.37 Whitehead's Lemmas and the Rigidity of Semisimple Algebras

This appendix proves Proposition 14.77 of Chapter 14: the first and second Chevalley–Eilenberg cohomology groups of a finite-dimensional semisimple Lie algebra, with coefficients in any finite-dimensional module, vanish. The second of these — Whitehead's second lemma — is what Section 14.4 needs, since by Proposition 14.76 the vanishing of $H^2(\mathfrak{g}, \mathbb{R})$ says exactly that every central extension of $\mathfrak{g}$ is trivial. The corollary for the Lorentz algebra follows from Corollary A.344.

Nothing is quoted. The engine is the Casimir operator of the representation: it is invertible on any module with no trivial summand, and averaging a cocycle against it produces the coboundary that was sought. Semisimplicity enters through Appendix A.31 at three points — the Killing form and the trace form are nondegenerate, ideals split off orthogonally, and $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$.

Section 14.4.1 defines $H^2(\mathfrak{g}, \mathbb{R})$ directly, by writing down the cocycle condition Equation (14.121) and the coboundaries Equation (14.122) for the one case it needs; the complex those live in is named there but not built. It is built here (Appendix A.37.1), because the proof needs coefficients in an arbitrary module and needs degree three to say what a 2-cocycle is. Appendix A.37.2 constructs the Casimir of a representation, Appendix A.37.3 carries out the averaging, Appendix A.37.4 proves the complete reducibility needed to reduce a general module to irreducible ones, Appendix A.37.5 treats the trivial coefficients that averaging cannot reach, and Appendix A.37.6 assembles the lemmas and draws the corollary.

Throughout, $\mathbb{K}$ is a field of characteristic zero with algebraic closure $\mathbb{F}$, $\mathfrak{g}$ is a finite-dimensional semisimple Lie algebra over $\mathbb{K}$ with basis $\{T_a\}$, and a *module* is a finite-dimensional vector space $\mathbb{V}$ with a homomorphism $\rho : \mathfrak{g} \to \mathfrak{gl}(\mathbb{V})$, that is, a linear map with $\rho([X, Y]) = [\rho(X), \rho(Y)]$. We write $\rho(X)v$ and $X \cdot v$ interchangeably, and $\rho_a = \rho(T_a)$.

A.37.1 The Chevalley–Eilenberg complex

Definition A.397 (Cochains and the differential)**.** Let $\mathbb{V}$ be a $\mathfrak{g}$-module. For $n \geq 1$ let $C^n(\mathfrak{g}, \mathbb{V})$ be the space of alternating n-linear maps $\mathfrak{g}^n \to \mathbb{V}$, and let $C^0(\mathfrak{g}, \mathbb{V}) = \mathbb{V}$. The

Chevalley–Eilenberg differential $\mathrm{d} : C^n \to C^{n+1}$ is

$$(\mathrm{d}f)(X_0,\dots,X_n) = \sum_{i=0}^{n} (-1)^i \rho(X_i)\, f\left(X_0,\dots,\widehat{X_i},\dots,X_n\right) + \sum_{0\le i<j\le n} (-1)^{i+j} f\left([X_i,X_j],X_0,\dots,\widehat{X_i},\dots,\widehat{X_j},\dots,X_n\right), \quad \text{(A.659)}$$

the hat marking an omitted argument. Explicitly, in the three degrees used below,

$$(\mathrm{d}v)(X) = \rho(X)v\,, \quad \text{(A.660)}$$

$$(\mathrm{d}f)(X,Y) = \rho(X)f(Y) - \rho(Y)f(X) - f([X,Y])\,, \quad \text{(A.661)}$$

$$(\mathrm{d}c)(X,Y,Z) = \rho(X)c(Y,Z) - \rho(Y)c(X,Z) + \rho(Z)c(X,Y) - c([X,Y],Z) + c([X,Z],Y) - c([Y,Z],X)\,. \quad \text{(A.662)}$$

A cochain with $\mathrm{d}f = 0$ is a *cocycle*, one of the form $\mathrm{d}h$ a *coboundary*, and $H^n(\mathfrak{g},\mathbb{V})$ is the quotient of the cocycles by the coboundaries in degree n.

Rests on Definition 14.75.

Lemma A.398 (The differential squares to zero in the degrees used)**.** $\mathrm{d}(\mathrm{d}v) = 0$ *for* $v \in C^0$ *and* $\mathrm{d}(\mathrm{d}f) = 0$ *for* $f \in C^1$*. Hence* $H^1(\mathfrak{g},\mathbb{V})$ *and* $H^2(\mathfrak{g},\mathbb{V})$ *are defined.*

Rests on Definition A.397.

Proof. *Derives Lemma A.398.*

In degree zero, Equations (A.660) and (A.661) give

$$(\mathrm{d}\mathrm{d}v)(X,Y) = \rho(X)\rho(Y)v - \rho(Y)\rho(X)v - \rho([X,Y])\,v = 0\,, \quad \text{(A.663)}$$

which is precisely the statement that ρ is a homomorphism.

In degree one, write $g = \mathrm{d}f$ and expand $(\mathrm{d}g)(X,Y,Z)$ by Equation (A.662), substituting Equation (A.661) in each of the six terms. Suppressing ρ and writing Xu for $\rho(X)u$, the six terms are

$$\begin{aligned}
Xg(Y,Z) &= XYf(Z) - XZf(Y) - Xf([Y,Z])\,,\\
-Yg(X,Z) &= -YXf(Z) + YZf(X) + Yf([X,Z])\,,\\
Zg(X,Y) &= ZXf(Y) - ZYf(X) - Zf([X,Y])\,,\\
-g([X,Y],Z) &= -[X,Y]f(Z) + Zf([X,Y]) + f([[X,Y],Z])\,,\\
g([X,Z],Y) &= [X,Z]f(Y) - Yf([X,Z]) - f([[X,Z],Y])\,,\\
-g([Y,Z],X) &= -[Y,Z]f(X) + Xf([Y,Z]) + f([[Y,Z],X])\,.
\end{aligned} \quad \text{(A.664)}$$

Add them and collect. The terms carrying $f(Z)$ are $XY - YX - [X,Y]$, which vanish because ρ is a homomorphism; likewise those carrying $f(Y)$, namely $-XZ + ZX + [X,Z]$, and those carrying $f(X)$, namely $YZ - ZY - [Y,Z]$. The terms $\pm Xf([Y,Z])$, $\pm Yf([X,Z])$ and $\pm Zf([X,Y])$ cancel in pairs. What remains is

$$f([[X,Y],Z] - [[X,Z],Y] + [[Y,Z],X]) = 0\,, \quad \text{(A.665)}$$

because $-[[X,Z],Y] = [[Z,X],Y]$ and the argument is then the Jacobi identity. □

Remark A.399 (Agreement with the chapter's definitions). Take $\mathbb{V} = \mathbb{R}$ with the trivial action, $\rho = 0$. Then Equation (A.662) reduces to

$$(\mathrm{d}c)(X,Y,Z) = -[c([X,Y],Z) + c([Y,Z],X) + c([Z,X],Y)]\,, \quad \text{(A.666)}$$

using the antisymmetry of c to write $c([X,Z],Y) = -c([Z,X],Y)$; so $\mathrm{d}c = 0$ is exactly the cocycle condition Equation (14.121) of Proposition 14.73. And Equation (A.661) reduces to $(\mathrm{d}b)(X,Y) = -b([X,Y])$, whose image is the space of coboundaries Equation (14.122) of Definition 14.74 — the overall sign changes no subspace. The group $H^2(\mathfrak{g},\mathbb{R})$ of Definition 14.75 is therefore the group H^2 of Definition A.397, and Proposition 14.76 may be applied to it without further comment.

A.37.2 The Casimir operator of a representation

Definition A.400 (Invariant form, dual basis, Casimir)**.** A symmetric bilinear form β on $\mathfrak{g}$ is *invariant* if

$$\beta\left([X,Y],Z\right)=\beta\left(X,[Y,Z]\right)\qquad\text{for all } X,Y,Z\in\mathfrak{g}\,, \tag{A.667}$$

which for the Killing form is Equation (14.16) rewritten with the antisymmetry of the bracket. If β is in addition nondegenerate, the *dual basis* $\{T^a\}$ is defined by $\beta\left(T_a,T^b\right)=\delta_a{}^b$, and the *Casimir operator of the representation* ρ relative to β is

$$\boxed{\Gamma=\rho\left(T^a\right)\rho\left(T_a\right)\in\mathrm{End}(\mathbb{V})}\,, \tag{A.668}$$

summed over a. Taking $\beta=\kappa$ and ρ the identity map of $\mathfrak{g}$ into $U(\mathfrak{g})$ recovers $C_2=\kappa^{ab}T_aT_b$ of Equation (14.20).

Rests on Definition 14.10 and Lemma 14.11.

Lemma A.401 (The invariance identity)**.** *Let β be invariant and nondegenerate with dual basis $\{T^a\}$. Then for every $Y\in\mathfrak{g}$*

$$\sum_a\left([T^a,Y]\otimes T_a+T^a\otimes[T_a,Y]\right)=0\qquad\textit{in }\mathfrak{g}\otimes\mathfrak{g}\,. \tag{A.669}$$

Consequently Γ commutes with $\rho(Y)$ for every $Y\in\mathfrak{g}$.

Rests on Definition A.400.

Proof. *Derives Lemma A.401.*

Expand $[T_a,Y]=\mu_a{}^bT_b$ and $[T^a,Y]=\nu^a{}_bT^b$, so that $\mu_a{}^b=\beta\left([T_a,Y],T^b\right)$ and $\nu^a{}_b=\beta\left([T^a,Y],T_b\right)$. By Equation (A.667) and the symmetry of β,

$$\mu_a{}^b=\beta\left(T_a,\left[Y,T^b\right]\right)=-\beta\left(T_a,\left[T^b,Y\right]\right)=-\nu^b{}_a\,. \tag{A.670}$$

The left-hand side of Equation (A.669) is therefore $\sum_{a,b}\left(\nu^a{}_b+\mu_b{}^a\right)T^b\otimes T_a$, after renaming the summation indices in the second term, and each coefficient vanishes by Equation (A.670).

For the last statement apply the bilinear map $(u,w)\mapsto\rho(u)\rho(w)$ to Equation (A.669):

$$[\Gamma,\rho(Y)]=\rho\left(T^a\right)\left[\rho\left(T_a\right),\rho(Y)\right]+\left[\rho\left(T^a\right),\rho(Y)\right]\rho\left(T_a\right)=\rho\left(T^a\right)\rho\left([T_a,Y]\right)+\rho\left([T^a,Y]\right)\rho\left(T_a\right)\,, \tag{A.671}$$

which is the image of Equation (A.669) and hence zero. □

Lemma A.402 (The trace form is nondegenerate for a faithful representation)**.** *Let $\mathfrak{g}$ be semisimple and ρ faithful. Then*

$$\beta_{\mathbb{V}}(X,Y)=\mathrm{tr}\left(\rho(X)\rho(Y)\right) \tag{A.672}$$

is a symmetric invariant nondegenerate bilinear form on $\mathfrak{g}$, and the corresponding Casimir operator satisfies $\mathrm{tr}_{\mathbb{V}}\,\Gamma=\dim\mathfrak{g}$.

Rests on Theorems A.330 and A.340.

Proof. *Derives Lemma A.402.*

Symmetry is cyclicity of the trace, and invariance is the computation Equation (A.605): $\mathrm{tr}\left(\rho\left([X,Y]\right)\rho(Z)\right)=\mathrm{tr}\left(\rho(X)\rho\left([Y,Z]\right)\right)$. Let $\mathfrak{s}$ be the radical of $\beta_{\mathbb{V}}$; it is an ideal by the argument of Equation (A.606) with $\beta_{\mathbb{V}}$ in place of κ. For $W\in[\mathfrak{s},\mathfrak{s}]$ and $Y\in\mathfrak{s}$, $\mathrm{tr}\left(\rho(W)\rho(Y)\right)=\beta_{\mathbb{V}}(W,Y)=0$, so Theorem A.340 makes $\rho(\mathfrak{s})$ solvable; as ρ is faithful, $\mathfrak{s}$ is a solvable ideal of a semisimple algebra, hence zero (Remark A.333). Finally

$$\mathrm{tr}_{\mathbb{V}}\,\Gamma=\sum_a\mathrm{tr}\left(\rho\left(T^a\right)\rho\left(T_a\right)\right)=\sum_a\beta_{\mathbb{V}}\left(T^a,T_a\right)=\sum_a\delta_a{}^a=\dim\mathfrak{g}\,. \tag{A.673}$$

□

A.37.3 Averaging a cocycle against the Casimir

Theorem A.403 (Vanishing when the Casimir is invertible)**.** *Let β be an invariant nondegenerate symmetric form on $\mathfrak{g}$ and let the Casimir Γ of Equation (A.668) be invertible on $\mathbb{V}$. Then*

$$H^1(\mathfrak{g},\mathbb{V}) = H^2(\mathfrak{g},\mathbb{V}) = 0\,. \tag{A.674}$$

Rests on Lemma A.401 and Definition A.397.

Proof. *Derives Theorem A.403.*

Degree one. Let f be a 1-cocycle, so that by Equation (A.661)

$$\rho(X)f(Y) - \rho(Y)f(X) - f\left([X,Y]\right) = 0 \qquad \text{for all } X, Y\,. \tag{A.675}$$

Put $X = T_a$, apply $\rho\left(T^a\right)$ and sum over a. The first term gives $\Gamma f(Y)$. In the second, commute $\rho\left(T^a\right)$ past $\rho(Y)$,

$$\rho\left(T^a\right)\rho(Y) = \rho(Y)\rho\left(T^a\right) + \rho\left([T^a,Y]\right)\,, \tag{A.676}$$

and set

$$v = \rho\left(T^a\right) f\left(T_a\right) \in \mathbb{V}\,. \tag{A.677}$$

The result is

$$\Gamma f(Y) - \rho(Y)v - \rho\left([T^a,Y]\right) f\left(T_a\right) - \rho\left(T^a\right) f\left([T_a,Y]\right) = 0\,. \tag{A.678}$$

The last two terms are the image of the invariance identity Equation (A.669) under the bilinear map $(u,w) \mapsto \rho(u)f(w)$, hence vanish. So $\Gamma f(Y) = \rho(Y)v$ for every Y. Since Γ commutes with every $\rho(Y)$ (Lemma A.401) so does Γ^{-1}, and therefore

$$f(Y) = \rho(Y)\,\Gamma^{-1}v = \left(\mathrm{d}\left(\Gamma^{-1}v\right)\right)(Y)\,, \tag{A.679}$$

by Equation (A.660): every 1-cocycle is a coboundary.

Degree two. Let c be a 2-cocycle. Put $X = T_a$ in $(\mathrm{d}c)\,(X,Y,Z) = 0$, apply $\rho\left(T^a\right)$, sum over a, and set

$$b(X) = \rho\left(T^a\right) c\left(T_a, X\right) \in \mathbb{V}\,, \qquad b \in C^1(\mathfrak{g},\mathbb{V})\,. \tag{A.680}$$

The six terms of Equation (A.662) contribute as follows. The first gives $\Gamma c(Y,Z)$ outright. In the second and the third, Equation (A.676) moves $\rho\left(T^a\right)$ past $\rho(Y)$ and $\rho(Z)$, producing $-\rho(Y)b(Z)$ and $+\rho(Z)b(Y)$ together with one commutator term each. The fourth and fifth are unchanged. The sixth is $b\left([Y,Z]\right)$, since $-c\left([Y,Z]\,,T_a\right) = +c\left(T_a,[Y,Z]\right)$. Collecting,

$$\begin{aligned} 0 &= \rho\left(T^a\right)(\mathrm{d}c)\left(T_a,Y,Z\right) \\ &= \Gamma c(Y,Z) - \rho(Y)b(Z) + \rho(Z)b(Y) + b\left([Y,Z]\right) + R(Y,Z)\,, \end{aligned} \tag{A.681}$$

where the leftover is

$$\begin{aligned} R(Y,Z) = &-\rho\left([T^a,Y]\right) c\left(T_a,Z\right) - \rho\left(T^a\right) c\left([T_a,Y]\,,Z\right) \\ &+ \rho\left([T^a,Z]\right) c\left(T_a,Y\right) + \rho\left(T^a\right) c\left([T_a,Z]\,,Y\right)\,. \end{aligned} \tag{A.682}$$

Both lines of Equation (A.682) vanish: applying the bilinear map $(u,w) \mapsto \rho(u)c(w,Z)$ to Equation (A.669) gives

$$\rho\left([T^a,Y]\right) c\left(T_a,Z\right) + \rho\left(T^a\right) c\left([T_a,Y]\,,Z\right) = 0\,, \tag{A.683}$$

and the same computation with Y and Z exchanged kills the second line. So Equation (A.681) leaves

$$\Gamma c(Y,Z) - \rho(Y)b(Z) + \rho(Z)b(Y) + b\left([Y,Z]\right) = 0\,, \tag{A.684}$$

that is, by Equation (A.661), $\Gamma c = \mathrm{d}b$. As in degree one, Γ^{-1} commutes with every $\rho(Y)$ and therefore with d, so $c = \mathrm{d}\left(\Gamma^{-1}b\right)$: every 2-cocycle is a coboundary. □

Remark A.404 (What the averaging does). The name is worth justifying. For a *compact group* the standard proof of the same statement integrates a cochain over the group with the Haar measure, and the invariant so produced is a coboundary. There is no group and no measure here, but Γ plays the same role: it is the one operator built from the algebra alone that commutes with everything (Lemma A.401), and Equations (A.678) and (A.684) say that contracting a cocycle with the Casimir tensor $T^a \otimes T_a$ produces the cochain whose coboundary it is. What has to be paid for is the invertibility of Γ, and the price is exactly the trivial summands of $\mathbb{V}$, on which Γ vanishes identically. Those are dealt with separately in Appendix A.37.5, and they are the reason the proof needs complete reducibility as well.

A.37.4 Complete reducibility

Lemma A.405 (Schur, algebra form)**.** *Let $\mathbb{V}$ be a finite-dimensional irreducible module over $\mathbb{F}$ algebraically closed, and let $A \in \mathrm{End}(\mathbb{V})$ commute with $\rho(X)$ for every $X \in \mathfrak{g}$. Then $A = \lambda \mathbb{1}$ for some $\lambda \in \mathbb{F}$.*

Rests on Theorem 5.153.

Proof. *Derives Lemma A.405.*

The argument is that of Theorem 5.153, with the family of operators $\{U(a)\}$ replaced by $\{\rho(X)\}$: the characteristic polynomial of A has a root λ because $\mathbb{F}$ is algebraically closed and $\dim \mathbb{V} \geq 1$, so $\ker(A - \lambda\mathbb{1}) \neq 0$; that kernel is a submodule, because $A - \lambda\mathbb{1}$ commutes with the action; and irreducibility forces it to be all of $\mathbb{V}$. □

Lemma A.406 (Splitting a submodule of codimension one)**.** *Let $\mathfrak{g}$ be semisimple over $\mathbb{F}$ and let $\mathbb{V}$ be a module with a submodule $\mathbb{W}$ of codimension one. Then $\mathbb{V} = \mathbb{W} \oplus \mathbb{X}$ for some one-dimensional submodule $\mathbb{X}$.*

Rests on Lemmas A.402 and A.405 and Corollary A.342.

Proof. *Derives Lemma A.406.*

First, $\rho(\mathfrak{g})\mathbb{V} \subseteq \mathbb{W}$: the quotient $\mathbb{V}/\mathbb{W}$ is a one-dimensional module, so the action on it is a homomorphism of $\mathfrak{g}$ into an abelian algebra, which annihilates $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$ by Corollary A.342.

Next, we may assume ρ faithful. Otherwise put $\mathfrak{k} = \ker \rho$, an ideal; by Corollary A.342, $\mathfrak{g}/\mathfrak{k} \cong \mathfrak{k}^\perp$ is again semisimple, and $\mathbb{V}$ is a module over it with the same submodules. If $\mathfrak{g}/\mathfrak{k} = 0$ the action is trivial, every subspace is a submodule and any complementary line serves.

Induct on $\dim \mathbb{W}$.

Case 1: $\mathbb{W}$ has a submodule $\mathbb{W}'$ with $0 \neq \mathbb{W}' \neq \mathbb{W}$. In $\mathbb{V}/\mathbb{W}'$ the submodule $\mathbb{W}/\mathbb{W}'$ has codimension one and smaller dimension, so by induction there is a one-dimensional submodule $\widetilde{\mathbb{X}}/\mathbb{W}'$ with $\mathbb{V}/\mathbb{W}' = \mathbb{W}/\mathbb{W}' \oplus \widetilde{\mathbb{X}}/\mathbb{W}'$. Inside $\widetilde{\mathbb{X}}$ the submodule $\mathbb{W}'$ has codimension one and $\dim \mathbb{W}' < \dim \mathbb{W}$, so by induction again $\widetilde{\mathbb{X}} = \mathbb{W}' \oplus \mathbb{X}$ with $\mathbb{X}$ a one-dimensional submodule. Then $\mathbb{X} \cap \mathbb{W} \subseteq \widetilde{\mathbb{X}} \cap \mathbb{W} = \mathbb{W}'$ and $\mathbb{X} \cap \mathbb{W}' = 0$, so $\mathbb{X} \cap \mathbb{W} = 0$ and dimensions give $\mathbb{V} = \mathbb{W} \oplus \mathbb{X}$.

Case 2: $\mathbb{W}$ is irreducible. If $\rho(\mathfrak{g})\mathbb{W} = 0$ then $\mathbb{W}$ is one-dimensional, $\dim \mathbb{V} = 2$, and every product of two elements of $\rho(\mathfrak{g})$ vanishes because $\rho(\mathfrak{g})\mathbb{V} \subseteq \mathbb{W}$; so $\rho(\mathfrak{g})$ is abelian and $\rho(\mathfrak{g}) = \rho([\mathfrak{g}, \mathfrak{g}]) = 0$, and any complementary line serves. So assume $\rho(\mathfrak{g})\mathbb{W} \neq 0$, and let Γ be the Casimir built from the trace form $\beta_{\mathbb{V}}$ of Lemma A.402, which is available because ρ is faithful. By Lemma A.401 Γ commutes with the action; $\Gamma\mathbb{V} \subseteq \rho(\mathfrak{g})\mathbb{V} \subseteq \mathbb{W}$, since Γ is a sum of products of two operators $\rho(\cdot)$; hence Γ acts as zero on $\mathbb{V}/\mathbb{W}$ and

$$\mathrm{tr}_{\mathbb{W}} \Gamma = \mathrm{tr}_{\mathbb{V}} \Gamma - \mathrm{tr}_{\mathbb{V}/\mathbb{W}} \Gamma = \dim \mathfrak{g}\,, \tag{A.685}$$

by Lemma A.402. On the irreducible $\mathbb{W}$, Lemma A.405 gives $\Gamma|_{\mathbb{W}} = \lambda\mathbb{1}$ with $\lambda \dim \mathbb{W} = \dim \mathfrak{g}$, so $\lambda \neq 0$ in characteristic zero and $\Gamma|_{\mathbb{W}}$ is invertible. Therefore $\Gamma\mathbb{V} = \mathbb{W}$, $\dim \ker \Gamma = \dim \mathbb{V} - \dim \mathbb{W} = 1$, and $\ker \Gamma \cap \mathbb{W} = 0$. As Γ commutes with the action, $\ker \Gamma$ is a submodule, and $\mathbb{V} = \mathbb{W} \oplus \ker \Gamma$. □

Theorem A.407 (Weyl's complete reducibility theorem)**.** *Let $\mathfrak{g}$ be semisimple over $\mathbb{F}$ algebraically closed of characteristic zero. Every finite-dimensional $\mathfrak{g}$-module is a direct sum of irreducible submodules.*

Rests on Lemma A.406 and Corollary A.342.

Proof. *Derives Theorem A.407.*

It suffices to show that every submodule $\mathbb{W} \subseteq \mathbb{V}$ has a complementary submodule; the decomposition then follows by induction on $\dim \mathbb{V}$. Assume $0 \neq \mathbb{W} \neq \mathbb{V}$ and let $\mathcal{H} = \mathrm{Hom}(\mathbb{V}, \mathbb{W})$, a $\mathfrak{g}$-module under

$$(X \cdot \phi)(u) = \rho(X)\phi(u) - \phi(\rho(X)u) \,. \tag{A.686}$$

Let $\mathcal{V} \subseteq \mathcal{H}$ be the set of ϕ whose restriction to $\mathbb{W}$ is a scalar multiple of the identity, and $\mathcal{W} \subseteq \mathcal{V}$ the set of ϕ vanishing on $\mathbb{W}$. Both are subspaces, and the map $\mathcal{V} \to \mathbb{F}$ sending ϕ to its scalar is linear, surjective — any projection of $\mathbb{V}$ onto $\mathbb{W}$ lies in $\mathcal{V}$ with scalar 1 — and has kernel $\mathcal{W}$, so $\mathcal{W}$ has codimension one in $\mathcal{V}$.

They are submodules: if $\phi|_{\mathbb{W}} = \lambda \mathbb{1}$ then for $w \in \mathbb{W}$,

$$(X \cdot \phi)(w) = \rho(X)\lambda w - \lambda \rho(X) w = 0 \,, \tag{A.687}$$

using that $\mathbb{W}$ is a submodule; so $X \cdot \mathcal{V} \subseteq \mathcal{W} \subseteq \mathcal{V}$.

Apply Lemma A.406 to $\mathcal{W} \subseteq \mathcal{V}$: there is a one-dimensional submodule $\mathbb{F}\phi_0$ complementary to $\mathcal{W}$, and ϕ_0 may be normalized so that $\phi_0|_{\mathbb{W}} = \mathbb{1}$. Being a one-dimensional submodule with $X \cdot \mathcal{V} \subseteq \mathcal{W}$, it satisfies $X \cdot \phi_0 \in \mathbb{F}\phi_0 \cap \mathcal{W} = 0$: that is, ϕ_0 is a homomorphism of modules. Its kernel is therefore a submodule; it meets $\mathbb{W}$ in 0 and has dimension $\dim \mathbb{V} - \dim \mathbb{W}$ because ϕ_0 is onto $\mathbb{W}$. Hence $\mathbb{V} = \mathbb{W} \oplus \ker \phi_0$. □

A.37.5 Trivial coefficients

Proposition A.408 (The two groups vanish for the trivial module)**.** *Let $\mathfrak{g}$ be semisimple over $\mathbb{F}$ and let $\mathbb{F}$ carry the trivial action. Then $H^1(\mathfrak{g}, \mathbb{F}) = H^2(\mathfrak{g}, \mathbb{F}) = 0$.*

Rests on Theorem A.407, Corollary A.342, and Proposition 14.76.

Proof. *Derives Proposition A.408.*

Degree one. With $\rho = 0$, Equation (A.661) makes a 1-cocycle a linear map f with $f([X, Y]) = 0$, that is, a linear map vanishing on $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$ (Corollary A.342); so $f = 0$. The coboundaries are zero as well, and $H^1 = 0$.

Degree two. Let c be a 2-cocycle. By the construction in the proof of Proposition 14.76 — "every class is realized" — the vector space $\tilde{\mathfrak{g}} = \mathfrak{g} \oplus \mathbb{F}Z$ with Z central and the bracket Equation (14.120) is a Lie algebra, the Jacobi identity being Equation (14.121), which by Remark A.399 is $\mathrm{d}c = 0$.

Make $\tilde{\mathfrak{g}}$ a $\mathfrak{g}$-module. For $\xi \in \mathfrak{g}$ and $u \in \tilde{\mathfrak{g}}$ set $\xi \cdot u = \left[\tilde{\xi}, u\right]$, where $\tilde{\xi}$ is any preimage of ξ under the projection $\pi : \tilde{\mathfrak{g}} \to \mathfrak{g}$. This is well defined, because two preimages differ by a multiple of the central Z; and it is a module structure, because $\left[\widetilde{[\xi, \eta]}, u\right] = \left[\left[\tilde{\xi}, \tilde{\eta}\right], u\right]$ — the two lifts differ by a central element — and the right-hand side expands by the Jacobi identity of $\tilde{\mathfrak{g}}$ into $\left[\tilde{\xi}, [\tilde{\eta}, u]\right] - \left[\tilde{\eta}, \left[\tilde{\xi}, u\right]\right]$.

The line $\mathbb{F}Z$ is a submodule, Z being central. By Theorem A.407 it has a complementary submodule $\mathfrak{m}$, so $\tilde{\mathfrak{g}} = \mathbb{F}Z \oplus \mathfrak{m}$ with $[\tilde{\mathfrak{g}}, \mathfrak{m}] \subseteq \mathfrak{m}$; in particular $\mathfrak{m}$ is a subalgebra, and $\pi|_{\mathfrak{m}} : \mathfrak{m} \to \mathfrak{g}$ is a bijective homomorphism of Lie algebras. Its inverse $s = (\pi|_{\mathfrak{m}})^{-1}$ is a linear splitting of π which is also a homomorphism, so the cocycle it defines through Equation (14.120) is $[s(\xi), s(\eta)] - s([\xi, \eta]) = 0$.

Finally, two linear splittings of π differ by a linear map $\mathfrak{g} \to \mathbb{F}$, and by the first part of the proof of Proposition 14.76 the cocycles they define differ by a coboundary. The splitting $\xi \mapsto (\xi, 0)$ defines c and the splitting s defines 0; hence c is a coboundary and $H^2(\mathfrak{g}, \mathbb{F}) = 0$. □

A.37.6 The Whitehead lemmas

Lemma A.409 (A Casimir invertible on a nontrivial irreducible module)**.** *Let $\mathfrak{g}$ be semisimple over $\mathbb{F}$ and let $\mathbb{V}$ be an irreducible module with $\rho \neq 0$. Then there is an invariant nondegenerate symmetric form β on $\mathfrak{g}$ whose Casimir operator Equation (A.668) is invertible on $\mathbb{V}$.*

Rests on Corollaries A.342 and A.343 and Lemma A.405.

Proof. *Derives Lemma A.409.*

Write $\mathfrak{g} = \mathfrak{g}_1 \oplus \cdots \oplus \mathfrak{g}_s$ as a direct sum of simple ideals (Corollary A.342) and let κ_j be the Killing form of $\mathfrak{g}_j$, extended to $\mathfrak{g}$ by declaring the other summands orthogonal to $\mathfrak{g}_j$ and to each other. Each κ_j is invariant and, by Corollary A.343, $\kappa = \sum_j \kappa_j$. For nonzero scalars $t_1, \dots, t_s$ put $\beta_t = \sum_j t_j \kappa_j$, an invariant symmetric form, nondegenerate because each κ_j is nondegenerate on $\mathfrak{g}_j$. Choosing a basis adapted to the decomposition, the β_t-dual basis is t_j^{-1} times the κ-dual basis in block j, so

$$\Gamma_t = \sum_j t_j^{-1} \Gamma_j \,, \qquad \Gamma_j = \sum_{a \in j} \rho\left(T^a\right) \rho\left(T_a\right) \,, \tag{A.688}$$

the inner sum running over a basis of $\mathfrak{g}_j$ with its κ_j-dual. Each Γ_j commutes with $\rho(\mathfrak{g}_j)$ by Lemma A.401 applied inside $\mathfrak{g}_j$, and with $\rho(\mathfrak{g}_i)$ for $i \neq j$ because $[\mathfrak{g}_i, \mathfrak{g}_j] = 0$; so it commutes with the whole action, and Lemma A.405 gives $\Gamma_j = \lambda_j \mathbb{1}$ on the irreducible $\mathbb{V}$.

The λ_j are not all zero. Let $\beta_{\mathbb{V}}$ be the trace form Equation (A.672), which is invariant but need not be nondegenerate. Its restriction to the simple ideal $\mathfrak{g}_j$ is a multiple of κ_j: writing $\beta_{\mathbb{V}}(X, Y) = \kappa_j(AX, Y)$ for $X, Y \in \mathfrak{g}_j$ with $A \in \operatorname{End}(\mathfrak{g}_j)$, invariance of both forms gives

$$\kappa_j\left(A\left[Z, X\right], Y\right) = \beta_{\mathbb{V}}\left(\left[Z, X\right], Y\right) = -\beta_{\mathbb{V}}\left(X, \left[Z, Y\right]\right) = -\kappa_j\left(AX, \left[Z, Y\right]\right) = \kappa_j\left(\left[Z, AX\right], Y\right) \tag{A.689}$$

for all $Z \in \mathfrak{g}_j$, so A commutes with every ad_Z. The adjoint representation of a simple algebra is irreducible — its submodules are its ideals — so Lemma A.405 gives $A = c_j \mathbb{1}$ and $\beta_{\mathbb{V}}|_{\mathfrak{g}_j} = c_j \kappa_j$. Now

$$\lambda_j \dim \mathbb{V} = \operatorname{tr}_{\mathbb{V}} \Gamma_j = \sum_{a \in j} \beta_{\mathbb{V}}\left(T^a, T_a\right) = c_j \sum_{a \in j} \kappa_j\left(T^a, T_a\right) = c_j \dim \mathfrak{g}_j \,. \tag{A.690}$$

If $c_j = 0$ then $\beta_{\mathbb{V}}$ vanishes on $\mathfrak{g}_j$, so $\operatorname{tr}(\rho(X)\rho(Y)) = 0$ for all $X \in [\mathfrak{g}_j, \mathfrak{g}_j] = \mathfrak{g}_j$ and $Y \in \mathfrak{g}_j$, and Theorem A.340 makes $\rho(\mathfrak{g}_j)$ solvable; being a homomorphic image of the simple $\mathfrak{g}_j$ it is either isomorphic to $\mathfrak{g}_j$, which is not solvable, or zero. So $c_j = 0$ forces $\rho(\mathfrak{g}_j) = 0$. Since $\rho \neq 0$, some $\mathfrak{g}_{j_0}$ acts nontrivially and $\lambda_{j_0} \neq 0$.

Choice of t. The eigenvalue of Γ_t on $\mathbb{V}$ is $\sum_j t_j^{-1} \lambda_j$. If $\sum_j \lambda_j \neq 0$ take every $t_j = 1$. Otherwise take $t_{j_0} = 2$ and $t_j = 1$ for $j \neq j_0$; the eigenvalue is then $\sum_j \lambda_j - \frac{1}{2}\lambda_{j_0} = -\frac{1}{2}\lambda_{j_0} \neq 0$. In either case Γ_t is a nonzero scalar on $\mathbb{V}$, hence invertible. □

Theorem A.410 (Whitehead's first and second lemmas)**.** *Let $\mathfrak{g}$ be a finite-dimensional semisimple Lie algebra over a field of characteristic zero and let $\mathbb{V}$ be any finite-dimensional $\mathfrak{g}$-module. Then*

$$\boxed{H^1(\mathfrak{g}, \mathbb{V}) = H^2(\mathfrak{g}, \mathbb{V}) = 0} \,. \tag{A.691}$$

Rests on Theorems A.403 and A.407 and Lemma A.409.

Proof. *Derives Theorem A.410.*

Reduction to an algebraically closed field. Each $C^n(\mathfrak{g}, \mathbb{V})$ is a finite-dimensional $\mathbb{K}$-vector space and d is $\mathbb{K}$-linear; extending scalars to $\mathbb{F}$ gives $C^n(\mathfrak{g}, \mathbb{V}) \otimes_{\mathbb{K}} \mathbb{F} = C^n(\mathfrak{g}_{\mathbb{F}}, \mathbb{V}_{\mathbb{F}})$, with the same differential, because a multilinear map is fixed by its values on a $\mathbb{K}$-basis. The

ranks of d in each degree are unchanged by field extension, so $H^n(\mathfrak{g},\mathbb{V})\otimes_{\mathbb{K}}\mathbb{F}=H^n\left(\mathfrak{g}_{\mathbb{F}},\mathbb{V}_{\mathbb{F}}\right)$, and a space vanishes if and only if its extension does. Moreover $\mathfrak{g}_{\mathbb{F}}$ is again semisimple by Remark A.341. It therefore suffices to prove the theorem over $\mathbb{F}$.

Reduction to irreducible coefficients. By Theorem A.407, $\mathbb{V}=\mathbb{V}_1\oplus\cdots\oplus\mathbb{V}_r$ with each $\mathbb{V}_i$ irreducible. A cochain with values in a direct sum is the same thing as a tuple of cochains with values in the summands, and d acts componentwise, so

$$H^n\left(\mathfrak{g},\mathbb{V}\right)=\bigoplus_{i=1}^{r}H^n\left(\mathfrak{g},\mathbb{V}_i\right)\,. \tag{A.692}$$

The two cases. If $\mathfrak{g}$ acts as zero on the irreducible $\mathbb{V}_i$, then every subspace is a submodule, so $\dim\mathbb{V}_i=1$ and $\mathbb{V}_i\cong\mathbb{F}$ with the trivial action; Proposition A.408 gives $H^1=H^2=0$. Otherwise Lemma A.409 supplies an invariant nondegenerate form whose Casimir is invertible on $\mathbb{V}_i$, and Theorem A.403 gives $H^1=H^2=0$. Summing over i in Equation (A.692) completes the proof. □

Corollary A.411 (Semisimple algebras admit no nontrivial extension)**.** *A finite-dimensional semisimple Lie algebra $\mathfrak{g}$ over a field of characteristic zero satisfies $H^1(\mathfrak{g},\mathbb{R})=H^2(\mathfrak{g},\mathbb{R})=0$ and admits no nontrivial central extension. In particular, for $D=p+q\geq 3$ the algebra $\mathfrak{so}(p,q)$ — and hence the Lorentz algebra $\mathfrak{so}(D-1,1)$ of Equation (14.90), and the rotation algebra $\mathfrak{so}(3)$ — admits no nontrivial central extension.*

Rests on Theorem A.410, Corollary A.344, and Proposition 14.76.

Proof. *Derives Corollary A.411.*

Take $\mathbb{V}=\mathbb{R}$ with the trivial action in Theorem A.410; by Remark A.399 the group H^2 so computed is the group $H^2(\mathfrak{g},\mathbb{R})$ of Definition 14.75, and by Proposition 14.76 its vanishing says that every central extension of $\mathfrak{g}$ by $\mathbb{R}$ is trivial, that is, isomorphic to $\mathfrak{g}\oplus\mathbb{R}$ as a Lie algebra. That $\mathfrak{so}(p,q)$ is semisimple for $D\geq 3$ is Corollary A.344, proved there by computing its Killing form and applying Theorem A.330. □

Remark A.412 (Why the hypotheses cannot be dropped). Each hypothesis of Theorem A.410 earns its place, and the counterexamples are the physical cases of Section 14.4.2. *Semisimplicity*: the abelian algebra $\mathbb{R}^{2f}$ of Example 14.79 has $H^2=\Lambda^2\left(\mathbb{R}^{2f}\right)^*$, as far from zero as possible, and the extension is the Heisenberg algebra with $Z=\mathrm{i}\hbar$; the proof fails at the first step, since $[\mathfrak{g},\mathfrak{g}]=0$ and the Killing form vanishes identically. *Finite dimension*: the Witt algebra of Example 14.81 has a one-dimensional H^2 whose extension is the Virasoro algebra, and the Casimir Equation (A.668) is an infinite sum there, with no meaning as an operator on the algebra. *Characteristic zero*: the division by λ in Lemma A.406 and by $\dim\mathbb{V}$ in Equation (A.690) both fail in characteristic p, and so does the theorem.

The Galilei algebra of Example 14.80 is the instructive intermediate case: it is finite-dimensional over $\mathbb{R}$ but not semisimple — it has the translations and the boosts as abelian subalgebras and is a contraction (Section 14.5) of an algebra that is — and it carries the mass as a central charge. So the mass is not an accident of the nonrelativistic limit that a better choice of generators could remove; it is an obstruction that exists because the algebra fails the hypothesis of this theorem.

Appendix A.37 discharges the proof obligation of Proposition 14.77. Its statement is used at Remark 14.78 to localize where nontrivial central charges can live — not in semisimple algebras, hence only in abelian algebras, in inhomogeneous semidirect sums such as Equation (14.99), in contractions (Section 14.5) and in infinite-dimensional algebras — and at Remark 14.85, where the cohomological reading of a contraction is set out. Two by-products proved along the way have independent value and are used tacitly elsewhere in the chapter: Weyl's complete reducibility theorem (Theorem A.407), which is why a finite-dimensional representation of a semisimple algebra may always be decomposed into

irreducibles as Theorem 14.47 does for $\mathfrak{su}(2)$, and the vanishing of $H^1(\mathfrak{g}, \mathbb{V})$, which is the statement that a semisimple algebra admits no nontrivial deformation of a module structure.

A.38 The Second Cohomology of the Galilei Algebra

This appendix proves the assertion left owed in Example 14.80 of Chapter 14: the second Chevalley–Eilenberg cohomology $H^2(\mathfrak{g}, \mathbb{R})$ of the ten-generator Galilei algebra of three space dimensions is *one-dimensional*, so that up to a scale the mass extension is the only central extension the algebra admits.

Exactly one half of that statement is proved elsewhere in this treatise. Proposition 25.14, in the chapter bridging Poisson brackets and quantization, shows that the mass cocycle $c(K_i, P_j) = m\,\delta_{ij}$ of Equation (25.15) is not a coboundary — that its class in H^2 is *nonzero*. It says nothing about how large H^2 is, and the chapter records the difference honestly. What is added here is the other half: that the class of the mass cocycle *exhausts* the cohomology. Nothing in Chapter 14 or in Part III needs the stronger statement, so this appendix closes a gap in the mathematics rather than supplying a missing physical input.

The computation is finite and is carried out in full. The space of alternating bilinear forms on a ten-dimensional algebra has dimension $\binom{10}{2} = 45$; the cocycle condition Equation (14.121) cuts it down to 10, the coboundaries account for 9 of those, and the survivor is the mass.

A.38.1 The algebra, and the two spaces

Definition A.413 (The Galilei algebra of three space dimensions)**.** Let $\mathfrak{g}$ be the real ten-dimensional Lie algebra spanned by rotations J_i, boosts K_i, space translations P_i $(i = 1, 2, 3)$ and the time translation H, with the isotropy relations Equation (A.161),

$$[J_i, J_j] = \epsilon_{ijk} J_k\,, \quad [J_i, K_j] = \epsilon_{ijk} K_k\,, \quad [J_i, P_j] = \epsilon_{ijk} P_k\,, \quad [J_i, H] = 0\,, \tag{A.693}$$

and the single further nonvanishing bracket

$$[K_i, H] = P_i\,, \qquad [K_i, K_j] = [P_i, P_j] = [K_i, P_j] = [P_i, H] = 0\,. \tag{A.694}$$

This is the case $(\gamma, \mu, \nu) = (0, *, 0)$ of the enumeration in Appendix A.10.5, with $\alpha = -\gamma\mu = 0$ and $\beta = \gamma\nu = 0$; the normalisation $\mu = 1$ is the one fixed there. All indices are Cartesian, repeated indices are summed, and ϵ_{ijk} is the Levi-Civita symbol.

Rests on Equations (A.161) *and* (A.165) *and Theorem 14.90.*

Theorem A.414 (H^2 of the Galilei algebra)**.** $\dim_{\mathbb{R}} H^2(\mathfrak{g}, \mathbb{R}) = 1$*, and the class is represented by the* mass cocycle

$$c_M(K_i, P_j) = \delta_{ij}\,, \qquad c_M = 0 \text{ on every other pair of basis generators}\,. \tag{A.695}$$

Every central extension of $\mathfrak{g}$ by $\mathbb{R}$ is therefore equivalent to the one with bracket $\left[\tilde{K}_i, \tilde{P}_j\right] = m\,\delta_{ij} Z$ for a single real number m, and it is trivial exactly when $m = 0$.

Rests on Definitions 14.75 and A.413 and Proposition 14.76.

A cocycle is fixed by its values on pairs of basis generators. Grouping those values by type, an alternating bilinear $c : \mathfrak{g} \times \mathfrak{g} \to \mathbb{R}$ is exactly the datum

$$\begin{aligned} &A_{ij} = c(J_i, J_j)\,, \quad B_{ij} = c(J_i, K_j)\,, \quad C_{ij} = c(J_i, P_j)\,, \quad u_i = c(J_i, H)\,, \\ &D_{ij} = c(K_i, K_j)\,, \quad E_{ij} = c(K_i, P_j)\,, \quad v_i = c(K_i, H)\,, \quad F_{ij} = c(P_i, P_j)\,, \\ &w_i = c(P_i, H)\,, \end{aligned} \tag{A.696}$$

with A, D, F antisymmetric and B, C, E unrestricted: $3+9+9+3+3+9+3+3+3 = 45$ real parameters, as it must be.

A.38.2 Three tensor lemmas

The cocycle conditions that involve one or two rotation generators are statements about SO(3)-covariance of the arrays in Equation (A.696), and all of them reduce to two identities. They are isolated here so that the enumeration below reads cleanly. Throughout, the only inputs are the standard contractions

$$\epsilon_{ijm}\epsilon_{ijl} = 2\delta_{ml}\,, \qquad \epsilon_{ijm}\epsilon_{ikl} = \delta_{jk}\delta_{ml} - \delta_{jl}\delta_{mk}\,, \qquad \epsilon_{ijl}\epsilon_{lkm} = \delta_{ik}\delta_{jm} - \delta_{im}\delta_{jk}\,. \tag{A.697}$$

Lemma A.415 (Invariant two-tensors)**.** *A real array T_{jk} satisfies*

$$\epsilon_{ijl}T_{lk} + \epsilon_{ikl}T_{jl} = 0 \qquad \text{for all } i, j, k \tag{A.698}$$

if and only if $T_{jk} = t\,\delta_{jk}$ for a real number t.
Rests on Equation (A.697)*.*

Proof. *Derives Lemma A.415.*

Necessity. Contract Equation (A.698) with ϵ_{ijm} and sum over i and j. By the first identity of Equation (A.697) the first term gives $2T_{mk}$; by the second, the remaining term gives

$$\epsilon_{ijm}\epsilon_{ikl}T_{jl} = (\delta_{jk}\delta_{ml} - \delta_{jl}\delta_{mk})\,T_{jl} = T_{km} - \delta_{mk}\,\operatorname{tr}T\,,$$

where $\operatorname{tr}T = T_{jj}$. Hence

$$2T_{mk} + T_{km} - \delta_{mk}\,\operatorname{tr}T = 0\,. \tag{A.699}$$

Exchanging m and k in Equation (A.699) and subtracting gives $T_{mk} - T_{km} = 0$, so T is symmetric; adding instead gives $3\,(T_{mk} + T_{km}) = 2\delta_{mk}\operatorname{tr}T$, that is $T_{mk} = \frac{1}{3}\,(\operatorname{tr}T)\,\delta_{mk}$.

Sufficiency. For $T = \delta$ the left side of Equation (A.698) is $\epsilon_{ijk} + \epsilon_{ikj} = 0$. □

Lemma A.416 (The mixed covariance condition)**.** *A real array T_{jk} satisfies*

$$\epsilon_{ijl}T_{lk} + \epsilon_{ikl}T_{jl} - \epsilon_{jkl}T_{il} = 0 \qquad \text{for all } i, j, k \tag{A.700}$$

if and only if T is antisymmetric, that is $T_{jk} = \epsilon_{jkm}t_m$ for a real vector t.
Rests on Equation (A.697)*.*

Proof. *Derives Lemma A.416.*

Necessity. Contract with ϵ_{ijm} as before. The first two terms give $2T_{mk}$ and $T_{km} - \delta_{mk}\operatorname{tr}T$ exactly as in Lemma A.415. For the third, use $\epsilon_{ijm}\epsilon_{jkl} = -\epsilon_{jim}\epsilon_{jkl} = -\,(\delta_{ik}\delta_{ml} - \delta_{il}\delta_{mk})$, so that

$$-\epsilon_{ijm}\epsilon_{jkl}T_{il} = (\delta_{ik}\delta_{ml} - \delta_{il}\delta_{mk})\,T_{il} = T_{km} - \delta_{mk}\,\operatorname{tr}T\,.$$

Summing the three,

$$2T_{mk} + 2T_{km} - 2\delta_{mk}\,\operatorname{tr}T = 0\,, \qquad \text{i.e.} \qquad T_{mk} + T_{km} = \delta_{mk}\,\operatorname{tr}T\,. \tag{A.701}$$

Setting $m = k$ and summing gives $2\operatorname{tr}T = 3\operatorname{tr}T$, so $\operatorname{tr}T = 0$, and then Equation (A.701) reads $T_{mk} = -T_{km}$.

Sufficiency. Put $T_{jk} = \epsilon_{jkm}t_m$ and evaluate the three terms of Equation (A.700) with the third identity of Equation (A.697), applied after cycling the symbols into the shape $\epsilon_{abl}\epsilon_{lcd}$:

$$\begin{aligned}
\epsilon_{ijl}T_{lk} &= \epsilon_{ijl}\epsilon_{lkm}t_m = \delta_{ik}t_j - \delta_{jk}t_i\,,\\
\epsilon_{ikl}T_{jl} &= \epsilon_{ikl}\epsilon_{lmj}t_m = \delta_{im}\delta_{kj}t_m - \delta_{ij}\delta_{km}t_m = \delta_{kj}t_i - \delta_{ij}t_k\,,\\
\epsilon_{jkl}T_{il} &= \epsilon_{jkl}\epsilon_{lmi}t_m = \delta_{jm}\delta_{ki}t_m - \delta_{ji}\delta_{km}t_m = \delta_{ki}t_j - \delta_{ji}t_k\,.
\end{aligned}$$

The first two add to $\delta_{ik}t_j - \delta_{ij}t_k$, which is the third; the alternating sum vanishes. □

Remark A.417 (Where the three space dimensions enter). Lemmas A.415 and A.416 are the only steps of the computation that are not pure bookkeeping, and both are statements about the invariant two-tensors of the rotation group of *three* space dimensions: they use the Levi-Civita symbol with three indices and the contraction identities Equation (A.697), which are three-dimensional. The theorem proved here is accordingly a theorem about the Galilei algebra of the observed three space dimensions, which is the only case this treatise instantiates; no claim is made, and none is needed, for any other number of space dimensions.

A.38.3 Solving the cocycle condition

Write $\delta c(x, y, z)$ for the left-hand side of Equation (14.121),

$$\delta c(x, y, z) = c\left([x, y], z\right) + c\left([y, z], x\right) + c\left([z, x], y\right) , \tag{A.702}$$

which is alternating in (x, y, z), so it suffices to impose $\delta c = 0$ on unordered triples of basis generators; a triple containing H twice gives nothing, since there is only one H. Sixteen types remain, and every one is evaluated below.

Triples with no bracket. For (K_i, K_j, K_k), (P_i, P_j, P_k), (K_i, K_j, P_k), (K_i, P_j, P_k) and (P_i, P_j, H) every bracket appearing in Equation (A.702) vanishes by Equation (A.694), so the condition is empty.

The triple (J_i, J_j, J_k). Writing $A_{ij} = \epsilon_{ijm} a_m$ and using the third identity of Equation (A.697), $\epsilon_{ijl} A_{lk} = \delta_{ik} a_j - \delta_{jk} a_i$, so the cyclic sum is

$$(\delta_{ik} a_j - \delta_{jk} a_i) + (\delta_{ji} a_k - \delta_{ki} a_j) + (\delta_{kj} a_i - \delta_{ij} a_k) = 0$$

identically. *The array A is unconstrained*, and it appears in no other triple: a term $c([\cdot, \cdot], J)$ can involve A only when the bracket produces a rotation, and only $[J, J]$ does.

The triple (J_i, J_j, K_k). Using Equation (A.693) three times,

$$\delta c = \epsilon_{ijl} B_{lk} + \epsilon_{jkl} c(K_l, J_i) - \epsilon_{ikl} c(K_l, J_j) = \epsilon_{ijl} B_{lk} + \epsilon_{ikl} B_{jl} - \epsilon_{jkl} B_{il} ,$$

which is Equation (A.700). By Lemma A.416,

$$B_{jk} = \epsilon_{jkm} \beta_m \tag{A.703}$$

for an arbitrary vector β, and nothing more.

The triple (J_i, J_j, P_k). Identical in form, with C in place of B: it forces C to be antisymmetric and imposes nothing else. It will be subsumed by the stronger condition below.

The triple (J_i, J_j, H). Since $[J_i, H] = [H, J_j] = 0$, only the first term survives: $\delta c = \epsilon_{ijl} u_l = 0$ for all i, j, hence

$$u_i = c(J_i, H) = 0 . \tag{A.704}$$

The triple (J_i, K_j, K_k). Here $[K_j, K_k] = 0$, so

$$\delta c = \epsilon_{ijl} D_{lk} - \epsilon_{ikl} D_{lj} = \epsilon_{ijl} D_{lk} + \epsilon_{ikl} D_{jl} ,$$

the second step using the antisymmetry of D. This is Equation (A.698), so $D = t\delta$ by Lemma A.415; but D is antisymmetric and δ is not, so $t = 0$ and

$$D_{ij} = c(K_i, K_j) = 0 . \tag{A.705}$$

The triple (J_i, P_j, P_k). Word for word the same with F in place of D, giving $F_{ij} = c(P_i, P_j) = 0$.

The triple (J_i, K_j, H). Now the bracket $[K_j, H] = P_j$ contributes:

$$\delta c = \epsilon_{ijl} c(K_l, H) + c(P_j, J_i) + c(0, K_j) = \epsilon_{ijl} v_l - C_{ij}\,,$$

so that

$$C_{ij} = c(J_i, P_j) = \epsilon_{ijl} v_l\,. \tag{A.706}$$

The array C is thus determined by the vector v, and is automatically antisymmetric — consistent with, and stronger than, the (J, J, P) condition.

The triple (J_i, P_j, H). Both $[P_j, H]$ and $[H, J_i]$ vanish, leaving $\delta c = \epsilon_{ijl} w_l = 0$, hence

$$w_i = c(P_i, H) = 0\,. \tag{A.707}$$

The triple (J_i, K_j, P_k). With $[K_j, P_k] = 0$,

$$\delta c = \epsilon_{ijl} E_{lk} - \epsilon_{ikl} c(P_l, K_j) = \epsilon_{ijl} E_{lk} + \epsilon_{ikl} E_{jl}\,,$$

which is Equation (A.698) again. By Lemma A.415,

$$E_{ij} = c(K_i, P_j) = e\,\delta_{ij}\,, \qquad e \in \mathbb{R}\,. \tag{A.708}$$

The triple (K_i, K_j, H). Only the brackets with H survive:

$$\delta c = c(P_j, K_i) + c(-P_i, K_j) = -E_{ij} + E_{ji} = 0\,,$$

which says E is symmetric — already true of Equation (A.708), so no new information.

The triple (K_i, P_j, H). Here $[K_i, P_j] = 0$ and $[P_j, H] = 0$, and $[H, K_i] = -P_i$, so $\delta c = c(-P_i, P_j) = -F_{ij} = 0$: again already known.

That exhausts the sixteen types. Collecting Equations (A.703) to (A.708) and the two vanishing statements for F:

Proposition A.418 (The cocycles)**.** *A bilinear form on* $\mathfrak{g}$ *is a* 2*-cocycle in the sense of Proposition 14.73 if and only if*

$$A_{ij} = \epsilon_{ijm} a_m\,, \quad B_{ij} = \epsilon_{ijm} \beta_m\,, \quad C_{ij} = \epsilon_{ijm} v_m\,, \quad E_{ij} = e\,\delta_{ij}\,, \quad c(K_i, H) = v_i\,, \tag{A.709}$$

with $u = w = 0$ *and* $D = F = 0$*, for arbitrary vectors* $a, \beta, v \in \mathbb{R}^3$ *and an arbitrary real number* e*. Hence* $\dim_{\mathbb{R}} Z^2(\mathfrak{g}, \mathbb{R}) = 3 + 3 + 3 + 1 = 10$.

Rests on Definition A.413 and Lemmas A.415 and A.416.

A.38.4 The coboundaries, and the quotient

Proposition A.419 (The coboundaries)**.** $\dim_{\mathbb{R}} B^2(\mathfrak{g}, \mathbb{R}) = 9$*, and a coboundary is exactly a cocycle Equation* (A.709) *with* $e = 0$.

Rests on Definitions 14.74 and A.413 and Proposition A.418.

Proof. *Derives Proposition A.419.*

Let $b : \mathfrak{g} \to \mathbb{R}$ be linear, with components $b(J_i) = \beta_i^J$, $b(K_i) = \beta_i^K$, $b(P_i) = \beta_i^P$ and $b(H) = \eta$. Evaluating $(\partial b)(x, y) = b([x, y])$ of Equation (14.122) on each pair of basis generators with Equations (A.693) and (A.694),

$$\begin{aligned} &(\partial b)(J_i, J_j) = \epsilon_{ijk} \beta_k^J\,, \quad (\partial b)(J_i, K_j) = \epsilon_{ijk} \beta_k^K\,, \quad (\partial b)(J_i, P_j) = \epsilon_{ijk} \beta_k^P\,, \\ &(\partial b)(K_i, H) = \beta_i^P\,, \qquad (\partial b)(J_i, H) = 0\,, \qquad (\partial b)(K_i, P_j) = 0\,, \end{aligned} \tag{A.710}$$

and $(\partial b)(K_i, K_j) = (\partial b)(P_i, P_j) = (\partial b)(P_i, H) = 0$. In the coordinates of Equation (A.709) this is

$$a = \beta^J\,, \qquad \beta = \beta^K\,, \qquad v = \beta^P\,, \qquad e = 0\,, \tag{A.711}$$

where the third entry is consistent in both places it occurs: $C_{ij} = \epsilon_{ijk}\beta^P_k$ and $c(K_i, H) = \beta^P_i$ are exactly the pair required by Equation (A.709). Two consequences follow at once. First, every coboundary *is* a cocycle with $e = 0$ — which also re-proves Proposition 14.73 for this algebra. Second, the map $b \mapsto \partial b$ carries $\left(\beta^J, \beta^K, \beta^P\right)$ onto the whole nine-dimensional subspace $\{e = 0\}$ of Z^2, bijectively, while η is annihilated. Hence $\dim B^2 = 9$.

The kernel is visible structurally as well: $\partial b = 0$ if and only if b vanishes on the derived algebra $[\mathfrak{g}, \mathfrak{g}]$, which by Equations (A.693) and (A.694) is the span of the J_i, K_i and P_i — the P_i arising both from $[J, P]$ and from $[K, H]$ — of dimension 9. No bracket produces H, so H is not in the derived algebra and the kernel is the line $\left\{\beta^J = \beta^K = \beta^P = 0\right\}$, of dimension $10 - 9 = 1$. □

Proof of Theorem A.414. *Derives Theorem A.414.*

By Propositions A.418 and A.419 the coboundaries are the hyperplane $\{e = 0\}$ inside the ten-dimensional space of cocycles, so

$$\dim_{\mathbb{R}} H^2(\mathfrak{g}, \mathbb{R}) = \dim Z^2 - \dim B^2 = 10 - 9 = 1\,, \tag{A.712}$$

and the coordinate e descends to an isomorphism $H^2(\mathfrak{g}, \mathbb{R}) \cong \mathbb{R}$. The cocycle c_M of Equation (A.695) has $a = \beta = v = 0$ and $e = 1$: it satisfies Equation (A.709), hence is a cocycle, and its class is a generator. The final statement is Proposition 14.76: central extensions of $\mathfrak{g}$ by $\mathbb{R}$ correspond to classes in H^2, so they form a one-parameter family $m\,[c_M]$, realized by $\left[\tilde{K}_i, \tilde{P}_j\right] = m\,\delta_{ij} Z$ through Equation (14.120), and by Equation (14.123) the extension is trivial precisely for the zero class, $m = 0$. □

Remark A.420 (What was already known, and what is new). Proposition 25.14 proves that the class of c_M is nonzero, by exhibiting the Poisson-bracket realisation of a free particle and showing that no shift of the generators by constants removes the defect Equation (25.15). The present appendix reproves that as the single line $e \neq 0$ in Equation (A.712) — coboundaries have $e = 0$, c_M has $e = 1$ — and adds the converse, that there is nothing else: the whole of H^2 is the mass.

Two structural readings of the computation are worth recording, because they explain the answer rather than merely producing it. The first is that the surviving parameter is the one component of c that no coboundary can reach, namely $c(K_i, P_j)$, and it is unreachable because $[K_i, P_j] = 0$ in $\mathfrak{g}$: a coboundary is a linear functional composed with the bracket, and there is no bracket there to compose with. That is exactly the argument of Proposition 25.14, and the enumeration above shows it is the *only* such place. The second is that the answer is forced to be at most one-dimensional by rotational covariance: Lemma A.415 says the surviving array $c(K_i, P_j)$ has nowhere to live but the invariant δ_{ij}, so the mass is a single number and not a tensor.

Example 14.80 may therefore be read without its final caveat, and Remark 14.88 — which describes how the mass emerges from the Inönü–Wigner limit of a trivially extended Poincaré algebra — is now accompanied by an independent statement of where the limit must land: H^2 of the contracted algebra is one-dimensional, so any nontrivial extension produced by any route is a multiple of c_M. This is a concrete instance of the phenomenon named in Remark 14.78: a contraction can create a cohomology class that the parent semisimple algebra, rigid by Proposition 14.77, did not have.

A.39 Central Extensions of Three Infinite-Dimensional Algebras

Example 14.81 of Chapter 14 lists three infinite-dimensional algebras whose central extensions are nontrivial: the Witt algebra of vector fields on the circle, whose extension is the Virasoro

algebra; the loop algebra of a simple Lie algebra, whose extension is affine Kac–Moody; and the equal-time current algebra of a field theory, whose extension is the Schwinger term. This appendix supplies the three computations.

They are of two different kinds, and the difference is the point of the section. The first two are finite algebraic computations in the sense of Definition 14.75: a cocycle is written down, the cocycle condition Equation (14.121) is solved, and the coboundaries Equation (14.122) are divided out. The third is not. There is no finite computation of a Schwinger term, because the object whose commutator is asked for — a product of operator-valued distributions at a point — does not exist until it is regularized. What can be proved, and is proved below, is a *no-go*: in any theory with a Hamiltonian bounded below, a normalizable ground state and a local conserved current, the equal-time commutator of the charge density with the current density cannot vanish, and the leading term it must carry is a central one of definite sign. Remark A.434 states exactly which hypotheses that argument consumes and which question it leaves open.

Throughout, the ground field is written F and may be taken to be $\mathbb{R}$ or $\mathbb{C}$; Definitions 14.72, 14.74 and 14.75 and Propositions 14.73 and 14.76 are stated in Section 14.4.1 for $\mathbb{R}$, and their statements and proofs use nothing but the field axioms in characteristic zero, so they hold verbatim over $\mathbb{C}$. Rule 1 of this treatise applies here as everywhere: what follows is mathematics about Lie algebras, and no claim is made about any physical theory in which these algebras have been proposed to act.

A.39.1 The Virasoro cocycle

Definition A.421 (Witt algebra)**.** The *Witt algebra* $\mathfrak{w}$ is the Lie algebra over F with basis $\{L_n\}_{n\in\mathbb{Z}}$ and brackets

$$[L_m, L_n] = (m-n)\,L_{m+n}\,, \qquad m,n\in\mathbb{Z}\,. \tag{A.713}$$

It is the algebra of polynomial vector fields on the circle: with $t=\mathrm{e}^{\mathrm{i}\theta}$ and $L_n=-t^{n+1}\frac{\mathrm{d}}{\mathrm{d}t}$ acting on $F[t,t^{-1}]$, a direct computation reproduces Equation (A.713). The Jacobi identity is the identity $(m-n)(m+n-k)+(n-k)(n+k-m)+(k-m)(k+m-n)=0$, which holds because every quadratic monomial cancels in pairs.

Rests on Definition 14.72.

Note that $\mathfrak{w}$ is infinite-dimensional, so a bilinear form on it is an arbitrary array of values $c(L_m,L_n)$; bilinearity constrains only finite linear combinations, and no continuity or boundedness is imposed anywhere below.

Theorem A.422 (The Virasoro extension)**.** *$H^2(\mathfrak{w},F)$ is one-dimensional, generated by the class of*

$$\boxed{\omega\,(L_m,L_n)=\frac{m^3-m}{12}\,\delta_{m+n,0}}\,. \tag{A.714}$$

Every central extension of $\mathfrak{w}$ by F is therefore equivalent, for exactly one $C\in F$, to the Virasoro algebra

$$[L_m,L_n]=(m-n)L_{m+n}+\frac{C}{12}\left(m^3-m\right)\delta_{m+n,0}\,Z\,, \qquad [Z,\cdot]=0\,. \tag{A.715}$$

Rests on Definitions 14.75 and A.421 and Proposition 14.76.

Proof. *Derives Theorem A.422.*

Let c be a 2-cocycle on $\mathfrak{w}$.

Step 1: normalize against L_0. Define a linear functional b on $\mathfrak{w}$ by

$$b\,(L_n)=\frac{c\,(L_n,L_0)}{n}\quad (n\neq 0)\,, \qquad b\,(L_0)=0\,. \tag{A.716}$$

Its coboundary is $(\partial b)(L_m, L_n) = b\left([L_m, L_n]\right) = (m-n)\, b(L_{m+n})$, so for $n \neq 0$

$$(\partial b)\left(L_n, L_0\right) = n\, b\left(L_n\right) = c\left(L_n, L_0\right) ,$$

while $(\partial b)(L_0, L_0) = 0 = c(L_0, L_0)$ by antisymmetry. Replacing c by $c - \partial b$ — which changes nothing in H^2 — we may and do assume

$$c\left(L_n, L_0\right) = 0 \qquad \text{for every } n \in \mathbb{Z} . \tag{A.717}$$

Step 2: only the diagonal $m+n=0$ survives. Impose the cocycle condition Equation (14.121) on the triple (L_0, L_m, L_n). With $[L_0, L_m] = -mL_m$, $[L_n, L_0] = nL_n$ and $[L_m, L_n] = (m-n)L_{m+n}$,

$$0 = -m\, c\left(L_m, L_n\right) + (m-n)\, c\left(L_{m+n}, L_0\right) + n\, c\left(L_n, L_m\right) = -(m+n)\, c\left(L_m, L_n\right) ,$$

the middle term vanishing by Equation (A.717) and the last by antisymmetry. Hence

$$c\left(L_m, L_n\right) = h(m)\, \delta_{m+n,0} , \qquad h(m) = c\left(L_m, L_{-m}\right) , \tag{A.718}$$

and antisymmetry gives $h(-m) = -h(m)$, with $h(0) = 0$.

Step 3: the recursion. A cocycle of the shape Equation (A.718) makes every term of Equation (14.121) vanish on a triple (L_m, L_n, L_k) with $m+n+k \neq 0$, since each term pairs L_{m+n} with L_k and the two indices sum to $m+n+k$. For $m+n+k=0$ the three terms are

$$\begin{aligned} (m-n)\, c\left(L_{-k}, L_k\right) + (n-k)\, c\left(L_{-m}, L_m\right) + (k-m)\, c\left(L_{-n}, L_n\right) \\ = -\left[(m-n)h(k) + (n-k)h(m) + (k-m)h(n)\right] , \end{aligned}$$

so the cocycle condition is $(m-n)h(k) + (n-k)h(m) + (k-m)h(n) = 0$ with $k = -m-n$, that is

$$(m-n)\, h(m+n) = (m+2n)\, h(m) - (2m+n)\, h(n) . \tag{A.719}$$

Setting $n = 1$,

$$(m-1)\, h(m+1) = (m+2)\, h(m) - (2m+1)\, h(1) , \tag{A.720}$$

which for $m \geq 2$ determines $h(m+1)$ from $h(m)$ and $h(1)$. Since $h(2)$ is not reached by Equation (A.720) — at $m = 1$ the left-hand side vanishes identically, and indeed the relation then reads $3h(1) = 3h(1)$ — the solution space has dimension at most 2, the free data being $h(1)$ and $h(2)$; negative arguments follow from $h(-m) = -h(m)$.

Two solutions are exhibited directly. For $h(m) = m$,

$$(m-n)(m+n) = m^2 - n^2 = \left(m^2 + 2mn\right) - \left(2mn + n^2\right) = (m+2n)m - (2m+n)n ,$$

and for $h(m) = m^3$, expanding both sides,

$$(m-n)(m+n)^3 = m^4 + 2m^3 n - 2mn^3 - n^4 = (m+2n)m^3 - (2m+n)n^3 ,$$

where the middle expression is obtained by multiplying out $(m-n)\left(m^3 + 3m^2 n + 3mn^2 + n^3\right)$ and cancelling the two $3m^2 n^2$ terms. The two are linearly independent, so the solution space of Equation (A.719) is exactly $\left\{h(m) = \alpha m + \beta m^3\right\}$.

Step 4: the linear solution is a coboundary. Take $b_\gamma\left(L_n\right) = \gamma\, \delta_{n,0}$. Then $\left(\partial b_\gamma\right)\left(L_m, L_n\right) = (m-n)\gamma\, \delta_{m+n,0} = 2m\gamma\, \delta_{m+n,0}$, which is Equation (A.718) with $h(m) = 2\gamma m$; and it respects Equation (A.717), since $\left(\partial b_\gamma\right)\left(L_n, L_0\right) = n\, b_\gamma(L_n) = 0$ for $n \neq 0$. Choosing $\gamma = \alpha/2$ removes the α term, so every class is represented by a multiple of $h(m) = m^3$, and equally by a multiple of Equation (A.714), since $(m^3 - m)/12$ is such a combination. Thus $\dim H^2(\mathfrak{w}, F) \leq 1$.

Step 5: ω is not a coboundary. Every coboundary satisfies $(\partial b)\left(L_m, L_{-m}\right) = b\left([L_m, L_{-m}]\right) = 2m\, b\left(L_0\right)$, a linear function of m. If $\omega = \partial b$ then $(m^3 - m)/12 = 2m\, b(L_0)$

for every m; at $m = 2$ this gives $b(L_0) = \frac{1}{8}$ and at $m = 3$ it gives $b(L_0) = \frac{1}{3}$, a contradiction. Hence $\dim H^2(\mathfrak{w}, F) = 1$ and ω generates it.

The last assertion is Proposition 14.76: extensions up to equivalence are in bijection with classes, the class of $C\,\omega$ corresponds to the bracket Equation (A.715) through Equation (14.120), and by Equation (14.123) the extension is trivial precisely for $C = 0$. □

Remark A.423 (Why the normalization $(m^3 - m)/12$). Step 4 shows that the representative of a class is fixed only up to a multiple of $h(m) = m$, so the choice Equation (A.714) is a convention and not a result. It is the convention that makes ω vanish on the three-dimensional subalgebra spanned by L_{-1}, L_0, L_1: those are the values $m \in \{-1, 0, 1\}$, at which $m^3 - m = 0$. That subalgebra is closed — $[L_1, L_{-1}] = 2L_0$, $[L_0, L_{\pm 1}] = \mp L_{\pm 1}$ — and is isomorphic to $\mathfrak{sl}(2, F)$, which by Proposition 14.77 admits no nontrivial central extension at all; the normalization therefore makes the restriction of the extension to it not merely trivial in cohomology but zero on the nose. The factor 12 is chosen so that the coefficient C in Equation (A.715) takes the value 1 for the smallest extension usually singled out; nothing in the mathematics depends on it.

A.39.2 The Kac–Moody cocycle

Definition A.424 (Loop algebra and residue)**.** Let $\mathfrak{g}$ be a finite-dimensional simple Lie algebra over $\mathbb{C}$ with Killing form κ (Definition 14.10). The *loop algebra* is

$$L\mathfrak{g} = \mathfrak{g} \otimes \mathbb{C}\left[t, t^{-1}\right], \qquad [x \otimes f, y \otimes g] = [x, y] \otimes fg\,, \tag{A.721}$$

which is a Lie algebra because the bracket of $\mathfrak{g}$ is and $\mathbb{C}[t, t^{-1}]$ is commutative and associative. For a Laurent polynomial $f = \sum_n a_n t^n$ the *residue* is $\operatorname{Res} f = a_{-1}$.

Rests on Definitions 14.10 and 14.72.

Lemma A.425 (The residue of a derivative vanishes)**.** $\operatorname{Res} f' = 0$ *for every* $f \in \mathbb{C}[t, t^{-1}]$, *and consequently* $\operatorname{Res}(f'g) = -\operatorname{Res}(fg')$ *for all* f, g.

Rests on Definition A.424.

Proof. *Derives Lemma A.425.*

$(t^n)' = n\,t^{n-1}$, whose exponent is -1 only for $n = 0$, where the coefficient n vanishes; so no t^{-1} term is ever produced and $\operatorname{Res} f' = 0$ by linearity. Applying this to fg and using the Leibniz rule, $0 = \operatorname{Res}(fg)' = \operatorname{Res}(f'g) + \operatorname{Res}(fg')$. □

Theorem A.426 (The Kac–Moody cocycle)**.** *The bilinear map*

$$\boxed{\omega_\kappa\,(x \otimes f,\ y \otimes g) = \kappa(x, y)\,\operatorname{Res}(f'g)}, \qquad \textit{that is} \qquad \omega_\kappa\,(x \otimes t^m,\ y \otimes t^n) = m\,\delta_{m+n,0}\,\kappa(x, y)\,, \tag{A.722}$$

is an antisymmetric 2-cocycle on $L\mathfrak{g}$ *and is not a coboundary. The corresponding central extension is the affine Kac–Moody algebra*

$$[x \otimes t^m, y \otimes t^n] = [x, y] \otimes t^{m+n} + \ell\, m\, \delta_{m+n,0}\, \kappa(x, y)\, Z\,, \tag{A.723}$$

with $\ell \in \mathbb{C}$ *the* level *and* Z *central.*

Rests on Definition A.424 and Lemmas 14.11 and A.425.

Proof. *Derives Theorem A.426.*

The two forms agree. For $f = t^m$, $g = t^n$ one has $f'g = m\,t^{m+n-1}$, whose residue is $m\,\delta_{m+n,0}$.

Antisymmetry. κ is symmetric (Definition 14.10), so

$$\omega_\kappa\,(y \otimes g,\ x \otimes f) = \kappa(x, y)\operatorname{Res}(g'f) = -\kappa(x, y)\operatorname{Res}(f'g) = -\omega_\kappa\,(x \otimes f,\ y \otimes g)$$

by Lemma A.425.

The cocycle condition. Let $u = x \otimes f$, $v = y \otimes g$, $w = z \otimes h$. By Equation (A.721),

$$\omega_\kappa\left([u,v],w\right) = \kappa\left([x,y],z\right)\operatorname{Res}\left((fg)'h\right)\ ,$$

and likewise for the two cyclic permutations. The three Killing-form factors are equal. Indeed Lemma 14.11 reads $\kappa\left([X,Y],Z\right) = -\kappa\left(Y,[X,Z]\right)$; taking $X = y$, $Y = x$, $Z = z$ and using antisymmetry of the bracket gives $\kappa\left([x,y],z\right) = \kappa\left(x,[y,z]\right)$, and the symmetry of κ turns the right side into $\kappa\left([y,z],x\right)$; applying the same step again gives $\kappa\left([z,x],y\right)$. Write $\mathcal{K}$ for the common value. The cyclic sum is therefore

$$\mathcal{K}\operatorname{Res}\left[(fg)'h + (gh)'f + (hf)'g\right]\ . \tag{A.724}$$

Expanding each derivative by the Leibniz rule,

$$(fg)'h + (gh)'f + (hf)'g = (f'gh + fg'h) + (fg'h + fgh') + (fgh' + f'gh) = 2\,(fgh)'\ ,$$

and Res of a derivative vanishes by Lemma A.425. So Equation (A.724) is zero, which is Equation (14.121).

It is not a coboundary. Since $\mathfrak{g}$ is simple, $[\mathfrak{g},\mathfrak{g}]$ is a nonzero ideal, hence all of $\mathfrak{g}$; so $\kappa \neq 0$, and there is $x \in \mathfrak{g}$ with $\kappa(x,x) \neq 0$ — were $\kappa(x,x) = 0$ for every x, the polarization identity $2\kappa(x,y) = \kappa(x+y,x+y) - \kappa(x,x) - \kappa(y,y)$ would make κ vanish identically, contradicting its nondegeneracy (Theorem 14.13). Put $u = x \otimes t$ and $v = x \otimes t^{-1}$. Then $[u,v] = [x,x] \otimes 1 = 0$, so $(\partial b)\,(u,v) = b\,([u,v]) = 0$ for every linear b, whereas $\omega_\kappa(u,v) = 1 \cdot \kappa(x,x) \neq 0$ by Equation (A.722). The extension Equation (A.723) is then Equation (14.120) for the cocycle $\ell\,\omega_\kappa$, and is nontrivial for $\ell \neq 0$ by Equation (14.123). □

The structure of Equation (A.722) is worth naming: it is the product of the unique invariant pairing on $\mathfrak{g}$ with the unique invariant pairing on $\mathbb{C}[t,t^{-1}]$ that is antisymmetric, namely $(f,g) \mapsto \operatorname{Res}(f'g)$. The next proposition makes the first "unique" precise and shows that the product is forced.

Lemma A.427 (Invariant bilinear forms on a simple algebra)**.** *Let $\mathfrak{g}$ be finite-dimensional and simple over $\mathbb{C}$. Every bilinear $B : \mathfrak{g} \times \mathfrak{g} \to \mathbb{C}$ with*

$$B\left([z,x],y\right) + B\left(x,[z,y]\right) = 0 \qquad \text{for all } x,y,z \tag{A.725}$$

is a multiple of the Killing form; in particular it is symmetric.

Rests on Lemma 14.11 and Theorems 5.153 and 14.13.

Proof. *Derives Lemma A.427.*

Give $\mathfrak{g}^*$ the coadjoint action $(z \cdot \phi)\,(y) = -\phi\,([z,y])$. For a bilinear B satisfying Equation (A.725) the map $\Phi_B : \mathfrak{g} \to \mathfrak{g}^*$, $\Phi_B(x) = B(x,\cdot)$, obeys

$$\Phi_B\left([z,x]\right)(y) = B\left([z,x],y\right) = -B\left(x,[z,y]\right) = \left(z \cdot \Phi_B(x)\right)(y)\,,$$

so Φ_B intertwines the adjoint and coadjoint actions. The Killing form satisfies Equation (A.725) by Lemma 14.11 and is nondegenerate by Theorem 14.13, so Φ_κ is an isomorphism of $\mathfrak{g}$-modules. Hence $T = \Phi_\kappa^{-1} \circ \Phi_B$ is an endomorphism of $\mathfrak{g}$ commuting with every ad_z. The adjoint representation of a simple algebra on the finite-dimensional complex space $\mathfrak{g}$ is irreducible — an invariant subspace is an ideal — so Schur's lemma, in the form of Theorem 5.153 (whose proof uses only that T commutes with an irreducible family of operators on a finite-dimensional complex space), gives $T = \lambda\,\mathbb{1}$ and therefore $B = \lambda\,\kappa$. □

Proposition A.428 (Uniqueness in degree zero)**.** *Every 2-cocycle c on $L\mathfrak{g}$ that is homogeneous of degree zero — meaning $c\left(x \otimes t^m, y \otimes t^n\right) = 0$ whenever $m + n \neq 0$ — is cohomologous to a multiple of ω_κ.*

Rests on Theorem A.426, Lemma A.427, and Proposition 14.77.

Proof. *Derives Proposition A.428.*

Write $B_m(x,y) = c\,(x \otimes t^m, y \otimes t^{-m})$. Antisymmetry of c gives

$$B_{-m}(u,v) = -B_m(v,u)\,, \tag{A.726}$$

so B_0 is an antisymmetric bilinear form on $\mathfrak{g} \otimes 1 \cong \mathfrak{g}$.

Step 1: kill B_0. The restriction of a cocycle to a subalgebra is a cocycle, so B_0 is a 2-cocycle on $\mathfrak{g}$; by Proposition 14.77, $H^2(\mathfrak{g},\mathbb{C}) = 0$, so $B_0 = \partial\beta$ for a linear β on $\mathfrak{g}$. Extend β to b on $L\mathfrak{g}$ by $b(x \otimes t^0) = \beta(x)$ and $b(x \otimes t^n) = 0$ for $n \neq 0$. Then $(\partial b)\,(x \otimes t^m, y \otimes t^n) = \delta_{m+n,0}\,\beta\,([x,y])$, which is homogeneous of degree zero, so $c - \partial b$ is still homogeneous of degree zero and now has $B_0 = 0$. Replace c by it.

Step 2: each B_m is invariant. Apply Equation (14.121) to $(z \otimes t^0,\ x \otimes t^m,\ y \otimes t^{-m})$:

$$B_m\,([z,x]\,,y) + B_0\,([x,y]\,,z) + B_{-m}\,([y,z]\,,x) = 0\,,$$

and $B_0 = 0$ while $B_{-m}\,([y,z]\,,x) = -B_m\,(x,[y,z]) = B_m\,(x,[z,y])$ by Equation (A.726). What is left is exactly Equation (A.725), so $B_m = \lambda_m\,\kappa$ by Lemma A.427. Feeding that back into Equation (A.726) and using the symmetry of κ,

$$\lambda_{-m} = -\lambda_m\,, \qquad \lambda_0 = 0\,. \tag{A.727}$$

Step 3: λ is additive. Apply Equation (14.121) to $\left(x \otimes t^a,\ y \otimes t^b,\ z \otimes t^e\right)$ with $a+b+e = 0$. The three terms pair $t^{a+b} = t^{-e}$ with t^e, and so on, giving

$$\lambda_{-e}\,\kappa\,([x,y]\,,z) + \lambda_{-a}\,\kappa\,([y,z]\,,x) + \lambda_{-b}\,\kappa\,([z,x]\,,y) = (\lambda_{-a} + \lambda_{-b} + \lambda_{-e})\,\mathcal{K} = 0\,,$$

the three Killing factors being the common value $\mathcal{K}$ computed in the proof of Theorem A.426. Some choice of x, y, z makes $\mathcal{K} \neq 0$: $\mathfrak{g} = [\mathfrak{g},\mathfrak{g}]$ for a simple algebra, so a nonzero w with $\kappa(w,z) \neq 0$ — available by nondegeneracy — is a sum of brackets, one of which must pair nontrivially with z. Hence $\lambda_p + \lambda_q + \lambda_r = 0$ whenever $p+q+r = 0$; putting $r = -p-q$ and using Equation (A.727), $\lambda_{p+q} = \lambda_p + \lambda_q$, so $\lambda_m = m\,\lambda_1$.

Therefore $c\,(x \otimes t^m, y \otimes t^n) = \lambda_1\,m\,\delta_{m+n,0}\,\kappa(x,y)$, which is $\lambda_1\omega_\kappa$ by Equation (A.722). □

Remark A.429 (The reduction to homogeneous cocycles, and what is left open). The hypothesis of Proposition A.428 is less restrictive than it looks, and it is worth saying how much less. Define, for a cocycle c and $k \in \mathbb{Z}$, the component c_k by $c_k\,(x \otimes t^m, y \otimes t^n) = c\,(x \otimes t^m, y \otimes t^n)$ when $m + n = k$ and 0 otherwise. Each c_k is bilinear and antisymmetric, and each is separately a cocycle: every term of Equation (14.121) evaluated on $\left(x \otimes t^a, y \otimes t^b, z \otimes t^e\right)$ pairs two factors whose exponents sum to $a + b + e$, so the cocycle condition never mixes degrees. Since only one component is nonzero on any given pair of basis elements, $c = \sum_k c_k$ makes sense with no convergence question. So the classification of cocycles reduces exactly to the classification of homogeneous ones, and Proposition A.428 settles the degree-zero part.

What is *not* proved here is that a homogeneous cocycle of nonzero degree is a coboundary — equivalently, that $\dim H^2(L\mathfrak{g},\mathbb{C}) = 1$. That is a theorem of Garland (*The arithmetic theory of loop groups*, Publications mathématiques de l'IHÉS **52**, 1980, 5–136), reproved in several places since. Its proof at degree $k \neq 0$ turns on the nonvanishing of the third cohomology of a simple Lie algebra, which this treatise does not develop; the present appendix therefore establishes that the level is *a* central charge and the only one that survives the natural grading, not that it is the only one whatever.

A.39.3 The Schwinger term

The third case is different in kind, and the difference is not a defect of the exposition. A current algebra is a bracket relation between operator-valued distributions at equal times. The naive canonical evaluation of that bracket multiplies two distributions at one point, which is not an operation, so the "formal" current algebra is not a Lie algebra whose cohomology one can compute; the question is instead whether the central term that any legitimate definition produces can be made to vanish. Schwinger's answer is that it cannot. The argument below is his; it is stated in J. Schwinger, *Field theory commutators*, Physical Review Letters **3** (1959), 296–297, a work for which this treatise carries no bibliography entry, so the attribution is made here in prose and the reader is warned that it is uncited.

Throughout this subsection the setting is $3+1$ dimensions, in accordance with rule 7: the argument is dimension-independent, but it is instantiated where the evidence is.

Definition A.430 (The hypotheses)**.** Assume a quantum theory on a Hilbert space $\mathcal{H}$ with

1. a self-adjoint Hamiltonian $\hat{H}$ bounded below, with a normalizable ground state $|0\rangle$, $\hat{H}\,|0\rangle = E_0\,|0\rangle$, and a complete orthonormal family of eigenstates $\{|n\rangle\}$ with $\hat{H}\,|n\rangle = E_n\,|n\rangle$ and $E_n \geq E_0$;
2. a conserved current: Hermitian operator-valued distributions $\hat{\jmath}^0(\boldsymbol{x},t)$ (charge density, of SI unit $\mathrm{C/m^3}$) and $\hat{\jmath}^k(\boldsymbol{x},t)$ (current density, $\mathrm{C/(m^2\,s)}$), satisfying the operator continuity equation
$$\frac{\partial \hat{\jmath}^0}{\partial t} + \partial_k \hat{\jmath}^k = 0\,. \tag{A.728}$$

For a real smooth compactly supported f on $\mathbb{R}^3$, dimensionless, write
$$\hat{A}_f = \int \mathrm{d}^3x\; f(\boldsymbol{x})\,\hat{\jmath}^0(\boldsymbol{x},0)\,, \tag{A.729}$$
a Hermitian operator of SI unit C.

Rests on Definition 14.72.

Lemma A.431 (Positivity of the double commutator)**.** *Under Definition A.430,*
$$\langle 0|\left[\hat{A}_f,\left[\hat{H},\hat{A}_f\right]\right]|0\rangle = 2\sum_n (E_n - E_0)\left|\langle n|\,\hat{A}_f\,|0\rangle\right|^2 \;\geq\; 0\,, \tag{A.730}$$
with equality if and only if $\hat{A}_f\,|0\rangle$ lies in the ground eigenspace of $\hat{H}$.

Rests on Definition A.430.

Proof. *Derives Lemma A.431.*

Abbreviate $\hat{A} = \hat{A}_f$. Expanding the nested commutator,
$$\left[\hat{A},\left[\hat{H},\hat{A}\right]\right] = \hat{A}\hat{H}\hat{A} - \hat{A}^2\hat{H} - \hat{H}\hat{A}^2 + \hat{A}\hat{H}\hat{A} = 2\hat{A}\hat{H}\hat{A} - \hat{A}^2\hat{H} - \hat{H}\hat{A}^2\,.$$
Taking the ground-state expectation and using $\hat{H}\,|0\rangle = E_0\,|0\rangle$ on the last two terms,
$$\langle 0|\left[\hat{A},\left[\hat{H},\hat{A}\right]\right]|0\rangle = 2\,\langle 0|\,\hat{A}\hat{H}\hat{A}\,|0\rangle - 2E_0\,\langle 0|\,\hat{A}^2\,|0\rangle\;.$$
Inserting the resolution of the identity $\sum_n |n\rangle\langle n|$ between the operators and using $\hat{A}^\dagger = \hat{A}$, which gives $\langle 0|\,\hat{A}\,|n\rangle = \overline{\langle n|\,\hat{A}\,|0\rangle}$,
$$\langle 0|\left[\hat{A},\left[\hat{H},\hat{A}\right]\right]|0\rangle = 2\sum_n (E_n - E_0)\left|\langle n|\,\hat{A}\,|0\rangle\right|^2\,,$$
which is Equation (A.730). Every summand is nonnegative because $E_n \geq E_0$, so the sum vanishes only if $\langle n|\,\hat{A}\,|0\rangle = 0$ for every n with $E_n > E_0$, that is only if $\hat{A}\,|0\rangle$ has no component outside the ground eigenspace. $\square$

Theorem A.432 (Schwinger's theorem)**.** *Under Definition A.430, suppose the equal-time commutator of the charge density with the current density has the vacuum expectation*

$$\langle 0|\left[\hat{\jmath}^0(\boldsymbol{y},0),\hat{\jmath}^k(\boldsymbol{x},0)\right]|0\rangle = \mathrm{i}\,S\,\partial_k^{(\boldsymbol{x})}\,\delta^3(\boldsymbol{x}-\boldsymbol{y})\ , \tag{A.731}$$

with S a constant (the Schwinger term*), of SI unit $\mathrm{C}^2/(\mathrm{m\,s})$. Then*

$$\hbar\,S\int \mathrm{d}^3x\ |\nabla f|^2 = 2\sum_n (E_n-E_0)\left|\langle n|\,\hat{A}_f\,|0\rangle\right|^2\ \geq\ 0 \tag{A.732}$$

for every test function f. Hence $S\geq 0$, and $S=0$ forces $\hat{\jmath}^0(\boldsymbol{x},0)\,|0\rangle$ to lie in the ground eigenspace for every $\boldsymbol{x}$ — that is, the charge density must have no vacuum fluctuations at all. In any theory in which it does, the equal-time commutator cannot vanish.

Rests on Lemma A.431 and Definition A.430.

Proof. *Derives Theorem A.432.*

The Heisenberg equation $\mathrm{i}\hbar\,\partial_t\hat{O}=\left[\hat{O},\hat{H}\right]$ and the continuity equation Equation (A.728) give

$$\left[\hat{H},\hat{\jmath}^0(\boldsymbol{x},0)\right] = -\mathrm{i}\hbar\,\frac{\partial\hat{\jmath}^0}{\partial t}(\boldsymbol{x},0) = \mathrm{i}\hbar\,\partial_k\hat{\jmath}^k(\boldsymbol{x},0)\,,$$

so that, integrating by parts against the compactly supported f,

$$\left[\hat{H},\hat{A}_f\right] = \mathrm{i}\hbar\int \mathrm{d}^3x\ f(\boldsymbol{x})\,\partial_k\hat{\jmath}^k(\boldsymbol{x},0) = -\mathrm{i}\hbar\int \mathrm{d}^3x\ \left(\partial_k f\right)(\boldsymbol{x})\,\hat{\jmath}^k(\boldsymbol{x},0)\,. \tag{A.733}$$

Therefore

$$\left[\hat{A}_f,\left[\hat{H},\hat{A}_f\right]\right] = -\mathrm{i}\hbar\int \mathrm{d}^3y\int \mathrm{d}^3x\ f(\boldsymbol{y})\left(\partial_k f\right)(\boldsymbol{x})\ \left[\hat{\jmath}^0(\boldsymbol{y},0),\hat{\jmath}^k(\boldsymbol{x},0)\right]\ ,$$

whose ground-state expectation, by Equation (A.731) and $\int \mathrm{d}^3y\,f(\boldsymbol{y})\,\partial_k^{(\boldsymbol{x})}\delta^3(\boldsymbol{x}-\boldsymbol{y}) = \left(\partial_k f\right)(\boldsymbol{x})$, is

$$-\mathrm{i}\hbar\int \mathrm{d}^3x\ \left(\partial_k f\right)(\boldsymbol{x})\ \mathrm{i}S\ \left(\partial_k f\right)(\boldsymbol{x}) = \hbar S\int \mathrm{d}^3x\ |\nabla f|^2\ ,$$

the two imaginary units combining to $-\mathrm{i}\cdot\mathrm{i}=1$. Equating this with Equation (A.730) gives Equation (A.732). Since $\int|\nabla f|^2>0$ for any nonconstant f, the sign of S follows; and $S=0$ makes the right-hand side vanish for every f, so by the equality clause of Lemma A.431 the vector $\hat{A}_f\,|0\rangle$ lies in the ground eigenspace for every f, which is the stated conclusion.

The dimensions check: $\hat{A}_f$ carries C, so the right-hand side of Equation (A.732) carries $\mathrm{J}\times\mathrm{C}^2$; on the left, $\hbar S\int|\nabla f|^2\,\mathrm{d}^3x$ carries $\mathrm{J\,s}\times\mathrm{C}^2/(\mathrm{m\,s})\times\mathrm{m}$, the same. □

Remark A.433 (The extension is central, and it is not removable)*.* Two features of Equation (A.731) put it in the frame of Section 14.4. First, the right-hand side is a *c*-number: it commutes with every operator of the theory, so as an element of the algebra generated by the smeared densities it is central, exactly as the Z of Equation (14.120). Second, it cannot be removed by redefining the generators. The would-be redefinition is a shift of each smeared density by a constant, $\hat{A}_f\mapsto\hat{A}_f+b(f)$, which changes the commutator by b of a bracket — the coboundary of Equation (14.122) — and the naive algebra has $\left[\hat{\jmath}^0,\hat{\jmath}^k\right]=0$, so that bracket is zero and no shift produces anything. The situation is formally identical to the Galilei mass of Example 14.80: a central term sitting on a pair of generators whose classical bracket vanishes, and therefore beyond the reach of any coboundary (Appendix A.38).

What Theorem A.432 adds, and what has no analogue in the two finite computations above, is that the term is *forced*. There the question was which cocycles exist; here the cocycle that appears to be zero is proved to be nonzero by an argument that never computes it — only positivity of the energy above the ground state and conservation of the current are used.

Remark A.434 (What is quoted here). Three things are imported in this section, and the three are of different weight.

(i) Whitehead's second lemma, used once, in Step 1 of Proposition A.428: $H^2(\mathfrak{g}, F) = 0$ for a finite-dimensional semisimple $\mathfrak{g}$ in characteristic zero. It is Proposition 14.77 of the chapter, whose own derivation is recorded there and not repeated here. It is used only to remove the degree-zero-in-t piece B_0, and nothing else in this appendix depends on it.

(ii) Garland's theorem, that $\dim H^2(L\mathfrak{g}, \mathbb{C}) = 1$ for simple $\mathfrak{g}$, which would upgrade Proposition A.428 from the degree-zero statement to the full one. It is not proved here and it is not used: Theorem A.426 constructs the cocycle and proves it nontrivial without it, and Remark A.429 says precisely which case remains. The Virasoro uniqueness, by contrast, is proved in full — Theorem A.422 imports nothing.

(iii) The framework in which Theorem A.432 is stated. This is the substantial one, and it is an assumption rather than a theorem. It is assumed that the charge and current densities are operator-valued distributions on a common dense domain, smearable as in Equation (A.729), and that the equal-time commutator of two of them is supported on the diagonal $\boldsymbol{x} = \boldsymbol{y}$, so that it is a finite sum of derivatives of $\delta^3(\boldsymbol{x} - \boldsymbol{y})$ with operator coefficients; the leading such term with a c-number coefficient is what Equation (A.731) isolates. That framework is the Wightman axiomatization [1297], which this treatise uses but does not construct. The spectral hypothesis of Definition A.430 — a self-adjoint Hamiltonian bounded below with a complete family of eigenstates — is likewise the standard spectral theorem for a self-adjoint operator [1481], applied in the form in which the sum over n is legitimate.

What is *proved* above, given that framework, is the inequality Equation (A.732) and its two consequences: $S \geq 0$, and $S = 0$ only in the degenerate case where the charge density does not fluctuate in the ground state. What is *not* proved, and is not attempted, is the evaluation of S in any particular theory. That evaluation requires a regularization — a point splitting of the product of two field operators, followed by the removal of the splitting — and it is a computation in quantum field theory, not in Lie algebra cohomology. This appendix therefore establishes the statement made in Example 14.81 in the exact form "a formally central extension survives regularization": the extension is central by Remark A.433, and it survives because Theorem A.432 forbids the value zero that the unregularized computation returns.

Finally, the evidential status of all three algebras is settled in Remark 14.82 and is not reopened here. Writing down a consistent central extension is a mathematical act; that remark records what is and is not measured, in particular that the Schwinger term reaches experiment only through the anomalies [1388, 1391] and their measured consequences [2361], and that this treatise records no measurement of a Virasoro central charge.

Remark A.435. Theorems A.422, A.426 and A.432 discharge the three obligations recorded after Example 14.81 of Chapter 14. Read together with Example 14.79 — where every antisymmetric form on an abelian algebra is a cocycle and none is a coboundary — and with Appendix A.38, they fill in Remark 14.78's list of the four places a central charge can live: abelian algebras, inhomogeneous algebras and their contractions, and infinite-dimensional algebras. The last is the case treated here, and the three instances differ instructively. The Witt algebra has a one-dimensional H^2 proved outright; the loop algebra has a distinguished class built from the Killing form and the residue, whose uniqueness is settled here only within the natural grading; and the current algebra has an extension that no cohomological computation produces, because the bracket it extends does not exist until the theory is regularized.

A.40 The Strengthened Jacobi Condition and Conjugate Points

This appendix proves the equivalence asserted in Remark 16.57 of Chapter 16: for a regular problem, the hypothesis of Theorem 16.56 — the existence of a solution of the Jacobi accessory equation Equation (16.49) with no zero on the closed interval $[a, b]$ — holds if and only if no point of $(a, b]$ is conjugate to a in the sense of Definition 16.53. That equivalence is what joins Jacobi's necessary condition (Theorem 16.55) to his sufficient one, and without it the two theorems test different hypotheses.

Three ingredients are needed and all three are proved here: global existence and uniqueness for the accessory equation, which is Corollary 9.9 of Chapter 9 once the equation is written as a first-order system; the continuous dependence of a solution on the point at which its initial data are posed, which is proved from the integral equation Equation (9.7) underlying the Picard–Lindelöf theorem Theorem 9.8 together with a Grönwall estimate; and the Sturm separation theorem, which is proved from the Lagrange identity Lemma 9.56. The notation is that of Section 16.4.2 throughout: P and Q are the coefficients Equation (16.46) evaluated along the extremal, the problem is *regular*, meaning $P > 0$ on $[a, b]$, and the accessory equation is

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(P(x)\,\frac{\mathrm{d}u}{\mathrm{d}x}\right) - Q(x)\,u = 0\,. \tag{A.734}$$

Only the continuity of P and Q and the positivity of P are used; in particular P is never differentiated, which is why the results below hold for every C^3 integrand and C^2 extremal without a further regularity hypothesis.

A.40.1 Statement

Theorem A.436 (Strengthened Jacobi condition and conjugate points)**.** *Let $P, Q \in C^0[a, b]$ with $P > 0$ on $[a, b]$, and let u_a denote the solution of Equation (A.734) determined by*

$$u_a(a) = 0\,, \qquad \left(P\,\frac{\mathrm{d}u_a}{\mathrm{d}x}\right)(a) = 1\,. \tag{A.735}$$

The following three statements are equivalent.

1. *There exists a solution w of Equation (A.734) with $w(x) \neq 0$ for every $x \in [a, b]$* (the strengthened Jacobi condition)*.*
2. *No point of $(a, b]$ is conjugate to a* (Definition 16.53)*.*
3. *$u_a > 0$ on $(a, b]$.*

Rests on Definition 16.53 and Equation (16.49).

Remark A.437 (The two normalisations agree). Definition 16.53 normalises the Jacobi solution by $u'(a) = 1$, whereas Equation (A.735) normalises the quantity $P\,\mathrm{d}u/\mathrm{d}x$, which is the natural momentum variable of Equation (A.734). The two differ by the positive factor $P(a)$, and a nonzero constant multiple of a solution has exactly the same zeros, so the set of points conjugate to a is the same under either normalisation. The momentum normalisation is used below because P is not assumed differentiable and $P\,\mathrm{d}u/\mathrm{d}x$, unlike $\mathrm{d}u/\mathrm{d}x$ alone, is a component of the first-order system Equation (A.736).

The proof occupies the rest of the section: the system form and the elementary structure of the zeros (Appendix A.40.2), continuous dependence on the initial point (Appendix A.40.3), the Sturm separation theorem (Appendix A.40.4), and the equivalence itself (Appendix A.40.5), after which Appendix A.40.6 recovers the exact phrasing used in the chapter, in which the initial point is displaced to the left of a.

A.40.2 The accessory equation as a first-order system

Definition A.438 (Momentum variable and the accessory system)**.** For a solution u of Equation (A.734) put $v = P\,\mathrm{d}u/\mathrm{d}x$ and $\boldsymbol{Y} = (u, v)$. Then $\boldsymbol{Y}$ solves the linear system

$$\frac{\mathrm{d}\boldsymbol{Y}}{\mathrm{d}x} = A(x)\,\boldsymbol{Y}\,, \qquad A(x) = \begin{pmatrix} 0 & 1/P(x) \\ Q(x) & 0 \end{pmatrix}, \tag{A.736}$$

whose coefficient matrix is continuous on $[a, b]$ because P and Q are continuous and P does not vanish. Conversely, if $\boldsymbol{Y} = (u, v)$ solves Equation (A.736) then u is of class C^1 with $P\,\mathrm{d}u/\mathrm{d}x = v$ of class C^1, and $\mathrm{d}\left(P\,\mathrm{d}u/\mathrm{d}x\right)/\mathrm{d}x = Qu$, which is Equation (A.734).

Rests on Equations (9.3) and (16.49).

Lemma A.439 (Existence, uniqueness, and the structure of the zeros)**.** *For every $s \in [a, b]$ and every $\boldsymbol{Y}_0 \in \mathbb{R}^2$ the system Equation (A.736) has exactly one solution on the whole of $[a, b]$ with $\boldsymbol{Y}(s) = \boldsymbol{Y}_0$. Consequently, for a solution u of Equation (A.734) that does not vanish identically:*

1. *at every zero x_0 of u one has $\left(P\,\mathrm{d}u/\mathrm{d}x\right)(x_0) \neq 0$, so the zero is simple;*
2. *the zeros of u in $[a, b]$ are isolated, and there are finitely many of them.*

Rests on Definition A.438 and Corollary 9.9.

Proof. *Derives Lemma A.439.*

The system Equation (A.736) is linear with a continuous coefficient matrix on the compact interval $[a, b]$, so Corollary 9.9 applies verbatim with $\boldsymbol{g} = \boldsymbol{0}$: there is exactly one solution on all of $[a, b]$ for each choice of s and $\boldsymbol{Y}_0$.

(1) If $u(x_0) = 0$ and $\left(P\,\mathrm{d}u/\mathrm{d}x\right)(x_0) = 0$ then $\boldsymbol{Y}(x_0) = \boldsymbol{0}$; but $\boldsymbol{Y} \equiv \boldsymbol{0}$ is a solution with those data, so uniqueness forces $\boldsymbol{Y} \equiv \boldsymbol{0}$ and hence $u \equiv 0$, which was excluded.

(2) Let x_0 be a zero. By (1) and $P > 0$ we have $\left(\mathrm{d}u/\mathrm{d}x\right)(x_0) \neq 0$, so by the definition of the derivative there is a punctured neighbourhood of x_0 in which $u(x)/(x - x_0)$ keeps the sign of $\left(\mathrm{d}u/\mathrm{d}x\right)(x_0)$ and in particular $u(x) \neq 0$: the zero is isolated. If there were infinitely many zeros in $[a, b]$, Bolzano–Weierstrass (Theorem 7.7) would give a sequence of distinct zeros converging to some $x_* \in [a, b]$; continuity of u makes x_* a zero as well, and it is not isolated — a contradiction. □

A.40.3 Continuous dependence on the initial point

The estimate below is the one place where a quantitative statement is needed rather than a qualitative one, and it is obtained from the integral form Equation (9.7) of the initial value problem, which is where the proof of Theorem 9.8 begins. Throughout, $\|\cdot\|$ is any fixed norm on $\mathbb{R}^2$ and $\|A(x)\|$ the induced operator norm, so that $\|A(x)\boldsymbol{Z}\| \le \|A(x)\|\,\|\boldsymbol{Z}\|$; put

$$L = \max_{x \in [a,b]} \|A(x)\|\,, \tag{A.737}$$

which is finite because A is continuous on a compact interval (Theorem 7.24).

Lemma A.440 (Grönwall's inequality)**.** *Let $\varphi \in C^0[a, b]$ be nonnegative, let $x_0 \in [a, b]$, and let $C \ge 0$ and $L \ge 0$ satisfy*

$$\varphi(x) \le C + L\left|\int_{x_0}^{x} \varphi(t)\,\mathrm{d}t\right| \qquad \text{for every } x \in [a, b]\,. \tag{A.738}$$

Then $\varphi(x) \le C\,\mathrm{e}^{L|x - x_0|}$ on $[a, b]$.

Rests on Theorems 7.35 and 7.42.

Proof. *Derives Lemma A.440.*

Take first $x \geq x_0$ and set $\Phi(x) = C + L\int_{x_0}^{x} \varphi(t)\,\mathrm{d}t$, which is of class C^1 with $\Phi' = L\varphi$ by the fundamental theorem of calculus (Theorem 7.42). Hypothesis Equation (A.738) reads $\varphi \leq \Phi$ on $[x_0, b]$, so $\Phi' = L\varphi \leq L\Phi$ there, and therefore

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\Phi(x)\,\mathrm{e}^{-L(x-x_0)}\right] = \left[\Phi'(x) - L\,\Phi(x)\right]\mathrm{e}^{-L(x-x_0)} \leq 0\,.$$

By the mean value theorem (Theorem 7.35), a function whose derivative is nowhere positive on an interval satisfies $g(x) - g\,(x_0) = g'(\xi)\,(x - x_0) \leq 0$ for some interior ξ whenever $x > x_0$; applied to $g(x) = \Phi(x)\,\mathrm{e}^{-L(x-x_0)}$ this gives $\Phi(x)\,\mathrm{e}^{-L(x-x_0)} \leq \Phi(x_0) = C$, that is $\varphi(x) \leq \Phi(x) \leq C\,\mathrm{e}^{L(x-x_0)}$.

For $x \leq x_0$ apply what has just been proved to $\tilde{\varphi}(t) = \varphi\,(x_0 - t)$ on $[0, x_0 - a]$, which satisfies the same hypothesis with the same constants because the substitution $t \longmapsto x_0 - t$ turns $\left|\int_{x_0}^{x} \varphi\right|$ into $\left|\int_{0}^{x_0 - x} \tilde{\varphi}\right|$. □

Theorem A.441 (Continuous dependence on the data and on the initial point)**.** *For $s \in [a, b]$ and $\boldsymbol{Y}_0 \in \mathbb{R}^2$ write $\boldsymbol{Y}\,(\cdot\,; s, \boldsymbol{Y}_0)$ for the solution of Equation (A.736) with $\boldsymbol{Y}(s) = \boldsymbol{Y}_0$. Then, with L as in Equation (A.737), for all $s, \tilde{s} \in [a, b]$ and all $\boldsymbol{Y}_0, \tilde{\boldsymbol{Y}}_0$,*

$$\max_{x\in[a,b]} \left\|\boldsymbol{Y}\,(x; s, \boldsymbol{Y}_0) - \boldsymbol{Y}\left(x; \tilde{s}, \tilde{\boldsymbol{Y}}_0\right)\right\| \leq \left(\left\|\boldsymbol{Y}_0 - \tilde{\boldsymbol{Y}}_0\right\| + L\,\mathrm{e}^{L(b-a)}\left\|\tilde{\boldsymbol{Y}}_0\right\|\,|s - \tilde{s}|\right)\mathrm{e}^{L(b-a)}\,. \tag{A.739}$$

In particular the map $(x, s, \boldsymbol{Y}_0) \longmapsto \boldsymbol{Y}\,(x; s, \boldsymbol{Y}_0)$ is continuous on $[a, b] \times [a, b] \times \mathbb{R}^2$.

Rests on Lemmas A.439 and A.440 and Equation (9.7).

Proof. *Derives Theorem A.441.*

Write $\boldsymbol{Y} = \boldsymbol{Y}\,(\cdot\,; s, \boldsymbol{Y}_0)$ and $\boldsymbol{Z} = \boldsymbol{Y}\left(\cdot\,; \tilde{s}, \tilde{\boldsymbol{Y}}_0\right)$. Both are continuous and, by the equivalence of the initial value problem with its integral form Equation (9.7) — which is the first step of the proof of Theorem 9.8 and uses only the fundamental theorem of calculus —

$$\boldsymbol{Y}(x) = \boldsymbol{Y}_0 + \int_{s}^{x} A(t)\,\boldsymbol{Y}(t)\,\mathrm{d}t\,, \qquad \boldsymbol{Z}(x) = \tilde{\boldsymbol{Y}}_0 + \int_{\tilde{s}}^{x} A(t)\,\boldsymbol{Z}(t)\,\mathrm{d}t\,. \tag{A.740}$$

A bound for $\boldsymbol{Z}$ alone. Taking norms in the second identity of Equation (A.740),

$$\|\boldsymbol{Z}(x)\| \leq \left\|\tilde{\boldsymbol{Y}}_0\right\| + L\left|\int_{\tilde{s}}^{x} \|\boldsymbol{Z}(t)\|\,\mathrm{d}t\right|\,,$$

so Lemma A.440 with $\varphi = \|\boldsymbol{Z}\|$, $C = \left\|\tilde{\boldsymbol{Y}}_0\right\|$ and $x_0 = \tilde{s}$ gives

$$\max_{x\in[a,b]} \|\boldsymbol{Z}(x)\| \leq \left\|\tilde{\boldsymbol{Y}}_0\right\|\,\mathrm{e}^{L(b-a)}\,. \tag{A.741}$$

The difference. Subtracting the two identities of Equation (A.740) and splitting the second integral at s,

$$\boldsymbol{Y}(x) - \boldsymbol{Z}(x) = \left(\boldsymbol{Y}_0 - \tilde{\boldsymbol{Y}}_0\right) + \int_{s}^{x} A(t)\,[\boldsymbol{Y}(t) - \boldsymbol{Z}(t)]\,\mathrm{d}t - \int_{\tilde{s}}^{s} A(t)\,\boldsymbol{Z}(t)\,\mathrm{d}t\,.$$

The last term is bounded in norm by $L\,|s - \tilde{s}|\,\max_{[a,b]}\|\boldsymbol{Z}\|$, which Equation (A.741) bounds by $L\,\mathrm{e}^{L(b-a)}\left\|\tilde{\boldsymbol{Y}}_0\right\|\,|s - \tilde{s}|$. Hence, writing $\varphi = \|\boldsymbol{Y} - \boldsymbol{Z}\|$ and $C = \left\|\boldsymbol{Y}_0 - \tilde{\boldsymbol{Y}}_0\right\| + L\,\mathrm{e}^{L(b-a)}\left\|\tilde{\boldsymbol{Y}}_0\right\|\,|s - \tilde{s}|$,

$$\varphi(x) \leq C + L\left|\int_{s}^{x} \varphi(t)\,\mathrm{d}t\right|\,,$$

and Lemma A.440 gives Equation (A.739).

For the final assertion, fix $(x_1, s_1, \boldsymbol{Y}_1)$. The estimate Equation (A.739) shows that $\boldsymbol{Y}(\cdot\,; s, \boldsymbol{Y}_0)$ converges *uniformly in* x to $\boldsymbol{Y}(\cdot\,; s_1, \boldsymbol{Y}_1)$ as $(s, \boldsymbol{Y}_0) \longrightarrow (s_1, \boldsymbol{Y}_1)$, and each solution is continuous in x; a uniform limit of continuous functions is continuous, and the triangle inequality

$$\begin{aligned}\|\boldsymbol{Y}(x; s, \boldsymbol{Y}_0) - \boldsymbol{Y}(x_1; s_1, \boldsymbol{Y}_1)\| \le \max_{[a,b]} \|\boldsymbol{Y}(\cdot\,; s, \boldsymbol{Y}_0) - \boldsymbol{Y}(\cdot\,; s_1, \boldsymbol{Y}_1)\| \\ + \|\boldsymbol{Y}(x; s_1, \boldsymbol{Y}_1) - \boldsymbol{Y}(x_1; s_1, \boldsymbol{Y}_1)\|\end{aligned}$$

makes both contributions small. □

A.40.4 The Sturm separation theorem

Lemma A.442 (The Wronskian of the accessory equation is constant)**.** *Let u_1, u_2 solve Equation (A.734) on $[a, b]$ and put*

$$W = P\left(u_1 \frac{\mathrm{d}u_2}{\mathrm{d}x} - u_2 \frac{\mathrm{d}u_1}{\mathrm{d}x}\right). \tag{A.742}$$

Then W is constant on $[a, b]$, and $W = 0$ if and only if u_1 and u_2 are linearly dependent.
Rests on Lemmas 9.56 and A.439.

Proof. *Derives Lemma A.442.*

Equation (A.734) is the Sturm–Liouville equation $L[u] = 0$ of Lemma 9.56 with the data $p = P$ and $q = -Q$. The Lagrange identity Equation (9.56) therefore gives

$$\frac{\mathrm{d}W}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}\left[p\left(u_1 \frac{\mathrm{d}u_2}{\mathrm{d}x} - u_2 \frac{\mathrm{d}u_1}{\mathrm{d}x}\right)\right] = u_1\, L[u_2] - u_2\, L[u_1] = 0\,,$$

so W is constant. In the notation of Definition A.438, W is the determinant of the matrix with columns $\boldsymbol{Y}_1 = (u_1, v_1)$ and $\boldsymbol{Y}_2 = (u_2, v_2)$, since $u_1 v_2 - u_2 v_1 = P(u_1 u_2' - u_2 u_1')$. If $W = 0$, the two vectors are parallel at every point; picking a point x_0 at which $\boldsymbol{Y}_1(x_0) \neq \boldsymbol{0}$ — one exists unless $u_1 \equiv 0$, in which case the dependence is trivial — gives $\boldsymbol{Y}_2(x_0) = \kappa\, \boldsymbol{Y}_1(x_0)$ for a scalar κ, and uniqueness in Lemma A.439 then forces $\boldsymbol{Y}_2 = \kappa \boldsymbol{Y}_1$ identically, hence $u_2 = \kappa u_1$. Conversely $u_2 = \kappa u_1$ makes Equation (A.742) vanish outright. □

Theorem A.443 (Sturm separation theorem)**.** *Let u_1, u_2 be linearly independent solutions of Equation (A.734) on $[a, b]$. Then between any two consecutive zeros of u_1 there lies exactly one zero of u_2, and u_1, u_2 have no zero in common. In particular the zeros of the two solutions strictly interlace.*
Rests on Lemma A.442 and Theorem 7.23.

Proof. *Derives Theorem A.443.*

By Lemma A.442 the constant $W = P(u_1 u_2' - u_2 u_1')$ is nonzero. Neither solution vanishes identically, so Lemma A.439 applies to both.

No common zero. If $u_1(x_0) = u_2(x_0) = 0$ then both terms of Equation (A.742) vanish at x_0 and $W = 0$, a contradiction.

At least one zero in between. Let $x_1 < x_2$ be consecutive zeros of u_1 in $[a, b]$, so that $u_1 \neq 0$ on (x_1, x_2). Evaluating Equation (A.742) at the two ends, where u_1 vanishes,

$$W = -P(x_1)\, u_2(x_1) \left(\frac{\mathrm{d}u_1}{\mathrm{d}x}\right)(x_1) = -P(x_2)\, u_2(x_2) \left(\frac{\mathrm{d}u_1}{\mathrm{d}x}\right)(x_2)\,. \tag{A.743}$$

Both derivatives are nonzero by Lemma A.439(1), and they have opposite signs: u_1 keeps one sign, say positive, throughout (x_1, x_2), so $(\mathrm{d}u_1/\mathrm{d}x)(x_1) = \lim_{x \to x_1^+} u_1(x)/(x - x_1) \ge 0$ and hence is > 0, while $(\mathrm{d}u_1/\mathrm{d}x)(x_2) = \lim_{x \to x_2^-} u_1(x)/(x - x_2) \le 0$ and hence is < 0; if $u_1 < 0$ in between, both signs reverse. Since $W \neq 0$ is the *same* number at both ends

of Equation (A.743) and $P > 0$, the values $u_2(x_1)$ and $u_2(x_2)$ are both nonzero and of opposite sign. The intermediate value theorem (Theorem 7.23) gives a zero of u_2 in (x_1, x_2).

At most one. Suppose u_2 had two zeros $y_1 < y_2$ in (x_1, x_2); choosing them consecutive among the zeros of u_2 — possible because by Lemma A.439(2) there are finitely many — and applying the previous paragraph with the roles of u_1 and u_2 exchanged produces a zero of u_1 in $(y_1, y_2) \subset (x_1, x_2)$, contradicting the assumption that x_1 and x_2 are consecutive zeros of u_1. □

A.40.5 Proof of the equivalence

Proof of Theorem A.436. *Derives Theorem A.436.*

The three implications are proved in the cycle (1) ⇒ (2) ⇒ (3) ⇒ (1).

(1) ⇒ (2). Let w be a solution with no zero on $[a, b]$ and suppose, for contradiction, that some $c \in (a, b]$ is conjugate to a, i.e. $u_a(c) = 0$. The solution u_a does not vanish identically, since $(P\,\mathrm{d}u_a/\mathrm{d}x)(a) = 1 \neq 0$. Moreover u_a and w are linearly independent: a multiple of w cannot vanish at a as u_a does, unless the multiple is zero. By Lemma A.439(2) the zeros of u_a in $[a, b]$ are finite in number, so among those lying in $(a, c]$ there is a smallest one, c_1; then a and c_1 are consecutive zeros of u_a. The Sturm separation theorem Theorem A.443 places a zero of w in $(a, c_1) \subset [a, b]$, contradicting the hypothesis on w.

(2) ⇒ (3). Since $(P\,\mathrm{d}u_a/\mathrm{d}x)(a) = 1$ and $P(a) > 0$, we have $(\mathrm{d}u_a/\mathrm{d}x)(a) = 1/P(a) > 0$, and with $u_a(a) = 0$ this makes $u_a(x) > 0$ for $x > a$ close enough to a. By (2), u_a has no zero in $(a, b]$; were u_a negative at some point of $(a, b]$, the intermediate value theorem (Theorem 7.23) would produce a zero between that point and one where $u_a > 0$. Hence $u_a > 0$ throughout $(a, b]$.

(3) ⇒ (1). Let z be the solution of Equation (A.734) with

$$z(a) = 1\,, \qquad \left(P\,\frac{\mathrm{d}z}{\mathrm{d}x}\right)(a) = 0\,,$$

which exists by Lemma A.439, and consider the one-parameter family of solutions

$$w_\delta = u_a + \delta\, z\,, \qquad \delta > 0\,, \tag{A.744}$$

each of which solves Equation (A.734) by linearity. We show that $w_\delta > 0$ on $[a, b]$ for all small $\delta > 0$.

Write $v_a = P\,\mathrm{d}u_a/\mathrm{d}x$, a continuous function with $v_a(a) = 1$, and recall $z(a) = 1$ with z continuous. Choose $\eta > 0$ so small that

$$v_a \geq \tfrac{1}{2} \quad \text{and} \quad z \geq \tfrac{1}{2} \qquad \text{on } [a, a+\eta] \cap [a, b]\,. \tag{A.745}$$

On that subinterval $\mathrm{d}u_a/\mathrm{d}x = v_a/P > 0$, so $u_a(x) = \int_a^x (v_a/P)\,\mathrm{d}t \geq 0$ by Theorem 7.43, and therefore

$$w_\delta(x) = u_a(x) + \delta\, z(x) \geq 0 + \tfrac{\delta}{2} > 0\,, \qquad x \in [a, a+\eta] \cap [a, b]\,.$$

If $a + \eta \geq b$ this already proves the claim for every $\delta > 0$. Otherwise, on the compact interval $[a + \eta, b]$ the continuous function u_a is strictly positive by (3), so by the extreme value theorem (Theorem 7.24) it has a minimum $m > 0$ there; let $M = \max_{[a,b]} |z|$, also finite by Theorem 7.24. Choosing any δ with $0 < \delta < m/(M+1)$ gives, for $x \in [a + \eta, b]$,

$$w_\delta(x) \geq m - \delta M > m - \frac{m\,M}{M+1} = \frac{m}{M+1} > 0\,.$$

So $w = w_\delta$ has no zero on $[a, b]$, which is (1). □

Remark A.444 (What the equivalence buys). With Theorem A.436 the two halves of the second-order theory close on each other. For a regular problem, Theorem 16.55 says that a weak minimum admits no conjugate point in the *open* interval (a, b), and Theorem 16.56 says that a solution without zeros on the *closed* interval makes $\delta^2 J$ positive definite; by statement (2) of the theorem just proved, the hypothesis of the second is exactly the absence of conjugate points in $(a, b]$. The one remaining gap between them is the single point $x = b$: an extremal whose first conjugate point is the right endpoint itself satisfies the necessary condition and fails the sufficient one, and the second variation is then positive semidefinite with the Jacobi solution u_a in its null space — which is precisely the computation carried out in the proof of Theorem 16.55, where $\delta^2 J\,[y;\eta] = 0$ for the broken variation built from u_a. The threshold case is not a defect of the theory: it is the kinetic focus, and Remark 16.74 exhibits it as the exact instant $T = \pi\sqrt{m/k}$ at which the harmonic oscillator's action stops being a minimum.

A.40.6 The displaced initial point

Remark 16.57 states the equivalence in the form in which it is usually met: the Jacobi solution vanishing slightly to the *left* of a has no zero in $[a, b]$ exactly when the solution vanishing at a has none in $(a, b]$. That form presupposes that the coefficients are defined on a slightly larger interval, which happens whenever the extremal itself extends; under that hypothesis it is a corollary of Theorems A.436 and A.441, and it is the form in which continuous dependence on the initial point does real work.

Corollary A.445 (Displacing the initial point)**.** *Suppose P and Q extend continuously to $[a - \varepsilon_0, b]$ for some $\varepsilon_0 > 0$, with $P > 0$ there, and for $s \in [a - \varepsilon_0, b)$ let u_s be the solution of Equation (A.734) with $u_s(s) = 0$ and $(P\,\mathrm{d}u_s/\mathrm{d}x)\,(s) = 1$. Then the strengthened Jacobi condition holds on $[a, b]$ if and only if there is an $\varepsilon \in (0, \varepsilon_0)$ such that $u_{a-\varepsilon}$ has no zero in $[a, b]$ — and in that case no zero in $[a, b]$ for every smaller ε as well.*

Rests on Theorems A.436 and A.441.

Proof. *Derives Corollary A.445.*

Sufficiency is immediate: $u_{a-\varepsilon}$ restricted to $[a, b]$ is a solution of Equation (A.734) without zeros there, which is statement (1) of Theorem A.436.

Necessity. Assume the strengthened Jacobi condition, hence statement (3): $u_a > 0$ on $(a, b]$. Work on the extended interval $I = [a - \varepsilon_0, b]$, on which Theorem A.441 holds with the constant L of Equation (A.737) computed over I. Write $v_s = P\,\mathrm{d}u_s/\mathrm{d}x$, so that the initial vector is $\boldsymbol{Y}_0 = (0, 1)$ for every s and only the initial *point* varies; Equation (A.739) then reads

$$\max_{x \in I} \left(|u_s(x) - u_a(x)| + |v_s(x) - v_a(x)| \right) \le L\, \mathrm{e}^{2L(b-a+\varepsilon_0)}\ |s - a| \equiv \Lambda\ |s - a|\ , \tag{A.746}$$

where the norm on $\mathbb{R}^2$ has been taken to be the sum of the moduli of the components and $\|\boldsymbol{Y}_0\| = 1$.

Since v_a is continuous on I with $v_a(a) = 1$, fix $\eta > 0$ with $\eta < \varepsilon_0$ and

$$v_a \ge \tfrac{3}{4} \qquad \text{on } \ [a - \eta, a + \eta] \cap I\,.$$

If $a + \eta < b$, the function u_a is continuous and strictly positive on the compact interval $[a + \eta, b]$ and so has a minimum $m > 0$ there (Theorem 7.24); if $a + \eta \ge b$ put $m = +\infty$ and ignore the second condition below. Now choose $\varepsilon > 0$ subject to

$$\varepsilon < \eta\,, \qquad \Lambda\,\varepsilon < \tfrac{1}{4}\,, \qquad \Lambda\,\varepsilon < \tfrac{m}{2}\,.$$

By Equation (A.746) the solution $u = u_{a-\varepsilon}$ and its momentum $v = v_{a-\varepsilon}$ then satisfy

$$v \ge v_a - \tfrac{1}{4} \ge \tfrac{1}{2} \quad \text{on } \ [a - \eta, a + \eta] \cap I\,, \qquad u \ge u_a - \tfrac{m}{2} \ge \tfrac{m}{2} > 0 \quad \text{on } \ [a + \eta, b]\ .$$

On $[a-\varepsilon, a+\eta]$ the first inequality gives $\mathrm{d}u/\mathrm{d}x = v/P > 0$, so u is strictly increasing there and, since $u\,(a-\varepsilon) = 0$, it is strictly positive on $(a-\varepsilon, a+\eta]$ — in particular on $[a, a+\eta]$, because $a > a-\varepsilon$. Together with the second inequality, $u > 0$ on all of $[a, b]$, so $u_{a-\varepsilon}$ has no zero there. The argument used only the three smallness conditions on ε, each of which persists when ε is decreased, which proves the last clause. □

Remark A.446 (The geometric reading). Corollary A.445 is the analytic content of the picture in Remark 16.54. A solution of Equation (A.734) is the derivative $\partial y/\partial\alpha$ of a one-parameter family of extremals through a common initial point, so u_s describes, to first order, the spreading of the pencil of extremals issuing from the point s. The corollary says that the pencil issuing from a stays spread out over $[a, b]$ if and only if a pencil issuing from a slightly earlier point does, and Theorem A.441 is what makes "slightly earlier" a legitimate perturbation: the focusing distance depends continuously on the point of emission. On the unit sphere, where the extremals are great circles and every pencil refocuses at the antipode, the corollary reproduces the familiar statement that an arc is a minimising geodesic exactly while it is shorter than half a great circle.

Remark A.447. Theorem A.436 discharges the equivalence asserted in Remark 16.57 of Section 16.4.2, and Corollary A.445 supplies the displaced-point phrasing used there. The two conditions the equivalence links are Jacobi's necessary condition (Theorem 16.55) and the sufficiency of the second variation (Theorem 16.56); with the equivalence in hand, a regular extremal with no conjugate point in $(a, b]$ has a positive definite second variation, which is the form in which the result is used in the field-theoretic sufficiency argument of Section 16.4.4 and in the discussion of the kinetic focus of Hamilton's principle in Section 16.7.1.

A.41 Tonelli's Existence Theorem for an Integrand Depending on the Function

This appendix proves Theorem 16.73 of Chapter 16: an integrand $f\,(x, y, q)$ that is convex in the slope q and coercive in it makes the functional Equation (16.11) attain its minimum in the Sobolev class with prescribed endpoint values. The abstract half of the argument is already in the chapter — the direct method Theorem 16.70 and the convexity lemma Lemma 16.72 — and what that lemma leaves owed is the passage from an integrand $f\,(x, q)$ to one that also depends on y. That passage is what is carried out here, and it turns on one fact about the one-dimensional Sobolev space: a bounded set of $W^{1,r}(a, b)$ is precompact in the continuous functions with the uniform norm. That compact embedding is proved below in full (Theorem A.455) from the fundamental theorem of calculus, Hölder's inequality and the Arzelà–Ascoli theorem, the last of which is not available anywhere else in this treatise and is therefore also proved here (Lemma A.454).

Uniform convergence of the competitors is exactly the leverage the y-dependence needs: it makes the values $f\,(x, u_n(x), q)$ converge pointwise to $f\,(x, u(x), q)$ and confines them to a compact range of y, on which the growth hypothesis on $\partial f/\partial q$ can be applied uniformly. With that in hand the classical Scorza–Dragoni argument — which is the route taken when no such growth hypothesis is assumed — is not needed; Remark A.458 states precisely what it would supply and why it is avoided here, and Remark A.459 lists the analytic inputs that are quoted rather than proved, all of them from the Lebesgue theory and the duality of L^r which this treatise declares as imports in Remark 12.1.

A.41.1 Statement

Throughout, $1 < r < \infty$ and $r' = r/(r-1)$ is the conjugate exponent, so that $1/r + 1/r' = 1$; $a < b$ are real and y_a, y_b are the prescribed endpoint values of Equation (16.10).

Theorem A.448 (Tonelli)**.** *Let $f : [a,b] \times \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ satisfy*

(H1) f is continuous, and continuously differentiable with respect to its third argument, with derivative written $f_q = \partial f / \partial q$;

(H2) $q \longmapsto f(x,y,q)$ is convex for every (x,y);

(H3) $f(x,y,q) \geq \alpha |q|^r - \beta$ for constants $\alpha > 0$ and $\beta \in \mathbb{R}$ and all (x,y,q);

(H4) for every $R > 0$ there is a constant C_R with $|f_q(x,y,q)| \leq C_R \left(1 + |q|^{r-1}\right)$ whenever $x \in [a,b]$, $|y| \leq R$ and $q \in \mathbb{R}$.

Then the functional Equation (16.11) attains its minimum on

$$\mathcal{A}_r = \left\{ u \in W^{1,r}(a,b) \;\middle|\; u(a) = y_a,\ u(b) = y_b \right\}, \tag{A.747}$$

that is, there is $u_* \in \mathcal{A}_r$ with $J[u_*] = \inf_{\mathcal{A}_r} J$, and the infimum is finite.

Rests on Theorem 16.70, Lemma 16.72, and Equation (16.11).

Remark A.449 (Where each hypothesis comes from). (H1)–(H4) are the hypotheses of Theorem 16.73 as the chapter states it. (H2) and (H3) — convexity and coercivity in the slope — are Tonelli's own. (H1) and (H4) are the hypotheses under which the chapter's convexity lemma Lemma 16.72 was already stated in the y-free case: there f is of class C^1 in q and $|\partial f/\partial q(x,q)| \leq C\left(1 + |q|^{r-1}\right)$. (H4) is that same bound made *locally uniform in* y, which is the least that can be asked once y enters, and it is automatic for every integrand met in this book — for a mechanical Lagrangian $\mathcal{L} = \frac{1}{2} m_{ij}(q)\dot{q}^i\dot{q}^j - V(q)$ the derivative in the velocities is linear in them, so (H4) holds with $r = 2$ and C_R the maximum of the mass matrix over $|y| \leq R$. The theorem as Tonelli stated it [432] assumes only continuity and convexity; Remark A.458 says what closing that last gap costs.

A.41.2 The Sobolev space in one dimension

Definition A.450 (The space $W^{1,r}(a,b)$)**.** For $1 < r < \infty$, $W^{1,r}(a,b)$ is the set of functions $u : [a,b] \longrightarrow \mathbb{R}$ for which there exists $v \in L^r(a,b)$ with

$$u(x) = u(a) + \int_a^x v(t)\,\mathrm{d}t \qquad \text{for every } x \in [a,b], \tag{A.748}$$

normed by $\|u\|_{W^{1,r}} = \|u\|_{L^r} + \|v\|_{L^r}$. The function v is unique as an element of L^r and is written u'; every $u \in W^{1,r}(a,b)$ is continuous on $[a,b]$, by Equation (A.748) and the absolute continuity of the Lebesgue integral.

Rests on Definition 10.85 and Theorem 7.42.

Remark A.451 (Why the space is defined by the integral formula). Definition 10.85 introduces $H^1 = W^{1,2}$ through *distributional* derivatives, and in one dimension the two definitions agree: if $u \in L^r(a,b)$ has a distributional derivative $v \in L^r$, then $w(x) = \int_a^x v$ has the same distributional derivative, so $u - w$ has distributional derivative zero and is therefore almost everywhere equal to a constant — which is the du Bois-Reymond lemma Lemma 16.19 in its integrable form, the one step of this identification that needs Lebesgue theory rather than the continuous argument given in the chapter. Nothing below uses the distributional definition, so Definition A.450 is taken as the definition and the identification is recorded only to place the space where the reader expects it. In particular each element of $W^{1,r}(a,b)$ is here a genuine continuous function, not an equivalence class, so the boundary conditions in Equation (A.747) have their literal meaning.

Lemma A.452 (Young and Hölder)**.** *Let $1 < r < \infty$ and $r' = r/(r-1)$. For all $s, t \geq 0$,*

$$s\,t \leq \frac{s^r}{r} + \frac{t^{r'}}{r'}\,, \tag{A.749}$$

and consequently, for measurable g, h on (a, b),

$$\int_a^b |g\,h|\,\mathrm{d}x \leq \|g\|_{L^{r'}}\,\|h\|_{L^r}\,. \tag{A.750}$$

Rests on Theorem 7.38 and Definition A.450.

Proof. *Derives Lemma A.452.*

Young. For $s = 0$ or $t = 0$ the inequality is trivial, so let $s, t > 0$ and put $\lambda = 1/r \in (0, 1)$, so $1 - \lambda = 1/r'$. The function $\Phi = -\log$ is of class C^2 on $(0, \infty)$ with $\Phi''(\tau) = \tau^{-2} > 0$, hence convex there: writing $c = \lambda\,\sigma + (1-\lambda)\,\tau$ for $\sigma, \tau > 0$, Taylor's theorem with Lagrange remainder (Theorem 7.38) about c gives $\Phi(\sigma) = \Phi(c) + \Phi'(c)\,(\sigma - c) + \frac{1}{2}\Phi''(\xi)\,(\sigma - c)^2 \geq \Phi(c) + \Phi'(c)\,(\sigma - c)$ and the same with τ in place of σ; multiplying the first by λ, the second by $1 - \lambda$ and adding annihilates the linear terms, because $\lambda\,(\sigma - c) + (1-\lambda)\,(\tau - c) = 0$, and leaves $\lambda\,\Phi(\sigma) + (1-\lambda)\,\Phi(\tau) \geq \Phi(c)$. Taking $\sigma = s^r$ and $\tau = t^{r'}$ and undoing the minus sign,

$$\log\left(\frac{s^r}{r} + \frac{t^{r'}}{r'}\right) \geq \frac{1}{r}\log s^r + \frac{1}{r'}\log t^{r'} = \log(s\,t)\,,$$

and log is increasing, which is Equation (A.749).

Hölder. If $\|g\|_{L^{r'}} = 0$ or $\|h\|_{L^r} = 0$ then $gh = 0$ almost everywhere and there is nothing to prove; if either norm is infinite the right side is $+\infty$. Otherwise replace g by $g/\|g\|_{L^{r'}}$ and h by $h/\|h\|_{L^r}$, which reduces the claim to the case $\|g\|_{L^{r'}} = \|h\|_{L^r} = 1$. Apply Equation (A.749) pointwise with $s = |h(x)|$, $t = |g(x)|$ and integrate:

$$\int_a^b |g\,h|\,\mathrm{d}x \leq \frac{1}{r}\int_a^b |h|^r\,\mathrm{d}x + \frac{1}{r'}\int_a^b |g|^{r'}\,\mathrm{d}x = \frac{1}{r} + \frac{1}{r'} = 1\,.$$

□

Lemma A.453 (Uniform bound and uniform Hölder continuity)**.** *Let $u \in W^{1,r}(a, b)$. Then for all $x, x' \in [a, b]$,*

$$|u(x) - u(x')| \leq \|u'\|_{L^r}\,|x - x'|^{1/r'}\,, \tag{A.751}$$

and consequently

$$\max_{[a,b]} |u| \leq |u(a)| + \|u'\|_{L^r}\,(b-a)^{1/r'}\,. \tag{A.752}$$

Rests on Definition A.450 and Lemma A.452.

Proof. *Derives Lemma A.453.*

By Equation (A.748), for $x' < x$, $u(x) - u(x') = \int_{x'}^x u'(t)\,\mathrm{d}t$. Apply Equation (A.750) on the interval (x', x) with $h = u'$ and g the constant function 1, whose $L^{r'}$ norm over that interval is $(x - x')^{1/r'}$:

$$|u(x) - u(x')| \leq \int_{x'}^x |u'(t)|\,\mathrm{d}t \leq (x - x')^{1/r'}\left(\int_{x'}^x |u'|^r\,\mathrm{d}t\right)^{1/r} \leq \|u'\|_{L^r}\,(x - x')^{1/r'}\,.$$

Taking $x' = a$ and adding $|u(a)|$ gives Equation (A.752). □

A.41.3 The Arzelà–Ascoli theorem and the compact embedding

Lemma A.454 (Arzelà–Ascoli on an interval)**.** *Let (u_n) be a sequence of continuous functions on $[a,b]$ that is*

1. uniformly bounded*: $|u_n(x)| \le K$ for all n and all x; and*
2. equicontinuous*: for every $\varepsilon > 0$ there is $\delta > 0$, independent of n, such that $|x - x'| < \delta$ implies $|u_n(x) - u_n(x')| < \varepsilon$ for every n.*

Then some subsequence converges uniformly on $[a,b]$, and its limit is continuous.
Rests on Theorems 7.7 and 7.8.

Proof. *Derives Lemma A.454.*

A subsequence converging on a countable dense set. Enumerate the rational points of $[a,b]$ together with a and b as $q_1, q_2, \ldots$; the set is countable and dense in $[a,b]$. The numerical sequence $(u_n(q_1))_n$ is bounded by K, so by Bolzano–Weierstrass (Theorem 7.7) there is a subsequence $\left(u_n^{(1)}\right)$ of (u_n) for which $\left(u_n^{(1)}(q_1)\right)$ converges. Applying the same argument to $\left(u_n^{(1)}(q_2)\right)$ produces a subsequence $\left(u_n^{(2)}\right)$ of $\left(u_n^{(1)}\right)$ converging at q_1 and q_2, and so on, giving nested subsequences $\left(u_n^{(j)}\right)_n$, the j-th converging at $q_1, \ldots, q_j$. The diagonal sequence $w_k = u_k^{(k)}$ is, from its j-th term onwards, a subsequence of $\left(u_n^{(j)}\right)_n$, so $(w_k(q_j))_k$ converges for every j.

Uniform convergence. Let $\varepsilon > 0$ and take $\delta > 0$ from equicontinuity for $\varepsilon/3$. Partition $[a,b]$ into finitely many subintervals of length less than δ and pick in each one a point of the dense set, obtaining $q_{j_1}, \ldots, q_{j_m}$ such that every $x \in [a,b]$ lies within δ of some q_{j_i}. Because each of the m sequences $(w_k(q_{j_i}))_k$ converges, it is Cauchy (Theorem 7.8), and there is an N — the largest of the m thresholds — with $|w_k(q_{j_i}) - w_l(q_{j_i})| < \varepsilon/3$ for all $k, l \ge N$ and all i. For an arbitrary $x \in [a,b]$ choose i with $|x - q_{j_i}| < \delta$; then, for $k, l \ge N$,

$$|w_k(x) - w_l(x)| \le |w_k(x) - w_k(q_{j_i})| + |w_k(q_{j_i}) - w_l(q_{j_i})| + |w_l(q_{j_i}) - w_l(x)| < \varepsilon\,,$$

the outer terms by equicontinuity. The bound does not involve x, so (w_k) is uniformly Cauchy; at each fixed x it therefore converges (Theorem 7.8), to a limit $u(x)$, and letting $l \to \infty$ in the display gives $|w_k(x) - u(x)| \le \varepsilon$ for every $k \ge N$ and every x — uniform convergence.

Continuity of the limit. Given $\varepsilon > 0$, pick k with $\max_{[a,b]} |w_k - u| < \varepsilon/3$ and δ from equicontinuity for $\varepsilon/3$; then $|x - x'| < \delta$ gives

$$|u(x) - u(x')| \le |u(x) - w_k(x)| + |w_k(x) - w_k(x')| + |w_k(x') - u(x')| < \varepsilon\,.$$

□

Theorem A.455 (Compact embedding of $W^{1,r}(a,b)$ into the continuous functions)**.** *Let $(u_n) \subset W^{1,r}(a,b)$ with*

$$\sup_n \|u_n'\|_{L^r} \le M < \infty \qquad \text{and} \qquad \sup_n |u_n(a)| \le M_0 < \infty\,. \tag{A.753}$$

Then a subsequence of (u_n) converges uniformly on $[a,b]$ to a continuous limit u. If in addition $u_n' \rightharpoonup v$ weakly in $L^r(a,b)$ along that subsequence — that is, $\int_a^b g\,u_n'\,\mathrm{d}x \longrightarrow \int_a^b g\,v\,\mathrm{d}x$ for every $g \in L^{r'}(a,b)$ — then $u \in W^{1,r}(a,b)$ with $u' = v$.
Rests on Lemmas A.453 and A.454.

Proof. *Derives Theorem A.455.*

Precompactness. By Equation (A.752) the sequence is uniformly bounded by $K = M_0 + M\,(b-a)^{1/r'}$, and by Equation (A.751)

$$\left|u_n(x) - u_n\left(x'\right)\right| \le M\ \left|x - x'\right|^{1/r'}$$

for every n, so the sequence is equicontinuous: given $\varepsilon > 0$, the choice $\delta = (\varepsilon/M)^{r'}$ serves every n at once, and it is here that $r > 1$ is used, since $1/r' > 0$. Lemma A.454 now supplies a uniformly convergent subsequence, which we do not relabel, with continuous limit u.

Identification of the derivative. Fix $x \in [a,b]$ and let χ_x be the function equal to 1 on (a,x) and to 0 on (x,b). It belongs to $L^{r'}(a,b)$, being bounded on a bounded interval, so weak convergence gives

$$\int_a^x u_n'(t)\,\mathrm{d}t = \int_a^b \chi_x\, u_n'\,\mathrm{d}t \longrightarrow \int_a^b \chi_x\, v\,\mathrm{d}t = \int_a^x v(t)\,\mathrm{d}t\,.$$

On the other hand Equation (A.748) for u_n says the left side equals $u_n(x) - u_n(a)$, which converges to $u(x) - u(a)$ by uniform — indeed pointwise — convergence. Hence $u(x) = u(a) + \int_a^x v$ for every x, which is Equation (A.748) for u with v in the role of the derivative; since $v \in L^r$ by hypothesis, $u \in W^{1,r}(a,b)$ and $u' = v$. □

A.41.4 Weak lower semicontinuity with y present

Theorem A.456 (Convexity in q implies weak lower semicontinuity)**.** *Let f satisfy (H1)–(H4) of Theorem A.448. Let $(u_n) \subset W^{1,r}(a,b)$ and $u \in W^{1,r}(a,b)$ be such that*

1. $\sup_n \|u_n'\|_{L^r} < \infty$;
2. $u_n \longrightarrow u$ *uniformly on* $[a,b]$;
3. $u_n' \rightharpoonup u'$ *weakly in* $L^r(a,b)$.

Then

$$\int_a^b f\left(x,u,u'\right)\mathrm{d}x \le \liminf_{n\to\infty} \int_a^b f\left(x,u_n,u_n'\right)\mathrm{d}x\,. \tag{A.754}$$

Rests on Lemma 16.72, Equation (16.62), and Theorem A.455.

Proof. *Derives Theorem A.456.*

Write $M = \sup_n \|u_n'\|_{L^r}$ and $R = \sup_n \max_{[a,b]} |u_n|$, the latter finite because a uniformly convergent sequence of continuous functions is uniformly bounded; enlarge R if necessary so that also $\max_{[a,b]} |u| \le R$.

Step 1: the tangent inequality, taken at the limit slope. Convexity of a C^1 function of q puts its graph above every tangent, which is Equation (16.62); applied at the fixed value $y = u_n(x)$ of the second argument, with $q_1 = u'(x)$ and $q_2 = u_n'(x)$, it gives, for almost every x,

$$f\left(x,u_n,u_n'\right) \ge f\left(x,u_n,u'\right) + g_n(x)\left[u_n'(x) - u'(x)\right]\,, \qquad g_n = f_q\left(\cdot\,,u_n,u'\right)\,. \tag{A.755}$$

Note that the tangent is taken at the *limit* slope u' but at the *approximating* value u_n of the function; this is the one departure from the y-free argument of Lemma 16.72, and it is what makes both terms on the right tractable. Integrating Equation (A.755),

$$\int_a^b f\left(x,u_n,u_n'\right)\mathrm{d}x \ge \int_a^b f\left(x,u_n,u'\right)\mathrm{d}x + \int_a^b g_n\left(u_n' - u'\right)\mathrm{d}x\,. \tag{A.756}$$

Both integrals on the right are well defined: the second because $g_n \in L^{r'}$ by Step 3 below and $u_n' - u' \in L^r$, so Equation (A.750) applies; the first because its integrand is bounded below by $-\beta$ through (H3), so the integral exists in $(-\infty, +\infty]$.

Step 2: the first term. By (H1) the map $y \longmapsto f(x, y, u'(x))$ is continuous, and $u_n(x) \longrightarrow u(x)$ for every x, so

$$f(x, u_n(x), u'(x)) \longrightarrow f(x, u(x), u'(x)) \qquad \text{for almost every } x\,.$$

By (H3) the functions $f(x, u_n, u') + \beta$ are nonnegative and measurable, so Fatou's lemma applies to them and gives

$$\int_a^b f(x, u, u')\,\mathrm{d}x \le \liminf_{n\to\infty} \int_a^b f(x, u_n, u')\,\mathrm{d}x\,, \tag{A.757}$$

the constants $\beta(b-a)$ cancelling from both sides. The inequality holds in $(-\infty, +\infty]$, so nothing is assumed about the finiteness of either side.

Step 3: the second term vanishes. Put $g = f_q(\cdot, u, u')$. By (H4) with the constant C_R,

$$|g_n(x)| \le C_R\left(1 + |u'(x)|^{r-1}\right) \qquad \text{and the same bound for } |g|\,, \tag{A.758}$$

and the right-hand side lies in $L^{r'}(a,b)$, because $\left(|u'|^{r-1}\right)^{r'} = |u'|^r$ is integrable and constants are integrable on a bounded interval. In particular $g_n, g \in L^{r'}$, as promised in Step 1. Moreover $g_n \longrightarrow g$ pointwise almost everywhere, by the continuity of f_q in its second argument (H1) and the pointwise convergence $u_n \to u$. Since $|g_n - g|^{r'} \le (2C_R)^{r'}\left(1 + |u'|^{r-1}\right)^{r'} \in L^1$ and tends to 0 almost everywhere, the dominated convergence theorem gives

$$\|g_n - g\|_{L^{r'}} \longrightarrow 0\,. \tag{A.759}$$

Now split

$$\int_a^b g_n\,(u_n' - u')\,\mathrm{d}x = \int_a^b g\,(u_n' - u')\,\mathrm{d}x + \int_a^b (g_n - g)\,(u_n' - u')\,\mathrm{d}x\,.$$

The first integral tends to 0: $g \in L^{r'}$ and $u_n' \rightharpoonup u'$ in L^r, which is exactly the statement that $\int g\,u_n' \to \int g\,u'$. The second is bounded, by Hölder's inequality Equation (A.750), by

$$\|g_n - g\|_{L^{r'}}\ \|u_n' - u'\|_{L^r} \le \left(M + \|u'\|_{L^r}\right)\|g_n - g\|_{L^{r'}}\,,$$

which tends to 0 by Equation (A.759) and hypothesis (1). Hence

$$\int_a^b g_n\,(u_n' - u')\,\mathrm{d}x \longrightarrow 0\,. \tag{A.760}$$

Conclusion. Take the lower limit in Equation (A.756). The lower limit of a sum is at least the lower limit of the first summand plus the limit of the second when the latter exists, so by Equations (A.757) and (A.760)

$$\liminf_n \int_a^b f(x, u_n, u_n')\,\mathrm{d}x \ge \liminf_n \int_a^b f(x, u_n, u')\,\mathrm{d}x + 0 \ge \int_a^b f(x, u, u')\,\mathrm{d}x\,,$$

which is Equation (A.754). □

A.41.5 Proof of the theorem

Proof of Theorem A.448. *Derives Theorem A.448.*

The infimum is finite. The affine function $\ell(x) = y_a + (y_b - y_a)\,(x-a)\,/\,(b-a)$ lies in $\mathcal{A}_r$, with constant derivative, and f is continuous, so $J[\ell]$ is finite and $m = \inf_{\mathcal{A}_r} J \le J[\ell] < \infty$. By (H3), $J[u] \ge -\beta\,(b-a)$ for every admissible u, so $m > -\infty$.

Coercivity bounds a minimising sequence. Let $(u_n) \subset \mathcal{A}_r$ satisfy $J[u_n] \longrightarrow m$; discarding finitely many terms we may assume $J[u_n] \le m+1$ for every n. By (H3),

$$m + 1 \ge J[u_n] \ge \alpha \int_a^b |u_n'|^r \,\mathrm{d}x - \beta\,(b-a)\,, \tag{A.761}$$

so $\|u_n'\|_{L^r}^r \le [m+1+\beta\,(b-a)]\,/\alpha =: M^r$, a bound independent of n. Since $u_n(a) = y_a$ for every n, the hypotheses Equation (A.753) of Theorem A.455 hold with $M_0 = |y_a|$, and by Equation (A.752) the sequence is also bounded in L^r, hence bounded in $W^{1,r}(a,b)$.

Extraction of a limit. The sequence (u_n') is bounded in $L^r(a,b)$, a reflexive Banach space for $1 < r < \infty$, so some subsequence converges weakly, $u_{n_k}' \rightharpoonup v$ in $L^r(a,b)$ — this is the compactness input of Theorem 16.70, quoted there in the same form (see Remark A.459). Passing to a further subsequence, not relabelled, Theorem A.455 makes u_{n_k} converge uniformly on $[a,b]$ to a continuous u_*, and identifies $u_* \in W^{1,r}(a,b)$ with $u_*' = v$.

The limit is admissible. Uniform convergence is in particular pointwise at the endpoints, so $u_*(a) = \lim_k u_{n_k}(a) = y_a$ and $u_*(b) = \lim_k u_{n_k}(b) = y_b$: the class $\mathcal{A}_r$ is closed under this mode of convergence, which is the second place where the compact embedding earns its keep — weak convergence in $W^{1,r}$ alone says nothing about the value of a function at a point.

The limit minimises. The three hypotheses of Theorem A.456 hold for (u_{n_k}) and u_*, so

$$J[u_*] \le \liminf_k J[u_{n_k}] = m\,,$$

while $J[u_*] \ge m$ because $u_* \in \mathcal{A}_r$. Hence $J[u_*] = m$ and the infimum is attained. □

Remark A.457 (The shape of the argument, against the abstract direct-method theorem). The proof just given is the abstract direct method with each of its three ingredients made concrete, and it is worth seeing which concrete statement plays which abstract role. Coercivity of J in the sense of Theorem 16.70 is Equation (A.761), and it controls only the *derivative*; that this suffices to control the function as well is Equation (A.752), i.e. the boundary condition plus Hölder's inequality. Sequential weak lower semicontinuity is Theorem A.456. The compactness that the abstract theorem imports from reflexivity is used here in two different topologies at once — weakly in L^r for the derivatives, uniformly in C^0 for the functions — and the second is not a convenience but a necessity: without it neither the endpoint values nor the y-dependence of the integrand would survive the limit. That is the precise sense in which Lemma 16.72 left something owed.

Remark A.458 (The route not taken: Scorza-Dragoni). If (H4) is dropped and f is asked only to be continuous and convex in q — Tonelli's own hypotheses — Step 3 above breaks down, because $|f_q(x, u_n, u')|$ can then grow with n faster than any fixed $L^{r'}$ function, and the classical repair is a different one. It rests on the theorem of Scorza-Dragoni: a Carathéodory integrand — measurable in x, continuous in (y,q) — is, after the deletion from $[a,b]$ of a set of arbitrarily small measure, continuous *jointly* in all its arguments on what remains; on that set the modulus of continuity in y is then uniform in q over compacta, which is what Step 3 needs, and the deleted set is handled by the coercivity bound (H3). This treatise's bibliography carries no entry for Scorza-Dragoni's paper, and the statement above is therefore quoted here *without a citation* — the honest record of a result that is named but neither used nor verified in this book. Nothing in Theorem A.448 depends on it: the theorem proved above is the theorem under (H1)–(H4), and every integrand this treatise minimises satisfies (H4) (Remark A.449).

Remark A.459 (What is quoted here). Three analytic inputs are used and not proved in this treatise. Each belongs to the measure theory and functional analysis that Remark 12.1 of Chapter 12 already declares as imports, and all three are in Reed and Simon [1481].

1. *The Lebesgue integral and its convergence theorems.* Fatou's lemma is used once, in Equation (A.757), and the dominated convergence theorem once, in Equation (A.759). The absolute continuity of the integral, used in Definition A.450 to see that an element of $W^{1,r}$ is continuous, is of the same family.

2. *The duality* $(L^r)^* = L^{r'}$ *and the reflexivity of* L^r *for* $1 < r < \infty$, from which a bounded sequence in L^r has a weakly convergent subsequence. This is the only compactness assumption in the proof, and it is the same one Theorem 16.70 makes in the chapter, where it is named as the Eberlein–Smulian theorem; the special case $r = 2$ is the Hilbert-space statement of Chapter 12.

3. *The du Bois-Reymond lemma for integrable functions*, used only in Remark A.451 to identify Definition A.450 with the distributional definition Definition 10.85. No step of the proof uses it; the continuous version proved in the chapter, Lemma 16.19, is not strong enough for the identification, which is why the space is defined here by Equation (A.748).

Everything else — the Arzelà–Ascoli theorem, the compact embedding, Young's and Hölder's inequalities, and the semicontinuity theorem itself — is proved above from Chapter 7 and the chapter.

Remark A.460. Theorem A.448 discharges Theorem 16.73 of Section 16.5.2, completing the direct method of Theorem 16.70 for the concrete functional Equation (16.11). Its hypotheses are what Remark 16.74 tests against the Lagrangians of mechanics, and the answer given there is unchanged by anything proved here: convexity in the velocities holds for a kinetic energy with a positive definite mass matrix, while coercivity (H3) demands a potential bounded above, so the theorem certifies existence for a free or repelled particle and not for a confining one. The failure of convexity is equally untouched — the functional Equation (16.64) satisfies (H1), (H3) and (H4) but not (H2), and has no minimiser.

A.42 Completeness of the Sturm–Liouville Eigenfunctions in the Energy Norm

This appendix proves the statement quoted in Remark 16.77 of Chapter 16: the eigenfunctions of a regular Sturm–Liouville problem are complete not merely in the weighted L^2 space, but in the *energy* norm built from the quadratic form of the problem, and with that completeness the inequality Equation (16.70) sharpens into the identity Equation (16.71) for the energy form. The identity is what the Rayleigh–Ritz error estimate Corollary 16.78 needs; the inequality, which is what the variational characterisation Theorem 16.76 needs, is proved in the chapter from Bessel's inequality alone and is *independent* of everything below — a point Remark A.473 returns to, since a circular appendix would silently destroy the chapter's argument.

The route is the one the subject was born from [388]: invert the differential operator. The inverse of a regular Sturm–Liouville operator with Dirichlet data is an integral operator whose kernel is the Green function of the problem; that kernel is continuous, so the operator is compact and self-adjoint, and the Hilbert–Schmidt theorem Theorem 12.44, proved in Appendix A.23, hands back an orthonormal basis of eigenfunctions. Completeness in the weighted L^2 space — which Theorem 16.76 *assumes* — then comes out as a theorem, and completeness in the energy norm follows from it by one identity relating the two inner products.

Throughout, the data are those of Definition 16.75: p, q, r are continuous and real on $[a,b]$ with $p>0$ and $r>0$, and the boundary conditions are $y(a)=y(b)=0$. As in the proof of Theorem 16.76, fix once and for all a constant K with

$$q + Kr \geq 0 \qquad \text{on } [a,b]\,, \tag{A.762}$$

which is possible because q is continuous and $r>0$ on a compact interval, and write

$$\langle u, v\rangle_r = \int_a^b r\,u\,v\,\mathrm{d}x\,, \qquad B_K[u,v] = \int_a^b \left[p\,u'v' + (q+Kr)\,u\,v\right]\mathrm{d}x\,. \tag{A.763}$$

The operator of the problem, shifted by K, is

$$L_K[u] = -\frac{\mathrm{d}}{\mathrm{d}x}\left(p\,\frac{\mathrm{d}u}{\mathrm{d}x}\right) + (q+Kr)\,u\,, \tag{A.764}$$

so that y solves Equation (16.66) with eigenvalue λ if and only if $L_K[y] = (\lambda+K)\,r\,y$.

A.42.1 Statement

Theorem A.461 (Completeness in the weighted and in the energy norm)**.** *Let $p,q,r \in C^0[a,b]$ with $p>0$ and $r>0$. Then:*

1. *the Dirichlet problem Equation (16.66) has an infinite sequence of real eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots$ with $\lambda_n \longrightarrow +\infty$, each of finite multiplicity, and eigenfunctions y_n that are orthonormal in $\langle\cdot,\cdot\rangle_r$ and form an orthonormal basis of the weighted space $L^2_r(a,b)$;*
2. *the rescaled eigenfunctions $y_n/\sqrt{\lambda_n+K}$ form an orthonormal basis of the energy space H_E of Definition A.462 with the inner product B_K;*
3. *consequently, for every $y \in H_E$ — in particular for every $y \in C^1_0[a,b]$ — the partial sums $S_N = \sum_{n\leq N} c_n y_n$, $c_n = \langle y, y_n\rangle_r$, satisfy $B_K\left[y-S_N, y-S_N\right] \longrightarrow 0$, and*

$$\int_a^b \left(p\,y'^2 + q\,y^2\right)\mathrm{d}x = \sum_{n\geq 1} \lambda_n\,c_n^2\,, \tag{A.765}$$

the series on the right converging.

Rests on Definition 16.75, Theorem 12.44, and Equation (16.66).

Equation (A.765) is Equation (16.71). The proof occupies the rest of the section: the energy space (Appendix A.42.2), the Green function and the inverse operator (Appendix A.42.3), its compactness and the spectral decomposition (Appendix A.42.4), and the transfer to the energy norm (Appendix A.42.5).

A.42.2 The energy space

Definition A.462 (The weighted space and the energy space)**.** $L^2_r(a,b)$ is the space of square-integrable functions with the inner product $\langle\cdot,\cdot\rangle_r$ of Equation (A.763), and

$$H_E = \left\{ u \in W^{1,2}(a,b) \;\mid\; u(a) = u(b) = 0 \right\} \tag{A.766}$$

is the *energy space*, where $W^{1,2}(a,b)$ is the space of Definition A.450 with $r=2$: the continuous functions u for which there is $u' \in L^2(a,b)$ with $u(x) = u(a) + \int_a^x u'$. On H_E the energy form B_K of Equation (A.763) is used as an inner product, with norm $\|u\|_E = B_K[u,u]^{1/2}$.

Rests on Definitions 16.75 and A.450.

Lemma A.463 (Poincaré inequality; B_K is an inner product)**.** *Write* $p_- = \min_{[a,b]} p > 0$, $p_+ = \max_{[a,b]} p$, $r_- = \min_{[a,b]} r > 0$, $r_+ = \max_{[a,b]} r$ *and* $\kappa_+ = \max_{[a,b]} (q + Kr)$. *For every* $u \in H_E$,

$$\max_{[a,b]} |u| \le (b-a)^{1/2} \|u'\|_{L^2} , \qquad \langle u, u \rangle_r \le r_+ (b-a)^2 \|u'\|_{L^2}^2 , \tag{A.767}$$

and consequently

$$p_- \|u'\|_{L^2}^2 \le B_K[u,u] \le \left[p_+ + \kappa_+ (b-a)^2 \right] \|u'\|_{L^2}^2 . \tag{A.768}$$

In particular B_K *is an inner product on* H_E*: it is bilinear and symmetric by inspection, and* $B_K[u,u] = 0$ *forces* $u = 0$.

Rests on Definition A.462 and Lemma A.452.

Proof. *Derives Lemma A.463.*

For $u \in H_E$ and $x \in [a,b]$, $u(a) = 0$ gives $u(x) = \int_a^x u'$, and the Cauchy–Schwarz case $r = 2$ of Hölder's inequality Equation (A.750) bounds this by $(x-a)^{1/2} \|u'\|_{L^2} \le (b-a)^{1/2} \|u'\|_{L^2}$, which is the first inequality of Equation (A.767). Integrating its square against r over $[a,b]$ gives the second.

For Equation (A.768), the lower bound drops the term $(q + Kr)\, u^2 \ge 0$, which is nonnegative by Equation (A.762), and bounds $p \ge p_-$; the upper bound uses $p \le p_+$, $q + Kr \le \kappa_+$ and the second inequality of Equation (A.767) with r replaced by the constant 1, i.e. $\int u^2 \le (b-a)^2 \|u'\|_{L^2}^2$. If $B_K[u,u] = 0$ then $\|u'\|_{L^2} = 0$ by the lower bound, so $u' = 0$ almost everywhere and $u(x) = u(a) + \int_a^x u' = 0$. □

Lemma A.464 (H_E is a Hilbert space)**.** (H_E, B_K) *is complete, hence a Hilbert space in the sense of Definition 12.2, and the inclusion* $H_E \subset L^2_r(a,b)$ *is continuous.*

Rests on Lemma A.463 and Theorem 12.12.

Proof. *Derives Lemma A.464.*

Continuity of the inclusion is the second inequality of Equation (A.767) combined with the lower bound in Equation (A.768): $\langle u, u \rangle_r \le r_+ (b-a)^2 p_-^{-1} B_K[u,u]$.

Let (u_k) be Cauchy in $\|\cdot\|_E$. By the lower bound in Equation (A.768) the sequence (u_k') is Cauchy in $L^2(a,b)$, which is complete (Theorem 12.12), so $u_k' \longrightarrow v$ in L^2. By the first inequality of Equation (A.767) applied to $u_k - u_l$, the sequence (u_k) is uniformly Cauchy on $[a,b]$, hence converges uniformly to a continuous u with $u(a) = u(b) = 0$. Passing to the limit in $u_k(x) = \int_a^x u_k'$ — legitimate because $|\int_a^x (u_k' - v)| \le (b-a)^{1/2} \|u_k' - v\|_{L^2} \longrightarrow 0$ by Equation (A.750) — gives $u(x) = \int_a^x v$, so $u \in H_E$ with $u' = v$. Finally the upper bound in Equation (A.768) applied to $u_k - u$ gives $\|u_k - u\|_E \longrightarrow 0$. □

Remark A.465 (This is the space the chapter's trial functions live in). $C_0^1[a,b] \subset H_E$, since a continuously differentiable function vanishing at both ends satisfies Equation (A.748) with u' its classical derivative. So every statement below applies to the trial functions of Theorem 16.76 and Proposition 16.79 without further comment. The space H_E is the $H_0^1(a,b)$ of Definition 10.85: the identification of the two descriptions is Remark A.451 together with the fact that a $W^{1,2}$ function vanishing at both endpoints is a uniform limit of smooth functions of compact support, which is not needed here and is not proved.

A.42.3 The Green function and the inverse operator

Lemma A.466 (The homogeneous Dirichlet problem is trivial)**.** *Let* φ *solve* $L_K[\varphi] = 0$ *on* $[a,b]$ *with* $\varphi(a) = \varphi(b) = 0$. *Then* $\varphi \equiv 0$.

Rests on Lemma A.463 and Equation (A.764).

Proof. *Derives Lemma A.466.*

Multiply $L_K[\varphi] = 0$ by φ and integrate over $[a, b]$; one integration by parts (Equation (7.28)), whose boundary term $[p\,\varphi\,\varphi']_a^b$ vanishes because φ does at both ends, gives $B_K[\varphi, \varphi] = 0$. Since $\varphi \in C^1[a, b]$ vanishes at a, it lies in H_E, and Lemma A.463 forces $\varphi = 0$. □

Definition A.467 (Green function of the shifted problem)**.** Let ϕ and ψ be the solutions of $L_K[u] = 0$ determined by

$$\phi(a) = 0\,, \quad (p\,\phi')\,(a) = 1\,, \qquad \psi(b) = 0\,, \quad (p\,\psi')\,(b) = 1\,, \tag{A.769}$$

which exist, are unique and are defined on the whole of $[a, b]$ by Lemma A.439 applied to the system form of $L_K[u] = 0$ — the equation $(p\,u')' = (q + Kr)\,u$ is Equation (A.734) with $P = p$ and $Q = q + Kr$. Put

$$C = p\,(\phi\,\psi' - \psi\,\phi')\,, \tag{A.770}$$

a nonzero constant, and define the *Green function*

$$G\,(x, s) = -\frac{1}{C} \times \begin{cases} \phi(x)\,\psi(s) & a \le x \le s \le b\,, \\ \phi(s)\,\psi(x) & a \le s \le x \le b\,. \end{cases} \tag{A.771}$$

Rests on Lemmas A.442 and A.466.

Lemma A.468 (The kernel is well defined, symmetric and Lipschitz)**.** *The constant C of Equation (A.770) is independent of x and nonzero; G is well defined by Equation (A.771) (the two branches agree on $x = s$), symmetric, $G\,(x, s) = G\,(s, x)$, continuous on $[a, b] \times [a, b]$, and there is a constant L_G with*

$$\left|G\,(x, s) - G\,(x', s)\right| \le L_G\,\left|x - x'\right| \qquad \text{for all } x, x', s \in [a, b]\,. \tag{A.772}$$

Rests on Definition A.467 and Theorem 7.35.

Proof. *Derives Lemma A.468.*

That C is constant is Lemma A.442 applied with $P = p$ and $Q = q + Kr$. If $C = 0$ the same lemma makes ϕ and ψ linearly dependent, so ϕ would vanish at b as ψ does, and ϕ would be a nontrivial solution of the homogeneous Dirichlet problem — nontrivial because $(p\phi')\,(a) = 1$ — contradicting Lemma A.466. Hence $C \neq 0$.

The two branches of Equation (A.771) coincide when $x = s$, so G is well defined, and exchanging x and s exchanges the two branches, which is the symmetry. Both ϕ and ψ are of class C^1 on the compact interval, hence Lipschitz there by the mean value theorem (Theorem 7.35) with constants $\max|\phi'|$ and $\max|\psi'|$; each branch of Equation (A.771) is therefore Lipschitz in x uniformly in s, with $L_G = |C|^{-1} \max\{\max|\phi'|\,\max|\psi|\,, \max|\phi|\,\max|\psi'|\}$, and since the branches agree where they meet, the bound Equation (A.772) holds for x, x' on opposite sides of s as well, by splitting the increment at s. Continuity follows, jointly in (x, s), from Equation (A.772) and the symmetry. □

Theorem A.469 (The Green operator inverts L_K)**.** *Define*

$$(Tf)\,(x) = \int_a^b G\,(x, s)\,f(s)\,r(s)\,\mathrm{d}s\,, \qquad f \in L^2_r(a, b)\,. \tag{A.773}$$

Then:

1. *for continuous f, $u = Tf$ is of class C^1 with $p\,u'$ of class C^1, and it is the unique solution of*

$$L_K[u] = r\,f\,, \qquad u(a) = u(b) = 0\,. \tag{A.774}$$

2. *T maps $L^2_r(a,b)$ into H_E, with $\|Tf\|_E \le C_P \|f\|_r$ where $C_P = \left(r_+ (b-a)^2 / p_-\right)^{1/2}$, and*
$$B_K[Tf, \chi] = \langle f, \chi \rangle_r \qquad \text{for every } \chi \in H_E\,. \tag{A.775}$$

3. *T is self-adjoint and positive on $L^2_r(a,b)$, and injective.*

Rests on Definition A.467 and Lemmas A.463 and A.466.

Proof. *Derives Theorem A.469.*
(1) Split Equation (A.773) at x using Equation (A.771):
$$-C\,u(x) = \psi(x) \int_a^x \phi\, f\, r\, \mathrm{d}s + \phi(x) \int_x^b \psi\, f\, r\, \mathrm{d}s\,, \tag{A.776}$$
the first branch of Equation (A.771) governing $s \ge x$ and the second $s \le x$. Both integrals have continuous integrands, so by the fundamental theorem of calculus (Theorem 7.42) the right side is differentiable and
$$-C\,u'(x) = \psi'(x) \int_a^x \phi\, f\, r\, \mathrm{d}s + \phi'(x) \int_x^b \psi\, f\, r\, \mathrm{d}s + [\psi(x)\phi(x) - \phi(x)\psi(x)]\, f(x) r(x)\,,$$
the last bracket vanishing identically: the two boundary terms produced by differentiating the limits of integration cancel. Hence
$$-C\,p(x)u'(x) = (p\psi')(x) \int_a^x \phi\, f\, r\, \mathrm{d}s + (p\phi')(x) \int_x^b \psi\, f\, r\, \mathrm{d}s\,,$$
which is again differentiable because $p\phi'$ and $p\psi'$ are of class C^1, with $(p\phi')' = (q + Kr)\,\phi$ and likewise for ψ. Differentiating,
$$\begin{aligned} -C\,(p\,u')'(x) &= (q+Kr)\,\psi(x) \int_a^x \phi\, f\, r\, \mathrm{d}s + (q+Kr)\,\phi(x) \int_x^b \psi\, f\, r\, \mathrm{d}s \\ &\quad + [(p\psi')(x)\phi(x) - (p\phi')(x)\psi(x)]\, f(x)\, r(x) \\ &= -C\,(q+Kr)\,u(x) + C\,f(x)\,r(x)\,, \end{aligned}$$
using Equation (A.776) for the first two terms and Equation (A.770) for the bracket, which is $\phi\,(p\psi') - \psi\,(p\phi') = p\,(\phi\psi' - \psi\phi') = C$. Dividing by $-C$ gives $(p\,u')' = (q+Kr)\,u - r\,f$, that is Equation (A.774). The boundary conditions hold because $x = a$ kills the first integral in Equation (A.776) and leaves $\phi(a) = 0$ in the second, and symmetrically at $x = b$ with $\psi(b) = 0$. Uniqueness is Lemma A.466 applied to the difference of two solutions.

(2) Let first f be continuous and $u = Tf$. Then $u \in C^1[a,b]$ with $u(a) = u(b) = 0$, so $u \in H_E$, and for $\chi \in C^1_0[a,b]$ one integration by parts (Equation (7.28)) applied to Equation (A.774) gives $B_K[u,\chi] = \langle f, \chi \rangle_r$, the boundary term $[p\,u'\chi]_a^b$ vanishing because χ does. The same identity for arbitrary $\chi \in H_E$ follows because both sides are continuous in χ for $\|\cdot\|_E$ — the left side by the Cauchy–Schwarz inequality for the inner product B_K, the right by Lemma A.464 — and, by the argument of Lemma A.464 run backwards, each $\chi \in H_E$ is an $\|\cdot\|_E$-limit of members of $C^1_0[a,b]$: take $\chi_k(x) = \int_a^x v_k$ with v_k continuous, $\int_a^b v_k = 0$ and $v_k \longrightarrow \chi'$ in L^2, which is possible because the continuous functions are dense in L^2 (Remark A.474) and the mean may be subtracted off with a loss tending to zero.

Taking $\chi = u$ in Equation (A.775) and using Equation (A.767),
$$\|u\|_E^2 = \langle f, u \rangle_r \le \|f\|_r \, \|u\|_r \le C_P \, \|f\|_r \, \|u\|_E\,, \tag{A.777}$$
so $\|Tf\|_E \le C_P \|f\|_r$ for continuous f. For general $f \in L^2_r$ take continuous $f_k \longrightarrow f$ in L^2_r; by Equation (A.777) the sequence (Tf_k) is Cauchy in H_E, hence convergent

there (Lemma A.464), while by Equation (A.773) and the Cauchy–Schwarz inequality $|Tf_k(x) - Tf(x)| \le \max|G|\,(r_+(b-a))^{1/2}\,\|f_k - f\|_r$ tends to zero uniformly. The two limits agree, so $Tf \in H_E$, the bound persists, and Equation (A.775) passes to the limit.

(3) For $f, g \in L^2_r$ put $u = Tf$, $w = Tg$. Then Equation (A.775) twice, with the symmetry of B_K in between, gives

$$\langle Tf, g\rangle_r = \langle g, u\rangle_r = B_K[w,u] = B_K[u,w] = \langle f, w\rangle_r = \langle f, Tg\rangle_r \ ,$$

so T is self-adjoint; and $\langle Tf, f\rangle_r = B_K[u,u] \ge 0$, so T is positive. If $Tf = 0$ then $B_K[0,\chi] = \langle f, \chi\rangle_r = 0$ for every $\chi \in H_E$, and since H_E is dense in L^2_r — it contains the functions χ_k built above from an arbitrary continuous v, and the continuous functions are dense — this forces $f = 0$. Hence T is injective. □

A.42.4 Compactness and the spectral decomposition

Theorem A.470 (T is compact)**.** *T maps every bounded sequence of $L^2_r(a,b)$ to a sequence with a convergent subsequence in $L^2_r(a,b)$; that is, T is compact in the sense of Definition 12.41.*

Rests on Lemmas A.454 and A.468.

Proof. *Derives Theorem A.470.*

Let $\|f_n\|_r \le M$. By the Cauchy–Schwarz inequality in L^2_r applied to Equation (A.773),

$$|(Tf_n)(x)| \le \left(\int_a^b G(x,s)^2\, r(s)\,\mathrm{d}s\right)^{1/2} \|f_n\|_r \le M \max|G|\,(r_+(b-a))^{1/2} \ ,$$

so the sequence (Tf_n) is uniformly bounded, and by the Lipschitz estimate Equation (A.772),

$$|(Tf_n)(x) - (Tf_n)(x')| \le \int_a^b |G(x,s) - G(x',s)|\,|f_n(s)|\,r(s)\,\mathrm{d}s \le L_G\,(r_+(b-a))^{1/2}\,M\ |x - x'| \ ,$$

so it is equicontinuous, with a modulus independent of n. By the Arzelà–Ascoli theorem Lemma A.454 a subsequence converges uniformly on $[a,b]$, and uniform convergence implies convergence in L^2_r, since $\|w\|_r^2 \le r_+(b-a)\max|w|^2$. □

Theorem A.471 (Spectral decomposition and completeness in L^2_r)**.** *There is an orthonormal basis $\{y_n\}_{n\ge1}$ of $L^2_r(a,b)$ and numbers $\mu_n > 0$ with $\mu_n \longrightarrow 0$ such that $Ty_n = \mu_n y_n$. Setting $\lambda_n = \mu_n^{-1} - K$, each y_n is a classical solution of the Sturm–Liouville problem Equation (16.66) with eigenvalue λ_n, and $\lambda_n \longrightarrow +\infty$. Every eigenvalue has finite multiplicity, and the λ_n may be enumerated in nondecreasing order.*

Rests on Theorems 12.44, A.469 and A.470.

Proof. *Derives Theorem A.471.*

$L^2_r(a,b)$ is a Hilbert space: its inner product differs from that of $L^2(a,b)$ by the factor r, which is bounded between the positive constants r_- and r_+, so the two norms are equivalent and completeness transfers from Theorem 12.12. On it, T is bounded (by Equation (A.777) and Equation (A.767)), self-adjoint and compact, by Theorems A.469 and A.470. The Hilbert–Schmidt theorem Theorem 12.44, proved in Appendix A.23, therefore supplies real numbers $\mu_n \longrightarrow 0$ and an orthonormal system $\{y_n\}$ with

$$Tf = \sum_n \mu_n\,\langle y_n, f\rangle_r\ y_n \qquad \text{for every } f \in L^2_r(a,b)\,. \tag{A.778}$$

The system is a basis. If f is orthogonal to every y_n then Equation (A.778) gives $Tf = 0$, and T is injective by Theorem A.469(3), so $f = 0$. By Definition 12.29 the system is maximal, i.e. an orthonormal basis, and Theorem 12.30 applies to it.

The eigenvalues are positive. Taking $f = y_n$ in Equation (A.778) gives $Ty_n = \mu_n y_n$, and $\mu_n = \langle Ty_n, y_n\rangle_r \geq 0$ by positivity; $\mu_n = 0$ is excluded by injectivity. Hence $\mu_n > 0$ and $\lambda_n = \mu_n^{-1} - K$ is well defined, with $\lambda_n \longrightarrow +\infty$ because $\mu_n \longrightarrow 0^+$.

The eigenfunctions solve the differential equation. From $y_n = \mu_n^{-1} T y_n$ and Theorem A.469(2), $y_n \in H_E$; in particular y_n is continuous, so Theorem A.469(1) applies to $f = y_n$ and makes $u = Ty_n = \mu_n y_n$ a classical solution of $L_K[u] = r\, y_n$ with $u(a) = u(b) = 0$. Dividing by μ_n,

$$L_K\,[y_n] = \mu_n^{-1} r\, y_n = (\lambda_n + K)\, r\, y_n\,,$$

which, on subtracting $K\, r\, y_n$ from both sides, is Equation (16.66) with eigenvalue λ_n and Dirichlet data.

Finite multiplicity and ordering. An eigenvalue λ of Equation (16.66) corresponds to the eigenvalue $\mu = (\lambda + K)^{-1}$ of T, whose eigenspace is finite dimensional by Theorem 12.44. Since $\lambda_n \longrightarrow +\infty$, only finitely many λ_n lie below any given bound, so the sequence may be reordered nondecreasingly, and the y_n with it. □

A.42.5 From the weighted norm to the energy norm

Lemma A.472 (The pairing identity)**.** *For every $\chi \in H_E$ and every n,*

$$B_K\,[y_n, \chi] = (\lambda_n + K)\,\langle y_n, \chi\rangle_r\,. \tag{A.779}$$

In particular $B_K\,[y_m, y_n] = (\lambda_n + K)\,\delta_{mn}$, and $\lambda_n + K > 0$ for every n.
Rests on Theorems A.469 and A.471.

Proof. *Derives Lemma A.472.*

Apply Equation (A.775) with $f = y_n$: $B_K\,[Ty_n, \chi] = \langle y_n, \chi\rangle_r$. Since $Ty_n = \mu_n y_n$ and B_K is bilinear, the left side is $\mu_n B_K\,[y_n, \chi]$; dividing by $\mu_n > 0$ and writing $\mu_n^{-1} = \lambda_n + K$ gives Equation (A.779). Taking $\chi = y_m$ and using orthonormality in $\langle\cdot,\cdot\rangle_r$ gives the second statement, and $\lambda_n + K = \mu_n^{-1} > 0$. □

Proof of Theorem A.461. *Derives Theorem A.461.*

Part (1) is Theorem A.471. For (2), put

$$\tilde{y}_n = \frac{y_n}{\sqrt{\lambda_n + K}}\,. \tag{A.780}$$

By Lemma A.472 the system $\{\tilde{y}_n\}$ is orthonormal for the inner product B_K:

$$B_K\,[\tilde{y}_m, \tilde{y}_n] = \frac{(\lambda_n + K)\,\delta_{mn}}{\sqrt{(\lambda_m + K)\,(\lambda_n + K)}} = \delta_{mn}\,.$$

It is maximal in H_E: if $\chi \in H_E$ satisfies $B_K\,[\tilde{y}_n, \chi] = 0$ for every n, then Equation (A.779) gives $\langle y_n, \chi\rangle_r = 0$ for every n, and $\{y_n\}$ is an orthonormal basis of L^2_r by part (1), so $\chi = 0$ as an element of L^2_r; being continuous, χ vanishes identically. Since (H_E, B_K) is a Hilbert space (Lemma A.464), maximality is completeness: Theorem 12.30 applies and $\{\tilde{y}_n\}$ is an orthonormal basis of H_E.

For (3), let $y \in H_E$ and $c_n = \langle y, y_n\rangle_r$. The energy-Fourier coefficients of y are, by Equation (A.779),

$$B_K\,[\tilde{y}_n, y] = \frac{(\lambda_n + K)\,\langle y_n, y\rangle_r}{\sqrt{\lambda_n + K}} = \sqrt{\lambda_n + K}\; c_n\,, \tag{A.781}$$

so the N-th partial sum of the expansion of y in the basis $\{\tilde{y}_n\}$ is

$$\sum_{n\leq N} B_K\,[\tilde{y}_n, y]\,\tilde{y}_n = \sum_{n\leq N} \sqrt{\lambda_n + K}\; c_n\, \frac{y_n}{\sqrt{\lambda_n + K}} = \sum_{n\leq N} c_n y_n = S_N\,,$$

exactly the partial sum formed in the chapter. By Equation (12.18) the expansion converges in the norm of H_E, which says $B_K\,[y - S_N, y - S_N] \longrightarrow 0$, and by Parseval's identity Equation (12.19) together with Equation (A.781),

$$B_K[y,y] = \sum_{n\geq 1} (\lambda_n + K)\, c_n^2\,, \tag{A.782}$$

the series converging because it is a Parseval sum. Finally, $\{y_n\}$ is an orthonormal basis of L^2_r, so Parseval's identity there gives $\langle y, y\rangle_r = \sum_n c_n^2$; subtracting K times that from Equation (A.782) and using $B_K[y,y] = \int \left(p\,y'^2 + q\,y^2\right) \mathrm{d}x + K\,\langle y, y\rangle_r$ leaves Equation (A.765). Both series converge separately — the first as a Parseval sum, the second because $\sum_n c_n^2 = \langle y, y\rangle_r < \infty$ — so the subtraction is legitimate. □

Remark A.473 (The chapter's inequality is untouched). Remark 16.77 insists that the *inequality* Equation (16.70), which is what the variational characterisation Theorem 16.76 rests on, is independent of the identity proved here, and that remains true: nothing above is used in the chapter's proof of Theorem 16.76, which argues from Equation (16.68) and the pairing Equation (16.69) to $B_K[y,y] \geq \sum_{n\leq N} (\lambda_n + K)\, c_n^2$ for every finite N — Bessel's inequality in the energy inner product, which needs no completeness at all. Conversely nothing in this appendix uses Theorem 16.76, so there is no circle. What this appendix does remove is a hypothesis: Theorem 16.76 assumes that the eigenfunctions are complete in L^2_r, and Theorem A.471 proves it. The difference between the two statements is exactly the difference between Bessel's inequality Equation (12.15) and Parseval's identity Equation (12.19), which is the four-way criterion Theorem 12.30 in the one place where the distinction has a consequence a reader can feel: by Corollary 16.78 the Rayleigh–Ritz error is second order in the deviation *measured in the energy norm*, and the estimate is vacuous without Equation (A.765).

Remark A.474 (What is quoted here). Two inputs are used and not proved in this section, and one further result is proved elsewhere in this appendix.

1. *The completeness of* L^2 (Theorem 12.12) and, with it, the density of the continuous functions in $L^2(a,b)$. The first is quoted in Chapter 12 itself (Remark 12.1) as resting on the convergence theorems of Lebesgue integration, which this treatise does not develop; the second belongs to the same package and is used twice above, both times to pass from continuous f to $f \in L^2_r$ in Theorem A.469. Reed and Simon [1481] is the reference of record for both.

2. *Nothing else.* The Hilbert–Schmidt theorem Theorem 12.44 is not an import: it is proved, from the numerical-radius formula and sequential compactness, in Appendix A.23 of this appendix. The Arzelà–Ascoli theorem used for the compactness of T is proved in Lemma A.454, and the existence and uniqueness theory for the second-order equation defining ϕ and ψ is Lemma A.439, both in this appendix.

Remark A.475. Theorem A.461 discharges the derivation owed in Remark 16.77 of Section 16.5.3: the eigenfunctions of a regular Sturm–Liouville problem are complete in the energy norm, and Equation (16.71) holds for every admissible trial function. The identity is used in the chapter to make the error of a Ritz estimate second order (Corollary 16.78), and through that corollary it underwrites the convergence of the Ritz scheme Equation (16.72) and of its descendants — the finite element method, the linear variational method for molecular orbitals, and the variational principle for the ground-state energy of Chapter 82. The eigenvalue problem itself is that of Chapter 9, whose orthogonality theorem Theorem 9.57 is here recovered as a by-product: eigenfunctions belonging to different eigenvalues are orthogonal in $\langle\cdot,\cdot\rangle_r$ because they are eigenvectors of a self-adjoint operator for different eigenvalues.

A.43 The Plateau Problem: the Architecture of Douglas' Proof

This appendix accompanies Theorem 16.81 of Chapter 16: every rectifiable Jordan curve in $\mathbb{R}^3$ bounds a disc-type minimal surface. It differs from every other section of this appendix in what it claims. Douglas' proof [678] and Radó's [667] are each of monograph length and neither is reproduced; what is given here is the *architecture* of Douglas' argument, with every step that can be carried out inside this treatise carried out in full and every step that cannot stated precisely, as a displayed theorem, and named as an import. Remark A.495 lists the imports exactly, and the reader who wants the honest one-sentence summary should read that remark first: this section reduces Theorem 16.81 to two named results it does not prove, namely the solvability of the Dirichlet problem for the disc by the Poisson integral and Douglas' inner-variation argument for conformality, together with the classification of the conformal automorphisms of the disc.

What *is* proved here, from scratch, is the chain that makes those two imports the only ones: that the Dirichlet integral dominates the area with equality exactly for conformal maps (Proposition A.478); that the Dirichlet integral is conformally invariant, which is what makes the non-compactness of the conformal group a real obstruction rather than a technicality (Lemma A.479); that Douglas' boundary functional equals the Dirichlet integral of the harmonic extension, with the constant made explicit (Theorem A.482); that it is lower semicontinuous (Lemma A.484); that the three-point normalisation restores compactness, through the Courant–Lebesgue estimate and the Arzelà–Ascoli theorem (Appendix A.43.5); and that a harmonic conformal map has vanishing mean curvature, so that a minimiser really is a minimal surface (Proposition A.493).

A.43.1 Setting and statement

Definition A.476 (Disc-type surfaces and the two functionals)**.** Let $B = \{z \in \mathbb{C} \mid |z| < 1\}$ be the open unit disc, with $z = u + \mathrm{i}v = \rho\,\mathrm{e}^{\mathrm{i}\theta}$, and $S^1 = \partial B$ its boundary circle. A *disc-type surface* is a map $\boldsymbol{X} : \overline{B} \longrightarrow \mathbb{R}^3$, continuous on $\overline{B}$ and of class C^1 on B. Its first fundamental coefficients (Definition 13.22) are

$$E = \boldsymbol{X}_u \cdot \boldsymbol{X}_u\,, \qquad F = \boldsymbol{X}_u \cdot \boldsymbol{X}_v\,, \qquad G = \boldsymbol{X}_v \cdot \boldsymbol{X}_v\,, \tag{A.783}$$

and its *area* and *Dirichlet* functionals are

$$A\,[\boldsymbol{X}] = \int_B \sqrt{EG - F^2}\,\mathrm{d}u\,\mathrm{d}v\,, \qquad D\,[\boldsymbol{X}] = \frac{1}{2}\int_B (E+G)\,\mathrm{d}u\,\mathrm{d}v\,. \tag{A.784}$$

$\boldsymbol{X}$ is *conformal* at a point if $E = G$ and $F = 0$ there.

Rests on Definition 13.22 and Equation (16.60).

Theorem A.477 (Douglas; Radó)**.** *Let $\Gamma \subset \mathbb{R}^3$ be a rectifiable Jordan curve. Then there exists a disc-type surface $\boldsymbol{X}$, harmonic and conformal on B, whose restriction to S^1 is a monotone parametrisation of Γ and which minimises the area among all disc-type surfaces spanning Γ. Its image is a minimal surface.*

Rests on Equation (16.7) and Theorem 16.70.

The strategy is not to minimise A, which is invariant under *all* reparametrisations of the disc and therefore has no compactness whatever, but to minimise D, which is invariant only under the conformal ones — and then to arrange that the minimiser be conformal, so that the two functionals agree on it by Proposition A.478. Douglas' innovation is to push the minimisation onto the boundary circle, where the competitors are parametrisations of Γ and the functional is Equation (A.786).

A.43.2 The Dirichlet integral dominates the area

Proposition A.478 ($D \geq A$, with equality exactly for conformal maps)**.** *For every disc-type surface, $D[\boldsymbol{X}] \geq A[\boldsymbol{X}]$, with equality if and only if $\boldsymbol{X}$ is conformal almost everywhere on B.*

Rests on Definition A.476.

Proof. *Derives Proposition A.478.*

Pointwise on B, the arithmetic–geometric mean inequality $\left(\sqrt{E}-\sqrt{G}\right)^2 \geq 0$ gives $E+G \geq 2\sqrt{EG}$, with equality if and only if $E=G$; and $\sqrt{EG} \geq \sqrt{EG-F^2}$, with equality if and only if $F=0$, since $F^2 \geq 0$ and both radicands are nonnegative — indeed $EG-F^2 = |\boldsymbol{X}_u|^2\,|\boldsymbol{X}_v|^2 - (\boldsymbol{X}_u \cdot \boldsymbol{X}_v)^2 = |\boldsymbol{X}_u \times \boldsymbol{X}_v|^2 \geq 0$ by Lagrange's identity. Chaining the two,

$$\frac{1}{2}(E+G) \geq \sqrt{EG} \geq \sqrt{EG-F^2}\,, \tag{A.785}$$

and integrating over B gives $D \geq A$. Equality of the integrals of two continuous functions ordered pointwise forces equality of the functions almost everywhere, and by the two equality cases just identified that means $E=G$ and $F=0$ almost everywhere, i.e. conformality. □

Lemma A.479 (Conformal invariance of the Dirichlet integral)**.** *Let $\tau : B \longrightarrow B$ be a bijective holomorphic map, written in real coordinates as $\tau(u,v) = (\sigma(u,v), \varrho(u,v))$. Then $D[\boldsymbol{X} \circ \tau] = D[\boldsymbol{X}]$ for every disc-type surface $\boldsymbol{X}$.*

Rests on Definition A.476 and Proposition 7.31.

Proof. *Derives Lemma A.479.*

Write $\boldsymbol{Y} = \boldsymbol{X} \circ \tau$. By the chain rule (Proposition 7.31), for each of the three components,

$$\boldsymbol{Y}_u = \sigma_u\,\boldsymbol{X}_\sigma + \varrho_u\,\boldsymbol{X}_\varrho\,, \qquad \boldsymbol{Y}_v = \sigma_v\,\boldsymbol{X}_\sigma + \varrho_v\,\boldsymbol{X}_\varrho\,,$$

the derivatives of $\boldsymbol{X}$ being evaluated at $\tau(u,v)$. The Cauchy–Riemann equations for τ read $\sigma_u = \varrho_v$ and $\sigma_v = -\varrho_u$, so with $J = \sigma_u^2 + \varrho_u^2$ — which is $|\tau'|^2$ and also the Jacobian determinant $\sigma_u\varrho_v - \sigma_v\varrho_u$ — one finds

$$\begin{aligned} |\boldsymbol{Y}_u|^2 + |\boldsymbol{Y}_v|^2 &= \left(\sigma_u^2+\sigma_v^2\right)|\boldsymbol{X}_\sigma|^2 + \left(\varrho_u^2+\varrho_v^2\right)|\boldsymbol{X}_\varrho|^2 + 2\left(\sigma_u\varrho_u + \sigma_v\varrho_v\right)\boldsymbol{X}_\sigma \cdot \boldsymbol{X}_\varrho \\ &= J\left(|\boldsymbol{X}_\sigma|^2 + |\boldsymbol{X}_\varrho|^2\right), \end{aligned}$$

because $\sigma_u^2+\sigma_v^2 = \varrho_u^2+\varrho_v^2 = J$ and $\sigma_u\varrho_u + \sigma_v\varrho_v = \sigma_u\varrho_u - \varrho_u\sigma_u = 0$ by Cauchy–Riemann. Hence the integrand of $D[\boldsymbol{Y}]$ is the integrand of $D[\boldsymbol{X}]$ composed with τ and multiplied by the Jacobian J, and the change-of-variables formula for a bijective C^1 substitution gives $D[\boldsymbol{Y}] = D[\boldsymbol{X}]$. □

Remark A.480 (Why the invariance is the whole difficulty)*.* The area functional is invariant under every diffeomorphism of the disc, an infinite-dimensional group, and a minimising sequence for A may therefore be dragged about by reparametrisations without any change in the value being minimised: no compactness argument can survive that. Passing to D cuts the invariance group down to the conformal automorphisms of B, which form only a three-parameter family — but it is still a *non-compact* one, and that residual non-compactness is exactly the difficulty Douglas' three-point normalisation removes. Lemma A.485 makes the statement precise.

A.43.3 The Douglas functional

Definition A.481 (Douglas' boundary functional)**.** For a bounded measurable $\boldsymbol{g} : S^1 \longrightarrow \mathbb{R}^3$, written as a function of the angle, $\boldsymbol{g} = \boldsymbol{g}(\theta)$, put

$$\mathcal{D}[\boldsymbol{g}] = \frac{1}{16\pi} \int_0^{2\pi} \int_0^{2\pi} \frac{|\boldsymbol{g}(\theta) - \boldsymbol{g}(\varphi)|^2}{\sin^2((\theta-\varphi)/2)} \, \mathrm{d}\theta \, \mathrm{d}\varphi \quad \in [0, +\infty] \, . \tag{A.786}$$

Rests on Definition A.476.

Theorem A.482 (Douglas' identity)**.** *Let $\boldsymbol{g} \in L^2(S^1; \mathbb{R}^3)$ have Fourier coefficients $\boldsymbol{c}_n = (2\pi)^{-1} \int_0^{2\pi} \boldsymbol{g}(\theta) \, \mathrm{e}^{-\mathrm{i}n\theta} \, \mathrm{d}\theta \in \mathbb{C}^3$, and let*

$$\boldsymbol{h}(\rho, \theta) = \sum_{n \in \mathbb{Z}} \boldsymbol{c}_n \, \rho^{|n|} \, \mathrm{e}^{\mathrm{i}n\theta} \tag{A.787}$$

be its harmonic extension to B. Then

$$\mathcal{D}[\boldsymbol{g}] = \pi \sum_{n \in \mathbb{Z}} |n| \, |\boldsymbol{c}_n|^2 = D[\boldsymbol{h}] \, , \tag{A.788}$$

both sides being $+\infty$ together.

Rests on Definition A.481 and Equation (12.19).

Proof. *Derives Theorem A.482.*

Throughout, $\left\{(2\pi)^{-1/2} \, \mathrm{e}^{\mathrm{i}n\theta}\right\}_{n \in \mathbb{Z}}$ is the standard orthonormal basis of $L^2(S^1)$ (Section 9.7.4), and Parseval's identity Equation (12.19) is applied componentwise to the three components of $\boldsymbol{g}$, so that $|\boldsymbol{c}_n|^2$ denotes $\sum_{k=1}^{3} \left|c_n^k\right|^2$.

Step 1: the Dirichlet integral of the extension. In polar coordinates $\mathrm{d}u \, \mathrm{d}v = \rho \, \mathrm{d}\rho \, \mathrm{d}\theta$ and

$$E + G = |\nabla \boldsymbol{h}|^2 = |\partial_\rho \boldsymbol{h}|^2 + \frac{1}{\rho^2} |\partial_\theta \boldsymbol{h}|^2 \, .$$

Differentiating Equation (A.787) term by term — legitimate on every disc $\rho \le \rho_0 < 1$, where the series and each of its derived series are dominated termwise by $|n|^2 \, |\boldsymbol{c}_n| \, \rho_0^{|n|-2}$, whose sum converges because $|\boldsymbol{c}_n|$ is bounded and $\rho_0 < 1$, so that the Weierstrass M-test Lemma 9.6 gives uniform convergence — and integrating in θ by orthogonality,

$$\int_0^{2\pi} |\partial_\rho \boldsymbol{h}|^2 \, \mathrm{d}\theta = 2\pi \sum_n n^2 \rho^{2|n|-2} \, |\boldsymbol{c}_n|^2 \, , \qquad \int_0^{2\pi} \frac{|\partial_\theta \boldsymbol{h}|^2}{\rho^2} \, \mathrm{d}\theta = 2\pi \sum_n n^2 \rho^{2|n|-2} \, |\boldsymbol{c}_n|^2 \, .$$

The two are equal — a first sign that the extension is conformally balanced — so

$$D[\boldsymbol{h}] = \frac{1}{2} \int_0^1 4\pi \sum_n n^2 \rho^{2|n|-1} \, |\boldsymbol{c}_n|^2 \, \mathrm{d}\rho = 2\pi \sum_n n^2 \, |\boldsymbol{c}_n|^2 \, \frac{1}{2\,|n|} = \pi \sum_n |n| \, |\boldsymbol{c}_n|^2 \, , \tag{A.789}$$

the term-by-term integration in ρ being legitimate because every term is nonnegative.

Step 2: the Douglas integral. Substitute $\theta = \varphi + t$ and integrate first in φ at fixed t. The Fourier coefficients of $\varphi \longmapsto \boldsymbol{g}(\varphi + t) - \boldsymbol{g}(\varphi)$ are $\boldsymbol{c}_n \, (\mathrm{e}^{\mathrm{i}nt} - 1)$, so Parseval Equation (12.19) gives

$$\int_0^{2\pi} |\boldsymbol{g}(\varphi + t) - \boldsymbol{g}(\varphi)|^2 \, \mathrm{d}\varphi = 2\pi \sum_n |\boldsymbol{c}_n|^2 \left|\mathrm{e}^{\mathrm{i}nt} - 1\right|^2 = 4\pi \sum_n |\boldsymbol{c}_n|^2 \, (1 - \cos nt) \, ,$$

whence

$$\mathcal{D}[\boldsymbol{g}] = \frac{1}{16\pi} \int_0^{2\pi} \frac{4\pi \sum_n |\boldsymbol{c}_n|^2 \, (1 - \cos nt)}{\sin^2(t/2)} \, \mathrm{d}t = \frac{1}{4} \sum_n |\boldsymbol{c}_n|^2 \int_0^{2\pi} \frac{1 - \cos nt}{\sin^2(t/2)} \, \mathrm{d}t \, , \tag{A.790}$$

the interchange of sum and integral being legitimate because every term is nonnegative.

Step 3: the kernel integral. Using $\sin^2(t/2) = (1-\cos t)/2$ and, for $n \geq 1$, the Dirichlet kernel $\Delta_n(t) = \sum_{j=0}^{n-1} \mathrm{e}^{\mathrm{i}jt} = (\mathrm{e}^{\mathrm{i}nt}-1)/(\mathrm{e}^{\mathrm{i}t}-1)$,

$$\frac{1-\cos nt}{\sin^2(t/2)} = \frac{2(1-\cos nt)}{1-\cos t} = 2\,\frac{\left|\mathrm{e}^{\mathrm{i}nt}-1\right|^2}{\left|\mathrm{e}^{\mathrm{i}t}-1\right|^2} = 2\,\left|\Delta_n(t)\right|^2 , \tag{A.791}$$

where $\left|\mathrm{e}^{\mathrm{i}\alpha}-1\right|^2 = 2(1-\cos\alpha)$ has been used twice. By orthogonality of the exponentials, $\int_0^{2\pi} |\Delta_n|^2 \,\mathrm{d}t = 2\pi n$, so the integral in Equation (A.790) equals $4\pi|n|$ — the case $n=0$ giving 0, and negative n the same value as $|n|$ because cos is even. Substituting,

$$\mathcal{D}[\boldsymbol{g}] = \frac{1}{4}\sum_n |\boldsymbol{c}_n|^2 \, 4\pi|n| = \pi \sum_n |n|\,|\boldsymbol{c}_n|^2 ,$$

which with Equation (A.789) is Equation (A.788). Both computations produce the same nonnegative series, so one side is infinite exactly when the other is. □

Remark A.483 (What the identity accomplishes). Equation (A.788) converts a variational problem over maps of a two-dimensional domain into one over functions on a circle: by Proposition A.478 the area of a disc-type surface is at most its Dirichlet integral, and by Equation (A.788) the Dirichlet integral of the harmonic extension of a given boundary parametrisation is computable from that parametrisation alone. The harmonic extension is moreover the *minimiser* of D among all extensions of $\boldsymbol{g}$, which is the Dirichlet principle of Definition 16.67 for the disc; so the boundary functional $\mathcal{D}$ is exactly the value of the two-dimensional problem, and Douglas' competitors are parametrisations, not surfaces.

Lemma A.484 (Lower semicontinuity of the Douglas functional). *Let $\boldsymbol{g}_k \longrightarrow \boldsymbol{g}$ pointwise almost everywhere on S^1. Then $\mathcal{D}[\boldsymbol{g}] \leq \liminf_k \mathcal{D}[\boldsymbol{g}_k]$.*

Rests on Definition A.481.

Proof. *Derives Lemma A.484.*

The integrand of Equation (A.786) is nonnegative, and for almost every pair (θ,φ) it converges to the integrand built from $\boldsymbol{g}$, the denominator being independent of k and the numerator continuous in the values of the map. Fatou's lemma — quoted, as throughout this appendix (Remark A.459) — applied on the square $[0,2\pi]^2$ gives the claim. □

A.43.4 The conformal group and the three-point normalisation

Lemma A.485 (Conformal automorphisms of the disc). *For $\alpha \in B$ and $\vartheta \in \mathbb{R}$ define*

$$M_{\alpha,\vartheta}(z) = \mathrm{e}^{\mathrm{i}\vartheta}\,\frac{z-\alpha}{1-\overline{\alpha}\,z} . \tag{A.792}$$

Each $M_{\alpha,\vartheta}$ is a holomorphic bijection of B onto itself, extending to a homeomorphism of $\overline{B}$ that maps S^1 onto S^1; these maps form a group under composition; the group is not *compact; and it acts* simply transitively *on the ordered triples of distinct points of S^1 that are in positive cyclic order. The first three assertions are proved below; the fourth is quoted (Remark A.495).*

Rests on Lemma A.479.

Proof. *Derives Lemma A.485.*

The circle is preserved. For $|z|=1$, $\overline{z} = z^{-1}$, so

$$|1-\overline{\alpha}z| = |z|\,\left|z^{-1}-\overline{\alpha}\right| = |\overline{z}-\overline{\alpha}| = |z-\alpha| ,$$

whence $|M_{\alpha,\vartheta}(z)| = 1$. The denominator never vanishes on $\overline{B}$, since $|\overline{\alpha}z| \leq |\alpha| < 1$, so $M_{\alpha,\vartheta}$ is holomorphic on a neighbourhood of $\overline{B}$ and continuous on $\overline{B}$.

The disc is preserved, and the map is bijective. A direct computation gives, for $|z| < 1$,

$$1 - |M_{\alpha,\vartheta}(z)|^2 = \frac{|1-\overline{\alpha}z|^2 - |z-\alpha|^2}{|1-\overline{\alpha}z|^2} = \frac{\left(1-|\alpha|^2\right)\left(1-|z|^2\right)}{|1-\overline{\alpha}z|^2} > 0\,,$$

the middle equality by expanding both squared moduli, so $M_{\alpha,\vartheta}$ maps B into B. Its inverse is $M_{\beta,\eta}$ with $\beta = -\mathrm{e}^{-\mathrm{i}\vartheta}\alpha$ suitably normalised — explicitly, solving $w = \mathrm{e}^{\mathrm{i}\vartheta}(z-\alpha)/(1-\overline{\alpha}z)$ for z gives $z = \left(\mathrm{e}^{-\mathrm{i}\vartheta}w + \alpha\right)/\left(1+\overline{\alpha}\mathrm{e}^{-\mathrm{i}\vartheta}w\right)$, which is again of the form Equation (A.792); so each $M_{\alpha,\vartheta}$ is a bijection of B and of S^1, and the family is closed under inversion. It is closed under composition as well, since a composition of two quotients of linear polynomials is again one and it maps B bijectively onto B — that such a map must again be of the form Equation (A.792) is the classification quoted in the fourth assertion.

Non-compactness. Take $\vartheta = 0$ and $\alpha = 1 - 1/k$ along the real axis. For any fixed $z \in S^1$ with $z \neq 1$,

$$M_{\alpha,0}(z) = \frac{z-\alpha}{1-\alpha z} \longrightarrow \frac{z-1}{1-z} = -1 \qquad \text{as } k \longrightarrow \infty\,,$$

whereas $M_{\alpha,0}(1) = 1$ for every k. The sequence therefore has no subsequence converging uniformly on S^1 to a homeomorphism: its pointwise limit collapses the whole circle except one point onto the single value -1. This is precisely the degeneration a minimising sequence for $\mathcal{D}$ may undergo at no cost, since by Lemma A.479 and Equation (A.788) the value of $\mathcal{D}$ is unchanged by such a reparametrisation of the boundary. □

Definition A.486 (The normalised admissible class)**.** Fix a homeomorphism $\gamma : S^1 \longrightarrow \Gamma$ and three points $\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3$ on Γ, in positive cyclic order. A *monotone parametrisation* of Γ is a map $\boldsymbol{g} = \gamma \circ \psi$, where $\psi : [0, 2\pi] \longrightarrow \mathbb{R}$ is continuous and nondecreasing with $\psi(2\pi) = \psi(0) + 2\pi$, angles being read modulo 2π. The *normalised class* $\mathcal{C}^*(\Gamma)$ consists of those monotone parametrisations with

$$\boldsymbol{g}(1) = \boldsymbol{Q}_1\,, \qquad \boldsymbol{g}(\mathrm{i}) = \boldsymbol{Q}_2\,, \qquad \boldsymbol{g}(-1) = \boldsymbol{Q}_3\,. \tag{A.793}$$

Rests on Lemma A.485 and Definition A.481.

Lemma A.487 (The normalisation costs nothing)**.** $\inf\{\mathcal{D}[\boldsymbol{g}] \mid \boldsymbol{g} \in \mathcal{C}^*(\Gamma)\} = \inf\{\mathcal{D}[\boldsymbol{g}] \mid \boldsymbol{g}$ *a monotone parametrisation of* $\Gamma\}$.

Rests on Lemmas A.479 and A.485.

Proof. *Derives Lemma A.487.*

The class $\mathcal{C}^*$ is contained in the larger class, so the left infimum is at least the right one. Conversely let $\boldsymbol{g}$ be any monotone parametrisation. The three points $\boldsymbol{g}^{-1}(\boldsymbol{Q}_i)$ may be chosen as three points $\zeta_1, \zeta_2, \zeta_3$ of S^1 in positive cyclic order, because $\boldsymbol{g}$ is monotone of degree one; by the simple transitivity in Lemma A.485 there is a conformal automorphism M of B carrying $(1, \mathrm{i}, -1)$ to $(\zeta_1, \zeta_2, \zeta_3)$. Then $\boldsymbol{g} \circ M$ is again a monotone parametrisation, now satisfying Equation (A.793), and its harmonic extension is $\boldsymbol{h} \circ M$ — the composition of a harmonic map with a holomorphic one is harmonic, since $\nabla^2(\boldsymbol{h} \circ M) = |M'|^2 (\nabla^2 \boldsymbol{h}) \circ M$ by the computation of Lemma A.479 applied to second derivatives, and its boundary values are $\boldsymbol{g} \circ M$, so it *is* the harmonic extension of $\boldsymbol{g} \circ M$ by the uniqueness of the solution of the Dirichlet problem (Corollary 10.76). Hence, by Equation (A.788) and Lemma A.479,

$$\mathcal{D}[\boldsymbol{g} \circ M] = D[\boldsymbol{h} \circ M] = D[\boldsymbol{h}] = \mathcal{D}[\boldsymbol{g}]\,,$$

so every value attained on the larger class is attained on $\mathcal{C}^*$. □

A.43.5 Compactness: the Courant–Lebesgue estimate

Lemma A.488 (Small chords cut off small arcs)**.** *Let Γ be a Jordan curve with homeomorphic parametrisation $\gamma : S^1 \longrightarrow \Gamma$. For every $\varepsilon > 0$ there is $\delta > 0$ such that any two points $\boldsymbol{P}, \boldsymbol{P}'$ of Γ with $|\boldsymbol{P} - \boldsymbol{P}'| < \delta$ divide Γ into two arcs, at least one of which has diameter less than ε.*

Rests on Theorems 6.31 and 7.25.

Proof. *Derives Lemma A.488.*

By definition a Jordan curve is the image of a homeomorphism γ of S^1, so γ^{-1} exists and is continuous, and Γ is compact, being a continuous image of a compact set.

Both maps are uniformly continuous. For γ this is Theorem 7.25 applied componentwise to the angle parametrisation on $[0, 2\pi]$. For γ^{-1}, suppose not: there are $\varepsilon_0 > 0$ and points $\boldsymbol{P}_k, \boldsymbol{P}'_k \in \Gamma$ with $|\boldsymbol{P}_k - \boldsymbol{P}'_k| \longrightarrow 0$ and $|\gamma^{-1}(\boldsymbol{P}_k) - \gamma^{-1}(\boldsymbol{P}'_k)| \geq \varepsilon_0$. The circle is compact, hence sequentially compact (Theorem 6.31), so along a subsequence $\gamma^{-1}(\boldsymbol{P}_k) \longrightarrow \zeta$ and $\gamma^{-1}(\boldsymbol{P}'_k) \longrightarrow \zeta'$ with $|\zeta - \zeta'| \geq \varepsilon_0$, so $\zeta \neq \zeta'$. Applying the continuous γ, $\boldsymbol{P}_k \longrightarrow \gamma(\zeta)$ and $\boldsymbol{P}'_k \longrightarrow \gamma(\zeta')$, and $|\boldsymbol{P}_k - \boldsymbol{P}'_k| \longrightarrow 0$ then forces $\gamma(\zeta) = \gamma(\zeta')$, contradicting injectivity.

Given $\varepsilon > 0$, uniform continuity of γ supplies $\eta > 0$ such that arcs of S^1 of length below η have images of diameter below ε; uniform continuity of γ^{-1} then supplies $\delta > 0$ such that $|\boldsymbol{P} - \boldsymbol{P}'| < \delta$ forces $\gamma^{-1}(\boldsymbol{P})$ and $\gamma^{-1}(\boldsymbol{P}')$ to be at distance below $\eta/(2\pi)$ along S^1, so that the shorter of the two arcs they determine has length below η. Its image is one of the two arcs of Γ, of diameter below ε. □

Lemma A.489 (Courant–Lebesgue)**.** *Let $\boldsymbol{h}$ be harmonic on B and continuous on $\overline{B}$ with $D[\boldsymbol{h}] \leq M$, let $z_0 \in S^1$, and let $0 < \delta < 1$. Write $C_\varsigma = \{z \in B \mid |z - z_0| = \varsigma\}$. Then there is a radius $\varsigma \in \left(\delta, \sqrt{\delta}\right)$ for which the image $\boldsymbol{h}(C_\varsigma)$ has diameter at most*

$$\left(\frac{4\pi M}{\log(1/\delta)}\right)^{1/2}. \tag{A.794}$$

Rests on Definition A.476 and Lemma A.452.

Proof. *Derives Lemma A.489.*

Use polar coordinates (ς, ω) centred at z_0, in which $\mathrm{d}u\,\mathrm{d}v = \varsigma\,\mathrm{d}\varsigma\,\mathrm{d}\omega$ and $|\nabla \boldsymbol{h}|^2 = |\partial_\varsigma \boldsymbol{h}|^2 + \varsigma^{-2}|\partial_\omega \boldsymbol{h}|^2$. Writing I_ς for the set of angles ω with $z_0 + \varsigma \mathrm{e}^{\mathrm{i}\omega} \in B$ and discarding the radial term,

$$2M \geq 2D[\boldsymbol{h}] \geq \int_\delta^{\sqrt{\delta}} \left(\int_{I_\varsigma} |\partial_\omega \boldsymbol{h}|^2 \,\mathrm{d}\omega\right) \frac{\mathrm{d}\varsigma}{\varsigma} \equiv \int_\delta^{\sqrt{\delta}} \zeta(\varsigma)\,\frac{\mathrm{d}\varsigma}{\varsigma}.$$

Since $\int_\delta^{\sqrt{\delta}} \mathrm{d}\varsigma/\varsigma = \tfrac{1}{2}\log(1/\delta)$, the mean value of ζ against the measure $\mathrm{d}\varsigma/\varsigma$ on that range is at most $4M/\log(1/\delta)$, so there exists $\varsigma \in \left(\delta, \sqrt{\delta}\right)$ with $\zeta(\varsigma) \leq 4M/\log(1/\delta)$ — otherwise the integral would exceed $2M$. For that radius the length of the image curve is bounded by the Cauchy–Schwarz inequality Equation (A.750),

$$\int_{I_\varsigma} |\partial_\omega \boldsymbol{h}| \,\mathrm{d}\omega \leq |I_\varsigma|^{1/2} \left(\int_{I_\varsigma} |\partial_\omega \boldsymbol{h}|^2 \,\mathrm{d}\omega\right)^{1/2} \leq (2\pi)^{1/2} \left(\frac{4M}{\log(1/\delta)}\right)^{1/2},$$

which is Equation (A.794), and the diameter of a curve is at most its length. □

Theorem A.490 (Equicontinuity of the normalised class)**.** *Let $(\boldsymbol{g}_k) \subset \mathcal{C}^*(\Gamma)$ satisfy $\mathcal{D}[\boldsymbol{g}_k] \leq M$ for every k. Then the family is equicontinuous on S^1, and a subsequence converges uniformly to some $\boldsymbol{g} \in \mathcal{C}^*(\Gamma)$.*

Rests on Lemmas A.454, A.488 and A.489.

Proof. *Derives Theorem A.490.*

Equicontinuity. Let $\varepsilon > 0$ and take δ_0 from Lemma A.488 for this ε; shrink ε if necessary so that 2ε is smaller than the mutual distances of $\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3$ and of the arcs between them. Choose $\delta \in (0,1)$ with $(4\pi M/\log(1/\delta))^{1/2} < \delta_0$ and $\sqrt{\delta}$ smaller than one third of the least of those arc lengths. Let $\boldsymbol{h}_k$ be the harmonic extension of $\boldsymbol{g}_k$, which by Equation (A.788) has $D[\boldsymbol{h}_k] = \mathcal{D}[\boldsymbol{g}_k] \le M$, and which is continuous on $\overline{B}$ with boundary values $\boldsymbol{g}_k$ by the quoted solvability of the Dirichlet problem (Theorem A.491). For $z_0 \in S^1$, Lemma A.489 produces a radius $\varsigma_k \in \left(\delta, \sqrt{\delta}\right)$ for which the arc C_{ς_k} has image of diameter below δ_0; its two endpoints lie on S^1 and their images $\boldsymbol{P}, \boldsymbol{P}'$ are values of $\boldsymbol{g}_k$ at distance below δ_0. By Lemma A.488 one of the two arcs of Γ they cut off has diameter below ε; it is the one whose preimage under the monotone map $\boldsymbol{g}_k$ is the short boundary arc cut off by C_{ς_k}, because the other boundary arc has length above $2\pi - 2\sqrt{\delta}$ and its image therefore contains at least two of $\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3$ — which by the choice of ε cannot lie in a set of diameter ε. Hence $\boldsymbol{g}_k$ oscillates by less than ε on the boundary arc cut off at z_0, which contains all points of S^1 within δ of z_0; and δ was chosen independently of k and of z_0. That is equicontinuity.

Extraction of a limit. Write $\boldsymbol{g}_k = \gamma \circ \psi_k$ as in Definition A.486. Equicontinuity of $\boldsymbol{g}_k$ transfers to ψ_k through the uniform continuity of γ^{-1} (Lemma A.488 and its proof), and the ψ_k are uniformly bounded once normalised by $\psi_k(0) \in [0, 2\pi)$. The Arzelà–Ascoli theorem Lemma A.454 gives a uniformly convergent subsequence $\psi_k \longrightarrow \psi$; the limit is continuous, nondecreasing and satisfies $\psi(2\pi) = \psi(0) + 2\pi$, these being closed conditions under uniform convergence. Hence $\boldsymbol{g} = \gamma \circ \psi$ is again a monotone parametrisation, it satisfies Equation (A.793) because each $\boldsymbol{g}_k$ does and convergence is pointwise, and $\boldsymbol{g}_k \longrightarrow \boldsymbol{g}$ uniformly by the uniform continuity of γ. □

A.43.6 The two quoted inputs, and the conclusion

Theorem A.491 (Dirichlet problem for the disc; quoted)**.** *For every continuous $\boldsymbol{g} : S^1 \longrightarrow \mathbb{R}^3$ the Poisson integral*

$$\boldsymbol{h}(\rho,\theta) = \frac{1}{2\pi}\int_0^{2\pi} \frac{1-\rho^2}{1-2\rho\cos(\theta-\varphi)+\rho^2}\,\boldsymbol{g}(\varphi)\,\mathrm{d}\varphi \tag{A.795}$$

is harmonic on B, continuous on $\overline{B}$, and equal to $\boldsymbol{g}$ on S^1; it coincides with the series Equation (A.787), and it minimises D among all disc-type surfaces with boundary values $\boldsymbol{g}$. This treatise does not prove it: Chapter 10 develops the maximum principles and the fundamental solution but not the Poisson kernel of the disc, and Chapter 8 explicitly reserves conformal mapping. See [1222].

Rests on Definitions 10.66 and 16.67.

Theorem A.492 (Douglas' conformality theorem; quoted)**.** *Let $\boldsymbol{g}_*$ minimise $\mathcal{D}$ over $\mathcal{C}^*(\Gamma)$ and let $\boldsymbol{X} = \boldsymbol{h}_{\boldsymbol{g}_*}$ be its harmonic extension. Then $\boldsymbol{X}$ is conformal on B, that is $E = G$ and $F = 0$, except possibly at isolated branch points where $\boldsymbol{X}_u = \boldsymbol{X}_v = \boldsymbol{0}$. Equivalently, the Hopf differential $\Phi = |\boldsymbol{X}_u|^2 - |\boldsymbol{X}_v|^2 - 2\mathrm{i}\,\boldsymbol{X}_u \cdot \boldsymbol{X}_v$ vanishes identically. The proof is by* inner *variations — reparametrising the disc rather than moving the surface, which is legitimate precisely because $\mathcal{C}^*$ is a full gauge slice by Lemma A.487 — and it is not reproduced here [678].*

Rests on Definition A.481 and Lemma A.487.

Proposition A.493 (A harmonic conformal map is a minimal surface)**.** *Let $\boldsymbol{X}$ be harmonic and conformal on B, with $E = G = \Lambda^2 > 0$ and $F = 0$ at a point. Then the mean curvature $H = \frac{1}{2} g^{ij} b_{ij}$ of the surface vanishes there. Conversely a conformal parametrisation of a surface of vanishing mean curvature is harmonic.*

Rests on Definition 13.25 and Proposition 16.8.

Proof. *Derives Proposition A.493.*

Work at a point where $E = G = \Lambda^2 > 0$ and $F = 0$, with unit normal $\hat{n}$ (Equation (13.80)). Differentiating the conformality relations,

$$\boldsymbol{X}_{uu} \cdot \boldsymbol{X}_u = \tfrac{1}{2} E_u \,, \qquad \boldsymbol{X}_{vv} \cdot \boldsymbol{X}_u = \partial_v \left(\boldsymbol{X}_v \cdot \boldsymbol{X}_u \right) - \boldsymbol{X}_v \cdot \boldsymbol{X}_{uv} = F_v - \tfrac{1}{2} G_u = -\tfrac{1}{2} E_u \,,$$

using $F \equiv 0$ and $G \equiv E$. Adding, $(\boldsymbol{X}_{uu} + \boldsymbol{X}_{vv}) \cdot \boldsymbol{X}_u = 0$, and the same computation with u and v exchanged gives $(\boldsymbol{X}_{uu} + \boldsymbol{X}_{vv}) \cdot \boldsymbol{X}_v = 0$. So the Laplacian $\nabla^2 \boldsymbol{X} = \boldsymbol{X}_{uu} + \boldsymbol{X}_{vv}$ is orthogonal to both tangent vectors, hence parallel to $\hat{n}$, and by Equation (13.80) its normal component is

$$\hat{n} \cdot \nabla^2 \boldsymbol{X} = b_{11} + b_{22} \,.$$

Since $F = 0$ and $E = G = \Lambda^2$, the inverse metric is $g^{ij} = \Lambda^{-2} \delta^{ij}$ (Definition 13.23), so $2H = g^{ij} b_{ij} = \Lambda^{-2} \left(b_{11} + b_{22} \right)$ and therefore

$$\nabla^2 \boldsymbol{X} = 2 \Lambda^2 H \, \hat{n} \,. \tag{A.796}$$

If $\boldsymbol{X}$ is harmonic the left side vanishes and, Λ^2 being positive, $H = 0$; conversely $H = 0$ makes the right side vanish. By Proposition 16.8 the vanishing of the mean curvature is the minimal-surface equation Equation (16.7) wherever the surface is a graph, so the image is a minimal surface in the sense of Section 16.1.3. □

Proof of Theorem A.477, granting the two quoted theorems. *Derives Theorem A.477.*

Let $m = \inf \{ \mathcal{D}[\boldsymbol{g}] \mid \boldsymbol{g} \in \mathcal{C}^*(\Gamma) \}$. It is finite: Γ is rectifiable, and a parametrisation proportional to arc length has $|\boldsymbol{g}(\theta) - \boldsymbol{g}(\varphi)| \le \ell \, |\theta - \varphi| \, / \, (2\pi)$ with ℓ the length of Γ, so the integrand of Equation (A.786) is bounded by $\ell^2 \, |\theta - \varphi|^2 \, / \, (4\pi^2 \sin^2 ((\theta - \varphi)/2))$, which is bounded on the square, and $\mathcal{D} < \infty$ there; this is the only place where rectifiability is used, and it is used exactly to make the infimum finite.

Take a minimising sequence $(\boldsymbol{g}_k) \subset \mathcal{C}^*$, which by discarding finitely many terms satisfies $\mathcal{D}[\boldsymbol{g}_k] \le m + 1$. By Theorem A.490 a subsequence converges uniformly to some $\boldsymbol{g}_* \in \mathcal{C}^*(\Gamma)$, and by lower semicontinuity Lemma A.484, $\mathcal{D}[\boldsymbol{g}_*] \le \liminf_k \mathcal{D}[\boldsymbol{g}_k] = m$; the reverse inequality holds because $\boldsymbol{g}_* \in \mathcal{C}^*$. So the infimum is attained.

Let $\boldsymbol{X}$ be the harmonic extension of $\boldsymbol{g}_*$, which exists and is continuous on $\overline{B}$ by Theorem A.491. By Theorem A.492 it is conformal, so by Proposition A.478 $A[\boldsymbol{X}] = D[\boldsymbol{X}] = \mathcal{D}[\boldsymbol{g}_*] = m$, using Equation (A.788) for the last equality.

It remains to see that no disc-type surface spanning Γ has smaller area. Let $\boldsymbol{Y}$ be one, with boundary parametrisation $\boldsymbol{g}$, which by Lemma A.487 may be taken in $\mathcal{C}^*(\Gamma)$. The area is *reparametrisation invariant*, since $\sqrt{EG - F^2}$ transforms with the Jacobian under any C^1 change of variables, and by Proposition A.478 $A[\boldsymbol{Y}] = A[\boldsymbol{Y} \circ \tau] \le D[\boldsymbol{Y} \circ \tau]$ for every such τ; the quoted approximation step — item (4) of Remark A.495 — upgrades this to the equality

$$A[\boldsymbol{Y}] = \inf_{\tau} D[\boldsymbol{Y} \circ \tau] \,, \tag{A.797}$$

the infimum being over the reparametrisations of the disc. Each $\boldsymbol{Y} \circ \tau$ has the same boundary values as $\boldsymbol{Y}$ up to a reparametrisation of S^1, and by the Dirichlet principle of Theorem A.491 its Dirichlet integral is at least that of the harmonic extension of those boundary values, which by Equation (A.788) and Lemma A.487 is $\mathcal{D}[\boldsymbol{g}] \ge m$. Hence $A[\boldsymbol{Y}] \ge m = A[\boldsymbol{X}]$: the surface $\boldsymbol{X}$ minimises area among disc-type surfaces spanning Γ. Finally Proposition A.493 makes the image a minimal surface away from branch points. □

Remark A.494 (Radó's route, for contrast). Radó's independent proof [667], published a year before Douglas', never forms a boundary functional. It approximates Γ by inscribed *polygons* Γ_k; for a polygon the minimal surface is produced by solving finitely many Dirichlet problems on the disc and controlling the resulting sequence with the maximum principle, and

the areas A_k converge to the infimum for Γ. The compactness that Douglas buys with the three-point normalisation is bought here by a condition on the curve — Radó first assumes Γ bounds *some* surface of finite area, which for a rectifiable curve is automatic — and the limit surface is obtained from the Γ_k by a normal-family argument in the function-theoretic sense. The two proofs meet at the same place: both must produce a conformal parametrisation, and neither can avoid the Dirichlet principle for the disc. What Douglas gains is that his functional is intrinsic to the boundary, which is why it, and not Radó's construction, generalises to higher connectivity and higher genus.

Remark A.495 (What is quoted here). This section does not prove Theorem 16.81. It reduces it to named results, and the reduction is only worth what the honesty of this list is worth.

1. *The Dirichlet problem for the disc* (Theorem A.491): that the Poisson integral of a continuous boundary function is harmonic inside, continuous up to the boundary, and minimises the Dirichlet integral among all extensions. Used twice — in Theorem A.490 to have a continuous harmonic extension to apply Courant–Lebesgue to, and in the final proof. Not proved anywhere in this treatise.

2. *Douglas' conformality theorem* (Theorem A.492): that a minimiser of $\mathcal{D}$ over the normalised class has a conformal harmonic extension. This is the technical heart of the whole subject and the step for which Douglas was awarded one of the first two Fields Medals; it is stated precisely above and not proved [678].

3. *Simple transitivity of the conformal group on boundary triples* (the fourth assertion of Lemma A.485), which belongs to the conformal mapping theory that Chapter 8 explicitly reserves. Everything else in that lemma — that the maps Equation (A.792) are automorphisms and that the family is non-compact — is proved.

4. *The comparison of A with D over reparametrisations* used at the end of the final proof, i.e. that the infimum of D over all reparametrisations of a given disc-type surface equals its area. This is Douglas' and Radó's approximation argument and is quoted with them.

5. *Lebesgue integration*, as everywhere in this appendix: Fatou's lemma in Lemma A.484, and the polar- coordinate rewriting of a double integral in Lemma A.489.

Proved here in full, and used above: the pointwise inequality $D \geq A$ with its equality case, the conformal invariance of D, the identity $\mathcal{D} = D$ of the harmonic extension with its constant, lower semicontinuity, the group property and non-compactness of the Möbius family, the arc lemma, the Courant–Lebesgue estimate, the equicontinuity of the normalised class and the extraction of a limit, and the identity Equation (A.796) that turns a harmonic conformal map into a minimal surface. The reader should also carry forward Remark 16.82 of the chapter, which is about what the *theorem* claims — disc type, minimisation among disc-type maps, possible branch points — as opposed to what this section proves about it.

Remark A.496. This section accompanies Theorem 16.81 of Section 16.5.4, replacing the pending derivation there with a proof architecture and an explicit list of imports rather than with a complete derivation — the one place in this appendix where that is so, and it is flagged as such in Remark A.495. The minimal surfaces the theorem produces are the soap films of Section 16.1.3, whose equation Equation (16.7) was derived there for a graph and is recovered here in the conformal form Equation (A.796); the differential geometry is that of Chapter 13, and the physics of the surface tension that realises the area as an energy belongs to Chapter 31.

A.44 Change of Variables in a Multiple Integral

This appendix proves Theorem 7.97 of Chapter 7: that a C^1 change of coordinates $\boldsymbol{\Phi}$ transports a multiple integral with the modulus of its Jacobian determinant as the weight, Equation (7.113). It is the one property of the multiple integral that the physical parts use on almost every page — every passage to polar, cylindrical or spherical coordinates is an instance — and it is the only one Chapter 7 states without proof.

The derivation is elementary throughout: no measure theory, no Lebesgue convergence theorem, no fixed-point theorem, and nothing from outside this treatise. Two ingredients that the chapters state only in a weaker form are proved here rather than quoted — the multiplicativity of the determinant, which Chapter 5 does not record, and iterated integration in $\mathbb{R}^N$, which Remark 7.96 states for $N = 2$ and $N = 3$. One ingredient is genuinely constructed here and was not foreseen in the plan for this section: a *continuous* partition of unity, built from distance functions in Lemma A.505. Without it the passage from a local statement to a global one cannot be made for a Riemann integral over a general region, and saying that no partition of unity is used would be false. It costs half a page and quotes nothing; see Remark A.515.

Throughout, points of $\mathbb{R}^N$ are written $\boldsymbol{u} = (u^1, \dots, u^N)$, and when the last coordinate is singled out we write $\boldsymbol{u} = (\boldsymbol{u}', t)$ with $\boldsymbol{u}' \in \mathbb{R}^{N-1}$ and $t \in \mathbb{R}$. A *box* is a product $[a_1, b_1] \times \cdots \times [a_N, b_N]$ and a *cube* is a box with all edges equal. The support of a function f on an open $V \subseteq \mathbb{R}^N$ is the closure of $\{\boldsymbol{x} \mid f(\boldsymbol{x}) \neq 0\}$; we write $f \in C_c(V)$ when f is continuous and that closure is a compact subset of V, in which case the extension of f by zero is continuous on the whole of $\mathbb{R}^N$ and vanishes off a box, so $\int_{\mathbb{R}^N} f$ means the integral of that extension over any box containing the support (Definition 7.93), and is independent of the box.

A.44.1 What the derivation uses

From Chapter 7: the Darboux construction of the multiple integral (Definition 7.93), the integrability of a continuous function on a box (Theorem 7.40), uniform continuity on a compact set (Theorem 7.25), the one-variable substitution rule Equation (7.27), the mean value theorem (Theorem 7.35), the chain rule (Proposition 7.72) and the inverse function theorem (Corollary 7.81). From Chapter 5: the Leibniz expansion Equation (5.19) and the two properties read off it there — linearity in each column and antisymmetry under exchange of two columns. From Chapter 6: compactness (Definition 6.9), the Heine–Borel theorem in $\mathbb{R}^N$ (Theorem 6.12) and the Lebesgue number lemma (Lemma 6.30).

Lemma A.497 (The determinant is multiplicative)**.** *For real $N \times N$ matrices A and B, $\det(AB) = \det A \det B$.*

Rests on Equation (5.19).

Proof. *Derives Lemma A.497.*

Write $(AB)^i{}_j = \sum_k A^i{}_k B^k{}_j$ and expand Equation (5.19), distributing the product over j:

$$\det(AB) = \sum_{\sigma} \operatorname{sgn}(\sigma) \prod_{j=1}^{N} \sum_{k_j=1}^{N} A^{\sigma(j)}{}_{k_j} B^{k_j}{}_j = \sum_{k_1,\dots,k_N} \left(\prod_{j=1}^{N} B^{k_j}{}_j \right) \sum_{\sigma} \operatorname{sgn}(\sigma) \prod_{j=1}^{N} A^{\sigma(j)}{}_{k_j} , \tag{A.798}$$

the interchange being a finite rearrangement. The inner sum is, by Equation (5.19) again, the determinant of the matrix whose j-th column is the k_j-th column of A. If two of the k_j coincide that matrix has a repeated column and the determinant vanishes, so only the tuples with $k_j = \tau(j)$ for a permutation τ survive; and bringing the columns back into their

natural order costs $\operatorname{sgn}(\tau)$, by the antisymmetry recorded after Equation (5.19). Hence the inner sum equals $\operatorname{sgn}(\tau)\det A$ and

$$\det(AB) = \det A \sum_{\tau} \operatorname{sgn}(\tau) \prod_{j=1}^{N} B^{\tau(j)}{}_{j} = \det A \, \det B \,. \tag{A.799}$$

□

A.44.2 Sets of zero content

The Darboux construction ignores sets that can be covered by finitely many boxes of arbitrarily small total volume. That notion — *zero content* — is strictly stronger than the measure zero of Section 7.12, which allows countably many covering sets; it is the one a Riemann sum can see, and it is the one used here.

Definition A.498 (Zero content)**.** A bounded set $Z \subset \mathbb{R}^N$ has *zero content* if for every $\varepsilon > 0$ there are finitely many cubes of total volume less than ε whose union contains Z.
Rests on Definitions 6.9 and 7.93.

Covering by cubes rather than by boxes is no restriction: a box with edges $a_1, \dots, a_N$ is covered by cubes of side h in number at most $\prod_i (a_i/h + 1)$, of total volume at most $\prod_i (a_i + h)$, which tends to the volume of the box as $h \to 0$. A finite union of sets of zero content has zero content.

Lemma A.499 (What zero content buys)**.** *Let R be a box, $Z \subset R$ of zero content, and $f : R \to \mathbb{R}$ bounded with $|f| \le M$.*

1. *If f is continuous at every point of $R \setminus Z$, then f is integrable on R.*
2. *If f vanishes off Z, then f is integrable with $\int_R f = 0$.*
3. *If $E_1, \dots, E_m \subseteq R$ are such that each characteristic function χ_{E_i} is integrable, the E_i cover a set E and the pairwise intersections $E_i \cap E_j$ $(i \neq j)$ have zero content, then $\int_E f = \sum_i \int_{E_i} f$ for every f for which the left-hand side is defined.*

Rests on Definitions 7.93 and A.498 and Theorem 7.25.

Proof. *Derives Lemma A.499.*
(1) Given $\varepsilon > 0$, cover Z by finitely many open cubes of total volume less than ε and let K be the complement in R of their union: K is closed and bounded, hence compact (Theorem 6.12). Note that f is continuous, as a function on R, at every point of K.

We claim there is $\lambda > 0$ such that any subset of R of diameter less than λ which meets K carries an oscillation of f of at most ε. If not, then for each j there are $\boldsymbol{x}_j \in K$ and $\boldsymbol{y}_j, \boldsymbol{z}_j \in R$ within $1/j$ of $\boldsymbol{x}_j$ with $|f(\boldsymbol{y}_j) - f(\boldsymbol{z}_j)| > \varepsilon$; by compactness a subsequence $\boldsymbol{x}_j$ converges to some $\boldsymbol{x} \in K$, the corresponding $\boldsymbol{y}_j$ and $\boldsymbol{z}_j$ converge to $\boldsymbol{x}$ as well, and the continuity of f at $\boldsymbol{x}$ makes both $f(\boldsymbol{y}_j)$ and $f(\boldsymbol{z}_j)$ tend to $f(\boldsymbol{x})$ — a contradiction.

Now partition R with mesh smaller than $\lambda/\sqrt{N}$. Writing A for the union of the covering cubes, every sub-box either lies inside A or meets $K = R \setminus A$. Those of the first kind have total volume at most $\operatorname{vol}(A) < \varepsilon$ and contribute at most $2M\varepsilon$ to the gap between the upper and the lower sum; on those of the second kind the oscillation of f is at most ε, contributing at most $\varepsilon \operatorname{vol}(R)$. The gap can therefore be made as small as desired, and f is integrable (Definition 7.93).

(2) By (1), f is integrable, since it is continuous (with value zero) off Z. Cover Z by finitely many cubes of total volume less than ε and partition R with a mesh small compared with the smallest of those cubes. Every sub-box on which f is not identically zero meets Z,

hence lies in the union of the cubes enlarged by one mesh on every side, whose volume is then at most 2ε; so $|\int_R f| \le 2M\varepsilon$ for every $\varepsilon > 0$.

(3) The function $\chi_E - \sum_i \chi_{E_i}$ vanishes off $\bigcup_{i<j} (E_i \cap E_j)$, a set of zero content, and is bounded by m; multiplying by f and applying (2) gives the statement. □

Lemma A.500 (Graphs and C^1 images have zero content)**.** *1. Let $B \subset \mathbb{R}^{N-1}$ be a box and $g : B \to \mathbb{R}$ continuous. Then the graph $\{(\boldsymbol{u}', g(\boldsymbol{u}')) \mid \boldsymbol{u}' \in B\}$ has zero content in $\mathbb{R}^N$; so does its intersection with any coordinate hyperplane, and so does a face of a box.*

2. Let $W \subseteq \mathbb{R}^N$ be open, $\boldsymbol{F} : W \to \mathbb{R}^N$ of class C^1, and $Z \subset W$ a compact set of zero content. Then $\boldsymbol{F}(Z)$ has zero content.

Rests on Lemma A.499 and Theorems 7.25 and 7.35.

Proof. *Derives Lemma A.500.*

(1) Let $\varepsilon > 0$. By uniform continuity (Theorem 7.25) there is $\delta > 0$ such that g varies by less than ε on any subset of B of diameter less than δ. Cover B by cubes of side $h < \min\left\{\delta/\sqrt{N-1}, \varepsilon\right\}$, in number at most $C_B h^{-(N-1)}$ with C_B depending on B alone. Over each such cube the graph lies in a box of base that cube and height ε, which is covered by $\lceil \varepsilon/h \rceil$ cubes of side h, of total volume at most $(\varepsilon + h)\, h^{N-1} \le 2\varepsilon h^{N-1}$. Summing gives a total volume at most $2C_B\varepsilon$. A face of a box is the graph of a constant function.

(2) Cover Z by finitely many closed balls contained in W; on each, the entries of $D\boldsymbol{F}$ are bounded, and the mean value theorem (Theorem 7.35) applied to each component along the segment joining two points of the ball — which stays in the ball, the ball being convex — gives a constant L with $|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{y})| \le L\,|\boldsymbol{x} - \boldsymbol{y}|$ there. Take L to be the largest of the finitely many constants. Now cover Z by cubes of side h and total volume less than ε, refining h so small that each such cube meeting Z lies in one of the balls. A cube of side h has diameter $\sqrt{N}h$, so its image has diameter at most $L\sqrt{N}h$ and lies in a cube of that side, of volume $\left(L\sqrt{N}\right)^N h^N$. The images therefore cover $\boldsymbol{F}(Z)$ with total volume at most $\left(L\sqrt{N}\right)^N \varepsilon$. □

A.44.3 Iterated integration in $\mathbb{R}^N$

Lemma A.501 (Iterated integration over a box)**.** *Let $R = R' \times [a, b]$ be a box in $\mathbb{R}^N = \mathbb{R}^{N-1} \times \mathbb{R}$ and let $f : R \to \mathbb{R}$ be bounded and integrable, and such that $t \longmapsto f(\boldsymbol{u}', t)$ is integrable on $[a, b]$ for every $\boldsymbol{u}' \in R'$. Then*

$$\int_R f \,\mathrm{d}^N u = \int_{R'} \left[\int_a^b f(\boldsymbol{u}', t)\,\mathrm{d}t \right] \mathrm{d}^{N-1} u' \,, \tag{A.800}$$

the inner integral being an integrable function of $\boldsymbol{u}'$. The same holds with any one of the N coordinates in the role of t.

Rests on Definitions 7.39 and 7.93 and Theorem 7.40.

Proof. *Derives Lemma A.501.*

Write $F(\boldsymbol{u}') = \int_a^b f(\boldsymbol{u}', t)\,\mathrm{d}t$. Let P' be a partition of R' into sub-boxes S' and P_t a partition $a = t_0 < \cdots < t_n = b$; together they partition R into the sub-boxes $S = S' \times [t_{i-1}, t_i]$. Fix $\boldsymbol{u}' \in S'$. For each i and every $t \in [t_{i-1}, t_i]$ one has $f(\boldsymbol{u}', t) \ge \inf_S f$, so

$$F(\boldsymbol{u}') \ge \sum_{i=1}^{n} \left(\inf_{S' \times [t_{i-1}, t_i]} f \right) (t_i - t_{i-1}) \,, \tag{A.801}$$

the right-hand side being independent of $\boldsymbol{u}' \in S'$; hence it is a lower bound for $\inf_{S'} F$. Multiplying by $\mathrm{vol}(S')$ and summing over S' gives $L(F, P') \ge L(f, P)$, where L denotes the

lower Darboux sum Equation (7.22). The mirror argument gives $U(F,P') \le U(f,P)$. Since f is integrable, the outer sums squeeze: for every $\varepsilon > 0$ a partition P with $U(f,P) - L(f,P) < \varepsilon$ yields $U(F,P') - L(F,P') < \varepsilon$, so F is integrable on R', and both $\int_{R'} F$ and $\int_R f$ lie between $L(f,P)$ and $U(f,P)$; letting $\varepsilon \to 0$ makes them equal. Relabelling the coordinates puts any chosen one in the role of t. □

Corollary A.502 (All iterated orders agree)**.** *Let $f \in C_c(\mathbb{R}^N)$ and let σ be a permutation of $\{1,\dots,N\}$. Then $\int_{\mathbb{R}^N} f$ equals the iterated integral taken in the order $u^{\sigma(N)}, u^{\sigma(N-1)}, \dots, u^{\sigma(1)}$ from innermost to outermost; consequently $\int_{\mathbb{R}^N} f \circ \sigma = \int_{\mathbb{R}^N} f$, where σ also denotes the map $\boldsymbol{u} \mapsto \left(u^{\sigma(1)}, \dots, u^{\sigma(N)}\right)$.*

Rests on Lemma A.501 and Theorem 7.40.

Proof. *Derives Corollary A.502.*

Take a box R containing the support in its interior. A continuous function is integrable on R and on every slice (Theorem 7.40), so Lemma A.501 applies with any chosen coordinate innermost; the inner integral is again continuous, by the uniform continuity of f on the compact R (Theorem 7.25), so the lemma may be applied again to it, and N applications reduce $\int_R f$ to the iterated integral in the chosen order. All $N!$ orders therefore give the same number. Finally $f \circ \sigma$ is continuous with compact support, and its iterated integral in the order $u^1, \dots, u^N$ becomes, on renaming the integration variable of the j-th integration from u^j to $u^{\sigma(j)}$, the iterated integral of f in the order $u^{\sigma(1)}, \dots, u^{\sigma(N)}$, which is $\int_{\mathbb{R}^N} f$. □

A.44.4 Two reductions

Definition A.503 (The substitution property)**.** Let $U, V \subseteq \mathbb{R}^N$ be open and $\boldsymbol{\Phi} : U \to V$ a C^1 diffeomorphism. We say $\boldsymbol{\Phi}$ *has the substitution property* if

$$\int_{\mathbb{R}^N} f = \int_{\mathbb{R}^N} (f \circ \boldsymbol{\Phi}) \, |\det D\boldsymbol{\Phi}| \qquad \text{for every } f \in C_c(V)\,, \tag{A.802}$$

both integrands being understood as extended by zero; the right-hand one belongs to $C_c(U)$, its support being $\boldsymbol{\Phi}^{-1}(\operatorname{supp} f)$.

Rests on Corollary 7.81 and Definition 7.93.

Lemma A.504 (Transitivity)**.** *If $\boldsymbol{\Phi} : U \to V$ and $\boldsymbol{\Psi} : V \to W$ are C^1 diffeomorphisms with the substitution property, so is $\boldsymbol{\Psi} \circ \boldsymbol{\Phi}$.*

Rests on Definition A.503, Proposition 7.72, and Lemma A.497.

Proof. *Derives Lemma A.504.*

Let $g \in C_c(W)$. Then $f := (g \circ \boldsymbol{\Psi}) \, |\det D\boldsymbol{\Psi}|$ lies in $C_c(V)$, and applying Equation (A.802) first to $\boldsymbol{\Psi}$ and then to $\boldsymbol{\Phi}$,

$$\int_{\mathbb{R}^N} g = \int_{\mathbb{R}^N} f = \int_{\mathbb{R}^N} (g \circ \boldsymbol{\Psi} \circ \boldsymbol{\Phi}) \, (|\det D\boldsymbol{\Psi}| \circ \boldsymbol{\Phi}) \, |\det D\boldsymbol{\Phi}| \,. \tag{A.803}$$

By the chain rule (Proposition 7.72), $D\,(\boldsymbol{\Psi} \circ \boldsymbol{\Phi})\,(\boldsymbol{u}) = D\boldsymbol{\Psi}\,(\boldsymbol{\Phi}(\boldsymbol{u}))\, D\boldsymbol{\Phi}(\boldsymbol{u})$, and Lemma A.497 turns the product of the two moduli into $|\det D\,(\boldsymbol{\Psi} \circ \boldsymbol{\Phi})|$. □

Lemma A.505 (A continuous partition of unity)**.** *Let $K \subset \mathbb{R}^N$ be compact and $W_1, \dots, W_m$ open sets with $K \subseteq W_1 \cup \cdots \cup W_m$. There are continuous functions $\lambda_1, \dots, \lambda_m$ on $\mathbb{R}^N$ with $0 \le \lambda_j \le 1$, with $\operatorname{supp} \lambda_j$ compact and contained in W_j, and with $\sum_j \lambda_j = 1$ on a neighbourhood of K.*

Rests on Definitions 6.6, 6.9 and 6.24.

Proof. *Derives Lemma A.505.*

For each $\boldsymbol{x} \in K$ choose $j(\boldsymbol{x})$ with $\boldsymbol{x} \in W_{j(\boldsymbol{x})}$ and $r(\boldsymbol{x}) > 0$ with the closed ball of radius $2r(\boldsymbol{x})$ about $\boldsymbol{x}$ inside $W_{j(\boldsymbol{x})}$. The open balls of radius $r(\boldsymbol{x})$ cover K; extract a finite subcover

(Definition 6.9) with centres $\boldsymbol{x}_1, \dots, \boldsymbol{x}_p$ and let K_j be the union of the closed balls $\overline{B}_{r(\boldsymbol{x}_i)}(\boldsymbol{x}_i)$ with $j(\boldsymbol{x}_i) = j$. Each K_j is compact and contained in W_j, and $K \subseteq \bigcup_j K_j$. Let $\rho_j > 0$ be smaller than the distance from K_j to the complement of W_j — positive because K_j is compact and the complement closed — and put

$$\psi_j(\boldsymbol{x}) = \max\{0,\ \rho_j - \operatorname{dist}(\boldsymbol{x}, K_j)\}\ , \tag{A.804}$$

which is continuous, non-negative, strictly positive exactly on the open set $\{\operatorname{dist}(\cdot, K_j) < \rho_j\}$, and supported in its closure, a compact subset of W_j. On the open set $O = \{\sum_i \psi_i > 0\}$, which contains K, set $\lambda_j = \psi_j / \sum_i \psi_i$; off O set $\lambda_j = 0$. Each λ_j is continuous on O and vanishes on a neighbourhood of every point of $\mathbb{R}^N \setminus \operatorname{supp}\psi_j$, so it is continuous everywhere; its support lies in that of ψ_j; and $\sum_j \lambda_j = 1$ on O. □

Lemma A.506 (Locality)**.** *Let $\boldsymbol{\Phi} : U \to V$ be a C^1 diffeomorphism and suppose every $\boldsymbol{p} \in U$ has an open neighbourhood $N_{\boldsymbol{p}} \subseteq U$ such that the restriction $\boldsymbol{\Phi}|_{N_{\boldsymbol{p}}} : N_{\boldsymbol{p}} \to \boldsymbol{\Phi}(N_{\boldsymbol{p}})$ has the substitution property. Then $\boldsymbol{\Phi}$ has it.*

Rests on Lemma A.505 and Definitions 6.9 and A.503.

Proof. *Derives Lemma A.506.*

Let $f \in C_c(V)$ with $K = \operatorname{supp} f$. The set $\boldsymbol{\Phi}^{-1}(K)$ is compact, so finitely many $N_{\boldsymbol{p}_1}, \dots, N_{\boldsymbol{p}_m}$ cover it, and the open sets $W_j = \boldsymbol{\Phi}(N_{\boldsymbol{p}_j}) \subseteq V$ cover K. Take $\lambda_1, \dots, \lambda_m$ as in Lemma A.505. Each $\lambda_j f$ is continuous, vanishes off $\operatorname{supp}\lambda_j \cap K$ — a compact subset of W_j — and hence lies in $C_c(W_j)$; and $\sum_j \lambda_j f = f$, since the two sides agree on the neighbourhood of K where $\sum_j \lambda_j = 1$ and both vanish off K. Applying the substitution property of $\boldsymbol{\Phi}|_{N_{\boldsymbol{p}_j}}$ to $\lambda_j f$ and summing over j,

$$\int_{\mathbb{R}^N} f = \sum_{j=1}^{m} \int_{\mathbb{R}^N} \lambda_j f = \sum_{j=1}^{m} \int_{\mathbb{R}^N} ((\lambda_j f) \circ \boldsymbol{\Phi})\, |\det D\boldsymbol{\Phi}| = \int_{\mathbb{R}^N} (f \circ \boldsymbol{\Phi})\, |\det D\boldsymbol{\Phi}|\ , \tag{A.805}$$

the first and last steps by the linearity of the integral over a box containing every support involved. □

A.44.5 Permutations and primitive maps

Lemma A.507 (Coordinate permutations)**.** *For a permutation σ of $\{1, \dots, N\}$, the linear map $\boldsymbol{u} \longmapsto \left(u^{\sigma(1)}, \dots, u^{\sigma(N)}\right)$ of $\mathbb{R}^N$ onto itself has the substitution property.*

Rests on Corollary A.502 and Equation (5.19).

Proof. *Derives Lemma A.507.*

Its matrix has exactly one 1 in each row and column, so Equation (5.19) leaves a single term and the determinant is $\operatorname{sgn}(\sigma) = \pm 1$; hence $|\det| = 1$ and Equation (A.802) reduces to $\int f = \int f \circ \sigma$, which is Corollary A.502. This is the only step of the whole derivation at which the *sign* of the determinant is disposed of, and it is why Equation (7.113) carries a modulus while Equation (5.19) does not. □

Definition A.508 (Primitive map)**.** A C^1 diffeomorphism $\boldsymbol{\Phi} : U \to V$ of open subsets of $\mathbb{R}^N$ is *primitive* if it changes only the last coordinate:

$$\boldsymbol{\Phi}(\boldsymbol{u}', t) = \left(\boldsymbol{u}',\ \Phi^N(\boldsymbol{u}', t)\right)\ , \qquad \det D\boldsymbol{\Phi} = \frac{\partial \Phi^N}{\partial t}\ , \tag{A.806}$$

the determinant being that of a triangular matrix with $N - 1$ unit diagonal entries.

Rests on Definition 7.66 and Equation (5.19).

Lemma A.509 (Primitive maps have the substitution property)**.** *Every primitive C^1 diffeomorphism has the substitution property.*

Rests on Definition A.508, Equation (7.27), and Lemma A.501.

Proof. *Derives Lemma A.509.*

Let $f \in C_c(V)$ and write $g = (f \circ \mathbf{\Phi}) \left|\partial_t \Phi^N\right| \in C_c(U)$. Both f and g, extended by zero, are continuous with compact support, so by Corollary A.502 each integral is the iterated integral with t innermost. It therefore suffices to prove, for every fixed $\boldsymbol{u}' \in \mathbb{R}^{N-1}$,

$$\int_{\mathbb{R}} f(\boldsymbol{u}', s)\,\mathrm{d}s = \int_{\mathbb{R}} f\left(\boldsymbol{u}', \Phi^N(\boldsymbol{u}', t)\right) \left|\frac{\partial \Phi^N}{\partial t}(\boldsymbol{u}', t)\right| \mathrm{d}t\,. \tag{A.807}$$

Fix $\boldsymbol{u}'$ and put $U_{\boldsymbol{u}'} = \{t \mid (\boldsymbol{u}', t) \in U\}$ and $V_{\boldsymbol{u}'} = \{s \mid (\boldsymbol{u}', s) \in V\}$, both open in $\mathbb{R}$. Because $\mathbf{\Phi}$ leaves $\boldsymbol{u}'$ untouched and is a bijection of U onto V, the map $\varphi(t) = \Phi^N(\boldsymbol{u}', t)$ is a bijection of $U_{\boldsymbol{u}'}$ onto $V_{\boldsymbol{u}'}$; it is C^1 with $\varphi' = \partial_t \Phi^N \neq 0$, hence strictly monotone on each connected component of $U_{\boldsymbol{u}'}$, mapping that component onto a component of $V_{\boldsymbol{u}'}$.

The integrand on the right of Equation (A.807) vanishes off the compact set $\mathbf{\Phi}^{-1}(\operatorname{supp} f)$, whose slice at $\boldsymbol{u}'$ is a compact subset T of $U_{\boldsymbol{u}'}$. The components of $U_{\boldsymbol{u}'}$ are open and disjoint and cover T, so only finitely many of them, $I_1, \ldots, I_q$, meet T; choose in each a compact interval $[\alpha_r, \beta_r] \subset I_r$ containing $T \cap I_r$, with $\alpha_r < \beta_r$. On I_r the substitution rule Equation (7.27), applied to the C^1 function φ and the continuous $s \mapsto f(\boldsymbol{u}', s)$, gives

$$\int_{\alpha_r}^{\beta_r} f\left(\boldsymbol{u}', \varphi(t)\right) \varphi'(t)\,\mathrm{d}t = \int_{\varphi(\alpha_r)}^{\varphi(\beta_r)} f(\boldsymbol{u}', s)\,\mathrm{d}s\,. \tag{A.808}$$

If φ increases on I_r then $\varphi' = |\varphi'|$ and the right-hand side is the integral over $[\varphi(\alpha_r), \varphi(\beta_r)]$; if φ decreases then $\varphi' = -|\varphi'|$ and the right-hand side is minus the integral over $[\varphi(\beta_r), \varphi(\alpha_r)]$. In both cases

$$\int_{\alpha_r}^{\beta_r} f\left(\boldsymbol{u}', \varphi(t)\right) |\varphi'(t)|\,\mathrm{d}t = \int_{\varphi([\alpha_r, \beta_r])} f(\boldsymbol{u}', s)\,\mathrm{d}s\,. \tag{A.809}$$

Outside $[\alpha_r, \beta_r]$ the left integrand vanishes, so the left-hand side is the integral over the whole of I_r; and $f(\boldsymbol{u}', \cdot)$ vanishes on $\varphi(I_r) \setminus \varphi\left([\alpha_r, \beta_r]\right)$, since a point s there has $\varphi^{-1}(s) \notin T$, so the right-hand side is the integral over the component $\varphi(I_r)$ of $V_{\boldsymbol{u}'}$. Summing over $r = 1, \ldots, q$ and noting that $f(\boldsymbol{u}', \cdot)$ vanishes on the components of $V_{\boldsymbol{u}'}$ not among the $\varphi(I_r)$, and off $V_{\boldsymbol{u}'}$ altogether, gives Equation (A.807). □

A.44.6 The local factorisation

Lemma A.510 (Every diffeomorphism factorises locally)**.** *Let $\mathbf{\Phi}$ be a C^1 diffeomorphism of an open subset of $\mathbb{R}^N$ onto an open subset, and let $\boldsymbol{p}$ be a point of its domain. Then some neighbourhood of $\boldsymbol{p}$ carries a factorisation of $\mathbf{\Phi}$ as a composition of finitely many primitive maps and coordinate permutations.*

Rests on Corollary 7.81, Definition A.508, and Proposition 7.72.

Proof. *Derives Lemma A.510.*

Say that a C^1 diffeomorphism $\mathbf{\Theta}$ defined near a point is *m-adapted* there, $1 \leq m \leq N+1$, if $\Theta^i(\boldsymbol{u}) = u^i$ for every $i < m$ throughout a neighbourhood. Every diffeomorphism is 1-adapted (the condition is empty), and an $(N+1)$-adapted one is the identity. We show that an m-adapted $\mathbf{\Theta}$ is, near the point, the composition of an $(m+1)$-adapted diffeomorphism with one primitive map and one transposition of coordinates; N applications then prove the lemma, since a primitive map in the m-th coordinate is a transposition conjugate of one in the last coordinate.

Let $\mathbf{\Theta}$ be m-adapted at $\boldsymbol{q}$. Its Jacobian there has the block form

$$D\mathbf{\Theta}(\boldsymbol{q}) = \begin{pmatrix} \mathbb{1} & 0 \\ * & C \end{pmatrix}, \tag{A.810}$$

with $\mathbb{1}$ of size $m-1$ and C the matrix $\partial\Theta^i/\partial u^j$ with $i,j \geq m$; expanding the determinant along the first $m-1$ rows shows $\det C = \det D\boldsymbol{\Theta} \neq 0$, so the rows of C are independent and in particular the row $\left(\partial\Theta^m/\partial u^j\right)_{j\geq m}$ is non-zero: there is $j \geq m$ with $\partial\Theta^m/\partial u^j(\boldsymbol{q}) \neq 0$.

Let τ be the transposition of the coordinates m and j, and set

$$\boldsymbol{\varsigma}(\boldsymbol{u}) = \left(u^1, \dots, u^{m-1}, \Theta^m\left(\tau(\boldsymbol{u})\right), u^{m+1}, \dots, u^N\right), \tag{A.811}$$

a map changing only the m-th coordinate. At $\boldsymbol{q}' = \tau(\boldsymbol{q})$ its m-th partial derivative in the m-th coordinate is $\partial\Theta^m/\partial u^j(\boldsymbol{q}) \neq 0$, so $\det D\boldsymbol{\varsigma}(\boldsymbol{q}') \neq 0$ and, by the inverse function theorem (Corollary 7.81), $\boldsymbol{\varsigma}$ is a C^1 diffeomorphism of a neighbourhood of $\boldsymbol{q}'$ onto an open set; it is a primitive map in the m-th coordinate.

Put $\boldsymbol{\Xi} = \boldsymbol{\Theta} \circ \tau \circ \boldsymbol{\varsigma}^{-1}$, so that $\boldsymbol{\Theta} = \boldsymbol{\Xi} \circ \boldsymbol{\varsigma} \circ \tau$ near $\boldsymbol{q}$ (τ being its own inverse). For $i < m$, neither τ nor $\boldsymbol{\varsigma}$ — hence nor $\boldsymbol{\varsigma}^{-1}$ — touches the i-th coordinate, and $\boldsymbol{\Theta}$ preserves it, so $\Xi^i(\boldsymbol{v}) = v^i$. For $i = m$: writing $\boldsymbol{u} = \boldsymbol{\varsigma}^{-1}(\boldsymbol{v})$ one has $v^m = \varsigma^m(\boldsymbol{u}) = \Theta^m\left(\tau(\boldsymbol{u})\right) = \Xi^m(\boldsymbol{v})$. Hence $\boldsymbol{\Xi}$ is $(m+1)$-adapted, as claimed. □

Proposition A.511 (The substitution property is universal)**.** *Every C^1 diffeomorphism $\boldsymbol{\Phi} : U \to V$ of open subsets of $\mathbb{R}^N$ has the substitution property: for every $f \in C_c(V)$,*

$$\int_V f(\boldsymbol{x})\, \mathrm{d}^N x = \int_U f\left(\boldsymbol{\Phi}(\boldsymbol{u})\right) \left|\det D\boldsymbol{\Phi}(\boldsymbol{u})\right| \mathrm{d}^N u\,. \tag{A.812}$$

Rests on Lemmas A.506, A.509 and A.510.

Proof. *Derives Proposition A.511.*

By Lemma A.510 each point of U has a neighbourhood on which $\boldsymbol{\Phi}$ is a composition of primitive maps and permutations; each factor has the substitution property by Lemmas A.507 and A.509, so the composition has it by Lemma A.504. Thus $\boldsymbol{\Phi}$ has the property locally, and Lemma A.506 globalises it. □

A.44.7 From compact support to the theorem

Proposition A.511 carries no hypothesis on the shape of U and V. To reach Theorem 7.97 as stated — an integrand merely continuous and bounded on V, not vanishing near the boundary — something must be assumed about that boundary, and about the behaviour of $\boldsymbol{\Phi}$ there; otherwise neither integral need exist. The hypotheses below hold for every use made of the formula in this treatise.

Lemma A.512 (The boundary strip is thin)**.** *Let $V \subset \mathbb{R}^N$ be open and bounded and suppose its boundary is contained in finitely many graphs of continuous functions over boxes, in the sense of Lemma A.500(1) with the coordinates permuted — the N-dimensional reading of Definition 7.95. Define the boundary strip*

$$V_\delta = \{\boldsymbol{x} \in V \mid \operatorname{dist}(\boldsymbol{x}, \partial V) < \delta\}\,. \tag{A.813}$$

Then the content of V_δ tends to zero as $\delta \to 0$.

Rests on Lemma A.500, Theorem 7.25, and Definition 7.95.

Proof. *Derives Lemma A.512.*

It is enough to treat one graph Γ, of a continuous g over a box $B \subset \mathbb{R}^{N-1}$, and to bound the content of its δ-neighbourhood. By uniform continuity (Theorem 7.25) split B into cubes of side h on each of which g varies by less than δ. A point within distance δ of Γ and lying over such a cube has last coordinate within 2δ of the value of g at the cube's centre, and lies over the cube enlarged by δ; so the neighbourhood is covered by boxes of total volume at most $4\delta\left(\operatorname{vol}(B) + C\delta\right)$ with C depending on B alone. Letting first $h \to 0$ and then $\delta \to 0$ gives the claim; the finitely many graphs are summed. □

Theorem A.513 (Change of variables)**.** *Let $U, V \subseteq \mathbb{R}^N$ be open and bounded and let $\mathbf{\Phi} : U \to V$ be a bijection of class C^1 with $\det D\mathbf{\Phi} \neq 0$ everywhere. Assume in addition that $\mathbf{\Phi}$ is the restriction of a C^1 diffeomorphism of an open set containing $\overline{U}$, and that ∂U and ∂V are as in Lemma A.512. Then for every bounded continuous f on V both integrals in Equation (7.113) exist and are equal.*

Rests on Proposition A.511 and Lemmas A.499 and A.512.

Proof. *Derives Theorem A.513.*

Existence. Extend f by zero to a box $R \supseteq \overline{V}$. The extension is bounded and continuous off ∂V, which has zero content by Lemma A.500(1); so it is integrable by Lemma A.499(1). The same argument applies on the U side: $(f \circ \mathbf{\Phi})\,|\det D\mathbf{\Phi}|$ is bounded there, because $D\mathbf{\Phi}$ extends continuously to the compact $\overline{U}$, and continuous off ∂U.

Equality. Let M bound $|f|$ and J bound $|\det D\mathbf{\Phi}|$ on $\overline{U}$. For $\delta > 0$ put

$$\eta_\delta(\boldsymbol{x}) = \min\left\{1,\ \max\left\{0,\ \frac{\operatorname{dist}(\boldsymbol{x}, \mathbb{R}^N \setminus V) - \delta}{\delta}\right\}\right\}, \tag{A.814}$$

a continuous function, zero where the distance is at most δ and one where it is at least 2δ. The set where $\eta_\delta \neq 0$ has closure inside $\left\{\operatorname{dist}(\cdot, \mathbb{R}^N \setminus V) \geq \delta\right\}$, which is closed and bounded, hence a compact subset of V; so $f_\delta = \eta_\delta f$ lies in $C_c(V)$ and Proposition A.511 gives

$$\int_V f_\delta = \int_U (f_\delta \circ \mathbf{\Phi})\,|\det D\mathbf{\Phi}|\ . \tag{A.815}$$

On the left, $f - f_\delta$ vanishes outside $V_{2\delta}$ and is bounded by M, so the difference of the two integrals is at most M times the content of $V_{2\delta}$, which tends to zero by Lemma A.512. On the right, the integrands differ only on $\mathbf{\Phi}^{-1}(V_{2\delta})$ and by at most MJ. The extension of $\mathbf{\Phi}^{-1}$ is C^1 on a neighbourhood of the compact $\overline{V}$, hence Lipschitz there with some constant L by the argument of Lemma A.500(2); covering $V_{2\delta}$ by cubes of total volume ε and taking images covers $\mathbf{\Phi}^{-1}(V_{2\delta})$ with total volume at most $\left(L\sqrt{N}\right)^N \varepsilon$, so its content also tends to zero. Letting $\delta \to 0$ in Equation (A.815) gives Equation (7.113). □

Corollary A.514 (Degeneracy on a negligible set)**.** *Let U, V be open and bounded, let $Z \subset U$ be closed in U and contained in finitely many graphs of continuous functions over boxes, and let $\mathbf{\Phi} : U \to V$ be C^1, mapping $U \setminus Z$ bijectively onto $V \setminus \mathbf{\Phi}(Z)$ with non-vanishing Jacobian there. If $\mathbf{\Phi}$ extends C^1 to a neighbourhood of $\overline{U}$, then Equation (7.113) holds for every $f \in C_c(V)$, and for bounded continuous f under the boundary hypotheses of Theorem A.513 applied to $U \setminus Z$ and $V \setminus \mathbf{\Phi}(Z)$.*

Rests on Theorem A.513 and Lemmas A.499 and A.500.

Proof. *Derives Corollary A.514.*

$U \setminus Z$ is open, and $\mathbf{\Phi}|_{U \setminus Z}$ is a C^1 diffeomorphism onto the open set $V \setminus \mathbf{\Phi}(Z)$; note that $\mathbf{\Phi}(Z \cap K)$ has zero content for every compact $K \subset \overline{U}$, by Lemma A.500. Apply Proposition A.511, or Theorem A.513, to that restriction. Both integrals change by nothing when the sets Z and $\mathbf{\Phi}(Z)$ are restored, because the integrands are bounded and the two sets have zero content (Lemma A.499(2)). □

The corollary is what licenses the two Jacobians of Example 7.98. The polar map $\mathbf{\Phi}(\rho, \varphi) = (\rho\cos\varphi, \rho\sin\varphi)$ of Equation (7.114) has $\det D\mathbf{\Phi} = \rho$, which vanishes on the segment $\rho = 0$, and it fails to be injective unless φ is confined to an interval of length 2π, so that a ray must be removed from the image; segment and ray are each contained in the graph of a continuous function, and neither contributes to either integral. The spherical map is degenerate on the polar axis in the same way.

A.44.8 Reading the result

Remark A.515 (What is quoted here). Nothing outside this treatise, and no result of Chapter 7 beyond the ones listed in Appendix A.44.1. Three points are worth recording, because each is a place where the chapters state slightly less than the derivation needs and the gap is filled here rather than passed over.

First, Chapter 5 records the Leibniz expansion Equation (5.19) and reads off column linearity and antisymmetry, but never states that the determinant is multiplicative; that is Lemma A.497, proved above from the expansion alone. It is a debt of Part II, small but real: the composition rule for Jacobians is used wherever coordinates are changed twice.

Second, Remark 7.96 states the reduction of a multiple integral to an iterated one for $N = 2$ and $N = 3$ only. The N-dimensional statement is Lemma A.501, proved here by the same Darboux squeeze. This too is a Part II debt, and it is the reason Theorem 7.97 could not simply be quoted from the chapter's own machinery.

Third — and this corrects the plan under which this section was written — the passage from a local identity to a global one does use a partition of unity. It is a *continuous* one, constructed in Lemma A.505 out of distance functions and nothing else; no smooth bump function is needed, because f is only assumed continuous. There is no fixed-point theorem, no convergence theorem of Lebesgue integration and no measure theory anywhere in the section: every set that is discarded is discarded because it has zero *content* (Definition A.498), which a Darboux sum can see, and not merely measure zero in the sense of Section 7.12.

Remark A.516 (The hypotheses of the global form). Proposition A.511, the compactly supported form, is the theorem proper: it holds for arbitrary open U and V and is where all the work is. The extra hypotheses of Theorem A.513 — a boundary made of finitely many graphs, and a C^1 extension across $\overline{U}$ — are not artefacts of the method. Without some such assumption the two integrals in Equation (7.113) need not exist at all: a bounded continuous function on a bounded open set with a sufficiently wild boundary is not Riemann integrable, and the statement of Theorem 7.97 that "both integrals exist" is then false rather than unproved. The chapter's phrasing should be read with these hypotheses supplied; they hold in every application made in this book, where the regions are simple in the sense of Definition 7.95 and the maps are restrictions of diffeomorphisms defined on larger open sets.

Rests on Theorem A.513, Proposition A.511, and Definition 7.95.

Remark A.517. Appendix A.44 discharges the derivation owed at Theorem 7.97 of Chapter 7. The result is cashed in three places. It is what makes the surface integral Equation (7.109) independent of the parametrization used to define it, as Remark 7.102 states explicitly: two parametrizations of one surface differ by a C^1 diffeomorphism of their parameter domains, and the Jacobian factor supplied by Equation (7.113) cancels the corresponding factor in $\partial_u \boldsymbol{r} \times \partial_v \boldsymbol{r}$ — so the flux, and with it Theorem 7.101 and Theorem 7.100, is well defined. It supplies the polar and spherical volume elements of Example 7.98, used throughout the physical parts and, in this appendix alone, in Appendix A.45. And it is the statement behind the invariance of phase-space volume, Corollary 22.24 of Part III, where the relevant map is the Hamiltonian flow and the Jacobian determinant is the one Liouville's theorem shows to be unity.

A.45 The Helmholtz Decomposition

This appendix proves Theorem 7.105 of Chapter 7: every smooth compactly supported vector field on $\mathbb{R}^3$ splits as in Equation (7.129) into a gradient and a curl, with the potentials given explicitly by Equation (7.130), and the two summands are unique among fields that vanish at infinity. It is the theorem that licenses the routine physical move of treating the irrotational and the solenoidal parts of a field separately — the longitudinal and transverse

parts of the electromagnetic field, the P and S waves of a seismogram, the potential and vortical parts of a flow.

The paragraph following the proof link in Section 7.11.5 already gives the mechanism, and it is correct: each Cartesian component of $\boldsymbol{w}$ is a Newtonian potential, so $\nabla^2 \boldsymbol{w} = \boldsymbol{u}$, and Equation (7.105) read backwards turns that Laplacian into the required sum. What the chapter does not supply, and what this section owes, is the analysis: that the singular integral Equation (7.130) converges at all, that it may be differentiated under the integral sign as often as one pleases, that the resulting field decays, and that the decay is enough to force uniqueness. That is the whole content of what follows.

One organisational point should be made at the outset. The argument uses two results of Chapter 10 — the fundamental-solution property of the Newtonian potential (Proposition 10.67) and the weak maximum principle (Theorem 10.75) — and Chapter 10 comes *after* Chapter 7. Nothing is circular: neither of those two results uses Theorem 7.105, or anything proved here. But the forward dependence is real and is why the derivation sits in the appendix rather than in the chapter, where it would have to be either postponed or duplicated. See Remark A.526.

Throughout, $\boldsymbol{u}$ is a smooth vector field on $\mathbb{R}^3$ whose support is contained in the ball $|\boldsymbol{y}| \le R_0$, and we write

$$\Phi(\boldsymbol{z}) = -\frac{1}{4\pi\,|\boldsymbol{z}|}\,, \tag{A.816}$$

for the Newtonian kernel Equation (10.63). The section is pure vector analysis, but the SI bookkeeping is worth recording once because the physical parts read the decomposition off it: Φ has the dimension of a reciprocal length and $\mathrm{d}^3 z$ that of a volume, so $\boldsymbol{w}$ carries the unit of $\boldsymbol{u}$ multiplied by m^2; ϕ and $\boldsymbol{\Psi}$, being first derivatives of $\boldsymbol{w}$, carry that of $\boldsymbol{u}$ multiplied by m; and $\nabla\phi$ and $\nabla \times \boldsymbol{\Psi}$ carry the unit of $\boldsymbol{u}$ itself, as Equation (7.129) requires.

A.45.1 What the derivation uses

From Chapter 7: the Leibniz integral rule (Theorem 7.77), both halves of the fundamental theorem of calculus (Theorems 7.42 and 7.43), uniform continuity (Theorem 7.25), the Cauchy criterion (Theorem 7.8), the second-order identities of the nabla calculus (Proposition 7.91), the class C^1 (Definition 7.66) and the theorem that a C^1 map is differentiable (Theorem 7.68), and the spherical volume element of Example 7.98 — which is itself an instance of Appendix A.44. From Chapter 10: Proposition 10.67 and Theorem 10.75, both proved there.

A.45.2 The kernel is integrable, and uniformly so

Lemma A.518 (The excised ball)**.** *For $0 \le \delta < \varepsilon$,*

$$\int_{\delta \le |\boldsymbol{z}| \le \varepsilon} \frac{\mathrm{d}^3 z}{|\boldsymbol{z}|} = 2\pi \left(\varepsilon^2 - \delta^2\right)\,, \tag{A.817}$$

so that for every bounded g on $\mathbb{R}^3$,

$$\left| \int_{\delta \le |\boldsymbol{z}| \le \varepsilon} \Phi(\boldsymbol{z})\, g(\boldsymbol{x} - \boldsymbol{z})\, \mathrm{d}^3 z \right| \le \frac{1}{2}\,(\sup |g|)\,\varepsilon^2 \qquad \text{for every } \boldsymbol{x} \in \mathbb{R}^3\,. \tag{A.818}$$

Rests on Example 7.98, Theorem 7.97, and Definition 7.93.

Proof. *Derives Lemma A.518.*

The shell $\delta \le |\boldsymbol{z}| \le \varepsilon$ is a compact region decomposing into finitely many simple regions (Definition 7.95), and $|\boldsymbol{z}|^{-1}$ is continuous on it, hence integrable. Passing to

spherical coordinates, whose Jacobian determinant is $r^2 \sin\theta$ (Example 7.98, licensed by Theorem 7.97),

$$\int_{\delta\le|\boldsymbol{z}|\le\varepsilon} \frac{\mathrm{d}^3 z}{|\boldsymbol{z}|} = \int_0^{2\pi} \mathrm{d}\varphi \int_0^{\pi} \sin\theta \,\mathrm{d}\theta \int_\delta^\varepsilon \frac{r^2}{r}\,\mathrm{d}r = 2\pi\cdot 2\cdot \frac{\varepsilon^2-\delta^2}{2}\,, \tag{A.819}$$

which is Equation (A.817). Bounding the integrand of Equation (A.818) by $(4\pi\,|\boldsymbol{z}|)^{-1}\sup|g|$ and dropping δ gives the estimate, and the bound does not involve $\boldsymbol{x}$. □

Proposition A.519 (The convolution exists)**.** *Let g be continuous on $\mathbb{R}^3$ with support in $|\boldsymbol{y}|\le R_0$. For $\varepsilon>0$ put*

$$(\Phi * g)_\varepsilon(\boldsymbol{x}) = \int_{|\boldsymbol{z}|\ge\varepsilon} \Phi(\boldsymbol{z})\, g(\boldsymbol{x}-\boldsymbol{z})\,\mathrm{d}^3 z\,, \tag{A.820}$$

a proper integral over a compact region for each $\boldsymbol{x}$. As $\varepsilon\to 0$ these converge uniformly on $\mathbb{R}^3$ to a continuous limit $\Phi * g$, and

$$(\Phi * g)(\boldsymbol{x}) = -\frac{1}{4\pi}\int_{\mathbb{R}^3} \frac{g(\boldsymbol{y})}{|\boldsymbol{x}-\boldsymbol{y}|}\,\mathrm{d}^3 y\,, \tag{A.821}$$

the right-hand side being understood in the same excised sense.
Rests on Lemma A.518 and Theorems 7.8 and 7.97.

Proof. *Derives Proposition A.519.*

Fix $\boldsymbol{x}$ in a bounded set. The integrand of Equation (A.820) vanishes unless $\boldsymbol{x}-\boldsymbol{z}$ lies in the support of g, i.e. unless $\boldsymbol{z}$ lies in a fixed bounded set; so the integration is over a compact shell, on which the integrand is continuous, and the integral exists. For $0<\varepsilon'<\varepsilon$ the difference is the integral over the shell $\varepsilon'\le|\boldsymbol{z}|\le\varepsilon$, bounded by $\frac{1}{2}(\sup|g|)\,\varepsilon^2$ uniformly in $\boldsymbol{x}$ by Equation (A.818). The family is therefore uniformly Cauchy, so it converges pointwise (Theorem 7.8) and the convergence is uniform.

Each truncation is continuous in $\boldsymbol{x}$: on a compact shell and for $\boldsymbol{x}$ in a compact ball the integrand is uniformly continuous (Theorem 7.25), so the integral is continuous in the parameter — this is the last clause of Theorem 7.77, whose proof needs only that bound. A uniform limit of continuous functions is continuous: given $\varepsilon>0$ choose ε_0 with $\left|(\Phi*g)-(\Phi*g)_{\varepsilon_0}\right|<\varepsilon/3$ everywhere and then use the continuity of the truncation at the point in question, the three errors summing to ε; this is the argument already made for series in Lemma 9.6. Finally Equation (A.821) is the substitution $\boldsymbol{y}=\boldsymbol{x}-\boldsymbol{z}$, a translation composed with a reflection, whose Jacobian determinant has modulus one (Theorem 7.97). □

A.45.3 Differentiating under the integral sign

The singular factor in Equation (A.820) does not depend on $\boldsymbol{x}$. That is the whole trick: every derivative in $\boldsymbol{x}$ falls on g, where it is harmless, and the singularity is never differentiated.

Lemma A.520 (A C^1 limit)**.** *Let $O\subseteq\mathbb{R}^n$ be open and let $f_j : O\to\mathbb{R}$ be of class C^1, converging pointwise on O to f, with each $\partial_i f_j$ converging uniformly on O to a function h_i. Then f is of class C^1 on O and $\partial_i f = h_i$.*
Rests on Theorems 7.42 and 7.43 and Definition 7.66.

Proof. *Derives Lemma A.520.*

Each h_i is continuous, being a uniform limit of continuous functions (the $\varepsilon/3$ argument recalled in the proof of Proposition A.519). Fix $\boldsymbol{x}\in O$ and $\sigma>0$ with the segment $\boldsymbol{x}+s\boldsymbol{e}_i$,

$|s| \le \sigma$, inside O. By the fundamental theorem of calculus (Theorem 7.43) applied to the one-variable C^1 function $s \mapsto f_j(\boldsymbol{x} + s\boldsymbol{e}_i)$,

$$f_j\left(\boldsymbol{x} + t\boldsymbol{e}_i\right) - f_j(\boldsymbol{x}) = \int_0^t \partial_i f_j\left(\boldsymbol{x} + s\boldsymbol{e}_i\right) \mathrm{d}s\,, \qquad |t| \le \sigma\,. \tag{A.822}$$

The left-hand side tends to $f(\boldsymbol{x} + t\boldsymbol{e}_i) - f(\boldsymbol{x})$; on the right the error is at most $|t| \sup_O |\partial_i f_j - h_i|$, which tends to zero. Hence $f(\boldsymbol{x} + t\boldsymbol{e}_i) - f(\boldsymbol{x}) = \int_0^t h_i(\boldsymbol{x} + s\boldsymbol{e}_i)\,\mathrm{d}s$, and since h_i is continuous, differentiating at $t = 0$ (Theorem 7.42) gives $\partial_i f(\boldsymbol{x}) = h_i(\boldsymbol{x})$. All partial derivatives of f exist and are continuous, so f is of class C^1 (Definition 7.66) and differentiable (Theorem 7.68). □

Proposition A.521 (Every derivative passes onto g)**.** *Let g be of class C^k, $k \ge 1$, with compact support. Then $\Phi * g$ is of class C^k and, for every multi-index of order at most k,*

$$\partial_{i_1} \cdots \partial_{i_m} \left(\Phi * g\right) = \Phi * \left(\partial_{i_1} \cdots \partial_{i_m} g\right)\,, \qquad m \le k\,. \tag{A.823}$$

In particular $\Phi * g$ is smooth when g is.

Rests on Lemmas A.518 and A.520 and Theorem 7.77.

Proof. *Derives Proposition A.521.*

It suffices to prove the case $m = 1$ and to iterate, since $\partial_i g$ is again C^{k-1} with compact support.

Fix a ball $|\boldsymbol{x}| < \rho$. For $\boldsymbol{x}$ in it, the integrand of Equation (A.820) vanishes unless $\boldsymbol{z}$ lies in the compact shell $D_\varepsilon = \{\varepsilon \le |\boldsymbol{z}| \le \rho + R_0\}$, which decomposes into finitely many simple regions. On $D_\varepsilon \times (-\rho, \rho)$ the function $(\boldsymbol{z}, x^i) \mapsto \Phi(\boldsymbol{z}) g(\boldsymbol{x} - \boldsymbol{z})$ is continuous with continuous derivative $-\Phi(\boldsymbol{z})\left(\partial_i g\right)(\boldsymbol{x} - \boldsymbol{z})$ in the parameter x^i — the kernel is bounded away from its singularity on D_ε — so the Leibniz integral rule (Theorem 7.77) applies in each variable x^i separately and gives

$$\partial_i \left(\Phi * g\right)_\varepsilon (\boldsymbol{x}) = \int_{|\boldsymbol{z}| \ge \varepsilon} \Phi(\boldsymbol{z})\left(\partial_i g\right)(\boldsymbol{x} - \boldsymbol{z})\,\mathrm{d}^3 z = \left(\Phi * \partial_i g\right)_\varepsilon (\boldsymbol{x})\,, \tag{A.824}$$

the sign of the chain rule being absorbed because the derivative is taken with respect to $\boldsymbol{x}$ and not $\boldsymbol{z}$: writing $g(\boldsymbol{x} - \boldsymbol{z})$ and differentiating in x^i produces $\left(\partial_i g\right)(\boldsymbol{x} - \boldsymbol{z})$ with no sign at all.

By Proposition A.519 applied to g and to $\partial_i g$, both sides of Equation (A.824) converge uniformly as $\varepsilon \to 0$, the first to $\Phi * g$ and the second to $\Phi * \partial_i g$; each truncation is C^1 on the ball by the same Leibniz rule. Lemma A.520 therefore gives $\partial_i\left(\Phi * g\right) = \Phi * \partial_i g$ there, and ρ was arbitrary. □

A.45.4 Poisson's equation

Proposition A.522 (The Newtonian potential inverts the Laplacian)**.** *Let $\boldsymbol{u}$ be smooth with compact support and let $\boldsymbol{w} = \Phi * \boldsymbol{u}$, meaning the convolution taken componentwise. Then $\boldsymbol{w}$ is smooth and*

$$\nabla^2 \boldsymbol{w} = \boldsymbol{u} \qquad \textit{everywhere on } \mathbb{R}^3\,. \tag{A.825}$$

Rests on Propositions 10.67 and A.521 and Definition 17.38.

Proof. *Derives Proposition A.522.*

Smoothness and $\nabla^2 w_i = \Phi * \left(\nabla^2 u_i\right)$ are Proposition A.521. Fix $\boldsymbol{x}$ and put $\varphi(\boldsymbol{z}) = u_i(\boldsymbol{x} - \boldsymbol{z})$, a smooth function of compact support and hence a test function in the sense of Definition 17.38. Its Laplacian in $\boldsymbol{z}$ is $\left(\nabla^2 u_i\right)(\boldsymbol{x} - \boldsymbol{z})$, the two sign changes of the chain rule cancelling in a second derivative. By Proposition 10.67, which says precisely that $\int_{\mathbb{R}^3} \Phi\, \nabla^2 \varphi\, \mathrm{d}^3 z = \varphi(\boldsymbol{0})$ for every test function,

$$\left[\Phi * \left(\nabla^2 u_i\right)\right](\boldsymbol{x}) = \int_{\mathbb{R}^3} \Phi(\boldsymbol{z})\,\left(\nabla^2 u_i\right)(\boldsymbol{x} - \boldsymbol{z})\,\mathrm{d}^3 z = \varphi(\boldsymbol{0}) = u_i(\boldsymbol{x})\,, \tag{A.826}$$

which is Equation (A.825) componentwise. Note that the integral on the left is the excised one of Proposition A.519 and the one in Proposition 10.67 is the same limit, the kernel being absolutely integrable near the origin by Equation (A.817). □

A.45.5 Existence of the decomposition

Theorem A.523 (Helmholtz decomposition: existence)**.** *Let $\boldsymbol{u}$ be smooth on $\mathbb{R}^3$ with compact support and let $\boldsymbol{w} = \Phi * \boldsymbol{u}$. Then $\phi = \nabla \cdot \boldsymbol{w}$ and $\boldsymbol{\Psi} = -\nabla \times \boldsymbol{w}$ are smooth, satisfy Equation (7.130), and*

$$\boldsymbol{u} = \nabla\phi + \nabla \times \boldsymbol{\Psi}\,, \qquad \nabla \cdot \boldsymbol{\Psi} = 0\,. \tag{A.827}$$

Rests on Propositions 7.91 and A.522 and Equation (7.105).

Proof. *Derives Theorem A.523.*

$\boldsymbol{w}$ is smooth by Proposition A.521, so ϕ and $\boldsymbol{\Psi}$ are, and Equation (A.821) identifies $\boldsymbol{w}$ with the field displayed in Equation (7.130). The identity Equation (7.105) of Proposition 7.91, applied to the C^2 field $\boldsymbol{w}$ and rearranged, reads

$$\nabla^2 \boldsymbol{w} = \nabla\left(\nabla \cdot \boldsymbol{w}\right) - \nabla \times \left(\nabla \times \boldsymbol{w}\right) = \nabla\phi + \nabla \times \boldsymbol{\Psi}\,, \tag{A.828}$$

and the left-hand side is $\boldsymbol{u}$ by Equation (A.825). Finally $\nabla \cdot \boldsymbol{\Psi} = -\nabla \cdot (\nabla \times \boldsymbol{w}) = 0$ by Equation (7.104), and $\nabla \times \nabla\phi = \boldsymbol{0}$ by Equation (7.103), so the first summand is irrotational and the second solenoidal, as Theorem 7.105 asserts. □

A.45.6 Decay at infinity

The uniqueness clause of Theorem 7.105 speaks of fields that tend to zero at infinity. The potentials just constructed do, and with a definite rate.

Proposition A.524 (Rate of decay)**.** *With $\boldsymbol{u}$ supported in $|\boldsymbol{y}| \le R_0$ and $\boldsymbol{w} = \Phi * \boldsymbol{u}$ as above, there is a constant C, depending only on R_0 and $\sup|\boldsymbol{u}|$, such that for $|\boldsymbol{x}| \ge 2R_0$*

$$|\boldsymbol{w}(\boldsymbol{x})| \le \frac{C}{|\boldsymbol{x}|}\,, \qquad |\partial_i \boldsymbol{w}(\boldsymbol{x})| \le \frac{C}{|\boldsymbol{x}|^2}\,, \qquad |\partial_i\partial_j \boldsymbol{w}(\boldsymbol{x})| \le \frac{C}{|\boldsymbol{x}|^3}\,. \tag{A.829}$$

In particular $\nabla\phi$ and $\nabla \times \boldsymbol{\Psi}$ tend to zero at infinity, and so do ϕ and $\boldsymbol{\Psi}$ themselves.

Rests on Propositions A.519 and A.521 and Theorem 7.77.

Proof. *Derives Proposition A.524.*

Use the form Equation (A.821), in which the integration runs over the fixed compact ball $|\boldsymbol{y}| \le R_0$ and the singular factor now depends on $\boldsymbol{x}$. For $|\boldsymbol{x}| \ge 2R_0$ and $\boldsymbol{y}$ in that ball,

$$|\boldsymbol{x} - \boldsymbol{y}| \ge |\boldsymbol{x}| - R_0 \ge \tfrac{1}{2}\,|\boldsymbol{x}|\,, \tag{A.830}$$

so the integrand is bounded and no singularity is present: the integral is proper, and the Leibniz rule (Theorem 7.77) may be applied directly in each x^i, differentiating the kernel. Doing so,

$$\partial_i \frac{1}{|\boldsymbol{x} - \boldsymbol{y}|} = -\frac{x^i - y^i}{|\boldsymbol{x} - \boldsymbol{y}|^3}\,, \qquad \left|\partial_i \frac{1}{|\boldsymbol{x} - \boldsymbol{y}|}\right| \le \frac{1}{|\boldsymbol{x} - \boldsymbol{y}|^2}\,, \tag{A.831}$$

and similarly the second derivatives of the kernel are bounded by $3\,|\boldsymbol{x} - \boldsymbol{y}|^{-3}$. Inserting Equation (A.830) and bounding the integral by the volume $\frac{4}{3}\pi R_0^3$ of the ball containing the support times the supremum of the integrand gives all three estimates of Equation (A.829) with the single constant $C = 8R_0^3 \sup|\boldsymbol{u}|$: the three bounds so obtained are $\frac{2}{3}R_0^3 \sup|\boldsymbol{u}|$, $\frac{4}{3}R_0^3 \sup|\boldsymbol{u}|$ and $8R_0^3 \sup|\boldsymbol{u}|$ respectively, the factor $(4\pi)^{-1}$ cancelling against $\frac{4}{3}\pi$ and the remaining powers of 2 coming from Equation (A.830). The bound is not sharp and is not meant to be. Since $\nabla\phi$ and $\nabla \times \boldsymbol{\Psi}$ are built from second derivatives of $\boldsymbol{w}$, they are $O\left(|\boldsymbol{x}|^{-3}\right)$; ϕ and $\boldsymbol{\Psi}$ are built from first derivatives and are $O\left(|\boldsymbol{x}|^{-2}\right)$. □

A.45.7 Uniqueness

Theorem A.525 (Helmholtz decomposition: uniqueness)**.** *Let ϕ_1, ϕ_2 be C^3 scalar fields and $\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2$ be C^3 vector fields on $\mathbb{R}^3$ with*

$$\nabla\phi_1 + \nabla \times \boldsymbol{\Psi}_1 = \nabla\phi_2 + \nabla \times \boldsymbol{\Psi}_2\,, \tag{A.832}$$

and suppose each of the four fields $\nabla\phi_1, \nabla \times \boldsymbol{\Psi}_1, \nabla\phi_2, \nabla \times \boldsymbol{\Psi}_2$ tends to zero as $|\boldsymbol{x}| \to \infty$. Then $\nabla\phi_1 = \nabla\phi_2$ and $\nabla \times \boldsymbol{\Psi}_1 = \nabla \times \boldsymbol{\Psi}_2$.

Rests on Proposition 7.91, Theorem 10.75, and Equation (7.105).

Proof. *Derives Theorem A.525.*

Put $\boldsymbol{h} = \nabla\phi_1 - \nabla\phi_2 = \nabla \times \boldsymbol{\Psi}_2 - \nabla \times \boldsymbol{\Psi}_1$, a C^2 field tending to zero at infinity. Read from the left it is a gradient, so $\nabla \times \boldsymbol{h} = \boldsymbol{0}$ by Equation (7.103); read from the right it is a curl, so $\nabla \cdot \boldsymbol{h} = 0$ by Equation (7.104). Hence Equation (7.105) gives

$$\nabla^2\boldsymbol{h} = \nabla\,(\nabla \cdot \boldsymbol{h}) - \nabla \times (\nabla \times \boldsymbol{h}) = \boldsymbol{0}\,, \tag{A.833}$$

so each Cartesian component h_i is harmonic on the whole of $\mathbb{R}^3$ (Definition 10.66).

Fix $\boldsymbol{x}_0$ and let $R > |\boldsymbol{x}_0|$. On the open ball B_R the function h_i is harmonic and continuous on the closure, so by the weak maximum principle (Theorem 10.75) it attains both its extremes on the sphere $|\boldsymbol{x}| = R$:

$$\min_{|\boldsymbol{x}|=R} h_i \le h_i(\boldsymbol{x}_0) \le \max_{|\boldsymbol{x}|=R} h_i\,. \tag{A.834}$$

Both bounds tend to zero as $R \to \infty$, because $\boldsymbol{h} \to \boldsymbol{0}$ at infinity: given $\varepsilon > 0$ there is R_1 with $|\boldsymbol{h}| < \varepsilon$ outside B_{R_1}, so for $R > R_1$ the two extremes lie in $(-\varepsilon, \varepsilon)$. Hence $|h_i(\boldsymbol{x}_0)| < \varepsilon$ for every $\varepsilon > 0$, and $\boldsymbol{h} = \boldsymbol{0}$ identically. □

Proof of Theorem 7.105. *Derives Theorem 7.105.*

Existence, with the potentials Equation (7.130), is Theorem A.523; the two summands are respectively curl-free and divergence-free by the same theorem, which is what Proposition 7.91 contributes; and the uniqueness clause is Theorem A.525. That the potentials constructed here do satisfy the decay hypothesis of the uniqueness clause is Proposition A.524. □

A.45.8 Reading the result

Remark A.526 (What is quoted here)*.* Nothing from outside this treatise. Two results are taken from a *later* chapter, and that is worth stating plainly rather than letting a reader discover it: Proposition 10.67, which identifies Equation (A.816) as the fundamental solution of the Laplacian, and Theorem 10.75, the weak maximum principle. Both are proved in Chapter 10, both are proved there without any appeal to Theorem 7.105 or to anything in this section, and neither proof uses the Helmholtz decomposition anywhere in the book. So the forward reference is a matter of exposition, not of logic. It is also the reason the derivation belongs in an appendix: placed in Chapter 7 it would have to import a chapter that has not yet been written, and placed in Chapter 10 it would be far from the vector calculus it completes.

The remaining ingredients are all from Chapter 7 and are listed in Appendix A.45.1. Three things are deliberately *not* used: no Lebesgue convergence theorem — the passage to the limit $\varepsilon \to 0$ is uniform, by Lemma A.518, and needs only the Cauchy criterion; no Fourier transform, although the decomposition has a one-line proof in Fourier variables that hides every question of convergence; and no theory of distributions beyond the single statement of Proposition 10.67, which is quoted as an identity between numbers, tested on one explicit test function.

Remark A.527 (The hypotheses, and what happens without them). Compact support is a convenience, not a necessity. The same argument runs whenever $\boldsymbol{u}$ decays faster than $|\boldsymbol{x}|^{-2}$ with first derivatives decaying faster than $|\boldsymbol{x}|^{-3}$: the excision estimate Equation (A.818) is untouched, and the only new work is to check that the integral over the far region converges, which those rates supply.

Some decay is indispensable, and the failure is not subtle. The constant field $\boldsymbol{u} = \hat{\boldsymbol{z}}$ is a gradient, $\boldsymbol{u} = \nabla z$; it is also a curl, $\boldsymbol{u} = \nabla \times (x\,\hat{\boldsymbol{y}})$, as one checks from Equation (7.87) by evaluating the three components of the curl of $(0, x, 0)$. So the split Equation (7.129) exists for it in two ways — everything in the gradient, or everything in the curl — and nothing in the local differential identities can choose between them. The decay condition in the uniqueness clause of Theorem 7.105 is what excludes such a field, and Theorem A.525 shows it is exactly enough.

The regularity stated in Theorem A.525 is a small strengthening of the chapter's phrasing, which leaves it implicit: the argument differentiates the potentials three times, so C^3 potentials — equivalently C^2 summands — are what Equation (A.833) needs. For the potentials constructed here the point is moot, since they are smooth.

Rests on Theorems 7.105 and A.525 and Equation (7.87).

Remark A.528. Appendix A.45 discharges the derivation owed at Theorem 7.105 of Chapter 7. It also settles a debt recorded in Part III: Remark 30.46 of Chapter 30 states that the textbook route to the separation of P and S waves needs exactly this theorem — existence and uniqueness of the split, with the decay conditions that make it unique — and that Part II did not carry it, which is why the chapter re-routed the proof of Phenomenon 30.43 onto the acoustic tensor and plane waves instead. That re-routing stands as an independent derivation and is not superseded; what changes is that the alternative route is now available. The second site named there, Proposition 30.48, could not be re-routed: its displacement potentials are introduced as an ansatz sufficient for an existence claim, and the chapter says so. With Theorem A.523 in hand that ansatz can be upgraded to a decomposition, and the Rayleigh wave shown to be the general plane-strain surface motion rather than one exhibited example.

A.46 The Quotient Manifold Theorem

This appendix proves Theorem 13.67 of Chapter 13: when a Lie group G of dimension d acts smoothly, freely and properly on a manifold M_n, the set of orbits carries exactly one smooth structure making the projection Equation (13.156) a submersion, its dimension is $n - d$, its orbits are embedded copies of G, and a function on it is smooth exactly when its pullback is. It is the second of the two standard ways a manifold is produced — the first, Theorem 13.62, cuts a manifold out with equations; this one makes one by identifying points — and it is the one physics reaches for whenever a symmetry is quotiented away.

The consumer that makes the theorem urgent is Theorem 24.47 of Chapter 24: the Marsden–Weinstein reduced phase space is a level set of a momentum map divided by the isotropy group of that level, and the entire smooth structure of the reduced space comes from the statement proved here. The same theorem is what makes a configuration space of *shapes* out of a configuration space of frames, and what turns a gauge orbit space into a manifold wherever the action is free.

The proof runs in five movements: the orbit space is Hausdorff and second countable (Appendix A.46.2); orbits are embedded submanifolds diffeomorphic to G (Appendix A.46.3); through every point there is a slice, a transversal meeting nearby orbits at most once (Appendix A.46.4); the slices are the charts (Appendix A.46.5); and the universal property follows, with it the uniqueness of the structure (Appendix A.46.6). The heart is the slice, and it is built with the constant rank theorem Theorem 13.63 and the inverse function

theorem Corollary A.292 and nothing else. In particular no Lie algebra, no exponential map and no fundamental vector field is used: only that G is a manifold on which multiplication and inversion are smooth.

A.46.1 What the derivation uses, and what it must state itself

From Chapter 13: the definition of the action and of freeness and properness (Definition 13.66), smooth maps and their differentials (Definitions 13.49 and 13.52), immersions, submersions and embeddings (Definition 13.53), embedded submanifolds (Definition 13.54), the differentiable structure (Definition 13.47) and the constant rank theorem (Theorem 13.63) together with the observation of Remark 13.65 that it transfers to manifolds by being read in charts. From Appendix A.28: the inverse function theorem Corollary A.292. From Chapter 6: compactness (Definition 6.9), the continuous image of a compact set (Proposition 6.10) and homeomorphisms (Definition 6.7). From Chapter 14: nothing beyond the definition of a Lie group as a manifold with smooth multiplication and inversion.

Two separation properties are named in the statement of Theorem 13.67 and are defined nowhere in Part II; Chapter 6 carries neither. They are stated here, and the gap is recorded as a Part II debt in Remark A.541.

Definition A.529 (Hausdorff; second countable; locally compact)**.** A topological space X is *Hausdorff* if any two distinct points have disjoint open neighbourhoods; *second countable* if its topology has a countable basis, i.e. a countable family of open sets of which every open set is a union; and *locally compact* if every point has a compact neighbourhood.

Rests on Definitions 6.1, 6.2 and 6.9.

Every manifold in the sense of Definition 13.48 is assumed Hausdorff and second countable, as is standard and as the statement of Theorem 13.67 presupposes when it asserts those properties of the quotient. Every manifold is also locally compact and first countable, both because it is locally homeomorphic to $\mathbb{R}^n$: a closed coordinate ball is a compact neighbourhood, and the coordinate balls of rational radius about a point form a countable neighbourhood basis. First countability is what licenses the sequential arguments used below, in which a property is established by testing it on convergent sequences.

Lemma A.530 (Two elementary facts about compactness)**.** *Let X be Hausdorff.*

1. *Every compact $C \subseteq X$ is closed.*
2. *If X is in addition locally compact and $F : Y \longrightarrow X$ is continuous and proper — the preimage of every compact set is compact — then F carries closed sets to closed sets.*

Rests on Definitions 6.9 and A.529 and Proposition 6.10.

Proof. *Derives Lemma A.530.*

(1) Let $q \notin C$. For each $c \in C$ choose disjoint open $V_c \ni q$ and $W_c \ni c$; the W_c cover C, so finitely many $W_{c_1}, \dots, W_{c_m}$ do, and $V_{c_1} \cap \dots \cap V_{c_m}$ is an open neighbourhood of q disjoint from C. Hence the complement of C is open.

(2) Let $A \subseteq Y$ be closed and let q lie in the closure of $F(A)$. Choose a compact neighbourhood K of q. Then $F^{-1}(K)$ is compact by properness, so $A \cap F^{-1}(K)$ is compact — a closed subset of a compact set — and $F\left(A \cap F^{-1}(K)\right) = F(A) \cap K$ is compact by Proposition 6.10, hence closed by (1). Every neighbourhood of q contained in the interior of K meets $F(A)$, and the intersection lies in K; so q lies in the closure of $F(A) \cap K$, which is $F(A) \cap K$ itself. Thus $q \in F(A)$. □

Throughout the rest of the section, G is a Lie group of dimension d, $M = M_n$ a manifold, and the action is smooth, free and proper in the sense of Definition 13.66. We write $L_g : M \to M$ for the diffeomorphism $P \mapsto g \cdot P$, with inverse $L_{g^{-1}}$, and

$$\Theta : G \times M \longrightarrow M \times M\,, \qquad \Theta(g, P) = (g \cdot P,\, P)\,, \tag{A.835}$$

for the map whose properness is the hypothesis. We write $\theta_P : G \to M$, $\theta_P(g) = g \cdot P$, for the orbit map, and $k = n - d$.

A.46.2 The orbit space is Hausdorff and second countable

Lemma A.531 (The projection is open)**.** *For every open $U \subseteq M$ the set $\pi(U)$ is open in M/G, and π is continuous and surjective.*
Rests on Definitions 6.2, 6.6 and 13.66.

Proof. *Derives Lemma A.531.*
Continuity and surjectivity are the definition of the quotient topology. For openness,

$$\pi^{-1}(\pi(U)) = \{Q \in M \mid \exists\, g \in G :\ Q \in g \cdot U\} = \bigcup_{g \in G} L_g(U)\,, \tag{A.836}$$

a union of open sets, since each L_g is a diffeomorphism. A set whose preimage is open is open in the quotient topology. □

Proposition A.532 (Separation and countability of the quotient)**.** *The orbit space M/G is Hausdorff and second countable, and every orbit is closed in M.*
Rests on Lemmas A.530 and A.531 and Definition 13.66.

Proof. *Derives Proposition A.532.*
The orbit relation is closed. $M \times M$ is Hausdorff and locally compact, being a manifold, and Θ of Equation (A.835) is continuous and proper by hypothesis. Write

$$\mathcal{R} = \Theta(G \times M) = \{(Q, P) \in M \times M \mid Q \in G \cdot P\}\,. \tag{A.837}$$

By Lemma A.530(2) this set is closed — $G \times M$ being closed in itself. Fixing P and intersecting with $M \times \{P\}$ shows each orbit closed in M.

Hausdorff. Let $\pi(P) \neq \pi(Q)$, so $(Q, P) \notin \mathcal{R}$. Since $\mathcal{R}$ is closed and the products of open sets form a basis of $M \times M$, there are open $U \ni P$ and $V \ni Q$ with $(V \times U) \cap \mathcal{R} = \varnothing$: no point of V lies on the orbit of a point of U. Then $\pi(U)$ and $\pi(V)$ are open by Lemma A.531, contain $\pi(P)$ and $\pi(Q)$, and are disjoint — a common point would be an orbit meeting both U and V, which is what was excluded.

Second countable. Let $\{B_i\}$ be a countable basis of M. Each $\pi(B_i)$ is open. If $W \subseteq M/G$ is open and $\pi(P) \in W$, then $P \in \pi^{-1}(W)$, which is open, so $P \in B_i \subseteq \pi^{-1}(W)$ for some i, whence $\pi(P) \in \pi(B_i) \subseteq \pi(\pi^{-1}(W)) = W$, the last equality by surjectivity of π. So $\{\pi(B_i)\}$ is a countable basis. □

A.46.3 Orbits are embedded submanifolds

Lemma A.533 (The orbit map has constant rank d)**.** *For every $P \in M$ the map $\theta_P : G \to M$ has the same rank at every point of G, and that rank is d; that is, θ_P is an injective immersion.*
Rests on Theorem 13.63 and Definitions 13.53 and 13.66.

Proof. *Derives Lemma A.533.*
Constancy. Write $\ell_g : G \to G$ for left translation $h \mapsto gh$, a diffeomorphism. The associativity of the action gives $\theta_P(gh) = g \cdot (h \cdot P)$, that is

$$\theta_P \circ \ell_g = L_g \circ \theta_P\,. \tag{A.838}$$

Differentiating at the identity e and using the chain rule for differentials of smooth maps (Definition 13.52), $(\theta_P)_{*g} \circ (\ell_g)_{*e} = (L_g)_{*P} \circ (\theta_P)_{*e}$. Both $(\ell_g)_{*e}$ and $(L_g)_{*P}$ are isomorphisms, so the ranks of $(\theta_P)_{*g}$ and $(\theta_P)_{*e}$ agree.

The rank is d. Suppose it were $\rho < d$. Read in charts, θ_P is a C^∞ map of an open subset of $\mathbb{R}^d$ into $\mathbb{R}^n$ of constant rank ρ, so by the constant rank theorem (Theorem 13.63,

transferred to manifolds as in Remark 13.65) there are coordinates about e and about P in which it reads $\left(u^1,\dots,u^d\right)\mapsto(u^1,\dots,u^\rho,0,\dots,0)$, which is Equation (13.152). Two distinct nearby points differing only in the coordinate $u^{\rho+1}$ then have the same image, so θ_P is not injective on any neighbourhood of e — and θ_P *is* injective, because $g\cdot P=h\cdot P$ gives $\left(h^{-1}g\right)\cdot P=P$ and hence $g=h$ by freeness (Definition 13.66). The rank is therefore d, and $(\theta_P)_{*g}$ is injective for every g. □

Proposition A.534 (Every orbit is an embedded copy of G)**.** *For every $P\in M$ the orbit map θ_P is an embedding, so $G\cdot P$ is an embedded submanifold of M of dimension d, diffeomorphic to G.*

Rests on Lemma A.533 and Definitions 13.53 and 13.54.

Proof. *Derives Proposition A.534.*

By Lemma A.533 the map is an injective immersion; by Definition 13.53 what remains is that it be a homeomorphism onto its image with the topology inherited from M. Equivalently, since G is first countable, that $g_j\cdot P\to g\cdot P$ in M implies $g_j\to g$ in G.

Let $g_j\cdot P\to Q=g\cdot P$. The set $C=\{g_j\cdot P\mid j\}\cup\{Q\}$ is compact — a convergent sequence with its limit — and so is $C\times\{P\}$; properness makes $\Theta^{-1}(C\times\{P\})$ compact, and it contains every (g_j,P). Hence the g_j lie in a compact subset K of G. Since G is first countable, a compact subset of it is sequentially compact, so every subsequence of (g_j) has a further subsequence converging in K; and if $g_{j_m}\to g'$, then continuity of the action gives $g'\cdot P=\lim g_{j_m}\cdot P=Q=g\cdot P$, whence $g'=g$ by freeness. A sequence in a compact set all of whose convergent subsequences have the limit g converges to g: otherwise some neighbourhood of g would be left by infinitely many terms, and those terms would have a convergent subsequence whose limit, lying in the complement of that neighbourhood, could not be g. So $g_j\to g$.

The image of an embedding is an embedded submanifold and the map is a diffeomorphism onto it, as recorded after Definition 13.54. □

A.46.4 The slice

Proposition A.535 (Existence of a slice)**.** *Let $P\in M$. There are an open neighbourhood $O\subseteq G$ of e, an embedded k-dimensional submanifold $S\subseteq M$ with $P\in S$ carried by a single chart $\varsigma:S\to B\subseteq\mathbb{R}^k$, and an open neighbourhood $\Omega=O\cdot S$ of P in M, such that*

1. *the map $F:O\times S\to\Omega$, $F(g,Q)=g\cdot Q$, is a diffeomorphism; and*
2. *S meets every orbit at most once: if $Q,Q'\in S$ and $Q'=g\cdot Q$ for some $g\in G$, then $g=e$ and $Q=Q'$.*

Rests on Lemma A.533, Corollary A.292, and Definition 13.66.

Proof. *Derives Proposition A.535.*

Choice of a transversal. By Lemma A.533 the image $V_P=(\theta_P)_{*e}(T_eG)$ is a d-dimensional subspace of T_PM. Choose a chart (ϕ,W) about P with $\phi(P)=0$; composing ϕ with a linear automorphism of $\mathbb{R}^n$ — itself a change of chart — we may assume $\phi_{*P}(V_P)=\mathbb{R}^d\times\{0\}$. For a small ball $B\subseteq\mathbb{R}^k$ put

$$S_0=\phi^{-1}(\{0\}\times B)\,,\qquad \varsigma=(\text{last } k \text{ coordinates of } \phi)\big|_{S_0}\,, \tag{A.839}$$

an embedded k-submanifold through P with $T_PS_0\oplus V_P=T_PM$, since ϕ_{*P} carries the two summands onto the complementary $\{0\}\times\mathbb{R}^k$ and $\mathbb{R}^d\times\{0\}$.

The map F is a local diffeomorphism. On the n-manifold $G\times S_0$ define $F(g,Q)=g\cdot Q$, smooth as a restriction of the action. Its differential at (e,P) acts on $T_eG\oplus T_PS_0$ by

$$F_{*(e,P)}(\xi,v)=(\theta_P)_{*e}\,\xi+v\,, \tag{A.840}$$

because $F(\cdot, P) = \theta_P$ and $F(e, \cdot)$ is the inclusion of S_0. By the choice of S_0 this is a linear isomorphism onto $T_P M$. Read in charts, F is a smooth map between open subsets of $\mathbb{R}^n$ with invertible derivative at the point, so the inverse function theorem (Corollary A.292) makes it a diffeomorphism of a neighbourhood of (e, P) onto an open neighbourhood of P. Shrinking, that neighbourhood contains a product $O \times S$ with $O \ni e$ open in G and $S \ni P$ open in S_0; put $\Omega = F(O \times S) = O \cdot S$, open in M. This is (1).

Shrinking to a genuine slice. Suppose (2) failed for every choice of S. Take $S_j = \phi^{-1}\left(\{0\} \times B_{1/j}\right)$, $j \in \mathbb{N}$, a basis of neighbourhoods of P in S_0. For each j there would then be $Q_j, Q_j' \in S_j$ and $g_j \neq e$ with $g_j \cdot Q_j = Q_j'$. Both sequences converge to P, so $C = \{Q_j\} \cup \left\{Q_j'\right\} \cup \{P\}$ is compact and $\Theta^{-1}(C \times C)$ is compact by properness; it contains every (g_j, Q_j), so the g_j lie in a compact subset of G and a subsequence converges, $g_{j_m} \to g$. Continuity of the action gives $g \cdot P = \lim g_{j_m} \cdot Q_{j_m} = \lim Q_{j_m}' = P$, so $g = e$ by freeness. Hence $g_{j_m} \in O$ and $Q_{j_m}, Q_{j_m}' \in S$ for m large. But then

$$F(g_{j_m}, Q_{j_m}) = Q_{j_m}' = F\left(e, Q_{j_m}'\right), \tag{A.841}$$

and the injectivity of F on $O \times S$ forces $g_{j_m} = e$, contradicting $g_j \neq e$. So some S_j satisfies (2); replace S by it and shrink O and Ω accordingly, which preserves (1). □

A.46.5 The charts and the smooth structure

Lemma A.536 (A slice is a chart domain downstairs)**.** *With O, S, Ω as in Proposition A.535, the set $\pi(S) = \pi(\Omega)$ is open in M/G and $\pi|_S : S \to \pi(S)$ is a homeomorphism. Writing $\varrho : \Omega \to S$ for the second component of F^{-1}, a smooth map, one has $\varrho = (\pi|_S)^{-1} \circ \pi$ on Ω.*
Rests on Proposition A.535, Lemma A.531, and Definition 6.7.

Proof. *Derives Lemma A.536.*

Every point of Ω is $g \cdot Q$ with $Q \in S$, so $\pi(\Omega) = \pi(S)$, and $\pi(\Omega)$ is open by Lemma A.531. The map $\pi|_S$ is continuous, surjective onto $\pi(S)$ by definition, and injective by Proposition A.535(2).

For $R \in \Omega$ write $R = F(g, Q) = g \cdot Q$ with $(g, Q) \in O \times S$ unique; then $\varrho(R) = Q$ lies on the orbit of R and in S, and by (2) it is the *only* point of S on that orbit. Hence $\varrho = (\pi|_S)^{-1} \circ \pi$ on Ω, and ϱ is smooth because F^{-1} is.

It remains to see that $(\pi|_S)^{-1}$ is continuous. The restriction $\pi|_\Omega : \Omega \to \pi(S)$ is continuous, surjective and open (Lemma A.531 applied to open subsets of Ω), hence a quotient map: a subset of $\pi(S)$ is open exactly when its preimage is. The map ϱ is continuous and constant on the fibres of $\pi|_\Omega$, by the previous paragraph; so for $U \subseteq S$ open the set $\left((\pi|_S)^{-1}\right)^{-1}(U) = \pi\left(\varrho^{-1}(U)\right)$ has preimage $\varrho^{-1}(U)$, which is open, and is therefore open. Thus $(\pi|_S)^{-1}$ is continuous and $\pi|_S$ is a homeomorphism onto the open set $\pi(S)$. □

Theorem A.537 (The smooth structure on the orbit space)**.** *The maps*

$$u = \varsigma \circ (\pi|_S)^{-1} : \pi(S) \longrightarrow B \subseteq \mathbb{R}^k, \tag{A.842}$$

one for each slice of Proposition A.535, form a k-dimensional atlas on M/G. With the differentiable structure it generates (Definition 13.47), M/G is a smooth manifold of dimension $k = n - d$, which is Equation (13.157), and π is a smooth surjective submersion.
Rests on Lemma A.536, Proposition A.532, and Definition 13.53.

Proof. *Derives Theorem A.537.*

The charts cover and are homeomorphisms. Each u is a homeomorphism of the open set $\pi(S)$ onto the open $B \subseteq \mathbb{R}^k$, being the composition of the homeomorphism $(\pi|_S)^{-1}$ of Lemma A.536 with the chart ς of S; and every orbit meets some slice, namely one built at any of its points.

Smooth compatibility. Let S and S' be two slices with $\pi(S)\cap\pi(S')\neq\varnothing$ and let $\pi(Q_0)$ lie in the intersection, $Q_0\in S$. The transition map is $u'\circ u^{-1}=\varsigma'\circ\tau\circ\varsigma^{-1}$, where $\tau=(\pi|_{S'})^{-1}\circ\pi|_S$ sends a point of S to the unique point of S' on its orbit. Put $Q_0'=\tau(Q_0)\in S'$ and let $g_0\in G$ with $Q_0'=g_0\cdot Q_0$. The set $L_{g_0}^{-1}(\Omega')$ is an open neighbourhood of Q_0 in M, and on it the map $\varrho'\circ L_{g_0}$ is smooth and takes each point to the unique point of S' on its orbit — which is τ where both are defined. Restricted to the embedded submanifold S, a smooth map remains smooth, so τ is smooth near Q_0 as a map $S\to S'$, and the transition map is smooth as a map between open subsets of $\mathbb{R}^k$. Exchanging the roles of S and S' gives smoothness of the inverse. Together with Proposition A.532 — Hausdorff and second countable — this makes M/G a smooth k-manifold.

π is a submersion. Fix $P\in M$, take the slice at P and use $F^{-1}:\Omega\to O\times S$ to give M the chart $(\kappa,\varsigma)\circ F^{-1}$ near P, where κ is any chart of G about e. In these coordinates, and with the chart u downstairs, π reads

$$(a,y)\longmapsto y\,,\qquad a\in\kappa(O)\subseteq\mathbb{R}^d\,,\ y\in B\subseteq\mathbb{R}^k\,, \tag{A.843}$$

because $\pi\left(F(g,Q)\right)=\pi(Q)$ and $u\left(\pi(Q)\right)=\varsigma(Q)$. A projection has surjective differential, so π is a submersion at P (Definition 13.53); and P was arbitrary. □

A.46.6 The universal property and uniqueness

Lemma A.538 (Submersions have smooth local sections)**.** *Let $p:M\to X$ be a smooth submersion between manifolds and let $P\in M$. There are an open $W\ni p(P)$ in X and a smooth $s:W\to M$ with $p\circ s=\mathrm{id}_W$ and $s\left(p(P)\right)=P$.*

Rests on Theorem 13.63 and Definitions 13.49 and 13.53.

Proof. *Derives Lemma A.538.*

A submersion has constant rank $\dim X$, so Theorem 13.63, read in charts as in Remark 13.65, supplies coordinates $\left(u^1,\dots,u^{\dim M}\right)$ about P and $\left(y^1,\dots,y^{\dim X}\right)$ about $p(P)$, both vanishing at the point, in which p reads $(u^1,\dots)\mapsto\left(u^1,\dots,u^{\dim X}\right)$. In those coordinates set $s(y)=(y,0)$, which is smooth and satisfies both requirements on the coordinate neighbourhood. □

Proposition A.539 (Universal property)**.** *Let $p:M\to X$ be a smooth surjective submersion and N a manifold. A map $f:X\to N$ is smooth if and only if $f\circ p$ is smooth.*

Rests on Lemma A.538 and Definition 13.49.

Proof. *Derives Proposition A.539.*

If f is smooth then so is $f\circ p$, a composition of smooth maps. Conversely let $f\circ p$ be smooth and let $x\in X$; pick P with $p(P)=x$, which exists by surjectivity, and a smooth local section s on $W\ni x$ from Lemma A.538. On W, $f=f\circ p\circ s$ is a composition of smooth maps, hence smooth; and f is continuous there for the same reason. Smoothness is local, so f is smooth. □

Corollary A.540 (Uniqueness of the smooth structure)**.** *There is at most one differentiable structure on the topological space M/G for which π is a smooth submersion.*

Rests on Proposition A.539 and Definition 13.47.

Proof. *Derives Corollary A.540.*

Let X and X' denote the same topological space M/G carrying two such structures. Since $\pi:M\to X'$ is smooth and $\pi:M\to X$ is a smooth surjective submersion, Proposition A.539 applied to $f=\mathrm{id}$ shows $\mathrm{id}:X\to X'$ is smooth. Exchanging the two gives $\mathrm{id}:X'\to X$ smooth. The identity is therefore a diffeomorphism, and two differentiable structures related by the identity diffeomorphism have the same smooth functions and hence the same maximal atlas. □

Proof of Theorem 13.67. *Derives Theorem 13.67.*

Assembling. The orbit space is Hausdorff and second countable by Proposition A.532; it carries a smooth structure of dimension $n - d$ for which Equation (13.156) is a submersion, by Theorem A.537, and Equation (13.157) is that dimension count; the structure is the only one with that property, by Corollary A.540; every orbit is an embedded submanifold diffeomorphic to G, by Proposition A.534; and a map f on the quotient is smooth exactly when $f \circ \pi$ is, by Proposition A.539 applied to the surjective submersion π. □

A.46.7 Reading the result

Remark A.541 (What is quoted here). No theorem from outside this treatise is used, and the two analytic inputs — the inverse function theorem Corollary A.292 and the constant rank theorem Theorem 13.63 — are proved in Appendix A.28 and in Chapter 13 respectively. What *is* missing from Part II, and is supplied here rather than cited, is point-set vocabulary: Chapter 6 defines neither the Hausdorff property nor second countability nor local compactness, although Theorem 13.67 asserts the first two of the quotient and Definition 13.48 tacitly assumes all of them of a manifold. They are stated in Definition A.529, and the two facts about them that the proof consumes — a compact subset of a Hausdorff space is closed, and a proper map into a locally compact Hausdorff space is closed — are proved in Lemma A.530. That is a Part II debt of about a page, and it is the only one this section incurs.

Two things are deliberately *not* used, and the omission is what keeps the section self-contained. There is no Lie algebra anywhere: the standard proof shows the orbit map to be an immersion by computing its differential as the fundamental vector field of an element of $\mathfrak{g}$ and appealing to freeness, which would import the machinery of Chapter 14; instead Lemma A.533 gets the same conclusion from the constant rank theorem alone, by observing that a constant-rank map of rank less than d cannot be injective near a point. And there is no partition of unity and no smooth-structure transport theorem: the charts are the slices, and the transitions are computed from the action itself.

Remark A.542 (Where each hypothesis is spent). Freeness is used twice and properness three times, and it is worth saying where, because Example 13.68 shows what happens when either fails but not which step breaks.

Freeness makes the orbit map injective, which is what forces its rank to be full in Lemma A.533 — and hence what gives every orbit the same dimension d, without which no dimension count is possible. It is used again in Proposition A.535 to identify the limiting group element as e. When it fails, orbits through different points have different dimensions, and the quotient is a stratified space: this is the SO(2) action of Example 13.68, whose orbit space is a half-line with a bad endpoint.

Properness gives the closedness of the orbit relation (Proposition A.532), and with it the Hausdorff property of the quotient; it gives the embedding property of the orbit map (Proposition A.534); and it gives the second shrinking in Proposition A.535, which is what turns a local transversal into a genuine slice. The irrational flow on the torus in Example 13.68 fails at the first of these, and every later step fails with it: no slice exists, because every orbit returns arbitrarily close to every point.

Rests on Propositions A.532 and A.535 and Example 13.68.

Remark A.543. Appendix A.46 discharges the derivation owed at Theorem 13.67 of Chapter 13. The statement is used at once in the same chapter, where it is the second of the two manufacturing theorems beside Theorem 13.62 and Corollary 13.64, and it is what makes Example 13.68 a statement about manifolds rather than about sets. Its principal consumer in the physical parts is Theorem 24.47 of Chapter 24: symplectic reduction divides a level set of a momentum map — a manifold by Theorem 13.62 — by the isotropy group of that

level, and every smooth statement about the reduced phase space, from the existence of its symplectic form to the descent of the dynamics, rests on the projection being a submersion with the universal property Proposition A.539. The one hypothesis to watch there is the same one watched here: the action must be free, and where it is not the reduced space is singular and none of this applies.

A.47 The Liouville–Arnold Theorem

This appendix proves Theorem 23.31 of Chapter 23: a system of f degrees of freedom possessing f functionally independent integrals $F_1 = \mathcal{H}, F_2, \dots, F_f$ in involution has each compact connected component of a regular level set diffeomorphic to the torus T^f, carries quasi-periodic motion on it, and admits action–angle variables in a neighbourhood of it. The last clause is what makes the theorem worth its length: it establishes Theorem 23.30 without the hypothesis of separability under which that theorem was proved, so that the action–angle description of a bounded integrable motion is a consequence of the involution relations alone and not of a lucky choice of coordinates.

Exactly one ingredient is imported, and Remark 23.32 already tells the reader which: that a discrete subgroup of $\mathbb{R}^f$ with compact quotient is a lattice of full rank. It is stated below as Theorem A.550 and used once, in Appendix A.47.3. Everything else is proved here from the symplectic material of Chapter 24 and the manifold material of Chapter 13; in particular Frobenius' theorem, which the argument needs twice, is Theorem 13.133 and is not reproved.

A.47.1 Setting and statement

Let (M, ω) be a symplectic manifold of dimension $2f$ (Definition 24.8), let $F_1, \dots, F_f$ be smooth real functions on M in involution,

$$\{F_a, F_b\} = 0\,, \qquad a, b = 1, \dots, f\,, \tag{A.844}$$

which is Equation (23.54), and write $\boldsymbol{F} = (F_1, \dots, F_f) : M \longrightarrow \mathbb{R}^f$. Fix $\boldsymbol{c} \in \mathbb{R}^f$ and suppose that $\boldsymbol{c}$ is a regular value, so that the differentials $\mathrm{d}F_1, \dots, \mathrm{d}F_f$ are linearly independent at every point of $M_{\boldsymbol{c}} = \boldsymbol{F}^{-1}(\boldsymbol{c})$. Denote by $X_a = X_{F_a}$ the Hamiltonian vector field of F_a, defined by $\iota_{X_a}\omega = \mathrm{d}F_a$ as in Equation (24.5); by Equation (24.7) it acts on functions as $X_a[g] = \{g, F_a\}$.

Remark A.544 (Units). Nothing here fixes the SI dimension of $F_2, \dots, F_f$: each is whatever conserved quantity it happens to be, and by Remark 24.11 the vector field X_a carries $[F_a]/\mathrm{J\,s}$, so its flow parameter carries $\mathrm{J\,s}/[F_a]$ — seconds when F_a is the energy, radians when it is an angular momentum. The action variables of Appendix A.47.4 carry $\mathrm{J\,s}$ without exception, because $\oint p_a \mathrm{d}q^a$ does; the angles are pure numbers. That is the whole of the dimensional bookkeeping, and it is worth stating because the lattice Γ of Appendix A.47.3 lives in a copy of $\mathbb{R}^f$ whose f axes carry f different units.

Theorem A.545 (Liouville–Arnold)**.** *Under the hypotheses above, let N be a compact connected component of $M_{\boldsymbol{c}}$. Then*

1. *N is an f-dimensional embedded submanifold of M, diffeomorphic to the torus $T^f = \mathbb{R}^f/\mathbb{Z}^f$, and it is Lagrangian: ω vanishes on $T_P N$ for every $P \in N$;*
2. *in the angular coordinates supplied by that diffeomorphism the flow of $\mathcal{H} = F_1$ is linear, so the motion is quasi-periodic with f frequencies;*
3. *on a neighbourhood of N there exist action–angle variables (I_a, θ^a), canonical, with $\mathcal{H} = \mathcal{H}(I)$ and with the equations of motion Equation (23.51) and Equation (23.52).*

Rests on Equations (23.54) and (24.8) and Theorem 13.133.

A.47.2 The commuting fields and the level set

Lemma A.546 (The Hamiltonian fields of the integrals)**.** *On the open set $U \subseteq M$ where $\mathrm{d}F_1, \dots, \mathrm{d}F_f$ are linearly independent:*

1. *$X_a[F_b] = 0$ for all a, b, so every X_a is tangent to every level set of $\boldsymbol{F}$;*
2. *$[X_a, X_b] = 0$ as vector fields;*
3. *$X_1, \dots, X_f$ are linearly independent at every point of U, and $\omega(X_a, X_b) = 0$.*

Rests on Equations (24.5), (24.8) and (A.844).

Proof. *Derives Lemma A.546.*

(1) By Equation (24.7), $X_a[F_b] = \{F_b, F_a\}$, which vanishes by Equation (A.844). A vector annihilating every $\mathrm{d}F_b$ is tangent to the level set through its base point.

(2) Equation (24.8) states that $X_{\{u,v\}} = -[X_u, X_v]$, so $[X_a, X_b] = -X_{\{F_a,F_b\}} = -X_0 = 0$. This is the one place where involution does its real work, and it is worth saying what it buys: the flows of the f integrals commute, which is what will make them into an action of an abelian group.

(3) Suppose $\sum_a \lambda^a X_a(P) = 0$ at some $P \in U$. Contracting with ω and using Equation (24.5) gives $\sum_a \lambda^a \,\mathrm{d}F_a(P) = 0$, whence every λ^a vanishes by the independence of the differentials. Finally $\omega(X_a, X_b) = \{F_a, F_b\} = 0$ by Equation (24.7) and Equation (A.844). □

Proposition A.547 (The components of the level set are the leaves of an integrable distribution)**.** *Let $D_P = \operatorname{span}\{X_1(P), \dots, X_f(P)\}$ for $P \in U$. Then D is a smooth distribution of constant rank f (Definition 13.130), it is involutive, and it is integrable. Every connected component N of $M_{\boldsymbol{c}} \cap U$ is an embedded f-dimensional submanifold with $T_P N = D_P$, hence an integral manifold of D, and it is Lagrangian.*

Rests on Lemma A.546 and Theorems 13.59 and 13.133.

Proof. *Derives Proposition A.547.*

Smoothness and constant rank are Lemma A.546(3); involutivity is Lemma A.546(2), since the bracket of two spanning fields vanishes and so lies in D trivially. Integrability then follows from Theorem 13.133, which also supplies, around each point of U, a chart in which D is spanned by the first f coordinate fields Equation (13.281), so that U is foliated by f-dimensional integral manifolds.

Since $\boldsymbol{c}$ is a regular value, $\boldsymbol{F}$ restricted to U is a submersion onto a neighbourhood of $\boldsymbol{c}$, and by the regular value theorem Theorem 13.59 the set $M_{\boldsymbol{c}} \cap U$ is an embedded submanifold of dimension $2f - f = f$ with $T_P(M_{\boldsymbol{c}} \cap U) = \ker \mathrm{d}\boldsymbol{F}_P$. By Lemma A.546(1) every $X_a(P)$ lies in that kernel, and by (3) the f of them are independent; a subspace of dimension f inside a space of dimension f is the whole of it, so $T_P(M_{\boldsymbol{c}} \cap U) = D_P$. Each connected component is therefore a connected integral manifold of D. That it is Lagrangian is Lemma A.546(3) again: ω evaluated on two vectors of $T_P N = D_P$ is a combination of the numbers $\omega(X_a, X_b)$, all zero. □

A.47.3 The translation action and the torus

From here on N is a *compact* connected component of $M_{\boldsymbol{c}}$.

Lemma A.548 (The joint flow is a translation action on the level set)**.** *Each X_a restricts to a complete vector field on N, and*

$$\Phi(\boldsymbol{t}) = \phi^1_{t^1} \circ \phi^2_{t^2} \circ \cdots \circ \phi^f_{t^f}\,, \qquad \boldsymbol{t} = (t^1, \dots, t^f) \in \mathbb{R}^f\,, \tag{A.845}$$

where ϕ^a is the flow of X_a, is independent of the order of the factors and defines a smooth action of the additive group $\mathbb{R}^f$ on N.

Rests on Lemma A.546, Proposition 13.131, and Theorem 13.125.

Proof. *Derives Lemma A.548.*

By Lemma A.546(1) each X_a is tangent to N, so its integral curves through points of N stay in N as long as they exist. A smooth vector field on a compact manifold is complete: by Theorem 13.125 each point has a neighbourhood on which the flow exists for a time at least ε depending on the neighbourhood, compactness extracts a finite subcover and hence a single $\varepsilon > 0$ that works everywhere on N, and the flow is then extended to all times by composing it with itself. Commutativity of the flows is Proposition 13.131 applied to $[X_a, X_b] = 0$, which makes the composition Equation (A.845) independent of the order and gives $\Phi(\boldsymbol{t})\Phi(\boldsymbol{t}') = \Phi(\boldsymbol{t}+\boldsymbol{t}')$; $\Phi(\mathbf{0})$ is the identity, and smoothness in $(\boldsymbol{t}, P)$ jointly is that of each flow. □

Lemma A.549 (The action is transitive)**.** *For every $P \in N$ the orbit map $\boldsymbol{t} \longmapsto \Phi(\boldsymbol{t})P$ is a local diffeomorphism of $\mathbb{R}^f$ onto its image; the orbit is open in N; and since N is connected there is exactly one orbit, so the action is transitive.*

Rests on Lemma A.548, Corollary A.292, and Proposition A.547.

Proof. *Derives Lemma A.549.*

Differentiating Equation (A.845) at $\boldsymbol{t} = \mathbf{0}$ sends the a-th basis vector of $\mathbb{R}^f$ to $X_a(P)$, and by Lemma A.546(3) together with Proposition A.547 those f vectors are a basis of $T_P N$. The differential of the orbit map at the origin is therefore an isomorphism onto $T_P N$, and the inverse function theorem (Corollary A.292, read in a chart of N) makes the orbit map a diffeomorphism of a neighbourhood of $\mathbf{0}$ onto a neighbourhood of P in N. Applying this at an arbitrary point $\Phi(\boldsymbol{t}_0)P$ of the orbit — which is legitimate because $\Phi(\boldsymbol{t}_0)$ is a diffeomorphism of N carrying the orbit map based at P to the one based at $\Phi(\boldsymbol{t}_0)P$ — shows that the orbit contains a neighbourhood of each of its points, i.e. is open. Distinct orbits are disjoint, so the orbits form a partition of N into open sets; connectedness leaves only one. □

Theorem A.550 (Quoted: discrete subgroups of a real vector space)**.** *Let $\Gamma \subseteq \mathbb{R}^f$ be a subgroup that is discrete, in the sense that some neighbourhood of $\mathbf{0}$ meets Γ only in $\mathbf{0}$. Then there are linearly independent vectors $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_k \in \mathbb{R}^f$, $k \le f$, with*

$$\Gamma = \left\{ \sum_{a=1}^{k} n^a \boldsymbol{\lambda}_a \;\middle|\; n^a \in \mathbb{Z} \right\}, \tag{A.846}$$

and the quotient group $\mathbb{R}^f/\Gamma$ is compact if and only if $k = f$, in which case it is diffeomorphic to the torus T^f.

Rests on Definitions 4.20 and 6.9.

Remark A.551 (What is quoted here). Theorem A.550 is the only statement in this section that is not proved, and Remark 23.32 of Chapter 23 names it as the debt in the same words. It is elementary — an induction on the dimension of the linear span of Γ, at each step choosing a shortest non-zero element of the part of Γ lying in a new direction and showing that discreteness makes it generate that direction's contribution — but it is a theorem about subgroups of $\mathbb{R}^f$, which is algebra and topology rather than mechanics, and it belongs with the material of Chapters 4 and 6, which do not carry it. The treatise's bibliography holds no source for it either, so the attribution is made in words: it is standard lattice theory, and the form used here is the one Arnold states in *Mathematical Methods of Classical Mechanics* (2nd edition, 1989), a work for which Chapter 24 also records that there is no key. Nothing else below is quoted.

Proposition A.552 (The component is a torus)**.** *Fix $P \in N$ and let $\Gamma = \left\{ \boldsymbol{t} \in \mathbb{R}^f \mid \Phi(\boldsymbol{t})P = P \right\}$. Then Γ is a discrete subgroup of $\mathbb{R}^f$ of rank exactly f, the induced map*

$$\mathbb{R}^f/\Gamma \longrightarrow N, \qquad \boldsymbol{t} + \Gamma \longmapsto \Phi(\boldsymbol{t})P, \tag{A.847}$$

is a diffeomorphism, and N is diffeomorphic to T^f.
Rests on Lemmas A.548 and A.549 and Theorem A.550.

Proof. *Derives Proposition A.552.*

Γ is a subgroup because Φ is an action, and it is discrete because Lemma A.549 makes the orbit map injective on some neighbourhood B of $\mathbf{0}$, so $\Gamma \cap B = \{\mathbf{0}\}$. The orbit map is constant on cosets of Γ and, by transitivity, descends to a bijection Equation (A.847) of $\mathbb{R}^f/\Gamma$ onto N; it is a local diffeomorphism by the same lemma, hence a diffeomorphism. Since N is compact so is $\mathbb{R}^f/\Gamma$, and Theorem A.550 then gives Γ of rank f and $\mathbb{R}^f/\Gamma \cong T^f$. □

Corollary A.553 (Quasi-periodic motion)**.** *Under the identification Equation (A.847), the Hamiltonian flow of $\mathcal{H} = F_1$ starting at P is $t \longmapsto t\,\boldsymbol{e}_1 + \Gamma$: a straight line in $\mathbb{R}^f$ read modulo the lattice. The motion is therefore quasi-periodic, with f frequencies fixed by the position of $\boldsymbol{e}_1$ relative to the lattice basis, and it is periodic exactly when the line closes, that is when $T\boldsymbol{e}_1 \in \Gamma$ for some $T > 0$.*
Rests on Proposition A.552 and Lemma A.548.

Proof. *Derives Corollary A.553.*

The Hamiltonian vector field of F_1 is X_1, whose flow is $\Phi(t\boldsymbol{e}_1)$ by Equation (A.845). The identification Equation (A.847) carries it to translation by $t\boldsymbol{e}_1$ on $\mathbb{R}^f/\Gamma$, and a translation flow on a torus closes if and only if its generator is commensurable with the lattice. Remark 23.33 records what happens to these tori under perturbation, and Theorem 32.51 is the statement that most of them survive. □

A.47.4 The action variables

Let $\boldsymbol{\lambda}_1(\boldsymbol{c}), \dots, \boldsymbol{\lambda}_f(\boldsymbol{c})$ be a basis of the lattice Γ supplied by Proposition A.552, chosen to depend smoothly on $\boldsymbol{c}$ — which is possible because a lattice basis is locally rigid: it is determined by finitely many periods, and a nearby torus has nearby periods, so the basis is continued uniquely. Write $\gamma_a(\boldsymbol{c})$ for the closed curve $t \in [0,1] \mapsto \Phi(t\boldsymbol{\lambda}_a)P$ on the torus $N(\boldsymbol{c})$; its homology class is independent of the base point P, and $\gamma_1, \dots, \gamma_f$ generate the first homology of N because they are the images of the lattice generators under Equation (A.847).

Definition A.554 (Actions on the torus)**.** With θ the tautological one-form of Equation (24.3), that is $\theta = p_a \mathrm{d}q^a$ in any canonical chart, put

$$I_a(\boldsymbol{c}) = \frac{1}{2\pi} \oint_{\gamma_a(\boldsymbol{c})} p_b \,\mathrm{d}q^b\,, \qquad a = 1, \dots, f\,. \tag{A.848}$$

Rests on Proposition A.552 and Equations (23.49) and (24.3).

Lemma A.555 (The actions are well defined)**.** *The integral Equation (A.848) depends only on the homology class of γ_a in N, not on the representing curve, and it depends smoothly on $\boldsymbol{c}$.*
Rests on Proposition A.547, Definition A.554, and Theorem A.311.

Proof. *Derives Lemma A.555.*

N is Lagrangian by Proposition A.547, so the restriction of $\omega = -\mathrm{d}\theta$ to N vanishes, that is the restriction of θ to N is a closed one-form. Two homologous closed curves on N bound a two-chain there, and Stokes' theorem (Theorem A.311) turns the difference of the two circuit integrals into the integral of $\mathrm{d}\theta = -\omega$ over that chain, which is zero. Smoothness in $\boldsymbol{c}$ follows because the curves $\gamma_a(\boldsymbol{c})$ were chosen to depend smoothly on $\boldsymbol{c}$ and θ is smooth. □

The next statement is the one the chapter's own proof of Theorem 23.30 assumed without proving: that the actions may replace the constants $\boldsymbol{c}$ as labels of the tori. Here it is a theorem, and its content is an exact identity between the Jacobian and the lattice.

Proposition A.556 (The Jacobian of the actions is the lattice matrix)**.** *With the conventions above,*

$$\frac{\partial I_a}{\partial c_b} = \frac{1}{2\pi}\left(\boldsymbol{\lambda}_a\right)^b , \qquad \textit{hence} \qquad \det\left(\frac{\partial I_a}{\partial c_b}\right) = \frac{\det(\boldsymbol{\lambda}_a)}{(2\pi)^f} \neq 0 , \tag{A.849}$$

so $\boldsymbol{c} \longmapsto \boldsymbol{I}(\boldsymbol{c})$ is a local diffeomorphism and the I_a are independent functions on a neighbourhood of N.

Rests on Definition A.554 and Propositions 22.18 and A.552.

Proof. *Derives Proposition A.556.*

Let S be the union of the tori $N(\boldsymbol{c}')$ for $\boldsymbol{c}'$ near $\boldsymbol{c}$, a neighbourhood of N in M, and on it define

$$W = \int^{\cdot} p_b \, \mathrm{d}q^b , \tag{A.850}$$

the integral taken along a path from a fixed reference point. Since the restriction of θ to each torus is closed (Lemma A.555), W is well defined up to the periods, and by Equation (A.848) carrying the path once around γ_a increases it by

$$\Delta_a W = 2\pi I_a(\boldsymbol{c}) . \tag{A.851}$$

Read W as a generating function of the second kind (Proposition 22.18) with the constants c_b as new momenta; the conjugate new coordinates are then $\tau^b = \partial W/\partial c_b$, and by the second of Hamilton's equations in the new variables (Theorem 22.6) each τ^b advances at unit rate along the flow of F_b and is constant along the flow of every other F_a. In other words τ^b is the b-th time coordinate of the action Equation (A.845). Carrying the path once around γ_a means applying $\Phi(\boldsymbol{\lambda}_a)$, so

$$\Delta_a \tau^b = \left(\boldsymbol{\lambda}_a\right)^b . \tag{A.852}$$

On the other hand $\tau^b = \partial W/\partial c_b$, and the circuit γ_a lies inside a single torus, so the constants c_b are fixed along it and the two differentiations may be exchanged under the circuit integral, exactly as in Equation (23.53):

$$\Delta_a \tau^b = \frac{\partial}{\partial c_b}\left(\Delta_a W\right) = \frac{\partial\left(2\pi I_a\right)}{\partial c_b} . \tag{A.853}$$

Comparing Equation (A.852) with Equation (A.853) gives Equation (A.849). The determinant of the lattice generators is non-zero because they are linearly independent (Theorem A.550). □

Example A.557 (The check that fixes the convention)**.** For one degree of freedom and a bounded orbit of energy E, $f = 1$ and $\boldsymbol{\lambda}_1$ is the period T of the motion, so Equation (A.849) reads $\mathrm{d}I/\mathrm{d}E = T/2\pi = 1/\omega$, which is exactly Equation (23.57), derived there from Hamilton's equations alone. For the harmonic oscillator, $I = E/\omega$ with ω independent of E, and both sides are $1/\omega$.

Rests on Proposition A.556 and Equation (23.57).

A.47.5 The angles, and the proof of the theorem

Proof of Theorem A.545. *Derives Theorem A.545.*

Part (1) is Proposition A.547 together with Proposition A.552; part (2) is Corollary A.553. It remains to build the action–angle variables.

By Proposition A.556 the map $\boldsymbol{c} \mapsto \boldsymbol{I}$ may be inverted on a neighbourhood, so W of Equation (A.850) may be regarded as a function $W(q, I)$ of the position coordinates and

the actions. It is again a generating function of the second kind, now with the I_a as new momenta, so

$$p_a = \frac{\partial W}{\partial q^a}\,, \qquad \theta^a = \frac{\partial W}{\partial I_a}\,, \tag{A.854}$$

which is Equation (23.50), and the transformation $(q,p) \mapsto (\theta, I)$ is canonical by Proposition 22.18. Repeating the exchange of differentiations that produced Equation (A.853), now with I_a in place of c_b,

$$\Delta_b \theta^a = \frac{\partial}{\partial I_a}\left(\Delta_b W\right) = \frac{\partial\left(2\pi I_b\right)}{\partial I_a} = 2\pi\,\delta^a_b\,, \tag{A.855}$$

so each θ^a is well defined modulo 2π and the f of them are angular coordinates on the torus. The I_a are functions of the c_b alone and are therefore constants of the motion; conversely $\mathcal{H} = F_1 = c_1$ is a function of the c_b, hence of the I_a, and contains no angle. Theorem 22.6 in the new variables then gives $\dot I_a = -\partial\mathcal{H}/\partial\theta^a = 0$, which is Equation (23.51), and $\dot\theta^a = \partial\mathcal{H}/\partial I_a = \omega^a(I)$, constant along the motion, which integrates to Equation (23.52). That is part (3), and with it Theorem 23.30 freed of the separability hypothesis under which it was originally proved. □

Remark A.558 (Where the projection to position space fails, and why that is the Maslov index). One step above deserves to be stated with its caveat rather than glossed. Writing $W = W(q,I)$ presumes that the f position coordinates q^a are good coordinates on the torus, i.e. that $N \to \mathbb{R}^f$, $P \mapsto q(P)$, is a local diffeomorphism. A Lagrangian torus need not project that way everywhere: the projection degenerates on the *caustics*, which for the one-dimensional bounded orbit are the two turning points, where $\partial p/\partial q$ blows up and the orbit folds back. The repair is local and is the one Proposition 22.18 already provides: near a point where some subset of the q^a fails, Legendre-transform in those variables and use a generating function of a different kind, whose arguments are the surviving q's and the corresponding p's. The circuit integrals Equation (A.848) and Equation (A.855) are unaffected, because $\oint\theta$ is defined on N without reference to any projection. What the caustics do affect is the semiclassical use of W: each fold crossed contributes a phase, and the count of them around γ_a is the Maslov index μ_a of Equation (23.62). The classical theorem is indifferent to it; the quantization condition is not.

Remark A.559. Appendix A.47 discharges the derivation owed at Theorem 23.31 of Chapter 23, and Remark 23.32 there states the single imported ingredient, restated above as Theorem A.550. Three connections back into the book are worth making explicit. First, the chapter observes that nothing in it depends on the theorem, because Proposition 23.20 establishes on its own that a separable system is integrable; what the appendix adds is the converse direction of usefulness — action–angle variables now exist for an integrable system whether or not anyone can separate its Hamilton–Jacobi equation, which is Theorem 23.30 with its hypothesis removed. Second, the tori built here are the objects whose fate under perturbation is the subject of Theorem 32.51, and Remark 23.33 is the honest statement that the hypotheses of Theorem 23.31 are met by almost no system. Third, the Lagrangian property established in Proposition A.547 is what Remark 24.13 of Chapter 24 has in mind when it lists the Liouville–Arnold tori among the statements of symplectic geometry that carry physical content: they are global objects, and a symplectic manifold has nothing local to say.

A.48 Marsden–Weinstein Reduction

This appendix proves Theorem 24.47 of Chapter 24: when a Lie group acts freely and properly on a symplectic manifold by symplectomorphisms with an equivariant momentum

map, and μ is a regular value of that map, the quotient Equation (24.31) of the level set by the isotropy group of μ is again a symplectic manifold, of the dimension Equation (24.32), and every invariant Hamiltonian descends to it with its dynamics intact.

The chapter states the geometric core of the argument and does not prove it: the tangent space to the level set of the momentum map is the symplectic orthogonal of the group orbit (Lemma 24.46, which *is* proved there), and the degenerate directions of the restricted form turn out to be exactly the orbit directions of the isotropy group G_μ, so that quotienting by that group removes the degeneracy and removes nothing else. That sentence is Appendix A.48.4 below. Everything around it — that the level set is a manifold, that the quotient is one, that the descended form is well defined, closed and nondegenerate — is what makes the theorem long, and it is written out here.

Remark A.560 (One quoted input less than the chapter says). The paragraph following the proof link in Chapter 24 names one input as quoted rather than proved: the quotient-manifold theorem, that a free and proper action of a Lie group on a manifold has a smooth quotient with the projection a submersion, described there as differential topology belonging in Chapter 13, "which does not yet carry it". That sentence is now out of date, and the correction is worth making rather than leaving as a discrepancy between two pages of the same book: Chapter 13 carries the statement as Theorem 13.67, with its own proof, and this section therefore *cites* it rather than quoting it. Nothing in Appendix A.48 is imported from outside the treatise. The role of that theorem is unchanged and remains confined to producing the smooth structure on M_μ: the symplectic content — existence, uniqueness, closedness and nondegeneracy of the reduced form, and the dimension count — rests on Lemma 24.46 and on Appendix A.48.2 alone.

A.48.1 Setting and statement

Throughout, (M,ω) is a symplectic manifold (Definition 24.8), G is a Lie group with Lie algebra $\mathfrak{g}$ acting smoothly on M by symplectomorphisms (Definition 24.17), and ξ_M denotes the vector field generating the action of $\xi \in \mathfrak{g}$, so that $\xi_M(x) = \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} \exp(t\xi)\cdot x$. The action is assumed *free* and *proper* in the sense of Definition 13.66. The momentum map $\boldsymbol{J} : M \longrightarrow \mathfrak{g}^*$ is the one of Definition 24.43, defined by Equation (24.28), and the interior product is written ι_X, as in Chapter 24; the same operation is written i_X in Chapter 13.

Definition A.561 (Coadjoint action and equivariance)**.** The *coadjoint action* of G on $\mathfrak{g}^*$ is

$$\left\langle \mathrm{Ad}^*_g \mu, \eta \right\rangle = \left\langle \mu, \mathrm{Ad}_{g^{-1}} \eta \right\rangle , \qquad \mu \in \mathfrak{g}^* , \quad \eta \in \mathfrak{g} , \tag{A.856}$$

a left action, and its infinitesimal form is $\mathrm{ad}^*_\xi = \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} \mathrm{Ad}^*_{\exp(t\xi)}$, so that

$$\left\langle \mathrm{ad}^*_\xi \, \mu, \eta \right\rangle = - \left\langle \mu, [\xi, \eta] \right\rangle . \tag{A.857}$$

A momentum map is *equivariant* if $\boldsymbol{J}(g\cdot x) = \mathrm{Ad}^*_g \boldsymbol{J}(x)$ for every g and x. For $\mu \in \mathfrak{g}^*$ write $G_\mu = \left\{ g \in G \mid \mathrm{Ad}^*_g \mu = \mu \right\}$ for its isotropy group, a closed subgroup, and $\mathfrak{g}_\mu$ for the Lie algebra of G_μ.

Rests on Definitions 4.20, 14.2 and 24.43.

Remark A.562 (Conventions and units). Two sign conventions circulate for Equation (A.856), differing by $g \leftrightarrow g^{-1}$; with the one chosen here the coadjoint action is a left action and equivariance reads $\boldsymbol{J} \circ g = \mathrm{Ad}^*_g \circ \boldsymbol{J}$ without an inverse. Sources that define Ad^* by $\langle \mathrm{Ad}^*_g \mu, \eta \rangle = \langle \mu, \mathrm{Ad}_g \eta \rangle$ write the same statement with g^{-1} on the right, and the two agree.

This section is pure geometry and the only dimensional statement it needs is the one Remark 24.11 already fixes: ω carries J s, and a component J_ξ of the momentum map carries

the SI unit of the conserved quantity it generates — so $\boldsymbol{J}$ is not a single dimensioned object but a covector whose f components may carry f different units, exactly as the lattice of Appendix A.47 does. The reduced form ω_μ inherits J s from ω, because it is defined by pullback.

Theorem A.563 (Marsden–Weinstein reduction)**.** *Let G act freely and properly on (M,ω) by symplectomorphisms with an equivariant momentum map $\boldsymbol{J}$, and let $\mu \in \mathfrak{g}^*$. Write $N = \boldsymbol{J}^{-1}(\mu)$, let $i : N \hookrightarrow M$ be the inclusion and $\pi : N \longrightarrow M_\mu = N/G_\mu$ the projection. Then*

1. *every value of $\boldsymbol{J}$ is regular, and N is a closed embedded submanifold of M of dimension $\dim M - \dim G$ with $T_xN = \ker \mathrm{d}\boldsymbol{J}_x$;*
2. *G_μ acts freely and properly on N, and M_μ is a smooth manifold with π a surjective submersion and*
$$\dim M_\mu = \dim M - \dim G - \dim G_\mu\,, \tag{A.858}$$
which is Equation (24.32);
3. *there is exactly one two-form ω_μ on M_μ with $\pi^*\omega_\mu = i^*\omega$, and it is symplectic;*
4. *every G-invariant $H \in C^\infty(M)$ descends to $H_\mu \in C^\infty(M_\mu)$ with $H_\mu \circ \pi = H \circ i$; the flow of X_H preserves N; and π carries it to the flow of X_{H_μ} on (M_μ,ω_μ).*

Rests on Lemma 24.46, Theorem 13.67, and Definition A.561.

A.48.2 Two facts of symplectic linear algebra

Lemma A.564 (Dimension and double orthogonal)**.** *Let (V,Ω) be a finite-dimensional real vector space with a nondegenerate antisymmetric bilinear form, and for a subspace $W \subseteq V$ let $W^\Omega = \{v \in V \mid \Omega(v,w) = 0 \text{ for all } w \in W\}$. Then*

$$\dim W + \dim W^\Omega = \dim V\,, \qquad \left(W^\Omega\right)^\Omega = W\,. \tag{A.859}$$

Rests on Definition 24.2 and Theorem 5.38.

Proof. *Derives Lemma A.564.*

Nondegeneracy makes $\flat : v \longmapsto \Omega(v,\cdot)$ an injective, hence bijective, linear map $V \longrightarrow V^*$. By definition $W^\Omega = \flat^{-1}(W^\circ)$, where $W^\circ \subseteq V^*$ is the annihilator of W; and $\dim W^\circ = \dim V - \dim W$ by rank–nullity (Theorem 5.38) applied to the restriction map $V^* \longrightarrow W^*$, which is surjective. That is the first identity. For the second, $W \subseteq (W^\Omega)^\Omega$ is immediate from antisymmetry, and applying the first identity twice gives $\dim (W^\Omega)^\Omega = \dim V - \dim W^\Omega = \dim W$, so the inclusion is an equality. □

A.48.3 The level set is a manifold

Proposition A.565 (Freeness makes every value regular)**.** *If the action is free then $\mathrm{d}\boldsymbol{J}_x$ is surjective at every $x \in M$. Consequently every $\mu \in \mathfrak{g}^*$ is a regular value, $N = \boldsymbol{J}^{-1}(\mu)$ is a closed embedded submanifold of dimension $\dim M - \dim G$, and $T_xN = \ker \mathrm{d}\boldsymbol{J}_x = \left(T_x(G\cdot x)\right)^\omega$.*

Rests on Lemmas 24.46 and A.564 and Theorem 13.59.

Proof. *Derives Proposition A.565.*

Fix x and write $W = T_x(G\cdot x)$, the span of the values $\xi_M(x)$, $\xi \in \mathfrak{g}$. Freeness makes the map $\xi \mapsto \xi_M(x)$ injective: if $\xi_M(x) = 0$ then the curve $t \mapsto \exp(t\xi)\cdot x$ has vanishing velocity and, being an integral curve of ξ_M through x, is constant by the uniqueness clause

of Theorem 13.125; so $\exp(t\xi)\cdot x = x$ for all t, whence $\exp(t\xi) = e$ by freeness and $\xi = 0$. Hence $\dim W = \dim \mathfrak{g} = \dim G$.

By Lemma 24.46, $\ker \mathrm{d}\boldsymbol{J}_x = W^\omega$, so Equation (A.859) gives $\dim \ker \mathrm{d}\boldsymbol{J}_x = \dim M - \dim W = \dim M - \dim G$, and by rank–nullity the rank of $\mathrm{d}\boldsymbol{J}_x$ is $\dim G = \dim \mathfrak{g}^*$: the differential is surjective. The regular value theorem Theorem 13.59 then makes $N = \boldsymbol{J}^{-1}(\mu)$ an embedded submanifold of dimension $\dim M - \dim G$, closed because it is the preimage of a point under a continuous map, with tangent space the kernel. □

Proposition A.566 (The isotropy group acts, and the quotient is smooth)**.** *G_μ maps N into itself, and its action on N is free and proper. Hence $M_\mu = N/G_\mu$ carries exactly one smooth structure making $\pi : N \longrightarrow M_\mu$ a surjective submersion; it is Hausdorff and second countable, the fibres of π are the G_μ-orbits, and Equation (A.858) holds.*

Rests on Definition A.561, Theorem 13.67, and Proposition A.565.

Proof. *Derives Proposition A.566.*

For $g \in G_\mu$ and $x \in N$, equivariance gives $\boldsymbol{J}(g\cdot x) = \mathrm{Ad}^*_g \boldsymbol{J}(x) = \mathrm{Ad}^*_g \mu = \mu$, so $g\cdot x \in N$. The restricted action is free because the G-action is, and it is proper: G_μ is a closed subgroup of G and N is a closed subset of M, so the preimage of a compact set under $(g,x) \mapsto (g\cdot x, x)$ on $G_\mu \times N$ is a closed subset of the corresponding preimage on $G \times M$, hence compact. Theorem 13.67 applied to this action supplies the smooth structure, the submersion, the separation properties and the fibres, together with the dimension $\dim M_\mu = \dim N - \dim G_\mu$, which with Proposition A.565 is Equation (A.858). □

A.48.4 The degenerate directions of the restricted form

Write $\omega_N = i^*\omega$ for the restriction of ω to N. It is a closed two-form, but it is not symplectic: it has a radical, and identifying that radical is the heart of the theorem.

Lemma A.567 (The differential of the momentum map along the orbit)**.** *Let $x \in N$, so that $\boldsymbol{J}(x) = \mu$. Then for every $\xi \in \mathfrak{g}$*

$$\mathrm{d}\boldsymbol{J}_x(\xi_M(x)) = \mathrm{ad}^*_\xi \mu\,, \tag{A.860}$$

and consequently $\xi_M(x) \in \ker \mathrm{d}\boldsymbol{J}_x$ if and only if $\xi \in \mathfrak{g}_\mu$.

Rests on Definition A.561, Proposition A.565, and Theorem 13.125.

Proof. *Derives Lemma A.567.*

Apply equivariance along the one-parameter subgroup $g(t) = \exp(t\xi)$ and differentiate at $t = 0$. The left-hand side of $\boldsymbol{J}(g(t)\cdot x) = \mathrm{Ad}^*_{g(t)}\mu$ differentiates by the chain rule to $\mathrm{d}\boldsymbol{J}_x(\xi_M(x))$, since $\frac{\mathrm{d}}{\mathrm{d}t}\big|_0 g(t)\cdot x = \xi_M(x)$ by definition; the right-hand side differentiates to $\mathrm{ad}^*_\xi \mu$ by Definition A.561. That is Equation (A.860).

For the second claim, $\mathrm{ad}^*_\xi \mu = 0$ certainly holds when $\xi \in \mathfrak{g}_\mu$, since then $\mathrm{Ad}^*_{\exp(t\xi)}\mu = \mu$ for all t. Conversely suppose $\mathrm{ad}^*_\xi \mu = 0$ and set $\nu(t) = \mathrm{Ad}^*_{\exp(t\xi)}\mu$. Differentiating Equation (A.856) in t and using the group law gives the linear differential equation $\dot\nu(t) = \mathrm{ad}^*_\xi \nu(t)$ with $\nu(0) = \mu$. The constant function $\nu \equiv \mu$ solves it, because $\mathrm{ad}^*_\xi \mu = 0$, and the solution of a linear equation with given initial value is unique (Theorem 9.8); so $\nu(t) = \mu$ for all t, that is $\exp(t\xi) \in G_\mu$ for all t, that is $\xi \in \mathfrak{g}_\mu$. □

Theorem A.568 (The radical is the isotropy orbit)**.** *For every $x \in N$,*

$$\ker (\omega_N)_x := \{v \in T_xN \;\mid\; (\omega_N)_x(v,w) = 0 \text{ for all } w \in T_xN\} = T_x(G_\mu \cdot x)\,. \tag{A.861}$$

Rests on Lemmas 24.46, A.564 and A.567.

Proof. *Derives Theorem A.568.*

Write $W = T_x(G \cdot x)$ as before. By Proposition A.565 and Equation (24.30), $T_xN = \ker \mathrm{d}\boldsymbol{J}_x = W^\omega$. Since ω_N is the restriction of ω, the radical of ω_N at x is $T_xN \cap (T_xN)^\omega$, the second symplectic orthogonal being taken in the symplectic vector space (T_xM, ω_x). Hence, using the second identity of Equation (A.859),

$$\ker (\omega_N)_x = W^\omega \cap (W^\omega)^\omega = W^\omega \cap W = T_x (G \cdot x) \cap \ker \mathrm{d}\boldsymbol{J}_x \,. \tag{A.862}$$

A vector of $T_x(G \cdot x)$ is $\xi_M(x)$ for exactly one $\xi \in \mathfrak{g}$, by the injectivity established in Proposition A.565, and it lies in $\ker \mathrm{d}\boldsymbol{J}_x$ precisely when $\xi \in \mathfrak{g}_\mu$, by Lemma A.567. The intersection is therefore $\{\xi_M(x) \mid \xi \in \mathfrak{g}_\mu\} = T_x(G_\mu \cdot x)$. □

Remark A.569 (Why the quotient is by a group and not merely by a foliation). The radical of a closed two-form of constant rank is always an involutive distribution, so it is integrable by Theorem 13.133 and the manifold is foliated by its leaves. The check is two lines with Cartan's magic formula: if $\iota_X\omega_N = 0$ and $\iota_Y\omega_N = 0$ then $\mathcal{L}_X\omega_N = \mathrm{d}\,\iota_X\omega_N + \iota_X\mathrm{d}\omega_N = 0$ by Equation (13.276) and $\mathrm{d}\omega_N = i^*\mathrm{d}\omega = 0$, and then Equation (13.277) gives $\iota_{[X,Y]}\omega_N = \mathcal{L}_X\iota_Y\omega_N - \iota_Y\mathcal{L}_X\omega_N = 0$, so $[X, Y]$ lies in the radical too. What Theorem A.568 adds is that the leaves of that foliation are exactly the G_μ-orbits. Without it one could still divide N by the characteristic foliation and get a symplectic quotient, but the quotient would be a leaf space with no reason to be Hausdorff or to be a manifold at all. Identifying the leaves as orbits of a group acting freely and properly is what lets Theorem 13.67 supply the smooth structure — and it is also what makes the reduced space computable in practice, since a physicist knows the symmetry group and does not know the foliation.

A.48.5 Descent of the form

Proposition A.570 (Existence and uniqueness of the reduced form)**.** *There is exactly one two-form ω_μ on M_μ with $\pi^*\omega_\mu = \omega_N$, it is smooth, and it is closed and nondegenerate.*
Rests on Theorems 24.16 and A.568 and Proposition A.566.

Proof. *Derives Proposition A.570.*

Definition. Let $[x] \in M_\mu$ and $u, v \in T_{[x]}M_\mu$. Since π is a submersion, $\mathrm{d}\pi_x$ is onto, so there are $\tilde{u}, \tilde{v} \in T_xN$ with $\mathrm{d}\pi_x\tilde{u} = u$ and $\mathrm{d}\pi_x\tilde{v} = v$; set

$$(\omega_\mu)_{[x]}(u, v) = (\omega_N)_x (\tilde{u}, \tilde{v}) \,. \tag{A.863}$$

This is independent of the lifts. The fibres of π are the G_μ-orbits (Proposition A.566), so $\ker \mathrm{d}\pi_x = T_x(G_\mu \cdot x)$, which is exactly the radical of $(\omega_N)_x$ by Theorem A.568; changing a lift changes it by an element of that radical, which ω_N does not see.

Independence of the point in the fibre. Let $g \in G_\mu$ and replace x by $g \cdot x$, lifting through $\mathrm{d}(g\cdot)_x\tilde{u}$ and $\mathrm{d}(g\cdot)_x\tilde{v}$, which are legitimate lifts because $\pi \circ g = \pi$. Since G acts by symplectomorphisms, $g^*\omega = \omega$, and g maps N to N, so $g^*\omega_N = \omega_N$ and the two evaluations agree.

Smoothness. A submersion admits local smooth sections: near any $[x]$ choose s with $\pi \circ s = \mathrm{id}$; then $\omega_\mu = s^*\omega_N$ on that neighbourhood by Equation (A.863), a pullback of a smooth form along a smooth map.

Uniqueness. If $\pi^*\omega_\mu = \pi^*\omega'_\mu$ then, $\mathrm{d}\pi_x$ being surjective at every x and π being surjective, the two forms agree on every pair of tangent vectors at every point.

Closedness. $\pi^*\mathrm{d}\omega_\mu = \mathrm{d}\pi^*\omega_\mu = \mathrm{d}\,i^*\omega = i^*\mathrm{d}\omega = 0$, the exterior derivative commuting with pullback; and π^* is injective on forms because $\mathrm{d}\pi$ is surjective. Hence $\mathrm{d}\omega_\mu = 0$.

Nondegeneracy. Suppose $(\omega_\mu)_{[x]}(u, \cdot) = 0$. Lift u to $\tilde{u} \in T_xN$; then $(\omega_N)_x(\tilde{u}, \tilde{v}) = 0$ for every $\tilde{v} \in T_xN$, because every v is $\mathrm{d}\pi_x\tilde{v}$. So $\tilde{u}$ lies in the radical, which is $\ker \mathrm{d}\pi_x$, and therefore $u = \mathrm{d}\pi_x\tilde{u} = 0$. □

A.48.6 Descent of the dynamics

Proposition A.571 (Invariant Hamiltonians descend with their flows)**.** *Let $H \in C^\infty(M)$ be G-invariant. Then $H \circ i$ is G_μ-invariant and descends to a smooth H_μ on M_μ; the flow of X_H preserves N; and $\mathrm{d}\pi_x(X_H(x)) = X_{H_\mu}([x])$ for every $x \in N$.*

Rests on Theorem 24.45, Proposition A.570, and Equation (24.5).

Proof. *Derives Proposition A.571.*

Invariance under G implies invariance under the subgroup G_μ, and by the last clause of Theorem 13.67 a function on M_μ is smooth exactly when its composition with π is; the G_μ-invariant function $H \circ i$ therefore defines a smooth H_μ with $H_\mu \circ \pi = H \circ i$.

That the flow of X_H preserves N is Noether's theorem in the form Theorem 24.45: G acts by symplectomorphisms and preserves H, so $\boldsymbol{J}$ is constant along the flow, and a trajectory starting on $\boldsymbol{J}^{-1}(\mu)$ stays there. In particular $X_H(x) \in T_xN$ for $x \in N$.

For the last claim, let $v \in T_{[x]}M_\mu$ and lift it to $\tilde{v} \in T_xN$. Using Equation (A.863), then Equation (24.5), then $H_\mu \circ \pi = H \circ i$:

$$(\omega_\mu)_{[x]}\left(\mathrm{d}\pi_x X_H(x),\, v\right) = (\omega_N)_x\left(X_H(x), \tilde{v}\right) = \mathrm{d}H_x(\tilde{v}) = \mathrm{d}\left(H_\mu\right)_{[x]}(v)\,. \tag{A.864}$$

Since v was arbitrary and ω_μ is nondegenerate, the vector $\mathrm{d}\pi_x X_H(x)$ is the unique one satisfying Equation (24.5) for H_μ, that is $X_{H_\mu}([x])$. □

Proof of Theorem A.563. *Derives Theorem A.563.*

Part (1) is Proposition A.565, part (2) is Proposition A.566, part (3) is Proposition A.570 and part (4) is Proposition A.571. Since $M_\mu = \boldsymbol{J}^{-1}(\mu)/G_\mu$ is Equation (24.31) and Equation (A.858) is Equation (24.32), this is Theorem 24.47. □

A.48.7 The example the chapter names

Example A.572 (Rotational reduction of the central-force problem)**.** Take $M = T^*\mathbb{R}^3$ with $\omega = \mathrm{d}q^i \wedge \mathrm{d}p_i$ and $G = \mathrm{SO}(3)$ acting by $(\boldsymbol{q}, \boldsymbol{p}) \mapsto (R\boldsymbol{q}, R\boldsymbol{p})$. This is an action by symplectomorphisms, and Example 24.44 computes its momentum map: $\boldsymbol{J} = \boldsymbol{q} \times \boldsymbol{p} = \boldsymbol{L}$, the angular momentum, in J s.

The action is *not* free on all of M — a rotation about $\boldsymbol{q}$ fixes any point with $\boldsymbol{p}$ parallel to $\boldsymbol{q}$, and every rotation fixes the origin — so the hypotheses hold only on the open, $\mathrm{SO}(3)$-invariant subset $M^\times = \{(\boldsymbol{q}, \boldsymbol{p}) \mid \boldsymbol{q} \times \boldsymbol{p} \neq \boldsymbol{0}\}$, where a rotation fixing two independent vectors is the identity. That restriction is not a technicality to be waved through: it is exactly the collinear and radial motions, which the reduced picture cannot describe and which Section 27.4 treats separately.

On $M^\times$ fix $\mu = \boldsymbol{L}$ with $\ell = |\boldsymbol{L}| > 0$. The isotropy group of μ under the coadjoint action of $\mathrm{SO}(3)$ — which for $\mathrm{SO}(3)$ is the ordinary rotation action on $\mathbb{R}^3$, the algebra carrying an invariant inner product — is the group $\mathrm{SO}(2)$ of rotations about the axis of $\boldsymbol{L}$, of dimension 1. So Equation (A.858) gives

$$\dim M_\mu = 6 - 3 - 1 = 2\,, \tag{A.865}$$

and the reduced space is the two-dimensional radial phase space (r, p_r). For $H = |\boldsymbol{p}|^2/2m + V(r)$, resolving $\boldsymbol{p}$ into its radial and transverse parts on $\boldsymbol{J}^{-1}(\mu)$ gives $|\boldsymbol{p}|^2 = p_r^2 + \ell^2/r^2$, so

$$H_\mu(r, p_r) = \frac{p_r^2}{2m} + \frac{\ell^2}{2mr^2} + V(r)\,, \tag{A.866}$$

which is the effective potential of Equation (27.32), and the centrifugal term $\ell^2/2mr^2$ carries J as it must. This is the identification Remark 24.48 asserts: the effective potential *is* the reduced Hamiltonian at the fixed value of the angular momentum, and the disappearance of

the two angular coordinates is not a trick of the coordinate system but the reduction of six dimensions to two.

Rests on Theorem A.563, Example 24.44, and Equation (27.32).

Example A.573 (The abelian case, and eliminating a cyclic coordinate)**.** If G is abelian the coadjoint action is trivial, so $G_\mu = G$ for every μ and Equation (A.858) reads $\dim M_\mu = \dim M - 2\dim G$: two dimensions are lost per generator. For $G = \mathbb{R}$ acting by translation in a cyclic coordinate q^1, the momentum map is p_1, fixing μ freezes that momentum, and dividing by the group deletes q^1 — the pair (q^1, p_1) disappears together, which is precisely the elementary manoeuvre Remark 24.48 says every physicist performs without naming it.

Rests on Theorem A.563 and Remark 24.48.

A.48.8 What is quoted here

Remark A.574 (What is quoted here)*.* Nothing. The inputs are Lemma 24.46, proved in Chapter 24; the regular value theorem Theorem 13.59, the quotient-manifold theorem Theorem 13.67, Frobenius' theorem Theorem 13.133 and the Cartan identities Equation (13.276) and Equation (13.277), all proved in Chapter 13; rank–nullity Theorem 5.38 from Chapter 5; and the uniqueness of solutions of an ordinary differential equation, Theorem 9.8. Remark A.560 records that the chapter's own prose still describes the quotient-manifold theorem as owed by Part II, which it no longer is.

Remark A.575 (What each hypothesis does, and what is not treated)*.* The three hypotheses divide the work cleanly, and it is worth recording which of them fails first in practice.

Freeness is used twice: in Proposition A.565, to make every value of $\boldsymbol{J}$ regular and N a manifold, and in Proposition A.566, as a hypothesis of Theorem 13.67. *Properness* is used once, also through Theorem 13.67, and what it buys is that the quotient topology is Hausdorff — Example 13.68 exhibits a free but improper action whose orbit space is not. *Equivariance* is used once, in Lemma A.567, and it is what makes $\mathfrak{g}_\mu$ rather than $\mathfrak{g}$ appear in Equation (A.861); without it the radical would not be the tangent space to an orbit of anything.

When freeness fails — which, as Example A.572 shows, is the normal state of affairs at the symmetric configurations — M_μ is not a manifold but a stratified space, a union of manifolds of different dimensions glued along their boundaries. That is *singular* reduction, and it is not treated in this treatise. The honest summary is that the theorem proved here describes the generic stratum and says nothing about the symmetric configurations, which for the central-force problem are exactly the collinear orbits.

Remark A.576 (Attribution)*.* The theorem is Marsden and Weinstein's, from *Reduction of symplectic manifolds with symmetry* (Reports on Mathematical Physics **5**, 1974, 121–130). That paper has no entry in this treatise's bibliography — Chapter 24 records the omission in its own header, together with those of Abraham–Marsden and Arnold — so the attribution is made in words, on the footing of Darboux's memoir in Remark A.81. Nothing above rests on it: the derivation is carried out in full from results proved in this book.

Remark A.577*.* Appendix A.48 discharges the derivation owed at Theorem 24.47 of Chapter 24. Two threads run back into the book from here. Remark 24.48 lists four familiar manoeuvres — dropping a cyclic pair, passing to the centre-of-mass frame, reducing the two-body problem to the radial one, and using the effective potential — and says that each is this theorem applied to a particular group; Example A.572 and Example A.573 above verify two of them, including the identification of Equation (27.32) as the reduced Hamiltonian. And the same remark points at the Lie–Poisson bracket Definition 24.40 as reduction with the whole group divided out, which is why the free rigid body of Proposition 24.41 lives on three variables rather than the six a cotangent bundle would demand: the dimension count is Equation (A.858) once more, with $M = T^*\,\mathrm{SO}(3)$ of dimension six, $G = \mathrm{SO}(3)$ of

dimension three, and a generic $G_\mu = \mathrm{SO}(2)$ removing one further dimension to leave the two-dimensional coadjoint orbit — a sphere of constant $|\boldsymbol{L}|$ — on which the Euler equations Equation (24.27) run.

A.49 The Stone–von Neumann Theorem

This appendix proves Theorem 25.34 of Chapter 25: for finitely many degrees of freedom an irreducible representation of the canonical commutation relations in Weyl's exponentiated form is unitarily equivalent to the Schrödinger representation on $L^2(\mathbb{R}^f)$, so that the quantum kinematics of a system with finitely many canonical pairs is fixed, up to equivalence, by the classical bracket algebra alone.

Nothing in the argument is physics, and the statement it proves stands on the books twice. Theorem 12.114 of Chapter 12 states it with the care about domains and irreducibility that the subject needs, and Remarks 12.115 and 25.35 both record that the proof was owed and that it belongs beside the Hilbert-space material rather than in a chapter of mechanics. This section discharges both debts at once. It is therefore written throughout in the notation of Definition 12.109 — dimensionful parameters, generators $\boldsymbol{Q}$ and $\boldsymbol{P}$, the relation in the form Equation (12.71) — so that moving it into Chapter 12 would change nothing but its address. Two by-products are worth naming in advance: the argument produces the vacuum vector explicitly, and it proves that the Schrödinger system is irreducible, a fact that Example 12.110 currently takes from [1481] without proof.

A.49.1 Statement, conventions and units

Throughout, $\mathcal{H}$ is a separable complex Hilbert space (Definition 12.2), f is a finite positive integer, and (U, V) is a Weyl system of f degrees of freedom in the sense of Definition 12.109: two families of strongly continuous unitary groups

$$U(\boldsymbol{\alpha}) = \exp\!\left(\frac{\mathrm{i}\,\boldsymbol{\alpha}\cdot\boldsymbol{Q}}{\hbar}\right), \qquad V(\boldsymbol{\beta}) = \exp\!\left(-\frac{\mathrm{i}\,\boldsymbol{\beta}\cdot\boldsymbol{P}}{\hbar}\right), \qquad \boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^f, \tag{A.867}$$

whose members belonging to different degrees of freedom commute, and which satisfy

$$U(\boldsymbol{\alpha})\,V(\boldsymbol{\beta}) = \mathrm{e}^{\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}/\hbar}\,V(\boldsymbol{\beta})\,U(\boldsymbol{\alpha}). \tag{A.868}$$

Remark A.578 (Units, and the dictionary to the form used in Part III). The components of $\boldsymbol{Q}$ carry m and those of $\boldsymbol{P}$ carry kg m/s, so the parameter $\boldsymbol{\alpha}$ carries kg m/s and $\boldsymbol{\beta}$ carries m; both $\boldsymbol{\alpha}\cdot\boldsymbol{Q}/\hbar$ and $\boldsymbol{\alpha}\cdot\boldsymbol{\beta}/\hbar$ are then pure numbers, since $\hbar$ carries J s. Every exponent written below is dimensionless as it stands, and $\hbar$ is never set to unity. Equation (25.39) of Chapter 25 writes the same relation with parameters of reciprocal dimension, a in 1/m and b in s/(kg m); the substitution $a = \alpha/\hbar$, $b = -\beta/\hbar$ turns Equation (A.868) into Equation (25.39) and back, and no statement below depends on which of the two is used. Finally, a fixed length $s > 0$, in m, enters in Appendix A.49.4 as the width of a Gaussian weight; it is arbitrary, and Remark A.588 records what changes when it is changed.

Theorem A.579 (Stone–von Neumann)**.** *Let (U, V) be a Weyl system of $f < \infty$ degrees of freedom on a separable Hilbert space $\mathcal{H} \neq \{0\}$.*

1. *If the system acts irreducibly (Definition 12.90) there is a unitary $T : \mathcal{H} \longrightarrow L^2(\mathbb{R}^f)$ carrying it to the Schrödinger system Equation (12.72), and T is unique up to a factor of modulus one.*

2. *Without the irreducibility hypothesis, $\mathcal{H}$ is an orthogonal direct sum (Definition 12.84) of at most countably many closed subspaces, each invariant under the system and each carrying a copy of the Schrödinger system.*

Rests on Definitions 12.90 and 12.109 and Example 12.110.

The proof occupies the rest of this section: Appendix A.49.2 assembles the Weyl operators and their composition law, Appendix A.49.3 the two analytic tools, Appendix A.49.4 the Gaussian average and the single identity on which everything turns, Appendix A.49.5 the vacuum vector, Appendix A.49.6 the unitary, and Appendix A.49.7 the Schrödinger system itself.

A.49.2 The Weyl operators and their composition law

Definition A.580 (Weyl operator)**.** For $z = (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{R}^{2f}$ put

$$W(z) = \mathrm{e}^{-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}/2\hbar}\, U(\boldsymbol{\alpha})\, V(\boldsymbol{\beta})\,, \tag{A.869}$$

and let

$$\sigma(z, z') = \boldsymbol{\alpha}\cdot\boldsymbol{\beta}' - \boldsymbol{\alpha}'\cdot\boldsymbol{\beta}\,, \qquad z = (\boldsymbol{\alpha}, \boldsymbol{\beta})\,, \quad z' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')\,, \tag{A.870}$$

be the standard symplectic form on $\mathbb{R}^{2f}$, of dimension J s.

Rests on Definitions 12.64 and 12.109 and Equation (12.71).

Lemma A.581 (Composition law)**.** *Every $W(z)$ is unitary, $W(0) = \mathbb{1}$, and*

$$W(z)\, W(z') = \mathrm{e}^{\mathrm{i}\sigma(z,z')/2\hbar}\, W(z + z')\,, \qquad W(z)^{\dagger} = W(-z) = W(z)^{-1}\,. \tag{A.871}$$

Consequently $W(z)W(z') = \mathrm{e}^{\mathrm{i}\sigma(z,z')/\hbar} W(z')W(z)$, and the family $\{W(z)\}_{z\in\mathbb{R}^{2f}}$ is self-adjoint in the sense of Definition 12.90. Moreover $W(\boldsymbol{\alpha}, \mathbf{0}) = U(\boldsymbol{\alpha})$ and $W(\mathbf{0}, \boldsymbol{\beta}) = V(\boldsymbol{\beta})$, so a closed subspace is invariant under every $W(z)$ if and only if it is invariant under every $U(\boldsymbol{\alpha})$ and every $V(\boldsymbol{\beta})$.

Rests on Definitions 12.64 and A.580 and Equation (12.71).

Proof. *Derives Lemma A.581.*

Unitarity and $W(0) = \mathbb{1}$ are immediate from Equation (A.869), a phase times a product of unitaries. Rearranging Equation (A.868) gives the form in which the relation will be used,

$$V(\boldsymbol{\beta})\, U(\boldsymbol{\alpha}') = \mathrm{e}^{-\mathrm{i}\boldsymbol{\alpha}'\cdot\boldsymbol{\beta}/\hbar}\, U(\boldsymbol{\alpha}')\, V(\boldsymbol{\beta})\,. \tag{A.872}$$

Now compute, using Equation (A.872) once to move $U(\boldsymbol{\alpha}')$ to the left of $V(\boldsymbol{\beta})$, and then the group laws of U and of V separately:

$$\begin{aligned} W(z)W(z') &= \mathrm{e}^{-\mathrm{i}(\boldsymbol{\alpha}\cdot\boldsymbol{\beta} + \boldsymbol{\alpha}'\cdot\boldsymbol{\beta}')/2\hbar}\, U(\boldsymbol{\alpha})V(\boldsymbol{\beta})U(\boldsymbol{\alpha}')V(\boldsymbol{\beta}') \\ &= \mathrm{e}^{-\mathrm{i}(\boldsymbol{\alpha}\cdot\boldsymbol{\beta} + \boldsymbol{\alpha}'\cdot\boldsymbol{\beta}')/2\hbar}\, \mathrm{e}^{-\mathrm{i}\boldsymbol{\alpha}'\cdot\boldsymbol{\beta}/\hbar}\, U(\boldsymbol{\alpha} + \boldsymbol{\alpha}')\, V(\boldsymbol{\beta} + \boldsymbol{\beta}')\,. \end{aligned} \tag{A.873}$$

On the other hand $W(z + z') = \mathrm{e}^{-\mathrm{i}(\boldsymbol{\alpha}+\boldsymbol{\alpha}')\cdot(\boldsymbol{\beta}+\boldsymbol{\beta}')/2\hbar} U(\boldsymbol{\alpha} + \boldsymbol{\alpha}')V(\boldsymbol{\beta} + \boldsymbol{\beta}')$, so the ratio of Equation (A.873) to $W(z + z')$ is a phase whose exponent, multiplied by $\hbar/\mathrm{i}$, is

$$\begin{aligned} &-\frac{\boldsymbol{\alpha}\cdot\boldsymbol{\beta} + \boldsymbol{\alpha}'\cdot\boldsymbol{\beta}'}{2} - \boldsymbol{\alpha}'\cdot\boldsymbol{\beta} + \frac{(\boldsymbol{\alpha} + \boldsymbol{\alpha}')\cdot(\boldsymbol{\beta} + \boldsymbol{\beta}')}{2} \\ &\quad = \frac{\boldsymbol{\alpha}\cdot\boldsymbol{\beta}' - \boldsymbol{\alpha}'\cdot\boldsymbol{\beta}}{2} = \frac{\sigma(z, z')}{2}\,, \end{aligned} \tag{A.874}$$

every term in $\boldsymbol{\alpha}\cdot\boldsymbol{\beta}$ and in $\boldsymbol{\alpha}'\cdot\boldsymbol{\beta}'$ cancelling. That is Equation (A.871). Since $\sigma(z, -z) = 0$ we get $W(z)W(-z) = W(0) = \mathbb{1}$, so $W(z)^{-1} = W(-z)$, and unitarity turns the inverse into the adjoint. Interchanging z and z' in Equation (A.871) and using $\sigma(z', z) = -\sigma(z, z')$ gives the commutation form. Setting $\boldsymbol{\beta} = \mathbf{0}$ or $\boldsymbol{\alpha} = \mathbf{0}$ in Equation (A.869) recovers U and V; conversely $W(z)$ is a phase times a product of one U and one V, so the two families have the same invariant subspaces. □

It is convenient to remove the units once and for all. Fix a length $s > 0$ and put

$$\zeta = (\boldsymbol{a}, \boldsymbol{b}) = \left(\frac{s\boldsymbol{\alpha}}{\hbar}, \; \frac{\boldsymbol{\beta}}{s} \right) \in \mathbb{R}^{2f} , \qquad \sigma(z, z') = \hbar \, \sigma_0(\zeta, \zeta') , \tag{A.875}$$

where $\sigma_0(\zeta, \zeta') = \boldsymbol{a} \cdot \boldsymbol{b}' - \boldsymbol{a}' \cdot \boldsymbol{b}$ is the same form built from the dimensionless coordinates, and $\mathrm{d}^f\alpha \, \mathrm{d}^f\beta = \hbar^f \, \mathrm{d}^{2f}\zeta$. Writing $W(\zeta)$ for $W(z)$ under this substitution, Equation (A.871) becomes

$$W(\zeta) \, W(\zeta') = \mathrm{e}^{\mathrm{i}\sigma_0(\zeta,\zeta')/2} \, W(\zeta + \zeta') . \tag{A.876}$$

It is also useful to write $\sigma_0(\zeta, \zeta') = \zeta \cdot J\zeta'$, where $J(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{b}, -\boldsymbol{a})$ is a real orthogonal map of $\mathbb{R}^{2f}$ with $J^2 = -\mathbb{1}$; then $\zeta \cdot J\zeta = 0$ and $|J\zeta| = |\zeta|$ for every ζ.

Lemma A.582 (Joint strong continuity)**.** *For every $x \in \mathcal{H}$ the map $\zeta \longmapsto W(\zeta)x$ is continuous from $\mathbb{R}^{2f}$ to $\mathcal{H}$, and $\|W(\zeta)x\| = \|x\|$.*

Rests on Lemma A.581 and Definition 12.64.

Proof. *Derives Lemma A.582.*

The norm statement is unitarity. For continuity it is enough, by Equation (A.869) and continuity of the phase, to show that $\boldsymbol{\alpha} \mapsto U(\boldsymbol{\alpha})x$ and $\boldsymbol{\beta} \mapsto V(\boldsymbol{\beta})x$ are continuous. Take U; the argument for V is identical. By hypothesis the f one-parameter groups $\alpha_j \mapsto U(\alpha_j \boldsymbol{e}_j)$ commute and each is strongly continuous, and $U(\boldsymbol{\alpha}) = U(\alpha_1 \boldsymbol{e}_1) \cdots U(\alpha_f \boldsymbol{e}_f)$. Telescoping the difference of two such products and using that every factor is an isometry,

$$\big\| U(\boldsymbol{\alpha})x - U(\boldsymbol{\alpha}')x \big\| \le \sum_{j=1}^{f} \Big\| U(\alpha_j \boldsymbol{e}_j) y_j - U(\alpha'_j \boldsymbol{e}_j) y_j \Big\| , \tag{A.877}$$

with $y_j = U(\alpha'_{j+1}\boldsymbol{e}_{j+1}) \cdots U(\alpha'_f \boldsymbol{e}_f)x$ a fixed vector for each j once $\boldsymbol{\alpha}'$ is fixed. Each summand tends to 0 as $\alpha_j \to \alpha'_j$ by strong continuity of the j-th group, which is continuity at $\boldsymbol{\alpha}'$. □

A.49.3 Two analytic tools

Lemma A.583 (Absolutely convergent operator-valued integrals)**.** *Let $G : \mathbb{R}^{2f} \longrightarrow \mathbb{C}$ be continuous with $\int_{\mathbb{R}^{2f}} |G| < \infty$. Then for every $x \in \mathcal{H}$ the $\mathcal{H}$-valued integral*

$$\Pi_G \, x = \int_{\mathbb{R}^{2f}} G(\zeta) \, W(\zeta) x \, \mathrm{d}^{2f}\zeta \tag{A.878}$$

exists as the limit of the integrals over the balls $|\zeta| \le R$; Π_G is a bounded operator with $\|\Pi_G\| \le \int |G|$; and

$$\langle w | \Pi_G x \rangle = \int_{\mathbb{R}^{2f}} G(\zeta) \, \langle w | W(\zeta) x \rangle \, \mathrm{d}^{2f}\zeta , \qquad (\Pi_G)^\dagger = \Pi_{G^\sharp} , \quad G^\sharp(\zeta) = \overline{G(-\zeta)} . \tag{A.879}$$

If G and G' both satisfy the hypotheses then

$$\Pi_G \, \Pi_{G'} = \Pi_{G''} , \qquad G''(u) = \int_{\mathbb{R}^{2f}} G(\zeta) \, G'(u - \zeta) \, \mathrm{e}^{\mathrm{i}\sigma_0(\zeta,u)/2} \, \mathrm{d}^{2f}\zeta . \tag{A.880}$$

Finally, a closed subspace invariant under every $W(\zeta)$ is invariant under Π_G.

Rests on Lemmas A.254 and A.582 and Proposition 12.4.

Proof. *Derives Lemma A.583.*

On each ball the integrand is a continuous $\mathcal{H}$-valued function of compact support, so the Riemann construction of Lemma A.254 applies verbatim in $2f$ variables — it uses only uniform continuity on a compact set and completeness of $\mathcal{H}$, neither of which cares how many variables there are. For $R' > R$ the difference of the two truncated integrals

is bounded in norm by $\|x\| \int_{R<|\zeta|\le R'} |G|$, which tends to zero as $R \to \infty$ because $|G|$ is integrable; the truncations therefore form a Cauchy net and converge. The bound on $\|\Pi_G\|$ and the first identity in Equation (A.879) pass to the limit from Equation (A.471). For the adjoint, use that identity twice together with $W(\zeta)^\dagger = W(-\zeta)$: $\langle \Pi_G w | x\rangle = \overline{\langle x | \Pi_G w\rangle} = \int \overline{G(\zeta)}\, \langle W(\zeta) w | x\rangle \, \mathrm{d}^{2f}\zeta = \int \overline{G(\zeta)}\, \langle w | W(-\zeta) x\rangle \, \mathrm{d}^{2f}\zeta$, and substituting $\zeta \to -\zeta$ identifies the result as $\langle w | \Pi_{G^\sharp} x\rangle$. Invariance of a closed invariant subspace is inherited from the truncated integrals, whose Riemann sums lie in it, the subspace being closed.

For Equation (A.880), take $w, x \in \mathcal{H}$ and expand $\langle w | \Pi_G \Pi_{G'} x\rangle$ by the first identity of Equation (A.879), applied once to Π_G acting on the vector $\Pi_{G'} x$ and once inside. The result is the scalar double integral

$$\iint G(\zeta) G'(\zeta')\, \langle w | W(\zeta) W(\zeta') x\rangle \, \mathrm{d}^{2f}\zeta \, \mathrm{d}^{2f}\zeta' \,, \tag{A.881}$$

absolutely convergent because the matrix element is bounded by $\|w\|\,\|x\|$ and G, G' are integrable. Substituting Equation (A.876) and changing variables to $u = \zeta + \zeta'$ at fixed ζ — a translation, which preserves $\mathrm{d}^{2f}\zeta'$ — and then exchanging the order of integration, which absolute convergence and Equation (7.112) on an exhausting sequence of boxes permit, turns Equation (A.881) into $\int G''(u)\, \langle w | W(u) x\rangle \, \mathrm{d}^{2f}u$, using $\sigma_0(\zeta, u-\zeta) = \sigma_0(\zeta, u)$. Since w was arbitrary this is $\langle w | \Pi_{G''} x\rangle$, and G'' is continuous and integrable because $|G''| \le |G| * |G'|$. □

The second tool is the injectivity of the Fourier transform in several variables. Corollary 17.33 states it on the line; the proof of Theorem 17.32 carries over unchanged, and it is worth writing out, because Appendix A.50 needs the same statement.

Lemma A.584 (Fourier injectivity in n variables)**.** *Let $h : \mathbb{R}^n \longrightarrow \mathbb{C}$ be continuous with $\int_{\mathbb{R}^n} |h| < \infty$, and suppose $\int_{\mathbb{R}^n} h(\xi)\, \mathrm{e}^{\mathrm{i}\eta\cdot\xi}\, \mathrm{d}^n\xi = 0$ for every $\eta \in \mathbb{R}^n$. Then $h \equiv 0$.*
Rests on Theorem 17.32, Equation (17.42), and Definition 17.29.

Proof. *Derives Lemma A.584.*

Write $\hat{h}(\eta) = (2\pi)^{-n/2} \int h(\xi) \mathrm{e}^{-\mathrm{i}\eta\cdot\xi} \mathrm{d}^n\xi$, which vanishes identically by hypothesis, with η replaced by $-\eta$. For $\varepsilon > 0$ put

$$I_\varepsilon(\xi) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} \hat{h}(\eta)\, \mathrm{e}^{-\varepsilon|\eta|^2/2}\, \mathrm{e}^{\mathrm{i}\eta\cdot\xi}\, \mathrm{d}^n\eta = 0 \,. \tag{A.882}$$

Inserting the definition of $\hat{h}$ and exchanging the order of integration — legitimate because $|h(\xi')|\, \mathrm{e}^{-\varepsilon|\eta|^2/2}$ is integrable over $\mathbb{R}^n \times \mathbb{R}^n$, so Equation (7.112) applies on an exhausting sequence of boxes — gives $I_\varepsilon(\xi) = \int h(\xi') g_\varepsilon(\xi - \xi') \mathrm{d}^n\xi'$ with

$$g_\varepsilon(u) = (2\pi\varepsilon)^{-n/2}\, \mathrm{e}^{-|u|^2/2\varepsilon} \,, \qquad \int_{\mathbb{R}^n} g_\varepsilon = 1 \,, \qquad g_\varepsilon \ge 0 \,, \tag{A.883}$$

the inner η-integral factorising into n one-dimensional Gaussian integrals, each evaluated by Equation (17.42). The family g_ε is an approximate identity exactly as in the proof of Theorem 17.32: non-negative, of unit mass, and concentrating on $|u| < \delta$ as $\varepsilon \to 0$. Since h is continuous at the fixed point ξ, bounded near it and integrable at infinity, $I_\varepsilon(\xi) \to h(\xi)$. But $I_\varepsilon \equiv 0$, so $h(\xi) = 0$. □

A.49.4 The Gaussian average

Definition A.585 (The Gaussian average of the Weyl operators)**.** With $s > 0$ fixed and ζ as in Equation (A.875), put

$$\Pi = (2\pi)^{-f} \int_{\mathbb{R}^{2f}} \mathrm{e}^{-|\zeta|^2/4}\, W(\zeta)\, \mathrm{d}^{2f}\zeta = \frac{1}{(2\pi\hbar)^f} \int_{\mathbb{R}^{2f}} \mathrm{e}^{-\frac{1}{4}\left(\frac{s^2|\boldsymbol{\alpha}|^2}{\hbar^2} + \frac{|\boldsymbol{\beta}|^2}{s^2}\right)} W(\boldsymbol{\alpha}, \boldsymbol{\beta})\, \mathrm{d}^f\alpha\, \mathrm{d}^f\beta \,, \tag{A.884}$$

a bounded operator by Lemma A.583. The two expressions agree by Equation (A.875), and the second exhibits the weight as a Gaussian of width $\hbar/s$ in momentum and s in position.
Rests on Lemma A.583 and Definition A.580.

Everything rests on one identity, which is a Gaussian integral and nothing else.

Lemma A.586 (Gaussian integral with a complex linear term)**.** *For every $n \geq 1$ and every $v \in \mathbb{C}^n$,*

$$\int_{\mathbb{R}^n} \mathrm{e}^{-|\zeta|^2/2+\zeta\cdot v}\,\mathrm{d}^n\zeta = (2\pi)^{n/2}\,\mathrm{e}^{v\cdot v/2}\,, \tag{A.885}$$

where $\zeta\cdot v=\sum_k \zeta_k v_k$ and $v\cdot v=\sum_k v_k^2$ are bilinear, not Hermitian.
Rests on Equations (7.112), (17.35) and (17.42).

Proof. *Derives Lemma A.586.*

Split $v = p+\mathrm{i}q$ with $p,q\in\mathbb{R}^n$ and complete the square in the real part, $-|\zeta|^2/2+\zeta\cdot p = |p|^2/2 - |\zeta-p|^2/2$. Substituting $u=\zeta-p$,

$$\int_{\mathbb{R}^n} \mathrm{e}^{-|\zeta|^2/2+\zeta\cdot v}\mathrm{d}^n\zeta = \mathrm{e}^{|p|^2/2}\,\mathrm{e}^{\mathrm{i}p\cdot q}\int_{\mathbb{R}^n}\mathrm{e}^{-|u|^2/2}\,\mathrm{e}^{\mathrm{i}u\cdot q}\,\mathrm{d}^n u\,. \tag{A.886}$$

The remaining integral factorises into n one-dimensional integrals by Equation (7.112), and each is Equation (17.42) with $a=1$, giving $(2\pi)^{n/2}\mathrm{e}^{-|q|^2/2}$. Collecting the three exponents, $|p|^2/2+\mathrm{i}p\cdot q-|q|^2/2 = v\cdot v/2$, which is Equation (A.885). □

Theorem A.587 (The Gaussian average is a non-zero projector that absorbs the Weyl operators)**.** *Π is self-adjoint, $\Pi^2=\Pi$, and for every $\eta\in\mathbb{R}^{2f}$*

$$\Pi\,W(\eta)\,\Pi = \mathrm{e}^{-|\eta|^2/4}\,\Pi\,. \tag{A.887}$$

Moreover $\Pi\neq 0$ whenever $\mathcal{H}\neq\{0\}$.
Rests on Definition A.585 and Lemmas A.584 and A.586.

Proof. *Derives Theorem A.587.*

Self-adjointness. The weight $G(\zeta)=(2\pi)^{-f}\mathrm{e}^{-|\zeta|^2/4}$ is real and even, so $G^\sharp = G$ and the second identity of Equation (A.879) gives $\Pi^\dagger=\Pi$.

Idempotence. By Equation (A.880) it suffices to show that

$$\int_{\mathbb{R}^{2f}} G(\zeta)\,G(u-\zeta)\,\mathrm{e}^{\mathrm{i}\sigma_0(\zeta,u)/2}\,\mathrm{d}^{2f}\zeta = G(u)\,. \tag{A.888}$$

Expand the product of the two Gaussians,

$$-\frac{|\zeta|^2}{4}-\frac{|u-\zeta|^2}{4} = -\frac{|\zeta|^2}{2}+\frac{\zeta\cdot u}{2}-\frac{|u|^2}{4}\,, \tag{A.889}$$

and write the phase as $\mathrm{i}\sigma_0(\zeta,u)/2 = \mathrm{i}\,\zeta\cdot Ju/2$. The integral in Equation (A.888) is therefore $(2\pi)^{-2f}\mathrm{e}^{-|u|^2/4}$ times the integral Equation (A.885) in $n=2f$ variables with

$$v=\frac{u+\mathrm{i}Ju}{2}\,,\qquad v\cdot v=\frac{1}{4}\left(u\cdot u+2\mathrm{i}\,u\cdot Ju-Ju\cdot Ju\right)=0\,, \tag{A.890}$$

because $u\cdot Ju=\sigma_0(u,u)=0$ and $Ju\cdot Ju=|u|^2$ by orthogonality of J. Hence the integral equals $(2\pi)^{-2f}\mathrm{e}^{-|u|^2/4}(2\pi)^f = (2\pi)^{-f}\mathrm{e}^{-|u|^2/4} = G(u)$, which is Equation (A.888). The exponent $-|\zeta|^2/4$ and the coefficient $(2\pi)^{-f}$ are fixed by this requirement and by nothing else; that is where they come from.

The absorption identity. Expanding as in Equation (A.881), and with the same justification,

$$\Pi W(\eta)\Pi = \iint G(\zeta)G(\zeta')\,W(\zeta)W(\eta)W(\zeta')\,\mathrm{d}^{2f}\zeta\,\mathrm{d}^{2f}\zeta'\,, \tag{A.891}$$

and two applications of Equation (A.876) give

$$W(\zeta)W(\eta)W(\zeta') = \exp\Big\{\frac{\mathrm{i}}{2}\big[\sigma_0(\zeta,\eta)+\sigma_0(\zeta,\zeta')+\sigma_0(\eta,\zeta')\big]\Big\}\,W(\zeta+\zeta'+\eta)\,. \tag{A.892}$$

Substitute $u=\zeta+\zeta'$ at fixed ζ. Using $\sigma_0(\zeta,u-\zeta)=\sigma_0(\zeta,u)$ and $\sigma_0(\eta,u-\zeta)=\sigma_0(\eta,u)+\sigma_0(\zeta,\eta)$, the bracket becomes $2\sigma_0(\zeta,\eta)+\sigma_0(\zeta,u)+\sigma_0(\eta,u)$, so

$$\Pi W(\eta)\Pi = \int \mathrm{e}^{\mathrm{i}\sigma_0(\eta,u)/2}\,W(u+\eta)\left[\int G(\zeta)G(u-\zeta)\,\mathrm{e}^{\mathrm{i}\sigma_0(\zeta,\,u+2\eta)/2}\,\mathrm{d}^{2f}\zeta\right]\mathrm{d}^{2f}u\,. \tag{A.893}$$

The inner integral is Equation (A.885) again, with the same Gaussian Equation (A.889) but with $v=\big[u+\mathrm{i}J(u+2\eta)\big]/2=p+\mathrm{i}q$, where $p=u/2$ and $q=J(u/2+\eta)$. Now

$$v\cdot v = |p|^2-|q|^2+2\mathrm{i}\,p\cdot q = \frac{|u|^2}{4}-\left|\frac{u}{2}+\eta\right|^2+\mathrm{i}\,\sigma_0(u,\eta) = -u\cdot\eta-|\eta|^2+\mathrm{i}\,\sigma_0(u,\eta)\,, \tag{A.894}$$

using $|J\cdot|=|\cdot|$ and $p\cdot q=(u/2)\cdot J(u/2+\eta)=\sigma_0(u/2,\eta)$. Hence the inner integral equals $G(u)\,\mathrm{e}^{-u\cdot\eta/2-|\eta|^2/2}\,\mathrm{e}^{\mathrm{i}\sigma_0(u,\eta)/2}$, and its phase cancels the prefactor $\mathrm{e}^{\mathrm{i}\sigma_0(\eta,u)/2}$ of Equation (A.893) exactly, because $\sigma_0(\eta,u)+\sigma_0(u,\eta)=0$. What remains is a pure Gaussian in u,

$$-\frac{|u|^2}{4}-\frac{u\cdot\eta}{2}-\frac{|\eta|^2}{2} = -\frac{|u+\eta|^2}{4}-\frac{|\eta|^2}{4}\,, \tag{A.895}$$

so that, substituting $w=u+\eta$, $\Pi W(\eta)\Pi=\mathrm{e}^{-|\eta|^2/4}(2\pi)^{-f}\int\mathrm{e}^{-|w|^2/4}W(w)\mathrm{d}^{2f}w=\mathrm{e}^{-|\eta|^2/4}\Pi$. That is Equation (A.887), and at $\eta=0$ it reproduces $\Pi^2=\Pi$, a check on every sign above.

Non-vanishing. Suppose $\Pi=0$. Conjugating a single Weyl operator by another, Equation (A.876) twice gives $W(\eta)W(\zeta)W(\eta)^\dagger=\mathrm{e}^{\mathrm{i}\sigma_0(\eta,\zeta)}W(\zeta)$, hence

$$W(\eta)\,\Pi\,W(\eta)^\dagger = \int_{\mathbb{R}^{2f}} G(\zeta)\,\mathrm{e}^{\mathrm{i}\sigma_0(\eta,\zeta)}\,W(\zeta)\,\mathrm{d}^{2f}\zeta\,, \tag{A.896}$$

which vanishes for every η if Π does. Fix $x,y\in\mathcal{H}$ and set $h(\zeta)=G(\zeta)\,\langle y|W(\zeta)x\rangle$, continuous by Lemma A.582 and absolutely integrable because $|h|\le G\,\|x\|\,\|y\|$. Taking the matrix element of Equation (A.896) gives $\int h(\zeta)\mathrm{e}^{\mathrm{i}\eta\cdot J\zeta}\mathrm{d}^{2f}\zeta=0$ for every η; since $\eta\cdot J\zeta=-(J\eta)\cdot\zeta$ and J is a bijection of $\mathbb{R}^{2f}$, that is the hypothesis of Lemma A.584, so $h\equiv 0$. At $\zeta=0$, $h(0)=(2\pi)^{-f}\,\langle y|x\rangle$, so $\langle y|x\rangle=0$ for all x and y — impossible unless $\mathcal{H}=\{0\}$. □

Remark A.588 (The width is a choice, not an input). Nothing above fixes s, and no statement below depends on it: the dimensionless variable Equation (A.875) absorbs it completely, so Π and every identity satisfied by it are the same for every s. What does change with s is the *vector* Π projects onto, which in the Schrödinger picture is the Gaussian of width s, Equation (A.902). The theorem is a statement about a representation, so it cannot depend on s; the vacuum is a statement about a vector, so it must.

A.49.5 The vacuum vector

Proposition A.589 (Cyclic subspaces and the rank of the average)**.** *Let $\Omega\in\operatorname{ran}\Pi$ with $\|\Omega\|=1$, and let M_Ω be the closed linear span of $\big\{W(\zeta)\Omega:\zeta\in\mathbb{R}^{2f}\big\}$. Then*

1. *M_Ω reduces the Weyl system, and the restriction of the system to M_Ω is again a Weyl system whose own Gaussian average is the restriction of Π;*
2. *that restricted average has range exactly $\mathbb{C}\,\Omega$, and the restricted system is irreducible;*
3. *if the original system is irreducible then $M_\Omega=\mathcal{H}$ and $\operatorname{ran}\Pi=\mathbb{C}\,\Omega$.*

Rests on Theorem A.587, Definition 12.90, and Proposition 12.89.

Proof. *Derives Proposition A.589.*

(1) M_Ω is invariant under every $W(\eta)$, since $W(\eta)W(\zeta)\Omega$ is a multiple of $W(\eta+\zeta)\Omega$ by Equation (A.876), and the family is self-adjoint by Lemma A.581; so $M_\Omega^\perp$ is invariant too and M_Ω reduces the system in the sense of Proposition 12.89. The restricted groups are again strongly continuous unitary groups on M_Ω satisfying Equation (A.868), and the restricted average is given by the same formula Equation (A.884), so it is Π restricted to M_Ω, which by the last clause of Lemma A.583 maps M_Ω into itself.

(2) $\Omega = \Pi\Omega \in M_\Omega$ lies in the range of the restricted average. Let $u \in M_\Omega$ satisfy $\Pi u = u$ and $\langle \Omega | u \rangle = 0$. Then, for every η,

$$\langle u | W(\eta)\Omega \rangle = \langle \Pi u | W(\eta)\Pi\Omega \rangle = \langle u | \Pi W(\eta)\Pi\,\Omega \rangle = \mathrm{e}^{-|\eta|^2/4}\,\langle u|\Omega\rangle = 0\,, \tag{A.897}$$

by Equation (A.887) and $\Pi = \Pi^\dagger$. So u is orthogonal to every generator of M_Ω, hence to M_Ω, hence to itself: $u = 0$. The range is therefore the line $\mathbb{C}\Omega$. For irreducibility, let $N \subseteq M_\Omega$ be a closed subspace reducing the restricted system, and let N' be its orthogonal complement inside M_Ω. By the last clause of Lemma A.583 both N and N' are invariant under Π, and their images are orthogonal subspaces of the line $\mathbb{C}\Omega$, so at most one of them is non-zero; and at least one is, because $\Pi\Omega = \Omega \neq 0$. Say $\Pi N' = \{0\}$ and write $\Omega = m + m'$ with $m \in N$, $m' \in N'$; then $\Omega = \Pi\Omega = \Pi m \in N$, so N contains every $W(\zeta)\Omega$ and equals M_Ω. In the other case the same argument gives $N' = M_\Omega$, i.e. $N = \{0\}$.

(3) $M_\Omega \neq \{0\}$ contains Ω, so irreducibility of the original system forces $M_\Omega = \mathcal{H}$, and then (2) is the statement about Π itself. □

Lemma A.590 (The Gram matrix is universal)**.** *Let the system act irreducibly and let Ω be a unit vector spanning* $\operatorname{ran}\Pi$*. Then for all ζ, ζ'*

$$\big\langle W(\zeta)\Omega \big| W(\zeta')\Omega \big\rangle = \mathrm{e}^{-\mathrm{i}\sigma_0(\zeta,\zeta')/2}\,\mathrm{e}^{-|\zeta'-\zeta|^2/4}\,. \tag{A.898}$$

The right-hand side involves the Weyl algebra alone: it is the same number in every irreducible Weyl system of f degrees of freedom.

Rests on Proposition A.589, Theorem A.587, and Lemma A.581.

Proof. *Derives Lemma A.590.*

First, $\Pi\Omega = \Omega$ and Equation (A.887) give $\langle \Omega | W(\eta)\Omega \rangle = \langle \Omega | \Pi W(\eta)\Pi\Omega \rangle = \mathrm{e}^{-|\eta|^2/4}$. Then, by Equation (A.876) and $W(\zeta)^\dagger = W(-\zeta)$,

$$\big\langle W(\zeta)\Omega \big| W(\zeta')\Omega \big\rangle = \big\langle \Omega \big| W(-\zeta)W(\zeta')\Omega \big\rangle = \mathrm{e}^{\mathrm{i}\sigma_0(-\zeta,\zeta')/2}\,\big\langle \Omega \big| W(\zeta'-\zeta)\Omega \big\rangle\,, \tag{A.899}$$

which is Equation (A.898). □

A.49.6 The intertwining unitary

Proposition A.591 (Any two irreducible Weyl systems are equivalent)**.** *Let (U, V) on $\mathcal{H}$ and (U', V') on $\mathcal{H}'$ be irreducible Weyl systems of the same finite number f of degrees of freedom. Then there is a unitary $T : \mathcal{H} \longrightarrow \mathcal{H}'$ with $TU(\boldsymbol{\alpha}) = U'(\boldsymbol{\alpha})T$ and $TV(\boldsymbol{\beta}) = V'(\boldsymbol{\beta})T$ for all $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and T is unique up to a factor of modulus one.*

Rests on Lemma A.590, Proposition A.589, and Theorem 12.91.

Proof. *Derives Proposition A.591.*

Choose unit vectors $\Omega \in \operatorname{ran}\Pi$ and $\Omega' \in \operatorname{ran}\Pi'$, which exist and span those ranges by Theorem A.587 and Proposition A.589. On the dense subspace $D \subset \mathcal{H}$ of finite linear combinations $x = \sum_{k=1}^{N} c_k W(\zeta_k)\Omega$ — dense because $M_\Omega = \mathcal{H}$ — define

$$T\left(\sum_k c_k W(\zeta_k)\Omega\right) = \sum_k c_k W'(\zeta_k)\Omega'\,. \tag{A.900}$$

By Lemma A.590 applied in each space,

$$\left\| \sum_k c_k W'(\zeta_k)\Omega' \right\|^2 = \sum_{k,l} \overline{c_k} c_l \, \langle W'(\zeta_k)\Omega' | W'(\zeta_l)\Omega' \rangle = \sum_{k,l} \overline{c_k} c_l \, \langle W(\zeta_k)\Omega | W(\zeta_l)\Omega \rangle = \|x\|^2 \,, \tag{A.901}$$

the two Gram matrices being equal entry by entry. In particular the right-hand side of Equation (A.900) vanishes whenever x does, so T is well defined on D, and it is isometric. Its range contains every $W'(\zeta)\Omega'$ and is therefore dense in $\mathcal{H}'$, again by Proposition A.589. An isometry with dense domain and dense range extends by continuity to a unitary of $\mathcal{H}$ onto $\mathcal{H}'$. Intertwining holds on D by construction — $TW(\eta)W(\zeta_k)\Omega$ and $W'(\eta)TW(\zeta_k)\Omega$ are both $\mathrm{e}^{\mathrm{i}\sigma_0(\eta,\zeta_k)/2}W'(\eta+\zeta_k)\Omega'$ — and extends by continuity; by the last clause of Lemma A.581 intertwining the W's is the same as intertwining the U's and the V's.

For uniqueness, let T_1 and T_2 both intertwine. Then $S = T_2 T_1^{-1}$ is a unitary of $\mathcal{H}'$ commuting with every $W'(\zeta)$. The family $\{W'(\zeta)\}$ is self-adjoint and acts irreducibly, so Theorem 12.91 gives $S = c\,\mathbb{1}$ with $c \in \mathbb{C}$, and $|c| = 1$ because S is unitary. □

A.49.7 The Schrödinger system

It remains to exhibit one irreducible Weyl system and to identify its vacuum. Here the Gaussian average can be computed in closed form, and the computation is the shortest route to both facts at once.

Proposition A.592 (The vacuum of the Schrödinger system)**.** *On $\mathcal{H} = L^2(\mathbb{R}^f)$ take the Schrödinger system Equation (12.72), read componentwise for f degrees of freedom, and build Π from it by Equation (A.884). Then, with*

$$\Omega_s(\boldsymbol{x}) = \left(\pi s^2\right)^{-f/4} \exp\!\left(-\frac{|\boldsymbol{x}|^2}{2s^2}\right), \qquad \|\Omega_s\| = 1\,, \tag{A.902}$$

the average is the rank-one orthogonal projector onto that vector,

$$\Pi = |\Omega_s\rangle\langle\Omega_s| \,, \qquad \textit{that is} \qquad (\Pi\psi)(\boldsymbol{x}) = \Omega_s(\boldsymbol{x}) \int_{\mathbb{R}^f} \Omega_s(\boldsymbol{x}')\,\psi(\boldsymbol{x}')\,\mathrm{d}^f x' \,. \tag{A.903}$$

Consequently $\{W(\zeta)\Omega_s\}$ is total in $L^2(\mathbb{R}^f)$ and the Schrödinger system acts irreducibly.
Rests on Example 12.110, Theorem A.587, and Lemma A.586.

Proof. *Derives Proposition A.592.*

The kernel. By Equation (12.72) and Equation (A.869),

$$(W(\boldsymbol{\alpha},\boldsymbol{\beta})\psi)(\boldsymbol{x}) = \mathrm{e}^{-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}/2\hbar}\,\mathrm{e}^{\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{x}/\hbar}\,\psi(\boldsymbol{x}-\boldsymbol{\beta})\,. \tag{A.904}$$

In the dimensionless variables Equation (A.875), $\boldsymbol{\alpha}\cdot\boldsymbol{\beta}/2\hbar = \boldsymbol{a}\cdot\boldsymbol{b}/2$, $\boldsymbol{\alpha}\cdot\boldsymbol{x}/\hbar = \boldsymbol{a}\cdot\boldsymbol{x}/s$ and $\boldsymbol{\beta} = s\boldsymbol{b}$, so

$$(\Pi\psi)(\boldsymbol{x}) = (2\pi)^{-f} \iint \mathrm{e}^{-(|\boldsymbol{a}|^2+|\boldsymbol{b}|^2)/4}\,\mathrm{e}^{-\mathrm{i}\boldsymbol{a}\cdot\boldsymbol{b}/2}\,\mathrm{e}^{\mathrm{i}\boldsymbol{a}\cdot\boldsymbol{x}/s}\,\psi(\boldsymbol{x}-s\boldsymbol{b})\,\mathrm{d}^f a\,\mathrm{d}^f b\,. \tag{A.905}$$

Do the $\boldsymbol{a}$ integral first, at fixed $\boldsymbol{b}$. It is Equation (A.885) in $n = f$ variables, after the rescaling $\boldsymbol{a} = \sqrt{2}\,\boldsymbol{u}$, and gives

$$\int_{\mathbb{R}^f} \mathrm{e}^{-|\boldsymbol{a}|^2/4 + \mathrm{i}\boldsymbol{a}\cdot(\boldsymbol{x}/s - \boldsymbol{b}/2)}\,\mathrm{d}^f a = (4\pi)^{f/2} \exp\!\left(-\left|\frac{\boldsymbol{x}}{s} - \frac{\boldsymbol{b}}{2}\right|^2\right). \tag{A.906}$$

Substituting $\boldsymbol{x}' = \boldsymbol{x} - s\boldsymbol{b}$, so that $\boldsymbol{b} = (\boldsymbol{x}-\boldsymbol{x}')/s$ and $\mathrm{d}^f b = s^{-f}\mathrm{d}^f x'$, the two exponents combine:

$$-\frac{|\boldsymbol{x}-\boldsymbol{x}'|^2}{4s^2} - \frac{|\boldsymbol{x}+\boldsymbol{x}'|^2}{4s^2} = -\frac{|\boldsymbol{x}|^2+|\boldsymbol{x}'|^2}{2s^2}\,, \tag{A.907}$$

because $\boldsymbol{x}/s - \boldsymbol{b}/2 = (\boldsymbol{x} + \boldsymbol{x}')/2s$ and the parallelogram identity turns the sum of the two squares into twice the sum of the squares of $\boldsymbol{x}$ and $\boldsymbol{x}'$. Hence

$$(\Pi\psi)(\boldsymbol{x}) = \frac{(4\pi)^{f/2}}{(2\pi)^f\, s^f}\, \mathrm{e}^{-|\boldsymbol{x}|^2/2s^2} \int_{\mathbb{R}^f} \mathrm{e}^{-|\boldsymbol{x}'|^2/2s^2}\, \psi(\boldsymbol{x}')\, \mathrm{d}^f x' \,, \tag{A.908}$$

and the constant is $(4\pi)^{f/2}(2\pi)^{-f} s^{-f} = (\pi s^2)^{-f/2}$, which is exactly what turns the two bare Gaussians into the normalised Equation (A.902) twice over. That is Equation (A.903), and it is manifestly an orthogonal projector of rank one, in agreement with Theorem A.587.

Totality. Let K be the closed linear span of $\{W(\zeta)\Omega_s\}$; it reduces the system by Proposition A.589(1), so $K^\perp$ is invariant and carries a Weyl system of its own, whose Gaussian average is the restriction of Π to $K^\perp$. But that restriction is zero: for $x \in K^\perp$, Equation (A.903) gives $\Pi x = \langle \Omega_s | x \rangle\, \Omega_s = 0$, because $\Omega_s = W(0)\Omega_s \in K$. By the last clause of Theorem A.587 a Weyl system on a non-zero space has a non-zero average, so $K^\perp = \{0\}$ and $K = L^2(\mathbb{R}^f)$.

Irreducibility. Ω_s spans $\operatorname{ran}\Pi$ and $M_{\Omega_s} = K = L^2(\mathbb{R}^f)$, so Proposition A.589(2) applied to M_{Ω_s} says that the system is irreducible. □

Proof of Theorem A.579. *Derives Theorem A.579.*

Part (1) is Proposition A.591 applied to the given system and to the Schrödinger system, which is an irreducible Weyl system by Proposition A.592.

Part (2). Let $\mathcal{H} \neq \{0\}$ carry a Weyl system, not assumed irreducible. By Theorem A.587 the operator Π is a non-zero orthogonal projector; choose a unit vector Ω_1 in its range and put $M_1 = M_{\Omega_1}$. By Proposition A.589, M_1 reduces the system and the restriction to it is an irreducible Weyl system, hence carries a copy of the Schrödinger system by part (1). Now $M_1^\perp$ is invariant and carries a Weyl system of its own; if it is not $\{0\}$ its average is again non-zero, and the construction produces $M_2 \subseteq M_1^\perp$, and so on. The subspaces so produced are mutually orthogonal and each is infinite-dimensional — it is a copy of $L^2(\mathbb{R}^f)$ — so separability of $\mathcal{H}$ (Definition 12.2) allows at most countably many of them, and the process exhausts $\mathcal{H}$: the orthogonal complement of their closed sum is invariant and carries a vanishing average, hence is $\{0\}$. That is the direct-sum statement, the sum being the one of Definition 12.84. □

A.49.8 What is quoted here

Remark A.593 (What is quoted here). Nothing, beyond the standing declaration of Remark 12.1. Every step above is a Gaussian integral, an application of Equation (17.42) through Lemma A.586 and Lemma A.584, or an application of Definition 12.90 and Theorem 12.91 — all proved in Chapter 12 or in Chapter 17. Two points deserve naming rather than leaving implicit.

First, the passage between the exponentiated form and the unbounded generators $\boldsymbol{Q}$ and $\boldsymbol{P}$ is Theorem 12.66, proved in Appendix A.25; it is what makes Theorem A.579 a statement about the canonical commutation relations at all, and Remark 25.36 explains why the unexponentiated relation would not do. The vector-valued integral of Lemma A.583 is the one built there, in Lemma A.254, extended from a compact interval to $\mathbb{R}^{2f}$ by absolute convergence.

Second, the irreducibility of the Schrödinger system is here *proved*, in Proposition A.592, by computing the Gaussian average in closed form. Example 12.110 takes it instead from [1481], through the maximal-abelian property of L^∞ acting on L^2; that route is shorter but rests on a fact about operator algebras which this treatise does not build, so the argument given here is the one on which Theorem A.579 stands.

Remark A.594 (Attribution). The exponentiated form of the commutation relations is Weyl's, from *Quantenmechanik und Gruppentheorie* (Zeitschrift für Physik **46**, 1927, 1–46), and

the uniqueness theorem is von Neumann's, from *Die Eindeutigkeit der Schrödingerschen Operatoren* (Mathematische Annalen **104**, 1931, 570–578); Stone had announced the result, and the theorem carries both names. Neither of those two papers has an entry in this treatise's bibliography — and the existing key for Weyl in 1929 is a different work of his, on the electron and gravitation, which must not be used for it — so that part of the attribution is made in words, on the footing of Darboux's memoir in Remark A.81. Nothing above rests on it, the proof being carried out in full. Stone's own paper on one-parameter unitary groups [707] is in the bibliography and is cited where it is used, in Appendix A.25; the modern treatment against which every step here can be checked is [1481], theorem VIII.14.

Remark A.595. Appendix A.49 discharges the derivation owed at Theorem 25.34 of Chapter 25, and with it the debt recorded in Remark 12.115 against Theorem 12.114 of Chapter 12 — the two are the same statement, and Remark 25.35 says that the proof belongs to Part II. Where the result is cashed is worth repeating. In Chapter 25 it is hypothesis (3) of Theorem 25.38: the theorem is what makes "the" Schrödinger representation a legitimate phrase, so that the Groenewold–van Hove obstruction obstructs a *quantization rule* and not merely one choice of Hilbert space. Its failure for infinitely many degrees of freedom, which Remark 25.37 and Remark 12.116 both record, is visible in the proof at exactly one place: the Gaussian weight of Equation (A.884) is a product over the f degrees of freedom, and for $f = \infty$ that product is not a measure on any space on which the argument could be run. Nothing here can be repaired to cover that case, and the physics of inequivalent vacua is the reason it should not be.

A.50 Hudson's Theorem: the Pure States of Non-Negative Wigner Function

This appendix proves Theorem 25.55 of Chapter 25: a pure state has a Wigner function that is non-negative everywhere if and only if its wavefunction is Gaussian. Together with Corollary 25.54 it is what makes negativity of the Wigner function an exact witness of departure from the Gaussian family, and the chapter uses it in that role.

Remark A.596 (A correction to the debt the chapter records). Remark 25.56 states that the converse half of the theorem "reduces it to one imported theorem — the Hadamard factorization of an entire function of order at most two" and records the absence of any order-and-genus theory in Chapter 8 as a debt of Part II. That is the standard route, and it is more than is needed. The only consequence of Hadamard's theorem this proof uses is that an entire function of one variable whose *real part* is bounded above by $A + B\,|\lambda|^2$ is a polynomial of degree at most two, and that statement follows in a dozen lines from the Taylor expansion Theorem 8.20 and the coefficient formula Equation (8.13), both of which Chapter 8 proves. It is Lemma A.608 below. Consequently this section quotes nothing, Part II is not in arrears on this account, and Remark 25.56 overstates the debt — which is worth saying plainly, because an unnecessary debt on the books is as misleading as a hidden one.

The other two statements of Remark 25.56 stand unchanged and are repeated in Remark A.614: the theorem is about pure states only, and Hudson's 1974 paper has no entry in this bibliography, so the attribution is made without a source the reader can check — which is exactly why the proof is written out.

A.50.1 Statement and conventions

Throughout, the system is one particle in three-dimensional space, as in Definition 25.52, and $\psi \in L^2(\mathbb{R}^3)$ is normalised. Its Wigner function is Equation (25.53) with $\rho = \psi\psi^*$, that

is

$$W(\boldsymbol{x},\boldsymbol{p}) = \frac{1}{(2\pi\hbar)^3}\int_{\mathbb{R}^3} \psi\Big(\boldsymbol{x}+\frac{\boldsymbol{y}}{2}\Big)\,\psi^*\Big(\boldsymbol{x}-\frac{\boldsymbol{y}}{2}\Big)\,\mathrm{e}^{-\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{y}/\hbar}\,\mathrm{d}^3y\,, \tag{A.909}$$

carrying $1/(\mathrm{J}^3\,\mathrm{s}^3)$. For a complex symmetric 3×3 matrix A write $A_\mathrm{r} = \operatorname{Re} A$ and $A_\mathrm{i} = \operatorname{Im} A$, both real symmetric, and $A_\mathrm{r} > 0$ for positive definiteness. Bilinear, never Hermitian, products are meant throughout: $\boldsymbol{u}\cdot A\boldsymbol{v} = \sum_{jk} u_j A_{jk} v_k$ and $\boldsymbol{u}\cdot\boldsymbol{v} = \sum_j u_j v_j$, even for complex arguments.

Remark A.597 (Units, and what the constant in the exponent carries). In the Gaussian form $\psi(\boldsymbol{x}) = \exp(-\boldsymbol{x}\cdot A\boldsymbol{x} + \boldsymbol{b}\cdot\boldsymbol{x} + c)$ the exponent must be dimensionless, so A carries $1/\mathrm{m}^2$ and $\boldsymbol{b}$ carries $1/\mathrm{m}$; e^c is then whatever makes $|\psi|^2$ carry $1/\mathrm{m}^3$, so c is fixed only once a unit of length is chosen and it is cleaner to read the Gaussian family as $\psi = \mathcal{N}\exp(-\boldsymbol{x}\cdot A\boldsymbol{x} + \boldsymbol{b}\cdot\boldsymbol{x})$ with $\mathcal{N}$ the normalisation. Nothing below depends on which of the two readings is used, and $\hbar$ appears explicitly at every step where it appears at all: in the kernel of Equation (A.909), and in the auxiliary length s of Appendix A.50.4, which enters the combination $s^2\boldsymbol{p}/\hbar$ — a length, as it must be.

Theorem A.598 (Hudson). *Let $\psi \in L^2(\mathbb{R}^3)$ be normalised. Then $W \geq 0$ everywhere if and only if there are a complex symmetric matrix A with $A_\mathrm{r} > 0$, a vector $\boldsymbol{b}\in\mathbb{C}^3$ and a constant $c\in\mathbb{C}$ with*

$$\psi(\boldsymbol{x}) = \exp(-\boldsymbol{x}\cdot A\boldsymbol{x} + \boldsymbol{b}\cdot\boldsymbol{x} + c) \tag{A.910}$$

for almost every $\boldsymbol{x}$. When that holds, W is a strictly positive Gaussian on phase space.

Rests on Equations (25.53) and (25.56) and Theorem 8.20.

A.50.2 Gaussian integrals

Lemma A.599 (One variable, complex coefficient). *Let $a\in\mathbb{C}$ with $\operatorname{Re} a > 0$ and $w\in\mathbb{C}$. Then*

$$\int_{\mathbb{R}} \mathrm{e}^{-az^2+wz}\,\mathrm{d}z = \sqrt{\frac{\pi}{a}}\;\mathrm{e}^{w^2/4a}\,, \tag{A.911}$$

the square root being the principal branch, which is defined and holomorphic on the half-plane $\operatorname{Re} a > 0$.

Rests on Equation (17.42) and Theorems 7.77 and 8.6.

Proof. *Derives Lemma A.599.*

Both integrals converge absolutely, since $\left|\mathrm{e}^{-az^2+wz}\right| = \mathrm{e}^{-(\operatorname{Re} a)z^2+(\operatorname{Re} w)z}$.

The case $w = 0$. Put $I(a) = \int_{\mathbb{R}} \mathrm{e}^{-az^2}\mathrm{d}z$. Differentiating under the integral sign, which Theorem 7.77 licenses on any compact subset of the half-plane because the differentiated integrand is dominated there by $z^2\mathrm{e}^{-\varepsilon z^2}$ for some $\varepsilon > 0$, and checking the Cauchy–Riemann equations Equation (8.5) on the integrand, I is holomorphic with $I'(a) = -\int z^2\mathrm{e}^{-az^2}\mathrm{d}z$. Integrating that by parts, $\int z\cdot(z\mathrm{e}^{-az^2})\mathrm{d}z = \int z\cdot(-\frac{1}{2a})\frac{\mathrm{d}}{\mathrm{d}z}\mathrm{e}^{-az^2}\mathrm{d}z = \frac{1}{2a}\int \mathrm{e}^{-az^2}\mathrm{d}z$, the boundary terms vanishing. Hence

$$I'(a) = -\frac{I(a)}{2a}\,, \qquad \text{so} \qquad \frac{\mathrm{d}}{\mathrm{d}a}\Big(\sqrt{a}\,I(a)\Big) = 0 \tag{A.912}$$

on the half-plane, which is connected; and $\sqrt{1}\,I(1) = \sqrt{\pi}$ by the Gauss integral, itself the case $\omega = 0$ of Equation (17.42). So $I(a) = \sqrt{\pi/a}$.

General w. Put $J(w) = \int_{\mathbb{R}} \mathrm{e}^{-az^2+wz}\mathrm{d}z$ at fixed a. The same argument makes J entire with $J'(w) = \int z\,\mathrm{e}^{-az^2+wz}\mathrm{d}z$, and the same integration by parts, now carrying the factor e^{wz}, gives $J'(w) = \frac{w}{2a}J(w)$. With $J(0) = \sqrt{\pi/a}$ this integrates to Equation (A.911). □

Lemma A.600 (The real part of an inverse)**.** *Let M and S be real symmetric $n\times n$ matrices with $M>0$. Then $M+\mathrm{i}S$ is invertible and*

$$\operatorname{Re}\left[(M+\mathrm{i}S)^{-1}\right]=\left(M+SM^{-1}S\right)^{-1}>0\,. \tag{A.913}$$

Rests on Theorems 5.38 and 5.80.

Proof. *Derives Lemma A.600.*

Look for the inverse in the form $X+\mathrm{i}Y$ with X,Y real. The two equations $MX-SY=\mathbb{1}$ and $MY+SX=0$ give $Y=-M^{-1}SX$ and then $\left(M+SM^{-1}S\right)X=\mathbb{1}$. The matrix $M+SM^{-1}S$ is real symmetric and positive definite — it is a sum of a positive definite matrix and the positive semi-definite $SM^{-1}S$, since $v\cdot SM^{-1}Sv=(Sv)\cdot M^{-1}(Sv)\geq 0$ and $M^{-1}>0$ — hence invertible, and $X=(M+SM^{-1}S)^{-1}$ solves the pair. That exhibits an inverse, which is unique, and its real part is X. □

Lemma A.601 (Several variables, complex symmetric matrix)**.** *Let N be a complex symmetric $n\times n$ matrix with $N_{\mathrm{r}}>0$, and let $\boldsymbol{w}\in\mathbb{C}^n$. Then N is invertible, $\operatorname{Re}(N^{-1})>0$, and*

$$\int_{\mathbb{R}^n}\mathrm{e}^{-\boldsymbol{x}\cdot N\boldsymbol{x}+\boldsymbol{w}\cdot\boldsymbol{x}}\,\mathrm{d}^n x=C_N\,\mathrm{e}^{\boldsymbol{w}\cdot N^{-1}\boldsymbol{w}/4}\,, \tag{A.914}$$

with $C_N\neq 0$ depending on N alone; if N is real then $C_N=\pi^{n/2}\,(\det N)^{-1/2}>0$.

Rests on Lemmas A.599 and A.600 and Theorem 5.80.

Proof. *Derives Lemma A.601.*

Invertibility and $\operatorname{Re}(N^{-1})>0$ are Lemma A.600 with $M=N_{\mathrm{r}}$ and $S=N_{\mathrm{i}}$. For the integral, let R be the real symmetric positive-definite square root of N_{r}, which exists by Theorem 5.80 — diagonalise N_{r} orthogonally and take the positive square roots of its eigenvalues — and substitute $\boldsymbol{x}=R^{-1}\boldsymbol{y}$. Then $\boldsymbol{x}\cdot N\boldsymbol{x}=\boldsymbol{y}\cdot(\mathbb{1}+\mathrm{i}T)\boldsymbol{y}$ with $T=R^{-1}N_{\mathrm{i}}R^{-1}$ real symmetric. Diagonalise T orthogonally, $T=O^{\mathsf{T}}DO$ with $D=\operatorname{diag}(d_1,\dots,d_n)$, and substitute $\boldsymbol{y}=O^{\mathsf{T}}\boldsymbol{z}$; the quadratic form becomes $\sum_j(1+\mathrm{i}d_j)z_j^2$ and the linear form becomes $\boldsymbol{w}'\cdot\boldsymbol{z}$ with $\boldsymbol{w}'=OR^{-1}\boldsymbol{w}$. The integral factorises into n copies of Equation (A.911) with $a_j=1+\mathrm{i}d_j$, each of which has $\operatorname{Re}a_j=1>0$:

$$\int_{\mathbb{R}^n}\mathrm{e}^{-\boldsymbol{x}\cdot N\boldsymbol{x}+\boldsymbol{w}\cdot\boldsymbol{x}}\mathrm{d}^n x=\frac{1}{\det R}\prod_{j=1}^{n}\sqrt{\frac{\pi}{a_j}}\,\exp\left(\sum_j\frac{\left(w_j'\right)^2}{4a_j}\right)\,, \tag{A.915}$$

the Jacobian being $\left|\det(R^{-1}O^{\mathsf{T}})\right|=1/\det R$. The exponent reassembles as $\boldsymbol{w}'\cdot(\mathbb{1}+\mathrm{i}D)^{-1}\boldsymbol{w}'/4=\boldsymbol{w}\cdot R^{-1}O^{\mathsf{T}}(\mathbb{1}+\mathrm{i}D)^{-1}OR^{-1}\boldsymbol{w}/4=\boldsymbol{w}\cdot N^{-1}\boldsymbol{w}/4$, because $N=R(\mathbb{1}+\mathrm{i}T)R$. The prefactor is a non-zero constant determined by N, and for real N every d_j is zero and it is $\pi^{n/2}(\det N)^{-1/2}$. □

A.50.3 The easy direction

Proposition A.602 (The Wigner function of a Gaussian is a positive Gaussian)**.** *Let ψ be of the form Equation (A.910) with $A_{\mathrm{r}}>0$. Then, with $\boldsymbol{w}=-2A_{\mathrm{i}}\boldsymbol{x}+\operatorname{Im}\boldsymbol{b}-\boldsymbol{p}/\hbar$,*

$$W(\boldsymbol{x},\boldsymbol{p})=\frac{(2\pi)^{3/2}}{(2\pi\hbar)^3\sqrt{\det A_{\mathrm{r}}}}\,\exp\left(-2\boldsymbol{x}\cdot A_{\mathrm{r}}\boldsymbol{x}+2\operatorname{Re}\boldsymbol{b}\cdot\boldsymbol{x}+2\operatorname{Re}c-\tfrac{1}{2}\boldsymbol{w}\cdot A_{\mathrm{r}}^{-1}\boldsymbol{w}\right)\,, \tag{A.916}$$

which is real, strictly positive at every point of phase space, and Gaussian in $(\boldsymbol{x},\boldsymbol{p})$ jointly.

Rests on Equations (25.53) and (A.910) and Lemma A.601.

Proof. *Derives Proposition A.602.*

Put $\boldsymbol{u} = \boldsymbol{x} + \boldsymbol{y}/2$ and $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{y}/2$, so that the integrand of Equation (A.909) is $\exp(-\boldsymbol{u}\cdot A\boldsymbol{u} + \boldsymbol{b}\cdot\boldsymbol{u} + c - \boldsymbol{v}\cdot\overline{A}\boldsymbol{v} + \overline{\boldsymbol{b}}\cdot\boldsymbol{v} + \overline{c})$ times $\mathrm{e}^{-\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{y}/\hbar}$. Expanding by symmetry of A,

$$\begin{aligned} -\boldsymbol{u}\cdot A\boldsymbol{u} - \boldsymbol{v}\cdot\overline{A}\boldsymbol{v} &= -\boldsymbol{x}\cdot\left(A+\overline{A}\right)\boldsymbol{x} - \boldsymbol{x}\cdot\left(A-\overline{A}\right)\boldsymbol{y} - \tfrac{1}{4}\boldsymbol{y}\cdot\left(A+\overline{A}\right)\boldsymbol{y} \\ &= -2\boldsymbol{x}\cdot A_{\mathrm{r}}\boldsymbol{x} - 2\mathrm{i}\,\boldsymbol{x}\cdot A_{\mathrm{i}}\boldsymbol{y} - \tfrac{1}{2}\boldsymbol{y}\cdot A_{\mathrm{r}}\boldsymbol{y}\,, \end{aligned} \tag{A.917}$$

while $\boldsymbol{b}\cdot\boldsymbol{u} + \overline{\boldsymbol{b}}\cdot\boldsymbol{v} = 2\,\mathrm{Re}\,\boldsymbol{b}\cdot\boldsymbol{x} + \mathrm{i}\,\mathrm{Im}\,\boldsymbol{b}\cdot\boldsymbol{y}$ and $c + \overline{c} = 2\,\mathrm{Re}\,c$. Collecting the terms carrying $\boldsymbol{y}$ and adding the Fourier kernel, the whole $\boldsymbol{y}$ dependence is $-\frac{1}{2}\boldsymbol{y}\cdot A_{\mathrm{r}}\boldsymbol{y} + \mathrm{i}\,\boldsymbol{y}\cdot\boldsymbol{w}$ with $\boldsymbol{w}$ as stated, a *real* vector. The remaining integral is Equation (A.914) with the real positive-definite $N = A_{\mathrm{r}}/2$ and $\boldsymbol{w} \to \mathrm{i}\boldsymbol{w}$, giving $(2\pi)^{3/2}(\det A_{\mathrm{r}})^{-1/2}\exp(-\frac{1}{2}\boldsymbol{w}\cdot A_{\mathrm{r}}^{-1}\boldsymbol{w})$. That is Equation (A.916). Every factor is real and positive: the exponential of a real number, and a positive constant. Since $\boldsymbol{w}$ is an affine function of $(\boldsymbol{x},\boldsymbol{p})$, the exponent is a real quadratic in $(\boldsymbol{x},\boldsymbol{p})$ and W is a Gaussian on phase space. □

A.50.4 Coherent states, and where the Wigner function cannot vanish

Fix once and for all a length $s > 0$ and put, for $\boldsymbol{x}_0, \boldsymbol{p}_0 \in \mathbb{R}^3$,

$$\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}(\boldsymbol{x}) = \left(\pi s^2\right)^{-3/4}\exp\!\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}_0|^2}{2s^2} + \frac{\mathrm{i}\,\boldsymbol{p}_0\cdot\boldsymbol{x}}{\hbar}\right), \tag{A.918}$$

a normalised element of $L^2(\mathbb{R}^3)$ of the form Equation (A.910), with $A = \mathbb{1}/2s^2$, $\boldsymbol{b} = \boldsymbol{x}_0/s^2 + \mathrm{i}\boldsymbol{p}_0/\hbar$ and the appropriate c. It is exactly the vector $W(\zeta)\Omega_s$ of Proposition A.592, up to a phase: by Equation (A.904), $\big(W(\boldsymbol{\alpha},\boldsymbol{\beta})\Omega_s\big)(\boldsymbol{x}) = \mathrm{e}^{-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}/2\hbar}\varphi_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\boldsymbol{x})$. That identification is used once, at the end of Appendix A.50.7.

Lemma A.603 (Overlaps and the resolution of unity)**.** *For any normalised ψ,*

$$|\langle\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}|\psi\rangle|^2 = (2\pi\hbar)^3\int W_{\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}}\,W_\psi\,\mathrm{d}^3x\,\mathrm{d}^3p\,, \tag{A.919}$$

and if $W_\psi \geq 0$ everywhere then

$$\int_{\mathbb{R}^3}\!\int_{\mathbb{R}^3}|\langle\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}|\psi\rangle|^2\;\mathrm{d}^3x_0\,\mathrm{d}^3p_0 = (2\pi\hbar)^3\,. \tag{A.920}$$

Rests on Equations (25.56) and (A.918) and Theorem 25.53.

Proof. *Derives Lemma A.603.*

For pure states $\hat\rho_1 = |\psi_1\rangle\langle\psi_1|$ and $\hat\rho_2 = |\psi_2\rangle\langle\psi_2|$ one has $\mathrm{tr}(\hat\rho_1\hat\rho_2) = \langle\psi_1|\psi_2\rangle\,\langle\psi_2|\psi_1\rangle = |\langle\psi_1|\psi_2\rangle|^2$, so Equation (25.56) is Equation (A.919).

For the second statement, note first that Equation (A.918) is $\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}(\boldsymbol{x}) = \mathrm{e}^{\mathrm{i}\boldsymbol{p}_0\cdot\boldsymbol{x}/\hbar}\,\varphi_{\boldsymbol{0},\boldsymbol{0}}(\boldsymbol{x}-\boldsymbol{x}_0)$, and that substituting this into Equation (A.909) makes the two exponential factors combine as $\mathrm{e}^{\mathrm{i}\boldsymbol{p}_0\cdot\boldsymbol{y}/\hbar}$ and shifts the argument, so that

$$W_{\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}}(\boldsymbol{x},\boldsymbol{p}) = W_{\varphi_{\boldsymbol{0},\boldsymbol{0}}}(\boldsymbol{x}-\boldsymbol{x}_0,\;\boldsymbol{p}-\boldsymbol{p}_0)\,. \tag{A.921}$$

By Proposition A.602 that function is a strictly positive Gaussian, and by Theorem 25.53 it integrates to 1 over phase space; hence $\iint W_{\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}}(\boldsymbol{x},\boldsymbol{p})\mathrm{d}^3x_0\mathrm{d}^3p_0 = 1$ for every $(\boldsymbol{x},\boldsymbol{p})$. Integrating Equation (A.919) over $(\boldsymbol{x}_0,\boldsymbol{p}_0)$ and exchanging the order of integration — legitimate because *every* integrand in sight is non-negative, which is where the hypothesis $W_\psi \geq 0$ is used — gives $(2\pi\hbar)^3\int W_\psi\mathrm{d}^3x\,\mathrm{d}^3p$, which is $(2\pi\hbar)^3$ by the normalisation clause of Theorem 25.53. □

Proposition A.604 (Non-negativity forbids a vanishing overlap)**.** *If $W_\psi \geq 0$ everywhere then $\langle \varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}|\psi\rangle \neq 0$ for every $\boldsymbol{x}_0$ and every $\boldsymbol{p}_0$.*

Rests on Lemma A.603, Proposition A.602, and Equation (25.54).

Proof. *Derives Proposition A.604.*

Suppose the overlap vanishes for some $(\boldsymbol{x}_0, \boldsymbol{p}_0)$. By Equation (A.919) the integral of $W_\varphi W_\psi$ over phase space is then zero. Both factors are non-negative — the first strictly positive everywhere by Proposition A.602, the second by hypothesis — so the product vanishes identically, and therefore W_ψ does. Integrating W_ψ over $\boldsymbol{p}$ and using Equation (25.54) gives $|\psi(\boldsymbol{x})|^2 = 0$ for every $\boldsymbol{x}$, contradicting $\|\psi\| = 1$. □

A.50.5 An entire function without zeros

Definition A.605 (The analytic continuation of the overlaps)**.** For $\zeta \in \mathbb{C}^3$ put

$$F(\zeta) = \int_{\mathbb{R}^3} \psi(\boldsymbol{x}) \, \exp\!\left(-\frac{(\boldsymbol{x}-\zeta)\cdot(\boldsymbol{x}-\zeta)}{2s^2}\right) \mathrm{d}^3x \,, \tag{A.922}$$

the product in the exponent being the bilinear one.

Rests on Equation (A.918) and Definition 25.52.

Lemma A.606 (F is entire, and is the overlap in disguise)**.** *The integral Equation (A.922) converges absolutely for every $\zeta \in \mathbb{C}^3$ and defines a function holomorphic in each variable separately and jointly continuous, hence entire on $\mathbb{C}^3$. Writing $\zeta = \boldsymbol{x}_0 - \mathrm{i}s^2\boldsymbol{p}_0/\hbar$ with $\boldsymbol{x}_0, \boldsymbol{p}_0 \in \mathbb{R}^3$,*

$$F(\zeta) = \kappa(\boldsymbol{x}_0,\boldsymbol{p}_0)\, \langle \varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}|\psi\rangle \,, \qquad \kappa = \left(\pi s^2\right)^{3/4} \mathrm{e}^{|\operatorname{Im}\zeta|^2/2s^2}\, \mathrm{e}^{\mathrm{i}\boldsymbol{x}_0\cdot\boldsymbol{p}_0/\hbar} \neq 0 \,, \tag{A.923}$$

and consequently

$$|F(\zeta)| \leq \left(\pi s^2\right)^{3/4} \exp\!\left(\frac{|\operatorname{Im}\zeta|^2}{2s^2}\right) . \tag{A.924}$$

Rests on Definition A.605, Theorem 8.6, and Proposition 12.4.

Proof. *Derives Lemma A.606.*

Write $\zeta = \boldsymbol{\xi} + \mathrm{i}\boldsymbol{\eta}$. Then $(\boldsymbol{x}-\zeta)\cdot(\boldsymbol{x}-\zeta) = |\boldsymbol{x}-\boldsymbol{\xi}|^2 - |\boldsymbol{\eta}|^2 - 2\mathrm{i}(\boldsymbol{x}-\boldsymbol{\xi})\cdot\boldsymbol{\eta}$, so the modulus of the kernel is $\mathrm{e}^{|\boldsymbol{\eta}|^2/2s^2}\mathrm{e}^{-|\boldsymbol{x}-\boldsymbol{\xi}|^2/2s^2}$, a Gaussian in $\boldsymbol{x}$; its product with $|\psi|$ is integrable by Cauchy–Schwarz (Proposition 12.4), and the bound is locally uniform in ζ. Differentiating under the integral sign in the real and imaginary parts of each ζ_k — licensed by Theorem 7.77 on bounded regions together with the locally uniform domination just displayed, which controls the tails — shows that F has continuous first partial derivatives and that they satisfy the Cauchy–Riemann equations Equation (8.5) in each variable, because the integrand does. Hence F is holomorphic in each variable, and being continuous it is entire on $\mathbb{C}^3$.

For Equation (A.923), substitute $\zeta = \boldsymbol{x}_0 - \mathrm{i}s^2\boldsymbol{p}_0/\hbar$, so that $\boldsymbol{x}-\zeta = (\boldsymbol{x}-\boldsymbol{x}_0) + \mathrm{i}s^2\boldsymbol{p}_0/\hbar$ and

$$-\frac{(\boldsymbol{x}-\zeta)\cdot(\boldsymbol{x}-\zeta)}{2s^2} = -\frac{|\boldsymbol{x}-\boldsymbol{x}_0|^2}{2s^2} - \frac{\mathrm{i}(\boldsymbol{x}-\boldsymbol{x}_0)\cdot\boldsymbol{p}_0}{\hbar} + \frac{s^2|\boldsymbol{p}_0|^2}{2\hbar^2} \,. \tag{A.925}$$

The first two terms are the exponent of $\overline{\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}}$ up to the constant $(\pi s^2)^{-3/4}$ and the phase $\mathrm{e}^{\mathrm{i}\boldsymbol{x}_0\cdot\boldsymbol{p}_0/\hbar}$, and the third is $|\operatorname{Im}\zeta|^2/2s^2$ because $\operatorname{Im}\zeta = -s^2\boldsymbol{p}_0/\hbar$. That is Equation (A.923). Finally $|\langle\varphi|\psi\rangle| \leq \|\varphi\|\,\|\psi\| = 1$ by Cauchy–Schwarz, which with $|\kappa|$ as displayed gives Equation (A.924); and every $\zeta \in \mathbb{C}^3$ arises from exactly one pair $(\boldsymbol{x}_0,\boldsymbol{p}_0)$, namely $\boldsymbol{x}_0 = \operatorname{Re}\zeta$ and $\boldsymbol{p}_0 = -\hbar\operatorname{Im}\zeta/s^2$. □

Lemma A.607 (A zero-free entire function is an exponential)**.** *Let F be entire on $\mathbb{C}^n$ with $F(\zeta) \neq 0$ for every ζ. Then there is an entire g on $\mathbb{C}^n$ with $F = \mathrm{e}^g$.*

Rests on Lemma A.606 and Theorems 7.77 and 8.6.

Proof. *Derives Lemma A.607.*

Choose $c_0 \in \mathbb{C}$ with $\mathrm{e}^{c_0} = F(0)$, possible because $F(0) \neq 0$, and set

$$g(\zeta) = c_0 + \int_0^1 \sum_{j=1}^{n} \zeta_j \, \frac{(\partial_j F)(t\zeta)}{F(t\zeta)} \, \mathrm{d}t \,. \tag{A.926}$$

The integrand is continuous in t and holomorphic in ζ, since F is entire and non-vanishing, so g is entire by the same combination of Theorem 7.77 and Equation (8.5) used in Lemma A.606. Fix ζ and substitute $u = vt$ in Equation (A.926) written at the point $t\zeta$; the result is $g(t\zeta) = c_0 + \int_0^t \sum_j \zeta_j (\partial_j F)(u\zeta)/F(u\zeta)\,\mathrm{d}u$, so

$$\frac{\mathrm{d}}{\mathrm{d}t}\, g(t\zeta) = \sum_j \zeta_j \frac{(\partial_j F)(t\zeta)}{F(t\zeta)} = \frac{1}{F(t\zeta)} \frac{\mathrm{d}}{\mathrm{d}t} F(t\zeta) \,, \tag{A.927}$$

the last step by the chain rule. Hence $\frac{\mathrm{d}}{\mathrm{d}t}\big[F(t\zeta)\mathrm{e}^{-g(t\zeta)}\big] = 0$, so $F(\zeta)\mathrm{e}^{-g(\zeta)} = F(0)\mathrm{e}^{-c_0} = 1$. □

Lemma A.608 (An entire function with a quadratic bound on its real part)**.** *Let u be entire on $\mathbb{C}$ and suppose there are real constants A and B with $\operatorname{Re} u(\lambda) \leq A + B\,|\lambda|^2$ for all λ. Then u is a polynomial of degree at most two.*

Rests on Theorems 8.18 and 8.20 and Equation (8.13).

Proof. *Derives Lemma A.608.*

By Theorem 8.20 the Taylor series $u(\lambda) = \sum_{k\geq 0} a_k \lambda^k$ converges on every disc. Fix $r > 0$. The series converges uniformly on the circle $|\lambda| = r$, so it may be integrated term by term against $\mathrm{e}^{-\mathrm{i}k\vartheta}$, and for $k \geq 0$

$$a_k r^k = \frac{1}{2\pi} \int_0^{2\pi} u\Big(r\mathrm{e}^{\mathrm{i}\vartheta}\Big)\, \mathrm{e}^{-\mathrm{i}k\vartheta}\, \mathrm{d}\vartheta \,, \qquad 0 = \frac{1}{2\pi} \int_0^{2\pi} u\Big(r\mathrm{e}^{\mathrm{i}\vartheta}\Big)\, \mathrm{e}^{+\mathrm{i}k\vartheta}\, \mathrm{d}\vartheta \quad (k \geq 1) \,, \tag{A.928}$$

the first being Equation (8.13) written on the circle and the second holding because every term of the series then carries a strictly positive power of $\mathrm{e}^{\mathrm{i}\vartheta}$. Conjugating the second identity and adding it to the first gives, for $k \geq 1$,

$$a_k r^k = \frac{1}{\pi} \int_0^{2\pi} \operatorname{Re} u\Big(r\mathrm{e}^{\mathrm{i}\vartheta}\Big)\, \mathrm{e}^{-\mathrm{i}k\vartheta}\, \mathrm{d}\vartheta \,. \tag{A.929}$$

Write $\mu(r) = A + Br^2$, so that $\operatorname{Re} u(r\mathrm{e}^{\mathrm{i}\vartheta}) \leq \mu(r)$ by hypothesis, and subtract the constant $\mu(r)$ inside Equation (A.929), which changes nothing because $\int_0^{2\pi} \mathrm{e}^{-\mathrm{i}k\vartheta}\mathrm{d}\vartheta = 0$ for $k \geq 1$. Bounding the modulus of the integral by the integral of the modulus, and using that $\mu(r) - \operatorname{Re} u$ is non-negative so that its modulus is itself,

$$|a_k|\, r^k \leq \frac{1}{\pi} \int_0^{2\pi} \Big(\mu(r) - \operatorname{Re} u\Big(r\mathrm{e}^{\mathrm{i}\vartheta}\Big)\Big)\, \mathrm{d}\vartheta = 2\,(\mu(r) - \operatorname{Re} a_0) \,, \tag{A.930}$$

the mean of $\operatorname{Re} u$ over the circle being $\operatorname{Re} a_0$ by the case $k = 0$ of Equation (A.928). Hence for every $k \geq 1$

$$|a_k| \leq \frac{2\,(A + Br^2 - \operatorname{Re} a_0)}{r^k} \,, \tag{A.931}$$

and for $k \geq 3$ the right-hand side tends to 0 as $r \to \infty$. So $a_k = 0$ for $k \geq 3$. □

Proposition A.609 (The exponent is a quadratic polynomial)**.** *If $W_\psi \geq 0$ everywhere then there are a complex symmetric 3×3 matrix Γ, a vector $\boldsymbol{\lambda} \in \mathbb{C}^3$ and a constant $\nu \in \mathbb{C}$ with*

$$F(\zeta) = \exp(-\zeta \cdot \Gamma\zeta + \boldsymbol{\lambda} \cdot \zeta + \nu) \,, \qquad \zeta \in \mathbb{C}^3 \,. \tag{A.932}$$

Rests on Proposition A.604 and Lemmas A.607 and A.608.

Proof. *Derives Proposition A.609.*

By Lemma A.606 and Proposition A.604 the entire function F has no zeros: $F(\zeta)$ is a non-zero multiple of an overlap that never vanishes. So $F = \mathrm{e}^{g}$ with g entire, by Lemma A.607, and Equation (A.924) gives

$$\operatorname{Re} g(\zeta) = \log|F(\zeta)| \le \tfrac{3}{4}\log\!\left(\pi s^2\right) + \frac{|\zeta|^2}{2s^2}\,, \tag{A.933}$$

since $|\operatorname{Im}\zeta| \le |\zeta|$.

Fix $\boldsymbol{\beta} \in \mathbb{C}^3$ and consider $u_{\boldsymbol{\beta}}(\lambda) = g(\lambda\boldsymbol{\beta})$, entire in the single variable λ because g is entire and $\lambda \mapsto \lambda\boldsymbol{\beta}$ is holomorphic. By Equation (A.933), $\operatorname{Re} u_{\boldsymbol{\beta}}(\lambda) \le \frac{3}{4}\log(\pi s^2) + |\boldsymbol{\beta}|^2\,|\lambda|^2/2s^2$, so Lemma A.608 makes $u_{\boldsymbol{\beta}}$ a polynomial of degree at most two:

$$g(\lambda\boldsymbol{\beta}) = g(0) + \lambda\,\frac{\mathrm{d}}{\mathrm{d}\lambda}\bigg|_0 g(\lambda\boldsymbol{\beta}) + \frac{\lambda^2}{2}\,\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\bigg|_0 g(\lambda\boldsymbol{\beta})\,, \tag{A.934}$$

exactly, with no remainder. The two derivatives are computed by the chain rule: $\mathrm{d}g(\lambda\boldsymbol{\beta})/\mathrm{d}\lambda|_0 = \sum_j \beta_j\,\partial_j g(0)$ and $\mathrm{d}^2 g(\lambda\boldsymbol{\beta})/\mathrm{d}\lambda^2|_0 = \sum_{j,k}\beta_j\beta_k\,\partial_j\partial_k g(0)$. Setting $\lambda = 1$ and renaming $\boldsymbol{\beta}$ as ζ, which is legitimate because $\boldsymbol{\beta}$ was arbitrary,

$$g(\zeta) = g(0) + \sum_j \zeta_j\,\partial_j g(0) + \frac{1}{2}\sum_{j,k}\zeta_j\zeta_k\,\partial_j\partial_k g(0)\,, \tag{A.935}$$

which is Equation (A.932) with $\Gamma_{jk} = -\frac{1}{2}\partial_j\partial_k g(0)$, symmetric because mixed holomorphic partial derivatives commute, $\lambda_j = \partial_j g(0)$ and $\nu = g(0)$. Note that no theorem about Taylor expansions in several variables was used: the expansion was performed along one complex line at a time, and Lemma A.608 guaranteed that on every such line it terminates. □

A.50.6 What normalisability forces

The quadratic form Γ is not arbitrary, and the constraint it obeys is exactly the one that will make the recovered wavefunction square integrable. Write $\Gamma_{\mathrm{r}} = \operatorname{Re}\Gamma$ and $\Gamma_{\mathrm{i}} = \operatorname{Im}\Gamma$.

Lemma A.610 (Block positivity)**.** *Let P_{11}, P_{12} and P_{22} be real $n \times n$ matrices with P_{11} and P_{22} symmetric. The real symmetric form $Q(\boldsymbol{\xi},\boldsymbol{\eta}) = \boldsymbol{\xi}\cdot P_{11}\boldsymbol{\xi} + 2\boldsymbol{\xi}\cdot P_{12}\boldsymbol{\eta} + \boldsymbol{\eta}\cdot P_{22}\boldsymbol{\eta}$ on $\mathbb{R}^{2n}$ is positive definite if and only if $P_{22} > 0$ and $P_{11} - P_{12}P_{22}^{-1}P_{12}^{\mathsf{T}} > 0$.*

Rests on Theorems 5.38 and 5.80.

Proof. *Derives Lemma A.610.*

If $P_{22} > 0$, complete the square: with $\boldsymbol{\eta}' = \boldsymbol{\eta} + P_{22}^{-1}P_{12}^{\mathsf{T}}\boldsymbol{\xi}$,

$$Q = \boldsymbol{\eta}'\cdot P_{22}\boldsymbol{\eta}' + \boldsymbol{\xi}\cdot\left(P_{11} - P_{12}P_{22}^{-1}P_{12}^{\mathsf{T}}\right)\boldsymbol{\xi}\,, \tag{A.936}$$

as one checks by expanding. The map $(\boldsymbol{\xi},\boldsymbol{\eta}) \mapsto (\boldsymbol{\xi},\boldsymbol{\eta}')$ is a linear bijection of $\mathbb{R}^{2n}$, so $Q > 0$ if and only if the right-hand side is positive for every non-zero $(\boldsymbol{\xi},\boldsymbol{\eta}')$, which is the stated pair of conditions. Conversely $Q > 0$ forces $P_{22} > 0$, by setting $\boldsymbol{\xi} = \mathbf{0}$, after which the displayed identity applies. □

Proposition A.611 (The constraint on Γ)**.** *With Γ as in Equation (A.932) and $W_\psi \ge 0$, the real symmetric form*

$$Q(\boldsymbol{\xi},\boldsymbol{\eta}) = \boldsymbol{\xi}\cdot\Gamma_{\mathrm{r}}\boldsymbol{\xi} - 2\,\boldsymbol{\xi}\cdot\Gamma_{\mathrm{i}}\boldsymbol{\eta} + \boldsymbol{\eta}\cdot\left(\frac{\mathbb{1}}{2s^2} - \Gamma_{\mathrm{r}}\right)\boldsymbol{\eta} \tag{A.937}$$

is positive definite on $\mathbb{R}^6$. Equivalently, $\Gamma_{\mathrm{r}} < \mathbb{1}/2s^2$ and

$$\Gamma_{\mathrm{r}} - \Gamma_{\mathrm{i}}\left(\frac{\mathbb{1}}{2s^2} - \Gamma_{\mathrm{r}}\right)^{-1}\Gamma_{\mathrm{i}} > 0\,. \tag{A.938}$$

Rests on Lemmas A.603 and A.610 and Proposition A.609.

Proof. *Derives Proposition A.611.*

By Equation (A.923), $|\langle\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}|\psi\rangle| = (\pi s^2)^{-3/4}\,|F(\zeta)|\,\mathrm{e}^{-|\mathrm{Im}\,\zeta|^2/2s^2}$ with $\zeta = \boldsymbol{\xi} + \mathrm{i}\boldsymbol{\eta}$, $\boldsymbol{\xi} = \boldsymbol{x}_0$ and $\boldsymbol{\eta} = -s^2\boldsymbol{p}_0/\hbar$. That substitution has Jacobian $\mathrm{d}^3x_0\,\mathrm{d}^3p_0 = (\hbar/s^2)^3\mathrm{d}^3\xi\,\mathrm{d}^3\eta$, so Equation (A.920) says that

$$\int_{\mathbb{R}^6} \exp\left(2\,\mathrm{Re}\,g(\boldsymbol{\xi}+\mathrm{i}\boldsymbol{\eta}) - \frac{|\boldsymbol{\eta}|^2}{s^2}\right)\mathrm{d}^3\xi\,\mathrm{d}^3\eta = (2\pi\hbar)^3\left(\pi s^2\right)^{3/2}\left(\frac{s^2}{\hbar}\right)^3 < \infty\,. \tag{A.939}$$

Now compute the real part of the exponent from Equation (A.932). Since $\zeta\cdot\Gamma\zeta = \boldsymbol{\xi}\cdot\Gamma\boldsymbol{\xi} - \boldsymbol{\eta}\cdot\Gamma\boldsymbol{\eta} + 2\mathrm{i}\,\boldsymbol{\xi}\cdot\Gamma\boldsymbol{\eta}$,

$$\mathrm{Re}\,g(\boldsymbol{\xi}+\mathrm{i}\boldsymbol{\eta}) = -\boldsymbol{\xi}\cdot\Gamma_\mathrm{r}\boldsymbol{\xi} + \boldsymbol{\eta}\cdot\Gamma_\mathrm{r}\boldsymbol{\eta} + 2\,\boldsymbol{\xi}\cdot\Gamma_\mathrm{i}\boldsymbol{\eta} + L(\boldsymbol{\xi},\boldsymbol{\eta})\,, \tag{A.940}$$

with L real affine. The exponent of Equation (A.939) is therefore $-2Q(\boldsymbol{\xi},\boldsymbol{\eta}) + 2L$, with Q as in Equation (A.937). An integral $\int_{\mathbb{R}^6}\mathrm{e}^{-2Q+2L}$ over the whole space is finite if and only if the quadratic part Q is positive definite: if some $(\boldsymbol{\xi}_0,\boldsymbol{\eta}_0)\neq 0$ had $Q(\boldsymbol{\xi}_0,\boldsymbol{\eta}_0)\leq 0$ then along the ray $t(\boldsymbol{\xi}_0,\boldsymbol{\eta}_0)$ the integrand would be at least e^{2tL_0} with L_0 the value of the linear part, which is bounded below on the ray only if $L_0\geq 0$ and in either case the integral over a solid cone about the ray diverges, Q being continuous and hence bounded above on a neighbourhood of the ray. So $Q>0$, and Lemma A.610 with $P_{22} = \mathbb{1}/2s^2 - \Gamma_\mathrm{r}$ and $P_{12} = -\Gamma_\mathrm{i}$ turns that into Equation (A.938) together with $P_{22}>0$. □

A.50.7 Recovering the wavefunction

Proposition A.612 (The dictionary between A and Γ)**.** *Let A be complex symmetric with $A_\mathrm{r}>0$ and put $N = A+\mathbb{1}/2s^2$, so that $N_\mathrm{r}>0$. Then the function $\psi_0 = \exp(-\boldsymbol{x}\cdot A\boldsymbol{x}+\boldsymbol{b}\cdot\boldsymbol{x}+c)$ has*

$$F_0(\zeta) = \exp(-\zeta\cdot\Gamma\zeta + \boldsymbol{\lambda}\cdot\zeta + \nu)\,,\qquad \Gamma = \frac{\mathbb{1}}{2s^2} - \frac{N^{-1}}{4s^4}\,, \tag{A.941}$$

for suitable $\boldsymbol{\lambda}$ and ν. Conversely, a complex symmetric Γ arises this way from exactly one such A if and only if it satisfies the conditions of Proposition A.611.

Rests on Lemmas A.600 and A.601 and Proposition A.611.

Proof. *Derives Proposition A.612.*

Forward. Insert ψ_0 into Equation (A.922). Expanding $(\boldsymbol{x}-\zeta)\cdot(\boldsymbol{x}-\zeta) = |\boldsymbol{x}|_*^2 - 2\zeta\cdot\boldsymbol{x}+\zeta\cdot\zeta$, where $|\boldsymbol{x}|_*^2 = \boldsymbol{x}\cdot\boldsymbol{x}$, the exponent is $-\boldsymbol{x}\cdot N\boldsymbol{x} + (\boldsymbol{b}+\zeta/s^2)\cdot\boldsymbol{x} + c - \zeta\cdot\zeta/2s^2$, and Equation (A.914) evaluates the integral as $C_N\exp\left[(\boldsymbol{b}+\zeta/s^2)\cdot N^{-1}(\boldsymbol{b}+\zeta/s^2)/4 + c - \zeta\cdot\zeta/2s^2\right]$. The part quadratic in ζ is $\zeta\cdot N^{-1}\zeta/4s^4 - \zeta\cdot\zeta/2s^2$, which is $-\zeta\cdot\Gamma\zeta$ with Γ as in Equation (A.941); the rest is affine in ζ and supplies $\boldsymbol{\lambda}$ and ν.

Converse. Solving Equation (A.941) for N gives $N^{-1} = K$ with

$$K = 4s^4\left(\frac{\mathbb{1}}{2s^2} - \Gamma\right) = 2s^2\mathbb{1} - 4s^4\Gamma\,,\qquad K_\mathrm{r} = 4s^4\left(\frac{\mathbb{1}}{2s^2} - \Gamma_\mathrm{r}\right)\,,\quad K_\mathrm{i} = -4s^4\Gamma_\mathrm{i}\,. \tag{A.942}$$

The first condition of Proposition A.611, $\Gamma_\mathrm{r} < \mathbb{1}/2s^2$, is exactly $K_\mathrm{r}>0$, which by Lemma A.600 makes K invertible; set $N = K^{-1}$ and $A = N - \mathbb{1}/2s^2$, complex symmetric and uniquely determined. It remains to show that $A_\mathrm{r}>0$ is equivalent to Equation (A.938). By Equation (A.913),

$$N_\mathrm{r} = \mathrm{Re}\left(K^{-1}\right) = \left(K_\mathrm{r} + K_\mathrm{i}K_\mathrm{r}^{-1}K_\mathrm{i}\right)^{-1}\,, \tag{A.943}$$

so, $A_\mathrm{r} = N_\mathrm{r} - \mathbb{1}/2s^2$ and both matrices being real symmetric with the inverse order-reversing on positive definite matrices,

$$A_\mathrm{r}>0 \iff N_\mathrm{r} > \frac{\mathbb{1}}{2s^2} \iff K_\mathrm{r} + K_\mathrm{i}K_\mathrm{r}^{-1}K_\mathrm{i} < 2s^2\mathbb{1}\,. \tag{A.944}$$

Substituting Equation (A.942), the left-hand side is $2s^2\mathbb{1} - 4s^4\Gamma_{\mathrm{r}} + 4s^4\Gamma_{\mathrm{i}}(\mathbb{1}/2s^2 - \Gamma_{\mathrm{r}})^{-1}\Gamma_{\mathrm{i}}$, the factors $4s^4$ cancelling against the inverse of K_{r}; so the last inequality of Equation (A.944) is $\Gamma_{\mathrm{i}}(\mathbb{1}/2s^2 - \Gamma_{\mathrm{r}})^{-1}\Gamma_{\mathrm{i}} < \Gamma_{\mathrm{r}}$, which is Equation (A.938). □

Proof of Theorem A.598. *Derives Theorem A.598.*

The "if" half is Proposition A.602, which also gives the last sentence of the theorem.

For the "only if" half, let $W_\psi \geq 0$ everywhere. By Proposition A.609 the entire function F of Equation (A.922) is e^g with g a quadratic polynomial, of matrix Γ; by Proposition A.611 that Γ satisfies the positivity conditions; and by Proposition A.612 there is therefore exactly one complex symmetric A with $A_{\mathrm{r}} > 0$, together with a $\boldsymbol{b}$ and a c matching the affine part, such that the Gaussian $\psi_0 = \exp(-\boldsymbol{x}\cdot A\boldsymbol{x} + \boldsymbol{b}\cdot\boldsymbol{x} + c)$ has $F_0 = F$. Because $A_{\mathrm{r}} > 0$, ψ_0 lies in $L^2(\mathbb{R}^3)$.

Both ψ and ψ_0 therefore have the same overlaps with every coherent state: by Equation (A.923) applied to each of them, and with the same non-vanishing factor κ,

$$\langle\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}|\psi\rangle = \frac{F(\zeta)}{\kappa} = \frac{F_0(\zeta)}{\kappa} = \langle\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}|\psi_0\rangle\ , \tag{A.945}$$

for every $\boldsymbol{x}_0$ and $\boldsymbol{p}_0$. Hence $\psi - \psi_0 \in L^2(\mathbb{R}^3)$ is orthogonal to every $\varphi_{\boldsymbol{x}_0,\boldsymbol{p}_0}$, that is to every $W(\zeta)\Omega_s$, and those vectors are total in $L^2(\mathbb{R}^3)$ by Proposition A.592. So $\psi = \psi_0$ in L^2, which is Equation (A.910) almost everywhere. □

A.50.8 Scope, and what is quoted

Remark A.613 (What is quoted here). Nothing. Remark A.596 explains why the Hadamard factorization theorem, named as the import in Remark 25.56, is not needed: Lemma A.608 replaces it and is proved from Theorem 8.20 and Equation (8.13). The other inputs are the spectral theorem for a real symmetric matrix (Theorem 5.80), differentiation under the integral sign (Theorem 7.77), the Cauchy–Riemann criterion (Theorem 8.6), the Gaussian transform pair (Equation (17.42)), the Wigner-function properties (Theorem 25.53, in particular the overlap identity Equation (25.56)), and the totality of the coherent states, proved in Proposition A.592 of Appendix A.49. Every one of them is proved in this book.

The auxiliary width s is arbitrary, exactly as in Remark A.588: it enters the definitions of φ, F and Γ and cancels out of the conclusion, which mentions only A. A reader who repeats the argument with a different s obtains a different Γ and the same ψ.

Remark A.614 (What the theorem does not say). Two limits, both stated in Remark 25.56 and both worth repeating where the proof can be pointed at.

The theorem is about *pure* states, and the restriction is essential rather than technical: every step from Proposition A.604 onwards uses that W_ψ is built from a single wavefunction, through the identity Equation (A.919) in which the right-hand side is $|\langle\varphi|\psi\rangle|^2$ and not merely a trace. Mixed states with everywhere non-negative Wigner function that are not mixtures of Gaussians do exist, so "non-negative Wigner function" and "classical" are not synonyms in general, and Corollary 25.54 is an exact witness only inside the pure family.

And the attribution stands without a source the reader of this book can check: Hudson's 1974 paper, *When is the Wigner quasi-probability density non-negative?* (Reports on Mathematical Physics **6**, 249–252), has no entry in this bibliography, and the several-variable form used here is due to Soto and Claverie (1983). Both are cited in words only, on the footing of Darboux's memoir in Remark A.81. That is precisely why the proof is written out in full: nothing in Chapter 25 rests on the attribution.

Remark A.615. Appendix A.50 discharges the derivation owed at Theorem 25.55 of Chapter 25. Its use there is as the converse half of Corollary 25.54: that corollary exhibits states whose Wigner function reaches the most negative value Equation (25.57) permits, and the theorem proved here says that the ones which avoid negativity altogether are exactly the

Gaussians — so that within the pure states, negativity of W and non-Gaussianity are the same property, and an experiment that reconstructs a Wigner function and finds it negative has measured a departure from the Gaussian family and nothing weaker. The two sections of this appendix that serve Chapter 25 are linked in one direction: Appendix A.49 supplies the totality of the coherent states used at the end of the proof above, and both are built on the same Gaussian, Equation (A.902) there and Equation (A.918) here.

A.51 The Jacobi Identity for the Dirac Bracket

This appendix proves the one clause of Theorem 26.21 of Chapter 26 that the chapter defers: the Dirac bracket Equation (26.25) obeys the Jacobi identity Equation (22.46). Bilinearity, antisymmetry, the Leibniz rule Equation (24.24), the strong vanishing Equation (26.26) and the two statements Equations (26.27) and (26.28) are one-line computations and are done in the chapter; the identity below is what earns the object the name *bracket*, because without it the correspondence rule Equation (25.31) would send it to something that is not a commutator.

Two independent routes are written out, as the chapter's prose promises. Appendix A.51.2 is the direct expansion: it uses nothing but the Poisson Jacobi identity Equation (22.46), the derivation property Proposition 25.4 and the antisymmetry of C^{-1}, and it is the route that shows *where* the identity comes from — every one of the four orders in C^{-1} cancels for its own reason, and three of the four reasons are the Poisson Jacobi identity applied to a different triple. Appendix A.51.3 is the structural route: the Dirac bracket is the Poisson bracket of the symplectic manifold that Proposition 26.25 constructs on the second-class surface, and the identity is then read off the closure of a two-form. The second is three paragraphs long and is the one to remember; it is written second because it consumes Proposition 26.25, which is proved in the chapter after the Dirac bracket is introduced.

Nothing outside this treatise is used anywhere in this section.

A.51.1 Notation, and the four identities consumed

Throughout, χ_α, $\alpha = 1, \dots, S$, is a complete set of second-class constraints in the sense of Definition 26.12, $C_{\alpha\beta} = \{\chi_\alpha, \chi_\beta\}$ is the matrix Equation (26.24), which is invertible on a neighbourhood of the constraint surface by Proposition 26.19 and continuity, and

$$D^{\alpha\beta} := \left(C^{-1}\right)^{\alpha\beta}, \qquad D^{\alpha\gamma}C_{\gamma\beta} = \delta^\alpha_\beta, \qquad D^{\alpha\beta} = -D^{\beta\alpha}. \tag{A.946}$$

The antisymmetry of D is proved in the chapter, in the antisymmetry step of Theorem 26.21: $C^{\mathsf{T}} = -C$ gives $\left(C^{-1}\right)^{\mathsf{T}} = \left(C^{\mathsf{T}}\right)^{-1} = -C^{-1}$. That S is even, so that such a C can exist at all, is Proposition 26.19. With this notation Equation (26.25) reads

$$\{A, B\}_{\mathrm{D}} = \{A, B\} + A_\alpha D^{\alpha\beta} B_\beta, \qquad \text{where} \qquad A_\alpha := \{A, \chi_\alpha\}, \tag{A.947}$$

the sign having flipped because $\{\chi_\beta, B\} = -B_\beta$. Two further abbreviations carry the whole computation:

$$A_{\alpha\beta} := \{\{A, \chi_\alpha\}, \chi_\beta\} = \{A_\alpha, \chi_\beta\}, \qquad \widehat{A}^\alpha := A_\beta D^{\beta\alpha}. \tag{A.948}$$

Note that $A_{\alpha\beta}$ is *not* symmetric — its antisymmetric part is Equation (A.949) below — and that the index of $\widehat{A}$ is raised from the left, so that $A_\alpha D^{\alpha\beta} B_\beta = \widehat{A}^\beta B_\beta$ while $D^{\alpha\beta} B_\beta = -\widehat{B}^\alpha$.

Dimensionally there is nothing to watch: $D^{\alpha\beta}$ carries the inverse of the dimension of $C_{\alpha\beta}$, so the correction term of Equation (A.947) has exactly the dimension of $\{A, B\}$, whatever SI dimensions the individual χ_α happen to have. Nothing below depends on the constraints being dimensionally homogeneous among themselves.

Lemma A.616 (The four identities)**.** *For all phase-space functions A, B, E, wherever C^{-1} is defined,*

$$\{A, C_{\alpha\beta}\} = A_{\alpha\beta} - A_{\beta\alpha}\,, \tag{A.949}$$

$$\{\{B, E\}, \chi_\delta\} = \{B, E_\delta\} - \{E, B_\delta\}\,, \tag{A.950}$$

$$\left\{A, D^{\alpha\beta}\right\} = -D^{\alpha\mu}\{A, C_{\mu\nu}\}D^{\nu\beta}\,, \tag{A.951}$$

$$T_{\alpha\beta\gamma} + T_{\beta\gamma\alpha} + T_{\gamma\alpha\beta} = 0\,, \qquad T_{\alpha\beta\gamma} := \{C_{\alpha\beta}, \chi_\gamma\}\,. \tag{A.952}$$

Rests on Equations (22.46) and (26.24) and Proposition 25.4.

Proof. *Derives Lemma A.616.*

Three of the four are the Poisson Jacobi identity Equation (22.46) applied to a different triple, and the fourth is the derivative of an inverse.

Equation (A.949). Take the triple $(A, \chi_\alpha, \chi_\beta)$ in Equation (22.46):

$$\{A, \{\chi_\alpha, \chi_\beta\}\} + \{\chi_\alpha, \{\chi_\beta, A\}\} + \{\chi_\beta, \{A, \chi_\alpha\}\} = 0\,. \tag{A.953}$$

The second term is $\{\chi_\alpha, -A_\beta\} = +A_{\beta\alpha}$ and the third is $\{\chi_\beta, A_\alpha\} = -A_{\alpha\beta}$, which gives Equation (A.949).

Equation (A.950). Take the triple (χ_δ, B, E): $\{\chi_\delta, \{B, E\}\} + \{B, \{E, \chi_\delta\}\} + \{E, \{\chi_\delta, B\}\} = 0$, that is $-\{\{B, E\}, \chi_\delta\} + \{B, E_\delta\} - \{E, B_\delta\} = 0$.

Equation (A.951). The map $X \longmapsto \{A, X\}$ is a derivation (Proposition 25.4), so applying it to $D^{\alpha\mu}C_{\mu\beta} = \delta^\alpha_\beta$ gives $\{A, D^{\alpha\mu}\}C_{\mu\beta} + D^{\alpha\mu}\{A, C_{\mu\beta}\} = 0$; contracting with $D^{\beta\nu}$ isolates $\{A, D^{\alpha\nu}\}$. This is the *only* fact about the inverse matrix that the computation needs, and it is why D may be kept as an unexpanded symbol throughout.

Equation (A.952). Take the triple $(\chi_\gamma, \chi_\alpha, \chi_\beta)$: each of the three terms of Equation (22.46) is minus one of the T's, in the cyclic order stated. □

A.51.2 Route A: the direct expansion

Theorem A.617 (Jacobi identity for the Dirac bracket)**.** *Wherever $C_{\alpha\beta}$ is invertible, and for all phase-space functions A, B, E,*

$$\{A, \{B, E\}_D\}_D + \{B, \{E, A\}_D\}_D + \{E, \{A, B\}_D\}_D = 0\,, \tag{A.954}$$

strongly — *as an identity of functions, not merely on the constraint surface.*

Rests on Definition 26.20, Equation (22.46), and Lemma A.616.

Proof. *Derives Theorem A.617.*

Write $\mathfrak{S}$ for the sum over the three cyclic images $(A, B, E) \to (B, E, A) \to (E, A, B)$, so that the left-hand side of Equation (A.954) is $\mathfrak{S}\,\{A, \{B, E\}_{\mathrm{D}}\}_{\mathrm{D}}$.

Step 1: expand once and sort by the number of D factors. Put $F := \{B, E\}_{\mathrm{D}} = \{B, E\} + B_\alpha D^{\alpha\beta} E_\beta$. By Equation (A.947) applied a second time, $\{A, F\}_{\mathrm{D}} = \{A, F\} + A_\gamma D^{\gamma\delta} F_\delta$ with $F_\delta = \{F, \chi_\delta\}$. Expanding both pieces by the Leibniz rule (Proposition 25.4) and using Equation (A.950) on $\{\{B, E\}, \chi_\delta\}$,

$$\{A, F\} = \{A, \{B, E\}\} + \{A, B_\alpha\} D^{\alpha\beta} E_\beta + B_\alpha D^{\alpha\beta}\{A, E_\beta\} + B_\alpha \left\{A, D^{\alpha\beta}\right\} E_\beta\,, \tag{A.955}$$

$$F_\delta = \{B, E_\delta\} - \{E, B_\delta\} + B_{\alpha\delta} D^{\alpha\beta} E_\beta + B_\alpha D^{\alpha\beta} E_{\beta\delta} + B_\alpha \left\{D^{\alpha\beta}, \chi_\delta\right\} E_\beta\,. \tag{A.956}$$

Assembling, and using Equation (A.951) on $\left\{A, D^{\alpha\beta}\right\}$ and on $\left\{D^{\alpha\beta}, \chi_\delta\right\}$, the eight surviving terms sort by the number of factors of D they carry:

$$\begin{aligned}\{A, \{B, E\}_{\mathrm{D}}\}_{\mathrm{D}} = {} & \underbrace{\{A, \{B, E\}\}}_{\text{order 0}} + \underbrace{\{A, B_\alpha\} D^{\alpha\beta} E_\beta + B_\alpha D^{\alpha\beta}\{A, E_\beta\} + A_\gamma D^{\gamma\delta}\left(\{B, E_\delta\} - \{E, B_\delta\}\right)}_{\text{order 1}} \\ & + \underbrace{\widehat{B}^\mu \widehat{E}^\nu \{A, C_{\mu\nu}\} - \widehat{A}^\delta \widehat{E}^\alpha B_{\alpha\delta} + \widehat{A}^\delta \widehat{B}^\beta E_{\beta\delta}}_{\text{order 2}} + \underbrace{\widehat{A}^\delta \widehat{B}^\mu \widehat{E}^\nu T_{\mu\nu\delta}}_{\text{order 3}}\,.\end{aligned} \tag{A.957}$$

The three order-two terms and the order-three term are obtained by substituting Equation (A.951) and then absorbing each $A_\gamma D^{\gamma\delta}$ and $B_\alpha D^{\alpha\mu}$ into the hatted abbreviation of Equation (A.948), remembering from it that $D^{\nu\beta}E_\beta = -\widehat{E}^\nu$, which is where the sign of the first order-two term and of the order-three term comes from.

Step 2: order zero. $\mathfrak{S}\,\{A,\{B,E\}\} = 0$ is the Poisson Jacobi identity Equation (22.46) itself.

Step 3: order one. Introduce the trilinear object

$$G(U,V,W) := \{U, V_\alpha\}\, D^{\alpha\beta} W_\beta\,. \tag{A.958}$$

The first order-one term is $G(A,B,E)$. For the second, relabel $\alpha \leftrightarrow \beta$ and use $D^{\beta\alpha} = -D^{\alpha\beta}$: $B_\alpha D^{\alpha\beta}\,\{A, E_\beta\} = -G(A,E,B)$. For the third, the same manoeuvre on each of the two pieces gives $A_\gamma D^{\gamma\delta}\,\{B, E_\delta\} = -G(B,E,A)$ and $-A_\gamma D^{\gamma\delta}\,\{E, B_\delta\} = +G(E,B,A)$. The order-one part of Equation (A.957) is therefore

$$G(A,B,E) - G(A,E,B) - G(B,E,A) + G(E,B,A)\,. \tag{A.959}$$

Now apply $\mathfrak{S}$. A cyclic sum of G is a sum of three terms that depends only on which of the two *orientations* of the triple its arguments carry, so $\mathfrak{S}G(B,E,A) = \mathfrak{S}G(A,B,E)$ and $\mathfrak{S}G(E,B,A) = \mathfrak{S}G(A,E,B)$: in each case the three summands are the same three terms in a different order. Hence the first term of Equation (A.959) cancels the third and the second cancels the fourth, and the cyclic sum vanishes. Concretely, the summand $G(A,B,E)$ produced by the first term of the (A,B,E) copy is cancelled by the summand $-G(A,B,E)$ produced by the third term of the (E,A,B) copy; the pairing of the second and fourth terms is the same one cycle over. This is the step that a reader cannot reconstruct without being told which term meets which, and it is the reason Equation (A.950) had to be used in Step 1: without it the third order-one term is a bracket of a bracket and does not have the shape Equation (A.958) at all.

Step 4: order two. Replace $\{A, C_{\mu\nu}\}$ by $A_{\mu\nu} - A_{\nu\mu}$, which is Equation (A.949), and relabel so that every term carries the free index pair in the order (p,q):

$$X(A,B,E) = \widehat{B}^p\widehat{E}^q A_{pq} - \widehat{B}^q\widehat{E}^p A_{pq} - \widehat{A}^q\widehat{E}^p B_{pq} + \widehat{A}^q\widehat{B}^p E_{pq}\,, \tag{A.960}$$

the second term coming from $-\widehat{B}^p\widehat{E}^q A_{qp}$ by the interchange $p \leftrightarrow q$. Now collect the cyclic sum by which of A_{pq}, B_{pq}, E_{pq} each term carries. The coefficient of A_{pq} receives $\widehat{B}^p\widehat{E}^q - \widehat{B}^q\widehat{E}^p$ from $X(A,B,E)$, $+\widehat{B}^q\widehat{E}^p$ from the fourth term of $X(B,E,A)$ and $-\widehat{E}^q\widehat{B}^p$ from the third term of $X(E,A,B)$; the four contributions cancel in pairs. The coefficients of B_{pq} and of E_{pq} are the same computation read one and two cycles on, so $\mathfrak{S}X = 0$.

Step 5: order three. By Equation (A.952) the totally contracted coefficient is a cyclic sum of T. Writing the order-three term of Equation (A.957) as $\widehat{A}^p\widehat{B}^q\widehat{E}^r T_{qrp}$ and applying $\mathfrak{S}$, which permutes $\widehat{A},\widehat{B},\widehat{E}$ cyclically, the three summands are $\widehat{A}^p\widehat{B}^q\widehat{E}^r$ times, in turn, T_{qrp}, T_{rpq} and T_{pqr}. Their sum vanishes by Equation (A.952).

All four orders vanish separately, so Equation (A.954) holds. No constraint was set to zero anywhere in the argument, so the identity is strong, as claimed. □

Remark A.618 (What each order used). It is worth recording the audit, because the four cancellations look alike and are not. Order zero is the Poisson Jacobi identity on the triple (A,B,E). Order one is the Poisson Jacobi identity on (χ_δ,B,E), which turns a bracket of a bracket into two terms of the shape Equation (A.958), plus the antisymmetry of D. Order two is the Poisson Jacobi identity on $(A,\chi_\alpha,\chi_\beta)$, which is what makes the derivative of the inverse matrix collapse onto the second brackets $A_{\alpha\beta}$ already present in the sum, plus the antisymmetry of D. Order three is the Poisson Jacobi identity on three constraints $(\chi_\gamma,\chi_\alpha,\chi_\beta)$. The antisymmetry of D enters at orders one, two and three, and never on its

own: at every order it does no more than allow the terms to be written with their indices in a common order, and the vanishing is always supplied by Equation (22.46). So the Dirac bracket inherits its Jacobi identity from the Poisson bracket four times over, once for each way of feeding a constraint into the triple.

A.51.3 Route B: the bracket of the induced symplectic form

The second proof is shorter, gives the reason rather than the verification, and is the one worth remembering. Its input is Proposition 26.25, proved in the chapter: with $C_{\alpha\beta}$ invertible at x, the tangent space splits as $T_xM = W \oplus T_x\Sigma_\chi$ along the projector Equation (26.42), the two summands are ω-orthogonal, and the restriction ω_Σ of the symplectic form to $T_x\Sigma_\chi$ is non-degenerate, so that $(\Sigma_\chi, \omega_\Sigma)$ is a symplectic manifold.

Proposition A.619 (The Dirac bracket is a projected Hamiltonian flow)**.** *Let X_A be the Hamiltonian vector field of A in the ambient space, Equation (24.5), and let P be the projector Equation (26.42). Then, for all A and B,*

$$\{B, A\}_D = \mathrm{d}B\left(P\left(X_A\right)\right) . \tag{A.961}$$

On Σ_χ the vector field $P(X_A)$ is the Hamiltonian vector field of the restriction $A|_{\Sigma_\chi}$ with respect to ω_Σ, and consequently

$$\{A, B\}_D\Big|_{\Sigma_\chi} = \{A|_{\Sigma_\chi}, B|_{\Sigma_\chi}\}_{\omega_\Sigma} . \tag{A.962}$$

Rests on Proposition 26.25 and Equations (24.7) and (26.42).

Proof. *Derives Proposition A.619.*

By Equation (24.7), $\{u, v\} = \mathrm{d}u\left(X_v\right)$ for every pair of functions. Hence $\{B, A\} = \mathrm{d}B\left(X_A\right)$, $\{B, \chi_\alpha\} = \mathrm{d}B\left(X_{\chi_\alpha}\right)$ and $\{\chi_\beta, A\} = \mathrm{d}\chi_\beta\left(X_A\right)$, so that Equation (26.25) reads

$$\{B, A\}_{\mathrm{D}} = \mathrm{d}B\left(X_A\right) - \mathrm{d}B\left(X_{\chi_\alpha}\right)\left(C^{-1}\right)^{\alpha\beta}\mathrm{d}\chi_\beta\left(X_A\right) = \mathrm{d}B\Big(X_A - X_{\chi_\alpha}\left(C^{-1}\right)^{\alpha\beta}\mathrm{d}\chi_\beta\left(X_A\right)\Big) , \tag{A.963}$$

and the argument of $\mathrm{d}B$ is exactly $P\left(X_A\right)$ of Equation (26.42). That is the whole of Equation (A.961), and it is the whole point: the Dirac bracket differs from the Poisson bracket only in that the Hamiltonian vector field is projected onto the constraint surface before it is used.

For the second statement, let $v \in T_x\Sigma_\chi$. The ω-orthogonality of the splitting gives $\omega(X_{\chi_\alpha}, v) = \mathrm{d}\chi_\alpha(v) = 0$, so

$$\omega_\Sigma(P\left(X_A\right), v) = \omega(X_A, v) - \left(C^{-1}\right)^{\alpha\beta}\mathrm{d}\chi_\beta\left(X_A\right)\omega(X_{\chi_\alpha}, v) = \mathrm{d}A(v) . \tag{A.964}$$

Since $P\left(X_A\right)$ is tangent to Σ_χ and $\mathrm{d}A(v)$ for tangent v depends on A only through $A|_{\Sigma_\chi}$, this says precisely that $P\left(X_A\right)$ is the ω_Σ-Hamiltonian vector field of $A|_{\Sigma_\chi}$. Feeding that back into Equation (A.961) with the roles of A and B exchanged, and using antisymmetry of both brackets, gives Equation (A.962). □

Second derivation of the Jacobi identity. *Derives Theorem A.617.*

This is a second, independent proof of Theorem A.617. Fix a point x at which $C_{\alpha\beta}$ is invertible, and let $c_\alpha := \chi_\alpha(x)$. The shifted functions $\chi_\alpha - c_\alpha$ have the same brackets as the χ_α, hence the same invertible matrix $C_{\alpha\beta}$, and they define the level surface Σ_c through x. Everything in Proposition 26.25 and in Proposition A.619 therefore applies verbatim to Σ_c, and the Dirac bracket built from $\chi_\alpha - c_\alpha$ is, term by term in Equation (26.25), the Dirac bracket built from the χ_α: constants drop out of every bracket. So the level surfaces

of the constraints foliate a neighbourhood of Σ_χ by symplectic manifolds (Σ_c, ω_c), and by Equation (A.962) the Dirac bracket of two functions restricted to Σ_c is the Poisson bracket of (Σ_c, ω_c) applied to their restrictions.

Now ω_c is the pullback $i_c^*\omega$ of the ambient symplectic form along the inclusion, so $\mathrm{d}\omega_c = i_c^*\mathrm{d}\omega = 0$: it is closed because ω is, and non-degenerate by Proposition 26.25. A symplectic manifold's Poisson bracket satisfies the Jacobi identity — that is Equation (22.46) on (Σ_c, ω_c), which by Theorem 24.12 may be computed in canonical coordinates, where it is the same computation as in the ambient space. Hence the cyclic sum Equation (A.954), restricted to Σ_c, vanishes.

Finally, the value of the cyclic sum at x is determined by its restriction to the leaf through x: by Equation (A.962) each of its three terms is, at x, the corresponding double bracket of (Σ_c, ω_c) applied to restrictions. Since x was an arbitrary point of the neighbourhood on which C^{-1} exists, Equation (A.954) holds there identically. □

Remark A.620 (Why the leaves, and not the surface alone). The last paragraph is the step at which a shorter-looking argument fails, and the failure is worth naming. It is tempting to say that Equation (26.26) lets one add any multiple of a constraint to either argument without changing the Dirac bracket, so that an identity holding on Σ_χ holds everywhere. It does not: by the Leibniz rule, $\{A, c^\alpha\chi_\alpha\}_{\mathrm{D}} = \chi_\alpha \{A, c^\alpha\}_{\mathrm{D}}$, which vanishes *on* Σ_χ and not off it. What Equation (26.26) does say — and this is the right reading — is that every χ_α is a Casimir of the Dirac bracket, so its level sets are unions of symplectic leaves, and the foliation used above is not an artifice but the canonical structure the bracket itself defines. The Dirac bracket makes a neighbourhood of Σ_χ a Poisson manifold in the sense of Definition 24.36, whose symplectic leaves are the Σ_c; Theorem A.617 is the assertion that it is a Poisson bracket at all.

Remark A.621 (What is quoted here). Nothing. Both routes are carried out from material proved in this treatise: the Poisson Jacobi identity Equation (22.46), the derivation property Proposition 25.4, the relation between the bracket and the symplectic form Equation (24.7), Darboux's theorem Theorem 24.12 in the form proved in Appendix A.12, and the splitting of Proposition 26.25. The one hypothesis that is not proved but assumed is the regularity of the constraint set (Remark 26.7), which is what makes Σ_χ a submanifold in the first place; it is an assumption about the system under study and is stated as such in the chapter.

Remark A.622. Appendix A.51 discharges the derivation owed at Theorem 26.21 of Chapter 26, whose proof establishes every other property of the Dirac bracket and defers this one. With it the object of Definition 26.20 is a bracket in the full sense, so that Postulate 26.47 may send it to a commutator, and the three facts the chapter draws together are seen to be one fact in three costumes: the Jacobi identity proved here, the degree-of-freedom count of Theorem 26.17, and the reduced Liouville measure of Proposition 26.25 are the tangential, the dimensional and the volumetric readings of the single statement that $(\Sigma_\chi, \omega_\Sigma)$ is a symplectic manifold. Remark 26.26 makes the same point from the measure side; Example 26.22 is the worked instance, where the leaves are the tangent bundles of a surface in $\mathbb{R}^3$ and the reduced bracket is the one Chapter 21 obtains by eliminating a coordinate.

A.52 The Gauss–Codazzi Form of the Einstein–Hilbert Lagrangian

This appendix proves Proposition 26.41 of Chapter 26: that the $3+1$ decomposition Equation (26.78) gives $\sqrt{|g|} = N\sqrt{h}$, that the four-dimensional curvature scalar splits into the intrinsic curvature of the leaf, a quadratic in the extrinsic curvature and two divergences, and — the conclusion the chapter actually consumes — that no time derivative of the lapse or of the shift survives anywhere in the Einstein–Hilbert Lagrangian.

The mathematics is the Gauss–Codazzi decomposition of the Riemann tensor of an ambient manifold along a hypersurface. The difficulty is not the decomposition, which is four projections of one identity; it is the *signature bookkeeping*, and this treatise sits in exactly the place where that bookkeeping is worst. The metric of Chapter 39 is mostly minus, so the unit normal to a spacelike leaf satisfies $n \cdot n = +1$; the curvature conventions of Theorem 13.152 and Definition 13.153 are the ones for which the round two-sphere has $R > 0$; and Definition 26.40 takes h_{ij} to be the *positive-definite* Riemannian metric of the leaf, so that the metric g_{ij} actually induced on the leaf by Equation (26.78) is $-h_{ij}$. Three sign conventions, and every formula in the literature is written with a different combination of them. Nothing below is inherited from a mostly-plus source. The discipline used instead is this. Write

$$\epsilon := n^\mu n_\mu \in \{+1, -1\} \,, \tag{A.965}$$

carry ϵ as a symbol through every step, do the whole computation with the *induced* metric $\gamma_{\mu\nu} = g_{\mu\nu} - \epsilon n_\mu n_\nu$ — whatever its sign — and only at the very end substitute $\epsilon = +1$ and $\gamma_{ij} = -h_{ij}$. A reader checking against a mostly-plus text sets $\epsilon = -1$ and $\gamma_{ij} = +h_{ij}$ and compares line by line; Remark A.633 records what happens when the substitution is made, which is the one place where the answer differs from the printed statement of Proposition 26.41 and the reason the chapter's prose insists that the sign be settled here rather than assumed.

Throughout, the ambient connection ∇ is the Levi-Civita connection of $g_{\mu\nu}$ (Theorem 13.150 with vanishing torsion), ${R^\lambda}_{\rho\mu\nu}$ is Equation (13.306), and $R_{\mu\nu}$, R are the contractions Equation (13.307). Indices $\mu, \nu, \ldots$ run over $0, 1, 2, 3$ and $i, j, k, \ldots$ over $1, 2, 3$; $x^0 = ct$, as in Notation 26.28. Since h_{ij} and N, N^i are dimensionless and the x^i carry metres, every term of Equation (26.80) has the SI dimension $1/\mathrm{m}^2$, and the identity is dimensionally homogeneous as it stands; no factor of c or G enters until the Lagrangian is multiplied by $1/2\kappa$.

A.52.1 The kit: normal, inverse metric, projector, determinant

Lemma A.623 (The adapted frame of the $3+1$ split)**.** *Write Equation (26.78) in terms of the coframe*

$$\theta^0 = \mathrm{d}x^0 \,, \qquad \theta^i = \mathrm{d}x^i + N^i \,\mathrm{d}x^0 \,, \qquad \textit{so that} \qquad \mathrm{d}s^2 = N^2 \theta^0 \theta^0 - h_{ij} \theta^i \theta^j \,. \tag{A.966}$$

Then, with $N > 0$:

1. *the dual frame is $e_0 = \partial_0 - N^i \partial_i$, $e_i = \partial_i$, and the inverse metric is*

$$g^{00} = \frac{1}{N^2} \,, \qquad g^{0i} = -\frac{N^i}{N^2} \,, \qquad g^{ij} = \frac{N^i N^j}{N^2} - h^{ij} \,, \tag{A.967}$$

   with h^{ij} the inverse of h_{ij};

2. *the future-directed unit normal to the leaves $x^0 = \text{const}$ is*

$$n_\mu = N \delta^0_\mu \,, \qquad n^\mu = \left(\frac{1}{N}, -\frac{N^i}{N} \right) = \frac{1}{N} e_0 \,, \qquad \epsilon = n^\mu n_\mu = +1 \,, \tag{A.968}$$

   so the normal is timelike and the leaves are spacelike;

3. $\det g = -N^2 h$ *with* $h = \det h_{ij} > 0$*, hence*

$$\sqrt{|g|} = N \sqrt{h} \,. \tag{A.969}$$

Rests on Definitions 13.117 and 26.40 and Equation (26.78).

Proof. *Derives Lemma A.623.*

(i). That Equation (A.966) reproduces Equation (26.78) is the definition of θ^i. For the dual frame, $\theta^0(e_0) = 1$, $\theta^i(e_0) = -N^i + N^i = 0$, $\theta^0(e_i) = 0$ and $\theta^j(e_i) = \delta^j_i$, so $\{e_0, e_i\}$ is indeed dual to $\{\theta^0, \theta^i\}$. In that basis the metric is the block array $\operatorname{diag}(N^2, -h_{ij})$, whose inverse is the block array $\operatorname{diag}(N^{-2}, -h^{ij})$; hence

$$g^{-1} = \frac{1}{N^2}\, e_0 \otimes e_0 - h^{ij}\, e_i \otimes e_j \,, \tag{A.970}$$

and expanding $e_0 = \partial_0 - N^k \partial_k$ in coordinates gives Equation (A.967).

(ii). A covector annihilating every vector tangent to a leaf — that is, every ∂_i — is proportional to δ^0_μ. Normalizing, $g^{\mu\nu}\left(N\delta^0_\mu\right)\left(N\delta^0_\nu\right) = N^2 g^{00} = 1$ by Equation (A.967), so Equation (A.968) is a unit covector and $\epsilon = +1$. Raising with Equation (A.967), $n^\mu = N g^{\mu 0}$ gives the components stated, and $n^0 = 1/N > 0$ makes it future directed.

(iii). This is the step Proposition 26.41 asserts and does not prove. The change of coframe Equation (A.966) has the matrix $\theta^a = \Lambda^a{}_\mu \mathrm{d}x^\mu$ with $\Lambda^0{}_0 = 1$, $\Lambda^0{}_i = 0$, $\Lambda^i{}_0 = N^i$, $\Lambda^i{}_j = \delta^i_j$: it is lower triangular with unit diagonal, so $\det \Lambda = 1$. Since $g_{\mu\nu} = \Lambda^a{}_\mu \hat{g}_{ab} \Lambda^b{}_\nu$ with $\hat{g} = \operatorname{diag}(N^2, -h_{ij})$,

$$\det g = (\det \Lambda)^2 \det \hat{g} = N^2\,(-1)^3 \det h_{ij} = -N^2 h \,, \tag{A.971}$$

the factor $(-1)^3$ being the three minus signs the signature puts on the spatial block. Completing the square in $\mathrm{d}x^i + N^i \mathrm{d}x^0$ is thus exactly the statement that the shift can be removed from the determinant by a unimodular change of coframe. □

Definition A.624 (Projector and induced metric)**.** The *projector* onto a leaf and the *induced metric* are

$$\gamma^\mu{}_\nu := \delta^\mu_\nu - \epsilon\, n^\mu n_\nu \,, \qquad \gamma_{\mu\nu} := g_{\mu\nu} - \epsilon\, n_\mu n_\nu \,. \tag{A.972}$$

A tensor is *tangential* if every index is annihilated by contraction with n, equivalently if projecting every index returns it. For a tangential tensor T the *intrinsic* covariant derivative is

$$D_\mu T_{\nu\ldots} := \gamma^\alpha{}_\mu \gamma^\beta{}_\nu \cdots \nabla_\alpha T_{\beta\ldots} \,. \tag{A.973}$$

Rests on Equation (A.968) and Definition 13.145.

Three facts about Equation (A.972) are used without further comment. First, $\gamma^\mu{}_\nu n^\nu = n^\mu - \epsilon^2 n^\mu = 0$ and $\gamma^\mu{}_\nu v^\nu = v^\mu$ for v tangent to a leaf, since such a v has $n_\nu v^\nu = 0$; so γ is the projection along n, and $\gamma^\mu{}_\alpha \gamma^\alpha{}_\nu = \gamma^\mu{}_\nu$. Second, D is the Levi-Civita connection of γ on the leaf: it is torsion free, and $D_\mu \gamma_{\nu\rho} = \gamma\gamma\gamma\nabla\,(g - \epsilon nn) = 0$, the first term because $\nabla g = 0$ and the second because each projector annihilates the n it meets. Third — and this is the sign that will have to be paid for at the end — in the coordinates of Lemma A.623 one has $n_i = 0$, so

$$\gamma_{ij} = g_{ij} = -h_{ij} \,, \qquad \gamma^{ij} = -h^{ij} \,, \tag{A.974}$$

where γ^{ij} means the inverse of γ_{ij} *as a three-by-three matrix*, which is also $\gamma^{\mu\nu}$ restricted to spatial indices. The induced metric of Equation (26.78) is negative definite; the positive-definite h_{ij} of Definition 26.40 is its negative.

A.52.2 The extrinsic curvature, and the dictionary

Definition A.625 (Four-dimensional extrinsic curvature)**.** Put

$$\mathcal{K}_{\mu\nu} := \gamma^\alpha{}_\mu \gamma^\beta{}_\nu \nabla_\alpha n_\beta \,, \qquad \mathcal{K} := \gamma^{\mu\nu}\mathcal{K}_{\mu\nu} \,, \qquad a_\mu := n^\lambda \nabla_\lambda n_\mu \,. \tag{A.975}$$

Rests on Equation (A.972) and Definition 13.145.

Lemma A.626 (Decomposition of ∇n, and the acceleration)**.** *$\mathcal{K}_{\mu\nu}$ and a_μ are tangential, $\mathcal{K}$ is symmetric, and*

$$\nabla_\mu n_\nu = \mathcal{K}_{\mu\nu} + \epsilon\, n_\mu a_\nu \,, \tag{A.976}$$

$$\nabla_\mu n^\mu = \mathcal{K} \,, \tag{A.977}$$

$$(\nabla_\mu n^\nu)\,(\nabla_\nu n^\mu) = \mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} \,, \tag{A.978}$$

$$a_\mu = -\epsilon\, D_\mu \ln N \,. \tag{A.979}$$

Rests on Definition A.625, Lemma A.623, and Equation (13.295).

Proof. *Derives Lemma A.626.*

Tangentiality of $\mathcal{K}$ is built into Equation (A.975). For a, differentiate $n^\nu n_\nu = \epsilon$: $n^\nu \nabla_\mu n_\nu = 0$, and contracting with n^μ gives $n^\nu a_\nu = 0$.

Equation (A.976). Insert $\delta^\alpha_\mu = \gamma^\alpha{}_\mu + \epsilon n^\alpha n_\mu$ in both slots of $\nabla_\mu n_\nu$. The $\gamma\gamma$ term is $\mathcal{K}_{\mu\nu}$. Any term carrying n^β in the second slot vanishes, because $n^\beta \nabla_\alpha n_\beta = 0$. The remaining term is $\epsilon n^\alpha n_\mu \gamma^\beta{}_\nu \nabla_\alpha n_\beta = \epsilon n_\mu \gamma^\beta{}_\nu a_\beta = \epsilon n_\mu a_\nu$, since a is tangential.

Symmetry of $\mathcal{K}$ follows: the leaves are level sets of the coordinate x^0, so $n_\mu/N = \partial_\mu x^0$ is a gradient and $\nabla_{[\mu}\left(n_{\nu]}/N\right) = 0$, whence

$$\nabla_\mu n_\nu - \nabla_\nu n_\mu = n_\nu \partial_\mu \ln N - n_\mu \partial_\nu \ln N \,. \tag{A.980}$$

Projecting both slots kills the right-hand side, so $\mathcal{K}_{\mu\nu} = \mathcal{K}_{\nu\mu}$.

Equations (A.977) and (A.978). Trace Equation (A.976) with $g^{\mu\nu}$: the second term gives $\epsilon n^\nu a_\nu = 0$ and the first gives $g^{\mu\nu}\mathcal{K}_{\mu\nu} = \gamma^{\mu\nu}\mathcal{K}_{\mu\nu} = \mathcal{K}$, because $\mathcal{K}$ is tangential and g and γ agree on tangential tensors. Squaring Equation (A.976),

$$(\mathcal{K}_\mu{}^\nu + \epsilon n_\mu a^\nu)\,(\mathcal{K}_\nu{}^\mu + \epsilon n_\nu a^\mu) = \mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} \,, \tag{A.981}$$

each cross term carrying an n contracted with a tangential index and the last term carrying $a^\nu n_\nu = 0$.

Equation (A.979). Compare Equation (A.980) with the antisymmetric part of Equation (A.976), which is $\epsilon\,(n_\mu a_\nu - n_\nu a_\mu)$, and contract with n^μ. The left-hand side gives $\epsilon\,(\epsilon a_\nu - 0) = a_\nu$ and the right-hand side gives $n_\nu\,(n^\mu \partial_\mu \ln N) - \epsilon \partial_\nu \ln N$. Projecting with γ, which fixes the tangential a_ν and kills n_ν, leaves Equation (A.979). □

Lemma A.627 (The dictionary between $\mathcal{K}, \gamma$ and K, h)**.** *In the coordinates of Lemma A.623, with K_{ij} the extrinsic curvature Equation (26.79) of Definition 26.40, $K = h^{ij}K_{ij}$, $K^{ij} = h^{ik}h^{jl}K_{kl}$ and ${}^{(3)}R$ the Ricci scalar of h_{ij}:*

$$\mathcal{K}_{ij} = -K_{ij} \,, \qquad \mathcal{K} = K \,, \qquad \mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} = K_{ij}K^{ij} \,, \qquad R[\gamma] = -{}^{(3)}R \,, \tag{A.982}$$

$R[\gamma]$ denoting the curvature scalar Equation (13.307) of the induced metric γ_{ij} of Equation (A.974). The Levi-Civita connections of γ_{ij} and of h_{ij} coincide, so D_i of Equation (A.973) is the D_i of Definition 26.40.

Rests on Equations (26.79) and (A.974) and Definition 13.153.

Proof. *Derives Lemma A.627.*

Connections and Riemann tensors. $\gamma_{ij} = -h_{ij}$ differs from h_{ij} by a constant factor, and the Christoffel symbols Equation (13.299) are homogeneous of degree zero in the metric — one inverse metric against one derivative of the metric — so they are the same for γ and for h. By Equation (13.306) the Riemann tensors $R^i{}_{jkl}$ and hence the Ricci tensors R_{jl} agree. The scalars do not: $R[\gamma] = \gamma^{jl}R_{jl} = -h^{jl}R_{jl} = -{}^{(3)}R$, which is the fourth relation.

Extrinsic curvature. Since $n_i = 0$, ${\gamma^\alpha}_i = \delta^\alpha_i$, so $\mathcal{K}_{ij} = \nabla_i n_j = \partial_i \left(N\delta^0_j\right) - {\Gamma^\lambda}_{ij} n_\lambda = -N{\Gamma^0}_{ij}$. Compute ${\Gamma^0}_{ij}$ from Equation (13.299) with $g_{0i} = -N_i$, $g_{ij} = -h_{ij}$ and Equation (A.967):

$$\begin{aligned} {\Gamma^0}_{ij} &= \tfrac{1}{2} g^{00} \left(\partial_i g_{0j} + \partial_j g_{0i} - \partial_0 g_{ij}\right) + \tfrac{1}{2} g^{0k} \left(\partial_i g_{kj} + \partial_j g_{ki} - \partial_k g_{ij}\right) \\ &= \frac{1}{2N^2} \left[\partial_0 h_{ij} - \partial_i N_j - \partial_j N_i + N^k \left(\partial_i h_{kj} + \partial_j h_{ki} - \partial_k h_{ij}\right)\right] , \end{aligned} \tag{A.983}$$

every term having picked up one overall minus from $g_{ij} = -h_{ij}$, $g_{0i} = -N_i$ and, in the second group, a further minus from $g^{0k} = -N^k/N^2$. Now the Christoffel symbols of h_{ij} give $D_i N_j + D_j N_i = \partial_i N_j + \partial_j N_i - 2{\Gamma^{(3)k}}_{ij} N_k$ and $2{\Gamma^{(3)k}}_{ij} N_k = N^l \left(\partial_i h_{lj} + \partial_j h_{li} - \partial_l h_{ij}\right)$, so the bracket of Equation (A.983) is $\partial_0 h_{ij} - D_i N_j - D_j N_i$ and

$${\Gamma^0}_{ij} = \frac{1}{2N^2} \left(\partial_0 h_{ij} - D_i N_j - D_j N_i\right) = \frac{K_{ij}}{N} , \tag{A.984}$$

by Equation (26.79). Hence $\mathcal{K}_{ij} = -K_{ij}$, the first relation. This is not a choice: it is forced, because $\mathcal{K}$ is one half the Lie derivative of γ along n while K is one half the Lie derivative of h, and $\gamma = -h$.

Trace and square. $\mathcal{K} = \gamma^{ij}\mathcal{K}_{ij} = \left(-h^{ij}\right)\left(-K_{ij}\right) = K$: the two minus signs cancel, so the traces agree even though the tensors differ in sign. Likewise $\mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} = \gamma^{ik}\gamma^{jl}\mathcal{K}_{ij}\mathcal{K}_{kl} = h^{ik}h^{jl}K_{ij}K_{kl} = K_{ij}K^{ij}$, four minus signs cancelling in pairs. Only the curvature scalar, which carries one inverse metric rather than an even number, survives the substitution with its sign changed — and it is exactly that asymmetry which Remark A.633 is about. □

A.52.3 The Gauss equation

Lemma A.628 (Gauss equation, arbitrary signature)**.** *For a hypersurface with unit normal n, $\epsilon = n \cdot n$,*

$$R[\gamma]^\kappa{}_{\nu\rho\mu} = {\gamma^\kappa}_\alpha {\gamma^\tau}_\nu {\gamma^\lambda}_\rho {\gamma^\sigma}_\mu \, R^\alpha{}_{\tau\lambda\sigma} - \epsilon \left(\mathcal{K}_{\rho\nu}\mathcal{K}_\mu{}^\kappa - \mathcal{K}_{\mu\nu}\mathcal{K}_\rho{}^\kappa\right) , \tag{A.985}$$

$R[\gamma]$ being the Riemann tensor of the induced metric.
Rests on Equations (13.305) and (A.973) and Definition A.625.

Proof. *Derives Lemma A.628.*

Let ω_μ be a tangential covector field. By Equation (A.973), $(D\omega)_{\alpha\beta} = {\gamma^\sigma}_\alpha {\gamma^\tau}_\beta \nabla_\sigma \omega_\tau$, and applying D once more,

$$D_\rho D_\mu \omega_\nu = {\gamma^\lambda}_\rho {\gamma^\alpha}_\mu {\gamma^\beta}_\nu \nabla_\lambda \left[{\gamma^\sigma}_\alpha {\gamma^\tau}_\beta \nabla_\sigma \omega_\tau\right] . \tag{A.986}$$

Expand the derivative by the Leibniz rule. Because $\nabla_\lambda {\gamma^\sigma}_\alpha = -\epsilon \left(n^\sigma \nabla_\lambda n_\alpha + n_\alpha \nabla_\lambda n^\sigma\right)$, and because every n_α or n_β meets a projector and dies, the two terms in which ∇ falls on a projector are

$$-\epsilon \, \mathcal{K}_{\rho\mu} \, n^\sigma {\gamma^\tau}_\nu \nabla_\sigma \omega_\tau - \epsilon \, \mathcal{K}_{\rho\nu} \, {\gamma^\sigma}_\mu n^\tau \nabla_\sigma \omega_\tau \, . \tag{A.987}$$

The first is symmetric in $\rho \leftrightarrow \mu$ and therefore drops out of the commutator $D_\rho D_\mu - D_\mu D_\rho$. In the second, differentiate $\omega_\tau n^\tau = 0$ to get $n^\tau \nabla_\sigma \omega_\tau = -\omega_\tau \nabla_\sigma n^\tau$, and then ${\gamma^\sigma}_\mu \omega_\tau \nabla_\sigma n^\tau = \omega_\tau \mathcal{K}_\mu{}^\tau$ because ω is tangential; so the second term equals $+\epsilon \, \mathcal{K}_{\rho\nu}\mathcal{K}_\mu{}^\tau \omega_\tau$.

The remaining term of Equation (A.986) is ${\gamma^\lambda}_\rho {\gamma^\sigma}_\mu {\gamma^\tau}_\nu \nabla_\lambda \nabla_\sigma \omega_\tau$. Antisymmetrizing in $\rho \leftrightarrow \mu$ and using the Ricci identity Equation (13.305) in its covector form, $[\nabla_\lambda, \nabla_\sigma]\,\omega_\tau = -R^\kappa{}_{\tau\lambda\sigma}\omega_\kappa$ (which follows from Equation (13.305) by applying the commutator to the scalar $\omega_\lambda V^\lambda$), and the same identity on the leaf for D,

$$-R[\gamma]^\kappa{}_{\nu\rho\mu}\omega_\kappa = \epsilon \left(\mathcal{K}_{\rho\nu}\mathcal{K}_\mu{}^\tau - \mathcal{K}_{\mu\nu}\mathcal{K}_\rho{}^\tau\right)\omega_\tau - {\gamma^\lambda}_\rho {\gamma^\sigma}_\mu {\gamma^\tau}_\nu R^\kappa{}_{\tau\lambda\sigma}\omega_\kappa \, . \tag{A.988}$$

Both sides are tangential in the free index κ and ω was an arbitrary tangential covector, so the coefficients agree after projecting κ, which is Equation (A.985). □

Remark A.629 (A check at $\epsilon = +1$ that uses no relativity). The sign of the quadratic term in Equation (A.985) can be checked on the one case every reader already knows, and in the signature that matters here. Take the sphere of radius r in Euclidean $\mathbb{R}^3$: the ambient is flat, $\epsilon = +1$ because the normal is a unit vector of a positive-definite metric, and $\mathcal{K}_{ij} = \nabla_i n_j = \gamma_{ij}/r$ for the outward normal. Then Equation (A.985), with indices lowered, gives $R[\gamma]_{\kappa\nu\rho\mu} = r^{-2}\left(\gamma_{\mu\nu}\gamma_{\rho\kappa} - \gamma_{\rho\nu}\gamma_{\mu\kappa}\right)$, whose double contraction is $R[\gamma] = 2/r^2$. That is 2 times the Gaussian curvature, which is what Definition 13.153 says a two-dimensional curvature scalar must be, and it is Gauss's *theorema egregium*: the sphere's intrinsic curvature is fixed by its second fundamental form alone. A sign error in Equation (A.985) would make the sphere intrinsically hyperbolic.

A.52.4 The Codazzi equation

Lemma A.630 (Codazzi equation)**.** *With the same notation,*

$$D_\mu \mathcal{K}_{\nu\rho} - D_\nu \mathcal{K}_{\mu\rho} = -\gamma^\alpha{}_\mu \gamma^\beta{}_\nu \gamma^\lambda{}_\rho \, n_\kappa R^\kappa{}_{\lambda\alpha\beta} \,, \tag{A.989}$$

and contracting ν with ρ,

$$D_\nu \left(\mathcal{K}^\nu{}_\mu - \mathcal{K}\gamma^\nu{}_\mu\right) = \gamma^\alpha{}_\mu R_{\alpha\kappa} n^\kappa \,. \tag{A.990}$$

No factor of ϵ appears in either.

Rests on Lemma A.626, Equation (13.305), and Definition A.625.

Proof. *Derives Lemma A.630.*

By Equation (A.976), $\mathcal{K}_{\beta\lambda} = \nabla_\beta n_\lambda - \epsilon n_\beta a_\lambda$, so

$$D_\mu \mathcal{K}_{\nu\rho} = \gamma^\alpha{}_\mu \gamma^\beta{}_\nu \gamma^\lambda{}_\rho \nabla_\alpha \nabla_\beta n_\lambda - \epsilon \, a_\rho \mathcal{K}_{\mu\nu} \,, \tag{A.991}$$

the second term because $\gamma^\beta{}_\nu \nabla_\alpha \left(n_\beta a_\lambda\right) = a_\lambda \gamma^\beta{}_\nu \nabla_\alpha n_\beta$, the other piece dying on $\gamma^\beta{}_\nu n_\beta = 0$. Antisymmetrizing in $\mu \leftrightarrow \nu$, the term $a_\rho \mathcal{K}_{\mu\nu}$ is symmetric and drops, and the Ricci identity in covector form leaves Equation (A.989).

For the contraction, apply $\gamma^{\nu\rho}$. On the left it gives $D_\mu \mathcal{K} - D_\nu \mathcal{K}^\nu{}_\mu$, which is minus the left-hand side of Equation (A.990). On the right, $\gamma^{\beta\lambda} = g^{\beta\lambda} - \epsilon n^\beta n^\lambda$; the $g^{\beta\lambda}$ piece contracts the second and fourth slots of $R_{\kappa\lambda\alpha\beta}$, which by the antisymmetry within each index pair (Proposition 13.154) is the same as contracting the first and third, that is again the Ricci tensor, giving $n^\kappa R_{\kappa\alpha}$, while the $n^\beta n^\lambda$ piece vanishes because $R_{\kappa\lambda\alpha\beta} n^\kappa n^\lambda = 0$ by antisymmetry in the first pair. Rearranging gives Equation (A.990). □

Equation (A.990) is worth reading before it is used. Under the dictionary Equation (A.982) the mixed-index extrinsic curvature carries two inverse metrics' worth of sign and is unchanged, $\mathcal{K}^j{}_i = \gamma^{jk}\mathcal{K}_{ki} = \left(-h^{jk}\right)\left(-K_{ki}\right) = K^j{}_i$, so the left-hand side is $D_j K^j{}_i - D_i K$. By Equation (26.82) that is $2\kappa h^{-1/2} D_j \pi^j{}_i = -\kappa h^{-1/2} \mathcal{H}_i$ with $\mathcal{H}_i$ the momentum constraint Equation (26.85); and the right-hand side is a projection of the Ricci tensor onto one normal and one tangential index. So the momentum constraint of Theorem 26.42 is the contracted Codazzi equation, which is the precise content of the first paragraph of Remark 26.44: the constraints are the $G^0{}_\mu$ components of the field equations, and they are conditions on data laid down on one leaf because Codazzi's equation contains no derivative off the leaf.

A.52.5 The twice-normal contraction

The double contraction of Equation (A.985) will leave behind the term $R_{\mu\nu} n^\mu n^\nu$, which is neither intrinsic nor a projection of anything on the leaf. Converting it is the step that produces the total derivatives, and it is the one usually passed over. It is not passed over here.

Lemma A.631 (The normal–normal Ricci contraction)**.** *For any unit normal field,*

$$R_{\mu\nu}n^\mu n^\nu = \nabla_\mu\left(a^\mu - \mathcal{K}n^\mu\right) - \left(\mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} - \mathcal{K}^2\right) . \tag{A.992}$$

Rests on Equation (13.305), Lemma A.626, and Definition 13.153.

Proof. *Derives Lemma A.631.*

Apply the Ricci identity Equation (13.305) to n^λ, set $\lambda = \mu$ and contract: since $R^\mu{}_{\rho\mu\nu} = R_{\rho\nu}$ by Equation (13.307),

$$\nabla_\mu\nabla_\nu n^\mu - \nabla_\nu\nabla_\mu n^\mu = R_{\rho\nu}n^\rho . \tag{A.993}$$

Contract with n^ν and rewrite each side as a divergence minus the term the Leibniz rule left over:

$$n^\nu\nabla_\mu\nabla_\nu n^\mu = \nabla_\mu\left(n^\nu\nabla_\nu n^\mu\right) - \left(\nabla_\mu n^\nu\right)\left(\nabla_\nu n^\mu\right) , \tag{A.994}$$

$$n^\nu\nabla_\nu\left(\nabla_\mu n^\mu\right) = \nabla_\nu\left(n^\nu\nabla_\mu n^\mu\right) - \left(\nabla_\nu n^\nu\right)\left(\nabla_\mu n^\mu\right) . \tag{A.995}$$

Subtracting, and inserting Equations (A.977) and (A.978) for the two quadratic terms and $a^\mu = n^\nu\nabla_\nu n^\mu$ for the first divergence, gives Equation (A.992). □

Equation (A.992) is the trace of what is called the Ricci, or Mainardi, equation — the twice-normal projection of the Riemann tensor. Only the trace is needed below, and only the trace is proved: the untraced equation would be an identity for $\gamma^\alpha{}_\mu\gamma^\beta{}_\nu n^\lambda n^\sigma R_{\alpha\lambda\beta\sigma}$ in terms of the Lie derivative of $\mathcal{K}$ along n, and nothing in this appendix or in Chapter 26 consumes it.

A.52.6 Assembling the Lagrangian

Theorem A.632 (The 3+1 identity for the curvature scalar)**.** *For a foliation by hypersurfaces with unit normal n and $\epsilon = n\cdot n$,*

$$R = R[\gamma] - \epsilon\left(\mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} - \mathcal{K}^2\right) + 2\epsilon\,\nabla_\mu\left(a^\mu - \mathcal{K}n^\mu\right) . \tag{A.996}$$

Specialized to Equation (26.78) by Lemmas A.623 and A.627, that is $\epsilon = +1$, $\gamma_{ij} = -h_{ij}$,

$$\sqrt{|g|}\,R = -N\sqrt{h}\left({}^{(3)}R + K_{ij}K^{ij} - K^2\right) + 2\,\partial_\mu\left[\sqrt{|g|}\left(a^\mu - Kn^\mu\right)\right] , \tag{A.997}$$

with $\sqrt{|g|} = N\sqrt{h}$. In particular ${}^{(3)}R$ contains h_{ij} and its spatial derivatives only, and K_{ij} contains $\partial_0 h_{ij}$ linearly and no other time derivative, so neither $\partial_0 N$ nor $\partial_0 N^i$ occurs anywhere on the right.

Rests on Lemmas A.627, A.628 and A.631.

Proof. *Derives Theorem A.632.*

Step 1: contract the Gauss equation once. Put $\rho = \kappa$ in Equation (A.985) and sum. On the right, $\gamma^\kappa{}_\alpha\gamma^\lambda{}_\kappa = \gamma^\lambda{}_\alpha$, so, lowering the free index,

$$R[\gamma]_{\nu\mu} = \gamma^{\lambda\alpha}\gamma^\tau{}_\nu\gamma^\sigma{}_\mu R_{\alpha\tau\lambda\sigma} - \epsilon\left(\mathcal{K}_\mu{}^\rho\mathcal{K}_{\rho\nu} - \mathcal{K}\mathcal{K}_{\mu\nu}\right) . \tag{A.998}$$

Writing $\gamma^{\lambda\alpha} = g^{\lambda\alpha} - \epsilon n^\lambda n^\alpha$ and using $g^{\lambda\alpha}R_{\alpha\tau\lambda\sigma} = R_{\tau\sigma}$ from Equation (13.307),

$$R[\gamma]_{\nu\mu} = \gamma^\tau{}_\nu\gamma^\sigma{}_\mu\left[R_{\tau\sigma} - \epsilon\, n^\alpha n^\lambda R_{\alpha\tau\lambda\sigma}\right] - \epsilon\left(\mathcal{K}_\mu{}^\rho\mathcal{K}_{\rho\nu} - \mathcal{K}\mathcal{K}_{\mu\nu}\right) . \tag{A.999}$$

Step 2: contract again. Apply $\gamma^{\nu\mu}$. The first bracket gives $\gamma^{\tau\sigma}R_{\tau\sigma} = R - \epsilon R_{\tau\sigma}n^\tau n^\sigma$. The second gives $\gamma^{\tau\sigma}n^\alpha n^\lambda R_{\alpha\tau\lambda\sigma} = R_{\alpha\lambda}n^\alpha n^\lambda$: the $g^{\tau\sigma}$ part is again a Ricci contraction on

the second and fourth slots, and the $n^\tau n^\sigma$ part vanishes because $R_{\alpha\tau\lambda\sigma}n^\alpha n^\tau = 0$. The quadratic terms give $\mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} - \mathcal{K}^2$. Hence

$$R[\gamma] = R - 2\epsilon\, R_{\mu\nu}n^\mu n^\nu - \epsilon\left(\mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} - \mathcal{K}^2\right) . \tag{A.1000}$$

Step 3: eliminate the normal–normal Ricci term. Substitute Equation (A.992) into Equation (A.1000) and solve for R:

$$\begin{aligned} R &= R[\gamma] + \epsilon\left(\mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} - \mathcal{K}^2\right) + 2\epsilon\, R_{\mu\nu}n^\mu n^\nu \\ &= R[\gamma] + \epsilon\left(\mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} - \mathcal{K}^2\right) + 2\epsilon\left[\nabla_\mu\left(a^\mu - \mathcal{K}n^\mu\right) - \left(\mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} - \mathcal{K}^2\right)\right] , \end{aligned} \tag{A.1001}$$

which is Equation (A.996): the quadratic term appears twice with opposite weight and what survives is $-\epsilon\left(\mathcal{K}\mathcal{K} - \mathcal{K}^2\right)$. It is worth noticing that the sign of the quadratic term in the final answer is opposite to the sign it carries in the Gauss equation, and that the flip is entirely the work of Lemma A.631; this is the step at which a derivation that quotes the Ricci equation instead of proving it can go wrong without leaving a trace.

Step 4: substitute. By Lemma A.623, $\epsilon = +1$; by Lemma A.627, $R[\gamma] = -{}^{(3)}R$, $\mathcal{K}_{\mu\nu}\mathcal{K}^{\mu\nu} = K_{ij}K^{ij}$ and $\mathcal{K} = K$. So

$$R = -{}^{(3)}R - \left(K_{ij}K^{ij} - K^2\right) + 2\nabla_\mu\left(a^\mu - Kn^\mu\right) . \tag{A.1002}$$

Multiplying by $\sqrt{|g|} = N\sqrt{h}$ and using the standard identity $\sqrt{|g|}\,\nabla_\mu V^\mu = \partial_\mu\left(\sqrt{|g|}V^\mu\right)$ for the divergence of a vector field — which follows from $\Gamma^\mu{}_{\mu\nu} = \partial_\nu \ln\sqrt{|g|}$, itself a contraction of Equation (13.299) — gives Equation (A.997).

Step 5: the velocities. ${}^{(3)}R$ is built from h_{ij} and its spatial derivatives alone. K_{ij} is given by Equation (26.79), in which ∂_0 acts on h_{ij} and on nothing else: N and N^i enter algebraically and through *spatial* derivatives D_iN_j only. The divergence term contains ∂_0 of $\sqrt{|g|}\left(a^0 - Kn^0\right)$, hence of N, N^i and h_{ij}; but it is a total derivative and is removed by the boundary term of Appendix A.52.7, so no time derivative of the lapse or the shift survives in the bulk Lagrangian. □

Remark A.633 (The overall sign, settled). Equation (A.997) carries an overall minus sign that Equation (26.80) does not display, and it is worth saying plainly where it comes from and what it costs, because this is exactly the bookkeeping the chapter's prose defers to this appendix.

The source is isolated in Lemma A.627 and is a single asymmetry: of the four quantities in the dictionary, three are built with an even number of inverse metrics and are blind to $\gamma_{ij} = -h_{ij}$, while the curvature *scalar* carries exactly one and changes sign. So with the curvature conventions of Theorem 13.152 and Definition 13.153 — the conventions for which the round two-sphere has $R = 2/r^2$, checked in Remark A.629 — and the mostly-minus signature of Chapter 39, the four-scalar that equals $N\sqrt{h}\left({}^{(3)}R + K_{ij}K^{ij} - K^2\right)$ up to divergences is $-\sqrt{|g|}\,R$, and not $+\sqrt{|g|}\,R$.

Nothing physical turns on this, and three things should be said about it. First, the gravitational Lagrangian density in this signature is therefore $-\sqrt{|g|}\,R/2\kappa$; that is the sign carried, for precisely this reason, by every mostly-minus treatment of general relativity, and it is the combination whose ADM form is $+N\sqrt{h}\left({}^{(3)}R + K_{ij}K^{ij} - K^2\right)/2\kappa$. Second, that is exactly the Lagrangian L used in the proof of Theorem 26.42 — read the first display of that proof — so the momentum Equation (26.82), the constraint densities Equations (26.84) and (26.85) and the degree-of-freedom count all stand unchanged. Third, and this is why Proposition 26.41 can be used as it is written, neither of the two facts the chapter draws from it is touched: $\sqrt{|g|} = N\sqrt{h}$ is Equation (A.969), and the absence of $\partial_0 N$ and $\partial_0 N^i$

is Step 5 above, and both are independent of the overall sign. A reader working in a mostly-plus convention sets $\epsilon = -1$ and $\gamma_{ij} = +h_{ij}$ in Equation (A.996) and recovers the familiar $R = {}^{(3)}\!R + K_{ij}K^{ij} - K^2 - 2\nabla_\mu\left(a^\mu - Kn^\mu\right)$ with no minus sign in front, which is the form printed in the standard references [1221, 1509, 1746] and the reason the sign is so easy to import unexamined.

A.52.7 The boundary term

The divergence in Equation (A.997) is a genuine coordinate divergence, so over a region $\mathcal{R}$ bounded by two leaves it contributes only a surface integral. Evaluate it. The acceleration is tangential, so $n_\mu a^\mu = Na^0 = 0$ and $a^0 = 0$: the acceleration has no flux through a leaf. And $\sqrt{|g|}\,n^0 = N\sqrt{h}\cdot N^{-1} = \sqrt{h}$ by Equations (A.968) and (A.969). Hence the $\mu = 0$ component of the divergence's argument is

$$\sqrt{|g|}\left(a^0 - Kn^0\right) = -\sqrt{h}\,K\,, \tag{A.1003}$$

and, integrating Equation (A.997) over $\mathcal{R}$,

$$\int_{\mathcal{R}} \mathrm{d}^4x\left(-\sqrt{|g|}\,R\right) = \int_{\mathcal{R}} \mathrm{d}^4x\,N\sqrt{h}\left({}^{(3)}\!R + K_{ij}K^{ij} - K^2\right) + 2\oint_{\partial\mathcal{R}} \mathrm{d}^3x\,\sqrt{h}\,K\,, \tag{A.1004}$$

the sign of the last term following the outward orientation of $\partial\mathcal{R}$ once that orientation is fixed. Dividing by 2κ, the surface term is $\kappa^{-1}\oint\sqrt{h}\,K$, which is the Gibbons–Hawking–York boundary action [1490, 1596].

Two statements are made about it here and only the first is proved. The first is Equation (A.1004) itself: the bulk ADM Lagrangian and the gravitational action differ by exactly that surface integral, which is why Proposition 26.41 may write "plus total derivatives" and Theorem 26.42 may drop them. The second is quoted: adding the Gibbons–Hawking–York term to the action makes the variational problem well posed with h_{ij} — and only h_{ij} — held fixed on $\partial\mathcal{R}$, because the unwanted normal derivatives of the metric variation that $\sqrt{|g|}\,R$ produces on the boundary are precisely what it cancels. That is proved in [1596] and not here; the constrained-dynamics consequence of the same term, that the numerical value of the Hamiltonian of an isolated system is carried entirely by a surface integral, is Remark 26.45.

Remark A.634 (What is quoted here). Two imports, and one debt of Chapter 13.

Quoted. That the Gibbons–Hawking–York term makes the variational problem well posed, as stated in the paragraph above; it is used nowhere in Chapter 26 except as the justification for discarding total derivatives, which Equation (A.1004) establishes independently.

Owed by Chapter 13. A general-signature Gauss–Codazzi theory of hypersurfaces. That chapter carries the classical version Equation (13.92), but it is the theory of a surface in *flat Euclidean* $\mathbb{R}^3$, built from a position vector $\boldsymbol{x}$ and its derivatives, and it is not usable for a spacelike leaf of a Lorentzian spacetime: there is no position vector, the ambient is curved, and the normal's square is ϵ rather than $+1$. The chapter's later treatment of a quadric in a flat ambient of signature (p, q) carries the right sign structure but is again an embedding into a *flat* space given by an explicit position vector. Lemmas A.628, A.630 and A.631 are therefore proved here, from Theorem 13.152 and the projector alone, and the general statement belongs in the manifolds chapter. This is recorded as a Part II debt, in the same terms as Remarks 26.4 and 26.18 record theirs.

Everything else — the inverse metric, the determinant, the extrinsic curvature in coordinates, the dictionary, the two contractions of the Gauss equation and the elimination of the normal–normal Ricci term — is carried out above from the Levi-Civita connection of Theorem 13.150 and the Riemann tensor of Theorem 13.152.

Remark A.635. Appendix A.52 discharges the derivation owed at Proposition 26.41 of Section 26.6.5 in Chapter 26. What the chapter takes from it is narrow and is now established: $\sqrt{|g|} = N\sqrt{h}$ (Equation (A.969)), and that the Einstein–Hilbert Lagrangian contains no time derivative of the lapse or the shift (Step 5 of Theorem A.632). Those two facts are what make π_N and π_i vanish identically in Equation (26.81), hence what make N and N^i Lagrange multipliers rather than dynamical fields, hence what make $\mathcal{H}_\perp$ and $\mathcal{H}_i$ secondary constraints by outcome 2 of Proposition 26.10 — so the whole of Theorem 26.42, and with it the count of two degrees of freedom per point, rests on this section. The overall sign of Equation (26.80) is settled in Remark A.633 and changes none of that. The brackets of the constraints so obtained are computed in Appendix A.53.

A.53 The Hypersurface-Deformation Algebra

This appendix proves Proposition 26.43 of Chapter 26: the three brackets Equations (26.89) to (26.91) of the smeared constraints of Theorem 26.42. With them the eight constraints of general relativity are shown to be first class in the sense of Definition 26.12, so that Equation (26.21) applies and the count of two propagating degrees of freedom per point of space is established; and the last of the three exhibits, on the right-hand side, the inverse spatial metric h^{ij} — a function on phase space where a Lie algebra would carry a constant.

Everything is computed with *smeared* constraints. That is not a matter of taste. The unsmeared densities have brackets proportional to $\delta^3(\boldsymbol{x}-\boldsymbol{y})$ and to its first and second derivatives, and Equation (26.91) in unsmeared form is an identity between distributions in which the two sides must be compared after integrating twice by parts against test functions anyway. Smearing does that integration once and for all, at the start; every step below is then an ordinary integration by parts on smooth compactly supported data.

Two things are assumed rather than proved and are named here. The first is the constraint densities themselves, Equations (26.84) and (26.85), which Theorem 26.42 derives; nothing below depends on the overall sign discussed in Remark A.633, because the brackets are computed from the densities as the chapter displays them. The second is that all fields fall off fast enough for the surface terms of the integrations by parts to vanish — on a region with a boundary they do not, and what they contribute is exactly the ADM charges of Remark 26.45.

A.53.1 Conventions: weights, smearing functions, and the bracket

Notation A.636 (Densities and smearing). The canonical pair is (h_{ij}, π^{ij}) of Equation (26.82), with the field bracket Equation (26.47) read as

$$\{F,G\} = \int \mathrm{d}^3x \left(\frac{\delta F}{\delta h_{ij}(\boldsymbol{x})} \frac{\delta G}{\delta \pi^{ij}(\boldsymbol{x})} - \frac{\delta F}{\delta \pi^{ij}(\boldsymbol{x})} \frac{\delta G}{\delta h_{ij}(\boldsymbol{x})} \right) , \tag{A.1005}$$

functional derivatives with respect to h_{ij} and π^{ij} being taken with both regarded as symmetric. The *weights* are: h_{ij} is a tensor of weight 0; π^{ij} is a tensor *density* of weight 1, as Equation (26.82) shows through its factor $\sqrt{h}$; $\mathcal{H}_\perp$ of Equation (26.84) is a scalar density of weight 1, its kinetic part carrying $\pi\pi/\sqrt{h}$ and its curvature part $\sqrt{h}\,{}^{(3)}R$; and $\mathcal{H}_i$ of Equation (26.85) is a covector density of weight 1. The smearing functions f, g are scalars and ξ^i, ζ^i vectors, all of weight 0, all smooth, all of compact support, and — this is used constantly and is easy to forget — all *independent of the canonical variables*, so that they pass through every functional derivative untouched. Finally

$$H_\perp[f] = \int \mathrm{d}^3x\, f\, \mathcal{H}_\perp \,, \qquad H_\parallel[\xi] = \int \mathrm{d}^3x\, \xi^i \mathcal{H}_i \,, \tag{A.1006}$$

both of which are ordinary numbers because a weight-one density integrates without a metric factor.

Rests on Notation 26.28, Equation (26.82), and Theorem 26.42.

Getting a weight wrong is the commonest way this computation fails, so the one fact used about them is stated once and checked: for a weight-one vector density V^i,

$$D_i V^i = \partial_i V^i \,, \qquad \text{hence} \qquad \int \mathrm{d}^3x \; D_i V^i = 0 \,, \tag{A.1007}$$

for V of compact support. This is immediate from the covariant derivative of a density, $D_i V^i = \partial_i V^i + \Gamma^i{}_{ik} V^k - \Gamma^k{}_{ki} V^i$, the last term being the weight term; the two cancel. Every integration by parts below is an application of Equation (A.1007). For a weight-*zero* vector the corresponding statement carries a $\sqrt{h}$, $\sqrt{h} D_i V^i = \partial_i \left(\sqrt{h} V^i \right)$, and that form is used where the integrand carries an explicit $\sqrt{h}$.

The sign convention for the commutator of vector fields is

$$[\xi, \zeta]^i = \xi^j \partial_j \zeta^i - \zeta^j \partial_j \xi^i = (\mathcal{L}_\xi \zeta)^i \,, \tag{A.1008}$$

which is what makes the right-hand side of Equation (26.89) carry a plus sign rather than a minus.

A.53.2 The momentum constraint generates the spatial drag

Lemma A.637 (Closed form of $H_\parallel$, and what it generates)**.** *For every smearing vector ξ,*

$$H_\parallel[\xi] = \int \mathrm{d}^3x \; \pi^{ij} \, \mathcal{L}_\xi h_{ij} \,, \qquad \mathcal{L}_\xi h_{ij} = D_i \xi_j + D_j \xi_i \,, \tag{A.1009}$$

and consequently

$$\left\{ h_{ij}, H_\parallel[\xi] \right\} = \mathcal{L}_\xi h_{ij} \,, \qquad \left\{ \pi^{ij}, H_\parallel[\xi] \right\} = \mathcal{L}_\xi \pi^{ij} \,, \tag{A.1010}$$

the second Lie derivative being that of a weight-one contravariant density, $\mathcal{L}_\xi \pi^{ij} = \xi^k \partial_k \pi^{ij} - \pi^{kj} \partial_k \xi^i - \pi^{ik} \partial_k \xi^j + \pi^{ij} \partial_k \xi^k$.

Rests on Equations (26.85) and (A.1005) and Notation A.636.

Proof. *Derives Lemma A.637.*

Equation (A.1009). By Equation (26.85), $H_\parallel[\xi] = -2 \int \mathrm{d}^3x \; \xi^i D_j \pi^j{}_i = -2 \int \mathrm{d}^3x \; \xi_k D_j \pi^{jk}$, the index having been moved through D because $Dh = 0$. The quantity $\xi_k \pi^{jk}$ is a weight-one vector density, so Equation (A.1007) gives $\int D_j \left(\xi_k \pi^{jk} \right) = 0$, whence

$$H_\parallel[\xi] = 2 \int \mathrm{d}^3x \; (D_j \xi_k) \, \pi^{jk} = \int \mathrm{d}^3x \; \pi^{jk} \, (D_j \xi_k + D_k \xi_j) \,, \tag{A.1011}$$

the symmetrization being free because π^{jk} is symmetric. That $D_j \xi_k + D_k \xi_j = \mathcal{L}_\xi h_{jk}$ is Equation (13.275).

First of Equation (A.1010). Take $F = h_{ij}(\boldsymbol{x})$ in Equation (A.1005): only the first term survives, and $\{h_{ij}(\boldsymbol{x}), G\} = \delta G / \delta \pi^{ij}(\boldsymbol{x})$. Applied to Equation (A.1009), in which π appears linearly and undifferentiated, this is $\mathcal{L}_\xi h_{ij}$.

Second. Likewise $\{\pi^{ij}(\boldsymbol{x}), G\} = -\delta G / \delta h_{ij}(\boldsymbol{x})$. Write out Equation (A.1009) in partial derivatives, $\mathcal{L}_\xi h_{ab} = \xi^k \partial_k h_{ab} + h_{kb} \partial_a \xi^k + h_{ak} \partial_b \xi^k$, and vary. The first term contributes, after one integration by parts, $-\partial_k \left(\xi^k \pi^{ij} \right)$; the second contributes $\pi^{kj} \partial_k \xi^i$ and the third $\pi^{ik} \partial_k \xi^j$. Hence

$$\frac{\delta H_\parallel[\xi]}{\delta h_{ij}} = -\xi^k \partial_k \pi^{ij} - \pi^{ij} \partial_k \xi^k + \pi^{kj} \partial_k \xi^i + \pi^{ik} \partial_k \xi^j = -\mathcal{L}_\xi \pi^{ij} \,, \tag{A.1012}$$

and the minus of the bracket cancels it. Note where the weight shows itself: the term $\pi^{ij}\partial_k\xi^k$, which distinguishes the Lie derivative of a density from that of a tensor, comes from the integration by parts in the first term and from nowhere else. Had π^{ij} been treated as a weight-zero tensor, the density term would be spurious and Equation (26.89) would fail. □

A.53.3 The first two brackets

Both follow from Lemma A.637 and one observation, which is worth isolating because it does the work twice.

Lemma A.638 (Naturality)**.** *Let Φ be a functional of the canonical variables and of one smearing field σ (a scalar f or a vector ξ),*

$$\Phi[\sigma; h, \pi] = \int \mathrm{d}^3x\, \mathcal{F}(\sigma, h, \pi)\ , \tag{A.1013}$$

with $\mathcal{F}$ a scalar density of weight one built covariantly from its arguments. Then

$$\Big\{\Phi[\sigma], H_\parallel[\zeta]\Big\} = -\Phi[\mathcal{L}_\zeta\sigma]\ . \tag{A.1014}$$

Rests on Lemma A.637 and Equation (A.1005)*.*

Proof. *Derives Lemma A.638.*

By Equation (A.1005) and Equation (A.1010),

$$\Big\{\Phi, H_\parallel[\zeta]\Big\} = \int \mathrm{d}^3x \left(\frac{\delta\Phi}{\delta h_{ij}}\mathcal{L}_\zeta h_{ij} + \frac{\delta\Phi}{\delta\pi^{ij}}\mathcal{L}_\zeta\pi^{ij}\right) = \delta_\zeta\Phi\,, \tag{A.1015}$$

the change of Φ when the canonical fields alone are dragged by ζ. But Φ is invariant when *everything* is dragged: if all of σ, h, π are transported, $\mathcal{F}$ changes by $\mathcal{L}_\zeta\mathcal{F} = \partial_k\left(\zeta^k\mathcal{F}\right)$, because $\mathcal{F}$ is a weight-one scalar density, and that integrates to zero. Hence dragging the fields alone gives minus the result of dragging σ alone, which is Equation (A.1014). □

Proposition A.639 (The first two brackets)**.** *Equations* (26.89) *and* (26.90) *hold:*

$$\Big\{H_\parallel[\xi], H_\parallel[\zeta]\Big\} = H_\parallel[[\xi,\zeta]]\ , \qquad \Big\{H_\parallel[\xi], H_\perp[f]\Big\} = H_\perp\Big[\xi^i\partial_i f\Big]\ . \tag{A.1016}$$

Rests on Lemma A.638 and Equations (26.84) *and* (A.1008)*.*

Proof. *Derives Proposition A.639.*

Both integrands satisfy the hypothesis of Lemma A.638: $\xi^i\mathcal{H}_i$ and $f\mathcal{H}_\perp$ are weight-one scalar densities built covariantly from the smearing field and the canonical pair, by Notation A.636.

For the first, apply Equation (A.1014) with $\Phi = H_\parallel$, $\sigma = \xi$: $\Big\{H_\parallel[\xi], H_\parallel[\zeta]\Big\} = -H_\parallel[\mathcal{L}_\zeta\xi] = -H_\parallel[[\zeta,\xi]] = H_\parallel[[\xi,\zeta]]$ by Equation (A.1008). Nothing about the detailed form of $\mathcal{H}_i$ was used beyond Lemma A.637; the identity is $[\mathcal{L}_\xi, \mathcal{L}_\zeta] = \mathcal{L}_{[\xi,\zeta]}$ in disguise.

For the second, apply it with $\Phi = H_\perp$, $\sigma = f$ a scalar, so $\mathcal{L}_\xi f = \xi^i\partial_i f$: $\Big\{H_\perp[f], H_\parallel[\xi]\Big\} = -H_\perp[\xi^i\partial_i f]$, and the antisymmetry of the bracket gives the stated form. The content is that $\mathcal{H}_\perp$ is a scalar density of weight one — that dragging $f\mathcal{H}_\perp$ produces $\mathcal{L}_\xi\left(f\mathcal{H}_\perp\right)$, whose total-divergence part integrates away and whose remainder is $\left(\xi^i\partial_i f\right)\mathcal{H}_\perp$. □

A.53.4 The variation of the intrinsic curvature

The third bracket needs the piece a reader cannot supply from the chapter: how $\int \sqrt{h}\,{}^{(3)}R$ responds to a variation of h_{ij}.

Lemma A.640 (Variation of the integrated three-curvature)**.** *For any smooth f of compact support, independent of h_{ij},*

$$\frac{\delta}{\delta h_{ij}} \int \mathrm{d}^3y\, f\sqrt{h}\,{}^{(3)}R = -f\sqrt{h}\,{}^{(3)}G^{ij} + \sqrt{h}\left(D^iD^j - h^{ij}D^2\right) f\,, \tag{A.1017}$$

with ${}^{(3)}G^{ij} = {}^{(3)}R^{ij} - \frac{1}{2}h^{ij}\,{}^{(3)}R$ the three-dimensional Einstein tensor and $D^2 = h^{ij}D_iD_j$.

Rests on Definition 13.153, Equation (13.299), and Theorem 13.152.

Proof. *Derives Lemma A.640.*

Three pieces. First, $\delta\sqrt{h} = \frac{1}{2}\sqrt{h}\,h^{ij}\delta h_{ij}$, from $\delta \ln\det h = \mathrm{tr}\,(h^{-1}\delta h)$. Second, $\delta h^{ij} = -h^{ik}h^{jl}\delta h_{kl}$, from varying $h^{ik}h_{kj} = \delta^i_j$; so $\delta\left({}^{(3)}R\right) = {}^{(3)}R_{ij}\delta h^{ij} + h^{ij}\delta\,{}^{(3)}R_{ij} = -{}^{(3)}R^{ij}\delta h_{ij} + h^{ij}\delta\,{}^{(3)}R_{ij}$.

Third, the term $h^{ij}\delta\,{}^{(3)}R_{ij}$. The variation of a Christoffel symbol is a tensor, and from Equation (13.299) with vanishing torsion,

$$\delta\Gamma^k{}_{ij} = \tfrac{1}{2}h^{kl}\left(D_i\delta h_{lj} + D_j\delta h_{li} - D_l\delta h_{ij}\right)\,, \tag{A.1018}$$

which one checks by evaluating both sides in normal coordinates at a point, where $\Gamma = 0$ and $D = \partial$. Then Equation (13.306) gives $\delta\,{}^{(3)}R_{ij} = D_k\delta\Gamma^k{}_{ij} - D_i\delta\Gamma^k{}_{kj}$, so $h^{ij}\delta\,{}^{(3)}R_{ij} = D_kv^k$ with $v^k = h^{ij}\delta\Gamma^k{}_{ij} - h^{kj}\delta\Gamma^i{}_{ij}$. Evaluate the two pieces from Equation (A.1018). In the second, the first and third terms of the bracket cancel on contraction, leaving $\delta\Gamma^i{}_{ij} = \frac{1}{2}D_j\left(h^{il}\delta h_{il}\right)$. In the first, the two symmetric terms combine, leaving $h^{ij}\delta\Gamma^k{}_{ij} = h^{kl}D^j\delta h_{lj} - \frac{1}{2}D^k\left(h^{ij}\delta h_{ij}\right)$. Hence

$$v^k = h^{kl}D^j\delta h_{lj} - D^k\left(h^{ij}\delta h_{ij}\right)\,, \qquad D_kv^k = \left(D^iD^j - h^{ij}D^2\right)\delta h_{ij}\,. \tag{A.1019}$$

Collecting,

$$\delta\left(\sqrt{h}\,{}^{(3)}R\right) = -\sqrt{h}\,{}^{(3)}G^{ij}\delta h_{ij} + \sqrt{h}\left(D^iD^j - h^{ij}D^2\right)\delta h_{ij}\,, \tag{A.1020}$$

the first term assembling $\frac{1}{2}h^{ij}\,{}^{(3)}R - {}^{(3)}R^{ij}$. Multiply by f, integrate, and move the two derivatives of the second term onto f using $\sqrt{h}\,D_kV^k = \partial_k\left(\sqrt{h}V^k\right)$ twice. The result is Equation (A.1017); note that D^iD^jf is automatically symmetric, f being a scalar. □

A.53.5 The third bracket

Theorem A.641 (The normal–normal bracket)**.** *Equation (26.91) holds:*

$$\{H_\perp[f], H_\perp[g]\} = H_\parallel\left[h^{ij}\left(f\partial_jg - g\partial_jf\right)\right]\,. \tag{A.1021}$$

Rests on Lemma A.640 and Equations (26.84) and (26.85).

Proof. *Derives Theorem A.641.*

Step 1: the two functional derivatives. From Equation (26.84), in which π occurs only algebraically,

$$\frac{\delta H_\perp[f]}{\delta\pi^{ij}} = f\,B_{ij}\,, \qquad B_{ij} := \frac{4\kappa}{\sqrt{h}}\left(\pi_{ij} - \tfrac{1}{2}\pi h_{ij}\right)\,, \tag{A.1022}$$

since $\delta\left(\pi_{kl}\pi^{kl}\right)/\delta\pi^{ij} = 2\pi_{ij}$ and $\delta\left(\pi^2\right)/\delta\pi^{ij} = 2\pi h_{ij}$. From Lemma A.640 together with the variation of the kinetic part — which involves h_{ij} *algebraically only*, through π_{ij}, π and $\sqrt{h}$, and which therefore produces no derivative of f —

$$\frac{\delta H_\perp[f]}{\delta h_{ij}} = f\,A^{ij} - \frac{\sqrt{h}}{2\kappa}\,\Delta^{ij} f\,, \qquad \Delta^{ij} := D^i D^j - h^{ij} D^2\,, \tag{A.1023}$$

where A^{ij} is a local expression in h and π whose exact form is never needed: it collects the kinetic variation and the term $+\sqrt{h}\,{}^{(3)}G^{ij}/2\kappa$, and the only property used is that it multiplies f undifferentiated.

Step 2: what cancels, and why. Insert Equations (A.1022) and (A.1023) into Equation (A.1005):

$$\begin{aligned}\{H_\perp[f], H_\perp[g]\} &= \int \mathrm{d}^3x \left[\left(f A^{ij} - \frac{\sqrt{h}}{2\kappa}\Delta^{ij} f\right) g B_{ij} - f B_{ij}\left(g A^{ij} - \frac{\sqrt{h}}{2\kappa}\Delta^{ij} g\right)\right] \\ &= \frac{1}{2\kappa}\int \mathrm{d}^3x\,\sqrt{h}\,B_{ij}\left(f\,\Delta^{ij} g - g\,\Delta^{ij} f\right)\,,\end{aligned} \tag{A.1024}$$

because the terms $fgA^{ij}B_{ij}$ are symmetric under $f \leftrightarrow g$ and cancel between the two halves of the bracket. This is the load-bearing observation of the whole computation and it deserves to be stated in words: *everything ultralocal in the smearing functions drops out.* The Einstein-tensor part of Equation (A.1017) goes with it, and so does the entire kinetic variation; what survives is only the group Δ^{ij}, which is precisely the part carrying *two spatial derivatives.* That is why the answer will carry a metric where a structure constant would sit: two derivatives must have their indices raised, and h^{ij} is what raises them.

Step 3: contract. With $\sqrt{h}B_{ij}/2\kappa = 2\left(\pi_{ij} - \frac{1}{2}\pi h_{ij}\right)$,

$$\left(\pi_{ij} - \tfrac{1}{2}\pi h_{ij}\right)\Delta^{ij} f = \pi_{ij} D^i D^j f - \tfrac{1}{2}\pi D^2 f - \pi D^2 f + \tfrac{3}{2}\pi D^2 f = \pi_{ij} D^i D^j f\,, \tag{A.1025}$$

the three trace terms cancelling exactly, the last carrying $h_{ij}h^{ij} = 3$. Hence

$$\{H_\perp[f], H_\perp[g]\} = 2\int \mathrm{d}^3x\;\pi^{ij}\left(f\,D_i D_j g - g\,D_i D_j f\right)\,. \tag{A.1026}$$

Step 4: one integration by parts. π^{ij} is a weight-one density, so Equation (A.1007) applies directly to $\pi^{ij} f D_j g$:

$$\int \pi^{ij} f\,D_i D_j g = -\int \left[\left(D_i \pi^{ij}\right) f + \pi^{ij} D_i f\right] D_j g\,, \tag{A.1027}$$

and likewise with f and g exchanged. In the difference, the term $\pi^{ij}\left(D_i f\,D_j g - D_i g\,D_j f\right)$ is antisymmetric in i, j contracted with the symmetric π^{ij}, and vanishes. What remains is

$$\{H_\perp[f], H_\perp[g]\} = -2\int \mathrm{d}^3x\;\left(D_j \pi^{ji}\right)\left(f\,D_i g - g\,D_i f\right)\,. \tag{A.1028}$$

Step 5: recognize the momentum constraint. Lower the index on π and raise it on the bracket, which is legitimate because $Dh = 0$: $\left(D_j\pi^{ji}\right)\left(fD_i g - gD_i f\right) = \left(D_j \pi^j{}_k\right) h^{ki}\left(f\partial_i g - g\partial_i f\right)$, the covariant derivatives of the scalars f, g being ordinary ones. By Equation (26.85), $-2D_j\pi^j{}_k = \mathcal{H}_k$, so

$$\{H_\perp[f], H_\perp[g]\} = \int \mathrm{d}^3x\;\mathcal{H}_k\,h^{ki}\left(f\partial_i g - g\partial_i f\right) = H_\parallel\left[h^{ij}\left(f\partial_j g - g\partial_j f\right)\right]\,, \tag{A.1029}$$

which is Equation (A.1021). □

Corollary A.642 (The constraints are first class)**.** *Every right-hand side in Equations* (A.1016) *and* (A.1021) *is a smeared constraint, and the primary constraints Equation* (26.81) *have vanishing bracket with everything, since no $\mathcal{H}$ contains N or N^i. So all eight constraints of Theorem 26.42 are first class in the sense of Definition 26.12, $F = 8$ and $S = 0$, and Equation* (26.21) *gives $2 \times 10 - 2 \times 8 = 4$ per point of space.*

Rests on Theorems 26.17 and A.641 and Proposition A.639.

Proof. *Derives Corollary A.642.*

By Proposition 26.13 it suffices that the brackets of the generators close on the constraint set, which is what the three displayed identities say — each right-hand side is $H_\perp$ or $H_\parallel$ of some smearing field, hence weakly zero. The smearing fields on the right of Equation (A.1021) depend on h_{ij}, but that is irrelevant to the weak vanishing: a phase-space-dependent coefficient times a constraint still vanishes on Σ. The count is Equation (26.21) with $2n = 20$ per point. □

A.53.6 The geometric reading

Remark A.643 (Why the metric sits where a structure constant would). The three brackets are the composition law for deformations of a spacelike surface embedded in a Lorentzian spacetime, and read that way they explain themselves.

$H_\parallel[\xi]$ deforms the surface *within itself*: it slides the points along ξ and leaves the surface, as a subset of spacetime, where it was. Two such slides compose as diffeomorphisms of a three-manifold compose, which is Equation (A.1016) left, and their commutator is again a slide. That part of the algebra *is* a Lie algebra — the Lie algebra of vector fields on the leaf — and it is the only part that is.

$H_\perp[f]$ pushes the surface off itself, by a proper time f/c per point along the normal. The mixed bracket says only that f is a scalar: sliding and then pushing differs from pushing and then sliding by pushing with the dragged function $\xi^i \partial_i f$.

The third bracket is the one with content. Push by f, then by g; push by g, then by f. Each pair of normal pushes lands on the same surface — to first order the two orders of deformation reach the same place — but not on the same *points* of it: the two routes differ by a slide. The slide is easy to see. The normal at a point of the first intermediate surface is not the normal at the corresponding point of the second, because the two intermediate surfaces are tilted relative to one another by an amount proportional to the gradient of the pushing function; and tilting a normal is exactly what turns a normal deformation into a tangential one. The resulting tangent vector is $f\,\nabla g - g\,\nabla f$ — the gradient of one function weighted by the other, antisymmetrized — and a gradient is a covector, so to be a deformation vector it must have its index raised. There is only one object available to raise it: the metric of the surface itself. Hence h^{ij} in Equation (A.1021).

And hence the sentence that closes Proposition 26.43. The coefficient on the right of the third bracket is a *function on phase space*, not a number, so Equation (26.16) for general relativity has structure functions and the first-class algebra is not a Lie algebra. This is not a technicality: it is the reason the constraint algebra of gravity cannot be treated as the algebra of a symmetry group with a fixed structure, why the BRST charge of Appendix A.54 needs terms beyond the two displayed in Equation (26.97), and why the comparison with Yang–Mills of Section 26.6.4, where the same brackets do close on constants, is the sharpest single contrast the chapter draws.

Remark A.644 (What is quoted here). Nothing outside the treatise. The constraint densities Equations (26.84) and (26.85) are taken from Theorem 26.42, which proves them; the Levi-Civita apparatus is Theorems 13.150 and 13.152, and the component form of the Lie derivative is Proposition 13.127. Two assumptions are made and are not hidden. The smearing fields are independent of the canonical variables — for field-dependent smearings

the brackets acquire extra terms and Equation (A.1016) is false as it stands. And the fields fall off fast enough that every integration by parts loses its surface term; on a region with a boundary they do not, the constraints are not even functionally differentiable without supplementary surface integrals, and those integrals are the ADM energy and momentum of Remark 26.45.

Remark A.645. Appendix A.53 discharges the derivation owed at Proposition 26.43 of Section 26.6.5 in Chapter 26. It completes the last step of Theorem 26.42, which asserts that the eight constraints are first class and points forward to Equation (26.89) for the proof; with Corollary A.642 the count of two degrees of freedom per point of space — the same count Proposition 26.32 gives for light, and the two polarizations detected in Chapter 47 — rests on nothing unproved. The Lagrangian from which those constraints were obtained is established in Appendix A.52.

A.54 Existence and Uniqueness of a Nilpotent BRST Charge

This appendix proves Theorem 26.51 of Section 26.7.3 in Chapter 26: for every regular, irreducible set of first-class constraints the series Equation (26.97) can be continued to a Grassmann-odd function Ω of ghost number $+1$ satisfying Equation (26.98) exactly; two such functions with the same leading term differ by a canonical transformation of the extended phase space; and when the structure functions of Equation (26.16) are constants the series stops at the two terms Equation (26.97) displays.

This is the statement that makes BRST a method rather than a device. Without it the definition Definition 26.50 would be a recipe that happens to work for Yang–Mills and is silent about general relativity, whose structure functions are genuinely functions (Proposition 26.43, proved in Appendix A.53); with it, the cohomological characterization Equation (26.99) of the physical state space is available for every first-class system.

Scope. BRST is in this treatise because it is the machinery behind the perturbative calculations of Chapter 102 and the path-integral development of Chapter 106, whose predictions are measured; it is not here as a formal programme. Accordingly this section stops where the chapter stops — at the Hamiltonian BRST charge of a finite-dimensional constrained system — and does not go on to the antifield or Batalin–Vilkovisky formulation, which is a different construction with different inputs and is used nowhere in this book.

Attribution. The existence of a nilpotent charge for a general first-class system is due to Fradkin and Vilkovisky and, in the form with structure functions, to Batalin, Fradkin and Vilkovisky; the homological proof organized around the Koszul–Tate differential, which is the one given below, is due to Henneaux and Teitelboim. None of these works has an entry in this treatise's bibliography, so — exactly as Remark 26.2 records for the rest of the chapter — the attribution is made in words and nothing below rests on it. The original BRST papers [1564, 1571] are cited in the chapter.

A.54.1 The extended phase space and its two gradings

Notation A.646 (Gradings and brackets). Let z^c, $c = 1, \dots, 2n$, be the original phase-space variables and γ_A, $A = 1, \dots, F$, the first-class constraints, with Equation (26.16), $\{\gamma_A, \gamma_B\} = f_{AB}{}^C \gamma_C$. Adjoin the ghost pairs $\left(\eta^A, \mathcal{P}_A\right)$ of Definition 26.50. Functions of $(z, \eta, \mathcal{P})$ carry a Grassmann parity $\varepsilon \in \{0, 1\}$ and the graded bracket obeys

$$\{F, G\} = -(-1)^{\varepsilon_F \varepsilon_G} \{G, F\} , \qquad \{F, GH\} = \{F, G\} H + (-1)^{\varepsilon_F \varepsilon_G} G \{F, H\} , \tag{A.1030}$$

together with the graded Jacobi identity. It splits as

$$\{F, G\} = \{F, G\}_M + \{F, G\}_{\text{gh}} , \tag{A.1031}$$

where $\{\cdot,\cdot\}_M$ differentiates only the z^c — so that the ghosts pass through it as constants — and $\{\cdot,\cdot\}_{\mathrm{gh}}$ only the ghost pair, with $\left\{\eta^A,\mathcal{P}_B\right\}=-\delta^A_B$.

Three gradings are used. The *pure ghost number* counts the η's; the *antighost number* counts the $\mathcal{P}$'s,

$$\operatorname{antigh}\mathcal{P}_A=1\,,\qquad \operatorname{antigh}\eta^A=\operatorname{antigh}z^c=0\,, \tag{A.1032}$$

and the *ghost number* of Definition 26.50 is their difference, $\operatorname{gh}=\operatorname{pgh}-\operatorname{antigh}$, so that $\operatorname{gh}\eta^A=+1$ and $\operatorname{gh}\mathcal{P}_A=-1$. Parity is pgh + antigh modulo two.

Rests on Definition 26.50 and Equations (26.16) and (26.97).

Two grading facts are used constantly and are recorded once. $\{\cdot,\cdot\}_M$ preserves the antighost number, while $\{\cdot,\cdot\}_{\mathrm{gh}}$ removes one η and one $\mathcal{P}$ and so *lowers it by one*; both add ghost numbers. And an Ω that is odd with $\operatorname{gh}\Omega=+1$ decomposes as

$$\Omega=\sum_{k\geq 0}\Omega_{(k)}\,,\qquad \operatorname{antigh}\Omega_{(k)}=k\,, \tag{A.1033}$$

in which $\Omega_{(k)}$ has pure ghost number $k+1$: it is a polynomial carrying exactly $k+1$ ghosts and k ghost momenta, hence $2k+1$ Grassmann-odd factors, hence odd, as it must be. The first two terms are the ones Equation (26.97) displays,

$$\Omega_{(0)}=\eta^A\gamma_A\,,\qquad \Omega_{(1)}=-\tfrac{1}{2}\eta^B\eta^A f_{AB}{}^C\mathcal{P}_C\,. \tag{A.1034}$$

A.54.2 The Koszul–Tate differential

Definition A.647 (Koszul–Tate differential)**.** δ is the odd derivation of the algebra of functions on the extended phase space fixed by

$$\delta\mathcal{P}_A=-\gamma_A\,,\qquad \delta z^c=0\,,\qquad \delta\eta^A=0\,. \tag{A.1035}$$

It lowers the antighost number by one and raises the ghost number by one.

Rests on Notation A.646 and Equation (26.16).

Lemma A.648 (δ is the ghost bracket with $\Omega_{(0)}$, and $\delta^2=0$)**.** *For every X, $\delta X=\left\{X,\Omega_{(0)}\right\}_{gh}$, and $\delta^2=0$.*

Rests on Definition A.647 and Equations (A.1030) and (A.1034).

Proof. *Derives Lemma A.648.*

$\Omega_{(0)}$ contains no $\mathcal{P}$, so $\left\{\cdot,\Omega_{(0)}\right\}_{\mathrm{gh}}$ annihilates z^c and η^A, as δ does. On $\mathcal{P}_A$, the graded Leibniz rule Equation (A.1030) gives $\left\{\mathcal{P}_A,\eta^B\gamma_B\right\}_{\mathrm{gh}}=\left\{\mathcal{P}_A,\eta^B\right\}\gamma_B$, and graded antisymmetry between two odd objects gives $\left\{\mathcal{P}_A,\eta^B\right\}=\left\{\eta^B,\mathcal{P}_A\right\}=-\delta^B_A$, so the value is $-\gamma_A=\delta\mathcal{P}_A$. Both maps are odd derivations agreeing on generators, hence equal.

For nilpotency, δ^2 is the square of an odd derivation and is therefore an even derivation; it vanishes on z and η trivially and on $\mathcal{P}_A$ because $\delta^2\mathcal{P}_A=-\delta\gamma_A=0$, the constraints being functions of z alone. A derivation vanishing on generators is zero. □

Remark A.649 (The regularity hypothesis, in the form the proof needs). Remark 26.7 assumes two things about the constraint set, and Theorem 26.51 adds a third word, *irreducible*. Spelled out, what is assumed is: (i) the differentials $\mathrm{d}\gamma_A$ are linearly independent at every point of the surface Σ they define; (ii) every smooth function vanishing on Σ is $c^A\gamma_A$ with smooth c^A; and (iii) every λ^A with $\lambda^A\gamma_A=0$ identically is of the form $\lambda^A=\mu^{AB}\gamma_B$ with $\mu^{AB}=-\mu^{BA}$ — there are no relations among the constraints beyond the trivial ones. Item (iii) is what *irreducible* means; a reducible set, such as the one obtained by writing a two-form constraint in components with a Bianchi identity among them, fails it, and the construction below then needs ghosts for ghosts and is not carried out in this treatise.

Lemma A.650 shows that (i) already implies (ii) and (iii) locally, so the three are not independent assumptions; the content is entirely in (i), which is the same regularity that Proposition 26.5 and Theorem 26.17 rest on.

Lemma A.650 (Flattening the constraints)**.** *About each point of Σ there are local coordinates $\left(y^A, w^\alpha\right)$ on phase space in which $\gamma_A = y^A$. In such coordinates, (ii) and (iii) of Remark A.649 hold.*

Rests on Remarks 26.4 and 26.7 and Theorem A.286.

Proof. *Derives Lemma A.650.*

The map $z \longmapsto (\gamma_1(z), \dots, \gamma_F(z))$ has surjective differential at each point of Σ by (i), hence constant maximal rank on a neighbourhood; the constant-rank theorem stated in Remark 26.4, itself a consequence of Theorem A.286, supplies coordinates in which it is the projection onto the first F of them. Those coordinates are the $\left(y^A, w^\alpha\right)$.

For (ii), let $F(y, w)$ vanish at $y = 0$. By the fundamental theorem of calculus applied along the segment $t \longmapsto (ty, w)$,

$$F(y,w) = y^A \lambda_A(y,w)\,, \qquad \lambda_A(y,w) = \int_0^1 \left(\frac{\partial F}{\partial y^A}\right)(ty,w)\,\mathrm{d}t\,, \tag{A.1036}$$

and λ_A is smooth by differentiation under the integral sign. That is (ii). Item (iii) is the case $k = 1$ of Lemma A.651 below, whose proof uses only Equation (A.1036) and the coordinates just constructed, so nothing is circular. □

Lemma A.651 (Acyclicity of δ)**.** *On the coordinate patch of Lemma A.650: $H_k(\delta) = 0$ for every $k \geq 1$ — every δ-closed function of antighost number $k \geq 1$ is δ of a function of antighost number $k+1$ — and $H_0(\delta)$ is the algebra of smooth functions on Σ (tensored with the ghosts).*

Rests on Lemma A.650, Definition A.647, and Remark A.649.

Proof. *Derives Lemma A.651.*

Work in the coordinates of Lemma A.650, so that $\delta \mathcal{P}_A = -y^A$. The η's play no part: δ annihilates them and they simply multiply everything, so the complex is a free module over the Grassmann algebra they generate and it suffices to treat η-independent coefficients.

A homotopy. Let σ be the odd derivation with

$$\sigma y^A = \mathcal{P}_A\,, \qquad \sigma \mathcal{P}_A = 0\,, \qquad \sigma w^\alpha = 0\,, \qquad \sigma \eta^A = 0\,, \tag{A.1037}$$

so that on functions $\sigma f = \mathcal{P}_A\, \partial f / \partial y^A$. The anticommutator of two odd derivations is an even derivation, so

$$\mathcal{N} := -\left(\delta\sigma + \sigma\delta\right)\,, \tag{A.1038}$$

is a derivation, and it is determined by its values on generators: $\mathcal{N} y^A = -\delta \mathcal{P}_A = y^A$, $\mathcal{N}\mathcal{P}_A = -\sigma\left(-y^A\right) = \mathcal{P}_A$, and $\mathcal{N} w^\alpha = \mathcal{N}\eta^A = 0$. So $\mathcal{N}$ is the Euler operator counting the joint degree in y and $\mathcal{P}$. Because $\delta^2 = 0$, $\mathcal{N}$ commutes with δ.

Inverting $\mathcal{N}$ in positive antighost degree. A general element of antighost number k is $a = \frac{1}{k!} a^{A_1 \dots A_k}(y,w)\, \mathcal{P}_{A_1} \cdots \mathcal{P}_{A_k}$, on which $\mathcal{N} = k + \mathcal{N}_y$ with $\mathcal{N}_y = y^A \partial / \partial y^A$. For $k \geq 1$ define

$$\left(\mathcal{N}^{-1} a\right)(y,w) := \int_0^1 t^{k-1}\, a(ty,w)\,\mathrm{d}t\,, \tag{A.1039}$$

acting on the coefficient functions and leaving the $\mathcal{P}$'s alone; it is smooth by differentiation under the integral sign. It is the inverse: from $\frac{\mathrm{d}}{\mathrm{d}t}\left[t^k a(ty,w)\right] = t^{k-1}\left[k\, a + \mathcal{N}_y a\right](ty,w)$, integrating from 0 to 1 — where the boundary term at $t = 0$ vanishes because $k \geq 1$ — gives $a = \mathcal{N}^{-1}\left(k + \mathcal{N}_y\right) a = \mathcal{N}^{-1}\mathcal{N} a$. Since $\mathcal{N}$ commutes with δ and is invertible in every antighost degree ≥ 1, so does $\mathcal{N}^{-1}$.

Acyclicity. Let $\delta a = 0$ with $\operatorname{antigh} a = k \geq 1$. Then by Equation (A.1038), $\mathcal{N}a = -\delta\sigma a - \sigma\delta a = -\delta\sigma a$, and applying $\mathcal{N}^{-1}$, which commutes with δ,

$$a = -\delta\left(\mathcal{N}^{-1}\sigma a\right), \tag{A.1040}$$

so a is δ-exact, with a primitive of antighost number $k+1$.

Degree zero. At antighost number 0 every element is closed, and the image of δ from degree one is $\left\{\lambda^A\gamma_A\right\}$, which by Equation (A.1036) is exactly the ideal of functions vanishing on Σ. Hence $H_0(\delta) = C^\infty(\Sigma)$.

This is the mathematical heart of the section, and it is worth naming what was consumed: the coordinates supplied by regularity, and nothing else. Irreducibility — item (iii) of Remark A.649 — is the case $k = 1$ of what has just been proved, since $\lambda^A\mathcal{P}_A$ is δ-closed precisely when $\lambda^A\gamma_A = 0$, and a primitive of antighost number 2 is an antisymmetric μ^{AB} with $\lambda^A = \mu^{AB}\gamma_B$. □

A.54.3 The recursion

Lemma A.652 (Nilpotency, order by order)**.** *With Ω expanded as in Equation (A.1033), the antighost-k component of $\{\Omega,\Omega\}$ is*

$$\{\Omega,\Omega\}_{(k)} = 2\,\delta\Omega_{(k+1)} - 2D_{(k)}\,, \tag{A.1041}$$

where

$$D_{(k)} := -\frac{1}{2}\left[\sum_{\substack{i+j=k+1\\ i,j\geq 1}}\left\{\Omega_{(i)},\Omega_{(j)}\right\}_{gh} + \sum_{i+j=k}\left\{\Omega_{(i)},\Omega_{(j)}\right\}_{M}\right], \tag{A.1042}$$

depends on $\Omega_{(0)},\dots,\Omega_{(k)}$ only. So Equation (26.98) is the sequence of equations $\delta\Omega_{(k+1)} = D_{(k)}$, $k \geq 0$. Explicitly

$$D_{(0)} = -\tfrac{1}{2}\,\eta^A\eta^B f_{AB}{}^C\gamma_C\,, \tag{A.1043}$$

$$D_{(1)} = -\tfrac{1}{2}\left\{\Omega_{(1)},\Omega_{(1)}\right\}_{gh} - \left\{\Omega_{(0)},\Omega_{(1)}\right\}_{M}\,, \tag{A.1044}$$

and $\Omega_{(1)}$ of Equation (A.1034) solves $\delta\Omega_{(1)} = D_{(0)}$.

Rests on Lemma A.648, Notation A.646, and Equation (26.98).

Proof. *Derives Lemma A.652.*

Expand $\{\Omega,\Omega\}$ bilinearly and sort by antighost number, using that $\{\cdot,\cdot\}_M$ preserves it and $\{\cdot,\cdot\}_{\mathrm{gh}}$ lowers it by one. The terms of the ghost sum with $i = 0$ or $j = 0$ are $\left\{\Omega_{(0)},\Omega_{(k+1)}\right\}_{\mathrm{gh}} + \left\{\Omega_{(k+1)},\Omega_{(0)}\right\}_{\mathrm{gh}}$, and the bracket of two odd functions is graded-symmetric by Equation (A.1030), so these are equal and their sum is $2\left\{\Omega_{(k+1)},\Omega_{(0)}\right\}_{\mathrm{gh}} = 2\delta\Omega_{(k+1)}$ by Lemma A.648. Everything else is Equation (A.1042). That $D_{(k)}$ involves no $\Omega_{(j)}$ with $j > k$ is immediate from the ranges of the two sums.

For Equation (A.1043): since $\{\cdot,\cdot\}_M$ differentiates only the z^c, the ghosts pass through it, and moving one odd η past the even object $\partial_c\gamma_A$ costs no sign, so

$$\left\{\Omega_{(0)},\Omega_{(0)}\right\}_M = \eta^A\eta^B\left\{\gamma_A,\gamma_B\right\} = \eta^A\eta^B f_{AB}{}^C\gamma_C\,, \tag{A.1045}$$

by Equation (26.16). Halving and negating gives Equation (A.1043). Equation (A.1044) is the case $k = 1$ of Equation (A.1042), the ghost sum having only the term $i = j = 1$ and the M sum only the two equal terms $\left\{\Omega_{(0)},\Omega_{(1)}\right\}_M$.

Finally, δ passes through the even coefficient $-\frac{1}{2}\eta^B\eta^A f_{AB}{}^C$ without a sign and $\delta\mathcal{P}_C = -\gamma_C$, so

$$\delta\Omega_{(1)} = \tfrac{1}{2}\eta^B\eta^A f_{AB}{}^C\gamma_C = -\tfrac{1}{2}\eta^A\eta^B f_{AB}{}^C\gamma_C = D_{(0)}\,, \tag{A.1046}$$

the middle step by anticommutativity of the ghosts. So the term Equation (26.97) displays is not a guess: it is the unique solution of the first equation of the recursion, up to the δ-closed ambiguity that Appendix A.54.6 disposes of. □

Notice what the first step already required. $D_{(0)}$ has antighost number 0, where δ is *not* acyclic, so $\delta\Omega_{(1)} = D_{(0)}$ is solvable only if $D_{(0)}$ lies in the image of δ — that is, by Lemma A.651, only if it vanishes on Σ. Equation (A.1043) shows it is proportional to the constraints, and that is precisely the statement that the constraints are first class. The whole construction is powered by Equation (26.16), and it fails at the first step for a set that is not first class.

A.54.4 Existence

Theorem A.653 (Existence of a nilpotent BRST charge)**.** *Let γ_A be regular and irreducible in the sense of Remark A.649. Then on a neighbourhood of Σ there is an odd Ω of ghost number $+1$, of the form Equation (A.1033) with the first two terms Equation (A.1034), satisfying $\{\Omega,\Omega\} = 0$.*

Rests on Lemmas A.651 and A.652 and Equation (26.16).

Proof. *Derives Theorem A.653.*

Induction on the antighost number. By Equation (A.1046), $\Omega_{(0)}$ and $\Omega_{(1)}$ are constructed and $\{\Omega,\Omega\}$ has no antighost-0 component. Suppose $\Omega_{(0)},\dots,\Omega_{(m)}$ have been found, $m \geq 1$, such that $R := \left\{\Omega^{[m]},\Omega^{[m]}\right\}$, with $\Omega^{[m]} := \sum_{j\leq m}\Omega_{(j)}$, has vanishing components of antighost number $\leq m-1$.

The obstruction is δ-closed. For an odd Ω the graded Jacobi identity applied to the triple $\left(\Omega^{[m]},\Omega^{[m]},\Omega^{[m]}\right)$ collapses to the identity

$$\left\{\Omega^{[m]},\left\{\Omega^{[m]},\Omega^{[m]}\right\}\right\} = 0\,, \tag{A.1047}$$

which holds whatever $\Omega^{[m]}$ is, nilpotent or not: the three cyclic terms of the graded Jacobi identity are equal, and their common sign is such that three times one of them must vanish. Take the antighost-$(m-1)$ component of Equation (A.1047). The M-part contributes $\sum_{i+j=m-1}\left\{\Omega_{(i)},R_{(j)}\right\}_M$, in which every j is at most $m-1$ and every $R_{(j)}$ therefore vanishes by hypothesis. The ghost part contributes $\sum_{i+j=m}\left\{\Omega_{(i)},R_{(j)}\right\}_{\mathrm{gh}}$, in which the only surviving term is $i=0$, $j=m$. Hence $\left\{\Omega_{(0)},R_{(m)}\right\}_{\mathrm{gh}} = 0$, and since R is even and $\Omega_{(0)}$ odd this is $-\delta R_{(m)}$. So $\delta R_{(m)} = 0$.

Solve. $R_{(m)}$ has antighost number $m \geq 1$, so by Lemma A.651 there is a c with $\delta c = R_{(m)}$ and $\operatorname{antigh} c = m+1$. Counting gradings: R is even with ghost number 2, so $R_{(m)}$ has pure ghost number $m+2$; c therefore has pure ghost number $m+2$ and antighost number $m+1$, hence ghost number $+1$ and odd parity — exactly the type of an $\Omega_{(m+1)}$. Put $\Omega_{(m+1)} := -\frac{1}{2}c$. Adding it to $\Omega^{[m]}$ changes nothing of antighost number $< m$, because the new term enters $\{\Omega,\Omega\}$ only at antighost number $\geq m$, and it changes the antighost-m component by $2\delta\Omega_{(m+1)} = -\delta c = -R_{(m)}$, cancelling it. This is the statement $\delta\Omega_{(m+1)} = D_{(m)}$ of Lemma A.652.

The induction produces $\Omega_{(k)}$ for every k. It terminates for a reason of degree: $\Omega_{(k)}$ carries k factors $\mathcal{P}_A$, which anticommute, so $\Omega_{(k)} = 0$ identically for $k > F$. The series Equation (A.1033) is therefore a finite sum, and Ω is a genuine function on the extended phase space, not a formal series. □

A.54.5 Termination at the displayed terms

Proposition A.654 (The Yang–Mills case)**.** *If the ${f_{AB}}^C$ of Equation (26.16) are constants, then $D_{(1)} = 0$ and $\Omega = \Omega_{(0)} + \Omega_{(1)}$ is exactly nilpotent: the series Equation (26.97) stops at the two terms displayed.*

Rests on Lemma A.652 and Equations (22.46) and (26.16).

Proof. *Derives Proposition A.654.*

Both terms of Equation (A.1044) vanish.

The second does so for a trivial reason: with constant f, the function $\Omega_{(1)}$ contains no z^c at all, and $\{\cdot,\cdot\}_M$ differentiates only the z^c, so $\left\{\Omega_{(0)}, \Omega_{(1)}\right\}_M = 0$. This is exactly the place where structure *functions* would obstruct: for non-constant f this bracket is a nonzero function of antighost number 1, and it is what forces $\Omega_{(2)}$ to exist.

The first is the Jacobi identity. The ghost bracket pairs the $\partial/\partial\eta$ of one $\Omega_{(1)}$ with the $\partial/\partial\mathcal{P}$ of the other, producing a term with three η's, one $\mathcal{P}$ and two factors f; because the three ghosts anticommute, the coefficient is the total antisymmetrization ${f_{[AB}}^D {f_{C]D}}^E$. That vanishes: applying the Poisson Jacobi identity Equation (22.46) to the triple $(\gamma_A, \gamma_B, \gamma_C)$ and using Equation (26.16) twice with constant coefficients gives ${f_{[AB}}^D {f_{C]D}}^E \gamma_E = 0$, and the γ_E are independent by irreducibility, so the coefficient itself vanishes.

With $D_{(1)} = 0$ one may take $\Omega_{(2)} = 0$, and then every later $D_{(k)}$ is built from $\Omega_{(0)}$, $\Omega_{(1)}$ and vanishing terms and is itself zero by the same two arguments, so $\Omega_{(k)} = 0$ for all $k \geq 2$ solves the recursion. That is the case of Section 26.6.4; general relativity, whose structure functions are genuinely functions by Proposition 26.43, is not it, and the omitted terms of Equation (26.97) are exactly what Theorem A.653 supplies there. □

A.54.6 Uniqueness

Proposition A.655 (Uniqueness up to a canonical transformation)**.** *Let Ω and Ω' both satisfy the hypotheses and conclusion of Theorem A.653 with the same $\Omega_{(0)} = \Omega'_{(0)} = \eta^A \gamma_A$. Then there is a canonical transformation of the extended phase space, generated by an even function of ghost number 0, carrying Ω to Ω'.*

Rests on Theorem A.653 and Lemmas A.651 and A.652.

Proof. *Derives Proposition A.655.*

Induction on the lowest order at which the two differ. Suppose $\Omega_{(j)} = \Omega'_{(j)}$ for $j < k$ and put $\Delta_{(k)} := \Omega'_{(k)} - \Omega_{(k)}$, with $k \geq 1$. By Lemma A.652 both satisfy $\delta\Omega_{(k)} = D_{(k-1)}$ with the *same* right-hand side, since $D_{(k-1)}$ is built from the orders below k, which agree. Hence $\delta\Delta_{(k)} = 0$.

By Lemma A.651 there is a K with $\delta K = \Delta_{(k)}$ and $\operatorname{antigh} K = k + 1$. Its type is forced: $\Delta_{(k)}$ is odd with ghost number $+1$ and pure ghost number $k + 1$, so K has pure ghost number $k + 1$ and antighost number $k + 1$, hence ghost number 0 and even parity — a legitimate generator of a canonical transformation preserving both gradings.

Let T_K be the canonical transformation generated by K,

$$T_K\Omega = \Omega + \{\Omega, K\} + \tfrac{1}{2}\{\{\Omega, K\}, K\} + \cdots, \tag{A.1048}$$

which is a finite sum here because K has positive antighost number and each bracket lowers it by at most one while raising the number of η's, so the series terminates. Since K has antighost number $k + 1$ and the ghost bracket lowers antighost number by one, $\{\Omega, K\}$ has antighost number $\geq k$, and its antighost-k part is $\left\{\Omega_{(0)}, K\right\}_{\text{gh}} = -\delta K = -\Delta_{(k)}$. So $T_K\Omega'$ agrees with Ω through antighost number k and, being the image of a nilpotent charge under a canonical transformation, is itself nilpotent — a canonical transformation preserves the graded bracket, so it carries $\{\Omega', \Omega'\} = 0$ to $\{T_K\Omega', T_K\Omega'\} = 0$.

Repeat at the next order at which the two still differ. The generators so produced have strictly increasing antighost number, and all vanish beyond F by the degree argument of Theorem A.653, so their composition is a single canonical transformation carrying Ω' to Ω. □

Remark A.656 (What is quoted here). Two things, and one limitation.

Quoted. The constant-rank theorem, in the form Remark 26.4 states it, is used once, in Lemma A.650, to flatten the constraints. That theorem is owed by Chapters 7 and 13 and is a standing debt of Chapter 26, not a new one; the analytic input from which it is proved, Theorem A.286, is in this appendix.

Not quoted, though it usually is. The acyclicity of the Koszul complex in positive degree (Lemma A.651) is normally imported from homological algebra, where it is the statement that the Koszul complex of a regular sequence is a resolution. This treatise carries no homological algebra, and opening a chapter of it to prove one lemma would be out of proportion; the case at hand is therefore proved here outright, by an explicit contracting homotopy Equations (A.1037) and (A.1039), and the general theorem is recorded as a debt of Part II rather than used.

The limitation. Lemma A.650 is local: the coordinates in which $\gamma_A = y^A$ exist on a neighbourhood of a point of Σ, so Lemma A.651 and with it Theorem A.653 are local statements. Passing to a global Ω requires patching the local solutions with a partition of unity, which works whenever the γ_A are globally defined and the surface admits a tubular neighbourhood — the case in every application in this book, where the constraints are given by global formulas — but it is a further step and it is not carried out here. This is the same locality that Remark 26.27 records for gauge fixing, and for the same reason: what is easy near a point of the constraint surface can be obstructed in the large.

Remark A.657. Appendix A.54 discharges the derivation owed at Theorem 26.51 of Section 26.7.3 in Chapter 26. Three consequences are worth carrying back. The construction shows that the two terms Equation (26.97) displays are not an ansatz but the first two solutions of a recursion, the second of them forced; Proposition A.654 shows that the series stops there exactly when Equation (26.16) closes on constants, which is the Yang–Mills case of Section 26.6.4 and is why the chapter can display two terms and say "$+\cdots$"; and Proposition A.655 is what makes Equation (26.99) well posed, since a cohomology defined by a charge that was only unique up to nothing in particular would not be a property of the theory. The first-class algebra that the whole construction consumes is Proposition 26.13; for general relativity it is the hypersurface-deformation algebra of Appendix A.53, whose structure functions are the reason a general existence theorem was needed at all.

A.55 Bertrand's Theorem

This appendix proves Theorem 27.41 of Chapter 27: among all central forces there are exactly two for which *every* bounded orbit closes, namely the inverse-square attraction $f(r) = -k/r^2$ and the linear attraction $f(r) = -\kappa r$ of the isotropic oscillator. Here k carries the unit $\mathrm{N\,m^2}$ and κ the unit $\mathrm{N/m}$, so that f is a force in newtons in both cases.

Nothing is quoted here. The whole argument is carried out from the Binet equation Equation (27.30) of Chapter 27, Taylor's theorem with remainder (Theorem 7.38) and the elementary fact that a continuous rational-valued function on an interval is constant. There is no imported theorem and therefore no *What is quoted here* remark.

What the section does contain, and what the reader should watch for, is the division of labour between its two stages. The chapter's roadmap after Theorem 27.41 states it, and it is worth stating again because the point is easy to get wrong: the first-order stage does *not* isolate the two force laws. It narrows the candidates from all central forces to a

one-parameter family of power laws, and no more. The third-order stage is where the two laws are picked out of that family, and it is the real content of the theorem.

A.55.1 The orbit equation and its circular solutions

Throughout, the motion is that of a particle of mass m (in kg; for a two-body problem read the reduced mass) in a central field $f(r)$, with angular momentum $L = mr^2\dot{\phi} \neq 0$ of Equation (27.20), in $\mathrm{kg\,m^2/s}$. Write $u := 1/r$, in $1/\mathrm{m}$.

Definition A.658 (The orbit function)**.** For a central force f and a fixed angular momentum $L \neq 0$, the *orbit function* is

$$J(u) := -\frac{m}{L^2u^2}\, f\!\left(\frac{1}{u}\right), \tag{A.1049}$$

defined for those $u > 0$ at which f is defined, and carrying the unit $1/\mathrm{m}$. Combining $ma_r = f(r)$ from Equation (19.4) with the Binet equation Equation (27.30) puts the shape of every orbit of that angular momentum in the form

$$\frac{\mathrm{d}^2u}{\mathrm{d}\phi^2} + u = J(u)\,. \tag{A.1050}$$

The force is attractive where $f < 0$, that is where $J > 0$.

Rests on Equations (19.4) *and* (27.30) *and Proposition 27.23.*

The time has been eliminated: Equation (A.1050) is an autonomous second-order equation for the shape alone, and closure of an orbit is the statement that its solution $u(\phi)$ is periodic in ϕ with a period commensurable with 2π.

Definition A.659 (Circular orbit and its index)**.** A *circular orbit* of the field at angular momentum L is a constant solution $u \equiv u_0 > 0$ of Equation (A.1050), that is a root of

$$u_0 = J(u_0)\,. \tag{A.1051}$$

Its *index* is the dimensionless number

$$\beta^2 := 1 - J'(u_0)\,. \tag{A.1052}$$

Rests on Definition A.658 and Equation (A.1050)*.*

Lemma A.660 (The index in terms of the force)**.** *Let $r_0 = 1/u_0$ be the radius of a circular orbit. Then*

$$\beta^2 = 3 + \frac{r_0 f'(r_0)}{f(r_0)}\,, \tag{A.1053}$$

which depends on the force law alone and not on L or m except through the radius at which it is evaluated.

Rests on Definition A.659 and Equation (A.1049)*.*

Proof. *Derives Lemma A.660.*

Differentiate Equation (A.1049), writing f and f' for the values at $r = 1/u$ and using $\mathrm{d}(1/u)/\mathrm{d}u = -u^{-2}$:

$$J'(u) = -\frac{m}{L^2}\left[-2u^{-3}f - u^{-2}f'\,u^{-2}\right] = \frac{m}{L^2}\left[2u^{-3}f + u^{-4}f'\right]. \tag{A.1054}$$

At a circular orbit Equation (A.1051) reads $u_0 = -\,(m/L^2)\,u_0^{-2}f(r_0)$, so

$$\frac{m}{L^2} = -\frac{u_0^3}{f(r_0)}\,, \tag{A.1055}$$

which incidentally requires $f(r_0) < 0$: a circular orbit exists only where the force attracts. Substituting Equation (A.1055) into Equation (A.1054) at $u = u_0$,

$$J'(u_0) = -\frac{u_0^3}{f}\left[2u_0^{-3}f + u_0^{-4}f'\right] = -2 - \frac{f'(r_0)}{u_0 f(r_0)} = -2 - \frac{r_0 f'(r_0)}{f(r_0)}\,, \tag{A.1056}$$

and Equation (A.1052) gives $\beta^2 = 1 - J'(u_0) = 3 + r_0 f'(r_0)/f(r_0)$, which is Equation (A.1053). □

A.55.2 First stage: closure near a circle forces a power law

Proposition A.661 (The first-order narrowing)**.** *Suppose every bounded orbit of the field f closes, and suppose f admits circular orbits at every radius of some interval $0 < r_- < r < r_+$ — which is the case as L ranges over an interval, by Equation (A.1055). Then β is one fixed rational number on that interval and*

$$f(r) = -C\, r^{\beta^2 - 3}\,, \qquad C > 0\,, \tag{A.1057}$$

where $C > 0$ carries whatever unit makes the right-hand side a force in N*, namely* N *times a length raised to the power $3 - \beta^2$. Equivalently, in the orbit variable,*

$$J(u) = A\, u^{1-\beta^2}\,, \qquad A = \frac{mC}{L^2} > 0\,. \tag{A.1058}$$

This is as far as the first-order calculation reaches: it leaves a one-parameter family*, indexed by β, and does not select any member of it.*

Rests on Lemma A.660 and Equation (27.39).

Proof. *Derives Proposition A.661.*

Perturb the circular orbit: put $u = u_0 + \delta$ in Equation (A.1050) and expand J about u_0 by Theorem 7.38. Using Equation (A.1051) to cancel the constant terms and Equation (A.1052) to name the coefficient of δ,

$$\frac{\mathrm{d}^2\delta}{\mathrm{d}\phi^2} + \beta^2\delta = \tfrac{1}{2}J''(u_0)\,\delta^2 + \tfrac{1}{6}J'''(u_0)\,\delta^3 + O\!\left(\delta^4\right)\,. \tag{A.1059}$$

To first order in δ the right-hand side is absent and the solution is $\delta = a\cos(\beta\phi)$ with the origin of ϕ at an apsis. If $\beta^2 \le 0$ the solution grows without bound and there are no bounded orbits in the neighbourhood at all, so $\beta^2 > 0$ and δ oscillates. Successive turning points of r are successive extrema of δ, separated by $\Delta\phi = \pi/\beta$; that is, the apsidal angle Equation (27.39) of these orbits satisfies

$$\lim_{a\to 0}\Phi = \frac{\pi}{\beta}\,. \tag{A.1060}$$

An orbit closes exactly when the total angle swept between one apsis and the next repetition of the whole configuration is a whole multiple of 2π, that is when Φ/π is rational. So β is rational for every circular radius in the interval. By Equation (A.1053), β is a continuous function of r_0 wherever f is continuously differentiable and non-vanishing; a continuous function on an interval taking only rational values has connected image contained in $\mathbb{Q}$, hence is constant. So β is one fixed rational number.

With β constant, Equation (A.1053) is the separable first-order equation $rf'/f = \beta^2 - 3$ on the interval, whose solution is $\log|f| = (\beta^2 - 3)\log r + \text{const}$, that is Equation (A.1057), with $C > 0$ because $f < 0$ at a circular orbit. Substituting into Equation (A.1049),

$$J(u) = \frac{m}{L^2}u^{-2}\,C\,u^{-(\beta^2-3)} = \frac{mC}{L^2}\,u^{1-\beta^2}\,, \tag{A.1061}$$

which is Equation (A.1058). Every $\beta > 0$ produces a power law satisfying the hypotheses used so far, so nothing has yet been excluded. □

Remark A.662 (Why the first stage cannot finish the proof). The condition just used is closure of the orbits *infinitesimally near* a circle, which is strictly weaker than closure of all bounded orbits. Equation (A.1060) is a limit as the amplitude a tends to zero, and it fixes the apsidal angle only in that limit; at finite amplitude the apsidal angle of a power law generally depends on a, and an orbit whose apsidal angle drifts with amplitude cannot close at every amplitude even though it closes in the limit. The second stage is precisely the computation of that drift.

Lemma A.663 (Derivatives of the power law)**.** *For J of the form Equation (A.1058) with circular orbit at u_0,*

$$J'(u_0) = 1 - \beta^2 \,, \tag{A.1062}$$

$$u_0 J''(u_0) = -\beta^2 \left(1 - \beta^2\right) , \tag{A.1063}$$

$$u_0^2 J'''(u_0) = \beta^2 \left(1 - \beta^2\right) \left(1 + \beta^2\right) . \tag{A.1064}$$

Rests on Equations (A.1051) and (A.1058).

Proof. *Derives Lemma A.663.*

Equation (A.1051) applied to Equation (A.1058) reads $u_0 = A u_0^{1-\beta^2}$, that is

$$A\, u_0^{-\beta^2} = 1 \,. \tag{A.1065}$$

Differentiating Equation (A.1058) three times,

$$J'(u) = A \left(1 - \beta^2\right) u^{-\beta^2} \,,$$
$$J''(u) = A \left(1 - \beta^2\right) \left(-\beta^2\right) u^{-\beta^2 - 1} \,,$$
$$J'''(u) = A \left(1 - \beta^2\right) \left(-\beta^2\right) \left(-\beta^2 - 1\right) u^{-\beta^2 - 2} \,,$$

and evaluating at u_0 with Equation (A.1065) gives Equations (A.1062) to (A.1064) in turn. Equation (A.1062) is a consistency check on Equation (A.1052): it returns $\beta^2 = 1 - J'(u_0)$, as it must. □

A.55.3 Second stage: the third-order solvability condition

The apsidal angle at finite amplitude is computed by the standard device for a nonlinear oscillator: allow the frequency itself to depend on the amplitude, and fix that dependence by demanding that the solution stay periodic. The alternative bookkeeping — keeping the frequency at β and requiring that no term proportional to $\phi \sin(\beta\phi)$ appear — gives the same condition, because such a term is exactly the first-order Taylor expansion of the frequency shift.

Theorem A.664 (Bertrand's condition)**.** *Let J be the power law Equation (A.1058) with $\beta^2 > 0$, and let the radial oscillation about the circular orbit u_0 have amplitude a. Then its angular frequency is*

$$\Omega^2 = \beta^2 + a^2 \omega_2 + O\left(a^4\right) , \qquad \omega_2 = -\frac{5\,(J'')^2}{24\beta^2} - \frac{J'''}{8} \,, \tag{A.1066}$$

with J'', J''' evaluated at u_0, and the apsidal angle is $\Phi = \pi/\Omega + O(a^4)$. If every bounded orbit closes then $\omega_2 = 0$, and

$$5\,(J'')^2 + 3\beta^2 J''' = 0 \quad \Longleftrightarrow \quad 2\beta^4 \left(1 - \beta^2\right) \left(4 - \beta^2\right) = 0 \,, \tag{A.1067}$$

so that $\beta^2 = 1$ or $\beta^2 = 4$.

Rests on Lemma A.663, Proposition A.661, and Equation (A.1059).

Proof. *Derives Theorem A.664.*

Introduce the stretched angle $\theta = \Omega\phi$, so that $\mathrm{d}/\mathrm{d}\phi = \Omega\,\mathrm{d}/\mathrm{d}\theta$ and δ is sought as a 2π-periodic function of θ. Writing a dot for $\mathrm{d}/\mathrm{d}\theta$, Equation (A.1059) becomes

$$\Omega^2\ddot{\delta} + \beta^2\delta = \tfrac{1}{2}J''\delta^2 + \tfrac{1}{6}J'''\delta^3 + O\left(\delta^4\right) . \tag{A.1068}$$

Expand in the amplitude,

$$\delta = a\delta_1 + a^2\delta_2 + a^3\delta_3 + \cdots , \qquad \Omega^2 = \beta^2 + a^2\omega_2 + O\left(a^4\right) , \tag{A.1069}$$

the frequency carrying only even powers of a because Equation (A.1068) is unchanged by $a \mapsto -a$ together with $\theta \mapsto \theta + \pi$.

Order a. $\beta^2\left(\ddot{\delta}_1 + \delta_1\right) = 0$, so $\delta_1 = \cos\theta$, the amplitude being carried by a and the phase fixed by putting an apsis at $\theta = 0$.

Order a^2. Using $\cos^2\theta = \tfrac{1}{2}\left(1 + \cos 2\theta\right)$,

$$\beta^2\left(\ddot{\delta}_2 + \delta_2\right) = \tfrac{1}{2}J''\cos^2\theta = \frac{J''}{4}\left(1 + \cos 2\theta\right) , \tag{A.1070}$$

which has no resonant forcing — neither 1 nor $\cos 2\theta$ is a solution of the homogeneous equation — and is solved by

$$\delta_2 = \frac{J''}{4\beta^2}\left(1 - \tfrac{1}{3}\cos 2\theta\right) , \tag{A.1071}$$

as is checked by applying $\ddot{} + 1$: the constant reproduces itself, and $-\tfrac{1}{3}\cos 2\theta$ returns $-\tfrac{1}{3}\left(-4 + 1\right)\cos 2\theta = \cos 2\theta$. The homogeneous solution that could be added here is absorbed into the definition of a.

Order a^3. Collecting the terms of that order in Equation (A.1068), and using $\ddot{\delta}_1 = -\cos\theta$,

$$\beta^2\left(\ddot{\delta}_3 + \delta_3\right) = \omega_2\cos\theta + J''\delta_1\delta_2 + \tfrac{1}{6}J'''\cos^3\theta\,. \tag{A.1072}$$

Reduce the two nonlinear terms with $\cos\theta\cos 2\theta = \tfrac{1}{2}\left(\cos 3\theta + \cos\theta\right)$ and $\cos^3\theta = \tfrac{3}{4}\cos\theta + \tfrac{1}{4}\cos 3\theta$:

$$\begin{aligned} J''\delta_1\delta_2 &= \frac{\left(J''\right)^2}{4\beta^2}\left[\cos\theta - \tfrac{1}{3}\cos\theta\cos 2\theta\right] \\ &= \frac{\left(J''\right)^2}{4\beta^2}\left[\tfrac{5}{6}\cos\theta - \tfrac{1}{6}\cos 3\theta\right] , \end{aligned} \tag{A.1073}$$

$$\tfrac{1}{6}J'''\cos^3\theta = \frac{J'''}{8}\cos\theta + \frac{J'''}{24}\cos 3\theta\,. \tag{A.1074}$$

The operator $\ddot{} + 1$ annihilates $\cos\theta$, so Equation (A.1072) has a 2π-periodic solution only if the total coefficient of $\cos\theta$ on its right-hand side vanishes. By Equations (A.1073) and (A.1074) that coefficient is

$$\omega_2 + \frac{5\left(J''\right)^2}{24\beta^2} + \frac{J'''}{8} = 0\,, \tag{A.1075}$$

which is the second member of Equation (A.1066). (The $\cos 3\theta$ terms are non-resonant and merely fix δ_3; they play no part in what follows.) The apsidal angle is half the period of δ in ϕ, that is π/Ω.

Now impose closure. For each amplitude a the orbit closes only if $\Phi/\pi = 1/\Omega(a)$ is rational; Ω is a continuous function of a by Equation (A.1066), and a continuous rational-valued function on an interval is constant — the same step used in Proposition A.661, now

applied in the amplitude rather than in the radius. Hence Ω is independent of a and $\omega_2 = 0$. Multiplying Equation (A.1075) by $-24\beta^2$ gives the first member of Equation (A.1067).

It remains to evaluate it. Substituting Equations (A.1063) and (A.1064),

$$\begin{aligned} 5\left(J''\right)^2 + 3\beta^2 J''' &= \frac{5\beta^4\left(1-\beta^2\right)^2}{u_0^2} + \frac{3\beta^4\left(1-\beta^2\right)\left(1+\beta^2\right)}{u_0^2} \\ &= \frac{\beta^4\left(1-\beta^2\right)}{u_0^2}\left[5\left(1-\beta^2\right) + 3\left(1+\beta^2\right)\right] \\ &= \frac{2\beta^4\left(1-\beta^2\right)\left(4-\beta^2\right)}{u_0^2}\,, \end{aligned} \tag{A.1076}$$

since $5 - 5\beta^2 + 3 + 3\beta^2 = 8 - 2\beta^2$. With $\beta^2 > 0$ the factor β^4 cannot vanish, so $\beta^2 = 1$ or $\beta^2 = 4$. □

Remark A.665 (The two roots are not accidents of the algebra). Each root switches off the condition for a different reason, and it is worth seeing both. At $\beta^2 = 1$ the power law is $J(u) = A$, a constant: *every* derivative of J vanishes, the right-hand side of Equation (A.1059) is empty to all orders, and the orbit equation is exactly linear — which is Equation (27.47), and is why the Kepler problem has an apsidal angle independent of amplitude rather than merely to third order. At $\beta^2 = 4$ the derivatives do not vanish; by Equations (A.1063) and (A.1064), $u_0 J'' = 12$ and $u_0^2 J''' = -60$, and the two terms of Equation (A.1067) cancel against each other: $u_0^2\left[5\left(J''\right)^2 + 3\beta^2 J'''\right] = 5{\cdot}12^2 + 3{\cdot}4{\cdot}(-60) = 720 - 720 = 0$. The oscillator therefore closes by a cancellation between the second-order feedback of Equation (A.1071) and the direct cubic term, not because its orbit equation is linear.

A.55.4 The two laws, and that they really do close

The argument so far is one of *necessity*: no force law other than these two can have every bounded orbit closed. Sufficiency has to be checked separately, and must be checked for all bounded orbits and not only for those near a circle, since that is the property claimed.

Proposition A.666 (Both exceptional laws close every bounded orbit)**.** *For $f(r) = -k/r^2$ every bounded orbit is an ellipse with the force centre at a focus, with apsidal angle $\Phi = \pi$; for $f(r) = -\kappa r$ every orbit is an ellipse with the force centre at its* centre*, with apsidal angle $\Phi = \pi/2$. In both cases Φ is independent of the energy and of the angular momentum.*

Rests on Equations (27.47) and (A.1053).

Proof. *Derives Proposition A.666.*

The inverse square. With $f = -k/r^2$, $rf'/f = r\,(2k/r^3)\,/\,(-k/r^2) = -2$, so Equation (A.1053) gives $\beta^2 = 1$, consistent with Theorem A.664. The orbit is Equation (27.47), $1/r = (mk/L^2)\,(1 + e\cos\phi)$, obtained there without any expansion in amplitude. It is 2π-periodic in ϕ for every e, and bounded precisely for $e < 1$; r is least at $\phi = 0$ and greatest at $\phi = \pi$, so $\Phi = \pi$ exactly, at every eccentricity. The force centre sits at $r = 0$, which is a focus of the conic.

The linear law. With $f = -\kappa r$, $rf'/f = r\,(-\kappa)\,/\,(-\kappa r) = 1$, so $\beta^2 = 4$ and $\beta = 2$. Here the direct route is better than the orbit equation. The force is $\boldsymbol{F} = -\kappa\boldsymbol{x}$, so each Cartesian component of Equation (19.4) is an independent harmonic oscillator of the *same* angular frequency $\omega_0 = \sqrt{\kappa/m}$, in rad/s. In the plane of the motion,

$$x(t) = a_1\cos\left(\omega_0 t\right)\,, \qquad y(t) = a_2\cos\left(\omega_0 t - \gamma\right)\,, \tag{A.1077}$$

which is a closed ellipse centred on the origin, traversed once per period $2\pi/\omega_0$, for every a_1, a_2 and phase γ. Every orbit is bounded and every orbit closes. Its radius $r = \sqrt{x^2 + y^2}$ is a

function of $\cos^2$ and $\sin^2$ of $\omega_0 t$ and therefore has period π/ω_0, half the orbital period: the particle passes two pericentres and two apocentres per revolution, one pair at each end of each principal axis, and the apsidal angle is a quarter of a revolution,

$$\Phi = \frac{2\pi}{4} = \frac{\pi}{2} = \frac{\pi}{\beta}\,, \tag{A.1078}$$

in agreement with Equation (A.1060) — and here the agreement holds at every amplitude, not merely in the limit, which is the content of $\omega_2 = 0$.

The contrast between the two cases is not a detail of geometry. For the inverse square the centre of force is a *focus* of the ellipse and the two apsidal distances differ; for the oscillator it is the *centre*, the two axes are traversed symmetrically, and the orbit closes after half as much angle. That is the same fact as $\beta = 1$ against $\beta = 2$. □

Remark A.667 (What "closed" is doing in the statement). Three distinct properties are in play and only the strongest is Bertrand's. (i) *Some* bounded orbit closes: true for every attractive power law, since the circular orbits themselves close. (ii) Every orbit *infinitesimally near* a circle closes: true for every power law with β rational — a countable but infinite family, including $f \propto r^{-5/4}$ with $\beta = \frac{1}{2}$ and $f \propto r^6$ with $\beta = 3$. (iii) Every bounded orbit closes: true only for $\beta = 1$ and $\beta = 2$. The whole distance between (ii) and (iii) is Equation (A.1075), which measures how the apsidal angle drifts as the orbit is made less circular. For any other power law, Φ varies continuously with the amplitude, takes irrational multiples of π on a dense set of amplitudes, and those orbits fill the annulus $r_- \le r \le r_+$ without ever repeating, exactly as Theorem 27.41 asserts.

Remark A.668 (Attribution). The theorem is Bertrand's, announced in the *Comptes rendus* of the Paris Academy in 1873. That memoir is not among the sources verified for this treatise and has no key in `references.bib`, so the attribution is made here in words, on the same footing as Darboux's memoir in Remark A.81. Nothing above rests on it: the proof given is self-contained, and the third-order stage in particular is carried out from Equation (A.1059) alone.

Remark A.669. Appendix A.55 discharges the derivation owed at Theorem 27.41 of Chapter 27. The chapter uses the theorem twice. It uses it to say what the closure of planetary orbits is evidence *for* — the inverse-square law and nothing weaker — and it uses it, through Phenomenon 27.42, to turn Mercury's residual perihelion advance into a measurement: by Equation (A.1053) an apsidal angle differing from π is a force law differing from r^{-2}, and the size of the difference is the size of the departure. The reader returning there should carry back the division of labour set out at the head of this section, since the chapter's roadmap states it and this appendix is where it is earned: Proposition A.661 narrows all central forces to one parameter, and Theorem A.664 spends that parameter.

A.56 Isotropic and Cubic Elastic Tensors

This appendix proves Lemma 30.26 of Chapter 30 and the statement the chapter takes from it: that the twenty-one independent constants of Proposition 30.25 collapse to *two* for an isotropic solid, the Lamé constants of Equation (30.21). It also supplies the intermediate result the chapter never states, and which is the more useful of the two in a laboratory: a solid invariant only under the rotation group of the cube has *three* independent elastic constants, and the single number by which the general cubic tensor fails to be isotropic is directly measurable.

A.56.1 What is proved here, and what is proved elsewhere

Remark A.670 (The classification itself belongs to Part II, and is now there). When Chapter 30 was written, the classification of the isotropic Cartesian tensors of low rank was carried by no chapter of Part II, and the chapter recorded the gap as a debt owed by the Cartesian-tensor algebra of Chapter 13. That debt is discharged, and not where it was expected: the classification for ranks one to four is Theorem 5.128 of Chapter 5 — the representation chapter, not the manifolds one, because the statement is invariant theory for SO(3) rather than tensor algebra — proved there in full from the definition Equation (5.164) of an isotropic Cartesian tensor. Section 13.2.2 carries the pointer and the constitutive-law reading. Remark 30.27 has been rewritten accordingly and now records what the chapter *rests on* rather than what it is owed; the Helmholtz decomposition named there is likewise in place, as Theorem 7.105.

This section therefore does *not* reprove the classification. It quotes Theorem 5.128, adds the two things that theorem does not contain — the independence of the three delta products, and the strictly weaker classification under the finite rotation group of the cube — and then carries out the specialisation to elasticity, which is physics and belongs here.

For reference, the result quoted is this. A Cartesian tensor T of rank N is *isotropic* (Definition 5.126) when its components are unchanged by every $R \in \mathrm{SO}(3)$,

$$T_{i_1 \dots i_N} = R_{i_1 j_1} \cdots R_{i_N j_N}\, T_{j_1 \dots j_N} \qquad \text{for every } R \in \mathrm{SO}(3)\,, \tag{A.1079}$$

and Theorem 5.128 states that the isotropic tensors of rank one, two, three and four are respectively 0, the multiples of δ_{ij}, the multiples of ε_{ijk}, and the linear combinations

$$T_{iklm} = a\,\delta_{ik}\delta_{lm} + b\,\delta_{il}\delta_{km} + c\,\delta_{im}\delta_{kl}\,, \qquad a, b, c \in \mathbb{R}\,, \tag{A.1080}$$

which is Equation (30.20) and hence Lemma 30.26. Nothing below reopens that proof.

Lemma A.671 (The three products are independent). *The three arrays $\delta_{ik}\delta_{lm}$, $\delta_{il}\delta_{km}$ and $\delta_{im}\delta_{kl}$ are linearly independent, so the space of isotropic rank-four Cartesian tensors has dimension exactly three and the constants a, b, c of Equation* (A.1080) *are uniquely determined by T.*

Rests on Theorem 5.128 and Equation (5.168).

Proof. *Derives Lemma A.671.*

Evaluate Equation (A.1080) on the three index quadruples $(1,1,2,2)$, $(1,2,1,2)$ and $(1,2,2,1)$. A product of two Kronecker deltas is 1 when both pairs it joins carry equal values and 0 otherwise, so

$$T_{1122} = a\,, \qquad T_{1212} = b\,, \qquad T_{1221} = c\,. \tag{A.1081}$$

Each probe isolates one coefficient and annihilates the other two; hence a vanishing combination has $a = b = c = 0$, and the coefficients of a given T are read off by Equation (A.1081). □

A.56.2 The rotation group of the cube leaves four constants

A crystal is not isotropic. Its elasticity tensor is invariant only under the finite point group of its lattice, and for the cubic classes — which include the two commonest structural metals, body-centred iron and face-centred aluminium — that group is the rotation group of the cube. The classification under this smaller group is a strictly weaker statement than Theorem 5.128, and the difference between the two is one number.

Definition A.672 (The octahedral rotation group). The *octahedral group* $\mathcal{O} \subset \mathrm{SO}(3)$ is the group of the 24 rotations carrying a cube centred at the origin, with faces normal to the coordinate axes, to itself. In the standard frame its elements are exactly the 3×3 matrices

carrying a single entry ± 1 in each row and each column and having determinant $+1$; such a matrix acts as $R_{ij} = s_i\,\delta_{i\,\pi(j)}$ for a permutation π of $\{1,2,3\}$ and signs $s_i = \pm 1$ whose product is $\operatorname{sgn}\pi$. A tensor obeying Equation (A.1079) for every $R \in \mathcal{O}$, but not necessarily for every $R \in \mathrm{SO}(3)$, is called *cubic*.

Rests on Definitions 5.126 and 14.26.

Proposition A.673 (Cubic tensors of rank four)**.** *The rank-four cubic Cartesian tensors form a four-dimensional space, spanned by the three products of Equation* (A.1080) *together with the* cubic array

$$\Delta_{iklm} = \begin{cases} 1 & \text{if } i = k = l = m\,, \\ 0 & \text{otherwise} \end{cases}\,, \tag{A.1082}$$

whose components are taken in the crystal frame of Definition A.672. Every cubic T is therefore

$$T_{iklm} = a\,\delta_{ik}\delta_{lm} + b\,\delta_{il}\delta_{km} + c\,\delta_{im}\delta_{kl} + f\,\Delta_{iklm}\,, \tag{A.1083}$$

and T is isotropic if and only if $f = 0$.

Rests on Definition A.672 and Lemma A.671.

Proof. *Derives Proposition A.673.*

Step 1: the half-turn filter. For each $p \in \{1,2,3\}$ the diagonal matrix with $+1$ in the p-th place and -1 in the other two is a half-turn about the p-axis; it has determinant $+1$ and is a signed permutation, so it lies in $\mathcal{O}$. The argument of Lemma 5.127 uses these three matrices and nothing else, so its conclusion holds verbatim for a cubic tensor: a component T_{iklm} vanishes unless every value p occurs among its four indices an even number of times. At rank four the multiplicities (m_1, m_2, m_3) sum to four, so the surviving components have multiplicities $(4,0,0)$ or $(2,2,0)$ in some order. These are the three components T_{pppp} and, for each of the six *ordered* pairs $p \neq q$, the three arrangements T_{ppqq}, T_{pqpq}, T_{pqqp}: twenty-one components in all, and $81 - 21 = 60$ vanishing ones.

Step 2: the three-fold axis. The cyclic matrix C with $C_{21} = C_{32} = C_{13} = 1$ and every other entry zero is a signed permutation of determinant $+1$ — a three-cycle is even — so it too lies in $\mathcal{O}$, and it is the rotation by $2\pi/3$ about the body diagonal $(1,1,1)/\sqrt{3}$. Substituted into Equation (A.1079) it says that the components are unchanged when every index *value* is relabelled by the cycle $\sigma : 1 \mapsto 3 \mapsto 2 \mapsto 1$. Hence $T_{1111} = T_{2222} = T_{3333}$, a common value d, and the ordered pairs (p,q) fall into the two orbits $\{(1,2),(3,1),(2,3)\}$ and $\{(2,1),(1,3),(3,2)\}$, within each of which the three arrangements have common values.

Step 3: the four-fold axis. The quarter turn Q about the third axis, with $Q_{21} = 1$, $Q_{12} = -1$, $Q_{33} = 1$, is a signed permutation of determinant $+1$ and lies in $\mathcal{O}$. It exchanges the values 1 and 2 and multiplies a component by $(-1)^{m_1}$. The components T_{1122}, T_{1212}, T_{1221} each have $m_1 = 2$ and so carry the sign $+1$, giving $T_{1122} = T_{2211}$, $T_{1212} = T_{2121}$, $T_{1221} = T_{2112}$: the two orbits of Step 2 are joined to one another. The three numbers

$$a = T_{ppqq}\,, \qquad b = T_{pqpq}\,, \qquad c = T_{pqqp} \qquad (p \neq q)\,, \tag{A.1084}$$

are therefore independent of which ordered pair is used, and together with $d = T_{pppp}$ they determine every component: the invariant space has dimension at most four.

Step 4: it has dimension exactly four, and Equation (A.1083) *is its general element.* The right-hand side of Equation (A.1083) is cubic: Δ is unchanged by any relabelling of values and, having all four indices equal on each nonvanishing component, by any sign change; and the three delta products are isotropic by Theorem 5.128, hence a fortiori cubic. Reading off its components with Equation (A.1081) gives $T_{ppqq} = a$, $T_{pqpq} = b$, $T_{pqqp} = c$ and $T_{pppp} = a + b + c + f$, so the four constants (a,b,c,f) of the ansatz reproduce the four constants (a,b,c,d) of Steps 2 and 3 through

$$f = d - (a + b + c)\,. \tag{A.1085}$$

The map $(a, b, c, f) \mapsto (a, b, c, d)$ is a bijection, so the four constants of the ansatz are in one-to-one correspondence with those found in Steps 2 and 3. The four arrays are independent: the three delta products are independent of one another by Lemma A.671, and none of them is a multiple of Δ, since Δ alone changes T_{1111} without changing T_{1122}, T_{1212} or T_{1221}. Hence the dimension is exactly four.

Step 5: isotropy is $f = 0$. If $f = 0$ then T is a combination of the three isotropic products, hence isotropic. Conversely, an isotropic T is of the form Equation (A.1080) by Theorem 5.128, and its coefficients are those read off by Equation (A.1081), so $d = T_{1111} = a + b + c$ and $f = 0$ by Equation (A.1085). □

Remark A.674 (Where the extra relation comes from). The whole difference between the cubic and the isotropic classifications is the single relation $d = a + b + c$, and it is worth being clear about which rotations supply it. Steps 1 to 3 used only the finite group: three half-turns, one three-fold axis, one four-fold axis. Those fix every component in terms of four numbers and can do no more, because the cubic array Equation (A.1082) is genuinely invariant under all 24 of them. The relation is produced instead by a rotation through a *general* angle — in the proof of Theorem 5.128 it is the rotation by θ about the third axis, which forces $2\cos^2\theta\sin^2\theta\,(a + b + c - d) = 0$ and hence, at $\theta = \pi/4$, $d = a + b + c$. So the elementary argument stopped one step early does not fail: it delivers cubic anisotropy, which is a physically meaningful and experimentally accessible intermediate result rather than a defective version of isotropy.

A.56.3 From twenty-one constants to three, and then to two

We now impose on Equation (A.1083) the symmetries that an elasticity tensor has by Proposition 30.25: the minor symmetries $C_{iklm} = C_{kilm} = C_{ikml}$, which follow from the symmetry of the strain and of the stress, and the major symmetry $C_{iklm} = C_{lmik}$, which follows from the existence of the energy density of Definition 30.24.

Theorem A.675 (Three constants for a cubic crystal, two for an isotropic solid)**.** *Let C_{iklm} be an elasticity tensor in the sense of Definition 30.24, so that it obeys the minor and major symmetries of Proposition 30.25.*

1. *If C is cubic, then in the crystal frame*

$$C_{iklm} = \lambda\,\delta_{ik}\delta_{lm} + \mu\,(\delta_{il}\delta_{km} + \delta_{im}\delta_{kl}) + f\,\Delta_{iklm}\,, \tag{A.1086}$$

   three independent constants, all of SI dimension Pa*. The minor symmetries are what force $b = c = \mu$; the major symmetry is then automatic and imposes nothing further.*

2. *If C is isotropic, then $f = 0$ and $\sigma_{ik} = C_{iklm}u_{lm}$ is Equation (30.21), with two constants.*

Rests on Propositions 30.25 and A.673 and Definition 30.24.

Proof. *Derives Theorem A.675.*

The minor symmetries. Exchange i and k in Equation (A.1083). The first term is unchanged, since δ_{ik} is symmetric; the cubic array is unchanged, since it is symmetric under every permutation of its indices; and the middle two terms are exchanged,

$$b\,\delta_{il}\delta_{km} + c\,\delta_{im}\delta_{kl} \;\longmapsto\; b\,\delta_{kl}\delta_{im} + c\,\delta_{km}\delta_{il}\,. \tag{A.1087}$$

Equality of Equation (A.1087) with the original, together with the independence of the two arrays $\delta_{il}\delta_{km}$ and $\delta_{im}\delta_{kl}$ established in Lemma A.671, forces $b = c$. Writing $a = \lambda$ and $b = c = \mu$ gives Equation (A.1086). Exchanging l and m instead produces the same equation and hence no new condition.

The major symmetry. Send $(ik) \mapsto (lm)$ in Equation (A.1086). The first term becomes $\lambda\delta_{lm}\delta_{ik}$, which is itself; the second becomes $\mu(\delta_{li}\delta_{mk} + \delta_{lk}\delta_{mi})$, which is itself because each Kronecker delta is symmetric; and Δ is symmetric under every index permutation. So the major symmetry is satisfied identically, and a cubic crystal has exactly the three constants (λ, μ, f).

Isotropy. If C is isotropic then $f = 0$ by Proposition A.673, and contracting Equation (A.1086) with the symmetric strain u_{lm} gives

$$\sigma_{ik} = C_{iklm}u_{lm} = \lambda\,\delta_{ik}u_{ll} + \mu\,(u_{ik} + u_{ki}) = \lambda\,u_{ll}\,\delta_{ik} + 2\mu\,u_{ik}\,, \tag{A.1088}$$

which is Equation (30.21), the second equality using $u_{ik} = u_{ki}$. Counting: the general tensor has 21 constants by Proposition 30.25, a cubic crystal three, an isotropic solid two. □

Remark A.676 (The third cubic constant is measured, and it is not small). In the Voigt notation, in which an index pair (ik) is coded as one of $1, \dots, 6$ by $11 \mapsto 1$, $22 \mapsto 2$, $33 \mapsto 3$, $23 \mapsto 4$, $13 \mapsto 5$, $12 \mapsto 6$, the three constants of Equation (A.1086) are

$$C_{11} = C_{1111} = \lambda + 2\mu + f\,, \qquad C_{12} = C_{1122} = \lambda\,, \qquad C_{44} = C_{2323} = \mu\,, \tag{A.1089}$$

so that

$$f = C_{11} - C_{12} - 2C_{44}\,, \qquad A \equiv \frac{2C_{44}}{C_{11} - C_{12}} = \frac{2\mu}{2\mu + f}\,, \tag{A.1090}$$

and $A = 1$ exactly when $f = 0$, that is exactly when the crystal is elastically isotropic. A is the Zener anisotropy ratio, and the measured values [1455] show that the isotropic idealisation is a poor description of a single crystal and an accidentally good one for a few materials: for body-centred iron $C_{11} = 230\,\text{GPa}$, $C_{12} = 135\,\text{GPa}$, $C_{44} = 117\,\text{GPa}$ give $A = 2.5$; for face-centred aluminium $C_{11} = 108\,\text{GPa}$, $C_{12} = 61\,\text{GPa}$, $C_{44} = 28.5\,\text{GPa}$ give $A = 1.2$; and for tungsten $C_{11} = 523\,\text{GPa}$, $C_{12} = 203\,\text{GPa}$, $C_{44} = 160\,\text{GPa}$ give $A = 1.00$, so that tungsten is isotropic to the precision of the measurement — a coincidence of the numbers, not a consequence of any symmetry. A *polycrystalline* specimen of iron, on the other hand, is isotropic on scales large against the grain, because the orientations average; that is why Equation (30.21) describes structural steel and would not describe a single iron whisker.

Rests on Theorem A.675 and Definition 30.31.

Remark A.677 (The lower ranks, and why an isotropic solid is not piezoelectric). The two lower-rank cases of Theorem 5.128 carry physics of their own and are used elsewhere in Part III. *Rank two*: an isotropic tensor is $a\,\delta_{ik}$, so every rank-two material property of an isotropic body — thermal expansion, thermal conductivity, electrical conductivity — is a single scalar times the identity, and the response is parallel to the drive. *Rank three*: an isotropic tensor is $a\,\varepsilon_{ikl}$, and ε is a *pseudo*tensor: it is invariant under SO(3) but changes sign under the inversion $R = -\mathbb{1}$, by Equation (5.59). A material property that is a true rank-three tensor and is invariant under the full orthogonal group, inversion included, therefore vanishes identically. The piezoelectric coupling $P_i = d_{ikl}\sigma_{kl}$ is such a property, so no centrosymmetric medium is piezoelectric — quartz is not centrosymmetric and is, a fact used in every quartz oscillator, and no isotropic solid can be. The same parity argument is why an odd-rank isotropic tensor of rank one is zero.

Rests on Theorem 5.128 and Remark 5.129.

Remark A.678 (A representation-theoretic check, and why it is not the proof). The dimension three found in Lemma A.671 can be checked against the character theory of Chapter 5 without recomputing a single component. Write $V = \mathbb{R}^3$ for the defining representation of SO(3). An invariant of $V^{\otimes 4}$ is, after using the invariant inner product to identify V with its dual, an SO(3)-equivariant endomorphism of $V \otimes V$; and the Clebsch–Gordan series Theorem 14.47 decomposes $V \otimes V$ into the trace, the antisymmetric part and the

symmetric traceless part, of dimensions 1, 3 and 5, each occurring *once*. Schur's lemmas (Theorems 5.153 and 5.155) then say that an equivariant map is a scalar on each summand and zero between distinct ones, so the space of such maps has dimension $1 + 1 + 1 = 3$.

This is offered as a check and not as the proof, for a reason worth stating. Over the complex numbers Schur's second lemma delivers a scalar; over the reals it delivers only that the endomorphism algebra of an irreducible is a division algebra, which may be $\mathbb{R}$, $\mathbb{C}$ or the quaternions, and the multiplicity count above is correct only because each real integer-spin irreducible of SO(3) has endomorphism algebra $\mathbb{R}$. That last step is not carried by Chapter 5, whereas the elementary argument of Theorem 5.128 is complete as it stands. Where the two routes agree, as here, the agreement is evidence that neither has dropped a constant.

Remark A.679. Appendix A.56 discharges the derivation owed at Lemma 30.26 of Chapter 30. The classification itself is Theorem 5.128 of Chapter 5 and is not reproved here (Remark A.670); what this section adds is the uniqueness of the three coefficients (Lemma A.671), the cubic classification (Proposition A.673) that shows exactly which rotations the isotropy of an elastic solid actually consumes, and the reduction $21 \to 3 \to 2$ of Theorem A.675, which is the step the derivation of Phenomenon 30.28 takes for granted when it writes $b = c$ and $a = \lambda$, $b = \mu$. Everything downstream of Equation (30.21) in that chapter — the moduli of Definition 30.31, the elastic waves, the plate of Appendix A.57 and the contact of Appendix A.58 — rests on the two-constant form proved here.

A.57 Reduction of Three-Dimensional Elasticity to the Kirchhoff Plate Equation

This appendix proves Theorem 30.64 of Chapter 30: that the transverse deflection $w(x, y, t)$ of a thin isotropic plate obeys $\rho h\, \partial_t^2 w + D\nabla^4 w = q$ with the flexural rigidity Equation (30.57), and that a free edge carries exactly *two* boundary conditions rather than the three a naive force balance demands. The reduction has three stages and each is a separate piece of work: a kinematic hypothesis about how the material moves, a constitutive statement that the plate is in plane *stress* and not plane strain, and a variational calculation whose boundary terms are the interesting part of the answer.

Everything used is in the book. The constitutive law is Equation (30.21), proved in Appendix A.56; the conversions among the moduli are Proposition 30.33; the flexural rigidity is Definition 30.63; the variational machinery is that of Chapter 16, and in particular the fundamental lemma Lemma 16.18 and the natural boundary conditions Theorem 16.31. Nothing is quoted from outside the treatise. One point of bookkeeping is worth making in advance: Theorem 16.38 states the Euler–Poisson equation for a functional of *one* independent variable, while the plate functional depends on two, so the corresponding equation is derived here by carrying out the two integrations by parts explicitly — which is necessary anyway, because the boundary terms those integrations produce are the content of Appendix A.57.5.

Throughout, Greek indices α, β, γ run over the in-plane values x, y and are summed when repeated; Latin indices run over x, y, z. The mid-surface is $z = 0$, the thickness is h and the material occupies $|z| \le h/2$. Overdots are ∂_t.

A.57.1 Kinematics: the Kirchhoff hypothesis

Definition A.680 (Kirchhoff kinematics)**.** The plate is said to deform *in the Kirchhoff manner* when the mid-surface points move only transversely and material lines initially normal to the mid-surface remain straight, normal and unstretched. Writing $w(x, y, t)$ for

the transverse displacement of the mid-surface, this is

$$u_z = w(x,y,t)\,, \qquad u_\alpha = -z\,\partial_\alpha w(x,y,t)\,, \tag{A.1091}$$

so that a cross-section rotates through the angle $\partial_\alpha w$, as in the rod of Proposition 30.51. All three displacements carry m and w is assumed small enough that the strain Definition 30.6 may be linearised.

Rests on Definition 30.6 and Proposition 30.51.

Lemma A.681 (Strains under the Kirchhoff hypothesis)**.** *Under Equation* (A.1091) *the in-plane strains are proportional to the curvatures of the deflected mid-surface and the transverse shear strains vanish identically,*

$$u_{\alpha\beta} = -z\,\partial_\alpha\partial_\beta w\,, \qquad u_{\alpha z} = 0\,, \tag{A.1092}$$

while u_{zz} is not *determined by the hypothesis and is fixed instead by the constitutive statement of Appendix A.57.2.*

Rests on Definitions 30.6 and A.680.

Proof. *Derives Lemma A.681.*

From Definition 30.6, $u_{\alpha\beta} = \frac{1}{2}\left(\partial_\alpha u_\beta + \partial_\beta u_\alpha\right) = -\frac{1}{2}z\left(\partial_\alpha\partial_\beta w + \partial_\beta\partial_\alpha w\right) = -z\,\partial_\alpha\partial_\beta w$, the two mixed partials being equal for a twice continuously differentiable w. For the transverse shear,

$$u_{\alpha z} = \tfrac{1}{2}\left(\partial_\alpha u_z + \partial_z u_\alpha\right) = \tfrac{1}{2}\left(\partial_\alpha w - \partial_\alpha w\right) = 0\,. \tag{A.1093}$$

That the shear vanishes is exactly the statement that normals stay normal, and it is the hypothesis rather than a consequence: a real plate carries a transverse shear stress, and *must*, since the transverse load has to reach the supports somehow. What Equation (A.1093) asserts is that the shear *strain* it produces is negligible against the bending strain, which is true to relative order $(h/L)^2$ for a plate of lateral dimension L.

The remaining component, $u_{zz} = \partial_z u_z$, would be zero if Equation (A.1091) were read as an exact statement about every material point. It must not be so read. The ansatz is a statement about the in-plane displacement field and about the rotation of normals; the thickness of the plate does change, by an amount of the same order as the in-plane strains, and u_{zz} is left free here so that the correct constitutive condition can determine it. □

A.57.2 Plane stress, not plane strain

This is the step at which a plate differs from a two-dimensional block of material, and it is where the factor $\left(1-\nu^2\right)^{-1}$ of Equation (30.57) is born.

Proposition A.682 (The plane-stress law)**.** *Let the faces $z = \pm h/2$ be free of traction. Then throughout a thin plate σ_{zz} is negligible, and eliminating u_{zz} from Equation* (30.21) *by $\sigma_{zz} = 0$ gives*

$$u_{zz} = -\frac{\lambda}{\lambda+2\mu}\,u_{\gamma\gamma}\,, \qquad \sigma_{\alpha\beta} = \frac{E}{1+\nu}\left[u_{\alpha\beta} + \frac{\nu}{1-\nu}\,u_{\gamma\gamma}\,\delta_{\alpha\beta}\right]. \tag{A.1094}$$

Both λ,μ and E,ν are the constants of Definition 30.31, related by Proposition 30.33, and every quantity in Equation (A.1094) *except ν carries* Pa.

Rests on Equation (30.21)*, Proposition 30.33, and Definition 30.63.*

Proof. *Derives Proposition A.682.*

Why σ_{zz} and not u_{zz} vanishes. The traction on the faces is $\sigma_{iz}n_z$ with $n_z = \pm 1$, so $\sigma_{zz} = 0$ exactly on $z = \pm h/2$. Between the faces σ_{zz} is of the order of the applied load q, whereas the in-plane stresses are of order qL^2/h^2, larger by the square of the aspect ratio; setting $\sigma_{zz} = 0$ throughout is therefore an error of relative order $(h/L)^2$, the same order

already discarded in Equation (A.1093). Imposing $u_{zz} = 0$ instead — plane strain — would be a different and *wrong* theory: it describes a slab confined between rigid platens, and it would put $\lambda + 2\mu$ where Equation (30.57) has $E/(1-\nu^2)$, overstating the flexural rigidity by $(\lambda + 2\mu)(1-\nu^2)/E = (1-\nu)^2/(1-2\nu)$, a factor 1.22 at $\nu = 0.3$.

The elimination. The zz component of Equation (30.21) reads $\sigma_{zz} = \lambda(u_{\gamma\gamma} + u_{zz}) + 2\mu u_{zz}$, where $u_{\gamma\gamma} = u_{xx} + u_{yy}$ is the in-plane trace. Setting this to zero gives the first of Equation (A.1094). Substituting it back into the in-plane components,

$$\sigma_{\alpha\beta} = \lambda(u_{\gamma\gamma} + u_{zz})\,\delta_{\alpha\beta} + 2\mu u_{\alpha\beta} = \frac{2\lambda\mu}{\lambda + 2\mu}\,u_{\gamma\gamma}\delta_{\alpha\beta} + 2\mu\,u_{\alpha\beta}\,, \tag{A.1095}$$

since $\lambda[1 - \lambda/(\lambda+2\mu)] = 2\lambda\mu/(\lambda+2\mu)$.

In engineering constants. By Equation (30.28), $\lambda = 2\mu\nu/(1-2\nu)$ and $2\mu = E/(1+\nu)$, so

$$\lambda + 2\mu = 2\mu\,\frac{\nu + (1-2\nu)}{1-2\nu} = 2\mu\,\frac{1-\nu}{1-2\nu}\,, \qquad \frac{2\lambda\mu}{\lambda+2\mu} = 2\mu\,\frac{\nu}{1-\nu}\,. \tag{A.1096}$$

Substituting Equation (A.1096) into Equation (A.1095) and pulling out $2\mu = E/(1+\nu)$ gives the second of Equation (A.1094). □

Remark A.683 (Where the factor $1-\nu^2$ comes from). Contracting the second of Equation (A.1094) for a state of pure uniaxial in-plane *strain*, $u_{xx} = \epsilon$ and $u_{yy} = 0$, gives $\sigma_{xx} = [E/(1+\nu)][1 + \nu/(1-\nu)]\,\epsilon = E\epsilon/(1-\nu^2)$: the plate is stiffer than a bar of the same material by $(1-\nu^2)^{-1}$, because its own width prevents the transverse contraction that a bar is free to make. That is the sentence Definition 30.63 states in words, and it is the whole difference between Equation (30.57) and the EI of Equation (30.49).

Rests on Proposition A.682 and Definition 30.63.

A.57.3 Integrating the energy across the thickness

Proposition A.684 (The plate energy)**.** *Under Definition A.680 and Proposition A.682, the elastic energy of a plate occupying the region Ω of the mid-plane is*

$$U[w] = \frac{D}{2}\int_\Omega \left[\left(\nabla^2 w\right)^2 + 2(1-\nu)\left((\partial_x\partial_y w)^2 - \partial_x^2 w\,\partial_y^2 w\right)\right]\mathrm{d}A\,, \tag{A.1097}$$

with $D = Eh^3/[12(1-\nu^2)]$ the flexural rigidity Equation (30.57), of SI dimension N m*; U is then in* J*.*

Rests on Proposition A.682, Lemma A.681, and Definition 30.63.

Proof. *Derives Proposition A.684.*

The density. By Equation (30.22) and Equation (30.19) the energy density is $F = \frac{1}{2}\sigma_{ik}u_{ik}$. Of the nine terms, $\sigma_{\alpha z}u_{\alpha z}$ vanishes because the strain does (Equation (A.1093)) and $\sigma_{zz}u_{zz}$ vanishes because the stress does (Proposition A.682), so only the in-plane block survives:

$$F = \tfrac{1}{2}\sigma_{\alpha\beta}u_{\alpha\beta} = \frac{E}{2(1+\nu)}\left[u_{\alpha\beta}u_{\alpha\beta} + \frac{\nu}{1-\nu}\,(u_{\gamma\gamma})^2\right]. \tag{A.1098}$$

Note that the plate stores no energy in the eliminated component: the work done by σ_{zz} through u_{zz} is zero because σ_{zz} is zero, which is the reason the elimination costs nothing.

Through the thickness. Insert Equation (A.1092). Every strain carries one factor of z, so every term of Equation (A.1098) carries z^2, and

$$\int_{-h/2}^{h/2} z^2\,\mathrm{d}z = \frac{h^3}{12}\,. \tag{A.1099}$$

Writing $w_{,\alpha\beta} = \partial_\alpha \partial_\beta w$ and abbreviating

$$P = w_{,\alpha\beta} w_{,\alpha\beta} = \left(\partial_x^2 w\right)^2 + 2\left(\partial_x \partial_y w\right)^2 + \left(\partial_y^2 w\right)^2 , \qquad Q = \left(\nabla^2 w\right)^2 , \tag{A.1100}$$

the integrated energy is

$$U = \frac{h^3}{12} \frac{E}{2\,(1+\nu)} \int_\Omega \left[P + \frac{\nu}{1-\nu} Q \right] \mathrm{d}A \,. \tag{A.1101}$$

The rearrangement. Expanding Q and comparing with P,

$$P - Q = 2\left(\partial_x \partial_y w\right)^2 - 2\,\partial_x^2 w\,\partial_y^2 w \,, \tag{A.1102}$$

because $Q = \left(\partial_x^2 w\right)^2 + 2\,\partial_x^2 w\,\partial_y^2 w + \left(\partial_y^2 w\right)^2$. Write $2G$ for the right-hand side of Equation (A.1102), so that $P = Q + 2G$ with $G = \left(\partial_x \partial_y w\right)^2 - \partial_x^2 w\,\partial_y^2 w$. Then the bracket of Equation (A.1101) is

$$P + \frac{\nu}{1-\nu} Q = \left(1 + \frac{\nu}{1-\nu}\right) Q + 2G = \frac{Q + 2\,(1-\nu)\,G}{1-\nu} \,, \tag{A.1103}$$

and the prefactor of Equation (A.1101) combines with the $(1-\nu)^{-1}$ to give

$$\frac{h^3}{12} \frac{E}{2\,(1+\nu)\,(1-\nu)} = \frac{1}{2} \frac{E h^3}{12\,(1-\nu^2)} = \frac{D}{2} \,, \tag{A.1104}$$

which is Equation (A.1097). □

Lemma A.685 (The Gaussian-curvature term is a null Lagrangian)**.** *The integrand* $G = \left(\partial_x \partial_y w\right)^2 - \partial_x^2 w\,\partial_y^2 w$ *of Equation* (A.1097) *is an exact divergence,*

$$G = \partial_x \left(\partial_y w\,\partial_x \partial_y w\right) - \partial_y \left(\partial_y w\,\partial_x^2 w\right) \,, \tag{A.1105}$$

so that $\int_\Omega G\,\mathrm{d}A$ *depends only on* w *and its first derivatives on* $\partial\Omega$*. It therefore contributes nothing to the field equation, and nothing at all when the edge is clamped.*

Rests on Proposition A.684.

Proof. *Derives Lemma A.685.*

Differentiate the two products: $\partial_x \left(\partial_y w\,\partial_x \partial_y w\right) = \left(\partial_x \partial_y w\right)^2 + \partial_y w\,\partial_x^2 \partial_y w$ and $\partial_y \left(\partial_y w\,\partial_x^2 w\right) = \partial_y^2 w\,\partial_x^2 w + \partial_y w\,\partial_y \partial_x^2 w$. The two third-derivative terms are equal and cancel in the difference, leaving G. By Theorem 7.101 in the plane, $\int_\Omega G\,\mathrm{d}A = \oint_{\partial\Omega} \left(\partial_y w\,\partial_x \partial_y w\,n_x - \partial_y w\,\partial_x^2 w\,n_y\right) \mathrm{d}s$, an expression in the boundary values of w and ∇w alone. A variation δw vanishing together with $\nabla \delta w$ on $\partial\Omega$ therefore does not change it, so it cannot enter the Euler equation; and for a clamped edge, where w and $\partial_n w$ are prescribed, the whole term is a constant that may be discarded from the functional. □

Remark A.686 (Why the operator comes out as ∇^4). Lemma A.685 is what makes the plate equation as simple as it is, and a reader who does not see it will suspect an error: the energy Equation (A.1097) is manifestly *not* proportional to $\int (\nabla^2 w)^2$, it depends on Poisson's ratio, and yet the field equation contains only the biharmonic operator and no ν at all. The resolution is that the ν-dependent part is a boundary functional. Geometrically, $-G$ is the Gaussian curvature of the deflected mid-surface to leading order in ∇w — the determinant of the Hessian — so Equation (A.1105) is the small-deflection shadow of the Gauss–Bonnet theorem: the total Gaussian curvature of a surface is determined by its boundary. Poisson's ratio does not disappear from the problem; it survives exactly where the boundary functional lives, in the edge conditions of Appendix A.57.5, which is why two plates of the same D and different ν have the same equation and different free-edge behaviour.

Rests on Lemma A.685 and Definition 30.63.

A.57.4 The field equation

Lemma A.687 (Kinetic energy)**.** *To leading order in h/L the kinetic energy of the plate is*

$$T[\dot w] = \frac{\rho h}{2}\int_\Omega \dot w^2 \,\mathrm{d}A\,, \tag{A.1106}$$

with ρ the density in $\mathrm{kg/m^3}$*, so that ρh is a mass per unit area in* $\mathrm{kg/m^2}$*. The discarded term is the rotatory inertia of the cross-sections, $(\rho h^3/24)\int_\Omega |\nabla\dot w|^2\,\mathrm{d}A$.*

Rests on Definition A.680.

Proof. *Derives Lemma A.687.*

Differentiating Equation (A.1091) in time gives $\dot u_z = \dot w$ and $\dot u_\alpha = -z\,\partial_\alpha \dot w$, so

$$T = \frac{\rho}{2}\int_\Omega\int_{-h/2}^{h/2}\left(\dot w^2 + z^2\,|\nabla\dot w|^2\right)\mathrm{d}z\,\mathrm{d}A = \frac{\rho h}{2}\int_\Omega \dot w^2\mathrm{d}A + \frac{\rho h^3}{24}\int_\Omega |\nabla\dot w|^2\,\mathrm{d}A\,, \tag{A.1107}$$

using Equation (A.1099). For a disturbance of lateral scale L one has $|\nabla\dot w|^2 \sim \dot w^2/L^2$, so the second term is smaller than the first by $(h/L)^2/12$, the order already discarded twice above, and is dropped. Retaining it, together with the transverse shear strain discarded in Equation (A.1093), gives the Mindlin–Reissner theory, whose flexural waves have a finite limiting speed instead of the unbounded $\omega \propto k^2$ that Equation (30.58) predicts; the two agree for $kh \ll 1$, which is the regime in which the plate theory is being used. □

Theorem A.688 (The Kirchhoff plate equation)**.** *Let w extremise the action*

$$S[w] = \int_{t_1}^{t_2}\left(T[\dot w] - U[w] + \int_\Omega q\,w\,\mathrm{d}A\right)\mathrm{d}t\,, \tag{A.1108}$$

built from Equation (A.1097) and Equation (A.1106), with q the transverse load per unit area in Pa*. Then in the interior of Ω*

$$\rho h\,\partial_t^2 w + D\,\nabla^4 w = q\,, \tag{A.1109}$$

which is Equation (30.58), and the variation leaves on $\partial\Omega$ the boundary integral Equation (A.1115) below.

Rests on Proposition A.684 and Lemmas 16.18 and A.687.

Proof. *Derives Theorem A.688.*

Moments. It shortens every formula to introduce the bending moments per unit length of edge,

$$M_{\alpha\beta} = -D\left[(1-\nu)\,\partial_\alpha\partial_\beta w + \nu\,\nabla^2 w\,\delta_{\alpha\beta}\right]\,, \tag{A.1110}$$

of SI dimension $\mathrm{N\,m/m} = \mathrm{N}$. With the notation Equation (A.1100), the energy Equation (A.1097) is

$$U = \frac{D}{2}\int_\Omega\left[(1-\nu)\,P + \nu Q\right]\mathrm{d}A = -\frac{1}{2}\int_\Omega M_{\alpha\beta}\,\partial_\alpha\partial_\beta w\,\mathrm{d}A\,, \tag{A.1111}$$

the first equality being Equation (A.1103) read backwards, $Q + 2(1-\nu)G = (1-\nu)P + \nu Q$.

The variation. U is a quadratic form in the second derivatives of w and $M_{\alpha\beta}$ is linear in them, so $\delta U = -\int_\Omega M_{\alpha\beta}\,\partial_\alpha\partial_\beta\delta w\,\mathrm{d}A$. Integrating by parts twice with Theorem 7.101,

$$\delta U = -\int_\Omega\left(\partial_\alpha\partial_\beta M_{\alpha\beta}\right)\delta w\,\mathrm{d}A - \oint_{\partial\Omega} M_{\alpha\beta}n_\alpha\,\partial_\beta\delta w\,\mathrm{d}s + \oint_{\partial\Omega}\left(\partial_\alpha M_{\alpha\beta}\right)n_\beta\,\delta w\,\mathrm{d}s\,, \tag{A.1112}$$

with $\boldsymbol{n}$ the outward unit normal of the edge. Taking the double divergence of Equation (A.1110),

$$\partial_\alpha\partial_\beta M_{\alpha\beta} = -D\left[(1-\nu)\,\nabla^4 w + \nu\,\nabla^4 w\right] = -D\,\nabla^4 w\,, \tag{A.1113}$$

the ν cancelling exactly as Remark A.686 promised. The kinetic term contributes $\delta \int T \mathrm{d}t = -\int \mathrm{d}t \int_\Omega \rho h\, \partial_t^2 w\, \delta w \,\mathrm{d}A$ after one integration by parts in t, the endpoint terms vanishing because δw is held zero at t_1 and t_2. Collecting,

$$\delta S = -\int \mathrm{d}t \left\{ \int_\Omega \left(\rho h\, \partial_t^2 w + D \nabla^4 w - q \right) \delta w \,\mathrm{d}A + \mathcal{B}\left[\delta w\right] \right\} , \tag{A.1114}$$

where $\mathcal{B}$ is the edge integral

$$\mathcal{B}\left[\delta w\right] = -\oint_{\partial\Omega} M_{\alpha\beta} n_\alpha\, \partial_\beta \delta w \,\mathrm{d}s + \oint_{\partial\Omega} \left(\partial_\alpha M_{\alpha\beta}\right) n_\beta\, \delta w \,\mathrm{d}s \,. \tag{A.1115}$$

For variations supported strictly inside Ω, $\mathcal{B}$ vanishes and Lemma 16.18 applied to the area integral gives Equation (A.1109). Since D carries N m and $\nabla^4 w$ carries $1/\mathrm{m}^3$, each term of Equation (A.1109) is a pressure, as q is. □

A.57.5 The free edge: three conditions are one too many

Definition A.689 (Edge resultants)**.** Let $\boldsymbol{n}$ and $\boldsymbol{s}$ be the outward unit normal and the unit tangent of $\partial\Omega$, forming a right-handed pair, and let $\partial_n = n_\alpha \partial_\alpha$ and $\partial_s = s_\alpha \partial_\alpha$. Define

$$M_{nn} = M_{\alpha\beta} n_\alpha n_\beta \,, \qquad \text{bending moment, N}\,, \tag{A.1116}$$

$$M_{ns} = M_{\alpha\beta} n_\alpha s_\beta \,, \qquad \text{twisting moment, N}\,, \tag{A.1117}$$

$$Q_n = \left(\partial_\alpha M_{\alpha\beta}\right) n_\beta \,, \qquad \text{transverse shear, N/m}\,. \tag{A.1118}$$

Explicitly, from Equation (A.1110), $M_{nn} = -D\left[\partial_n^2 w + \nu\, \partial_s^2 w\right]$ for a straight edge, and $M_{ns} = -D\,(1-\nu)\,\partial_n \partial_s w$, the trace term dropping out of the second because $\delta_{\alpha\beta} n_\alpha s_\beta = \boldsymbol{n}\cdot\boldsymbol{s} = 0$: the twisting moment measures the *twist* $\partial_n \partial_s w$ of the mid-surface at the edge and nothing else.

Rests on Equation (A.1110) *and Proposition A.682.*

Remark A.690 (Poisson's count, and why it cannot be right). Read as a force balance, a free edge — one at which nothing is attached — carries no bending moment, no twisting moment and no transverse shear, which is three conditions: $M_{nn} = 0$, $M_{ns} = 0$, $Q_n = 0$. This is Poisson's count, and Remark 30.67 records that it was believed for two decades. It cannot be right. Equation (A.1109) is of fourth order in the two space variables, so along each edge exactly *two* conditions may be imposed; three over-determine the problem and in general leave it with no solution at all. The resolution is not to discard one of the three arbitrarily but to observe that they are not independent, which is what the variational form Equation (A.1115) makes visible and a force balance cannot.

Rests on Definition A.689 and Theorem A.688.

Theorem A.691 (Kirchhoff's edge conditions and the corner force)**.** *Let $\partial\Omega$ be a closed piecewise-smooth curve with corners $c_1, \dots, c_N$. Then the edge integral Equation* (A.1115) *may be written*

$$\mathcal{B}\left[\delta w\right] = -\oint_{\partial\Omega} M_{nn}\, \partial_n \delta w \,\mathrm{d}s + \oint_{\partial\Omega} V_n\, \delta w \,\mathrm{d}s + \sum_{j=1}^{N} F_j\, \delta w\left(c_j\right) \,, \tag{A.1119}$$

with the effective *(Kirchhoff) shear and the corner force*

$$V_n = Q_n + \partial_s M_{ns} \quad (\mathrm{N/m}) \,, \qquad F_j = \left[\!\left[M_{ns} \right]\!\right]_{c_j} \quad (\mathrm{N}) \,, \tag{A.1120}$$

where $\left[\!\left[M_{ns} \right]\!\right]_{c_j}$ is the jump in M_{ns} across the corner. Consequently, at an edge where nothing is prescribed the natural boundary conditions (Theorem 16.31) are the two *statements*

$$M_{nn} = 0 \,, \qquad V_n = Q_n + \partial_s M_{ns} = 0 \,, \tag{A.1121}$$

together with the vanishing of F_j at every free corner. These are the conditions asserted by Theorem 30.64.

Rests on Theorems 16.31 and A.688 and Definition A.689.

Proof. *Derives Theorem A.691.*

Resolving the normal derivative. Along the edge, $\{\boldsymbol{n}, \boldsymbol{s}\}$ is an orthonormal frame of the plane, so $\partial_\beta \delta w = n_\beta\, \partial_n \delta w + s_\beta\, \partial_s \delta w$ and

$$M_{\alpha\beta} n_\alpha \partial_\beta \delta w = M_{nn}\, \partial_n \delta w + M_{ns}\, \partial_s \delta w\,. \tag{A.1122}$$

Substituting Equation (A.1122) and Equation (A.1118) into Equation (A.1115) gives the three-term form

$$\mathcal{B}\,[\delta w] = \oint_{\partial\Omega} \left[-M_{nn}\, \partial_n \delta w - M_{ns}\, \partial_s \delta w + Q_n\, \delta w\right] \mathrm{d}s\,, \tag{A.1123}$$

which is where Poisson's three conditions appear to come from.

The three are two. The quantities δw and $\partial_n \delta w$ may be prescribed independently on the edge, but $\partial_s \delta w$ may not: it is the derivative of δw *along* the edge, and is determined by δw there. Integrate that term by parts along the curve. On each smooth arc,

$$-\int_{\text{arc}} M_{ns}\, \partial_s \delta w \,\mathrm{d}s = \int_{\text{arc}} (\partial_s M_{ns})\, \delta w \,\mathrm{d}s - \Big[M_{ns}\, \delta w\Big]_{\text{start}}^{\text{end}}\,. \tag{A.1124}$$

Summing Equation (A.1124) over the arcs of the closed curve, the endpoint terms of consecutive arcs meet at each corner c_j and combine into $-M_{ns}^{-}\delta w(c_j) + M_{ns}^{+}\delta w(c_j)$, where $\pm$ denote the limits taken along the outgoing and incoming arc; that is $F_j\, \delta w(c_j)$ with F_j the jump of Equation (A.1120). On a smooth closed curve, M_{ns} is continuous and every such term cancels. Collecting with the $Q_n \delta w$ term of Equation (A.1123) gives Equation (A.1119).

The conditions. In Equation (A.1119) the three factors $\partial_n \delta w$ on the edge, δw on the edge and δw at each corner are now independently assignable. By Lemma 16.18 applied on the edge and then at the finitely many corners, $\delta S = 0$ for all admissible variations requires each coefficient to vanish wherever its variation is free. Where the plate is clamped, $\delta w = \partial_n \delta w = 0$ and nothing is required of M_{nn} or V_n; where it is simply supported, $\delta w = 0$ but $\partial_n \delta w$ is free, giving the single natural condition $M_{nn} = 0$; where it is free, both are free and Equation (A.1121) follows, with $F_j = 0$ at a free corner. In every case the count is two conditions per edge, as the order of Equation (A.1109) requires. □

Remark A.692 (The corner force is real, and you can feel it). F_j is not an artefact of the variational bookkeeping. Take a square plate, simply supported on all four edges — resting on knife edges, free to rotate — and load it uniformly. Along each edge $\delta w = 0$, so the edge integral of Equation (A.1119) imposes only $M_{nn} = 0$; but at a corner the tangent turns through a right angle, the roles of n and s are exchanged, and M_{ns} changes sign, so $F_j = [\![M_{ns}]\!] = 2M_{ns}$ does not vanish. A downward concentrated force of that magnitude must be supplied at each corner. If the supports cannot pull down — if the plate merely rests on them — the corners lift off, which is exactly what a uniformly loaded square glass pane does, and why a concrete floor slab supported on four walls is reinforced diagonally at its corners. The corner force also explains the sign of the effect: the plate is attempting to become anticlastic, curving up along one diagonal while it curves down along the other, and it is M_{ns} that measures that twist.

Rests on Theorem A.691 and Equation (A.1120).

Remark A.693 (Attribution, and what the hypotheses cost). The fourth-order equation is Germain's [70] and the two correct edge conditions are Kirchhoff's [130], obtained exactly as above, from the variation of the energy rather than from a force balance; Poisson's three-condition count is in his memoir on elastic bodies [86]. Three approximations were

made and each is of relative order $(h/L)^2$: the transverse shear strain was set to zero (Equation (A.1093)), the transverse normal stress was set to zero (Proposition A.682), and the rotatory inertia was dropped (Lemma A.687). Two of the three are removed by the Mindlin–Reissner theory, and none of them is a small-deflection assumption: the linearisation of the strain in Definition A.680 is a separate hypothesis, and it is the one that fails first for a thin plate, since a deflection of order h already stretches the mid-surface and brings in the membrane terms of Remark 30.68.

Remark A.694. Appendix A.57 discharges the derivation owed at Theorem 30.64 of Chapter 30. The three stages announced in that chapter are Appendix A.57.2 (plane stress, and the birth of the factor $1-\nu^2$ in Equation (30.57)), Appendix A.57.3 together with Appendix A.57.4 (the thickness integration and the variation, giving Equation (30.58)), and Appendix A.57.5 (Kirchhoff's resolution of the edge count, with the concentrated corner force). Everything the chapter builds on Equation (30.58) — the Chladni figures of Phenomenon 30.65, the scaling Equation (30.59), and the historical reading of Remark 30.67 — rests on the equation proved here, and the isotropic constitutive law it starts from is Equation (30.21), proved in Appendix A.56.

A.58 Hertz's Solution for the Elastic Half-Space Under an Axisymmetric Pressure

This appendix completes the derivation of Phenomenon 30.70 of Chapter 30. It is deliberately *not* a second derivation of that phenomenon. The inline derivation in the chapter already establishes the exponent, and establishes it exactly: from the geometry of two nearly touching spheres, the single strain scale of the contact region and the linearity of the elastic law it obtains $F \propto E^* R^{1/2} \delta^{3/2}$, with $a = \sqrt{R\delta}$, and nothing about that argument is a guess. What it cannot reach is the pure number in front and the shape of the pressure distribution, because both require actually solving the elastic problem. Those two things — the coefficient $\frac{4}{3}$ of Equation (30.60) and the elliptic profile $p(r) = p_0\sqrt{1-r^2/a^2}$ — are what this section supplies, together with a proof of the two-body combination rule for E^*, which the chapter's derivation asserts in a clause and proves nowhere.

A.58.1 What is quoted here

One result is imported, and everything else is derived from it.

Theorem A.695 (Boussinesq's point-force solution, quoted)**.** *Let an isotropic elastic half-space $z \geq 0$ with Young's modulus E and Poisson's ratio ν (Definition 30.31) be loaded by a normal force P concentrated at the origin of its free surface, the rest of the surface being free of traction and the displacement vanishing at infinity. Then the normal displacement of the surface at distance r from the load is*

$$u_z(r) = \frac{1-\nu^2}{\pi E}\,\frac{P}{r} = \frac{P}{\pi E^* r}\,, \qquad \frac{1}{E^*} \equiv \frac{1-\nu^2}{E}\,, \tag{A.1125}$$

directed into the half-space [1778]. Here P carries N*, E and E^* carry* Pa *and u_z carries* m.
Rests on Equation (30.21), Definition 30.31, and Theorem 30.18.

Remark A.696 (What is quoted here). Theorem A.695 is Boussinesq's, published in 1885, and it is the one step of this section that is not carried out from the material of this treatise. Obtaining it needs a general representation of the solutions of the Navier–Cauchy equations of elastostatics in terms of harmonic functions — the Papkovich–Neuber representation — and the potential theory of Section 10.4.3 carries the harmonic functions but not that representation. It is therefore a debt owed by Chapter 10, and it is named here rather than

passed over; the statement above is proved in Landau and Lifshitz's treatment of the elastic half-space [1778], from which the form Equation (A.1125) is taken.

Two features of Equation (A.1125) can be checked here without the representation, and are worth checking, because between them they fix everything except the pure number $1/\pi$. *Dimensions*: $P/(E^* r)$ is $\mathrm{N}/(\mathrm{Pa}\cdot\mathrm{m}) = \mathrm{m}$, as a displacement must be. *Scale invariance*: the problem has no length in it at all — a half-space is scale-invariant, a point load introduces no length, and linear elasticity has no intrinsic scale — so u_z must be a homogeneous function of r, and linearity in P together with the dimensions forces the exponent -1. The modulus combination $E/(1-\nu^2)$ appears rather than E because the material beneath a surface load is laterally confined, which is the same statement as the one made for a plate in Remark A.683.

A.58.2 Superposition and the contact conditions

Lemma A.697 (Surface displacement under a distributed pressure)**.** *Let a pressure $p(\rho) \geq 0$ act on the disc $\rho \leq a$ of the free surface of the half-space, and vanish outside it. Then the surface normal displacement is*

$$u_z(r) = \frac{1}{\pi E^*}\,\Phi(r)\,, \qquad \Phi(r) = \int_{\rho\leq a} \frac{p\,(\rho)}{s}\,\mathrm{d}A'\,, \tag{A.1126}$$

where s is the distance in the surface plane from the field point to the element $\mathrm{d}A'$. The integral converges although the kernel is singular, because $\mathrm{d}A' = s\,\mathrm{d}s\,\mathrm{d}\varphi$ in polar coordinates centred on the field point.

Rests on Theorem A.695.

Proof. *Derives Lemma A.697.*

The elastic problem is linear (Equation (30.21)) and the boundary conditions are linear in the applied traction, so displacements superpose. Treating the load on the element $\mathrm{d}A'$ as a point force $P = p\,\mathrm{d}A'$ and summing Equation (A.1125) gives Equation (A.1126). In polar coordinates (s,φ) about the field point the element is $s\,\mathrm{d}s\,\mathrm{d}\varphi$, so the integrand is $p\,\mathrm{d}s\,\mathrm{d}\varphi$ with no singularity at all; the integral exists for any bounded p. □

Definition A.698 (The Hertz contact problem)**.** Two elastic bodies of radii R_1 and R_2, with moduli (E_1,ν_1) and (E_2,ν_2), are pressed together along their line of centres by a normal force F, their centres approaching by δ. Let a be the radius of the circle of contact and $p(r)$ the contact pressure. The problem is to find p, a and the relation between F and δ, subject to

1. *contact inside the circle*: for $r \leq a$ the sum of the two surface displacements closes the initial gap,
$$u_z^{(1)}(r) + u_z^{(2)}(r) = \delta - \frac{r^2}{2R}\,, \qquad \frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}\,, \tag{A.1127}$$
2. *separation outside*: $p(r) = 0$ for $r > a$, and the surfaces there do not overlap;
3. *no adhesion and no edge singularity*: $p(r) \geq 0$ throughout, and p is bounded as $r \to a$.

The contact is taken to be *frictionless*, so that the traction transmitted across it is purely normal; and $a \ll R$, so that near the contact each body is indistinguishable from a half-space and Theorem A.695 applies to it.

Rests on Theorem A.695 and Phenomenon 30.70.

Remark A.699 (The third condition is not decoration). Conditions (1) and (2) alone do not determine the answer. The same geometry admits the flat-punch solution $p(r) \propto (1 - r^2/a^2)^{-1/2}$, which is a perfectly good solution of the elastic equations, produces a

constant surface displacement inside the circle, and diverges at the rim; it is the right answer for a rigid flat cylindrical punch pressed into a half-space, where the contact radius is fixed by the punch and the material at the rim is genuinely singular. What excludes it here is condition (3): the contact circle of two smooth bodies is not fixed externally — it is whatever radius the load produces — so the surfaces must meet the rim tangentially and the pressure must go to zero there, not to infinity. Condition (3) also excludes the tensile solutions that would hold the surfaces together outside the circle, which is the assumption of *no adhesion*; it fails for very small or very compliant bodies, where surface energy matters, and Hertz's relation is then replaced by an adhesive contact theory. Stating the third condition is what turns the profile below from a lucky guess into the unique answer.

Rests on Definition A.698.

A.58.3 The elliptic pressure and its displacement

The whole solution rests on one integral, and it is worth doing in full.

Lemma A.700 (The displacement produced by an elliptic pressure)**.** *Let* $p(\rho) = p_0\sqrt{1-\rho^2/a^2}$ *on* $\rho \le a$ *and zero outside. Then for every field point inside the circle,* $r \le a$*,*

$$\Phi(r) = \frac{\pi^2 p_0}{4a}\left(2a^2 - r^2\right), \qquad \textit{hence} \qquad u_z(r) = \frac{\pi p_0}{4E^* a}\left(2a^2 - r^2\right). \tag{A.1128}$$

The displacement is therefore quadratic *in* r *inside the contact circle — which is exactly the shape Equation* (A.1127) *demands.*

Rests on Lemma A.697 and Definition A.698.

Proof. *Derives Lemma A.700.*

Place the field point at $(r, 0)$ with $0 \le r \le a$ and use polar coordinates (s, φ) centred on it, so that the element at (s, φ) sits at $(r + s\cos\varphi,\ s\sin\varphi)$ and is at distance

$$\rho^2 = r^2 + 2rs\cos\varphi + s^2, \tag{A.1129}$$

measured from the centre of the disc. By Lemma A.697 the kernel cancels the Jacobian, so

$$\Phi(r) = \int_0^{2\pi} \mathrm{d}\varphi \int_0^{\ell(\varphi)} p\,(\rho)\,\mathrm{d}s = \frac{p_0}{a}\int_0^{2\pi} \mathrm{d}\varphi \int_0^{\ell(\varphi)} \sqrt{a^2 - r^2 - 2rs\cos\varphi - s^2}\,\mathrm{d}s, \tag{A.1130}$$

where $\ell(\varphi)$ is the distance from the field point to the rim along the direction φ, that is the value of s at which $\rho = a$.

Completing the square. Put

$$u = s + r\cos\varphi, \qquad A^2 = a^2 - r^2\sin^2\varphi, \qquad a^2 - r^2 - 2rs\cos\varphi - s^2 = A^2 - u^2, \tag{A.1131}$$

which is checked by expanding $A^2 - u^2$. The rim, $\rho = a$, is $u = A$; the field point, $s = 0$, is $u = r\cos\varphi$. Since u and s differ by a constant at fixed φ, $\mathrm{d}u = \mathrm{d}s$, and

$$\int_0^{\ell(\varphi)} \sqrt{A^2 - u^2}\,\mathrm{d}s = \int_{r\cos\varphi}^{A} \sqrt{A^2 - u^2}\,\mathrm{d}u = \left[\frac{u}{2}\sqrt{A^2 - u^2} + \frac{A^2}{2}\arcsin\frac{u}{A}\right]_{u=r\cos\varphi}^{u=A}, \tag{A.1132}$$

the antiderivative being elementary. At the upper limit the first term vanishes and $\arcsin 1 = \pi/2$, giving $\pi A^2/4$. At the lower limit,

$$A^2 - r^2\cos^2\varphi = a^2 - r^2\sin^2\varphi - r^2\cos^2\varphi = a^2 - r^2, \tag{A.1133}$$

a constant independent of φ — the geometric fact that makes the calculation work. Writing $B = \sqrt{a^2 - r^2}$, the inner integral is

$$\frac{\pi A^2}{4} - \frac{rB\cos\varphi}{2} - \frac{A^2}{2}\arcsin\frac{r\cos\varphi}{A}\,. \tag{A.1134}$$

The angular integral. Integrate Equation (A.1134) over $\varphi \in [0, 2\pi)$ term by term. The second term integrates to zero, $\int_0^{2\pi}\cos\varphi\,\mathrm{d}\varphi = 0$. The third also integrates to zero: under $\varphi \mapsto \pi - \varphi$ the quantity A is unchanged, since it depends on $\sin^2\varphi$, while $\cos\varphi$ changes sign and arcsin is odd, so the integrand is odd about that involution. The first term gives

$$\frac{\pi}{4}\int_0^{2\pi}\left(a^2 - r^2\sin^2\varphi\right)\mathrm{d}\varphi = \frac{\pi}{4}\left(2\pi a^2 - \pi r^2\right) = \frac{\pi^2}{4}\left(2a^2 - r^2\right)\,, \tag{A.1135}$$

using $\int_0^{2\pi}\sin^2\varphi\,\mathrm{d}\varphi = \pi$. Multiplying by the prefactor p_0/a of Equation (A.1130) gives the first of Equation (A.1128), and Equation (A.1126) then gives the second. □

A.58.4 Matching, and the coefficient $\frac{4}{3}$

Proposition A.701 (The two-body compliance rule)**.** *Under Definition A.698 the two bodies carry the* same *contact pressure $p(r)$, so their surface displacements add, and*

$$u_z^{(1)}(r) + u_z^{(2)}(r) = \frac{\Phi(r)}{\pi E^*}\,, \qquad \frac{1}{E^*} = \frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2}\,, \tag{A.1136}$$

which is the combination appearing in Equation (30.60). The geometry likewise combines: the gap between the two undeformed surfaces is $r^2/2R$ with $1/R = 1/R_1 + 1/R_2$, so the pair is equivalent to a single sphere of radius R pressed against a rigid plane by the same force.
Rests on Lemma A.697 and Definition A.698.

Proof. *Derives Proposition A.701.*

The pressures are equal. By Newton's third law the traction each body exerts on the other is equal and opposite at every point of the common contact area, so the same scalar $p(r)$ appears in Equation (A.1126) for both. Applying that formula to each body with its own modulus and adding,

$$u_z^{(1)} + u_z^{(2)} = \left(\frac{1-\nu_1^2}{\pi E_1} + \frac{1-\nu_2^2}{\pi E_2}\right)\Phi(r)\,, \tag{A.1137}$$

which is Equation (A.1136). It is the *compliances* $1/E_i^*$ that add and not the moduli, because the two bodies are loaded in series by a common force and their deflections accumulate; that is the clause the chapter's inline derivation asserts without proof, and Equation (A.1137) is the proof. A rigid body has $E_i \to \infty$ and contributes nothing, so a single elastic sphere against a rigid plane has $E^* = E/(1-\nu^2)$.

The geometry. Measure heights from the plane of first contact. To leading order in r/R_i the surface of a sphere of radius R_i lies at height $r^2/2R_i$ below its own pole — the expansion $R_i - \sqrt{R_i^2 - r^2} = r^2/2R_i + O(r^4/R_i^3)$ — so the gap between the two undeformed surfaces at distance r from the axis is $r^2/2R_1 + r^2/2R_2 = r^2/2R$. Both the elastic response and the geometry therefore depend on the pair only through E^* and R. □

Theorem A.702 (Hertz's solution)**.** *The unique solution of Definition A.698 is the elliptic pressure distribution*

$$p(r) = p_0\sqrt{1 - \frac{r^2}{a^2}}\,, \qquad p_0 = \frac{2E^*a}{\pi R} = \frac{3F}{2\pi a^2}\,, \tag{A.1138}$$

with contact radius, approach and force related by

$$a = \sqrt{R\delta}\,, \qquad \delta = \frac{a^2}{R}\,, \qquad F = \frac{4}{3}\,E^* R^{1/2}\delta^{3/2}\,, \tag{A.1139}$$

which is Equation (30.60) complete with its coefficient.
Rests on Lemma A.700, Proposition A.701, and Definition A.698.

Proof. *Derives Theorem A.702.*

Matching. Insert Equation (A.1128) into Equation (A.1136) and set the result equal to the required displacement Equation (A.1127):

$$\frac{\pi p_0}{4E^* a}\left(2a^2 - r^2\right) = \delta - \frac{r^2}{2R} \qquad \text{for all } r \le a\,. \tag{A.1140}$$

Both sides are polynomials in r^2 of degree one, so Equation (A.1140) holds identically if and only if the constant terms and the coefficients of r^2 agree separately:

$$\frac{\pi p_0 a}{2E^*} = \delta\,, \qquad \frac{\pi p_0}{4E^* a} = \frac{1}{2R}\,. \tag{A.1141}$$

That an elliptic pressure produces a displacement of *exactly* the required quadratic form, with two free constants p_0 and a to match two coefficients, is what makes the problem solvable in closed form; it is the content of Lemma A.700 and it is why Hertz's guess was the right one.

Solving. The second of Equation (A.1141) gives $p_0 = 2E^* a/(\pi R)$, the first of Equation (A.1138). Substituting into the first of Equation (A.1141),

$$\delta = \frac{\pi a}{2E^*} \cdot \frac{2E^* a}{\pi R} = \frac{a^2}{R}\,, \tag{A.1142}$$

which is $a = \sqrt{R\delta}$, the contact radius quoted in Phenomenon 30.70.

The force. Integrating the pressure over the contact circle, with the substitution $t = r^2/a^2$,

$$F = \int_0^a p_0 \sqrt{1 - \frac{r^2}{a^2}}\, 2\pi r \,\mathrm{d}r = \pi a^2 p_0 \int_0^1 \sqrt{1-t}\,\mathrm{d}t = \frac{2}{3}\pi a^2 p_0\,, \tag{A.1143}$$

so $p_0 = 3F/(2\pi a^2)$: the peak pressure is $\frac{3}{2}$ times the mean pressure $F/\pi a^2$, which is the second equality of Equation (A.1138). Eliminating p_0 between Equation (A.1143) and Equation (A.1138), and then a by Equation (A.1142),

$$F = \frac{2}{3}\pi a^2 \cdot \frac{2E^* a}{\pi R} = \frac{4}{3}\,\frac{E^* a^3}{R} = \frac{4}{3}\,\frac{E^*\,(R\delta)^{3/2}}{R} = \frac{4}{3}\,E^* R^{1/2}\delta^{3/2}\,. \tag{A.1144}$$

Uniqueness and admissibility. The profile Equation (A.1138) is non-negative and vanishes at $r = a$, so it satisfies condition (3) of Definition A.698; outside the circle the deformed surfaces separate, since the elastic displacement there falls faster than the parabolic gap grows. That it is the *only* admissible solution follows from linearity together with the positive definiteness of the elastic energy. Suppose two admissible pressure distributions produced the same δ over the same contact circle. Their difference drives an elastic field whose surface traction is purely normal (the contact is frictionless), whose normal displacement vanishes inside the circle, and whose traction vanishes outside it. Its stored energy is $\frac{1}{2}\oint \sigma_{ik} n_k u_i \,\mathrm{d}S$ by Equation (30.18) with $w_i = u_i$ and no body force, and every term of that integral has one vanishing factor, so the energy is zero. Since the energy density is positive definite (Proposition 30.30), the difference field has zero strain and the two pressures coincide.

The coefficient $\frac{4}{3}$ of Equation (A.1144) is exactly and only what the scaling argument of the chapter left open, and it is assembled from three pure numbers: the π of the Boussinesq kernel Equation (A.1125), the $\pi^2/4$ of the integral Equation (A.1128), and the $\frac{2}{3}$ of Equation (A.1143). □

A.58.5 Checks and consequences

Example A.703 (The numbers of the chapter, recomputed)**.** Example 30.71 presses two steel balls of radius 10 mm together with 1 N, with $E = 210\,\mathrm{GPa}$ and $\nu = 0.29$. Then Equation (A.1136) gives

$$\frac{1}{E^*} = \frac{2\,(1-\nu^2)}{E} = \frac{2\,(1-0.0841)}{2.10\times10^{11}\,\mathrm{Pa}} = 8.72\times10^{-12}/\mathrm{Pa}\,, \qquad E^* = 1.15\times10^{11}\,\mathrm{Pa}\,, \tag{A.1145}$$

and $R = 5.0\,\mathrm{mm}$ by Equation (A.1127). Inverting Equation (A.1144) with $\delta = a^2/R$ gives $a = [3FR/(4E^*)]^{1/3}$, so

$$a = \left[\frac{3\,(1\,\mathrm{N})\,(5.0\times10^{-3}\,\mathrm{m})}{4\,(1.15\times10^{11}\,\mathrm{Pa})}\right]^{1/3} = 3.2\times10^{-5}\,\mathrm{m} = 32\,\mu\mathrm{m}\,, \tag{A.1146}$$

whence $\delta = a^2/R = 2.0\times10^{-7}\,\mathrm{m} = 0.20\,\mu\mathrm{m}$ and, by Equation (A.1138), $p_0 = 3F/(2\pi a^2) = 4.7\times10^8\,\mathrm{Pa} = 470\,\mathrm{MPa}$. All three reproduce Example 30.71, which is the arithmetic check a reader can run in a minute. Note also the ratio $a/R = 6.4\times10^{-3}$, which is what licenses the half-space approximation of Definition A.698.

Rests on Theorem A.702 and Phenomenon 30.70.

Remark A.704 (E^* in two roles, and one trap). The symbol $E/(1-\nu^2)$ occurs twice in Chapter 30 and means two different things, which is worth stating plainly because the two are easily read as one quantity. In Definition 30.63 it is a *single-body* stiffness: the plane-strain modulus of the plate material, appearing because a bent plate cannot contract transversely. In Equation (30.60) the same expression is one term of a *two-body* sum, Equation (A.1136), and the contact modulus E^* is the harmonic-type combination of both bodies' plane-strain moduli. The underlying modulus is the same and its physical reason is the same — lateral confinement — but a contact between two steel bodies has E^* equal to *half* the plate value for steel, and substituting one for the other misstates the contact stiffness by a factor of two.

Rests on Theorem A.702 and Definition 30.63.

Remark A.705 (The dangerous stress is below the surface). The peak *pressure* is at the centre of the contact, on the surface. The peak *shear* stress is not: evaluating the subsurface stress field from the same Boussinesq kernel — an elementary but long computation, not carried out here — places the maximum of $\frac{1}{2}(\sigma_1 - \sigma_3)$ on the axis at a depth of about $0.48\,a$ for $\nu \approx 0.3$, where it reaches about $0.31\,p_0$. Since ductile yielding is governed by the maximum shear stress and not by the pressure (Proposition 30.74), first yield under a Hertzian contact begins at that depth, inside apparently undamaged material. For the balls of Example A.703 the peak shear is $0.31\times470\,\mathrm{MPa} = 146\,\mathrm{MPa}$ at a depth of 15 µm. That is the mechanism behind rolling-contact fatigue: a bearing race accumulates plastic damage at a fixed small depth under every pass of a ball, and eventually a subsurface crack reaches the surface and detaches a flake. It is why such a failure is a spall and not wear, and why it is invisible until the moment it is not — the point made at the end of Example 30.71 and taken up in Section 30.8.

Rests on Theorem A.702 and Proposition 30.74.

Remark A.706. Appendix A.58 discharges the derivation owed at Phenomenon 30.70 of Chapter 30, and only that part of it which the inline derivation cannot reach: the elliptic pressure profile Equation (A.1138), the coefficient $\frac{4}{3}$ of Equation (30.60), and the two-body compliance rule of Proposition A.701. The exponent $\frac{3}{2}$ and the relation $a = \sqrt{R\delta}$ are proved in the chapter, by an argument that needs no elasticity solution at all, and are not reproved here. One result is imported and named as such (Remark A.696): Boussinesq's point-force solution Equation (A.1125), a debt owed by Chapter 10, whose Papkovich–Neuber representation this treatise does not carry. Everything else — the potential integral of Lemma A.700 in particular — is carried out in full from the isotropic Hooke law Equation (30.21) proved in Appendix A.56.

A.59 D'Alembert's Paradox for a Body of Arbitrary Shape

This appendix proves Phenomenon 31.30 of Chapter 31 in the generality in which it was stated. The chapter derives the result for a sphere, where the fore-and-aft symmetry of Equation (31.28) makes the cancellation visible in one line; what is owed, and is discharged here, is the statement for a body of *any* shape, where no such symmetry is available and the vanishing of the force has to be extracted from the structure of the far field instead. The conclusion is stronger than the chapter's phrasing: not only the drag but the whole resultant vanishes, so an arbitrary body in a steady, simply connected, irrotational flow of an ideal fluid feels no lift either. That is what makes Remark 31.31 exhaustive — the only way to buy a transverse force is to break the hypotheses, which is what Theorem 31.68 does.

A.59.1 Statement and hypotheses

Theorem A.707 (D'Alembert's paradox)**.** *Let S be the smooth closed surface of a finite rigid body held fixed in an unbounded incompressible ideal fluid of uniform density ρ, in* $\mathrm{kg/m^3}$*, occupying the exterior region Ω, and let the motion be steady and irrotational with a single-valued potential ϕ on Ω, satisfying*

$$\nabla^2\phi = 0 \text{ in } \Omega\,, \qquad \partial_n\phi = 0 \text{ on } S\,, \qquad \boldsymbol{\nabla}\phi \longrightarrow U\hat{\boldsymbol{z}} \text{ as } |\boldsymbol{x}| \longrightarrow \infty\,. \tag{A.1147}$$

Then the resultant force exerted on the body by the fluid pressure vanishes identically,

$$\boldsymbol{F} = -\oint_S p\,\boldsymbol{N}\,\mathrm{d}S = \boldsymbol{0}\,, \tag{A.1148}$$

$\boldsymbol{N}$ being the unit normal on S directed into the fluid. The statement holds for every shape of S, for every value of U in m/s*, and componentwise: there is no drag and no lift.*

Rests on Proposition 31.28 and Theorem 31.24.

Three hypotheses carry the whole argument and each is named in Remark 31.31 as a way out. *Steadiness* is what licenses the momentum balance of Appendix A.59.3. *Single-valuedness of ϕ* is the simple-connectivity assumption; it is what forbids a circulation, and dropping it is the plane-flow escape of Section 31.7.1. *Attachment* — that Ω really is the whole exterior, with the fluid following the surface everywhere — is what nature declines to provide, and its failure is Section 31.7.2.

A.59.2 The far field carries no source

Write

$$\phi = Uz + \chi\,, \qquad \boldsymbol{u}' := \boldsymbol{\nabla}\chi\,, \tag{A.1149}$$

so that χ is harmonic on Ω and $\boldsymbol{u}'$, the disturbance velocity, tends to zero at infinity. The first step is that χ has no monopole term, and this is a consequence of impermeability alone.

Lemma A.708 (The disturbance flux vanishes)**.** *Let S_R be a sphere of radius R large enough to contain S, with outward unit normal $\boldsymbol{n}$. Then*

$$\oint_{S_R} \boldsymbol{u}' \cdot \boldsymbol{n}\,\mathrm{d}S = 0 \tag{A.1150}$$

for every such R.

Rests on Theorem 7.101 and Equation (A.1149).

Proof. *Derives Lemma A.708.*

Apply the divergence theorem (Theorem 7.101) to $\boldsymbol{u}'$ on the region bounded by S and S_R. Since $\boldsymbol{\nabla}\cdot\boldsymbol{u}' = \nabla^2\chi = 0$ there,

$$\oint_{S_R} \boldsymbol{u}'\cdot\boldsymbol{n}\,\mathrm{d}S = \oint_S \boldsymbol{u}'\cdot\boldsymbol{N}\,\mathrm{d}S = \oint_S \boldsymbol{\nabla}\phi\cdot\boldsymbol{N}\,\mathrm{d}S - U\oint_S \hat{\boldsymbol{z}}\cdot\boldsymbol{N}\,\mathrm{d}S\,.$$

The first integral is zero because $\partial_n\phi = 0$ on S: the body is impermeable. The second is zero because the integral of the unit normal over any closed surface vanishes — take $\boldsymbol{a}$ constant in the divergence theorem and read $\oint_S \boldsymbol{a}\cdot\boldsymbol{N}\,\mathrm{d}S = \int \boldsymbol{\nabla}\cdot\boldsymbol{a}\,\mathrm{d}V = 0$. □

Theorem A.709 (Exterior expansion in solid harmonics; quoted)**.** *Let χ be harmonic in the exterior of a ball of radius R_0 in three dimensions and tend to zero at infinity. Then, in spherical coordinates (r,θ,φ) centred in that ball,*

$$\chi(\boldsymbol{x}) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} \frac{c_{lm}}{r^{\,l+1}}\,Y_{lm}(\theta,\varphi)\,, \tag{A.1151}$$

the series converging uniformly, together with the series of its term-by-term derivatives of every order, on $r \geq R_1$ for each $R_1 > R_0$. Here Y_{lm} is the spherical harmonic Equation (9.163), and the powers r^{-l-1} are the decaying radial solutions of Equation (10.34) at $k = 0$ (Example 10.33).

Rests on Equations (9.163) and (10.34).

Remark A.710 (What is quoted here). Theorem A.709 is the only statement in this section that is not derived. Part II supplies one half of it and not the other. The half it supplies is the *basis*: separation of Laplace's equation in spherical coordinates, carried out at Example 10.33, produces exactly the radial factors r^l and r^{-l-1} against the spherical harmonics of Section 9.9.3, and the decay condition selects the second family. The half it does not supply is *completeness with convergence* — that every exterior harmonic function decaying at infinity is the sum of such a series, uniformly and differentiably. That is a theorem of potential theory, resting on the mean-value property (Appendix A.16) and on the analyticity of harmonic functions, and this treatise does not build it; it is recorded here as a debt against Chapter 10, in the same spirit as the vector identities recorded at Remark 31.5. Its classical use in exactly the present setting — the far field of a body in a stream as a source, a dipole and higher terms — is Lamb's [701]. Nothing below uses more of it than the leading two terms.

Corollary A.711 (The disturbance is a dipole at leading order)**.** *There is a constant vector $\boldsymbol{A}$, of dimension $\mathsf{L}^4\mathsf{T}^{-1}$, such that*

$$\chi = \frac{\boldsymbol{A}\cdot\boldsymbol{x}}{r^3} + O\Big(r^{-3}\Big)\,, \qquad u_i' = \frac{1}{r^3}\left(A_i - 3\,(\boldsymbol{A}\cdot\boldsymbol{n})\,n_i\right) + O\Big(r^{-4}\Big)\,, \tag{A.1152}$$

with $\boldsymbol{n} = \boldsymbol{x}/r$.

Rests on Lemma A.708 and Theorem A.709.

Proof. *Derives Corollary A.711.*

In Equation (A.1151) the $l = 0$ term is $c_{00}Y_{00}/r$, a point source of strength $-4\pi c_{00}Y_{00}$; integrating its gradient over S_R gives that strength, and every $l \geq 1$ term integrates to zero over the sphere by the orthogonality of the spherical harmonics (Theorem 9.119). So Lemma A.708 forces $c_{00} = 0$. The three $l = 1$ terms are linear combinations of x_i/r^3, which is the first expression in Equation (A.1152); the remainder is $l \geq 2$, i.e. $O(r^{-3})$. Differentiating,

$$\partial_i \frac{A_j x_j}{r^3} = \frac{A_i}{r^3} - \frac{3A_j x_j x_i}{r^5} = \frac{1}{r^3}\left(A_i - 3\,(\boldsymbol{A}\cdot\boldsymbol{n})\,n_i\right)\,,$$

and term-by-term differentiation of the remainder, licensed by Theorem A.709, gives $O(r^{-4})$. □

The physical reading of Lemma A.708 is worth one sentence, because it is the only place where the body enters at all: a rigid impermeable body neither creates nor destroys fluid, so it cannot look like a source from far away, and the slowest decay it can produce is that of a dipole. The sphere of Example 31.29 is the worked case, $\chi = Ua^3\cos\theta/2r^2$, with $\boldsymbol{A} = \frac{1}{2}Ua^3\hat{\boldsymbol{z}}$.

A.59.3 The momentum theorem on a large sphere

Lemma A.712 (The force is a far-field integral)**.** *For every R large enough that S_R encloses S,*

$$F_i = -\oint_{S_R} \left[p\, n_i + \rho v_i\,(\boldsymbol{v}\cdot\boldsymbol{n})\right] \mathrm{d}S\,, \tag{A.1153}$$

with $\boldsymbol{v} = \boldsymbol{\nabla}\phi$ and p the pressure. In particular the right-hand side is independent of R.

Rests on Theorems 7.101 and 31.22.

Proof. *Derives Lemma A.712.*

Let V be the fluid region between S and S_R and let $\boldsymbol{m}$ be the outward normal of V, equal to $\boldsymbol{n}$ on S_R and to $-\boldsymbol{N}$ on S. In steady flow of an ideal fluid with no body force, the Euler equation (Theorem 31.22) combined with $\boldsymbol{\nabla}\cdot\boldsymbol{v} = 0$ reads $\partial_j\left(\rho v_i v_j + p\,\delta_{ij}\right) = 0$, since $\rho v_j \partial_j v_i = \partial_j(\rho v_i v_j)$. Integrating this over V and applying the divergence theorem (Theorem 7.101) to each i,

$$0 = \oint_{\partial V} \left[p\, m_i + \rho v_i\,(\boldsymbol{v}\cdot\boldsymbol{m})\right] \mathrm{d}S = \oint_{S_R} \left[p\, n_i + \rho v_i\,(\boldsymbol{v}\cdot\boldsymbol{n})\right] \mathrm{d}S - \oint_{S} p\, N_i \,\mathrm{d}S\,,$$

where the momentum flux over S dropped because $\boldsymbol{v}\cdot\boldsymbol{N} = 0$ there. The last integral is $-F_i$ by Equation (A.1148), which is the assertion. □

The pressure is supplied by Bernoulli's theorem (Theorem 31.24) at constant height. With $q^2 = |\boldsymbol{v}|^2 = U^2 + 2Uu'_z + |\boldsymbol{u}'|^2$,

$$p = p_\infty + \tfrac{1}{2}\rho U^2 - \tfrac{1}{2}\rho q^2 = p_\infty - \rho U u'_z - \tfrac{1}{2}\rho\,|\boldsymbol{u}'|^2\,. \tag{A.1154}$$

By Corollary A.711 the last term is $O(R^{-6})$ on S_R; multiplied by the area element it contributes $O(R^{-4})$ to Equation (A.1153) and vanishes in the limit. The same estimate disposes of the quadratic part of the momentum flux:

$$\rho v_i\,(\boldsymbol{v}\cdot\boldsymbol{n}) = \rho U^2\delta_{i3}n_z + \rho U\delta_{i3}\,(\boldsymbol{u}'\cdot\boldsymbol{n}) + \rho U u'_i n_z + O\Big(R^{-6}\Big)\,. \tag{A.1155}$$

A.59.4 Every surviving term integrates to zero

Assemble Equations (A.1153) to (A.1155) and take the four contributions in turn.

The uniform terms. $\oint_{S_R} p_\infty n_i\,\mathrm{d}S = 0$ and $\rho U^2\delta_{i3}\oint_{S_R} n_z\,\mathrm{d}S = 0$, both because the integral of the unit normal over a closed surface vanishes. A constant pressure and a uniform stream, taken alone, push a closed surface nowhere.

The disturbance flux. $\rho U\delta_{i3}\oint_{S_R}\boldsymbol{u}'\cdot\boldsymbol{n}\,\mathrm{d}S = 0$ by Lemma A.708 — exactly, at every R, not merely in the limit.

The two surviving dipole terms. What is left is

$$F_i = \rho U \oint_{S_R} \left[u'_z n_i - u'_i n_z\right] \mathrm{d}S + O\Big(R^{-2}\Big)\,, \tag{A.1156}$$

the first term coming from the pressure through Equation (A.1154) and the second from the momentum flux, with the signs of Equation (A.1153) already applied. Insert Equation (A.1152) and write $\mathrm{d}S = R^2 \mathrm{d}\Omega$ with $\mathrm{d}\Omega$ the solid angle:

$$\begin{aligned}\oint_{S_R} \left[u'_z n_i - u'_i n_z\right] \mathrm{d}S &= \frac{1}{R} \oint \Big[\left(A_3 - 3\left(\boldsymbol{A}\cdot\boldsymbol{n}\right) n_3\right) n_i - \left(A_i - 3\left(\boldsymbol{A}\cdot\boldsymbol{n}\right) n_i\right) n_3\Big] \mathrm{d}\Omega \\ &= \frac{1}{R} \oint \left[A_3 n_i - A_i n_3\right] \mathrm{d}\Omega \\ &= \frac{1}{R}\left[A_3 \oint n_i \,\mathrm{d}\Omega - A_i \oint n_3 \,\mathrm{d}\Omega\right] = 0\,,\end{aligned} \tag{A.1157}$$

because the two terms in $\boldsymbol{A}\cdot\boldsymbol{n}$ cancel identically — they enter the two integrands as $-3(\boldsymbol{A}\cdot\boldsymbol{n})n_3 n_i$ and $+3(\boldsymbol{A}\cdot\boldsymbol{n})n_i n_3$, the same quantity with opposite signs — and because $\oint n_i \,\mathrm{d}\Omega = 0$ over the whole sphere.

That cancellation is the mechanism, and it deserves to be read rather than merely verified. The dipole disturbance contributes to the pressure integral and to the momentum-flux integral separately, and neither contribution is zero; what is zero is their difference, at every R and for every orientation of $\boldsymbol{A}$, hence for every shape of body.

Proof. *Derives Theorem A.707.*

By Lemma A.712 the quantity F_i defined by Equation (A.1153) does not depend on R. By Equations (A.1156) and (A.1157) it equals $O(R^{-2})$. A constant that is $O(R^{-2})$ for arbitrarily large R is zero, so $\boldsymbol{F} = \boldsymbol{0}$ — exactly, and for all three components at once. □

Remark A.713 (Where the sphere's fore-and-aft symmetry went). The chapter's proof for the sphere runs on the symmetry of Equation (31.28) about $\theta = \pi/2$, and a general body has no such symmetry: the pressure distribution over an aerofoil at incidence is grossly asymmetric, and every individual surface element carries a real force. What survives in general is weaker and sufficient — the *resultant* is fixed by the far field alone, and the far field of any impermeable body is a dipole, whose two contributions to Equation (A.1153) cancel. The near field, where all the physics of shape lives, never enters the calculation. This is also why the theorem is so unforgiving: nothing about the body is used except that it is closed and that the fluid does not pass through it.

Remark A.714. Appendix A.59 discharges the general case of Phenomenon 31.30 in Chapter 31, where the sphere is done inline and the arbitrary body is deferred here. Read it next to Remark 31.31, whose claim that there are exactly two escapes is now exact rather than rhetorical, because Theorem A.707 kills the resultant and not only its streamwise component. The first escape is multiple connectivity: in the plane, dropping single-valuedness of ϕ admits a circulation and produces the transverse force of Theorem 31.68 — which still leaves the drag zero, so the paradox survives its own resolution for lift, and the appendix computing the circulation itself is Appendix A.62. The second is that real flow is neither steady nor attached, which is Section 31.7.2, quantified by Theorem 31.71 and measured in Section 36.5.

A.60 The Coefficient 6π in Stokes' Drag Law

This appendix supplies the one thing Phenomenon 31.45 of Chapter 31 does not derive. The chapter's inline argument settles the *form* of Equation (31.39) completely: linearity of

Equation (31.38) makes the drag exactly proportional to U, and Theorem D.2 then leaves $\mu a U$ as the only force that can be built, so $F = C\mu a U$ with C a pure number. What is owed is C. It is obtained here by solving the creeping-flow problem for the sphere in closed form and integrating the surface traction, and the calculation returns rather more than a number: it says how the drag divides between pressure and friction, which is the physically informative part and the part the chapter has no room for.

A.60.1 The stream function and the equation it obeys

Work in the frame of the sphere, spherical coordinates (r, θ, φ) with the polar axis along the stream, and take the flow axisymmetric and without swirl, $v_\varphi = 0$. Boundary conditions: no slip on the sphere, and a uniform stream at infinity,

$$v_r = v_\theta = 0 \text{ at } r = a\,, \qquad \boldsymbol{v} \longrightarrow U\hat{\boldsymbol{z}} \text{ as } r \longrightarrow \infty\,. \tag{A.1158}$$

Definition A.715 (Stokes stream function)**.** The *Stokes stream function* $\psi(r, \theta)$, in m^3/s, is defined by

$$v_r = \frac{1}{r^2 \sin\theta}\frac{\partial \psi}{\partial \theta}\,, \qquad v_\theta = -\frac{1}{r \sin\theta}\frac{\partial \psi}{\partial r}\,. \tag{A.1159}$$

Rests on Example 31.15 and Equation (7.92).

Incompressibility is then an identity rather than a constraint: in spherical coordinates $\boldsymbol{\nabla}\cdot\boldsymbol{v} = r^{-2}\partial_r(r^2 v_r) + (r\sin\theta)^{-1}\partial_\theta(\sin\theta\, v_\theta)$, and substituting Equation (A.1159) gives $(r^2\sin\theta)^{-1}(\partial_r\partial_\theta\psi - \partial_\theta\partial_r\psi) = 0$. This is the axisymmetric counterpart of the plane construction of Example 31.15, and it is why the whole problem collapses to one scalar equation.

Lemma A.716 (The creeping-flow equation for ψ)**.** *Under Definition A.715 the Stokes equations Equation (31.38) are equivalent to*

$$E^4\psi = 0\,, \qquad E^2 := \partial_r^2 + \frac{\sin\theta}{r^2}\,\partial_\theta\left(\frac{1}{\sin\theta}\,\partial_\theta\right)\,, \tag{A.1160}$$

together with a pressure recovered by quadrature.

Rests on Definitions 31.43 and A.715.

Proof. *Derives Lemma A.716.*

Take the curl of $\mathbf{0} = -\boldsymbol{\nabla}p + \mu\nabla^2\boldsymbol{v}$. The pressure disappears because the curl of a gradient vanishes (Equation (7.128)), leaving $\nabla^2\boldsymbol{\omega} = \mathbf{0}$ with $\boldsymbol{\omega} = \boldsymbol{\nabla}\times\boldsymbol{v}$. In an axisymmetric flow without swirl the vorticity has only an azimuthal component, and substituting Equation (A.1159) into $\omega_\varphi = r^{-1}[\partial_r(r v_\theta) - \partial_\theta v_r]$ gives

$$\omega_\varphi = -\frac{1}{r\sin\theta}\,E^2\psi\,. \tag{A.1161}$$

The vector Laplacian acting on a purely azimuthal field of this form reduces, term by term in the spherical expression of Equation (7.94), to $(\nabla^2\boldsymbol{\omega})_\varphi = -(r\sin\theta)^{-1}E^2(E^2\psi)$, which is Equation (A.1160). Conversely, given ψ with $E^4\psi = 0$, the momentum equation reads $\boldsymbol{\nabla}p = \mu\nabla^2\boldsymbol{v} = -\mu\boldsymbol{\nabla}\times\boldsymbol{\omega}$ by Equation (31.6) and $\boldsymbol{\nabla}\cdot\boldsymbol{v} = 0$, and its right-hand side is then curl-free, so p exists and is obtained by integrating along any path. □

A.60.2 Solution for the sphere

The condition at infinity fixes the angular dependence once and for all. A uniform stream has $\psi \longrightarrow \tfrac{1}{2}Ur^2\sin^2\theta$ — read Equation (A.1159) backwards — and both the equation and the boundary conditions are then satisfied by the single separated form

$$\psi(r, \theta) = f(r)\sin^2\theta\,. \tag{A.1162}$$

Acting with E^2 on Equation (A.1162),

$$E^2\psi = \left[f'' - \frac{2f}{r^2}\right]\sin^2\theta\,,$$

since $\partial_\theta\left(\sin^{-1}\theta\,\partial_\theta\sin^2\theta\right) = \partial_\theta\left(2\cos\theta\right) = -2\sin\theta$. Applying E^2 a second time to $g := f'' - 2f/r^2$ in the same way and collecting,

$$f'''' - \frac{4f''}{r^2} + \frac{8f'}{r^3} - \frac{8f}{r^4} = 0\,. \tag{A.1163}$$

Lemma A.717 (The radial solution)**.** *The general solution of Equation* (A.1163) *is* $f = A/r + Br + Cr^2 + Dr^4$*, and the conditions Equation* (A.1158) *select*

$$D = 0\,, \qquad C = \tfrac{1}{2}U\,, \qquad A = \tfrac{1}{4}Ua^3\,, \qquad B = -\tfrac{3}{4}Ua\,. \tag{A.1164}$$

Rests on Lemma A.716 and Equation (A.1158)*.*

Proof. *Derives Lemma A.717.*

Equation (A.1163) is equidimensional: every term scales as r^{n-4} under $f = r^n$, so the substitution turns it into the algebraic equation

$$n(n-1)(n-2)(n-3) - 4n(n-1) + 8n - 8 = 0\,,$$

whose left-hand side factorizes as $(n+1)\,(n-1)\,(n-2)\,(n-4)$; one checks the four roots directly, $n = -1$ giving $24 - 8 - 16 = 0$, $n = 1$ giving $0 - 0 + 8 - 8 = 0$, $n = 2$ giving $0 - 8 + 16 - 8 = 0$ and $n = 4$ giving $24 - 48 + 32 - 8 = 0$. The four powers are independent, so they span the solution space.

The condition at infinity requires $f \longrightarrow \frac{1}{2}Ur^2$, which forbids r^4 and fixes $C = U/2$. No slip requires both components of Equation (A.1159) to vanish at $r = a$, that is $f(a) = 0$ and $f'(a) = 0$:

$$\frac{A}{a} + Ba + \frac{Ua^2}{2} = 0\,, \qquad -\frac{A}{a^2} + B + Ua = 0\,.$$

Multiplying the second by a and subtracting it from the first eliminates B and leaves $2A/a - Ua^2/2 = 0$, whence $A = Ua^3/4$; back-substitution gives $B = A/a^2 - Ua = -3Ua/4$. □

Reading the velocity field off Equations (A.1159) and (A.1164),

$$v_r = U\cos\theta\left(1 - \frac{3a}{2r} + \frac{a^3}{2r^3}\right), \qquad v_\theta = -U\sin\theta\left(1 - \frac{3a}{4r} - \frac{a^3}{4r^3}\right). \tag{A.1165}$$

Both vanish at $r = a$ and both tend to the free stream, as they must.

A.60.3 The pressure, and a check

From Equation (A.1164), $f'' - 2f/r^2 = 3Ua/2r$, so by Equation (A.1161)

$$\omega_\varphi = -\frac{3aU\sin\theta}{2r^2}\,. \tag{A.1166}$$

The radial component of $\boldsymbol{\nabla}p = -\mu\boldsymbol{\nabla}\times\boldsymbol{\omega}$ is then

$$\frac{\partial p}{\partial r} = -\frac{\mu}{r\sin\theta}\,\partial_\theta\left(\sin\theta\,\omega_\varphi\right) = \frac{3\mu aU}{r\sin\theta}\,\partial_\theta\left(\frac{\sin^2\theta}{2r^2}\right) = \frac{3\mu aU\cos\theta}{r^3}\,,$$

and integrating from infinity,

$$p = p_\infty - \frac{3\mu aU\cos\theta}{2r^2}\,. \tag{A.1167}$$

It is worth confirming that this is harmonic, since $\nabla^2 p = 0$ follows from taking the divergence of Equation (31.38): the function $\cos\theta/r^2$ is the $l = 1$ exterior solid harmonic of Example 10.33, and it is. The pressure is high in front ($\theta = \pi$, taking the stream along $+\hat{\boldsymbol{z}}$ so that $\theta = \pi$ faces upstream) and low behind, by equal amounts — the fore-and-aft antisymmetry that would give zero drag in an ideal fluid (Phenomenon 31.30). Here it does not, because the pressure is no longer the whole traction.

A.60.4 The traction and its integral

For a Newtonian fluid Equation (31.30) gives, in spherical coordinates,

$$\sigma_{rr} = -p + 2\mu\,\frac{\partial v_r}{\partial r}\,, \qquad \sigma_{r\theta} = \mu\left[r\,\partial_r\!\left(\frac{v_\theta}{r}\right) + \frac{1}{r}\,\frac{\partial v_r}{\partial\theta}\right]. \tag{A.1168}$$

Evaluate both at $r = a$ from Equations (A.1165) and (A.1167). For the normal component, $\partial_r v_r = U\cos\theta\,(3a/2r^2 - 3a^3/2r^4)$, which vanishes at $r = a$; the entire normal traction is therefore the pressure,

$$\sigma_{rr}|_{r=a} = -p_\infty + \frac{3\mu U\cos\theta}{2a}\,. \tag{A.1169}$$

For the tangential component, v_r vanishes identically on $r = a$ so its θ derivative does too, and $v_\theta/r = -U\sin\theta\,(1/r - 3a/4r^2 - a^3/4r^4)$ gives $\partial_r\,(v_\theta/r) = -U\sin\theta\,(-1/r^2 + 3a/2r^3 + a^3/r^5)$, equal to $-3U\sin\theta/2a^2$ at $r = a$; hence

$$\sigma_{r\theta}|_{r=a} = -\frac{3\mu U\sin\theta}{2a}\,. \tag{A.1170}$$

Both are in Pa, as $\mu U/a$ has the dimensions $\mathsf{ML^{-1}T^{-2}}$.

Theorem A.718 (Stokes' drag law with its coefficient)**.** *The force exerted on the sphere by the fluid is directed along the stream and has magnitude*

$$F = 6\pi\mu a U\,, \tag{A.1171}$$

of which $2\pi\mu aU$ — one third — comes from the pressure and $4\pi\mu aU$ — two thirds — from the tangential viscous traction.

Rests on Lemma A.717 and Theorem 31.36.

Proof. *Derives Theorem A.718.*

The component along the stream of the traction on the surface element $\mathrm{d}S$ is $(\sigma_{rr}\cos\theta - \sigma_{r\theta}\sin\theta)\,\mathrm{d}S$, the projections of the radial and meridional directions on $\hat{\boldsymbol{z}}$. Substituting Equations (A.1169) and (A.1170),

$$\sigma_{rr}\cos\theta - \sigma_{r\theta}\sin\theta = -p_\infty\cos\theta + \frac{3\mu U}{2a}\left(\cos^2\theta + \sin^2\theta\right) = -p_\infty\cos\theta + \frac{3\mu U}{2a}\,.$$

The uniform pressure integrates to zero over the closed surface, and what remains is a constant traction $3\mu U/2a$ over the whole sphere:

$$F = \frac{3\mu U}{2a}\oint \mathrm{d}S = \frac{3\mu U}{2a}\cdot 4\pi a^2 = 6\pi\mu a U\,.$$

For the split, keep the two sources separate. The pressure contributes

$$F^{(p)} = \oint (-p)\cos\theta\,\mathrm{d}S = \frac{3\mu U}{2a}\,2\pi a^2\int_0^\pi \cos^2\theta\sin\theta\,\mathrm{d}\theta = \frac{3\mu U}{2a}\,2\pi a^2\cdot\frac{2}{3} = 2\pi\mu a U\,,$$

and the tangential traction contributes

$$F^{(\tau)} = \oint (-\sigma_{r\theta})\sin\theta\,\mathrm{d}S = \frac{3\mu U}{2a}\,2\pi a^2\int_0^\pi \sin^3\theta\,\mathrm{d}\theta = \frac{3\mu U}{2a}\,2\pi a^2\cdot\frac{4}{3} = 4\pi\mu a U\,,$$

using $\int_0^\pi \cos^2\theta\sin\theta\,\mathrm{d}\theta = 2/3$ and $\int_0^\pi \sin^3\theta\,\mathrm{d}\theta = 4/3$. Their sum is Equation (A.1171), which is Equation (31.39) with $C = 6\pi$ [135]. □

Remark A.719 (Why the split is worth printing). The one-third–two-thirds division is a fact about creeping flow that survives no other regime and is easy to get backwards. At high Reynolds number the drag of a bluff body is almost entirely pressure drag, because the wake destroys the rear pressure recovery (Phenomenon 31.55); at low Reynolds number the pressure distribution Equation (A.1167) is exactly the fore-and-aft antisymmetric one, and it contributes only because it is multiplied by $\cos\theta$ and integrated — it is not the residue of a broken symmetry but the symmetric distribution itself doing work against the motion. The larger part of the resistance is skin friction, which is the sense in which Stokes flow is the opposite of the ideal flow of Section 31.3.3: there the friction was absent and the pressure did nothing, here the friction dominates and the pressure does a third.

Remark A.720 (The slow decay, and what it costs). Equation (A.1165) decays only as a/r: the term $-3aU\cos\theta/2r$ in v_r is the leading disturbance, in contrast with the a^3/r^3 dipole of the ideal-flow solution Equation (31.27). A sphere in creeping flow drags a very large volume of fluid with it, which is why suspensions interact hydrodynamically at many diameters' separation, and why the drag is proportional to a rather than to a^2: the resisted region has the size of the disturbance, not the size of the body. The same slow decay is the origin of the non-uniformity discussed at Remark 31.47 — the neglected inertial term falls off as r^{-3} while the retained viscous term falls off as r^{-4}, so beyond a distance of order a/Re the approximation inverts itself. Whitehead's paradox is the failure of the naive second approximation to satisfy the condition at infinity, and Oseen's repair, Equation (31.41), keeps the advection by the uniform stream in the outer region from the start.

Remark A.721. Appendix A.60 discharges the numerical coefficient of Phenomenon 31.45 in Chapter 31, where Equation (31.39) is derived up to the pure number C from linearity and Theorem D.2 and the calculation of C is deferred here. With $C = 6\pi$ in hand, Example 31.46 converts an observed fall speed into a radius, which is how the oil droplets of [407] were sized, and Remark 31.47 states the first correction. Nothing in this section is valid outside $\mathrm{Re} \ll 1$ (Definition 31.43), and the honest statement of its range is Equation (31.41): the leading correction to Equation (A.1171) is a relative $\frac{3}{16}\mathrm{Re}$, so the law is good to one percent only up to $\mathrm{Re} \approx 0.05$.

A.61 The Kármán Spacing Ratio of the Vortex Street

This appendix proves Equation (31.54) of Phenomenon 31.61 in Chapter 31: among all staggered double rows of point vortices, exactly one ratio of row separation to streamwise spacing is not destroyed by its own induced motion, namely

$$\frac{h}{a} = \frac{1}{\pi}\cosh^{-1}\sqrt{2} = \frac{1}{\pi}\log\left(1+\sqrt{2}\right) = 0.280\,55\ldots, \tag{A.1172}$$

[375]. What the calculation settles and what it does not is stated in the chapter at Remark 31.62, and the reader should have that remark in view throughout: nothing here bears on the Strouhal number Equation (31.53), and the stability obtained is neutral, not asymptotic. The closing Remark A.731 returns to the point with the calculation in hand.

Throughout, the fluid is ideal, unbounded and two-dimensional; the vortices are point vortices in the sense of Section 31.7.1, each of circulation $\pm\Gamma$ in m^2/s, moving with the local velocity induced by all the others. The complex variable is $z = x + \mathrm{i}y$, the complex potential w is that of Equation (31.57), and a vortex of circulation Γ counterclockwise at z_0 has $w = -(\mathrm{i}\Gamma/2\pi)\log(z - z_0)$, so that

$$\frac{\mathrm{d}w}{\mathrm{d}z} = v_x - \mathrm{i}v_y = -\frac{\mathrm{i}\Gamma}{2\pi\,(z - z_0)}, \qquad \frac{\mathrm{d}\bar{z}_k}{\mathrm{d}t} = \sum_{j\neq k}\frac{-\mathrm{i}\Gamma_j}{2\pi\,(z_k - z_j)}, \tag{A.1173}$$

the second equation being the statement that each vortex is carried by the flow of the others.

A.61.1 One summation formula does all the work

Every lattice sum below is a special case of a single identity, which is proved from the Fourier series of an elementary function and needs nothing else.

Lemma A.722 (The basic lattice sum)**.** *Let $\zeta \in \mathbb{C} \setminus \mathbb{Z}$ and $0 < \varphi < 2\pi$. Then*

$$\sum_{n\in\mathbb{Z}} \frac{\mathrm{e}^{\mathrm{i}n\varphi}}{n+\zeta} = \frac{\pi}{\sin \pi\zeta}\, \mathrm{e}^{\mathrm{i}(\pi-\varphi)\zeta}\,, \tag{A.1174}$$

and consequently, differentiating with respect to ζ,

$$\Sigma\,(\varphi,\zeta) := \sum_{n\in\mathbb{Z}} \frac{\mathrm{e}^{\mathrm{i}n\varphi}}{(n+\zeta)^2} = \pi\, \mathrm{e}^{\mathrm{i}(\pi-\varphi)\zeta} \left[\frac{\pi\cos\pi\zeta}{\sin^2\pi\zeta} - \frac{\mathrm{i}\,(\pi-\varphi)}{\sin\pi\zeta}\right], \tag{A.1175}$$

whose limit as $\varphi \to 0^+$ is $\Sigma\,(0,\zeta) = \pi^2/\sin^2\pi\zeta$.

Rests on Theorem 8.6 and Equation (8.3).

Proof. *Derives Lemma A.722.*

Let $F(\varphi) := (\pi/\sin\pi\zeta)\,\mathrm{e}^{\mathrm{i}(\pi-\varphi)\zeta}$ on $(0,2\pi)$ and compute its Fourier coefficients:

$$c_n = \frac{1}{2\pi}\int_0^{2\pi} F(\varphi)\,\mathrm{e}^{-\mathrm{i}n\varphi}\,\mathrm{d}\varphi = \frac{1}{2}\,\frac{\mathrm{e}^{\mathrm{i}\pi\zeta}}{\sin\pi\zeta}\cdot\frac{1-\mathrm{e}^{-2\pi\mathrm{i}(\zeta+n)}}{\mathrm{i}\,(\zeta+n)}\,,$$

the elementary integral of $\mathrm{e}^{-\mathrm{i}(\zeta+n)\varphi}$. Since n is an integer, $\mathrm{e}^{-2\pi\mathrm{i}(\zeta+n)} = \mathrm{e}^{-2\pi\mathrm{i}\zeta}$, and $1 - \mathrm{e}^{-2\pi\mathrm{i}\zeta} = \mathrm{e}^{-\mathrm{i}\pi\zeta}\left(\mathrm{e}^{\mathrm{i}\pi\zeta} - \mathrm{e}^{-\mathrm{i}\pi\zeta}\right) = 2\mathrm{i}\,\mathrm{e}^{-\mathrm{i}\pi\zeta}\sin\pi\zeta$ by Equation (8.3). The two exponentials and the two sines cancel, leaving $c_n = 1/(n+\zeta)$, which is Equation (A.1174).

The series Equation (A.1175) converges absolutely and uniformly in ζ on compact subsets of $\mathbb{C}\setminus\mathbb{Z}$, so it may be obtained from Equation (A.1174) by differentiating term by term, using $\mathrm{d}\,(n+\zeta)^{-1}/\mathrm{d}\zeta = -\,(n+\zeta)^{-2}$ and the product rule on the right-hand side. For the limit, set $\varphi = 0$ in Equation (A.1175): the bracket becomes $\pi\,(\cos\pi\zeta - \mathrm{i}\sin\pi\zeta)\,/\sin^2\pi\zeta = \pi\,\mathrm{e}^{-\mathrm{i}\pi\zeta}/\sin^2\pi\zeta$, which cancels the prefactor $\mathrm{e}^{\mathrm{i}\pi\zeta}$ exactly. □

Corollary A.723 (Potential of one infinite row)**.** *The complex potential of vortices of circulation Γ placed at $z = na$ for every $n \in \mathbb{Z}$ is*

$$w(z) = -\frac{\mathrm{i}\Gamma}{2\pi}\,\log\sin\frac{\pi z}{a}\,, \qquad \frac{\mathrm{d}w}{\mathrm{d}z} = -\frac{\mathrm{i}\Gamma}{2a}\cot\frac{\pi z}{a}\,, \tag{A.1176}$$

and it is the only such potential whose velocity is bounded as $|\operatorname{Im} z| \to \infty$. The induced velocity tends to $\mp\Gamma/2a$ in the x direction as $\operatorname{Im} z \to \pm\infty$.

Rests on Equation (A.1173) and Theorem 8.18.

Proof. *Derives Corollary A.723.*

The function $\cot\,(\pi z/a)$ is meromorphic with simple poles exactly at $z = na$, where $\sin\,(\pi z/a)$ has its simple zeros, and near $z = na$ its principal part is $a/\pi\,(z-na)$. Hence $\mathrm{d}w/\mathrm{d}z$ in Equation (A.1176) behaves near $z = na$ as $-\,(\mathrm{i}\Gamma/2a)\cdot a/\pi\,(z-na) = -\mathrm{i}\Gamma/2\pi\,(z-na)$, which by Equation (A.1173) is precisely a vortex of circulation Γ there, and it is holomorphic elsewhere. As $\operatorname{Im} z \to \pm\infty$, $\cot\,(\pi z/a) \to \mp\mathrm{i}$, so $\mathrm{d}w/\mathrm{d}z \to \mp\Gamma/2a$; since $\mathrm{d}w/\mathrm{d}z = v_x - \mathrm{i}v_y$, the row drives the fluid backwards above it and forwards below it, at the stated speed. Any other candidate differs from Equation (A.1176) by a function that is entire (the singularities coincide) and bounded (both velocities are), hence constant by Liouville's theorem (Theorem 8.18). □

A.61.2 The street and its rigid translation

Definition A.724 (The staggered street)**.** Fix $a > 0$ and $h > 0$, both in m. The *staggered vortex street* consists of the rows

$$A_n = na \quad \text{with circulation } +\Gamma\,, \qquad B_n = \left(n + \tfrac{1}{2}\right) a - \mathrm{i}h \quad \text{with circulation } -\Gamma\,, \tag{A.1177}$$

for all $n \in \mathbb{Z}$. Write $s := \pi h/a$, $c := \cosh s$ and $t := \tanh s$ throughout.
Rests on Corollary A.723.

Proposition A.725 (The street translates rigidly)**.** *The configuration Equation (A.1177) is an exact solution of Equation (A.1173): every vortex moves with the same velocity*

$$U_{\mathrm{s}} = \frac{\Gamma}{2a} \tanh \frac{\pi h}{a} = \frac{\Gamma t}{2a}\,, \tag{A.1178}$$

parallel to the rows, so the street is carried downstream without change of shape.
Rests on Definition A.724 and Corollary A.723.

Proof. *Derives Proposition A.725.*

Take the vortex $A_0 = 0$. Its own row induces nothing there: the contributions of A_n and A_{-n} in Equation (A.1173) are equal and opposite. Row B, of circulation $-\Gamma$ and origin $a/2 - \mathrm{i}h$, contributes by Equation (A.1176)

$$\frac{\mathrm{d}\bar{z}}{\mathrm{d}t} = +\frac{\mathrm{i}\Gamma}{2a} \cot \frac{\pi\,(0 - a/2 + \mathrm{i}h)}{a} = \frac{\mathrm{i}\Gamma}{2a} \cot\left(-\frac{\pi}{2} + \mathrm{i}s\right) = -\frac{\mathrm{i}\Gamma}{2a} \tan(\mathrm{i}s) = \frac{\Gamma}{2a} \tanh s\,,$$

using $\cot(\theta - \pi/2) = -\tan\theta$ and $\tan(\mathrm{i}s) = \mathrm{i}\tanh s$. The result is real and positive, so $v_x = \Gamma t/2a$ and $v_y = 0$. Repeating at $B_0 = a/2 - \mathrm{i}h$ under row A gives $-(\mathrm{i}\Gamma/2a)\cot(\pi/2 - \mathrm{i}s) = -(\mathrm{i}\Gamma/2a)\tan(\mathrm{i}s) = \Gamma t/2a$, the same velocity; and again the vortex's own row contributes nothing. Every vortex therefore moves with Equation (A.1178). □

Equation (A.1178) is worth reading before the stability calculation, because it is directly visible in the photographs: the street moves *more slowly* than the stream that made it, since U_{s} is the velocity relative to the fluid at infinity and it is directed against the free stream in the frame of the cylinder. In the frame translating with the street the base state is an equilibrium, and since the base velocity is a constant it drops out of the linearized equations altogether; the analysis below may be read in either frame.

A.61.3 Linearization

Displace every vortex,

$$z_n^A = na + \alpha_n(t)\,, \qquad z_n^B = \left(n + \tfrac{1}{2}\right) a - \mathrm{i}h + \beta_n(t)\,, \tag{A.1179}$$

with $|\alpha_n|, |\beta_n| \ll a$. Expanding Equation (A.1173) to first order with $(D + \delta)^{-1} = D^{-1} - \delta D^{-2} + O(\delta^2)$, and using Proposition A.725 to cancel the zeroth-order terms against the rigid motion, gives for each k

$$\frac{\mathrm{d}\bar{\alpha}_k}{\mathrm{d}t} = \frac{\mathrm{i}\Gamma}{2\pi} \left[\sum_{n \neq k} \frac{\alpha_k - \alpha_n}{(A_k - A_n)^2} - \sum_n \frac{\alpha_k - \beta_n}{(A_k - B_n)^2} \right], \tag{A.1180}$$

$$\frac{\mathrm{d}\bar{\beta}_k}{\mathrm{d}t} = \frac{\mathrm{i}\Gamma}{2\pi} \left[-\sum_{n \neq k} \frac{\beta_k - \beta_n}{(B_k - B_n)^2} + \sum_n \frac{\beta_k - \alpha_n}{(B_k - A_n)^2} \right], \tag{A.1181}$$

the signs following from the circulations $\pm\Gamma$ carried by the inducing vortices.

Remark A.726 (Why one Fourier mode is not enough). Equations (A.1180) and (A.1181) are linear over $\mathbb{R}$ but *not* over $\mathbb{C}$: the left-hand sides carry the complex conjugates of the unknowns while the right-hand sides carry the unknowns themselves. The single-mode ansatz $\alpha_n = \alpha\,\mathrm{e}^{\mathrm{i}n\varphi}$ is therefore inconsistent, since $\bar{\alpha}_n = \bar{\alpha}\,\mathrm{e}^{-\mathrm{i}n\varphi}$ belongs to the mode $-\varphi$: matching the k dependence of the two sides would force $\varphi \equiv 0$. The wavenumbers φ and $-\varphi$ close on each other and on nothing else, so the smallest closed system is four-dimensional. Overlooking this is the one way the calculation can be got wrong while looking right, and it changes the answer: the two-dimensional system obtained by ignoring it makes the street unstable at every spacing, Equation (A.1172) included.

Accordingly set, for a fixed $\varphi \in (0, 2\pi)$,

$$\alpha_n = \alpha\,\mathrm{e}^{\mathrm{i}n\varphi} + \gamma\,\mathrm{e}^{-\mathrm{i}n\varphi}\,, \qquad \beta_n = \beta\,\mathrm{e}^{\mathrm{i}n\varphi} + \delta\,\mathrm{e}^{-\mathrm{i}n\varphi}\,. \tag{A.1182}$$

Three lattice sums appear. With $m = n - k$ and $\zeta = \frac{1}{2} - \mathrm{i}h/a$, so that $\pi\zeta = \pi/2 - \mathrm{i}s$ and therefore $\sin\pi\zeta = \cosh s = c$ and $\cos\pi\zeta = \mathrm{i}\sinh s$:

$$P(\varphi) := \sum_{m\neq 0} \frac{1 - \mathrm{e}^{\mathrm{i}m\varphi}}{m^2} = \frac{\varphi\,(2\pi - \varphi)}{2} = \frac{\pi^2 - \psi^2}{2}\,, \qquad \psi := \pi - \varphi\,, \tag{A.1183}$$

$$S_0 := \sum_n \frac{1}{\left(\left(n + \frac{1}{2}\right)a - \mathrm{i}h\right)^2} = \frac{1}{a^2}\,\Sigma\,(0,\zeta) = \frac{\pi^2}{a^2 c^2}\,, \tag{A.1184}$$

$$S_1 := \sum_n \frac{\mathrm{e}^{\mathrm{i}n\varphi}}{\left(\left(n + \frac{1}{2}\right)a - \mathrm{i}h\right)^2} = \frac{\mathrm{i}\pi}{a^2}\,\mathrm{e}^{\mathrm{i}\psi/2}\,E\,M\,, \tag{A.1185}$$

where the closed forms come from Lemma A.722 and

$$E := \mathrm{e}^{\psi s/\pi}\,, \qquad M := \frac{\pi t - \psi}{c}\,, \qquad M' := \frac{\pi t + \psi}{c}\,. \tag{A.1186}$$

For Equation (A.1183) use $\sum_{m\geq 1} \cos m\varphi/m^2 = \pi^2/6 - \pi\varphi/2 + \varphi^2/4$ on $[0, 2\pi]$; for Equation (A.1185) substitute $\sin\pi\zeta = c$, $\cos\pi\zeta = \mathrm{i}\sinh s$ and $\mathrm{e}^{\mathrm{i}\psi\zeta} = \mathrm{e}^{\mathrm{i}\psi/2}\mathrm{e}^{\psi s/\pi}$ in Equation (A.1175). The companion sum with φ replaced by $-\varphi$, i.e. ψ by $-\psi$, is

$$\tilde{S}_1 := \sum_n \frac{\mathrm{e}^{-\mathrm{i}n\varphi}}{\left(\left(n + \frac{1}{2}\right)a - \mathrm{i}h\right)^2} = \frac{\mathrm{i}\pi}{a^2}\,\mathrm{e}^{-\mathrm{i}\psi/2}\,\frac{M'}{E}\,. \tag{A.1187}$$

Two further sums reduce to these: replacing n by $-n$ shows that $\sum_n \left(\left(n - \frac{1}{2}\right)a + \mathrm{i}h\right)^{-2} = S_0$ and $\sum_n \mathrm{e}^{\mathrm{i}n\varphi}\left(\left(n - \frac{1}{2}\right)a + \mathrm{i}h\right)^{-2} = \tilde{S}_1$.

Substituting Equation (A.1182) into Equations (A.1180) and (A.1181) and matching the coefficients of $\mathrm{e}^{\mathrm{i}k\varphi}$ and $\mathrm{e}^{-\mathrm{i}k\varphi}$ separately — noting $P(-\varphi) = P(\varphi)$ and $S_1(-\varphi) = \tilde{S}_1(\varphi)$ — yields, with $K := \Gamma/2\pi$ and

$$Q := \frac{P}{a^2} - S_0 = \frac{1}{a^2}\left[\frac{\pi^2 - \psi^2}{2} - \frac{\pi^2}{c^2}\right], \tag{A.1188}$$

the closed four-dimensional system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} p \\ q \end{pmatrix} = \mathsf{B}\begin{pmatrix} r \\ u \end{pmatrix}, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} r \\ u \end{pmatrix} = \mathsf{C}\begin{pmatrix} p \\ q \end{pmatrix}, \tag{A.1189}$$

in the variables $p := \alpha$, $q := \beta$, $r := \bar{\gamma}$, $u := \bar{\delta}$, with

$$\mathsf{B} = \mathrm{i}K\begin{pmatrix} -Q & -\overline{\tilde{S}_1} \\ \overline{S_1} & Q \end{pmatrix}, \qquad \mathsf{C} = \mathrm{i}K\begin{pmatrix} Q & S_1 \\ -\tilde{S}_1 & -Q \end{pmatrix}. \tag{A.1190}$$

A.61.4 The growth rates

Solutions of Equation (A.1189) proportional to $\mathrm{e}^{\sigma t}$ have σ^2 an eigenvalue of BC. Three real quantities suffice to write it down:

$$X := |S_1|^2 = \frac{\pi^2}{a^4} E^2 M^2 \,, \qquad Y := \left|\tilde{S}_1\right|^2 = \frac{\pi^2}{a^4} \frac{M'^2}{E^2} \,, \qquad \Pi := S_1 \tilde{S}_1 = -\frac{\pi^2}{a^4} M M' \,, \tag{A.1191}$$

the last being real because the two phases in Equations (A.1185) and (A.1187) cancel and $(\mathrm{i}\pi)^2 = -\pi^2$; note $XY = \Pi^2$.

Lemma A.727 (Characteristic equation)**.** *The eigenvalues $\lambda = \sigma^2$ of BC satisfy*

$$\lambda^2 - K^2 \left(2Q^2 - X - Y\right) \lambda + K^4 \left(Q^2 - \Pi\right)^2 = 0 \,. \tag{A.1192}$$

Rests on Equations (A.1190) and (A.1191).

Proof. *Derives Lemma A.727.*

Multiplying the matrices of Equation (A.1190),

$$\mathsf{BC} = -K^2 \begin{pmatrix} -Q^2 + \overline{\tilde{S}_1} \tilde{S}_1 & Q\left(\overline{\tilde{S}_1} - S_1\right) \\ Q\left(\overline{S_1} - \tilde{S}_1\right) & \overline{S_1} S_1 - Q^2 \end{pmatrix} ,$$

whose trace is $-K^2\left(-2Q^2 + Y + X\right)$, i.e. the coefficient displayed. Its determinant is the product of the two determinants. Now $\overline{S_1} = \mathrm{e}^{\mathrm{i}\varphi} S_1$ — replace n by $-n-1$ in the series defining $\overline{S_1}$ and compare — and likewise $\overline{\tilde{S}_1} = \mathrm{e}^{-\mathrm{i}\varphi} \tilde{S}_1$, so the determinant of B is $-K^2\left(-Q^2 + \overline{\tilde{S}_1}\,\overline{S_1}\right) = K^2\left(Q^2 - \Pi\right)$, and that of C is the same. The characteristic polynomial of a 2×2 matrix is $\lambda^2 - (\text{trace})\,\lambda + (\text{determinant})$, which is Equation (A.1192). □

Proposition A.728 (Criterion for neutrality)**.** *The mode φ neither grows nor decays — all four roots σ are purely imaginary — if and only if*

$$2\,|Q| \le \frac{\pi}{a^2} \left|EM - \frac{M'}{E}\right| = \frac{2\pi}{a^2 c} \left|\pi t \sinh u - \psi \cosh u\right| \,, \qquad u := \frac{\psi s}{\pi} \,, \tag{A.1193}$$

and it grows otherwise.

Rests on Lemma A.727.

Proof. *Derives Proposition A.728.*

Since $\sigma^2 = \lambda$, the four exponents are purely imaginary exactly when both roots of Equation (A.1192) are real and non-positive; if either root is complex, or real and positive, some σ has positive real part. Write $\mathcal{T} = 2Q^2 - X - Y$ and $\mathcal{D} = \left(Q^2 - \Pi\right)^2 \ge 0$ for the two coefficients divided by K^2 and K^4. The discriminant factorizes,

$$\mathcal{T}^2 - 4\mathcal{D} = \left[2\Pi - X - Y\right] \left[4Q^2 - X - Y - 2\Pi\right] \,,$$

because $\mathcal{T} \pm 2\left(Q^2 - \Pi\right)$ are the two brackets. Put $\pi_+ := \sqrt{X}$, $\pi_- := \sqrt{Y}$, so $\Pi = \pm \pi_+ \pi_-$ by $XY = \Pi^2$, the sign being that of $-MM'$. If $MM' > 0$ then $\Pi = -\pi_+\pi_-$ and the factorization reads $-\left(\pi_+ + \pi_-\right)^2 \left[4Q^2 - \left(\pi_+ - \pi_-\right)^2\right]$, non-negative exactly when $2\,|Q| \le |\pi_+ - \pi_-|$; if $MM' < 0$ then $\Pi = +\pi_+\pi_-$ and it reads $-\left(\pi_+ - \pi_-\right)^2 \left[4Q^2 - \left(\pi_+ + \pi_-\right)^2\right]$, non-negative exactly when $2\,|Q| \le \pi_+ + \pi_-$. Both cases are covered by $2\,|Q| \le |EM - M'/E|\,\pi/a^2$, since EM and M'/E have the same sign as M and M' respectively; and the closed form on the right of Equation (A.1193) follows from Equation (A.1186),

$$EM - \frac{M'}{E} = \frac{1}{c}\left[(\pi t - \psi)\,\mathrm{e}^{u} - (\pi t + \psi)\,\mathrm{e}^{-u}\right] = \frac{2}{c}\left[\pi t \sinh u - \psi \cosh u\right] \,.$$

Finally the second requirement, $\mathcal{T} \le 0$, is implied by the first: in either case $4Q^2 \le \left(\pi_+ \pm \pi_-\right)^2 \le 2\left(X + Y\right)$, so $2Q^2 \le X + Y$. □

A.61.5 Only one spacing survives

Theorem A.729 (Kármán's spacing ratio)**.** *The staggered street of Definition A.724 is neutrally stable to every infinitesimal disturbance Equation (A.1182) if and only if*

$$\cosh\frac{\pi h}{a}=\sqrt{2}\,,\qquad \textit{that is}\qquad \frac{h}{a}=\frac{1}{\pi}\log\left(1+\sqrt{2}\right)=0.280\,55\,, \tag{A.1194}$$

which is Equation (31.54). For every other ratio some mode grows exponentially.
Rests on Proposition A.728 and Definition A.724.

Proof. *Derives Theorem A.729.*

Necessity. Take $\varphi=\pi$, that is $\psi=0$: neighbouring vortices of a row are then displaced in opposite senses, the disturbance of longest wavelength that is not a rigid translation. Then $u=0$, $E=1$ and $M=M'=\pi t/c$, so the right-hand side of Equation (A.1193) vanishes identically and the criterion demands $Q=0$. By Equation (A.1188) with $\psi=0$ this is $\pi^2/2=\pi^2/c^2$, i.e. Equation (A.1194). Solving, $\sinh s=\sqrt{c^2-1}=1$ and $s=\log\left(c+\sinh s\right)=\log\left(1+\sqrt{2}\right)$, whence Equation (A.1172).

Sufficiency. Let $c=\sqrt{2}$, so $t=1/\sqrt{2}$, $\sinh s=1$ and $s=\log\left(1+\sqrt{2}\right)$. Then $Q=-\psi^2/2a^2$ by Equation (A.1188), and Equation (A.1193) becomes, after multiplying by a^2 and using $c=\sqrt{2}$,

$$\psi^2\le\pi\left|\pi\sinh u-\sqrt{2}\,\psi\cosh u\right|\,,\qquad u=\frac{\psi s}{\pi}\,. \tag{A.1195}$$

Both sides are even in ψ, so take $0\le\psi\le\pi$, where the bracket is negative and Equation (A.1195) is the assertion $G(u)\ge0$ for $0\le u\le s$, with $\psi=\pi u/s$ and

$$G(u):=\frac{\cosh s}{s}\,u\cosh u-\sinh u-\frac{u^2}{s^2}\,,$$

obtained by dividing Equation (A.1195) by π^2 and substituting $\sqrt{2}=\cosh s$, $1=\sinh s$. Now $G(0)=0$ and $G(s)=\cosh^2 s-\sinh s-1=2-1-1=0$; moreover

$$\begin{aligned}G'(u)&=\frac{\cosh s}{s}\left(\cosh u+u\sinh u\right)-\cosh u-\frac{2u}{s^2}\,,\\ G''(u)&=\left(\frac{2\cosh s}{s}-1\right)\sinh u+\frac{\cosh s}{s}\,u\cosh u-\frac{2}{s^2}\,.\end{aligned}$$

Since $2\cosh s/s>1$, every u-dependent term of G'' is strictly increasing on $[0,s]$, so G'' is; and $G''(0)=-2/s^2<0$ while $G''(s)=2\sqrt{2}/s+1-2/s^2>0$ numerically ($s=0.881\,37$ gives $3.209+1-2.575$). Hence G'' changes sign exactly once, so G' decreases and then increases. As $G'(0)=\cosh s/s-1>0$ and $G'(s)=\cosh^2 s/s+\cosh s\sinh s-\cosh s-2/s=0$ — the two terms $\cosh^2 s/s=2/s$ and $-2/s$ cancel, as do $\cosh s\sinh s=\sqrt{2}$ and $-\cosh s=-\sqrt{2}$ — the function G' is positive on $(0,u^*)$ and negative on (u^*,s) for a single u^*. Therefore G rises from $G(0)=0$ and falls back to $G(s)=0$ without crossing zero in between, so $G\ge0$ on $[0,s]$, which is Equation (A.1195). Equality holds only at $\psi=0$ and $\psi=\pi$, the mode of Equation (A.1194) and the rigid translation.

Instability elsewhere. If $c\ne\sqrt{2}$ then $Q\ne0$ at $\psi=0$, while $X=Y=-\Pi=\pi^4t^2/a^4c^2$ there, so the discriminant computed in Proposition A.728 equals $[4\Pi]\,[4Q^2]=16\Pi Q^2<0$: the roots λ are a complex conjugate pair, their square roots are not purely imaginary, and the mode $\varphi=\pi$ grows. □

Proposition A.730 (The unstaggered street is always unstable)**.** *If the second row is placed at $B_n=na-\mathrm{i}h$, directly beneath the first, the mode $\varphi=\pi$ grows for every $h>0$.*
Rests on Theorem A.729.

Proof. *Derives Proposition A.730.*

The only change is in the offset, so Lemma A.722 is applied at $\zeta = -\mathrm{i}h/a$ instead of $\frac{1}{2} - \mathrm{i}h/a$, giving $\sin\pi\zeta = -\mathrm{i}\sinh s$ and $\cos\pi\zeta = \cosh s$; hence $S_0 = -\pi^2/a^2\sinh^2 s$, and S_1, $\tilde{S}_1$ come out *real*. At $\psi = 0$ they are equal, with $\sqrt{X} = \pi^2\cosh s/a^2\sinh^2 s$ and $\Pi = +X$, so the second case of Proposition A.728 applies and neutrality would require $|Q| \le \sqrt{X}$. Here $Q = (\pi^2/a^2)\left(\frac{1}{2} + 1/\sinh^2 s\right)$, so the requirement reads $\frac{1}{2}\sinh^2 s + 1 \le \cosh s$, i.e. $(\cosh s - 1)^2 \le 0$ after using $\sinh^2 s = \cosh^2 s - 1$. That holds only at $s = 0$, where the two rows coincide. For every real separation the unstaggered arrangement therefore grows, and no spacing rescues it — which is why only the staggered pattern is ever photographed. □

Remark A.731 (What the calculation settles, and what it does not). Four limitations are built into the statement proved above and none of them is repaired by any refinement of the algebra.

First, the stability is *neutral*. Theorem A.729 says that at the critical ratio no linear mode grows; it does not say that any decays, and it says nothing at second order, where the street is in fact unstable. The observed consequence is that the photographed spacing drifts slowly downstream instead of locking to Equation (A.1172), and the agreement with the photographs — close to a few percent — is better than a neutral result has any right to expect.

Second, the vortices are points in an unbounded ideal fluid. Real vortices have cores of finite size, diffuse by Equation (31.52), and decay; the calculation has no viscosity in it anywhere.

Third, and most important, *the cylinder never appears*. The configuration Equation (A.1177) is postulated, not derived: nothing here explains why two boundary layers separating from a bluff body should roll up into two staggered rows, nor with what strength Γ, nor at what rate. In particular nothing here predicts the Strouhal number, and Remark 31.62 says so at length: Equation (31.53) is a measurement, disciplined by the similarity argument of Phenomenon 31.49 into the form $\mathrm{St} = \mathrm{St}(\mathrm{Re})$ and no further.

Fourth, the analysis is of a rigid infinite pattern already formed. The selection it performs is therefore of the same kind as the selection of a preferred wavelength in Section 31.8.1: it says which configurations can persist, not which one nature will build.

Remark A.732. Appendix A.61 discharges Equation (31.54) of Phenomenon 31.61 in Chapter 31, and only that equation: Equation (31.53), the Strouhal number that shares the phenomenon with it, is reported there as measurement and is not derived here or anywhere else in this treatise. The two halves of that phenomenon are of different epistemic kinds and Remark 31.62 is where the difference is set out; the reader who has followed the algebra above should return to it, because the calculation just performed makes concrete how much idealization purchased the one number that could be derived. The measured record for the cylinder wake is collected in Chapter 36.

A.62 The Joukowski Map, the Kutta Condition and the 2π Lift Slope

This appendix proves Proposition 31.70 of Chapter 31 — that the Kutta condition selects Equation (31.60) from the one-parameter family of irrotational flows past a thin profile, giving a lift-curve slope of exactly 2π per radian. The chapter proves Theorem 31.68, which converts a circulation into a force and says nothing about what the circulation is; Remark 31.69 names the physical selection principle and defers the computation here. Three things are established below: the flow past a circle with arbitrary circulation (Appendix A.62.1), the conformal transfer to the profile with the trailing edge as the distinguished point (Appendix A.62.2), and the value of Γ that the finiteness of the velocity there enforces (Appendix A.62.3).

The sign convention is the chapter's throughout: Γ is the circulation measured counterclockwise, Equation (31.55), and the transverse force per unit span is $Y = -\rho U\Gamma$, Equation (31.59). The complex velocity is Equation (31.57), $\mathrm{d}w/\mathrm{d}z = v_x - \mathrm{i}v_y$.

A.62.1 Flow past a circle with circulation

Lemma A.733 (The circle with incidence and circulation)**.** *Let $U > 0$ in* m/s*, $R > 0$ in* m*, $\alpha \in (-\pi/2, \pi/2)$ and $\Gamma \in \mathbb{R}$ in* m^2/s*. The function*

$$w(\zeta) = U\left(\zeta\,\mathrm{e}^{-\mathrm{i}\alpha} + \frac{R^2\mathrm{e}^{\mathrm{i}\alpha}}{\zeta}\right) - \frac{\mathrm{i}\Gamma}{2\pi}\log\zeta \tag{A.1196}$$

is holomorphic on $|\zeta| > R$ apart from the branch cut of the logarithm, has $|\zeta| = R$ as a streamline, tends to the uniform stream $U\mathrm{e}^{\mathrm{i}\alpha}$ at infinity, and carries circulation Γ around the circle. Its derivative on the circle is

$$\left.\frac{\mathrm{d}w}{\mathrm{d}\zeta}\right|_{\zeta=R\mathrm{e}^{\mathrm{i}\theta}} = \mathrm{i}\,\mathrm{e}^{-\mathrm{i}\theta}\left[2U\sin(\theta-\alpha) - \frac{\Gamma}{2\pi R}\right], \tag{A.1197}$$

so the stagnation points on the circle are at the angles θ with

$$\sin(\theta-\alpha) = \frac{\Gamma}{4\pi U R}. \tag{A.1198}$$

Rests on Proposition 31.28 and Equation (31.57).

Proof. *Derives Lemma A.733.*

On $\zeta = R\mathrm{e}^{\mathrm{i}\theta}$ the first bracket of Equation (A.1196) is $UR\left(\mathrm{e}^{\mathrm{i}(\theta-\alpha)} + \mathrm{e}^{-\mathrm{i}(\theta-\alpha)}\right) = 2UR\cos(\theta-\alpha)$, which is real, and $\log\zeta = \log R + \mathrm{i}\theta$, so $\operatorname{Im} w = -(\Gamma/2\pi)\log R$, a constant: the circle is a streamline. As $\zeta \to \infty$, $\mathrm{d}w/\mathrm{d}\zeta \to U\mathrm{e}^{-\mathrm{i}\alpha}$, which by Equation (31.57) is the stream $\boldsymbol{v} = U(\cos\alpha, \sin\alpha)$. The circulation is $\oint(\mathrm{d}w/\mathrm{d}\zeta)\,\mathrm{d}\zeta$ around the circle; the first bracket is single valued and contributes nothing, and $-(\mathrm{i}\Gamma/2\pi)\oint\mathrm{d}\zeta/\zeta = -(\mathrm{i}\Gamma/2\pi)\,2\pi\mathrm{i} = \Gamma$ by Theorem 8.24.

For Equation (A.1197), differentiate Equation (A.1196) and set $\zeta = R\mathrm{e}^{\mathrm{i}\theta}$:

$$\frac{\mathrm{d}w}{\mathrm{d}\zeta} = U\left(\mathrm{e}^{-\mathrm{i}\alpha} - \frac{R^2\mathrm{e}^{\mathrm{i}\alpha}}{\zeta^2}\right) - \frac{\mathrm{i}\Gamma}{2\pi\zeta} = U\mathrm{e}^{-\mathrm{i}\theta}\left(\mathrm{e}^{\mathrm{i}(\theta-\alpha)} - \mathrm{e}^{-\mathrm{i}(\theta-\alpha)}\right) - \frac{\mathrm{i}\Gamma\mathrm{e}^{-\mathrm{i}\theta}}{2\pi R},$$

and the bracket is $2\mathrm{i}\sin(\theta-\alpha)$ by Equation (8.3). Factoring out $\mathrm{i}\mathrm{e}^{-\mathrm{i}\theta}$ gives Equation (A.1197), which vanishes exactly at the angles Equation (A.1198).

Uniqueness among flows with the stated data is the Neumann uniqueness of Proposition 31.28: two such potentials differ by a function whose velocity vanishes at infinity, whose normal derivative vanishes on the circle, and around which the circulation is zero, hence by the maximum principle (Theorem 10.77) a constant. □

Note that Equation (A.1198) has a solution only for $|\Gamma| \le 4\pi UR$; larger circulations lift the stagnation points off the circle entirely. Nothing in the potential problem picks a value — every Γ gives a legitimate irrotational flow that does not penetrate the body, which is exactly the indeterminacy stated at Remark 31.69.

A.62.2 The Joukowski map and the cusped edge

Definition A.734 (The Joukowski map)**.** For $b > 0$ in m, the map

$$z = Z(\zeta) := \zeta + \frac{b^2}{\zeta} \tag{A.1199}$$

carries $|\zeta| > b$ bijectively onto the complement of the segment $[-2b, 2b]$ of the real axis, with inverse

$$\zeta = \tfrac{1}{2}\left(z + \sqrt{z^2 - 4b^2}\right), \tag{A.1200}$$

the branch of the square root behaving as z at infinity.

Rests on Theorem 8.6.

That Equation (A.1200) inverts Equation (A.1199) is the quadratic formula applied to $\zeta^2 - z\zeta + b^2 = 0$, whose two roots have product b^2; the branch named is the one of modulus greater than b, and it is holomorphic off the segment because $z^2 - 4b^2$ vanishes only at $z = \pm 2b$. The circle $|\zeta| = b$ itself maps two-to-one onto the segment: $\zeta = b\mathrm{e}^{\mathrm{i}\theta}$ gives $z = 2b\cos\theta$. The chord is therefore $c = 4b$, and the point $\zeta = b$ maps to the trailing edge $z = 2b$.

Lemma A.735 (Transfer of the flow)**.** *Let w be as in Equation (A.1196) with $R = b$, and put $W(z) := w(\zeta(z))$ with Equation (A.1200). Then W is a complex potential for the flow past the segment: it is holomorphic off the segment, the segment is a streamline, the free stream at infinity is $U\mathrm{e}^{\mathrm{i}\alpha}$, the circulation around the segment is Γ, and the physical velocity is*

$$\frac{\mathrm{d}W}{\mathrm{d}z} = \frac{\mathrm{d}w/\mathrm{d}\zeta}{\mathrm{d}z/\mathrm{d}\zeta} = \frac{\mathrm{d}w/\mathrm{d}\zeta}{1 - b^2/\zeta^2}\,. \tag{A.1201}$$

Rests on Definition A.734 and Lemma A.733.

Proof. *Derives Lemma A.735.*

A composition of holomorphic functions is holomorphic and its derivative is given by the chain rule; both statements follow from Theorem 8.6 and Definition 8.5 exactly as for real functions, and $\mathrm{d}z/\mathrm{d}\zeta = 1 - b^2/\zeta^2$ is non-zero on $|\zeta| > b$, which gives Equation (A.1201). Since W takes the same values as w, its imaginary part is constant on the image of the circle, i.e. on the segment, and its behaviour at infinity is that of w because $\zeta(z) = z + O\left(z^{-1}\right)$ there. The circulation is unchanged for the same reason: it is the real part of $\oint_{C_z} \mathrm{d}W$, and $\mathrm{d}W = \mathrm{d}w$ under the substitution $z \mapsto \zeta(z)$, which carries a circuit around the segment to a circuit around the circle. □

Remark A.736 (What conformal mapping buys, and what it is). The step just taken is the only one where a general theorem might have been invoked, and it has been avoided. What is usually quoted here is that a holomorphic bijection carries harmonic functions to harmonic functions and streamlines to streamlines, so a solved problem in one plane is a solved problem in the other; the general statement needs the holomorphic inverse function theorem, and beyond it, to know that a suitable map exists at all, the Riemann mapping theorem. Part II's complex-analysis chapter carries neither: Chapter 8 ends at Theorem 8.24 and has no conformal-mapping material, which is a real gap and is recorded as such. Nothing above needs it, because the map Equation (A.1199) and its inverse Equation (A.1200) are both written down explicitly, and the only property used is that a composition of holomorphic functions is holomorphic — which is Theorem 8.6 and the chain rule.

The transformation is degenerate at $\zeta = \pm b$, where $\mathrm{d}z/\mathrm{d}\zeta = 0$. This is not an accident of the map but the definition of a sharp edge: the interior angle of the profile at $z = 2b$ is zero, and a smooth curve is being folded onto itself. Equation (A.1201) then says that the physical velocity at the trailing edge is infinite — *unless* the numerator vanishes there too.

A.62.3 The Kutta condition and the lift

Theorem A.737 (The lift-curve slope of the flat plate)**.** *Of the one-parameter family of flows Equation (A.1196) with $R = b$, exactly one has a finite velocity at the trailing edge $z = 2b$, namely that with*

$$\Gamma = -4\pi U b \sin\alpha = -\pi c U \sin\alpha\,, \qquad c = 4b\,, \tag{A.1202}$$

and for it Equation (31.59) *gives the lift per unit span and lift coefficient*

$$Y = \pi\rho U^2 c \sin\alpha\,, \qquad C_{\mathrm{L}} = \frac{Y}{\frac{1}{2}\rho U^2 c} = 2\pi\sin\alpha \approx 2\pi\alpha\,, \tag{A.1203}$$

which is Equation (31.60).

Rests on Lemma A.735 and Theorem 31.68.

Proof. *Derives Theorem A.737.*

By Equation (A.1201) the velocity at $z = 2b$ is the quotient of $\mathrm{d}w/\mathrm{d}\zeta$ and $1 - b^2/\zeta^2$ as $\zeta \to b$ along the circle. The denominator has a simple zero there: with $\zeta = b\mathrm{e}^{\mathrm{i}\theta}$, $1 - \mathrm{e}^{-2\mathrm{i}\theta} = 2\mathrm{i}\mathrm{e}^{-\mathrm{i}\theta}\sin\theta$, which vanishes linearly in θ. The numerator is Equation (A.1197) with $R = b$. The quotient is therefore

$$\frac{\mathrm{d}W}{\mathrm{d}z} = \frac{\mathrm{i}\mathrm{e}^{-\mathrm{i}\theta}\left[2U\sin(\theta-\alpha) - \Gamma/2\pi b\right]}{2\mathrm{i}\mathrm{e}^{-\mathrm{i}\theta}\sin\theta} = \frac{2U\sin(\theta-\alpha) - \Gamma/2\pi b}{2\sin\theta}\,,$$

and its limit as $\theta \to 0$ is finite if and only if the numerator vanishes at $\theta = 0$, that is $-2U\sin\alpha = \Gamma/2\pi b$, which is Equation (A.1202); every other value of Γ gives a velocity diverging like $1/\theta$. (When the condition holds, l'Hôpital's rule gives the finite edge velocity $U\cos\alpha$, the component of the stream along the plate — the flow leaves the edge smoothly, along the chord.)

For the force, Lemma A.735 places a circulation Γ around a body in a uniform stream of speed U, so Theorem 31.68 applies in axes aligned with the stream: the drag vanishes and the transverse force is $-\rho U\Gamma = \pi\rho U^2 c\sin\alpha$, directed perpendicular to the stream and, since $\Gamma < 0$ for $\alpha > 0$, upwards. Dividing by $\frac{1}{2}\rho U^2 c$ gives Equation (A.1203), and for small incidence $\sin\alpha = \alpha + O(\alpha^3)$, so $\Gamma \to -\pi c U\alpha$ and $C_{\mathrm{L}} \to 2\pi\alpha$ as in Equation (31.60). □

Remark A.738 (Why the circulation may be read off in either plane). The step that makes the calculation possible is that Γ is the same number in the two planes. It is worth isolating, because it is what licenses computing the circulation where the geometry is easy — a circle — and applying Theorem 31.68 where the geometry is the one that matters. The reason is that the circulation is a contour integral of an exact differential, $\oint \mathrm{d}W$, and a change of variable in a contour integral is a substitution and nothing more: no property of the map enters beyond its being holomorphic and one-to-one. The same remark explains why the Blasius formula Equation (31.58) may be evaluated on a large circle rather than on the profile — Corollary 8.15, used already in the chapter's proof of Lemma 31.67.

Remark A.739 (What the 2π is worth). Equation (A.1203) is a genuine prediction with no adjustable content, and measured slopes for thin two-dimensional sections at small incidence do come within a few percent of 2π per radian. Four qualifications are owed.

Thickness and camber. The circle of Lemma A.733 was concentric with the origin, which is what degenerated the profile to a segment. Displacing its centre gives thickness (a shift along the real axis) and camber (a shift along the imaginary axis), and the same argument then yields $C_{\mathrm{L}} = 2\pi\sin(\alpha - \alpha_0)$: camber moves the zero-lift angle α_0 and leaves the slope alone, which is why the number 2π is so much more robust than anything else in aerofoil theory.

Finite span. A real wing sheds trailing vorticity, whose downwash reduces the effective incidence, and the slope falls with aspect ratio. This is a three-dimensional effect and nothing in a plane flow can see it.

The Kutta condition is a viscous fact. It has been imposed here, not derived. Its justification is that the boundary layer of Section 31.7.2 cannot negotiate the infinite adverse pressure gradient implied by an infinite edge velocity and separates there instead — Proposition 31.73 — and its consistency with the conservation of circulation Corollary 31.65 is secured by the starting vortex described at Remark 31.69. An inviscid theory that selects

its solution by a viscous criterion is not a closed theory, and it should not be presented as one.

No stall. Equation (A.1203) rises without bound with α, which is false: beyond ten to fifteen degrees the flow separates from the upper surface and the lift collapses. The theory contains no mechanism for this whatever, because it contains no separation. The measured lift curve, its linear range and its breakdown are in Chapter 36.

Remark A.740. Appendix A.62 discharges Proposition 31.70 of Chapter 31, whose statement Equation (31.60) is proved here as Equation (A.1203), and completes Remark 31.69 by supplying the value of Γ that the chapter's discussion of the Kutta condition leaves open. It should be read against Appendix A.59: the circulation is the first of the two escapes from d'Alembert's theorem catalogued at Remark 31.31, and what this section computes is precisely how much of it a sharp trailing edge is worth. The drag is still exactly zero, here as there — Theorem 31.68 gives $X = 0$ — so the escape buys lift and nothing else. Kutta's own computation is [307]; the general circulation theorem is Joukowski's [362].

A.63 The Kármán–Howarth Relation and the Four-Fifths Law

This appendix proves Equation (31.74) of Remark 31.86 in Chapter 31:

$$\left\langle \left(\delta u_{\parallel}\right)^{3} \right\rangle = -\tfrac{4}{5}\,\varepsilon\, r \tag{A.1204}$$

in the inertial range of homogeneous isotropic turbulence, with no adjustable constant. The chapter calls this the one nontrivial statement about turbulence that is *derived* from the Navier–Stokes equations rather than assumed and tested, and sets it deliberately against the dimensional argument of Phenomenon 31.84, which is not derived. The whole weight of that claim rests on the pages below, and the section is written accordingly: everything is proved, the symmetry hypotheses are used explicitly at the four places where they are needed, and Remark A.750 states at the end exactly how much the word *exact* is buying. It is buying less than it sounds, and more than anything else in the subject.

A.63.1 Setting

Definition A.741 (The statistical setting)**.** Let $\boldsymbol{v}(\boldsymbol{x},t)$ be a random velocity field on $\mathbb{R}^3$ obeying the incompressible Navier–Stokes equations Equation (31.32) with kinematic viscosity ν in m^2/s and a body force $\boldsymbol{F}$ per unit mass, and let $\langle\,\cdot\,\rangle$ denote the ensemble average. The turbulence is assumed

1. *homogeneous*: every joint moment is invariant under translations of $\mathbb{R}^3$;
2. *isotropic in the strong sense*: every joint moment is invariant under the full orthogonal group $\mathrm{O}(3)$, rotations and reflections alike;
3. *statistically steady*: every moment is independent of t;
4. *forced at large scales only*: $\boldsymbol{F}$ is a random field with the same symmetries whose correlation with $\boldsymbol{v}$ varies on a single length L, the integral scale.

Write $\boldsymbol{v} := \boldsymbol{v}(\boldsymbol{x},t)$, $\boldsymbol{v}' := \boldsymbol{v}(\boldsymbol{x}+\boldsymbol{r},t)$, $\delta\boldsymbol{v} := \boldsymbol{v}' - \boldsymbol{v}$, $r = |\boldsymbol{r}|$, $\boldsymbol{n} = \boldsymbol{r}/r$, and $\delta u_{\parallel} := \delta\boldsymbol{v}\cdot\boldsymbol{n}$ for the longitudinal increment. Define

$$u^2 := \left\langle v_1^2 \right\rangle, \quad R(r) := \langle \boldsymbol{v}\cdot\boldsymbol{v}' \rangle, \quad S_2(r) := \left\langle \left(\delta u_{\parallel}\right)^{2} \right\rangle, \quad S_3(r) := \left\langle \left(\delta u_{\parallel}\right)^{3} \right\rangle, \tag{A.1205}$$

and let

$$\varepsilon := \nu \, \langle \partial_j v_i \, \partial_j v_i \rangle \tag{A.1206}$$

be the mean dissipation rate per unit mass, in W/kg=m^2/s^3.

Rests on Theorem 31.37 and Definition 11.4.

Hypothesis (2) is stronger than rotational invariance alone and the difference matters: a turbulence with a mean helicity is invariant under SO(3) but not under reflection, and the pseudo-tensor terms that reflection invariance removes below would then survive.

A.63.2 Isotropic tensor functions of one vector

Everything in this section rests on one algebraic lemma, which is proved here rather than quoted because the whole reduction from a field theory to two ordinary differential equations is contained in it.

Lemma A.742 (Isotropic representation)**.** *Let T be a Cartesian tensor field on $\mathbb{R}^3 \setminus \{0\}$, depending on the single vector $\boldsymbol{r}$, and covariant under* O(3)*: $T_{i\ldots}(\mathsf{O}\boldsymbol{r}) = \mathsf{O}_{ii'} \cdots T_{i'\ldots}(\boldsymbol{r})$ for every orthogonal O. Then, with $\boldsymbol{n} = \boldsymbol{r}/r$,*

$$\textit{rank } 1: \quad T_i = \phi(r)\, n_i \,, \tag{A.1207}$$

$$\textit{rank } 2,\ \textit{symmetric}: \quad T_{ij} = (\phi_1 - \phi_2)\, n_i n_j + \phi_2\, \delta_{ij} \,, \tag{A.1208}$$

$$\textit{rank } 3,\ \textit{symmetric in } i,j: \quad T_{ijk} = \mathcal{A}\, n_i n_j n_k + \mathcal{B}\, \delta_{ij} n_k + \mathcal{C}\, (\delta_{ik} n_j + \delta_{jk} n_i) \,, \tag{A.1209}$$

with scalar functions of r alone. A constant tensor with these symmetries and no $\boldsymbol{r}$ dependence is zero at odd rank and a multiple of δ_{ij} at rank two.

Rests on Theorem 5.128 and Definition 14.26.

Proof. *Derives Lemma A.742.*

Fix $r > 0$. Covariance under the rotations that move $\boldsymbol{r}$ shows that the components in a frame adapted to $\boldsymbol{r}$ depend on r only, so it suffices to work at one point. Choose axes with $\boldsymbol{n} = \hat{\boldsymbol{e}}_3$ and let Latin indices a, b, c run over the transverse values $1, 2$. The subgroup of O(3) fixing $\boldsymbol{r}$ is the group O(2) acting on the transverse plane, and each block of components must be an invariant tensor of that group.

At rank one: T_a is an invariant vector of O(2), hence zero (the element $-\mathsf{I}$ of O(2) reverses it), and T_3 is free. This is Equation (A.1207).

At rank two, symmetric: T_{3a} is again an invariant transverse vector, hence zero; T_{ab} is an invariant symmetric two-dimensional tensor, hence $\phi_2 \delta_{ab}$, since the traceless part carries the spin-two representation of SO(2), which has no invariant; and $T_{33} = \phi_1$ is free. Writing $\delta_{ab} = \delta_{ij} - n_i n_j$ in covariant form gives Equation (A.1208).

At rank three, symmetric in the first two indices, four blocks occur. T_{33a} is an invariant transverse vector: zero. T_{ab3} is an invariant symmetric transverse two-tensor: $\mathcal{B}\, \delta_{ab}$. $T_{3ab} = T_{a3b}$ is an invariant transverse two-tensor, not required symmetric, hence again $\mathcal{C}\, \delta_{ab}$ — the antisymmetric ϵ_{ab} is invariant under SO(2) but changes sign under a reflection of the plane, and hypothesis (2) of Definition A.741 excludes it. T_{abc} is an invariant transverse tensor of odd rank, and $-\mathsf{I} \in$ SO(2) reverses it: zero. Assembling with $\delta_{ab} = \delta_{ij} - n_i n_j$ and renaming the free coefficient of $n_i n_j n_k$ gives Equation (A.1209).

For a constant tensor the same argument applies at every $\boldsymbol{r}$ simultaneously, so the surviving coefficients must be independent of the direction $\boldsymbol{n}$; at odd rank every term in Equations (A.1207) and (A.1209) carries an odd number of factors n, and no such expression is direction independent unless it vanishes. □

Corollary A.743 (Two scalar functions instead of a tensor field)**.** *Write $B_{ij}(\boldsymbol{r}) := \left\langle v_i v'_j \right\rangle$. Then there is a single scalar f, with $f(0) = 1$, such that*

$$B_{ij} = u^2 \left[(f - g)\, n_i n_j + g\, \delta_{ij} \right] , \qquad g = f + \tfrac{1}{2} r f' \,, \tag{A.1210}$$

and $R(r) = B_{ii} = u^2\,(3f + rf')$*; moreover* $\langle v_i v_j v_k' \rangle$ *is determined by a single scalar* k*, namely* $\left\langle v_\parallel^2 v_\parallel' \right\rangle = u^3 k(r)$*, and* $\langle v_i v_j v_k \rangle = 0$*.*

Rests on Lemma A.742 and Corollary 31.13.

Proof. *Derives Corollary A.743.*

B_{ij} is symmetric under the simultaneous exchange $i \leftrightarrow j$, $\boldsymbol{r} \to -\boldsymbol{r}$, and by Equation (A.1208) it takes the displayed form with $f = B_{\parallel\parallel}/u^2$ and $g = B_{\perp\perp}/u^2$, the longitudinal and transverse correlations; $f(0) = 1$ by the definition of u^2. Incompressibility acts through $\partial B_{ij}/\partial r_j = \left\langle v_i\, \partial_j' v_j' \right\rangle = 0$. Writing $h := (f - g)/r^2$ so that the first term is $u^2 h\, r_i r_j$,

$$\partial_j\,(h\, r_i r_j) = h'\, r\, r_i + 4h\, r_i\,, \qquad \partial_j\,(g\,\delta_{ij}) = g'\frac{r_i}{r}\,,$$

using $\partial_j r = r_j/r$ and $\partial_j\,(r_i r_j) = 4r_i$. Their sum vanishes for all $\boldsymbol{r}$, so $h'r + 4h + g'/r = 0$; substituting h and clearing r^2 leaves $rf' + 2\,(f - g) = 0$, which is the second relation in Equation (A.1210). Contracting the first relation gives $R = u^2\,(f - g + 3g) = u^2\,(f + 2g) = u^2\,(3f + rf')$.

For the triple correlation put $B_{ij,k} := \langle v_i v_j v_k' \rangle$; it is symmetric in i, j, so Equation (A.1209) applies with coefficients $\mathcal{A}, \mathcal{B}, \mathcal{C}$. Incompressibility at the far point gives $\partial B_{ij,k}/\partial r_k = \langle v_i v_j \partial_k' v_k' \rangle = 0$; carrying out the differentiation as above,

$$\partial_k\,(\mathcal{A} n_i n_j n_k) = n_i n_j \left(\mathcal{A}' + \frac{2\mathcal{A}}{r}\right), \quad \partial_k\,(\mathcal{B}\delta_{ij} n_k) = \delta_{ij}\left(\mathcal{B}' + \frac{2\mathcal{B}}{r}\right),$$

$$\partial_k\Big[\mathcal{C}\,(\delta_{ik} n_j + \delta_{jk} n_i)\Big] = 2\mathcal{C}' n_i n_j + \frac{2\mathcal{C}}{r}\,(\delta_{ij} - n_i n_j)\,,$$

and since $n_i n_j$ and δ_{ij} are independent the two coefficients must vanish separately:

$$\mathcal{A}' + \frac{2\mathcal{A}}{r} + 2\mathcal{C}' - \frac{2\mathcal{C}}{r} = 0\,, \qquad \mathcal{B}' + \frac{2\mathcal{B}}{r} + \frac{2\mathcal{C}}{r} = 0\,. \tag{A.1211}$$

Two equations for three functions leave one free function, which may be taken to be $u^3 k := B_{\parallel\parallel,\parallel} = \mathcal{A} + \mathcal{B} + 2\mathcal{C}$. Finally $\langle v_i v_j v_k \rangle$ is a constant isotropic tensor of rank three, so it vanishes by the last clause of Lemma A.742. □

A.63.3 The pressure drops out

This is the step that makes the result exact rather than a closure, and it is the one a reader should doubt until it is shown, since a correlation involving the pressure is otherwise as intractable as anything in the subject.

Lemma A.744 (Vanishing of the pressure–velocity correlation)**.** *Under Definition A.741,* $\langle p\, v_i' \rangle = 0$ *identically, and likewise* $\left\langle |\boldsymbol{v}|^2\, v_i' \right\rangle = 0$*.*

Rests on Lemma A.742 and Corollary 31.13.

Proof. *Derives Lemma A.744.*

Both are isotropic vector functions of $\boldsymbol{r}$, so by Equation (A.1207) each equals $\phi(r) n_i$ for some scalar ϕ. Both are divergence free in $\boldsymbol{r}$: differentiating with respect to r_i is differentiating the far point, and

$$\partial_i\,\langle p\, v_i' \rangle = \langle p\, \partial_i' v_i' \rangle = 0\,, \qquad \partial_i \left\langle |\boldsymbol{v}|^2\, v_i' \right\rangle = \left\langle |\boldsymbol{v}|^2\, \partial_i' v_i' \right\rangle = 0\,,$$

because the field is incompressible (Corollary 31.13) and the near point is held fixed. For a radial field $\phi(r) n_i$ the divergence is $r^{-2}\,(r^2\phi)'$, so $r^2\phi$ is constant; the constant is zero because ϕ is bounded at the origin, both quantities being finite one-point moments there. Hence $\phi \equiv 0$. □

A.63.4 The two-point energy equation

Theorem A.745 (Kármán–Howarth–Monin relation)**.** *Under Definition A.741, and without any closure hypothesis whatever,*

$$\frac{\partial R}{\partial t} = \tfrac{1}{2}\,\boldsymbol{\nabla}_r \cdot \left\langle |\delta \boldsymbol{v}|^2\, \delta \boldsymbol{v} \right\rangle + 2\nu \nabla_r^2 R + \mathcal{F}(r)\,, \qquad \mathcal{F} := \left\langle \boldsymbol{F}\cdot\boldsymbol{v}' + \boldsymbol{v}\cdot\boldsymbol{F}' \right\rangle\,, \tag{A.1212}$$

and $\mathcal{F}(0) = 2\varepsilon$ in a steady state.
Rests on Theorem 31.37 and Lemma A.744.

Proof. *Derives Theorem A.745.*

By homogeneity every correlation depends on $\boldsymbol{r} = \boldsymbol{x}' - \boldsymbol{x}$ alone, so $\partial/\partial x_i = -\partial/\partial r_i$ and $\partial/\partial x_i' = +\partial/\partial r_i$ when applied to one. Write the Navier–Stokes equation Equation (31.32) at $\boldsymbol{x}$ and at $\boldsymbol{x}'$, multiply the first by v_i' and the second by v_i, add and average:

$$\frac{\partial R}{\partial t} = -\left\langle (v_j \partial_j v_i)\, v_i' \right\rangle - \left\langle v_i \left(v_j' \partial_j' v_i' \right) \right\rangle - \frac{1}{\rho}\left[\left\langle (\partial_i p)\, v_i' \right\rangle + \left\langle v_i \partial_i' p' \right\rangle \right] + \nu \left[\left\langle \left(\nabla^2 v_i \right) v_i' \right\rangle + \left\langle v_i \nabla'^2 v_i' \right\rangle \right] + \mathcal{F}\,.$$

The pressure bracket is $\partial_i \langle p\, v_i' \rangle - \partial_i \langle v_i p' \rangle$ after moving the derivatives outside the averages, and both terms vanish by Lemma A.744 — the second is the first with $\boldsymbol{r}$ reversed. The viscous bracket is $2\nu\nabla_r^2 R$, since each Laplacian acts on one factor and becomes ∇_r^2 on the correlation.

For the nonlinear terms use incompressibility to write $v_j \partial_j v_i = \partial_j (v_j v_i)$; then

$$-\left\langle \partial_j (v_j v_i)\, v_i' \right\rangle = +\partial_{r_j} \left\langle v_j v_i v_i' \right\rangle\,, \qquad -\left\langle v_i \partial_j' \left(v_j' v_i' \right) \right\rangle = -\partial_{r_j} \left\langle v_i v_j' v_i' \right\rangle\,,$$

so the nonlinear contribution is $\partial_{r_j}\left[\langle v_j v_i v_i' \rangle - \left\langle v_i v_j' v_i' \right\rangle \right]$. It remains to identify the bracket. Expand

$$\left\langle |\delta \boldsymbol{v}|^2\, \delta v_j \right\rangle = \left\langle \left(|\boldsymbol{v}'|^2 - 2\boldsymbol{v}\cdot\boldsymbol{v}' + |\boldsymbol{v}|^2 \right) \left(v_j' - v_j \right) \right\rangle\,.$$

The terms $\left\langle |\boldsymbol{v}'|^2\, v_j' \right\rangle$ and $\left\langle |\boldsymbol{v}|^2\, v_j \right\rangle$ are equal by homogeneity and cancel; $\left\langle |\boldsymbol{v}|^2\, v_j' \right\rangle$ and $\left\langle |\boldsymbol{v}'|^2\, v_j \right\rangle$ vanish by Lemma A.744; and what is left is

$$\left\langle |\delta \boldsymbol{v}|^2\, \delta v_j \right\rangle = 2\left\langle \boldsymbol{v}\cdot\boldsymbol{v}'\, v_j \right\rangle - 2\left\langle \boldsymbol{v}\cdot\boldsymbol{v}'\, v_j' \right\rangle = 2\left[\langle v_j v_i v_i' \rangle - \left\langle v_i v_j' v_i' \right\rangle \right]\,,$$

which is twice the bracket. This gives Equation (A.1212).

At $\boldsymbol{r} = \boldsymbol{0}$: the divergence term vanishes, because $\left\langle |\delta \boldsymbol{v}|^2\, \delta \boldsymbol{v} \right\rangle = O(r^3)$ there; the viscous term is $2\nu\nabla_r^2 R|_0 = 2\nu \langle v_i \nabla^2 v_i \rangle = -2\nu \langle \partial_j v_i \partial_j v_i \rangle = -2\varepsilon$ by Equation (A.1206) and homogeneity; and $\partial_t R(0) = 0$ in a steady state. Hence $\mathcal{F}(0) = 2\varepsilon$, which is the statement that in a steady state the force injects energy at exactly the rate viscosity destroys it. □

Proposition A.746 (The Kármán–Howarth equation in classical form)**.** *For unforced (decaying) homogeneous isotropic turbulence, Equation (A.1212) is equivalent to*

$$\frac{\partial}{\partial t}\left(u^2 f \right) = \frac{u^3}{r^4}\,\partial_r\left(r^4 k \right) + \frac{2\nu u^2}{r^4}\,\partial_r\left(r^4\, \partial_r f \right)\,, \tag{A.1213}$$

with f and k the scalars of Corollary A.743.
Rests on Theorem A.745 and Corollary A.743.

Proof. *Derives Proposition A.746.*

Write $T(r)$ for the scalar in $\left\langle |\delta \boldsymbol{v}|^2\, \delta v_j \right\rangle = T(r) n_j$, legitimate by Equation (A.1207); then $\boldsymbol{\nabla}_r \cdot \left\langle |\delta \boldsymbol{v}|^2\, \delta \boldsymbol{v} \right\rangle = r^{-2}\left(r^2 T \right)'$ and $\nabla_r^2 R = r^{-2}\left(r^2 R' \right)'$, so Equation (A.1212) unforced reads

$$\frac{\partial R}{\partial t} = \frac{1}{2r^2}\left(r^2 T \right)' + \frac{2\nu}{r^2}\left(r^2 R' \right)'\,. \tag{A.1214}$$

By Corollary A.743, $R = u^2\,(3f + rf') = u^2\,(r^3 f)'\,/r^2$, and Appendix A.63.5 below gives $T = u^3\,(8k + 2rk')$. Apply the operator $\mathcal{L}[\Phi] := r^{-2}\partial_r\,(r^3\Phi)$ to each term of Equation (A.1213). On the left, $\mathcal{L}\big[\partial_t(u^2 f)\big] = \partial_t R$. On the right,

$$\mathcal{L}\left[\frac{u^3}{r^4}\left(r^4 k\right)'\right] = \frac{u^3}{r^2}\left[4r^2 k + r^3 k'\right]' = \frac{u^3}{r^2}\left[8rk + 7r^2 k' + r^3 k''\right] = \frac{1}{2r^2}\left(r^2 T\right)' ,$$

the last step by direct differentiation of $r^2 T = u^3\,(8r^2 k + 2r^3 k')$; and

$$\mathcal{L}\left[\frac{2\nu u^2}{r^4}\left(r^4 f'\right)'\right] = \frac{2\nu u^2}{r^2}\left[4r^2 f' + r^3 f''\right]' = \frac{2\nu u^2}{r^2}\left[8rf' + 7r^2 f'' + r^3 f'''\right] = \frac{2\nu}{r^2}\left(r^2 R'\right)' ,$$

since $r^2 R' = u^2\,(4r^2 f' + r^3 f'')$. So Equation (A.1213) maps term by term onto Equation (A.1214); and $\mathcal{L}$ is injective on functions regular at the origin, since $\mathcal{L}[\Phi] = 0$ forces $r^3\Phi$ constant, hence $\Phi \propto r^{-3}$, hence $\Phi = 0$. □

Equation (A.1213) is the Kármán–Howarth relation, obtained by von Kármán and Howarth in 1938 (*Proceedings of the Royal Society of London* A **164**, 192–215). That paper is not in the bibliography of this treatise, and the statement is not being taken from it: everything above is derived here from Equation (31.32) and Definition A.741, so the attribution is historical and nothing rests on it. The equation is exact and it does not close: one equation relates the two unknown functions f and k, and nothing determines k. What Kolmogorov saw is that in one limit the undetermined function drops out.

A.63.5 Structure functions

Lemma A.747 (The third-order structure functions)**.** *With $S_{ijk} := \langle \delta v_i \delta v_j \delta v_k \rangle$ and $T(r)$ as above,*

$$S_3 = 6u^3 k\,, \qquad T = u^3\,(8k + 2rk') = S_3 + \tfrac{1}{3}\,(rS_3)'\ . \tag{A.1215}$$

Rests on Corollary A.743 and Lemma A.742.

Proof. *Derives Lemma A.747.*

Expand S_{ijk} in the eight products of v and v'. The two single-point terms $\langle v_i v_j v_k \rangle$ and $\langle v'_i v'_j v'_k \rangle$ vanish by Corollary A.743. Of the remaining six, three have two factors at $\boldsymbol{x}$ and three have two at $\boldsymbol{x} + \boldsymbol{r}$; using homogeneity to write the latter as B evaluated at $-\boldsymbol{r}$, and the fact that $B_{ij,k}$ is odd — every term of Equation (A.1209) carries an odd number of factors $\boldsymbol{n}$ — the two groups add rather than cancel:

$$S_{ijk} = 2\,[B_{ij,k} + B_{ik,j} + B_{jk,i}] = 6\mathcal{A}\,n_i n_j n_k + 2\,(\mathcal{B} + 2\mathcal{C})\,(\delta_{ij} n_k + \delta_{ik} n_j + \delta_{jk} n_i)\ ,$$

the second equality by adding the three copies of Equation (A.1209) with indices permuted. Write $W := \mathcal{A} + \mathcal{B} + 2\mathcal{C} = u^3 k$ and $V := \mathcal{B} + 2\mathcal{C}$. Contracting with $n_i n_j n_k$ gives $S_3 = 6\mathcal{A} + 6V = 6W$, the first relation in Equation (A.1215); contracting on $i = j$ and then with n_k gives $T = S_{iik} n_k = 6\mathcal{A} + 10V = S_3 + 4V$.

It remains to express V through W. Adding the two relations Equation (A.1211) and using $\mathcal{A} + \mathcal{B} = W - 2\mathcal{C}$,

$$W' + \frac{2\,(W - 2\mathcal{C})}{r} = 0 \quad \Longrightarrow \quad \mathcal{C} = \frac{rW' + 2W}{4}\,,$$

and substituting $\mathcal{B} = V - 2\mathcal{C}$ into the second of Equation (A.1211) gives $V' + 2V/r = 2\mathcal{C}' + 2\mathcal{C}/r$. With the expression just obtained for $\mathcal{C}$, the right-hand side is $[rW'' + 4W' + 2W/r]\,/2$, so

$$\left(r^2 V\right)' = r^2 V' + 2rV = \tfrac{1}{2}\left[r^3 W'' + 4r^2 W' + 2rW\right] = \tfrac{1}{2}\,\frac{\mathrm{d}}{\mathrm{d}r}\left[r^3 W' + r^2 W\right]\ ,$$

the last step because $(r^3W')' = 3r^2W' + r^3W''$ and $(r^2W)' = 2rW + r^2W'$. Integrating and discarding the r^{-2} homogeneous solution by regularity at the origin, $V = \frac{1}{2}(rW)'$. Hence $T = S_3 + 2(rW)' = S_3 + \frac{1}{3}(rS_3)'$, and in terms of k, $T = 6u^3k + 2u^3(k + rk') = u^3(8k + 2rk')$. □

Lemma A.748 (The second-order structure function)**.** $D(r) := \left\langle |\delta \boldsymbol{v}|^2 \right\rangle = 2[R(0) - R(r)] = 3S_2 + rS_2'$.

Rests on Corollary A.743.

Proof. *Derives Lemma A.748.*

$\left\langle |\delta \boldsymbol{v}|^2 \right\rangle = \left\langle |\boldsymbol{v}|^2 \right\rangle + \left\langle |\boldsymbol{v}'|^2 \right\rangle - 2\langle \boldsymbol{v} \cdot \boldsymbol{v}' \rangle$, and the first two are $R(0) = 3u^2$ each by homogeneity. For the second equality, Equation (A.1210) gives $S_2 = 2u^2(1 - f)$ and the transverse structure function $\left\langle (\delta u_\perp)^2 \right\rangle = 2u^2(1 - g) = S_2 + \frac{1}{2}rS_2'$, using $g = f + \frac{1}{2}rf'$; summing one longitudinal and two transverse contributions gives $D = 3S_2 + rS_2'$. □

A.63.6 The inertial range and the four-fifths law

Theorem A.749 (Kolmogorov's equation and the four-fifths law)**.** *Under Definition A.741, for every separation $r \ll L$,*

$$S_3(r) - 6\nu \frac{\mathrm{d}S_2}{\mathrm{d}r} = -\tfrac{4}{5}\varepsilon r, \tag{A.1216}$$

and consequently, in the inertial range $\eta \ll r \ll L$ where the viscous term is negligible,

$$\left\langle \left(\delta u_\parallel\right)^3 \right\rangle = -\tfrac{4}{5}\varepsilon r, \tag{A.1217}$$

which is Equation (31.74). Here $\eta = (\nu^3/\varepsilon)^{1/4}$ is the Kolmogorov length Equation (31.72).

Rests on Theorem A.745 and Lemmas A.747 and A.748.

Proof. *Derives Theorem A.749.*

In a steady state the left-hand side of Equation (A.1212) vanishes, and in radial form

$$0 = \frac{1}{2r^2}\left(r^2T\right)' + \frac{2\nu}{r^2}\left(r^2R'\right)' + \mathcal{F}(r).$$

Multiply by r^2 and integrate from 0 to r. The first term gives $\frac{1}{2}r^2T$, the boundary term at the origin vanishing because $T = O(r^3)$. The second gives $2\nu r^2 R'(r)$, likewise. For the third, $\mathcal{F}$ varies on the scale L by hypothesis (4), so for $r \ll L$ it may be replaced by $\mathcal{F}(0) = 2\varepsilon$, giving $\frac{2}{3}\varepsilon r^3$. Hence

$$0 = \tfrac{1}{2}r^2T + 2\nu r^2R' + \tfrac{2}{3}\varepsilon r^3, \qquad \text{i.e.} \qquad T = -4\nu R' - \tfrac{4}{3}\varepsilon r.$$

By Lemma A.748, $R = R(0) - \frac{1}{2}D$, so $R' = -\frac{1}{2}D'$ and

$$T = 2\nu D' - \tfrac{4}{3}\varepsilon r. \tag{A.1218}$$

This is already exact and already the whole content; the rest is translation. Substituting $T = (r^4S_3)'/3r^3$ — which is Equation (A.1215) rewritten, since $(r^4S_3)' = r^3(4S_3 + rS_3') = 3r^3\left[S_3 + \frac{1}{3}(rS_3)'\right]$ — and $D = (r^3S_2)'/r^2$ from Lemma A.748, one verifies that Equation (A.1216) solves Equation (A.1218): with $S_3 = 6\nu S_2' - \frac{4}{5}\varepsilon r$,

$$\frac{(r^4S_3)'}{3r^3} = \frac{6\nu\left(r^4S_2'\right)' - 4\varepsilon r^4}{3r^3} = 2\nu\left[4S_2' + rS_2''\right] - \tfrac{4}{3}\varepsilon r = 2\nu D' - \tfrac{4}{3}\varepsilon r,$$

because $D' = (3S_2 + rS_2')' = 4S_2' + rS_2''$. Since Equation (A.1218) determines S_3 given S_2 — it is a first-order linear equation with the regularity condition $S_3 = O(r^3)$ at the origin fixing the integration constant — Equation (A.1216) is the solution.

For Equation (A.1217): the viscous term $6\nu S_2'$ is of order $\nu S_2/r$, and in the inertial range $S_2 \sim (\varepsilon r)^{2/3}$, so the ratio of the viscous to the inertial term is of order $\nu\varepsilon^{2/3}r^{-4/3}/\varepsilon r = (\eta/r)^{4/3}$, which is negligible for $r \gg \eta$; and $r \ll L$ was used already. Dropping it leaves Equation (A.1217). □

The direction of the inequality is the physics. $S_3 < 0$ says that the odd moment of the longitudinal increment is negative: the fluid at $\boldsymbol{x}+\boldsymbol{r}$ is, on average, moving *towards* the fluid at $\boldsymbol{x}$ more strongly than away from it, in the cubic sense. That asymmetry is the cascade, and Equation (A.1217) measures it against ε with no free constant. Reversed sign would be an inverse cascade, which is what two-dimensional turbulence does and three-dimensional turbulence does not.

Remark A.750 (Exactly what was assumed). The chapter calls Equation (31.74) the one exact nontrivial result about turbulence, and the reader is owed the bill.

Homogeneity was used at every step where $\partial/\partial x_i = -\partial/\partial r_i$ was applied, and in Lemma A.748. No real flow is homogeneous — there is always a boundary and always a forcing region — so the statement is about an idealized ensemble that laboratory turbulence approximates locally.

Isotropy, including reflection, was used three times: in Lemma A.742 to reduce tensor fields to scalars, in Lemma A.744 to kill the pressure term, and in Lemma A.747. Anisotropy at the large scales survives into the inertial range more persistently than K41 supposes, and the four-fifths law is correspondingly harder to observe cleanly than its exactness suggests.

Stationarity set $\partial_t R = 0$. For decaying turbulence the same calculation retains a term in ∂_t and Equation (A.1217) acquires a correction that is small only when the decay time exceeds the eddy turnover time at separation r.

The ordering of the limits. Two neglects were made, $r \ll L$ and $r \gg \eta$, and they are compatible only when $L/\eta \gg 1$, that is at large Reynolds number by Equation (31.73). The four-fifths law is a statement about an asymptotic regime, and the inertial range at attainable Reynolds numbers is short.

Finite dissipation as $\nu \to 0$. The whole argument treats ε as a fixed quantity while the viscous term is dropped. That ε tends to a non-zero limit as $\nu \to 0$ — the dissipation anomaly — is an assumption, not a theorem; it is strongly supported by measurement and it is what makes the cascade picture consistent, but it is precisely the kind of statement whose proof would require knowing that the equations have the solutions Remark 31.48 says are not known to exist.

What survives all five qualifications is still remarkable, and the comparison with Phenomenon 31.84 is the point of stating it: the exponent $-5/3$ follows from hypotheses that measurement has falsified in detail — the intermittency corrections are real — while the coefficient $-4/5$ follows from the equations themselves, and measurement has found no correction to it. Remark 31.87 sets the whole position out, and Chapter 36 collects the measurements a future theory will be asked to reproduce.

Remark A.751 (Attribution). Equation (A.1217) is Kolmogorov's, and the paper that contains it is the companion of [857] — the dissipation paper, *Doklady Akademii Nauk SSSR* **32** (1941) 16–18, recorded in the bibliographic note of that entry. The entry [857] itself is the local-structure paper, which states the similarity hypotheses used for Equation (31.71). Remark 31.86 already warns that the K41 attributions must be read carefully, the spectral form being Obukhov's [861]; the same care applies within Kolmogorov's own output, and the two 1941 papers are not interchangeable.

Remark A.752. Appendix A.63 discharges Equation (31.74) inside Remark 31.86 of Chapter 31, and with it the chapter's claim that exactly one nontrivial consequence of Equation (31.32) about turbulence is derived rather than assumed. The reader returning there

should carry back two things this section supplies and the chapter cannot: that the Kármán–Howarth relation Equation (A.1213) is exact but does *not* close, one equation in two unknown functions, which is the closure problem of Remark 31.83 in its sharpest form; and that the four-fifths law survives the non-closure only because the undetermined function is exactly the one the inertial-range limit eliminates. Everything else in Section 31.9 is dimensional analysis and measurement, as Remark 31.87 says.

A.64 The Korteweg–de Vries Equation from the Free-Surface Problem

This appendix derives Equation (31.81) of Chapter 31 from the exact equations of an ideal free-surface layer. The chapter, in the derivation of Phenomenon 31.96, takes the Korteweg–de Vries equation as given, integrates it for a travelling wave, and obtains the profile Equation (31.82) with Russell's speed Equation (31.80); what it owes, and what is supplied here, is the equation itself [265]. The travelling-wave integration is *not* repeated below — Equation (31.82) stands where it is.

The derivation is an expansion in two small parameters, and the single most important thing about it is that the relation between them is *chosen*, not deduced. That choice is stated up front in Appendix A.64.2 and revisited at Remark A.753; a reader who takes it for a consequence has misunderstood the whole calculation.

A.64.1 The exact free-surface problem

Take an ideal, incompressible, irrotational layer of undisturbed depth h over a flat rigid bottom at $y = 0$, with the free surface at $y = h + \eta(x, t)$ and gravity g acting downwards. Surface tension is set to zero here and restored in Remark A.758. By Proposition 31.28 the velocity is $\boldsymbol{v} = \boldsymbol{\nabla}\phi$ with

$$\partial_x^2\phi + \partial_y^2\phi = 0 \quad \text{for } 0 < y < h + \eta\,, \qquad \partial_y\phi = 0 \text{ at } y = 0\,, \tag{A.1219}$$

the second being impermeability of the bottom. At the surface two conditions hold, both nonlinear and both imposed on a boundary whose position is itself unknown:

$$\partial_t\eta + \partial_x\phi\,\partial_x\eta = \partial_y\phi \qquad \text{at } y = h + \eta\,, \tag{A.1220}$$

$$\partial_t\phi + \tfrac{1}{2}\left[(\partial_x\phi)^2 + (\partial_y\phi)^2\right] + g\eta = 0 \qquad \text{at } y = h + \eta\,. \tag{A.1221}$$

Equation (A.1220) says the surface is material — a particle on it stays on it — and Equation (A.1221) is the unsteady Bernoulli equation Equation (31.21) evaluated at the surface, where the pressure equals the constant atmospheric value, absorbed into ϕ.

A.64.2 Two small parameters, and a choice

Let a be a typical wave amplitude and ℓ a typical horizontal length, and set

$$\epsilon := \frac{a}{h}\,, \qquad \delta^2 := \frac{h^2}{\ell^2}\,, \qquad c_0 := \sqrt{gh}\,. \tag{A.1222}$$

Nondimensionalize by

$$x = \ell\tilde{x}\,, \quad y = h\tilde{y}\,, \quad t = \frac{\ell}{c_0}\tilde{t}\,, \quad \eta = a\tilde{\eta}\,, \quad \phi = \frac{ga\ell}{c_0}\tilde{\phi}\,, \tag{A.1223}$$

the scale for ϕ being the one that makes Equation (A.1221) balance at leading order. Substituting and dropping tildes, Equations (A.1219) to (A.1221) become exactly

$$\delta^2 \partial_x^2 \phi + \partial_y^2 \phi = 0\,, \qquad \partial_y \phi = 0 \text{ at } y = 0\,, \tag{A.1224}$$

$$\partial_t \eta + \epsilon\, \partial_x \phi\, \partial_x \eta = \frac{1}{\delta^2}\, \partial_y \phi \qquad \text{at } y = 1 + \epsilon\eta\,, \tag{A.1225}$$

$$\partial_t \phi + \frac{\epsilon}{2}\left[(\partial_x \phi)^2 + \frac{1}{\delta^2}\,(\partial_y \phi)^2 \right] + \eta = 0 \qquad \text{at } y = 1 + \epsilon\eta\,. \tag{A.1226}$$

No approximation has yet been made: the three coefficients ϵ, $1/\delta^2$ and ϵ/δ^2 are what the substitution produces, using $ga/c_0^2 = a/h = \epsilon$ and $\ell^2/h^2 = 1/\delta^2$.

Remark A.753 (The distinguished limit is a choice). Nothing so far relates ϵ to δ^2, and different relations give different equations. Taking $\epsilon \ll \delta^2$ and keeping only δ^2 gives the linear dispersive equation whose relation is Equation (A.1239) below; taking $\delta^2 \ll \epsilon$ and keeping only ϵ gives the nonlinear non-dispersive equation, whose solutions steepen and break in finite time. The Korteweg–de Vries equation is the *distinguished limit*

$$\epsilon = O\!\left(\delta^2\right)\,, \qquad \text{equivalently} \qquad \frac{a\ell^2}{h^3} = O(1)\,, \tag{A.1227}$$

in which the two effects enter at the same order and can balance. This is the entire physical content of the solitary wave — steepening against spreading, as Phenomenon 31.96 says — and it is a hypothesis about the class of initial data being described, not a consequence of the equations. Data not satisfying Equation (A.1227) are described by one of the other two limits and do not produce solitary waves. The dimensionless group in Equation (A.1227) is the Ursell number; the solitary wave of Equation (31.82) has $\ell^2 = 4h^3/3a$, i.e. Ursell number $4/3$ exactly, which is the self-consistency of the whole construction.

Both parameters are treated as first order below: terms of order ϵ and δ^2 are retained, and terms of order ϵ^2, $\epsilon\delta^2$ and δ^4 are discarded.

A.64.3 Eliminating the depth

Lemma A.754 (Exact depth expansion)**.** *Every solution of Equation* (A.1224) *analytic in* y *is*

$$\phi(x,y,t) = \sum_{n=0}^{\infty} \frac{(-1)^n\, \delta^{2n} y^{2n}}{(2n)!}\, \partial_x^{2n} F(x,t)\,, \qquad F(x,t) := \phi(x,0,t)\,, \tag{A.1228}$$

which for the retained orders reads $\phi = F - \frac{1}{2}\delta^2 y^2 \partial_x^2 F + \frac{1}{24}\delta^4 y^4 \partial_x^4 F + O(\delta^6)$.

Rests on Equation (A.1224) *and Theorem 7.38.*

Proof. *Derives Lemma A.754.*

Write $\phi = \sum_{m\geq 0} y^m \phi_m(x,t)/m!$. The bottom condition kills ϕ_1, and Equation (A.1224) gives the recursion $\phi_{m+2} = -\delta^2 \partial_x^2 \phi_m$; hence every odd coefficient vanishes and $\phi_{2n} = \left(-\delta^2\right)^n \partial_x^{2n} \phi_0$, which is Equation (A.1228) with $\phi_0 = F$. □

The lemma is exact and is what makes the problem one-dimensional: the whole vertical structure is slaved to the single function F, and the long-wave parameter δ^2 is precisely the bookkeeping parameter of that slaving. Introduce the horizontal velocity at the bottom,

$$u(x,t) := \partial_x F\,, \tag{A.1229}$$

and read off from Equation (A.1228)

$$\frac{1}{\delta^2}\partial_y \phi = -y\, \partial_x u + \frac{\delta^2 y^3}{6}\, \partial_x^3 u + O\!\left(\delta^4\right)\,, \qquad \partial_x \phi = u + O\!\left(\delta^2\right)\,, \qquad \partial_t \phi = \partial_t F - \frac{\delta^2 y^2}{2}\, \partial_x \partial_t u + O\!\left(\delta^4\right)\,. \tag{A.1230}$$

Proposition A.755 (The Boussinesq pair)**.** *To first order in ϵ and in δ^2, the free-surface conditions Equations* (A.1225) *and* (A.1226) *reduce to the closed system*

$$\partial_t \eta + \partial_x u + \epsilon\, \partial_x (\eta u) - \frac{\delta^2}{6}\, \partial_x^3 u = 0\,, \tag{A.1231}$$

$$\partial_t u + \partial_x \eta + \epsilon\, u\, \partial_x u - \frac{\delta^2}{2}\, \partial_x^2 \partial_t u = 0\,. \tag{A.1232}$$

Rests on Lemma A.754 and Equation (A.1225)*.*

Proof. *Derives Proposition A.755.*

Evaluate Equation (A.1230) at $y = 1 + \epsilon\eta$ and substitute in Equation (A.1225). The right-hand side is

$$-(1+\epsilon\eta)\,\partial_x u + \frac{\delta^2}{6}\,(1+\epsilon\eta)^3\,\partial_x^3 u + O\Big(\epsilon\delta^2, \delta^4\Big) = -\partial_x u - \epsilon\eta\,\partial_x u + \frac{\delta^2}{6}\partial_x^3 u + \cdots\,,$$

while the left-hand side is $\partial_t \eta + \epsilon u\, \partial_x \eta$. Collecting and using $\epsilon u\, \partial_x \eta + \epsilon \eta\, \partial_x u = \epsilon\, \partial_x (\eta u)$ gives Equation (A.1231).

In Equation (A.1226) the term $\epsilon\,(\partial_y \phi)^2 / 2\delta^2$ is $O(\epsilon\delta^2)$ by Equation (A.1230) and is discarded, while $\epsilon\,(\partial_x \phi)^2 / 2 = \epsilon u^2/2 + O(\epsilon\delta^2)$. Evaluating $\partial_t \phi$ at $y = 1 + \epsilon\eta$ to the retained order,

$$\partial_t F - \frac{\delta^2}{2}\,\partial_x \partial_t u + \frac{\epsilon}{2} u^2 + \eta = 0\,,$$

and differentiating once in x, with $\partial_x F = u$, gives Equation (A.1232). □

A.64.4 Unidirectional reduction

At $\epsilon = \delta^2 = 0$ the pair Equations (A.1231) and (A.1232) is $\partial_t \eta + \partial_x u = 0$, $\partial_t u + \partial_x \eta = 0$, whence $\partial_t^2 \eta = \partial_x^2 \eta$: two waves, one running each way, with $u = \eta$ on the right-running one. Restrict to that branch and correct it.

Theorem A.756 (The Korteweg–de Vries equation)**.** *A right-running solution of Equations* (A.1231) *and* (A.1232)*, correct to first order in ϵ and δ^2, has*

$$u = \eta - \frac{\epsilon}{4}\eta^2 + \frac{\delta^2}{3}\,\partial_x^2 \eta + O\Big(\epsilon^2, \epsilon\delta^2, \delta^4\Big)\,, \tag{A.1233}$$

and η obeys

$$\partial_t \eta + \partial_x \eta + \tfrac{3}{2}\epsilon\,\eta\,\partial_x \eta + \tfrac{1}{6}\delta^2\,\partial_x^3 \eta = 0\,. \tag{A.1234}$$

In dimensional variables this is

$$\frac{\partial \eta}{\partial t} + c_0 \left(1 + \frac{3\eta}{2h}\right) \frac{\partial \eta}{\partial x} + \frac{c_0 h^2}{6}\,\frac{\partial^3 \eta}{\partial x^3} = 0\,, \qquad c_0 = \sqrt{gh}\,, \tag{A.1235}$$

which is Equation (31.81)*.*

Rests on Proposition A.755 and Equation (A.1227)*.*

Proof. *Derives Theorem A.756.*

Write $u = \eta + \epsilon A + \delta^2 B$ with A and B functions of η and its x derivatives, to be determined. Substituting into Equations (A.1231) and (A.1232) and discarding second-order terms,

$$\partial_t \eta + \partial_x \eta + \epsilon\,(\partial_x A + 2\eta\,\partial_x \eta) + \delta^2 \left(\partial_x B - \tfrac{1}{6}\partial_x^3 \eta\right) = 0\,, \tag{A.1236}$$

$$\partial_t \eta + \partial_x \eta + \epsilon\,(\partial_t A + \eta\,\partial_x \eta) + \delta^2 \left(\partial_t B - \tfrac{1}{2}\partial_x^2 \partial_t \eta\right) = 0\,. \tag{A.1237}$$

Inside a term already carrying a factor ϵ or δ^2 the leading-order relation $\partial_t = -\partial_x$ may be used, since the error is of second order; so $\partial_t A = -\partial_x A$, $\partial_t B = -\partial_x B$ and $\partial_x^2 \partial_t \eta = -\partial_x^3 \eta$, and Equation (A.1237) becomes

$$\partial_t \eta + \partial_x \eta + \epsilon \left(-\partial_x A + \eta\, \partial_x \eta\right) + \delta^2 \left(-\partial_x B + \tfrac{1}{2} \partial_x^3 \eta\right) = 0 \,.$$

The two equations must be *consistent*: a single function η cannot obey two different evolution laws. Subtracting them removes the leading part and leaves

$$\epsilon \left(2\partial_x A + \eta\, \partial_x \eta\right) + \delta^2 \left(2\partial_x B - \tfrac{2}{3} \partial_x^3 \eta\right) = 0 \,,$$

and since ϵ and δ^2 are independent parameters each bracket must vanish. Integrating, and taking the constants of integration to be zero so that u vanishes with η,

$$A = -\tfrac{1}{4} \eta^2 \,, \qquad B = \tfrac{1}{3} \partial_x^2 \eta \,, \tag{A.1238}$$

which is Equation (A.1233). Adding the two equations instead, the terms in A and B cancel identically and what survives is

$$2 \left(\partial_t \eta + \partial_x \eta\right) + 3\epsilon\, \eta\, \partial_x \eta + \tfrac{1}{3} \delta^2\, \partial_x^3 \eta = 0 \,,$$

which is Equation (A.1234).

For Equation (A.1235), undo Equation (A.1223): multiplying Equation (A.1234) by $a c_0 / \ell$ and using $\partial_{\tilde t} \tilde\eta = (\ell / a c_0)\, \partial_t \eta$, $\partial_{\tilde x} \tilde\eta = (\ell / a)\, \partial_x \eta$, $\epsilon \tilde\eta \partial_{\tilde x} \tilde\eta = (\ell / a h)\, \eta\, \partial_x \eta$ and $\delta^2 \partial_{\tilde x}^3 \tilde\eta = (h^2 \ell / a)\, \partial_x^3 \eta$, every occurrence of a and ℓ cancels and Equation (A.1235) results — as it must, since a and ℓ were bookkeeping devices and no physical statement may depend on them. □

Remark A.757 (No general perturbation theory was needed). The step just carried out is what is usually called the removal of secular terms, and it is often presented as an application of the method of multiple scales. Part II carries no such method — its asymptotic material is the WKB expansion of Section 9.6 and the matched expansions named at Definition 9.86 — and none is needed. The reason is visible in the proof: because Equations (A.1231) and (A.1232) are *two* equations for the same pair of unknowns, consistency alone fixes A and B, and the freedom that a general theory would have to organize into slow and fast variables has already been used up. Had the reduction been attempted on a single equation the situation would be different, and the apparatus would be owed.

A.64.5 Two checks

The chapter asserts, without demonstration, that the two nontrivial terms of Equation (31.81) are exactly the steepening and the dispersion. Both halves can be verified against results already proved.

Drop the third derivative. Equation (A.1235) becomes $\partial_t \eta + c_0 \left(1 + 3\eta/2h\right) \partial_x \eta = 0$, a quasilinear first-order equation whose solution is constant along the characteristics $\mathrm{d}x/\mathrm{d}t = c_0 \left(1 + 3\eta/2h\right)$. Higher parts of the profile therefore travel faster and overtake lower ones, the front steepens, and after a finite time the solution becomes multivalued: this is the simple-wave steepening of Remark 31.90, which in a compressible gas ends in a shock. Note that the amplitude-dependent speed $c_0 \left(1 + 3\eta/2h\right)$ is consistent with Russell's measured $c^2 = g\,(h + a)$ only through the factor $3/2$ that the expansion supplies and dimensional analysis cannot.

Drop the nonlinear term. Setting $\eta \propto \mathrm{e}^{\mathrm{i}(kx-\omega t)}$ in Equation (A.1235) without the $\eta\,\partial_x\eta$ term gives $-\mathrm{i}\omega + \mathrm{i}c_0 k + (c_0 h^2/6)\,(\mathrm{i}k)^3 = 0$, that is

$$\omega = c_0 k\left(1 - \tfrac{1}{6}\,(kh)^2\right) . \tag{A.1239}$$

Expanding the exact gravity-wave relation Equation (31.78) at $\gamma = 0$ with $\tanh(kh) = kh - \tfrac{1}{3}\,(kh)^3 + \cdots$,

$$\omega^2 = gk\tanh(kh) = ghk^2\left(1 - \tfrac{1}{3}\,(kh)^2 + \cdots\right) \quad\Longrightarrow\quad \omega = c_0 k\left(1 - \tfrac{1}{6}\,(kh)^2 + \cdots\right) ,$$

which is Equation (A.1239) exactly, coefficient included. The third-derivative term of Equation (31.81) is therefore the first dispersive correction to the shallow-water speed and nothing else, as the chapter says.

Remark A.758 (What is not in the equation). Four idealizations were made and each has an observable consequence.

No viscosity. Equation (A.1219) is the ideal-fluid problem, so the solitary wave of Equation (31.82) propagates without loss. Real waves in a channel decay, chiefly through the boundary layer on the bed and walls (Section 31.7.2), and Russell's own observation ended when his wave died away.

No surface tension. Restoring it means replacing the constant surface pressure in Equation (A.1221) by the Young–Laplace jump Equation (31.18), and the expansion then delivers the same equation with the dispersive coefficient $(c_0 h^2/2)\left(\tfrac{1}{3} - \tau\right)$, $\tau := \gamma/\rho g h^2$ the Bond parameter. The check above confirms the coefficient independently: expanding Equation (31.78) with γ retained gives $\omega = c_0 k\left[1 - \tfrac{1}{2}\,(kh)^2\left(\tfrac{1}{3} - \tau\right)\right]$. The sign of the dispersion therefore *reverses* at $\tau = 1/3$, that is for $h < \sqrt{3\gamma/\rho g}$, about 4.7 mm for clean water with $\gamma = 7.3 \times 10^{-2}\,\mathrm{N/m}$ and $\rho = 998\,\mathrm{kg/m^3}$. In a layer thinner than that there are no elevation solitary waves, and Equation (31.82) does not apply — a limitation invisible in the chapter's statement.

Irrotational and one-directional. The reduction of Appendix A.64.4 discards the left-running wave entirely. An initial disturbance at rest splits into two, and Equation (A.1235) describes one of them.

Small amplitude. The expansion is in $\epsilon = a/h$ with $\epsilon = O(\delta^2)$, so it is valid only for $a \ll h$ — Russell's regime, as the chapter notes in the derivation of Phenomenon 31.96. Solitary waves of large amplitude exist but are not described by Equation (31.81): they have a limiting height of about $0.8h$ and a corner at the crest, neither of which this equation knows about.

Remark A.759. Appendix A.64 discharges Equation (31.81) in the derivation of Phenomenon 31.96 in Chapter 31, which takes the equation as given and proceeds from it to the profile Equation (31.82) and to Russell's speed Equation (31.80) [116]. The reader returning there now knows what the two nontrivial terms are and where each comes from — the first from the ϵ ordering of the surface conditions, the second from the δ^2 ordering of the depth expansion — and, more importantly, that their coexistence was imposed by the choice Equation (A.1227) rather than discovered. The reading of Equation (31.82) as a soliton, with the elastic collisions and the infinite hierarchy of conserved quantities that follow, belongs to Chapter 32; nothing of that is visible in the derivation above, which is why Korteweg and de Vries did not see it [265].

A.65 The Lorenz System from Boussinesq Convection

This appendix proves Definition 32.62 of Chapter 32: that Equation (32.36), with the three parameters Equation (32.37), is what the equations of a convecting fluid layer become

when they are truncated to one velocity mode and two temperature modes. The point of writing it out is stated in Remark 32.2: the Lorenz system is almost always presented as a dimensionless curiosity, and it is only by carrying the reduction through in SI variables that σ, r and b can be read back as ratios of quantities a laboratory measures. Along the way the critical Rayleigh number $\mathrm{Ra_c} = 27\pi^4/4$ and the critical wavenumber $a^2 = \frac{1}{2}$, both quoted in Definition 32.62 and in Equation (31.68), are derived.

Nothing is imported from outside the treatise. The starting equations are those of Chapter 31 — the incompressible Navier–Stokes equations Equation (31.32) and the advection–diffusion equation for heat — specialised by the Boussinesq approximation, which is that chapter's business and is discussed at Phenomenon 31.78; the linear-stability calculation and its comparison with measurement are Section 36.4. The attribution of the mode set is to Saltzman, whose 1962 study of finite convective amplitude in the *Journal of the Atmospheric Sciences* supplied the expansion Lorenz truncated; that paper has no entry in this bibliography, so the attribution is made in words, and the truncation and its consequences are Lorenz's [1262].

A.65.1 The layer, in SI units

Definition A.760 (The convecting layer)**.** A horizontal layer of fluid occupies $0 \le z \le d$, with z measured upwards against gravity $\boldsymbol{g} = -g\hat{\boldsymbol{z}}$. The lower surface is held at temperature $T_0 + \Delta T$ and the upper at T_0, so that $\Delta T > 0$ means heating from below. The fluid is characterised by four measured properties, all functions of the working substance and of its mean state:

$$\nu \ \left(\mathrm{m^2/s}\right), \quad \kappa \ \left(\mathrm{m^2/s}\right), \quad \alpha_T \ \left(1/\mathrm{K}\right), \quad \rho_0 \ \left(\mathrm{kg/m^3}\right), \tag{A.1240}$$

the kinematic viscosity, the thermal diffusivity, the coefficient of thermal expansion and the reference density; d is in m, ΔT in K and g in $\mathrm{m/s^2}$.

Rests on Theorem 31.37 and Definition 31.51.

Proposition A.761 (The Boussinesq equations)**.** *Under the Boussinesq approximation — the density is treated as the constant ρ_0 everywhere except in the buoyancy force, where it is $\rho = \rho_0\left[1 - \alpha_T\left(T - T_0\right)\right]$, and the flow is treated as incompressible — the motion of the layer of Definition A.760 obeys*

$$\frac{\partial \boldsymbol{u}}{\partial t} + \left(\boldsymbol{u}\cdot\boldsymbol{\nabla}\right)\boldsymbol{u} = -\frac{1}{\rho_0}\boldsymbol{\nabla}p' + \nu\nabla^2\boldsymbol{u} + g\alpha_T\left(T - T_0\right)\hat{\boldsymbol{z}}\,, \tag{A.1241}$$

$$\frac{\partial T}{\partial t} + \left(\boldsymbol{u}\cdot\boldsymbol{\nabla}\right)T = \kappa\nabla^2 T\,, \qquad \boldsymbol{\nabla}\cdot\boldsymbol{u} = 0\,, \tag{A.1242}$$

with p' the departure of the pressure from its hydrostatic value. The state of rest, $\boldsymbol{u} = \boldsymbol{0}$, is a solution with the conducting *temperature profile*

$$T_{\mathrm{c}}(z) = T_0 + \Delta T\left(1 - \frac{z}{d}\right). \tag{A.1243}$$

Rests on Definition A.760, Theorem 31.37, and Phenomenon 31.78.

Proof. *Derives Proposition A.761.*

Equation (A.1241) is Equation (31.32) of Theorem 31.37 with the body force $\boldsymbol{g}$ retained and the density in that force alone allowed to vary. Writing $\rho\boldsymbol{g} = -\rho_0 g\hat{\boldsymbol{z}} + \rho_0 g\alpha_T(T - T_0)\hat{\boldsymbol{z}}$ and absorbing the first, constant, piece into the pressure — which is what p' means — leaves the buoyancy term displayed. Equation (A.1242) is the advection–diffusion equation for the temperature of an incompressible fluid, viscous heating being of order $\nu U^2/(\kappa\Delta T)$ relative to the terms kept and negligible in every laboratory convection experiment. For the conducting state, put $\boldsymbol{u} = \boldsymbol{0}$ and seek a steady T: Equation (A.1242) becomes $\nabla^2 T = 0$, whose solution with the imposed boundary values is the linear profile Equation (A.1243), and Equation (A.1241) is then satisfied by a p' depending on z alone. □

Remark A.762 (What is taken from Part III and not proved here). The Boussinesq approximation is an approximation and not a limit theorem: it is justified when $\alpha_T \Delta T \ll 1$ and when the layer is thin against the density scale height, and the honest statement of its domain belongs with the convection material of Chapter 31, not here. This is an internal dependency of Part III on itself rather than a debt owed by Part II, and it is the only thing in this section taken from elsewhere. For air at room temperature $\alpha_T \approx 3.4 \times 10^{-3}/\mathrm{K}$, so a layer driven by $\Delta T = 10\,\mathrm{K}$ has $\alpha_T \Delta T = 0.034$: the approximation is good to a few per cent, which is the accuracy at which Section 36.4 tests the threshold.

A.65.2 Rolls, the stream function, and the elimination of the pressure

Definition A.763 (Two-dimensional rolls)**.** Attention is restricted to motions independent of the horizontal coordinate y — *rolls* with axes along y. The velocity is then written in terms of a *stream function* $\psi(x, z, t)$ of SI dimension m^2/s,

$$\boldsymbol{u} = (-\partial_z \psi,\ 0,\ \partial_x \psi)\ , \tag{A.1244}$$

and the temperature by its departure from the conducting profile,

$$\theta(x, z, t) = T - T_{\mathrm{c}}(z)\ \ (\mathrm{K})\ . \tag{A.1245}$$

For any pair of functions the Jacobian is written

$$J\,(f, h) = \partial_x f\, \partial_z h - \partial_z f\, \partial_x h\,, \tag{A.1246}$$

so that $(\boldsymbol{u} \cdot \boldsymbol{\nabla})\, h = J(\psi, h)$ for the field Equation (A.1244).
Rests on Proposition A.761 and Equation (31.11).

Remark A.764 (The roll ansatz is a symmetry restriction, not a two-dimensional world). Definition A.763 restricts attention to solutions of the three-dimensional equations Equation (A.1241) and Equation (A.1242) that happen not to depend on one of the three spatial coordinates. It is a symmetry ansatz inside the observed $3 + 1$ spacetime, exactly like the axisymmetric ansatz used for Stokes flow, and it is not a model of a two-dimensional universe; the fluid, the layer and the rolls are all three-dimensional objects. The ansatz is a genuine restriction — real convection at large Ra is not roll-like — and that restriction, not the dimensionality of space, is what has to be watched: it is one of the approximations listed in Remark A.771.

Proposition A.765 (The roll equations)**.** *For motions of the form Equation (A.1244), the system Equation (A.1241)–Equation (A.1242) is equivalent to the pair*

$$\partial_t \nabla^2 \psi + J\left(\psi, \nabla^2 \psi\right) = \nu \nabla^4 \psi + g \alpha_T\, \partial_x \theta\,, \tag{A.1247}$$

$$\partial_t \theta + J\,(\psi, \theta) = \frac{\Delta T}{d}\, \partial_x \psi + \kappa \nabla^2 \theta\,, \tag{A.1248}$$

in which the pressure has disappeared and ∇^2, ∇^4 act in the (x, z) plane.
Rests on Definition A.763 and Proposition A.761.

Proof. *Derives Proposition A.765.*

Incompressibility. $\boldsymbol{\nabla} \cdot \boldsymbol{u} = -\partial_x \partial_z \psi + \partial_z \partial_x \psi = 0$ identically, so Equation (A.1244) solves the constraint of Equation (A.1242) and no Lagrange multiplier is needed.

Vorticity. For a y-independent flow the vorticity has only a y component,

$$\omega_y = \partial_z u_x - \partial_x u_z = -\partial_z^2 \psi - \partial_x^2 \psi = -\nabla^2 \psi\,. \tag{A.1249}$$

Taking the y component of the curl of Equation (A.1241) annihilates the pressure gradient, since the curl of a gradient vanishes; the buoyancy term $g \alpha_T \theta \hat{\boldsymbol{z}}$ — the z-dependent part

of $T - T_0$ having been absorbed into p' along with the hydrostatic piece — contributes $-g\alpha_T \partial_x \theta$; and the advective term contributes $(\boldsymbol{u}\cdot\boldsymbol{\nabla})\,\omega_y = J(\psi,\omega_y)$, the vortex-stretching term $(\boldsymbol{\omega}\cdot\boldsymbol{\nabla})\,\boldsymbol{u}$ vanishing because $\boldsymbol{\omega}$ points along y and nothing depends on y. Hence

$$\partial_t \omega_y + J(\psi,\omega_y) = \nu\nabla^2\omega_y - g\alpha_T \partial_x\theta\,, \tag{A.1250}$$

and substituting Equation (A.1249) and multiplying through by -1 gives Equation (A.1247).

Temperature. Substituting $T = T_c + \theta$ into Equation (A.1242), and using $\nabla^2 T_c = 0$ because Equation (A.1243) is linear in z,

$$\partial_t\theta + J(\psi,\theta) + u_z \frac{\mathrm{d}T_c}{\mathrm{d}z} = \kappa\nabla^2\theta\,, \tag{A.1251}$$

and $u_z = \partial_x\psi$ with $\mathrm{d}T_c/\mathrm{d}z = -\Delta T/d$ turns the third term into $-(\Delta T/d)\,\partial_x\psi$, which moved to the right is Equation (A.1248). The physical content of that term is the whole instability: fluid rising through the mean gradient ($\partial_x\psi > 0$) arrives warmer than its surroundings, which increases θ, which by Equation (A.1247) drives more rising motion. □

A.65.3 The scaling, and the two dimensionless groups

Proposition A.766 (Nondimensionalisation)**.** *Scale lengths by d, time by d^2/κ, the stream function by κ and the temperature departure by $\nu\kappa/(g\alpha_T d^3)$,*

$$x = d\tilde{x}\,,\quad z = d\tilde{z}\,,\quad t = \frac{d^2}{\kappa}\tilde{t}\,,\quad \psi = \kappa\tilde{\psi}\,,\quad \theta = \frac{\nu\kappa}{g\alpha_T d^3}\tilde{\theta}\,. \tag{A.1252}$$

Then Equation (A.1247)–Equation (A.1248) become, with the tildes dropped,

$$\frac{1}{\sigma}\left[\partial_t\nabla^2\psi + J\left(\psi,\nabla^2\psi\right)\right] = \nabla^4\psi + \partial_x\theta\,, \tag{A.1253}$$

$$\partial_t\theta + J(\psi,\theta) = \mathrm{Ra}\,\partial_x\psi + \nabla^2\theta\,, \tag{A.1254}$$

and exactly two dimensionless groups survive, the Prandtl and Rayleigh numbers

$$\sigma = \frac{\nu}{\kappa}\,,\qquad \mathrm{Ra} = \frac{g\alpha_T\,\Delta T\,d^3}{\nu\kappa}\,, \tag{A.1255}$$

which are Equation (32.37) and Equation (32.38).

Rests on Proposition A.765 and Definition 31.51.

Proof. *Derives Proposition A.766.*

Under Equation (A.1252), $\nabla^2 = d^{-2}\tilde{\nabla}^2$, $\partial_t = (\kappa/d^2)\,\partial_{\tilde{t}}$ and $J = d^{-2}\tilde{J}$ acting on the scaled fields. The four terms of Equation (A.1247) then carry the prefactors

$$\frac{\kappa^2}{d^4}\,,\quad \frac{\kappa^2}{d^4}\,,\quad \frac{\nu\kappa}{d^4}\,,\quad g\alpha_T\cdot\frac{1}{d}\cdot\frac{\nu\kappa}{g\alpha_T d^3} = \frac{\nu\kappa}{d^4}\,, \tag{A.1256}$$

so dividing by κ^2/d^4 leaves the factor $\nu/\kappa = \sigma$ on the right-hand pair and gives Equation (A.1253). The four terms of Equation (A.1248) carry

$$\frac{\nu\kappa^2}{g\alpha_T d^5}\,,\quad \frac{\nu\kappa^2}{g\alpha_T d^5}\,,\quad \frac{\Delta T\,\kappa}{d^2}\,,\quad \frac{\nu\kappa^2}{g\alpha_T d^5}\,, \tag{A.1257}$$

and dividing by the first leaves the third with the factor

$$\frac{\Delta T\,\kappa}{d^2}\cdot\frac{g\alpha_T d^5}{\nu\kappa^2} = \frac{g\alpha_T\Delta T d^3}{\nu\kappa} = \mathrm{Ra}\,, \tag{A.1258}$$

which is Equation (A.1254).

The dimension check. It is worth carrying out in full, because it is what allows σ and r to be read back as measurements:

$$[\mathrm{Ra}] = \frac{(\mathrm{m/s^2})\,(1/\mathrm{K})\,(\mathrm{K})\,(\mathrm{m^3})}{(\mathrm{m^2/s})\,(\mathrm{m^2/s})} = \frac{\mathrm{m^4/s^2}}{\mathrm{m^4/s^2}} = 1\,, \tag{A.1259}$$

and $\sigma = \nu/\kappa$ is a ratio of two quantities in $\mathrm{m^2/s}$, so both are pure numbers, as Definition 31.51 requires. That no third group appears is the substance of the reduction: the four material properties Equation (A.1240) and the three imposed quantities $g, \Delta T, d$ enter the scaled equations only through σ and Ra, and ρ_0 has dropped out entirely, having been absorbed into p'. □

A.65.4 Linear stability and the critical Rayleigh number

Theorem A.767 (Marginal stability of the conducting state)**.** *Let the boundaries $z = 0, 1$ be* stress-free *and perfectly conducting, so that $\psi = \partial_z^2 \psi = \theta = 0$ there. Then the conducting state is marginally stable to the roll disturbance of dimensionless horizontal wavenumber πa precisely when*

$$\mathrm{Ra} = \mathrm{Ra_c}(a) \equiv \frac{\pi^4 \left(1 + a^2\right)^3}{a^2}\,, \tag{A.1260}$$

and the minimum of Equation (A.1260) over a is attained at $a^2 = \frac{1}{2}$, where

$$\mathrm{Ra_c} = \frac{\pi^4\,(3/2)^3}{1/2} = \frac{27\pi^4}{4} = 657.5\,, \tag{A.1261}$$

the value quoted in Definition 32.62 and in Equation (31.68).

Rests on Proposition A.766 and Phenomenon 31.78.

Proof. *Derives Theorem A.767.*

Linearisation. Drop the Jacobians from Equation (A.1253)–Equation (A.1254); they are quadratic in the disturbance and cannot affect the threshold.

The modes. Take

$$\psi = \Psi \sin(\pi a x) \sin(\pi z)\,, \qquad \theta = \Theta \cos(\pi a x) \sin(\pi z)\,, \tag{A.1262}$$

which satisfy every boundary condition: $\sin(\pi z)$ and its second derivative vanish at $z = 0, 1$. Both are eigenfunctions of the Laplacian, $\nabla^2 \to -k^2$ with

$$k^2 = \pi^2 \left(1 + a^2\right)\,, \tag{A.1263}$$

the sum of the squared horizontal wavenumber πa and the squared vertical wavenumber π. Marginal stability means a steady disturbance, so set $\partial_t = 0$.

The two-by-two system. With $\partial_x \theta = -\pi a \Theta \sin(\pi a x) \sin(\pi z)$ and $\partial_x \psi = \pi a \Psi \cos(\pi a x) \sin(\pi z)$, the two linearised equations reduce to

$$0 = k^4 \Psi - \pi a \Theta\,, \qquad 0 = \mathrm{Ra}\,\pi a \Psi - k^2 \Theta\,. \tag{A.1264}$$

A nonzero solution exists exactly when the determinant vanishes, $k^4 \cdot k^2 = \mathrm{Ra}\,\pi^2 a^2$, that is

$$\mathrm{Ra} = \frac{k^6}{\pi^2 a^2} = \frac{\pi^6 \left(1 + a^2\right)^3}{\pi^2 a^2} = \frac{\pi^4 \left(1 + a^2\right)^3}{a^2}\,, \tag{A.1265}$$

which is Equation (A.1260).

The minimum. Write $X = a^2 > 0$ and $F(X) = (1+X)^3/X$. Then

$$F'(X) = \frac{3(1+X)^2 X - (1+X)^3}{X^2} = \frac{(1+X)^2(2X-1)}{X^2}, \tag{A.1266}$$

which is negative for $X < \frac{1}{2}$ and positive for $X > \frac{1}{2}$: the unique minimum is at $X = a^2 = \frac{1}{2}$, and substituting gives Equation (A.1261). Numerically $27\pi^4/4 = 657.51$. The corresponding horizontal wavelength is $2\pi/(\pi a) = 2/a = 2\sqrt{2}$ in units of d, which is the $\lambda_c = 2\sqrt{2}\,d$ of Equation (31.68): convection cells are about 1.4 times as wide as the layer is deep, a prediction with no free parameter that Section 36.4 tests. □

Remark A.768 (Two conventions for the wavenumber). Chapter 31 writes the marginal curve as $\mathrm{Ra}(a) = (\pi^2 + a^2)^3/a^2$ and Chapter 32 writes it as Equation (A.1260); the two are the same curve in different variables, and a reader comparing them will otherwise think one of them wrong. In Chapter 31 the symbol a is the horizontal wavenumber itself, in units of $1/d$; here and in Definition 32.62 it is that wavenumber in units of π/d, so that $a_{\text{ch. 14}} = \pi a$. Substituting turns one expression into the other, and the minimum sits at $a^2 = \pi^2/2$ in the first convention and $a^2 = \frac{1}{2}$ in the second. The second is used here because it is the one in which Equation (32.37) reads $b = 4/(1+a^2)$, giving Lorenz's $b = 8/3$.

Rests on Theorem A.767 and Equation (31.68).

A.65.5 The Galerkin truncation

Theorem A.769 (The Lorenz system)**.** *Fix a horizontal wavenumber a. Write $\tilde{x}$ for the dimensionless horizontal coordinate and z for the dimensionless vertical one, reserving $x(s)$, $y(s)$ and $z_*(s)$ for the three amplitudes — these last are the x, y and z of Equation (32.36), decorated here only so that no symbol does two jobs. Substitute into Equation (A.1253)–Equation (A.1254) the three-mode ansatz*

$$\psi = \frac{\sqrt{2}\,(1+a^2)}{a}\, x(s)\, \sin(\pi a \tilde{x}) \sin(\pi z), \tag{A.1267}$$

$$\theta = \frac{\mathrm{Ra}_c(a)}{\pi}\left[\sqrt{2}\, y(s) \cos(\pi a \tilde{x}) \sin(\pi z) - z_*(s) \sin(2\pi z)\right], \tag{A.1268}$$

with the dimensionless time

$$s = \pi^2\left(1+a^2\right)\tilde{t} = \frac{\pi^2(1+a^2)\,\kappa t}{d^2}, \tag{A.1269}$$

and project the resulting equations onto the three modes. Discarding the single term that leaves the retained set, the amplitudes obey

$$\dot{x} = \sigma(y - x), \qquad \dot{y} = x(r - z_*) - y, \qquad \dot{z}_* = xy - b z_*, \tag{A.1270}$$

with the dot denoting $\mathrm{d}/\mathrm{d}s$ *and*

$$\sigma = \frac{\nu}{\kappa}, \qquad r = \frac{\mathrm{Ra}}{\mathrm{Ra}_c(a)}, \qquad b = \frac{4}{1+a^2}. \tag{A.1271}$$

These are Equation (32.36) and Equation (32.37); at the critical wavenumber $a^2 = \frac{1}{2}$ they give $b = 8/3$.

Rests on Proposition A.766, Theorem A.767, and Definition 32.62.

Proof. *Derives Theorem A.769.*

Write $\tilde{x}$ for the horizontal coordinate throughout, to keep it apart from the amplitude $x(s)$, and z_* for the third amplitude, to keep it apart from the vertical coordinate z. Put

$$\psi = A\, x\, S_1, \qquad \theta = B\, y\, C_1 - C\, z_* S_2, \tag{A.1272}$$

with constants A, B, C to be determined and the abbreviations

$$S_1 = \sin(\pi a\tilde{x})\sin(\pi z)\ , \quad C_1 = \cos(\pi a\tilde{x})\sin(\pi z)\ , \quad S_2 = \sin(2\pi z)\ . \tag{A.1273}$$

All three are mutually orthogonal over one horizontal period and over $0 \le z \le 1$, and all satisfy the boundary conditions of Theorem A.767.

Step 1: the vorticity equation has no nonlinearity. Since $\nabla^2\psi = -k^2\psi$ with k^2 from Equation (A.1263), the Jacobian $J(\psi, \nabla^2\psi) = -k^2 J(\psi,\psi) = 0$ identically: a single roll does not advect its own vorticity. Using $\partial_{\tilde{t}} = k^2 \mathrm{d}/\mathrm{d}s$ from Equation (A.1269), Equation (A.1253) becomes

$$-\frac{k^4 A\dot{x}}{\sigma} S_1 = k^4 A\, x\, S_1 - \pi a B\, y\, S_1\,, \tag{A.1274}$$

because $\partial_{\tilde{x}} C_1 = -\pi a S_1$ and the S_2 term of θ is independent of $\tilde{x}$. Dividing by $-k^4 A/\sigma$,

$$\dot{x} = \sigma\left(\frac{\pi a B}{k^4 A}\, y - x\right)\ , \tag{A.1275}$$

which is the first of Equation (A.1270) provided

$$\pi a B = k^4 A\,. \tag{A.1276}$$

Step 2: the Jacobian in the temperature equation, in full. This is the step that decides whether three modes close, and it cannot be seen without doing it. From Equation (A.1272),

$$\begin{aligned} \partial_{\tilde{x}}\psi &= \pi a A\, x\, C_1\,, \qquad \partial_z\psi = \pi A\, x\, \sin(\pi a\tilde{x})\cos(\pi z)\ , \\ \partial_{\tilde{x}}\theta &= -\pi a B\, y\, S_1\,, \quad \partial_z\theta = \pi B\, y \cos(\pi a\tilde{x})\cos(\pi z) - 2\pi C\, z_* \cos(2\pi z)\ . \end{aligned} \tag{A.1277}$$

Assembling $J(\psi,\theta) = \partial_{\tilde{x}}\psi\,\partial_z\theta - \partial_z\psi\,\partial_{\tilde{x}}\theta$ gives three products. The two involving y combine, since

$$\pi^2 a A B\, xy\left[\cos^2(\pi a\tilde{x}) + \sin^2(\pi a\tilde{x})\right]\sin(\pi z)\cos(\pi z) = \frac{\pi^2 a A B}{2}\, xy\, S_2\,, \tag{A.1278}$$

using $\sin\phi\cos\phi = \frac{1}{2}\sin 2\phi$: the horizontal dependence cancels completely, and what is left is a mode with *twice* the vertical wavenumber and no horizontal structure. That is the physical origin of the third mode — a roll carrying warm fluid up on one side and cool fluid down on the other distorts the horizontally averaged temperature profile — and it is why z_* must be kept. The remaining product is

$$-2\pi^2 a A C\, x z_* \cos(\pi a\tilde{x})\sin(\pi z)\cos(2\pi z) = -\pi^2 a A C\, x z_* \cos(\pi a\tilde{x})\left[\sin(3\pi z) - \sin(\pi z)\right]\ , \tag{A.1279}$$

by $\sin\phi\cos 2\phi = \frac{1}{2}\left[\sin 3\phi - \sin\phi\right]$. Its $\sin(\pi z)$ half is $+\pi^2 a A C\, x z_* C_1$, a term in the retained set; its $\sin(3\pi z)$ half is not, and it is the *only* term produced anywhere in the calculation that leaves the three-mode set. Discarding it is the entire content of the truncation.

Step 3: projection. Insert Equation (A.1278), the retained half of Equation (A.1279), and $\partial_{\tilde{x}}\psi = \pi a A\, x\, C_1$ into Equation (A.1254), use $\nabla^2 C_1 = -k^2 C_1$ and $\nabla^2 S_2 = -4\pi^2 S_2$, and collect the coefficients of the two orthogonal modes. The C_1 component gives

$$k^2 B\dot{y} + \pi^2 a A C\, x z_* = \mathrm{Ra}\,\pi a A\, x - k^2 B\, y\,, \tag{A.1280}$$

and the S_2 component, remembering the minus sign carried by z_* in Equation (A.1272), gives

$$-k^2 C\dot{z}_* + \frac{\pi^2 a A B}{2}\, xy = 4\pi^2 C\, z_*\,. \tag{A.1281}$$

Step 4: fixing the constants. Dividing Equation (A.1280) by k^2B and Equation (A.1281) by $-k^2C$,

$$\dot{y} = \frac{\mathrm{Ra}\,\pi a A}{k^2 B}x - y - \frac{\pi^2 a A C}{k^2 B}xz_* \,, \qquad \dot{z}_* = \frac{\pi^2 a A B}{2k^2 C}xy - \frac{4\pi^2}{k^2}z_* \,. \tag{A.1282}$$

Comparison with Equation (A.1270) imposes three conditions, to be read together with Equation (A.1276):

$$\frac{\mathrm{Ra}\,\pi a A}{k^2 B} = r \,, \qquad \frac{\pi^2 a A C}{k^2 B} = 1 \,, \qquad \frac{\pi^2 a A B}{2k^2 C} = 1 \,. \tag{A.1283}$$

The last two give $C = k^2 B/\left(\pi^2 a A\right)$ and $C = \pi^2 a A B/\left(2k^2\right)$; equating them yields $\pi^4 a^2 A^2 = 2k^4$, hence

$$A = \frac{\sqrt{2}\,k^2}{\pi^2 a} = \frac{\sqrt{2}\,(1+a^2)}{a} \,, \tag{A.1284}$$

by Equation (A.1263), which is the coefficient in Equation (A.1267). Then Equation (A.1276) gives

$$B = \frac{k^4 A}{\pi a} = \frac{\sqrt{2}\,k^6}{\pi^3 a^2} = \frac{\sqrt{2}}{\pi}\,\frac{k^6}{\pi^2 a^2} = \frac{\sqrt{2}\,\mathrm{Ra_c}(a)}{\pi} \,, \tag{A.1285}$$

the last equality being Equation (A.1265), and $C = k^2 B/(\pi^2 a A) = B/\sqrt{2} = \mathrm{Ra_c}(a)/\pi$ after substituting Equation (A.1284). With $\theta = By\,C_1 - Cz_* S_2$ this is exactly Equation (A.1268). Finally the first of Equation (A.1283) reads, using Equation (A.1276),

$$r = \frac{\mathrm{Ra}\,\pi a A}{k^2 B} = \mathrm{Ra}\,\frac{\pi^2 a^2}{k^6} = \frac{\mathrm{Ra}}{\mathrm{Ra_c}(a)} \,, \tag{A.1286}$$

and the last term of Equation (A.1282) gives

$$b = \frac{4\pi^2}{k^2} = \frac{4\pi^2}{\pi^2\,(1+a^2)} = \frac{4}{1+a^2} \,, \tag{A.1287}$$

which completes Equation (A.1271). At $a^2 = \frac{1}{2}$, $b = 4/(3/2) = 8/3$. □

A.65.6 Reading the parameters back

Remark A.770 (What the three variables and three parameters are). Undoing the scaling Equation (A.1252) makes every symbol of Equation (32.36) a statement about the layer. *The variables.* x is the amplitude of the convective roll: the physical stream function is $\kappa\,A\,x\,S_1$, so the fluid speed is of order $(\kappa/d)\,(1+a^2)\,|x|$ — $2\times10^{-5}\,\mathrm{m^2/s}$ divided by a centimetre, of order 1 mm/s for air at $|x|$ of order unity. y measures the horizontal temperature variation at the roll's own wavenumber, and z_* the horizontally averaged distortion of the vertical temperature profile. Their common temperature scale is not arbitrary: substituting Equation (A.1285) into Equation (A.1252) gives

$$\frac{\nu\kappa}{g\alpha_T d^3}\cdot\frac{\mathrm{Ra_c}(a)}{\pi} = \frac{\Delta T_\mathrm{c}}{\pi} \,, \tag{A.1288}$$

where ΔT_c is the temperature difference at which that wavenumber first becomes unstable — so y and z_* are temperature departures measured in units of the critical driving difference divided by π. *The parameters.* $\sigma = \nu/\kappa$ is the Prandtl number of the working fluid alone: about 0.71 for air, 7 for water at 20 °C, of order 0.02 for liquid sodium and 10^2 or more for oils. Lorenz's $\sigma = 10$ is therefore a plausible fluid and not a dial setting. $r = \mathrm{Ra}/\mathrm{Ra_c}$ is the imposed temperature difference in units of the one that starts convection, so $r = 28$ means a layer driven 28 times past onset. $b = 4/(1+a^2)$ is a pure function of the horizontal wavenumber of the roll and of nothing else at all.

Rests on Theorem A.769 and Remark 32.2.

Remark A.771 (What was assumed, and where it fails). Five approximations were made, and they are of quite different standing. *One* is the Boussinesq approximation (Remark A.762), good to a few per cent in ordinary laboratory convection. *Two* is the restriction to rolls (Remark A.764), which is a fact about the observed flow only near onset. *Three* is stress-free boundaries, chosen because they make $\sin \pi z$ an exact eigenfunction; the laboratory case of two rigid plates gives $\mathrm{Ra_c} \approx 1708$ instead of 657.5 (Phenomenon 31.78) and no closed-form modes at all, so the tidy arithmetic above is bought with a boundary condition no experiment has. *Four* is the fixing of a single horizontal wavenumber a, which suppresses every interaction between cells of different sizes. *Five*, and the serious one, is the discarding of the $\sin(3\pi z)$ term at Equation (A.1279). That discarded term is small only while the convection is weak. At Lorenz's $r = 28$ it is not small, the neglected modes are not small, and Lorenz said so plainly in the paper that introduced the system [1262]. The projection is *exact* on the linear terms — which is why Theorem A.767 is a genuine result about fluids — and an approximation on the nonlinear ones, which is why Equation (32.36) beyond onset is a model of convection and not convection. Remark 32.63 makes the same point in the chapter, and the two statements are meant to stand together: what Equation (32.36) is good for is the study of a low-dimensional dissipative flow that happens to have been derived from a fluid, not the prediction of a heat flux.

Rests on Theorem A.769 and Remark 32.63.

Remark A.772. Appendix A.65 discharges the derivation owed at Definition 32.62 of Chapter 32. The four stages announced there are Appendix A.65.2 (the stream function, which eliminates the pressure), Appendix A.65.3 (the explicit scaling and the two dimensionless groups Equation (32.38)), Appendix A.65.4 (the linear stability calculation giving $\mathrm{Ra_c} = 27\pi^4/4$ at $a^2 = \frac{1}{2}$ for stress-free boundaries [438]), and Appendix A.65.5 (the Galerkin truncation producing Equation (32.36) with the parameters Equation (32.37)). Everything the chapter builds on Equation (32.36) — the volume contraction, the fixed points, the strange attractor — is derived there and does not depend on this section; what this section supplies is the licence to call σ, r and b measurements rather than dials, which is the demand Remark 32.2 makes of every model in that chapter.

A.66 Rayleigh Acoustic Streaming Above a Vibrating Surface

This appendix supplies the derivation owed at Phenomenon 35.8 of Chapter 35: the steady second-order flow that an oscillating boundary drives in the fluid above it, and the direction of its outer branch, which is what carries a light powder to the antinodes of a sounding plate where sand collects at the nodes [39, 88].

The scope is fixed in advance by Remark 35.9 of the chapter and is not widened here. What is derived is the streaming velocity, its coefficient, and the sense of the circulation. What is *not* derived is a threshold grain radius separating sand from lycopodium: that comparison turns on the grain's adhesion to the plate, which this treatise does not model. The reader should hold the chapter's bound in view throughout, because the calculation below is exact enough to invite being pushed further than it can go.

A.66.1 What is quoted here

No theorem. The whole argument is an expansion of the incompressible Navier–Stokes equations Equation (31.32) of Chapter 31 in the amplitude of the motion, and every step — the linear boundary-layer solution, the time averages, and five elementary integrals — is carried out here.

Three things are *assumed* rather than proved, and all three are statements about the physical regime rather than about mathematics. The fluid is treated as incompressible

within a plate wavelength, which requires the acoustic wavelength in air to be much the longer of the two (Lemma A.777 states the condition and Example A.779 checks it). The motion is assumed to have settled into a state that is steady in the mean, so that $\langle \partial_t \boldsymbol{v}_2 \rangle = 0$. And the expansion is assumed to be asymptotic in the amplitude, which is the usual and unproved status of such expansions in fluid mechanics.

The result derived below is Rayleigh's, published in 1884 in the *Philosophical Transactions* under the title "On the circulation of air observed in Kundt's tubes, and on some allied acoustical problems". That paper has no key in `references.bib` — `Rayleigh:1883` there is the paper on the equilibrium of a heavy fluid of variable density, a different work, and citing it here would be exactly the misattribution the editorial rules warn against — so the attribution is made in words, on the footing of Darboux's memoir in Remark A.81. Nothing below rests on it.

A.66.2 Why the effect cannot appear before second order

Write $\langle \cdot \rangle$ for the average over one period $2\pi/\omega$ of the driving, and expand every field in the dimensionless amplitude ϵ of the boundary motion:

$$\boldsymbol{v} = \epsilon\, \boldsymbol{v}_1 + \epsilon^2 \boldsymbol{v}_2 + \cdots , \qquad p = p_0 + \epsilon\, p_1 + \epsilon^2 p_2 + \cdots . \tag{A.1289}$$

At first order the equations are linear with coefficients independent of t and the forcing is $\propto \cos \omega t$, so every first-order field is a harmonic function of time of the same frequency and

$$\langle \boldsymbol{v}_1 \rangle = 0 . \tag{A.1290}$$

A tracer carried by the fluid therefore returns, at first order, to where it started after each period: there is no steady transport whatever. Steady transport first appears at order ϵ^2, and it appears there because the average of a *product* of two oscillating quantities of the same frequency need not vanish even though each factor averages to zero. That is the entire physical content of the phenomenon, and it is why Phenomenon 35.8 cannot be obtained from any first-order acoustic quantity, however cleverly combined.

A.66.3 The first-order Stokes layer

Proposition A.773 (The oscillating boundary layer)**.** *Let a plane wall occupy $y = 0$ with fluid in $y > 0$, and let the outer acoustic field impose, just outside the layer, the tangential velocity $u_\infty = U(x) \cos \omega t$, with U varying on a length scale much greater than the layer thickness. Then the first-order velocity satisfying no slip at the wall is*

$$u_1 = U(x) \left[\cos \omega t - \mathrm{e}^{-\eta} \cos\left(\omega t - \eta\right) \right] , \qquad \eta := \frac{y}{\delta} , \tag{A.1291}$$

with the Stokes-layer thickness

$$\delta := \sqrt{\frac{2\nu}{\omega}} , \tag{A.1292}$$

in m*, and the wall-normal component follows from continuity as*

$$v_1 = -\frac{\mathrm{d}U}{\mathrm{d}x}\, \delta\, \Re\left[H(\eta)\, \mathrm{e}^{\mathrm{i}\omega t} \right] , \qquad H(\eta) := \eta - \frac{h(\eta)}{1+\mathrm{i}} , \tag{A.1293}$$

where $h(\eta) := 1 - E(\eta)$ and $E(\eta) := \mathrm{e}^{-(1+\mathrm{i})\eta}$, so that $u_1 = U\, \Re\left[h\, \mathrm{e}^{\mathrm{i}\omega t} \right]$.
Rests on Equations (31.32) *and* (A.1289).

Proof. *Derives Proposition A.773.*

At first order Equation (31.32) linearizes to $\partial_t u_1 = -\rho^{-1} \partial_x p_1 + \nu\, \partial_y^2 u_1$, the convective terms being of second order and the streamwise viscous term smaller than the transverse

one by the square of the ratio of the layer thickness to the scale of U. The outer field carries no shear, so it satisfies $\partial_t u_\infty = -\rho^{-1}\partial_x p_1$; and by Equation (A.1289) applied to the pressure together with the thinness of the layer, p_1 has the same value inside the layer as just outside it. Subtracting,

$$\partial_t \left(u_1 - u_\infty\right) = \nu\, \partial_y^2 \left(u_1 - u_\infty\right) , \tag{A.1294}$$

since $\partial_y^2 u_\infty = 0$. Seek $u_1 - u_\infty = \Re\left[A(x)\, \mathrm{e}^{\mathrm{i}\omega t}\mathrm{e}^{-\gamma y}\right]$: then $\mathrm{i}\omega = \nu\gamma^2$, so $\gamma = \sqrt{\mathrm{i}\omega/\nu} = (1+\mathrm{i})/\delta$ with δ as in Equation (A.1292), the root with positive real part being the one that decays. No slip at $y = 0$ requires $u_1 = 0$ there, so $A = -U(x)$ and

$$u_1 = U\, \Re\left[\left(1 - \mathrm{e}^{-(1+\mathrm{i})\eta}\right) \mathrm{e}^{\mathrm{i}\omega t}\right] = U\, \Re\left[h\, \mathrm{e}^{\mathrm{i}\omega t}\right] , \tag{A.1295}$$

and expanding the real part with $\mathrm{e}^{-(1+\mathrm{i})\eta}\mathrm{e}^{\mathrm{i}\omega t} = \mathrm{e}^{-\eta}\mathrm{e}^{\mathrm{i}(\omega t - \eta)}$ gives Equation (A.1291).

For v_1, integrate continuity $\partial_x u_1 + \partial_y v_1 = 0$ upward from the wall, where $v_1 = 0$:

$$v_1 = -\int_0^y \partial_x u_1 \, \mathrm{d}y' = -\frac{\mathrm{d}U}{\mathrm{d}x}\, \delta\, \Re\left[\mathrm{e}^{\mathrm{i}\omega t} \int_0^\eta h \, \mathrm{d}\eta'\right] , \tag{A.1296}$$

and

$$\int_0^\eta h \, \mathrm{d}\eta' = \int_0^\eta \left(1 - \mathrm{e}^{-(1+\mathrm{i})\eta'}\right) \mathrm{d}\eta' = \eta - \frac{1-E}{1+\mathrm{i}} = \eta - \frac{h}{1+\mathrm{i}} = H , \tag{A.1297}$$

which is Equation (A.1293). Note that v_1 carries the factor $\mathrm{d}U/\mathrm{d}x$: a spatially uniform oscillation drives no normal motion, and it will drive no streaming either. □

A.66.4 The second-order mean flow and the slip velocity

Lemma A.774 (The layer is driven by the excess Reynolds stress)**.** *Let $\langle u_2 \rangle$ be the time-averaged second-order streamwise velocity. Then, to leading order in the thinness of the layer,*

$$\mu\, \partial_y^2 \langle u_2 \rangle = F(y) , \qquad F := \rho \left[\langle u_1 \partial_x u_1 + v_1 \partial_y u_1 \rangle - \langle u_\infty \partial_x u_\infty \rangle\right] , \tag{A.1298}$$

with $\langle u_2 \rangle = 0$ at $y = 0$ and $\partial_y \langle u_2 \rangle \to 0$ as $\eta \to \infty$. Consequently

$$u_s := \lim_{\eta\to\infty} \langle u_2 \rangle = -\frac{1}{\mu} \int_0^\infty y\, F(y) \, \mathrm{d}y . \tag{A.1299}$$

Rests on Proposition A.773 and Equations (31.32) and (A.1290).

Proof. *Derives Lemma A.774.*

Take the streamwise component of Equation (31.32) at order ϵ^2 and average over a period. The unsteady term averages to zero in a state steady in the mean; the convective terms contribute $\langle u_1 \partial_x u_1 + v_1 \partial_y u_1 \rangle$, since the second-order velocity's own convection is of fourth order; and of the viscous term only $\nu\, \partial_y^2 \langle u_2 \rangle$ survives, by the same comparison of transverse with streamwise gradients used in Proposition A.773. Hence

$$\mu\, \partial_y^2 \langle u_2 \rangle = \rho \langle u_1 \partial_x u_1 + v_1 \partial_y u_1 \rangle + \partial_x \langle p_2 \rangle . \tag{A.1300}$$

Now evaluate the same equation just outside the layer, where the field is the irrotational u_∞ and there is no mean shear. There

$$0 = \rho \langle u_\infty \partial_x u_\infty \rangle + \partial_x \langle p_2 \rangle , \tag{A.1301}$$

that is, the Reynolds stress of an irrotational oscillation is balanced entirely by the second-order mean pressure and drives nothing. The layer is thin, so $\langle p_2 \rangle$ is the same function of x inside it as immediately outside; subtracting Equation (A.1301) from Equation (A.1300)

eliminates the pressure and leaves Equation (A.1298). By construction $F \to 0$ as $\eta \to \infty$, and exponentially, since $u_1 \to u_\infty$ exponentially.

For the boundary conditions: no slip holds at every order, so $\langle u_2 \rangle (0) = 0$; and the mean shear at the outer edge of the layer belongs to the outer streaming, which varies on the acoustic length scale rather than on δ and is therefore negligible here. Then integrating Equation (A.1298) downward from infinity, $\mu\, \partial_y \langle u_2 \rangle (y) = -\int_y^\infty F \,\mathrm{d}y'$, and integrating that upward from the wall,

$$\mu\, u_s = -\int_0^\infty \int_y^\infty F(y')\,\mathrm{d}y'\,\mathrm{d}y = -\int_0^\infty F(y') \int_0^{y'} \mathrm{d}y\,\mathrm{d}y' = -\int_0^\infty y' F(y')\,\mathrm{d}y' \,, \tag{A.1302}$$

the exchange of the order of integration being legitimate because F decays exponentially. That is Equation (A.1299). □

Theorem A.775 (Rayleigh's limiting velocity)**.** *With U and δ as above,*

$$u_s = -\frac{3}{4\omega}\, U \frac{\mathrm{d}U}{\mathrm{d}x} = -\frac{3}{4\omega} \frac{\mathrm{d}\,\langle u_\infty^2 \rangle}{\mathrm{d}x} \,, \tag{A.1303}$$

in m/s. *The viscosity has cancelled: u_s does not contain ν, and the thickness δ of the layer that produced it does not appear.*

Rests on Lemma A.774 and Equations (A.1291) *and* (A.1293).

Proof. *Derives Theorem A.775.*

Write $g(\eta, t) := \Re\left[h\, \mathrm{e}^{\mathrm{i}\omega t}\right]$ and $G(\eta, t) := \Re\left[H\, \mathrm{e}^{\mathrm{i}\omega t}\right]$, so that by Proposition A.773

$$u_1 = U g \,, \qquad \partial_x u_1 = U' g \,, \qquad \partial_y u_1 = \frac{U}{\delta} \partial_\eta g \,, \qquad v_1 = -U' \delta\, G \,, \tag{A.1304}$$

with $U' = \mathrm{d}U/\mathrm{d}x$. The two Reynolds-stress terms are therefore

$$\langle u_1 \partial_x u_1 \rangle = U U' \left\langle g^2 \right\rangle \,, \qquad \langle v_1 \partial_y u_1 \rangle = -U U' \left\langle G\, \partial_\eta g \right\rangle \,, \tag{A.1305}$$

the factor δ cancelling between v_1 and $\partial_y u_1$ — which is already the reason the answer will not contain the layer thickness. Since $\langle u_\infty \partial_x u_\infty \rangle = U U' \left\langle \cos^2 \omega t \right\rangle = \frac{1}{2} U U'$,

$$F = \rho\, U U'\, \Psi(\eta) \,, \qquad \Psi := \left\langle g^2 \right\rangle - \left\langle G\, \partial_\eta g \right\rangle - \tfrac{1}{2} \,. \tag{A.1306}$$

Evaluating Ψ. For two quantities written as $\Re\left[A \mathrm{e}^{\mathrm{i}\omega t}\right]$ and $\Re\left[B \mathrm{e}^{\mathrm{i}\omega t}\right]$ the period average of the product is $\frac{1}{2} \Re\left[A \bar{B}\right]$. With $\partial_\eta g = \Re\left[h' \mathrm{e}^{\mathrm{i}\omega t}\right]$ and $h' = (1 + \mathrm{i})\, E$,

$$2\Psi = |h|^2 - 1 - \Re\left[H \overline{h'}\right] \,. \tag{A.1307}$$

Now $h = 1 - E$ with $E = \mathrm{e}^{-\eta} \mathrm{e}^{-\mathrm{i}\eta}$, so

$$|h|^2 - 1 = -2\Re E + |E|^2 = -2\mathrm{e}^{-\eta} \cos\eta + \mathrm{e}^{-2\eta} \,, \tag{A.1308}$$

and, using $\overline{h'} = (1 - \mathrm{i})\, \bar{E}$ together with $(1 - \mathrm{i}) \,/\, (1 + \mathrm{i}) = -\mathrm{i}$,

$$H \overline{h'} = \left[\eta - \frac{1 - E}{1 + \mathrm{i}}\right] (1 - \mathrm{i})\, \bar{E} = \eta\, (1 - \mathrm{i})\, \bar{E} + \mathrm{i} \bar{E} - \mathrm{i}\, |E|^2 \,. \tag{A.1309}$$

Taking real parts with $\bar{E} = \mathrm{e}^{-\eta}\, (\cos\eta + \mathrm{i} \sin\eta)$,

$$\Re\left[H \overline{h'}\right] = \eta \mathrm{e}^{-\eta}\, (\cos\eta + \sin\eta) - \mathrm{e}^{-\eta} \sin\eta \,, \tag{A.1310}$$

the last term of Equation (A.1309) being purely imaginary. Assembling Equations (A.1308) and (A.1310),

$$2\Psi = -2\mathrm{e}^{-\eta}\cos\eta + \mathrm{e}^{-2\eta} - \eta\mathrm{e}^{-\eta}\left(\cos\eta + \sin\eta\right) + \mathrm{e}^{-\eta}\sin\eta\,. \tag{A.1311}$$

The moment. By Equations (A.1299) and (A.1306), with $y = \delta\eta$,

$$u_s = -\frac{\rho U U'\delta^2}{\mu}\int_0^\infty \eta\,\Psi\,\mathrm{d}\eta = -\frac{2UU'}{\omega}\int_0^\infty \eta\,\Psi\,\mathrm{d}\eta\,, \tag{A.1312}$$

since $\rho\delta^2/\mu = \delta^2/\nu = 2/\omega$ by Equation (A.1292). The required integrals follow from

$$\int_0^\infty \eta^n \mathrm{e}^{-\eta}\mathrm{e}^{\mathrm{i}\eta}\,\mathrm{d}\eta = \frac{n!}{(1-\mathrm{i})^{n+1}}\,, \tag{A.1313}$$

with $(1-\mathrm{i})^2 = -2\mathrm{i}$ and $(1-\mathrm{i})^3 = -2-2\mathrm{i}$. For $n=1$ the right-hand side is $1/(-2\mathrm{i}) = \mathrm{i}/2$, so

$$\int_0^\infty \eta\mathrm{e}^{-\eta}\cos\eta\,\mathrm{d}\eta = 0\,, \qquad \int_0^\infty \eta\mathrm{e}^{-\eta}\sin\eta\,\mathrm{d}\eta = \tfrac{1}{2}\,, \tag{A.1314}$$

and for $n=2$ it is $2/(-2-2\mathrm{i}) = -1/(1+\mathrm{i}) = (-1+\mathrm{i})/2$, so

$$\int_0^\infty \eta^2\mathrm{e}^{-\eta}\cos\eta\,\mathrm{d}\eta = -\tfrac{1}{2}\,, \qquad \int_0^\infty \eta^2\mathrm{e}^{-\eta}\sin\eta\,\mathrm{d}\eta = \tfrac{1}{2}\,, \tag{A.1315}$$

while $\int_0^\infty \eta\,\mathrm{e}^{-2\eta}\mathrm{d}\eta = \frac{1}{4}$. Multiplying Equation (A.1311) by η and integrating term by term,

$$2\int_0^\infty \eta\,\Psi\,\mathrm{d}\eta = -2\cdot 0 + \tfrac{1}{4} - \left(-\tfrac{1}{2}+\tfrac{1}{2}\right) + \tfrac{1}{2} = \tfrac{3}{4}\,, \tag{A.1316}$$

so $\int_0^\infty \eta\Psi\,\mathrm{d}\eta = \frac{3}{8}$ and Equation (A.1312) gives $u_s = -(2/\omega)(3/8)UU'$, which is the first member of Equation (A.1303). The second member follows from $\langle u_\infty^2\rangle = \frac{1}{2}U^2$, whose x derivative is UU'. □

Proposition A.776 (The inner counter-flow)**.** *The mean flow at the wall itself runs the* other *way. Its shear there is*

$$\left(\partial_y\langle u_2\rangle\right)_{y=0} = \frac{1}{2\omega\delta}\,U\frac{\mathrm{d}U}{\mathrm{d}x}\,, \tag{A.1317}$$

which has the sign of $U\,\mathrm{d}U/\mathrm{d}x$*, opposite to that of* u_s *in Equation* (A.1303)*. The mean profile therefore reverses once within the layer, at a height of order* δ*.*

Rests on Lemma A.774, Equation (A.1311)*, and Theorem A.775.*

Proof. *Derives Proposition A.776.*

From the proof of Lemma A.774, $\mu\left(\partial_y\langle u_2\rangle\right)_{y=0} = -\int_0^\infty F\,\mathrm{d}y = -\rho UU'\delta\int_0^\infty \Psi\,\mathrm{d}\eta$. The zeroth moments needed are $\int_0^\infty \mathrm{e}^{-\eta}\cos\eta\,\mathrm{d}\eta = \int_0^\infty \mathrm{e}^{-\eta}\sin\eta\,\mathrm{d}\eta = \frac{1}{2}$, from Equation (A.1313) at $n=0$, together with $\int_0^\infty \mathrm{e}^{-2\eta}\mathrm{d}\eta = \frac{1}{2}$ and Equation (A.1314). Integrating Equation (A.1311),

$$2\int_0^\infty \Psi\,\mathrm{d}\eta = -2\cdot\tfrac{1}{2} + \tfrac{1}{2} - \left(0+\tfrac{1}{2}\right) + \tfrac{1}{2} = -\tfrac{1}{2}\,, \tag{A.1318}$$

so $\int_0^\infty \Psi\,\mathrm{d}\eta = -\frac{1}{4}$ and $\mu\left(\partial_y\langle u_2\rangle\right)_{y=0} = \frac{1}{4}\rho\delta UU'$. Dividing by $\mu = \rho\nu$ and using $\nu = \omega\delta^2/2$ gives Equation (A.1317). Since $\langle u_2\rangle$ starts from zero at the wall with a shear of one sign and ends at u_s of the other, it changes sign once, at a height that the explicit profile places within a small multiple of δ. □

A.66.5 The direction of the circulation over a plate

Everything so far holds for any oscillating tangential field. To reach Phenomenon 35.8 one further step is needed, and it is the step at which the answer is decided: the relation between the tangential air velocity $U(x)$ and the plate's *normal* motion.

Lemma A.777 (The tangential field above a bending plate)**.** *Let the plate vibrate with normal velocity* $v(x,0,t) = V_0 \sin(k_p x)\cos\omega t$*, so that its nodal lines are at* $\sin(k_p x) = 0$ *and its antinodes midway between them. Suppose the bending wavelength is much shorter than the acoustic wavelength in the fluid,* $k_p \gg \omega/c$*, and much longer than the Stokes layer,* $k_p\delta \ll 1$*. Then the air motion within a bending wavelength is incompressible and irrotational, with*

$$\phi = -\frac{V_0}{k_p}\sin(k_p x)\,\mathrm{e}^{-k_p y}\cos\omega t\,, \tag{A.1319}$$

and the tangential velocity imposed on the layer is

$$U(x) = -V_0\cos(k_p x)\ . \tag{A.1320}$$

Its magnitude is therefore greatest over the plate's nodal lines *and vanishes over its* antinodes.
Rests on Proposition 31.28 and Equations (31.25) and (A.1292).

Proof. *Derives Lemma A.777.*

When $k_p \gg \omega/c$ the compressible wave equation for the velocity potential, $\nabla^2\phi = c^{-2}\partial_t^2\phi$, reduces within a bending wavelength to Equation (31.25), because the term on the right is smaller than $\partial_x^2\phi$ by $(\omega/ck_p)^2$. Separating with the x dependence imposed by the plate and requiring decay as $y\to\infty$ gives $\phi \propto \sin(k_p x)\,\mathrm{e}^{-k_p y}$, and $\nabla^2\phi = \left(-k_p^2 + k_p^2\right)\phi = 0$ confirms it. Fixing the constant by $\partial_y\phi = v(x,0,t)$ at $y=0$ gives Equation (A.1319), whence

$$u = \partial_x\phi = -V_0\cos(k_p x)\,\mathrm{e}^{-k_p y}\cos\omega t\,. \tag{A.1321}$$

Since $k_p\delta \ll 1$, the exponential is indistinguishable from unity across the Stokes layer, so the field Equation (A.1321) is what Proposition A.773 sees as u_∞, and Equation (A.1320) follows.

The physical reading is worth one sentence, because the result is the opposite of the naive guess. Over an antinode the plate pushes air straight up and down and the horizontal motion vanishes by symmetry; over a nodal line the plate does not move at all, but the air expelled from the antinode on one side must cross to the other, and it is there that the horizontal velocity is greatest. □

Corollary A.778 (The powder collects at the antinodes)**.** *With Equation (A.1320), Equation (A.1303) gives*

$$u_s = \frac{3V_0^2 k_p}{8\omega}\sin(2k_p x)\ , \tag{A.1322}$$

which is directed from each nodal line towards the antinodes on either side of it. The outer branch of the circulation therefore sweeps the fluid just above the layer from the nodes to the antinodes, rising there and returning above; a grain carried by that branch collects at the antinodes, where sand does not, which is Phenomenon 35.8.
Rests on Theorem A.775, Lemma A.777, and Phenomenon 35.5.

Proof. *Derives Corollary A.778.*

$U^2 = V_0^2\cos^2(k_p x)$, so $\mathrm{d}\left(U^2\right)/\mathrm{d}x = -V_0^2 k_p\sin(2k_p x)$ and $u_s = -(3/8\omega)\,\mathrm{d}\left(U^2\right)/\mathrm{d}x$ is Equation (A.1322). Take a nodal line at $x=0$ and the neighbouring antinode at $x=\pi/2k_p$: on the interval between them $\sin(2k_p x) > 0$, so $u_s > 0$ and the flow is directed towards the antinode. By the reflection symmetry about the nodal line the same holds on the other side, so the nodal lines are lines of divergence of the slip flow and the antinodes are lines of

convergence. Mass conservation then requires the fluid to rise above each antinode and to descend over each node, closing the circulation with an outward branch aloft.

The statement Equation (A.1303) makes about $|U|$ is general; what is special to the plate is Lemma A.777, which places the maxima of $|U|$ over the nodes. The two together are what invert the figure. □

Example A.779 (The scales, in SI)**.** Take a plate sounding at $f = 500\,\mathrm{Hz}$ with a bending wavelength $\lambda_p = 0.10\,\mathrm{m}$, so $\omega = 3.14 \times 10^3/\mathrm{s}$ and $k_p = 63/\mathrm{m}$, in air of kinematic viscosity $1.5 \times 10^{-5}\,\mathrm{m}^2/\mathrm{s}$ and sound speed $343\,\mathrm{m/s}$. Then

$$\delta = \sqrt{\frac{2 \times 1.5 \times 10^{-5}\,\mathrm{m}^2/\mathrm{s}}{3.14 \times 10^3/\mathrm{s}}} = 9.8 \times 10^{-5}\,\mathrm{m} \approx 0.1\,\mathrm{mm}\,, \tag{A.1323}$$

against an acoustic wavelength $c/f = 0.69\,\mathrm{m}$. The two hypotheses of Lemma A.777 are met with room to spare in one case and by about a factor of seven in the other: $k_p\delta = 6.3 \times 10^{-3} \ll 1$, and $k_p/(\omega/c) = 63/9.2 = 6.8 \gg 1$. The second ratio is the one to watch, since it fails at the coincidence frequency where the bending wave becomes sonic, and above it the plate radiates and the near field is no longer evanescent.

With a plate displacement amplitude $A_p = 50\,\mu\mathrm{m}$, so $V_0 = \omega A_p = 0.16\,\mathrm{m/s}$, Equation (A.1322) gives a peak slip velocity

$$\frac{3V_0^2 k_p}{8\omega} = \tfrac{3}{8}\,\omega A_p^2 k_p = 1.9 \times 10^{-4}\,\mathrm{m/s} \approx 0.2\,\mathrm{mm/s}\,, \tag{A.1324}$$

which moves a grain across a five-centimetre cell in a few minutes — slow, steady and unmistakable, which is what the observation reports. The same amplitude gives a peak plate acceleration $\omega^2 A_p = 4.9 \times 10^2\,\mathrm{m/s}^2$, some fifty times g, so sand is thrown vigorously; the two mechanisms are operating simultaneously and on different powders, exactly as the chapter says.

Rests on Equations (A.1292) *and* (A.1322).

A.66.6 The same mechanism in Kundt's tube

Corollary A.780 (Dust ridges at the displacement nodes)**.** *In a gas column carrying a standing wave of wavenumber k along a tube, the tangential velocity at the wall is $U(x) = U_0 \sin(kx)$ with its maxima at the displacement antinodes. Equation* (A.1303) *then gives*

$$u_s = -\frac{3U_0^2 k}{8\omega}\sin(2kx) = -\frac{3U_0^2}{8c}\sin(2kx)\,, \tag{A.1325}$$

directed from each displacement antinode towards the displacement nodes on either side. A powder swept along the wall by that flow therefore accumulates at the displacement nodes, spaced by half a wavelength — which is Phenomenon 35.12.

Rests on Theorem A.775 and Equations (35.17) *and* (A.1303).

Proof. *Derives Corollary A.780.*

In a standing wave the velocity amplitude is greatest at the displacement antinodes and zero at the nodes, and the wall is parallel to the axis, so the tangential field imposed on the Stokes layer is $U_0 \sin(kx)$ with U_0 the velocity amplitude. Then $\mathrm{d}(U^2)/\mathrm{d}x = U_0^2 k \sin(2kx)$ and Equation (A.1303) gives Equation (A.1325), the second form using $\omega = ck$. Between a node at $x = 0$ and the antinode at $x = \pi/2k$, $\sin(2kx) > 0$ and $u_s < 0$: the flow runs from the antinode back to the node. So the displacement nodes are lines of convergence and the ridges form there.

Phenomenon 35.12 asserts this collection and uses the ridge spacing to measure a wavelength; the chapter takes the location of the ridges from the observation. It is worth

noting that the mechanism is the same as the plate's and the *apparent* reversal between them is not in the streaming at all but in Lemma A.777: in the tube the maxima of $|U|$ coincide with the visible antinodes of the wave, whereas over a plate they sit above its nodal lines. In both cases the powder goes where $|U|$ is least. □

A.66.7 What is not derived here

Remark A.781 (Three gaps, and why they are left open). *Which branch a given grain samples.* By Proposition A.776 the mean flow reverses within the layer, so a grain lying in contact with the plate is not in the same current as one riding above it. A lycopodium spore is some 30 µm across against the $\delta \approx 0.1$ mm of Equation (A.1323): it protrudes into the layer but does not clear it, so which branch dominates the drag on it is a quantitative question about a body of comparable size to the layer, and neither Theorem A.775 nor Proposition A.776 settles it. What is derived here is the outer branch and its direction, which is precisely the bound Remark 35.9 sets.

Adhesion. The competition the chapter names — between the streaming drag and the residual contact friction — requires a model of the grain's adhesion to the plate. There is none in this treatise, so no threshold radius separating the two powders is predicted, and none should be read into Equation (A.1324).

The figure itself. Equation (A.1322) was computed for a one-dimensional mode shape. A real Chladni figure is two-dimensional, and the streaming above it is a two-dimensional pattern of cells whose stagnation structure is not obtained by writing $\sin(2k_p x)$ twice. That the powder ends up on the antinodal regions follows from the sign of Equation (A.1303) along any line crossing a nodal curve, which is all that is claimed.

Remark A.782. Appendix A.66 discharges the derivation owed at Phenomenon 35.8 of Chapter 35, in the Chladni experiment Section 35.3. The reader returning there should carry back the reason the chapter keeps the phenomenon at all. The sand figure and the lycopodium figure are traced by the same mode of the same plate at the same amplitude, and they are complementary: one records where the *plate* is still, the other where the *air* is still. The first is a threshold in $\omega^2 A$ against g, obtained in Chapter 30; the second is Equation (A.1303), a second-order effect in a fluid that the plate merely stirs. Neither figure is the eigenfunction, and an experiment that images a field through a tracer measures the tracer as well as the field — which is the standing warning Chapter 35 attaches to Phenomenon 35.8, and which this derivation makes quantitative.

A.67 The Free-Streamline Jet: Kirchhoff's Contraction Coefficient

This appendix proves Proposition 36.3 of Chapter 36: the two-dimensional efflux of an ideal incompressible liquid through a sharp-edged slit of width a in a plane wall contracts to an asymptotic jet width a_j with

$$C_c = \frac{a_j}{a} = \frac{\pi}{\pi + 2} = 0.611\,, \tag{A.1326}$$

which is Equation (36.6), and does so with no empirical input whatever [172, 175].

The chapter's paragraph after the `prooflink` explains why the problem is hard and what the hodograph does about it; that discussion is not repeated here. This section begins where it stops: it builds the hodograph region explicitly, maps it, and integrates back.

A.67.1 What is quoted here

One theorem is imported, and it is the only one.

Theorem A.783 (Riemann mapping with boundary correspondence, quoted)**.** *Let $D \subsetneq \mathbb{C}$ be a simply connected domain whose boundary on the Riemann sphere is a Jordan curve. Then there is a holomorphic bijection of D onto the upper half plane $\mathbb{H} := \{t \in \mathbb{C} : \Im t > 0\}$, it extends to a homeomorphism of the closures, and it is unique once three boundary points and their cyclic order are prescribed. Consequently two such maps that agree on three boundary points agree everywhere.*

Rests on Definition 8.5 and Theorem 8.6.

Remark A.784 (What is quoted here)*.* Chapter 8 builds holomorphy (Definition 8.5), the Cauchy–Riemann equations (Theorem 8.6), Cauchy's theorem and the residue calculus, but no theory of conformal mapping: it has no Riemann mapping theorem, no boundary-correspondence theorem and no Schwarz–Christoffel formula. Theorem A.783 is therefore imported, and it is used for exactly one purpose — to identify the parameter t reached from the hodograph plane with the parameter t reached from the potential plane. Everything else below is elementary and is verified here: the mapping properties of exp, of $\chi \mapsto \chi^2$ and of the Joukowski map are established by tracking their boundary values corner by corner, so no general mapping theorem is needed for them, and the Schwarz–Christoffel formula, which the standard treatments reach for at this point, is not used at all.

The imported theorem has no key in `references.bib` — the only Riemann and Carathéodory entries there are a paper on electrodynamics and one on the foundations of thermodynamics — so its attribution is made in words, on the footing of Darboux's memoir in Remark A.81: Riemann's inaugural dissertation of 1851 for the mapping, Carathéodory's papers of 1913 for the extension to the boundary. A reader who grants Theorem A.783 is granted everything else in this section.

Two further limitations are honest to state here rather than at the end. The construction below is a *verification*: it exhibits a flow satisfying every condition of the problem and computes its contraction. That the free-boundary problem has no other solution is not proved, and the treatise does not carry the machinery to prove it. And gravity is neglected *within* the jet, so that the speed on the free streamline is the single constant q_0; this is exact only in the limit $a \ll h$, and the chapter's measurements are made in that regime.

A.67.2 Formulation

Let the wall occupy the plane $x = 0$ with a slit $|y| < a/2$, the reservoir the region $x < 0$, and the jet issue into $x > 0$; the flow is two-dimensional, steady, incompressible and irrotational, so by Proposition 31.28 it has a harmonic velocity potential Equation (31.25) and, with the stream function Equation (31.12), a holomorphic complex potential $w(z) = \phi + \mathrm{i}\psi$ of $z = x + \mathrm{i}y$, whose derivative is the complex velocity Equation (31.57),

$$\Omega := \frac{\mathrm{d}w}{\mathrm{d}z} = v_x - \mathrm{i}v_y = q\,\mathrm{e}^{-\mathrm{i}\theta}\,, \tag{A.1327}$$

with $q \geq 0$ the speed in m/s and θ the direction of the velocity.

Three conditions define the problem.

1. *Impermeability* on the wall: ψ is constant there.
2. *Free surface*: the jet is bounded by two streamlines on which the pressure equals the atmospheric pressure, so by Theorem 31.24 and Equation (36.2) the speed on them is the constant
$$q_0 = \sqrt{2gh}\,, \tag{A.1328}$$
h being the head in m and g the acceleration of free fall.
3. *Position of the free surface*: unknown. It is a streamline whose location is part of the answer.

The third condition is what makes this a free-boundary problem, and it is the reason no direct solution of Equation (31.25) is available: the domain on which the equation is to be solved is not given.

The configuration is symmetric about $y = 0$, so it suffices to treat the upper half. Its boundary consists of three arcs, on each of which the stream function is constant:

- the axis $y = 0$, $-\infty < x < +\infty$, on which $\psi = 0$;
- the wall $x = 0$, $y \geq a/2$, on which $\psi = m$;
- the free streamline, leaving the edge $E = (0, a/2)$ and running downstream to $x = +\infty$, $y = a_j/2$, on which also $\psi = m$.

Here m is half the volume flux per unit span; evaluating it far downstream, where the jet is uniform at speed q_0 and width a_j,

$$m = \tfrac{1}{2} q_0 a_j \,, \tag{A.1329}$$

in m^2/s. The image of the flow region in the w plane is therefore the strip $0 < \psi < m$: the reservoir at infinity is the end $\phi \to -\infty$, the jet at infinity the end $\phi \to +\infty$, and ϕ is monotone along every streamline.

A.67.3 The hodograph region is a half-strip

Lemma A.785 (The hodograph image)**.** *Put*

$$\zeta := \log \frac{q_0}{\Omega} = \log \frac{q_0}{q} + \mathrm{i}\theta =: \sigma + \mathrm{i}\theta \,. \tag{A.1330}$$

Then the flow region maps into the semi-infinite strip

$$S := \left\{ \sigma + \mathrm{i}\theta \;:\; \sigma > 0 \,,\; -\tfrac{\pi}{2} < \theta < 0 \right\} , \tag{A.1331}$$

with the three boundary arcs going to the three sides as follows: the free streamline to $\sigma = 0$, the wall to $\theta = -\pi/2$, the axis to $\theta = 0$, and the reservoir at infinity to $\sigma = +\infty$.

Rests on Equations (A.1327) *and* (A.1328) *and Definition 8.3.*

Proof. *Derives Lemma A.785.*

Take the boundary arcs in turn. On the free streamline $q = q_0$ by Equation (A.1328), so $\sigma = 0$ exactly; the direction turns from $\theta = -\pi/2$ at the edge, where the fluid is still moving down the inner face of the wall, to $\theta = 0$ far downstream, where the jet is parallel to the axis. On the wall $x = 0$, $y > a/2$, the velocity is tangent to the wall and directed towards the slit, that is in the $-y$ direction, so $\theta = -\pi/2$ throughout, while the speed rises from 0 deep in the reservoir to q_0 at the edge — so σ falls from $+\infty$ to 0. On the axis $\theta = 0$ by symmetry, and the speed rises from 0 far upstream to q_0 far downstream, so σ falls from $+\infty$ to 0.

Everywhere in the flow $0 < q < q_0$, since by Theorem 31.24 the speed is greatest where the pressure is least and the least pressure in the field is the atmospheric pressure on the free surface; and $-\pi/2 < \theta < 0$, since the flow turns monotonically from the wall direction to the axis direction between the two. Hence the image lies in S, and the three sides of S carry the three boundary arcs as listed. The corner $\zeta = 0$ is the edge E, and the corner $\zeta = -\mathrm{i}\pi/2$ at $\sigma = 0$ does not occur: the wall and the free streamline meet only at E. □

This is the step the chapter's prose promises, and it is worth naming what it has achieved. The unknown of the problem was the shape of the free surface. In the ζ plane that surface is the segment $\sigma = 0$, $-\pi/2 \leq \theta \leq 0$ — a known set, fixed before anything has been solved. The price is that the map from the physical plane to the hodograph plane is itself unknown; but that map is what the potential will supply.

A.67.4 Two maps onto the same half plane

Lemma A.786 (From the half-strip to the half plane)**.** *Define*

$$\chi = \mathrm{e}^{-\zeta} = \frac{\Omega}{q_0}\,, \qquad s = \chi^2\,, \qquad t = -\frac{1}{2}\left(s + \frac{1}{s}\right). \tag{A.1332}$$

Then $\zeta \mapsto t$ carries S holomorphically and bijectively onto $\mathbb{H}$, with the boundary correspondence

$$\textit{axis} \mapsto t \le -1\,, \qquad \textit{free streamline} \mapsto -1 \le t \le 1\,, \qquad \textit{wall} \mapsto t \ge 1\,, \tag{A.1333}$$

the edge E going to $t = 1$, the jet at infinity to $t = -1$ and the reservoir at infinity to $t = \infty$. On the free streamline the parametrization is

$$\theta = -\alpha\,, \qquad t = -\cos 2\alpha\,, \qquad 0 \le \alpha \le \frac{\pi}{2}\,. \tag{A.1334}$$

Rests on Lemma A.785, Proposition 8.4, and Definition 8.3.

Proof. *Derives Lemma A.786.*

Each map is holomorphic where used, and each is checked on the boundary.

The exponential. $\chi = \mathrm{e}^{-\zeta} = \mathrm{e}^{-\sigma}\mathrm{e}^{-\mathrm{i}\theta}$ has $|\chi| = \mathrm{e}^{-\sigma} \le 1$ and $\arg\chi = -\theta \in [0, \pi/2]$, and $\zeta \mapsto \chi$ is injective on S because the imaginary part of ζ spans an interval of length $\pi/2 < 2\pi$. So χ ranges over the open quarter disc $\{0 < |\chi| < 1,\ 0 < \arg\chi < \pi/2\}$, and the three sides of S go to the two radii and the arc.

The square. On the quarter disc $\chi \mapsto s = \chi^2$ is injective, since $\arg\chi$ spans an interval of length $\pi/2$ and doubling it spans $\pi < 2\pi$. Its image is the open half disc $\{0 < |s| < 1,\ 0 < \arg s < \pi\}$; the arc $|\chi| = 1$ goes to the arc $|s| = 1$, the radius $\arg\chi = 0$ to $s \in (0, 1)$ and the radius $\arg\chi = \pi/2$ to $s \in (-1, 0)$.

The Joukowski map. On the half disc, write $s = \varrho\mathrm{e}^{\mathrm{i}\vartheta}$ and

$$t = -\frac{1}{2}\left(s + \frac{1}{s}\right) = -\frac{1}{2}\left(\varrho + \frac{1}{\varrho}\right)\cos\vartheta - \frac{\mathrm{i}}{2}\left(\varrho - \frac{1}{\varrho}\right)\sin\vartheta\,. \tag{A.1335}$$

For $0 < \varrho < 1$ the bracket $\varrho - 1/\varrho$ is negative and $\sin\vartheta > 0$, so $\Im t > 0$: the half disc goes into $\mathbb{H}$. Injectivity is the observation that s and $1/s$ are the two preimages of a given t and that exactly one of them has modulus less than one. Surjectivity onto $\mathbb{H}$ follows because for each $t \in \mathbb{H}$ the quadratic $s^2 + 2ts + 1 = 0$ has roots with product 1 and neither on the unit circle, since a root of modulus one would make t real by Equation (A.1335).

On the boundary: $|s| = 1$, $s = \mathrm{e}^{\mathrm{i}\vartheta}$, gives $t = -\cos\vartheta \in [-1, 1]$; $s \in (0, 1)$ gives $t = -\frac{1}{2}(s + 1/s) \le -1$; and $s \in (-1, 0)$ gives $t \ge 1$. Tracing back through the two earlier maps, $s \in (0, 1)$ is the axis and $s \in (-1, 0)$ is the wall, which is Equation (A.1333). On the free streamline $\Omega = q_0\mathrm{e}^{-\mathrm{i}\theta}$, so $\chi = \mathrm{e}^{\mathrm{i}\alpha}$ with $\alpha = -\theta$, $s = \mathrm{e}^{2\mathrm{i}\alpha}$ and $t = -\cos 2\alpha$, which is Equation (A.1334). Its endpoints are $t = 1$ at $\alpha = \pi/2$ (the edge) and $t = -1$ at $\alpha = 0$ (downstream). □

Lemma A.787 (From the potential strip to the half plane)**.** *Fix the origin of the velocity potential at the edge E and define*

$$t = -1 - 2\,\mathrm{e}^{-\pi w/m}\,. \tag{A.1336}$$

Then $w \mapsto t$ carries the strip $0 < \psi < m$ holomorphically and bijectively onto $\mathbb{H}$, with the same boundary correspondence Equation (A.1333), and

$$\mathrm{d}w = -\frac{m}{\pi}\,\frac{\mathrm{d}t}{t + 1}\,. \tag{A.1337}$$

Rests on Equation (A.1329), Definition 8.3, and Lemma A.786.

Proof. *Derives Lemma A.787.*

For $0 < \psi < m$ the argument of $\mathrm{e}^{-\pi w/m}$ is $-\pi\psi/m \in (-\pi, 0)$, so the argument of $-2\mathrm{e}^{-\pi w/m}$ lies in $(0, \pi)$ and $\Im t > 0$; the map is injective because the imaginary part of $-\pi w/m$ spans an interval of length $\pi < 2\pi$, and it is onto $\mathbb{H}$ because $\mathrm{e}^{-\pi w/m}$ covers the lower half plane.

On $\psi = 0$, $t = -1 - 2\mathrm{e}^{-\pi\phi/m}$ decreases through $(-\infty, -1)$ as ϕ runs from $-\infty$ to $+\infty$: the reservoir goes to $t = \infty$ and the jet at infinity to $t = -1$. On $\psi = m$, $\mathrm{e}^{-\pi(\phi+\mathrm{i}m)/m} = -\mathrm{e}^{-\pi\phi/m}$, so $t = -1 + 2\mathrm{e}^{-\pi\phi/m}$, which is $+\infty$ deep in the reservoir, equals 1 at $\phi = 0$, and tends to -1 far downstream. The choice of origin therefore puts the edge at $t = 1$, matching Lemma A.786 arc for arc.

Differentiating $t + 1 = -2\mathrm{e}^{-\pi w/m}$ gives $\mathrm{d}t = -(\pi/m)(t+1)\,\mathrm{d}w$, which is Equation (A.1337). □

Corollary A.788 (The two parameters are the same parameter)**.** *Lemmas A.786 and A.787 are two conformal maps of the flow region onto $\mathbb{H}$ agreeing at the three boundary points E, the jet at infinity, and the reservoir at infinity. By Theorem A.783 they are the same map. Hence Ω and w are both explicit functions of one parameter t, and*

$$\mathrm{d}z = \frac{\mathrm{d}w}{\Omega} = -\frac{m}{\pi}\,\frac{1}{\Omega(t)}\,\frac{\mathrm{d}t}{t+1}\,. \tag{A.1338}$$

Rests on Theorem A.783 and Lemmas A.786 and A.787.

This is the whole solution. The physical plane has been eliminated in favour of t, and Equation (A.1338) recovers it by quadrature.

A.67.5 Integrating back along the free streamline

Theorem A.789 (Kirchhoff's contraction coefficient)**.** *The free streamline leaves the edge of the slit at height $a/2$ and approaches the height $a_j/2$ downstream, with*

$$\frac{a_j}{2} - \frac{a}{2} = -\frac{a_j}{\pi}\,, \tag{A.1339}$$

so that $a = a_j(\pi + 2)/\pi$ and Equation (A.1326) holds.
Rests on Corollary A.788 and Equations (A.1329) and (A.1334).

Proof. *Derives Theorem A.789.*

Parametrize the free streamline by α as in Equation (A.1334), so that $\alpha = \pi/2$ at the edge and $\alpha = 0$ downstream. Then

$$t + 1 = 1 - \cos 2\alpha = 2\sin^2\alpha\,, \qquad \mathrm{d}t = 2\sin 2\alpha\,\mathrm{d}\alpha\,, \tag{A.1340}$$

and Equation (A.1337) gives

$$\mathrm{d}w = -\frac{m}{\pi}\,\frac{2\sin 2\alpha\,\mathrm{d}\alpha}{2\sin^2\alpha} = -\frac{2m}{\pi}\cot\alpha\,\mathrm{d}\alpha\,, \tag{A.1341}$$

using $\sin 2\alpha = 2\sin\alpha\cos\alpha$. The sign is right: along the streamline α decreases and $\cot\alpha > 0$, so $\mathrm{d}\phi > 0$ and the potential increases downstream.

On the free streamline $\Omega = q_0\mathrm{e}^{-\mathrm{i}\theta} = q_0\mathrm{e}^{\mathrm{i}\alpha}$, so

$$\mathrm{d}z = \frac{\mathrm{d}w}{\Omega} = -\frac{2m}{\pi q_0}\cot\alpha\,\mathrm{e}^{-\mathrm{i}\alpha}\,\mathrm{d}\alpha\,, \tag{A.1342}$$

whose imaginary part is

$$\mathrm{d}y = -\frac{2m}{\pi q_0}\cot\alpha\,(-\sin\alpha)\,\mathrm{d}\alpha = \frac{2m}{\pi q_0}\cos\alpha\,\mathrm{d}\alpha\,. \tag{A.1343}$$

Integrating from the edge to the far downstream station,

$$\Delta y = \frac{2m}{\pi q_0}\int_{\pi/2}^{0}\cos\alpha\,\mathrm{d}\alpha = \frac{2m}{\pi q_0}\Big[\sin\alpha\Big]_{\pi/2}^{0} = -\frac{2m}{\pi q_0}\,, \tag{A.1344}$$

and Equation (A.1329) turns $2m/q_0$ into a_j, giving Equation (A.1339). Since the streamline starts at $y = a/2$ and ends at $y = a_j/2$,

$$\frac{a}{2} = \frac{a_j}{2} + \frac{a_j}{\pi} = a_j\,\frac{\pi+2}{2\pi} \quad\Longrightarrow\quad \frac{a_j}{a} = \frac{\pi}{\pi+2}\,, \tag{A.1345}$$

which is Equation (A.1326). Numerically $\pi/(\pi+2) = 0.611\,01$.

Note what has and has not entered. The head h appears only through q_0, and q_0 cancels between Equation (A.1329) and Equation (A.1344); so does the density, which never appeared at all. The result is a pure number, as Proposition 36.3 claims, and it is the same number at every head. □

Remark A.790 (Where the two constants come from). The π and the 2 of Equation (A.1326) have different origins, and the calculation is worth reading once with that in mind. The π is the width of the potential strip: it enters through the exponential map Equation (A.1336), whose exponent is $\pi w/m$ precisely because the strip has height m and must be opened into a half plane. It is therefore a statement about the *flux*. The 2 is $\int_0^{\pi/2}\cos\alpha\,\mathrm{d}\alpha = 1$ counted for the two halves of the jet — a statement about the *turning of the velocity through a right angle* at the edge. Nothing else survives. (No arctangent occurs anywhere in the evaluation, and no Schwarz–Christoffel integral is needed: the one quadrature to be done is Equation (A.1344), and it is elementary.)

Remark A.791 (Two checks on the solution). *The edge.* At $\alpha = \pi/2$ the speed is q_0, finite. A free-streamline solution is admissible only if the velocity is finite where the free surface leaves the solid boundary, and here that condition holds automatically, because $|\Omega| = q_0$ on the whole free streamline including its endpoint. Nothing had to be imposed to secure it, which is a feature of the sharp-edged geometry and is not true of a rounded mouthpiece.

The asymptote. The real part of Equation (A.1342) is $\mathrm{d}x = -\,(2m/\pi q_0)\,(\cos^2\alpha/\sin\alpha)\,\mathrm{d}\alpha$, whose integrand behaves as $1/\alpha$ as $\alpha \to 0$. The streamline therefore runs off to $x = +\infty$ logarithmically in α, while Equation (A.1343) shows the remaining drop in y is of order α^2: the jet approaches its asymptotic width exponentially in x, so "the" contraction is reached within a few slit widths and is a measurable quantity rather than a limit that is never attained.

A.67.6 What the theorem does not cover

Remark A.792 (The round hole is a different problem). The hodograph is a two-dimensional device. It works because the velocity of a plane irrotational flow is a holomorphic function of position, so that the map $z \mapsto \Omega$ can be treated as a conformal change of variable and an unknown boundary in one plane becomes a known boundary in another. In three dimensions there is no such structure: the velocity is a harmonic vector field, the "hodograph" is a map from a three-dimensional region to another and carries no conformal group beyond the Möbius transformations, and no closed solution of the axisymmetric free-boundary problem is known. The measured contraction coefficient for a sharp-edged circular orifice is near 0.61 — close enough to Equation (A.1326) that the coincidence is often quoted as though the theorem covered it. It does not, and Proposition 36.3 says so.

Remark A.793 (Why Borda's mouthpiece needs none of this). Proposition 36.2 obtains $C_c = 1/2$ exactly for the re-entrant mouthpiece, and it does so in half a page with no complex analysis at all. The contrast is instructive rather than embarrassing. What made

the flush slit hard was that the pressure over the wall around the orifice is unknown, so the momentum balance carries an undetermined term; the whole apparatus above exists to determine it by finding the flow. The re-entrant geometry removes the term instead: the tube is wetted on both faces, so it transmits no net force, and every other wetted surface carries a hydrostatic pressure. The unknown was arranged out of the problem rather than computed.

The two results are consistent and they bracket the observations: $\frac{1}{2}$ for the re-entrant mouthpiece, $\pi/(\pi+2)$ for the flush slit, and measurements in between for intermediate geometries. Neither is an approximation to the other.

Remark A.794. Appendix A.67 discharges the derivation owed at Proposition 36.3 of Chapter 36, in the efflux experiment Section 36.1. The reader returning there should carry back the division the chapter draws in its *Interpretation*: Bernoulli's theorem gives the speed Equation (36.2) exactly and gives the discharge not at all, because it says nothing about the area of the bundle of streamlines that reaches the jet. This section supplies that area for one geometry — and the reason it can is that the geometry is two-dimensional and sharp-edged, so the free surface has a known image in the hodograph plane. That is a narrower statement than "the ideal theory predicts the discharge", and the narrowness is the point.

A.68 Prandtl's Boundary-Layer Equations and the Blasius Similarity Solution

This appendix proves Proposition 36.15 of Chapter 36: that at large Reynolds number the Navier–Stokes system Equation (31.32) reduces, within the layer against a body, to Prandtl's equations Equation (36.26); that for a flat plate at zero incidence those equations admit a similarity solution governed by the Blasius ordinary differential equation Equation (36.27); and that the three constants quoted in Phenomenon 36.12 — the 5.0 of the thickness, the 0.664 of the local friction coefficient and the 1.328 of the plate drag — are consequences of two numbers read off the numerical solution of that equation [315, 341].

The chapter reaches Equation (36.23) by a dominant-balance estimate. This section derives what that estimate guessed: the exponent $-1/2$ is obtained rather than assumed, and the estimate's silence about the pressure is replaced by a statement of where the pressure comes from.

A.68.1 What is quoted here

Two things, of very different kinds.

The matching principle. The reduction below is a matched asymptotic expansion, and this treatise builds no theory of matched asymptotics. Section 9.6 of Chapter 9 carries the WKB approximation, which is a singular perturbation of the same family — a small parameter multiplying the highest derivative, an outer solution that cannot satisfy every boundary condition, and an inner solution on a stretched variable — and its connection formulae play the part that matching plays here. But WKB is a theorem about one linear second-order equation, not a general principle, and nothing in Part II licenses what follows. The principle is therefore stated outright and used as stated:

Postulate A.795 (Matching, quoted)**.** Let a problem carry a small parameter ε, let F_{out} be the leading term of an expansion valid away from the wall in the original variable $\hat{y}$, and let F_{in} be the leading term of an expansion valid near the wall in a stretched variable $Y=\hat{y}/\varepsilon^{p}$. Then the two expansions are assumed to possess a common domain of validity in which

$$\lim_{Y\to\infty} F_{\mathrm{in}}\left(\hat{x},Y\right)=\lim_{\hat{y}\to 0^{+}} F_{\mathrm{out}}\left(\hat{x},\hat{y}\right)\,, \tag{A.1346}$$

applied separately to each dependent variable.
Rests on Equation (31.32) *and Theorem 9.72.*

Two numbers. Equation (36.27) has no solution in closed form. The values

$$f''(0) = 0.3321\,, \qquad f'(\eta) = 0.99 \text{ at } \eta = 5.0\,, \tag{A.1347}$$

are quadratures of a numerical integration, quoted from [341] and not computed here. Every algebraic manipulation that turns them into the three constants of Phenomenon 36.12 *is* carried out below, so what is imported is exactly two decimal numbers and nothing structural.

Remark A.796 (What is quoted here). Postulate A.795 is an assumption about the solution, not a theorem: it asserts that the two expansions overlap. For the flat plate the assertion can be checked *a posteriori* — the similarity solution constructed in Appendix A.68.5 approaches its free-stream value exponentially in η, so the overlap region is genuine and wide — but that is a verification in one case, not a proof in general. The debt is Part II's, and it is named in the chapter's prose after the `prooflink` as well as here.

The two numbers of Equation (A.1347) are the only external inputs to the arithmetic. Note in particular that $f''(0)$ enters the friction coefficient linearly and the drag coefficient linearly, so the relation $1.328 = 2 \times 0.664$ between the last two constants is exact and independent of the numerics, as Appendix A.68.6 shows.

A.68.2 Scaling, and why the inviscid limit is singular

Take steady, two-dimensional, incompressible flow past a body of streamwise extent L in a stream of speed U, with gravity absorbed into the pressure. Equation (31.32) reads

$$u\,\partial_x u + v\,\partial_y u = -\frac{1}{\rho}\partial_x p + \nu\left(\partial_x^2 + \partial_y^2\right) u\,, \tag{A.1348}$$

$$u\,\partial_x v + v\,\partial_y v = -\frac{1}{\rho}\partial_y p + \nu\left(\partial_x^2 + \partial_y^2\right) v\,, \tag{A.1349}$$

$$\partial_x u + \partial_y v = 0\,. \tag{A.1350}$$

Scale on the body: $x = L\hat{x}$, $y = L\hat{y}$, $u = U\hat{u}$, $v = U\hat{v}$, $p = p_\infty + \rho U^2 \hat{p}$. Every hatted quantity is dimensionless and, by construction, of order unity in the outer flow. Dividing Equation (A.1348) by U^2/L and Equation (A.1349) likewise,

$$\hat{u}\,\partial_{\hat{x}}\hat{u} + \hat{v}\,\partial_{\hat{y}}\hat{u} = -\partial_{\hat{x}}\hat{p} + \frac{1}{\mathrm{Re}}\left(\partial_{\hat{x}}^2 + \partial_{\hat{y}}^2\right)\hat{u}\,, \tag{A.1351}$$

$$\hat{u}\,\partial_{\hat{x}}\hat{v} + \hat{v}\,\partial_{\hat{y}}\hat{v} = -\partial_{\hat{y}}\hat{p} + \frac{1}{\mathrm{Re}}\left(\partial_{\hat{x}}^2 + \partial_{\hat{y}}^2\right)\hat{v}\,, \tag{A.1352}$$

$$\partial_{\hat{x}}\hat{u} + \partial_{\hat{y}}\hat{v} = 0\,, \tag{A.1353}$$

with $\mathrm{Re} = UL/\nu$ the Reynolds number Equation (36.9). The small parameter is $\varepsilon = 1/\mathrm{Re}$, and it multiplies the highest derivative — which is the definition of a singular perturbation and the whole source of the difficulty.

Proposition A.797 (The outer limit loses a boundary condition)**.** *Setting $\varepsilon = 0$ in Equations* (A.1351) *and* (A.1352) *gives the Euler equations Equation* (31.19)*, whose solution for flow past a body is the potential flow of Proposition 31.28. That solution satisfies impermeability, $\hat{v} = 0$ on the surface, and cannot in general satisfy no slip, $\hat{u} = 0$ there. The limit $\mathrm{Re} \to \infty$ is therefore not uniform in $\hat{y}$.*
Rests on Theorem 31.22, Proposition 31.28, and Equation (A.1351)*.*

Proof. *Derives Proposition A.797.*

With $\varepsilon = 0$ the system is first order in $\hat{y}$ and admits one condition on the normal component; the potential solution of Equation (31.25) with $\partial\phi/\partial n = 0$ supplies it and leaves the tangential velocity at the wall determined — and non-zero, by Equation (31.57) and the geometry, except at stagnation points. Since the true solution has $\hat{u} = 0$ at the wall for every Re, however large, $\lim_{\mathrm{Re}\to\infty} \hat{u}$ is discontinuous at $\hat{y} = 0$ and the convergence cannot be uniform. This is d'Alembert's error stated as an analytic fact rather than a physical one, and Phenomenon 36.12 is its resolution. □

A.68.3 The inner expansion: Prandtl's equations

Theorem A.798 (Prandtl's layer equations)**.** *Introduce the stretched wall-normal coordinate and the rescaled normal velocity*

$$Y = \hat{y}\,\mathrm{Re}^{1/2}\,, \qquad \hat{v} = \mathrm{Re}^{-1/2}\,V\,, \tag{A.1354}$$

with Y and V of order unity. Then, to leading order in Re^{-1},

$$\hat{u}\,\partial_{\hat{x}}\hat{u} + V\,\partial_Y\hat{u} = -\frac{\mathrm{d}\hat{p}}{\mathrm{d}\hat{x}} + \partial_Y^2\hat{u}\,, \tag{A.1355}$$

$$\partial_Y\hat{p} = 0\,, \tag{A.1356}$$

$$\partial_{\hat{x}}\hat{u} + \partial_Y V = 0\,. \tag{A.1357}$$

Restoring dimensions, with $\delta \sim L\,\mathrm{Re}^{-1/2}$ the layer thickness of Equation (36.23), this is Equation (36.26). The exponent $\frac{1}{2}$ in Equation (A.1354) is forced: it is the only choice for which the viscous term survives the limit without dominating it.

Rests on Equations (36.23), (A.1351) and (A.1353).

Proof. *Derives Theorem A.798.*

The exponent. Try $Y = \hat{y}\,\mathrm{Re}^{p}$ with $p > 0$. In Equation (A.1351) the convective term $\hat{u}\,\partial_{\hat{x}}\hat{u}$ is of order unity, while the surviving viscous term is $\mathrm{Re}^{-1}\partial_{\hat{y}}^2\hat{u} = \mathrm{Re}^{2p-1}\partial_Y^2\hat{u}$. If $2p - 1 < 0$ the viscous term vanishes in the limit and the no-slip condition is again lost; if $2p - 1 > 0$ it dominates and the leading-order equation is $\partial_Y^2\hat{u} = 0$, whose solution is linear in Y and cannot match a bounded outer flow at $Y \to \infty$ while vanishing at the wall, except trivially. Only $p = \frac{1}{2}$ retains both, which is Equation (A.1354). Note that this *derives* $\delta/L \sim \mathrm{Re}^{-1/2}$, which Equation (36.23) obtained by comparing orders of magnitude.

The normal velocity. With Y as above, Equation (A.1353) reads $\partial_{\hat{x}}\hat{u} + \mathrm{Re}^{1/2}\partial_Y\hat{v} = 0$. For continuity to survive at leading order, $\hat{v}$ must itself be of order $\mathrm{Re}^{-1/2}$, which is the second member of Equation (A.1354), and then Equation (A.1357) follows exactly. The layer is thin *and* the flow through it is weak, in the same ratio.

Streamwise momentum. Substituting into Equation (A.1351),

$$\hat{u}\,\partial_{\hat{x}}\hat{u} + V\,\partial_Y\hat{u} = -\partial_{\hat{x}}\hat{p} + \mathrm{Re}^{-1}\partial_{\hat{x}}^2\hat{u} + \partial_Y^2\hat{u}\,, \tag{A.1358}$$

every term of which is of order unity except $\mathrm{Re}^{-1}\partial_{\hat{x}}^2\hat{u}$, which is smaller by Re^{-1} and is dropped. That is Equation (A.1355), the pressure gradient being written as a total derivative in anticipation of Equation (A.1356).

Normal momentum, and the pressure. This is the step usually waved through, and it carries the most important consequence, so it is done in full. Substituting $\hat{v} = \mathrm{Re}^{-1/2}V$ and $\partial_{\hat{y}} = \mathrm{Re}^{1/2}\partial_Y$ into Equation (A.1352), the convective terms are

$$\hat{u}\,\partial_{\hat{x}}\hat{v} + \hat{v}\,\partial_{\hat{y}}\hat{v} = \mathrm{Re}^{-1/2}\left(\hat{u}\,\partial_{\hat{x}}V + V\,\partial_Y V\right)\,, \tag{A.1359}$$

the pressure term is $-\partial_{\hat{y}}\hat{p} = -\mathrm{Re}^{1/2}\partial_Y\hat{p}$, and the viscous term is

$$\mathrm{Re}^{-1}\left(\partial_{\hat{x}}^2 + \mathrm{Re}\,\partial_Y^2\right)\mathrm{Re}^{-1/2}V = \mathrm{Re}^{-3/2}\partial_{\hat{x}}^2 V + \mathrm{Re}^{-1/2}\partial_Y^2 V\,. \tag{A.1360}$$

Multiplying the equation through by $\mathrm{Re}^{-1/2}$ collects the orders as

$$\partial_Y \hat{p} = \mathrm{Re}^{-1}\left(-\hat{u}\,\partial_{\hat{x}} V - V\,\partial_Y V + \partial_Y^2 V\right) + \mathrm{Re}^{-2}\partial_{\hat{x}}^2 V\,, \tag{A.1361}$$

so $\partial_Y \hat{p} = O\left(\mathrm{Re}^{-1}\right)$ and at leading order Equation (A.1356) holds. In words: the pressure is uniform *across* the layer. It is therefore not an unknown of the layer at all — it is whatever the outer flow imposes at the wall — and this is exactly the sentence Equation (36.26) makes when it writes $\partial_y p = 0$. □

A.68.4 Matching hands over the pressure

Corollary A.799 (The impressed pressure gradient)**.** *Let $U_e(x)$ be the tangential velocity of the outer potential solution evaluated at the surface. Then Postulate A.795 gives*

$$\lim_{Y\to\infty} \hat{u}\,(\hat{x}, Y) = \frac{U_e\,(x)}{U}\,, \tag{A.1362}$$

and, applied to the pressure,

$$-\frac{1}{\rho}\frac{\mathrm{d}p}{\mathrm{d}x} = U_e \frac{\mathrm{d}U_e}{\mathrm{d}x}\,, \tag{A.1363}$$

so the coefficient in Equation (A.1355) is a known function of x before the layer problem is posed.

Rests on Postulate A.795 and Theorems 31.24 and A.798.

Proof. *Derives Corollary A.799.*

Equation (A.1362) is Equation (A.1346) applied to $\hat{u}$: the inner solution at large Y must agree with the outer solution at small $\hat{y}$, and the latter is the wall value U_e of the potential flow, since the potential solution is smooth up to the wall. Applied to the pressure and combined with Equation (A.1356), which makes $\hat{p}$ independent of Y, it gives $\hat{p}\,(\hat{x}) = \hat{p}_{\mathrm{out}}\,(\hat{x}, 0)$. In the outer flow Theorem 31.24 holds along the wall streamline, so $p + \frac{1}{2}\rho U_e^2$ is constant there, which on differentiating is Equation (A.1363).

The logical shape is worth stating plainly, because it is what makes the boundary-layer method a method and not merely an approximation. Nothing has been solved simultaneously. The outer problem is solved first, in ignorance of the layer; its wall pressure is then handed to the inner problem as a coefficient; and the inner problem is solved afterwards. The coupling is one-way at this order, and it is Equation (A.1356) — the pressure being constant across the layer — that makes it so. □

A.68.5 The flat plate: similarity and the Blasius equation

For a flat plate aligned with a uniform stream the outer flow is undisturbed, $U_e = U$ constant, and Equation (A.1363) gives $\mathrm{d}p/\mathrm{d}x = 0$. Restoring dimensions, Equation (36.26) becomes

$$u\,\partial_x u + v\,\partial_y u = \nu\,\partial_y^2 u\,, \qquad \partial_x u + \partial_y v = 0\,, \tag{A.1364}$$

with $u = v = 0$ at $y = 0$ for $x > 0$ and $u \to U$ as $y \to \infty$.

Theorem A.800 (The similarity reduction)**.** *Put*

$$\eta = y\sqrt{\frac{U}{\nu x}}\,, \qquad \psi\,(x, y) = \sqrt{\nu x U}\, f\,(\eta)\,, \tag{A.1365}$$

with ψ the stream function Equation (31.12). Then Equation (A.1364) is satisfied identically in x if and only if f obeys the Blasius equation Equation (36.27),

$$2f''' + f f'' = 0\,, \qquad f(0) = f'(0) = 0\,, \qquad f'(\infty) = 1\,. \tag{A.1366}$$

Here η is dimensionless and ψ carries m^2/s, so f is dimensionless too.

Rests on Equations (31.12) and (A.1364) and Theorem A.798.

Proof. *Derives Theorem A.800.*

Continuity is satisfied identically by the use of ψ, with $u = \partial_y \psi$ and $v = -\partial_x \psi$. Compute the four derivatives that occur, using

$$\partial_y \eta = \sqrt{\frac{U}{\nu x}}\,, \qquad \partial_x \eta = -\frac{\eta}{2x}\,, \tag{A.1367}$$

the second because η carries $x^{-1/2}$. Then

$$u = \partial_y \psi = \sqrt{\nu x U}\, f' \sqrt{\frac{U}{\nu x}} = U f'\,, \tag{A.1368}$$

$$v = -\partial_x \psi = -\left[\tfrac{1}{2} \sqrt{\frac{\nu U}{x}} f + \sqrt{\nu x U}\, f' \left(-\frac{\eta}{2x} \right) \right] = \tfrac{1}{2} \sqrt{\frac{\nu U}{x}} \left(\eta f' - f \right)\,, \tag{A.1369}$$

$$\partial_x u = U f'' \, \partial_x \eta = -\frac{U \eta f''}{2x}\,, \tag{A.1370}$$

$$\partial_y u = U f'' \sqrt{\frac{U}{\nu x}}\,, \qquad \partial_y^2 u = \frac{U^2}{\nu x} f'''\,. \tag{A.1371}$$

Now assemble the two convective terms. The first is

$$u\, \partial_x u = U f' \left(-\frac{U \eta f''}{2x} \right) = -\frac{U^2}{2x}\, \eta f' f''\,, \tag{A.1372}$$

and the second, using $\sqrt{\nu U/x}\,\sqrt{U/\nu x} = U/x$,

$$v\, \partial_y u = \tfrac{1}{2} \sqrt{\frac{\nu U}{x}} \left(\eta f' - f \right) \cdot U f'' \sqrt{\frac{U}{\nu x}} = \frac{U^2}{2x} \left(\eta f' - f \right) f''\,. \tag{A.1373}$$

The terms in $\eta f' f''$ cancel between Equations (A.1372) and (A.1373) — this is the cancellation on which the whole reduction turns — leaving

$$u\, \partial_x u + v\, \partial_y u = -\frac{U^2}{2x}\, f f''\,. \tag{A.1374}$$

The viscous term is $\nu\, \partial_y^2 u = \nu U^2 f'''/\nu x = (U^2/x)\, f'''$ by Equation (A.1371). Equating,

$$-\frac{U^2}{2x}\, f f'' = \frac{U^2}{x}\, f'''\,, \tag{A.1375}$$

and the factor U^2/x divides out of both sides — *every* explicit occurrence of x has cancelled, which is precisely the statement that the ansatz Equation (A.1365) is a similarity solution. What remains is $f''' = -\frac{1}{2} f f''$, that is Equation (A.1366).

The boundary conditions transcribe as follows. At $y = 0$, $\eta = 0$: $u = 0$ gives $f'(0) = 0$ by Equation (A.1368), and then $v = 0$ gives $\eta f' - f = 0$ at $\eta = 0$, hence $f(0) = 0$ by Equation (A.1369). As $y \to \infty$, $u \to U$ gives $f'(\infty) = 1$. Three conditions for a third-order equation, which is why the problem is well posed and why $f''(0)$ is determined rather than free. □

A.68.6 The three constants

Proposition A.801 (Thickness, local friction, and plate drag)**.** *With* $\mathrm{Re}_x = Ux/\nu$ *and* $\mathrm{Re}_L = UL/\nu$*, the solution of Equation* (A.1366) *gives*

$$\delta_{99}(x) = 5.0\, x\, \mathrm{Re}_x^{-1/2}\,, \tag{A.1376}$$

$$c_f(x) := \frac{\tau_w(x)}{\frac{1}{2}\rho U^2} = 2 f''(0)\, \mathrm{Re}_x^{-1/2} = 0.664\, \mathrm{Re}_x^{-1/2}\,, \tag{A.1377}$$

$$C_D := \frac{F}{\frac{1}{2}\rho U^2 L} = 4 f''(0)\, \mathrm{Re}_L^{-1/2} = 1.328\, \mathrm{Re}_L^{-1/2}\,, \tag{A.1378}$$

the last for one side of a plate of length L and unit span, F being in N/m *of span.*
Rests on Equations (A.1347), (A.1366) *and* (A.1371).

Proof. *Derives Proposition A.801.*

Thickness. By Equation (A.1368), $u/U = f'(\eta)$, so the station at which the velocity has reached 99 % of the free stream is the station at which $\eta = 5.0$, by the second member of Equation (A.1347). Inverting Equation (A.1365),

$$\delta_{99} = 5.0\sqrt{\frac{\nu x}{U}} = 5.0\,x\sqrt{\frac{\nu}{Ux}} = 5.0\,x\,\mathrm{Re}_x^{-1/2}\,. \tag{A.1379}$$

Local friction. The wall stress is $\tau_w = \mu\,(\partial_y u)_{y=0}$, in Pa, and by Equation (A.1371)

$$\tau_w(x) = \mu U f''(0)\sqrt{\frac{U}{\nu x}}\,. \tag{A.1380}$$

Dividing by the dynamic pressure $\frac{1}{2}\rho U^2$ and using $\mu = \rho\nu$,

$$c_f = \frac{2\rho\nu U f''(0)}{\rho U^2}\sqrt{\frac{U}{\nu x}} = 2f''(0)\sqrt{\frac{\nu}{Ux}} = 2f''(0)\,\mathrm{Re}_x^{-1/2}\,, \tag{A.1381}$$

which with $f''(0) = 0.3321$ is 0.6642, the 0.664 of Phenomenon 36.12.

Plate drag. Integrate Equation (A.1380) along the plate:

$$F = \int_0^L \tau_w\,\mathrm{d}x = \mu U f''(0)\sqrt{\frac{U}{\nu}}\int_0^L x^{-1/2}\mathrm{d}x = 2\mu U f''(0)\sqrt{\frac{UL}{\nu}}\,, \tag{A.1382}$$

the integral converging at the leading edge despite the singularity of Equation (A.1380) there. Then

$$C_D = \frac{2F}{\rho U^2 L} = \frac{4\rho\nu U f''(0)}{\rho U^2 L}\,\mathrm{Re}_L^{1/2} = 4f''(0)\,\mathrm{Re}_L^{-1/2}\,, \tag{A.1383}$$

which is 1.3284. The relation $C_D = 2c_f(L)$ is exact and independent of the numerical value of $f''(0)$: it is the statement that the mean of $x^{-1/2}$ over $(0, L)$ is twice its endpoint value. □

Proposition A.802 (Consistency with the wake momentum deficit)**.** *The momentum thickness Equation* (36.25) *of the Blasius profile is*

$$\vartheta(x) = I\sqrt{\frac{\nu x}{U}}\,, \qquad I := \int_0^\infty f'\,(1 - f')\,\mathrm{d}\eta = 2f''(0)\,, \tag{A.1384}$$

so that Equation (36.25) *returns exactly the drag Equation* (A.1382) *computed from the wall stress.*
Rests on Equations (36.25) *and* (A.1366) *and Proposition A.801.*

Proof. *Derives Proposition A.802.*

The first member is Equation (A.1368) substituted into the definition of ϑ in Equation (36.25), with $\mathrm{d}y = \sqrt{\nu x/U}\,\mathrm{d}\eta$. For the identity, differentiate the product $f\,(1 - f')$ and use the Blasius equation Equation (A.1366) in the form $ff'' = -2f'''$:

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\left[f\,(1 - f')\right] = f'\,(1 - f') - ff'' = f'\,(1 - f') + 2f'''\,. \tag{A.1385}$$

Integrate from 0 to ∞. On the left, $f(0) = 0$ kills the lower limit, and at the upper limit $1 - f'$ vanishes exponentially while f grows only linearly, so the product vanishes; the left

side is therefore zero. On the right the first term is I and the second is $2\,[f']_0^\infty = -2f''(0)$, since $f'' \to 0$ with $1 - f'$. Hence $I = 2f''(0)$.

Substituting into Equation (36.25), $F = \rho U^2 \vartheta(L) = 2f''(0)\rho U^2\sqrt{\nu L/U} = 2\mu U f''(0)\sqrt{UL/\nu}$, which is Equation (A.1382). The two routes to the drag — integrating the stress the fluid exerts on the plate, and traversing the wake far downstream — agree exactly, as Proposition 36.14 requires, and neither uses the other in its derivation. □

Example A.803 (The chapter's worked number, in SI)**.** A plate 1 m long and of unit span, held edge-on in water of kinematic viscosity $1.0 \times 10^{-6}\,\mathrm{m^2/s}$ and density $1.0 \times 10^{3}\,\mathrm{kg/m^3}$, at 1 m/s. Then $\mathrm{Re}_L = 10^6$, $\mathrm{Re}_L^{-1/2} = 10^{-3}$, and the dynamic pressure is $\frac{1}{2}\rho U^2 = 500\,\mathrm{Pa}$, so

$$\begin{aligned} \delta_{99}(L) &= 5.0 \times 10^{-3} \times 1\,\mathrm{m} = 5\,\mathrm{mm}\,, \\ \tau_w(L) &= 0.664 \times 10^{-3} \times 500\,\mathrm{Pa} = 0.33\,\mathrm{Pa}\,, \\ F &= 1.328 \times 10^{-3} \times 500\,\mathrm{Pa} \times 1\,\mathrm{m} = 0.66\,\mathrm{N/m}\,, \end{aligned} \tag{A.1386}$$

per side. The layer is half a percent of the plate, which is the number Phenomenon 36.12 quotes, and the drag on a square metre of wetted surface is under a newton — a useful calibration of how small viscous friction is, and of how completely it nevertheless governs the flow when it separates.

Rests on Proposition A.801 and Phenomenon 36.12.

A.68.7 What these equations cannot describe

Remark A.804 (Three limits, all of them real). *Transition.* The solution above is laminar. The plate's own layer becomes turbulent at Re_x of order 5×10^5, so in Example A.803 the laminar solution is honest over roughly the first half metre and the quoted whole-plate drag is the laminar figure carried past its own domain of validity. The chapter records the same transition for the pipe at Section 36.3, and the threshold is no better determined here than it is there.

Separation. Equation (A.1355) is *parabolic* in x: it is marched downstream from an initial profile, and information travels only forwards. That is exactly what fails at separation. As the wall stress approaches zero under a rising external pressure, the marching solution develops a square-root singularity in x at the station where $(\partial_y u)_{y=0}$ vanishes — Goldstein's singularity, established in 1948 — and the computation cannot be continued through it. The layer equations therefore cannot describe the separated region whose existence is the principal observation of Phenomenon 36.12. This is why Proposition 36.13 is proved in the chapter by an expansion of the *full* equations at the wall, and not from Equation (36.26): the criterion Equation (36.24) has to be obtained by an argument that survives at the point where these equations break down. Goldstein's paper has no key in `references.bib` and the attribution here is made in words.

Leading edge. Equation (A.1380) diverges as $x \to 0$, where δ is comparable with x and the assumption $\partial_x^2 u \ll \partial_y^2 u$ used in Theorem A.798 fails. The divergence is integrable, so Equation (A.1382) is finite and the drag coefficient is unaffected at leading order; but the stress distribution near the leading edge is not given correctly by this theory.

Remark A.805. Appendix A.68 discharges the derivation owed at Proposition 36.15 of Chapter 36, in the boundary-layer experiment Section 36.5. Three things done here are used there and are worth carrying back. Theorem A.798 derives the exponent that Equation (36.23) estimated, so the chapter's resolution of the d'Alembert paradox rests on a reduction rather than on an order-of-magnitude argument. Corollary A.799 identifies what the outer flow gives the layer, which is the content of the chapter's remark that the pressure is "impressed" on the layer from outside and is the hypothesis under which Proposition 36.13 is stated. And Proposition A.802 closes the loop with Proposition 36.14: the drag measured by a

traverse of the wake and the drag computed from the stress at the wall are the same number, exactly, for the one flow in which both can be evaluated in closed form.

Appendix B

Physical Constants and SI Units

Since the revision that entered into force on 20 May 2019, the International System of Units (SI) is defined by assigning exact numerical values to seven defining constants of nature; every SI unit is then realised as a consequence of these fixed values [2333]. All other physical constants are determined experimentally, and this treatise uses throughout the recommended values of the CODATA 2018 adjustment [2392]. The present appendix collects, first, the seven defining constants with their exact values (Appendix B.2); second, the derived and measured constants that recur in the main text, quoted with their standard uncertainties (Appendix B.3); and third, the seven SI base units (Appendix B.5).

B.1 Mathematical constants

Two dimensionless numbers precede every physical constant in this appendix, since they are not measured but proven: $\mathrm{e} = 2.718\,281\,828\ldots$ and $\pi = 3.141\,592\,653\ldots$, both derived in full — e as the sum $\sum 1/n!$ and the limit $\lim(1+1/n)^n$ (Definition 7.53 and Proposition 7.57), π as twice the first positive zero of the cosine series and independently as the circle constant (Definition 7.61 and Proposition 7.62) — in Section 7.9.1 of Chapter 7. They are listed here only for numerical reference; their proofs are not repeated.

The digits are a separate claim from the existence of the constants, and worth separating explicitly. A definition that converges establishes that a number exists; it does not by itself justify writing down a single decimal. For e both steps are carried out: the truncation bound $\frac{1}{(n+1)!} \le \mathrm{e} - \sum_{k\le n} 1/k! < \frac{1}{n\cdot n!}$ (Lemma 7.54) traps e between two explicit rationals, and twelve terms already force the nine decimals quoted above (Corollary 7.55). The trailing … is a theorem too, not a typographical hedge: e is irrational (Proposition 7.56), so no finite decimal in this appendix equals it. Every entry below is a truncation; the only question is whether its error is stated.

The same two claims for π are asserted here but not yet derived.

Derivation pending #281

Digits and irrationality of π: an explicit error bound pinning $\pi = 3.141592653\ldots$, and Niven's proof that π is irrational. The latter needs $\sin\pi = 0$ and $\cos\pi = -1$, hence the addition formulas, which the real-analysis chapter does not yet establish — it defines π as twice the first positive zero of the cosine series and proves it is the circle constant, but computes no decimal of it.

B.2 The seven defining constants of the SI

The 2019 revision fixes the numerical values listed in Table B.1 exactly, by definition and without uncertainty [2333]. Each constant anchors one base unit, although the definitions are interlocking: the kilogram, for instance, is realised through the fixed value of the Planck constant together with the definitions of the second and the metre.

Table B.1: The seven defining constants of the SI (2019 revision). All values are exact by definition [2333].

Defining constant	Symbol	Exact value	Anchors
hyperfine transition frequency of ^{133}Cs	$\Delta\nu_{\mathrm{Cs}}$	$9\,192\,631\,770\,\mathrm{Hz}$	s
speed of light in vacuum	c	$299\,792\,458\,\mathrm{m/s}$	m
Planck constant	h	$6.626\,070\,15 \times 10^{-34}\,\mathrm{J\,s}$	kg
elementary charge	e	$1.602\,176\,634 \times 10^{-19}\,\mathrm{C}$	A
Boltzmann constant	k	$1.380\,649 \times 10^{-23}\,\mathrm{J/K}$	K
Avogadro constant	N_{A}	$6.022\,140\,76 \times 10^{23}/\mathrm{mol}$	mol
luminous efficacy of $540 \times 10^{12}\,\mathrm{Hz}$ radiation	K_{cd}	$683\,\mathrm{lm/W}$	cd

B.3 CODATA 2018 derived and measured constants

Table B.2 lists the constants used most frequently in this treatise, with the CODATA 2018 recommended values and standard uncertainties [2392]. Two entries deserve comment. The reduced Planck constant $\hbar = h/2\pi$ is exact because h is exact; its decimal representation is merely truncated in the table. The standard acceleration of free fall g_0 is not a constant of nature at all but a conventional exact value, adopted by the third Conférence Générale des Poids et Mesures in 1901 and retained in the SI Brochure [2333]; the local acceleration of free fall differs from it at the level of a few parts in 10^3 and is what absolute gravimeters actually measure [1906].

B.4 Planck units

The constants $\hbar$, c, and G together fix a length, mass, and time with no free numerical factor left undetermined by dimensional analysis alone.

Proposition B.1 (Planck units)**.** *The unique (up to a dimensionless constant) combinations of $\hbar$, c, G with dimensions of length, mass, and time are*

$$\ell_{\mathrm{P}} = \sqrt{\frac{\hbar G}{c^3}}\,, \qquad m_{\mathrm{P}} = \sqrt{\frac{\hbar c}{G}}\,, \qquad t_{\mathrm{P}} = \sqrt{\frac{\hbar G}{c^5}} = \frac{\ell_{\mathrm{P}}}{c}\,. \tag{B.1}$$

Rests on Theorem 2.3 and Axiom 2.2.

Proof. *Derives Proposition B.1.*

Write $\hbar^a c^b G^d$ and demand dimension L (length). In SI, $[\hbar] = \mathsf{M}\mathsf{L}^2\mathsf{T}^{-1}$, $[c] = \mathsf{L}\mathsf{T}^{-1}$, $[G] = \mathsf{M}^{-1}\mathsf{L}^3\mathsf{T}^{-2}$. Matching powers of $\mathsf{M}, \mathsf{L}, \mathsf{T}$ gives the linear system $a - d = 0$ (M), $2a + b + 3d = 1$ (L), $-a - b - 2d = 0$ (T); solving, $a = d = 1/2$, $b = -3/2$, i.e. $\hbar^{1/2} c^{-3/2} G^{1/2} = \sqrt{\hbar G/c^3}$. By the Buckingham π theorem (Theorem 2.3), three quantities built from three dimensionally independent constants leave zero dimensionless combinations, so this exponent solution is unique and no undetermined dimensionless prefactor can be fixed by dimensional analysis alone — exactly the freedom absorbed into the convention that defines ℓ_{P} with prefactor 1. The mass and time forms follow by the same linear system with target dimension M and T

Table B.2: Derived and measured constants, CODATA 2018 recommended values [2392]. The column u_{r} gives the relative standard uncertainty.

Quantity	Symbol	Value	u_{r}
Newtonian constant of gravitation	G	$(6.674\,30 \pm 0.000\,15) \times 10^{-11}\,\mathrm{m^3/(kg\,s^2)}$	2.2×10^{-5}
reduced Planck constant	$\hbar$	$1.054\,571\,817 \times 10^{-34}\,\mathrm{J\,s}$	exact[a]
electron mass	m_{e}	$(9.109\,383\,701\,5 \pm 0.000\,000\,002\,8) \times 10^{-31}\,\mathrm{kg}$	3.0×10^{-10}
proton mass	m_{p}	$(1.672\,621\,923\,69 \pm 0.000\,000\,000\,51) \times 10^{-27}\,\mathrm{kg}$	3.1×10^{-10}
vacuum magnetic permeability	μ_0	$(1.256\,637\,062\,12 \pm 0.000\,000\,000\,19) \times 10^{-6}\,\mathrm{N/A^2}$	1.5×10^{-10}
vacuum electric permittivity	ϵ_0	$(8.854\,187\,812\,8 \pm 0.000\,000\,001\,3) \times 10^{-12}\,\mathrm{F/m}$	1.5×10^{-10}
fine-structure constant	α	$(7.297\,352\,569\,3 \pm 0.000\,000\,001\,1) \times 10^{-3}$	1.5×10^{-10}
Bohr magneton	μ_{B}	$(9.274\,010\,078\,3 \pm 0.000\,000\,002\,8) \times 10^{-24}\,\mathrm{J/T}$	3.0×10^{-10}
standard acceleration of free fall	g_0	$9.806\,65\,\mathrm{m/s^2}$	exact[b]

[a] Exact because $\hbar = h/2\pi$ with h fixed; the decimal digits shown are truncated. [b] Conventional exact value, not a constant of nature [2333].

respectively (or, more quickly, from $m_{\mathrm{P}} = \hbar/(\ell_{\mathrm{P}} c)$ and $t_{\mathrm{P}} = \ell_{\mathrm{P}}/c$, using $[\hbar] = \mathsf{ML^2T^{-1}}$ to fix the mass combination and dimensional consistency to fix the time combination). □

Numerically, from the CODATA values of Table B.2,

$$\ell_{\mathrm{P}} \approx 1.616 \times 10^{-35}\,\mathrm{m}\,, \qquad m_{\mathrm{P}} \approx 2.176 \times 10^{-8}\,\mathrm{kg}\,, \qquad t_{\mathrm{P}} \approx 5.391 \times 10^{-44}\,\mathrm{s}\,, \tag{B.2}$$

the length and time scales at which quantum and gravitational effects are expected to become comparable in magnitude, and the mass scale against which the hierarchy problem of Appendix C is stated quantitatively.

B.5 SI base units

The seven base units of the SI, from which all coherent derived units are formed by products of powers, are collected in Table B.3 [2333]. Every quantity in this treatise is expressed in these units or in coherent derived units built from them, in accordance with the axiom stated in the front matter.

Table B.3: The seven SI base units [2333].

Base quantity	Base unit	Symbol
time	second	s
length	metre	m
mass	kilogram	kg
electric current	ampere	A
thermodynamic temperature	kelvin	K
amount of substance	mole	mol
luminous intensity	candela	cd

Appendix C

The Free Parameters of Physics

This appendix catalogs the complete set of dimensionless (or, where noted, dimensionful-but-irreducible) parameters that the evidence-based physics of this treatise — the Standard Model of particle physics and the base ΛCDM cosmological model — treats as free: numbers no confirmed theory yet predicts from anything more basic. The mathematical constants π and e (Section 7.9.1) and the dimensionful SI constants c, $\hbar$, G, ϵ_0, and the fine-structure constant α are catalogued separately in Appendix B, since they are either proven mathematical facts or exactly-defined/precisely-measured physical constants rather than free parameters in the sense used here.

Every entry below states, where one exists, how the quantity is *constructed* from more basic measured inputs; where no such construction exists, the entry says so plainly — the quantity is itself one of the truly free numbers of physics. Entries marked **(open)** are ones whose value, nature, or theoretical status is not yet settled by current evidence; each says explicitly what is and is not yet known, in keeping with this treatise's evidence-only scope (Chapter 1). None of the speculative theoretical programs sometimes proposed to explain these numbers — supersymmetry, extra dimensions, or similar — is invoked here as an explanation, since none is yet supported by direct evidence; where relevant this is stated explicitly, and the broader question is returned to in Chapter 133.

C.1 Electroweak and Higgs

The parameters catalogued here fix the electroweak sector of the Standard Model: the strength and mixing of the $SU(2)_L \times U(1)_Y$ gauge interactions, the scale of spontaneous symmetry breaking, and the mass and self-coupling of the Higgs boson that breaks it. The fine-structure constant α, the third traditional input of electroweak fits alongside G_F and m_Z, is already catalogued in Appendix B (`docs/latex/appendices/B-constants-units.tex`) and is not repeated here. All numerical values below are taken from the Particle Data Group's 2024 Review of Particle Physics. Because several of these quantities (the weak mixing angle, the gauge boson masses) are related by tree-level and loop-corrected Standard Model relations, they overdetermine the electroweak Lagrangian and their mutual consistency is itself a stringent test of the theory.

Fermi coupling constant G_F: $(1.166\,378\,7 \pm 0.000\,000\,6) \times 10^{-5}/\mathrm{GeV}^2$.

measured directly (muon lifetime: $1/\tau_\mu = \dfrac{G_F^2 m_\mu^5}{192\pi^3}(1 + \Delta q)$, Δq the known QED radiative correction).

The Fermi coupling constant G_F parametrizes the strength of charged-current weak interactions at energies far below the W-boson mass, in the effective four-fermion contact-interaction language that predates the discovery of the W and Z bosons. It is extracted

Table C.1: Electroweak and Higgs: summary of catalogued parameters.

Quantity	Symbol	Value
Fermi coupling constant	G_F	$(1.166\,378\,7 \pm 0.000\,000\,6) \times 10^{-5}/\mathrm{GeV}^2$
Higgs vacuum expectation value	v	$(246.219\,65 \pm 0.000\,06)$ GeV
Weak mixing angle	$\sin^2\hat{\theta}_W(m_Z)$	$0.231\,22 \pm 0.000\,04$
W boson mass	m_W	(80.3692 ± 0.0133) GeV
Z boson mass	m_Z	(91.1876 ± 0.0021) GeV
Higgs boson mass	m_H	(125.20 ± 0.11) GeV
Higgs quartic self-coupling	λ	0.1293 ± 0.0002

with extraordinary precision from the muon lifetime, measured most accurately by the MuLan experiment at PSI, combined with QED radiative corrections computed to several loop orders. Because G_F is known to a relative precision of order 5×10^{-7}, it – rather than v directly – is used as one of the primary electroweak input parameters, alongside α (Appendix B) and m_Z, in global electroweak fits. Through $v = (\sqrt{2}G_F)^{-1/2}$ it fixes the overall scale of electroweak symmetry breaking. [2427]

Higgs vacuum expectation value v: $(246.219\,65 \pm 0.000\,06)$ GeV.

$v \equiv (\sqrt{2}\,G_F)^{-1/2}$.

The vacuum expectation value v of the neutral component of the Higgs doublet sets the scale at which $SU(2)_L \times U(1)_Y$ electroweak symmetry breaks spontaneously to the electromagnetic $U(1)_{\mathrm{em}}$, and through the Yukawa couplings $y_f = \sqrt{2}\,m_f/v$ it fixes the overall normalization of every fermion mass in the Standard Model. It is not measured directly but derived from the far more precisely known Fermi constant via $v = (\sqrt{2}G_F)^{-1/2}$, since G_F and v describe the same low-energy strength of the charged-current weak interaction. Its numerical value, close to 246 GeV, together with the Higgs mass m_H fixes the Higgs quartic self-coupling λ (see below). [2427]

Weak mixing angle $\sin^2\hat{\theta}_W(m_Z)$: $0.231\,22 \pm 0.000\,04$.

$\overline{\mathrm{MS}}$ scheme: $\sin^2\hat{\theta}_W(\mu = m_Z) \equiv \dfrac{\hat{g}'^2}{\hat{g}^2 + \hat{g}'^2}\Big|_{\overline{\mathrm{MS}}}$; on-shell scheme: $\sin^2\theta_W \equiv 1 - m_W^2/m_Z^2$.

The weak mixing (Weinberg) angle θ_W rotates the neutral $SU(2)_L \times U(1)_Y$ gauge fields W_μ^3, B_μ into the physical Z boson and the photon, and its sine-squared is the single additional number, beyond α and G_F (or m_Z), needed to fix all electroweak neutral-current couplings at tree level. Because $\sin^2\theta_W$ is a ratio of couplings, its numerical value is renormalization-scheme dependent: the value quoted here is the $\overline{\mathrm{MS}}$ value evaluated at $\mu = m_Z$, obtained from global electroweak fits to Z-pole asymmetries and low-energy neutral-current data. The alternative on-shell definition, $\sin^2\theta_W \equiv 1 - m_W^2/m_Z^2$, gives a numerically different value – using the PDG-2024 m_W and m_Z listed below, $1 - m_W^2/m_Z^2 \approx 0.2232$ – because it absorbs electroweak radiative corrections differently order by order. [2427]

W boson mass m_W: (80.3692 ± 0.0133) GeV.

$m_W = \frac{1}{2}gv$, fixing the $SU(2)_L$ gauge coupling g once v is known; measured directly.

The $W^\pm$ boson mass is one of the two masses generated by the Higgs mechanism through the covariant derivative of the Higgs doublet, $m_W = \frac{1}{2}gv$, so its measurement, combined with the independently fixed v, determines the $SU(2)_L$ gauge coupling g. It is measured kinematically from the transverse-mass and transverse-momentum distributions

of leptons in $W \to \ell\nu$ decays at hadron colliders (Tevatron, LHC) and, less precisely, from $e^+e^- \to W^+W^-$ threshold scans at LEP2. The value quoted is the PDG-2024 world average; the 2022 CDF-II measurement ($(80.4335 \pm 0.0094)\,\mathrm{GeV}$) lies in tension of several standard deviations with this average and with the Standard Model electroweak fit, a discrepancy still under active experimental scrutiny following subsequent 2024 measurements (including a CMS result and a revised ATLAS analysis) that favor the lower, SM-consistent value. [2427]

Z boson mass m_Z: $(91.1876 \pm 0.0021)\,\mathrm{GeV}$.

$m_Z = \frac{1}{2}v\sqrt{g^2 + g'^2}$, fixing the combination $g^2 + g'^2$ once v is known; measured directly.

The Z boson mass is fixed, like m_W, by electroweak symmetry breaking, $m_Z = \frac{1}{2}v\sqrt{g^2 + g'^2}$, so its measurement fixes the combination of $SU(2)_L$ and $U(1)_Y$ gauge couplings $g^2 + g'^2$ once v is known from G_F. It is by far the most precisely known electroweak boson mass, extracted from the $e^+e^- \to Z$ resonance line shape scanned at LEP1 and at the SLC, where it appears as the peak of the total hadronic and leptonic cross sections. Its relative precision of about 2×10^{-5} makes it, together with α and G_F, one of the three canonical input parameters of the electroweak Lagrangian in most global-fit conventions. [2427]

Higgs boson mass m_H: $(125.20 \pm 0.11)\,\mathrm{GeV}$.

measured directly (resonance peak in $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$); fixes the quartic self-coupling via $m_H^2 = 2\lambda v^2$.

The Higgs boson mass, established in 2012 by ATLAS and CMS through the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$ decay channels, is measured directly as the invariant-mass peak of the decay products; unlike the fermion and gauge-boson masses, it is not predicted from a more fundamental parameter within the Standard Model but is itself a genuinely free input, fixing the strength of the Higgs field's quartic self-interaction λ through $m_H^2 = 2\lambda v^2$. The value quoted is the PDG-2024 world average combining the high-resolution di-photon and four-lepton mass measurements of both experiments. Once m_H and v are fixed, every other tree-level Higgs property – total width, branching ratios, and the self-coupling λ – follows with no further free parameters, which is precisely what ongoing LHC Higgs-coupling measurements continue to test. [2427]

Higgs quartic self-coupling λ: 0.1293 ± 0.0002.

$\lambda = m_H^2/(2v^2)$.

The Higgs quartic self-coupling λ multiplies the $\lambda(\Phi^\dagger\Phi)^2$ term of the Higgs potential and, together with v, sets both the shape of the electroweak vacuum and the physical Higgs boson mass, $m_H^2 = 2\lambda v^2$. Unlike m_H, λ is not itself measured to comparable precision: its value quoted here is inferred algebraically from the two directly measured quantities m_H and v, giving $\lambda \approx 0.129$. A genuinely independent, direct determination of λ requires observing Higgs pair production ($pp \to HH$) and extracting the Higgs trilinear self-coupling, a measurement that remains statistics-limited at the LHC and has so far only placed loose bounds around the Standard Model expectation. The measured smallness of λ relative to the top-quark Yukawa coupling squared is also central to the (unresolved) question of the stability of the electroweak vacuum under renormalization-group running to high energy scales. [2427]

C.2 Quantum Chromodynamics

Quantum chromodynamics (QCD), the $\mathrm{SU}(3)_c$ gauge sector of the Standard Model, is characterized by a running coupling α_s and six quark mass parameters whose numerical values are, unlike the electroweak sector, deeply entangled with the choice of renormalization

scheme and scale: light-quark masses are conventionally quoted in $\overline{\text{MS}}$ at a fixed scale of 2 GeV, the charm and bottom masses in $\overline{\text{MS}}$ at their own mass scale, and the top mass – uniquely among quarks, since it decays before hadronizing – as an approximate pole mass reconstructed from collider kinematics. This scheme-dependence is not a bookkeeping nuisance but reflects the underlying fact that confined color charges have no scheme-independent, asymptotic-state definition of mass in the way leptons do. The sector also harbors one of the cleanest open problems in physics: the QCD vacuum angle $\bar{\theta}_{\text{QCD}}$, which the symmetries of the Standard Model permit to take any value of order unity but which experiment constrains to be smaller than one part in 10^{10}, with no confirmed dynamical explanation.

Table C.2: Quantum Chromodynamics: summary of catalogued parameters.

Quantity	Symbol	Value
Strong coupling constant	$\alpha_s(M_Z)$	0.1180 ± 0.0009
Up-quark mass	m_u	$2.16^{+0.49}_{-0.26}$ MeV
Down-quark mass	m_d	$4.67^{+0.48}_{-0.17}$ MeV
Strange-quark mass	m_s	$93.4^{+8.6}_{-3.4}$ MeV
Charm-quark mass	$m_c(m_c)$	(1.27 ± 0.02) GeV
Bottom-quark mass	$m_b(m_b)$	$4.18^{+0.03}_{-0.02}$ GeV
Top-quark mass	m_t	(172.57 ± 0.29) GeV
QCD vacuum angle (strong CP problem) [(open)]	$\bar{\theta}_{\text{QCD}}$	10^{-10} (see text)

Strong coupling constant $\alpha_s(M_Z)$: 0.1180 ± 0.0009.

$\alpha_s \equiv g_s^2/(4\pi)$, where g_s is the QCD gauge coupling; the value is scale-dependent (renormalization-group running governed by the QCD β-function) and is conventionally quoted at $\mu = M_Z$, the world average from lattice QCD, τ-decay, deep-inelastic scattering, event-shape, and electroweak precision-fit determinations.

The strong coupling $\alpha_s(\mu)$ measures the strength of the color force at renormalization scale μ; unlike the electromagnetic fine-structure constant α, which is essentially scale-independent at laboratory energies, α_s runs rapidly and grows large at low energy, producing confinement, and shrinks logarithmically at high energy, producing asymptotic freedom – the property for which the 2004 Nobel Prize was awarded. By convention it is quoted at the reference scale $\mu = M_Z \approx 91.19$ GeV, where the PDG world average combines lattice-QCD sum rules, τ-lepton hadronic decays, deep-inelastic scattering structure functions, e^+e^- jet event shapes, and the running of α_s inferred from Z-pole electroweak fits. Its value sets the size of perturbative QCD corrections throughout the treatise (hadron collider cross sections, jet physics, hadronic decay widths) and, via dimensional transmutation, fixes the intrinsic QCD scale $\Lambda_{\text{QCD}} \sim 200$ MeV at which the coupling formally diverges and confinement sets in. [2427]

Up-quark mass m_u: $2.16^{+0.49}_{-0.26}$ MeV.

measured directly, as a Lagrangian mass parameter (coefficient of $\bar{u}u$) in the $\overline{\text{MS}}$ renormalization scheme at $\mu = 2$ GeV; extracted from lattice QCD combined with QCD sum rules, chiral perturbation theory mass ratios, and hadron spectroscopy.

The up-quark mass is a free parameter of the Standard Model Lagrangian, fixed by the Yukawa coupling of u to the Higgs field, and is never observed as a free-particle mass because of color confinement. It is instead extracted as a running $\overline{\text{MS}}$-scheme mass parameter,

conventionally quoted at $\mu = 2\,\mathrm{GeV}$, from lattice QCD combined with chiral perturbation theory constraints on light-quark mass ratios and QCD sum rules applied to hadron spectra. This scheme dependence is physically essential: quoting "the" up-quark mass without specifying scheme and scale is meaningless, since the value shifts under renormalization-group running exactly as α_s does. The smallness and asymmetric uncertainty of m_u relative to Λ_{QCD} (light-quark masses are a small perturbation on hadron masses, most of which arises from gluon binding energy) also underlies the near-exact isospin symmetry of nuclear physics. [2427]

Down-quark mass m_d: $4.67^{+0.48}_{-0.17}\,\mathrm{MeV}$.

measured directly, as a Lagrangian mass parameter in the $\overline{\mathrm{MS}}$ scheme at $\mu = 2\,\mathrm{GeV}$; extracted from lattice QCD, chiral perturbation theory, and QCD sum rules, jointly with m_u and m_s.

Like m_u, the down-quark mass is a free Standard-Model parameter quoted in the $\overline{\mathrm{MS}}$ scheme at $\mu = 2\,\mathrm{GeV}$, obtained jointly with the other light-quark masses from lattice QCD and chiral perturbation theory fits to pion, kaon, and eta meson masses and decay constants. That $m_d > m_u$ by roughly a factor of two is the dominant source of the proton-neutron mass difference (partially offset by electromagnetic self-energy), and is responsible for the neutron being heavier than the proton – a fact of considerable cosmological importance, since it permits free-neutron beta decay and shapes primordial nucleosynthesis. As with m_u, no experiment isolates a free down quark; all determinations are indirect, via QCD-mediated observables. [2427]

Strange-quark mass m_s: $93.4^{+8.6}_{-3.4}\,\mathrm{MeV}$.

measured directly, as a Lagrangian mass parameter in the $\overline{\mathrm{MS}}$ scheme at $\mu = 2\,\mathrm{GeV}$; extracted principally from lattice QCD and QCD sum rules applied to kaon and η/η' physics.

The strange-quark mass, quoted in the $\overline{\mathrm{MS}}$ scheme at the same reference scale $\mu = 2\,\mathrm{GeV}$ used for m_u and m_d, is roughly twenty times larger than the down-quark mass and is the dominant explicit breaking term for flavor SU(3) symmetry among the light hadrons. It is determined mainly from lattice QCD (moments of correlators, kaon masses and decay constants) and from QCD sum rules constrained by τ-decay and e^+e^- data; its uncertainty, while still asymmetric and scheme-dependent, has shrunk considerably as lattice methods have matured. Together, m_u, m_d, m_s set the size of chiral symmetry breaking effects that organize the light hadron spectrum (pions and kaons as pseudo-Goldstone bosons of spontaneously broken chiral symmetry). [2427]

Charm-quark mass $m_c(m_c)$: $(1.27 \pm 0.02)\,\mathrm{GeV}$.

measured directly, as a Lagrangian mass parameter in the $\overline{\mathrm{MS}}$ scheme evaluated at its own mass scale, $\mu = m_c$ (the conventional choice for heavy quarks, distinct from the fixed $\mu = 2\,\mathrm{GeV}$ used for u, d, s); extracted from charmonium sum rules and lattice QCD.

Unlike the light quarks, the charm-quark mass is conventionally quoted in the $\overline{\mathrm{MS}}$ scheme evaluated at its own mass, $\mu = m_c(m_c)$, rather than at the fixed 2 GeV reference used for u, d, s – a scheme choice that avoids large logarithms in the perturbative relation between the running mass and observables near the charm threshold. The dominant determinations come from QCD sum rules applied to $e^+e^- \to$ charm-quark-pair cross-section data and charmonium spectroscopy, cross-checked by lattice QCD. Because m_c sits close to Λ_{QCD} times a moderate factor, charm physics occupies an intermediate regime where both perturbative QCD and nonperturbative (lattice, effective field theory) methods are needed and mutually validated. [2427]

Bottom-quark mass $m_b(m_b)$: $4.18^{+0.03}_{-0.02}\,\mathrm{GeV}$.

measured directly, as a Lagrangian mass parameter in the $\overline{\text{MS}}$ scheme evaluated at its own mass scale, $\mu = m_b$; extracted from bottomonium sum rules, lattice QCD, and inclusive/exclusive B-meson semileptonic decay spectra.

The bottom-quark mass, like the charm mass, is quoted in the $\overline{\text{MS}}$ scheme at its own scale $\mu = m_b(m_b)$, which keeps the perturbative series for mass-dependent observables well behaved. It is determined from a combination of bottomonium (Υ-system) sum rules, lattice QCD, and QCD-based analyses of inclusive and exclusive semileptonic B-meson decay spectra, all of which now agree to sub-percent precision. The bottom mass is a key input to CKM-matrix determinations from B-physics (Part on flavor physics) and to Higgs-boson branching-ratio predictions, since $H \to b\bar{b}$ is the dominant Higgs decay channel and its rate depends quadratically on m_b. [2427]

Top-quark mass m_t: $(172.57 \pm 0.29)\,\text{GeV}$.

measured directly, as a pole mass (not an $\overline{\text{MS}}$ running mass, unlike the other five quarks) from kinematic reconstruction of top-antitop pair-production events at the Tevatron and LHC.

The top quark is the only quark reported as a pole mass rather than an $\overline{\text{MS}}$ running mass, because it decays via the electroweak interaction ($t \to W^+ b$) before it can hadronize – it is the one quark whose mass can be reconstructed directly from the kinematics of its (quasi-free) decay products, bypassing the scheme ambiguities inherent to confined quarks. The world average combines direct kinematic reconstructions from ATLAS, CMS, and the Tevatron experiments (CDF, D0). It should be noted that the "pole mass" extracted this way carries a residual, formally ill-defined nonperturbative ambiguity of order Λ_{QCD} (the renormalon ambiguity), so the quoted value is understood as an operationally defined Monte-Carlo mass closely related to, but not rigorously identical to, the theoretical pole mass; conversion to a well-defined short-distance mass scheme (e.g. $\overline{\text{MS}}$) shifts the central value down by roughly 9 GeV to 10 GeV (e.g. $\overline{m}_t(\overline{m}_t) \approx 163\,\text{GeV}$ versus the pole-like value above). At $\sim 173\,\text{GeV}$, comparable to the electroweak scale, the top mass gives the largest Yukawa coupling in the Standard Model ($y_t \approx 1$) and dominates radiative corrections to the Higgs mass and to electroweak precision observables. [2427]

QCD vacuum angle (strong CP problem) (open) $\bar{\theta}_{\text{QCD}}$: $\left|\bar{\theta}_{\text{QCD}}\right| \lesssim 10^{-10}$ (90% C.L., from the neutron electric dipole moment bound).

appears as the coefficient of the CP-violating topological term $\mathcal{L}_\theta = \theta_{\text{QCD}} \dfrac{g_s^2}{32\pi^2} G^a_{\mu\nu} \tilde{G}^{a\,\mu\nu}$ in the QCD Lagrangian; the physically observable combination $\bar{\theta}_{\text{QCD}} = \theta_{\text{QCD}} + \arg\det(M_u M_d M_s \cdots)$ absorbs an equal, in-principle-independent phase from the quark mass matrix, and is bounded from above by the non-observation of a neutron electric dipole moment.

The QCD Lagrangian admits a gauge-invariant, Lorentz-invariant, renormalizable term $\theta_{\text{QCD}}\,(g_s^2/32\pi^2)\,G\tilde{G}$ built from the gluon field strength and its dual, which violates parity and time-reversal (hence CP, by CPT) and is entirely permitted by the symmetries of the Standard Model; the physical angle $\bar{\theta}_{\text{QCD}}$ also receives a contribution from the phase of the determinant of the quark mass matrix, so it is not removable by a field redefinition once electroweak symmetry breaking generates quark masses. A nonzero $\bar{\theta}_{\text{QCD}}$ of order unity is the generic theoretical expectation, since it is an allowed, unsuppressed term with no symmetry reason to vanish, and it would produce an observably large neutron electric dipole moment $d_n \sim \bar{\theta}_{\text{QCD}} \times e\,10^{-16}\,\text{cm}$. Experimentally, the most sensitive constraint comes from ultracold-neutron electric-dipole-moment experiments, whose current upper bound on $|d_n|$ (of order $e\,10^{-26}\,\text{cm}$) translates into $\left|\bar{\theta}_{\text{QCD}}\right| \lesssim 10^{-10}$, a suppression by roughly ten orders of magnitude below its theoretically natural size. This unexplained fine-tuning – why a

free, dimensionless parameter of the strong interaction sits so close to zero when nothing in the Standard Model forces it to – constitutes the strong CP problem, one of the cleanest and longest-standing naturalness puzzles in particle physics; it remains open within the evidence-based scope of this treatise; proposed dynamical solutions (e.g. a spontaneously broken Peccei–Quinn symmetry and its associated axion) are not yet supported by direct experimental evidence and are noted here only for completeness, not asserted. [2427]

C.3 Charged Lepton Masses and Quark Mixing (CKM)

The charged-lepton sector of the Standard Model exhibits a mass hierarchy spanning more than three orders of magnitude between the electron, muon, and tau, with no accepted dynamical explanation for the specific ratios observed – this is the long-standing "flavor puzzle." In the quark sector, the mismatch between the mass eigenbasis and the weak-interaction eigenbasis of the up- and down-type quarks is encoded in the unitary Cabibbo–Kobayashi–Maskawa (CKM) matrix V_{CKM}, conventionally written in the standard parametrization as a product of three Euler-like rotations through mixing angles $\theta_{12}, \theta_{13}, \theta_{23}$ and a single physical complex phase δ_{13} that is the sole source of CP violation in the quark sector at tree level in the SM. All four parameters, together with the nine $|V_{ij}|$ moduli they generate, are extracted from a global fit to semileptonic decays, meson mixing, and CP-violating observables (e.g., ϵ_K, $\sin 2\beta$) tabulated by the Particle Data Group. The Jarlskog invariant J packages $\theta_{12}, \theta_{13}, \theta_{23}, \delta_{13}$ into the unique rephasing- and basis-independent scalar measure of CP violation, and its smallness relative to the maximal value permitted by unitarity ($J_{\max} = 1/(6\sqrt{3}) \approx 0.096$) is what makes CP-violating effects in flavor-conserving SM processes generically tiny.

Table C.3: Charged Lepton Masses and Quark Mixing (CKM): summary of catalogued parameters.

Quantity	Symbol	Value
Muon-to-electron mass ratio	m_μ/m_e	$206.768\,283\,0 \pm 0.000\,004\,6$
Tau-to-electron mass ratio	m_τ/m_e	3477.23 ± 0.23
Cabibbo angle	θ_{12}	$(13.02 \pm 0.04)^\circ$
CKM mixing angle (13)	θ_{13}	$(0.212 \pm 0.006)^\circ$
CKM mixing angle (23)	θ_{23}	$(2.40 \pm 0.05)^\circ$
CKM CP-violating phase	δ_{13}	(1.144 ± 0.027) rad
Jarlskog invariant	J	$(3.08 \pm 0.15) \times 10^{-5}$

Muon-to-electron mass ratio m_μ/m_e: $206.768\,283\,0 \pm 0.000\,004\,6$.

m_μ/m_e(ratio of the independently measured muon and electron pole masses).

The muon-to-electron mass ratio is one of the most precisely known numbers in particle physics, determined from the independently measured muon mass (itself pinned down via muonium hyperfine spectroscopy and muon spin rotation in magnetic fields) and the electron mass from Penning-trap g-factor measurements. Both leptons couple identically to the electroweak gauge fields and differ only through their Yukawa couplings to the Higgs field, so this ratio is, in the Standard Model, simply the ratio of two a priori independent, unrelated Yukawa parameters. No symmetry or dynamical principle within the SM fixes or explains why $m_\mu/m_e \approx 207$ rather than some other value; this is the leptonic instance of the flavor puzzle. The ratio enters directly in tests of lepton universality, muonium spectroscopy, and the theoretical prediction of the muon's anomalous magnetic moment. [2427]

Tau-to-electron mass ratio m_τ/m_e: 3477.23 ± 0.23.

m_τ/m_e(ratio of the measured tau and electron pole masses).

The tau lepton mass, and hence this ratio, is measured principally from the tau pair production threshold scan in e^+e^- collisions (BES, Belle) via precise reconstruction of the threshold cross-section shape, giving a much larger relative uncertainty than the muon case because the tau's short lifetime prevents the ultra-precise spectroscopic techniques available for the muon and electron. As with m_μ/m_e, the Standard Model provides three independent Yukawa couplings for the charged leptons with no built-in relation between them, so the value $m_\tau/m_e \approx 3477$ is an unexplained input rather than a derived consequence of any known symmetry. Together with m_μ/m_e, this ratio is a central piece of evidence for the flavor puzzle: three generations of otherwise identical particles (same electroweak quantum numbers) with masses set by numerically unrelated, hierarchically spread coupling constants that the SM does not predict. [2427]

Cabibbo angle θ_{12}: $(13.02 \pm 0.04)^\circ$.

$\sin\theta_{12} = |V_{us}| / \sqrt{|V_{ud}|^2 + |V_{us}|^2}$(standard parametrization of V_{CKM}).

The Cabibbo angle is the largest of the three CKM mixing angles, first introduced by Cabibbo in 1963 to explain the suppression of strangeness-changing weak decays relative to strangeness-conserving ones, well before the third quark generation was known to exist. It is extracted from a combination of superallowed nuclear β decays and pion/kaon semileptonic decays (fixing $|V_{ud}|$ and $|V_{us}|$) together with the CKM unitarity constraint. Its near-maximal size relative to θ_{13} and θ_{23} reflects the empirical fact that quark mixing is strongly hierarchical, concentrated between the first two generations, with the corresponding value $\sin\theta_{12} \approx 0.225$ often denoted the Wolfenstein parameter λ. [2427]

CKM mixing angle (13) θ_{13}: $(0.212 \pm 0.006)^\circ$.

$\sin\theta_{13} = |V_{ub}|$(standard parametrization of V_{CKM}).

This is the smallest of the three CKM mixing angles, governing the direct coupling between the first- and third-generation quarks. It is fixed essentially by $|V_{ub}|$, determined from semileptonic $b \to u\ell\nu$ transitions in both inclusive and exclusive ($B \to \pi\ell\nu$) decay channels; a persistent few-percent-level tension between the inclusive and exclusive extractions of $|V_{ub}|$ remains one of the more actively studied discrepancies in flavor physics. The extreme smallness of θ_{13} compared to θ_{12} is part of the same unexplained mixing hierarchy that characterizes the quark sector. [2427]

CKM mixing angle (23) θ_{23}: $(2.40 \pm 0.05)^\circ$.

$\sin\theta_{23} = |V_{cb}| / \sqrt{|V_{cb}|^2 + |V_{tb}|^2} \approx |V_{cb}|$(standard parametrization of V_{CKM}).

This angle governs mixing between the second and third quark generations and is fixed principally by $|V_{cb}|$, measured from semileptonic $B \to D^{(*)}\ell\nu$ exclusive decays and from inclusive $b \to c\ell\nu$ moment analyses. As with θ_{13}, mild tension persists between inclusive and exclusive determinations, currently at roughly the 1–2σ level and under continued scrutiny with improved lattice QCD form factors. Its intermediate size, between the large θ_{12} and the tiny θ_{13}, completes the hierarchical pattern of quark mixing angles that has no accepted first-principles explanation within the Standard Model. [2427]

CKM CP-violating phase δ_{13}: (1.144 ± 0.027) rad.

measured directly, as the complex phase of V_{CKM} in the standard (Chau–Keung) parametrization, from a global fit to CP-violating and CP-conserving flavor observables.

The phase δ_{13} is the single irreducible complex parameter of the CKM matrix once unphysical quark-field rephasings are removed, and it is the origin of all CP violation

observed in flavor-changing weak processes within the Standard Model, first evidenced in neutral kaon decays and later confirmed in the B-meson system by the BaBar and Belle experiments. It is extracted from a global fit combining CP-violating observables, most importantly the indirect CP violation parameter ϵ_K in the kaon system and the mixing-induced CP asymmetry $\sin 2\beta$ measured in $B^0 \to J/\psi K^0_S$ decays, together with the CP-conserving $|V_{ij}|$ measurements that fix the angles. That a single phase in a 3×3 unitary mixing matrix accounts for essentially all quark-sector CP violation measured to date is itself a strong, quantitatively successful prediction of the three-generation Standard Model, dating to Kobayashi and Maskawa's 1973 proposal. [2427]

Jarlskog invariant J: $(3.08 \pm 0.15)\times 10^{-5}$.

$$J = \Im\left(V_{us}V_{cb}V_{ub}^{*}V_{cs}^{*}\right) = \sin\theta_{12}\sin\theta_{13}\sin\theta_{23}\cos\theta_{12}\cos^{2}\theta_{13}\cos\theta_{23}\sin\delta_{13}$$

The Jarlskog invariant J is the unique combination of CKM angles and phase that is invariant under rephasing of the quark fields, and it appears as an overall multiplicative factor in every CP-violating asymmetry computable in the Standard Model quark sector, making it the single basis-independent scalar measure of how much CP violation the CKM mechanism can generate. Its value is obtained by propagating the fitted $\theta_{12},\theta_{13},\theta_{23},\delta_{13}$ (or equivalently the fitted V_{ij} moduli and phase convention) through the defining combination above. That $J \sim 3\times 10^{-5}$ is many orders of magnitude below both the naive $\mathcal{O}(1)$ scale set by the largest mixing angle and the unitarity bound $J_{\max} = 1/(6\sqrt{3}) \approx 0.096$ is the quantitative statement of why CP-violating effects, while firmly established, are numerically small throughout the flavor-conserving sector of the Standard Model – and why the observed matter-antimatter asymmetry of the universe cannot be explained by SM CP violation alone. [2427]

C.4 Neutrino Oscillations and the PMNS Matrix

Neutrino flavour oscillation, observed in solar, atmospheric, reactor, and accelerator beams, established that neutrinos are massive and that the weak-interaction flavour eigenstates ν_e,ν_μ,ν_τ are related to the mass eigenstates ν_1,ν_2,ν_3 by a unitary mixing matrix, the Pontecorvo–Maki–Nakagawa–Sakata (PMNS) matrix U, in direct analogy with the CKM matrix of the quark sector. In the standard three-flavour parametrization, U is written as a product of three rotations through the angles θ_{12}, θ_{13}, θ_{23}, one Dirac CP-violating phase δ_{CP}, and (if neutrinos are Majorana particles) two additional Majorana phases that do not affect oscillation probabilities. Oscillation probabilities depend on the mixing angles and on the two independent mass-squared splittings Δm^2_{21} and Δm^2_{32} (equivalently $\Delta m^2_{31} = \Delta m^2_{32} + \Delta m^2_{21}$), never on the absolute neutrino masses themselves, which remain a separate and still-open experimental question tied to the equally open question of whether the neutrino mass term is of Dirac or Majorana type. Values below are global-fit results as compiled in the Particle Data Group review of neutrino mixing.

Solar mixing angle θ_{12}: 33.41°.

$\theta_{12} = \arcsin\sqrt{\sin^2\theta_{12}}$, $\sin^2\theta_{12} = 0.303 \pm 0.011$ measured directly.

The solar mixing angle governs the ν_2–ν_1 admixture and, together with Δm^2_{21}, sets the survival probability $P(\nu_e \to \nu_e)$ observed in solar-neutrino experiments (Homestake, GALLEX/SAGE, Super-Kamiokande, SNO) and in the long-baseline reactor experiment KamLAND, which probes the same oscillation parameters in vacuum rather than through solar matter effects. Because θ_{12} is unambiguously bounded away from both 0 and 45°, it is often called the best-measured of the three angles once matter (MSW) effects inside the Sun

Table C.4: Neutrino Oscillations and the PMNS Matrix: summary of catalogued parameters.

Quantity	Symbol	Value
Solar mixing angle	θ_{12}	33.41°
Reactor mixing angle	θ_{13}	8.54°
Atmospheric mixing angle	θ_{23}	49.1°
Solar mass-squared splitting	Δm^2_{21}	$(7.53 \pm 0.18) \times 10^{-5}\,\mathrm{eV}^2$
Atmospheric mass-squared splitting	Δm^2_{32}	$(2.453 \pm 0.034) \times 10^{-3}\,\mathrm{eV}^2$
PMNS CP-violating phase (open)	δ_{CP}	197°
Absolute neutrino mass scale and Dirac/Majorana nature (open)	$m_i,\ \sum_i m_i$	0.12 eV (see text)

– confirmed decisively by SNO's neutral-current measurement – are accounted for. The value quoted is the current global-fit best fit to the combined solar and KamLAND data set; it is essentially identical for the normal and inverted mass orderings, since ordering sensitivity enters mainly through θ_{13} and Δm^2_{31}. [2427]

Reactor mixing angle θ_{13}: 8.54°.

$\theta_{13} = \arcsin\sqrt{\sin^2\theta_{13}}$, $\sin^2\theta_{13} = 0.0220 \pm 0.0007$ measured directly.

θ_{13} is the smallest of the three PMNS angles and was the last to be measured, established at high significance in 2012 by the short-baseline reactor experiments Daya Bay, RENO, and Double Chooz through the disappearance $P(\bar{\nu}_e \to \bar{\nu}_e) \simeq 1 - \sin^2(2\theta_{13})\sin^2(\Delta m^2_{31} L/4E)$ at $\mathcal{O}(1\,\mathrm{km})$ baselines. Although numerically small, a nonzero θ_{13} is what makes the CP-violating phase δ_{CP} observable in $\nu_\mu \to \nu_e$ appearance channels and what allows matter-effect sensitivity to the neutrino mass ordering in long-baseline accelerator experiments; its discovery reopened the experimental programme now pursued by T2K, NOvA, JUNO, DUNE, and Hyper-Kamiokande. The value is now known to about three percent precision and is essentially ordering-independent at this level. [2427]

Atmospheric mixing angle θ_{23}: 49.1°.

$\theta_{23} = \arcsin\sqrt{\sin^2\theta_{23}}$, $\sin^2\theta_{23} = 0.572 \pm 0.020$ measured (normal-ordering best fit).

The atmospheric angle controls ν_μ disappearance in atmospheric neutrinos (Super-Kamiokande, IceCube DeepCore) and in accelerator long-baseline experiments (T2K, NOvA, MINOS), and is close to, but not proven exactly equal to, the maximal-mixing value 45°. Current global fits place the best fit in the upper octant ($\theta_{23} > 45^\circ$, equivalently $\sin^2\theta_{23} > 0.5$) as quoted above, with a statistically nearly degenerate solution in the lower octant near $\sin^2\theta_{23} \approx 0.45$; resolving this octant ambiguity, which affects both δ_{CP} extraction and mass-ordering sensitivity, is a stated goal of NOvA, T2K-II, DUNE, and Hyper-Kamiokande. The value is only mildly ordering-dependent and is the least precisely known of the three angles, at roughly the four percent level. [2427]

Solar mass-squared splitting Δm^2_{21}: $(7.53 \pm 0.18) \times 10^{-5}\,\mathrm{eV}^2$.

$\Delta m^2_{21} \equiv m_2^2 - m_1^2 > 0$ (sign fixed by the MSW matter resonance in the Sun); measured directly.

Δm^2_{21} is the smaller of the two independent squared-mass splittings and, together with θ_{12}, fixes the long oscillation length responsible for solar-neutrino conversion and for the

KamLAND reactor-antineutrino oscillation pattern at $\mathcal{O}(100\,\mathrm{km})$ baselines. Unlike the atmospheric splitting, its sign is unambiguously determined (positive, i.e. $m_2 > m_1$) because solar matter effects (the MSW mechanism) break the vacuum degeneracy between the two sign choices; this is the only mass-ordering information oscillation experiments obtain without appeal to θ_{13}-driven matter effects on the atmospheric splitting. The quoted global-fit value combines solar data with the reactor measurement from KamLAND and is known to about 2 % precision. [2427]

Atmospheric mass-squared splitting Δm^2_{32}: $(2.453 \pm 0.034) \times 10^{-3}\,\mathrm{eV}^2$.

$\Delta m^2_{32} \equiv m_3^2 - m_2^2$; magnitude measured directly, value quoted for Normal Ordering ($m_3 > m_2 > m_1$); equivalently $\Delta m^2_{31} = \Delta m^2_{32} + \Delta m^2_{21} \approx 2.528 \times 10^{-3}\,\mathrm{eV}^2$.

Δm^2_{32} sets the fast oscillation frequency probed by atmospheric neutrinos and by accelerator long-baseline experiments and, together with θ_{23}, dominates ν_μ disappearance in that regime. Its magnitude is known to about 1.5 %, essentially identically whichever mass ordering is assumed, but its *sign* is not yet established at high statistical significance from oscillation data alone: this sign is precisely the neutrino mass-ordering (normal-vs-inverted) question. Combined global fits including reactor, atmospheric, and accelerator data presently favour Normal Ordering, the convention used above, at roughly the 2–3σ level; JUNO's precision reactor measurement, together with continued running of atmospheric (IceCube-Upgrade, Super-K/Hyper-K) and accelerator (DUNE, T2K-II/NOvA) experiments, is expected to resolve the ordering within the coming decade. [2427]

PMNS CP-violating phase (open) δ_{CP}: 197°.

measured via the CP-asymmetry between $\nu_\mu \to \nu_e$ and $\bar\nu_\mu \to \bar\nu_e$ appearance probabilities in long-baseline accelerator experiments, combined with the reactor constraint on θ_{13}.

The Dirac CP phase δ_{CP} is the PMNS analogue of the CKM phase and, if nonzero and different from π, allows leptonic CP violation in oscillations; it enters the appearance probabilities $P(\nu_\mu \to \nu_e)$ and $P(\bar\nu_\mu \to \bar\nu_e)$ measured by T2K and NOvA through terms proportional to $\sin\delta_{CP}\sin\theta_{13}$. This is genuinely one of the least well-determined parameters in the Standard Model sector: current global fits give a best fit near $\delta_{CP} \approx 197°$, closer to the CP-conserving value π (180°) than to maximal CP violation ($-\pi/2$, i.e. 270°); consistent with this, the CP-conserving values (0 or π) are disfavoured only around the

$$2\sigma$$

level, so this does not constitute a discovery of leptonic CP violation, and the allowed range widens substantially – covering most of the $[0, 2\pi)$ interval – once the mass-ordering degeneracy is marginalized over; T2K and NOvA show some mutual tension in δ_{CP} depending on the assumed ordering. Next-generation long-baseline experiments, principally DUNE and Hyper-Kamiokande, are designed to reach

$$5\sigma$$

sensitivity to CP violation over a substantial fraction of the δ_{CP} range; until such a measurement, the true value – and even the sign of CP violation – remains an open experimental question. [2427]

Absolute neutrino mass scale and Dirac/Majorana nature (open) m_i, $\sum_i m_i$: $\sum_i m_i < 0.12\,\mathrm{eV}$ (95% CL, cosmological bound, model-dependent).

not determined by oscillations, which fix only Δm^2_{21} and Δm^2_{32}; bounded independently by β-decay kinematics (KATRIN), cosmology (CMB+BAO), and $0\nu\beta\beta$ searches.

Oscillation phases depend only on $\Delta m^2_{ij} L/E$, so oscillation data are completely insensitive to the overall additive mass scale: they fix the two independent squared-mass splittings but

not any individual m_i, nor even the lightest mass m_{lightest}, which could in principle be zero. Three independent, complementary methods bound the absolute scale: direct kinematic spectroscopy of the tritium β-decay endpoint (KATRIN gives $m_\beta \lesssim 0.45\,\text{eV}$ at 90 % CL from its early data run, with a design sensitivity down to about 0.2 eV); cosmological structure formation and CMB lensing, which constrain the sum $\sum_i m_i$ under ΛCDM assumptions (Planck 2018 combined with baryon acoustic oscillation data gives the bound quoted above); and searches for neutrinoless double beta decay ($0\nu\beta\beta$), which additionally probe whether the neutrino mass term is of Majorana type. Whether neutrinos are Dirac or Majorana fermions is a genuinely open question with no experimental resolution to date: $0\nu\beta\beta$ would violate lepton number by two units and can occur only if neutrinos are massive Majorana particles, but current searches (KamLAND-Zen, GERDA, and successors) report no signal, placing only an upper bound on the effective Majorana mass parameter $m_{\beta\beta}$ at the few-tens-of-meV level (with sizeable nuclear-matrix-element uncertainty); a positive detection remains one of the primary open experimental goals of the field, bearing directly on the origin of neutrino mass (e.g. seesaw-type mechanisms) and on leptogenesis as a possible explanation of the cosmic baryon asymmetry. [2348, 2427]

C.5 Lambda-CDM Cosmological Parameters

The base ΛCDM model describes the observed universe with six free parameters fit jointly to the cosmic microwave background (and, for some derived quantities, external distance-ladder and large-scale-structure data): the physical baryon and cold-dark-matter densities $\Omega_b h^2$ and $\Omega_c h^2$, the amplitude and tilt of the primordial scalar power spectrum A_s and n_s, the reionization optical depth τ, and the acoustic angular scale (from which H_0 and Ω_Λ are derived under the flatness assumption). All values below are taken from the Planck 2018 final data release, specifically the TT,TE,EE+lowE+lensing combination reported in Planck Collaboration VI, unless otherwise noted. The sector also includes the baryon-to-photon ratio η, a derived cross-check quantity linking the CMB-era baryon density to the independent, much earlier Big Bang nucleosynthesis epoch, and it flags explicitly the Hubble tension, the unresolved several-sigma discrepancy between CMB-inferred and locally measured values of H_0 that remains one of the most actively debated open problems in the field.

Table C.5: Lambda-CDM Cosmological Parameters: summary of catalogued parameters.

Quantity	Symbol	Value
Baryon density parameter	$\Omega_b h^2$	$0.022\,37 \pm 0.000\,15$
Cold dark matter density parameter	$\Omega_c h^2$	0.1200 ± 0.0012
Dark energy density parameter	Ω_Λ	0.6847 ± 0.0073
Hubble constant (open)	H_0	$(67.36 \pm 0.54)\,\text{km}/(\text{s}\,\text{Mpc})$ (see text)
Scalar spectral index	n_s	0.9649 ± 0.0042
Amplitude of the primordial scalar power spectrum	A_s	2.101×10^{-9} (see text)
Optical depth to reionization	τ	0.0544 ± 0.0073
Baryon-to-photon ratio	η	$(6.12 \pm 0.04) \times 10^{-10}$

Baryon density parameter $\Omega_b h^2$: $0.022\,37 \pm 0.000\,15$.

$\Omega_b h^2 \equiv (\rho_b/\rho_{\mathrm{crit},0})\, h^2$; measured directly from the relative heights of the odd/even acoustic peaks in the CMB temperature and polarization power spectra.

The physical baryon density $\Omega_b h^2$ fixes the ratio of baryonic (ordinary) matter to the critical density, scaled by h^2 to remove the degeneracy with H_0 that would otherwise plague a bare Ω_b. It is pinned down by the CMB because baryons load the photon-baryon plasma before recombination, compressing odd acoustic peaks and suppressing even ones relative to a baryon-free plasma; the Planck 2018 TT,TE,EE+lowE+lensing fit gives $\Omega_b h^2 = 0.022\,37 \pm 0.000\,15$, in excellent agreement with the independent value inferred from primordial deuterium abundances via Big Bang nucleosynthesis. This parameter is one of the six free numbers of the base ΛCDM model and is measured to better than one percent precision. [2348]

Cold dark matter density parameter $\Omega_c h^2$: 0.1200 ± 0.0012.

$\Omega_c h^2 \equiv (\rho_c/\rho_{\mathrm{crit},0})\, h^2$; measured directly, chiefly from the amplitude of the third acoustic peak and the overall peak-height envelope of the CMB power spectrum.

$\Omega_c h^2$ is the physical density of cold (pressureless, collisionless) dark matter, the dominant matter component of the universe. In base ΛCDM all non-baryonic matter is assumed cold, so this parameter is equivalently quoted as the total non-baryonic dark-matter density $\Omega_{\mathrm{dm}} h^2$; some analyses distinguish Ω_c (cold) from a possible hot/warm component, but Planck finds no evidence requiring one. The Planck 2018 fit gives $\Omega_c h^2 = 0.1200 \pm 0.0012$, roughly 5.4 times the baryon density, driven observationally by the enhanced gravitational driving of the third and higher acoustic peaks and by the low-ℓ lensing potential power spectrum. Combined with $\Omega_b h^2$ this yields a present-day total matter density $\Omega_m = 0.3153 \pm 0.0073$. [2348]

Dark energy density parameter Ω_Λ: 0.6847 ± 0.0073.

$\Omega_\Lambda \equiv \rho_\Lambda/\rho_{\mathrm{crit},0}$; in a spatially flat universe,$\Omega_\Lambda = 1 - \Omega_m - \Omega_r$, derived from the CMB angular-diameter-distance constraint combined with the measured matter density.

Ω_Λ is the fractional contribution of the cosmological constant (or, phenomenologically, dark energy with equation of state consistent with $w = -1$) to the total energy density today. It is not measured by a single spectral feature but emerges from the combination of the CMB's precise measurement of the angular acoustic scale θ_* (which fixes the comoving distance to last scattering) together with the flatness prior and the measured matter density, and is corroborated independently by baryon acoustic oscillation (BAO) distance measurements and Type Ia supernova Hubble-diagram fits. Planck 2018 (TT,TE,EE+lowE+lensing) gives $\Omega_\Lambda = 0.6847 \pm 0.0073$, confirming that dark energy dominates the present energy budget by roughly a factor of two over matter. Its microphysical origin — vacuum energy, a slowly rolling field, or a modification of gravity — remains unidentified, though its phenomenological value is measured to sub-percent precision. [2348]

Hubble constant (open) H_0: $(67.36 \pm 0.54)\,\mathrm{km/(s\,Mpc)}$ (Planck 2018, TT,TE,EE+lowE+lensing, CMB-inferred within base ΛCDM); local distance-ladder determinations give values near $(73.0 \pm 1.0)\,\mathrm{km/(s\,Mpc)}$ [2407], in tension with the Planck value at the 4–5σ level.

$H_0 \equiv \dot{a}/a\big|_{t_0}$, the present-day cosmic expansion rate; measured either model-dependently from the CMB acoustic scale within an assumed ΛCDM background, or directly and locally from the distance-redshift relation of the cosmic distance ladder.

H_0 sets the present expansion rate of the universe and, through $1/H_0$, the characteristic age and size scale of the observable universe. Two broad classes of measurement exist: CMB/early-universe methods, which infer H_0 indirectly by fitting the full ΛCDM model (including $\Omega_b h^2$, $\Omega_c h^2$, and the acoustic scale θ_*) to the Planck power spectra and obtain

$H_0 = (67.36 \pm 0.54)\,\mathrm{km/(s\,Mpc)}$; and late-universe, model-independent distance-ladder methods (Cepheid-calibrated Type Ia supernovae, and independently the tip-of-the-red-giant-branch and Mira-variable calibrations), which cluster around $73\,\mathrm{km/(s\,Mpc)}$. The discrepancy between these two families of measurement — the *Hubble tension* — now exceeds 4–5σ in the highest-precision comparisons and has resisted resolution by known systematics on either side; it is one of the most actively studied open problems in cosmology, with proposed resolutions ranging from unrecognized systematics to genuine early- or late-time extensions of ΛCDM (e.g. early dark energy, modified recombination physics), none yet confirmed. This entry reports the Planck (CMB) value as the primary ΛCDM-model parameter but flags the tension explicitly rather than presenting a single settled number. [2348, 2407]

Scalar spectral index n_s: 0.9649 ± 0.0042.

power-law index of the primordial scalar (curvature) power spectrum,$\mathcal{P}_\mathcal{R}(k) \propto k^{n_s-1}$; measured directly from the scale dependence (tilt) of the CMB TT, TE, and EE power spectra across the observed multipole range.

n_s characterizes the near-scale-invariant tilt of the spectrum of primordial curvature perturbations laid down (in the standard picture) during an early inflationary epoch: $n_s = 1$ would correspond to exact scale invariance (Harrison-Zel'dovich), while the measured value slightly below unity indicates a small red tilt. Planck 2018 measures $n_s = 0.9649 \pm 0.0042$ from the shape of the CMB angular power spectrum, excluding exact scale invariance at high significance ($n_s = 1$ is disfavored by more than 8σ). This is one of the most precisely determined ΛCDM parameters and is a key discriminant among inflationary models, though the underlying inflaton dynamics (if any) remain outside the evidence-based scope of this treatise. [2348]

Amplitude of the primordial scalar power spectrum A_s: 2.101×10^{-9} at pivot scale $k_0 = 0.05/\mathrm{Mpc}$, conventionally reported as $\ln(10^{10} A_s) = 3.044 \pm 0.014$; the corresponding late-time, nonlinearly evolved amplitude is $\sigma_8 = 0.8111 \pm 0.0060$.

$A_s \equiv \mathcal{P}_\mathcal{R}(k_0)$, the amplitude of the dimensionless primordial curvature power spectrum at the pivot scale; measured directly from the overall (multipole-averaged) amplitude of the CMB temperature and polarization power spectra, degenerate with τ through the combination $A_s e^{-2\tau}$.

A_s fixes the overall normalization of the primordial curvature perturbations from which all later structure grows; because CMB anisotropies are damped by rescattering after reionization, the direct CMB observable is the combination $A_s e^{-2\tau}$, so an independent handle on τ (from large-scale polarization) is needed to break the degeneracy. Planck 2018 reports $\ln(10^{10} A_s) = 3.044 \pm 0.014$, i.e. $A_s \approx 2.101 \times 10^{-9}$. For late-time, nonlinear structure the amplitude is more commonly quoted as σ_8, the rms mass-density fluctuation in spheres of radius 8 Mpc/h obtained by evolving A_s forward through the transfer function and linear growth factor; Planck finds $\sigma_8 = 0.8111 \pm 0.0060$. This treatise uses A_s (with $\ln(10^{10} A_s)$) as the primary ΛCDM parameter and cites σ_8 as its late-time, structure-formation-friendly equivalent. [2348]

Optical depth to reionization τ: 0.0544 ± 0.0073.

$\tau \equiv \int_0^{t_0} n_e(t)\,\sigma_T\,c$, dt, the Thomson-scattering optical depth integrated from the observer back through the reionized intergalactic medium; measured directly from the suppression and large-angle bump of the CMB E-mode polarization power spectrum at low multipoles.

τ measures the integrated column of free electrons that CMB photons scatter from after the universe reionizes (following the formation of the first stars and galaxies), which damps the temperature and E-mode power spectra on scales inside the horizon at reionization and

generates a distinctive large-angle polarization bump. Planck 2018's low-ℓ EE likelihood gives $\tau = 0.0544 \pm 0.0073$, corresponding to an inferred midpoint reionization redshift around $z_{\rm re} \sim 7$–9 under simple reionization histories, consistent with the epoch inferred from quasar Gunn-Peterson trough studies. Because τ is degenerate with A_s in the small-scale CMB spectra (only the combination $A_s e^{-2\tau}$ is well constrained there), its independent determination from large-angle polarization is essential to breaking that degeneracy and obtaining A_s itself. [2348]

Baryon-to-photon ratio η: $(6.12 \pm 0.04) \times 10^{-10}$.

$\eta \equiv n_b/n_\gamma \approx 273.9\,\Omega_b h^2 \times 10^{-10}$, derived by combining the CMB-measured baryon density $\Omega_b h^2$ with the photon number density fixed by the measured CMB blackbody temperature.

The baryon-to-photon ratio η (equivalently often quoted as $\eta_{10} \equiv 10^{10}\eta$) measures the tiny excess of baryons over antibaryons that survived annihilation in the early universe, expressed relative to the number density of CMB photons; it is the single parameter controlling the light-element yields (D, ^{3}He, ^{4}He, ^{7}Li) of Big Bang nucleosynthesis. It is not measured independently in this catalog but is derived as a cross-check quantity from the CMB-fit baryon density $\Omega_b h^2 = 0.022\,37 \pm 0.000\,15$ combined with the precisely known photon number density, giving $\eta = (6.12 \pm 0.04) \times 10^{-10}$. The remarkable agreement between this CMB-derived value and the value independently required to match observed primordial deuterium and helium abundances is one of the strongest quantitative successes of standard Big Bang cosmology, cross-validating two entirely different physical epochs (nucleosynthesis at $t \sim 1$ s–200 s versus recombination at $t \sim 380\,000$ yr). [2348]

C.6 Gravitational Coupling, Planck Units, and the Hierarchy Problem

This sector collects the scales at which the treatise's two most disparate confirmed frameworks – quantum mechanics, governed by $\hbar$, and general relativity, governed by G and c – become comparably important, together with the dimensionless measures of just how far current particle-physics scales sit from that regime. The Planck length, mass, and time are not free parameters but are fixed by dimensional analysis alone: the Buckingham π theorem (Theorem 2.3), proved in Section 2.2.1 from the rank-nullity theorem applied to the matrix of dimensional exponents, shows that $\{\hbar, G, c\}$ admit no dimensionless combination among themselves, which is precisely the condition that makes the length, mass, and time built from them unique up to an unresolvable $O(1)$ prefactor. The gravitational coupling constant α_G then expresses, in the same dimensionless language as the fine-structure constant, how weak gravity is between two protons compared with the other measured interactions. The final three entries are mass ratios between confirmed particle-physics scales and the Planck mass; their extreme smallness is directly observed and uncontroversial, but the absence of any experimentally confirmed dynamical principle that explains why the scales are so separated – especially for the Higgs vacuum expectation value, whose radiative stability is not protected by any confirmed symmetry – constitutes the hierarchy (naturalness) problem, recorded here strictly as an open observational puzzle per this treatise's evidence-only scope.

Planck length ℓ_P: $(1.616\,255 \pm 0.000\,018) \times 10^{-35}$ m.

$\ell_P = \sqrt{\dfrac{\hbar G}{c^3}}$.

The Planck length is the unique length scale constructible from the reduced Planck constant $\hbar$, the Newtonian gravitational constant G, and the speed of light c – the three constants that respectively govern quantum mechanics, gravitation, and relativity. Its

Table C.6: Gravitational Coupling, Planck Units, and the Hierarchy Problem: summary of catalogued parameters.

Quantity	Symbol	Value
Planck length	ℓ_P	$(1.616\,255 \pm 0.000\,018) \times 10^{-35}\,\mathrm{m}$
Planck mass	M_P	$(2.176\,434 \pm 0.000\,024) \times 10^{-8}\,\mathrm{kg}$
Planck time	t_P	$(5.391\,247 \pm 0.000\,060) \times 10^{-44}\,\mathrm{s}$
Gravitational coupling constant	α_G	5.91×10^{-39}
Electron-to-Planck mass ratio (open)	m_e/M_P	4.19×10^{-23}
Proton-to-Planck mass ratio (open)	m_p/M_P	7.69×10^{-20}
Higgs vacuum expectation value to Planck mass ratio (open)	v/M_P	2.01×10^{-17}

dimensional exponents are fixed, not merely suggested, by the Buckingham π theorem (Theorem 2.3): among the three quantities $\{\hbar, G, c\}$, whose dimension exponents span the full $r = 3$-dimensional space of mechanical base dimensions $\{\mathsf{M}, \mathsf{L}, \mathsf{T}\}$, the theorem's exponent matrix M has trivial kernel ($n - r = 3 - 3 = 0$), so no dimensionless combination of $\hbar$, G, c alone exists; augmenting M with the target dimension L then yields a 3×3 linear system in the exponents of $\hbar^a G^b c^d$ with a unique solution $a = \frac{1}{2}$, $b = \frac{1}{2}$, $d = -\frac{3}{2}$, i.e. $\ell_P = \sqrt{\hbar G/c^3}$. The same rank-nullity argument that proves the theorem shows explicitly what it cannot deliver: any dimensionless numerical prefactor multiplying ℓ_P is invisible to dimensional analysis and is fixed here by convention to 1. Numerically, using the CODATA 2018 values of $\hbar$, G, c tabulated in Appendix B, $\ell_P \approx 1.6 \times 10^{-35}\,\mathrm{m}$, some fifteen orders of magnitude below the smallest length probed directly by any collider experiment to date. [2392]

Planck mass M_P: $(2.176\,434 \pm 0.000\,024) \times 10^{-8}\,\mathrm{kg}$.

$$M_P = \sqrt{\frac{\hbar c}{G}}.$$

By the same application of Theorem 2.3 used for the Planck length – the exponent system for a quantity of dimension M built from $\{\hbar, G, c\}$ has a unique solution because the three quantities admit no dimensionless combination among themselves – the Planck mass is fixed up to the same unresolvable $O(1)$ prefactor as $M_P = \sqrt{\hbar c/G}$. Equivalently $M_P = \hbar/(\ell_P c)$. Its numerical value, about 22 µg, is unremarkable on laboratory scales – comparable to a fine dust grain – which is itself notable: the mass scale conventionally associated with quantum-gravitational effects becoming significant is macroscopic in mass even though the associated length (ℓ_P) and energy ($M_P c^2 \approx 1.22 \times 10^{19}\,\mathrm{GeV}$) are wildly inaccessible to any existing or foreseeable particle experiment. [2392]

Planck time t_P: $(5.391\,247 \pm 0.000\,060) \times 10^{-44}\,\mathrm{s}$.

$$t_P = \sqrt{\frac{\hbar G}{c^5}} = \ell_P/c.$$

Completing the triad, the Planck time is the unique time scale formed from $\{\hbar, G, c\}$, fixed by exactly the same rank-3 exponent argument furnished by Theorem 2.3: $t_P = \sqrt{\hbar G/c^5} = \ell_P/c$, again with a convention-fixed prefactor of 1 that dimensional analysis alone cannot supply. Numerically $t_P \approx 5.4 \times 10^{-44}\,\mathrm{s}$, conventionally quoted as the timescale below which no confirmed physical theory – classical general relativity least of all, since it

treats spacetime as a smooth continuum – is expected to remain a reliable description; no experiment has probed timescales anywhere near this regime. [2392]

Gravitational coupling constant α_G: 5.91×10^{-39}.

$\alpha_G \equiv \dfrac{Gm_p^2}{\hbar c}$.

In direct structural analogy with the electromagnetic fine-structure constant $\alpha = e^2/(4\pi\epsilon_0\hbar c)$, which measures the strength of the electromagnetic coupling between two elementary charges in dimensionless form, the gravitational coupling constant α_G measures the strength of the Newtonian gravitational attraction between two protons in the same dimensionless units. It is built entirely from constants already tabulated in Appendix B: G, the proton mass m_p, $\hbar$, and c. Its value, $\alpha_G \approx 5.9 \times 10^{-39}$, is smaller than α by roughly thirty-six orders of magnitude, making α_G the quantitative, dimensionless measure of the empirical fact that gravity is by far the weakest of the interactions accessible to laboratory and particle-physics measurement. [2392]

Electron-to-Planck mass ratio (open) m_e/M_P: 4.19×10^{-23}.

m_e/M_P, m_e measured, $M_P = \sqrt{\hbar c/G}$.

The electron mass, measured directly, is smaller than the Planck mass by a factor of about 4.2×10^{-23}. For a chiral fermion such as the electron, standard effective-field-theory power counting makes this smallness *technically natural*: radiative corrections to a fermion mass are proportional to the mass itself (protected by the approximate chiral symmetry restored as $m_e \to 0$) rather than being pulled up quadratically toward the ultraviolet cutoff, so no fine-tuning is strictly required to keep m_e small under quantum corrections. The ratio is nonetheless recorded here as part of the same hierarchy-of-scales puzzle: no confirmed dynamical principle explains why the Yukawa coupling that ultimately sets m_e (via the Higgs mechanism, Part XI) takes the minuscule value it does rather than some other value of order unity; that the coupling itself is unexplained – not that it is radiatively unstable – is the open question here. [2392]

Proton-to-Planck mass ratio (open) m_p/M_P: 7.69×10^{-20}.

m_p/M_P, m_p measured, $M_P = \sqrt{\hbar c/G}$.

The proton mass is smaller than the Planck mass by a factor of about 7.7×10^{-20}, some three orders of magnitude larger a ratio than the electron's because most of the proton's mass arises not from the Higgs mechanism but from the strong-interaction confinement scale Λ_{QCD} (Part XI), with the current-quark Higgs-generated masses contributing only a small fraction. The ratio is nonetheless part of the same observed hierarchy: no confirmed dynamical mechanism explains why Λ_{QCD}, and hence the mass of ordinary baryonic matter, sits so far below the Planck scale rather than comparably close to it, even though the running of the strong coupling (asymptotic freedom, itself experimentally confirmed) makes the smallness only logarithmically – not quadratically – sensitive to the ultraviolet completion, unlike the scalar case below. [2392]

Higgs vacuum expectation value to Planck mass ratio (open) v/M_P: 2.01×10^{-17}.

$v/(M_Pc^2)$, $v \approx 246\,\mathrm{GeV}$(measured), $M_Pc^2 \approx 1.22 \times 10^{19}\,\mathrm{GeV}$.

The Higgs vacuum expectation value $v \approx 246\,\mathrm{GeV}$, which sets the electroweak scale and, through it, the masses of the W and Z bosons and (via Yukawa couplings) the charged fermions, is smaller than the Planck energy $M_Pc^2 \approx 1.22 \times 10^{19}\,\mathrm{GeV}$ by a factor of about 2.0×10^{-17}. This is the ratio at the heart of the hierarchy (naturalness) problem: unlike a fermion mass, the Higgs boson is a fundamental scalar, and nothing within the confirmed Standard Model protects its mass-squared parameter from receiving radiative corrections

that scale quadratically with whatever ultraviolet cutoff the theory is embedded in – if that cutoff is the Planck scale, the corrections are technically expected to be some seventeen orders of magnitude larger than the observed electroweak scale itself, requiring an extreme and unexplained cancellation to keep v as small as it is measured to be. No experimentally confirmed symmetry or dynamical mechanism protecting the electroweak scale from such corrections has been observed to date (in particular, no signal of any such protecting structure has been seen at the LHC), so the smallness of v/M_P remains, strictly on the evidence, an open naturalness puzzle rather than a solved problem. [2392, 2427]

Appendix D

The Dimensionless Constants

Appendix C tabulates the free parameters of the Standard Model and of ΛCDM with their measured values. This appendix asks a different question about the same numbers: *which* of them mean anything independently of us. The answer is sharp — only the dimensionless ones — and it has a formal statement, a complete construction, and a consequence that reaches beyond physics into what could in principle be compared with a civilization that shares none of our history.

D.1 Why only dimensionless numbers carry content

The speed of light in vacuum is exactly 299 792 458 m/s. That statement contains no physics whatever. Since 1983 the metre has been *defined* as the distance light travels in 1/299792458 of a second (Chapter 2), so the number records a decision of the Conférence Générale des Poids et Mesures, not a property of light. Had the committee chosen differently, c would read differently and not one experimental outcome would change.

This is not a peculiarity of the SI. Any dimensional constant carries the same defect in weaker form: its numerical value is a statement about the relation between the quantity and an arbitrarily chosen yardstick.

Remark D.1 (Convention and content). The distinction is the one drawn in Section 1.3.2 between convention and content, and it is why that section insists the conventions be declared. A theory's empirical content is what survives every allowed change of convention. For units, the allowed changes are rescalings of the base quantities, and what survives them is exactly the dimensionless combinations.

D.1.1 The dimension map

Make the claim precise. Let $q_1, \dots, q_n$ be physical quantities and let there be k base dimensions — in mechanics $\mathsf{M}, \mathsf{L}, \mathsf{T}$, so $k = 3$; in the SI, seven. Each q_a has a dimension

$$[q_a] = \mathsf{X}_1^{d_{1a}} \mathsf{X}_2^{d_{2a}} \cdots \mathsf{X}_k^{d_{ka}} \,, \tag{D.1}$$

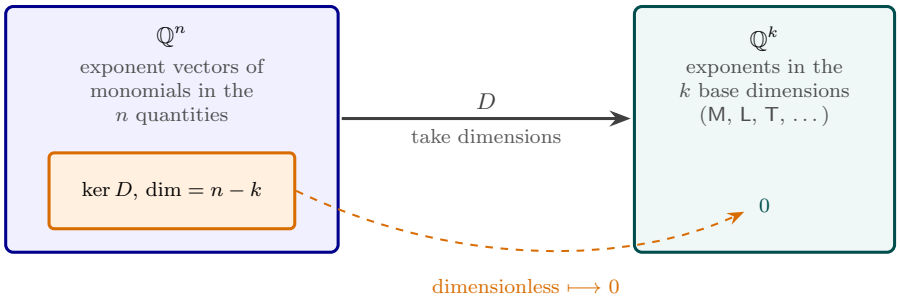
with rational exponents d_{ia}. A monomial in the quantities,

$$\Pi = q_1^{p_1} q_2^{p_2} \cdots q_n^{p_n} \,, \qquad p \in \mathbb{Q}^n \,, \tag{D.2}$$

then has dimension $\mathsf{X}_i^{\sum_a d_{ia} p_a}$. Writing D for the $k \times n$ matrix of exponents d_{ia}, the monomial is dimensionless precisely when

$$Dp = 0 \,, \tag{D.3}$$

so the dimensionless combinations are the kernel of a linear map. Taking logarithms has turned a multiplicative question into linear algebra, and Figure D.1 is the whole content of the subject in one picture.



A change of units rescales each base dimension: it acts on $\mathbb{Q}^n$ by a shift that is invisible on $\ker D$. Only the elements of $\ker D$ therefore carry information independent of the choice of units — which is why the numbers a distant civilisation could compare with ours are exactly the ones living in that kernel.

Figure D.1: Dimensional analysis as linear algebra. Monomials in the n quantities are represented by their exponent vectors in $\mathbb{Q}^n$; the dimension map D records the exponents of the k base dimensions. The dimensionless combinations are exactly $\ker D$. A change of units acts on $\mathbb{Q}^n$ in a way that leaves the kernel pointwise fixed, which is the precise sense in which only dimensionless numbers are convention-free.

Theorem D.2 (Buckingham's Π theorem)**.** *Let n quantities involve k base dimensions, and let r be the rank of the dimension matrix D. Then the dimensionless monomials form a $\mathbb{Q}$-vector space of dimension $n - r$, and any physical law relating the n quantities can be written as a relation among $n - r$ independent dimensionless groups [413].*

Rests on Equation (D.3).

Remark D.3 (Two footnotes on the theorem's name and its statement). The attribution is conventional rather than historical: the general statement was published by Vaschy in 1892 and, apparently independently, by Federman and Riabouchinsky in 1911, and Buckingham himself later ceded priority to Riabouchinsky. Second, the theorem is often quoted with r replaced by the *number* of base dimensions. That is the imprecise version, and it is not Buckingham's: he defines the subtracted quantity as the size of a maximal independent set, which is the rank. Where the dimensional matrix is rank-deficient the two differ, and the careless form undercounts the groups. Bridgman's monograph [485] is the standard early treatment.

Proof. *Derives Theorem D.2.*

That the dimensionless monomials are $\ker D$ is Equation (D.3), and the rank–nullity theorem (Chapter 5) gives $\dim \ker D = n - r$ at once. Choose a basis $\Pi_1, \dots, \Pi_{n-r}$ of the kernel.

For the second statement, let $f(q_1, \dots, q_n) = 0$ be a law. A change of units multiplies each q_a by $\prod_i \lambda_i^{d_{ia}}$ for positive scale factors λ_i. If the law is to hold in every system of units — which is what it means for it to be a law about nature rather than about a bureau of standards — then f must be invariant under the whole k-parameter group of such rescalings. The orbits of that group are the level sets of the map $q \longmapsto (\Pi_1, \dots, \Pi_{n-r})$, since two configurations have the same Π's exactly when a rescaling carries one to the other. A function constant on orbits factors through the orbit space, so f is a function of the Π's alone. □

Example D.4 (The theorem doing work)**.** A pendulum's period T can depend on its length ℓ, its mass m and g. Here $n = 4$ and the dimension matrix in $(\mathsf{M}, \mathsf{L}, \mathsf{T})$ has rank 3, so there is $4 - 3 = 1$ independent group. It is $\Pi = T\sqrt{g/\ell}$, whence $T = C\sqrt{\ell/g}$ with C a pure number — and the mass cannot appear. Two lines of linear algebra have produced the form of the answer, the independence of the period from the mass, and the impossibility of any

other scaling, without solving an equation of motion (Chapter 28 supplies $C = 2\pi$ for small amplitude).

D.1.2 How many dimensionful constants are fundamental?

If dimensional constants are conventions, why keep any? The question was put sharply in a three-way exchange between Duff, Okun and Veneziano, who argued respectively for zero, three and two fundamental dimensionful constants [2026]. The disagreement is not empirical: all three agree on every measurable prediction. It is about what one calls fundamental, and it resolves once the roles are separated.

One caveat on that source: Veneziano's count of two is explicitly conditional on superstring theory, which this treatise excludes from scope by editorial rule 1. His position is reported here as a position held in the literature, not as a result. And the paper cannot be cited in support of any particular count — its three authors never converged, which is precisely its value: it establishes that the count is not fixed by physics.

Constants such as c, $\hbar$ and G do two jobs. As *conversion factors* they relate quantities we had historically thought distinct — c converts time to length, $\hbar$ converts frequency to energy, G converts mass to length — and in that role they are pure convention, and can be set to 1, which is what natural units do. As *markers of a regime* they say where new physics appears: relativity when speeds approach c, quantum mechanics when actions approach $\hbar$, and strong gravity near the Planck scale built from all three. That second role is physical, but it is a statement about *ratios* — v/c, $S/\hbar$ — and ratios are dimensionless. Nothing is lost by the reduction.

D.2 The systematic construction

Theorem D.2 says how many independent dimensionless numbers exist; this section constructs them for the quantities physics actually has.

D.2.1 Planck units and the mass ratios

From $\hbar$, c and G alone — three quantities, three dimensions, rank three — the kernel is trivial: there is *no* dimensionless combination of the three. That is precisely why they can be used to build a unique system of units,

$$m_{\mathrm{P}} = \sqrt{\frac{\hbar c}{G}}\,, \qquad \ell_{\mathrm{P}} = \sqrt{\frac{\hbar G}{c^3}}\,, \qquad t_{\mathrm{P}} = \sqrt{\frac{\hbar G}{c^5}}\,, \tag{D.4}$$

and it is why the Planck scale is the natural yardstick against which every other scale becomes a pure number. Adjoin any fourth quantity and a dimensionless group appears immediately: adjoining a mass m gives the single group m/m_{P}.

Every particle mass therefore contributes one dimensionless number. Two families of ratio are worth separating, because they are physically very different:

1. **Ratios to the Planck mass**, such as

$$\frac{m_e}{m_{\mathrm{P}}} \approx 4.19 \times 10^{-23}\,, \qquad \frac{m_p}{m_{\mathrm{P}}} \approx 7.69 \times 10^{-20}\,, \tag{D.5}$$

which measure how feebly gravity acts on elementary particles. Their extreme smallness is the hierarchy problem (Appendix C.6). Equivalently one writes the *gravitational coupling*

$$\alpha_G = \frac{G m_e^2}{\hbar c} = \left(\frac{m_e}{m_{\mathrm{P}}}\right)^2 \approx 1.75 \times 10^{-45}\,, \tag{D.6}$$

the exact gravitational analogue of Equation (D.8) below. The ratio $\alpha/\alpha_G \approx 4 \times 10^{42}$ is the number Dirac found suspicious enough to build a cosmology on [812]; that cosmology, which required G to vary with time, is excluded by the bounds of Appendix D.3.4.

2. **Ratios among the particles themselves**, which involve no gravity at all:

$$\frac{m_p}{m_e} \approx 1836.15\,, \qquad \frac{m_\mu}{m_e} \approx 206.768\,, \qquad \frac{m_\tau}{m_\mu} \approx 16.817\,. \tag{D.7}$$

These are the numbers that fix chemistry, the size of atoms relative to nuclei, and the pattern of the three generations. In the Standard Model they are not independent inputs but consequences of the Yukawa couplings, themselves dimensionless (Appendix C.1).

D.2.2 The couplings

Definition D.5 (Fine-structure constant)**.**

$$\alpha := \frac{e^2}{4\pi\varepsilon_0 \hbar c}\,. \tag{D.8}$$

Every dimensional quantity in Equation (D.8) cancels: α is a pure number, the same in SI, Gaussian or natural units, and the same for any observer anywhere. It measures the strength of the electromagnetic interaction — more precisely, it is the square of the electric charge in units where $\hbar = c = 1$, and it is the expansion parameter of the perturbation series of quantum electrodynamics (Chapter 100). Its smallness is why that series works.

The recommended value is

$$\alpha^{-1} = 137.035\,999\,177 \pm 0.000\,000\,021\,, \tag{D.9}$$

a relative uncertainty of 1.6×10^{-10} [2431].

Remark D.6 (The last digits are less settled than they look). Equation (D.9) is not a straight statistical combination. The two most precise determinations — caesium recoil at Berkeley, $\alpha^{-1} = 137.035\,999\,046 \pm 0.000\,000\,027$ [2316], and rubidium recoil at the Kastler–Brossel laboratory, $\alpha^{-1} = 137.035\,999\,206 \pm 0.000\,000\,011$ [2363] — disagree with each other by more than 5σ, and the disagreement is unresolved. The recommended uncertainty in Equation (D.9) has been inflated by an expansion factor to cover it. Quoting ten significant figures without that caveat would misrepresent the state of the measurement, which is precisely the kind of thing the availability grading of Chapter 1 exists to prevent.

The other interactions contribute their own couplings, α_s for the strong force and the electroweak pair g, g'; unlike α they are not small at all accessible energies, which is why perturbation theory fails for the strong interaction at low energy (Chapter 102).

D.2.3 The full catalogue

Collecting: the dimensionless numbers of established physics are the gauge couplings, the Yukawa couplings (equivalently the mass ratios), the CKM and PMNS mixing angles and phases, the strong CP angle θ_{QCD}, the Higgs self-coupling and the ratio of the Higgs vacuum expectation value to the Planck mass, together with the cosmological parameters — the density fractions Ω_i, the baryon-to-photon ratio η, the scalar spectral index n_s. Their measured values occupy Appendix C; the count is about twenty-six for the Standard Model plus a handful for cosmology.

Remark D.7 (What a "theory of everything" would have to do). The catalogue is the honest statement of what physics does not explain. Feynman's verdict on α — "a magic number that comes to us with no understanding by man" [1752] — applies to every entry. A theory that derived even one of them from something more basic would be a decisive advance, and none has. This is worth stating plainly because it is often obscured: the Standard Model is not incomplete in the sense of failing a measurement, but in the sense that roughly two dozen numbers must be handed to it from outside. Programmes that claim to derive them are assessed on that claim and no other, and Chapter 135 records that none currently succeeds.

D.3 Universality: the numbers a stranger could check

D.3.1 Why α and not a metre

Suppose we wished to establish, with a civilization that shares nothing of our history, that we had found the same physics. Almost everything we might send fails. A length in metres is meaningless without the metre. A temperature in kelvin, a mass in kilograms, an energy in joules: all encode the Earth, the water molecule, a platinum cylinder, a French committee. Worse, much of what we would think to describe is *local*: the luminosity of our star, the surface gravity of our planet, the composition of our atmosphere, the strength of our magnetic field. A recipient orbiting a different star has different values for every one of them, and no disagreement about those would indicate any disagreement about physics.

What survives is $\ker D$. The dimensionless constants are the same for them as for us — not by convention but because they are what remains when all conventions are quotiented away. If their measurement of α disagreed with ours, that would be a discovery of the first order; if their measurement of the local gravitational acceleration disagreed with ours, that would be a weather report.

D.3.2 The message that needs no dictionary

α can be communicated with no shared units, no shared numerals and no shared language, because it is a *ratio of counts*, and counting is not a convention. Figure D.2 draws it: one stroke, a divider, one hundred and thirty-seven strokes.

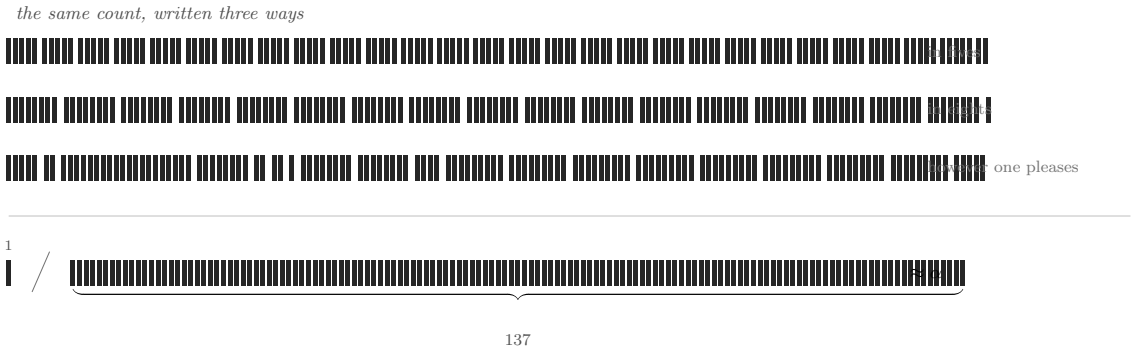


Figure D.2: The fine-structure constant drawn without conventions. Above: the same count of 137 grouped in fives, in eights, and irregularly — the grouping is notation and varies freely, the total does not. Below: the invariant itself, one stroke against 137. A recipient needs no unit, no base, no numeral system and no language, only the ability to count. The approximation is discussed in Remark D.8.

This is not an idle conceit. The Pioneer plaque used exactly this strategy, taking as

its unit the hyperfine transition of neutral hydrogen — the most common object in the universe, with a frequency any radio astronomer anywhere must know — and expressing every other quantity as a multiple of it [1482]. The design principle is the one above: build the message out of $\ker D$, never out of a chosen yardstick.

D.3.3 A primer built from nothing

A single ratio is not a message. To make α mean anything to a recipient one must first establish what a count is, what a ratio is, and what the two things are whose coupling is being reported — and each of those must be established using only what has already been established. Figures D.3 and D.4 carry that construction out.

The governing constraint is strict and is what makes the exercise instructive: *no glyph may be used before a panel has defined it by exhibiting instances.* Nothing is translated, because there is no shared language to translate into; everything is shown. That forces an order. Counting must come before equality, since equality is defined as sameness of count. Equality must come before the marks for TRUE and FALSE, which are introduced by attaching them to one true and one false statement — an ostensive definition, and the only kind available. Only then can anything be *asserted*, so only then can the naturals, sets, division and the reals follow.

Two features of the second board deserve comment. First, α is introduced *as a coupling before it is a number*: panel 10 shows the vertex at which an electron emits or absorbs a photon, and identifies α as how strongly that joining happens. This is the right order, because Equation (D.8) is not a definition a stranger could parse — it presupposes e, ε_0, $\hbar$ and c, none of which is available — whereas the vertex is a picture. Second, the alphabet is introduced last and is anchored: panel 13 pairs each written word with the thing already drawn, because a list of 26 shapes carries no information whatever until at least one of them is tied to something the recipient can independently identify. That is the same requirement Frege's sense and reference imposes (Section 3.7.1), arriving here as an engineering constraint rather than a philosophical one.

What the boards cannot do is worth stating as plainly as what they can. They convey $\alpha \approx 1/137$ and not $\alpha^{-1} = 137.035\,999\,177 \pm 0.000\,000\,021$; they do not specify the energy scale at which the coupling was measured, for the reason given in Remark D.8; and they assume the recipient shares our arithmetic, which is the one assumption the construction cannot itself justify. They are a demonstration of what a convention-free message would have to look like, not a protocol ready to transmit.

Remark D.8 (Three honest qualifications). The picture is a good deal less tidy than the slogan, and the qualifications matter more than the slogan does.

First, 137 *is not the value.* The measured quantity is $\alpha^{-1} = 137.035\,999$ to the digits shown, and no tally can draw a non-integer. What Figure D.2 transmits is a rational approximation good to about four parts in 10^5 — impressive for a row of scratches, useless for physics. Eddington's insistence that the value must be exactly an integer, first 136 and then 137 when measurement moved, is the cautionary tale here [889]: the integer was the artefact, and the digits after the decimal point were the physics.

Second, α runs. The number $1/137$ is the low-energy limit, measured at zero momentum transfer. At the mass of the Z boson the same coupling is near $1/128$ (Chapter 108). A civilization measuring at a different scale and reporting the result without saying so would appear to disagree with us while agreeing perfectly. Any serious message would have to specify the scale — which can itself be done dimensionlessly, as a ratio to a particle mass.

Third, agreement is expected, not guaranteed. That α takes the same value everywhere is an empirical claim, and Appendix D.3.4 reports how well it is established. It is exactly the kind of claim that would be most interesting to find false.

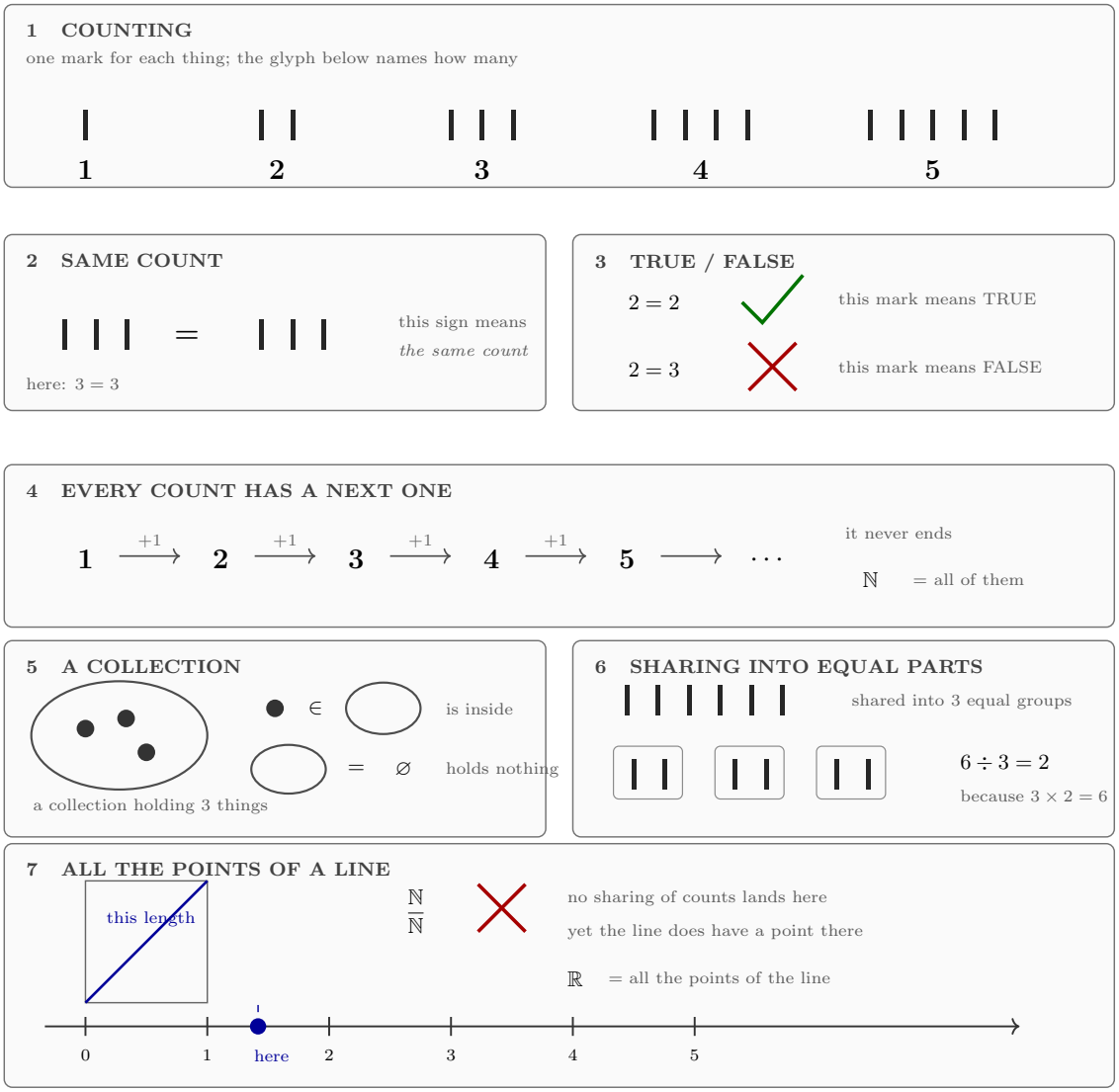


Figure D.3: An ostensive primer, first board: mathematics. Counting is established by tally, then the equality sign as sameness of count, then the marks for TRUE and FALSE by exhibiting a true and a false statement. Those three panels make assertion possible; the rest follow. The unit square in panel 7 is drawn to scale, its side equal to the 0-to-1 spacing of the line, so the transferred diagonal really does land at $\sqrt{2}$ — a point no sharing of counts reaches, which is what forces $\mathbb{R}$ beyond $\mathbb{N}/\mathbb{N}$.

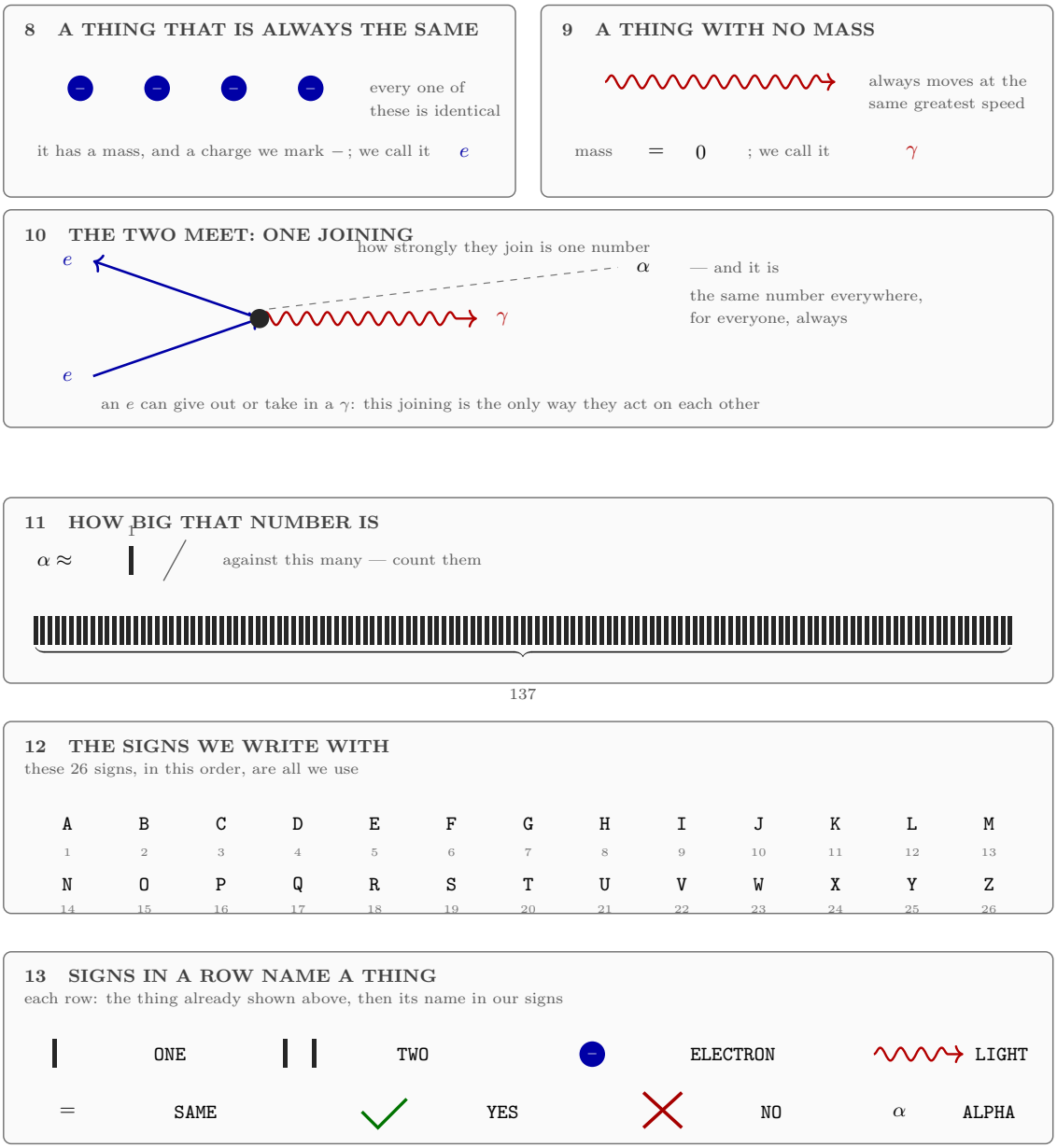


Figure D.4: Second board: physics and language. The electron is introduced as a thing of which every instance is identical, the photon as a thing with no mass moving always at the greatest speed, and α as the strength of the single joining by which the two act on each other — which is what α *is*, before it is a number. Panel 11 then gives its size as a count. Panels 12 and 13 attach a written language to referents already fixed on the boards: an alphabet in isolation is uninterpretable, so the words are grounded in things the recipient has just been shown.

D.3.4 Is it actually constant?

The universality argument is worth only as much as the evidence that α does not vary. Three independent probes bound it.

1. **Laboratory clocks** give the cleanest bound, with no astrophysical modelling at all. Comparing an aluminium-ion with a mercury-ion optical clock over a year — transitions whose frequencies depend on α with different powers — gives

$$\frac{\dot{\alpha}}{\alpha} = (-1.6 \pm 2.3) \times 10^{-17}/\mathrm{yr}\,, \tag{D.10}$$

   consistent with zero [2134, 2183].

2. **The Oklo natural reactor.** Isotope ratios left by a natural fission reactor that operated in Gabon about 2×10^9 years ago constrain the position of a samarium neutron resonance, and through it α at that epoch [1919]:

$$-0.9 \times 10^{-7} < \frac{\Delta\alpha}{\alpha} < 1.2 \times 10^{-7} \tag{D.11}$$

   at two standard deviations.

3. **Quasar absorption spectra** probe α at large redshift and in distant directions — the only probe that tests *spatial* universality. Here the evidence is contested, and must be reported as such. The many-multiplet analyses of Keck and VLT data are mutually inconsistent: Keck/HIRES gives $\Delta\alpha/\alpha = (-0.57 \pm 0.11) \times 10^{-5}$, a nominal five-sigma detection of variation, while VLT/UVES on the same method gives $(-0.06 \pm 0.06) \times 10^{-5}$, consistent with none [2183]. The proposed reconciliation — that α varies spatially, the two telescopes viewing different hemispheres — was reported as a 4.2σ dipole [2184]. That claim has since been substantially undercut, and from inside the original collaboration: long-range wavelength-scale distortions of order $\pm 200\,\mathrm{m/s}$ per 100 nm are ubiquitous in both instruments' archives, and reproduce important features of the reported signal [2261].

The laboratory and geophysical bounds are robust and consistent with no variation; the astrophysical detection is contested and is recorded here as contested rather than omitted, because a claim of that kind is exactly what the universality argument stands or falls on. Reporting only the comfortable evidence would be a failure of the editorial rule, and the contested item is in any case the more interesting one: it is the one that could still turn out to matter.

The universality of Appendix D.3.1 is therefore an experimentally supported statement about this universe, not an assumption — and it is falsifiable, which by Remark 1.2 is the most that can be asked of it.

D.4 Relations to mathematical constants

If the dimensionless constants are pure numbers, it is natural to ask whether they are *recognizable* numbers — whether α or the mass ratios are built from π, from e, from small integers. The history of that question is mostly a history of error, and it is worth setting out carefully, because the failure mode is seductive and the criterion for distinguishing sense from nonsense is sharp.

D.4.1 When a mathematical constant legitimately appears

Mathematical constants appear throughout physics with impeccable credentials. The 2π in the pendulum period of Example D.4 comes from integrating the equation of motion; the π in the Gaussian integral of Chapter 11 comes from the integral; the 4π in Equation (D.8) comes from the surface area of a sphere in Gauss's law. In every case the constant enters as the *output of a derivation*, and one can point to the step that produced it.

Definition D.9 (The criterion)**.** A proposed relation between a measured dimensionless constant and a mathematical one is physics if and only if it is *derived* — if there is an argument from stated premises whose conclusion is the relation, and which would have predicted the value before it was measured. A relation obtained by searching for a formula that reproduces a known value is *numerology*, however many digits it matches.

The asymmetry is the same one that separates a prediction from a fit (Section 1.4), and the reason for the strictness is counting: the supply of short expressions built from π, e and small integers is enormous, so agreement to a few digits with *some* such expression is expected by chance and carries no information.

D.4.2 Eddington, and the failure mode

Eddington argued that α^{-1} must be exactly the integer 136, on a group-theoretic counting argument; when measurements moved, he produced a revised derivation giving exactly 137 [889]. The modern value is 137.035 999, and neither derivation survives. The instructive part is not that he was wrong but *how*: the argument was adjusted to track the data while retaining the appearance of derivation, which is the signature Definition D.9 is designed to catch.

D.4.3 The Koide relation

The most serious modern case is Koide's [1691, 1723]. Form, from the three charged-lepton masses, the ratio

$$Q := \frac{m_e + m_\mu + m_\tau}{\left(\sqrt{m_e} + \sqrt{m_\mu} + \sqrt{m_\tau}\right)^2} \,. \tag{D.12}$$

Q is dimensionless by construction, and it is bounded: writing $v_i = \sqrt{m_i}$, Equation (D.12) is $|\boldsymbol{v}|^2 \,/\, (\boldsymbol{v}\cdot\boldsymbol{u})^2$ with $\boldsymbol{u} = (1,1,1)/\sqrt{3}$ up to normalization, so by Cauchy–Schwarz

$$\frac{1}{3} \le Q \le 1 \,, \tag{D.13}$$

the lower bound attained when the three masses are equal and the upper when one dominates. The measured value sits almost exactly at the midpoint of that range: with the current pole masses,

$$Q_{\text{obs}} = 0.666\,664\,5 \pm 0.000\,005\,1 \,, \qquad \frac{2}{3} = 0.666\,666\,7 \,, \tag{D.14}$$

a deviation of 0.43σ, or agreement at the level of a few parts in 10^6.

What should be made of it? Three observations, in the spirit of Definition D.9.

1. The agreement is genuinely striking, and *much* better than the bounds Equation (D.13) alone would make likely. It has the predictive character Definition D.9 demands: the relation was written down when the τ mass was poorly known, and it survived the subsequent measurements — indeed the agreement *improved* when the world average moved after the Belle II determination. That is the reason it is discussed here rather than dismissed.

2. It has no accepted derivation. Lepton masses in the Standard Model are free Yukawa parameters with no reason to satisfy any relation, and no mechanism in or beyond it produces Equation (D.14).

3. **The precision is a fact about pole masses, and only about pole masses** — which is the crux, and is usually stated carelessly. Masses run with scale (Chapter 108), so "the mass of the electron" is a function of where it is defined. Two results must be held together. Li and Ma showed the *running* Koide ratio is very nearly scale-independent, drifting by less than one part in 10^5 from 1 GeV to 2×10^{16} GeV [2099] — but the value it is frozen at is about 0.19 % *above* 2/3, not 2/3. Sumino showed the same thing from the other side: the QED self-energy term $-\frac{3\alpha}{4\pi} m_i \log m_i^2$, coming from the infrared region of the loop, rotates the vector $(\sqrt{m_e}, \sqrt{m_\mu}, \sqrt{m_\tau})$, so that if the running masses satisfied the relation at a high scale the pole masses would violate it by roughly 0.1 % — about 120 times the experimental error [2150].

   The two are not in conflict, and it is worth being explicit because they are routinely quoted against each other: the running ratio is stable but sits off 2/3, while the pole ratio sits on 2/3. Nothing carries the running value onto 2/3. And that is precisely what is suspicious — pole masses are infrared-sensitive, low-energy quantities that no known symmetry singles out, whereas a fundamental relation would be expected to hold at a high scale where a symmetry is unbroken.

4. The extension to quarks does not work, and fails asymmetrically: $Q \approx 0.669$ for the heavy triplet and $Q \approx 0.57$ for the light one, and it must be applied to *running* masses there — the opposite prescription from the charged leptons. A relation needing a different mass definition in each sector to look good is a relation whose status is unresolved.

The honest verdict is that the Koide relation is an unexplained numerical regularity of the right shape to be meaningful, which in four decades has been neither derived nor explained away. Koide himself allows that it "may be an accidental coincidence". This treatise records it as exactly that, and Definition D.9 says what would have to be supplied to settle it: a derivation, from stated premises, of why the relation should hold for pole masses.

Remark D.10 (A cautionary detail about which paper says what). The relation $Q = 2/3$ is that of Koide's earlier "model I" [1691, 1723]. The often-cited 1983 Brief Report [1724] proposes a *modified* relation with an extra parameter $\delta = 0.040\,026$, giving $Q = (2 + \delta^2)/3$ and predicting $m_\tau = 1786.45$ MeV — which the measured (1776.93 ± 0.09) MeV excludes overwhelmingly. Citing that paper for $Q = 2/3$, as is common, attributes the surviving relation to the one Koide paper whose headline prediction is dead.

D.4.4 Fine-tuning and the anthropic move

Several of the constants appear to lie in narrow windows outside which complex structure would not form. Weinberg's bound on the cosmological constant is the sharpest instance: he argued, before the measurement, that Λ could not greatly exceed a value set by the requirement that galaxies form, and the measured value (Appendix C.5) is of that order [1799].

Whether such reasoning explains anything is disputed, and this treatise does not adjudicate it. What can be said within its editorial rules is narrow and worth saying: an anthropic argument requires an ensemble of universes with varying constants for the selection to act on, and there is at present no observational evidence that such an ensemble exists. It therefore falls under the same rule as the programmes named in Chapter 135 — it may be named,

and it is named here, but it is not evidence-based physics and is not presented as such. The measured values in Appendix C stand as measurements whatever their eventual explanation.

D.5 Summary

The content of physics that is independent of every human choice is exactly $\ker D$: the dimensionless combinations. There are about thirty of them in established physics; none is derived; and they are the same throughout the observable universe to the precision of Appendix D.3.4. They are therefore simultaneously the deepest unexplained facts we possess and the only quantitative statements we could compare, without any shared convention, against a measurement made by anyone else anywhere.

Figure D.2 is the smallest complete expression of that idea: one stroke, then one hundred and thirty-seven. It is not the value of α, and it does not say at what scale it was measured. But it is the part of our physics that needs no dictionary — and the part that a stranger could confirm or refute.

Appendix E

Notation

This appendix is the treatise's single, canonical reference for notation: general typographic conventions (Appendix E.1), the macro-to-meaning dictionary (Appendix E.3), and a pointer to the dimension-and-index ladder used throughout the geometric chapters (Appendix E.2). The editorial *rules* that govern how the treatise is written — the SI axiom, the proofs-always rule, the experiment-detail rule — are stated once, in the front matter (*Editorial Conventions*); this appendix does not repeat them, only the symbols.

E.1 General conventions

- Vectors are set in bold, $\boldsymbol{x}$; operators wear no hats unless ambiguity demands them.
- The upright d denotes the differential, i the imaginary unit, e Euler's number — never the italic d, i, e used for ordinary variables (Chapter 1 macro policy).
- Einstein summation over repeated indices is understood unless stated otherwise.
- The metric signature in relativistic chapters is $(-,+,+,+)$. Geometric chapters of Part II are written for general dimension $D = p{+}q$ with signature (p, q) (Appendix E.2).
- A tensor's *type* is written (r, s): r contravariant (upper) and s covariant (lower) slots. The pair (p, q) is reserved for metric signature, never for tensor type.
- Label prefixes identify what a cross-reference points to: `ch:` chapter, `sec:` section, `eq:` equation, `fig:` figure, `tab:` table, `exp:` experiment, `phen:` phenomenon, `app:` appendix section, `part:` part, `thm:`/`def:`/`prop:`/ `cor:`/`lem:` the corresponding theorem-family environment.

E.2 Index conventions for general dimension

Chapters that develop geometry in general dimension $D = p + q$ (Part II, Chapters 13 and 14) fix a specific ladder of index letters, embedding spaces, and symmetry groups — $\mu, \nu, \ldots$ versus $a, b, \ldots$ for diffeomorphism- and local-Lorentz-covariant objects, extended to $m, n, \ldots / A, B, \ldots$ and $M, N, \ldots / I, J, \ldots$ for the $(p{+}q{+}1)$- and $(p{+}q{+}2)$-dimensional embeddings of the (anti-)de Sitter and conformal geometries. Rather than duplicate that material, this appendix points to it: the full statement, with the symmetry-ladder table (Poincaré, Lorentz, translations, anti-de Sitter, de Sitter, conformal), is Section 13.1 of Chapter 13.

E.3 Macro-to-meaning dictionary

Every macro used in the mathematics of this treatise is defined once, in `preamble/macros.tex`, and used identically in every chapter; none is redefined locally. Table E.1 is the complete dictionary.

Remark E.1 (Banned redefinitions). The kernel commands `\d`, `\i`, `\c`, `\L` are never redefined in this treatise (CLAUDE.md macro policy): old ported sources that redefined them as differential/imaginary-unit/speed-of-light/ Lagrangian have been translated on port to `\dd`, `\ii`, plain c, and `\Lag` respectively. The `physics` package is banned outright, since its `\qty` collides with siunitx v3's.

Table E.1: Macro-to-meaning dictionary. Every symbol in this treatise's mathematics resolves to exactly one row of this table.

Macro	Renders as	Meaning
`\dd`	d	differential (upright)
`\ii`	i	imaginary unit (upright)
`\ee`	e	Euler's number (upright)
`\pp`	∂	partial derivative symbol
`\dv{f}{x}`	$\frac{\mathrm{d}f}{\mathrm{d}x}$	total derivative
`\pdv{f}{x}`	$\frac{\partial f}{\partial x}$	partial derivative
`\abs{x}`	$\lvert x \rvert$	absolute value / modulus
`\norm{x}`	$\lVert x \rVert$	norm
`\set{x}`	$\{x\}$	set-builder braces
`\gen{x}`	$\langle x \rangle$	generated subgroup/subspace
`\avg{x}`	$\langle x \rangle$	ensemble or expectation average
`\ket{\psi}`	$\lvert \psi \rangle$	quantum ket
`\bra{\psi}`	$\langle \psi \rvert$	quantum bra
`\braket{\phi}{\psi}`	$\langle \phi \vert \psi \rangle$	inner product
`\ketbra{\phi}{\psi}`	$\lvert \phi \rangle\langle \psi \rvert$	outer product
`\ev{A}{\psi}`	$\langle \psi \vert A \vert \psi \rangle$	expectation value
`\comm{A}{B}`	$[A, B]$	commutator
`\acomm{A}{B}`	$\{A, B\}$	anticommutator
`\pb{A}{B}`	$\{A, B\}$	Poisson bracket
`\tr`	tr	trace
`\transpose`	$^{\mathsf{T}}$ (superscript)	matrix transpose
`\sgn`	sgn	sign function
`\diag`	diag	diagonal matrix
`\im`	im	image of a map
`\id`	id	identity map
`\grad`, `\curl`, `\Div`	$\mathrm{grad}, \mathrm{curl}, \mathrm{div}$	gradient, curl, divergence
`\SO`, `\SU`, `\ISO`	$\mathrm{SO}, \mathrm{SU}, \mathrm{ISO}$	special orthogonal, unitary, inhomogeneous
`\U`, `\Ogrp`, `\GL`, `\SL`	$\mathrm{U}, \mathrm{O}, \mathrm{GL}, \mathrm{SL}$	unitary, orthogonal, general/special linear
`\Sp`, `\Spin`	$\mathrm{Sp}, \mathrm{Spin}$	symplectic, spin group
`\R`, `\C`, `\N`, `\Z`, `\Q`	$\mathbb{R}, \mathbb{C}, \mathbb{N}, \mathbb{Z}, \mathbb{Q}$	reals, complexes, naturals, integers, rationals
`\identity`	$\mathbb{1}$	identity operator
`\Lag`	$\mathcal{L}$	Lagrangian (never `\L`)
`\Ham`	$\mathcal{H}$	Hamiltonian
`\vect{x}`	$\boldsymbol{x}$	vector (bold)
`\ec`	,	equation-final comma
`\ep`	.	equation-final period
`\nn`	(no output)	suppress an equation number

Bibliography

[1] W. Gilbert. *De Magnete, Magneticisque Corporibus, et de Magno Magnete Tellure; Physiologia noua, plurimis & argumentis, & experimentis demonstrata.* The title page reads 'Guilielmi Gilberti Colcestrensis, medici Londinensis, De Magnete...'; the form above is the standard short title as printed. First edition London (Londini), Peter Short, 1600. No DOI; scans are free at the Smithsonian and Library of Congress, so it can grade GREEN with a manifest URL. Surname is Gilbert (Latinised Gilbertus on the title page). London: Petrus Short, 1600.

[2] M. Mersenne. *Harmonie universelle, contenant la théorie et la pratique de la musique.* Empirical laws of the vibrating string (frequency vs length, tension, mass). Paris: Sébastien Cramoisy, 1636.

[3] R. Descartes. *Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences. Plus la Dioptrique, les Météores et la Géométrie, qui sont des essais de cette méthode.* Leyde (Leiden), 1637. The chapter cites it for the sine law of refraction, which is in the second discours ('De la réfraction') of *La Dioptrique* – one of the three essays bound with the Discours, not a separate publication, so cite the 1637 volume and name the essay in a note. Title given in modernised orthography; the title page reads 'DISCOURS DE LA METHODE Pour bien conduire sa raison, & chercher la verité dans les sciences. PLUS LA DIOPTRIQUE. LES METEORES. ET LA GEOMETRIE. Qui sont des essais de cete METHODE.' and the imprint 'De l'Imprimerie de Ian Maire' (Ian = Jan). No DOI; facsimiles are freely available (Gallica, e-rara), which gives a real access route. Jan Maire, 1637.

[4] G. Galilei. *Discorsi e dimostrazioni matematiche intorno a due nuove scienze.* English translation: H. Crew and A. de Salvio, Macmillan (1914). Leiden: Elzevir, 1638.

[5] E. Torricelli. *Opera geometrica.* Title page prints 'Opera geometrica Evangelistae Torricellii'. The only work published in Torricelli's lifetime; three separately signed parts. The efflux law is in the second part, 'De motu gravium naturaliter descendentium, et proiectorum libri duo', book II ('De motu aquarum') – this is the 'memoir on the motion of heavy bodies of 1644' the chapter names. No DOI (1644 imprint); a full scan is public domain at archive.org/details/operageometrica00torr, which would make an honest manifest access route. As an @book it is auto-excused from the DOI rule and grades AMBER without one. Florentiae: Typis Amatoris Massae & Laurentij de Landis, 1644.

[6] R. Boyle. *A Defence of the Doctrine Touching the Spring and Weight of the Air.* Appendix to the second edition of New Experiments Physico-Mechanicall, Touching the Spring of the Air; contains the pressure-volume data now called Boyle's law (reply to Linus and Hobbes). Oxford: Thomas Robinson, 1662.

[7] B. Pascal. *Traitez de l'équilibre des liqueurs et de la pesanteur de la masse de l'air*. Posthumous (Pascal died 1662). Pascal's principle and the hydraulic press. Paris: Guillaume Desprez, 1663.

[8] F. M. Grimaldi. *Physico-mathesis de lumine, coloribus, et iride, aliisque adnexis libri duo*. Posthumous; the work in which diffraction is named and reported, which is exactly the claim cited. The title page reads 'adnexis'; many modern citations modernise it to 'annexis' – prefer the printed 'adnexis'. Imprint is sometimes given as 'heirs of Vittorio Benacci for Girolamo Bernia'. Pre-1990 book, no DOI. Bologna: Ex typographia haeredis Victorij Benatij, 1665.

[9] E. Bartholinus. *Experimenta crystalli Islandici disdiaclastici quibus mira & insolita refractio detegitur*. The report of double refraction in Iceland spar. Published at Hafniae (Copenhagen), 1669, 4to, [4] + 60 pp. The title page prints the Latinised name (Erasmi Bartholini / Erasmus Bartholinus); the Danish form Rasmus Bartholin and the anglicised 'Bartholin' used in the chapter prose are the same man. Keep the ampersand '&' as printed. Public-domain scans exist, so a URL can make this GREEN. Sumptibus Danielis Paulli, 1669.

[10] I. Newton. "A letter of Mr. Isaac Newton, Professor of the Mathematicks in the University of Cambridge; containing his new theory about light and colors: sent by the author to the publisher from Cambridge, Febr. 6. 1671/72; in order to be communicated to the R. Society". In: *Philosophical Transactions of the Royal Society of London* 6.80 (1672). Issue no. 80, dated 19 February 1671/72 – Old Style, i.e. 1672 in modern reckoning, which is why the DOI string carries 1671 while the year is 1672. Volume 6 spans 1671–1672. The long title above is the one Crossref and the Royal Society print; a shortened 'New Theory about Light and Colours' is customary in prose but is not the published title., pp. 3075–3087. DOI: 10.1098/rstl.1671.0072.

[11] C. Huygens. *Horologium Oscillatorium sive de motu pendulorum ad horologia aptato demonstrationes geometricae*. Cycloidal pendulum, centre of oscillation/compound pendulum; theorems on centrifugal force stated at the end. Paris: F. Muguet, 1673.

[12] O. Rømer. "Démonstration touchant le mouvement de la lumière trouvé par M. Römer de l'Académie Royale des Sciences". In: *Journal des Sçavans* (1676). The first evidence that light propagates at a finite speed, from the timing of the eclipses of Io. No DOI; digitised by the Bibliothèque nationale de France, pp. 233–236.

[13] R. Hooke. *Lectures de Potentia Restitutiva, or of Spring Explaining the Power of Springing Bodies*. States 'ut tensio, sic vis', announced as the anagram ceiiinosssttuv in 1676. London: John Martyn, 1678.

[14] I. Newton. *Philosophiæ Naturalis Principia Mathematica*. London: Royal Society, 1687.

[15] C. Huygens. *Traité de la lumière*. Leiden: Pierre van der Aa, 1690.

[16] J. Bernoulli. "Solutio problematis funicularii". In: *Acta Eruditorum* (1691). Catenary solution, June 1691 issue; Leibniz and Huygens published solutions in the same issue. Author is Johann Bernoulli (same surname as the 1696 entry, different problem).

[17] J. Bernoulli. "Problema novum ad cuius solutionem Mathematici invitantur". In: *Acta Eruditorum* (1696). The brachistochrone challenge, June 1696 issue; solutions by Newton, Leibniz, l'Hopital and the Bernoullis appeared in the May 1697 issue., p. 269.

[18] I. Newton. *Opticks: or, A Treatise of the Reflexions, Refractions, Inflexions and Colours of Light. Also Two Treatises of the Species and Magnitude of Curvilinear Figures*. First edition, London, 1704, published anonymously (Newton's name does not appear on the title page). The two mathematical treatises appended in 1704 were dropped from the 1717 second edition, so an edition field is worth carrying: cite the 1704 first edition, since Newton's rings are described in Book II of that edition. Address: London. Sam. Smith and Benj. Walford, Printers to the Royal Society, 1704.

[19] J. Bradley. "A letter from the Reverend Mr. James Bradley Savilian Professor of Astronomy at Oxford, and F. R. S. to Dr. Edmond Halley Astronom. Reg. &c. giving an account of a new discovered motion of the fix'd stars". In: *Philosophical Transactions of the Royal Society of London* 35.406 (1728). The discovery of stellar aberration. Crossref dates the issue 1728-12-31 and gives vol. 35, pp. 637-661, so year 1728 is the publisher's dating and matches the key. The DOI suffix reads 'rstl.1727' because Phil. Trans. vol. 35 spans the years 1727-1728 – that is not an error, do not 'correct' it to 1728. Crossref prefixes the title with the Royal Society's modern article number 'IV.', which is a digitisation artefact and is omitted here; the DOI title check should still match. Article number 406., pp. 637–661. DOI: 10.1098/rstl.1727.0064.

[20] H. Pitot. "Description d'une machine pour mesurer la vitesse des eaux courantes et le sillage des vaisseaux". In: *Mémoires de l'Académie Royale des Sciences (Paris)*. Volume for the year 1732, printed 1735; the series is numbered by year, not by volume (academy-proceedings rule). The Pitot tube. 1732, pp. 363–376.

[21] G. Hadley. "Concerning the cause of the general trade-winds". In: *Philosophical Transactions of the Royal Society of London* 39.437 (1735). Explanation of the trade winds by the Earth's rotation. Volume 39 spans 1735-1736; customarily dated 1735., pp. 58–62. DOI: 10.1098/rstl.1735.0014.

[22] D. Bernoulli. *Hydrodynamica, sive de viribus et motibus fluidorum commentarii*. The pressure-velocity relation appears via a vis viva argument; the modern local form of Bernoulli's theorem is Euler's (see Euler:1757). The chapter records this attribution explicitly. Strasbourg (Argentorati): Johann Reinhold Dulsecker, 1738.

[23] L. Euler. *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes, sive solutio problematis isoperimetrici latissimo sensu accepti*. Founding monograph of the calculus of variations; Additamentum II states the least-action principle for mechanics (relevant to the Maupertuis misattribution remark). Lausanne and Geneva: Marcum-Michaelem Bousquet, 1744.

[24] P. L. M. de Maupertuis. "Les loix du mouvement et du repos déduites d'un principe métaphysique". In: *Histoire de l'Académie Royale des Sciences et des Belles-Lettres de Berlin*. Volume for the year 1746 (printed 1748); customarily cited as 1746. His 1744 Paris memoir treats only the refraction of light – the mechanical least-action claim is here, and the mathematically correct statement is Euler:1744 Additamentum II (misattribution noted in the chapter prose). 1746, pp. 267–294.

[25] J. le Rond d'Alembert. "Recherches sur la courbe que forme une corde tendue mise en vibration". In: *Histoire de l'Académie royale des sciences et des belles-lettres de Berlin* 3 (1747). The wave equation and its travelling-wave solution. Volume for 1747, printed 1749; keyed by the customary/nominal volume year (no DOI exists to arbitrate). A continuation follows at pp. 220-249., pp. 214–219.

[26] J. Bradley. “A letter to the Right honourable George Earl of Macclesfield concerning an apparent motion observed in some of the fixed stars; by James Bradley D. D. F. R. S”. In: *Philosophical Transactions of the Royal Society of London* 45.485 (1748). Discovery of the nutation of the Earth’s axis., pp. 1–41. DOI: 10.1098/rstl.1748.0002.

[27] J. le Rond d’Alembert. *Essai d’une nouvelle théorie de la résistance des fluides*. Source of the d’Alembert paradox: steady potential flow exerts no drag. Paris: David l’aîné, 1752.

[28] D. Bernoulli. “Réflexions et éclaircissemens sur les nouvelles vibrations des cordes exposées dans les mémoires de l’Académie de 1747 et 1748”. In: *Histoire de l’Académie royale des sciences et des belles-lettres de Berlin* 9 (1753). Superposition of trigonometric modes on the vibrating string. Volume for 1753, printed 1755; keyed by the nominal volume year, matching customary citation., pp. 147–172.

[29] J. A. Segner. *Specimen theoriae turbinum*. Discovery that every rigid body has three mutually orthogonal principal axes. Halle: Gebauer, 1755.

[30] L. Euler. “Principes généraux du mouvement des fluides”. In: *Mémoires de l’Académie Royale des Sciences et des Belles-Lettres de Berlin*. Vol. 11. Presented 1755; the volume for 1755 was printed in 1757. Contains the Euler equations and the continuity equation in field form. 1757, pp. 274–315.

[31] J. L. Lagrange. “Essai d’une nouvelle méthode pour déterminer les maxima et les minima des formules intégrales indéfinies”. In: *Miscellanea Taurinensia* 2 (1762). Volume for 1760–61, printed 1762; customarily dated 1760/1762. Introduces the delta-calculus derivation of the Euler–Lagrange equation., pp. 173–195.

[32] L. Euler. *Theoria motus corporum solidorum seu rigidorum*. Enestrom E289. Euler’s equations, the inertia tensor and moments of inertia, free precession of the symmetric top, prediction of the Earth’s free wobble (period near 305 days). Rostock and Greifswald: A. F. Röse, 1765.

[33] N. Maskelyne. “An account of observations made on the mountain Schehallien for finding its attraction”. In: *Philosophical Transactions of the Royal Society of London* 65.65 (1775). The Schiehallion plumb-line experiment; the paper spells the mountain ’Schehallien’., pp. 500–542. DOI: 10.1098/rstl.1775.0050.

[34] L. Euler. “Formulae generales pro translatione quacunque corporum rigidorum”. In: *Novi Commentarii Academiae Scientiarum Imperialis Petropolitanae* 20 (1776). Enestrom E478. Euler’s rotation theorem and the general rotation formulae; volume for 1775, published 1776., pp. 189–207.

[35] C. Hutton. “An account of the calculations made from the survey and measures taken at Schehallien, in order to ascertain the mean density of the Earth”. In: *Philosophical Transactions of the Royal Society of London* 68.68 (1778). Reduction of the Schiehallion survey; mean density near 4.5 times water., pp. 689–788. DOI: 10.1098/rstl.1778.0034.

[36] C.-A. de Coulomb. “Recherches théoriques et expérimentales sur la force de torsion et sur l’élasticité des fils de métal”. In: *Mémoires de l’Académie Royale des Sciences (Paris)*. Volume for the year 1784, printed 1787; the series is numbered by year, not volume (academy-proceedings rule), and the customary citation year 1784 is kept as key since no DOI arbitrates. Torsion law of metal fibres; independent invention of the torsion balance. 1784, pp. 229–269.

[37] J. Michell. “On the Means of Discovering the Distance, Magnitude, &c. of the Fixed Stars, in Consequence of the Diminution of the Velocity of Their Light”. In: *Philosophical Transactions of the Royal Society of London* 74 (1784). The dark-star argument. Metadata confirmed against Crossref., pp. 35–57. DOI: 10.1098/rstl.1784.0008.

[38] C. A. d. Coulomb. “Premier mémoire sur l’électricité et le magnétisme”. In: *Histoire de l’Académie Royale des Sciences, avec les Mémoires de Mathématique et de Physique, Année 1785*. Academy proceedings: @inproceedings with a booktitle and NO volume – the series has none, do not invent one. DATING: the volume bears ’Année 1785’ on its title page (the memoir was read to the Académie in 1785), but it was actually printed in 1788; the customary citation, and the key, follow 1785. Record the 1788 imprint in a bib note. The full printed title continues ’Construction & usage d’une balance électrique, fondée sur la propriété qu’ont les fils de métal, d’avoir une force de réaction de torsion proportionnelle à l’angle de torsion.’ – the torsion-balance inverse-square measurement the chapter cites. Pages 569-577 verified. No DOI; scans on Gallica/Internet Archive (Coulomb, *Mémoires sur l’électricité et le magnétisme*) give a real access route. 1785, pp. 569–577.

[39] E. F. F. Chladni. *Entdeckungen über die Theorie des Klanges*. The sand figures of vibrating plates. Leipzig: Weidmanns Erben und Reich, 1787.

[40] J.-L. Lagrange. *Méchanique analitique*. Original spelling of the first edition; later editions titled Mécanique analytique. General theory of small oscillations and the heavy symmetric top. Paris: la Veuve Desaint, 1788.

[41] A.-M. Legendre. “Mémoire sur la manière de distinguer les maxima des minima dans le calcul des variations”. In: *Mémoires de l’Académie Royale des Sciences de Paris*. Read to the Academy in 1786; printed in the volume for the year 1786, which appeared in 1788. Customarily cited as Legendre 1786 – keyed by publication year per house rule. 1788, pp. 7–37.

[42] G. B. Venturi. *Recherches expérimentales sur le principe de la communication latérale du mouvement dans les fluides appliqué à l’explication de différens phénomènes hydrauliques*. The Venturi effect and meter. Paris: Houel et Ducros, 1797.

[43] C. R. Benjamin Thompson. “An inquiry concerning the source of the heat which is excited by friction”. In: *Philosophical Transactions of the Royal Society of London* 88 (1798), pp. 80–102. DOI: 10.1098/rstl.1798.0006.

[44] H. Cavendish. “Experiments to determine the density of the earth”. In: *Philosophical Transactions of the Royal Society of London* 88 (1798). The torsion-balance experiment; credits John Michell for conceiving and building the apparatus., pp. 469–526. DOI: 10.1098/rstl.1798.0022.

[45] P.-S. Laplace. *Traité de mécanique céleste, Tome premier*. The potential equation appears in Book II; the equation itself dates to his 1780s memoirs on spheroidal attraction. Paris: J. B. M. Duprat, 1799.

[46] J. Dalton. “Experimental Essays on the Constitution of Mixed Gases; on the Force of Steam or Vapour from Water and Other Liquids in Different Temperatures, Both in a Torricellian Vacuum and in Air; on Evaporation; and on the Expansion of Gases by Heat”. In: *Memoirs of the Literary and Philosophical Society of Manchester* 5 (1802). Society memoirs but carries a volume number; the partial-pressure law and the gas-expansion law., pp. 535–602.

[47] J. L. Gay-Lussac. “Recherches sur la dilatation des gaz et des vapeurs”. In: *Annales de Chimie* 43 (1802). Credits earlier unpublished work by Charles (customary ’Charles’s law’)., pp. 137–175.

[48] T. Young. “The Bakerian Lecture. On the theory of light and colours”. In: *Philosophical Transactions of the Royal Society of London* 92 (1802). Verified via Crossref. Two cautions. (1) The leading "II." is the Royal Society's article number as Crossref stores it; the 1802 printing reads "The Bakerian Lecture. On the Theory of Light and Colours". I give Crossref's exact string so the machine title-check cannot reject the DOI — drop the numeral in the printed bibliography if the house style prefers. (2) The lecture was READ 12 November 1801 and is often cited as 1801; the issue year of Phil. Trans. vol. 92 is 1802, which is what the key uses, so no discrepancy to record. Distinct from Young's other 1802 Phil. Trans. paper, "An account of some cases of the production of colours, not hitherto described" (vol. 92, 387-397) — if that one is ever needed it takes key Young:1802b., pp. 12–48. DOI: 10.1098/rstl.1802.0004.

[49] T. Young. “The Bakerian Lecture. Experiments and calculations relative to physical optics”. In: *Philosophical Transactions of the Royal Society of London* 94 (1804). Verified via Crossref. Same leading-numeral caveat as Young:1802: "I." is the Royal Society article number carried by Crossref, not part of the 1804 printed title. This is the paper carrying the two-source interference experiments and the wavelength calculations that both citing chapters attribute to it., pp. 1–16. DOI: 10.1098/rstl.1804.0001.

[50] T. Young. “An essay on the cohesion of fluids”. In: *Philosophical Transactions of the Royal Society of London* 95 (1805). Surface tension and contact angles; Young's half of the Young-Laplace equation (Laplace's half is the 1806 supplement to Mécanique Céleste, Book X)., pp. 65–87. DOI: 10.1098/rstl.1805.0005.

[51] M.-A. P. des Chênes. “Mémoire sur les séries et sur l'intégration complète d'une équation aux différences partielles linéaires du second ordre, à coefficients constants”. In: *Mémoires présentés à l'Institut des Sciences, Lettres et Arts, par divers savans, Sciences mathématiques et physiques*. Vol. 1. Read in 1799, printed 1806; customarily dated 1799. Full surname 'Parseval des Chênes'. 1806, pp. 638–648.

[52] J. L. Gay-Lussac. “Premier essai pour déterminer les variations de température qu'éprouvent les gaz en changeant de volume, et considérations sur leur capacité pour le calorique”. In: *Mémoires de Physique et de Chimie de la Société d'Arcueil*. Vol. 1. Free-expansion experiment. 1807.

[53] T. Young. *A Course of Lectures on Natural Philosophy and the Mechanical Arts*. Two volumes. Introduces the modulus of elasticity (Young's modulus). London: Joseph Johnson, 1807.

[54] É. L. Malus. “Sur une propriété de la lumière réfléchie”. In: *Mémoires de physique et de chimie de la Société d'Arcueil*. Vol. 2. Tome 2 of the Société d'Arcueil memoirs, 1809. Malus made the observation in late 1808; the memoir is 1809, so the key year is correct. A short prior notice appeared in Nouveau Bulletin des Sciences par la Société Philomathique de Paris 1 (1809) 266-269 — the Arcueil memoir is the substantive paper and the one containing the cos^2 law. No DOI; French Wikisource carries the full text, which is the access route to declare. 1809, pp. 143–158.

[55] A. Avogadro. “Essai d'une manière de déterminer les masses relatives des molécules élémentaires des corps, et les proportions selon lesquelles elles entrent dans ces combinaisons”. In: *Journal de Physique, de Chimie, d'Histoire Naturelle et des Arts* 73 (1811), pp. 58–76.

[56] P.-S. Laplace. *Théorie analytique des probabilités*. Cited as the origin of the (Laplace) generating-function transform. Paris: Veuve Courcier, 1812.

[57] S. D. Poisson. “Remarques sur une équation qui se présente dans la théorie des attractions des sphéroïdes”. In: *Nouveau Bulletin des Sciences, par la Société Philomatique de Paris* 3 (1813). The Poisson equation., pp. 388–392.

[58] D. Brewster. “On the laws which regulate the polarisation of light by reflexion from transparent bodies. By David Brewster, LL. D. F. R. S. Edin. and F. S. A. Edin. In a letter addressed to Right Hon. Sir Joseph Banks, Bart. K. B. P. R. S”. In: *Philosophical Transactions of the Royal Society of London* 105.105 (1815). Crossref confirmed: vol. 105, pp. 125-159, issued 1815. The Crossref record carries the article number ’IX.’ and the trailing dedication ’In a letter addressed to Right Hon. Sir Joseph Banks, Bart. K. B. P. R. S.’; both are omitted here as digitisation/period boilerplate. This is the paper establishing Brewster’s angle (tan i = n)., pp. 125–159. DOI: 10.1098/rstl.1815.0010.

[59] P.-S. Laplace. “Sur la vitesse du son dans l’air et dans l’eau”. In: *Annales de Chimie et de Physique* 3 (1816). The adiabatic correction to Newton’s isothermal sound speed., pp. 238–241.

[60] J. Fraunhofer. “Bestimmung des Brechungs- und des Farbenzerstreungs-Vermögens verschiedener Glasarten, in Bezug auf die Vervollkommnung achromatischer Fernröhre”. In: *Annalen der Physik*. Vol. 56. 7. Academy Denkschriften, not a journal article. Umlauts: Vermoegens = Vermögens, Fernroehre = Fernröhre, Koeniglichen = Königlichen, Muenchen = München. 1817, pp. 264–313. DOI: 10.1002/andp.18170560706.

[61] A. Fresnel. “Lettre de M. Fresnel à M. Arago sur l’influence du mouvement terrestre dans quelques phénomènes d’optique”. In: *Annales de chimie et de physique* 9 (1818). The partial-drag coefficient (1 - 1/n^2), September 1818 issue of tome 9; a short addendum by Fresnel follows at pp. 286-287 of the same tome (November 1818). Journal article, not an academy memoir – kind is article. No DOI; freely readable on Gallica., pp. 57–66.

[62] H. Kater. “An Account of experiments for determining the length of the pendulum vibrating seconds in the latitude of London”. In: *Philosophical Transactions of the Royal Society of London* 108.108 (1818). Read to the Royal Society 29 January 1818; the reversible (Kater) pendulum and the first absolute determination of g. The Royal Society prints the title with the volume article number ’IV.’ prefixed; dropped here per the house convention already used for Cavendish:1798 (’XXI. Experiments to determine the Density of the Earth’). Typed @article, not @inproceedings: Philosophical Transactions is a numbered journal volume and the sibling Phil. Trans. entry Cavendish:1798 is @article with volume 88 and an rstl DOI. Volume 108 is printed on the volume (JSTOR ’Vol. 108, 1818’; Royal Society ’Phil. Trans. R. Soc. (1818) 108: 33-102’) and is not invented – Crossref’s record for this DOI stores 108 in the ’issue’ field and leaves ’volume’ empty, which is a metadata quirk of the pre-1900 rstl deposits. TRAP: Kater serialized the same memoir in The Philosophical Magazine 52 (1818) as four separate DOI-bearing pieces (pp. 90-100, 173-182, 364-370, 415-431), all with near-identical titles; the memoir read to the Royal Society is the Phil. Trans. paper cited here., pp. 33–102. DOI: 10.1098/rstl.1818.0006.

[63] F. Arago and A. Fresnel. “Mémoire sur l’action que les rayons de lumière polarisée exercent les uns sur les autres”. In: *Annales de chimie et de physique* 10 (1819). The joint Arago–Fresnel memoir establishing the Fresnel–Arago laws (perpendicularly polarized beams do not interfere), which is exactly what the citing sentence claims. Série 2, tome 10, 1819; the journal prints it ’par MM. Arago et Fresnel’, hence keying by Arago is right. Some dealer catalogues give the range as 288–306; 288–305

is the more commonly recorded. No DOI; digitised on Gallica. English translation by H. Crew in 'The Wave Theory of Light' (1900)., pp. 288–305.

[64] P. L. Dulong and A. T. Petit. "Recherches sur quelques points importants de la théorie de la chaleur". In: *Annales de Chimie et de Physique* 10 (1819), pp. 395–413.

[65] A.-T. Petit and P.-L. Dulong. "Recherches sur quelques points importants de la Théorie de la Chaleur". In: *Annales de Chimie et de Physique* 10 (1819). The Dulong–Petit law; keyed by first listed author, Petit., pp. 395–413.

[66] A.-M. Ampère. "Mémoire présenté à l'Académie royale des Sciences, le 2 octobre 1820, où se trouve compris le résumé de ce qui avait été lu à la même Académie les 18 et 25 septembre 1820, sur les effets des courans électriques". In: *Annales de chimie et de physique* 15 (1820). The memoir announcing the force between two currents. Published in two instalments in the same volume (série 2, tome 15): pp. 59–76 and, as the 'suite', pp. 170–218; the CNRS Ampère catalogue gives the first instalment as 59–74, most physics bibliographies as 59–76 – both refer to the same instalment, so keep the compound range rather than picking one. Note the period spelling 'courans', not 'courants'. No DOI; digitised on Gallica and at ampere.cnrs.fr., pp. 59–76, 170–218.

[67] J.-B. Biot and F. Savart. "Note sur le magnétisme de la pile de Volta". In: *Annales de chimie et de physique* 15 (1820). The two-page note reporting the 1/r force of a long straight current on a magnetic pole – the measurement the chapter describes. Série 2, tome 15, 1820; presented to the Académie 30 October 1820. The co-author is Félix Savart (not the acoustician's namesake Nicolas Savart, whom some catalogues wrongly list). No DOI; digitised on Gallica and at ampere.cnrs.fr., pp. 222–223.

[68] H. C. Ørsted. *Experimenta circa effectum conflictus electrici in acum magneticam.* A privately printed four-page Latin pamphlet (2 leaves, paginated [1] 2–4), dated at the end 21 July 1820, which Ørsted mailed to natural philosophers across Europe – hence kind=misc, not article: it appeared in no journal. Surname: the pamphlet itself uses the Latinised 'Oersted' and auction/library catalogues (Christie's, WorldCat) file it under OERSTED; the Danish form Ørsted is what the treatise prose uses. Contemporary reprints, if a retrievable route is wanted for the availability grade: Journal für Chemie und Physik (Schweigger's Journal) 29, 275–281 (1820), and the German version 'Versuche über die Wirkung des elektrischen Conflicts auf die Magnetnadel', Annalen der Physik 66 (1820) 295–304 – I have not verified a DOI for either, so do not attach one. Scans are open at the Internet Archive (Experimentacirc00Orst) and the Smithsonian Libraries, and an English translation is in Jelved, Jackson & Knudsen (eds.), Selected Scientific Works of Hans Christian Oersted (Princeton, 1998), pp. 413–420. 1820.

[69] J. Fraunhofer. "Neue Modifikation des Lichtes durch gegenseitige Einwirkung und Beugung der Strahlen, und Gesetze derselben". In: *Denkschriften der Königlichen Akademie der Wissenschaften zu München.* Vol. 8. Academy memoirs, so inproceedings; the Denkschriften do number their Bände, and Band 8 covers the years 1821-1822 – the volume number is real, not invented. This is the grating/diffraction memoir (wire and ruled gratings, the grating equation, first absolute wavelength determinations). No DOI. Distinct from Fraunhofer:1817, the solar dark-line catalogue, which the same chapter cites separately. 1821, pp. 3–76.

[70] S. Germain. *Recherches sur la théorie des surfaces élastiques.* The prize essay on the vibrating-plate equation, answering Chladni's figures. Paris: Veuve Courcier, 1821.

[71] F. Arago. “Résultats des expériences faites, par ordre du Bureau des Longitudes, pour la détermination de la vitesse du son dans l’atmosphère”. In: *Annales de Chimie et de Physique* 20 (1822). Série 2, tome 20 (1822). Verified page-by-page against the scanned volume (archive.org item s3id13207880): the memoir opens on p. 210 immediately after Wollaston’s note, runs through the printed page markers 211-223, and closes with Arago’s Annales siglum ’(Ar.)’; the volume’s own Table du vingtième volume lists it at p. 210. Unsigned in the heading, but written in the first person as one of the commission (’MM. de Prony, Bouvard, Mathieu et moi’), which identifies Arago as the reporter. Commission: Prony, Bouvard, Mathieu and Arago, joined by Humboldt and Gay-Lussac; the reciprocal Villejuif-Montlhéry cannon firings of 21-22 June 1822. Pre-DOI: no DOI exists, so this needs a manifest "access": "open" entry pointing at https://archive.org/details/s3id13207880 or it grades RED as a DOI-less @article. Do not confuse with Laplace:1816 (already in the bibliography) or with Laplace’s separate ’Note sur la vitesse du son’ in the same tome 20., pp. 210–223.

[72] J. B. J. Fourier. *Théorie analytique de la chaleur*. Fourier series and the decomposition of arbitrary initial data. Paris: Firmin Didot, 1822.

[73] A.-L. Cauchy. “Recherches sur l’équilibre et le mouvement intérieur des corps solides ou fluides, élastiques ou non élastiques”. In: *Bulletin de la Société Philomathique de Paris* (1823). Read to the Académie 30 September 1822 and universally cited as Cauchy (1822); keyed by the printing year per the house publisher’s-year rule. Announces the stress and strain tensors; the full development followed in the Exercices de mathématiques (1827-1828). The Bulletin has no volume numbers., pp. 9–13.

[74] J. Fraunhofer. “Kurzer Bericht von den Resultaten neuerer Versuche über die Gesetze des Lichtes, und die Theorie derselben”. In: *Annalen der Physik* 74.8 (1823). Verified against Crossref (issue 8 of volume 74; author printed ’Jos. Fraunhofer’). Volume 74 of the Annalen is Gilbert’s Annalen der Physik – record it as Annalen der Physik, which is what Wiley and Crossref use. This is the follow-up to the 1821 Munich memoir and is the more accessible of the two., pp. 337–378. DOI: 10.1002/andp.18230740802.

[75] C.-L. Navier. “Mémoire sur les lois du mouvement des fluides”. In: *Mémoires de l’Académie Royale des Sciences de l’Institut de France*. Vol. 6. Read 1822; the volume for 1823 was printed in 1827. Universally cited as 1823, and keyed so here since no DOI exists to arbitrate; record the dating in the entry note. 1823, pp. 389–440.

[76] S. Carnot. *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance*. Paris: Bachelier, 1824.

[77] A.-M. Ampère. *Théorie des phénomènes électro-dynamiques, uniquement déduite de l’expérience*. The book the chapter names in italics. Paris, Méquignon-Marvis, 1826 (first edition; LoC and Internet Archive hold it). Do NOT retitle it ’Théorie mathématique des phénomènes électro-dynamiques...’ – that is the 1827 Mémoires de l’Académie version and the 1883 second edition; the 1826 title page has no ’mathématique’. Public-domain scans exist, so this can grade GREEN with a URL. Méquignon-Marvis, 1826.

[78] A. Fresnel. “Mémoire sur la diffraction de la lumière, où l’on examine particulièrement le phénomène des franges colorées que présentent les ombres des corps éclairés par un point lumineux”. In: *Mémoires de l’Académie royale des sciences de l’Institut de France*. Vol. 5. The prize memoir: submitted 29 July 1818, crowned 15 March 1819, PRINTED only in 1826 in tome V (covering 1821-1822) – so the key year 1826 is

the publisher's year and is correct, while the work is customarily dated 1818/1819. This is the memoir containing the Poisson–Arago spot, the Fresnel integrals, and the biprism and two-mirror arrangements, which is why the same key is cited in four different places in 10-exp-wave-optics.tex. Academy memoirs, hence inproceedings; the tome number is genuine. Reprinted in Œuvres complètes d'Augustin Fresnel, tome I (1866), pp. 247-383, which is the freely available route. 1826, pp. 339–475.

[79] T. J. Seebeck. "Ueber die magnetische Polarisation der Metalle und Erze durch Temperatur-Differenz". In: *Annalen der Physik* 82.1 (1826), pp. 1–20. DOI: 10.1002/andp.18260820102.

[80] C.-L. Navier. "Mémoire sur les lois de l'équilibre et du mouvement des corps solides élastiques". In: *Mémoires de l'Académie Royale des Sciences de l'Institut de France*. Vol. 7. Read 14 May 1821; customarily cited as Navier (1821). Molecular derivation of the elastodynamic (Navier-Cauchy) equations. 1827, pp. 375–393.

[81] G. S. Ohm. *Die galvanische Kette, mathematisch bearbeitet*. T. H. Riemann, 1827.

[82] R. Brown. "A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies". In: *The Philosophical Magazine* 4.21 (1828). Series 2 (new series); also privately printed 1828., pp. 161–173. DOI: 10.1080/14786442808674769.

[83] G. Green. *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism*. Privately printed essay; origin of the Green's function and the reciprocity theorem. Nottingham: Printed for the author by T. Wheelhouse, 1828.

[84] W. R. Hamilton. "Theory of Systems of Rays". In: *The Transactions of the Royal Irish Academy* 15 (1828). Hamilton's characteristic-function optics, the work the eikonal section names. Read before the Royal Irish Academy 23 April 1827, printed in the volume dated 1828. Only Part First appeared in his lifetime; three Supplements followed in Trans. R. Irish Acad. 16 and 17 (1830, 1831, 1837). No DOI; JSTOR stable id 30078906. Transactions carries a volume number, so @article is appropriate despite being an academy series., pp. 69–174.

[85] P. G. L. Dirichlet. "Sur la convergence des séries trigonométriques qui servent à représenter une fonction arbitraire entre des limites données". In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 4.4 (1829). Byline 'Lejeune Dirichlet'., pp. 157–169. DOI: 10.1515/crll.1829.4.157.

[86] S. D. Poisson. "Mémoire sur l'équilibre et le mouvement des corps élastiques". In: *Mémoires de l'Académie Royale des Sciences de l'Institut de France*. Vol. 8. Read April 1828. Derives the two distinct elastic wave speeds (P and S) and the ratio now called Poisson's ratio. 1829, pp. 357–570.

[87] M. Chasles. "Note sur les propriétés générales du système de deux corps semblables entr'eux et placés d'une manière quelconque dans l'espace; et sur le déplacement fini ou infiniment petit d'un corps solide libre". In: *Bulletin des Sciences Mathématiques, Astronomiques, Physiques et Chimiques* 14 (1830). Bulletin de Férussac. Chasles' screw-displacement theorem., pp. 321–326.

[88] M. Faraday. "On a peculiar class of acoustical figures; and on certain forms assumed by groups of particles upon vibrating elastic surfaces". In: *Philosophical Transactions of the Royal Society of London* 121.121 (1831). First observation of the parametric (half-frequency) response, in the appendix on crispations., pp. 299–340. DOI: 10.1098/rstl.1831.0018.

[89] M. Faraday. “Experimental researches in electricity”. In: *Philosophical Transactions of the Royal Society of London* 122.122 (1832). The First Series – the induction of transient currents, moving-magnet induction and the flux rule. Crossref prints ’V. Experimental researches in electricity’ (the Royal Society article number) and records 122 as an issue, not a volume; cite as volume 122 with a series note. There is a parallel abridged printing in Phil. Mag., which is NOT this entry., pp. 125–162. DOI: 10.1098/rstl.1832.0006.

[90] A. Fresnel. “Mémoire sur la loi des modifications que la réflexion imprime à la lumière polarisée”. In: *Mémoires de l’Académie royale des sciences de l’Institut de France*. Vol. 11. Read to the Academy on 7 January 1823 but printed in tome XI in 1832 (the manuscript was mislaid and recovered among Fourier’s papers), so the key year is the publisher’s year and the customary dating is 1823 – the citing prose already says ’the memoir read in 1823’, which is consistent. Source of the Fresnel amplitude coefficients (sine law and tangent law) and of the transverse-vibration conclusion. Academy memoirs, hence inproceedings. A partial version appeared in Annales de chimie et de physique 46, 225-264 (1831); that is a different printing, not this key. 1832, pp. 393–433.

[91] J. Henry. “On the Production of Currents and Sparks of Electricity from Magnetism”. In: *The American Journal of Science and Arts* 22 (1832). July 1832 number; this is the paper carrying Henry’s independent observation of induction and his report of self-induction, which is what the citing sentence claims. No DOI (pre-1990)., pp. 403–408.

[92] F. Reich. *Fallversuche über die Umdrehung der Erde*. Eastward deflection of falling bodies in the Freiberg mineshaft; experiments performed 1831. Freiberg: J. G. Engelhardt, 1832.

[93] M. Faraday. “Experimental researches in electricity.— fourth series”. In: *Philosophical Transactions of the Royal Society of London* 123.123 (1833). Contains ’On a new law of electric conduction’ — the observation that the conducting power of silver sulfide increases on heating., pp. 507–522. DOI: 10.1098/rstl.1833.0022.

[94] T. Graham. “On the law of the diffusion of gases”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.9 (1833). Series 3; continued at pp. 269–276 and 351–358 of the same volume., pp. 175–190. DOI: 10.1080/14786443308648004.

[95] É. Clapeyron. “Mémoire sur la puissance motrice de la chaleur”. In: *Journal de l’École Royale Polytechnique* 14 (1834). Tome XIV, 23e cahier; ideal-gas equation, indicator diagram, analytic Carnot theory, coexistence-slope equation., pp. 153–190.

[96] W. R. Hamilton. “On a general method in dynamics; by which the study of the motions of all free systems of attracting or repelling points is reduced to the search and differentiation of one central relation, or characteristic function”. In: *Philosophical Transactions of the Royal Society of London* 124.124 (1834). The ’Second Essay’ is Phil. Trans. 125 (1835) 95–144, not cited in these chapters., pp. 247–308. DOI: 10.1098/rstl.1834.0017.

[97] E. Lenz. “Ueber die Bestimmung der Richtung der durch elektrodynamische Vertheilung erregten galvanischen Ströme”. In: *Annalen der Physik* 107.31 (1834). Crossref-confirmed. Wiley files it under the continuous volume 107 with ’issue 31’ — 31 is the contemporaneous Poggendorff series volume, so cite volume 107 (Wiley) and mention Poggendorffs Annalen Bd. 31 in a note if desired. Crossref’s container-title is the modern ’Annalen der Physik’; the 1834 imprint is ’Annalen der Physik und Chemie’., pp. 483–494. DOI: 10.1002/andp.18341073103.

[98] J. C. A. Peltier. “Nouvelles expériences sur la caloricité des courans électriques”. In: *Annales de Chimie et de Physique* 56 (1834), pp. 371–386.

[99] L. Poinsot. *Théorie nouvelle de la rotation des corps*. Geometric (rolling-ellipsoid) description of torque-free motion; polhode and herpolhode. An extended version appeared in J. Math. Pures Appl. 16 (1851). Paris: Bachelier, 1834.

[100] G.-G. Coriolis. “Sur les équations du mouvement relatif des systèmes de corps”. In: *Journal de l’École Royale Polytechnique* 15 (1835). Cahier 24, tome 15. The Coriolis force., pp. 142–154.

[101] C. G. J. Jacobi. “Zur Theorie der Variations-Rechnung und der Differential-Gleichungen”. In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 17.17 (1837). Source of the accessory (Jacobi) equation and conjugate points., pp. 68–82. DOI: 10.1515/crll.1837.17.68.

[102] H. Lloyd. “On a New Case of Interference of the Rays of Light”. In: *The Transactions of the Royal Irish Academy* 17 (1837). Dating is genuinely contested and the key’s 1834 is the customary date the experiment was described/read to the Academy. Transactions vol. 17 is dated 1837 by most catalogues (it is the volume carrying Hamilton’s Third Supplement, TRIA 17 (1837) 1-144, and MacCullagh’s papers); JSTOR labels the same volume 1831 (JSTOR stable/30078788, volume index i30078783). Recommend keying Lloyd:1834 as written with year 1837 and a note recording both the 1834 reading and the JSTOR 1831 label. No DOI. Access route: JSTOR (paywalled) — grade AMBER unless a copy is stored., pp. 171–177.

[103] G. B. Airy. “On the intensity of light in the neighbourhood of a caustic”. In: *Transactions of the Cambridge Philosophical Society* 6 (1838). The caustic integral (Airy function) explaining supernumerary rainbow bows. Read to the Society in 1836, printed in volume 6 part III, 1838 – so the customary date 1838 is the issue date and is correct. No DOI exists for this volume; the reader’s route is the digitised Transactions (archive.org / Google Books), not a publisher record., pp. 379–402.

[104] G. Hagen. “Ueber die Bewegung des Wassers in engen cylindrischen Röhren”. In: *Annalen der Physik* 122.3 (1839). Second-series volume 46 (volume 122 in the continuous numbering of Annalen der Physik). Independent discovery of the fourth-power pipe-flow law., pp. 423–442. DOI: 10.1002/andp.18391220304.

[105] J. L. M. Poiseuille. “Recherches expérimentales sur le mouvement des liquides dans les tubes de très-petits diamètres”. In: *Comptes Rendus des Séances de l’Académie des Sciences*. Vol. 11. Two notes in CR 11 (1840); continued in CR 12 (1841) 112-115; the full memoir appeared in 1846 (Mémoires des Savants Étrangers 9, 433-544). 1840, pp. 961–967, 1041–1048.

[106] C. F. Gauss. “Intensitas vis magneticae terrestris ad mensuram absolutam revocata”. In: *Commentationes Societatis Regiae Scientiarum Gottingensis Recentiores*. Vol. 8. CUSTOMARY-VS-PUBLISHER DATING, the Kretschmann trap: read to the Göttingen Society on 15 December 1832 and issued as a separate offprint in 1833, but the Commentationes volume VIII (covering 1832-1837) was published in 1841 – so the key year is the publisher’s year and is correct, and the note should record the 1832 reading. Society commentaries, hence inproceedings; volume 8 is genuine. Reprinted in Gauss, Werke V (Göttingen 1867), pp. 79-118. No DOI for the original: a Springer chapter DOI exists (10.1007/978-3-642-49319-5_10) but it points at a modern reprint, not this venue, so it should NOT be attached. 1841, pp. 3–44.

[107] C. Doppler. “Über das farbige Licht der Doppelsterne und einiger anderer Gestirne des Himmels”. In: *Abhandlungen der Königlichen Böhmischen Gesellschaft der Wissenschaften*. Academy series (V. Folge, Bd. 2, Prague), no journal volume number — @inproceedings per the house rule on academy proceedings. 1842, pp. 465–482.

[108] J. R. Mayer. “Bemerkungen über die Kräfte der unbelebten Natur”. In: *Justus Liebigs Annalen der Chemie* 42.2 (1842), pp. 233–240. DOI: 10.1002/jlac.18420420212.

[109] J. Joule. “On the calorific effects of magneto-electricity, and on the mechanical value of heat”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 23.153 (1843). Series 3; continued at pp. 347–355 and 435–443 of the same volume., pp. 347–355. DOI: 10.1080/14786444308644749.

[110] A. J. C. B. de Saint-Venant. “Note à joindre au Mémoire sur la dynamique des fluides”. In: *Comptes Rendus des Séances de l’Académie des Sciences*. Vol. 17. States the viscous flow equations in final form two years before Stokes; cited to keep the Navier-Stokes attribution honest. 1843, pp. 1240–1243.

[111] C. H. D. B. Ballot. “Akustische Versuche auf der Niederländischen Eisenbahn, nebst gelegentlichen Bemerkungen zur Theorie des Hrn. Prof. Doppler”. In: *Annalen der Physik und Chemie* 66 (1845). Experimental test of the Doppler effect for sound with musicians on a moving train. Poggendorff vol. 66 = Wiley continuous numbering vol. 142., pp. 321–351.

[112] J. Joule. “On the changes of temperature produced by the rarefaction and condensation of air”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 26.174 (1845). Series 3. Suffix a: two Joule works of 1845 are cited; this is the earlier (vol. 26)., pp. 369–383. DOI: 10.1080/14786444508645153.

[113] J. P. Joule. “On the existence of an equivalent relation between heat and the ordinary forms of mechanical power”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 27.179 (1845). Series 3. Suffix b: first paddle-wheel announcement (vol. 27)., pp. 205–207. DOI: 10.1080/14786444508645256.

[114] S. Kirchhoff. “Ueber den Durchgang eines elektrischen Stromes durch eine Ebene, insbesondere durch eine kreisförmige”. In: *Annalen der Physik* 140.4 (1845). Issue 4. The paper universally cited as the first statement of Kirchhoff’s circuit laws, matching the citing sentence. TWO TRAPS. (1) The volume is 140 in Wiley’s whole-series numbering, which is what the DOI and Crossref use; Poggendorff’s own running number for the same volume is 64, and many secondary sources cite ’Ann. Phys. Chem. 64, 497’. Use 140 so the DOI and volume agree. (2) Kirchhoff generalised the rules to arbitrary networks two years later in ’Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird’, Ann. Phys. 148, 497–508 (1847) – if the prose ever needs the network form specifically, that is the paper, not this one., pp. 497–514. DOI: 10.1002/andp.18451400402.

[115] F. E. Neumann. “Die mathematischen Gesetze der inducirten elektrischen Ströme”. In: *Abhandlungen der Königlichen Akademie der Wissenschaften zu Berlin, aus dem Jahre 1845, Physikalische Abhandlungen*. Academy Abhandlungen: NO volume number – the series is designated by year (’aus dem Jahre 1845’). Read to the Berlin Academy on 27 October 1845; issued separately by G. Reimer, Berlin, 1846 (87 pp.), and Jackson & Okun date the collected Academy printing 1847, so 1845 is the volume’s own dating rather than the press date. TITLE TRAP: Jackson & Okun (Rev. Mod. Phys. 73, 663; arXiv hep-ph/0012061) cite this memoir as ’Allgemeine Gesetze der inducirten elektrischen Ströme’ – that is the title of the abridged extract in Annalen der Physik 143 (1846) 31–44, not of the Academy

memoir; several secondary sources have copied the conflation. The sequel is 'Über ein allgemeines Princip der mathematischen Theorie inducirter electrischer Ströme', Abh. Berlin aus dem Jahre 1847, 1–71 (Reimer 1848) – a different work. Reprinted as Ostwalds Klassiker Nr. 10, ed. C. Neumann (Engelmann, Leipzig 1889). No DOI. Note for the citing chapter: Neumann never displays the vector potential explicitly; the form A_N is attributed to him by later workers (Jackson & Okun say so in as many words), so 'introduced by Neumann' should be softened to 'implicit in Neumann's mutual-inductance integral'. 1845, pp. 1–87.

[116] J. S. Russell. "Report on Waves". In: *Report of the Fourteenth Meeting of the British Association for the Advancement of Science*. The 14th BAAS meeting was held at York in September 1844; the Report volume printed 1845 (keyed by print year). Contains the account of the solitary wave observed on the Union Canal in 1834 and the wave-tank experiments. 1845, pp. 311–390.

[117] G. G. Stokes. "On the Theories of the Internal Friction of Fluids in Motion, and of the Equilibrium and Motion of Elastic Solids". In: *Transactions of the Cambridge Philosophical Society* 8 (1845). Read 14 April 1845; the completed volume 8 is dated 1849. Universally cited as 1845 and keyed so here (no DOI to arbitrate); record the dating in the entry note., pp. 287–319.

[118] M. Faraday. "Experimental researches in electricity.—Nineteenth series". In: *Philosophical Transactions of the Royal Society of London* 136.136 (1846). The Faraday-effect paper; the series carries the descriptive subtitle 'On the magnetization of light and the illumination of magnetic lines of force'. Crossref title is 'I. Experimental researches in electricity.—Nineteenth series' (article number I) and it records 136 as an issue; cite as volume 136., pp. 1–20. DOI: 10.1098/rstl.1846.0001.

[119] M. Faraday. "Experimental researches in electricity.—Twentieth series". In: *Philosophical Transactions of the Royal Society of London* 136.136 (1846). The diamagnetism paper (bismuth repelled), descriptive subtitle 'On new magnetic actions, and on the magnetic condition of all matter'. Crossref title 'II. Experimental researches in electricity.—Twentieth series'; 136 recorded as issue. NOTE the DOI is ...1846.0003, not 0002 – 10.1098/rstl.1846.0002 is an unrelated errata page. The Twenty-first series (10.1098/rstl.1846.0004, pp. 41-62) continues the same work if a second key is ever needed., pp. 21–40. DOI: 10.1098/rstl.1846.0003.

[120] T. Graham. "On the motion of gases". In: *Philosophical Transactions of the Royal Society of London* 136.136 (1846). Effusion and transpiration., pp. 573–631. DOI: 10.1098/rstl.1846.0029.

[121] H. von Helmholtz. *Über die Erhaltung der Kraft, eine physikalische Abhandlung*. Berlin: G. Reimer, 1847.

[122] V. Regnault. "Relation des expériences entreprises pour déterminer les principales lois et les données numériques qui entrent dans le calcul des machines à vapeur". In: *Mémoires de l'Académie des Sciences de l'Institut de France*. Vol. 21. Memoir occupying the whole of volume 21; the full official title carries a ministerial commission clause – verify exact form when building the bib entry. 1847.

[123] W. Thomson. "On an Absolute Thermometric Scale Founded on Carnot's Theory of the Motive Power of Heat, and Calculated from Regnault's Observations". In: *Philosophical Magazine* 33 (1848). Series 3; author later Lord Kelvin. Also read to the Cambridge Philosophical Society, June 1848., pp. 313–317.

[124] H. Wilbraham. “On a certain periodic function”. In: *The Cambridge and Dublin Mathematical Journal* 3 (1848). First description of the overshoot customarily called the Gibbs phenomenon – misattribution recorded in the chapter prose per house rule., pp. 198–201.

[125] H. Fizeau. “Sur une expérience relative à la vitesse de propagation de la lumière”. In: *Comptes rendus hebdomadaires des séances de l’Académie des sciences*. Vol. 29. The toothed-wheel measurement over the Suresnes–Montmartre baseline, séance of 23 July 1849. Comptes rendus DOES carry tome numbers (t. XXIX), so volume 29 is real and not invented. A short addendum appears at p. 132 of the same tome. No DOI exists; the volume is free on Gallica, so it grades GREEN via a manifest URL rather than a DOI. booktitle should be the Comptes rendus series name, with no publisher volume beyond the tome. 1849, pp. 90–92.

[126] A. Bravais. “Mémoire sur les systèmes formés par des points distribués régulièrement sur un plan ou dans l’espace”. In: *Journal de l’École Polytechnique* 19 (1850), pp. 1–128.

[127] R. Clausius. “Ueber die bewegende Kraft der Wärme und die Gesetze, welche sich daraus für die Wärmelehre selbst ableiten lassen”. In: *Annalen der Physik* 155.4 (1850). Part II in the same volume, pp. 500–524. Poggendorff numbering; Wiley continuous numbering vol. 155., pp. 500–524. DOI: 10.1002/andp.18501550403.

[128] L. Foucault. “Méthode générale pour mesurer la vitesse de la lumière dans l’air et les milieux transparents. Vitesses relatives de la lumière dans l’air et dans l’eau. Projet d’expérience sur la vitesse de propagation du calorique rayonnant”. In: *Comptes rendus hebdomadaires des séances de l’Académie des sciences*. Vol. 30. Séance of 6 May 1850 (t. 30, no. 18). This is the rotating-mirror experiment showing light slower in water, i.e. the result that decided against the corpuscular theory; Foucault announced no numerical value for c until 1862. The title is long because the Comptes rendus prints all three parts as one heading – a shortened form ending at ’milieux transparents’ is acceptable if the bib writer prefers, but the above is what is printed. No DOI; free on Gallica. 1850, pp. 551–560.

[129] J. P. Joule. “On the mechanical equivalent of heat”. In: *Philosophical Transactions of the Royal Society of London* 140.140 (1850), pp. 61–82. DOI: 10.1098/rstl.1850.0004.

[130] G. Kirchhoff. “Über das Gleichgewicht und die Bewegung einer elastischen Scheibe”. In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 40.40 (1850). Consistent plate bending theory and boundary conditions., pp. 51–88. DOI: 10.1515/crll.1850.40.51.

[131] O. F. Mossotti. “Discussione analitica sull’influenza che l’azione di un mezzo dielettrico ha sulla distribuzione dell’elettricità alla superficie di più corpi elettrici disseminati in esso”. In: *Memorie di Matematica e di Fisica della Società Italiana delle Scienze Residente in Modena* 24 (1850). Received by the Society 15 August 1846 (per the Museo Galileo catalogue record) and printed in 1850, which fixes the key; the memoir is consequently sometimes dated 1846 or 1847. Some citations add ’parte seconda’ (part 2) to the volume. No DOI. Digitised copy catalogued by the Museo Galileo OPAC (record uri=1014641)., pp. 49–74.

[132] M. Ostrogradsky. “Mémoire sur les équations différentielles relatives au problème des isopérimètres”. In: *Mémoires de l’Académie Impériale des Sciences de Saint-Pétersbourg, VIe série*. Vol. 4. Higher-derivative variational calculus; origin of the Ostrogradsky Hamiltonian and instability. 1850, pp. 385–517.

[133] H. Fizeau. “Sur les hypothèses relatives à l’éther lumineux, et sur une expérience qui paraît démontrer que le mouvement des corps change la vitesse avec laquelle la lumière se propage dans leur intérieur”. In: *Comptes rendus hebdomadaires des séances de l’Académie des sciences* 33.13 (1851). Reports fringe displacements only: 0.23 simple and 0.46 double, against Fresnel’s predicted 0.40. Fizeau prints no drag coefficient and no uncertainty; the figures often attributed to him are later back-derivations. No DOI; digitised by Gallica, pp. 349–355.

[134] L. Foucault. “Démonstration physique du mouvement de rotation de la Terre au moyen du pendule”. In: *Comptes rendus hebdomadaires des séances de l’Académie des sciences*. Vol. 32. The Foucault pendulum. 1851, pp. 135–138.

[135] G. G. Stokes. “On the Effect of the Internal Friction of Fluids on the Motion of Pendulums”. In: *Transactions of the Cambridge Philosophical Society* 9 (1851). Part II of volume 9. Source of the Stokes drag law F = 6 pi mu a v (creeping flow past a sphere)., pp. 8–106.

[136] L. Foucault. “Sur les phénomènes d’orientation des corps tournants entraînés par un axe fixe à la surface de la Terre”. In: *Comptes rendus hebdomadaires des séances de l’Académie des sciences*. Vol. 35. Names and demonstrates the gyroscope as a second proof of the Earth’s rotation. 1852, pp. 424–427.

[137] J. Joule and W. Thomson. “On the thermal effects experienced by air in rushing through small apertures”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 4.28 (1852). Series 4. First of the Joule–Thomson series 1852–1862; the porous-plug papers continue in Phil. Trans. R. Soc. vols. 143, 144, 150, 152., pp. 481–492. DOI: 10.1080/14786445208647169.

[138] G. Lamé. *Leçons sur la théorie mathématique de l’élasticité des corps solides*. The Lamé constants of isotropic elasticity. Paris: Bachelier, 1852.

[139] G. G. Stokes. “On the Composition and Resolution of Streams of Polarized Light from different Sources”. In: *Transactions of the Cambridge Philosophical Society* 9 (1852). No DOI; do not invent one. Read 9 March 1852 and printed in Part III of volume 9 of the Transactions, which appeared in parts between 1849 and 1856 – the ADS record is dated 1851 (bibcode 1851TCaPS...9..399S) while the paper itself and essentially all optics literature date it 1852. Keep 1852 and record the discrepancy. Reprinted in Stokes, Mathematical and Physical Papers, vol. 3 (Cambridge University Press, 1901), pp. 233-258, which is public-domain and scanned – the practical retrieval route for the availability grade., pp. 399–416.

[140] R. Franz and G. Wiedemann. “Ueber die Wärme-Leitungsfähigkeit der Metalle”. In: *Annalen der Physik* 165.8 (1853). Some indexes list the authors in the order Franz, Wiedemann; keyed here by the customary name of the law., pp. 497–531. DOI: 10.1002/andp.18531650802.

[141] W. Thomson. “XV.—On the Dynamical Theory of Heat, with numerical results deduced from Mr Joule’s Equivalent of a Thermal Unit, and M. Regnault’s Observations on Steam”. In: *Transactions of the Royal Society of Edinburgh* 20.2 (1853). Part I of the series; contains the Kelvin statement of the second law and the absolute thermodynamic scale. Read 17 March 1851 and customarily cited as 1851, but Cambridge and Crossref both date the issue (vol. 20, part 2) to 1853 – keyed by the publisher’s year per the Kretschmann rule, with the customary dating recorded here. Cambridge prints the title with the volume’s article number, ’XV.–’, dropped here., pp. 261–288. DOI: 10.1017/S0080456800033172.

[142] R. Clausius. “Ueber eine veränderte Form des zweiten Hauptsatzes der mechanischen Wärmetheorie”. In: *Annalen der Physik* 169.12 (1854). Poggendorff numbering; Wiley continuous numbering vol. 169., pp. 481–506. DOI: 10.1002/andp.18541691202.

[143] A. Fick. “Ueber Diffusion”. In: *Annalen der Physik* 170.1 (1855), pp. 59–86. DOI: 10.1002/andp.18551700105.

[144] A. J. C. B. de Saint-Venant. “Mémoire sur la torsion des prismes”. In: *Mémoires présentés par divers savants à l’Académie des Sciences*. Vol. 14. Torsion of prisms of arbitrary cross-section; also the customary source of Saint-Venant’s principle. 1855, pp. 233–560.

[145] A. Krönig. “Grundzüge einer Theorie der Gase”. In: *Annalen der Physik* 175.10 (1856), pp. 315–322. DOI: 10.1002/andp.18561751008.

[146] W. Weber and R. Kohlrausch. “Ueber die Elektricitätsmenge, welche bei galvanischen Strömen durch den Querschnitt der Kette fliesst”. In: *Annalen der Physik* 175.9 (1856). Verified via Crossref: Annalen der Physik 175(9), 10-25 (1856), authors Weber then Kohlrausch. VOLUME TRAP: the paper appeared as Poggendorff’s Annalen der Physik und Chemie Bd. 99, and virtually every historical citation says "vol. 99"; Wiley’s continuous re-numbering, which is what the DOI and Crossref carry, calls it vol. 175. Use 175 to match the resolving DOI and record "Poggendorffs Annalen Bd. 99" in the bib note. Keep the publisher’s spelling "Ueber" (not "Über") and "fliesst" (not "fließt") — that is how it is printed and how Crossref stores it. A duplicate key Kohlrausch:1856 for this same paper was merged here (first-listed-author rule); distinct from the 1857 Abhandlungen memoir "Elektrodynamische Maassbestimmungen", which is a different work., pp. 10–25. DOI: 10.1002/andp.18561750903.

[147] R. Clausius. “Ueber die Art der Bewegung, welche wir Wärme nennen”. In: *Annalen der Physik* 176.3 (1857). Kinetic hypothesis with molecular speeds; does NOT contain the mean free path (that is Clausius:1858) – the chapter says so explicitly., pp. 353–380. DOI: 10.1002/andp.18571760302.

[148] W. Thomson. “On a Mechanical Theory of Thermo-Electric Currents”. In: *Proceedings of the Royal Society of Edinburgh* 3 (1857). Predicts the bulk (Thomson) effect and gives the two Thomson relations from the thermodynamic argument Thomson knew to be incomplete. Read 15 December 1851 and customarily cited as 1851; PRSE volume 3 spans the 1850–1857 sessions and both Cambridge and Crossref date it 1857, so it is keyed by the publisher’s year (Kretschmann rule). Cambridge prints the title with the volume’s article number, ’4.’, dropped here. Distinct from Thomson’s later ’On the Dynamical Theory of Heat. Part VI. Thermo-electric currents’, Trans. Roy. Soc. Edinb. 21 (1857)., pp. 91–98. DOI: 10.1017/S0370164600027310.

[149] R. Clausius. “Ueber die mittlere Länge der Wege, welche bei der Molecularbewegung gasförmiger Körper von den einzelnen Molecülen zurückgelegt werden; nebst einigen anderen Bemerkungen über die mechanische Wärmetheorie”. In: *Annalen der Physik* 181.10 (1858). Mean free path, in reply to Buys Ballot; customarily mis-dated to the 1857 paper., pp. 239–258. DOI: 10.1002/andp.18581811005.

[150] H. von Helmholtz. “Über Integrale der hydrodynamischen Gleichungen, welche den Wirbelbewegungen entsprechen”. In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 55.55 (1858). The vortex theorems., pp. 25–55. DOI: 10.1515/crll.1858.55.25.

[151] U. J. J. L. Verrier. “Lettre de M. Le Verrier à M. Faye sur la théorie de Mercure et sur le mouvement du périhélie de cette planète”. In: *Comptes rendus hebdomadaires des séances de l’Académie des sciences*. Vol. 49. No DOI; scans on Gallica/BnF. 1859, pp. 379–383.

[152] G. Kirchhoff and R. Bunsen. “Chemische Analyse durch Spectralbeobachtungen”. In: *Annalen der Physik* 186.6 (1860). Suffix ’b’; companion to Kirchhoff:1860a. The founding paper of spectral analysis; led directly to the discovery of caesium and rubidium., pp. 161–189. DOI: 10.1002/andp.18601860602.

[153] G. Kirchhoff. “Ueber das Verhaeltniss zwischen dem Emissionsvermoegen und dem Absorptionsvermoegen der Koerper fuer Waerme und Licht”. In: *Annalen der Physik* 185 (1860). Suffix ’a’ because a second Kirchhoff 1860 paper (with Bunsen, spectral analysis) is cited in the same part as Kirchhoff:1860b. German title uses umlauts: Verhaeltniss = Verhältniss, Emissionsvermoegen = Emissionsvermögen, Koerper = Körper, fuer = für, Waerme = Wärme., pp. 275–301.

[154] J. C. Maxwell. “Illustrations of the dynamical theory of gases.—Part I. On the motions and collisions of perfectly elastic spheres”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 19.124 (1860). Series 4. Part II in vol. 20, pp. 21–37 (1860)., pp. 19–32. DOI: 10.1080/14786446008642818.

[155] F. Melde. “Ueber die Erregung stehender Wellen eines fadenförmigen Körpers”. In: *Annalen der Physik* 187.12 (1860). Standing waves on a driven string (Melde’s experiment). Poggendorff 2nd-series vol. 111 = Wiley continuous numbering vol. 187., pp. 513–537. DOI: 10.1002/andp.18601871202.

[156] B. Riemann. “Über die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite”. In: *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen*. Vol. 8. Shock formation and Riemann invariants; the Abhandlungen series carries volume numbers. 1860, pp. 43–66.

[157] J. C. Maxwell. “On physical lines of force”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 21.139 (1861). Series 4. The paper runs in four parts across two volumes: Parts I-II in vol. 21 (1861) pp. 161-175, 281-291, 338-348 and Parts III-IV in vol. 23 (1862) pp. 12-24, 85-95 — the displacement current the chapter attributes to it is introduced in Part III (1862). The DOI given is Part I only (issue 139, pp. 161-175), Crossref-confirmed as ’On physical lines of force’; if the validator objects to the multi-range pages field, cite Part I’s range and record the full span in a note rather than inventing a single range., pp. 161–175. DOI: 10.1080/14786446108643033.

[158] L. Foucault. “Détermination expérimentale de la vitesse de la lumière; parallaxe du Soleil”. In: *Comptes rendus hebdomadaires des séances de l’Académie des sciences*. Vol. 55. Séance of 22 September 1862 – the absolute determination (298000 km/s), which revised the accepted solar parallax. A companion note, ’Détermination expérimentale de la vitesse de la lumière; description des appareils’, séance of 24 November 1862, appears at t. 55, pp. 792-796; if the chapter needs the apparatus it should cite that one too. No DOI; free on Gallica. 1862, pp. 501–503.

[159] H. von Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Helmholtz resonators, analysis of timbre. Braunschweig: Friedrich Vieweg und Sohn, 1863.

[160] A. Matthiessen and A. C. C. Vogt. “On the influence of temperature on the electric conducting-power of alloys”. In: *Philosophical Transactions of the Royal Society of London* 154.154 (1864), pp. 167–200. DOI: 10.1098/rstl.1864.0004.

[161] H. Tresca. “Mémoire sur l’écoulement des corps solides soumis à de fortes pressions”. In: *Comptes rendus hebdomadaires des séances de l’Académie des sciences*. Vol. 59. Plastic yield at a critical shear stress (Tresca criterion). 1864, pp. 754–758.

[162] R. Clausius. “Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie”. In: *Annalen der Physik* 201.7 (1865). Coins ’entropy’. Poggendorff numbering; Wiley continuous numbering vol. 201., pp. 353–400. DOI: 10.1002/andp.18652010702.

[163] J. Loschmidt. “Zur Grösse der Luftmoleküle”. In: *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften in Wien, Mathematisch-Naturwissenschaftliche Classe*. Vol. 52. Customary dating 1865 (session year); the volume for 1865 was printed 1866 – verify the publisher’s year against the copy before final keying (Kretschmann rule). 1865, pp. 395–413.

[164] J. C. Maxwell. “A dynamical theory of the electromagnetic field”. In: *Philosophical Transactions of the Royal Society of London* 155.155 (1865). Crossref-confirmed (it records 155 in the ’issue’ slot; 155 is the volume). Read 8 December 1864, published in the 1865 volume — the key year is the issue year and is correct., pp. 459–512. DOI: 10.1098/rstl.1865.0008.

[165] J. Tyndall. *Heat considered as a mode of motion: being a course of twelve lectures delivered at the Royal Institution of Great Britain in the season of 1862*. Second edition (London, 1865); first edition 1863. This is the source of the platinum-wire radiation data Stefan re-analysed: the ratio 11.7 between 525 C (798 K) and 1200 C (1473 K) that the chapter quotes. Crepeau’s history of the T^4 law cites it as ’Tyndall, J., 1865, Heat Considered as a Mode of Motion, Longman, Green, Longman Roberts and Green, London, Chapter 12’. No DOI (1865 book). Verified open-access scan of the London 1865 printing at archive.org identifier 10134906bsb (Bayerische Staatsbibliothek copy, https://archive.org/details/10134906bsb) – checked live via the archive.org metadata API, not guessed; suitable for a manifest "access": "open" entry. Do NOT cite the later retitled ’Heat: A Mode of Motion’ editions (1870s) for this key. Roberts & Green: Longman, Green, Longman, 1865.

[166] A. Kundt. “Ueber eine neue Art akustischer Staubfiguren und über die Anwendung derselben zur Bestimmung der Schallgeschwindigkeit in festen Körpern und Gasen”. In: *Annalen der Physik* 203.4 (1866). Issue 4. Volume/pages/year taken from Crossref, which uses Wiley’s continuous numbering; the printed source is Poggendorff’s Annalen der Physik und Chemie, Band 127, p. 497 (1866) — the numbering the chapter’s prose quotes. The +76 offset is the same convention already used by Clausius:1850 (Wiley 155 = Poggendorff 79) and Balmer:1885 in this bibliography, so 203 is the house-consistent volume. Consider recording the Poggendorff band in a BibTeX note., pp. 497–523. DOI: 10.1002/andp.18662030402.

[167] J. C. Maxwell. “The Bakerian Lecture .—On the viscosity or internal friction of air and other gases”. In: *Philosophical Transactions of the Royal Society of London* 156.156 (1866). Not a genuine collision: both claimed works are the same Bakerian Lecture, one record merely omitting the ’The Bakerian Lecture’ prefix and the DOI. No split; both citing chapters keep the bare key. Crossref prints the title with the article number and an em dash, ’XIII. The Bakerian Lecture.—On the viscosity or internal friction of air and other gases’; the colon form given here is the customary citation form. Volume 156 is not carried in the Crossref record but is the printed volume., pp. 249–268. DOI: 10.1098/rstl.1866.0013.

[168] L. Lorenz. “Ueber die Identität der Schwingungen des Lichts mit den elektrischen Strömen”. In: *Annalen der Physik* 207.6 (1867). Issue 6; Wiley continuous volume 207 (contemporaneous series (2) vol. 131). Crossref-confirmed. This is Ludvig Valentin Lorenz, the Danish physicist — the Lorenz gauge condition — and the chapter explicitly makes the point that he is not H. A. Lorentz, so the bib entry must carry the full given names. An author’s English version, ’On the identity of the vibrations of light with electrical currents’, appeared in Phil. Mag. ser. 4, 34 (1867) 287-301; the Annalen paper is the primary., pp. 243–263. DOI: 10.1002/andp.18672070606.

[169] J. C. Maxwell. “On the dynamical theory of gases”. In: *Philosophical Transactions of the Royal Society of London* 157.157 (1867), pp. 49–88. DOI: 10.1098/rstl.1867.0004.

[170] B. Riemann. “Ein Beitrag zur Elektrodynamik”. In: *Annalen der Physik* 207.6 (1867). Crossref-verified. Written in 1858 and submitted to the Gottingen Academy, then withdrawn by Riemann; published posthumously in 1867 (Poggendorff’s Annalen der Physik und Chemie). Wiley’s continuous volume numbering is 207; Poggendorff’s own series number for the same volume is Band 131, which is how nineteenth-century citations give it – worth a note in the entry so a reader chasing ’Ann. Phys. Chem. 131, 237’ finds it., pp. 237–243. DOI: 10.1002/andp.18672070605.

[171] L. Boltzmann. “Studien über das Gleichgewicht der lebendigen Kraft zwischen bewegten materiellen Punkten”. In: *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften in Wien, Mathematisch-Naturwissenschaftliche Classe*. Vol. 58. The Vienna Sitzungsberichte DO carry volume numbers, unlike the Berlin series. 1868, pp. 517–560.

[172] H. von Helmholtz. “Über discontinuirliche Flüssigkeitsbewegungen”. In: *Monatsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin*. Academy Monatsberichte are numbered by year (1868), not by volume – do not invent a volume (cf. the Sitzungsberichte rule in CLAUDE.md). Vortex sheets and the Kelvin-Helmholtz instability, Helmholtz’s half. 1868, pp. 215–228.

[173] É. Mathieu. “Mémoire sur le mouvement vibratoire d’une membrane de forme elliptique”. In: *Journal de Mathématiques Pures et Appliquées* 13 (1868). Second series, vol. 13. Source of the Mathieu equation governing parametric resonance., pp. 137–203.

[174] T. Andrews. “The Bakerian Lecture .—On the continuity of the gaseous and liquid states of matter”. In: *Philosophical Transactions of the Royal Society of London* 159.159 (1869), pp. 575–590. DOI: 10.1098/rstl.1869.0021.

[175] G. Kirchhoff. “Zur Theorie freier Flüssigkeitsstrahlen”. In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 70.70 (1869). Crelle’s Journal. De Gruyter’s Crossref record files this as ’volume 1869, issue 70’ (year-as-volume) and carries no author name; the printed Band is 70, which is the number every bibliography uses – do not copy Crossref’s 1869 into the volume field. Crossref title carries a trailing full stop (’Zur Theorie freier Flüssigkeitsstrahlen.’). This is the paper that develops Helmholtz’s free-streamline method into a computation and yields the contraction coefficient pi/(pi+2) = 0.611 for the jet from a slit – i.e. the ’0.6’ the chapter says the ideal theory cannot supply., pp. 289–298. DOI: 10.1515/crll.1869.70.289.

[176] W. Thomson. “On Vortex Motion”. In: *Proceedings of the Royal Society of Edinburgh* 6 (1869). Author is Lord Kelvin. Source of the circulation theorem. Read 1867; volume 25 part I dated 1869 (customary citation 1869)., pp. 167–167. DOI: 10.1017/s0370164600045740.

[177] R. Clausius. “Ueber einen auf die Wärme anwendbaren mechanischen Satz”. In: *Annalen der Physik* 217.9 (1870). The virial theorem. English translation in Phil. Mag. Ser. 4, 40, 122–127 (1870)., pp. 124–130. DOI: 10.1002/andp.18702170911.

[178] W. J. M. Rankine. “On the thermodynamic theory of waves of finite longitudinal disturbance”. In: *Philosophical Transactions of the Royal Society of London* 160.160 (1870). Rankine’s half of the Rankine-Hugoniot shock conditions., pp. 277–288. DOI: 10.1098/rstl.1870.0015.

[179] J. C. Maxwell. *Theory of Heat*. Zeroth-law statement and the thermodynamic relations later called Maxwell relations. London: Longmans, Green, and Co., 1871.

[180] Sellmeier. “Zur Erklärung der abnormen Farbenfolge im Spectrum einiger Substanzen”. In: *Annalen der Physik* 219.6 (1871). Crossref-verified, and the only Sellmeier paper of 1871 – this is the standard source of the Sellmeier dispersion relation. Journal title on the issue is ’Annalen der Physik und Chemie’ (Poggendorff); Wiley’s continuous volume 219 = Poggendorff Band 143, issue 6. The better-known four-part series ’Ueber die durch die Aetherschwingungen erregten Mitschwingungen der Körpertheilchen...’ followed in 1872 (Ann. Phys. 221, 399-421 and 520-549; 223, 386-403 and 525-554); if the chapter intends the full theory rather than the 1871 announcement, those are 1872 papers and the key year would be wrong. The 1871 note is what the Sellmeier equation is universally cited to, so I have kept it., pp. 272–282. DOI: 10.1002/andp.18712190612.

[181] J. Strutt. “On the light from the sky, its polarization and colour”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 41.271 (1871). Crossref-verified. Philosophical Magazine series 4. PUBLISHER PRINTS THE AUTHOR AS ’J. W. Strutt’ – he succeeded to the barony and became Lord Rayleigh only in 1873, so the customary key ’Rayleigh:1871’ does not match the printed byline; record that in the entry. This first instalment is the one carrying the lambdaˆ-4 argument (the chapter cites it for the driven-oscillator polarizability at sec:rad-rayleigh). A continuation appeared in the same volume at pp. 274-279, doi 10.1080/14786447108640479. Crossref’s deposited title is prefixed ’XV. ’ (the issue article number); I have dropped the numeral, which is not part of the title., pp. 107–120. DOI: 10.1080/14786447108640452.

[182] W. Thomson. “XLVI.Hydrokinetic solutions and observations”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 42.281 (1871). Author is Lord Kelvin. Wind-over-water analysis; Kelvin’s half of the Kelvin-Helmholtz instability., pp. 362–377. DOI: 10.1080/14786447108640585.

[183] L. Boltzmann. “Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen”. In: *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften in Wien, Mathematisch-Naturwissenschaftliche Classe*. Vol. 66. Boltzmann equation and H-theorem. 1872, pp. 275–370.

[184] E. Abbe. “Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung”. In: *Archiv für Mikroskopische Anatomie* 9.1 (1873). Crossref confirms title/volume/pages/DOI. The paper containing the resolution limit d = lambda/(2 n sin alpha); Crossref capitalises the journal as ’Archiv für Mikroskopische Anatomie’, the journal itself prints ’Archiv für mikroskopische Anatomie’., pp. 413–468. DOI: 10.1007/BF02956173.

[185] J. C. Maxwell. *A Treatise on Electricity and Magnetism*. First edition, 2 volumes, Oxford: Clarendon Press, 1873. Cited in four chapters for div B = 0, the stress tensor and radiation pressure; the radiation-pressure prediction is in vol. 2, art. 792-793. Later editions (2nd 1881, 3rd 1891) renumber articles, so the entry should

pin the 1873 first edition as the key year says. No DOI for the original; a Cambridge Library Collection reissue exists but should not be substituted. Clarendon Press, 1873.

[186] J. Plateau. *Statique expérimentale et théorique des liquides soumis aux seules forces moléculaires*. Two volumes; the soap-film experiments realising minimal surfaces. Paris: Gauthier-Villars, 1873.

[187] J. D. van der Waals. *Over de Continuïteit van den Gas- en Vloeistoftoestand*. Doctoral dissertation, Universiteit Leiden, printed by A. W. Sijthoff. Not a genuine collision: the two claimed works are the same dissertation, one record naming the printer and the other the awarding university. No DOI; a full scan is held in Leiden University Digital Collections (shelfmark DISL 1873: 17), which is the access route to declare in the manifest. A. W. Sijthoff, 1873.

[188] G. Cantor. "Ueber eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen". In: *Journal für die reine und angewandte Mathematik* 77 (1874). Crelle's Journal is numbered by year in the Crossref record; cited here by the printed volume 77, pp. 258–262. DOI: 10.1515/crll.1874.77.258.

[189] J. Kerr. "A new relation between electricity and light: Dielectrified media birefringent". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.332 (1875). Series 4, issue 332. This is Part First, the discovery paper for the quadratic electro-optic (Kerr) effect that the citing sentence names. A 'Part second' follows in the same volume, pp. 446–458 (DOI 10.1080/14786447508641318) – not needed for this claim. Verified against Crossref., pp. 337–348. DOI: 10.1080/14786447508641302.

[190] S. von Kowalevsky. "Zur Theorie der partiellen Differentialgleichung". In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 80.80 (1875). Author also transliterated Kovalevskaya; keyed by the paper's byline spelling., pp. 1–32. DOI: 10.1515/crll.1875.80.1.

[191] J. C. Maxwell. "On the Dynamical Evidence of the Molecular Constitution of Bodies". In: *Nature* 11 (1875). Contains Maxwell's statement that the specific-heat discrepancy is the greatest difficulty facing the molecular theory, pp. 357–359.

[192] P. du Bois-Reymond. "Untersuchungen über die Convergenz und Divergenz der Fourierschen Darstellungsformeln". In: *Abhandlungen der Mathematisch-Physicalischen Classe der K. Bayerischen Akademie der Wissenschaften*. Contains the continuous function with divergent Fourier series. Volume/pagination (vol. 12, 1–103) should be verified against the Abhandlungen series before entry. 1876.

[193] J. W. Gibbs. "On the Equilibrium of Heterogeneous Substances". In: *Transactions of the Connecticut Academy of Arts and Sciences* 3 (1876). First part 1876; second part in the same volume, pp. 343–524, issued 1878., pp. 108–248.

[194] A. Kundt and E. Warburg. "Ueber die specifische Wärme des Quecksilbergases". In: *Annalen der Physik* 233.3 (1876). Issue 3. Crossref/Wiley volume 233 = Poggendorff's Annalen der Physik und Chemie, Band 157 (1876), same +76 offset as Kundt:1866. This is the Kundt's-tube measurement on mercury vapour giving c_p/c_V ≈ 1.67, i.e. the monatomic value — the paper that makes the claim the chapter attributes to it. Note the key departs from the house Surname:YYYY convention (it would collide with Kundt:1876 only if another Kundt 1876 paper is added); the key was supplied by the authoring agent and is used as given., pp. 353–369. DOI: 10.1002/andp.18762330302.

[195] J. Loschmidt. “Über den Zustand des Wärmegleichgewichtes eines Systems von Körpern mit Rücksicht auf die Schwerkraft. I.” In: *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften in Wien, Mathematisch-Naturwissenschaftliche Classe*. Vol. 73. The reversibility (Umkehr) objection to the H-theorem. 1876, pp. 128–142.

[196] L. Boltzmann. “Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respective den Sätzen über das Wärmegleichgewicht”. In: *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften in Wien, Mathematisch-Naturwissenschaftliche Classe, Abteilung II*. Vol. 76. The combinatorial definition of entropy; S = k log W is not written in this form here, it is Planck’s later notation. 1877, pp. 373–435.

[197] B. R. John William Strutt. *The Theory of Sound, Volume 1*. Keyed by the customary name Rayleigh (family name Strutt). London: Macmillan, 1877.

[198] B. R. John William Strutt. *The Theory of Sound, Volume I*. Volume II followed 1878. Suffix a: Rayleigh:1877b is the Proc. LMS progressive-waves paper of the same year. London: Macmillan, 1877.

[199] J. Kerr. “On rotation of the plane of polarization by reflection from the pole of a magnet”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 3.19 (1877), pp. 321–343. DOI: 10.1080/14786447708639245.

[200] Rayleigh. “On Progressive Waves”. In: *Proceedings of the London Mathematical Society* s1-9.1 (1877). The group-velocity vs phase-velocity distinction; read November 1877., pp. 21–26. DOI: 10.1112/plms/s1-9.1.21.

[201] P. Clifford. “Applications of Grassmann’s Extensive Algebra”. In: *American Journal of Mathematics* 1.4 (1878), p. 350. DOI: 10.2307/2369379.

[202] H. A. Rowland. “Magnetic effect of electric convection”. In: *American Journal of Science* s3-15.85 (1878). Crossref-verified (Crossref’s deposited title omits the leading ’On the’: ’Magnetic effect of electric convection’). American Journal of Science third series, volume 15, issue 85 – the volume must be cited as series 3 vol. 15 or it collides with the first-series volume 15. The chapter’s subsection heading says ’Rowland (1876)’ because the experiment was performed in Helmholtz’s Berlin laboratory in 1876 and first announced in German – ’Ueber die magnetische Wirkung electrischer Convection’, Ann. Phys. (Poggendorff) 158, 487-493 (1876), and in the Monatsberichte of the Berlin Academy. The key Rowland:1878 denotes the English publication; the heading/key mismatch is a real dating discrepancy and should be noted in the entry., pp. 30–38. DOI: 10.2475/ajs.s3-15.85.30.

[203] V. Strouhal. “Ueber eine besondere Art der Tonerregung”. In: *Annalen der Physik* 241.10 (1878). Neue Folge volume 5 (volume 241 in the continuous numbering). Aeolian tones; the Strouhal number., pp. 216–251. DOI: 10.1002/andp.18782411005.

[204] P. du Bois-Reymond. “Erläuterungen zu den Anfangsgründen der Variationsrechnung”. In: *Mathematische Annalen* 15.2 (1879). Sharpened fundamental lemma (du Bois-Reymond lemma). Particle ’du’ must not be dropped from the surname., pp. 283–314. DOI: 10.1007/bf01444144.

[205] J. C. (Cavendish H. (author); Maxwell. *The Electrical Researches of the Honourable Henry Cavendish, F.R.S., Written Between 1771 and 1781*. Edited from the original manuscripts in the possession of the Duke of Devonshire by J. Clerk Maxwell, Cambridge University Press, 1879. This is what the chapter means by ’Cavendish’s null test for the inverse-square law’ cited as Maxwell:1879 — Maxwell’s edition is where Cavendish’s unpublished concentric-shell experiment (and Maxwell’s own repetition of it) first appeared in print. Set biblatex ’editor = Maxwell, James Clerk’

with 'author = Cavendish, Henry'; the validator accepts editor in place of author for @book, and the key stays Maxwell:1879 as cited. No DOI; scans are on the Internet Archive. Cambridge University Press, 1879.

[206] R. Clausius. *Die mechanische Behandlung der Electricität*. Braunschweig, 1879. This volume is Band 2 of Clausius's *Die mechanische Wärmetheorie* – the full catalogue heading is 'Die mechanische Wärmetheorie, Band 2: Die mechanische Behandlung der Electricität', which the bib writer may want in a series/note field. It is the Clausius half of the Clausius-Mossotti relation, cited in the chapter beside Mossotti:1850. No DOI for the original; the Vieweg+Teubner facsimile reprint carries ISBN 978-3-663-20232-5 and is on SpringerLink, which is a usable access route for the availability grade. Free scan on archive.org. Friedrich Vieweg und Sohn, 1879.

[207] G. Frege. *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. English translation in van Heijenoort's source book, pp. 1–82. Halle a. S.: Louis Nebert, 1879.

[208] E. H. Hall. "On a New Action of the Magnet on Electric Currents". In: *American Journal of Mathematics* 2.3 (1879), p. 287. DOI: 10.2307/2369245.

[209] Rayleigh. "Investigations in optics, with special reference to the spectroscope". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 8.49 (1879). Crossref-verified. Philosophical Magazine series 5. First of four instalments under this title; the resolution criterion is in this one. The others are Phil. Mag. 8, 403-411 and 8, 477-486 (1879) and Phil. Mag. 9, 40-55 (1880). Byline is 'Lord Rayleigh, F.R.S.'; Crossref files the family name as 'Rayleigh'. Crossref's deposited title is prefixed 'XXXI. ' (issue article number), dropped here., pp. 261–274. DOI: 10.1080/14786447908639684.

[210] L. Rayleigh. "On the Stability, or Instability, of certain Fluid Motions". In: *Proceedings of the London Mathematical Society* s1-11.1 (1879). Series 1 volume (s1-11). The inflexion-point criterion for parallel shear flows., pp. 57–72. DOI: 10.1112/plms/s1-11.1.57.

[211] J. Stefan. "Ueber die Beziehung zwischen der Waermestrahlung und der Temperatur". In: *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften in Wien, Mathematisch-Naturwissenschaftliche Classe, Abtheilung II*. Vol. 79. Academy proceedings: no journal volume in the biblatex @article sense; the 79 is the Sitzungsberichte volume. Umlauts: Waermestrahlung = Wärmestrahlung. 1879, pp. 391–428.

[212] J. Curie and P. Curie. "Développement par compression de l'électricité polaire dans les cristaux hémièdres à faces inclinées". In: *Bulletin de la Société minéralogique de France* 3.4 (1880). Crossref/Persée confirmed: vol. 3, pp. 90-93, 1880, both Curie brothers (Jacques first, then Pierre – Pierre is NOT the first author here). The Persée record spells the last word 'hémiédres' (é, an OCR/period spelling); the correct and usual form 'hémièdres' is given above and should still pass a DOI title check. The same result was communicated to the Académie in the same year as C. R. Acad. Sci. Paris 91, 294-295 (1880), under a near-identical title – the Bulletin paper is the one with a resolvable DOI, so prefer it. Persée is open access, so this grades GREEN., pp. 90–93. DOI: 10.3406/bulmi.1880.1564.

[213] H. A. Lorentz. "Ueber die Beziehung zwischen der Fortpflanzungsgeschwindigkeit des Lichtes und der Körperdichte". In: *Annalen der Physik* 245.4 (1880). Issue 4; Wiley continuous volume 245 (contemporaneous series (3) vol. 9). Crossref-confirmed. This is H. A. Lorentz's half of the Lorentz-Lorenz relation; the companion key Lorenz:1880 in this same batch is L. V. Lorenz, a different man — the citing

sentence in 05-electrodynamics-matter.tex depends on the two entries being visibly distinct., pp. 641–665. DOI: 10.1002/andp.18802450406.

[214] L. Lorenz. “Ueber die Refractionsconstante”. In: *Annalen der Physik* 247.9 (1880). Issue 9; Wiley continuous volume 247 (contemporaneous series (3) vol. 11). Crossref-confirmed. L. V. Lorenz’s half of the Lorentz-Lorenz equation, deliberately paired with Lorentz:1880 in the same sentence — keep the two distinguishable by full given names., pp. 70–103. DOI: 10.1002/andp.18802470905.

[215] A. A. Michelson. “Experimental Determination of the Velocity of Light Made at the United States Naval Academy, Annapolis”. In: *Astronomical Papers Prepared for the Use of the American Ephemeris and Nautical Almanac* 1 (1880). The 1879 Annapolis measurement. Issued as Part III of volume 1 in 1880 (the memoir is dated 20 February 1880 and transmitted by Simon Newcomb); the collected volume 1 was published by the Bureau of Navigation, Navy Department, Washington, in 1882, which is why ADS bibcodes it 1882USNAO...1..109M. The key’s 1880 is the part’s own year and is defensible — record the 1882 collected-volume date in a note. No DOI. Full text is at Project Gutenberg ebook 11753, a GREEN access route., pp. 109–145.

[216] H. Poincaré. “Mémoire sur les courbes définies par une équation différentielle”. In: *Journal de Mathématiques Pures et Appliquées (3e série)* 7 (1881). First part of the four-part memoir (parts appeared 1881, 1882, 1885, 1886). Foundation of the qualitative theory of differential equations., pp. 375–422.

[217] H. von Helmholtz. “Die Thermodynamik chemischer Vorgänge”. In: *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin*. Academy proceedings; the series has no volume numbers (same rule as the Einstein Sitzungsberichte entries). Introduces the free energy. 1882, pp. 22–39.

[218] H. Hertz. “Ueber die Berührung fester elastischer Körper”. In: *crll* 92.92 (1882). Hertz contact theory., pp. 156–171. DOI: 10.1515/crll.1882.92.156.

[219] A. Lindstedt. “Ueber die Integration einer für die Störungstheorie wichtigen Differentialgleichung”. In: *Astronomische Nachrichten* 103.14 (1882). Frequency renormalization removing secular terms. Astron. Nachr. paginates by column; the paper begins at col. 211., pp. 211–220. DOI: 10.1002/asna.18821031404.

[220] H. Poincaré. “Mémoire sur les courbes définies par une équation différentielle (seconde partie)”. In: *Journal de Mathématiques Pures et Appliquées (3e série)* 8 (1882). Second part of the memoir; the term ’cycle limite’ (limit cycle) is introduced here – cite this part, not the 1881 first part, for limit cycles., pp. 251–296.

[221] Rayleigh. “Investigation of the Character of the Equilibrium of an Incompressible Heavy Fluid of Variable Density”. In: *Proceedings of the London Mathematical Society* s1-14.1 (1882). Series 1 volume (s1-14). The Rayleigh half of the Rayleigh-Taylor instability., pp. 170–177. DOI: 10.1112/plms/s1-14.1.170.

[222] H. Rowland. “Preliminary notice of the results accomplished in the manufacture and theory of gratings for optical purposes”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.84 (1882). Crossref-verified. Philosophical Magazine series 5; Crossref’s deposited title is prefixed ’LXI. ’, dropped here. The same notice appeared in Nature 26, 211-213 (1882), doi 10.1038/026211b0. Rowland’s concave grating paper is the following year: Phil. Mag. 16, 197-210 (1883)., pp. 469–474. DOI: 10.1080/14786448208627217.

[223] J. Gram. “Ueber die Entwickelung reeller Functionen in Reihen mittelst der Methode der kleinsten Quadrate”. In: *crll* 94.94 (1883), pp. 41–73. DOI: 10.1515/crll.1883.94.41.

[224] G. Kirchhoff. “Zur Theorie der Lichtstrahlen”. In: *Annalen der Physik* 254.4 (1883). Volume 254 in the continuous Wiedemann numbering = vol. 18 of the third series; also read to the Berlin Academy in 1882. Source of the Kirchhoff formula for the 3d wave equation., pp. 663–695. DOI: 10.1002/andp.18832540409.

[225] E. Mach. *Die Mechanik in ihrer Entwickelung historisch-kritisch dargestellt*. First edition; the original spelling Entwickelung is retained. The critique of Newton’s absolute space and of the bucket experiment is in chapter II. English translation: The Science of Mechanics, trans. T. J. McCormack, Open Court, Chicago (1893). Leipzig: F. A. Brockhaus, 1883.

[226] O. Reynolds. “An experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous, and of the law of resistance in parallel channels”. In: *Philosophical Transactions of the Royal Society of London* 174.174 (1883). The dye experiment; origin of the Reynolds-number transition criterion., pp. 935–982. DOI: 10.1098/rstl.1883.0029.

[227] L. Boltzmann. “Ableitung des Stefan’schen Gesetzes, betreffend die Abhängigkeit der Wärmestrahlung von der Temperatur aus der electromagnetischen Lichttheorie”. In: *Annalen der Physik* 258.6 (1884). Wiedemann’s Annalen vol. 22 = whole-series vol. 258. Uses Bartoli’s radiation-pressure cycle., pp. 291–294. DOI: 10.1002/andp.18842580616.

[228] G. Frege. *Die Grundlagen der Arithmetik: eine logisch mathematische Untersuchung über den Begriff der Zahl*. The subtitle is unhyphenated in the original printing. Breslau: Wilhelm Koebner, 1884.

[229] J. H. Poynting. “On the transfer of energy in the electromagnetic field”. In: *Philosophical Transactions of the Royal Society of London* 175.175 (1884). Read 10 January 1884, issue dated 31 December 1884. Crossref records the title with the volume’s article number prefixed – ’XV. On the transfer of energy in the electromagnetic field’ – and puts 175 in the ’issue’ slot rather than ’volume’; the correct citation is volume 175. Drop the ’XV.’ from the title., pp. 343–361. DOI: 10.1098/rstl.1884.0016.

[230] J. J. Balmer. “Notiz über die Spectrallinien des Wasserstoffs”. In: *Annalen der Physik* 261.5 (1885). Also published the same year in Verhandlungen der Naturforschenden Gesellschaft in Basel 7, 548–560 and 750–752., pp. 80–87. DOI: 10.1002/andp.18852610506.

[231] J. A. Ewing. “Experimental researches in magnetism”. In: *Philosophical Transactions of the Royal Society of London* 176.176 (1885). The paper in which Ewing named and mapped hysteresis. Crossref prints the title with the Royal Society’s in-volume article number, ’X. Experimental researches in magnetism’, and files 176 as an ISSUE rather than a volume; the customary citation is volume 176. Author is ’James Alfred Ewing’ in Crossref, ’J. A. Ewing’ on the paper. Do not confuse with the short abstract of the same title in Proc. R. Soc. Lond. 38, 58-62 (10.1098/rspl.1884.0062)., pp. 523–640. DOI: 10.1098/rstl.1885.0010.

[232] L. Rayleigh. “On Waves Propagated along the Plane Surface of an Elastic Solid”. In: *Proceedings of the London Mathematical Society* s1-17.1 (1885). Rayleigh surface waves., pp. 4–11. DOI: 10.1112/plms/s1-17.1.4.

[233] L. Rayleigh. “On the self-induction and resistance of straight conductors”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 21.132 (1886). Crossref-verified. Philosophical Magazine series 5. This is the skin-effect paper the chapter means (high-frequency resistance of a straight conductor). Crossref’s deposited title is prefixed ’LII. ’, dropped here. Not to be confused with Rayleigh’s ’Notes on electricity and magnetism.–II. The self-induction and

resistance of compound conductors', Phil. Mag. 22, 469-500 (1886)., pp. 381–394. DOI: 10.1080/14786448608627863.

[234] H. Hertz. "Ueber einen Einfluss des ultravioletten Lichtes auf die electrische Entladung". In: *Annalen der Physik* 267.8 (1887). The photoelectric effect noticed as an unwanted disturbance during the experiments that produced the first electromagnetic waves., pp. 983–1000. DOI: 10.1002/andp.18872670827.

[235] E. Mach and P. Salcher. "Photographische Fixirung der durch Projectile in der Luft eingeleiteten Vorgänge". In: *Annalen der Physik* 268.10 (1887). The Vienna Sitzungsberichte DO carry volume numbers (unlike the Prussian ones), so @article with volume 95 is correct. Shock-cone photographs; origin of the Mach angle. Distinct from existing key Mach:1883., pp. 277–291. DOI: 10.1002/andp.18872681008.

[236] A. A. Michelson and E. W. Morley. "On the Relative Motion of the Earth and the Luminiferous Ether". In: *American Journal of Science* 34.203 (1887). Series 3. The bound is stated only as a fraction of the Earth's orbital velocity — probably below one sixth, certainly below one quarter — and no figure in kilometres per second appears in the paper, pp. 333–345. DOI: 10.2475/ajs.s3-34.203.333.

[237] R. Dedekind. *Was sind und was sollen die Zahlen?* 1st ed. Erklärung 64 defines a system as infinite when it is similar to a proper part of itself. No DOI. Beware the archive.org item WasSindUndWasSollenDieZahlen: it is a 2012 setting of the 4th edition, not the original. Braunschweig: Friedrich Vieweg und Sohn, 1888.

[238] W. Hallwachs. "Ueber den Einfluss des Lichtes auf electrostatisch geladene Körper". In: *Annalen der Physik* 269.2 (1888), pp. 301–312. DOI: 10.1002/andp.18882690206.

[239] H. Hertz. "Ueber electrodynamische Wellen im Luftraume und deren Reflexion". In: *Annalen der Physik* 270.8A (1888). Issue 8A. Chosen over Hertz's other 1888 papers because both citing passages are about standing waves: the experiment chapter describes reflection off a zinc sheet and mapping nodes/antinodes, which is this paper. Note for the writer: the prism-refraction and wire-grid-polarization demonstrations mentioned in the same reserved passage are from 'Ueber Strahlen electrischer Kraft', Ann. Phys. 272(4), 769–783 (1889), DOI 10.1002/andp.18892720402 – a separate key if that detail is ever written out. Crossref prints the old orthography 'electrodynamische' (with c)., pp. 609–623. DOI: 10.1002/andp.18882700802.

[240] F. Reinitzer. "Beiträge zur Kenntniss des Cholesterins". In: *Monatshefte für Chemie - Chemical Monthly* 9.1 (1888), pp. 421–441. DOI: 10.1007/BF01516710.

[241] H. Hertz. "Die Kräfte electrischer Schwingungen, behandelt nach der Maxwell'schen Theorie". In: *Annalen der Physik* 272.1 (1889). Issue 1. Hertz's dipole-radiation paper – the source for the multipole/finite-size-emitter claim in the citing sentence. TRAP: the DOI string reads '1888' although Crossref (and the volume) date the issue 1889; the DOI resolves and its title matches, so do not 'correct' it to 1889. Original orthography: 'electrischer', 'Kräfte' with umlaut., pp. 1–22. DOI: 10.1002/andp.18882720102.

[242] P.-H. Hugoniot. "Mémoire sur la propagation du mouvement dans les corps et spécialement dans les gaz parfaits (deuxième partie)". In: *Journal de l'École Polytechnique* 58 (1889). Posthumous. First part in volume 57 (1887) 3-97; the shock jump relations are in this second part., pp. 1–125.

[243] S. Kovalevskaya. "Sur le problème de la rotation d'un corps solide autour d'un point fixe". In: *Acta Mathematica* 12 (1889). Printed byline 'Sophie Kowalevski'. The Kovalevskaya integrable case of the heavy top., pp. 177–232.

[244] O. Lehmann. “Über fliessende Krystalle”. In: *Zeitschrift für Physikalische Chemie* 4U.1 (1889), pp. 462–472. DOI: 10.1515/zpch-1889-0434.

[245] G. Peano. *Arithmetices principia, nova methodo exposita*. In Latin; introduces the Peano axioms. Turin: Fratres Bocca, 1889.

[246] M. Couette. “Études sur le frottement des liquides”. In: *Annales de Chimie et de Physique (Série 6)* 21 (1890). The rotating-cylinder viscometry after which Couette flow is named., pp. 433–510.

[247] H. Poincaré. “Sur le problème des trois corps et les équations de la dynamique”. In: *Acta Mathematica* 13 (1890). The King Oscar II prize memoir, in its corrected second printing (the flawed first printing was withdrawn). Non-existence of further analytic integrals; homoclinic tangle., pp. 1–270.

[248] J. R. Rydberg. “On the structure of the line-spectra of the chemical elements”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 29.179 (1890). The full memoir is Recherches sur la constitution des spectres d’emission des elements chimiques, Kongliga Svenska Vetenskaps-Akademiens Handlingar 23 (11), 1890; the Philosophical Magazine paper is the accessible statement of the formula., pp. 331–337. DOI: 10.1080/14786449008619945.

[249] G. Cantor. “Ueber eine elementare Frage der Mannigfaltigkeitslehre”. In: *Jahresbericht der Deutschen Mathematiker-Vereinigung* 1 (1891). The diagonal argument. No DOI exists; digitised by GDZ Göttingen. Band 1 is dated 1890/91 by the publisher and was issued in 1892, pp. 75–78.

[250] S. C. Chandler. “On the variation of latitude,I”. In: *The Astronomical Journal* 11 (1891). Discovery of the 427-day polar motion (Chandler wobble)., p. 59. DOI: 10.1086/101603.

[251] A. Schoenflies. *Krystallsysteme und Krystallstructur*. B. G. Teubner, 1891.

[252] O. Heaviside. *Electrical Papers*. Two volumes; first edition 1892 (second 1894). This is the collection the citing sentence describes – it reprints the 1885 ’Electrician’ series in which Heaviside reduced Maxwell’s system to the four vector equations. Consider recording ’volumes = 2’. Macmillan, Co., London, and New York, 1892.

[253] A. M. Lyapunov. *Obshchaya zadacha ob ustoichivosti dvizheniya (The General Problem of the Stability of Motion)*. Doctoral dissertation, in Russian. French translation: Annales de la Faculté des Sciences de Toulouse (2) 9 (1907); English translation: International Journal of Control 55 (1992) 531-773. Kharkov: Kharkov Mathematical Society, 1892.

[254] S. Newcomb. “On the Dynamics of the Earth’s Rotation, with Respect to the Periodic Variations of Latitude”. In: *Monthly Notices of the Royal Astronomical Society* 52.5 (1892). Explains the Euler-to-Chandler period lengthening by the Earth’s elasticity., pp. 336–336. DOI: 10.1093/mnras/52.5.336.

[255] H. Poincaré. *Les méthodes nouvelles de la mécanique céleste*. Three volumes: I (1892), II (1893), III (1899); keyed by volume I. Poincaré sections and return maps; the homoclinic analysis is in volume III. Paris: Gauthier-Villars, 1892.

[256] J. J. Waterston and J. W. Strutt. “On the physics of media that are composed of free and perfectly elastic molecules in a state of motion”. In: *Philosophical Transactions of the Royal Society of London. (A.)* 183.183 (1892). Submitted 1845 and refused; printed 1892 with an introduction by Lord Rayleigh. Keyed by publication year per house rule., pp. 1–79. DOI: 10.1098/rsta.1892.0001.

[257] G. Frege. *Grundgesetze der Arithmetik, begriffsschriftlich abgeleitet*. Vol. 1. Band I 1893, Band II 1903. Basic Law V, the value-range axiom on which the logicist reduction rests, is the principle Russell's paradox refutes; Band II carries Frege's appendix on it. Partial English translation in van Heijenoort's source book. Jena: Hermann Pohle, 1893.

[258] O. Heaviside. *Electromagnetic Theory, Volume 1*. Operational calculus practised before justification. London: The Electrician Printing and Publishing Company, 1893.

[259] W. Wien. "Eine neue Beziehung der Strahlung schwarzer Koerper zum zweiten Hauptsatz der Waermetheorie". In: *Sitzungsberichte der Koeniglich Preussischen Akademie der Wissenschaften zu Berlin*. The displacement law. Academy proceedings, no volume number for the series. 1893, pp. 55–62.

[260] P. d. Fermat. *Œuvres de Fermat. Tome deuxième: Correspondance*. Ed. by P. Tannery and C. Henry. The 1662 letters to Marin Cureau de la Chambre, in which Fermat derives the sine law from least time, exist only as manuscripts; this Tannery-Henry critical edition is the standard printed source and is what the key must denote. Kind is book with an EDITOR and no author in the biblatex sense – editor satisfies the author requirement for @book under the project's validator. Freely readable on Gallica (ark:/12148/bpt6k6213616t) and archive.org (oeuvresdefermat02ferm), so it can be graded GREEN with a manifest URL. I did not verify the page span of the individual letters and have deliberately given none. Paris: Gauthier-Villars, 1894.

[261] J. H. Poynting. *The Mean Density of the Earth*. Adams Prize essay (1893). Reviews Cavendish and successors; records that Cavendish's printed mean 5.48 contains an arithmetic slip, his data giving 5.448. London: Charles Griffin and Company, 1894.

[262] C. V. Boys. "On the Newtonian constant of gravitation". In: *Philosophical Transactions of the Royal Society of London. (A.)* 186.186 (1895). Quartz-fibre small balance; reframes the Cavendish experiment as a measurement of G rather than of the Earth's density., pp. 1–72. DOI: 10.1098/rsta.1895.0001.

[263] H. Bruns. "Das Eikonal". In: *Abhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften, Mathematisch-Physische Classe*. Vol. 21. Pages given as customarily cited (this is the form Born & Wolf use: Abh. Kgl. Sächs. Ges. Wiss., math.-phys. Kl. 21, 321-436). Library catalogues (SLUB Dresden) record 323-435 for the same item, so if the bib writer prefers a catalogue-verifiable range, use 323-435 – but do not silently mix the two. Simultaneously issued as a separate monograph: H. Bruns, *Das Eikonal*, S. Hirzel, Leipzig, 1895. No DOI. This is the work the eikonal equation is named after. 1895, pp. 321–436.

[264] P. Curie. "Propriétés magnétiques des corps à diverses températures". In: *Annales de Chimie et de Physique* 5 (1895). 7th series; Curie's doctoral thesis., pp. 289–405.

[265] D. J. Korteweg and G. de Vries. "On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 39.240 (1895). The KdV equation., pp. 422–443. DOI: 10.1080/14786449508620739.

[266] A. A. Michelson. "Détermination expérimentale de la valeur du mètre en longueurs d'ondes lumineuses". In: *Travaux et Mémoires du Bureau International des Poids et Mesures*. Vol. 11. BIPM Travaux et Mémoires tome 11, 1895. The title page reads 'par M. Albert-A. Michelson; mémoire traduit de l'anglais par M. J.-René Benoît' — Michelson is the sole named author and Benoît the translator, although the etalon work was carried out jointly with Benoît at the BIPM, which is why the chapter says 'Michelson and Benoit's determination'. Record Benoît as translator in a note rather

than as a second author. This is the source of the 1 m = 1553163.5 red-cadmium wavelengths figure quoted in 10-exp-wave-optics.tex. No DOI; digitised on Gallica (ark:/12148/bpt6k9693613c), a GREEN access route. 1895, pp. 1–85.

[267] S. Newcomb. *The Elements of the Four Inner Planets and the Fundamental Constants of Astronomy*. Washington, DC. Newcomb's re-reduction of the inner-planet observations with improved planetary masses; this is the work that moved Le Verrier's 38 arcsec/century Mercury residual to the modern 43 (Newcomb's own value was 41.24 +/- 2.09). No DOI (1895 government monograph). Full text is on the Internet Archive (elementsoffourin00newcrich and cihm_16774), so it can be declared "access": "open" with that URL rather than grading RED. Cited at the Newcomb sentence in sec:clt-mercury. U.S. Government Printing Office, 1895.

[268] O. Reynolds. "On the dynamical theory of incompressible viscous fluids and the determination of the criterion". In: *Philosophical Transactions of the Royal Society of London. (A.)* 186.186 (1895). Introduces the Reynolds decomposition and Reynolds stresses., pp. 123–164. DOI: 10.1098/rsta.1895.0004.

[269] W. C. Roentgen. "Ueber eine neue Art von Strahlen (Vorlaeufige Mittheilung)". In: *Sitzungsberichte der Physikalisch-Medicinischen Gesellschaft zu Wuerzburg*. Key transliterates the umlaut as in the existing Roemer:1676 and Goedel:1931 entries. Surname Röntgen; title Ueber = Über, Vorlaeufige = Vorläufige; venue Würzburg. 1895, pp. 132–141.

[270] K. Weierstrass. *Über das sogenannte Dirichlet'sche Princip*. Paper read to the Berlin Academy on 14 July 1870; first printed in the collected works (1895). Best entered as @incollection in Mathematische Werke, Band 2. 1895.

[271] H. Becquerel. "Sur les radiations emises par phosphorescence". In: *Comptes Rendus Hebdomadaires des Seances de l'Academie des Sciences*. Vol. 122. 1896, pp. 420–421.

[272] W. Wien. "Ueber die Energievertheilung im Emissionsspectrum eines schwarzen Körpers". In: *Annalen der Physik* 294.8 (1896). Wien's distribution law, falsified at long wavelength by Lummer:1899 and Rubens:1900., pp. 662–669. DOI: 10.1002/andp.18962940803.

[273] E. Zermelo. "Ueber einen Satz der Dynamik und die mechanische Wärmetheorie". In: *Annalen der Physik* 293.3 (1896). Not a genuine collision: the two claimed works are one paper under two volume numberings – Wiedemann's Annalen der Physik und Chemie vol. 57 is Wiley's continuous Annalen der Physik vol. 293, same pages. Crossref issue 3, January 1896. The 1896 printing sets the title in the old orthography ('Ueber'); modernised to 'Über' per house rule. Zermelo's recurrence objection to the H-theorem; same author as Zermelo:1908., pp. 485–494. DOI: 10.1002/andp.18962930314.

[274] Archimedes. *The Works of Archimedes*. Ed. by T. L. Heath. Edited and translated by Thomas L. Heath; the editor satisfies the author requirement. Cited for On Floating Bodies (buoyancy, Book I, Props. 5-7); primary text is 3rd century BC. Cambridge University Press, 1897.

[275] J. Larmor. "On the theory of the magnetic influence on spectra; and on the radiation from moving ions". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 44.271 (1897). Series 5, issue 271 (December 1897). This is the paper containing the Larmor radiated-power formula (radiation from moving ions), which is what the citing sentence in 07-electrodynamics/08-radiation-scattering.tex needs. Not to be confused with Larmor's 'A dynamical theory of the electric and luminiferous medium, Part III', Phil. Trans. R. Soc. A 190 (1897) 205-300, which is

sometimes cited for the same result. Crossref confirms title/volume/pages against the DOI., pp. 503–512. DOI: 10.1080/14786449708621095.

[276] J. J. Thomson. “Cathode Rays”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 44.269 (1897). Discovery of the electron. No collision with Thomson:1927 (G. P. Thomson), which is a different year., pp. 293–316. DOI: 10.1080/14786449708621070.

[277] P. Zeeman. “On the influence of magnetism on the nature of the light emitted by a substance”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 43.262 (1897). A shorter announcement appeared in Nature 55 (1897) 347., pp. 226–239. DOI: 10.1080/14786449708620985.

[278] P. Curie and M. Curie. “Sur une substance nouvelle radio-active, contenue dans la pechblende”. In: *Comptes Rendus Hebdomadaires des Seances de l’Academie des Sciences*. Vol. 127. 1898, pp. 175–178.

[279] P. Curie, M. Curie, and G. Bémont. “Sur une nouvelle substance fortement radio-active, contenue dans la pechblende”. In: *Comptes Rendus Hebdomadaires des Seances de l’Academie des Sciences*. Vol. 127. The radium paper (December 1898). The companion key Curie:1898 is the July polonium paper; the two are distinct discoveries in the same journal volume. 1898, pp. 1215–1217.

[280] J. Dewar. “Preliminary note on the liquefaction of hydrogen and helium”. In: *Proceedings of the Royal Society of London* 63.389-400 (1898), pp. 256–258. DOI: 10.1098/rspl.1898.0031.

[281] A. Liénard. “Champ électrique et magnétique produit par une charge électrique concentrée en un point et animée d’un mouvement quelconque”. In: *L’Éclairage Électrique* 16 (1898). Published in three instalments, nos. 27, 28 and 29 of tome 16 (July-September 1898). No DOI; the journal is digitised by CNAM (cnum.cnam.fr, P91.16) which is the access route to declare. Pages must be given as the three ranges — a single range would misreport the source., pp. 5–14, 53–59, 106–112.

[282] O. Lummer and F. Kurlbaum. “Der elektrisch gegluehte absolut schwarze Koerper und seine Temperaturmessung”. In: *Verhandlungen der Physikalischen Gesellschaft zu Berlin*. Vol. 17. The Reichsanstalt cavity radiator. Published title puts ’absolut schwarze’ in quotation marks. The Berlin society’s Verhandlungen restarted numbering as the Deutsche Physikalische Gesellschaft series vol. 1 in 1899. 1898, pp. 106–111.

[283] J. W. GIBBS. “Fourier’s Series”. In: *Nature* 59.1539 (1899). Letter correcting his 1898 letter (Nature 59, 200); this is where the overshoot value is stated., pp. 606–606. DOI: 10.1038/059606a0.

[284] O. Lummer and E. Pringsheim. “Die Vertheilung der Energie im Spectrum des schwarzen Koerpers”. In: *Verhandlungen der Deutschen Physikalischen Gesellschaft*. Vol. 1. First clear deviations from Wien’s law. 1899, pp. 23–41.

[285] M. Planck. “Über irreversible Strahlungsvorgänge”. In: *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin*. Academy proceedings: the series carries no volume number. Introduces the natural units built from h, c and G. No DOI. 1899, pp. 440–480.

[286] L. Rayleigh. “On the transmission of light through an atmosphere containing small particles in suspension, and on the origin of the blue of the sky”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 47.287 (1899). Crossref-verified. Philosophical Magazine series 5. This is the later paper the chapter cites for the demonstration that the air molecules themselves suffice, so the blue of

the sky needs no dust. Crossref deposits this title with no roman-numeral prefix., pp. 375–384. DOI: 10.1080/14786449908621276.

[287] E. Rutherford. “Uranium radiation and the electrical conduction produced by it”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 47.284 (1899), pp. 109–163. DOI: 10.1080/14786449908621245.

[288] H. Bénard. “Étude expérimentale des courants de convection dans une nappe liquide. — Régime permanent : tourbillons cellulaires”. In: *Journal de Physique Théorique et Appliquée* 9.1 (1900). Two instalments in volume 11. The cellular-convection experiments., pp. 513–524. DOI: 10.1051/jphystap:019000090051300.

[289] P. Drude. “Zur Elektronentheorie der Metalle”. In: *Annalen der Physik* 306.3 (1900), pp. 566–613. DOI: 10.1002/andp.19003060312.

[290] P. Drude. “Zur Elektronentheorie der Metalle; II. Teil. Galvanomagnetische und thermomagnetische Effecte”. In: *Annalen der Physik* 308.11 (1900), pp. 369–402. DOI: 10.1002/andp.19003081102.

[291] D. Hilbert. “Mathematische Probleme”. In: *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse*. The twenty-three problems. The Göttinger Nachrichten carry no volume numbers. English translation by M. W. Newson, Bulletin of the AMS 8 (1902), 437–479. 1900, pp. 253–297.

[292] O. Lummer and E. Pringsheim. “Über die Strahlung des schwarzen Körpers für lange Wellen”. In: *Verhandlungen der Deutschen Physikalischen Gesellschaft*. Vol. 2. The 1900 continuation of Lummer:1899, extending the measurement to about 18 um and to source temperatures up to 1772 C. Same series as the existing Lummer:1899 entry (Verhandlungen der Deutschen Physikalischen Gesellschaft, vol. 2 for 1900), so @inproceedings with booktitle, matching the house treatment of Lummer:1898/1899. No DOI exists. NOTE on encoding: the two neighbouring entries in references.bib transliterate umlauts (Koerper, gegluehte); if that local convention is kept, this title becomes ’Ueber die Strahlung des schwarzen Koerpers fuer lange Wellen’ – but the published title is the Unicode form given here and pdflatex+T1 handles it. Placed only where the chapter claims the push to long wavelength AND high temperature; deliberately NOT added to the phenomenon, whose ’8 um / 300-1650 K’ numbers are the 1899 measurement, not this one. 1900, pp. 163–180.

[293] R. D. Oldham. “On the propagation of earthquake motion to great distances”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 194.252-261 (1900). Identification of P, S and surface-wave phases in distant seismograms., pp. 135–174. DOI: 10.1098/rsta.1900.0015.

[294] M. Planck. “Ueber eine Verbesserung der Wienschen Spectralgleichung”. In: *Verhandlungen der Deutschen Physikalischen Gesellschaft*. Vol. 2. Suffix ’a’: the interpolation of 19 October 1900. Distinct from Planck:1900b, the quantum derivation of 14 December 1900, in the same volume. 1900, pp. 202–204.

[295] M. Planck. “Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum”. In: *Verhandlungen der Deutschen Physikalischen Gesellschaft*. Vol. 2. Suffix ’b’: read 14 December 1900, the combinatorial derivation introducing energy elements h*nu and the relation S = k log W. 1900, pp. 237–245.

[296] L. Rayleigh. “Remarks upon the law of complete radiation”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 49.301 (1900). Published under the title ’Lord Rayleigh’; the paper states the nu^2 T form and inserts an exponential cut-off by hand. Rayleigh’s numerical factor was later corrected by Jeans:1905., pp. 539–540. DOI: 10.1080/14786440009463878.

[297] H. Rubens and F. Kurlbaum. “Ueber die Emission langwelliger Waermestrahlen durch den schwarzen Koerper bei verschiedenen Temperaturen”. In: *Sitzungsberichte der Koeniglich Preussischen Akademie der Wissenschaften zu Berlin*. Residual-ray measurements at 24 and 51.2 micrometres; communicated to Planck on 7 October 1900, the immediate cause of Planck:1900a. An expanded version appeared in Annalen der Physik 309 (1901) 649–666. 1900, pp. 929–941.

[298] I. Bendixson. “Sur les courbes définies par des équations différentielles”. In: *Acta Mathematica* 24.0 (1901). The Poincaré-Bendixson theorem in its general form., pp. 1–88. DOI: 10.1007/bf02403068.

[299] P. Lebedew. “Untersuchungen über die Druckkräfte des Lichtes”. In: *Annalen der Physik* 311.11 (1901). Keyed Lebedew (the published spelling) rather than the modern transliteration Lebedev, per the Surname:YYYY convention. Wiley volume 311 = series 4, vol. 6, confirmed against Crossref. A duplicate key Lebedev:1901 was merged here. Umlauts: ueber = über, Druckkraefte = Druckkräfte., pp. 433–458. DOI: 10.1002/andp.19013111102.

[300] M. Planck. “Ueber das Gesetz der Energieverteilung im Normalspectrum”. In: *Annalen der Physik* 309.3 (1901). DOI supplied by the task prompt. Whole-series vol. 309 = 4th series vol. 4., pp. 553–563. DOI: 10.1002/andp.19013090310.

[301] E. Wiechert. “Elektrodynamische Elementargesetze”. In: *Annalen der Physik* 309.4 (1901). DATING TRAP, of the Kretschmann kind but sharper: the text first appeared in 1900 in the Lorentz jubilee volume — "Recueil de travaux offerts par les auteurs a H. A. Lorentz", Archives Neerlandaises des Sciences Exactes et Naturelles, 2nd series, vol. 5 (Nijhoff, The Hague, 1900), pp. 549-573 — which is the year the citing author used and the year the key records. That 1900 printing has no DOI. The same paper was republished in Annalen der Physik 309(4), 667-689 (1901), and only that version has a resolvable DOI (fetched and title-checked: "Elektrodynamische Elementargesetze", author Wiechert). Recommendation: keep the citekey Wiechert:1900 (it is fixed by the citation in 08-radiation-scattering.tex), give the entry the Annalen 1901 fields plus DOI so the reader can actually retrieve it, and record the 1900 Festschrift first appearance in the note field — do NOT put the 1901 DOI on 1900 Archives Neerlandaises fields, and do not invent a DOI for the 1900 printing., pp. 667–689. DOI: 10.1002/andp.19013090403.

[302] L. Bianchi. “Sui simboli a quattro indici e sulla curvatura di Riemann”. In: *Rendiconti della Reale Accademia Nazionale dei Lincei*. Vol. 11. Series 5, vol. 11. Attribution is customary: the identities were known earlier to Voss (1880) and in Ricci’s school; the citing chapter says so explicitly. 1902, pp. 3–7.

[303] H. L. Callendar. “Continuous electrical calorimetry”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 199.312-320 (1902). Companion paper by H. T. Barnes on the heat capacity of water in the same volume, pp. 149–263., pp. 55–148. DOI: 10.1098/rsta.1902.0013.

[304] J. W. Gibbs. *Elementary Principles in Statistical Mechanics, Developed with Especial Reference to the Rational Foundation of Thermodynamics*. New York: Charles Scribner’s Sons, 1902.

[305] J. Hadamard. “Sur les problèmes aux dérivées partielles et leur signification physique”. In: *Princeton University Bulletin* 13 (1902). The well-posedness definition., pp. 49–52.

[306] J. H. Jeans. “The Stability of a Spherical Nebula”. In: *Philosophical Transactions of the Royal Society of London A* 199 (1902). Metadata confirmed against Crossref., pp. 1–53. DOI: 10.1098/rsta.1902.0012.

[307] W. M. Kutta. “Auftriebskräfte in strömenden Flüssigkeiten”. In: *Illustrierte Aeronautische Mitteilungen* 6 (1902). Thin-aerofoil lift computation; the Kutta half of Kutta-Joukowski., pp. 133–135.

[308] P. Lenard. “Ueber die lichtelektrische Wirkung”. In: *Annalen der Physik* 313.5 (1902). Establishes that the maximum electron energy depends on frequency and not on intensity, and that there is a threshold frequency., pp. 149–198. DOI: 10.1002/andp.19023130510.

[309] H. Poincaré. *La Science et l’Hypothèse*. Source of the conventionalist thesis: parts of a physical theory are stipulations rather than measured facts. English translation: Science and Hypothesis, Walter Scott, London (1905). Paris: Ernest Flammarion, 1902.

[310] L. Fejér. “Untersuchungen über Fouriersche Reihen”. In: *Mathematische Annalen* 58.1-2 (1903), pp. 51–69. DOI: 10.1007/bf01447779.

[311] E. F. Nichols and G. F. Hull. “The Pressure due to Radiation”. In: *The Astrophysical Journal* 17 (1903). The independent and more accurate measurement of radiation pressure. A companion pair of papers appeared the same year in Physical Review (Series I) 17, 26–50 and 91–104., p. 315. DOI: 10.1086/141035.

[312] E. Rutherford and F. Soddy. “Radioactive Change”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 5.29 (1903), pp. 576–591. DOI: 10.1080/14786440309462960.

[313] M. Abraham. “Zur Theorie der Strahlung und des Strahlungsdruckes”. In: *Annalen der Physik* 319.7 (1904). Cited for the Abraham–Lorentz radiation-reaction force. Volume 319 is the continuous Annalen count; contemporary citations give ’Band 14’ (4th series), so a source saying Ann. Phys. 14, 236 (1904) is the same paper. Crossref-verified., pp. 236–287. DOI: 10.1002/andp.19043190703.

[314] D. Hilbert. “Über das Dirichletsche Prinzip”. In: *Mathematische Annalen* 59.1-2 (1904). The rehabilitation of the Dirichlet principle; an earlier short version appeared in Jahresbericht der DMV 8 (1900)., pp. 161–186. DOI: 10.1007/bf01444753.

[315] L. Prandtl. “Über Flüssigkeitsbewegung bei sehr kleiner Reibung”. In: *Verhandlungen des dritten Internationalen Mathematiker-Kongresses in Heidelberg 1904*. Delivered at the Heidelberg ICM in August 1904; the Verhandlungen volume was printed in 1905. Universally cited as 1904 and keyed so here (no DOI to arbitrate); record the dating in the entry note. The boundary-layer paper. 1904, pp. 484–491.

[316] J. Thomson. “On the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 7.39 (1904). The ’plum pudding’ model, cited as a genuine falsifiable dynamical model rather than as a straw man., pp. 237–265. DOI: 10.1080/14786440409463107.

[317] C. G. Barkla. “Polarised röntgen radiation”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 204.372-386 (1905). The double-scattering experiment showing X-rays are transverse. Crossref prefixes the title with the article number, "XIII. Polarised röntgen radiation", and lowercases it; the volume as printed reads "Polarised Röntgen Radiation". Verified against the Crossref record for the DOI (volume 204, pages 467-479, 1905)., pp. 467–479. DOI: 10.1098/rsta.1905.0013.

[318] A. Einstein. “Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen”. In: *Annalen der Physik* 322.8 (1905). Ser. 4, vol. 17, issue 8; received 11 May 1905. The Brownian-motion paper. Takes the next free suffix d, since a/b are the relativity and mass–energy papers and c is claimed for the light-quantum paper by a directive in the photoelectric chapter., pp. 549–560. DOI: 10.1002/andp.19053220806.

[319] A. Einstein. “Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt”. In: *Annalen der Physik* 322.6 (1905). Ser. 4, vol. 17, issue 6; received 18 March 1905. The light-quantum (photoelectric) paper. Suffix c because Einstein:1905a/b are the relativity and mass–energy papers; the Brownian-motion paper is Einstein:1905d. A prior version of this entry duplicated the Brownian-motion metadata under this key while the photon chapter cited it for the light quantum – wrong work under a right key, caught by a reader on the bibliography page. Metadata confirmed against Crossref., pp. 132–148. DOI: 10.1002/andp.19053220607.

[320] A. Einstein. “Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig?” In: *Annalen der Physik* 323.13 (1905). Ser. 4, vol. 18. Received 27 September 1905. The mass–energy paper; the equation never appears as $E = mc^2$, but as a mass change L/V^2, quoted numerically as $L/(9 \times 10^{20})$ with energy in erg and mass in gram, pp. 639–641. DOI: 10.1002/andp.19053231314.

[321] A. Einstein. “Zur Elektrodynamik bewegter Körper”. In: *Annalen der Physik* 322.10 (1905). Ser. 4, vol. 17. Received 30 June 1905. The founding paper of special relativity; Einstein writes V for the speed of light throughout, pp. 891–921. DOI: 10.1002/andp.19053221004.

[322] J. Jeans. “On the partition of energy between matter and Æther”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 10.55 (1905). Published title spells the second noun with the ligature ’Aether’ (AE ligature)., pp. 91–98. DOI: 10.1080/14786440509463348.

[323] P. Langevin. “Magnétisme et théorie des électrons”. In: *Annales de Chimie et de Physique* 5 (1905). 8th series., pp. 70–127.

[324] W. Sutherland. “A dynamical theory of diffusion for non-electrolytes and the molecular mass of albumin”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 9.54 (1905). Not a genuine collision: both candidates are this one paper, one carrying the Phil. Mag. running article number ’LXXV.’ before the title. Series 6, issue 54. Title given in the form the DOI record prints it, without the running number., pp. 781–785. DOI: 10.1080/14786440509463331.

[325] P. Duhem. *La Théorie physique: son objet, sa structure*. The holism thesis later generalized by Quine: an experiment tests a hypothesis only together with its auxiliaries. Second, enlarged edition 1914; English translation The Aim and Structure of Physical Theory, trans. P. P. Wiener, Princeton University Press (1954). Paris: Chevalier et Rivière, 1906.

[326] A. Einstein. “Zur Theorie der Brownschen Bewegung”. In: *Annalen der Physik* 324.2 (1906). The second Brownian paper; contains the rotational diffusion result, pp. 371–381. DOI: 10.1002/andp.19063240208.

[327] T. Lyman. “The Spectrum of Hydrogen in the Region of Extremely Short Wave-Length”. In: *Memoirs of the American Academy of Arts and Sciences* 13.3 (1906), p. 125. DOI: 10.2307/25058084.

[328] W. Nernst. “Über die Berechnung chemischer Gleichgewichte aus thermischen Messungen”. In: *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse*. Academy proceedings, year-based issues, no volume number; the heat theorem (third law). 1906, pp. 1–40.

[329] H. Poincaré. “Sur la dynamique de l’électron”. In: *Rendiconti del Circolo Matematico di Palermo* 21.1 (1906). Submitted July 1905 and universally cited as 1905; the volume is dated 1906, which fixes the key. Introduces the Lorentz group as a group and the invariant $x^2 + y^2 + z^2 - c^2t^2$, pp. 129–176. DOI: 10.1007/BF03013466.

[330] K. Schwarzschild. “Über das Gleichgewicht der Sonnenatmosphäre”. In: *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse*. Vol. 1906. Karl Schwarzschild; radiative equilibrium and the convective stability criterion; no DOI. 1906, pp. 41–53.

[331] M. von Smoluchowski. “Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen”. In: *Annalen der Physik* 326.14 (1906), pp. 756–780. DOI: 10.1002/andp.19063261405.

[332] J. J. Thomson. *Conduction of Electricity through Gases*. Second edition, 1906 (first edition 1903) – the edition is load-bearing: the scattering calculation giving sigma_T = (8 pi/3) r_e^2, which is what both citations in ch. 08-radiation-scattering rest on, is in the second edition, so the entry must carry edition = 2. No DOI. The later two-volume rewrite with G. P. Thomson (1928, 1933) is a different work. Public-domain scans exist (Internet Archive, conductionofele00thomuoft), which is a declarable access route. Cambridge University Press, 1906.

[333] A. Einstein. “Die Plancksche Theorie der Strahlung und die Theorie der spezifischen Wärme”. In: *Annalen der Physik* 327.1 (1907). Series 4, volume 22 = continuous volume 327, issue 1. The DOI string encodes 1906 (Wiley numbered it from the volume’s opening), but Crossref records published-print 1907 for the issue, which is also the customary citation — verified against the Crossref record, title, volume, issue and pages all matching. Introduces the quantized-oscillator model of the specific heats of solids., pp. 180–190. DOI: 10.1002/andp.19063270110.

[334] A. Einstein. “Über das Relativitätsprinzip und die aus demselben gezogenen Folgerungen”. In: *Jahrbuch der Radioaktivität und Elektronik* 4 (1907). Review article whose section V gives the first statement of the equivalence principle and of the gravitational redshift. The volume is dated 1907 and the issue appeared in early 1908; the customary citation is 1907. Corrections printed as Berichtigungen, Jahrbuch der Radioaktivität und Elektronik 5 (1908), 98–99. No DOI exists., pp. 411–462.

[335] E. Fischer. “Sur la convergence en moyenne”. In: *Comptes Rendus de l’Académie des Sciences (Paris)*. Vol. 144. The Riesz–Fischer theorem (completeness of L2). 1907, pp. 1022–1024.

[336] F. Riesz. “Sur une espèce de géométrie analytique des systèmes de fonctions sommables”. In: *Comptes Rendus de l’Académie des Sciences (Paris)*. Vol. 144. The L2 representation theorem. Riesz published several C.R. notes in 1907; if another is ever cited, suffix then. Fréchet’s simultaneous note customarily shares credit. 1907, pp. 1409–1411.

[337] H. J. Round. “A Note on Carborundum”. In: *Electrical World* 49 (1907), p. 309.

[338] E. Schmidt. “Zur Theorie der linearen und nichtlinearen Integralgleichungen”. In: *Mathematische Annalen* 63.4 (1907). The orthogonalization procedure and the expansion theorem for compact kernels., pp. 433–476. DOI: 10.1007/bf01449770.

[339] P. Weiss. “L’hypothèse du champ moléculaire et la propriété ferromagnétique”. In: *Journal de Physique Théorique et Appliquée* 6.1 (1907), pp. 661–690. DOI: 10.1051/jphystap:019070060066100.

[340] C. G. Barkla and C. A. Sadler. “Homogeneous secondary Röntgen radiations”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 16.94 (1908). Series 6, issue 94 (October 1908). Discovery of homogeneous radiation characteristic of the radiating element. Crossref prefixes the title with the article number "XLVIII."; dropped here as in the existing Moseley entries., pp. 550–584. DOI: 10.1080/14786441008636537.

[341] H. Blasius. “Grenzschichten in Flüssigkeiten mit kleiner Reibung”. In: *Zeitschrift für Mathematik und Physik* 56 (1908). The flat-plate boundary-layer similarity solution., pp. 1–37.

[342] G. E. Hale. “On the Probable Existence of a Magnetic Field in Sun-Spots”. In: *The Astrophysical Journal* 28 (1908). The founding observation of astrophysical magnetism, as the citing sentence says. Crossref confirms title, ApJ vol. 28, first page 315, year 1908; the article runs to p. 343., p. 315. DOI: 10.1086/141602.

[343] H. Kamerlingh Onnes. “The Liquefaction of Helium”. In: *Science* 28.714 (1908). Leiden Communication no. 108; also Proceedings KNAW Amsterdam 11, 168 (1908)., pp. 316–316. DOI: 10.1126/science.28.714.316.a.

[344] P. Langevin. “Sur la théorie du mouvement brownien”. In: *Comptes Rendus de l’Académie des Sciences (Paris)*. Vol. 146. 1908, pp. 530–533.

[345] G. Mie. “Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen”. In: *Annalen der Physik* 330.3 (1908). Wiley’s continuous volume numbering is 330, issue 3; the original title page reads 4. Folge, Band 25. Cite volume 330 (that is what the DOI and Crossref carry) and record ’ser. 4, vol. 25’ in a note if desired., pp. 377–445. DOI: 10.1002/andp.19083300302.

[346] H. Minkowski. “Die Grundgleichungen für die elektromagnetischen Vorgänge in bewegten Körpern”. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*. Academy Nachrichten: NO volume number – the series is dated by year only. Presented in the session of 21 December 1907 and therefore cited as 1907 by some sources; printed in the 1908 Nachrichten, which fixes the key. This is the Minkowski electromagnetic energy–momentum tensor paper the citing sentence needs (Minkowski vs Abraham momentum), distinct from the existing bib entry Minkowski:1909 (’Raum und Zeit’). The Gesellschaft was styled ’Königliche’ on the 1908 title page; the form given here is the one universally used and matches EUDML. Reprinted in Math. Ann. 68 (1910) 472–525, DOI 10.1007/BF01455871 – do NOT attach that DOI to the 1908 Nachrichten entry. No DOI exists for the Nachrichten printing. 1908, pp. 53–111.

[347] F. Paschen. “Zur Kenntnis ultraroter Linienspektra. I. (Normalwellenlängen bis 27000 Å.-E.)” In: *Annalen der Physik* 332.13 (1908). The infrared hydrogen series. Title contains the Angstrom symbol and an umlaut: Normalwellenlängen bis 27000 ångström-Einheiten., pp. 537–570. DOI: 10.1002/andp.19083321303.

[348] W. Ritz. “On a New Law of Series Spectra”. In: *The Astrophysical Journal* 28 (1908). The combination principle. A German version, Ueber ein neues Gesetz der Serienspektren, appeared in Physikalische Zeitschrift 9 (1908) 521–529., p. 237. DOI: 10.1086/141591.

[349] M. v. Smoluchowski. “Molekular-kinetische Theorie der Opaleszenz von Gasen im kritischen Zustande, sowie einiger verwandter Erscheinungen”. In: *Annalen der Physik* 330.2 (1908), pp. 205–226. DOI: 10.1002/andp.19083300203.

[350] E. Zermelo. “Untersuchungen über die Grundlagen der Mengenlehre. I”. In: *Mathematische Annalen* 65.2 (1908), pp. 261–281. DOI: 10.1007/BF01449999.

[351] M. Abraham. “Zur Elektrodynamik bewegter Körper”. In: *Rendiconti del Circolo Matematico di Palermo* 28.1 (1909). The origin of the Abraham momentum in the Abraham–Minkowski controversy. Same title as Einstein’s 1905 paper – do not confuse them. Crossref-verified., pp. 1–28. DOI: 10.1007/BF03018208.

[352] C. Barkla and C. Sadler. “The absorption of Röntgen rays”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 17.101 (1909). Series 6, issue 101 (May 1909). Mass absorption coefficients of the characteristic radiations through the periodic table, and the excitation threshold. Crossref prefixes the title with the article number "LXIX."., pp. 739–760. DOI: 10.1080/14786440508636650.

[353] C. Carathéodory. “Untersuchungen über die Grundlagen der Thermodynamik”. In: *Mathematische Annalen* 67.3 (1909), pp. 355–386. DOI: 10.1007/bf01450409.

[354] H. Geiger and E. Marsden. “On a diffuse reflection of the alpha-particles”. In: *Proceedings of the Royal Society of London A* 82 (1909). Published title uses the Greek letter: On a Diffuse Reflection of the α-Particles., pp. 495–500.

[355] M. Knudsen. “Die Gesetze der Molekularströmung und der inneren Reibungsströmung der Gase durch Röhren”. In: *Annalen der Physik* 333.1 (1909). Wiley continuous numbering; customary citation vol. 28 (series 4)., pp. 75–130. DOI: 10.1002/andp.19093330106.

[356] H. A. Lorentz. *The Theory of Electrons and Its Applications to the Phenomena of Light and Radiant Heat*. First edition, Leipzig: B. G. Teubner; New York: G. E. Stechert & Co., 1909. Subtitle: ’A Course of Lectures Delivered in Columbia University, New York, in March and April 1906’. A second edition appeared in 1916 — the electron-theory programme the chapters cite is in both, but the key’s year fixes the 1909 first edition. No DOI. Full scans are on the Internet Archive (archive.org/details/theoryofelectron00loreuoft), which is a GREEN access route. B. G. Teubner, 1909.

[357] H. Minkowski. “Raum und Zeit”. In: *Physikalische Zeitschrift* 10.3 (1909). Address delivered at the 80th Assembly of German Natural Scientists and Physicians, Cologne, 21 September 1908, and customarily cited as 1908; printed in the issue of 1 February 1909, which fixes the key. No DOI exists: Physikalische Zeitschrift was never registered, pp. 104–111.

[358] J. Perrin. “Mouvement brownien et réalité moléculaire”. In: *Annales de Chimie et de Physique* 18 (1909). 8th series., pp. 5–114.

[359] W. Ritz. “Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik”. In: *crll* 135.135 (1909), pp. 1–61. DOI: 10.1515/crll.1909.135.1.

[360] G. I. Taylor. “Interference fringes with feeble light”. In: *Proceedings of the Cambridge Philosophical Society* 15 (1909). No DOI – Cambridge’s back-file DOIs for this series do not reach 1909, and the Crossref hits under this title are later reprints (Concepts of Quantum Optics, 1983, pp. 91-92, doi 10.1016/b978-0-08-029160-4.50015-4). Do not use the reprint DOI for the 1909 entry. The journal was renamed ’Mathematical Proceedings of the Cambridge Philosophical Society’ only in 1975, so the 1909 venue name is the shorter one., pp. 114–115.

[361] A. Einstein. “Theorie der Opaleszenz von homogenen Flüssigkeiten und Flüssigkeitsgemischen in der Nähe des kritischen Zustandes”. In: *Annalen der Physik* 338.16 (1910), pp. 1275–1298. DOI: 10.1002/andp.19103381612.

[362] N. Joukowski. “Über die Konturen der Tragflächen der Drachenflieger”. In: *Zeitschrift für Flugtechnik und Motorluftschiffahrt* 1 (1910). Name also transliterated Zhukovsky. The circulation lift theorem was first published in Russian in 1906 (’O prisoedinennykh vikhryakh’, Trudy Otdeleniya Fizicheskikh Nauk); this German paper is the accessible primary and the chapter notes the 1906 origin., pp. 281–284.

[363] M. Plancherel and M. Leffler. “Contribution À LÉtude de la reprÉsentation D’une fonction arbitraire par des intÉgrales dÉfinies”. In: *Rendiconti del Circolo Matematico di Palermo* 30.1 (1910), pp. 289–335. DOI: 10.1007/bf03014877.

[364] W. Voigt. *Lehrbuch der Kristallphysik (mit Ausschluss der Kristalloptik)*. Anisotropic elasticity, the 21 constants, Voigt notation; experimental resolution of the rari-constant vs multi-constant (Cauchy relations) controversy. Leipzig: B. G. Teubner, 1910.

[365] A. N. Whitehead and B. Russell. *Principia Mathematica*. Vol. 1. Volumes II and III followed in 1912 and 1913; second edition 1925–1927. Cambridge: Cambridge University Press, 1910.

[366] T. Wulf. “Über die in der Atmosphäre vorhandene Strahlung von hoher Durchdringungsfähigkeit”. In: *Physikalische Zeitschrift* 11 (1910), pp. 811–813.

[367] C. G. Barkla. “The spectra of the fluorescent Röntgen radiations”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 22.129 (1911). Series 6. The paper in which the two series are named K and L, letters chosen mid-alphabet to leave room on both sides. Barkla’s 1917 Nobel Prize (awarded 1918) rests on this line of work., pp. 396–412. DOI: 10.1080/14786440908637137.

[368] N. Bohr. *Studier over metallernes elektrontheori*. English translation in Niels Bohr Collected Works, vol. 1. 1911.

[369] P. Ehrenfest. “Welche Züge der Lichtquantenhypothese spielen in der Theorie der Wärmestrahlung eine wesentliche Rolle?” In: *Annalen der Physik* 341.11 (1911). Ser. 4, vol. 36; Wiley numbering 341. The paper that coins ’Ultraviolettkatastrophe’, a decade after Planck. Sole-author, so it takes ’a’., pp. 91–118. DOI: 10.1002/andp.19113411106.

[370] P. Ehrenfest and T. Ehrenfest. “Begriffliche Grundlagen der statistischen Auffassung in der Mechanik”. In: *Encyklopädie der mathematischen Wissenschaften, Band IV, Teil 32*. Encyclopaedia article, Leipzig 1911; the Encyklopädie numbers by Band and Artikel and has no volume number, so none is given. English translation 1959 as ’The Conceptual Foundations of the Statistical Approach in Mechanics’ (Cornell University Press). Some booksellers date the fascicle 1912. 1911, pp. 1–90.

[371] A. Einstein. “Über den Einfluß der Schwerkraft auf die Ausbreitung des Lichtes”. In: *Annalen der Physik* 340.10 (1911). Volume 340 continuous numbering (= ser. 4 vol. 35). Half-value light deflection and redshift prediction., pp. 898–908. DOI: 10.1002/andp.19113401005.

[372] H. Geiger and J. Nuttall. “The ranges of the α particles from various radioactive substances and a relation between range and period of transformation”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 22.130 (1911), pp. 613–621. DOI: 10.1080/14786441008637156.

[373] E. Hertzsprung. “Über die Verwendung photographischer effektiver Wellenlängen zur Bestimmung von Farbenäquivalenten”. In: *Publikationen des Astrophysikalischen Observatoriums zu Potsdam* 22 (1911). No. 63; colour–magnitude diagrams of the Pleiades and Hyades; no DOI.

[374] H. Kamerlingh Onnes. “Further experiments with liquid helium. D. On the change of the electrical resistance of pure metals at very low temperatures, etc. V. The disappearance of the resistance of mercury”. In: *Communications from the Physical Laboratory of the University of Leiden* 122b (1911). Leiden Communication no. 122b; also Proceedings KNAW Amsterdam 14 (1911). No DOI exists for the Leiden Communications series.

[375] T. von Kármán. “Über den Mechanismus des Widerstandes, den ein bewegter Körper in einer Flüssigkeit erfährt”. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*. The Göttingen Nachrichten are numbered by year (1911), not by volume (academy-proceedings rule). Stability of the vortex street; second part in the 1912 volume, 547-556. 1911, pp. 509–517.

[376] A. E. H. Love. *Some Problems of Geodynamics*. Love waves in a layered half-space. Cambridge: Cambridge University Press, 1911.

[377] R. A. Millikan. “The Isolation of an Ion, a Precision Measurement of its Charge, and the Correction of Stokes’s Law”. In: *Physical Review (Series I)* 32.4 (1911). Physical Review Series I, issue 4 (April 1911). This is the paper carrying the low-pressure correction to Stokes drag that the citing sentence refers to. Journal name must include ’(Series I)’: the same volume/page pair exists in Series II., pp. 349–397. DOI: 10.1103/PhysRevSeriesI.32.349.

[378] W. Nernst. “Der Energieinhalt fester Stoffe”. In: *Annalen der Physik* 341.12 (1911), pp. 395–439. DOI: 10.1002/andp.19113411207.

[379] E. Rutherford. “The scattering of α and β particles by matter and the structure of the atom”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 21.125 (1911). Published title uses Greek letters: The Scattering of α and β Particles by Matter and the Structure of the Atom., pp. 669–688. DOI: 10.1080/14786440508637080.

[380] O. Sackur. “Die Anwendung der kinetischen Theorie der Gase auf chemische Probleme”. In: *Annalen der Physik* 341.15 (1911), pp. 958–980. DOI: 10.1002/andp.19113411505.

[381] M. Born and T. von Kármán. “Über Schwingungen in Raumgittern”. In: *Physikalische Zeitschrift* 13 (1912), pp. 297–309.

[382] P. Debye. “Zur Theorie der spezifischen Wärmen”. In: *Annalen der Physik* 344.14 (1912). Ser. 4, vol. 39, issue 14; Wiley numbering 344. The T^3 law and the Debye temperature. Printed as ’P. Debye’., pp. 789–839. DOI: 10.1002/andp.19123441404.

[383] P. Debye. “Einige Resultate einer kinetischen Theorie der Isolatoren”. In: *Physikalische Zeitschrift* 13 (1912). Vorläufige Mitteilung, Heft 3. Orientation polarization of permanent molecular dipoles and the temperature dependence of the static dielectric constant. NOT one of the two supplied candidates — both of those were the specific-heats paper — but the electrodynamics-of-matter chapter plainly cites Debye 1912 for polar molecules, so this work is supplied. Physikalische Zeitschrift is not in Crossref; no DOI exists, which is acceptable for pre-1990 work., pp. 97–100.

[384] G. Eiffel. “Sur la résistance des sphères dans l’air en mouvement”. In: *Comptes Rendus des Séances de l’Académie des Sciences*. Vol. 155. Discovery of the drag crisis of the sphere. 1912, pp. 1597–1599.

[385] W. Friedrich, P. Knipping, and M. Laue. “Interferenz-Erscheinungen bei Roentgenstrahlen”. In: *Sitzungsberichte der Koeniglich Bayerischen Akademie der Wissenschaften zu Muenchen*. First X-ray diffraction by a crystal, keyed by the first listed author. Umlauts: Röntgenstrahlen, Königlich, München. 1912, pp. 303–322.

[386] E. Grüneisen. “Theorie des festen Zustandes einatomiger Elemente”. In: *Annalen der Physik* 344.12 (1912), pp. 257–306. DOI: 10.1002/andp.19123441202.

[387] V. F. Hess. “Über Beobachtungen der durchdringenden Strahlung bei sieben Freiballonfahrten”. In: *Physikalische Zeitschrift* 13 (1912). Discovery of cosmic rays; no DOI exists for Physikalische Zeitschrift of this period., pp. 1084–1091.

[388] D. Hilbert. *Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen*. Collects the six Göttinger Nachrichten communications of 1904–1910; origin of the word ’spectrum’ and of l2. Leipzig and Berlin: B. G. Teubner, 1912.

[389] M. von Laue. “Eine quantitative Prüfung der Theorie für die Interferenzerscheinungen bei Röntgenstrahlen”. In: *Sitzungsberichte der Mathematisch-Physikalischen Classe der Königlich-Bayerischen Akademie der Wissenschaften zu München*. 1912, pp. 363–373.

[390] H. S. Leavitt and E. C. Pickering. “Periods of 25 Variable Stars in the Small Magellanic Cloud”. In: *Harvard College Observatory Circular* 173 (1912). Observatory circular signed by Pickering, reporting Leavitt’s work; commonly cited under Leavitt. No DOI., pp. 1–3.

[391] F. Paschen and E. Back. “Normale und anomale Zeemaneffekte”. In: *Annalen der Physik* 344.15 (1912). The systematic study of the anomalous Zeeman patterns and of their collapse in strong fields (the Paschen-Back effect)., pp. 897–932. DOI: 10.1002/andp.19123441502.

[392] G. A. Schott. *Electromagnetic Radiation and the Mechanical Reactions Arising from It, being an Adams Prize Essay in the University of Cambridge*. No DOI (1912 monograph). This is the classical calculation of radiation from a charge on a circular orbit, decades before synchrotron light was seen – exactly what sec:rad-synchrotron cites it for. Public domain and scanned on the Internet Archive (electromagneticr00schouoft), which is a declarable open access route for the availability grade. Cambridge University Press, 1912.

[393] H. Tetrode. “Die chemische Konstante der Gase und das elementare Wirkungsquantum”. In: *Annalen der Physik* 343.7 (1912). Contains the comparison with mercury-vapour data, pp. 434–442. DOI: 10.1002/andp.19123430708.

[394] C. T. R. Wilson. “On an expansion apparatus for making visible the tracks of ionising particles in gases and some results obtained by its use”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 87.595 (1912). C. T. R. Wilson; distinct from Kenneth G. Wilson of Wilson:1969, pp. 277–292. DOI: 10.1098/rspa.1912.0081.

[395] N. Bohr. “On the constitution of atoms and molecules”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 26.151 (1913). Part I – the one that derives the Balmer formula and R_infinity from hbar, m and e, which is the claim in the citing sentence. Series 6, vol. 26, issue 151. Crossref records the title with the Phil. Mag. article-number prefix (’I. On the constitution of atoms and molecules’), so a strict string comparison against the DOI will differ by that prefix

and by capitalisation only. Parts II (26, 476–502, DOI 10.1080/14786441308634993) and III (26, 857–875, DOI 10.1080/14786441308635031) are separate works – keep them as Bohr:1913b/c if ever cited., pp. 1–25. DOI: 10.1080/14786441308634955.

[396] N. Bohr. "On the constitution of atoms and molecules". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 26.155 (1913). Part I of the trilogy; suffix a/b/c distinguishes the three 1913 instalments., pp. 857–875. DOI: 10.1080/14786441308635031.

[397] N. Bohr. "On the constitution of atoms and molecules. Part II. Systems containing only a single nucleus". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 6* 26 (1913), pp. 476–502.

[398] N. Bohr. "On the constitution of atoms and molecules. Part III. Systems containing several nuclei". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 6* 26 (1913), pp. 857–875.

[399] W. L. Bragg. "The Diffraction of Short Electromagnetic Waves by a Crystal". In: *Proceedings of the Cambridge Philosophical Society* 17 (1913). Read 11 November 1912, published in the 1913 volume. The paper in which n lambda = 2 d sin theta first appears, recasting Laue's conditions as reflection from lattice planes. No DOI: the Cambridge Philosophical Society proceedings of this period are not registered., pp. 43–57.

[400] W. H. Bragg and W. L. Bragg. "The reflection of X-rays by crystals". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 88.605 (1913). The two-author paper describing the X-ray (ionization) spectrometer and the reflection measurements made with it. DOI verified against Crossref, which titles the container "Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character"., pp. 428–438. DOI: 10.1098/rspa.1913.0040.

[401] L. E. J. Brouwer. "Intuitionism and Formalism". In: *Bulletin of the American Mathematical Society* 20.2 (1913). Translated by Arnold Dresden from the inaugural address at the University of Amsterdam, 14 October 1912, pp. 81–96. DOI: 10.1090/S0002-9904-1913-02440-6.

[402] P. Debye. "Über die Intensitätsverteilung in den mit Röntgenstrahlen erzeugten Interferenzbildern". In: *Verhandlungen der Deutschen Physikalischen Gesellschaft.* Vol. 15. 1913, pp. 678–689.

[403] A. Eucken. "Über das Wärmeleitvermögen, die spezifische Wärme und die innere Reibung der Gase". In: *Physikalische Zeitschrift* 14 (1913). NO DOI given deliberately: Physikalische Zeitschrift (ceased 1945) is not in Crossref, so any DOI here would be invented. Volume 14 is the 1913 volume, so the issue year and the customary year agree. Volume/page range corroborated across four independent citing-literature searches, all giving 'Physik. Z. 14 (1913) 324-332'; note that several modern reference lists mis-expand the abbreviation as 'Z. Phys. 14:324', which is impossible (Zeitschrift für Physik began in 1920) – the venue is Physikalische Zeitschrift. This is the paper that established the Eucken factor kappa*m/(eta*c_V) = (9*gamma-5)/4, equal to 5/2 for a monatomic gas. Being pre-1990 the validator will not demand a DOI, but with no DOI and no stored copy it will grade RED on availability: it needs a docs/library/manifest.json entry, honestly 'access': 'offline' naming volume 14, pages 324-332, unless a verified open scan URL is found (I did not verify one, so I am not asserting it)., pp. 324–332.

[404] H. Geiger and E. Marsden. “The laws of deflexion of a particles through large angles”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 25.148 (1913). Published title uses the Greek letter: The Laws of Deflexion of α Particles through Large Angles., pp. 604–623. DOI: 10.1080/14786440408634197.

[405] W. Kolhörster. “Messungen der durchdringenden Strahlung im Freiballon in größeren Höhen”. In: *Physikalische Zeitschrift* 14 (1913), pp. 1153–1156.

[406] A. Lo Surdo. “Sul fenomeno analogo a quello di Zeeman nel campo elettrico”. In: *Rendiconti della Reale Accademia dei Lincei, Classe di scienze fisiche, matematiche e naturali*. Vol. 22. Series 5, vol. 22; presented at the session of 21 December 1913. This is the priority note that establishes the independent discovery the chapter claims, so it is the right work for the key. No DOI. A fuller version with the same title appeared as Il Nuovo Cimento 7 (1914) 335-337, DOI 10.1007/BF02958602 — that DOI belongs to the Nuovo Cimento record and must NOT be attached to the Rendiconti entry; mention it in a note (or add a separate LoSurdo:1914 entry) so the reader has a retrievable route. 1913, pp. 664–666.

[407] R. A. Millikan. “On the Elementary Electrical Charge and the Avogadro Constant”. In: *Physical Review* 2 (1913). Physical Review Series II, vol. 2, issue 2, 1 August 1913. Crossref prints the author family name as ’Millikan.’ with a stray period; the correct form is Millikan., pp. 109–143.

[408] H. Moseley. “The high-frequency spectra of the elements”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 26.156 (1913), pp. 1024–1034. DOI: 10.1080/14786441308635052.

[409] J. Perrin. *Les Atomes*. Paris: Félix Alcan, 1913.

[410] G. Sagnac. “L’éther lumineux démontré par l’effet du vent relatif d’éther dans un interféromètre en rotation uniforme”. In: *Comptes rendus hebdomadaires des séances de l’Académie des Sciences*. Vol. 157. No DOI. Comptes rendus tome 157, séance of 27 October 1913 (issue no. 17) – Comptes rendus does carry tome numbers, so 157 is real and not invented. This first note announces the effect. The chapter also promises ’the measured fringe shift versus rotation rate’, which is in Sagnac’s companion note in the SAME tome: ’Sur la preuve de la réalité de l’éther lumineux par l’expérience de l’interférographe tournant’, C. R. Acad. Sci. 157, 1410-1413 (1913). If the measurement is to be cited, that second note is the source and deserves its own key (Sagnac:1913b). 1913, pp. 708–710.

[411] A. VAN DER BROEK. “Intra-atomic Charge”. In: *Nature* 92.2300 (1913). Issue 2300. The hypothesis that the nuclear charge equals the serial number of the element, which Moseley then tested. Crossref’s deposited metadata misspells the author as "A. Van der Broek"; Nature printed "A. van den Broek" — an instance of the corrupt-Crossref-metadata trap in CLAUDE.md, so take the name from the journal and the DOI/volume/pages from Crossref. Key follows the existing vonLaue:1912 convention for particled surnames. Do not confuse with his other 1913 Nature note, "Intra-atomic Charge and the Structure of the Atom", 92, 476-478, doi 10.1038/092476b0., pp. 372–373. DOI: 10.1038/092372c0.

[412] L. Brillouin. “Über die Fortpflanzung des Lichtes in dispergierenden Medien”. In: *Annalen der Physik* 349.10 (1914). TRAP: Sommerfeld’s companion paper in the very same issue carries the IDENTICAL title and is DOI 10.1002/andp.19143491002, pp. 177-202. Brillouin’s is the SECOND of the pair, pp. 203-240, DOI ...1003 – both Crossref-confirmed. Annalen vol. 349 = 4th series vol. 44; issue 10. The citing chapter cites Sommerfeld:1914 and Brillouin:1914 together, so getting the two the right way round matters., pp. 203–240. DOI: 10.1002/andp.19143491003.

[413] E. Buckingham. “On Physically Similar Systems; Illustrations of the Use of Dimensional Equations”. In: *Physical Review* 4.4 (1914). The Π theorem. Priority is contested: the general statement was published by Vaschy (1892) and, apparently independently, by Federman and Riabouchinsky (both 1911), and Buckingham himself later ceded priority to Riabouchinsky. Note also that Buckingham defines k as the size of a maximal independent set of quantities, i.e. the rank of the dimensional matrix, not the number of base dimensions, pp. 345–376. DOI: 10.1103/PhysRev.4.345.

[414] J. Chadwick. “Intensitatsverteilung im magnetischen Spektrum der beta-Strahlen von Radium B+C”. In: *Verhandlungen der Deutschen Physikalischen Gesellschaft*. Vol. 16. No DOI; first demonstration that the beta spectrum is continuous. 1914, pp. 383–391.

[415] A. D. Fokker. “Die mittlere Energie rotierender elektrischer Dipole im Strahlungsfeld”. In: *Annalen der Physik* 348.5 (1914), pp. 810–820. DOI: 10.1002/andp.19143480507.

[416] J. Franck and G. Hertz. “Ueber die Erregung der Quecksilberresonanzlinie 253,6 Mikromikrometer durch Elektronenstoesse”. In: *Verhandlungen der Deutschen Physikalischen Gesellschaft*. Vol. 16. Suffix ’b’; the companion paper reporting the 253.6 nm resonance line excited by electron impact. Published title writes the wavelength as 253,6 mu-mu (micromicrons), i.e. 253.6 nm in SI. Umlaut/eszett: Elektronenstoße. 1914, pp. 512–517.

[417] J. Franck and G. Hertz. “Ueber Zusammenstoesse zwischen Elektronen und den Molekuelen des Quecksilberdampfes und die Ionisierungsspannung desselben”. In: *Verhandlungen der Deutschen Physikalischen Gesellschaft*. Vol. 16. Suffix ’a’ because a second Franck-Hertz paper of 1914 is cited as Franck:1914b. Society proceedings, not a journal article. The title itself carries the authors’ misreading of the 4.9 V threshold as an ionization potential, corrected in Bohr:1915. Umlauts and eszett: Zusammenstöße, Molekülen. 1914, pp. 457–467.

[418] H. Moseley. “The high-frequency spectra of the elements. Part II”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 27.160 (1914), pp. 703–713. DOI: 10.1080/14786440408635141.

[419] L. S. Ornstein and F. Zernike. “Accidental deviations of density and opalescence at the critical point of a single substance”. In: *Proceedings of the Royal Netherlands Academy of Arts and Sciences*. Vol. 17. 1914, pp. 793–806.

[420] H. N. Russell. “Relations Between the Spectra and other Characteristics of the Stars ast;” in: *Nature* 93.2322 (1914). No DOI, pp. 227–230. DOI: 10.1038/093227b0.

[421] A. Sommerfeld. “Über die Fortpflanzung des Lichtes in dispergierenden Medien”. In: *Annalen der Physik* 349.10 (1914). Crossref-verified. Brillouin’s companion paper cited in the same sentence carries the IDENTICAL title and immediately follows it in the same volume: Brillouin, L., Ann. Phys. 349, 203-240 (1914), doi 10.1002/andp.19143491003 – if Brillouin:1914 is also orphaned, that is the entry, and the two must be distinguished by author and pages, not by title. Wiley volume 349 = Band 44., pp. 177–202. DOI: 10.1002/andp.19143491002.

[422] J. Stark. “Beobachtungen über den Effekt des elektrischen Feldes auf Spektrallinien. I. Quereffekt”. In: *Annalen der Physik* 348.7 (1914). The discovery was announced in the Sitzungsberichte of the Berlin Academy in November 1913; this Annalen paper of 1914 is the full account, and the entry is keyed by its publication year per the treatise convention., pp. 965–982. DOI: 10.1002/andp.19143480702.

[423] W. S. Adams. “THE SPECTRUM OF THE COMPANION OF SIRIUS”. In: *Publications of the Astronomical Society of the Pacific* 27.161 (1915). First spectrum of a white dwarf, pp. 236–236. DOI: 10.1086/122440.

[424] S. J. Barnett. “Magnetization by Rotation”. In: *Physical Review* 6.4 (1915), pp. 239–270. DOI: 10.1103/PhysRev.6.239.

[425] N. Bohr. “On the Quantum theory of radiation and the structure of the atom”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 30.177 (1915). Contains the correction of Franck and Hertz’s reading of their own 4.9 V threshold as an ionization rather than an excitation potential., pp. 394–415. DOI: 10.1080/14786440908635413.

[426] A. Einstein and W. J. de Haas. “Experimenteller Nachweis der Ampèreschen Molekularströme”. In: *Verhandlungen der Deutschen Physikalischen Gesellschaft*. Vol. 17. First measurement of the gyromagnetic ratio of iron (the Einstein–de Haas effect). Also issued in English by the KNAW (Amsterdam) proceedings. Keyed with suffix c, not a, because Einstein:1915a (perihelion) and Einstein:1915b (field equations) are the two November 1915 Sitzungsberichte papers already cited throughout Part V; renumbering them would have rewritten correct references. The volume number 17 is the genuine Jahrgang of the series, not an invented field. No DOI exists. 1915, pp. 152–170.

[427] A. Einstein. “Die Feldgleichungen der Gravitation”. In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)*. Academy proceedings, series has NO volume number – @inproceedings. The final field equations, presented 25 November 1915. 1915, pp. 844–847.

[428] A. Einstein. “Erklärung der Perihelbewegung des Merkur aus der allgemeinen Relativitätstheorie”. In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)*. Academy proceedings, series has NO volume number – @inproceedings with booktitle, never invent a volume. Suffix a/b: two November 1915 Sitzungsberichte papers; a = perihelion (18 Nov), b = field equations (25 Nov). Also contains the full 1.7 arcsec light deflection. 1915, pp. 831–839.

[429] A. Einstein. “Zur allgemeinen Relativitätstheorie”. In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)*. Academy proceedings, series has NO volume number. Presented 4 November 1915. Suffix d: a = perihelion, b = field equations, c = Einstein–de Haas; cite by content, never by letter. 1915, pp. 778–786.

[430] A. Einstein. “Zur allgemeinen Relativitätstheorie (Nachtrag)”. In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)*. Academy proceedings, series has NO volume number. Presented 11 November 1915; the Nachtrag to Einstein:1915d. 1915, pp. 799–801.

[431] D. Hilbert. “Die Grundlagen der Physik. (Erste Mitteilung.)” In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse*. Communicated 20 November 1915; the issue was printed in 1916 but the Nachrichten volume and customary citation year are 1915. Proceedings series, no volume number – @inproceedings. 1915, pp. 395–407.

[432] L. Tonelli. “Sur une méthode directe du calcul des variations”. In: *Rendiconti del Circolo Matematico di Palermo* 39.1 (1915). Founding paper of the direct method (lower semicontinuity + coercivity + convexity)., pp. 233–264. DOI: 10.1007/bf03015981.

[433] E. T. Whittaker. “XVIII.—On the Functions which are represented by the Expansions of the Interpolation-Theory”. In: *Proceedings of the Royal Society of Edinburgh* 35 (1915). The cardinal series; earliest form of the sampling theorem., pp. 181–194. DOI: 10.1017/s0370164600017806.

[434] S. Chapman. “On the law of distribution of molecular velocities, and on the theory of viscosity and thermal conduction, in a non-uniform simple monatomic gas”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 216.538-548 (1916), pp. 279–348. DOI: 10.1098/rsta.1916.0006.

[435] A. Einstein. “Die Grundlage der allgemeinen Relativitätstheorie”. In: *Annalen der Physik* 354.7 (1916). Suffix a because key Einstein:1916 is already taken by the Naeherungsweise-Integration GW paper. Volume 354 continuous numbering (= ser. 4 vol. 49). The 1916 review presentation of general relativity., pp. 769–822. DOI: 10.1002/andp.19163540702.

[436] A. Einstein. “Näherungsweise Integration der Feldgleichungen der Gravitation”. In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)*. Vol. 1916, Erster Halbband. Read before the Academy on 22 June 1916. The Academy proceedings are issued annually in half-volumes and carry no journal volume number. 1916, pp. 688–696.

[437] R. A. Millikan. “A Direct Photoelectric Determination of Planck’s "h"”. In: *Physical Review* 7.3 (1916). Published title puts h in quotation marks: A Direct Photoelectric Determination of Planck’s "h"., pp. 355–388. DOI: 10.1103/PhysRev.7.355.

[438] L. Rayleigh. “On convection currents in a horizontal layer of fluid, when the higher temperature is on the under side”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 32.192 (1916). The critical-Rayleigh-number calculation for Bénard convection., pp. 529–546. DOI: 10.1080/14786441608635602.

[439] H. Reissner. “Über die Eigengravitation des elektrischen Feldes nach der Einsteinschen Theorie”. In: *Annalen der Physik* 355.9 (1916). Volume 355 continuous numbering (= ser. 4 vol. 50)., pp. 106–120. DOI: 10.1002/andp.19163550905.

[440] K. Schwarzschild. “Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie”. In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)*. Academy proceedings, series has NO volume number – @inproceedings. The exterior point-mass solution; the interior (incompressible-fluid) companion paper is pp. 424–434 of the same year, not entered here. 1916, pp. 189–196.

[441] A. Sommerfeld. “Zur Quantentheorie der Spektrallinien”. In: *Annalen der Physik* 356 (1916), pp. 1–94. DOI: 10.1002/andp.19163561702.

[442] T. J. I. Bromwich. “Normal Coordinates in Dynamical Systems”. In: *Proceedings of the London Mathematical Society* s2-15.1 (1917). Second series (s2-15). Source of the Bromwich inversion contour justifying Heaviside’s operational calculus., pp. 401–448. DOI: 10.1112/plms/s2-15.1.401.

[443] B. Davis and F. S. Goucher. “Ionization and Excitation of Radiation by Electron Impact in Mercury Vapor and Hydrogen”. In: *Physical Review* 10.2 (1917). Separated excitation from ionization experimentally, confirming Bohr’s reinterpretation of the Franck-Hertz threshold., pp. 101–115. DOI: 10.1103/PhysRev.10.101.

[444] J. Droste. “The field of a single centre in Einstein’s theory of gravitation, and the motion of a particle in that field”. In: *Koninklijke Nederlandsche Akademie van Wetenschappen Proceedings*. Vol. 19. Communicated May 1916, volume printed 1917. Independent derivation in the now-standard coordinates. 1917, pp. 197–215.

[445] A. Einstein. “Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie”. In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)*. Read before the Prussian Academy on 8 February 1917. Introduces the cosmological constant and the static Einstein universe. Academy proceedings: the series carries no journal volume number. No DOI exists. 1917, pp. 142–152.

[446] A. Einstein. “Zur Quantentheorie der Strahlung”. In: *Physikalische Zeitschrift* 18 (1917). The A and B coefficients, spontaneous and stimulated emission from detailed balance, and the momentum h*nu/c carried by an emitted quantum. First printed in Mitteilungen der Physikalischen Gesellschaft Zürich 18 (1916), 47–62; the 1917 Physikalische Zeitschrift printing is the one universally cited. No DOI exists., pp. 121–128.

[447] D. Enskog. *Kinetische Theorie der Vorgänge in mässig verdünnten Gasen*. Doctoral dissertation, Uppsala. Uppsala: Almqvist & Wiksell, 1917.

[448] M. Planck. “Über einen Satz der statistischen Dynamik und seine Erweiterung in der Quantentheorie”. In: *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*. 1917, pp. 324–341.

[449] J. Radon. “Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten”. In: *Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*. Vol. 69. Academy proceedings series that does carry volume numbers (69 is real, not invented). 1917, pp. 262–277.

[450] W. de Sitter. “On Einstein’s Theory of Gravitation and its Astronomical Consequences. Third Paper”. In: *Monthly Notices of the Royal Astronomical Society* 78.1 (1917). Third of the three 1916–1917 MNRAS papers; contains the de Sitter solution., pp. 3–28. DOI: 10.1093/mnras/78.1.3.

[451] V. M. Slipher. “Theory of Rotation of Spiral Nebulae”. In: *Scientific American* 84.2182supp (1917). First systematic radial velocities of spiral nebulae. No DOI., pp. 261–261. DOI: 10.1038/scientificamerican10271917-261csupp.

[452] G. Duffing. *Erzwungene Schwingungen bei veränderlicher Eigenfrequenz und ihre technische Bedeutung*. The Duffing oscillator: amplitude-dependent frequency, bent resonance curve, jump phenomenon. Friedrich Vieweg und Sohn, Braunschweig (Sammlung Vieweg, Heft 41/42), 1918.

[453] A. Einstein. “Über Gravitationswellen”. In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)*. Vol. 1918, Erster Halbband. Read before the Academy on 31 January 1918. Gives the quadrupole radiation formula in its correct form, repairing a computational slip of Einstein (1916). 1918, pp. 154–167.

[454] E. Kretschmann. “Über den physikalischen Sinn der Relativitätspostulate, A. Einsteins neue und seine ursprüngliche Relativitätstheorie”. In: *Annalen der Physik* 358.16 (1918). Widely cited as Kretschmann (1917), the date of the argument; the publisher’s record dates the issue (Vierte Folge, Band 53) to January 1918, pp. 575–614. DOI: 10.1002/andp.19183581602.

[455] J. Lense and H. Thirring. “Über den Einfluß der Eigenrotation der Zentralkörper auf die Bewegung der Planeten und Monde nach der Einsteinschen Gravitationstheorie”. In: *Physikalische Zeitschrift* 19 (1918). The frame-dragging (Lense–Thirring) paper proper: the effect of a rotating central body on the orbits of planets and moons. Physikalische Zeitschrift is defunct and has no DOIs; do NOT attach 10.1007/BF00762913 or similar, which belong to the 1984 English translation by Mashhoon, Hehl and Theiss in Gen. Relativ. Gravit. 16, 727–741, a different record.

ADS bibcode 1918PhyZ...19..156L. Being pre-1990 it is DOI-exempt, but it needs an "access": "offline" (or open, via a scan) manifest entry naming volume 19, pp. 156–163, or it grades RED as a DOI-less @article., pp. 156–163.

[456] E. Noether. “Invariante Variationsprobleme”. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse*. Academy proceedings series without volume numbers – @inproceedings per house convention (same as the Einstein Sitzungsberichte entries). Contains BOTH Noether theorems. 1918, pp. 235–257.

[457] G. Nordström. “On the Energy of the Gravitation Field in Einstein’s Theory”. In: *Koninklijke Nederlandsche Akademie van Wetenschappen Proceedings*. Vol. 20. Completes the charged static solution (Reissner–Nordstrom). 1918, pp. 1238–1245.

[458] H. Thirring. “Über die Wirkung rotierender ferner Massen in der Einsteinschen Gravitationstheorie”. In: *Physikalische Zeitschrift* 19 (1918). Thirring’s companion paper (sole author), the gravitomagnetic field inside a rotating mass shell. ADS bibcode 1918PhyZ...19...33T. Errata/corrections appeared in Phys. Z. 22, 29–30 (1921). No DOI for the original; the 1984 English translation is Gen. Relativ. Gravit. 16, 712–725 and is a separate record. An open scan of the original offprint is at PHAIDRA, University of Vienna (o:137123), which is a usable GREEN access route. Cited together with Lense:1918 at the frame-dragging sentence opening sec:clt-gyroscopes., pp. 33–39.

[459] H. Barkhausen. “Zwei mit Hilfe der neuen Verstärker entdeckte Erscheinungen”. In: *Physikalische Zeitschrift* 20 (1919), pp. 401–403.

[460] J. Larmor. “How could a rotating body such as the Sun become a magnet?” In: *Report of the British Association for the Advancement of Science*. 87th Meeting of the British Association (Bournemouth, 1917 meeting series; report published 1919). The series has no journal volume number — the ’87’ seen in some citations is the meeting number, not a volume; do not put it in a volume field, put it in the booktitle as ’87th Meeting’. No DOI exists. Full text is on Wikisource, which is a usable access route for the manifest. 1919, pp. 159–160.

[461] E. Rutherford. “Collision of α particles with light atoms. IV. An anomalous effect in nitrogen”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 37.222 (1919). Series 6, issue 222, June 1919; the first artificial transmutation. Crossref returns the title with the volume article number prefixed and italicised: ’LIV. Collision of alpha particles with light atoms. IV. An anomalous effect in nitrogen’. The ’LIV.’ is dropped per the house convention already applied to Cavendish:1798, and the Greek letter is written α as in the sibling entry Rutherford:1911. This is paper IV of a set of four in the same volume (I Hydrogen, 537–561, DOI 10.1080/14786440608635916; II 562–571, ...917; III 571–580, ...918) with near-identical titles – cite IV, the nitrogen paper, and check the roman numeral before reusing any of the other three DOIs. Verified against api.crossref.org 2026-08-13., pp. 581–587. DOI: 10.1080/14786440608635919.

[462] M. Siegbahn. “Precision-measurements in the X-ray spectra”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 37.222 (1919). Series 6, issue 222 (June 1919); sole author, confirmed against Crossref. Crossref prefixes the title with the article number "LVIII.". Parts II and III are 10.1080/14786441108635992 and 10.1080/14786441108635993 (volume 38, pages 639-646 and 647-651, 1919) if a fuller set is wanted., pp. 601–612. DOI: 10.1080/14786440608635923.

[463] N. Bohr. “Ueber die Serienspektra der Elemente”. In: *Zeitschrift fuer Physik* 2.5 (1920). The correspondence principle stated as such. Umlaut: Über die Serienspektra der Elemente., pp. 423–469. DOI: 10.1007/BF01329978.

[464] J. Chadwick. “The charge on the atomic nucleus and the law of force”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 40.240 (1920). Series 6, issue 238. The direct scattering determination of the nuclear charge of copper, silver and platinum, agreeing with the atomic numbers to about one per cent. Crossref prefixes the title with the article number "LXXXIV." and records the page range as 734-745; some secondary sources give 734-746. Note the existing Chadwick keys already use the plain-year form (Chadwick:1914, :1930, :1934), so Chadwick:1920 does not collide., pp. 734–745. DOI: 10.1080/14786441208636183.

[465] F. W. Dyson, A. S. Eddington, and C. Davidson. “A determination of the deflection of light by the sun’s gravitational field, from observations made at the total eclipse of May 29, 1919”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 220.571-581 (1920), pp. 291–333. DOI: 10.1098/rsta.1920.0009.

[466] A. S. Eddington. “The Internal Constitution of the Stars”. In: *Science* 52.1341 (1920). BA address proposing subatomic energy as the stellar source; same title as the 1926 book; no DOI, pp. 233–240. DOI: 10.1126/science.52.1341.233.

[467] J. Franck and E. Einsporn. “Ueber die Anregungspotentiale des Quecksilberdampfes”. In: *Zeitschrift fuer Physik* 2.1 (1920). Resolved the metastable 4.67 eV level from the radiating 4.89 eV level; the paper in which Franck accepts Bohr’s correction., pp. 18–29. DOI: 10.1007/bf01333785.

[468] W. Lenz. “Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern”. In: *Physikalische Zeitschrift* 21 (1920), pp. 613–615.

[469] A. A. Michelson. “On the Application of Interference Methods to Astronomical Measurements”. In: *The Astrophysical Journal* 51 (1920). ADS 1920ApJ....51..257M. Crossref-confirmed title/volume/first page; the range 257-262 is corroborated independently. A short companion abstract with the same title ran in PNAS 6 (1920) 474-475, DOI 10.1073/pnas.6.8.474 — do not swap the DOIs. This is the stellar beam interferometer proposal; the measurement itself is Michelson:1921., p. 257. DOI: 10.1086/142550.

[470] M. N. Saha. “Ionization in the solar chromosphere”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 40.238 (1920). Crossref-verified. Philosophical Magazine series 6; Crossref’s deposited title is prefixed ’LIII. ’, dropped here. The byline is printed ’Megh Nad Saha’. A short companion appeared as ’Ionisation in the Solar Chromosphere’, Nature 105, 232-233 (1920), doi 10.1038/105232b0 – the Phil. Mag. paper is the substantive one and is the standard source of the Saha equation., pp. 472–488. DOI: 10.1080/14786441008636148.

[471] O. Stern. “Eine direkte Messung der thermischen Molekulargeschwindigkeit”. In: *Zeitschrift für Physik* 2.1 (1920), pp. 49–56. DOI: 10.1007/bf01333787.

[472] E. Bessel-Hagen. “Über die Erhaltungssätze der Elektrodynamik”. In: *Mathematische Annalen* 84.3-4 (1921). The conformal currents of electrodynamics extracted from Noether’s theorems, exactly the claim in the citing sentence. The surname is hyphenated in print (’Bessel-Hagen’), which is why the key has no punctuation; keep the hyphen in the author field. Crossref-verified., pp. 258–276. DOI: 10.1007/BF01459410.

[473] A. Einstein. “Geometrie und Erfahrung”. In: *Sitzungsberichte der Preußischen Akademie der Wissenschaften (Berlin)*. Festvortrag of 27 January 1921. The Academy proceedings carry no volume number. An expanded version appeared as a separate booklet, Julius Springer, Berlin (1921). 1921, pp. 123–130.

[474] P. P. Ewald. “Die Berechnung optischer und elektrostatischer Gitterpotentiale”. In: *Annalen der Physik* 369.3 (1921), pp. 253–287. DOI: 10.1002/andp.19213690304.

[475] A. A. Griffith. “The phenomena of rupture and flow in solids”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 221.582-593 (1921). Energy criterion for brittle fracture., pp. 163–198. DOI: 10.1098/rsta.1921.0006.

[476] J. T. Jebsen. “Über die allgemeinen kugelsymmetrischen Lösungen der Einsteinschen Gravitationsgleichungen im Vakuum”. In: *Arkiv för Matematik, Astronomi och Fysik* 15 (1921). No. 18 of vol. 15. Proves the theorem customarily attributed to Birkhoff (1923) two years earlier; cited for priority, as the chapter states., pp. 1–9.

[477] C. Kuratowski. “Sur la notion de l’ordre dans la Théorie des Ensembles”. In: *Fundamenta Mathematicae* 2.1 (1921). The set-theoretic definition of the ordered pair. Published under the French form of the author’s name; Polish form Kazimierz Kuratowski, pp. 161–171. DOI: 10.4064/fm-2-1-161-171.

[478] A. Landé. “Über den anomalen Zeemaneffekt (Teil I)”. In: *Zeitschrift für Physik* 5 (1921), pp. 231–241.

[479] A. A. Michelson and F. G. Pease. “Measurement of the diameter of alpha Orionis with the interferometer”. In: *The Astrophysical Journal* 53 (1921). ADS 1921ApJ....53..249M, May 1921. Crossref renders the title with ’alpha Orionis’ spelled out — if the DOI title check is character-strict, that is the form it will see; the journal prints the Greek letter. The measured Betelgeuse diameter the chapter quotes is here. A separate PNAS note, ’Measurement of the Diameter of Alpha-Orionis by the Interferometer’, PNAS 7 (1921) 143-146, DOI 10.1073/pnas.7.5.143, is a different record — do not merge them., p. 249. DOI: 10.1086/142603.

[480] C. Ramsauer. “Über den Wirkungsquerschnitt der Gasmoleküle gegenüber langsamen Elektronen”. In: *Annalen der Physik* 369.6 (1921). Volume 369 in the continuous Annalen numbering (Band 64 of the old series)., pp. 513–540. DOI: 10.1002/andp.19213690603.

[481] O. Stern. “Ein Weg zur experimentellen Prüfung der Richtungsquantelung im Magnetfeld”. In: *Zeitschrift für Physik* 7 (1921), pp. 249–253. DOI: 10.1007/BF01332793.

[482] S. Timoshenko. “On the correction for shear of the differential equation for transverse vibrations of prismatic bars”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 41.245 (1921). Series 6. Shear correction to Euler-Bernoulli beam theory., pp. 744–746. DOI: 10.1080/14786442108636264.

[483] J. Valasek. “Piezo-Electric and Allied Phenomena in Rochelle Salt”. In: *Physical Review* 17.4 (1921). Verified via Crossref (Phys. Rev. 17(4), 1 Apr 1921). This is the paper reporting ferroelectric hysteresis in Rochelle salt, which is exactly the claim made at sec:emmat-ferroelectric., pp. 475–481. DOI: 10.1103/PhysRev.17.475.

[484] H.-J. Van Leeuwen. “Problèmes de la théorie électronique du magnétisme”. In: *Journal de Physique et le Radium* 2.12 (1921), pp. 361–377. DOI: 10.1051/jphysrad:01921002012036100.

[485] P. W. Bridgman. *Dimensional Analysis*. Distinct from the same author’s 1927 The Logic of Modern Physics. New Haven: Yale University Press, 1922.

[486] R. von Eötvös, D. Pekár, and E. Fekete. “Beiträge zum Gesetze der Proportionalität von Trägheit und Gravität”. In: *Annalen der Physik* 373 (1922). Experiments performed 1906–1909, published posthumously. Volume 373 in the continuous Annalen numbering (= ser. 4 vol. 68); key by the Crossref volume., pp. 11–66.

[487] R. A. Fisher. “On the mathematical foundations of theoretical statistics”. In: *Philosophical Transactions of the Royal Society of London A* 222 (1922). Crossref-verified: author, year, volume and pages all confirmed against the DOI. The paper that introduces the terms parameter, statistic, likelihood, sufficiency, consistency and efficiency, and defines the information later named after Fisher. Cited in the probability chapter for the score and the information., pp. 309–368. DOI: 10.1098/rsta.1922.0009.

[488] A. Fraenkel. “Zu den Grundlagen der Cantor-Zermeloschen Mengenlehre”. In: *Mathematische Annalen* 86.3-4 (1922), pp. 230–237. DOI: 10.1007/BF01457986.

[489] G. Friedel. “Les états mésomorphes de la matière”. In: *Annales de Physique* 9.18 (1922), pp. 273–474. DOI: 10.1051/anphys/192209180273.

[490] A. Friedmann. “Über die Krümmung des Raumes”. In: *Zeitschrift für Physik* 10 (1922), pp. 377–386.

[491] W. Gerlach and O. Stern. “Das magnetische Moment des Silberatoms”. In: *Zeitschrift für Physik* 9 (1922), pp. 353–355. DOI: 10.1007/BF01326984.

[492] W. Gerlach and O. Stern. “Der experimentelle Nachweis der Richtungsquantelung im Magnetfeld”. In: *Zeitschrift für Physik* 9 (1922), pp. 349–352. DOI: 10.1007/BF01326983.

[493] J. W. Lindeberg. “Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung”. In: *Mathematische Zeitschrift* 15 (1922), pp. 211–225. DOI: 10.1007/BF01494395.

[494] L. F. Richardson. *Weather Prediction by Numerical Process*. Contains the energy-cascade picture (’big whorls have little whorls...’). Cambridge University Press, 1922.

[495] J. Townsend and V. Bailey. “The motion of electrons in argon”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 43.255 (1922). Series 6. DOI not supplied – not confident enough., pp. 593–600. DOI: 10.1080/14786442208633916.

[496] L. Wittgenstein. *Tractatus Logico-Philosophicus*. Bilingual first English edition, translated by C. K. Ogden with F. P. Ramsey, introduction by Bertrand Russell. The German original appeared as Logisch-philosophische Abhandlung, Annalen der Naturphilosophie 14 (1921), 185–262, printed without the author’s corrections. London: Kegan Paul, Trench, Trubner and Co., 1922.

[497] G. D. Birkhoff. *Relativity and Modern Physics*. Written with the cooperation of R. E. Langer. Customary source of Birkhoff’s theorem; priority belongs to Jebsen (1921), which the chapter says. Harvard University Press, 1923.

[498] N. BOHR. “The Structure of the Atom1”. In: *Nature* 112.2801 (1923). Bohr’s Nobel lecture (delivered December 1922, published 1923); the building-up principle and the periodic system. The earlier German statement is Z. Phys. 9, 1 (1922)., pp. 29–44. DOI: 10.1038/112029a0.

[499] A. H. Compton. “A Quantum Theory of the Scattering of X-rays by Light Elements”. In: *Physical Review* 21.5 (1923), pp. 483–502. DOI: 10.1103/PhysRev.21.483.

[500] L. DE BROGLIE. “Waves and Quanta”. In: *Nature* 112.2815 (1923). Key follows the existing deAzcarraga:2003 pattern for a name with a particle., pp. 540–540. DOI: 10.1038/112540a0.

[501] P. Debye. “Zerstreuung von Roentgenstrahlen und Quantentheorie”. In: *Physikalische Zeitschrift* 24 (1923). Independent derivation of the Compton relation, published within weeks of Compton:1923 and without knowledge of the experiment. Umlaut: Roentgenstrahlen = Röntgenstrahlen., pp. 161–166.

[502] J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Settles the Huygens-principle dichotomy between odd and even space dimensions. New Haven: Yale University Press, 1923.

[503] H. A. Kramers. “On the theory of X-ray absorption and of the continuous X-ray spectrum”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 46.275 (1923). Source of the Kramers opacity law, pp. 836–871. DOI: 10.1080/14786442308565244.

[504] A. Smekal. “Zur Quantentheorie der Dispersion”. In: *Die Naturwissenschaften* 11.43 (1923), pp. 873–875. DOI: 10.1007/BF01576902.

[505] G. I. Taylor. “Stability of a viscous liquid contained between two rotating cylinders”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 223.605-615 (1923). G. I. Taylor – distinct from the existing keys Taylor:1982 (J. H. Taylor) and Taylor:1997 (J. R. Taylor). Theory and experiment for Taylor-Couette stability., pp. 289–343. DOI: 10.1098/rsta.1923.0008.

[506] I. Waller. “Zur Frage der Einwirkung der Wärmebewegung auf die Interferenz von Röntgenstrahlen”. In: *Zeitschrift für Physik* 17.1 (1923), pp. 398–408. DOI: 10.1007/BF01328696.

[507] N. Wiener. “Differential-Space”. In: *Journal of Mathematics and Physics* 2.1-4 (1923), pp. 131–174. DOI: 10.1002/sapm192321131.

[508] N. Bohr, H. Kramers, and J. Slater. “The quantum theory of radiation”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 47.281 (1924). The BKS proposal, in which energy and momentum are conserved only statistically; falsified by Bothe:1925 and Compton:1925., pp. 785–802. DOI: 10.1080/14786442408565262.

[509] Bose. “Plancks Gesetz und Lichtquantenhypothese”. In: *Zeitschrift fuer Physik* 26.1 (1924). Translated into German and submitted by Einstein. Venue with umlaut: Zeitschrift für Physik., pp. 178–181. DOI: 10.1007/BF01327326.

[510] A. S. EDDINGTON. “A Comparison of Whitehead's and Einstein's Formulæ”. In: *Nature* 113.2832 (1924). Issue of 9 February 1924; takes 'a' as the earlier of the two Eddington 1924 papers. A one-page note that introduces what are now called Eddington–Finkelstein coordinates, without the horizon interpretation. Nature prints the ligature 'Formulæ'., pp. 192–192. DOI: 10.1038/113192a0.

[511] A. S. Eddington. “On the Relation between the Masses and Luminosities of the Stars.: (Plate 8.)” In: *Monthly Notices of the Royal Astronomical Society* 84.5 (1924). Issue 5, March 1924. Crossref records the end page as 333 and appends ': (Plate 8.)' to the title; ADS and the customary citation give 308–332., pp. 308–333. DOI: 10.1093/mnras/84.5.308.

[512] A. Einstein. “Quantentheorie des einatomigen idealen Gases”. In: *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-mathematische Klasse*. First paper of the pair; the condensation conclusion is in the 1925 sequel. 1924, pp. 261–267.

[513] A. Friedmann. “Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes”. In: *Zeitschrift für Physik* 21.1 (1924), pp. 326–332. DOI: 10.1007/BF01328280.

[514] W. Lenz. “Über den Bewegungsverlauf und die Quantenzustände der gestörten Keplerbewegung”. In: *Zeitschrift für Physik* 24.1 (1924). Issue 1. Crossref-confirmed. This is Wilhelm Lenz (Runge-Lenz vector in the old quantum theory) — a different man from Emil Lenz of Lenz:1834, and both keys sit in this batch; the bibliography should make the given names explicit., pp. 197–207. DOI: 10.1007/BF01327245.

[515] R. S. MULLIKEN. “The Band Spectrum of Boron Monoxide”. In: *Nature* 114.2862 (1924). Nature issue 2862, September 1924. This is the note in which Mulliken states the half-quantum (zero-point) vibrational offset, and it is the one cited as the first direct evidence for zero-point energy – NOT his other 1924 Nature note ’Isotope Effects in the Band Spectra of Boron Monoxide and Silicon Nitride’, Nature 113, 423–424, DOI 10.1038/113423a0, and not the full report ’The Isotope Effect in Band Spectra, II: The Spectrum of Boron Monoxide’, Phys. Rev. 25, 259–294 (1925), DOI 10.1103/PhysRev.25.259. If the chapter later wants the detailed data rather than the announcement, the 1925 Phys. Rev. paper is the right key, and it would be Mulliken:1925., pp. 349–350. DOI: 10.1038/114349a0.

[516] W. Pauli. “Zur Frage der theoretischen Deutung der Satelliten einiger Spektrallinien und ihrer Beeinflussung durch magnetische Felder”. In: *Naturwissenschaften* 12.37 (1924), pp. 741–743. DOI: 10.1007/bf01504828.

[517] E. V. APPLETON and M. A. F. BARNETT. “Local Reflection of Wireless Waves from the Upper Atmosphere”. In: *Nature* 115.2888 (1925). The ionospheric-reflection demonstration by Appleton and Barnett named in the prose. Crossref-verified. Distinct from their 1926 Proc. R. Soc. A 113, 450 paper on the same experiment., pp. 333–334. DOI: 10.1038/115333a0.

[518] P. M. S. Blackett. “The ejection of protons from nitrogen nuclei, photographed by the Wilson method”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 107.742 (1925). The cloud-chamber identification of the reaction Rutherford:1919 had only inferred: forked tracks with exactly two branches and no third show the alpha particle was captured, fixing the reaction as N-14 + He-4 → O-17 + H-1. Distinct from the existing Blackett:1933 (Blackett and Occhialini, pair production). Verified against api.crossref.org 2026-08-13., pp. 349–360. DOI: 10.1098/rspa.1925.0029.

[519] W. Bothe and H. Geiger. “Ueber das Wesen des Comptoneffekts; ein experimenteller Beitrag zur Theorie der Strahlung”. In: *Zeitschrift fuer Physik* 32.1 (1925). The founding use of the coincidence method., pp. 639–663. DOI: 10.1007/BF01331702.

[520] A. H. Compton and A. W. Simon. “Directed Quanta of Scattered X-Rays”. In: *Physical Review* 26.3 (1925), pp. 289–299. DOI: 10.1103/PhysRev.26.289.

[521] L. De Broglie. “Recherches sur la théorie des Quanta”. In: *Annales de Physique* 10.3 (1925). Doctoral thesis defended in November 1924 and printed in Annales de Physique in 1925; keyed by the publication year per the treatise convention, with the customary 1924 dating recorded here. Accent: théorie., pp. 22–128. DOI: 10.1051/anphys/192510030022.

[522] A. Einstein. “Quantentheorie des einatomigen idealen Gases. Zweite Abhandlung”. In: *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-mathematische Klasse*. Academy proceedings: no volume number exists for the series; @inproceedings with booktitle, not @article. This is the paper containing the condensation prediction. 1925, pp. 3–14.

[523] W. Elsasser. “Bemerkungen zur Quantenmechanik freier Elektronen”. In: *Die Naturwissenschaften* 13.33 (1925). One-page note predicting electron diffraction by crystals two years before Davisson:1927 and Thomson:1927., pp. 711–711. DOI: 10.1007/bf01558853.

[524] F. Hund. “Zur Deutung verwickelter Spektren, insbesondere der Elemente Scandium bis Nickel”. In: *Zeitschrift für Physik* 33.1 (1925), pp. 345–371. DOI: 10.1007/BF01328319.

[525] E. Ising. “Beitrag zur Theorie des Ferromagnetismus”. In: *Zeitschrift für Physik* 31.1 (1925), pp. 253–258. DOI: 10.1007/BF02980577.

[526] H. Jeffreys. “On Certain Approximate Solutions of Lineae Differential Equations of the Second Order*”. In: *Proceedings of the London Mathematical Society* s2-23.1 (1925). The ’earlier purely mathematical treatment’ that puts the J in JWKB. The paper was communicated in 1923, so it is occasionally cited as 1923 or 1924; the issue is dated 1925 and that is what Crossref carries. TRAP for the DOI title check: Crossref’s metadata contains a typo, ’Lineae Differential Equations’ for ’Linear’ – the title above is the published one, so expect a one-word diff against the fetched record. Volume must be written ’s2-23’ (series 2), not ’23’., pp. 428–436. DOI: 10.1112/plms/s2-23.1.428.

[527] W. Kuhn. “Über die Gesamtstärke der von einem Zustande ausgehenden Absorptionslinien”. In: *Zeitschrift für Physik* 33.1 (1925). Issue 1. The Kuhn half of the Thomas–Reiche–Kuhn sum rule, as cited. Byline is ’W. Kuhn’ (Werner Kuhn – not Thomas Kuhn). Verified against Crossref; title begins with U-umlaut., pp. 408–412. DOI: 10.1007/BF01328322.

[528] O. Laporte and W. F. Meggers. “Some Rules of Spectral Structure*”. In: *Journal of the Optical Society of America* 11.5 (1925), p. 459. DOI: 10.1364/josa.11.000459.

[529] W. Pauli. “Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren”. In: *Zeitschrift für Physik* 31.1 (1925), pp. 765–783. DOI: 10.1007/BF02980631.

[530] H. N. Russell and F. A. Saunders. “New Regularities in the Spectra of the Alkaline Earths”. In: *The Astrophysical Journal* 61 (1925), p. 38. DOI: 10.1086/142872.

[531] M. Siegbahn. *The Spectroscopy of X-Rays*. English edition, translated with the author’s additions by George A. Lindsay, of "Spektroskopie der Röntgenstrahlen" (Springer, Berlin, 1924). xii+287 pp. No DOI; a printed monograph grades AMBER under classify_access(), which auto-excuses @book. Year and publisher confirmed from dealer and library records; page count quoted at second hand and may be dropped. London: Oxford University Press (Humphrey Milford), 1925.

[532] G. E. Uhlenbeck and S. Goudsmit. “Ersetzung der Hypothese vom unmechanischen Zwang durch eine Forderung bezüglich des inneren Verhaltens jedes einzelnen Elektrons”. In: *Die Naturwissenschaften* 13 (1925), pp. 953–954.

[533] M. Born. “Quantenmechanik der Stoßvorgänge”. In: *Zeitschrift für Physik* 38.11-12 (1926). The extended treatment; this is the paper that develops what is now called the Born approximation., pp. 803–827. DOI: 10.1007/BF01397184.

[534] M. Born. “Zur Quantenmechanik der Stoßvorgänge”. In: *Zeitschrift für Physik* 37.12 (1926). The probability interpretation appears here (famously in an added footnote), together with the first-order collision amplitude. Suffix a/b needed: Born published two collision papers in 1926., pp. 863–867. DOI: 10.1007/BF01397477.

[535] L. Brillouin. “La mécanique ondulatoire de Schrödinger; une méthode générale de résolution par approximations successives”. In: *Comptes rendus hebdomadaires des séances de l’Académie des sciences*. Vol. 183. Academy proceedings – @inproceedings with booktitle, not @article. Unlike the Sitzungsberichte it DOES carry a volume (tome 183). No DOI (the Comptes rendus of this period are on Gallica, not Crossref). This is the ’B’ of WKB, cited alongside Wentzel:1926 and Kramers:1926. The original punctuation is a semicolon after ’Schrödinger’; some catalogues print a colon. 1926, pp. 24–26.

[536] E. Condon. “A Theory of Intensity Distribution in Band Systems”. In: *Physical Review* 28.6 (1926), pp. 1182–1201. DOI: 10.1103/PhysRev.28.1182.

[537] P. A. M. Dirac. “On the theory of quantum mechanics”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 112.762 (1926). Source both for Fermi-Dirac statistics and for the exchange-energy spin coupling; cited for both in ch. IX.7., pp. 661–677. DOI: 10.1098/rspa.1926.0133.

[538] A. S. Eddington. *The Internal Constitution of the Stars*. Cambridge University Press, 1926.

[539] E. Fermi. “Zur Quantelung des idealen einatomigen Gases”. In: *Zeitschrift für Physik* 36.11-12 (1926), pp. 902–912. DOI: 10.1007/BF01400221.

[540] V. Fock. “Über die invariante Form der Wellen- und der Bewegungsgleichungen für einen geladenen Massenpunkt”. In: *Zeitschrift für Physik* 39.2-3 (1926). Verified against Crossref, umlauts confirmed as printed. This is the minimal-coupling / gauge-principle paper cited for p_mu → p_mu + i q A_mu, not Fock’s later self-consistent-field work., pp. 226–232. DOI: 10.1007/BF01321989.

[541] R. H. Fowler. “On Dense Matter”. In: *Monthly Notices of the Royal Astronomical Society* 87.2 (1926). This is Ralph H. Fowler, not W. A. Fowler of B2FH, pp. 114–122. DOI: 10.1093/mnras/87.2.114.

[542] J. Franck and E. G. Dymond. “Elementary processes of photochemical reactions”. In: *Transactions of the Faraday Society* 21.February (1926). The classical half of the Franck–Condon principle. Published in the February 1926 issue, earlier in the year than the Nobel lecture, hence the a suffix. The Crossref record lists E. G. Dymond as a second author; he is the translator of the English text, and the printed byline is Franck alone., p. 536. DOI: 10.1039/TF9262100536.

[543] J. Franck. *Transformations of kinetic energy of free electrons into excitation energy of atoms by impacts*. Nobel Lecture delivered in Stockholm on 11 December 1926 for the 1925 Nobel Prize in Physics, shared with Gustav Hertz; the year in the key is the year of delivery. Collected in Nobel Lectures, Physics 1922–1941 (Elsevier, Amsterdam, 1965) and published open access by the Nobel Foundation at nobelprize.org. Contains Franck’s own account of how he came to accept Bohr’s reinterpretation of the Franck–Hertz result. No DOI exists. 1926.

[544] W. Gordon. “Der Comptoneffekt nach der Schrödingerschen Theorie”. In: *Zeitschrift für Physik* 40.1-2 (1926). NOT a genuine collision: the two claimed records are the same paper with the same DOI, one of them with the umlauts stripped (Schrodingerschen, Zeitschrift fur Physik) — so the bare key is kept for all three citing chapters. Crossref record verified: volume 40, issue 1-2, pages 117–133, 1926., pp. 117–133. DOI: 10.1007/BF01390840.

[545] W. Heisenberg. “Mehrkörperproblem und Resonanz in der Quantenmechanik”. In: *Zeitschrift für Physik* 38.6-7 (1926), pp. 411–426. DOI: 10.1007/BF01397160.

[546] D. Hilbert. “Über das Unendliche”. In: *Mathematische Annalen* 95.1 (1926). Text of the Münster lecture of June 1925 setting out Hilbert’s programme, pp. 161–190. DOI: 10.1007/BF01206605.

[547] O. Klein. “Quantentheorie und fünfdimensionale Relativitätstheorie”. In: *Zeitschrift für Physik* 37.12 (1926). Cited only for the wave equation it contains; its five-dimensional unification framing is outside the treatise’s evidence base and is flagged as such in the chapter., pp. 895–906. DOI: 10.1007/BF01397481.

[548] A. N. Kolmogorov. “Une série de Fourier–Lebesgue divergente partout”. In: *Comptes Rendus de l’Académie des Sciences (Paris)*. Vol. 183. Byline ’A. Kolmogoroff’. 1926, pp. 1327–1328.

[549] H. A. Kramers. “Wellenmechanik und halbzahlige Quantisierung”. In: *Zeitschrift für Physik* 39.10-11 (1926). Issue 10–11. The K of WKB, cited alongside Wentzel:1926 and Brillouin:1926. Verified against Crossref (byline ’H. A. Kramers’)., pp. 828–840. DOI: 10.1007/BF01451751.

[550] R. de L. Kronig. “On the Theory of Dispersion of X-Rays”. In: *Journal of the Optical Society of America* 12.6 (1926), p. 547. DOI: 10.1364/JOSA.12.000547.

[551] G. N. LEWIS. “The Conservation of Photons”. In: *Nature* 118.2981 (1926). The paper that coined the word ’photon’, for an entity quite unlike the one it now names., pp. 874–875. DOI: 10.1038/118874a0.

[552] R. A. Millikan and G. H. Cameron. “High Frequency Rays of Cosmic Origin III. Measurements in Snow-Fed Lakes at High Altitudes”. In: *Physical Review* 28.5 (1926), pp. 851–868. DOI: 10.1103/PhysRev.28.851.

[553] W. Pauli. “Über das Wasserstoffspektrum vom Standpunkt der neuen Quantenmechanik”. In: *Zeitschrift für Physik A Hadrons and nuclei* 36.5 (1926). Issue 5, May 1926 (submitted 17 January 1926, i.e. before Schrödinger’s first communication – which is the point the citing sentence makes). Springer’s Crossref record names the container by the journal’s later title, ’Zeitschrift für Physik A Hadrons and nuclei’; the 1926 journal was simply ’Zeitschrift für Physik’., pp. 336–363. DOI: 10.1007/BF01450175.

[554] J. Perrin. *Discontinuous Structure of Matter*. Nobel Prize in Physics 1926. 1926.

[555] B. van der Pol. “LXXXVIII.On “relaxation-oscillations””. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1926). The van der Pol oscillator and relaxation oscillations., pp. 978–992. DOI: 10.1080/14786442608564127.

[556] E. Schrödinger. “Der stetige Übergang von der Mikro- zur Makromechanik”. In: *Die Naturwissenschaften* 14 (1926). Crossref-verified. NOT one of the four Annalen communications: this is the short Naturwissenschaften note that constructs the coherent state of the oscillator, which is exactly the ’where Schrödinger himself left them’ of sec:qsimple. Issue 28, 664-666., pp. 664–666.

[557] E. Schrödinger. “Quantisierung als Eigenwertproblem”. In: *Annalen der Physik* 384.4 (1926). Crossref-verified. The ’first communication’ the hydrogen chapter follows. Crossref deposits the bare title ’Quantisierung als Eigenwertproblem’ for all four communications – they are distinguished only by volume and pages – so the ’(Erste Mitteilung)’ parenthetical here is the published designation, not a Crossref string. Wiley’s continuous volume 384 corresponds to the fourth-series Band 79 that contemporary citations use., pp. 361–376. DOI: 10.1002/andp.19263840404.

[558] E. Schrödinger. “Quantisierung als Eigenwertproblem”. In: *Annalen der Physik* 384.6 (1926). Crossref-verified. Second communication – the one containing the harmonic-oscillator/Hermite-function solution cited at sec:qsimple-oscillator-analytic. Same Wiley volume 384 (= Band 79) as the first communication, different issue. Crossref deposits the bare title without ’(Zweite Mitteilung)’., pp. 489–527. DOI: 10.1002/andp.19263840602.

[559] E. Schrödinger. “Quantisierung als Eigenwertproblem”. In: *Annalen der Physik* 386.18 (1926). Crossref-verified. Fourth communication – time-dependent perturbation theory, which is what sec on time-dependent methods cites it for alongside Dirac:1927. Wiley volume 386 = Band 81. Crossref deposits only ’Quantisierung als Eigenwertproblem’., pp. 109–139. DOI: 10.1002/andp.19263861802.

[560] E. Schrödinger. “Quantisierung als Eigenwertproblem (Dritte Mitteilung: Störungstheorie, mit Anwendung auf den Starkeffekt der Balmerlinien)”. In: *Annalen der Physik* 385 (1926). Crossref-verified. Third communication; the subtitle names the Stark effect of the Balmer lines, which is precisely why the hydrogen and approximation-methods chapters cite this key next to Stark:1914. Wiley volume 385 = Band 80. Crossref deposits only ’Quantisierung als Eigenwertproblem’., pp. 437–490.

[561] L. H. THOMAS. “The Motion of the Spinning Electron”. In: *Nature* 117.2945 (1926), pp. 514–514. DOI: 10.1038/117514a0.

[562] G. E. UHLENBECK and S. GOUDSMIT. “Spinning Electrons and the Structure of Spectra”. In: *Nature* 117.2938 (1926). Verified via Crossref (Nature 117, Feb 1926; Crossref prints the names in caps, UHLENBECK/GOUDSMIT). This is the English Nature note; the companion key Uhlenbeck:1925 is the earlier German announcement in Die Naturwissenschaften 13, 953-954 (1925). The chapter cites both together, so the two entries must stay distinct., pp. 264–265. DOI: 10.1038/117264a0.

[563] G. Wentzel. “Eine Verallgemeinerung der Quantenbedingungen für die Zwecke der Wellenmechanik”. In: *Zeitschrift für Physik* 38.6-7 (1926). Verified via Crossref (Z. Phys. 38, June 1926). This is the W of WKB; the sibling keys Kramers:1926 and Brillouin:1926 cited alongside it are Z. Phys. 39, 828 (1926) and C. R. Acad. Sci. Paris 183, 24 (1926) respectively — the latter is Comptes rendus and therefore @inproceedings with no volume, if it also turns out to be an orphan., pp. 518–529. DOI: 10.1007/BF01397171.

[564] M. Born and R. Oppenheimer. “Zur Quantentheorie der Molekeln”. In: *Annalen der Physik* 389.20 (1927), pp. 457–484. DOI: 10.1002/andp.19273892002.

[565] P. W. Bridgman. *The Logic of Modern Physics*. The operational analysis of physical concepts. Bridgman’s identification of a concept with one set of operations was abandoned; the demand that a quantity possess a measurement procedure was not. New York: Macmillan, 1927.

[566] L. Brillouin. “Les moments de rotation et le magnétisme dans la mécanique ondulatoire”. In: *Journal de Physique et le Radium* 8.2 (1927), pp. 74–84. DOI: 10.1051/jphysrad:019270080207400.

[567] L. de Broglie. “La mécanique ondulatoire et la structure atomique de la matière et du rayonnement”. In: *Journal de Physique et le Radium* 8.5 (1927), pp. 225–241. DOI: 10.1051/jphysrad:0192700805022500.

[568] Ø. Burrau. “Berechnung des Energiewertes des Wasserstoffmolekel-Ions (H 2 +) im Normalzustand”. In: *Die Naturwissenschaften* 15.1 (1927), pp. 16–17. DOI: 10.1007/bf01504875.

[569] J. Clay. “Penetrating Radiation”. In: *Proceedings of the Royal Academy of Sciences at Amsterdam*. Vol. 30. 1927, pp. 1115–1127.

[570] C. G. Darwin. “Free motion in the wave mechanics”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 117.776 (1927). Crossref confirmed: vol. 117, pp. 258-293, 1 December 1927. This is C. G. Darwin (Charles Galton Darwin, the physicist), and it is the paper containing the first computation of the free spreading of a Gaussian wave packet that the chapter attributes to him., pp. 258–293. DOI: 10.1098/rspa.1927.0179.

[571] C. Davisson and L. H. Germer. “Diffraction of Electrons by a Crystal of Nickel”. In: *Physical Review* 30.6 (1927), pp. 705–740. DOI: 10.1103/PhysRev.30.705.

[572] P. A. M. Dirac. “The quantum theory of the emission and absorption of radiation”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 114.767 (1927), pp. 243–265. DOI: 10.1098/rspa.1927.0039.

[573] P. Ehrenfest. “Bemerkung über die angenäherte Gültigkeit der klassischen Mechanik innerhalb der Quantenmechanik”. In: *Zeitschrift für Physik* 45.7-8 (1927). Crossref confirmed: Z. Phys. vol. 45, issue 7-8, pp. 455-457, July 1927. Three pages – Ehrenfest's theorem in its original short note., pp. 455–457. DOI: 10.1007/BF01329203.

[574] C. D. Ellis and W. A. Wooster. “The average energy of disintegration of radium E”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 117.776 (1927), pp. 109–123. DOI: 10.1098/rspa.1927.0168.

[575] H. Faxén and J. Holtsmark. “Beitrag zur Theorie des Durchganges langsamer Elektronen durch Gase”. In: *Zeitschrift für Physik* 45 (1927), pp. 307–324.

[576] W. Heisenberg. “Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik”. In: *Zeitschrift für Physik* 43.3-4 (1927). Not a genuine collision: the two claimed works are the same paper, one record having ASCII-mangled umlauts (Ueber ... Zeitschrift fuer Physik) and the same DOI. No split; all three citing chapters keep the bare key. Metadata confirmed against Crossref., pp. 172–198. DOI: 10.1007/BF01397280.

[577] W. Heitler and F. London. “Wechselwirkung neutraler Atome und homöopolare Bindung nach der Quantenmechanik”. In: *Zeitschrift für Physik* 44.6-7 (1927), pp. 455–472. DOI: 10.1007/BF01397394.

[578] E. Hopf. “Elementare Bemerkungen über die Lösungen partieller Differentialgleichungen zweiter Ordnung vom elliptischen Typus”. In: *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-Mathematische Klasse*. Academy proceedings without volume number – @inproceedings per house convention. The strong maximum principle. 1927, pp. 147–152.

[579] P. Jordan and O. Klein. “Zum Mehrkörperproblem der Quantentheorie”. In: *Zeitschrift für Physik* 45.11-12 (1927), pp. 751–765. DOI: 10.1007/BF01329553.

[580] E. H. Kennard. “Zur Quantenmechanik einfacher Bewegungstypen”. In: *Zeitschrift für Physik* 44.4-5 (1927). First proof of the inequality sigma_x sigma_p ≥ hbar/2; Heisenberg:1927 argues it physically., pp. 326–352. DOI: 10.1007/BF01391200.

[581] H. A. Kramers. “La diffusion de la lumière par les atomes”. In: *Atti del Congresso Internazionale dei Fisici, Como*. Volume 2 of the Como congress proceedings; a conference volume with no journal volume number. 1927, pp. 545–557.

[582] L. Landau. “Das Dämpfungsproblem in der Wellenmechanik”. In: *Zeitschrift für Physik* 45.5-6 (1927), pp. 430–441. DOI: 10.1007/BF01343064.

[583] G. Lemaître. “Un Univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extra-galactiques”. In: *Annales de la Société Scientifique de Bruxelles A* 47 (1927). Derives the velocity–distance law and the first expansion-rate value, two years before Hubble; the 1931 MNRAS translation omitted those paragraphs. The chapter records the priority. No DOI for the original., pp. 49–59.

[584] A. A. Michelson. “Measurement of the Velocity of Light Between Mount Wilson and Mount San Antonio”. In: *The Astrophysical Journal* 65 (1927). January 1927 issue (vol. LXV, no. 1); Mount Wilson Observatory Contribution No. 329. The paper states it is a continuation of Michelson, ApJ 60, 256 (1924), and credits F. Pearson as assistant, so the byline is Michelson alone. Page range: the Smithsonian catalogue and the scanned issue give 1–13 (running head ’12 A. A. MICHELSON’ near the end); some secondary lists say 1–22 – do not copy that without checking the scan., p. 1. DOI: 10.1086/143021.

[585] J. von Neumann. “Wahrscheinlichkeitstheoretischer Aufbau der Quantenmechanik”. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*. Vol. 1927. Second of the three 1927 Göttingen papers; introduces the statistical operator independently of Landau. 1927, pp. 245–272.

[586] W. Pauli. “Über Gasentartung und Paramagnetismus”. In: *Zeitschrift für Physik A Hadrons and nuclei* 41.2-3 (1927). Pauli paramagnetism of a degenerate electron gas. Signed W. Pauli jr. on the paper., pp. 81–102. DOI: 10.1007/BF01391920.

[587] W. Pauli. “Zur Quantenmechanik des magnetischen Elektrons”. In: *Zeitschrift für Physik* 43.9-10 (1927). The two-component spin theory and the Pauli matrices. Signed W. Pauli jr. on the paper., pp. 601–623. DOI: 10.1007/BF01397326.

[588] L. H. Thomas. “The calculation of atomic fields”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 23.5 (1927), pp. 542–548. DOI: 10.1017/S0305004100011683.

[589] G. P. THOMSON and A. REID. “Diffraction of Cathode Rays by a Thin Film”. In: *Nature* 119.3007 (1927). G. P. Thomson, son of the J. J. Thomson of Thomson:1897; no key collision, the years differ., pp. 890–890. DOI: 10.1038/119890a0.

[590] B. VAN DER POL and J. VAN DER MARK. “Frequency Demultiplication”. In: *Nature* 120.3019 (1927). The driven neon-tube circuit whose ’irregular noise’ is retrospectively read as the first laboratory encounter with deterministic chaos., pp. 363–364. DOI: 10.1038/120363a0.

[591] K. Weierstrass. *Vorlesungen über Variationsrechnung (Mathematische Werke, Band 7)*. Berlin lectures of the 1870s–80s published posthumously; primary source for the excess function, strong variations and the field-theoretic sufficiency proof. Leipzig: Akademische Verlagsgesellschaft, 1927.

[592] E. Wigner. “Über die Erhaltungssätze in der Quantenmechanik”. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*. Vol. 1927. No DOI; pre-1990 academy proceedings volume, cited for the introduction of parity as a conserved quantum number. 1927, pp. 375–381.

[593] N. BOHR. “The Quantum Postulate and the Recent Development of Atomic Theory1”. In: *Nature* 121.3050 (1928), pp. 580–590. DOI: 10.1038/121580a0.

[594] M. Born and V. Fock. “Beweis des Adiabatensatzes”. In: *Zeitschrift für Physik* 51.3-4 (1928). Issue 3-4, March 1928. Crossref confirmed: title, volume 51, pp. 165-180, both authors. This is the paper the adiabatic theorem is named for; keyed by first listed author Born per the collaboration rule., pp. 165–180. DOI: 10.1007/BF01343193.

[595] C. G. Darwin. “The wave equations of the electron”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 118.780 (1928), pp. 654–680. DOI: 10.1098/rspa.1928.0076.

[596] P. A. M. Dirac. “The quantum theory of the electron”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 117.778 (1928). Key mandated unsuffixed by the task brief; the 1928 collision with Part II is resolved by giving Part II the suffix b (see Dirac:1928b), not by re-keying this entry as 1928a., pp. 610–624. DOI: 10.1098/rspa.1928.0023.

[597] P. A. M. Dirac. “The quantum theory of the Electron. Part II”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 118.779 (1928). Suffix b chosen because the brief fixes the unsuffixed key Dirac:1928 for Part I., pp. 351–361. DOI: 10.1098/rspa.1928.0056.

[598] E. Fermi. “Eine statistische Methode zur Bestimmung einiger Eigenschaften des Atoms und ihre Anwendung auf die Theorie des periodischen Systems der Elemente”. In: *Zeitschrift für Physik* 48.1-2 (1928). The 1927 Rendiconti Lincei note is the earlier announcement; this is the published paper the hint names., pp. 73–79. DOI: 10.1007/BF01351576.

[599] R. H. Fowler and L. Nordheim. “Electron emission in intense electric fields”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 119.781 (1928). Verified against Crossref (author given there as ’Ralph Howard Fowler’). The Fowler–Nordheim cold-emission current–field law. The long series title is what Crossref holds; ’Proceedings of the Royal Society A’ is the usual short form., pp. 173–181. DOI: 10.1098/rspa.1928.0091.

[600] G. Gamow. “Zur Quantentheorie des Atomkernes”. In: *Zeitschrift für Physik* 51.3-4 (1928), pp. 204–212. DOI: 10.1007/BF01343196.

[601] W. Gordon. “Die Energieniveaus des Wasserstoffatoms nach der Diracschen Quantentheorie des Elektrons”. In: *Zeitschrift für Physik* 48.1-2 (1928). Split from the former shared key Gordon:1928. Suffixed a/b because both are Walter Gordon papers in Zeitschrift für Physik volume 48 and neither sense is secondary; a/b follow the page order within the volume. Metadata confirmed against Crossref., pp. 11–14. DOI: 10.1007/BF01351570.

[602] W. Gordon. “Über den Stoß zweier Punktladungen nach der Wellenmechanik”. In: *Zeitschrift für Physik* 48.3-4 (1928). Split from the former shared key Gordon:1928; the exact Coulomb-scattering solution in parabolic coordinates. Metadata confirmed against Crossref., pp. 180–191. DOI: 10.1007/BF01351302.

[603] R. W. GURNEY and E. U. CONDON. “Wave Mechanics and Radioactive Disintegration”. In: *Nature* 122.3073 (1928), pp. 439–439. DOI: 10.1038/122439a0.

[604] D. R. Hartree. “The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.1 (1928), pp. 89–110. DOI: 10.1017/S0305004100011919.

[605] W. Heisenberg. “Zur Theorie des Ferromagnetismus”. In: *Zeitschrift für Physik* 49.9-10 (1928), pp. 619–636. DOI: 10.1007/BF01328601.

[606] D. Hilbert and W. Ackermann. *Grundzüge der theoretischen Logik*. Die Grundlehren der mathematischen Wissenschaften 27. First edition, viii+120 pp.; poses the Entscheidungsproblem. The Springer DOI 10.1007/978-3-662-41928-1 belongs to the second edition of 1938 and is deliberately not given here. Berlin: Julius Springer, 1928.

[607] F. Hund. “Zur Deutung der Molekelspektren. IV”. In: *Zeitschrift für Physik* 51.11-12 (1928), pp. 759–795. DOI: 10.1007/bf01400239.

[608] J. B. Johnson. “Thermal Agitation of Electricity in Conductors”. In: *Physical Review* 32.1 (1928). J. B. Johnson; distinct from Thomas H. Johnson of Johnson:1933, pp. 97–109. DOI: 10.1103/PhysRev.32.97.

[609] P. Jordan and E. Wigner. “Über das Paulische Äquivalenzverbot”. In: *Zeitschrift für Physik* 47.9-10 (1928). Not a genuine collision: both claimed works are the Jordan–Wigner anticommutator paper, one record with ASCII-mangled umlauts and abbreviated given names, same DOI. No split. Metadata confirmed against Crossref., pp. 631–651. DOI: 10.1007/BF01331938.

[610] G. Landsberg and L. Mandelstam. “Eine neue Erscheinung bei der Lichtzerstreuung in Krystallen”. In: *Die Naturwissenschaften* 16.28 (1928), pp. 557–558. DOI: 10.1007/BF01506807.

[611] I. Langmuir. “Oscillations in Ionized Gases”. In: *Proceedings of the National Academy of Sciences* 14.8 (1928). Issue 8. The paper in which Langmuir names the plasma and studies its collective oscillation, matching the citing sentence exactly; cited there together with Tonks:1929. Verified against Crossref. PNAS is open access for this vintage, so this should grade GREEN with a URL in the manifest., pp. 627–637. DOI: 10.1073/pnas.14.8.627.

[612] R. A. Millikan and C. C. Lauritsen. “Relations of Field-Currents to Thermionic-Currents”. In: *Proceedings of the National Academy of Sciences* 14.1 (1928). Issue 1, 15 January 1928. Verified against Crossref and against the content claim: this is the paper that plots log i against 1/F and finds it straight – the plot is named the ’Millikan-Lauritsen plot’ after it, and it is the cold-emission data Fowler & Nordheim (Fowler:1928) analysed months later. Do not confuse it with Millikan & Eyring, Phys. Rev. 27, 51 (1926), which plotted log i against the square root of the field and got a curve. Four other Millikan keys already exist (1911, 1913, 1916, 1926); none collides., pp. 45–49. DOI: 10.1073/pnas.14.1.45.

[613] R. S. Mulliken. “The Assignment of Quantum Numbers for Electrons in Molecules. I”. In: *Physical Review* 32.2 (1928), pp. 186–222. DOI: 10.1103/PhysRev.32.186.

[614] H. Nyquist. “Certain Topics in Telegraph Transmission Theory”. In: *Transactions of the American Institute of Electrical Engineers* 47.2 (1928). The sampling-rate paper. Split from the former shared key Nyquist:1928; suffixed a/b because the telegraph-transmission and thermal-noise papers are both standard references and neither is secondary, with a/b in publication-date order within 1928. Metadata confirmed against Crossref., pp. 617–644. DOI: 10.1109/T-AIEE.1928.5055024.

[615] H. Nyquist. “Thermal Agitation of Electric Charge in Conductors”. In: *Physical Review* 32.1 (1928). The Johnson–Nyquist noise derivation, published in the same Physical Review issue as Johnson’s measurement (Johnson:1928) — which is what the ’in the same issue’ phrasing in 12-statistical-mechanics/07-nonequilibrium-transport.tex refers to. Split from the former shared key Nyquist:1928. Metadata confirmed against Crossref., pp. 110–113. DOI: 10.1103/PhysRev.32.110.

[616] J. R. Oppenheimer. “Three Notes on the Quantum Theory of Aperiodic Effects”. In: *Physical Review* 31.1 (1928). Issue 1, 1 January 1928. The first of the three notes treats field ionization of hydrogen in a uniform electric field – the result the citing sentence attributes to Oppenheimer (no strictly bound state; the Stark levels are long-lived resonances)., pp. 66–81. DOI: 10.1103/PhysRev.31.66.

[617] W. Pauli. “Über das H-Theorem vom Anwachsen der Entropie vom Standpunkt der neuen Quantenmechanik”. In: *Probleme der Modernen Physik: Arnold Sommerfeld zum 60. Geburtstage*. 1928, pp. 30–45.

[618] C. V. RAMAN and K. S. KRISHNAN. “A New Type of Secondary Radiation”. In: *Nature* 121.3048 (1928), pp. 501–502. DOI: 10.1038/121501c0.

[619] H. Reichenbach. *Philosophie der Raum-Zeit-Lehre*. Source of the coordinative definition and of the conventionality of distant simultaneity. English translation: The Philosophy of Space and Time, trans. M. Reichenbach and J. Freund, Dover, New York (1958). Berlin: Walter de Gruyter, 1928.

[620] A. Sommerfeld. “Zur Elektronentheorie der Metalle auf Grund der Fermischen Statistik”. In: *Zeitschrift für Physik* 47.1-2 (1928), pp. 1–32. DOI: 10.1007/BF01391052.

[621] G. P. Thomson. “Experiments on the diffraction of cathode rays”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 117.778 (1928). Issue 778. SINGLE AUTHOR – G. P. Thomson alone. Reid co-signed only the 1927 Nature letter (the existing Thomson:1927), so an author field of ’Thomson and Reid’ would be wrong even though the chapter’s prose says ’the fuller papers’. Received 4 November 1927 but the issue is dated 1 February 1928, so the ISSUE year 1928 keys the entry. Content verified against the article’s own pages, not just metadata: section 5 reads ’Specimens of the results obtained with films of aluminium, gold, celluloid, and a substance X (at first thought to be platinum)’, which is exactly the claim cited – the metal films that made the result independent of target chemistry – and its section 1 footnote back-references ’Thomson and Reid, Nature, vol. 119, p. 890 (1927)’. Kept as @article, not @inproceedings: unlike Sitzungsberichte or Comptes rendus, Proc. Roy. Soc. A carries a real volume and issue number, and the house bib already files Bragg:1913b, Bethe:1934 and Blackett:1933 in this same container as @article. Container title matches the string Crossref and the existing Bragg:1913b use. This is G. P. Thomson, distinct from the J. J. Thomson of Thomson:1897 and the William Thomson entries; no key collision, the nine existing Thomson:* keys are all other years. Verified twice independently: Crossref, and the JSTOR title page of the scanned article (’Vol. 117, No. 778 (Feb. 1, 1928), pp. 600-609’)., pp. 600–609. DOI: 10.1098/rspa.1928.0022.

[622] R. d. E. Atkinson and F. G. Houtermans. “Zur Frage der Aufbaumöglichkeit der Elemente in Sternen”. In: *Zeitschrift für Physik* 54.9-10 (1929), pp. 656–665. DOI: 10.1007/BF01341595.

[623] F. Bloch. “Über die Quantenmechanik der Elektronen in Kristallgittern”. In: *Zeitschrift für Physik* 52.7-8 (1929), pp. 555–600. DOI: 10.1007/BF01339455.

[624] R. Carnap, H. Hahn, and O. Neurath. *Wissenschaftliche Weltauffassung: Der Wiener Kreis*. The Vienna Circle manifesto, published unsigned by the Verein Ernst Mach; the Geleitwort is signed by Hahn, Neurath and Carnap, and the drafting was chiefly Neurath's. Vienna: Artur Wolf Verlag, 1929.

[625] R. W. Gurney and E. U. Condon. "Quantum Mechanics and Radioactive Disintegration". In: *Physical Review* 33.2 (1929). Issue 2. The full paper, distinct from the companion note Gurney:1928 ('Wave Mechanics and Radioactive Disintegration', Nature 122, 439, 1928) that the same sentence cites alongside it. Verified against Crossref., pp. 127–140. DOI: 10.1103/PhysRev.33.127.

[626] W. Heisenberg and W. Pauli. "Zur Quantendynamik der Wellenfelder". In: *Zeitschrift fur Physik* 56.1-2 (1929), pp. 1–61. DOI: 10.1007/BF01340129.

[627] E. Hubble. "A relation between distance and radial velocity among extra-galactic nebulae". In: *Proceedings of the National Academy of Sciences* 15.3 (1929), pp. 168–173. DOI: 10.1073/pnas.15.3.168.

[628] E. A. Hylleraas. "Neue Berechnung der Energie des Heliums im Grundzustande, sowie des tiefsten Terms von Ortho-Helium". In: *Zeitschrift für Physik* 54.5-6 (1929), pp. 347–366. DOI: 10.1007/BF01375457.

[629] O. Klein. "Die Reflexion von Elektronen an einem Potentialsprung nach der relativistischen Dynamik von Dirac". In: *Zeitschrift für Physik* 53.3-4 (1929). The Klein paradox. Split from the former shared key Klein:1929. Metadata confirmed against Crossref., pp. 157–165. DOI: 10.1007/BF01339716.

[630] O. Klein and Y. Nishina. "Über die Streuung von Strahlung durch freie Elektronen nach der neuen relativistischen Quantendynamik von Dirac". In: *Zeitschrift für Physik* 52.11-12 (1929). The Klein–Nishina cross-section, keyed by first listed author. Split from the former shared key Klein:1929; suffixed a/b because neither the Klein–Nishina paper nor the Klein-paradox paper is secondary, with a/b in volume order (52 before 53). Two of the three input candidates were the same paper (one ASCII-mangled). Metadata confirmed against Crossref., pp. 853–868. DOI: 10.1007/BF01366453.

[631] J. E. Lennard-Jones. "The electronic structure of some diatomic molecules". In: *Transactions of the Faraday Society* 25 (1929). Hyphen dropped in the key (Lennard-Jones) so the key stays a single word., p. 668. DOI: 10.1039/TF9292500668.

[632] P. M. Morse. "Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels". In: *Physical Review* 34.1 (1929), pp. 57–64. DOI: 10.1103/PhysRev.34.57.

[633] N. F. Mott. "The scattering of fast electrons by atomic nuclei". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 124.794 (1929). This is the relativistic spin-1/2 Coulomb cross section, NOT the identical-particle result. The task hint said 'Mott 1929 identical-particle scattering'; the identical-particle calculation is Mott:1930 below, and ch. IX.10 states the customary misattribution explicitly rather than silently using the right key., pp. 425–442. DOI: 10.1098/rspa.1929.0127.

[634] R. Peierls. "Zur kinetischen Theorie der Wärmeleitung in Kristallen". In: *Annalen der Physik* 395.8 (1929). Umklapp processes and the finite thermal conductivity of an insulating crystal. The DOI in the batch candidate (10.1002/andp.19293951006) returns 404; the resolving DOI is issue 8, article 3, as given here., pp. 1055–1101. DOI: 10.1002/andp.19293950803.

[635] R. Peierls. “Zur Theorie der galvanomagnetischen Effekte”. In: *Zeitschrift für Physik* 53.3-4 (1929). Introduces the effective-mass tensor and the hole concept. DOI verified against Crossref (title, volume and pages all match); the batch candidate carried no DOI., pp. 255–266. DOI: 10.1007/BF01339727.

[636] H. P. Robertson. “The Uncertainty Principle”. In: *Physical Review* 34.1 (1929), pp. 163–164. DOI: 10.1103/PhysRev.34.163.

[637] J. C. Slater. “The Theory of Complex Spectra”. In: *Physical Review* 34.10 (1929), pp. 1293–1322. DOI: 10.1103/PhysRev.34.1293.

[638] L. Szilard. “Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen”. In: *Zeitschrift für Physik* 53.11-12 (1929), pp. 840–856. DOI: 10.1007/BF01341281.

[639] L. Tonks and I. Langmuir. “Oscillations in Ionized Gases”. In: *Physical Review* 33.2 (1929). Verified via Crossref (Phys. Rev. 33(2), 1 Feb 1929). Tonks is first author, so the key is correctly Tonks:1929 even though the prose says "Langmuir ... analysed with Tonks". Do not confuse with the short follow-up Tonks & Langmuir, Phys. Rev. 33, 990 (1929)., pp. 195–210. DOI: 10.1103/PhysRev.33.195.

[640] B. L. van der Waerden. “Spinoranalyse”. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*. Founding paper of two-component (dotted/undotted) spinor calculus. 1929, pp. 100–109.

[641] H. Weyl. “Elektron und Gravitation. I”. In: *Zeitschrift für Physik* 56.5-6 (1929). Likely also introduced by the Weyl-equation chapter of the same part., pp. 330–352. DOI: 10.1007/BF01339504.

[642] F. Bloch. “Zum elektrischen Widerstandsgesetz bei tiefen Temperaturen”. In: *Zeitschrift für Physik* 59.3-4 (1930), pp. 208–214. DOI: 10.1007/bf01341426.

[643] F. Bloch. “Zur Theorie des Ferromagnetismus”. In: *Zeitschrift für Physik* 61.3-4 (1930), pp. 206–219. DOI: 10.1007/bf01339661.

[644] L. Brillouin. “Les électrons libres dans les métaux et le role des réflexions de Bragg”. In: *Journal de Physique et le Radium* 1.11 (1930), pp. 377–400. DOI: 10.1051/jphysrad:01930001011037700.

[645] J. Chadwick. “The scattering of α-particles in helium”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 128.807 (1930). Experimental confirmation of Mott’s identical-particle interference term. DOI not supplied – not confident enough., pp. 114–122. DOI: 10.1098/rspa.1930.0098.

[646] P. A. M. Dirac. *The Principles of Quantum Mechanics*. First edition (1930). Source for the delta function and for the algebraic ladder-operator method; the bra–ket notation itself is later, so do not cite this edition for it – cite Dirac:1939, the paper that introduces it. Clarendon Press, 1930.

[647] P. A. M. Dirac. “A theory of electrons and protons”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 126.801 (1930). Issue of 1 January 1930; takes ’a’ as the earlier of the two Dirac 1930 works. Hole theory with the hole wrongly identified as the proton — the citing chapters say so explicitly and cite Dirac:1931 for the anti-electron proper., pp. 360–365. DOI: 10.1098/rspa.1930.0013.

[648] C. Eckart. “The Application of Group theory to the Quantum Dynamics of Monatomic Systems”. In: *Reviews of Modern Physics* 2.3 (1930). Crossref confirmed: RMP vol. 2, pp. 305-380, 1 July 1930. Note the lower-case ’theory’ in the published title – that is how APS prints it, not a typo. This is the Eckart half of the Wigner-Eckart theorem, cited beside Wigner:1931., pp. 305–380. DOI: 10.1103/RevModPhys.2.305.

[649] A. S. Eddington. “On the Instability of Einstein’s Spherical World”. In: *Monthly Notices of the Royal Astronomical Society* 90 (1930). Metadata confirmed against Crossref., pp. 668–678. DOI: 10.1093/mnras/90.7.668.

[650] R. Eisenschitz and F. London. “Über das Verhältnis der van der Waalsschen Kräfte zu den homöopolaren Bindungskräften”. In: *Zeitschrift für Physik* 60 (1930), pp. 491–527.

[651] I. Estermann and O. Stern. “Beugung von Molekularstrahlen”. In: *Zeitschrift fuer Physik* 61.1-2 (1930). Diffraction of neutral helium atoms and hydrogen molecules from a lithium fluoride cleavage face., pp. 95–125. DOI: 10.1007/BF01340293.

[652] E. Fermi. “Über die magnetischen Momente der Atomkerne”. In: *Zeitschrift für Physik* 60.5-6 (1930), pp. 320–333. DOI: 10.1007/bf01339933.

[653] V. Fock. “Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems”. In: *Zeitschrift für Physik* 61 (1930), pp. 126–148.

[654] K. Gödel. “Die Vollständigkeit der Axiome des logischen Funktionenkalküls”. In: *Monatshefte für Mathematik und Physik* 37 (1930). The completeness theorem for first-order logic, pp. 349–360. DOI: 10.1007/BF01696781.

[655] W. J. de Haas and P. M. van Alphen. “The dependence of the susceptibility of diamagnetic metals upon the field”. In: *Proceedings of the Royal Netherlands Academy of Arts and Sciences*. Vol. 33. Also issued as Communications from the Physical Laboratory, University of Leiden. 1930, pp. 1106–1118.

[656] W. Heisenberg and W. Pauli. “Zur Quantentheorie der Wellenfelder. II”. In: *Zeitschrift fur Physik* 59.3-4 (1930). Sequel to the 1929 paper; DOI not asserted here, pp. 168–190. DOI: 10.1007/bf01341423.

[657] P. Heyl. “A redetermination of the constant of gravitation”. In: *Bureau of Standards Journal of Research* 5.6 (1930). Time-of-swing determination at the US Bureau of Standards., p. 1243. DOI: 10.6028/jres.005.074.

[658] D. Hilbert. “Naturerkennen und Logik”. In: *Die Naturwissenschaften* 18.47-49 (1930). Address delivered at Königsberg on 8 September 1930, ending Wir müssen wissen, wir werden wissen, pp. 959–963. DOI: 10.1007/BF01492194.

[659] H. A. Kramers. “Théorie générale de la rotation paramagnétique dans les cristaux”. In: *Proceedings of the Royal Netherlands Academy of Arts and Sciences*. Vol. 33. 1930, pp. 959–972.

[660] L. Landau. “Diamagnetismus der Metalle”. In: *Zeitschrift für Physik* 64.9-10 (1930), pp. 629–637. DOI: 10.1007/BF01397213.

[661] F. London. “Zur Theorie und Systematik der Molekularkräfte”. In: *Zeitschrift für Physik* 63.3-4 (1930), pp. 245–279. DOI: 10.1007/BF01421741.

[662] L. Meitner and W. Orthmann. “Über eine absolute Bestimmung der Energie der primären beta-Strahlen von Radium E”. In: *Zeitschrift für Physik* 60.3-4 (1930), pp. 143–155. DOI: 10.1007/BF01339819.

[663] N. F. Mott. “The collision between two electrons”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 126.801 (1930). The identical-particle interference result. See the note on Mott:1929., pp. 259–267. DOI: 10.1098/rspa.1930.0006.

[664] J. von Neumann. “Allgemeine Eigenwerttheorie Hermitescher Funktionaloperatoren”. In: *Mathematische Annalen* 102 (1930). Often cited as 1929 (submission); Mathematische Annalen 102 is dated 1930 – keyed by the publisher’s year per house rule. Spectral theorem for unbounded self-adjoint operators, deficiency indices, Cayley transform., pp. 49–131.

[665] W. Pauli. *Offener Brief an die Gruppe der Radioaktiven bei der Gauvereins-Tagung zu Tubingen*. The neutrino proposal was never published as a paper; the letter is the primary source. 1930.

[666] M. Presburger. “Über die Vollständigkeit eines gewissen Systems der Arithmetik ganzer Zahlen, in welchem die Addition als einzige Operation hervortritt”. In: *Sprawozdanie z I Kongresu Matematyków Krajów Słowiańskich, Warszawa 1929*. Ed. by F. Leja. The congress met 23–27 September 1929; the proceedings volume appeared in 1930, which fixes the key. An unnumbered supplement leaf after p. 394 carries an addendum, cited by some authors as p. 395. Warsaw: Książnica-Atlas, 1930, pp. 92–101.

[667] T. Radó. “On Plateau’s problem”. In: *Annals of Mathematics* 31 (1930). Annals of Mathematics, second series., pp. 457–469.

[668] B. Rossi. “On the Magnetic Deflection of Cosmic Rays”. In: *Physical Review* 36.3 (1930), pp. 606–606. DOI: 10.1103/PhysRev.36.606.

[669] E. Schrödinger. “Über die kräftefreie Bewegung in der relativistischen Quantenmechanik”. In: *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-mathematische Klasse*. Academy proceedings: no volume number exists for the series (CLAUDE.md rule 5). The Zitterbewegung paper. 1930, pp. 418–428.

[670] L. W. Shubnikov and W. J. de Haas. “Magnetische Widerstandsvergrösserung in Einkristallen von Wismut bei tiefen Temperaturen”. In: *Proceedings of the Royal Netherlands Academy of Arts and Sciences*. Vol. 33. 1930.

[671] J. C. Slater. “Note on Hartree’s Method”. In: *Physical Review* 35.2 (1930), pp. 210–211. DOI: 10.1103/PhysRev.35.210.2.

[672] G. E. Uhlenbeck and L. S. Ornstein. “On the Theory of the Brownian Motion”. In: *Physical Review* 36.5 (1930), pp. 823–841. DOI: 10.1103/PhysRev.36.823.

[673] N. Wiener. “Generalized harmonic analysis”. In: *Acta Mathematica* 55.0 (1930), pp. 117–258. DOI: 10.1007/BF02546511.

[674] G. D. Birkhoff. “Proof of the Ergodic Theorem”. In: *Proceedings of the National Academy of Sciences* 17.12 (1931), pp. 656–660. DOI: 10.1073/pnas.17.2.656.

[675] F. Bitter. “On Inhomogeneities in the Magnetization of Ferromagnetic Materials”. In: *Physical Review* 38.10 (1931), pp. 1903–1905. DOI: 10.1103/PhysRev.38.1903.

[676] S. Chandrasekhar. “The Maximum Mass of Ideal White Dwarfs”. In: *The Astrophysical Journal* 74 (1931), p. 81. DOI: 10.1086/143324.

[677] P. A. M. Dirac. “Quantised singularities in the electromagnetic field”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 133.821 (1931). This, not Dirac:1930, is the paper containing the anti-electron prediction; cited only for that, never for the magnetic monopole, which has no evidence., pp. 60–72. DOI: 10.1098/rspa.1931.0130.

[678] J. Douglas. “Solution of the problem of Plateau”. In: *Transactions of the American Mathematical Society* 33.1 (1931), pp. 263–321. DOI: 10.1090/S0002-9947-1931-1501590-9.

[679] P. Ehrenfest and J. R. Oppenheimer. “Note on the Statistics of Nuclei”. In: *Physical Review* 37.4 (1931), pp. 333–338. DOI: 10.1103/PhysRev.37.333.

[680] J. Frenkel. “On the Transformation of light into Heat in Solids. I”. In: *Physical Review* 37.1 (1931), pp. 17–44. DOI: 10.1103/PhysRev.37.17.

[681] K. Gödel. “Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I”. In: *Monatshefte für Mathematik und Physik* 38 (1931). The trailing I is part of the title; the announced part II never appeared. English translation in Gödel, Collected Works I (Oxford, 1986), pp. 144–195, pp. 173–198. DOI: 10.1007/BF01700692.

[682] E. Kappler. “Versuche zur Messung der Avogadro-Loschmidtschen Zahl aus der Brownschen Bewegung einer Drehwaage”. In: *Annalen der Physik* 403.2 (1931), pp. 233–256. DOI: 10.1002/andp.19314030208.

[683] R. D. L. Kronig and W. G. Penney. “Quantum mechanics of electrons in crystal lattices”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 130.814 (1931), pp. 499–513. DOI: 10.1098/rspa.1931.0019.

[684] L. Onsager. “Reciprocal Relations in Irreversible Processes. I”. In: *Physical Review* 37.4 (1931), pp. 405–426. DOI: 10.1103/PhysRev.37.405.

[685] L. Onsager. “Reciprocal Relations in Irreversible Processes. II”. In: *Physical Review* 38.12 (1931), pp. 2265–2279. DOI: 10.1103/PhysRev.38.2265.

[686] L. Pauling. “THE NATURE OF THE CHEMICAL BOND. APPLICATION OF RESULTS OBTAINED FROM THE QUANTUM MECHANICS AND FROM A THEORY OF PARAMAGNETIC SUSCEPTIBILITY TO THE STRUCTURE OF MOLECULES”. In: *Journal of the American Chemical Society* 53.4 (1931), pp. 1367–1400. DOI: 10.1021/ja01355a027.

[687] E. Wigner. *Gruppentheorie und ihre Anwendung auf die Quantenmechanik der Atomspektren*. English translation: Group Theory and Its Application to the Quantum Mechanics of Atomic Spectra, Academic Press, New York (1959). Braunschweig: Friedrich Vieweg und Sohn, 1931.

[688] A. H. Wilson. “The theory of electronic semi-conductors”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 133.822 (1931), pp. 458–491. DOI: 10.1098/rspa.1931.0162.

[689] A. H. Wilson. “The theory of electronic semi-conductors. - II”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 134.823 (1931), pp. 277–287. DOI: 10.1098/rspa.1931.0196.

[690] I. F. Zartman. “A Direct Measurement of Molecular Velocities”. In: *Physical Review* 37.4 (1931). Issue 4. Crossref-confirmed. The abstract matches the chapter’s description of the apparatus: bismuth beam, glass plate on a cylinder rotating at 241 r.p.s., molecules from 168 to 673 m/s spread over a 3 cm band, vapour found to be 40% Bi and 60% Bi2 at 851 degC. First of the two rotating-drum velocity analysers named in exp:kt-miller-kusch., pp. 383–391. DOI: 10.1103/PhysRev.37.383.

[691] C. D. Anderson. “The Apparent Existence of Easily Deflectable Positives”. In: *Science* 76.1967 (1932). First announcement of the positron., pp. 238–239. DOI: 10.1126/science.76.1967.238.

[692] J. Chadwick. “Possible Existence of a Neutron”. In: *Nature* 129.3252 (1932). The one-page discovery letter of 27 February 1932; the announcement that dates the discovery., pp. 312–312. DOI: 10.1038/129312a0.

[693] J. Chadwick. “The existence of a neutron”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 136.830 (1932). The full account, with the recoil kinematics from protons and nitrogen nuclei and the neutron mass determination. Crossref gives the author as James Chadwick., pp. 692–708. DOI: 10.1098/rspa.1932.0112.

[694] J. D. COCKCROFT and E. T. S. WALTON. “Disintegration of Lithium by Swift Protons”. In: *Nature* 129.3261 (1932). The announcement letter, issue of 30 April 1932; carries the first observation at about 125 kV and the alpha range of about 8.4 cm of air. Keyed ’a’ because three Cockcroft-Walton 1932 papers are cited in this chapter; a/b/c are in publication order (Nature April, Proc. Roy. Soc. A 136 June, Proc. Roy. Soc. A 137 July). Crossref prints the authors in full caps (COCKCROFT, WALTON), normalised here. Verified against api.crossref.org 2026-08-13., pp. 649–649. DOI: 10.1038/129649a0.

[695] J. D. Cockcroft and E. T. S. Walton. “Experiments with high velocity positive ions. II. -The disintegration of elements by high velocity protons”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 137.831 (1932). Issue 831, 1 July 1932. The full account: yield against accelerating potential, the range measurement, and the coincidence test showing the two alpha particles are emitted back to back – i.e. the quantitative content of the first accelerator-driven nuclear reaction and of the first nuclear test of E=mc^2. TRAP: Crossref stores the title as ’...positive ions. II. -The disintegration...’, where the stray hyphen is the tail of the same U+2015 dash artefact as in Cockcroft:1932b; removed here. Verified against api.crossref.org 2026-08-13., pp. 229–242. DOI: 10.1098/rspa.1932.0133.

[696] J. D. Cockcroft and E. T. S. Walton. “Experiments with high velocity positive ions.—(I) Further developments in the method of obtaining high velocity positive ions”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 136.830 (1932). The apparatus paper: the cascade voltage multiplier (Cockcroft-Walton generator) and the sectioned accelerating tube. TRAP: Crossref stores the title with a U+2015 HORIZONTAL BAR after the first full stop (’...positive ions.<U+2015>(I) Further developments...’), which is a non-ASCII byte fatal under pdflatex + T1; it is normalised to a plain space here. If a title-similarity check is run against Crossref, expect the dash to be the only difference. Verified against api.crossref.org 2026-08-13., pp. 619–630. DOI: 10.1098/rspa.1932.0107.

[697] E. Feenberg. “The Scattering of Slow Electrons by Neutral Atoms”. In: *Physical Review* 40.1 (1932). Contains the optical theorem for potential scattering., pp. 40–54. DOI: 10.1103/PhysRev.40.40.

[698] V. Fock. “Konfigurationsraum und zweite Quantelung”. In: *Zeitschrift für Physik* 75.9-10 (1932), pp. 622–647. DOI: 10.1007/BF01344458.

[699] W. Heisenberg. “Über den Bau der Atomkerne. I”. In: *Zeitschrift für Physik* 77.1-2 (1932), pp. 1–11. DOI: 10.1007/BF01342433.

[700] R. J. Kennedy and E. M. Thorndike. “Experimental Establishment of the Relativity of Time”. In: *Physical Review* 42.3 (1932). Unequal-arm interferometer; tests the velocity dependence that Michelson–Morley alone cannot, and with it the time-dilation factor, pp. 400–418. DOI: 10.1103/PhysRev.42.400.

[701] H. Lamb. *Hydrodynamics*. 6th edition (the standard one). Classical reference for potential flow and surface waves. Cambridge University Press, 1932.

[702] L. D. Landau. “On the theory of stars”. In: *Physikalische Zeitschrift der Sowjetunion* 1 (1932). Written 1931, published 1932; venue has no DOIs, pp. 285–288.

[703] E. O. Lawrence and M. S. Livingston. “The Production of High Speed Light Ions Without the Use of High Voltages”. In: *Physical Review* 40.1 (1932). The cyclotron paper (Nobel Prize to Lawrence, 1939): resonant acceleration to protons above 1.2 MeV from a few kilovolts on the dees. Verified against api.crossref.org 2026-08-13., pp. 19–35. DOI: 10.1103/PhysRev.40.19.

[704] C. Møller. “Zur Theorie des Durchgangs schneller Elektronen durch Materie”. In: *Annalen der Physik* 406.5 (1932), pp. 531–585. DOI: 10.1002/andp.19324060506.

[705] J. v. Neumann. “Proof of the Quasi-Ergodic Hypothesis”. In: *Proceedings of the National Academy of Sciences* 18.1 (1932). Suffix a introduced by this batch. The bare key vonNeumann:1932 is used by six other citation sites (02-mathematical-methods/10-hilbert-spaces.tex, 09-quantum-mechanics/09-interpretations.tex, 09-quantum-mechanics/11-open-systems-decoherence.tex) for the BOOK ’Mathematische Grundlagen der Quantenmechanik’, which does not contain the mean ergodic theorem. This entry is the PNAS paper that does. The orchestrator may prefer to rename the book to vonNeumann:1932b for a/b symmetry., pp. 70–82. DOI: 10.1073/pnas.18.1.70.

[706] J. von Neumann. *Mathematische Grundlagen der Quantenmechanik*. Die Grundlehren der mathematischen Wissenschaften, Bd. 38. Title page prints the author as ’J. v. Neumann’; the particle is kept in the surname to match the key. Keeps the bare key as the clearly primary sense of ’von Neumann 1932’ (three of the four citing chapters mean it). No DOI for the 1932 original – Springer’s 10.1007/978-3-642-61409-5 is the 1996 reissue and must not be attached to this entry. Julius Springer, 1932.

[707] M. H. Stone. “On One-Parameter Unitary Groups in Hilbert Space”. In: *The Annals of Mathematics* 33.3 (1932). Annals of Mathematics, second series. Stone’s book ’Linear Transformations in Hilbert Space’ (AMS Colloquium 15, also 1932) is a different work – suffix if both are ever needed., p. 643. DOI: 10.2307/1968538.

[708] H. C. Urey, F. G. Brickwedde, and G. M. Murphy. “A Hydrogen Isotope of Mass 2”. In: *Physical Review* 39.1 (1932). Deuterium found by seeking the reduced-mass-shifted Balmer lines predicted by the Bohr model., pp. 164–165. DOI: 10.1103/PhysRev.39.164.

[709] J. H. Van Vleck. *The Theory of Electric and Magnetic Susceptibilities*. Oxford University Press, 1932.

[710] E. Wigner. “Über die Operation der Zeitumkehr in der Quantenmechanik”. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*. 1932, pp. 546–559.

[711] C. Zener. “Non-adiabatic crossing of energy levels”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 137.833 (1932). Verified via Crossref (Proc. R. Soc. Lond. A 137, issue 833, 1 Sep 1932). Crossref’s container title is the full 1932 masthead given above; "Proceedings of the Royal Society of London A" is the usual short form. This is the paper giving the exp(-2*pi*gamma) linear-sweep transition probability quoted at sec:approx-landau-zener., pp. 696–702. DOI: 10.1098/rspa.1932.0165.

[712] L. Alvarez and A. H. Compton. “A Positively Charged Component of Cosmic Rays”. In: *Physical Review* 43.10 (1933), pp. 835–836. DOI: 10.1103/PhysRev.43.835.

[713] C. D. Anderson. “The Positive Electron”. In: *Physical Review* 43.6 (1933). The full analysis, including the apparatus description and the 2 August 1932 photograph., pp. 491–494. DOI: 10.1103/PhysRev.43.491.

[714] P. M. S. Blackett and G. P. S. Occhialini. “Some photographs of the tracks of penetrating radiation”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 139.839 (1933). First photographs of electron-positron pair production., pp. 699–726. DOI: 10.1098/rspa.1933.0048.

[715] N. Bohr and L. Rosenfeld. “Zur Frage der Messbarkeit der elektromagnetischen Feldgrossen”. In: *Kongelige Danske Videnskabernes Selskab, Matematisk-fysiske Meddelelser* 12 (1933). Issue no. 8; no DOI, pp. 1–65.

[716] A. H. Compton. “A Geographic Study of Cosmic Rays”. In: *Physical Review* 43.6 (1933), pp. 387–403. DOI: 10.1103/PhysRev.43.387.

[717] T. G. Cowling. “The Magnetic Field of Sunspots”. In: *Monthly Notices of the Royal Astronomical Society* 94.1 (1933). Crossref confirmed: MNRAS vol. 94, issue 1, pp. 39-48, 10 November 1933. The anti-dynamo theorem (no steady axisymmetric dynamo) is in this paper, despite the title being about sunspots – that is the standard citation the chapter wants., pp. 39–48. DOI: 10.1093/mnras/94.1.39.

[718] P. A. M. Dirac. “The Lagrangian in Quantum Mechanics”. In: *Physikalische Zeitschrift der Sowjetunion* 3 (1933), pp. 64–72.

[719] P. Ehrenfest. “Phasenumwandlungen im ueblichen und erweiterten Sinn, classifiziert nach den entsprechenden Singularitaeten des thermodynamischen Potentiales”. In: *Proceedings of the Royal Academy of Sciences at Amsterdam*. Vol. 36. Also issued as Communications from the Physical Laboratory of the University of Leiden, Supplement 75b. 1933, pp. 153–157.

[720] E. Grüneisen. “Die Abhängigkeit des elektrischen Widerstandes reiner Metalle von der Temperatur”. In: *Annalen der Physik* 408.5 (1933), pp. 530–540. DOI: 10.1002/andp.19334080504.

[721] A. Haar. “Der Massbegriff in der Theorie der kontinuierlichen Gruppen”. In: *Annals of Mathematics* 34.1 (1933). Crossref-verified: author, year, volume and first page confirmed against the DOI. Existence of the invariant measure on a locally compact topological group, which is what lets Weyl’s averaging argument for complete reducibility pass from a finite group to a compact one. This treatise imports the theorem and does not construct the measure; the linear-algebra chapter says so where it uses it., pp. 147–169. DOI: 10.2307/1968346.

[722] T. H. Johnson. “The Azimuthal Asymmetry of the Cosmic Radiation”. In: *Physical Review* 43.10 (1933). Thomas H. Johnson; distinct from J. B. Johnson of Johnson:1928, pp. 834–835. DOI: 10.1103/PhysRev.43.834.

[723] G. Lemaître and M. S. Vallarta. “On Compton’s Latitude Effect of Cosmic Radiation”. In: *Physical Review* 43 (1933), pp. 87–91.

[724] W. Meissner and R. Ochsenfeld. “Ein neuer Effekt bei Eintritt der Supraleitfähigkeit”. In: *Die Naturwissenschaften* 21.44 (1933), pp. 787–788. DOI: 10.1007/BF01504252.

[725] J. Neyman and E. S. Pearson. “On the problem of the most efficient tests of statistical hypotheses”. In: *Philosophical Transactions of the Royal Society of London A* 231 (1933). Crossref-verified. TRAP: Crossref stores the title prefixed with the issue article number, as ’IX. On the problem...’; the numeral is not part of the title and is dropped here, the same artefact recorded for Rayleigh:1871. NOT to be confused with any other Pearson key in this bibliography: Pearson:1949 is Pearson

and Bardeen on silicon and Pearson:1958 is J. R. A. Pearson on convection cells, neither of them Egon Pearson., pp. 289–337. DOI: 10.1098/rsta.1933.0009.

[726] J. R. Oppenheimer and M. S. Plesset. “On the Production of the Positive Electron”. In: *Physical Review* 44.1 (1933), pp. 53–55. DOI: 10.1103/physrev.44.53.2.

[727] B. W. Sargent. “The maximum energy of the β-Rays from uranium X and other bodies”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 139.839 (1933), pp. 659–673. DOI: 10.1098/rspa.1933.0045.

[728] E. Wigner and F. Seitz. “On the Constitution of Metallic Sodium”. In: *Physical Review* 43.10 (1933), pp. 804–810. DOI: 10.1103/PhysRev.43.804.

[729] F. Zwicky. “Die Rotverschiebung von extragalaktischen Nebeln”. In: *Helvetica Physica Acta* 6 (1933). Coma-cluster virial discrepancy. No DOI., pp. 110–127.

[730] W. Baade and F. Zwicky. “Cosmic Rays from Super-Novae”. In: *Proceedings of the National Academy of Sciences* 20.5 (1934). Contains the neutron-star proposal; companion paper On Super-novae is PNAS 20, 254, pp. 259–263. DOI: 10.1073/pnas.20.5.259.

[731] W. H. Bennett. “Magnetically Self-Focussing Streams”. In: *Physical Review* 45.12 (1934). Source of the Bennett pinch relation. The published title uses the British double-s spelling ’Self-Focussing’ – do not normalise it to ’Focusing’. Crossref-verified., pp. 890–897. DOI: 10.1103/PhysRev.45.890.

[732] H. Bethe and W. Heitler. “On the stopping of fast particles and on the creation of positive electrons”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 146.856 (1934), pp. 83–112. DOI: 10.1098/rspa.1934.0140.

[733] W. L. Bragg and E. J. Williams. “The effect of thermal agitation on atomic arrangement in alloys”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 145.855 (1934), pp. 699–730. DOI: 10.1098/rspa.1934.0132.

[734] J. Chadwick and M. Goldhaber. “A Nuclear Photo-effect: Disintegration of the Diplon by -Rays”. In: *Nature* 134.3381 (1934), pp. 237–238. DOI: 10.1038/134237a0.

[735] P. A. Cherenkov. “Vidimoe svechenie chistykh zhidkostei pod deistviem γ-radiatsii”. In: *Doklady Akademii nauk SSSR* 2 (1934). THREE things the bib writer must decide, all flagged deliberately. (1) SCRIPT: the treatise builds with pdflatex, which will NOT compile Cyrillic without extra packages – do not paste the Russian title into references.bib unless the preamble is prepared for it. The work is universally cited in the Latin/English form printed in the parallel foreign-language edition (Comptes Rendus (Doklady) de l’Académie des Sciences de l’URSS): ’Visible emission of clean liquids by action of γ radiation’; that is the safe title field, with the Russian original recorded in a bib note. (2) PAGES: the standard citation is ’Dokl. Akad. Nauk SSSR 2, 451 (1934)’ – a single start page. Some bibliographies print 451-454; I could not verify the end page, so only 451 is asserted here. Do not invent 451-454. (3) NAME: the 1934 parallel edition prints ’Čerenkov’; the Russian original ’Cherenkov’. ’Cherenkov’ is given to match the citation key and the Nobel-era Latin usage. No DOI for the 1934 note; it was reprinted as Usp. Fiz. Nauk 93, 385 (1967), DOI 10.3367/UFNr.0093.196710n.0385 – that DOI belongs to the REPRINT, not to this entry, so do not attach it to a 1934-dated record without saying so. Russian-language original: the title and journal are given here in scientific transliteration, because the treatise builds with pdflatex under T1 encoding, which cannot typeset Cyrillic., p. 451.

[736] P. van Cittert. “Die Wahrscheinliche Schwingungsverteilung in Einer von Einer Lichtquelle Direkt Oder Mittels Einer Linse Beleuchteten Ebene”. In: *Physica* 1.1-6 (1934). Verified via Crossref: Physica 1, issue 1-6, 201-210 (1934); Crossref files the family name as "van Cittert" and title-cases the German title, the lower-case form given here is the printed one. BibTeX trap: write the author as van Cittert, P. H. (braced particle) or biber will alphabetize and abbreviate on "Cittert"; the same applies to vandeHulst:1945., pp. 201–210. DOI: 10.1016/S0031-8914(34)90026-4.

[737] I. Curie and F. Joliot. “Un nouveau type de radioactivité”. In: *Comptes Rendus Hebdomadaires des Séances de l’Académie des Sciences*. Vol. 198. Customarily cited as Joliot-Curie, but the paper lists Irène Curie first, so the key follows the first listed author. 1934, pp. 254–256.

[738] E. Fermi. “Versuch einer Theorie der ?-Strahlen. I”. In: *Zeitschrift fur Physik* 88.3-4 (1934). An Italian version appeared as Il Nuovo Cimento 11 (1934) 1-19, pp. 161–177. DOI: 10.1007/BF01351864.

[739] C. Gorter and H. Casimir. “On supraconductivity I”. In: *Physica* 1.1-6 (1934), pp. 306–320. DOI: 10.1016/S0031-8914(34)90037-9.

[740] A. Y. Khinchin. “Korrelationstheorie der stationären stochastischen Prozesse”. In: *Mathematische Annalen* 109 (1934). Byline ’A. Khintchine’., pp. 604–615.

[741] O. Klemperer. “On the annihilation radiation of the positron”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 30.3 (1934), pp. 347–354. DOI: 10.1017/s0305004100012536.

[742] C. C. Ko. “The heat of dissociation of Bi2 determined by the method of molecular beams”. In: *Journal of the Franklin Institute* 217.2 (1934). Issue 2. Crossref-confirmed (title as deposited by Elsevier is sentence case with ’Bi2’ unsubscripted; the printed title sets Bi_2 as a subscript, so a .bib entry should write Bi_2 and brace-protect it). WARNING for the title check: the title names the dissociation heat, not the velocity distribution, so it does not look like a velocity-analyser paper. It is one: the abstract states Ko used Zartman’s velocity analyser with improved technique to obtain the combined velocity spectrum of Bi atoms and Bi2 molecules, and derived the dissociation heat from it. This is the ’Ko’ of the standard Zartman–Ko pairing, and the paper the chapter’s phrase ’Zartman’s and Ko’s’ denotes., pp. 173–199. DOI: 10.1016/S0016-0032(34)91029-2.

[743] T. Koopmans. “Über die Zuordnung von Wellenfunktionen und Eigenwerten zu den Einzelnen Elektronen Eines Atoms”. In: *Physica* 1.1-6 (1934), pp. 104–113. DOI: 10.1016/S0031-8914(34)90011-2.

[744] H. Kramers. “L’interaction Entre les Atomes Magnétogènes dans un Cristal Paramagnétique”. In: *Physica* 1.1-6 (1934), pp. 182–192. DOI: 10.1016/s0031-8914(34)90023-9.

[745] J. Leray. “Sur le mouvement d’un liquide visqueux emplissant l’espace”. In: *Acta Mathematica* 63.0 (1934). Global weak solutions of the Navier-Stokes equations., pp. 193–248. DOI: 10.1007/BF02547354.

[746] M. Morse. *The Calculus of Variations in the Large*. Morse index theorem. Author is Marston Morse – distinct from Philip M. Morse of Morse:1953. New York: American Mathematical Society (Colloquium Publications 18), 1934.

[747] A. L. Patterson. “A Fourier Series Method for the Determination of the Components of Interatomic Distances in Crystals”. In: *Physical Review* 46.5 (1934). Crossref-verified: title, first author Patterson, Physical Review vol. 46, pp. 372-376, 1934. Cited for the vector (Patterson) map, phen:phase-patterson., pp. 372–376. DOI: 10.1103/PhysRev.46.372.

[748] W. Pauli and V. Weisskopf. “Über die Quantisierung der skalaren relativistischen Wellengleichung”. In: *Helvetica Physica Acta* 7 (1934). The reinterpretation of the Klein–Gordon equation as a quantized field. Helvetica Physica Acta of this period has no DOI; scans are on e-periodica., pp. 709–731.

[749] S. I. Vavilov. “O vozmozhnykh prichinakh sinego γ-svecheniya zhidkostei”. In: *Doklady Akademii Nauk SSSR* 2 (1934). Companion piece to Cherenkov’s paper in the SAME issue: Doklady AN SSSR, vol. 2, no. 8 (1934) — Cherenkov pp. 451-457, Vavilov pp. 457-461. No DOI exists; do not invent one (honest AMBER/RED on availability). BUILD WARNING: pdflatex cannot typeset Cyrillic, so references.bib must carry the transliteration, not the string above: title = O vozmozhnykh prichinakh sinego gamma-svecheniya zhidkostei [On the possible causes of the blue gamma-glow of liquids], journal = Doklady Akademii Nauk SSSR. Doklady is the USSR academy’s Comptes Rendus; it does carry volume and issue numbers, so @article with volume 2 is correct here rather than @inproceedings — but if the project prefers the academy-proceedings treatment, note that the volume is genuine and must not be dropped. Russian-language original: the title and journal are given here in scientific transliteration, because the treatise builds with pdflatex under T1 encoding, which cannot typeset Cyrillic., pp. 457–461.

[750] C. Zener. “A theory of the electrical breakdown of solid dielectrics”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 145.855 (1934), pp. 523–529. DOI: 10.1098/rspa.1934.0116.

[751] E. U. Condon and G. H. Shortley. *The Theory of Atomic Spectra*. Cited in the angular-momentum chapter as a standard treatment, alongside Sakurai:2017 and CohenTannoudji:1977 – unmistakably Condon & Shortley, the source of the Condon-Shortley phase convention. First edition Cambridge, 1935; reprinted many times (1951, 1959, 1963 ...) and reissued in paperback 1970. Two authors: an @book still owes both. Printed monograph, no DOI – honest AMBER on availability. Cambridge University Press, 1935.

[752] T. De Donder. *Théorie invariantive du calcul des variations*. Nouvelle édition, Paris, 1935 – the edition the key’s year refers to and the one universally cited for the De Donder-Weyl covariant Hamiltonian formalism. A first edition appeared in 1930 (same publisher); record the edition explicitly so the two are not conflated. NAME: the title page prints ’Th. De Donder’ but library catalogues file him as ’Donder, Théophile de’ – biblatex should be given De Donder, Th. with the surname braced so it is not split. Printed monograph, no DOI: honest AMBER unless a scan is stored. Gauthier-Villars, 1935.

[753] A. Einstein, B. Podolsky, and N. Rosen. “Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?” In: *Physical Review* 47.10 (1935), pp. 777–780. DOI: 10.1103/PhysRev.47.777.

[754] H. Eyring. “The Activated Complex in Chemical Reactions”. In: *The Journal of Chemical Physics* 3.2 (1935). Issue 2. The transition-state concept that the femtochemistry measurements put under test. Confirmed against the Crossref record for this DOI., pp. 107–115. DOI: 10.1063/1.1749604.

[755] V. Fock. “Zur Theorie des Wasserstoffatoms”. In: *Zeitschrift für Physik* 98.3-4 (1935). Verified against Crossref. The hidden SO(4) symmetry of the Coulomb problem via momentum-space stereographic projection; Bargmann’s 1936 companion paper (Z. Phys. 99, 576) is the separate Bargmann:1936 key., pp. 145–154. DOI: 10.1007/BF01336904.

[756] L. Landau and E. Lifshitz. “On the theory of the dispersion of magnetic permeability in ferromagnetic bodies”. In: *Physikalische Zeitschrift der Sowjetunion* 8 (1935), pp. 153–169.

[757] F. London and H. London. “The electromagnetic equations of the supraconductor”. In: *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* 149.866 (1935), pp. 71–88. DOI: 10.1098/rspa.1935.0048.

[758] H. P. Robertson. “Kinematics and World-Structure”. In: *The Astrophysical Journal* 82 (1935). Parts II and III follow in ApJ 83 (1936)., p. 284. DOI: 10.1086/143681.

[759] E. Schrödinger. “Discussion of Probability Relations between Separated Systems”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 31.4 (1935), pp. 555–563. DOI: 10.1017/S0305004100013554.

[760] E. Schrödinger. “Die gegenwärtige Situation in der Quantenmechanik”. In: *Die Naturwissenschaften* 23 (1935). Three-part article: pp. 807-812, 823-828, 844-849. The DOI resolves the first installment; the term ’Verschränkung’ and the cat both appear here., pp. 807–812.

[761] A. Tarski. “Der Wahrheitsbegriff in den formalisierten Sprachen”. In: *Studia Philosophica* 1 (1935). German translation by L. Blaustein, with a new Nachwort, of the Polish original of 1933. The volume is dated 1935 and is cited so, but did not appear until 1936, pp. 261–405.

[762] E. A. Uehling. “Polarization Effects in the Positron Theory”. In: *Physical Review* 48.1 (1935), pp. 55–63. DOI: 10.1103/PhysRev.48.55.

[763] C. F. von Weizsäcker. “Zur Theorie der Kernmassen”. In: *Zeitschrift für Physik* 96 (1935), pp. 431–458.

[764] H. Weyl. “Geodesic Fields in the Calculus of Variation for Multiple Integrals”. In: *The Annals of Mathematics* 36.3 (1935). Title, journal, volume 36 issue 3, start page 607 and year 1935 verified via Crossref; Crossref stores only the start page, the end page 629 is the standard JSTOR range. Crossref’s container title is "The Annals of Mathematics" (the journal’s name in 1935) — either form is defensible. This is the Weyl half of the De Donder-Weyl formalism cited at sec:lagft-covariant-hamiltonian., p. 607. DOI: 10.2307/1968645.

[765] H. Yukawa. “On the Interaction of Elementary Particles. I”. In: *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series* 17 (1935), pp. 48–57.

[766] V. Bargmann. “Zur Theorie des Wasserstoffatoms”. In: *Zeitschrift für Physik* 99.7-8 (1936). The SO(4) hidden symmetry of the Coulomb problem. The printed article carries the subtitle ’Bemerkungen zur gleichnamigen Arbeit von V. Fock’ – add it to the bib entry if desired, but note that Crossref records only the main title, so the automated DOI/title check should be run against the short form given here. The main title is IDENTICAL to Fock:1935 (Z. Phys. 98, 145, DOI 10.1007/BF01336904) – do not let the two entries collapse., pp. 576–582. DOI: 10.1007/BF01338811.

[767] R. A. Beth. “Mechanical Detection and Measurement of the Angular Momentum of Light”. In: *Physical Review* 50 (1936), pp. 115–125. DOI: 10.1103/PhysRev.50.115.

[768] H. A. Bethe and R. F. Bacher. “Nuclear Physics A. Stationary States of Nuclei”. In: *Reviews of Modern Physics* 8.2 (1936), pp. 82–229. DOI: 10.1103/RevModPhys.8.82.

[769] H. J. Bhabha. “The scattering of positrons by electrons with exchange on Dirac’s theory of the positron”. In: *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* 154.881 (1936), pp. 195–206. DOI: 10.1098/rspa.1936.0046.

[770] N. Bohr. “Neutron Capture and Nuclear Constitution”. In: *Nature* 137.3461 (1936), pp. 344–348. DOI: 10.1038/137344a0.

[771] G. Breit and E. Wigner. “Capture of Slow Neutrons”. In: *Physical Review* 49.7 (1936), pp. 519–531. DOI: 10.1103/PhysRev.49.519.

[772] A. Church. “A Note on the Entscheidungsproblem”. In: *Journal of Symbolic Logic* 1.1 (1936). Correction: Journal of Symbolic Logic 1 (1936), 101–102, doi 10.2307/2269030, pp. 40–41. DOI: 10.2307/2269326.

[773] A. Church. “An Unsolvable Problem of Elementary Number Theory”. In: *American Journal of Mathematics* 58.2 (1936), pp. 345–363. DOI: 10.2307/2371045.

[774] A. Einstein. “Lens-Like Action of a Star by the Deviation of Light in the Gravitational Field”. In: *Science* 84.2188 (1936), pp. 506–507. DOI: 10.1126/science.84.2188.506.

[775] G. Gamow and E. Teller. “Selection Rules for theβ-Disintegration”. In: *Physical Review* 49.12 (1936), pp. 895–899. DOI: 10.1103/PhysRev.49.895.

[776] G. Gentzen. “Die Widerspruchsfreiheit der reinen Zahlentheorie”. In: *Mathematische Annalen* 112 (1936). Consistency of Peano arithmetic by transfinite induction up to the ordinal epsilon-nought, pp. 493–565. DOI: 10.1007/BF01565428.

[777] W. Heisenberg and H. Euler. “Folgerungen aus der Diracschen Theorie des Positrons”. In: *Zeitschrift fur Physik* 98.11-12 (1936). Source of the Euler-Heisenberg effective Lagrangian, pp. 714–732. DOI: 10.1007/BF01343663.

[778] S. C. Kleene. “General Recursive Functions of Natural Numbers”. In: *Mathematische Annalen* 112 (1936), pp. 727–742. DOI: 10.1007/BF01565439.

[779] F. N. D. Kurie, J. R. Richardson, and H. C. Paxton. “The Radiations Emitted from Artificially Produced Radioactive Substances. I. The Upper Limits and Shapes of theβ-Ray Spectra from Several Elements”. In: *Physical Review* 49.5 (1936), pp. 368–381. DOI: 10.1103/PhysRev.49.368.

[780] E. Madelung. *Die mathematischen Hilfsmittel des Physikers*. 3rd edition; the n+l filling rule appears in the tables appendix. Klechkovskii restated it independently in the 1950s. Springer, 1936.

[781] L. Néel. “Propriétés magnétiques de l’état métallique et énergie d’interaction entre atomes magnétiques”. In: *Annales de physique* 11.5 (1936). 11th series., pp. 232–279. DOI: 10.1051/anphys/193611050232.

[782] R. Peierls. “On Ising’s model of ferromagnetism”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 32.3 (1936), pp. 477–481. DOI: 10.1017/S0305004100019174.

[783] S. J. Plimpton and W. E. Lawton. “A Very Accurate Test of Coulomb’s Law of Force Between Charges”. In: *Physical Review* 50.11 (1936). Issue 11, 1 December 1936., pp. 1066–1071. DOI: 10.1103/PhysRev.50.1066.

[784] E. L. Post. “Finite Combinatory Processes—Formulation 1”. In: *Journal of Symbolic Logic* 1.3 (1936), pp. 103–105. DOI: 10.2307/2269031.

[785] A. Proca. “Sur la théorie ondulatoire des électrons positifs et négatifs”. In: *Journal de Physique et le Radium* 7.8 (1936). Série 7, issue 8, 1936. The DOI contains a colon, which is legal in a DOI but must be brace-protected in the .bib field and left unescaped in the URL. Author’s name is printed ’Al. Proca’ on the paper; the full form is Alexandru Proca., pp. 347–353. DOI: 10.1051/jphysrad:0193600708034700.

[786] B. Rosser. “Extensions of Some Theorems of Gödel and Church”. In: *Journal of Symbolic Logic* 1.3 (1936), pp. 87–91. DOI: 10.2307/2269028.

[787] S. L. Sobolev. “Méthode nouvelle à résoudre le problème de Cauchy pour les équations linéaires hyperboliques normales”. In: *Matematicheskii Sbornik* 1(43) (1936). Byline ’S. Soboleff’. Volume carries the double numbering 1(43)., pp. 39–72.

[788] F. Bloch and A. Nordsieck. “Note on the Radiation Field of the Electron”. In: *Physical Review* 52.2 (1937), pp. 54–59. DOI: 10.1103/PhysRev.52.54.

[789] P. A. Cherenkov. “Visible Radiation Produced by Electrons Moving in a Medium with Velocities Exceeding that of Light”. In: *Physical Review* 52 (1937), pp. 378–379.

[790] M. Fierz. “Zur Fermischen Theorie des ?-Zerfalls”. In: *Zeitschrift für Physik* 104.7-8 (1937). Source of the Fierz rearrangement identities for spinor bilinears., pp. 553–565. DOI: 10.1007/bf01330070.

[791] S. E. Forbush. “On the Effects in Cosmic-Ray Intensity Observed During the Recent Magnetic Storm”. In: *Physical Review* 51.12 (1937), pp. 1108–1109. DOI: 10.1103/physrev.51.1108.3.

[792] I. Frank and I. Tamm. “Coherent Visible Radiation of Fast Electrons Passing Through Matter”. In: *Comptes Rendus (Doklady) de l’Académie des Sciences de l’URSS*. Vol. 14. 1937, pp. 109–114.

[793] W. H. Furry. “A Symmetry Theorem in the Positron Theory”. In: *Physical Review* 51.2 (1937), pp. 125–129. DOI: 10.1103/PhysRev.51.125.

[794] H. A. Jahn and E. Teller. “Stability of polyatomic molecules in degenerate electronic states - I—Orbital degeneracy”. In: *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* 161.905 (1937), pp. 220–235. DOI: 10.1098/rspa.1937.0142.

[795] L. D. Landau. “On the theory of phase transitions”. In: *Zhurnal Eksperimental’noi i Teoreticheskoi Fiziki* 7 (1937), pp. 19–32.

[796] E. Majorana. “Teoria simmetrica dell’elettrone e del positrone”. In: *Il Nuovo Cimento* 14.4 (1937), pp. 171–184. DOI: 10.1007/BF02961314.

[797] S. H. Neddermeyer and C. D. Anderson. “Note on the Nature of Cosmic-Ray Particles”. In: *Physical Review* 51.10 (1937). Cited as the counter-example: the same apparatus and argument style yielding a particle that was misidentified for a decade., pp. 884–886. DOI: 10.1103/PhysRev.51.884.

[798] I. I. Rabi. “Space Quantization in a Gyrating Magnetic Field”. In: *Physical Review* 51.8 (1937). Issue 8, 15 April 1937. Single-author paper – this is the exact two-level solution and the Rabi formula, distinct from Rabi:1938 (the molecular-beam magnetic-resonance method, Rabi, Zacharias, Millman and Kusch), also cited in the same sentence., pp. 652–654. DOI: 10.1103/PhysRev.51.652.

[799] G. Racah. “Sulla Simmetria Tra Particelle e Antiparticelle”. In: *Il Nuovo Cimento* 14.7 (1937), pp. 322–328. DOI: 10.1007/bf02961321.

[800] T. Schmidt. “Über die magnetischen Momente der Atomkerne”. In: *Zeitschrift für Physik* 106.5-6 (1937), pp. 358–361. DOI: 10.1007/bf01338744.

[801] J. Street and E. Stevenson. “New Evidence for the Existence of a Particle of Mass Intermediate Between the Proton and Electron”. In: *Physical Review* 52.9 (1937), pp. 1003–1004. DOI: 10.1103/PhysRev.52.1003.

[802] E. C. Titchmarsh. *Introduction to the Theory of Fourier Integrals*. Canonical monograph for the transform, convolution, Poisson summation, Hankel and Mellin material. Oxford: Clarendon Press, 1937.

[803] A. M. Turing. “On Computable Numbers, with an Application to the Entscheidungsproblem”. In: *Proceedings of the London Mathematical Society* s2-42.1 (1937). Read 12 November 1936 and universally cited as 1936; pages 230–240 appeared in part 3 of volume s2-42, issued 30 November 1936, and the remainder in part 4, issued 23 December 1936, but the publisher dates the bound volume s2-42 to 1937, which fixes the key. Correction: Proceedings of the London Mathematical Society s2-43 (1938), 544–546, doi 10.1112/plms/s2-43.6.544, pp. 230–265. DOI: 10.1112/plms/s2-42.1.230.

[804] A. G. Walker. “On Milne’s Theory of World-Structure*”. In: *Proceedings of the London Mathematical Society* s2-42.1 (1937). Series 2, vol. 42., pp. 90–127. DOI: 10.1112/plms/s2-42.1.90.

[805] G. H. Wannier. “The Structure of Electronic Excitation Levels in Insulating Crystals”. In: *Physical Review* 52.3 (1937), pp. 191–197. DOI: 10.1103/PhysRev.52.191.

[806] J. A. Wheeler. “On the Mathematical Description of Light Nuclei by the Method of Resonating Group Structure”. In: *Physical Review* 52.11 (1937). The paper in which the scattering matrix is introduced and named., pp. 1107–1122. DOI: 10.1103/PhysRev.52.1107.

[807] F. Zwicky. “On the Masses of Nebulae and of Clusters of Nebulae”. In: *The Astrophysical Journal* 86 (1937), p. 217. DOI: 10.1086/143864.

[808] J. F. ALLEN and H. JONES. “New Phenomena Connected with Heat Flow in Helium II”. In: *Nature* 141.3562 (1938), pp. 243–244. DOI: 10.1038/141243a0.

[809] J. F. ALLEN and A. D. MISENER. “Flow of Liquid Helium II”. In: *Nature* 141.3558 (1938), pp. 75–75. DOI: 10.1038/141075a0.

[810] H. A. Bethe and C. L. Critchfield. “The Formation of Deuterons by Proton Combination”. In: *Physical Review* 54.4 (1938), pp. 248–254. DOI: 10.1103/PhysRev.54.248.

[811] B. Davydov. “The Rectifying Action of Semiconductors”. In: *Technical Physics of the USSR* 5 (1938). No DOI; Soviet journal. One of the three independent barrier theories of rectification, pp. 87–95.

[812] P. A. M. Dirac. “A new basis for cosmology”. In: *Proceedings of the Royal Society of London. Series A* 165.921 (1938). The large-numbers hypothesis, which requires G to vary with cosmic time and is excluded by the bounds of the appendix, pp. 199–208. DOI: 10.1098/rspa.1938.0053.

[813] P. A. M. Dirac. “Classical theory of radiating electrons”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 167.929 (1938). Crossref confirmed: vol. 167, pp. 148-169, 5 August 1938. The ’a’ suffix is deliberate and is documented in a comment at the top of the citing chapter: the key Dirac:1938 is ALREADY TAKEN in references.bib by the large-numbers cosmology note in Nature 139. Do not merge or renumber them. This is the Lorentz-Dirac equation paper (covariant radiation reaction, runaway and pre-accelerating solutions)., pp. 148–169. DOI: 10.1098/rspa.1938.0124.

[814] R. J. Duffin. “On The Characteristic Matrices of Covariant Systems”. In: *Physical Review* 54.12 (1938), pp. 1114–1114. DOI: 10.1103/PhysRev.54.1114.

[815] K. Gödel. “The Consistency of the Axiom of Choice and of the Generalized Continuum-Hypothesis”. In: *Proceedings of the National Academy of Sciences of the United States of America* 24.12 (1938). Announcement; expanded in Proceedings of the National Academy of Sciences 25 (1939), 220–224, and in the 1940 Princeton monograph, pp. 556–557. DOI: 10.1073/pnas.24.12.556.

[816] H. E. Ives and G. R. Stilwell. “An Experimental Study of the Rate of a Moving Atomic Clock”. In: *Journal of the Optical Society of America* 28.7 (1938). The transverse (second-order) Doppler shift in hydrogen canal rays: the first laboratory measurement of time dilation. The authors quote no single ratio with an uncertainty; a later assessment by Mandelberg and Witten, J. Opt. Soc. Am. 52, 529 (1962), puts the experimental uncertainty near 10–15%, pp. 215–226. DOI: 10.1364/JOSA.28.000215.

[817] P. Kapitza. “Viscosity of Liquid Helium below the λ-Point”. In: *Nature* 141.3558 (1938), pp. 74–74. DOI: 10.1038/141074a0.

[818] F. LONDON. “The λ-Phenomenon of Liquid Helium and the Bose-Einstein Degeneracy”. In: *Nature* 141.3571 (1938), pp. 643–644. DOI: 10.1038/141643a0.

[819] I. I. Rabi, J. R. Zacharias, S. Millman, and P. Kusch. “A New Method of Measuring Nuclear Magnetic Moment”. In: *Physical Review* 53.4 (1938), pp. 318–318. DOI: 10.1103/PhysRev.53.318.

[820] E. C. Stoner. “Collective electron ferromagnetism”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 165.922 (1938), pp. 372–414. DOI: 10.1098/rspa.1938.0066.

[821] E. C. G. Stueckelberg. “Die Wechselwirkungskräfte in der Elektrodynamik und in der Feldtheorie der Kernkräfte”. In: *Helvetica Physica Acta* 11 (1938), pp. 225–244.

[822] L. TISZA. “Transport Phenomena in Helium II”. In: *Nature* 141.3577 (1938), pp. 913–913. DOI: 10.1038/141913a0.

[823] A. A. Vlasov. “O vibratsionnykh svoistvakh elektronnogo gaza”. In: *Zhurnal eksperimentalnoi i teoreticheskoi fiziki* 8 (1938). ZhETF 8, no. 3, 291-318 (1938); the start page 291 is confirmed by several independent citations and the 291-318 range by the Springer/TMF reference list. The original has no DOI. BUILD WARNING: transliterate for pdflatex — title = O vibratsionnykh svoistvakh elektronnogo gaza [On the vibrational properties of an electron gas], journal = Zhurnal Eksperimental’noi i Teoreticheskoi Fiziki. Two retrievable descendants exist and give the reader a real access route, so at least one belongs in the entry’s note or as a related entry: Russian reprint Uspekhi Fiz. Nauk 93, 444-470 (1967), doi:10.3367/UFNr.0093.196711f.0444; English translation "The vibrational properties of an electron gas", Soviet Physics Uspekhi 10, 721-733 (1968), doi:10.1070/PU1968v010n06ABEH003709 (both DOIs resolved and title-checked). Russian-language original: the title and journal are given here in scientific transliteration, because the treatise builds with pdflatex under T1 encoding, which cannot typeset Cyrillic., pp. 291–318.

[824] C. F. von Weizsäcker. “Über Elementumwandlungen im Innern der Sterne. II”. In: *Physikalische Zeitschrift* 39 (1938). CNO cycle proposed independently of Bethe; no DOI, pp. 633–646.

[825] S. S. Wilks. “The large-sample distribution of the likelihood ratio for testing composite hypotheses”. In: *The Annals of Mathematical Statistics* 9.1 (1938). Crossref-verified. The asymptotic chi-squared distribution of twice the log likelihood ratio, with degrees of freedom equal to the number of constrained parameters. The hypotheses matter and are stated where it is used: the result fails at a parameter boundary (Chernoff:1954) and when a parameter is unidentified under the null, which is the case in a bump search and is why the look-elsewhere effect exists., pp. 60–62. DOI: 10.1214/aoms/1177732360.

[826] F. Zernike. “The concept of degree of coherence and its application to optical problems”. In: *Physica* 5.8 (1938). Verified via Crossref (Physica 5(8), Aug 1938). This is the Zernike half of the van Cittert-Zernike theorem, paired in the prose with vanCittert:1934., pp. 785–795. DOI: 10.1016/S0031-8914(38)80203-2.

[827] P. Auger, P. Ehrenfest, R. Maze, J. Daudin, and R. A. Fréon. “Extensive Cosmic-Ray Showers”. In: *Reviews of Modern Physics* 11.3-4 (1939), pp. 288–291. DOI: 10.1103/RevModPhys.11.288.

[828] H. A. Bethe. “Energy Production in Stars”. In: *Physical Review* 55.5 (1939), pp. 434–456. DOI: 10.1103/PhysRev.55.434.

[829] N. BOHR, R. PEIERLS, and G. PLACZEK. “Nuclear Reactions in the Continuous Energy Region”. In: *Nature* 144.3639 (1939). The general, inelastic-channel statement of the optical theorem., pp. 200–201. DOI: 10.1038/144200a0.

[830] N. Bohr and J. A. Wheeler. “The Mechanism of Nuclear Fission”. In: *Physical Review* 56.5 (1939). Suffixed to avoid a collision: 09-quantum-mechanics/10-scattering-theory.tex uses Bohr:1939 for Bohr, Peierls and Placzek, ’Nuclear Reactions in the Continuous Energy Region’, Nature 144, 200-201, doi 10.1038/144200a0. Two different 1939 Bohr papers., pp. 426–450. DOI: 10.1103/PhysRev.56.426.

[831] S. Chandrasekhar. *An Introduction to the Study of Stellar Structure*. University of Chicago Press, 1939.

[832] P. A. M. Dirac. “A new notation for quantum mechanics”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 35.3 (1939). Crossref-verified (resolved record read: Dirac, P. A. M., Math. Proc. Cambridge Philos. Soc. 35(3) 416–418, 1939). The paper that introduces the bra–ket notation and the names “bra” and “ket”; it is the correct citation for the notation, not the 1930 first edition of the Principles (Dirac:1930b). The notation enters the book in the third edition (1947), and Dirac:1958 is the fourth., pp. 416–418. DOI: 10.1017/S0305004100021162.

[833] R. P. Feynman. “Forces in Molecules”. In: *Physical Review* 56.4 (1939). Verified against Crossref. This is the Feynman half of the Hellmann–Feynman theorem and the source of the electrostatic theorem for molecular binding; Hellmann’s 1937 Einführung in die Quantenchemie is the other half and would need its own key., pp. 340–343. DOI: 10.1103/PhysRev.56.340.

[834] M. Fierz and W. E. Pauli. “On relativistic wave equations for particles of arbitrary spin in an electromagnetic field”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 173.953 (1939), pp. 211–232. DOI: 10.1098/rspa.1939.0140.

[835] R. H. Fowler and E. A. Guggenheim. *Statistical Thermodynamics*. Source of the name ’zeroth law’. Cambridge University Press, 1939.

[836] W. H. Furry. “On Transition Probabilities in Double Beta-Disintegration”. In: *Physical Review* 56.12 (1939), pp. 1184–1193. DOI: 10.1103/PhysRev.56.1184.

[837] O. Hahn and F. Strassmann. “Über den Nachweis und das Verhalten der bei der Bestrahlung des Urans mittels Neutronen entstehenden Erdalkalimetalle”. In: *Die Naturwissenschaften* 27.1 (1939), pp. 11–15. DOI: 10.1007/BF01488241.

[838] O. Halpern and M. H. Johnson. “On the Magnetic Scattering of Neutrons”. In: *Physical Review* 55.10 (1939). Issue 10, 15 May 1939. Crossref-verified. This is the theory paper for the dipole coupling of the neutron to unpaired-electron moments, with the magnetic form factor and geometrical factor that the chapter’s derivation packs into p; placed at the claim in sec:nscat-why that magnetic order produces diffraction peaks of ordinary strength (line 131). No collision: the only other Halpern

in the bib is M. Halpern as a WMAP coauthor in Bennett:2013., pp. 898–923. DOI: 10.1103/PhysRev.55.898.

[839] D. Hilbert and P. Bernays. *Grundlagen der Mathematik*. Vol. 2. Die Grundlehren der mathematischen Wissenschaften 50. First edition, xii+498 pp.; source of the derivability conditions for the second incompleteness theorem. Berlin: Julius Springer, 1939.

[840] J. M. B. Kellogg, I. I. Rabi, N. F. Ramsey, and J. R. Zacharias. "An Electrical Quadrupole Moment of the Deuteron". In: *Physical Review* 55.3 (1939), pp. 318–319. DOI: 10.1103/PhysRev.55.318.

[841] N. Kemmer. "The particle aspect of meson theory". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 173.952 (1939), pp. 91–116. DOI: 10.1098/rspa.1939.0131.

[842] L. Meitner and O. R. Frisch. "Disintegration of Uranium by Neutrons: a New Type of Nuclear Reaction". In: *Nature* 143.3615 (1939), pp. 239–240. DOI: 10.1038/143239a0.

[843] N. F. Mott. "The theory of crystal rectifiers". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 171.944 (1939), pp. 27–38. DOI: 10.1098/rspa.1939.0051.

[844] J. R. Oppenheimer and H. Snyder. "On Continued Gravitational Contraction". In: *Physical Review* 56.5 (1939), pp. 455–459. DOI: 10.1103/PhysRev.56.455.

[845] J. R. Oppenheimer and G. M. Volkoff. "On Massive Neutron Cores". In: *Physical Review* 55.4 (1939). Suffix a: two Oppenheimer 1939 papers are cited (see Oppenheimer:1939b); do not create an unsuffixed Oppenheimer:1939, pp. 374–381. DOI: 10.1103/PhysRev.55.374.

[846] N. Osborne, H. Stimson, and D. Ginnings. "Measurements of heat capacity and heat of vaporization of water in the range 0 degrees to 100 degrees C". In: *Journal of Research of the National Bureau of Standards* 23.2 (1939), p. 197. DOI: 10.6028/jres.023.008.

[847] W. Schottky. "Zur Halbleitertheorie der Sperrschicht- und Spitzengleichrichter". In: *Zeitschrift für Physik* 113.5-6 (1939), pp. 367–414. DOI: 10.1007/bf01340116.

[848] R. C. Tolman. "Static Solutions of Einstein's Field Equations for Spheres of Fluid". In: *Physical Review* 55.4 (1939), pp. 364–373. DOI: 10.1103/PhysRev.55.364.

[849] E. P. Wigner. "On Unitary Representations of the Inhomogeneous Lorentz Group". In: *Annals of Mathematics* 40 (1939), pp. 149–204. DOI: 10.2307/1968551.

[850] F. Belinfante. "On the current and the density of the electric charge, the energy, the linear momentum and the angular momentum of arbitrary fields". In: *Physica* 7.5 (1940). The improved (Belinfante–Rosenfeld) energy–momentum tensor cited in the chapter. Crossref-verified; the DOI's check digit-looking tail 'X' is part of the suffix, not a typo. Distinct from his 1939 Physica 6, 887 paper on meson spin., pp. 449–474. DOI: 10.1016/S0031-8914(40)90091-X.

[851] C. Herring. "A New Method for Calculating Wave Functions in Crystals". In: *Physical Review* 57.12 (1940), pp. 1169–1177. DOI: 10.1103/PhysRev.57.1169.

[852] T. Holstein and H. Primakoff. "Field Dependence of the Intrinsic Domain Magnetization of a Ferromagnet". In: *Physical Review* 58.12 (1940), pp. 1098–1113. DOI: 10.1103/PhysRev.58.1098.

[853] H. Kramers. "Brownian motion in a field of force and the diffusion model of chemical reactions". In: *Physica* 7.4 (1940), pp. 284–304. DOI: 10.1016/S0031-8914(40)90098-2.

[854] W. Pauli. “The Connection Between Spin and Statistics”. In: *Physical Review* 58.8 (1940), pp. 716–722. DOI: 10.1103/PhysRev.58.716.

[855] L. Rosenfeld. “Sur le tenseur d’impulsion-énergie”. In: *Mémoires de l’Académie Royale de Belgique, Classe des Sciences*. Vol. 18. No DOI exists; do not fabricate one – this will grade AMBER on availability unless a scan is stored. Full designation: Académie Royale de Belgique, Classe des Sciences, Mémoires, Collection in-8, tome 18, fascicule 6 (1940), separately paginated 1-30 (hence citations of the form ’18(6), 1-30’). The fascicule number is worth recording. English translation: ’On the energy-momentum tensor’, in Selected Papers of Léon Rosenfeld, ed. R. S. Cohen and J. J. Stachel (Reidel, Dordrecht, 1979), pp. 711-735 – a retrievable route for a reader who cannot get the Mémoires. Cited alongside Belinfante:1940 for the improvement term. 1940, pp. 1–30.

[856] A. C. Berry. “The accuracy of the Gaussian approximation to the sum of independent variates”. In: *Transactions of the American Mathematical Society* 49 (1941), pp. 122–136. DOI: 10.1090/S0002-9947-1941-0003498-3.

[857] A. N. Kolmogorov. “The Local Structure of Turbulence in Incompressible Viscous Fluid for Very Large Reynolds Numbers”. In: *Doklady Akademii Nauk SSSR* 30 (1941). In Russian. English translation: Proceedings of the Royal Society of London A 434 (1991) 9-13, DOI 10.1098/rspa.1991.0075 (that DOI belongs to the translation, not this entry). Companion dissipation paper: Dokl. Akad. Nauk SSSR 32 (1941) 16-18., pp. 299–303.

[858] L. Landau. “Theory of the Superfluidity of Helium II”. In: *Physical Review* 60.4 (1941), pp. 356–358. DOI: 10.1103/PhysRev.60.356.

[859] R. H. Lyddane, R. G. Sachs, and E. Teller. “On the Polar Vibrations of Alkali Halides”. In: *Physical Review* 59.8 (1941), pp. 673–676. DOI: 10.1103/PhysRev.59.673.

[860] A. McKellar. “Molecular Lines from the Lowest States of Diatomic Molecules Composed of Atoms Probably Present in Interstellar Space”. In: *Publications of the Dominion Astrophysical Observatory Victoria* 7 (1941). CN excitation temperature 2.3 K; no DOI, pp. 251–272.

[861] A. M. Obukhov. “On the Distribution of Energy in the Spectrum of Turbulent Flow”. In: *Izvestiya Akademii Nauk SSSR, Seriya Geograficheskaya i Geofizicheskaya* 5 (1941). In Russian. Short announcement also in Dokl. Akad. Nauk SSSR 32 (1941) 22-24. Source of the spectral k^(-5/3) form; the chapter records the split attribution with K41., pp. 453–466.

[862] W. Rarita and J. Schwinger. “On a Theory of Particles with Half-Integral Spin”. In: *Physical Review* 60.1 (1941), pp. 61–61. DOI: 10.1103/PhysRev.60.61.

[863] B. Rossi and D. B. Hall. “Variation of the Rate of Decay of Mesotrons with Momentum”. In: *Physical Review* 59.3 (1941). Cosmic-ray muon survival between Echo Lake and Denver; yields a proper lifetime $(2.4 \pm 0.3) \times 10^{-6}$ s. “Mesotron” is the contemporary name for the muon, pp. 223–228. DOI: 10.1103/PhysRev.59.223.

[864] E. C. G. Stueckelberg. “Remarque à propos de la création de paires de particules en théorie de relativité”. In: *Helvetica Physica Acta* 14 (1941). Not a genuine collision: both candidates describe this single paper, one with the accents stripped. Source of the Stueckelberg–Feynman reading of antiparticles. Name also written Stückelberg; the paper itself is signed E. C. G. Stueckelberg. Helvetica Physica Acta of this period has no DOI; scans are at e-periodica.ch., pp. 588–594.

[865] H. ALFVÉN. “Existence of Electromagnetic-Hydrodynamic Waves”. In: *Nature* 150.3805 (1942). The frozen-in field / Alfvén wave letter. Nature prints the title with a hyphen (’Electromagnetic-Hydrodynamic’), not an en dash. Surname carries an acute accent: Alfvén. Crossref-verified., pp. 405–406. DOI: 10.1038/150405d0.

[866] E. Hopf. “Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems”. In: *Berichte der Mathematisch-Physischen Klasse der Sächsischen Akademie der Wissenschaften zu Leipzig*. Vol. 94. The Hopf bifurcation. English translation in Marsden and McCracken, The Hopf Bifurcation and Its Applications (Springer, 1976). 1942, pp. 1–22.

[867] G. Racah. “Theory of Complex Spectra. II”. In: *Physical Review* 62.9-10 (1942). Issue 9–10, 1 November 1942. Part II is the one carrying the closed-form (Racah) expression for the general Clebsch–Gordan coefficient and introducing the W coefficient; Part I is Phys. Rev. 61, 186 (1942) and Part III, with the 6j/9j apparatus, is Phys. Rev. 63, 367 (1943). The author signs ’G. Racah’; forename Giulio (Hebrew: Yoel)., pp. 438–462. DOI: 10.1103/PhysRev.62.438.

[868] M. Schönberg and S. Chandrasekhar. “On the Evolution of the Main-Sequence Stars”. In: *The Astrophysical Journal* 96 (1942), p. 161. DOI: 10.1086/144444.

[869] S. Chandrasekhar. “Stochastic Problems in Physics and Astronomy”. In: *Reviews of Modern Physics* 15.1 (1943), pp. 1–89. DOI: 10.1103/RevModPhys.15.1.

[870] M. Fréchet. “Sur l’extension de certaines évaluations statistiques au cas de petits échantillons”. In: *Revue de l’Institut International de Statistique* 11.3/4 (1943). Crossref-verified. The earliest of the three independent statements of the information inequality, ahead of Rao:1945 and of Cramer’s 1946 book; the treatise keeps the customary name Cramer-Rao and records the priority here. Crossref lists the author twice, once accented and once not, which is a deposit artefact and not two people., pp. 182–205. DOI: 10.2307/1401114.

[871] I. Gelfand and M. Neumark. “On the imbedding of normed rings into the ring of operators in Hilbert space”. In: *Matematicheskii Sbornik* 12(54) (1943), pp. 197–213.

[872] W. Heisenberg. “Die "beobachtbaren Größen" in der Theorie der Elementarteilchen”. In: *Zeitschrift für Physik* 120.7-10 (1943), pp. 513–538. DOI: 10.1007/BF01329800.

[873] G. Racah. “Theory of Complex Spectra. III”. In: *Physical Review* 63.9-10 (1943). Crossref-verified. Third instalment of the series; this is the paper that introduces the coefficient for the recoupling of four angular momenta (the modern 9j symbol). Its companion Racah:1942, cited in the same sentence, is ’Theory of Complex Spectra. II’, Phys. Rev. 62, 438-462 (1942), doi 10.1103/PhysRev.62.438, where the W (6j) coefficient appears., pp. 367–382. DOI: 10.1103/PhysRev.63.367.

[874] R. L. Goodstein. “On the restricted ordinal theorem”. In: *The Journal of Symbolic Logic* 9.2 (1944). Proves what is now called Goodstein’s theorem; its unprovability in PA is Kirby–Paris 1982, pp. 33–41. DOI: 10.2307/2268019.

[875] L. Onsager. “Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition”. In: *Physical Review* 65.3-4 (1944), pp. 117–149. DOI: 10.1103/PhysRev.65.117.

[876] V. Peshkov. “Second sound in helium II”. In: *Journal of Physics USSR* 8 (1944). No DOI., p. 381.

[877] S. O. Rice. “Mathematical analysis of random noise”. In: *Bell System Technical Journal* 23.3 (1944). Crossref-verified. Parts I (Shot Effect) and II (Power Spectra and Correlation Functions). This instalment does NOT contain the level-crossing formula: the expected number of upcrossings of a stationary process is Part III, sec. 3.3, in the companion Rice:1945. Cite Rice:1945 for the crossing count., pp. 282–332. DOI: 10.1002/j.1538-7305.1944.tb00874.x.

[878] H. B. G. Casimir. “On Onsager’s Principle of Microscopic Reversibility”. In: *Reviews of Modern Physics* 17.2-3 (1945), pp. 343–350. DOI: 10.1103/RevModPhys.17.343.

[879] E. A. Guggenheim. “The Principle of Corresponding States”. In: *The Journal of Chemical Physics* 13.7 (1945), pp. 253–261. DOI: 10.1063/1.1724033.

[880] H. C. van de Hulst. “Radiogolven uit het wereldruim. II. Herkomst der radiogolven”. In: *Nederlandsch Tijdschrift voor Natuurkunde* 11 (1945). The wartime 21 cm prediction: presented to the Nederlandsche Astronomen Club, Leiden, 15 April 1944, published in the December 1945 issue (jaargang XI, no. 12). It is part II of a two-part article; part I, "Ontvangst der radiogolven", pp. 201-210, is by C. J. Bakker, which is why some catalogues cite the whole as Bakker & van de Hulst, Ned. Tijdschr. Natuurkd. 11, 201-221 (1945). Only pp. 210-221 are van de Hulst’s and only those contain the cited prediction. TITLE VARIANT: the Dictionary of Scientific Biography bibliography prints the series title as "Radiostraling uit het wereldruim"; the radio-astronomy literature and the Dutch sources agree on "Radiogolven", which is also the form consistent with both part titles, so that is what I give. No DOI exists (do not invent one); the retrievable access route for a reader is the English translation, paper 34 in W. T. Sullivan III (ed.), Classics in Radio Astronomy (Reidel, Dordrecht, 1982), pp. 302-316 — worth recording in the bib note or the library manifest, since the Dutch original is otherwise unverifiable., pp. 210–221.

[881] C. Møller. “General Properties of the Characteristic Matrix in the Theory of Elementary Particles. I”. In: *Det Kongelige Danske Videnskabernes Selskab, Matematisk-fysiske Meddelelser* 23 (1945). Issue no. 1 of volume 23. Danish academy communications; unlikely to be in Crossref, no DOI supplied., pp. 1–48.

[882] C. R. Rao. “Information and the accuracy attainable in the estimation of statistical parameters”. In: *Bulletin of the Calcutta Mathematical Society* 37 (1945). The information inequality, independently obtained by Frechet:1943 and by Cramer in 1946; the treatise states it as the Cramer-Rao bound and records the joint attribution. The 1945 original is not deposited with Crossref (the Bulletin of the Calcutta Mathematical Society has no DOIs for that period), so no DOI is given here rather than an invented one. It is reprinted, with an introduction by P. K. Pathak, in Breakthroughs in Statistics vol. I, Springer Series in Statistics (1992), pp. 235–247, doi 10.1007/978-1-4612-0919-5_16, which is the practical route to the text and was verified against the API., pp. 81–91.

[883] S. O. Rice. “Mathematical analysis of random noise”. In: *Bell System Technical Journal* 24.1 (1945). Crossref-verified (resolved record read: Rice, S. O., Bell System Technical Journal 24(1) 46–156, 1945). Parts III (Statistical Properties of Random Noise Currents) and IV of the paper begun in Rice:1944. Section 3.3 carries the level-crossing formula for the expected number of zeros and upcrossings per second, which is what fixes how the trials factor of the look-elsewhere effect scales with the threshold., pp. 46–156. DOI: 10.1002/j.1538-7305.1945.tb00453.x.

[884] J. A. Wheeler and R. P. Feynman. “Interaction with the Absorber as the Mechanism of Radiation”. In: *Reviews of Modern Physics* 17.2-3 (1945). Verified via Crossref (Rev. Mod. Phys. 17(2-3), Apr 1945). This is the absorber-theory paper the prose names. The 1949 sequel, "Classical Electrodynamics in Terms of Direct Interparticle Action", Rev. Mod. Phys. 21, 425 (1949), is a different work and would need its own key., pp. 157–181. DOI: 10.1103/RevModPhys.17.157.

[885] E. Zavoisky. “Spin-magnetic resonance in paramagnetics”. In: *Journal of Physics USSR* 9 (1945), pp. 211–216.

[886] E. L. Andronikashvili. “A direct observation of two kinds of motion in helium II”. In: *Journal of Physics USSR* 10 (1946). Russian original Zh. Eksp. Teor. Fiz. 16, 780 (1946); no DOI., pp. 201–206.

[887] F. Bloch, W. W. Hansen, and M. Packard. “Nuclear Induction”. In: *Physical Review* 69.3-4 (1946), pp. 127–127. DOI: 10.1103/PhysRev.69.127.

[888] R. H. Dicke. “The Measurement of Thermal Radiation at Microwave Frequencies”. In: *Review of Scientific Instruments* 17.7 (1946). The Dicke switched radiometer, pp. 268–275. DOI: 10.1063/1.1770483.

[889] A. S. Eddington. *Fundamental Theory*. Published posthumously. The attempted derivation of $1/\alpha$ as an exact integer, revised from 136 to 137 as measurements moved, is the standard cautionary tale against numerology. Cambridge: Cambridge University Press, 1946.

[890] V. L. Ginzburg and I. M. Frank. “Izluchenie ravnomerno dvizhushchegosya elektrona, voznikayushchee pri ego perekhode iz odnoi sredy v druguyu”. In: *Zhurnal eksperimentalnoi i teoreticheskoi fiziki* 16 (1946). The transition-radiation prediction. TWO THINGS THE BIB WRITER MUST KNOW. (1) The treatise builds with pdflatex, which cannot typeset Cyrillic – use the transliteration ’Izluchenie ravnomerno dvizhushchegosya elektrona, voznikayushchee pri ego perekhode iz odnoi sredy v druguyu’, venue ’Zhurnal Eksperimental’noi i Teoreticheskoi Fiziki (ZhETF)’, and put the Cyrillic in a note if wanted. (2) An English-language version was published FIRST: Ginzburg and Frank, ’Radiation of a uniformly moving electron due to its transition from one medium into another’, J. Phys. (USSR) 9, 353-362 (1945) – that is the retrievable one and the better entry if the chapter is willing to carry the 1945 dating; keeping the key Ginzburg:1946 while citing the 1945 English printing would misreport the year, so pick one and say so in the note. Pages: only the opening page 15 is attested by the sources I checked; the range is often quoted as 15-28 but I did not confirm it, so no range is given here. No DOI. Russian-language original: the title and journal are given here in scientific transliteration, because the treatise builds with pdflatex under T1 encoding, which cannot typeset Cyrillic., p. 15.

[891] L. D. Landau. “On the vibration of the electronic plasma”. In: *Journal of Physics (USSR)* 10 (1946). The collisionless (Landau) damping paper cited in the plasma chapter. Published in English in the Soviet Academy’s Journal of Physics (USSR), vol. 10 no. 1; the Russian original is Zh. Eksp. Teor. Fiz. 16, 574 (1946). Title variants exist in the literature (’On the vibrations of the electronic plasma’, used by the 1967 UFN reprint and by mathnet); the form above is the one carried by the J. Phys. USSR printing and by the ter Haar Collected Papers reprint (Pergamon, 1965, pp. 445–460). No DOI exists; the ter Haar volume or the UFN reprint is the practical access route for a reader., pp. 25–34.

[892] E. M. Lifshitz. “On the gravitational stability of the expanding universe”. In: *Journal of Physics (USSR)* 10 (1946). Pre-DOI original; access route recorded in the library manifest (Golden Oldie republication, Gen. Rel. Grav. 49, 18 (2017))., pp. 116–129.

[893] E. M. Purcell. “Spontaneous Emission Probabilities at Radio Frequencies”. In: *Physical Review* 69 (1946). Meeting abstract B10, printed under ’Proceedings of the American Physical Society’; the origin of the Purcell factor. No DOI of its own: the only DOI covering it, 10.1103/PhysRev.69.674.2, is titled ’Proceedings of the American Physical Society’ and would not match this title, so it is deliberately omitted., p. 681.

[894] E. M. Purcell, H. C. Torrey, and R. V. Pound. “Resonance Absorption by Nuclear Magnetic Moments in a Solid”. In: *Physical Review* 69.1-2 (1946). NMR detected by absorption in bulk paraffin; the counterpart of Bloch’s induction method., pp. 37–38. DOI: 10.1103/PhysRev.69.37.

[895] S. Tomonaga. “On a Relativistically Invariant Formulation of the Quantum Theory of Wave Fields*”. In: *Progress of Theoretical Physics* 1.2 (1946), pp. 27–42. DOI: 10.1143/PTP.1.27.

[896] J. Bardeen. “Surface States and Rectification at a Metal Semi-Conductor Contact”. In: *Physical Review* 71.10 (1947), pp. 717–727. DOI: 10.1103/PhysRev.71.717.

[897] H. A. Bethe. “The Electromagnetic Shift of Energy Levels”. In: *Physical Review* 72.4 (1947), pp. 339–341. DOI: 10.1103/PhysRev.72.339.

[898] N. N. Bogoliubov. “On the theory of superfluidity”. In: *Journal of Physics (USSR)* 11 (1947), pp. 23–32.

[899] M. Conversi, E. Pancini, and O. Piccioni. “On the Disintegration of Negative Mesons”. In: *Physical Review* 71.3 (1947), pp. 209–210. DOI: 10.1103/PhysRev.71.209.

[900] F. R. Elder, A. M. Gurewitsch, R. V. Langmuir, and H. C. Pollock. “Radiation from Electrons in a Synchrotron”. In: *Physical Review* 71.11 (1947). Crossref confirmed: Phys. Rev. vol. 71, pp. 829-830, 1 June 1947, all four authors. The DOI’s trailing ’.5’ is REAL – it disambiguates the fifth Letter to the Editor beginning on p. 829; 10.1103/PhysRev.71.829 alone does not resolve to this paper. Do not ’clean’ it. First observation of synchrotron radiation; cited in the chapter beside Schott:1912., pp. 829–830. DOI: 10.1103/PhysRev.71.829.5.

[901] W. E. Lamb and R. C. Retherford. “Fine Structure of the Hydrogen Atom by a Microwave Method”. In: *Physical Review* 72.3 (1947), pp. 241–243. DOI: 10.1103/PhysRev.72.241.

[902] C. M. G. LATTES, H. MUIRHEAD, G. P. S. OCCHIALINI, and C. F. POWELL. “PROCESSES INVOLVING CHARGED MESONS”. In: *Nature* 159.4047 (1947), pp. 694–697. DOI: 10.1038/159694a0.

[903] G. D. Rochester and C. C. Butler. “Evidence for the Existence of New Unstable Elementary Particles”. In: *Nature* 160 (1947), pp. 855–857.

[904] I. E. Segal. “Postulates for General Quantum Mechanics”. In: *The Annals of Mathematics* 48.4 (1947), p. 930. DOI: 10.2307/1969387.

[905] R. A. Alpher, H. Bethe, and G. Gamow. “The Origin of Chemical Elements”. In: *Physical Review* 73.7 (1948). The alpha-beta-gamma paper. Suffix a/b: two 1948 works by first author Alpher; b is the Alpher–Herman Nature CMB prediction., pp. 803–804. DOI: 10.1103/PhysRev.73.803.

[906] R. A. ALPHER and R. HERMAN. “Evolution of the Universe”. In: *Nature* 162.4124 (1948). Prediction of the relic radiation at a few kelvin. See suffix note on Alpher:1948a., pp. 774–775. DOI: 10.1038/162774b0.

[907] J. Bardeen and W. H. Brattain. “The Transistor, A Semi-Conductor Triode”. In: *Physical Review* 74.2 (1948), pp. 230–231. DOI: 10.1103/PhysRev.74.230.

[908] V. Bargmann and E. P. Wigner. “Group Theoretical Discussion of Relativistic Wave Equations”. In: *Proceedings of the National Academy of Sciences* 34.5 (1948), pp. 211–223. DOI: 10.1073/pnas.34.5.211.

[909] H. B. G. Casimir. “On the attraction between two perfectly conducting plates”. In: *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*. Vol. 51. Academy proceedings (Series B, Physical Sciences), issued in numbered volumes, so volume 51 is printed and not invented; no DOI exists for it. Treated as @article because it is a numbered periodical with volume and pages. 1948, pp. 793–795.

[910] H. B. G. Casimir and D. Polder. “The Influence of Retardation on the London-van der Waals Forces”. In: *Physical Review* 73.4 (1948). Retarded dispersion forces: Rˆ-7 between two atoms, zˆ-4 between an atom and a conducting wall., pp. 360–372. DOI: 10.1103/PhysRev.73.360.

[911] R. P. Feynman. “Space-Time Approach to Non-Relativistic Quantum Mechanics”. In: *Reviews of Modern Physics* 20.2 (1948), pp. 367–387. DOI: 10.1103/RevModPhys.20.367.

[912] H. M. Foley and P. Kusch. “On the Intrinsic Moment of the Electron”. In: *Physical Review* 73.4 (1948). The earlier of the two 1948 papers (received February 1948); the customary single citation Kusch–Foley 1948 refers to the later Phys. Rev. 74, 250., pp. 412–412. DOI: 10.1103/PhysRev.73.412.

[913] D. GABOR. “A New Microscopic Principle”. In: *Nature* 161.4098 (1948). Crossref-verified. Crossref records the title as ’A New Microscopic Principle’; the paper is very widely cited as ’A new microscope principle’, so the customary form is wrong and the Crossref form is used. Nobel Prize in Physics 1971. Cited for the holography contrast in the interpretation., pp. 777–778. DOI: 10.1038/161777a0.

[914] G. Gamow. “The Origin of Elements and the Separation of Galaxies”. In: *Physical Review* 74 (1948). Letter; the DOI carries the .2 suffix because the page is shared. Metadata confirmed against Crossref., pp. 505–506. DOI: 10.1103/PhysRev.74.505.2.

[915] C. G. Hempel and P. Oppenheim. “Studies in the Logic of Explanation”. In: *Philosophy of Science* 15.2 (1948). The deductive-nomological account of explanation: the explanandum is derived from laws together with stated initial conditions, pp. 135–175. DOI: 10.1086/286983.

[916] P. Kusch and H. M. Foley. “The Magnetic Moment of the Electron”. In: *Physical Review* 74.3 (1948), pp. 250–263. DOI: 10.1103/PhysRev.74.250.

[917] L. D. Landau. “On the angular momentum of a system of two photons”. In: *Doklady Akademii Nauk SSSR* 60 (1948), pp. 207–209.

[918] M. L. Néel. “Propriétés magnétiques des ferrites ; ferrimagnétisme et antiferromagnétisme”. In: *Annales de Physique* 12.3 (1948). DOI not verified; omitted deliberately, pp. 137–198. DOI: 10.1051/anphys/194812030137.

[919] J. Schwinger. “On Quantum-Electrodynamics and the Magnetic Moment of the Electron”. In: *Physical Review* 73.4 (1948), pp. 416–417. DOI: 10.1103/PhysRev.73.416.

[920] J. Schwinger. “Quantum Electrodynamics. I. A Covariant Formulation”. In: *Physical Review* 74.10 (1948), pp. 1439–1461. DOI: 10.1103/PhysRev.74.1439.

[921] C. E. Shannon. “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[922] E. P. Wigner. “On the Behavior of Cross Sections Near Thresholds”. In: *Physical Review* 73.9 (1948), pp. 1002–1009. DOI: 10.1103/PhysRev.73.1002.

[923] E. O. Wollan and C. G. Shull. “The Diffraction of Neutrons by Crystalline Powders”. In: *Physical Review* 73.8 (1948). Issue 8. Crossref-verified. The Oak Ridge graphite-reactor powder-diffractometer method paper (monochromating crystal, Soller collimators, BF3 counter on a rotating arm) that the Apparatus subsection of exp:nscat-shull-mno describes; added to the existing cite group there (line 158). Wollan is first author here, Shull second, so the key is correctly Wollan:YYYY and does not collide with Shull:1949/Shull:1951., pp. 830–841. DOI: 10.1103/PhysRev.73.830.

[924] H. A. Bethe. “Theory of the Effective Range in Nuclear Scattering”. In: *Physical Review* 76.1 (1949), pp. 38–50. DOI: 10.1103/PhysRev.76.38.

[925] D. Bohm and E. P. Gross. “Theory of Plasma Oscillations. A. Origin of Medium-Like Behavior”. In: *Physical Review* 75.12 (1949). The Bohm–Gross dispersion relation cited in the chapter is in part A. Part B (’Excitation and Damping of Oscillations’, Phys. Rev. 75, 1864–1876, DOI 10.1103/PhysRev.75.1864) is a separate entry – if the damping discussion is ever cited, it needs key Bohm:1949b. Crossref-verified., pp. 1851–1864. DOI: 10.1103/PhysRev.75.1851.

[926] D. Crowfoot, C. W. Bunn, B. W. Rogers-Low, and A. Turner-Jones. “The X-Ray Crystallographic Investigation of the Structure of Penicillin”. In: *Chemistry of Penicillin*. Book chapter – the correct biblatex type is @incollection (the schema here has no such option; do NOT enter as @inproceedings without a booktitle). booktitle = The Chemistry of Penicillin, editors H. T. Clarke, J. R. Johnson and R. Robinson, Princeton University Press 1949, pp. 310-366. Crossref prints the title with its chapter numeral, ’XI. The X-Ray Crystallographic Investigation of the Structure of Penicillin’; a JSTOR record of the same chapter carries DOI 10.2307/j.ctt183px0j.14. Dorothy Crowfoot is Dorothy Crowfoot Hodgkin, publishing under her maiden name. 1949, pp. 310–366. DOI: 10.1515/9781400874910-012.

[927] F. J. Dyson. “The Radiation Theories of Tomonaga, Schwinger, and Feynman”. In: *Physical Review* 75.3 (1949). Crossref confirmed: Phys. Rev. vol. 75, pp. 486-502, 1 February 1949. This is the paper that introduces the time-ordered expansion now called the Dyson series, which is what the citing sentence claims, so it is the right one of the two 1949 Dyson papers. The companion – ’The S Matrix in Quantum Electrodynamics’, Phys. Rev. 75, 1736-1755 (1949), DOI 10.1103/PhysRev.75.1736 – is sometimes cited for the same material; if the QED chapter later needs it, key it Dyson:1949b and leave this one unsuffixed., pp. 486–502. DOI: 10.1103/PhysRev.75.486.

[928] F. J. Dyson. “TheSMatrix in Quantum Electrodynamics”. In: *Physical Review* 75.11 (1949), pp. 1736–1755. DOI: 10.1103/PhysRev.75.1736.

[929] E. Fermi. “On the Origin of the Cosmic Radiation”. In: *Physical Review* 75.8 (1949), pp. 1169–1174. DOI: 10.1103/PhysRev.75.1169.

[930] R. P. Feynman. “Space-Time Approach to Quantum Electrodynamics”. In: *Physical Review* 76.6 (1949), pp. 769–789. DOI: 10.1103/PhysRev.76.769.

[931] R. P. Feynman. “The Theory of Positrons”. In: *Physical Review* 76.6 (1949). Suffix a/b needed: two Feynman papers of 1949 are cited., pp. 749–759. DOI: 10.1103/PhysRev.76.749.

[932] D. Gabor. “Microscopy by reconstructed wave-fronts”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 197.1051 (1949). Issue 1051. The full theory plus the in-line mercury-arc experiments, including the reconstruction of the transparency bearing the names Huygens, Young and Fresnel, and the electron-microscope aberration-correction programme that motivated the work., pp. 454–487. DOI: 10.1098/rspa.1949.0075.

[933] O. Haxel, J. H. D. Jensen, and H. E. Suess. “On the "Magic Numbers" in Nuclear Structure”. In: *Physical Review* 75.11 (1949), pp. 1766–1766. DOI: 10.1103/physrev.75.1766.2.

[934] L. van Hove. “Quelques Propriétés Générales De L’intégrale De Configuration D’un Système De Particules Avec Interaction”. In: *Physica* 15.11-12 (1949), pp. 951–961. DOI: 10.1016/0031-8914(49)90059-2.

[935] M. Kac. “On distributions of certain Wiener functionals”. In: *Transactions of the American Mathematical Society* 65.1 (1949), pp. 1–13. DOI: 10.1090/S0002-9947-1949-0027960-X.

[936] A. Y. Khinchin. *Mathematical Foundations of Statistical Mechanics*. Translated by G. Gamow. New York: Dover Publications, 1949.

[937] N. M. Kroll and W. E. Lamb. “On the Self-Energy of a Bound Electron”. In: *Physical Review* 75.3 (1949), pp. 388–398. DOI: 10.1103/PhysRev.75.388.

[938] N. Levinson. “On the uniqueness of the potential in a Schrödinger equation for a given asymptotic phase”. In: *Kongelige Danske Videnskabernes Selskab, Matematisk-Fysiske Meddelelser* 25 (1949). Issue no. 9 of volume 25. Danish academy communications; no DOI supplied., pp. 1–29.

[939] S. Lundquist. “Experimental Investigations of Magneto-Hydrodynamic Waves”. In: *Physical Review* 76.12 (1949). Issue 12. Crossref-confirmed. The liquid-mercury Alfven-wave experiment the plasma chapter cites., pp. 1805–1809. DOI: 10.1103/PhysRev.76.1805.

[940] M. G. Mayer. “On Closed Shells in Nuclei. II”. In: *Physical Review* 75.12 (1949), pp. 1969–1970. DOI: 10.1103/PhysRev.75.1969.

[941] N. F. Mott. “The Basis of the Electron Theory of Metals, with Special Reference to the Transition Metals”. In: *Proceedings of the Physical Society. Section A* 62.7 (1949), pp. 416–422. DOI: 10.1088/0370-1298/62/7/303.

[942] T. D. Newton and E. P. Wigner. “Localized States for Elementary Systems”. In: *Reviews of Modern Physics* 21.3 (1949), pp. 400–406. DOI: 10.1103/RevModPhys.21.400.

[943] L. Onsager. “Statistical hydrodynamics”. In: *Il Nuovo Cimento* 6.S2 (1949). Supplement 2 to volume 6. Contains the remark that circulation in superfluid helium is quantized in units of h/m; the two 1949 Onsager papers are split a/b because neither sense is secondary., pp. 279–287. DOI: 10.1007/BF02780991.

[944] L. Onsager. “THE EFFECTS OF SHAPE ON THE INTERACTION OF COLLOIDAL PARTICLES”. In: *Annals of the New York Academy of Sciences* 51.4 (1949). Origin of the purely entropic, excluded-volume isotropic–nematic transition for hard rods., pp. 627–659. DOI: 10.1111/j.1749-6632.1949.tb27296.x.

[945] W. Pauli and F. Villars. “On the Invariant Regularization in Relativistic Quantum Theory”. In: *Reviews of Modern Physics* 21.3 (1949), pp. 434–444. DOI: 10.1103/RevModPhys.21.434.

[946] G. L. Pearson and J. Bardeen. “Electrical Properties of Pure Silicon and Silicon Alloys Containing Boron and Phosphorus”. In: *Physical Review* 75.5 (1949), pp. 865–883. DOI: 10.1103/PhysRev.75.865.

[947] C. Shannon. “Communication in the Presence of Noise”. In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21. DOI: 10.1109/JRPROC.1949.232969.

[948] W. Shockley. “The Theory ofp-nJunctions in Semiconductors andp-nJunction Transistors”. In: *Bell System Technical Journal* 28.3 (1949), pp. 435–489. DOI: 10.1002/j.1538-7305.1949.tb03645.x.

[949] C. G. Shull and J. S. Smart. “Detection of Antiferromagnetism by Neutron Diffraction”. In: *Physical Review* 76.8 (1949), pp. 1256–1257. DOI: 10.1103/physrev.76.1256.2.

[950] A. Wald. “Note on the consistency of the maximum likelihood estimate”. In: *The Annals of Mathematical Statistics* 20.4 (1949). Crossref-verified. Consistency of the maximum-likelihood estimator under conditions weaker than the classical ones. DISTINGUISH from Wald:1984 and Wald:1994, which are Robert M. Wald’s General Relativity and Quantum Field Theory in Curved Spacetime: a different Wald entirely., pp. 595–601. DOI: 10.1214/aoms/1177729952.

[951] H. Alfvén. *Cosmical Electrodynamics*. Cited as a standard plasma monograph alongside Spitzer:1962. First edition, Oxford, 1950, in the International Series of Monographs on Physics. A much-revised 2nd edition (1963) is co-authored with C.-G. Fälthammar – if the 2nd edition is ever meant, the author list and year both change, so keep this entry pinned to the 1950 first edition. Printed monograph: expect an AMBER availability grade. Clarendon Press, 1950.

[952] P. W. Anderson. “Antiferromagnetism. Theory of Superexchange Interaction”. In: *Physical Review* 79.2 (1950), pp. 350–356. DOI: 10.1103/PhysRev.79.350.

[953] K. Bleuler. “Eine neue Methode zur Behandlung der longitudinalen und skalaren Photonen”. In: *Helvetica Physica Acta* 23 (1950). No DOI; Helvetica Physica Acta is not indexed by Crossref for this period, pp. 567–586.

[954] E. Fermi. *Nuclear Physics: A Course Given by Enrico Fermi at the University of Chicago*. Lecture notes from Fermi’s 1949 Chicago course; this is where the transition-rate formula is christened ’Golden Rule No. 2’. A revised edition appeared in 1950/1951 – the citing prose only needs the naming, so the 1950 first printing is correct for the key. No DOI. notes compiled by Orear, Jay and Rosenfeld, A. H. and Schluter, R. A. Chicago: University of Chicago Press, 1950.

[955] L. L. Foldy and S. A. Wouthuysen. “On the Dirac Theory of Spin 1/2 Particles and Its Non-Relativistic Limit”. In: *Physical Review* 78.1 (1950), pp. 29–36. DOI: 10.1103/PhysRev.78.29.

[956] H. Fröhlich. “Theory of the Superconducting State. I. The Ground State at the Absolute Zero of Temperature”. In: *Physical Review* 79.5 (1950), pp. 845–856. DOI: 10.1103/PhysRev.79.845.

[957] V. L. Ginzburg and L. D. Landau. “On the theory of superconductivity”. In: *Zhurnal Eksperimental’noi i Teoreticheskoi Fiziki* 20 (1950). Russian original; no DOI. English translation in Collected Papers of L. D. Landau (1965)., pp. 1064–1082.

[958] S. N. Gupta. “Theory of Longitudinal Photons in Quantum Electrodynamics”. In: *Proceedings of the Physical Society. Section A* 63.7 (1950), pp. 681–691. DOI: 10.1088/0370-1298/63/7/301.

[959] E. L. Hahn. “Spin Echoes”. In: *Physical Review* 80.4 (1950), pp. 580–594. DOI: 10.1103/PhysRev.80.580.

[960] C. Hayashi. “Proton-Neutron Concentration Ratio in the Expanding Universe at the Stages preceding the Formation of the Elements”. In: *Progress of Theoretical Physics* 5 (1950). Metadata confirmed against Crossref., pp. 224–235. DOI: 10.1143/ptp/5.2.224.

[961] G. Herzberg. *Molecular Spectra and Molecular Structure. I. Spectra of Diatomic Molecules*. 2nd edition. Van Nostrand, 1950.

[962] J. Karle and H. Hauptman. “The phases and magnitudes of the structure factors”. In: *Acta Crystallographica* 3.3 (1950). Crossref-verified (issue 3). Keyed by first listed author Karle per house convention; the pair is Karle and Hauptman, Nobel 1985., pp. 181–187. DOI: 10.1107/S0365110X50000446.

[963] B. A. Lippmann and J. Schwinger. “Variational Principles for Scattering Processes. I”. In: *Physical Review* 79.3 (1950), pp. 469–480. DOI: 10.1103/PhysRev.79.469.

[964] E. Maxwell. “Isotope Effect in the Superconductivity of Mercury”. In: *Physical Review* 78.4 (1950), pp. 477–477. DOI: 10.1103/PhysRev.78.477.

[965] L. Michel. “Interaction between Four Half-Spin Particles and the Decay of the -Meson”. In: *Proceedings of the Physical Society. Section A* 63.5 (1950), pp. 514–531. DOI: 10.1088/0370-1298/63/5/311.

[966] E. M. Purcell and N. F. Ramsey. “On the Possibility of Electric Dipole Moments for Elementary Particles and Nuclei”. In: *Physical Review* 78.6 (1950), pp. 807–807. DOI: 10.1103/PhysRev.78.807.

[967] J. Rainwater. “Nuclear Energy Level Argument for a Spheroidal Nuclear Model”. In: *Physical Review* 79.3 (1950), pp. 432–434. DOI: 10.1103/PhysRev.79.432.

[968] N. F. Ramsey. “A Molecular Beam Resonance Method with Separated Oscillating Fields”. In: *Physical Review* 78.6 (1950), pp. 695–699. DOI: 10.1103/PhysRev.78.695.

[969] C. A. Reynolds, B. Serin, W. H. Wright, and L. B. Nesbitt. “Superconductivity of Isotopes of Mercury”. In: *Physical Review* 78.4 (1950), pp. 487–487. DOI: 10.1103/PhysRev.78.487.

[970] M. N. Rosenbluth. “High Energy Elastic Scattering of Electrons on Protons”. In: *Physical Review* 79.4 (1950), pp. 615–619. DOI: 10.1103/PhysRev.79.615.

[971] L. Schwartz. *Théorie des distributions*. Tome I 1950, Tome II 1951 (Actualités scientifiques et industrielles). Cited from the Fourier, PDE and Hilbert-space chapters. Paris: Hermann, 1950.

[972] W. Shockley. *Electrons and Holes in Semiconductors, with Applications to Transistor Electronics*. D. Van Nostrand, 1950.

[973] J. C. Ward. “An Identity in Quantum Electrodynamics”. In: *Physical Review* 78.2 (1950). One-page letter, pp. 182–182. DOI: 10.1103/PhysRev.78.182.

[974] G. C. Wick. “The Evaluation of the Collision Matrix”. In: *Physical Review* 80.2 (1950), pp. 268–272. DOI: 10.1103/PhysRev.80.268.

[975] C. N. Yang. “Selection Rules for the Dematerialization of a Particle into Two Photons”. In: *Physical Review* 77.2 (1950), pp. 242–245. DOI: 10.1103/PhysRev.77.242.

[976] L. Biermann. “Kometenschweife und solare Korpuskularstrahlung”. In: *Zeitschrift für Astrophysik* 29 (1951). Biermann’s inference of a continuous solar corpuscular stream from comet ion tails – the claim in the citing sentence. Zeitschrift für Astrophysik ceased in 1968 and has no Crossref DOIs; the retrieval route is the ADS scanned article, bibcode 1951ZA.....29..274B. Do not invent a DOI for it., pp. 274–286.

[977] J. M. BIJVOET, A. F. PEERDEMAN, and A. J. van BOMMEL. “Determination of the Absolute Configuration of Optically Active Compounds by Means of X-Rays”. In: *Nature* 168.4268 (1951). Crossref-verified. Cited for the measurable breakdown of Friedel’s law near an absorption edge, which is the experimental consequence of the derivation in phen:phase-patterson., pp. 271–272. DOI: 10.1038/168271a0.

[978] D. Bohm. *Quantum Theory*. The spin-1/2 singlet reformulation of the EPR argument is in ch. 22. New York: Prentice-Hall, 1951.

[979] H. B. Callen and T. A. Welton. “Irreversibility and Generalized Noise”. In: *Physical Review* 83.1 (1951), pp. 34–40. DOI: 10.1103/PhysRev.83.34.

[980] M. Deutsch. “Evidence for the Formation of Positronium in Gases”. In: *Physical Review* 82.3 (1951), pp. 455–456. DOI: 10.1103/PhysRev.82.455.

[981] H. I. Ewen and E. M. Purcell. “Observation of a Line in the Galactic Radio Spectrum: Radiation from Galactic Hydrogen at 1,420 Mc./sec”. In: *Nature* 168.4270 (1951), pp. 356–356. DOI: 10.1038/168356a0.

[982] I. M. Gel’fand and B. M. Levitan. “On the determination of a differential equation from its spectral function”. In: *Izvestiya Akademii Nauk SSSR, Seriya Matematicheskaya* 15 (1951). Russian original; the AMS Translations version (ser. 2, vol. 1, 1955, pp. 253-304) is the accessible route., pp. 309–360.

[983] J. R. Haynes and W. Shockley. “The Mobility and Life of Injected Holes and Electrons in Germanium”. In: *Physical Review* 81.5 (1951), pp. 835–843. DOI: 10.1103/PhysRev.81.835.

[984] C. A. MULLER and J. H. OORT. “Observation of a Line in the Galactic Radio Spectrum: The Interstellar Hydrogen Line at 1,420 Mc./sec., and an Estimate of Galactic Rotation”. In: *Nature* 168.4270 (1951). Nature issue 4270, 1 September 1951. The Dutch detection of the 21 cm line, companion to Ewen & Purcell in the same issue. ’Muller’ has no umlaut and no diaeresis – C. A. Muller is Dutch, not German. Nature prints the authors in caps (MULLER, OORT); the correct casing is given here., pp. 357–358. DOI: 10.1038/168357a0.

[985] W. K. H. Panofsky, R. L. Aamodt, and J. Hadley. “The Gamma-Ray Spectrum Resulting from Capture of Negativeπ-Mesons in Hydrogen and Deuterium”. In: *Physical Review* 81.4 (1951), pp. 565–574. DOI: 10.1103/PhysRev.81.565.

[986] E. M. Purcell and R. V. Pound. “A Nuclear Spin System at Negative Temperature”. In: *Physical Review* 81.2 (1951), pp. 279–280. DOI: 10.1103/PhysRev.81.279.

[987] W. V. Quine. “Two Dogmas of Empiricism”. In: *The Philosophical Review* 60.1 (1951). Reprinted with revisions in From a Logical Point of View (Harvard, 1953), pp. 20–43. DOI: 10.2307/2181906.

[988] J. Schwinger. “On Gauge Invariance and Vacuum Polarization”. In: *Physical Review* 82.5 (1951). Source of the critical field for spontaneous pair creation, which the Dirac-equation chapter labels an unconfirmed prediction., pp. 664–679. DOI: 10.1103/PhysRev.82.664.

[989] J. Schwinger. “On the Green’s functions of quantized fields. I”. In: *Proceedings of the National Academy of Sciences* 37.7 (1951). Suffixed to avoid a collision: 10-relativistic-qm/02-dirac-equation.tex uses Schwinger:1951 for ’On Gauge Invariance and Vacuum Polarization’, Phys. Rev. 82, 664-679, doi 10.1103/PhysRev.82.664. This entry is the source/generating-functional paper., pp. 452–455. DOI: 10.1073/pnas.37.7.452.

[990] W. Shockley, M. Sparks, and G. K. Teal. “p−nJunction Transistors”. In: *Physical Review* 83.1 (1951), pp. 151–162. DOI: 10.1103/PhysRev.83.151.

[991] C. G. Shull, W. A. Strauser, and E. O. Wollan. “Neutron Diffraction by Paramagnetic and Antiferromagnetic Substances”. In: *Physical Review* 83.2 (1951), pp. 333–345. DOI: 10.1103/PhysRev.83.333.

[992] A. Tarski. *A Decision Method for Elementary Algebra and Geometry*. Report R-109. Prepared for publication by J. C. C. McKinsey. First issued 1 August 1948, revised May 1951; the freely available RAND scan is the second printing of 1957. Also published by the University of California Press, Berkeley and Los Angeles, 1951. Santa Monica, CA: RAND Corporation, 1951.

[993] C. Zener. “Interaction between thed-Shells in the Transition Metals. II. Ferromagnetic Compounds of Manganese with Perovskite Structure”. In: *Physical Review* 82.3 (1951), pp. 403–405. DOI: 10.1103/PhysRev.82.403.

[994] H. L. Anderson, E. Fermi, E. A. Long, and D. E. Nagle. “Total Cross Sections of Positive Pions in Hydrogen”. In: *Physical Review* 85.5 (1952), pp. 936–936. DOI: 10.1103/PhysRev.85.936.

[995] D. Bohm. “A Suggested Interpretation of the Quantum Theory in Terms of "Hidden" Variables. I”. In: *Physical Review* 85.2 (1952), pp. 166–179. DOI: 10.1103/PhysRev.85.166.

[996] D. Bohm. “A Suggested Interpretation of the Quantum Theory in Terms of "Hidden" Variables. II”. In: *Physical Review* 85.2 (1952), pp. 180–193. DOI: 10.1103/PhysRev.85.180.

[997] A. P. Calderón and A. Zygmund. “On the existence of certain singular integrals”. In: *Acta Mathematica* 88 (1952). Crossref-verified (resolved record read: Acta Math. 88 85–139, 1952). The Calderón–Zygmund theory of singular integral operators, which supplies the Hölder estimate for the Beurling transform on which the Korn–Lichtenstein theorem – and with it the reduction of an elliptic second-order operator to canonical form at merely Hölder coefficients – depends., pp. 85–139. DOI: 10.1007/BF02392130.

[998] F. J. Dyson. “Divergence of Perturbation Theory in Quantum Electrodynamics”. In: *Physical Review* 85.4 (1952), pp. 631–632. DOI: 10.1103/PhysRev.85.631.

[999] Y. Fourès-Bruhat. “Théorème d’existence pour certains systèmes d’équations aux dérivées partielles non linéaires”. In: *Acta Mathematica* 88 (1952). Published under the author’s earlier name Foures-Bruhat; keyed by her later universal name Choquet-Bruhat – the citing chapter states the printed name. Re-key to the printed surname if strict surname-at-publication is preferred., pp. 141–225.

[1000] D. A. Glaser. “Some Effects of Ionizing Radiation on the Formation of Bubbles in Liquids”. In: *Physical Review* 87.4 (1952). A one-page Letter to the Editor (Crossref gives the page range as 665-665), issue of 15 August 1952: radiation-induced nucleation in superheated diethyl ether, the origin of the bubble chamber (Nobel Prize 1960). NOTE for the chapter prose: this letter reports the shortening of the boiling delay by a radiation source; the photographs of bubble tracks came in a following paper, and the chapter is worded accordingly. Verified against api.crossref.org 2026-08-13., pp. 665–665. DOI: 10.1103/PhysRev.87.665.

[1001] R. N. Hall. “Electron-Hole Recombination in Germanium”. In: *Physical Review* 87.2 (1952), pp. 387–387. DOI: 10.1103/PhysRev.87.387.

[1002] E. Hopf. “A remark on linear elliptic differential equations of second order”. In: *Proceedings of the American Mathematical Society* 3.5 (1952). Crossref-verified (resolved record read: Hopf, Eberhard, Proc. Amer. Math. Soc. 3(5) 791–793, 1952). The boundary-point lemma – that the outward normal derivative is strictly positive at a boundary maximum – which is a different and later result from the strong maximum principle of Hopf:1927, and is not contained in that paper. Obtained independently by O. A. Oleinik the same year (Mat. Sbornik 30(72) (1952) 695–702)., pp. 791–793. DOI: 10.1090/S0002-9939-1952-0050126-X.

[1003] G. Källén. “On the definition of the renormalization constants in quantum electrodynamics”. In: *Helvetica Physica Acta* 25 (1952), pp. 417–434.

[1004] T. D. Lee and C. N. Yang. “Statistical Theory of Equations of State and Phase Transitions. II. Lattice Gas and Ising Model”. In: *Physical Review* 87.3 (1952), pp. 410–419. DOI: 10.1103/PhysRev.87.410.

[1005] G. W. Mackey. “Induced Representations of Locally Compact Groups I”. In: *Annals of Mathematics* 55.1 (1952). Crossref-verified (resolved record read: Ann. of Math. 55(1) 101, 1952; printed range 101–139). The systematic theory of induced representations, of which Wigner:1939 is the case that classifies the unitary representations of the Poincare group; it supplies the irreducibility and exhaustiveness statements that turn a little group into a complete set of labels., pp. 101–139. DOI: 10.2307/1969423.

[1006] S. P. W. Merrill. “Spectroscopic Observations of Stars of Class”. In: *The Astrophysical Journal* 116 (1952). Technetium in S stars; also reported in Science 115, 484 (1952), p. 21. DOI: 10.1086/145589.

[1007] L. Mestel. “On the Theory of White Dwarf Stars: I. The Energy Sources of White Dwarfs”. In: *Monthly Notices of the Royal Astronomical Society* 112.6 (1952), pp. 583–597. DOI: 10.1093/mnras/112.6.583.

[1008] L. Onsager. “Interpretation of the de Haas-van Alphen effect”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 43.344 (1952), pp. 1006–1008. DOI: 10.1080/14786440908521019.

[1009] G. W. Rodeback and J. S. Allen. “Neutrino Recoils Following the Capture of Orbital Electrons inA37”. In: *Physical Review* 86.4 (1952), pp. 446–450. DOI: 10.1103/PhysRev.86.446.

[1010] E. E. Salpeter. “Nuclear Reactions in Stars Without Hydrogen”. In: *The Astrophysical Journal* 115 (1952). Triple-alpha process, p. 326. DOI: 10.1086/145546.

[1011] D. Sayre. “The squaring method: a new method for phase determination”. In: *Acta Crystallographica* 5.1 (1952). Crossref-verified. Source of Sayre’s equation, eq:phase-sayre., pp. 60–65. DOI: 10.1107/S0365110X52000137.

[1012] W. Shockley and W. T. Read. “Statistics of the Recombinations of Holes and Electrons”. In: *Physical Review* 87.5 (1952), pp. 835–842. DOI: 10.1103/PhysRev.87.835.

[1013] C. N. Yang. “The Spontaneous Magnetization of a Two-Dimensional Ising Model”. In: *Physical Review* 85.5 (1952), pp. 808–816. DOI: 10.1103/PhysRev.85.808.

[1014] C. N. Yang and T. D. Lee. “Statistical Theory of Equations of State and Phase Transitions. I. Theory of Condensation”. In: *Physical Review* 87.3 (1952), pp. 404–409. DOI: 10.1103/PhysRev.87.404.

[1015] A. Bohr and B. R. Mottelson. “Collective and Individual-Particle Aspects of Nuclear Structure”. In: *Matematisk-fysiske Meddelelser Kongelige Danske Videnskabernes Selskab* 27 (1953). Number 16, pp. 1–174.

[1016] R. Dalitz. “On the analysis of τ-meson data and the nature of the τ-meson”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 44.357 (1953). Series 7, volume 44, issue 357, pp. 1068–1080. DOI: 10.1080/14786441008520365.

[1017] M. Gell-Mann. “Isotopic Spin and New Unstable Particles”. In: *Physical Review* 92.3 (1953). Issue 3, November 1953. Introduces strangeness, independently of Nakano and Nishijima. Crossref record verified., pp. 833–834. DOI: 10.1103/PhysRev.92.833.

[1018] M. Gell-Mann and M. L. Goldberger. “The Formal Theory of Scattering”. In: *Physical Review* 91.2 (1953). Issue 2, July 1953 — earlier in the year than GellMann:1953b, hence the a suffix. Crossref record verified., pp. 398–408. DOI: 10.1103/PhysRev.91.398.

[1019] H. S. Green. “A Generalized Method of Field Quantization”. In: *Physical Review* 90.2 (1953). Green:1955 already exists in the bib (a different author, on the Q factor); no collision., pp. 270–273. DOI: 10.1103/PhysRev.90.270.

[1020] H. Hauptman and J. Karle. *Solution of the Phase Problem I. The Centrosymmetric Crystal*. ACA Monograph 3. The monograph introducing the direct methods that later won the 1985 Nobel Prize in Chemistry. A monograph in a society series, with no DOI; identified by publisher and series number. Ann Arbor: American Crystallographic Association, 1953.

[1021] E. Inönü and E. P. Wigner. “On the Contraction of Groups and Their Representations”. In: *Proceedings of the National Academy of Sciences* 39.6 (1953), pp. 510–524. DOI: 10.1073/pnas.39.6.510.

[1022] E. J. Konopinski and H. M. Mahmoud. “The Universal Fermi Interaction”. In: *Physical Review* 92.4 (1953), pp. 1045–1049. DOI: 10.1103/PhysRev.92.1045.

[1023] P. M. Morse and H. Feshbach. *Methods of Theoretical Physics*. Two parts. Resolves the second TODO-cite of the PDE chapter. Author is Philip M. Morse – distinct from Marston Morse of Morse:1934. New York: McGraw-Hill, 1953.

[1024] T. Nakano and K. Nishijima. “Charge Independence forV-particles”. In: *Progress of Theoretical Physics* 10.5 (1953), pp. 581–582. DOI: 10.1143/PTP.10.581.

[1025] A. B. Pippard and W. L. Bragg. “An experimental and theoretical study of the relation between magnetic field and current in a superconductor”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 216.1127 (1953), pp. 547–568. DOI: 10.1098/rspa.1953.0040.

[1026] F. Reines and C. L. Cowan. “Detection of the Free Neutrino”. In: *Physical Review* 92.3 (1953), pp. 830–831. DOI: 10.1103/PhysRev.92.830.

[1027] L. Spitzer and R. Härm. “Transport Phenomena in a Completely Ionized Gas”. In: *Physical Review* 89.5 (1953). Crossref-verified. Source of the Spitzer resistivity and its T^3/2 scaling cited at sec:plasma-collisions. Co-author’s name carries an umlaut: Härm., pp. 977–981. DOI: 10.1103/PhysRev.89.977.

[1028] E. C. G. Stueckelberg and A. Petermann. “La normalisation des constantes dans la théorie des quanta”. In: *Helvetica Physica Acta* 26 (1953). Not a genuine collision: both candidates are this one paper, one with the accents stripped. Earliest statement of the renormalization group (the groupe de normalisation). No DOI; Helvetica Physica Acta scans are at e-periodica.ch., pp. 499–520.

[1029] V. Bargmann. “On Unitary Ray Representations of Continuous Groups”. In: *Annals of Mathematics* 59 (1954), pp. 1–46. DOI: 10.2307/1969831.

[1030] M. Born and K. Huang. *Dynamical Theory of Crystal Lattices*. Oxford University Press, 1954.

[1031] R. H. Brown and R. Twiss. “A new type of interferometer for use in radio astronomy”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 45.366 (1954), pp. 663–682. DOI: 10.1080/14786440708520475.

[1032] D. M. Chapin, C. S. Fuller, and G. L. Pearson. “A New Silicon p-n Junction Photocell for Converting Solar Radiation into Electrical Power”. In: *Journal of Applied Physics* 25.5 (1954), pp. 676–677. DOI: 10.1063/1.1721711.

[1033] H. Chernoff. “On the distribution of the likelihood ratio”. In: *The Annals of Mathematical Statistics* 25.3 (1954). Crossref-verified. The correction to Wilks:1938 when the parameter lies on the boundary of its range: the asymptotic law becomes a half-and-half mixture of a point mass at zero and a chi-squared. This is the case that matters whenever a signal strength cannot be negative., pp. 573–578. DOI: 10.1214/aoms/1177728725.

[1034] W. S. Corak, B. B. Goodman, C. B. Satterthwaite, and A. Wexler. “Exponential Temperature Dependence of the Electronic Specific Heat of Superconducting Vanadium”. In: *Physical Review* 96.5 (1954), pp. 1442–1444. DOI: 10.1103/physrev.96.1442.2.

[1035] S. Deser, M. L. Goldberger, K. Baumann, and W. Thirring. “Energy Level Displacements in Pi-Mesonic Atoms”. In: *Physical Review* 96.3 (1954), pp. 774–776. DOI: 10.1103/PhysRev.96.774.

[1036] H. Feshbach, C. E. Porter, and V. F. Weisskopf. “Model for Nuclear Reactions with Neutrons”. In: *Physical Review* 96.2 (1954), pp. 448–464. DOI: 10.1103/PhysRev.96.448.

[1037] S. E. Forbush. “World-wide cosmic ray variations, 1937–1952”. In: *Journal of Geophysical Research* 59.4 (1954). The eleven-year anticorrelation of cosmic-ray intensity with solar activity. Distinct from Forbush:1937, which is the transient magnetic-storm decrease. Metadata confirmed against Crossref., pp. 525–542. DOI: 10.1029/JZ059i004p00525.

[1038] M. Gell-Mann and F. E. Low. “Quantum Electrodynamics at Small Distances”. In: *Physical Review* 95.5 (1954), pp. 1300–1312. DOI: 10.1103/PhysRev.95.1300.

[1039] J. P. Gordon, H. J. Zeiger, and C. H. Townes. “Molecular Microwave Oscillator and New Hyperfine Structure in the Microwave Spectrum of NH3”. In: *Physical Review* 95.1 (1954), pp. 282–284. DOI: 10.1103/PhysRev.95.282.

[1040] D. W. Green, V. M. Ingram, and M. F. Perutz. “The structure of haemoglobin - IV. Sign determination by the isomorphous replacement method”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 225.1162 (1954). Crossref-verified. Key suffixed ’b’ because Green:1954 already exists in references.bib for Melville S. Green’s Markoff-process paper – a different author entirely; the existing key was NOT renamed. Keyed by first listed author Green per house convention although it is Perutz’s method. Cited in the chapter only for what it actually did: sign determination by isomorphous replacement in centrosymmetric projections. The general twofold phase ambiguity requiring two derivatives is stated in the prose without attributing it to this paper., pp. 287–307. DOI: 10.1098/rspa.1954.0203.

[1041] M. S. Green. “Markoff Random Processes and the Statistical Mechanics of Time-Dependent Phenomena. II. Irreversible Processes in Fluids”. In: *The Journal of Chemical Physics* 22.3 (1954), pp. 398–413. DOI: 10.1063/1.1740082.

[1042] S. N. Gupta. “Gravitation and Electromagnetism”. In: *Physical Review* 96.6 (1954), pp. 1683–1685. DOI: 10.1103/PhysRev.96.1683.

[1043] W. Heitler. *The Quantum Theory of Radiation*. 3rd edition. Oxford University Press, 1954.

[1044] F. Hoyle. “On Nuclear Reactions Occuring in Very Hot STARS.I. the Synthesis of Elements from Carbon to Nickel”. In: *The Astrophysical Journal Supplement Series* 1 (1954). Original printed title spells ’Occuring’; prediction of the 7.65 MeV state in carbon-12, p. 121. DOI: 10.1086/190005.

[1045] A. N. Kolmogorov. “On Conservation of Conditionally Periodic Motions for a Small Change in Hamilton’s Function”. In: *Doklady Akademii Nauk SSSR* 98 (1954). In Russian (’O sokhranenii uslovno-periodicheskikh dvizhenii pri malom izmenenii funktsii Gamil’tona’). English translation in Lecture Notes in Physics 93 (1979) 51-56. The K of KAM., pp. 527–530.

[1046] M. D. Kruskal and M. Schwarzschild. “Some instabilities of a completely ionized plasma”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 223.1154 (1954). Issue 1154. The Kruskal–Schwarzschild analysis of interchange/kink instabilities named in the citing sentence. Verified against Crossref. Distinct from the Kruskal–Shafranov safety-factor limit mentioned in the same passage, which has no single canonical paper., pp. 348–360. DOI: 10.1098/rspa.1954.0120.

[1047] L. D. Landau, A. A. Abrikosov, and I. M. Khalatnikov. “The removal of infinities in quantum electrodynamics”. In: *Doklady Akademii Nauk SSSR* 95 (1954). Russian original; no DOI. Source of the Landau pole, pp. 497–499.

[1048] H. Lehmann. “Über Eigenschaften von Ausbreitungsfunktionen und Renormierungskonstanten quantisierter Felder”. In: *Il Nuovo Cimento* 11.4 (1954), pp. 342–357. DOI: 10.1007/BF02783624.

[1049] G. Lüders. “On the equivalence of invariance under time reversal and under particle-antiparticle conjugation for relativistic field theories”. In: *Matematisk-fysiske Meddelelser Kongelige Danske Videnskabernes Selskab* 28 (1954). Number 5, pp. 1–17.

[1050] M. A. Ruderman and C. Kittel. “Indirect Exchange Coupling of Nuclear Magnetic Moments by Conduction Electrons”. In: *Physical Review* 96.1 (1954), pp. 99–102. DOI: 10.1103/PhysRev.96.99.

[1051] J. C. Slater and G. F. Koster. “Simplified LCAO Method for the Periodic Potential Problem”. In: *Physical Review* 94.6 (1954), pp. 1498–1524. DOI: 10.1103/PhysRev.94.1498.

[1052] L. Van Hove. “Correlations in Space and Time and Born Approximation Scattering in Systems of Interacting Particles”. In: *Physical Review* 95.1 (1954). Issue 1, 1 July 1954. Crossref-verified. The paper that defines the scattering law S(Q,omega) and the space-time correlation functions; placed at ’What is measured directly is a scattering law’ in sec:nscat-summary (line 750). NAME CAUTION: the publisher prints ’Léon Van Hove’ and Crossref’s family field is ’Van Hove’ (capital V), but the existing bib entry vanHove:1949 writes the author as ’van Hove, L.’. Pick one spelling for the author field so the two entries do not print as two different people; the key vanHove:1954 already follows the house convention and there is no key collision., pp. 249–262. DOI: 10.1103/PhysRev.95.249.

[1053] G. C. Wick. “Properties of Bethe-Salpeter Wave Functions”. In: *Physical Review* 96.4 (1954), pp. 1124–1134. DOI: 10.1103/PhysRev.96.1124.

[1054] C. N. Yang and R. L. Mills. “Conservation of Isotopic Spin and Isotopic Gauge Invariance”. In: *Physical Review* 96.1 (1954), pp. 191–195. DOI: 10.1103/PhysRev.96.191.

[1055] J. S. Bell. “Time reversal in field theory”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 231.1187 (1955), pp. 479–495. DOI: 10.1098/rspa.1955.0189.

[1056] B. N. Brockhouse and A. T. Stewart. “Scattering of Neutrons by Phonons in an Aluminum Single Crystal”. In: *Physical Review* 100.2 (1955), pp. 756–757. DOI: 10.1103/PhysRev.100.756.

[1057] O. Chamberlain, E. Segrè, C. Wiegand, and T. Ypsilantis. “Observation of Antiprotons”. In: *Physical Review* 100.3 (1955), pp. 947–950. DOI: 10.1103/PhysRev.100.947.

[1058] C. Chevalley. “Invariants of Finite Groups Generated by Reflections”. In: *American Journal of Mathematics* 77.4 (1955). Crossref-verified (resolved record read: Amer. J. Math. 77(4) 778, 1955; printed range 778–782). That the invariants of a finite reflection group form a free polynomial algebra on as many generators as the rank – the last step of Racah’s count of independent Casimir operators. The Chevalley RESTRICTION theorem, quoted alongside it, is a different result and is not in this paper., pp. 778–782. DOI: 10.2307/2372597.

[1059] W. Cochran. “Relations between the phases of structure factors”. In: *Acta Crystallographica* 8.8 (1955). Crossref-verified by bibliographic query. This is the non-centrosymmetric (von Mises) phase distribution, distinct from Cochran:1955a which treats signs; the two are cited for different statements and must not be merged., pp. 473–478. DOI: 10.1107/S0365110X55001485.

[1060] W. Cochran and M. M. Woolfson. “The theory of sign relations between structure factors”. In: *Acta Crystallographica* 8.1 (1955). Crossref-verified by bibliographic query. Suffixed ’a’ because Cochran:1955b (below) is a second 1955 Acta Cryst paper by the same first author. Cited for the tanh sign probability, eq:phase-tanh., pp. 1–12. DOI: 10.1107/S0365110X55000017.

[1061] W. S. Corak, M. P. Garfunkel, C. B. Satterthwaite, and A. Wexler. “Atomic Heats of Copper, Silver, and Gold from 1°K to 5°K”. In: *Physical Review* 98.6 (1955), pp. 1699–1707. DOI: 10.1103/PhysRev.98.1699.

[1062] R. Davis. “Attempt to Detect the Antineutrinos from a Nuclear Reactor by theCl37(ν^{-}, e−)A37Reaction”. In: *Physical Review* 97.3 (1955), pp. 766–769. DOI: 10.1103/PhysRev.97.766.

[1063] G. Dresselhaus, A. F. Kip, and C. Kittel. “Cyclotron Resonance of Electrons and Holes in Silicon and Germanium Crystals”. In: *Physical Review* 98.2 (1955), pp. 368–384. DOI: 10.1103/PhysRev.98.368.

[1064] L. ESSEN and J. V. L. PARRY. “An Atomic Standard of Frequency and Time Interval: A Cæsium Resonator”. In: *Nature* 176.4476 (1955), pp. 280–282. DOI: 10.1038/176280a0.

[1065] E. Fermi, J. Pasta, and S. Ulam. *Studies of Non Linear Problems*. Mary Tsingou performed the computations and is now customarily credited (FPUT). 1955.

[1066] R. P. Feynman. “Chapter II Application of Quantum Mechanics to Liquid Helium”. In: *Progress in Low Temperature Physics*. Vol. 1. 1955, pp. 17–53. DOI: 10.1016/S0079-6417(08)60077-3.

[1067] M. Gell-Mann and A. Pais. “Behavior of Neutral Particles under Charge Conjugation”. In: *Physical Review* 97.5 (1955), pp. 1387–1389. DOI: 10.1103/PhysRev.97.1387.

[1068] R. J. Glauber. “Cross Sections in Deuterium at High Energies”. In: *Physical Review* 100.1 (1955), pp. 242–248. DOI: 10.1103/PhysRev.100.242.

[1069] J. B. Goodenough. “Theory of the Role of Covalence in the Perovskite-Type Manganites[La, M(II)]MnO3”. In: *Physical Review* 100.2 (1955), pp. 564–573. DOI: 10.1103/PhysRev.100.564.

[1070] J. P. Gordon, H. J. Zeiger, and C. H. Townes. “The Maser—New Type of Microwave Amplifier, Frequency Standard, and Spectrometer”. In: *Physical Review* 99.4 (1955), pp. 1264–1274. DOI: 10.1103/PhysRev.99.1264.

[1071] E. I. Green. “The Story of Q”. In: *American Scientist* 43 (1955). History and engineering origin of the quality factor Q., pp. 584–594.

[1072] R. Haag. “On quantum field theories”. In: *Kongelige Danske Videnskabernes Selskab, Matematisk-fysiske Meddelelser* 29 (1955). Issue no. 12; no DOI. Source of Haag’s theorem, pp. 1–37.

[1073] H. Lehmann, K. Symanzik, and W. Zimmermann. “Zur Formulierung quantisierter Feldtheorien”. In: *Il Nuovo Cimento* 1.1 (1955), pp. 205–225. DOI: 10.1007/BF02731765.

[1074] M. H. Löb. “Solution of a Problem of Leon Henkin”. In: *Journal of Symbolic Logic* 20.2 (1955), pp. 115–118. DOI: 10.2307/2266895.

[1075] P.-O. Löwdin. “Quantum Theory of Many-Particle Systems. I. Physical Interpretations by Means of Density Matrices, Natural Spin-Orbitals, and Convergence Problems in the Method of Configurational Interaction”. In: *Physical Review* 97.6 (1955). Key transliterates the umlaut (Loewdin) to stay ASCII., pp. 1474–1489. DOI: 10.1103/PhysRev.97.1474.

[1076] R. C. Miller and P. Kusch. “Velocity Distributions in Potassium and Thallium Atomic Beams”. In: *Physical Review* 99.4 (1955). Direct measurement of the Maxwell speed distribution, pp. 1314–1321. DOI: 10.1103/PhysRev.99.1314.

[1077] A. Pais and O. Piccioni. “Note on the Decay and Absorption of theθ0”. In: *Physical Review* 100.5 (1955), pp. 1487–1489. DOI: 10.1103/PhysRev.100.1487.

[1078] E. N. Parker. “Hydromagnetic Dynamo Models”. In: *The Astrophysical Journal* 122 (1955). September 1955. The cyclonic-turbulence (alpha-effect) mean-field dynamo paper the citing sentence needs. Crossref prints a trailing period in the title (’Hydromagnetic Dynamo Models.’); drop it., p. 293. DOI: 10.1086/146087.

[1079] W. Pauli. “Exclusion principle, Lorentz group and reflection of space-time and charge”. In: *Niels Bohr and the Development of Physics*. 1955, pp. 30–51.

[1080] W. F. Stinespring. “Positive functions on C^*-algebras”. In: *Proceedings of the American Mathematical Society* 6.2 (1955), pp. 211–216. DOI: 10.1090/S0002-9939-1955-0069403-4.

[1081] E. P. Wigner. “Lower Limit for the Energy Derivative of the Scattering Phase Shift”. In: *Physical Review* 98.1 (1955), pp. 145–147. DOI: 10.1103/PhysRev.98.145.

[1082] M. J. BLOCK. “Surface Tension as the Cause of Bénard Cells and Surface Deformation in a Liquid Film”. In: *Nature* 178.4534 (1956). Issue 4534. Crossref prints the author as ’MYRON J. BLOCK’ (Nature’s period house style, all caps); the surname is Block. The experimental half of the pair showing Bénard’s cells are surface-tension driven., pp. 650–651. DOI: 10.1038/178650a0.

[1083] R. H. BROWN and R. Q. TWISS. “Correlation between Photons in two Coherent Beams of Light”. In: *Nature* 177.4497 (1956), pp. 27–29. DOI: 10.1038/177027a0.

[1084] L. N. Cooper. “Bound Electron Pairs in a Degenerate Fermi Gas”. In: *Physical Review* 104.4 (1956), pp. 1189–1190. DOI: 10.1103/PhysRev.104.1189.

[1085] C. L. Cowan, F. Reines, F. B. Harrison, H. W. Kruse, and A. D. McGuire. “Detection of the Free Neutrino: a Confirmation”. In: *Science* 124.3212 (1956), pp. 103–104. DOI: 10.1126/science.124.3212.103.

[1086] R. HANBURY BROWN and R. Q. TWISS. “A Test of a New Type of Stellar Interferometer on Sirius”. In: *Nature* 178.4541 (1956), pp. 1046–1048. DOI: 10.1038/1781046a0.

[1087] D. C. HODGKIN, J. KAMPER, M. MACKAY, J. PICKWORTH, K. N. TRUEBLOOD, and J. G. WHITE. “Structure of Vitamin B12”. In: *Nature* 178.4524 (1956). Crossref-verified: title, six family names in this order, Nature 178, 64-66, 1956. Given-name initials are reconstructed and should be confirmed against the paper; the family names and their order are from Crossref. Typeset the title as ’Structure of Vitamin B_12’ in the bib so the subscript prints. A related 1955 paper exists (Nature 176, 325-328) – distinct work., pp. 64–66. DOI: 10.1038/178064a0.

[1088] R. Hofstadter. “Electron Scattering and Nuclear Structure”. In: *Reviews of Modern Physics* 28.3 (1956), pp. 214–254. DOI: 10.1103/RevModPhys.28.214.

[1089] J. Karle and H. Hauptman. “A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P12, 3P22”. In: *Acta Crystallographica* 9.8 (1956). Crossref-verified. Crossref stores the last two space-group symbols with Unicode subscript characters; written here with TeX math subscripts so the entry is ASCII and survives pdflatex + T1. The non-centrosymmetric extension of the direct methods, which is what the biological molecules needed., pp. 635–651. DOI: 10.1107/S0365110X56001741.

[1090] K. Lande, E. T. Booth, J. Impeduglia, L. M. Lederman, and W. Chinowsky. “Observation of Long-Lived NeutralVParticles”. In: *Physical Review* 103.6 (1956), pp. 1901–1904. DOI: 10.1103/PhysRev.103.1901.

[1091] T. D. Lee and C. N. Yang. “Question of Parity Conservation in Weak Interactions”. In: *Physical Review* 104.1 (1956), pp. 254–258. DOI: 10.1103/PhysRev.104.254.

[1092] I. M. Lifshitz and A. M. Kosevich. “Theory of Magnetic Susceptibility in Metals at Low Temperatures”. In: *Soviet Physics JETP* 2 (1956). Russian original: Zh. Eksp. Teor. Fiz. 29, 730 (1955)., pp. 636–645.

[1093] S. Pancharatnam. “Generalized theory of interference, and its applications”. In: *Proceedings of the Indian Academy of Sciences - Section A* 44.5 (1956). Issue 5, November 1956. The title as printed on the paper carries the series designation ’Part I. Coherent pencils’; Crossref registers only the stem given above, so the DOI title check will match the stem. If the bib writer wants the full printed form, write ’Generalized theory of interference, and its applications. Part I. Coherent pencils’ and expect the automated title comparison to be looser., pp. 247–262. DOI: 10.1007/BF03046050.

[1094] O. Penrose and L. Onsager. “Bose-Einstein Condensation and Liquid Helium”. In: *Physical Review* 104.3 (1956), pp. 576–584. DOI: 10.1103/PhysRev.104.576.

[1095] E. M. PURCELL. “The Question of Correlation between Photons in Coherent Light Rays”. In: *Nature* 178.4548 (1956), pp. 1449–1450. DOI: 10.1038/1781449a0.

[1096] N. F. Ramsey. “Thermodynamics and Statistical Mechanics at Negative Absolute Temperatures”. In: *Physical Review* 103.1 (1956), pp. 20–28. DOI: 10.1103/PhysRev.103.20.

[1097] F. REINES and C. L. COWANjun. “The Neutrino”. In: *Nature* 178.4531 (1956), pp. 446–449. DOI: 10.1038/178446a0.

[1098] H. E. Suess and H. C. Urey. “Abundances of the Elements”. In: *Reviews of Modern Physics* 28.1 (1956), pp. 53–74. DOI: 10.1103/RevModPhys.28.53.

[1099] A. S. Wightman. “Quantum Field Theory in Terms of Vacuum Expectation Values”. In: *Physical Review* 101.2 (1956), pp. 860–866. DOI: 10.1103/PhysRev.101.860.

[1100] A. A. Abrikosov. “On the magnetic properties of superconductors of the second group”. In: *Soviet Physics JETP* 5 (1957). Translation of Zh. Eksp. Teor. Fiz. 32, 1442 (1957); no DOI., pp. 1174–1182.

[1101] E. Ambler, R. W. Hayward, D. D. Hoppes, R. P. Hudson, and C. S. Wu. “Further Experiments onβDecay of Polarized Nuclei”. In: *Physical Review* 106.6 (1957), pp. 1361–1363. DOI: 10.1103/PhysRev.106.1361.

[1102] J. Bardeen, L. N. Cooper, and J. R. Schrieffer. “Theory of Superconductivity”. In: *Physical Review* 108.5 (1957), pp. 1175–1204. DOI: 10.1103/PhysRev.108.1175.

[1103] N. N. Bogoliubow and O. S. Parasiuk. “Über die Multiplikation der Kausalfunktionen in der Quantentheorie der Felder”. In: *Acta Mathematica* 97 (1957). Not a real collision: both candidate records are this paper (same DOI); they differ only in that one lost the umlaut on Über. Acta Mathematica prints the byline in German transliteration as N. N. Bogoliubow; the citation key keeps the customary Bogoliubov., pp. 227–266.

[1104] B. N. Brockhouse. “Scattering of Neutrons by Spin Waves in Magnetite”. In: *Physical Review* 106.5 (1957), pp. 859–864. DOI: 10.1103/PhysRev.106.859.

[1105] E. M. Burbidge, G. R. Burbidge, W. A. Fowler, and F. Hoyle. “Synthesis of the Elements in Stars”. In: *Reviews of Modern Physics* 29.4 (1957), pp. 547–650. DOI: 10.1103/RevModPhys.29.547.

[1106] A. G. W. Cameron. “Nuclear Reactions in Stars and Nucleogenesis”. In: *Publications of the Astronomical Society of the Pacific* 69 (1957), p. 201. DOI: 10.1086/127051.

[1107] H. Everett. “"Relative State" Formulation of Quantum Mechanics”. In: *Reviews of Modern Physics* 29.3 (1957), pp. 454–462. DOI: 10.1103/RevModPhys.29.454.

[1108] H. Frauenfelder et al. “Parity and the Polarization of Electrons fromCo60”. In: *Physical Review* 106.2 (1957), pp. 386–387. DOI: 10.1103/PhysRev.106.386.

[1109] J. I. Friedman and V. L. Telegdi. “Nuclear Emulsion Evidence for Parity Nonconservation in the Decay Chain$\pi+-\mu+-$e+”. In: *Physical Review* 105.5 (1957), pp. 1681–1682. DOI: 10.1103/physrev.105.1681.2.

[1110] R. L. Garwin, L. M. Lederman, and M. Weinrich. “Observations of the Failure of Conservation of Parity and Charge Conjugation in Meson Decays: the Magnetic Moment of the Free Muon”. In: *Physical Review* 105.4 (1957), pp. 1415–1417. DOI: 10.1103/PhysRev.105.1415.

[1111] A. Gleason. “Measures on the Closed Subspaces of a Hilbert Space”. In: *Indiana University Mathematics Journal* 6.4 (1957), pp. 885–893. DOI: 10.1512/iumj.1957.6.56050.

[1112] R. E. Glover and M. Tinkham. “Conductivity of Superconducting Films for Photon Energies between 0.3 and40kTc”. In: *Physical Review* 108.2 (1957), pp. 243–256. DOI: 10.1103/PhysRev.108.243.

[1113] D. Hall and A. S. Wightman. “A theorem on invariant analytic functions with applications to relativistic quantum field theory”. In: *Kongelige Danske Videnskabernes Selskab, Matematisk-fysiske Meddelelser* 31 (1957). Number 5 of the series. No DOI., pp. 1–41.

[1114] E. T. Jaynes. “Information Theory and Statistical Mechanics”. In: *Physical Review* 106.4 (1957). Part I; Part II is Phys. Rev. 108, 171 (1957), not cited here, pp. 620–630. DOI: 10.1103/PhysRev.106.620.

[1115] R. Jost. “Eine Bemerkung zum CTP-Theorem”. In: *Helvetica Physica Acta* 30 (1957), pp. 409–416.

[1116] E. O. Kane. “Band structure of indium antimonide”. In: *Journal of Physics and Chemistry of Solids* 1.4 (1957), pp. 249–261. DOI: 10.1016/0022-3697(57)90013-6.

[1117] R. Kubo. “Statistical-Mechanical Theory of Irreversible Processes. I. General Theory and Simple Applications to Magnetic and Conduction Problems”. In: *Journal of the Physical Society of Japan* 12.6 (1957), pp. 570–586. DOI: 10.1143/JPSJ.12.570.

[1118] L. Landau. “On the conservation laws for weak interactions”. In: *Nuclear Physics* 3.1 (1957). The two-component neutrino and combined CP parity. Printed as ’L. Landau’. Metadata confirmed against Crossref. A key Landau:1957a once existed carrying a byte-identical copy of this entry, same DOI and all; the intended split was a = this weak-interaction paper and b = Landau’s Fermi-liquid paper of the same year, which is present as Landau:1957b; what went wrong is that the unsuffixed key was left behind as well, so one work stood under two identical keys and the bibliography would have printed it twice. The suffixed duplicate was deleted and its citations repointed here. Landau:1957b is a different paper and is unaffected., pp. 127–131. DOI: 10.1016/0029-5582(57)90061-5.

[1119] L. D. Landau. “The Theory of a Fermi Liquid”. In: *Soviet Physics JETP* 3 (1957). English translation of Zh. Eksp. Teor. Fiz. 30, 1058 (1956); the translation volume carries the 1957 issue year used for the key. Soviet Physics JETP articles of this period have no DOI, so none is given rather than guessed. Split from the former shared key Landau:1957., pp. 920–925.

[1120] J. D. Lawson. “Some Criteria for a Power Producing Thermonuclear Reactor”. In: *Proceedings of the Physical Society. Section B* 70.1 (1957). Issue 1. Crossref-confirmed. The ’criterion written down in 1957’ of the plasma chapter. Lawson circulated it as classified AERE report GP/R 1807 in 1955; the 1957 Proc. Phys. Soc. B paper is the published version and the correct citation., pp. 6–10. DOI: 10.1088/0370-1301/70/1/303.

[1121] T. D. Lee and C. N. Yang. “Parity Nonconservation and a Two-Component Theory of the Neutrino”. In: *Physical Review* 105.5 (1957), pp. 1671–1675. DOI: 10.1103/PhysRev.105.1671.

[1122] H. Lewy. “An Example of a Smooth Linear Partial Differential Equation Without Solution”. In: *The Annals of Mathematics* 66.1 (1957). Annals of Mathematics, second series., p. 155. DOI: 10.2307/1970121.

[1123] G. Lüders. “Proof of the TCP theorem”. In: *Annals of Physics* 2.1 (1957), pp. 1–15. DOI: 10.1016/0003-4916(57)90032-5.

[1124] R. W. Morse and H. V. Bohm. “Superconducting Energy Gap from Ultrasonic Attenuation Measurements”. In: *Physical Review* 108.4 (1957), pp. 1094–1096. DOI: 10.1103/PhysRev.108.1094.

[1125] E. N. Parker. “Sweet’s mechanism for merging magnetic fields in conducting fluids”. In: *Journal of Geophysical Research* 62.4 (1957). Issue 4, December 1957. This is the quantitative current-sheet model giving the R_m^-1/2 rate; the companion Sweet:1958 is the IAU Symposium 6 paper, a different kind of source (inproceedings)., pp. 509–520. DOI: 10.1029/JZ062i004p00509.

[1126] A. B. Pippard. “An experimental determination of the Fermi surface in copper”. In: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 250.979 (1957), pp. 325–357. DOI: 10.1098/rsta.1957.0023.

[1127] B. Pontecorvo. “Mesonium and antimesonium”. In: *Zhurnal Eksperimentalnoi i Teoreticheskoi Fiziki* 33 (1957). Russian original; English translation in Soviet Physics JETP 6 (1958) 429. No DOI, pp. 549–551.

[1128] A. G. Redfield. “On the Theory of Relaxation Processes”. In: *IBM Journal of Research and Development* 1.1 (1957), pp. 19–31. DOI: 10.1147/rd.11.0019.

[1129] M. N. Rosenbluth, W. M. MacDonald, and D. L. Judd. “Fokker-Planck Equation for an Inverse-Square Force”. In: *Physical Review* 107.1 (1957). Crossref-verified. Source of the Rosenbluth potentials cited at sec:plasma-collisions. Three authors; keyed by first listed author per the project convention., pp. 1–6. DOI: 10.1103/PhysRev.107.1.

[1130] A. Salam. “On parity conservation and neutrino mass”. In: *Il Nuovo Cimento* 5.1 (1957), pp. 299–301. DOI: 10.1007/BF02812841.

[1131] H. Schopper. “Circular polarization of γ-rays: Further proof for parity failure in β decay”. In: *Philosophical Magazine* 2.17 (1957), pp. 710–713. DOI: 10.1080/14786435708242717.

[1132] Y. Takahashi. “On the generalized ward identity”. In: *Il Nuovo Cimento* 6.2 (1957), pp. 371–375. DOI: 10.1007/BF02832514.

[1133] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson. “Experimental Test of Parity Conservation in Beta Decay”. In: *Physical Review* 105.4 (1957), pp. 1413–1415. DOI: 10.1103/PhysRev.105.1413.

[1134] P. W. Anderson. “Absence of Diffusion in Certain Random Lattices”. In: *Physical Review* 109.5 (1958), pp. 1492–1505. DOI: 10.1103/PhysRev.109.1492.

[1135] I. B. Bernstein, E. A. Frieman, M. D. Kruskal, and R. M. Kulsrud. “An energy principle for hydromagnetic stability problems”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 244.1236 (1958). The MHD energy principle cited for linear stability from the sign of a quadratic form. Crossref-verified; issued 25 Feb 1958., pp. 17–40. DOI: 10.1098/rspa.1958.0023.

[1136] E. Böhm-Vitense. “Über die Wasserstoffkonvektionszone in Sternen verschiedener Effektivtemperaturen und Leuchtkräfte”. In: *Zeitschrift für Astrophysik* 46 (1958). Mixing-length theory; no DOI, pp. 108–143.

[1137] B. N. Brockhouse and P. K. Iyengar. “Normal Modes of Germanium by Neutron Spectrometry”. In: *Physical Review* 111.3 (1958), pp. 747–754. DOI: 10.1103/PhysRev.111.747.

[1138] O. Buneman. “Instability, Turbulence, and Conductivity in Current-Carrying Plasma”. In: *Physical Review Letters* 1.1 (1958). Crossref confirmed: PRL vol. 1 (the journal’s first volume), pp. 8-9, 1 July 1958. The Buneman instability. A longer companion treatment exists as O. Buneman, ’Dissipation of Currents in Ionized Media’, Phys. Rev. 115, 503 (1959) – the key’s year points to the 1958 PRL, which is the standard citation., pp. 8–9. DOI: 10.1103/PhysRevLett.1.8.

[1139] N. Burgoyne. “On the connection of spin with statistics”. In: *Il Nuovo Cimento* 8.4 (1958), pp. 607–609. DOI: 10.1007/bf02828775.

[1140] P. A. M. Dirac. *The Principles of Quantum Mechanics*. Fourth edition. Oxford: Clarendon Press, 1958.

[1141] L. Esaki. “New Phenomenon in Narrow Germaniump−nJunctions”. In: *Physical Review* 109.2 (1958), pp. 603–604. DOI: 10.1103/PhysRev.109.603.

[1142] H. Feshbach. “Unified theory of nuclear reactions”. In: *Annals of Physics* 5.4 (1958). Issue 4, December 1958. The projection-operator formalism behind what are now called Feshbach resonances. Crossref record verified., pp. 357–390. DOI: 10.1016/0003-4916(58)90007-1.

[1143] H. FESHBACH and F. VILLARS. “Elementary Relativistic Wave Mechanics of Spin 0 and Spin 1/2 Particles”. In: *Reviews of Modern Physics* 30.1 (1958). Issue 1, January 1958 — earlier in the year than Feshbach:1958b, hence the a suffix. Crossref record verified: title, authors, volume, pages all match., pp. 24–45. DOI: 10.1103/RevModPhys.30.24.

[1144] R. P. Feynman and M. Gell-Mann. “Theory of the Fermi Interaction”. In: *Physical Review* 109.1 (1958), pp. 193–198. DOI: 10.1103/PhysRev.109.193.

[1145] D. Finkelstein. “Past-Future Asymmetry of the Gravitational Field of a Point Particle”. In: *Physical Review* 110.4 (1958), pp. 965–967. DOI: 10.1103/PhysRev.110.965.

[1146] F. C. Frank. “Liquid crystals. On the theory of liquid crystals”. In: *Discussions of the Faraday Society* 25 (1958), p. 19. DOI: 10.1039/df9582500019.

[1147] M. L. Goldberger and S. B. Treiman. “Decay of the Pi Meson”. In: *Physical Review* 110.5 (1958), pp. 1178–1184. DOI: 10.1103/PhysRev.110.1178.

[1148] M. Goldhaber, L. Grodzins, and A. W. Sunyar. “Helicity of Neutrinos”. In: *Physical Review* 109.3 (1958), pp. 1015–1017. DOI: 10.1103/PhysRev.109.1015.

[1149] H. Grad and H. Rubin. “Hydromagnetic Equilibria and Force-Free Fields”. In: *Proceedings of the Second United Nations International Conference on the Peaceful Uses of Atomic Energy, Geneva, 1–13 September 1958, Volume 31: Theoretical and Experimental Aspects of Controlled Nuclear Fusion*. Vol. 31. The origin of the Grad–Shafranov equation, as the citing sentence requires. UN conference proceedings: @inproceedings with booktitle; the ’31’ is the printed volume of the proceedings set, not a journal volume. Commonly cited as ’Vol. 31, p. 190’; the paper runs to p. 197. No DOI exists for this item. 1958, pp. 190–197.

[1150] J. B. Keller. “Corrected bohr-sommerfeld quantum conditions for nonseparable systems”. In: *Annals of Physics* 4.2 (1958). Issue 2. The extension of Bohr–Sommerfeld to non-separable systems with the general (Maslov) index counting, which is precisely the citing claim. Crossref lowercases the title as ’Corrected bohr-sommerfeld ...’; the form above is the published capitalisation. Note the parentheses inside the DOI – they are part of it (Elsevier legacy PII form) and must be percent-encoded if it is ever fetched by URL., pp. 180–188. DOI: 10.1016/0003-4916(58)90032-0.

[1151] G. Lüders and B. Zumino. “Connection between Spin and Statistics”. In: *Physical Review* 110.6 (1958). Key transliterates the umlaut (Lueders) to keep the citekey ASCII., pp. 1450–1453. DOI: 10.1103/PhysRev.110.1450.

[1152] G. G. Macfarlane, T. P. McLean, J. E. Quarrington, and V. Roberts. “Fine Structure in the Absorption-Edge Spectrum of Si”. In: *Physical Review* 111.5 (1958), pp. 1245–1254. DOI: 10.1103/PhysRev.111.1245.

[1153] W. Maier and A. Saupe. “Eine einfache molekular-statistische Theorie der nematischen kristallinflüssigen Phase. Teil I”. In: *Zeitschrift für Naturforschung A* 13 (1958), pp. 564–566.

[1154] R. L. Mössbauer. “Kernresonanzfluoreszenz von Gammastrahlung in Ir^{191}”. In: *Zeitschrift für Physik* 151 (1958), pp. 124–143.

[1155] S. Nakajima. “On Quantum Theory of Transport Phenomena”. In: *Progress of Theoretical Physics* 20.6 (1958), pp. 948–959. DOI: 10.1143/PTP.20.948.

[1156] E. N. Parker. “Dynamics of the Interplanetary Gas and Magnetic Fields”. In: *The Astrophysical Journal* 128 (1958). November 1958. The solar-wind prediction paper. Crossref prints a trailing period in the title; drop it. Distinct from the already-present bib entry Parker:2018., p. 664. DOI: 10.1086/146579.

[1157] L. J. van der Pauw. “A method of measuring specific resistivity and Hall effect of discs of arbitrary shape”. In: *Philips Research Reports* 13 (1958), pp. 1–9.

[1158] J. R. A. Pearson. “On convection cells induced by surface tension”. In: *Journal of Fluid Mechanics* 4.5 (1958). Issue 5. The theoretical half: linear stability of a layer with a temperature-dependent surface tension, giving the Marangoni threshold. Note a distinct existing key Pearson:1949 is already in references.bib (different author/work) – no collision. Crossref returns the DOI lower-cased (10.1017/s0022112058000616); the publisher’s canonical form is upper-case S., pp. 489–500. DOI: 10.1017/S0022112058000616.

[1159] B. Pontecorvo. “Mesonium and Antimesonium”. In: *Soviet Physics JETP* 6 (1958). English translation of Zh. Eksp. Teor. Fiz. 33, 549 (1957); keyed by the year of the translated issue. No DOI., p. 429.

[1160] A. L. Schawlow and C. H. Townes. “Infrared and Optical Masers”. In: *Physical Review* 112.6 (1958), pp. 1940–1949. DOI: 10.1103/PhysRev.112.1940.

[1161] M. Sparnaay. “Measurements of attractive forces between flat plates”. In: *Physica* 24.6-10 (1958), pp. 751–764. DOI: 10.1016/S0031-8914(58)80090-7.

[1162] E. C. G. Sudarshan and R. E. Marshak. “Chirality Invariance and the Universal Fermi Interaction”. In: *Physical Review* 109.5 (1958), pp. 1860–1862. DOI: 10.1103/physrev.109.1860.2.

[1163] P. A. Sweet. “The neutral point theory of solar flares”. In: *Symposium - International Astronomical Union*. Vol. 6. Crossref-verified. IAU Symposium 6, edited by B. Lehnert; Crossref’s container-title is the series name ’Symposium - International Astronomical Union’ and its deposited article title carries the contribution number (’14. The neutral point theory of solar flares’), dropped here. The editor (Lehnert) belongs in the entry. Dungey’s reply to Cowling follows at pp. 135-140. 1958, pp. 123–134. DOI: 10.1017/S0074180900237704.

[1164] Y. Aharonov and D. Bohm. “Significance of Electromagnetic Potentials in the Quantum Theory”. In: *Physical Review* 115.3 (1959), pp. 485–491. DOI: 10.1103/PhysRev.115.485.

[1165] V. Bargmann, L. Michel, and V. L. Telegdi. “Precession of the Polarization of Particles Moving in a Homogeneous Electromagnetic Field”. In: *Physical Review Letters* 2.10 (1959). Bargmann:1954 and Bargmann:1964 already exist in references.bib; 1959 is free, no suffix needed., pp. 435–436. DOI: 10.1103/PhysRevLett.2.435.

[1166] N. N. Bogoliubov and D. V. Shirkov. *Introduction to the Theory of Quantized Fields*. Interscience Publishers, 1959.

[1167] H. A. Buchdahl. “General Relativistic Fluid Spheres”. In: *Physical Review* 116.4 (1959), pp. 1027–1034. DOI: 10.1103/PhysRev.116.1027.

[1168] L. P. Gor'kov. "Microscopic derivation of the Ginzburg-Landau equations in the theory of superconductivity". In: *Soviet Physics JETP* 9 (1959). Translation of Zh. Eksp. Teor. Fiz. 36, 1918 (1959); no DOI., pp. 1364–1367.

[1169] L. C. Hebel and C. P. Slichter. "Nuclear Spin Relaxation in Normal and Superconducting Aluminum". In: *Physical Review* 113.6 (1959), pp. 1504–1519. DOI: 10.1103/PhysRev.113.1504.

[1170] J. Kanamori. "Superexchange interaction and symmetry properties of electron orbitals". In: *Journal of Physics and Chemistry of Solids* 10.2-3 (1959), pp. 87–98. DOI: 10.1016/0022-3697(59)90061-7.

[1171] G. V. Kulikov and G. B. Khristiansen. "On the Size Spectrum of Extensive Air Showers". In: *Soviet Physics JETP* 8 (1959), pp. 441–444.

[1172] J. C. Phillips and L. Kleinman. "New Method for Calculating Wave Functions in Crystals and Molecules". In: *Physical Review* 116.2 (1959), pp. 287–294. DOI: 10.1103/PhysRev.116.287.

[1173] K. R. Popper. *The Logic of Scientific Discovery*. Original German edition: Logik der Forschung, Springer, Vienna (1934). London: Hutchinson, 1959.

[1174] T. Regge. "Introduction to complex orbital momenta". In: *Il Nuovo Cimento* 14.5 (1959), pp. 951–976. DOI: 10.1007/BF02728177.

[1175] H. F. Trotter. "On the product of semi-groups of operators". In: *Proceedings of the American Mathematical Society* 10.4 (1959), pp. 545–551. DOI: 10.1090/S0002-9939-1959-0108732-6.

[1176] J. A. Van Allen. "The geomagnetically trapped corpuscular radiation". In: *Journal of Geophysical Research* 64.11 (1959). Verified via Crossref (JGR 64(11), Nov 1959; Crossref files the family name as "Van Allen"). Chosen over the other 1959 candidate, Van Allen & Frank, "Radiation around the Earth to a radial distance of 107,400 km", Nature 183, 430-434 (1959), doi:10.1038/183430a0, because the key is single-author and the cited claim is the trapped-belt/loss-cone population, which is this paper's exact subject. The true discovery paper is earlier: Van Allen, Ludwig, Ray & McIlwain, Jet Propulsion 28, 588-592 (1958) — if the chapter later wants the discovery rather than the characterization, that is the entry to add under a new key., pp. 1683–1689. DOI: 10.1029/JZ064i011p01683.

[1177] H. .-. Borchers. "Über die mannigfaltigkeit der interpolierenden felder zu einer kausalenS-Matrix". In: *Il Nuovo Cimento* 15.5 (1960), pp. 784–794. DOI: 10.1007/bf02732693.

[1178] R. G. Chambers. "Shift of an Electron Interference Pattern by Enclosed Magnetic Flux". In: *Physical Review Letters* 5.1 (1960). Crossref-verified (resolved record read: Phys. Rev. Lett. 5(1) 3–5, 1960). First measurement of the electromagnetic Aharonov–Bohm phase, using an iron whisker behind a biprism. The leakage-field objection to this experiment is closed by Tonomura:1986., pp. 3–5. DOI: 10.1103/PhysRevLett.5.3.

[1179] I. Giaever. "Electron Tunneling Between Two Superconductors". In: *Physical Review Letters* 5.10 (1960). Issue 10, 15 November 1960. This is the SECOND of Giaever's two 1960 PRLs and the reason the key needs the 'b': the existing Giaever:1960 is 'Energy Gap in Superconductors Measured by Electron Tunneling', PRL 5 (4), 147-148, the normal-superconductor junction that the chapter's experiment section reports. This one is the superconductor-insulator-superconductor junction, where the single-particle current acquires the larger threshold – which is the exact claim it was added to., pp. 464–466. DOI: 10.1103/PhysRevLett.5.464.

[1180] I. Giaever. “Energy Gap in Superconductors Measured by Electron Tunneling”. In: *Physical Review Letters* 5.4 (1960), pp. 147–148. DOI: 10.1103/PhysRevLett.5.147.

[1181] V. L. Ginzburg. “Some remarks on phase transitions of the second kind and the microscopic theory of ferroelectric materials”. In: *Soviet Physics Solid State* 2 (1960), pp. 1824–1834.

[1182] H. M. Goldenberg, D. Kleppner, and N. F. Ramsey. “Atomic Hydrogen Maser”. In: *Physical Review Letters* 5.8 (1960), pp. 361–362. DOI: 10.1103/PhysRevLett.5.361.

[1183] P. Hartman. “A lemma in the theory of structural stability of differential equations”. In: *Proceedings of the American Mathematical Society* 11.4 (1960). The Hartman side of the Hartman-Grobman linearization theorem (Grobman’s paper is Dokl. Akad. Nauk SSSR 128 (1959) 880-881)., pp. 610–620. DOI: 10.1090/s0002-9939-1960-0121542-7.

[1184] V. W. Hughes, D. W. McColm, K. Ziock, and R. Prepost. “Formation of Muonium and Observation of its Larmor Precession”. In: *Physical Review Letters* 5.2 (1960). Issue 2. Exactly the claim cited: muonium observed through its Larmor precession. Verified against Crossref., pp. 63–65. DOI: 10.1103/PhysRevLett.5.63.

[1185] M. D. Kruskal. “Maximal Extension of Schwarzschild Metric”. In: *Physical Review* 119.5 (1960), pp. 1743–1745. DOI: 10.1103/PhysRev.119.1743.

[1186] T. H. MAIMAN. “Stimulated Optical Radiation in Ruby”. In: *Nature* 187.4736 (1960), pp. 493–494. DOI: 10.1038/187493a0.

[1187] Y. Nambu. “Quasi-Particles and Gauge Invariance in the Theory of Superconductivity”. In: *Physical Review* 117.3 (1960), pp. 648–663. DOI: 10.1103/PhysRev.117.648.

[1188] L. Pauling. *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry*. 3rd edition. Cornell University Press, 1960.

[1189] R. V. Pound and G. A. Rebka. “Apparent Weight of Photons”. In: *Physical Review Letters* 4.7 (1960), pp. 337–341. DOI: 10.1103/PhysRevLett.4.337.

[1190] L. I. Schiff. “On Experimental Tests of the General Theory of Relativity”. In: *American Journal of Physics* 28.4 (1960), pp. 340–343. DOI: 10.1119/1.1935800.

[1191] L. I. Schiff. “Possible New Experimental Test of General Relativity Theory”. In: *Physical Review Letters* 4.5 (1960). THE GYROSCOPE PROPOSAL – the paper that became Gravity Probe B, giving both the geodetic and the frame-dragging precession rates for an orbiting gyroscope. Schiff has three 1960 papers and only this one makes the claim: the existing key Schiff:1960 is "On Experimental Tests of the General Theory of Relativity", Am. J. Phys. 28, 340 (the Schiff-conjecture paper, cited in 01-equivalence-classical-tests.tex), and there is a third, "Motion of a Gyroscope According to Einstein’s Theory of Gravitation", Proc. Natl. Acad. Sci. USA 46, 871–882 (1960), which is the longer companion to this one. Crossref confirmed title, volume, pages and year against the DOI. Cited at the two precession rates in phen:clt-gyroscope., pp. 215–217. DOI: 10.1103/PhysRevLett.4.215.

[1192] F. T. Smith. “Lifetime Matrix in Collision Theory”. In: *Physical Review* 118.1 (1960), pp. 349–356. DOI: 10.1103/PhysRev.118.349.

[1193] P. Suppes. “A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences”. In: *Synthese* 12.2-3 (1960). The semantic view of theories: a theory is presented as a class of models, not as a set of sentences, pp. 287–301. DOI: 10.1007/BF00485107.

[1194] G. Szekeres. “On the singularities of a Riemannian manifold”. In: *Publicationes Mathematicae Debrecen* 7 (1960). Independent of Kruskal. Reprinted Gen. Relativ. Gravit. 34, 2001–2016 (2002)., pp. 285–301.

[1195] E. P. Wigner. “The Unreasonable Effectiveness of Mathematics in the Natural Sciences”. In: *Communications on Pure and Applied Mathematics* 13 (1960), pp. 1–14. DOI: 10.1002/cpa.3160130102.

[1196] J. M. Ziman. *Electrons and Phonons: The Theory of Transport Phenomena in Solids*. Oxford University Press, 1960.

[1197] R. Zwanzig. “Ensemble Method in the Theory of Irreversibility”. In: *The Journal of Chemical Physics* 33.5 (1960), pp. 1338–1341. DOI: 10.1063/1.1731409.

[1198] J. Bardeen. “Tunnelling from a Many-Particle Point of View”. In: *Physical Review Letters* 6.2 (1961). Issue 2, 15 January 1961. Crossref-verified. The published title uses the British ’Tunnelling’ although the journal is American and the same volume spells it ’Tunneling’ elsewhere – do not normalise it. This is the transfer-Hamiltonian paper, written to account for Giaever’s junctions: it gives the elastic current as the convolution of the two densities of states weighted by the difference of the Fermi functions. Cited here in two proofpending notes (the Giaever N-I-S current, and the STM transfer-Hamiltonian step that Tersoff:1985 then reduces)., pp. 57–59. DOI: 10.1103/PhysRevLett.6.57.

[1199] S. Chandrasekhar. *Hydrodynamic and Hydromagnetic Stability*. The International Series of Monographs on Physics; xix + 652 pp. No DOI for the 1961 Clarendon edition (the 1981 Dover reprint has none either). This is the standard source for the rigid-boundary Bénard problem: Ra_c = 1707.762 and a_c = 3.117, i.e. the ’near 1708, with a correspondingly larger critical wavenumber’ clause. Existing keys Chandrasekhar:1931/1939/1943 are unrelated works – no collision. Oxford: Clarendon Press, 1961.

[1200] G. F. Chew and S. C. Frautschi. “Principle of Equivalence for all Strongly Interacting Particles within theS-Matrix Framework”. In: *Physical Review Letters* 7.10 (1961), pp. 394–397. DOI: 10.1103/PhysRevLett.7.394.

[1201] B. S. Deaver and W. M. Fairbank. “Experimental Evidence for Quantized Flux in Superconducting Cylinders”. In: *Physical Review Letters* 7.2 (1961), pp. 43–46. DOI: 10.1103/PhysRevLett.7.43.

[1202] R. Doll and M. Näbauer. “Experimental Proof of Magnetic Flux Quantization in a Superconducting Ring”. In: *Physical Review Letters* 7.2 (1961), pp. 51–52. DOI: 10.1103/PhysRevLett.7.51.

[1203] J. W. Dungey. “Interplanetary Magnetic Field and the Auroral Zones”. In: *Physical Review Letters* 6.2 (1961). Crossref confirmed: PRL vol. 6, pp. 47-48, 15 January 1961. Two pages; this is the open-magnetosphere / Dungey cycle paper the chapter cites., pp. 47–48. DOI: 10.1103/PhysRevLett.6.47.

[1204] U. Fano. “Effects of Configuration Interaction on Intensities and Phase Shifts”. In: *Physical Review* 124.6 (1961), pp. 1866–1878. DOI: 10.1103/PhysRev.124.1866.

[1205] M. Gell-Mann. “THE EIGHTFOLD WAY: A THEORY OF STRONG INTERACTION SYMMETRY”. In: *California Institute of Technology Synchrotron Laboratory Report CTSL-20* (1961). Unpublished laboratory report; reprinted in Gell-Mann and Ne’eman, The Eightfold Way (Benjamin, 1964). DOI: 10.2172/4008239.

[1206] S. L. Glashow. “Partial-symmetries of weak interactions”. In: *Nuclear Physics* 22.4 (1961), pp. 579–588. DOI: 10.1016/0029-5582(61)90469-2.

[1207] J. Goldstone. “Field theories with « Superconductor » solutions”. In: *Il Nuovo Cimento* 19.1 (1961), pp. 154–164. DOI: 10.1007/BF02812722.

[1208] E. P. Gross. “Structure of a quantized vortex in boson systems”. In: *Il Nuovo Cimento* 20.3 (1961), pp. 454–477. DOI: 10.1007/BF02731494.

[1209] C. Hayashi. “Stellar Evolution in Early Phases of Gravitational Contraction”. In: *Publications of the Astronomical Society of Japan* 13.4 (1961). No DOI, pp. 450–452. DOI: 10.1093/pasj/13.4.450.

[1210] D. G. Henshaw and A. D. B. Woods. “Modes of Atomic Motions in Liquid Helium by Inelastic Scattering of Neutrons”. In: *Physical Review* 121.5 (1961), pp. 1266–1274. DOI: 10.1103/PhysRev.121.1266.

[1211] K. Johnson and E. Sudarshan. “Inconsistency of the local field theory of charged spin32 particles”. In: *Annals of Physics* 13.1 (1961), pp. 126–145. DOI: 10.1016/0003-4916(61)90030-6.

[1212] C. Jönsson. “Elektroneninterferenzen an mehreren künstlich hergestellten Feinspalten”. In: *Zeitschrift für Physik* 161 (1961). Issue 4. The first true multi-slit electron interference experiment, which is what the citing sentence claims. Surname carries an umlaut as the publisher prints it (Jönsson) even though the citekey is ASCII 'Jonsson:1961' – keep the key, spell the name with the umlaut in the entry., pp. 454–474.

[1213] R. Landauer. “Irreversibility and Heat Generation in the Computing Process”. In: *IBM Journal of Research and Development* 5.3 (1961), pp. 183–191. DOI: 10.1147/rd.53.0183.

[1214] E. Nagel. *The Structure of Science: Problems in the Logic of Scientific Explanation.* The classical account of intertheoretic reduction: a superseded theory survives as a limit case of its successor under stated bridge principles. New York: Harcourt, Brace and World, 1961.

[1215] Y. Nambu and G. Jona-Lasinio. “Dynamical Model of Elementary Particles Based on an Analogy with Superconductivity. I”. In: *Physical Review* 122.1 (1961), pp. 345–358. DOI: 10.1103/PhysRev.122.345.

[1216] Y. Ne'eman. “Derivation of strong interactions from a gauge invariance”. In: *Nuclear Physics* 26.2 (1961), pp. 222–229. DOI: 10.1016/0029-5582(61)90134-1.

[1217] L. P. Pitaevskii. “Vortex lines in an imperfect Bose gas”. In: *Soviet Physics JETP* 13 (1961). English translation of Zh. Eksp. Teor. Fiz. 40, 646 (1961)., pp. 451–454.

[1218] H. Reeh and S. Schlieder. “Bemerkungen zur unitäräquivalenz von lorentzinvarianten feldern”. In: *Il Nuovo Cimento* 22.5 (1961), pp. 1051–1068. DOI: 10.1007/BF02787889.

[1219] W. Shockley and H. J. Queisser. “Detailed Balance Limit of Efficiency ofp-nJunction Solar Cells”. In: *Journal of Applied Physics* 32.3 (1961), pp. 510–519. DOI: 10.1063/1.1736034.

[1220] W. F. Vinen. “The detection of single quanta of circulation in liquid helium II”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 260.1301 (1961), pp. 218–236. DOI: 10.1098/rspa.1961.0029.

[1221] R. Arnowitt, S. Deser, and C. W. Misner. “The Dynamics of General Relativity”. In: *Gravitation: An Introduction to Current Research (ed. Louis Witten).* Book chapter (@incollection in biblatex terms). Republished as Gen. Relativ. Gravit. 40, 1997–2027 (2008), DOI of the reprint 10.1007/s10714-008-0661-1 – do not attach the 2008 DOI to the 1962 entry. 1962, pp. 227–265.

[1222] R. Courant and D. Hilbert. *Methods of Mathematical Physics, Volume II: Partial Differential Equations*. Resolves the first TODO-cite of the PDE chapter; also cited from the calculus-of-variations chapter. New York: Interscience Publishers, 1962.

[1223] G. Danby et al. "Observation of High-Energy Neutrino Reactions and the Existence of Two Kinds of Neutrinos". In: *Physical Review Letters* 9.1 (1962), pp. 36–44. DOI: 10.1103/PhysRevLett.9.36.

[1224] Y. N. Denisyuk. "On the reproduction of the optical properties of an object by the wave field of its scattered radiation". In: *Soviet Physics Doklady* 7 (1962). English translation of Doklady Akademii Nauk SSSR 144, 1275 (1962). The reflection (white-light) hologram, arrived at independently of Gabor and of Leith and Upatnieks. Soviet Physics Doklady carries no DOI; verifiable by volume and page., pp. 543–545.

[1225] R. Giacconi, H. Gursky, F. R. Paolini, and B. B. Rossi. "Evidence for x Rays From Sources Outside the Solar System". In: *Physical Review Letters* 9.11 (1962), pp. 439–443. DOI: 10.1103/PhysRevLett.9.439.

[1226] J. Goldstone, A. Salam, and S. Weinberg. "Broken Symmetries". In: *Physical Review* 127.3 (1962), pp. 965–970. DOI: 10.1103/PhysRev.127.965.

[1227] H. L. Grant, R. W. Stewart, and A. Moilliet. "Turbulence spectra from a tidal channel". In: *Journal of Fluid Mechanics* 12.2 (1962). The Discovery Passage measurement confirming the five-thirds spectrum., pp. 241–268. DOI: 10.1017/s002211206200018x.

[1228] S. R. de Groot and P. Mazur. *Non-Equilibrium Thermodynamics*. North-Holland, 1962.

[1229] R. N. Hall, G. E. Fenner, J. D. Kingsley, T. J. Soltys, and R. O. Carlson. "Coherent Light Emission From GaAs Junctions". In: *Physical Review Letters* 9.9 (1962), pp. 366–368. DOI: 10.1103/PhysRevLett.9.366.

[1230] N. Holonyak and S. F. Bevacqua. "COHERENT (VISIBLE) LIGHT EMISSION FROM Ga(As1−xPx) JUNCTIONS". In: *Applied Physics Letters* 1.4 (1962), pp. 82–83. DOI: 10.1063/1.1753706.

[1231] B. Josephson. "Possible new effects in superconductive tunnelling". In: *Physics Letters* 1.7 (1962), pp. 251–253. DOI: 10.1016/0031-9163(62)91369-0.

[1232] T. Kinoshita. "Mass Singularities of Feynman Amplitudes". In: *Journal of Mathematical Physics* 3.4 (1962), pp. 650–677. DOI: 10.1063/1.1724268.

[1233] T. S. Kuhn. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1962.

[1234] R. B. Leighton, R. W. Noyes, and G. W. Simon. "Velocity Fields in the Solar Atmosphere. I. Preliminary Report". In: *The Astrophysical Journal* 135 (1962). Discovery of the solar five-minute oscillations, p. 474. DOI: 10.1086/147285.

[1235] E. N. Leith and J. Upatnieks. "Reconstructed Wavefronts and Communication Theory*". In: *Journal of the Optical Society of America* 52.10 (1962). Issue 10. The off-axis (carrier-frequency) reference wave that separates the four reconstructed waves in direction and removes the twin image., p. 1123. DOI: 10.1364/JOSA.52.001123.

[1236] W. A. Little and R. D. Parks. "Observation of Quantum Periodicity in the Transition Temperature of a Superconducting Cylinder". In: *Physical Review Letters* 9.1 (1962), pp. 9–12. DOI: 10.1103/PhysRevLett.9.9.

[1237] Z. Maki, M. Nakagawa, and S. Sakata. "Remarks on the Unified Model of Elementary Particles". In: *Progress of Theoretical Physics* 28.5 (1962), pp. 870–880. DOI: 10.1143/PTP.28.870.

[1238] J. Moser. “On Invariant Curves of Area-Preserving Mappings of an Annulus”. In: *Nachrichten der Akademie der Wissenschaften in Göttingen, II. Mathematisch-Physikalische Klasse*. The Göttingen Nachrichten are numbered by year, not volume (1962, no. 1) – academy-proceedings rule, do not invent a volume. The twist theorem; the M of KAM. 1962, pp. 1–20.

[1239] M. Neugebauer and C. W. Snyder. “Solar Plasma Experiment”. In: *Science* 138.3545 (1962). Science issue 3545, 7 December 1962; one of the ’Mariner II: Preliminary Reports on Measurements of Venus’ items, which is why the printed title is only two words. Crossref holds a duplicate registration 10.1126/science.138.3545.1095-a with identical metadata; both resolve, but the ’.b’ form redirects to the live science.org landing page while ’-a’ redirects to the legacy sciencemag.org lookup – use ’.b’., pp. 1095–1097. DOI: 10.1126/science.138.3545.1095.b.

[1240] S. Okubo. “Note on Unitary Symmetry in Strong Interactions”. In: *Progress of Theoretical Physics* 27.5 (1962), pp. 949–966. DOI: 10.1143/PTP.27.949.

[1241] J. Schwinger. “Gauge Invariance and Mass”. In: *Physical Review* 125.1 (1962), pp. 397–398. DOI: 10.1103/PhysRev.125.397.

[1242] J. Spitzer Lyman. *Physics of Fully Ionized Gases*. Second revised edition, 1962; Interscience Tracts on Physics and Astronomy, number 3; 170 pp. The first edition is 1956, so the edition field matters – the chapter cites this as a standard monograph alongside Alfven:1950. No DOI; printed monograph, so an honest AMBER on availability. New York: Interscience Publishers (John Wiley & Sons), 1962.

[1243] E. Teller. “On the Stability of Molecules in the Thomas-Fermi Theory”. In: *Reviews of Modern Physics* 34.4 (1962), pp. 627–631. DOI: 10.1103/RevModPhys.34.627.

[1244] C. N. Yang. “Concept of Off-Diagonal Long-Range Order and the Quantum Phases of Liquid He and of Superconductors”. In: *Reviews of Modern Physics* 34.4 (1962), pp. 694–704. DOI: 10.1103/RevModPhys.34.694.

[1245] P. W. Anderson. “Plasmons, Gauge Invariance, and Mass”. In: *Physical Review* 130.1 (1963), pp. 439–442. DOI: 10.1103/PhysRev.130.439.

[1246] P. W. Anderson and J. M. Rowell. “Probable Observation of the Josephson Superconducting Tunneling Effect”. In: *Physical Review Letters* 10.6 (1963), pp. 230–232. DOI: 10.1103/PhysRevLett.10.230.

[1247] V. I. Arnold. “Proof of a Theorem of A. N. Kolmogorov on the Preservation of Conditionally Periodic Motions under a Small Perturbation of the Hamiltonian”. In: *Uspekhi Matematicheskikh Nauk* 18 (1963). Issue 5(113), in Russian. The DOI resolves the English translation, Russian Mathematical Surveys 18:5 (1963) 9-36, same year. The A of KAM; his ’Small denominators’ survey is Uspekhi 18:6 (1963) 91-192., pp. 13–40.

[1248] N. Cabibbo. “Unitary Symmetry and Leptonic Decays”. In: *Physical Review Letters* 10.12 (1963), pp. 531–533. DOI: 10.1103/PhysRevLett.10.531.

[1249] P. J. Cohen. “The Independence of the Continuum Hypothesis”. In: *Proceedings of the National Academy of Sciences of the United States of America* 50.6 (1963). Part II, completing the proof: Proceedings of the National Academy of Sciences 51 (1964), 105–110, doi 10.1073/pnas.51.1.105, pp. 1143–1148. DOI: 10.1073/pnas.50.6.1143.

[1250] A. M. Cormack. “Representation of a Function by Its Line Integrals, with Some Radiological Applications”. In: *Journal of Applied Physics* 34.9 (1963). Part II is J. Appl. Phys. 35 (1964) 2908–2913, not cited., pp. 2722–2727. DOI: 10.1063/1.1729798.

[1251] R. Feynman and F. Vernon. “The theory of a general quantum system interacting with a linear dissipative system”. In: *Annals of Physics* 24 (1963), pp. 118–173. DOI: 10.1016/0003-4916(63)90068-X.

[1252] J. File and R. G. Mills. “Observation of Persistent Current in a Superconducting Solenoid”. In: *Physical Review Letters* 10.3 (1963), pp. 93–96. DOI: 10.1103/PhysRevLett.10.93.

[1253] D. H. Frisch and J. H. Smith. “Measurement of the Relativistic Time Dilation Using μ-Mesons”. In: *American Journal of Physics* 31.5 (1963). Mount Washington to Cambridge muon survival; observed dilation factor 8.8 ± 0.8 against a predicted effective 8.4 ± 2, pp. 342–355. DOI: 10.1119/1.1969508.

[1254] I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Translated and edited by R. A. Silverman; canonical monograph for the chapter. NJ: Prentice-Hall, Englewood Cliffs, 1963.

[1255] R. J. Glauber. “Coherent and Incoherent States of the Radiation Field”. In: *Physical Review* 131.6 (1963). Verified against Crossref. The citing sentence is about the coherent states themselves (’rediscovered by Glauber as the states of the coherent radiation field’), which is this paper – not the companion ’The Quantum Theory of Optical Coherence’, Phys. Rev. 130, 2529 (1963), 10.1103/PhysRev.130.2529, which is about correlation functions and would be a separate key if the chapter later needs it., pp. 2766–2788. DOI: 10.1103/PhysRev.131.2766.

[1256] R. J. Glauber. “The Quantum Theory of Optical Coherence”. In: *Physical Review* 130.6 (1963), pp. 2529–2539. DOI: 10.1103/PhysRev.130.2529.

[1257] J. Hubbard. “Electron correlations in narrow energy bands”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 276.1365 (1963), pp. 238–257. DOI: 10.1098/rspa.1963.0204.

[1258] E. Jaynes and F. Cummings. “Comparison of quantum and semiclassical radiation theories with application to the beam maser”. In: *Proceedings of the IEEE* 51.1 (1963), pp. 89–109. DOI: 10.1109/PROC.1963.1664.

[1259] R. P. Kerr. “Gravitational Field of a Spinning Mass as an Example of Algebraically Special Metrics”. In: *Physical Review Letters* 11.5 (1963), pp. 237–238. DOI: 10.1103/PhysRevLett.11.237.

[1260] H. Kroemer. “A proposed class of hetero-junction injection lasers”. In: *Proceedings of the IEEE* 51.12 (1963), pp. 1782–1783. DOI: 10.1109/proc.1963.2706.

[1261] J. Linsley. “Evidence for a Primary Cosmic-Ray Particle with Energy1020eV”. In: *Physical Review Letters* 10.4 (1963), pp. 146–148. DOI: 10.1103/PhysRevLett.10.146.

[1262] E. N. Lorenz. “Deterministic Nonperiodic Flow”. In: *Journal of the Atmospheric Sciences* 20.2 (1963). The Lorenz system and sensitive dependence; the DOI contains angle brackets and must be preserved verbatim., pp. 130–141. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

[1263] S. Okubo. “φ-meson and unitary symmetry model”. In: *Physics Letters* 5.2 (1963). the O of the OZI rule, pp. 165–168. DOI: 10.1016/S0375-9601(63)92548-9.

[1264] G. S. Rushbrooke. “On the Thermodynamics of the Critical Region for the Ising Problem”. In: *The Journal of Chemical Physics* 39.3 (1963), pp. 842–843. DOI: 10.1063/1.1734338.

[1265] M. SCHMIDT. “3C 273 : A Star-Like Object with Large Red-Shift”. In: *Nature* 197.4872 (1963), pp. 1040–1040. DOI: 10.1038/1971040a0.

[1266] S. Shapiro. “Josephson Currents in Superconducting Tunneling: The Effect of Microwaves and Other Observations”. In: *Physical Review Letters* 11.2 (1963), pp. 80–82. DOI: 10.1103/PhysRevLett.11.80.

[1267] E. C. G. Sudarshan. “Equivalence of Semiclassical and Quantum Mechanical Descriptions of Statistical Light Beams”. In: *Physical Review Letters* 10.7 (1963), pp. 277–279. DOI: 10.1103/PhysRevLett.10.277.

[1268] D. T. Wilkinson and H. R. Crane. “Precision Measurement of thegFactor of the Free Electron”. In: *Physical Review* 130.3 (1963), pp. 852–863. DOI: 10.1103/PhysRev.130.852.

[1269] V. Bargmann. “Note on Wigner’s Theorem on Symmetry Operations”. In: *Journal of Mathematical Physics* 5 (1964), pp. 862–868. DOI: 10.1063/1.1704188.

[1270] V. E. Barnes et al. “Observation of a Hyperon with Strangeness Minus Three”. In: *Physical Review Letters* 12.8 (1964), pp. 204–206. DOI: 10.1103/PhysRevLett.12.204.

[1271] J. S. Bell. “On the Einstein Podolsky Rosen Paradox”. In: *Physics Physique Fizika* 1.3 (1964). The short-lived journal Physics ran 1964–1968; APS hosts the issue free of charge. The title is printed without commas, pp. 195–200. DOI: 10.1103/PhysicsPhysiqueFizika.1.195.

[1272] W. Bertozzi. “Speed and Kinetic Energy of Relativistic Electrons”. In: *American Journal of Physics* 32.7 (1964). Time-of-flight speed against kinetic energy for electrons up to 15 MeV; the speed saturates at c while the energy grows without bound, pp. 551–555. DOI: 10.1119/1.1970770.

[1273] J. D. Bjorken and S. D. Drell. *Relativistic Quantum Mechanics*. Standard monograph for Part X. McGraw-Hill, 1964.

[1274] J. H. Christenson, J. W. Cronin, V. L. Fitch, and R. Turlay. “Evidence for the2πDecay of theK20Meson”. In: *Physical Review Letters* 13.4 (1964). Not a genuine collision: the two claimed candidates are the same paper, differing only in how the title’s math was transliterated. All three citing chapters keep the bare key., pp. 138–140. DOI: 10.1103/PhysRevLett.13.138.

[1275] F. Englert and R. Brout. “Broken Symmetry and the Mass of Gauge Vector Mesons”. In: *Physical Review Letters* 13.9 (1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[1276] I. M. Gelfand and N. Y. Vilenkin. *Generalized Functions, Volume 4: Applications of Harmonic Analysis*. Russian original 1961, translated by Amiel Feinstein. Rigged Hilbert spaces / Gelfand triples and the nuclear spectral theorem. New York: Academic Press, 1964.

[1277] M. Gell-Mann. “A schematic model of baryons and mesons”. In: *Physics Letters* 8.3 (1964), pp. 214–215. DOI: 10.1016/S0031-9163(64)92001-3.

[1278] O. W. Greenberg. “Spin and Unitary-Spin Independence in a Paraquark Model of Baryons and Mesons”. In: *Physical Review Letters* 13.20 (1964), pp. 598–602. DOI: 10.1103/PhysRevLett.13.598.

[1279] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. “Global Conservation Laws and Massless Particles”. In: *Physical Review Letters* 13.20 (1964), pp. 585–587. DOI: 10.1103/PhysRevLett.13.585.

[1280] R. Haag and D. Kastler. “An Algebraic Approach to Quantum Field Theory”. In: *Journal of Mathematical Physics* 5.7 (1964), pp. 848–861. DOI: 10.1063/1.1704187.

[1281] M. Hénon and C. Heiles. “The Applicability of the Third Integral of Motion: Some Numerical Experiments”. In: *The Astronomical Journal* 69 (1964). The Hénon-Heiles system; numerical discovery of mixed regular/chaotic phase space., pp. 73–79.

[1282] P. Higgs. “Broken symmetries, massless particles and gauge fields”. In: *Physics Letters* 12.2 (1964), pp. 132–133. DOI: 10.1016/0031-9163(64)91136-9.

[1283] P. W. Higgs. “Broken Symmetries and the Masses of Gauge Bosons”. In: *Physical Review Letters* 13.16 (1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[1284] P. Hohenberg and W. Kohn. “Inhomogeneous Electron Gas”. In: *Physical Review* 136.3B (1964), B864–B871. DOI: 10.1103/PhysRev.136.B864.

[1285] R. C. Jaklevic, J. Lambe, A. H. Silver, and J. E. Mercereau. “Quantum Interference Effects in Josephson Tunneling”. In: *Physical Review Letters* 12.7 (1964), pp. 159–160. DOI: 10.1103/PhysRevLett.12.159.

[1286] G. Jona-Lasinio. “Relativistic field theories with symmetry-breaking solutions”. In: *Il Nuovo Cimento* 34.6 (1964), pp. 1790–1795. DOI: 10.1007/BF02750573.

[1287] W. E. Lamb. “Theory of an Optical Maser”. In: *Physical Review* 134.6A (1964), A1429–A1450. DOI: 10.1103/PhysRev.134.A1429.

[1288] T. D. Lee and M. Nauenberg. “Degenerate Systems and Mass Singularities”. In: *Physical Review* 133.6B (1964), B1549–B1562. DOI: 10.1103/PhysRev.133.B1549.

[1289] E. N. Leith and J. Upatnieks. “Wavefront Reconstruction with Diffused Illumination and Three-Dimensional Objects*”. In: *Journal of the Optical Society of America* 54.11 (1964). Issue 11. The laser realization: diffusely illuminated solid objects, full parallax, and the redundancy of the record (a fragment reconstructs the whole scene)., p. 1295. DOI: 10.1364/JOSA.54.001295.

[1290] J. H. Malmberg and C. B. Wharton. “Collisionless Damping of Electrostatic Plasma Waves”. In: *Physical Review Letters* 13.6 (1964). Issue 6. Crossref-confirmed. The first laboratory measurement of Landau damping., pp. 184–186. DOI: 10.1103/PhysRevLett.13.184.

[1291] A. M. L. Messiah and O. W. Greenberg. “Symmetrization Postulate and Its Experimental Foundation”. In: *Physical Review* 136.1B (1964), B248–B267. DOI: 10.1103/PhysRev.136.B248.

[1292] E. Nelson. “Feynman Integrals and the Schrödinger Equation”. In: *Journal of Mathematical Physics* 5.3 (1964), pp. 332–343. DOI: 10.1063/1.1704124.

[1293] H. E. Petschek. “Magnetic Field Annihilation”. In: *AAS-NASA Symposium on the Physics of Solar Flares*. Conference proceedings, NASA Special Publication SP-50, edited by W. N. Hess; the symposium was held at Goddard Space Flight Center, 28–30 October 1963, proceedings published 1964. No volume (it is an SP number, not a series volume) and no DOI. ADS bibcode 1964NASSP..50..425P. Carry the editor and the SP-50 report number so a reader can find it – the NASA Technical Reports Server holds a scan. 1964, pp. 425–439.

[1294] P. Roll, R. Krotkov, and R. Dicke. “The equivalence of inertial and passive gravitational mass”. In: *Annals of Physics* 26.3 (1964), pp. 442–517. DOI: 10.1016/0003-4916(64)90259-3.

[1295] A. Salam and J. Ward. “Electromagnetic and weak interactions”. In: *Physics Letters* 13.2 (1964), pp. 168–171. DOI: 10.1016/0031-9163(64)90711-5.

[1296] I. I. Shapiro. “Fourth Test of General Relativity”. In: *Physical Review Letters* 13.26 (1964), pp. 789–791. DOI: 10.1103/PhysRevLett.13.789.

[1297] R. F. Streater and A. S. Wightman. *PCT, Spin and Statistics, and All That*. W. A. Benjamin, 1964.

[1298] S. Weinberg. “Feynman Rules for Any Spin”. In: *Physical Review* 133.5B (1964), B1318–B1332. DOI: 10.1103/PhysRev.133.B1318.

[1299] S. Weinberg. “Photons and Gravitons inS-Matrix Theory: Derivation of Charge Conservation and Equality of Gravitational and Inertial Mass”. In: *Physical Review* 135.4B (1964), B1049–B1056. DOI: 10.1103/PhysRev.135.B1049.

[1300] L. Wolfenstein. “Violation ofCPInvariance and the Possibility of Very Weak Interactions”. In: *Physical Review Letters* 13.18 (1964), pp. 562–564. DOI: 10.1103/PhysRevLett.13.562.

[1301] G. Zweig. *An SU(3) model for strong interaction symmetry and its breaking*. Unpublished CERN preprint (a second part, TH.412, followed); reprinted in Developments in the Quark Theory of Hadrons, vol. 1. 1964.

[1302] F. T. Arecchi. “Measurement of the Statistical Distribution of Gaussian and Laser Sources”. In: *Physical Review Letters* 15.24 (1965), pp. 912–916. DOI: 10.1103/PhysRevLett.15.912.

[1303] S. I. Braginskii. “Transport Processes in a Plasma”. In: *Reviews of Plasma Physics, Volume 1*. NOT a proceedings paper – it is a chapter in an edited volume, so the right biblatex type is @incollection: booktitle = Reviews of Plasma Physics, volume = 1, editor = Leontovich, M. A., publisher = Consultants Bureau, address = New York, pages = 205–311. 'inproceedings' is only the closest kind the schema offers. Authorised English translation from the Russian by H. Lashinsky; the Russian original is 'Yavleniya perenosa v plazme' in *Voprosy teorii plazmy*, vyp. 1 (Gosatomizdat, Moskva, 1963) – I do not have its page range, so do not state one. No DOI exists; ADS bibcode 1965RvPP....1..205B. This is the source of the standard anisotropic transport-coefficient tables cited in the plasma chapter. 1965, pp. 205–311.

[1304] J. W. Cooley and J. W. Tukey. “An algorithm for the machine calculation of complex Fourier series”. In: *Mathematics of Computation* 19.90 (1965), pp. 297–301. DOI: 10.1090/S0025-5718-1965-0178586-1.

[1305] R. H. Dicke, P. J. E. Peebles, P. G. Roll, and D. T. Wilkinson. “Cosmic Black-Body Radiation”. In: *The Astrophysical Journal* 142 (1965). Companion interpretation paper, same issue as Penzias–Wilson., p. 414. DOI: 10.1086/148306.

[1306] R. P. Feynman and A. R. Hibbs. *Quantum Mechanics and Path Integrals*. McGraw-Hill, 1965.

[1307] L. Gårding and A. S. Wightman. “Fields as operator-valued distributions in relativistic quantum theory”. In: *Arkiv för Fysik* 28 (1965), pp. 129–184.

[1308] R. B. Griffiths. “Thermodynamic Inequality Near the Critical Point for Ferromagnets and Fluids”. In: *Physical Review Letters* 14.16 (1965), pp. 623–624. DOI: 10.1103/PhysRevLett.14.623.

[1309] M. Y. Han and Y. Nambu. “Three-Triplet Model with DoubleSU(3)Symmetry”. In: *Physical Review* 139.4B (1965), B1006–B1010. DOI: 10.1103/PhysRev.139.B1006.

[1310] L. Hedin. “New Method for Calculating the One-Particle Green's Function with Application to the Electron-Gas Problem”. In: *Physical Review* 139.3A (1965), A796–A823. DOI: 10.1103/PhysRev.139.A796.

[1311] D. C. Hodgkin. “The X-ray Analysis of Complicated Molecules”. In: *Science* 150.3699 (1965). Crossref-verified (issue 3699). Nobel Lecture in Chemistry 1964, published in Science the following year. No collision with Hodgkin:1956., pp. 979–988. DOI: 10.1126/science.150.3699.979.

[1312] R. Jost. *The General Theory of Quantized Fields*. American Mathematical Society, 1965.

[1313] I. M. Khalatnikov. *An Introduction to the Theory of Superfluidity*. W. A. Benjamin, 1965.

[1314] W. Kohn and L. J. Sham. "Self-Consistent Equations Including Exchange and Correlation Effects". In: *Physical Review* 140.4A (1965), A1133–A1138. DOI: 10.1103/PhysRev.140.A1133.

[1315] G. E. Moore. "Cramming more components onto integrated circuits". In: *Electronics* 38 (1965), pp. 114–117.

[1316] H. Mori. "Transport, Collective Motion, and Brownian Motion". In: *Progress of Theoretical Physics* 33.3 (1965), pp. 423–455. DOI: 10.1143/PTP.33.423.

[1317] J. Moser. "On the volume elements on a manifold". In: *Transactions of the American Mathematical Society* 120.2 (1965). Crossref-verified (resolved record read: Trans. Amer. Math. Soc. 120(2), 1965; Crossref deposits the first page twice as the range, the printed range is 286–294). The deformation argument now called Moser's trick, by which Darboux's theorem and the analogous normal-form results are proved. Distinct from Moser:1962, which is the twist theorem of KAM., pp. 286–294. DOI: 10.1090/S0002-9947-1965-0182927-5.

[1318] E. T. Newman, E. Couch, K. Chinnapared, A. Exton, A. Prakash, and R. Torrence. "Metric of a Rotating, Charged Mass". In: *Journal of Mathematical Physics* 6.6 (1965), pp. 918–919. DOI: 10.1063/1.1704351.

[1319] R. Penrose. "Gravitational Collapse and Space-Time Singularities". In: *Physical Review Letters* 14.3 (1965), pp. 57–59. DOI: 10.1103/PhysRevLett.14.57.

[1320] A. A. Penzias and R. W. Wilson. "A Measurement of Excess Antenna Temperature at 4080 Mc/s". In: *The Astrophysical Journal* 142 (1965), p. 419. DOI: 10.1086/148307.

[1321] R. V. Pound and J. L. Snider. "Effect of Gravity on Gamma Radiation". In: *Physical Review* 140.3B (1965), B788–B803. DOI: 10.1103/PhysRev.140.B788.

[1322] S. Weinberg. "Infrared Photons and Gravitons". In: *Physical Review* 140.2B (1965), B516–B524. DOI: 10.1103/PhysRev.140.B516.

[1323] B. Widom. "Equation of State in the Neighborhood of the Critical Point". In: *The Journal of Chemical Physics* 43.11 (1965), pp. 3898–3905. DOI: 10.1063/1.1696618.

[1324] J. S. BELL. "On the Problem of Hidden Variables in Quantum Mechanics". In: *Reviews of Modern Physics* 38.3 (1966). Written in 1964, published 1966; keyed by the publication year per house convention., pp. 447–452. DOI: 10.1103/RevModPhys.38.447.

[1325] F. A. Berezin. *The Method of Second Quantization*. Academic Press, 1966.

[1326] R. Berger. "The undecidability of the domino problem". In: *Memoirs of the American Mathematical Society* 66 (1966). Berger's 1964 Harvard thesis; the series is cited by number, not volume. 72 pp. DOI: 10.1090/memo/0066.

[1327] L. Carleson. "On convergence and growth of partial sums of Fourier series". In: *Acta Mathematica* 116.0 (1966), pp. 135–157. DOI: 10.1007/BF02392815.

[1328] S. A. Colgate and R. H. White. "The Hydrodynamic Behavior of Supernovae Explosions". In: *The Astrophysical Journal* 143 (1966), p. 626. DOI: 10.1086/148549.

[1329] A. B. Fowler, F. F. Fang, W. E. Howard, and P. J. Stiles. "Magneto-Oscillatory Conductance in Silicon Surfaces". In: *Physical Review Letters* 16.20 (1966), pp. 901–903. DOI: 10.1103/PhysRevLett.16.901.

[1330] K. Greisen. “End to the Cosmic-Ray Spectrum?” In: *Physical Review Letters* 16.17 (1966), pp. 748–750. DOI: 10.1103/PhysRevLett.16.748.

[1331] K. Hepp. “Proof of the Bogoliubov-Parasiuk theorem on renormalization”. In: *Communications in Mathematical Physics* 2.1 (1966), pp. 301–326. DOI: 10.1007/BF01773358.

[1332] P. W. Higgs. “Spontaneous Symmetry Breakdown without Massless Bosons”. In: *Physical Review* 145.4 (1966), pp. 1156–1163. DOI: 10.1103/PhysRev.145.1156.

[1333] J. Iizuka. “A Systematics and Phenomenology of Meson Family”. In: *Progress of Theoretical Physics Supplement* 37 (1966), pp. 21–34. DOI: 10.1143/PTPS.37.21.

[1334] L. P. Kadanoff. “Scaling laws for ising models near Tc”. In: *Physics Physique Fizika* 2.6 (1966), pp. 263–272. DOI: 10.1103/PhysicsPhysiqueFizika.2.263.

[1335] T. Kato. *Perturbation Theory for Linear Operators*. First edition, 1966; Grundlehren der mathematischen Wissenschaften, Band 132. This is the source of the rigorous analytic-families-of-operators statement of perturbation theory that the citing sentence invokes. Given deliberately without a DOI: the widely indexed Springer DOI belongs to the 1976 second edition / 1995 reprint, not to the 1966 first edition, and I will not guess which one a validator would resolve. If the writer prefers a checkable route, cite the 1995 Classics in Mathematics reprint explicitly rather than attaching its DOI to a 1966 entry. Springer-Verlag, Berlin, Heidelberg and New York, 1966.

[1336] R. Kubo. “The fluctuation-dissipation theorem”. In: *Reports on Progress in Physics* 29.1 (1966), pp. 255–284. DOI: 10.1088/0034-4885/29/1/306.

[1337] N. D. Mermin and H. Wagner. “Absence of Ferromagnetism or Antiferromagnetism in One- or Two-Dimensional Isotropic Heisenberg Models”. In: *Physical Review Letters* 17.22 (1966), pp. 1133–1136. DOI: 10.1103/PhysRevLett.17.1133.

[1338] J. L. Synge. “The Escape of Photons from Gravitationally Intense Stars”. In: *Monthly Notices of the Royal Astronomical Society* 131.3 (1966). Photon escape cone; basis of the shadow prediction., pp. 463–466. DOI: 10.1093/mnras/131.3.463.

[1339] G. T. Zatsepin and V. A. Kuz’min. “Upper Limit of the Spectrum of Cosmic Rays”. In: *JETP Letters* 4 (1966), pp. 78–80.

[1340] G. K. Batchelor. *An Introduction to Fluid Dynamics*. Canonical monograph for the fluid-dynamics chapter. Cambridge University Press, 1967.

[1341] E. Bishop. *Foundations of Constructive Analysis*. McGraw-Hill Series in Higher Mathematics. Constructive redevelopment of analysis. Successor volume: E. Bishop and D. Bridges, Constructive Analysis, Grundlehren 279, Springer (1985), doi 10.1007/978-3-642-61667-9. New York: McGraw-Hill, 1967.

[1342] R. H. Boyer and R. W. Lindquist. “Maximal Analytic Extension of the Kerr Metric”. In: *Journal of Mathematical Physics* 8.2 (1967), pp. 265–281. DOI: 10.1063/1.1705193.

[1343] S. Coleman and J. Mandula. “All Possible Symmetries of theSMatrix”. In: *Physical Review* 159.5 (1967), pp. 1251–1256. DOI: 10.1103/PhysRev.159.1251.

[1344] B. S. DeWitt. “Quantum Theory of Gravity. I. The Canonical Theory”. In: *Physical Review* 160.5 (1967), pp. 1113–1148. DOI: 10.1103/PhysRev.160.1113.

[1345] F. J. Dyson and A. Lenard. “Stability of Matter. I”. In: *Journal of Mathematical Physics* 8.3 (1967), pp. 423–434. DOI: 10.1063/1.1705209.

[1346] U. Essmann and H. Träuble. “The direct observation of individual flux lines in type II superconductors”. In: *Physics Letters A* 24.10 (1967), pp. 526–527. DOI: 10.1016/0375-9601(67)90819-5.

[1347] L. Faddeev and V. Popov. “Feynman diagrams for the Yang-Mills field”. In: *Physics Letters B* 25.1 (1967), pp. 29–30. DOI: 10.1016/0370-2693(67)90067-6.

[1348] M. E. Fisher. “The theory of equilibrium critical phenomena”. In: *Reports on Progress in Physics* 30.2 (1967), pp. 615–730. DOI: 10.1088/0034-4885/30/2/306.

[1349] J. van Heijenoort, ed. *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Contains Russell’s letter to Frege of 16 June 1902 and Frege’s reply, with translations of the Frege, Peano, Zermelo and Gödel papers cited here. Cambridge, MA: Harvard University Press, 1967.

[1350] P. C. Hohenberg. “Existence of Long-Range Order in One and Two Dimensions”. In: *Physical Review* 158.2 (1967), pp. 383–386. DOI: 10.1103/PhysRev.158.383.

[1351] W. Israel. “Event Horizons in Static Vacuum Space-Times”. In: *Physical Review* 164.5 (1967), pp. 1776–1779. DOI: 10.1103/PhysRev.164.1776.

[1352] T. W. B. Kibble. “Symmetry Breaking in Non-Abelian Gauge Theories”. In: *Physical Review* 155.5 (1967), pp. 1554–1561. DOI: 10.1103/PhysRev.155.1554.

[1353] S. Kochen and E. Specker. “The Problem of Hidden Variables in Quantum Mechanics”. In: *Indiana University Mathematics Journal* 17.1 (1967). Customarily dated 1967 (volume 17, issue 1). The journal was renamed Indiana University Mathematics Journal and the registered DOI carries a 1968 stamp – the same trap as Kretschmann:1918. If the validator prefers the publisher year, re-key as Kochen:1968 and record the customary 1967 dating in a note., pp. 59–87. DOI: 10.1512/iumj.1968.17.17004.

[1354] C. A. Kocher and E. D. Commins. “Polarization Correlation of Photons Emitted in an Atomic Cascade”. In: *Physical Review Letters* 18.15 (1967), pp. 575–577. DOI: 10.1103/PhysRevLett.18.575.

[1355] F. PACINI. “Energy Emission from a Neutron Star”. In: *Nature* 216.5115 (1967), pp. 567–568. DOI: 10.1038/216567a0.

[1356] I. Prigogine. *Introduction to Thermodynamics of Irreversible Processes*. 3rd edition. Interscience Publishers, 1967.

[1357] R. K. Sachs and A. M. Wolfe. “Perturbations of a Cosmological Model and Angular Variations of the Microwave Background”. In: *The Astrophysical Journal* 147 (1967), p. 73. DOI: 10.1086/148982.

[1358] A. D. Sakharov. “Violation of CP invariance, C asymmetry, and baryon asymmetry of the universe”. In: *JETP Letters* 5 (1967). Russian original Pis’ma Zh. Eksp. Teor. Fiz. 5, 32 (1967). No DOI for the original; the 1991 Soviet Physics Uspekhi reprint is a different publication and is deliberately not used as the key., pp. 24–27.

[1359] S. Smale. “Differentiable dynamical systems”. In: *Bulletin of the American Mathematical Society* 73.6 (1967). The horseshoe and symbolic dynamics survey., pp. 747–817. DOI: 10.1090/S0002-9904-1967-11798-1.

[1360] Y. Varshni. “Temperature dependence of the energy gap in semiconductors”. In: *Physica* 34.1 (1967), pp. 149–154. DOI: 10.1016/0031-8914(67)90062-6.

[1361] R. V. Wagoner, W. A. Fowler, and F. Hoyle. “On the Synthesis of Elements at Very High Temperatures”. In: *The Astrophysical Journal* 148 (1967). Metadata confirmed against Crossref., pp. 3–49. DOI: 10.1086/149126.

[1362] S. Weinberg. “A Model of Leptons”. In: *Physical Review Letters* 19.21 (1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264.

[1363] R. B. N. (editor). *The Stanford Two-Mile Accelerator*. Edited volume; editor satisfies the author requirement for @book. New York: W. A. Benjamin, 1968.

[1364] H. Bacry and J.-M. Lévy-Leblond. “Possible Kinematics”. In: *Journal of Mathematical Physics* 9.10 (1968), pp. 1605–1614. DOI: 10.1063/1.1664490.

[1365] J. N. Bahcall, N. A. Bahcall, and G. Shaviv. “Present Status of the Theoretical Predictions for theCl37Solar-Neutrino Experiment”. In: *Physical Review Letters* 20.21 (1968). Printed title formats the chlorine isotope as a superscript-37 Cl, pp. 1209–1212. DOI: 10.1103/PhysRevLett.20.1209.

[1366] B. Carter. “Global Structure of the Kerr Family of Gravitational Fields”. In: *Physical Review* 174.5 (1968). Carter constant; integrability of Kerr geodesics., pp. 1559–1571. DOI: 10.1103/PhysRev.174.1559.

[1367] G. Charpak, R. Bouclier, T. Bressani, J. Favier, and Č. Zupančič. “The use of multiwire proportional counters to select and localize charged particles”. In: *Nuclear Instruments and Methods* 62.3 (1968). The multiwire proportional chamber (Nobel Prize to Charpak, 1992): each anode wire an independent proportional counter with its own amplifier, replacing the photograph with an electronic address. The last author’s name carries carons (Zupancic) and must be written with TeX accents, č, not as UTF-8 bytes. Verified against api.crossref.org 2026-08-13., pp. 262–268. DOI: 10.1016/0029-554X(68)90371-6.

[1368] R. Davis, D. S. Harmer, and K. C. Hoffman. “Search for Neutrinos from the Sun”. In: *Physical Review Letters* 20.21 (1968), pp. 1205–1209. DOI: 10.1103/PhysRevLett.20.1205.

[1369] D. J. DeRosier and A. Klug. “Reconstruction of Three Dimensional Structures from Electron Micrographs”. In: *Nature* 217 (1968). Crossref-verified. Crossref records the first author as ’DE ROSIER’, which is how Nature set it in 1968; keyed DeRosier to match Crowther:1970, where Crossref itself prints ’DeRosier’, and the later literature. The T4 bacteriophage tail reconstruction., pp. 130–134.

[1370] M. Gell-Mann, R. J. Oakes, and B. Renner. “Behavior of Current Divergences underSU3×SU3”. In: *Physical Review* 175.5 (1968), pp. 2195–2199. DOI: 10.1103/PhysRev.175.2195.

[1371] J. Glimm and A. Jaffe. “A$\lambda\phi$4Quantum Field Theory without Cutoffs. I”. In: *Physical Review* 176.5 (1968), pp. 1945–1951. DOI: 10.1103/PhysRev.176.1945.

[1372] T. GOLD. “Rotating Neutron Stars as the Origin of the Pulsating Radio Sources”. In: *Nature* 218.5143 (1968), pp. 731–732. DOI: 10.1038/218731a0.

[1373] A. HEWISH, S. J. BELL, J. D. H. PILKINGTON, P. F. SCOTT, and R. A. COLLINS. “Observation of a Rapidly Pulsating Radio Source”. In: *Nature* 217.5130 (1968), pp. 709–713. DOI: 10.1038/217709a0.

[1374] W. Israel. “Event Horizons in Static Electrovac Space-Times”. In: *Communications in Mathematical Physics* 8 (1968). The static electrovac uniqueness theorem; the vacuum case is Israel:1967. Metadata confirmed against Crossref., pp. 245–260. DOI: 10.1007/BF01645859.

[1375] D. Lynden-Bell, R. Wood, and A. Royal. “The Gravo-Thermal Catastrophe in Isothermal Spheres and the Onset of Red-Giant Structure for Stellar Systems”. In: *Monthly Notices of the Royal Astronomical Society* 138.4 (1968), pp. 495–525. DOI: 10.1093/mnras/138.4.495.

[1376] N. F. MOTT. “Metal-Insulator Transition”. In: *Reviews of Modern Physics* 40.4 (1968), pp. 677–683. DOI: 10.1103/RevModPhys.40.677.

[1377] K. Nordtvedt. “Equivalence Principle for Massive Bodies. II. Theory”. In: *Physical Review* 169.5 (1968). Part I (definitions) is Phys. Rev. 169, 1014, same issue; II carries the effect., pp. 1017–1025. DOI: 10.1103/PhysRev.169.1017.

[1378] V. I. Oseledets. “A Multiplicative Ergodic Theorem. Lyapunov Characteristic Numbers for Dynamical Systems”. In: *Trudy Moskovskogo Matematicheskogo Obshchestva* 19 (1968). In Russian. English translation: Transactions of the Moscow Mathematical Society 19 (1968) 197-231. Name also transliterated Oseledec., pp. 179–210.

[1379] W. K. H. Panofsky. “Low q^2 electrodynamics, elastic and inelastic electron (and muon) scattering”. In: *Proceedings of the 14th International Conference on High Energy Physics, Vienna*. First public announcement of the SLAC-MIT inelastic results. 1968, pp. 23–39.

[1380] L. Parker. “Particle Creation in Expanding Universes”. In: *Physical Review Letters* 21.8 (1968), pp. 562–564. DOI: 10.1103/PhysRevLett.21.562.

[1381] A. Salam. “Weak and Electromagnetic Interactions”. In: *Elementary Particle Theory: Relativistic Groups and Analyticity (Proceedings of the Eighth Nobel Symposium)*. Nobel Symposium series has no volume number; edited by N. Svartholm. No DOI. 1968, pp. 367–377.

[1382] I. I. Shapiro et al. “Fourth Test of General Relativity: Preliminary Results”. In: *Physical Review Letters* 20.22 (1968), pp. 1265–1269. DOI: 10.1103/PhysRevLett.20.1265.

[1383] J. Silk. “Cosmic Black-Body Radiation and Galaxy Formation”. In: *The Astrophysical Journal* 151 (1968), p. 459. DOI: 10.1086/149449.

[1384] D. H. Staelin and E. C. Reifenstein. “Pulsating Radio Sources near the Crab Nebula”. In: *Science* 162.3861 (1968), pp. 1481–1483. DOI: 10.1126/science.162.3861.1481.

[1385] F. W. Stecker. “Effect of Photomeson Production by the Universal Radiation Field on High-Energy Cosmic Rays”. In: *Physical Review Letters* 21.14 (1968), pp. 1016–1018. DOI: 10.1103/PhysRevLett.21.1016.

[1386] H. M. Van Horn. “Crystallization of White Dwarfs”. In: *The Astrophysical Journal* 151 (1968). Metadata confirmed against Crossref., p. 227. DOI: 10.1086/149432.

[1387] M. J. ADAMS et al. “Structure of Rhombohedral 2 Zinc Insulin Crystals”. In: *Nature* 224.5218 (1969). Crossref-verified: title, ten family names in this order, Nature 224, 491-495, 1969. Keyed by first listed author Adams per house convention, although it is Hodgkin’s insulin structure. No collision: existing keys are Adams:1915 and Adams:2005., pp. 491–495. DOI: 10.1038/224491a0.

[1388] S. L. Adler. “Axial-Vector Vertex in Spinor Electrodynamics”. In: *Physical Review* 177.5 (1969), pp. 2426–2438. DOI: 10.1103/PhysRev.177.2426.

[1389] S. L. Adler and W. A. Bardeen. “Absence of Higher-Order Corrections in the Anomalous Axial-Vector Divergence Equation”. In: *Physical Review* 182.5 (1969), pp. 1517–1536. DOI: 10.1103/PhysRev.182.1517.

[1390] G. BAYM, C. PETHICK, and D. PINES. “Superfluidity in Neutron Stars”. In: *Nature* 224.5220 (1969), pp. 673–674. DOI: 10.1038/224673a0.

[1391] J. S. Bell and R. Jackiw. “A PCAC puzzle: $\pi 0 \to \gamma\gamma$ in the σ-model”. In: *Il Nuovo Cimento A* 60.1 (1969), pp. 47–61. DOI: 10.1007/BF02823296.

[1392] C. M. Bender and T. T. Wu. “Anharmonic Oscillator”. In: *Physical Review* 184.5 (1969). The paper establishing that the Rayleigh–Schrödinger series for the anharmonic oscillator has zero radius of convergence – the ’perturbation series are asymptotic and divergent’ claim in the citing sentence. Crossref-verified., pp. 1231–1260. DOI: 10.1103/PhysRev.184.1231.

[1393] J. D. Bjorken. “Asymptotic Sum Rules at Infinite Momentum”. In: *Physical Review* 179.5 (1969). Suffix a: two 1969 Bjorken papers are cited; this is the scaling prediction, Bjorken:1969b the parton-model paper with Paschos. A bare key Bjorken:1969 was a byte-identical duplicate of this entry and was deleted (one work, one key)., pp. 1547–1553. DOI: 10.1103/PhysRev.179.1547.

[1394] J. D. Bjorken and E. A. Paschos. “Inelastic Electron-Proton andγ-Proton Scattering and the Structure of the Nucleon”. In: *Physical Review* 185.5 (1969). Suffix b: the parton-distribution formulation, pp. 1975–1982. DOI: 10.1103/PhysRev.185.1975.

[1395] E. D. Bloom et al. “High-Energy Inelastice−pScattering at 6° and 10°”. In: *Physical Review Letters* 23.16 (1969). Not a real collision: the two candidate records carry the same DOI and describe this one paper, differing only in author formatting and in spelling out "degrees". Title uses a minus sign in e−p and degree signs, per Crossref., pp. 930–934. DOI: 10.1103/PhysRevLett.23.930.

[1396] A. Bohr and B. R. Mottelson. *Nuclear Structure, Volume I: Single-Particle Motion*. W. A. Benjamin, 1969.

[1397] M. Breidenbach et al. “Observed Behavior of Highly Inelastic Electron-Proton Scattering”. In: *Physical Review Letters* 23.16 (1969), pp. 935–939. DOI: 10.1103/PhysRevLett.23.935.

[1398] C. G. Callan and D. J. Gross. “High-Energy Electroproduction and the Constitution of the Electric Current”. In: *Physical Review Letters* 22.4 (1969), pp. 156–159. DOI: 10.1103/PhysRevLett.22.156.

[1399] Y. Choquet-Bruhat and R. Geroch. “Global aspects of the Cauchy problem in general relativity”. In: *Communications in Mathematical Physics* 14.4 (1969), pp. 329–335. DOI: 10.1007/BF01645389.

[1400] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt. “Proposed Experiment to Test Local Hidden-Variable Theories”. In: *Physical Review Letters* 23.15 (1969), pp. 880–884. DOI: 10.1103/PhysRevLett.23.880.

[1401] R. P. Feynman. “Very High-Energy Collisions of Hadrons”. In: *Physical Review Letters* 23.24 (1969), pp. 1415–1417. DOI: 10.1103/PhysRevLett.23.1415.

[1402] J. Lamb Willis E. and M. O. Scully. “The photoelectric effect without photons”. In: *Polarisation, matiere et rayonnement: volume jubilaire en l’honneur d’Alfred Kastler*. Contributed chapter in a Festschrift for Alfred Kastler; cited for the honest caveat that a quantized atom in a classical field reproduces the photoelectric phenomenology. French title word with accent: matière. 1969, pp. 363–369.

[1403] D. LYNDEN-BELL. “Galactic Nuclei as Collapsed Old Quasars”. In: *Nature* 223.5207 (1969), pp. 690–694. DOI: 10.1038/223690a0.

[1404] L. W. MO and Y. S. TSAI. “Radiative Corrections to Elastic and Inelasticepandup-Scattering”. In: *Reviews of Modern Physics* 41.1 (1969), pp. 205–235. DOI: 10.1103/RevModPhys.41.205.

[1405] R. Penrose. “Gravitational Collapse: The Role of General Relativity”. In: *Rivista del Nuovo Cimento* 1 (1969). Numero Speciale I. Cosmic censorship and the Penrose process. Reprinted Gen. Relativ. Gravit. 34, 1141–1165 (2002)., pp. 252–276.

[1406] V. RADHAKRISHNAN and R. N. MANCHESTER. “Detection of a Change of State in the Pulsar PSR 0833-45”. In: *Nature* 222.5190 (1969). Vela glitch discovery letter; verify exact printed title when entering, pp. 228–229. DOI: 10.1038/222228a0.

[1407] D. Ruelle. *Statistical Mechanics: Rigorous Results*. New York: W. A. Benjamin, 1969.

[1408] G. Velo and D. Zwanziger. “Propagation and Quantization of Rarita-Schwinger Waves in an External Electromagnetic Potential”. In: *Physical Review* 186.5 (1969), pp. 1337–1341. DOI: 10.1103/PhysRev.186.1337.

[1409] G. Velo and D. Zwanzinger. “Noncausality and Other Defects of Interaction Lagrangians for Particles with Spin One and Higher”. In: *Physical Review* 188.5 (1969), pp. 2218–2222. DOI: 10.1103/PhysRev.188.2218.

[1410] K. G. Wilson. “Non-Lagrangian Models of Current Algebra”. In: *Physical Review* 179.5 (1969), pp. 1499–1512. DOI: 10.1103/PhysRev.179.1499.

[1411] W. Zimmermann. “Convergence of Bogoliubov’s method of renormalization in momentum space”. In: *Communications in Mathematical Physics* 15.3 (1969), pp. 208–234. DOI: 10.1007/BF01645676.

[1412] A. Ashkin. “Acceleration and Trapping of Particles by Radiation Pressure”. In: *Physical Review Letters* 24.4 (1970), pp. 156–159. DOI: 10.1103/PhysRevLett.24.156.

[1413] G. Backenstoss. “Pionic Atoms”. In: *Annual Review of Nuclear Science* 20.1 (1970), pp. 467–508. DOI: 10.1146/annurev.ns.20.120170.002343.

[1414] H. Bacry, P. Combe, and J. L. Richard. “Group-theoretical analysis of elementary particles in an external electromagnetic field”. In: *Il Nuovo Cimento A* 67.2 (1970), pp. 267–299. DOI: 10.1007/BF02725178.

[1415] C. G. Callan. “Broken Scale Invariance in Scalar Field Theory”. In: *Physical Review D* 2.8 (1970), pp. 1541–1547. DOI: 10.1103/PhysRevD.2.1541.

[1416] S. Chapman and T. G. Cowling. *The Mathematical Theory of Non-uniform Gases*. 3rd edition. Cambridge University Press, 1970.

[1417] D. Christodoulou. “Reversible and Irreversible Transformations in Black-Hole Physics”. In: *Physical Review Letters* 25 (1970). Metadata confirmed against Crossref., pp. 1596–1597. DOI: 10.1103/PhysRevLett.25.1596.

[1418] R. A. Crowther, D. J. DeRosier, and A. Klug. “The reconstruction of a three-dimensional structure from projections and its application to electron microscopy”. In: *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 317.1530 (1970). Crossref-verified; container-title copied verbatim from Crossref (note ’London. A.’, not ’London. Series A.’, for this record). Source of the sampling criterion in eq:phase-views., pp. 319–340. DOI: 10.1098/rspa.1970.0119.

[1419] H. van Dam and M. Veltman. “Massive and mass-less Yang-Mills and gravitational fields”. In: *Nuclear Physics B* 22.2 (1970), pp. 397–411. DOI: 10.1016/0550-3213(70)90416-5.

[1420] S. Deser. “Self-interaction and gauge invariance”. In: *General Relativity and Gravitation* 1.1 (1970), pp. 9–18. DOI: 10.1007/BF00759198.

[1421] S. D. Drell and T.-M. Yan. “Massive Lepton-Pair Production in Hadron-Hadron Collisions at High Energies”. In: *Physical Review Letters* 25.5 (1970), pp. 316–320. DOI: 10.1103/PhysRevLett.25.316.

[1422] L. Esaki and R. Tsu. “Superlattice and Negative Differential Conductivity in Semiconductors”. In: *IBM Journal of Research and Development* 14.1 (1970), pp. 61–65. DOI: 10.1147/rd.141.0061.

[1423] S. L. Glashow, J. Iliopoulos, and L. Maiani. “Weak Interactions with Lepton-Hadron Symmetry”. In: *Physical Review D* 2.7 (1970), pp. 1285–1292. DOI: 10.1103/PhysRevD.2.1285.

[1424] E. R. Harrison. “Fluctuations at the Threshold of Classical Cosmology”. In: *Physical Review D* 1.10 (1970), pp. 2726–2730. DOI: 10.1103/PhysRevD.1.2726.

[1425] S. W. Hawking and R. Penrose. “The singularities of gravitational collapse and cosmology”. In: *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 314.1519 (1970), pp. 529–548. DOI: 10.1098/rspa.1970.0021.

[1426] P. K. Kabir. “Test of T Invariance”. In: *Physical Review D* 2 (1970), pp. 540–542.

[1427] S. N. Kružkov. “First order quasilinear equations in several independent variables”. In: *Mathematics of the USSR-Sbornik* 10.2 (1970). Crossref-verified (resolved record read: Math. USSR-Sbornik 10(2) 217–243, 1970). The entropy-solution theory of scalar conservation laws: existence and uniqueness in the class defined by the entropy inequality, by the doubling-of-variables argument. Crossref’s deposited author string carries a replacement character for the caron; the name is transliterated here as Kružkov, and is also written Kruzhkov. Translated from Mat. Sbornik 81(123) (1970) 228–255., pp. 217–243. DOI: 10.1070/SM1970v010n02ABEH002156.

[1428] I. Lakatos. “Falsification and the Methodology of Scientific Research Programmes”. In: *Criticism and the Growth of Knowledge*. Ed. by I. Lakatos and A. Musgrave. Proceedings of the International Colloquium in the Philosophy of Science, London 1965, volume 4. Replaces the single falsifiable theory by the progressive or degenerating research programme. Cambridge: Cambridge University Press, 1970, pp. 91–196. DOI: 10.1017/cbo9781139171434.009.

[1429] Y. V. Matiyasevich. “Diofantovost’ perechislimykh mnozhestv”. In: *Doklady Akademii Nauk SSSR* 191.2 (1970). In Russian; the last step in the solution of Hilbert’s tenth problem. English translation: Enumerable sets are Diophantine, Soviet Mathematics Doklady 11 (1970), 354–357, translated by A. Doohovskoy, with errata ibid. 11, no. 6, p. vi. No DOI exists for either version; the Russian original is freely readable at Math-Net.Ru, pp. 279–282.

[1430] P. M. Pearle. “Hidden-Variable Example Based upon Data Rejection”. In: *Physical Review D* 2.8 (1970), pp. 1418–1425. DOI: 10.1103/PhysRevD.2.1418.

[1431] P. J. E. Peebles and J. T. Yu. “Primeval Adiabatic Perturbation in an Expanding Universe”. In: *The Astrophysical Journal* 162 (1970), p. 815. DOI: 10.1086/150713.

[1432] V. C. Rubin and J. Ford W. Kent. “Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions”. In: *The Astrophysical Journal* 159 (1970), p. 379. DOI: 10.1086/150317.

[1433] Y. G. Sinai. “Dynamical systems with elastic reflections”. In: *Russian Mathematical Surveys* 25.2 (1970), pp. 137–189. DOI: 10.1070/RM1970v025n02ABEH003794.

[1434] R. A. Sunyaev and Y. B. Zeldovich. “Small-scale fluctuations of relic radiation”. In: *Astrophysics and Space Science* 7.1 (1970), pp. 3–19. DOI: 10.1007/BF00653471.

[1435] K. Symanzik. “Small distance behaviour in field theory and power counting”. In: *Communications in Mathematical Physics* 18.3 (1970), pp. 227–246. DOI: 10.1007/BF01649434.

[1436] M. Takesaki. *Tomita’s Theory of Modular Hilbert Algebras and its Applications*. Springer, 1970.

[1437] R. K. Ulrich. “The Five-Minute Oscillations on the Solar Surface”. In: *The Astrophysical Journal* 162 (1970), p. 993. DOI: 10.1086/150731.

[1438] C. V. Vishveshwara. “Scattering of Gravitational Radiation by a Schwarzschild Black-hole”. In: *Nature* 227 (1970). Metadata confirmed against Crossref., pp. 936–938. DOI: 10.1038/227936a0.

[1439] V. I. Zakharov. “Linearized gravitation theory and the graviton mass”. In: *JETP Letters* 12 (1970). Russian original: Pis’ma Zh. Eksp. Teor. Fiz. 12, 447 (1970). No DOI., pp. 312–314.

[1440] H. D. Zeh. “On the interpretation of measurement in quantum theory”. In: *Foundations of Physics* 1.1 (1970), pp. 69–76. DOI: 10.1007/BF00708656.

[1441] G. Baym, C. Pethick, and P. Sutherland. “The Ground State of Matter at High Densities: Equation of State and Stellar Models”. In: *The Astrophysical Journal* 170 (1971). Metadata confirmed against Crossref., p. 299. DOI: 10.1086/151216.

[1442] V. L. Berezinskii. “Destruction of Long-range Order in One-dimensional and Two-dimensional Systems Possessing a Continuous Symmetry Group. I. Classical Systems”. In: *Soviet Physics JETP* 32 (1971), pp. 493–500.

[1443] B. Carter. “Axisymmetric Black Hole Has Only Two Degrees of Freedom”. In: *Physical Review Letters* 26.6 (1971), pp. 331–333. DOI: 10.1103/PhysRevLett.26.331.

[1444] P. G. De Gennes. “Short Range Order Effects in the Isotropic Phase of Nematics and Cholesterics”. In: *Molecular Crystals and Liquid Crystals* 12.3 (1971), pp. 193–214. DOI: 10.1080/15421407108082773.

[1445] S. Doplicher, R. Haag, and J. E. Roberts. “Local observables and particle statistics I”. In: *Communications in Mathematical Physics* 23.3 (1971), pp. 199–230. DOI: 10.1007/BF01877742.

[1446] W. Feller. *An Introduction to Probability Theory and Its Applications, Volume II*. 2nd ed. New York: Wiley, 1971.

[1447] S. W. Hawking. “Gravitational Radiation from Colliding Black Holes”. In: *Physical Review Letters* 26.21 (1971). The area theorem., pp. 1344–1346. DOI: 10.1103/PhysRevLett.26.1344.

[1448] G. Hooft. “Renormalizable Lagrangians for massive Yang-Mills fields”. In: *Nuclear Physics B* 35.1 (1971), pp. 167–188. DOI: 10.1016/0550-3213(71)90139-8.

[1449] G. Hooft. “Renormalization of massless Yang-Mills fields”. In: *Nuclear Physics B* 33.1 (1971). Crossref-verified: author, year, journal, volume and pages all confirmed against the DOI. THE PAIR IS EASY TO CONFUSE AND THE LETTERS DO NOT FOLLOW THE ORDER OF PUBLICATION. This is the massless paper, Nucl. Phys. B 33 (1971) 173, which appeared first; tHooft:1971a is the massive one, Nucl. Phys. B 35 (1971) 167. The key letters were fixed before this entry existed and are kept, because tHooft:1971a already carries six citations; cite by content, not by letter. A previous key tHooft:1971b was a byte-identical duplicate of the massive paper and was deleted, which is how the absence of the massless one came to light: the electroweak chapter had been citing the massive paper for the massless result., pp. 173–199. DOI: 10.1016/0550-3213(71)90395-6.

[1450] K. Kraus. “General state changes in quantum theory”. In: *Annals of Physics* 64.2 (1971), pp. 311–335. DOI: 10.1016/0003-4916(71)90108-4.

[1451] D. Lovelock. “The Einstein Tensor and Its Generalizations”. In: *Journal of Mathematical Physics* 12.3 (1971), pp. 498–501. DOI: 10.1063/1.1665613.

[1452] R. Lynds. “The Absorption-Line Spectrum of 4c 05.34”. In: *The Astrophysical Journal* 164 (1971). Letters section, 1 March 1971 (ADS bibcode 1971ApJ...164L..73L). Crossref prints the object name lowercased as ’4c 05.34’; the journal sets it ’4C 05.34’ — either form should match on a DOI title check. This is the discovery paper of the Lyman-alpha absorption forest., p. L73. DOI: 10.1086/180695.

[1453] D. Ruelle and F. Takens. “On the nature of turbulence”. In: *Communications in Mathematical Physics* 20.3 (1971). Coins ’strange attractor’; proposes the quasiperiodic-breakdown route to turbulence. Addendum: Comm. Math. Phys. 23 (1971) 343-344., pp. 167–192. DOI: 10.1007/BF01646553.

[1454] R. Ruffini and J. A. Wheeler. “Introducing the black hole”. In: *Physics Today* 24.1 (1971). Issue 1 (January 1971). Source of the no-hair slogan., pp. 30–41. DOI: 10.1063/1.3022513.

[1455] G. Simmons and H. Wang. *Single Crystal Elastic Constants and Calculated Aggregate Properties: A Handbook*. Compilation of measured single-crystal elastic constants; the data anchor for the moduli. MIT Press, Cambridge, MA; 2nd edition, 1971.

[1456] H. E. Stanley. *Introduction to Phase Transitions and Critical Phenomena*. Oxford University Press, 1971.

[1457] S. J. Takiff. “Rings of invariant polynomials for a class of Lie algebras”. In: *Transactions of the American Mathematical Society* 160 (1971), pp. 249–262. DOI: 10.1090/s0002-9947-1971-0281839-9.

[1458] J. Taylor. “Ward identities and charge renormalization of the Yang-Mills field”. In: *Nuclear Physics B* 33.2 (1971), pp. 436–444. DOI: 10.1016/0550-3213(71)90297-5.

[1459] J. Wess and B. Zumino. “Consequences of anomalous ward identities”. In: *Physics Letters B* 37.1 (1971), pp. 95–97. DOI: 10.1016/0370-2693(71)90582-X.

[1460] E. R. Williams, J. E. Faller, and H. A. Hill. “New Experimental Test of Coulomb’s Law: A Laboratory Upper Limit on the Photon Rest Mass”. In: *Physical Review Letters* 26.12 (1971). Verified via Crossref (PRL 26(12), 22 Mar 1971). Exactly the claim made in both citing chapters: laboratory bound on a deviation from the Coulomb exponent and the resulting photon-mass limit., pp. 721–724. DOI: 10.1103/PhysRevLett.26.721.

[1461] K. G. Wilson. “Renormalization Group and Critical Phenomena. I. Renormalization Group and the Kadanoff Scaling Picture”. In: *Physical Review B* 4.9 (1971), pp. 3174–3183. DOI: 10.1103/PhysRevB.4.3174.

[1462] K. G. Wilson. “Renormalization Group and Critical Phenomena. II. Phase-Space Cell Analysis of Critical Behavior”. In: *Physical Review B* 4.9 (1971), pp. 3184–3205. DOI: 10.1103/PhysRevB.4.3184.

[1463] G. ’t Hooft and M. Veltman. “Regularization and renormalization of gauge fields”. In: *Nuclear Physics B* 44.1 (1972), pp. 189–213. DOI: 10.1016/0550-3213(72)90279-9.

[1464] J. M. Bardeen, W. H. Press, and S. A. Teukolsky. “Rotating Black Holes: Locally Nonrotating Frames, Energy Extraction, and Scalar Synchrotron Radiation”. In: *The Astrophysical Journal* 178 (1972). Systematic treatment of circular orbits, ISCO and photon orbits for Schwarzschild and Kerr., p. 347. DOI: 10.1086/151796.

[1465] C. G. Bollini and J. J. Giambiagi. “Dimensional renorinalization : The number of dimensions as a regularizing parameter”. In: *Il Nuovo Cimento B* 12.1 (1972), pp. 20–26. DOI: 10.1007/BF02895558.

[1466] C. T. BOLTON. “Identification of Cygnus X-1 with HDE 226868”. In: *Nature* 235.5336 (1972), pp. 271–273. DOI: 10.1038/235271b0.

[1467] C. Bouchiat, J. Iliopoulos, and P. Meyer. “An anomaly-free version of Weinberg’s model”. In: *Physics Letters B* 38.7 (1972), pp. 519–523. DOI: 10.1016/0370-2693(72)90532-1.

[1468] D. G. Boulware and S. Deser. “Can Gravitation Have a Finite Range?” In: *Physical Review D* 6.12 (1972), pp. 3368–3382. DOI: 10.1103/PhysRevD.6.3368.

[1469] V. B. Braginsky and V. I. Panov. “Verification of the equivalence of inertial and gravitational mass”. In: *Soviet Physics JETP* 34 (1972). English translation; Russian original Zh. Eksp. Teor. Fiz. 61, 873 (1971). No DOI for the translation journal., pp. 463–466.

[1470] K. M. Evenson et al. “Speed of Light from Direct Frequency and Wavelength Measurements of the Methane-Stabilized Laser”. In: *Physical Review Letters* 29.19 (1972). Verified against Crossref: title, volume, pages and 1972 issue year all match. This is the frequency-times-wavelength determination that fixed c to 1.1e-9 and led to the 1983 SI definition of the metre., pp. 1346–1349. DOI: 10.1103/PhysRevLett.29.1346.

[1471] S. J. Freedman and J. F. Clauser. “Experimental Test of Local Hidden-Variable Theories”. In: *Physical Review Letters* 28.14 (1972), pp. 938–941. DOI: 10.1103/PhysRevLett.28.938.

[1472] J. I. Friedman and H. W. Kendall. “Deep Inelastic Electron Scattering”. In: *Annual Review of Nuclear Science* 22.1 (1972), pp. 203–254. DOI: 10.1146/annurev.ns.22.120172.001223.

[1473] D. Gabor. “Holography, 1948-1971”. In: *Science* 177 (1972). Gabor’s own Nobel-lecture-era review of holography from the 1948 proposal to the laser era; cited for the history rather than for the original proposal., pp. 299–313. DOI: 10.1126/science.177.4046.299.

[1474] V. N. Gribov and L. N. Lipatov. “Deep inelastic ep scattering in perturbation theory”. In: *Soviet Journal of Nuclear Physics* 15 (1972), pp. 438–450.

[1475] D. J. Gross and R. Jackiw. “Effect of Anomalies on Quasi-Renormalizable Theories”. In: *Physical Review D* 6.2 (1972), pp. 477–493. DOI: 10.1103/PhysRevD.6.477.

[1476] J. C. Hafele and R. E. Keating. “Around-the-World Atomic Clocks: Observed Relativistic Time Gains”. In: *Science* 177.4044 (1972). The companion paper to Hafele:1972a: caesium clocks flown east and west around the world. The comparison mixes the kinematic effect of Part IV with the gravitational one of Part V, pp. 168–170. DOI: 10.1126/science.177.4044.168.

[1477] J. C. Hafele and R. E. Keating. “Around-the-World Atomic Clocks: Predicted Relativistic Time Gains”. In: *Science* 177.4044 (1972), pp. 166–168. DOI: 10.1126/science.177.4044.166.

[1478] A. J. Leggett. “Interpretation of Recent Results onHe3below 3 mK: A New Liquid Phase?” In: *Physical Review Letters* 29.18 (1972), pp. 1227–1230. DOI: 10.1103/PhysRevLett.29.1227.

[1479] D. Lovelock. “The Four-Dimensionality of Space and the Einstein Tensor”. In: *Journal of Mathematical Physics* 13.6 (1972), pp. 874–876. DOI: 10.1063/1.1666069.

[1480] D. D. Osheroff, R. C. Richardson, and D. M. Lee. “Evidence for a New Phase of SolidHe3”. In: *Physical Review Letters* 28.14 (1972). The title says solid; the anomalies were later identified as transitions of the liquid, which the chapter records explicitly., pp. 885–888. DOI: 10.1103/PhysRevLett.28.885.

[1481] M. Reed and B. Simon. *Methods of Modern Mathematical Physics I: Functional Analysis*. Revised and enlarged edition is 1980; cited here in the 1972 first edition. Volume I covers chapters I–IX (topological and Hilbert spaces, bounded operators, the spectral theorem). Deficiency indices, self-adjoint extensions and the self-adjointness of Schroedinger operators are Volume II: cite Reed:1975 for those. New York: Academic Press, 1972.

[1482] C. Sagan, L. S. Sagan, and F. Drake. "A Message from Earth". In: *Science* 175.4024 (1972). The Pioneer 10 plaque, whose unit of length and time is the hyperfine transition of neutral hydrogen — a physical invariant rather than a chosen yardstick, pp. 881–884. DOI: 10.1126/science.175.4024.881.

[1483] R. Schrader. "The Maxwell group and the quantum theory of particles in classical homogeneous electromagnetic fields". In: *Fortschritte der Physik* 20.12 (1972), pp. 701–734. DOI: 10.1002/prop.19720201202.

[1484] A. A. Slavnov. "Ward identities in gauge theories". In: *Theoretical and Mathematical Physics* 10.2 (1972), pp. 99–104. DOI: 10.1007/BF01090719.

[1485] R. A. Sunyaev and Y. B. Zel'dovich. "The Observation of Relic Radiation as a Test of the Nature of X-Ray Radiation from the Clusters of Galaxies". In: *Comments on Astrophysics and Space Physics* 4 (1972). Canonical SZ-effect citation; venue has no DOIs, pp. 173–178.

[1486] J. R. Taylor. *Scattering Theory: The Quantum Theory of Nonrelativistic Collisions*. Same author as the existing Taylor:1997 (error analysis) but a different year, so no key collision. New York: John Wiley & Sons, 1972.

[1487] A. Vainshtein. "To the problem of nonvanishing gravitation mass". In: *Physics Letters B* 39.3 (1972), pp. 393–394. DOI: 10.1016/0370-2693(72)90147-5.

[1488] B. L. WEBSTER and P. MURDIN. "Cygnus X-1—a Spectroscopic Binary with a Heavy Companion ?" In: *Nature* 235.5332 (1972), pp. 37–38. DOI: 10.1038/235037a0.

[1489] K. G. Wilson and M. E. Fisher. "Critical Exponents in 3.99 Dimensions". In: *Physical Review Letters* 28.4 (1972), pp. 240–243. DOI: 10.1103/PhysRevLett.28.240.

[1490] J. W. York. "Role of Conformal Three-Geometry in the Dynamics of Gravitation". In: *Physical Review Letters* 28.16 (1972), pp. 1082–1085. DOI: 10.1103/PhysRevLett.28.1082.

[1491] Y. B. Zel'dovich. "A hypothesis, unifying the structure and the entropy of the Universe". In: *Monthly Notices of the Royal Astronomical Society* 160 (1972). MNRAS short-communication (pink) pages., 1P–3P.

[1492] G. 't Hooft. "Dimensional regularization and the renormalization group". In: *Nuclear Physics B* 61 (1973), pp. 455–468. DOI: 10.1016/0550-3213(73)90376-3.

[1493] J. M. Bardeen, B. Carter, and S. W. Hawking. "The four laws of black hole mechanics". In: *Communications in Mathematical Physics* 31.2 (1973). First author James M. Bardeen, not John Bardeen of Bardeen:1957., pp. 161–170. DOI: 10.1007/BF01645742.

[1494] J. D. Bekenstein. "Black Holes and Entropy". In: *Physical Review D* 7.8 (1973), pp. 2333–2346. DOI: 10.1103/PhysRevD.7.2333.

[1495] S. Coleman and E. Weinberg. "Radiative Corrections as the Origin of Spontaneous Symmetry Breaking". In: *Physical Review D* 7.6 (1973), pp. 1888–1910. DOI: 10.1103/PhysRevD.7.1888.

[1496] M. Davis. “Hilbert’s Tenth Problem is Unsolvable”. In: *The American Mathematical Monthly* 80.3 (1973). Exposition of the Davis–Putnam–Robinson–Matiyasevich theorem. The JSTOR alias 10.2307/2318447 redirects to this DOI, pp. 233–269. DOI: 10.1080/00029890.1973.11993265.

[1497] T. Eichten et al. “Measurement of the neutrino-nucleon and antineutrino-nucleon total cross sections”. In: *Physics Letters B* 46.2 (1973), pp. 274–280. DOI: 10.1016/0370-2693(73)90702-8.

[1498] H. Epstein and V. Glaser. “The role of locality in perturbation theory”. In: *Annales de l’Institut Henri Poincaré, Section A* 19 (1973), pp. 211–295.

[1499] H. Fritzsch, M. Gell-Mann, and H. Leutwyler. “Advantages of the color octet gluon picture”. In: *Physics Letters B* 47.4 (1973), pp. 365–368. DOI: 10.1016/0370-2693(73)90625-4.

[1500] J. Glimm and A. Jaffe. “Positivity of the $\phi 34$ Hamiltonian”. In: *Fortschritte der Physik* 21.7 (1973), pp. 327–376. DOI: 10.1002/prop.19730210702.

[1501] D. J. Gross and F. Wilczek. “Ultraviolet Behavior of Non-Abelian Gauge Theories”. In: *Physical Review Letters* 30.26 (1973), pp. 1343–1346. DOI: 10.1103/PhysRevLett.30.1343.

[1502] F. Hasert et al. “Observation of neutrino-like interactions without muon or electron in the gargamelle neutrino experiment”. In: *Physics Letters B* 46.1 (1973), pp. 138–140. DOI: 10.1016/0370-2693(73)90499-1.

[1503] F. Hasert et al. “Search for elastic muon-neutrino electron scattering”. In: *Physics Letters B* 46.1 (1973), pp. 121–124. DOI: 10.1016/0370-2693(73)90494-2.

[1504] S. W. Hawking and G. F. R. Ellis. *The Large Scale Structure of Space-Time*. Cambridge Monographs on Mathematical Physics; the DOI is the CUP ebook edition. Cambridge University Press, 1973. DOI: 10.1017/CBO9780511524646.

[1505] M. Kobayashi and T. Maskawa. “CP-Violation in the Renormalizable Theory of Weak Interaction”. In: *Progress of Theoretical Physics* 49.2 (1973), pp. 652–657. DOI: 10.1143/PTP.49.652.

[1506] J. M. Kosterlitz and D. J. Thouless. “Ordering, metastability and phase transitions in two-dimensional systems”. In: *Journal of Physics C: Solid State Physics* 6.7 (1973), pp. 1181–1203. DOI: 10.1088/0022-3719/6/7/010.

[1507] P. C. LAUTERBUR. “Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance”. In: *Nature* 242.5394 (1973). Issue 5394. Crossref-confirmed. This is the MRI spatial-encoding-by-field-gradients paper the chapter cites., pp. 190–191. DOI: 10.1038/242190a0.

[1508] E. H. Lieb and M. B. Ruskai. “Proof of the strong subadditivity of quantum-mechanical entropy”. In: *Journal of Mathematical Physics* 14.12 (1973), pp. 1938–1941. DOI: 10.1063/1.1666274.

[1509] C. W. Misner, K. S. Thorne, and J. A. Wheeler. *Gravitation*. San Francisco: W. H. Freeman, 1973.

[1510] K. Osterwalder and R. Schrader. “Axioms for Euclidean Green’s functions”. In: *Communications in Mathematical Physics* 31.2 (1973), pp. 83–112. DOI: 10.1007/BF01645738.

[1511] H. D. Politzer. “Reliable Perturbative Results for Strong Interactions?” In: *Physical Review Letters* 30.26 (1973), pp. 1346–1349. DOI: 10.1103/PhysRevLett.30.1346.

[1512] N. I. Shakura and R. A. Sunyaev. “Black Holes in Binary Systems: Observational Appearances”. In: *X- and Gamma-Ray Astronomy* 24 (1973). Pre-2001 A&A; no Crossref DOI, pp. 155–164. DOI: 10.1007/978-94-010-2585-0_13.

[1513] L. Smarr. “Mass Formula for Kerr Black Holes”. In: *Physical Review Letters* 30 (1973). Erratum: Phys. Rev. Lett. 30, 521 (1973). Metadata confirmed against Crossref., pp. 71–73. DOI: 10.1103/PhysRevLett.30.71.

[1514] G. ’t Hooft and M. Veltman. “One-loop divergencies in the theory of gravitation”. In: *Annales de l’Institut Henri Poincaré A: Physique Théorique* 20 (1974). No DOI; the journal predates DOI registration for this series., pp. 69–94.

[1515] J. J. Aubert et al. “Experimental Observation of a Heavy ParticleJ”. In: *Physical Review Letters* 33.23 (1974), pp. 1404–1406. DOI: 10.1103/PhysRevLett.33.1404.

[1516] J. .-. Augustin et al. “Discovery of a Narrow Resonance ine+e−Annihilation”. In: *Physical Review Letters* 33.23 (1974), pp. 1406–1408. DOI: 10.1103/PhysRevLett.33.1406.

[1517] B. Balick and R. L. Brown. “Intense sub-arcsecond structure in the galactic center”. In: *The Astrophysical Journal* 194 (1974). Discovery of Sgr A*, p. 265. DOI: 10.1086/153242.

[1518] C. Becchi, A. Rouet, and R. Stora. “The abelian Higgs Kibble model, unitarity of the S-operator”. In: *Physics Letters B* 52.3 (1974), pp. 344–346. DOI: 10.1016/0370-2693(74)90058-6.

[1519] M. Bouchiat and C. Bouchiat. “Weak neutral currents in atomic physics”. In: *Physics Letters B* 48.2 (1974), pp. 111–114. DOI: 10.1016/0370-2693(74)90656-X.

[1520] B. J. Carr and S. W. Hawking. “Black Holes in the Early Universe”. In: *Monthly Notices of the Royal Astronomical Society* 168.2 (1974), pp. 399–415. DOI: 10.1093/mnras/168.2.399.

[1521] J. F. Clauser. “Experimental distinction between the quantum and classical field-theoretic predictions for the photoelectric effect”. In: *Physical Review D* 9.4 (1974). Single-author paper, issue of 15 February 1974; takes the ’a’ suffix as the earlier of the two Clauser 1974 papers. The coincidence test of the photoelectric effect against semiclassical (classical-field) theory., pp. 853–860. DOI: 10.1103/PhysRevD.9.853.

[1522] J. F. Clauser and M. A. Horne. “Experimental consequences of objective local theories”. In: *Physical Review D* 10.2 (1974). Issue of 15 July 1974. The Clauser–Horne inequality, formulated for detection probabilities rather than correlations., pp. 526–535. DOI: 10.1103/PhysRevD.10.526.

[1523] E. B. Davies. “Markovian master equations”. In: *Communications in Mathematical Physics* 39.2 (1974), pp. 91–110. DOI: 10.1007/BF01608389.

[1524] R. Dingle, W. Wiegmann, and C. H. Henry. “Quantum States of Confined Carriers in Very ThinAlxGa1−xAs-GaAs-AlxGa1−xAsHeterostructures”. In: *Physical Review Letters* 33.14 (1974), pp. 827–830. DOI: 10.1103/PhysRevLett.33.827.

[1525] S. Doplicher, R. Haag, and J. E. Roberts. “Local observables and particle statistics II”. In: *Communications in Mathematical Physics* 35.1 (1974), pp. 49–85. DOI: 10.1007/BF01646454.

[1526] M. E. Fisher. “The renormalization group in the theory of critical behavior”. In: *Reviews of Modern Physics* 46.4 (1974), pp. 597–616. DOI: 10.1103/RevModPhys.46.597.

[1527] M. K. Gaillard and B. W. Lee. “Rare decay modes of the K mesons in gauge theories”. In: *Physical Review D* 10.3 (1974). The charm-quark mass estimated from the K_L–K_S mass difference through the GIM box. Metadata confirmed against Crossref., pp. 897–916. DOI: 10.1103/PhysRevD.10.897.

[1528] G. C. Hegerfeldt. “Remark on causality and particle localization”. In: *Physical Review D* 10.10 (1974), pp. 3320–3321. DOI: 10.1103/PhysRevD.10.3320.

[1529] J. P. Ostriker, P. J. E. Peebles, and A. Yahil. “The size and mass of galaxies, and the mass of the universe”. In: *The Astrophysical Journal* 193 (1974), p. L1. DOI: 10.1086/181617.

[1530] A. W. Overhauser and R. Colella. “Experimental Test of Gravitationally Induced Quantum Interference”. In: *Physical Review Letters* 33.20 (1974). Issue 20; published 11 November 1974. This is the PROPOSAL, not the measurement: the existing Colella:1975 (PRL 34, 1472) is the observation a year later. Author order is reversed between the two papers – the proposal is by Overhauser and Colella, the observation by Colella, Overhauser and Werner – so keying each by its own first listed author gives Overhauser:1974 and Colella:1975 with no collision. The chapter cites it at the Interpretation sentence ’The 1974 proposal that preceded the measurement is ...’ in exp:expmw-cow. Verified via Crossref (title, volume 33, issue 20, pp. 1237-1239, issued 1974)., pp. 1237–1239. DOI: 10.1103/PhysRevLett.33.1237.

[1531] D. N. Page and K. S. Thorne. “Disk-Accretion onto a Black Hole. Time-Averaged Structure of Accretion Disk”. In: *The Astrophysical Journal* 191 (1974). The time-averaged radial flux profile of a steady thin disc with a zero-torque inner boundary at the innermost stable circular orbit. Crossref deposits only the first page. Metadata confirmed against Crossref., p. 499. DOI: 10.1086/152990.

[1532] H. Rauch, W. Treimer, and U. Bonse. “Test of a single crystal neutron interferometer”. In: *Physics Letters A* 47.5 (1974), pp. 369–371. DOI: 10.1016/0375-9601(74)90132-7.

[1533] C. E. Rhoades and R. Ruffini. “Maximum Mass of a Neutron Star”. In: *Physical Review Letters* 32.6 (1974), pp. 324–327. DOI: 10.1103/PhysRevLett.32.324.

[1534] L. P. S. Singh and C. R. Hagen. “Lagrangian formulation for arbitrary spin. I. The boson case”. In: *Physical Review D* 9.4 (1974), pp. 898–909. DOI: 10.1103/PhysRevD.9.898.

[1535] J. Wess and B. Zumino. “Supergauge transformations in four dimensions”. In: *Nuclear Physics B* 70.1 (1974). Cited solely to name an excluded programme and record the collider nulls against it, per editorial scope rule 1., pp. 39–50. DOI: 10.1016/0550-3213(74)90355-1.

[1536] K. Wilson. “The renormalization group and the ε expansion”. In: *Physics Reports* 12.2 (1974), pp. 75–199. DOI: 10.1016/0370-1573(74)90023-4.

[1537] K. G. Wilson. “Confinement of quarks”. In: *Physical Review D* 10.8 (1974). Formerly duplicated under a second key Wilson:1974a, merged here (one work, one key)., pp. 2445–2459. DOI: 10.1103/PhysRevD.10.2445.

[1538] T. Appelquist and J. Carazzone. “Infrared singularities and massive fields”. In: *Physical Review D* 11.10 (1975), pp. 2856–2861. DOI: 10.1103/PhysRevD.11.2856.

[1539] A. Belavin, A. Polyakov, A. Schwartz, and Y. Tyupkin. “Pseudoparticle solutions of the Yang-Mills equations”. In: *Physics Letters B* 59.1 (1975), pp. 85–87. DOI: 10.1016/0370-2693(75)90163-X.

[1540] J. J. Bisognano and E. H. Wichmann. “On the duality condition for a Hermitian scalar field”. In: *Journal of Mathematical Physics* 16.4 (1975), pp. 985–1007. DOI: 10.1063/1.522605.

[1541] M.-D. Choi. “Completely positive linear maps on complex matrices”. In: *Linear Algebra and its Applications* 10.3 (1975), pp. 285–290. DOI: 10.1016/0024-3795(75)90075-0.

[1542] R. Colella, A. W. Overhauser, and S. A. Werner. “Observation of Gravitationally Induced Quantum Interference”. In: *Physical Review Letters* 34.23 (1975). The COW experiment, keyed by its first listed author to avoid any collision with the existing Werner:1975 entry (the neutron spin-precession paper of the same year)., pp. 1472–1474. DOI: 10.1103/PhysRevLett.34.1472.

[1543] F.-L. Deubner. “Observations of low wavenumber nonradial eigenmodes of the sun”. In: *Astronomy and Astrophysics* 44 (1975). Pre-2001 A&A; no Crossref DOI, pp. 371–375.

[1544] P. Feyerabend. *Against Method: Outline of an Anarchistic Theory of Knowledge*. The case that no proposed rule of scientific method has gone unviolated in some successful episode. London: New Left Books, 1975.

[1545] J. P. Gollub and H. L. Swinney. “Onset of Turbulence in a Rotating Fluid”. In: *Physical Review Letters* 35.14 (1975). Taylor-Couette spectra deciding for the Ruelle-Takens picture against Landau’s., pp. 927–930. DOI: 10.1103/PhysRevLett.35.927.

[1546] T. W. Hänsch, S. A. Lee, R. Wallenstein, and C. Wieman. “Doppler-Free Two-Photon Spectroscopy of Hydrogen1S−2S”. In: *Physical Review Letters* 34.6 (1975). Suffix a/b needed: two Hänsch 1975 papers cited., pp. 307–309. DOI: 10.1103/PhysRevLett.34.307.

[1547] T. Hänsch and A. Schawlow. “Cooling of gases by laser radiation”. In: *Optics Communications* 13.1 (1975). Suffix a/b needed: two Hänsch 1975 papers cited. Key transliterates the umlaut., pp. 68–69. DOI: 10.1016/0030-4018(75)90159-5.

[1548] G. Hanson et al. “Evidence for Jet Structure in Hadron Production bye+e−Annihilation”. In: *Physical Review Letters* 35.24 (1975), pp. 1609–1612. DOI: 10.1103/PhysRevLett.35.1609.

[1549] S. W. Hawking. “Particle creation by black holes”. In: *Communications In Mathematical Physics* 43.3 (1975). Cited with an explicit unobserved-prediction status in the chapter, per editorial rule 1., pp. 199–220. DOI: 10.1007/BF02345020.

[1550] R. A. Hulse and J. H. Taylor. “Discovery of a pulsar in a binary system”. In: *The Astrophysical Journal* 195 (1975), pp. L51–L53. DOI: 10.1086/181708.

[1551] M. Jullière. “Tunneling between ferromagnetic films”. In: *Physics Letters A* 54 (1975), pp. 225–226.

[1552] L. D. Landau and E. M. Lifshitz. *The Classical Theory of Fields*. Course of Theoretical Physics, vol. 2. Distinct from existing key Landau:1976 (Mechanics, vol. 1). Pergamon Press, 4th revised English edition, 1975.

[1553] T.-Y. Li and J. A. Yorke. “Period Three Implies Chaos”. In: *The American Mathematical Monthly* 82.10 (1975). Origin of the mathematical term ’chaos’. Distinct from existing key Li:2006 (Nan Li) and new key Li:2018 (Qing Li)., pp. 985–992. DOI: 10.1080/00029890.1975.11994008.

[1554] E. H. Lieb and W. E. Thirring. “Bound for the Kinetic Energy of Fermions Which Proves the Stability of Matter”. In: *Physical Review Letters* 35.11 (1975), pp. 687–689. DOI: 10.1103/PhysRevLett.35.687.

[1555] K. Osterwalder and R. Schrader. “Axioms for Euclidean Green’s functions II”. In: *Communications in Mathematical Physics* 42.3 (1975), pp. 281–305. DOI: 10.1007/BF01608978.

[1556] M. L. Perl et al. “Evidence for Anomalous Lepton Production ine+−e−Annihilation”. In: *Physical Review Letters* 35.22 (1975), pp. 1489–1492. DOI: 10.1103/PhysRevLett.35.1489.

[1557] M. E. Phelps, E. J. Hoffman, N. A. Mullani, and M. M. Ter-Pogossian. “Application of annihilation coincidence detection to transaxial reconstruction tomography”. In: *Journal of Nuclear Medicine* 16 (1975). Companion paper to TerPogossian:1975., pp. 210–224.

[1558] H. Rauch, A. Zeilinger, G. Badurek, A. Wilfing, W. Bauspiess, and U. Bonse. “Verification of coherent spinor rotation of fermions”. In: *Physics Letters A* 54.6 (1975). Independent of the existing Werner:1975 entry, which reports the same 4-pi periodicity from the Missouri group., pp. 425–427. DOI: 10.1016/0375-9601(75)90798-7.

[1559] M. Reed and B. Simon. *Methods of Modern Mathematical Physics II: Fourier Analysis, Self-Adjointness*. Volume II of the series begun in Reed:1972. Carries the Fourier transform on tempered distributions, von Neumann’s theory of self-adjoint extensions and deficiency indices (chapter X), and the potential-by-potential self-adjointness results (Kato–Rellich and successors). Printed monograph, no DOI. New York: Academic Press, 1975.

[1560] D. C. Robinson. “Uniqueness of the Kerr Black Hole”. In: *Physical Review Letters* 34.14 (1975), pp. 905–906. DOI: 10.1103/PhysRevLett.34.905.

[1561] D. Ross and M. Veltman. “Neutral currents and the Higgs mechanism”. In: *Nuclear Physics B* 95.1 (1975), pp. 135–147. DOI: 10.1016/0550-3213(75)90485-X.

[1562] R. F. Streater. “Outline of axiomatic relativistic quantum field theory”. In: *Reports on Progress in Physics* 38.7 (1975), pp. 771–846. DOI: 10.1088/0034-4885/38/7/001.

[1563] M. M. Ter-Pogossian, M. E. Phelps, E. J. Hoffman, and N. A. Mullani. “A Positron-Emission Transaxial Tomograph for Nuclear Imaging (PETT)”. In: *Radiology* 114.1 (1975). Key drops the hyphen to keep the citekey alphanumeric., pp. 89–98. DOI: 10.1148/114.1.89.

[1564] I. V. Tyutin. *Gauge Invariance in Field Theory and Statistical Physics in Operator Formalism*. Reprinted as arXiv:0812.0580. 1975.

[1565] S. Weinberg. “The U(1) problem”. In: *Physical Review D* 11.12 (1975). Source of the bound $m \leq \sqrt{3}\, m_\pi$ on a would-be $U(1)_A$ Goldstone boson. Metadata confirmed against Crossref., pp. 3583–3593. DOI: 10.1103/PhysRevD.11.3583.

[1566] S. A. Werner, R. Colella, A. W. Overhauser, and C. F. Eagen. “Observation of the Phase Shift of a Neutron Due to Precession in a Magnetic Field”. In: *Physical Review Letters* 35 (1975), pp. 1053–1055. DOI: 10.1103/PhysRevLett.35.1053.

[1567] K. G. Wilson. “The renormalization group: Critical phenomena and the Kondo problem”. In: *Reviews of Modern Physics* 47.4 (1975), pp. 773–840. DOI: 10.1103/RevModPhys.47.773.

[1568] G. ’t Hooft. “Computation of the quantum effects due to a four-dimensional pseudoparticle”. In: *Physical Review D* 14.12 (1976), pp. 3432–3450. DOI: 10.1103/PhysRevD.14.3432.

[1569] G. ’t Hooft. “Symmetry Breaking through Bell-Jackiw Anomalies”. In: *Physical Review Letters* 37.1 (1976), pp. 8–11. DOI: 10.1103/PhysRevLett.37.8.

[1570] N. W. Ashcroft and N. D. Mermin. *Solid State Physics*. Rinehart and Winston: Holt, 1976.

[1571] C. Becchi, A. Rouet, and R. Stora. "Renormalization of gauge theories". In: *Annals of Physics* 98.2 (1976), pp. 287–321. DOI: 10.1016/0003-4916(76)90156-1.

[1572] J. J. Bisognano and E. H. Wichmann. "On the duality condition for quantum fields". In: *Journal of Mathematical Physics* 17.3 (1976), pp. 303–321. DOI: 10.1063/1.522898.

[1573] C. Callan, R. Dashen, and D. Gross. "The structure of the gauge theory vacuum". In: *Physics Letters B* 63.3 (1976), pp. 334–340. DOI: 10.1016/0370-2693(76)90277-X.

[1574] H. J. Carmichael and D. F. Walls. "Proposal for the measurement of the resonant Stark effect by photon correlation techniques". In: *Journal of Physics B: Atomic and Molecular Physics* 9.4 (1976), pp. L43–L46. DOI: 10.1088/0022-3700/9/4/001.

[1575] M. P. do Carmo. *Differential Geometry of Curves and Surfaces*. Englewood Cliffs, New Jersey: Prentice-Hall, 1976.

[1576] J. F. Clauser. "Experimental Investigation of a Polarization Correlation Anomaly". In: *Physical Review Letters* 36.21 (1976), pp. 1223–1226. DOI: 10.1103/PhysRevLett.36.1223.

[1577] B. d'Espagnat. *Conceptual Foundations of Quantum Mechanics*. 2nd edition. Source of the proper/improper mixture distinction. W. A. Benjamin, 1976.

[1578] E. S. Fry and R. C. Thompson. "Experimental Test of Local Hidden-Variable Theories". In: *Physical Review Letters* 37.8 (1976). Shares its title with Freedman–Clauser 1972; distinct experiment (mercury cascade)., pp. 465–468. DOI: 10.1103/PhysRevLett.37.465.

[1579] V. Gorini, A. Kossakowski, and E. C. G. Sudarshan. "Completely positive dynamical semigroups of N-level systems". In: *Journal of Mathematical Physics* 17.5 (1976), pp. 821–825. DOI: 10.1063/1.522979.

[1580] S. W. Hawking. "Breakdown of predictability in gravitational collapse". In: *Physical Review D* 14.10 (1976), pp. 2460–2473. DOI: 10.1103/PhysRevD.14.2460.

[1581] M. Hénon. "A two-dimensional mapping with a strange attractor". In: *Communications in Mathematical Physics* 50.1 (1976). The Hénon map., pp. 69–77. DOI: 10.1007/BF01608556.

[1582] R. Jackiw and C. Rebbi. "Vacuum Periodicity in a Yang-Mills Quantum Theory". In: *Physical Review Letters* 37.3 (1976), pp. 172–175. DOI: 10.1103/PhysRevLett.37.172.

[1583] D. Kahng. "A historical perspective on the development of MOS transistors and related devices". In: *IEEE Transactions on Electron Devices* 23.7 (1976), pp. 655–657. DOI: 10.1109/t-ed.1976.18468.

[1584] J. Kilby. "Invention of the integrated circuit". In: *IEEE Transactions on Electron Devices* 23.7 (1976), pp. 648–654. DOI: 10.1109/t-ed.1976.18467.

[1585] L. D. Landau and E. M. Lifshitz. *Mechanics*. 3rd ed. Vol. 1. Course of Theoretical Physics. Oxford: Butterworth-Heinemann, 1976.

[1586] G. Lindblad. "On the generators of quantum dynamical semigroups". In: *Communications in Mathematical Physics* 48.2 (1976), pp. 119–130. DOI: 10.1007/BF01608499.

[1587] R. M. May. "Simple mathematical models with very complicated dynamics". In: *Nature* 261.5560 (1976). The logistic-map review that made the map canonical., pp. 459–467. DOI: 10.1038/261459a0.

[1588] P. G. Merli, G. F. Missiroli, and G. Pozzi. “On the statistical aspect of electron interference phenomena”. In: *American Journal of Physics* 44.3 (1976). Priority for single-electron buildup of an interference pattern, thirteen years before Tonomura:1989., pp. 306–307. DOI: 10.1119/1.10184.

[1589] B. W. Roberts. “Survey of superconductive materials and critical evaluation of selected properties”. In: *Journal of Physical and Chemical Reference Data* 5.3 (1976), pp. 581–822. DOI: 10.1063/1.555540.

[1590] W. G. Unruh. “Notes on black-hole evaporation”. In: *Physical Review D* 14.4 (1976), pp. 870–892. DOI: 10.1103/PhysRevD.14.870.

[1591] G. Altarelli and G. Parisi. “Asymptotic freedom in parton language”. In: *Nuclear Physics B* 126.2 (1977), pp. 298–318. DOI: 10.1016/0550-3213(77)90384-4.

[1592] J. Bailey et al. “Measurements of relativistic time dilatation for positive and negative muons in a circular orbit”. In: *Nature* 268.5618 (1977). CERN Muon Storage Ring at $\gamma = 29.33$: dilated lifetimes $\tau_+ = (64.419 \pm 0.058)\,\mu s$ and $\tau_- = (64.368 \pm 0.029)\,\mu s$, confirming the dilation factor to 2×10^{-3} at 95% confidence, pp. 301–305. DOI: 10.1038/268301a0.

[1593] C. Cohen-Tannoudji, B. Diu, and F. Laloë. *Quantum Mechanics*. Two volumes. New York: Wiley, 1977.

[1594] R. B. Davies. “Hypothesis testing when a nuisance parameter is present only under the alternative”. In: *Biometrika* 64.2 (1977). Crossref-verified. The bound on the global p-value in terms of the expected number of upcrossings, for the case where a parameter (a mass, in a bump search) has no meaning under the null hypothesis, so Wilks:1938 does not apply. DISTINGUISH from Davies:1974, which is E. B. Davies on Markovian master equations: a different author and a different subject., pp. 247–254. DOI: 10.2307/2335690.

[1595] Dokshitzer. “Calculation of the Structure Functions for Deep Inelastic Scattering and e+e- Annihilation by Perturbation Theory in Quantum Chromodynamics”. In: *Soviet Physics JETP* 46 (1977). Russian original Zh. Eksp. Teor. Fiz. 73, 1216 (1977); no DOI, pp. 641–653.

[1596] G. W. Gibbons and S. W. Hawking. “Action integrals and partition functions in quantum gravity”. In: *Physical Review D* 15.10 (1977), pp. 2752–2756. DOI: 10.1103/PhysRevD.15.2752.

[1597] S. W. Herb et al. “Observation of a Dimuon Resonance at 9.5 GeV in 400-GeV Proton-Nucleus Collisions”. In: *Physical Review Letters* 39.5 (1977), pp. 252–255. DOI: 10.1103/PhysRevLett.39.252.

[1598] H. J. Kimble, M. Dagenais, and L. Mandel. “Photon Antibunching in Resonance Fluorescence”. In: *Physical Review Letters* 39.11 (1977), pp. 691–695. DOI: 10.1103/PhysRevLett.39.691.

[1599] L. D. Landau and E. M. Lifshitz. *Quantum Mechanics: Non-Relativistic Theory*. Course of Theoretical Physics, Volume 3; third edition, revised and enlarged, 1977, translated from the Russian by J. B. Sykes and J. S. Bell. Cited as one of the standard treatments alongside Sakurai:2017 and CohenTannoudji:1977, so the English Pergamon edition is the right target even though the original language is Russian – record edition = 3 and the translators. Printed monograph with no DOI: expect an honest AMBER on availability. Oxford: Pergamon Press, 1977.

[1600] J. C. Le Guillou and J. Zinn-Justin. “Critical Exponents for then-Vector Model in Three Dimensions from Field Theory”. In: *Physical Review Letters* 39.2 (1977), pp. 95–98. DOI: 10.1103/PhysRevLett.39.95.

[1601] B. W. Lee, C. Quigg, and H. B. Thacker. “Weak interactions at very high energies: The role of the Higgs-boson mass”. In: *Physical Review D* 16.5 (1977). Tree-level unitarity of longitudinal W scattering and the resulting upper bound on the Higgs mass; the full paper following the companion letter PRL 38, 883 (1977). Split from the former shared key Lee:1977. Metadata confirmed against Crossref., pp. 1519–1531. DOI: 10.1103/PhysRevD.16.1519.

[1602] B. W. Lee and S. Weinberg. “Cosmological Lower Bound on Heavy-Neutrino Masses”. In: *Physical Review Letters* 39.4 (1977). The thermal-relic (Lee–Weinberg) bound. Split from the former shared key Lee:1977; suffixed a/b because both are landmark Benjamin W. Lee papers of 1977 and neither is secondary, with a/b in publication-date order (PRL July, PRD September). Metadata confirmed against Crossref., pp. 165–168. DOI: 10.1103/PhysRevLett.39.165.

[1603] J. M. Leinaas and J. Myrheim. “On the theory of identical particles”. In: *Il Nuovo Cimento B Series 11* 37.1 (1977), pp. 1–23. DOI: 10.1007/BF02727953.

[1604] E. H. Lieb and B. Simon. “The Thomas-Fermi theory of atoms, molecules and solids”. In: *Advances in Mathematics* 23.1 (1977), pp. 22–116. DOI: 10.1016/0001-8708(77)90108-6.

[1605] P. Minkowski. “$\mu \to e\gamma$ at a rate of one out of 109 muon decays?” In: *Physics Letters B* 67.4 (1977). Cited in the treatise only as a hypothesis explicitly flagged as lacking evidence, pp. 421–428. DOI: 10.1016/0370-2693(77)90435-X.

[1606] B. Misra and E. C. G. Sudarshan. “The Zeno’s paradox in quantum theory”. In: *Journal of Mathematical Physics* 18.4 (1977), pp. 756–763. DOI: 10.1063/1.523304.

[1607] D. R. Nelson and J. M. Kosterlitz. “Universal Jump in the Superfluid Density of Two-Dimensional Superfluids”. In: *Physical Review Letters* 39.19 (1977), pp. 1201–1205. DOI: 10.1103/PhysRevLett.39.1201.

[1608] J. Paris and L. Harrington. “A Mathematical Incompleteness in Peano Arithmetic”. In: *Handbook of Mathematical Logic*. Ed. by J. Barwise. Vol. 90. Studies in Logic and the Foundations of Mathematics. Amsterdam: North-Holland, 1977, pp. 1133–1142. DOI: 10.1016/S0049-237X(08)71130-3.

[1609] R. D. Peccei and H. R. Quinn. “CPConservation in the Presence of Pseudoparticles”. In: *Physical Review Letters* 38.25 (1977). Cited only alongside an explicit statement that no axion has been detected, pp. 1440–1443. DOI: 10.1103/PhysRevLett.38.1440.

[1610] G. F. Smoot, M. V. Gorenstein, and R. A. Muller. “Detection of Anisotropy in the Cosmic Blackbody Radiation”. In: *Physical Review Letters* 39.14 (1977). CMB dipole from U-2 aircraft, pp. 898–901. DOI: 10.1103/PhysRevLett.39.898.

[1611] G. Sterman and S. Weinberg. “Jets from Quantum Chromodynamics”. In: *Physical Review Letters* 39.23 (1977). The infrared- and collinear-safe jet cross-section definition. Metadata confirmed against Crossref., pp. 1436–1439. DOI: 10.1103/PhysRevLett.39.1436.

[1612] M. Veltman. “Limit on mass differences in the Weinberg model”. In: *Nuclear Physics B* 123.1 (1977), pp. 89–99. DOI: 10.1016/0550-3213(77)90342-X.

[1613] W. A. Bardeen, A. J. Buras, D. W. Duke, and T. Muta. “Deep-inelastic scattering beyond the leading order in asymptotically free gauge theories”. In: *Physical Review D* 18.11 (1978), pp. 3998–4017. DOI: 10.1103/PhysRevD.18.3998.

[1614] A. R. Bell. “The acceleration of cosmic rays in shock fronts - I”. In: *Monthly Notices of the Royal Astronomical Society* 182.2 (1978). A. R. Bell; distinct from J. S. Bell of Bell:1964, pp. 147–156. DOI: 10.1093/mnras/182.2.147.

[1615] D. J. Bishop and J. D. Reppy. “Study of the Superfluid Transition in Two-DimensionalHe4Films”. In: *Physical Review Letters* 40.26 (1978), pp. 1727–1730. DOI: 10.1103/PhysRevLett.40.1727.

[1616] R. D. Blandford and J. P. Ostriker. “Particle acceleration by astrophysical shocks”. In: *The Astrophysical Journal* 221 (1978), p. L29. DOI: 10.1086/182658.

[1617] L. M. Brown. “The idea of the neutrino”. In: *Physics Today* 31.9 (1978), pp. 23–28. DOI: 10.1063/1.2995181.

[1618] R. Dingle, H. L. Störmer, A. C. Gossard, and W. Wiegmann. “Electron mobilities in modulation-doped semiconductor heterojunction superlattices”. In: *Applied Physics Letters* 33.7 (1978), pp. 665–667. DOI: 10.1063/1.90457.

[1619] P. H. Eberhard. “Bell’s theorem and the different concepts of locality”. In: *Il Nuovo Cimento B Series 11* 46.2 (1978), pp. 392–419. DOI: 10.1007/BF02728628.

[1620] J. Fang and C. Fronsdal. “Massless fields with half-integral spin”. In: *Physical Review D* 18.10 (1978), pp. 3630–3633. DOI: 10.1103/PhysRevD.18.3630.

[1621] M. J. Feigenbaum. “Quantitative universality for a class of nonlinear transformations”. In: *Journal of Statistical Physics* 19.1 (1978). The universal constants delta and alpha; sequel in J. Stat. Phys. 21 (1979) 669-706., pp. 25–52. DOI: 10.1007/BF01020332.

[1622] C. Fronsdal. “Massless fields with integer spin”. In: *Physical Review D* 18.10 (1978), pp. 3624–3629. DOI: 10.1103/PhysRevD.18.3624.

[1623] V. Gribov. “Quantization of non-Abelian gauge theories”. In: *Nuclear Physics B* 139.1-2 (1978), pp. 1–19. DOI: 10.1016/0550-3213(78)90175-X.

[1624] N. K. Nielsen and P. Olesen. “An unstable Yang-Mills field mode”. In: *Nuclear Physics B* 144.2–3 (1978). The unstable $n = 0$ mode of a constant chromomagnetic field, the caveat to magnetic-vacuum readings of asymptotic freedom. Not the Nielsen–Ninomiya theorem, which is Nielsen:1981. Metadata confirmed against Crossref., pp. 376–396. DOI: 10.1016/0550-3213(78)90377-2.

[1625] C. Prescott et al. “Parity non-conservation in inelastic electron scattering”. In: *Physics Letters B* 77.3 (1978), pp. 347–352. DOI: 10.1016/0370-2693(78)90722-0.

[1626] I. M. Singer. “Some remarks on the Gribov ambiguity”. In: *Communications in Mathematical Physics* 60.1 (1978), pp. 7–12. DOI: 10.1007/BF01609471.

[1627] G. L. Squires. *Introduction to the Theory of Thermal Neutron Scattering*. First edition, Cambridge University Press, Cambridge, 1978; vii+260 pp.; ISBN 0-521-21884-5. NO DOI GIVEN DELIBERATELY: the 1978 first edition predates DOI registration. The CUP third edition (2012) carries 10.1017/CBO9781139107808 and a Dover reprint appeared in 1996, but attaching either DOI to a key dated 1978 would misreport the issue year, so the entry should grade AMBER on publisher/edition as an @book, which classify_access() auto-excuses. Placed at the textbook claim in sec:nscat-why that the nuclear scattering length is angle-independent and varies erratically with element and isotope (line 126) - the one paragraph of the chapter that was pure textbook material with no source at all. Cambridge University Press, 1978.

[1628] S. Weinberg. “A New Light Boson?” In: *Physical Review Letters* 40.4 (1978), pp. 223–226. DOI: 10.1103/PhysRevLett.40.223.

[1629] J. A. Wheeler. “The “Past” and the “Delayed-Choice” Double-Slit Experiment”. In: *Mathematical Foundations of Quantum Theory*. Contribution to the volume edited by A. R. Marlow; no journal volume exists. 1978, pp. 9–48. DOI: 10.1016/b978-0-12-473250-6.50006-6.

[1630] F. Wilczek. “Problem of StrongPandTInvariance in the Presence of Instantons”. In: *Physical Review Letters* 40.5 (1978), pp. 279–282. DOI: 10.1103/PhysRevLett.40.279.

[1631] L. Wolfenstein. “Neutrino oscillations in matter”. In: *Physical Review D* 17.9 (1978), pp. 2369–2374. DOI: 10.1103/PhysRevD.17.2369.

[1632] J. Bailey et al. “Final report on the CERN muon storage ring including the anomalous magnetic moment and the electric dipole moment of the muon, and a direct test of relativistic time dilation”. In: *Nuclear Physics B* 150 (1979). Distinct from the already-present Bailey:1977 (Nature 268, 301), which reports the time-dilation result only; the anomalous moment value is in this final report., pp. 1–75. DOI: 10.1016/0550-3213(79)90292-X.

[1633] V. Baluni. “CP-nonconserving effects in quantum chromodynamics”. In: *Physical Review D* 19.7 (1979), pp. 2227–2230. DOI: 10.1103/PhysRevD.19.2227.

[1634] R. Brandelik et al. “Evidence for planar events in e+e− annihilation at high energies”. In: *Physics Letters B* 86.2 (1979), pp. 243–249. DOI: 10.1016/0370-2693(79)90830-X.

[1635] A. Brillet and J. L. Hall. “Improved Laser Test of the Isotropy of Space”. In: *Physical Review Letters* 42.9 (1979). A laser-cavity descendant of Michelson–Morley, pp. 549–552. DOI: 10.1103/PhysRevLett.42.549.

[1636] B. V. Chirikov. “A universal instability of many-dimensional oscillator systems”. In: *Physics Reports* 52.5 (1979). The standard map and the resonance-overlap criterion., pp. 263–379. DOI: 10.1016/0370-1573(79)90023-1.

[1637] R. Crewther, P. Di Vecchia, G. Veneziano, and E. Witten. “Chiral estimate of the electric dipole moment of the neutron in quantum chromodynamics”. In: *Physics Letters B* 88.1-2 (1979), pp. 123–127. DOI: 10.1016/0370-2693(79)90128-X.

[1638] K. Fujikawa. “Path-Integral Measure for Gauge-Invariant Fermion Theories”. In: *Physical Review Letters* 42.18 (1979), pp. 1195–1198. DOI: 10.1103/PhysRevLett.42.1195.

[1639] J. G. H. Groot et al. “Inclusive interactions of high-energy neutrinos and antineutrinos in iron”. In: *Zeitschrift für Physik C* 1.2 (1979), pp. 143–162. DOI: 10.1007/BF01445406.

[1640] G. M. Harvey. “Gravitational deflection of light: a re-examination of the observations of the solar eclipse of 1919”. In: *The Observatory* 99 (1979). The 1978 Royal Greenwich Observatory re-measurement (set in motion by Andrew Murray) of the original 1919 plates on modern measuring machines: five of the seven Sobral 4-inch plates and all 16 Sobral astrographic plates, the latter with much improved accuracy and consistent with the full deflection. This is exactly the claim in the eclipse Interpretation subsection. ADS bibcode 1979Obs....99..195H. The Observatory of that era has no DOIs – the ADS scanned-article link is the access route, so this needs an "access" manifest entry or it grades RED as a DOI-less @article. Note the ADS title form uses a dash rather than a colon before the subtitle; the sense is identical., pp. 195–198.

[1641] T. Kugo and I. Ojima. “Local Covariant Operator Formalism of Non-Abelian Gauge Theories and Quark Confinement Problem”. In: *Progress of Theoretical Physics Supplement* 66 (1979), pp. 1–130. DOI: 10.1143/PTPS.66.1.

[1642] J.-P. Luminet. “Image of a spherical black hole with thin accretion disk”. In: *Astronomy and Astrophysics* 75 (1979). Pre-2001 A&A article; no DOI exists., pp. 228–235.

[1643] C. Y. Prescott et al. “Further measurements of parity non-conservation in inelastic electron scattering”. In: *Physics Letters B* 84.4 (1979). The follow-up to Prescott:1978 with the full kinematic scan; source of $\sin^2\theta_W = 0.224(20)$, where the 1978 letter gave 0.20(3). Metadata confirmed against Crossref., pp. 524–528. DOI: 10.1016/0370-2693(79)91253-X.

[1644] R. D. Reasenberg et al. “Viking relativity experiment - Verification of signal retardation by solar gravity”. In: *The Astrophysical Journal* 234 (1979). Astrophysical Journal Letters pages., p. L219. DOI: 10.1086/183144.

[1645] D. Walsh, R. F. Carswell, and R. J. Weymann. “0957 + 561 A, B: twin quasistellar objects or gravitational lens?” In: *Nature* 279 (1979), pp. 381–384.

[1646] S. Weinberg. “Baryon- and Lepton-Nonconserving Processes”. In: *Physical Review Letters* 43.21 (1979), pp. 1566–1570. DOI: 10.1103/PhysRevLett.43.1566.

[1647] S. Weinberg. “Phenomenological Lagrangians”. In: *Physica A: Statistical Mechanics and its Applications* 96.1-2 (1979). Crossref container title is the modern ’Physica A: Statistical Mechanics and its Applications’; issue 1-2, April 1979. Founding paper of effective field theory and of chiral perturbation theory. This key is already in use in the tree (10-renormalization-group.tex) for this same paper., pp. 327–340. DOI: 10.1016/0378-4371(79)90223-1.

[1648] S. Weinberg. “Ultraviolet divergences in quantum theories of gravitation”. In: *General Relativity: An Einstein Centenary Survey*. Chapter 16 of the volume edited by S. W. Hawking and W. Israel (Cambridge, 1979); origin of the asymptotic-safety conjecture. No volume number and no DOI for the 1979 chapter. Suffixed ’c’ rather than ’b’ because Weinberg:1979b is already claimed in 11-qft-standard-model/10-renormalization-group.tex for a third 1979 paper, ’Baryon- and Lepton-Nonconserving Processes’ (Phys. Rev. Lett. 43, 1566), which supplies the dimension-five neutrino-mass operator. 1979, pp. 790–831.

[1649] D. P. Wong. “Theoretical justification of Madelung’s rule”. In: *Journal of Chemical Education* 56.11 (1979), p. 714. DOI: 10.1021/ed056p714.

[1650] B. S. Cirel’son. “Quantum generalizations of Bell’s inequality”. In: *Letters in Mathematical Physics* 4.2 (1980). Published under the transliteration ’Cirel’son’; the author is now usually written Tsirelson, and the prose of ch. IX.8 uses that spelling while the key follows the published byline., pp. 93–100. DOI: 10.1007/BF00417500.

[1651] M. Creutz. “Asymptotic-Freedom Scales”. In: *Physical Review Letters* 45.5 (1980). Issue of 4 August 1980. The string tension of SU(2) lattice gauge theory shown to follow asymptotic-freedom scaling, i.e. the lattice theory joining onto the continuum., pp. 313–316. DOI: 10.1103/PhysRevLett.45.313.

[1652] B. C. van Fraassen. *The Scientific Image*. Constructive empiricism: accepting a theory involves believing only that it is empirically adequate. Oxford: Clarendon Press, 1980. DOI: 10.1093/0198244274.001.0001.

[1653] G. C. Ghirardi, A. Rimini, and T. Weber. “A general argument against superluminal transmission through the quantum mechanical measurement process”. In: *Lettere al Nuovo Cimento* 27.10 (1980), pp. 293–298. DOI: 10.1007/BF02817189.

[1654] C. Itzykson and J.-B. Zuber. *Quantum Field Theory*. McGraw-Hill, 1980.

[1655] K. v. Klitzing, G. Dorda, and M. Pepper. “New Method for High-Accuracy Determination of the Fine-Structure Constant Based on Quantized Hall Resistance”. In: *Physical Review Letters* 45.6 (1980), pp. 494–497. DOI: 10.1103/PhysRevLett.45.494.

[1656] L. D. Landau and E. M. Lifshitz. *Statistical Physics, Part 1*. Course of Theoretical Physics, Volume 5; 3rd edition, revised by E. M. Lifshitz and L. P. Pitaevskii. Oxford: Pergamon Press, 1980.

[1657] Y. Pomeau and P. Manneville. "Intermittent transition to turbulence in dissipative dynamical systems". In: *Communications in Mathematical Physics* 74.2 (1980). The intermittency route to chaos., pp. 189–197. DOI: 10.1007/BF01197757.

[1658] V. C. Rubin, N. Thonnard, and J. Ford W. K. "Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 /R = 4kpc/ to UGC 2885 /R = 122 kpc/". In: *The Astrophysical Journal* 238 (1980), p. 471. DOI: 10.1086/158003.

[1659] P. Sikivie, L. Susskind, M. Voloshin, and V. Zakharov. "Isospin breaking in technicolor models". In: *Nuclear Physics B* 173.2 (1980), pp. 189–207. DOI: 10.1016/0550-3213(80)90214-X.

[1660] R. F. C. Vessot et al. "Test of Relativistic Gravitation with a Space-Borne Hydrogen Maser". In: *Physical Review Letters* 45.26 (1980). Gravity Probe A., pp. 2081–2084. DOI: 10.1103/PhysRevLett.45.2081.

[1661] S. Weinberg and E. Witten. "Limits on massless particles". In: *Physics Letters B* 96.1-2 (1980), pp. 59–62. DOI: 10.1016/0370-2693(80)90212-9.

[1662] M. Aizenman. "Proof of the Triviality ofϕd4Field Theory and Some Mean-Field Features of Ising Models ford>4". In: *Physical Review Letters* 47.1 (1981), pp. 1–4. DOI: 10.1103/PhysRevLett.47.1.

[1663] A. Aspect, P. Grangier, and G. Roger. "Experimental Tests of Realistic Local Theories via Bell's Theorem". In: *Physical Review Letters* 47.7 (1981), pp. 460–463. DOI: 10.1103/PhysRevLett.47.460.

[1664] I. Batalin and G. Vilkovisky. "Gauge algebra and quantization". In: *Physics Letters B* 102.1 (1981), pp. 27–31. DOI: 10.1016/0370-2693(81)90205-7.

[1665] A. Bosma. "21-cm line studies of spiral galaxies. II. The distribution and kinematics of neutral hydrogen in spiral galaxies of various morphological types". In: *The Astronomical Journal* 86 (1981), p. 1825. DOI: 10.1086/113063.

[1666] A. O. Caldeira and A. J. Leggett. "Influence of Dissipation on Quantum Tunneling in Macroscopic Systems". In: *Physical Review Letters* 46.4 (1981), pp. 211–214. DOI: 10.1103/PhysRevLett.46.211.

[1667] C. M. Caves. "Quantum-mechanical noise in an interferometer". In: *Physical Review D* 23.8 (1981), pp. 1693–1708. DOI: 10.1103/PhysRevD.23.1693.

[1668] R. K. Ellis, D. A. Ross, and A. E. Terrano. "The perturbative calculation of jet structure in e^+e^- annihilation". In: *Nuclear Physics B* 178.3 (1981). The exact $O(\alpha_s)$ three-jet matrix element and distributions. Metadata confirmed against Crossref., pp. 421–456. DOI: 10.1016/0550-3213(81)90165-6.

[1669] R. L. Fork, B. I. Greene, and C. V. Shank. "Generation of optical pulses shorter than 0.1 psec by colliding pulse mode locking". In: *Applied Physics Letters* 38.9 (1981). Issue 9. The colliding-pulse mode-locked ring dye laser, the source that made the femtosecond timescale accessible. Confirmed against the Crossref record for this DOI., pp. 671–672. DOI: 10.1063/1.92500.

[1670] A. H. Guth. "Inflationary universe: A possible solution to the horizon and flatness problems". In: *Physical Review D* 23.2 (1981), pp. 347–356. DOI: 10.1103/PhysRevD.23.347.

[1671] R. B. Laughlin. “Quantized Hall conductivity in two dimensions”. In: *Physical Review B* 23.10 (1981), pp. 5632–5633. DOI: 10.1103/PhysRevB.23.5632.

[1672] V. F. Mukhanov and G. V. Chibisov. “Quantum fluctuations and a nonsingular universe”. In: *JETP Letters* 33 (1981). English translation of Pis’ma Zh. Eksp. Teor. Fiz. 33, 549 (1981). No DOI for the translation., pp. 532–535.

[1673] H. Nielsen and M. Ninomiya. “Absence of neutrinos on a lattice”. In: *Nuclear Physics B* 185.1 (1981), pp. 20–40. DOI: 10.1016/0550-3213(81)90361-8.

[1674] D. N. Page and C. D. Geilker. “Indirect Evidence for Quantum Gravity”. In: *Physical Review Letters* 47.14 (1981), pp. 979–982. DOI: 10.1103/PhysRevLett.47.979.

[1675] W. G. Unruh. “Experimental Black-Hole Evaporation?” In: *Physical Review Letters* 46.21 (1981), pp. 1351–1353. DOI: 10.1103/PhysRevLett.46.1351.

[1676] W. H. Zurek. “Pointer basis of quantum apparatus: Into what mixture does the wave packet collapse?” In: *Physical Review D* 24.6 (1981), pp. 1516–1525. DOI: 10.1103/PhysRevD.24.1516.

[1677] A. Albrecht and P. J. Steinhardt. “Cosmology for Grand Unified Theories with Radiatively Induced Symmetry Breaking”. In: *Physical Review Letters* 48 (1982). Metadata confirmed against Crossref. A. Albrecht (1982, inflation) and H. Albrecht (1987, ARGUS) are different people; do not fold the keys., pp. 1220–1223. DOI: 10.1103/PhysRevLett.48.1220.

[1678] T. Ando, A. B. Fowler, and F. Stern. “Electronic properties of two-dimensional systems”. In: *Reviews of Modern Physics* 54.2 (1982), pp. 437–672. DOI: 10.1103/RevModPhys.54.437.

[1679] A. Aspect, J. Dalibard, and G. Roger. “Experimental Test of Bell’s Inequalities Using Time- Varying Analyzers”. In: *Physical Review Letters* 49.25 (1982). Suffix a/b needed: two Aspect papers in 1982. ’b’ is the switched-analyzer measurement (S = 2.404 +- 0.080)., pp. 1804–1807. DOI: 10.1103/PhysRevLett.49.1804.

[1680] A. Aspect, P. Grangier, and G. Roger. “Experimental Realization of Einstein-Podolsky-Rosen-BohmGedankenexperiment: A New Violation of Bell’s Inequalities”. In: *Physical Review Letters* 49.2 (1982). Suffix a/b needed: two Aspect papers in 1982. ’a’ is the two-channel-polarizer measurement (S = 2.697 +- 0.015)., pp. 91–94. DOI: 10.1103/PhysRevLett.49.91.

[1681] D. C. Backer, S. R. Kulkarni, C. Heiles, M. M. Davis, and W. M. Goss. “A millisecond pulsar”. In: *Nature* 300.5893 (1982), pp. 615–618. DOI: 10.1038/300615a0.

[1682] G. Binnig, H. Rohrer, C. Gerber, and E. Weibel. “Surface Studies by Scanning Tunneling Microscopy”. In: *Physical Review Letters* 49.1 (1982). Two Binnig 1982 papers compete for this key: Appl. Phys. Lett. 40, 178 (tunnelling through a controllable vacuum gap) and this PRL. The citing prose says the instrument ’resolves single atoms’, which is the PRL (Si(111) 7x7), so the PRL is the right work. Crossref-verified; published 5 Jul 1982. Note the American ’Tunneling’ in the title against the British ’tunnelling’ in the chapter prose., pp. 57–61. DOI: 10.1103/PhysRevLett.49.57.

[1683] N. D. Birrell and P. C. W. Davies. *Quantum Fields in Curved Space*. Cambridge University Press, 1982.

[1684] D. Buchholz and K. Fredenhagen. “Locality and the structure of particle states”. In: *Communications in Mathematical Physics* 84.1 (1982), pp. 1–54. DOI: 10.1007/BF01208370.

[1685] B. Cabrera. “First Results from a Superconductive Detector for Moving Magnetic Monopoles”. In: *Physical Review Letters* 48.20 (1982). Crossref confirmed: PRL vol. 48, pp. 1378-1381, 17 May 1982. Single author. This is the paper reporting the single 14 February 1982 candidate event of one Dirac charge that the chapter describes as never repeated., pp. 1378–1381. DOI: 10.1103/PhysRevLett.48.1378.

[1686] D. Dieks. “Communication by EPR devices”. In: *Physics Letters A* 92.6 (1982), pp. 271–272. DOI: 10.1016/0375-9601(82)90084-6.

[1687] A. Fine. “Hidden Variables, Joint Probability, and the Bell Inequalities”. In: *Physical Review Letters* 48.5 (1982), pp. 291–295. DOI: 10.1103/PhysRevLett.48.291.

[1688] J. Fröhlich. “On the triviality of $\lambda\phi$d4 theories and the approach to the critical point in d$(-) > 4$ dimensions”. In: *Nuclear Physics B* 200.2 (1982), pp. 281–296. DOI: 10.1016/0550-3213(82)90088-8.

[1689] B. I. Halperin. “Quantized Hall conductance, current-carrying edge states, and the existence of extended states in a two-dimensional disordered potential”. In: *Physical Review B* 25.4 (1982), pp. 2185–2190. DOI: 10.1103/PhysRevB.25.2185.

[1690] L. Kirby and J. Paris. “Accessible Independence Results for Peano Arithmetic”. In: *Bulletin of the London Mathematical Society* 14.4 (1982). Proves Goodstein’s theorem and the hydra theorem unprovable in PA. Crossref still records Wiley as publisher; the DOI resolves to Oxford Academic on behalf of the London Mathematical Society, pp. 285–293. DOI: 10.1112/blms/14.4.285.

[1691] Y. Koide. “Fermion-boson two-body model of quarks and leptons and Cabibbo mixing”. In: *Lettere al Nuovo Cimento* 34.8 (1982). With Koide:1983a the source of the mass relation $Q = 2/3$ (the “model I” relation). NOT Koide:1983b, which proposes a modified relation now excluded, pp. 201–205. DOI: 10.1007/BF02817096.

[1692] H. Kragh. “Erwin Schrödinger and the Wave Equation: The Crucial Phase”. In: *Centaurus* 26.2 (1982). Historical source for the claim that Schrödinger derived and discarded the relativistic equation first., pp. 154–197. DOI: 10.1111/j.1600-0498.1982.tb00659.x.

[1693] A. Libchaber, C. Laroche, and S. Fauve. “Period doubling cascade in mercury, a quantitative measurement”. In: *Journal de Physique Lettres* 43.7 (1982). Quantitative recovery of the Feigenbaum constant in Rayleigh-Bénard convection of mercury; earlier helium experiments Maurer and Libchaber, J. Physique Lettres 40 (1979) L419., pp. 211–216. DOI: 10.1051/jphyslet:01982004307021100.

[1694] A. D. Linde. “A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems”. In: *Physics Letters B* 108 (1982). Metadata confirmed against Crossref., pp. 389–393. DOI: 10.1016/0370-2693(82)91219-9.

[1695] G. G. Luther and W. R. Towler. “Redetermination of the Newtonian Gravitational ConstantG”. In: *Physical Review Letters* 48.3 (1982). The reference torsion value underlying CODATA for a generation., pp. 121–123. DOI: 10.1103/PhysRevLett.48.121.

[1696] B. B. Mandelbrot. *The Fractal Geometry of Nature*. Updated and augmented edition of Fractals (1977). San Francisco: W. H. Freeman, 1982.

[1697] P. O. Mazur. “Proof of uniqueness of the Kerr–Newman black hole solution”. In: *Journal of Physics A: Mathematical and General* 15 (1982). Metadata confirmed against Crossref., pp. 3173–3180. DOI: 10.1088/0305-4470/15/10/021.

[1698] R. G. Newton. *Scattering Theory of Waves and Particles*. 2nd edition. Different author from the existing Newton:1687 (Isaac Newton); no key collision. New York: Springer-Verlag, 1982.

[1699] P. J. E. Peebles. "Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations". In: *The Astrophysical Journal* 263 (1982), p. L1. DOI: 10.1086/183911.

[1700] W. D. Phillips and H. Metcalf. "Laser Deceleration of an Atomic Beam". In: *Physical Review Letters* 48.9 (1982), pp. 596–599. DOI: 10.1103/PhysRevLett.48.596.

[1701] J. Schechter and J. W. F. Valle. "Neutrinoless double-βdecay in SU(2)×U(1) theories". In: *Physical Review D* 25.11 (1982), pp. 2951–2954. DOI: 10.1103/PhysRevD.25.2951.

[1702] M. O. Scully and K. Drühl. "Quantum eraser: A proposed photon correlation experiment concerning observation and "delayed choice" in quantum mechanics". In: *Physical Review A* 25.4 (1982), pp. 2208–2213. DOI: 10.1103/PhysRevA.25.2208.

[1703] J. H. Taylor and J. M. Weisberg. "A new test of general relativity: Gravitational radiation and the binary pulsar PSR 1913+16". In: *The Astrophysical Journal* 253 (1982), pp. 908–920. DOI: 10.1086/159690.

[1704] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs. "Quantized Hall Conductance in a Two-Dimensional Periodic Potential". In: *Physical Review Letters* 49.6 (1982), pp. 405–408. DOI: 10.1103/PhysRevLett.49.405.

[1705] D. C. Tsui, H. L. Stormer, and A. C. Gossard. "Two-Dimensional Magnetotransport in the Extreme Quantum Limit". In: *Physical Review Letters* 48.22 (1982), pp. 1559–1562. DOI: 10.1103/PhysRevLett.48.1559.

[1706] F. Wilczek. "Quantum Mechanics of Fractional-Spin Particles". In: *Physical Review Letters* 49.14 (1982), pp. 957–959. DOI: 10.1103/PhysRevLett.49.957.

[1707] E. Witten. "An SU(2) anomaly". In: *Physics Letters B* 117.5 (1982), pp. 324–328. DOI: 10.1016/0370-2693(82)90728-6.

[1708] W. K. Wootters and W. H. Zurek. "A single quantum cannot be cloned". In: *Nature* 299.5886 (1982), pp. 802–803. DOI: 10.1038/299802a0.

[1709] W. H. Zurek. "Environment-induced superselection rules". In: *Physical Review D* 26.8 (1982), pp. 1862–1880. DOI: 10.1103/PhysRevD.26.1862.

[1710] G. Arnison et al. "Experimental observation of isolated large transverse energy electrons with associated missing energy at $\sqrt{s}$ = 540 GeV". In: *Physics Letters B* 122.1 (1983). UA1 discovery of the W boson, pp. 103–116. DOI: 10.1016/0370-2693(83)91177-2.

[1711] G. Arnison et al. "Experimental observation of lepton pairs of invariant mass around 95 GeV/c2 at the CERN SPS collider". In: *Physics Letters B* 126.5 (1983). UA1 discovery of the Z boson, pp. 398–410. DOI: 10.1016/0370-2693(83)90188-0.

[1712] J. Aubert et al. "The ratio of the nucleon structure functions F2N for iron and deuterium". In: *Physics Letters B* 123.3-4 (1983), pp. 275–278. DOI: 10.1016/0370-2693(83)90437-9.

[1713] P. Bagnaia et al. "Evidence for Z0→e+e− at the CERN p collider". In: *Physics Letters B* 129.1-2 (1983), pp. 130–140. DOI: 10.1016/0370-2693(83)90744-X.

[1714] M. Banner et al. "Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN p collider". In: *Physics Letters B* 122.5-6 (1983), pp. 476–485. DOI: 10.1016/0370-2693(83)91605-2.

[1715] G. Binnig, H. Rohrer, C. Gerber, and E. Weibel. “7 × 7 Reconstruction on Si(111) Resolved in Real Space”. In: *Physical Review Letters* 50.2 (1983). Issue 2, 10 January 1983. Same four authors as the existing Binnig:1982. The title carries U+00D7 MULTIPLICATION SIGN, which is fatal under pdflatex+T1 – write the .bib title as 7 × 7 Reconstruction on Si(111) Resolved in Real Space. Caution: the note on the existing Binnig:1982 entry claims that PRL (49, 57) is the Si(111) 7x7 paper; it is not – the 7x7 in real space is this 1983 PRL, and the chapter prose (’within a year the same instrument’) is correct. That stale note on Binnig:1982 should be fixed separately., pp. 120–123. DOI: 10.1103/PhysRevLett.50.120.

[1716] A. Caldeira and A. Leggett. “Path integral approach to quantum Brownian motion”. In: *Physica A: Statistical Mechanics and its Applications* 121.3 (1983), pp. 587–616. DOI: 10.1016/0378-4371(83)90013-4.

[1717] S. Chandrasekhar. *The Mathematical Theory of Black Holes*. International Series of Monographs on Physics 69. Oxford: Clarendon Press, 1983.

[1718] R. W. P. Drever et al. “Laser phase and frequency stabilization using an optical resonator”. In: *Applied Physics B Photophysics and Laser Chemistry* 31.2 (1983), pp. 97–105. DOI: 10.1007/BF00702605.

[1719] P. Goy, J. M. Raimond, M. Gross, and S. Haroche. “Observation of Cavity-Enhanced Single-Atom Spontaneous Emission”. In: *Physical Review Letters* 50.24 (1983), pp. 1903–1906. DOI: 10.1103/PhysRevLett.50.1903.

[1720] P. Grassberger and I. Procaccia. “Characterization of Strange Attractors”. In: *Physical Review Letters* 50.5 (1983). The correlation-dimension algorithm; companion long paper Physica D 9 (1983) 189-208., pp. 346–349. DOI: 10.1103/PhysRevLett.50.346.

[1721] R. A. de Groot, F. M. Mueller, P. G. v. Engen, and K. H. J. Buschow. “New Class of Materials: Half-Metallic Ferromagnets”. In: *Physical Review Letters* 50.25 (1983), pp. 2024–2027. DOI: 10.1103/PhysRevLett.50.2024.

[1722] A. Klug. “From Macromolecules to Biological Assemblies (Nobel Lecture)”. In: *Angewandte Chemie International Edition in English* 22.8 (1983). Crossref-verified. Nobel Lecture in Chemistry 1982. Note the exact title is ’From Macromolecules to Biological Assemblies’, not ’...From Macromolecular Structures...’; a Bioscience Reports version (3, 395-430) exists as a separate record., pp. 565–582. DOI: 10.1002/anie.198305653.

[1723] Y. Koide. “A fermion-boson composite model of quarks and leptons”. In: *Physics Letters B* 120.1-3 (1983), pp. 161–165. DOI: 10.1016/0370-2693(83)90644-5.

[1724] Y. Koide. “New view of quark and lepton mass hierarchy”. In: *Physical Review D* 28.1 (1983). A Brief Report proposing a MODIFIED relation with $\delta = 0.040026$, predicting $m_\tau = 1786.45\,\mathrm{MeV}$; that prediction is excluded by the measured $(1776.93 \pm 0.09)\,\mathrm{MeV}$. It is not the source of $Q = 2/3$, which it calls “the old prediction” of model I, pp. 252–254. DOI: 10.1103/PhysRevD.28.252.

[1725] R. B. Laughlin. “Anomalous Quantum Hall Effect: An Incompressible Quantum Fluid with Fractionally Charged Excitations”. In: *Physical Review Letters* 50.18 (1983), pp. 1395–1398. DOI: 10.1103/PhysRevLett.50.1395.

[1726] M. Milgrom. “A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis”. In: *The Astrophysical Journal* 270 (1983), p. 365. DOI: 10.1086/161130.

[1727] S. L. Shapiro and S. A. Teukolsky. *Black Holes, White Dwarfs, and Neutron Stars: The Physics of Compact Objects*. DOI is the Wiley online edition of the 1983 book. Wiley, 1983. DOI: 10.1002/9783527617661.

[1728] R. Short and L. Mandel. "Observation of Sub-Poissonian Photon Statistics". In: *Physical Review Letters* 51.5 (1983), pp. 384–387. DOI: 10.1103/PhysRevLett.51.384.

[1729] P. Sikivie. "Experimental Tests of the "Invisible" Axion". In: *Physical Review Letters* 51.16 (1983). Issue 16. The resonant-cavity haloscope proposal: Primakoff conversion of a halo axion into a single photon in a static magnetic field. Cited in the ADMX derivation at the conversion sentence. Two bib-authoring points: (a) the published title contains typographic double quotes around Invisible — write them as "Invisible" in the .bib, never as a raw " character; (b) there is an erratum, Phys. Rev. Lett. 52, 695 (1984), doi 10.1103/PhysRevLett.52.695.2, worth recording in a note field. Distinct from the existing key Sikivie:1980 (technicolor isospin breaking)., pp. 1415–1417. DOI: 10.1103/PhysRevLett.51.1415.

[1730] D. F. Walls. "Squeezed states of light". In: *Nature* 306.5939 (1983), pp. 141–146. DOI: 10.1038/306141a0.

[1731] J. A. Wheeler and W. H. Zurek. *Quantum Theory and Measurement*. Edited source book; per the validator's rule an editor satisfies the author requirement for @book/@collection. Princeton: Princeton University Press, 1983.

[1732] J. Wisdom. "Chaotic behavior and the origin of the 31 Kirkwood gap". In: *Icarus* 56.1 (1983). Chaotic origin of the 3/1 Kirkwood gap., pp. 51–74. DOI: 10.1016/0019-1035(83)90127-6.

[1733] L. Wolfenstein. "Parametrization of the Kobayashi-Maskawa Matrix". In: *Physical Review Letters* 51.21 (1983), pp. 1945–1947. DOI: 10.1103/PhysRevLett.51.1945.

[1734] M. V. Berry. "Quantal phase factors accompanying adiabatic changes". In: *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 392.1802 (1984). Berry's phase. Crossref records the author as 'Michael Victor Berry'; the paper itself is signed M. V. Berry. Crossref-verified; issued 8 Mar 1984., pp. 45–57. DOI: 10.1098/rspa.1984.0023.

[1735] G. R. Blumenthal, S. M. Faber, J. R. Primack, and M. J. Rees. "Formation of galaxies and large-scale structure with cold dark matter". In: *Nature* 311.5986 (1984), pp. 517–525. DOI: 10.1038/311517a0.

[1736] R. J. Boyd. "A quantum mechanical explanation for Hund's multiplicity rule". In: *Nature* 310.5977 (1984), pp. 480–481. DOI: 10.1038/310480a0.

[1737] J. Gasser and H. Leutwyler. "Chiral perturbation theory to one loop". In: *Annals of Physics* 158.1 (1984), pp. 142–210. DOI: 10.1016/0003-4916(84)90242-2.

[1738] R. B. Griffiths. "Consistent histories and the interpretation of quantum mechanics". In: *Journal of Statistical Physics* 36.1-2 (1984). Different author from the existing Griffiths:2018 (David J. Griffiths, textbook); no key collision., pp. 219–272. DOI: 10.1007/BF01015734.

[1739] F. Halzen and A. D. Martin. *Quarks and Leptons: An Introductory Course in Modern Particle Physics*. New York: John Wiley and Sons, 1984.

[1740] M. Heideman, D. Johnson, and C. Burrus. "Gauss and the history of the fast fourier transform". In: *IEEE ASSP Magazine* 1.4 (1984). Issue 4. Documents Gauss's 1805 anticipation of the FFT., pp. 14–21. DOI: 10.1109/massp.1984.1162257.

[1741] A. M. Hillas. “The Origin of Ultra-High-Energy Cosmic Rays”. In: *Annual Review of Astronomy and Astrophysics* 22.1 (1984), pp. 425–444. DOI: 10.1146/annurev.aa.22.090184.002233.

[1742] T. Matsumoto. “A chaotic attractor from Chua’s circuit”. In: *IEEE Transactions on Circuits and Systems* 31.12 (1984). The double-scroll attractor of Chua’s circuit., pp. 1055–1058. DOI: 10.1109/TCS.1984.1085459.

[1743] J. Polchinski. “Renormalization and effective lagrangians”. In: *Nuclear Physics B* 231.2 (1984), pp. 269–295. DOI: 10.1016/0550-3213(84)90287-6.

[1744] D. Shechtman, I. Blech, D. Gratias, and J. W. Cahn. “Metallic Phase with Long-Range Orientational Order and No Translational Symmetry”. In: *Physical Review Letters* 53.20 (1984), pp. 1951–1953. DOI: 10.1103/PhysRevLett.53.1951.

[1745] D. Shoenberg. *Magnetic Oscillations in Metals*. Cambridge University Press, 1984.

[1746] R. M. Wald. *General Relativity*. Chicago: University of Chicago Press, 1984.

[1747] T. S. van Albada, J. N. Bahcall, K. Begeman, and R. Sancisi. “Distribution of dark matter in the spiral galaxy NGC 3198”. In: *The Astrophysical Journal* 295 (1985), p. 305. DOI: 10.1086/163375.

[1748] R. Baltrusaitis et al. “The Utah Fly’s Eye detector”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 240.2 (1985), pp. 410–428. DOI: 10.1016/0168-9002(85)90658-8.

[1749] H. B. Callen. *Thermodynamics and an Introduction to Thermostatistics*. 2nd ed. New York: Wiley, 1985.

[1750] S. Chu, L. Hollberg, J. E. Bjorkholm, A. Cable, and A. Ashkin. “Three-dimensional viscous confinement and cooling of atoms by resonance radiation pressure”. In: *Physical Review Letters* 55.1 (1985), pp. 48–51. DOI: 10.1103/PhysRevLett.55.48.

[1751] S. Coleman. *Aspects of Symmetry: Selected Erice Lectures*. Cambridge University Press, 1985.

[1752] R. P. Feynman. *QED: The Strange Theory of Light and Matter*. Source of the remark that α is “one of the greatest damn mysteries of physics: a magic number that comes to us with no understanding by man”. Princeton, NJ: Princeton University Press, 1985.

[1753] M. W. Goodman and E. Witten. “Detectability of certain dark-matter candidates”. In: *Physical Review D* 31.12 (1985). Issue 12. The founding proposal that halo particles be sought via coherent elastic nuclear recoil. Cited at sec:expdm-kinematics, first sentence ("The signature sought underground is a single nuclear recoil")., pp. 3059–3063. DOI: 10.1103/PhysRevD.31.3059.

[1754] A. M. Hillas. “Cerenkov Light Images of EAS Produced by Primary Gamma Rays and by Nuclei”. In: *Proceedings of the 19th International Cosmic Ray Conference, La Jolla*. Vol. 3. 1985, pp. 445–448.

[1755] R. G. Hulet, E. S. Hilfer, and D. Kleppner. “Inhibited Spontaneous Emission by a Rydberg Atom”. In: *Physical Review Letters* 55.20 (1985), pp. 2137–2140. DOI: 10.1103/PhysRevLett.55.2137.

[1756] C. Jarlskog. “Commutator of the Quark Mass Matrices in the Standard Electroweak Model and a Measure of MaximalCPNonconservation”. In: *Physical Review Letters* 55.10 (1985), pp. 1039–1042. DOI: 10.1103/PhysRevLett.55.1039.

[1757] E. Joos and H. D. Zeh. “The emergence of classical properties through interaction with the environment”. In: *Zeitschrift für Physik B Condensed Matter* 59.2 (1985), pp. 223–243. DOI: 10.1007/BF01725541.

[1758] V. Kuzmin, V. Rubakov, and M. Shaposhnikov. “On anomalous electroweak baryon-number non-conservation in the early universe”. In: *Physics Letters B* 155.1-2 (1985), pp. 36–42. DOI: 10.1016/0370-2693(85)91028-7.

[1759] A. J. Leggett and A. Garg. “Quantum mechanics versus macroscopic realism: Is the flux there when nobody looks?” In: *Physical Review Letters* 54.9 (1985), pp. 857–860. DOI: 10.1103/PhysRevLett.54.857.

[1760] S. van der Meer. “Stochastic cooling and the accumulation of antiprotons”. In: *Reviews of Modern Physics* 57.3 (1985), pp. 689–697. DOI: 10.1103/RevModPhys.57.689.

[1761] A. L. Migdall, J. V. Prodan, W. D. Phillips, T. H. Bergeman, and H. J. Metcalf. “First Observation of Magnetically Trapped Neutral Atoms”. In: *Physical Review Letters* 54.24 (1985), pp. 2596–2599. DOI: 10.1103/PhysRevLett.54.2596.

[1762] T. J. Quinn and J. E. Martin. “A radiometric determination of the Stefan-Boltzmann constant and thermodynamic temperatures between -40 °C and +100 °C”. In: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 316.1536 (1985). Issue 1536, 26 November 1985. The founding absolute-cryogenic-radiometry paper: total radiant exitance of a blackbody measured with a cryogenic radiometer, giving sigma = 5.66967(76)e-8 W mˆ-2 Kˆ-4 and thermodynamic temperatures from 233 K to 373 K. Crossref types it journal-article with a real volume, issue, pages and DOI, so it is @article, not @inproceedings – the academy-proceedings rule is aimed at the volume-less 18th/19th-century series, and modern Phil. Trans. A is a journal. TYPESETTING TRAP for the .bib: the title carries two degree signs (U+00B0). Under pdflatex+T1 write them as ° (and keep the ASCII hyphens; Crossref’s stored title uses hyphen-minus, not U+2212 – do not ’improve’ it to a Unicode minus, which is fatal outside math mode)., pp. 85–189. DOI: 10.1098/rsta.1985.0058.

[1763] R. E. Slusher, L. W. Hollberg, B. Yurke, J. C. Mertz, and J. F. Valley. “Observation of Squeezed States Generated by Four-Wave Mixing in an Optical Cavity”. In: *Physical Review Letters* 55.22 (1985), pp. 2409–2412. DOI: 10.1103/PhysRevLett.55.2409.

[1764] I. Tanihata et al. “Measurements of Interaction Cross Sections and Nuclear Radii in the Lightp-Shell Region”. In: *Physical Review Letters* 55.24 (1985), pp. 2676–2679. DOI: 10.1103/PhysRevLett.55.2676.

[1765] J. Tersoff and D. R. Hamann. “Theory of the scanning tunneling microscope”. In: *Physical Review B* 31.2 (1985). Issue 2, 15 January 1985. Crossref-verified; the published title is lowercase after the first word. This is the full Phys. Rev. B paper, not the earlier short version (Tersoff & Hamann, PRL 50, 1998 (1983), ’Theory and Application for the Scanning Tunneling Microscope’). The chapter’s proofpending describes the s-wave tip model reducing the current to the sample local density of states at the tip position, which is the PRB paper’s central result, so the 1985 key is the right work for the 1985 claim., pp. 805–813. DOI: 10.1103/PhysRevB.31.805.

[1766] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano. “Determining Lyapunov exponents from a time series”. In: *Physica D: Nonlinear Phenomena* 16.3 (1985). Standard algorithm for Lyapunov exponents from experimental data., pp. 285–317. DOI: 10.1016/0167-2789(85)90011-9.

[1767] A. Ashkin, J. M. Dziedzic, J. E. Bjorkholm, and S. Chu. “Observation of a single-beam gradient force optical trap for dielectric particles”. In: *Optics Letters* 11.5 (1986), p. 288. DOI: 10.1364/OL.11.000288.

[1768] J. G. Bednorz and K. A. Müller. “Possible highT c superconductivity in the Ba?La?Cu?O system”. In: *Zeitschrift für Physik B Condensed Matter* 64.2 (1986), pp. 189–193. DOI: 10.1007/BF01303701.

[1769] L. S. Brown and G. Gabrielse. “Geonium theory: Physics of a single electron or ion in a Penning trap”. In: *Reviews of Modern Physics* 58.1 (1986), pp. 233–311. DOI: 10.1103/RevModPhys.58.233.

[1770] W. Buchmüller and D. Wyler. “Effective lagrangian analysis of new interactions and flavour conservation”. In: *Nuclear Physics B* 268.3-4 (1986), pp. 621–653. DOI: 10.1016/0550-3213(86)90262-2.

[1771] A. K. Drukier, K. Freese, and D. N. Spergel. “Detecting cold dark-matter candidates”. In: *Physical Review D* 33.12 (1986). Issue 12. The annual-modulation prediction (period one year, phase fixed by the Earth’s orbit, amplitude of order a per cent) plus the standard halo model. Cited in the DAMA/LIBRA inline derivation, at the three-predictions sentence. NOT Drukier & Stodolsky 1984, which is the neutral-current detector paper., pp. 3495–3508. DOI: 10.1103/PhysRevD.33.3495.

[1772] G. C. Ghirardi, A. Rimini, and T. Weber. “Unified dynamics for microscopic and macroscopic systems”. In: *Physical Review D* 34.2 (1986), pp. 470–491. DOI: 10.1103/PhysRevD.34.470.

[1773] M. H. Goroff and A. Sagnotti. “The ultraviolet behavior of Einstein gravity”. In: *Nuclear Physics B* 266.3-4 (1986), pp. 709–736. DOI: 10.1016/0550-3213(86)90193-8.

[1774] P. Grangier, G. Roger, and A. Aspect. “Experimental Evidence for a Photon Anticorrelation Effect on a Beam Splitter: A New Light on Single-Photon Interferences”. In: *Europhysics Letters (EPL)* 1.4 (1986). Second author’s given name carries an accent: Gérard., pp. 173–179. DOI: 10.1209/0295-5075/1/4/004.

[1775] H. Hauptman. “Direct Methods and Anomalous Dispersion (Nobel Lecture)”. In: *Angewandte Chemie International Edition in English* 25.7 (1986). Crossref-verified. Nobel Lecture in Chemistry 1985. A German version exists as Angew. Chem. 98, 600-610 (10.1002/ange.19860980703) – do not key both., pp. 603–613. DOI: 10.1002/anie.198606031.

[1776] H. F. Hess. “Evaporative cooling of magnetically trapped and compressed spin-polarized hydrogen”. In: *Physical Review B* 34.5 (1986), pp. 3476–3479. DOI: 10.1103/PhysRevB.34.3476.

[1777] J. Karle. “Recovering Phase Information from Intensity Data (Nobel Lecture)”. In: *Angewandte Chemie International Edition in English* 25.7 (1986). Crossref-verified; pages 614-629, not 614-626 as sometimes cited. Nobel Lecture in Chemistry 1985. A shorter Science version exists (Science 232, 837-843) – different work, do not conflate., pp. 614–629. DOI: 10.1002/anie.198606141.

[1778] L. D. Landau and E. M. Lifshitz. *Theory of Elasticity*. Third English edition prepared with A. M. Kosevich and L. P. Pitaevskii. Distinct from existing key Landau:1976 (Mechanics, vol. 1). Pergamon Press, Oxford; 3rd English edition (Course of Theoretical Physics, Volume 7), 1986.

[1779] S. P. Mikheyev and A. Y. Smirnov. “Resonant amplification of ν oscillations in matter and solar-neutrino spectroscopy”. In: *Il Nuovo Cimento C* 9.1 (1986). Russian original; English translation in Soviet Journal of Nuclear Physics 42 (1985) 913. No DOI, pp. 17–26. DOI: 10.1007/bf02508049.

[1780] W. Nagourney, J. Sandberg, and H. Dehmelt. “Shelved optical electron amplifier: Observation of quantum jumps”. In: *Physical Review Letters* 56.26 (1986), pp. 2797–2799. DOI: 10.1103/PhysRevLett.56.2797.

[1781] B. Paczyński. “Gravitational microlensing by the galactic halo”. In: *The Astrophysical Journal* 304 (1986), pp. 1–5.

[1782] J. V. Sengers and J. M. H. L. Sengers. “Thermodynamic Behavior of Fluids Near the Critical Point”. In: *Annual Review of Physical Chemistry* 37.1 (1986), pp. 189–222. DOI: 10.1146/annurev.pc.37.100186.001201.

[1783] A. Tomita and R. Y. Chiao. “Observation of Berry’s Topological Phase by Use of an Optical Fiber”. In: *Physical Review Letters* 57.8 (1986). Verified against Crossref CSL metadata (PRL 57(8), 25 Aug 1986). This is the helically-wound single-mode fibre experiment the prose describes. A companion erratum/comment exists but is not what is cited., pp. 937–940. DOI: 10.1103/PhysRevLett.57.937.

[1784] A. Tonomura et al. “Evidence for Aharonov–Bohm effect with magnetic field completely shielded from electron wave”. In: *Physical Review Letters* 56.8 (1986). Crossref-verified (resolved record read: Phys. Rev. Lett. 56(8) 792–795, 1986). The definitive measurement of the electromagnetic Aharonov–Bohm phase: the toroidal magnet is covered by a superconducting shield, so the leakage-field objection to earlier experiments does not apply. Distinct from Tonomura:1989, which is the single-electron interference buildup., pp. 792–795. DOI: 10.1103/PhysRevLett.56.792.

[1785] L.-A. Wu, H. J. Kimble, J. L. Hall, and H. Wu. “Generation of Squeezed States by Parametric Down Conversion”. In: *Physical Review Letters* 57.20 (1986), pp. 2520–2523. DOI: 10.1103/PhysRevLett.57.2520.

[1786] Y. Aharonov and J. Anandan. “Phase change during a cyclic quantum evolution”. In: *Physical Review Letters* 58.16 (1987). The non-adiabatic (Aharonov–Anandan) generalization of Berry’s phase, which is what the citing sentence claims. Crossref-verified., pp. 1593–1596. DOI: 10.1103/PhysRevLett.58.1593.

[1787] H. Albrecht et al. “Observation of B0-B0 mixing”. In: *Physics Letters B* 192.1-2 (1987), pp. 245–252. DOI: 10.1016/0370-2693(87)91177-4.

[1788] J. S. Bell. *Speakable and Unspeakable in Quantum Mechanics*. Bell:1964 already exists in the bib for the original paper; this is the collected-papers volume. Cambridge: Cambridge University Press, 1987.

[1789] R. M. Bionta et al. “Observation of a neutrino burst in coincidence with supernova 1987A in the Large Magellanic Cloud”. In: *Physical Review Letters* 58.14 (1987), pp. 1494–1496. DOI: 10.1103/PhysRevLett.58.1494.

[1790] M. Dantus, M. J. Rosker, and A. H. Zewail. “Real-time femtosecond probing of “transition states” in chemical reactions”. In: *The Journal of Chemical Physics* 87.4 (1987). Issue 4. Title, first author, volume, pages and year confirmed against the Crossref record for this DOI. The published title carries typographic quotation marks around ’transition states’; rendered here with TeX quote ligatures., pp. 2395–2397. DOI: 10.1063/1.453122.

[1791] R. B. Davies. “Hypothesis testing when a nuisance parameter is present only under the alternative”. In: *Biometrika* 74.1 (1987). Crossref-verified. The sequel to Davies:1977, extending the upcrossing bound; the two share a title and are distinct papers ten years apart, so both keys are needed and neither supersedes the other. Not E. B. Davies of Davies:1974., pp. 33–43. DOI: 10.1093/biomet/74.1.33.

[1792] J. Glimm and A. Jaffe. *Quantum Physics: A Functional Integral Point of View*. Second edition. Springer, 1987.

[1793] K. Hirata et al. “Observation of a neutrino burst from the supernova SN1987A”. In: *Physical Review Letters* 58.14 (1987), pp. 1490–1493. DOI: 10.1103/PhysRevLett.58.1490.

[1794] C. K. Hong, Z. Y. Ou, and L. Mandel. “Measurement of subpicosecond time intervals between two photons by interference”. In: *Physical Review Letters* 59.18 (1987), pp. 2044–2046. DOI: 10.1103/PhysRevLett.59.2044.

[1795] A. J. Leggett, S. Chakravarty, A. T. Dorsey, M. P. A. Fisher, A. Garg, and W. Zwerger. “Dynamics of the dissipative two-state system”. In: *Reviews of Modern Physics* 59.1 (1987), pp. 1–85. DOI: 10.1103/RevModPhys.59.1.

[1796] E. L. Raab, M. Prentiss, A. Cable, S. Chu, and D. E. Pritchard. “Trapping of Neutral Sodium Atoms with Radiation Pressure”. In: *Physical Review Letters* 59.23 (1987), pp. 2631–2634. DOI: 10.1103/PhysRevLett.59.2631.

[1797] G. Rempe, H. Walther, and N. Klein. “Observation of quantum collapse and revival in a one-atom maser”. In: *Physical Review Letters* 58.4 (1987), pp. 353–356. DOI: 10.1103/PhysRevLett.58.353.

[1798] R. S. Van Dyck, P. B. Schwinberg, and H. G. Dehmelt. “New high-precision comparison of electron and positrongfactors”. In: *Physical Review Letters* 59.1 (1987). Particle ’Van’ kept in the surname per the convention., pp. 26–29. DOI: 10.1103/PhysRevLett.59.26.

[1799] S. Weinberg. “Anthropic Bound on the Cosmological Constant”. In: *Physical Review Letters* 59.22 (1987), pp. 2607–2610. DOI: 10.1103/PhysRevLett.59.2607.

[1800] R. Willett, J. P. Eisenstein, H. L. Störmer, D. C. Tsui, A. C. Gossard, and J. H. English. “Observation of an even-denominator quantum number in the fractional quantum Hall effect”. In: *Physical Review Letters* 59.15 (1987), pp. 1776–1779. DOI: 10.1103/PhysRevLett.59.1776.

[1801] D. E. Winget et al. “An Independent Method for Determining the Age of the Universe”. In: *The Astrophysical Journal* 315 (1987). Metadata confirmed against Crossref., p. L77. DOI: 10.1086/184864.

[1802] M. K. Wu et al. “Superconductivity at 93 K in a new mixed-phase Y-Ba-Cu-O compound system at ambient pressure”. In: *Physical Review Letters* 58.9 (1987), pp. 908–910. DOI: 10.1103/PhysRevLett.58.908.

[1803] Y. Aharonov, D. Z. Albert, and L. Vaidman. “How the result of a measurement of a component of the spin of a spin-1/2particle can turn out to be 100”. In: *Physical Review Letters* 60.14 (1988), pp. 1351–1354. DOI: 10.1103/PhysRevLett.60.1351.

[1804] J. Ashman et al. “A measurement of the spin asymmetry and determination of the structure function g1 in deep inelastic muon-proton scattering”. In: *Physics Letters B* 206.2 (1988), pp. 364–370. DOI: 10.1016/0370-2693(88)91523-7.

[1805] M. N. Baibich et al. “Giant Magnetoresistance of (001)Fe/(001)Cr Magnetic Superlattices”. In: *Physical Review Letters* 61.21 (1988), pp. 2472–2475. DOI: 10.1103/PhysRevLett.61.2472.

[1806] H. Burkhardt et al. “First evidence for direct CP violation”. In: *Physics Letters B* 206.1 (1988), pp. 169–176. DOI: 10.1016/0370-2693(88)91282-8.

[1807] M. Büttiker. “Absence of backscattering in the quantum Hall effect in multiprobe conductors”. In: *Physical Review B* 38.14 (1988), pp. 9375–9389. DOI: 10.1103/PhysRevB.38.9375.

[1808] F. D. M. Haldane. “Model for a Quantum Hall Effect without Landau Levels: Condensed-Matter Realization of the "Parity Anomaly"”. In: *Physical Review Letters* 61.18 (1988), pp. 2015–2018. DOI: 10.1103/PhysRevLett.61.2015.

[1809] K. S. Krane. *Introductory Nuclear Physics*. John Wiley & Sons, 1988.

[1810] P. D. Lett, R. N. Watts, C. I. Westbrook, W. D. Phillips, P. L. Gould, and H. J. Metcalf. “Observation of Atoms Laser Cooled below the Doppler Limit”. In: *Physical Review Letters* 61.2 (1988), pp. 169–172. DOI: 10.1103/PhysRevLett.61.169.

[1811] N. Masuhara et al. “Evaporative Cooling of Spin-Polarized Atomic Hydrogen”. In: *Physical Review Letters* 61.8 (1988), pp. 935–938. DOI: 10.1103/PhysRevLett.61.935.

[1812] T. S. Rose, M. J. Rosker, and A. H. Zewail. “Femtosecond real-time observation of wave packet oscillations (resonance) in dissociation reactions”. In: *The Journal of Chemical Physics* 88.10 (1988). Issue 10. The sodium iodide experiment. Title, first author, volume, pages and year confirmed against the Crossref record for this DOI., pp. 6672–6673. DOI: 10.1063/1.454408.

[1813] M. J. Rosker, M. Dantus, and A. H. Zewail. “Femtosecond Clocking of the Chemical Bond”. In: *Science* 241.4870 (1988). Issue 4870. Title, first author, volume, pages and year confirmed against the Crossref record for this DOI., pp. 1200–1202. DOI: 10.1126/science.241.4870.1200.

[1814] A. Zeilinger, R. Gähler, C. G. Shull, W. Treimer, and W. Mampe. “Single- and double-slit diffraction of neutrons”. In: *Reviews of Modern Physics* 60.4 (1988). Issue 4; issued October 1988. A review-journal article reporting original measurements (cold neutrons at 20-100 micrometre single- and double-slit assemblies), which is what the chapter’s claim needs – the neutron single- and double-slit demonstration extending interference to a massive neutral particle. Second author is Gähler: the parts/ tree and references.bib are ASCII with accent macros, so the bib field needs Gähler, not the raw Unicode given here. Verified via Crossref (title, volume 60, issue 4, pp. 1067-1073, issued 1988). No collision: no Zeilinger:* key exists yet, though Zeilinger appears as a co-author inside the existing Rauch:1975., pp. 1067–1073. DOI: 10.1103/RevModPhys.60.1067.

[1815] A. Benvenuti et al. “A high statistics measurement of the proton structure functions F2(x, Q2) and R from deep inelastic muon scattering at high Q2”. In: *Physics Letters B* 223.3-4 (1989), pp. 485–489. DOI: 10.1016/0370-2693(89)91637-7.

[1816] G. Binasch, P. Grünberg, F. Saurenbach, and W. Zinn. “Enhanced magnetoresistance in layered magnetic structures with antiferromagnetic interlayer exchange”. In: *Physical Review B* 39.7 (1989), pp. 4828–4830. DOI: 10.1103/PhysRevB.39.4828.

[1817] J. C. Collins, D. E. Soper, and G. Sterman. “Factorization of hard processes in QCD”. In: *Perturbative Quantum Chromodynamics*. Ed. by A. H. Mueller. Vol. 5. Advanced Series on Directions in High Energy Physics. The all-orders collinear-factorization proof for hard processes. Metadata confirmed against Crossref. World Scientific, 1989, pp. 1–91. DOI: 10.1142/9789814503266_0001.

[1818] J. Dalibard and C. Cohen-Tannoudji. “Laser cooling below the Doppler limit by polarization gradients: simple theoretical models”. In: *Journal of the Optical Society of America B* 6.11 (1989), p. 2023. DOI: 10.1364/JOSAB.6.002023.

[1819] D. DeCamp et al. “Determination of the number of light neutrino species”. In: *Physics Letters B* 231.4 (1989), pp. 519–529. DOI: 10.1016/0370-2693(89)90704-1.

[1820] L. Diósi. “Models for universal reduction of macroscopic quantum fluctuations”. In: *Physical Review A* 40.3 (1989), pp. 1165–1174. DOI: 10.1103/PhysRevA.40.1165.

[1821] D. M. Greenberger, M. A. Horne, and A. Zeilinger. “Going Beyond Bell’s Theorem”. In: *Bell’s Theorem, Quantum Theory and Conceptions of the Universe*. Contribution to the volume edited by M. Kafatos; @incollection/@inproceedings with booktitle, no journal volume. 1989, pp. 69–72. DOI: 10.1007/978-94-017-0849-4_10.

[1822] K. S. Hirata et al. “Observation ofB8solar neutrinos in the Kamiokande-II detector”. In: *Physical Review Letters* 63.1 (1989), pp. 16–19. DOI: 10.1103/PhysRevLett.63.16.

[1823] J. K. Jain. “Composite-fermion approach for the fractional quantum Hall effect”. In: *Physical Review Letters* 63.2 (1989), pp. 199–202. DOI: 10.1103/PhysRevLett.63.199.

[1824] J. Laskar. “A numerical experiment on the chaotic behaviour of the Solar System”. In: *Nature* 338.6212 (1989). Inner solar system chaotic, Lyapunov time about 5 Myr., pp. 237–238. DOI: 10.1038/338237a0.

[1825] P. Pearle. “Combining stochastic dynamical state-vector reduction with spontaneous localization”. In: *Physical Review A* 39.5 (1989), pp. 2277–2289. DOI: 10.1103/PhysRevA.39.2277.

[1826] A. Tonomura, J. Endo, T. Matsuda, T. Kawasaki, and H. Ezawa. “Demonstration of single-electron buildup of an interference pattern”. In: *American Journal of Physics* 57.2 (1989), pp. 117–120. DOI: 10.1119/1.16104.

[1827] T. C. Weekes et al. “Observation of TeV gamma rays from the Crab nebula using the atmospheric Cerenkov imaging technique”. In: *The Astrophysical Journal* 342 (1989), p. 379. DOI: 10.1086/167599.

[1828] S. Weinberg. “The cosmological constant problem”. In: *Reviews of Modern Physics* 61.1 (1989), pp. 1–23. DOI: 10.1103/RevModPhys.61.1.

[1829] R. F. Werner. “Quantum states with Einstein-Podolsky-Rosen correlations admitting a hidden-variable model”. In: *Physical Review A* 40.8 (1989), pp. 4277–4281. DOI: 10.1103/PhysRevA.40.4277.

[1830] T. R. Anthony et al. “Thermal diffusivity of isotopically enrichedC12diamond”. In: *Physical Review B* 42.2 (1990), pp. 1104–1111. DOI: 10.1103/PhysRevB.42.1104.

[1831] G. C. Ghirardi, P. Pearle, and A. Rimini. “Markov processes in Hilbert space and continuous spontaneous localization of systems of identical particles”. In: *Physical Review A* 42.1 (1990), pp. 78–89. DOI: 10.1103/PhysRevA.42.78.

[1832] D. M. Greenberger, M. A. Horne, A. Shimony, and A. Zeilinger. “Bell’s theorem without inequalities”. In: *American Journal of Physics* 58.12 (1990), pp. 1131–1143. DOI: 10.1119/1.16243.

[1833] W. M. Itano, D. J. Heinzen, J. J. Bollinger, and D. J. Wineland. “Quantum Zeno effect”. In: *Physical Review A* 41.5 (1990), pp. 2295–2300. DOI: 10.1103/PhysRevA.41.2295.

[1834] J. C. Mather et al. “A preliminary measurement of the cosmic microwave background spectrum by the Cosmic Background Explorer (COBE) satellite”. In: *The Astrophysical Journal* 354 (1990). COBE/FIRAS; collaboration paper keyed by first listed author., p. L37. DOI: 10.1086/185717.

[1835] M. E. Peskin and T. Takeuchi. “New constraint on a strongly interacting Higgs sector”. In: *Physical Review Letters* 65.8 (1990), pp. 964–967. DOI: 10.1103/PhysRevLett.65.964.

[1836] R. E. Prange and S. M. (Girvin. *The Quantum Hall Effect*. 2nd edition; edited volume. Springer, 1990.

[1837] E. Ramberg and G. A. Snow. “Experimental limit on a small violation of the Pauli principle”. In: *Physics Letters B* 238.2-4 (1990), pp. 438–441. DOI: 10.1016/0370-2693(90)91762-Z.

[1838] J. A. Tyson, R. A. Wenk, and F. Valdes. “Detection of systematic gravitational lens galaxy image alignments - Mapping dark matter in galaxy clusters”. In: *The Astrophysical Journal* 349 (1990), p. L1. DOI: 10.1086/185636.

[1839] D. Vollhardt and P. Wölfle. *The Superfluid Phases of Helium 3*. Taylor and Francis, 1990.

[1840] S. Weinberg. “Nuclear forces from chiral lagrangians”. In: *Physics Letters B* 251.2 (1990), pp. 288–292. DOI: 10.1016/0370-2693(90)90938-3.

[1841] L. E. Ballentine. “Comment on “Quantum Zeno effect””. In: *Physical Review A* 43.9 (1991), pp. 5165–5167. DOI: 10.1103/PhysRevA.43.5165.

[1842] J. M. Deutsch. “Quantum statistical mechanics in a closed system”. In: *Physical Review A* 43.4 (1991), pp. 2046–2049. DOI: 10.1103/PhysRevA.43.2046.

[1843] A. K. Ekert. “Quantum cryptography based on Bell’s theorem”. In: *Physical Review Letters* 67.6 (1991), pp. 661–663. DOI: 10.1103/PhysRevLett.67.661.

[1844] J. I. Friedman. “Deep inelastic scattering: Comparisons with the quark model”. In: *Reviews of Modern Physics* 63.3 (1991). Nobel lecture, pp. 615–627. DOI: 10.1103/RevModPhys.63.615.

[1845] N. Gisin. “Bell’s inequality holds for all non-product states”. In: *Physics Letters A* 154.5-6 (1991), pp. 201–202. DOI: 10.1016/0375-9601(91)90805-I.

[1846] A. Hartland, K. Jones, J. M. Williams, B. L. Gallagher, and T. Galloway. “Direct comparison of the quantized Hall resistance in gallium arsenide and silicon”. In: *Physical Review Letters* 66.8 (1991), pp. 969–973. DOI: 10.1103/physrevlett.66.969.

[1847] D. W. Keith, C. R. Ekstrom, Q. A. Turchette, and D. E. Pritchard. “An interferometer for atoms”. In: *Physical Review Letters* 66.21 (1991), pp. 2693–2696. DOI: 10.1103/PhysRevLett.66.2693.

[1848] H. W. Kendall. “Deep inelastic scattering: Experiments on the proton and the observation of scaling”. In: *Reviews of Modern Physics* 63.3 (1991). Nobel lecture, pp. 597–614. DOI: 10.1103/RevModPhys.63.597.

[1849] R. E. Taylor. “Deep inelastic scattering: The early years”. In: *Reviews of Modern Physics* 63.3 (1991). Nobel lecture; Richard E. Taylor, distinct from the Taylors of Taylor:1982 and Taylor:1997, pp. 573–595. DOI: 10.1103/RevModPhys.63.573.

[1850] J. Casares, P. A. Charles, and T. Naylor. “A 6.5-day periodicity in the recurrent nova V404 Cygni implying the presence of a black hole”. In: *Nature* 355 (1992). Metadata confirmed against Crossref., pp. 614–617. DOI: 10.1038/355614a0.

[1851] J. Dalibard, Y. Castin, and K. Mølmer. “Wave-function approach to dissipative processes in quantum optics”. In: *Physical Review Letters* 68.5 (1992), pp. 580–583. DOI: 10.1103/PhysRevLett.68.580.

[1852] R. C. Duncan and C. Thompson. “Formation of very strongly magnetized neutron stars - Implications for gamma-ray bursts”. In: *The Astrophysical Journal* 392 (1992), p. L9. DOI: 10.1086/186413.

[1853] J. Feldmann et al. “Optical investigation of Bloch oscillations in a semiconductor superlattice”. In: *Physical Review B* 46.11 (1992), pp. 7252–7255. DOI: 10.1103/physrevb.46.7252.

[1854] N. Gisin and I. C. Percival. “The quantum-state diffusion model applied to open systems”. In: *Journal of Physics A: Mathematical and General* 25.21 (1992), pp. 5677–5691. DOI: 10.1088/0305-4470/25/21/023.

[1855] N. Goldenfeld. *Lectures on Phase Transitions and the Renormalization Group*. Addison-Wesley, 1992.

[1856] R. Haag. *Local Quantum Physics: Fields, Particles, Algebras*. Springer, 1992.

[1857] R. Omnès. “Consistent interpretations of quantum mechanics”. In: *Reviews of Modern Physics* 64.2 (1992), pp. 339–382. DOI: 10.1103/RevModPhys.64.339.

[1858] G. F. Smoot et al. “Structure in the COBE differential microwave radiometer first-year maps”. In: *The Astrophysical Journal* 396 (1992). COBE/DMR; collaboration paper keyed by first listed author., p. L1. DOI: 10.1086/186504.

[1859] R. J. Thompson, G. Rempe, and H. J. Kimble. “Observation of normal-mode splitting for an atom in an optical cavity”. In: *Physical Review Letters* 68.8 (1992), pp. 1132–1135. DOI: 10.1103/PhysRevLett.68.1132.

[1860] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters. “Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels”. In: *Physical Review Letters* 70.13 (1993), pp. 1895–1899. DOI: 10.1103/PhysRevLett.70.1895.

[1861] R. R. Du, H. L. Stormer, D. C. Tsui, L. N. Pfeiffer, and K. W. West. “Experimental evidence for new particles in the fractional quantum Hall effect”. In: *Physical Review Letters* 70.19 (1993), pp. 2944–2947. DOI: 10.1103/PhysRevLett.70.2944.

[1862] P. H. Eberhard. “Background level and counter efficiencies required for a loophole-free Einstein-Podolsky-Rosen experiment”. In: *Physical Review A* 47.2 (1993), R747–R750. DOI: 10.1103/PhysRevA.47.R747.

[1863] D. J. Evans, E. G. D. Cohen, and G. P. Morriss. “Probability of second law violations in shearing steady states”. In: *Physical Review Letters* 71.15 (1993), pp. 2401–2404. DOI: 10.1103/PhysRevLett.71.2401.

[1864] M. S. Fee et al. “Measurement of the positronium13S1-23S1interval by continuous-wave two-photon excitation”. In: *Physical Review Letters* 70.10 (1993), pp. 1397–1400. DOI: 10.1103/PhysRevLett.70.1397.

[1865] M. Gell-Mann and J. B. Hartle. “Classical equations for quantum systems”. In: *Physical Review D* 47.8 (1993), pp. 3345–3382. DOI: 10.1103/PhysRevD.47.3345.

[1866] P. G. de Gennes and J. Prost. *The Physics of Liquid Crystals*. 2nd edition. Oxford University Press, 1993.

[1867] H. Georgi. “Effective Field Theory”. In: *Annual Review of Nuclear and Particle Science* 43.1 (1993), pp. 209–252. DOI: 10.1146/annurev.ns.43.120193.001233.

[1868] L. P. Hughston, R. Jozsa, and W. K. Wootters. “A complete classification of quantum ensembles having a given density matrix”. In: *Physics Letters A* 183.1 (1993), pp. 14–18. DOI: 10.1016/0375-9601(93)90880-9.

[1869] J. D. Norton. “General covariance and the foundations of general relativity: eight decades of dispute”. In: *Reports on Progress in Physics* 56 (1993), pp. 791–858. DOI: 10.1088/0034-4885/56/7/001.

[1870] D. N. Page. “Information in black hole radiation”. In: *Physical Review Letters* 71.23 (1993), pp. 3743–3746. DOI: 10.1103/PhysRevLett.71.3743.

[1871] P. J. E. Peebles. *Principles of Physical Cosmology*. Princeton University Press, 1993.

[1872] A. Peres. *Quantum Theory: Concepts and Methods*. Dordrecht: Kluwer Academic Publishers, 1993.

[1873] M. M. Phillips. “The Absolute Magnitudes of Type Ia Supernovae”. In: *The Astrophysical Journal* 413 (1993). Metadata confirmed against Crossref., pp. L105–L108. DOI: 10.1086/186970.

[1874] V. Stoks, R. Timmermans, and J. J. de Swart. “Pion-nucleon coupling constant”. In: *Physical Review C* 47.2 (1993). Nijmegen partial-wave determination: $f^2 = 0.0745(6)$, i.e. $g^2/4\pi = 13.54$ and $g_{\pi N} = 13.05$. Metadata confirmed against Crossref., pp. 512–520. DOI: 10.1103/PhysRevC.47.512.

[1875] C. I. Sukenik, M. G. Boshier, D. Cho, V. Sandoghdar, and E. A. Hinds. “Measurement of the Casimir-Polder force”. In: *Physical Review Letters* 70.5 (1993), pp. 560–563. DOI: 10.1103/PhysRevLett.70.560.

[1876] C. Wetterich. “Exact evolution equation for the effective potential”. In: *Physics Letters B* 301.1 (1993), pp. 90–94. DOI: 10.1016/0370-2693(93)90726-X.

[1877] D. A. Wollman, D. J. Van Harlingen, W. C. Lee, D. M. Ginsberg, and A. J. Leggett. “Experimental determination of the superconducting pairing state in YBCO from the phase coherence of YBCO-Pb dc SQUIDs”. In: *Physical Review Letters* 71.13 (1993), pp. 2134–2137. DOI: 10.1103/PhysRevLett.71.2134.

[1878] M. Żukowski, A. Zeilinger, M. A. Horne, and A. K. Ekert. ““Event-ready-detectors” Bell experiment via entanglement swapping”. In: *Physical Review Letters* 71.26 (1993). Key drops the diacritic (Zukowski) to stay ASCII., pp. 4287–4290. DOI: 10.1103/PhysRevLett.71.4287.

[1879] M. Arneodo, A. Arvidson, B. Badelek, et al. “Reevaluation of the Gottfried sum”. In: *Physical Review D* 50.1 (1994). New Muon Collaboration. The Gottfried sum measured well below 1/3, i.e. an isospin-asymmetric light-quark sea. Metadata confirmed against Crossref., R1–R3. DOI: 10.1103/PhysRevD.50.R1.

[1880] J. F. Donoghue. “General relativity as an effective field theory: The leading quantum corrections”. In: *Physical Review D* 50.6 (1994), pp. 3874–3888. DOI: 10.1103/PhysRevD.50.3874.

[1881] T. F. Gallagher. *Rydberg Atoms*. Verified against Crossref as a monograph, Cambridge University Press, 1994. Cambridge Monographs on Atomic, Molecular and Chemical Physics, no. 3; ISBN 9780521385312. Paywalled but with a resolving DOI, so it grades AMBER honestly. Cambridge: Cambridge University Press, 1994. DOI: 10.1017/CBO9780511524530.

[1882] M. B. GAVELA, P. HERNÁNDEZ, J. ORLOFF, and O. PÈNE. “STANDARD MODEL CP-VIOLATION AND BARYON ASYMMETRY”. In: *Modern Physics Letters A* 09.09 (1994), pp. 795–809. DOI: 10.1142/S0217732394000629.

[1883] S. W. Hell and J. Wichmann. “Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy”. In: *Optics Letters* 19.11 (1994). Issue 11. The founding STED paper. Crossref records first page 780; the paper ends on 782., p. 780. DOI: 10.1364/OL.19.000780.

[1884] J. C. Mather et al. “Measurement of the cosmic microwave background spectrum by the COBE FIRAS instrument”. In: *The Astrophysical Journal* 420 (1994), p. 439. DOI: 10.1086/173574.

[1885] S. Nakamura, T. Mukai, and M. Senoh. “Candela-class high-brightness InGaN/AlGaN double-heterostructure blue-light-emitting diodes”. In: *Applied Physics Letters* 64.13 (1994), pp. 1687–1689. DOI: 10.1063/1.111832.

[1886] S. Popescu and D. Rohrlich. “Quantum nonlocality as an axiom”. In: *Foundations of Physics* 24.3 (1994), pp. 379–385. DOI: 10.1007/BF02058098.

[1887] M. Srednicki. “Chaos and quantum thermalization”. In: *Physical Review E* 50.2 (1994), pp. 888–901. DOI: 10.1103/PhysRevE.50.888.

[1888] S. H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Canonical modern text for the chapter. MA: Addison-Wesley, Reading, 1994.

[1889] C. C. Tsuei et al. “Pairing Symmetry and Flux Quantization in a Tricrystal Superconducting Ring ofYBa2Cu3O7−δ”. In: *Physical Review Letters* 73.4 (1994), pp. 593–596. DOI: 10.1103/PhysRevLett.73.593.

[1890] R. M. Wald. *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics*. University of Chicago Press, 1994.

[1891] S. Abachi et al. “Observation of the Top Quark”. In: *Physical Review Letters* 74.14 (1995), pp. 2632–2637. DOI: 10.1103/PhysRevLett.74.2632.

[1892] F. Abe et al. “Observation of Top Quark Production inp¯pCollisions with the Collider Detector at Fermilab”. In: *Physical Review Letters* 74.14 (1995), pp. 2626–2631. DOI: 10.1103/PhysRevLett.74.2626.

[1893] M. H. Anderson, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell. “Observation of Bose-Einstein Condensation in a Dilute Atomic Vapor”. In: *Science* 269.5221 (1995). M. H. Anderson; distinct from the Carl D. Anderson of Anderson:1933, pp. 198–201. DOI: 10.1126/science.269.5221.198.

[1894] P. Billingsley. *Probability and Measure*. 3rd ed. Lévy’s continuity theorem is Theorem 26.3; the Lindeberg–Feller theorem is Theorem 27.2. New York: Wiley, 1995.

[1895] D. J. Bird et al. “Detection of a cosmic ray with measured energy well beyond the expected spectral cutoff due to cosmic microwave radiation”. In: *The Astrophysical Journal* 441 (1995), p. 144. DOI: 10.1086/175344.

[1896] C. C. Bradley, C. A. Sackett, J. J. Tollett, and R. G. Hulet. “Evidence of Bose-Einstein Condensation in an Atomic Gas with Attractive Interactions”. In: *Physical Review Letters* 75.9 (1995), pp. 1687–1690. DOI: 10.1103/PhysRevLett.75.1687.

[1897] B. N. Brockhouse. “Slow neutron spectroscopy and the grand atlas of the physical world”. In: *Reviews of Modern Physics* 67.4 (1995). Nobel Lecture, 8 December 1994; issue 4. Crossref-verified: title, venue, vol. 67, pp. 735-751, published print October 1995. Placed at the sentence naming Brockhouse’s half of the 1994 prize (line 48)., pp. 735–751. DOI: 10.1103/RevModPhys.67.735.

[1898] K. B. Davis et al. “Bose-Einstein Condensation in a Gas of Sodium Atoms”. In: *Physical Review Letters* 75.22 (1995). K. B. Davis; distinct from the Martin Davis of Davis:1973, pp. 3969–3973. DOI: 10.1103/PhysRevLett.75.3969.

[1899] G. Gallavotti and E. G. D. Cohen. “Dynamical Ensembles in Nonequilibrium Statistical Mechanics”. In: *Physical Review Letters* 74.14 (1995), pp. 2694–2697. DOI: 10.1103/PhysRevLett.74.2694.

[1900] G. A. Glatzmaier and P. H. Roberts. “A three-dimensional self-consistent computer simulation of a geomagnetic field reversal”. In: *Nature* 377 (1995). Verified against Crossref. Crossref’s metadata misspells the first author as ’Glatzmaiers’; the published spelling is Glatzmaier, which is also what the citation key uses – do not copy the Crossref typo into the bib entry., pp. 203–209.

[1901] K. Kuroda. “Does the Time-of-Swing Method Give a Correct Value of the Newtonian Gravitational Constant?” In: *Physical Review Letters* 75.15 (1995). Fibre anelasticity as a bias of the time-of-swing method., pp. 2796–2798. DOI: 10.1103/PhysRevLett.75.2796.

[1902] P. G. Kwiat, K. Mattle, H. Weinfurter, A. Zeilinger, A. V. Sergienko, and Y. Shih. “New High-Intensity Source of Polarization-Entangled Photon Pairs”. In: *Physical Review Letters* 75.24 (1995), pp. 4337–4341. DOI: 10.1103/PhysRevLett.75.4337.

[1903] D. E. Lebach et al. “Measurement of the Solar Gravitational Deflection of Radio Waves Using Very-Long-Baseline Interferometry”. In: *Physical Review Letters* 75.8 (1995). Crossref confirmed title, all eight authors, volume, pages and year against the DOI. The VLBI deflection measurement that reaches the part-in-10^4 level quoted in phen:clt-deflection; cited there and again in the eclipse Interpretation alongside Fomalont:2009, replacing the second-hand reliance on Will:2014 alone., pp. 1439–1442. DOI: 10.1103/PhysRevLett.75.1439.

[1904] L. Mandel and E. Wolf. *Optical Coherence and Quantum Optics*. Cambridge University Press, 1995.

[1905] T. Maudlin. “Three measurement problems”. In: *Topoi* 14.1 (1995), pp. 7–15. DOI: 10.1007/BF00763473.

[1906] T. M. Niebauer, G. S. Sasagawa, J. E. Faller, R. Hilt, and F. Klopping. “A new generation of absolute gravimeters”. In: *Metrologia* 32 (1995), pp. 159–180. DOI: 10.1088/0026-1394/32/3/004.

[1907] M. E. Peskin and D. V. Schroeder. *An Introduction to Quantum Field Theory*. Addison-Wesley, 1995.

[1908] W. Petrich, M. H. Anderson, J. R. Ensher, and E. A. Cornell. “Stable, Tightly Confining Magnetic Trap for Evaporative Cooling of Neutral Atoms”. In: *Physical Review Letters* 74.17 (1995), pp. 3352–3355. DOI: 10.1103/PhysRevLett.74.3352.

[1909] P. W. Shor. “Scheme for reducing decoherence in quantum computer memory”. In: *Physical Review A* 52.4 (1995), R2493–R2496. DOI: 10.1103/PhysRevA.52.R2493.

[1910] C. G. Shull. “Early development of neutron scattering”. In: *Reviews of Modern Physics* 67.4 (1995). Nobel Lecture, 8 December 1994; issue 4, immediately following Brockhouse’s in the same issue. Crossref-verified. Placed at the sentence naming Shull’s half of the 1994 prize (line 48). Distinct from the existing Shull:1949 (Shull & Smart, Detection of Antiferromagnetism) and Shull:1951., pp. 753–757. DOI: 10.1103/RevModPhys.67.753.

[1911] Y. Tanaka et al. “Gravitationally redshifted emission implying an accretion disk and massive black hole in the active galaxy MCG–6–30–15”. In: *Nature* 375.6533 (1995), pp. 659–661. DOI: 10.1038/375659a0.

[1912] S. Weinberg. *The Quantum Theory of Fields. Volume I: Foundations*. Cambridge: Cambridge University Press, 1995.

[1913] R. B. Wiringa, V. G. J. Stoks, and R. Schiavilla. "Accurate nucleon-nucleon potential with charge-independence breaking". In: *Physical Review C* 51.1 (1995), pp. 38–51. DOI: 10.1103/PhysRevC.51.38.

[1914] M. R. Andrews, M.-O. Mewes, N. J. van Druten, D. S. Durfee, D. M. Kurn, and W. Ketterle. "Direct, Nondestructive Observation of a Bose Condensate". In: *Science* 273.5271 (1996), pp. 84–87. DOI: 10.1126/science.273.5271.84.

[1915] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters. "Mixed-state entanglement and quantum error correction". In: *Physical Review A* 54.5 (1996), pp. 3824–3851. DOI: 10.1103/PhysRevA.54.3824.

[1916] M. Brune et al. "Observing the Progressive Decoherence of the "Meter" in a Quantum Measurement". In: *Physical Review Letters* 77.24 (1996), pp. 4887–4890. DOI: 10.1103/PhysRevLett.77.4887.

[1917] M. Brune et al. "Quantum Rabi Oscillation: A Direct Test of Field Quantization in a Cavity". In: *Physical Review Letters* 76.11 (1996), pp. 1800–1803. DOI: 10.1103/PhysRevLett.76.1800.

[1918] J. Cardy. *Scaling and Renormalization in Statistical Physics*. Cambridge University Press, 1996.

[1919] T. Damour and F. Dyson. "The Oklo bound on the time variation of the fine-structure constant revisited". In: *Nuclear Physics B* 480.1-2 (1996), pp. 37–54. DOI: 10.1016/S0550-3213(96)00467-1.

[1920] B.-G. Englert. "Fringe Visibility and Which-Way Information: An Inequality". In: *Physical Review Letters* 77.11 (1996), pp. 2154–2157. DOI: 10.1103/PhysRevLett.77.2154.

[1921] J. R. Ensher, D. S. Jin, M. R. Matthews, C. E. Wieman, and E. A. Cornell. "Bose-Einstein Condensation in a Dilute Gas: Measurement of Energy and Ground-State Occupation". In: *Physical Review Letters* 77.25 (1996), pp. 4984–4987. DOI: 10.1103/PhysRevLett.77.4984.

[1922] D. J. Fixsen, E. S. Cheng, J. M. Gales, J. C. Mather, R. A. Shafer, and E. L. Wright. "The Cosmic Microwave Background Spectrum from the FullCOBEFIRAS Data Set". In: *The Astrophysical Journal* 473.2 (1996), pp. 576–587. DOI: 10.1086/178173.

[1923] M. Horodecki, P. Horodecki, and R. Horodecki. "Separability of mixed states: necessary and sufficient conditions". In: *Physics Letters A* 223.1-2 (1996), pp. 1–8. DOI: 10.1016/S0375-9601(96)00706-2.

[1924] D. S. Jin, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell. "Collective Excitations of a Bose-Einstein Condensate in a Dilute Gas". In: *Physical Review Letters* 77.3 (1996), pp. 420–423. DOI: 10.1103/PhysRevLett.77.420.

[1925] K. Kajantie, M. Laine, K. Rummukainen, and M. Shaposhnikov. "Is There a Hot Electroweak Phase Transition atmH$\gtrsim$mW?" In: *Physical Review Letters* 77.14 (1996), pp. 2887–2890. DOI: 10.1103/PhysRevLett.77.2887.

[1926] R. Penrose. "On Gravity's role in Quantum State Reduction". In: *General Relativity and Gravitation* 28.5 (1996), pp. 581–600. DOI: 10.1007/BF02105068.

[1927] J. P. Perdew, K. Burke, and M. Ernzerhof. "Generalized Gradient Approximation Made Simple". In: *Physical Review Letters* 77.18 (1996), pp. 3865–3868. DOI: 10.1103/PhysRevLett.77.3865.

[1928] A. Peres. "Separability Criterion for Density Matrices". In: *Physical Review Letters* 77.8 (1996), pp. 1413–1415. DOI: 10.1103/PhysRevLett.77.1413.

[1929] C. Rovelli. “Relational quantum mechanics”. In: *International Journal of Theoretical Physics* 35.8 (1996), pp. 1637–1678. DOI: 10.1007/BF02302261.

[1930] J. Slonczewski. “Current-driven excitation of magnetic multilayers”. In: *Journal of Magnetism and Magnetic Materials* 159.1-2 (1996), pp. L1–L7. DOI: 10.1016/0304-8853(96)00062-5.

[1931] A. M. Steane. “Error Correcting Codes in Quantum Theory”. In: *Physical Review Letters* 77.5 (1996), pp. 793–797. DOI: 10.1103/PhysRevLett.77.793.

[1932] M. Tinkham. *Introduction to Superconductivity*. Second edition. McGraw-Hill, 1996.

[1933] S. Weinberg. *The Quantum Theory of Fields. Volume II: Modern Applications*. Cambridge University Press, 1996.

[1934] M. R. Andrews, C. G. Townsend, H.-J. Miesner, D. S. Durfee, D. M. Kurn, and W. Ketterle. “Observation of Interference Between Two Bose Condensates”. In: *Science* 275.5300 (1997), pp. 637–641. DOI: 10.1126/science.275.5300.637.

[1935] D. Bouwmeester, J.-W. Pan, K. Mattle, M. Eibl, H. Weinfurter, and A. Zeilinger. “Experimental quantum teleportation”. In: *Nature* 390.6660 (1997), pp. 575–579. DOI: 10.1038/37539.

[1936] I. L. Chuang and M. A. Nielsen. “Prescription for experimental determination of the dynamics of a quantum black box”. In: *Journal of Modern Optics* 44.11-12 (1997), pp. 2455–2467. DOI: 10.1080/09500349708231894.

[1937] L. Corry, J. Renn, and J. Stachel. “Belated Decision in the Hilbert-Einstein Priority Dispute”. In: *Science* 278.5341 (1997), pp. 1270–1273. DOI: 10.1126/science.278.5341.1270.

[1938] C. Jarzynski. “Nonequilibrium Equality for Free Energy Differences”. In: *Physical Review Letters* 78.14 (1997), pp. 2690–2693. DOI: 10.1103/PhysRevLett.78.2690.

[1939] S. K. Lamoreaux. “Demonstration of the Casimir Force in the 0.6 to6μmRange”. In: *Physical Review Letters* 78.1 (1997), pp. 5–8. DOI: 10.1103/PhysRevLett.78.5.

[1940] M.-O. Mewes, M. R. Andrews, D. M. Kurn, D. S. Durfee, C. G. Townsend, and W. Ketterle. “Output Coupler for Bose-Einstein Condensed Atoms”. In: *Physical Review Letters* 78.4 (1997), pp. 582–585. DOI: 10.1103/PhysRevLett.78.582.

[1941] R. Narayan, M. R. Garcia, and J. E. McClintock. “Advection-dominated Accretion and Black Hole Event Horizons”. In: *The Astrophysical Journal* 478 (1997). Metadata confirmed against Crossref., pp. L79–L82. DOI: 10.1086/310554.

[1942] J. F. Navarro, C. S. Frenk, and S. D. M. White. “A Universal Density Profile from Hierarchical Clustering”. In: *The Astrophysical Journal* 490.2 (1997), pp. 493–508. DOI: 10.1086/304888.

[1943] R. de-Picciotto, M. Reznikov, M. Heiblum, V. Umansky, G. Bunin, and D. Mahalu. “Direct observation of a fractional charge”. In: *Nature* 389.6647 (1997), pp. 162–164. DOI: 10.1038/38241.

[1944] L. Saminadayar, D. C. Glattli, Y. Jin, and B. Etienne. “Observation of thee/3Fractionally Charged Laughlin Quasiparticle”. In: *Physical Review Letters* 79.13 (1997), pp. 2526–2529. DOI: 10.1103/PhysRevLett.79.2526.

[1945] J. R. Taylor. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. 2nd ed. Sausalito, California: University Science Books, 1997.

[1946] C. S. Wood et al. “Measurement of Parity Nonconservation and an Anapole Moment in Cesium”. In: *Science* 275.5307 (1997), pp. 1759–1763. DOI: 10.1126/science.275.5307.1759.

[1947] G. Amelino-Camelia, J. Ellis, N. E. Mavromatos, D. V. Nanopoulos, and S. Sarkar. "Tests of quantum gravity from observations of γ-ray bursts". In: *Nature* 393.6687 (1998), pp. 763–765. DOI: 10.1038/31647.

[1948] A. Angelopoulos et al. "First direct observation of time-reversal non-invariance in the neutral-kaon system". In: *Physics Letters B* 444.1-2 (1998), pp. 43–51. DOI: 10.1016/S0370-2693(98)01356-2.

[1949] S. Chu. "Nobel Lecture: The manipulation of neutral particles". In: *Reviews of Modern Physics* 70.3 (1998), pp. 685–706. DOI: 10.1103/RevModPhys.70.685.

[1950] B. T. Cleveland et al. "Measurement of the Solar Electron Neutrino Flux with the Homestake Chlorine Detector". In: *The Astrophysical Journal* 496.1 (1998), pp. 505–526. DOI: 10.1086/305343.

[1951] C. N. Cohen-Tannoudji. "Nobel Lecture: Manipulating atoms with photons". In: *Reviews of Modern Physics* 70.3 (1998), pp. 707–719. DOI: 10.1103/RevModPhys.70.707.

[1952] D. Colladay and V. A. Kostelecký. "Lorentz-violating extension of the standard model". In: *Physical Review D* 58.11 (1998), p. 116002. DOI: 10.1103/PhysRevD.58.116002.

[1953] D. G. Fried et al. "Bose-Einstein Condensation of Atomic Hydrogen". In: *Physical Review Letters* 81.18 (1998), pp. 3811–3814. DOI: 10.1103/PhysRevLett.81.3811.

[1954] Y. Fukuda et al. "Evidence for Oscillation of Atmospheric Neutrinos". In: *Physical Review Letters* 81.8 (1998), pp. 1562–1567. DOI: 10.1103/PhysRevLett.81.1562.

[1955] M. Göckeler, R. Horsley, V. Linke, P. Rakow, G. Schierholz, and H. Stüben. "Is There a Landau Pole Problem in QED?" In: *Physical Review Letters* 80.19 (1998), pp. 4119–4122. DOI: 10.1103/PhysRevLett.80.4119.

[1956] S. Inouye, M. R. Andrews, J. Stenger, H.-J. Miesner, D. M. Stamper-Kurn, and W. Ketterle. "Observation of Feshbach resonances in a Bose–Einstein condensate". In: *Nature* 392.6672 (1998), pp. 151–154. DOI: 10.1038/32354.

[1957] D. A. Lidar, I. L. Chuang, and K. B. Whaley. "Decoherence-Free Subspaces for Quantum Computation". In: *Physical Review Letters* 81.12 (1998), pp. 2594–2597. DOI: 10.1103/PhysRevLett.81.2594.

[1958] U. Mohideen and A. Roy. "Precision Measurement of the Casimir Force from 0.1 to0.9μm". In: *Physical Review Letters* 81.21 (1998), pp. 4549–4552. DOI: 10.1103/PhysRevLett.81.4549.

[1959] W. D. Phillips. "Nobel Lecture: Laser cooling and trapping of neutral atoms". In: *Reviews of Modern Physics* 70.3 (1998), pp. 721–741. DOI: 10.1103/RevModPhys.70.721.

[1960] J. Polchinski. *String Theory. Volume I: An Introduction to the Bosonic String*. Cited solely to name an excluded programme and record its zero-evidence status, per editorial scope rule 1. Cambridge University Press, 1998.

[1961] M. Reuter. "Nonperturbative evolution equation for quantum gravity". In: *Physical Review D* 57.2 (1998), pp. 971–985. DOI: 10.1103/PhysRevD.57.971.

[1962] A. G. Riess et al. "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant". In: *The Astronomical Journal* 116.3 (1998). High-z Supernova Search Team; keyed by first listed author. Distinct from existing key Riess:2022., pp. 1009–1038. DOI: 10.1086/300499.

[1963] G. Weihs, T. Jennewein, C. Simon, H. Weinfurter, and A. Zeilinger. “Violation of Bell’s Inequality under Strict Einstein Locality Conditions”. In: *Physical Review Letters* 81.23 (1998), pp. 5039–5043. DOI: 10.1103/PhysRevLett.81.5039.

[1964] W. K. Wootters. “Entanglement of Formation of an Arbitrary State of Two Qubits”. In: *Physical Review Letters* 80.10 (1998), pp. 2245–2248. DOI: 10.1103/PhysRevLett.80.2245.

[1965] A. Alavi-Harati et al. “Observation of DirectCPViolation inKS,L→ ππDecays”. In: *Physical Review Letters* 83.1 (1999), pp. 22–27. DOI: 10.1103/PhysRevLett.83.22.

[1966] M. Arndt, O. Nairz, J. Vos-Andreae, C. Keller, G. van der Zouw, and A. Zeilinger. “Wave–particle duality of C60 molecules”. In: *Nature* 401.6754 (1999). Published title uses an en dash and a subscript: Wave–particle duality of C_60 molecules., pp. 680–682. DOI: 10.1038/44348.

[1967] M. Born and E. Wolf. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. Also wanted eventually by the Hamilton–Jacobi chapter’s TODO-cite (eikonal limit). Cambridge University Press, 7th (expanded) edition, 1999. DOI: 10.1017/CBO9781139644181.

[1968] S. Coleman and S. L. Glashow. “High-energy tests of Lorentz invariance”. In: *Physical Review D* 59.11 (1999), p. 116008. DOI: 10.1103/PhysRevD.59.116008.

[1969] G. E. Crooks. “Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences”. In: *Physical Review E* 60.3 (1999), pp. 2721–2726. DOI: 10.1103/PhysRevE.60.2721.

[1970] F. Dalfovo, S. Giorgini, L. P. Pitaevskii, and S. Stringari. “Theory of Bose-Einstein condensation in trapped gases”. In: *Reviews of Modern Physics* 71.3 (1999), pp. 463–512. DOI: 10.1103/RevModPhys.71.463.

[1971] B. DeMarco and D. S. Jin. “Onset of Fermi Degeneracy in a Trapped Atomic Gas”. In: *Science* 285.5434 (1999), pp. 1703–1706. DOI: 10.1126/science.285.5434.1703.

[1972] D. Deutsch. “Quantum theory of probability and decisions”. In: *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 455.1988 (1999), pp. 3129–3137. DOI: 10.1098/rspa.1999.0443.

[1973] V. Fanti et al. “A new measurement of direct CP violation in two pion decays of the neutral kaon”. In: *Physics Letters B* 465.1-4 (1999), pp. 335–348. DOI: 10.1016/S0370-2693(99)01030-8.

[1974] W. Hampel et al. “GALLEX solar neutrino observations: results for GALLEX IV”. In: *Physics Letters B* 447.1-2 (1999), pp. 127–133. DOI: 10.1016/S0370-2693(98)01579-2.

[1975] M. Henny et al. “The Fermionic Hanbury Brown and Twiss Experiment”. In: *Science* 284.5412 (1999), pp. 296–298. DOI: 10.1126/science.284.5412.296.

[1976] J. D. Jackson. *Classical Electrodynamics*. 3rd ed. New York: Wiley, 1999.

[1977] M. Keilhacker et al. “High fusion performance from deuterium-tritium plasmas in JET”. In: *Nuclear Fusion* 39.2 (1999). Issue 2. The DTE1 campaign paper reporting the 16 MW peak fusion power quoted in the citing sentence. Verified against Crossref (title, vol. 39, pp. 209–234, 1999, first author M. Keilhacker). The author list is long and is given here from the published byline; if the bib writer prefers, truncate with ’and others’ – biblatex will render et al. Keyed by first listed author per the project convention., pp. 209–234. DOI: 10.1088/0029-5515/39/2/306.

[1978] W. Liu et al. “High Precision Measurements of the Ground State Hyperfine Structure Interval of Muonium and of the Muon Magnetic Moment”. In: *Physical Review Letters* 82.4 (1999), pp. 711–714. DOI: 10.1103/PhysRevLett.82.711.

[1979] M. R. Matthews, B. P. Anderson, P. C. Haljan, D. S. Hall, C. E. Wieman, and E. A. Cornell. “Vortices in a Bose-Einstein Condensate”. In: *Physical Review Letters* 83.13 (1999), pp. 2498–2501. DOI: 10.1103/PhysRevLett.83.2498.

[1980] Y. Nakamura, Y. A. Pashkin, and J. S. Tsai. “Coherent control of macroscopic quantum states in a single-Cooper-pair box”. In: *Nature* 398.6730 (1999), pp. 786–788. DOI: 10.1038/19718.

[1981] S. Perlmutter et al. “Measurements of Ω and Λ from 42 High-Redshift Supernovae”. In: *The Astrophysical Journal* 517.2 (1999). Supernova Cosmology Project; keyed by first listed author., pp. 565–586. DOI: 10.1086/307221.

[1982] C. Raman et al. “Evidence for a Critical Velocity in a Bose-Einstein Condensed Gas”. In: *Physical Review Letters* 83.13 (1999). First author Chandra Raman; unrelated to C. V. Raman of Raman:1928., pp. 2502–2505. DOI: 10.1103/PhysRevLett.83.2502.

[1983] L. Viola, E. Knill, and S. Lloyd. “Dynamical Decoupling of Open Quantum Systems”. In: *Physical Review Letters* 82.12 (1999), pp. 2417–2421. DOI: 10.1103/PhysRevLett.82.2417.

[1984] P. de Bernardis et al. “A flat Universe from high-resolution maps of the cosmic microwave background radiation”. In: *Nature* 404.6781 (2000). BOOMERanG first acoustic peak, pp. 955–959. DOI: 10.1038/35010035.

[1985] E. Burkel. “Phonon spectroscopy by inelastic x-ray scattering”. In: *Reports on Progress in Physics* 63.2 (2000), pp. 171–232. DOI: 10.1088/0034-4885/63/2/203.

[1986] V. Coffman, J. Kundu, and W. K. Wootters. “Distributed entanglement”. In: *Physical Review A* 61.5 (2000), p. 052306. DOI: 10.1103/PhysRevA.61.052306.

[1987] S. A. Diddams et al. “Direct Link between Microwave and Optical Frequencies with a 300 THz Femtosecond Laser Comb”. In: *Physical Review Letters* 84.22 (2000), pp. 5102–5105. DOI: 10.1103/PhysRevLett.84.5102.

[1988] H. Falcke, F. Melia, and E. Agol. “Viewing the Shadow of the Black Hole at the Galactic Center”. In: *The Astrophysical Journal* 528.1 (2000), pp. L13–L16. DOI: 10.1086/312423.

[1989] A. Gailitis et al. “Detection of a Flow Induced Magnetic Field Eigenmode in the Riga Dynamo Facility”. In: *Physical Review Letters* 84.19 (2000). Verified against Crossref. Both uses of the key in 09-plasmas-mhd.tex (the intro list ’a laboratory dynamo’ and the dynamo section ’a self-sustaining field eigenmode grew from a purely hydrodynamic flow’) describe THIS paper, the eigenmode-detection result – not the 2001 PRL 86, 3024 saturation paper, which would need its own key. Note the umlaut in Hänel., pp. 4365–4368. DOI: 10.1103/PhysRevLett.84.4365.

[1990] R. E. Grisenti, W. Schöllkopf, J. P. Toennies, G. C. Hegerfeldt, T. Köhler, and M. Stoll. “Determination of the Bond Length and Binding Energy of the Helium Dimer by Diffraction from a Transmission Grating”. In: *Physical Review Letters* 85.11 (2000), pp. 2284–2287. DOI: 10.1103/PhysRevLett.85.2284.

[1991] J. H. Gundlach and S. M. Merkowitz. “Measurement of Newton’s Constant Using a Torsion Balance with Angular Acceleration Feedback”. In: *Physical Review Letters* 85.14 (2000). The angular-acceleration-feedback method (University of Washington)., pp. 2869–2872. DOI: 10.1103/PhysRevLett.85.2869.

[1992] V. Halyo, P. Kim, E. R. Lee, I. T. Lee, D. Loomba, and M. L. Perl. “Search for Free Fractional Electric Charge Elementary Particles Using an Automated Millikan Oil Drop Technique”. In: *Physical Review Letters* 84.12 (2000), pp. 2576–2579. DOI: 10.1103/PhysRevLett.84.2576.

[1993] D. J. Jones et al. “Carrier-Envelope Phase Control of Femtosecond Mode-Locked Lasers and Direct Optical Frequency Synthesis”. In: *Science* 288.5466 (2000), pp. 635–639. DOI: 10.1126/science.288.5466.635.

[1994] Y.-H. Kim, R. Yu, S. P. Kulik, Y. Shih, and M. O. Scully. “Delayed “Choice” Quantum Eraser”. In: *Physical Review Letters* 84.1 (2000), pp. 1–5. DOI: 10.1103/PhysRevLett.84.1.

[1995] C. Kurtsiefer, S. Mayer, P. Zarda, and H. Weinfurter. “Stable Solid-State Source of Single Photons”. In: *Physical Review Letters* 85.2 (2000), pp. 290–293. DOI: 10.1103/PhysRevLett.85.290.

[1996] R. Loudon. *The Quantum Theory of Light*. 3rd edition. Oxford University Press, 2000.

[1997] B. Lounis and W. E. Moerner. “Single photons on demand from a single molecule at room temperature”. In: *Nature* 407.6803 (2000), pp. 491–493. DOI: 10.1038/35035032.

[1998] K. W. Madison, F. Chevy, W. Wohlleben, and J. Dalibard. “Vortex Formation in a Stirred Bose-Einstein Condensate”. In: *Physical Review Letters* 84.5 (2000), pp. 806–809. DOI: 10.1103/PhysRevLett.84.806.

[1999] V. Meyer et al. “Measurement of the1s−2sEnergy Interval in Muonium”. In: *Physical Review Letters* 84.6 (2000), pp. 1136–1139. DOI: 10.1103/PhysRevLett.84.1136.

[2000] P. Michler et al. “A Quantum Dot Single-Photon Turnstile Device”. In: *Science* 290.5500 (2000), pp. 2282–2285. DOI: 10.1126/science.290.5500.2282.

[2001] C. J. Myatt et al. “Decoherence of quantum superpositions through coupling to engineered reservoirs”. In: *Nature* 403.6767 (2000), pp. 269–273. DOI: 10.1038/35002001.

[2002] J.-W. Pan, D. Bouwmeester, M. Daniell, H. Weinfurter, and A. Zeilinger. “Experimental test of quantum nonlocality in three-photon Greenberger–Horne–Zeilinger entanglement”. In: *Nature* 403.6769 (2000), pp. 515–519. DOI: 10.1038/35000514.

[2003] Q. A. Turchette et al. “Decoherence and decay of motional quantum states of a trapped atom coupled to engineered reservoirs”. In: *Physical Review A* 62.5 (2000), p. 053807. DOI: 10.1103/PhysRevA.62.053807.

[2004] A. H. Zewail. “Femtochemistry: Atomic-Scale Dynamics of the Chemical Bond”. In: *The Journal of Physical Chemistry A* 104.24 (2000). Issue 24. The published version of the 1999 Nobel Lecture in Chemistry. Confirmed against the Crossref record for this DOI. A parallel Angewandte Chemie printing exists (39, 2586-2631, 2000) titled '...Using Ultrafast Lasers (Nobel Lecture)'; the J. Phys. Chem. A version is the one keyed here., pp. 5660–5694. DOI: 10.1021/jp001460h.

[2005] K. Abe et al. “Observation of LargeCPViolation in the NeutralBMeson System”. In: *Physical Review Letters* 87.9 (2001), p. 091802. DOI: 10.1103/PhysRevLett.87.091802.

[2006] J. R. Abo-Shaeer, C. Raman, J. M. Vogels, and W. Ketterle. “Observation of Vortex Lattices in Bose-Einstein Condensates”. In: *Science* 292.5516 (2001), pp. 476–479. DOI: 10.1126/science.1060182.

[2007] A. Aguilar et al. “Evidence for neutrino oscillations from the observation of$\bar{\nu}$eappearance in a$\bar{\nu}\mu$beam”. In: *Physical Review D* 64.11 (2001), p. 112007. DOI: 10.1103/PhysRevD.64.112007.

[2008] Q. R. Ahmad et al. “Measurement of the Rate ofνe+d→p+p+e−Interactions Produced byB8Solar Neutrinos at the Sudbury Neutrino Observatory”. In: *Physical Review Letters* 87.7 (2001), p. 071301. DOI: 10.1103/PhysRevLett.87.071301.

[2009] Z. I. Alferov. “Nobel Lecture: The double heterostructure concept and its applications in physics, electronics, and technology”. In: *Reviews of Modern Physics* 73.3 (2001), pp. 767–782. DOI: 10.1103/RevModPhys.73.767.

[2010] B. Aubert et al. “Observation ofCPViolation in theB0Meson System”. In: *Physical Review Letters* 87.9 (2001). B. Aubert (BaBar), distinct from J. J. Aubert 1974, p. 091801. DOI: 10.1103/PhysRevLett.87.091801.

[2011] S. Blundell. *Magnetism in Condensed Matter*. Oxford University Press, 2001.

[2012] M. C. Fischer, B. Gutiérrez-Medina, and M. G. Raizen. “Observation of the Quantum Zeno and Anti-Zeno Effects in an Unstable System”. In: *Physical Review Letters* 87.4 (2001), p. 040402. DOI: 10.1103/PhysRevLett.87.040402.

[2013] S. Hollands and R. M. Wald. “Local Wick Polynomials and Time Ordered Products¶of Quantum Fields in Curved Spacetime”. In: *Communications in Mathematical Physics* 223.2 (2001), pp. 289–326. DOI: 10.1007/s002200100540.

[2014] B. Jeckelmann and B. Jeanneret. “The quantum Hall effect as an electrical resistance standard”. In: *Reports on Progress in Physics* 64.12 (2001), pp. 1603–1655. DOI: 10.1088/0034-4885/64/12/201.

[2015] D. Kielpinski et al. “A Decoherence-Free Quantum Memory Using Trapped Ions”. In: *Science* 291.5506 (2001), pp. 1013–1015. DOI: 10.1126/science.1057357.

[2016] K. Kodama et al. “Observation of tau neutrino interactions”. In: *Physics Letters B* 504.3 (2001), pp. 218–224. DOI: 10.1016/S0370-2693(01)00307-0.

[2017] A. I. Lvovsky, H. Hansen, T. Aichele, O. Benson, J. Mlynek, and S. Schiller. “Quantum State Reconstruction of the Single-Photon Fock State”. In: *Physical Review Letters* 87.5 (2001), p. 050402. DOI: 10.1103/PhysRevLett.87.050402.

[2018] J. Nagamatsu, N. Nakagawa, T. Muranaka, Y. Zenitani, and J. Akimitsu. “Superconductivity at 39 K in magnesium diboride”. In: *Nature* 410.6824 (2001), pp. 63–64. DOI: 10.1038/35065039.

[2019] M. A. Rowe et al. “Experimental violation of a Bell’s inequality with efficient detection”. In: *Nature* 409.6822 (2001), pp. 791–794. DOI: 10.1038/35057215.

[2020] R. S. Towell, P. L. McGaughey, T. C. Awes, et al. “Improved measurement of the $\bar{d}/\bar{u}$ asymmetry in the nucleon sea”. In: *Physical Review D* 64.5 (2001). FNAL E866/NuSea. The Drell–Yan determination of the sea asymmetry that the Gottfried-sum deficit implies. Metadata confirmed against Crossref., p. 052002. DOI: 10.1103/PhysRevD.64.052002.

[2021] Q. R. Ahmad et al. “Direct Evidence for Neutrino Flavor Transformation from Neutral-Current Interactions in the Sudbury Neutrino Observatory”. In: *Physical Review Letters* 89.1 (2002), p. 011301. DOI: 10.1103/PhysRevLett.89.011301.

[2022] A. M. Beloborodov. “Gravitational Bending of Light Near Compact Objects”. In: *The Astrophysical Journal* 566 (2002). Metadata confirmed against Crossref., pp. L85–L88. DOI: 10.1086/339511.

[2023] H.-P. Breuer and F. Petruccione. *The Theory of Open Quantum Systems*. Oxford University Press, 2002.

[2024] J. Christensen-Dalsgaard. “Helioseismology”. In: *Reviews of Modern Physics* 74.4 (2002), pp. 1073–1129. DOI: 10.1103/RevModPhys.74.1073.

[2025] E. A. Cornell and C. E. Wieman. “Nobel Lecture: Bose-Einstein condensation in a dilute gas, the first 70 years and some recent experiments”. In: *Reviews of Modern Physics* 74.3 (2002), pp. 875–893. DOI: 10.1103/RevModPhys.74.875.

[2026] M. J. Duff, L. B. Okun, and G. Veneziano. “Trialogue on the number of fundamental constants”. In: *Journal of High Energy Physics* 2002.03 (2002). The three authors did not converge: Okun argues for three dimensionful constants, Veneziano for at most two (within superstring theory, which this treatise excludes by editorial rule 1), Duff for zero. The paper supports only the meta-claim that the count is convention-dependent, never a particular count, p. 023. DOI: 10.1088/1126-6708/2002/03/023.

[2027] H. Goldstein, C. P. Poole, and J. L. Safko. *Classical Mechanics*. 3rd ed. San Francisco: Addison-Wesley, 2002.

[2028] O. W. Greenberg. “CPTViolation Implies Violation of Lorentz Invariance”. In: *Physical Review Letters* 89.23 (2002), p. 231602. DOI: 10.1103/PhysRevLett.89.231602.

[2029] M. Greiner, O. Mandel, T. Esslinger, T. W. Hänsch, and I. Bloch. “Quantum phase transition from a superfluid to a Mott insulator in a gas of ultracold atoms”. In: *Nature* 415.6867 (2002), pp. 39–44. DOI: 10.1038/415039a.

[2030] W. Ketterle. “Nobel lecture: When atoms behave as waves: Bose-Einstein condensation and the atom laser”. In: *Reviews of Modern Physics* 74.4 (2002), pp. 1131–1151. DOI: 10.1103/RevModPhys.74.1131.

[2031] J. M. Kovac, E. M. Leitch, C. Pryke, J. E. Carlstrom, N. W. Halverson, and W. L. Holzapfel. “Detection of polarization in the cosmic microwave background using DASI”. In: *Nature* 420.6917 (2002), pp. 772–787. DOI: 10.1038/nature01269.

[2032] A. Kuhn, M. Hennrich, and G. Rempe. “Deterministic Single-Photon Source for Distributed Quantum Networking”. In: *Physical Review Letters* 89.6 (2002), p. 067901. DOI: 10.1103/PhysRevLett.89.067901.

[2033] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante. “Equilibrium Information from Nonequilibrium Measurements in an Experimental Test of Jarzynski’s Equality”. In: *Science* 296.5574 (2002), pp. 1832–1835. DOI: 10.1126/science.1071152.

[2034] V. V. Nesvizhevsky et al. “Quantum states of neutrons in the Earth’s gravitational field”. In: *Nature* 415.6869 (2002), pp. 297–299. DOI: 10.1038/415297a.

[2035] A. Pelissetto and E. Vicari. “Critical phenomena and renormalization-group theory”. In: *Physics Reports* 368.6 (2002), pp. 549–727. DOI: 10.1016/S0370-1573(02)00219-3.

[2036] A. L. Read. “Presentation of search results: theCLstechnique”. In: *Journal of Physics G: Nuclear and Particle Physics* 28.10 (2002), pp. 2693–2704. DOI: 10.1088/0954-3899/28/10/313.

[2037] R. H. Sanders and S. S. McGaugh. “Modified Newtonian Dynamics as an Alternative to Dark Matter”. In: *Annual Review of Astronomy and Astrophysics* 40.1 (2002), pp. 263–317. DOI: 10.1146/annurev.astro.40.060401.093923.

[2038] R. Schödel et al. “A star in a 15.2-year orbit around the supermassive black hole at the centre of the Milky Way”. In: *Nature* 419.6908 (2002), pp. 694–696. DOI: 10.1038/nature01121.

[2039] T. Udem, R. Holzwarth, and T. W. Hänsch. “Optical frequency metrology”. In: *Nature* 416.6877 (2002), pp. 233–237. DOI: 10.1038/416233a.

[2040] G. Vidal and R. F. Werner. “Computable measure of entanglement”. In: *Physical Review A* 65.3 (2002), p. 032314. DOI: 10.1103/PhysRevA.65.032314.

[2041] G. M. Wang, E. M. Sevick, E. Mittag, D. J. Searles, and D. J. Evans. "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales". In: *Physical Review Letters* 89.5 (2002), p. 050601. DOI: 10.1103/PhysRevLett.89.050601.

[2042] J. Zinn-Justin. *Quantum Field Theory and Critical Phenomena*. Fourth edition. Oxford University Press, 2002.

[2043] N. Ashby. "Relativity in the Global Positioning System". In: *Living Reviews in Relativity* 6.1 (2003). Article number 1; open access., p. 1. DOI: 10.12942/lrr-2003-1.

[2044] J. A. de Azcárraga, J. M. Izquierdo, M. Picón, and O. Varela. "Generating Lie and gauge free differential (super)algebras by expanding Maurer–Cartan forms and Chern–Simons supergravity". In: *Nuclear Physics B* 662.1-2 (2003), pp. 185–219. DOI: 10.1016/S0550-3213(03)00342-0.

[2045] R. Barate et al. "Search for the Standard Model Higgs boson at LEP". In: *Physics Letters B* 565 (2003). LEP Working Group for Higgs Boson Searches, ALEPH, DELPHI, L3 and OPAL Collaborations, pp. 61–75. DOI: 10.1016/S0370-2693(03)00614-2.

[2046] B. Bertotti, L. Iess, and P. Tortora. "A test of general relativity using radio links with the Cassini spacecraft". In: *Nature* 425.6956 (2003), pp. 374–376. DOI: 10.1038/nature01997.

[2047] M. Burgay et al. "An increased estimate of the merger rate of double neutron stars from observations of a highly relativistic system". In: *Nature* 426.6966 (2003). Double pulsar PSR J0737-3039 discovery, pp. 531–533. DOI: 10.1038/nature02124.

[2048] A. Damascelli, Z. Hussain, and Z.-X. Shen. "Angle-resolved photoemission studies of the cuprate superconductors". In: *Reviews of Modern Physics* 75.2 (2003), pp. 473–541. DOI: 10.1103/RevModPhys.75.473.

[2049] K. Eguchi et al. "First Results from KamLAND: Evidence for Reactor Antineutrino Disappearance". In: *Physical Review Letters* 90.2 (2003), p. 021802. DOI: 10.1103/PhysRevLett.90.021802.

[2050] B. Friedrich and D. Herschbach. "Stern and Gerlach: How a Bad Cigar Helped Reorient Atomic Physics". In: *Physics Today* 56.12 (2003), pp. 53–59. DOI: 10.1063/1.1650229.

[2051] M. Greiner, C. A. Regal, and D. S. Jin. "Emergence of a molecular Bose–Einstein condensate from a Fermi gas". In: *Nature* 426.6966 (2003), pp. 537–540. DOI: 10.1038/nature02199.

[2052] K. Hornberger, S. Uttenthaler, B. Brezger, L. Hackermüller, M. Arndt, and A. Zeilinger. "Collisional Decoherence Observed in Matter Wave Interferometry". In: *Physical Review Letters* 90.16 (2003), p. 160401. DOI: 10.1103/PhysRevLett.90.160401.

[2053] H. Katori, M. Takamoto, V. G. Pal'chikov, and V. D. Ovsiannikov. "Ultrastable Optical Clock with Neutral Atoms in an Engineered Light Shift Trap". In: *Physical Review Letters* 91.17 (2003), p. 173005. DOI: 10.1103/PhysRevLett.91.173005.

[2054] J. A. Lipa, J. A. Nissen, D. A. Stricker, D. R. Swanson, and T. C. P. Chui. "Specific heat of liquid helium in zero gravity very near the lambda point". In: *Physical Review B* 68.17 (2003), p. 174518. DOI: 10.1103/PhysRevB.68.174518.

[2055] W. H. Zurek. "Decoherence, einselection, and the quantum origins of the classical". In: *Reviews of Modern Physics* 75.3 (2003), pp. 715–775. DOI: 10.1103/RevModPhys.75.715.

[2056] S. N. Ahmed, A. E. Anthony, E. W. Beier, et al. “Measurement of the total active ^{8}B solar neutrino flux at the Sudbury Neutrino Observatory with enhanced neutral current sensitivity”. In: *Physical Review Letters* 92.18 (2004). SNO Collaboration, the salt phase. Keyed Ahmed (S. N. Ahmed) and distinct from Ahmad:2001 and Ahmad:2002, whose first author is Q. R. Ahmad – different people, not a transliteration variant. Metadata confirmed against Crossref., p. 181301. DOI: 10.1103/PhysRevLett.92.181301.

[2057] Y. Ashie, J. Hosaka, K. Ishihara, et al. “Evidence for an Oscillatory Signature in Atmospheric Neutrino Oscillations”. In: *Physical Review Letters* 93.10 (2004). Super-Kamiokande Collaboration. The L/E dip: the direct display of the oscillatory minimum, which a pure disappearance rate cannot establish. Metadata confirmed against Crossref., p. 101801. DOI: 10.1103/PhysRevLett.93.101801.

[2058] J. D. Bekenstein. “Relativistic gravitation theory for the modified Newtonian dynamics paradigm”. In: *Physical Review D* 70.8 (2004), p. 083509. DOI: 10.1103/PhysRevD.70.083509.

[2059] J. Clarke and A. I. (Braginski. *The SQUID Handbook. Volume I: Fundamentals and Technology of SQUIDs and SQUID Systems*. Edited volume; editor satisfies the author requirement. Wiley-VCH, 2004.

[2060] L. Hackermüller, K. Hornberger, B. Brezger, A. Zeilinger, and M. Arndt. “Decoherence of matter waves by thermal emission of radiation”. In: *Nature* 427.6976 (2004). Key transliterates the umlaut (Hackermueller) to stay ASCII., pp. 711–714. DOI: 10.1038/nature02276.

[2061] O. Madelung. *Semiconductors: Data Handbook*. 3rd edition. Springer, 2004.

[2062] M. Markevitch et al. “Direct Constraints on the Dark Matter Self-Interaction Cross Section from the Merging Galaxy Cluster 1E 0657−56”. In: *The Astrophysical Journal* 606.2 (2004), pp. 819–824. DOI: 10.1086/383178.

[2063] K. S. Novoselov et al. “Electric Field Effect in Atomically Thin Carbon Films”. In: *Science* 306.5696 (2004), pp. 666–669. DOI: 10.1126/science.1102896.

[2064] J. L. O’Brien et al. “Quantum Process Tomography of a Controlled-NOT Gate”. In: *Physical Review Letters* 93.8 (2004), p. 080502. DOI: 10.1103/PhysRevLett.93.080502.

[2065] C. A. Regal, M. Greiner, and D. S. Jin. “Observation of Resonance Condensation of Fermionic Atom Pairs”. In: *Physical Review Letters* 92.4 (2004), p. 040403. DOI: 10.1103/PhysRevLett.92.040403.

[2066] C. Rovelli. *Quantum Gravity*. Cited solely to name an excluded programme and record its zero-evidence status, per editorial scope rule 1. Cambridge University Press, 2004.

[2067] J. G. Williams, S. G. Turyshev, and D. H. Boggs. “Progress in Lunar Laser Ranging Tests of Relativistic Gravity”. In: *Physical Review Letters* 93.26 (2004), p. 261101. DOI: 10.1103/PhysRevLett.93.261101.

[2068] J. Adams et al. “Experimental and theoretical challenges in the search for the quark–gluon plasma: The STAR Collaboration’s critical assessment of the evidence from RHIC collisions”. In: *Nuclear Physics A* 757.1-2 (2005), pp. 102–183. DOI: 10.1016/j.nuclphysa.2005.03.085.

[2069] P. L. Anthony et al. “Precision Measurement of the Weak Mixing Angle in Møller Scattering”. In: *Physical Review Letters* 95.8 (2005), p. 081601. DOI: 10.1103/PhysRevLett.95.081601.

[2070] T. Antoni et al. “KASCADE measurements of energy spectra for elemental groups of cosmic rays: Results and open problems”. In: *Astroparticle Physics* 24.1-2 (2005), pp. 1–25. DOI: 10.1016/j.astropartphys.2005.04.001.

[2071] T. Araki, K. Eguchi, S. Enomoto, et al. “Measurement of Neutrino Oscillation with KamLAND: Evidence of Spectral Distortion”. In: *Physical Review Letters* 94.8 (2005). KamLAND Collaboration. The distortion of the reactor antineutrino spectrum, i.e. the oscillatory L/E dependence itself rather than a rate deficit. Metadata confirmed against Crossref., p. 081801. DOI: 10.1103/PhysRevLett.94.081801.

[2072] J. N. Bahcall, A. M. Serenelli, and S. Basu. “New Solar Opacities, Abundances, Helioseismology, and Neutrino Fluxes”. In: *The Astrophysical Journal* 621.1 (2005), pp. L85–L88. DOI: 10.1086/428929.

[2073] G. Bertone, D. Hooper, and J. Silk. “Particle dark matter: evidence, candidates and constraints”. In: *Physics Reports* 405.5-6 (2005), pp. 279–390. DOI: 10.1016/j.physrep.2004.08.031.

[2074] J. Charles et al. “CP violation and the CKM matrix: assessing the impact of the asymmetric B factories”. In: *The European Physical Journal C* 41.1 (2005), pp. 1–131. DOI: 10.1140/epjc/s2005-02169-1.

[2075] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, and C. Bustamante. “Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies”. In: *Nature* 437.7056 (2005), pp. 231–234. DOI: 10.1038/nature04061.

[2076] D. J. Eisenstein et al. “Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies”. In: *The Astrophysical Journal* 633.2 (2005). Collaboration paper keyed by first listed author., pp. 560–574. DOI: 10.1086/466512.

[2077] C. L. Kane and E. J. Mele. “Quantum Spin Hall Effect in Graphene”. In: *Physical Review Letters* 95.22 (2005), p. 226801. DOI: 10.1103/PhysRevLett.95.226801.

[2078] C. Kittel. *Introduction to Solid State Physics*. 8th edition. John Wiley and Sons, 2005.

[2079] J. Knödlseder et al. “The all-sky distribution of 511 keV electron-positron annihilation emission”. In: *Astronomy & Astrophysics* 441.2 (2005). Key transliterates the umlaut as ’oe’ to keep the citekey ASCII., pp. 513–532. DOI: 10.1051/0004-6361:20042063.

[2080] J. Matthews. “A Heitler model of extensive air showers”. In: *Astroparticle Physics* 22.5-6 (2005), pp. 387–397. DOI: 10.1016/j.astropartphys.2004.09.003.

[2081] K. S. Novoselov et al. “Two-dimensional gas of massless Dirac fermions in graphene”. In: *Nature* 438.7065 (2005), pp. 197–200. DOI: 10.1038/nature04233.

[2082] F. Pretorius. “Evolution of Binary Black-Hole Spacetimes”. In: *Physical Review Letters* 95 (2005), p. 121101. DOI: 10.1103/PhysRevLett.95.121101.

[2083] V. Springel et al. “Simulations of the formation, evolution and clustering of galaxies and quasars”. In: *Nature* 435.7042 (2005), pp. 629–636. DOI: 10.1038/nature03597.

[2084] M. Takamoto, F.-L. Hong, R. Higashi, and H. Katori. “An optical lattice clock”. In: *Nature* 435.7040 (2005), pp. 321–324. DOI: 10.1038/nature03541.

[2085] Y. Zhang, Y.-W. Tan, H. L. Stormer, and P. Kim. “Experimental observation of the quantum Hall effect and Berry’s phase in graphene”. In: *Nature* 438.7065 (2005), pp. 201–204. DOI: 10.1038/nature04235.

[2086] W. H. Zurek. “Probabilities from entanglement, Born’s rulepk=| ψk|2from envariance”. In: *Physical Review A* 71.5 (2005). The published title contains inline math ($p_k = |\psi_k|^2$); transcribed here in ASCII. The DOI is authoritative., p. 052105. DOI: 10.1103/PhysRevA.71.052105.

[2087] M. W. Zwierlein, J. R. Abo-Shaeer, A. Schirotzek, C. H. Schunck, and W. Ketterle. “Vortices and superfluidity in a strongly interacting Fermi gas”. In: *Nature* 435.7045 (2005), pp. 1047–1051. DOI: 10.1038/nature03858.

[2088] A. Achterberg et al. “First year performance of the IceCube neutrino telescope”. In: *Astroparticle Physics* 26.3 (2006), pp. 155–173. DOI: 10.1016/j.astropartphys.2006.06.007.

[2089] M. H. Ahn et al. “Measurement of neutrino oscillation by the K2K experiment”. In: *Physical Review D* 74.7 (2006), p. 072003. DOI: 10.1103/PhysRevD.74.072003.

[2090] C. A. Baker et al. “Improved Experimental Limit on the Electric Dipole Moment of the Neutron”. In: *Physical Review Letters* 97.13 (2006), p. 131801. DOI: 10.1103/PhysRevLett.97.131801.

[2091] G. W. Bennett et al. “Final report of the E821 muon anomalous magnetic moment measurement at BNL”. In: *Physical Review D* 73.7 (2006), p. 072003. DOI: 10.1103/PhysRevD.73.072003.

[2092] B. A. Bernevig, T. L. Hughes, and S.-C. Zhang. “Quantum Spin Hall Effect and Topological Phase Transition in HgTe Quantum Wells”. In: *Science* 314.5806 (2006), pp. 1757–1761. DOI: 10.1126/science.1133734.

[2093] D. Clowe et al. “A Direct Empirical Proof of the Existence of Dark Matter”. In: *The Astrophysical Journal* 648.2 (2006), pp. L109–L113. DOI: 10.1086/508162.

[2094] C. L. Fefferman. “Existence and Smoothness of the Navier-Stokes Equation”. In: *The Millennium Prize Problems*. Chapter in the volume edited by J. Carlson, A. Jaffe and A. Wiles. The official Millennium problem statement; the chapter records the problem as open. 2006, pp. 57–67.

[2095] J. W. T. Hessels, S. M. Ransom, I. H. Stairs, P. C. C. Freire, V. M. Kaspi, and F. Camilo. “A Radio Pulsar Spinning at 716 Hz”. In: *Science* 311 (2006). Metadata confirmed against Crossref., pp. 1901–1904. DOI: 10.1126/science.1123430.

[2096] F. Izaurieta, E. Rodríguez, and P. Salgado. “Expanding Lie (super)algebras through Abelian semigroups”. In: *Journal of Mathematical Physics* 47.12 (2006), p. 123512. DOI: 10.1063/1.2390659.

[2097] A. Jaffe and E. Witten. *Quantum Yang-Mills Theory*. Official Clay Millennium Prize problem statement; edited by J. Carlson, A. Jaffe and A. Wiles. 2006.

[2098] M. Kramer et al. “Tests of General Relativity from Timing the Double Pulsar”. In: *Science* 314.5796 (2006), pp. 97–102. DOI: 10.1126/science.1132305.

[2099] N. Li and B.-Q. Ma. “Energy scale independence of Koide’s relation for quark and lepton masses”. In: *Physical Review D* 73.1 (2006). Shows the running-mass Koide ratio is nearly scale-independent from 1 GeV to 2×10^{16} GeV — but frozen at about 0.19% above 2/3, not at 2/3, p. 013009. DOI: 10.1103/PhysRevD.73.013009.

[2100] D. G. Michael et al. “Observation of Muon Neutrino Disappearance with the MINOS Detectors in the NuMI Neutrino Beam”. In: *Physical Review Letters* 97.19 (2006), p. 191801. DOI: 10.1103/PhysRevLett.97.191801.

[2101] B. Odom, D. Hanneke, B. D’Urso, and G. Gabrielse. “New Measurement of the Electron Magnetic Moment Using a One-Electron Quantum Cyclotron”. In: *Physical Review Letters* 97.3 (2006), p. 030801. DOI: 10.1103/PhysRevLett.97.030801.

[2102] G. Rapior, K. Sengstock, and V. Baev. “New features of the Franck-Hertz experiment”. In: *American Journal of Physics* 74.5 (2006). Quantifies the drift of the apparent spacing with maximum number and the multi-level systematics of the neon variant., pp. 423–428. DOI: 10.1119/1.2174033.

[2103] R. A. Remillard and J. E. McClintock. “X-Ray Properties of Black-Hole Binaries”. In: *Annual Review of Astronomy and Astrophysics* 44.1 (2006), pp. 49–92. DOI: 10.1146/annurev.astro.44.051905.092532.

[2104] S. Schael et al. “Precision electroweak measurements on the Z resonance”. In: *Physics Reports* 427.5-6 (2006). ALEPH, DELPHI, L3, OPAL and SLD Collaborations, LEP Electroweak Working Group, SLD Electroweak and Heavy Flavour Groups, pp. 257–454. DOI: 10.1016/j.physrep.2005.12.006.

[2105] A. Vikhlinin et al. “ChandraSample of Nearby Relaxed Galaxy Clusters: Mass, Gas Fraction, and Mass-Temperature Relation”. In: *The Astrophysical Journal* 640.2 (2006), pp. 691–709. DOI: 10.1086/500288.

[2106] C. Giunti and C. W. Kim. *Fundamentals of Neutrino Physics and Astrophysics*. Oxford University Press, 2007.

[2107] C. Guerlin et al. “Progressive field-state collapse and quantum non-demolition photon counting”. In: *Nature* 448.7156 (2007), pp. 889–893. DOI: 10.1038/nature06057.

[2108] V. Jacques et al. “Experimental Realization of Wheeler’s Delayed-Choice Gedanken Experiment”. In: *Science* 315.5814 (2007), pp. 966–968. DOI: 10.1126/science.1136303.

[2109] T. Jeltes et al. “Comparison of the Hanbury Brown–Twiss effect for bosons and fermions”. In: *Nature* 445.7126 (2007), pp. 402–405. DOI: 10.1038/nature05513.

[2110] M. König et al. “Quantum Spin Hall Insulator State in HgTe Quantum Wells”. In: *Science* 318 (2007), pp. 766–770.

[2111] E. R. Scerri. *The Periodic Table: Its Story and Its Significance*. Oxford University Press, 2007.

[2112] M. Schlosshauer. *Decoherence and the Quantum-to-Classical Transition*. Berlin: Springer, 2007.

[2113] S. M. Sze and K. K. Ng. *Physics of Semiconductor Devices*. 3rd edition. Wiley, 2007.

[2114] P. Tisserand et al. “Limits on the Macho content of the Galactic Halo from the EROS-2 Survey of the Magellanic Clouds”. In: *Astronomy & Astrophysics* 469.2 (2007), pp. 387–404. DOI: 10.1051/0004-6361:20066017.

[2115] R. U. Abbasi et al. “First Observation of the Greisen-Zatsepin-Kuzmin Suppression”. In: *Physical Review Letters* 100.10 (2008), p. 101101. DOI: 10.1103/PhysRevLett.100.101101.

[2116] J. Abraham et al. “Observation of the Suppression of the Flux of Cosmic Rays above4×1019 eV”. In: *Physical Review Letters* 101.6 (2008), p. 061101. DOI: 10.1103/PhysRevLett.101.061101.

[2117] R. Bernabei et al. “First results from DAMA/LIBRA and the combined results with DAMA/NaI”. In: *The European Physical Journal C* 56.3 (2008), pp. 333–355. DOI: 10.1140/epjc/s10052-008-0662-y.

[2118] I. Bloch, J. Dalibard, and W. Zwerger. “Many-body physics with ultracold gases”. In: *Reviews of Modern Physics* 80.3 (2008), pp. 885–964. DOI: 10.1103/RevModPhys.80.885.

[2119] T. A. Collaboration et al. “The ATLAS Experiment at the CERN Large Hadron Collider”. In: *Journal of Instrumentation* 3.08 (2008), S08003–S08003. DOI: 10.1088/1748-0221/3/08/S08003.

[2120] T. C. Collaboration et al. “The CMS experiment at the CERN LHC”. In: *Journal of Instrumentation* 3.08 (2008), S08004–S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[2121] S. Deléglise et al. “Reconstruction of non-classical cavity field states with snapshots of their decoherence”. In: *Nature* 455.7212 (2008), pp. 510–514. DOI: 10.1038/nature07288.

[2122] S. Durr et al. “Ab Initio Determination of Light Hadron Masses”. In: *Science* 322 (2008). First author’s name is written Durr (with umlaut) in the original, pp. 1224–1227.

[2123] L. Evans and P. Bryant. “LHC Machine”. In: *Journal of Instrumentation* 3.08 (2008), S08001–S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[2124] É. É. Flanagan and T. Hinderer. “Constraining neutron-star tidal Love numbers with gravitational-wave detectors”. In: *Physical Review D* 77 (2008). Metadata confirmed against Crossref., p. 021502. DOI: 10.1103/PhysRevD.77.021502.

[2125] A. M. Ghez et al. “Measuring Distance and Properties of the Milky Way’s Central Supermassive Black Hole with Stellar Orbits”. In: *The Astrophysical Journal* 689.2 (2008), pp. 1044–1062. DOI: 10.1086/592738.

[2126] D. Hanneke, S. Fogwell, and G. Gabrielse. “New Measurement of the Electron Magnetic Moment and the Fine Structure Constant”. In: *Physical Review Letters* 100.12 (2008), p. 120801. DOI: 10.1103/PhysRevLett.100.120801.

[2127] T. Hinderer. “Tidal Love Numbers of Neutron Stars”. In: *The Astrophysical Journal* 677 (2008). Erratum: Astrophys. J. 697 (2009) 964 corrects some tabulated Love numbers; the range quoted in the treatise survives. Metadata confirmed against Crossref., pp. 1216–1220. DOI: 10.1086/533487.

[2128] D. Hsieh et al. “A topological Dirac insulator in a quantum spin Hall phase”. In: *Nature* 452.7190 (2008), pp. 970–974. DOI: 10.1038/nature06843.

[2129] Joint Committee for Guides in Metrology. *Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement*. JCGM 100:2008 (GUM 1995 with minor corrections). 2008.

[2130] Y. Kamihara, T. Watanabe, M. Hirano, and H. Hosono. “Iron-Based Layered Superconductor La[O1-xFx]FeAs (x = 0.05−0.12) with Tc = 26 K”. In: *Journal of the American Chemical Society* 130.11 (2008), pp. 3296–3297. DOI: 10.1021/ja800073m.

[2131] A. E. Lita, A. J. Miller, and S. W. Nam. “Counting near-infrared single-photons with 95% efficiency”. In: *Optics Express* 16.5 (2008), p. 3032. DOI: 10.1364/OE.16.003032.

[2132] M. Maggiore. *Gravitational Waves, Volume 1: Theory and Experiments*. Oxford: Oxford University Press, 2008.

[2133] C. J. Pethick and H. Smith. *Bose-Einstein Condensation in Dilute Gases*. 2nd edition. Cambridge University Press, 2008.

[2134] T. Rosenband et al. “Frequency Ratio of Al^+ and Hg^+ Single-Ion Optical Clocks; Metrology at the 17th Decimal Place”. In: *Science* 319.5871 (2008), pp. 1808–1812. DOI: 10.1126/science.1154622.

[2135] S. Schlamminger, K.-Y. Choi, T. A. Wagner, J. H. Gundlach, and E. G. Adelberger. "Test of the Equivalence Principle Using a Rotating Torsion Balance". In: *Physical Review Letters* 100.4 (2008), p. 041101. DOI: 10.1103/PhysRevLett.100.041101.

[2136] G. J. Snyder and E. S. Toberer. "Complex thermoelectric materials". In: *Nature Materials* 7.2 (2008), pp. 105–114. DOI: 10.1038/nmat2090.

[2137] A. A. Abdo et al. "A limit on the variation of the speed of light arising from quantum gravity effects". In: *Nature* 462.7271 (2009), pp. 331–334. DOI: 10.1038/nature08574.

[2138] J. N. Abdurashitov et al. "Measurement of the solar neutrino capture rate with gallium metal. III. Results for the 2002–2007 data-taking period". In: *Physical Review C* 80.1 (2009), p. 015807. DOI: 10.1103/PhysRevC.80.015807.

[2139] O. Adriani et al. "An anomalous positron abundance in cosmic rays with energies 1.5–100 GeV". In: *Nature* 458.7238 (2009), pp. 607–609. DOI: 10.1038/nature07942.

[2140] M. Asplund, N. Grevesse, A. J. Sauval, and P. Scott. "The Chemical Composition of the Sun". In: *Annual Review of Astronomy and Astrophysics* 47.1 (2009), pp. 481–522. DOI: 10.1146/annurev.astro.46.060407.145222.

[2141] W. B. Atwood et al. "THE LARGE AREA TELESCOPE ON THEFERMI GAMMA-RAY SPACE TELESCOPEMISSION". In: *The Astrophysical Journal* 697.2 (2009), pp. 1071–1102. DOI: 10.1088/0004-637X/697/2/1071.

[2142] S. Bethke. "The 2009 world average of α s". In: *The European Physical Journal C* 64.4 (2009), pp. 689–703. DOI: 10.1140/epjc/s10052-009-1173-1.

[2143] H.-P. Breuer, E.-M. Laine, and J. Piilo. "Measure for the Degree of Non-Markovian Behavior of Quantum Processes in Open Systems". In: *Physical Review Letters* 103.21 (2009), p. 210401. DOI: 10.1103/PhysRevLett.103.210401.

[2144] D. J. Fixsen. "THE TEMPERATURE OF THE COSMIC MICROWAVE BACKGROUND". In: *The Astrophysical Journal* 707.2 (2009), pp. 916–920. DOI: 10.1088/0004-637X/707/2/916.

[2145] E. Fomalont, S. Kopeikin, G. Lanyi, and J. Benson. "PROGRESS IN MEASUREMENTS OF THE GRAVITATIONAL BENDING OF RADIO WAVES USING THE VLBA". In: *The Astrophysical Journal* 699.2 (2009). Issue 2. Crossref confirmed title, authors, volume, issue, pages and year against the DOI; Crossref stores the title in all caps ("PROGRESS IN MEASUREMENTS OF THE GRAVITATIONAL BENDING OF RADIO WAVES USING THE VLBA") as an artefact of the IOP deposit – the published title is in title case as given here. The 2005 VLBA campaign giving gamma-1 = (-2.0 +/- 3.0) x 10^-4. Cited with Lebach:1995., pp. 1395–1402. DOI: 10.1088/0004-637X/699/2/1395.

[2146] S. Gillessen et al. "MONITORING STELLAR ORBITS AROUND THE MASSIVE BLACK HOLE IN THE GALACTIC CENTER". In: *The Astrophysical Journal* 692.2 (2009), pp. 1075–1109. DOI: 10.1088/0004-637X/692/2/1075.

[2147] S. Herrmann, A. Senger, K. Möhle, M. Nagel, E. V. Kovalchuk, and A. Peters. "Rotating optical cavity experiment testing Lorentz invariance at the 10^{-17} level". In: *Physical Review D* 80.10 (2009), p. 105011. DOI: 10.1103/PhysRevD.80.105011.

[2148] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki. "Quantum entanglement". In: *Reviews of Modern Physics* 81.2 (2009), pp. 865–942. DOI: 10.1103/RevModPhys.81.865.

[2149] S. G. Simpson. *Subsystems of Second Order Arithmetic*. 2nd ed. Perspectives in Logic. The standard reference for reverse mathematics. First edition: Springer (1999), doi 10.1007/978-3-642-59971-2. Cambridge: Cambridge University Press, 2009. DOI: 10.1017/CBO9780511581007.

[2150] Y. Sumino. "Family gauge symmetry as an origin of Koide's mass formula and charged lepton spectrum". In: *Journal of High Energy Physics* 2009.05 (2009). Identifies the QED self-energy logarithm as the obstruction to the relation holding for running masses, and proposes a family gauge symmetry to cancel it. The proposed gauge bosons have not been observed; the construction is reported here as the reason the relation is contested, not as evidence for it, p. 075. DOI: 10.1088/1126-6708/2009/05/075.

[2151] W. H. Zurek. "Quantum Darwinism". In: *Nature Physics* 5.3 (2009), pp. 181–188. DOI: 10.1038/nphys1202.

[2152] F. D. Aaron et al. "Combined measurement and QCD analysis of the inclusive e ± p scattering cross sections at HERA". In: *Journal of High Energy Physics* 2010.1 (2010), p. 109. DOI: 10.1007/JHEP01(2010)109.

[2153] C. Chin, R. Grimm, P. Julienne, and E. Tiesinga. "Feshbach resonances in ultracold gases". In: *Reviews of Modern Physics* 82.2 (2010), pp. 1225–1286. DOI: 10.1103/RevModPhys.82.1225.

[2154] C. W. Chou, D. B. Hume, T. Rosenband, and D. J. Wineland. "Optical Clocks and Relativity". In: *Science* 329.5999 (2010), pp. 1630–1633. DOI: 10.1126/science.1192720.

[2155] P. B. Demorest, T. Pennucci, S. M. Ransom, M. S. E. Roberts, and J. W. T. Hessels. "A two-solar-mass neutron star measured using Shapiro delay". In: *Nature* 467.7319 (2010), pp. 1081–1083. DOI: 10.1038/nature09466.

[2156] R. Genzel, F. Eisenhauer, and S. Gillessen. "The Galactic Center massive black hole and nuclear star cluster". In: *Reviews of Modern Physics* 82.4 (2010), pp. 3121–3195. DOI: 10.1103/RevModPhys.82.3121.

[2157] A. S. Goldhaber and M. M. Nieto. "Photon and graviton mass limits". In: *Reviews of Modern Physics* 82.1 (2010). Issue 1. Verified against Crossref: title, volume, pages and 2010 issue year all match. This is the standard review the treatise means by 'the far sharper bounds from planetary and galactic magnetic fields reviewed in'., pp. 939–979. DOI: 10.1103/RevModPhys.82.939.

[2158] E. Gross and O. Vitells. "Trial factors for the look elsewhere effect in high energy physics". In: *The European Physical Journal C* 70.1 (2010). Crossref-verified. Davies' upcrossing bound turned into the practical trials factor used in searches: the global p-value follows from the expected number of upcrossings of the test statistic at a low reference level, estimated from a modest number of toy experiments. DISTINGUISH from Gross:1961 (E. P. Gross, quantized vortices) and Gross:1972, Gross:1973 (D. J. Gross): three different physicists named Gross in this bibliography., pp. 525–530. DOI: 10.1140/epjc/s10052-010-1470-8.

[2159] B. Grzadkowski, M. Iskrzyński, M. Misiak, and J. Rosiek. "Dimension-six terms in the Standard Model Lagrangian". In: *Journal of High Energy Physics* 2010.10 (2010), p. 085. DOI: 10.1007/JHEP10(2010)085.

[2160] T. Li, S. Kheifets, D. Medellin, and M. G. Raizen. "Measurement of the Instantaneous Velocity of a Brownian Particle". In: *Science* 328.5986 (2010). Tongcang Li; distinct from the Nan Li of Li:2006, pp. 1673–1675. DOI: 10.1126/science.1189403.

[2161] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge: Cambridge University Press, 2010. DOI: 10.1017/CBO9780511976667.

[2162] Y. T. Oganessian et al. "Synthesis of a New Element with Atomic NumberZ=117". In: *Physical Review Letters* 104.14 (2010), p. 142502. DOI: 10.1103/PhysRevLett.104.142502.

[2163] S. Pironio et al. "Random numbers certified by Bell's theorem". In: *Nature* 464.7291 (2010), pp. 1021–1024. DOI: 10.1038/nature09008.

[2164] R. Pohl et al. "The size of the proton". In: *Nature* 466.7303 (2010), pp. 213–216. DOI: 10.1038/nature09250.

[2165] Á. Rivas, S. F. Huelga, and M. B. Plenio. "Entanglement and Non-Markovianity of Quantum Evolutions". In: *Physical Review Letters* 105.5 (2010), p. 050403. DOI: 10.1103/PhysRevLett.105.050403.

[2166] T. Scheidl et al. "Violation of local realism with freedom of choice". In: *Proceedings of the National Academy of Sciences* 107.46 (2010), pp. 19708–19713. DOI: 10.1073/pnas.1002780107.

[2167] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano. "Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality". In: *Nature Physics* 6.12 (2010), pp. 988–992. DOI: 10.1038/nphys1821.

[2168] K. Abe et al. "Indication of Electron Neutrino Appearance from an Accelerator-Produced Off-Axis Muon Neutrino Beam". In: *Physical Review Letters* 107.4 (2011), p. 041801. DOI: 10.1103/PhysRevLett.107.041801.

[2169] K. Avila, D. Moxey, A. de Lozar, M. Avila, D. Barkley, and B. Hof. "The Onset of Turbulence in Pipe Flow". In: *Science* 333.6039 (2011). Issue 6039, 8 July 2011. First author K. Avila is distinct from the fourth author M. Avila – key by the first listed author per house convention. Establishes the two-state picture quantitatively: turbulence is transient below and sustained above a critical point Re = 2040 +/- 10, set by the balance of puff decay against puff splitting., pp. 192–196. DOI: 10.1126/science.1203223.

[2170] R. Bayes et al. "Experimental Constraints on Left-Right Symmetric Models from Muon Decay". In: *Physical Review Letters* 106.4 (2011), p. 041804. DOI: 10.1103/PhysRevLett.106.041804.

[2171] M. Bazzi et al. "A new measurement of kaonic hydrogen X-rays". In: *Physics Letters B* 704.3 (2011). SIDDHARTA collaboration; keyed by first listed author., pp. 113–117. DOI: 10.1016/j.physletb.2011.09.011.

[2172] G. Cowan, K. Cranmer, E. Gross, and O. Vitells. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71.2 (2011). Crossref-verified. The reference that fixes the conventions a discovery claim actually uses: the profile-likelihood ratio as test statistic, its asymptotic distribution including the boundary case of Chernoff:1954, the one-sided convention Z = sqrt(q) converting it into a significance in sigma, and the Asimov data set for median sensitivity without toys. An erratum appears at Eur. Phys. J. C 73 (2013) 2501., p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0.

[2173] C. W. F. Everitt et al. "Gravity Probe B: Final Results of a Space Experiment to Test General Relativity". In: *Physical Review Letters* 106.22 (2011), p. 221101. DOI: 10.1103/PhysRevLett.106.221101.

[2174] S. Kocsis et al. “Observing the Average Trajectories of Single Photons in a Two-Slit Interferometer”. In: *Science* 332.6034 (2011), pp. 1170–1173. DOI: 10.1126/science.1202218.

[2175] Z. Komargodski and A. Schwimmer. “On renormalization group flows in four dimensions”. In: *Journal of High Energy Physics* 2011.12 (2011), p. 099. DOI: 10.1007/JHEP12(2011)099.

[2176] V. A. Kostelecký and N. Russell. “Data tables for Lorentz andCPTviolation”. In: *Reviews of Modern Physics* 83.1 (2011), pp. 11–31. DOI: 10.1103/RevModPhys.83.11.

[2177] R. Machleidt and D. Entem. “Chiral effective field theory and nuclear forces”. In: *Physics Reports* 503.1 (2011), pp. 1–75. DOI: 10.1016/j.physrep.2011.02.001.

[2178] J. A. Orosz et al. “The Mass of the Black Hole in Cygnus X-1”. In: *The Astrophysical Journal* 742 (2011). Metadata confirmed against Crossref., p. 84. DOI: 10.1088/0004-637X/742/2/84.

[2179] C. G. Parthey et al. “Improved Measurement of the Hydrogen1S–2STransition Frequency”. In: *Physical Review Letters* 107.20 (2011), p. 203001. DOI: 10.1103/PhysRevLett.107.203001.

[2180] R. K. Pathria and P. D. Beale. *Statistical Mechanics*. 3rd edition. Amsterdam: Academic Press, 2011.

[2181] C. de Rham, G. Gabadadze, and A. J. Tolley. “Resummation of Massive Gravity”. In: *Physical Review Letters* 106.23 (2011), p. 231101. DOI: 10.1103/PhysRevLett.106.231101.

[2182] R. van Rooij et al. “Frequency Metrology in Quantum Degenerate Helium: Direct Measurement of the 2 3 S 1 $\rightarrow$ 2 1 S 0 Transition”. In: *Science* 333.6039 (2011), pp. 196–198. DOI: 10.1126/science.1205163.

[2183] J.-P. Uzan. “Varying Constants, Gravitation and Cosmology”. In: *Living Reviews in Relativity* 14 (2011). Living Reviews are living documents; this is the 2011 version. Distinct from the same author’s Rev. Mod. Phys. 75 (2003) 403, p. 2. DOI: 10.12942/lrr-2011-2.

[2184] J. K. Webb, J. A. King, M. T. Murphy, V. V. Flambaum, R. F. Carswell, and M. B. Bainbridge. “Indications of a Spatial Variation of the Fine Structure Constant”. In: *Physical Review Letters* 107.19 (2011). CONTESTED. The paper’s own systematic analysis reduces the dipole significance from 4.2σ to as little as 2.2σ; see Whitmore:2015. The widely quoted amplitude 0.97×10^{-5} belongs to the companion paper King et al., MNRAS 422 (2012) 3370, not to this Letter, whose figure gives $(1.02 \pm 0.21) \times 10^{-5}$, p. 191101. DOI: 10.1103/PhysRevLett.107.191101.

[2185] S. Weinfurtner, E. W. Tedford, M. C. J. Penrice, W. G. Unruh, and G. A. Lawrence. “Measurement of Stimulated Hawking Emission in an Analogue System”. In: *Physical Review Letters* 106.2 (2011). Analogue (surface-wave) system; flagged as such in the chapter., p. 021302. DOI: 10.1103/PhysRevLett.106.021302.

[2186] G. Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Physics Letters B* 716.1 (2012). ATLAS collaboration; keyed by first listed author. Likely also introduced by the Higgs-discovery chapter., pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020.

[2187] T. Aaltonen et al. “Evidence for a Particle Produced in Association with Weak Bosons and Decaying to a Bottom-Antibottom Quark Pair in Higgs Boson Searches at the Tevatron”. In: *Physical Review Letters* 109.7 (2012), p. 071804. DOI: 10.1103/PhysRevLett.109.071804.

[2188] J. K. Ahn et al. “Observation of Reactor Electron Antineutrinos Disappearance in the RENO Experiment”. In: *Physical Review Letters* 108.19 (2012), p. 191802. DOI: 10.1103/PhysRevLett.108.191802.

[2189] F. P. An et al. “Observation of Electron-Antineutrino Disappearance at Daya Bay”. In: *Physical Review Letters* 108.17 (2012), p. 171803. DOI: 10.1103/PhysRevLett.108.171803.

[2190] T. Aoyama, M. Hayakawa, T. Kinoshita, and M. Nio. “Tenth-Order QED Contribution to the Electrong−2and an Improved Value of the Fine Structure Constant”. In: *Physical Review Letters* 109.11 (2012), p. 111807. DOI: 10.1103/PhysRevLett.109.111807.

[2191] A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz. “Experimental verification of Landauer’s principle linking information and thermodynamics”. In: *Nature* 483.7388 (2012), pp. 187–189. DOI: 10.1038/nature10872.

[2192] S. Chatrchyan et al. “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”. In: *Physics Letters B* 716.1 (2012). CMS collaboration; keyed by first listed author., pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021.

[2193] G. Degrassi et al. “Higgs mass and vacuum stability in the Standard Model at NNLO”. In: *Journal of High Energy Physics* 2012.8 (2012). Issue 08, article 098., p. 098. DOI: 10.1007/JHEP08(2012)098.

[2194] K. Hinterbichler. “Theoretical aspects of massive gravity”. In: *Reviews of Modern Physics* 84.2 (2012), pp. 671–710. DOI: 10.1103/RevModPhys.84.671.

[2195] T. Kessler et al. “A sub-40-mHz-linewidth laser based on a silicon single-crystal optical cavity”. In: *Nature Photonics* 6.10 (2012), pp. 687–692. DOI: 10.1038/nphoton.2012.217.

[2196] R. Kippenhahn, A. Weigert, and A. Weiss. *Stellar Structure and Evolution*. Springer, 2nd edition, 2012. DOI: 10.1007/978-3-642-30304-3.

[2197] J. P. Lees et al. “Observation of Time-Reversal Violation in theB0Meson System”. In: *Physical Review Letters* 109.21 (2012), p. 211801. DOI: 10.1103/PhysRevLett.109.211801.

[2198] T. W. Murphy et al. “APOLLO: millimeter lunar laser ranging”. In: *Classical and Quantum Gravity* 29.18 (2012). Issue 18; article number 184005 (no page range). Crossref confirmed title, journal, volume, issue, article number and year against the DOI, and returned eight of the nine authors – it omits E. L. Michelsen, who is on the published byline and on the ADS record, so the list above is the full published one. Chosen over Murphy’s review "Lunar laser ranging: the millimeter challenge", which is Rep. Prog. Phys. 76, 076901 with an ISSUE year of 2013 and would therefore be keyed Murphy:2013, not Murphy:2012. Cited at the lunar-laser-ranging apparatus paragraph (single-photon returns timed against an atomic clock) alongside Williams:2004., p. 184005. DOI: 10.1088/0264-9381/29/18/184005.

[2199] M. F. Pusey, J. Barrett, and T. Rudolph. “On the reality of the quantum state”. In: *Nature Physics* 8.6 (2012), pp. 475–478. DOI: 10.1038/nphys2309.

[2200] U. Seifert. “Stochastic thermodynamics, fluctuation theorems and molecular machines”. In: *Reports on Progress in Physics* 75.12 (2012), p. 126001. DOI: 10.1088/0034-4885/75/12/126001.

[2201] G. Aad et al. “Evidence for the spin-0 nature of the Higgs boson using ATLAS data”. In: *Physics Letters B* 726.1-3 (2013), pp. 120–144. DOI: 10.1016/j.physletb.2013.08.026.

[2202] R. Aaij et al. “First Observation ofCPViolation in the Decays ofBs0Mesons”. In: *Physical Review Letters* 110.22 (2013), p. 221601. DOI: 10.1103/PhysRevLett.110.221601.

[2203] M. G. Aartsen and others (IceCube Collaboration). “Evidence for High-Energy Extraterrestrial Neutrinos at the IceCube Detector”. In: *Science* 342.6161 (2013), p. 1242856. DOI: 10.1126/science.1242856.

[2204] J. Aasi et al. “Enhanced sensitivity of the LIGO gravitational wave detector by using squeezed states of light”. In: *Nature Photonics* 7.8 (2013), pp. 613–619. DOI: 10.1038/nphoton.2013.177.

[2205] M. Ackermann et al. “Detection of the Characteristic Pion-Decay Signature in Supernova Remnants”. In: *Science* 339.6121 (2013), pp. 807–811. DOI: 10.1126/science.1231160.

[2206] R. Agnese et al. “Silicon Detector Dark Matter Results from the Final Exposure of CDMS II”. In: *Physical Review Letters* 111.25 (2013). Issue 25; article number 251301. CDMS Collaboration — collaboration goes in ‘note‘, not ‘author‘. This is the silicon result with three candidate events against a smaller expected background, reported as not constituting a detection, which is the claim the bolometer section makes. Cited in the CDMS clause of sec:expdm-bolometers., p. 251301. DOI: 10.1103/PhysRevLett.111.251301.

[2207] M. Aguilar et al. “First Result from the Alpha Magnetic Spectrometer on the International Space Station: Precision Measurement of the Positron Fraction in Primary Cosmic Rays of 0.5–350 GeV”. In: *Physical Review Letters* 110.14 (2013), p. 141102. DOI: 10.1103/PhysRevLett.110.141102.

[2208] B. Aharmim, S. N. Ahmed, A. E. Anthony, et al. “Combined analysis of all three phases of solar neutrino data from the Sudbury Neutrino Observatory”. In: *Physical Review C* 88.2 (2013). SNO Collaboration; the final combined three-phase result. Metadata confirmed against Crossref., p. 025501. DOI: 10.1103/PhysRevC.88.025501.

[2209] I. Angeli and K. Marinova. “Table of experimental nuclear ground state charge radii: An update”. In: *Atomic Data and Nuclear Data Tables* 99.1 (2013), pp. 69–95. DOI: 10.1016/j.adt.2011.12.006.

[2210] A. Antognini et al. “Proton Structure from the Measurement of 2S-2P Transition Frequencies of Muonic Hydrogen”. In: *Science* 339.6118 (2013), pp. 417–420. DOI: 10.1126/science.1230016.

[2211] J. Antoniadis et al. “A Massive Pulsar in a Compact Relativistic Binary”. In: *Science* 340.6131 (2013), p. 1233232. DOI: 10.1126/science.1233232.

[2212] A. Bassi, K. Lochan, S. Satin, T. P. Singh, and H. Ulbricht. “Models of wave-function collapse, underlying theories, and experimental tests”. In: *Reviews of Modern Physics* 85.2 (2013), pp. 471–527. DOI: 10.1103/RevModPhys.85.471.

[2213] C. L. Bennett et al. “NINE-YEAR WILKINSON MICROWAVE ANISOTROPY PROBE (WMAP) OBSERVATIONS: FINAL MAPS AND RESULTS”. In: *The Astrophysical Journal Supplement Series* 208.2 (2013), p. 20. DOI: 10.1088/0067-0049/208/2/20.

[2214] S. Braun et al. “Negative Absolute Temperature for Motional Degrees of Freedom”. In: *Science* 339.6115 (2013), pp. 52–55. DOI: 10.1126/science.1227831.

[2215] D. Buttazzo et al. “Investigating the near-criticality of the Higgs boson”. In: *Journal of High Energy Physics* 2013.12 (2013), p. 089. DOI: 10.1007/JHEP12(2013)089.

[2216] C.-Z. Chang et al. “Experimental Observation of the Quantum Anomalous Hall Effect in a Magnetic Topological Insulator”. In: *Science* 340.6129 (2013), pp. 167–170. DOI: 10.1126/science.1234414.

[2217] S. Chatrchyan et al. “Study of the Mass and Spin-Parity of the Higgs Boson Candidate via Its Decays toZBoson Pairs”. In: *Physical Review Letters* 110.8 (2013), p. 081803. DOI: 10.1103/PhysRevLett.110.081803.

[2218] A. Gando, Y. Gando, H. Hanakago, et al. “Reactor on-off antineutrino measurement with KamLAND”. In: *Physical Review D* 88.3 (2013). KamLAND Collaboration; the final reactor analysis, taken across the Japanese reactor shutdown. Metadata confirmed against Crossref., p. 033001. DOI: 10.1103/PhysRevD.88.033001.

[2219] D. Hanson et al. “Detection ofB-Mode Polarization in the Cosmic Microwave Background with Data from the South Pole Telescope”. In: *Physical Review Letters* 111.14 (2013), p. 141301. DOI: 10.1103/PhysRevLett.111.141301.

[2220] S. Haroche. “Nobel Lecture: Controlling photons in a box and exploring the quantum to classical boundary”. In: *Reviews of Modern Physics* 85.3 (2013), pp. 1083–1102. DOI: 10.1103/RevModPhys.85.1083.

[2221] A. Matveev et al. “Precision Measurement of the Hydrogen1S−2SFrequency via a 920-km Fiber Link”. In: *Physical Review Letters* 110.23 (2013), p. 230801. DOI: 10.1103/PhysRevLett.110.230801.

[2222] T. Quinn, H. Parks, C. Speake, and R. Davis. “Improved Determination ofGUsing Two Methods”. In: *Physical Review Letters* 111.10 (2013). BIPM balance in servo and Cavendish modes; erratum Phys. Rev. Lett. 113 (2014) 039901. Canonically discrepant with Gundlach:2000., p. 101102. DOI: 10.1103/PhysRevLett.111.101102.

[2223] S. Schael et al. “Electroweak measurements in electron–positron collisions at W-boson-pair energies at LEP”. In: *Physics Reports* 532.4 (2013). ALEPH, DELPHI, L3 and OPAL Collaborations, pp. 119–244. DOI: 10.1016/j.physrep.2013.07.004.

[2224] D. Steppenbeck et al. “Evidence for a new nuclear ‘magic number’ from the level structure of 54Ca”. In: *Nature* 502.7470 (2013), pp. 207–210. DOI: 10.1038/nature12522.

[2225] V. Tishchenko et al. “Detailed report of the MuLan measurement of the positive muon lifetime and determination of the Fermi constant”. In: *Physical Review D* 87.5 (2013), p. 052003. DOI: 10.1103/PhysRevD.87.052003.

[2226] A. Aab et al. “Depth of maximum of air-shower profiles at the Pierre Auger Observatory. I. Measurements at energies above1017.8 eV”. In: *Physical Review D* 90.12 (2014), p. 122005. DOI: 10.1103/PhysRevD.90.122005.

[2227] P. A. R. Ade et al. “Detection ofB-Mode Polarization at Degree Angular Scales by BICEP2”. In: *Physical Review Letters* 112.24 (2014). Primordial claim later attributed to Galactic dust; see Ade:2015, p. 241101. DOI: 10.1103/PhysRevLett.112.241101.

[2228] M. Baak et al. “The global electroweak fit at NNLO and prospects for the LHC and ILC”. In: *The European Physical Journal C* 74.9 (2014), p. 3046. DOI: 10.1140/epjc/s10052-014-3046-5.

[2229] G. Bellini and others (Borexino Collaboration). “Neutrinos from the primary proton–proton fusion process in the Sun”. In: *Nature* 512.7515 (2014), pp. 383–386. DOI: 10.1038/nature13702.

[2230] J. Billard, E. Figueroa-Feliciano, and L. Strigari. “Implication of neutrino backgrounds on the reach of next generation dark matter direct detection experiments”. In: *Physical Review D* 89.2 (2014). Issue 2; article number 023524. The original neutrino-floor computation: coherent neutrino-nucleus scattering scales with exposure as the signal does, so reach improves at best as the square root. Cited in the XENONnT derivation., p. 023524. DOI: 10.1103/PhysRevD.89.023524.

[2231] L. Blanchet. “Gravitational Radiation from Post-Newtonian Sources and Inspiralling Compact Binaries”. In: *Living Reviews in Relativity* 17 (2014), p. 2. DOI: 10.12942/lrr-2014-2.

[2232] N. Brunner, D. Cavalcanti, S. Pironio, V. Scarani, and S. Wehner. “Bell nonlocality”. In: *Reviews of Modern Physics* 86.2 (2014), pp. 419–478. DOI: 10.1103/RevModPhys.86.419.

[2233] C. A. Fuchs, N. D. Mermin, and R. Schack. “An introduction to QBism with an application to the locality of quantum mechanics”. In: *American Journal of Physics* 82.8 (2014), pp. 749–754. DOI: 10.1119/1.4874855.

[2234] S. Kheifets, A. Simha, K. Melin, T. Li, and M. G. Raizen. “Observation of Brownian Motion in Liquids at Short Times: Instantaneous Velocity and Memory Loss”. In: *Science* 343.6178 (2014), pp. 1493–1496. DOI: 10.1126/science.1248091.

[2235] J.-Å. Larsson. “Loopholes in Bell inequality tests of local realism”. In: *Journal of Physics A: Mathematical and Theoretical* 47.42 (2014), p. 424003. DOI: 10.1088/1751-8113/47/42/424003.

[2236] Real Academia Española. *Diccionario de la lengua española*. 23rd ed. Edición del Tricentenario. Freely consultable online at dle.rae.es, where the text is revised periodically. Espasa. Madrid, 2014.

[2237] G. Rosi, F. Sorrentino, L. Cacciapuoti, M. Prevedelli, and G. M. Tino. “Precision measurement of the Newtonian gravitational constant using cold atoms”. In: *Nature* 510.7506 (2014). Atom-interferometric determination; systematics independent of torsion fibres., pp. 518–521. DOI: 10.1038/nature13433.

[2238] I. G. Shevtsova. “On the absolute constants in the Berry–Esseen-type inequalities”. In: *Doklady Mathematics* 89.3 (2014). Source of the constant $C \leq 0.4690$ for identically distributed summands. The frequently cited 2011 preprint arXiv:1111.6554 proves the weaker $C < 0.4748$, pp. 378–381. DOI: 10.1134/S1064562414030338.

[2239] S. El-Showk, M. F. Paulos, D. Poland, S. Rychkov, D. Simmons-Duffin, and A. Vichi. “Solving the 3d Ising Model with the Conformal Bootstrap II. c c -Minimization and Precise Critical Exponents”. In: *Journal of Statistical Physics* 157.4-5 (2014), pp. 869–914. DOI: 10.1007/s10955-014-1042-7.

[2240] C. M. Will. “The Confrontation between General Relativity and Experiment”. In: *Living Reviews in Relativity* 17.1 (2014). Article number 4; open access., p. 4. DOI: 10.12942/lrr-2014-4.

[2241] A. Aab and others (Pierre Auger Collaboration). “The Pierre Auger Cosmic Ray Observatory”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 798 (2015), pp. 172–213. DOI: 10.1016/j.nima.2015.06.058.

[2242] G. Aad et al. “Combined Measurement of the Higgs Boson Mass inppCollisions ats=7and 8 TeV with the ATLAS and CMS Experiments”. In: *Physical Review Letters* 114.19 (2015), p. 191803. DOI: 10.1103/PhysRevLett.114.191803.

[2243] J. Aasi et al. “Advanced LIGO”. In: *Classical and Quantum Gravity* 32 (2015). LIGO Scientific Collaboration, p. 074001. DOI: 10.1088/0264-9381/32/7/074001.

[2244] H. Abramowicz et al. “Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data”. In: *The European Physical Journal C* 75.12 (2015), p. 580. DOI: 10.1140/epjc/s10052-015-3710-4.

[2245] M. Ackermann et al. “Searching for Dark Matter Annihilation from Milky Way Dwarf Spheroidal Galaxies with Six Years of Fermi Large Area Telescope Data”. In: *Physical Review Letters* 115.23 (2015). Issue 23; article number 231301. Fermi-LAT Collaboration — put the collaboration name in ‘note‘, never in ‘author‘ (the comma trap makes biber skip the whole entry). This is the stacked dwarf-spheroidal joint-likelihood analysis whose bound crosses the thermal relic value below 100 GeV. Cited in phen:expdm-fermi. Distinct from the existing key Ackermann:2013 (pion-decay signature in supernova remnants, Science 339)., p. 231301. DOI: 10.1103/PhysRevLett.115.231301.

[2246] P. A. R. Ade et al. “Joint Analysis of BICEP2/Keck ArrayandPlanckData”. In: *Physical Review Letters* 114.10 (2015), p. 101301. DOI: 10.1103/PhysRevLett.114.101301.

[2247] N. Agafonova et al. “Discovery ofτNeutrino Appearance in the CNGS Neutrino Beam with the OPERA Experiment”. In: *Physical Review Letters* 115.12 (2015), p. 121802. DOI: 10.1103/PhysRevLett.115.121802.

[2248] M. Agostini et al. “Test of Electric Charge Conservation with Borexino”. In: *Physical Review Letters* 115.23 (2015). Source of the e $\to$ nu gamma lifetime bound $>$ 6.6e28 yr quoted in the chapter. Collaboration paper keyed by first listed author. ’231802’ is an article number. Issued 3 Dec 2015., p. 231802. DOI: 10.1103/PhysRevLett.115.231802.

[2249] A. Aguilar-Arevalo et al. “Improved Measurement of the$\pi \to$eνBranching Ratio”. In: *Physical Review Letters* 115.7 (2015), p. 071801. DOI: 10.1103/PhysRevLett.115.071801.

[2250] E. Aver, K. A. Olive, and E. D. Skillman. “The effects of He I λ10830 on helium abundance determinations”. In: *Journal of Cosmology and Astroparticle Physics* 2015.07 (2015). Metadata confirmed against Crossref., p. 011. DOI: 10.1088/1475-7516/2015/07/011.

[2251] T. S. Cubitt, D. Perez-Garcia, and M. M. Wolf. “Undecidability of the spectral gap”. In: *Nature* 528.7581 (2015). Short version; the full proofs occupy arXiv:1502.04573, pp. 207–211. DOI: 10.1038/nature16059.

[2252] A. P. Drozdov, M. I. Eremets, I. A. Troyan, V. Ksenofontov, and S. I. Shylin. “Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system”. In: *Nature* 525.7567 (2015), pp. 73–76. DOI: 10.1038/nature14964.

[2253] M. Giustina et al. “Significant-Loophole-Free Test of Bell’s Theorem with Entangled Photons”. In: *Physical Review Letters* 115.25 (2015), p. 250401. DOI: 10.1103/PhysRevLett.115.250401.

[2254] B. Hensen et al. “Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres”. In: *Nature* 526.7575 (2015), pp. 682–686. DOI: 10.1038/nature15759.

[2255] O. James, E. von Tunzelmann, P. Franklin, and K. S. Thorne. “Gravitational lensing by spinning black holes in astrophysics, and in the movie Interstellar”. In: *Classical and Quantum Gravity* 32.6 (2015). The DNGR ray-bundle renderer built for the film, and the source for the statement that the Doppler and gravitational frequency shifts were deliberately omitted from the rendered images. Crossref’s deposited title carries JATS italic markup around the film name, which is dropped here. Metadata confirmed against Crossref., p. 065001. DOI: 10.1088/0264-9381/32/6/065001.

[2256] A. D. Ludlow, M. M. Boyd, J. Ye, E. Peik, and P. O. Schmidt. "Optical atomic clocks". In: *Reviews of Modern Physics* 87.2 (2015), pp. 637–701. DOI: 10.1103/RevModPhys.87.637.

[2257] B. Q. Lv et al. "Experimental Discovery of Weyl Semimetal TaAs". In: *Physical Review X* 5.3 (2015), p. 031013. DOI: 10.1103/PhysRevX.5.031013.

[2258] M. Nagel et al. "Direct terrestrial test of Lorentz symmetry in electrodynamics to 10^{-18}". In: *Nature Communications* 6 (2015). The modern descendant of Michelson–Morley: rotating cryogenic microwave cavities. Not to be confused with Nagel:1961, Ernest Nagel's monograph, p. 8174. DOI: 10.1038/ncomms9174.

[2259] T. Ortín. *Gravity and Strings*. 2nd ed. Cambridge Monographs on Mathematical Physics. Cambridge: Cambridge University Press, 2015.

[2260] L. K. Shalm et al. "Strong Loophole-Free Test of Local Realism". In: *Physical Review Letters* 115.25 (2015), p. 250402. DOI: 10.1103/PhysRevLett.115.250402.

[2261] J. B. Whitmore and M. T. Murphy. "Impact of instrumental systematic errors on fine-structure constant measurements with quasar spectra". In: *Monthly Notices of the Royal Astronomical Society* 447.1 (2015). Advance access 18 December 2014; the issue of record is 11 February 2015, which is the universal citation and fixes the key. Murphy co-authored Webb:2011, so this is a self-critique from inside the original collaboration. It claims the evidence is significantly weakened, not refuted, pp. 446–462. DOI: 10.1093/mnras/stu2420.

[2262] S.-Y. Xu et al. "Discovery of a Weyl fermion semimetal and topological Fermi arcs". In: *Science* 349.6248 (2015), pp. 613–617. DOI: 10.1126/science.aaa9297.

[2263] G. Aad et al. "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV". In: *Journal of High Energy Physics* 2016.8 (2016). Issue 08, article 045., p. 045. DOI: 10.1007/JHEP08(2016)045.

[2264] B. P. Abbott et al. "Observation of Gravitational Waves from a Binary Black Hole Merger". In: *Physical Review Letters* 116 (2016). LIGO Scientific Collaboration and Virgo Collaboration, p. 061102. DOI: 10.1103/PhysRevLett.116.061102.

[2265] A. Accardi et al. "Electron-Ion Collider: The next QCD frontier". In: *The European Physical Journal A* 52.9 (2016), p. 268. DOI: 10.1140/epja/i2016-16268-9.

[2266] M. Aguilar et al. "Precision Measurement of the Boron to Carbon Flux Ratio in Cosmic Rays from 1.9 GV to 2.6 TV with the Alpha Magnetic Spectrometer on the International Space Station". In: *Physical Review Letters* 117.23 (2016), p. 231102. DOI: 10.1103/PhysRevLett.117.231102.

[2267] G. Angloher et al. "Results on light dark matter particles with a low-threshold CRESST-II detector". In: *The European Physical Journal C* 76.1 (2016). Issue 1; article number 25 (that is the article number, not a first page of a range). CRESST Collaboration — collaboration in 'note', not 'author'. The CRESST-II low-threshold light-dark-matter exclusion. Cited with Abdelhameed:2019 at the CRESST sentence in sec:expdm-bolometers., p. 25. DOI: 10.1140/epjc/s10052-016-3877-3.

[2268] J. L. Burch et al. "Electron-scale measurements of magnetic reconnection in space". In: *Science* 352.6290 (2016). Crossref confirmed: Science vol. 352, issue 6290, article number aaf2939 (there is no conventional page range – use eid/article number, not a fabricated page). The MMS four-spacecraft electron-diffusion-region measurement; large author list, keyed by first listed author Burch per the collaboration convention., aaf2939. DOI: 10.1126/science.aaf2939.

[2269] A. S. Chou et al. “First Measurements of High Frequency Cross-Spectra from a Pair of Large Michelson Interferometers”. In: *Physical Review Letters* 117.11 (2016), p. 111102. DOI: 10.1103/PhysRevLett.117.111102.

[2270] R. H. Cyburt, B. D. Fields, K. A. Olive, and T.-H. Yeh. “Big bang nucleosynthesis: Present status”. In: *Reviews of Modern Physics* 88.1 (2016), p. 015004. DOI: 10.1103/RevModPhys.88.015004.

[2271] T. K. Gaisser, R. Engel, and E. Resconi. *Cosmic Rays and Particle Physics*. 2nd edition. Cambridge University Press, 2016. DOI: 10.1017/CBO9781139192194.

[2272] A. M. Kaufman et al. “Quantum thermalization through entanglement in an isolated many-body system”. In: *Science* 353.6301 (2016), pp. 794–800. DOI: 10.1126/science.aaf6725.

[2273] S. S. McGaugh, F. Lelli, and J. M. Schombert. “Radial Acceleration Relation in Rotationally Supported Galaxies”. In: *Physical Review Letters* 117.20 (2016), p. 201101. DOI: 10.1103/PhysRevLett.117.201101.

[2274] J. Steinhauer. “Observation of quantum Hawking radiation and its entanglement in an analogue black hole”. In: *Nature Physics* 12.10 (2016), pp. 959–965. DOI: 10.1038/nphys3863.

[2275] J. M. Weisberg and Y. Huang. “Relativistic Measurements from Timing the Binary Pulsar PSR B1913+16”. In: *The Astrophysical Journal* 829 (2016), p. 55. DOI: 10.3847/0004-637X/829/1/55.

[2276] A. Aab et al. “Observation of a large-scale anisotropy in the arrival directions of cosmic rays above 8 × 10 18 eV”. In: *Science* 357.6357 (2017), pp. 1266–1270. DOI: 10.1126/science.aan4338.

[2277] Aaboud. “Evidence for light-by-light scattering in heavy-ion collisions with the ATLAS detector at the LHC”. In: *Nature Physics* 13.9 (2017). ATLAS Collaboration; first listed author M. Aaboud, not G. Aad, pp. 852–858. DOI: 10.1038/nphys4208.

[2278] B. P. Abbott et al. “GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral”. In: *Physical Review Letters* 119 (2017). LIGO Scientific Collaboration and Virgo Collaboration, p. 161101. DOI: 10.1103/PhysRevLett.119.161101.

[2279] R. Abuter et al. “First light for GRAVITY: Phase referencing optical interferometry for the Very Large Telescope Interferometer”. In: *Astronomy & Astrophysics* 602 (2017). Collaboration paper keyed by first listed author per house convention, A94. DOI: 10.1051/0004-6361/201730838.

[2280] M. Ahmadi et al. “Observation of the 1S–2S transition in trapped antihydrogen”. In: *Nature* 541.7638 (2017). ALPHA collaboration; keyed by first listed author., pp. 506–510. DOI: 10.1038/nature21040.

[2281] A. Beyer et al. “The Rydberg constant and proton size from atomic hydrogen”. In: *Science* 358.6359 (2017), pp. 79–85. DOI: 10.1126/science.aah6677.

[2282] S. Bose et al. “Spin Entanglement Witness for Quantum Gravity”. In: *Physical Review Letters* 119.24 (2017), p. 240401. DOI: 10.1103/PhysRevLett.119.240401.

[2283] J. Handsteiner et al. “Cosmic Bell Test: Measurement Settings from Milky Way Stars”. In: *Physical Review Letters* 118.6 (2017), p. 060401. DOI: 10.1103/PhysRevLett.118.060401.

[2284] C. Marletto and V. Vedral. “Gravitationally Induced Entanglement between Two Massive Particles is Sufficient Evidence of Quantum Effects in Gravity”. In: *Physical Review Letters* 119.24 (2017), p. 240402. DOI: 10.1103/PhysRevLett.119.240402.

[2285] W. Rosenfeld et al. “Event-Ready Bell Test Using Entangled Atoms Simultaneously Closing Detection and Locality Loopholes”. In: *Physical Review Letters* 119.1 (2017), p. 010402. DOI: 10.1103/PhysRevLett.119.010402.

[2286] C. Rothleitner and S. Schlamminger. “Invited Review Article: Measurements of the Newtonian constant of gravitation, G”. In: *Review of Scientific Instruments* 88.11 (2017). Review of all modern G determinations and their mutual discrepancies., p. 111101. DOI: 10.1063/1.4994619.

[2287] J. J. Sakurai and J. Napolitano. *Modern Quantum Mechanics*. 2nd ed. Cambridge: Cambridge University Press, 2017.

[2288] C. Smorra et al. “A parts-per-billion measurement of the antiproton magnetic moment”. In: *Nature* 550.7676 (2017), pp. 371–374. DOI: 10.1038/nature24048.

[2289] P. Touboul et al. “MICROSCOPE Mission: First Results of a Space Test of the Equivalence Principle”. In: *Physical Review Letters* 119.23 (2017). Collaboration paper keyed by first listed author., p. 231101. DOI: 10.1103/PhysRevLett.119.231101.

[2290] J. Yin et al. “Satellite-based entanglement distribution over 1200 kilometers”. In: *Science* 356.6343 (2017), pp. 1140–1144. DOI: 10.1126/science.aan3211.

[2291] X. Zheng, Y. R. Sun, J.-J. Chen, W. Jiang, K. Pachucki, and S.-M. Hu. “Measurement of the Frequency of the2 S3−2 P3Transition ofHe4”. In: *Physical Review Letters* 119.26 (2017), p. 263002. DOI: 10.1103/PhysRevLett.119.263002.

[2292] M. G. Aartsen, M. Ackermann, J. Adams, et al. “Measurement of Atmospheric Neutrino Oscillations at 6–56 GeV with IceCube DeepCore”. In: *Physical Review Letters* 120.7 (2018). IceCube Collaboration. Suffix a because the bare key Aartsen:2018 is already in use for the TXS 0506+056 blazar paper in Science 361, 147 – two different works of the same collaboration and year; cite by content, not by letter. Metadata confirmed against Crossref., p. 071801. DOI: 10.1103/PhysRevLett.120.071801.

[2293] M. Aartsen et al. “Neutrino emission from the direction of the blazar TXS 0506+056 prior to the IceCube-170922A alert”. In: *Science* 361.6398 (2018), pp. 147–151. DOI: 10.1126/science.aat2890.

[2294] B. P. Abbott et al. “GW170817: Measurements of Neutron Star Radii and Equation of State”. In: *Physical Review Letters* 121.16 (2018), p. 161101. DOI: 10.1103/PhysRevLett.121.161101.

[2295] R. Abuter et al. “Detection of the gravitational redshift in the orbit of the star S2 near the Galactic centre massive black hole”. In: *Astronomy & Astrophysics* 615 (2018), p. L15. DOI: 10.1051/0004-6361/201833718.

[2296] G. Adhikari and others (COSINE-100 Collaboration). “An experiment to search for dark-matter interactions using sodium iodide detectors”. In: *Nature* 564.7734 (2018), pp. 83–86. DOI: 10.1038/s41586-018-0739-1.

[2297] M. Agostini et al. “Comprehensive measurement of pp-chain solar neutrinos”. In: *Nature* 562.7728 (2018), pp. 505–510. DOI: 10.1038/s41586-018-0624-y.

[2298] A. A. Aguilar-Arevalo, B. C. Brown, L. Bugel, et al. “Significant Excess of Electron-like Events in the MiniBooNE Short-Baseline Neutrino Experiment”. In: *Physical Review Letters* 121.22 (2018). MiniBooNE Collaboration; the low-energy excess. Distinct from AguilarArevalo:2015, which is the PIENU pion-decay-branching measurement. Metadata confirmed against Crossref., p. 221801. DOI: 10.1103/PhysRevLett.121.221801.

[2299] M. Ahmadi et al. “Characterization of the 1S–2S transition in antihydrogen”. In: *Nature* 557.7703 (2018), pp. 71–75. DOI: 10.1038/s41586-018-0017-2.

[2300] V. Andreev and others (ACME Collaboration). “Improved limit on the electric dipole moment of the electron”. In: *Nature* 562.7727 (2018), pp. 355–360. DOI: 10.1038/s41586-018-0599-8.

[2301] D. Androić and others (Qweak Collaboration). “Precision measurement of the weak charge of the proton”. In: *Nature* 557.7704 (2018), pp. 207–211. DOI: 10.1038/s41586-018-0096-0.

[2302] T. Aoyama, T. Kinoshita, and M. Nio. “Revised and improved value of the QED tenth-order electron anomalous magnetic moment”. In: *Physical Review D* 97.3 (2018), p. 036001. DOI: 10.1103/PhysRevD.97.036001.

[2303] P. Bierhorst et al. “Experimentally generated randomness certified by the impossibility of superluminal signals”. In: *Nature* 556.7700 (2018), pp. 223–226. DOI: 10.1038/s41586-018-0019-0.

[2304] M. A. Ciampini, G. Pinna, P. Mataloni, and M. Paternostro. “Experimental signature of quantum Darwinism in photonic cluster states”. In: *Physical Review A* 98.2 (2018), p. 020101. DOI: 10.1103/PhysRevA.98.020101.

[2305] J. M. Cole et al. “Experimental Evidence of Radiation Reaction in the Collision of a High-Intensity Laser Pulse with a Laser-Wakefield Accelerated Electron Beam”. In: *Physical Review X* 8.1 (2018). Crossref confirmed: PRX vol. 8, article number 011020, 7 February 2018 (article number, not a page range). PRX is fully open access, so this should grade GREEN on availability with the DOI alone. Cited in the chapter together with Poder:2018, the companion radiation-reaction measurement., p. 011020. DOI: 10.1103/PhysRevX.8.011020.

[2306] I. Collaboration et al. “Multimessenger observations of a flaring blazar coincident with high-energy neutrino IceCube-170922A”. In: *Science* 361 (2018). Keyed by the collaboration because the article’s author field is a list of collaborations, not of individuals; companion paper is Aartsen:2018, eaat1378.

[2307] T. B. B. T. Collaboration. “Challenging local realism with human choices”. In: *Nature* 557.7704 (2018). Corporate author with no first listed individual; keyed on the collaboration name, as BIPM:2019 and JCGM:2008 already are., pp. 212–216. DOI: 10.1038/s41586-018-0085-3.

[2308] R. J. Cooke, M. Pettini, and C. C. Steidel. “One Percent Determination of the Primordial Deuterium Abundance”. In: *The Astrophysical Journal* 855 (2018). Metadata confirmed against Crossref., p. 102. DOI: 10.3847/1538-4357/aaab53.

[2309] N. Du et al. “Search for Invisible Axion Dark Matter with the Axion Dark Matter Experiment”. In: *Physical Review Letters* 120.15 (2018), p. 151301. DOI: 10.1103/PhysRevLett.120.151301.

[2310] J. Fischer et al. “The Boltzmann project”. In: *Metrologia* 55.2 (2018). Issue 2 (April 2018 print; online 5 February 2018). The review of the k_B determinations (acoustic and dielectric-constant gas thermometry, Johnson noise, Doppler broadening) on which the 2019 fixing of k_B = 1.380 649e-23 J/K rests – i.e. the paper behind the chapter’s claim that the kelvin is now defined by fixing that number. 32 authors in the Crossref record; truncate with ’and others’ rather than listing all. WARNING per the house rule: the author list contains ’B. Darquié’, ’M. C. Martín’ and ’D. del Campo’ – keep the accents as Unicode and never let a parenthetical collaboration list with commas into the author field., R1–R20. DOI: 10.1088/1681-7575/aaa790.

[2311] H. Fleurbaey et al. “New Measurement of the 1S−3S Transition Frequency of Hydrogen: Contribution to the Proton Charge Radius Puzzle”. In: *Physical Review Letters* 120.18 (2018), p. 183001. DOI: 10.1103/PhysRevLett.120.183001.

[2312] D. Frauchiger and R. Renner. “Quantum theory cannot consistently describe the use of itself”. In: *Nature Communications* 9.1 (2018), p. 3711. DOI: 10.1038/s41467-018-05739-8.

[2313] D. J. Griffiths and D. F. Schroeter. *Introduction to Quantum Mechanics*. 3rd ed. Cambridge: Cambridge University Press, 2018.

[2314] Q. Li et al. “Measurements of the gravitational constant using two independent methods”. In: *Nature* 560.7720 (2018). Time-of-swing and angular-acceleration-feedback determinations, each near 12 ppm. Key does not collide with existing Li:2006 (Nan Li, different author and year)., pp. 582–588. DOI: 10.1038/s41586-018-0431-5.

[2315] W. F. McGrew et al. “Atomic clock performance enabling geodesy below the centimetre level”. In: *Nature* 564.7734 (2018), pp. 87–90. DOI: 10.1038/s41586-018-0738-2.

[2316] R. H. Parker, C. Yu, W. Zhong, B. Estey, and H. Müller. “Measurement of the fine-structure constant as a test of the Standard Model”. In: *Science* 360.6385 (2018), pp. 191–195. DOI: 10.1126/science.aap7706.

[2317] K. Poder et al. “Experimental Signatures of the Quantum Nature of Radiation Reaction in the Field of an Ultraintense Laser”. In: *Physical Review X* 8.3 (2018). Issue 3, 5 July 2018; article number 031004 (no page range). Open access (PRX is fully OA), so the availability grade is GREEN with the publisher URL. Companion to Cole:2018, cited alongside it in the same sentence., p. 031004. DOI: 10.1103/PhysRevX.8.031004.

[2318] D. Rauch et al. “Cosmic Bell Test Using Random Measurement Settings from High-Redshift Quasars”. In: *Physical Review Letters* 121.8 (2018), p. 080403. DOI: 10.1103/PhysRevLett.121.080403.

[2319] F. Riehle, P. Gill, F. Arias, and L. Robertsson. “The CIPM list of recommended frequency standard values: guidelines and procedures”. In: *Metrologia* 55.2 (2018), pp. 188–200. DOI: 10.1088/1681-7575/aaa302.

[2320] A. M. Sirunyan et al. “Observation of $t\bar{t}H$ Production”. In: *Physical Review Letters* 120.23 (2018), p. 231801. DOI: 10.1103/PhysRevLett.120.231801.

[2321] S. Weyers et al. “Advances in the accuracy, stability, and reliability of the PTB primary fountain clocks”. In: *Metrologia* 55.6 (2018), pp. 789–805. DOI: 10.1088/1681-7575/aae008.

[2322] G. Aad et al. “Observation of Light-by-Light Scattering in Ultraperipheral Pb+Pb Collisions with the ATLAS Detector”. In: *Physical Review Letters* 123.5 (2019), p. 052001. DOI: 10.1103/PhysRevLett.123.052001.

[2323] R. Aaij et al. “Observation of CP Violation in Charm Decays”. In: *Physical Review Letters* 122.21 (2019), p. 211803. DOI: 10.1103/PhysRevLett.122.211803.

[2324] B. P. Abbott et al. “GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs”. In: *Physical Review X* 9.3 (2019), p. 031040. DOI: 10.1103/PhysRevX.9.031040.

[2325] A. H. Abdelhameed et al. “First results from the CRESST-III low-mass dark matter program”. In: *Physical Review D* 100.10 (2019). Issue 10; article number 102002. CRESST Collaboration — collaboration in ‘note‘, not ‘author‘. CRESST-III is the run that reaches the lowest candidate masses of any direct-detection experiment, which is the specific claim it is attached to. Cited with Angloher:2016 in sec:expdm-bolometers., p. 102002. DOI: 10.1103/PhysRevD.100.102002.

[2326] V. A. Acciari et al. “Teraelectronvolt emission from the γ-ray burst GRB 190114C”. In: *Nature* 575.7783 (2019), pp. 455–458. DOI: 10.1038/s41586-019-1750-x.

[2327] M. Aguilar, L. Ali Cavasonza, G. Ambrosi, L. Arruda, et al. “Towards Understanding the Origin of Cosmic-Ray Positrons”. In: *Physical Review Letters* 122.4 (2019). AMS Collaboration. The positron flux to 1 TeV, showing the excess and its high-energy turnover; the companion Aguilar:2013 covers only 0.5–350 GeV and must not be stretched to cover the turnover. Metadata confirmed against Crossref., p. 041102. DOI: 10.1103/PhysRevLett.122.041102.

[2328] K. Akiyama et al. “First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole”. In: *The Astrophysical Journal Letters* 875.1 (2019). Collaboration paper keyed by first listed author per house convention, p. L1. DOI: 10.3847/2041-8213/ab0ec7.

[2329] D. Antypas, A. Fabricant, J. E. Stalnaker, K. Tsigutkin, V. V. Flambaum, and D. Budker. “Isotopic variation of parity violation in atomic ytterbium”. In: *Nature Physics* 15.2 (2019), pp. 120–123. DOI: 10.1038/s41567-018-0312-8.

[2330] T. Aoyama, T. Kinoshita, and M. Nio. “Theory of the Anomalous Magnetic Moment of the Electron”. In: *Atoms* 7.1 (2019), p. 28. DOI: 10.3390/atoms7010028.

[2331] N. Bezginov, T. Valdez, M. Horbatsch, A. Marsman, A. C. Vutha, and E. A. Hessels. “A measurement of the atomic hydrogen Lamb shift and the proton charge radius”. In: *Science* 365.6457 (2019), pp. 1007–1012. DOI: 10.1126/science.aau7807.

[2332] S. M. Brewer et al. “Al+27 Quantum-Logic Clock with a Systematic Uncertainty below 10−18”. In: *Physical Review Letters* 123.3 (2019), p. 033201. DOI: 10.1103/PhysRevLett.123.033201.

[2333] Bureau International des Poids et Mesures. *The International System of Units (SI)*. 9th ed. Sèvres, 2019.

[2334] Consultative Committee for Thermometry. *Mise en pratique for the definition of the kelvin in the SI*. MeP-K-19, version 2.0, dated 20 May 2019; the CCT document that makes absolute radiometry (and the other listed methods) a recognised primary realization of the kelvin. Corporate author, so brace-protect it and key it on the committee exactly as BIPM:2019 and JCGM:2008 already are in this bibliography. No DOI: Crossref’s 10.59161/* DOIs under BIPM cover CCT *Recommendations* (e.g. 10.59161/cct2005rec3e, ’Creation of a mise en pratique of the definition of the kelvin’), not the MeP document itself – do not attach one of those to this entry. Access route for the manifest ("access": "open"): https://www.bipm.org/documents/20126/41489682/SI-App2-kelvin.pdf, which is served as Appendix 2 of the 9th-edition SI Brochure. Address Sevres, matching BIPM:2019. 2019.

[2335] H. T. Cromartie et al. “Relativistic Shapiro delay measurements of an extremely massive millisecond pulsar”. In: *Nature Astronomy* 4.1 (2019), pp. 72–76. DOI: 10.1038/s41550-019-0880-2.

[2336] T. Do, A. Hees, A. Ghez, G. D. Martinez, D. S. Chu, et al. “Relativistic redshift of the star S0-2 orbiting the Galactic Center supermassive black hole”. In: *Science* 365 (2019). Metadata confirmed against Crossref., pp. 664–668. DOI: 10.1126/science.aav8137.

[2337] Y. Y. Fein et al. “Quantum superposition of molecules beyond 25 kDa”. In: *Nature Physics* 15.12 (2019). Fourth author’s surname carries a Polish barred l: Kiałka., pp. 1242–1245. DOI: 10.1038/s41567-019-0663-9.

[2338] M. C. Miller et al. “PSR J0030+0451 Mass and Radius from NICER Data and Implications for the Properties of Neutron Star Matter”. In: *The Astrophysical Journal Letters* 887.1 (2019), p. L24. DOI: 10.3847/2041-8213/ab50c5.

[2339] T. E. Riley et al. “A NICER View of PSR J0030+0451: Millisecond Pulsar Parameter Estimation”. In: *The Astrophysical Journal Letters* 887.1 (2019), p. L21. DOI: 10.3847/2041-8213/ab481c.

[2340] M. Somayazulu et al. “Evidence for Superconductivity above 260 K in Lanthanum Superhydride at Megabar Pressures”. In: *Physical Review Letters* 122.2 (2019), p. 027001. DOI: 10.1103/PhysRevLett.122.027001.

[2341] P.-E. Tremblay et al. “Core crystallization and pile-up in the cooling sequence of evolving white dwarfs”. In: *Nature* 565 (2019). Metadata confirmed against Crossref., pp. 202–205. DOI: 10.1038/s41586-018-0791-x.

[2342] M. Tse et al. “Quantum-Enhanced Advanced LIGO Detectors in the Era of Gravitational-Wave Astronomy”. In: *Physical Review Letters* 123.23 (2019), p. 231107. DOI: 10.1103/PhysRevLett.123.231107.

[2343] S. Volkov. “Calculating the five-loop QED contribution to the electron anomalous magnetic moment: Graphs without lepton loops”. In: *Physical Review D* 100.9 (2019), p. 096004. DOI: 10.1103/PhysRevD.100.096004.

[2344] W. Xiong et al. “A small proton charge radius from an electron–proton scattering experiment”. In: *Nature* 575.7781 (2019), pp. 147–150. DOI: 10.1038/s41586-019-1721-2.

[2345] K. Abe et al. “Constraint on the matter–antimatter symmetry-violating phase in neutrino oscillations”. In: *Nature* 580.7803 (2020), pp. 339–344. DOI: 10.1038/s41586-020-2177-0.

[2346] C. Abel et al. “Measurement of the Permanent Electric Dipole Moment of the Neutron”. In: *Physical Review Letters* 124.8 (2020), p. 081803. DOI: 10.1103/PhysRevLett.124.081803.

[2347] R. Abuter et al. “Detection of the Schwarzschild precession in the orbit of the star S2 near the Galactic centre massive black hole”. In: *Astronomy & Astrophysics* 636 (2020), p. L5. DOI: 10.1051/0004-6361/202037813.

[2348] N. Aghanim et al. “Planck 2018 results. VI. Cosmological parameters”. In: *Astronomy & Astrophysics* 641 (2020). Planck Collaboration, A6. DOI: 10.1051/0004-6361/201833910.

[2349] M. Agostini and others (Borexino Collaboration). “Experimental evidence of neutrinos produced in the CNO fusion cycle in the Sun”. In: *Nature* 587.7835 (2020). Nature bylines the paper to The Borexino Collaboration; keyed by the first listed author M. Agostini per the collaboration-paper convention., pp. 577–582. DOI: 10.1038/s41586-020-2934-0.

[2350] M. Agostini et al. “Final Results of GERDA on the Search for Neutrinoless Double-β Decay”. In: *Physical Review Letters* 125.25 (2020). Article number 252502. Title verified against Crossref: Physical Review Letters prints the Greek beta, not "beta"., p. 252502. DOI: 10.1103/PhysRevLett.125.252502.

[2351] S. Aiola et al. “The Atacama Cosmology Telescope: DR4 maps and cosmological parameters”. In: *Journal of Cosmology and Astroparticle Physics* 2020.12 (2020). Issue 12, pp. 047–047. DOI: 10.1088/1475-7516/2020/12/047.

[2352] Y. Akrami et al. “Planck 2018 results. I. Overview and the cosmological legacy of Planck”. In: *Astronomy & Astrophysics* 641 (2020). Planck Collaboration; source of the solar-dipole amplitude and direction. Keyed by the printed byline’s first listed author (Akrami); Crossref’s deposited author array lists Aghanim first, a metadata artifact. The parameters paper (VI) of the same series is Aghanim:2020. Metadata confirmed against Crossref., A1. DOI: 10.1051/0004-6361/201833880.

[2353] T. Aoyama et al. “The anomalous magnetic moment of the muon in the Standard Model”. In: *Physics Reports* 887 (2020), pp. 1–166. DOI: 10.1016/j.physrep.2020.07.006.

[2354] H. Bartolomei et al. “Fractional statistics in anyon collisions”. In: *Science* 368.6487 (2020), pp. 173–177. DOI: 10.1126/science.aaz5601.

[2355] T. Braine et al. “Extended Search for the Invisible Axion with the Axion Dark Matter Experiment”. In: *Physical Review Letters* 124.10 (2020), p. 101303. DOI: 10.1103/PhysRevLett.124.101303.

[2356] I. Esteban, M. Gonzalez-Garcia, M. Maltoni, T. Schwetz, and A. Zhou. “The fate of hints: updated global analysis of three-flavor neutrino oscillations”. In: *Journal of High Energy Physics* 2020.9 (2020). Issue 09, article 178., p. 178. DOI: 10.1007/JHEP09(2020)178.

[2357] B. D. Fields, K. A. Olive, T.-H. Yeh, and C. Young. “Big-Bang Nucleosynthesis after Planck”. In: *Journal of Cosmology and Astroparticle Physics* 2020.03 (2020). Issue 03, article 010. An erratum appeared as JCAP 11 (2020) E02., pp. 010–010. DOI: 10.1088/1475-7516/2020/03/010.

[2358] A. Grinin et al. “Two-photon frequency comb spectroscopy of atomic hydrogen”. In: *Science* 370.6520 (2020), pp. 1061–1066. DOI: 10.1126/science.abc7776.

[2359] L. Gurung, T. J. Babij, S. D. Hogan, and D. B. Cassidy. “Precision Microwave Spectroscopy of the Positronium n=2 Fine Structure”. In: *Physical Review Letters* 125.7 (2020), p. 073002. DOI: 10.1103/PhysRevLett.125.073002.

[2360] K. von Klitzing et al. “40 years of the quantum Hall effect”. In: *Nature Reviews Physics* 2.8 (2020), pp. 397–401. DOI: 10.1038/s42254-020-0209-1.

[2361] I. Larin et al. “Precision measurement of the neutral pion lifetime”. In: *Science* 368.6490 (2020), pp. 506–509. DOI: 10.1126/science.aay6641.

[2362] J. G. Lee, E. G. Adelberger, T. S. Cook, S. M. Fleischer, and B. R. Heckel. “New Test of the Gravitational 1/r2 Law at Separations down to 52 μm”. In: *Physical Review Letters* 124.10 (2020), p. 101101. DOI: 10.1103/PhysRevLett.124.101101.

[2363] L. Morel, Z. Yao, P. Cladé, and S. Guellati-Khélifa. “Determination of the fine-structure constant with an accuracy of 81 parts per trillion”. In: *Nature* 588.7836 (2020). Disagrees with Parker:2018 at more than 5σ; the conflict is unresolved and is absorbed into the CODATA uncertainty by an expansion factor, pp. 61–65. DOI: 10.1038/s41586-020-2964-7.

[2364] J. Nakamura, S. Liang, G. C. Gardner, and M. J. Manfra. “Direct observation of anyonic braiding statistics”. In: *Nature Physics* 16.9 (2020), pp. 931–936. DOI: 10.1038/s41567-020-1019-1.

[2365] A. Takenaka et al. “Search for proton decay via p$\to$e+π0 and p$\to$ μ+π0 with an enlarged fiducial volume in Super-Kamiokande I-IV”. In: *Physical Review D* 102.11 (2020), p. 112011. DOI: 10.1103/PhysRevD.102.112011.

[2366] A. Vinante et al. “Narrowing the Parameter Space of Collapse Models with Ultracold Layered Force Sensors”. In: *Physical Review Letters* 125.10 (2020), p. 100404. DOI: 10.1103/PhysRevLett.125.100404.

[2367] G. Aad et al. “Search for new phenomena in events with an energetic jet and missing transverse momentum in pp collisions at s=13 TeV with the ATLAS detector”. In: *Physical Review D* 103.11 (2021), p. 112006. DOI: 10.1103/PhysRevD.103.112006.

[2368] B. Abi et al. “Measurement of the Positive Muon Anomalous Magnetic Moment to 0.46 ppm”. In: *Physical Review Letters* 126.14 (2021), p. 141801. DOI: 10.1103/PhysRevLett.126.141801.

[2369] P. A. R. Ade et al. “Improved Constraints on Primordial Gravitational Waves using Planck , WMAP, and BICEP/ Keck Observations through the 2018 Observing Season”. In: *Physical Review Letters* 127.15 (2021). Source of the r < 0.036 bound; collaboration paper keyed by first listed author., p. 151301. DOI: 10.1103/PhysRevLett.127.151301.

[2370] D. Adhikari et al. “Accurate Determination of the Neutron Skin Thickness of Pb208 through Parity-Violation in Electron Scattering”. In: *Physical Review Letters* 126.17 (2021), p. 172502. DOI: 10.1103/PhysRevLett.126.172502.

[2371] M. Aizenman and H. Duminil-Copin. “Marginal triviality of the scaling limits of critical 4D Ising and $\lambda\phi_4^4$ models”. In: *Annals of Mathematics* 194.1 (2021), pp. 163–235. DOI: 10.4007/annals.2021.194.1.3.

[2372] K. Akiyama et al. “First M87 Event Horizon Telescope Results. VII. Polarization of the Ring”. In: *The Astrophysical Journal Letters* 910.1 (2021), p. L12. DOI: 10.3847/2041-8213/abe71d.

[2373] J. Amaré et al. “Annual modulation results from three-year exposure of ANAIS-112”. In: *Physical Review D* 103.10 (2021). Issue 10; article number 102005. ANAIS-112 Collaboration — collaboration in ‘note‘, not ‘author‘. The publisher prints the first-author surname with an acute accent, Amaré; the citekey folds it to ASCII as Amare:2021 per the house convention, and the .bib author field should carry Amaré, Cebrián. Second same-target (NaI(Tl)) replication that does not reproduce the DAMA/LIBRA modulation — cited alongside the existing Adhikari:2018 (COSINE-100) in phen:expdm-dama, where the prose says "experiments" in the plural but previously cited only one., p. 102005. DOI: 10.1103/PhysRevD.103.102005.

[2374] M. Asgari et al. “KiDS-1000 cosmology: Cosmic shear constraints and comparison between two point statistics”. In: *Astronomy & Astrophysics* 645 (2021), A104. DOI: 10.1051/0004-6361/202039070.

[2375] S. Borsanyi et al. “Leading hadronic contribution to the muon magnetic moment from lattice QCD”. In: *Nature* 593.7857 (2021), pp. 51–55. DOI: 10.1038/s41586-021-03418-1.

[2376] Z. Cao et al. “Ultrahigh-energy photons up to 1.4 petaelectronvolts from 12 γ-ray Galactic sources”. In: *Nature* 594.7861 (2021), pp. 33–36. DOI: 10.1038/s41586-021-03498-z.

[2377] S. Donadi, K. Piscicchia, C. Curceanu, L. Diósi, M. Laubenstein, and A. Bassi. “Underground test of gravity-related wave function collapse”. In: *Nature Physics* 17.1 (2021), pp. 74–78. DOI: 10.1038/s41567-020-1008-4.

[2378] E. Fonseca, H. T. Cromartie, T. T. Pennucci, et al. “Refined Mass and Geometric Measurements of the High-mass PSR J0740+6620”. In: *The Astrophysical Journal Letters* 915 (2021). Metadata confirmed against Crossref., p. L12. DOI: 10.3847/2041-8213/ac03b8.

[2379] W. L. Freedman. “Measurements of the Hubble Constant: Tensions in Perspective*”. In: *The Astrophysical Journal* 919.1 (2021), p. 16. DOI: 10.3847/1538-4357/ac0e95.

[2380] F. M. Gonzalez et al. “Improved Neutron Lifetime Measurement with UCNτ”. In: *Physical Review Letters* 127.16 (2021), p. 162501. DOI: 10.1103/PhysRevLett.127.162501.

[2381] M. Isi, W. M. Farr, M. Giesler, M. A. Scheel, and S. A. Teukolsky. “Testing the Black-Hole Area Law with GW150914”. In: *Physical Review Letters* 127.1 (2021), p. 011103. DOI: 10.1103/PhysRevLett.127.011103.

[2382] F. Kondev, M. Wang, W. Huang, S. Naimi, and G. Audi. “The NUBASE2020 evaluation of nuclear physics properties *”. In: *Chinese Physics C* 45.3 (2021). Issue 3, 1 March 2021; 030001 is an article number, not a page range. Crossref returns the title with a trailing footnote asterisk ('... properties *') which is not part of the title – drop it. NUBASE2020 is the canonical tabulation behind the half-lives and alpha decay energies of tab:tun-alpha; the chapter states the values are standard tabulated ones and explicitly disclaims Geiger:1911 as their source, so this is the entry that closes that gap., p. 030001. DOI: 10.1088/1674-1137/abddae.

[2383] M. Kramer, I. H. Stairs, R. N. Manchester, et al. “Strong-Field Gravity Tests with the Double Pulsar”. In: *Physical Review X* 11 (2021). Metadata confirmed against Crossref., p. 041050. DOI: 10.1103/PhysRevX.11.041050.

[2384] J. J. Krauth et al. “Measuring the α-particle charge radius with muonic helium-4 ions”. In: *Nature* 589.7843 (2021), pp. 527–531. DOI: 10.1038/s41586-021-03183-1.

[2385] R. Lange et al. “Improved Limits for Violations of Local Position Invariance from Atomic Clock Comparisons”. In: *Physical Review Letters* 126.1 (2021), p. 011102. DOI: 10.1103/PhysRevLett.126.011102.

[2386] M. C. Miller, F. K. Lamb, A. J. Dittmann, et al. “The Radius of PSR J0740+6620 from NICER and XMM-Newton Data”. In: *The Astrophysical Journal Letters* 918 (2021). Metadata confirmed against Crossref., p. L28. DOI: 10.3847/2041-8213/ac089b.

[2387] J. C. A. Miller-Jones et al. “Cygnus X-1 contains a 21–solar mass black hole—Implications for massive star winds”. In: *Science* 371.6533 (2021), pp. 1046–1049. DOI: 10.1126/science.abb3363.

[2388] C. A. J. O’Hare. “New Definition of the Neutrino Floor for Direct Dark Matter Searches”. In: *Physical Review Letters* 127 (2021). Issue 25; article number 251802. Single author. The redefinition from floor to fog — sensitivity degrades progressively rather than being forbidden — which is the exact claim it is attached to in the XENONnT derivation. Surname carries an apostrophe as printed: O’Hare., p. 251802.

[2389] T. E. Riley, A. L. Watts, P. S. Ray, et al. “A NICER View of the Massive Pulsar PSR J0740+6620 Informed by Radio Timing and XMM-Newton Spectroscopy”. In: *The Astrophysical Journal Letters* 918 (2021). Metadata confirmed against Crossref., p. L27. DOI: 10.3847/2041-8213/ac0a81.

[2390] A. M. Sirunyan et al. “Evidence for Higgs boson decay to a pair of muons”. In: *Journal of High Energy Physics* 2021.1 (2021). Issue 01, article 148., p. 148. DOI: 10.1007/JHEP01(2021)148.

[2391] J. A. Sobota, Y. He, and Z.-X. Shen. “Angle-resolved photoemission studies of quantum materials”. In: *Reviews of Modern Physics* 93.2 (2021), p. 025006. DOI: 10.1103/RevModPhys.93.025006.

[2392] E. Tiesinga, P. J. Mohr, D. B. Newell, and B. N. Taylor. “CODATA recommended values of the fundamental physical constants: 2018”. In: *Reviews of Modern Physics* 93 (2021), p. 025010. DOI: 10.1103/RevModPhys.93.025010.

[2393] M. Wang, W. Huang, F. Kondev, G. Audi, and S. Naimi. “The AME 2020 atomic mass evaluation (II). Tables, graphs and references*”. In: *Chinese Physics C* 45.3 (2021), p. 030003. DOI: 10.1088/1674-1137/abddaf.

[2394] G. Aad et al. “A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery”. In: *Nature* 607.7917 (2022), pp. 52–59. DOI: 10.1038/s41586-022-04893-w.

[2395] T. Aaltonen et al. “High-precision measurement of the W boson mass with the CDF II detector”. In: *Science* 376.6589 (2022), pp. 170–176. DOI: 10.1126/science.abk1781.

[2396] T. M. C. Abbott et al. “Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing”. In: *Physical Review D* 105.2 (2022). First author T. M. C. Abbott (DES), not B. P. Abbott (LIGO) of Abbott:2016 and Abbott:2017., p. 023520. DOI: 10.1103/PhysRevD.105.023520.

[2397] P. Abratenko et al. “Search for an Excess of Electron Neutrino Interactions in MicroBooNE Using Multiple Final-State Topologies”. In: *Physical Review Letters* 128.24 (2022), p. 241801. DOI: 10.1103/PhysRevLett.128.241801.

[2398] M. A. Acero et al. “Improved measurement of neutrino oscillation parameters by the NOvA experiment”. In: *Physical Review D* 106.3 (2022), p. 032004. DOI: 10.1103/PhysRevD.106.032004.

[2399] M. Aker et al. “Direct neutrino-mass measurement with sub-electronvolt sensitivity”. In: *Nature Physics* 18.2 (2022), pp. 160–166. DOI: 10.1038/s41567-021-01463-1.

[2400] M. J. Borchert et al. “A 16-parts-per-trillion measurement of the antiproton-to-proton charge–mass ratio”. In: *Nature* 601.7891 (2022), pp. 53–57. DOI: 10.1038/s41586-021-04203-w.

[2401] T. Bothwell et al. “Resolving the gravitational redshift across a millimetre-scale atomic sample”. In: *Nature* 602.7897 (2022), pp. 420–424. DOI: 10.1038/s41586-021-04349-7.

[2402] D. Brout et al. “The Pantheon+ Analysis: Cosmological Constraints”. In: *The Astrophysical Journal* 938.2 (2022), p. 110. DOI: 10.3847/1538-4357/ac8e04.

[2403] E. H. T. Collaboration et al. “First Sagittarius A* Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole in the Center of the Milky Way”. In: *The Astrophysical Journal Letters* 930.2 (2022), p. L12. DOI: 10.3847/2041-8213/ac6674.

[2404] D. P. Nadlinger et al. “Experimental quantum key distribution certified by Bell’s theorem”. In: *Nature* 607.7920 (2022), pp. 682–686. DOI: 10.1038/s41586-022-04941-5.

[2405] F. Napolitano et al. “Testing the Pauli Exclusion Principle with the VIP-2 Experiment”. In: *Symmetry* 14.5 (2022). VIP-2 continuation of the Ramberg–Snow search, run at the Gran Sasso underground laboratory. DOI supplied with moderate confidence; author list is long and given here in the VIP-2 collaboration order (first listed author keys the entry)., p. 893. DOI: 10.3390/sym14050893.

[2406] C. Overstreet, P. Asenbaum, J. Curti, M. Kim, and M. A. Kasevich. “Observation of a gravitational Aharonov-Bohm effect”. In: *Science* 375.6577 (2022), pp. 226–229. DOI: 10.1126/science.abl7152.

[2407] A. G. Riess et al. “A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km/s/Mpc Uncertainty from the Hubble Space Telescope and the SH0ES Team”. In: *The Astrophysical Journal Letters* 934 (2022), p. L7. DOI: 10.3847/2041-8213/ac5c5b.

[2408] The Royal Swedish Academy of Sciences. *Scientific Background on the Nobel Prize in Physics 2022: For experiments with entangled photons, establishing the violation of Bell inequalities and pioneering quantum information science*. Institutional report, no DOI; retrievable from nobelprize.org. Keyed on the academy (KVA) like BIPM:2019. 2022.

[2409] P. Touboul et al. "MICROSCOPE Mission: Final Results of the Test of the Equivalence Principle". In: *Physical Review Letters* 129.12 (2022). Collaboration paper keyed by first listed author., p. 121102. DOI: 10.1103/PhysRevLett.129.121102.

[2410] A. Tumasyan et al. "A portrait of the Higgs boson by the CMS experiment ten years after the discovery". In: *Nature* 607.7917 (2022). CMS Collaboration; the ten-year Higgs coupling portrait, the CMS companion to Aad:2022. A prior version of this entry carried the metadata of the ATLAS paper's one-page publisher correction (Nature 612, E24), a wrong-work trap caught in review. Metadata confirmed against Crossref., pp. 60–68. DOI: 10.1038/s41586-022-04892-x.

[2411] G. Aad et al. "Evidence of off-shell Higgs boson production from ZZ leptonic decay channels and constraints on its total width with the ATLAS detector". In: *Physics Letters B* 846 (2023), p. 138223. DOI: 10.1016/j.physletb.2023.138223.

[2412] R. Aaij et al. "Test of Lepton Universality in b→sℓ+ℓ− Decays". In: *Physical Review Letters* 131.5 (2023), p. 051803. DOI: 10.1103/PhysRevLett.131.051803.

[2413] J. Aalbers et al. "First Dark Matter Search Results from the LUX-ZEPLIN (LZ) Experiment". In: *Physical Review Letters* 131.4 (2023), p. 041002. DOI: 10.1103/PhysRevLett.131.041002.

[2414] S. Abe et al. "Search for the Majorana Nature of Neutrinos in the Inverted Mass Ordering Region with KamLAND-Zen". In: *Physical Review Letters* 130.5 (2023), p. 051801. DOI: 10.1103/PhysRevLett.130.051801.

[2415] R. Acharya et al. "Suppressing quantum errors by scaling a surface code logical qubit". In: *Nature* 614.7949 (2023). Collaboration paper keyed by the first listed author, per the house convention., pp. 676–681. DOI: 10.1038/s41586-022-05434-1.

[2416] D. P. Aguillard et al. "Measurement of the Positive Muon Anomalous Magnetic Moment to 0.20 ppm". In: *Physical Review Letters* 131.16 (2023), p. 161802. DOI: 10.1103/PhysRevLett.131.161802.

[2417] E. K. Anderson et al. "Observation of the effect of gravity on the motion of antimatter". In: *Nature* 621 (2023). ALPHA-g collaboration; keyed by first listed author. No collision with Anderson:1932/Anderson:1933 (different years), though the first author is a different Anderson., pp. 716–722.

[2418] E. Aprile et al. "First Dark Matter Search with Nuclear Recoils from the XENONnT Experiment". In: *Physical Review Letters* 131.4 (2023), p. 041003. DOI: 10.1103/PhysRevLett.131.041003.

[2419] X. Fan, T. G. Myers, B. A. D. Sukra, and G. Gabrielse. "Measurement of the Electron Magnetic Moment". In: *Physical Review Letters* 130.7 (2023), p. 071801. DOI: 10.1103/PhysRevLett.130.071801.

[2420] A. Kramida, Y. Ralchenko, J. Reader, and NIST ASD Team. *NIST Atomic Spectra Database (version 5.11)*. Living database; version 5.11 released 2023. The DOI is version-independent, so the note in references.bib must record the version and access date actually used for the mercury level energies. 2023.

[2421] T. S. Roussy et al. "An improved bound on the electron's electric dipole moment". In: *Science* 381.6653 (2023), pp. 46–50. DOI: 10.1126/science.adg4084.

[2422] S. Storz et al. “Loophole-free Bell inequality violation with superconducting circuits”. In: *Nature* 617.7960 (2023), pp. 265–270. DOI: 10.1038/s41586-023-05885-0.

[2423] H. Abu-Shawareb et al. “Achievement of Target Gain Larger than Unity in an Inertial Fusion Experiment”. In: *Physical Review Letters* 132.6 (2024). NIF ignition result exceeding the Lawson criterion. Collaboration paper keyed by the first listed author, per rule 5. The publisher prints the surname hyphenated, ’Abu-Shawareb’ (Crossref files it that way too). ’065102’ is an article number, not a page range. Issued 5 Feb 2024., p. 065102. DOI: 10.1103/PhysRevLett.132.065102.

[2424] K. Afanaciev et al. “A search for $\mu^+ \to e^+\gamma$ with the first dataset of the MEG II experiment”. In: *The European Physical Journal C* 84.3 (2024), p. 216. DOI: 10.1140/epjc/s10052-024-12416-2.

[2425] P. Concha, D. Pino, L. Ravera, and E. Rodríguez. “Extended kinematical 3D gravity theories”. In: *Journal of High Energy Physics* 2024.1 (2024), p. 40. DOI: 10.1007/JHEP01(2024)040.

[2426] F. V. Ignatov et al. “Measurement of the e+e− → π+π− cross section from threshold to 1.2 GeV with the CMD-3 detector”. In: *Physical Review D* 109.11 (2024), p. 112002. DOI: 10.1103/PhysRevD.109.112002.

[2427] S. Navas et al. “Review of Particle Physics”. In: *Physical Review D* 110 (2024). Particle Data Group, p. 030001. DOI: 10.1103/PhysRevD.110.030001.

[2428] J. Tiedau et al. “Laser Excitation of the Th-229 Nucleus”. In: *Physical Review Letters* 132.18 (2024), p. 182501. DOI: 10.1103/PhysRevLett.132.182501.

[2429] C. Zhang et al. “Frequency ratio of the 229mTh nuclear isomeric transition and the 87Sr atomic clock”. In: *Nature* 633.8028 (2024), pp. 63–70. DOI: 10.1038/s41586-024-07839-6.

[2430] A. G. Adame and others (DESI Collaboration). “DESI 2024 VI: cosmological constraints from the measurements of baryon acoustic oscillations”. In: *Journal of Cosmology and Astroparticle Physics* 2025 (2025). Customarily cited as 2024 after its report title and preprint; the journal issue is 2025 (JCAP 02). Keyed by the publisher’s year per the house rule. DOI omitted pending confirmation., p. 021.

[2431] P. J. Mohr, D. B. Newell, B. N. Taylor, and E. Tiesinga. “CODATA recommended values of the fundamental physical constants: 2022”. In: *Reviews of Modern Physics* 97.2 (2025). Universally called “CODATA 2022” after its data-closing date of 31 December 2022; published 2025, which fixes the key, exactly as Tiesinga:2021 keys the 2018 report. Co-published as J. Phys. Chem. Ref. Data 54 (2025) 033105. The recommended α carries an uncertainty inflated by an expansion factor of 2.5 to cover the unresolved conflict between Parker:2018 and Morel:2020, p. 025002. DOI: 10.1103/RevModPhys.97.025002.

[2432] P. Concha, N. Gallegos, E. Rodríguez, and S. Salgado. *Maxwell kinematical algebras and 3D gravities*. Preprint, submitted 24 February 2026; not peer-reviewed at the time of writing. 2026. DOI: 10.48550/arXiv.2602.21038. arXiv: 2602.21038 [hep-th].

[2433] N. Gonzalez. “An explicit isomorphism between property-graph and relational presentations of a validated knowledge base, with a treatise of physics as worked example”. Manuscript in preparation; not publicly available at the time of writing. The worked example is the Treatise of Physics, https://nbow.io/en/products/ai-powered-research/treatise-of-physics/. Not to be confused with Gonzalez:2026, on Lie algebra expansions. 2026.

[2434] N. Gonzalez and S. Salgado. “Classification of tensor product Lie algebras from $\mathfrak{so}(p, q+1)$”. Manuscript in preparation; not publicly available at the time of writing. Cited for the organising construction only: every result taken from it is reproved from the definitions in the chapter on expansions of Lie algebras and its appendix. Not to be confused with Gonzalez:2026a, on the graph–relational correspondence. 2026.